**FEE vs ADMISSION**

**By**
Jaya Bhargav Burugupalli
Morampudi Ramya
Akkiraju Sri Pragna

**Group Number:1**

## OVERVIEW

The Data contains the information of a college or university. It includes weather it is public vs private sector based on locations and the students joining in the colleges with amount of tuition fees they were paying to the colleges or universities for that we are implementing the machine learning models to variate on public versus private. And finding the different ways to interpret the results. We are using different types of techniques.

The Techniques we are implementing the techniques of machine learning and statistical methods to come out an informative result. Based on these results we will get the expenditures of student in colleges. And we will find the graduates from public and private universities which has the higher and cost of rooms for public and private. These insights we can get from our analysis part.

## TOOLS AND TECHNIQUES

The tools we are using for our analysis part are IDE Jupyter Notebook. It is used in such a way that to generate graphs, plots and to omit the null values by cleaning the data. In place of the null values, we have done mean values based on the given data. Graphs such as Histograms, Line Plots, Box Plots, using some packages of matplotlib.

Python Libraries like NumPy, Pandas are used to analyze the data. We also performed EDA to build some of the models like linear regression, multi linear regression.

## DESCRIPTION OF DATASET

### A. *Dataset Link*

*https://docs.google.com/spreadsheets/d/1fTsB9vGzfGjKwFZAPZGlbgX0sXKsxe8U/edit?usp=sharing&ouid=106712751114517514446&rtpof=true&sd=true*

### B. *More Details of the Dataset*

The data set contains with 1302 records of universities and colleges.

The data set have 20 columns

1) College Name: It is string format. It contains names of colleges and universities.
2) State: It is categorical data contains the state and the universities' locations.
3) Public/Private: This column indicates the college weather it is public or private limited. And it is indicated as 1's and 2's, 1-Public and 2-Private.
4) Application received to colleges (numeric data)
5) Applications accepted by the colleges (numeric data)
6) New students enrolled to colleges (numeric data)
7) New students from top 10 universities (numeric data)
8) New students from top 25 universities (numeric data)
9) Two columns are of undergraduates contains count (numeric data)
10) Two columns are having the fees of in state tuition fees and out state tuition fees of universities (numeric data).

11) Add. Fees: this column contains the additional fees paid by students (numeric data).
12) Estim. Book costs: It is having the estimated price of books (numeric data)
13) Estim. personal: It have personal expenditures (numeric data)
14) Having the ratio of faculty and PHD (numeric data)
15) Having the ratio of students and faculty (numeric data)
16) Graduation rate: It have results of graduation rate (numeric data).

*Exploratory Data Analysis*

Before we proceed to the EDA there are few steps need to follow, they are:

*a). Importing Required libraries*

```
In [2]: M import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        import warnings
        warnings.filterwarnings("ignore", category=FutureWarning)
        %matplotlib inline
        sns.set(color_codes=True)
```

*b). Loading data into data frame*

```
M data = pd.read_excel('Dataset.xlsx')
  data.head(5)
```

| | College Name | State | Public (1)/ Private (2) | # appli. rec'd | # appl. accepted | # new stud. enrolled | % new stud. from top 10% | % new stud. from top 25% | # FT undergrad | # PT undergrad | in-state tuition | out-of-state tuition | room | board | add. fees | estim. book costs | estim. personal $ | % fa w/PH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Alaska Pacific University | AK | 2 | 193.0 | 146.0 | 55.0 | 16.0 | 44.0 | 249.0 | 869.0 | 7560.0 | 7560.0 | 1620.0 | 2500.0 | 130.0 | 800.0 | 1500.0 | 76. |
| 1 | University of Alaska at Fairbanks | AK | 1 | 1852.0 | 1427.0 | 928.0 | NaN | NaN | 3885.0 | 4519.0 | 1742.0 | 5226.0 | 1800.0 | 1790.0 | 155.0 | 650.0 | 2304.0 | 67. |
| 2 | University of Alaska Southeast | AK | 1 | 146.0 | 117.0 | 89.0 | 4.0 | 24.0 | 492.0 | 1849.0 | 1742.0 | 5226.0 | 2514.0 | 2250.0 | 34.0 | 500.0 | 1162.0 | 39. |
| 3 | University of Alaska at Anchorage | AK | 1 | 2065.0 | 1598.0 | 1162.0 | NaN | NaN | 6209.0 | 10537.0 | 1742.0 | 5226.0 | 2600.0 | 2520.0 | 114.0 | 580.0 | 1260.0 | 48. |
| 4 | Alabama Agri. & Mech. Univ. | AL | 1 | 2817.0 | 1920.0 | 984.0 | NaN | NaN | 3958.0 | 305.0 | 1700.0 | 3400.0 | 1108.0 | 1442.0 | 155.0 | 500.0 | 850.0 | 53. |

*c). Checking the types of data:*

```
data.dtypes
```

```
College Name                  object
State                         object
Public (1)/ Private (2)        int64
# appli. rec'd               float64
# appl. accepted             float64
# new stud. enrolled         float64
% new stud. from top 10%     float64
% new stud. from top 25%     float64
# FT undergrad               float64
# PT undergrad               float64
in-state tuition             float64
out-of-state tuition         float64
room                         float64
board                        float64
add. fees                    float64
estim. book costs            float64
estim. personal $            float64
% fac. w/PHD                 float64
stud./fac. ratio             float64
Graduation rate              float64
dtype: object
```

```
data.shape
```

```
(1302, 20)
```

*d). Dropping irrelevant column and renaming the columns*

As we do not have any irrelevant columns and all the columns have been named properly, we can skip this step.

*e). Dropping the duplicate rows*

After running the code to identify the duplicates. we have found there are no duplicates in the data. So, we do not need to delete any.

*f). Dropping the missing or null values*

Let's find if there are any null values:

We found there are over 500 null values in total consider different rows and columns like room, new students from top 10, new students from top 25 etc. If there are under 100 values for data set of 1000 we can ignore and delete the values but for this case, it shows effect on the analysis of the data so let's replace these null values with mean values.

```
#filling the data with mean values
#fillna is the command used in python to fill null values
data["# appli. rec'd"]=data["# appli. rec'd"].fillna(data["# appli. rec'd"].mean())
data["# appl. accepted"]=data["# appl. accepted"].fillna(data["# appl. accepted"].mean())
data["# new stud. enrolled"]=data["# new stud. enrolled"].fillna(data["# new stud. enrolled"].mean())
data["% new stud. from top 10%"]=data["% new stud. from top 10%"].fillna(data["% new stud. from top 10%"].mean())
data["% new stud. from top 25%"]=data["% new stud. from top 25%"].fillna(data["% new stud. from top 25%"].mean())
data["# FT undergrad"]=data["# FT undergrad"].fillna(data["# FT undergrad"].mean())
data["# PT undergrad"]=data["# PT undergrad"].fillna(data["# PT undergrad"].mean())
data["in-state tuition"]=data["in-state tuition"].fillna(data["in-state tuition"].mean())
data["out-of-state tuition"]=data["out-of-state tuition"].fillna(data["out-of-state tuition"].mean())
data["room"]=data["room"].fillna(data["room"].mean())
data["board"]=data["board"].fillna(data["board"].mean())
data["add. fees"]=data["add. fees"].fillna(data["add. fees"].mean())
data["estim. book costs"]=data["estim. book costs"].fillna(data["estim. book costs"].mean())
data["estim. personal $"]=data["estim. personal $"].fillna(data["estim. personal $"].mean())
data["% fac. w/PHD"]=data["% fac. w/PHD"].fillna(data["% fac. w/PHD"].mean())
data["stud./fac. ratio"]=data["stud./fac. ratio"].fillna(data["stud./fac. ratio"].mean())
data["Graduation rate"]=data["Graduation rate"].fillna(data["Graduation rate"].mean())
```

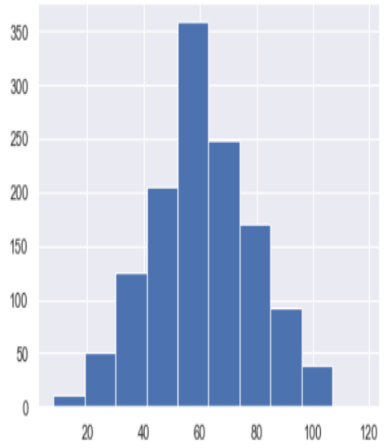
*g). Plotting the data*

*Histogram:* We need to find totally how many students after they got acceptance letter from university have enrolled into the course and finished the graduation.

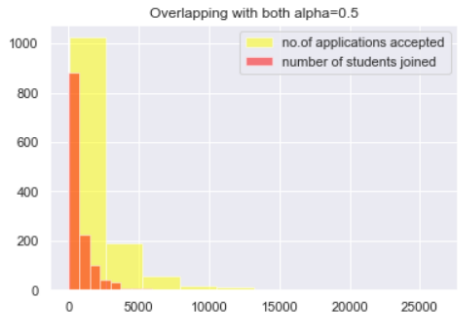
For that we can do two histograms:
1). One with graduation rate
2). One with the application acceptance and new student enrolled.

```
import matplotlib.pyplot as plt
plt.hist(data["Graduation rate"])
```

```
(array([ 11.,  51., 125., 205., 359., 248., 171.,  92.,  39.,   1.]),
 array([  8.,  19.,  30.,  41.,  52.,  63.,  74.,  85.,  96., 107., 118.]),
 <BarContainer object of 10 artists>)
```




```
plt.hist(data["# appl. accepted"],
         alpha=0.5, # the transaparency parameter
         label='no.of applications accepted',
        color='yellow')

plt.hist(data["# new stud. enrolled"],
         alpha=0.5,
         label='number of students joined',
        color='red')

plt.legend(loc='upper right')
plt.title('Overlapping with both alpha=0.5')
plt.show()
```

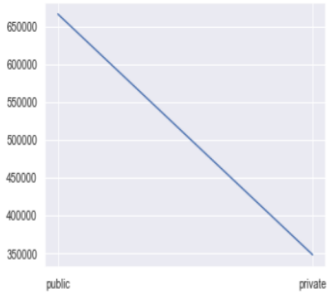



If we observe this plot closely, we can see in the first overlap plot. Out of 1000 applications that have been accepted. We see around 850 that have enrolled in the program. Similar is the case as we proceed there is a **huge** variation the acceptance to enrolled. We are not sure, yet the fee is the key component of the affect yet.

*Line plot:*

From the histogram we have seen there is a huge variation of student admission to the college even after the admission has been granted, now to move to the next steps let's see whether the college being public or private shows any effect on the student admission. To understand this data better, we are doing a line plot between total students enrolled for public universities and total students enrolled for private universities.

```
In [116]: total_public=data.loc[data['Public (1)/ Private (2)'] == 1,"# new stud. enrolled"].sum()
          total_public

Out[116]: 666077.5219737857

In [ ]:

In [117]: total_private=data.loc[data['Public (1)/ Private (2)'] == 2,"# new stud. enrolled"].sum()
          total_private

Out[117]: 348024.8804934464

In [118]: x=[total_public,total_private]
          y=["public","private"]
          plt.plot(y,x)
          plt.show()
```



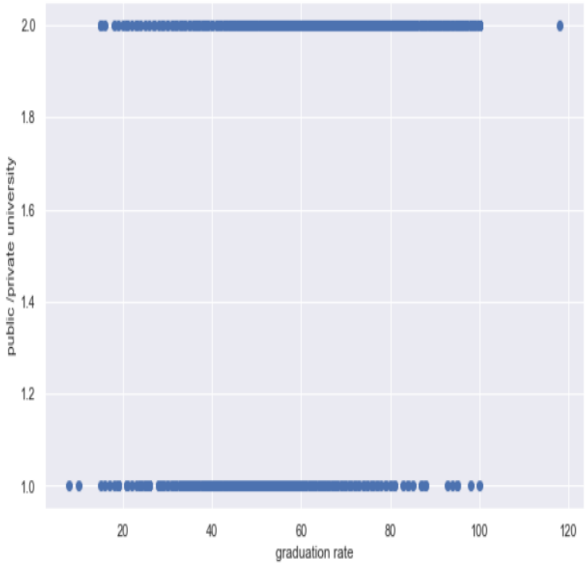

If we observe the plot carefully, we can see more students enrolling for public universities rather than private universities.

*Scatter plot:*

From the above line we have assumed that most students choose public university over private universities, but we need to check any other variables are showing any effect on students choosing public university like graduation rate. May be students are choosing public as they are more people graduating from public rather than private.

So, we are doing scatter plot between public/private university and graduation rate.

```
# Plotting a scatter plot
fig, ax = plt.subplots(figsize=(10,6))
ax.scatter( data['Graduation rate'],data['Public (1)/ Private (2)'])
ax.set_ylabel('public /private university')
ax.set_xlabel('graduation rate')
plt.show()
```




Through the plot we can clearly say both the university have equal graduation rate except one outlier with 118 of Cazenovia College located at New York City.

## *DETECTING THE OUTLIERS*

Now, we have better understood the data so before we could test the data for any variable dependencies. We need to detect the outliers of the data, so they do not affect the variable dependencies. Few methods to identify outliers are:

*1). Sorting the data:*

Our main concern is the admission fee, graduation rate, application received and accepted so let's sort these columns to see if there are any outliers in the data.

*Graduation rate:*

```
data.sort_values(by=['Graduation rate'], inplace=True, ascending=True)
data['Graduation rate']
```

```
1177      8.0
1158     10.0
670      15.0
0        15.0
15       15.0
        ...
451     100.0
77      100.0
824     100.0
979     100.0
771     118.0
Name: Graduation rate, Length: 1302, dtype: float64
```

*Application accepted:*

```
data.sort_values(by=["# appl. accepted"], inplace=True, ascending=True)
data["# appl. accepted"]
```

```
726       35.0
114       36.0
360       44.0
889       55.0
756       61.0
         ...
340    13243.0
104    14141.0
542    15096.0
352    18744.0
750    26330.0
Name: # appl. accepted, Length: 1302, dtype: float64
```

*Application received:*

```
data.sort_values(by=["# appli. rec'd"], inplace=True, ascending=True)
data["# appli. rec'd"]
```

```
726       35.0
114       52.0
889       57.0
360       69.0
622       75.0
         ...
98     19873.0
440    20192.0
352    21804.0
101    22165.0
750    48094.0
Name: # appli. rec'd, Length: 1302, dtype: float64
```

*Instate tuition fee:*

```
data.sort_values(by=["in-state tuition"], inplace=True, ascending=True)
data["in-state tuition"]
```

```
240      480.0
656      556.0
646      608.0
673      628.0
1161     647.0
         ...
484    20655.0
1225   21700.0
975    24940.0
511    25180.0
1233   25750.0
Name: in-state tuition, Length: 1302, dtype: float64
```

No outliers are found in graduation rate and instate tuition fee, but outliers are found in application accepted, application received.

Rutgers at New Brunswick accepted 26330 applications whereas rest accepted around 10000-17000, it has received 48094 applications where rest all have received 15000-20000.

So, we can't completely consider it as an outlier. To consider whether it is an outlier. we have conducted acceptance percentage and found a completely different result.

***The Missouri Baptist college has accepted 1870 applications when it has received only 308 applications. Which is completely absurd so this is an outlier that should be treated.***

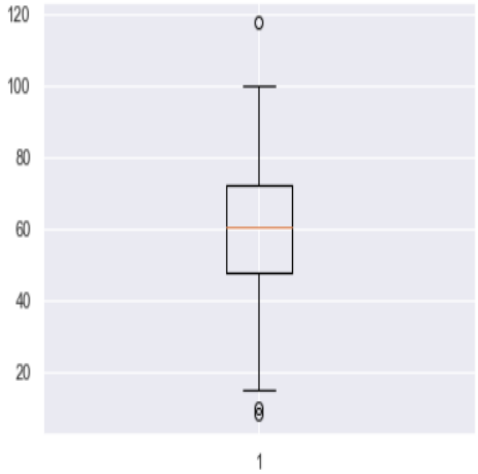
*Box plots:*

In the above we have noticed that the acceptance rate is way different from the data. But if we observe closely the number of students enrolled is

Is again 110 so we can't completely consider it as outlier. so, to better understand the outliers. It is better to box plot the values.

*Graduation Rate:*

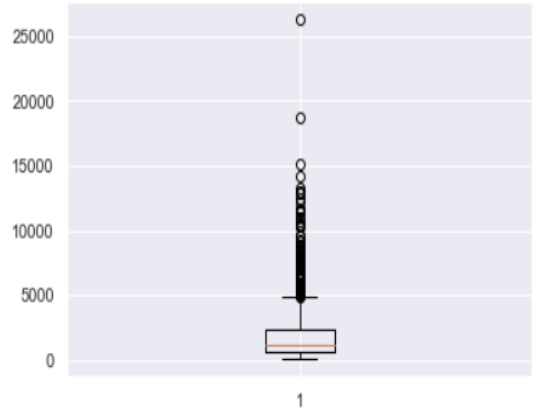
```
plt.boxplot(data["Graduation rate"])
plt.show()
```




Canzovea College of New York is the outlier with value 118.

*Application Accepted:*

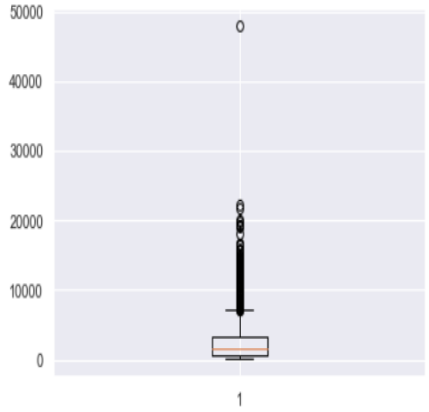
```
plt.boxplot(data["# appl. accepted"])
plt.show()
```




Rutgers at New Brunswick accepted 26330 applications we have identified this before through sorting and it is showing in the box plot as well clearly. So, we can consider it as an outlier.

*Application received:*

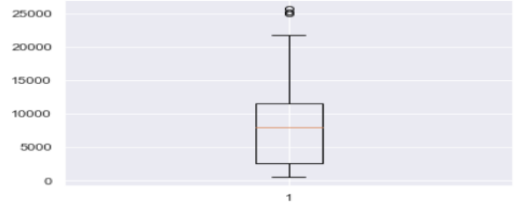
```
plt.boxplot(data["# appli. rec'd"])
plt.show()
```




Just like we have found through sorting Rutgers at New Brunswick accepted 48094 applications is acting like an outlier.

*In State tuition fee:*

```
plt.boxplot(data["in-state tuition"])
plt.show()
```

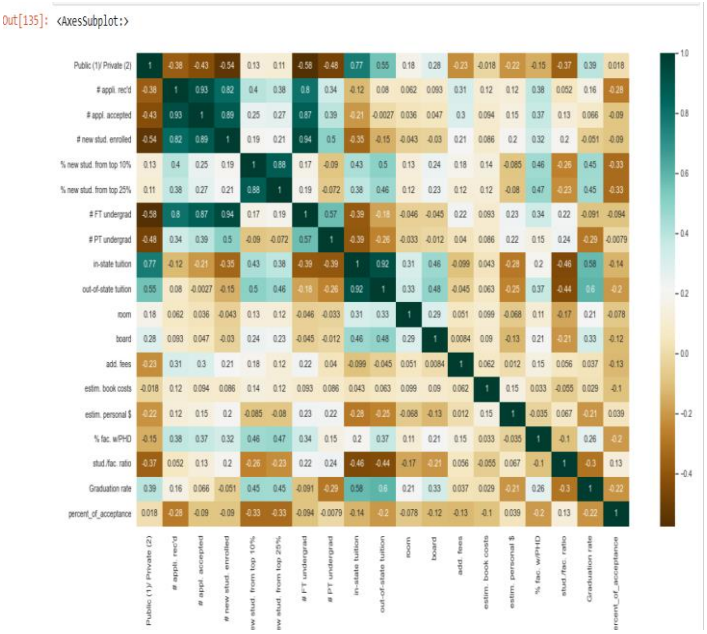



Every university has decent in-state tuition, so we have not come across any outlier in this case.

*Data Cleaning and Transformation*

Now, that we have identified the outliers, it is time to take care of those outliers by deleting them. But, in general 0.3% of the total data for outliers is acceptable.so, we have 1302 records we can have 4 outliers. According to our analysis we have only one. So, we can ignore it. Once the data is cleaned it's time to find any dependencies between the data. One of the methods to do that is heat map.

*Heat Map:*




Since the main goal of the project is conclude whether the new student enrollment depends on the tuition fee. We are particularly paying attention to those dependencies.

The Dependencies:

***New students enrolled:***

a). The number of acceptance of applications.

b). The number of fulltime undergraduate students enrolled

***In-state tuition:***

a). Whether the university is public or private

b). Out -of -state- tuition.

c). Graduation rate

*STATISTICAL ANALYSIS*

*Research Questions and Hypothesis*

From the data analysis we have found some insights to get the questions we have to solve.

1. In the data given the details of universities and expenditures for education to students. The student's weather they are getting the admissions in top colleges/universities.
2. The newly applying students were joining in the public or private Universities.
3. Weather the public or private which has most admissions.
4. Weather the students were joining the inner state or outer of the state.
5. If the students joining inner state means what is the tuition fees they were paying when compared to outer states, weather it is higher from inner state or same.
6. How many graduates we are getting from all the universities?
7. Which university public or private has the most full-time undergraduate students.
8. Which university public or private has the most part-time undergraduate students.
9. Does the university fee depend on whether the university is public or private?
10. Which college is the hardest to get the admission and which university is the easiest to get admit?

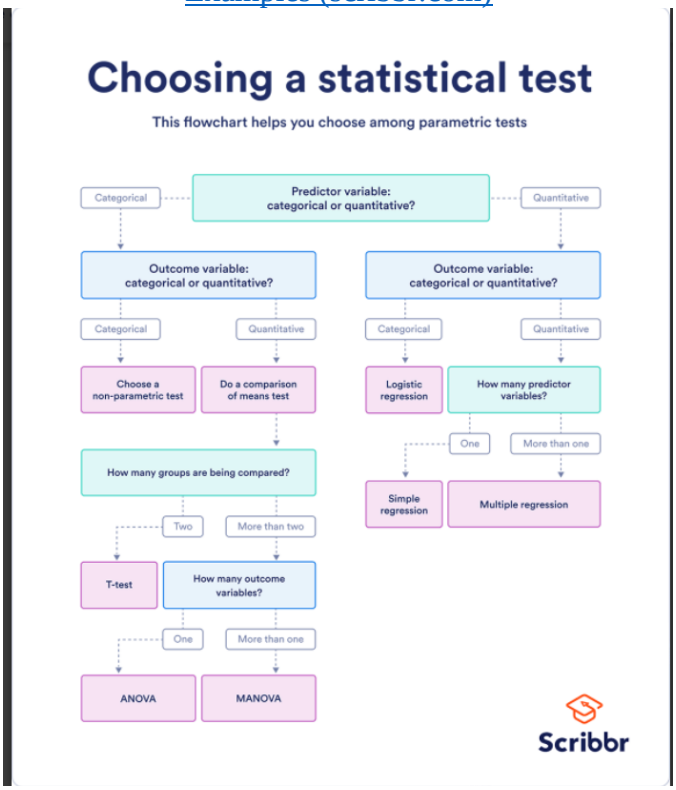
Primarily before performing the test. We need to categorize the hypothesis to map them to corresponding categories:

i. Regression tests
ii. Comparison tests
iii. Correlation tests
iv. Non-parametric tests

And perform necessary transformations of data.

*MAPPING*

[Choosing the Right Statistical Test | Types and Examples (scribbr.com)]




*Insight:*

i. Linear regression.
ii. Linear Regression
iii. Chi- square test
iv. multi-linear regression
v. T-test
vi. Numerical Test
vii. linear regression
viii. linear regression
ix. correlation
x. Numerical Test

Since we are model hypothesis towards logistic regression. We are stating the steps followed for one model and follow the same procedure for rest.

*Linear Regression:*

*Train data:* we have few data like college name, state are string variables which we would be using for later models. So, we transform the data to numeric values.

*HYPOTHESIS*

1). In the data given the details of universities and expenditures for education to students. The student's weather they are getting the admissions in top colleges/universities?

Ans: Yes. most of the students are getting admissions in top universities.
This is multi-linear regression where we consider the universities, in-state-tuition, out-of-state tuition. Since we have already transformed the data. We can go ahead and perform regression.

The value $R^2 = 1$ corresponds to SSR = 0, that is to the **perfect fit** since the values predicted and actual responses fit completely to each other.

We considered multi-linear regression model as we need to compare different variable against one out come.
From the R square value, we can conclude that yes most of the students are getting admits in top universities.

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **R-squared:** | 0.802 |
| **Model:** | OLS | **Adj. R-squared:** | 0.799 |
| **Method:** | Least Squares | **F-statistic:** | 287.8 |
| **Date:** | Wed, 08 Dec 2021 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 05:51:31 | **Log-Likelihood:** | 160.12 |
| **No. Observations:** | 1302 | **AIC:** | -282.2 |
| **Df Residuals:** | 1283 | **BIC:** | -184.0 |
| **Df Model:** | 18 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 1.6695 | 0.054 | 30.647 | 0.000 | 1.563 | 1.776 |
| **x[0]** | 0.0004 | 0.000 | 0.941 | 0.347 | -0.000 | 0.001 |
| **x[1]** | -9.992e-06 | 5.29e-06 | -1.888 | 0.059 | -2.04e-05 | 3.92e-07 |
| **x[2]** | 2.812e-05 | 9.67e-06 | 2.908 | 0.004 | 9.15e-06 | 4.71e-05 |
| **x[3]** | -7.373e-05 | 2.28e-05 | -3.232 | 0.001 | -0.000 | -2.9e-05 |
| **x[4]** | -0.0027 | 0.001 | -3.260 | 0.001 | -0.004 | -0.001 |
| **x[5]** | 0.0010 | 0.001 | 1.465 | 0.143 | -0.000 | 0.002 |
| **x[6]** | 3.062e-06 | 4.49e-06 | 0.682 | 0.495 | -5.75e-06 | 1.19e-05 |
| **x[7]** | -1.383e-05 | 4.84e-06 | -2.857 | 0.004 | -2.33e-05 | -4.33e-06 |
| **x[8]** | 0.0001 | 3.95e-06 | 35.028 | 0.000 | 0.000 | 0.000 |
| **x[9]** | -9.964e-05 | 4.85e-06 | -20.524 | 0.000 | -0.000 | -9.01e-05 |
| **x[10]** | -6.478e-06 | 6.51e-06 | -0.995 | 0.320 | -1.92e-05 | 6.29e-06 |
| **x[11]** | 8.642e-06 | 1.36e-05 | 0.636 | 0.525 | -1.8e-05 | 3.53e-05 |
| **x[12]** | -9.785e-05 | 1.57e-05 | -6.234 | 0.000 | -0.000 | -6.71e-05 |
| **x[13]** | -3.055e-05 | 3.77e-05 | -0.810 | 0.418 | -0.000 | 4.34e-05 |
| **x[14]** | 2.132e-05 | 9.75e-06 | 2.187 | 0.029 | 2.19e-06 | 4.04e-05 |
| **x[15]** | -0.0024 | 0.000 | -5.386 | 0.000 | -0.003 | -0.002 |
| **x[16]** | -0.0034 | 0.001 | -2.570 | 0.010 | -0.006 | -0.001 |
| **x[17]** | 0.0013 | 0.000 | 3.019 | 0.003 | 0.000 | 0.002 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 108.322 | **Durbin-Watson:** | 1.770 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 428.145 |
| **Skew:** | -0.305 | **Prob(JB):** | 1.07e-93 |
| **Kurtosis:** | 5.742 | **Cond. No.** | 1.36e+05 |

2). The newly applying students were joining in the public or private Universities.

Ans: we will be using simple linear regression in such cases since we are comparing one variable against other for one outcome.

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Y | **R-squared:** | 1.000 |
| **Model:** | OLS | **Adj. R-squared:** | 1.000 |
| **Method:** | Least Squares | **F-statistic:** | 1.468e+24 |
| **Date:** | Wed, 27 Apr 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 19:10:09 | **Log-Likelihood:** | 32625. |
| **No. Observations:** | 1302 | **AIC:** | -6.521e+04 |
| **Df Residuals:** | 1281 | **BIC:** | -6.510e+04 |
| **Df Model:** | 20 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -3.862e-12 | 1.16e-12 | -3.316 | 0.001 | -6.15e-12 | -1.58e-12 |
| **X[0]** | 7.315e-16 | 2.49e-16 | 2.940 | 0.003 | 2.43e-16 | 1.22e-15 |
| **X[1]** | -3.572e-15 | 6.61e-15 | -0.540 | 0.589 | -1.65e-14 | 9.4e-15 |
| **X[2]** | 1.0000 | 4.13e-13 | 2.42e+12 | 0.000 | 1.000 | 1.000 |
| **X[3]** | 6.687e-17 | 8.57e-17 | 0.781 | 0.435 | -1.01e-16 | 2.35e-16 |
| **X[4]** | 1.563e-17 | 1.54e-16 | 0.101 | 0.919 | -2.87e-16 | 3.18e-16 |
| **X[5]** | 8.391e-16 | 3.39e-16 | 2.476 | 0.013 | 1.74e-16 | 1.5e-15 |
| **X[6]** | 1.938e-15 | 6.31e-15 | 0.307 | 0.759 | -1.04e-14 | 1.43e-14 |
| **X[7]** | 3.377e-16 | 6.67e-17 | 5.065 | 0.000 | 2.07e-16 | 4.68e-16 |
| **X[8]** | 2.679e-16 | 7.21e-17 | 3.718 | 0.000 | 1.27e-16 | 4.09e-16 |
| **X[9]** | -3.184e-16 | 8.13e-17 | -3.914 | 0.000 | -4.78e-16 | -1.59e-16 |
| **X[10]** | 2.539e-16 | 8.32e-17 | 3.052 | 0.002 | 9.07e-17 | 4.17e-16 |

Based on R-value =1. we can most admissions are done to public university. We have also found this

in our line graph. where public universities have more admit than private.
3). Weather the public or private which has most admissions?

Ans: We already have concluded public universities have more admission than private through line graph, linear regression but better conclude it we are using chi-square testing were
**H0:** public universities have more admissions
**HA:** private universities have more admissions.

[Python - Pearson's Chi-Square Test - GeeksforGeeks]

```
from scipy.stats import chi2_contingency
data = data["Public (1)/ Private (2)"]
stat, p, dof, expected = chi2_contingency(data)

# interpret p-value
alpha = 0.05
print("p value is " + str(p))
if p <= alpha:
    print('Dependent (reject H0)')
else:
    print('Independent (H0 holds true)')
```

```
p value is 1.0
Independent (H0 holds true)
```

Since p value is 1, we consider null hypothesis.

4). Weather the students were joining the inner state or outer of the state.
Ans: We are conducting a multi linear regression consider the hypothesis as the students are enrolling in state. After the test result based on the r-square value, which is 1 we can conclude, yes, the student is joining in the in-state.

```
x=data.drop(['out-of-state tuition','in-state tuition'],axis=1)
y=data['# new stud. enrolled']
model=smf.ols("y~x",data=data).fit()
model.summary()
```

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **R-squared:** | 1.000 |
| **Model:** | OLS | **Adj. R-squared:** | 1.000 |
| **Method:** | Least Squares | **F-statistic:** | 4.594e+30 |
| **Date:** | Wed, 27 Apr 2022 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 20:14:46 | **Log-Likelihood:** | 12064. |
| **No. Observations:** | 474 | **AIC:** | -2.409e+04 |
| **Df Residuals:** | 455 | **BIC:** | -2.401e+04 |
| **Df Model:** | 18 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -2.757e-12 | 1.29e-12 | -2.141 | 0.033 | -5.29e-12 | -2.27e-13 |
| **x[0]** | 5.759e-16 | 2.91e-16 | 1.978 | 0.048 | 3.83e-18 | 1.15e-15 |
| **x[1]** | 1.599e-14 | 8.07e-15 | 1.982 | 0.048 | 1.36e-16 | 3.18e-14 |
| **x[2]** | 9.521e-13 | 3.71e-13 | 2.567 | 0.011 | 2.23e-13 | 1.68e-12 |
| **x[3]** | 1.323e-16 | 8.75e-17 | 1.511 | 0.131 | -3.97e-17 | 3.04e-16 |
| **x[4]** | 3.192e-16 | 1.78e-16 | 1.791 | 0.074 | -3.09e-17 | 6.69e-16 |
| **x[5]** | 1.0000 | 5.38e-16 | 1.86e+15 | 0.000 | 1.000 | 1.000 |
| **x[6]** | -1.177e-14 | 1.55e-14 | -0.761 | 0.447 | -4.21e-14 | 1.86e-14 |
| **x[7]** | -4.274e-15 | 1.26e-14 | -0.339 | 0.735 | -2.91e-14 | 2.05e-14 |

5). If the students joining inner state means what is the tuition fees they were paying when compared to outer states, weather it is higher from inner state or same.
Ans: We are performing two tailed T-test for this hypothesis.
H0: the inner-state tuition is more than the outer-state- tuition.
HA: The outer state tuition is more than the inner state tuition.

```
In [319]: import pingouin as pg

res = pg.ttest(data['out-of-state tuition'],data['in-state tuition'], correction=False)
display(res)
```

| | T | dof | alternative | p-val | CI95% | cohen-d | BF10 | power |
|---|---|---|---|---|---|---|---|---|
| **T-test** | 7.272643 | 2552 | two-sided | 4.666808e-13 | [1007.65, 1751.62] | 0.287816 | 9.017e+09 | 1.0 |

Since the p> is greater than 0.5 the test has failed, and we consider the alternative hypothesis as true.
The outer state fee is more than the inner state tuition.

6). How many graduates we are getting from all the universities?

Ans: we have not found appropriate test for this hypothesis, so we solved it through general pandas library python coding.

```
total_public=data.loc[data['Public (1)/ Private (2)'] == 1,"# new stud. enrolled"].sum()
total_public
```

```
666077.5219737857
```

```
total_enrolled = data["Graduation rate"].sum()
total_enrolled
```

```
72728.0
```

Out of 666077 students enrolled 72728 students have graduated.

7). Which university public or private has the most full-time undergraduate students.

Ans: We are conducting linear regression test let's assume the hypothesis as the public has more full-time undergraduate students than private.

```
x=data["Public (1)/ Private (2)"]
y=data['# FT undergrad']
model=smf.ols("y~x",data=data).fit()
model.summary()
```

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **R-squared:** | 0.335 |
| **Model:** | OLS | **Adj. R-squared:** | 0.335 |
| **Method:** | Least Squares | **F-statistic:** | 654.1 |
| **Date:** | Wed, 27 Apr 2022 | **Prob (F-statistic):** | 3.80e-117 |
| **Time:** | 20:50:54 | **Log-Likelihood:** | -12517. |
| **No. Observations:** | 1299 | **AIC:** | 2.504e+04 |
| **Df Residuals:** | 1297 | **BIC:** | 2.505e+04 |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 1.269e+04 | 366.357 | 34.626 | 0.000 | 1.2e+04 | 1.34e+04 |
| **x** | -5481.7556 | 214.340 | -25.575 | 0.000 | -5902.247 | -5061.264 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 701.029 | **Durbin-Watson:** | 1.882 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 6412.449 |
| **Skew:** | 2.348 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 12.820 | **Cond. No.** | 8.05 |

Consider the r value is less than 1 we can conclude that our hypothesis is wrong and private university has more full-time graduates than public.

8). Which university public or private has the most part-time undergraduate students.

Ans: We are conducting linear regression test let's assume the hypothesis as the public has more part-time undergraduate students than private.

```
x=data["Public (1)/ Private (2)"]
y=data['# PT undergrad']
model=smf.ols("y~x",data=data).fit()
model.summary()
```

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **R-squared:** | 0.238 |
| **Model:** | OLS | **Adj. R-squared:** | 0.238 |
| **Method:** | Least Squares | **F-statistic:** | 396.7 |
| **Date:** | Wed, 27 Apr 2022 | **Prob (F-statistic):** | 5.19e-77 |
| **Time:** | 20:55:07 | **Log-Likelihood:** | -11055. |
| **No. Observations:** | 1270 | **AIC:** | 2.211e+04 |
| **Df Residuals:** | 1268 | **BIC:** | 2.212e+04 |
| **Df Model:** | 1 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 3859.3288 | 145.363 | 26.550 | 0.000 | 3574.150 | 4144.508 |
| **x** | -1696.8776 | 85.198 | -19.917 | 0.000 | -1864.023 | -1529.732 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 1156.162 | **Durbin-Watson:** | 1.751 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 63354.389 |
| **Skew:** | 4.020 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 36.654 | **Cond. No.** | 8.01 |

Notes:

Consider the r value is less than 1 we can conclude that our hypothesis is wrong and private university has more part -time graduates than public.

9). Does the university fee depend on whether the university is public or private?

Ans: We are doing correlation testing to find out the dependency.

[How to Calculate Correlation Between Variables in Python (machinelearningmastery.com)]

```
# calculate the Pearson's correlation between two variables
from numpy.random import randn
from numpy.random import seed
from scipy.stats import pearsonr

# calculate Pearson's correlation
corr, _ = pearsonr(data['in-state tuition'], data['Public (1)/ Private (2)'])
print('Pearsons correlation: %.3f' % corr)
```

```
Pearsons correlation: 0.770
```

```
# calculate the Pearson's correlation between two variables
from numpy.random import randn
from numpy.random import seed
from scipy.stats import pearsonr

# calculate Pearson's correlation
corr, _ = pearsonr(data['out-of-state tuition'], data['Public (1)/ Private (2)'])
print('Pearsons correlation: %.3f' % corr)
```

```
Pearsons correlation: 0.552
```

As we can both the co-relation values are greater than 0.5, we can say there is strong correlation between university fee and whether the university is public or private.

10). Which college is the hardest to get the admission and which university is the easiest to get admit?

Ans: The college with lowest acceptance rate is the hardest admit achieving.

The college with highest acceptance rate is the easiest admit obtaining.

We are general pandas' library to solve this hypothesis.

```
uni_hard=data.loc[data['percent_of_acceptance'] <7]
uni_hard
```

| | College Name | State | Public (1)/ Private (2) | # appli. rec'd | # appl. accepted | # new stud. enrolled | % new stud. from top 10% | % new stud. from top 25% | # FT undergrad | # PT undergrad | ... | out-of-state tuition | room | board | add. fees | estim. book costs | estim. personal $ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 641 | Marygrove College | MI | 2 | 2752.097523 | 177.0 | 125.0 | 10.0 | 75.0 | 602.0 | 488.0 | ... | 8094.0 | 1100.0 | 880.0 | 48.0 | 566.0 | 1389.291704 |

1 rows × 21 columns

```
rslt_df = data.loc[data['percent_of_acceptance'] > 500]
rslt_df
```

| | College Name | State | Public (1)/ Private (2) | # appli. rec'd | # appl. accepted | # new stud. enrolled | % new stud. from top 10% | % new stud. from top 25% | # FT undergrad | # PT undergrad | ... | out-of-state tuition | room | board | add. fees | estim. book costs | es... person |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 618 | Missouri Baptist College | MO | 2 | 308.0 | 1870.683191 | 110.0 | 5.0 | 23.0 | 606.0 | 1142.0 | ... | 5780.0 | 2650.0 | 2060.983831 | 180.0 | 549.972887 | 1389.29 |

1 rows × 21 columns

Therefore, the hardest college to achieve the admit is Marygrove College and the easiest is Missouri Baptist College.

## *CONCLUSION*

Now, we have better understood our data through analysis and statistical test so we can conclude that the tuition fee is not key component considered for the student acceptance towards a university. It is one of the key components along other component whether the university is public or private. Through our analysis we have found that the variable of whether the university is public or private played a dominating over the tuition fee.

## *BARRIERS*

Lack of python knowledge in team members.

Understanding the algorithms

Making sure the dependencies does not affect the test results

Making sure the outliers do not affect the test results.

Time crunch of the project deadline.

## *REFERENCES*


1. Exploratory data analysis in Python. | by Tanu N Prabhu | Towards Data Science
2. Exploratory Data Analysis(EDA) from Scratch | With Pythin Implementation (analyticsvidhya.com)
3. Linear Regression in Python – Real Python
4. Choosing the Right Statistical Test | Types and Examples (scribbr.com)
5. Building A Logistic Regression in Python, Step by Step | by Susan Li | Towards Data Science
6. How to Calculate Correlation Between Variables in Python (machinelearningmastery.com)
7. DSS - Interpreting Regression Output. (n.d.). Retrieved from Princeton University: https://dss.princeton.edu/online_help/analysis/interpreting_regression.htm
8. Hagerman, I. (2017, December 19). Residual Plots Part 1 - Residuals vs. Fitted Plot. Retrieved from Medium: https://medium.com/data-distilled/residual-plots-part-1-residuals-vs-fitted-plot-f069849616b1.
9. Interpreting the Intercept in a Regression Model. (2020, January 16). Retrieved from The Analysis Factor: https://www.theanalysisfactor.com/interpreting-the-intercept-in-a-regression-model/
10. McCall, B. P. (1995). The Impact of Unemployment Insurance Benefit Levels on Recipiency. Journal of Business & Economic Statistics, Vol. 13, No. 2, JBES Symposium onProgram and Policy Evaluation, 189-198.