Business Insights in the Indian Restaurant Market

Banda Nikitha, Escolano Cyrill Ostonal, Prajs Sean Samuel - Singapore Management University

ABSTRACT

This research is explanatory in nature. It aims to provide insights to understand the dining behavior and preferences of people in India by employing interactive data insights and restaurant segmentation through cluster analysis. The IDEA analysis reveals restaurant rating differs based on the state where the restaurant is located. This paper also highlights that rating is dependent on cost. These discovery points are further supported by clustering analysis wherein some clusters are homogenized as top-rated high-end and mid-rangers. The unsupervised learning algorithm behind latent class analysis formed clusters depending on their online visibility.

INTRODUCTION

The rise of the application of information technology in the food sector has been changing consumer behavior and how the food industry do their business. Aside from the restaurants' websites, consumers can now reserve a table or place their order remotely using different online platforms. One of the most widely used applications for food delivery and reservation is Zomato. This company started in India in the year 2008 as a food delivery start up. To date, it has now expanded its presence to 24 countries worldwide.

REVIEW OF RELATED LITERATURE

India is a country of many tastes. Given its population of 1.37 billion people (World Bank, 2019), it boasts a large restaurant landscape. With the ongoing globalization more and more foreign cuisines are coming to India.

There are several trends shaping India's food and beverage market. For one, India is rediscovering itself. Meals from its diverse regions are gaining popularity and their ingredients are used in the creation of novel food styles. These new creations often include ingredients from other Asian cuisines with comparable gusto. On the other hand, so-called "food hedonism" is also gaining traction. This means that there will be deeper indulgence and a wider palette of international flavors found in Indian restaurants in the future (Lobo, 2019).

Nevertheless, no matter what trends will hold true in the future, India already commands an extensive count of restaurants serving foreign cuisines. Among the most popular of these are Italian, Chinese, and American restaurants. Clearly, Indians do have a taste not only for their own food but are open to ingredients used around the globe (Business Insider, 2019).

A large food market like India's also comes with a vast array of preferences. In fact, there are 41 distinct dietary patterns that can be found in the country. Even though customary vegetarian diets are found in many of the dietary groups, meat and high-sugar/fat consumption also comprises a significant part of the population's eating habits (Green, Milner, Joy, Agrawal, & Dangour, 2016).

Resulting from the rising popularity of global food in India is a business case. However, before opening a restaurant in India, it is imperative to find relevant patterns in consumption. Therefore, this paper looks into the question "What type of restaurant is most likely to be successful in India?"

(Word Count: 2998)

METHODOLOGY

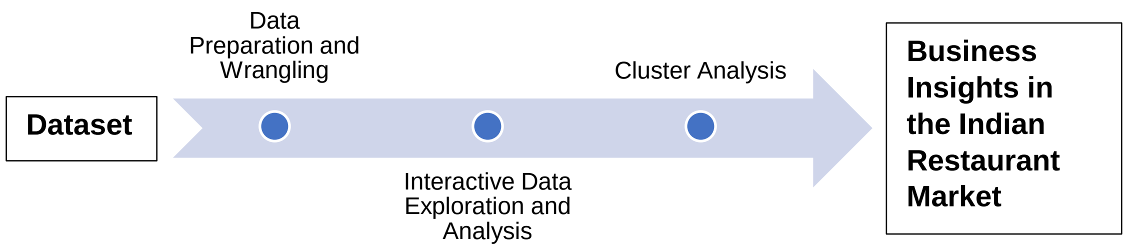


Figure 1: Steps Performed in Carrying Out the Generation of Business Insights

DATASET

Two data sources were taken for this research paper. The main data set is sourced form kaggle.com. It is the record of all the restaurants under Zomato food aggregator in India and contains restaurant names, locations, cuisines served, services offered, ratings, costs, and more variables.

The second data source strictly concerns the creation of the map of India for analysis. It was extracted from the KMP community, where it was posted by user markbailey.

DATA PREPARATION

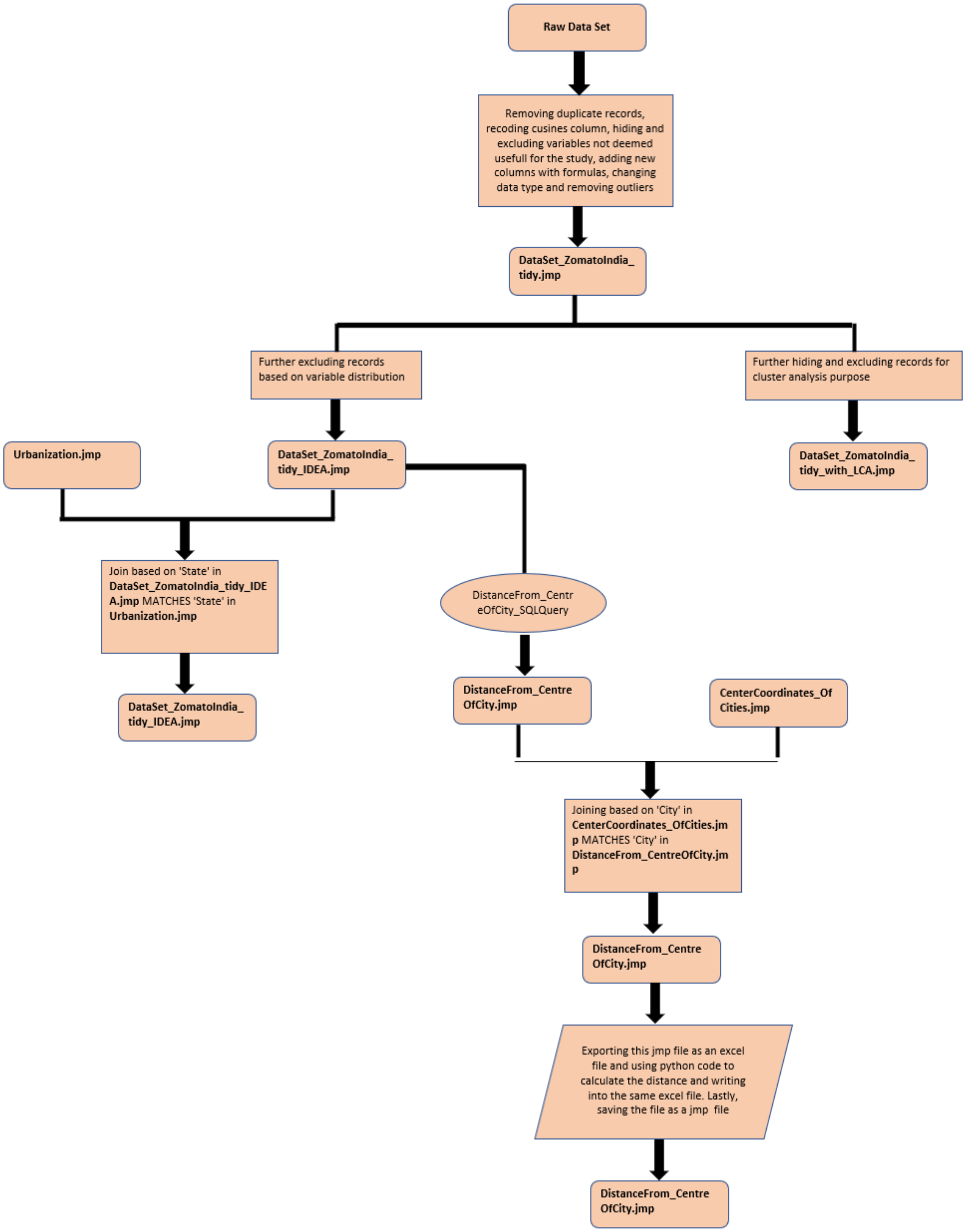


Figure 2. Data Preparation Flow Chart

Data cleaning

Our data preparation is divided into three main parts:

i. General Data Preparation

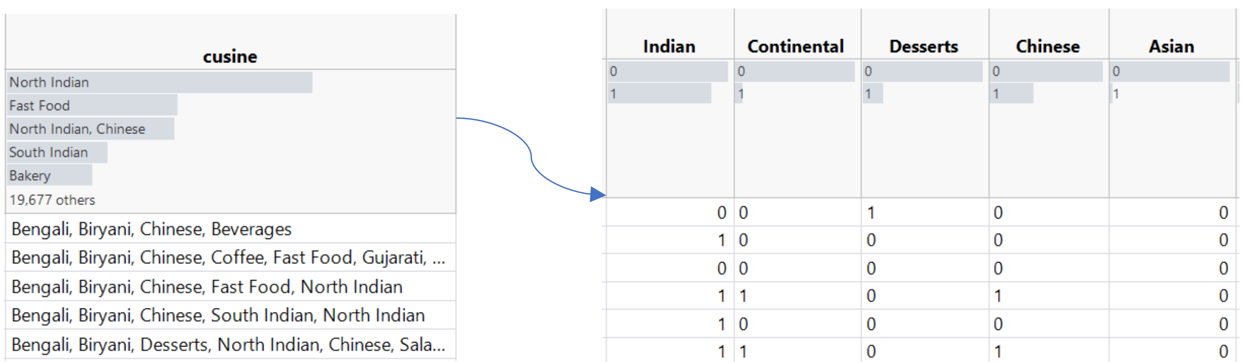
ii. Data Preparation for Cluster Analysis
iii. Data Preparation for IDEA Analysis

We first started off with basic changes like correcting the data type, removing outliers, adding new columns, recoding variables etc. The file were all such changes were made is called 'DataSet_ZomatoIndia.jmp' and then saved as 'DataSet_ZomatoIndia_tidy'. This file is considered the bases file for all our analysis. Detailed steps on data preparation can be found in Annex 1 '**Data preparation change log- General Changes**'.

Taking 'DataSet_ZomatoIndia_tidy.jmp' file as the base, further data cleaning was performed to conduct cluster analysis and IDEA analysis. Detailed steps that describe the data cleaning involved in clustering analysis can be found in Annex 1 '**Data preparation change log- Cluster Analysis**' and steps for IDEA analysis can be found in '**Data preparation change log- IDEA Analysis'.** Major data preparation changes are discussed in the follow up section**.**

Identifying Cuisines

The original dataset consisted of a column named 'cusine' that contained the list of all the cuisines a restaurant serves. This made it difficult to identify exactly which type of cuisines a given restaurant serves. Hence, distinct cuisines were identified, similar cuisines were grouped together and separate columns for each type of cusine were created as seen in the below image.

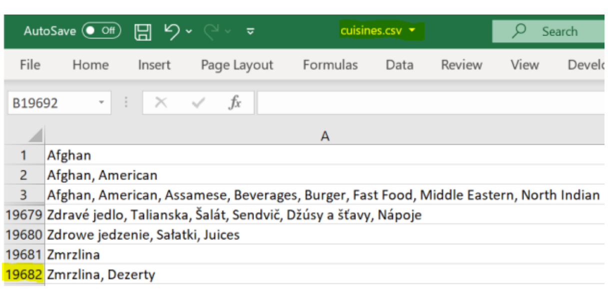


Indian	Continental	Desserts	Chinese	Asian
0	0	1	0	0
1	0	0	0	0
0	0	0	0	0
1	1	0	1	0
1	0	0	0	0
1	1	0	1	0

Figure 3: Initial dataset column and after data processing

Process of converting:

All distinct cuisine records from the column 'cusine' were saved to a separate excel file called 'cuisine.csv' as shown in the figure below. It contains 19,682 unique lists of cuisines. Using the python code in figure, all distinct cuisines were identified which were then grouped based on similarity and added as 14 columns to the dataset. Each column was then updated with a formula that represents '1' if the restaurat serves any of the cuisines mentioned in the formula. It should be noted that some of the cuisine names were in different languages such as Italian, Polish, and Czech.




```
import numpy as np
import pandas as pd
test = pd.read_csv('cuisines.csv',encoding = "ISO-8859-1")
lst=[]
for row in test['ï»¿Afghan']:
    if ',' in row:
                lst2=row.split(",")
                for row2 in lst2:
                    if (row2 or ""+row2) not in lst:
                        lst.append(row2)
    else:
        if (row or ""+row) not in lst:
            lst.append(row)
    lst2=[]
print(lst)
```

Figure 4: File used for python code and python code for identifying distinct cuisines

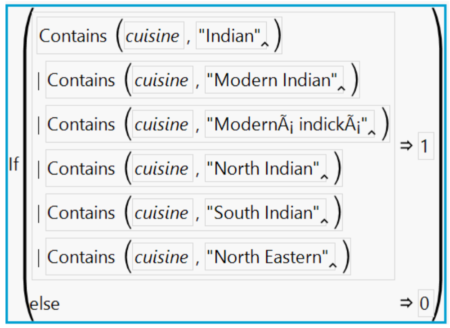


Figure 5: Formula for one of the newly added cuisine column

Urbanization

As part of our analysis for comparison between cost, ratings and urbanization as explained in 'Geographical Insights' under Findings and Discussions we required data on urbanization. For this purpose, a separate excel called 'Urbanization.xlsx' was created with data on the urbanization rate for each distinct city that existed in the dataset. This excel file was then imported and saved as a jmp file and a join was performed between 'DataSet_ZomatoIndia_tidy_IDEA' and 'Urbanization' with 'city=City' as a match criteria as shown in the figure below.

State	Urbanization
Delhi	97.5
Puducherry	68.31
Goa	62.17
Tamil Nadu	48.45
Maharashtra	45.23
Gujarat	42.58
Karnataka	38.57
Punjab	37.49
West Bengal	31.89

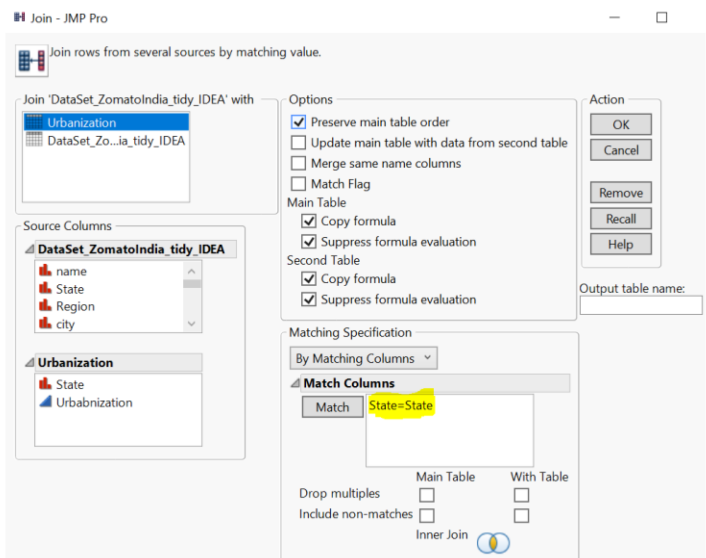


Figure 6: Urbanization file and join criteria used

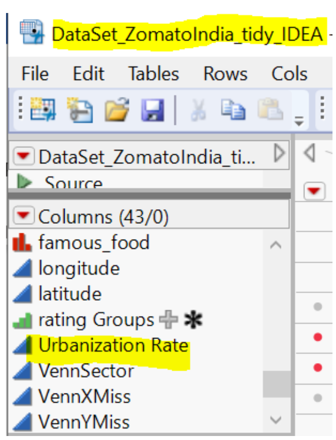


Figure 7: jmp file after the join

Distance from Centre of City

To identify the distance of each restaurant from the center of its city, the latitudes and longitudes coordinates of all distinct cities in the dataset were sourced and saved into an excel file called 'CentreCoordinates_OfCitis.xlsx' which was then saved as a .jmp file. A query shown below was used to extract specific columns required for the analysis and these two file were then joined based on the criteria 'city=City'.

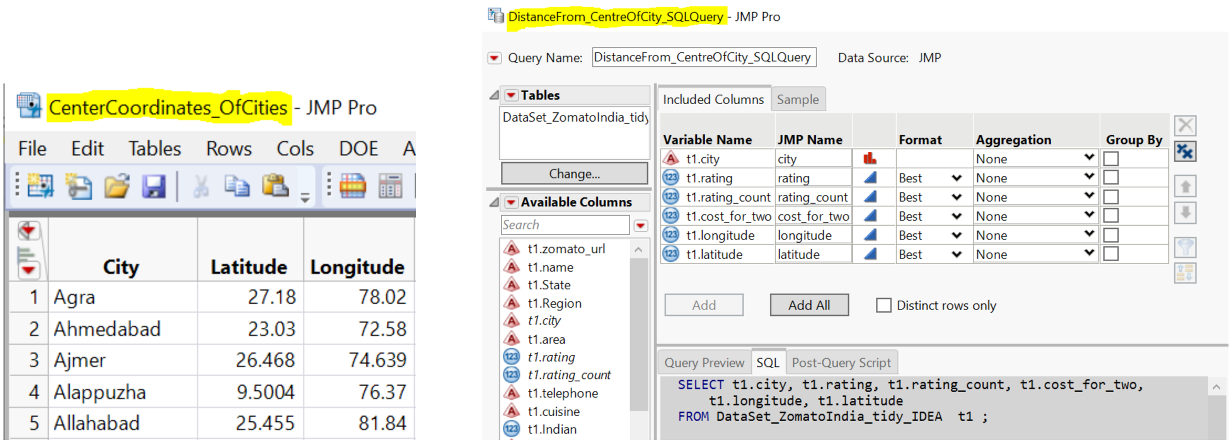


Figure 8: Center Coordinates jmp file and the query to extract required variables for analysis

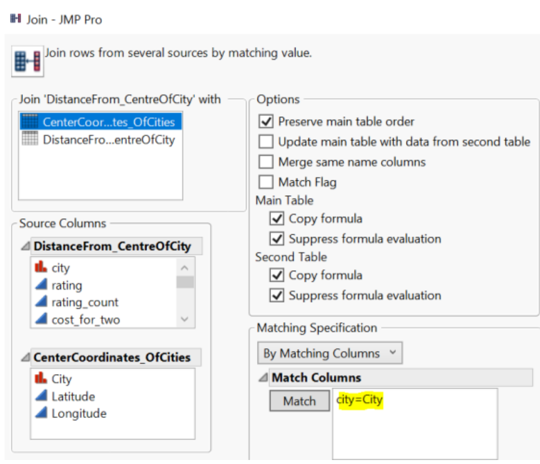


Figure 9: Join criteria used to combine the IDEA.jmp file and Center Coordinates.jmp file

city	rating	rating_count	cost_for_two	latitude	longitude	City	Center of City Latitude	Center of City Longitude
Mumbai	•	•	250	-35.885412...	-83.788490...	Mumbai	18.9667	72.8333
Kolkata	•	•	200	32.25759773	-26.919911...	Kolkata	22.5411	88.3378
Guwahati	2.9	7	250	0	0	Guwahati	26.1667	91.7667
Guwahati	•	•	200	0	0	Guwahati	26.1667	91.7667
Guwahati	•	•	1400	0	0	Guwahati	26.1667	91.7667
Guwahati	•	•	550	0	0	Guwahati	26.1667	91.7667
Guwahati	3	2	200	0	0	Guwahati	26.1667	91.7667
Guwahati	•	•	800	0	0	Guwahati	26.1667	91.7667

Figure 10: The result of the join

The resultant jmp file was then exported as an excel file and the below python code was executed to calculate the distance between the restaurant from the center of the city. The difference in distance was then written into the excel file under the column 'Distance to Town Centre (km)'. Using the interactive binning, the distances were grouped into 5 ranges as shown in the figure below. Notably, the physical formula of converting latitude and longitude differences into kilometers is too complex to display, and as such can be seen only in the python code below.

Python Code:

```
import pandas as pd
from math import sin, cos, sqrt, atan2, radians
import csv
```

```
test = pd.read_csv('Calculate Distance.csv',encoding = "ISO-8859-1")

# approximate radius of earth in km
R = 6373.0
for row in range(len(test)):
    lat1 = radians(test['latitude'][row])
    lon1 = radians(test['longitude'][row])
    lat2 = radians(test['Center of City Latitude'][row])
    lon2 = radians(test['Center of City Longitude'][row])

    dlon = lon2 - lon1
    dlat = lat2 - lat1

    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))

    distance = R * c
    test['Difference in Distance'][row]=distance
temp=[]
x=[]
for row in range(len(test)):
    x.append(test['Difference in Distance'][row])
    temp.append(x)
    x=[]
with open('new1.csv','w') as wobj:
w=csv.writer(wobj)
w.writerows(temp)
```

Figure 11: Python code to calculate the distance from the center of the city to the restaurant

Distance to Town Centre (km)
17057.86802
10832.95171
10187.04996
10187.04996
10187.04996
10187.04996
10187.04996
10187.04996
10187.04996

```
If
As Column(Column("Distance to Town Centre (km)")) ≤ 2                ⇒ "Core"
2 < As Column(Column("Distance to Town Centre (km)")) <= 5           ⇒ "Centre"
5 < As Column(Column("Distance to Town Centre (km)")) <= 10          ⇒ "Near Centre"
10 < As Column(Column("Distance to Town Centre (km)")) <= 20         ⇒ "Suburbs"
20 < As Column(Column("Distance to Town Centre (km)"))               ⇒ "Outskirts"
Is Missing(As Column(Column("Distance to Town Centre (km)")))        ⇒ "Missing"
```

Distance to Town Centre (km)	Distance to Town Centre (km) Groups
14.4121221	Suburbs
0.933536755	Core
3.600890464	Centre
8.73569513	Near Centre
3.253322164	Centre
3.253322164	Centre
3.219448504	Centre
4.956661706	Centre
3.25982393	Centre
38.52483928	Outskirts
3.848482637	Centre
9.193611158	Near Centre
3.606879386	Centre

Figure 12: The distance obtained for the execution of python code and the grouping of distances into 5 groups

States for Mapping

Since only the cities of each restaurant were provided in the dataset, we have included the 'State' that each city is located in. For this processing, we created an excel file consisting of all distinct cities and correctly identifying the State that the city comes under. After that, the excel file was imported into JMP Pro 15.2.0 and saved as 'Indian States.jmp' after which a join was executed to join the tidy jmp file and the Indian state jmp file based on the criteria 'city=City'.

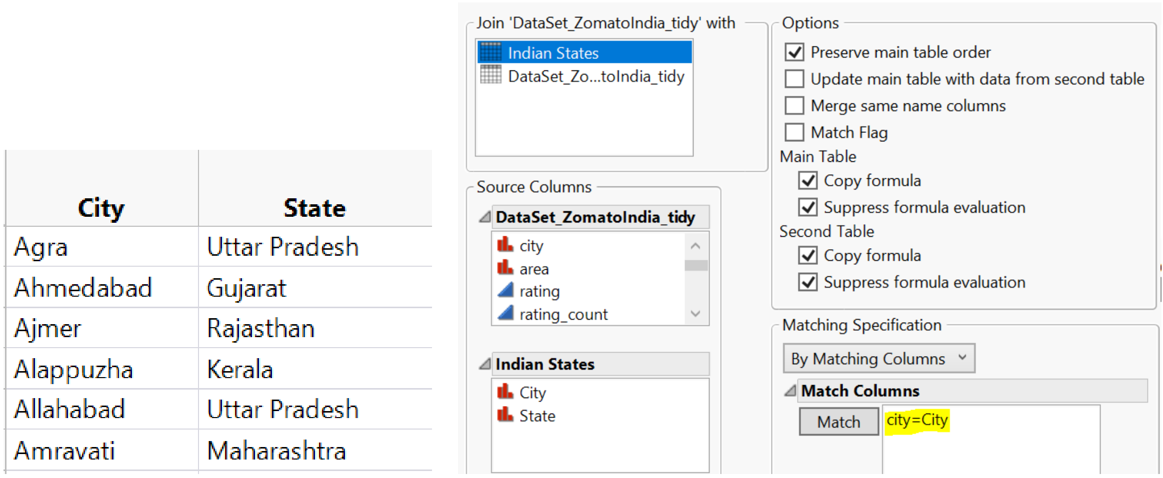


Figure 13: Indian States.jmp file and the join used

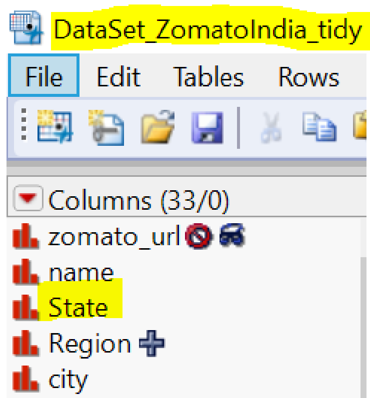


Figure 14: Final tidy file after join

After the states were included, a new column called 'Region' was added and a formula shown below was applied to identify if that state falls under 'North', 'South' or 'North East' India.

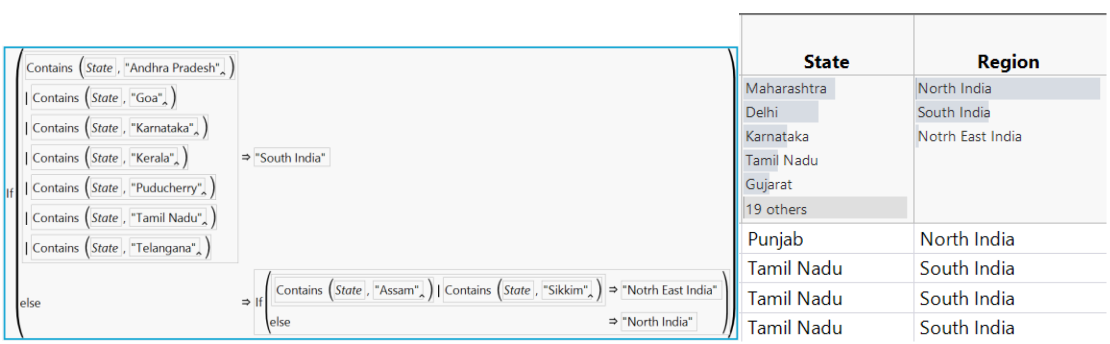


Figure 15: Formula used on 'Region' column and the resultant of that column

METHODOLOGY FOR INTERACTIVE DATA EXPLORATION AND ANALYSIS (IDEA)

This section includes all the steps performed in doing interactive data exploration and analysis. It covers distribution analysis, data visualization and relevant tests performed to verify different hypotheses.

1. Distribution analysis
   This analysis is performed to further check data quality and their appropriateness for the succeeding analysis. We do this to see the presence of outliers and to guide whether or not there is a need for transformation.

2. Venn Diagrams
   Create a Venn diagram to see relationships and possible convergence on the number of restaurants with similar amenities: online order, table reservation, and delivery.

3. Statistical Table
   Tabulate descriptive statistical values to reveal meaningful discovery points that could help in the analysis of the dataset.

4. Heat Maps
   This will be used to reveal geographical distribution of clusters and other relevant variables such as cost for two and rating.

5. Anderson Darling Test
   This non-parametric test is used for testing goodness-of-fit of the distribution of the rating variable.

6. Kruskal-Wallis Test
   This non-parametric test is used to determine if there are statistical differences between state and restaurant rating.

7. Geographical Segmentation
   We do geographical segmentation to reveal insights on possible association of geographical location, ratings, and cost per state.

METHODOLOGY FOR CLUSTER ANALYSIS

Based on the variables available in the dataset, clustering will be done using latent class analysis. We do this by performing the following steps:

1. Generate business questions to determine the needed clustering groups and techniques.
2. Identify variables needed for each clustering technique.
3. Refer to Interactive Data Exploration and Analysis part to check variable distribution.
4. Perform necessary variable transformation.
5. Use JMP add-in interactive binning to divide variable values and transform them to categorical variables.

Questions generated would determine the number of clustering that would be performed when deciding which variables would be used for the different clustering techniques. Below are the business questions that would be answered by cluster analysis.

<table>
<tr><th>Business Questions</th><th>Clustering Technique</th><th>Variables to Be Used</th><th>Required Transformation</th></tr>
<tr><td rowspan="5">What kind of restaurants can be found in India?</td><td rowspan="5">Latent Class Analysis</td><td>rating</td><td>Use interactive binning to divide the values into three bins:
<3: Not so good
3-4: Good enough
>4: Best option</td></tr>
<tr><td>cost_for_two</td><td>1. Transform the variable using log transformation to improve distribution.
2. Use interactive binning to divide values using percentile cut points.</td></tr>
<tr><td>online_order</td><td>No need for transformation.</td></tr>
<tr><td>table_reservation</td><td>No need for transformation.</td></tr>
<tr><td>delivery_only</td><td>No need for transformation.</td></tr>
</table>

What cuisines can be found in the Indian restaurant market?	Two Separate Latent Class Analyses	Indian	No need for transformation.
		Chinese	No need for transformation.
		Asian	No need for transformation.
		Italian	No need for transformation.
		American	No need for transformation.
		Continental	No need for transformation.
		Desserts	No need for transformation.
		Beverages	No need for transformation.
		Fast Food	No need for transformation.
		Seafood	No need for transformation.

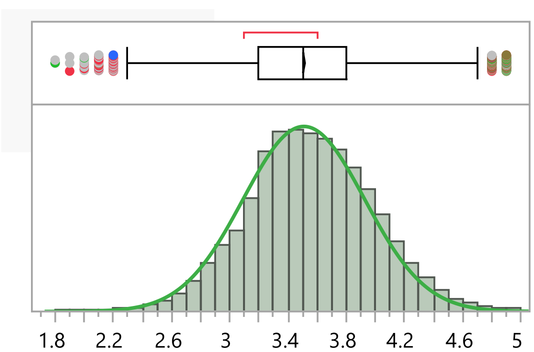
Figure 16 Summary of Techniques for Cluster Analysis

FINDINGS AND DISCUSSIONS

INTERACTIVE DATA EXPLORATION AND ANALYSIS

General Data Insights

Looking into IDEA, the first variable that has to be examined is "rating." It gives the average rating of any restaurant included in the data. First, looking at the distribution of the variable, it seems not to have too many outliers as seen in the figure below.



Mean	3.5088275
Std Dev	0.4175528
Std Err Mean	0.0011367
Upper 95% Mean	3.5110555
Lower 95% Mean	3.5065996

Figure 17. Summary Statistics of Rating

The factor shows a distribution that looks approximately normal. This can be seen by the proximity of the normal fit line (green) to the distribution. However, to validate this an Anderson-Darling Goodness of Fit test needs to be carried out. The hypotheses are the following:

H_0: rating follows a normal distribution

H_1: rating does not follow a normal distribution

The test yields a p-value smaller than 0.0001. Thus, despite similarity, rating does not follow a normal distribution.

Continuing with a general look on the data. Three of the main characteristics of restaurants in the data are whether online orders or table reservations are available and whether the restaurant is delivery only.

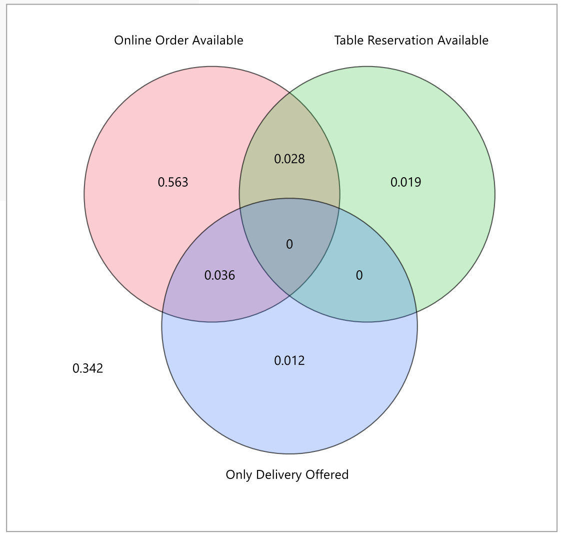


Figure 18. Venn Diagram of Proportions of Restaurant Offerings

The first important trend seen in the Venn Diagram is that there are no restaurants that are delivery only and offer table reservation. This is natural as delivery only restaurants do not have seating. Out of the three characteristics, by far the most common one is online orders. Around 56% of restaurants in the data have that. Comparably smaller are the proportions of restaurants that have table reservations or are delivery only (1.9% and 1.2% respectively). Interestingly, around a third of restaurants do not have any of the three characteristics. It is important to note these trends as there is a cost-benefit trade-off when deciding whether to adopt these characteristics.

Geographical Insights

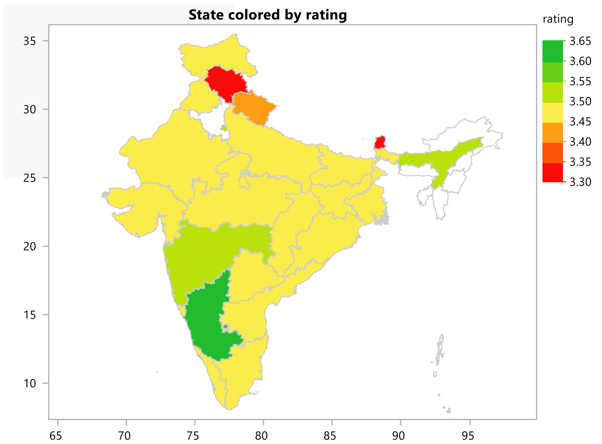
Next, a deeper dive into the data. India is a large country with numerous states. Evidently, the question whether certain states have higher rated restaurants than others comes to mind. Therefore, the Kruskal-Wallis Test needs to be run between variables “state” and “rating.” Specifications for the null and alternative hypotheses are:

H_0: rating does not differ based on state
H_1: rating differs based on state

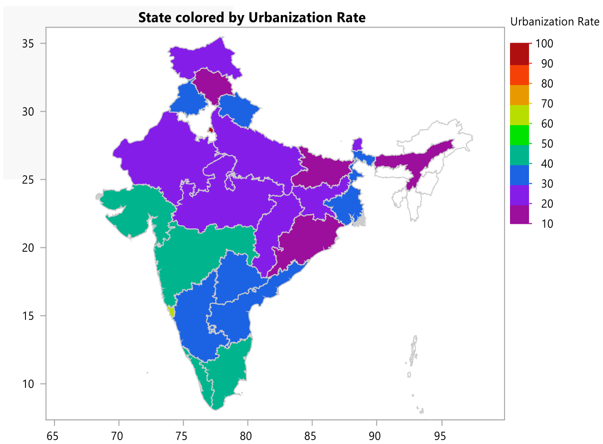
ChiSquare	DF	Prob>ChiSq
1663.2986	23	<.0001*

Figure 19. Results of Kruskal-Wallis Test between rating and state

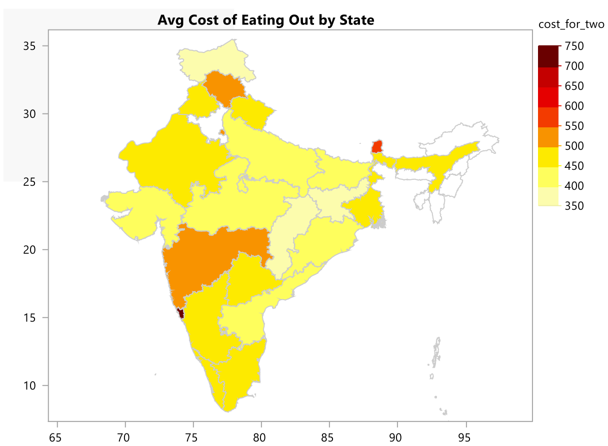
Given the small p-value, it can be concluded that rating is dependent on the state the restaurant is in. This could have different reasons.



As can be seen here, Karnataka (dark green) has the highest average rating, while Sikkim (smaller deep red) has the lowest average rating for its restaurants.



One of the reasons for different average rating in states could be the varying urbanisation. However, when comparing the trend in urbanisation rates of Indian states as compared to their average rating, no clear pattern emerges.



Another potential reason for the difference in average rating could be the different average cost for two people eating out in a restaurant. Similarly to the urbanisation rate though, there is no clear correlation of the average cost of eating out in a state and its average restaurant rating.

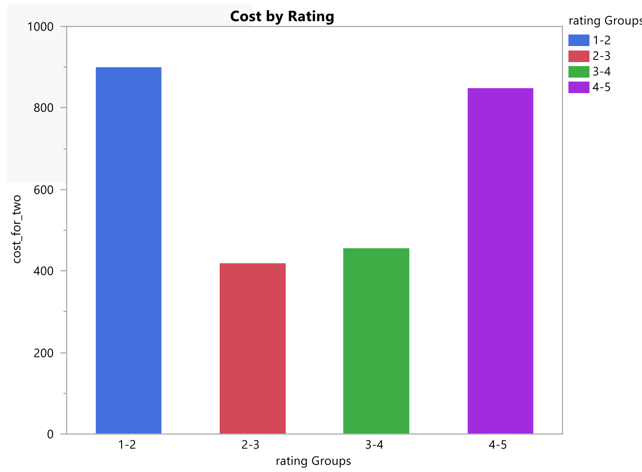
Figure 20. Rating, Urbanization, and Cost by State

Nevertheless, cost does seem to have something to do with the rating. When carrying out a Kruskal-Wallis test between the two variables a clear dependency can be observed. The data for the test is in the table below. The hypotheses are:

H_0: rating does not differ based on cost
H_1: rating differs based on cost

ChiSquare	DF	Prob>ChiSq
6266.9895	3	<.0001*

Figure 21. Results of Kruskal-Wallis Test between rating groups and cost



	rating Groups			
	1-2	2-3	3-4	4-5
cost_for_two (mean)	900	418.1464889	454.91443603	848.68091211

Figure 22. Cost by Rating Bar Chart

However, the trend between rating and cost does not follow any particular direction as seen above.

Distance Analysis

When opening a restaurant, location can be everything. Being far away or close to the center of a city both have their advantages and disadvantages. Every city in India has its unique mix of restaurants belonging to either category. Therefore, the following table shows these unique mixes. Larger cities like Delhi and Mumbai have a lot of restaurants that are further away from the center, while smaller cities seem to have more restaurants at their core.

The distance from the city center of each category is as follows:

- Core: 0-2km
- Centre: 2-5km
- Near Centre: 5-10km
- Suburbs: 10-20km
- Outskirts: 20km +

The three tables below are further divided into three regions in India. The map in Annex II shows the allocation of states into North India, North East India, and South India.

It should be noted that for some of the restaurants Goa was named as the city of locality. Since Goa is not a city and therefore taking its center is not useful, it is not included in the following tables. Furthermore, the aforementioned exclusion by certain criteria (in data preparation) has excluded all rows of 5 cities, which are Alappuzha, Amravati, Junagadh, Kharagpur, and Palakkad.

City	Core	Centre	Near Centre	Suburbs	Outskirts
Agra	7.36%	70.89%	21.63%	0.12%	0.00%
Ahmedabad	6.92%	25.32%	51.43%	14.19%	2.13%
Ajmer	46.85%	45.45%	7.69%	0.00%	0.00%
Allahabad	53.51%	28.31%	17.77%	0.41%	0.00%
Amritsar	3.27%	78.98%	16.76%	0.85%	0.14%
Aurangabad	29.51%	54.78%	14.21%	1.50%	0.00%
Bhopal	10.53%	47.53%	40.08%	1.86%	0.00%
Bhubaneswar	14.30%	45.09%	31.23%	9.21%	0.18%
Chandigarh	12.00%	30.04%	35.63%	22.30%	0.04%
Cuttack	11.11%	69.44%	19.44%	0.00%	0.00%
Darjeeling	100.00%	0.00%	0.00%	0.00%	0.00%
Dehradun	24.35%	44.01%	28.65%	2.99%	0.00%
Delhi NCR	0.83%	6.74%	16.04%	48.58%	27.81%
Dharamshala	38.68%	54.72%	4.72%	0.00%	1.89%
Gorakhpur	52.47%	41.67%	5.56%	0.31%	0.00%
Gwalior	38.00%	53.31%	8.32%	0.38%	0.00%
Haridwar	22.22%	39.39%	37.37%	1.01%	0.00%
Indore	7.17%	49.27%	41.15%	2.36%	0.05%
Jabalpur	57.48%	32.24%	10.05%	0.23%	0.00%
Jaipur	0.03%	13.20%	42.15%	44.18%	0.44%
Jalandhar	43.61%	49.91%	5.11%	0.68%	0.68%
Jammu	32.18%	44.44%	21.06%	1.62%	0.69%
Jamnagar	20.00%	80.00%	0.00%	0.00%	0.00%
Jamshedpur	35.22%	49.36%	15.17%	0.26%	0.00%
Jhansi	76.95%	22.71%	0.34%	0.00%	0.00%
Jodhpur	23.19%	61.51%	14.35%	0.95%	0.00%
Kanpur	23.89%	48.53%	26.45%	1.14%	0.00%
Kolhapur	100.00%	0.00%	0.00%	0.00%	0.00%
Kolkata	7.81%	28.93%	33.36%	25.76%	4.14%
Kota	7.48%	61.22%	31.29%	0.00%	0.00%
Lucknow	14.94%	28.70%	52.51%	3.84%	0.00%
Ludhiana	26.00%	56.31%	17.44%	0.19%	0.06%
Manali	0.00%	43.75%	51.56%	0.00%	4.69%
Meerut	35.98%	38.41%	25.39%	0.22%	0.00%
Mumbai	3.68%	9.91%	7.22%	33.69%	45.49%
Mussoorie	23.19%	76.81%	0.00%	0.00%	0.00%
Nagpur	20.81%	46.82%	30.56%	1.75%	0.05%
Nainital	100.00%	0.00%	0.00%	0.00%	0.00%
Nashik	25.10%	53.54%	19.31%	2.06%	0.00%
Neemrana	16.67%	33.33%	16.67%	16.67%	16.67%

Patiala	34.17%	45.49%	15.30%	0.00%	5.03%
Patna	36.32%	40.61%	18.92%	4.16%	0.00%
Pune	6.25%	15.81%	46.59%	29.88%	1.47%
Pushkar	97.37%	2.63%	0.00%	0.00%	0.00%
Raipur	25.66%	60.71%	13.63%	0.00%	0.00%
Rajkot	46.91%	46.91%	5.76%	0.41%	0.00%
Ranchi	38.69%	46.58%	12.50%	2.08%	0.15%
Rishikesh	2.94%	39.22%	57.84%	0.00%	0.00%
Shimla	89.13%	10.87%	0.00%	0.00%	0.00%
Siliguri	25.37%	71.69%	2.94%	0.00%	0.00%
Srinagar	4.44%	86.67%	6.67%	2.22%	0.00%
Surat	5.29%	38.41%	52.58%	3.53%	0.19%
Udaipur	45.38%	49.34%	5.01%	0.13%	0.13%
Vadodara	9.45%	71.20%	19.07%	0.12%	0.17%
Varanasi	14.63%	70.74%	14.18%	0.30%	0.15%

Figure 23. Restaurant Proximity by City (North India)

City	Core	Centre	Near Centre	Suburbs	Outskirts
Gangtok	95.74%	4.26%	0.00%	0.00%	0.00%
Guwahati	34.59%	42.27%	20.76%	2.38%	0.00%

Figure 24. Restaurant Proximity by City (North East India)

City	Core	Centre	Near Centre	Suburbs	Outskirts
Bengaluru	4.94%	18.53%	44.21%	31.18%	1.14%
Chennai	2.88%	16.05%	33.88%	33.56%	13.64%
Coimbatore	26.69%	33.54%	34.96%	4.72%	0.08%
Guntur	41.23%	57.89%	0.88%	0.00%	0.00%
Hyderabad	1.77%	12.24%	35.05%	47.47%	3.46%
Kochi	21.37%	34.12%	31.32%	9.29%	3.91%
Madurai	27.18%	64.08%	8.74%	0.00%	0.00%
Mangalore	1.58%	72.54%	20.77%	5.11%	0.00%
Manipal	23.04%	41.74%	35.22%	0.00%	0.00%
Mysore	28.32%	54.40%	17.27%	0.00%	0.00%
Ooty	93.33%	4.00%	1.33%	1.33%	0.00%
Puducherry	61.80%	21.09%	13.15%	3.97%	0.00%
Salem	10.67%	70.22%	15.73%	3.37%	0.00%
Thrissur	69.70%	25.25%	4.55%	0.51%	0.00%
Tirupati	0.00%	100.00%	0.00%	0.00%	0.00%
Trichy	47.10%	44.37%	8.53%	0.00%	0.00%
Trivandrum	0.29%	18.44%	78.96%	2.02%	0.29%

Vellore	27.75%	38.15%	34.10%	0.00%	0.00%
Vijayawada	20.09%	54.59%	24.02%	1.31%	0.00%
Visakhapatnam	39.03%	31.08%	15.90%	12.81%	1.19%

Figure 25. Restaurant Proximity by City (South India)

Cuisines Analysis

Lastly, the food offered in a restaurant is extremely important. The following figure offers insights to the average rating of restaurants containing a certain cuisine ("Yes") compared to restaurants not offering the cuisine ("No").

Cuisine	Mean Rating "No"	Mean Rating "Yes"
Indian	3.53	3.49
Continental	3.49	3.74
Desserts	3.5	3.56
Chinese	3.52	3.48
Asian	3.5	3.86
Neighboring Cuisines	3.5	3.57
Other	3.5	3.55
Italian	3.49	3.68
Non-Indian	3.49	3.7
Middle Eastern	3.5	3.61
American	3.5	3.71
Beverages	3.5	3.58
Fast Food	3.52	3.49
Seafood	3.51	3.55

Figure 26. Rating by Cuisine Inclusion

The red colored rows offer a view of the cuisines whose inclusion generates a lower average rating. However, the differences in these rows are marginal and therefore negligible. Nevertheless, there are certain cuisines which provide significant improvement to average restaurant rating. These are:

- Continental
- Asian
- Italian
- Non-Indian
- American

Thus, to conclude, while no cuisine will significantly hurt a restaurant, some food types are likely to increase the popularity of a restaurant.

CLUSTER ANALYSIS

Restaurants by Amenities

There are six clusters generated using latent class analysis. Percentage distribution of each category for each variable is shown in the table below. Clusters are generally homogenous. Based on the data presented below, we can classify the clusters as follows: mixed mid-rangers, online exotic bites, offline budget-friendly, local luxury delights, top-rated high-end, and take-aways.

Cluster Comparison

NCluster	-LogLikelihood	BIC	AIC	Best
3	342484	685240	685014	
4	341951	684268	683963	
5	341822	684104	683721	
6	341710	683976	683515	Smallest BIC
7	341702	684054	683514	
8	341657	684058	683440	Smallest AIC
9	341659	684156	683460	

Parameter Estimates

		online_order		table_reservation		delivery_only		rating Groups			Log[cost_for_two] Groups		
Cluster	Overall	False	True	False	True	False	True	Not so good	Good enough	Best option	Cheap eats	Midrange	High end
Cluster 1	0.33029	0.4182	0.5818	0.9994	0.0006	0.9269	0.0731	0.2546	0.7327	0.0127	0.1617	0.7798	0.0585
Cluster 2	0.26745	0.0809	0.9191	0.9995	0.0005	0.9877	0.0123	0.0001	0.9153	0.0847	0.2710	0.6813	0.0476
Cluster 3	0.22663	0.7626	0.2374	0.9997	0.0003	0.9904	0.0096	0.1952	0.7715	0.0333	0.3186	0.4644	0.2170
Cluster 4	0.08901	0.1091	0.8909	0.9262	0.0738	0.9999	0.0001	0.0218	0.7096	0.2686	0.0001	0.7701	0.2298
Cluster 5	0.06625	0.4537	0.5463	0.3946	0.6054	0.9996	0.0004	0.0098	0.4569	0.5333	0.0000	0.0000	1.0000
Cluster 6	0.02035	0.0324	0.9676	0.9998	0.0002	0.0758	0.9242	0.0192	0.7549	0.2259	0.0007	0.7035	0.2957

Figure 27 Results of Cluster Analysis

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
online_order						
False	76.37%	0.00%	100.00%	0.65%	41.80%	0.25%
True	23.63%	100.00%	0.00%	99.35%	58.20%	99.75%
table_reservation						
False	99.98%	99.99%	99.98%	93.68%	34.47%	100.00%
True	0.02%	0.01%	0.02%	6.32%	65.53%	0.00%
delivery_only						
False	93.01%	99.97%	100.00%	100.00%	99.98%	0.00%
True	6.99%	0.03%	0.00%	0.00%	0.02%	100.00%
rating						
Not so good	39.17%	0.00%	17.93%	0.16%	0.69%	0.00%
Good enough	60.77%	98.40%	75.12%	59.79%	30.48%	84.57%
Best options	0.06%	1.60%	6.95%	40.05%	68.84%	15.43%
Cost_for_two						
Cheap eats	5.87%	24.46%	58.11%	0.00%	0.00%	0.00%
Midrange	91.47%	75.54%	5.21%	44.77%	0.00%	79.23%
High end	2.67%	0.00%	36.68%	55.23%	100.00%	20.77%

Figure 28 Percentage Distribution of Variables in Each Cluster

Indian	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
No	51.07%	56.45%	60.98%	44.56%	38.38%	46.80%
Yes	48.93%	43.55%	39.02%	55.44%	61.62%	53.20%

Figure 29 Distribution of Indian Food in Each Cluster

Although the variable Indian is not used for clustering, the extra information obtained by comparing the clusters to the inclusion of Indian food provides valuable information on what type of restaurant is the most likely to serve the national cuisine. Latent class analysis was able to cluster restaurants depending on their amenities, ratings, and cost for two.

Clusters	Labels	Characteristics
1	Mixed Mid-Rangers	1. Most of them do not accept online order. 2. Most of them do not offer table reservation. 3. Most of them offer dine-in service. 4. Most are classified as “Good enough” restaurants. 5. Most are midrange. 6. They have an almost equal mix of restaurants that serve Indian and Non-Indian food.
2	Online Exotic Bites	1. All of them accept online order. 2. Most of them do not offer table reservation. 3. Most of them offer dine-in service. 4. Most are classified as “Good enough” restaurants. 5. Most are midrange. None of them are high end. 6. Most of them serve only Non-Indian food.
3	Offline Budget-friendly	1. All of them do not accept online order. 2. Most of them do not offer table reservation. 3. All of them offer dine-in service. 4. Most are classified as “Good enough” restaurants. 5. More than half are classified “Cheap eats”. 6. Most of them serve only Non-Indian food.
4	Local Luxury Delights	1. Most of them accept online order. 2. Most of them do not offer table reservation. 3. All of them offer dine-in service. 4. Although most of them are classified “Good enough”, significant percentage are in the “Best options”. 5. Most of them are classified “High end” with significant mix of “Midrange” ones. 6. Most of them serve only Indian food.
5	Top rated high-end	1. Most of them accept online order with significant percentage of those that do not accept. 2. Most of them offer table reservation. 3. Most of them offer dine-in service. 4. Most of them are classified “Best options”. 5. All of them are classified “High end”. 6. Most of them serve only Indian food.
6	Take aways	1. Most of them accept online order. 2. All of them do not offer table reservation. 3. All of them are delivery only. 4. Most of them are classified “Good enough”. 5. Most of them are classified “Midrange”. 6. Most of them serve only Indian food.

Figure 30 Summary of Cluster Characteristics

Using the variable State, the clusters generated can give an idea on which kind of restaurants can be found in each state. Among all other states, Goa and Sikkim stand out for having the most proportion of restaurants classified as offline budget friendly.

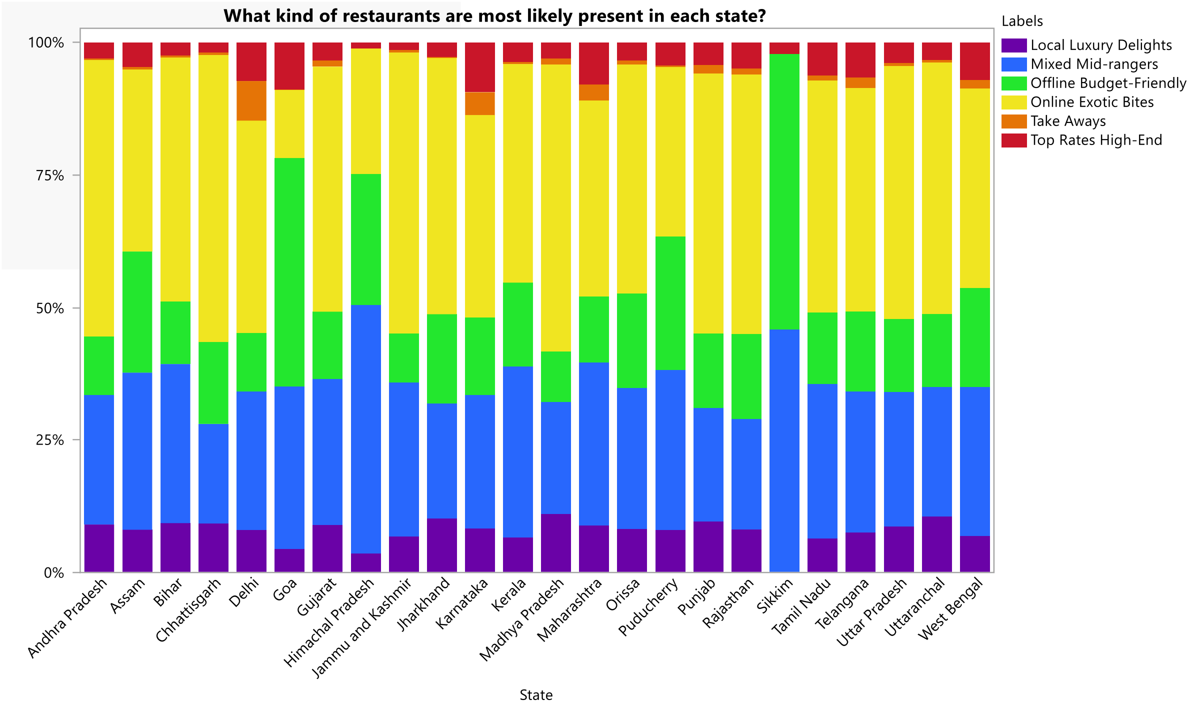


Figure 31 Visual Representation of Cluster Distribution per State

State	Mixed Mid-rangers	Online Exotic Bites	Offline Budget-Friendly	Local Luxury Delights	Top Rated High-End	Take Aways
Andhra Pradesh	24.43%	52.23%	11.10%	9.06%	3.00%	0.19%
Assam	29.66%	34.38%	22.85%	8.07%	4.61%	0.42%
Bihar	30.02%	46.00%	11.86%	9.32%	2.42%	0.36%
Chhattisgarh	18.74%	54.10%	15.52%	9.31%	1.88%	0.44%
Delhi	26.10%	40.05%	11.11%	8.03%	7.23%	7.48%
Goa	30.68%	12.90%	43.05%	4.51%	8.81%	0.05%
Gujarat	27.60%	46.25%	12.65%	8.98%	3.39%	1.13%
Himachal Pradesh	46.91%	23.64%	24.73%	3.64%	1.09%	0.00%
Jammu and Kashmir	29.03%	53.02%	9.27%	6.85%	1.41%	0.40%
Jharkhand	21.79%	48.38%	16.81%	10.16%	2.68%	0.18%
Karnataka	25.15%	38.14%	14.69%	8.38%	9.34%	4.31%
Kerala	32.28%	41.27%	15.81%	6.63%	3.66%	0.35%
Madhya Pradesh	21.06%	54.16%	9.52%	11.13%	3.01%	1.13%
Maharashtra	30.69%	36.97%	12.45%	8.96%	7.84%	3.08%
Orissa	26.57%	43.29%	17.83%	8.25%	3.36%	0.70%
Puducherry	30.17%	32.02%	25.21%	8.06%	4.34%	0.21%
Punjab	21.44%	49.12%	14.00%	9.67%	4.16%	1.61%
Rajasthan	20.87%	49.03%	16.01%	8.12%	4.82%	1.15%
Sikkim	45.83%	0.00%	52.08%	0.00%	2.08%	0.00%
Tamil Nadu	29.18%	43.74%	13.47%	6.46%	6.14%	1.01%
Telangana	26.56%	42.16%	15.11%	7.61%	6.54%	2.01%
Uttar Pradesh	25.40%	47.72%	13.73%	8.73%	3.86%	0.56%
Uttaranchal	24.49%	47.52%	13.75%	10.58%	3.25%	0.41%
West Bengal	28.11%	37.61%	18.72%	6.92%	7.02%	1.62%

Figure 32 Percentage Distribution of Each Cluster per State

Restaurants by Country Cuisines

Below are the results of restaurant clustering based on area-specific cuisines that they offer. The resulting clusters are significantly homogenous ranging from restaurants that only serve Indian food, Chinese and Asian food, and the western food from Italy and America.

Cuisines	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Indian					
No	65.79%	100.00%	0.00%	8.72%	76.46%
Yes	34.21%	0.00%	100.00%	91.28%	23.54%
Chinese					
No	91.50%	92.07%	0.00%	75.47%	0.00%
Yes	8.50%	7.93%	100.00%	24.53%	100.00%
Asian					
No	98.71%	99.71%	100.00%	70.54%	0.00%
Yes	1.29%	0.29%	0.00%	29.46%	100.00%
Italian					
No	99.98%	0.00%	100.00%	0.00%	99.42%
Yes	0.02%	100.00%	0.00%	100.00%	0.58%
American					
No	100.00%	63.50%	100.00%	68.06%	100.00%
Yes	0.00%	36.50%	0.00%	31.94%	0.00%

Figure 33 Percentage Distribution of Each Country Cuisines per Cluster

Clusters	Labels	Characteristics
1	Classic Indian	1. They do not serve Italian and American food. 2. They are more likely to serve Indian food.
2	Westerners	1. They do not serve Indian food and rarely serve Chinese and Asian food. 2. They serve Italian and American food.
3	Indo-Chinese	1. They mostly serve Indian and Chinese food. 2. Most of them do not serve Asian, Italian and American food.
4	Around the World	1. They mostly serve Indian but also serve Chinese, Asian, Italian and American.
5	Asian Delights	1. They mostly serve Chinese and Asian food. 2. They do not serve Italian and American food. 3. Some serve Indian food.

Figure 34 Summary of Cluster Characteristics Considering Country Cuisines

The figure below shows the distribution of each cluster in each state. Most cluster of restaurants can be found in Delhi, Karnataka and Maharashtra.

State	Around the World	Asian Delights	Classic Indian	Indo-Chinese	Westerners
Andhra Pradesh	0.70%	0.38%	65.24%	26.91%	6.76%
Assam	0.84%	2.20%	59.64%	31.66%	5.66%
Bihar	0.61%	0.36%	58.96%	35.71%	4.36%
Chhattisgarh	1.44%	0.33%	76.27%	17.18%	4.77%
Delhi	1.46%	1.80%	64.75%	21.35%	10.63%
Goa	3.09%	1.78%	55.37%	30.05%	9.70%
Gujarat	1.37%	0.40%	73.16%	15.00%	10.08%
Himachal Pradesh	0.36%	1.09%	65.45%	25.09%	8.00%
Jammu and Kashmir	3.63%	0.81%	57.06%	24.40%	14.11%
Jharkhand	2.68%	0.74%	57.89%	32.13%	6.56%
Karnataka	1.64%	1.57%	62.70%	24.68%	9.41%
Kerala	0.40%	0.50%	71.39%	16.47%	11.24%

Madhya Pradesh	1.57%	0.41%	72.13%	19.83%	6.07%
Maharashtra	1.85%	2.14%	61.97%	25.50%	8.53%
Orissa	0.70%	1.47%	61.19%	31.47%	5.17%
Puducherry	2.48%	0.83%	64.67%	20.25%	11.78%
Punjab	2.27%	0.55%	63.52%	20.27%	13.39%
Rajasthan	2.34%	0.29%	72.81%	18.60%	5.95%
Sikkim	0.00%	4.17%	60.42%	22.92%	12.50%
Tamil Nadu	1.27%	0.76%	66.92%	22.16%	8.89%
Telangana	0.91%	0.71%	70.03%	21.02%	7.34%
Uttar Pradesh	1.26%	0.32%	69.29%	21.00%	8.14%
Uttaranchal	1.71%	1.06%	63.79%	25.39%	8.06%
West Bengal	1.42%	2.41%	66.16%	22.88%	7.12%

Figure 35 Distribution of Country Cuisine per State

Restaurants by Non-Country Cuisines

Below are the results of restaurant clustering based on non-area-specific cuisines that they offer. Latent class analysis technique was able to identify clusters almost uniformly.

Cuisines	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Continental					
No	90.91%	97.89%	100.00%	97.69%	81.25%
Yes	9.09%	2.11%	0.00%	2.31%	18.75%
Desserts					
No	100.00%	100.00%	0.00%	39.47%	89.69%
Yes	0.00%	0.00%	100.00%	60.53%	10.31%
Beverages					
No	100.00%	100.00%	71.98%	41.13%	6.65%
Yes	0.00%	0.00%	28.02%	58.87%	93.35%
Fast Food					
No	100.00%	26.89%	94.69%	16.37%	100.00%
Yes	0.00%	73.11%	5.31%	83.63%	0.00%
Seafood					
No	95.45%	99.43%	100.00%	100.00%	95.90%
Yes	4.55%	0.57%	0.00%	0.00%	4.10%

Figure 36 Distribution of Non-Country Cuisine per Cluster

Clusters	Labels	Characteristics
1	Western Seas	1. They only serve continental and seafood.
2	Fast Food	1. They are mostly fast food restaurants. 2. Some of them serve continental and seafood.
3	Sugar Rush	1. All of them serve dessert. 2. Some of them serve beverages.
4	Calorie Combo	1. Most of them are fast food restaurants. 2. They also serve dessert and beverages.
5	Thirst Quenchers	1. They mostly serve beverages. 2. Some of them serve continental and desserts.

Figure 37 Summary of Cluster Characteristics Considering Non-Country Cuisines

State	Calorie Combo	Fast Food	Sugar Rush	Thirst Quenchers	Western Seas
Andhra Pradesh	7.40%	11.22%	14.99%	5.87%	60.52%
Assam	4.51%	20.65%	7.65%	1.68%	65.51%
Bihar	6.05%	20.82%	12.11%	1.82%	59.20%

Chhattisgarh	9.98%	28.49%	8.98%	3.22%	49.33%
Delhi	8.91%	22.34%	7.90%	2.41%	58.45%
Goa	3.25%	10.23%	5.77%	2.46%	78.29%
Gujarat	8.39%	33.21%	15.52%	2.94%	39.94%
Himachal Pradesh	7.27%	16.00%	2.91%	1.45%	72.36%
Jammu and Kashmir	11.09%	23.19%	7.06%	4.23%	54.44%
Jharkhand	6.93%	17.08%	14.22%	3.60%	58.17%
Karnataka	8.91%	16.65%	12.51%	4.63%	57.30%
Kerala	3.51%	10.59%	8.84%	10.69%	66.37%
Madhya Pradesh	10.91%	31.68%	8.31%	3.13%	45.96%
Maharashtra	9.56%	24.69%	11.82%	2.73%	51.20%
Orissa	4.62%	22.52%	6.78%	2.31%	63.78%
Puducherry	8.26%	10.12%	9.30%	7.02%	65.29%
Punjab	9.93%	28.48%	8.25%	3.23%	50.11%
Rajasthan	8.08%	27.59%	7.58%	3.02%	53.73%
Sikkim	2.08%	22.92%	0.00%	0.00%	75.00%
Tamil Nadu	9.74%	17.44%	9.91%	4.42%	58.50%
Telangana	8.92%	17.60%	13.51%	4.89%	55.09%
Uttar Pradesh	6.59%	24.62%	10.68%	2.95%	55.16%
Uttaranchal	8.06%	23.11%	5.94%	2.77%	60.13%
West Bengal	4.10%	23.37%	8.06%	2.49%	61.98%

Figure 38 Distribution of Non-Country Cuisine per State

FUTURE WORK

This paper aims to generate business insights in the Indian restaurant market. Although the IDEA and cluster analysis provides useful information that could make potential business impact, several future works are still recommended.

1. Find possible datapoints that could be used for identifying the course of meals that each restaurant serves (i.e., breakfast, lunch, and dinner).
2. Instead of performing analysis at a country level to obtain an overall analysis of restaurants at a national level, one can analyze data on restaurants constrained to the city of interest.
3. Do further analysis to discover reasons behind association of states to restaurant ratings.
4. Explore on the possibility of using different modelling techniques to predict which restaurants would most likely get higher ratings considering the cuisines they serve, cost and amenities.

CONCLUSION

India seems to have a wide variety of preferences when it comes to restaurants. There are several important factors to be addressed by business venturers having the prospect of opening a restaurant in the country. The following are the trends in factors that we observed:

1. State
    a. The State a restaurant is found in has a significant effect on the future rating success that locality will have.
2. Cost
    a. The cost of a restaurant must be appropriate to its standards and the amenities it offers.
3. Cluster Trends
    a. It is important to serve customer needs as shown in the different clusters of restaurants shown in the analysis. Trying to serve cross-cluster with one restaurant can be dissatisfactory to all customer segments involved.

REFERENCES

- Business Insider India. (2019, March 29). If you're dreaming of opening a restaurant, here are the most popular foreign cuisines likely to work in India. Retrieved December 05, 2020, from https://www.businessinsider.in/slideshows/most-popular-foreign-foods-likely-to-work-in-india/slidelist/68628019.cms
- Gokul. (2020, March 19). Zomato India Restaurants(2 Lakh+ restaurants data). Retrieved December 06, 2020, from https://www.kaggle.com/ngokulakannan/zomato-india-restaurants2-lakh-restaurants-data
- Green, R., Milner, J., Joy, E. J., Agrawal, S., & Dangour, A. D. (2016). Dietary patterns in India: A systematic review. British Journal of Nutrition, 116(1), 142-148. doi:10.1017/S0007114516001598
- Haridwar, Uttarakhand, India. (n.d.). Retrieved December 06, 2020, from https://www.latlong.net/place/haridwar-uttarakhand-india-7228.html
- Igi. (n.d.). Mussoorie, India latitude longitude. Retrieved December 06, 2020, from https://latitudelongitude.org/in/mussoorie/
- Igi. (n.d.). Tiruchirappalli, India latitude longitude. Retrieved December 06, 2020, from https://latitudelongitude.org/in/tiruchirappalli/
- Junagadh, Gujarat, India. (n.d.). Retrieved December 06, 2020, from https://www.latlong.net/place/junagadh-gujarat-india-1049.html
- Latitude, Longitude and GPS coordinates of Goa. (n.d.). Retrieved December 06, 2020, from https://www.gps-latitude-longitude.com/gps-coordinates-of-goa
- Lobo, G. (2019, January). The transforming foodscape of India. Retrieved December 05, 2020, from https://www.ipsos.com/en/transforming-foodscape-india
- markbailey. (2019, September 10). Is there a Name and shape file for country INDIA (including its states) so that we can create Geo map using data for India. Retrieved December 06, 2020, from https://community.jmp.com/t5/Discussions/Is-there-a-Name-and-shape-file-for-country-INDIA-including-its/m-p/225228
- Neemrana, Rajasthan, India. (n.d.). Retrieved December 06, 2020, from https://www.latlong.net/place/neemrana-rajasthan-india-9930.html
- Population, total. (n.d.). Retrieved December 05, 2020, from https://data.worldbank.org/indicator/SP.POP.TOTL?end=2019
- Rishikesh, Uttarakhand, India. (n.d.). Retrieved December 06, 2020, from https://www.latlong.net/place/rishikesh-uttarakhand-india-665.html
- Russ_wolfinger. (2014, May 28). Venn Diagram. Retrieved December 04, 2020, from https://community.jmp.com/t5/JMP-Add-Ins/Venn-Diagram/ta-p/22390
- Siliguri, West Bengal, India. (n.d.). Retrieved December 06, 2020, from https://www.latlong.net/place/siliguri-west-bengal-india-18183.html
- Udupi, Karnataka, India. (n.d.). Retrieved December 06, 2020, from https://www.latlong.net/place/udupi-karnataka-india-19505.html
- World Cities Database. (n.d.). Retrieved December 06, 2020, from https://simplemaps.com/data/world-cities

ACKNOWLEDGEMENTS

We are deeply thankful to our professor, Dr. Kam Tin Seong[1], for giving us the skills and guidance needed to complete our research.

[1] Associate Professor of Information Systems (Practice) at SMU (Singapore Management University)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Banda Nikitha
Enterprise: Singapore Management University
E-mail: nikithab.2020@mitb.smu.edu.sg

Name: Escolano Cyrill Ostonal
Enterprise: Singapore Management University
E-mail: coescolano.2020@mitb.smu.edu.sg

Name: Prajs Sean Samuel
Enterprise: Singapore Management University
E-mail: ssprajs.2020@mitb.smu.edu.sg

ANNEX I

DATA PREPARATION CHANGE LOG- GENERAL CHANGES

S No	Data Table File	Column	Issue	Action
1.	DataSet_ZomatoIndia_tidy.jmp	cost_for_two	There exist outliers	The 10 records with the value for cost_for_two too small (2 to 15) or too large (>=10,000) were hid and excluded.
2.	DataSet_ZomatoIndia_tidy.jmp	New Columns: "Indian", "Continental", "Desserts" etc.	The column "cuisine" is a string of cuisines separated by "," and it was not possible to identify the cuisines that restaurant provides.	All the distinct cuisines were identified, grouped accordingly and added 14 new columns to identify the cuisines that the restaurant provides.
3.	DataSet_ZomatoIndia_tidy.jmp	"Indian", "Continental", "Desserts" etc.	To identify if that restaurant provides the cuisine or not	A similar formula was applied on all these newly created columns where the columns take a value of "1" if that restaurant provides that cuisine or else "0".
4.	DataSet_ZomatoIndia_tidy.jmp		Duplicate records	There exist one duplicate record that has been hid and excluded from all further analysis.
5.	DataSet_ZomatoIndia_tidy.jmp	zomato_url, telephone, address	Not required for the study	Hid and excluded from all further analysis purpose.
6.	DataSet_ZomatoIndia_tidy.jmp	"Indian", "Continental", "Desserts" etc.	Wrong data type and modeling type	Changed data type from "Numeric" to "character" and modeling type from "continuous" to "Nominal".
7.	DataSet_ZomatoIndia_tidy.jmp	New Column: "State"	State Locations of Restaurant needed for geographic clustering	Assigning a state to each city. Steps shown in table: "Creating a Map of India"
8.	DataSet_ZomatoIndia_tidy.jmp	New Column: "Region"	Grouping the states in India into 3 regions	3 main regions "South", "North" and "North East" were identified and using the formula, states were recognized where they belong and labelled accordingly.
9.	DataSet_ZomatoIndia_tidy.jmp		One of the records has a different "State" name	Recoded "Jharkahand" to "Jarkhand"

CREATING A MAP OF INDIA

Step	Action
1.	Download India Map files from https://community.jmp.com/t5/Discussions/Is-there-a-Name-and-shape-file-for-country-INDIA-including-its/m-p/225228
2.	Place XY and Name map shape files of India into SAS JMPPRO Map folder
3.	Create “India States.xlsx” with ‘City’ and ‘State’ columns for each city in data set
4.	Group city column in original data file and extract city names into “India States.xlsx”
5.	Assign correct state to each city in “India States.xlsx”
6.	Upload “India States.xlsx” into JMP to create “India States.jmp”
7.	Join original data set with “India States.jmp” by matching ‘City’ and ‘city’ columns to create “Data Set Zomato India.jmp”
8.	Move ‘State’ and ‘City’ column to after ‘name’ column and delete ‘city’ column to avoid duplicate columns

DATA PREPARATION CHANGE LOG – CLUSTER ANALYSIS

S No	Data Table File	Column	Issue	Action
1.	DataSet_ZomatoIndia_tidy_with_LCA.jmp	New jmp file		Using the “DataSet_ZomatoIndia_tidy” generated a new jmp file called “DataSet_ZomatoIndia_tidy_with_LCA.jmp”
2.	DataSet_ZomatoIndia_tidy_with_LCA.jmp	rating_count	Extremely low rating count	Hid and exclude rows with rating count less than 5
3.	DataSet_ZomatoIndia_tidy_with_LCA.jmp	rating	Ordinal data must be binned prior to latent class analysis	Use interactive binning and create a new column named “Log (rating) Groups.” Bins were divided in percentile.
4.	DataSet_ZomatoIndia_tidy_with_LCA.jmp	cost_for_two	Skewed distribution	Transform the column using log transformation to improve distribution.
5.	DataSet_ZomatoIndia_tidy_with_LCA.jmp	cost_for_two	Numeric, continuous data must be binned prior to latent class analysis	Use interactive binning and creating a new column named “Log (cost_for_two) Groups.” Bins were divided in percentile.

DATA PREPARATION CHANGE LOG – IDEA ANALYSIS

S No	Data Table File	Column	Issue	Action
1.	Urbanization.jmp	New jmp file	No columns that define the percentage of urbanization for each state	Created a new jmp file the consists of the states and their respective urbanization levels.
2.	DataSet_ZomatoINdia_tidy_IDEA	New jmp file		Joined ‘DataSet_ZomatoIndia_tidy.jmp’ with ‘Urbanization.jmp’ based on ‘State’=’State’ match to generate ‘DataSet_ZomatoIndia_tidy_IDEA’
3.	DataSet_ZomatoINdia_tidy_IDEA		Duplicate columns	Deleted one of the two ‘State’ columns from the jmp file as they are duplicate obtained by the join from the above mention step
4.	DataSet_ZomatoIndia_tidy_IDEA	rating	Group for cost analysis	Use interactive binning tool to split ratings into following groups in a new column “rating Groups”: 1. 1-2 2. 2-3 3. 3-4 4. 4-5 Note: There are no ratings registered below 1 or above 5
5.	CenterCoordinates_OfCities.jmp	New jmp file		Created a new jmp file containing the center of all distinct coordinates (latitude and longitude)

6.	DistanceFrom_CentreOfCity.jmp	New jmp file		Executed 'DistanceFrom_CentreOfCity_SQLQuery.jmpquery' query to extract limited variables from DataSet_ZomatoINdia_tidy file and joined it with CenterCoordinates_OfCities.jmp based on the criteria 'city=City'
7.	DistanceFrom_CentreOfCity	Distance to Town Centre (km)	Needed in categories for further analysis	Group into the following distances (km): • 0-2: Core • 2-5: Centre • 5-10: Near Centre • 10-20: Suburbs • 20+: Outskirts
8.	DistanceFrom_CentreOfCity			All records with either latitude =0 or longitude=0 or both latitude and longitude missing are hid and excluded from all further analysis.
9.	DataSet_ZomatoIndia_tidy_IDEA	rating_count	Extremely low rating count	Hid and exclude rows with rating count less than 5
10.	N/A	N/A	N/A	Download and install venn diagram add-in from https://community.jmp.com/t5/JMP-Add-Ins/Venn-Diagram/ta-p/22390
11.	DataSet_ZomatoIndia_tidy.jmp_IDEA	online_order	Need to prepare for venn diagram	Recode into online_order recoded with True =1 and False = 0
12.	DataSet_ZomatoIndia_tidy.jmp_IDEA	table_reservation	Need to prepare for venn diagram	Recode into table_reservation recoded with True =1 and False = 0
13.	DataSet_ZomatoIndia_tidy.jmp_IDEA	delivery_only	Need to prepare for venn diagram	Recode into delivery_only recoded with True =1 and False = 0
14.	DataSet_ZomatoIndia_tidy.jmp_IDEA	VennSector, VennXMiss, VennYMiss, VennX, VennY	Implementation of Venn Diagram add-in	Venn Diagram analysis creates these five columns to be able to create the venn diagram graph used

ANNEX II

INDIAN STATES BY REGION