# CDS6224: Statistical Data Analysis

Mohamad Adriel Hakiem
Student ID: 243UC2462B

2025-04-29

## Assignment 1: Descriptive Statistics & Data Visualization Analysis

### Dataset Selection and Justification

Dataset selection is a critical aspect of any analytical study, as it determines the quality and relevance of the findings. In this essay, the House Prices in Malaysia 2025 dataset from Kaggle [1] has been chosen due to its real-world property data that directly pertains to Malaysia's economic and social context. This dataset contains essential numerical variables such as Median_Price and Median_PSF, along with categorical variables like Township, Area, State, Tenure, and Type, meeting the required criteria for analysis.

The selected dataset allows for the exploration of significant research inquiries, such as:

- Consistency of property prices across different regions.

- Correlation between price per square foot and total property price.

A deeper understanding of the variability, distribution patterns, and relationships within Malaysia's housing market can be achieved by delving into these questions. Moreover, the dataset facilitates the application of appropriate statistical techniques and the critical interpretation of results.

In conclusion, the House Prices in Malaysia 2025 dataset's comprehensive nature and relevance to Malaysia's real estate sector make it an ideal choice for exploring various facets of the housing market. Through rigorous analysis and interpretation, this dataset has the potential to yield valuable insights for this assignment.

```
# Set working directory
setwd("/Users/ad/Desktop/Studies/MMU/Trimester 2/Statistical Data Analysis/Assignment 1")

# Load dataset
houseData <- read.csv("malaysia_house_price_data_2025.csv")

# View first few rows for checking
head(houseData)
```

```
##                     Township          Area  State    Tenure
## 1         SCIENTEX SUNGAI DUA Tasek Gelugor Penang  Freehold
## 2               BANDAR PUTRA         Kulai  Johor  Freehold
## 3 TAMAN LAGENDA TROPIKA TAPAH   Chenderiang  Perak  Freehold
## 4      SCIENTEX JASIN MUTIARA        Bemban Melaka  Freehold
## 5          TAMAN LAGENDA AMAN         Tapah  Perak Leasehold
## 6           TAMAN IMPIAN EMAS        Tebrau  Johor  Freehold
##                           Type Median_Price Median_PSF Transactions
## 1               Terrace House       331800        304          593
```

```
## 2 Cluster House, Terrace House       590900        322          519
## 3                Terrace House       229954        130          414
## 4                Terrace House       255600        218          391
## 5                Terrace House       219300        168          363
## 6        Terrace House, Semi D       738000        328          349
```

## Part A: Exploratory Descriptive Analysis

### Variable Selection

The two variables selected for analysis are 'Median_Price' and 'Median_PSF' from the dataset.

```
# Variable 1
medianPrice <- houseData$Median_Price

# Variable 2
medianPSF <- houseData$Median_PSF
```

### Descriptive Statistics

In this section, descriptive statistical measures are calculated for the two selected numerical variables: Median_Price and Median_PSF.

The measures include mean, median, mode, standard deviation, variance, range, quartile, interquartile range (IQR), and coefficient of variation (CV).

These calculations provide insights into the central tendency, spread, and overall distribution characteristics of the property price data.

```
round(mean(medianPrice, na.rm = TRUE), 2)
```

```
## [1] 490685.4
```

```
round(mean(medianPSF, na.rm = TRUE), 2)
```

```
## [1] 328.86
```

The mean values of Median_Price (RM 490,685.40) and Median_PSF (RM 328.86 per square foot) indicate the typical overall property price and price density across townships.

```
round(median(medianPrice, na.rm = TRUE), 2)
```

```
## [1] 390000
```

```
round(median(medianPSF, na.rm = TRUE), 2)
```

```
## [1] 293
```

The median values of Median_Price (RM390,000) and Median_PSF (RM293 per square foot) represent the central tendency of overall property prices and price density, offering a robust measure unaffected by extreme values.

```r
get_mode <- function(x) {
  uniqueX <- unique(x)
  mode_value <- uniqueX[which.max(tabulate(match(x, uniqueX)))]
  format(mode_value, scientific = FALSE)
  }

# Calculate Mode for Median_Price and Median_PSF
get_mode(medianPrice)
```

```
## [1] "300000"
```

```r
get_mode(medianPSF)
```

```
## [1] "179"
```

The mode of Median_Price is RM300,000, and the mode of Median_PSF is RM179 per square foot, indicating the most common sale price and price density in the sample.

```r
round(sd(medianPrice, na.rm = TRUE), 2)
```

```
## [1] 468632.2
```

```r
round(sd(medianPSF, na.rm = TRUE), 2)
```

```
## [1] 193.28
```

The standard deviation of Median_Price is RM468,632.20 and of Median_PSF is RM193.28 per square foot, reflecting the typical dispersion around their respective averages.

```r
round(var(medianPrice, na.rm = TRUE), 2)
```

```
## [1] 219616147587
```

```r
round(var(medianPSF, na.rm = TRUE), 2)
```

```
## [1] 37357.83
```

The variance of Median_Price is 219,616,147,587 and of Median_PSF is 37,357.83, indicating the squared dispersion of property prices and price density in the sample.

```r
round(range(medianPrice, na.rm = TRUE), 2)
```

```
## [1]    27049 11420500
```

```r
round(range(medianPSF, na.rm = TRUE), 2)
```

```
## [1]   38 3017
```

The range for Median_Price spans from RM27,049 to RM11,420,500, indicating a large spread in property prices. The range for Median_PSF spans from RM38 to RM3,017 per square foot, showing the substantial version in price density across different properties.

```r
round(quantile(medianPrice, probs = c(0.25, 0.5, 0.75), na.rm = TRUE), 2)
```

```
##    25%    50%    75%
## 269950 390000 573500
```

```r
round(quantile(medianPSF, probs = c(0.25, 0.5, 0.75), na.rm = TRUE), 2)
```

```
## 25% 50% 75%
## 201 293 412
```

The quartiles of Median_Price are RM269,950 (25th percentile), RM390,000 (50th percentile), and RM573,500 (75th percentile), while for Median_PSF they are RM201, RM293, RM412 per square foot, indicating how the middle 50% of values are distributed around the median.

```r
round(IQR(medianPrice, na.rm = TRUE), 2)
```

```
## [1] 303550
```

```r
round(IQR(medianPSF, na.rm = TRUE), 2)
```

```
## [1] 211
```

The interquartile range (IQR) of Median_Price is RM303,550 and for Median_PSF is RM211 per square foot, representing the spread of the middle 50% of values and helping identify potential outliers.

```r
round(sd(medianPrice, na.rm = TRUE) / mean(medianPrice, na.rm = TRUE), 4)
```

```
## [1] 0.9551
```

```r
round(sd(medianPSF, na.rm = TRUE) / mean(medianPSF, na.rm = TRUE), 4)
```

```
## [1] 0.5877
```

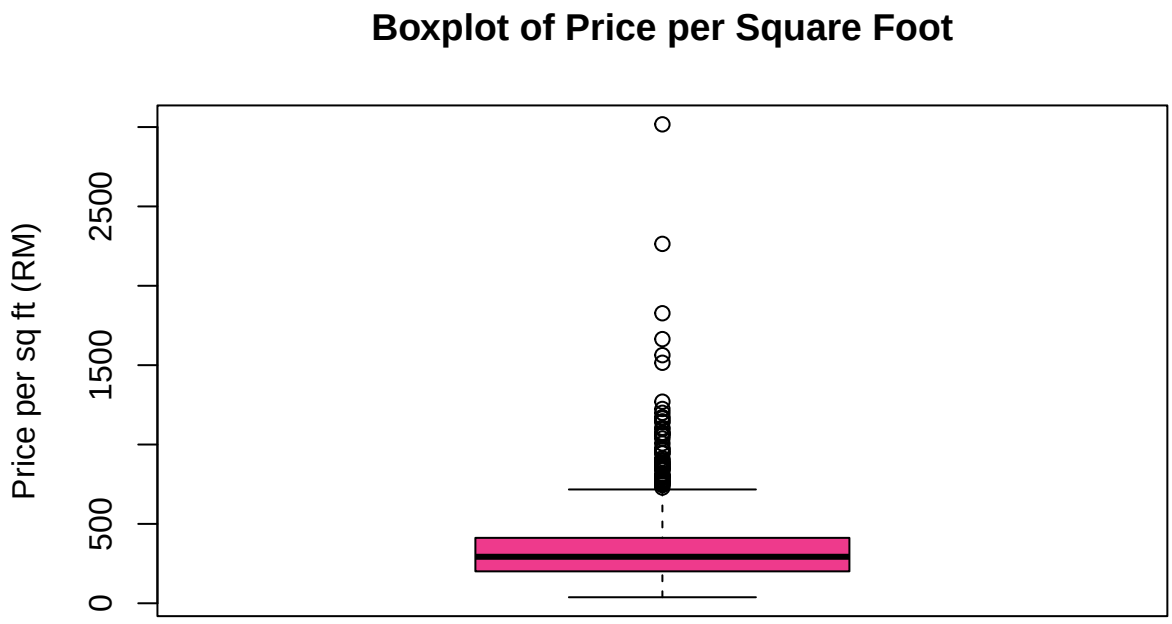
The coefficient of variation for Median_Price is 0.9551 and for Median_PSF is 0.5877, indicating the relative variability of each variable; a lower CV suggests more consistency in values.

Median_Price shows a greater level of dispersion with a standard deviation of RM468,632.20, compared to RM193.28 for Median_PSF. When considering relative variability, the coefficient of variation further confirms that Median_PSF is more consistent (CV = 0.5877) than Median_Price (CV = 0.9551). This suggests that property prices fluctuate more widely than price per square foot across the sample.

### Graphical Analysis

```r
boxplot(medianPrice,
        main = "Boxplot of Property Prices",
        ylab = "Price (RM)",
        col = "yellow2")
```


```
boxplot(medianPSF,
        main = "Boxplot of Price per Square Foot",
        ylab = "Price per sq ft (RM)",
        col = "violetred2")
```

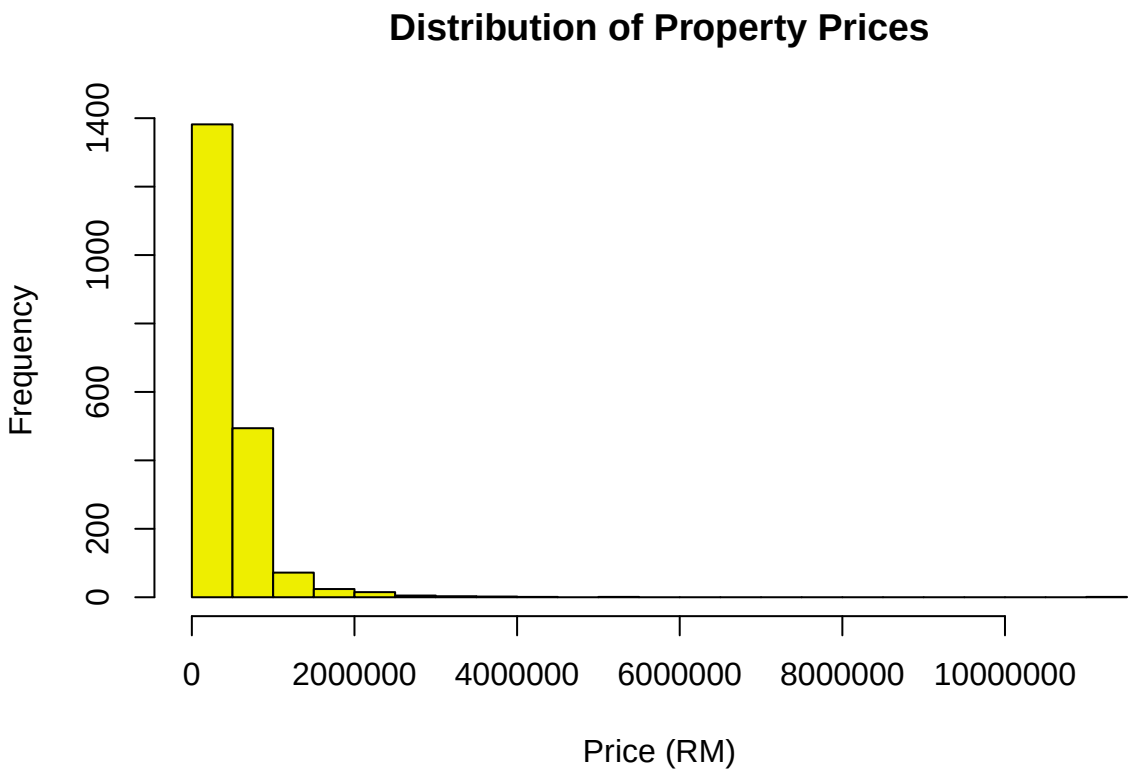



The boxplots show that both property prices and price per square foot are right-skewed with the presence of several outliers. In the case of property prices, the skew is more extreme, with a long upper whisker and a large number of high-end outliers indicating that most properties are priced on the lower end, while a few are significantly more expensive.

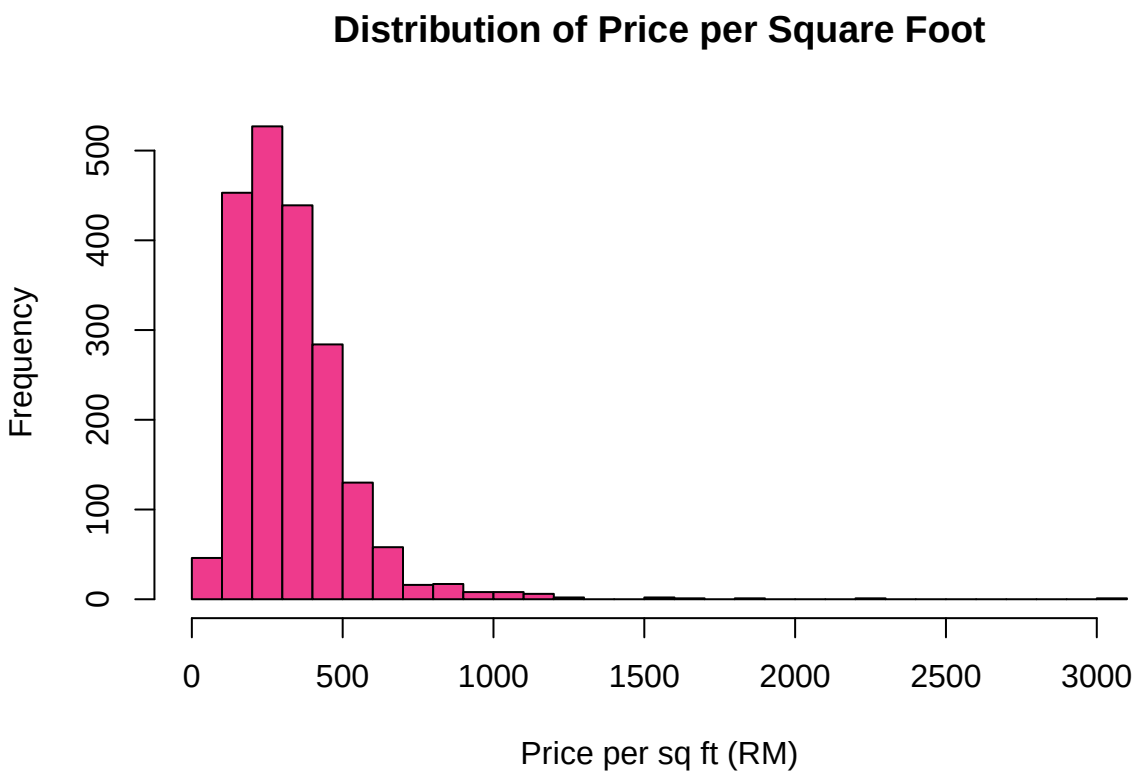
Similarly, the distribution of price per square foot shows moderate skewness, with values clustered toward the lower range and a smaller number of higher-priced outliers. Both variables display asymmetry and variability, especially in premium property segments.

```
options(scipen = 999)

hist(medianPrice,
     main = "Distribution of Property Prices",
     xlab = "Price (RM)",
     col = "yellow2",
     breaks = 30)
```




```r
hist(medianPSF,
     main = "Distribution of Price per Square Foot",
     xlab = "Price per sq ft (RM)",
     col = "violetred2",
     breaks = 30)
```




Both histograms reveal that Malaysia's housing market has a concentration of properties in the lower to middle price ranges, with fewer properties in the higher price brackets. The price per square per foot distribution is more concentrated than the total price distribution, suggesting that while property sizes vary significantly, the price per square foot tends to be more consistent across the market.

These distributions align with what we saw in the earlier boxplots, confirming the presence of outliers and the right-skewed nature of the data. The histograms provide a more detailed view of how the values are distributed across different price ranges.

## Part B: Distributional Assessment

### Variable Selection

For the distributional assessment in Part B, the chosen variable is **Median_Price** from Part A. This variable represents the median property price in Malaysia, which is a key measure for understanding the overall distribution of property prices in the dataset.

### Empirical Rule

To apply the empirical rule, we first calculate the mean and standard deviation of the **median property price (Median_Price)**.

```
# Calculate the mean and standard deviation for **Median_Price**
mean_medianPrice <- round(mean(medianPrice, na.rm = TRUE), 2)
sd_medianPrice <- round(sd(medianPrice, na.rm = TRUE), 2)
```

```
mean_medianPrice
```

```
## [1] 490685.4
```

```
sd_medianPrice
```

```
## [1] 468632.2
```

The mean of the median property price (Median_Price) is RM490,685.40, while the standard deviation is RM468,632.20. These values will be used to evaluate the spread of data according to the empirical rule.

```
# Define the intervals
within1sd <- medianPrice > (mean_medianPrice - sd_medianPrice) &
  medianPrice < (mean_medianPrice + sd_medianPrice)
within2sd <- medianPrice > (mean_medianPrice - 2*sd_medianPrice) &
  medianPrice < (mean_medianPrice + 2*sd_medianPrice)
within3sd <- medianPrice > (mean_medianPrice - 3*sd_medianPrice) &
  medianPrice < (mean_medianPrice + 3*sd_medianPrice)

# Calculate proportions
prop_within1sd <- round(mean(within1sd, na.rm = TRUE) * 100, 2)
prop_within2sd <- round(mean(within2sd, na.rm = TRUE) * 100, 2)
prop_within3sd <- round(mean(within3sd, na.rm = TRUE) * 100, 2)

# Display results
prop_within1sd
```

```
## [1] 93.05
```

```
prop_within2sd
```

```
## [1] 96.85
```

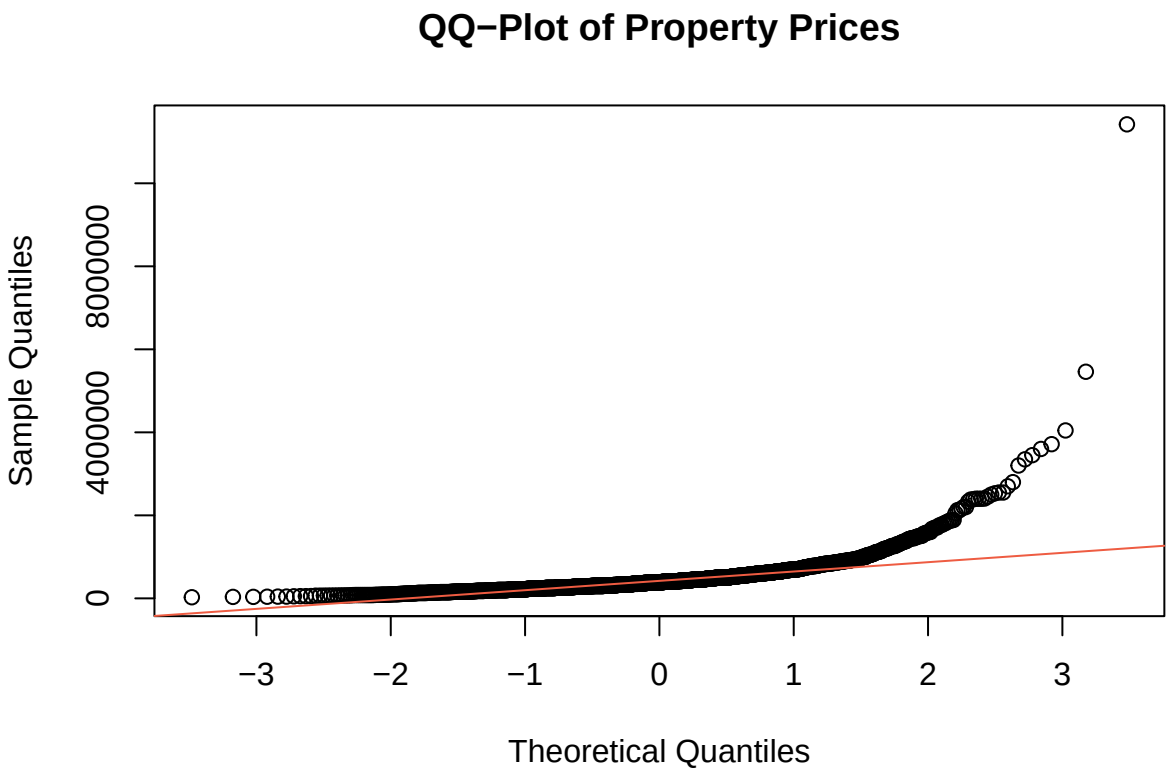
```
prop_within3sd
```

```
## [1] 98.6
```

Based on the empirical rule analysis, approximately 93.05% of property prices within one standard deviation of the mean, 96.85% within two standard deviations, and 98.60% within three. These proportions are higher than the expected 68%-95%-99.7% for a normal distribution, suggesting that the data is tightly clustered around the mean but not perfectly normal, possibly due to the influence of outliers or skewness.

### Normality Assessment

```
# QQ-plot
qqnorm(medianPrice, main = "QQ-Plot of Property Prices")
qqline(medianPrice, col = "tomato2")
```




```
# Shapiro-Wilk Test
shapiro.test(medianPrice)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  medianPrice
## W = 0.5425, p-value < 0.00000000000000022
```

The QQ-plot shows clear deviation from the reference line, especially at the upper tail, indicating that property prices are right-skewed and do not follow a normal distribution. The presence of extreme values further confirms this.

The Shapiro-Wilk test yields a $\mathbf{W = 0.5425}$ and a $\mathbf{p\text{-value} < 2.2e\text{-}16}$, which is far below the 0.05 significance threshold. This provides strong statistical evidence to reject the null hypothesis of normality. Therefore, the distribution of property prices is **not normal**.

### Method Justification

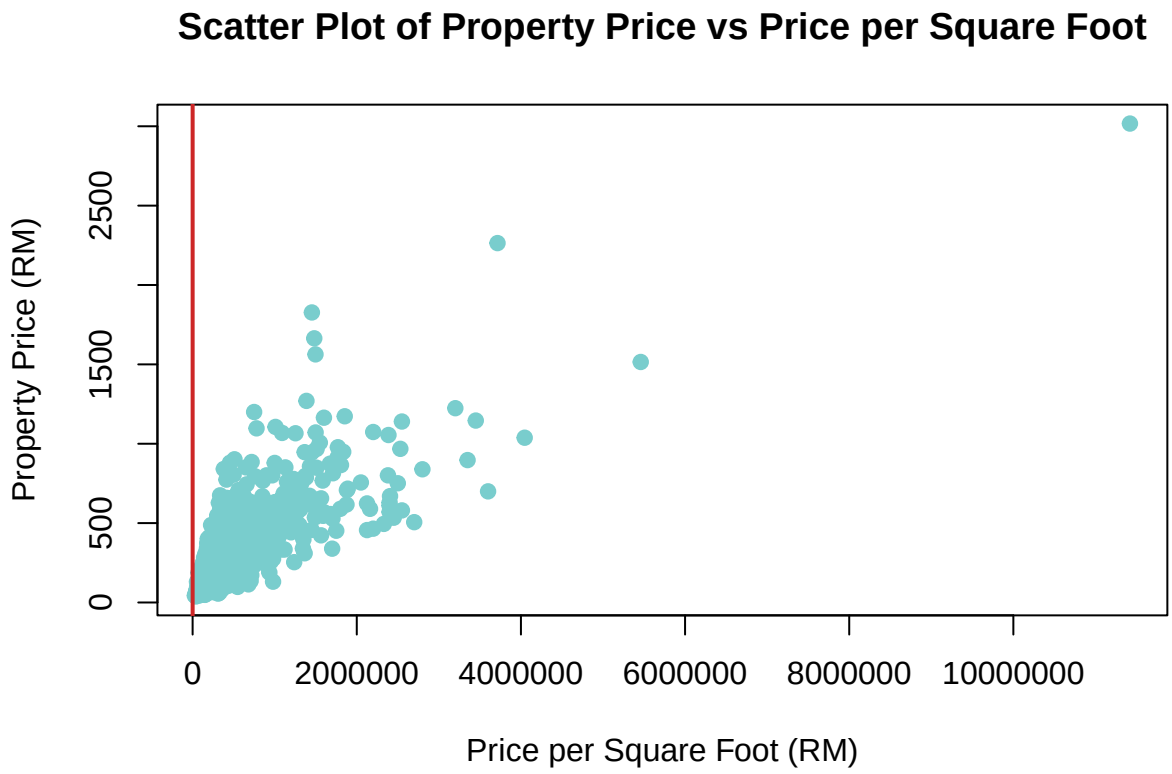
The Shapiro-Wilk test is the most practical and reliable of the three techniques used to determine normality in this situation. The Shapiro-Wilk test delivers a formal statistical test with a clear decision rule based on the p-value, while the empirical rule and QQ-plot offer useful visual and descriptive information. Using a statistical test guarantees a more accurate and objective conclusion on the distribution of property prices because of the huge sample size, skewness, and outliers.

## Part C: Bivariate Relationship Analysis

### Variable Selection

```
plot(medianPrice, medianPSF,
     main = "Scatter Plot of Property Price vs Price per Square Foot",
     xlab = "Price per Square Foot (RM)",
     ylab = "Property Price (RM)",
     col = "darkslategray3", pch = 19)

abline(lm(medianPrice ~ medianPSF), col = "firebrick3", lwd = 2)
```




The scatter plot displays the relationship between property price (RM) and price per square foot (RM). Most properties cluster in the lower left corner, suggesting that the majority of the dataset consists of properties with lower per-square-foot costs and lower overall prices.

The data shows a generally positive correlation, with property prices tending to increase as price per square foot increases. Several outliers are visible, particularly in the upper right region of the plot, where a few properties show extremely high per-square-foot costs (approaching RM11,000,000) relative to their total prices. The wide dispersion of points indicates considerable variability in this relationship.

```
pearsonCorr <- cor(medianPrice, medianPSF, use = "complete.obs")
pearsonCorr
```

```
## [1] 0.7466473
```

The Pearson correlation coefficient between **Median_Price** and **Median_PSF** is **0.7466**, indicating a moderate to strong positive linear relationship between the two variables. This suggests that as the price per square foot (Median_PSF) increases, the total property price (Median_Price) tends to increase as well, though the relationship is not perfectly linear. The correlation value reflects a reasonably strong association, meaning that changes in the price per square foot are often accompanied by corresponding changes in the total price of the property.

```
lmModel <- lm(medianPrice ~ medianPSF)
summary(lmModel)
```

```
##
## Call:
## lm(formula = medianPrice ~ medianPSF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1749653  -134578     2557    81428  6063409
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -104662.95   13763.38  -7.604   0.0000000000000437 ***
## medianPSF      1810.33      36.08  50.170 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 311800 on 1998 degrees of freedom
## Multiple R-squared:  0.5575, Adjusted R-squared:  0.5573
## F-statistic:  2517 on 1 and 1998 DF,  p-value: < 0.00000000000000022
```

The linear regression analysis suggests that there is a strong and statistically significant positive relationship between **price per square foot (RM)** and **total property price (RM)**. For each unit increase in the price per square foot, the total property price increases by **RM 1,810.33**. The model explains **55.75%** of the variability in **total property price (RM)**, indicating a decent fit, though other factors not included in the model may also influence property prices. The low p-value ($< \mathbf{0.0001}$) for both the intercept and slope suggests that these results are highly significant.

### Conclusion

There is a **moderate to strong positive relationship** between **price per square foot (RM)** and **total property price (RM)**, as indicated by the **Pearson correlation coefficient** of **0.75**. This suggests that properties with higher price per square foot tend to have higher overall prices. The **simple linear regression model** further supports this relationship, showing that for every **RM1 increase in price per square foot**, the **total property price increases by approximately RM1,810**. The model explains about **56% of the variability** in total price (**R-squared = 0.5575**), indicating a reasonably good fit. Given the **statistical significance** of both the slope and intercept (**p-values $<$ 0.001**), the linear model appears to be **appropriate** for describing this association, although other unaccounted factors may still influence property prices.

## Part D: Application & Critical Reflection

This study explored property prices in Malaysia using real market data. We analyzed two key metrics: the total price of a property and the price per square foot. On average, properties cost around RM490,000, and each square foot costs about RM329. However, there's a wide range in prices — from affordable homes to high-end properties — and some extreme values drive this variation. We found that property prices increase as price per square foot increases, meaning that more expensive land or locations tend to lead to higher overall property costs. Through charts and statistics, we also discovered that the data is not evenly distributed — most properties are priced lower, but a few outliers are extremely high-priced. This uneven spread affects how we interpret averages and trends. The visual tools helped make these patterns easier to understand at a glance.

A key limitation of descriptive statistics is that it only summarizes the current dataset without letting us make predictions or generalizations. For instance, while we know the average property price in this dataset, we cannot confidently say this applies to the entire country. Inferential statistics, like confidence intervals or hypothesis testing, would allow us to estimate national property trends or test if price differences between regions are statistically significant.

Visualizations like boxplots, histograms, and scatter plots reveal insights that raw numbers can't. For example, a boxplot shows outliers and skewness clearly, helping identify abnormal pricing. Histograms show how property prices are concentrated, and scatter plots reveal relationships between variables. These visuals make complex data more accessible and intuitive, especially for non-experts.

## Reference

[1] Kaggle, "House Prices in Malaysia 2025," available at: https://www.kaggle.com/datasets/lyhatt/house-prices-in-malaysia-2025.