STATISTICS FOR DATA SCIENCE
UE19CS203
PROJECT

RESTAURANTS OF BANGALORE

Team Members:

1. Sai Ramya Paturi — PES1UG19CS419
2. Sanjana B Atrey — PES1UG19CS429
3. Sheetal S — PES1UG19CS455
4. Shreyas Devaraj — PES1UG19CS466

Section: G

Faculty: Dr Uma D

RESTAURANTS OF BANGALORE

PES UNIVERSITY

- This project is focussed on the diverse food culture of Bangalore.
  With approximately 12,000 restaurants serving cuisines from all over India and abroad, it truly is a foodie's paradise.
- New eateries are coming up every day and face challenges such as high real estate costs, rising food costs, shortage of quality manpower, fragmented supply chain, over-licensing and stiff competition from established restaurants.
- Through this project, we hope to study and analyse the local demographic preferences that will help restaurants decide their theme, menu, cuisine and cost to cater to a wide range of customers.

DATASET

Title: **Zomato Bangalore Restaurants**

Source: **Kaggle**

https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants

Zomato is an Indian restaurant aggregator and food delivery start-up. It provides information, menus and user-reviews of restaurants as well as food delivery options from partner restaurants in select cities.

The data was scraped for educational purposes from the Zomato website (https://www.zomato.com/bangalore/restaurants) and is accurate until 15 March 2019.

A Snippet of the Dataset

address	name	online_order	book_table	rate	votes	phone	location	rest_type	dish_liked	cuisines	approx_co...	reviews_list	menu_item	listed_in(t...
2/1, 7th Main, Dwarakangar, Hosakeregalli, Banashankari, Bangalore	Foodiction	Yes	No	2.8/5	506	+91 9916107070	Banashankari	Quick Bites	Burgers, Lassi, Chicken Grill, Naan, Momos, Chicken Burger, Biryani	North Indian, Fast Food, Chinese, Burger	500	[('Rated 1.0', "RATED\n Worst restaurant ever , Veg Manchurian worth ?130 doesn't even fill a tiny ...	[]	Delivery
101 Ground Floor, Manjunatha Complex, 22nd Main Road, Banashankari, Bangalore	Sweet Truth	Yes	No	3.9/5	35	+91 7710055553	Banashankari	Delivery		Bakery, Desserts	500	[('Rated 4.0', 'RATED\n I had the opportunity to try the desserts and I was very	['Chocolate Fantasy (Pack Of 5)', 'Pan Cake (Pack Of 6)', 'Gulab Jamun (Pack Of	Delivery
101, Ground Floor, Manjunatha Complex, 22nd Main Road, 2nd Stage, Banashankari, Bangalore	Ovenstory Pizza	Yes	No	3.9/5	172	+91 7738								
80, BDA Complex, 2nd Stage, Banashankari, Bangalore	Faasos	Yes	No	4.2/5	415	+91 7708								
101, Ground Floor, Manjunatha Complex, 22nd Main Road, 2nd Stage, Banashankari, Bangalore	Behrouz Biryani	Yes	No	3.9/5	230	+91 7022 +91 7022								
32, 7th Main Road, Dwaraka Nagar,Near PES College, Banashankari, Bangalore	Fast And Fresh	Yes	No	2.8/5	91	+91 9742								

	name	online_order	book_table	rate	votes	location	rest_type	dish_liked	cuisines	cost
0	Jalsa	True	True	4.1	775	Banashankari	Casual Dining	pasta, lunch buffet, masala papad, paneer laja...	North Indian, Mughlai, Chinese	800.0
1	Spice Elephant	True	False	4.1	787	Banashankari	Casual Dining	momos, lunch buffet, chocolate nirvana, thai g...	Chinese, North Indian, Thai	800.0
2	San Churro Cafe	True	False	3.8	918	Banashankari	Cafe, Casual Dining	churros, cannelloni, minestrone soup, hot choc...	Cafe, Mexican, Italian	800.0
3	Addhuri Udupi Bhojana	False	False	3.7	88	Banashankari	Quick Bites	masala dosa	South Indian, North Indian	300.0
4	Grand Village	False	False	3.8	166	Basavanagudi	Casual Dining	panipuri, gol gappe	North Indian, Rajasthani	600.0

Columns Description

#	Column	Type	Description
1	url	categorical	url of the restaurant in the website
2	address	categorical	address of the restaurant in Bengaluru
3	name	categorical	name of the restaurant
4	online_order	binary	online ordering is possible or not
5	book_table	binary	table booking option available or not
6	rate	continuous numerical	overall rating of the restaurant out of 5
7	votes	discrete numerical	total number of ratings provided
8	phone	categorical	phone number of the restaurant
9	location	categorical	neighborhood in which it is located
10	rest_type	categorical	restaurant type
11	dish_liked	categorical	dishes people liked in the restaurant
12	cuisines	categorical	food styles, separated by comma
13	approx_cost(for two people)	discrete numerical	approximate cost of a meal for two people
14	reviews_list	categorical	list containing customer reviews
15	menu_item	categorical	list of items served in the restaurant
16	listed_in(type)	categorical	type of meal
17	listed_in(city)	categorical	neighborhood in which it is listed

DATA CLEANING

01	Duplicate Data	• Removed columns with duplicate observations.
02	Missing Data	• Extracted rate and dish_liked from reviews_list. • Replaced NaN values with default values.
03	Unwanted Observations	• Deleted unnecessary/repetitive columns such as address, url, phone number.
04	Typos & Inconsistent Capitalization	• Changed all text to lowercase. • Stripped special symbols from the numeric data.
05	Column Name and Type	• Renamed columns with appropriate titles. • Converted data types from object to bool, int and float.

#	Column	%Null	Dtype
0	url	0.00	object
1	address	0.00	object
2	name	0.00	object
3	online_order	0.00	object
4	book_table	0.00	object
5	rate	15.03	object
6	votes	0.00	int64
7	phone	2.34	object
8	location	0.04	object
9	rest_type	0.44	object
10	dish_liked	54.29	object
11	cuisines	0.09	object
12	approx_cost(for two people)	0.67	object
13	reviews_list	0.00	object
14	menu_item	0.00	object
15	listed_in(type)	0.00	object
16	listed_in(city)	0.00	object

Observations: 51717 Variables: 17

BEFORE

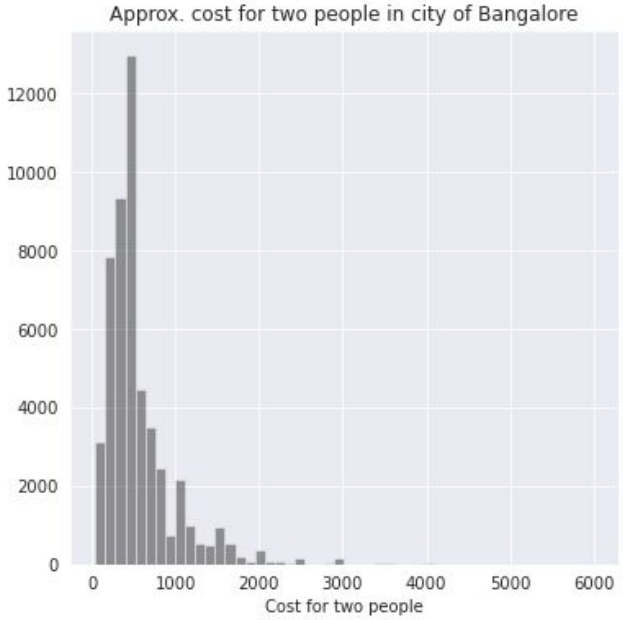
#	Column	Count	Dtype
1	name	42741	object
2	online_order	42741	bool
3	book_table	42741	bool
4	rate	42741	float64
5	votes	42741	int64
6	location	42741	object
7	rest_type	42741	object
8	dish_liked	42741	object
9	cuisines	42741	object
10	cost	42741	float64
11	menu_item	42741	object
12	meal_type	42741	object
13	reviews_text	42741	object

Observations: 42741 Variables: 13

AFTER

PES UNIVERSITY

VISUALIZATION AND INSIGHTS

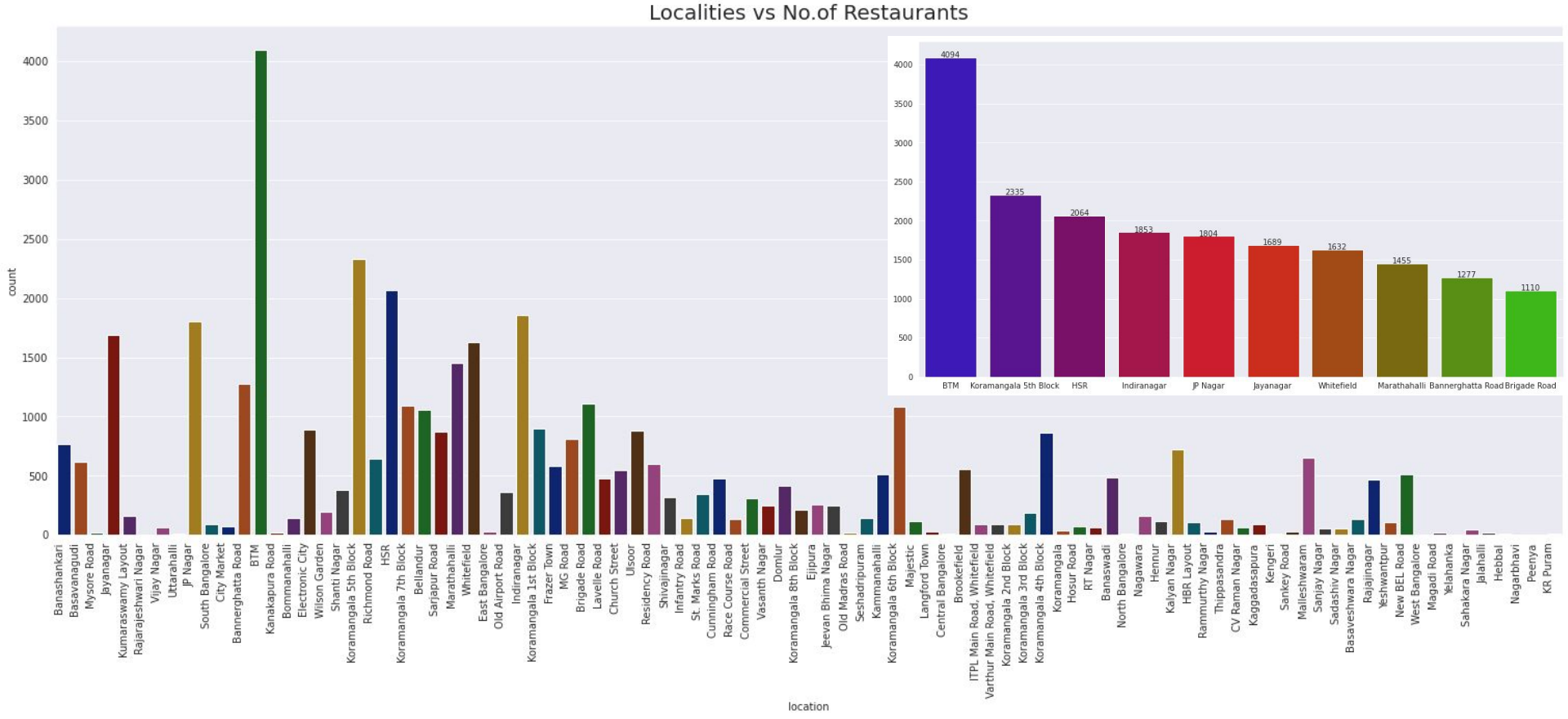


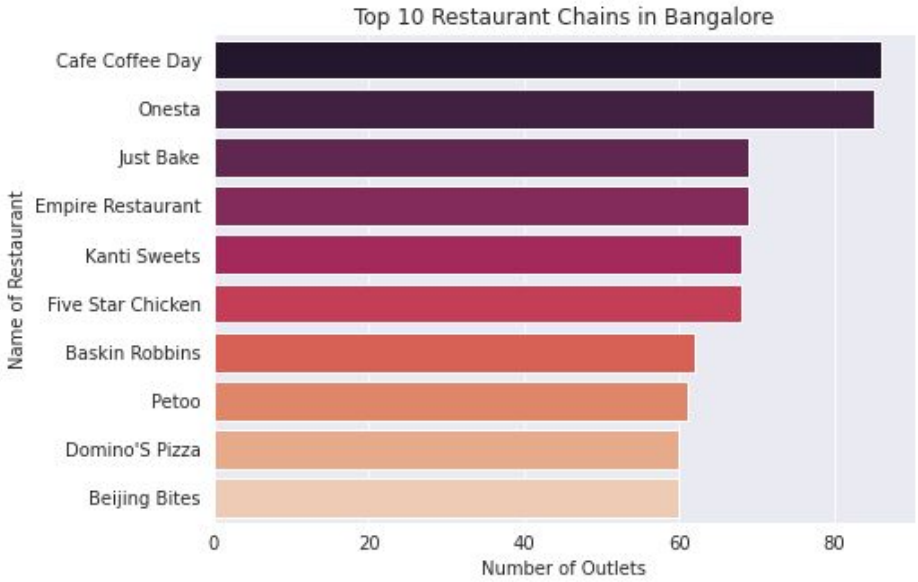
Positively skewed: Distribution has a large number of values on the lower side of x-axis. Most costs lie in range Rs. 300-600, outliers exist beyond Rs. 2000.

Negatively skewed: Distribution has a large number of values on the upper side of x-axis.

Most restaurants are rated 3-4.

With a large no.of restaurants, commercial hubs like BTM Layout, Koramangala and JP Nagar are the thriving food hotspots of the IT Capital with a plethora of options for everyone's taste buds, while it is not so on the outskirts, industrial areas.



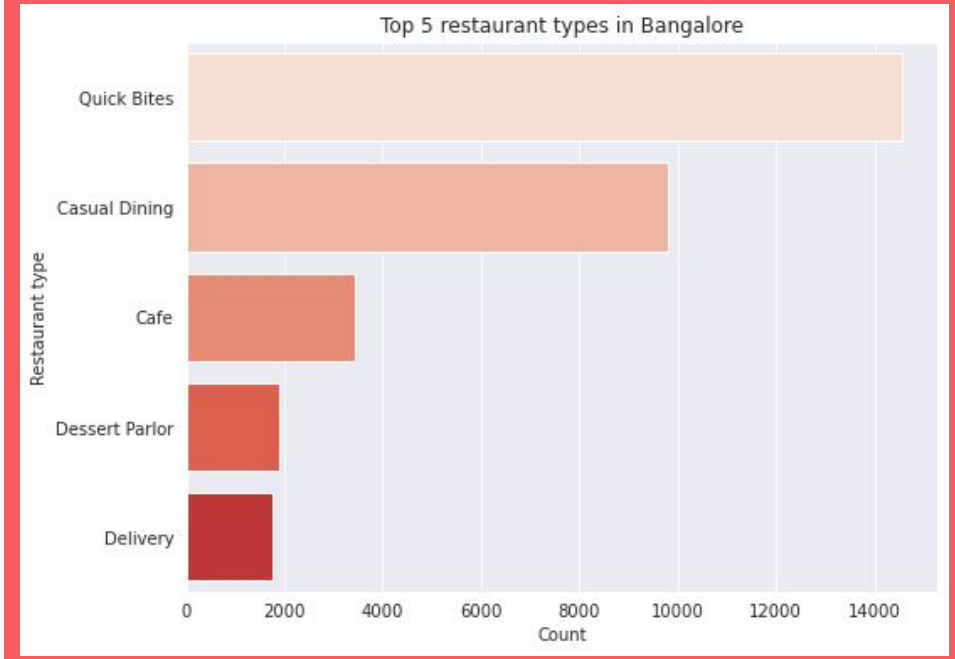


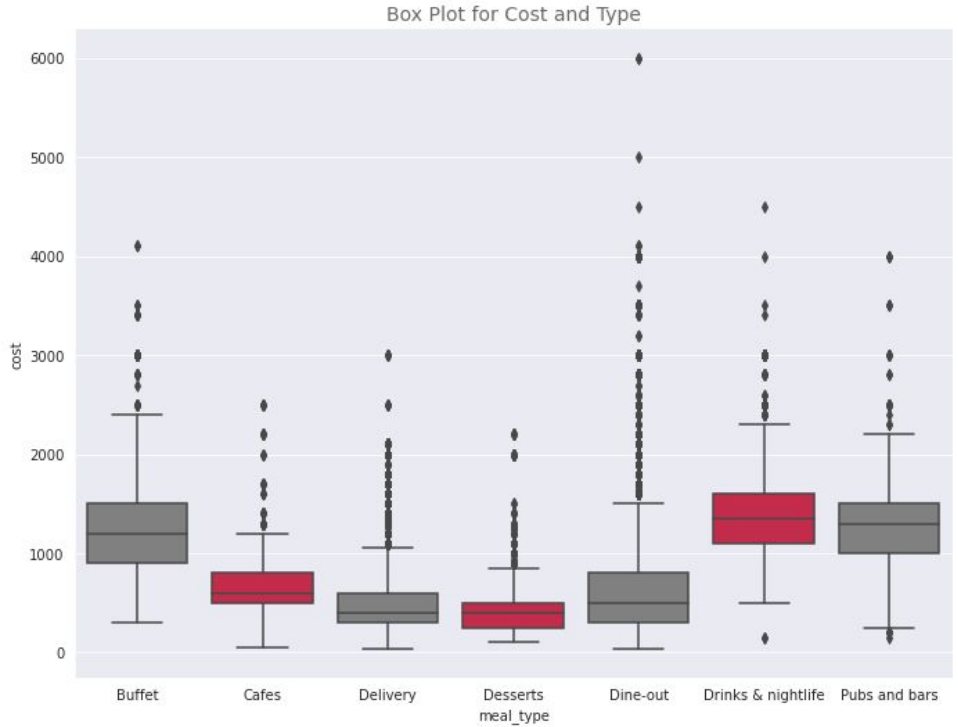
Cafe Coffee Day has the most no.of outlets in the city.

Others in the top 10 include pizzerias, non-veg speciality restaurants and dessert shops.
This goes on to show that the people also have an adventurous taste to complement their love for the classic coffee.

The most popular restaurant type is Quick Bites.

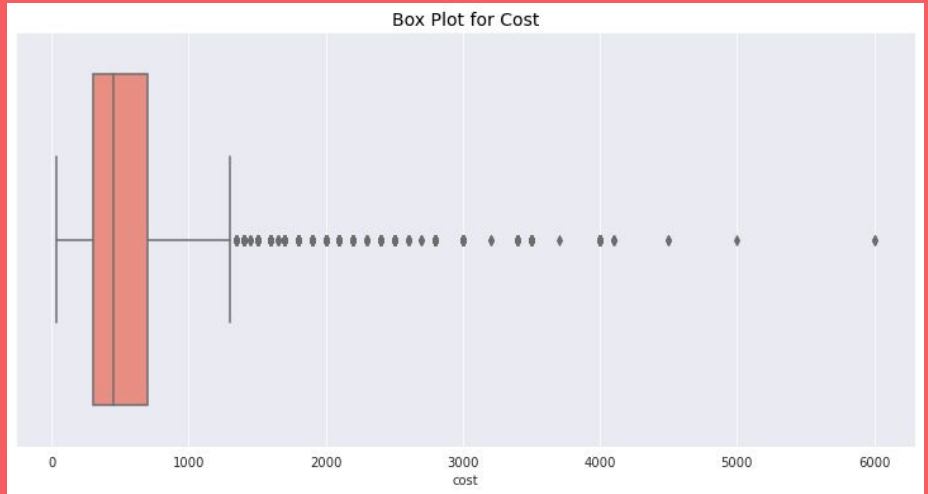
Mostly young and working professionals, Bangaloreans are always on the go and prefer snacks over heavy, extravagant meals.





Value	Cost
Min	40.00
Q1	300.00
Median	450.00
Q3	700.00
Max	6000.00

There exist non-anomalous outliers since fancy 5-star restaurants and those serving authentic international cuisines are extremely high-priced. Hence, they aren't filtered out.



Buffets and pubs are different from cafes, dine-outs, delivery and desserts and also have greater variability.

Dine-out has options for all budget ranges.

NORMALIZATION

Often used interchangeably, normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

WHAT IS NORMALIZATION?

Normalization is a technique often applied as part of data preparation for machine learning. Similarly, the goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, normalization is required only when features have different ranges.

WHAT ARE THE BENEFITS?

There are two primary advantages of having a highly normalized data schema:

- Increased consistency : Information is stored in one place and one place only, reducing the possibility of inconsistent data.
- Variables that are measured at the same scales contribute equally to the analysis and help avoid bias, the result is not intrinsically influenced by variability in the values.

MEASUREMENTS

From the previous graphs, it can be observed that the data is not normalized as there is skewness.

Hence, the numeric columns - cost, rate and votes were normalized.

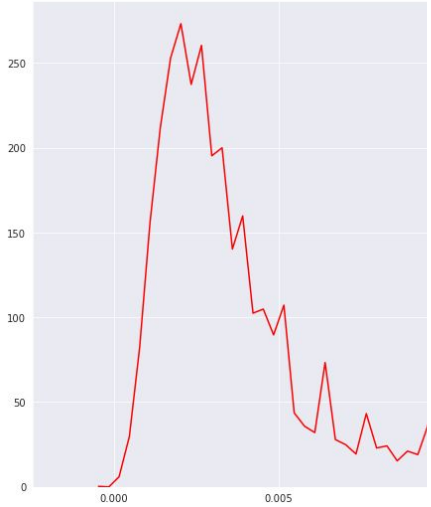
Before normalization:

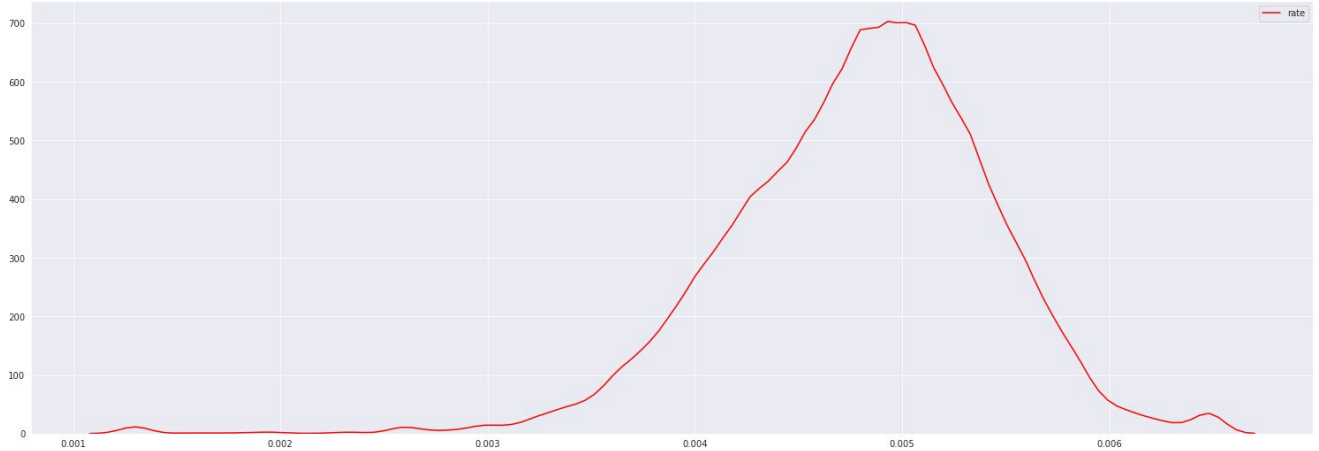
	rate	votes	cost
count	42741.000000	42741.000000	42741.000000
mean	3.698334	340.938490	597.772771
std	0.486326	871.364993	460.918542

After normalization:

	rate	votes	cost
count	42741.000000	42741.000000	42741.000000
mean	0.004796	0.001762	0.003831
std	0.000631	0.004505	0.002954

PLOTS FOR NORMALIZED DATA





After normalizing the dataset, we obtain a bell curve plot which depicts a normal distribution.

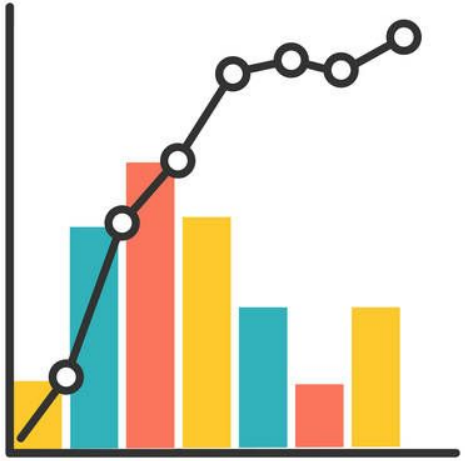
HYPOTHESIS TESTING

- We performed hypothesis testing on cost and rate.
- The test performed was to check if the rating of restaurant is 3.5 when the mean cost is 600.
- Considering a sample where rating = 3.5 yields a mean cost, x_bar = 600. We are testing if this hypothesis is from our dataset using z-test. First we state the null hypothesis and alternative hypothesis like this:
- H0: The sample is from the zomato restaurants, x_bar = μ.
- HA: The sample is not from the zomato restaurants, x_bar != μ.
- We used a two tail test where the confidence interval was 95% and the significance value was 0.05.
- z_critical = 1.96 *# alpha level of 0.05 and two-tailed test*
- **Since z_stat = 0.009 is less than z_critical, we accept the null hypothesis.**

CORRELATION

- We use this to evaluate the strength of relationship between two variables.

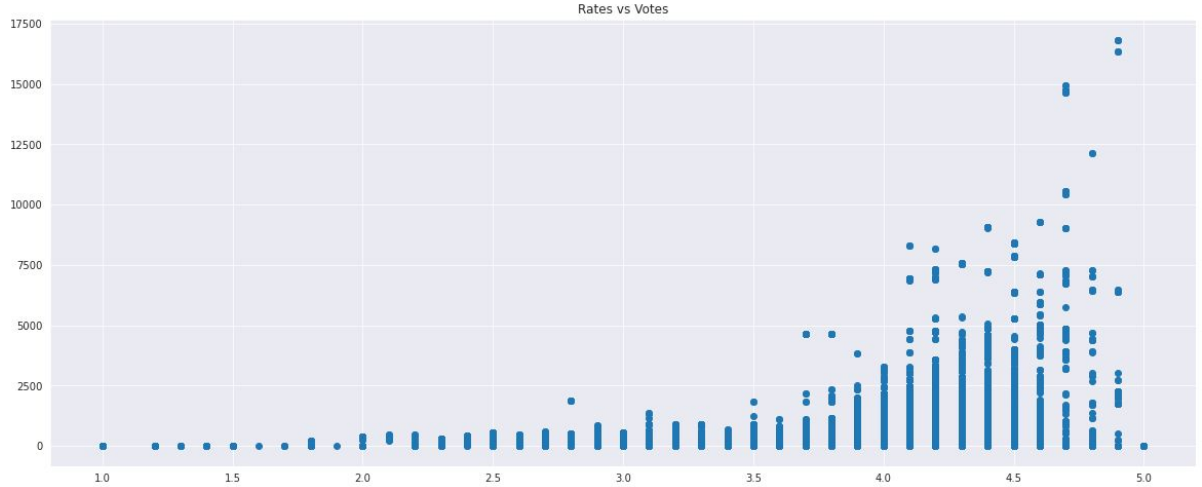
- Our objectives :
  - Find variables with correlation
  - Find the correlation coefficient
  - Obtain a scatter plot and heat map to visually analyze the same

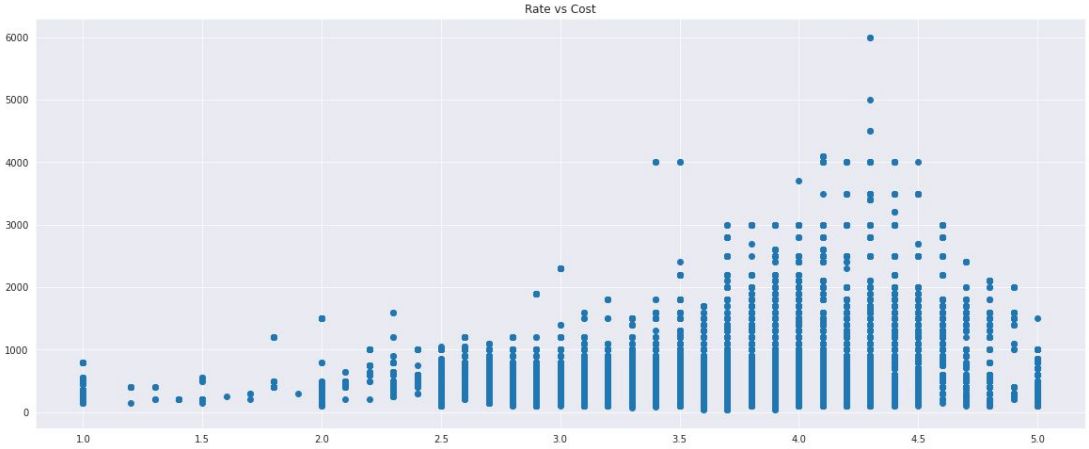


POSITIVE CORRELATION

- Consider Rates vs Votes,
  The coefficient of correlation = 0.38889696, giving a good positive correlation.

  We then observe the scatter plot :





- Consider the above two plots,
  As the rating of the restaurants increases, the cost/votes also increases.

NEGATIVE CORRELATION

- Consider relation between Cost with Online_Order,
  The coefficient of correlation = -0.12131347, giving a negative correlation.

HEATMAP

- The following heatmap explains the correlation between different variables from our dataset.
- The lighter ones depict the variables that are positively related.
- The darker ones are the negatively related.
- The neutral ones show weak or no correlation.

THANK YOU!