Restaurant Revenue Prediction – Data Collection

Project: Restaurant Revenue Prediction
Goal: Predict restaurant monthly revenue based on restaurant characteristics and performance indicators.
Learning Type: Supervised Learning
Model Type: Regression

About Dataset

This dataset contains information about various restaurants and aims to predict the revenue based on several features. Each row represents a unique restaurant with various attributes that may influence its revenue.

Columns

- **Name**: The name of the restaurant.
- **Location**: The location of the restaurant (e.g., Rural, Downtown).
- **Cuisine**: The type of cuisine offered (e.g., Japanese, Mexican, Italian).
- **Rating**: The average rating of the restaurant.
- **Seating Capacity**: The number of seats available in the restaurant.
- **Average Meal Price**: The average price of a meal at the restaurant.
- **Marketing Budget**: The marketing budget allocated for the restaurant.
- **Social Media Followers**: The number of social media followers.
- **Chef Experience Years**: The number of years of experience of the head chef.
- **Number of Reviews**: The total number of reviews the restaurant has received.
- **Avg Review Length**: The average length of reviews.
- **Ambience Score**: A score representing the ambience of the restaurant.
- **Service Quality Score:** A score representing the quality of service.
- **Parking Availability:** Indicates if parking is available (Yes/No).
- **Weekend Reservations:** The number of reservations made on weekends.
- **Weekday Reservations:** The number of reservations made on weekdays.
- **Revenue:** The total revenue generated by the restaurant.

Objective

- The main objective of this dataset is to **predict** the **revenue** of a restaurant based on the given features.

Data Format

- File Type: CSV (.csv)

- Number of Rows: 8368
- Number of Columns: 17(including target variable)

Univariate

1. Rating

→ Median = 4.0; Range = 3–5; Skew = -0.01
→ Ratings are centered around 4, with most restaurants rated between 3 and 5 and no heavy skew.

2. Seating Capacity

→ Median = 60; Range = 30–90; Skew = -0.01
→ Most restaurants have around 60 seats, showing a balanced and evenly spread capacity distribution.

3. Average Meal Price

→ Median = ₹45.5; Range = ₹25–₹76; Skew = 0.31
→ Meal prices mostly fall between ₹35–₹60, with a slight skew due to some higher-priced menus.

4. Marketing Budget

→ Median = ₹2846; Range = ₹604–₹9978; Skew = 1.12
→ Budgets vary widely, with a few restaurants spending significantly more, causing strong right skew.

5. Social Media Followers

→ Median = 32,518; Range = 5,277–103,777; Skew = 1.08
→ Follower counts vary greatly, and a few highly-popular restaurants create a noticeable right skew.

6. Chef Experience (Years)

→ Median = 10 years; Range = 1–19; Skew = -0.008
→ Chef experience is generally between 5–15 years and is evenly distributed without skew.

7. Number of Reviews

→ Median = 528; Range = 50–999; Skew = 0.0017
→ Review counts vary between 50 and 1,000 but remain fairly balanced overall.

8. Average Review Length

→ Median = 173.9 words; Range = 50–299; Skew = 0.004
→ Review length is consistent, with most reviews around 150–250 words and no skew.

9. Ambience Score

→ Median = 5.5; Range = 1–10; Skew = -0.019
→ Ambience ratings cluster around 5–7, showing no strong skew.

10. Service Quality Score

→ Median = 5.6; Range = 1–10; Skew = -0.011
→ Service scores are mostly mid-range (5–7) with evenly distributed values.

11. Weekend Reservations

→ Median = 27; Range = 0–88; Skew = 0.57
→ Reservation counts increase sharply for some busy restaurants, causing moderate right skew.

12. Weekday Reservations

→ Median = 26; Range = 0–88; Skew = 0.58
→ Weekday reservations follow a similar pattern to weekends, with a few high-traffic outliers.

13. Revenue

→ Median = ₹604,242; Range = ₹184,708–₹1,531,867; Skew = 0.73
→ Revenue shows large variation, with a few restaurants earning much more, causing strong right skew.

Bivariate

Covariance:

1. Rating

→ Very small covariance values with all variables.
→ Rating does not move significantly with any other feature.
→ Changes in rating are mostly independent of restaurant size or revenue.

2. Seating Capacity

→ High positive covariance with Revenue — as capacity increases, revenue also increases.
→ Also shows positive covariance with reservations and followers.
→ Bigger restaurants tend to show larger combined growth in other business metrics.

3. Average Meal Price

→ Positive covariance with Revenue — higher prices generally move with higher income.
→ Also shows covariance with cuisine indicators (premium cuisines → higher price).
→ Price changes move moderately along with marketing and review metrics.

4. Marketing Budget

→ Very high covariance with Revenue — a strong directional relationship.
→ Moves strongly with social media followers (more marketing → more followers).
→ As marketing increases, related business indicators rise together.

5. Social Media Followers

→ Large positive covariance with Revenue — both increase together.
→ Also moves with marketing spend and reservations.
→ Influencer growth and revenue grow side-by-side.

6. Chef Experience

→ Small positive covariance with Revenue.
→ Experience increases slightly along with reviews and ratings.
→ Not a major driver but still moves in a positive direction.

7. Number of Reviews

→ Moderate positive covariance with Revenue.
→ Shops with more reviews also tend to have more footfall, followers, and reservations.
→ Reviews grow with business scale.

8. Average Review Length

→ Low covariance with almost all variables.
→ Review length does not meaningfully move with revenue or ratings.
→ Mostly independent.

9. Ambience Score

→ Very small covariance with revenue-like features.
→ Ambience changes do not strongly shift other numeric features.
→ Impact is minimal.

10. Service Quality Score

→ Small positive covariance with Revenue and Reservations.
→ Better service moves slightly with business growth, but not strongly.
→ Weak directional relationship.

11. Weekend/Weekday Reservations

→ Strong positive covariance with Revenue — more bookings → more income.
→ Also move with seating capacity and followers.
→ Reservation counts scale with restaurant size.

12. Revenue

→ Shows the highest covariance with marketing, followers, price, and reservations.
→ Means all these variables tend to move in the same direction as revenue.
→ Revenue is the central scaling variable of the dataset.

Correlation:

1. Rating

→ Very weak correlations with all features.
→ Ratings do not strongly depend on capacity, pricing, or marketing.

2. Seating Capacity

→ Strong positive correlation with Revenue (0.68) — bigger restaurants earn more.
→ Also strongly correlates with Weekday Reservations (0.41) and Weekend Reservations (0.43).
→ Strong negative correlation with Location_Rural (–0.81) — rural restaurants tend to be smaller.

3. Average Meal Price

→ Strong positive correlation with Revenue (0.69).
→ Also highly correlated with Japanese cuisine (0.68) and French cuisine (0.40).
→ Strong negative correlation with Mexican cuisine (–0.54) and Indian cuisine (–0.38).

4. Marketing Budget

→ Very strong correlation with Social Media Followers (0.99).
→ Moderate positive correlation with Revenue (0.36).
→ Negative with rural/suburban locations (≈ –0.31), meaning higher-budget restaurants are urban.

5. Social Media Followers

→ Strongly related to Marketing Budget (0.99) and Seating Capacity (0.49).
→ Moderate positive correlation with Revenue (0.35).
→ Negative correlation with rural locations (–0.30).

6. Chef Experience

→ Very weak correlations across all features.
→ Slight positive effect on Revenue (0.02), but negligible.

7. Number of Reviews

→ Extremely weak correlations overall.
→ Slightly positive with Rating (0.01) and Ambience (0.01).
→ No meaningful link to revenue.

8. Avg Review Length

→ Weak correlations throughout.
→ Slightly higher in restaurants with French/Japanese/Italian cuisines.
→ No strong business impact.

9. Ambience Score

→ Very weak correlations.
→ Slightly positive with Rating and Service Score.
→ No major influence on reservations or revenue.

10. Service Quality Score

→ Very weak correlations with business metrics.
→ Slight positive link with Japanese cuisine (0.02).
→ Otherwise independent.

11. Weekend Reservations

→ Moderately correlated with Seating Capacity (0.43).
→ Some correlation with Revenue (0.29).
→ Negative with Location_Rural (–0.35).

12. Weekday Reservations

→ Strong correlation with Seating Capacity (0.41).
→ Moderate correlation with Revenue (0.27).
→ Negative with Location_Rural (–0.33).

13. Revenue

→ Strong correlations with:

- Seating Capacity (0.68)
- Average Meal Price (0.69)
- Japanese Cuisine (0.46)

→ Negative correlations with Rural location (–0.54) and Mexican cuisine (–0.38).

14. Location_Rural

→ Strong negative correlation with Seating Capacity (–0.81).
→ Negative with Revenue (–0.54).
→ Rural locations tend to be smaller and earn less.

15. Location_Suburban

→ Moderate negative correlation with Marketing Budget and Social Followers (≈ –0.30).
→ Very weak relationship with revenue or pricing.

16. Cuisine_French

→ Strongly linked with high meal prices (0.40).
→ Negative correlation with other cuisine flags (as expected).
→ Slight positive correlation with Revenue (0.28).

17. Cuisine_Indian

→ Strong negative correlation with high meal prices (–0.38).
→ Negative with revenue (–0.26).
→ Typical for lower-priced cuisine categories.

18. Cuisine_Italian

→ Mild positive correlation with Revenue (0.06).
→ Mostly independent of other features.

19. Cuisine_Japanese

→ Strong positive correlation with meal price (0.68).
→ Strong positive correlation with Revenue (0.46).
→ Japanese cuisine is associated with premium pricing and higher earnings.

20. Cuisine_Mexican

→ Strong negative correlation with meal price (–0.54).
→ Strong negative correlation with Revenue (–0.38).
→ Lower-priced cuisine category.

21. Parking Availability (Yes)

→ Very weak correlations across all features.
→ Slight positive with Service Quality and Social Followers.
→ Minimal impact on revenue.

Multicollinearity: (VIF)

VIF analysis was conducted to assess multicollinearity among the independent variables. Most variables showed VIF values in the acceptable range, while a few had moderately high scores. However, none of the variables crossed the critical threshold after refinement. Therefore, all selected predictors were considered suitable and retained for further model development.

Feature Selection Summary (RFE, SelectKBest, PCA) & Best Model Results

1. PCA Results (Simple, Clear Explanation)

Principal Component Analysis (PCA) was applied to reduce dimensionality and identify the most informative components. Models such as Linear Regression, SVM, Decision Tree, and Random Forest were tested across PCA_5, PCA_6, and PCA_8 components.

Across all PCA configurations, **Random Forest consistently produced the highest R^2 scores**, with the best result achieved using:

- **Best Model:** Random Forest
- **Best PCA Component:** PCA_8
- **Best R^2 Score: 0.9704**

This indicates that **retaining 8 components preserved maximum variance**, enabling Random Forest to capture complex relationships effectively.

---

2. RFE Results (Recursive Feature Elimination)

Recursive Feature Elimination (RFE) was used to select the most relevant subset of features. Multiple iterations (#3, #6, #8) were evaluated using Linear, SVM, Decision Tree, and Random Forest models.

Across all RFE-selected feature sets:

- **Random Forest consistently ranked as the best-performing model**
- R^2 values were extremely high (around **0.9987 – 0.9991**)
- The selected features via RFE proved highly predictive and stable across models

This shows that **RFE effectively identified strong predictors**, and Random Forest was able to leverage them with minimal performance variation.

---

3. SelectKBest (Chi-Square) Results

SelectKBest (Chi-square) was used to select statistically significant features. Similar to RFE, multiple iterations (#6, #8, #10) were tested on different models.

The findings were again consistent:

- **Random Forest achieved the highest accuracy/R^2 in every iteration**
- Performance remained highly stable with scores above **0.9990**
- Linear and Decision Tree models performed well, but Random Forest remained superior

This indicates that **Chi-Square–selected features are also highly relevant**, and Random Forest utilizes them most effectively.

Overall Conclusion:

➡️ **Random Forest is the best-performing and most reliable model across all feature selection methods.**

➡️ **RFE and Chi-Square produced stronger results than PCA**, but PCA still worked well with Random Forest.

➡️ These results confirm that the dataset contains strong predictive patterns, and Random Forest is highly effective in capturing them.

Final Summary

- RFE was selected as the final feature-selection method because it provided the highest model accuracy ($R^2 \approx 0.99$) and selected the most meaningful features for predicting restaurant revenue.
- Compared to PCA and SelectKBest, RFE produced a simpler, more interpretable, and more accurate model.
- The final Random Forest model trained on RFE-selected features was saved and deployed successfully.