Mall Shop Rent Prediction – Data Collection

Project: Mall Shop Rent Prediction
Goal: Predict the monthly rent of mall shops based on shop and mall features.
Learning Type: Supervised Learning
Model Type: Regression

Dataset Description

- **Shop_ID:** Unique ID of the shop
- **Mall_Name:** Name of the mall
- **City:** City of the mall
- **Floor:** Floor number (0 = Ground)
- **Shop_Size_sqft:** Area of the shop in square feet
- **Footfall_per_day:** Average number of visitors per day
- **Nearby_Brands:** Number of branded shops nearby
- **Has_Food_Court:** 1 if near food court, 0 otherwise
- **Monthly_Sales:** Average monthly sales (₹)
- **Rent:** Target variable (monthly rent in ₹)

Data Source

- The dataset was created synthetically for academic purposes to simulate mall rental prediction scenarios.

Data Format

- File Type: CSV (.csv)
- Number of Rows: 200
- Number of Columns: 10 (including target variable)

Target Variable

- Rent (monthly rent in ₹)

Univariate

1. Floor:

→ Median = 2; Range = 0–4; Skew = -0.12
→ Most shops are located on the 2nd floor, with floor distribution fairly balanced across levels.

2. Shop_Size_sqft:
→ Mean = 1542 sq.ft; Range = 159–2995; Skew = 0.09
→ Average shop size is around 1500 sq.ft, with a few larger shops. Distribution is almost symmetrical.

3. Footfall_per_day:
→ Mean = 5001 visitors; Range = 505–9946; Skew = 0.08
→ Daily footfall varies widely but is fairly symmetric; most shops see about 4,000–5,000 visitors per day.

4. Nearby_Brands:
→ Mean = 10; Range = 2–19; Skew = 0.09
→ Shops usually have around 10 nearby brands. Variation across shops is moderate.

5. Has_Food_Court:
→ Mean = 0.45; Mode = 0; Skew = 0.18
→ About 45% of shops are in malls that have a food court; slightly more are in malls without one.

6. Monthly_Sales:
→ Mean = ₹10.1 lakh; Range = 0–₹19.9 lakh; Skew = 0.09
→ Sales are moderately spread, with average around ₹10 lakh per month; distribution is almost normal.

7. Rent:
→ Mean = ₹51,942; Range = -₹7676–₹1,34,727; Skew = 0.22
→ Rent values show moderate variation; some data errors or unusual entries may exist (negative rent).

8. Mall_Name_Forum_Mall:
→ Mean = 0.18
→ Around 18% of shops belong to Forum Mall.

9. Mall_Name_Lulu_Mall:
→ Mean = 0.25
→ About 25% of shops are located in Lulu Mall.

10. Mall_Name_Phoenix_Marketcity:
→ Mean = 0.57
→ More than half of the shops are in Phoenix Marketcity.

---

1. Frequency Table for Monthly_Sales:

The table shows:

- **Unique_values:** each different sales number.
- **Frequency:** how many times it appears (here it's mostly 1 for each).
- **Relative_Frequency:** percentage of how common it is.
- **Cumsum:** adds up all the relative frequencies till 100%.

This means every sales value is **almost unique** — there are no repeating numbers. So, Monthly_Sales looks like continuous data (different for every shop or month).

2. Rent Probability Graph:

checked how likely the **Rent** is between **25,000 and 50,000**.

The results say:

- **Average (Mean) Rent = 51,942**
- **Standard Deviation = 25,874** (this shows how spread out the rent values are)
- The chance (probability) that rent is between 25,000 and 50,000 is **about 32%**.

In the graph:

- The **green bars** show how rent values are spread (histogram).
- The **blue line** is the smooth "bell curve" (normal distribution).
- The **two red lines** mark 25,000 and 50,000 rent values — the shaded area between them is about **32% of the data**.

Bivariate

Covariance:

- **Shop_Size_sqft and Rent** → covariance = 9.64114e+06
  → This is a **large positive covariance**, meaning as *shop size increases*, *rent* tends to increase as well.
- **Footfall_per_day and Monthly_Sales** → covariance = 7.811129e+07
  → Indicates a **strong positive relationship**. Higher *daily footfall* is associated with higher *monthly sales*.
- **Nearby_Brands and Footfall_per_day** → covariance = -3.977333e+02
  → **Negative covariance** suggests that as the *number of nearby brands increases*, *footfall per day* slightly decreases — possibly due to competition.
- **Shop_Size_sqft and Has_Food_Court** → covariance = 34.184997
  → Small positive covariance — *shops in malls with food courts* tend to have *slightly larger sizes*.
- **Monthly_Sales and Rent** → covariance = -2.564048e+09
  → **Negative covariance**, meaning *higher rent* does not necessarily lead to *higher sales*; possibly high-rent shops are not always the top performers.

Correlation:

- **Shop_Size_sqft and Rent →** correlation = 0.438540
  → Moderate positive correlation — as shop size increases, rent tends to increase, but not perfectly.
- **Footfall_per_day and Rent →** correlation = 0.159894
  → Weak positive correlation — footfall has a slight positive relationship with rent.
- **Shop_Size_sqft and Monthly_Sales →** correlation = -0.146357
  → Weak negative correlation — larger shops don't necessarily generate higher sales**.**
- **Has_Food_Court and Rent →** correlation = 0.232621
  → Weak positive correlation — malls with food courts tend to have slightly higher rent values.
- **Nearby_Brands and Footfall_per_day →** correlation = -0.026701
  → Very weak negative correlation — number of nearby brands doesn't strongly affect footfall.
- **Mall_Name_Lulu_Mall and Mall_Name_Phoenix_Marketcity →** correlation = -0.298122
  → Negative correlation due to one-hot encoding — being in one mall implies not being in another.

Multicollinearity: (VIF)

VIF analysis was performed to check multicollinearity among independent variables. All variables had VIF values below 5, indicating no serious multicollinearity issues. Therefore, all selected features were retained for model building.

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

1. Introduction

The objective of this analysis was to evaluate multiple feature-selection techniques and compare model performance across several regression models.
Three different feature-selection/pre-processing methods were tested:

1. SelectKBest (Chi-Square)
2. RFE (Recursive Feature Elimination)
3. PCA (Principal Component Analysis)

2. SelectKBest (Chi-Square)

- In all three cases, Linear Regression produced the highest R^2 scores.
- The best score achieved using SelectKBest was 0.245430.
- Other models (SVM, Decision, Random Forest) had lower scores and sometimes negative R^2, indicating poor fit.

Summary Table from SelectKBest

Result_No	Best_Model	Best_Score
#5	Linear	0.245430
#6	Linear	0.225526
#7	Linear	0.185558

3. RFE (Recursive Feature Elimination)

- Linear Regression again performed the best for each result.
- Overall R^2 scores were lower than those from SelectKBest.
- The maximum score from RFE was 0.184371, indicating weaker model performance with RFE.

Summary Table from RFE

Result_No	Best_Model	Best_Score
#5	Linear	0.184371
#6	Linear	0.177310
#7	Linear	0.171788

4. PCA (Principal Component Analysis)

PCA was applied with multiple component sizes (PCA_2, PCA_3, PCA_5, PCA_10).
Each PCA output was evaluated across multiple models.

- PCA_3 components gave the highest performance.
- Linear Regression remained the strongest model.
- Performance was better than RFE, but still lower than SelectKBest.

The code identified:

- Best_Model → Linear
- Best PCA Component → PCA_3
- Best R^2 Score → 0.1988

5. Comparison of All Methods

Method	Best Model	Best Score	Notes
SelectKBest	Linear	0.245430	Highest performance overall
RFE	Linear	0.184371	Lowest performance
PCA	Linear	0.1988	Moderate performance; PCA_3 best

FINAL SUMMARY:

In every experiment, Linear Regression performed better than SVM, Decision Tree, or Random Forest models.

Among all methods, **SelectKBest (Chi-Square)** produced the **highest R^2 value (0.245430)**, making it the most effective feature-selection strategy for this dataset. PCA achieved moderate performance, with PCA_3 components giving an R^2 of 0.1988, while RFE performed the weakest, with a maximum score of 0.184371.
