Pisa Data Exploratory Analysis(Part 1) (1)

November 25, 2022

1 (PISA DATA 2012)

1.1 by (Oluwashina Dedenuola)

1.2 Introduction

An international study known as PISA (Programme for International Student Assessment) started in the year 2000. By assessing the abilities and knowledge of 15–16-year-old students in participating nations/economies, it seeks to assess educational systems around the world.

The fifth survey for the program is PISA 2012. It evaluated the skills of 15-year-olds in 65 nations and economies in reading, mathematics, and science (with a focus on mathematics). Around 85 000 students participated in an optional test of creative problem-solving in 44 of those countries and economies, and students' financial literacy was evaluated in 18 of those nations and economies.

28 million 15-year-olds worldwide were represented by the approximately 510 000 students between the ages of 15 years, 3 months and 16 years, 2 months who took part in PISA 2012.

1.3 Gathering Data

The data is readily availablein a csv file, although very large but is readable using the pandas function.

```
In [1]: # import all packages and set plots to be embedded inline
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sb

        %matplotlib inline
        import warnings
        warnings.simplefilter("ignore")
```

```
In [2]: #load the csv file
        pisa_data = pd.read_csv('pisa2012.csv',encoding='unicode_escape')
```

```
In [3]: #checking the first set of columns across the dataset.
        pisa_data.head(5)
```

```
Out[3]:    Unnamed: 0      CNT  SUBNATIO  STRATUM      OECD       NC  SCHOOLID  \
        0           1  Albania     80000  ALB0006  Non-OECD  Albania         1
        1           2  Albania     80000  ALB0006  Non-OECD  Albania         1
        2           3  Albania     80000  ALB0006  Non-OECD  Albania         1
        3           4  Albania     80000  ALB0006  Non-OECD  Albania         1
        4           5  Albania     80000  ALB0006  Non-OECD  Albania         1

           STIDSTD  ST01Q01  ST02Q01  ...  W_FSTR75  W_FSTR76 W_FSTR77 W_FSTR78  \
        0        1       10      1.0  ...   13.7954   13.9235  13.1249  13.1249
        1        2       10      1.0  ...   13.7954   13.9235  13.1249  13.1249
        2        3        9      1.0  ...   12.7307   12.7307  12.7307  12.7307
        3        4        9      1.0  ...   12.7307   12.7307  12.7307  12.7307
        4        5        9      1.0  ...   12.7307   12.7307  12.7307  12.7307

           W_FSTR79 W_FSTR80 WVARSTRR VAR_UNIT SENWGT_STU  VER_STU
        0    4.3389  13.0829       19        1     0.2098  22NOV13
        1    4.3389  13.0829       19        1     0.2098  22NOV13
        2    4.2436  12.7307       19        1     0.1999  22NOV13
        3    4.2436  12.7307       19        1     0.1999  22NOV13
        4    4.2436  12.7307       19        1     0.1999  22NOV13

        [5 rows x 636 columns]
```

As it is important to know the types of variables in each of the columns, I checked the dictionary attached to the project and first chose columns of interest based on the name of the column. Then I had a check into if the data in the column is qualitative or quantitative.

STRUCTURE OF THE DATASET

```
In [4]: pisa_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 485490 entries, 0 to 485489
Columns: 636 entries, Unnamed: 0 to VER_STU
dtypes: float64(250), int64(18), object(368)
memory usage: 2.3+ GB
```

The structure of the dataset shows Six Hundred and Thirt Six(636) columns and Four Hundred and Eighty Five Thousand, Four Hundred and Ninety Columns with data types which consists of floats, integers and objects.

NOTE: To start with, I want to ensure that the majority of my analysis has a high level of accuracy and as such, there would be alot of instances where values might need to be dropped. I would have implemented the fillna or using mean or mode to fill up null values but this is only acceptable statistically where the null values aren't more than 5-10% of a column in a dataset. The dataset has very high amount of valuable information and it is best to work with the information that is most accurate and expresses the honesty and sincerity of the subsets of the entire student population.

MAIN FEATURES OF THE DATASET

Hence, I will be doing some form of cleaning and visualization intermittently while trying to utilize the clean data and also reducing as comparison are being made with two or more datasets with missing values.

Also, I am keenly interested in the type of variables I am dealing with, whether it is categorical(nominal or ordinal) or numeric(discrete or continuous). So I will be checking the dictionary for selected code meaning and check what type of variables are present.

I decided to use the value count function in pandas to check the type of data in some of the columns as the data wasn't totally visualized due to its volume.

```
In [5]: pisa_data.value_counts('ST55Q02')

Out[5]: ST55Q02
        I do not attend <out-of-school time lessons> in this subject    169786
        Less than 2 hours a week                                         59714
        2 or more but less than 4 hours a week                           46296
        4 or more but less than 6 hours a week                           22066
        6 or more hours a week                                           10309
        dtype: int64

In [6]: pisa_data.ST55Q02.value_counts()

Out[6]: I do not attend <out-of-school time lessons> in this subject    169786
        Less than 2 hours a week                                         59714
        2 or more but less than 4 hours a week                           46296
        4 or more but less than 6 hours a week                           22066
        6 or more hours a week                                           10309
        Name: ST55Q02, dtype: int64

In [7]: pisa_data.ST55Q03.value_counts()

Out[7]: I do not attend <out-of-school time lessons> in this subject    201614
        Less than 2 hours a week                                         48468
        2 or more but less than 4 hours a week                           33571
        4 or more but less than 6 hours a week                           14679
        6 or more hours a week                                            7758
        Name: ST55Q03, dtype: int64

In [8]: pisa_data.ST55Q01.value_counts()

Out[8]: I do not attend <out-of-school time lessons> in this subject    206705
        Less than 2 hours a week                                         46825
        2 or more but less than 4 hours a week                           32313
        4 or more but less than 6 hours a week                           14860
        6 or more hours a week                                            7058
        Name: ST55Q01, dtype: int64

In [9]: pisa_data.ST86Q01.value_counts()
```

```
Out[9]: Agree                180918
        Strongly agree        81459
        Disagree              43810
        Strongly disagree      7036
        Name: ST86Q01, dtype: int64

In [10]: pisa_data.ST03Q02.value_counts()

Out[10]: 1996    451476
         1997     34014
         Name: ST03Q02, dtype: int64

In [11]: pisa_data.ST04Q01.value_counts()

Out[11]: Female    245064
         Male      240426
         Name: ST04Q01, dtype: int64

In [12]: pisa_data.ST08Q01.value_counts()

Out[12]: None                     306065
         One or two times         124380
         Three or four times       29817
         Five or more times        18881
         Name: ST08Q01, dtype: int64

In [13]:  pisa_data.HOMSCH.value_counts()

Out[13]: -0.4477    21172
          0.0526    20992
         -0.0911    20223
         -0.2549    19149
         -0.6852    18737
                    ...
          0.2306        1
         -0.6255        1
          1.4309        1
         -0.5384        1
          1.3762        1
         Name: HOMSCH, Length: 685, dtype: int64

In [14]:  pisa_data.ST85Q01.value_counts()

Out[14]: Agree                158751
         Strongly agree       107494
         Disagree              36861
         Strongly disagree      9368
         Name: ST85Q01, dtype: int64

In [15]: pisa_data.ST81Q01.info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 485490 entries, 0 to 485489
Series name: ST81Q01
Non-Null Count   Dtype
--------------   -----
313982 non-null  object
dtypes: object(1)
memory usage: 3.7+ MB
```

```
In [16]: pisa_data.MMINS.value_counts()
```

```
Out[16]: 180.0     43751
         200.0     30096
         225.0     25487
         240.0     23729
         250.0     17390
                   ...
         1395.0        1
         2040.0        1
         1140.0        1
         1700.0        1
         1218.0        1
         Name: MMINS, Length: 397, dtype: int64
```

```
In [17]:  pisa_data.OCOD1.value_counts()
```

```
Out[17]: Housewife                                                    74358
         Missing                                                      27044
         Shop sales assistants                                        13124
         Primary school teachers                                      10320
         Secretaries (general)                                         9869
                                                                       ...
         Insulation workers                                               3
         Shotfirers and blasters                                          3
         Market-oriented skilled forestry, fishery and hunting worker     3
         Underwater divers                                                2
         Drivers of animal-drawn vehicles and machinery                   1
         Name: OCOD1, Length: 588, dtype: int64
```

```
In [18]: pisa_data.OCOD1.info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 485490 entries, 0 to 485489
Series name: OCOD1
Non-Null Count   Dtype
--------------   -----
483887 non-null  object
dtypes: object(1)
```

```
memory usage: 3.7+ MB
```

```
In [19]:   pisa_data.OUTHOURS.value_counts()
```

```
Out[19]: 6.0      22290
         4.0      20770
         5.0      20567
         3.0      19935
         2.0      18701
                  ...
         153.0        1
         151.0        1
         116.0        1
         124.0        1
         142.0        1
         Name: OUTHOURS, Length: 143, dtype: int64
```

```
In [20]: pisa_data.OUTHOURS.info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 485490 entries, 0 to 485489
Series name: OUTHOURS
Non-Null Count   Dtype
--------------   -----
308799 non-null  float64
dtypes: float64(1)
memory usage: 3.7 MB
```

```
In [21]: pisa_data.OPENPS.value_counts()
```

```
Out[21]:  0.0521    28434
          0.4639    28114
         -0.5433    28098
          0.2542    27711
         -0.1465    27686
                    ...
         -1.0985        1
         -2.5837        1
         -1.0452        1
         -2.1249        1
         -3.4543        1
         Name: OPENPS, Length: 274, dtype: int64
```

```
In [22]: pisa_data.INTMAT.value_counts()
```

```
Out[22]: -0.34    43119
          0.91    41948
```

```
 0.00    35614
 0.30    33344
 0.58    32430
         ...
-1.02        3
-0.86        3
-1.06        2
-1.68        1
-1.71        1
Name: INTMAT, Length: 96, dtype: int64
```

```
In [23]: pisa_data.ST42Q01.value_counts()
```

```
Out[23]: Agree                134489
         Disagree              83359
         Strongly agree        68696
         Strongly disagree     27311
         Name: ST42Q01, dtype: int64
```

```
In [24]: pisa_data.USEMATH.info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 485490 entries, 0 to 485489
Series name: USEMATH
Non-Null Count   Dtype
--------------   -----
290260 non-null  float64
dtypes: float64(1)
memory usage: 3.7 MB
```

MAIN FEATURES OF THE DATASET (Continued)

As the aim of the pisa project is to understand how well students have learnt and understood their curriculum, the focus is on their subject and scores and what possible relationship exists between the different data within the dataset as well. Also, I am focused on finding possible linear relationships between how well they performed as it relates to their gender, country and also the distribution between the gender based on their scores.

Questions to be asked and visualizations that proceeds them are:

1. What is the distribution of the students like based on their Gender?
2. Which Top ten(10) countries has the highest number of participants?
3. What's the Top ten(10) occupation of the mothers of the students?
4. What's the Top ten(10) occupation of the Fathers of the students?
5. How interested are the students as it relates to Maths?
6. What is the Gender distribution of the students that enjoy Maths?

7. What is the Gender distribution of the students that enjoy Maths Lesson?
8. What is the Gender distribution of the students that have interest in Maths?
9. Relationship of the Students between Problem Solving and Maths Score?
10. What is the correlation between Students performance in Reading and Maths exams as well as their Gender Distribution?
11. What is the correlation between Students performance in Science and Maths exams as well as their Gender Distribution?
12. What is the correlation between Students performance in Science and Reading exams as well as their Gender Distribution?
13. Does a student interest in Maths influence their Maths score?
14. Are there similar trends across the entire student scores?

Firstly, I'll make a copy of the dataset

```
In [25]: pisa_data = pisa_data.copy()
```

I'll select the columns of interest and drop the other columns. I'll drop the columns I don't need by creating a list out of the 636 columns.

```
In [26]: pisa_data = pisa_data[['STIDSTD','AGE','ST03Q02','ST04Q01','ICTRES','INTMAT','OCOD1','O
         'PV1MATH','PV2MATH','PV3MATH','PV4MATH','PV5MATH','PV1READ','PV2READ','PV3READ', 'PV4RE
```

```
In [27]: pisa_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 485490 entries, 0 to 485489
Data columns (total 29 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   STIDSTD  485490 non-null  int64
 1   AGE      485374 non-null  float64
 2   ST03Q02  485490 non-null  int64
 3   ST04Q01  485490 non-null  object
 4   ICTRES   477754 non-null  float64
 5   INTMAT   316708 non-null  float64
 6   OCOD1    483887 non-null  object
 7   OCOD2    482936 non-null  object
 8   OPENPS   312766 non-null  float64
 9   PV1MATH  485490 non-null  float64
 10  PV2MATH  485490 non-null  float64
 11  PV3MATH  485490 non-null  float64
 12  PV4MATH  485490 non-null  float64
 13  PV5MATH  485490 non-null  float64
 14  PV1READ  485490 non-null  float64
```

```
 15  PV2READ  485490 non-null  float64
 16  PV3READ  485490 non-null  float64
 17  PV4READ  485490 non-null  float64
 18  PV5READ  485490 non-null  float64
 19  PV1SCIE  485490 non-null  float64
 20  PV2SCIE  485490 non-null  float64
 21  PV3SCIE  485490 non-null  float64
 22  PV4SCIE  485490 non-null  float64
 23  PV5SCIE  485490 non-null  float64
 24  ST29Q01  315911 non-null  object
 25  ST29Q03  314928 non-null  object
 26  ST29Q04  314737 non-null  object
 27  ST29Q06  314746 non-null  object
 28  CNT      485490 non-null  object
dtypes: float64(19), int64(2), object(8)
memory usage: 107.4+ MB
```

```
In [28]: #IN ORDER TO OBTAIN THE AVERAGE SCORE OF THE DIFFERENT SUBJECTS, I'LL ADD THEIR SCORES

         pisa_data['Std Maths Score'] = (pisa_data.loc[:, ['PV1MATH', 'PV2MATH','PV3MATH','PV4MA
         pisa_data['Std Reading Score'] = (pisa_data.loc[:, ['PV1READ', 'PV2READ','PV3READ','PV4
         pisa_data['Std Science Score'] = (pisa_data.loc[:, ['PV1SCIE', 'PV2SCIE','PV3SCIE','PV4
```

```
In [29]: #DROP THE COLUMNS MATHS, SCIENCE AND READING COLUMNS AND LEAVE THE NEWLY ADDED COLUMN

         pisa_data.drop(['PV1MATH','PV2MATH','PV3MATH','PV4MATH','PV5MATH', 'PV1READ','PV2READ',
```

```
In [30]: #CHECKING IF THE COLUMNS HAS BEEN DROPPED.
         pisa_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 485490 entries, 0 to 485489
Data columns (total 17 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   STIDSTD            485490 non-null  int64
 1   AGE                485374 non-null  float64
 2   ST03Q02            485490 non-null  int64
 3   ST04Q01            485490 non-null  object
 4   ICTRES             477754 non-null  float64
 5   INTMAT             316708 non-null  float64
 6   OCOD1              483887 non-null  object
 7   OCOD2              482936 non-null  object
 8   OPENPS             312766 non-null  float64
 9   ST29Q01            315911 non-null  object
 10  ST29Q03            314928 non-null  object
 11  ST29Q04            314737 non-null  object
 12  ST29Q06            314746 non-null  object
```

```
 13  CNT                485490 non-null  object
 14  Std Maths Score    485490 non-null  float64
 15  Std Reading Score  485490 non-null  float64
 16  Std Science Score  485490 non-null  float64
dtypes: float64(7), int64(2), object(8)
memory usage: 63.0+ MB
```

In [31]:
```
#I would have to transpose to be able to view the columns properly.
pisa_data.head().transpose()
```

Out[31]:

	0	1
STIDSTD	1	2
AGE	16.17	16.17
ST03Q02	1996	1996
ST04Q01	Female	Female
ICTRES	-3.16	1.15
INTMAT	0.91	0.0
OCOD1	Building architects	Tailors, dressmakers, furriers and hatters
OCOD2	Primary school teachers	Building construction labourers
OPENPS	0.0521	-0.9492
ST29Q01	Agree	Disagree
ST29Q03	Agree	Disagree
ST29Q04	Agree	Disagree
ST29Q06	Agree	Agree
CNT	Albania	Albania
Std Maths Score	366.18634	470.56396
Std Reading Score	261.01424	384.68832
Std Science Score	371.91348	478.12382

	2	\
STIDSTD	3	
AGE	15.58	
ST03Q02	1996	
ST04Q01	Female	
ICTRES	-0.4	
INTMAT	1.23	
OCOD1	Housewife	
OCOD2	Bricklayers and related workers	
OPENPS	0.9383	
ST29Q01	Agree	
ST29Q03	Agree	
ST29Q04	Agree	
ST29Q06	Strongly agree	
CNT	Albania	
Std Maths Score	505.53824	
Std Reading Score	405.18154	
Std Science Score	486.60946	

		3	\
STIDSTD		4	
AGE		15.67	
ST03Q02		1996	
ST04Q01		Female	
ICTRES		-0.4	
INTMAT		NaN	
OCOD1		Housewife	
OCOD2	Cleaners and helpers in offices, hotels and ot...		
OPENPS		NaN	
ST29Q01		NaN	
ST29Q03		NaN	
ST29Q04		NaN	
ST29Q06		NaN	
CNT		Albania	
Std Maths Score		449.45476	
Std Reading Score		477.46376	
Std Science Score		453.9724	

	4
STIDSTD	5
AGE	15.5
ST03Q02	1996
ST04Q01	Female
ICTRES	0.24
INTMAT	0.3
OCOD1	Housewife
OCOD2	Economists

```
OPENPS                     1.2387
ST29Q01                 Disagree
ST29Q03                 Disagree
ST29Q04                 Disagree
ST29Q06           Strongly agree
CNT                      Albania
Std Maths Score        385.50398
Std Reading Score       256.0101
Std Science Score      367.15778
```

Another observation I have interest in has to do with the impact of anxiety on the performance of students as it relates to Mathematics.

```
In [32]: #Checking the shape of the data
         pisa_data.shape
```

```
Out[32]: (485490, 17)
```

```
In [33]: #USE THE MEAN METHOD TO FILL THE NAN SPACE IN THE AGE COLUMN
         pisa_data.groupby('AGE').mean()
```

```
Out[33]:            STIDSTD      ST03Q02    ICTRES    INTMAT    OPENPS  \
         AGE
         15.17  2418.000000  1997.000000 -0.470000 -0.660000 -2.094000
         15.25  6882.237193  1996.303187 -0.375051  0.199221 -0.058975
         15.33  5703.734923  1996.254532 -0.312708  0.204914 -0.022491
         15.42  6061.875856  1996.166283 -0.350864  0.216946  0.020288
         15.50  6167.231094  1996.149860 -0.369128  0.229704  0.018638
         15.58  6265.557738  1996.137458 -0.381140  0.226635  0.028528
         15.67  6175.292636  1996.042004 -0.354541  0.219992  0.040730
         15.75  6107.479719  1996.038666 -0.347944  0.200466  0.029678
         15.83  6172.407082  1996.026931 -0.342810  0.205163  0.044370
         15.92  6054.832709  1996.000414 -0.333388  0.206071  0.054851
         16.00  6102.177252  1996.000000 -0.341600  0.211223  0.060399
         16.08  6184.362603  1996.000000 -0.341353  0.210280  0.050431
         16.17  6282.967828  1996.000000 -0.353115  0.206516  0.065971
         16.25  5589.130571  1996.000000 -0.335809  0.204942  0.073945
         16.33  6881.254836  1996.000000 -0.467345  0.224800  0.149616

                Std Maths Score  Std Reading Score  Std Science Score
         AGE
         15.17       338.923520         375.156600         395.505400
         15.25       464.674602         469.977145         472.328336
         15.33       471.847503         473.127493         477.482047
         15.42       465.168331         467.643772         471.873831
         15.50       466.611744         468.829560         472.997031
         15.58       465.924751         468.773940         472.231948
         15.67       468.076500         470.368384         474.121578
         15.75       469.069550         471.585300         474.970162
```

```
          15.83       469.620003         472.189075         475.830182
          15.92       472.389195         474.982916         478.519362
          16.00       472.140566         474.231229         478.301272
          16.08       474.239685         476.214738         480.097149
          16.17       474.296844         476.557746         480.300278
          16.25       474.883250         474.959454         479.843280
          16.33       454.972127         461.334054         463.145229
```

```
In [34]: pisa_data.groupby(['AGE'], as_index=False)['AGE'].mean()
```

```
Out[34]:       AGE
          0   15.17
          1   15.25
          2   15.33
          3   15.42
          4   15.50
          5   15.58
          6   15.67
          7   15.75
          8   15.83
          9   15.92
          10  16.00
          11  16.08
          12  16.17
          13  16.25
          14  16.33
```

```
In [35]: pisa_data['AGE'] = pisa_data['AGE'].fillna(pisa_data['AGE'].mean())
```

```
In [36]: pisa_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 485490 entries, 0 to 485489
Data columns (total 17 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   STIDSTD            485490 non-null  int64
 1   AGE                485490 non-null  float64
 2   ST03Q02            485490 non-null  int64
 3   ST04Q01            485490 non-null  object
 4   ICTRES             477754 non-null  float64
 5   INTMAT             316708 non-null  float64
 6   OCOD1              483887 non-null  object
 7   OCOD2              482936 non-null  object
 8   OPENPS             312766 non-null  float64
 9   ST29Q01            315911 non-null  object
 10  ST29Q03            314928 non-null  object
 11  ST29Q04            314737 non-null  object
 12  ST29Q06            314746 non-null  object
```

```
 13  CNT                485490 non-null  object
 14  Std Maths Score    485490 non-null  float64
 15  Std Reading Score  485490 non-null  float64
 16  Std Science Score  485490 non-null  float64
dtypes: float64(7), int64(2), object(8)
memory usage: 63.0+ MB
```

```
In [37]: pisa_data.value_counts('ICTRES')
```

```
Out[37]: ICTRES
          0.24    122169
         -0.40    112592
          1.15     93383
         -1.13     68455
         -3.16     41241
         -1.99     19944
         -0.80      3872
          0.07      3103
          1.01      2989
         -2.47      1497
         -2.94      1412
          0.26      1405
         -1.02      1305
         -1.11      1201
          0.03       523
         -1.91       489
          0.99       485
         -0.47       328
         -3.12       300
         -0.83       281
         -1.45       197
          0.20       180
         -2.96       140
          1.12       127
          0.21        58
         -1.39        42
         -2.42        36
         dtype: int64
```

```
In [38]: # Filling up the null spaces in the ICT Resources with mean values.
         pisa_data['ICTRES'] = pisa_data['ICTRES'].fillna(pisa_data['ICTRES'].mean())
```

```
In [39]: #checking to see the null values in the ICT Resources has been removed.
         pisa_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 485490 entries, 0 to 485489
Data columns (total 17 columns):
```

```
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   STIDSTD            485490 non-null  int64
 1   AGE                485490 non-null  float64
 2   ST03Q02            485490 non-null  int64
 3   ST04Q01            485490 non-null  object
 4   ICTRES             485490 non-null  float64
 5   INTMAT             316708 non-null  float64
 6   OCOD1              483887 non-null  object
 7   OCOD2              482936 non-null  object
 8   OPENPS             312766 non-null  float64
 9   ST29Q01            315911 non-null  object
 10  ST29Q03            314928 non-null  object
 11  ST29Q04            314737 non-null  object
 12  ST29Q06            314746 non-null  object
 13  CNT                485490 non-null  object
 14  Std Maths Score    485490 non-null  float64
 15  Std Reading Score  485490 non-null  float64
 16  Std Science Score  485490 non-null  float64
dtypes: float64(7), int64(2), object(8)
memory usage: 63.0+ MB
```

I'll be renaming the columns header except for the Anxiety columns(ST42Q01 - ST42Q10) and the Interest of students in Maths column(ST29Q01-ST29Q06) which I'll merge later after removing their respective null values.

```
In [40]: #checking null values
         pisa_data.isna().sum()

Out[40]: STIDSTD                  0
         AGE                      0
         ST03Q02                  0
         ST04Q01                  0
         ICTRES                   0
         INTMAT              168782
         OCOD1                 1603
         OCOD2                 2554
         OPENPS              172724
         ST29Q01             169579
         ST29Q03             170562
         ST29Q04             170753
         ST29Q06             170744
         CNT                      0
         Std Maths Score          0
         Std Reading Score        0
         Std Science Score        0
         dtype: int64
```

```
In [41]: #I'll like to rename the other columns heading for better understanding of what the col
         pisa_data.rename(columns = {'ST04Q01':'Gender','ICTRES':'ICT_resources','ST03Q02':'Birt

In [42]: #checking null values
         pisa_data.isna().sum()

Out[42]: STIDSTD                      0
         AGE                          0
         Birth_Year                   0
         Gender                       0
         ICT_resources                0
         Math_interest           168782
         Mother_Occupa             1603
         Father_Occupa             2554
         Openess_Problem_Solving  172724
         Enjoy_Reading_Maths      169579
         Enjoy_Maths_Lesson       170562
         Enjoy_Maths              170753
         Interest_in_Maths        170744
         Country                      0
         Std Maths Score              0
         Std Reading Score            0
         Std Science Score            0
         dtype: int64

In [43]: # Checking random values in the dataset
         pisa_data.sample(30)

Out[43]:         STIDSTD    AGE  Birth_Year  Gender  ICT_resources  Math_interest  \
         116391     2293  16.17        1996  Female          -3.16           0.58
         340348    24646  15.58        1996  Female          -1.13            NaN
         301800      997  15.58        1996  Female           0.24            NaN
         317680     1978  15.83        1996    Male          -1.13           0.58
         473416     3179  15.42        1997    Male           1.15           0.58
         276814     5152  16.25        1996  Female           1.15          -0.34
         242378     1789  15.83        1996  Female           1.15          -0.34
         132780     5007  16.00        1996    Male           0.24           0.91
         290104     5053  15.75        1996  Female          -1.13           0.00
         166822    21240  15.92        1996  Female           1.15          -0.66
         447050     3456  15.67        1996    Male           1.15            NaN
         210297     3397  16.17        1996    Male           0.24           0.58
         251517    10928  15.75        1996  Female           0.24            NaN
         450301     1971  16.25        1996    Male           1.15          -0.95
         456805     2429  16.00        1996  Female          -1.99            NaN
         461922      940  16.17        1996  Female          -0.40            NaN
         304263     3460  15.67        1996    Male           0.24           0.00
         365071     1162  15.33        1996    Male           1.15           0.00
         1583       1584  16.00        1996    Male          -1.13            NaN
         358830     4578  15.75        1996  Female          -1.13            NaN
```

17554	1312	15.33	1997	Female	1.01	-1.27
53324	3341	15.92	1996	Male	-0.40	-1.78
132873	5100	15.58	1996	Female	1.15	0.30
426239	3464	15.92	1996	Male	-0.40	0.30
477075	1523	15.33	1997	Female	-0.40	-1.27
52162	2179	15.50	1996	Female	1.15	NaN
38112	1481	16.17	1996	Male	-1.13	0.58
225998	4610	16.25	1996	Male	-3.16	0.30
388147	4619	15.92	1996	Female	0.24	0.58
129552	1779	16.08	1996	Male	0.24	NaN

	Mother_Occupa \
116391	Kitchen helpers
340348	Housewife
301800	Secondary education teachers
317680	Chefs
473416	Cleaners and helpers in offices, hotels and ot...
276814	Housewife
242378	Housewife
132780	Personal care workers in health services not e...
290104	General office clerks
166822	Building and related electricians
447050	Nursing professionals
210297	Cashiers and ticket clerks
251517	Sales and purchasing agents and brokers
450301	Teaching professionals
456805	Field crop and vegetable growers
461922	Administrative and executive secretaries
304263	Primary school teachers
365071	Business and administration professionals
1583	Social beneficiary (unemployed, retired, sickn...
358830	Sales and marketing managers
17554	Advertising and public relations managers
53324	Cooks
132873	Nursing professionals
426239	Plant and machine operators, and assemblers
477075	Credit and loans officers
52162	Missing
38112	Missing
225998	Housewife
388147	Cleaners and helpers in offices, hotels and ot...
129552	Stock clerks

	Father_Occupa \
116391	Bartenders
340348	Security guards
301800	Sales and marketing managers
317680	Legal professionals

```
473416                                  Police officers
276814               Retail and wholesale trade managers
242378       Plant and machine operators, and assemblers
132780                   Chemical engineering technicians
290104                                 Finance managers
166822                  Building and related electricians
447050  Information and communications technology user...
210297     Building finishers and related trades workers
251517                                          Missing
450301                               Software developers
456805                   Field crop and vegetable growers
461922      Legal professionals not elsewhere classified
304263                      Customs and border inspectors
365071  Artistic, cultural and culinary associate prof...
1583    Social beneficiary (unemployed, retired, sickn...
358830          University and higher education teachers
17554                    Commercial sales representatives
53324   Social beneficiary (unemployed, retired, sickn...
132873                     Computer network professionals
426239                Electrical engineering technicians
477075                                 Security guards
52162                                           Missing
38112                                           Missing
225998                                          Missing
388147                                   Prison guards
129552                                          Waiters

        Openess_Problem_Solving Enjoy_Reading_Maths Enjoy_Maths_Lesson  \
116391                   0.4639               Agree           Disagree
340348                      NaN                 NaN                NaN
301800                      NaN                 NaN                NaN
317680                  -0.5433               Agree              Agree
473416                  -0.7446            Disagree              Agree
276814                   0.9383               Agree           Disagree
242378                   0.9383      Strongly agree  Strongly disagree
132780                   0.0521               Agree              Agree
290104                  -0.1465               Agree           Disagree
166822                  -0.9492            Disagree  Strongly disagree
447050                      NaN                 NaN                NaN
210297                  -0.5433            Disagree              Agree
251517                      NaN                 NaN                NaN
450301                   0.2542            Disagree  Strongly disagree
456805                      NaN                 NaN                NaN
461922                      NaN                 NaN                NaN
304263                   2.4465   Strongly disagree              Agree
365071                   2.4465            Disagree           Disagree
1583                        NaN                 NaN                NaN
358830                      NaN                 NaN                NaN
```

```
17554                  -0.5433   Strongly disagree  Strongly disagree
53324                   0.4639   Strongly disagree  Strongly disagree
132873                  0.2542            Disagree              Agree
426239                 -0.5433            Disagree              Agree
477075                  0.9383   Strongly disagree  Strongly disagree
52162                      NaN                 NaN                NaN
38112                  -0.7446               Agree              Agree
225998                 -0.1465               Agree           Disagree
388147                  0.0521            Disagree              Agree
129552                     NaN                 NaN                NaN

             Enjoy_Maths  Interest_in_Maths                   Country  \
116391             Agree              Agree                  Colombia
340348               NaN                NaN                    Mexico
301800               NaN                NaN                Luxembourg
317680          Disagree              Agree                    Mexico
473416             Agree              Agree                   Uruguay
276814  Strongly disagree           Disagree                    Jordan
242378  Strongly disagree           Disagree                     Italy
132780             Agree              Agree            Czech Republic
290104          Disagree           Disagree                Kazakhstan
166822          Disagree           Disagree                     Spain
447050               NaN                NaN                    Sweden
210297             Agree              Agree           Hong Kong-China
251517               NaN                NaN                     Italy
450301  Strongly disagree           Disagree            Chinese Taipei
456805               NaN                NaN                  Thailand
461922               NaN                NaN                   Tunisia
304263             Agree           Disagree                Luxembourg
365071          Disagree              Agree                    Norway
1583                 NaN                NaN                   Albania
358830               NaN                NaN                  Malaysia
17554   Strongly disagree           Disagree                 Argentina
53324   Strongly disagree  Strongly disagree                  Bulgaria
132873             Agree           Disagree            Czech Republic
426239          Disagree              Agree                 Singapore
477075  Strongly disagree           Disagree  United States of America
52162                NaN                NaN                  Bulgaria
38112           Disagree              Agree                   Austria
225998             Agree           Disagree                 Indonesia
388147             Agree              Agree                  Portugal
129552               NaN                NaN            Czech Republic

        Std Maths Score  Std Reading Score  Std Science Score
116391        422.34770          429.40798          465.44202
340348        465.26720          536.79870          500.50356
301800        475.31548          523.77206          504.23348
317680        354.11280          379.62254          292.18574
```

```
473416        386.12712          258.60902          414.99442
276814        460.43776          471.10926          492.20440
242378        386.20502          435.12700          398.95560
132780        572.29318          547.06868          597.76202
290104        479.52172          430.04342          440.63782
166822        516.67704          561.26346          505.25922
447050        596.90764          661.42606          648.67588
210297        432.39602          412.42210          417.51214
251517        448.67582          501.84910          458.54160
450301        779.02322          680.27174          699.49644
456805        386.51658          501.05478          443.99476
461922        463.00826          521.54800          451.08170
304263        601.42546          504.16462          625.27042
365071        508.26452          566.87674          546.38194
1583          408.01528          514.10872          491.64490
358830        422.73718          337.90348          393.73368
17554         457.16622          532.19172          488.19472
53324         386.90604          278.33688          445.95300
132873        588.65088          606.30088          579.76502
426239        526.72534          495.42342          536.59084
477075        563.64700          631.00392          590.02238
52162         352.86650          345.84658          362.68182
38112         537.16310          495.02242          470.01118
225998        424.60664          386.92024          390.28346
388147        531.47686          562.45496          523.44274
129552        586.08040          561.82446          626.76240
```

In [44]:
```
#Checking the Column for the Countries for errors.
pisa_data.value_counts('Country').head(30)
```

Out[44]:
```
Country
Mexico                  33806
Italy                   31073
Spain                   25313
Canada                  21544
Brazil                  19204
Australia               14481
United Kingdom          12659
United Arab Emirates    11500
Switzerland             11229
Qatar                   10966
Colombia                 9073
Finland                  8829
Belgium                  8597
Denmark                  7481
Jordan                   7038
Chile                    6856
Thailand                 6606
```

```
Japan                    6351
Chinese Taipei           6046
Peru                     6035
Slovenia                 5911
Argentina                5908
Kazakhstan               5808
Portugal                 5722
Indonesia                5622
Singapore                5546
Macao-China              5335
Czech Republic           5327
Uruguay                  5315
Bulgaria                 5282
dtype: int64
```

In [45]:
```
#Checking the Column for the Countries for errors.
pisa_data.value_counts('Country').tail(30)
```

Out[45]:
```
Country
Ireland                      5016
Croatia                      5008
Germany                      5001
United States of America     4978
Vietnam                      4959
Turkey                       4848
Hungary                      4810
Estonia                      4779
Austria                      4755
Montenegro                   4744
Albania                      4743
Sweden                       4736
Norway                       4686
Serbia                       4684
Slovak Republic              4678
Hong Kong-China              4670
Lithuania                    4618
France                       4613
Poland                       4607
Costa Rica                   4602
Netherlands                  4460
Tunisia                      4407
Latvia                       4306
New Zealand                  4291
Iceland                      3508
Florida (USA)                1896
Perm(Russian Federation)     1761
Massachusetts (USA)          1723
Connecticut (USA)            1697
```

```
Liechtenstein                293
dtype: int64
```

```
In [46]: #Renaming the states meant to be part of USA in the Country column to United States of
         pisa_data.Country = pisa_data.Country.replace({'Florida (USA)': 'United States of Ameri
```

```
In [47]: pisa_data.value_counts('Country').head(15)
```

```
Out[47]: Country
         Mexico                      33806
         Italy                       31073
         Spain                       25313
         Canada                      21544
         Brazil                      19204
         Australia                   14481
         United Kingdom              12659
         United Arab Emirates        11500
         Switzerland                 11229
         Qatar                       10966
         United States of America    10294
         Colombia                     9073
         Finland                      8829
         Belgium                      8597
         Denmark                      7481
         dtype: int64
```

I'll be using the Transpose feature which would help visualize the columns better.

```
In [48]: pisa_data.sample(40).transpose()
```

```
Out[48]:                                              347888  \
         STIDSTD                                       32186
         AGE                                           16.17
         Birth_Year                                     1996
         Gender                                       Female
         ICT_resources                                 -1.13
         Math_interest                                  0.91
         Mother_Occupa                           Housewife
         Father_Occupa           Car, taxi and van drivers
         Openess_Problem_Solving                      0.9383
         Enjoy_Reading_Maths                          Agree
         Enjoy_Maths_Lesson                           Agree
         Enjoy_Maths                                  Agree
         Interest_in_Maths                            Agree
         Country                                     Mexico
         Std Maths Score                          463.47562
         Std Reading Score                        516.14668
         Std Science Score                        471.40992
```

	480663	\	
STIDSTD	133		
AGE	16.08		
Birth_Year	1996		
Gender	Female		
ICT_resources	-3.16		
Math_interest	0.91		
Mother_Occupa	Shop keepers		
Father_Occupa	Building construction labourers		
Openess_Problem_Solving	-1.5946		
Enjoy_Reading_Maths	Agree		
Enjoy_Maths_Lesson	Agree		
Enjoy_Maths	Agree		
Interest_in_Maths	Agree		
Country	Vietnam		
Std Maths Score	449.53264		
Std Reading Score	520.67422		
Std Science Score	533.32712		

	357926	\	
STIDSTD	3674		
AGE	15.67		
Birth_Year	1996		
Gender	Female		
ICT_resources	1.15		
Math_interest	-0.34		
Mother_Occupa	Bank tellers and related clerks		
Father_Occupa	Information and communications technology serv...		
Openess_Problem_Solving	-1.158		
Enjoy_Reading_Maths	Disagree		
Enjoy_Maths_Lesson	Disagree		
Enjoy_Maths	Disagree		
Interest_in_Maths	Disagree		
Country	Malaysia		
Std Maths Score	478.04176		
Std Reading Score	466.97888		
Std Science Score	418.25812		

	181063	\	
STIDSTD	5389		
AGE	15.75		
Birth_Year	1996		
Gender	Female		
ICT_resources	-1.13		
Math_interest	-1.78		
Mother_Occupa	Kitchen helpers		
Father_Occupa	Car, taxi and van drivers		
Openess_Problem_Solving	-0.1465		

Enjoy_Reading_Maths	Strongly disagree			
Enjoy_Maths_Lesson	Strongly disagree			
Enjoy_Maths	Strongly disagree			
Interest_in_Maths	Strongly disagree			
Country	Finland			
Std Maths Score	375.4557			
Std Reading Score	423.92726			
Std Science Score	461.15256			

				\
			407280	\
STIDSTD			126	
AGE			15.92	
Birth_Year			1996	
Gender			Male	
ICT_resources			-0.4	
Math_interest			NaN	
Mother_Occupa	Bleaching, dyeing and fabric cleaning machine ...			
Father_Occupa	Messengers, package deliverers and luggage por...			
Openess_Problem_Solving			NaN	
Enjoy_Reading_Maths			NaN	
Enjoy_Maths_Lesson			NaN	
Enjoy_Maths			NaN	
Interest_in_Maths			NaN	
Country			United States of America	
Std Maths Score			437.537	
Std Reading Score			454.12324	
Std Science Score			413.96868	

				\
		240965	127465	\
STIDSTD		376	4294	
AGE		15.83	15.67	
Birth_Year		1996	1996	
Gender		Female	Female	
ICT_resources		-1.13	0.24	
Math_interest		NaN	0.91	
Mother_Occupa		Pharmacists	Secretaries (general)	
Father_Occupa	Administration professionals		Civil engineers	
Openess_Problem_Solving		NaN	0.4639	
Enjoy_Reading_Maths		NaN	Agree	
Enjoy_Maths_Lesson		NaN	Agree	
Enjoy_Maths		NaN	Agree	
Interest_in_Maths		NaN	Agree	
Country		Italy	Costa Rica	
Std Maths Score		368.44526	533.57998	
Std Reading Score		425.5953	573.97244	
Std Science Score		410.89146	559.43676	

		\
	157307	\

STIDSTD	11725			
AGE	15.42			
Birth_Year	1996			
Gender	Male			
ICT_resources	-1.13			
Math_interest	0.3			
Mother_Occupa	Beauticians and related workers			
Father_Occupa	House builders			
Openess_Problem_Solving	1.2387			
Enjoy_Reading_Maths	Disagree			
Enjoy_Maths_Lesson	Agree			
Enjoy_Maths	Disagree			
Interest_in_Maths	Agree			
Country	Spain			
Std Maths Score	509.9003			
Std Reading Score	505.04676			
Std Science Score	536.31106			

	369630	157604	...	\
STIDSTD	1035	12022	...	
AGE	15.33	15.58	...	
Birth_Year	1997	1996	...	
Gender	Male	Male	...	
ICT_resources	-0.4	-1.13	...	
Math_interest	-1.78	NaN	...	
Mother_Occupa	Other arts teachers	Cooks	...	
Father_Occupa	Receptionists (general)	Bartenders	...	
Openess_Problem_Solving	-0.5433	NaN	...	
Enjoy_Reading_Maths	Strongly disagree	NaN	...	
Enjoy_Maths_Lesson	Strongly disagree	NaN	...	
Enjoy_Maths	Strongly disagree	NaN	...	
Interest_in_Maths	Strongly disagree	NaN	...	
Country	New Zealand	Spain	...	
Std Maths Score	462.463	634.5303	...	
Std Reading Score	402.1572	565.75398	...	
Std Science Score	454.25212	625.73668	...	

	183529	\
STIDSTD	7855	
AGE	16.17	
Birth_Year	1996	
Gender	Male	
ICT_resources	0.24	
Math_interest	NaN	
Mother_Occupa	Mechanical engineers	
Father_Occupa	Motor vehicle mechanics and repairers	
Openess_Problem_Solving	NaN	
Enjoy_Reading_Maths	NaN	

Enjoy_Maths_Lesson			NaN
Enjoy_Maths			NaN
Interest_in_Maths			NaN
Country			Finland
Std Maths Score			407.5479
Std Reading Score			352.3564
Std Science Score			411.3577

	274064	\	
STIDSTD	2402		
AGE	16.25		
Birth_Year	1996		
Gender	Male		
ICT_resources	1.15		
Math_interest	2.29		
Mother_Occupa	Housewife		
Father_Occupa	Financial analysts		
Openess_Problem_Solving	2.4465		
Enjoy_Reading_Maths	Strongly agree		
Enjoy_Maths_Lesson	Strongly agree		
Enjoy_Maths	Strongly agree		
Interest_in_Maths	Strongly agree		
Country	Jordan		
Std Maths Score	484.03958		
Std Reading Score	415.87046		
Std Science Score	498.26556		

		20331	\
STIDSTD		4089	
AGE		15.67	
Birth_Year		1996	
Gender		Female	
ICT_resources		-0.4	
Math_interest		NaN	
Mother_Occupa		Kitchen helpers	
Father_Occupa	Vague(a good job, a quiet job, a well paid job...		
Openess_Problem_Solving		NaN	
Enjoy_Reading_Maths		NaN	
Enjoy_Maths_Lesson		NaN	
Enjoy_Maths		NaN	
Interest_in_Maths		NaN	
Country		Argentina	
Std Maths Score		387.7629	
Std Reading Score		396.20582	
Std Science Score		449.58974	

		456409	\
STIDSTD		2033	

```
AGE                                          16.0
Birth_Year                                   1996
Gender                                       Male
ICT_resources                                0.24
Math_interest                                1.51
Mother_Occupa                          Do not know
Father_Occupa           Commercial sales representatives
Openess_Problem_Solving                      1.2387
Enjoy_Reading_Maths                  Strongly agree
Enjoy_Maths_Lesson                   Strongly agree
Enjoy_Maths                                  Agree
Interest_in_Maths                            Agree
Country                                   Thailand
Std Maths Score                          381.84296
Std Reading Score                        356.60672
Std Science Score                         365.4793

                                            445020  \
STIDSTD                                       1426
AGE                                          15.25
Birth_Year                                    1996
Gender                                      Female
ICT_resources                                 1.15
Math_interest                                -0.66
Mother_Occupa           Home-based personal care workers
Father_Occupa                 Bus and tram drivers
Openess_Problem_Solving                    -0.9492
Enjoy_Reading_Maths              Strongly disagree
Enjoy_Maths_Lesson                           Agree
Enjoy_Maths                      Strongly disagree
Interest_in_Maths                         Disagree
Country                                     Sweden
Std Maths Score                          416.97306
Std Reading Score                        518.37072
Std Science Score                        419.28384

                                           357043                5762    \
STIDSTD                                      2791                1020
AGE                                         15.75               15.83
Birth_Year                                   1996                1996
Gender                                       Male              Female
ICT_resources                               -1.13               -1.99
Math_interest                                 NaN                0.91
Mother_Occupa                           Housewife           Housewife
Father_Occupa           Stall and market salespersons                 NaN
Openess_Problem_Solving                       NaN             -0.3443
Enjoy_Reading_Maths                           NaN               Agree
Enjoy_Maths_Lesson                            NaN      Strongly agree
```

Enjoy_Maths		NaN	Disagree
Interest_in_Maths		NaN	Agree
Country		Malaysia	United Arab Emirates
Std Maths Score		505.38246	357.54012
Std Reading Score		380.02354	417.17562
Std Science Score		459.2876	406.88176

	124130	\
STIDSTD	959	
AGE	15.58	
Birth_Year	1996	
Gender	Male	
ICT_resources	-1.13	
Math_interest	NaN	
Mother_Occupa	Housewife	
Father_Occupa	Hand packers	
Openess_Problem_Solving	NaN	
Enjoy_Reading_Maths	NaN	
Enjoy_Maths_Lesson	NaN	
Enjoy_Maths	NaN	
Interest_in_Maths	NaN	
Country	Costa Rica	
Std Maths Score	297.56194	
Std Reading Score	292.85208	
Std Science Score	259.73516	

	436253	\
STIDSTD	3248	
AGE	15.33	
Birth_Year	1996	
Gender	Male	
ICT_resources	1.15	
Math_interest	-0.34	
Mother_Occupa	Accounting associate professionals	
Father_Occupa	Managing directors and chief executives	
Openess_Problem_Solving	0.4639	
Enjoy_Reading_Maths	Disagree	
Enjoy_Maths_Lesson	Disagree	
Enjoy_Maths	Disagree	
Interest_in_Maths	Disagree	
Country	Slovak Republic	
Std Maths Score	574.4742	
Std Reading Score	493.81952	
Std Science Score	548.71314	

	130979
STIDSTD	3206
AGE	15.58

```
Birth_Year                                                          1996
Gender                                                              Male
ICT_resources                                                       1.15
Math_interest                                                        NaN
Mother_Occupa            Vague(a good job, a quiet job, a well paid job...
Father_Occupa            Vague(a good job, a quiet job, a well paid job...
Openess_Problem_Solving                                              NaN
Enjoy_Reading_Maths                                                  NaN
Enjoy_Maths_Lesson                                                   NaN
Enjoy_Maths                                                          NaN
Interest_in_Maths                                                    NaN
Country                                                    Czech Republic
Std Maths Score                                                439.01696
Std Reading Score                                              420.04056
Std Science Score                                               454.3454

[17 rows x 40 columns]
```

I'll like to have a general overview of the null values across the entire columns.

```
In [49]: #checking null values
         pisa_data.isna().sum()

Out[49]: STIDSTD                    0
         AGE                        0
         Birth_Year                 0
         Gender                     0
         ICT_resources              0
         Math_interest         168782
         Mother_Occupa           1603
         Father_Occupa           2554
         Openess_Problem_Solving 172724
         Enjoy_Reading_Maths     169579
         Enjoy_Maths_Lesson      170562
         Enjoy_Maths             170753
         Interest_in_Maths       170744
         Country                    0
         Std Maths Score            0
         Std Reading Score          0
         Std Science Score          0
         dtype: int64

In [50]: pisa_data.value_counts('Father_Occupa').head(40)

Out[50]: Father_Occupa
         Missing                                                    36559
         Vague(a good job, a quiet job, a well paid job, an office jo    14716
         Heavy truck and lorry drivers                              11816
         Bricklayers and related workers                            10536
```

```
Car, taxi and van drivers                                    9917
Social beneficiary (unemployed, retired, sickness, etc.)     8732
Motor vehicle mechanics and repairers                        8453
Police officers                                              7107
Managing directors and chief executives                      6820
House builders                                               6670
Shop keepers                                                 6638
Crop farm labourers                                          6250
Do not know                                                  6056
Security guards                                              5358
Sales and marketing managers                                 5246
Retail and wholesale trade managers                          5066
General office clerks                                        4934
Secondary education teachers                                 4274
Carpenters and joiners                                       4115
Building and related electricians                            4020
Cooks                                                        3910
Building construction labourers                              3785
Bus and tram drivers                                         3686
Accountants                                                  3551
Shop sales assistants                                        3543
Field crop and vegetable growers                             3401
Welders and flamecutters                                     3350
Subsistence crop farmers                                     3282
Commercial sales representatives                             3264
Plumbers and pipe fitters                                    3245
Invalid                                                      3206
Armed forces occupations, other ranks                        2736
Civil engineers                                              2554
Construction managers                                        2544
Managers                                                     2542
Painters and related workers                                 2537
Sales workers                                                2489
Lawyers                                                      2447
Construction supervisors                                     2410
Office supervisors                                           2401
dtype: int64
```

There are names/occupations that are erroneous such as Missing, Do not know etc.

```
In [51]: # remove the erroneous names
         pisa_data.Father_Occupa = pisa_data.Father_Occupa.replace({'Missing': None, 'Do not kno

In [52]: pisa_data.value_counts('Mother_Occupa').head(40)

Out[52]: Mother_Occupa
         Housewife                                           74358
         Missing                                             27044
         Shop sales assistants                               13124
```

```
Primary school teachers                                          10320
Secretaries (general)                                             9869
Nursing professionals                                             9787
Cleaners and helpers in offices, hotels and other establishm      9248
Cooks                                                             8991
Vague(a good job, a quiet job, a well paid job, an office jo      8841
Secondary education teachers                                      8834
Kitchen helpers                                                   8481
Domestic cleaners and helpers                                     8123
General office clerks                                             7499
Early childhood educators                                         6950
Social beneficiary (unemployed, retired, sickness, etc.)          6877
Child care workers                                                6244
Shop keepers                                                      6168
Accountants                                                       6135
Nursing associate professionals                                   6129
Hairdressers                                                      5443
Shop salespersons                                                 4453
Elementary workers not elsewhere classified                       4242
Tailors, dressmakers, furriers and hatters                        4205
Waiters                                                           4032
Cashiers and ticket clerks                                        3729
Accounting associate professionals                                3694
Domestic housekeepers                                             3480
Accounting and bookkeeping clerks                                 3216
Health care assistants                                            3098
Do not know                                                       3045
Sales and marketing managers                                      2984
Sewing, embroidery and related workers                            2936
Home-based personal care workers                                  2895
Retail and wholesale trade managers                               2836
Sales workers                                                     2786
Shop supervisors                                                  2712
Bank tellers and related clerks                                   2554
Administrative and executive secretaries                          2533
Teachers aides                                                    2425
Beauticians and related workers                                   2268
dtype: int64
```

There are names/occupations that are erroneous such as Missing, Do not know etc.

```
In [53]: pisa_data.Mother_Occupa = pisa_data.Mother_Occupa.replace({'Missing': None, 'Do not kno
```

```
In [54]: pisa_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 485490 entries, 0 to 485489
Data columns (total 17 columns):
 #   Column                   Non-Null Count   Dtype
```

```
---  ------                    --------------   -----
 0   STIDSTD                   485490 non-null  int64
 1   AGE                       485490 non-null  float64
 2   Birth_Year                485490 non-null  int64
 3   Gender                    485490 non-null  object
 4   ICT_resources             485490 non-null  float64
 5   Math_interest             316708 non-null  float64
 6   Mother_Occupa             451682 non-null  object
 7   Father_Occupa             437115 non-null  object
 8   Openess_Problem_Solving   312766 non-null  float64
 9   Enjoy_Reading_Maths       315911 non-null  object
 10  Enjoy_Maths_Lesson        314928 non-null  object
 11  Enjoy_Maths               314737 non-null  object
 12  Interest_in_Maths         314746 non-null  object
 13  Country                   485490 non-null  object
 14  Std Maths Score           485490 non-null  float64
 15  Std Reading Score         485490 non-null  float64
 16  Std Science Score         485490 non-null  float64
dtypes: float64(7), int64(2), object(8)
memory usage: 63.0+ MB
```

```
In [55]: #checking null values
         pisa_data.isna().sum()
```

```
Out[55]: STIDSTD                       0
         AGE                           0
         Birth_Year                    0
         Gender                        0
         ICT_resources                 0
         Math_interest            168782
         Mother_Occupa             33808
         Father_Occupa             48375
         Openess_Problem_Solving  172724
         Enjoy_Reading_Maths      169579
         Enjoy_Maths_Lesson       170562
         Enjoy_Maths              170753
         Interest_in_Maths        170744
         Country                       0
         Std Maths Score               0
         Std Reading Score             0
         Std Science Score             0
         dtype: int64
```

```
In [56]: pisa_data = pisa_data.dropna(subset=['Mother_Occupa', 'Father_Occupa'])
         pisa_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 422692 entries, 0 to 485489
```

```
Data columns (total 17 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   STIDSTD                  422692 non-null  int64
 1   AGE                      422692 non-null  float64
 2   Birth_Year               422692 non-null  int64
 3   Gender                   422692 non-null  object
 4   ICT_resources            422692 non-null  float64
 5   Math_interest            278603 non-null  float64
 6   Mother_Occupa            422692 non-null  object
 7   Father_Occupa            422692 non-null  object
 8   Openess_Problem_Solving  275449 non-null  float64
 9   Enjoy_Reading_Maths      278013 non-null  object
 10  Enjoy_Maths_Lesson       277268 non-null  object
 11  Enjoy_Maths              277087 non-null  object
 12  Interest_in_Maths        277110 non-null  object
 13  Country                  422692 non-null  object
 14  Std Maths Score          422692 non-null  float64
 15  Std Reading Score        422692 non-null  float64
 16  Std Science Score        422692 non-null  float64
dtypes: float64(7), int64(2), object(8)
memory usage: 58.0+ MB
```

```
In [57]: #checking null values
         pisa_data.isna().sum()
```

```
Out[57]: STIDSTD                       0
         AGE                           0
         Birth_Year                    0
         Gender                        0
         ICT_resources                 0
         Math_interest            144089
         Mother_Occupa                 0
         Father_Occupa                 0
         Openess_Problem_Solving  147243
         Enjoy_Reading_Maths      144679
         Enjoy_Maths_Lesson       145424
         Enjoy_Maths              145605
         Interest_in_Maths        145582
         Country                       0
         Std Maths Score               0
         Std Reading Score             0
         Std Science Score             0
         dtype: int64
```

While I'll explore the datasets with the level of cleaning I have, I'll still do further cleaning when I want to explore the other columns with null values. I'll also put into consideration the qualitative and quantitative variables.

2 VISUALIZATIONS

2.1 Univariate Exploration

In this exploration, I'll like to establish basic univariate plots of some of the datasets like the distribution of the students on the basis of their gender, and their scores in their respective subjects. I'll be using matplotlib & seaborn features which are referenced at the end of this project.

QUESTION 1

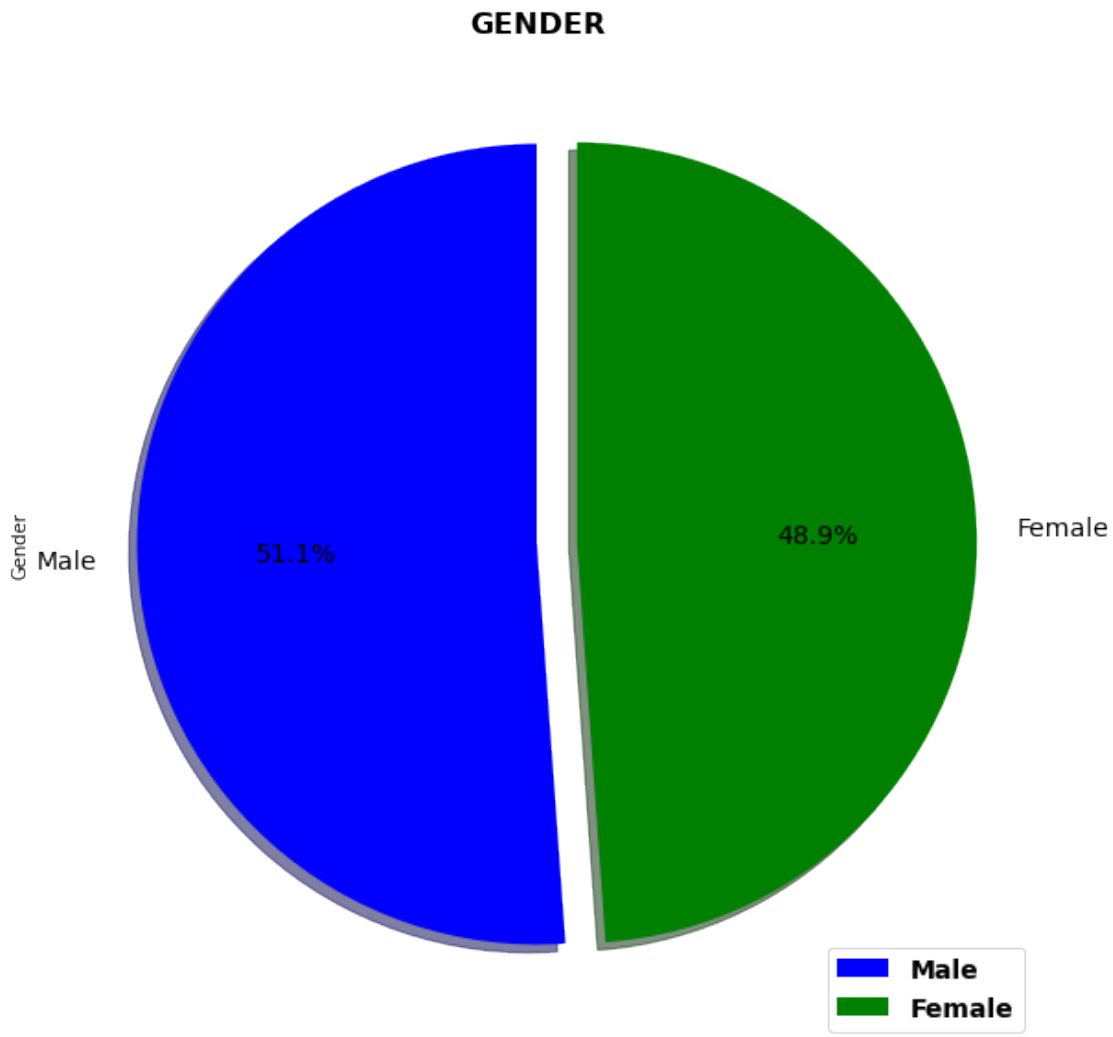
What is the distribution of the genders?

Let's establish the gender distribution using a pie chart.

VISUALIZATION

I'll be establishing the gender relationship with using pie chart

```
In [58]: #Expressing the no show relationship in percentages.
         label = ( 'Male', 'Female')
         colors = ('blue', 'green')
         pisa_data['Gender'].value_counts().plot(kind='pie', figsize=(10,10),
                                          autopct='%1.1f%%', explode=(0, 0.1),
                                          shadow=True, startangle=90, colors = colors, labels =
                                          textprops={'fontsize': 14})
         plt.title('GENDER', fontsize= 16, weight = "bold")
         plt.legend(loc="lower right", prop={'size': 14,'weight':'bold'})
         plt.show()
```




```
In [59]: pisa_data.value_counts('Gender')

Out[59]: Gender
         Female    216158
         Male      206534
         dtype: int64
```

OBSERVATION

The Gender shows that there are more Females than Males in the overall population of students.Although, distribution of the gender is almost evenly distributed

QUESTION 2

Which Top ten(10) countries has the highest number of participants?

```
In [60]: Top_Countries = pisa_data.value_counts('Country').head(10)
         Top_Countries
```

```
Out[60]: Country
         Mexico                  30022
         Italy                   29043
         Spain                   22875
         Canada                  18873
         Brazil                  16860
         Australia               12056
         United Kingdom          10985
         Switzerland             10466
         United Arab Emirates    10304
         Qatar                    8786
         dtype: int64
```

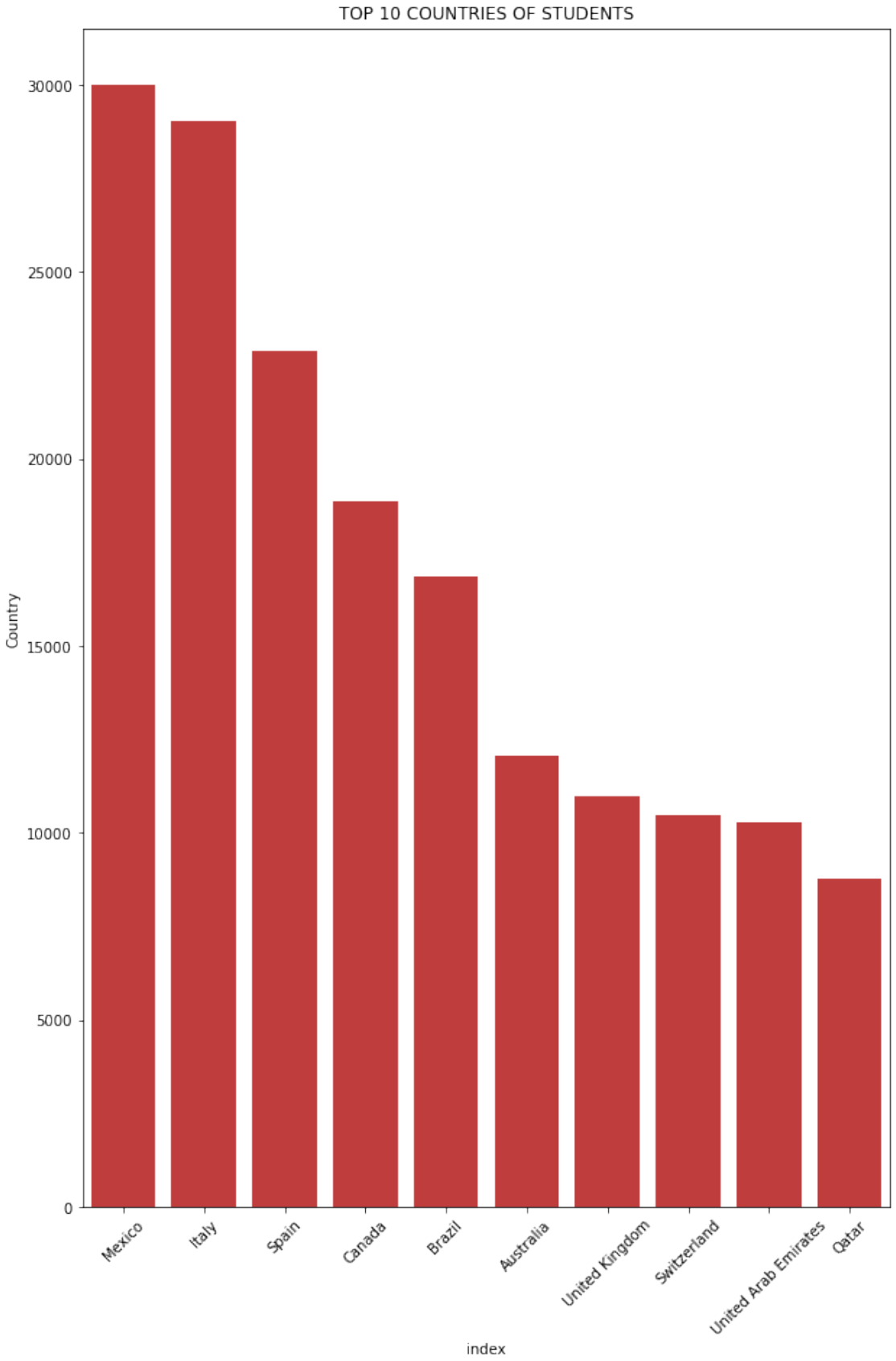
```
In [61]: Top10Countries = pisa_data.Country.value_counts().iloc[:10]. reset_index()
         Top10Countries
```

```
Out[61]:                   index  Country
         0                Mexico    30022
         1                 Italy    29043
         2                 Spain    22875
         3                Canada    18873
         4                Brazil    16860
         5             Australia    12056
         6        United Kingdom    10985
         7           Switzerland    10466
         8  United Arab Emirates    10304
         9                 Qatar     8786
```

VISUALIZATION

I'll be visualizing the top countries with the aid of a barplot

```
In [62]: def bar_plot():
             base_color = sb.color_palette()[3]

             sb.barplot(data=Top10Countries, x='index' , y = 'Country', color = base_color)
         plt.figure(figsize=(10,15))
         plt.xticks(rotation=45)
         plt.xlabel('Countries')
         plt.ylabel('Frequency')
         plt.title('TOP 10 COUNTRIES OF STUDENTS')
         bar_plot()
```


TOP 10 COUNTRIES OF STUDENTS
Country
30000
25000
20000
15000
10000
5000
0
Mexico
Italy
Spain
Canada
Brazil
Australia
United Kingdom
Switzerland
United Arab Emirates
Qatar
index

OBSERVATION

Mexico has the highest number of students that participated in the PISA programme. This was followed closely by Italy while Qatar is the tenth country with number of students that participated.

QUESTION 3

What's the Top ten(10) occupation of the mothers of the students?

```
In [63]: Mother_O = pisa_data.value_counts('Mother_Occupa').head(10)
         Mother_O
```

```
Out[63]: Mother_Occupa
         Housewife                                                    70497
         Shop sales assistants                                        12257
         Primary school teachers                                       9895
         Secretaries (general)                                         9330
         Nursing professionals                                         9163
         Secondary education teachers                                  8499
         Cleaners and helpers in offices, hotels and other establishm  8353
         Cooks                                                         8205
         Vague(a good job, a quiet job, a well paid job, an office jo  7875
         Kitchen helpers                                               7759
         dtype: int64
```

```
In [64]: Top10motheroccupa = pisa_data.Mother_Occupa.value_counts().iloc[:10]. reset_index()
         Top10motheroccupa
```

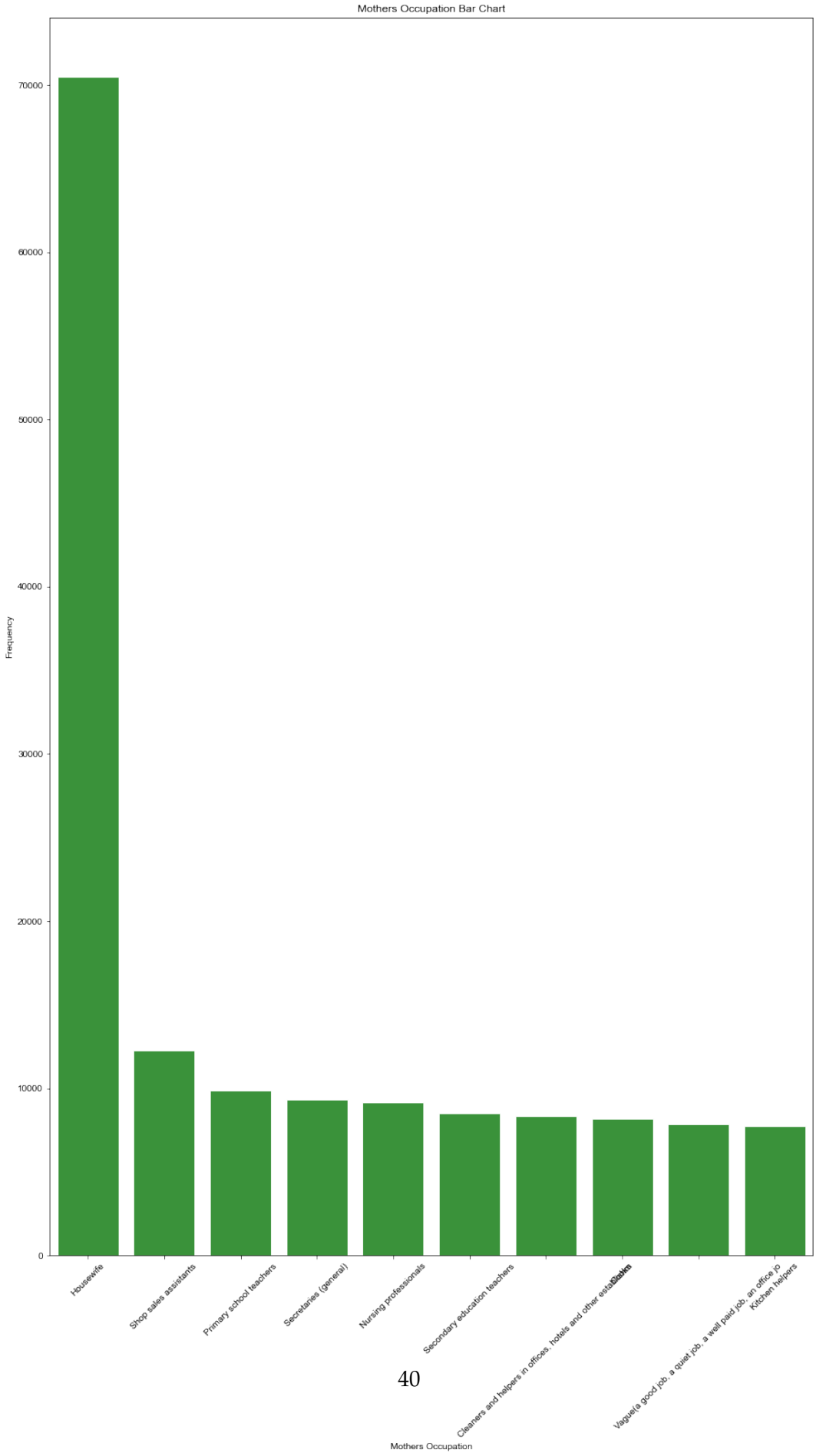
```
Out[64]:                                          index  Mother_Occupa
         0                                    Housewife          70497
         1                      Shop sales assistants          12257
         2                    Primary school teachers           9895
         3                      Secretaries (general)           9330
         4                      Nursing professionals           9163
         5                Secondary education teachers           8499
         6  Cleaners and helpers in offices, hotels and ot...     8353
         7                                       Cooks          8205
         8  Vague(a good job, a quiet job, a well paid job...     7875
         9                           Kitchen helpers           7759
```

VISUALIZATION

I'll be visualizing the occupation of the student mothers with a barplot

```
In [65]: plt.figure(figsize=(15,25))
         plt.title('Mothers Occupation Bar Chart')
         base_color = sb.color_palette()[2]
         sb.set_style(style='white')
         sb.barplot(data=Top10motheroccupa, x='index' , y = 'Mother_Occupa', color = base_color)
         plt.xticks(rotation=45)
```

```
plt.xlabel('Mothers Occupation')
plt.ylabel('Frequency')
plt.show()
```


Mothers Occupation Bar Chart
Frequency
70000
60000
50000
40000
30000
20000
10000
0
Housewife
Shop sales assistants
Primary school teachers
Secretaries (general)
Nursing professionals
Secondary education teachers
Cleaners and helpers in offices, hotels and other estab
Vague(a good job, a quiet job, a well paid job, an office jo
Kitchen helpers
Mothers Occupation

OBSERVATION

The vast majority of the mothers of the students are Housewives i.e they are unemployed. There is a wide margin between the housewives and actual occupation of the women. Sales assistant happens to be the most popular occupation of the women.

QUESTION 4

What's the Top ten(10) occupation of the Fathers of the students?

```
In [66]: Father_O = pisa_data.value_counts('Father_Occupa').head(10)
         Father_O
```

```
Out[66]: Father_Occupa
         Vague(a good job, a quiet job, a well paid job, an office jo    13903
         Heavy truck and lorry drivers                                  11410
         Bricklayers and related workers                                10166
         Car, taxi and van drivers                                       9565
         Social beneficiary (unemployed, retired, sickness, etc.)        8396
         Motor vehicle mechanics and repairers                           8165
         Police officers                                                 6839
         Managing directors and chief executives                         6638
         Shop keepers                                                    6474
         House builders                                                  6414
         dtype: int64
```

```
In [67]: Top10Fatheroccupa = pisa_data.Father_Occupa.value_counts().iloc[:10]. reset_index()
         Top10Fatheroccupa
```

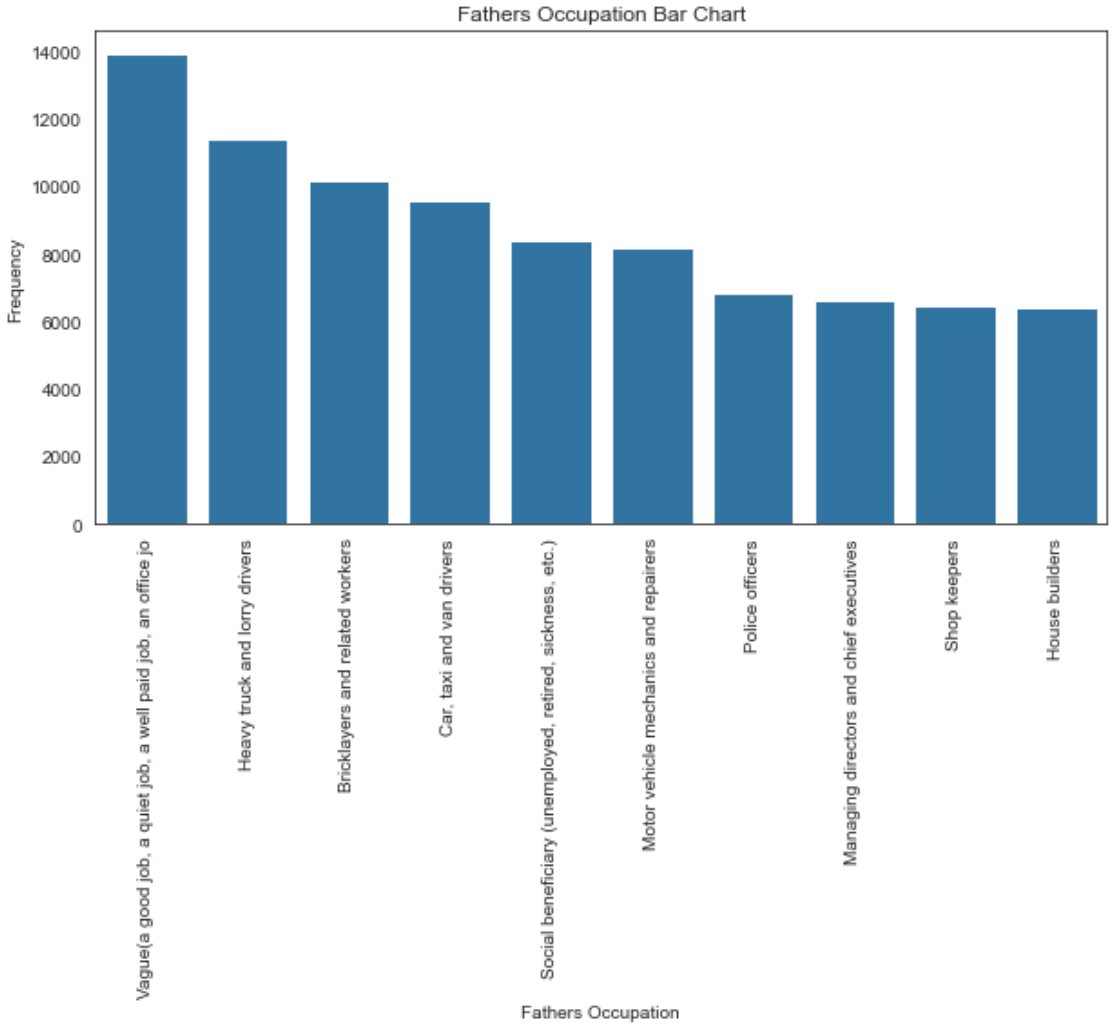
```
Out[67]:                                              index  Father_Occupa
         0  Vague(a good job, a quiet job, a well paid job...          13903
         1                      Heavy truck and lorry drivers          11410
         2                   Bricklayers and related workers          10166
         3                        Car, taxi and van drivers            9565
         4  Social beneficiary (unemployed, retired, sickn...           8396
         5            Motor vehicle mechanics and repairers           8165
         6                                  Police officers           6839
         7          Managing directors and chief executives           6638
         8                                     Shop keepers           6474
         9                                   House builders           6414
```

VISUALIZATION

I'll be establishing the fathers occupation visualization with a barplot

```
In [68]: plt.figure(figsize=(10,5))
         plt.title('Fathers Occupation Bar Chart')
         base_color = sb.color_palette()[0]
         sb.set_style(style='white')
         sb.barplot(data=Top10Fatheroccupa, x="index", y='Father_Occupa', color=base_color)
         plt.xlabel('Fathers Occupation')
```

```
plt.ylabel('Frequency')
plt.xticks(rotation=90)
plt.show()
```



OBSERVATION

Safe to say that majority of the fathers do white collar job. This is followed by heavy duty drivers.

The Next stage of this univariate exploration has to do with how well are the students interest as it relates to Mathematics.

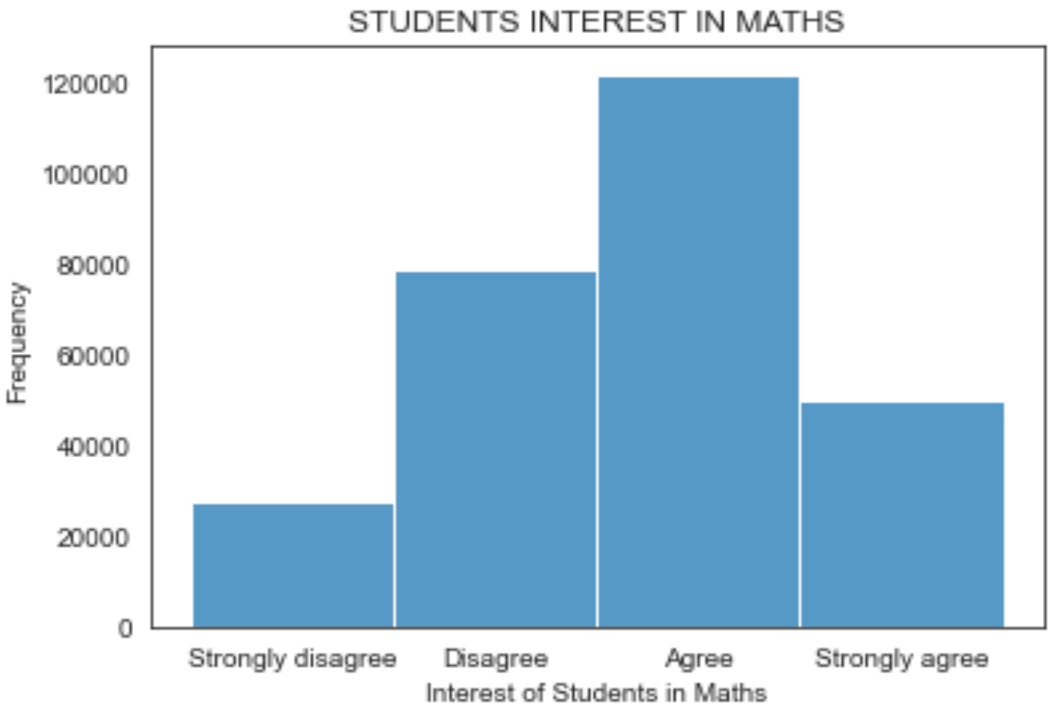
I will be defining a function that arrange the ordinal data in a sequential mannaer(Strongly Disagree, Disagree, Agree, Strongly Agree)

```
In [70]: # converting Enjoy_Reading_Maths, Enjoy_Maths_Lesson, Enjoy_Maths, Interest_in_Maths in
         O_Maths_dict = {'Enjoy_Maths': ['Strongly disagree', 'Disagree', 'Agree', 'Strongly agr
                        'Enjoy_Reading_Maths':['Strongly disagree', 'Disagree', 'Agree', 'Strong
                        'Enjoy_Maths_Lesson':['Strongly disagree', 'Disagree', 'Agree', 'Strongl
```

```
                    'Interest_in_Maths':['Strongly disagree', 'Disagree', 'Agree', 'Strongly

        for maths in O_Maths_dict:
            ordered_math = pd.api.types.CategoricalDtype(ordered = True,
                                                         categories = O_Maths_dict[maths])
            pisa_data[maths] = pisa_data[maths].astype(ordered_math)
```

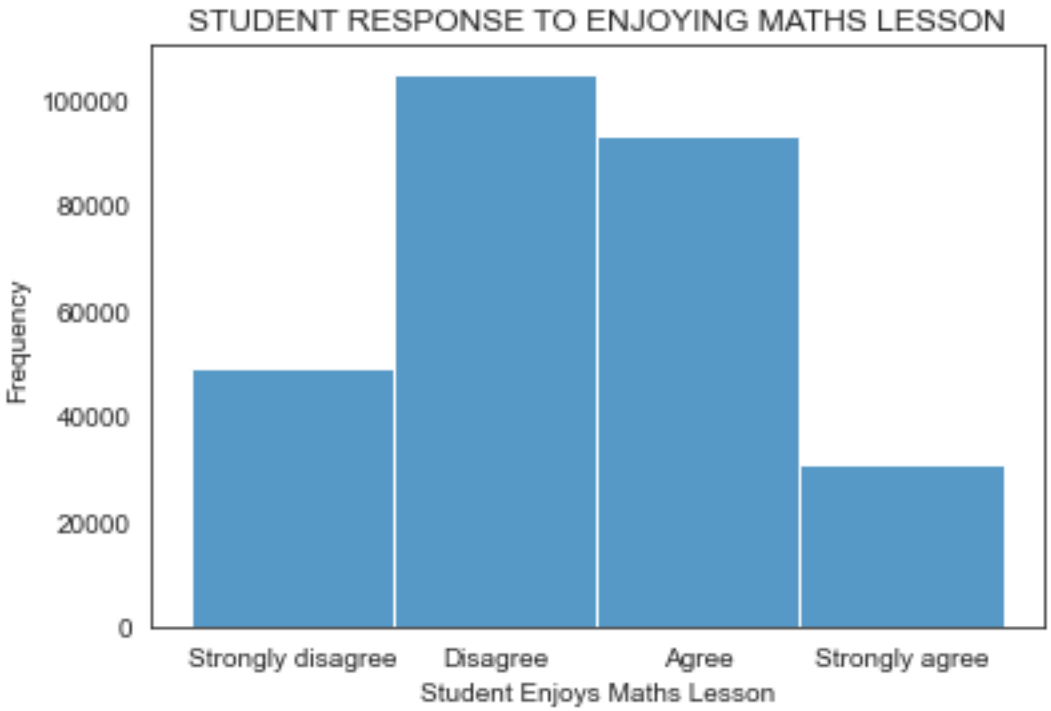
QUESTION 5
How interested are the students as it relates to Maths?
VISUALIZATION
I'll be visualizing the interest of the students in Maths with a histplot

```
In [71]: def hist_plot():
             base_color = sb.color_palette()[0]
             sb.histplot(data = pisa_data, x = 'Interest_in_Maths',color = base_color)
         plt.xlabel('Interest of Students in Maths')
         plt.ylabel('Frequency')
         plt.title('STUDENTS INTEREST IN MATHS')
         hist_plot()
```



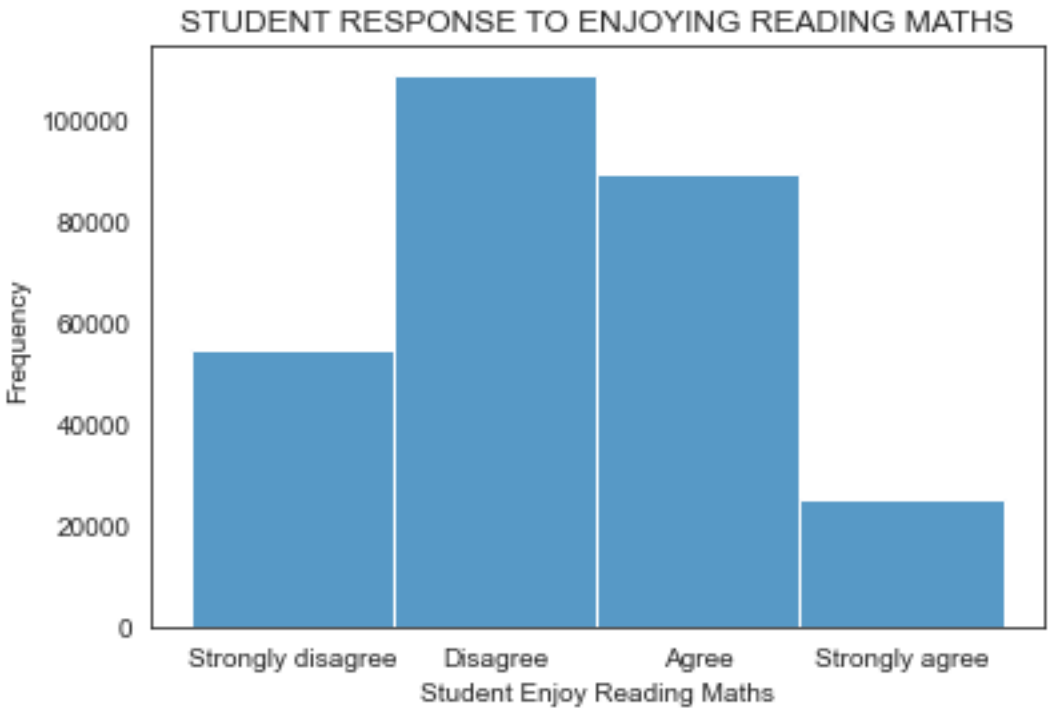
OBSERVATION
Majority of the students agree to having interest in Maths.
VISUALIZATION
I'll be visualizing the students enjoying Maths Lessons with a histplot

```
In [72]: sb.histplot(data=pisa_data, x="Enjoy_Maths_Lesson", bins = 30)

         plt.xlabel('Student Enjoys Maths Lesson')
         plt.ylabel('Frequency')
         plt.title('STUDENT RESPONSE TO ENJOYING MATHS LESSON')
         plt.show()
```



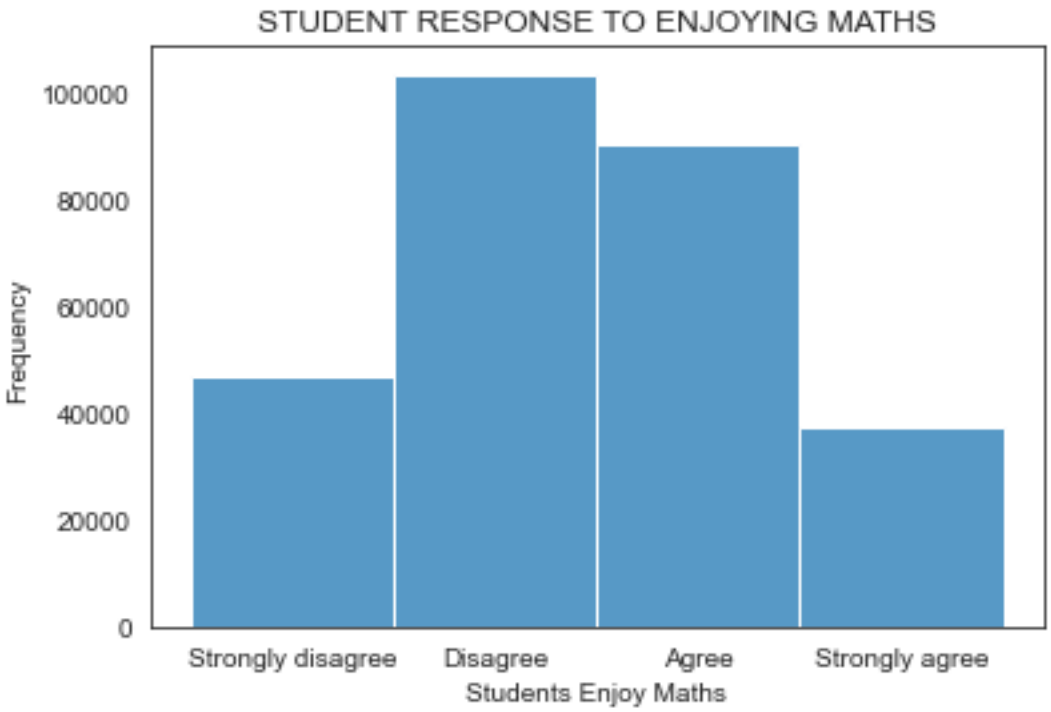
OBSERVATION
Majority of the students disagree to enjoying Maths lessons.
VISUALIZATION
I'll be visualizing the students enjoy Reading Maths with a histplot

```
In [73]: sb.histplot(data=pisa_data, x="Enjoy_Reading_Maths", bins = 30)

         plt.xlabel('Student Enjoy Reading Maths')
         plt.ylabel('Frequency')
         plt.title('STUDENT RESPONSE TO ENJOYING READING MATHS')
         plt.show()
```



OBSERVATION
Majority of the students disagree to enjoy reading Maths.
VISUALIZATION
I'll be visualizing the students enjoying Maths with a histplot

```
In [74]: sb.histplot(data=pisa_data, x="Enjoy_Maths", bins = 30)

         plt.xlabel('Students Enjoy Maths')
         plt.ylabel('Frequency')
         plt.title('STUDENT RESPONSE TO ENJOYING MATHS')

         plt.show()
```



OBSERVATION

Majority of the students disagree to enjoy Maths.

Note In this Univariate stage, for the sake of clean visualization, I decided to limit the number of bins/bars in the barplot as the values are quite enornmous so I restricted the visualization to the first ten(10) values of interest. The Gender distribution is evenly distributed but there might be variation during the Bivariate plots where I will be dropping more values due to the null values in some of the datasets.

It is quite interesting to see that at the first instance, majority of the students agreed to having interest in Maths but when the question were further asked, we had a major votes disagreeing to liking Math lessons or enjoying Maths.

Note

I tried visualizing the datasets that have no null values attached to them with the maximum number of students. While progressing to the bivariate plots especially where relationships between two variables would be established, I need to removed the null values in the other variables that would be needed for the bivariate visualization.

```
In [75]: #checking the non null values in the dataset
         pisa_data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 422692 entries, 0 to 485489
Data columns (total 17 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   STIDSTD                  422692 non-null  int64
```

```
 1   AGE                       422692 non-null  float64
 2   Birth_Year                422692 non-null  int64
 3   Gender                    422692 non-null  object
 4   ICT_resources             422692 non-null  float64
 5   Math_interest             278603 non-null  float64
 6   Mother_Occupa             422692 non-null  object
 7   Father_Occupa             422692 non-null  object
 8   Openess_Problem_Solving   275449 non-null  float64
 9   Enjoy_Reading_Maths       278013 non-null  category
 10  Enjoy_Maths_Lesson        277268 non-null  category
 11  Enjoy_Maths               277087 non-null  category
 12  Interest_in_Maths         277110 non-null  category
 13  Country                   422692 non-null  object
 14  Std Maths Score           422692 non-null  float64
 15  Std Reading Score         422692 non-null  float64
 16  Std Science Score         422692 non-null  float64
dtypes: category(4), float64(7), int64(2), object(4)
memory usage: 46.8+ MB
```

```
In [76]: #checking null values
         pisa_data.isna().sum()
```

```
Out[76]: STIDSTD                       0
         AGE                           0
         Birth_Year                    0
         Gender                        0
         ICT_resources                 0
         Math_interest            144089
         Mother_Occupa                 0
         Father_Occupa                 0
         Openess_Problem_Solving  147243
         Enjoy_Reading_Maths      144679
         Enjoy_Maths_Lesson       145424
         Enjoy_Maths              145605
         Interest_in_Maths        145582
         Country                       0
         Std Maths Score               0
         Std Reading Score             0
         Std Science Score             0
         dtype: int64
```

2.2 Bivariate Exploration

In this Bivariate exploration stage, I would need to ensure that the plots are as accurate as possible. One way to achieve this is to ensure that the non-null values are equal across all the columns while expressing the plots and its derivatives. In the Univariate exploration stage, I plotted charts with available data but in this Bivariate stage, I would be finding the relationship between two variables.

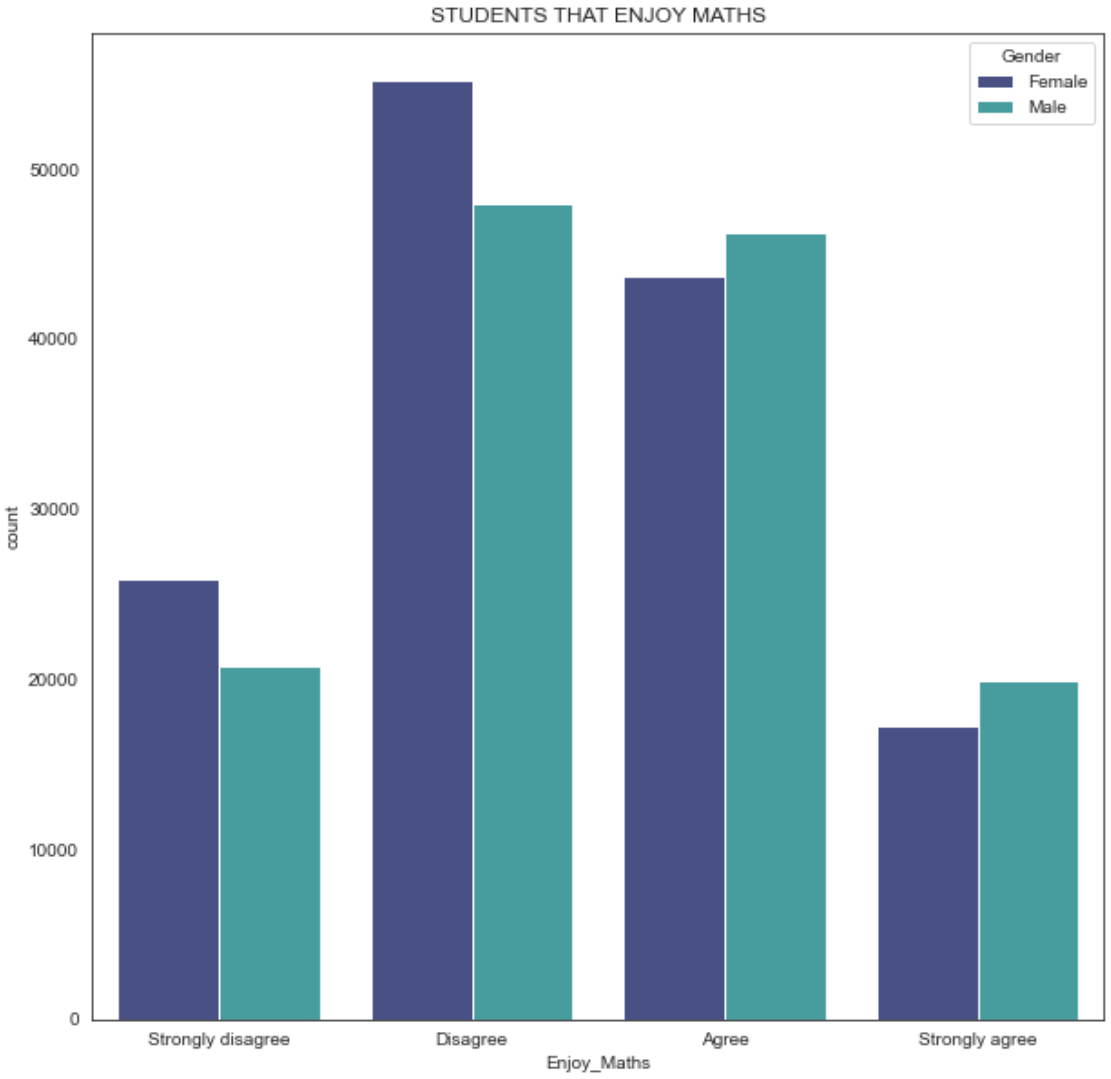
QUESTION 6

What is the Gender distribution of the students that enjoy Maths?

VISUALIZATION

I will be plotting a countplot to express the gender distribution of the students that enjoy Maths.

```
In [77]: plt.figure(figsize=(10,10))
         sb.countplot(data = pisa_data, x = "Enjoy_Maths", palette="mako", hue = 'Gender')
         plt.title ("STUDENTS THAT ENJOY MATHS")
         plt.show ()
```



OBSERVATION

Generally, majority of the students disagree to enjoying Maths while the majority are the females. There are more males that agree to enjoying Maths than females.

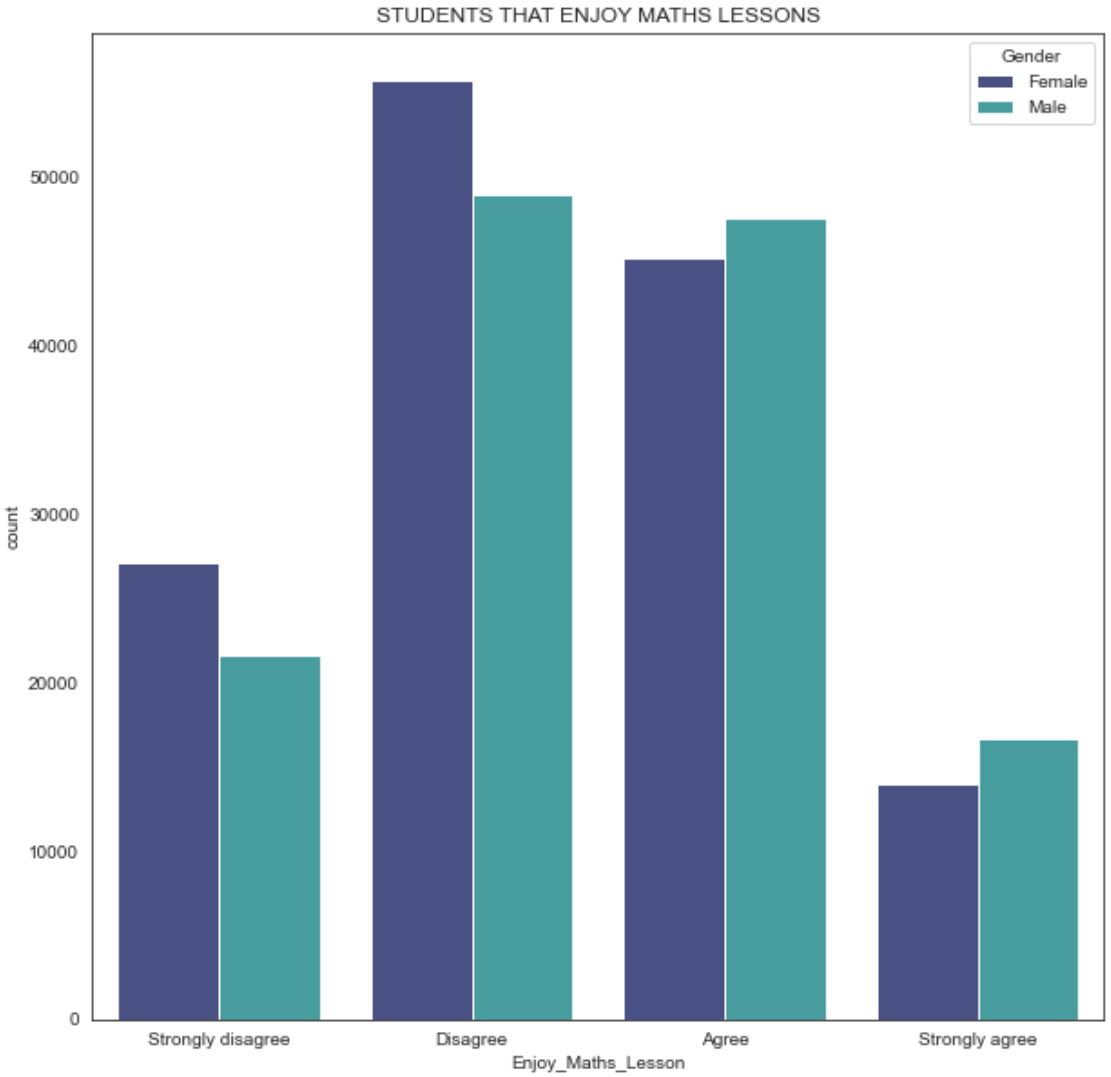
QUESTION 7

What is the Gender distribution of the students that enjoy Maths Lesson?

VISUALIZATION

I will be plotting a countplot to express the gender distribution of the students that enjoy Maths lessons.

```
In [78]: plt.figure(figsize=(10,10))
         sb.countplot(data = pisa_data, x = "Enjoy_Maths_Lesson", palette="mako", hue = 'Gender'
         plt.title ("STUDENTS THAT ENJOY MATHS LESSONS")
         plt.show ()
```



OBSERVATION

Generally, majority of the students disagree to enjoying Maths lessons while the majority are the females. There are more females that disagree to enjoying Maths than females.

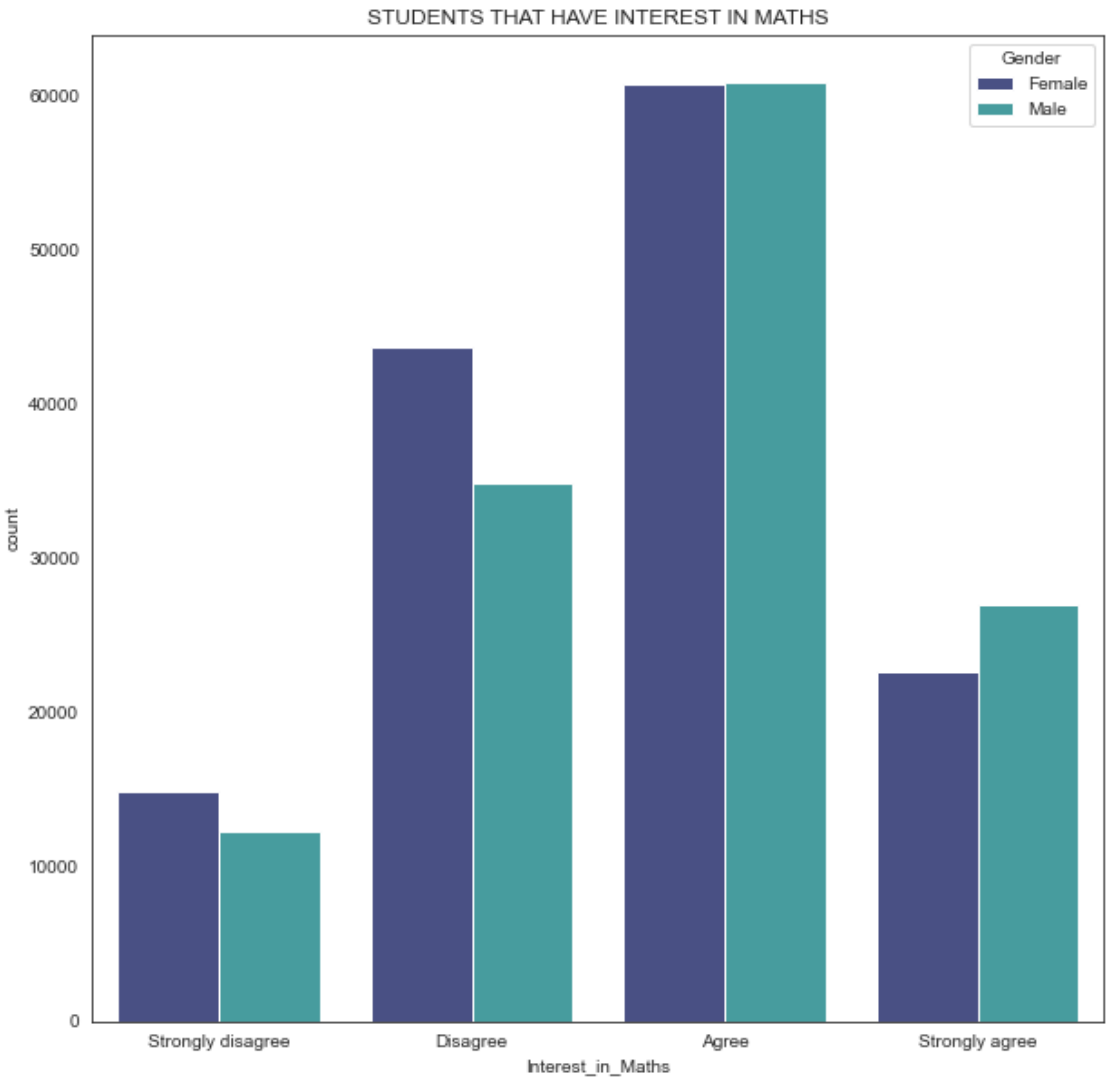
QUESTION 8

What is the Gender distribution of the students that have interest in Maths?

VISUALIZATION

I will be plotting a countplot to express the gender distribution of the students that have interest in Maths.

```
In [79]: plt.figure(figsize=(10,10))
         sb.countplot(data = pisa_data, x = "Interest_in_Maths", palette="mako", hue = 'Gender')
         plt.title ("STUDENTS THAT HAVE INTEREST IN MATHS")
         plt.show ()
```



OBSERVATION

The result above shows that the females and males performed similarly with a strong positive correlation relationship between their reading and math scores. The scatter plot also shows that a male scored the highest in Science as well as Maths.

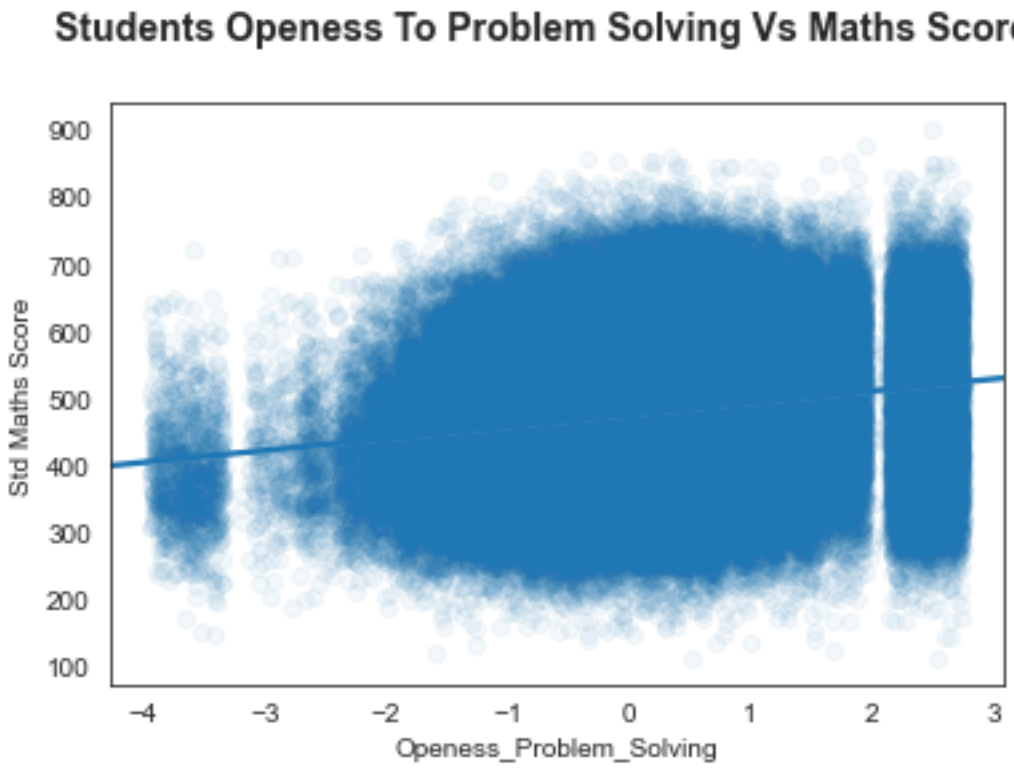
QUESTION 9

Relationship of the Students between Problem Solving and Maths Score?

VISUALIZATION

I will be plotting a regplot to express the openess of the students to problem solving and their Maths Score.

```
In [80]: sb.regplot(data = pisa_data, x = 'Openess_Problem_Solving', y = 'Std Maths Score', trun
         plt.suptitle("STUDENTS OPENESS TO PROBLEM SOLVING VS MATHS SCORE".title(), y = 1, fonts
```



OBSERVATION

The regplot implies a positive weak regression line which implies that there are some of the students that are open minded to problem solving tend to perform considerably well in their Maths Assessment.

So in this exploratory stage, I decided to dig deeper into the interest of the students in Maths on gender basis as well and I found out that despite having more males agreeing to enjoying maths, the females performed better in their maths assesssments.

2.3 Multivariate Exploration

In the Multivariate Exploratory stage, I would be establishing relationships between three variables(either Quantitative or Qualitative and establish possible relationships that exists amongst them and also carry out correlation and regression analysis.

I will also be expressing relationships between the various scores of the students in the three subject that they were tested on while also factoring in their gender.

```
In [81]: #Dropping more rows that have null values for uniformity sake.
         pisa_data = pisa_data.dropna(subset=['Enjoy_Reading_Maths','Enjoy_Maths','Interest_in_M
         pisa_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 270588 entries, 0 to 485489
Data columns (total 17 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   STIDSTD                  270588 non-null  int64
 1   AGE                      270588 non-null  float64
 2   Birth_Year               270588 non-null  int64
 3   Gender                   270588 non-null  object
 4   ICT_resources            270588 non-null  float64
 5   Math_interest            270588 non-null  float64
 6   Mother_Occupa            270588 non-null  object
 7   Father_Occupa            270588 non-null  object
 8   Openess_Problem_Solving  270588 non-null  float64
 9   Enjoy_Reading_Maths      270588 non-null  category
 10  Enjoy_Maths_Lesson       270588 non-null  category
 11  Enjoy_Maths              270588 non-null  category
 12  Interest_in_Maths        270588 non-null  category
 13  Country                  270588 non-null  object
 14  Std Maths Score          270588 non-null  float64
 15  Std Reading Score        270588 non-null  float64
 16  Std Science Score        270588 non-null  float64
dtypes: category(4), float64(7), int64(2), object(4)
memory usage: 29.9+ MB
```

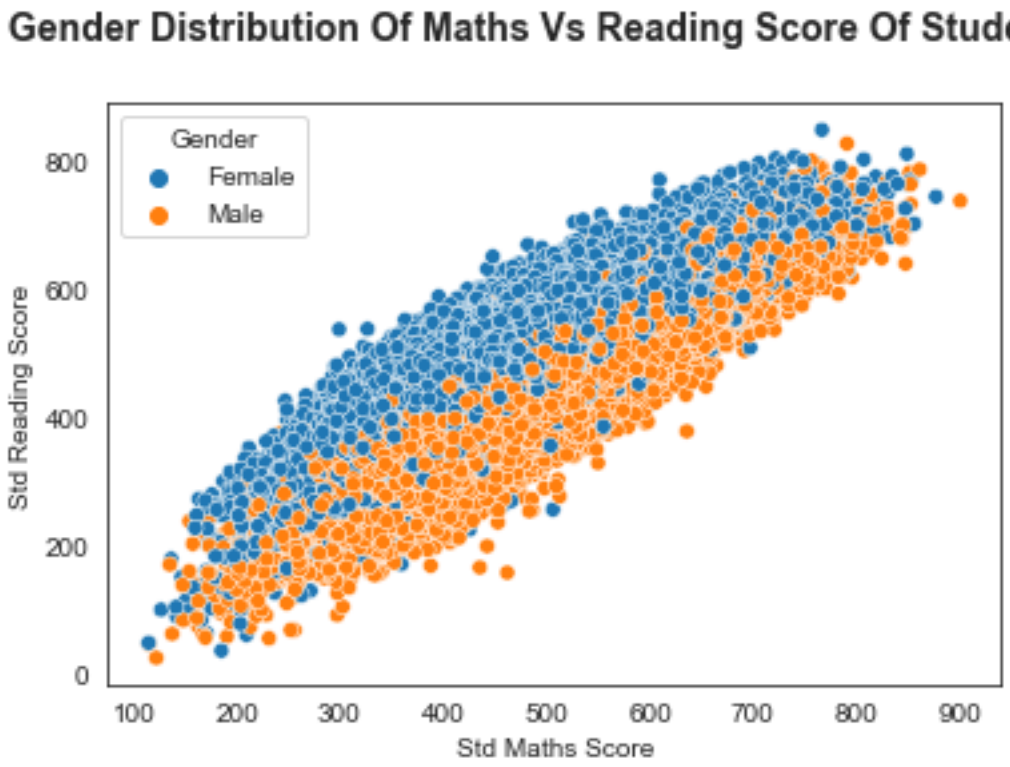
QUESTION 10

What is the correlation between Students performance in Reading and Maths exams as well as their Gender Distribution?

VISUALIZATION

I will be plotting a scatterplot to express the gender distribution of the students and the relationship between their Maths and Reading Score.

```
In [82]: sb.scatterplot(data=pisa_data, x="Std Maths Score", y="Std Reading Score", hue="Gender"
         plt.suptitle(" GENDER DISTRIBUTION OF MATHS VS READING SCORE OF STUDENTS".title(), y =
```



OBSERVATION

The result above shows that the females performed better than the males generally with a strong positive correlation relationship between their reading and math scores. This however shows that despite the majority of the female students disagreeing to enjoying maths, they did better than the males. Also the scatter plot shows that a female scored the highest in reading while a male scored highest in Maths.

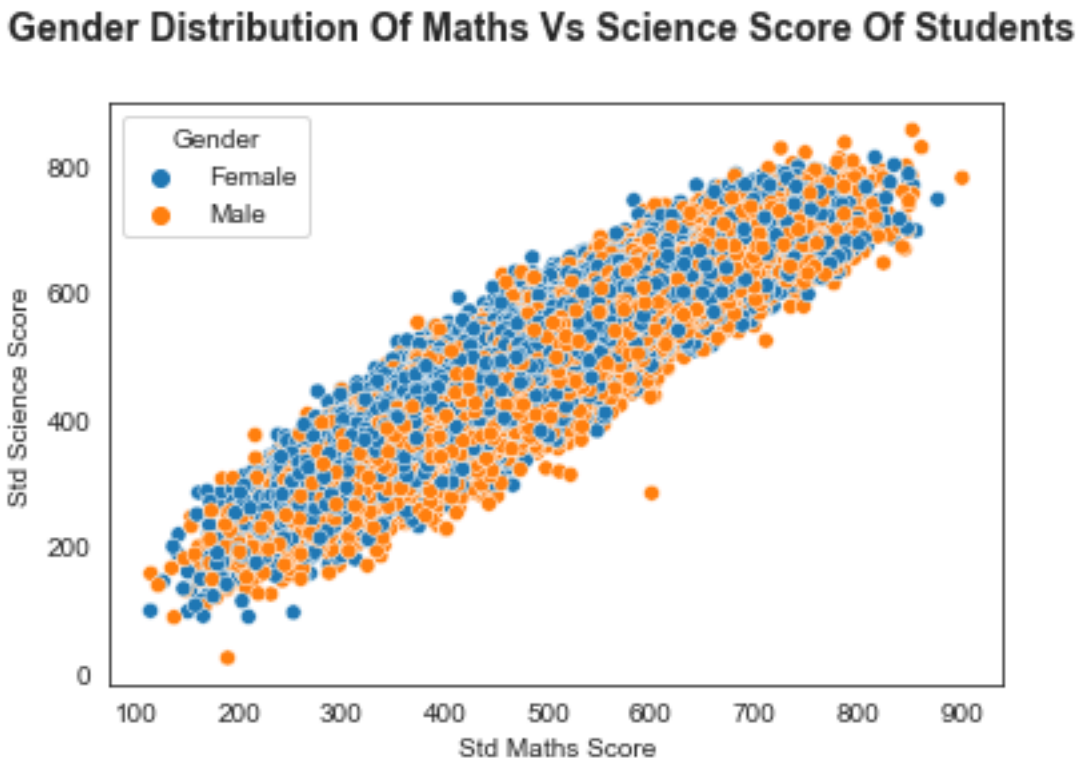
QUESTION 11

What is the correlation between Students performance in Science and Maths exams as well as their Gender Distribution?

VISUALIZATION

I will be plotting a scatterplot to express the gender distribution of the students and the relationship between their Maths and Science Score.

```
In [83]: sb.scatterplot(data=pisa_data, x="Std Maths Score", y="Std Science Score", hue="Gender"
         plt.suptitle("GENDER DISTRIBUTION OF MATHS VS SCIENCE SCORE OF STUDENTS".title(), y = 1
```



OBSERVATION

As earlier stated, we can see that there is a very close tie between the male and females as it relates to having interest in Maths. This however doesn't coincide with how much they enjoy maths lessons or enjoy maths.

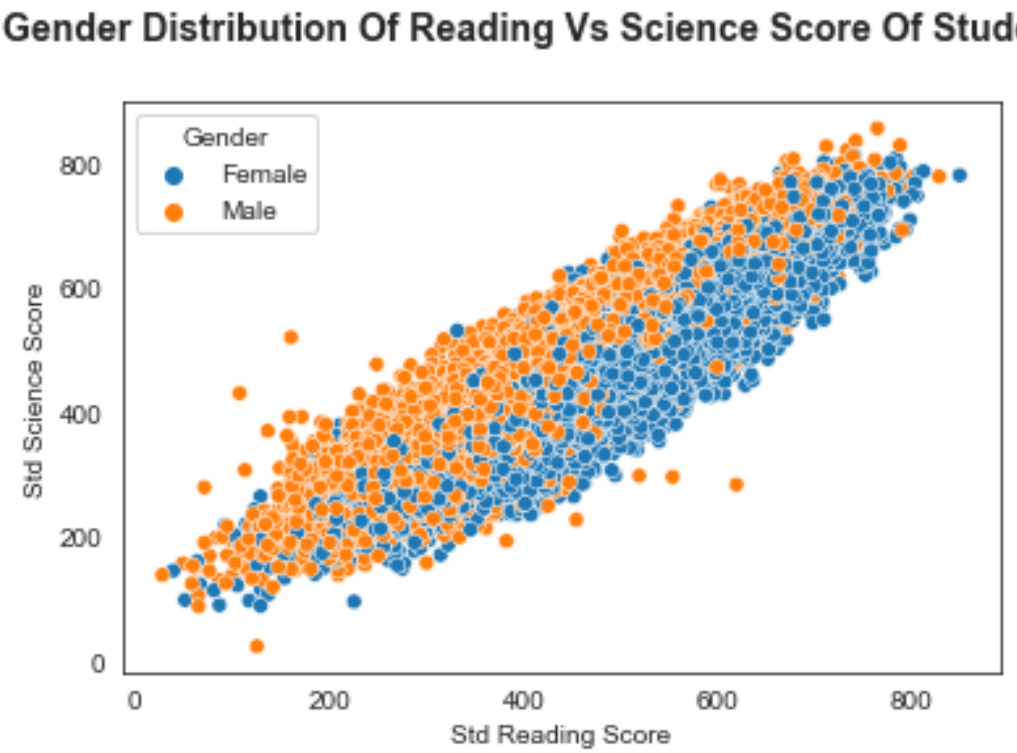
QUESTION 12

What is the correlation between Students performance in Science and Reading exams as well as their Gender Distribution?

VISUALIZATION

I will be plotting a scatterplot to express the gender distribution of the students and the relationship between their Reading and Science Score.

```
In [84]: sb.scatterplot(data=pisa_data, x="Std Reading Score", y="Std Science Score", hue="Gende
         plt.suptitle("GENDER DISTRIBUTION OF READING VS SCIENCE SCORE OF STUDENTS".title(), y =
```



OBSERVATION

Generally, The result above shows that the females and males performed similarly with a strong positve correlation relationship between their reading and science scores. The scatter plot also shows that a male scored the highest in Science while a female scored the highest in Reading.

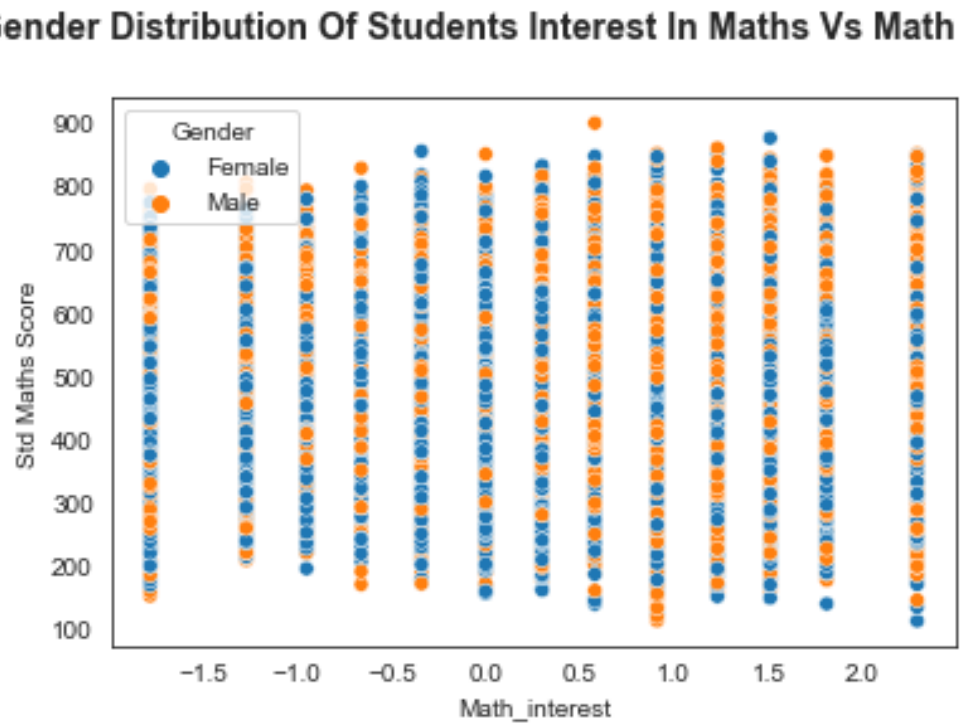
QUESTION 13

Does a student interest in Maths affects their Maths score?

VISUALIZATION

I will be plotting a scatterplot to express the relationship of the Interest of the students in Maths and their Maths Score.

```
In [85]: sb.scatterplot(data=pisa_data, x="Math_interest", y="Std Maths Score", hue="Gender")
         plt.suptitle("GENDER DISTRIBUTION OF STUDENTS INTEREST IN MATHS VS MATH SCORE".title(),
```



OBSERVATION

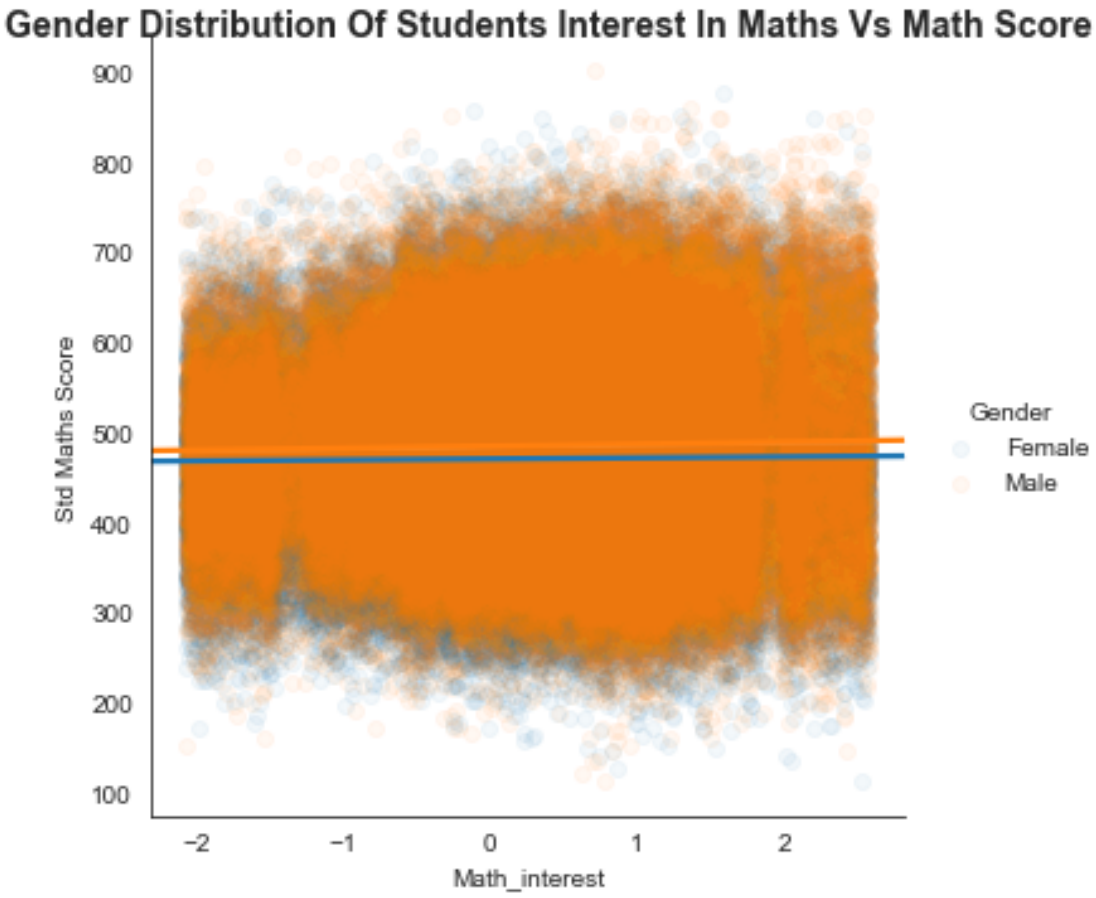
Prior to plotting, the heading of this column seems to express that there is a relationship between the two quantitative variables as it relates to Maths. But, There seems not to be a correlation between these two variables.

However, I would try the regplot/lmplot in my next analysis to see if there is

VISUALIZATION

I will be plotting a lmplot to express the relationship of the Interest of the students in Maths and their Maths Score.

```
In [86]: #I decided to try the lmplot to be able to utilize the hue feature and see if theres is
         sb.lmplot(data = pisa_data, x = 'Math_interest', y = 'Std Maths Score', truncate=False,
         plt.suptitle("GENDER DISTRIBUTION OF STUDENTS INTEREST IN MATHS VS MATH SCORE".title(),
```



OBSERVATION

The lmplot further validates that there is no relationship between the two variables.

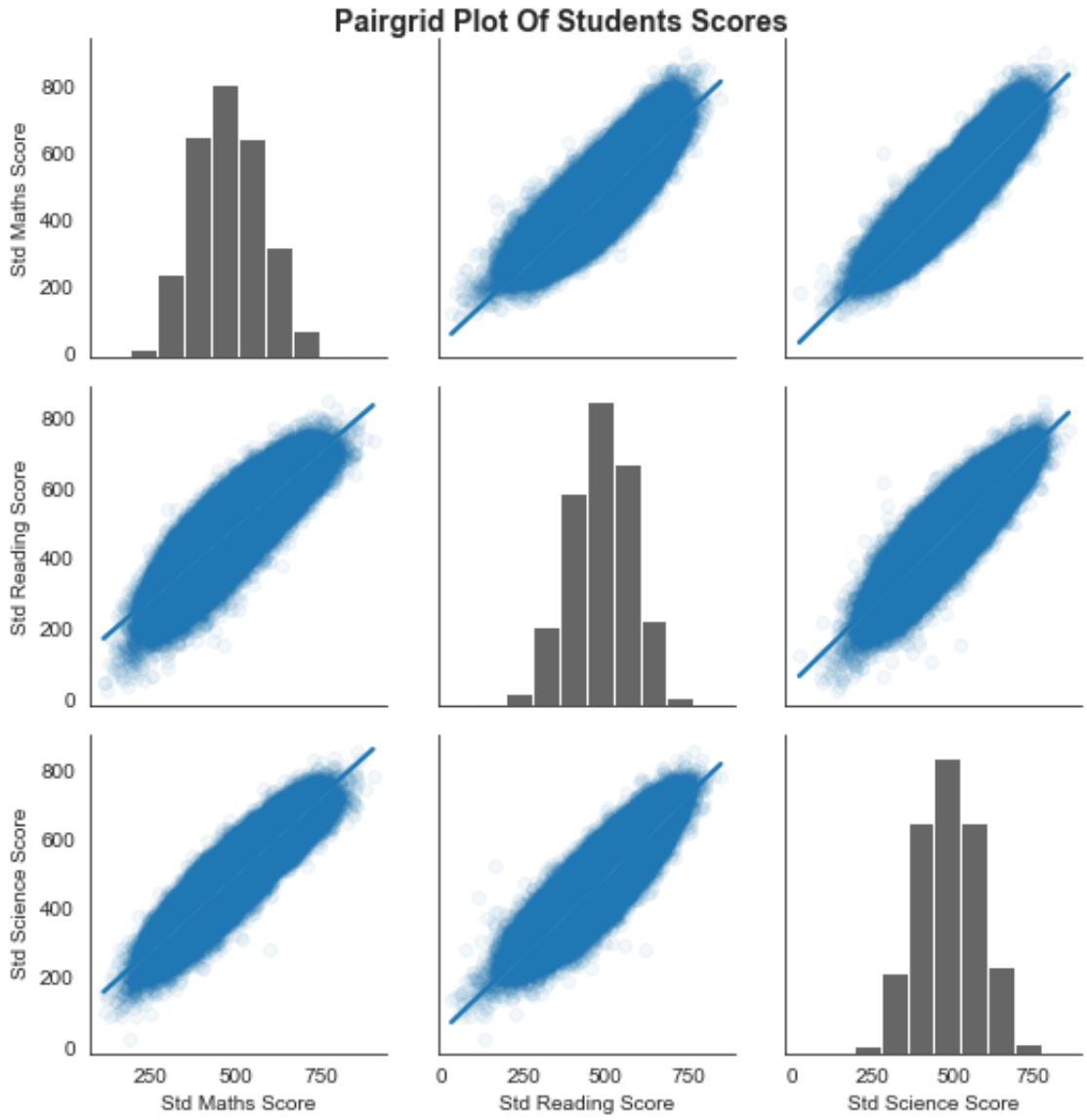
QUESTION 14

Are there similar trends across the entire student scores?

VISUALIZATION

I will be plotting a Pairgrid to express the relationship between the student scores and further validate the results from the Bivariate exploration.

```
In [87]: g = sb.PairGrid(data = pisa_data, vars = ['Std Maths Score', 'Std Reading Score', 'Std
         g.map_diag(plt.hist, color=".4")
         g.map_offdiag(sb.regplot, x_jitter = 0.3, scatter_kws={'alpha':1/20});
         plt.suptitle("PAIRGRID PLOT OF STUDENTS SCORES".title(), y = 1, fontsize = 14, weight =
```



The Pairgrid plot expressed the relationship between three(3) quantitative variables and further validates the scores across the Maths, Science, and Reading of the students. The scatter plot shows positive correlation across the students scores while the histogram shows a peak of 800 across the entire student scores. The histogram also shows a symmetric short tailed distribution across the entire student scores.

2.4 Conclusions

In conclusion, I had to follow the preliminary method of data analysis, assessing, cleaning/wrangling phase. I tried to remove null values to ensure that there were no discrepancies in the analysis. I then visualized using several plots to gain insight into the dataset. I was able to establish the following:

The gender distribution of students reveals that there were more female students than male students. even though it is almost evenly distributed by gender.

The majority of students who took part in the PISA program were from Mexico. Italy came in second place, and Qatar came in tenth place overall in terms of the number of students who participated.

As was already mentioned, it is clear that male and female share a similar level of interest in mathematics. However, this is not consistent with how much they enjoy math or math lessons.

According to the above result, female generally outperformed male in both reading and math, with a strong positive correlation between the two. This demonstrates that, despite the fact that most female students denied enjoying math, they performed better than the male. The scatter plot also reveals that a female outperformed a male in math while a female outperformed a male in reading.

The regplot suggests a positive weak regression line, which suggests that some students who are open to problem-solving have a tendency to perform very well on their math assessments.

The student test scores in math, science, and reading were further validated by the Pairgrid plot, which also expressed the relationship between three quantitative variables. While the histogram shows an 800 peak across all student scores, the scatter plot demonstrates positive correlation across the student scores. A symmetric short-tailed distribution across all student scores is also visible in the histogram.

References

https://www.statology.org/pandas-rename-columns/

https://stackoverflow.com/questions/18171739/unicodedecodeerror-when-reading-csv-file-in-pandas-with-python

https://www.enjoyalgorithms.com/blog/univariate-bivariate-multivariate-analysis

https://stackoverflow.com/questions/42063716/pandas-sum-up-multiple-columns-into-one-column-without-last-column

https://www.w3schools.com/python/pandas/ref_df_drop.asp#:~:text=The%20drop()%20method%20remo

https://linuxhint.com/seaborn-pie-chart/

https://seaborn.pydata.org/generated/seaborn.barplot.html

https://seaborn.pydata.org/generated/seaborn.countplot.html

https://seaborn.pydata.org/generated/seaborn.histplot.html

https://seaborn.pydata.org/generated/seaborn.lmplot.html

https://seaborn.pydata.org/generated/seaborn.PairGrid.html

https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html

https://www.edureka.co/community/584/how-can-replace-values-with-none-in-a-dataframe-using-pandas

https://stackoverflow.com/questions/72450191/i-am-using-pandas-for-a-sentiment-analysis-problem-i-want-to-change-my-categori

https://stackoverflow.com/questions/67551309/any-idea-on-how-to-plot-just-true-values-of-boolean

https://www.geeksforgeeks.org/how-to-extract-the-value-names-and-counts-from-value_counts-in-pandas/?tab=article

https://www.oecd.org/pisa/keyfindings/pisa-2012-results.htm#:~:text=PISA%202012%20is%20the%20prog

https://stackoverflow.com/questions/70133501/how-to-adjust-the-ticks-and-label-size-of-a-pandas-plot-with-secondary-y

3 THANK YOU