Clustering of Cars dataset using Principal Component Analysis and k-means

Arvind Murugan | MUR0071

Sweety Wadhwa | WAD0002

*VSB – Technical University of Ostrava*

# Aim

The main objective is to cluster the cars dataset using Principal Component Analysis (PCA) and K-means algorithm.

# Introduction

## Principal Component Analysis:

Principal component analysis (PCA) is a widely used statistical technique for unsupervised dimension reduction.

It is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. It reduces attribute space from a larger number of variables to a smaller number of factors and as such is a "non-dependent" procedure (that is, it does not assume a dependent variable is specified).

The goal is dimension reduction and there is no guarantee that the dimensions are interpretable (a fact often not appreciated by (amateur) statisticians). To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component.

## K – MEANS CLUSTERING:

The K-means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields.

If k is given, the K-means algorithm can be executed in the following steps:

* Partition of objects into k non-empty subsets
* Identifying the cluster centroids (mean point) of the current partition.
* Assigning each point to a specific cluster
* Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
* After re-allotting the points, find the centroid of the new cluster formed.

# Dataset

Cars Data has Information about 3 brands/make of cars. Namely US, Japan, Europe. Target of the data set to find the brand of a car using the parameters such as horsepower, Cubic inches, Make year, etc.

### Rows: 261

### Columns: 8

# Implementation

## Applying K-means without PCA:

A picture containing screenshot

Description automatically generated

## A screenshot of a cell phone Description automatically generated

## For pca = 3:

A screenshot of a cell phone

Description automatically generated

#### K-means for PCA = 3:

A screenshot of a cell phone

Description automatically generated

## For pca = 4:

A screenshot of a social media post

Description automatically generated

#### K-means for PCA = 4:

A screenshot of a social media post

Description automatically generated

## For pca = 7:

A screenshot of a cell phone

Description automatically generated

#### K-means for PCA = 7:

A screenshot of a cell phone

Description automatically generated

# Conclusion

After calculating the PCA, we see that the explained variance ratio for the first two components contain the important information for clustering.

So, we conclude that with the help of PCA we’ve reduced the dimension of the dataset and from that itself we where able to cluster the whole document.