Flipkart Sample Data Analysis

by Prithvi Shetty

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

Flipkart Sample Data Analysis

by Prithvi Shetty

Table of contents:

List of Tables & Figures.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

Flipkart Sample Data Analysis

by Prithvi Shetty

1. Abstract

Ecommerce, also known as electronic commerce or internet commerce, refers to the buying and selling of goods or services using the internet, and the transfer of money and data to execute these transactions.

This is a sample data set from Kaggle, Flipkart is one of the high growing e-commerce sectors in the recent years. This is dataset includes details of the products. We are going to build some meaningful information from the given data.

Various tasks have to be done to get better insights regarding the data, hence data has to be cleaned, transformed, descriptive and inferential statistics has to be applied. Handling of missing values, detecting and removing outliers in the data set etc.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

2. Introduction

2.1 Motivation

Flipkart is an Indian e-commerce company, headquartered in Bangalore, Karnataka, India, and incorporated in Singapore as a private limited company. The company initially focused on online book sales before expanding into other product categories such as consumer electronics, fashion, home essentials, groceries, and lifestyle products. The service competes primarily with Amazon's Indian subsidiary and domestic rival Snapdeal.

As of March 2017, Flipkart held a 39.5% market share of India's e-commerce industry. Flipkart has a dominant position in the apparel segment, bolstered by its acquisition of Myntra, and was described as being "neck and neck" with Amazon in the sale of electronics and mobile phones. Flipkart also owns PhonePe, a mobile payments service based on the Unified Payments Interface.

In August 2018, U.S.-based retail chain Walmart acquired an 77% controlling stake in Flipkart for US$16 billion, valuing Flipkart at around $20 billion.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

2.2 Project Scope

In this project we are going to analyze, interpret the key results where the organization can earn potential income. We are going to analyze the data using various statistical methods, python programming language, machine learning libraries, stats libraries and data visualization is being done in tableau, excel and powerBI.

This dataset contains the records of products sold in Flipkart platform. Before going ahead with the analysis, data has to be cleaned, and transformed in case of machine learning algorithm.

2.3 Project Goal

- Our goal is to predict retail price and discounted price using machine learning model.
- Finding out the key areas for development purposes.
- Reducing risks/errors.
- Using unsupervised learning classification algorithm to predict the main category of the product.
- Descriptive analysis of the data.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

2.4 Organisation of the report

- Understanding the business.
- Basic Data understanding.
- Exploratory Data Analysis.
- Feature Engineering.
- Data Visualization.
- Statistical Analysis.
- Model Interpretation.
- Model Validation.
- Performance Tuning.
- Interpreting Results.
- Drawing Conclusion.
- End recommendations for developers.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

3. Project Description

3.1 Dataset understanding

This dataset is taken from Kaggle, this data is regarding the products details bought by the customers. The excel dataset is named as flipkart_ecommerce_sampledata.csv. This dataset contains 20000 rows and 19 columns. Each attribute contains the details such as name of the product, ratings, overall rating, description, price, discounted price, categories of numerous products.

The dataset contains missing values, unwanted characters and these must be treated accordingly. Using various statistical methods, we can interpret the relationship between different attributes.

Data transformation must be done for categorical columns for modeling purposes.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

3.2 Data Limitations

The dataset contains various missing values. Conversion of the data, removal of unwanted characters in the dataset.

Feature Engineering to reduce the unnecessary columns in the dataset and finding the relationship between the independent features of the application. Data Transformation of the data and outliers (the extreme values must be treated).

This will improve the accuracy of the model and interpret better results.

The dataset does not include review of the customers, hence natural language processing can't be done to get the sentimental analysis. But here I will consider product description as review to show the demo of text analysis.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

4. Exploratory Data Analysis

4.1 Data cleaning

- Dropped unnecessary columns such as uniq_id, product_url, product_category_tree, pid,image, description, product_specifications.

- Converted product_category_tree into 4 different columns as using string functions as main_category, send_sub_category, third_sub_category, fourth_sub_category.

- As the dataset contained outliers, we detected it and removed using z-score and IQR.

- Converted timestamp column to date time format for computation.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

4.2 Data Transformation

- In order to implement machine learning to the dataset the data has to be transformed in such a way that the algorithm can study/train the data in efficient manner.

- In this project the dataset there are different type of attributes that is categorical, numerical and Boolean values.

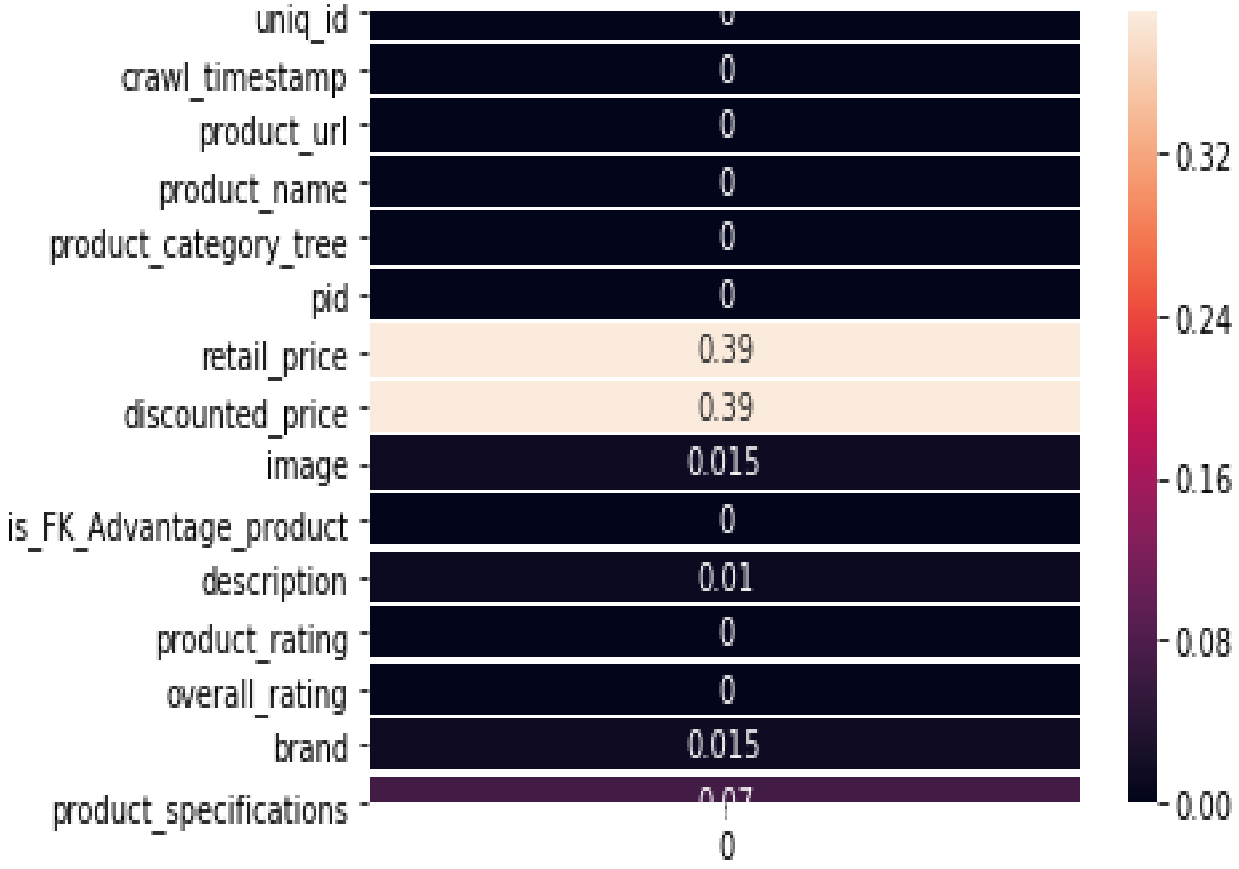
- However, here the categorical data is transformed to numerical.

- We use Sklearn Pre-processing package and the encoder for the data transformation that is applied is Label Encoder.

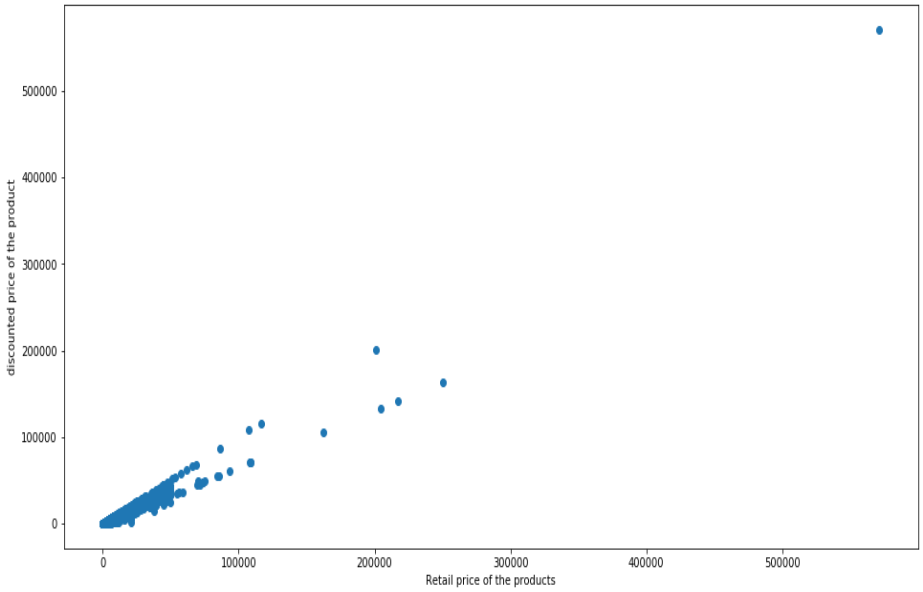
- As a result, we can observe that the categorical data is now converted to numerical data and then made available for implementing machine learning model for prediction.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

This is a heat map that represents the missing values present in the dataset:

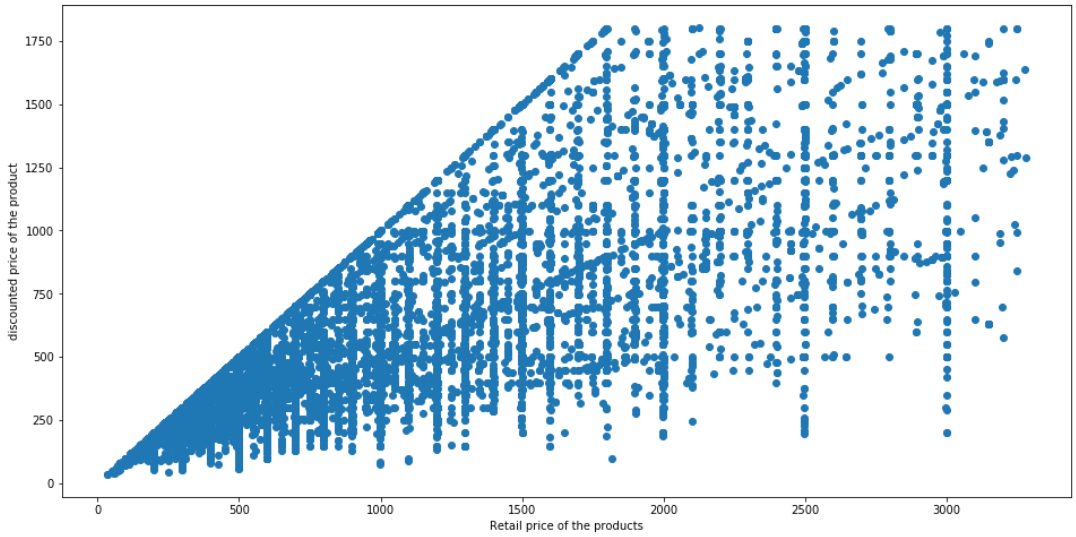


This scatter plot represents the spread of discounted price and retail price of the product. These are highly correlated hence when retail price increases discounted price increases and vice-versa .



Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

This scatter plot represents the data after removing in the dataset.



This screen shot is linear regression machine learning model that predicts the price of the products. As shown below the accuracy of the model to predict right information is 92%.

jupyter Untitled Last Checkpoint: 21 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run Code

In [150]:
```
# train the model using the training sets
reg.fit(X_train, y_train)

# regression coefficients
print('Coefficients: ', reg.coef_)

# variance score: 1 means perfect prediction
print('Variance score: {}'.format(reg.score(X_test, y_test)))
```

```
Coefficients:  [ 1.20504601e+00 -1.33995229e+01 -1.33995229e+01 -1.83876024e+00
 -1.12472380e-01 -9.90995002e-02 -5.50212325e-02  2.92288917e-02
 -2.94431323e+02]
Variance score: 0.9632097883584211
```

In [155]:
```
y_pred = reg.predict(X_test)
y_pred
```

Out[155]:
```
array([1401.08384741, 2023.79957899,  769.34823546, ..., 2719.7030969 ,
        502.8151233 , 1022.89190184])
```

In [156]:
```
from sklearn.metrics import r2_score
r2 = r2_score(y_pred, y_test)
print('r2 score for perfect model is', r2)
```

```
r2 score for perfect model is 0.9614176224907944
```


Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

4.3 Data Visualization using excel:

Created excel dashboards and applied complex functions and created macro file, macro file is basically transforming the unprepared data to the given format. From the below dashboard we can see and analyze the data.

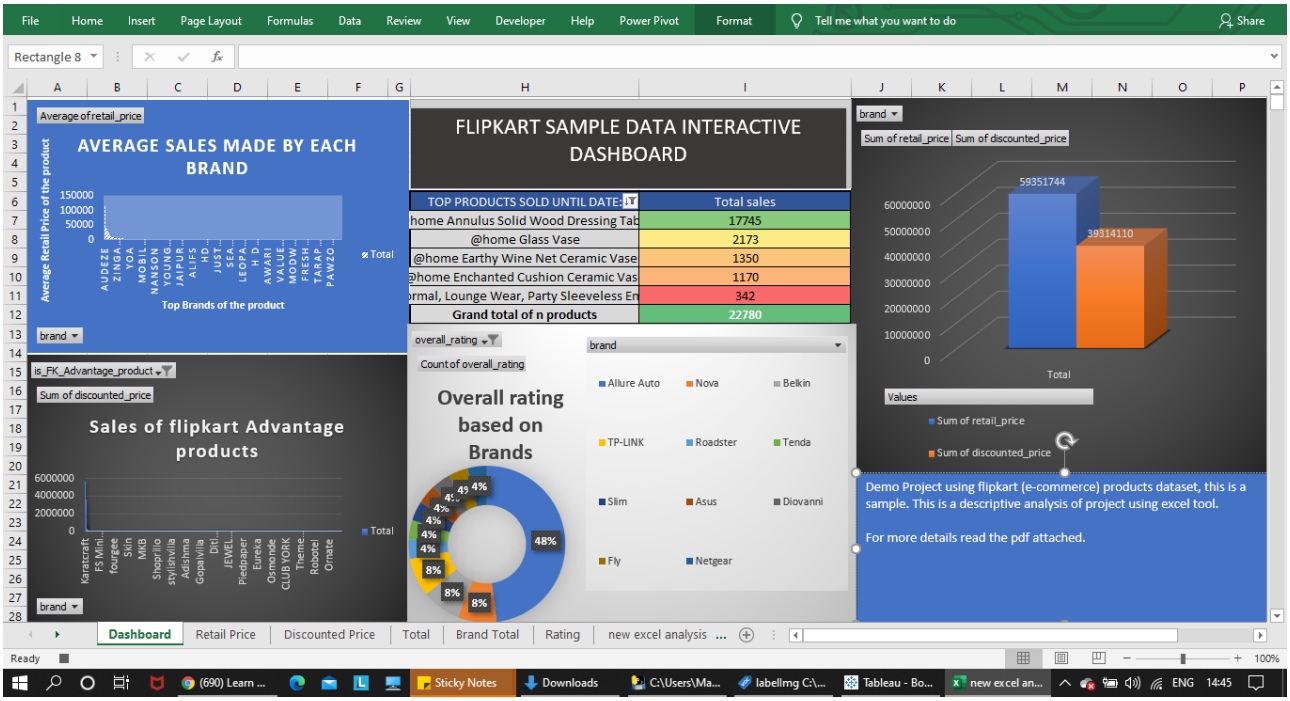
Further Descriptive analysis:

With the help of this dashboard, we can:

1. Identify best-selling and least selling products and their sales.
2. Overall and product rating of the product by each brand.
3. Identify the below average rated products with respective of sales above average and vice-versa.
4. Top and least performing brands, discount percentage of each brand.

Lot more analysis can be done using this dashboard (using filters). This shows immediate results based on the data used.

Please refer to the below image:



Excel Data: Click here..

Excel Macro File: Click here..

Note: For better functionality use excel to read the file.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

4.4 Data Visualization using Tableau:

Created Tableau dashboard with 3 types of charts.

Here we can see that on 2016 January there was a huge demand for jewellery products and after January the demand drastically decreases, this might be due to various factors i.e., covid lockdown. People intent to be at home hence there is no demand for the product in the market.

Likewise, we can compare the category of the products that is best suited to either promote or advertise to generate potential income in the organization.

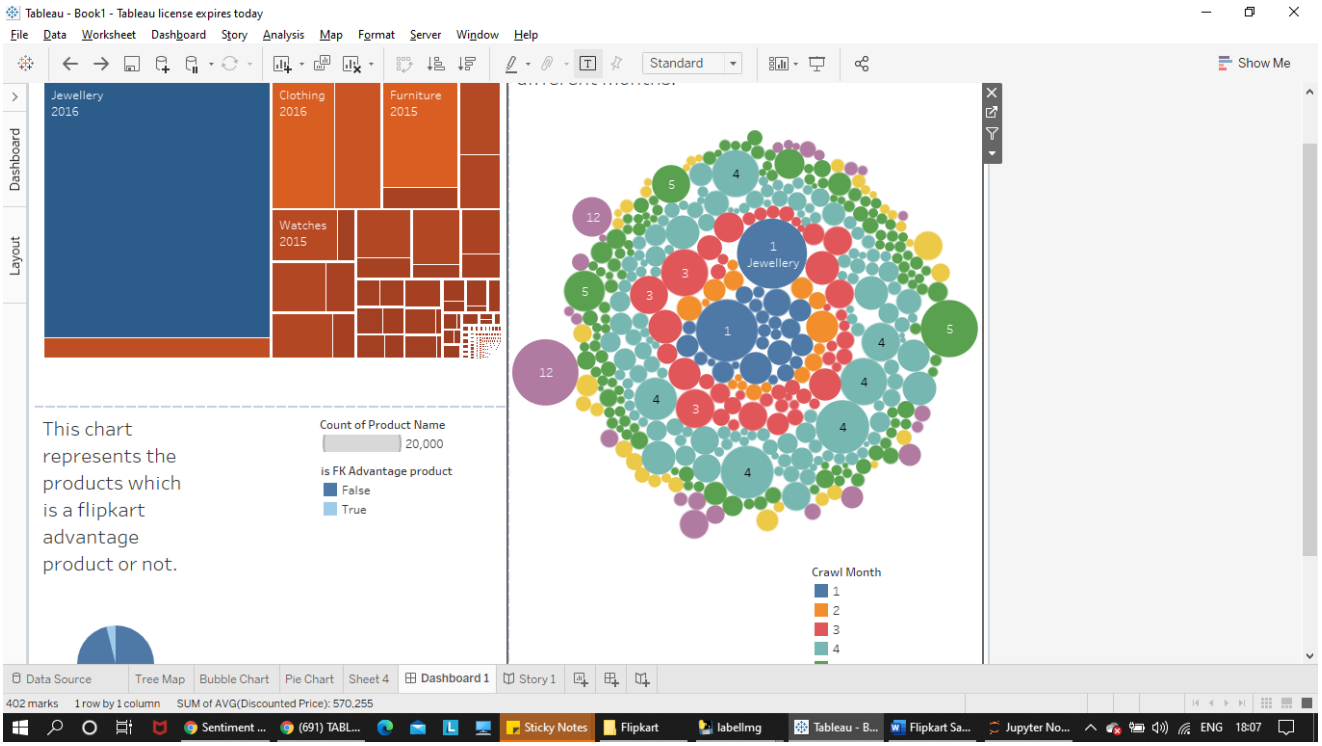


Tableau Worksheet: [Click here..]

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

4.5 Future Work:

- Following lean six sigma methodology DMAIC.
- Gathering the information additional information from the customer such as reviews and applying text analytics and natural language processing techniques and creating a sentimental analysis.
- This benefits us to know the errors and we can use effective techniques to control it.
- Focusing of the current economic trend and past data, knowing the difference.
- Prediction using Logistic Regression algorithm to predict whether the product is Flipkart advantage product or not.
- Aggressive marketing on social medias, blogs etc will lead to potential growth in sales. Taking benefits of affiliate marketers.
- Track and recording feedbacks of frequent buyers of the product.
- For potential growth of the business or the process we definitely need to know the key information of the product/service i.e., quality, performance, features, reliability, durability, serviceability.

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

5. References:

Dataset: Kaggle.

Link: Click here..

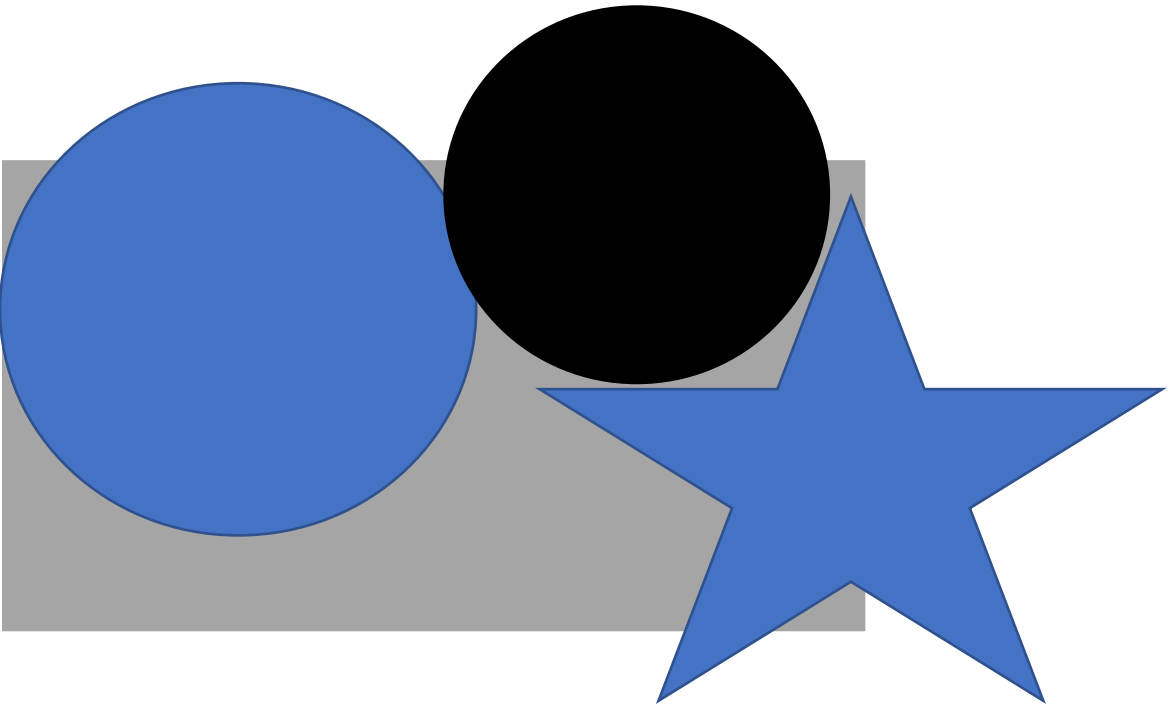
Python Programming language code: Click here..

(Converted as HTML)

Note : Download and open in Goocgle Chrome

Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480

THANK YOU



Email: shettyprithvi16@gmail.com
Phone: +91 888 078 1480