Untitled1

May 14, 2020

1 DATA ANALYSIS

2 AUTHOR

2.1 S SAI SURYATEJA

2.2 Vellore Institute of Technology, Vellore

2.3 Why do we need to perform Exploratory Data Analysis?

```
* To Maximise the insight into dataset.
* To understand the connection between the variables and to uncover the underlying structure
* To extract the import Variables
* To detect anomalies
* To test the underlying assumptions.
```

2.4 Objective of this kernel:

To understand the how the student's performance (test scores) is affected by the other variables (Gender, Ethnicity, Parental level of education, Lunch, Test preparation course)

2.4.1 Lets import the required libraries

[1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
```

2.4.2 Read the Dataset

[2]:

```
df=pd.read_csv("https://raw.githubusercontent.com/sakurusurya2000/
 ↪VERZEO_MINI_PROJECT/master/StudentsPerformance.csv")
```

2.4.3 Information of the Dataset

[3]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64
 6   reading score                1000 non-null   int64
 7   writing score                1000 non-null   int64
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

Here, you can see all the column names, total values and type of the values.

2.5 We have 2 types of variables.

- Numerical variables : which contains number as values
- Categorical variables : which contains descriptions of groups or things.

2.5.1 In this Dataset,

- Numerical Variables are Math score, Reading score and Writing score.
- Categorical Variables are Gender, Race/ethnicity, Parental level of education, Lunch and Test preparation course.

2.5.2 Statistics the numerical variables

[4]:
```
df.describe()
```

[4]:
```
       math score  reading score  writing score
count  1000.00000    1000.000000    1000.000000
mean     66.08900      69.169000      68.054000
std      15.16308      14.600192      15.195657
min       0.00000      17.000000      10.000000
25%      57.00000      59.000000      57.750000
50%      66.00000      70.000000      69.000000
75%      77.00000      79.000000      79.000000
max     100.00000     100.000000     100.000000
```

You can see the descriptive statistics of numerical variables such as total count, mean, standard deviation, minimum and maximum values and three quantiles of the data (25%,50%,75%).

2.5.3 Count the number of rows and columns

```
[5]: df.shape
```

```
[5]: (1000, 8)
```

2.5.4 Null Value Check

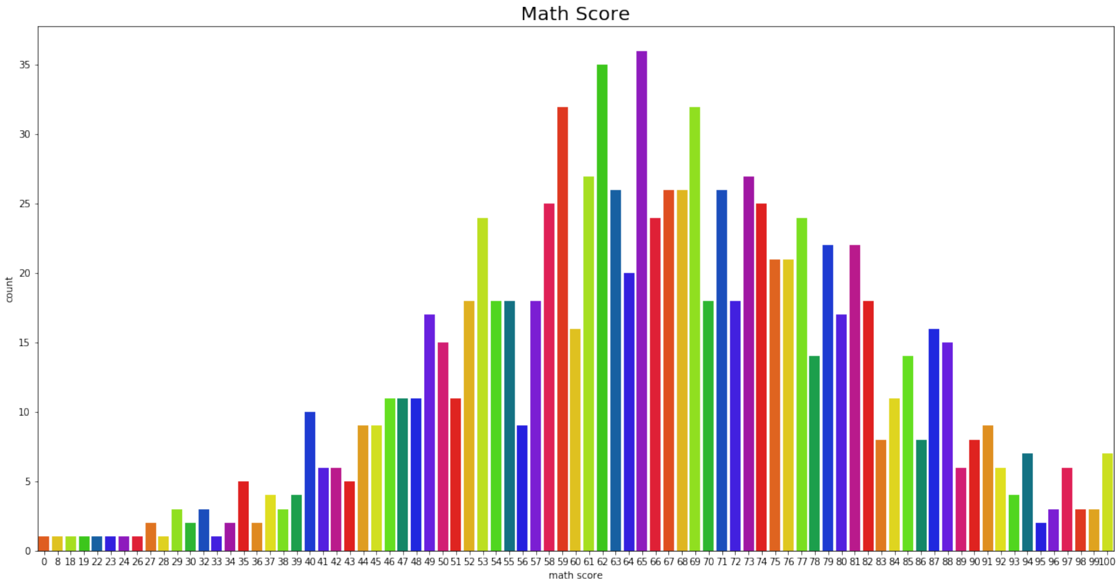
```
[6]: df.isnull().sum()
```

```
[6]: gender                         0
     race/ethnicity                 0
     parental level of education    0
     lunch                          0
     test preparation course        0
     math score                     0
     reading score                  0
     writing score                  0
     dtype: int64
```

2.6 Plots of Numerical Variables:

2.6.1 Maths Score Distribution

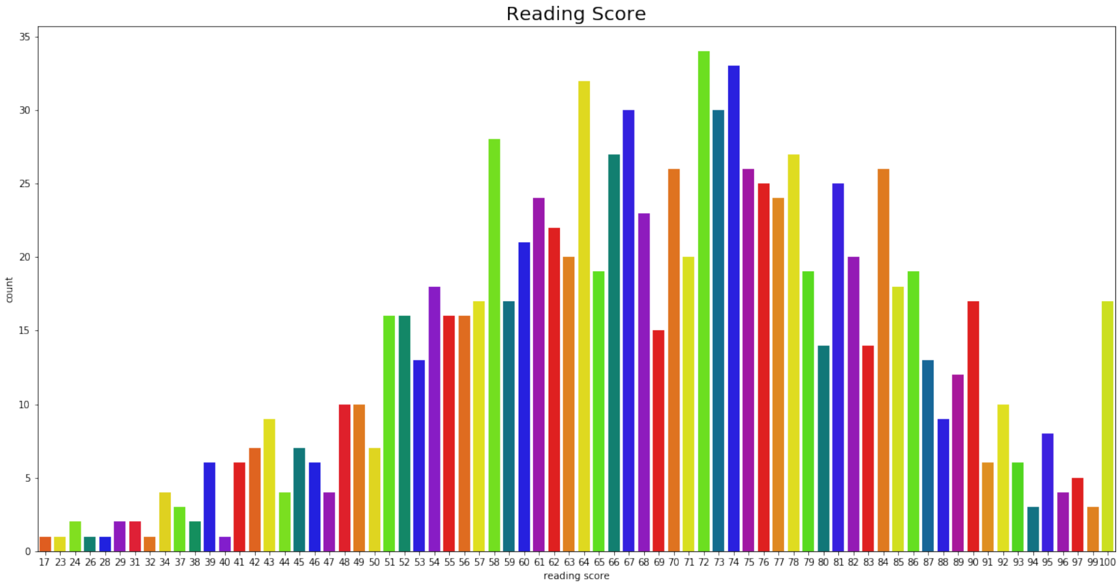
```
[7]: plt.rcParams['figure.figsize'] = (20, 10)
     sns.countplot(df['math score'], palette = 'prism')
     plt.title('Math Score',fontsize = 20)
     plt.show()
```



2.6.2 Reading Score Distribution

[8]:

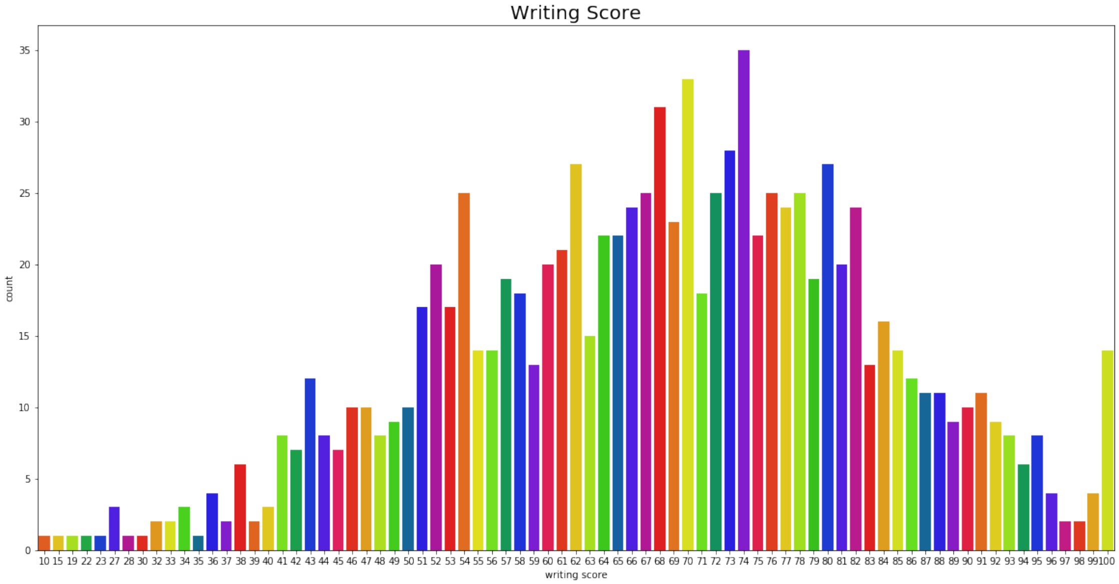
```
plt.rcParams['figure.figsize'] = (20, 10)
sns.countplot(df['reading score'], palette = 'prism')
plt.title('Reading Score',fontsize = 20)
plt.show()
```



2.6.3 Writing Score Distribution

[9]:

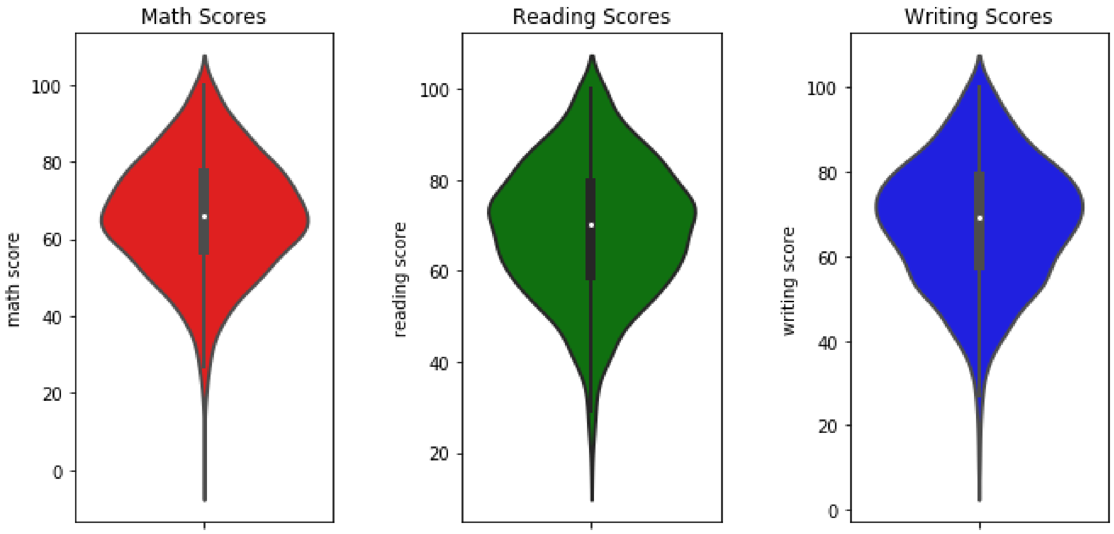
```
plt.rcParams['figure.figsize'] = (20, 10)
sns.countplot(df['writing score'], palette = 'prism')
plt.title('Writing Score',fontsize = 20)
plt.show()
```



2.6.4 Statistical Distribution

[10]:

```
plt.figure(figsize=(15,5))
plt.subplots_adjust(left=0.125, bottom=0.1, right=0.9, top=0.9,
                      wspace=0.5, hspace=0.2)
plt.subplot(141)
plt.title('Math Scores')
sns.violinplot(y='math score',data=df,color='r',linewidth=2)
plt.subplot(142)
plt.title('Reading Scores')
sns.violinplot(y='reading score',data=df,color='g',linewidth=2)
plt.subplot(143)
plt.title('Writing Scores')
sns.violinplot(y='writing score',data=df,color='b',linewidth=2)
plt.show()
```

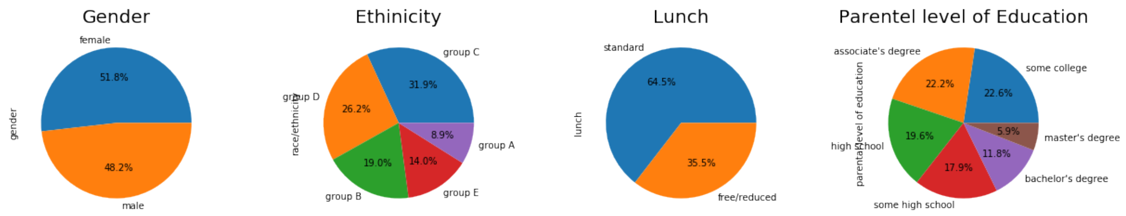


From the above plots, we can see that the maximum number of students have scored 60-80 in all three subjects i.e., math, reading and writing.

2.7 Plots of Categorical Variables

[11]:

```
plt.figure(figsize=(20,10))
plt.subplots_adjust(left=0.125, bottom=0.1, right=0.9, top=0.9,
                    wspace=0.5, hspace=0.2)
plt.subplot(141)
plt.title('Gender',fontsize = 20)
df['gender'].value_counts().plot.pie(autopct="%1.1f%%")

plt.subplot(142)
plt.title('Ethinicity',fontsize = 20)
df['race/ethnicity'].value_counts().plot.pie(autopct="%1.1f%%")

plt.subplot(143)
plt.title('Lunch',fontsize = 20)
df['lunch'].value_counts().plot.pie(autopct="%1.1f%%")

plt.subplot(144)
plt.title('Parentel level of Education',fontsize = 20)
df['parental level of education'].value_counts().plot.pie(autopct="%1.1f%%")
plt.show()
```



2.8 Observations:

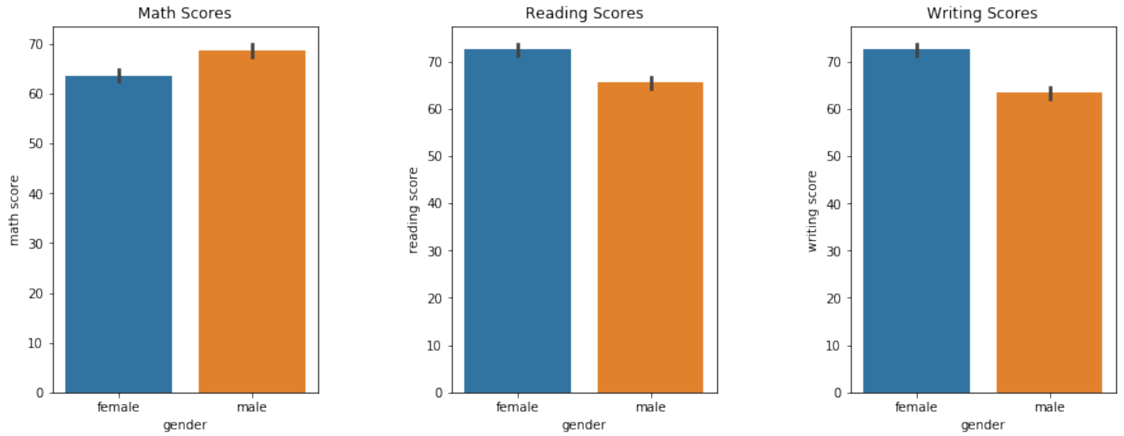
- The proportion of male and female are almost same
- Highest number of students belong to Group C ethinicity followed by Group D
- Highest proportion of the students have standard lunch
- Highest proportion of parentel level of Education is ‘Some college’, ‘associate’s degreee’ and ‘high school’

2.9 Division of data using different categories for subject scores:

2.9.1 Gender

[12]:

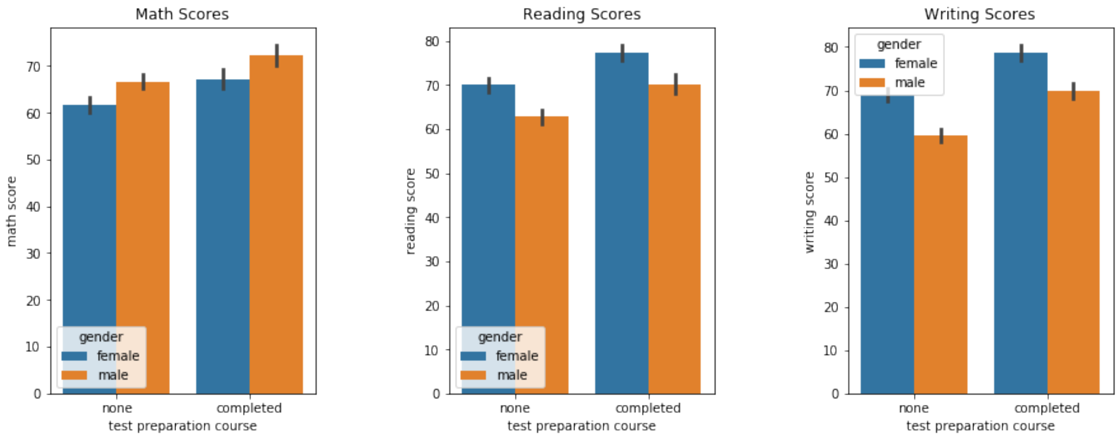
```
plt.figure(figsize=(15,5))
plt.subplots_adjust(left=0.125, bottom=0.1, right=0.9, top=0.9,
                    wspace=0.5, hspace=0.2)
plt.subplot(131)
plt.title('Math Scores')
sns.barplot(x="gender", y="math score", data=df)
plt.subplot(132)
plt.title('Reading Scores')
sns.barplot(x="gender", y="reading score", data=df)
plt.subplot(133)
plt.title('Writing Scores')
sns.barplot(x="gender", y="writing score", data=df)
plt.show()
```



We can see that male students scored higher in Maths where as female students scored higher in Reading and writing

2.9.2 Gender and Test Preparation Course

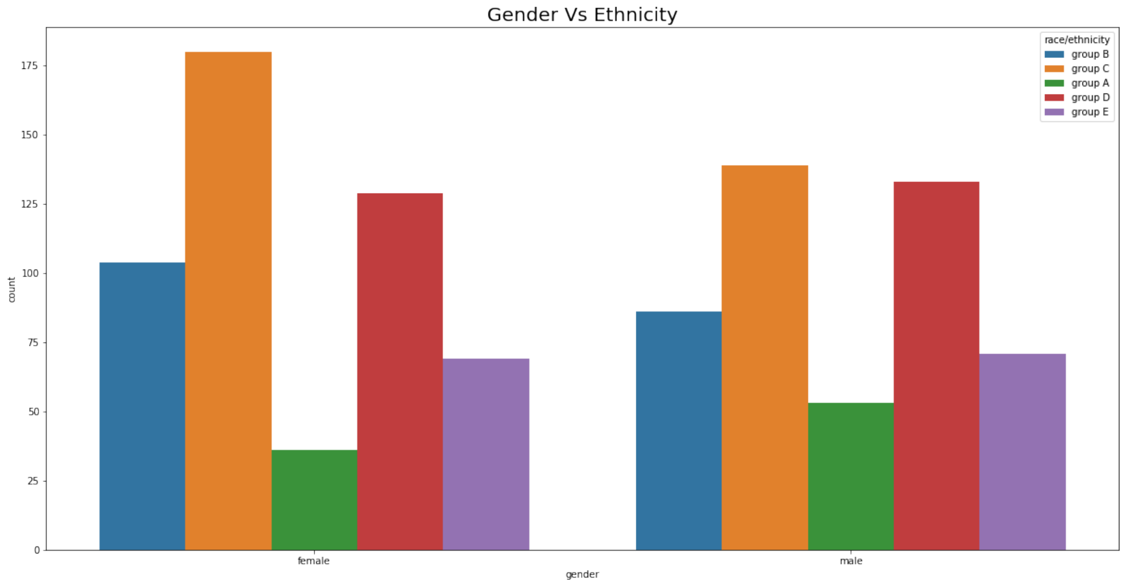
```
[13]: plt.figure(figsize=(15,5))
      plt.subplots_adjust(left=0.125, bottom=0.1, right=0.9, top=0.9,
                          wspace=0.5, hspace=0.2)
      plt.subplot(131)
      plt.title('Math Scores')
      sns.barplot(hue="gender", y="math score", x="test preparation course", data=df)
      plt.subplot(132)
      plt.title('Reading Scores')
      sns.barplot(hue="gender", y="reading score", x="test preparation course",␣
       ↪data=df)
      plt.subplot(133)
      plt.title('Writing Scores')
      sns.barplot(hue="gender", y="writing score", x="test preparation course",␣
       ↪data=df)
      plt.show()
```



So the students (male and female) who completed the test preparation course scored higher in all three subjects.

2.9.3 Gender and Ethnicity

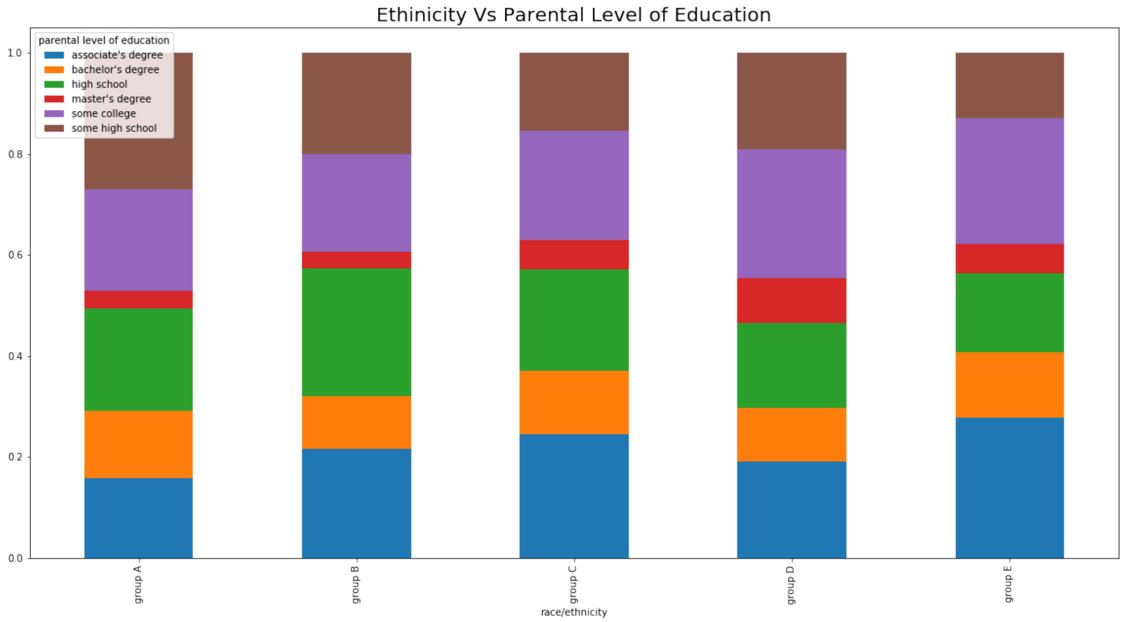
```
[14]: plt.title('Gender Vs Ethnicity',fontsize = 20)
      sns.countplot(x="gender", hue="race/ethnicity", data=df)
      plt.show()
```



2.9.4 Ethinicity and Parental Level of Education

[15]:

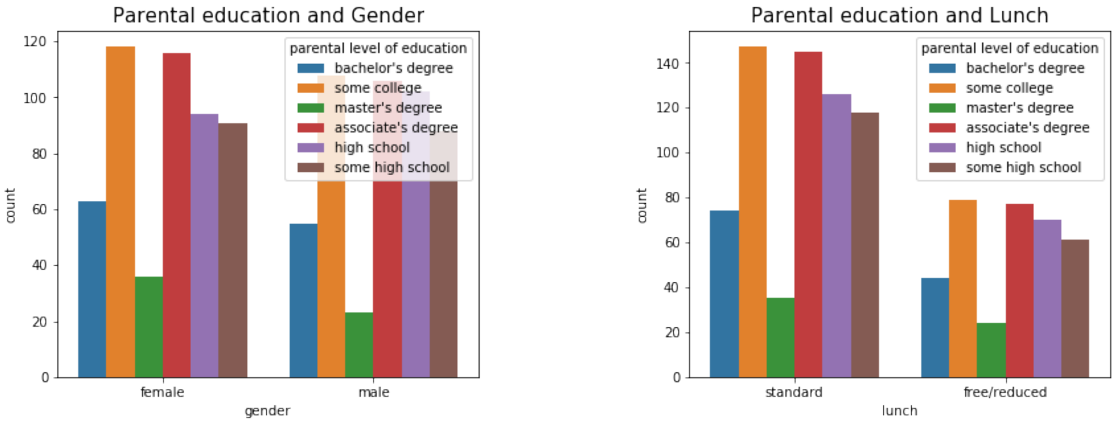
```
pr=pd.crosstab(df['race/ethnicity'],df['parental level of␣
 ↪education'],normalize=0)

pr.plot.bar(stacked=True)
plt.title('Ethinicity Vs Parental Level of Education',fontsize = 20)
plt.show()
```



2.9.5 Parental education, Lunch and Gender

[16]:

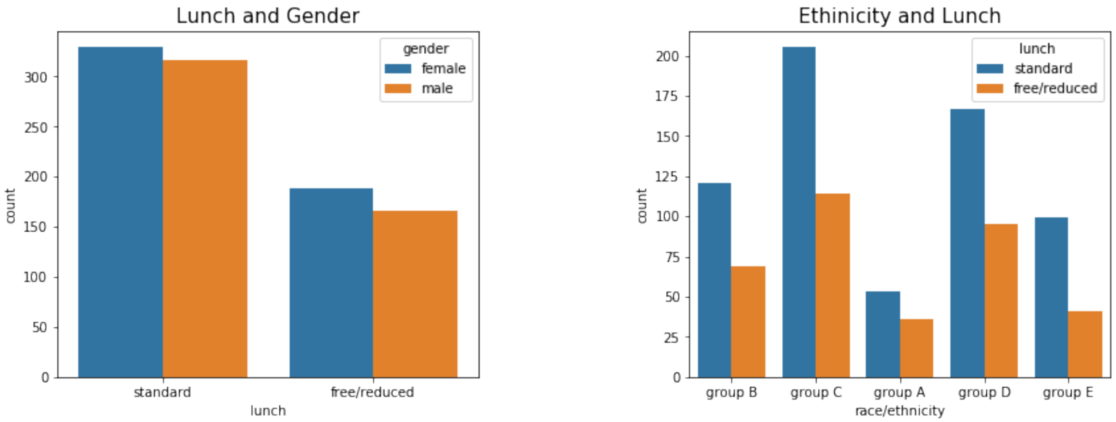
```
plt.figure(figsize=(40,10))
plt.subplots_adjust(left=0.125, bottom=0.1, right=0.9, top=0.9,
                      wspace=0.5, hspace=0.2)
plt.subplot(251)
plt.title('Parental education and Gender',fontsize=15)
sns.countplot(hue="parental level of education", x="gender", data=df)
plt.subplot(252)
plt.title('Parental education and Lunch',fontsize=15)
sns.countplot(hue="parental level of education", x="lunch", data=df)

plt.show()
```



2.9.6 Gender, Lunch and Ethenicity

[17]:

```
plt.figure(figsize=(40,10))
plt.subplots_adjust(left=0.125, bottom=0.1, right=0.9, top=0.9,
                      wspace=0.5, hspace=0.2)
plt.subplot(251)
plt.title('Lunch and Gender',fontsize=15)
sns.countplot(x="lunch", hue="gender", data=df)
plt.subplot(252)
plt.title('Ethinicity and Lunch',fontsize=15)
sns.countplot(x="race/ethnicity", hue="lunch", data=df)
plt.show()
```

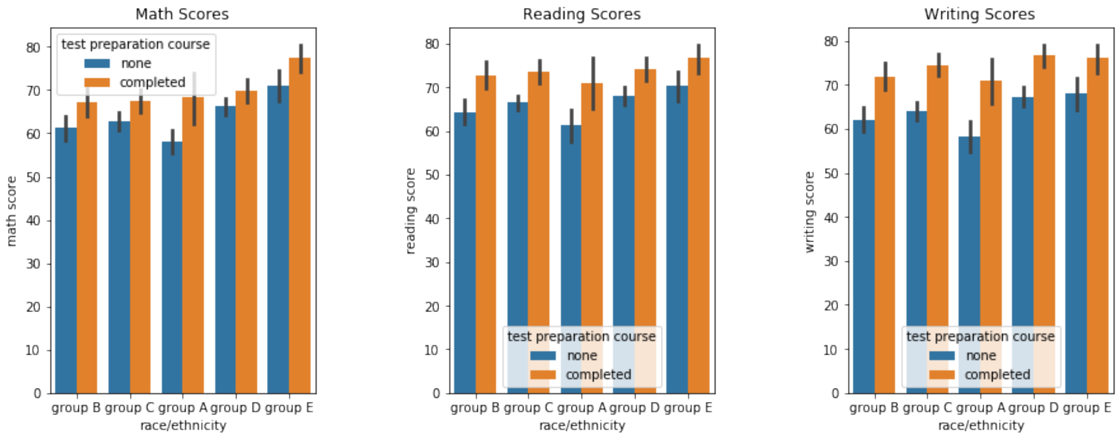


So, the students with standard lunch were better performers when compared with free lunch.

So, the students in group C performs better than other races.

2.9.7 Gender, Test Preparation Course and Ethnicity

[18]:
```
plt.figure(figsize=(15,5))
plt.subplots_adjust(left=0.125, bottom=0.1, right=0.9, top=0.9,
                      wspace=0.5, hspace=0.2)
plt.subplot(131)
plt.title('Math Scores')
sns.barplot(hue="test preparation course", y="math score", x="race/ethnicity",␣
 ↪data=df)
plt.subplot(132)
plt.title('Reading Scores')
sns.barplot(hue="test preparation course", y="reading score", x="race/
 ↪ethnicity", data=df)
plt.subplot(133)
plt.title('Writing Scores')
sns.barplot(hue="test preparation course", y="writing score", x= 'race/
 ↪ethnicity',data=df)

plt.show()
```



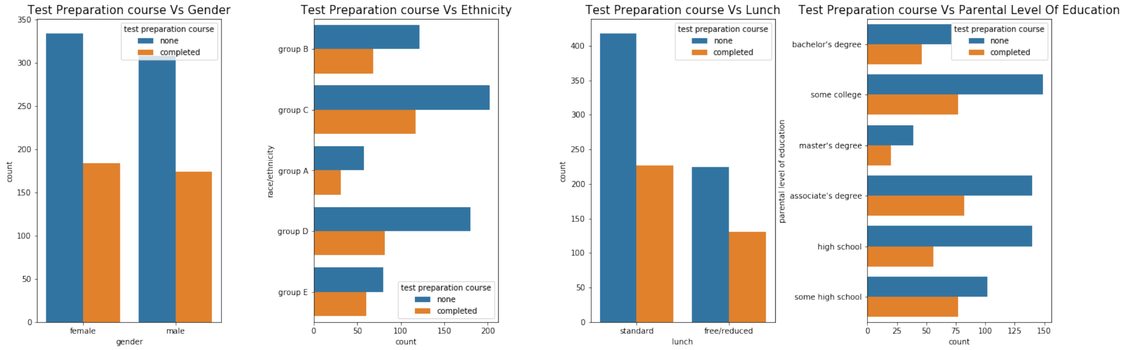
Highest number of Students who belongs to Group E has completed the test preperation course in Math and Reading and scored highest.

Highest number of Students who belongs to Group D and E has completed the test preperation course in Writing and scored highest.

2.9.8 Test Preparation Course vs. All Other Categorial Variables

[19]:

```
plt.figure(figsize=(30,15))
plt.subplots_adjust(left=0.125, bottom=0.1, right=0.9, top=0.9,
                      wspace=0.5, hspace=0.2)
plt.subplot(251)
plt.title('Test Preparation course Vs Gender',fontsize = 15)
sns.countplot(hue="test preparation course", x="gender", data=df)

plt.subplot(252)
plt.title('Test Preparation course Vs Ethnicity',fontsize = 15)
sns.countplot(hue="test preparation course", y="race/ethnicity", data=df)

plt.subplot(253)
plt.title('Test Preparation course Vs Lunch',fontsize = 15)
sns.countplot(hue="test preparation course", x="lunch", data=df)

plt.subplot(254)
plt.title('Test Preparation course Vs Parental Level Of Education',fontsize =␣
 ↪15)
sns.countplot(hue="test preparation course", y="parental level of education",␣
 ↪data=df)

plt.show()
```



2.9.9 Observations:

- Most of the students have not completed the test preparation course.
- Highest number Students who belong to group C ethinicity have completed the test preparation course.
- Standard lunch students have completed the test preparation course
- Students whos parental level of education is ‘some college, ’associate’s degree’, and high school have completed the test preparation course.

3 Statistical Study

3.0.1 To analyse the data in more deeper way, lets few new columns: Total marks, Percentage and Grades.

[20]:
```
df['total marks']=df['math score']+df['reading score']+df['writing score']
df['percentage']=df['total marks']/300*100
```

3.0.2 Grading System

- 85-100 : Grade A
- 70-84 : Grade B
- 55-69 : Grade C
- 35-54 : Grade D
- 0-35 : Grade E

[21]:
```
def determine_grade(scores):
    if scores >= 85 and scores <= 100:
        return 'Grade A'
    elif scores >= 70 and scores < 85:
        return 'Grade B'
    elif scores >= 55 and scores < 70:
        return 'Grade C'
    elif scores >= 35 and scores < 55:
        return 'Grade D'
    elif scores >= 0 and scores < 35:
        return 'Grade E'

df['grades']=df['percentage'].apply(determine_grade)
```

Now the columns “total marks”, “percentage” and “grades” are created

[22]:

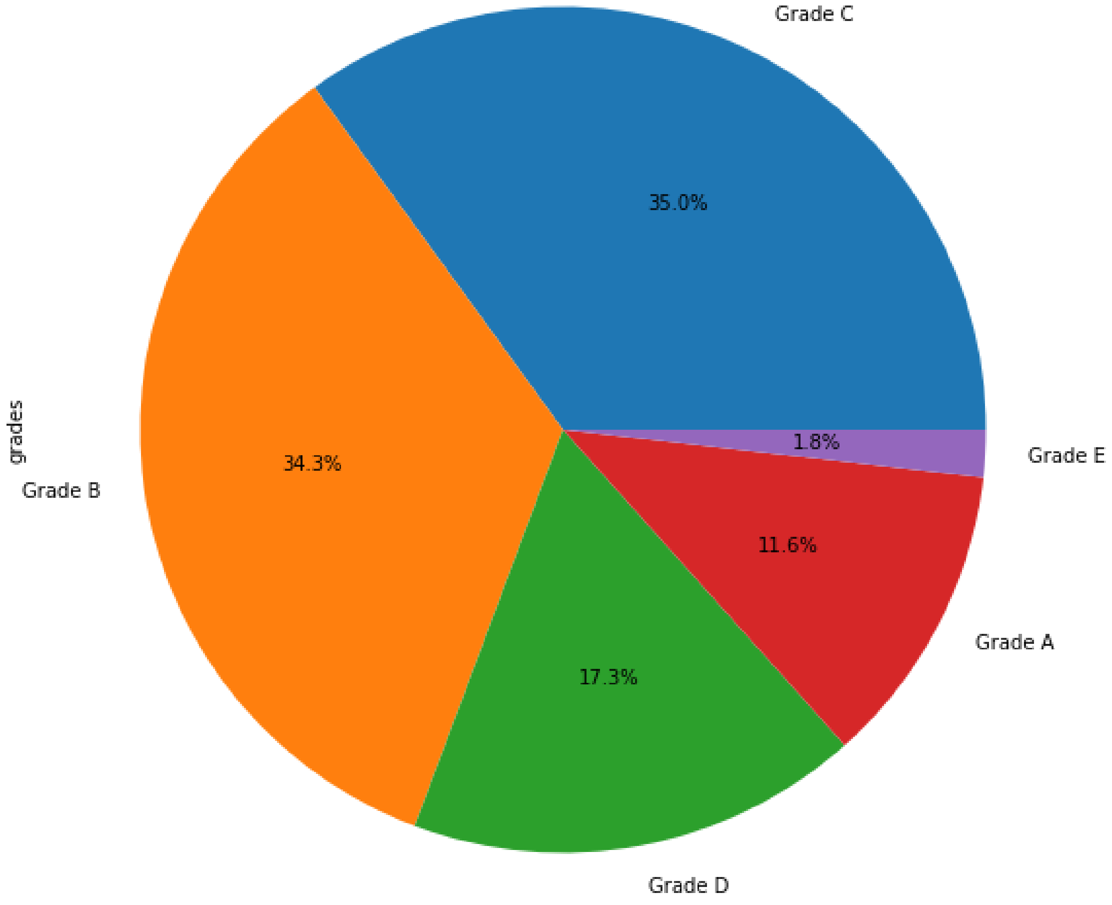
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64
 6   reading score                1000 non-null   int64
 7   writing score                1000 non-null   int64
 8   total marks                  1000 non-null   int64
 9   percentage                   1000 non-null   float64
 10  grades                       1000 non-null   object
dtypes: float64(1), int64(4), object(6)
memory usage: 86.1+ KB
```

3.0.3 Plot for grades of all the students

[23]:

```
df['grades'].value_counts().plot.pie(autopct="%1.1f%%")
plt.show()
```

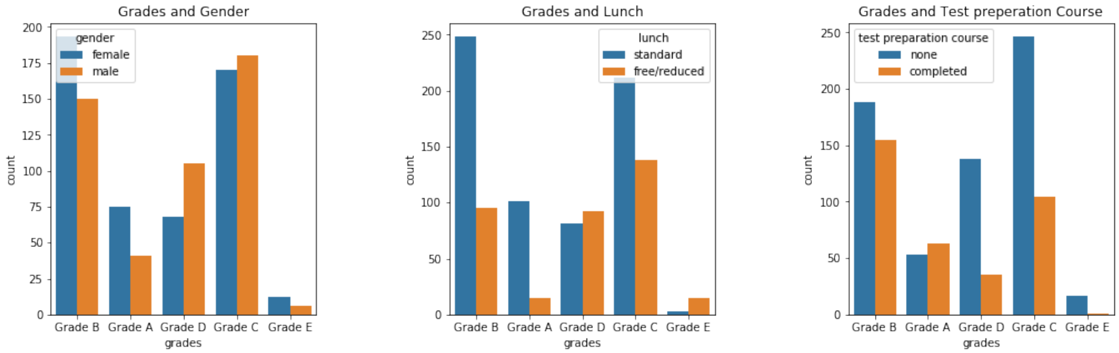


Most of the students got Grade B and Grade C.

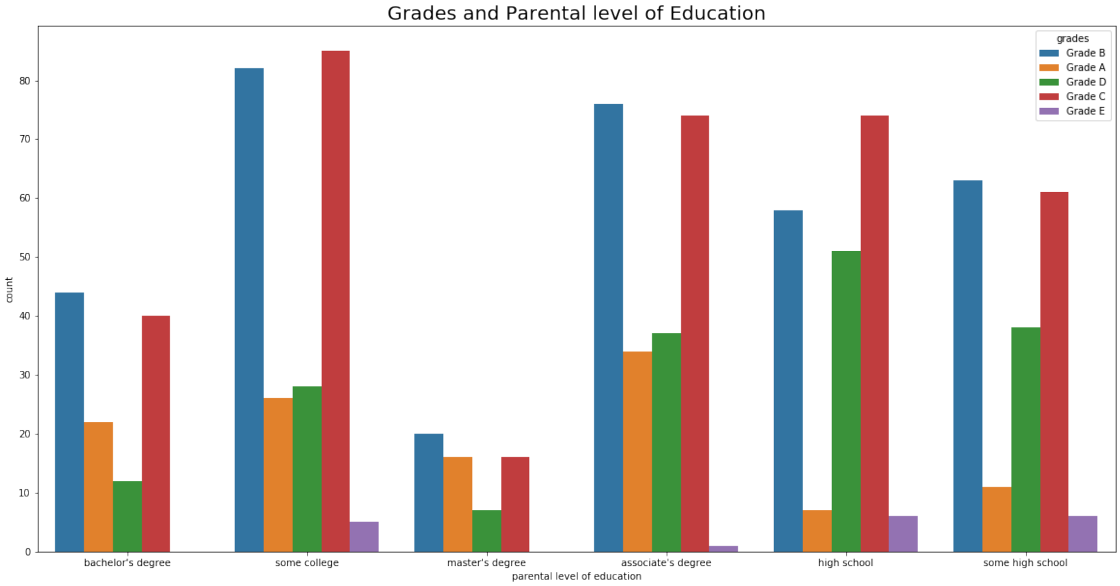
3.0.4 Grades vs. All Other Categorial Variables

[24]:
```
plt.figure(figsize=(30,10))
plt.subplots_adjust(left=0.125, bottom=0.1, right=0.9, top=0.9,
                    wspace=0.5, hspace=0.2)
plt.subplot(251)
plt.title('Grades and Gender')
sns.countplot(hue="gender", x="grades", data=df)

plt.subplot(252)
plt.title('Grades and Lunch')
sns.countplot(hue="lunch", x="grades", data=df)
```

```
plt.subplot(253)
plt.title('Grades and Test preperation Course')
sns.countplot(hue="test preparation course", x="grades", data=df)

plt.show()
```

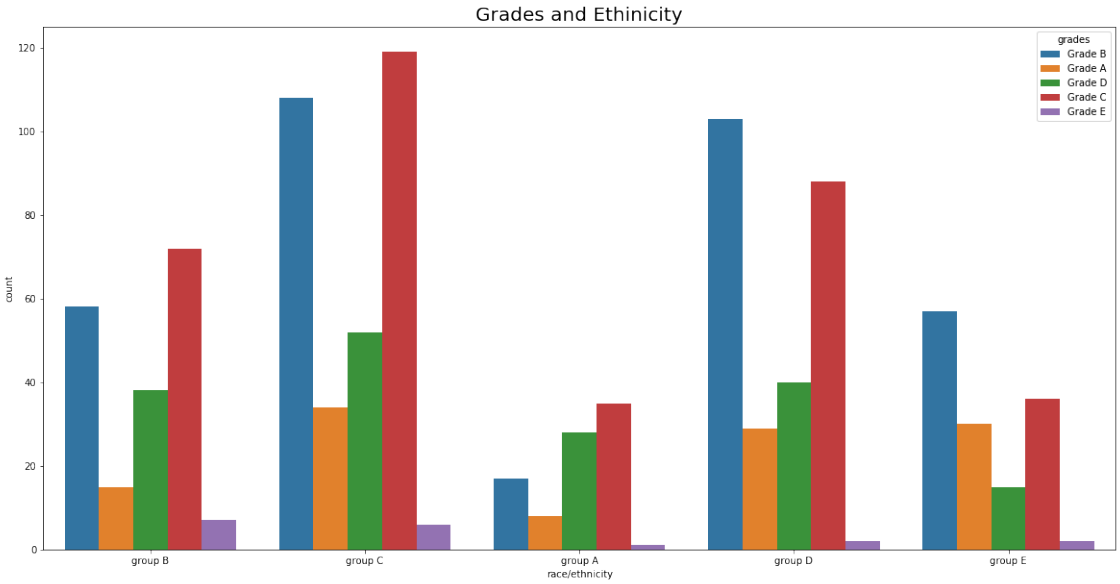


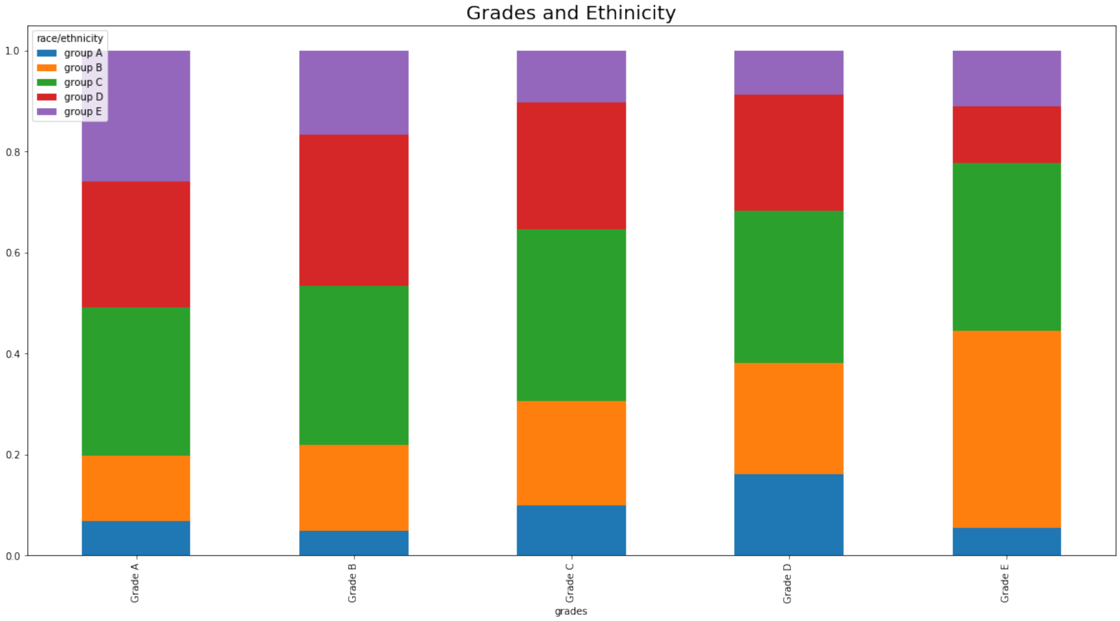
[25]:
```
plt.title('Grades and Parental level of Education',fontsize=20)
sns.countplot(x="parental level of education", hue="grades", data=df)
plt.show()
```



[26]:
```
plt.title('Grades and Ethinicity',fontsize=20)
sns.countplot(x="race/ethnicity", hue="grades", data=df)
```

```
gr=pd.crosstab(df['grades'],df['race/ethnicity'],normalize=0) #normalized values
gr.plot.bar(stacked=True)
plt.title('Grades and Ethinicity',fontsize=20)
plt.show()
```





3.0.5 Conclusion

- Most male students performed well in maths and females in literature, however considering the total scores females have an upper hand
- Parents with better degrees didn't send their children for any prep course.
- Most of the students got Grade B and Grade C.
- Most of the students have not completed the test preparation course.
- Highest number Students who belong to group C ethinicity have completed the test preparation course.
- Standard lunch students have completed the test preparation course
- Students whos parental level of education is 'some college, 'associate's degree', and high school have completed the test preparation course.