IBM Machine Learning Professional Certificate

- Unsupervised Learning - (Clustering)

Choe C.S

October 2021

Main Objective

Objective

- This report aims to perform **Clustering** on data set from Universities into **2 cluster** groups, **Private** and **Public**.

Data Set

- The data set used for this analysis was obtained from Kaggle.com.

Steps Involved

1) Perform Data preprocessing to – Data Formatting to convert data types to correct format.
2) EDA – To gain better understanding of each feature variables and its affects on the target data.
3) Perform Feature Engineering remove unwanted columns, perform Feature Scaling and Log Transformation for normalization.
4) Perform Machine Learning on the data sets using Clustering algorithm
   - (K-Means Cluster and Agglomerative Clustering).

.

Import Data

- The data set was import from Kaggle.com.
- The table at the bottom shows the top 5 rows for the data set

	Unnamed: 0	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rati
0	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18
1	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12
2	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12
3	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7
4	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11

Data Preprocessing

- The figure on the right shows the overall summary info of data set.
- There are a total of 777 rows with 19 column of data .
- There is a total of 1 “float64”, 1 “int64” and 2 “object” columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 19 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   777 non-null    object
 1   Private      777 non-null    object
 2   Apps         777 non-null    int64
 3   Accept       777 non-null    int64
 4   Enroll       777 non-null    int64
 5   Top10perc    777 non-null    int64
 6   Top25perc    777 non-null    int64
 7   F.Undergrad  777 non-null    int64
 8   P.Undergrad  777 non-null    int64
 9   Outstate     777 non-null    int64
 10  Room.Board   777 non-null    int64
 11  Books        777 non-null    int64
 12  Personal     777 non-null    int64
 13  PhD          777 non-null    int64
 14  Terminal     777 non-null    int64
 15  S.F.Ratio    777 non-null    float64
 16  perc.alumni  777 non-null    int64
 17  Expend       777 non-null    int64
 18  Grad.Rate    777 non-null    int64
dtypes: float64(1), int64(16), object(2)
```

Data Preprocessing – Data Formatting (Column Name)

- As the original column names are difficult to interpret, Data Formatting for renaming of columns was performed.
- This was done to avoid confusion of the data each column represents.

	Unnamed: 0	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Rati
0	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18
1	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12
2	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12
3	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7
4	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11

Before Renaming

Column Renaming

	College	Private	Apps	Accept	Enroll	Top 10%	Top 25%	Full-time-Undergrad	Part-time-Undergrad	Outstate	Boarding	Books	Personal	PhD	Terminal	S.F.Ratio	Alumni %	Expend
0	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041
1	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527
2	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735
3	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016
4	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922

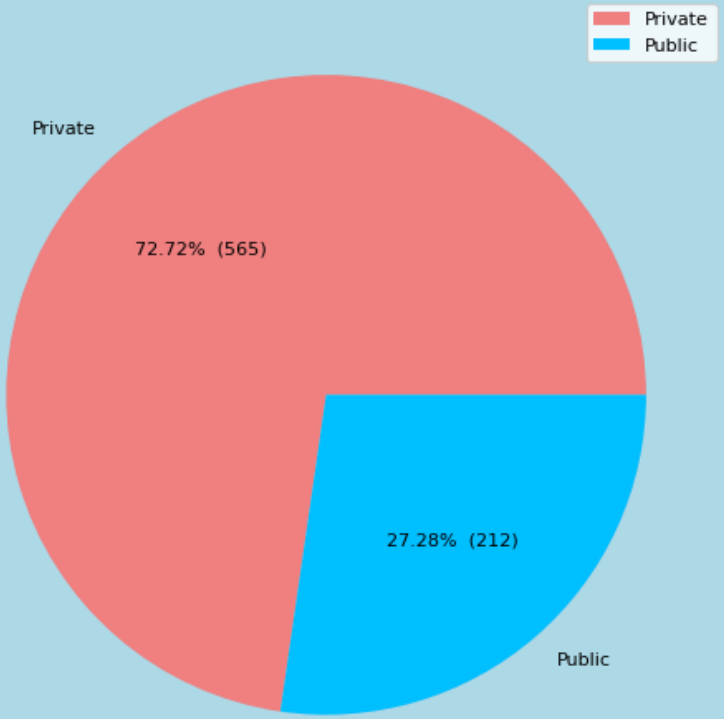
After Renaming

EDA - Descriptive analysis

	Apps	Accept	Enroll	Top 10%	Top 25%	Full-time-Undergrad	Part-time-Undergrad	Outstate	Boarding	Books	Personal	
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384	549.380952	1340.642214	72
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416	165.105360	677.071454	16
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	96.000000	250.000000	8
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000	850.000000	62
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000	1200.000000	75
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000	1700.000000	85
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000	6800.000000	103

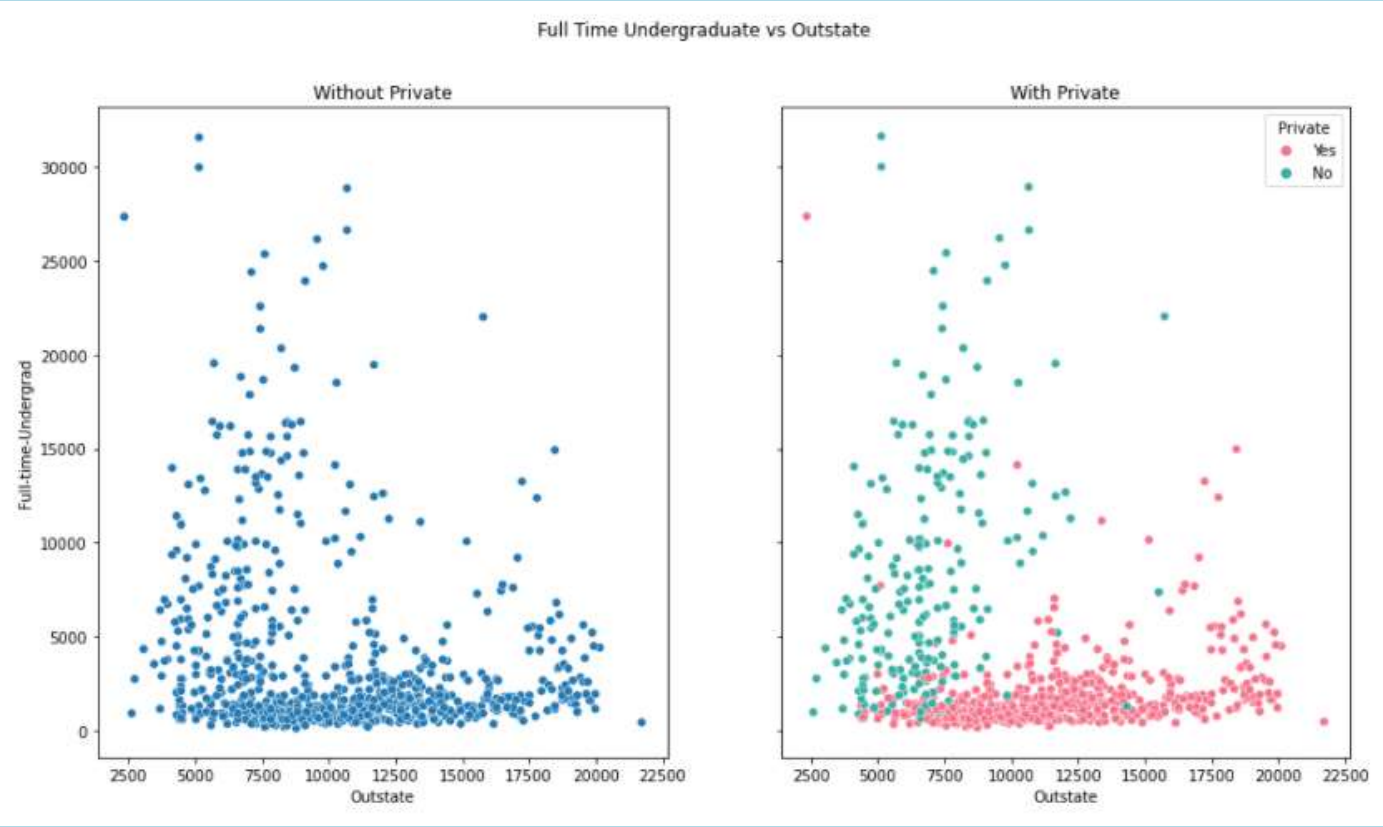
- The table on the top shows the descriptive analysis summary for data set.
- The mean value for Apps is 3001.6, Accept is 2018.8, Enroll is 779397, Top 10% is 27.58 , Top 25% is 55.8, Full-time-Undergrad is 3699.9, Part-time-Undergrad is 855.3 Outstate is 10440.67, Boarding is 4357.53, Books is 549.38, Personal is 1340.64 etc.
- The minimum value for Apps is 81, Accept is 72, Enroll is 35, Top 10% is 1, Top 25% is 9, Full-time-Undergrad is 139, Part-time-Undergrad is 1, Outstate is 2340, Boarding is 1780, Books is 96, Personal is 250 etc.
- The maximum value for Apps is 48094, Accept is 26330, Enroll is 6392, Top 10% is 96, Top 25% is 100, Full-time-Undergrad is 31643, Part-time-Undergrad is 21836, Outstate is 21700, Boarding is 8124, Books is 2340, Personal is 6800 etc.

EDA – Visual exploration



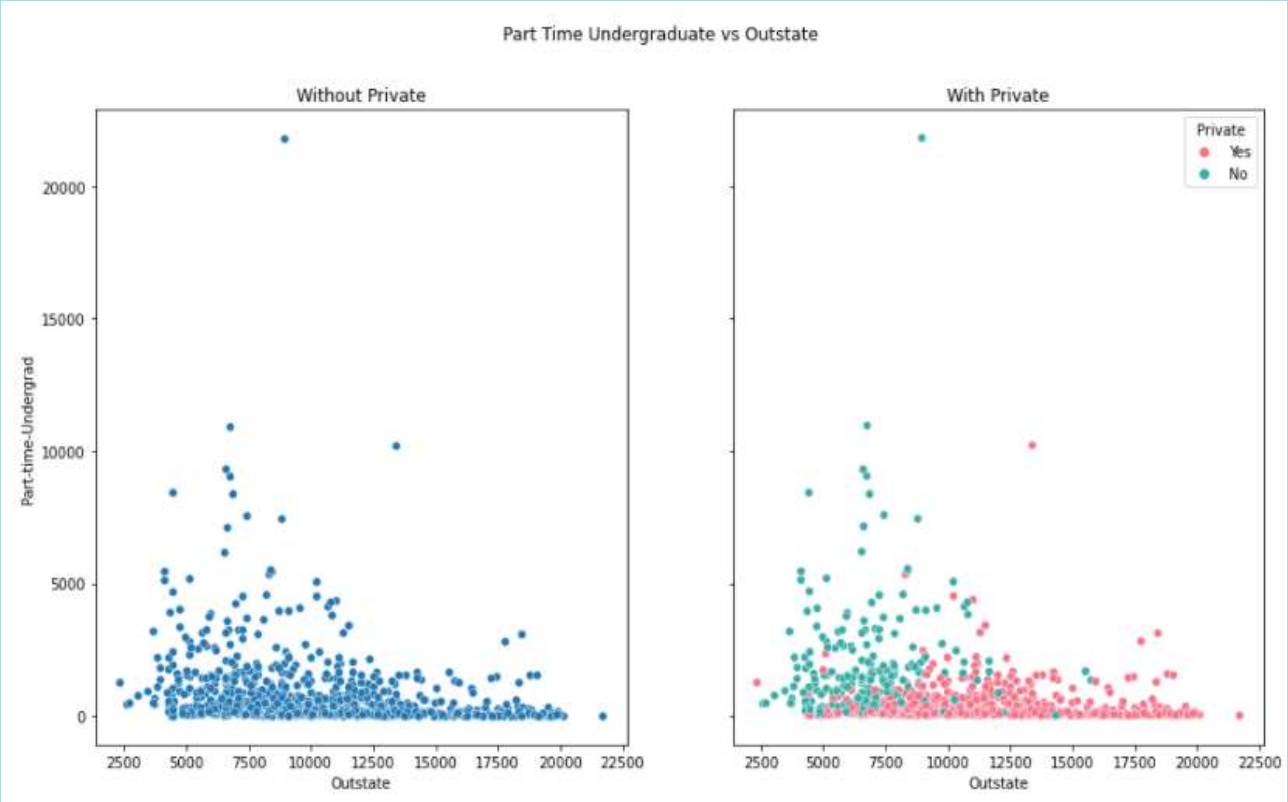
- As we are interested in analyzing cluster groups of Private and Public Universities, it is important to know the amount of Private and Public universities involved.
- A pie plot was plotted for easier and better view of the amount of Private and Public universities in data set.
- There are a total of 565 Private Universities and 212 Public Universities involved.
- Private universities is at 72.72% while Public universitities at 27.28%

EDA – Visual exploration (scatter-plot with hue)



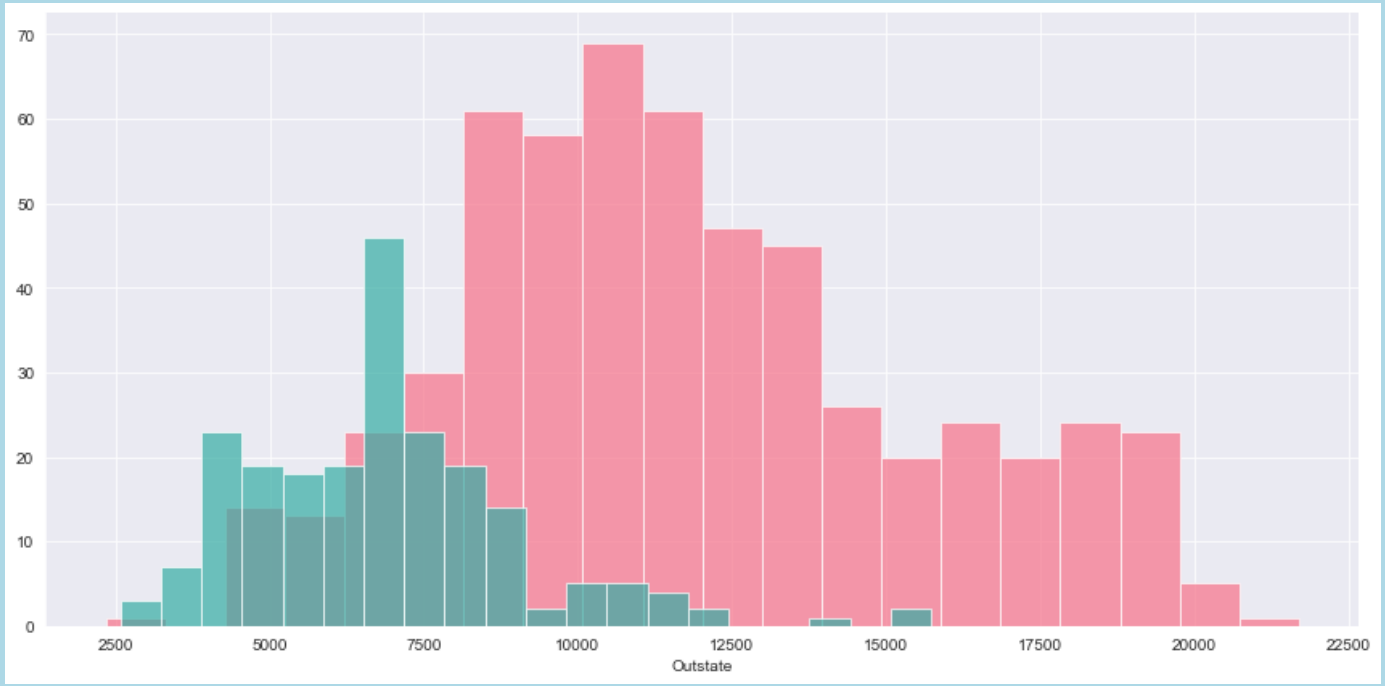
- The scatter plot on the right shows relationship between **Full-time-Undergrads** vs **Out of State** column**.**
- The first plot shows a rather general overview with higher **Full-time-Undergrads** at **Out of State** value between of 2500 to 10000.
- At the bottom of the graph, shows high concentration of data point between 0 to 5000 for **Full-time-Undergrads.**
- The second plot includes hue of **Private** category and the results shows, the lower the **Out of State** value the higher the **Full-time-Undegrads.**

EDA – Visual exploration (scatter-plot with hue)



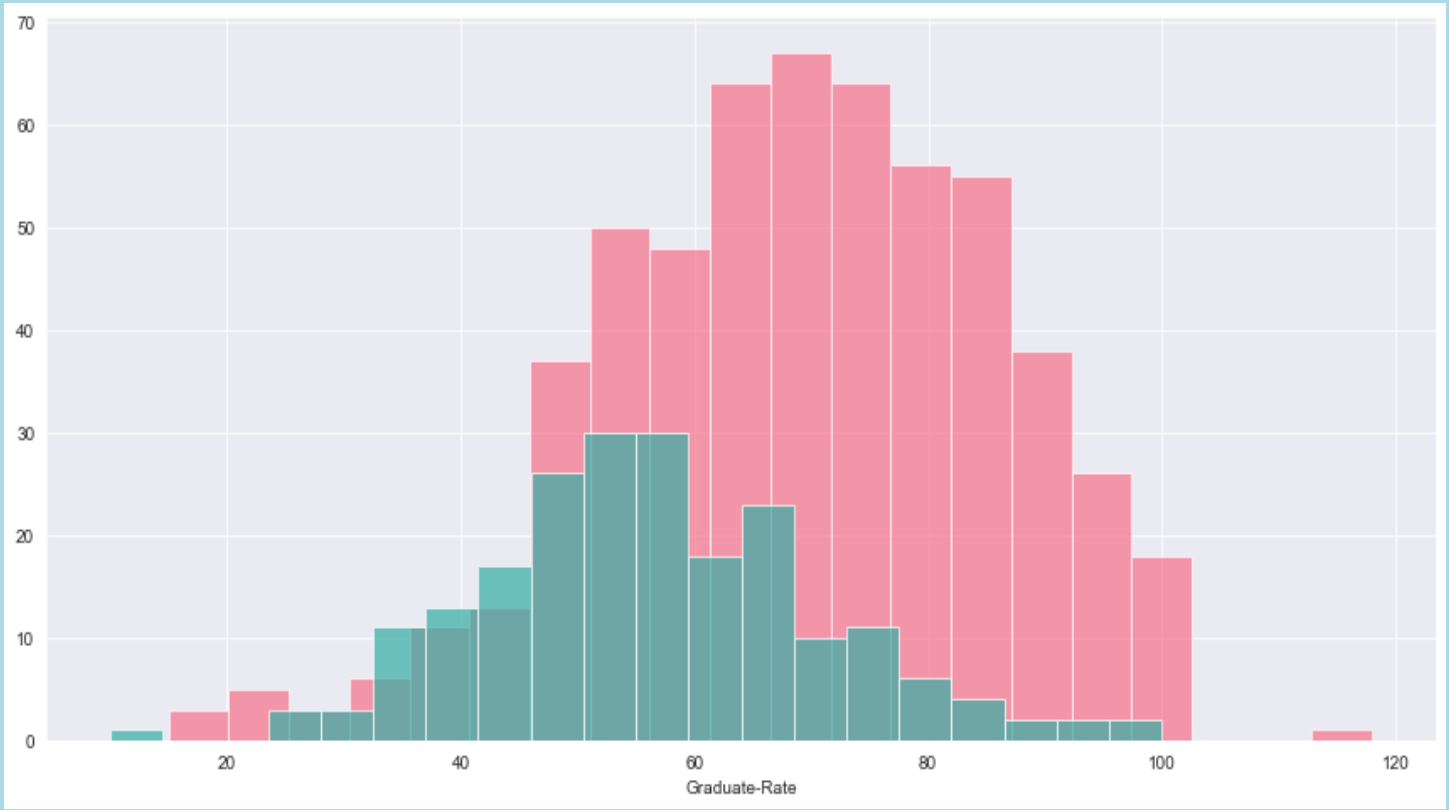
- The scatter plot on the right shows relationship between **Part-time-Undergrads** vs **Out of State** column**.**
- The first plot shows a rather general overview with data points mostly scattered at he the bottom half of the graph.
- The second plot includes hue of **Private** category and the results shows, higher **Part-time-Undergrads** values at **Out of State** value between 2500 to 10000 for **Private** undergrads**.**
- The comparison between **Full-time-Undergrad** is a lot higher than **Part-time-Undergrads** against **Out of State** values recorded.

EDA – Visual exploration (Histogram)



- The Histogram on the right shows the **Out of State** vs **Private** column.
- From the histogram, more students from Out od State enrolled in **Private** universities compared to public.
- The number of **Private** Universities records higher values between **Out of State** value at 7500 to 12500.
- While **Public** universities recorded higher **Out of State** value at 6500 to 7500.

EDA – Visual exploration (Histogram)



- The Histogram on the right shows the **Out of State** vs **Graduate Rate** column.
- From the histogram, **Private** Institute gain higher **Graduate Rate** compared to **Public** Institutes.
- The number of **Private** Universities records higher for **Graduate Rate** at **60- 80%.**
- The number of **Private** Universities records higher for **Graduate Rate** at **40-80%.**

Feature Engineering – Drop column

Drop unrequired column

University Column was removed from data set before machine learning process takes place.

Before

	College	Private	Apps	Accept	Enroll	Top 10%	Top 25%	Full-time-Undergrad	Part-time-Undergrad	Outstate	Boarding	Books	Personal	PhD	Terminal	S.F.Ratio	Alumni %	Expend
0	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041
1	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527
2	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735
3	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016
4	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922

After

	Private	Apps	Accept	Enroll	Top 10%	Top 25%	Full-time-Undergrad	Part-time-Undergrad	Outstate	Boarding	Books	Personal	PhD	Terminal	S.F.Ratio	Alumni %	Expend	Graduate Rat
0	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	6
1	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	5
2	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	5
3	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	5
4	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	1

Feature Engineering – Log Transformation

Log Transformation

- Log transformation was applied to columns with skew value more than 0.75.
- This was done reduce outliers due to overly right skewed distribution of certain data.

Part-time-Undergrad	5.692353
Apps	3.723750
Books	3.485025
Expend	3.459322
Accept	3.417727
Enroll	2.690465
Full-time-Undergrad	2.610458
Personal	1.742497
Top 10%	1.413217
S.F.Ratio	0.667435
Alumni %	0.606891
Outstate	0.509278
Boarding	0.477356
Top 25%	0.259340
Graduate-Rate	-0.113777
PhD	-0.768170
Terminal	-0.816542
dtype: float64	

Before Transformation

Log Transform →

Expend	0.845072
S.F.Ratio	0.667435
Alumni %	0.606891
Full-time-Undergrad	0.517054
Outstate	0.509278
Boarding	0.477356
Enroll	0.373329
Top 25%	0.259340
Apps	0.188485
Accept	0.179966
Personal	-0.105722
Graduate-Rate	-0.113777
Part-time-Undergrad	-0.362271
Books	-0.366866
Top 10%	-0.433738
PhD	-0.768170
Terminal	-0.816542
dtype: float64	

After Transformation

Feature Engineering – Feature Scaling

Standard Scaler

- Feature Scaling was perform on the numeric feature variables to normalize the ranges of the data.
- Standard Scaler from Scikit Learn preprocessing was used for this process.
- The results obtained is shown in the table below.

	Private	Apps	Accept	Enroll	Top 10%	Top 25%	Full-time-Undergrad	Part-time-Undergrad	Outstate	Boarding	Books	Personal	PhD	Terminal	S.F.Rat
0	Yes	-0.011583	0.006320	0.427055	0.014037	-0.191827	0.323264	0.361674	-0.746356	-0.964905	-0.601556	1.259401	-0.163028	-0.115729	1.0137
1	Yes	0.245031	0.457067	0.067493	-0.535636	-1.353911	0.252545	0.875509	0.457496	1.909208	1.286817	0.469963	-2.675646	-3.378176	-0.4777
2	Yes	-0.151913	-0.111012	-0.374601	-0.053802	-0.292878	-0.674259	-0.685997	0.201305	-0.554317	-1.036697	-0.050896	-1.204845	-0.931341	-0.3007
3	Yes	-1.298516	-1.267855	-1.313967	1.500947	1.677612	-1.363987	-0.963863	0.626633	0.996791	-0.601556	-0.640670	1.185206	1.175657	-1.6152
4	Yes	-2.014530	-2.145621	-2.262878	-0.535636	-0.596031	-2.060724	0.660925	-0.716508	-0.216723	1.525504	0.469963	0.204672	-0.523535	-0.5535

Machine Learning - (Clustering)

Unsupervised Learning (Clustering)

- A total of 4 models were build to analyze the clusters of Private and Public universities.
- 2 models from each **K-means** and **Agglomerative Clustering** algorithm were used for this analysis.
- The results obtained were compared and discussed in the next section.

K-Means Cluster

- There were a total of 2 Models created using **K-Means Cluster**.
- The 1st Model is **K-Means Cluster** with **initialization method** = '**random**'.
- The 2nd Model is **K-Means Cluster** with **initialization method** = '**k-means++**'.

Agglomerative Clustering

- There were a total of 2 Models created using **Agglomerative Clustering**.
- The 1st Model is **Agglomerative Clustering** with **linkage method** = '**complete**'.
- The 2nd Model is **Agglomerative Clustering** with **linkage method** = '**ward**

Results and Discussion

K-Means

initialization = 'random'

kmeans	Private	Total
0	No	102
	Yes	321
1	No	110
	Yes	244

initialization = 'k-means ++'

kmeans	Private	Total
0	No	112
	Yes	243
1	No	100
	Yes	322

Agglomerative Clustering

linkage = 'complete'

aggc_complete	Private	Total
0	No	166
	Yes	304
1	No	46
	Yes	261

linkage = 'ward'

aggc_ward	Private	Total
0	No	202
	Yes	252
1	No	10
	Yes	313

Discussion

- The results from **K-Means** cluster **random** initialization identifies 102 Public and 321 Private for Public records and 110 Public and 244 Private for Private records.
- The results from **K-Means** cluster **K-means ++** initialization identifies 112 Public and 243 Private for Public records and 100 Public and 322 Private for Private records.
- The results from **Agglomerative Clustering** cluster linkage **complete** at 116 Public and 304 Private for Public records and 46 Public and 261 Private for Private records.
- The results from **Agglomerative Clustering** cluster linkage **ward** at 202 Public and 252 Private for Public records and 10 Public and 313 Private for Private records.

Results and Discussion

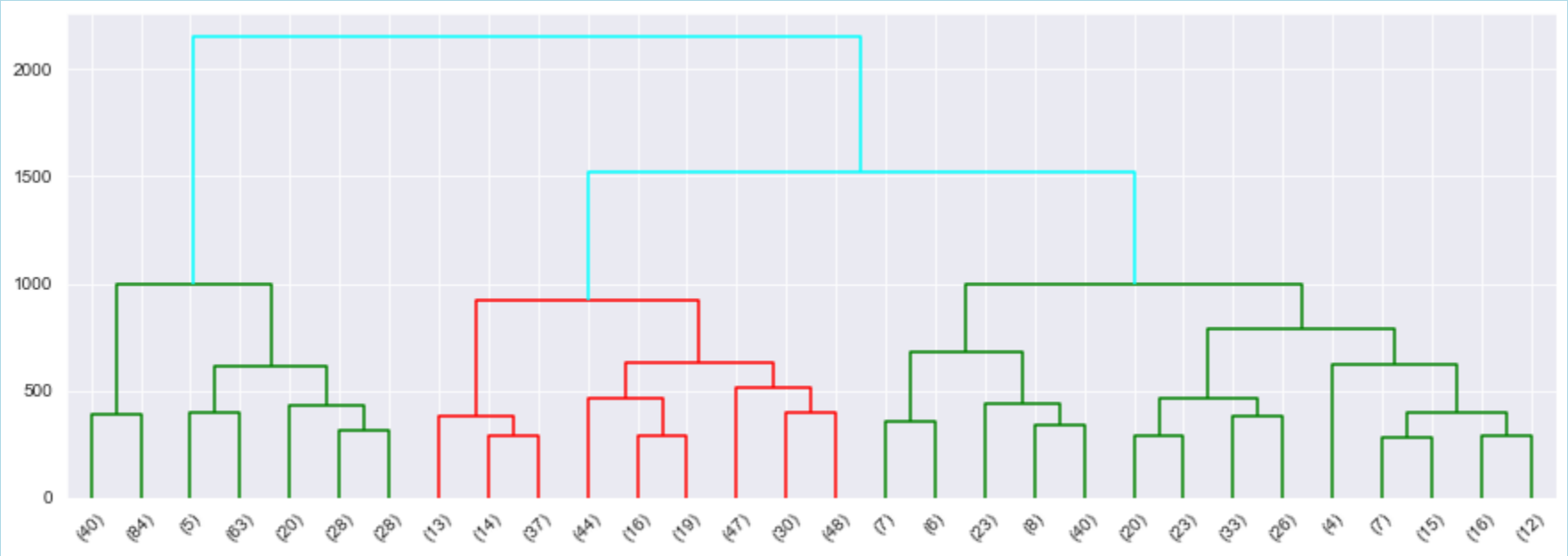
Table Summary for all Models

- All 4 model result were join together into a single table for summary view of the results

Private	aggc_complete	aggc_ward	kmeans	Total
No	0	0	0	104
			1	54
		1	0	8
	1	0	1	44
		1	1	2
Yes	0	0	0	14
			1	22
		1	0	222
			1	46
	1	0	0	2
			1	214
		1	0	5
			1	40

Results – Hierarchy plot

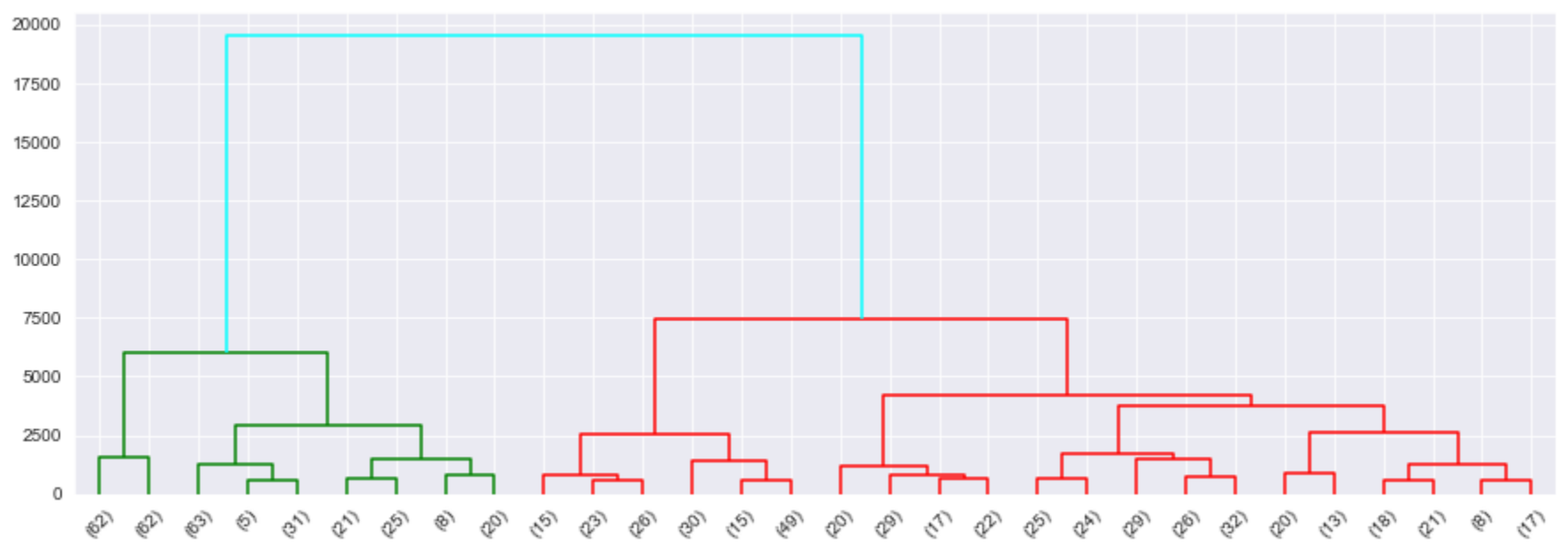
(Agglomerative Clustering – “Complete”)



The graph above shows the result of **Agglomerative Clustering** with **linkage Complete.**

Results – Hierarchy plot

(Agglomerative Clustering - ward)



The graph above shows the result of **Agglomerative Clustering** with **linkage Ward.**

Conclusion & Improvements

Conclusion

- From the results obtained better prediction was achieved for Private compared to Public universities due to more weightage of data provided (72.72% Private - 1 & 27.28% Public – 0.).

- The most suitable model selection is **Agglomerative Clustering** with linkage **Ward.**

Improvement

- Future methods to improve better prediction results can be achieve through **Dimensionality Reduction** for better feature selection and to reduce curse of dimensionality.

- One of the examples of Dimensionality Reduction that can be use is **PCA**.

- Other hyperparameters for K-Means and Agglomerative Clustering could also be chosen for analysis.

Conclusion & Improvements

Link to Code

https://github.com/cs-robot-collab/IBM-ML-DL/blob/master/IBM%20Machine%20Learning%20Clustering%20Report.ipynb