Department of Computer Engineering
College of Engineering
Polytechnic University of the Philippines Sta. Mesa

CMPE 40153: Big Data Laboratory

Data Manipulation of Student Mental Health by MD Shariful Islam

Submitted by:

Jamco, Kenneth A.
BSCOE 3-2

Submitted to:

EDCEL B. ARTIFICIO

I. Introduction

Student mental health refers to the psychological well-being and emotional resilience of students in educational settings. It includes aspects of students' life such as mental, emotional, and social aspect. This is important for their academic success, growth, and overall quality of life outside school. The domain I have chosen, Mental health of a student, addresses mental health issues of students from different year level and courses, as well as if they are getting some professional help. Poor mental health can significantly affect a student's academic performance. Mental health issues such as anxiety, depression and panic attack are focused by the given dataset. Recognizing and addressing these issues is important in creating a safe and nurturing environment where students can learn optimally, make them feel safe, and not alone. By conducting an analysis of the dataset, the objective is to gain a better understanding of student mental health and help to develop different ways to support the well-being of students in educational settings.

II. Data Dictionary

The dataset have 11 columns and 101 rows, containing information about student mental health and academic performance. The dataset includes variables such as Timestamp, Gender, Age, Course, Study Year, CGPA, Marital Status, and indicators for Depression, Anxiety, Panic Attack, and Specialist Treatment.

The Timestamp variable is about data and time when the survey response was conducted. The Gender variable represents the age. The Age variable shows their age. The Course variable is about what program are each student is enrolled. The Study Year variable indicates the current study year of the student. The CGPA variable (Cumulative Grade Point Average) is about their academic performance. The Marital Status variable is about the relationship status of the students. The variables "Has Depression", "Has anxiety", "Has Panic Attack", and "Specialist Treament", which are answerable by yes or no, are indicators whether the students have experienced depression, anxiety, panic attack or received professional help, respectively.

III. Data Manipulation using an RDD

A. Write the function that you use and describe what you did.

- filtered_rddyear1 = data_rdd.filter(lambda row: row[4] == "year 1" or row[4] == "Year 1")

```
[7]  filtered_rddyear1 = data_rdd.filter(lambda row: row[4] == "year 1" or row[4] == "Year 1")
```

The function I used here is the **filter()** function. The goal here is to filter or show only the students who are at Year 1 when they answered the survey. It helps to narrow down the data to only include records from the first year of study.

- filtered_rddyear3 = data_rdd.filter(lambda row: row[4] == "year 3" or row[4] == "Year 3")

```
[11] filtered_rddyear3 = data_rdd.filter(lambda row: row[4] == "year 3" or row[4] == "Year 3")
```

The function I used here is the **filter()** function. The goal here is to filter or show only the students who are at Year 3 when they answered the survey. It helps to narrow down the data to only include records from the third year of study.

- combined_rdd = filtered_rddyear1.union(filtered_rddyear3)

```
[14] combined_rdd = filtered_rddyear1.union(filtered_rddyear3)
```

The function I used here is the **union()** function. The goal here is to merged the data from filtered_rddyear1 and filtered_rddyear3 together into a single RDD. This allows for further analysis on the combined dataset that is deemed necessary for the analysis.

- print(first_row)
  for row in filtered_rddyear1.collect():
  print(row)

  print(first_row)
  for row in filtered_rddyear3.collect():
  print(row)

  print(first_row)
  for row in combined_rdd.collect():
  print(row)

  I used **collect()** function here to retrieve and print the year1, year2, and combined_rdd consists of students who answered year1 and year2. Allowing me to inspect the data and verify the results from the transformation of the dataset I made.

B. Add a screenshot
Pre-RDD transformation (part 1)

```
all_rows = data_rdd.collect()
for row in all_rows:
    print(row)
```

```
['Timestamp', 'Choose your gender', 'Age', 'What is your course?', 'Your current year of Study', 'What is your CGPA?', 'Marital status', 'Do you have Depression?', 'Do you have Anxi
['8/7/2020 12:02', 'Female', '18', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'Yes', 'No']
['8/7/2020 12:04', 'Male', '21', 'Islamic education', 'year 2', '3.00 - 3.49', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 12:05', 'Male', '19', 'BIT', 'Year 1', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 12:06', 'Female', '22', 'Laws', 'year 3', '3.00 - 3.49', 'Yes', 'Yes', 'No', 'No', 'No']
['8/7/2020 12:13', 'Male', '23', 'Mathemathics', 'year 4', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 12:31', 'Male', '19', 'Engineering', 'Year 2', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 12:32', 'Female', '23', 'Pendidikan islam', 'year 2', '3.50 - 4.00 ', 'Yes', 'Yes', 'No', 'Yes', 'No']
['8/7/2020 12:33', 'Female', '18', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 12:35', 'Female', '19', 'Human Resources', 'Year 2', '2.50 - 2.99', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 12:39', 'Male', '18', 'Irkhs', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'Yes', 'No']
['8/7/2020 12:39', 'Female', '20', 'Psychology', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 12:39', 'Female', '24', 'Engineering', 'Year 3', '3.50 - 4.00', 'Yes', 'Yes', 'No', 'No', 'No']
['8/7/2020 12:40', 'Female', '18', 'BCS', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'No', 'No']
['8/7/2020 12:41', 'Male', '19', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 12:43', 'Female', '18', 'KENMS', 'Year 2', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 12:43', 'Male', '24', 'BCS', 'Year 3', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 12:46', 'Female', '24', 'Accounting ', 'year 3', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 12:52', 'Female', '24', 'ENM', 'year 4', '3.00 - 3.49', 'Yes', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 13:05', 'Female', '20', 'BIT', 'Year 2', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 13:07', 'Female', '18', 'Marine science', 'year 2', '3.50 - 4.00', 'Yes', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 13:12', 'Female', '19', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 13:13', 'Female', '18', 'KOE', 'Year 2', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:13', 'Female', '24', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:15', 'Female', '24', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:17', 'Female', '23', 'BCS', 'Year 3', '3.50 - 4.00', 'No', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 13:29', 'Female', '18', 'Banking Studies', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:35', 'Female', '19', 'Engineering', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:41', 'Male', '18', 'Engineering', 'Year 2', '3.00 - 3.49', 'Yes', 'Yes', 'Yes', 'No', 'No']
```

Pre-RDD transformation (part 2)

```
all_rows = data_rdd.collect()
for row in all_rows:
    print(row)
```

```
['8/7/2020 15:08', 'Male', '23', 'TAASL', 'year 2', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 15:09', 'Male', '18', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'Yes', 'No']
['8/7/2020 15:12', 'Female', '19', 'Engineering', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 15:14', 'Female', '18', 'Engine', 'year 4', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 15:14', 'Male', '24', 'BCS', 'year 2', '3.00 - 3.49', 'No', 'Yes', 'No', 'No', 'No']
['8/7/2020 15:18', 'Female', '24', 'BCS', 'year 3', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 15:27', 'Female', '23', 'ALA', 'year 1', '2.50 - 2.99', 'Yes', 'Yes', 'No', 'Yes', 'Yes']
['8/7/2020 15:37', 'Female', '18', 'BCS', 'year 2', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 15:47', 'Female', '19', 'Biomedical science', 'year 3', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 15:48', 'Female', '20', 'koe', 'year 3', '3.00 - 3.49', 'Yes', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 15:57', 'Female', '19', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'No', 'Yes', 'Yes']
['8/7/2020 15:58', 'Male', '21', 'BCS', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:08', 'Male', '23', 'Kirkhs', 'Year 3', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:21', 'Female', '20', 'BENL', 'Year 3', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'No', 'No']
['8/7/2020 16:22', 'Female', '18', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:34', 'Female', '23', 'Benl', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:34', 'Female', '18', 'IT', 'Year 3', '3.00 - 3.49', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 16:53', 'Female', '19', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 17:05', 'Female', '18', 'CTS', 'Year 1', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 17:37', 'Female', '24', 'engin', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 17:46', 'Female', '24', 'Engine', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 17:50', 'Female', '23', 'Econs', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'Yes', 'No', 'No']
['8/7/2020 18:10', 'Female', '18', 'KOE', 'Year 3', '3.00 - 3.49', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 18:11', 'Male', '19', 'MHSC', 'Year 3', '3.00 - 3.49', 'Yes', 'Yes', 'No', 'Yes', 'No']
['8/7/2020 19:05', 'Female', '18', 'Malcom', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'No', 'No', 'No']
['8/7/2020 19:32', 'Female', '24', 'Kop', 'year 4', '3.00 - 3.49', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 20:36', 'Female', '24', 'Biomedical science', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 21:21', 'Female', '18', 'Laws', 'Year 3', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 22:35', 'Female', '19', 'BIT', 'Year 3', '3.00 - 3.49', 'Yes', 'Yes', 'No', 'No', 'No']
```

Post-RDD transformation filtered_rddyear1

```
print(first_row)
for row in filtered_rddyear1.collect():
    print(row)
```

```
['Timestamp', 'Choose your gender', 'Age', 'What is your course?', 'Your current year of Study', 'What is your CGPA?', 'Marital status', 'Do you have Depression?', 'Do you have Anxiety?', 'Do you have Panic attack?', 'Did you seek any specialist for a treatment?']
['8/7/2020 12:02', 'Female', '18', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'Yes', 'No']
['8/7/2020 12:05', 'Male', '19', 'BIT', 'Year 1', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 12:33', 'Female', '18', 'ECS', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 12:39', 'Male', '18', 'Irkhs', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'Yes', 'No']
['8/7/2020 12:39', 'Female', '20', 'Psychology', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 12:40', 'Female', '18', 'BCS', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'No', 'No']
['8/7/2020 12:41', 'Male', '19', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:12', 'Female', '19', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 13:13', 'Female', '24', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:15', 'Female', '24', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:29', 'Female', '18', 'Banking Studies', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:35', 'Female', '19', 'Engineering', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 14:29', 'Male', '19', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 14:41', 'Female', '19', 'BIT', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 14:43', 'Female', '18', 'Engineering', 'year 1', '2.00 - 2.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 14:45', 'Female', '19', 'BIT', 'year 1', '2.50 - 2.99', 'No', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 14:47', 'Female', '18', 'KIRKHS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 15:07', 'Male', '', 'BIT', 'year 1', '0 - 1.99', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 15:09', 'Male', '18', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'Yes', 'No']
['8/7/2020 15:12', 'Female', '19', 'Engineering', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 15:27', 'Female', '23', 'ALA', 'year 1', '2.50 - 2.99', 'Yes', 'Yes', 'No', 'Yes', 'Yes']
['8/7/2020 15:57', 'Female', '19', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'No', 'Yes', 'Yes']
['8/7/2020 15:58', 'Male', '21', 'BCS', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:22', 'Female', '18', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:34', 'Female', '23', 'Benl', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:53', 'Female', '19', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 17:05', 'Female', '18', 'CTS', 'Year 1', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 17:37', 'Female', '24', 'engin', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 17:46', 'Female', '24', 'Engine', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 17:50', 'Female', '23', 'Econs', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'Yes', 'No', 'No']
['8/7/2020 19:05', 'Female', '18', 'Malcom', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'No', 'No', 'No']
['8/7/2020 20:36', 'Female', '24', 'Biomedical science', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['9/7/2020 6:57', 'Male', '18', 'Biomedical science', 'year 1', '0 - 1.99', 'No', 'No', 'No', 'No', 'No']
['9/7/2020 11:57', 'Female', '24', 'KOE', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'Yes', 'No']
['9/7/2020 13:15', 'Female', '23', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'No', 'No']
['13/07/2020 10:12:26', 'Female', '19', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'No', 'No']
['13/07/2020 10:14:46', 'Male', '23', 'Radiography', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['13/07/2020 10:33:47', 'Female', '18', 'psychology', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'Yes', 'No', 'Yes']
['13/07/2020 11:46:13', 'Female', '18', 'psychology', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'Yes', 'Yes', 'No']
['13/07/2020 11:49:02', 'Male', '24', 'BIT', 'year 1', '3.00 - 3.49', 'No', 'No', 'Yes', 'No', 'No']
['13/07/2020 16:15:13', 'Female', '18', 'BENL', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'No', 'No']
['13/07/2020 19:08:32', 'Female', '18', 'Islamic Education', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['13/07/2020 19:56:49', 'Female', '21', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
```

Post-RDD transformation filtered_rddyear3

```
[61] print(first_row)
     for row in filtered_rddyear3.collect():
         print(row)
```

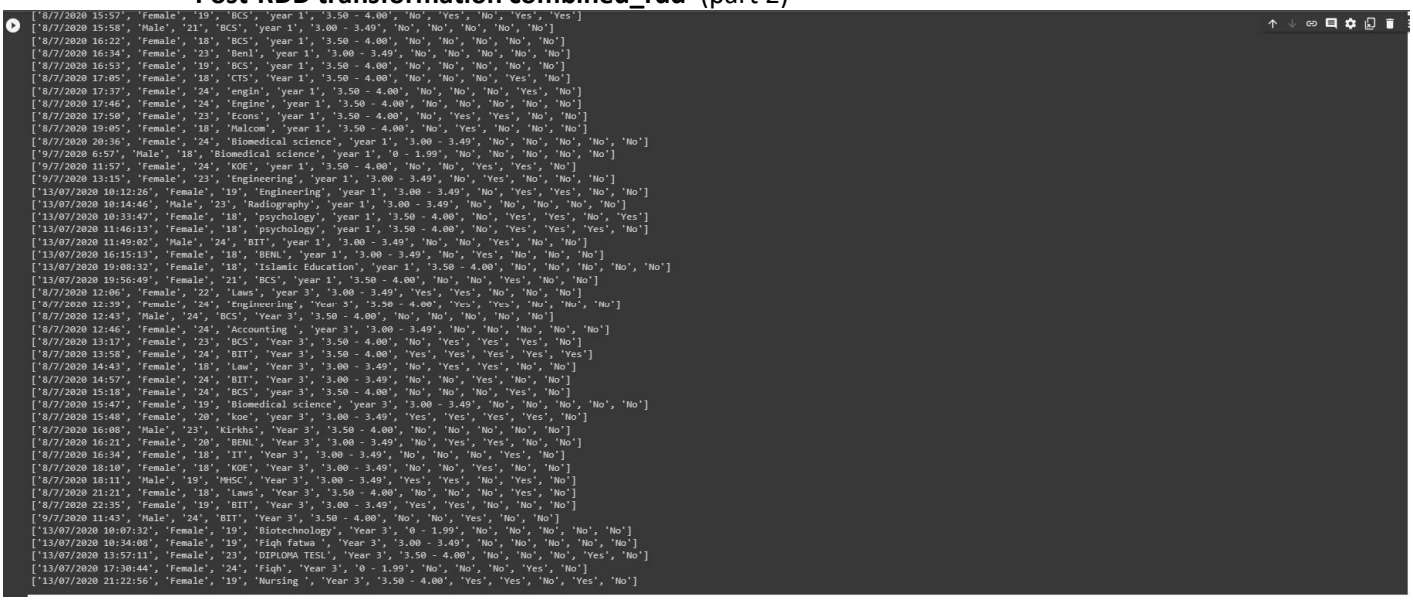
```
['Timestamp', 'Choose your gender', 'Age', 'What is your course?', 'Your current year of Study', 'What is your CGPA?', 'Marital status', 'Do you have Depression?', 'Do you have Anxiety?', 'Do you have Panic
['8/7/2020 12:06', 'Female', '22', 'Laws', 'year 3', '3.00 - 3.49', 'Yes', 'Yes', 'No', 'No', 'No']
['8/7/2020 12:39', 'Female', '24', 'Engineering', 'Year 3', '3.50 - 4.00', 'Yes', 'Yes', 'No', 'No', 'No']
['8/7/2020 12:43', 'Male', '24', 'BCS', 'Year 3', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 12:46', 'Female', '24', 'Accounting ', 'year 3', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:17', 'Female', '23', 'BCS', 'Year 3', '3.50 - 4.00', 'No', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 13:58', 'Female', '24', 'BIT', 'Year 3', '3.50 - 4.00', 'Yes', 'Yes', 'Yes', 'Yes', 'Yes']
['8/7/2020 14:43', 'Female', '18', 'Law', 'Year 3', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'No', 'No']
['8/7/2020 14:57', 'Female', '24', 'BIT', 'Year 3', '3.00 - 3.49', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 15:18', 'Female', '24', 'BCS', 'year 3', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 15:47', 'Female', '19', 'Biomedical science', 'year 3', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 15:48', 'Female', '20', 'koe', 'year 3', '3.00 - 3.49', 'Yes', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 16:08', 'Male', '23', 'Kirkhs', 'Year 3', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:21', 'Female', '20', 'BENL', 'Year 3', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'No', 'No']
['8/7/2020 16:34', 'Female', '18', 'IT', 'Year 3', '3.00 - 3.49', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 18:10', 'Female', '18', 'KOE', 'Year 3', '3.00 - 3.49', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 18:11', 'Male', '19', 'MHSC', 'Year 3', '3.00 - 3.49', 'Yes', 'Yes', 'No', 'Yes', 'No']
['8/7/2020 21:21', 'Female', '18', 'Laws', 'Year 3', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 22:35', 'Female', '19', 'BIT', 'Year 3', '3.00 - 3.49', 'Yes', 'Yes', 'No', 'No', 'No']
['9/7/2020 11:43', 'Male', '24', 'BIT', 'Year 3', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['13/07/2020 10:07:32', 'Female', '19', 'Biotechnology', 'Year 3', '0 - 1.99', 'No', 'No', 'No', 'No', 'No']
['13/07/2020 10:34:08', 'Female', '19', 'Fiqh fatwa ', 'Year 3', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['13/07/2020 13:57:11', 'Female', '23', 'DIPLOMA TESL', 'Year 3', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['13/07/2020 17:30:44', 'Female', '24', 'Fiqh', 'Year 3', '0 - 1.99', 'No', 'No', 'No', 'Yes', 'No']
['13/07/2020 21:22:56', 'Female', '19', 'Nursing ', 'Year 3', '3.50 - 4.00', 'Yes', 'Yes', 'No', 'Yes', 'No']
```

Post-RDD transformation combined_rdd (part 1)

```
print(first_row)
for row in combined_rdd.collect():
    print(row)
```

```
['Timestamp', 'Choose your gender', 'Age', 'What is your course?', 'Your current year of Study', 'What is your CGPA?', 'Marital status', 'Do you have Depression?', 'Do you have Anxiety?', 'Do you have Panic attack?', 'Did you seek any specialist fo
['8/7/2020 12:02', 'Female', '18', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'Yes', 'No']
['8/7/2020 12:05', 'Male', '19', 'BIT', 'Year 1', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 12:33', 'Female', '18', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 12:39', 'Male', '18', 'Irkhs', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'Yes', 'No']
['8/7/2020 12:39', 'Female', '20', 'Psychology', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 12:40', 'Female', '18', 'BCS', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'No', 'No']
['8/7/2020 12:41', 'Male', '19', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:12', 'Female', '19', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 13:13', 'Female', '24', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:15', 'Female', '24', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:29', 'Female', '18', 'Banking Studies', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 13:35', 'Female', '19', 'Engineering', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 14:29', 'Male', '19', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 14:41', 'Female', '19', 'BIT', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 14:43', 'Female', '18', 'Engineering', 'year 1', '2.00 - 2.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 14:45', 'Female', '19', 'BIT', 'year 1', '2.50 - 2.99', 'No', 'Yes', 'Yes', 'Yes', 'No']
['8/7/2020 14:47', 'Female', '18', 'KIRKHS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 15:07', 'Male', '', 'BIT', 'year 1', '0 - 1.99', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 15:09', 'Male', '18', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'Yes', 'No']
['8/7/2020 15:17', 'Female', '19', 'Engineering', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 15:27', 'Female', '23', 'ALA', 'year 1', '2.50 - 2.99', 'Yes', 'Yes', 'No', 'Yes', 'Yes']
['8/7/2020 15:57', 'Female', '19', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'No', 'Yes', 'Yes']
['8/7/2020 15:58', 'Male', '21', 'BCS', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:22', 'Female', '18', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:34', 'Female', '23', 'Beni', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 16:53', 'Female', '19', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 17:05', 'Female', '18', 'CTS', 'Year 1', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 17:37', 'Female', '24', 'engin', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'Yes', 'No']
['8/7/2020 17:46', 'Female', '24', 'Engine', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['8/7/2020 17:50', 'Female', '23', 'Econs', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'Yes', 'No', 'No']
['8/7/2020 19:05', 'Female', '18', 'Malcom', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'No', 'No', 'No']
['8/7/2020 20:36', 'Female', '24', 'Biomedical science', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['9/7/2020 6:57', 'Male', '18', 'Biomedical science', 'year 1', '0 - 1.99', 'No', 'No', 'No', 'No', 'No']
['9/7/2020 11:57', 'Female', '24', 'KOE', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'Yes', 'No']
['9/7/2020 13:15', 'Female', '23', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'No', 'No']
['13/07/2020 10:12:26', 'Female', '19', 'Engineering', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'Yes', 'No', 'No']
['13/07/2020 10:14:46', 'Male', '23', 'Radiography', 'year 1', '3.00 - 3.49', 'No', 'No', 'No', 'No', 'No']
['13/07/2020 10:33:47', 'Female', '18', 'psychology', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'Yes', 'No', 'Yes']
['13/07/2020 11:46:13', 'Female', '18', 'psychology', 'year 1', '3.50 - 4.00', 'No', 'Yes', 'Yes', 'Yes', 'No']
['13/07/2020 11:49:02', 'Male', '24', 'BIT', 'year 1', '3.00 - 3.49', 'No', 'No', 'Yes', 'No', 'No']
['13/07/2020 16:15:13', 'Female', '18', 'BENL', 'year 1', '3.00 - 3.49', 'No', 'Yes', 'No', 'No', 'No']
['13/07/2020 19:08:32', 'Female', '18', 'Islamic Education', 'year 1', '3.50 - 4.00', 'No', 'No', 'No', 'No', 'No']
['13/07/2020 19:56:49', 'Female', '21', 'BCS', 'year 1', '3.50 - 4.00', 'No', 'No', 'Yes', 'No', 'No']
['8/7/2020 12:06', 'Female', '22', 'Laws', 'year 3', '3.00 - 3.49', 'Yes', 'Yes', 'No', 'No', 'No']
['8/7/2020 12:39', 'Female', '24', 'Engineering', 'Year 3', '3.50 - 4.00', 'Yes', 'Yes', 'No', 'No', 'No']
```

Post-RDD transformation combined_rdd (part 2)



C. Reflection about data manipulation using an RDD

Data manipulation in PySpark using RDD involves different functions in order to transform dataset into our liking or goal. The function I used are collect(), filter(), and union() are essential in this process. The collect() function collects all data from an RDD and bringing it to the user for further analysis. The filter() function creates a new RDD by selecting elements that meet conditions given by the user or analyst. Its goal is to reduce the size of the dataset for more focused analysis. The union() function combines multiple RDD by concatenating their contents. It is used to merge dataset in order to conduct an easier analysis of mutiple RDDs.

Overall, working with RDDs and using functions like collect(), filter(), and union() have helped me manipulate my chosen dataset. I can see these functions to be useful and will help me conduct analysis and meaningful insights throughout this semester. Looking ahead, further exploration of RDDs in PySpark promises to unlock additional possibilities for complex data manipulation tasks and deeper understanding of the underlying data in future if I decided to pursue this track.

IV. Data manipulation using a Dataframe

A. Write the function that you use and describe what you did.

- new_column_names = {
  "Choose your gender": "Gender",
  "Your current year of Study": "Study Year",
  "Do you have Depression?": "Has Depression",
  "Do you have Anxiety?": "Has Anxiety",
  "Do you have Panic attack?": "Has Panic Attack",
  "Did you seek any specialist for a treatment?": "Specialist Treatment",
  "What is your course?": "Course",
  "What is your CGPA?" : "CGPA"
  }

  for old_name, new_name in new_column_names.items():

df = df.withColumnRenamed(old_name, new_name)
df.show()

```
[20] new_column_names = {
        "Choose your gender": "Gender",
        "Your current year of Study": "Study Year",
        "Do you have Depression?": "Has Depression",
        "Do you have Anxiety?": "Has Anxiety",
        "Do you have Panic attack?": "Has Panic Attack",
        "Did you seek any specialist for a treatment?": "Specialist Treatment",
        "What is your course?": "Course",
        "What is your CGPA?" : "CGPA"
    }

    for old_name, new_name in new_column_names.items():
        df = df.withColumnRenamed(old_name, new_name)

    df.show()
```

The function I used here is the **withColumnRenamed()** function. The goal here is to change the name of the column from the dataset because I think it is long and some are still in question form. This will help me read column names faster and will give me convenience when I use column names in a code.

- df.select("Course", "Study Year", "Has Depression", "Specialist Treatment").show(n=1000)

```
[21] df.select("Course", "Study Year", "Has Depression", "Specialist Treatment").show(n=1000)
```

The function I used here is the **select()** function. The goal here is to select columns that I need in my analysis. In this case, I choose to only show the column Course, Study Year, Has Depression, and Specialist Treatment because I want to see if there is any pattern that can help me form a hypothesis.

- df.orderBy("Study Year").show(n=1000)

```
[23] df.orderBy("Study Year").show(n=1000)
```

The function I used here is the **orderBy()** function. The goal here is to arrange the dataset based on what year the students are. By default the order is ascending, so the data will be arranged by Year 1 then Year 2 and so on. This function can help clean and organize data depending on how the user wants it.

B. Add a screenshot

Pre-df transformation part 1

```
df.show(n=1000)
```

Timestamp	Gender	Age	Course	Study Year	CGPA	Marital status	Has Depression	Has Anxiety	Has Panic Attack	Specialist Treatment
8/7/2020 12:02	Female	18	Engineering	Year 1	3.00 - 3.49	No	Yes	No	Yes	No
8/7/2020 12:04	Male	21	Islamic education	Year 2	3.00 - 3.49	No	No	Yes	No	No
8/7/2020 12:05	Male	19	BIT	Year 1	3.00 - 3.49	No	Yes	Yes	Yes	No
8/7/2020 12:06	Female	22	Laws	Year 3	3.00 - 3.49	Yes	Yes	No	No	No
8/7/2020 12:13	Male	23	Mathemathics	Year 4	3.00 - 3.49	No	No	No	No	No
8/7/2020 12:31	Male	19	Engineering	Year 2	3.50 - 4.00	No	No	No	Yes	No
8/7/2020 12:32	Female	23	Pendidikan islam	Year 2	3.50 - 4.00	Yes	Yes	No	Yes	No
8/7/2020 12:33	Female	18	BCS	Year 1	3.50 - 4.00	No	No	Yes	No	No
8/7/2020 12:35	Female	19	Human Resources	Year 2	2.50 - 2.99	No	No	No	No	No
8/7/2020 12:39	Male	18	Irkhs	Year 1	3.50 - 4.00	No	No	Yes	Yes	No
8/7/2020 12:39	Female	20	Psychology	Year 1	3.50 - 4.00	No	No	No	No	No
8/7/2020 12:39	Female	24	Engineering	Year 3	3.50 - 4.00	Yes	Yes	No	No	No
8/7/2020 12:40	Female	18	BCS	Year 1	3.00 - 3.49	No	Yes	No	No	No
8/7/2020 12:41	Male	19	Engineering	Year 1	3.00 - 3.49	No	No	No	No	No
8/7/2020 12:43	Female	18	KENMS	Year 2	3.50 - 4.00	No	No	Yes	No	No
8/7/2020 12:43	Male	24	BCS	Year 3	3.50 - 4.00	No	No	No	No	No
8/7/2020 12:46	Female	24	Accounting	Year 3	3.00 - 3.49	No	No	No	No	No
8/7/2020 12:52	Female	24	ENM	Year 4	3.00 - 3.49	Yes	Yes	Yes	Yes	No
8/7/2020 13:05	Female	20	BIT	Year 2	3.50 - 4.00	No	No	Yes	No	No
8/7/2020 13:07	Female	18	Marine science	Year 2	3.50 - 4.00	Yes	Yes	Yes	Yes	No
8/7/2020 13:12	Female	19	Engineering	Year 1	3.00 - 3.49	No	No	No	Yes	No
8/7/2020 13:13	Female	18	KOE	Year 2	3.00 - 3.49	No	No	No	No	No
8/7/2020 13:13	Female	24	BCS	Year 1	3.50 - 4.00	No	No	No	No	No
8/7/2020 13:15	Female	24	Engineering	Year 1	3.00 - 3.49	No	No	No	No	No
8/7/2020 13:17	Female	23	BCS	Year 3	3.50 - 4.00	No	Yes	Yes	Yes	No
8/7/2020 13:29	Female	18	Banking Studies	Year 1	3.50 - 4.00	No	No	No	No	No
8/7/2020 13:35	Female	19	Engineering	Year 1	3.50 - 4.00	No	No	No	No	No
8/7/2020 13:41	Male	18	Engineering	Year 2	3.00 - 3.49	Yes	Yes	Yes	No	No
8/7/2020 13:58	Female	24	BIT	Year 3	3.50 - 4.00	Yes	Yes	Yes	Yes	Yes
8/7/2020 14:05	Female	24	BCS	Year 4	3.50 - 4.00	No	No	No	No	No
8/7/2020 14:27	Female	23	Business Administ...	Year 2	3.00 - 3.49	No	No	No	No	No
8/7/2020 14:29	Male	18	BCS	Year 2	3.00 - 3.49	No	No	No	No	No
8/7/2020 14:29	Male	19	BCS	Year 1	3.50 - 4.00	No	No	No	Yes	No
8/7/2020 14:31	Male	18	BCS	Year 2	3.50 - 4.00	Yes	Yes	Yes	No	Yes
8/7/2020 14:41	Female	19	BIT	Year 1	3.00 - 3.49	No	Yes	Yes	Yes	No
8/7/2020 14:43	Female	18	Engineering	Year 1	2.00 - 2.49	No	No	No	No	No
8/7/2020 14:43	Female	18	Law	Year 3	3.00 - 3.49	No	Yes	Yes	No	No
8/7/2020 14:45	Female	19	BIT	Year 1	2.50 - 2.99	No	Yes	Yes	Yes	No
8/7/2020 14:47	Female	18	KIRKHS	Year 1	3.50 - 4.00	No	No	No	No	No
8/7/2020 14:56	Female	24	Engineering	Year 2	2.50 - 2.99	Yes	Yes	No	Yes	Yes
8/7/2020 14:57	Female	24	BIT	Year 3	3.00 - 3.49	No	No	Yes	No	No
8/7/2020 14:57	Female	22	Engineering	Year 4	3.50 - 4.00	No	No	No	No	No
8/7/2020 14:58	Female	20	Usuluddin	Year 2	3.00 - 3.49	No	Yes	No	No	No
8/7/2020 15:07	Male	null	BIT	Year 1	0 - 1.99	No	No	No	No	No
8/7/2020 15:08	Male	23	TAASL	Year 2	3.50 - 4.00	No	No	No	Yes	No
8/7/2020 15:09	Male	18	BCS	Year 1	3.50 - 4.00	No	No	Yes	Yes	No
8/7/2020 15:12	Female	19	Engineering	Year 1	3.50 - 4.00	No	No	Yes	No	No
8/7/2020 15:14	Female	18	Engine	Year 4	3.50 - 4.00	No	No	No	No	No
8/7/2020 15:14	Male	24	BCS	Year 2	3.00 - 3.49	No	Yes	No	No	No
8/7/2020 15:18	Female	24	BCS	Year 3	3.50 - 4.00	No	No	No	Yes	No
8/7/2020 15:27	Female	23	ALA	Year 1	2.50 - 2.99	Yes	Yes	No	Yes	Yes
8/7/2020 15:37	Female	18	BCS	Year 2	3.50 - 4.00	No	No	Yes	No	No
8/7/2020 15:47	Female	19	Biomedical science	Year 3	3.00 - 3.49	No	No	No	No	No
8/7/2020 15:48	Female	20	koe	Year 3	3.00 - 3.49	Yes	Yes	Yes	Yes	No

Pre-df transformation part 2

```
df.show(n=1000)
```

```
|     8/7/2020 15:09|  Male|  18|                BCS|    Year 1| 3.50 - 4.00|          No|          No|       Yes|          Yes|                No|
|     8/7/2020 15:12|Female|  19|        Engineering |    Year 1| 3.50 - 4.00|          No|          No|       Yes|           No|                No|
|     8/7/2020 15:14|Female|  18|             Engine |    Year 4| 3.50 - 4.00|          No|          No|        No|           No|                No|
|     8/7/2020 15:14|  Male|  24|                BCS|    Year 2| 3.00 - 3.49|          No|         Yes|        No|           No|                No|
|     8/7/2020 15:18|Female|  24|                BCS|    Year 3| 3.50 - 4.00|          No|          No|        No|          Yes|                No|
|     8/7/2020 15:27|Female|  23|                ALA|    Year 1| 2.50 - 2.99|         Yes|         Yes|        No|          Yes|               Yes|
|     8/7/2020 15:37|Female|  18|                BCS|    Year 2| 3.50 - 4.00|          No|          No|       Yes|           No|                No|
|     8/7/2020 15:47|Female|  19|  Biomedical science|    Year 3| 3.00 - 3.49|          No|          No|        No|           No|                No|
|     8/7/2020 15:48|Female|  20|               koe  |    Year 3| 3.00 - 3.49|         Yes|         Yes|       Yes|          Yes|                No|
|     8/7/2020 15:57|Female|  19|                BCS|    Year 1| 3.50 - 4.00|          No|         Yes|        No|          Yes|               Yes|
|     8/7/2020 15:58|  Male|  21|                BCS|    Year 1| 3.00 - 3.49|          No|          No|        No|           No|                No|
|     8/7/2020 16:08|  Male|  23|             Kirkhs |    Year 3| 3.50 - 4.00|          No|          No|        No|           No|                No|
|     8/7/2020 16:21|Female|  20|               BENL |    Year 3| 3.00 - 3.49|          No|         Yes|       Yes|           No|                No|
|     8/7/2020 16:22|Female|  18|                BCS|    Year 1| 3.50 - 4.00|          No|          No|        No|           No|                No|
|     8/7/2020 16:34|Female|  23|               Benl |    Year 1| 3.00 - 3.49|          No|          No|        No|           No|                No|
|     8/7/2020 16:34|Female|  18|                 IT |    Year 3| 3.00 - 3.49|          No|          No|        No|          Yes|                No|
|     8/7/2020 16:53|Female|  19|                BCS|    Year 1| 3.50 - 4.00|          No|          No|        No|           No|                No|
|     8/7/2020 17:05|Female|  18|                CTS|    Year 1| 3.50 - 4.00|          No|          No|        No|          Yes|                No|
|     8/7/2020 17:37|Female|  24|              engin |    Year 1| 3.50 - 4.00|          No|          No|        No|          Yes|                No|
|     8/7/2020 17:46|Female|  24|             Engine |    Year 1| 3.50 - 4.00|          No|          No|        No|           No|                No|
|     8/7/2020 17:50|Female|  23|              Econs |    Year 1| 3.50 - 4.00|          No|         Yes|       Yes|           No|                No|
|     8/7/2020 18:10|Female|  18|               KOE  |    Year 3| 3.00 - 3.49|          No|          No|       Yes|           No|                No|
|     8/7/2020 18:11|  Male|  19|               MHSC |    Year 3| 3.00 - 3.49|         Yes|         Yes|        No|          Yes|                No|
|     8/7/2020 19:05|Female|  18|             Malcom |    Year 1| 3.50 - 4.00|          No|         Yes|        No|           No|                No|
|     8/7/2020 19:32|Female|  24|                Kop |    Year 4| 3.00 - 3.49|          No|          No|       Yes|           No|                No|
|     8/7/2020 20:36|Female|  24|  Biomedical science|    Year 1| 3.00 - 3.49|          No|          No|        No|           No|                No|
|     8/7/2020 21:21|Female|  18|               Laws |    Year 3| 3.50 - 4.00|          No|          No|        No|          Yes|                No|
|     8/7/2020 22:35|Female|  19|                BIT |    Year 3| 3.00 - 3.49|         Yes|         Yes|        No|           No|                No|
|      9/7/2020 6:57|  Male|  18|  Biomedical science|    Year 1|    0 - 1.99|          No|          No|        No|           No|                No|
|     9/7/2020 11:43|  Male|  24|                BIT |    Year 3| 3.50 - 4.00|          No|          No|       Yes|           No|                No|
|     9/7/2020 11:57|Female|  24|                KOE |    Year 1| 3.50 - 4.00|          No|          No|       Yes|          Yes|                No|
|     9/7/2020 13:15|Female|  23|        Engineering |    Year 1| 3.00 - 3.49|          No|         Yes|        No|           No|                No|
|     9/7/2020 18:24|Female|  18|    Human Sciences  |    Year 2| 3.00 - 3.49|          No|          No|        No|          Yes|                No|
|13/07/2020 10:07:32|Female|  19|      Biotechnology |    Year 3|    0 - 1.99|          No|          No|        No|           No|                No|
|13/07/2020 10:10:30|Female|  18|        Engineering |    Year 4| 3.50 - 4.00|          No|          No|        No|           No|                No|
|13/07/2020 10:11:26|Female|  24|     Communication  |    Year 2| 3.50 - 4.00|         Yes|         Yes|       Yes|          Yes|                No|
|13/07/2020 10:12:18|Female|  24|    Diploma Nursing |    Year 2| 3.50 - 4.00|          No|          No|        No|           No|                No|
|13/07/2020 10:12:26|Female|  19|        Engineering |    Year 1| 3.00 - 3.49|          No|         Yes|       Yes|           No|                No|
|13/07/2020 10:12:28|Female|  19|  Pendidikan Islam  |    Year 2| 3.00 - 3.49|          No|          No|        No|           No|                No|
|13/07/2020 10:14:46|  Male|  23|        Radiography |    Year 1| 3.00 - 3.49|          No|          No|        No|           No|                No|
|13/07/2020 10:33:47|Female|  18|         psychology |    Year 1| 3.50 - 4.00|          No|         Yes|       Yes|           No|               Yes|
|13/07/2020 10:34:08|Female|  19|         Fiqh fatwa |    Year 3| 3.00 - 3.49|          No|          No|        No|           No|                No|
|13/07/2020 11:46:13|Female|  18|         psychology |    Year 1| 3.50 - 4.00|          No|         Yes|       Yes|          Yes|                No|
|13/07/2020 11:49:02|  Male|  24|                BIT |    Year 1| 3.00 - 3.49|          No|          No|       Yes|           No|                No|
|13/07/2020 11:54:58|  Male|  24|        Engineering |    Year 2| 2.00 - 2.49|          No|          No|        No|          Yes|                No|
|13/07/2020 13:57:11|Female|  23|       DIPLOMA TESL |    Year 3| 3.50 - 4.00|          No|          No|        No|          Yes|                No|
|13/07/2020 14:38:12|  Male|  18|                Koe |    Year 2| 3.00 - 3.49|          No|          No|       Yes|           No|                No|
|13/07/2020 14:48:05|Female|  19|               KOE  |    Year 2| 3.00 - 3.49|         Yes|         Yes|        No|           No|                No|
|13/07/2020 16:15:13|Female|  18|               BENL |    Year 1| 3.00 - 3.49|          No|         Yes|        No|           No|                No|
|13/07/2020 17:30:44|Female|  24|                Fiqh|    Year 3|    0 - 1.99|          No|          No|        No|          Yes|                No|
|13/07/2020 19:08:32|Female|  18|  Islamic Education |    Year 1| 3.50 - 4.00|          No|          No|        No|           No|                No|
|13/07/2020 19:56:49|Female|  21|                BCS|    Year 1| 3.50 - 4.00|          No|          No|       Yes|           No|                No|
|13/07/2020 21:21:42|  Male|  18|        Engineering |    Year 2| 3.00 - 3.49|          No|         Yes|       Yes|           No|                No|
|13/07/2020 21:22:56|Female|  19|           Nursing  |    Year 3| 3.50 - 4.00|         Yes|         Yes|        No|          Yes|                No|
|13/07/2020 21:23:57|Female|  23|   Pendidikan Islam |    Year 4| 3.50 - 4.00|          No|          No|        No|           No|                No|
|18/07/2020 20:16:21|  Male|  20|  Biomedical science|    Year 2| 3.00 - 3.49|          No|          No|        No|           No|                No|
+-------------------+------+----+--------------------+----------+------------+------------+------------+----------+-------------+--------------------+
```

Post-df transformation withColumnRenamed()

```
new_column_names = {
    "Choose your gender": "Gender",
    "Your current year of Study": "Study Year",
    "Do you have Depression?": "Has Depression",
    "Do you have Anxiety?": "Has Anxiety",
    "Do you have Panic attack?": "Has Panic Attack",
    "Did you seek any specialist for a treatment?": "Specialist Treatment",
    "What is your course?": "Course",
    "What is your CGPA?" : "CGPA"
}

for old_name, new_name in new_column_names.items():
    df = df.withColumnRenamed(old_name, new_name)

df.show()
```

```
+--------------+------+---+-----------------+----------+------------+--------------+--------------+-----------+----------------+--------------------+
|     Timestamp|Gender|Age|           Course|Study Year|        CGPA|Marital status|Has Depression|Has Anxiety|Has Panic Attack|Specialist Treatment|
+--------------+------+---+-----------------+----------+------------+--------------+--------------+-----------+----------------+--------------------+
|8/7/2020 12:02|Female| 18|      Engineering|    year 1| 3.00 - 3.49|            No|           Yes|         No|             Yes|                  No|
|8/7/2020 12:04|  Male| 21|Islamic education|    year 2| 3.00 - 3.49|            No|            No|        Yes|              No|                  No|
|8/7/2020 12:05|  Male| 19|              BIT|    Year 1| 3.00 - 3.49|            No|           Yes|        Yes|             Yes|                  No|
|8/7/2020 12:06|Female| 22|             Laws|    year 3| 3.00 - 3.49|           Yes|           Yes|         No|              No|                  No|
|8/7/2020 12:13|  Male| 23|     Mathemathics|    year 4| 3.00 - 3.49|            No|            No|         No|              No|                  No|
|8/7/2020 12:31|  Male| 19|      Engineering|    Year 2| 3.50 - 4.00|            No|            No|         No|             Yes|                  No|
|8/7/2020 12:32|Female| 23| Pendidikan islam|    year 2|3.50 - 4.00 |           Yes|           Yes|         No|             Yes|                  No|
|8/7/2020 12:33|Female| 18|              BCS|    year 1| 3.50 - 4.00|            No|            No|        Yes|              No|                  No|
|8/7/2020 12:35|Female| 19|  Human Resources|    Year 2| 2.50 - 2.99|            No|            No|         No|              No|                  No|
|8/7/2020 12:39|  Male| 18|             Irkhs|    year 1| 3.50 - 4.00|            No|            No|        Yes|             Yes|                  No|
|8/7/2020 12:39|Female| 20|       Psychology|    year 1| 3.50 - 4.00|            No|            No|         No|              No|                  No|
|8/7/2020 12:39|Female| 24|      Engineering|    Year 3| 3.50 - 4.00|           Yes|           Yes|         No|              No|                  No|
|8/7/2020 12:40|Female| 18|              BCS|    year 1| 3.00 - 3.49|            No|           Yes|         No|              No|                  No|
|8/7/2020 12:41|  Male| 19|      Engineering|    year 1| 3.00 - 3.49|            No|            No|         No|              No|                  No|
|8/7/2020 12:43|Female| 18|            KENMS|    Year 2| 3.50 - 4.00|            No|            No|        Yes|              No|                  No|
|8/7/2020 12:43|  Male| 24|              BCS|    Year 3| 3.50 - 4.00|            No|            No|         No|              No|                  No|
|8/7/2020 12:46|Female| 24|      Accounting |    year 3| 3.00 - 3.49|            No|            No|         No|              No|                  No|
|8/7/2020 12:52|Female| 24|              ENM|    year 4| 3.00 - 3.49|           Yes|           Yes|        Yes|             Yes|                  No|
|8/7/2020 13:05|Female| 20|              BIT|    Year 2| 3.50 - 4.00|            No|            No|        Yes|              No|                  No|
|8/7/2020 13:07|Female| 18|   Marine science|    year 2| 3.50 - 4.00|           Yes|           Yes|        Yes|             Yes|                  No|
+--------------+------+---+-----------------+----------+------------+--------------+--------------+-----------+----------------+--------------------+
only showing top 20 rows
```

Post-df transformation select()

```
df.select("Course", "Study Year", "Has Depression", "Specialist Treatment").show(n=1000)
```

```
+--------------------+----------+--------------+--------------------+
|              Course|Study Year|Has Depression|Specialist Treatment|
+--------------------+----------+--------------+--------------------+
|         Engineering|    year 1|           Yes|                  No|
|   Islamic education|    year 2|            No|                  No|
|                 BIT|    Year 1|           Yes|                  No|
|                Laws|    year 3|           Yes|                  No|
|        Mathemathics|    year 4|            No|                  No|
|         Engineering|    Year 2|            No|                  No|
|    Pendidikan islam|    year 2|           Yes|                  No|
|                 BCS|    year 1|            No|                  No|
|     Human Resources|    Year 2|            No|                  No|
|               Irkhs|    year 1|            No|                  No|
|          Psychology|    year 1|            No|                  No|
|         Engineering|    Year 3|           Yes|                  No|
|                 BCS|    year 1|           Yes|                  No|
|         Engineering|    year 1|            No|                  No|
|               KENMS|    Year 2|            No|                  No|
|                 BCS|    Year 3|            No|                  No|
|         Accounting |    year 3|            No|                  No|
|                 ENM|    year 4|           Yes|                  No|
|                 BIT|    Year 2|            No|                  No|
|      Marine science|    year 2|           Yes|                  No|
|         Engineering|    year 1|            No|                  No|
|                 KOE|    Year 2|            No|                  No|
|                 BCS|    year 1|            No|                  No|
|         Engineering|    year 1|            No|                  No|
|                 BCS|    Year 3|           Yes|                  No|
|     Banking Studies|    year 1|            No|                  No|
|         Engineering|    year 1|            No|                  No|
|         Engineering|    Year 2|           Yes|                  No|
|                 BIT|    Year 3|           Yes|                 Yes|
|                 BCS|    year 4|            No|                  No|
|Business Administ...|    Year 2|            No|                  No|
|                 BCS|    year 2|            No|                  No|
|                 BCS|    year 1|            No|                  No|
|                 BCS|    Year 2|           Yes|                 Yes|
|                 BIT|    year 1|           Yes|                  No|
|         Engineering|    year 1|            No|                  No|
|                 Law|    Year 3|           Yes|                  No|
|                 BIT|    year 1|           Yes|                  No|
|               KIRKHS|    year 1|            No|                  No|
|         Engineering|    Year 2|           Yes|                 Yes|
|                 BIT|    Year 3|            No|                  No|
```

Post-df transformation orderBy() part1

```
ordered_df = sentence_df.orderBy("Study Year").show(n=1000)
```

Course	Study Year	Has Depression	Specialist Treatment
Engineering	Year 1	Yes	No
BIT	Year 1	Yes	No
BCS	Year 1	No	No
Irkhs	Year 1	No	No
Psychology	Year 1	No	No
BCS	Year 1	Yes	No
Engineering	Year 1	No	No
Engineering	Year 1	No	No
BCS	Year 1	No	No
Engineering	Year 1	No	No
Banking Studies	Year 1	No	No
Engineering	Year 1	No	No
BCS	Year 1	No	No
BIT	Year 1	Yes	No
Engineering	Year 1	No	No
BIT	Year 1	Yes	No
KIRKHS	Year 1	No	No
BIT	Year 1	No	No
BCS	Year 1	No	No
Engineering	Year 1	No	No
ALA	Year 1	Yes	Yes
BCS	Year 1	Yes	Yes
BCS	Year 1	No	No
BCS	Year 1	No	No
Benl	Year 1	No	No
BCS	Year 1	No	No
CTS	Year 1	No	No
engin	Year 1	No	No
Engine	Year 1	No	No
Econs	Year 1	Yes	No
Malcom	Year 1	Yes	No
Biomedical science	Year 1	No	No
Biomedical science	Year 1	No	No
KOE	Year 1	No	No
Engineering	Year 1	Yes	No
Engineering	Year 1	Yes	No
Radiography	Year 1	No	No
psychology	Year 1	Yes	Yes
psychology	Year 1	Yes	No
BIT	Year 1	No	No
BENL	Year 1	Yes	No
Islamic Education	Year 1	No	No
BCS	Year 1	No	No
Islamic education	Year 2	No	No
Engineering	Year 2	No	No
Pendidikan islam	Year 2	Yes	No
Human Resources	Year 2	No	No
KENMS	Year 2	No	No
BIT	Year 2	No	No
Marine science	Year 2	Yes	No
KOE	Year 2	No	No
Engineering	Year 2	Yes	No
Business Administ...	Year 2	No	No
BCS	Year 2	No	No
BCS	Year 2	Yes	Yes

Post-df transformation orderBy() part 2

```
ordered_df = sentence_df.orderBy("Study Year").show(n=1000)
```

```
|    Pendidikan islam|    Year 2|          Yes|                  No|
|     Human Resources|    Year 2|           No|                  No|
|               KENMS|    Year 2|           No|                  No|
|                 BIT|    Year 2|           No|                  No|
|      Marine science|    Year 2|          Yes|                  No|
|                 KOE|    Year 2|           No|                  No|
|         Engineering|    Year 2|          Yes|                  No|
|Business Administ...|    Year 2|           No|                  No|
|                 BCS|    Year 2|           No|                  No|
|                 BCS|    Year 2|          Yes|                 Yes|
|         Engineering|    Year 2|          Yes|                 Yes|
|          Usuluddin |    Year 2|          Yes|                  No|
|                TAASL|    Year 2|           No|                  No|
|                 BCS|    Year 2|          Yes|                  No|
|                 BCS|    Year 2|           No|                  No|
|     Human Sciences |    Year 2|           No|                  No|
|      Communication |    Year 2|          Yes|                  No|
|     Diploma Nursing|    Year 2|           No|                  No|
|   Pendidikan Islam |    Year 2|           No|                  No|
|         Engineering|    Year 2|           No|                  No|
|                 Koe|    Year 2|           No|                  No|
|                 KOE|    Year 2|          Yes|                  No|
|         Engineering|    Year 2|          Yes|                  No|
|  Biomedical science|    Year 2|           No|                  No|
|                Laws|    Year 3|          Yes|                  No|
|         Engineering|    Year 3|          Yes|                  No|
|                 BCS|    Year 3|           No|                  No|
|         Accounting |    Year 3|           No|                  No|
|                 BCS|    Year 3|          Yes|                  No|
|                 BIT|    Year 3|          Yes|                 Yes|
|                 Law|    Year 3|          Yes|                  No|
|                 BIT|    Year 3|           No|                  No|
|                 BCS|    Year 3|           No|                  No|
|  Biomedical science|    Year 3|           No|                  No|
|                 koe|    Year 3|          Yes|                  No|
|              Kirkhs|    Year 3|           No|                  No|
|                BENL|    Year 3|          Yes|                  No|
|                  IT|    Year 3|           No|                  No|
|                 KOE|    Year 3|           No|                  No|
|                MHSC|    Year 3|          Yes|                  No|
|                Laws|    Year 3|           No|                  No|
|                 BIT|    Year 3|          Yes|                  No|
|                 BIT|    Year 3|           No|                  No|
|       Biotechnology|    Year 3|           No|                  No|
|         Fiqh fatwa |    Year 3|           No|                  No|
|        DIPLOMA TESL|    Year 3|           No|                  No|
|                Fiqh|    Year 3|           No|                  No|
|            Nursing |    Year 3|          Yes|                  No|
|        Mathemathics|    Year 4|           No|                  No|
|                 ENM|    Year 4|          Yes|                  No|
|                 BCS|    Year 4|           No|                  No|
|         Engineering|    Year 4|           No|                  No|
|              Engine|    Year 4|           No|                  No|
|                 Kop|    Year 4|           No|                  No|
|         Engineering|    Year 4|           No|                  No|
|    Pendidikan Islam|    Year 4|           No|                  No|
+--------------------+----------+-------------+--------------------+
```

C. Reflection about manipulating data from a dataframe using SQL

Reflect on why you choose your dataset and understand its structure and available columns. Then, identify three questions you have in mind about the data and list them down.

I chose this dataset because mental health is as important as our physical health. As of year 2023, we are slowly recognizing that mental health is also part of our self and should also be taken care of. I want to analyze this dataset by MD Shariful Islaam from 2020 because I want to know how many students have poor mental health and are suffering from mental health issues, and if it is affecting their academic performance.

Three questions I have in mind are:

1. What is the percentage of students dealing with mental health issues seeked for treatment?
2. Does higher year level contribute to mental health issues of the students?
3. Which gender have more mental health issues?

After identifying your questions, kindly run an SQL query to help you answer these questions.

1. What is the percentage of students dealing with mental health issues seeked for treatment?

```
[64] df_with_count = df.withColumn("Yes_Count",
                                   (col("Has Depression") == "Yes").cast("integer") +
                                   (col("Has Anxiety") == "Yes").cast("integer") +
                                   (col("Has Panic attack") == "Yes").cast("integer"))

[65] filtered_df = df_with_count.filter(col("Yes_Count") >= 2)
```

I made a code where there is a counter for students who answered at least 2 Yeses in the column Has Depression, Has Anxiety, and Has Panic Attack. Those who only answered only 1 Yes will be filtered out. Then, a column is created with a name Yes Count.

```
[66] filtered_df.show(n=1000)
```

Timestamp	Gender	Age	Course	Study Year	CGPA	Marital status	Has Depression	Has Anxiety	Has Panic Attack	Specialist Treatment	Yes_Count
8/7/2020 12:02	Female	18	Engineering	Year 1	3.00 - 3.49	No	Yes	No	Yes	No	2
8/7/2020 12:05	Male	19	BIT	Year 1	3.00 - 3.49	No	Yes	Yes	Yes	No	3
8/7/2020 12:32	Female	23	Pendidikan islam	Year 2	3.50 - 4.00	Yes	Yes	No	Yes	No	2
8/7/2020 12:39	Male	18	Irkhs	Year 1	3.50 - 4.00	No	No	Yes	Yes	No	2
8/7/2020 12:52	Female	24	ENM	Year 4	3.00 - 3.49	Yes	Yes	Yes	Yes	No	3
8/7/2020 13:07	Female	18	Marine science	Year 2	3.50 - 4.00	Yes	Yes	Yes	Yes	No	3
8/7/2020 13:17	Female	23	BCS	Year 3	3.50 - 4.00	No	Yes	Yes	Yes	No	3
8/7/2020 13:41	Male	18	Engineering	Year 2	3.00 - 3.49	Yes	Yes	Yes	No	No	2
8/7/2020 13:58	Female	24	BIT	Year 3	3.50 - 4.00	Yes	Yes	Yes	Yes	Yes	3
8/7/2020 14:31	Male	18	BCS	Year 2	3.50 - 4.00	Yes	Yes	Yes	No	Yes	2
8/7/2020 14:41	Female	19	BIT	Year 1	3.00 - 3.49	No	Yes	Yes	Yes	No	3
8/7/2020 14:43	Female	18	Law	Year 3	3.00 - 3.49	No	Yes	Yes	No	No	2
8/7/2020 14:45	Female	19	BIT	Year 1	2.50 - 2.99	No	Yes	Yes	Yes	No	3
8/7/2020 14:56	Female	24	Engineering	Year 2	2.50 - 2.99	Yes	Yes	No	Yes	Yes	2
8/7/2020 15:09	Male	18	BCS	Year 1	3.50 - 4.00	No	No	Yes	Yes	No	2
8/7/2020 15:27	Female	23	ALA	Year 1	2.50 - 2.99	Yes	Yes	No	Yes	Yes	2
8/7/2020 15:48	Female	20	koe	Year 3	3.00 - 3.49	Yes	Yes	Yes	Yes	No	3
8/7/2020 15:57	Female	19	BCS	Year 1	3.50 - 4.00	No	Yes	No	Yes	Yes	2
8/7/2020 16:21	Female	20	BENL	Year 3	3.00 - 3.49	No	Yes	Yes	No	No	2
8/7/2020 17:50	Female	23	Econs	Year 1	3.50 - 4.00	No	Yes	Yes	No	No	2
8/7/2020 18:11	Male	19	MHSC	Year 3	3.00 - 3.49	Yes	Yes	No	Yes	No	2
9/7/2020 11:57	Female	24	KOE	Year 1	3.50 - 4.00	No	No	Yes	Yes	No	2
13/07/2020 10:11:26	Female	24	Communication	Year 2	3.50 - 4.00	Yes	Yes	Yes	Yes	No	3
13/07/2020 10:12:26	Female	19	Engineering	Year 1	3.00 - 3.49	No	Yes	Yes	No	No	2
13/07/2020 10:33:47	Female	18	psychology	Year 1	3.50 - 4.00	No	Yes	Yes	No	Yes	2
13/07/2020 11:46:13	Female	18	psychology	Year 1	3.50 - 4.00	No	Yes	Yes	Yes	No	3
13/07/2020 21:21:42	Male	18	Engineering	Year 2	3.00 - 3.49	No	Yes	Yes	No	No	2
13/07/2020 21:22:56	Female	19	Nursing	Year 3	3.50 - 4.00	Yes	Yes	No	Yes	No	2

This is the result. All these students are having more than 2 mental health issues. I will then count how many students answered Yes in the column Specialist Treatment. Then get the percentage from filtered_df.

There are a total of 28 students that have 2 or more mental health issues, and 6 students answered Yes in the column Specialist Treatment. **Therefore, 21.43% of students who are having 2 or more mental health issues sought for help from a specialist.**

The result are what I expected since in 2020, all students are doing online schooling because of COVID-19 which added to the stressors that they are already dealing. Still, a lot of them did not seek any from professionals which means that mental health still needs more awareness from people. This finding raises concerns about the barriers and challenges that may prevent students from seeking professional help for their mental health issues. There could be reasons including lack of awareness about available resources, stigma associated with seeking help, or limited access to mental health services.

2. Does higher year level contribute to mental health issues of the students?

```
[68] study_year_counts = filtered_df.groupBy("Study Year").count().orderBy("Study Year")
     study_year_counts.show()
```

Study Year	count
Year 1	13
Year 2	7
Year 3	7
Year 4	1

I made a code using **groupBy()** function to group different year level then I used **count()** function to count the Year 1 to Year 4. Then used **orderBy()** function to arrange it in ascending order.

From this result, it shows that a lot of Year 1 students are having 2 or more mental health issue while Year 4 are the lowest at 1. However, we need to put context to it so I also counted the overall responses from each year level.

```
study_year_counts = df.groupBy("Study Year").count().orderBy("Study Year")
study_year_counts.show()
```

```
+----------+-----+
|Study Year|count|
+----------+-----+
|    Year 1|   43|
|    Year 2|   26|
|    Year 3|   24|
|    Year 4|    8|
+----------+-----+
```

Year 1 students are the majority while only 8 come from Year 4. **From this data, It is safe to say that younger students are more aware about their mental health and are more open on admitting that they are having mental health issues. But, this fails to give enough data to conclude that higher year level means more mental health issue because of lack of sample size from year level 4, 3, and 2.**

3. Which gender have more mental health issues?

```
[73] panic_attack_count = df.filter(col("Has Panic Attack") == "Yes").groupBy("Gender").count()
     depression_count = df.filter(col("Has Depression") == "Yes").groupBy("Gender").count()
     anxiety_count = df.filter(col("Has Anxiety") == "Yes").groupBy("Gender").count()
```

I made a code using **filter()** function to gather data from 3 columns that contains mental health issues and then used groupBy() function to group it based on gender Male or Female. Then **count()** to count the filtered data.

```
Panic Attack count by gender:
+------+-----+
|Gender|count|
+------+-----+
|Female|   25|
|  Male|    8|
+------+-----+

Depression count by gender:
+------+-----+
|Gender|count|
+------+-----+
|Female|   29|
|  Male|    6|
+------+-----+

Anxiety count by gender:
+------+-----+
|Gender|count|
+------+-----+
|Female|   24|
|  Male|   10|
+------+-----+
```

The result is that 75.75% that are having panic attack are female. 82% that are having depression are female. And 70.60% that are having anxiety are female. This shows that female are having more mental health issue in all 3 topics.

V. Synthesis and Moving Forward

A. What are the things that you learned about Big Data?

As a 3rd year student having its first subject about Big Data as an elective, I learned how these numbers can convey messages. And in the right hands, can either be helpful or dangerous. My knowledge in SQL improved especially in PySpark. The DataCamp course is very helpful in conducting this laboratory. Having the idea on how to use the functions made my life easier in relearning how it works when I started doing the laboratory. I also learned that not all functions can apply on a data set. sortByKey() and groupByKey() cannot be applied to the dataset that I imported. I also learned about data analysis along the way since we are required to give a reflection and to say something about the findings that we have. Learning these things will surely help me grasp the concept of Big Data as a whole.

B. How do you plan to use the things you learned moving forward?

I am planning to use the things that I learned in improving more. Building the base knowledge about Big Data is just the start. But right now, I will start some Big Data projects online and in Kaggle in order to practice and enhance my skills. Since I am also interested in business part of the analytics, I will focus on improving my data-driven and data-informed decision making. In conclusion, the goal is to continue exploring and deepen my understanding of Big Data and then stick to the part and aspect of Big Data that excites me the most.

C. What are your areas for improvement?

Doing this laboratory made me realize that making a conclusion and insight can sometimes be unpredictable. The data does not give us the insight that we assumed at the start. Formulating insights and conclusions is one of the area that I can still improve. Another area for improvement is the programming. As this language is new to me, there is a large room of improvement in terms of being creative on how to code and use libraries and frameworks such as PySpark.