RESTAURANT RECOMMENDATION AND RATING PREDICTOR

Team ABCD

- A – Anubhav Ghosh
- B – Bibekananda Mishra
- C – Chiranjeevi Pippalla
- D – Dinesh Dandamudi

OUTLINE

What are the topics we are going to see now?

- Introduction
- EDA
- Regression
- Decision Tree
- Random Forest
- Recommendations
- Anomalies
- Learning Outcomes.

INTRODUCTION

What is this project about?

- Basic idea of analyzing this dataset is to get a fair idea about the factors affecting the establishment of different types of restaurants at different places in Bangalore.
- Predicting the ratings of the restaurants in Bangalore is mainly based on
  - Location of the restaurant
  - Most liked cuisine
  - Most liked dish
  - Average cost for two people

INTRODUCTION

Role of Data-Science

- Explore the factors that play important roles in business
- Use classification / regression techniques to make predictions about ratings
- Make a Recommendation System

DATA EXPLORATION

DATASET

Features

- url
- address
- name
- Online order
- Votes
- Location
- Restaurant type
- Dish liked
- Cuisines
- Listed in (type)
- Listed in (city)

DATA EXPLORATION

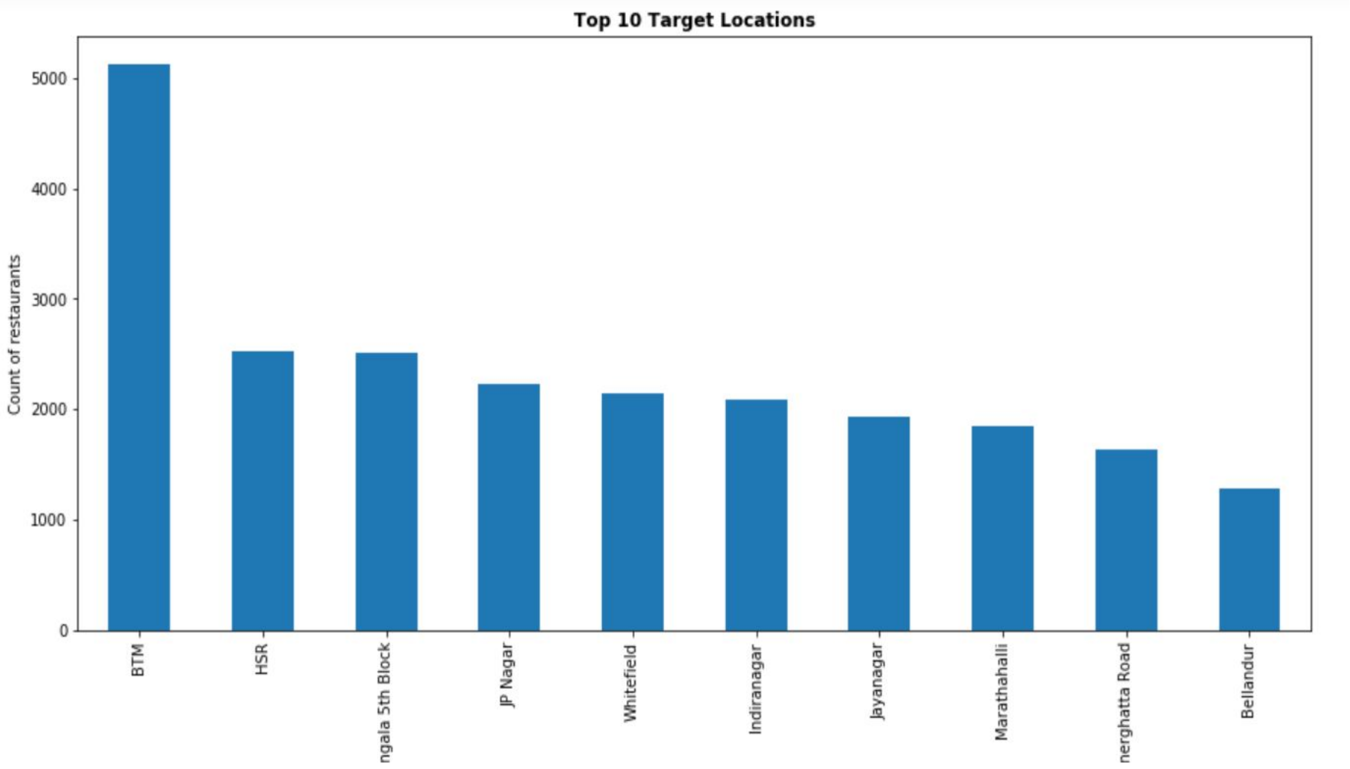
Data after dropping unnecessary columns and rows

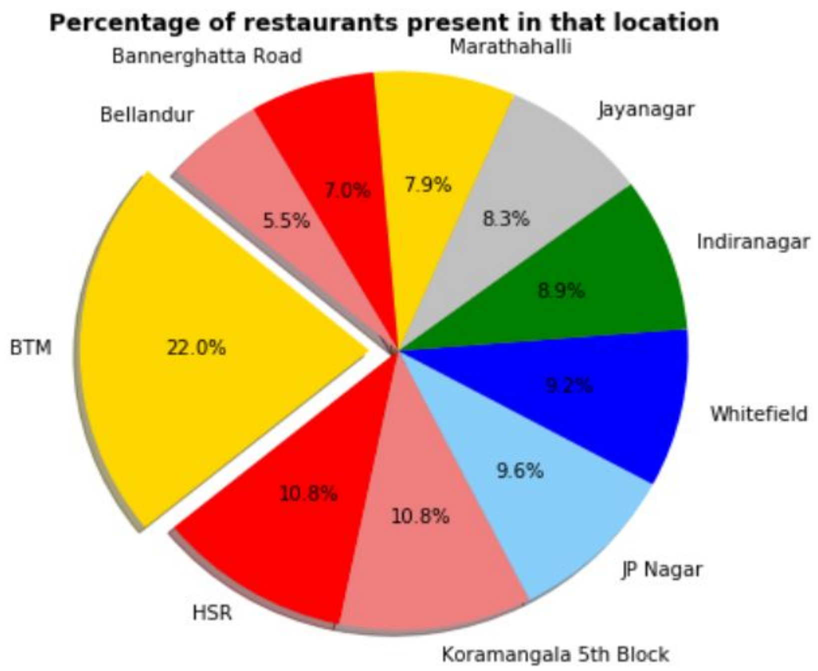
```
df.head()
```

	name	online_order	book_table	rate	votes	location	rest_type	dish_liked	cuisines	average_cost	reviews_list	menu_item	listed_in(type)
0	Jalsa	Yes	Yes	4.1/5	775	Banashankari	Casual Dining	Pasta, Lunch Buffet, Masala Papad, Paneer Laja...	North Indian, Mughlai, Chinese	800	[('Rated 4.0', 'RATED\n A beautiful place to ...	[]	Buffet
1	Spice Elephant	Yes	No	4.1/5	787	Banashankari	Casual Dining	Momos, Lunch Buffet, Chocolate Nirvana, Thai G...	Chinese, North Indian, Thai	800	[('Rated 4.0', 'RATED\n Had been here for din...	[]	Buffet
2	San Churro Cafe	Yes	No	3.8/5	918	Banashankari	Cafe, Casual Dining	Churros, Cannelloni, Minestrone Soup, Hot Choc...	Cafe, Mexican, Italian	800	[('Rated 3.0', "RATED\n Ambience is not that ...	[]	Buffet
3	Addhuri Udupi Bhojana	No	No	3.7/5	88	Banashankari	Quick Bites	Masala Dosa	South Indian, North Indian	300	[('Rated 4.0', "RATED\n Great food and proper...	[]	Buffet

DATA EXPLORATION

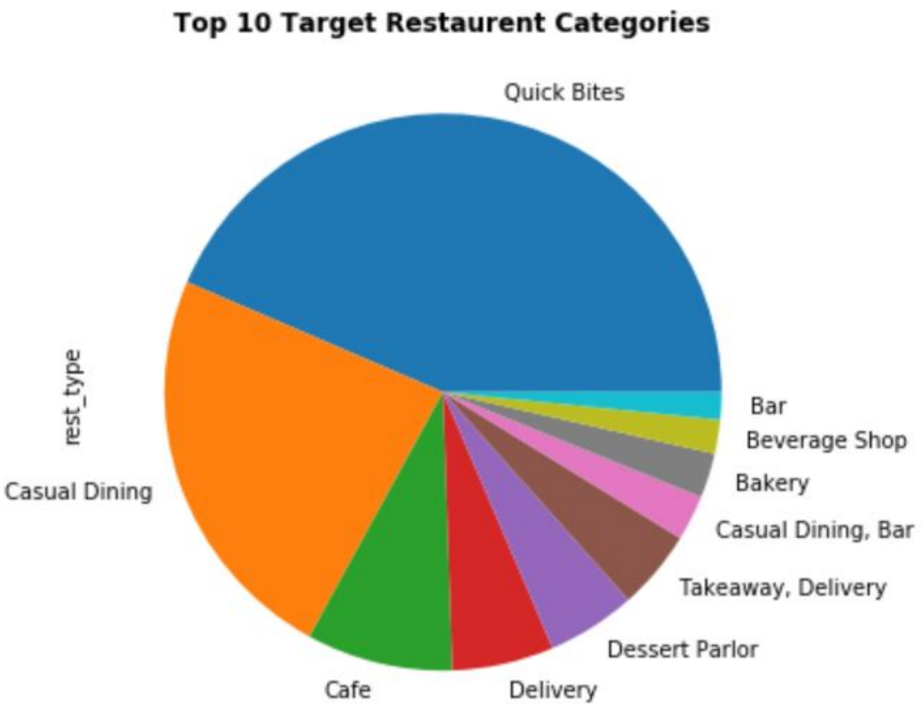
TOP 10 LOCATIONS

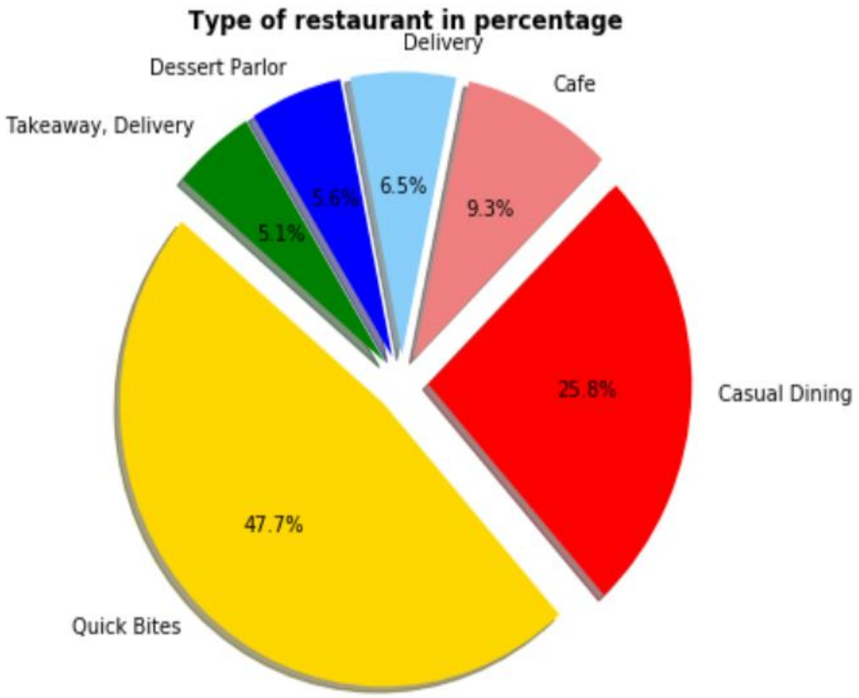




DATA EXPLORATION

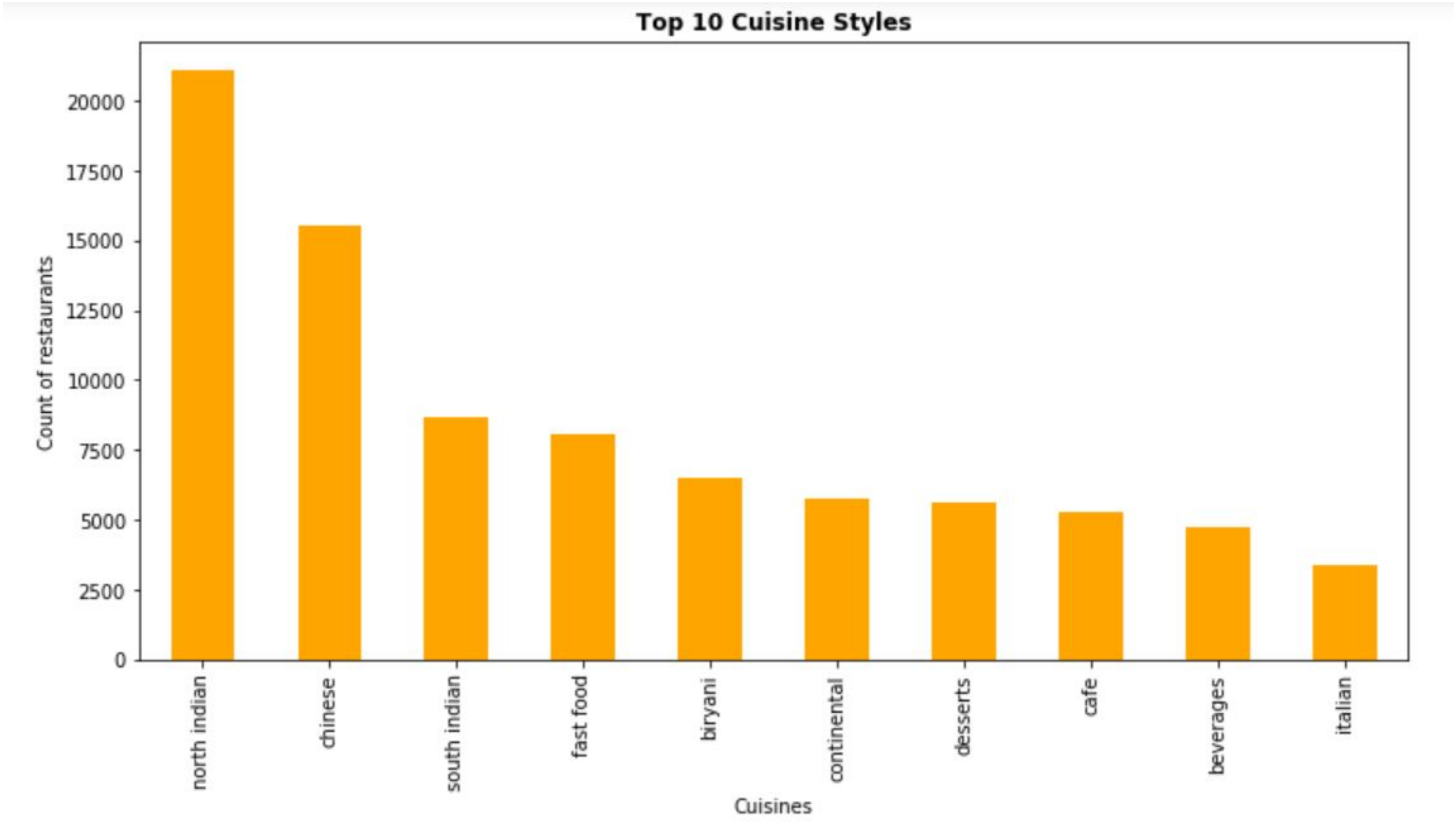
TOP 10 RESTAURANT CATEGORIES





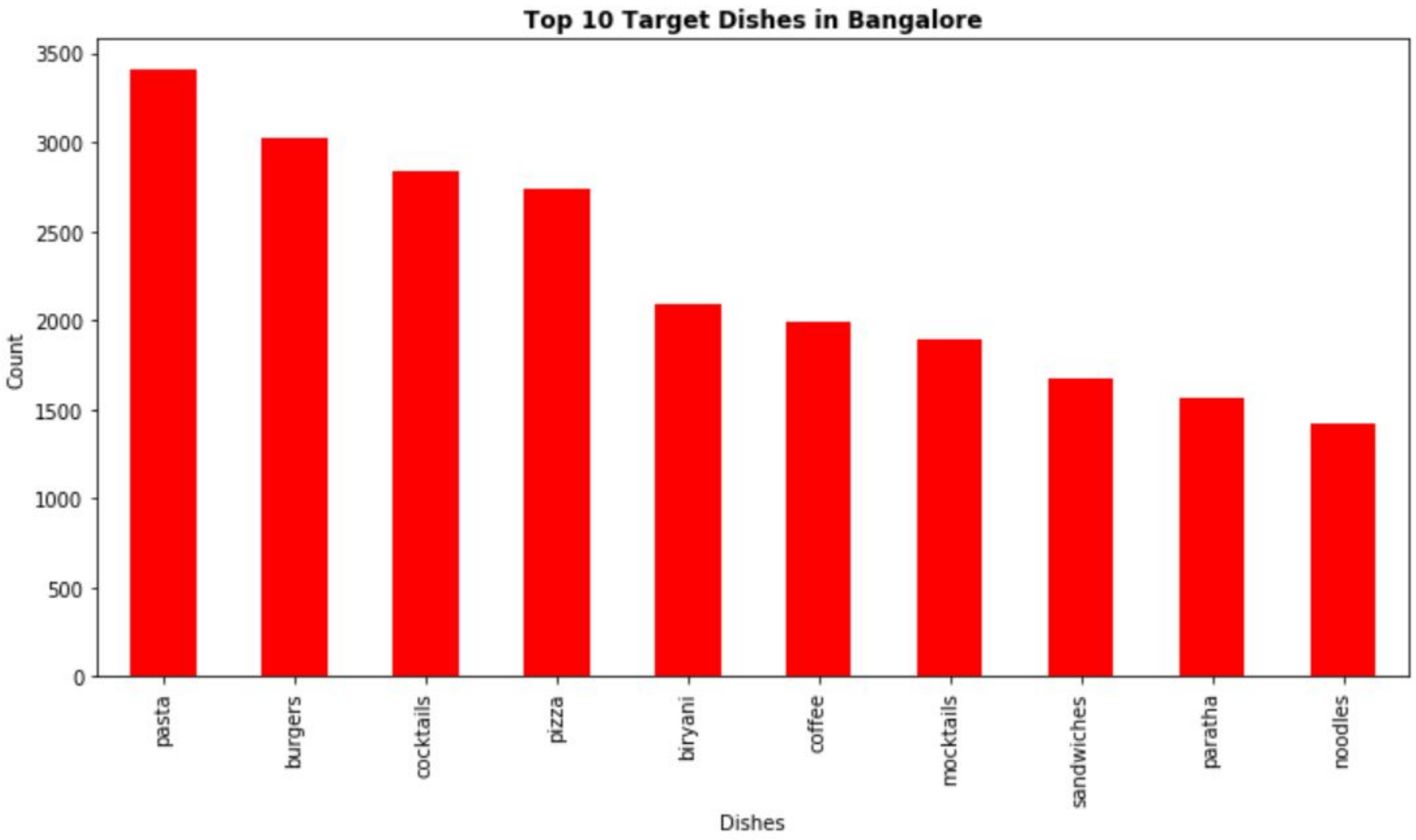
DATA EXPLORATION

TOP 10 CUISINE STYLES



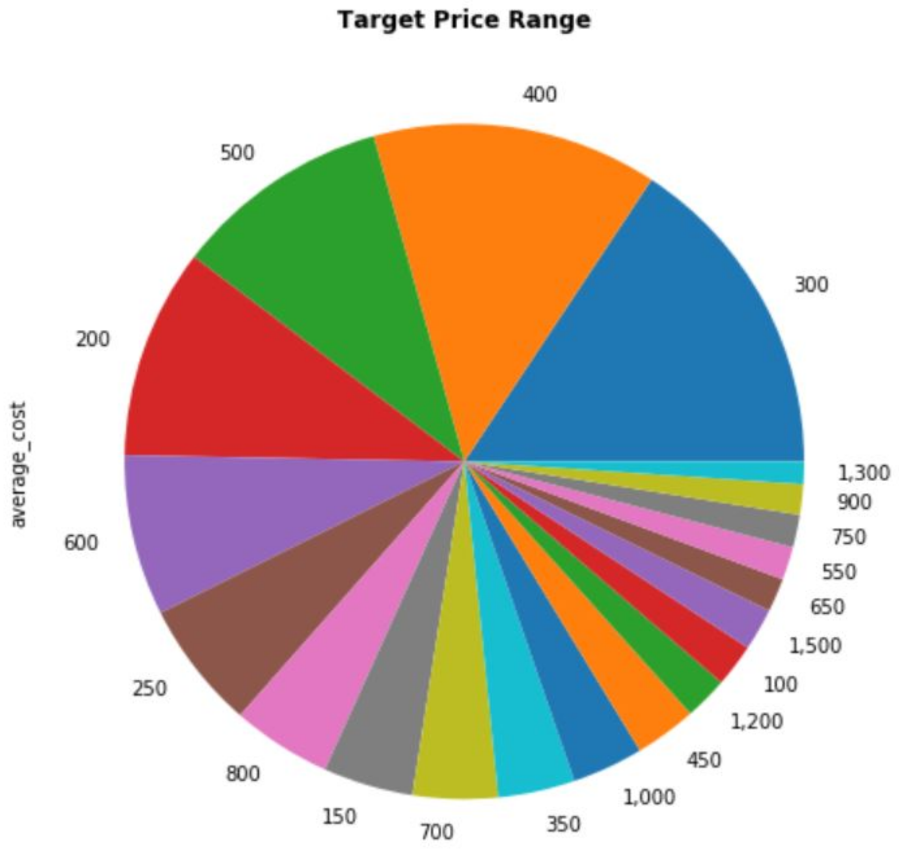
DATA EXPLORATION

TOP 10 DISHES



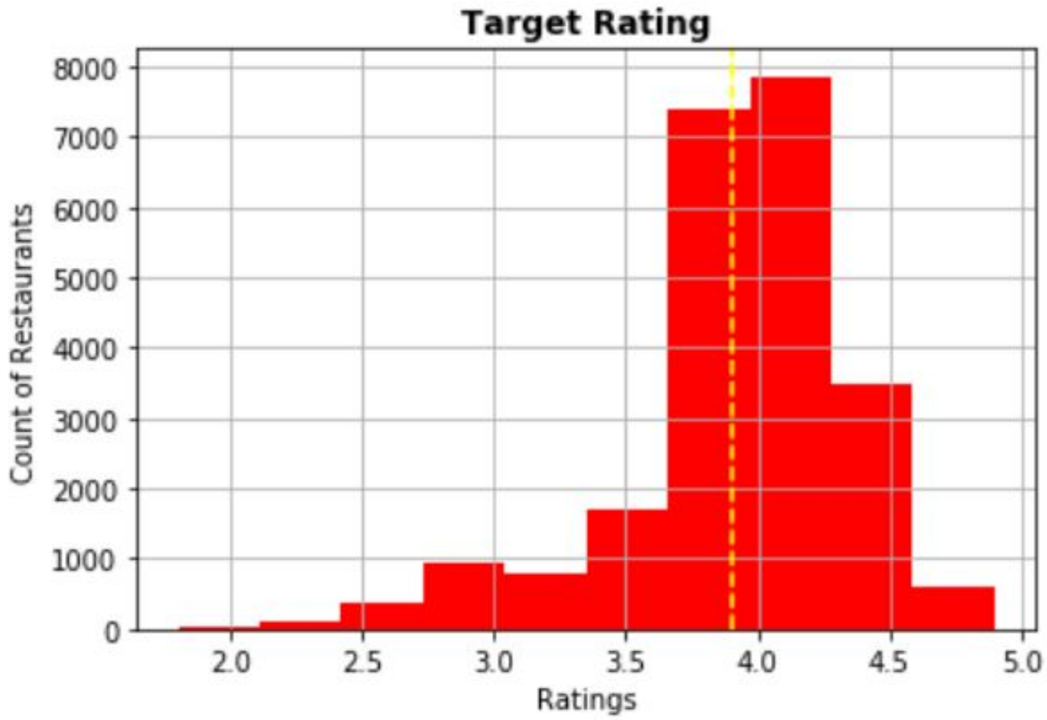
DATA EXPLORATION

AVERAGE PRICE RANGE



DATA EXPLORATION

MEAN RATING

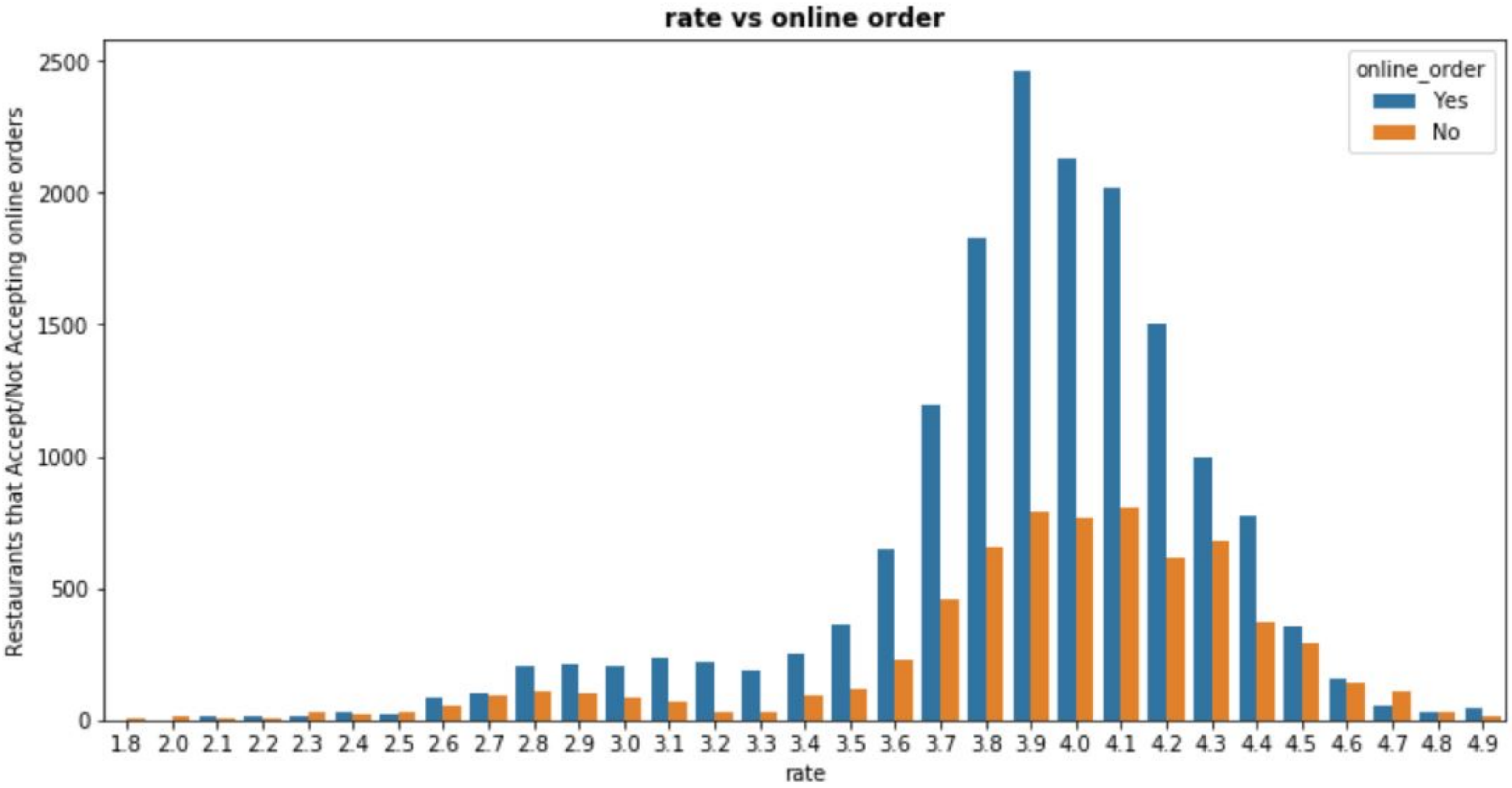


DATA EXPLORATION

RATIO OF RESTAURANTS ACCEPTING ONLINE ORDERS AND NOT

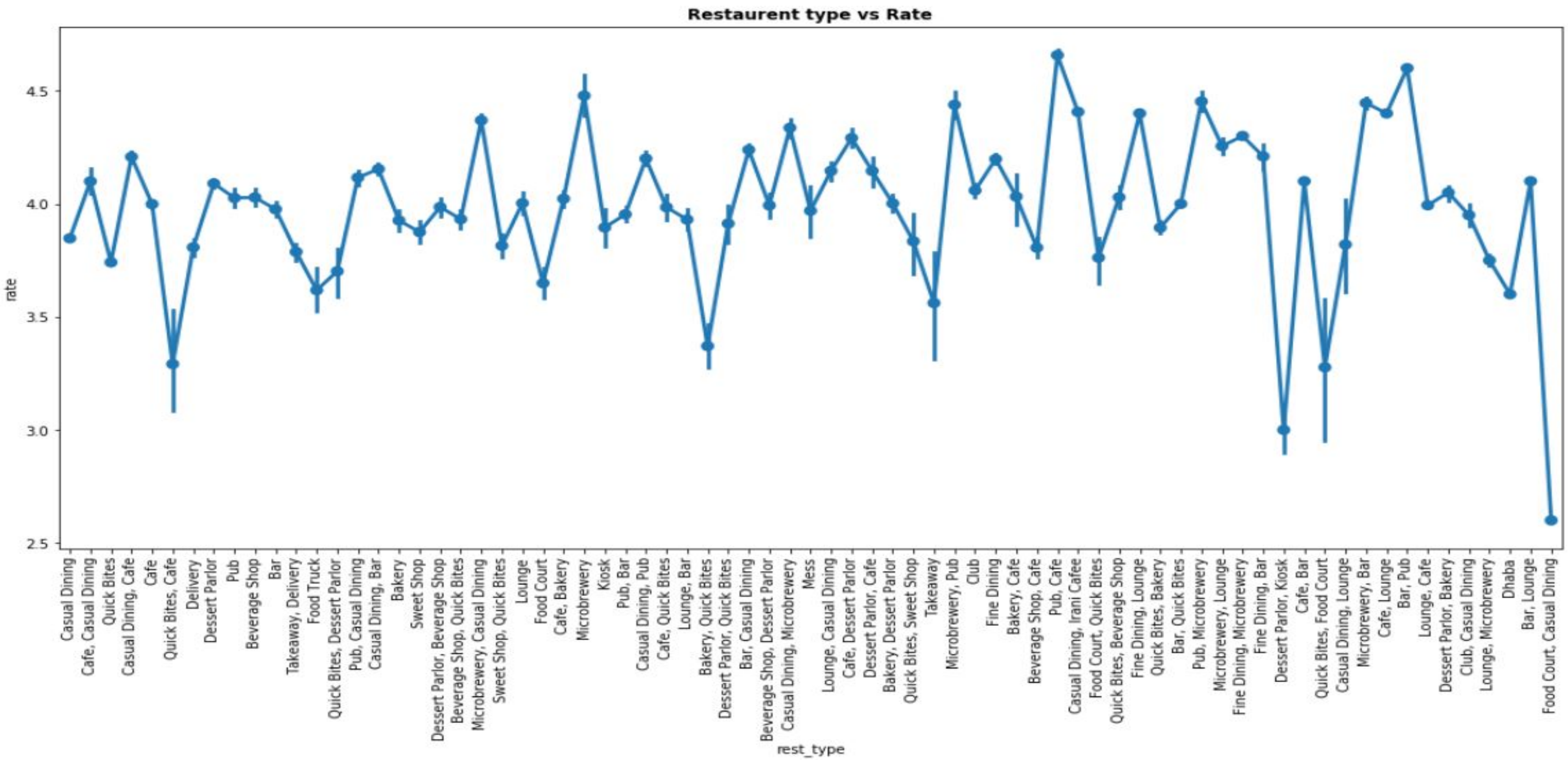
DATA EXPLORATION

RATING VS ONLINE ORDER



DATA EXPLORATION

RESTAURANT TYPE vs. RATING



DATA EXPLORATION

INFERENCES

Based on the EDA, we have the following inferences.

1. Target location - BTM
2. Target Restaurant Category - Quick Bytes
3. Target Cuisine Style - North Indian
4. Target Dish - Pasta
5. Target Price - 300
6. Target Rating - 3.9

LINEAR REGRESSION

DATASET: LINEAR REGRESSION

Analysis

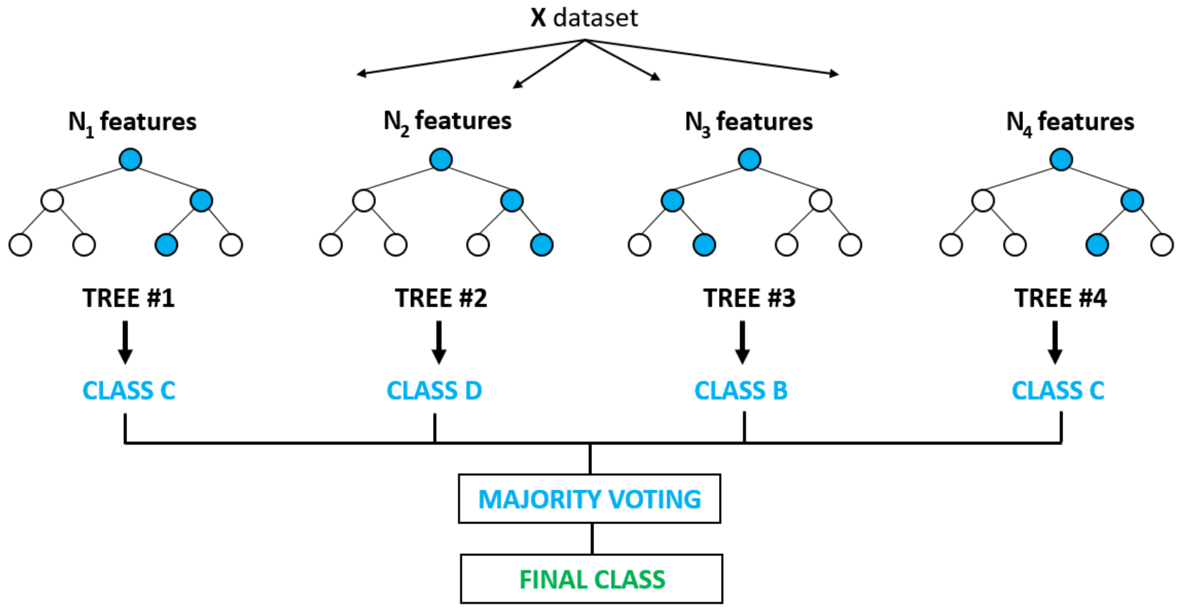
- Feature selection
- Accuracy is very low

RANDOM FOREST CLASSIFIER

DATASET: RANDOM FOREST REGRESSOR

Short Notes

- We use the technique of Bagging which is also called Bootstrap Aggregator
- Boot strapping
- Aggregator



DATASET: RANDOM FOREST REGRESSOR

Analysis

- Feature selection
- Accuracy is good

In [233]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 23259 entries, 0 to 51715
Data columns (total 9 columns):
name            23259 non-null object
online_order    23259 non-null uint8
book_table      23259 non-null uint8
rate            23259 non-null float64
votes           23259 non-null int64
location        23259 non-null int32
rest_type       23259 non-null int32
cuisines        23259 non-null int32
average_cost    23259 non-null float64
dtypes: float64(2), int32(3), int64(1), object(1), uint8(2)
memory usage: 1.8+ MB
```

DATASET: RANDOM FOREST CLASSIFIER

Analysis

- Feature selection
- Accuracy is good

	actual	pred
17577	4.3	4.29
12395	4.1	4.10
41620	3.6	3.40
11719	2.7	2.91
42483	4.0	3.97

DECISION TREE CLASSIFIER

DATASET: DECISION TREE CLASSIFIER

Analysis

- Feature selection
- Accuracy

	actual	pred
17577	4.3	4.30
12395	4.1	4.10
41620	3.6	3.40
11719	2.7	2.70
42483	4.0	3.97

RESTAURANT RECOMMENDER SYSTEM

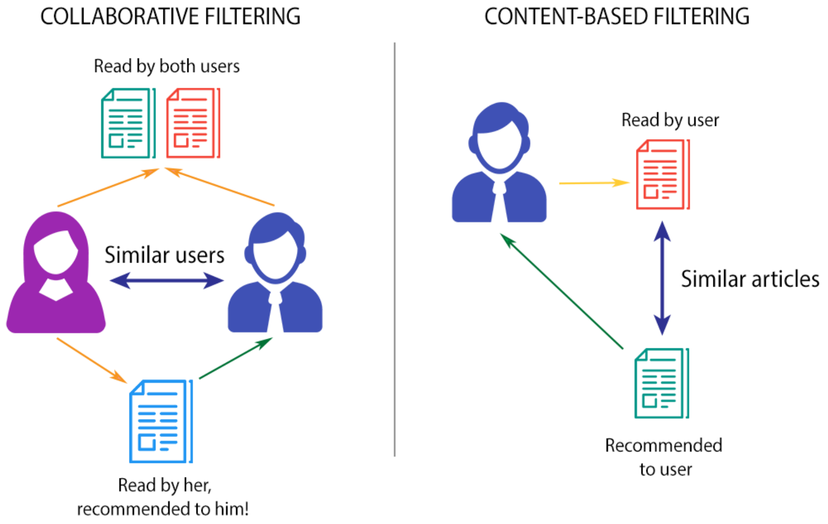
Purpose of the recommender system

The main purpose of this part of is to recommend 2-3 restaurants which are very similar to the restaurant which we choose as a parameter.

RESTAURANT RECOMMENDER SYSTEM

What is the recommendation system we are using and Why?

Our dataset is fit for content based



RESTAURANT RECOMMENDER SYSTEM

Preparing the data for the recommendation system

Out of the 18 features, we only keep 6, as others do not play any role in analyzing the data for the development of a recommendation system.

Trimming the data

We will drop the features which are not required for a recommendation system.

In [35]:
```
drop_cols = ['url','address','phone','book_table','location','reviews_list','listed_in(type)','menu_item','listed_in(ci
dataframe.drop(drop_cols, axis=1, inplace = True)
```

In [36]:
```
dataframe.rename({'approx_cost(for two people)':'cost_for_two'},axis = 1, inplace = True)
dataframe.head()
```

Out[36]:

	name	online_order	rate	votes	dish_liked	cuisines	cost_for_two
0	Jalsa	Yes	4.1/5	775	Pasta, Lunch Buffet, Masala Papad, Paneer Laja...	North Indian, Mughlai, Chinese	800
1	Spice Elephant	Yes	4.1/5	787	Momos, Lunch Buffet, Chocolate Nirvana, Thai G...	Chinese, North Indian, Thai	800
2	San Churro Cafe	Yes	3.8/5	918	Churros, Cannelloni, Minestrone Soup, Hot Choc...	Cafe, Mexican, Italian	800
3	Addhuri Udupi Bhojana	No	3.7/5	88	Masala Dosa	South Indian, North Indian	300
4	Grand Village	No	3.8/5	166	Panipuri, Gol Gappe	North Indian, Rajasthani	600

RESTAURANT RECOMMENDER SYSTEM

Preparing the data for the recommendation system

Identifying and formatting the unstructured data

cost_for_two
1,600
1,600
1,600
1,600
1,600

cost_for_two
1600.0
1600.0
1600.0
1600.0
1600.0

rate
4.1/5
4.1/5
3.8/5
3.7/5
3.8/5

rate
4.1
4.1
3.8
3.7
3.8

RESTAURANT RECOMMENDER SYSTEM

Preparing the data for the recommendation system

What about the missing data?

#Quick Question: What should we do with the restaurants for whom the either the rate, vote or the cost is missing?

Let's think in terms of business…

Missing values of the cost column can be replaced by the mean. However, can we do the same with rate and vote? Think, if the missing rate (which could have been 1 or 2) is replaced by the mean (which is 3.6), and recommend you that restaurant(may be that is your first date), what will you do???

Definitely throw tomatoes at us!!

The unrated restaurants are not taken into account.

RESTAURANT RECOMMENDER SYSTEM

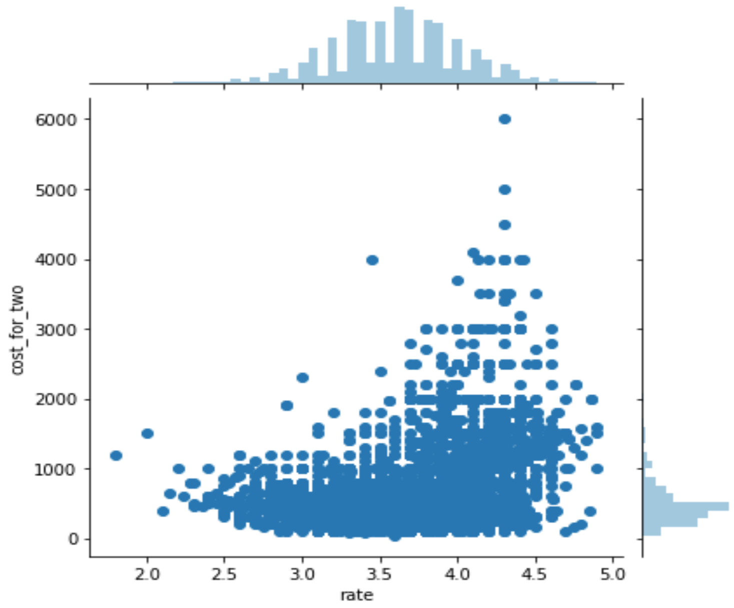
Information from different features

Does Costlier Means Better?

The jointplot below tells us that the most highly rated restaurants(above 4.5 ratings) are not the costliest ones.

In [204]:
```
sns.jointplot(x='rate', y='cost_for_two', data=dataframe)
```

Out[204]: <seaborn.axisgrid.JointGrid at 0x1b0bdc2048>



RESTAURANT RECOMMENDER SYSTEM

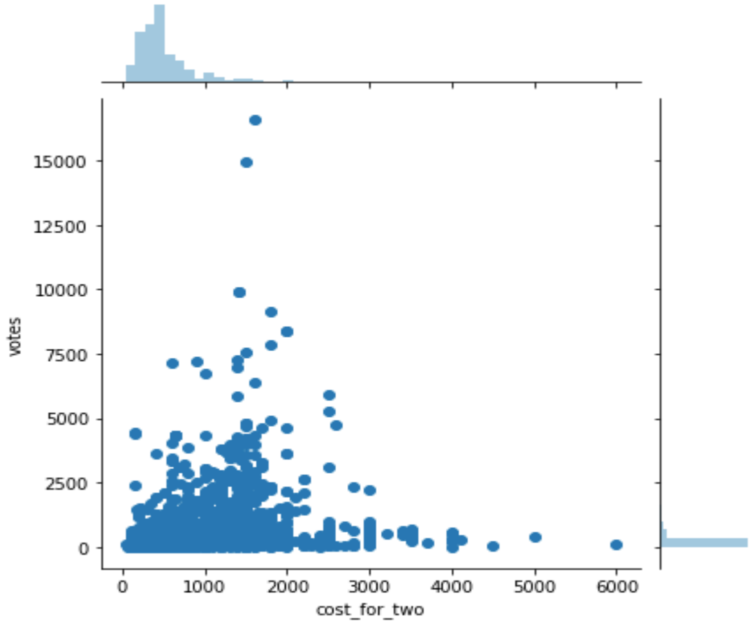
Information from different features

Does cheaper means more popular?

Here is a jointplot comparision which shows the city population is really cost conscious

In [205]:
```
sns.jointplot(x='cost_for_two', y='votes', data=dataframe)
```

Out[205]: <seaborn.axisgrid.JointGrid at 0x1b0b7824a8>



RESTAURANT RECOMMENDER SYSTEM

Developing the recommendation system

STEP-1: TOKENIZE THE CUISINES WITH TF-IDF

#Quick Question: Why TF-IDF, Why not Count-Vectorizer?

RESTAURANT RECOMMENDER SYSTEM

Developing the recommendation system

STEP 2: K-MEANS

The cluster size is 5

#How do you select the cluster size?

RESTAURANT RECOMMENDER SYSTEM

Developing the recommendation system

STEP 2: K-MEANS

The cluster size is 5

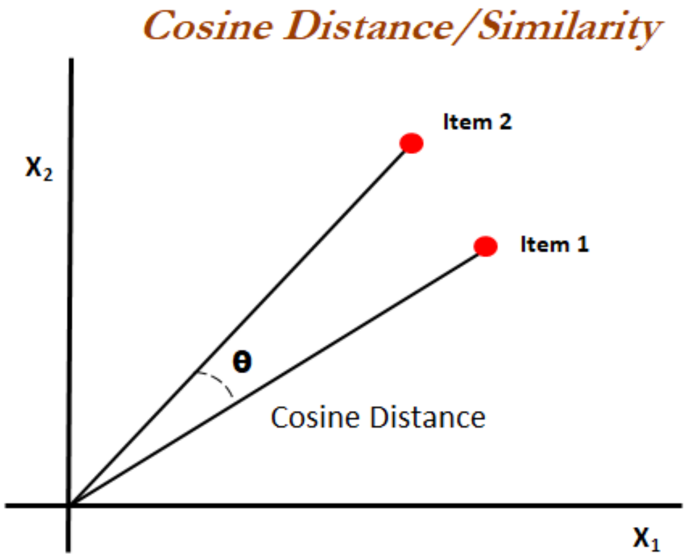
#How do you select the cluster size?

RESTAURANT RECOMMENDER SYSTEM

Developing the recommendation system

STEP 3: Cosine Similarity

What? Why?



RESTAURANT RECOMMENDER SYSTEM

Developing the recommendation system

STEP 3:Results

id							
0	Byg Brewski Brewing Company	1	4.900000	16588.500000	['Continental', ' North Indian', ' Italian', '...	1600.000000	26541600
1	SantÃÂÂÂÂÂÂÂÂÂÂÂÂÂÂÂÂÂÂÂÂÂÂÂ...	0	4.900000	246.000000	['Healthy Food', ' Salad', ' Mediterranean']	1000.000000	246000
2	Asia Kitchen By Mainland China	1	4.900000	2223.727273	['Asian', ' Chinese', ' Thai', ' Momos']	1500.000000	3335590
3	Punjab Grill	1	4.866667	1286.666667	['North Indian']	2000.000000	2573333
4	Belgian Waffle Factory	1	4.850000	890.785714	['Desserts']	400.000000	356314
5	Flechazo	0	4.833333	4301.000000	['Asian', ' Mediterranean', ' North Indian', '...	1400.000000	6021400
6	The Pizza Bakery	1	4.800000	1763.333333	['Italian', ' Pizza', ' Beverages']	1200.000000	2116000
7	O.G. Variar & Sons	0	4.800000	1158.500000	['Bakery', ' Desserts']	200.000000	231700
8	AB's - Absolute Barbecues	0	4.790909	4069.250000	['European', ' Mediterranean', ' North Indian'...	1563.636364	6362827
9	Biergarten	0	4.766667	2639.111111	['Continental', ' European', ' BBQ', ' Chinese...	2200.000000	5806044

```
In [231]: rest_recommendations('O.G. Variar & Sons').head(4)

Out[231]: id
3240       Lassi Darbar
3239     Karachi Bakery
3238      Paratha Plaza
3237           Calvin's
Name: name, dtype: object
```

RESTAURANT RECOMMENDER SYSTEM

Developing the recommendation system

STEP 3:Results

```
3240      Lassi Darbar
3239    Karachi Bakery
3238     Paratha Plaza
3237          Calvin's
```

Why this order?

5059	Karachi Bakery	1	3.624325	22.384615	['Bakery', ' Desserts']	423.076923	9470
5050	Calvin's	0	3.628748	240.181818	['Desserts', ' Italian', ' Pizza']	763.636364	183411

OUTLIERS

Choice between Supervised or Unsupervised.

Dataset is basically unlabeled (there are no labels)

OUTLIERS

Isolation Forest

K Nearest Neighbors (KNN)

Histogram-base Outlier Detection (HBOS)

Angle-based Outlier Detector (ABOD)

OUTLIERS

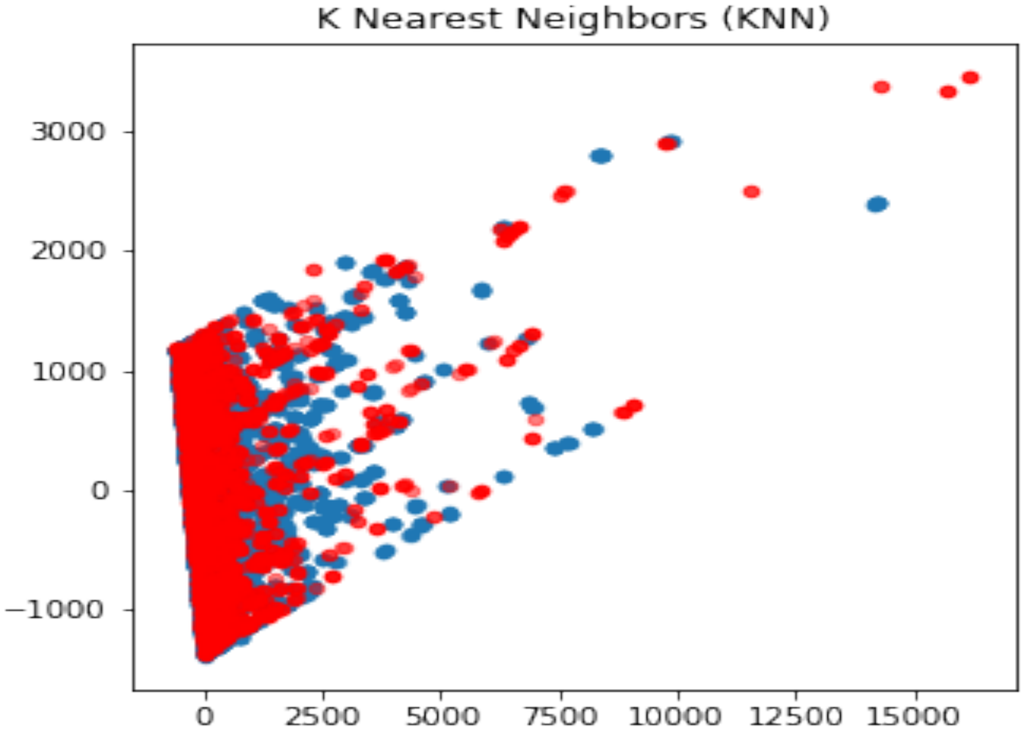
Feature Extraction

Reduce the dimension of feature space.

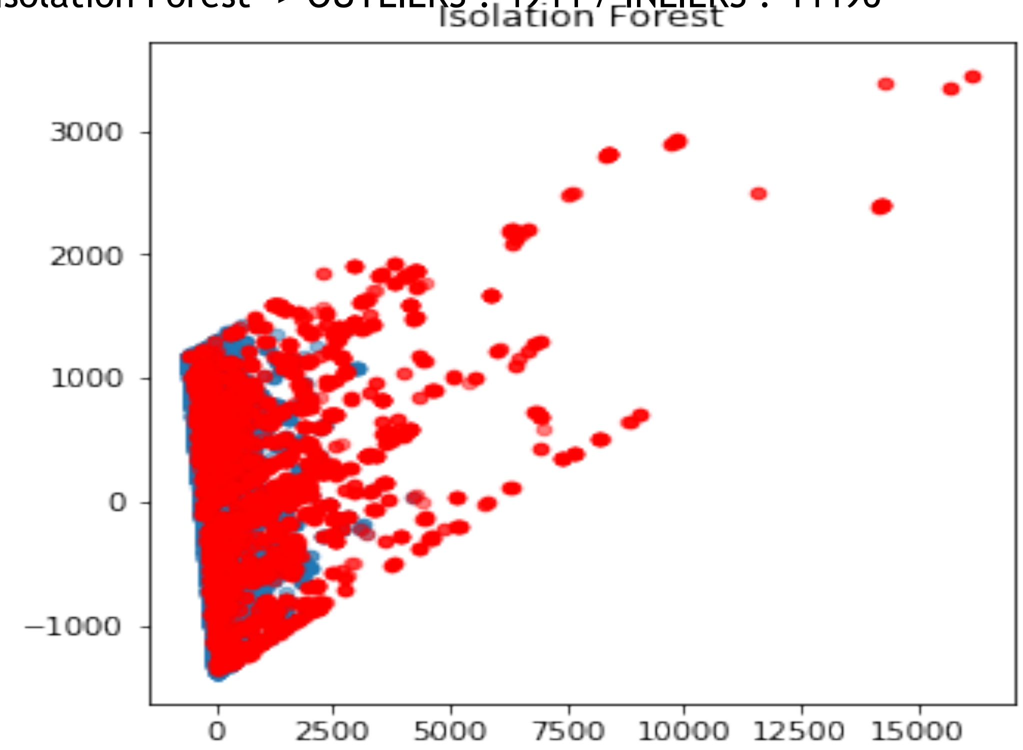
Principle component analysis – From ‘n’ Dimension to 2D

OUTLIERS

K Nearest Neighbors (KNN) -> OUTLIERS : 3706 / INLIERS : 45734

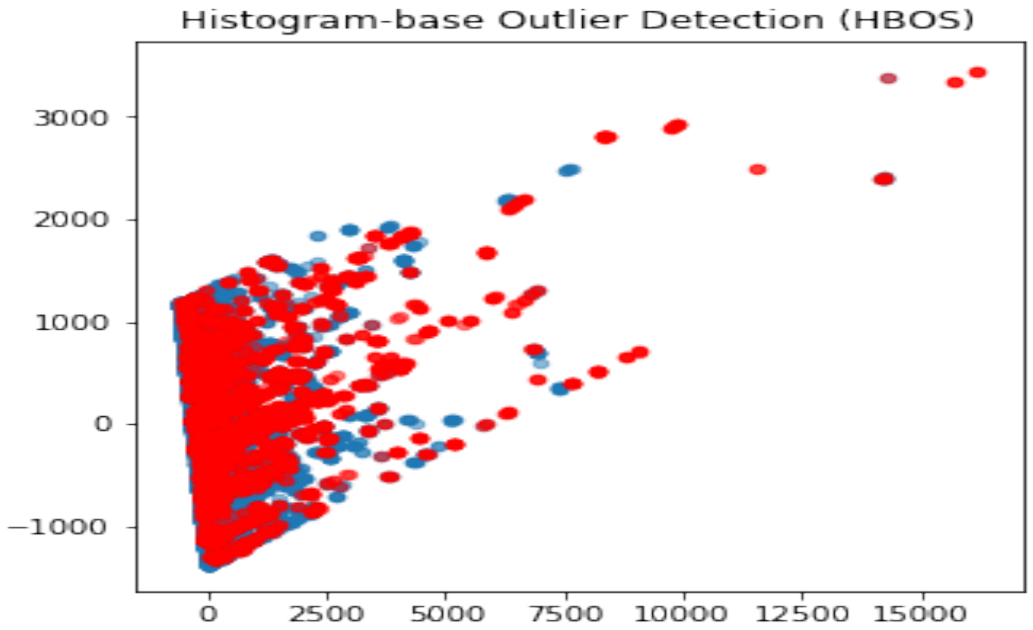


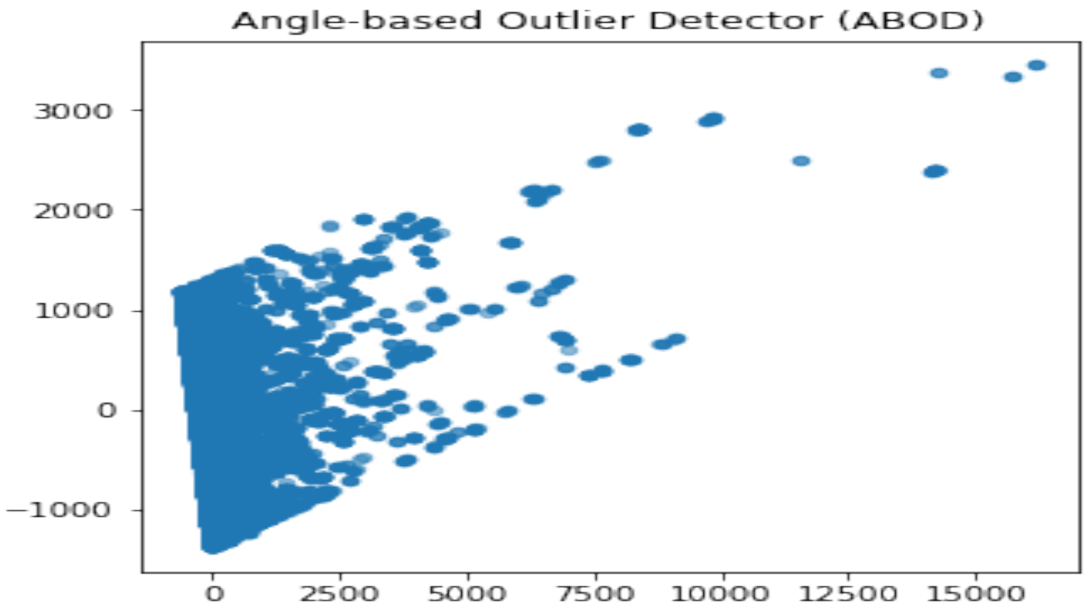
Isolation Forest -> OUTLIERS : 4944 / INLIERS : 44496



OUTLIERS

Histogram-base Outlier Detection (HBOS) -> OUTLIERS : 4939 / INLIERS : 44501 Angle-based Outlier Detector (ABOD) -> OUTLIERS : 0 / INLIERS : 49440

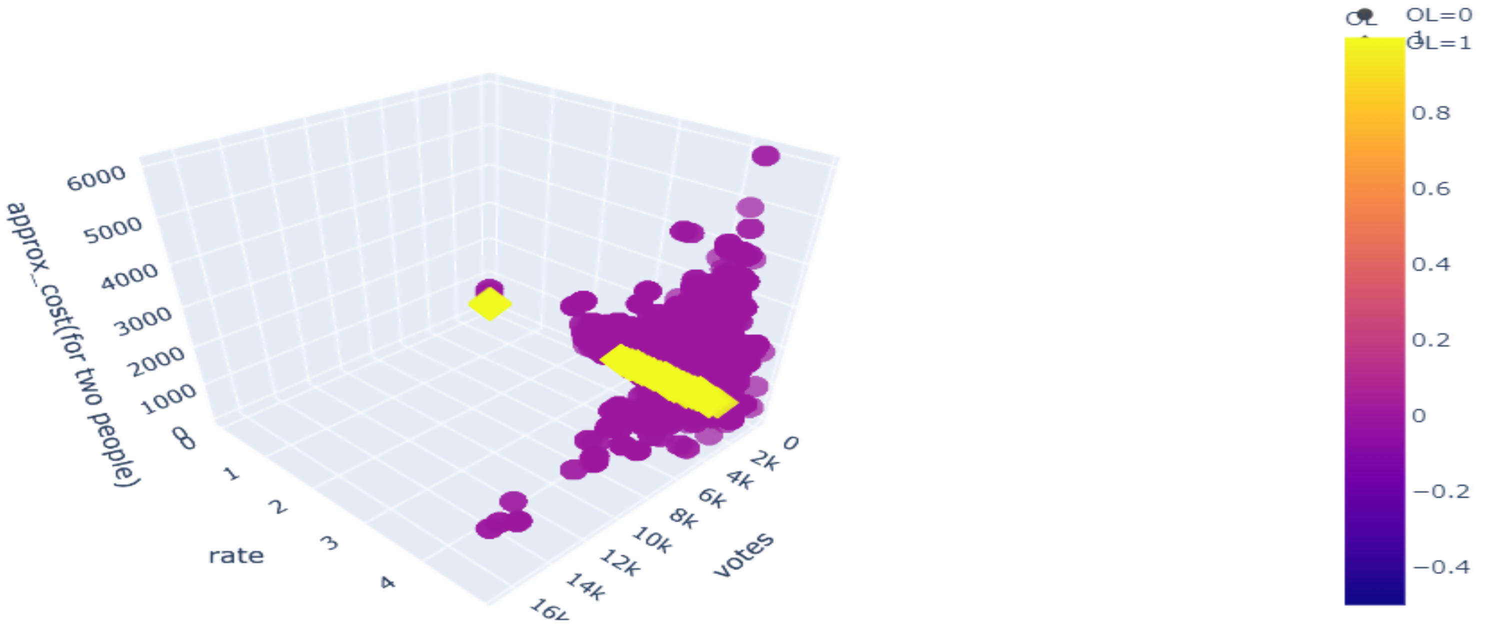




OUTLIERS

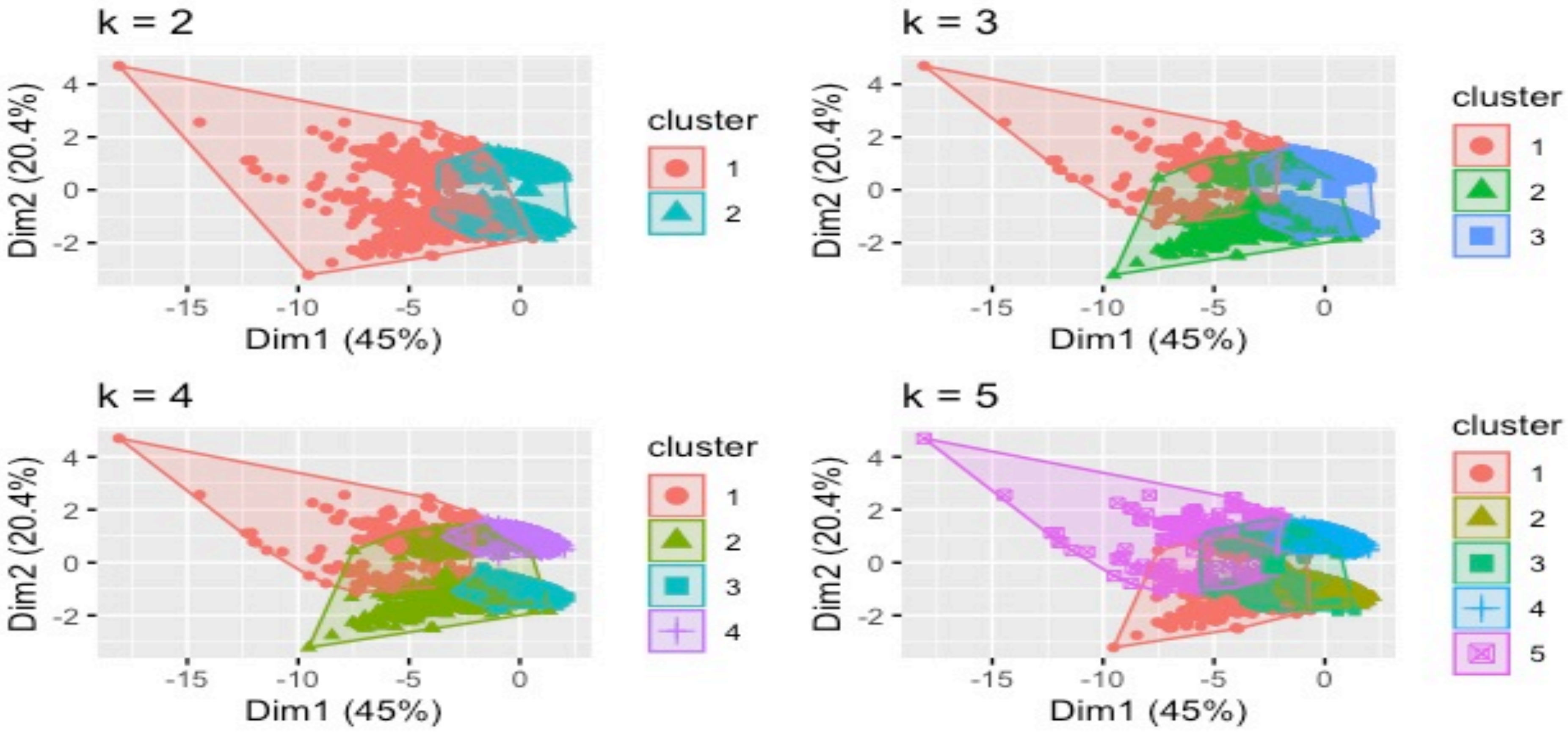
Basically a type of Support Vector Machines Algorithm

3D view to show Outliers w.r.t Cost, Rating & Vote

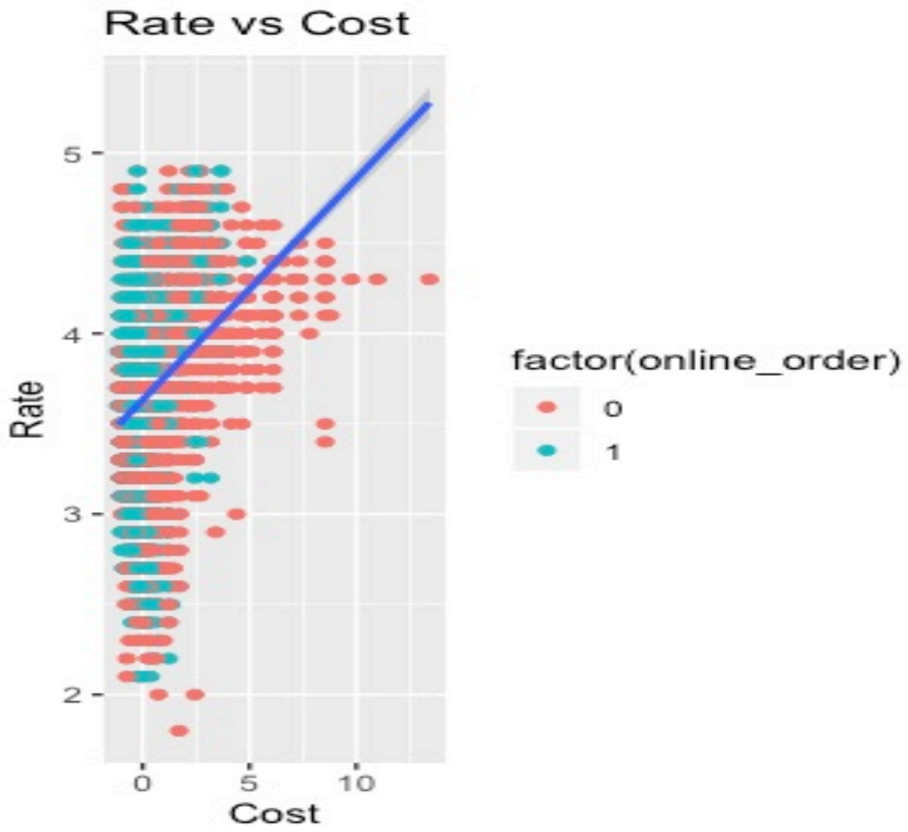


CLUSTERING

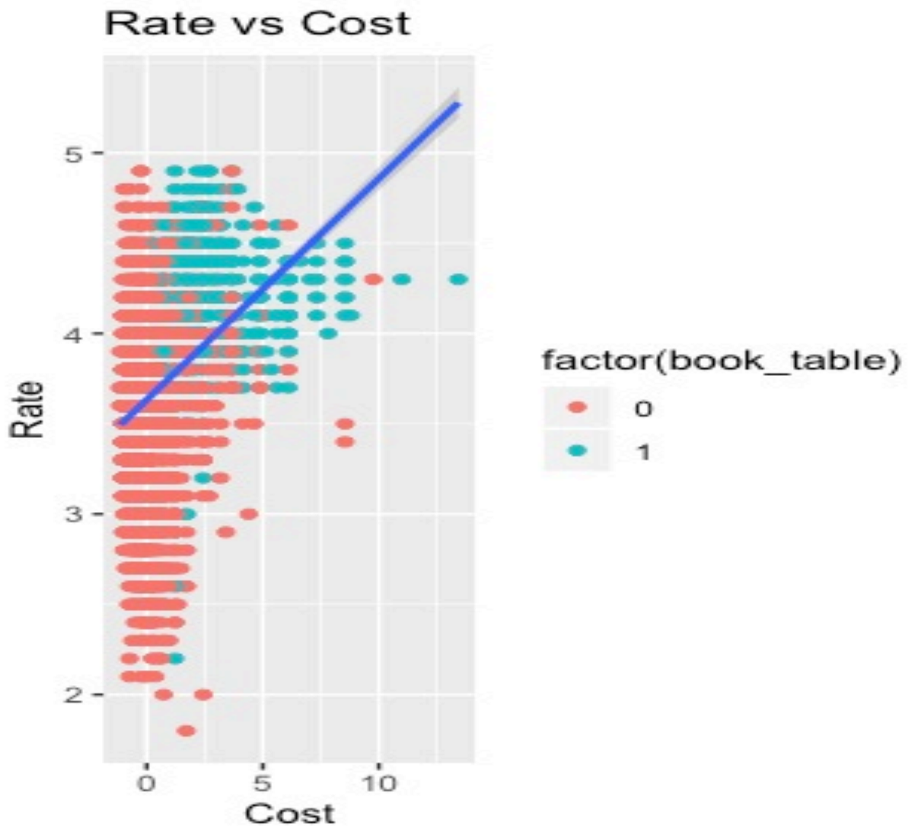
Analysis



REGRESSION

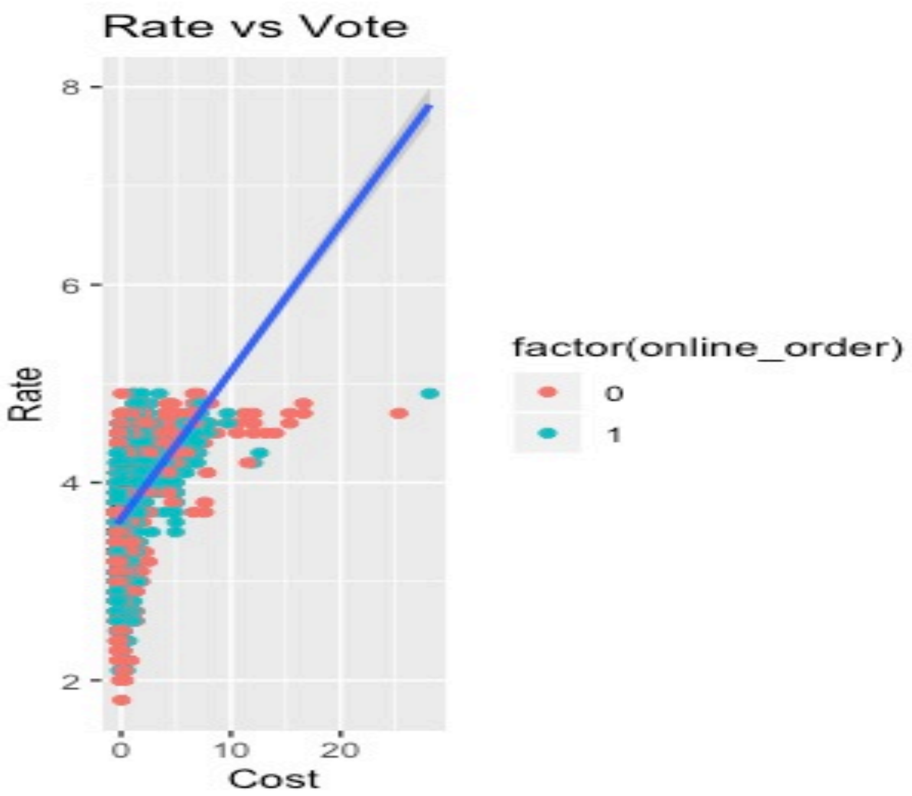


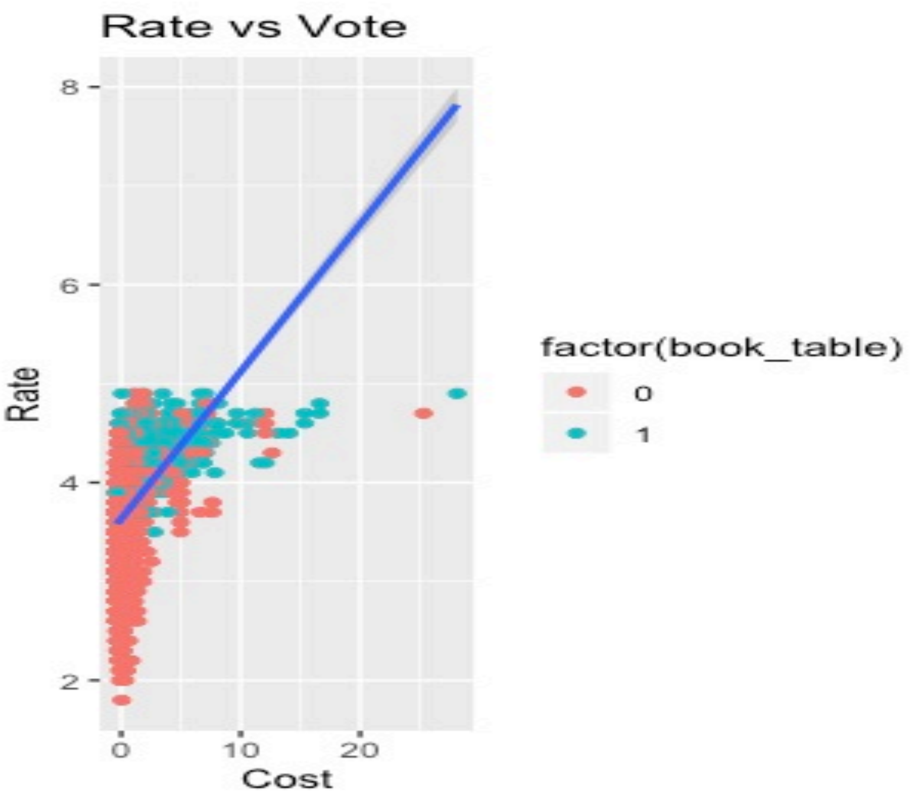
Rate vs Cost
Rate
Cost
5
4
3
2
0
5
10
factor(online_order)
0
1



Rate vs Cost
Rate
Cost
5
4
3
2
0
5
10
factor(book_table)
0
1

REGRESSION


Rate vs Vote
8
6
4
2
Rate
0
10
20
Cost
factor(online_order)
0
1


Rate vs Vote
8
6
4
2
Rate
0
10
20
Cost
factor(book_table)
0
1

LEARNING OUTCOMES

EECS-731

- Understood the **complications** involved in working on *datasets and making hypothesis.*
  - Spent **multiple weeks** on finding / exploring and (ultimately) discarding multiple datasets related to *restaurants and analysis* from a number of online sources
  - More clarity on how to collect **quality data** when working with *classifications, regression and recommendations.*
    - More finely tuned data -> better predicting models
  - Better understanding of the nuances involved in using these algorithms for answering *practical questions*