Buckman Internship Screening Test Report

Aswath Rao V S,

16PD05,

MSc Data Science,

PSG College of Technology

Coimbatore

**Q1.Prediction problem**

Problem Statement –

The dataset (Dataset Problem 1) is provided for the prediction of the noise pressure level. The data set has the following attributes

1) Frequency (Hz)

2) Angle (Degrees)

3) Chord Length (m)

4) Velocity – Free-stream velocity(m/s)

5) Displacement – Suction side displacement thickness (m)

You are required to predict the generated noise i.e., Noise Pressure (Decibels)

Analyse the patterns and insights from the following dataset and build a model which has the best accuracy for prediction. Explain all the steps from data pre-processing, model selection, model validation etc.

Objective

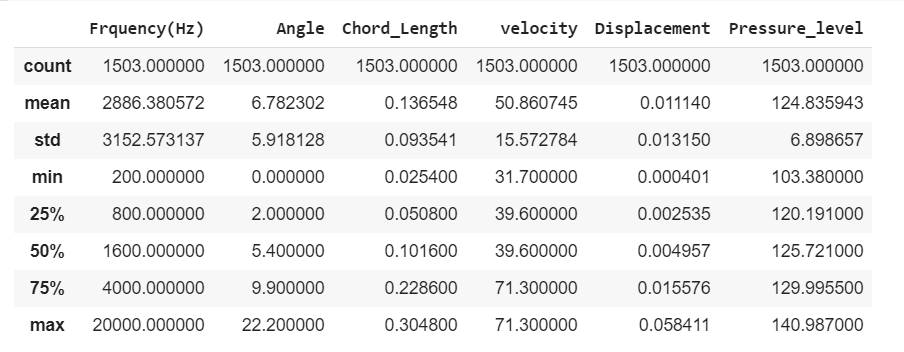
To predict the values of generated Noise.

Understanding the problem

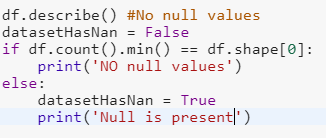
Since output is to be predicted from a set of inputs, Supervised learning algorithm must be used.

Exploratory Data Analysis

1. data.describe()



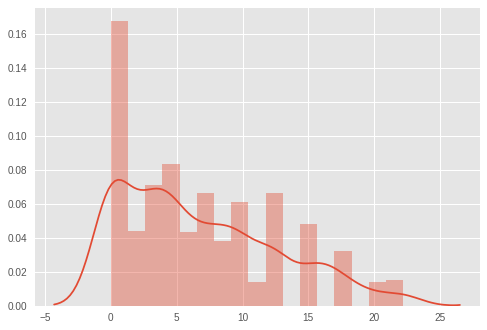
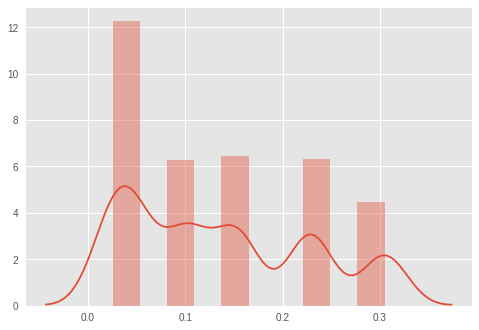
1. To check whether is there any null value in the data :



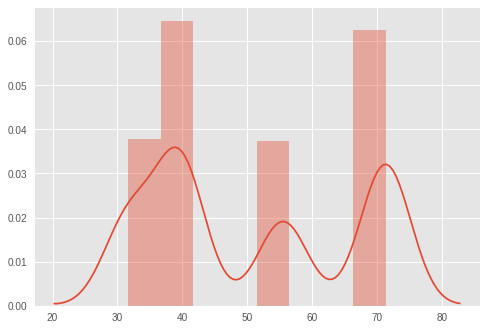
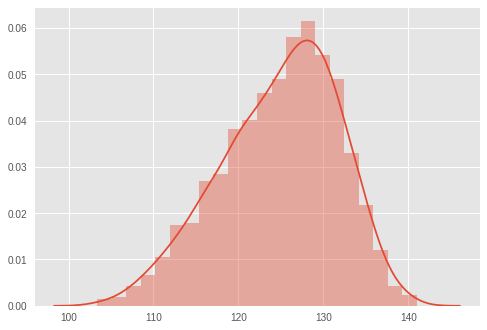
Inference : No null

1. Distribution of features

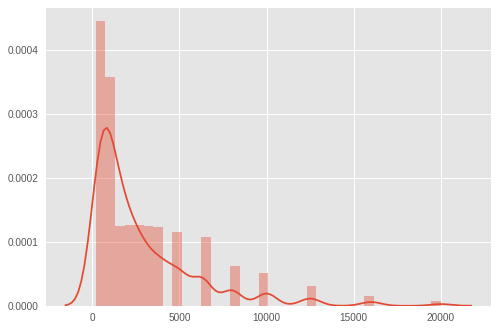
 

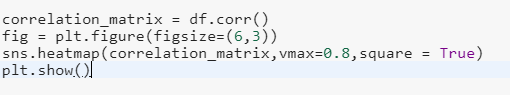
 

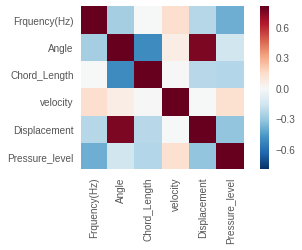
 

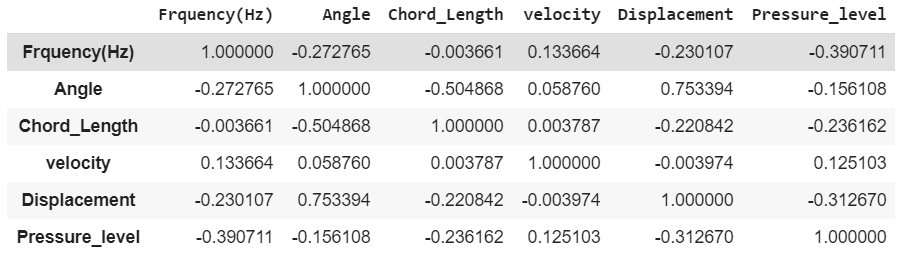




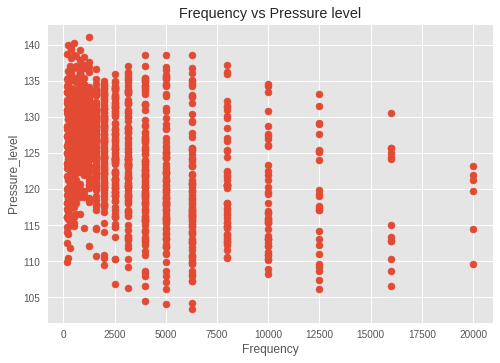
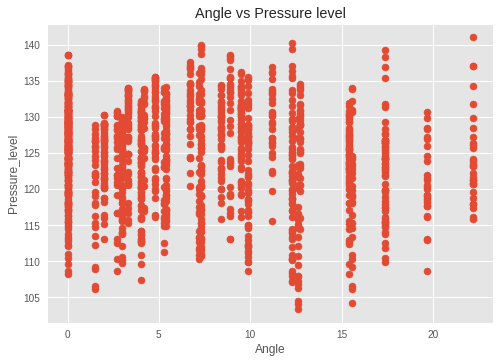
4.Corelation matrix

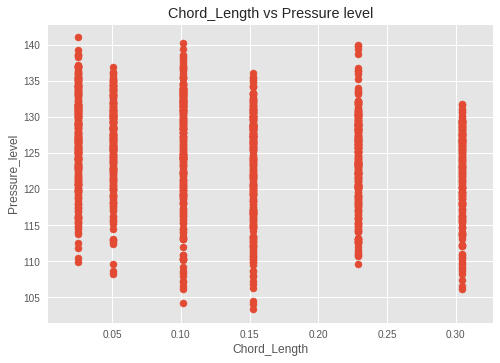
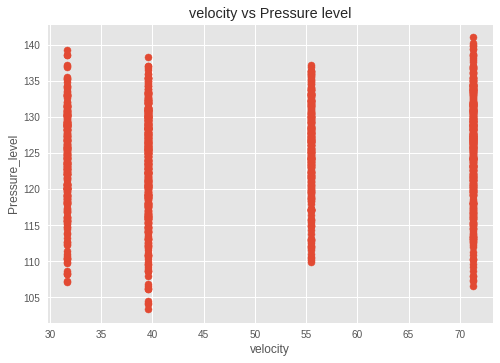


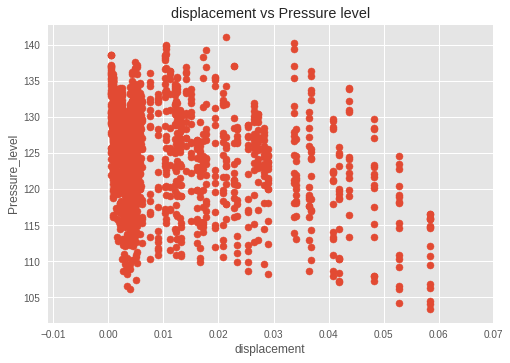




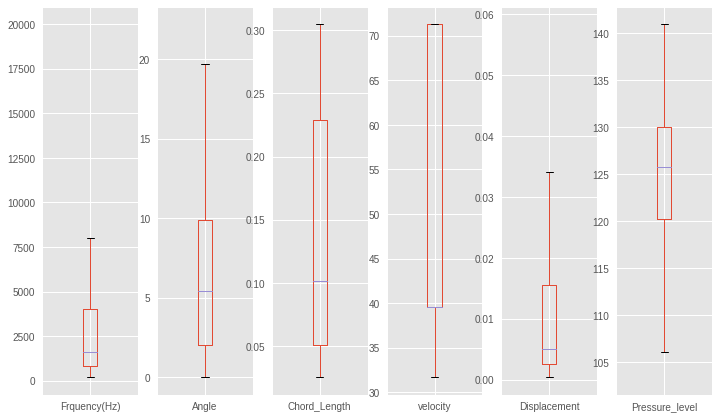
5.Scatter plot between pair of features



6. Outlier detection



Inference - There are no outliers

7.Splitting the data

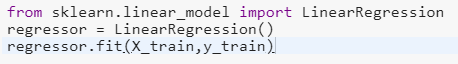
Train set and test set are separated by hold out method



8.Model building

Approach 1 :

Multiple Linear Regression





Model fitted :





Mean Absolute Error :



Mean Absolute Errror :



R-square value :

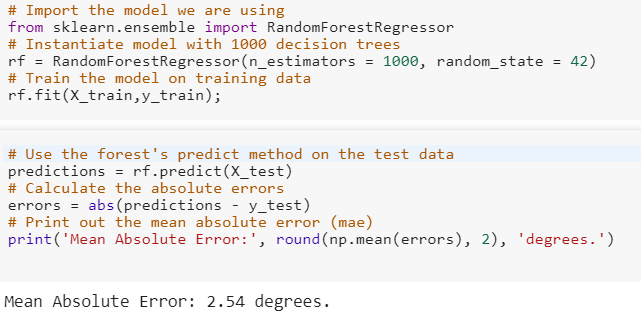




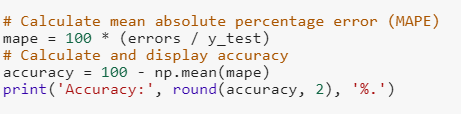
Approach 2 :

Using ensemble method

* Random forest



In terms of Error , Random forest has given better results

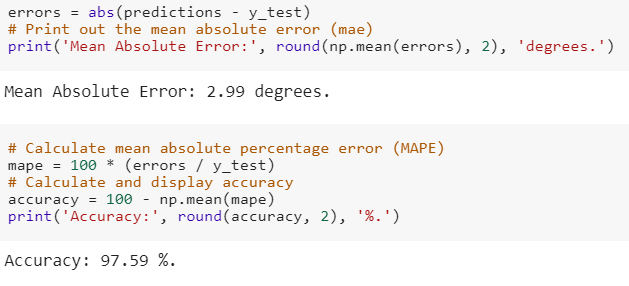


Accuracy achieved : 97.96%

To find the best parameter using grid Search and using that we do random forest



Results -



Q2.

Understanding the problem:

Given the 3 datasets which consists the details of Customers

1. Customer\_Demographics.xlsx – consists details of Customers
2. Customer\_Transaction.xlsx – consists Transaction details
3. Store\_Master.xlsx – consists details of the stores present

Data is imported in oracle

**Query 1** :

Write a SQL query to find out those customers and their demographics who have visited a store for more than ten times.

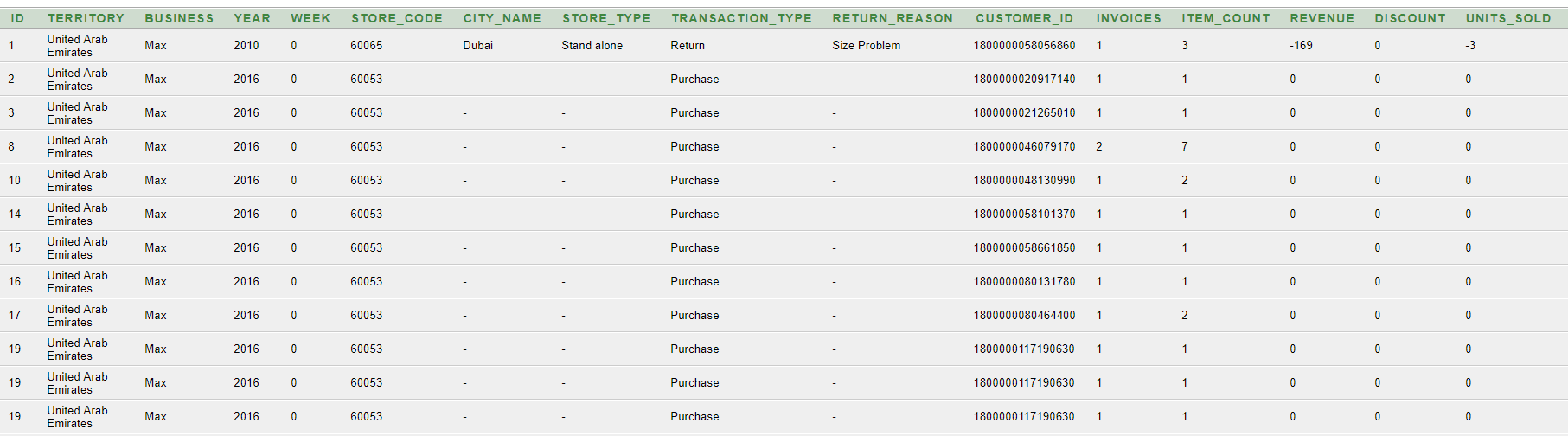
Solution:

Group by, count and Joins are used to solve this problem

Query:

select a.\*,b.Store\_Code from Customer\_Transaction a, (select count(\*),Customer\_ID, Store\_Code from Customer\_Transaction group by (Store\_Code, Customer\_ID) having count(\*) >=10) b where a.Customer\_ID=b.Customer\_ID

Sample Output:



**Query 2 :**

Write a SQL query to create a column called “customer rank” which will rank

the customers based on their frequent visits.

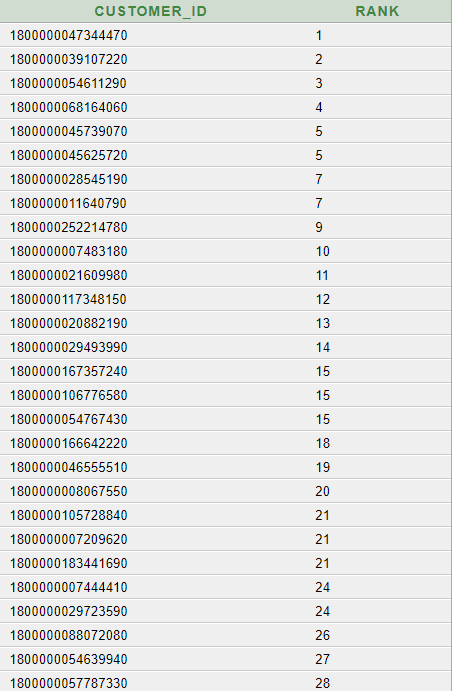
Solution :

Group by, count and Rank are used to solve this problem

Query :

select Customer\_ID, RANK() OVER(ORDER BY freq desc) as rank from (select Customer\_ID, count(\*) as freq from Customer\_Transaction group by Customer\_ID order by count(\*) desc)

Sample Output :



**Query 3 :**

Write a SQL Query which will create a master table (Combined all the three tables) for all the stores and customers taken into account.

Solution :

Inner Joins are used to solve this problem

Query :

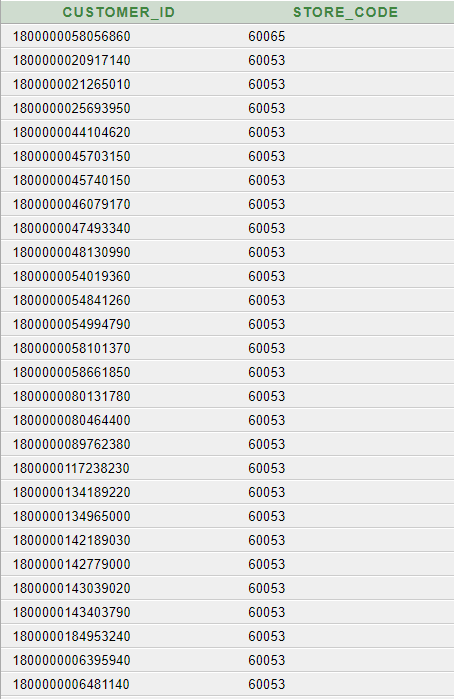
SELECT Customer\_Demographics.Customer\_ID, Customer\_Transaction .Store\_Code

FROM ((Customer\_Demographics

INNER JOIN FIRST ON Customer\_Demographics.Customer\_ID = Customer\_Transaction.Customer\_ID)

INNER JOIN Store\_Master ON Customer\_Transaction.Store\_Code = Store\_Master.Store\_Code);

Sample Output:



Q3. Time Series Analysis

Understanding the data set :

The data set Dataset Problem 3.xlsx consists of Time series data of power production from 2006-2017. The data set has 5 attributes with Date, Power, Consumption and Energy

Inference from data set

The data set consists of missing values for columns E1,E2 and E3 where missing values of E1 is lesser

We have complete data from 2012

Steps followed :

1. Question given
2. Understanding the problem
3. Approach
4. Code Snippet
5. Graph

Question given:

Plot the same graph given below for 2017 instead of the consumption rate take E1, E2 in the Y-axis (Note: Separate graph for E1 & E2 with the months in a year as X-axis)

Understanding the problem:

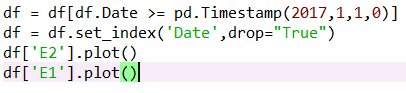
Plot the graph as given below by taking E1 for the year 2017

Plot the graph as given below by taking E2 for the year 2017

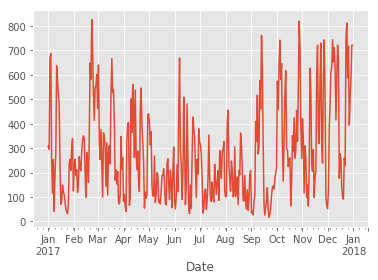
Approach:

The year 2017 is taken from Timestamp pandas ,Date is set as the Index and graph is plotted for E1 and E2

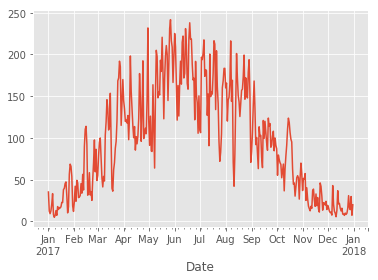
Code Snippet:



Graph for E1:



Graph for E2:



**Question given:**

Aggregate the data on a seasonal basis for E1 and E2 and plot the same for the year 2017.

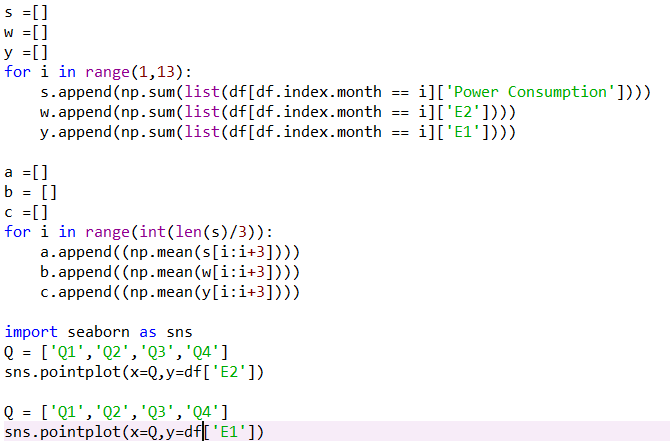
Understanding the problem:

To Infer the seasonal trend in 2017

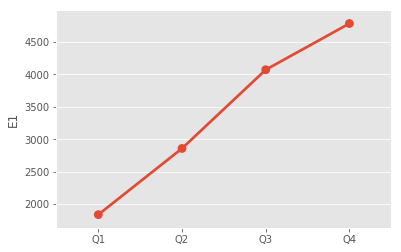
Approach:

The data in 2017 is aggregated into 4 season and trend is analyzed using the given plot

Code Snippet :



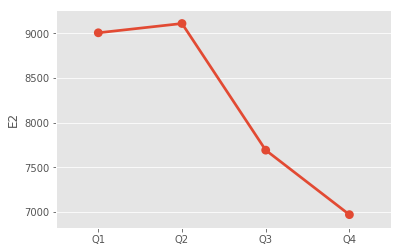
Graph for E1:



Observation:

There is an Increase trend for E1

Graph for E2:



Observation:

There is an Increase trend for E2

**Question given :**

Does the power consumption depend on E1 and/or E2? Justify your answer using

plots/tables?

Understanding the problem :

To find correlation between power and E1,E2

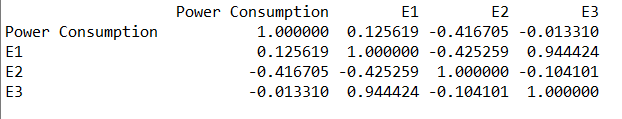
Approach:

Correlation matrix is tabulated

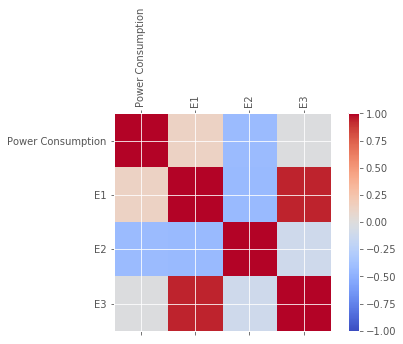
Code Snippet :



Table:



Graph:



* E2 is more depended on Power
* E2 and Power are negatively co-related

Q4. Problem on Statistics

Given problem –

From customer transaction data provided in dataset, find the probability of each customer visiting each store. Say, for a given customer what is the probability that he will visit a store? Include your output as probability.csv and explain the formula.

Insights from Data:

There are 1,00,000 Unique Customers and 32 Unique Stores.

Majority of the customers haven’t visited multiple stores.

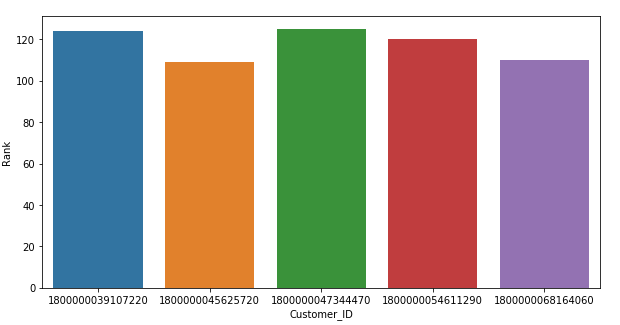
Solution:

Taking the frequency of visit to a store by a particular customer, the probability distributed table is generated. Say a customer visits 3 out of 32 stores with frequencies 10,23,18 the probability of the customer visiting it again would be 10/51, 23/51 and 18/51 respectively.

So the entire space (100000X32) is calculated as mentioned above.

P(visiting a store) = Frequency he visited that store / Total visits made to all the stores

Top 5 Customers with their ranks



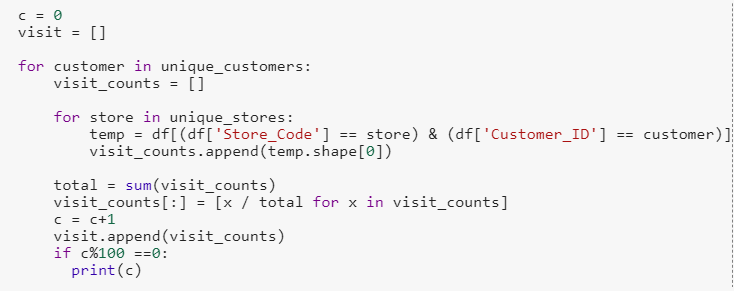
Coding Approach 1

Every permutation of customer, store is generated to find the frequency and followed probability. This will cost O(100000x32) . This is performed using nested loop.

Coding Approach 2

Customer and stores are grouped first and then the group object is iterated to find frequency of visits made. This reduced to O(1,84,000)

Coding snippet :



Sample Output :

