****

**Prepared By:** Sandeep Singh

**Subject:** Big Data Fundamentals

**Title:** 2018W-T1 BDM 1043 - Big Data Fundamentals 01 (DSMM Group 1)

**Sub Title:** Project 01w18

**Week:** 26 March – 01 April

**Submitted To:** Mr. Raed Karim

**Submission Date:** 02 April, 2018

Contents

[Project 4](#_Toc510401414)

[Introduction 4](#_Toc510401415)

[Problem Overview 4](#_Toc510401416)

[Solution 4](#_Toc510401417)

[Insights 4](#_Toc510401418)

[Screenshots 6](#_Toc510401419)

[Attachments 6](#_Toc510401420)

[Conclusion 6](#_Toc510401421)

[References 6](#_Toc510401422)

# Project

## Introduction

**Machine learning** is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. In this project, we are focusing on **K Means clustering** which is a type of unsupervised learning, and is used when we have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*.

## Problem Overview

In this project, we have a [sample dataset of delivery fleet driver data](https://raw.githubusercontent.com/datascienceinc/learn-data-science/master/Introduction-to-K-means-Clustering/Data/data_1024.csv). We have been provided with the distance feature and the speeding feature of various drivers. Distance vs. speed graph must be plotted for various drivers as the initial part on the dataset. Further, we need to compute the clustering on the dataset using the selected K (where K is increased by a step of 2). Also, we need to determine the best K for the plot of squared error function for every number of K and the error value.

Formula to be used:

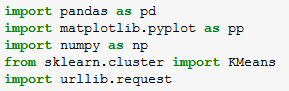
https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/images/image009.gif

## Solution

In order to resolve the problem described above, we’ll be using python3 as programming language. We may use any IDE that supports python3 (with the in use libraries installed), however Jupyter notebook was used during the coding of this project. The following sections provide a detailed view of the code.

### Insights

#### Python 3 libraries used:

// to perform dataframe operations

// to plot the various graphs used in this project

// to perform mathematical operations

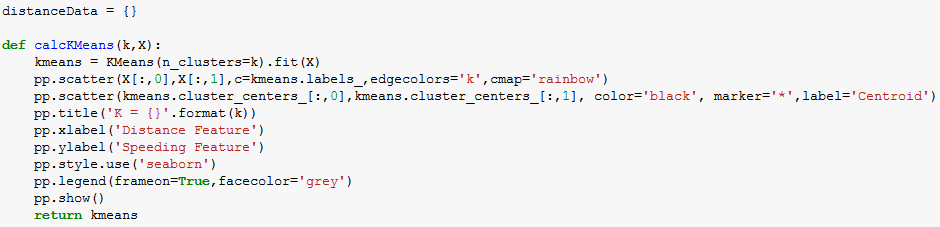
//to find K Means from given datasets

// to download the csv file

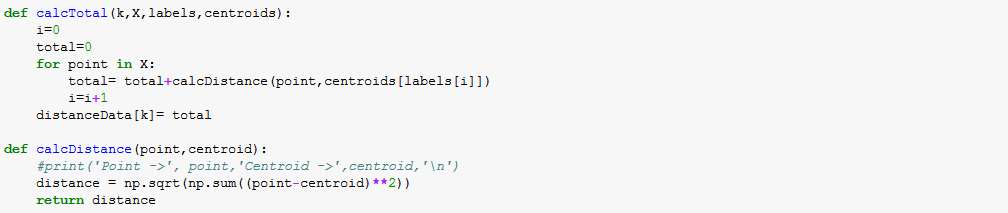
The program begins with the *Main()* being invoked first. This function downloads the information about distance and speed of the drivers as ‘*driversData.csv*’ file. A dataframe *df* is created to read the csv file. An array *X* is created with *distance\_feature* on x axis and *speeding\_feature* on y axis. Now, K means is calculated by changing the value of *k* starting from 2 till 50 (with a step interval of 2) and the graphs are plotted.



The function *calcKMeans(k,X)* is invoked from main() and computes the k means taking K and X as parameters. Desired graphs are plotted for each value of k taking care of the formatting part.



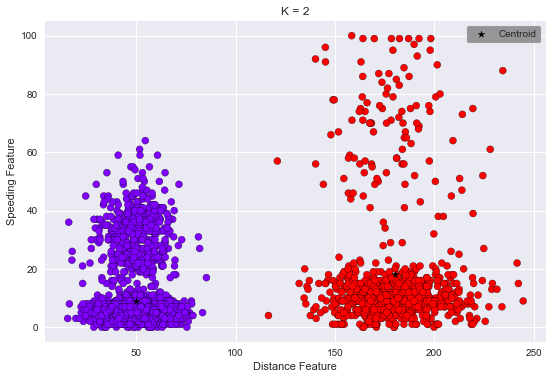
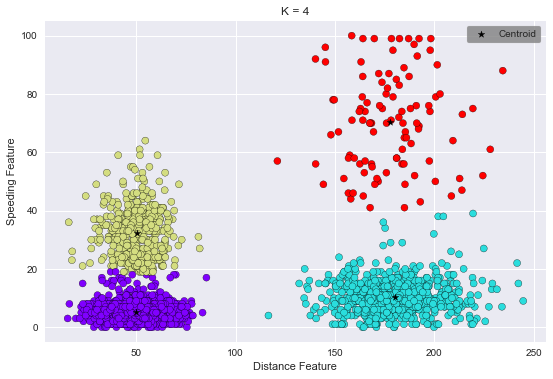
Next, we find out the error function from the available data. For this, we use the *sqrt* and *sum* function from numpy library. The distance from point to the centroid is returned using *calcDistance(). calcTotal(*) accumulates the values of all the Ks’ and respective distance from points and centroids in a dictionary called *distanceData*.

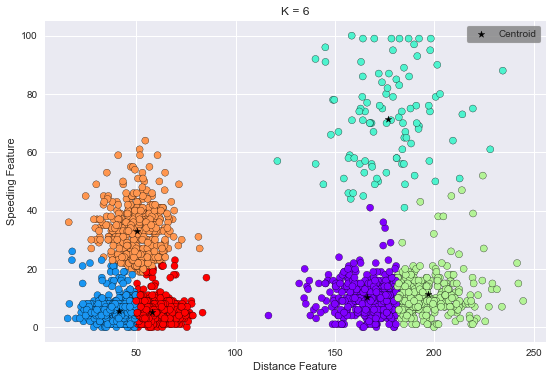
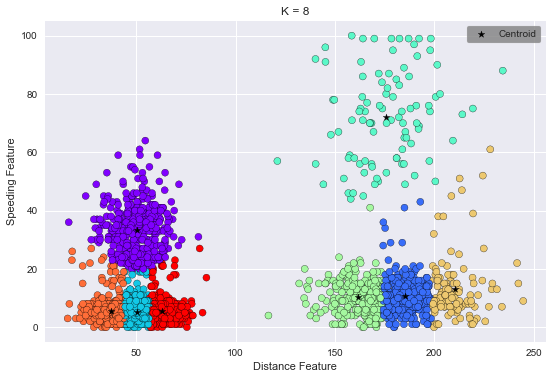


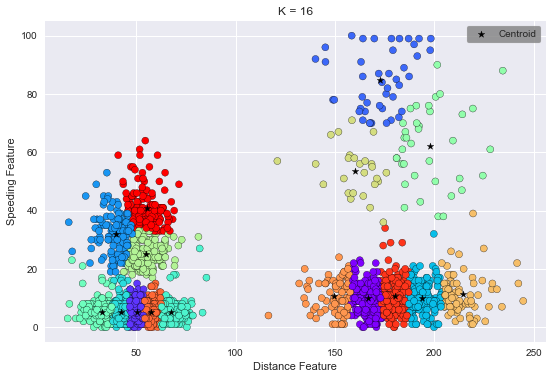
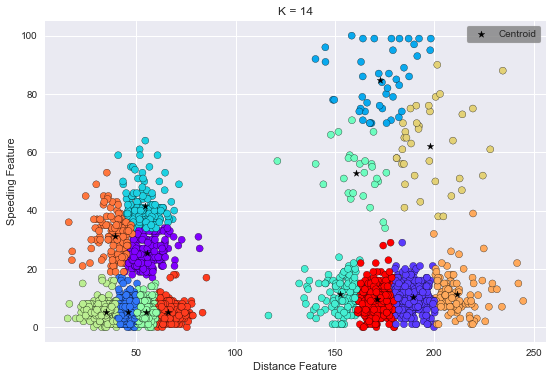
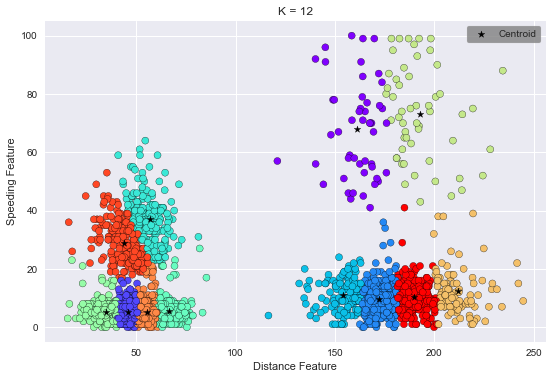
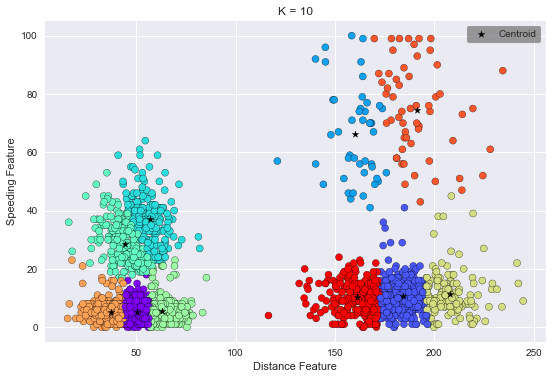
Finally, the Distance vs. Value of K graph is plotted to find the best value of the K.

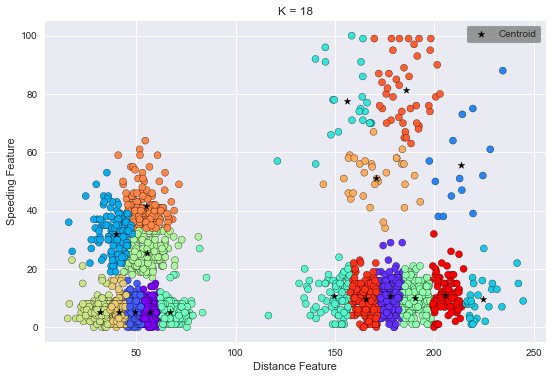
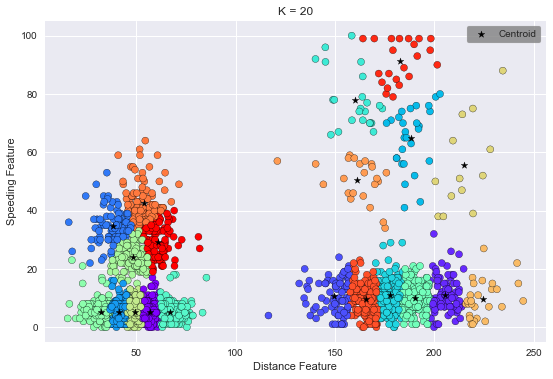
### Screenshots

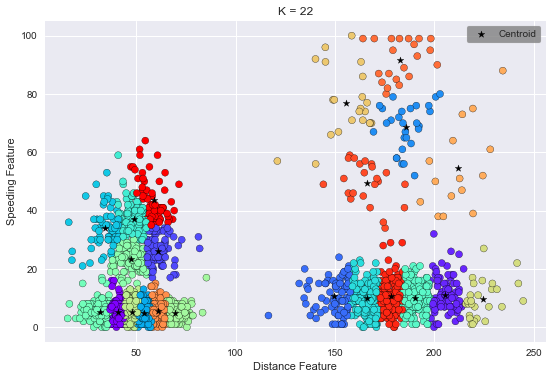
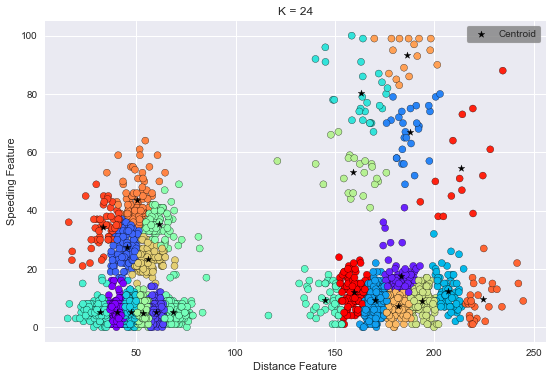
This section covers the screenshots of the graphs that illustrate the clustering in k means project.

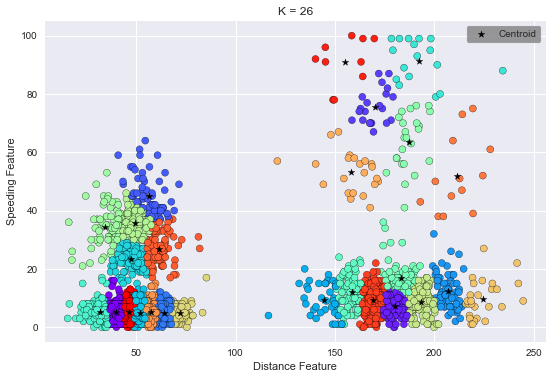
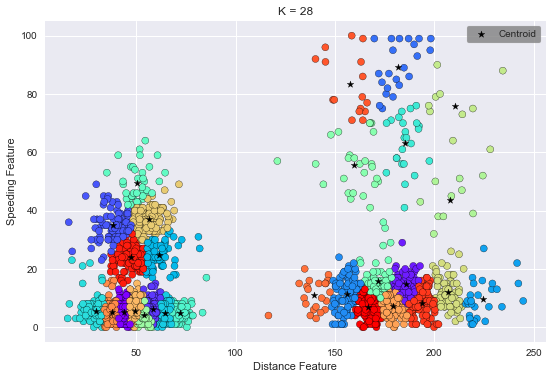
 

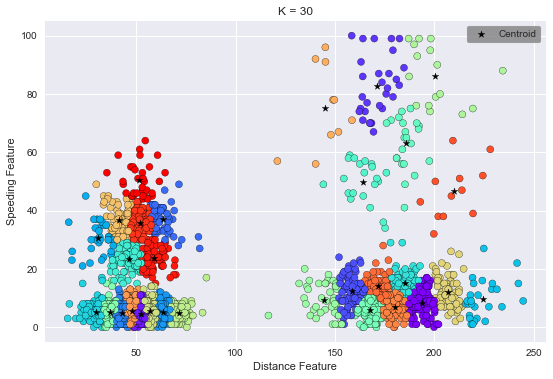
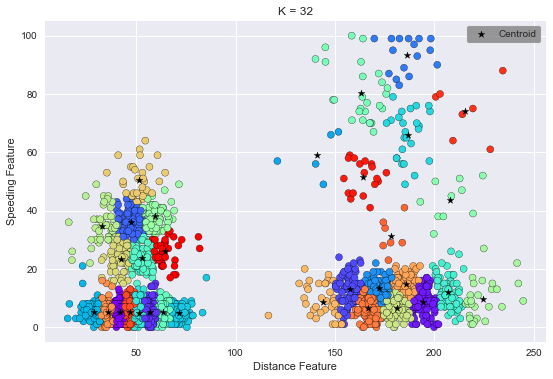
 



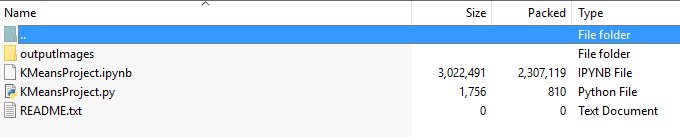
 

### 

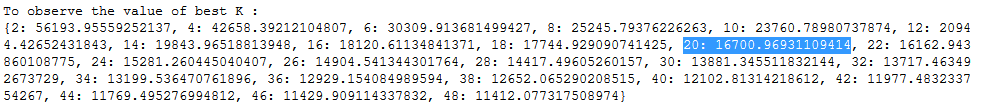
### Attachments

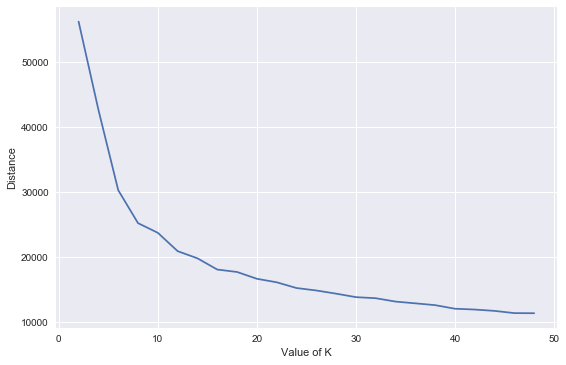
Project.zip has been attached which contains program code in .py extension as well as .ipynb extension (to be used with jupyter notebook)



## Conclusion

Values of distances for all Ks’ are collected in a dictionary. As observed, after *k = 20*, the bent in the curve is not significant. Also, from the dictionary data, we are able to find out that value of overall distance does not change prominently.





In order to conclude, we are able to judge that k=20 is the elbow point and happens to be the best value of k.

## References

* <https://www.datascience.com/blog/k-means-clustering>
* <https://raw.githubusercontent.com/datascienceinc/learn-data-science/master/Introduction-to-K-means-Clustering/Data/data_1024.csv>
* <https://www.youtube.com/watch?v=4R8nWDh-wA0>
* <http://scikit-learn.org/stable/modules/clustering.html#overview-of-clustering-methods>
* <http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html#sphx-glr-auto-examples-cluster-plot-cluster-iris-py>
* <http://www.numpy.org/>
* <https://matplotlib.org/index.html>
* <https://pandas.pydata.org/>

\*\*\*\*\*