Data Science Projects Summary

NAME: Shradha Shivanand Bhandari

INTERN ID: DS100143

COLLEGE: Nutan college of Engineering and Research,Talegaon Dabhade,

Pune-4120507

Week 1:

Project 2 - Movie Ratings Analysis

Objective:
Analyze a dataset of movie ratings to uncover patterns and insights, such as the most popular genres and average ratings trends. This project helps explore audience preferences and trends in the film industry.

Code Highlights:

1. **Data Loading and Cleaning:**
   - Imported the dataset containing movie names, ratings, and genres using Python's Pandas library.
   - Cleaned the dataset by removing duplicates and handling missing values to ensure accurate analysis.
2. **Genre Analysis:**
   - Calculated the average rating for each genre using group-by operations to identify popular genres.
3. **Trend Analysis:**
   - Extracted and plotted trends in average movie ratings over time by grouping data by release year.
4. **Visualization:**
   - Used Matplotlib to create bar charts for average ratings by genre and line plots for yearly trends.
   - Generated a word cloud to visualize frequent keywords in movie titles.

Explanation:
This project provides insights into the preferences of movie audiences by analyzing genres and trends in ratings. The word cloud visually represents recurring themes in movie titles, enhancing understanding of popular content. The analysis is particularly useful for industry stakeholders to gauge audience interest and market trends.

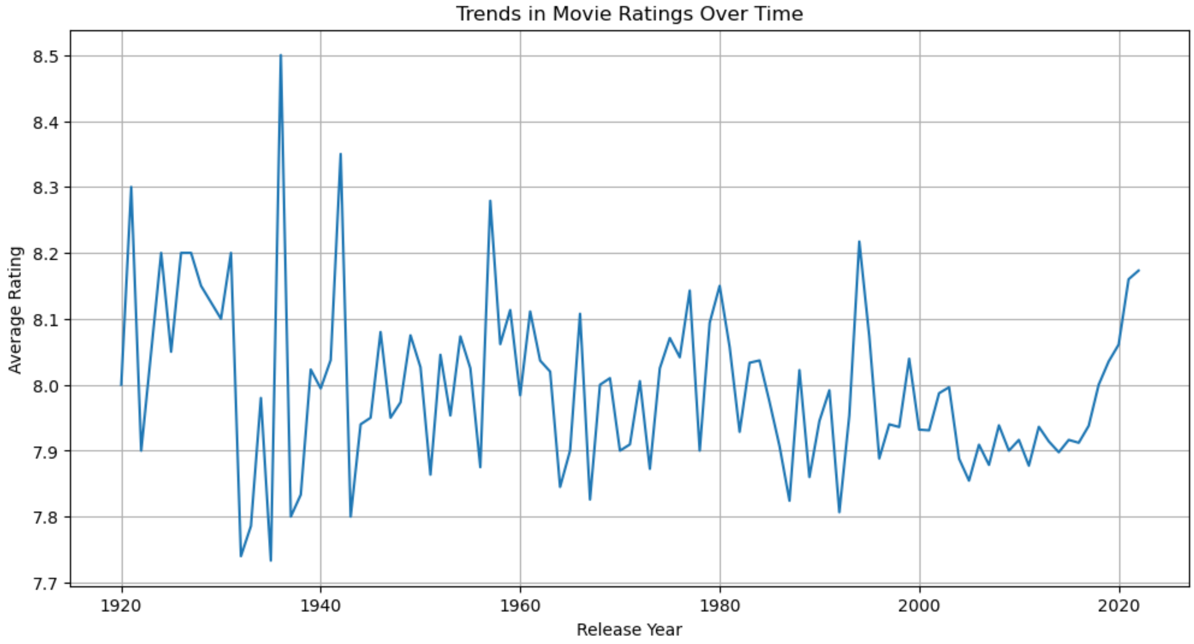
Tools and Technologies Used:

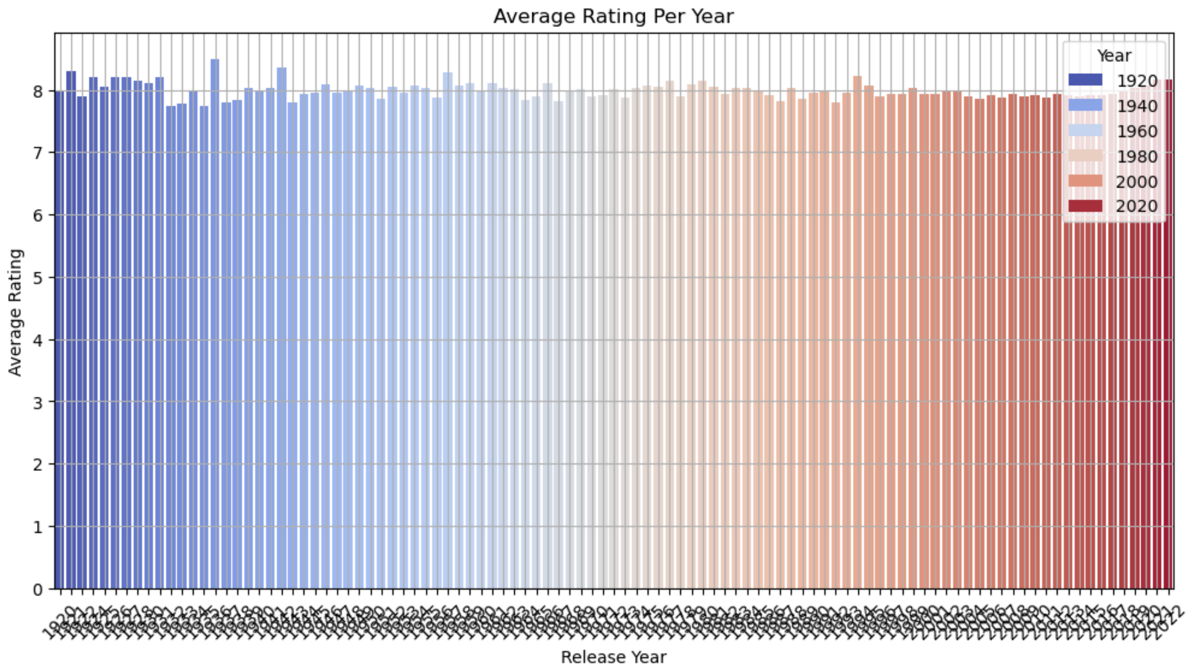
- **Programming Language:** Python

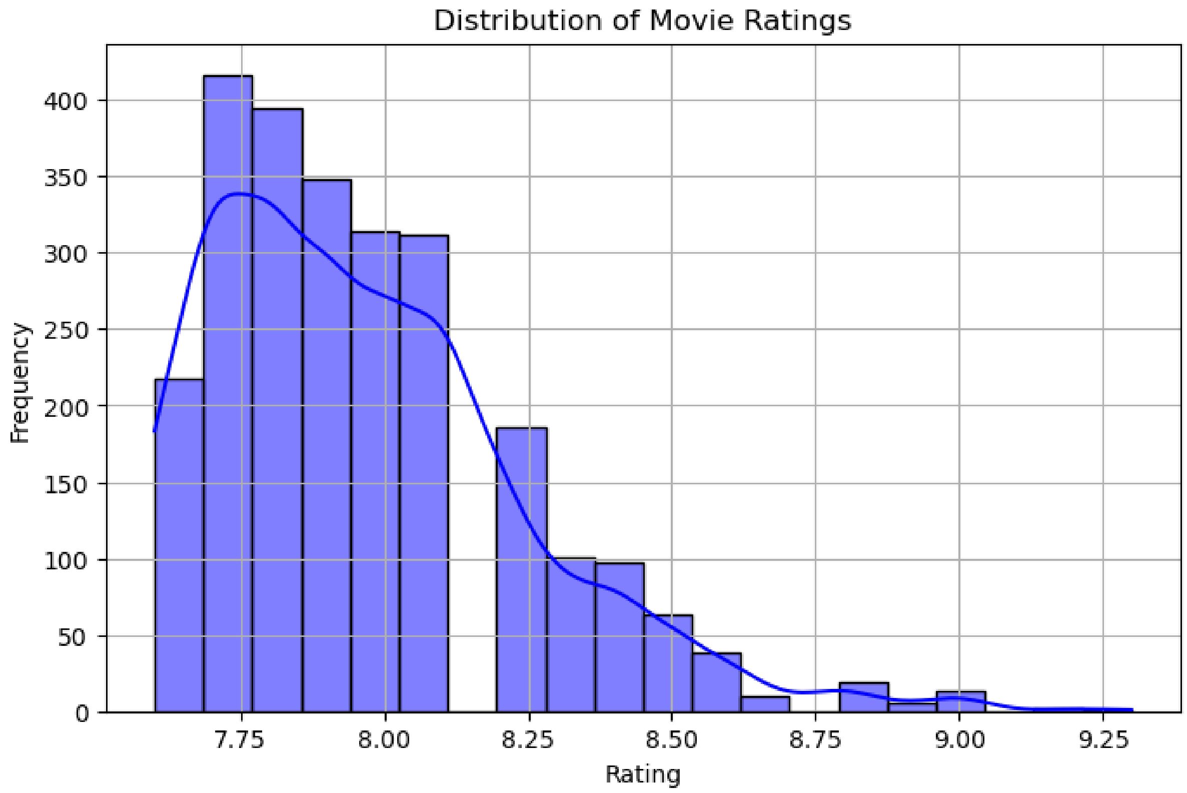
- **Libraries:** Pandas, Matplotlib, WordCloud
- **Dataset:** IMDB Movie Ratings Dataset from Kaggle

Output Description:

- **Bar Chart:** Displays average ratings for each genre, highlighting the most popular ones.
- **Line Plot:** Shows trends in average ratings over the years, providing historical insights.
- **Word Cloud:** Visualizes frequent keywords in movie titles to identify common themes.




Average Rating Per Year
Average Rating
0
1
2
3
4
5
6
7
8
Year
1920
1940
1960
1980
2000
2020
Release Year
1920
2022


Distribution of Movie Ratings
Frequency
0
50
100
150
200
250
300
350
400
7.75
8.00
8.25
8.50
8.75
9.00
9.25
Rating

Week 2:

Project 3 - Weather Data Visualization

Objective:
Analyze weather data to identify patterns, trends, and anomalies in temperature over time, providing insights for climate studies or seasonal planning.

Code Highlights:

1. **Data Loading and Cleaning:**

- Loaded the weather dataset and converted date columns to a datetime format for accurate time-series analysis.
- Handled missing values using interpolation methods to ensure data consistency.

2. **Trend Analysis:**

- Calculated monthly average temperatures to identify seasonal patterns.
- Detected anomalies by comparing data points to moving averages.

3. **Visualization:**

- Created line plots for temperature trends and scatter plots to highlight anomalies.
- Used Seaborn to create heatmaps for temperature distribution over months and years.

Explanation:
The project focuses on visualizing temperature data to identify long-term trends and unusual spikes or drops. Seasonal patterns and anomalies are presented using line and scatter plots, aiding in better understanding of weather behavior.

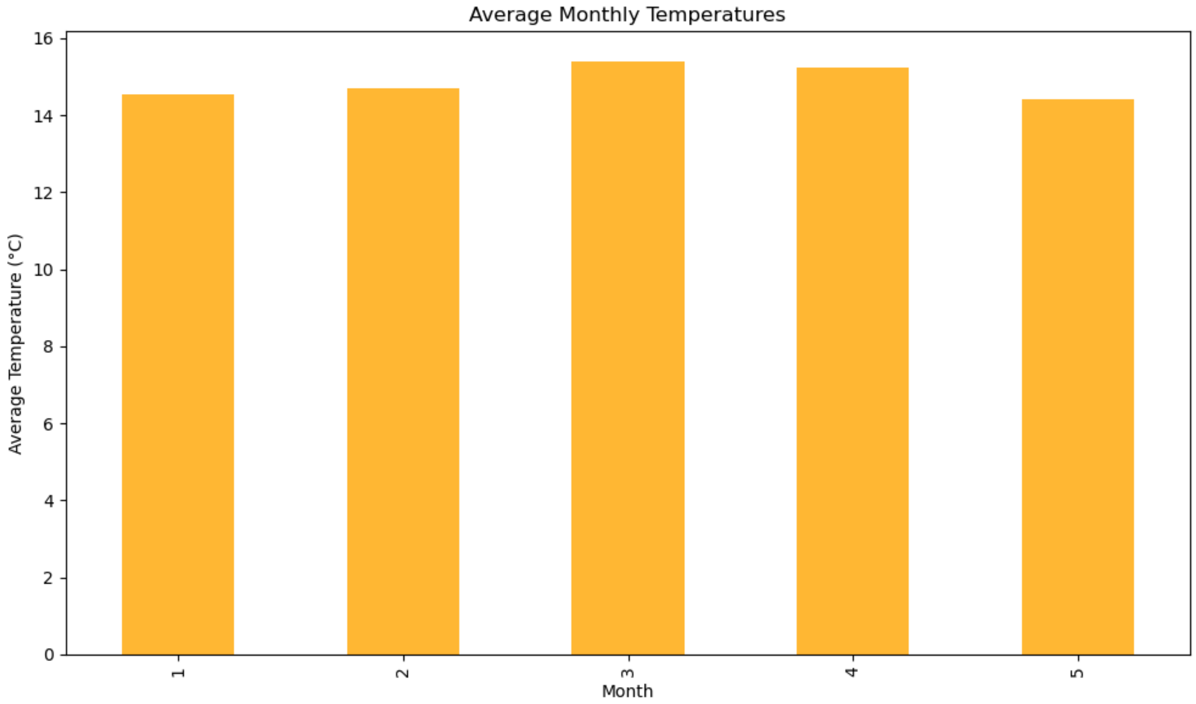
Tools and Technologies Used:

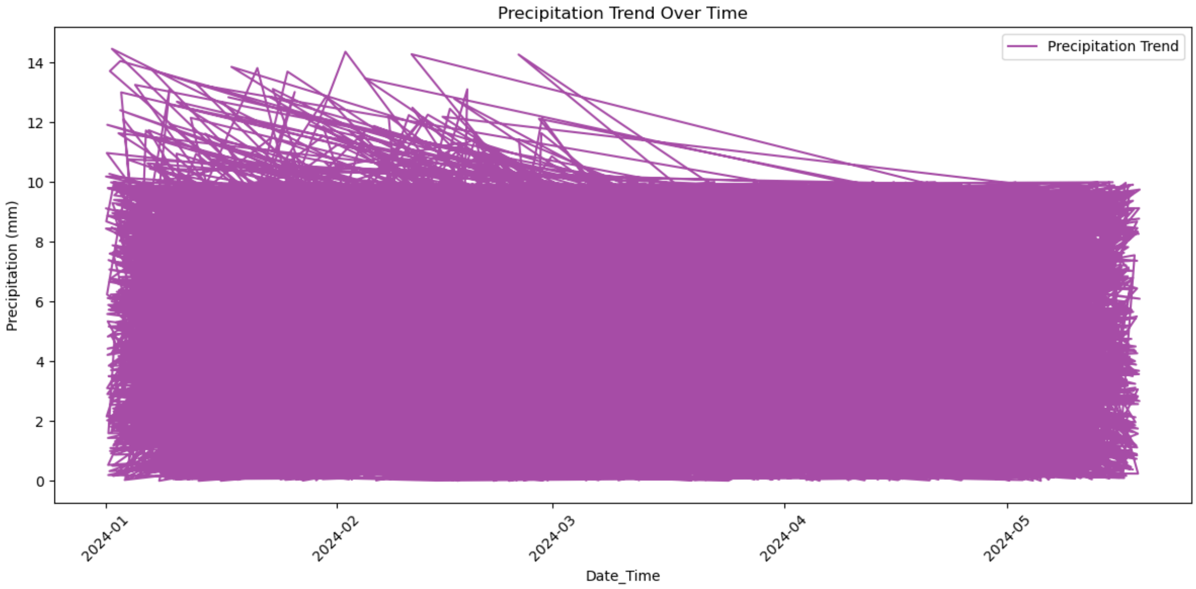
- **Programming Language:** Python

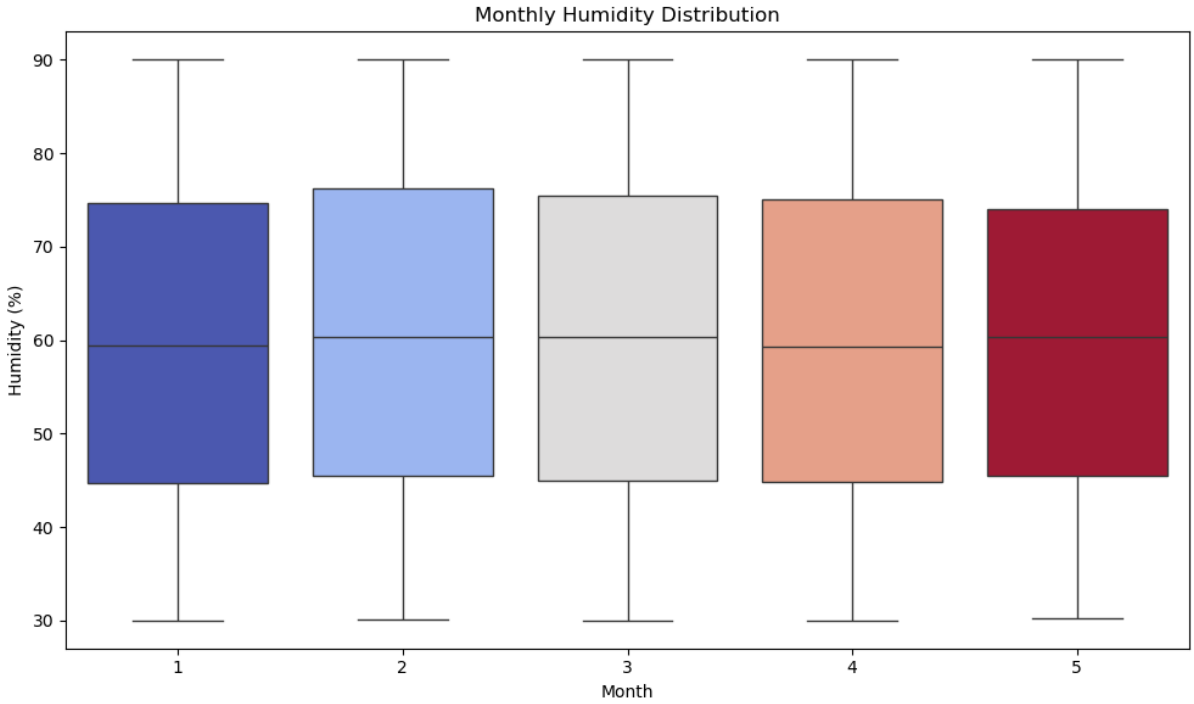
- **Libraries:** Pandas, Matplotlib, Seaborn
- **Dataset:** Weather Data from Kaggle

ge: **Output Description:**

- **Line Plot:** Depicts overall temperature trends, showing variations across seasons.
- **Scatter Plot:** Highlights anomalies, such as extreme temperature deviations.
- **Heatmap:** Visualizes temperature distributions across months and years.






Monthly Humidity Distribution
Humidity (%)
90
80
70
60
50
40
30
1
2
3
4
5
Month

Week 3:

Project 5 - Predicting House Prices

Objective:
Build a machine learning model to predict house prices based on features such as size, location, and amenities. This project demonstrates the practical application of regression modeling.

Code Highlights:

1. **Data Preprocessing:**
   - Cleaned the dataset by removing outliers and imputing missing values.
   - Encoded categorical variables like location using one-hot encodin
2. **Feature Selection:**
   - Selected relevant features for prediction using correlation analysis and domain knowledge.
3. **Model Training:**
   - Trained a Linear Regression model to predict house prices.
   - Split the data into training and testing sets for evaluation.
4. **Evaluation:**
   - Evaluated model performance using Mean Absolute Error (MAE) and R-squared (R^2).

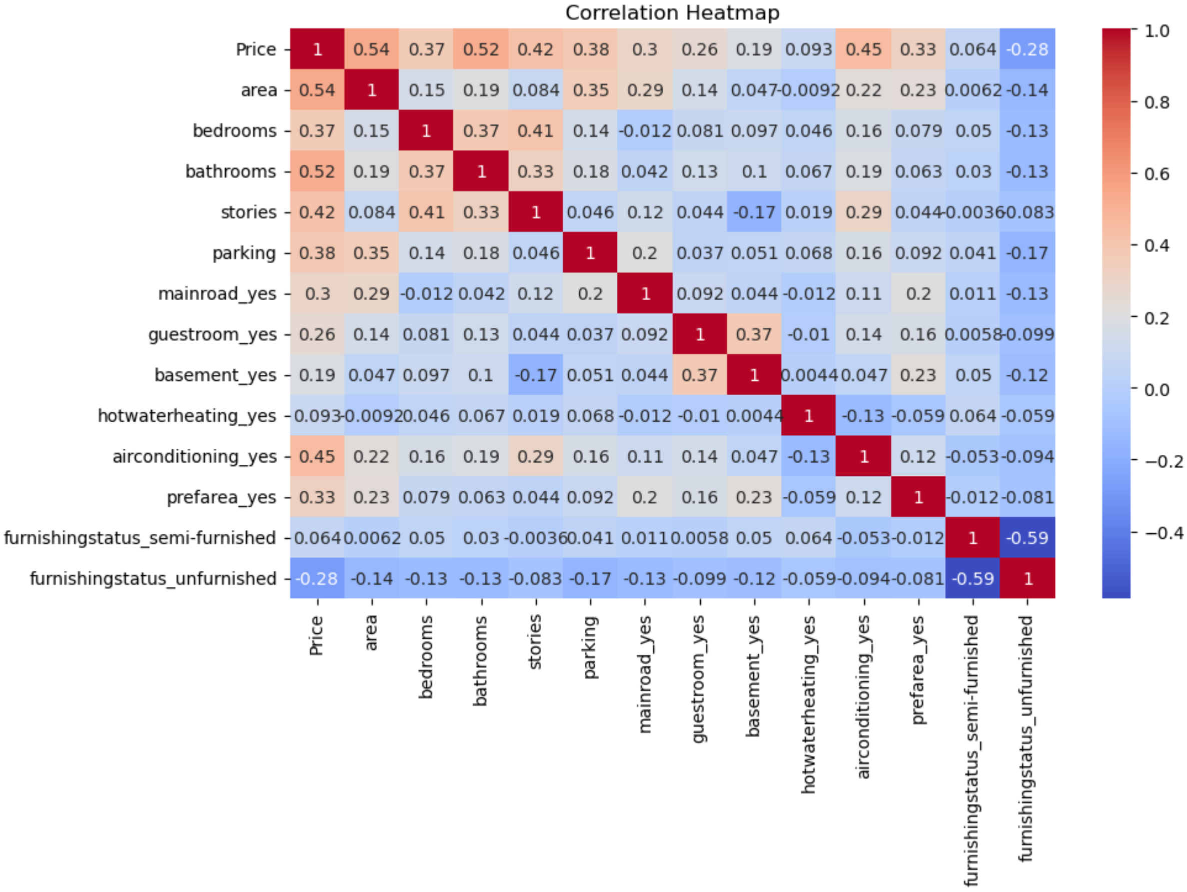
Explanation:
This project uses machine learning to predict house prices, demonstrating the role of regression modeling in real estate pricing. By analyzing features and training models, the project highlights how data-driven decisions can be applied to business problems.

Tools and Technologies Used:

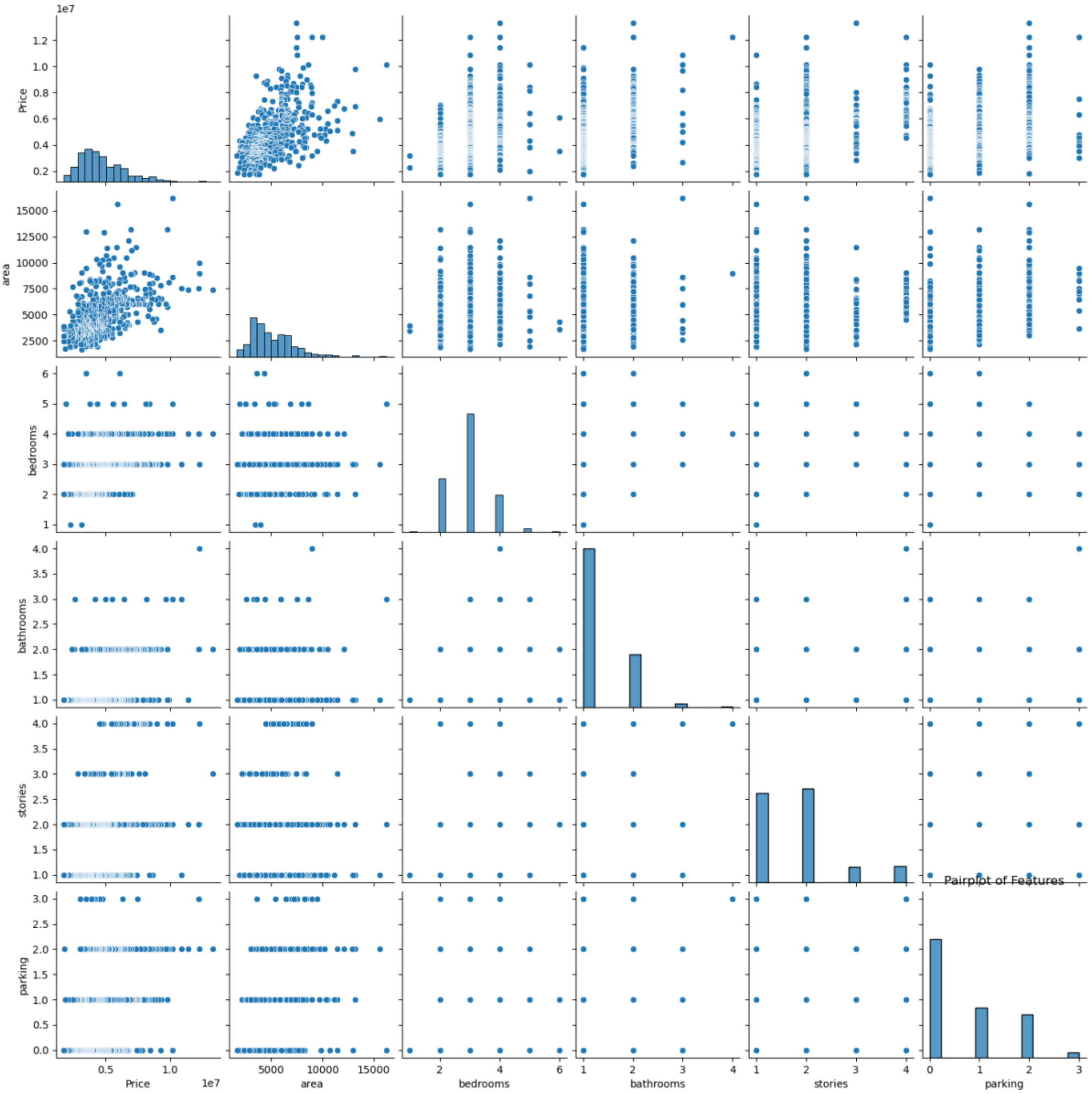
- **Programming Language:** Python
- **Libraries:** Pandas, Scikit-learn, Matplotlib
- **Dataset:** Housing Prices Dataset from Kaggle

Output Description:

- **Correlation Heatmap:** Shows relationships between features and price.
- **Scatter Plot:** Compares actual vs. predicted prices, indicating model accuracy.
- **Performance Metrics:** Includes MAE and R^2 values to assess model efficiency.



Correlation Heatmap



```
Linear Regression Model Performance:

Mean Absolute Error (MAE): 970043.4039201644

Mean Squared Error (MSE): 1754318687330.669

R-squared (R²): 0.6529242642153175


Random Forest Regressor Model Performance:

Mean Absolute Error (MAE): 1021546.0353211008

Mean Squared Error (MSE): 1961585044320.3433

R-squared (R²): 0.611918531405699

Best Random Forest Model R² Score:

0.6003865799717936
```