Personal Project #4

Housing Market Data

Perform EDA, Data Visualization, Hypothesis Testing and Linear Regression in R

Jihan Rana Ayunda Dewana

Contents.

Case Description

In real estate, the value of a home can be influenced by many factors, starting from crime rates, number of bedrooms, availability of public transportation access, and number of schools to median home value, which is data related to listings from highest to lowest prices. Later, several stages of analysis will begin to form a simple linear regression that predicts the market value of a home based on strong predictors.

#1. EDA

	Crime Rate	Average Rooms	Public Transport Access	Number of Schools	Median Home Value
1	*NA*	5.585324	10	3	47.90077
2	2.65433925	5.395206	3	6	41.53910
3	4.61922134	6.033965	9	4	48.51757
4	6.80757464	5.418335	10	5	42.50757
5	2.41461656	6.189320	2	4	51.39125
6	2.41465761	5.964833	6	4	49.64657
7	6.94803204	5.832736	7	4	48.76959
8	4.91858682	5.364705	9	4	38.21798
9	1.82631404	5.596260	3	6	44.69063
10	*NA*	6.528774	10	4	52.59876

Before starting the analysis, we need to do the EDA process. This stage functions to discover the condition of the data to be processed, both from the positive and negative sides

Number of Dataset: 506 rows and 5 columns

Variable on Dataset:

- **Crime.Rate**: Local crime rate per capita
- **Average.Rooms**: Average number of rooms in homes
- **Highway.Access**: Proximity to highways
- **Pupil.Teacher.Ratio**: Ratio of students to teachers in local schools
- **Median.Home.Value**: Median value of houses

After performing the EDA process, it was found that some data were blank. So, it is necessary to use a technique to input empty data, one of which is by the average value of the variable itself.

#2. Descriptive Statistic

The purpose of descriptive statistics is to summarize and make it coherent. Descriptive statistics can also provide an overview of the problem that will be analyzed.

Through descriptive statistics, we can discover mean, median, and mode values from the lowest and highest values. Moreover, the amount of missing data from each variable is also visible.

```
> print(summary(housing_data))
   Crime Rate         Average Rooms    Public Transport Access
 Min.   : 0.005305   Min.   :4.112    Min.   : 1.000
 1st Qu.: 1.299938   1st Qu.:5.598    1st Qu.: 3.000
 Median : 3.031481   Median :6.033    Median : 5.000
 Mean   : 3.137415   Mean   :6.026    Mean   : 5.421
 3rd Qu.: 4.584798   3rd Qu.:6.460    3rd Qu.: 8.000
 Max.   :12.631829   Max.   :7.801    Max.   :10.000
 NA's   :25          NA's   :15
 Number of Schools Median Home Value
 Min.   : 0.000    Min.   :31.55
 1st Qu.: 4.000    1st Qu.:43.23
 Median : 5.000    Median :46.91
 Mean   : 4.992    Mean   :47.10
 3rd Qu.: 6.000    3rd Qu.:50.85
 Max.   :10.000    Max.   :62.56
```

#3. Correlation

```
> correlation_matrix
                        Crime Rate Average Rooms Public Transport Access Number of Schools Median Home Value
Crime Rate              1.00000000   0.109411375              0.014246404       0.024421905       0.091610332
Average Rooms           0.10941138   1.000000000             -0.003768297       0.005000546       0.888351070
Public Transport Access 0.01424640  -0.003768297              1.000000000       0.035876982       0.010709022
Number of Schools       0.02442190   0.005000546              0.035876982       1.000000000       0.004667281
Median Home Value       0.09161033   0.888351070              0.010709022       0.004667281       1.000000000
```

The function of the correlation test is to see the relationship between variables. Based on the result, we spotted two variables that have the highest correlation. Specifically, there are average rooms and median home values of 0.888351070.

#4. Missing Data Imputation

As previously explained, to fill in the empty data, use the average value of the variable itself with R as below.

```
#Fill in missing values with the mean for each variable
names(housing_data)
#[1] "Crime Rate"
#[2] "Average Rooms"
#[3] "Public Transport Access"
#[4] "Number of Schools"
#[5] "Median Home Value"
housing_data$`Crime Rate`[is.na(housing_data$`Crime Rate`)] <- mean(housing_data$`Crime Rate`,na.rm = TRUE)
housing_data$`Average Rooms`[is.na(housing_data$`Average Rooms`)] <- mean(housing_data$`Average Rooms`,na.rm = TRUE)
housing_data$`Public Transport Access`[is.na(housing_data$`Public Transport Access`)] <- mean(housing_data$`Public Transport Access`,na.rm = TRUE)
housing_data$`Number of Schools`[is.na(housing_data$`Number of Schools`)] <- mean(housing_data$`Number of Schools`,na.rm = TRUE)
housing_data$`Median Home Value`[is.na(housing_data$`Median Home Value`)] <- mean(housing_data$`Median Home Value`,na.rm = TRUE)
```

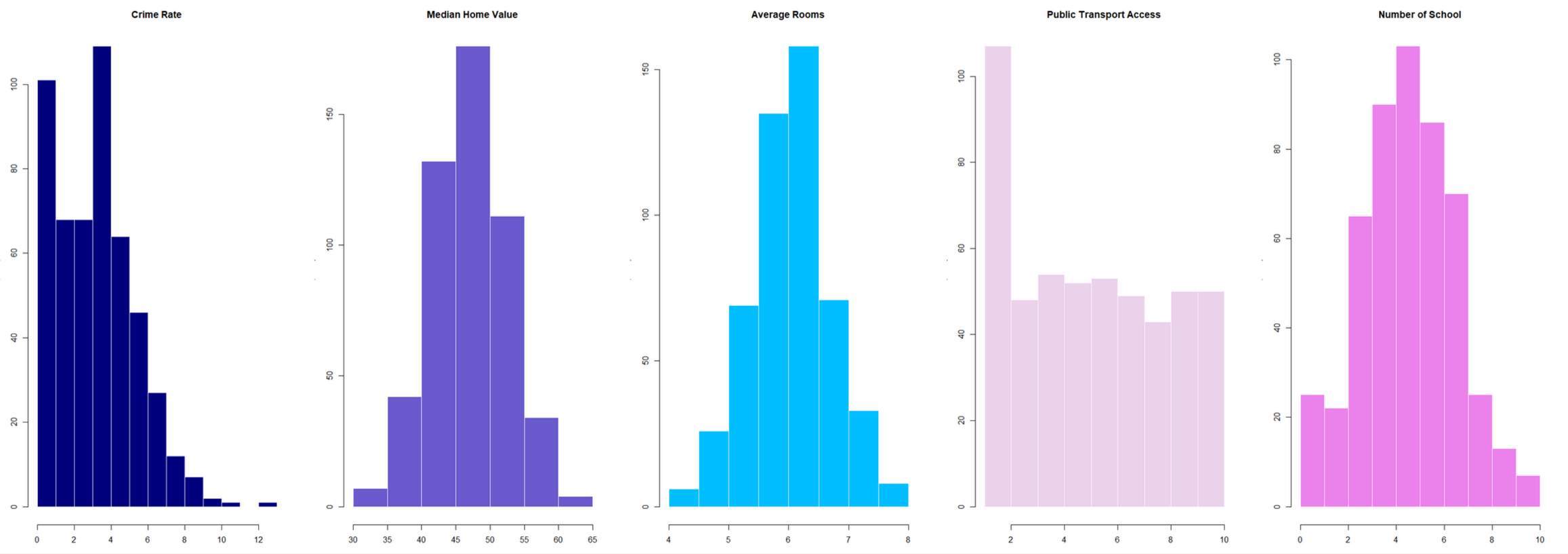
After filling in the data with the mean value, it is necessary to re-check it through descriptive statistics.

The results discover there are non-null data.

```
> summary (data)
  Crime Rate        Average Rooms   Public Transport Access
Min.   : 0.005305   Min.   :4.112   Min.   : 1.000
1st Qu.: 1.375937   1st Qu.:5.605   1st Qu.: 3.000
Median : 3.137415   Median :6.026   Median : 5.000
Mean   : 3.137415   Mean   :6.026   Mean   : 5.421
3rd Qu.: 4.533860   3rd Qu.:6.451   3rd Qu.: 8.000
Max.   :12.631829   Max.   :7.801   Max.   :10.000
Number of Schools Median Home Value Crime.Category
Min.   : 0.000    Min.   :31.55     Length:506
1st Qu.: 4.000    1st Qu.:43.23     Class :character
Median : 5.000    Median :46.91     Mode  :character
Mean   : 4.992    Mean   :47.10
3rd Qu.: 6.000    3rd Qu.:50.85
Max.   :10.000    Max.   :62.56
```

#5. Data Visualization and Hyphotesis Testing

Data Visualization: Histogram of Housing Data Variables



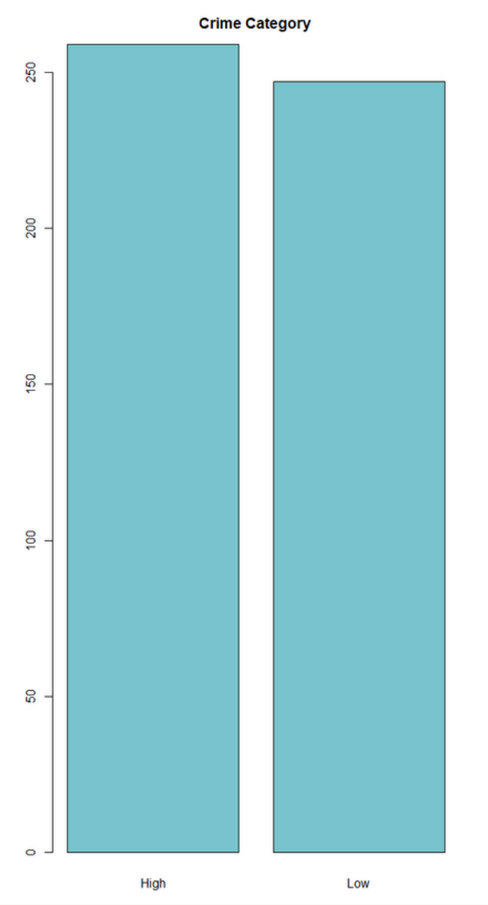
#5. Data Visualization and Hyphotesis Testing

```
> levene_test #semua variabel
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  1.1143 0.2917
      442
```

Levene's test measures how much variance among two or more different sets. The results of this test can indicate whether the axis indicates homogeneity or not. If the p-value is preponderant than 0.05, then the variances are not significantly different. It means that the assumption of homogeneity of variance is covered. If otherwise, then there is a significant difference between the variances.

The p-value based on the Levene test is 0.2917. It indicates that the data is homogeneous. These fulfill the prerequisites for the independent sample t-tests and ANOVA.

#5. Data Visualization and Hyphotesis Testing




```r
#Part 3 - Hypothesis testing
# 1. Define "high" and "low" crime rates based on the median of the Crime.Rate variable
# Categorize crime rate into two levels: Low and High
housing_data$crime_category <- ifelse(housing_data$`Crime Rate` < median(housing_data$`Crime Rate`), "Low", "High")
housing_data$crime_category

unique(housing_data$crime_category)
levels(housing_data$crime_category)

View(housing_data)

# Simple Bar Plot - Crime Category
counts <- table(housing_data$crime_category)
barplot(counts, main="Crime Category",
        col="cadetblue3",
        xlab="Number of Crime Category")
counts
```

In these steps, we want to classify the crime rate based on the median crime rating value divided into two categories: high and low. From 506 data collected, it was discovered that 259 houses had a high crime rate (51.2%) , and 247 houses had a low crime rate (48.8%)

#5. Data Visualization and Hyphotesis Testing

```
> shapiro_test_high

        Shapiro-Wilk normality test

data:  housing_data$`Median Home Value`[housing_data$crime_category == "High"]
W = 0.99612, p-value = 0.7717

> shapiro_test_low

        Shapiro-Wilk normality test

data:  housing_data$`Median Home Value`[housing_data$crime_category == "Low"]
W = 0.99737, p-value = 0.9598
```

The Shapiro test is one of the tests to determine the normality of data. If it is done based on the crime category, both high and low have a p-value above 0.05. Accordingly, it can be concluded that the data is normally distributed.

#6. Linear Regression

```
Call:
lm(formula = `Median Home Value` ~ `Average Rooms` + `Crime Rate` +
    `Public Transport Access` + `Number of Schools`, data = housing_data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5519 -1.7439 -0.0109  1.7734 14.0041

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.7286643  1.1658103   3.198  0.00147 **
`Average Rooms`            7.1926459  0.1820665  39.506  < 2e-16 ***
`Crime Rate`              -0.0124100  0.0558335  -0.222  0.82420
`Public Transport Access`  0.0140736  0.0421406   0.334  0.73854
`Number of Schools`       -0.0006922  0.0609198  -0.011  0.99094
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.707 on 501 degrees of freedom
Multiple R-squared:  0.7588,    Adjusted R-squared:  0.7569
F-statistic:   394 on 4 and 501 DF,  p-value: < 2.2e-16
```

- **Intercept (0.00147)**. The estimated value when the Average Rooms, Crime Rate, Public Transportation Access, and Number of Schools are zero—assuming all other variables are held constant. It acts as a baseline value for the Median Home Value.
- **Prob(F-statistic); also P-value.** It tells us the probability of obtaining an F-statistic **(< 2.2e-16)** as extreme as ours if the null hypothesis is true. Our p-value is 0.00, less than the commonly used significance level of 0.05, so we conclude that our model provides a significant fit to the data. In other words, all independent variables significantly explain the Median Home Value.
- **Coefficient of Determination (R-squared, 0.7588).** That means that all independent variables can explain about 75.88% of the variability in the Median Home Value. The remaining 24.12% could be explained by other factors not included in the model.
- **Besides all independent variables, only one has a significant relationship, with a p-value below 0.05, namely Average Rooms.**

Thank You!

Open Discussion and Feedback

Personal Project #4
Housing Data Market Analysis in R

jihan.dewana99@gmail.com

Jihan Dewana