# Assignment – Unsupervised ML

1. What is Clustering. How clustering helps in performing EDA? – 3 marks
2. What is the difference between K Means & Agglomerative Clustering. Explain it using some data in Jupyter Notebook through Coding. – 3 marks
3. Explain the Difference between the Silhoutte Score and Linkage Method – “Single”. – 3 marks
4. Consider the University Dataset with the 777 Rows and 18 Columns. Here, you need to identify the optimum cluster size by using K Means. – 7 marks

Here is the Data Dictionary

#### Private A factor with levels No and Yes indicating private or public university

#### Apps Number of applications received

#### Accept Number of applications accepted

#### Enroll Number of new students enrolled

#### Top10 perc Pct. new students from top 10% of H.S. class

#### Top25 perc Pct. new students from top 25% of H.S. class

#### F. Undergrad Number of fulltime undergraduates

#### P. Undergrad Number of Part Time undergraduates

#### Outstate Out-of-state tuition

#### Room. Board Room and board costs

#### Books Estimated book costs

#### Personal Estimated personal spending

#### PhD Pct. of faculty with Ph.D.’s

#### Terminal Pct. of faculty with terminal degree

#### S.F. Ratio Student/faculty ratio

#### Perc. Alumni Pct. alumni who donate

#### Expend Instructional expenditure per student

#### Grad. Rate Graduation rate[¶](http://localhost:8888/notebooks/Identifying%20the%20%23%20of%20Clusters%20-%20University%20Dataset.ipynb#Grad.Rate-Graduation-rate)

1. Consider the Titanic Train Set. Generate Clustering Model on the Training Dataset and find the following: - 7 marks

* Find out which clustering approach is better – K Means or Agglomerative Clustering.
* Compare the Labels with the Survived Column Values. The better accuracy score suggests which clustering explains the labels better.
* Also, find out if there is any need of cluster values In Agglomerative Clustering. Write a Note on that.

1. Consider a Kaggle Competition called Digit Recognizer under the following link:

<https://www.kaggle.com/c/digit-recognizer>

This is a Supervised ML Problem. You need to apply PCA, find the optimal features count and apply ML. (Algorithm Selection depends on you).

Note - We have already worked on this dataset while applying clustering in classroom. Its time to apply PCA and find out what we get when we submit the result on Kaggle.

* 7 marks