DSc Project

-By Ashwat Kumar and Shaurya Goel
https://github.com/shauryagoel/Data_Science_Project

Problem statement-

Analyse the Zomato Dataset to get an idea about the factors affecting the establishment of various types of restaurants at different places in **Bengaluru**, along with ratings, votes and reviews.
We also made predictive models about various parameters of restaurant.
We are also suggesting a user how much revenue he/she can get by opening a restaurant with given parameters.

Data acquisition-

We used Zomato datasets available on kaggle. There were two datasets which we combined together to get our final dataset.

In both datasets, we took common samples of restaurants in Bengaluru only and merged the two datasets.

	name	address	location	rest_type	cuisines	cost(per person)	online_order	book_table	rating	votes	revenue	reviews_list
	Jalsa	942, 21st Main Road, 2nd Stage, Banashankari, ...	Banashankari	Buffet	North Indian, Mughlai, Chinese	400	1	1	4.1	775	310000	[('Rated 4.0', 'RATED\n A beautiful place to ...
	Spice Elephant	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Banashankari	Buffet	Chinese, North Indian, Thai	400	1	0	4.1	787	314800	[('Rated 4.0', 'RATED\n Had been here for din...
	San Churro Cafe	1112, Next to KIMS Medical College, 17th Cross...	Banashankari	Buffet	Cafe, Mexican, Italian	400	1	0	3.8	918	367200	[('Rated 3.0', "RATED\n Ambience is not that ...
	Addhuri Udupi Bhojana	1st Floor, Annakuteera, 3rd Stage, Banashankar...	Banashankari	Buffet	South Indian, North Indian	150	0	0	3.7	88	13200	[('Rated 4.0', "RATED\n Great food and proper...
	Grand Village	10, 3rd Floor, Lakshmi Associates, Gandhi Baza...	Basavanagudi	Buffet	North Indian, Rajasthani	300	0	0	3.8	166	49800	[('Rated 4.0', 'RATED\n Very good restaurant ...

Data references-

1) https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants
2) https://www.kaggle.com/shrutimehta/zomato-restaurants-data

Final dataset-
https://drive.google.com/open?id=1oZR5CVixCDSUVG1rRex-DJxu34bqN7MM

Predictive Modeling

1) - Restaurant Rating Prediction

- Rating prediction of a restaurant given all other values such as cost per person, location, online_order, cuisines etc.
- Applied “Logit” Link Function to predict ‘rating’ between range 0 to 5.

Algorithm Used	**Train MAE**	**Test MAE**	**Train MSE**	**Test MSE**
Linear Regression	0.3045	0.3068	0.1431	0.1482
Lasso Regression	0.3138	0.3166	0.1500	0.1552
Ridge Regression	0.3045	0.3068	0.1431	0.1482
Random Forest Regressor	0.1365	0.1654	0.0243	0.0434

2) - Review Rating Prediction

-Did feature extraction from reviews using **BERT** (Bidirectional Encoder

Representations from Transformers).

-Trained various classifiers to predict **'rating'** of the review.

Algorithm Used	**Train Accuracy**	**Test Accuracy**
Random Forest	0.978	0.866
Decision Tree	0.977	0.809
Logistic Regression	0.77	0.695

3) - Revenue Prediction

- Given parameters like 'location', 'restaurant type', 'online order', 'booking table' etc. we predicted overall revenue of restaurant.

- Random Forest Regressor gave least error, hence we used this model for revenue prediction.

Algorithm Used	**Test MAE**	**Test RMSE**
Linear Regression	223771.3	530973.68
Ridge Regression	223766.46	530975.24
Lasso Regression	223771.08	530974.06
Random Forest Regressor	101921.13	351321.43

Hypothesis :-

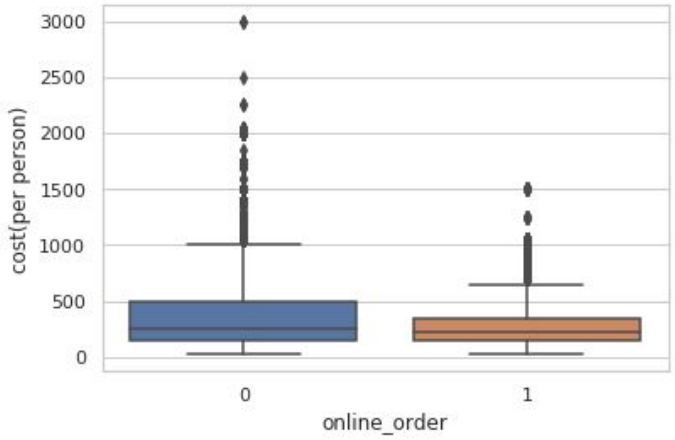
Hypothesis 1 - Does 'revenue' increase when we switch from 'offline' to

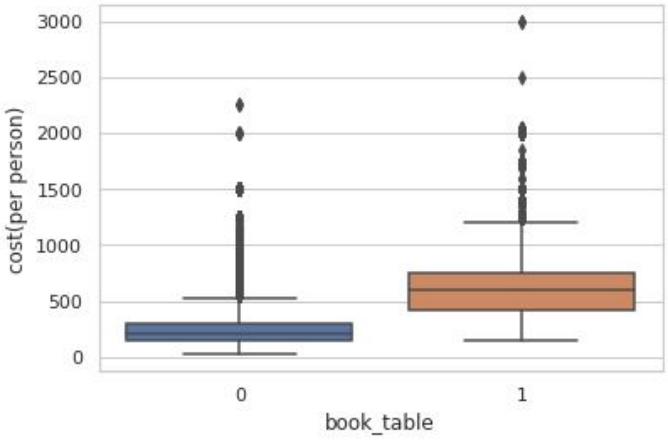
‘online’?

We used revenue predictive model to answer this.
We calculated difference between online order revenue and offline orders revenue. We saw,

‘9265.04’ - Difference between Online and Non-Online revenue for originally Non-Online.
This shows that revenue does increase when you switch to online orders.

‘-17362.65’ - Difference between Online and Non-Online revenue for originally Online.
This shows that revenue decreases when you switch to offline, your revenue increases too.





Conclusion - Hence, we **can’t be sure** whether the switch to take online orders will necessarily increase revenue. This is confirmed by above left graph.

Hypothesis 2 - Does ‘revenue’ increase when we switch from ‘not

booking tables’ to ‘booking tables’?

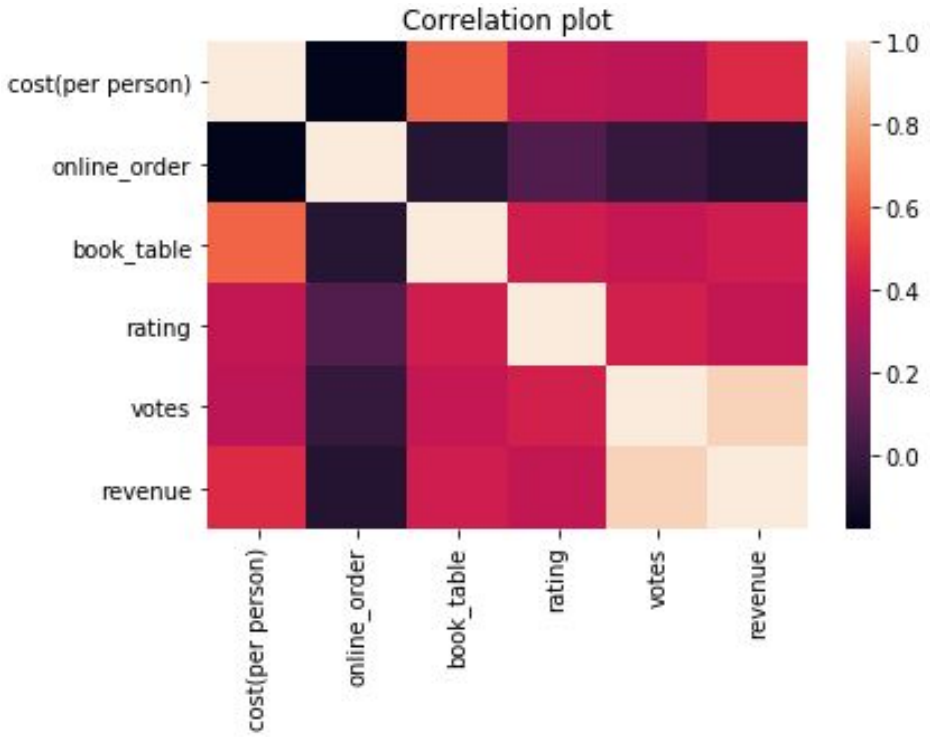
We used revenue predictive model to answer this.
We calculated difference between booking table revenue and not booking table revenue. We saw,

‘172646.94’ - Diff btw Booking Table and Non-Booking Table for originally Non-Booking Table
This shows that revenue does increase when you start doing booking table.

‘308789.35’ - Diff btw Booking Table and Non-Booking Table for originally Booking Table
This shows that revenue does increase when you start doing booking table.

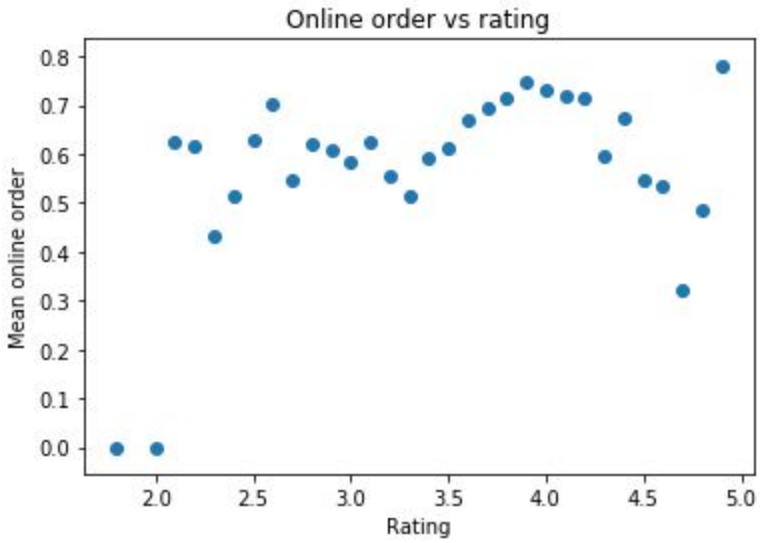
Conclusion - Hence, we **CAN BE** sure switching to doing booking table increases revenue. This is confirmed by above right graph.

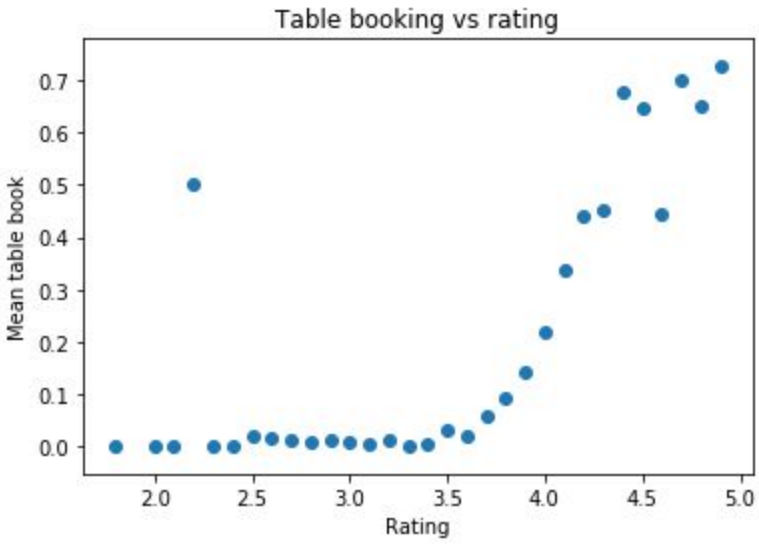
Hypothesis 3- Revenue should be highly correlated with votes and cost per person.



Conclusion- Votes is highly correlated with revenue but cost per person is only slightly correlated.

Hypothesis 4- Highly rated restaurants have higher chance of offering online services and book table facilities.





We can see from the left plot that low rating as well as high rating restaurants have high chances of offering online order services.

We can see from the right plot that ratings increases as table booking services become more probable.

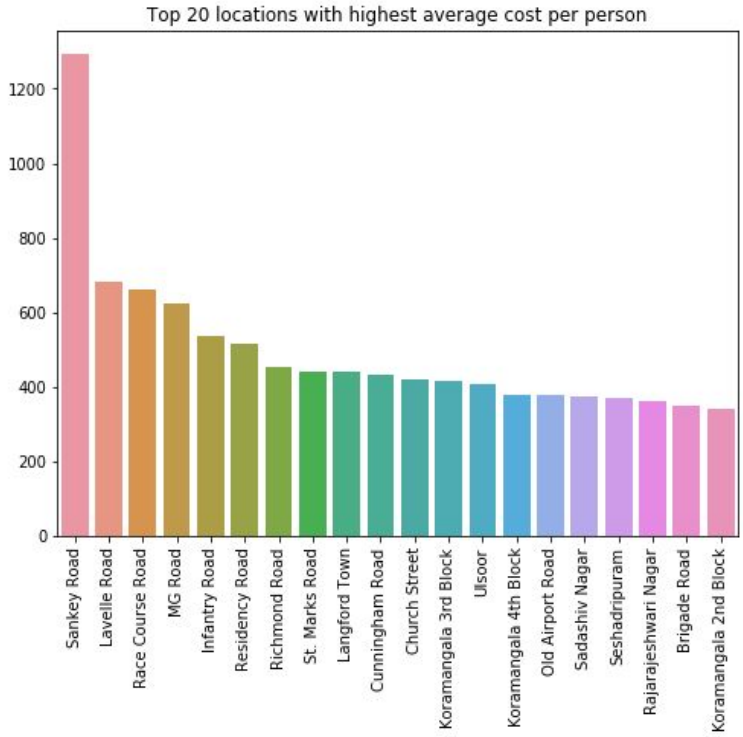
Conclusion- Offering Online order does not guarantee a high rating, but offering table booking service does increase ratings.

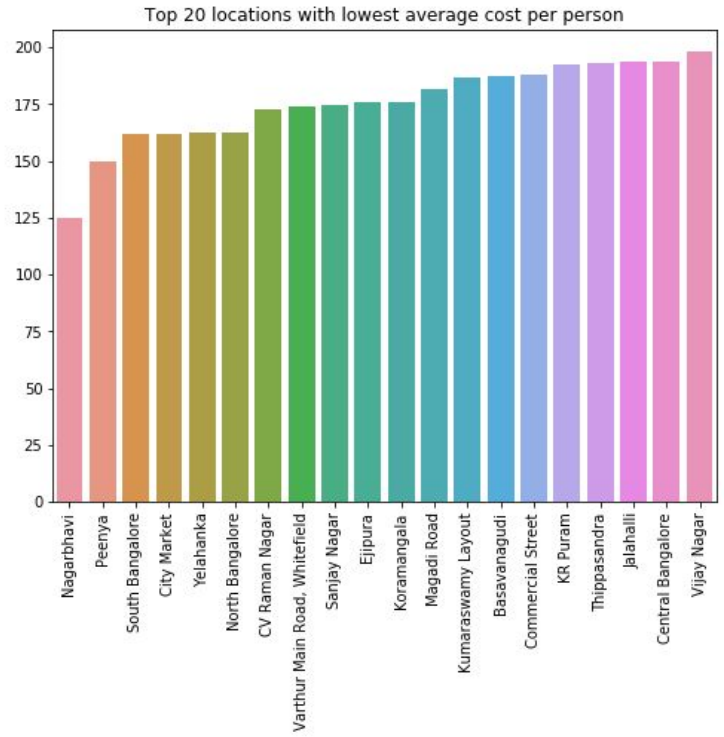
Larger font values means most occured words.

(left) - Word Cloud of Top 3 Restaurant reviews. We see ‘good’ , ‘service’ tags showing positive sentiment.

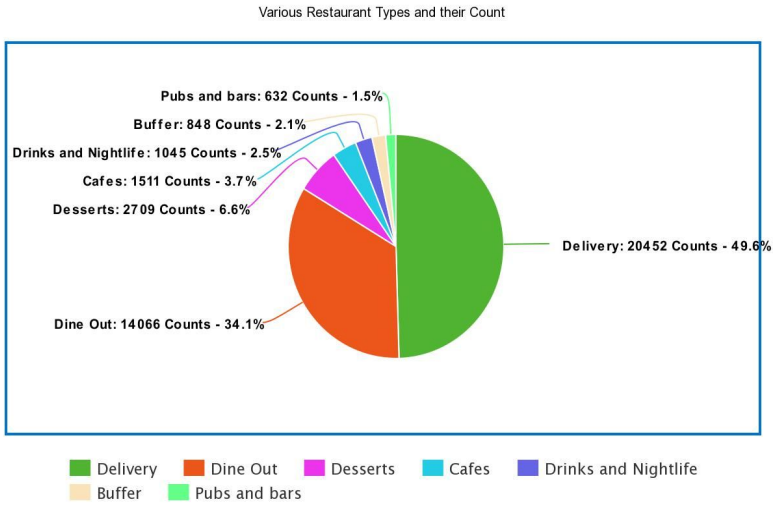
(right) - Word Cloud of Least 3 Restaurant reviews. We see ‘WTF’, ‘bad’ tags showing negative sentiment.

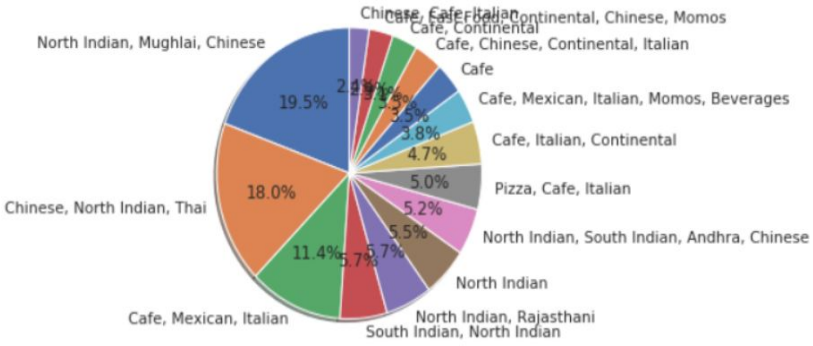
Visualizations



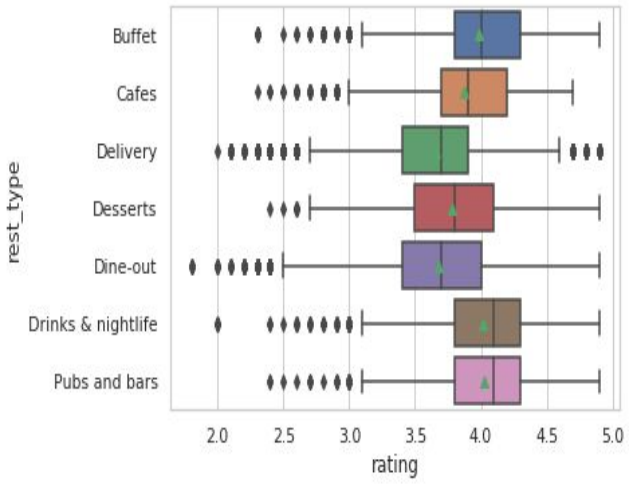


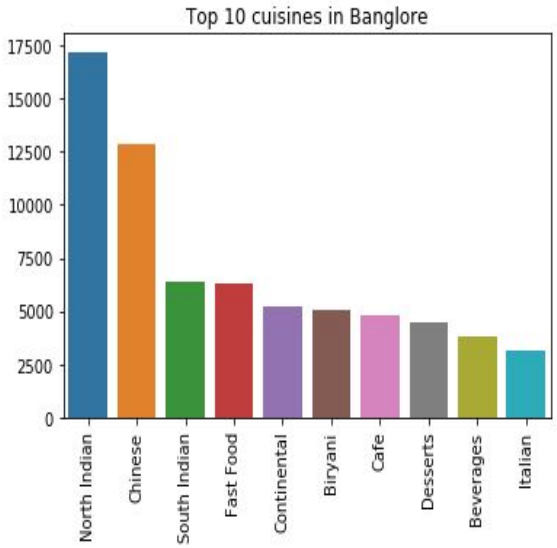
These plots show the most expensive and cheapest places to dine.





(left) This shows that ‘Delivery’ type restaurants are most common.
(right) North Indian cuisine is most popular

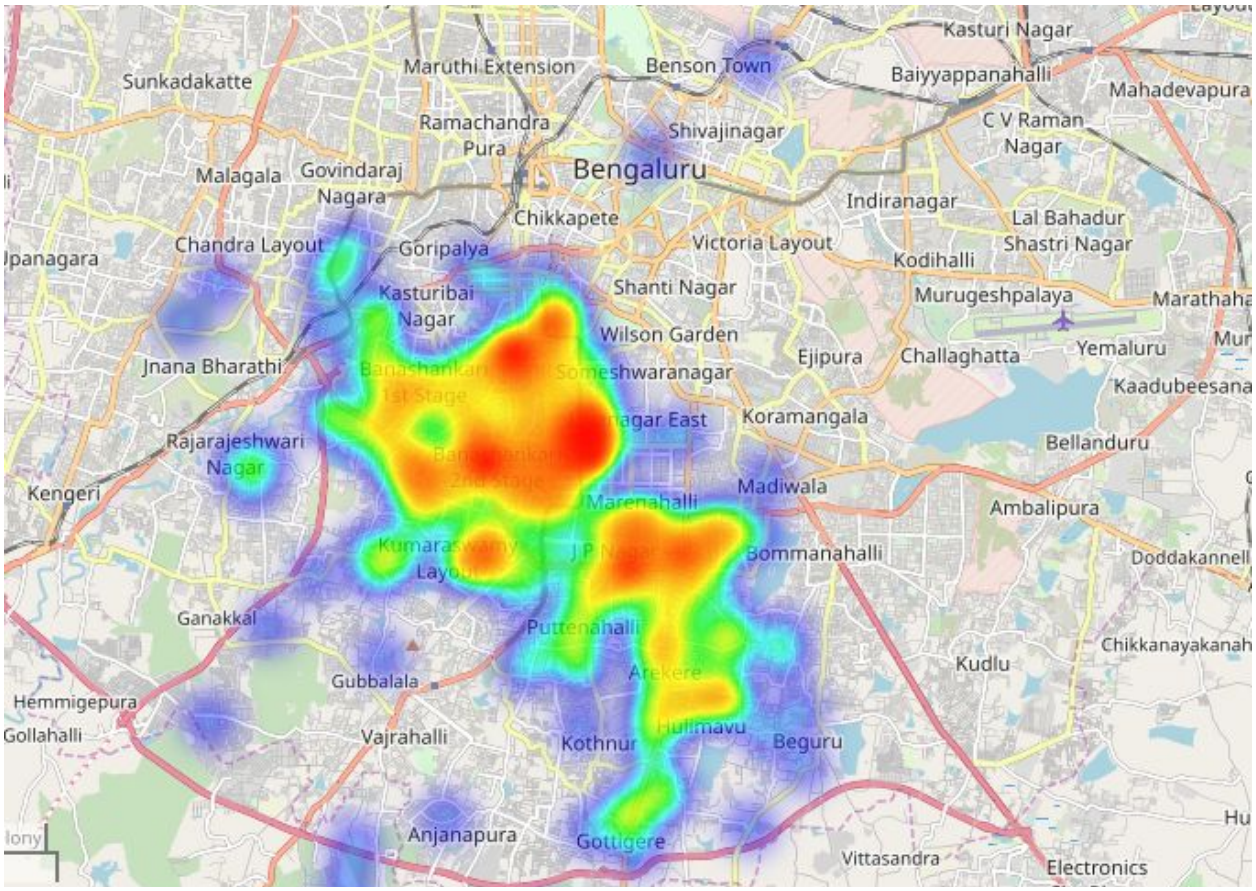




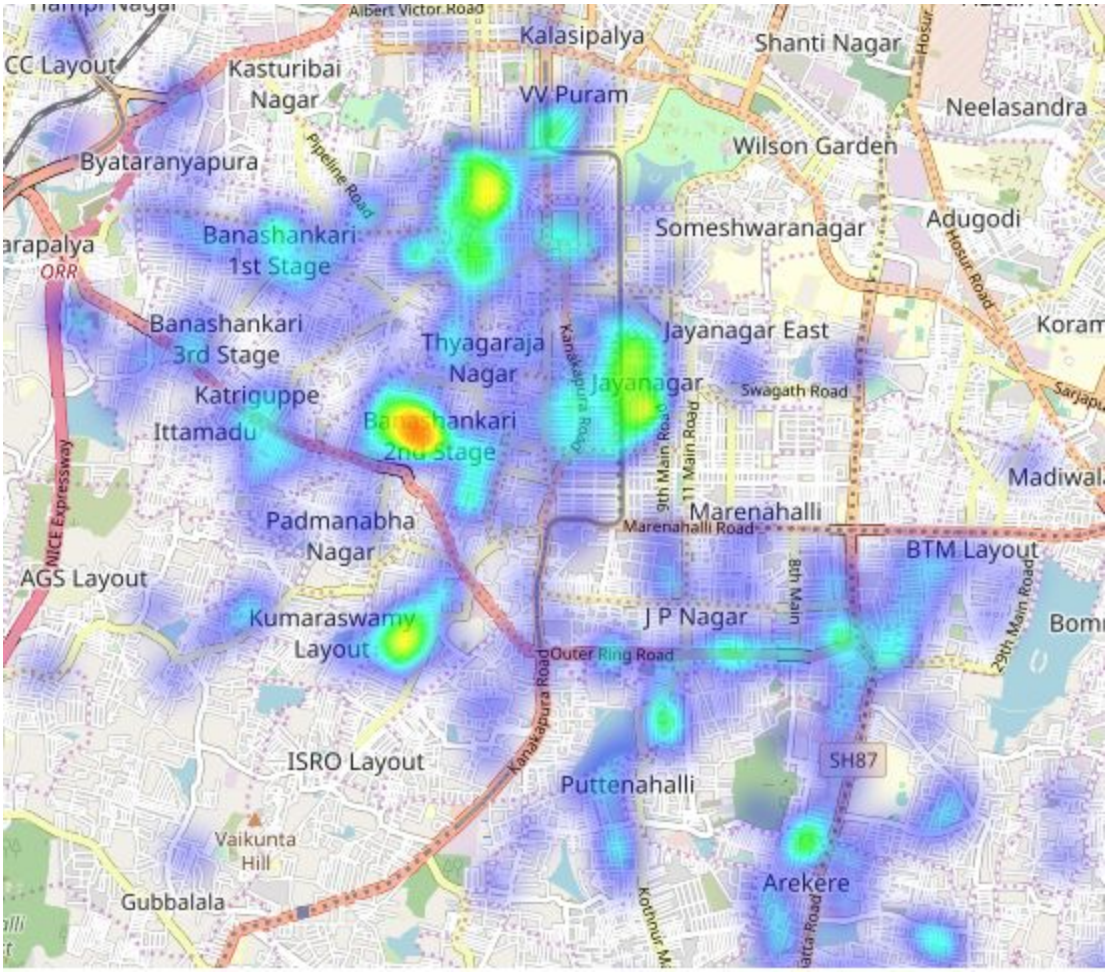
(left) - It shows BOX PLOT of each type of restaurant with it's rating. Green Triangle shows mean. We see mean values are comparable and similar to each other. With extreme high and low values being seen in 'delivery types'.

(right) - It shows frequency count of top cuisines with north indian and chinese being very much preferred.

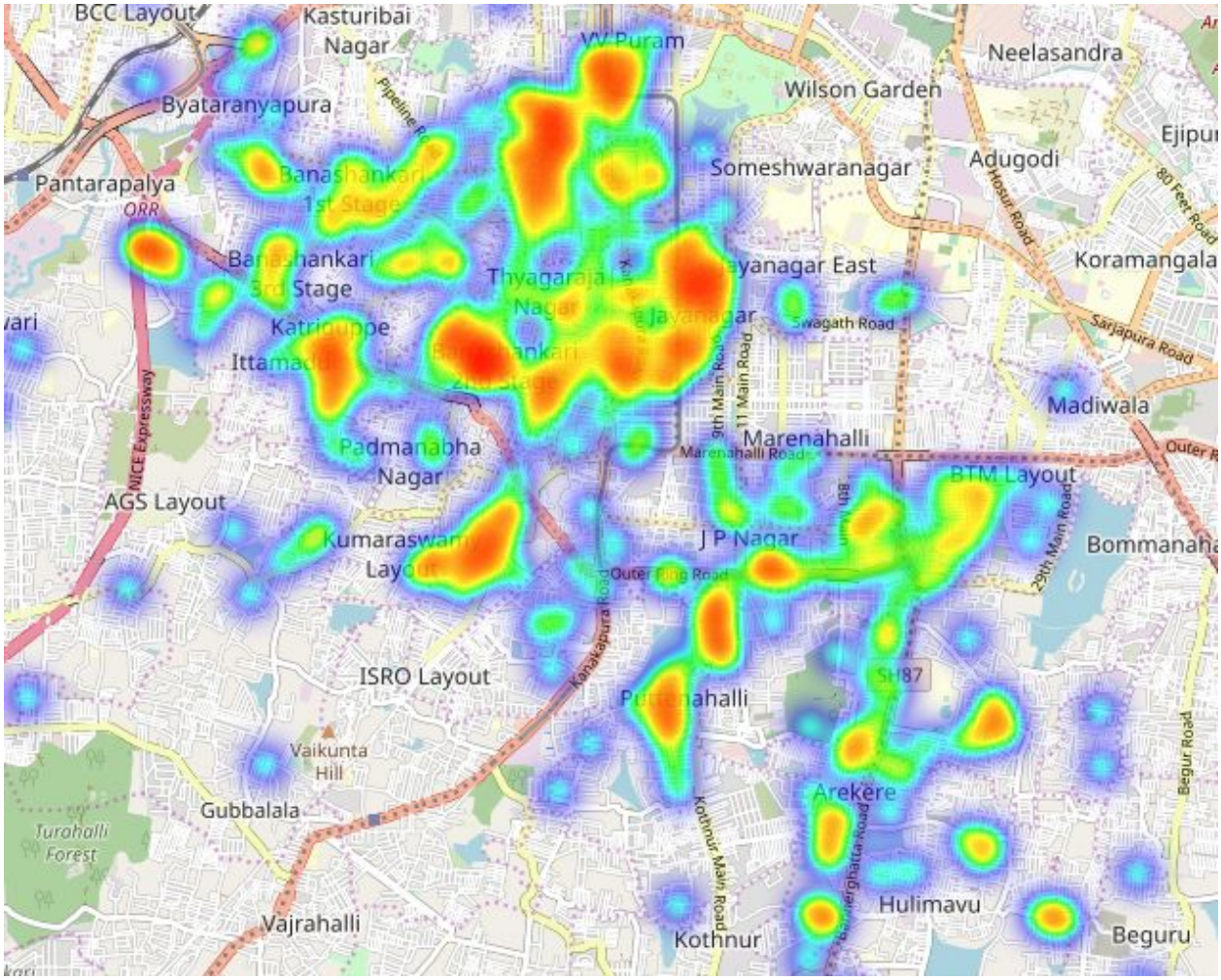
Maps



This is **location density of restaurants**. We can see that almost all restaurants are located near the center of Bengaluru.



This is **rating distribution of restaurants**. We can see that highly rated restaurants are clustered together.



This is **distribution of cost per person**. We can see that expensive restaurants are clustered together.