Feynn Labs: Project 2

Online Vehicle Booking Market

Contributors: Saniket, Vaidehi, Rahul, Gajendrasingh

Problem Statement

Task is to analyse the Online Vehicle Booking Market in India using Segmentation analysis and come up with a feasible strategy to enter the market, targeting the segments most likely to use their product in terms of Geographic, Demographic, Psychographic, Behavioural. In this report we analyse the Online Vehicle Booking Market in India using segments such as location, price, Trip start time, Trip end time, waiting time, pickup long time, temperature, whether, trip type and much more.

Data Collection

Some important columns explained below

1. ‘trip\_start\_time’ is time when trip is start
2. ‘trip\_end\_time’ is time when trip is end
3. ‘trip\_time’ is time taken to complete trip
4. ‘wait\_time’ tells how much time is waited to get trip
5. ‘total\_time’ is time required to complete trip with waiting time
6. ‘price\_rub’ is price for trip
7. ‘distance\_kms’ is trip distance
8. ‘precipitation’ is about at trip rain is came or not
9. ‘feels\_like’ is about customer feedback on trip
10. ‘temperature\_value’ is temperature at time of trip

Data Preprocessing

Steps to clean raw data

1. Drop duplicate values
2. Adjusting missing values
3. Drop unnecessary columns
4. Categorizing the data
5. price\_rub', 'cloudness’ change the format to float
6. ‘trip\_start\_date’, ‘trip\_start\_time’, ’trip\_end\_date’, ’trip\_end\_time’, ‘trip\_time’, ‘total\_time’,

‘wait\_time’ changes the format to datetime

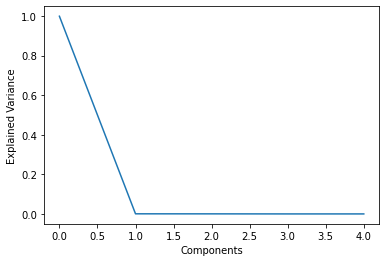
Principal Component Analysis

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

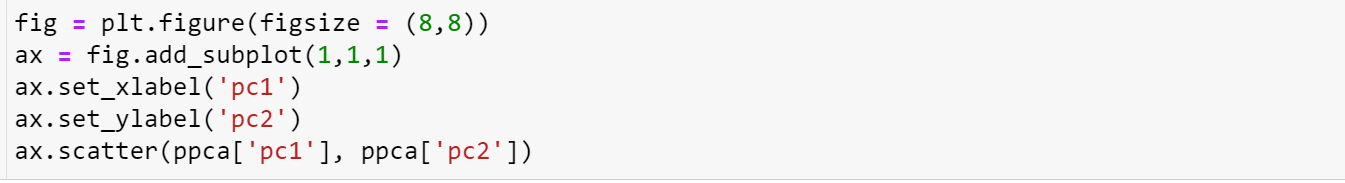
Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analysing data much easier and faster for machine learning algorithms without extraneous variables to process.

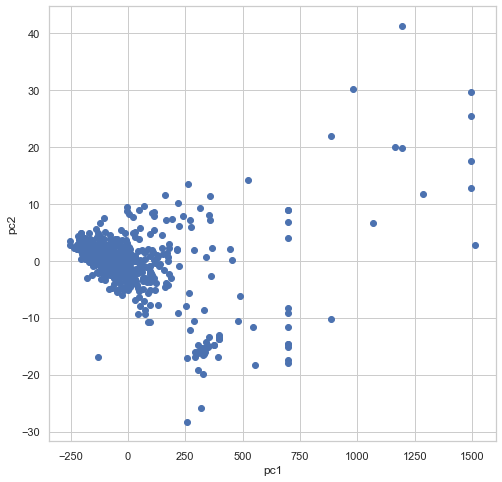
So, to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.





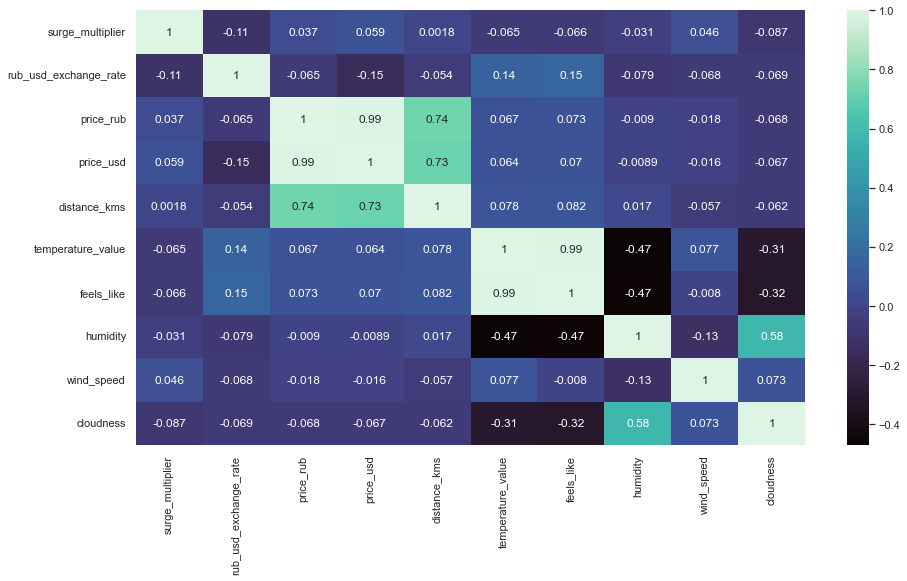
From above plot we can select important principle components. A vital part of using PCA in practice is the ability to estimate how many components are needed to describe the data. This can be determined by looking at the cumulative explained variance ratio as a function of the number of components





Above Scatter plot of principal components is shows the correlation between two principal components. There is no correlation between two principal components. It means that two principal components are independent.



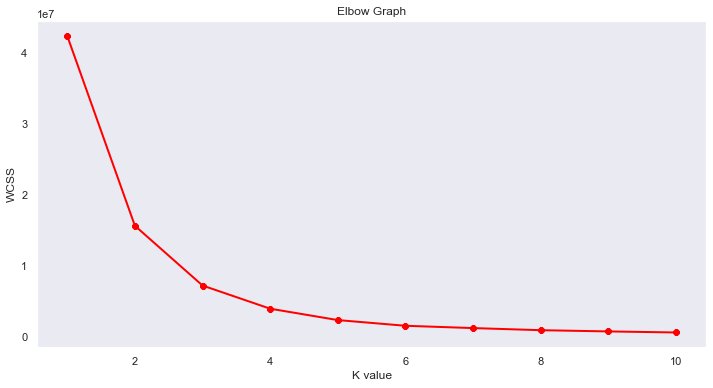


Above heatmap shows the correlation between variables. In heat map from Positive low correlation to Positive high correlation it represents with dark Green to light Green colour and Negative low correlation to Negative high correlation it represents with Light Blue to Dark Blue colour. From these correlation values we can relate the relation between variables

K-Means Clustering

The K-means clustering algorithm is an algorithm often used to draw insights into formats and differences within a database. In marketing, it is often used to build customer segments and understand the behaviour of these unique segments. Build an assembly model in Python's environment.





Using Elbow Inertia Method, Inertia measures the sum of squared distances of samples to their closest cluster centroid. With the same number of clusters, smaller the inertia indicates better clusters. The elbow method determines the optimal number of clusters by looking at the turning point (“elbow” point) of the graph below.

