ZOMATO DATA ANALYSIS

An Exploratory and Visualization Project

Team Member:
Priyanshu Sharma:
23SCSE1011750
Sanskriti Shukla:
23SCSE1011043

Submitted to:
Dr.Rakesh Sahu

Project Overview

- Zomato is a major food delivery and restaurant discovery platform in India.
- This project aims to explore and visualize key insights from this data.
- The analysis will help identify consumer trends, popular cuisines, and factors that influence customer ratings.

Objectives

- To uncover patterns and trends within Zomato's restaurant data to guide business insights.
- Analyze the distribution of restaurants across cities and localities.
- Identify the most popular cuisines and types of restaurants.
- Investigate how features like cost, location, and cuisine affect ratings.
- Understand price ranges and customer preferences in major cities.

Project Framework

- Data Collection and Preprocessing
- Handling Missing Values and Outliers
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Visual Insights and Dashboards
- Interpretation and Storytelling

Data Source & Dataset Overview

- **Dataset**: Zomato Dataset (CSV format from ZIP)
- **Source**: Kaggle or official Zomato dump
- **Encoding Used**: latin-1 for special characters
- **Total Records**: ~9,000+ Restaurants
- **Columns**: Restaurant Name, Location, Rating, Cost, Cuisines, Online Order, etc.
- **Objective**: Analyze restaurant data for patterns, pricing, and customer preferences.

Cleaning & Handling Missing Values

- Dropped unnecessary columns (e.g., URLs, irrelevant IDs)
- Handled null values in (Cuisines, Rating, Cost,) etc.
- Filled or removed missing values using mean/mode or dropped rows

🛠Feature Selection & Engineering

->Selected Features:

•Restaurant Name, Location, Online Order, Rating, Approx Cost, Cuisines

->Feature Engineering:

•Created new column: Cost Category (Low/Medium/High)
•Parsed location into City
•Grouped ratings: Excellent, Good, Average, Poor

Ensuring Data Integrity & Consistency

- Verified and corrected inconsistent spellings (e.g., "Bangalore" vs. "Bengaluru")
- Standardized string formats (trimmed whitespaces, title-cased names)
- Removed duplicate entries
- Unified Cost and Rating column types (float/int)

Patterns, Trends & Anomalies

Trends:

- High ratings cluster around restaurants that offer online ordering
- North Indian & Chinese cuisine most popular
- Bengaluru has the highest density of restaurants

Anomalies:

- Some restaurants have very high cost with poor ratings
- Suspiciously high votes for a few listings → possible spam?

Outliers & Data Transformations

Outliers:

- Used boxplots to detect extreme cost and vote values

Handling:

. Removed or capped extreme values beyond IQR Range

. Transformed Votes using log scale for better distribution

Normalization:

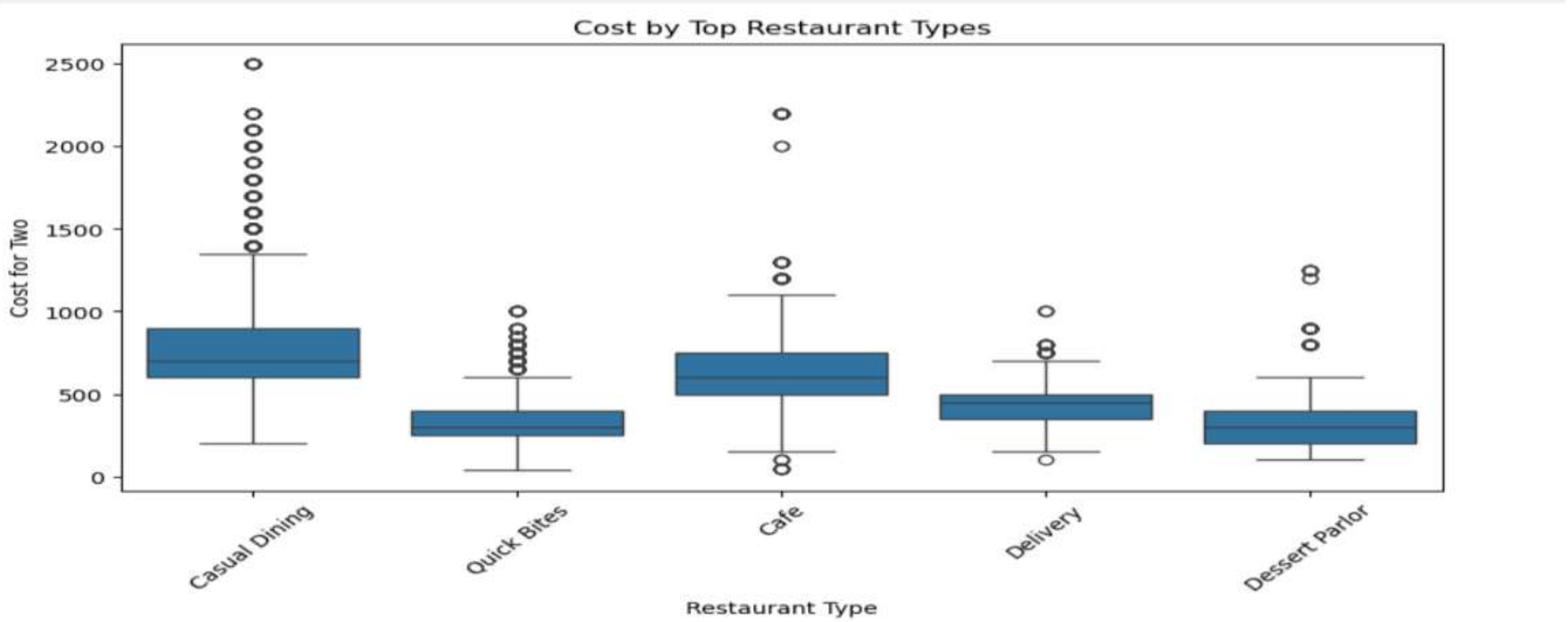
. Standardized cost and Votes for modeling readiness

Visual Representation of Key Findings

. Bar Chart: Online Order Availability

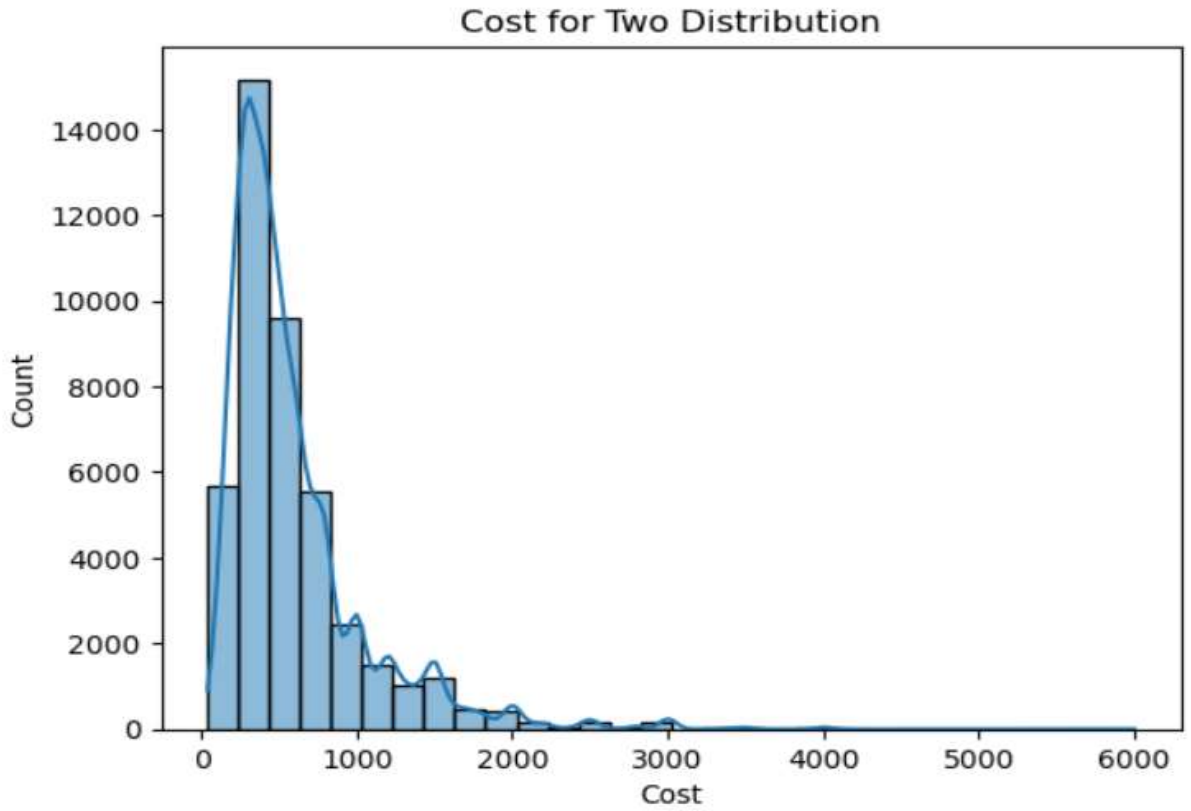
Visual Representation of Key Findings

. Box Plot: Cost by Top Restaurant Types



Visual Representation of Key Findings

. Heat Plot: Cost for Two Distribution



THANK YOU