

# ANALYSIS OF AMERICAN UNIVERSITIES

Major Project submitted in partial fulfilment of the requirements for the degree of  
Bachelor of Technology

in

**Computer Science and Engineering**

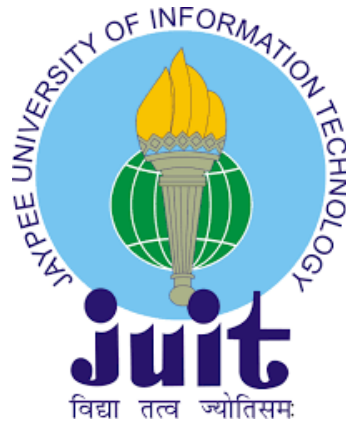
by

Anant Mishra (181367)

Utkarsh Pandit (181432)

**UNDER THE SUPERVISION OF**

Prof. Prateek Thakral



**Department of Computer Science & Engineering and Information  
Technology**

**Jaypee University of Information Technology, Wagnaghat, 173234,  
Himachal Pradesh, INDIA**

# DECLARATION

We hereby declare that this project has been done by us under the supervision of Prof. Prateek Thakral, Assistant Prof, Department of CSE Jaypee University Of Information Technology. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

**Supervised by:**

**Prof**

**Prateek Thakral**

Department of Computer Science & Engineering And Information Technology  
Jaypee University Of Information Technology

**Submitted by:**

**Utkarsh Pandit**

(181432)

**Anant Mishra**

(181367)

Department of CSE  
Jaypee University of Information Technology

# CERTIFICATE

This is to certify that the work which is being presented in the project report titled "ANALYSIS OF AMERICAN UNIVERSITIES" in partial fulfilment of the requirements for the award of the degree of B.Tech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by "Anant Mishra (181367)" And "Utkarsh Pandit (181352)" during the period from August 2021 to December 2021 under the supervision of Prof. Prateek Thakral, Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat.

Utkarsh Pandit  
(181432)

Anant Mishra  
(181367)

The above statement made is correct to the best of my knowledge.

**Professor Prateek Thakral**

Computer Science & Engineering And Information Technology  
Jaypee University of Information Technology, Waknaghat, India

# ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to Almighty God for His divine blessing makes us possible to complete the project successfully.

We are grateful and wish my profound indebtedness to Supervisor Prof. Prateek Thakral, Assistant Prof, Department of CSE Jaypee University Of Information Technology, Wakhnaghat. Deep Knowledge & keen interest of my supervisor in the field of "Exploratory Data Analysis" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express my heartiest gratitude to Prof. Prateek Thakral, Department of CSE, for his kind help to finish my project.

We would also generously welcome each one of those individuals who have helped us straightforwardly or in a roundabout way in making this project a win. In this unique situation, we might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, we must acknowledge with due respect the constant support and patients of my parents.

Anant Mishra  
(181367)

Utkarsh Pandit  
(181432)

# TABLES OF CONTENT

- Declaration
- Certificate
- Acknowledgement
- Abstract
- Introduction
- Literature review
- System Development
- Proposed Methodology
- Performance Analysis
- Result
- Conclusion
- Future Work
- References

# ABSTRACT

Information should be investigated in order to create great outcomes. Utilizing the outcome choice can be taken. For instance suggestion framework, positioning of the page, request guaging, forecast of acquisition of the item. There are a few driving organizations where the survey of the client assumes an extraordinary part in investigating the component which impacts the audit rating. We have utilized exploratory information examination (EDA) where information translations should be possible in line and section design. We have involved python for information examination. It is object situated ,deciphered and intelligent programming language. It is open source with rich arrangements of libraries like pandas, MATplotlib, seaborn and so on. We have utilized various sorts of outlines and different kinds of boundary to dissect Amazon survey informational collections which contains the audits of electronic information things. We have utilized python programming for the information examination

# INTRODUCTION

Information is becoming extremely quicker in this day and age. It isn't the case simple to physically deal with the information. Information investigation and perception programs take into consideration arriving at significantly more profound comprehension. The programming language Python, with its English orders and simple to-follow punctuation, offers an incredibly strong (and free!) open-source option in contrast to customary methods and applications. Data analysis allows organizations to figure out their effectiveness and execution, and eventually assists the business with pursuing more educated choices. For instance, an internet business organization may be keen on dissecting client credits to show designated advertisements for further developing deals. Information investigation can be applied to practically any part of a business assuming one comprehends the devices accessible to handle data. The online business organizations are investigating the surveys of client by utilizing legitimate perception technique. Exploratory Data Analysis (EDA) is a way to deal with sum up the information by taking their primary qualities and picture it with legitimate portrayals. EDA centers all the more barely around checking suppositions expected for model fitting and speculation testing, and dealing with missing qualities and making changes of factors depending on the situation. EDA includes IDA. EDA rapidly portrays the informational collections number of lines/sections, missing information, information types and review. Clean defiled information; handle missing information, invalid information types and inaccurate qualities. EDA imagine information conveyances; bar graphs, histograms, box plots. Compute and imagine connections (connections) between factors; heat map.

# LITERATURE REVIEW

## **#1 - A comprehensive review of tools for exploratory analysis of tabular industrial datasets**

**Aindrila Ghosh a,\* , Mona Nashaata, James Miller a, Shaikh Quader b, Chad Marstonc**

In this paper, there was an in-depth analysis of an industrial tabular dataset, they identified a set of additional exploratory requirements for large datasets, and also presented a comprehensive survey of the recent advancements in the emerging field of exploratory data analysis. They investigated 50 academic and non-academic visual data exploration tools with respect to their utility in the six fundamental steps of the exploratory data analysis process. They also examine the extent to which these modern data exploration tools fulfill the additional requirements for analyzing large datasets. At last they presented a set of research opportunities in the field of visual exploratory data analysis.

## **#2 John T. Behrens, “Principles and Procedures of Exploratory Data Analysis,” Psychological Methods, 1997, Vol. 2, No. 2, pp.131-160.**

With the help of above paper Author John T. Behrens has described about the difference between classical data analysis and exploratory data analysis using different visualisation methods. He has explained the need of EDA and how it provides conceptual and computational tools to foster hypothesis development and refinement. He has explained the role of graphical representation while comparing it with algebraic summaries. He has summarized the differences between all the graphs and how subtle information each provides.

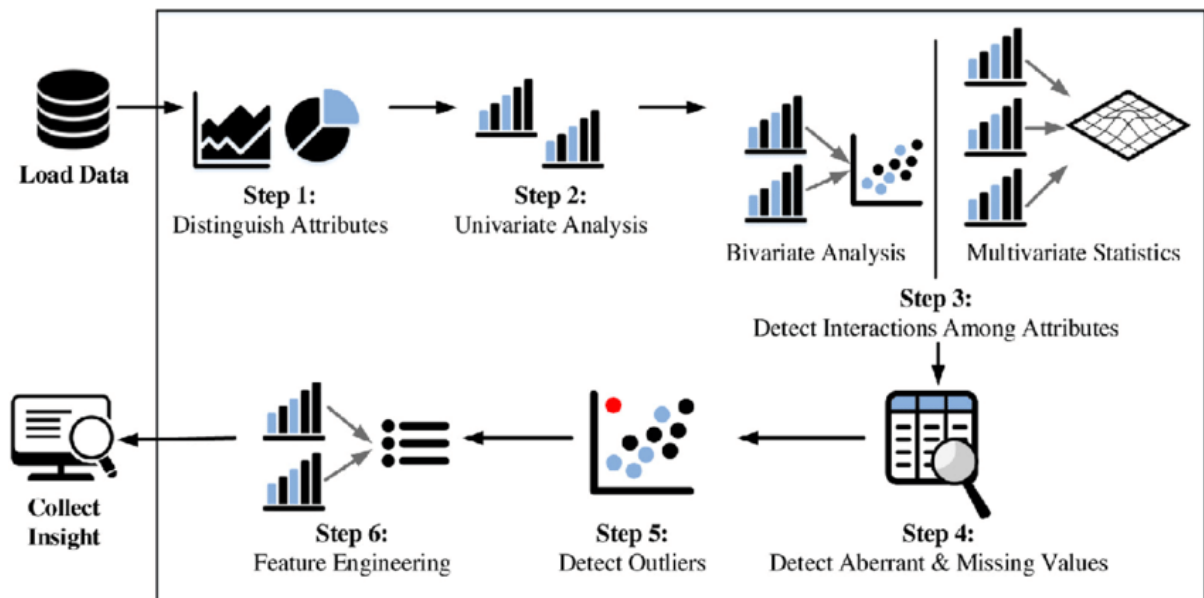


### **#3 Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities - Zeshan Fayyaz, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim and Rasha Kashef**


This paper provides the current landscape of recommender systems research and identifies directions in the field in various applications. This article provides an overview of the current state of the art in recommendation systems, their types, challenges, limitations, and business adoptions. To assess the quality of a recommendation system, qualitative evaluation metrics are discussed in the paper. In this paper, they have presented a detailed survey of RS that introduces different types of RSs as collaborative filtering, content-based, demographic-based, utility-based, knowledge-based, and hybrid-based. Different combination strategies of hybrid-based systems are also presented and categorized into weighted, mixed, switching, feature combination, feature augmentation, cascade, and meta-level.

---

# SYSTEM DEVELOPMENT



## # Tools and Technologies used

1. Python : 

It is a programming language with a wide range of capabilities. Its broad features make working with targeted programs (including meta-programming and meta-objects) simple, and it fully supports object-oriented programming. Python takes advantage of power typing as well as the integration of reference computation and waste management waste collecting.

2. Development Tool - Jupyter Notebook
3. Libraries - Pandas, Numpy, Matplotlib, Seaborn
4. Hardware/ Software Requirements :

Intel Core i3, 5, i7. and 2 GHz processor RAM must be at 512MB.  
Hard disc with a capacity of at least 10 GB  
Input Keyboard and Mouse are the devices that are used.

# TECHNIQUES FOR EXPLORATORY DATA ANALYSIS

Fundamentally, exploratory data analysis is a way to see what the information can convey us from the conventional displaying or theory testing task. EDA assists with investigating the informational collections to sum up their factual qualities zeroing in on four key angles, similar to, proportions of focal inclination (containing the mean, the mode and the middle), proportions of spread (including standard deviation and difference), the state of the dispersion and the presence of exceptions. In the accompanying passages, we have introduced a portrayal of these critical parts of EDA. At each progression of AI process, information examination and representation strategies are broadly being utilized. These methods are examined in as underneath

## #1 Data Exploration

It is the initial phase of data analysis. It is used to visualize data to uncover insights from the beginning and to identify patterns to look into more. It uses point and click data exploration and interactive dashboards so that users can better understand it and get better insights quicker.

### At Every Step of the Machine Learning Process, Data Analysis and Visualization Techniques are being Used

Data Exploration	Data Cleaning	Model Building	Present Results
<ul style="list-style-type: none"><li>• Visualization</li><li>• Find Missing</li><li>• Look for Correlations</li></ul>	<ul style="list-style-type: none"><li>• Check: did I fix the potential issues?</li></ul>	<ul style="list-style-type: none"><li>• Visualize Results</li><li>• Model Diagnostics</li><li>• Residual Diagnostics</li><li>• ROC Curves</li><li>• etc.</li></ul>	<ul style="list-style-type: none"><li>• Charts</li><li>• Graphs</li><li>• Tables</li><li>• Visualize to explain mode, explain results</li></ul>

## **#2 Data Cleaning**

It is process of removing or deleting corrupted data that might be incorrect, duplicated, wrongly formatted or incomplete within a dataset. Data often gets corrupted when combining different data sources so it is necessary to perform data cleaning on it to remove duplication and mislabelling.

## **#3 Model Building**

In this stage we develop statistical model or machine learning model for training, testing and manufacturing purposes. These models allow users to create analytical methods and train them, while also keeping aside some information for testing it. These can be supervised or unsupervised.

## **#4 Present Results**

Data can be graphically visualised in the form of graph, tables and charts. It is way easier for humans to understand that way in the form of graphs and charts. It is also an easy way to display intent. With it we can identify areas where improvement is needed very easily.

# Graphical EDA

In a general sense, graphical exploratory information investigation is only the graphical counterpart of the customary non-graphical EDA that examines the informational indexes to assist with summing up their factual attributes zeroing in on similar four critical angles, identical to, proportions of focal inclination, proportions of spread, the state of the appropriation and the presence of anomalies. Further, we have arranged GEDA into: Univariate GEDA, Bivariate GEDA, and Multivariate GEDA. In the accompanying passages, we have talked about these vital assortments and parts of GEDA

## Univariate Graphical EDA

Univariate GEDA gives factual synopsis to each field in the crude informational collection or the rundown just on one variable. Illustration of these kinds of GEDA incorporates total appropriation work (CDF), likelihood thickness work (PDF), Box plot and Violin plot. Not many of them are examined underneath:

I. Histograms: We can address the dissemination of mathematical information by the utilization of histogram. Histogram can connect with one variable instead of two factors. Here the whole scope of significant worth can be isolated into a series of span. Histograms are primarily utilized for consistent information. Histogram can be addressed as recurrence conveyance through square shape where a width addresses the class span and region relative to comparing frequencies. Level addresses the typical recurrence thickness. Apparent dissemination of computerized picture is a graphical portrayal which is called as picture histogram.

II. Stem Plots: It is generally called as leaf plot. Here the information is spitted in to two sections. The biggest digit addresses the stems and the littlest digit addresses the leaves. Somewhat more data is addressed by stem plot over histogram. It is likewise utilized for representation reason. It is a lot simpler here to Compare the information. The numbers are organized by place esteem. They are fundamentally utilized for featuring the mode .they are utilized for little informational indexes

III. Box plots: A decent graphical picture of the grouping of information can be addressed by the utilization of box plot. It shows the focal propensity, evenness, slant and anomaly. It tends to be developed from five qualities: the base, the main quartile, the middle, the third quartile and the greatest worth. These qualities are contrasted with show how close different information values are to them.

### **Bivariate Graphical EDA**

Bivariate GEDA is achieved to comprehend the associations between every variable in the dataset and the objective variable of interest or utilizing two factors and tracking down association among them. Instances of these sorts of GEDA incorporate Box plot and Violin plot.

### **Multivariate Graphical EDA**

Multivariate GEDA is achieved to comprehend the associations between various fields in the dataset or tracking down the associations between multiple factors. Illustration of these sorts of GEDA incorporates Pair plot and 3D Scatter plot. BARGRAPH plot is the most regularly utilized graphical strategy. These days Box plot is utilized to show the connection between two qualities. Sometimes Pair plot is utilized to show the perspective on all factor and their relationship.

I. Side by Side Box plots : For contrasting the levels of all potential qualities we utilize one next to the other box plot.it is utilized to think about two informational collections. it fundamentally sum up the information for every moment of clear cut variable

II. Dissipate plots : It is a kind of plot where Cartesian direction is utilized to show the qualities between two factors for a bunch of information. We can attract it by taking the variable worth X pivot and Y hub. The information are shown as an assortment of focuses. The worth of X hub and y hub gives the worth of the variable

III. Heat Maps and 3D Surface Plots : We can produce heat map taking the whole component variable. Feature factors are taken as line and section header and the variable versus itself on the inclining. Picturing the connection between factors in high layered space is extremely valuable.

# EDA IN PYTHON

We are using python for exploratory information examination. It is easy to learn. It has rich arrangements of libraries. Information taking care of limit are a lot higher. It is utilized as open source language. It has the ability to with all the outsider language .it can run on any stage. It can move the cycle starting with one stage then onto the next. It is not difficult to peruse. The designer can grasp the code. It offers an assortment of libraries and some of them utilizes incredible representation instrument. Representation interaction can make it more straightforward to make the reasonable report.

## **Pandas**

It is the most remarkable package for data analysis. We can clean, change and break down the information. Information can be put away in CSV design on Computer. Cleaning, picturing and putting away the information should be possible. It is based on the highest point of the NumPy bundle. Plotting capacities from Matplotlib and AI calculation in Scikit-learn.

## **Jupyter Notebook**

It gives ability the code in a specific cell. It gives the control center based approach for processing. It gives electronic application process. It incorporates information and result of the calculation. It gives rich media portrayal of the item

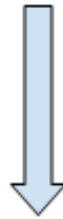
## **Applications of EDA**

1. It detects anomalies and mistakes easily.
2. It gives us insight in various kind of information.
3. Detects outliers in data.
4. EDA test assumptions very accurately.
5. EDA identifies important factors.
6. Explains the relationship between various data.
7. By using visualization it makes the data speak for itself .

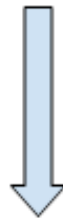


# PROPOSED METHODOLOGY / ALGORITHM

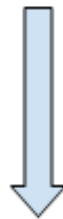
Select a large real world dataset from kaggle



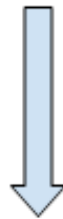
Perform Data preparation and cleaning using numpy and pandas



Perform exploratory analysis and visualization using matplotlib and seaborn



Ask and answer interesting questions about the data in jupyter notebook



Summarize your inferences and write conclusions based on it

# PERFORMANCE ANALYSIS

## WORKING WITH THE DATA SETS

Now is the ideal time to investigate the information and find out about it. The dataset we are utilizing contains plenty of data about American Universities. We will investigate the information with the conceivable arrangement of choices

### # 1. Reading the Dataset.

This portion presents the steps to read the dataset about American Universities.

### # 2. Data Preparation: Cleaning and Formatting.

In all information examination projects, the information planning step isn't simply essential yet additionally imperative to find and deal with highlights that could create a few issues while making the quantitative investigation or that could prompt low proficient coding. As per Alivia Smith[1], this progression for the most part takes up to 80% of the whole season of an informal investigation project.

Hence, missing, invalid, and conflicting qualities have been tended to.

At long last, this progression presents a code for changing the configuration of section names.

### # 3. Exploratory Data Analysis (EDA) and Visualization.

Quantitative and subjective examination (Asking and Answering Questions).

Albeit much of the time, the exploratory information investigation and the quantitative and subjective examination are isolated strides, in this particular venture, they have been joined.

This progression, past posing inquiries that could prompt arriving at the undertaking's point, presents worldwide valuable data about the various segments of the dataset. A few introductory assessments proceeded as an approach to starting tracking down designs, making speculations, and

validating early suspicions. Afterward, more profound examinations were portrayed as a feature of arriving at the task's objective.

Quantitative and qualitative analysis: Asking and Answering Questions.  
The hypotheses and questions generated to develop this projects are:

1. Do universities with a high number of applications are the preferred ones by students?; in other words, could the number of applications tell us that a university is one of the most preferred by students?.
2. Do students prefer universities that have a high rate of admission?, in other words, do students prefer a university where it is easier for them to be admitted?.
3. Do students prefer public or private universities?
4. Do students prefer universities with low tuition and fees?
5. Do students prefer a university for its low on-campus cost of living?
6. Do students prefer universities from highly populated states?
7. Do students prefer a university because it belongs to a state with a high GDP per capita?
8. Do students prefer a university based on the possibility of a higher, additional academic degree in the same university?

```
In [5]: universities_df.head()
```

Out[5]:

	ID number	Name	year	ZIP code	Highest degree offered	County name	Longitude location of institution	Latitude location of institution	Religious affiliation	Offers Less than one year certificate	...	Percent of freshmen receiving federal grant aid	Percent of freshmen receiving Pell grants	Percent of freshmen receiving other federal grant aid	Percent freshmen receive state/loc grant a
0	100654	Alabama A & M University	2013	35762	Doctor's degree - research/scholarship	Madison County	-865,685	3,478,337	Not applicable	Implied no	...	81.0	81.0	7.0	1
1	100663	University of Alabama at Birmingham	2013	35294-0110	Doctor's degree - research/scholarship and pro...	Jefferson County	-868,092	3,350,223	Not applicable	Implied no	...	36.0	36.0	10.0	0
2	100690	Amridge University	2013	36117-3553	Doctor's degree - research/scholarship and pro...	Montgomery County	-86,174	3,236,261	Churches of Christ	Implied no	...	90.0	90.0	0.0	40
3	100706	University of Alabama in Huntsville	2013	35899	Doctor's degree - research/scholarship and pro...	Madison County	-866,384	3,472,282	Not applicable	Yes	...	31.0	31.0	4.0	1
4	100724	Alabama State University	2013	36104-0271	Doctor's degree - research/scholarship and pro...	Montgomery County	-862,957	3,236,432	Not applicable	Implied no	...	76.0	76.0	13.0	11

5 rows x 145 columns

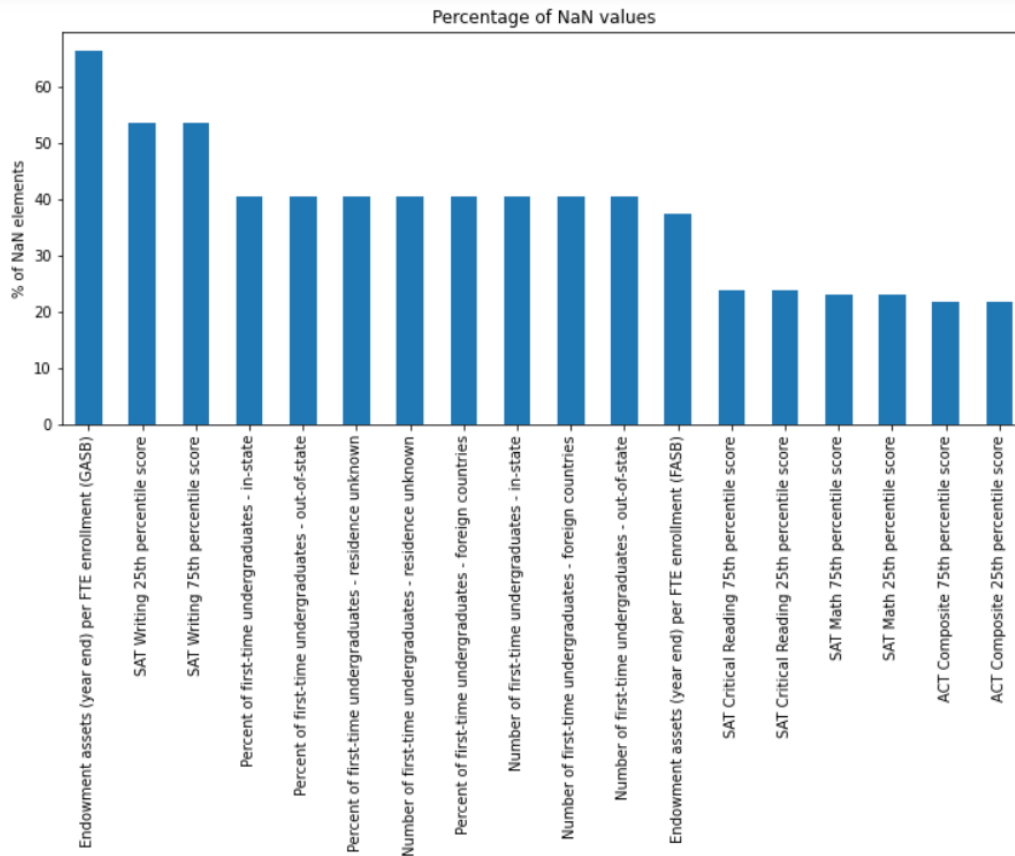
**Fig1 - head function showing top 5 rows of the dataset**

```
universities_df.info(max_cols=len(universities_df))
```

11	Offers Associate's degree	1532	non-null	object
12	Offers Two but less than 4 years certificate	1532	non-null	object
13	Offers Bachelor's degree	1532	non-null	object
14	Offers Postbaccalaureate certificate	1532	non-null	object
15	Offers Master's degree	1532	non-null	object
16	Offers Post-master's certificate	1532	non-null	object
17	Offers Doctor's degree - research/scholarship	1532	non-null	object
18	Offers Doctor's degree - professional practice	1532	non-null	object
19	Offers Doctor's degree - other	1532	non-null	object
20	Offers Other degree	1532	non-null	object
21	Applicants total	1377	non-null	float64
22	Admissions total	1377	non-null	float64
23	Enrolled total	1377	non-null	float64
24	Percent of freshmen submitting SAT scores	1257	non-null	float64
25	Percent of freshmen submitting ACT scores	1259	non-null	float64
26	SAT Critical Reading 25th percentile score	1169	non-null	float64
27	SAT Critical Reading 75th percentile score	1169	non-null	float64
28	SAT Math 25th percentile score	1182	non-null	float64
29	SAT Math 75th percentile score	1182	non-null	float64
30	SAT Writing 25th percentile score	714	non-null	float64

In this case, only three dtypes have been recognized: float64, int64, and object.

**Fig2 - showing data type for each column of data frame**



**Fig3 - Percentage of Null Values in each column**

## CODE:

```
In [26]: def remove_space(list_headers,charact): #charact should be: charact=[' - ',' ']  
new_headers=list()  
for header in list_headers:  
    for char in charact:  
        if char in header:  
            header=header.replace(char,'_')  
            header=header  
        new_headers.append(header)  
return new_headers
```

```
In [27]: def remove_sp_char(headers,chars):  
new_headers=list()  
for header in headers:  
    for char in chars:  
        if char=='-' or char=='/':  
            header=header.replace(char,'_')  
        if char in header:  
            header=header.replace(char,'')  
  
    header=header  
    new_headers.append(header)  
return new_headers
```

Besides removing spaces and replacing special characters, I'll change capitalized letters to avoid a typo of this kind.

```
In [28]: headers=remove_space(universitiesnw_df.columns,[' - ',' '])
```

```
In [29]: headers=remove_sp_char(headers,['"',',',';','-', '/'])
```

```
In [30]: list_new_header=list()  
  
for header in headers:  
    header=header.casefold() # All capitalized letters are changed.  
  
    if "degrese" in header: # One column name has a typo.  
        header=header.replace("degrese",'degrees')  
  
    list_new_header.append(header)
```

The next step is to replace the original column names with the new ones, which have the desired format.

```
In [32]: universitiesnw_df.columns=list_new_header
```

```
In [33]: universitiesnw_df.columns
```

```
Out[33]: Index(['id_number', 'name', 'year', 'zip_code', 'highest_degree_offered',  
              'county_name', 'longitude_location_of_institution',  
              'latitude_location_of_institution', 'religious_affiliation',  
              'offers_less_than_one_year_certificate',  
              ...  
              'percent_of_freshmen_receiving_any_financial_aid',  
              'percent_of_freshmen_receiving_federal_state_local_or_institutional_grant_aid',
```

```
with open('states_gdp.csv') as file:
    gdp_df=pd.read_csv(file)
```

```
gdp_df.head()
```

```
gdp_df.drop(columns=['code'],inplace=True)
```

```
universitiesnw_df=universitiesnw_df.merge(gdp_df,on='state')
```

```
universitiesnw_df[['state','gdp_million','population']].head()
```

As we have the GDP of each state but we want to work with GDP per capita, we need to calculate this value.

```
universitiesnw_df['gdp_capita']=universitiesnw_df.gdp_million/universitiesnw_df.population*1e6
```

Let's see the GDP per capita of each state.

```
gdp_state_df=universitiesnw_df.groupby('state')[['region','gdp_capita']].mean().sort_values('gdp_c
```

```
plt.figure(figsize=(16,7))
sns.scatterplot(x='gdp_capita',y='enrollment_rate',data=universitiesnw_df);
plt.plot([78000,78000], [0, 110], c='magenta',lw=3,marker='*',ls='--')
plt.title('GDP per Capita vs Enrollment Rate')
plt.grid()
plt.xlabel('GDP per Capita $')
plt.ylabel('Enrollment Rate %');
```

GDP per Capita vs Enrollment Rate

```
plt.figure(figsize=(16,7))
sns.scatterplot(x='gdp_capita',y='enrollment_rate',data=universitiesnw_df);
plt.axis([30000,80000,0, 101]);
plt.grid();
plt.title('GDP per Capita vs Enrollment Rate')
plt.xlabel('GDP per Capita $')
plt.ylabel('Enrollment Rate %'); sns.despine();
```

```
plt.figure(figsize=(10,8))
ax=sns.barplot(x=degree,y=degree.index)
ax.set_yticklabels(("Bachelor's Degree","Master's Degree",
                   "Doctor's Degree: Research/Scholarship",
                   "Doctor's Degree: Professional Practice"));
plt.title('Degrees Offered')
plt.xlabel('Universities')
plt.grid(axis='x');
```

```
hg_degree=universitiesnw_df.highest_degree_offered.value_counts()
```

```
plt.figure(figsize=(16,8))
plt.pie(hg_degree, labels=hg_degree.index,
        autopct='%1f%%', startangle=140, colors = ['violet', 'aqua', 'pink', 'lightsalmon', 'moccasin'],
plt.title('Highest Degree Offered');
```

```
plt.figure(figsize=(16,8))
ax=sns.scatterplot(y='highest_degree_offered',x='enrollment_rate',data=universitiesnw_df);
plt.title('Highest Degree Offered vs Enrollment Rate')
plt.ylabel('')
plt.xlabel('Enrollment Rate %')
plt.grid(axis='x')
ax.set_yticklabels(("Doctor's Degree:
Research/Scholarship",
"Doctor's Degree: Research/
Scholarship & Professional
Practice",
"Bachelor's Degree",
"Doctor's Degree:
Professional Practice",
"Master's Degree",
"Doctor's Degree: Other"));
```

```
: import matplotlib.pyplot as plt
import requests
import urllib.request
from IPython.core.debugger import Tracer

url_form = "http://thegradcafe.com/survey/index.php?q=u%2A&t=a&pp=250&o=d&p={0}"
DATA_DIR = './WebScraped_data/html/'

if __name__ == '__main__':
    for i in range(1691, 1948):
        url = url_form.format(i)
        handle = urllib.request.urlopen(url)
        html = handle.read()
        html = html.decode('utf8')
        #r = requests.get(url)
        fname = "{data_dir}/{page}.html".format(data_dir=DATA_DIR,page=str(i))
        with open(fname, 'wb') as f:
            f.write(html.encode('UTF-8'))
        print("getting {0}...".format(i))
```



```

def knn(trainingSet, testInstance, k):
    print(k)
    distances = {}
    sort = {}
    length = testInstance.shape[1]

    for x in range(len(trainingSet)):

        dist = euclideanDistance(testInstance, trainingSet.iloc[x], length)

        distances[x] = dist[0]

    sorted_d = sorted(distances.items(), key=lambda x: x[1])

    neighbors = []

    for x in range(k):
        neighbors.append(sorted_d[x][0])

    classVotes = {}

    for x in range(len(neighbors)):
        response = trainingSet.iloc[neighbors[x]][-1]
        if response in classVotes:
            classVotes[response] += 1
        else:
            classVotes[response] = 1

    sortedVotes = sorted(classVotes.items(), key=lambda x: x[1], reverse=True)

    return(sortedVotes, neighbors)

```

```

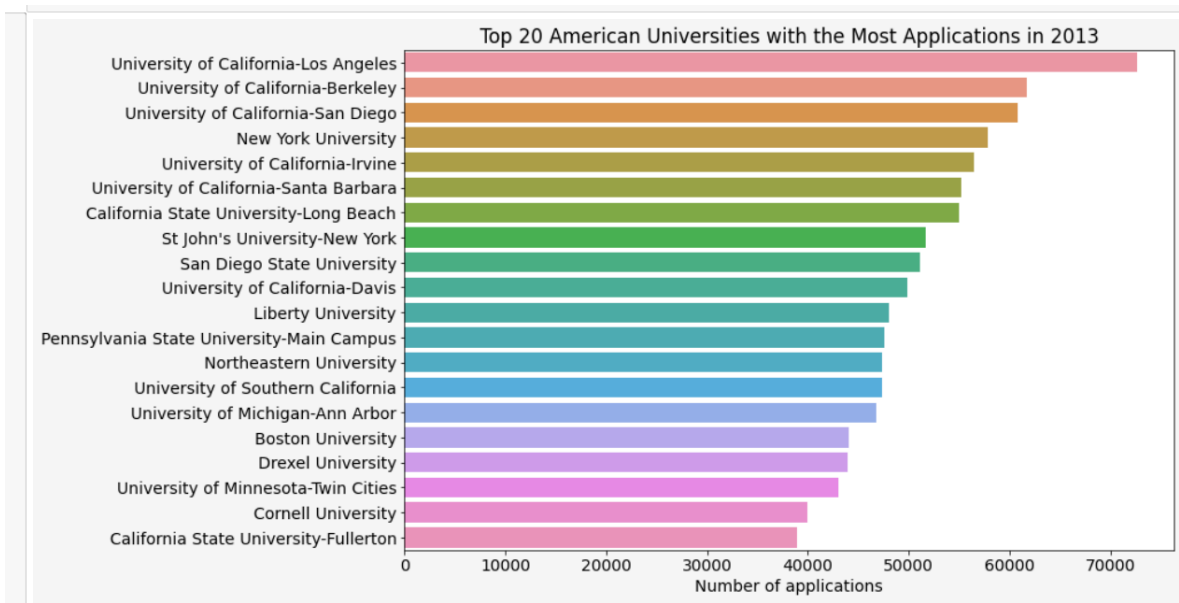
k = 7

result, neigh= knn(processed_data, test, k)

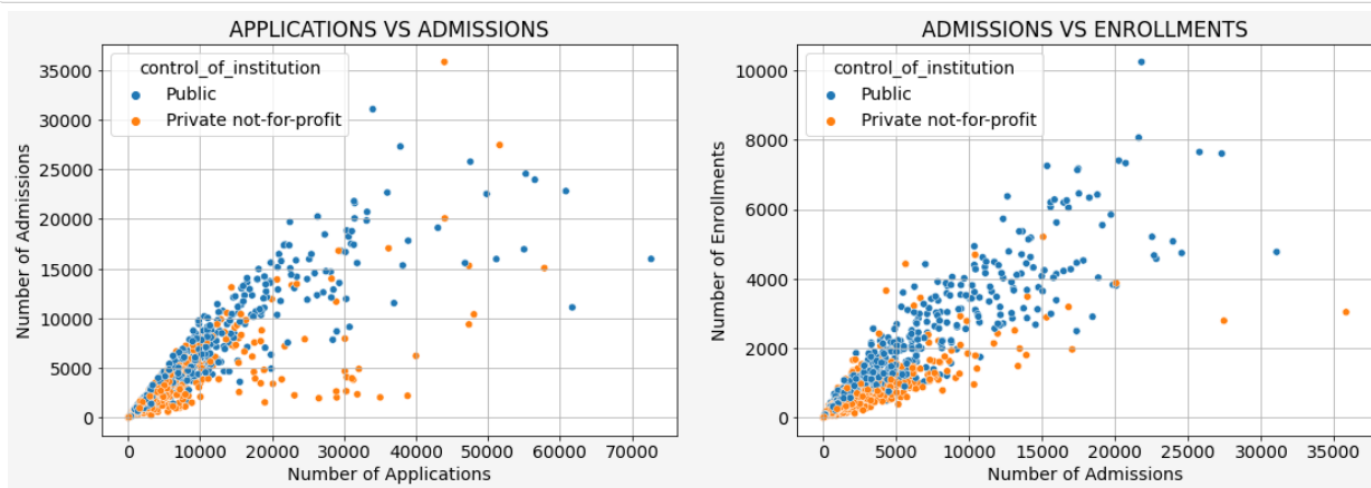
list1 = []
list2 = []
for i in result:
    list1.append(i[0])
    list2.append(i[1])
for i in list1:
    print(i)

```

# RESULT AND INFERENCES

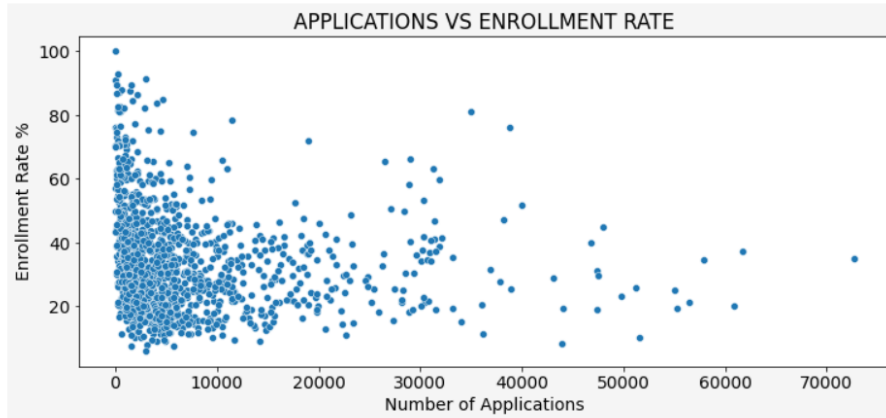


**Fig4 - Universities with most application**



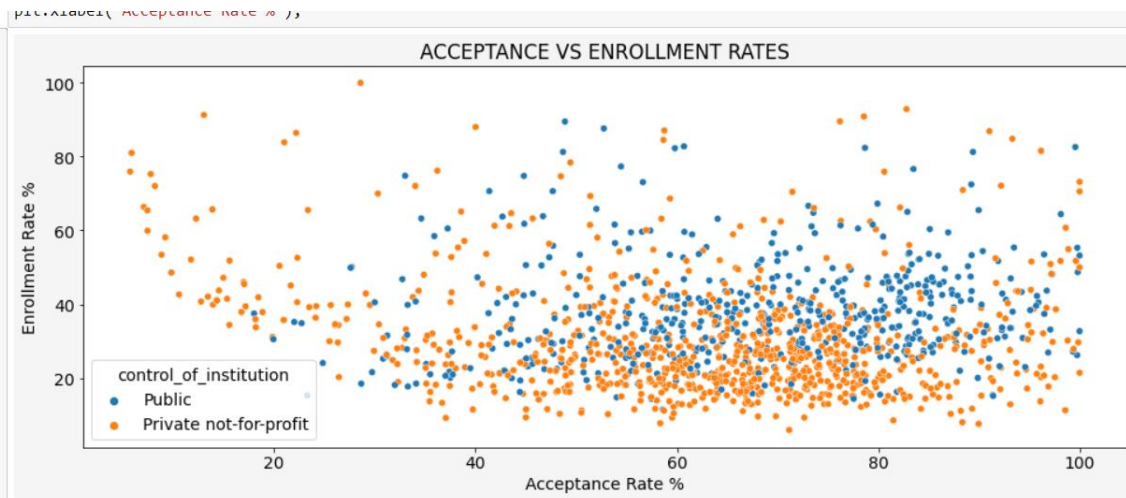
If we look at the left figure, we can see that, with a few exceptions, universities with a high number of applications also have a high number of admissions, and vice versa. However, the universities with the highest number of applications are not the ones with the highest number of admissions. Additionally, there is a batch of private universities with a high amount of applications, but their number of admissions is pretty low.

**Fig5 - Scatter plot**



This figure tells us that the universities which receive a lower number of applications are the ones with a higher enrollment rate. Obviously, there are some exceptions, but this is the strongest tendency. **Consequently, we can say that a high number of applications does not mean that a university is preferred among students.**

**Fig6 - Showing scatterplot between application vs enrollment rate**



We can see that for high acceptance rates, the enrollment rate vastly varies among public and private universities; nonetheless, there is a higher concentration where the enrollment rate is not high. That leads us to think that the acceptance rate is not a feature that strongly influences the student's preference for a

**Fig7 - Showing scatterplot between acceptance vs enrollment rate**

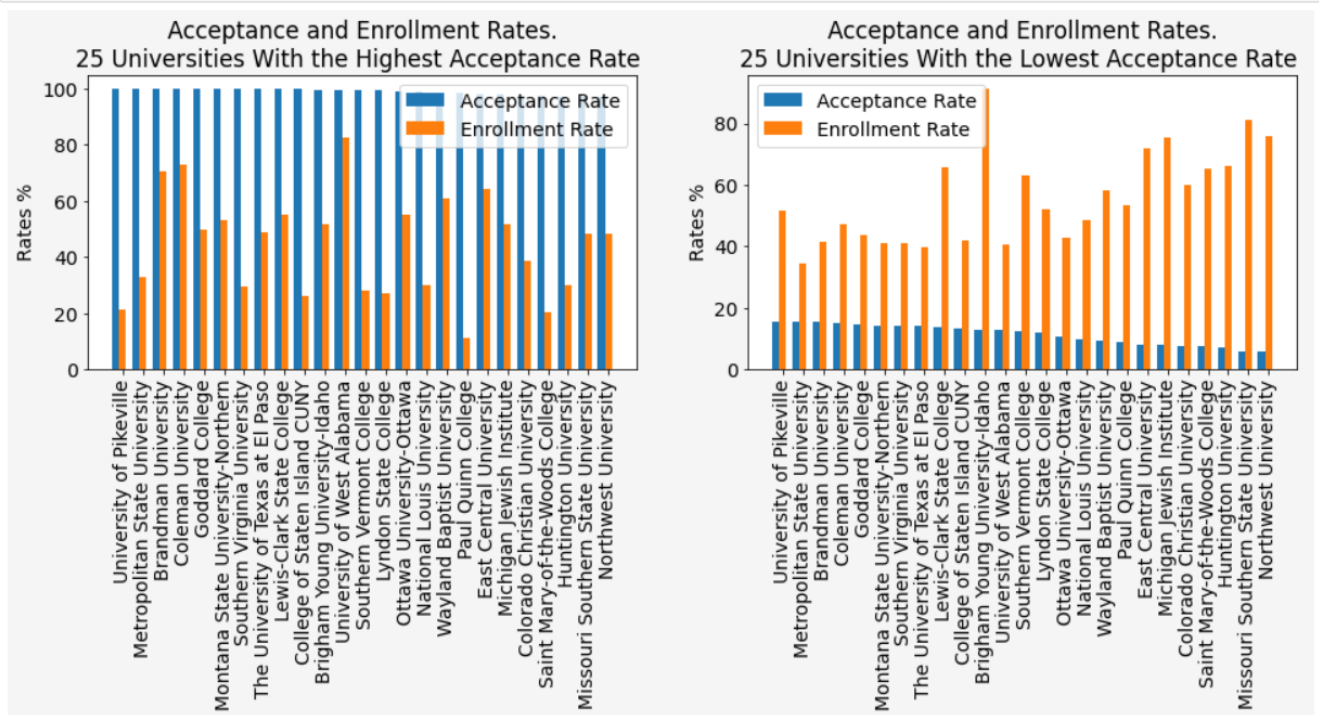
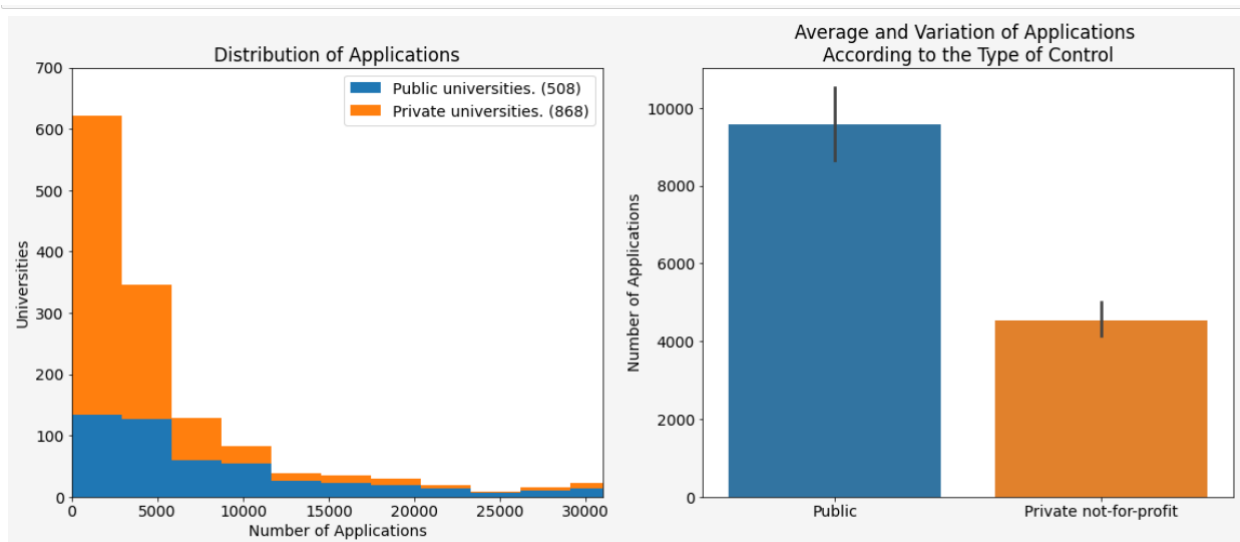


Fig8 - **Universities with highest and lowest acceptance rate**



According to these figures, the average of applications that public universities received in 2013 is virtually double the average of applications received by private universities in the same year (for public universities: around 9000, for private universities: around 4000). At this point, it's important to highlight that the

Fig9 - **Distribution of applications**

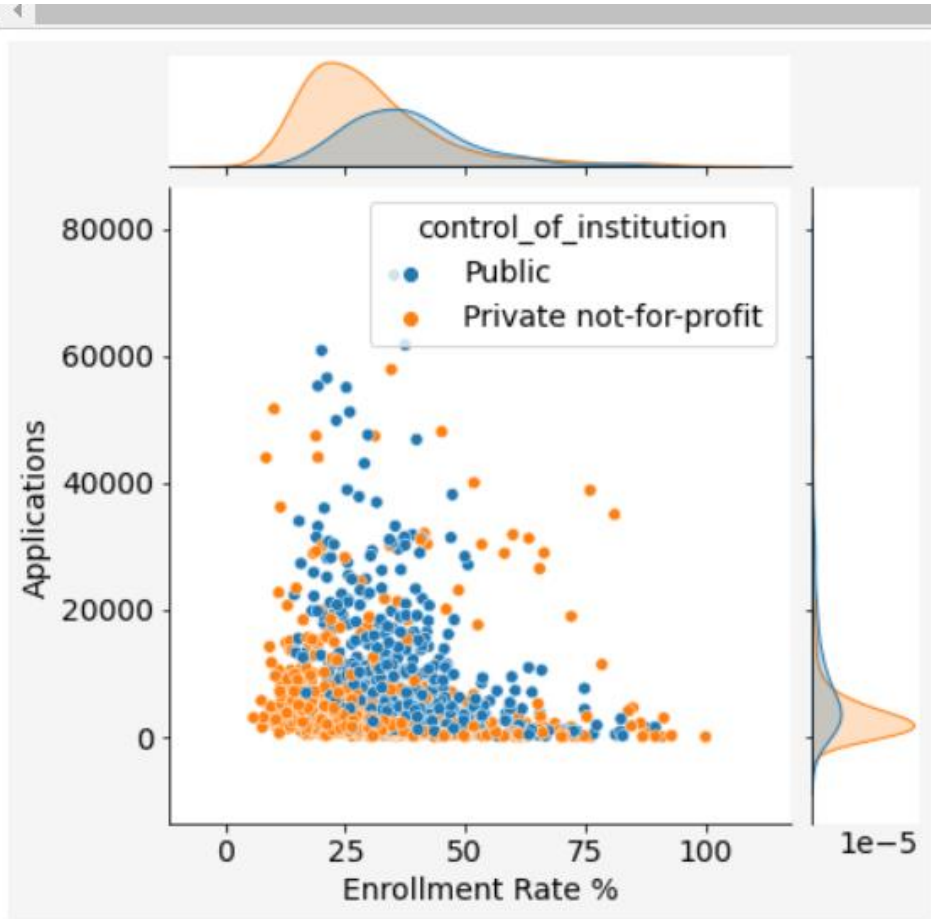


Fig10 - Scatter plot of public vs private universities

**Q: Do students prefer universities with low tuition and fees?**

```
!]: g=sns.jointplot(x=universitiesnw_df.tuition_and_fees_2013_14,y=univ  
g=(g.set_axis_labels("Tuition and Fees $","Applications"))
```

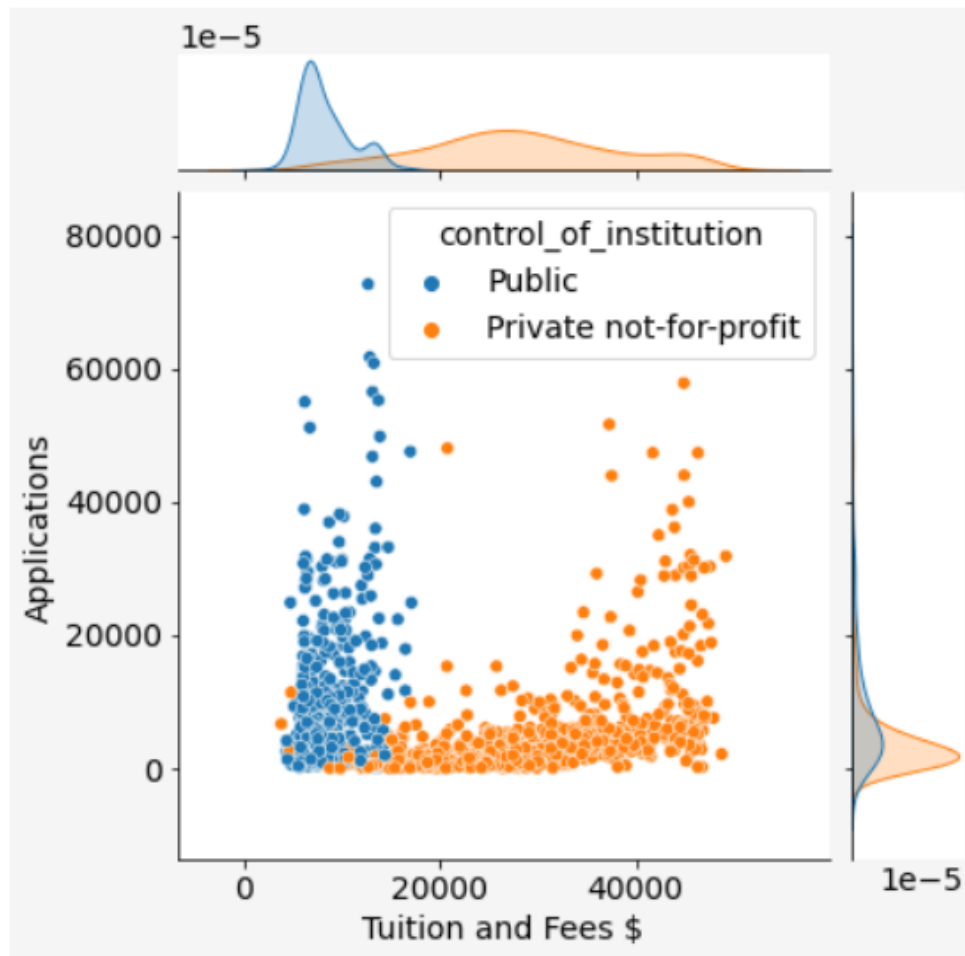


Fig11 - Scatter plot of Applications vs Tuition Fees

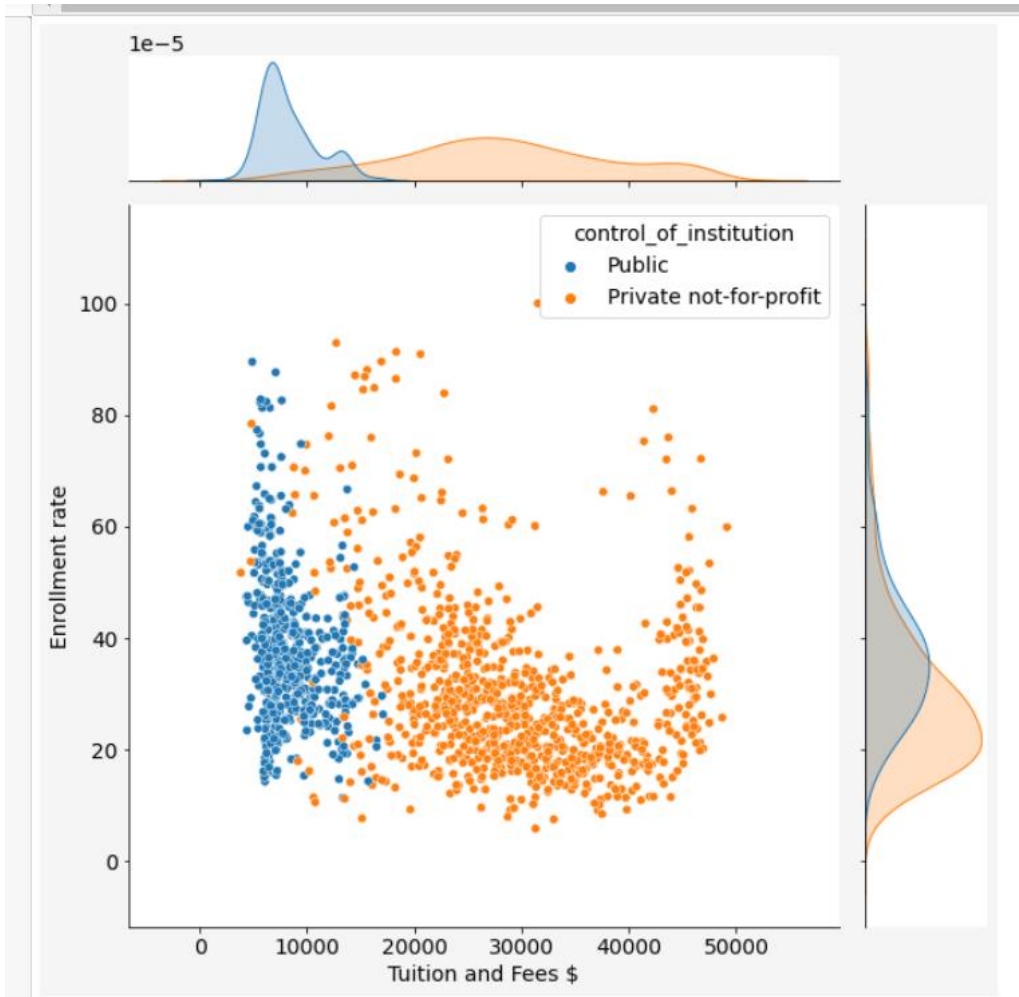


Fig12 - Scatter plot of Enrollment rate vs tuition fees

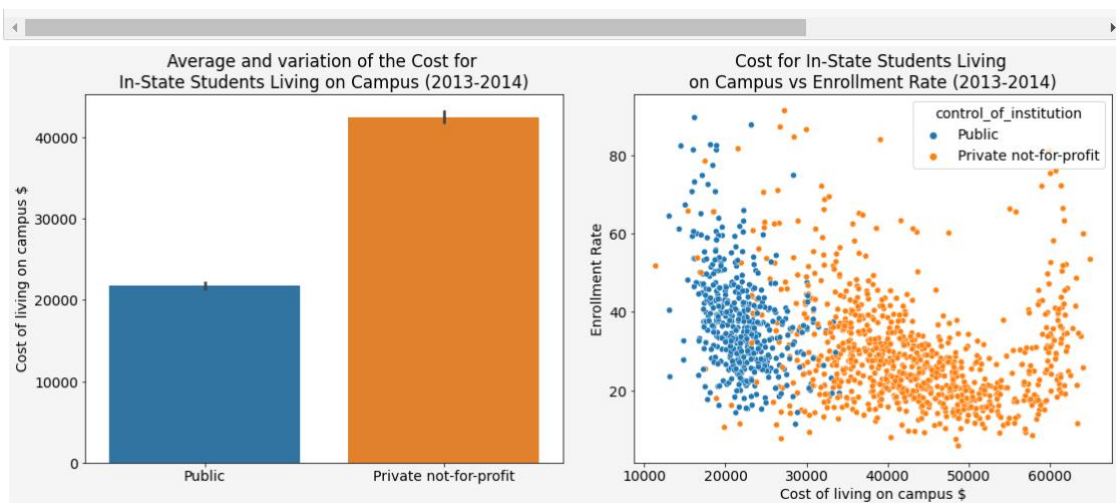
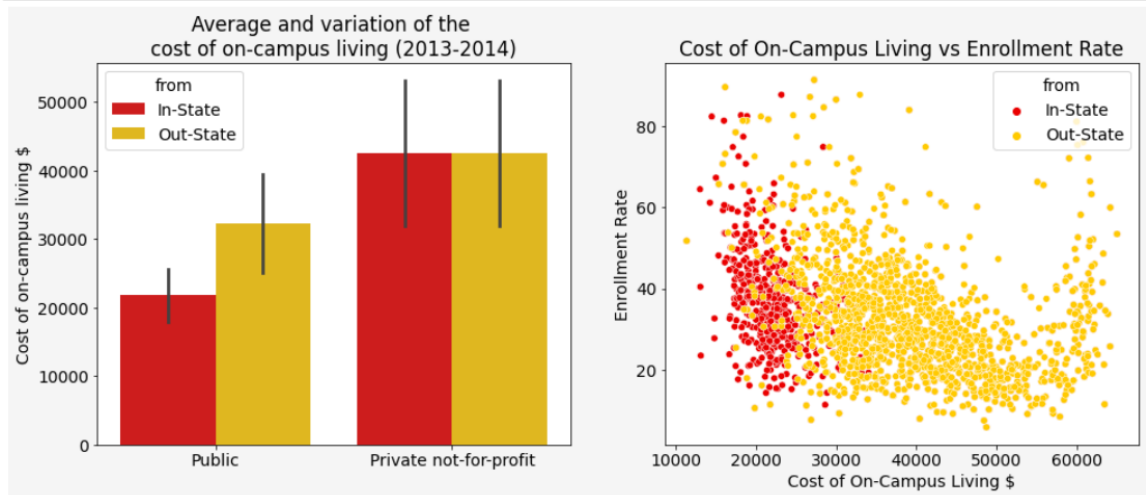
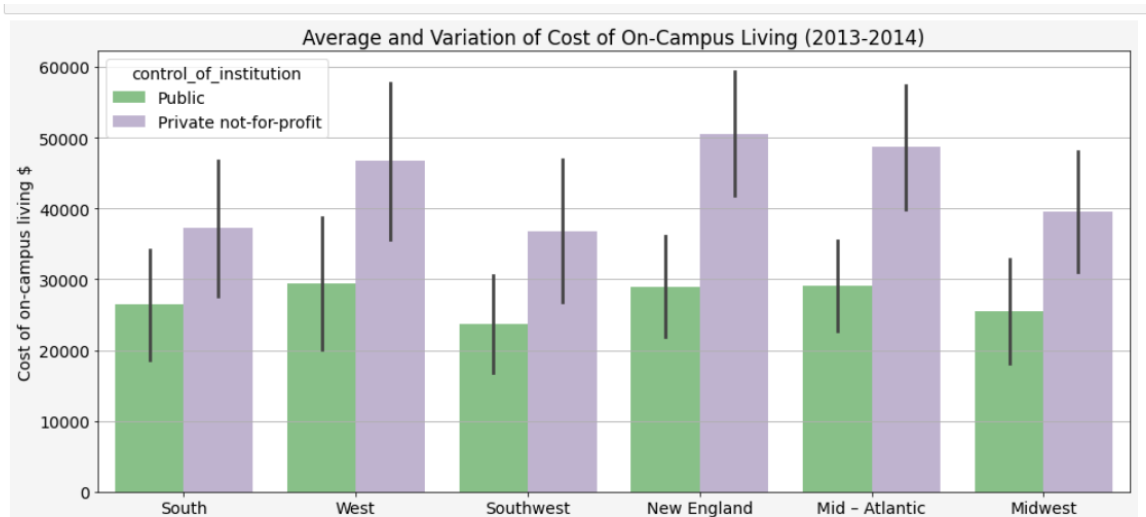


Fig13 - Barplot on Cost on living on campus and a scatter plot on Enrollment rate vs cost of living on campus



In general, we had found that *the average cost offered by public universities is lower than the cost offered by private universities*; the left figure verifies this. Now with all the costs side-by-side, we can elaborate on this by noting that *in the case of public universities, the average cost for out-state students is much higher than that for in-state students. On the other hand, in the case of private universities, the average cost for in-state and out-state students is the same.*

**Fig14 - A barplot and scatter plot on Cost of on campus living vs Enrollment rate**



We can see that the highest average cost corresponds to private universities in New England, followed by Mid-Atlantic. And the lowest average cost belongs to public universities in the Southwest.

**Fig15 - A barplot on Cost of on campus living vs location**



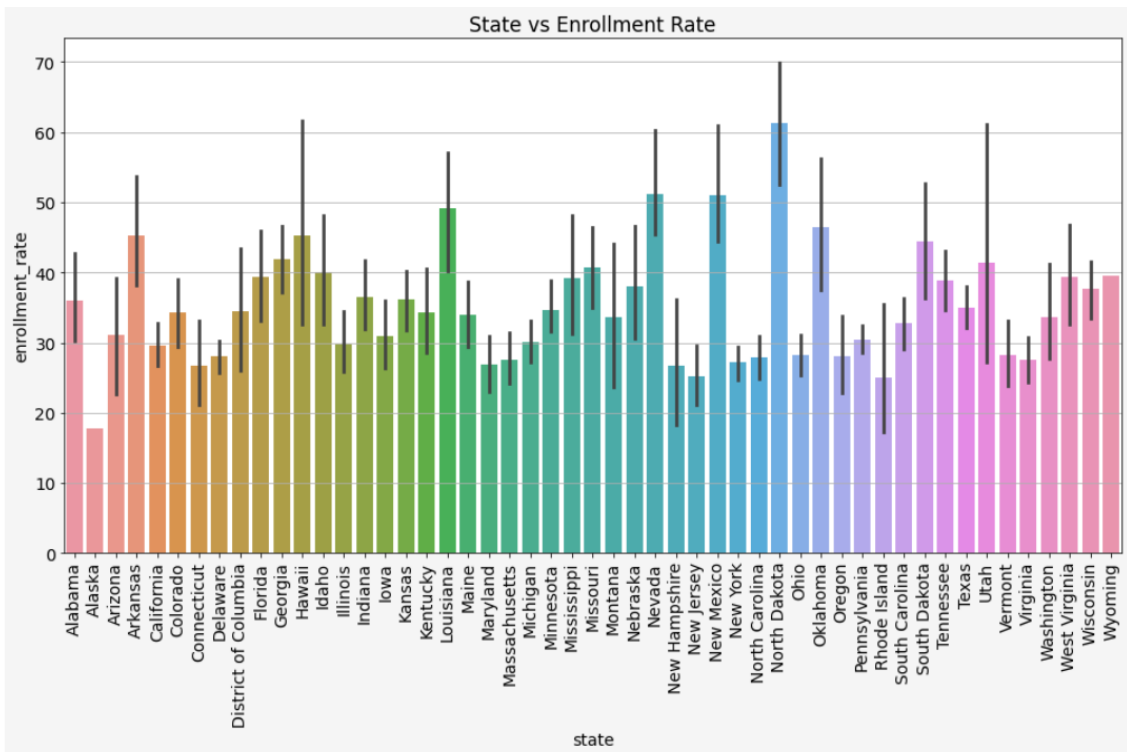


Fig16 - **Stacked bar chart on State vs Enrollment rate**

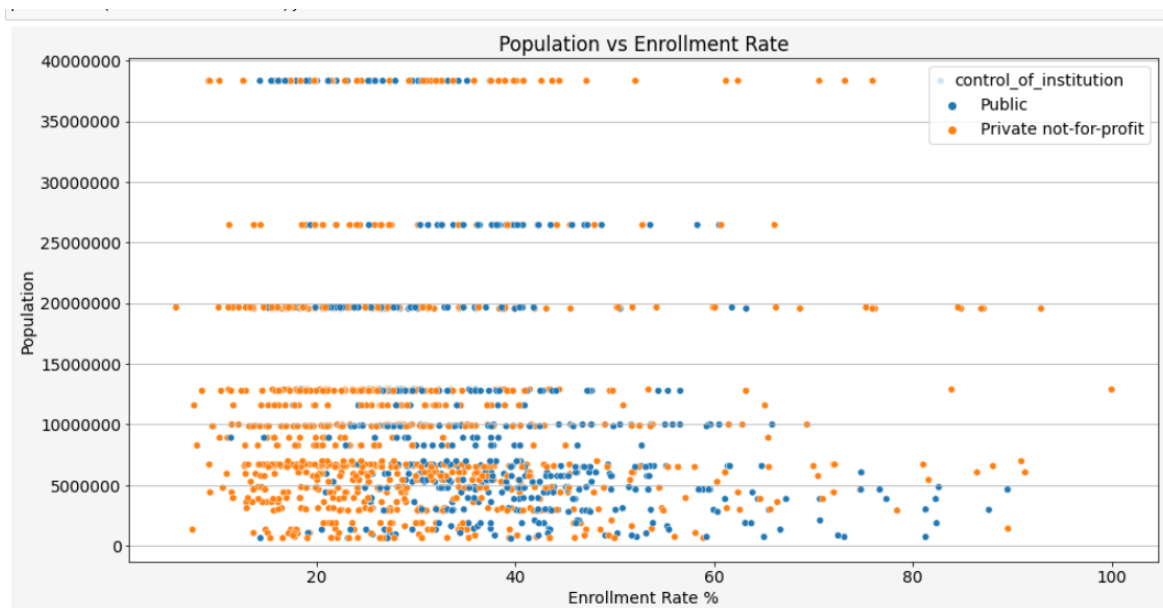


Fig17 - **A strip plot on Population vs Enrollment rate**

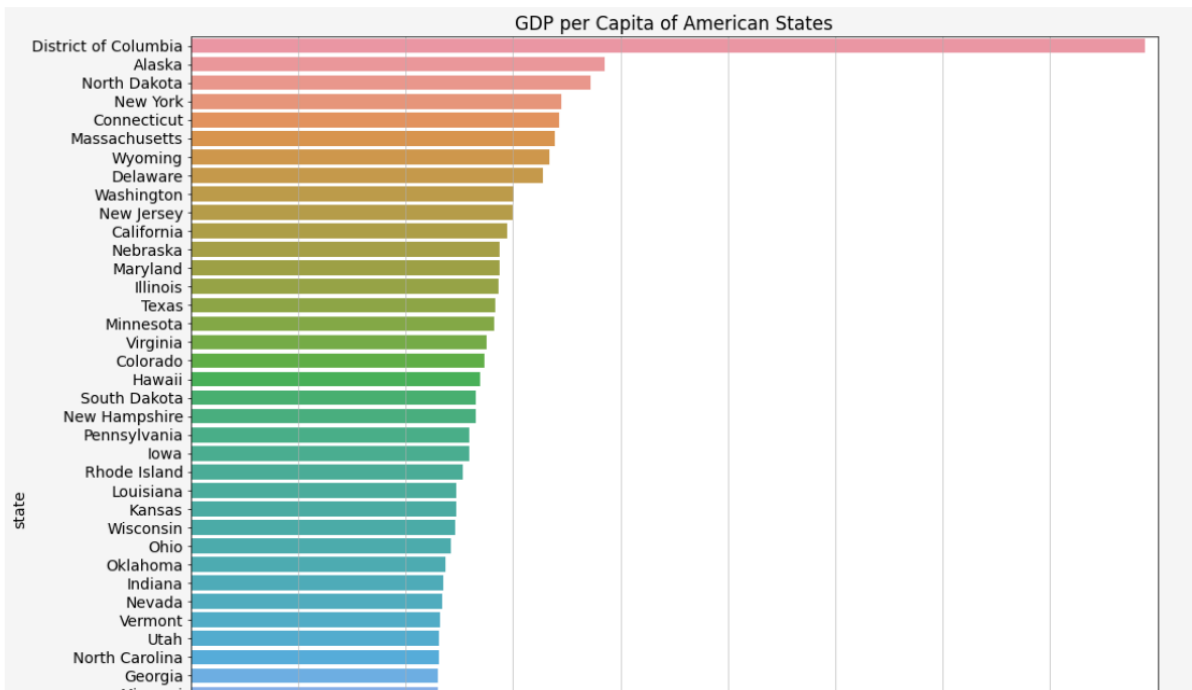


Fig18 - Horizontal bar plot on GDP per capita of American States

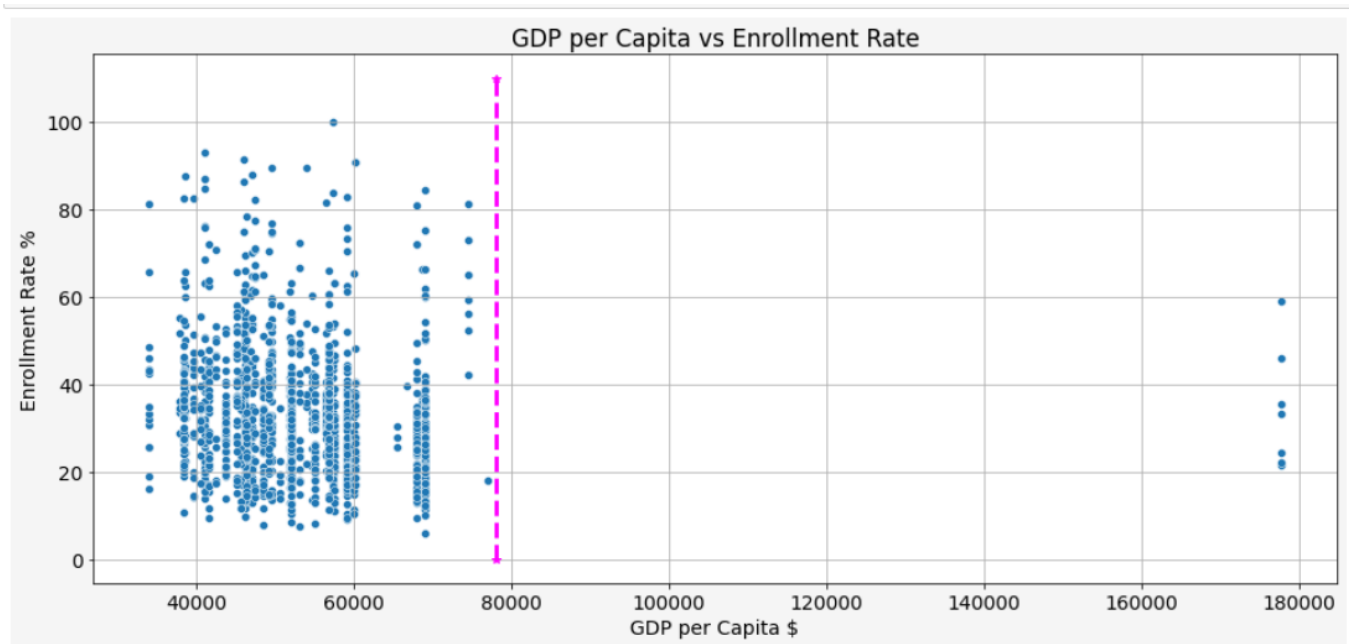


Fig19 - Strip plot on GDP per capita vs enrollment rate

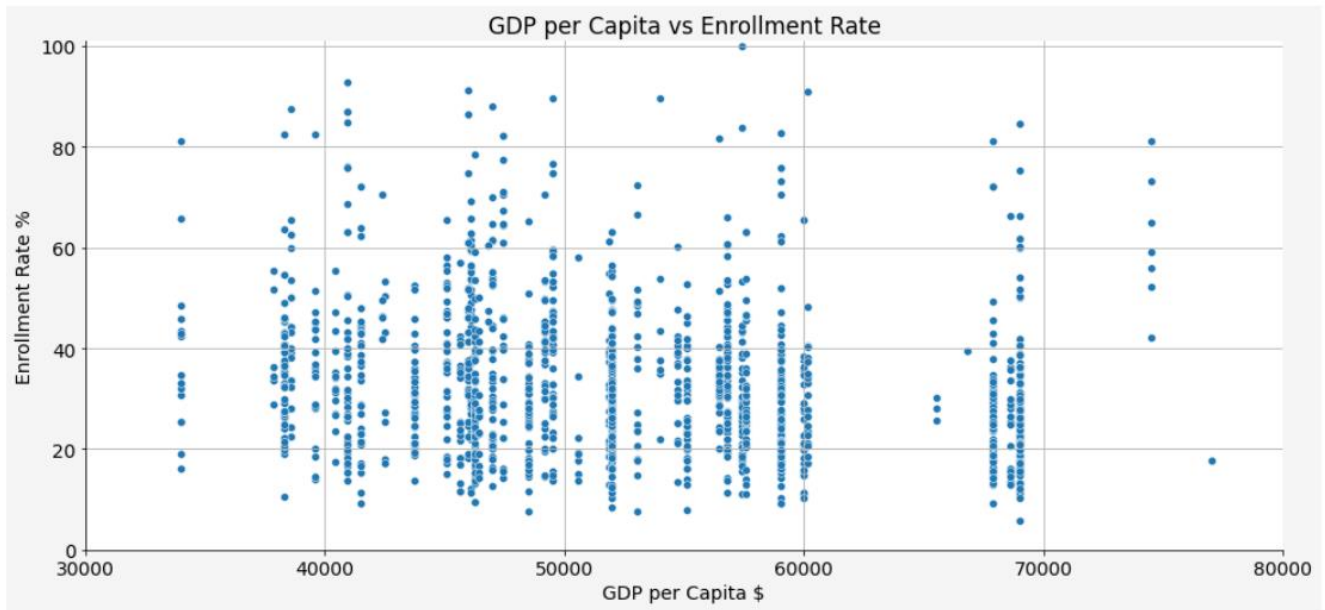


Fig20 - Scatter plot of public vs private universities

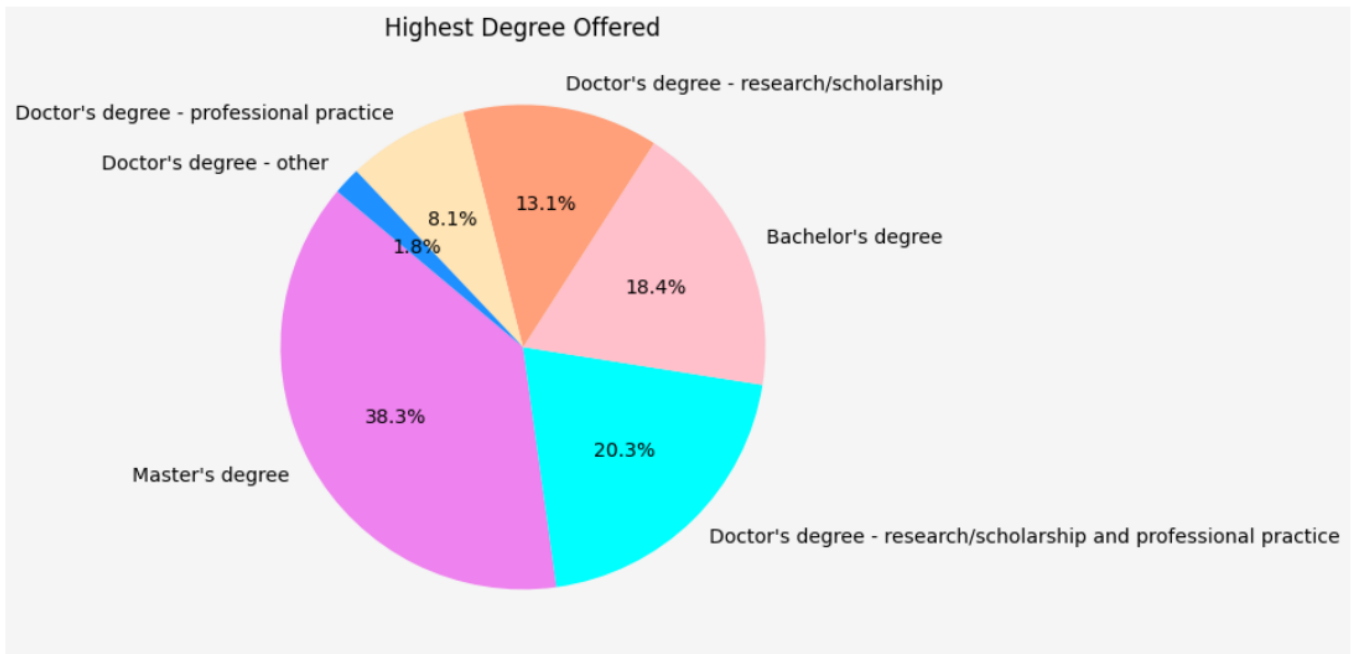


Fig21 - Pie plot on different types of degree

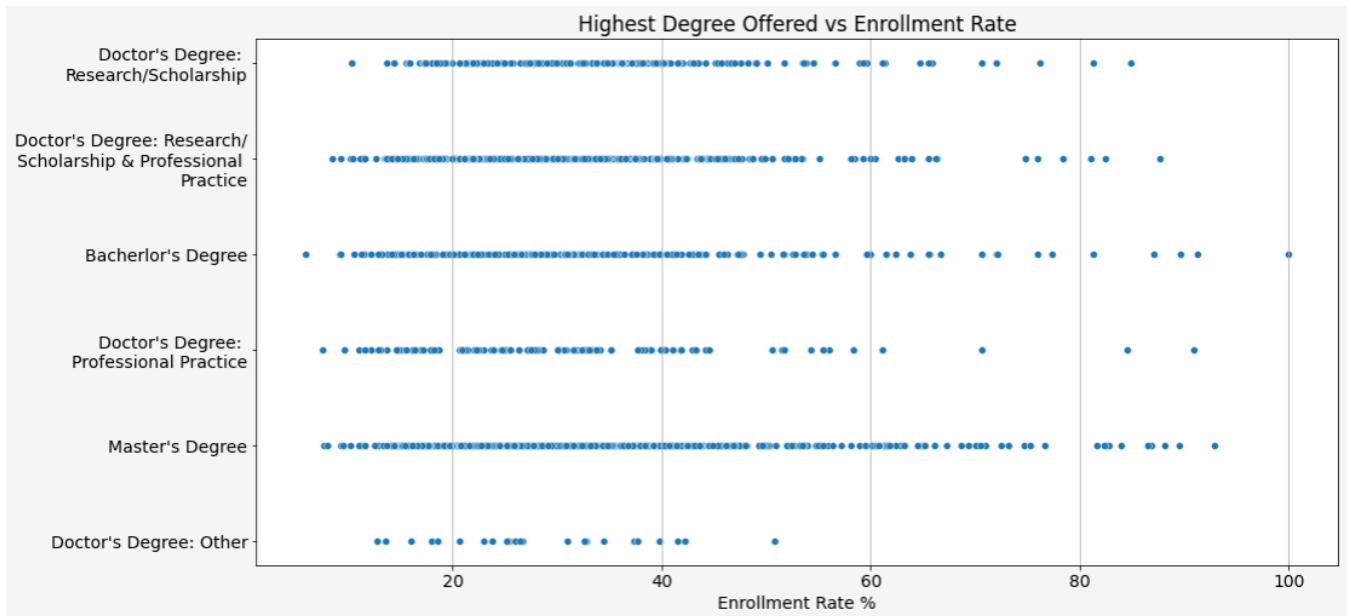


Fig22 - Strip plot on highest degree offered vs enrollment rate

The screenshot shows a web browser window with the address bar at "localhost:5000/undergraduate". The page title is "University Recommendation System" and the navigation menu includes "Home", "Undergraduate College", and "Graduate College". The main content area is titled "Under Graduate Universities" and contains the instruction "Please Enter your SAT score and Maximum Tution Fees". Below this are two input fields: "SAT Score:" and "Maximum Tution Fees:". A "Submit" button is located at the bottom of the form.

Fig 23 – Recommendation System

The screenshot shows a web browser at localhost:5000/undergraduate. The page has a dark header with the text "University Recommendation System" and navigation links for "Home", "Undergraduate College", and "Graduate College". The main content area is titled "Under Graduate Universities" and asks the user to "Please Enter your SAT score and Maximum Tution Fees". There are two input fields: "SAT Score:" with the value "1500" and "Maximum Tution Fees:" with the value "10000". A "Submit" button is located below the input fields.

Fig 24 – Input for Under Graduation recommendation

The screenshot shows the result page of the recommendation system. The browser address bar contains the URL localhost:5000/undergraduatealgo?sat=1500&tution=10000. The page header includes "Undergraduate Recommendations" and navigation links for "Home", "Undergraduate College", and "Graduate College". The main heading is "Under Graduate University Recommendation system". Below the heading, it states "The top recommended Universities based on your SAT Score & Maximum Tution Fee are". A table lists the top 5 recommended universities and their acceptance rates.

S.No	University	Acceptance Rate
1.	Brigham Young University-Idaho	0.8108348810996646
2.	Mississippi Valley State University	0.7441481637097037
3.	Alcorn State University	0.6810460877865222
4.	The University of Texas of the Permian Basin	0.6696032133488956
5.	Delta State University	0.637365259808179

Fig 25 – Output for Under Graduation recommendation

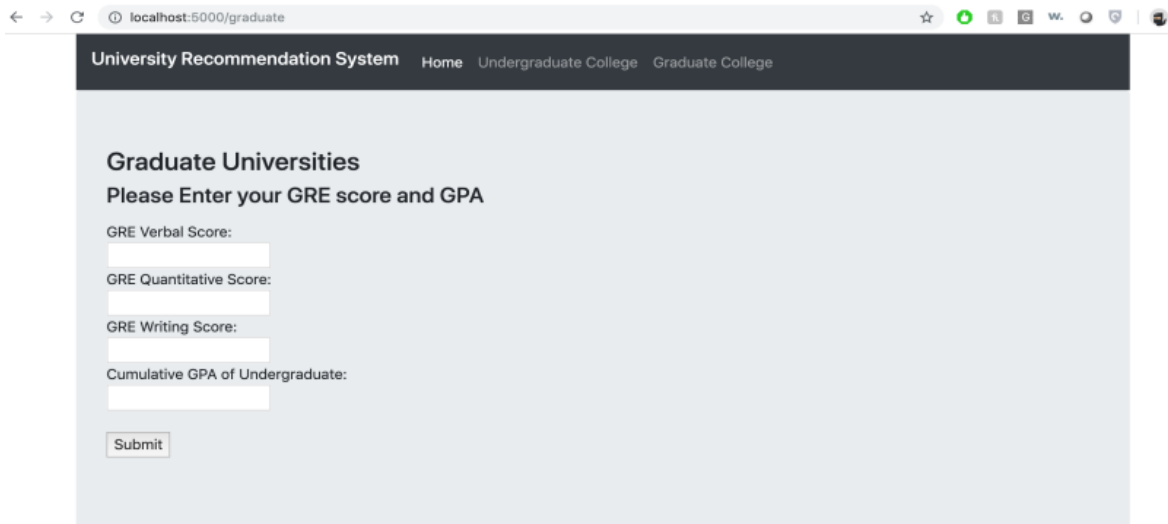


Fig 26 – Recommendation System for Graduation Universities

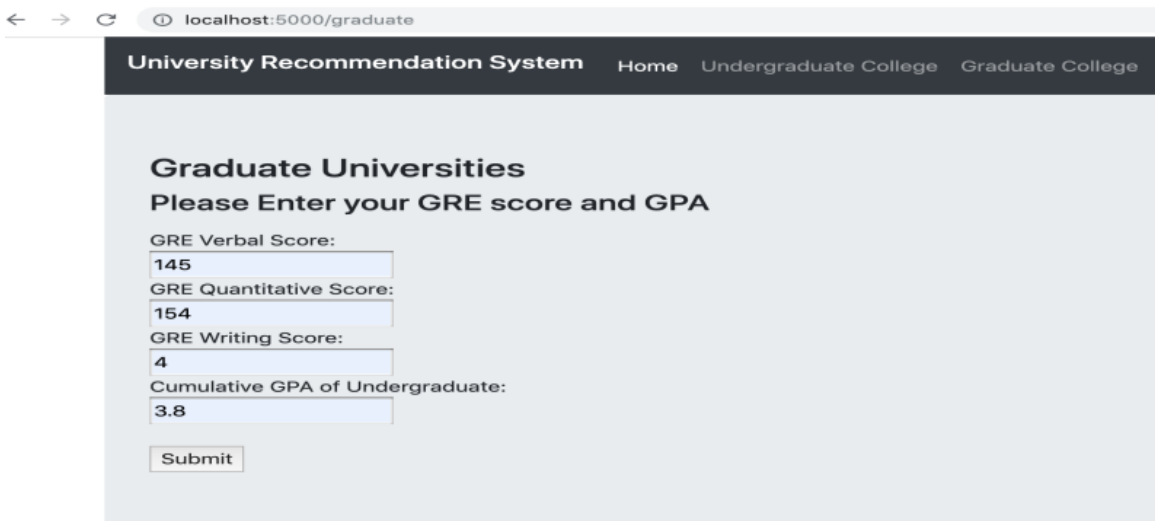
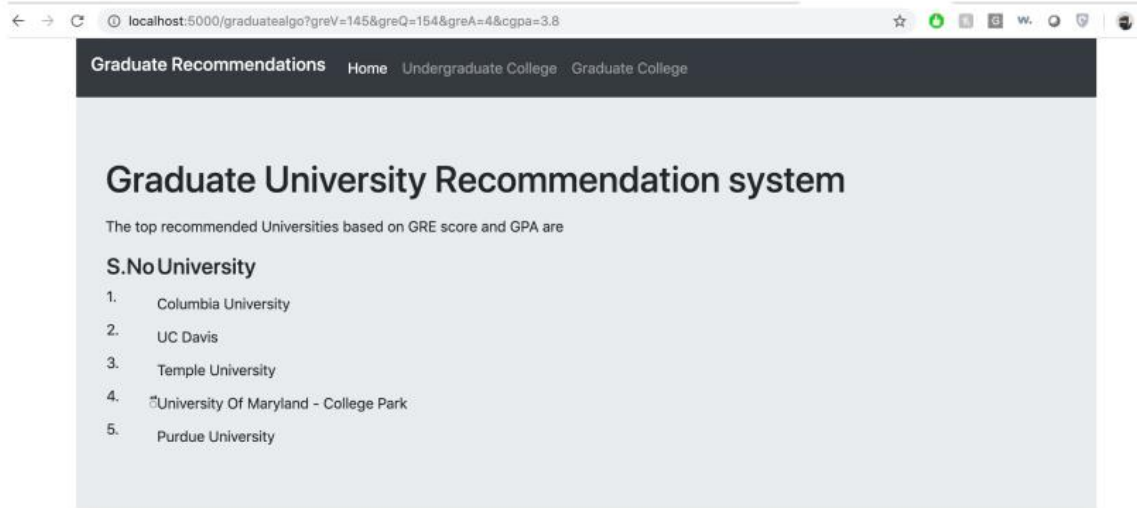


Fig 27 – Input for Graduation recommendation



---

Fig 28 – Output for Graduation recommendation

# CONCLUSION

A high number of applications does not imply that a university is preferred among students. In fact, the universities that receive a lower number of applications are the ones with a higher enrollment rate. Obviously, there are some exceptions, but this is the strongest tendency. Based on the lack of a strong pattern among admissions and the enrollment rate, we can say that students do not necessarily prefer a university because of its high acceptance rate or, in other words, the students' preference is not based on how easy it is for them to be admitted to a university. By analyzing the enrollment rate, we saw that this rate, on average, is higher for public universities than the average for private universities. So, there is a strong students' preference for public universities. When it comes to tuition and fees, students prefer affordable universities. Additionally, the reason for the students' preference for public universities is that public universities are much more affordable than the majority of private universities. In all the analyses made to find a pattern about costs for on-campus living, we found a high enrollment rate more frequently when costs are affordable. This means that students, in-state and out-state students, prefer universities with affordable costs of on-campus living. The majority of public universities offer a much more affordable price for in-state students than private universities.

The average cost of living for out-state students that public universities offer is higher than that for in-state students. However, the average cost that private universities offer does not make a distinction between in-state and out-state students. Since there was no firm trend when analyzing the state population with enrollment rates, we cannot say that students prefer universities of crowded states. Students do not prefer a university because of the GDP per capita of the state where the university is located. In other words, students do not choose a university based on the overall well-being of states. When students look for a university to study for a Bachelor's degree, they do not frequently choose the university thinking about a future possibility of pursuing a higher degree at the same university.



# FUTURE WORK

A possible future work could make a similar analysis but making a stronger consideration about how the different degrees offered by universities can modify the features studied in this project, for instance, the tuition and fees, the living on campus, etc.

It could also be interesting to see if the preference of undergraduate students differs from that of graduate students: a master's degree and a Ph.D.

Another future analysis could take into consideration which are the most preferred universities among the different students' races. The original dataset contains information about the percentage of enrollment according to different races; however, hundreds of universities have missing values in these categories.

A more accurate analysis should consider an enrollment rate for in-state students and another for out-state students.

A Recommender System which suggests universities

# REFERENCES

1. Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, and Chad Marston, "A Comprehensive Review of Tools for Exploratory Analysis of Tabular Industrial Datasets," Visual Informatics, Volume 2, Issue 4, December 2018, pp. 235-253.
2. John T. Behrens, "Principles and Procedures of Exploratory Data Analysis," Psychological Methods, 1997, Vol. 2, No. 2, pp.131-160
3. Exploratory data analysis – From Wikipedia, the free encyclopedia [Online], Available: [https://en.wikipedia.org/wiki/Exploratory\\_data\\_analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis)
4. <https://seaborn.pydata.org/>
5. <https://pandas.pydata.org/docs/>
6. <https://towardsdatascience.com/>
7. <https://stackoverflow.com/>
8. <https://matplotlib.org/>
9. <https://numpy.org/>
10. [www.kaggle.com](http://www.kaggle.com)