Histogram Analysis

Submitted By
Name: **Joan Jose M**
Register Number: **2248038**
Class: **2 MSc Data Science**

Sections

1. Importing Libraries
2. Loading the Dataset
3. Data Preprocessing
4. Exploratory Data Analysis
5. Conclusion

Dataset

- Taken from UG assignment

Importing Libraries

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

Loading the Dataset

In [2]:
```
df=pd.read_csv('ExamResults.csv')
df.head()
```

Out[2]:

	Sex	Country	ParentEducation	FinancialAid	TestPreparation	Science	Language	Communcation
0	Girl	UK	bachelor's degree	No Assistance	Minimum	75	74	75
1	Girl	USA	some college	No Assistance	Thorough	72	92	89
2	Girl	UK	master's degree	No Assistance	Minimum	93	97	94
3	Boy	India	associate's degree	Assisted	Minimum	50	59	45
4	Boy	USA	some college	No Assistance	Minimum	79	80	76

In [3]:
```
df.tail()
```

Out[3]:

	Sex	Country	ParentEducation	FinancialAid	TestPreparation	Science	Language	Communcation
494	Girl	UK	high school	No Assistance	Minimum	57	66	69
495	Boy	Australia	high school	No Assistance	Thorough	71	66	67
496	Girl	USA	some college	No Assistance	Minimum	57	50	53
497	Girl	Australia	some college	Assisted	Thorough	62	80	77
498	Girl	UK	some high school	No Assistance	Minimum	69	71	69

In [4]:
```
df.shape
```

Out[4]:
```
(499, 8)
```

In [5]:
```
df.columns
```

Out[5]:
```
Index(['Sex', 'Country', 'ParentEducation', 'FinancialAid', 'TestPreparation',
       'Science', 'Language', 'Communcation'],
      dtype='object')
```

About the Dataset:

The data contains data of students in an University

'ExamResults.csv' consists of 8 attributes that gives informations of Students:

1. Sex : The gender of the Student
2. Country : The Country that Student belongs to
3. ParentEductaion : The maximum Educational Qualification of the Parent
4. FinancialAid : The student has Financial Aid or not
5. TestPreparation : The preparation level of Student
6. Science : Score achieved in Science
7. Language : Score achieved in Language
8. Communication : Score achieved in Communication

Data Preprocessing

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 499 entries, 0 to 498
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Sex              499 non-null    object
 1   Country          499 non-null    object
 2   ParentEducation  499 non-null    object
 3   FinancialAid     499 non-null    object
 4   TestPreparation  499 non-null    object
 5   Science          499 non-null    int64 
 6   Language         499 non-null    int64 
 7   Communcation     499 non-null    int64 
dtypes: int64(3), object(5)
memory usage: 31.3+ KB
```

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
Sex               0
Country           0
ParentEducation   0
FinancialAid      0
TestPreparation   0
Science           0
Language          0
Communcation      0
dtype: int64
```

- No null values present in this data

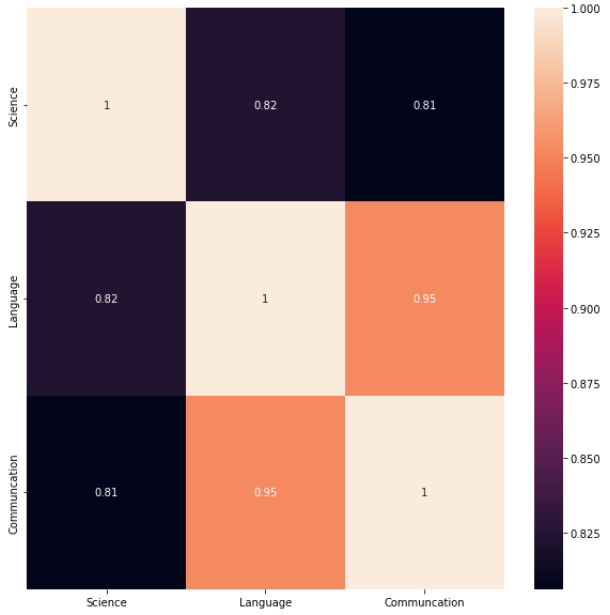
In [8]:
```
df.describe()
```

Out[8]:

	Science	Language	Communcation
count	499.000000	499.000000	499.000000
mean	68.655311	70.434870	68.234469
std	14.966275	14.664291	15.187950
min	3.000000	19.000000	11.000000
25%	60.000000	60.000000	58.000000
50%	69.000000	71.000000	69.000000
75%	79.000000	81.000000	79.000000
max	100.000000	99.000000	99.000000

In [9]:
```
plt.figure(figsize=(10,10))
sns.heatmap(df.corr(),annot=True)
```

Out[9]:
```
<AxesSubplot:>
```



- Language & Communication has correlation

In [10]:
```
# Calculate Total Marks obtained by the Students
df["TotalMarks"] = df["Science"] + df["Language"] + df["Communcation"]
```

In [11]:
```
# Calculate Percentage of Students
df["TotalPercentage"] = (df["TotalMarks"]/300)*100
df['TotalPercentage'] = df['TotalPercentage'].astype(int)
```

In [12]:
```
# Calculate Grade of Students
def grade(scores):
    if scores >=85 and scores <=100:
        return 'A'
    elif scores >=70 and scores <=85:
        return 'B'
    elif scores >=55 and scores <=70:
        return 'C'
    elif scores >=35 and scores <=55:
        return 'D'
    elif scores >=0 and scores <=35:
        return 'E'
df['Grades']=df['TotalPercentage'].apply(grade)
```

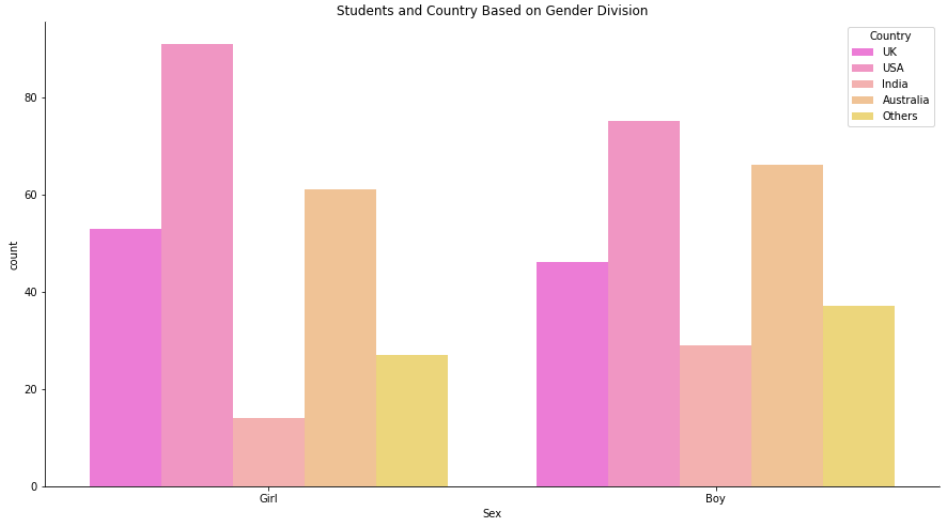
In [13]:
```
df.head()
```

Out[13]:

	Sex	Country	ParentEducation	FinancialAid	TestPreparation	Science	Language	Communcation	TotalMarks	TotalPercentage	Grades
0	Girl	UK	bachelor's degree	No Assistance	Minimum	75	74	75	224	74	B
1	Girl	USA	some college	No Assistance	Thorough	72	92	89	253	84	B
2	Girl	UK	master's degree	No Assistance	Minimum	93	97	94	284	94	A
3	Boy	India	associate's degree	Assisted	Minimum	50	59	45	154	51	D
4	Boy	USA	some college	No Assistance	Minimum	79	80	76	235	78	B

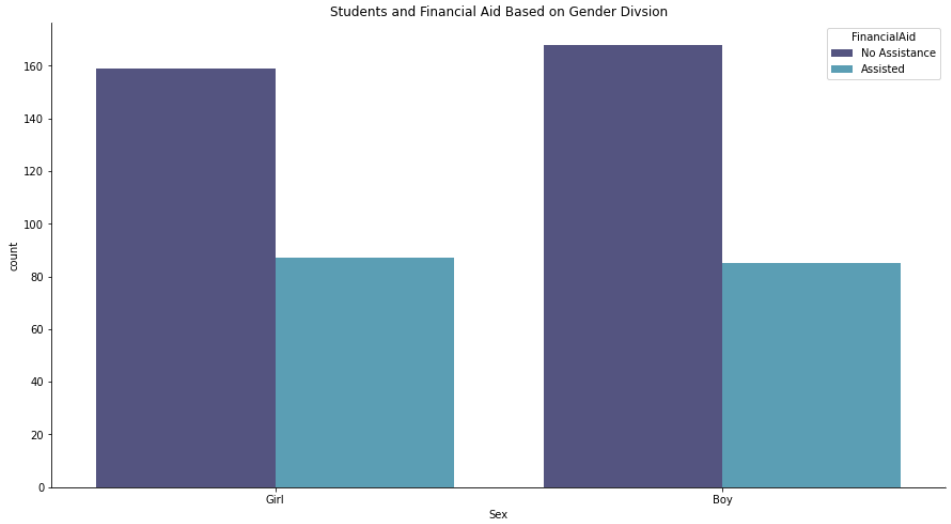
Exploratory Data Analysis (EDA)

In [14]:
```
plt.figure(figsize = [15,8])
sns.countplot(x = 'Sex', hue = "Country", palette = "spring", alpha = 0.7, data = df)
plt.title('Students and Country Based on Gender Division')
sns.despine()
```



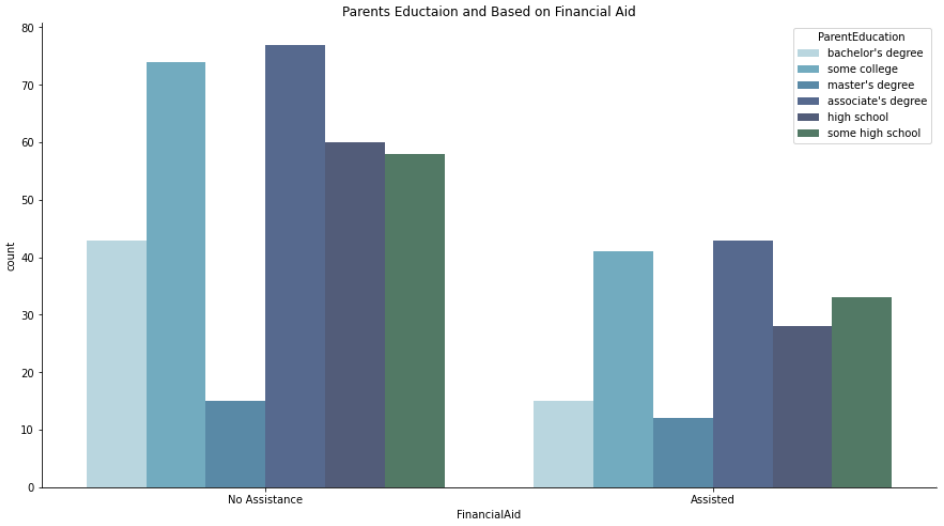
- Most of the Boys are from USA and least are from India same in the case of Girls as well

In [15]:
```
plt.figure(figsize = [15,8])
sns.countplot(x = 'Sex', hue = "FinancialAid", palette = "ocean", alpha = 0.7, data = df)
plt.title('Students and Financial Aid Based on Gender Divsion')
sns.despine()
```



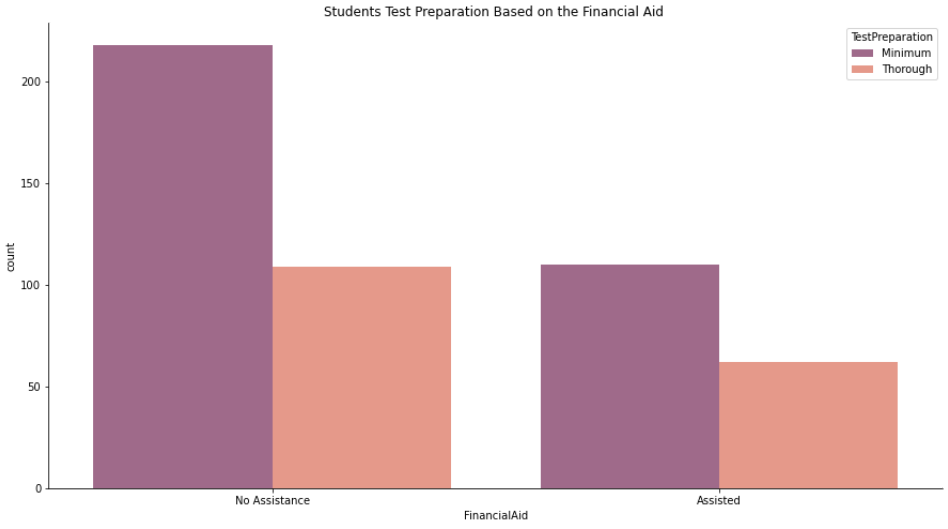
- Comparitevly both Boys and Girls are equally Financially Assisted while there is a light increase in Boys section who has not any Assistance than Girls

In [16]:
```
plt.figure(figsize = [15,8])
sns.countplot(x = 'FinancialAid', hue = "ParentEducation", palette = "ocean_r", alpha = 0.7, data = df)
plt.title('Parents Eductaion and Based on Financial Aid')
sns.despine()
```



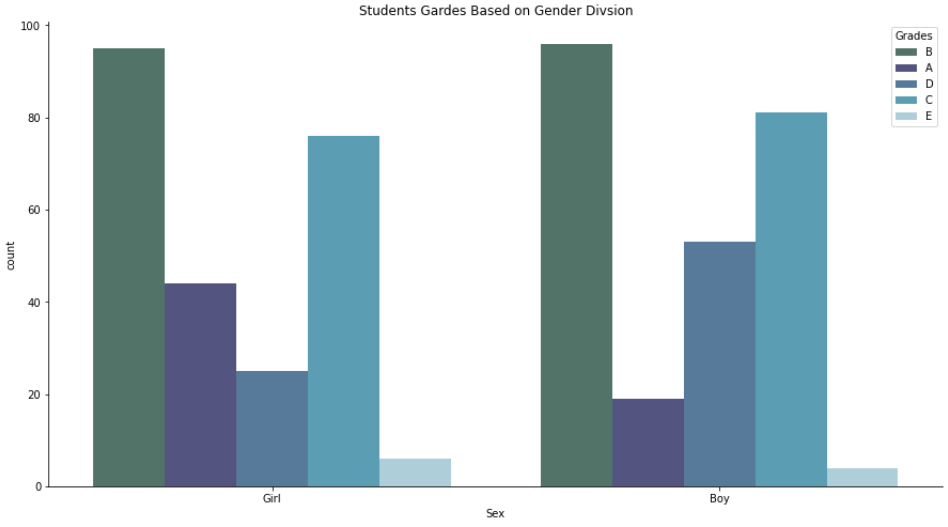
- Financial Assistance are given more to the Students whose parents have Associate Degree and Some college while Students whose parents have Master's degree

In [17]:
```
plt.figure(figsize = [15,8])
sns.countplot(x = 'FinancialAid', hue = "TestPreparation", palette = "rocket", alpha = 0.7, data = df)
plt.title('Students Test Preparation Based on the Financial Aid ')
sns.despine()
```



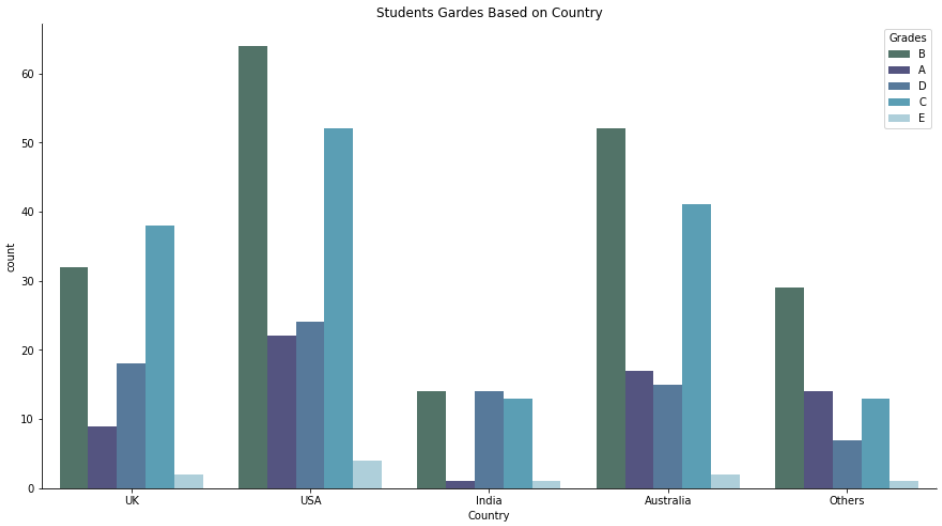
- Students who have Throughly prepared for exams are from Non Financialy assited group

In [19]:
```
plt.figure(figsize = [15,8])
sns.countplot(x = 'Sex', hue = "Grades", palette = "ocean", alpha = 0.7, data = df)
plt.title('Students Gardes Based on Gender Divsion')
sns.despine()
```



- Girls have more A and E grades than boys

In [20]:
```
plt.figure(figsize = [15,8])
sns.countplot(x = 'Country', hue = "Grades", palette = "ocean", alpha = 0.7, data = df)
plt.title('Students Gardes Based on Country')
sns.despine()
```



- India has least A grades while USA has highest A grades

Conclusion

- Most of the Boys are from USA and least are from India same in the case of Girls as well
- Comparitevly both Boys and Girls are equally Financially Assisted while there is a light increase in Boys section who has not any Assistance than Girls
- Financial Assistance are given more to the Students whose parents have Associate Degree and Some college while Students whose parents have Master's degree
- Students who have Throughly prepared for exams are from Non Financialy assited group
- Girls have more A and E grades than boys
- India has least A grades while USA has highest A grades
