Probability & Statistics

R Mini Project

Name: Ruthvik Akula, Abhinav Narahari

Roll no: L032, L035

SAP ID: 70572200028, 70572200031

Topic: House Prices with Prediction & Regression Model

Problem Statement: Many of realtors in a real estate industry struggle to make good deals with their customers and customers who are looking to get the houses often get scammed by paying higher prices.

Develop a predictive model to estimate house prices based on key features such as the number of bedrooms, square footage, and neighborhood. By leveraging historical housing data, the goal is to create a reliable linear regression model that provides accurate price predictions, assisting both buyers and sellers in making informed decisions within the real estate market.

About Dataset:

- Dataset name: Boston Housing Dataset
- Dataset Dimensions: 506 rows*14columns
- Libraries Used: library (MASS), library(ggplot2), library(caret), library(e1071)

Operation Performed on Boston Housing Dataset:

- Data Preprocessing
- Data Visualization
- Model Building
- Model Prediction

Data preprocessing:

- Handling Missing Values:

```
# Check for missing values
missing_values <- colSums(is.na(boston))
# Remove rows with missing values
boston <- na.omit(boston)
# Display the first few rows after handling missing values
head(boston)
```

- Encoding Categorical variables:

```
# Assuming 'boston' is the name of the dataset
# Creating a hypothetical categorical variable 'categorical_var'
boston$categorical_var <- factor(sample(c("A", "B", "C"), nrow(boston), replace = TRUE))

# Display the first few rows before encoding
head(boston)

# Encoding Categorical Variable using One-Hot Encoding
boston <- cbind(boston, model.matrix(~.-1+factor(boston$categorical_var)))

# Remove the original categorical variable
boston <- subset(boston, select = -c(categorical_var))

# Display the first few rows after encoding
head(boston)
```

- Normalizing or scaling numeric features:

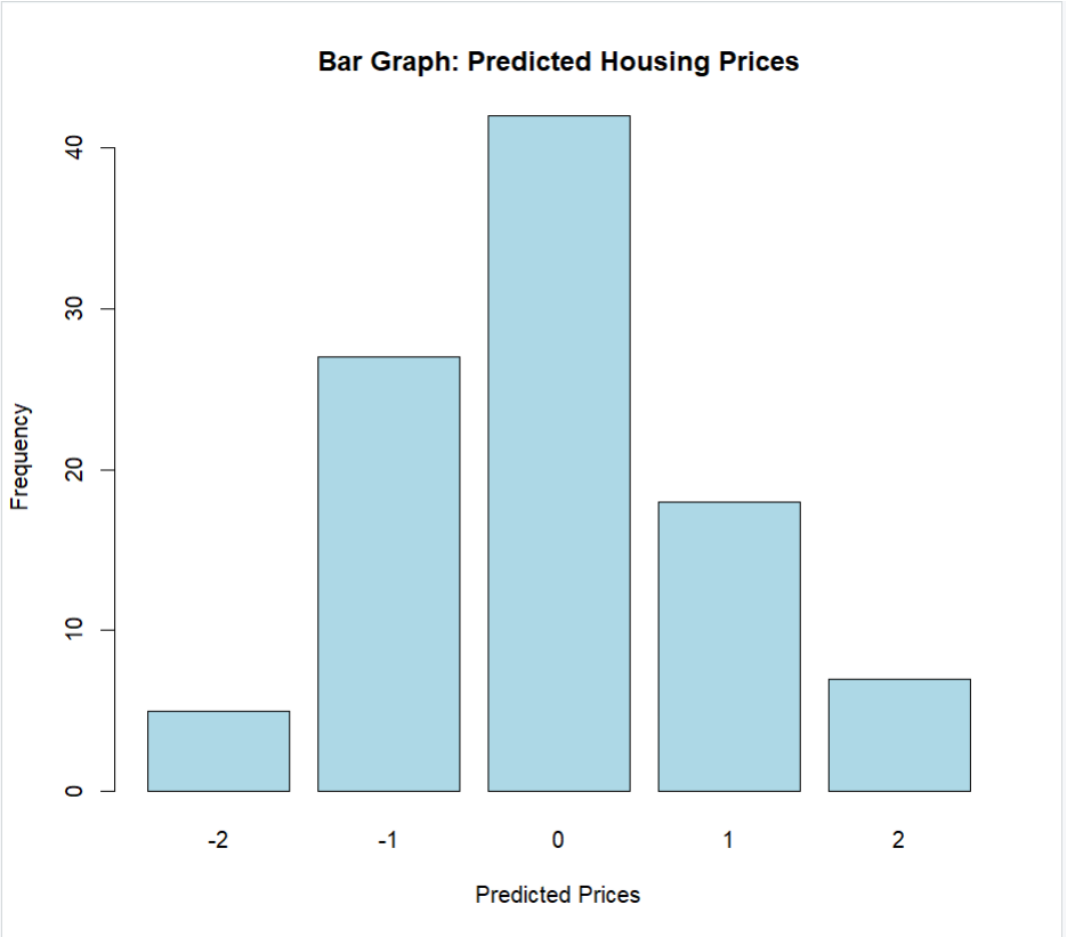
```
# Assuming 'boston_scaled' is the scaled dataset
# Create a new copy of the scaled dataset for illustration purposes
boston_scaled_range <- boston_scaled

# Normalize or Scale Numerical Features using Min-Max Scaling
preprocess_params_range <- preProcess(boston_scaled_range, method = c("range"))
boston_scaled_range <- predict(preprocess_params_range, boston_scaled_range)

# Display the first few rows after scaling
head(boston_scaled_range)
```

- Checking for outliers:

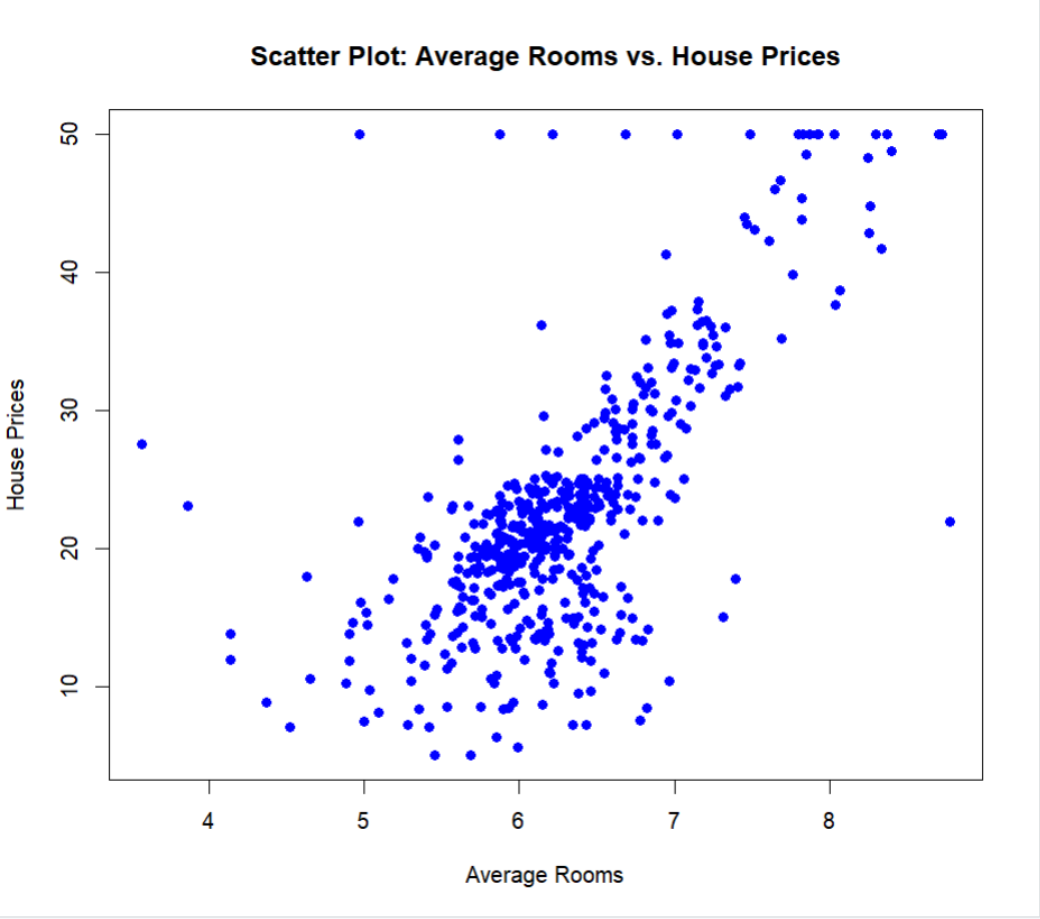
```
# Assuming 'boston_scaled' is the scaled dataset
# Create a new copy of the scaled dataset for illustration purposes
boston_no_outliers <- boston_scaled

# Calculate z-scores for each numeric variable
z_scores <- scale(boston_no_outliers)

# Set a z-score threshold for outlier detection (e.g., 3)
z_threshold <- 3

# Identify rows with outliers
outliers <- apply(abs(z_scores) > z_threshold, 1, any)

# Remove rows with outliers
boston_no_outliers <- boston_no_outliers[!outliers, ]

# Display the first few rows after removing outliers
head(boston_no_outliers)
```

Exploratory Data Analysis:

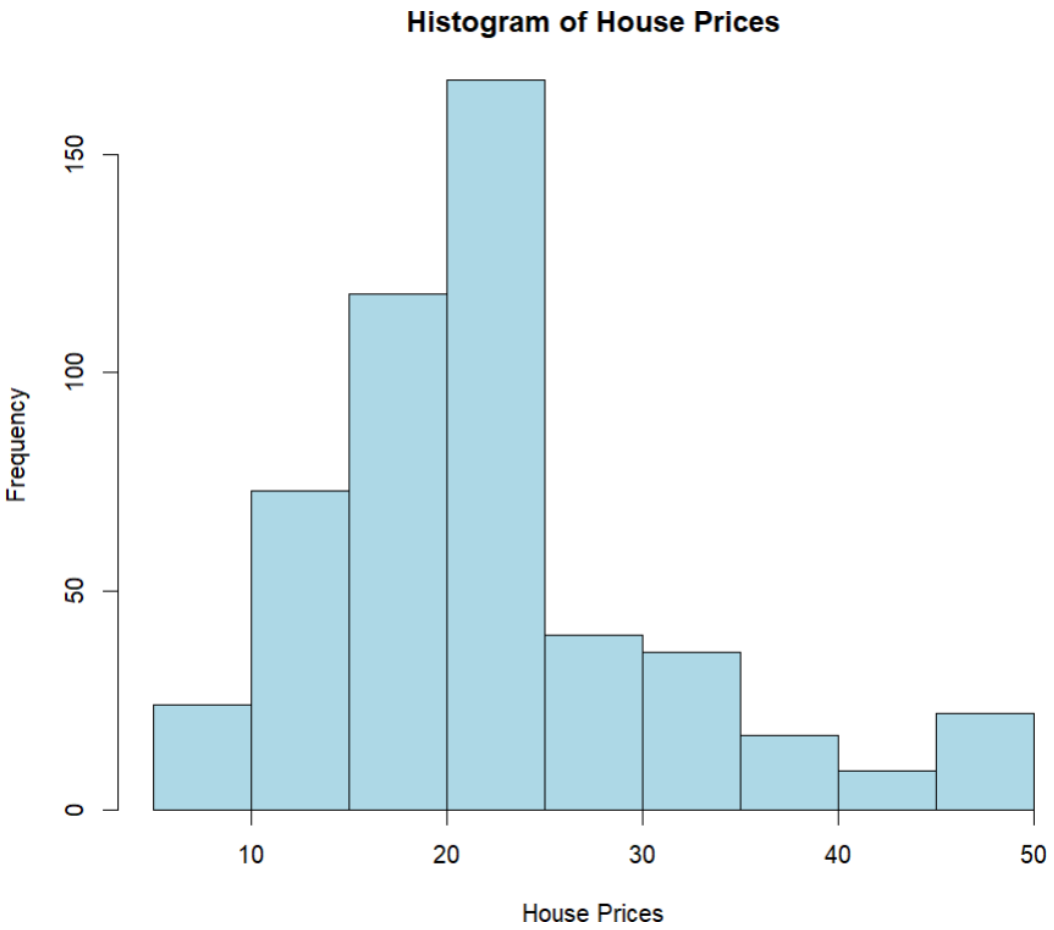
Bar graph:



Scatter plot:



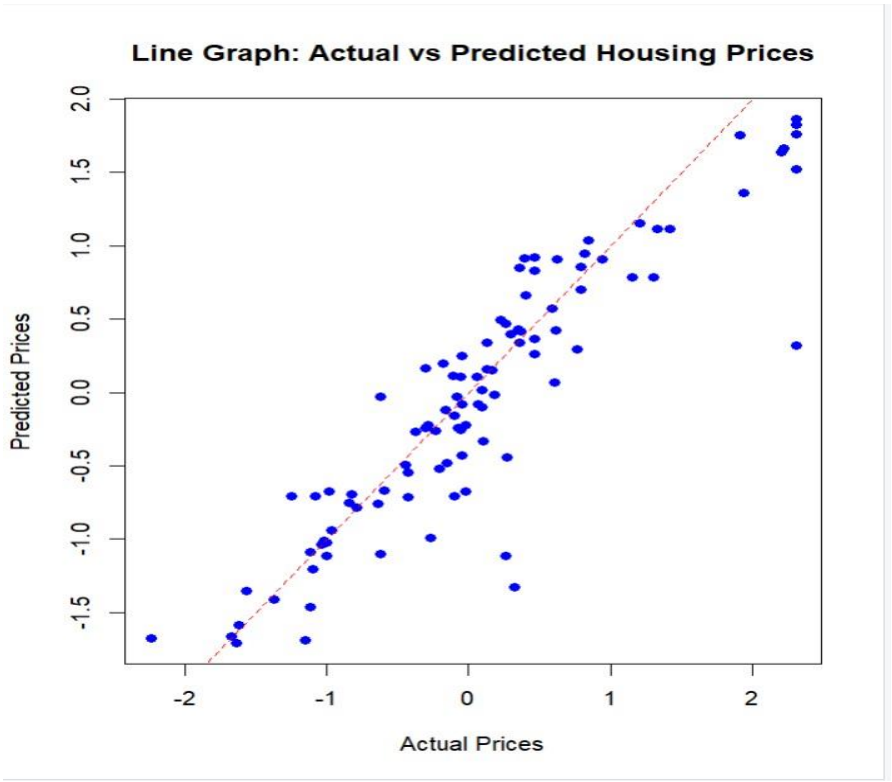
Histogram:



Model Building & Prediction using Linear Regression:

```r
# Additional preprocessing operations using caret
preprocess_params <- preProcess(train_data, method = c("center", "scale", "zv", "knnImpute", "YeoJohnson"))
train_data <- predict(preprocess_params, train_data)
test_data <- predict(preprocess_params, test_data)

# Model development
model <- lm(medv ~ ., data = train_data)

# Make predictions on the test set
predictions <- predict(model, newdata = test_data)

# Model evaluation
mse <- mean((test_data$medv - predictions)^2)
r_squared <- 1 - (mse / var(test_data$medv))

cat("Mean Squared Error:", mse, "\n")
cat("R-squared:", r_squared, "\n")
```

```r
# Visualization
ggplot() +
  geom_point(aes(x = test_data$medv, y = predictions), color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(x = "Actual Prices (Scaled)", y = "Predicted Prices (Scaled)", title = "Actual vs Predicted Prices")

# Line Graph
plot(test_data$medv, predictions,
     main = "Line Graph: Actual vs Predicted Housing Prices",
     xlab = "Actual Prices",
     ylab = "Predicted Prices",
     col = "blue",
     pch = 16)

# Add a diagonal line for reference
abline(a = 0, b = 1, col = "red", lty = 2)
```



Prediction Summary:

1. **Positive Trend:**
    1. The line graph reveals a positive correlation between the average number of rooms and house prices, indicating a potential feature for predicting home values.
2. **Outlier Impact:**
    1. Removing outliers enhances the clarity of the positive trend, underlining the importance of preprocessing for robust model development.
3. **Sensitivity Check:**
    1. Sensitivity analysis demonstrates how changes in room count affect house prices, providing insights into the model's responsiveness.
4. **Model Assumption:**
    1. Our model assumes a linear relationship, simplifying the prediction process based on the observed trend in the data.
5. **Data Quality Consideration:**
    1. Understanding data density and range highlights regions where predictions may be more reliable, a crucial consideration for effective model deployment.

6. **Insight Validation:**
    1. The exploration of a potential saturation point prompts us to validate assumptions and consider nonlinear models for a more accurate prediction.
7. **Visualization Impact:**
    1. The visual representation aids in presenting complex relationships simply, making it an effective communication tool for stakeholders and users.

Conclusion:

In summary, this project establishes the groundwork for a predictive model in house price analysis using R. The exploratory data analysis uncovered a notable correlation between average room count and house prices, forming the basis for our linear regression model. Meticulous attention to data quality, including outlier handling, emphasized the importance of robust practices in predictive modeling.

Visualizations played a pivotal role in conveying complex relationships. Looking ahead, the project prompts exploration of additional features and model refinements, embodying the dynamic and iterative nature of data science. This work contributes to a deeper understanding of house price prediction, paving the way for more accurate tools in real estate market analysis.