College Admission

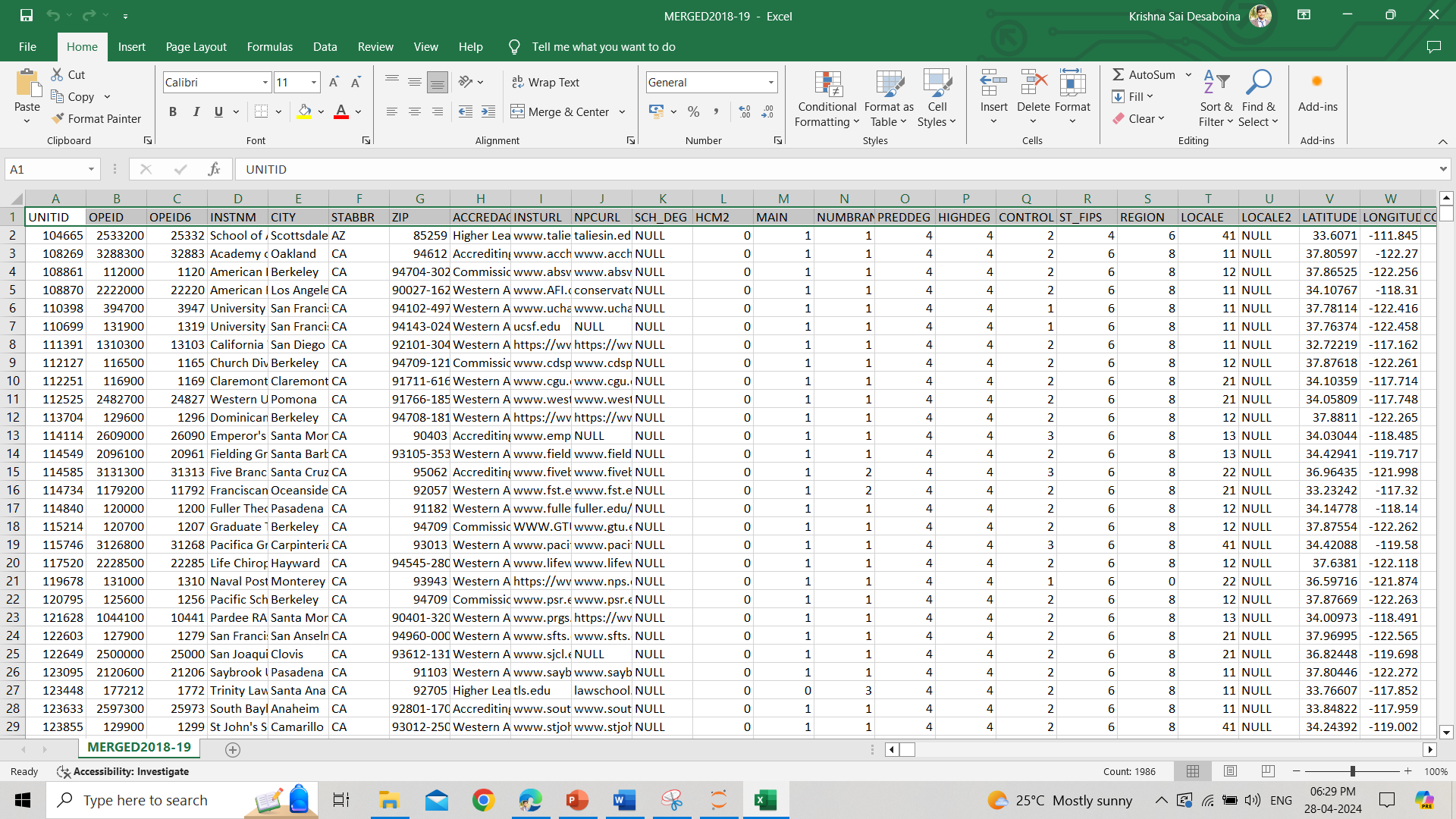
**Introduction:**

In this competitive world, pursuing higher education became a mode of opportunity and advancement. Countless high school seniors are navigating through different web applications to know their chance of getting to their preferred college.

The college admission rate, a measure that uses selectivity, has become a key factor of today's dynamics within higher education. Not only being a mere statistic, it also became a prestige of the University.

The College Admission Data Set (MERGED2018-19) consists of the following columns:

**UNITID,** **OPEID, OPEID6, INSTNM, CITY, REGION, ACCREDAGENCY, INSTURL, LOCALE, ADM\_RATE. ST\_FIPS, MAIN, NUMBRANCH, PREDDEG, HIGHDEG, CONTROL, etc.**



**df=pd.read\_csv("MERGED2018-19.csv")**

**MERGED2018-19.csv file data is converted to Data Frame by Pandas Package**

**CSV Structured file:**

CSV file, also called Comma Separated Values file, is a type of text file that stores the data in tabular format. It is an example of a structured format. Each line of data represents a row and in that row the values are differentiated by a delimiter comma ‘,’ or other delimiters like semi-colon ‘;’ or tab is used.

CSV files are used not only for storing and exchanging data but also for software applications. The storage is easy as they are simple and lightweight and read and written my many packages as pandas in python.

**Data Frame:**

The data frame is a fundamental data structure used by pandas in Python for some manipulating operations like dropping a row, counting the non-null values, and finding the type of data stored in each column, etc. A data Frame is data-organized in rows and columns and type of a structured data same as a spreadsheet or a database table.

**Data Cleaning:**

**Dropped all the ID’s (UNITID,** **OPEID, OPEID6) and Unwanted Columns:**

**df=df.drop(columns='UNITID')**

**df=df.drop(columns='OPEID')**

**df=df.drop(columns='OPEID6')**

**df=df.drop(columns='INSTURL')**

**df=df.drop(columns='NPCURL') etc,**

**Exploratory Data Analysis (EDA):**

**Dropped rows of Null values in ADM\_RATE**

df=df.dropna(subset=['ADM\_RATE'])

<class 'pandas.core.frame.DataFrame'>

Int64Index: 2006 entries, 772 to 6795

Data columns (total 15 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 INSTNM 2006 non-null object

1 CITY 2006 non-null object

2 STABBR 2006 non-null object

3 ZIP 2006 non-null object

4 ACCREDAGENCY 1983 non-null object

5 HCM2 2006 non-null int64

6 MAIN 2006 non-null int64

7 NUMBRANCH 2006 non-null int64

8 PREDDEG 2006 non-null int64

9 HIGHDEG 2006 non-null int64

10 CONTROL 2006 non-null int64

11 ST\_FIPS 2006 non-null int64

12 REGION 2006 non-null int64

13 LOCALE 2006 non-null float64

14 ADM\_RATE 2006 non-null float64

dtypes: float64(2), int64(8), object(5)

memory usage: 250.8+ KB

INSTNM – Institute Name (Categorical)

CITY – The City in which the institute is located. (Categorical)

ZIP – Zip Code

ACCREDAGENCY – Accreditor for the institution. (Categorical)

HCM2 - Schools that are on Heightened Cash Monitoring 2 by the Department of Education

MAIN - Flag for the main campus ()

|  |  |
| --- | --- |
| 0 | Not main campus |
| 1 | Main campus |

NUMBRANCH - Number of branch campuses

PREDDEG - Predominant undergraduate degree awarded  
 0 Not classified  
 1 Predominantly certificate-degree granting  
 2 Predominantly associate's-degree-granting  
 3 Predominantly bachelor's-degree-granting  
 4 Entirely graduate-degree granting

HIGHDEG - Highest degree awarded  
 0 Non-degree-granting  
 1 Certificate degree  
 2 Associate degree  
 3 Bachelor's degree  
 4 Graduate degree

CONTROL - Control of institution (IPEDS)

|  |  |
| --- | --- |
| 1 | Public |
| 2 | Private nonprofit |
| 3 | Private for-profit |

ST\_FIPS - FIPS code for state

REGION - Region (IPEDS)

LOCALE - Locale of institution

|  |  |
| --- | --- |
| 11 | City: Large (population of 250,000 or more) |
| 12 | City: Midsize (population of at least 100,000 but less than 250,000) |
| 13 | City: Small (population less than 100,000) |
| 21 | Suburb: Large (outside principal city, in urbanized area with population of 250,000 or more) |
| 22 | Suburb: Midsize (outside principal city, in urbanized area with population of at least 100,000 but less than 250,000) |
| 23 | Suburb: Small (outside principal city, in urbanized area with population less than 100,000) |
| 31 | Town: Fringe (in urban cluster up to 10 miles from an urbanized area) |
| 32 | Town: Distant (in urban cluster more than 10 miles and up to 35 miles from an urbanized area) |
| 33 | Town: Remote (in urban cluster more than 35 miles from an urbanized area) |
| 41 | Rural: Fringe (rural territory up to 5 miles from an urbanized area or up to 2.5 miles from an urban cluster) |
| 42 | Rural: Distant (rural territory more than 5 miles but up to 25 miles from an urbanized area or more than 2.5 and up to 10 miles from an urban cluster) |
| 43 | Rural: Remote (rural territory more than 25 miles from an urbanized area and more than 10 miles from an urban cluster) |

**Found that there are Outliers by using Tableau boxplot.**

**Outliers:**

Outliers are data points in a dataset that are different from other data points in a dataset. They are caused during the preparation of the dataset by measurement errors, rare events, or due to variability in the experiment. Identifying and eliminating the outliers is a prominent task in preprocessing and in EDA (Exploratory Data Analysis) so there should not be any miscalculations in statistics and any wrong predictions in Machine learning. These outliers can be identified by one of the methods of graph called Boxplot.

**Boxplot:**

The box plot is also called a box-and-whisker plot. It is a graphical representation of the distribution of a dataset based. Box-and-whisker plots mostly have five summary statistics. They are minimum, first quartile (Q1), median or second quartile (Q2), third quartile (Q3), and maximum.

**Minimum:** The smallest value in the dataset that excludes the outliers.

**First Quartile (Q1):** The value below the 25 percentile of data point.

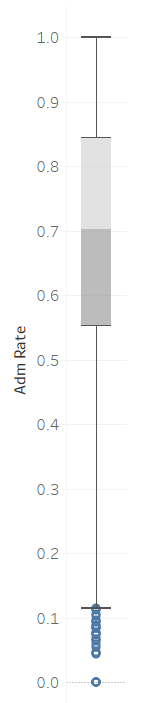
**Median (Q2):** The middle value of the dataset. It separates the lower 50 percentile and the upper 50 percentile of data.

**Third Quartile (Q3):** The value below the 75 percentile of the data.

**Maximum:** The largest value in the dataset that excludes the outliers.

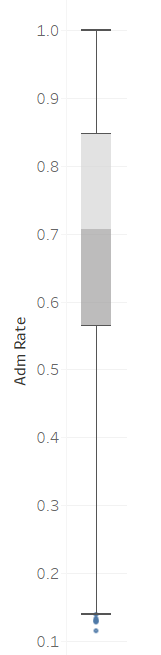
Boxplot can be used to identify the central tendency, and skewness and also for the detection of outliers.

The College Admission Data Set consists of Outliers

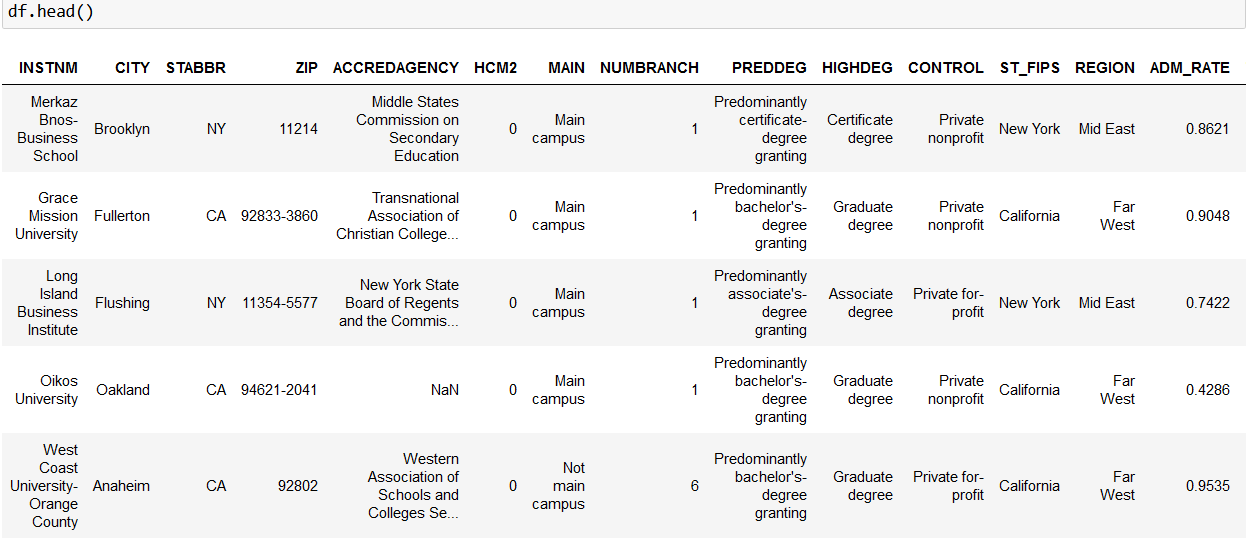


Removed the Outliers using Pandas

df=df[df.ADM\_RATE>0.115]



Converted Flags and Nominal Data values to Categorical Values by replacing those values with corresponding String values.



Converted LOCALE into two separate columns:

LOCALE - Locale of institution

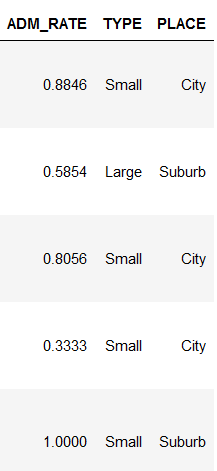
|  |  |
| --- | --- |
| 11 | City: Large (population of 250,000 or more) |
| 12 | City: Midsize (population of at least 100,000 but less than 250,000) |
| 13 | City: Small (population less than 100,000) |
| 21 | Suburb: Large (outside principal city, in urbanized area with population of 250,000 or more) |
| 22 | Suburb: Midsize (outside principal city, in urbanized area with population of at least 100,000 but less than 250,000) |
| 23 | Suburb: Small (outside principal city, in urbanized area with population less than 100,000) |
| 31 | Town: Fringe (in urban cluster up to 10 miles from an urbanized area) |
| 32 | Town: Distant (in urban cluster more than 10 miles and up to 35 miles from an urbanized area) |
| 33 | Town: Remote (in urban cluster more than 35 miles from an urbanized area) |
| 41 | Rural: Fringe (rural territory up to 5 miles from an urbanized area or up to 2.5 miles from an urban cluster) |
| 42 | Rural: Distant (rural territory more than 5 miles but up to 25 miles from an urbanized area or more than 2.5 and up to 10 miles from an urban cluster) |
| 43 | Rural: Remote (rural territory more than 25 miles from an urbanized area and more than 10 miles from an urban cluster) |

New columns:

PLACE, TYPE

PLACE - 'City', 'Suburb', 'Town', 'Rural'

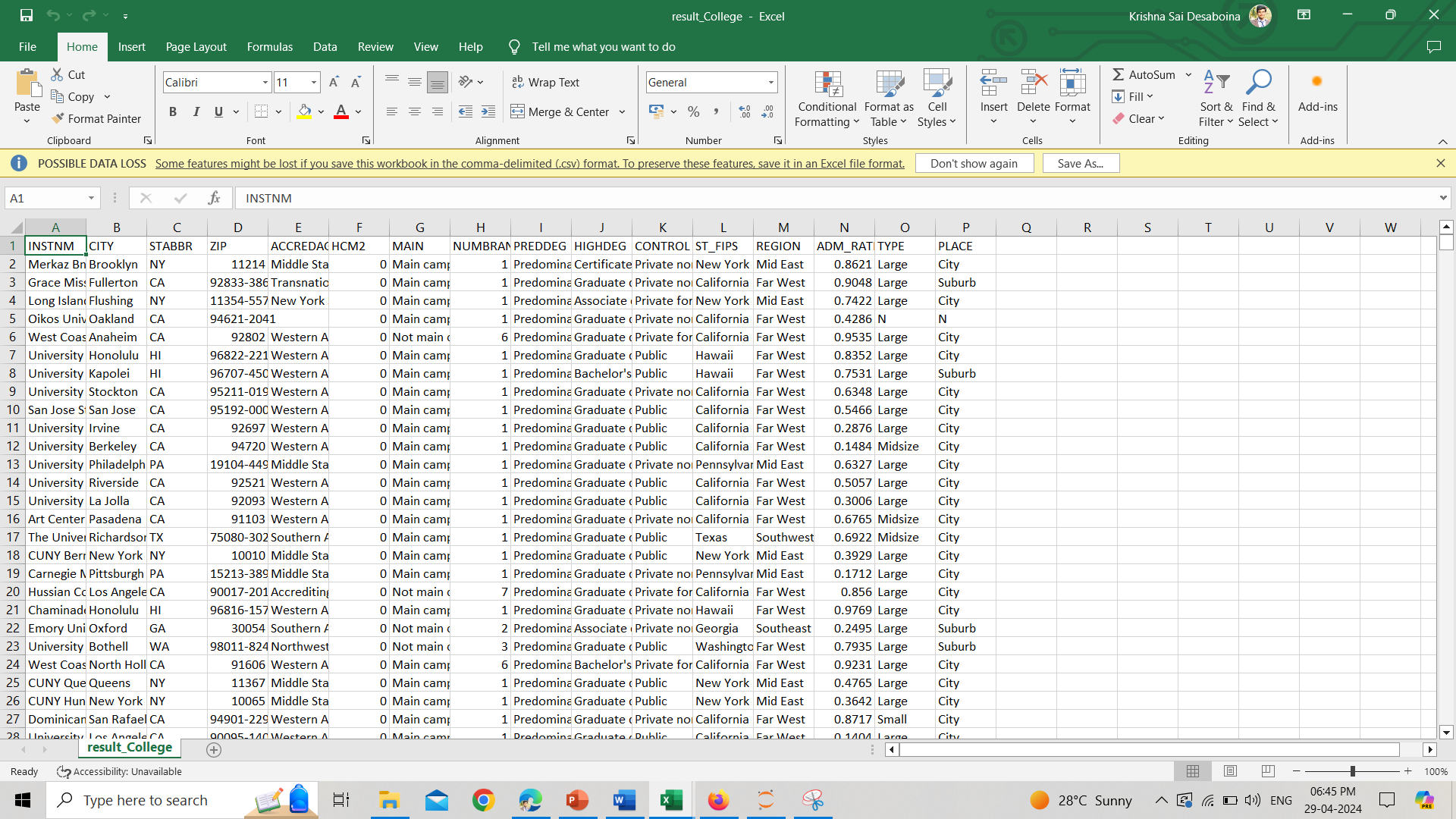
TYPE - 'Fringe', 'Distant', 'Remote', 'Large', 'Midsize', 'Small'



In the Data Visualization part, the Type and Place will be converted into Hire

**Finally Converting the Data Frame back to a new .csv file.**

df.to\_csv('result\_College.csv', index=False)

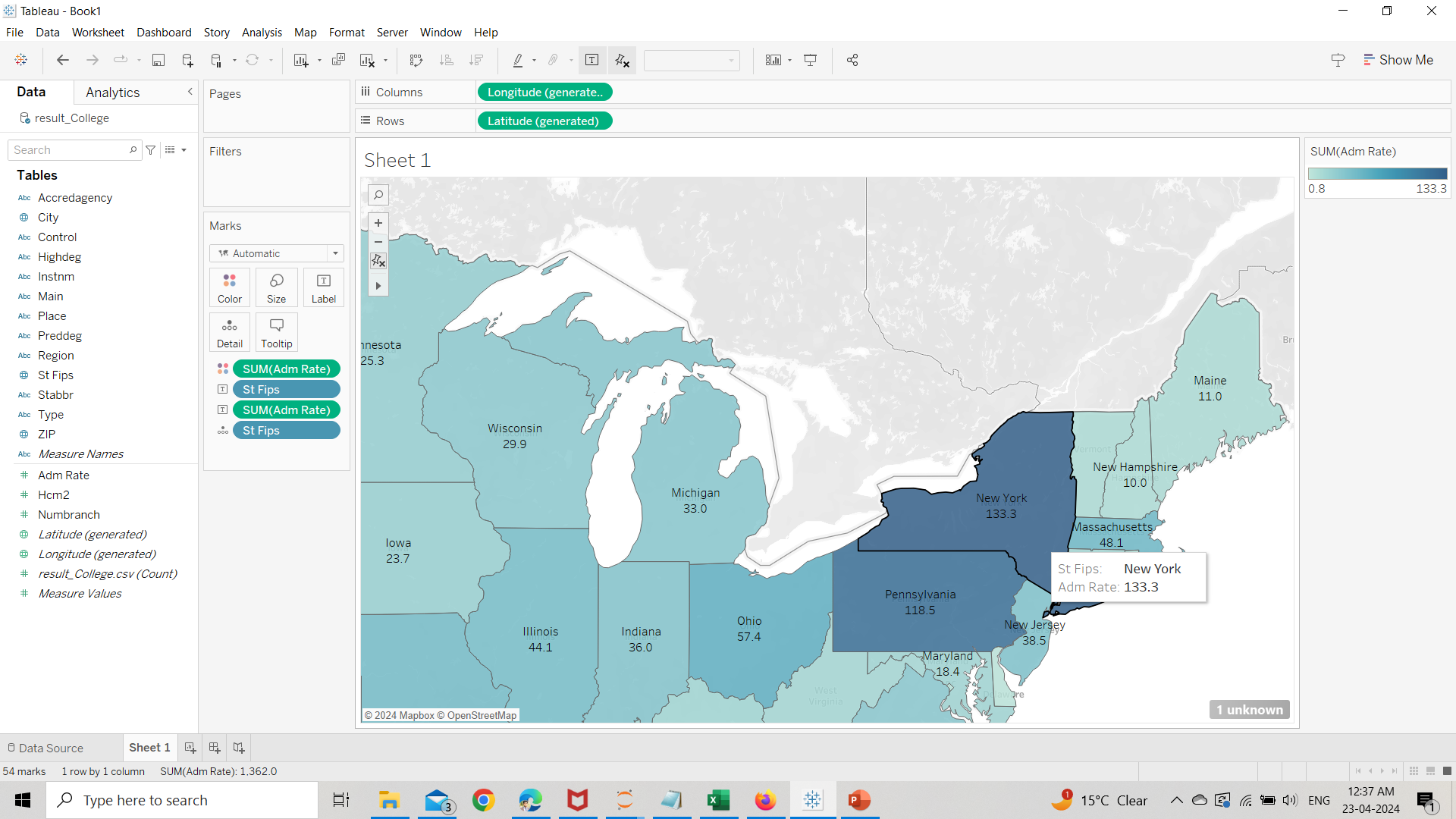


**Hypothesis:**

* Identifying the college with a high Admission rate compared to other colleges and in which state the Admission rate is high compared to other states in United States.
* What kind of place is the University situated, and has a high Admission Rate?
* Identifying the region for which the admission rate is high compared to other United States regions.
* Which University has the highest number of Branches compared to other Universities in the United States?

**Data Visualization:**

Which State has the Highest Admission Rate?



New York has the Highest Admission Rate.

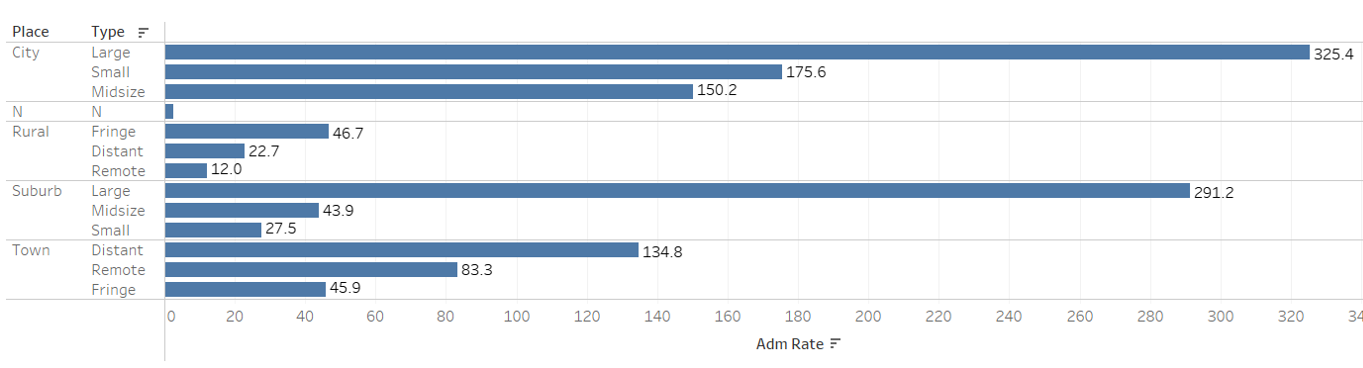
**Geographical Map:**

A geographical map is a visual representation of the Earth's surface in the form of a map. The geographical map depicts various features such as land, water, and boundaries.

Geographical maps are useful for many purposes like navigation, exploration, urban planning, and environmental management. There are different types of geographical maps. they are topographic maps, political maps, thematic maps, and digital maps.

Overall, geographical maps are very useful tools for understanding spatial relationships and patterns across the Earth's surface.

Which Places have the Highest Admission Rate?

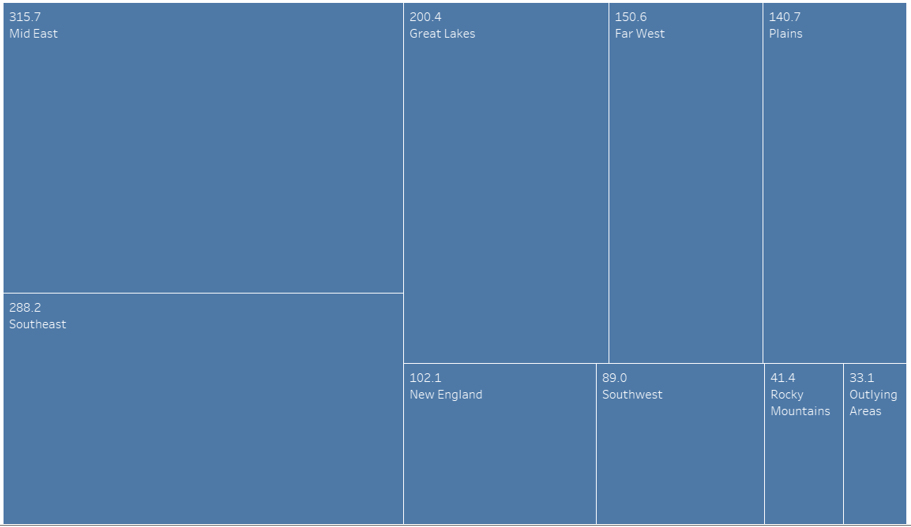


Large City places situated Universities have the Highest Admission Rates.

**Bar Graph:**

A bar graph is also called a bar chart. It is used to display categorical data. In a bar chart, rectangular bars of equal width are drawn along a horizontal or vertical axis. The larger the value in the dataset the more the length or height of each bar. Bar graphs are very effective in comparing discrete categories. Bar charts are either vertical or horizontal. In a vertical bar graph, the categories are shown along the x-axis (horizontally), while the numerical values are shown along the y-axis (vertically). In a horizontal bar graph, categories are drawn on the vertical axis, and the values are drawn along the horizontal axis.

Which Region has the Highest Admission Rate?

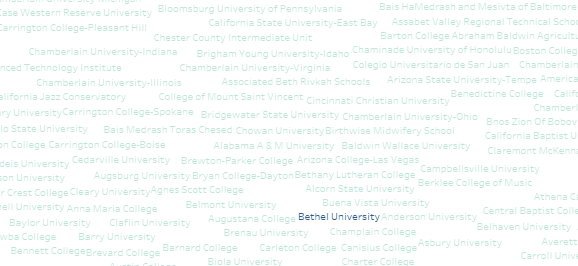


Mid East Region has the highest Admission Rate.

**Tree map:**

A tree map is a data visualization technique to display hierarchical data in a rectangular layout. In a tree map, the hierarchical structure of the data is represented by nested rectangles. The tree map is with each rectangle representing a specific data element or category. The more the category appears in the dataset, the larger the rectangle appears in the graph. The less number of times the category appears the smaller the rectangle size.

Which University has the Highest number of Branches?



Bethel University has the highest number of Branches.

**word map:**

A word map is also called a word cloud or tag cloud. It is a visual representation of text data where words are displayed in various sizes and colours according to their count number. In a word map, the more times a word appears in the text, the larger the word appears in the graph. The less number of times the word appears the smaller the font size. Word maps are mostly used to identify the key themes or keywords within a body of text.

Overall, word maps offer a concise and intuitive way to visualize textual information, allowing viewers to gain insights into the underlying patterns and trends present in the data.

**Conclusion:**

The examination of college admission rates unveils a complex interplay of factors that shape the educational landscape for countless aspiring students. Through this exploration, we've witnessed how admission rates reflect institutional selectivity. The data shows that college admission rates are not merely numerical representations but indicators of accessibility, equity, and opportunity within higher education.