

Principle Component Analysis And K- means Cluster

Unsupervised Learning project

By Joshua kabwanga




Table of content

- Objectives
- Dataset
- Requires Libraries
- PCA Libraries
- K-Means
- Visual data
- Conclusion
- Recommendation



Objectives

In this project I would demonstrate the PCA algorithm techniques, which is going to help me to reduce the number of the feature to the most relevant. than after I will be able to perform K-means cluster algorithm to segment customers expenditure pattern in the dataset selected





Problem solve

The company would like to know:

- Reduce the number of features in the dataset.
- Then apply K-mean clustering Algorithm.
- Understand customer purchase segment.
- The customer expenditure habit.
- You've been tasked with segmenting customers into groups depending on their purchasing habits.

Python libraries

- `import numpy as np`
- `import pandas as pd`
- `import matplotlib.pyplot as plt`
- `import seaborn as sns`
- `from sklearn.preprocessing import LabelEncoder`
- `from sklearn.preprocessing import StandardScaler`
- `from sklearn.decomposition import PCA, KernelPCA`
- `from sklearn.cluster import AgglomerativeClustering`
- `from yellowbrick.cluster import KElbowVisualizer`
- `from sklearn.preprocessing import LabelEncoder`
- `from sklearn.preprocessing import StandardScaler`






Dataset

Problem Description

A Chinese automobile company **Teclov_chinese** aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts. They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the customer purchase habit of cars in the American market, since those may be very different from the Chinese market. Essentially, the company wants to know:

- The company want to separate the customers segmenting into groups depending on their purchasing habits.

The dataset consist of the following features: car_ID, symboling, CarName, fueltype, aspiration, doornumber, carbody, drivewheel, enginelocation, wheelbase, carlength, carwidth, carheight, curbweight, enginetype, cylindernumber, enginesize, fuelsystem, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg, price.





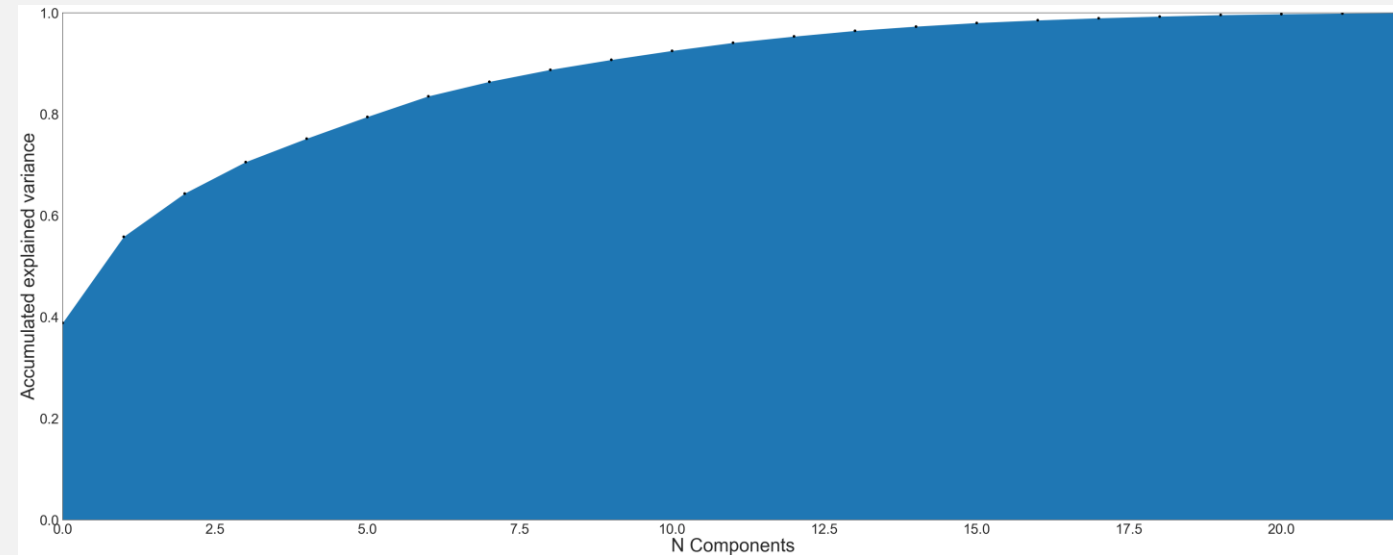
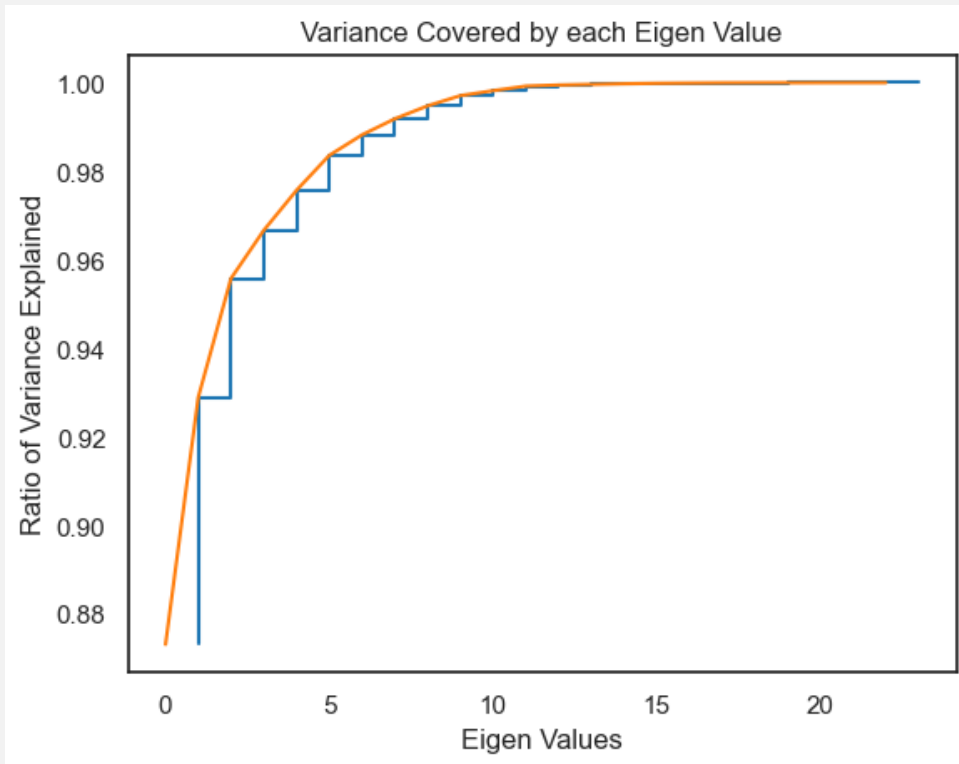
PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique that can be used to reduce a larger set of feature variables into a smaller set that still contains most of the variance in the larger set.

Use cases of PCA

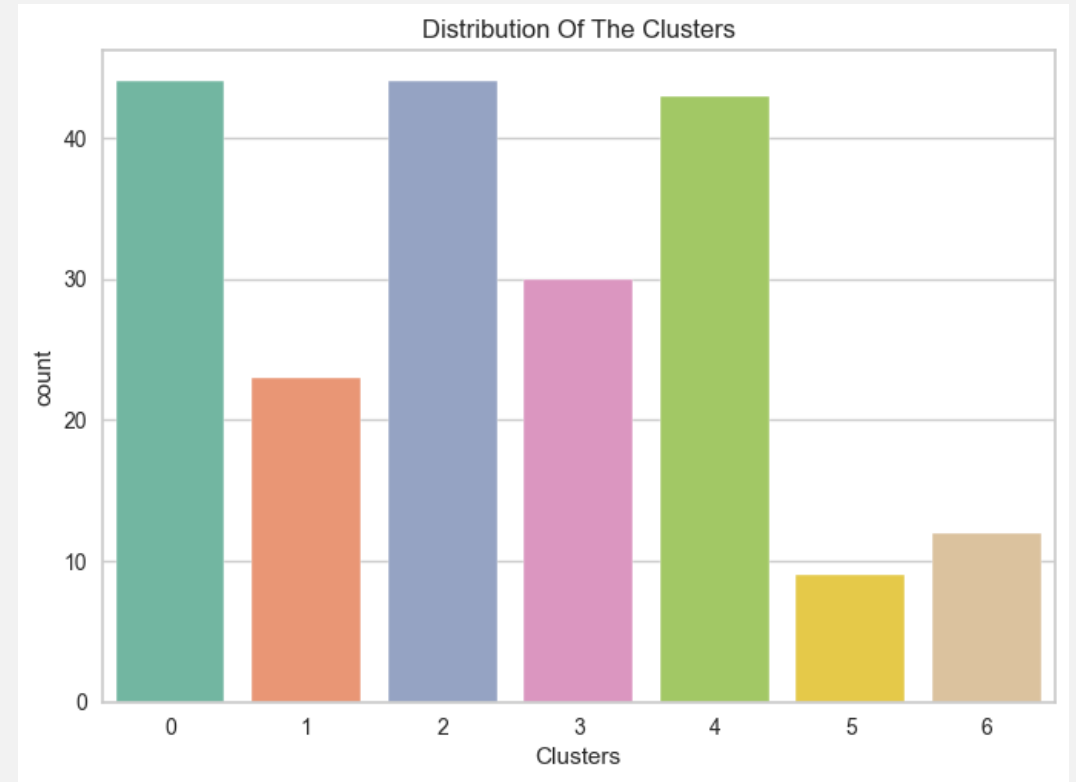
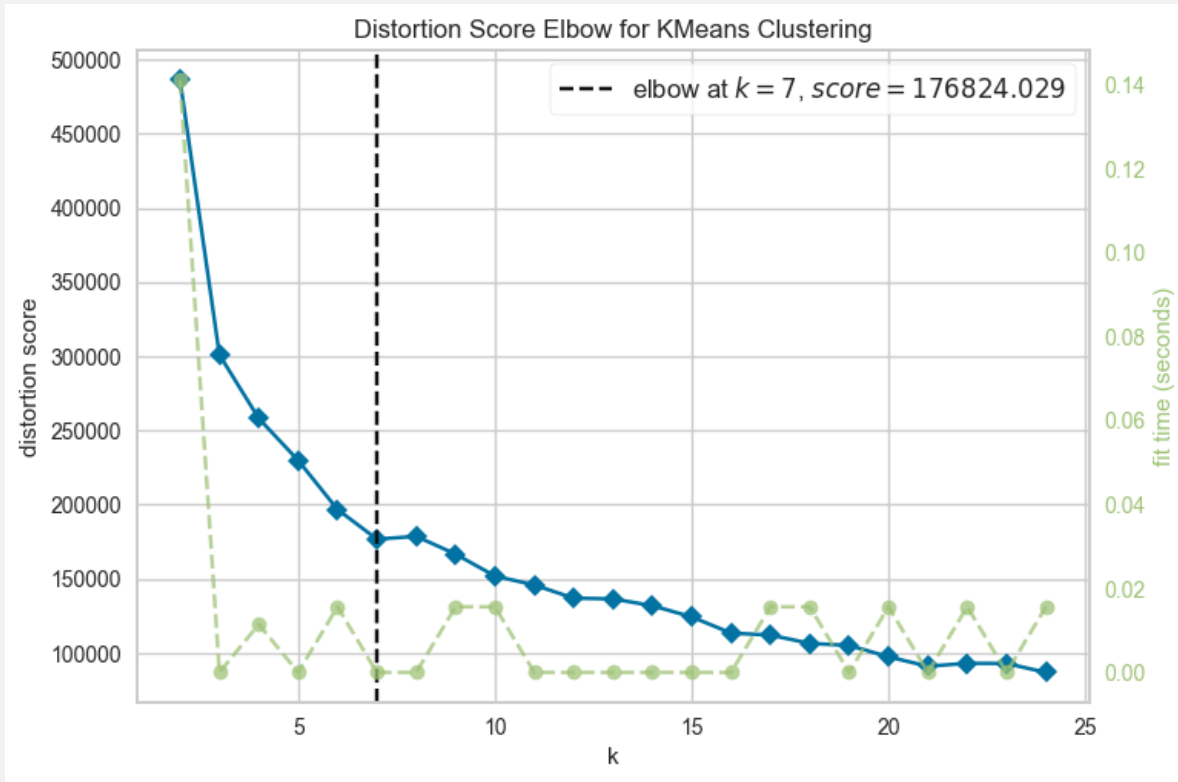
- Facial Recognition
 - Image Compression
 - Finding patterns in data of high dimension in the field of quantitative finance.
- 

Suppose a 99% threshold is sufficient for our task, let's see how many components (dimensions) we can drop:



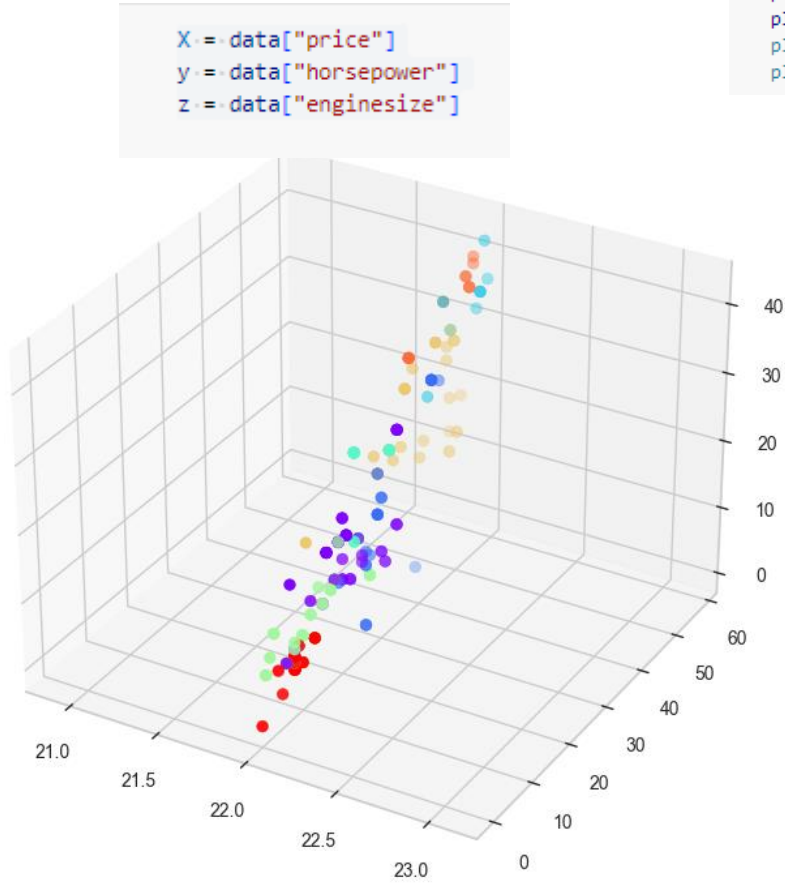
We can keep the first 19 components and discard the other 4, keeping $\geq 99.0\%$ of the explained variance!

Kmeans Clustering

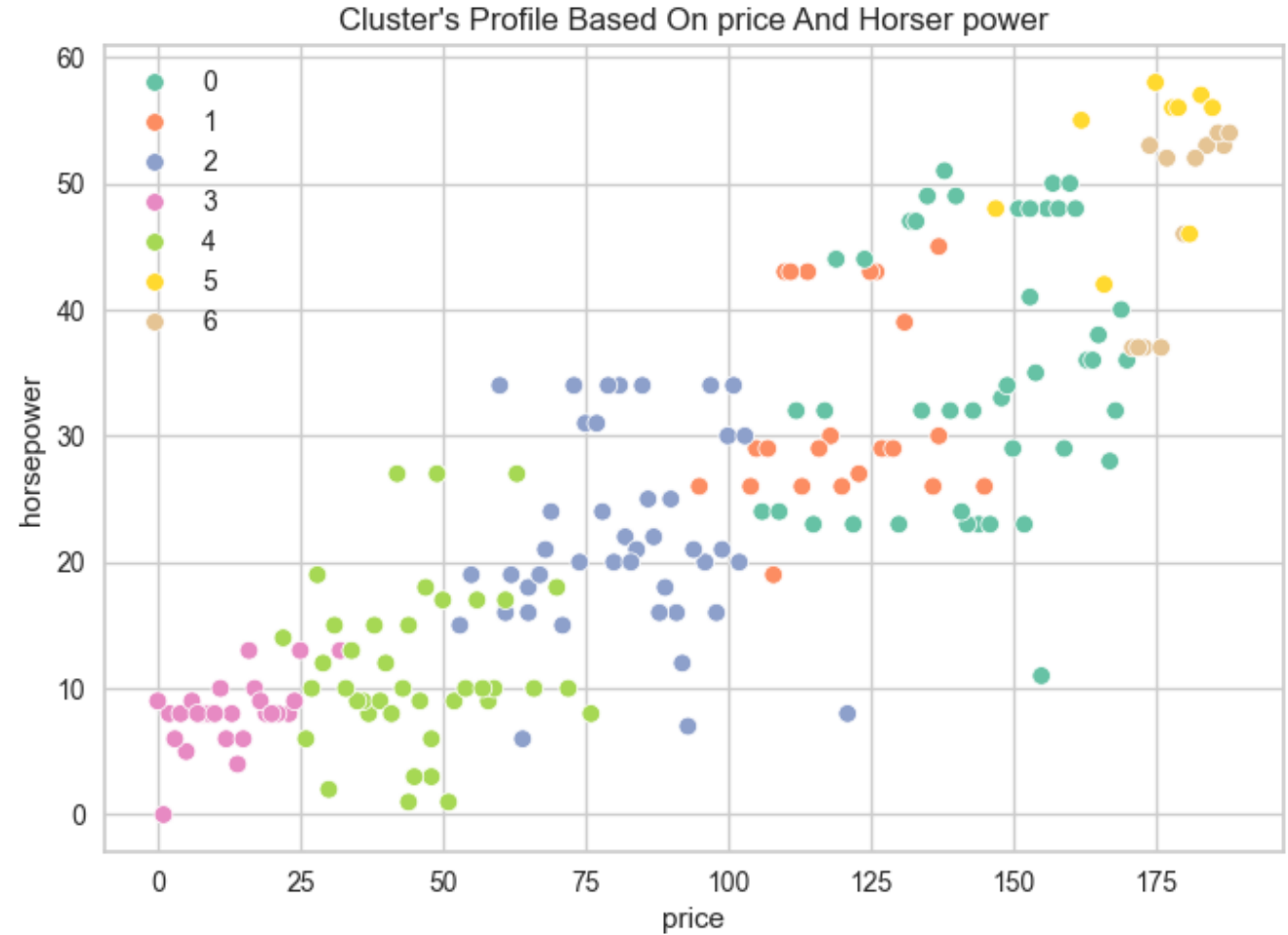


The above cell indicates that 7 will be an optimal number of clusters for this data.
Next, we will be fitting the Agglomerative Clustering Model to get the final clusters.

Plot the Clusters



```
pl = sns.scatterplot(data = data,x=data["price"], y=data["horsepower"],hue=data["Clusters"], palette= pal)  
pl.set_title("Cluster's Profile Based On price And Horser power")  
plt.legend()  
plt.show()
```



- The principal components analysis (PCA) has done very successful work accordingly and reduced the dataset features
- The K-means clustering algorithm has managed to detect 7 clusters, with 0 and 2 clusters being the most followed by th

Conclusion

In this project we apply PCA and the K-means clustering algorithm.

- The principal components analysis (PCA) has done very successful work accordingly and reduced the dataset features from 23 to 19 features with the 99% threshold set.
- The K-means clustering algorithm has managed to detect 7 clusters, with 0 and 2 clusters being the most followed by the 4, but the least of all is the 5 cluster.



Recommendation

Based on the founding of the Principal Components Analysis (PCA), I think we'll need set even more parameters to reduce the number of features from 19 even to a lower number for a better outcome.

For K-means clustering to perform even better.