Institute of Technology of Cambodia
Department of Applied Mathematics and Statistics

Restaurant Reviews / Zomato Dataset

Lecturer:

Course: **M. OL Say**

TP: **M. Min Sothearith**

Presented by group 9 :
Dataset:
https://www.kaggle.com/datasets/shrutimehta/zomato-restaurants-data

Group members:

KHUN Limchheang	e20230393
CHHAY Lyveng	e20230135

Contents

Introduction

Data Description

Data Cleaning Method

EDA and key insights

Conclusions

I. Introduction

- Brief overview of the dataset
  - Dataset: Zomato Restaurant Dataset (Kaggle)
  - Contains restaurant information across multiple countries, including location, pricing, services, and customer ratings
- Why this analysis is important
  - Helps understand customer preferences and service trends in the restaurant industry
  - Supports data-driven insights for business and platform decision-making

Objective

- Objective of This Project:
  - To analyze the distribution of restaurants across different countries and cities
  - To examine how restaurant ratings vary by price range
  - To investigate whether online delivery and table booking services are associated with higher customer ratings
  - To identify key factors that influence customer ratings in the Zomato dataset

II. Data Description

Initial Observation

- Distribution of restaurant ratings
- Presence/absence of missing values
- Early patterns noticed before cleaning

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9551 entries, 0 to 9550
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Restaurant ID         9551 non-null   int64
 1   Restaurant Name       9551 non-null   object
 2   Country Code          9551 non-null   int64
 3   City                  9551 non-null   object
 4   Address               9551 non-null   object
 5   Locality              9551 non-null   object
 6   Locality Verbose      9551 non-null   object
 7   Longitude             9551 non-null   float64
 8   Latitude              9551 non-null   float64
 9   Cuisines              9542 non-null   object
 10  Average Cost for two  9551 non-null   int64
 11  Currency              9551 non-null   object
 12  Has Table booking     9551 non-null   object
 13  Has Online delivery   9551 non-null   object
 14  Is delivering now     9551 non-null   object
 15  Switch to order menu  9551 non-null   object
 16  Price range           9551 non-null   int64
 17  Aggregate rating      9551 non-null   float64
 18  Rating color          9551 non-null   object
 19  Rating text           9551 non-null   object
 20  Votes                 9551 non-null   int64
dtypes: float64(3), int64(5), object(13)
memory usage: 1.5+ MB
```

- shape of dataset (9551,21):
  - Number of rows: 9,551 restaurants
  - Number of columns: 21 restaurant attributes

```
df.shape
```

```
(9551, 21)
```

- Data types:
  - Categorical variables: Country, City, Cuisines, Currency, Price Range, Online Delivery, Table Booking
  - Numerical variables: Aggregate Rating, Votes, Average Cost for Two, Latitude, Longitude

Column Name	Description
Restaurant ID	Unique identifier for each restaurant
Restaurant Name	Name of the restaurant
Country Code	Country identifier
City	City where the restaurant is located
Address	Full address
Locality	Area or neighborhood
Locality Verbose	Detailed locality description
Longitude	Geographic longitude
Latitude	Geographic latitude
Cuisines	Types of food offered
Average Cost for two	Estimated cost for two people
Currency	Currency used
Has Table booking	Whether table booking is available
Has Online delivery	Whether online delivery is available
Is delivering now	Current delivery status
Switch to order menu	Online menu availability
Price range	Price category (1 = low, 4 = high)
Aggregate rating	Average customer rating
Rating color	Color representing rating
Rating text	Textual rating (Excellent, Good, Poor)
Votes	Number of user votes

Descriptive Statistics

- Count
- mean
- standard deviation
- minimum
- Q1 or 25%
- median
- Q3 or 75%
- maximum

	Restaurant ID	Country Code	Longitude	Latitude	Average Cost for two	Price range	Aggregate rating	Votes
count	9.551000e+03	9551.000000	9551.000000	9551.000000	9551.000000	9551.000000	9551.000000	9551.000000
mean	9.051128e+06	18.365616	64.126574	25.854381	1199.210763	1.804837	2.666370	156.909748
std	8.791521e+06	56.750546	41.467058	11.007935	16121.183073	0.905609	1.516378	430.169145
min	5.300000e+01	1.000000	-157.948486	-41.330428	0.000000	1.000000	0.000000	0.000000
25%	3.019625e+05	1.000000	77.081343	28.478713	250.000000	1.000000	2.500000	5.000000
50%	6.004089e+06	1.000000	77.191964	28.570469	400.000000	2.000000	3.200000	31.000000
75%	1.835229e+07	1.000000	77.282006	28.642758	700.000000	2.000000	3.700000	131.000000
max	1.850065e+07	216.000000	174.832089	55.976980	800000.000000	4.000000	4.900000	10934.000000

III. Data Cleaning

Purpose of Data Cleaning:

- Improve data consistency and readability
- Ensure accurate analysis and visualization
- Prepare the dataset for exploratory data analysis

Data Quality Checks Performed

- Missing values
- Duplicate records
- Column naming consistency
- Country Merge
- Outliers Detection

Tools to uses:

- pandas
- numpy

• Handling Missing Values

```
df.isna().sum()
```

```
restaurant_id           0
restaurant_name         0
country_code            0
city                    0
address                 0
locality                0
locality_verbose        0
longitude               0
latitude                0
cuisines                9
average_cost_for_two    0
currency                0
has_table_booking       0
has_online_delivery     0
is_delivering_now       0
switch_to_order_menu    0
price_range             0
aggregate_rating        0
rating_color            0
rating_text             0
votes                   0
dtype: int64
```

Since missng is less than 1% of rows, then we can drop it using dropna

```
df = df.copy()
df.dropna(inplace=True)
df.shape
```

```
(9542, 21)
```

- # Check Duplicate

```
df.duplicated().sum()
```

```
np.int64(0)
```

No duplicate records were found, so no rows were removed.

• Column Name Standardization

```python
df.columns = (
    df.columns
    .str.strip()
    .str.lower()
    .str.replace(" ", "_")
)
df.columns
```

```
Index(['restaurant_id', 'restaurant_name', 'country_code', 'city', 'address',
       'locality', 'locality_verbose', 'longitude', 'latitude', 'cuisines',
       'average_cost_for_two', 'currency', 'has_table_booking',
       'has_online_delivery', 'is_delivering_now', 'switch_to_order_menu',
       'price_range', 'aggregate_rating', 'rating_color', 'rating_text',
       'votes', 'country', 'price_category', 'cuisine_list'],
      dtype='object')
```

• Country Merge

	Restaurant ID	Restaurant Name	Country Code	City	Address
9546	5915730	NamlÛ± Gurme	208	ÛÁstanbul	Kemanke□ô Karamustafa Pa□ôa Mahallesi, RÛ±htÛ±...
9547	5908749	Ceviz AÛôacÛ±	208	ÛÁstanbul	Ko□ôuyolu Mahallesi, Muhittin îstí_ndaÛô Cadd...
9548	5915807	Huqqa	208	ÛÁstanbul	Kuruí_e□ôme Mahallesi, Muallim Naci Caddesi, N...
9549	5916112	A□ô□ôk Kahve	208	ÛÁstanbul	Kuruí_e□ôme Mahallesi, Muallim Naci Caddesi, N...

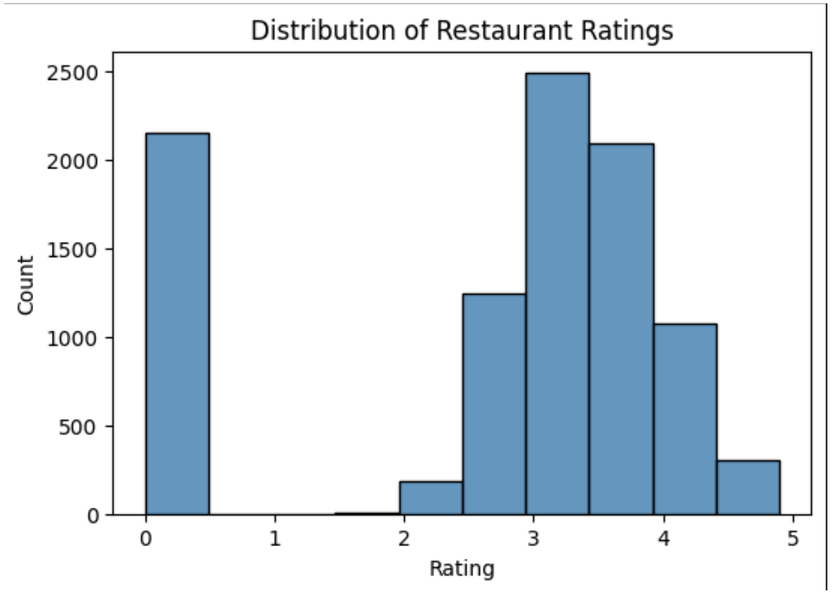
	country_code	country
0	162	Phillipines
1	162	Phillipines
2	162	Phillipines
3	162	Phillipines
4	162	Phillipines

• Outlier Detection

Column	Handle Outliers?	Reason
`average_cost_for_two`	Yes	Extremely high prices skew plots
`votes`	Yes	Few restaurants have huge votes
`aggregate_rating`	No	Ratings are bounded (0–5)
`price_range`	No	Categorical (1–4)
`latitude` / `longitude`	No	Geographic, not outliers
`country_code`	No	Identifier

IV. EDA and key Insights

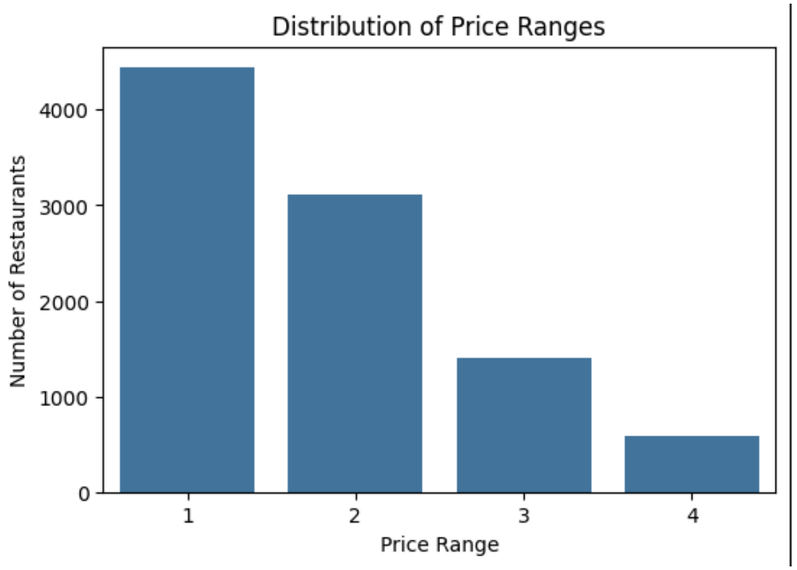
1. Rating Distribution



Histogram shows most ratings are between 3–4.

Insight: customers generally rate restaurants positively

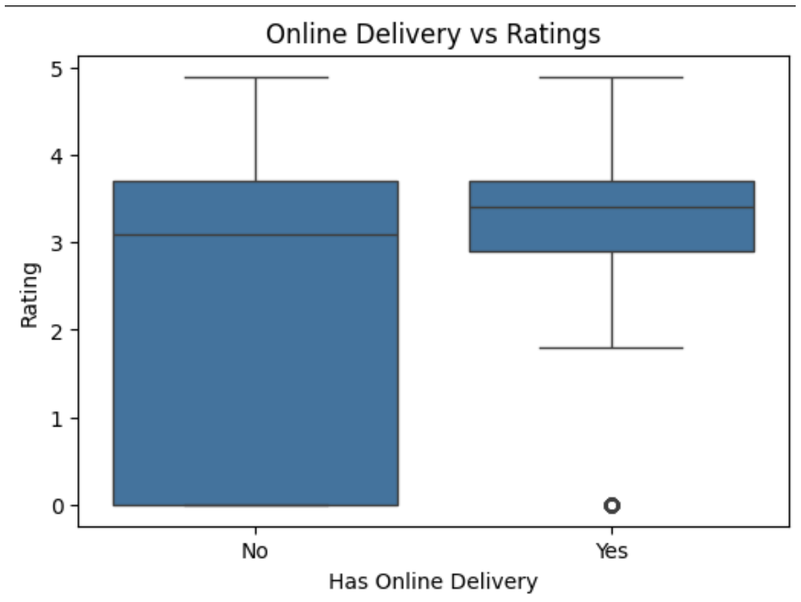
2. Price Range Distribution



Count plot shows most restaurants are in price range 1 (low cost).

Insight: most restaurants are at low cost

3. Online Delivery vs. Rating



Boxplot compares ratings between restaurants with/without online delivery.

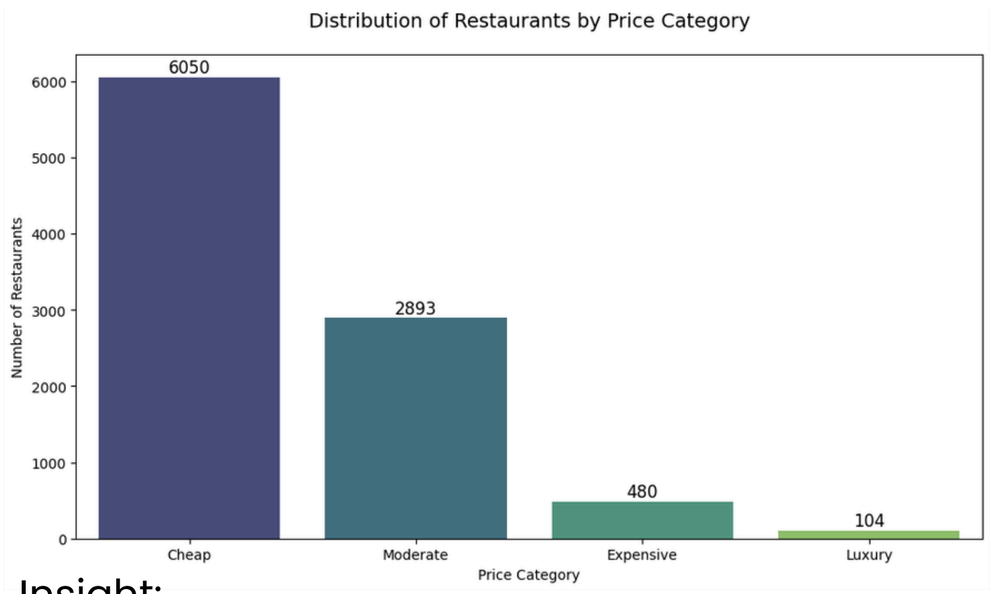
Insight:

- most restaurants have no online delivery
- tend to have lower ratings

4. Price Category Creation

```
df['price_category'] = pd.cut(
    df['average_cost_for_two'],
    bins=[0,500,1500,3000,10000],
    labels=['Cheap','Moderate','Expensive','Luxury']
)
```

Created price_category (Cheap, Moderate, Expensive, Luxury) based on cost for two.



Insight:

most restaurants for average cost for two people are cheap

5. Country-Level Summary

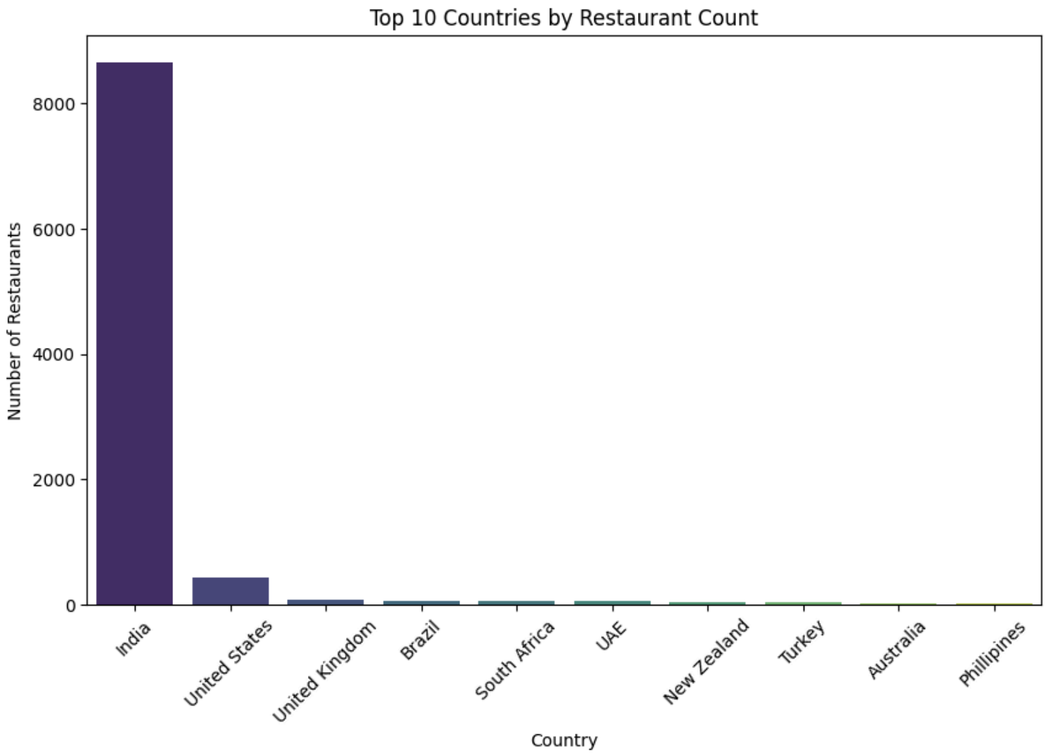
- Created country_summary DataFrame with counts, percentages, avg rating, avg cost.

country	restaurant_count	percentage	avg_rating	avg_cost_for_two
Australia	24	0.25	3.66	24.08
Brazil	60	0.63	3.76	134.67
Canada	4	0.04	3.58	36.25
India	8652	90.67	2.52	614.83
Indonesia	21	0.22	4.30	3418.00
New Zealand	40	0.42	4.26	69.75
Phillipines	22	0.23	4.47	1463.00
Qatar	20	0.21	4.06	223.75
Singapore	20	0.21	3.58	155.75
South Africa	60	0.63	4.21	419.73
Sri Lanka	20	0.21	3.87	2258.60
Turkey	34	0.36	4.30	84.85
UAE	60	0.63	4.23	166.42
United Kingdom	80	0.84	4.09	47.81
United States	425	4.45	4.01	26.35

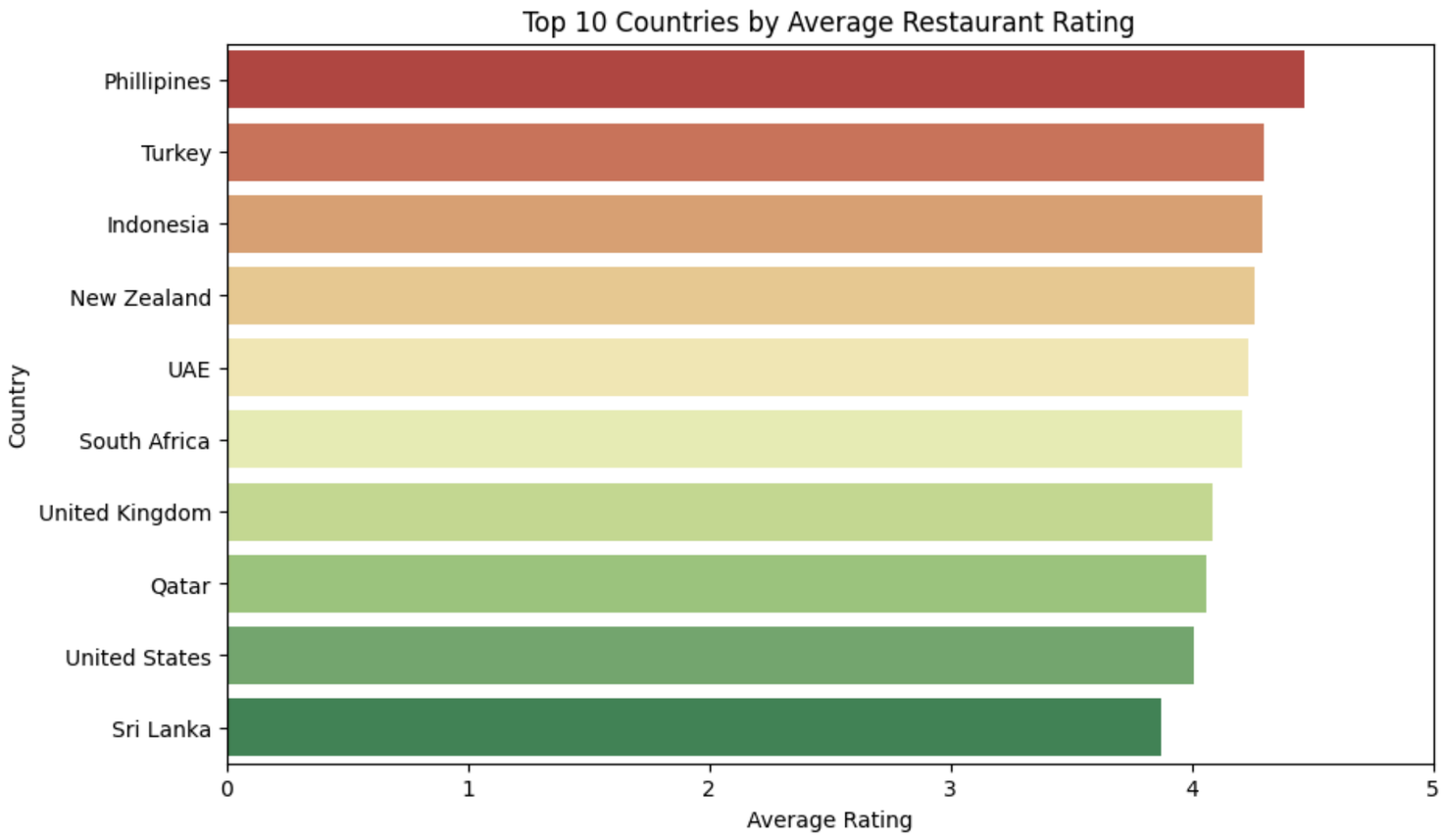
```
Total countries in dataset: 15
Country with most restaurants: India (8652 restaurants)
Country with fewest restaurants: Canada (4 restaurants)
```

insight:

most restaurants are in India.



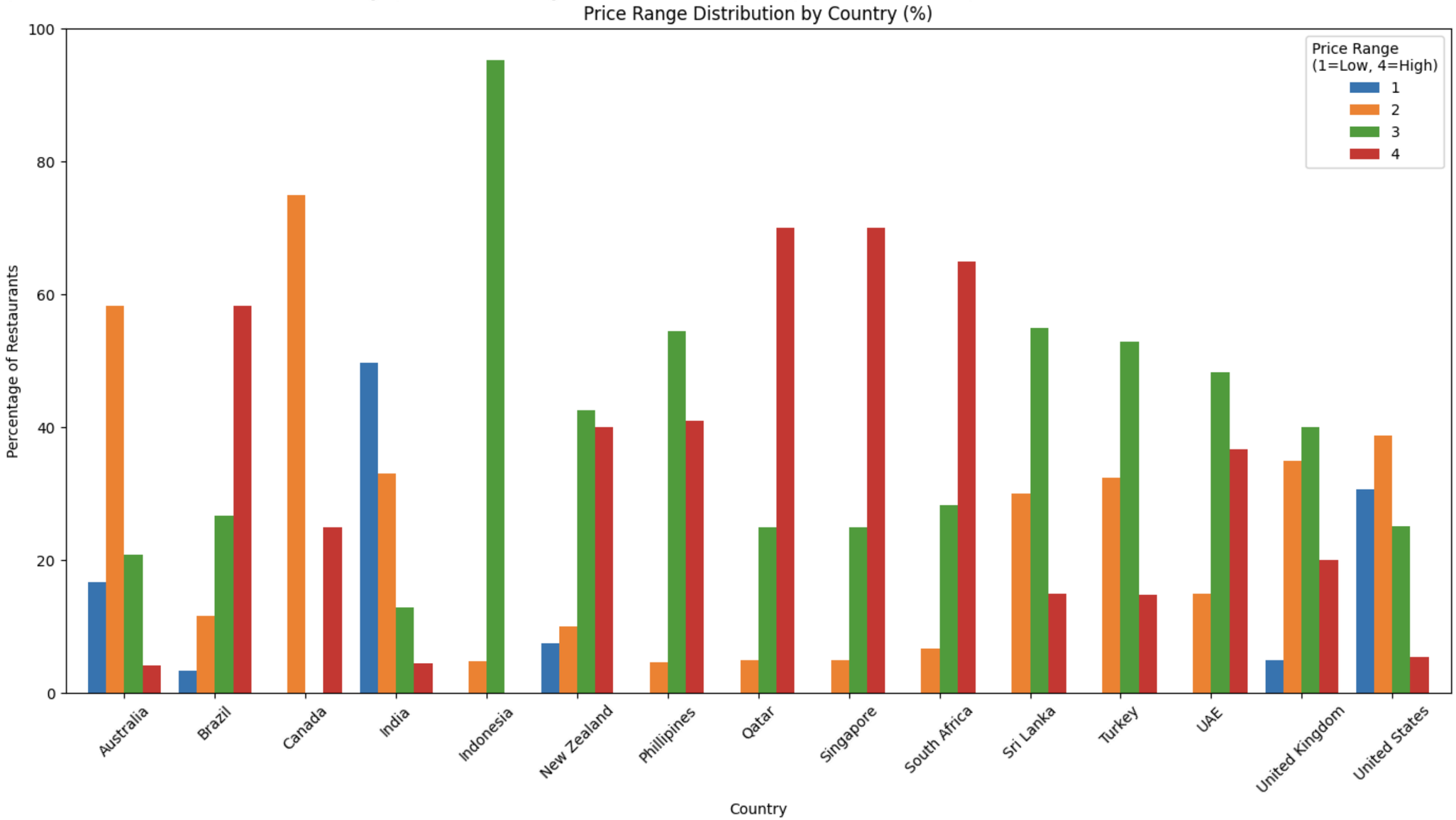
- Top Countries by average rating



Phillipines has the most rating

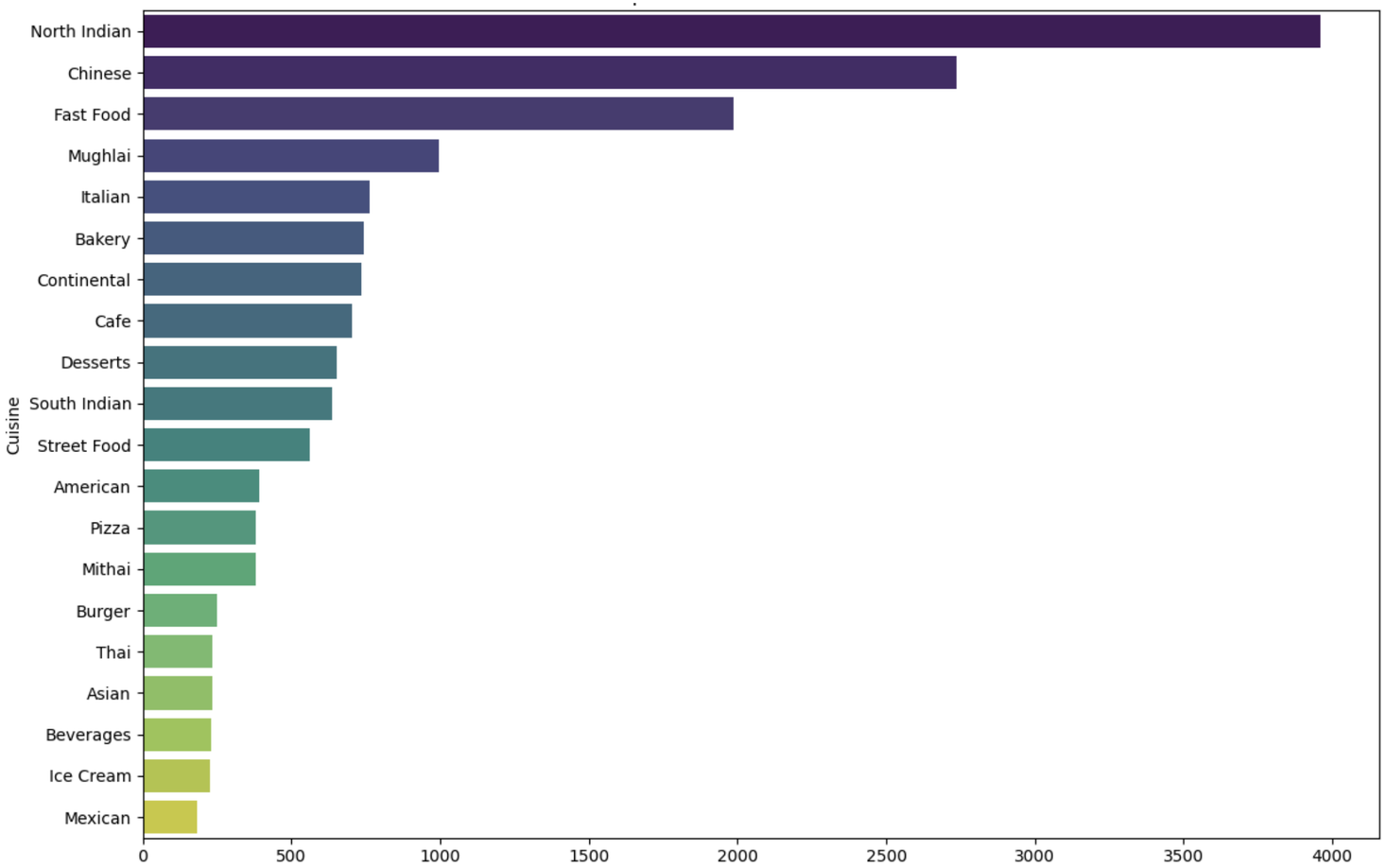
6. Price Range by Country

Grouped bar chart showing price range distribution per country.

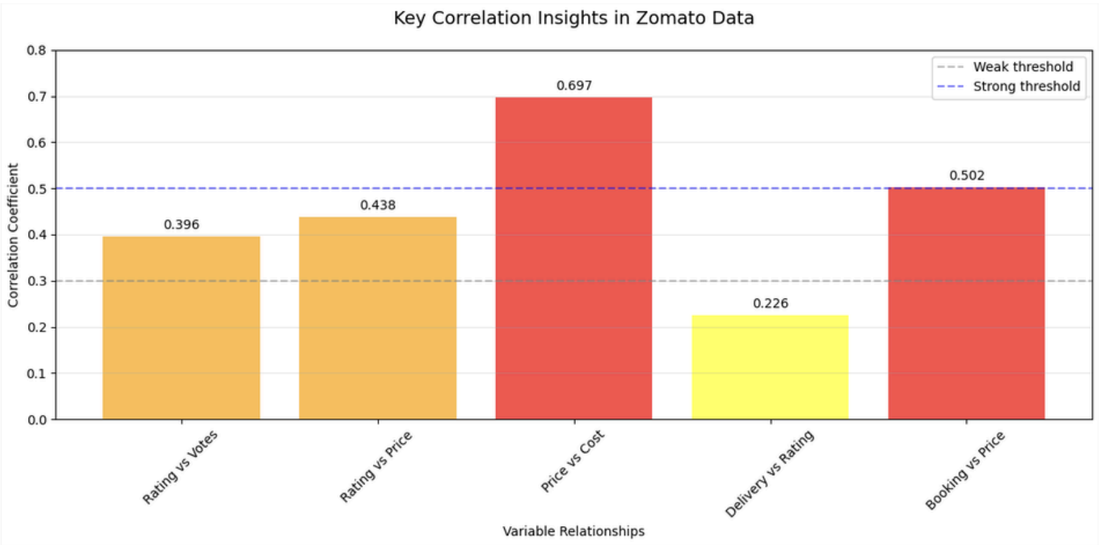


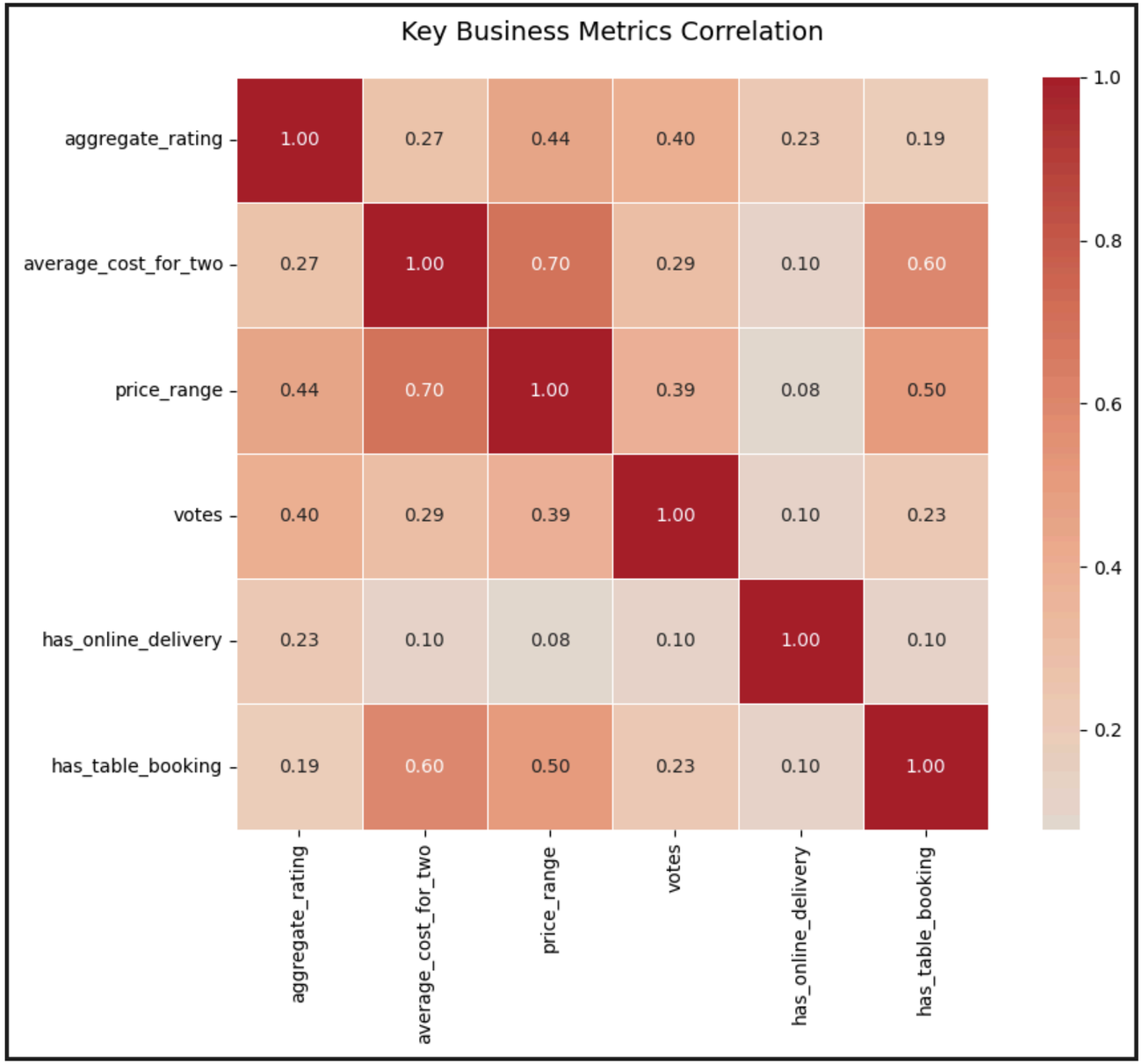
7. Cuisine Analysis

Top 20 cuisines visualized.



8. Correlation Analysis





Heatmap of numerical feature correlations.

Key insights:

- Rating vs. Votes: Moderate positive (0.396)
- Rating vs. Price Range: Moderate positive (0.438)
- Price Range vs. Cost: Strong positive (0.697)
- Online Delivery vs. Rating: Weak positive (0.226)
- Table Booking vs. Price Range: Strong positive (0.502)

VI. Conclusions

- Location: 15 countries, 141 cities analyzed → geographic patterns identified.
- Cuisine: 141 unique cuisines → popularity and pricing relationships mapped.
- Pricing: Strong correlation between price range and ratings when quality matches cost.
- Services: Delivery has slight positive impact; table booking strongly linked to higher-priced restaurants.
- Final Verdict: Objective fully met → analysis quantified how each factor influences restaurant success.

Thank You

For Your Attention