# CSC 5741: Jupyter Notebook—Auxiliary Data Preparation

Lighton Phiri
<lighton.phiri@unza.zm>

June 7 2021

## Contents

# Introduction

In this Jupyter Notebook, perform the following basic data preparation tasks:

1. Derivation of fields
2. Data formatting and scaling
3. Data integration and merging

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.

## Recap on CRISP-DM Data Preparation Phase

A reminder about the following key activities conducted as part of data preparation:

- Derivation of new attributes from existing ones
- Scaling of attributes using appropriate ranges
- Data transformation into a form expected by the estimators—this will be discussed in the subsequent notebook

## 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 joblib
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
```

## Load Save Pipelines

### Dataset #1: Initial Survey

[3]:
```
# Initial Survey
var_ict1110_survey_eda = joblib.load("var_ict1110_survey_eda_dataframe.pkl")
```

[4]:
```
# Inspect loaded dataframe
var_ict1110_survey_eda.columns
```

[4]:
```
Index(['Timestamp', 'StudentName', 'StudentID', 'HomeTown', 'MinorProgramme',
       'MinorProgrammeMotivation', 'MajorProgrammeMotivation',
       'DidComputerStudies', 'HasComputerTraining', 'ComputerTrainingType',
       'ExperienceWithComputers', 'HasComputerAccess', 'AboutMe', 'year'],
      dtype='object')
```

[5]:
```
len(var_ict1110_survey_eda)
var_ict1110_survey_eda.head(2).T
```

[5]:
```
90
```

[5]:

| | 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 |
| AboutMe | cycl everyday | day pass without joke feel... |
| year | 2019 | 2019 |

### Dataset #2: Demographic Details

[6]:
```
# Demographic Details
var_ict1110_demographics_eda = joblib.load("var_ict1110_demographics_eda_dataframe.
↪pkl")
```

[7]:
```
# Inspect loaded dataframe
var_ict1110_demographics_eda.columns
```

[7]:
```
Index(['StudentID', 'DateOfBirth', 'Gender', 'AcademicYear', 'YearOfStudy',
       'School', 'Program', 'MajorDescription', 'MinorDescription', 'Status',
       'Sponsor', 'Nationality', 'Accommodated'],
      dtype='object')
```

[8]:
```
var_ict1110_demographics_eda.head(2).T
```

[8]:

| | 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 #3: Assessment Scores

[9]:
```
# Quizzes
var_ict1110_assessments_quizzes = joblib.
 ↪load("var_ict1110_assessments_quizzes_dataframe.pkl")

# Tests
var_ict1110_assessments_tests = joblib.load("var_ict1110_assessments_tests_dataframe.
 ↪pkl")
#
# Examination
var_ict1110_assessments_examination = joblib.
 ↪load("var_ict1110_assessments_examination_dataframe.pkl")
```

[10]:
```
#
# Inspect loaded Quizzes dataframe
var_ict1110_assessments_quizzes.columns

#
# Inspect loaded Tests dataframe
var_ict1110_assessments_tests.columns

#
# Inspect loaded Examination dataframe
var_ict1110_assessments_examination.columns
```

[10]:
```
Index(['StudentID', 'Quiz1Score', 'Quiz2Score', 'Quiz3Score', 'Quiz4Score',
       'Quiz5Score', 'Quiz6Score', 'Quiz7Score', 'Quiz8Score', 'Quiz9Score',
```

```
       'Quiz10Score', 'Quiz11Score', 'Quiz12Score', 'Quiz13Score',
       'Quiz14Score', 'Quiz15Score', 'Quiz16Score', 'Quiz17Score',
       'Quiz18Score', 'Quiz19Score', 'Quiz20Score'],
      dtype='object')
```

[10]:
```
Index(['StudentID', 'Test1Score', 'Test2Score', 'Test3Score', 'Test4Score'],
dtype='object')
```

[10]:
```
Index(['StudentID', 'ExaminationScore'], dtype='object')
```

# Data Preparation

## Dataset #1: Initial Survey

- This should be done as an exercise
  - Identify attributes that require derivations

### Data Attribute Derivations

- The following attributes can be derived for the dataset
  - HomeTown—derive towns

[11]:
```
var_ict1110_survey_eda.head(1).T
```

[11]:

| | 0 |
|---|---|
| Timestamp | 2019/03/28 11:13:51 PM GMT+2 |
| StudentName | Participant1 |
| StudentID | NaN |
| HomeTown | Chudleigh/Lusaka/Lusaka |
| MinorProgramme | Data Mining |
| MinorProgrammeMotivation | love data |
| MajorProgrammeMotivation | love comput |
| DidComputerStudies | No |
| HasComputerTraining | Yes |
| ComputerTrainingType | I have studied Computer Sc... |
| ExperienceWithComputers | More than 5 years |
| HasComputerAccess | Yes |
| AboutMe | cycl everyday |
| year | 2019 |

[12]:
```
#
# Experiment with extracting the towns
#
var_ict1110_survey_eda["HomeTown"].str.replace("/", ",").str.split(",")
```

[12]:

| | HomeTown |
|---|---|
| 0 | [Chudleigh, Lusaka, Lusaka] |

Continued on next page

| | HomeTown |
|---|---|
| 1 | [Copperbelt, luanshya, Mpa... |
| 2 | [Mungule, senanga, western.] |
| 3 | [Lusaka] |
| 4 | [Lusaka ] |
| 5 | [shibuyunji, central pron... |
| 7 | [LUSAKA] |
| 8 | [Vorna valley, Lusaka, Lus... |
| 9 | [Tazara, mpika, muchinga] |
| 10 | [Ndeke village, kitwe, cop... |
| 12 | [Chazanga lusaka] |
| 13 | [Lusaka ] |
| 14 | [Pollen, kabwe, central] |
| 15 | [CHIPATA] |
| 16 | [KATETE, KATETE, EASTERN] |
| 18 | [Lusaka Garden compound ] |
| 19 | [Salama Park , Ibex Hill, ... |
| 20 | [Kaoma, Kaoma, Western] |
| 21 | [Mumbwa central] |
| 22 | [Kabangwe, Chibombo, Central] |
| 23 | [Mulonga, Mwense, Luapula] |
| 24 | [Lusaka] |
| 26 | [Lusaka] |
| 27 | [Nyimba, eastern province... |
| 28 | [Jesmondine, Lusaka, Lusaka] |
| 29 | [LUSAKA] |
| 30 | [DAMBWA, LIVINGSTONE] |
| 31 | [Chudleigh, Lusaka, Lusaka] |
| 32 | [Kanyama, Lusaka, Lusaka] |
| 34 | [Lusaka] |
| 36 | [8 miles , chibombo, Cen... |
| 37 | [Chamba valley, Lusaka] |
| 38 | [Airport, Sowezi, NWP] |
| 39 | [Libala lusaka] |
| 41 | [Rhodespark, Lusaka, Lusaka] |
| 42 | [State Lodge] |
| 43 | [Lusaka, Lusaka] |
| 44 | [Lusaka] |
| 45 | [Mandevu, Lusaka, Lusaka ] |
| 46 | [Hillcrest, Ndola, Copperb... |
| 47 | [Matero north, lusaka, lus... |
| 48 | [Lusaka] |
| 49 | [Lusaka, Lusaka, Lusaka] |
| 51 | [Mwembeshi, chilanga, lusaka] |
| 52 | [Masupe Estate Chipata Eas... |
| 53 | [(medium density, mansa, L... |
| 54 | [Chiwempala, Chingola, Cop... |
| 55 | [Lusaka] |
| 56 | [Meanwood ibex , Lusaka ,... |

Continued on next page

| | HomeTown |
|---|---|
| 57 | [Kamwala South, Lusaka, Lu... |
| 59 | [KABANANA, LUSAKA, LUSAKA ] |
| 60 | [Dambwa, Livingstone, Sout... |
| 61 | [New kanyama site and serv... |
| 62 | [Six miles, chibombo, cent... |
| 63 | [KUNGU, KASAMA, NORTHERN] |
| 64 | [Chama, Michinga Province] |
| 65 | [Lusaka, Lusaka, Lusaka] |
| 66 | [Kabanana] |
| 67 | [Hellen Kaunda, Lusaka, Lu... |
| 68 | [Chama, muchinga province] |
| 69 | [Chifubu] |
| 71 | [CHALAKWA, LUNDAZI, EASTERN] |
| 72 | [Lusaka] |
| 76 | [Garden compound Lusaka] |
| 77 | [lusaka\n\n] |
| 78 | [Chelstone, Lusaka, Lusaka] |
| 79 | [Ndola ] |
| 80 | [Livingstone ] |
| 81 | [Mazabuka\mazabuka, southern] |
| 82 | [Chipata Overapill, Lusaka... |
| 83 | [Lusaka ] |
| 84 | [Lusaka] |
| 85 | [Komboka, lusaka, lusaka] |
| 88 | [Kamwala south Lusaka] |
| 89 | [Matero, Lusaka, Lusaka ] |
| 90 | [Ndeke, Ndola, Copperbelt] |
| 92 | [Luapula, kazembe] |
| 94 | [Barlastone Villa, Lusaka,... |
| 95 | [Meanwood Chamba valley ] |
| 96 | [Mbala, northern province] |
| 97 | [Lilayi , Lusaka , Lusaka ] |
| 98 | [Barlaston park, Lusaka, L... |
| 100 | [Makeni Villa?Lusaka] |
| 101 | [Shampande , choma , south... |
| 102 | [Mumbwa District, Mumbwa T... |
| 106 | [Lusaka] |
| 107 | [Ibex, lusaka] |
| 108 | [Lusaka] |
| 110 | [Lusaka, Lusaka, Lusaka] |
| 111 | [Pamodzi Ndola] |

[13]:
```
#
# Implement function to filter out relevant town details
#
var_location_list = []
for var_location in var_ict1110_survey_eda["HomeTown"].str.replace("/", ",").str.
 ↪split(","):
    if len(var_location) > 1:
```

```
        var_location_list.append(var_location[1].strip().title())
    else:
        var_location_list.append(var_location[0].strip().title())

var_location_list[:10]
```

[13]:
```
['Lusaka',
 'Luanshya',
 'Senanga',
 'Lusaka',
 'Lusaka',
 'Central Pronvince',
 'Lusaka',
 'Lusaka',
 'Mpika',
 'Kitwe']
```

[14]:
```
#
# Implement function to filter out relevant provinces
#
var_location_list = []
for var_location in var_ict1110_survey_eda["HomeTown"].str.replace("/", ",").str.
 ↪split(","):
    if len(var_location) > 1:
        var_location_list.append(var_location[-1].strip().title())
    else:
        var_location_list.append(var_location[0].strip().title())

var_location_list[:10]

#
# Normalise provice names
# replace index entries where applicable
# Too tired to think so brute forcing things here [...]
#
for var_index, var_item in enumerate(var_location_list):
    if (var_item.lower().find("lusaka") != -1 or var_item.lower().find("state lodge") !
 ↪= -1 or var_item.lower().find("chongwe") != -1 or var_item.lower().find("kabanana") !
 ↪= -1 or var_item.lower().find("meanwood") != -1):
        var_location_list[var_index] = "Lusaka"
    elif (var_item.lower().find("ndola") != -1 or var_item.lower().find("mpatamato") !
 ↪= -1 or var_item.lower().find("chifubu") != -1):
        var_location_list[var_index] = "Copperbelt"
    elif (var_item.lower().find("central") != -1):
        var_location_list[var_index] = "Central"
    elif (var_item.lower().find("chinga") != -1):
        var_location_list[var_index] = "Muchinga"
    elif (var_item.lower().find("northern") != -1):
        var_location_list[var_index] = "Northern"
    elif (var_item.lower().find("western") != -1):
        var_location_list[var_index] = "Western"
```

```
    elif (var_item.lower().find("nwp") != -1):
        var_location_list[var_index] = "NorthWestern"
    elif (var_item.lower().find("chipata") != -1 or var_item.lower().find("eastern") !
↪= -1):
        var_location_list[var_index] = "Eastern"
    elif (var_item.lower().find("luapula") != -1 or var_item.lower().find("kazembe") !
↪= -1):
        var_location_list[var_index] = "Luapula"
    elif (var_item.lower().find("southern") != -1 or var_item.lower().
↪find("livingstone") != -1):
        var_location_list[var_index] = "Southern"
    else:
        var_location_list[var_index] = "MISSING DATA"

var_location_list[:10]
```

[14]:
```
['Lusaka',
 'Mpatamato',
 'Western.',
 'Lusaka',
 'Lusaka',
 'Central Pronvince',
 'Lusaka',
 'Lusaka',
 'Muchinga',
 'Copperbelt']
```

[14]:
```
['Lusaka',
 'Copperbelt',
 'Western',
 'Lusaka',
 'Lusaka',
 'Central',
 'Lusaka',
 'Lusaka',
 'Muchinga',
 'MISSING DATA']
```

[15]:
```
#
# Get unique town entries
#
# Central Pronvince: Kabwe
set(var_location_list)
```

[15]:
```
{'Central',
 'Copperbelt',
 'Eastern',
 'Luapula',
 'Lusaka',
 'MISSING DATA',
 'Muchinga',
```

```
'NorthWestern',
'Northern',
'Southern',
'Western'}
```

[16]:
```
#
# Create new DataFrame column
#
var_ict1110_survey_eda["StudentLocation"] = pd.Series(var_location_list)
```

### Data Attribute Formatting

[17]:
```
var_ict1110_survey_eda.head(3).T
```

[17]:

| | 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 | Languages |
| 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... | |
| year | 2019 | 2019 | 2019 |
| StudentLocation | Lusaka | Copperbelt | Western |

### Data Attribute Scaling

[18]:
```
var_ict1110_survey_eda.head(1).T
```

[18]:

| | 0 |
|---|---|
| Timestamp | 2019/03/28 11:13:51 PM GMT+2 |
| StudentName | Participant1 |
| StudentID | NaN |
| HomeTown | Chudleigh/Lusaka/Lusaka |
| MinorProgramme | Data Mining |
| MinorProgrammeMotivation | love data |
| MajorProgrammeMotivation | love comput |
| DidComputerStudies | No |
| HasComputerTraining | Yes |
| ComputerTrainingType | I have studied Computer Sc... |
| ExperienceWithComputers | More than 5 years |

Continued on next page

| | 0 |
|---|---|
| HasComputerAccess | Yes |
| AboutMe | cycl everyday |
| year | 2019 |
| StudentLocation | Lusaka |

### Dataset #2: Demographic Details

- This should be done as an exercise
  - Identify attributes that require derivations

#### Data Attribute Derivations

- The following attributes can be derived for the dataset
  - StudentAge—AcademicYear - DateOfBirth

[19]:
```
var_ict1110_demographics_eda.head(1).T
```

[19]:

| | 0 |
|---|---|
| StudentID | 9d5116a2451bc98c2b46b93acb... |
| DateOfBirth | 1998-09-14 |
| Gender | F |
| AcademicYear | 20191 |
| YearOfStudy | 2nd Year |
| School | EDUCATION |
| Program | BACHELOR OF INFORMATION AN... |
| MajorDescription | ICTs and Education |
| MinorDescription | RELIGIOUS STUDIES |
| Status | Registered |
| Sponsor | GRZ - 75 PERCENT |
| Nationality | ZAMBIAN |
| Accommodated | Yes |

[20]:
```
#
# Derive the student age
#
var_date_of_birth = var_ict1110_demographics_eda["DateOfBirth"].str[:4].apply(int)
var_year_of_study = var_ict1110_demographics_eda["AcademicYear"].apply(str).str[:4].
 ↪apply(int)
var_student_age = var_year_of_study - var_date_of_birth
var_student_age.head(3)
```

[20]:

| | 0 |
|---|---|
| 0 | 21 |
| 1 | 19 |
| 2 | 21 |

[21]:
```
var_ict1110_demographics_eda["StudentAge"] = var_student_age
```

### Data Attribute Formatting

[22]:
```
var_ict1110_demographics_eda.head(1).T
```

[22]:

| | 0 |
|---|---|
| StudentID | 9d5116a2451bc98c2b46b93acb... |
| DateOfBirth | 1998-09-14 |
| Gender | F |
| AcademicYear | 20191 |
| YearOfStudy | 2nd Year |
| School | EDUCATION |
| Program | BACHELOR OF INFORMATION AN... |
| MajorDescription | ICTs and Education |
| MinorDescription | RELIGIOUS STUDIES |
| Status | Registered |
| Sponsor | GRZ - 75 PERCENT |
| Nationality | ZAMBIAN |
| Accommodated | Yes |
| StudentAge | 21 |

### Data Attribute Scaling

[23]:
```
var_ict1110_demographics_eda.head(1).T
```

[23]:

| | 0 |
|---|---|
| StudentID | 9d5116a2451bc98c2b46b93acb... |
| DateOfBirth | 1998-09-14 |
| Gender | F |
| AcademicYear | 20191 |
| YearOfStudy | 2nd Year |
| School | EDUCATION |
| Program | BACHELOR OF INFORMATION AN... |
| MajorDescription | ICTs and Education |
| MinorDescription | RELIGIOUS STUDIES |
| Status | Registered |
| Sponsor | GRZ - 75 PERCENT |
| Nationality | ZAMBIAN |
| Accommodated | Yes |
| StudentAge | 21 |

## Dataset #3: Assessment Scores

### Data Attribute Scaling

- The assessment scores use different ranges
  - Quizzes: 0–10
  - Tests: 0–50

- Examination: 0–100

The quiz and test scores will need to be scaled to a uniform scale. In this case all absolute scores will be converted to their equivalent percentage scores

[24]:
```
#
def fxn_quiz_scaling(var_quiz_score):
    """
    Function to scale quiz scores
    """
    return round((var_quiz_score/10)*100, 2)

#
def fxn_test_scaling(var_test_score):
    """
    Function to scale test scores
    """
    return round((var_test_score/50)*100, 2)

# Testing the utility functions
var_sample_quiz_score = 6
var_sample_test_score = 29

print("Quiz score of ", var_sample_quiz_score, " is scaled to: ",␣
 ↪fxn_quiz_scaling(var_sample_quiz_score))
print("Test score of ", var_sample_test_score, " is scaled to: ",␣
 ↪fxn_test_scaling(var_sample_test_score))
```

```
Quiz score of  6  is scaled to:  60.0
Test score of  29  is scaled to:  58.0
```

[25]:
```
# Inspect current dataframe structure
var_ict1110_assessments_quizzes.head(1).T
```

[25]:

| | 0 |
|---|---|
| StudentID | 53b3c88ea00c4f0e137b4e6fe7... |
| Quiz1Score | 1 |
| Quiz2Score | 5 |
| Quiz3Score | 3 |
| Quiz4Score | 4 |
| Quiz5Score | 5.5 |
| Quiz6Score | 6 |
| Quiz7Score | 1 |
| Quiz8Score | 0 |
| Quiz9Score | 4 |
| Quiz10Score | 2 |
| Quiz11Score | 4 |
| Quiz12Score | 10 |
| Quiz13Score | 0 |
| Quiz14Score | 4 |

Continued on next page

| | 0 |
|---|---|
| Quiz15Score | 9 |
| Quiz16Score | 0 |
| Quiz17Score | 0 |
| Quiz18Score | 0 |
| Quiz19Score | 9 |
| Quiz20Score | 7 |

[26]:
```
var_ict1110_assessments_quizzes["Quiz1Score"].apply(fxn_quiz_scaling)
```

[26]:

| | Quiz1Score |
|---|---|
| 0 | 10.0 |
| 1 | 20.0 |
| 2 | 20.0 |
| 3 | 10.0 |
| 4 | 10.0 |
| 5 | 30.0 |
| 6 | 10.0 |
| 7 | 50.0 |
| 8 | 0.0 |
| 9 | 0.0 |
| 10 | 0.0 |
| 11 | 70.0 |
| 12 | 60.0 |
| 13 | 60.0 |
| 14 | 50.0 |
| 15 | 40.0 |
| 16 | 10.0 |
| 17 | 40.0 |
| 18 | 20.0 |
| 19 | 70.0 |
| 20 | 60.0 |
| 21 | 60.0 |
| 22 | 40.0 |
| 23 | 10.0 |
| 24 | 50.0 |
| 25 | 50.0 |
| 26 | 20.0 |
| 27 | 80.0 |
| 28 | 70.0 |
| 29 | 30.0 |
| 30 | 50.0 |
| 31 | 40.0 |
| 32 | 60.0 |
| 33 | 20.0 |
| 34 | 60.0 |
| 35 | 60.0 |
| 36 | 20.0 |

Continued on next page

| | Quiz1Score |
|---|---|
| 37 | 10.0 |
| 38 | 60.0 |
| 39 | 50.0 |
| 40 | 50.0 |
| 41 | 20.0 |
| 42 | 10.0 |
| 43 | 40.0 |
| 44 | 60.0 |
| 45 | 40.0 |
| 46 | 40.0 |
| 47 | 50.0 |
| 48 | 60.0 |
| 49 | 50.0 |
| 50 | 20.0 |
| 51 | 70.0 |
| 52 | 70.0 |
| 53 | 30.0 |
| 54 | 50.0 |
| 55 | 10.0 |
| 56 | 100.0 |
| 57 | 40.0 |
| 58 | 40.0 |
| 59 | 0.0 |
| 60 | 0.0 |
| 61 | 0.0 |
| 62 | 0.0 |
| 63 | 0.0 |

[27]:
```
#
# Loop through all quiz columns and create a new column with scaled score
#
for var_quiz in range(1, 21, 1):
    var_quiz_score_column = "Quiz"+str(var_quiz)+"Score" # dynamic column with score
    var_quiz_scaled_column = "Quiz"+str(var_quiz)+"PCT" # dynamic column with scaled␣
 ↪score
    var_ict1110_assessments_quizzes[var_quiz_scaled_column] =␣
 ↪var_ict1110_assessments_quizzes[var_quiz_score_column].apply(fxn_quiz_scaling)
```

[28]:
```
var_ict1110_assessments_quizzes.columns
```

[28]:
```
Index(['StudentID', 'Quiz1Score', 'Quiz2Score', 'Quiz3Score', 'Quiz4Score',
       'Quiz5Score', 'Quiz6Score', 'Quiz7Score', 'Quiz8Score', 'Quiz9Score',
       'Quiz10Score', 'Quiz11Score', 'Quiz12Score', 'Quiz13Score',
       'Quiz14Score', 'Quiz15Score', 'Quiz16Score', 'Quiz17Score',
       'Quiz18Score', 'Quiz19Score', 'Quiz20Score', 'Quiz1PCT', 'Quiz2PCT',
       'Quiz3PCT', 'Quiz4PCT', 'Quiz5PCT', 'Quiz6PCT', 'Quiz7PCT', 'Quiz8PCT',
       'Quiz9PCT', 'Quiz10PCT', 'Quiz11PCT', 'Quiz12PCT', 'Quiz13PCT',
       'Quiz14PCT', 'Quiz15PCT', 'Quiz16PCT', 'Quiz17PCT', 'Quiz18PCT',
       'Quiz19PCT', 'Quiz20PCT'],
```

```
      dtype='object')
```

[29]:
```
# Inspect new dataframe structure
var_ict1110_assessments_quizzes.head(1).T
```

[29]:

| | 0 |
|---|---|
| StudentID | 53b3c88ea00c4f0e137b4e6fe7... |
| Quiz1Score | 1 |
| Quiz2Score | 5 |
| Quiz3Score | 3 |
| Quiz4Score | 4 |
| Quiz5Score | 5.5 |
| Quiz6Score | 6 |
| Quiz7Score | 1 |
| Quiz8Score | 0 |
| Quiz9Score | 4 |
| Quiz10Score | 2 |
| Quiz11Score | 4 |
| Quiz12Score | 10 |
| Quiz13Score | 0 |
| Quiz14Score | 4 |
| Quiz15Score | 9 |
| Quiz16Score | 0 |
| Quiz17Score | 0 |
| Quiz18Score | 0 |
| Quiz19Score | 9 |
| Quiz20Score | 7 |
| Quiz1PCT | 10 |
| Quiz2PCT | 50 |
| Quiz3PCT | 30 |
| Quiz4PCT | 40 |
| Quiz5PCT | 55 |
| Quiz6PCT | 60 |
| Quiz7PCT | 10 |
| Quiz8PCT | 0 |
| Quiz9PCT | 40 |
| Quiz10PCT | 20 |
| Quiz11PCT | 40 |
| Quiz12PCT | 100 |
| Quiz13PCT | 0 |
| Quiz14PCT | 40 |
| Quiz15PCT | 90 |
| Quiz16PCT | 0 |
| Quiz17PCT | 0 |
| Quiz18PCT | 0 |
| Quiz19PCT | 90 |
| Quiz20PCT | 70 |

[30]:
```
# Inspect current dataframe structure
var_ict1110_assessments_tests.head(2).T
```

[30]:

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

[31]:
```
#
# Loop through all test columns and create a new column with scaled score
#
for var_test in range(1, 5, 1):
    var_test_score_column = "Test"+str(var_test)+"Score" # dynamic column with score
    var_test_scaled_column = "Test"+str(var_test)+"PCT" # dynamic column with scaled␣
 ↪score
    var_ict1110_assessments_tests[var_test_scaled_column] =␣
 ↪var_ict1110_assessments_tests[var_test_score_column].apply(fxn_test_scaling)
```

[32]:
```
var_ict1110_assessments_tests.columns
```

[32]:
```
Index(['StudentID', 'Test1Score', 'Test2Score', 'Test3Score', 'Test4Score',
       'Test1PCT', 'Test2PCT', 'Test3PCT', 'Test4PCT'],
      dtype='object')
```

[33]:
```
# Inspect current dataframe structure
var_ict1110_assessments_tests.head(2).T
```

[33]:

| | 0 | 1 |
|---|---|---|
| StudentID | 07f3ca235faaa1c9ad16facef5... | 921855f753932de762b780405a... |
| Test1Score | 31.5 | 40 |
| Test2Score | 29.5 | 16 |
| Test3Score | 34 | 24 |
| Test4Score | 29.5 | 22 |
| Test1PCT | 63 | 80 |
| Test2PCT | 59 | 32 |
| Test3PCT | 68 | 48 |
| Test4PCT | 59 | 44 |

[34]:
```
# Inspect current dataframe structure
var_ict1110_assessments_examination.head(2).T
```

[34]:

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

[35]:
```
#
#Create a new column with scaled examination score
#
var_ict1110_assessments_examination["ExaminationPCT"] =␣
 ↪var_ict1110_assessments_examination["ExaminationScore"]
```

[36]:
```
# Inspect current dataframe structure
var_ict1110_assessments_examination.head(2).T
```

[36]:

| | 0 | 1 |
|---|---|---|
| StudentID | 9d5116a2451bc98c2b46b93acb... | e7400496f1ce70cb62c2c44ca2... |
| ExaminationScore | 46.5 | 48.5 |
| ExaminationPCT | 46.5 | 48.5 |

[37]:
```
%%html
<style>
  table {margin-left: 0 !important;}
</style>
```

```
<IPython.core.display.HTML object>
```

**Data Attribute Derivations**

- The assessment scores might need to be used to predict if students pass or fail and/or if students get a specified grade: D–A+
  - Pass mark for undergraduate courses is 45%
  - Thresholds for grading and new GPA scores are below

| Grade | Description | Grade Point | Score Range |
|---|---|---|---|
| A+ | Distinction | 5.0 | 90–100 |
| A | Distinction | 4.0 | 80–89 |
| B+ | Meritorious | 3.5 | 70–79 |
| B | Credit | 3.0 | 60–69 |
| C+ | Credit | 2.37 | 50–59 |
| C | Pass | 1.5 | 45–49 |
| D+ | Fail | 0.0 | 40–44 |
| D | Fail | 0.0 | < 40 |

The quiz, test and examination scores will need to have derivations associated with pass/fail status, grade and gpa score

**Utility Functions for Deriving Attributes**

[38]:
```
#
#
def fxn_pass_status(var_score_pct):
    """
    Function to return a pass/fail status given a percentage score
```

```
    """
    if (var_score_pct < 45):
        return "Fail"
    else:
        return "Pass"
#
#
def fxn_grade_gpa_classification(var_score_pct):
    """
    Function to return a grade classification, given a percentage score
    """
    if (var_score_pct >= 90):
        return ("A+", 5.0)
    elif (var_score_pct >= 80):
        return ("A", 4.0)
    elif (var_score_pct >= 70):
        return ("B+", 3.5)
    elif (var_score_pct >= 60):
        return ("B", 3.0)
    elif (var_score_pct >= 50):
        return ("C+", 2.37)
    elif (var_score_pct >= 45):
        return ("C", 1.5)
    elif (var_score_pct >= 40):
        return ("D+", 0.0)
    else:
        return ("D", 0.0)

# Testing the utility functions
var_sample_score = 59.5

var_sample_score_status = fxn_pass_status(var_sample_score)
var_sample_score_grade = fxn_grade_gpa_classification(var_sample_score)[0]
var_sample_score_gpa = fxn_grade_gpa_classification(var_sample_score)[1]

print("Score of ", var_sample_score, " is classified as follows : ", {
    "Pass Status": var_sample_score_status,
    "Grade Class": var_sample_score_grade,
    "GPA Score": var_sample_score_gpa,
})
```

```
Score of  59.5  is classified as follows :  {'Pass Status': 'Pass', 'Grade Class': 'C+',
'GPA Score': 2.37}
```

**Quiz Scores**

[39]:
```
# Inspect current dataframe structure
var_ict1110_assessments_quizzes.head(1).T
```

[39]:

| | 0 |
|---|---|
| StudentID | 53b3c88ea00c4f0e137b4e6fe7... |
| Quiz1Score | 1 |
| Quiz2Score | 5 |
| Quiz3Score | 3 |
| Quiz4Score | 4 |
| Quiz5Score | 5.5 |
| Quiz6Score | 6 |
| Quiz7Score | 1 |
| Quiz8Score | 0 |
| Quiz9Score | 4 |
| Quiz10Score | 2 |
| Quiz11Score | 4 |
| Quiz12Score | 10 |
| Quiz13Score | 0 |
| Quiz14Score | 4 |
| Quiz15Score | 9 |
| Quiz16Score | 0 |
| Quiz17Score | 0 |
| Quiz18Score | 0 |
| Quiz19Score | 9 |
| Quiz20Score | 7 |
| Quiz1PCT | 10 |
| Quiz2PCT | 50 |
| Quiz3PCT | 30 |
| Quiz4PCT | 40 |
| Quiz5PCT | 55 |
| Quiz6PCT | 60 |
| Quiz7PCT | 10 |
| Quiz8PCT | 0 |
| Quiz9PCT | 40 |
| Quiz10PCT | 20 |
| Quiz11PCT | 40 |
| Quiz12PCT | 100 |
| Quiz13PCT | 0 |
| Quiz14PCT | 40 |
| Quiz15PCT | 90 |
| Quiz16PCT | 0 |
| Quiz17PCT | 0 |
| Quiz18PCT | 0 |
| Quiz19PCT | 90 |
| Quiz20PCT | 70 |

[40]:
```
#
# Loop through all quiz columns and create a new column with pass status, grade␣
 ↪classification and GPA classification
#
for var_quiz in range(1, 21, 1):
    var_quiz_pct_column = "Quiz"+str(var_quiz)+"PCT" # dynamic column with score␣
 ↪percentage
```

```
    var_quiz_status_column = "Quiz"+str(var_quiz)+"Status" # dynamic column score␣
↪status
    var_quiz_grade_column = "Quiz"+str(var_quiz)+"Grade" # dynamic column score grade
    var_quiz_gpa_column = "Quiz"+str(var_quiz)+"GPA" # dynamic column score grade
    #
    # Assign values to derived columns
    var_ict1110_assessments_quizzes[var_quiz_status_column] =␣
↪var_ict1110_assessments_quizzes[var_quiz_pct_column].apply(fxn_pass_status)
    var_ict1110_assessments_quizzes[var_quiz_grade_column] =␣
↪var_ict1110_assessments_quizzes[var_quiz_pct_column].
↪apply(fxn_grade_gpa_classification).str[0]
    var_ict1110_assessments_quizzes[var_quiz_gpa_column] =␣
↪var_ict1110_assessments_quizzes[var_quiz_pct_column].
↪apply(fxn_grade_gpa_classification).str[1].astype(float)
```

[41]:
```
var_ict1110_assessments_quizzes.columns
```

[41]:
```
Index(['StudentID', 'Quiz1Score', 'Quiz2Score', 'Quiz3Score', 'Quiz4Score',
       'Quiz5Score', 'Quiz6Score', 'Quiz7Score', 'Quiz8Score', 'Quiz9Score',
       ...
       'Quiz17GPA', 'Quiz18Status', 'Quiz18Grade', 'Quiz18GPA', 'Quiz19Status',
       'Quiz19Grade', 'Quiz19GPA', 'Quiz20Status', 'Quiz20Grade', 'Quiz20GPA'],
      dtype='object', length=101)
```

[42]:
```
# Inspect new dataframe structure
var_ict1110_assessments_quizzes.head(1).T
```

[42]:

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

Continued on next page

| | 0 |
|---|---|
| Quiz1PCT | 10 |
| Quiz2PCT | 50 |
| Quiz3PCT | 30 |
| Quiz4PCT | 40 |
| Quiz5PCT | 55 |
| Quiz6PCT | 60 |
| Quiz7PCT | 10 |
| Quiz8PCT | 0 |
| Quiz9PCT | 40 |
| Quiz10PCT | 20 |
| Quiz11PCT | 40 |
| Quiz12PCT | 100 |
| Quiz13PCT | 0 |
| Quiz14PCT | 40 |
| Quiz15PCT | 90 |
| Quiz16PCT | 0 |
| Quiz17PCT | 0 |
| Quiz18PCT | 0 |
| Quiz19PCT | 90 |
| Quiz20PCT | 70 |
| Quiz1Status | Fail |
| Quiz1Grade | D |
| Quiz1GPA | 0 |
| Quiz2Status | Pass |
| Quiz2Grade | C+ |
| Quiz2GPA | 2.37 |
| Quiz3Status | Fail |
| Quiz3Grade | D |
| Quiz3GPA | 0 |
| Quiz4Status | Fail |
| Quiz4Grade | D+ |
| Quiz4GPA | 0 |
| Quiz5Status | Pass |
| Quiz5Grade | C+ |
| Quiz5GPA | 2.37 |
| Quiz6Status | Pass |
| Quiz6Grade | B |
| Quiz6GPA | 3 |
| Quiz7Status | Fail |
| Quiz7Grade | D |
| Quiz7GPA | 0 |
| Quiz8Status | Fail |
| Quiz8Grade | D |
| Quiz8GPA | 0 |
| Quiz9Status | Fail |
| Quiz9Grade | D+ |
| Quiz9GPA | 0 |
| Quiz10Status | Fail |

Continued on next page

| | 0 |
|---|---|
| Quiz10Grade | D |
| Quiz10GPA | 0 |
| Quiz11Status | Fail |
| Quiz11Grade | D+ |
| Quiz11GPA | 0 |
| Quiz12Status | Pass |
| Quiz12Grade | A+ |
| Quiz12GPA | 5 |
| Quiz13Status | Fail |
| Quiz13Grade | D |
| Quiz13GPA | 0 |
| Quiz14Status | Fail |
| Quiz14Grade | D+ |
| Quiz14GPA | 0 |
| Quiz15Status | Pass |
| Quiz15Grade | A+ |
| Quiz15GPA | 5 |
| Quiz16Status | Fail |
| Quiz16Grade | D |
| Quiz16GPA | 0 |
| Quiz17Status | Fail |
| Quiz17Grade | D |
| Quiz17GPA | 0 |
| Quiz18Status | Fail |
| Quiz18Grade | D |
| Quiz18GPA | 0 |
| Quiz19Status | Pass |
| Quiz19Grade | A+ |
| Quiz19GPA | 5 |
| Quiz20Status | Pass |
| Quiz20Grade | B+ |
| Quiz20GPA | 3.5 |

**Test Scores**

[43]:
```
# Inspect current tests dataframe structure
var_ict1110_assessments_tests.head(1).T
```

[43]:

| | 0 |
|---|---|
| StudentID | 07f3ca235faaa1c9ad16facef5... |
| Test1Score | 31.5 |
| Test2Score | 29.5 |
| Test3Score | 34 |
| Test4Score | 29.5 |
| Test1PCT | 63 |
| Test2PCT | 59 |
| Test3PCT | 68 |
| Test4PCT | 59 |

[44]:
```
#
# Loop through all test columns and create a new column with pass status, grade␣
 ↪classification and GPA classification
#
for var_test in range(1, 5, 1):
    var_test_pct_column = "Test"+str(var_test)+"PCT" # dynamic column with score␣
 ↪percentage
    var_test_status_column = "Test"+str(var_test)+"Status" # dynamic column score␣
 ↪status
    var_test_grade_column = "Test"+str(var_test)+"Grade" # dynamic column score grade
    var_test_gpa_column = "Test"+str(var_test)+"GPA" # dynamic column score grade
    #
    # Assign values to derived columns
    var_ict1110_assessments_tests[var_test_status_column] =␣
 ↪var_ict1110_assessments_tests[var_test_pct_column].apply(fxn_pass_status)
    var_ict1110_assessments_tests[var_test_grade_column] =␣
 ↪var_ict1110_assessments_tests[var_test_pct_column].
 ↪apply(fxn_grade_gpa_classification).str[0]
    var_ict1110_assessments_tests[var_test_gpa_column] =␣
 ↪var_ict1110_assessments_tests[var_test_pct_column].
 ↪apply(fxn_grade_gpa_classification).str[1].astype(float)
```

[45]:
```
var_ict1110_assessments_tests.columns
```

[45]:
```
Index(['StudentID', 'Test1Score', 'Test2Score', 'Test3Score', 'Test4Score',
       'Test1PCT', 'Test2PCT', 'Test3PCT', 'Test4PCT', 'Test1Status',
       'Test1Grade', 'Test1GPA', 'Test2Status', 'Test2Grade', 'Test2GPA',
       'Test3Status', 'Test3Grade', 'Test3GPA', 'Test4Status', 'Test4Grade',
       'Test4GPA'],
      dtype='object')
```

[46]:
```
# Inspect new tests dataframe structure
var_ict1110_assessments_tests.tail(2).T
```

[46]:

| | 57 | 58 |
|---|---|---|
| StudentID | 4234d1794dd33c1b6ed975eab5... | 232bf11cb81bcdb269f76a08fd... |
| Test1Score | 31 | 22.5 |
| Test2Score | 19.5 | 17 |
| Test3Score | 28 | 17 |
| Test4Score | 22 | 20.5 |
| Test1PCT | 62 | 45 |
| Test2PCT | 39 | 34 |
| Test3PCT | 56 | 34 |
| Test4PCT | 44 | 41 |
| Test1Status | Pass | Pass |
| Test1Grade | B | C |
| Test1GPA | 3 | 1.5 |
| Test2Status | Fail | Fail |
| Test2Grade | D | D |

Continued on next page

| | 57 | 58 |
|---|---|---|
| Test2GPA | 0 | 0 |
| Test3Status | Pass | Fail |
| Test3Grade | C+ | D |
| Test3GPA | 2.37 | 0 |
| Test4Status | Fail | Fail |
| Test4Grade | D+ | D+ |
| Test4GPA | 0 | 0 |

**Examination Scores**

[47]:
```
# Inspect current examination dataframe structure
var_ict1110_assessments_examination.head(1).T
```

[47]:

| | 0 |
|---|---|
| StudentID | 9d5116a2451bc98c2b46b93acb... |
| ExaminationScore | 46.5 |
| ExaminationPCT | 46.5 |

[48]:
```
#
var_ict1110_assessments_examination["ExaminationStatus"] =␣
 ↪var_ict1110_assessments_examination["ExaminationPCT"].apply(fxn_pass_status)
#
#
var_ict1110_assessments_examination["ExaminationGrade"] =␣
 ↪var_ict1110_assessments_examination["ExaminationPCT"].
 ↪apply(fxn_grade_gpa_classification).str[0]
#
#
var_ict1110_assessments_examination["ExaminationGPA"] =␣
 ↪var_ict1110_assessments_examination["ExaminationPCT"].
 ↪apply(fxn_grade_gpa_classification).str[1].astype(float)
```

[49]:
```
var_ict1110_assessments_examination.columns
```

[49]:
```
Index(['StudentID', 'ExaminationScore', 'ExaminationPCT', 'ExaminationStatus',
       'ExaminationGrade', 'ExaminationGPA'],
      dtype='object')
```

[50]:
```
# Inspect new examination dataframe structure
var_ict1110_assessments_examination.head(2).T
```

[50]:

| | 0 | 1 |
|---|---|---|
| StudentID | 9d5116a2451bc98c2b46b93acb... | e7400496f1ce70cb62c2c44ca2... |
| ExaminationScore | 46.5 | 48.5 |
| ExaminationPCT | 46.5 | 48.5 |

Continued on next page

| | 0 | 1 |
|---|---|---|
| ExaminationStatus | Pass | Pass |
| ExaminationGrade | C | C |
| ExaminationGPA | 1.5 | 1.5 |

### Data Attribute Formatting

- No attributes to format

## Merging All Datasets

- All datasets to be combined into one contiguous dataset
  - Dataset #1: Initial Survey—`var_ict1110_survey_eda`
  - Dataset #2: Demographic Details—`var_ict1110_demographics_eda`
  - Dataset #3: Assessment Scores
    - Quizzes—`var_ict1110_assessments_quizzes`
    - Tests—`var_ict1110_assessments_tests`
    - Examination—`var_ict1110_assessments_examination`

[51]:
```
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_student_performance_dataframes =␣
↪[var_ict1110_survey_eda,var_ict1110_demographics_eda,var_ict1110_assessments_quizzes,var_ict111

# Merge all the quizzes into one dataframe
var_ict1110_student_performance = reduce(lambda  left,right: pd.
↪merge(left,right,on=['StudentID'],
                                            how='outer'),␣
↪var_ict1110_student_performance_dataframes)
```

[52]:
```
var_ict1110_student_performance.columns
```

[52]:
```
Index(['Timestamp', 'StudentName', 'StudentID', 'HomeTown', 'MinorProgramme',
       'MinorProgrammeMotivation', 'MajorProgrammeMotivation',
       'DidComputerStudies', 'HasComputerTraining', 'ComputerTrainingType',
       ...
       'Test3Grade', 'Test3GPA', 'Test4Status', 'Test4Grade', 'Test4GPA',
       'ExaminationScore', 'ExaminationPCT', 'ExaminationStatus',
       'ExaminationGrade', 'ExaminationGPA'],
      dtype='object', length=153)
```

[53]:
```
var_ict1110_student_performance[var_ict1110_student_performance["ExaminationScore"].
↪notna()].head(2).T
```

[53]:

| | 1 | 2 |
|---|---|---|
| Timestamp | 2019/03/28 11:55:27 PM GMT+2 | 2019/03/29 8:00:53 PM GMT+2 |

Continued on next page

| | 1 | 2 |
|---|---|---|
| StudentName | Participant2 | Participant3 |
| StudentID | 742b8abe5776a6d942a92ce7dc... | 921855f753932de762b780405a... |
| HomeTown | Copperbelt,luanshya,Mpatamato | Mungule,senanga,western. |
| MinorProgramme | Mathematics | Languages |
| MinorProgrammeMotivation | find easi studi understand | best avail option |
| MajorProgrammeMotivation | want acquir knowledg ict c... | alway want ict relat program |
| DidComputerStudies | No | No |
| HasComputerTraining | No | No |
| ComputerTrainingType | NaN | NaN |
| ExperienceWithComputers | 1 to 2 years | No Experience |
| HasComputerAccess | Yes | Yes |
| AboutMe | day pass without joke feel... | |
| year | 2019 | 2019 |
| StudentLocation | Copperbelt | Western |
| DateOfBirth | 1999-03-23 | 1999-12-24 |
| Gender | F | M |
| AcademicYear | 20191 | 20181 |
| YearOfStudy | 2nd Year | 1st Year |
| School | EDUCATION | EDUCATION |
| Program | BACHELOR OF INFORMATION AN... | BACHELOR OF INFORMATION AN... |
| MajorDescription | ICTs and Education | ICTs and Education |
| MinorDescription | MATHEMATICS | FRENCH |
| Status | Registered | Registered |
| Sponsor | GRZ-FULLY SPONSORED | GRZ-FULLY SPONSORED |
| Nationality | ZAMBIAN | ZAMBIAN |
| Accommodated | No | Yes |
| StudentAge | 20 | 19 |
| Quiz1Score | 5 | 5 |
| Quiz2Score | 4 | 4 |
| Quiz3Score | 2 | 6 |
| Quiz4Score | 3 | 4 |
| Quiz5Score | 6 | 9 |
| Quiz6Score | 9 | 10 |
| Quiz7Score | 0 | 8 |
| Quiz8Score | 5 | 5 |
| Quiz9Score | 7 | 9.5 |
| Quiz10Score | 2 | 7 |
| Quiz11Score | 2 | 6 |
| Quiz12Score | 10 | 10 |
| Quiz13Score | 0 | 0.5 |
| Quiz14Score | 6 | 5 |
| Quiz15Score | 7 | 8 |
| Quiz16Score | 1 | 2 |
| Quiz17Score | 10 | 0 |
| Quiz18Score | 2 | 9 |
| Quiz19Score | 8 | 9 |
| Quiz20Score | 9 | 10 |
| Quiz1PCT | 50 | 50 |

Continued on next page

| | 1 | 2 |
|---|---|---|
| Quiz2PCT | 40 | 40 |
| Quiz3PCT | 20 | 60 |
| Quiz4PCT | 30 | 40 |
| Quiz5PCT | 60 | 90 |
| Quiz6PCT | 90 | 100 |
| Quiz7PCT | 0 | 80 |
| Quiz8PCT | 50 | 50 |
| Quiz9PCT | 70 | 95 |
| Quiz10PCT | 20 | 70 |
| Quiz11PCT | 20 | 60 |
| Quiz12PCT | 100 | 100 |
| Quiz13PCT | 0 | 5 |
| Quiz14PCT | 60 | 50 |
| Quiz15PCT | 70 | 80 |
| Quiz16PCT | 10 | 20 |
| Quiz17PCT | 100 | 0 |
| Quiz18PCT | 20 | 90 |
| Quiz19PCT | 80 | 90 |
| Quiz20PCT | 90 | 100 |
| Quiz1Status | Pass | Pass |
| Quiz1Grade | C+ | C+ |
| Quiz1GPA | 2.37 | 2.37 |
| Quiz2Status | Fail | Fail |
| Quiz2Grade | D+ | D+ |
| Quiz2GPA | 0 | 0 |
| Quiz3Status | Fail | Pass |
| Quiz3Grade | D | B |
| Quiz3GPA | 0 | 3 |
| Quiz4Status | Fail | Fail |
| Quiz4Grade | D | D+ |
| Quiz4GPA | 0 | 0 |
| Quiz5Status | Pass | Pass |
| Quiz5Grade | B | A+ |
| Quiz5GPA | 3 | 5 |
| Quiz6Status | Pass | Pass |
| Quiz6Grade | A+ | A+ |
| Quiz6GPA | 5 | 5 |
| Quiz7Status | Fail | Pass |
| Quiz7Grade | D | A |
| Quiz7GPA | 0 | 4 |
| Quiz8Status | Pass | Pass |
| Quiz8Grade | C+ | C+ |
| Quiz8GPA | 2.37 | 2.37 |
| Quiz9Status | Pass | Pass |
| Quiz9Grade | B+ | A+ |
| Quiz9GPA | 3.5 | 5 |
| Quiz10Status | Fail | Pass |
| Quiz10Grade | D | B+ |

Continued on next page

| | 1 | 2 |
|---|---|---|
| Quiz10GPA | 0 | 3.5 |
| Quiz11Status | Fail | Pass |
| Quiz11Grade | D | B |
| Quiz11GPA | 0 | 3 |
| Quiz12Status | Pass | Pass |
| Quiz12Grade | A+ | A+ |
| Quiz12GPA | 5 | 5 |
| Quiz13Status | Fail | Fail |
| Quiz13Grade | D | D |
| Quiz13GPA | 0 | 0 |
| Quiz14Status | Pass | Pass |
| Quiz14Grade | B | C+ |
| Quiz14GPA | 3 | 2.37 |
| Quiz15Status | Pass | Pass |
| Quiz15Grade | B+ | A |
| Quiz15GPA | 3.5 | 4 |
| Quiz16Status | Fail | Fail |
| Quiz16Grade | D | D |
| Quiz16GPA | 0 | 0 |
| Quiz17Status | Pass | Fail |
| Quiz17Grade | A+ | D |
| Quiz17GPA | 5 | 0 |
| Quiz18Status | Fail | Pass |
| Quiz18Grade | D | A+ |
| Quiz18GPA | 0 | 5 |
| Quiz19Status | Pass | Pass |
| Quiz19Grade | A | A+ |
| Quiz19GPA | 4 | 5 |
| Quiz20Status | Pass | Pass |
| Quiz20Grade | A+ | A+ |
| Quiz20GPA | 5 | 5 |
| Test1Score | 24.5 | 40 |
| Test2Score | 15.5 | 16 |
| Test3Score | 21 | 24 |
| Test4Score | 17.5 | 22 |
| Test1PCT | 49 | 80 |
| Test2PCT | 31 | 32 |
| Test3PCT | 42 | 48 |
| Test4PCT | 35 | 44 |
| Test1Status | Pass | Pass |
| Test1Grade | C | A |
| Test1GPA | 1.5 | 4 |
| Test2Status | Fail | Fail |
| Test2Grade | D | D |
| Test2GPA | 0 | 0 |
| Test3Status | Fail | Pass |
| Test3Grade | D+ | C |
| Test3GPA | 0 | 1.5 |

Continued on next page

| | 1 | 2 |
|---|---|---|
| Test4Status | Fail | Fail |
| Test4Grade | D | D+ |
| Test4GPA | 0 | 0 |
| ExaminationScore | 56 | 85 |
| ExaminationPCT | 56 | 85 |
| ExaminationStatus | Pass | Pass |
| ExaminationGrade | C+ | A |
| ExaminationGPA | 2.37 | 4 |

[54]:
```
# Check how many observations have survey details
len(var_ict1110_student_performance)
len(var_ict1110_student_performance[var_ict1110_student_performance["ExaminationScore"].
 ↪notna()]["Timestamp"].notna())
```

[54]:
```
123
```

[54]:
```
63
```

### Merging All Datasets

Ensure that all NULL values have a marker

[55]:
```
#
# Replace ALL NULL values
#
# MinorProgramme; MinorProgrammeMotivation; MajorProgrammeMotivation
#
var_ict1110_student_performance.columns

var_ict1110_student_performance[["StudentLocation", "MinorProgramme",␣
 ↪"MinorProgrammeMotivation", "MajorProgrammeMotivation", "DidComputerStudies",␣
 ↪"HasComputerTraining", "ComputerTrainingType", "ExperienceWithComputers",␣
 ↪"HasComputerAccess", "AboutMe"]] =␣
 ↪var_ict1110_student_performance[["StudentLocation", "MinorProgramme",␣
 ↪"MinorProgrammeMotivation", "MajorProgrammeMotivation", "DidComputerStudies",␣
 ↪"HasComputerTraining", "ComputerTrainingType", "ExperienceWithComputers",␣
 ↪"HasComputerAccess", "AboutMe"]].fillna("MISSING DATA")
```

[55]:
```
Index(['Timestamp', 'StudentName', 'StudentID', 'HomeTown', 'MinorProgramme',
       'MinorProgrammeMotivation', 'MajorProgrammeMotivation',
       'DidComputerStudies', 'HasComputerTraining', 'ComputerTrainingType',
       ...
       'Test3Grade', 'Test3GPA', 'Test4Status', 'Test4Grade', 'Test4GPA',
       'ExaminationScore', 'ExaminationPCT', 'ExaminationStatus',
       'ExaminationGrade', 'ExaminationGPA'],
      dtype='object', length=153)
```

# Exploratory Data Analysis

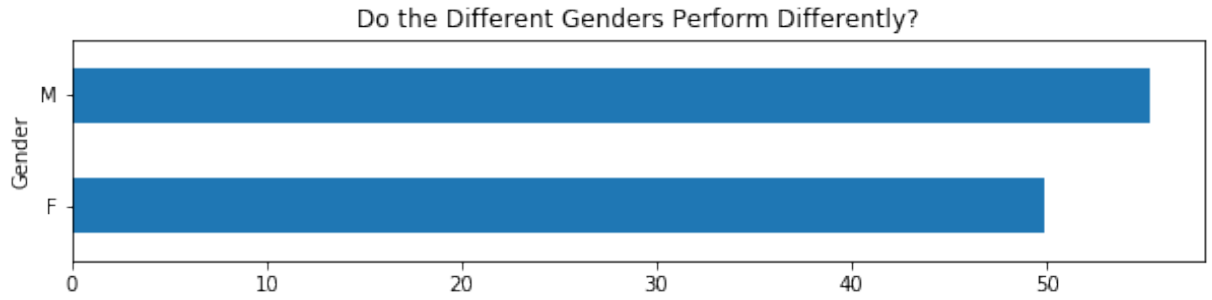
## Demographic Differences

### How are Exam Scores for the Different Genders

[56]:
```
fig, (ax1) = plt.subplots(1, 1, figsize=(10,2))

#
var_ict1110_student_performance.groupby(["Gender"])["ExaminationPCT"].mean().
 ↪plot(kind="barh", title="Do the Different Genders Perform Differently?", ax=ax1)
```

[56]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4cec56ef98>
```

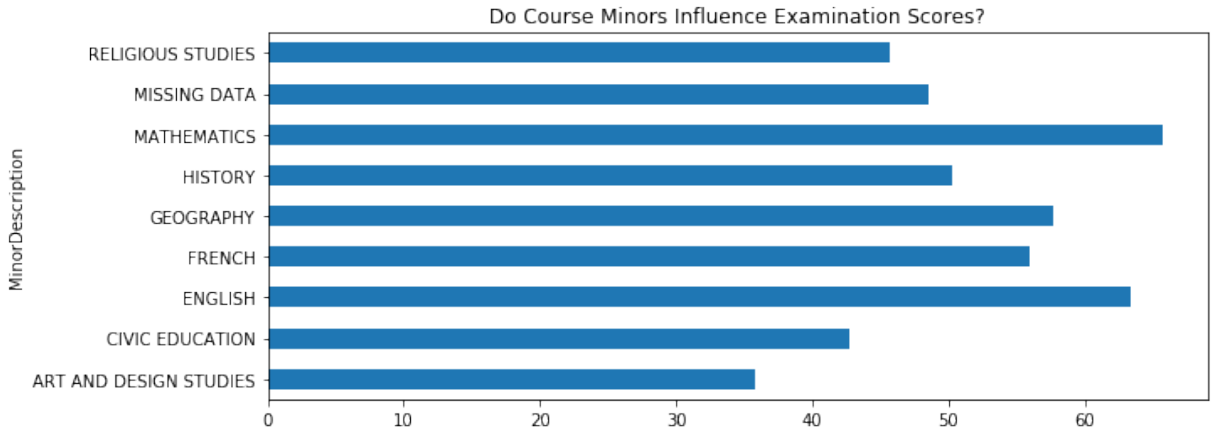



### Effect of Programme Minors

[57]:
```
fig, (ax1) = plt.subplots(1, 1, figsize=(10,4))

#
var_ict1110_student_performance.groupby(["MinorDescription"])["ExaminationPCT"].mean().
 ↪plot(kind="barh", title="Do Course Minors Influence Examination Scores?", ax=ax1)
```

[57]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce44a5400>
```



### Effect of Computing Experience on Examination Score

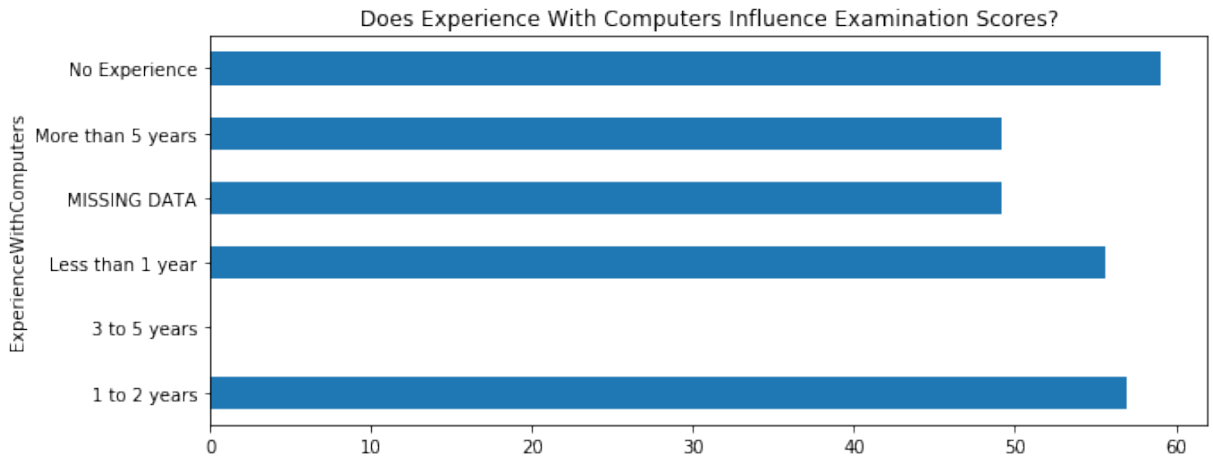
[58]:
```
var_ict1110_student_performance[var_ict1110_student_performance["ExaminationScore"].
 ↪notna()]["ExperienceWithComputers"].unique()
```

[58]:
```
array(['1 to 2 years', 'No Experience', 'Less than 1 year',
       'More than 5 years', 'MISSING DATA'], dtype=object)
```

[59]:
```
fig, (ax1) = plt.subplots(1, 1, figsize=(10,4))

#
var_ict1110_student_performance.groupby(["ExperienceWithComputers"])["ExaminationPCT"].
 ↪mean().plot(kind="barh", title="Does Experience With Computers Influence Examination␣
 ↪Scores?", ax=ax1)
```

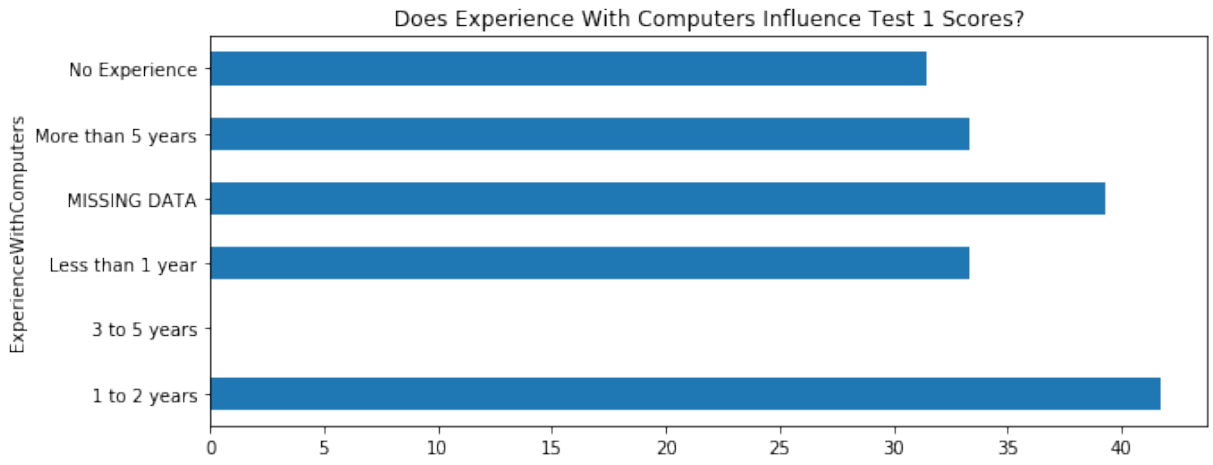
[59]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce43f35f8>
```




[60]:
```
fig, (ax1) = plt.subplots(1, 1, figsize=(10,4))

var_ict1110_student_performance.groupby(["ExperienceWithComputers"])["Quiz1PCT"].
 ↪mean().plot(kind="barh", title="Does Experience With Computers Influence Test 1␣
 ↪Scores?", ax=ax1)
```

[60]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce43b76a0>
```

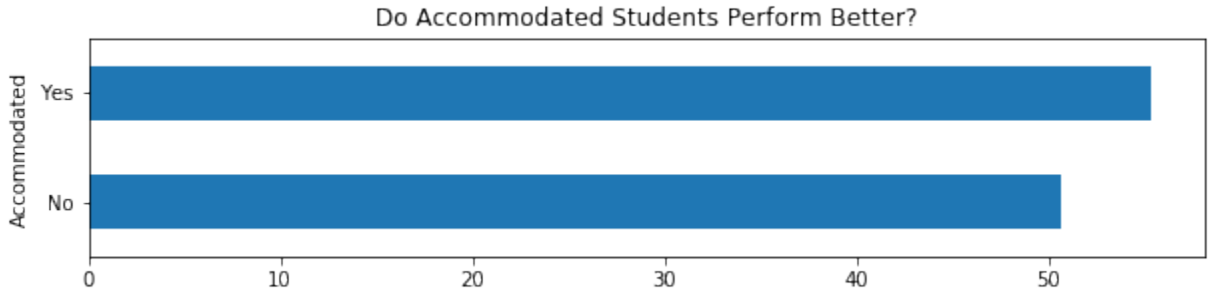



### Do Accommodated Students Perform Better?

[61]:
```
fig, (ax1) = plt.subplots(1, 1, figsize=(10,2))

#
var_ict1110_student_performance.groupby(["Accommodated"])["ExaminationPCT"].mean().
 ↪plot(kind="barh", title="Do Accommodated Students Perform Better?", ax=ax1)
```

[61]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce42bf0b8>
```




## Variable Correlations

### Test Scores vs Examination Correlations

[62]:
```
var_student_performance_eda_ =␣
 ↪var_ict1110_student_performance[var_ict1110_student_performance["ExaminationScore"].
 ↪notna()]

# Facet results by academic year

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

fig.suptitle('Test Scores and Examination Score Correlation')
```

```
sns.scatterplot(x='ExaminationPCT', y='Test1PCT', data=var_student_performance_eda_,␣
↪ax=ax1)
sns.scatterplot(x='ExaminationPCT', y='Test2PCT', data=var_student_performance_eda_,␣
↪ax=ax2)
sns.scatterplot(x='ExaminationPCT', y='Test3PCT', data=var_student_performance_eda_,␣
↪ax=ax3)
sns.scatterplot(x='ExaminationPCT', y='Test4PCT', data=var_student_performance_eda_,␣
↪ax=ax4)
```

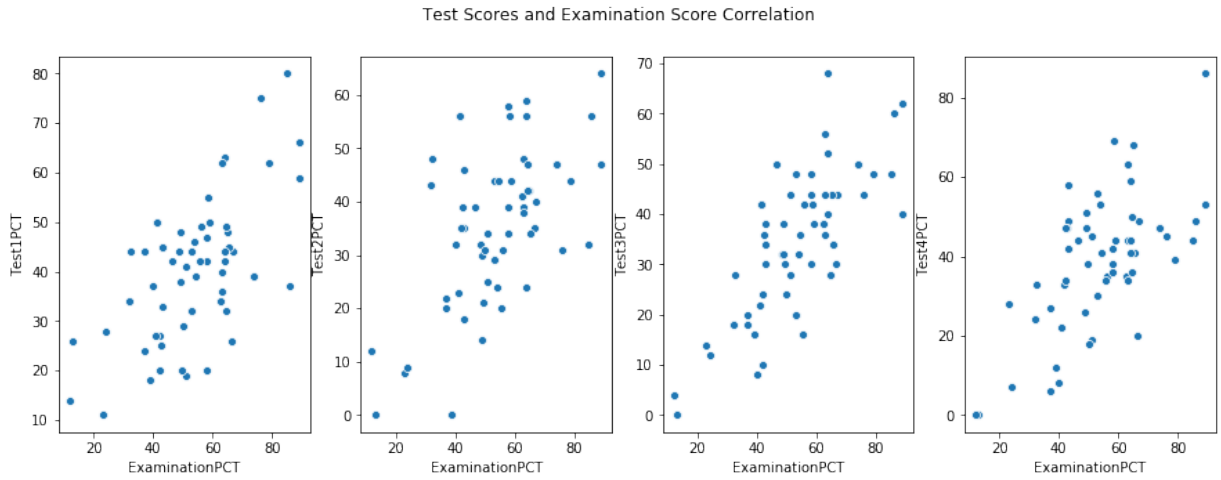
[62]: `Text(0.5, 0.98, 'Test Scores and Examination Score Correlation')`

[62]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f4cec599128>`

[62]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce424c5c0>`

[62]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce4276a20>`

[62]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce421cf98>`




**Quiz Scores vs Examination Correlations**

- Quizzes are mapped on to topics, so perhaps performance is certain topics influences examination scores

[63]:
```
# 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 vs Examination Correlation')

sns.scatterplot(x='ExaminationPCT', y='Quiz1PCT', data=var_student_performance_eda_,␣
↪ax=ax1)
```

```
sns.scatterplot(x='ExaminationPCT', y='Quiz2PCT', data=var_student_performance_eda_,␣
 ↪ax=ax2)
sns.scatterplot(x='ExaminationPCT', y='Quiz3PCT', data=var_student_performance_eda_,␣
 ↪ax=ax3)
sns.scatterplot(x='ExaminationPCT', y='Quiz4PCT', data=var_student_performance_eda_,␣
 ↪ax=ax4)
sns.scatterplot(x='ExaminationPCT', y='Quiz5PCT', data=var_student_performance_eda_,␣
 ↪ax=ax5)
sns.scatterplot(x='ExaminationPCT', y='Quiz6PCT', data=var_student_performance_eda_,␣
 ↪ax=ax6)
sns.scatterplot(x='ExaminationPCT', y='Quiz7PCT', data=var_student_performance_eda_,␣
 ↪ax=ax7)
sns.scatterplot(x='ExaminationPCT', y='Quiz8PCT', data=var_student_performance_eda_,␣
 ↪ax=ax8)
sns.scatterplot(x='ExaminationPCT', y='Quiz9PCT', data=var_student_performance_eda_,␣
 ↪ax=ax9)
sns.scatterplot(x='ExaminationPCT', y='Quiz10PCT', data=var_student_performance_eda_,␣
 ↪ax=ax10)
sns.scatterplot(x='ExaminationPCT', y='Quiz11PCT', data=var_student_performance_eda_,␣
 ↪ax=ax11)
sns.scatterplot(x='ExaminationPCT', y='Quiz12PCT', data=var_student_performance_eda_,␣
 ↪ax=ax12)
sns.scatterplot(x='ExaminationPCT', y='Quiz13PCT', data=var_student_performance_eda_,␣
 ↪ax=ax13)
sns.scatterplot(x='ExaminationPCT', y='Quiz14PCT', data=var_student_performance_eda_,␣
 ↪ax=ax14)
sns.scatterplot(x='ExaminationPCT', y='Quiz15PCT', data=var_student_performance_eda_,␣
 ↪ax=ax15)
sns.scatterplot(x='ExaminationPCT', y='Quiz16PCT', data=var_student_performance_eda_,␣
 ↪ax=ax16)
sns.scatterplot(x='ExaminationPCT', y='Quiz17PCT', data=var_student_performance_eda_,␣
 ↪ax=ax17)
sns.scatterplot(x='ExaminationPCT', y='Quiz18PCT', data=var_student_performance_eda_,␣
 ↪ax=ax18)
sns.scatterplot(x='ExaminationPCT', y='Quiz19PCT', data=var_student_performance_eda_,␣
 ↪ax=ax19)
sns.scatterplot(x='ExaminationPCT', y='Quiz20PCT', data=var_student_performance_eda_,␣
 ↪ax=ax20)
```

[63]: `Text(0.5, 0.98, 'Quiz Scores vs Examination Correlation')`

[63]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3fa3240>`

[63]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce40ff5c0>`

[63]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce4125b38>`

[63]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce40d70b8>`

[63]: `<matplotlib.axes._subplots.AxesSubplot at 0x7f4ce407d630>`

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce40a5ba8>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce4054160>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3ffc6a0>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce4023c50>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3f59208>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3f7e780>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3f27cf8>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3ed72b0>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3efe828>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3ea6da0>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3e56358>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3e7a8d0>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3e24e48>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3dd4400>

[63]: <matplotlib.axes._subplots.AxesSubplot at 0x7f4ce3dfc978>


Quiz Scores vs Examination Correlation
Quiz1PCT
Quiz2PCT
Quiz3PCT
Quiz4PCT
Quiz5PCT
Quiz6PCT
Quiz7PCT
Quiz8PCT
Quiz9PCT
Quiz10PCT
Quiz11PCT
Quiz12PCT
Quiz13PCT
Quiz14PCT
Quiz15PCT
Quiz16PCT
Quiz17PCT
Quiz18PCT
Quiz19PCT
Quiz20PCT
ExaminationPCT

## Save Pipeline

[64]:
```
#
# Save the merged DataFrame
joblib.dump(var_ict1110_student_performance, "var_ict1110_student_performance.pkl")
```

[64]:
```
['var_ict1110_student_performance.pkl']
```