Zomato Restaurant Rating Analysis

[Praveen Ilango]

1. Introduction

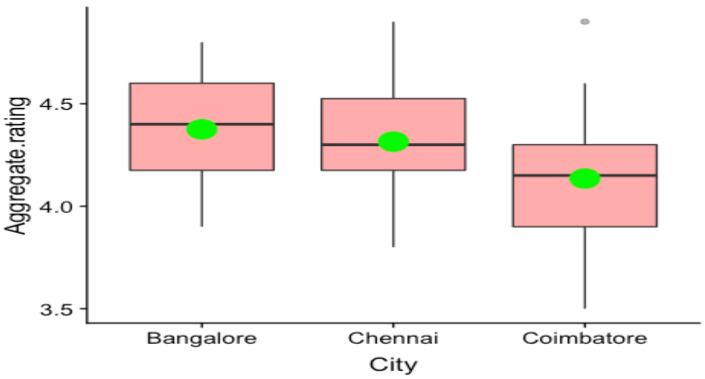
The data used in this project provides information about resteraunts and their associated attributes from around the world. The data has been collected from the Zomato API. Zomato is an Indian restaurant search and discovery service founded in 2008. It currently operates in 24 countries and provides information and reviews of restaurants. The primary purpose of this project is to compare and contrast the difference in aggregate ratings of restaurants located at key cities in South India. The three cities chosen for this analysis are:

- Chennai
- Coimbatore
- Bangalore

Research Question:

Do the aggregate ratings of restaurants at different key cities in the southern part of India differ from each other?

From the below boxplot, we can observe that the sample mean aggregate rating (green points) of restaurants in *Coimbatore* is slightly less than that of *Chennai* and *Bangalore*.



2. Methods: ANOVA & Tucky-Kramer (Multiple Comparison)

The ANOVA test is used to answer the primary research question and Tucky-Kramer is used to examines all pairwise differences between the groups. The following assumptions have to be met to carry out the tests:

1. ***Independence within groups***
2. ***Independence between groups***
3. ***Normality of populations***
4. ***Equal variances in all populations***

The objective of the research question is to evaluate the relationship between the means of three different population groups and the ANOVA test is designed exactly for this purpose, given that all the assumptions mentioned above are met.

Null hypothesis H_0: The population mean aggregate ratings of restaurants μ_j of all the groups are equal to each other.

H_0: $\mu_1 = \mu_2 = \mu_3$

Alternative Hypothesis H_A: At least one of the population mean aggregate ratings of a restaurant μ_j is not equal to the others.

3. Result [Analysis of Variance table]:

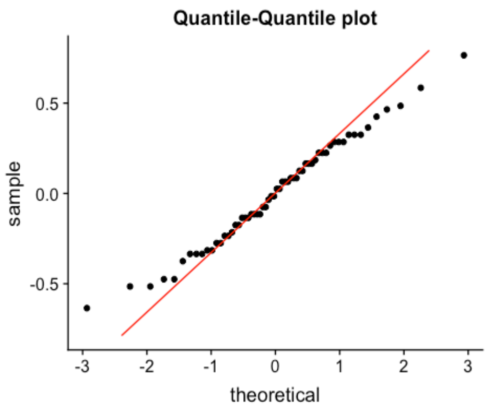
[Response: Aggregate Rating]	DF	Sun Sq	Mean Sq	F value	P value	Statistical Significance
City	2	0.6240	0.312000	3.4144	0.03976	Yes
Residuals	57	5.2085	0.091377			

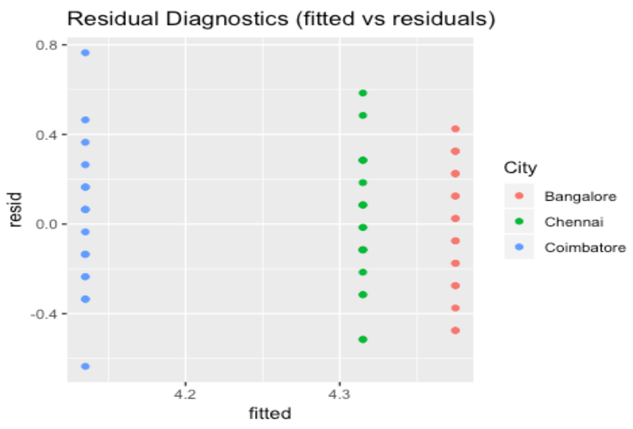
Conclusion: We reject the null hypothesis H_0 that the population mean aggregate ratings of restaurants in all the groups are equal at level $\alpha = 0.05$ in favor of the two-sided alternative hypothesis H_A. We have strong evidence that at least one of the groups have a population mean aggregate rating that is not equal to the others. $p - value = 0.03976 < \alpha = 0.05$.

As there is evidence that the population means are different, further evaluation of *pairwise* comparison is carried out using Tucky-Kramer procedure in order to determine which population means are different. Only the test for the group **Coimbatore - Bangalore** is significant (p-value = 0.039 & estimate = -0.24). The negative estimate indicates that the population aggregate rating of restaurants in Bangalore are higher than that of restaurants at Coimbatore.

4. Assessment

4.1 Normality and Equal variances in all populations





From the Q-Q plot, it is observed that the sample distribution is approximately normal. Also, from the residual diagnostics above, it is evident that there are no outliers and the variances are similar for all groups.

4.3 Independence within and between groups

It is assumed that the aggregate ratings of restaurants are independent within and between each city. Therefore, all of the assumptions for the *ANOVA* test have been met.

5. Conclusion

The analysis indicated that the restaurant ratings differ withing the three chosen cities in South India. The results were not surprising as two of the cities, Chennai and Banglore are IT hubs were restaurants have a high demand for cutomers. Whereas, Coimbatore being a conservative city does not have as high of a demand. Having been a resident at Chennai, it was surprising to see that there was no statistical significance in the comparison between Chennai and Coimbatore.

The analysis could be further extended to fit a linear regression model in order to predict the average cost for two depending on the aggregate restaurant rating.