Zomato
Behind Every **Rating**: The Data Science of Dining **Decisions**

By Group - 7

INTERIM CAPSTONE PRESENTATION

Table Of Contents

Problem Statement

- Restaurants receive ratings but lack clarity on what drives them.
- Zomato shows data but not actionable or predictive insights.
- Decisions on pricing, cuisine, and services are mostly intuition-based.
- Need for a data-driven approach to understand and predict restaurant success.

Business Problem Solved

- Identified key factors influencing restaurant ratings and popularity.
- Analyzed impact of price, cuisine, location, and services.
- Built models to predict restaurant performance.
- Enabled data-backed strategic decision-making.

Importance & Value Addition

- High competition makes restaurant decisions on pricing and offerings critical.
- Existing data lacks predictive and actionable insights.
- This project identifies key drivers of ratings and popularity.
- Builds predictive models for restaurant performance.
- Enables data-driven pricing, menu, and service optimization.

zomato

Dataset Overview (Zomato Restaurants India)

Dataset Size

- **Number of Rows:** 211,944 restaurants
- **Number of Columns:** 26 features

Types of Features

- **Restaurant Information:** Restaurant name, establishment type, URL, address
- **Location Details:** City, locality, latitude, longitude, zipcode
- **Cuisine Information:** Cuisines offered, highlights
- **Pricing Details:** Average cost for two, price range, currency
- **Customer Feedback:** Aggregate rating, rating text, number of votes
- **Service Availability:** Online delivery, takeaway, table booking support
- **Engagement Metrics:** Photo count, votes

Data Cleaning & Preprocessing

Duplicate Removal

Removed repeated restaurant entries to ensure each record represents a unique restaurant.

Missing Value Treatment

Handled missing and invalid values by appropriate imputation or removal to maintain data consistency and avoid bias.

Irrelevant Column Removal

Dropped non-informative and leakage columns such as IDs, URLs, addresses, and derived target fields to prevent noise in modeling

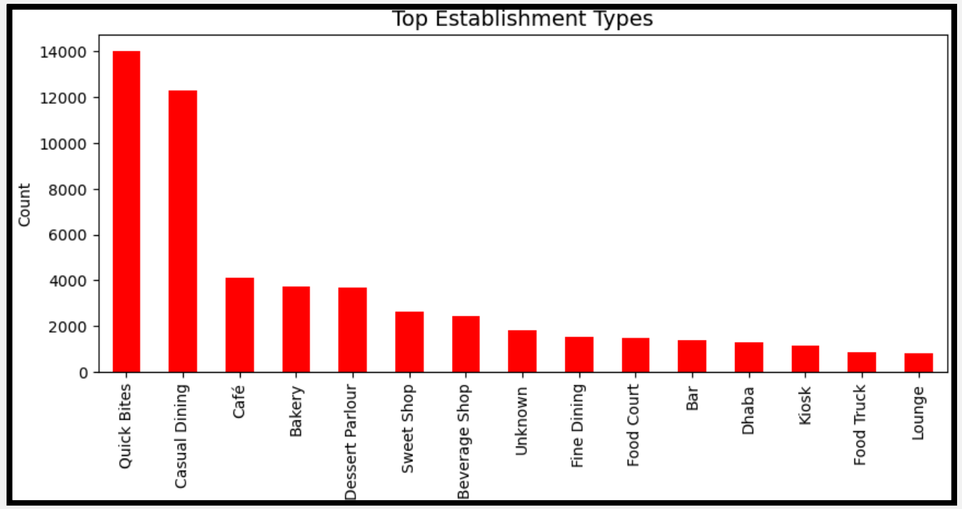
Data Type & Value Correction

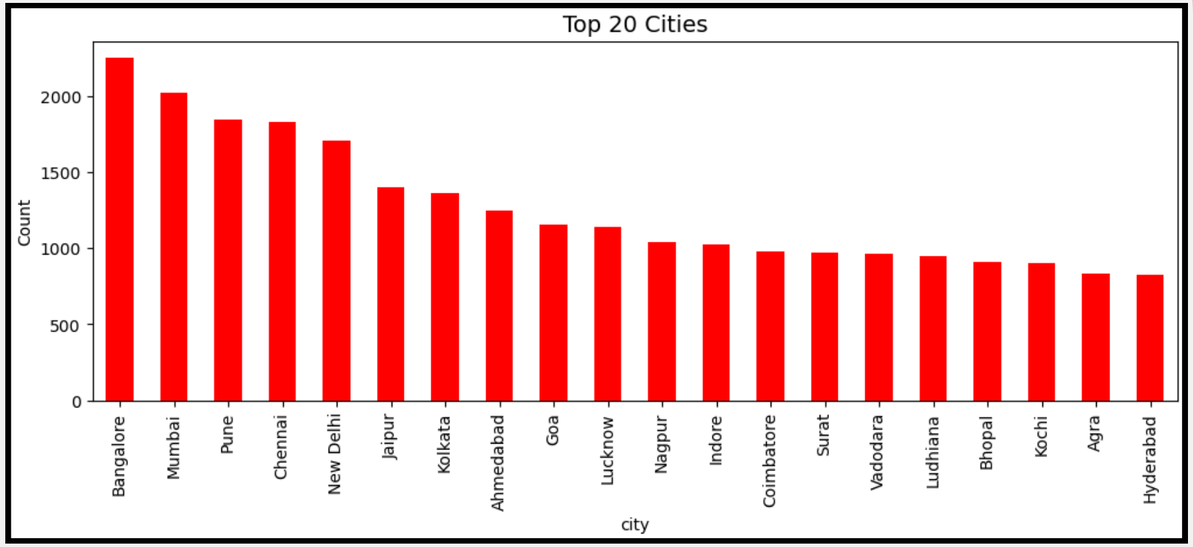
Converted incorrect encodings (e.g., delivery = -1) into meaningful numeric values and ensured all features had correct data types

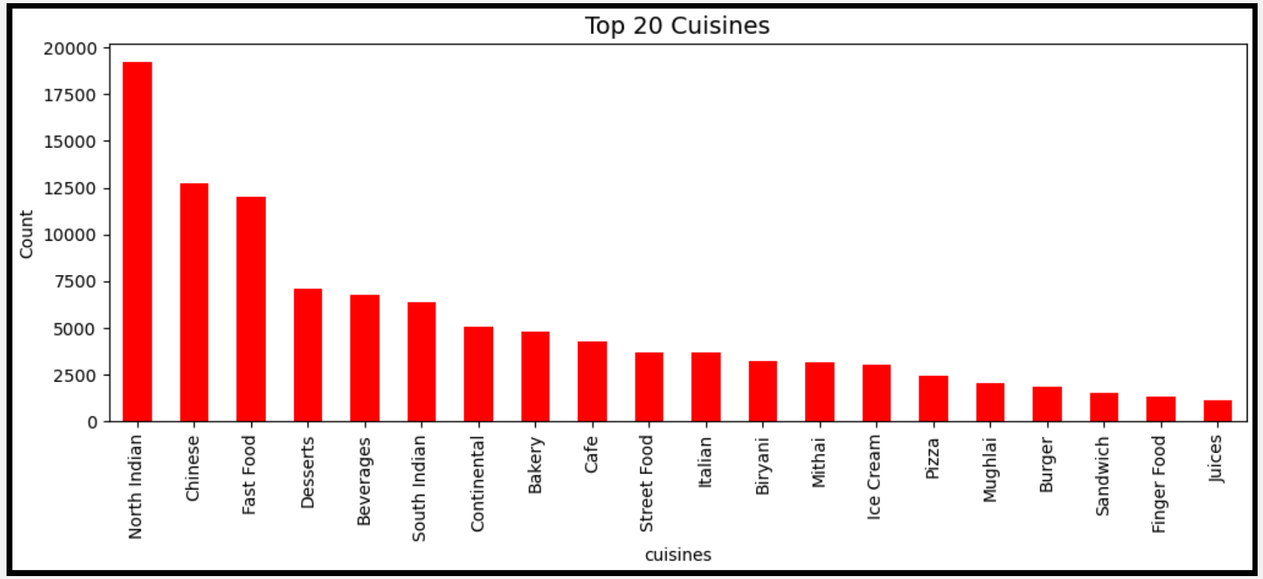
zomato

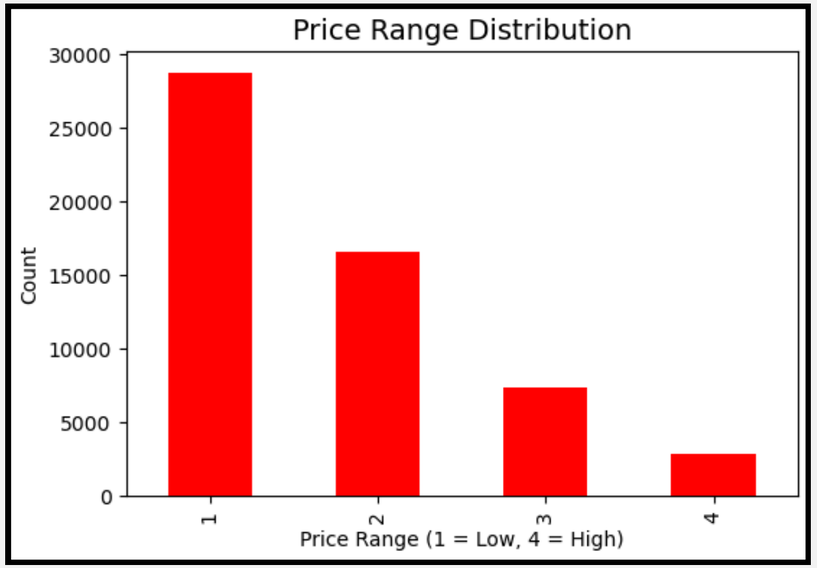
Exploratory Data Analysis

Univariate Analysis



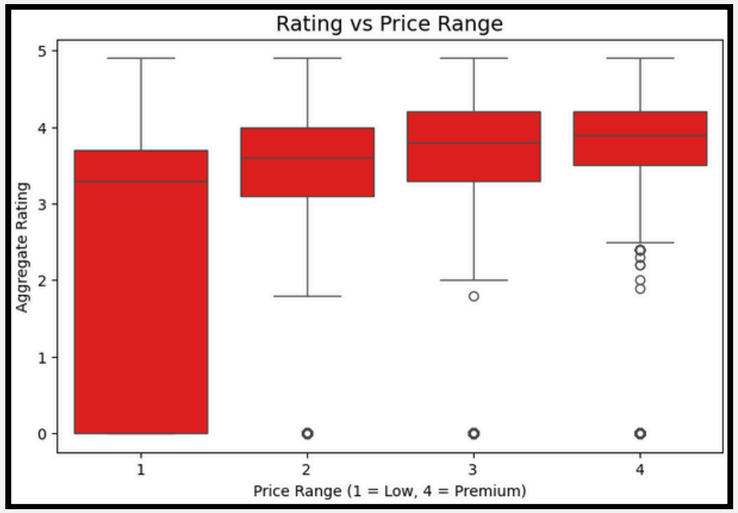


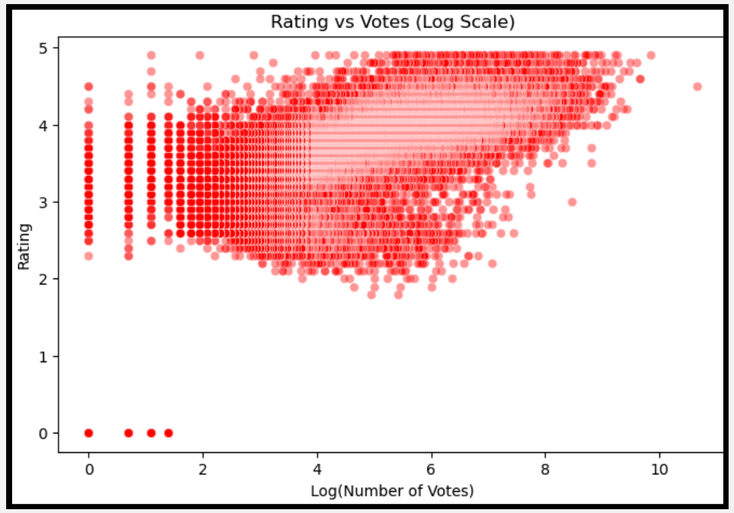


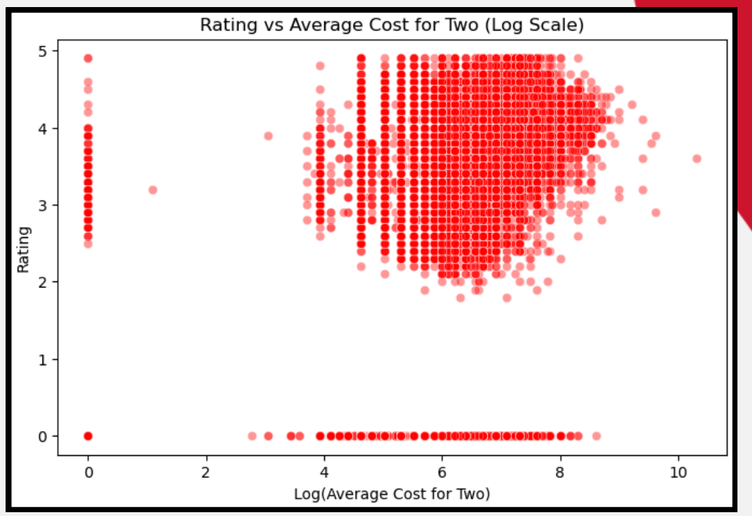


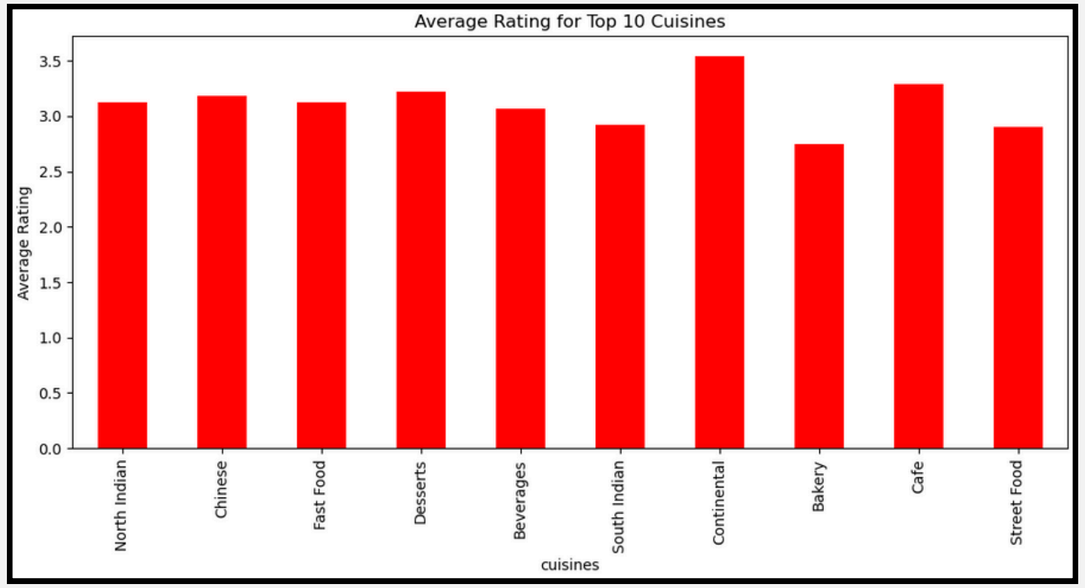
Exploratory Data Analysis

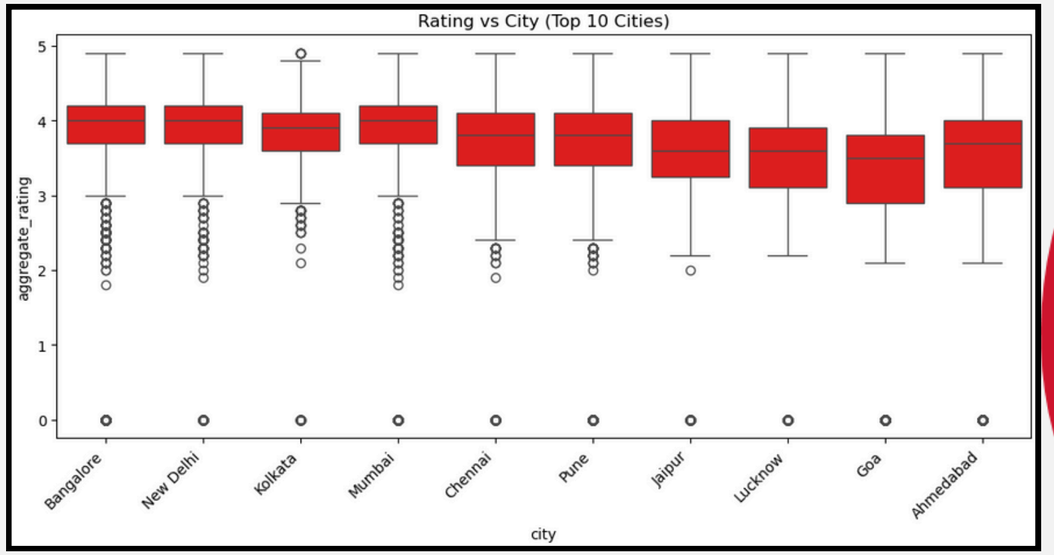
Bivariate EDA — Rating vs Key Features





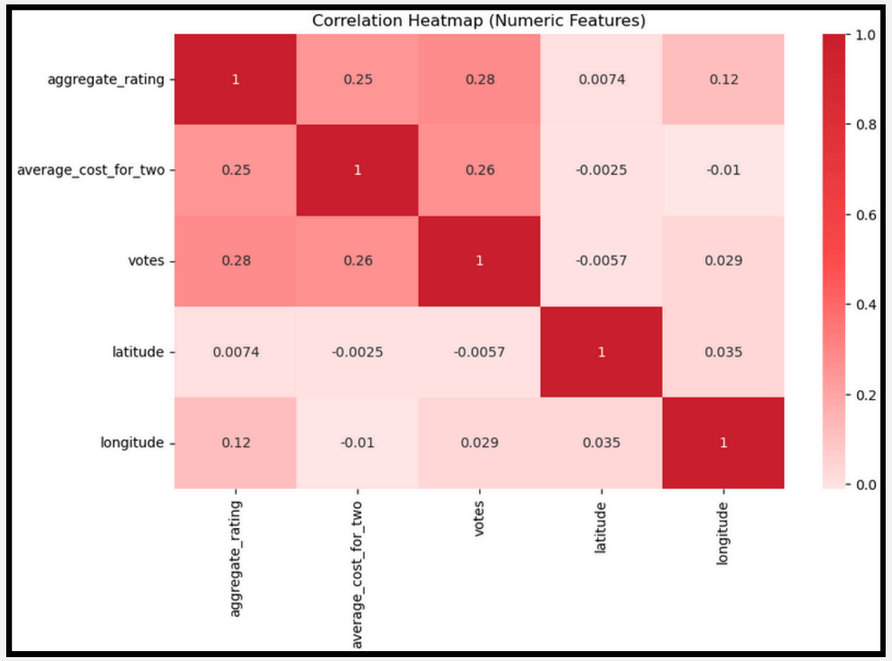


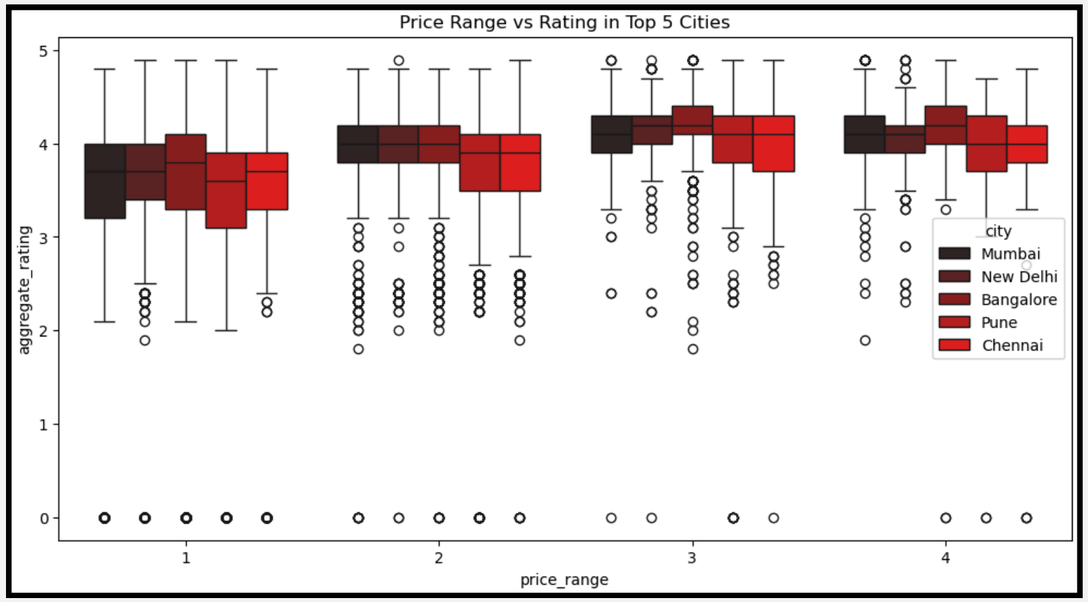


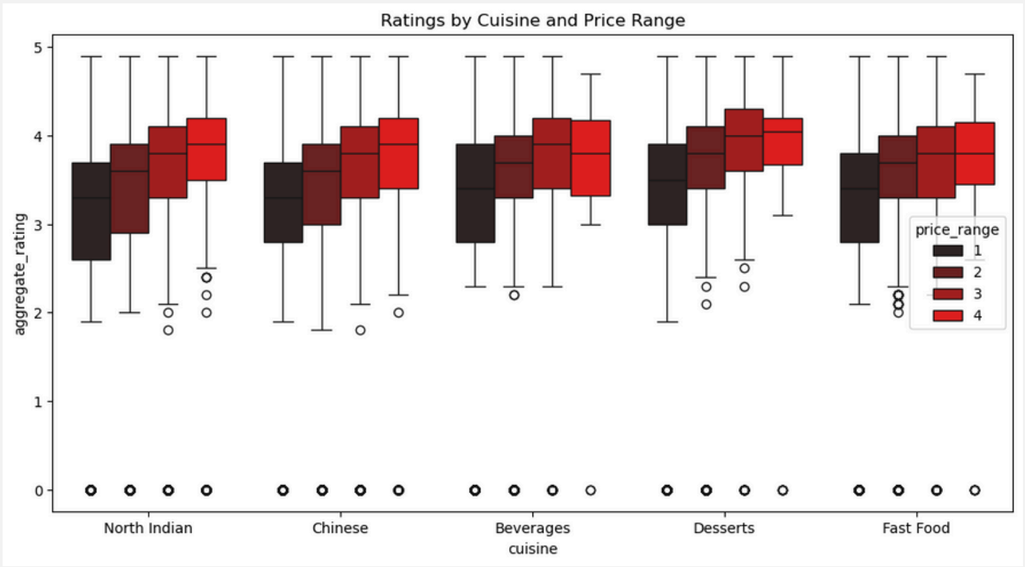


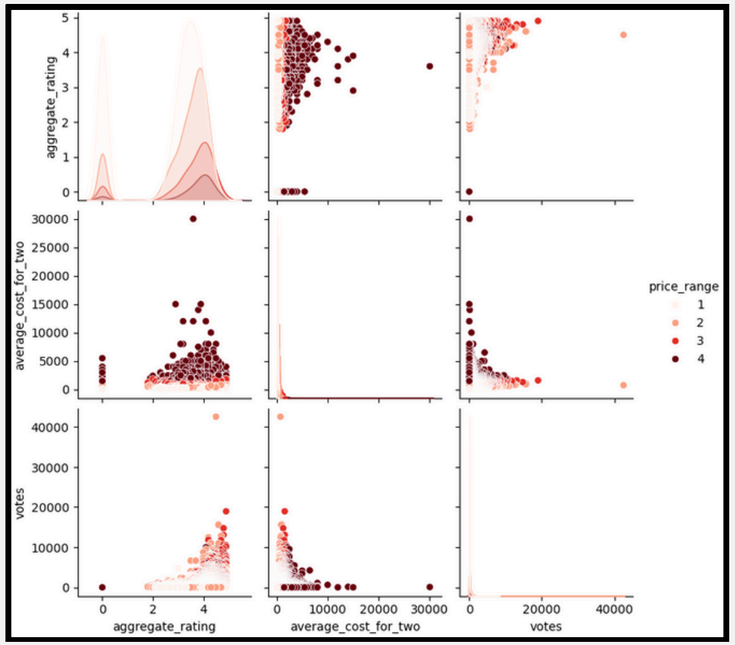
Exploratory Data Analysis

Multivariate Analysis









Statistical Tests and Feature Engineering

Statistical Tests Performed

- **Normality** tests show votes, cost, ratings are non-normal → use log transform for votes & non-parametric tests.
- **Spearman correlation** confirms votes and price_range correlate with rating.
- **ANOVA/Kruskal-Wallis** proves that rating differs significantly across price levels, cost bins, and cities.
- **Chi-square** tests show establishment type and cuisine have significant influence on rating category.
- **T-test** reveals restaurants offering delivery have different rating patterns.

Feature Engineering

- Log-transformed votes to reduce skewness
- Extracted binary service highlight features
- Engineered cuisine indicators and cuisine count
- One-hot encoded categorical variables
- Created location clusters from coordinates

Model Consideration

Problem Understanding

- Goal is to predict restaurant ratings, which is a continuous numeric value
- Problem is framed as a regression task
- Ratings are influenced by multiple factors such as popularity, price, location, cuisine, and services
- Relationships between features and ratings are complex and nonlinear

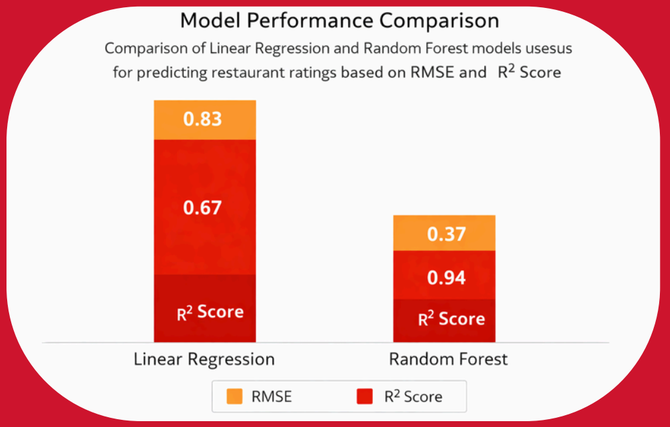
Insights from EDA & Statistical Analysis

- EDA revealed strong nonlinear patterns between predictors and ratings
- Votes and cost showed heavy right skewness
- Ratings varied significantly across cities, cuisines, and establishment types
- Statistical tests confirmed that price range, cuisines, and establishments significantly impact ratings

Models Considered

- **Linear Regression**
  - Selected as a baseline model
  - Helps check if linear relationships exist
  - Useful for validating preprocessing and feature engineering
- **Random Forest Regressor**
  - Ensemble-based tree model
  - Handles nonlinear relationships and feature interactions
  - Robust to outliers and high-dimensional encoded features

Model Building & Comparison



Model Building Approach

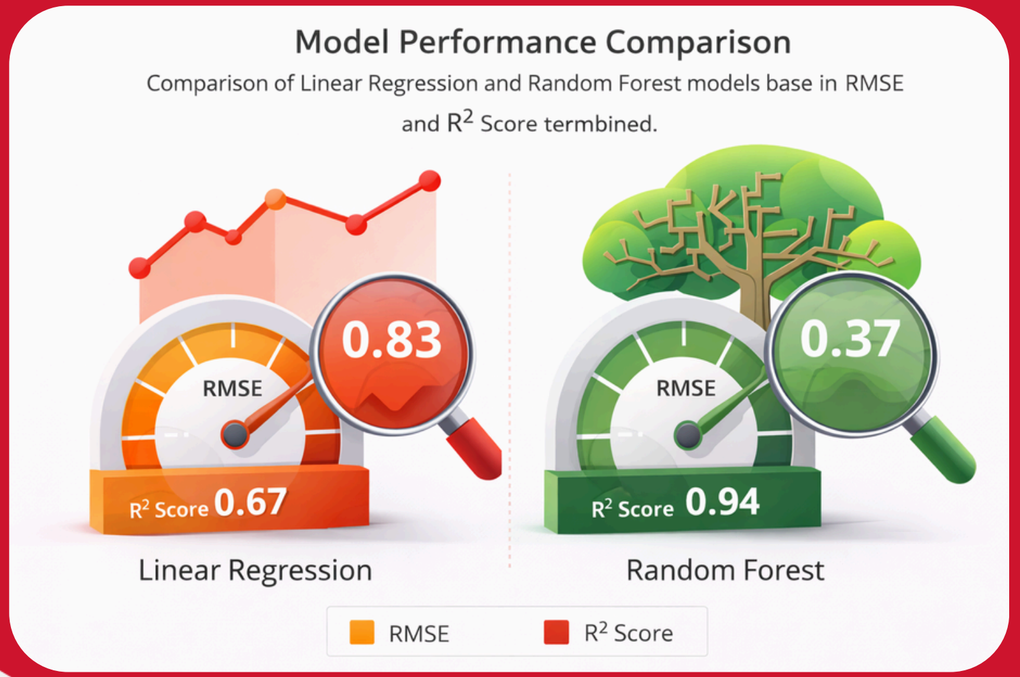
- Dataset split into 80% training and 20% testing
- Same engineered feature set used across models for fair comparison
- Numeric scaling applied only for linear models
- Tree-based models trained on unscaled numeric features

Evaluation Metrics

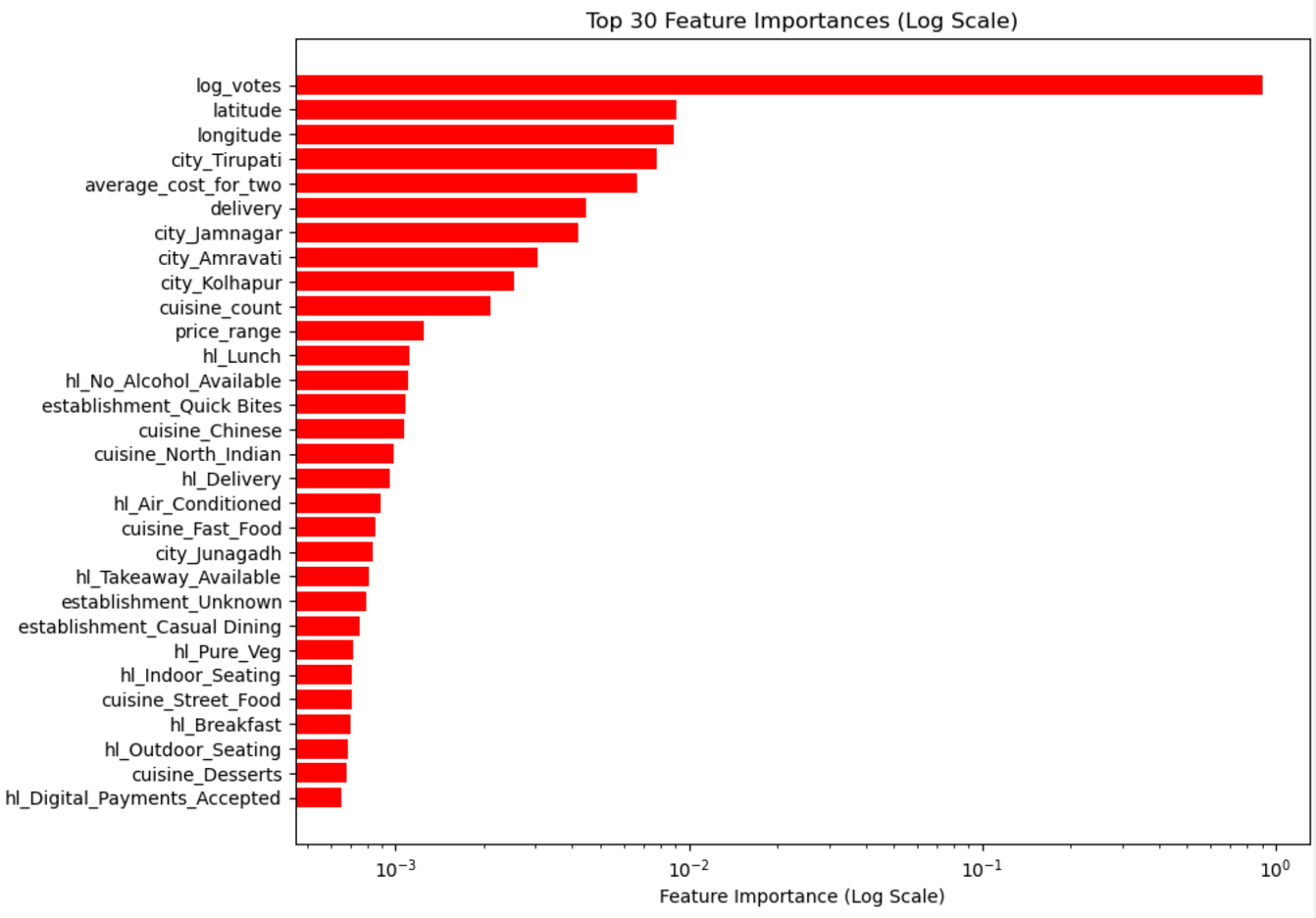
- RMSE used to measure average prediction error
- R^2 score used to measure variance explained by the model
- Combination of both metrics ensures robust evaluation

Model Performance Results

- **Linear Regression:**
  - RMSE ≈ 0.83
  - R^2 ≈ 0.67
  - Underperformed due to linear assumptions
- **Random Forest:**
  - RMSE ≈ 0.37
  - R^2 ≈ 0.94
  - Excellent predictive performance



Model Interpretation – Feature Importance



Interpretation

1. log_votes is the strongest predictor, reflecting restaurant popularity.
2. Location (latitude & longitude) significantly affects ratings.
3. Price-related features influence customer expectations.
4. City-level effects impact rating behavior.
5. Cuisine type contributes to rating differences.
6. Service highlights improve customer satisfaction.
7. Establishment type affects rating patterns.
8. Cuisine diversity has a moderate impact.
9. Individual service features add incremental value.
10. Ratings are influenced by multiple combined factors

Conclusion

- Successfully built a machine learning model to predict restaurant ratings using structured Zomato data.
- Comprehensive EDA and feature engineering helped uncover key drivers of restaurant ratings.
- Random Forest significantly outperformed Linear Regression due to its ability to capture nonlinear relationships.
- Popularity, location, pricing, cuisine type, and service features were identified as major contributors to ratings.
- The final model achieved high predictive accuracy and strong generalization performance.
- The project demonstrates practical application of data science for real-world business problems.

zomato

Future Enhancements & Extensions

zomato

Thank You!

BY GROUP-7

Jeevan Kolluri, Surya Prakash, Alekya, Preethin, Rahul
Parth Tiwari