Data Science Project
2021-22

Submitted to:
Dr. Ravindra Bhatt

Submitted by:
Kunika Sharma (191227)
Parul Sharma (191206)
Ria Mahajan (191236)

Contents

Data Set

(CollegePlace.csv)

We chose the data set from **KAGGLE.**

The link to the dataset is given below :

https://www.kaggle.com/tejashvi14/engineering-placements-prediction

The DataSet consists of the following parameters :

1. **AGE :** Age at the placement time
2. **GENDER :** Gender of the candidate.
3. **STREAM :** Engineering stream of the candidate. There are various streams like Computer Science, Information technology, Electrical And Electronics, etc.
4. **INTERNSHIPS :** Number of Internships undertaken during the course of studies, (not necessarily related to college studies rr stream.)
5. **CGPA :** CGPA till 6th semester.
6. **HOSTEL :** Whether a student lives in college accommodation or not. (values 1 if hostel facility availed and 0 if not)
7. **HISTORY OF BACKLOGS :** Whether a student ever had any backlogs during the course of study. (0 if no backlogs and 1 if there was any)
8. **PLACED OR NOT :** Target Variable. (value 1 means placed and 0 if not)

Data Cleaning

- **First upload the csv file and store it in a dataframe**

```
[1] from google.colab import files
    uploaded=files.upload()
```

```
Choose Files collegePlace.csv
• collegePlace.csv(application/vnd.ms-excel) - 109312 bytes, last modified: 10/23/2021 - 100% done
Saving collegePlace.csv to collegePlace.csv
```

```
[2] import pandas as pd
    df=pd.read_csv("collegePlace.csv")
```

- **Displaying DataSet (including anomalies)**

  (30 values from top)

```
print(df.head(30))
```

```
      Age  Gender  ... HistoryOfBacklogs  PlacedOrNot
0   22.0    Male  ...                 1          1.0
1   21.0  Female  ...                 1          1.0
2   22.0  Female  ...                 0          1.0
3   21.0    Male  ...                 1          1.0
4    NaN    Male  ...                 0          1.0
5   22.0    Male  ...                 0          0.0
6   21.0    Male  ...                 1          0.0
7   21.0    Male  ...                 0          0.0
8   21.0    Male  ...                 0          1.0
9   21.0  Female  ...                 0          0.0
10  22.0    Male  ...                 0          0.0
11  22.0  Female  ...                 1          1.0
12  21.0  Female  ...                 1          0.0
13  21.0    Male  ...                 1          1.0
14  21.0  Female  ...                 0          1.0
15  22.0    Male  ...                 0          1.0
16  22.0  Female  ...                 0          0.0
17  21.0    Male  ...                 0          0.0
18  21.0    Male  ...                 0          0.0
19  22.0    Male  ...                 0          0.0
20  22.0    Male  ...                 0          1.0
21  21.0    Male  ...                 0          0.0
22  22.0    Male  ...                 0          0.0
23  22.0    Male  ...                 1          0.0
24  22.0    Male  ...                 0          0.0
25  21.0    Male  ...                 0          1.0
26  22.0    Male  ...                 0          1.0
27  22.0    Male  ...                 0          1.0
28   NaN    Male  ...                 0          1.0
29  21.0    Male  ...                 0          1.0

[30 rows x 8 columns]
```

We can see there are some NULL values (NaN values) present indicating the dataset needs cleaning

- **Describe Dataset**

```
df.describe(include='O')
```

	Gender	Stream
count	2954	2948
unique	2	6
top	Male	Computer Science
freq	2467	772

Data has the maximum males and the stream as Computer Science.Hence we can replace null values with these values in their respective columns of Gender and Stream.

- **Dataset Info**

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2954 entries, 0 to 2965
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Age                2954 non-null   float64
 1   Gender             2954 non-null   object 
 2   Stream             2954 non-null   object 
 3   Internships        2954 non-null   int64  
 4   CGPA               2954 non-null   float64
 5   Hostel             2954 non-null   int64  
 6   HistoryOfBacklogs  2954 non-null   int64  
 7   PlacedOrNot        2954 non-null   float64
dtypes: float64(3), int64(3), object(2)
memory usage: 207.7+ KB
```

The dataset has different data types and some null values

- **Displaying Null Values In DataSet**

```
print(df[df.isnull().any(axis=1)])
```

```
       Age  Gender  ... HistoryOfBacklogs  PlacedOrNot
4      NaN    Male  ...                 0          1.0
27    22.0    Male  ...                 0          1.0
28     NaN    Male  ...                 0          1.0
82     NaN    Male  ...                 0          1.0
149   24.0    Male  ...                 0          1.0
167    NaN  Female  ...                 0          0.0
249   22.0    Male  ...                 0          NaN
279   21.0     NaN  ...                 1          1.0
334   22.0    Male  ...                 1          NaN
341   22.0    Male  ...                 0          1.0
398   21.0  Female  ...                 0          1.0
428    NaN  Female  ...                 0          1.0
437   21.0  Female  ...                 0          NaN
487   22.0    Male  ...                 0          0.0
508   21.0    Male  ...                 0          NaN
525   22.0    Male  ...                 1          0.0
526   22.0     NaN  ...                 0          0.0
533   22.0     NaN  ...                 0          0.0
603   21.0    Male  ...                 0          1.0
652   22.0     NaN  ...                 0          1.0
757   24.0    Male  ...                 0          1.0
797   22.0     NaN  ...                 0          1.0
882   22.0    Male  ...                 1          1.0
914   22.0  Female  ...                 0          NaN
971   21.0    Male  ...                 0          1.0
1056  22.0    Male  ...                 0          NaN
1082   NaN    Male  ...                 0          0.0
1102  21.0    Male  ...                 1          1.0
1124  22.0     NaN  ...                 0          1.0
1151  21.0    Male  ...                 0          NaN

[30 rows x 8 columns]
```

The above code shows that 30 rows in the Dataset have Null values somewhere in the 8 different columns.

- **Replacing NULL Values in "Age" Column with Mean of Age**

```
[ ] mAge=df['Age'].mean()
    df['Age'].fillna(mAge,inplace=True)
    print(df[df.isnull().any(axis=1)])
```

```
       Age  Gender  ... HistoryOfBacklogs  PlacedOrNot
27    22.0    Male  ...                 0          1.0
149   24.0    Male  ...                 0          1.0
249   22.0    Male  ...                 0          NaN
279   21.0     NaN  ...                 1          1.0
334   22.0    Male  ...                 1          NaN
341   22.0    Male  ...                 0          1.0
398   21.0  Female  ...                 0          1.0
437   21.0  Female  ...                 0          NaN
487   22.0    Male  ...                 0          0.0
508   21.0    Male  ...                 0          NaN
525   22.0    Male  ...                 1          0.0
526   22.0     NaN  ...                 0          0.0
533   22.0     NaN  ...                 0          0.0
603   21.0    Male  ...                 0          1.0
652   22.0     NaN  ...                 0          1.0
757   24.0    Male  ...                 0          1.0
797   22.0     NaN  ...                 0          1.0
882   22.0    Male  ...                 1          1.0
914   22.0  Female  ...                 0          NaN
971   21.0    Male  ...                 0          1.0
1056  22.0    Male  ...                 0          NaN
1102  21.0    Male  ...                 1          1.0
1124  22.0     NaN  ...                 0          1.0
1151  21.0    Male  ...                 0          NaN

[24 rows x 8 columns]
```

The missing age values have been replaced with the mean leaving 24 Null Values in the dataset.

- **Rows with NULL CGPA and PlacedOrNot Values are dropped from Dataset**

```
df.dropna(subset=['CGPA' ,'PlacedOrNot'],inplace=True)
print(df[df.isnull().any(axis=1)])
```

```
      Age Gender  ... HistoryOfBacklogs  PlacedOrNot
279   21.0    NaN  ...                 1          1.0
487   22.0   Male  ...                 0          0.0
526   22.0    NaN  ...                 0          0.0
533   22.0    NaN  ...                 0          0.0
603   21.0   Male  ...                 0          1.0
652   22.0    NaN  ...                 0          1.0
757   24.0   Male  ...                 0          1.0
797   22.0    NaN  ...                 0          1.0
882   22.0   Male  ...                 1          1.0
971   21.0   Male  ...                 0          1.0
1102  21.0   Male  ...                 1          1.0
1124  22.0    NaN  ...                 0          1.0

[12 rows x 8 columns]
```

Rows having Null CGPA and PlacedOrNot Values are dropped. We can now see that there are 6 rows which have Missing Values in the "Gender" and "Stream" columns that need to be replaced as the final steps in the data-cleaning process.

- **Replace missing "Gender" column values with "Male"**

```
df['Gender'].fillna("Male",inplace=True)
print(df[df.isnull().any(axis=1)])
```

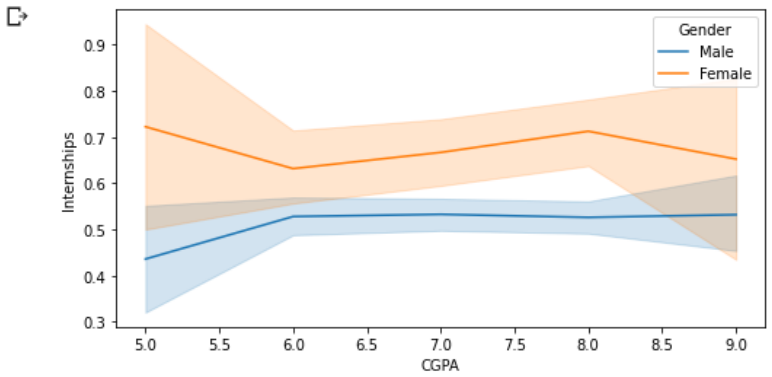
```
      Age Gender Stream  ...  Hostel  HistoryOfBacklogs  PlacedOrNot
487   22.0   Male    NaN  ...       0                  0          0.0
603   21.0   Male    NaN  ...       0                  0          1.0
757   24.0   Male    NaN  ...       1                  0          1.0
882   22.0   Male    NaN  ...       0                  1          1.0
971   21.0   Male    NaN  ...       0                  0          1.0
1102  21.0   Male    NaN  ...       1                  1          1.0

[6 rows x 8 columns]
```

All missing gender values replaced. Null values are now left only in the "Stream" Column.

Plots & Analysis

- **Features with respect to Internships:**
  - ➢ Internship Vs CGPA for different genders

```
import matplotlib.pyplot as plt
import seaborn as sns
colg = df.copy()
colg['Internships'] = colg['Internships'].apply(lambda x: 1 if x>0 else x)
plt.figure(figsize=(8,4))
sns.lineplot(x='CGPA',y='Internships',data=colg,hue='Gender')
plt.show()
```

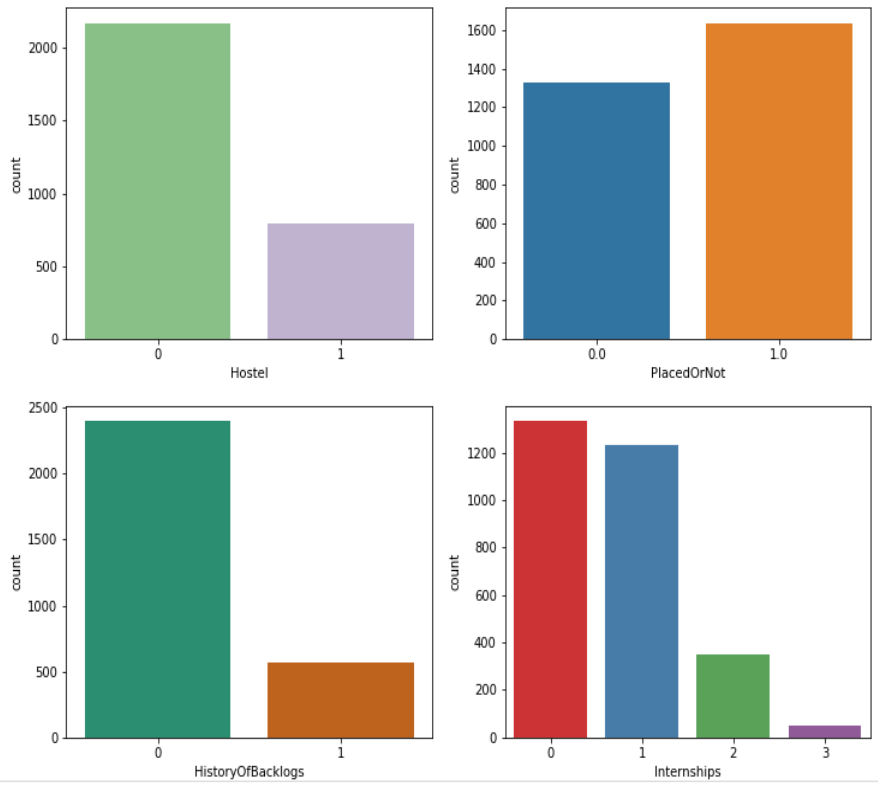


Analysis:

- ➔ Females do have a slight edge over Male in securing an internship irrespective of the CGPA
- ➔ In Male candidates, the higher the CGPA higher are the chances of getting an internship
- ➔ While this trend is not with females where chances increase initially and then slopes down.

➢ Plots with respect to counts

```
plt.subplots(2,2,figsize=(12,10))
plt.subplot(221)
sns.countplot(data=df, x='Hostel',palette='Accent')
plt.subplot(222)
sns.countplot(data=df, x='PlacedOrNot')
plt.subplot(223)
sns.countplot(data=df, x='HistoryOfBacklogs',palette='Dark2')
plt.subplot(224)
sns.countplot(data=df, x='Internships',palette='Set1')
plt.show()
```

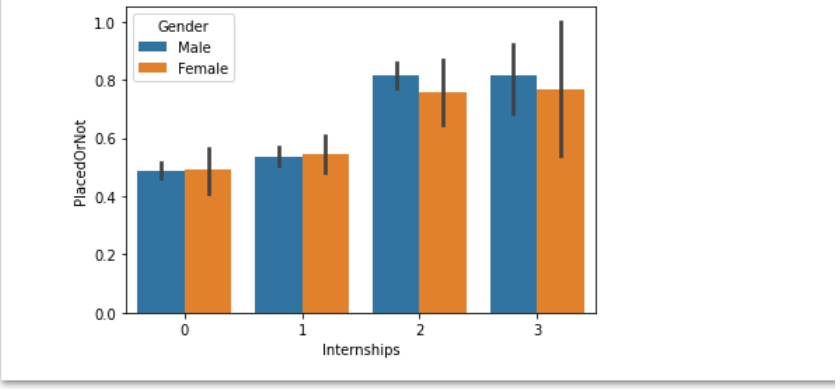


Analysis:

- ➔ Nearly 50% of students haven't done any internship while among the rest majority of them have done atleast 1 internship and some have even done 3.
- ➔ About 25% of the total students were residing in hostels.

➢ Internship Vs PlacedOrNot

```
sns.barplot(data=df,x='Internships',y='PlacedOrNot',hue='Gender')
plt.show()
```

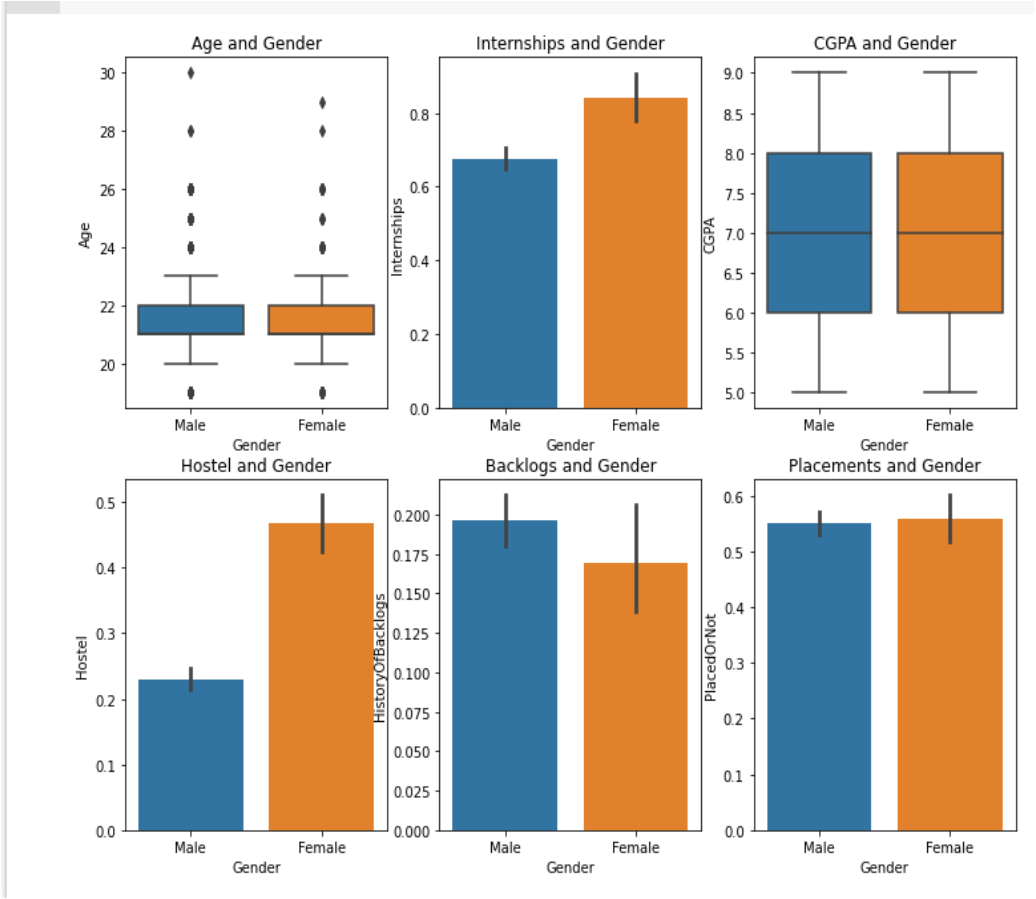


Analysis:

➔ If you have done more than 1 internship the chances of getting placed are actually good.

- **Features with respect to Gender**

```
plt.subplots(2,3,figsize=(12,10))
plt.subplot(231)
plt.title('Age and Gender')
sns.boxplot(y='Age',x='Gender',data=df)
plt.subplot(232)
plt.title('Internships and Gender')
sns.barplot(x='Gender',y='Internships',data=df)
plt.subplot(233)
plt.title('CGPA and Gender')
sns.boxplot(x='Gender',y='CGPA',data=df)
plt.subplot(234)
plt.title('Hostel and Gender')
sns.barplot(x='Gender',y='Hostel',data=df)
plt.subplot(235)
plt.title('Backlogs and Gender')
sns.barplot(x='Gender',y='HistoryOfBacklogs',data=df)
plt.subplot(236)
plt.title('Placements and Gender')
sns.barplot(x='Gender',y='PlacedOrNot',data=df)
plt.show()
```

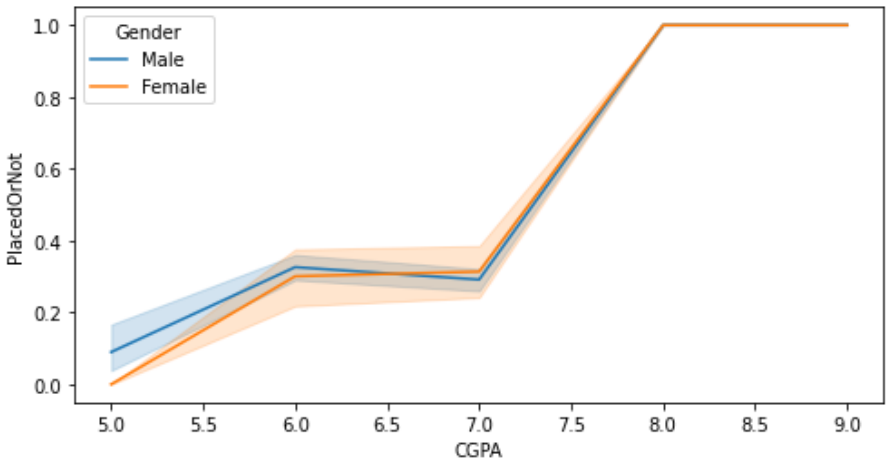


Analysis:

- ➔ The median value of Internships by Female is quite higher than that of the Male one's.
- ➔ Male as well as the Female candidates have the same median value of 7 CGPA.
- ➔ Females have more tendency to stay in hostels than the Males.
- ➔ On an average Male candidate has more backlogs then the Female ones.
- ➔ With respect to Placements, the chances of Female candidates being placed is just a fraction more than that of the Male candidates.

- **Features with respect to CGPA.**

```
[ ]  plt.figure(figsize=(8,4))
     sns.lineplot(x='CGPA',y='PlacedOrNot',data=df,hue='Gender')
     plt.show()
```



Analysis:

➔ If one scores more than 8 CGPA the chances of getting placed are actually extremely good than the rest, irrespective of their Gender.