# **COMPUTING FINAL PROJECT EXCECUTIVE REPORT**

# **Project Title: Data Analysis of different Universities in UK region.**

# **Submitted By**

1. Sai Lokesh Siddanathi
2. Divya Preksha Reddy Gali
3. Krishna Chaitanya MVS
4. Harshita Gangasani
5. Aditya Prakash Addanki

# 

A yellow and black logo

Description automatically generated with low confidence

DATE : 16TH DEC,2022

**University of Colorado Denver**

**(Aug 2022-May 2024)**

Table Of Contents

[Introduction: 1](#_Toc121855143)

[Attribute Information: 2](#_Toc121855144)

[Defining The Libraries: 3](#_Toc121855145)

[Read the Data and Import: 3](#_Toc121855146)

[Correlation Matrix and Usage: 4](#_Toc121855147)

[Correlation Matrix diagram: 4](#_Toc121855148)

[Findings – Result from Correlation Diagram: 5](#_Toc121855149)

[Plotting values for better insights: Scatter Plot – Public Vs Private 6](#_Toc121855150)

[Categorizing Colleges As Per Fees: 7](#_Toc121855151)

[Conclusions from Visuals: 8](#_Toc121855152)

[Scatter Plots: 9](#_Toc121855153)

[Line Graph: 10](#_Toc121855154)

[K Mean - Centroids: 10](#_Toc121855155)

[K- Means – Fitting Model: Test & Train 11](#_Toc121855156)

[K – Means: Identified Cluster & Array: 12](#_Toc121855157)

[Clustered Data 12](#_Toc121855158)

[Training and Testing Data – Split and Validation: 13](#_Toc121855159)

[Final Conclusion: 13](#_Toc121855160)

**Introduction:**

Data Analysis is one becoming one of the best approaches to figure out the best from the options available. The goal here of this research is to create some visuals using Python and find out the better result. The dataset was obtained from a known connect for working on list of colleges in UK. There are many factors that are required to rate the best college.

There are different visuals used here and K means cluster algorithm for finding the better result.

Here there are rankings and other components involved in the categorizing the universities. Here K means algorithm is used on this dataset to check the best college. This analysis helps in finding out the best universities.

**Attribute Information: Data Set Explained**

1. University: University Name

2. Location: University location

3. Year: Year founded

4. Tag Line: Main Motto or core value of the university

5. World Class Rank: World Ranking of the university

6. UK Rank: UK rank of the university

7. CWUR: Centre for world university rankings score

8. Min IELTS Score: Minimum IELTS requirement score

9. Avg. UG Tuition fees: UG Tuition fees in pounds

10. Avg. PG Tuition fees: PG Tuition fees in pounds

11. International student percent: Percentage of International students

studying in the university.

12. Enrolled students: Count enrolled students at university

13. Teaching staff: Count no. of staff members

14. Living Expenses: Cost of the living (Approx.) at university per year

(in pounds)

15. Campus Locality: Location of the campus (Rural/Urban/Suburban)

16. Latitude: Location Latitude

17. Longitude: Longitude Location

18. Website: University Website

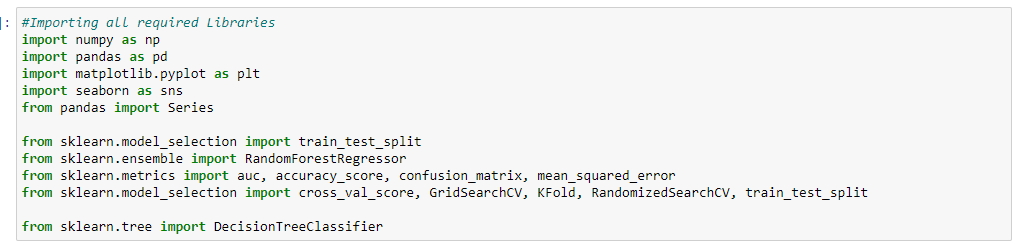
19. Type: Control type of university

20. Calendar: Academic calendar of the university.

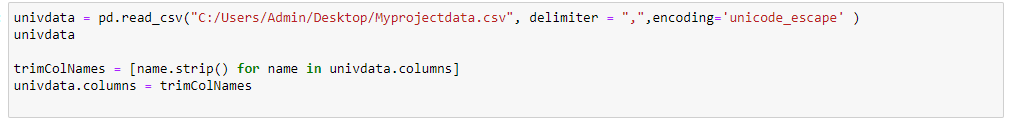
21. Student Satisfaction Percentage: Percentage of students who are satisfied

**Defining The Libraries:**

Below are the libraries that are used for this project. Numpy, Pandas for the Data Analysis and Matplotlib and Seaboarn for Visualisations. Skikit learn is also used here.

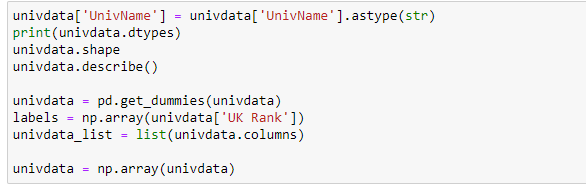


**Read the Data and Import:**

Here we read the data from the folder and define the delimiter as CSV and encode it as unicode. This Part also has trimmed columns, which is part of Data transformation and Data cleaning. Missing values and the uncleaned data often mis lead the results. So it is very important to clean the data, so the values can help models in predicting better results. In this step we have the trimmed the blank spaces .

**Data Exploration, Data Cleaning and Data Transformation:**

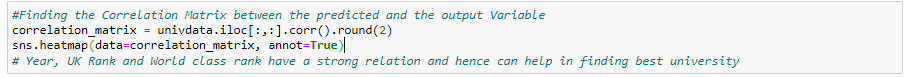
In continuation with previous step, we have shaped and tweaked the data further more. As part of Data Transformation, we have also defined the column datatype as it was captured delay.



Steps are taken to make sure that there are no missing values anymore after this step.

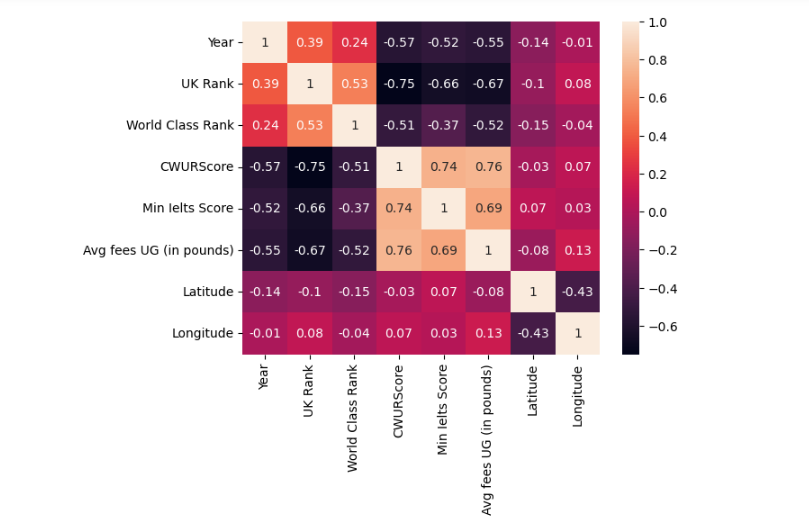
**Correlation Matrix and Usage:**

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. This is used for data summarization, input into a more complex analysis, and diagnostic for advanced analyses correlation matrix for the given dataset and below are the results.



**Correlation Matrix diagram:**

Below is the correlation matrix diagram that has been derived from the given dataset to identify the most Important factors that are contributing for deriving the better university.



**Findings – Result from Correlation Diagram:**

Year, UK Rank and World Class Rank are the most contributing factors from this diagram. Hence the UK Rank and World Class Rank areconsidered for our analysis.

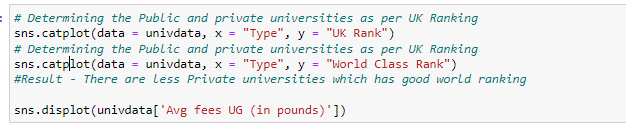
UK Rank – Local Rank in the UK for the university

World Class Rank – Global Rank for the university across world

Year – Year of establishment

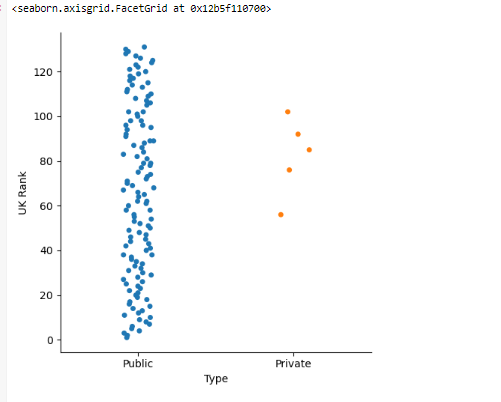
## Plotting values for better insights: Scatter Plot – Public Vs Private

Here Catplot has been used for drawing scatter plots across to define the universities which are public and Private. With this dataset containing 120 records, we were able to plot the colleges and below is the syntax that is used for defining this code.

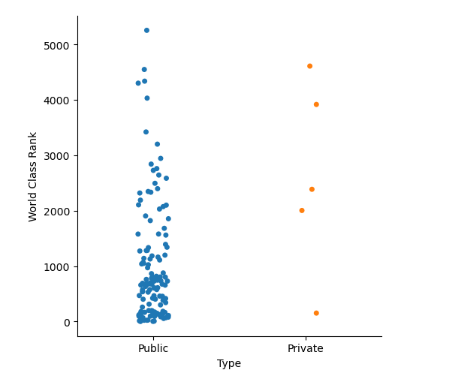


Code is written to depict the pictorial representation of the existing data for all the parameters. Below are the images that come as outputs after executing the above code.

**UK Rank with Public vs Private:**



**World Class Rank with Public Vs Private:**

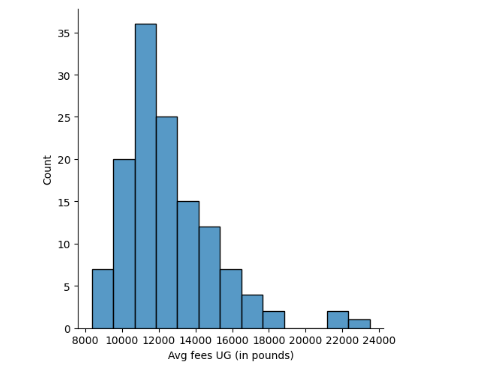


## Categorizing Colleges As Per Fees:

We need to categorize the college according to their fees structure to understand the college that is asking more fees and the other asking the low fees. Below is the code we used for the categorizing the lines, here displot is used for the quick visual. Histogram chart is developed for a better visualization.



**Count of Colleges vs Fees:**



## Conclusions from Visuals:

From these 3 visuals the below conclusions are drawn based on the considered dataset.

1. **UK Rankings** - Public Universities has the better ranking compared to Private and it is advised to go for a public university

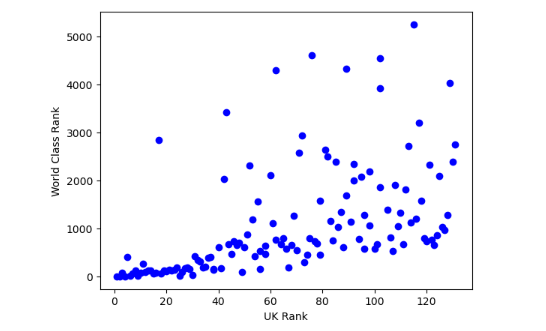
2. **World Class Rankings** – Public universities has better exposure and it is advised from the dataset to opt for the public universities that has better world class ranking.

3**. College Fees** - College Fees is represented pictorially to see the fees split. Most of the universities fall under the 10k – 14k Category. Out of 120 colleges considered for this analysis, 80% of colleges are falling under this category. So the fee for most of the colleges is expected to be ranging from 10K – 14K Pounds per year.

## Scatter Plots:

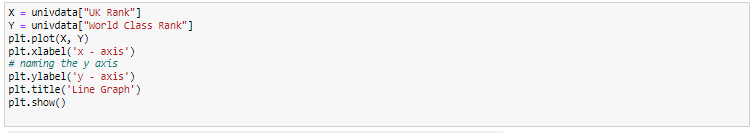
As Part of analysis now we focus on the scatter plots to identify the relation amongst and later we shall use these results for centroids which shall help us in finding centroid.

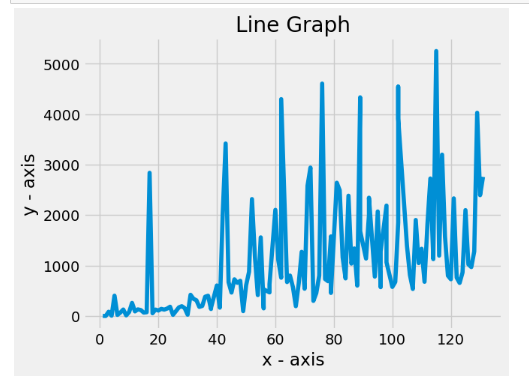
From Correlation matrix, it is understood that there are three factors that are contributing. Hence a chart is plotted between UK Rank and World Class Rank.



## Line Graph:

Here we tried to form a line graph to analyze the rankings to see if these visuals make much more insights.



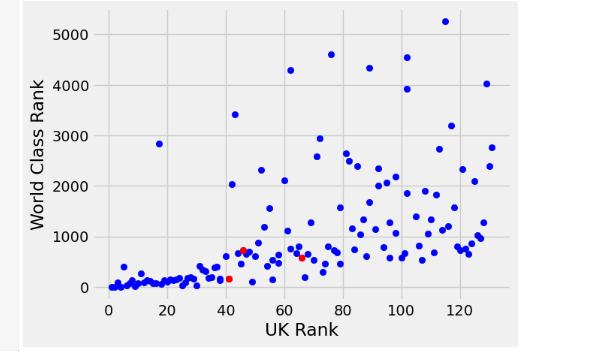


## K Mean - Model:

Now that the scattered plot is done. We need to find out the centroid of the scatters. We have considered K =3 and marked these are red colours.

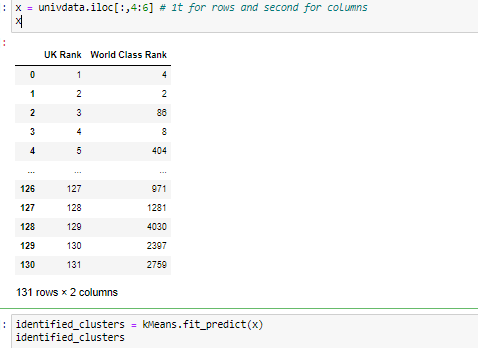


Once the above lines are built and executed then it shall plot the graphs to



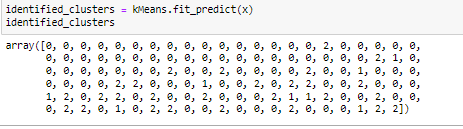
## K- Means – Fitting Model: Test & Train

Here we have already defined the columns that we have already targeted. So, the columns are selected from the dataset and now the K model is applied for the fitting and validated further.



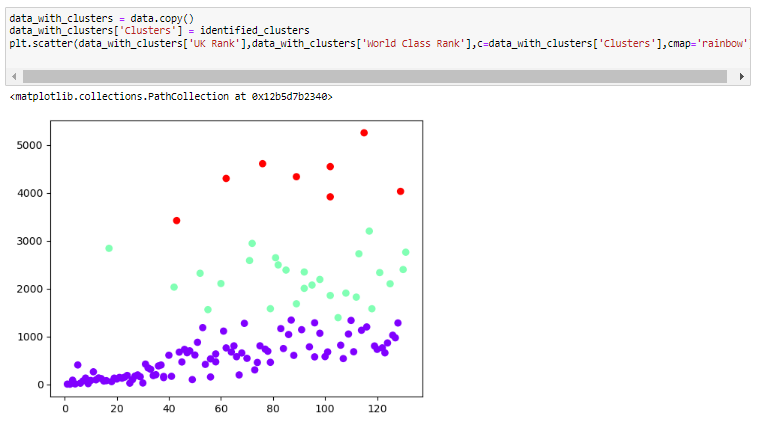
## K – Means: Identified Cluster & Array:

When it is tried to apply the fitment to the model, there are some predictions that are made by the model and below is the output for the same.



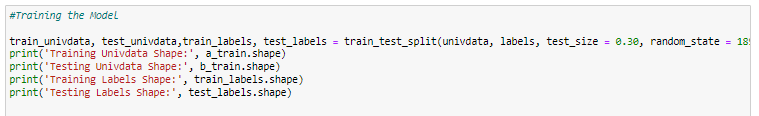
## Clustered Data

Below is the clustered data from the predictions.

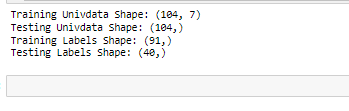


## Training and Testing Data – Split and Validation:

Here the dataset is split into Training and Testing Dataset. Random seed is defined as 1899 and test data is 30% and training data is 70%.

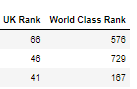


**Results of Test and Train Model:**



## Final Conclusion:

We have approached the final part of the analysis with this K means methods. We have found that these 3 are the best colleges that are predicted from the given dataset.



**Result Dataset:**

****