2020/21 CSC 5741: Data Mining and Warehousing Jupyter Notebook—Exploratory Data Analysis

Lighton Phiri
<lighton.phiri@unza.zm>

May 24, 2021

Contents

Introduction

In this Jupyter Notebook, we walk through practical examples in order to illustrate how to perform Exploratory Data Analysis (EDA). In all instances, you will notice two key operations:

1. Basic descriptive statistical analysis
2. Extensive use of plots, graphs and/or charts

While the pre-processing activity was descussed in the previous lecture series, we also include "some" aspects of it, to serve as a remind of tasks to be performed.

You will notice that the some examples use native Python features as opposed to libraries such as Pandas. This is done to highlight the flexibility that Python provides. In cases were they are not used, you are encouraged to explore how Pandas and other libraries can be used.

In all instances, you are encouraged to make reference to online documentation for the various tools. Additionally, you can exploit tools like Zeal Offline Documentation Browser to download and search through offline documentation. You are also encouraged to look up and explore other libraries, especially as you work towards the Mini Projects.

General Notebook Configuration

[1]:
```
# Aesthetics for pandas cell outsput
import pandas as pd

pd.set_option('display.latex.repr', True)
pd.set_option('display.latex.longtable', True)
pd.set_option('max_colwidth', 30)

# Show all Jupyter Notebook cell output
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

Python Packages for Data Pre-processing

[2]:
```
# Import all libraries and modules for use during lecture session code walkthrough
import matplotlib.pyplot as plt
import pandas as pd
import re
import seaborn as sns
import string

from collections import Counter
from IPython.core.interactiveshell import InteractiveShell
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from wordcloud import WordCloud
```

Implementing Core Functions

The generic functions in this section act as general utility functions, primarily for pre-processing. However, some of them perform specialised tasks.

[3]:
```
def fxn_case_folding(var_input):
    """
    Preprocessing: Case Folding
    """
    return var_input.lower()

def fxn_punctuation(var_input_text):
    """
    Preprocessing: Punctuation Removal
    """
    var_output_text = re.sub("[%s]" % re.escape(string.punctuation), " ",␣
 ↪var_input_text)
    var_output_text = re.sub("[%s]" % re.escape(string.punctuation), " ",␣
 ↪var_output_text)
    var_output_text = re.sub('\w*\d\w*', '', var_output_text) # HINT: lookup isalpha()␣
 ↪function
    return var_output_text

def fxn_stopwords(var_input_text):
    """
    Preprocessing: Stopwords Removal
    """
    var_etd_stop = " ".join([
        var_etd_word for var_etd_word in var_input_text.split()
        if var_etd_word not in stopwords.words('english')
    ])
    return var_etd_stop

def fxn_stem(var_input_text):
```

```
    """
    Preprocessing: Stemming
    """
    var_stemmer = PorterStemmer()
    var_output_text = " ".join([
        var_stemmer.stem(var_etd_word) for var_etd_word in var_input_text.split()
    ])
    return var_output_text

def fxn_normalise_ict1110_minors(var_input_minor):
    """
    Returns normalised ICT 1110 minor
    """
    var_ict1110_minors = ["Geography", "History", "Languages", "Mathematics", "Civic",␣
 ↪"Art", "Religious Studies"]
    if "civic" in var_input_minor.lower():
        var_output_minor = "Civic Education"
    elif "religious" in var_input_minor.lower() or "res" in var_input_minor.lower():
        var_output_minor = ""
    elif "history" in var_input_minor.lower():
        var_output_minor = "History"
    elif "art" in var_input_minor.lower():
        var_output_minor = "Art"
    elif "language" in var_input_minor.lower() or "french" in var_input_minor.lower():
        var_output_minor = "Languages"
    elif "geography" in var_input_minor.lower():
        var_output_minor = "Geography"
    elif "math" in var_input_minor.lower():
        var_output_minor = "Mathematics"
    elif "writing" in var_input_minor.lower():
        var_output_minor = "Writing Skill"
    else:
        var_output_minor = var_input_minor
    return var_output_minor.title()
```

[4]:
```
var_example_string = "This is an example string, used as part of CSC 5741 code␣
 ↪snippets."
```

[5]:
```
##fxn_stopwords(var_example_string)
fxn_case_folding(var_example_string)
fxn_stopwords(fxn_case_folding(var_example_string))
fxn_punctuation(fxn_stopwords(fxn_case_folding(var_example_string)))
fxn_stem(fxn_punctuation(fxn_stopwords(fxn_case_folding(var_example_string))))
```

[5]: 'this is an example string, used as part of csc 5741 code snippets.'

[5]: 'example string, used part csc 5741 code snippets.'

[5]: 'example string used part csc code snippets '

[5]: 'exampl string use part csc code snippet'

Dataset #1: ICT 1110 Information Survey

Data Preprocessing

Link to dataset

Students at enrolled into the "ICT 1110: Computer Systems and Architecture" course, at The University of Zambia, respond to a preliminary survey aimed at collecting background information about them. This is done using Google Forms.

Dataset Description

This dataset comprises of 25 student responses for the 2018/19 cohort and 73 responses for the 2019/20 cohort. The dataset has observations presented in CSV format, using "|" as the separator. In addition, each observation is associated with the following 13 data attributes: * Timestamp * Full Names * Student ID * Hometown (surburb/town/province—e.g. Kabwata/Lusaka/Lusaka) * What is your programme Minor (e.g. Mathematics, Languages) * What made you decide on your programme minor? * Why did you decide to major pursue the B.ICTs Ed. Programme? * Did you study Computer Studies at secondary school? * Have you undergone any computer related training? * If your response to the question above is year, please provide details of the type of course and/or training

[6]:
```
# Explore 2018/19 ICT 1110 survey
!cat -n db-unza21-csc5741-ict1110_2018_19-preliminary_survey.csv | head
```

```
     1  Timestamp|Full Names|Student ID|Hometown (surburb/town/province---e.g.
Kabwata/Lusaka/Lusaka)|What is your programme Minor (e.g. Mathematics, Languages)|What
made you decide on your programme minor?|Why did you decide to major pursue the B.ICTs
Ed. Programme?|Did you study Computer Studies at secondary school?|Have you undergone any
computer related training?|If your response to the question above is year, please provide
details of the type of course and/or training|How many years experience do you have using
computers?|Do you currently own a computer or have regular access to one?|List one
interesting fact about yourself (e.g. I cycle everyday!):
     2  2019/03/28 11:13:51 PM GMT+2|Participant1|#N/A|Chudleigh/Lusaka/Lusaka|Data
Mining|I love data|I love computers|No|Yes|I have studied Computer Science|More than 5
years|Yes|I cycle everyday!
     3  2019/03/28 11:55:27 PM GMT+2|Participant2|742b8abe5776a6d942a92ce7dc7d84a0|Copper
belt,luanshya,Mpatamato|Mathematics|I find it easy to study and understand|Wanted to
acquire more knowledge about ICTs and contribute to technology|No|No||1 to 2 years|Yes|A
day doesn't pass by without a joke,I feel laughing will make you feel like you are in
another world
     4  2019/03/29 8:00:53 PM
GMT+2|Participant3|921855f753932de762b780405a50bdf7|Mungule,senanga,western.|French|It
was the best of my available options |"i have always wanted to do an
     5  ICT related program."|No|No||No Experience|Yes|
     6  2019/03/30 11:25:30 AM
GMT+2|Participant4|07f3ca235faaa1c9ad16facef5526d8b|Lusaka|Religious studies|I just chose
it|Because my results met the requirements |No|No||Less than 1 year|Yes|I like the
internet
     7  2019/03/31 3:26:35 AM GMT+2|Participant5|4234d1794dd33c1b6ed975eab5148040|Lusaka
|Civic education |My first option was Chinese but it was a major and came with additional
courses increasing my courses to more than four. So I ended up picking civic education
because I found it easy in high school |I had written the same program twice on my
application form so the man collecting suggested B. ICTs Ed|No|No||Less than 1 year|No|I
```

```
enjoy indie music
     8  2019/03/31 1:49:53 PM
GMT+2|Participant6|9e7002d53d4db7bfad4f5cf419b0c126|shibuyunji/ central pronvince|civic
education|i want to know more of my rights and responsibilities as a zambian citizen and
take part in passing the knowledge to those who do not know much about their role in
democratic governance.|so that i can be part and parcel of the ever changing and
developing digital world, and to take part in the zambia 2030 vision of having a digital
zambia in all sectors of development.|No|No||No Experience|Yes|like exploring on IT
technology
     9  2019/03/31 1:51:38 PM
GMT+2|Participant7|9e7002d53d4db7bfad4f5cf419b0c126|shibuyunji/ central pronvince|civic
education|i want to know more of my rights and responsibilities as a zambian citizen and
take part in passing the knowledge to those who do not know much about their role in
democratic governance.|so that i can be part and parcel of the ever changing and
developing digital world, and to take part in the zambia 2030 vision of having a digital
zambia in all sectors of development.|No|No||No Experience|Yes|like exploring on IT
technology
    10  2019/04/01 7:12:07 PM
GMT+2|Participant8|fceb5af40df295d85851f390f4f8d78d|LUSAKA|RELIGIOUS STUDIES|TO HAVE
KNOWLEDGE APPRECIATE OTHER RELIGIONS|I HAVE ALWAYS WANTED STUDY THIS PROGRAM |No|No||Less
than 1 year|Yes|I ACCESS YOUTUBE ALMOST EVERYDAY
```

[7]:
```
# Count the number of observations in the 2018/19 ICT 1110 survey
!wc db-unza21-csc5741-ict1110_2018_19-preliminary_survey.csv
```

```
  43  1916 15361 db-unza21-csc5741-ict1110_2018_19-preliminary_survey.csv
```

[8]:
```
# Explore 2019/20 ICT 1110 survey
!cat -n db-unza21-csc5741-ict1110_2019_20-preliminary_survey.csv | head
```

```
     1  Timestamp|Full Names|Student ID|Hometown (surburb/town/province---e.g.
Kabwata/Lusaka/Lusaka)|What is your programme Minor (e.g. Mathematics, Languages)|What
made you decide on your programme minor?|Why did you decide to major pursue the B.ICTs
Ed. Programme?|Did you study Computer Studies at secondary school?|Have you undergone any
computer related training?|If your response to the question above is year, please provide
details of the type of course and/or training|How many years experience do you have using
computers?|Do you currently own a computer or have regular access to one?|List one
interesting fact about yourself (e.g. I cycle everyday!):
     2  2020/03/04 7:57:29 PM GMT+2|Participant43|aa34dad971bfc1edc090076ef05be225|Libala
lusaka|Mathematics |It proves to be a good combination and one of my strengths |Always
had interest  in technology |No|No||3 to 5 years|Yes|I can sing
     3  2020/03/04 7:57:35 PM GMT+2|Participant43|aa34dad971bfc1edc090076ef05be225|Libala
lusaka|Mathematics |It proves to be a good combination and one of my strengths |Always
had interest  in technology |No|No||3 to 5 years|Yes|I can sing
     4  2020/03/05 3:26:13 PM
GMT+2|Participant12|a41f2bfc1a4c22c8e0aaf518f42bac0b|Rhodespark/Lusaka/Lusaka|Mathematics
|A prefence of courses that involve more of solving to studying. |Interest in technology
and dream of being a established IT person. |No|No||More than 5 years|Yes|I love to keep
up to date with the latest tech gadgets that keep being released.
     5  2020/03/05 8:31:16 PM GMT+2|Participant36|ab7f1643a776fc9319422859c8869fd7|State
Lodge|Mathematics|Because I like challenging thing and I love numbers|Because I wanted to
know how computer works, know how to manipulate it and because when I was I kid always
```

```
wanted to use a computer but never had one home.|No|Yes|Office package|1 to 2 years|Yes|I
play the keyboard and little bit of singing
     6  2020/03/05 8:50:28 PM
GMT+2|Participant6|e997272aa3c790288c782a99f0e96b1d|Lusaka/Lusaka|Religious Studies|I
would want to know more about world Religion and understand more on how the study of
Religion can be the way of sorting out conflicts among Religious groups in
society.|Because it was one of my favourite in secondary, though I only learnt it in
grade 12 only...therefore, I would want to know more about it because we are now living
in the Digital world.|Yes|No||No Experience|No|I read the Novels
     7  2020/03/05 10:11:06 PM
GMT+2|Participant28|38e58148b7bc4c09777e880ec00d7aaf|Lusaka|Mathematics|The course seemed
to be a good combination with my major i.e ICT|Its interesting to me personally and its
diversity in the courses it has, is what inspires me the most. I get to choose anywhere
to work from and most vital, working for my self is also an option. |No|No||More than 5
years|Yes|"
     8  iLoveToExplore <Curiosity>"
     9  2020/03/06 1:08:04 PM
GMT+2|Participant5|4f4c3be8fb6ec9d0b157764c4abafd70|Mandevu/Lusaka/Lusaka |Civic
Education 1100|I thought it's a right course that can able me become one of agent of
development in Zambia, as well as being a voice of many other Zambians who can't come out
loud and so on.|Looking at Zambias plans to have a project on technology development so
by pursuing this program I'm very much sure that I will be helped to have knowledge as
information about technology studies, I'm hoping i will be one of the students to run the
developmental project.|No|No||Less than 1 year|No|I'm not good a speaker, but I'm very
fast to grasp information.
    10  2020/03/06 2:19:24 PM
GMT+2|Participant47|7f32c2c7fab80b6433f8818b8162c79c|Hillcrest/Ndola/Copperbelt
|Languages |To improve my vocabulary |To become a developer |No|No||More than 5
years|No|I repair/fix hardware and software to calm myself
```

[9]:
```
# Count the number of observations in the 2019/20 ICT 1110 survey
# NOTE: Careful with shell commands as processing CSV files can be problematic, e.g.␣
 ↪when counting records
#
!cat db-unza21-csc5741-ict1110_2019_20-preliminary_survey.csv | wc -l
```

```
91
```

[10]:
```
# Merge the 2018/19 and 2019/20 ICT 1110 survey results
#
# Spool contents of 2018/19 into input file
!cat db-unza21-csc5741-ict1110_2018_19-preliminary_survey.csv >␣
 ↪db-unza21-csc5741-ict1110_preliminary_survey.csv
```

[11]:
```
# Spool contents of 2019/20 into input file
# READ: https://stackoverflow.com/a/339941/664424
# One a UNIX-like OS, issue "man tail"
!tail +2 db-unza21-csc5741-ict1110_2019_20-preliminary_survey.csv >>␣
 ↪db-unza21-csc5741-ict1110_preliminary_survey.csv
```

[12]:
```
# Explore resulting merged input file
# NOTE: Careful with shell commands as processing CSV files can be problematic, e.g. when counting records
#
!cat db-unza21-csc5741-ict1110_preliminary_survey.csv | wc -l
```

```
133
```

Dataframe Creation

[13]:
```
# Create DataFrame of input dataset: ICT 1110 survey
#
var_ict1110_survey = pd.read_csv("db-unza21-csc5741-ict1110_preliminary_survey.csv", sep="|")
var_ict1110_survey.columns
```

[13]:
```
Index(['Timestamp', 'Full Names', 'Student ID',
       'Hometown (surburb/town/province---e.g. Kabwata/Lusaka/Lusaka)',
       'What is your programme Minor (e.g. Mathematics, Languages)',
       'What made you decide on your programme minor?',
       'Why did you decide to major pursue the B.ICTs Ed. Programme?',
       'Did you study Computer Studies at secondary school?',
       'Have you undergone any computer related training?',
       'If your response to the question above is year, please provide details of the type of course and/or training',
       'How many years experience do you have using computers?',
       'Do you currently own a computer or have regular access to one?',
       'List one interesting fact about yourself (e.g. I cycle everyday!):'],
      dtype='object')
```

[14]:
```
# Rename dataframe columns for easy processing
#
var_ict1110_survey.rename(columns={"Full Names": "StudentName",
                                   "Student ID": "StudentID",
                                   "Hometown (surburb/town/province---e.g. Kabwata/Lusaka/Lusaka)": "HomeTown",
                                   "What is your programme Minor (e.g. Mathematics, Languages)": "MinorProgramme",
                                   "What made you decide on your programme minor?": "MinorProgrammeMotivation",
                                   "Why did you decide to major pursue the B.ICTs Ed. Programme?": "MajorProgrammeMotivation",
                                   "Did you study Computer Studies at secondary school?": "DidComputerStudies",
                                   "Have you undergone any computer related training?": "HasComputerTraining",
                                   "If your response to the question above is year, please provide details of the type of course and/or training": "ComputerTrainingType",
                                   "How many years experience do you have using computers?": "ExperienceWithComputers",
```

```
                                    "Do you currently own a computer or have regular␣
↪access to one?": "HasComputerAccess",
                                     "List one interesting fact about yourself (e.g. I␣
↪cycle everyday!):": "AboutMe"}, inplace=True)

var_ict1110_survey.columns
```

[14]:
```
Index(['Timestamp', 'StudentName', 'StudentID', 'HomeTown', 'MinorProgramme',
       'MinorProgrammeMotivation', 'MajorProgrammeMotivation',
       'DidComputerStudies', 'HasComputerTraining', 'ComputerTrainingType',
       'ExperienceWithComputers', 'HasComputerAccess', 'AboutMe'],
      dtype='object')
```

[15]:
```
# Count records in dataframe
#
len(var_ict1110_survey)
```

[15]:
```
112
```

[16]:
```
# Inspect some of the records
#
var_ict1110_survey.head(3).T
```

[16]:

	0	1	2
Timestamp	2019/03/28 11:13:51 PM GMT+2	2019/03/28 11:55:27 PM GMT+2	2019/03/29 8
StudentName	Participant1	Participant2	Participant3
StudentID	NaN	742b8abe5776a6d942a92ce7dc...	921855f75393
HomeTown	Chudleigh/Lusaka/Lusaka	Copperbelt,luanshya,Mpatamato	Mungule,sen
MinorProgramme	Data Mining	Mathematics	French
MinorProgrammeMotivation	I love data	I find it easy to study an...	It was the be
MajorProgrammeMotivation	I love computers	Wanted to acquire more kno...	i have always
DidComputerStudies	No	No	No
HasComputerTraining	Yes	No	No
ComputerTrainingType	I have studied Computer Sc...	NaN	NaN
ExperienceWithComputers	More than 5 years	1 to 2 years	No Experien
HasComputerAccess	Yes	Yes	Yes
AboutMe	I cycle everyday!	A day doesn't pass by with...	NaN

Dataset Attributes

- Timestamp—Date
- StudentName—Text
- StudentID—Alphanumeric
- Hometown—Text
- MinorProgramme—Text
- MinorProgrammeMotivation—Text
- MajorProgrammeMotivation—Text
- DidComputerStudies—Categorical
- HasComputerTraining—Categorical
- ComputerTrainingType—Text

- ExperienceWithComputers—Ordinal
- HasComputerAccess—Categorical
- AboutMe—Text

Data Pre-processing Plan

- STEP 1: Remove duplicates using StudentID as unique field
- STEP 2: Apply case folding to all text attributes
- STEP 3: Remove punctuations from text attributes
- STEP 4: Remove stopwords from text attributes
- STEP 5: Stem all text attributes

NOTE: Null values to be handled on a case-by-case basis; e.g. null values in text attributes to be replaced with empty strings ""

[17]:
```
# STEP 1: Remove duplicates using StudentID as unique field
#
# Print duplicate records on StudentID
#
# Using Pandas, the df.duplicated() function can be used to identify duplicates
# http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.duplicated.
 ↪html
# var_ict1110_survey.duplicated(["StudentID"], keep=False)
#
var_ict1110_survey[var_ict1110_survey.duplicated(["StudentID"], keep=False)].
 ↪sort_values("StudentID").tail(2).T
```

[17]:

	0	17
Timestamp	2019/03/28 11:13:51 PM GMT+2	2019/04/03 1:10:20 PM GMT+2
StudentName	Participant1	Participant18
StudentID	NaN	NaN
HomeTown	Chudleigh/Lusaka/Lusaka	SINDA
MinorProgramme	Data Mining	MATHEMATICS
MinorProgrammeMotivation	I love data	I ENJOY IT
MajorProgrammeMotivation	I love computers	BECAUSE ITS MARKETABLE
DidComputerStudies	No	No
HasComputerTraining	Yes	No
ComputerTrainingType	I have studied Computer Sc...	NaN
ExperienceWithComputers	More than 5 years	No Experience
HasComputerAccess	Yes	No
AboutMe	I cycle everyday!	I LIKE SINGING

[18]:
```
# Remove duplicate records on StudentID
#
# drop_duplicates uses keep=[First, Last, False]
#
var_ict1110_survey.drop_duplicates(["StudentID"], keep="first", inplace=True)
```

[19]:
```
# Confirm duplicate records removal
len(var_ict1110_survey)
```

[19]:
```
90
```

[20]:
```
# STEP 2: Apply steps 2--5 to Text Attributes recursively
# Attributes: MinorProgrammeMotivation, MajorProgrammeMotivation and AboutMe
#
# Inspect dataframe before pre-processing
var_ict1110_survey.head(2).T
```

[20]:

	0	1
Timestamp	2019/03/28 11:13:51 PM GMT+2	2019/03/28 11:55:27 PM GMT+2
StudentName	Participant1	Participant2
StudentID	NaN	742b8abe5776a6d942a92ce7dc...
HomeTown	Chudleigh/Lusaka/Lusaka	Copperbelt,luanshya,Mpatamato
MinorProgramme	Data Mining	Mathematics
MinorProgrammeMotivation	I love data	I find it easy to study an...
MajorProgrammeMotivation	I love computers	Wanted to acquire more kno...
DidComputerStudies	No	No
HasComputerTraining	Yes	No
ComputerTrainingType	I have studied Computer Sc...	NaN
ExperienceWithComputers	More than 5 years	1 to 2 years
HasComputerAccess	Yes	Yes
AboutMe	I cycle everyday!	A day doesn't pass by with...

[21]:
```
# Handle null values---NaN
#
var_ict1110_survey["MinorProgrammeMotivation"].fillna("", inplace=True)

var_ict1110_survey["MajorProgrammeMotivation"].fillna("", inplace=True)

var_ict1110_survey["AboutMe"].fillna("", inplace=True)
```

[22]:
```
# Apply Case Folding
#
var_ict1110_survey["MinorProgrammeMotivation"] =␣
 ↪var_ict1110_survey["MinorProgrammeMotivation"].apply(fxn_case_folding)

var_ict1110_survey["MajorProgrammeMotivation"] =␣
 ↪var_ict1110_survey["MajorProgrammeMotivation"].apply(fxn_case_folding)

var_ict1110_survey["AboutMe"] = var_ict1110_survey["AboutMe"].apply(fxn_case_folding)
```

[23]:
```
# Inspect dataframe after pre-processing
var_ict1110_survey.head(2).T
```

[23]:

	0	1
Timestamp	2019/03/28 11:13:51 PM GMT+2	2019/03/28 11:55:27 PM GMT+2
StudentName	Participant1	Participant2

Continued on next page

	0	1
StudentID	NaN	742b8abe5776a6d942a92ce7dc...
HomeTown	Chudleigh/Lusaka/Lusaka	Copperbelt,luanshya,Mpatamato
MinorProgramme	Data Mining	Mathematics
MinorProgrammeMotivation	i love data	i find it easy to study an...
MajorProgrammeMotivation	i love computers	wanted to acquire more kno...
DidComputerStudies	No	No
HasComputerTraining	Yes	No
ComputerTrainingType	I have studied Computer Sc...	NaN
ExperienceWithComputers	More than 5 years	1 to 2 years
HasComputerAccess	Yes	Yes
AboutMe	i cycle everyday!	a day doesn't pass by with...

[24]:
```
# Apply Punction Removal
#
var_ict1110_survey["MinorProgrammeMotivation"] =␣
 ↪var_ict1110_survey["MinorProgrammeMotivation"].apply(fxn_punctuation)

var_ict1110_survey["MajorProgrammeMotivation"] =␣
 ↪var_ict1110_survey["MajorProgrammeMotivation"].apply(fxn_punctuation)

var_ict1110_survey["AboutMe"] = var_ict1110_survey["AboutMe"].apply(fxn_punctuation)
```

[25]:
```
# Inspect dataframe after pre-processing
var_ict1110_survey.head(2).T
```

[25]:

	0	1
Timestamp	2019/03/28 11:13:51 PM GMT+2	2019/03/28 11:55:27 PM GMT+2
StudentName	Participant1	Participant2
StudentID	NaN	742b8abe5776a6d942a92ce7dc...
HomeTown	Chudleigh/Lusaka/Lusaka	Copperbelt,luanshya,Mpatamato
MinorProgramme	Data Mining	Mathematics
MinorProgrammeMotivation	i love data	i find it easy to study an...
MajorProgrammeMotivation	i love computers	wanted to acquire more kno...
DidComputerStudies	No	No
HasComputerTraining	Yes	No
ComputerTrainingType	I have studied Computer Sc...	NaN
ExperienceWithComputers	More than 5 years	1 to 2 years
HasComputerAccess	Yes	Yes
AboutMe	i cycle everyday	a day doesn t pass by with...

[26]:
```
# Apply Stopwords Removal
#
var_ict1110_survey["MinorProgrammeMotivation"] =␣
 ↪var_ict1110_survey["MinorProgrammeMotivation"].apply(fxn_stopwords)

var_ict1110_survey["MajorProgrammeMotivation"] =␣
 ↪var_ict1110_survey["MajorProgrammeMotivation"].apply(fxn_stopwords)
```

```
var_ict1110_survey["AboutMe"] = var_ict1110_survey["AboutMe"].apply(fxn_stopwords)
```

[27]:
```
# Inspect dataframe after pre-processing
var_ict1110_survey.head(2).T
```

[27]:

	0	1
Timestamp	2019/03/28 11:13:51 PM GMT+2	2019/03/28 11:55:27 PM GMT+2
StudentName	Participant1	Participant2
StudentID	NaN	742b8abe5776a6d942a92ce7dc...
HomeTown	Chudleigh/Lusaka/Lusaka	Copperbelt,luanshya,Mpatamato
MinorProgramme	Data Mining	Mathematics
MinorProgrammeMotivation	love data	find easy study understand
MajorProgrammeMotivation	love computers	wanted acquire knowledge i...
DidComputerStudies	No	No
HasComputerTraining	Yes	No
ComputerTrainingType	I have studied Computer Sc...	NaN
ExperienceWithComputers	More than 5 years	1 to 2 years
HasComputerAccess	Yes	Yes
AboutMe	cycle everyday	day pass without joke feel...

[28]:
```
# Apply Stem Removal
#
var_ict1110_survey["MinorProgrammeMotivation"] =␣
↪var_ict1110_survey["MinorProgrammeMotivation"].apply(fxn_stem)

var_ict1110_survey["MajorProgrammeMotivation"] =␣
↪var_ict1110_survey["MajorProgrammeMotivation"].apply(fxn_stem)

var_ict1110_survey["AboutMe"] = var_ict1110_survey["AboutMe"].apply(fxn_stem)
```

[29]:
```
# Inspect dataframe after pre-processing
var_ict1110_survey.head(2).T
```

[29]:

	0	1
Timestamp	2019/03/28 11:13:51 PM GMT+2	2019/03/28 11:55:27 PM GMT+2
StudentName	Participant1	Participant2
StudentID	NaN	742b8abe5776a6d942a92ce7dc...
HomeTown	Chudleigh/Lusaka/Lusaka	Copperbelt,luanshya,Mpatamato
MinorProgramme	Data Mining	Mathematics
MinorProgrammeMotivation	love data	find easi studi understand
MajorProgrammeMotivation	love comput	want acquir knowledg ict c...
DidComputerStudies	No	No
HasComputerTraining	Yes	No
ComputerTrainingType	I have studied Computer Sc...	NaN
ExperienceWithComputers	More than 5 years	1 to 2 years
HasComputerAccess	Yes	Yes

Continued on next page

	0	1
AboutMe	cycl everyday	day pass without joke feel...

Exploratory Data Analysis

[30]:
```
# Describe the data
# Identify attributes to explore
var_ict1110_survey.head(3).T
```

[30]:

	0	1	2
Timestamp	2019/03/28 11:13:51 PM GMT+2	2019/03/28 11:55:27 PM GMT+2	2019/03/29 8
StudentName	Participant1	Participant2	Participant3
StudentID	NaN	742b8abe5776a6d942a92ce7dc...	921855f75393
HomeTown	Chudleigh/Lusaka/Lusaka	Copperbelt,luanshya,Mpatamato	Mungule,sen
MinorProgramme	Data Mining	Mathematics	French
MinorProgrammeMotivation	love data	find easi studi understand	best avail op
MajorProgrammeMotivation	love comput	want acquir knowledg ict c...	alway want i
DidComputerStudies	No	No	No
HasComputerTraining	Yes	No	No
ComputerTrainingType	I have studied Computer Sc...	NaN	NaN
ExperienceWithComputers	More than 5 years	1 to 2 years	No Experien
HasComputerAccess	Yes	Yes	Yes
AboutMe	cycl everyday	day pass without joke feel...	

Possible attributes to include in the EDA process

- Home Town
- Minor Programme
- Minor Programme Motivation
- Major Programme Motivation
- Computer Studies Elective in High School
- Prior Computing Training
- Prior Computing Training Type
- Experience Working With Computers
- Access to a Computer

[31]:
```
# Define variable to
var_ict1110_survey_eda = var_ict1110_survey
```

[32]:
```
# Create new column to mark different academic years
#
var_ict1110_survey_eda["year"] = var_ict1110_survey_eda["Timestamp"].str[:4]
```

Dataframe Statistical Information

[33]:
```
# Use describe function to get statistical information of dataset attributes
var_ict1110_survey_eda.describe(include='all').T
```

[33]:

	count	unique	top	freq
Timestamp	90	90	2020/03/16 6:44:54 PM GMT+2	1
StudentName	88	55	Participant21	2
StudentID	89	89	e2a96e074e1d8a6f6de56abbd4...	1
HomeTown	90	74	Lusaka	11
MinorProgramme	90	46	Mathematics	12
MinorProgrammeMotivation	90	89	passion	2
MajorProgrammeMotivation	90	88	love technolog	2
DidComputerStudies	90	2	No	78
HasComputerTraining	90	2	No	71
ComputerTrainingType	21	19	Computer networking and ha...	2
ExperienceWithComputers	90	5	No Experience	31
HasComputerAccess	90	2	Yes	64
AboutMe	90	84	sing	3
year	90	2	2020	57

Minor Programme

[34]:
```
# Minor Programme
#
# Apply attribute specific processing
var_ict1110_survey_eda["MinorProgramme"] = var_ict1110_survey_eda["MinorProgramme"].
 ↪apply(fxn_case_folding)

var_ict1110_survey_eda["MinorProgramme"] = var_ict1110_survey_eda["MinorProgramme"].
 ↪apply(fxn_punctuation)

var_ict1110_survey_eda["MinorProgramme"] = var_ict1110_survey_eda["MinorProgramme"].
 ↪apply(fxn_normalise_ict1110_minors)
```

[35]:
```
# Basic dataframe summaries
#
# Get unique entries
var_ict1110_survey_eda["MinorProgramme"].unique()

# Count observations
var_ict1110_survey_eda["MinorProgramme"].count()

# Get value counts
var_ict1110_survey_eda["MinorProgramme"].value_counts()
```

[35]:
```
array(['Data Mining', 'Mathematics', 'Languages', '', 'Civic Education',
       'History', 'Art', 'Geography', 'Writing Skill',
       'Drawing And Painting ', 'English', 'Ict ',
       'Physical Education And Sports Pes '], dtype=object)
```

[35]:
```
90
```

[35]:

	MinorProgramme
Civic Education	28
Mathematics	24
Languages	10
	9
History	8
Art	3
Geography	2
Physical Education And Spor...	1
Writing Skill	1
Drawing And Painting	1
Data Mining	1
Ict	1
English	1

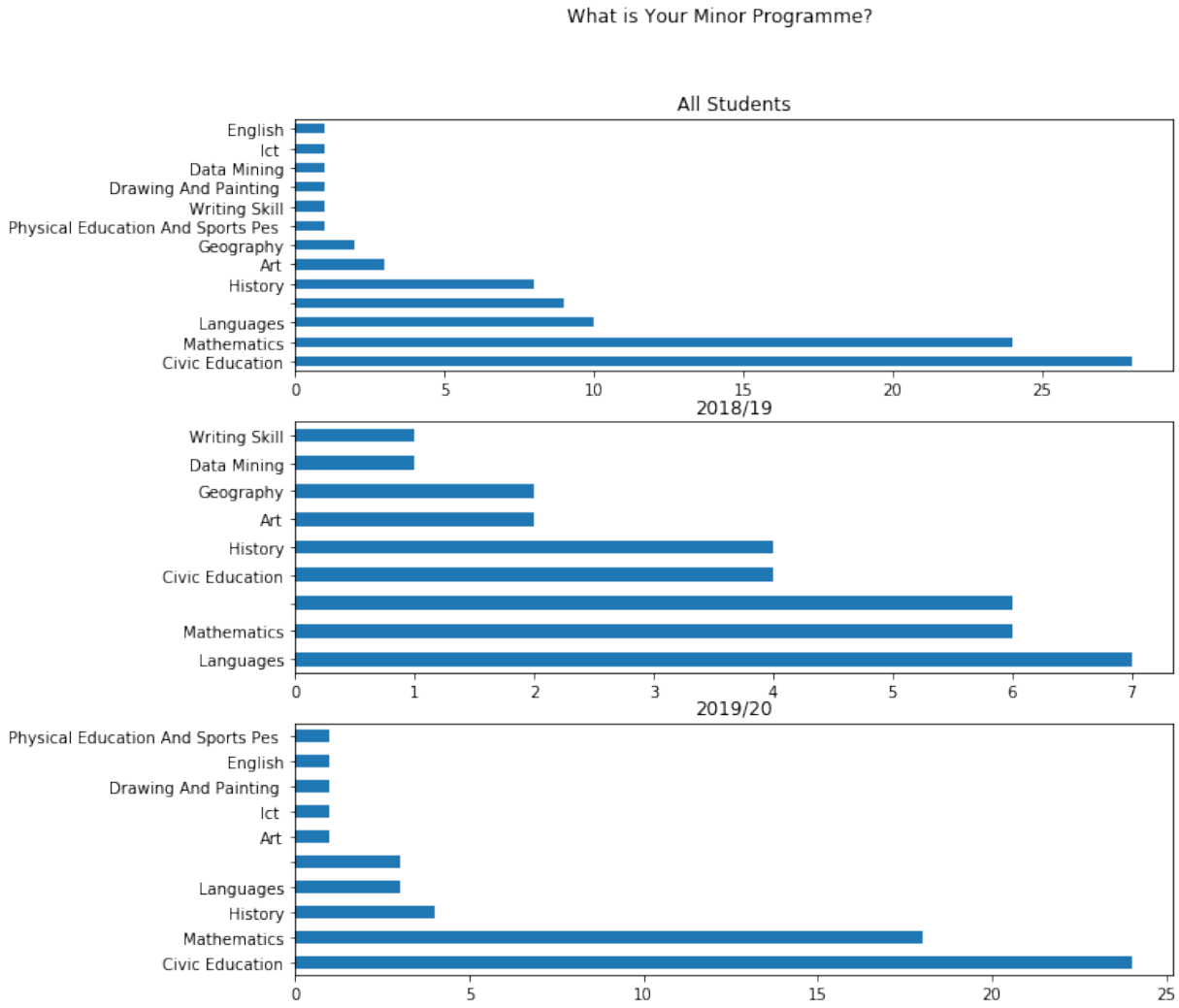
[36]:
```
# Facet results by academic year

fig, (ax1, ax2, ax3) = plt.subplots(3, 1, figsize=(10,10))

fig.suptitle('What is Your Minor Programme?')

#####var_ict1110_survey_eda.groupby(["ExperienceWithComputers"]).size().
 ↪plot(kind="barh", title="All Students", ax=ax1)
var_ict1110_survey_eda[var_ict1110_survey_eda["Timestamp"].str[:
 ↪1]=="2"]["MinorProgramme"].value_counts().plot(kind="barh", title="All Students",␣
 ↪ax=ax1)
var_ict1110_survey_eda[var_ict1110_survey_eda["Timestamp"].str[:
 ↪4]=="2019"]["MinorProgramme"].value_counts().plot(kind="barh", title="2018/19",␣
 ↪ax=ax2)
var_ict1110_survey_eda[var_ict1110_survey_eda["Timestamp"].str[:
 ↪4]=="2020"]["MinorProgramme"].value_counts().plot(kind="barh", title="2019/20",␣
 ↪ax=ax3)
```

[36]:
```
Text(0.5, 0.98, 'What is Your Minor Programme?')
```

[36]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f157139aac8>
```

[36]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1571361d30>
```

[36]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f15713122b0>
```



Computer Studies Elective in High School

[37]:
```
# 2. Computer Studies Elective in High School
#
# Basic dataframe summaries
#

# Get unique entries
var_ict1110_survey_eda["DidComputerStudies"].unique()

# Count observations
var_ict1110_survey_eda["DidComputerStudies"].count()

# Get value counts
var_ict1110_survey_eda["DidComputerStudies"].value_counts()
```

[37]:
```
array(['No', 'Yes'], dtype=object)
```

[37]:
```
90
```

[37]:

	DidComputerStudies
No	78
Yes	12

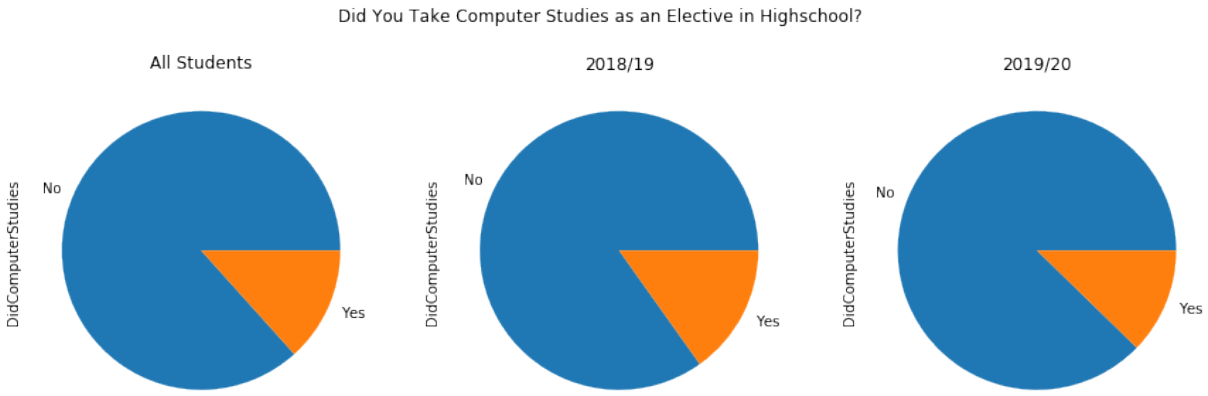
[38]:
```
# Facet results by academic year

fig, (ax1, ax2, ax3) = plt.subplots(1, 3, figsize=(15,5))

fig.suptitle('Did You Take Computer Studies as an Elective in Highschool?')

#####var_ict1110_survey_eda.groupby(["DidComputerStudies"]).size().plot(kind="pie",␣
↪title="All Students", ax=ax1)
var_ict1110_survey_eda[var_ict1110_survey_eda["Timestamp"].str[:
↪1]=="2"]["DidComputerStudies"].value_counts().plot(kind="pie", title="All Students",␣
↪ax=ax1)
var_ict1110_survey_eda[var_ict1110_survey_eda["Timestamp"].str[:
↪4]=="2019"]["DidComputerStudies"].value_counts().plot(kind="pie", title="2018/19",␣
↪ax=ax2)
var_ict1110_survey_eda[var_ict1110_survey_eda["Timestamp"].str[:
↪4]=="2020"]["DidComputerStudies"].value_counts().plot(kind="pie", title="2019/20",␣
↪ax=ax3)
```

[38]: `Text(0.5, 0.98, 'Did You Take Computer Studies as an Elective in Highschool?')`

[38]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f15707424e0>`

[38]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f1570753b00>`

[38]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f15711b5f98>`



Experience With Computers

[39]:
```
# 2. Computer Studies Elective in High School
#
# Basic dataframe summaries
#

# Get unique entries
var_ict1110_survey_eda["ExperienceWithComputers"].unique()

# Count observations
var_ict1110_survey_eda["ExperienceWithComputers"].count()

# Get value counts
var_ict1110_survey_eda["ExperienceWithComputers"].value_counts()
```

[39]:
```
array(['More than 5 years', '1 to 2 years', 'No Experience',
       'Less than 1 year', '3 to 5 years'], dtype=object)
```

[39]:
```
90
```

[39]:

	ExperienceWithComputers
No Experience	31
Less than 1 year	24
More than 5 years	21
1 to 2 years	11
3 to 5 years	3

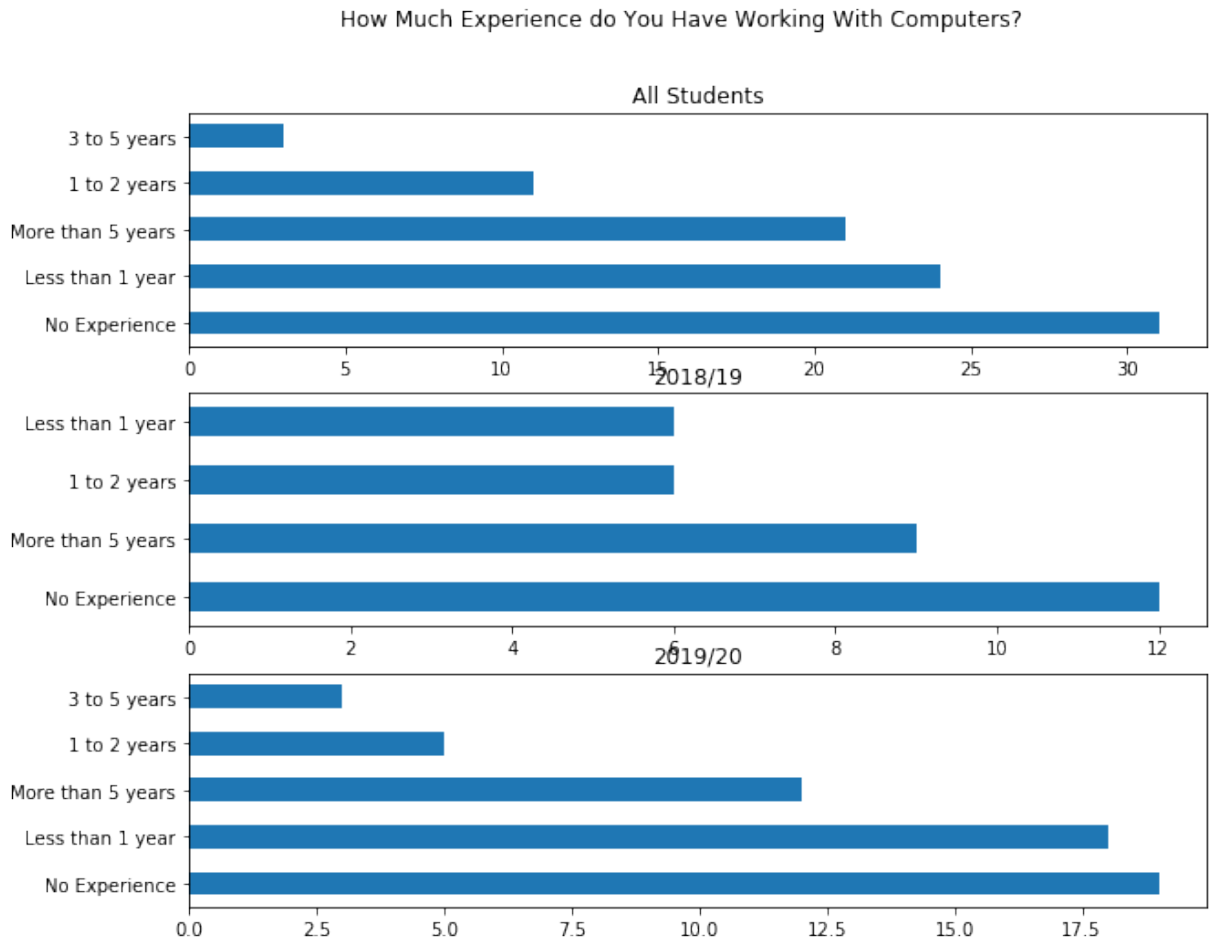
[40]:
```
# Facet results by academic year

fig, (ax1, ax2, ax3) = plt.subplots(3, 1, figsize=(10,8))

fig.suptitle('How Much Experience do You Have Working With Computers?')

#####var_ict1110_survey_eda.groupby(["ExperienceWithComputers"]).size().
 ↪plot(kind="barh", title="All Students", ax=ax1)
var_ict1110_survey_eda[var_ict1110_survey_eda["Timestamp"].str[:
 ↪1]=="2"]["ExperienceWithComputers"].value_counts().plot(kind="barh", title="All␣
 ↪Students", ax=ax1)
var_ict1110_survey_eda[var_ict1110_survey_eda["Timestamp"].str[:
 ↪4]=="2019"]["ExperienceWithComputers"].value_counts().plot(kind="barh", title="2018/
 ↪19", ax=ax2)
var_ict1110_survey_eda[var_ict1110_survey_eda["Timestamp"].str[:
 ↪4]=="2020"]["ExperienceWithComputers"].value_counts().plot(kind="barh", title="2019/
 ↪20", ax=ax3)
```

[40]:
```
Text(0.5, 0.98, 'How Much Experience do You Have Working With Computers?')
```

[40]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1571195ac8>
```

[40]: <matplotlib.axes._subplots.AxesSubplot at 0x7f157113d080>

[40]: <matplotlib.axes._subplots.AxesSubplot at 0x7f15711644e0>



Programme Major Motivation

[41]:
```
var_ict1110_major_motivation = WordCloud(stopwords=stopwords.words("english"),␣
↪background_color="white", colormap="Dark2", max_font_size=150, random_state=42)


var_ict1110_survey_eda_motivation = var_ict1110_survey_eda["MajorProgrammeMotivation"]
var_ict1110_major_motivation.generate(' '.join(var_ict1110_survey_eda_motivation))

plt.figure(figsize = (15,10))
plt.imshow(var_ict1110_major_motivation)
plt.axis("off")
```

[41]: <wordcloud.wordcloud.WordCloud at 0x7f15710d1b38>

[41]: <Figure size 1080x720 with 0 Axes>

[41]: <matplotlib.image.AxesImage at 0x7f1570523c18>

[41]: (-0.5, 399.5, 199.5, -0.5)

[42]:
```
#
# DataLab keywords
#
var_datalab = pd.read_csv("/home/lightonphiri/Dropbox/Projects/work/2021/
 ↪misc-unza21-datalab/db-unza21-datalab_output.csv", sep="|")

# Stopword removal; punctuation removal and case folding
var_datalab["Abstract"] = var_datalab["Abstract"].apply(fxn_punctuation).
 ↪apply(fxn_case_folding).apply(fxn_stopwords).str.title()

#
var_datalab_wordcloud = WordCloud(stopwords=stopwords.words("english"),␣
 ↪background_color="white", colormap="Dark2", max_font_size=150, random_state=42)


var_datalab_research = var_datalab["Abstract"]
var_datalab_wordcloud.generate(' '.join(var_datalab_research))

plt.figure(figsize = (15,5))
plt.imshow(var_datalab_wordcloud)
plt.axis("off")
```

[42]: <wordcloud.wordcloud.WordCloud at 0x7f15704ed940>

[42]: <Figure size 1080x360 with 0 Axes>

[42]: <matplotlib.image.AxesImage at 0x7f15705072b0>

[42]: (-0.5, 399.5, 199.5, -0.5)

Dataset #2: 2018/19 ICT 1110 Student Demographics

Using the UNZA 2018/19 ICT 1110 Student Demographics dataset, work towards the following: 1. Preprocess the datasets 3. Perform Exploratory Data Analysis on the merged dataset

Data Preprocessing

Dataset Description

Dataset Files

The files are categorised as follows: * db-unza20-csc5741-ict1110_student_demographics.csv * Student demographic details extracted from the Student Information System

Dataset Format

- Student Demographic Details
  - A total of 13 fields representing student demographics
  - The fields are pipe ("|") seperated

[43]:
```
# Datasets format
!cat -n db-unza21-csc5741-ict1110_student_demographics.csv | head
```

```
     1  Student ID|DateOfBirth|Gender|Academic Year|Year Of Study|School|Program|MajorDes
cription|MinorDescription|Status|Sponsor|Nationality|Accommodated
     2  9d5116a2451bc98c2b46b93acbc1b4f0|1998-09-14|F|20191|2nd Year|EDUCATION|BACHELOR
OF INFORMATION AND COMMUNICATION TECHNOLOGIES WITH EDUCATION (B.ICTs.Ed)|ICTs and
Education|RELIGIOUS STUDIES|Registered|GRZ - 75 PERCENT|ZAMBIAN|Yes
     3  e7400496f1ce70cb62c2c44ca2ddc469|2000-03-23|M|20191|2nd Year|EDUCATION|BACHELOR
```

```
OF INFORMATION AND COMMUNICATION TECHNOLOGIES WITH EDUCATION (B.ICTs.Ed)|ICTs and
Education|ART AND DESIGN STUDIES|Registered|GRZ-FULLY SPONSORED|ZAMBIAN|No
     4  cea34f6b4f356c28fc2b766ae46b6d6c|1996-06-06|M|20171|1st Year|EDUCATION|BACHELOR
OF INFORMATION AND COMMUNICATION TECHNOLOGIES WITH EDUCATION (B.ICTs.Ed)|ICTs and
Education|GEOGRAPHY|Registered|GRZ-FULLY SPONSORED|ZAMBIAN|Yes
     5  #N/A|#N/A|#N/A|#N/A|#N/A|#N/A|#N/A|#N/A|#N/A|#N/A|#N/A|#N/A|No
     6  6cd50fb3091b0a9d3c1ac2cf52441390|1999-03-03|M|20181|1st Year|EDUCATION|BACHELOR
OF INFORMATION AND COMMUNICATION TECHNOLOGIES WITH EDUCATION (B.ICTs.Ed)|ICTs and
Education|MATHEMATICS|Registered|GRZ-FULLY SPONSORED|ZAMBIAN|No
     7  e31959fe2842dacea4d16d36e9813620|1995-07-27|F|20171|1st Year|EDUCATION|BACHELOR
OF INFORMATION AND COMMUNICATION TECHNOLOGIES WITH EDUCATION (B.ICTs.Ed)|ICTs and
Education|HISTORY|Registered|GRZ-FULLY SPONSORED|ZAMBIAN|No
     8  97527dec0ae1a703599581d4f25dfbce|1995-06-08|F|20181|2nd Year|EDUCATION|BACHELOR
OF INFORMATION AND COMMUNICATION TECHNOLOGIES WITH EDUCATION (B.ICTs.Ed)|ICTs and
Education|CIVIC EDUCATION|Registered|GRZ-FULLY SPONSORED|ZAMBIAN|No
     9  9e7002d53d4db7bfad4f5cf419b0c126|1994-06-26|M|20181|1st Year|EDUCATION|BACHELOR
OF INFORMATION AND COMMUNICATION TECHNOLOGIES WITH EDUCATION (B.ICTs.Ed)|ICTs and
Education||Registered|GRZ-FULLY SPONSORED|ZAMBIAN|No
    10  74458a3d3e5f3074226b1f9fa23c9163|1999-04-04|M|20181|1st Year|EDUCATION|BACHELOR
OF INFORMATION AND COMMUNICATION TECHNOLOGIES WITH EDUCATION (B.ICTs.Ed)|ICTs and
Education|CIVIC EDUCATION|Registered|GRZ-FULLY SPONSORED|ZAMBIAN|Yes
```

[44]:
```
! cat db-unza21-csc5741-ict1110_student_demographics.csv | wc -l
```

```
61
```

Dataframe Creation

[45]:
```
# Create DataFrame of input dataset: ICT 1110 Demographics
#
var_ict1110_demographics = pd.read_csv("db-unza21-csc5741-ict1110_student_demographics.
 ↪csv", sep="|")
var_ict1110_demographics.columns
```

[45]:
```
Index(['Student ID', 'DateOfBirth', 'Gender', 'Academic Year', 'Year Of Study',
       'School', 'Program', 'MajorDescription', 'MinorDescription', 'Status',
       'Sponsor', 'Nationality', 'Accommodated'],
      dtype='object')
```

[46]:
```
var_ict1110_demographics.columns
var_ict1110_demographics.rename(columns={
    "Student ID": "StudentID",
    "Academic Year": "AcademicYear",
    "Year Of Study": "YearOfStudy"
}).head(2)

var_ict1110_demographics.columns

var_ict1110_demographics = var_ict1110_demographics.rename(columns={
    "Student ID": "StudentID",
    "Academic Year": "AcademicYear",
    "Year Of Study": "YearOfStudy"
```

```
})

var_ict1110_demographics.columns
```

[46]:
```
Index(['Student ID', 'DateOfBirth', 'Gender', 'Academic Year', 'Year Of Study',
       'School', 'Program', 'MajorDescription', 'MinorDescription', 'Status',
       'Sponsor', 'Nationality', 'Accommodated'],
      dtype='object')
```

[46]:

	StudentID	DateOfBirth	Gender	AcademicYear	YearOfStudy	School	Prog
0	9d5116a2451bc98c2b46b93acb...	1998-09-14	F	20191.0	2nd Year	EDUCATION	BAC
1	e7400496f1ce70cb62c2c44ca2...	2000-03-23	M	20191.0	2nd Year	EDUCATION	BAC

[46]:
```
Index(['Student ID', 'DateOfBirth', 'Gender', 'Academic Year', 'Year Of Study',
       'School', 'Program', 'MajorDescription', 'MinorDescription', 'Status',
       'Sponsor', 'Nationality', 'Accommodated'],
      dtype='object')
```

[46]:
```
Index(['StudentID', 'DateOfBirth', 'Gender', 'AcademicYear', 'YearOfStudy',
       'School', 'Program', 'MajorDescription', 'MinorDescription', 'Status',
       'Sponsor', 'Nationality', 'Accommodated'],
      dtype='object')
```

[47]:
```
# Rename dataframe columns for easy processing
#
# Function use: pd.rename([...])
# Important parameter: inplace=True
#
var_ict1110_demographics.rename(columns={
    "Student ID": "StudentID",
    "Academic Year": "AcademicYear",
    "Year Of Study": "YearOfStudy"
}, inplace=True)

var_ict1110_demographics.columns
```

[47]:
```
Index(['StudentID', 'DateOfBirth', 'Gender', 'AcademicYear', 'YearOfStudy',
       'School', 'Program', 'MajorDescription', 'MinorDescription', 'Status',
       'Sponsor', 'Nationality', 'Accommodated'],
      dtype='object')
```

[48]:
```
# Count records in dataframe
#
len(var_ict1110_demographics)
```

[48]:
```
60
```

[49]:
```
# Inspect some dataframe records
#
var_ict1110_demographics.head(2).T
```

[49]:

	0	1
StudentID	9d5116a2451bc98c2b46b93acb...	e7400496f1ce70cb62c2c44ca2...
DateOfBirth	1998-09-14	2000-03-23
Gender	F	M
AcademicYear	20191	20191
YearOfStudy	2nd Year	2nd Year
School	EDUCATION	EDUCATION
Program	BACHELOR OF INFORMATION AN...	BACHELOR OF INFORMATION AN...
MajorDescription	ICTs and Education	ICTs and Education
MinorDescription	RELIGIOUS STUDIES	ART AND DESIGN STUDIES
Status	Registered	Registered
Sponsor	GRZ - 75 PERCENT	GRZ-FULLY SPONSORED
Nationality	ZAMBIAN	ZAMBIAN
Accommodated	Yes	No

Dataset Attributes

- StudentID—Alphanumeric
- DateOfBirth—Date
- Gender—Text/Char
- AcademicYear—Text
- YearOfStudy—Text
- School—Text
- Programme—Text
- MajorDescription—Text
- MinorDescription—Text
- Status—Categorical (Dichotomus)
- Sponsor—Text
- Nationality—Text
- Accommodated—Categorical (Dichotomus)

Data Pre-processing Plan

- STEP 1: Remove all records with NULL StudentID values
- STEP 2: Remove duplicates using StudentID as unique field

NOTE: Null values to be handled on a case-by-case basis; e.g. null values in text attributes to be replaced with empty strings ""

[50]:
```
# STEP 1: Remove all records with NULL StudentID values
#
# Print records with NULL StudentID values
#
var_ict1110_demographics["StudentID"].isnull().head(2)
```

[50]:

	StudentID
0	False
1	False

[51]:
```
len(var_ict1110_demographics[var_ict1110_demographics["StudentID"].isnull()])
var_ict1110_demographics[var_ict1110_demographics["StudentID"].isnull()]
```

[51]: 2

[51]:

	StudentID	DateOfBirth	Gender	AcademicYear	YearOfStudy	School	Program	MajorDescription	
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
56	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

[52]:
```
# Remove records with NULL StudentID values
#
# var_ict1110_demographics.dropna(subset = ["StudentID"])
# var_ict1110_demographics =␣
 ↪var_ict1110_demographics[var_ict1110_demographics["StudentID"].notna()]
#
var_ict1110_demographics.dropna(subset = ["StudentID"], inplace=True)
```

[53]:
```
len(var_ict1110_demographics)
var_ict1110_demographics_ = var_ict1110_demographics.dropna(subset = ["StudentID"])
len(var_ict1110_demographics_)
```

[53]: 58

[53]: 58

[54]:
```
len(var_ict1110_demographics)
```

[54]: 58

[55]:
```
var_ict1110_demographics.duplicated?
```

[56]:
```
# STEP 2: Remove duplicates using StudentID as unique field
#
# Print duplicate records on StudentID
#
# Using Pandas, the df.duplicated() function can be used to identify duplicates
# http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.duplicated.
 ↪html
# var_ict1110_survey.duplicated(["StudentID"], keep=False)
#
var_ict1110_demographics[var_ict1110_demographics.duplicated(["StudentID"],␣
 ↪keep="first")].sort_values("StudentID").tail(2).T
```

[56]:

Empty DataFrame Columns: Int64Index([], dtype='int64') Index: Index(['StudentID', 'DateOfBirth', 'Gender', 'Acad

[57]:
```
# Remove duplicate records on StudentID: a record without a Student ID is useless
#
# drop_duplicates uses keep=[First, Last, False]
```

```
#
var_ict1110_survey.drop_duplicates(["StudentID"], keep="first", inplace=True)
```

[58]:
```
# Confirm duplicate records removal
len(var_ict1110_demographics)
```

[58]: 58

[59]:
```
# STEP 2: Apply steps 3--6 to Text Attributes recursively
# Attributes: MinorProgrammeMotivation, MajorProgrammeMotivation and AboutMe
#
# Inspect dataframe before pre-processing
var_ict1110_demographics.sample(2).T
```

[59]:

	21	15
StudentID	43f9f6077d6be60269ad8cfb3f...	4be25f9d27da71d4e98775668b...
DateOfBirth	1999-06-28	1983-08-25
Gender	F	F
AcademicYear	20191	20191
YearOfStudy	2nd Year	2nd Year
School	EDUCATION	EDUCATION
Program	BACHELOR OF INFORMATION AN...	BACHELOR OF INFORMATION AN...
MajorDescription	ICTs and Education	ICTs and Education
MinorDescription	HISTORY	RELIGIOUS STUDIES
Status	Registered	Registered
Sponsor	GRZ-FULLY SPONSORED	SELF-SPONSORED
Nationality	ZAMBIAN	ZAMBIAN
Accommodated	No	No

[60]:
```
# Handle null values---NaN
#
# DateOfBirth: Replace NaN with "MISSING DATA"
# Gender: Replace NaN with U, for Unknown
# AcademicYear: drop value
# YearOfStudy: drop value
# School: Replace with EDUCATION
# Programme: Replace with BACHELOR OF INFORMATION AND COMMUNICATION TECHNOLOGIES WITH␣
 ↪EDUCATION (B.ICTs.Ed)
# MajorDescription: Replace with ICTs and Education
# MinorDescription: Replace with "MISSING DATA"
# Status: Replace with "MISSING DATA"
# Sponsor: Replace with SELF-SPONSORED
# Nationality: Replace with ZAMBIAN
# Accommodated: Replace with No
#
```

[61]:
```
# ALWAYS check data types of columns to ensure you are replacing them with appropriate␣
 ↪values
#
```

```
var_ict1110_demographics.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 58 entries, 0 to 59
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   StudentID         58 non-null     object
 1   DateOfBirth       58 non-null     object
 2   Gender            58 non-null     object
 3   AcademicYear      58 non-null     float64
 4   YearOfStudy       58 non-null     object
 5   School            58 non-null     object
 6   Program           58 non-null     object
 7   MajorDescription  58 non-null     object
 8   MinorDescription  46 non-null     object
 9   Status            58 non-null     object
 10  Sponsor           58 non-null     object
 11  Nationality       58 non-null     object
 12  Accommodated      58 non-null     object
dtypes: float64(1), object(12)
memory usage: 6.3+ KB
```

[62]:
```
# DateOfBirth
var_ict1110_demographics["DateOfBirth"].fillna("MISSING DATA", inplace=True)

# Gender
var_ict1110_demographics["Gender"].fillna("U", inplace=True)

# AcademicYear
var_ict1110_demographics["AcademicYear"].fillna("MISSING DATA", inplace=True)

# YearOfStudy
var_ict1110_demographics["YearOfStudy"].fillna("MISSING DATA", inplace=True)

# School
var_ict1110_demographics["School"].fillna("EDUCATION", inplace=True)

# Program
var_ict1110_demographics["Program"].fillna("BACHELOR OF INFORMATION AND COMMUNICATION␣
 ↪TECHNOLOGIES WITH EDUCATION (B.ICTs.Ed)", inplace=True)

# MajorDescription
var_ict1110_demographics["MajorDescription"].fillna("MISSING DATA", inplace=True)

# MinorDescription
var_ict1110_demographics["MinorDescription"].fillna("MISSING DATA", inplace=True)

# Status
var_ict1110_demographics["Status"].fillna("", inplace=True)
```

```
# Sponsor
var_ict1110_demographics["Sponsor"].fillna("SELF-SPONSORED", inplace=True)

# Nationality
var_ict1110_demographics["Nationality"].fillna("MISSING DATA", inplace=True)

# Accommodated
var_ict1110_demographics["Accommodated"].fillna("MISSING DATA", inplace=True)
```

[63]:
```
# Inspect dataframe to check for new values
var_ict1110_demographics.sample(2).T
len(var_ict1110_demographics)
```

[63]:

	4	45
StudentID	6cd50fb3091b0a9d3c1ac2cf52...	b5bcce260d9e303ca0e63f0551...
DateOfBirth	1999-03-03	1999-05-02
Gender	M	F
AcademicYear	20181	20191
YearOfStudy	1st Year	2nd Year
School	EDUCATION	EDUCATION
Program	BACHELOR OF INFORMATION AN...	BACHELOR OF INFORMATION AN...
MajorDescription	ICTs and Education	ICTs and Education
MinorDescription	MATHEMATICS	HISTORY
Status	Registered	Registered
Sponsor	GRZ-FULLY SPONSORED	GRZ-FULLY SPONSORED
Nationality	ZAMBIAN	ZAMBIAN
Accommodated	No	Yes

[63]:
```
58
```

Exploratory Data Analysis

[64]:
```
# Describe the data
# Identify attributes to explore
var_ict1110_demographics.head(2).T
```

[64]:

	0	1
StudentID	9d5116a2451bc98c2b46b93acb...	e7400496f1ce70cb62c2c44ca2...
DateOfBirth	1998-09-14	2000-03-23
Gender	F	M
AcademicYear	20191	20191
YearOfStudy	2nd Year	2nd Year
School	EDUCATION	EDUCATION
Program	BACHELOR OF INFORMATION AN...	BACHELOR OF INFORMATION AN...
MajorDescription	ICTs and Education	ICTs and Education
MinorDescription	RELIGIOUS STUDIES	ART AND DESIGN STUDIES

Continued on next page

	0	1
Status	Registered	Registered
Sponsor	GRZ - 75 PERCENT	GRZ-FULLY SPONSORED
Nationality	ZAMBIAN	ZAMBIAN
Accommodated	Yes	No

[65]:
```
var_ict1110_demographics.describe(include="all")
```

[65]:

	StudentID	DateOfBirth	Gender	AcademicYear	YearOfStudy	School
count	58	58	58	58.000000	58	58
unique	58	58	2	NaN	2	1
top	e2a96e074e1d8a6f6de56abbd4...	1998-09-14	M	NaN	2nd Year	EDUCATION
freq	1	1	30	NaN	30	58
mean	NaN	NaN	NaN	20185.310345	NaN	NaN
std	NaN	NaN	NaN	5.956588	NaN	NaN
min	NaN	NaN	NaN	20171.000000	NaN	NaN
25%	NaN	NaN	NaN	20181.000000	NaN	NaN
50%	NaN	NaN	NaN	20181.000000	NaN	NaN
75%	NaN	NaN	NaN	20191.000000	NaN	NaN
max	NaN	NaN	NaN	20191.000000	NaN	NaN

Possible attributes to include in the EDA process

- DateOfBirth
- Gender
- MinorDescription
- Sponsor
- Accommodated

[66]:
```
# Define variable to
var_ict1110_demographics_eda = var_ict1110_demographics
```

Dataframe Statistical Information

[67]:
```
# Use describe function to get statistical information of dataset attributes
var_ict1110_demographics_eda.describe(include='all').T
```

[67]:

	count	unique	top	freq	mean	std	min
StudentID	58	58	e2a96e074e1d8a6f6de56abbd4...	1	NaN	NaN	NaN
DateOfBirth	58	58	1998-09-14	1	NaN	NaN	NaN
Gender	58	2	M	30	NaN	NaN	NaN
AcademicYear	58	NaN	NaN	NaN	20185.3	5.95659	20171
YearOfStudy	58	2	2nd Year	30	NaN	NaN	NaN
School	58	1	EDUCATION	58	NaN	NaN	NaN
Program	58	1	BACHELOR OF INFORMATION AN...	58	NaN	NaN	NaN
MajorDescription	58	1	ICTs and Education	58	NaN	NaN	NaN

	count	unique	top	freq	mean	std	min
MinorDescription	58	9	MISSING DATA	12	NaN	NaN	NaN
Status	58	2	Registered	56	NaN	NaN	NaN
Sponsor	58	4	GRZ-FULLY SPONSORED	38	NaN	NaN	NaN
Nationality	58	1	ZAMBIAN	58	NaN	NaN	NaN
Accommodated	58	2	No	33	NaN	NaN	NaN

Date of Birth

[68]:
```
# Basic dataframe summaries
#
# Get unique entries
var_ict1110_demographics_eda["DateOfBirth"].unique()

# Count observations
var_ict1110_demographics_eda["DateOfBirth"].count()

# Get value counts
var_ict1110_demographics_eda["DateOfBirth"].str[:4].value_counts()
```

[68]:
```
array(['1998-09-14', '2000-03-23', '1996-06-06', '1999-03-03',
       '1995-07-27', '1995-06-08', '1994-06-26', '1999-04-04',
       '1999-01-20', '2000-12-06', '1999-11-22', '1995-12-25',
       '1989-11-07', '1997-10-15', '1983-08-25', '1998-03-03',
       '1997-07-23', '1997-08-28', '1998-03-15', '1999-12-24',
       '1999-06-28', '1994-10-08', '1998-06-25', '2000-12-11',
       '2002-01-01', '1998-03-07', '1999-02-17', '1999-03-23',
       '1999-06-19', '2001-10-01', '2000-11-10', '1997-11-21',
       '1998-07-12', '1997-10-12', '1998-09-04', '1994-07-26',
       '2000-12-23', '1992-11-16', '1998-06-22', '1999-02-15',
       '1999-08-21', '1997-10-11', '1998-12-21', '1998-12-04',
       '1999-05-02', '1998-12-02', '1996-01-08', '2001-12-20',
       '2000-02-09', '1999-06-05', '1999-07-08', '1997-08-17',
       '1999-07-23', '1998-11-20', '2001-12-01', '2001-03-02',
       '2000-07-25', '1998-05-06'], dtype=object)
```

[68]:
```
58
```

[68]:

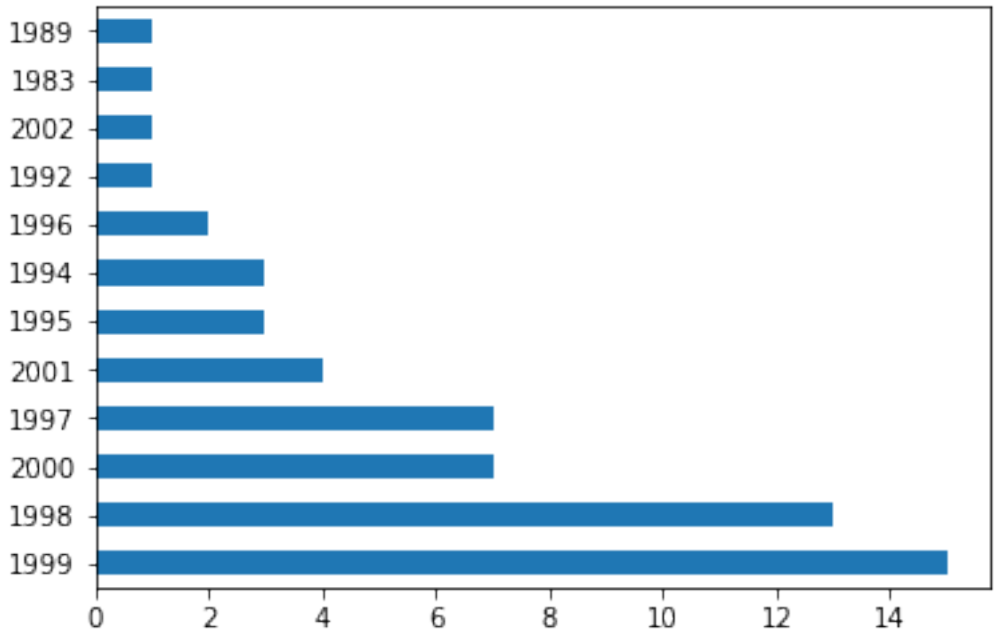
	DateOfBirth
1999	15
1998	13
2000	7
1997	7
2001	4
1995	3
1994	3
1996	2
1992	1

Continued on next page

	DateOfBirth
2002	1
1983	1
1989	1

[69]:
```
# Plot Date of Birth barplot
#
# Check the date type
type(var_ict1110_demographics_eda["DateOfBirth"].str[:4].value_counts())

#
var_ict1110_demographics_eda_dateofbirth = var_ict1110_demographics_eda["DateOfBirth"].
 ↪str[:4].value_counts()

#

#####sns.barplot(x=var_ict1110_demographics_eda_dateofbirth.index,␣
 ↪y=var_ict1110_demographics_eda_dateofbirth.values, palette='Spectral')
var_ict1110_demographics_eda_dateofbirth = var_ict1110_demographics_eda["DateOfBirth"].
 ↪str[:4].value_counts().plot(kind="barh")
```

[69]: pandas.core.series.Series



Gender

[70]:
```
# Basic dataframe summaries
#
# Get unique entries
var_ict1110_demographics_eda["Gender"].unique()

# Count observations
var_ict1110_demographics_eda["Gender"].count()

# Get value counts
var_ict1110_demographics_eda["Gender"].value_counts()
```

[70]:
```
array(['F', 'M'], dtype=object)
```

[70]:
```
58
```

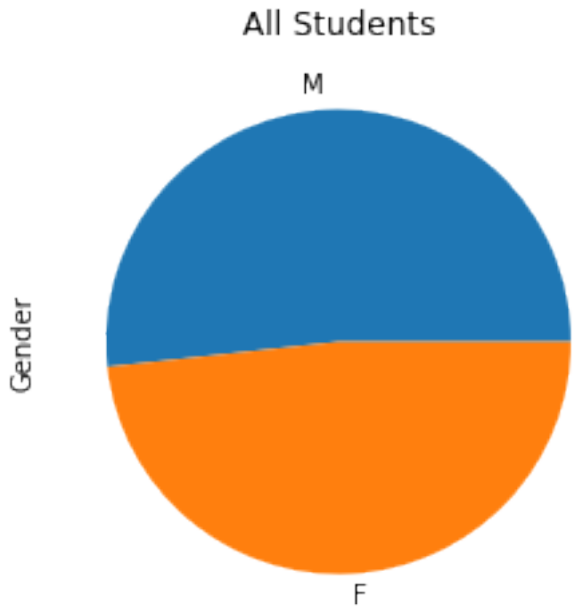
[70]:

	Gender
M	30
F	28

[71]:
```
# Plot Gender pie chart
#
# Check the date type
type(var_ict1110_demographics_eda["Gender"].value_counts())

#
var_ict1110_demographics_eda_gender = var_ict1110_demographics_eda["Gender"].
 ↪value_counts()

#

###sns.barplot(x=var_ict1110_demographics_eda_dateofbirth.index,␣
 ↪y=var_ict1110_demographics_eda_dateofbirth.values, palette='Spectral')

var_ict1110_demographics_eda["Gender"].value_counts().plot(kind="pie", title="All␣
 ↪Students")
```

[71]:
```
pandas.core.series.Series
```

[71]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1549044eb8>
```



Minor Description

[72]:
```
# Basic dataframe summaries
#
# Get unique entries
var_ict1110_demographics_eda["MinorDescription"].unique()

# Count observations
var_ict1110_demographics_eda["MinorDescription"].count()

# Get value counts
var_ict1110_demographics_eda["MinorDescription"].value_counts()
```

[72]:
```
array(['RELIGIOUS STUDIES', 'ART AND DESIGN STUDIES', 'GEOGRAPHY',
       'MATHEMATICS', 'HISTORY', 'CIVIC EDUCATION', 'MISSING DATA',
       'FRENCH', 'ENGLISH'], dtype=object)
```

[72]:
```
58
```

[72]:

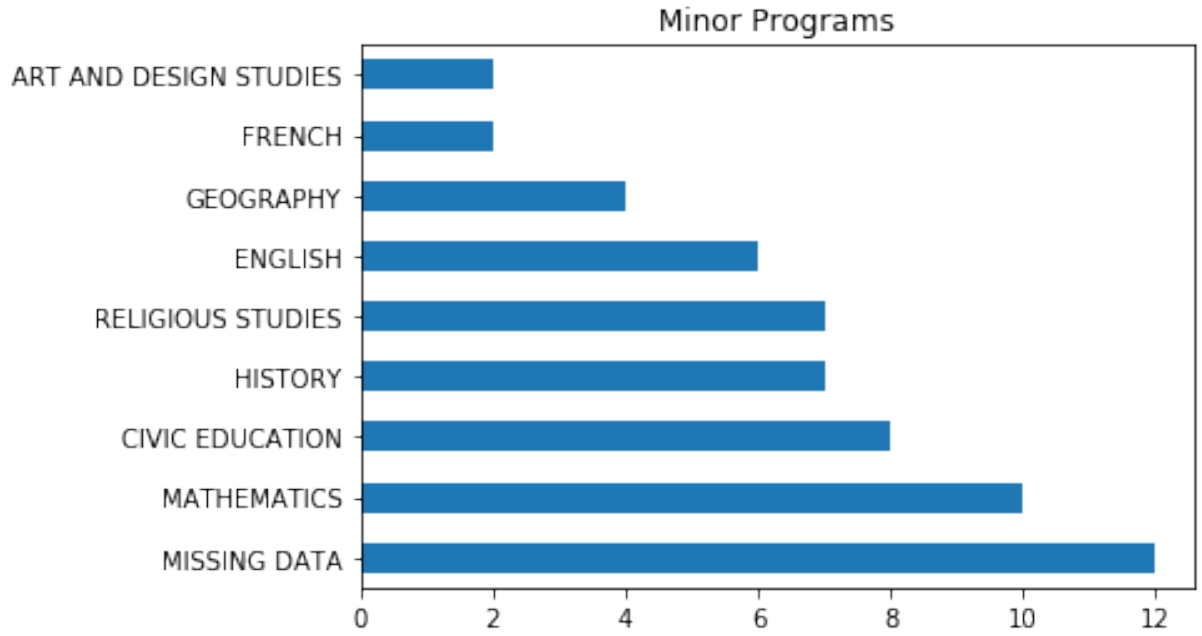
	MinorDescription
MISSING DATA	12
MATHEMATICS	10
CIVIC EDUCATION	8
HISTORY	7
RELIGIOUS STUDIES	7
ENGLISH	6
GEOGRAPHY	4

Continued on next page

	MinorDescription
FRENCH	2
ART AND DESIGN STUDIES	2

[73]:
```
# Plot Minors using barplot
#
#
var_ict1110_demographics_eda["MinorDescription"].value_counts().plot(kind="barh",␣
↪title="Minor Programs")
```

[73]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1549025438>
```



Sponsor

[74]:
```
# Basic dataframe summaries
#
# Get unique entries
var_ict1110_demographics_eda["Sponsor"].unique()

# Count observations
var_ict1110_demographics_eda["Sponsor"].count()

# Get value counts
var_ict1110_demographics_eda["Sponsor"].value_counts()
```

[74]:
```
array(['GRZ - 75 PERCENT', 'GRZ-FULLY SPONSORED',
       'TUITION WAIVER(DEPENDANTS)', 'SELF-SPONSORED'], dtype=object)
```

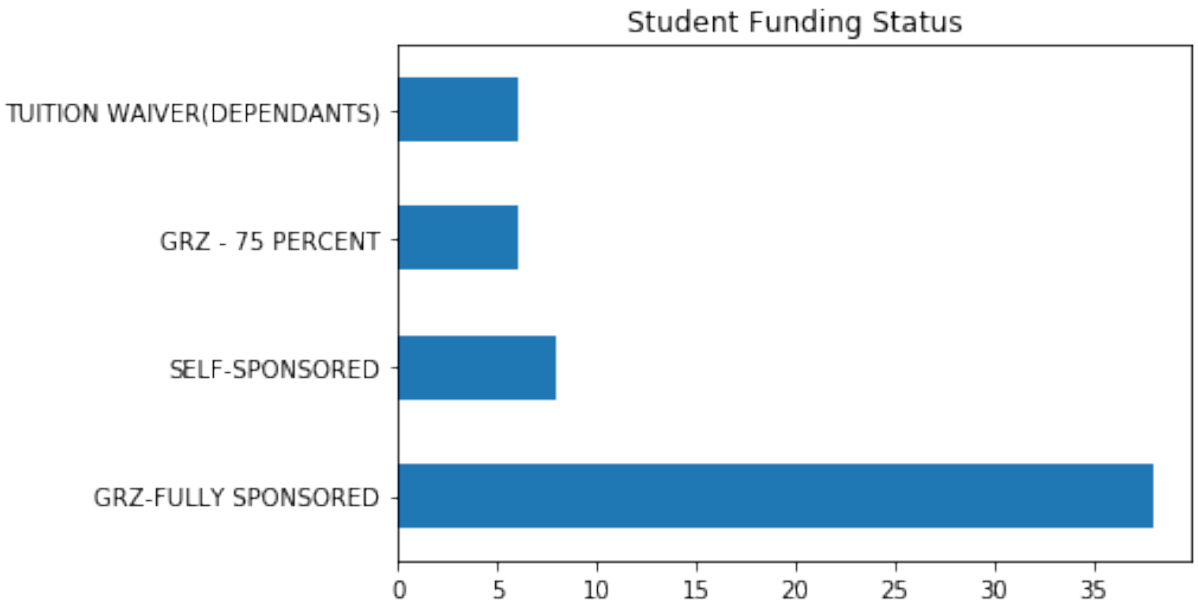
[74]: 58

[74]:

	Sponsor
GRZ-FULLY SPONSORED	38
SELF-SPONSORED	8
GRZ - 75 PERCENT	6
TUITION WAIVER(DEPENDANTS)	6

[75]:
```
# Plot Sponsors using barplot
#
#
var_ict1110_demographics_eda["Sponsor"].value_counts().plot(kind="barh",␣
 ↪title="Student Funding Status")
```

[75]: <matplotlib.axes._subplots.AxesSubplot at 0x7f1548ff4eb8>



Accommodated

[76]:
```
# Basic dataframe summaries
#
# Get unique entries
var_ict1110_demographics_eda["Accommodated"].unique()

# Count observations
var_ict1110_demographics_eda["Accommodated"].count()

# Get value counts
var_ict1110_demographics_eda["Accommodated"].value_counts()
```

[76]:
```
array(['Yes', 'No'], dtype=object)
```

[76]:
```
58
```

[76]:

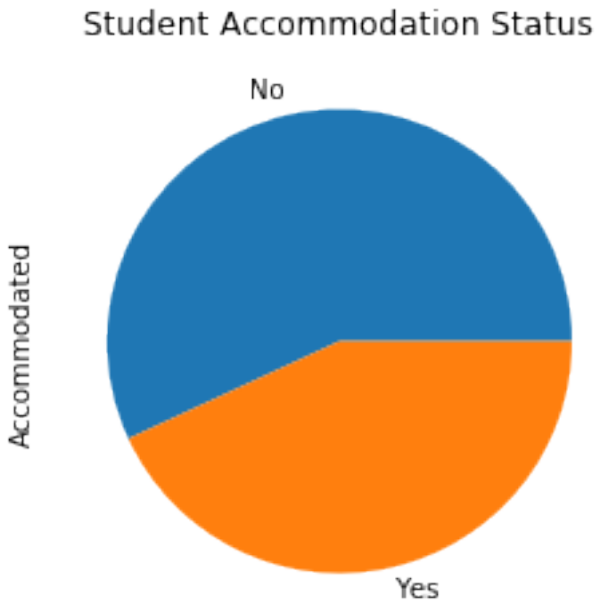
	Accommodated
No	33
Yes	25

[77]:
```
var_ict1110_demographics_eda.groupby("Accommodated")
```

[77]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x7f1548f6ee48>
```

[78]:
```
# Plot Accommodation status using pie chart
#
#
var_ict1110_demographics_eda["Accommodated"].value_counts().plot(kind="pie",␣
↪title="Student Accommodation Status")
```

[78]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1548f58048>
```



Dataset #3: 2018/19 ICT 1110 Assessment Scores

Using the UNZA 2018/19 ICT 1110 Assessment Scores dataset, work towards the following: 1. Pre-process the datasets 2. Merge the datasets 3. Perform Exploratory Data Analysis on the merged dataset

Data Preprocessing

Dataset Description

Dataset Files

The files are categorised as follows: * db-unza21-csc5741-ict1110_assessment_scores-quizN.csv * Weekly quiz scores gradded out of 10 * There are a total of 20 data files: quiz 1–20 * db-unza21-csc5741-ict1110_assessment_scores-testN.csv * Test scores graded out of 50 * There are a total of four data files: tests 1–4 * db-unza21-csc5741-ict1110_assessment_scores-makeup_test.csv * Make-up assessment for students that never scoore minimum continuous assessment score required to write the final examination * The scores are out of 50 * There is one data file * db-unza21-csc5741-ict1110_assessment_scores-final_examination.csv * Final Examination scores graded out of 100 * There is one data file

Dataset Format

- Assessment Scores Datasets
  - All data files have the same format; they are composed of two fields: StudentID (alphanumeric) and Mark (numeric)
  - The fileds are pipe ("|") seperated

[79]:
```
# Datasets format
!cat -n db-unza21-csc5741-ict1110_assessment_scores-quiz20.csv | head
```

```
     1  Student ID|Mark
     2  28765464efe1b6583610335965b4d75a|9
     3  74458a3d3e5f3074226b1f9fa23c9163|9
     4  9d5116a2451bc98c2b46b93acbc1b4f0|2
     5  #N/A|9
     6  94984a8c4896946d9bafd24959cb6181|10
     7  07f3ca235faaa1c9ad16facef5526d8b|10
     8  e7400496f1ce70cb62c2c44ca2ddc469|9
     9  e31959fe2842dacea4d16d36e9813620|9
    10  575b9408b6daa2ddcefbcf6d81c9b4c9|10
```

[80]:
```
!ls -l db-unza21-csc5741-ict1110_assessment_scores-quiz* | wc -l
!ls -l db-unza21-csc5741-ict1110_assessment_scores-quiz*
```

```
20
-rw-rw-r-- 1 lightonphiri lightonphiri 1949 May 30 16:08 db-
unza21-csc5741-ict1110_assessment_scores-quiz01.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 2028 May 30 16:08 db-
unza21-csc5741-ict1110_assessment_scores-quiz02.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 2009 May 30 16:08 db-
unza21-csc5741-ict1110_assessment_scores-quiz03.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1966 May 30 16:08 db-
unza21-csc5741-ict1110_assessment_scores-quiz04.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 2104 May 30 16:08 db-
unza21-csc5741-ict1110_assessment_scores-quiz05.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 2067 May 30 16:08 db-
unza21-csc5741-ict1110_assessment_scores-quiz06.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1920 May 30 16:08 db-
unza21-csc5741-ict1110_assessment_scores-quiz07.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1990 May 30 16:08 db-
```

```
unza21-csc5741-ict1110_assessment_scores-quiz08.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1986 May 30 16:08 db-
unza21-csc5741-ict1110_assessment_scores-quiz09.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1905 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz10.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1955 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz11.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1981 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz12.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1969 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz13.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1983 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz14.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1987 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz15.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1729 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz16.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1676 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz17.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1646 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz18.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1748 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz19.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 1862 May 31 15:06 db-
unza21-csc5741-ict1110_assessment_scores-quiz20.csv
```

[81]:
```
!ls -l db-unza21-csc5741-ict1110_assessment_scores-test* | wc -l
!ls -l db-unza21-csc5741-ict1110_assessment_scores-test*
!cat db-unza21-csc5741-ict1110_assessment_scores-test1.csv | head
```

```
4
-rw-rw-r-- 1 lightonphiri lightonphiri 2196 May 30 16:07 db-
unza21-csc5741-ict1110_assessment_scores-test1.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 2124 May 30 16:07 db-
unza21-csc5741-ict1110_assessment_scores-test2.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 2074 May 30 16:07 db-
unza21-csc5741-ict1110_assessment_scores-test3.csv
-rw-rw-r-- 1 lightonphiri lightonphiri 2088 May 30 16:07 db-
unza21-csc5741-ict1110_assessment_scores-test4.csv
Student ID|Total
07f3ca235faaa1c9ad16facef5526d8b|31.5
921855f753932de762b780405a50bdf7|40
1bdb61bdbf5bc9a2cdc5db037ab610e0|17
d2e4449b45608e33e472d939a73868f7|23.5
cea34f6b4f356c28fc2b766ae46b6d6c|16
d7fe73b6846dfc672236e636aa2cf6b5|13
8e4d9eeed250a9d065ac2bb8bdc67b30|23
bf2ec44b27fc25c6fd8a38792b9ef2a8|22
aa293d284f52b08da5ba7fe7792fe9c3|7
```

[82]:
```
!ls -l db-unza21-csc5741-ict1110_assessment_scores-final_examination* | wc -l
!ls -l db-unza21-csc5741-ict1110_assessment_scores-final_examination*
!cat db-unza21-csc5741-ict1110_assessment_scores-final_examination.csv | head
```

```
1
-rw-rw-r-- 1 lightonphiri lightonphiri 2147 May 30 16:06 db-
unza21-csc5741-ict1110_assessment_scores-final_examination.csv
Student ID|Total
9d5116a2451bc98c2b46b93acbc1b4f0|46.5
e7400496f1ce70cb62c2c44ca2ddc469|48.5
cea34f6b4f356c28fc2b766ae46b6d6c|53
#N/A|90
6cd50fb3091b0a9d3c1ac2cf52441390|64.5
e31959fe2842dacea4d16d36e9813620|41
97527dec0ae1a703599581d4f25dfbce|5
9e7002d53d4db7bfad4f5cf419b0c126|51
74458a3d3e5f3074226b1f9fa23c9163|67
```

Dataframe Creation

Final Examination Scores

Create Dataframes

[83]:
```
# Create DataFrame of input dataset: ICT 1110 Demographics
#
var_ict1110_assessments_examination = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-final_examination.csv",␣
 ↪sep="|")
var_ict1110_assessments_examination.columns
```

[83]:
```
Index(['Student ID', 'Total'], dtype='object')
```

Rename Dataframe Names

[84]:
```
# Rename dataframe columns for easy processing
#
# Function use: pd.rename([...])
# Important parameter: inplace=True
#
var_ict1110_assessments_examination.rename(columns={
    "Student ID": "StudentID",
    "Total": "ExaminationScore"
}, inplace=True)

var_ict1110_assessments_examination.columns
```

[84]:
```
Index(['StudentID', 'ExaminationScore'], dtype='object')
```

[85]:
```
# Count records in dataframe
#
len(var_ict1110_assessments_examination)
```

[85]:
```
60
```

[86]:
```
# Inspect some dataframe records
#
var_ict1110_assessments_examination.head(5)
```

[86]:

	StudentID	ExaminationScore
0	9d5116a2451bc98c2b46b93acb...	46.5
1	e7400496f1ce70cb62c2c44ca2...	48.5
2	cea34f6b4f356c28fc2b766ae4...	53.0
3	NaN	90.0
4	6cd50fb3091b0a9d3c1ac2cf52...	64.5

Drop Records With Null StudentID Values

[87]:
```
# Delete all test score entries with null StudentID values
#
#
print("Examination Records Before Drop Operation: ",␣
 ↪len(var_ict1110_assessments_examination))
var_ict1110_assessments_examination.dropna(subset = ["StudentID"], inplace=True)
print("Examination Records After Drop Operation: ",␣
 ↪len(var_ict1110_assessments_examination))
```

```
Examination Records Before Drop Operation:  60
Examination Records After Drop Operation:  58
```

Handle Examination Score Null Values

All null score entries will be replaced by 0

[88]:
```
var_ict1110_assessments_examination.fillna(0, inplace=True)
```

[89]:
```
# Inspect dataframe
var_ict1110_assessments_examination.head(2).T
```

[89]:

	0	1
StudentID	9d5116a2451bc98c2b46b93acb...	e7400496f1ce70cb62c2c44ca2...
ExaminationScore	46.5	48.5

Test Scores

- There are a total of four(4) test data files. Each of these will be loaded into a seperate dataframe.
- Test files will be subsequently merged into a single test scores dataframe

Create Dataframes

[90]:
```
# Create DataFrame of input dataset: ICT 1110 Test Score Results
#
var_ict1110_assessments_test1 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-test1.csv", sep="|")
```

```
var_ict1110_assessments_test1.columns

#
var_ict1110_assessments_test2 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-test2.csv", sep="|")
var_ict1110_assessments_test2.columns

#
var_ict1110_assessments_test3 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-test3.csv", sep="|")
var_ict1110_assessments_test3.columns

#
var_ict1110_assessments_test4 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-test4.csv", sep="|")
var_ict1110_assessments_test4.columns
```

[90]: `Index(['Student ID', 'Total'], dtype='object')`

[90]: `Index(['Student ID', 'Total'], dtype='object')`

[90]: `Index(['Student ID', 'Total'], dtype='object')`

[90]: `Index(['Student ID', 'Total'], dtype='object')`

Rename Dataframe Names

[91]:
```
# Rename dataframe columns for easy processing
#
# Function use: pd.rename([...])
# Important parameter: inplace=True
#
var_ict1110_assessments_test1.rename(columns={
    "Student ID": "StudentID",
    "Total": "Test1Score"
}, inplace=True)

var_ict1110_assessments_test1.columns

#
var_ict1110_assessments_test2.rename(columns={
    "Student ID": "StudentID",
    "Total": "Test2Score"
}, inplace=True)

var_ict1110_assessments_test2.columns

#
var_ict1110_assessments_test3.rename(columns={
    "Student ID": "StudentID",
    "Total": "Test3Score"
}, inplace=True)
```

```
var_ict1110_assessments_test3.columns

#
var_ict1110_assessments_test4.rename(columns={
    "Student ID": "StudentID",
    "Total": "Test4Score"
}, inplace=True)

var_ict1110_assessments_test4.columns
```

[91]: Index(['StudentID', 'Test1Score'], dtype='object')

[91]: Index(['StudentID', 'Test2Score'], dtype='object')

[91]: Index(['StudentID', 'Test3Score'], dtype='object')

[91]: Index(['StudentID', 'Test4Score'], dtype='object')

[92]:
```
# Count records in dataframes
#
len(var_ict1110_assessments_test1)

#
len(var_ict1110_assessments_test2)

#
len(var_ict1110_assessments_test3)

#
len(var_ict1110_assessments_test4)
```

[92]: 61

[92]: 59

[92]: 59

[92]: 57

[93]:
```
# Inspect some dataframe records
#
var_ict1110_assessments_test1.head(5)

#
var_ict1110_assessments_test2.head(5)

#
var_ict1110_assessments_test3.head(5)

#
```

```
var_ict1110_assessments_test4.head(5)
```

[93]:

	StudentID	Test1Score
0	07f3ca235faaa1c9ad16facef5...	31.5
1	921855f753932de762b780405a...	40.0
2	1bdb61bdbf5bc9a2cdc5db037a...	17.0
3	d2e4449b45608e33e472d939a7...	23.5
4	cea34f6b4f356c28fc2b766ae4...	16.0

[93]:

	StudentID	Test2Score
0	e2a96e074e1d8a6f6de56abbd4...	28.0
1	4234d1794dd33c1b6ed975eab5...	19.5
2	4be25f9d27da71d4e98775668b...	23.0
3	c89bc418c38da77213c6c6e03c...	15.5
4	b0aa0804e676a38255af4fd702...	15.5

[93]:

	StudentID	Test3Score
0	81975d05c61d8de83f46487739...	18
1	e99bb6b91ef51dbe4eec9340dc...	22
2	28765464efe1b6583610335965...	15
3	b2ce07e6bc0b55222aa66fcc1b...	18
4	NaN	1

[93]:

	StudentID	Test4Score
0	b2ce07e6bc0b55222aa66fcc1b...	17.0
1	afe1f0ece97fc8839cc4822a67...	6.0
2	921855f753932de762b780405a...	22.0
3	2d65f5236205dd23c6a8212627...	13.5
4	b0aa0804e676a38255af4fd702...	22.5

Drop Records With Null StudentID Values

[94]:
```
# Delete all test score entries with null StudentID values
#
#
print("Test #1 Records Before Drop Operation: ", len(var_ict1110_assessments_test1))
var_ict1110_assessments_test1.dropna(subset = ["StudentID"], inplace=True)
print("Test #1 Records After Drop Operation: ", len(var_ict1110_assessments_test1))
#
print("Test #2 Records Before Drop Operation: ", len(var_ict1110_assessments_test2))
var_ict1110_assessments_test2.dropna(subset = ["StudentID"], inplace=True)
print("Test #2 Records After Drop Operation: ", len(var_ict1110_assessments_test2))
#
print("Test #3 Records Before Drop Operation: ", len(var_ict1110_assessments_test3))
```

```
var_ict1110_assessments_test3.dropna(subset = ["StudentID"], inplace=True)
print("Test #3 Records After Drop Operation: ", len(var_ict1110_assessments_test3))
#
print("Test #4 Records Before Drop Operation: ", len(var_ict1110_assessments_test4))
var_ict1110_assessments_test4.dropna(subset = ["StudentID"], inplace=True)
print("Test #4 Records After Drop Operation: ", len(var_ict1110_assessments_test4))
```

```
Test #1 Records Before Drop Operation:  61
Test #1 Records After Drop Operation:  59
Test #2 Records Before Drop Operation:  59
Test #2 Records After Drop Operation:  57
Test #3 Records Before Drop Operation:  59
Test #3 Records After Drop Operation:  57
Test #4 Records Before Drop Operation:  57
Test #4 Records After Drop Operation:  56
```

Merge All Tests

[95]:
```
from functools import reduce
# READ: https://stackoverflow.com/a/44338256/664424
# Outlines how to merge multiple datasets
#
# Create a list to be used to hold all the test dataframes
var_ict1110_assessments_test_dataframes =␣
 ↪[var_ict1110_assessments_test1,var_ict1110_assessments_test2,var_ict1110_assessments_test3,var_

# Merge all the quizzes into one dataframe
var_ict1110_assessments_tests = reduce(lambda  left,right: pd.
 ↪merge(left,right,on=['StudentID'],
                                            how='outer'),␣
 ↪var_ict1110_assessments_test_dataframes)
```

[96]:
```
# Inspect merged dataframe records
#
var_ict1110_assessments_tests.tail(2).T
```

[96]:

	57	58
StudentID	4234d1794dd33c1b6ed975eab5...	232bf11cb81bcdb269f76a08fd...
Test1Score	31	22.5
Test2Score	19.5	17
Test3Score	28	17
Test4Score	22	20.5

Handle Test Score Null Values

All null score entries will be replaced by 0

[97]:
```
var_ict1110_assessments_tests.fillna(0, inplace=True)
```

[98]:
```
# Inspect dataframe
var_ict1110_assessments_tests.head(2).T
```

[98]:

	0	1
StudentID	07f3ca235faaa1c9ad16facef5...	921855f753932de762b780405a...
Test1Score	31.5	40
Test2Score	29.5	16
Test3Score	34	24
Test4Score	29.5	22

Quiz Scores

Create Dataframes

[99]:
```
#
# Instead of manually creating 20 dataframes, this can be done dynamically using a␣
 ↪dictionary
#
var_quiz_dataframes = {} # dictionary to hold all quiz dataframes
for var_quiz in range(1, 21, 1):
    var_quiz_dataframe_key = "var_ict1110_assessments_quiz_"+str(var_quiz)
    if var_quiz < 10:
        var_quiz_dataframe_input_file =␣
 ↪"db-unza21-csc5741-ict1110_assessment_scores-quiz0"+str(var_quiz)+".csv"
    else:
        var_quiz_dataframe_input_file =␣
 ↪"db-unza21-csc5741-ict1110_assessment_scores-quiz"+str(var_quiz)+".csv"
    var_quiz_dataframes[var_quiz_dataframe_key] = pd.
 ↪read_csv(var_quiz_dataframe_input_file, sep="|")
```

[100]:
```
var_quiz_dataframes["var_ict1110_assessments_quiz_20"].head(5)
```

[100]:

	Student ID	Mark
0	28765464efe1b6583610335965...	9
1	74458a3d3e5f3074226b1f9fa2...	9
2	9d5116a2451bc98c2b46b93acb...	2
3	NaN	9
4	94984a8c4896946d9bafd24959...	10

[101]:
```
# Create DataFrame of input dataset: ICT 1110 Quiz Score Results
#
# Code below can be processed intelligently
#
var_ict1110_assessments_quiz1 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz01.csv", sep="|")
var_ict1110_assessments_quiz1.columns

#
var_ict1110_assessments_quiz2 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz02.csv", sep="|")
var_ict1110_assessments_quiz2.columns
```

```
#
var_ict1110_assessments_quiz3 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz03.csv", sep="|")
var_ict1110_assessments_quiz3.columns

#
var_ict1110_assessments_quiz4 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz04.csv", sep="|")
var_ict1110_assessments_quiz4.columns

#
var_ict1110_assessments_quiz5 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz05.csv", sep="|")
var_ict1110_assessments_quiz5.columns

#
var_ict1110_assessments_quiz6 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz06.csv", sep="|")
var_ict1110_assessments_quiz6.columns

#
var_ict1110_assessments_quiz7 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz07.csv", sep="|")
var_ict1110_assessments_quiz7.columns

#
var_ict1110_assessments_quiz8 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz08.csv", sep="|")
var_ict1110_assessments_quiz8.columns

#
var_ict1110_assessments_quiz9 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz09.csv", sep="|")
var_ict1110_assessments_quiz9.columns

#
var_ict1110_assessments_quiz10 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz10.csv", sep="|")
var_ict1110_assessments_quiz10.columns

#
var_ict1110_assessments_quiz11 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz11.csv", sep="|")
var_ict1110_assessments_quiz11.columns

#
var_ict1110_assessments_quiz12 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz12.csv", sep="|")
var_ict1110_assessments_quiz12.columns
```

```
#
var_ict1110_assessments_quiz13 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz13.csv", sep="|")
var_ict1110_assessments_quiz13.columns

#
var_ict1110_assessments_quiz14 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz14.csv", sep="|")
var_ict1110_assessments_quiz14.columns

#
var_ict1110_assessments_quiz15 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz15.csv", sep="|")
var_ict1110_assessments_quiz15.columns

#
var_ict1110_assessments_quiz16 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz16.csv", sep="|")
var_ict1110_assessments_quiz16.columns

#
var_ict1110_assessments_quiz17 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz17.csv", sep="|")
var_ict1110_assessments_quiz17.columns

#
var_ict1110_assessments_quiz18 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz18.csv", sep="|")
var_ict1110_assessments_quiz18.columns

#
var_ict1110_assessments_quiz19 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz19.csv", sep="|")
var_ict1110_assessments_quiz19.columns

#
var_ict1110_assessments_quiz20 = pd.
 ↪read_csv("db-unza21-csc5741-ict1110_assessment_scores-quiz20.csv", sep="|")
var_ict1110_assessments_quiz20.columns
```

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

[101]: `Index(['Student ID', 'Mark'], dtype='object')`

Rename Dataframe Names

[102]:
```
#
# Loop through all quizzes and print first two records
#
for var_quiz in range(1, 21, 1):
    # dynamically create key for accessing dictionary value
    var_dataframe_key = "var_ict1110_assessments_quiz_"+str(var_quiz)
    var_quiz_dataframes[var_dataframe_key].rename(columns={
        "Student ID": "StudentID",
        "Mark": "Quiz"+str(var_quiz)+"Score"
    }, inplace=True)
```

[103]:
```
var_quiz_dataframes["var_ict1110_assessments_quiz_13"].head(5)
```

[103]:

	StudentID	Quiz13Score
0	e99bb6b91ef51dbe4eec9340dc...	1.0
1	cea34f6b4f356c28fc2b766ae4...	1.0
2	1bdb61bdbf5bc9a2cdc5db037a...	0.5
3	b14c3890a187a7798035ac60d8...	0.5

Continued on next page

	StudentID	Quiz13Score
4	9c05c40b81fd906b1585e231c0...	0.5

[104]:
```
# Rename dataframe columns for easy processing
#
# Function use: pd.rename([...])
# Important parameter: inplace=True
#
var_ict1110_assessments_quiz1.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz1Score"
}, inplace=True)

var_ict1110_assessments_quiz1.columns

#
var_ict1110_assessments_quiz2.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz2Score"
}, inplace=True)

var_ict1110_assessments_quiz2.columns

#
var_ict1110_assessments_quiz3.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz3Score"
}, inplace=True)

var_ict1110_assessments_quiz3.columns

#
var_ict1110_assessments_quiz4.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz4Score"
}, inplace=True)

var_ict1110_assessments_quiz4.columns

#
var_ict1110_assessments_quiz5.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz5Score"
}, inplace=True)

var_ict1110_assessments_quiz5.columns

#
var_ict1110_assessments_quiz6.rename(columns={
    "Student ID": "StudentID",
```

```
    "Mark": "Quiz6Score"
}, inplace=True)

var_ict1110_assessments_quiz6.columns

#
var_ict1110_assessments_quiz7.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz7Score"
}, inplace=True)

var_ict1110_assessments_quiz7.columns

#
var_ict1110_assessments_quiz8.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz8Score"
}, inplace=True)

var_ict1110_assessments_quiz8.columns

#
var_ict1110_assessments_quiz9.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz9Score"
}, inplace=True)

var_ict1110_assessments_quiz9.columns

#
var_ict1110_assessments_quiz10.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz10Score"
}, inplace=True)

var_ict1110_assessments_quiz10.columns

#
var_ict1110_assessments_quiz11.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz11Score"
}, inplace=True)

var_ict1110_assessments_quiz11.columns

#
var_ict1110_assessments_quiz12.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz12Score"
}, inplace=True)
```

```
var_ict1110_assessments_quiz12.columns

#
var_ict1110_assessments_quiz13.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz13Score"
}, inplace=True)

var_ict1110_assessments_quiz13.columns

#
var_ict1110_assessments_quiz14.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz14Score"
}, inplace=True)

var_ict1110_assessments_quiz14.columns

#
var_ict1110_assessments_quiz15.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz15Score"
}, inplace=True)

var_ict1110_assessments_quiz15.columns

#
var_ict1110_assessments_quiz16.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz16Score"
}, inplace=True)

var_ict1110_assessments_quiz16.columns

#
var_ict1110_assessments_quiz17.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz17Score"
}, inplace=True)

var_ict1110_assessments_quiz17.columns

#
var_ict1110_assessments_quiz18.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz18Score"
}, inplace=True)

var_ict1110_assessments_quiz18.columns

#
```

```
var_ict1110_assessments_quiz19.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz19Score"
}, inplace=True)

var_ict1110_assessments_quiz19.columns

#
var_ict1110_assessments_quiz20.rename(columns={
    "Student ID": "StudentID",
    "Mark": "Quiz20Score"
}, inplace=True)

var_ict1110_assessments_quiz20.columns
```

[104]: Index(['StudentID', 'Quiz1Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz2Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz3Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz4Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz5Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz6Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz7Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz8Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz9Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz10Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz11Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz12Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz13Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz14Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz15Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz16Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz17Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz18Score'], dtype='object')

[104]: Index(['StudentID', 'Quiz19Score'], dtype='object')

[104]:
```
Index(['StudentID', 'Quiz20Score'], dtype='object')
```

Drop Records With Null StudentID Values

[105]:
```
# Delete all quiz entries with null StudentID values
#
for var_quiz in range(1, 21, 1):
    # dynamically create key for accessing dictionary value
    var_dataframe_key = "var_ict1110_assessments_quiz_"+str(var_quiz)
    var_quiz_dataframes[var_dataframe_key].dropna(subset = ["StudentID"], inplace=True)
```

Merge All Quizzes

[106]:
```
from functools import reduce
# READ: https://stackoverflow.com/a/44338256/664424
# Outlines how to merge multiple datasets
#
# Create a list to be used to hold all the quiz dataframes
var_ict1110_assessments_quiz_dataframes = list(var_quiz_dataframes.values())

# Merge all the quizzes into one dataframe
var_ict1110_assessments_quizzes = reduce(lambda  left,right: pd.
↪merge(left,right,on=['StudentID'],
                                            how='outer'),␣
↪var_ict1110_assessments_quiz_dataframes)
```

[107]:
```
# Inspect merged dataframe records
#
var_ict1110_assessments_quizzes.tail(2).T
```

[107]:

	62	63
StudentID	aa293d284f52b08da5ba7fe779...	c89bc418c38da77213c6c6e03c...
Quiz1Score	NaN	NaN
Quiz2Score	NaN	NaN
Quiz3Score	0	NaN
Quiz4Score	1.5	1.5
Quiz5Score	2.5	6
Quiz6Score	4	4
Quiz7Score	0	2
Quiz8Score	3	4
Quiz9Score	NaN	4
Quiz10Score	NaN	3.5
Quiz11Score	1	3
Quiz12Score	5	10
Quiz13Score	0	0.5
Quiz14Score	NaN	3
Quiz15Score	7	NaN
Quiz16Score	NaN	1.5
Quiz17Score	NaN	10
Quiz18Score	NaN	1

Continued on next page

	62	63
Quiz19Score	NaN	9
Quiz20Score	NaN	10

Handle Quiz Score Null Values

All null score entries will be replaced by 0

[108]:
```
var_ict1110_assessments_quizzes.fillna(0, inplace=True)
```

[109]:
```
# Inspect dataframe
var_ict1110_assessments_quizzes.head(2).T
```

[109]:

	0	1
StudentID	53b3c88ea00c4f0e137b4e6fe7...	c03b1123e45fa00da3142e0424...
Quiz1Score	1	2
Quiz2Score	5	7
Quiz3Score	3	1
Quiz4Score	4	1
Quiz5Score	5.5	4.5
Quiz6Score	6	9
Quiz7Score	1	1
Quiz8Score	0	4
Quiz9Score	4	9
Quiz10Score	2	4.5
Quiz11Score	4	3
Quiz12Score	10	10
Quiz13Score	0	6
Quiz14Score	4	6
Quiz15Score	9	7
Quiz16Score	0	0
Quiz17Score	0	0
Quiz18Score	0	9
Quiz19Score	9	9
Quiz20Score	7	9

Dataset Attributes

- StudentID—Alphanumeric
- QuizNScore—Numeric
- TestNScore—Numeric
- ExaminationScore—Numeric

Data Pre-processing Plan

- STEP 1: Remove all records with NULL StudentID values
- STEP 2: Remove duplicates using StudentID as unique field

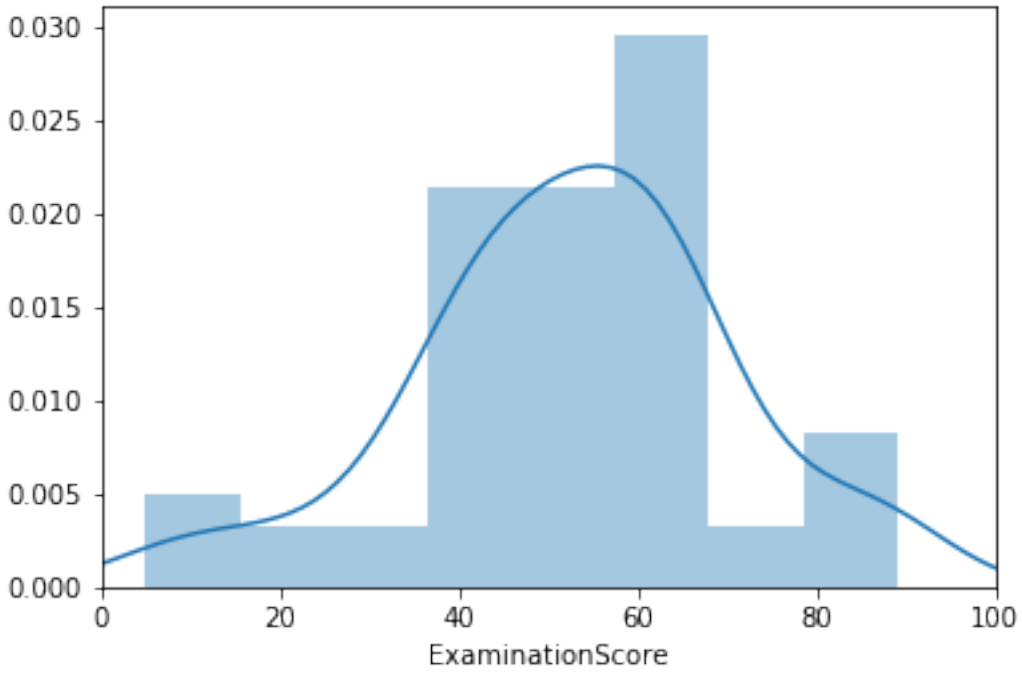
NOTE: Null values to be handled on a case-by-case basis; e.g. null values in text attributes to be replaced with empty strings ""

Exploratory Data Analysis

Examination Scores

[110]:
```
sns.distplot(var_ict1110_assessments_examination["ExaminationScore"]).set(xlim=(0,␣
↪100))
```

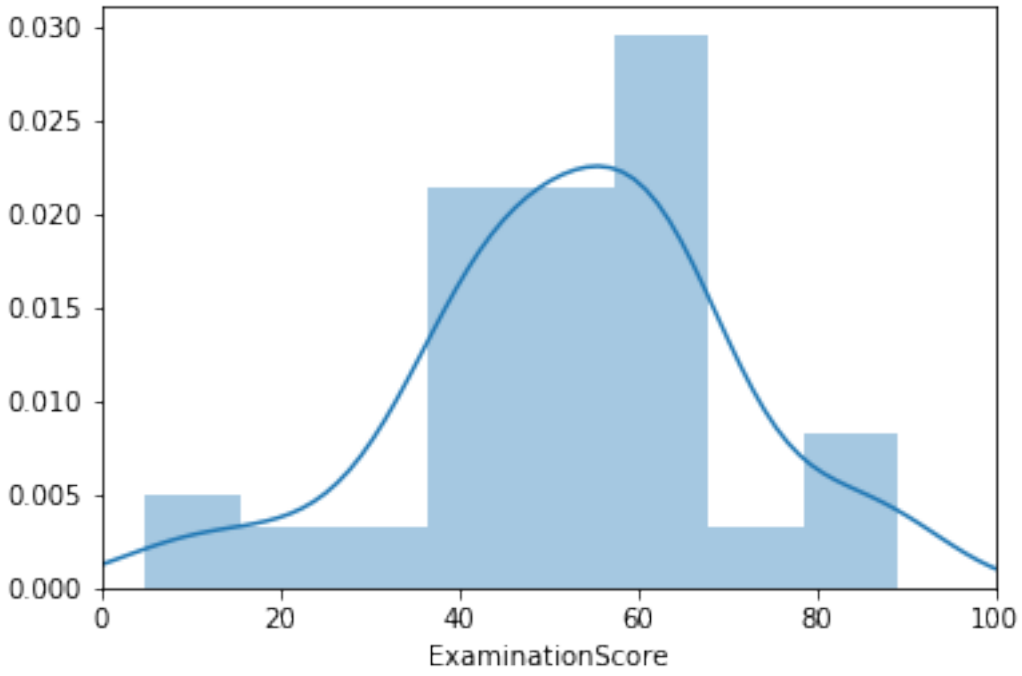
[110]:
```
[(0, 100)]
```



Test Scores

[111]:
```
sns.distplot(var_ict1110_assessments_examination["ExaminationScore"]).set(xlim=(0,␣
↪100))
```

[111]:
```
[(0, 100)]
```



[112]:
```
# Facet results by academic year

fig, (ax1, ax2, ax3, ax4) = plt.subplots(ncols=4, nrows=1, figsize=(15,5))

fig.suptitle('Test Scores Distributions')

sns.distplot(var_ict1110_assessments_tests["Test1Score"], ax=ax1).set(xlim=(0, 50))
sns.distplot(var_ict1110_assessments_tests["Test2Score"], ax=ax2).set(xlim=(0, 50))
sns.distplot(var_ict1110_assessments_tests["Test3Score"], ax=ax3).set(xlim=(0, 50))
sns.distplot(var_ict1110_assessments_tests["Test4Score"], ax=ax4).set(xlim=(0, 50))
```

[112]:
```
Text(0.5, 0.98, 'Test Scores Distributions')
```

[112]:
```
[(0, 50)]
```

[112]:
```
[(0, 50)]
```

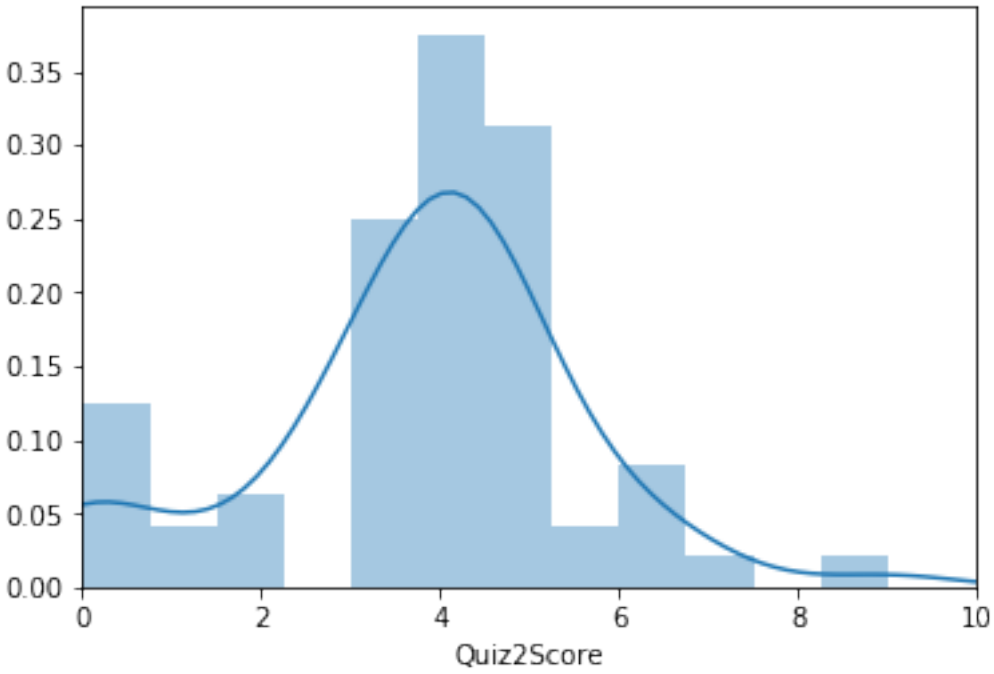
[112]:
```
[(0, 50)]
```

[112]:
```
[(0, 50)]
```

Quiz Scores

[113]:
```
sns.distplot(var_ict1110_assessments_quizzes["Quiz2Score"]).set(xlim=(0, 10))
```

[113]:
```
[(0, 10)]
```



[114]:
```
# Facet results by academic year

fig, ((ax1, ax2, ax3, ax4), (ax5, ax6, ax7, ax8), (ax9, ax10, ax11, ax12), (ax13,␣
 ↪ax14, ax15, ax16), (ax17, ax18, ax19, ax20)) = plt.subplots(ncols=4, nrows=5,␣
 ↪figsize=(15,20))
```

```
fig.suptitle('Quiz Scores Distributions')

sns.distplot(var_ict1110_assessments_quizzes["Quiz1Score"], ax=ax1).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz2Score"], ax=ax2).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz3Score"], ax=ax3).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz4Score"], ax=ax4).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz5Score"], ax=ax5).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz6Score"], ax=ax6).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz7Score"], ax=ax7).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz8Score"], ax=ax8).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz9Score"], ax=ax9).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz10Score"], ax=ax10).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz11Score"], ax=ax11).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz12Score"], ax=ax12).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz13Score"], ax=ax13).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz14Score"], ax=ax14).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz15Score"], ax=ax15).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz16Score"], ax=ax16).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz17Score"], ax=ax17).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz18Score"], ax=ax18).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz19Score"], ax=ax19).set(xlim=(0, 10))
sns.distplot(var_ict1110_assessments_quizzes["Quiz20Score"], ax=ax20).set(xlim=(0, 10))
```

[114]: Text(0.5, 0.98, 'Quiz Scores Distributions')

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: [(0, 10)]

[114]: `[(0, 10)]`

[114]: `[(0, 10)]`

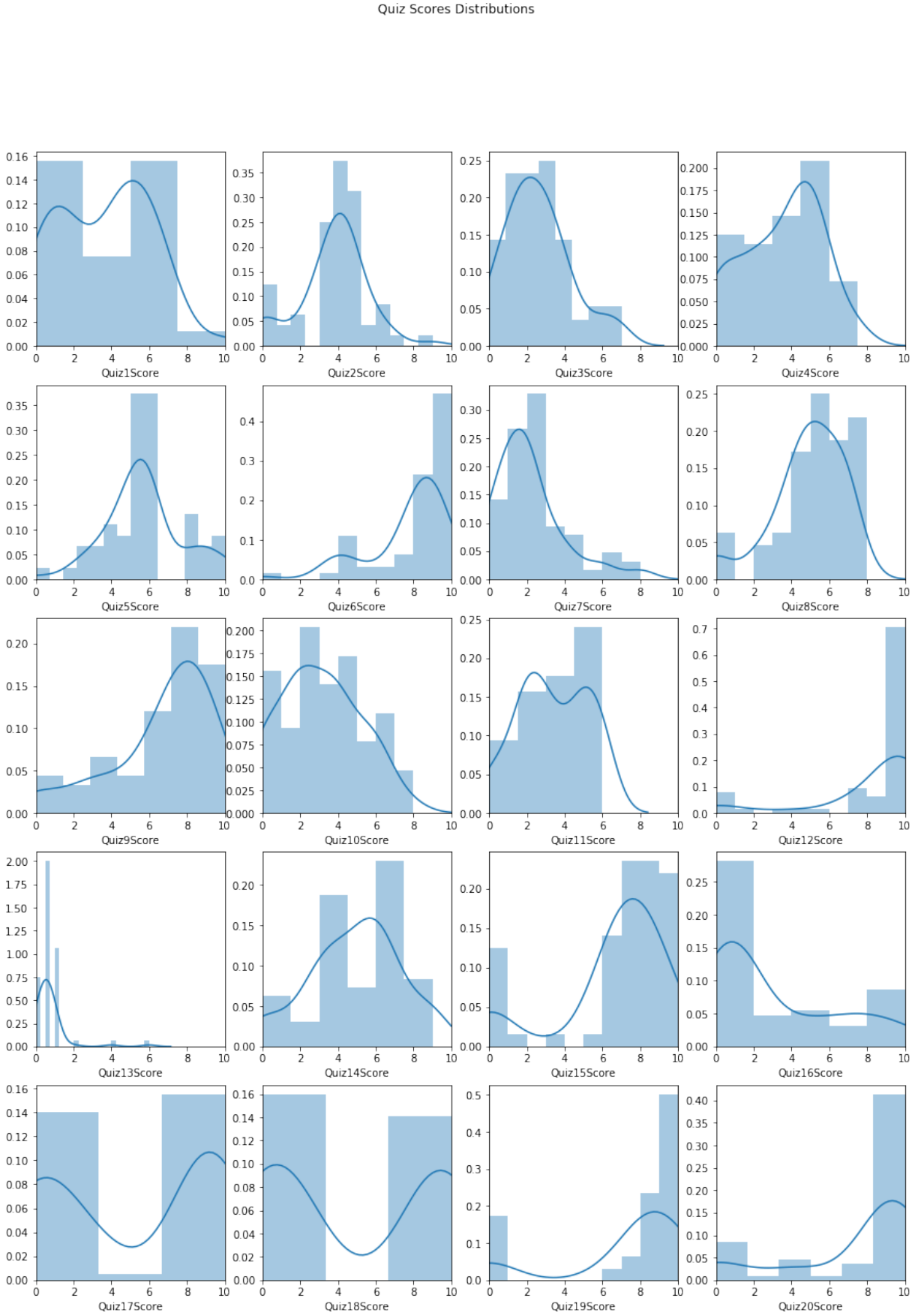
[114]: `[(0, 10)]`

[114]: `[(0, 10)]`

[114]: `[(0, 10)]`

[114]: `[(0, 10)]`

[114]: `[(0, 10)]`


Quiz Scores Distributions
Quiz1Score
Quiz2Score
Quiz3Score
Quiz4Score
Quiz5Score
Quiz6Score
Quiz7Score
Quiz8Score
Quiz9Score
Quiz10Score
Quiz11Score
Quiz12Score
Quiz13Score
Quiz14Score
Quiz15Score
Quiz16Score
Quiz17Score
Quiz18Score
Quiz19Score
Quiz20Score

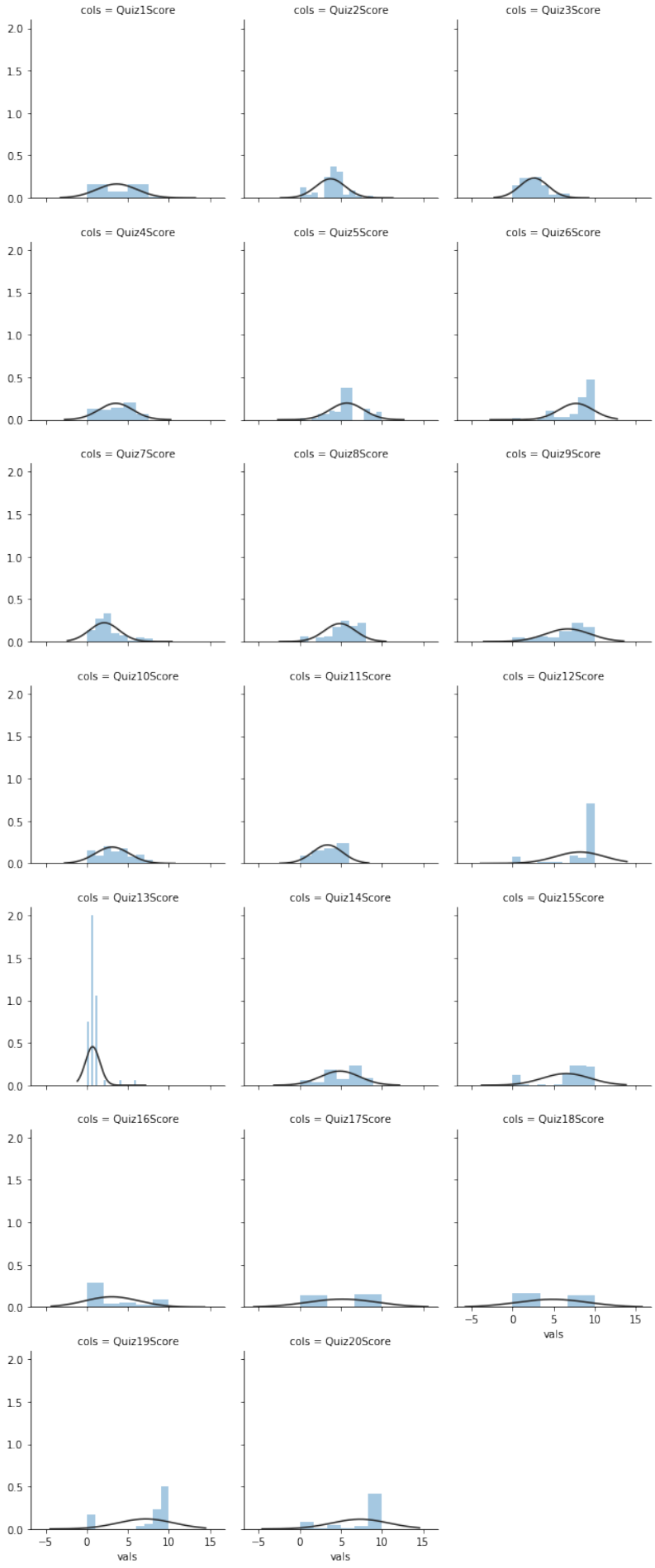
[115]:
```
#
#
#####var_ict1110_assessments_quizzes[["Quiz1Score", "Quiz2Score", "Quiz3Score",␣
 ↪"Quiz4Score", "Quiz5Score"]].plot(kind="hist")

var_ict1110_assessments_quizzes_ = var_ict1110_assessments_quizzes.melt('StudentID',␣
 ↪var_name='cols', value_name='vals')
```

[116]:
```
var_ict1110_assessments_quizzes_.head(5)
```

[116]:

	StudentID	cols	vals
0	53b3c88ea00c4f0e137b4e6fe7...	Quiz1Score	1.0
1	c03b1123e45fa00da3142e0424...	Quiz1Score	2.0
2	74458a3d3e5f3074226b1f9fa2...	Quiz1Score	2.0
3	e2a96e074e1d8a6f6de56abbd4...	Quiz1Score	1.0
4	e2a96e074e1d8a6f6de56abbd4...	Quiz1Score	1.0

[117]:
```
#
from scipy.stats import norm
# READ: https://seaborn.pydata.org/generated/seaborn.distplot.html
#
#
g = sns.FacetGrid(var_ict1110_assessments_quizzes_, col="cols", col_wrap=3)
g = g.map(sns.distplot, "vals", fit=norm, kde=False, hist=True)
```


cols = Quiz1Score
cols = Quiz2Score
cols = Quiz3Score
cols = Quiz4Score
cols = Quiz5Score
cols = Quiz6Score
cols = Quiz7Score
cols = Quiz8Score
cols = Quiz9Score
cols = Quiz10Score
cols = Quiz11Score
cols = Quiz12Score
cols = Quiz13Score
cols = Quiz14Score
cols = Quiz15Score
cols = Quiz16Score
cols = Quiz17Score
cols = Quiz18Score
cols = Quiz19Score
cols = Quiz20Score
vals
vals

Lightweight Pipelining With JobLib

There are usually a number of pipelining steps involved when working on data mining problems. Libraries like joblib make it possible for you to save the state of models.

> joblib.dump() and joblib.load() provide a replacement for pickle to work efficiently on arbitrary Python objects containing large data, in particular large numpy arrays. https://joblib.readthedocs.io/en/latest/persistence.html

In this instance, we will save the dataframes assocaited with all the data sources

- *Please also READ: https://stackoverflow.com/a/12617603/664424*

Save Initial Survey Dataframes

[118]:
```
# Import joblib
import joblib

# Initial Survey
joblib.dump(var_ict1110_survey_eda, "var_ict1110_survey_eda_dataframe.pkl")
```

[118]:
```
['var_ict1110_survey_eda_dataframe.pkl']
```

Save Demographic Dataframes

[119]:
```
# Import joblib
import joblib

# Demographic Details
joblib.dump(var_ict1110_demographics_eda, "var_ict1110_demographics_eda_dataframe.pkl")
```

[119]:
```
['var_ict1110_demographics_eda_dataframe.pkl']
```

Save Assessments Dataframes

[120]:
```
# Import joblib
#
import joblib

# Quizzes
joblib.dump(var_ict1110_assessments_quizzes,␣
 ↪"var_ict1110_assessments_quizzes_dataframe.pkl")
#
# Tests
joblib.dump(var_ict1110_assessments_tests, "var_ict1110_assessments_tests_dataframe.
 ↪pkl")
#
# Examination
```

```
joblib.dump(var_ict1110_assessments_examination,␣
 ↪"var_ict1110_assessments_examination_dataframe.pkl")
```

[120]: ['var_ict1110_assessments_quizzes_dataframe.pkl']

[120]: ['var_ict1110_assessments_tests_dataframe.pkl']

[120]: ['var_ict1110_assessments_examination_dataframe.pkl']

[121]:
```
var_example_saved_state = joblib.load("var_ict1110_assessments_examination_dataframe.
 ↪pkl")
```

[122]:
```
var_example_saved_state.columns
len(var_example_saved_state)
```

[122]: Index(['StudentID', 'ExaminationScore'], dtype='object')

[122]: 58