PISA

August 20, 2019

1 PISA 2012

1.1 by Jean-Pierre Distler

PISA is a global survey that examins how well 15 year old students are prepared for life after school. The survey is taken every three years.

[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 sns

%matplotlib inline
plt.style.use('fivethirtyeight')
```

For my analysis i want to have a look at how the gender impacts a students score.

[2]:
```
df = pd.read_csv('pisa2012.csv', encoding='latin1', index_col=0,␣
↪low_memory=False)
```

[3]:
```
df.head()
```

[3]:
```
       CNT  SUBNATIO  STRATUM      OECD       NC  SCHOOLID  STIDSTD  ST01Q01  \
1  Albania     80000  ALB0006  Non-OECD  Albania         1        1       10
2  Albania     80000  ALB0006  Non-OECD  Albania         1        2       10
3  Albania     80000  ALB0006  Non-OECD  Albania         1        3        9
4  Albania     80000  ALB0006  Non-OECD  Albania         1        4        9
5  Albania     80000  ALB0006  Non-OECD  Albania         1        5        9

   ST02Q01  ST03Q01  ...  W_FSTR75 W_FSTR76 W_FSTR77  W_FSTR78 W_FSTR79  \
1      1.0        2  ...   13.7954  13.9235  13.1249   13.1249   4.3389
2      1.0        2  ...   13.7954  13.9235  13.1249   13.1249   4.3389
3      1.0        9  ...   12.7307  12.7307  12.7307   12.7307   4.2436
4      1.0        8  ...   12.7307  12.7307  12.7307   12.7307   4.2436
5      1.0       10  ...   12.7307  12.7307  12.7307   12.7307   4.2436

  W_FSTR80 WVARSTRR VAR_UNIT SENWGT_STU  VER_STU
1  13.0829       19        1     0.2098  22NOV13
2  13.0829       19        1     0.2098  22NOV13
```

```
3  12.7307       19        1     0.1999  22NOV13
4  12.7307       19        1     0.1999  22NOV13
5  12.7307       19        1     0.1999  22NOV13

[5 rows x 635 columns]
```

1.2 Data Cleaning

The original dataset contains 634 features most of them are not of interest for my analysis, so i keep only some features of interest and will also change the column names to something more meaningful.

The featiures i'll keep are: * **Country** * **StudenId** * **InternationlGrade:** The class level of the student * **BirthMonth** * **BirthYear** * **Gender** * **Age** * **MathScore, ReadingScore and ScienceScore:** The scores have a range from 0 to 1000.

[4]:
```
columns = {
    'CNT': 'Country',
    'STIDSTD': 'StudentId',
    'ST01Q01': 'InternationalGrade',
    'ST03Q01': 'BirthMonth',
    'ST03Q02': 'BirthYear',
    'ST04Q01': 'Gender',
    'AGE': 'Age',
    'PV1MATH': 'MathScore',
    'PV1READ': 'ReadingScore',
    'PV1SCIE': 'ScienceScore'
}
df_clean = df.rename(columns=columns)
```

[5]:
```
df_clean = df_clean[list(columns.values())]
```

[6]:
```
df_clean.head()
```

[6]:
```
   Country  StudentId  InternationalGrade  BirthMonth  BirthYear  Gender  \
1  Albania          1                  10           2       1996  Female
2  Albania          2                  10           2       1996  Female
3  Albania          3                   9           9       1996  Female
4  Albania          4                   9           8       1996  Female
5  Albania          5                   9          10       1996  Female

     Age  MathScore  ReadingScore  ScienceScore
1  16.17   406.8469      249.5762      341.7009
2  16.17   486.1427      406.2936      548.9929
3  15.58   533.2684      401.2100      499.6643
4  15.67   412.2215      547.3630      438.6796
5  15.50   381.9209      311.7707      361.5628
```

[7]:
```
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Int64Index: 485490 entries, 1 to 485490
Data columns (total 10 columns):
Country               485490 non-null object
StudentId             485490 non-null int64
InternationalGrade    485490 non-null int64
BirthMonth            485490 non-null int64
BirthYear             485490 non-null int64
Gender                485490 non-null object
Age                   485374 non-null float64
MathScore             485490 non-null float64
ReadingScore          485490 non-null float64
ScienceScore          485490 non-null float64
dtypes: float64(4), int64(4), object(2)
memory usage: 40.7+ MB
```

Clean null values. I will fill missing Age values with the mean.

[8]:
```
df_clean.Age = df_clean[['Age']].fillna(df_clean.Age.mean())
```

[9]:
```
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 485490 entries, 1 to 485490
Data columns (total 10 columns):
Country               485490 non-null object
StudentId             485490 non-null int64
InternationalGrade    485490 non-null int64
BirthMonth            485490 non-null int64
BirthYear             485490 non-null int64
Gender                485490 non-null object
Age                   485490 non-null float64
MathScore             485490 non-null float64
ReadingScore          485490 non-null float64
ScienceScore          485490 non-null float64
dtypes: float64(4), int64(4), object(2)
memory usage: 40.7+ MB
```

[10]:
```
df_clean.describe()
```

[10]:

	StudentId	InternationalGrade	BirthMonth	BirthYear	\
count	485490.000000	485490.000000	485490.000000	485490.000000	
mean	6134.066201	9.813323	6.558512	1996.070061	
std	6733.144944	3.734726	3.705244	0.255250	
min	1.000000	7.000000	1.000000	1996.000000	
25%	1811.000000	9.000000	4.000000	1996.000000	
50%	3740.000000	10.000000	7.000000	1996.000000	
75%	7456.000000	10.000000	9.000000	1996.000000	
max	33806.000000	96.000000	99.000000	1997.000000	

	Age	MathScore	ReadingScore	ScienceScore
count	485490.000000	485490.000000	485490.000000	485490.000000
mean	15.784283	469.621653	472.004640	475.769824
std	0.290186	103.265391	102.505523	101.464426
min	15.170000	19.792800	0.083400	2.648300
25%	15.580000	395.318600	403.600700	404.457300
50%	15.750000	466.201900	475.455000	475.699400
75%	16.000000	541.057800	544.502500	547.780700
max	16.330000	962.229300	904.802600	903.338300

[11]:
```
# Save the prepared dataset for later use.
df_clean.to_csv('pisa2012_clean.csv')
```

[12]:
```
df_clean = pd.read_csv('pisa2012_clean.csv')
```

1.2.1 What is the structure of your dataset?

The dataset contains 485490 students and 10 features. Country and Gender are categorical variables.

1.2.2 What is/are the main feature(s) of interest in your dataset?

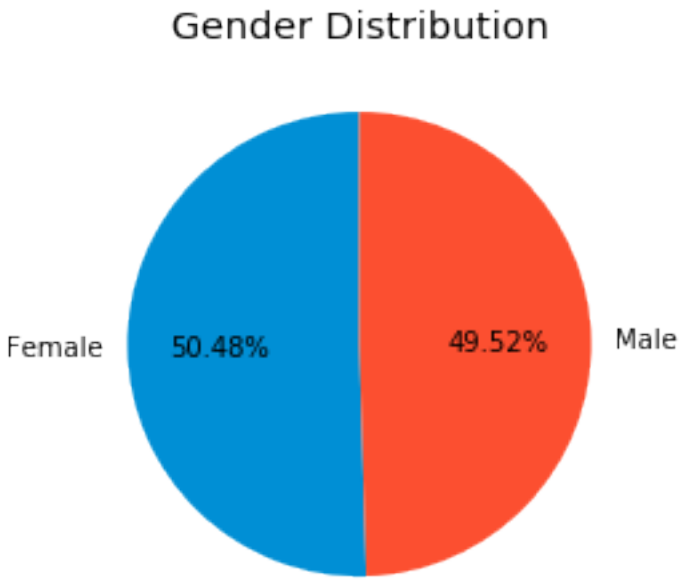
For my analysis gender and the three score variables are the ones of most interest.

1.3 Univariate Exploration

1.3.1 Figure 1:

First i want to have a look at the gender distribution

[13]:
```
values = df_clean.Gender.value_counts()
plt.pie(values, labels= values.index, startangle=90, autopct='%1.2f%%')
plt.title('Gender Distribution')
plt.show();
```



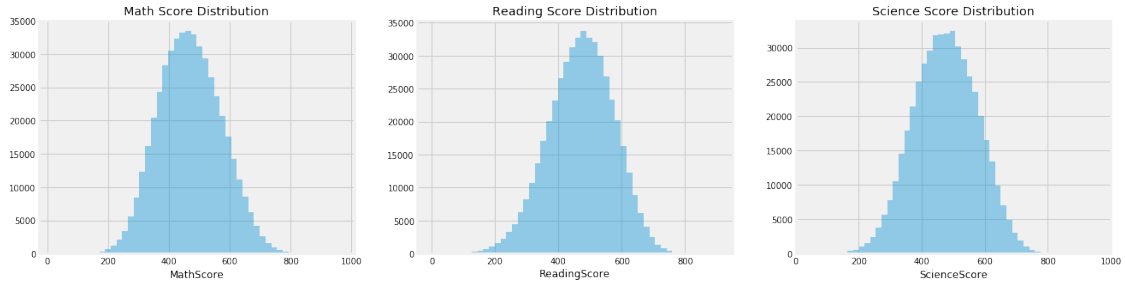
The distribution is almost even with a bit more femal students.

1.3.2 Figure 2:

Next i'll have a look at the distribution of the different scores.

[14]:
```
fig, ax = plt.subplots(figsize=(20,5), ncols=3, nrows=1)
plt.xlim(0, 1000)

sns.distplot(df_clean.MathScore, ax=ax[0], kde=False).set_title('Math Score␣
↪Distribution')
sns.distplot(df_clean.ReadingScore, ax=ax[1], kde=False).set_title('Reading␣
↪Score Distribution')
sns.distplot(df_clean.ScienceScore, ax=ax[2], kde=False).set_title('Science␣
↪Score Distribution');
```

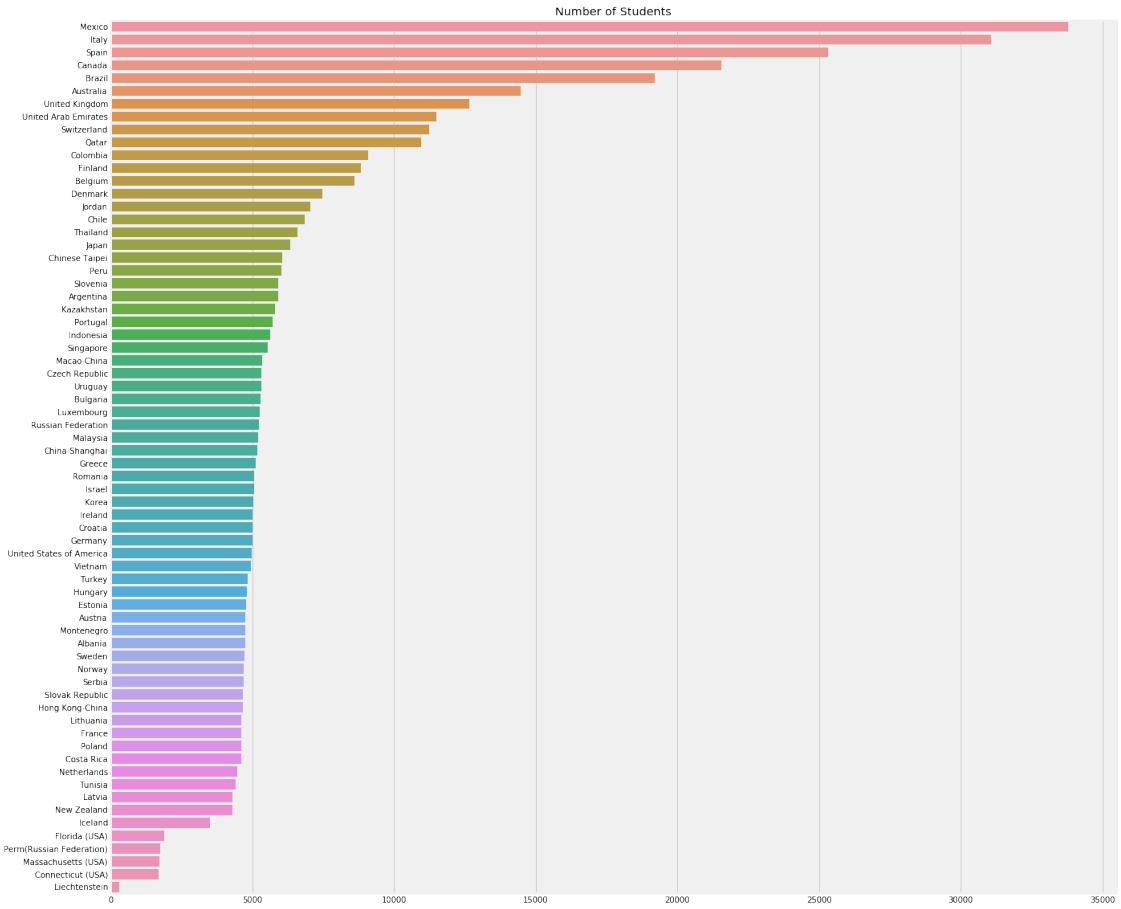


All three scores look normally distributes and most values are around 300 and 600 points.

1.3.3 Figure 3:

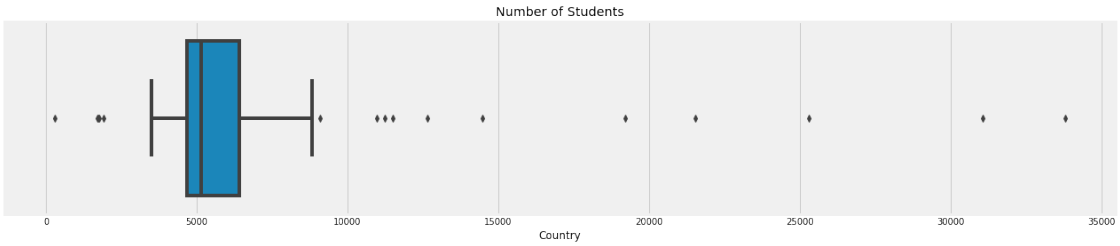
Just for interest i'll check how many students participated the survey per country

[15]:
```
plt.figure(figsize=[20,20])
students_count = df_clean.Country.value_counts().sort_values(ascending=False)
sns.barplot(students_count.values, students_count.index).set_title('Number of␣
↪Students');
```



1.3.4 Figure 4:

[16]:
```
plt.figure(figsize=[20,4])
sns.boxplot(df_clean.Country.value_counts())
plt.title('Number of Students');
```



[17]:
```
students_count
```

[17]:
```
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
                           ...
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
Name: Country, Length: 68, dtype: int64
```

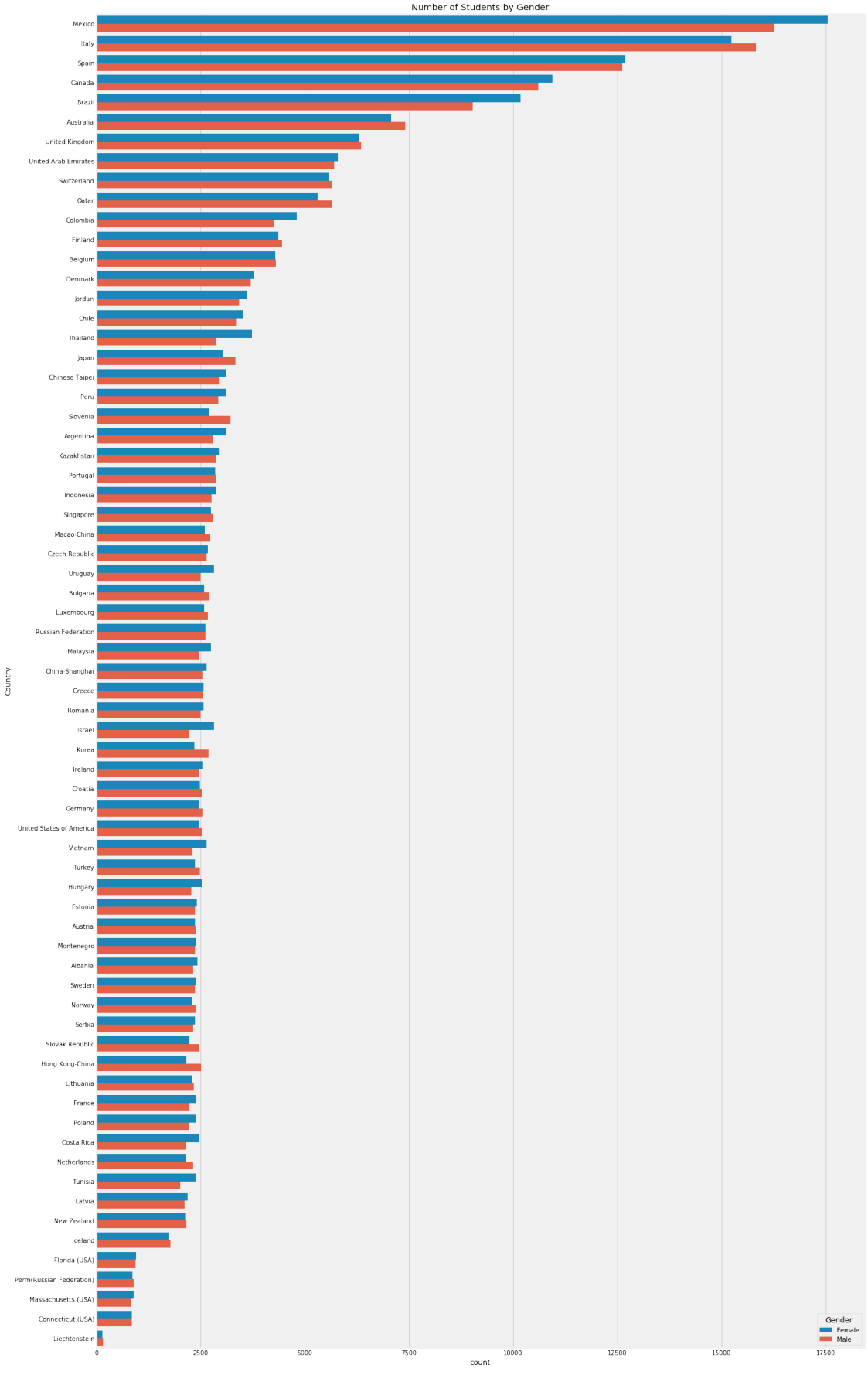
Figure 4 confirms the visible trend one could see in Figure 3. For most countries about 5000 students took the survey. The most extreme outliers are *Liechtenstein* with 283 students and *Mexico* with 33806 students.

1.4 Bivariate Exploration

1.4.1 Figure 5:

Gender per country.

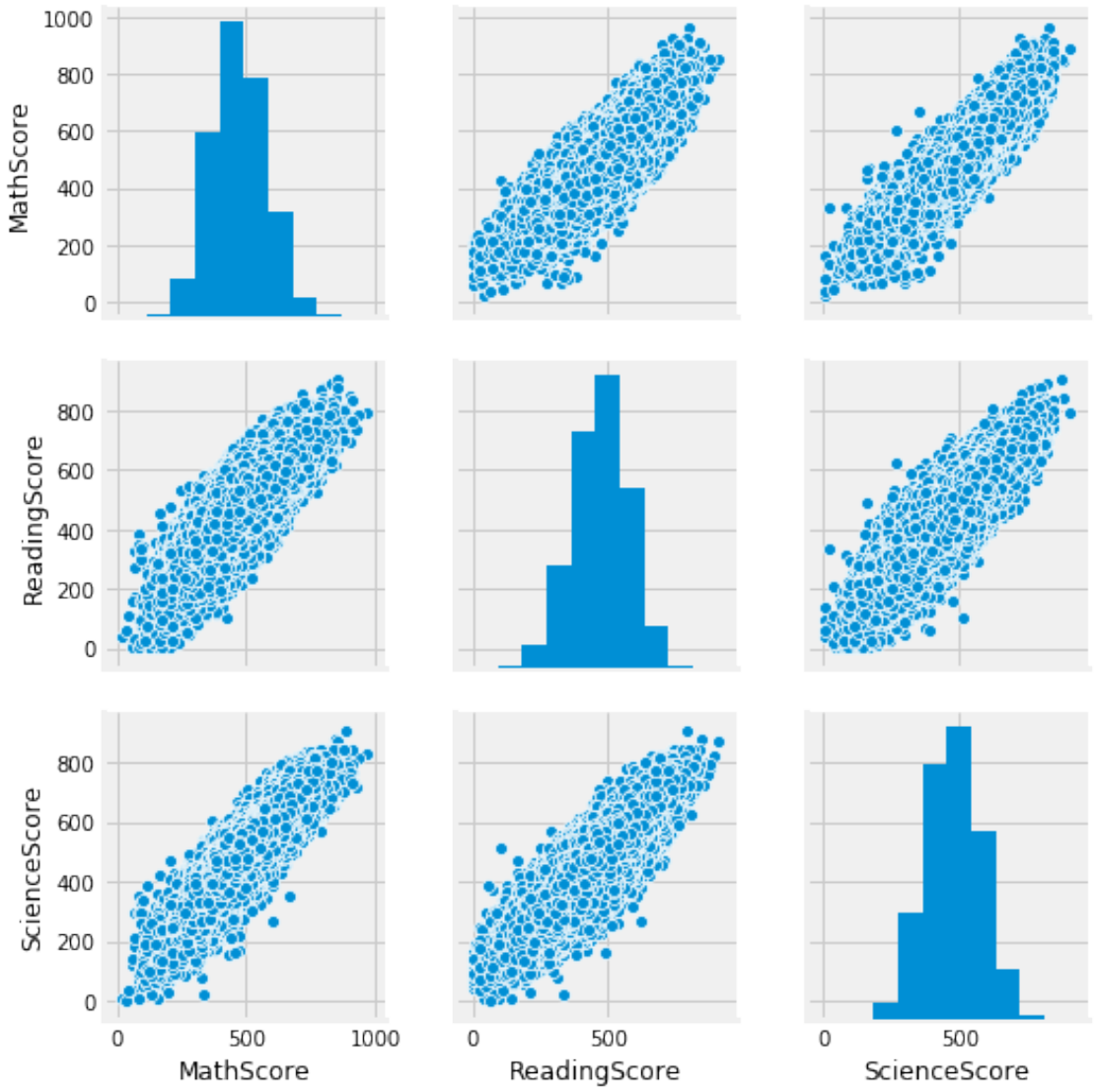
[18]:
```
plt.figure(figsize=[20,40])
sns.countplot(data=df_clean, y='Country', order=students_count.index,␣
 ↪hue='Gender').set_title('Number of Students by Gender');
```


Number of Students by Gender
Mexico
Italy
Spain
Canada
Brazil
Australia
United Kingdom
United Arab Emirates
Switzerland
Qatar
Colombia
Finland
Belgium
Denmark
Jordan
Chile
Thailand
Japan
Chinese Taipei
Peru
Slovenia
Argentina
Kazakhstan
Portugal
Indonesia
Singapore
Macao-China
Czech Republic
Uruguay
Bulgaria
Luxembourg
Russian Federation
Malaysia
China-Shanghai
Greece
Romania
Israel
Korea
Ireland
Croatia
Germany
United States of America
Vietnam
Turkey
Hungary
Estonia
Austria
Montenegro
Albania
Sweden
Norway
Serbia
Slovak Republic
Hong Kong-China
Lithuania
France
Poland
Costa Rica
Netherlands
Tunisia
Latvia
New Zealand
Iceland
Florida (USA)
Perm(Russian Federation)
Massachusetts (USA)
Connecticut (USA)
Liechtenstein
Country
0
2500
5000
7500
10000
12500
15000
17500
count
Gender
Female
Male

Most of the countries have a 50/50 gender distribution, but some like *Mexico* and *Italy* show an unbalanced distribution.

1.4.2 Figure 6:

[19]:
```
sns.pairplot(data=df_clean, vars=['MathScore', 'ReadingScore', 'ScienceScore']);
```



The pairplot shows that there is a positive correlation between the different scores. So it is unlikely that there are students that perform well in only one subject.

[20]:
```
df_male = df_clean[df_clean['Gender'] == 'Male']
df_female = df_clean[df_clean['Gender'] == 'Female']
```

[21]:
```
df_male.describe()
```

[21]:
```
          Unnamed: 0      StudentId  InternationalGrade     BirthMonth  \
count  240426.000000  240426.000000       240426.000000  240426.000000
mean   242791.738019    6153.358780            9.821267       6.549999
std    139782.984283    6787.831308            4.260150       3.667351
min         8.000000       1.000000            7.000000       1.000000
25%    122236.500000    1791.000000            9.000000       4.000000
50%    243518.500000    3728.000000           10.000000       7.000000
75%    364016.250000    7482.000000           10.000000       9.000000
max    485489.000000   33801.000000           96.000000      99.000000

           BirthYear            Age      MathScore   ReadingScore  \
count  240426.000000  240426.000000  240426.000000  240426.000000
mean     1996.069186      15.785047     475.317572     453.966386
std         0.253770       0.289944     106.071792     105.252145
min      1996.000000      15.250000      34.748400       0.083400
25%      1996.000000      15.580000     399.057500     382.509500
50%      1996.000000      15.750000     472.199700     457.010200
75%      1996.000000      16.000000     549.392400     529.024900
max      1997.000000      16.330000     962.229300     889.258700

        ScienceScore
count  240426.000000
mean      476.215567
std       105.287730
min         2.648300
25%       401.659800
50%       476.352100
75%       551.790400
max       903.338300
```

[22]:
```
df_female.describe()
```

[22]:
```
          Unnamed: 0      StudentId  InternationalGrade     BirthMonth  \
count  245064.000000  245064.000000       245064.000000  245064.000000
mean   242700.137066    6115.138747            9.805528       6.566864
std    140507.502483    6679.018142            3.134791       3.742035
min         1.000000       1.000000            7.000000       1.000000
25%    120551.750000    1830.000000            9.000000       4.000000
50%    241888.500000    3753.000000           10.000000       7.000000
75%    364219.250000    7423.000000           10.000000       9.000000
max    485490.000000   33806.000000           96.000000      99.000000

           BirthYear            Age      MathScore   ReadingScore  \
count  245064.000000  245064.000000  245064.000000  245064.000000
mean     1996.070920      15.783534     464.033534     489.701508
std         0.256692       0.290422     100.121737      96.515234
```

```
min      1996.000000      15.170000      19.792800       0.083400
25%      1996.000000      15.500000     392.124900     425.039300
50%      1996.000000      15.750000     460.983000     491.999600
75%      1996.000000      16.000000     533.034700     557.212500
max      1997.000000      16.330000     912.299400     904.802600

        ScienceScore
count  245064.000000
mean      475.332517
std        97.566233
min         6.844500
25%       407.254800
50%       475.326400
75%       544.050700
max       871.727000
```

1.4.3 Figure 6:

Male Score Difference

[23]:
```
fig, ax = plt.subplots(figsize=(20,1), ncols=3, nrows=1)
plt.xlim(0, 1000)
plt.subplot(1,3,1)
sns.boxplot(data=df_male, x='MathScore')
plt.subplot(1,3,2)
sns.boxplot(data=df_male, x='ReadingScore')
plt.subplot(1,3,3)
sns.boxplot(data=df_male, x='ScienceScore');
```

1.4.4 Figure 7:

Female Score Difference

[24]:
```
fig, ax = plt.subplots(figsize=(20,1), ncols=3, nrows=1)
plt.xlim(0, 1000)
plt.subplot(1,3,1)
sns.boxplot(data=df_female, x='MathScore')
plt.subplot(1,3,2)
sns.boxplot(data=df_female, x='ReadingScore')
plt.subplot(1,3,3)
sns.boxplot(data=df_female, x='ScienceScore');
```

Figure 6 and 7 show that male students perform a bit better in Math and female students in Reading. For science both perform more or less equal.

1.5 Multivariate Exploration

in this section i'll look at the different scores for each country.

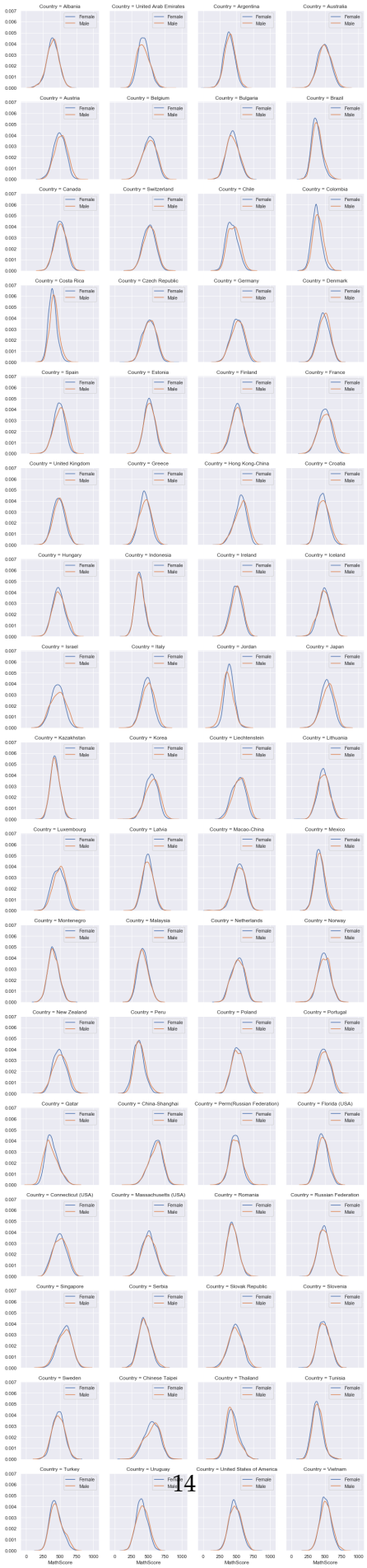
[25]:
```
def draw_gender_score(column):
    country=df_clean['Country'].unique().tolist()
    country.sort()
    sns.set(font_scale=1)
    g=sns.FacetGrid(data=df_clean, col='Country', hue='Gender', col_wrap=4)
    g.map(sns.kdeplot, column)

    for ax, c in zip(g.axes.flat, country):
        ax.legend()
```

1.5.1 Figure 8:

Math Score by Gender

[26]:
```
draw_gender_score('MathScore')
```

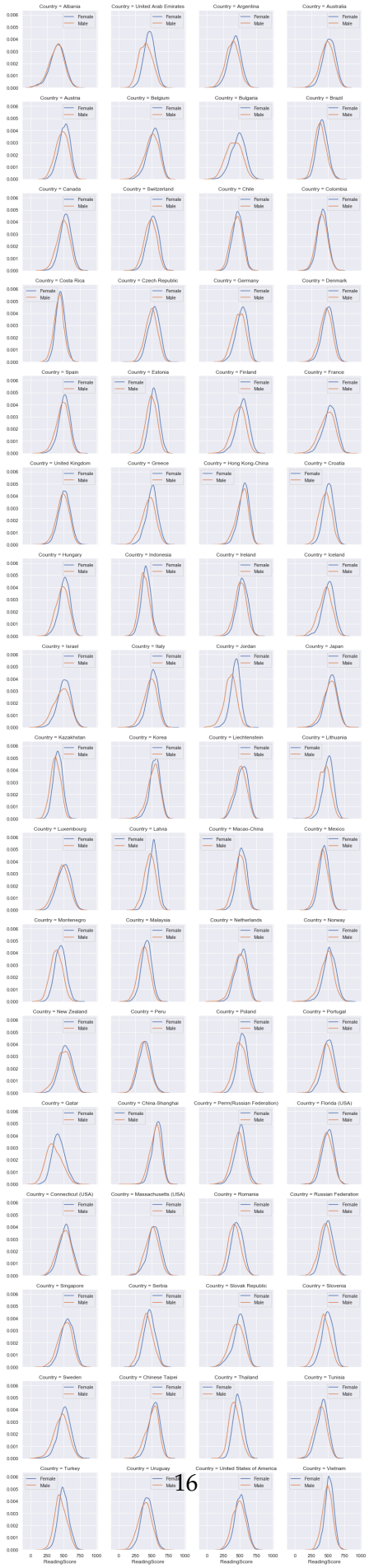

Country = Albania
Country = United Arab Emirates
Country = Argentina
Country = Australia
Country = Austria
Country = Belgium
Country = Bulgaria
Country = Brazil
Country = Canada
Country = Switzerland
Country = Chile
Country = Colombia
Country = Costa Rica
Country = Czech Republic
Country = Germany
Country = Denmark
Country = Spain
Country = Estonia
Country = Finland
Country = France
Country = United Kingdom
Country = Greece
Country = Hong Kong-China
Country = Croatia
Country = Hungary
Country = Indonesia
Country = Ireland
Country = Iceland
Country = Israel
Country = Italy
Country = Jordan
Country = Japan
Country = Kazakhstan
Country = Korea
Country = Liechtenstein
Country = Lithuania
Country = Luxembourg
Country = Latvia
Country = Macao-China
Country = Mexico
Country = Montenegro
Country = Malaysia
Country = Netherlands
Country = Norway
Country = New Zealand
Country = Peru
Country = Poland
Country = Portugal
Country = Qatar
Country = China-Shanghai
Country = Perm(Russian Federation)
Country = Florida (USA)
Country = Connecticut (USA)
Country = Massachusetts (USA)
Country = Romania
Country = Russian Federation
Country = Singapore
Country = Serbia
Country = Slovak Republic
Country = Slovenia
Country = Sweden
Country = Chinese Taipei
Country = Thailand
Country = Tunisia
Country = Turkey
Country = Uruguay
Country = United States of America
Country = Vietnam
Female
Male
MathScore

In most countries the male students achieve higher scores in Math.

1.5.2 Figure 9:

Reading Score by Gender

[27]:
```
draw_gender_score('ReadingScore')
```

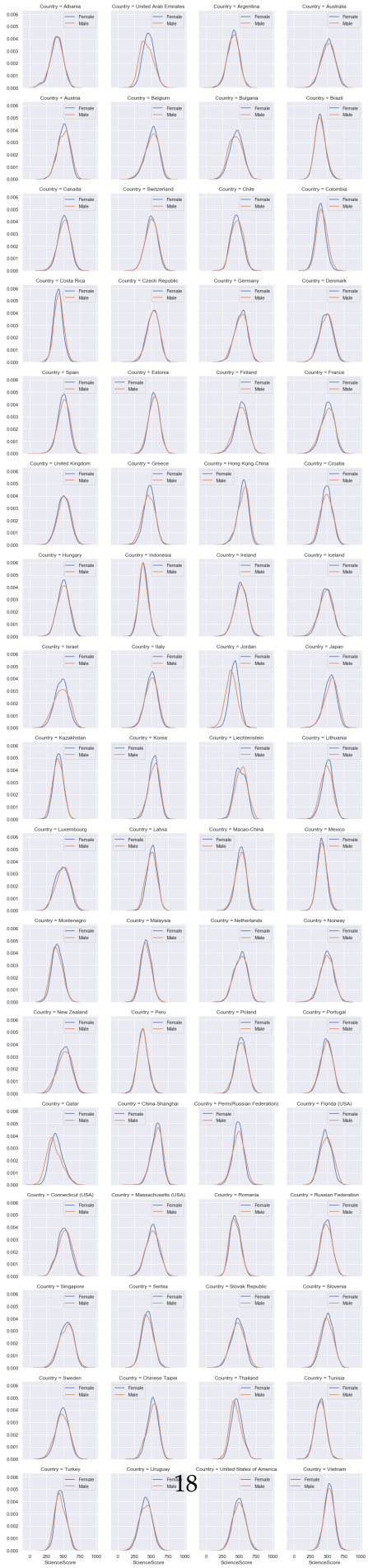

Country = Albania
Country = United Arab Emirates
Country = Argentina
Country = Australia
Country = Austria
Country = Belgium
Country = Bulgaria
Country = Brazil
Country = Canada
Country = Switzerland
Country = Chile
Country = Colombia
Country = Costa Rica
Country = Czech Republic
Country = Germany
Country = Denmark
Country = Spain
Country = Estonia
Country = Finland
Country = France
Country = United Kingdom
Country = Greece
Country = Hong Kong-China
Country = Croatia
Country = Hungary
Country = Indonesia
Country = Ireland
Country = Iceland
Country = Israel
Country = Italy
Country = Jordan
Country = Japan
Country = Kazakhstan
Country = Korea
Country = Liechtenstein
Country = Lithuania
Country = Luxembourg
Country = Latvia
Country = Macao-China
Country = Mexico
Country = Montenegro
Country = Malaysia
Country = Netherlands
Country = Norway
Country = New Zealand
Country = Peru
Country = Poland
Country = Portugal
Country = Qatar
Country = China-Shanghai
Country = Perm(Russian Federation)
Country = Florida (USA)
Country = Connecticut (USA)
Country = Massachusetts (USA)
Country = Romania
Country = Russian Federation
Country = Singapore
Country = Serbia
Country = Slovak Republic
Country = Slovenia
Country = Sweden
Country = Chinese Taipei
Country = Thailand
Country = Tunisia
Country = Turkey
Country = Uruguay
Country = United States of America
Country = Vietnam
Female
Male
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
250
500
750
1000
ReadingScore

As we saw earlier female students have a higher reading score than male. Interestingly this seems to be true for all countries.

1.5.3 Figure 10:

Science Score by Gender

[28]:
```
draw_gender_score('ScienceScore')
```


Country = Albania
Country = United Arab Emirates
Country = Argentina
Country = Australia
Country = Austria
Country = Belgium
Country = Bulgaria
Country = Brazil
Country = Canada
Country = Switzerland
Country = Chile
Country = Colombia
Country = Costa Rica
Country = Czech Republic
Country = Germany
Country = Denmark
Country = Spain
Country = Estonia
Country = Finland
Country = France
Country = United Kingdom
Country = Greece
Country = Hong Kong-China
Country = Croatia
Country = Hungary
Country = Indonesia
Country = Ireland
Country = Iceland
Country = Israel
Country = Italy
Country = Jordan
Country = Japan
Country = Kazakhstan
Country = Korea
Country = Liechtenstein
Country = Lithuania
Country = Luxembourg
Country = Latvia
Country = Macao-China
Country = Mexico
Country = Montenegro
Country = Malaysia
Country = Netherlands
Country = Norway
Country = New Zealand
Country = Peru
Country = Poland
Country = Portugal
Country = Qatar
Country = China-Shanghai
Country = Perm(Russian Federation)
Country = Florida (USA)
Country = Connecticut (USA)
Country = Massachusetts (USA)
Country = Romania
Country = Russian Federation
Country = Singapore
Country = Serbia
Country = Slovak Republic
Country = Slovenia
Country = Sweden
Country = Chinese Taipei
Country = Thailand
Country = Tunisia
Country = Turkey
Country = Uruguay
Country = United States of America
Country = Vietnam
Female
Male
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
250
500
750
1000
ScienceScore

The science score is almost euqal for many countries.