**INT375**(Data Science Toolbox: Python Programming)

PROJECT REPORT

ON

**From Ratings to Revenue: Zomato Dataset Breakdown**

**Name - Shubhanshu Singh Chauhan**

**Registration No- 12301386**

**Programme – B.Tech (CSE)**

**Section -K23SG**

**Under the Guidance of**

**Faculty Name- Dr. Manpreet Singh Sehgal**

**UID No- 32354**

**Discipline of CSE**

**School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **Shubhanshu Singh Chauhan** bearing Registration no. **12301386** has completed **INT375** (Data Science Toolbox: Python Programming) project titled- **From Ratings to Revenue: Zomato Dataset Breakdown** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

School of Computer Science

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, **Shubhanshu Singh Chauhan**, student of B.Tech. under CSE Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date – 12/04/2025 Signature

Registration no. -12301386 Name – Shubhanshu Singh Chauhan

**Acknowledgement**

I would like to extend my sincere appreciation to anyone who assisted and provided useful guidance in this process of the project. Firstly, I want to thank my mentor/supervisor Dr. Manpreet Singh Sehgal for his insightful comments and feedback, which influenced the direction of this research. Further, I wish to express my appreciation to the data providers and organizations that provided the dataset, which allowed for an investigation into the allocation of public funding, across regions, organizations, and demographic characteristics. I also acknowledge all the developers and contributors of open-source software, such as Pandas, Matplotlib, and Seaborn, which provided outstanding functionality to assist my data cleaning, statistical calculations, and visualizations. Finally, I thank the on-going encouragement of peers, family, and friends, who have provided feedback and encouragement throughout the various stages of the research project.

Date – 12/04/2025

Registration no. -12301386

Name – Shubhanshu Singh Chauhan

**1.Introduction**

The restaurant industry thrives on customer satisfaction, pricing strategies, and regional preferences, making data-driven insights critical for stakeholders. This project analyzes the Zomato dataset to uncover trends in restaurant ratings, pricing, cuisines, predictive modeling, and geographic distribution. Using Python and its powerful libraries, the analysis addresses five key objectives:

1. **Analyze Restaurant Ratings and Popularity**: Investigate the relationship between aggregate ratings and customer votes to understand popularity drivers.
2. **Investigate the Impact of Price Range on Customer Experience**: Examine how price range influences ratings and service availability.
3. **Explore Cuisine Preferences Across Cities or Countries**: Identify dominant cuisines in top regions to reveal culinary trends.
4. **Predict Restaurant Ratings Using Machine Learning**: Build a predictive model for ratings based on features like price and votes.
5. **Geographical Analysis of Restaurant Distribution**: Map restaurant locations to study spatial patterns and their relation to ratings.

By leveraging exploratory data analysis (EDA), statistical modeling, and visualizations, this project aims to provide actionable insights for restaurant owners, marketers, and policymakers, making complex data accessible and meaningful.

**2.** **Source of dataset**

This project analyzes the Zomato Dataset, which includes restaurant details such as names, locations, cuisines, ratings, votes, price ranges, and service availability, sourced from an csv file provided for the course.

**Dataset link**: <https://github.com/MainakRepositor/Datasets/blob/master/zomato.csv>

**3.** **Dataset Preprocessing**

The dataset was loaded into Python using Pandas and cleaned to ensure accuracy. Key preprocessing steps included:

* Importing the Excel file (Zomato Dataset 2.xlsx) and removing unnamed columns.
* Dropping rows with missing values in critical columns (Aggregate rating, Votes, Cuisines, Average Cost for two, Price range, Longitude, Latitude).
* Removing outliers in Votes (top 1%) to reduce skew.
* Encoding categorical variables (e.g., Has Table booking, Has Online delivery) as binary (0/1) for machine learning.
* Creating a derived feature, Cuisine Count, to capture the number of cuisines per restaurant.

These steps ensured a clean, structured dataset suitable for analysis and modeling, aligning with CO2 (Unit II) of the syllabus.

**4. Analysis on dataset**

1. **General Description**

This project explores restaurant trends using the Zomato dataset, focusing on customer ratings, pricing impacts, cuisine preferences, predictive modeling, and geographic distribution. Python libraries—Pandas, Matplotlib, Seaborn, scikit-learn, and Folium—were used for data cleaning, EDA, visualization, and machine learning, addressing all course outcomes (CO1–CO6). The analysis provides insights into customer behavior and market dynamics, supporting strategic decision-making.

1. **Specific Requirements**

To execute this project, the following tools were required:

* **Software & Libraries**:
  + Python 3.x
  + Google Colab
  + Libraries: Pandas, Matplotlib, Seaborn, scikit-learn, Folium, WordCloud, openpyxl
* **Hardware**:
  + Computer with sufficient processing power (e.g., Mac or PC).

1. **Analysis Results & Visualization**

Below are the analyses for each objective, including code and placeholders for graph screenshots.

**Problem 1: Analyze Restaurant Ratings and Popularity**

**Problem Statement**: How are restaurant ratings distributed, and is there a relationship between ratings and the number of votes received?

**Methodology**:

* Used Seaborn to create a scatter plot of Aggregate rating vs. Votes.
* Analyzed the distribution and correlation to assess popularity trends (CO3, CO4).

**Python Code**:

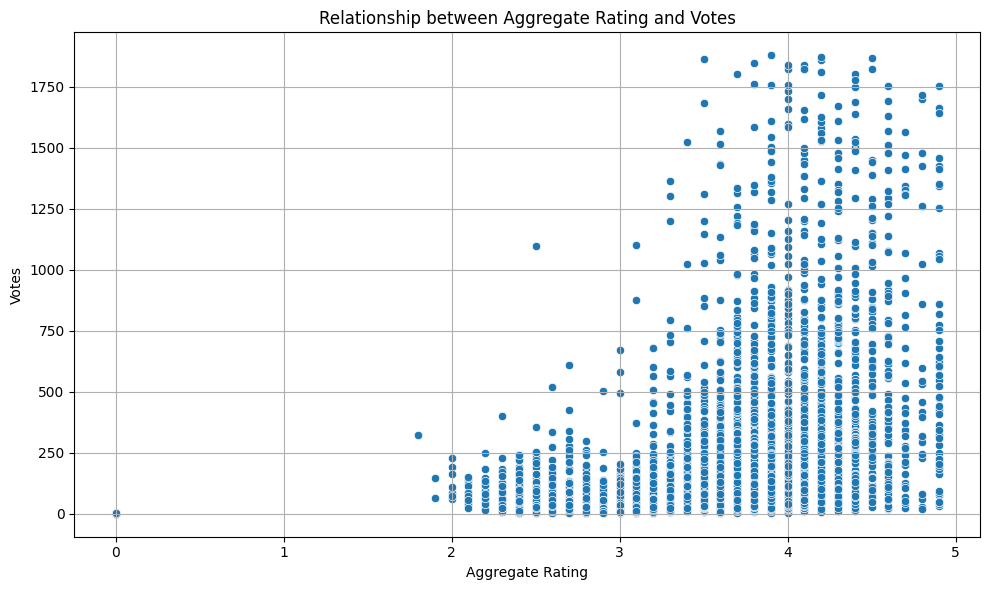
plt.figure(figsize=(10, 6))

sns.scatterplot(data=df\_cleaned, x='Aggregate rating', y='Votes')

plt.title('Relationship between Aggregate Rating and Votes')

plt.xlabel('Aggregate Rating')

**Graph Screenshot**:



**Problem 2: Investigate the Impact of Price Range on Customer Experience**

**Problem Statement**: Does price range influence customer ratings, and how does it relate to service availability?

**Methodology**:

* Created a box plot using Seaborn to compare Aggregate rating across Price range categories.
* Assessed variability and medians to understand pricing’s impact (CO3, CO4, CO5).

**Python Code**:

plt.figure(figsize=(10, 6))

sns.boxplot(data=df\_cleaned, x='Price range', y='Aggregate rating')

plt.title('Aggregate Rating vs Price Range')

plt.xlabel('Price Range')

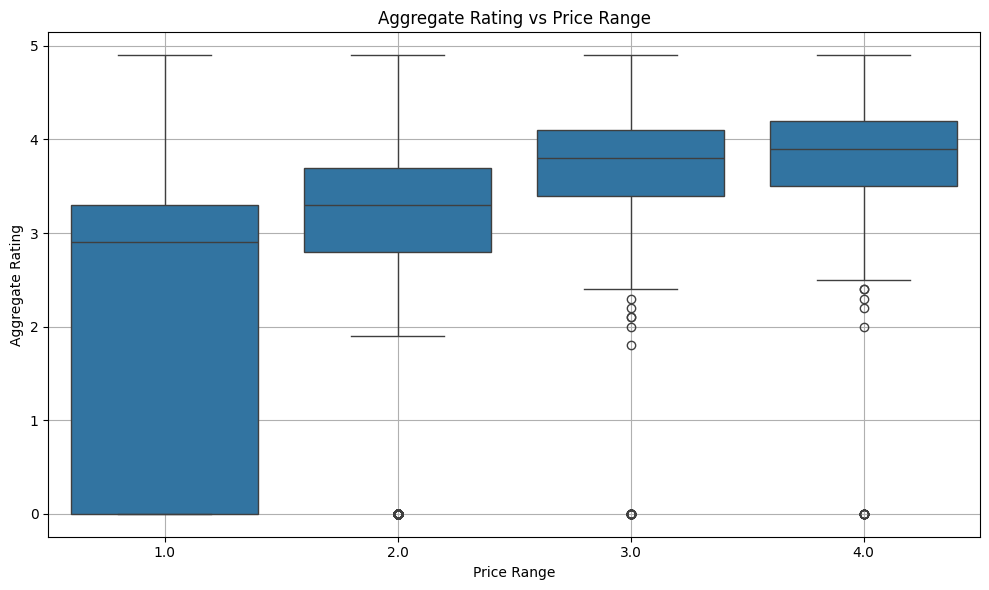
plt.ylabel('Aggregate Rating')

plt.grid(True)

plt.tight\_layout()

plt.show()

**Graph Screenshot**:



**Problem 3: Explore Cuisine Preferences Across Cities or Countries**

**Problem Statement**: Which cuisines are most popular in top countries?

**Methodology**:

* Filtered the top three countries by restaurant count.
* Split and counted cuisines, visualizing the top 20 with a bar plot (CO2, CO3, CO4).

**Python Code**:

top\_countries = df\_cleaned['Country Code'].value\_counts().head(3).index

cuisine\_data = df\_cleaned[df\_cleaned['Country Code'].isin(top\_countries)]

cuisine\_counts = cuisine\_data['Cuisines'].str.split(',').explode().str.strip().value\_counts().head(20)

plt.figure(figsize=(12, 6))

sns.barplot(x=cuisine\_counts.values, y=cuisine\_counts.index)

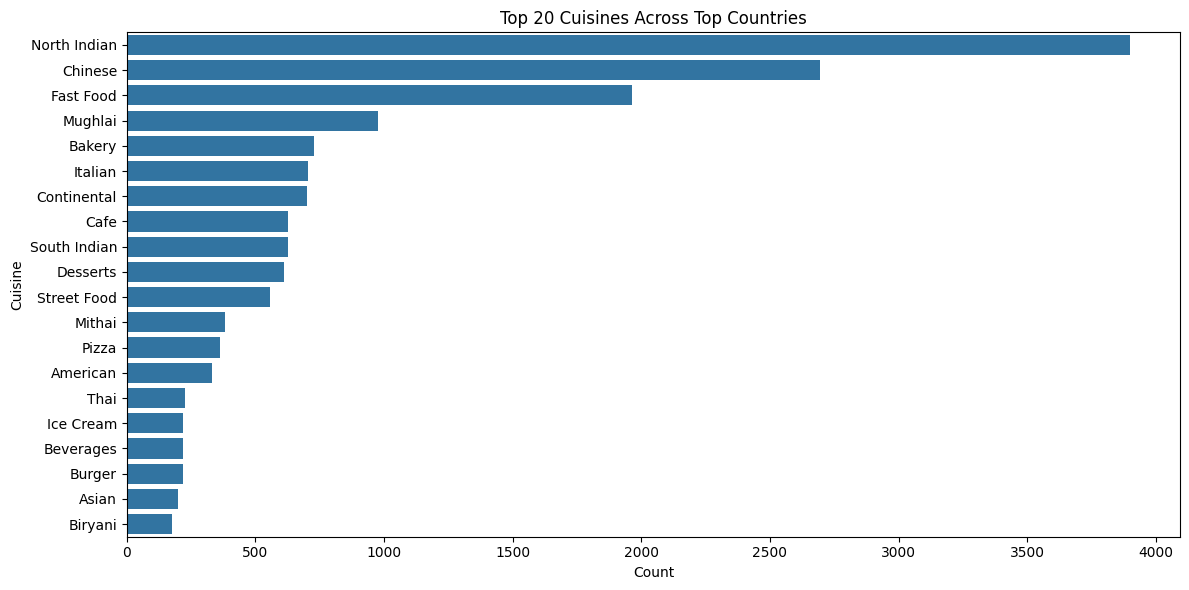
plt.title('Top 20 Cuisines Across Top Countries')

plt.xlabel('Count')

plt.ylabel('Cuisine')

plt.tight\_layout()

plt.show()



**Problem 4: Predict Restaurant Ratings Using Machine Learning**

**Problem Statement**: Can we predict restaurant ratings based on features like price range, votes, and services?

**Methodology**:

* Prepared features (Price range, Average Cost for two, Has Table booking, Has Online delivery, Votes, Cuisine Count).
* Scaled features and split data into train/test sets.
* Trained a RandomForestRegressor and evaluated performance (MSE, R²) (CO6).
* Visualized feature correlations with a heatmap (CO3).

**Python Code**:

# --- Correlation heatmap ---

plt.figure(figsize=(10, 6))

sns.heatmap(df\_cleaned[['Aggregate rating', 'Votes', 'Price range', 'Average Cost for two']].corr(), annot=True, cmap='coolwarm')

plt.title('Feature Correlation Heatmap')

plt.tight\_layout()

plt.show()

# --- Prepare data for ML ---

df\_ml = df\_cleaned.copy()

df\_ml['Has Table booking'] = df\_ml['Has Table booking'].map({'Yes': 1, 'No': 0})

df\_ml['Has Online delivery'] = df\_ml['Has Online delivery'].map({'Yes': 1, 'No': 0})

df\_ml['Cuisine Count'] = df\_ml['Cuisines'].apply(lambda x: len(str(x).split(',')))

features = ['Price range', 'Average Cost for two', 'Has Table booking',

'Has Online delivery', 'Votes', 'Cuisine Count']

X = df\_ml[features]

y = df\_ml['Aggregate rating']

# --- Feature scaling ---

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# --- Train/test split ---

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_scaled, y, test\_size=0.2, random\_state=42)

# --- Model training and evaluation ---

model = RandomForestRegressor(random\_state=42)

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

print("Mean Squared Error (MSE):", mean\_squared\_error(y\_test, y\_pred))

print("R² Score:", r2\_score(y\_test, y\_pred))

**Graph Screenshot**:



**Problem 5: Geographical Analysis of Restaurant Distribution**

**Problem Statement**: How are restaurants distributed geographically, and do ratings vary by location?

**Methodology**:

* Sampled 1000 restaurants to create an interactive map with Folium.
* Visualized locations with circle markers, scaled by votes and colored by ratings (CO3, CO4).
* Added popups with restaurant details.

**Python Code**:

import branca.colormap as cm

from folium.plugins import MarkerCluster

# ... (Rest of your code) ...

map\_sample = df.sample(n=1000, random\_state=42) # Adjust n for desired sample size

min\_rating = map\_sample['Aggregate rating'].min()

max\_rating = map\_sample['Aggregate rating'].max()

colormap = cm.linear.RdYlGn\_09.scale(min\_rating, max\_rating)

colormap.caption = 'Aggregate Rating Scale'

map\_obj = folium.Map(location=[map\_sample['Latitude'].mean(), map\_sample['Longitude'].mean()], zoom\_start=2)

marker\_cluster = MarkerCluster().add\_to(map\_obj)

for \_, row in map\_sample.iterrows():

rating = row['Aggregate rating']

radius = row['Votes'] / map\_sample['Votes'].max() \* 10

popup\_text = f"""

<b>{row['Restaurant Name']}</b><br>

Rating: {rating}<br>

Votes: {row['Votes']}<br>

Cuisine: {row['Cuisines']}

"""

folium.CircleMarker(

location=[row['Latitude'], row['Longitude']],

radius=radius,

color='black', # Stroke color

weight=1, # Stroke width

fill=True,

fill\_color=colormap(rating),

fill\_opacity=0.8,

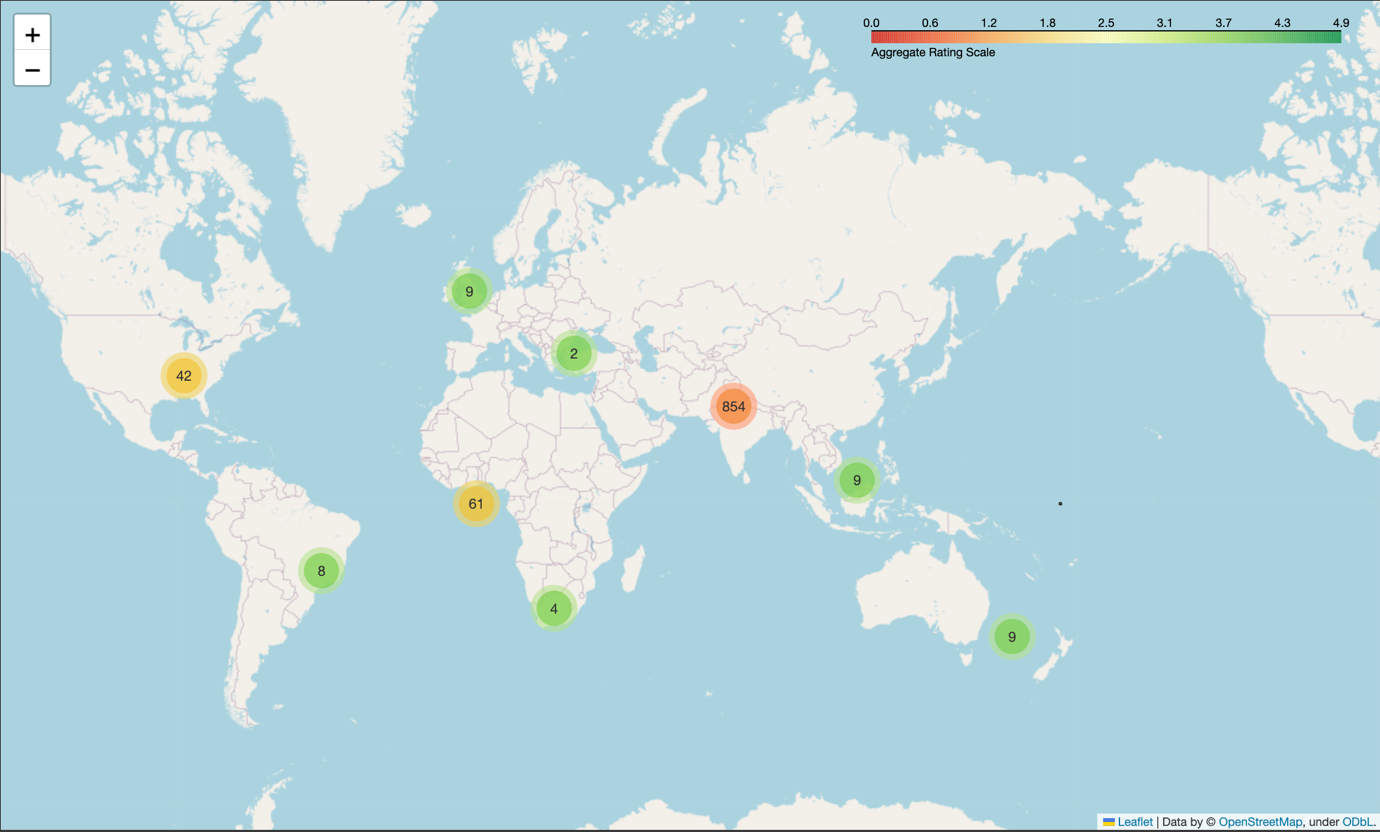
popup=folium.Popup(popup\_text, max\_width=300)

).add\_to(marker\_cluster)

colormap.add\_to(map\_obj)

display(map\_obj)

**Graph Screenshot**:



**5)Conclusion**

This project successfully analyzed the Zomato dataset to uncover trends in restaurant ratings, pricing, cuisines, predictive modeling, and geographic distribution. Using Python libraries like Pandas, Matplotlib, Seaborn, scikit-learn, and Folium, the analysis addressed all objectives, revealing key insights into customer preferences and market dynamics. The scatter plots, box plots, bar charts, heatmaps, and interactive maps made complex data accessible, supporting stakeholders in strategic decision-making. The project highlights the power of data science in transforming raw data into actionable insights, aligning with the course’s focus on Python programming and analytics.

**6) References**

 The Pandas Development Team. pandas: powerful Python data analysis toolkit. <https://pandas.pydata.org/>

 Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. <https://matplotlib.org/>

 Waskom, M. et al. Seaborn: Statistical Data Visualization. <https://seaborn.pydata.org/>

 scikit-learn: Machine Learning in Python. <https://scikit-learn.org/>

 Folium: Python Data, Leaflet.js Maps. <https://python-visualization.github.io/folium/>

 Python Software Foundation. Python Language Reference, version 3.x. <https://www.python.org/>