A·P·U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

GROUP ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PDFA

PROGRAMMING FOR DATA ANALYSIS

HAND OUT DATE: 18 DECEMBER 2023

HAND IN DATE: 20 FEBRUARY 2023

WEIGHTAGE: 50%

INTAKE: APD2F2311CS & APU2F2311CS

COURSEWORK TITLE: URBAN RESIDENTIAL PROPERTY PRICE TRENDS

STUDENT NAME	TP NUMBER
ABDUL MUHAIMIN AMAN	TP069510
AMEERA FAROOQ	TP071807
AISHATH ISHRAN SHAIFU	TP070329

Table of Contents

1.0 INTRODUCTION

In this group assignment, our team comprised of three real estate analysts will dive into a dataset comprising information on residential properties in Kuala Lumpur. Our goal is to analyse citywide property price trends and provide actionable insights and recommendations to real estate stakeholders. The dataset includes information like location, pricing, rooms, and parking.

We use advanced data analytics techniques such as exploration, manipulation, transformation, and visualisation. We prioritise proper programming methods, error-free R programme execution, and validation of user entries. The dataset will be cleaned and pre-processed entirely using R scripting, with no dependency on external tools.

The study will include a complete classification, examination, and explanation of methodologies, backed up by clear documentation, visualisations, and applicable R programming ideas. The goal of this project is to rebuild the dataset into meaningful representations, allowing for more informed decision-making in Kuala Lumpur's ever changing real estate sector.

1.1 Data Description

1.1.0 Location

This column categorises residences in Kuala Lumpur by region or neighbourhood. It gives critical geographical data, allowing us to examine changes in property values across regions. Entries include KLCC, Mont Kiara, Bukit Bintang, and Bukit Jalil etc. Understanding this column is critical for determining the spatial distribution of attributes and generating conclusions about location-specific trends.

1.1.1 Price (RM)

This column shows the monetary worth of residential homes in Kuala Lumpur. It gives critical insights into property expenses, allowing for the monitoring of pricing patterns. Entries may contain values such as MYR 500,000 or MYR 2,000,000. Understanding this column is critical for assessing affordability and variances in property prices across categories in Kuala Lumpur.

1.1.2 Rooms

The interior layout of residential dwellings is described in this column, along with numerical counts, alphanumeric codes, and layouts such as "3+1." It offers information on the spatial arrangement, including the quantity and kinds of rooms in each home. As an illustration, consider "2 Bedrooms," "Studio," and "4+1." The column provides important information on the makeup of residential spaces in the dataset and is essential for analysing property sizes and layouts.

1.1.3 Bathrooms

This column provides a simple numerical representation (e.g., 1, 2, 3), quantifying the number of bathrooms in each property. This data is essential for evaluating the utilities and amenities of a property and makes a substantial contribution to the dataset's study of residential spaces.

1.1.4 Car Parks

The "Car Parks" column counts the number of parking spots that are available for every residential property; entries in this column are expressed as numbers, e.g., 1, 2, or 3. This data is essential for evaluating the facilities on the property, especially the ease of use for car owners. The column contributes significantly to the overall examination of residential spaces by offering perceptions into the usefulness and desirability of every house in relation to its parking arrangements.

1.1.5 Property Type

The "Property Type" column categorizes residential properties based on their architectural or functional classifications, such as serviced residence, bungalows, condominiums, etc. This categorical information is crucial for understanding the diversity and distribution of property types within the dataset, playing a central role in analysing trends and assisting stakeholders in making informed decisions based on specific property classifications.

1.1.6 Type

This column of categories aids in differentiating between various metrics, including the total land area or the built-up space on the property. Values such as "Built-Up" or "Land" are examples of entries in this column that provide context for the numerical data that goes with it. For stakeholders who are interested in certain property dimensions, comprehending this column is crucial as it aids in the accurate analysis of the dataset's land or built-up regions.

1.1.7 Size

This column gives specific numbers on the area or physical dimensions of every residential property, revealing details about its spatial capacity. A numerical number reflecting the size of the corresponding unit is represented by each entry; examples of values include square footage measures. Gaining an understanding of this column is essential for stakeholders evaluating the capacity and physical area of properties. It also plays a major role in the study of patterns pertaining to property sizes in the dataset and supports well-informed decision-making about spatial dimensions.

1.1.8 Furnishing

This dataset's column classifies residential homes according on how furnished they are. The furnished, unfurnished, and partially furnished status of a property is indicated in this category column. The following are some sample values for this column: "Furnished," "Unfurnished," or "Partially Furnished." For stakeholders interested in the amenities of the property, understanding this column is essential as it facilitates the analysis of patterns pertaining to various furnishing levels within the dataset.

1.2 Initial Assumptions

Urban property prices in Kuala Lumpur are expected to show a general increasing trend, given the city's diversified and vibrant real estate sector. Urbanisation, economic expansion, and rising demand for residential real estate are all possible drivers of future property value rise. Differential pricing patterns depending on location and property attributes may also result from Kuala Lumpur's many neighbourhoods, amenities, and property types. Nonetheless, fluctuations in the market, outside economic variables, and governmental regulations may have an impact on the direction of real estate values. To verify these hypotheses and identify particular trends and patterns in Kuala Lumpur's urban real estate market, a more thorough examination of the dataset would be required.

1.3 Hypothesis

The properties in between 1000 to 10000 square feet that are located in KLCC cost 20% more than the properties in other regions that have the same area and have the same furnishing status.

1.4 Objectives

The primary objective of this project is to identify significant changes in Kuala Lumpur's residential property prices by thoroughly examining a dataset that contains such information. Our main objective as real estate analysts is to use up-to-date analytics techniques to thoroughly examine and analyse the information, accounting for important aspects like location, size, type, furnishing, and more. By going above and beyond the fundamentals of the course, we hope to provide real estate industry stakeholders insightful analysis and tactical suggestions. The task highlights the need of doing comprehensive data analysis, manipulation, transformation, and visualisation. Its goal is to rebuild the dataset into insightful representations that enable decision-makers to navigate the complexities of Kuala Lumpur's dynamic real estate market.

To prove that our hypothesis is accurate, we have divided the objective into 3 categories.

Student 1- Abdul Muhaimin Aman, TP069510
Objective: Impact of location on price.

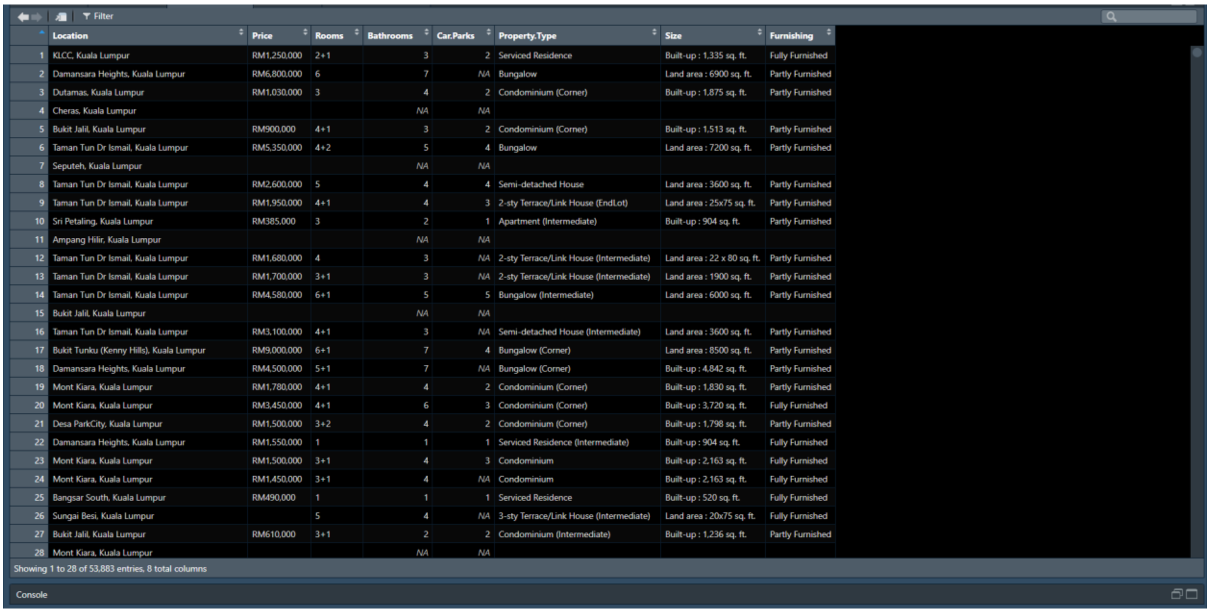
Student 2- Ameera Farooq, TP071807
Objective: Impact of sq. feet on price.

Student 3- Ishran Aishath Shaifu, TP070329
Objective: Impact of furnishing status on price.

2.0 DATA PREPARATION

Given the dataset's potential to provide rich insights into the real estate market in Kuala Lumpur, it is crucial to address the inconsistencies and prepare the data for analytical tasks. The dataset originally contained mixed-format data, missing values, and duplicates. This meant that a thorough cleaning and standardization had to be conducted to ensure that the data is meaningful. As such, this section will expand upon how the dataset was cleaned.

2.1 Data Importation



	Location	Price	Rooms	Bathrooms	Car.Parks	Property.Type	Size	Furnishing
1	KLCC, Kuala Lumpur	RM1,250,000	2+1	3	2	Serviced Residence	Built-up : 1,335 sq. ft.	Fully Furnished
2	Damansara Heights, Kuala Lumpur	RM6,800,000	6	7	NA	Bungalow	Land area : 6900 sq. ft.	Partly Furnished
3	Dutamas, Kuala Lumpur	RM1,030,000	3	4	2	Condominium (Corner)	Built-up : 1,875 sq. ft.	Partly Furnished
4	Cheras, Kuala Lumpur			NA	NA			
5	Bukit Jalil, Kuala Lumpur	RM900,000	4+1	3	2	Condominium (Corner)	Built-up : 1,513 sq. ft.	Partly Furnished
6	Taman Tun Dr Ismail, Kuala Lumpur	RM5,350,000	4+2	5	4	Bungalow	Land area : 7200 sq. ft.	Partly Furnished
7	Seputeh, Kuala Lumpur			NA	NA			
8	Taman Tun Dr Ismail, Kuala Lumpur	RM2,600,000	5	4	4	Semi-detached House	Land area : 3600 sq. ft.	Partly Furnished
9	Taman Tun Dr Ismail, Kuala Lumpur	RM1,950,000	4+1	4	3	2-sty Terrace/Link House (EndLot)	Land area : 25x75 sq. ft.	Partly Furnished
10	Sri Petaling, Kuala Lumpur	RM385,000	3	2	1	Apartment (Intermediate)	Built-up : 904 sq. ft.	Partly Furnished
11	Ampang Hilir, Kuala Lumpur			NA	NA			
12	Taman Tun Dr Ismail, Kuala Lumpur	RM1,680,000	4	3	NA	2-sty Terrace/Link House (Intermediate)	Land area : 22 x 80 sq. ft.	Partly Furnished
13	Taman Tun Dr Ismail, Kuala Lumpur	RM1,700,000	3+1	3	NA	2-sty Terrace/Link House (Intermediate)	Land area : 1900 sq. ft.	Partly Furnished
14	Taman Tun Dr Ismail, Kuala Lumpur	RM4,580,000	6+1	5	5	Bungalow (Intermediate)	Land area : 6000 sq. ft.	Partly Furnished
15	Bukit Jalil, Kuala Lumpur			NA	NA			
16	Taman Tun Dr Ismail, Kuala Lumpur	RM3,100,000	4+1	3	NA	Semi-detached House (Intermediate)	Land area : 3600 sq. ft.	Partly Furnished
17	Bukit Tunku (Kenny Hills), Kuala Lumpur	RM9,000,000	6+1	7	4	Bungalow (Corner)	Land area : 8500 sq. ft.	Partly Furnished
18	Damansara Heights, Kuala Lumpur	RM4,500,000	5+1	7	NA	Bungalow (Corner)	Built-up : 4,842 sq. ft.	Partly Furnished
19	Mont Kiara, Kuala Lumpur	RM1,780,000	4+1	4	2	Condominium (Corner)	Built-up : 1,830 sq. ft.	Partly Furnished
20	Mont Kiara, Kuala Lumpur	RM3,450,000	4+1	6	3	Condominium (Corner)	Built-up : 3,720 sq. ft.	Fully Furnished
21	Desa ParkCity, Kuala Lumpur	RM1,500,000	3+2	4	2	Condominium (Corner)	Built-up : 1,798 sq. ft.	Partly Furnished
22	Damansara Heights, Kuala Lumpur	RM1,550,000	1	1	1	Serviced Residence (Intermediate)	Built-up : 904 sq. ft.	Fully Furnished
23	Mont Kiara, Kuala Lumpur	RM1,500,000	3+1	4	3	Condominium	Built-up : 2,163 sq. ft.	Fully Furnished
24	Mont Kiara, Kuala Lumpur	RM1,450,000	3+1	4	NA	Condominium	Built-up : 2,163 sq. ft.	Fully Furnished
25	Bangsar South, Kuala Lumpur	RM490,000	1	1	1	Serviced Residence	Built-up : 520 sq. ft.	Fully Furnished
26	Sungai Besi, Kuala Lumpur		5	4	NA	3-sty Terrace/Link House (Intermediate)	Land area : 20x75 sq. ft.	Fully Furnished
27	Bukit Jalil, Kuala Lumpur	RM610,000	3+1	2	2	Condominium (Intermediate)	Built-up : 1,236 sq. ft.	Partly Furnished
28	Mont Kiara, Kuala Lumpur			NA	NA			

Showing 1 to 28 of 53,883 entries, 8 total columns

Figure 1.0: The Original Dataset kl_property_data

The above picture shows the original dataset, named kl_property_data, loaded into the R environment. It boasts of over 50,000 results within the Kuala Lumpur region. However, as one can see, the dataset is riddled with numerous inconsistencies. For the purpose of data cleaning, the kl_property_data was loaded into a dataframe named **filter_df.**

2.2 Data Cleaning

Several key packages from the R ecosystem were utilized.

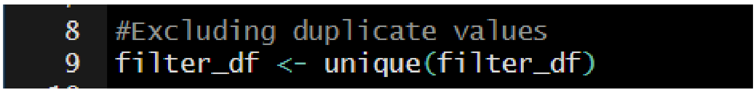
dplyr: Filtering, choosing, and altering the dataset effectively was made possible by dplyr's flexible data manipulation features. It is perfect for handling big datasets because of its performance optimizations and syntax.

stringr: The dataset's text data was managed and cleaned using this package. stringr offered reliable utilities for extracting and modifying textual data because mixed-format strings were included in columns like Price and Size column.

tidyr: Specifically used to divide the Size column into two separate columns, tidyr was included to the toolbox for its ability to tidy data. The readability and organization of the dataset was improved by dividing this column according to a designated delimiter using the separate() method. This will be explained in further detail within this section.

There are 5 main parts to the cleaning process.

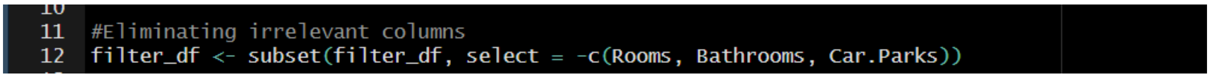
1. Removing Duplicates: Duplicate entries were removed as it can skew the analysis results if left alone.



```
#Excluding duplicate values
filter_df <- unique(filter_df)
```

Figure 1.1 Removing Duplicates

2. Dropping Irrelevant Columns: The columns ‘Rooms’, ‘Bathrooms’, and ‘Car.Parks’ were removed. The property price, location, type, size and furnishing status columns were identified to be the key factors influencing the property value by our team.



```
#Eliminating irrelevant columns
filter_df <- subset(filter_df, select = -c(Rooms, Bathrooms, Car.Parks))
```

Figure 1.2 Dropping Columns

3. Renaming Columns and Converting Data Types: To help facilitate analysis and enhance readability, columns names were standardized, and the data types were converted accordingly. Take note that the Size column for the kl_property_data was split into two columns.

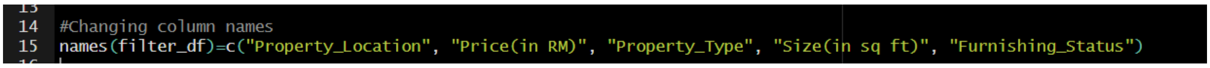
   Location -> Property_Location

Price -> Price(in RM)

Property.Type -> Property_Type

Size -> Type and Size(in sq ft)

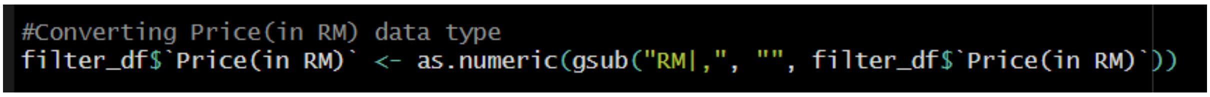
Furnishing -> Furnishing_Status



```
#Changing column names
names(filter_df)=c("Property_Location", "Price(in RM)", "Property_Type", "Size(in sq ft)", "Furnishing_Status")
```

Figure 1.3 Renaming Columns

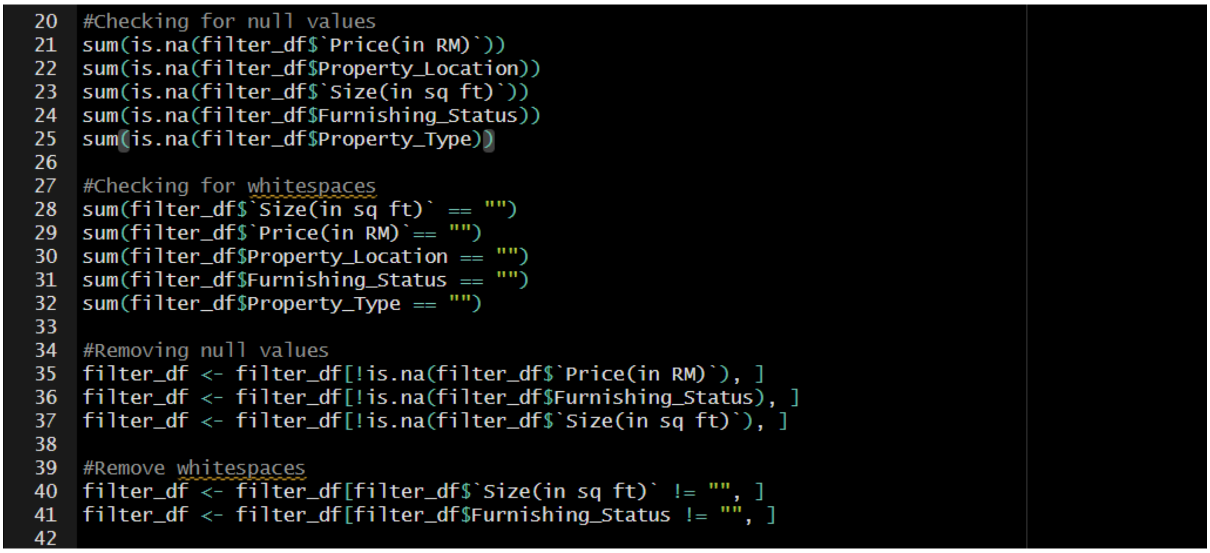
The columns Price(in RM) and Size(in sq ft) were converted to numeric.



```
#Converting Price(in RM) data type
filter_df$`Price(in RM)` <- as.numeric(gsub("RM|,", "", filter_df$`Price(in RM)`))
```

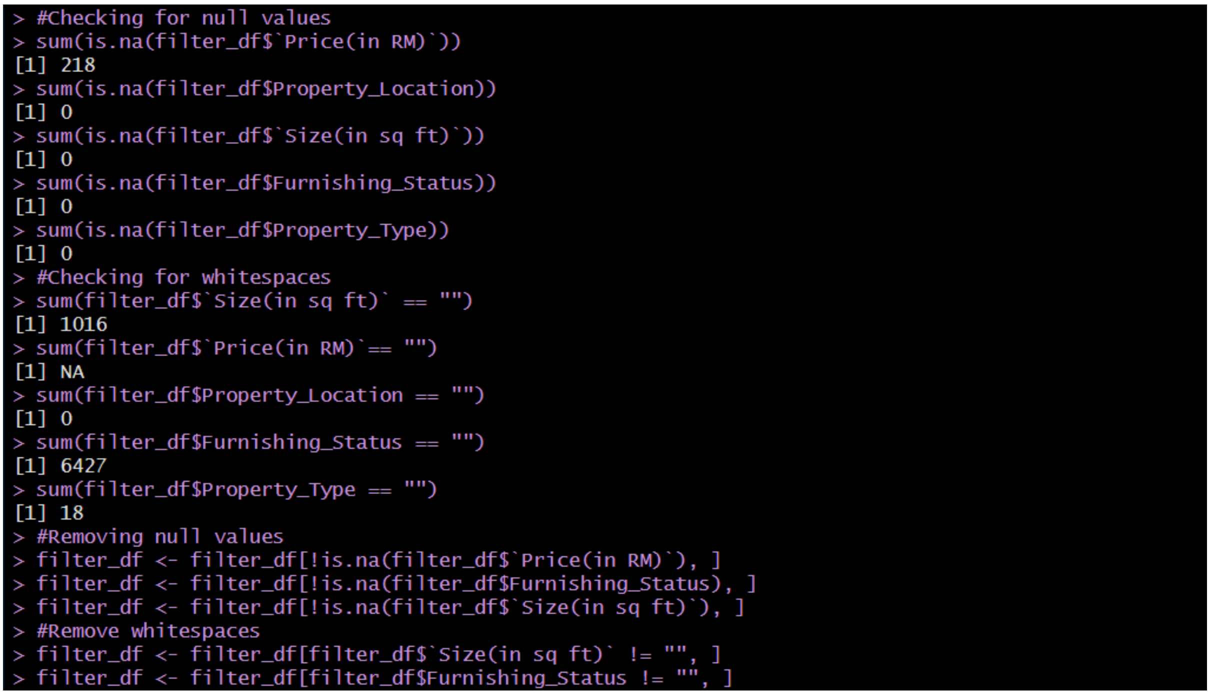
Figure 1.4 Changing datatype of Price(in RM)

4. Handling Missing and Fixing Improper Data Formats: Missing values and empty strings in the columns were identified and removed. Special attention was given to the Size(in sq ft) column, where a custom function was made to convert various size measurements into square feet. The process for the Size column will be clarified later on in the report. Some values in Furnishing_Status was marked with unknown were also removed.



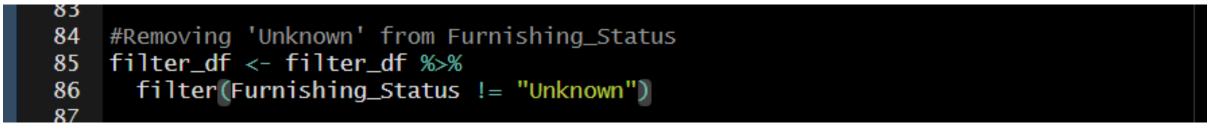
```
#Checking for null values
sum(is.na(filter_df$`Price(in RM)`))
sum(is.na(filter_df$Property_Location))
sum(is.na(filter_df$`Size(in sq ft)`))
sum(is.na(filter_df$Furnishing_Status))
sum(is.na(filter_df$Property_Type))

#Checking for whitespaces
sum(filter_df$`Size(in sq ft)` == "")
sum(filter_df$`Price(in RM)`== "")
sum(filter_df$Property_Location == "")
sum(filter_df$Furnishing_Status == "")
sum(filter_df$Property_Type == "")

#Removing null values
filter_df <- filter_df[!is.na(filter_df$`Price(in RM)`), ]
filter_df <- filter_df[!is.na(filter_df$Furnishing_Status), ]
filter_df <- filter_df[!is.na(filter_df$`Size(in sq ft)`), ]

#Remove whitespaces
filter_df <- filter_df[filter_df$`Size(in sq ft)` != "", ]
filter_df <- filter_df[filter_df$Furnishing_Status != "", ]

```

Figure 1.5: Handling Missing Data and Improper Data Formats



```
> #Checking for null values
> sum(is.na(filter_df$`Price(in RM)`))
[1] 218
> sum(is.na(filter_df$Property_Location))
[1] 0
> sum(is.na(filter_df$`Size(in sq ft)`))
[1] 0
> sum(is.na(filter_df$Furnishing_Status))
[1] 0
> sum(is.na(filter_df$Property_Type))
[1] 0
> #Checking for whitespaces
> sum(filter_df$`Size(in sq ft)` == "")
[1] 1016
> sum(filter_df$`Price(in RM)`== "")
[1] NA
> sum(filter_df$Property_Location == "")
[1] 0
> sum(filter_df$Furnishing_Status == "")
[1] 6427
> sum(filter_df$Property_Type == "")
[1] 18
> #Removing null values
> filter_df <- filter_df[!is.na(filter_df$`Price(in RM)`), ]
> filter_df <- filter_df[!is.na(filter_df$Furnishing_Status), ]
> filter_df <- filter_df[!is.na(filter_df$`Size(in sq ft)`), ]
> #Remove whitespaces
> filter_df <- filter_df[filter_df$`Size(in sq ft)` != "", ]
> filter_df <- filter_df[filter_df$Furnishing_Status != "", ]
```

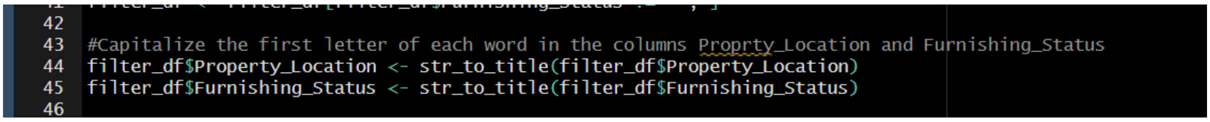
Figure 1.6: Output on Console for Code Snippet in Figure 6



```
#Removing 'Unknown' from Furnishing_Status
filter_df <- filter_df %>%
  filter(Furnishing_Status != "Unknown")
```

Figure 1.7: Removing Unkown from Furnishing_Status

5. <u>Standardizing Text</u>: The data in Property_Location and Furnishing status columns were capitalized to maintain consistency.

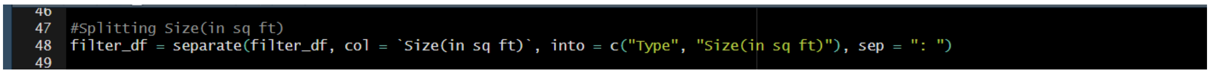


```
#Capitalize the first letter of each word in the columns Proprty_Location and Furnishing_Status
filter_df$Property_Location <- str_to_title(filter_df$Property_Location)
filter_df$Furnishing_Status <- str_to_title(filter_df$Furnishing_Status)
```

Figure 1.8: Standardizing Text

Challenged in the Size (in sq feet) column.

Initially the Size (in sq ft) column is divided into two different columns, Type and Size (in sq ft), using the separate() function from the “tidyr” package. The basis for this operation is a defined separator (": "), indicating that the original data has included a type designation or other information that comes before an actual size value and is separated by a colon and a space. Here, the objective is to separate the size-related numerical data and classify them based on the original type designation.

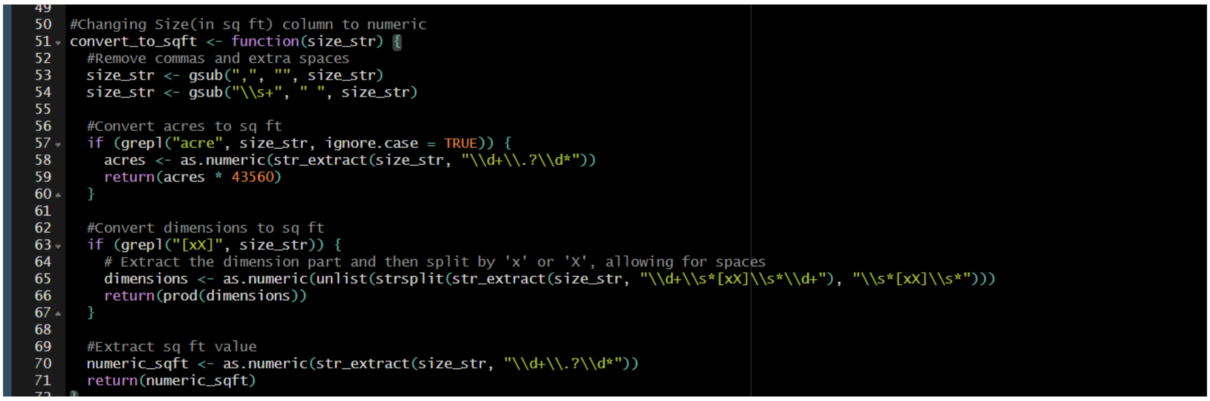


```r
#Splitting Size(in sq ft)
filter_df = separate(filter_df, col = `Size(in sq ft)`, into = c("Type", "Size(in sq ft)"), sep = ": ")
```

Figure 1.9: Separating the Size Column

To accommodate different size information formats and convert them all into a standard numeric format that represents square footage, a custom function called convert_to_sqft is defined.

- **Eliminating Commas and Extra Spaces**: The function begins by removing any commas and extra spaces from the input string. These characters are frequently found in formatted numbers but can cause issues when converting numbers.
- **Converting Acres to Square Feet**: The function uses the conversion factor (1 acre = 43,560 square feet) to convert the size information to square feet if it is provided in acres (which is indicated by the presence of "acre" in the string).
- **Converting Dimensions to Square Feet**: Rather than giving an exact area measurement, some data only include dimensions (such as "20x40"). In order to get the area in square feet, the function first finds these patterns, then extracts the numerical dimensions and multiplies them.
- **Direct Numerical Conversion**: The function extracts the numeric value and returns it in simple circumstances where the size is already given in square feet.



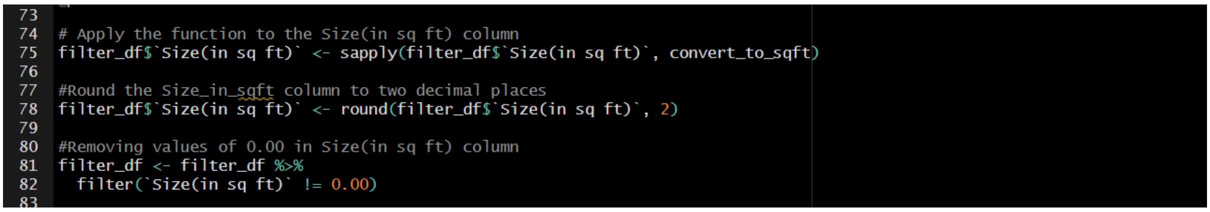
```r
#Changing Size(in sq ft) column to numeric
convert_to_sqft <- function(size_str) {
  #Remove commas and extra spaces
  size_str <- gsub(",", "", size_str)
  size_str <- gsub("\\s+", " ", size_str)

  #Convert acres to sq ft
  if (grepl("acre", size_str, ignore.case = TRUE)) {
    acres <- as.numeric(str_extract(size_str, "\\d+\\.?\\d*"))
    return(acres * 43560)
  }

  #Convert dimensions to sq ft
  if (grepl("[xX]", size_str)) {
    # Extract the dimension part and then split by 'x' or 'X', allowing for spaces
    dimensions <- as.numeric(unlist(strsplit(str_extract(size_str, "\\d+\\s*[xX]\\s*\\d+"), "\\s*[xX]\\s*")))
    return(prod(dimensions))
  }

  #Extract sq ft value
  numeric_sqft <- as.numeric(str_extract(size_str, "\\d+\\.?\\d*"))
  return(numeric_sqft)
```

Figure 1.10: Code Snippet of convert_to_sqft Function

Every value in the Size (in sq ft) column is subjected to the convert_to_sqft function, which transforms all size data into a standard numerical format.

After that, the values are rounded to two decimal places to enhance readability and ensure consistency in precision throughout the dataset.

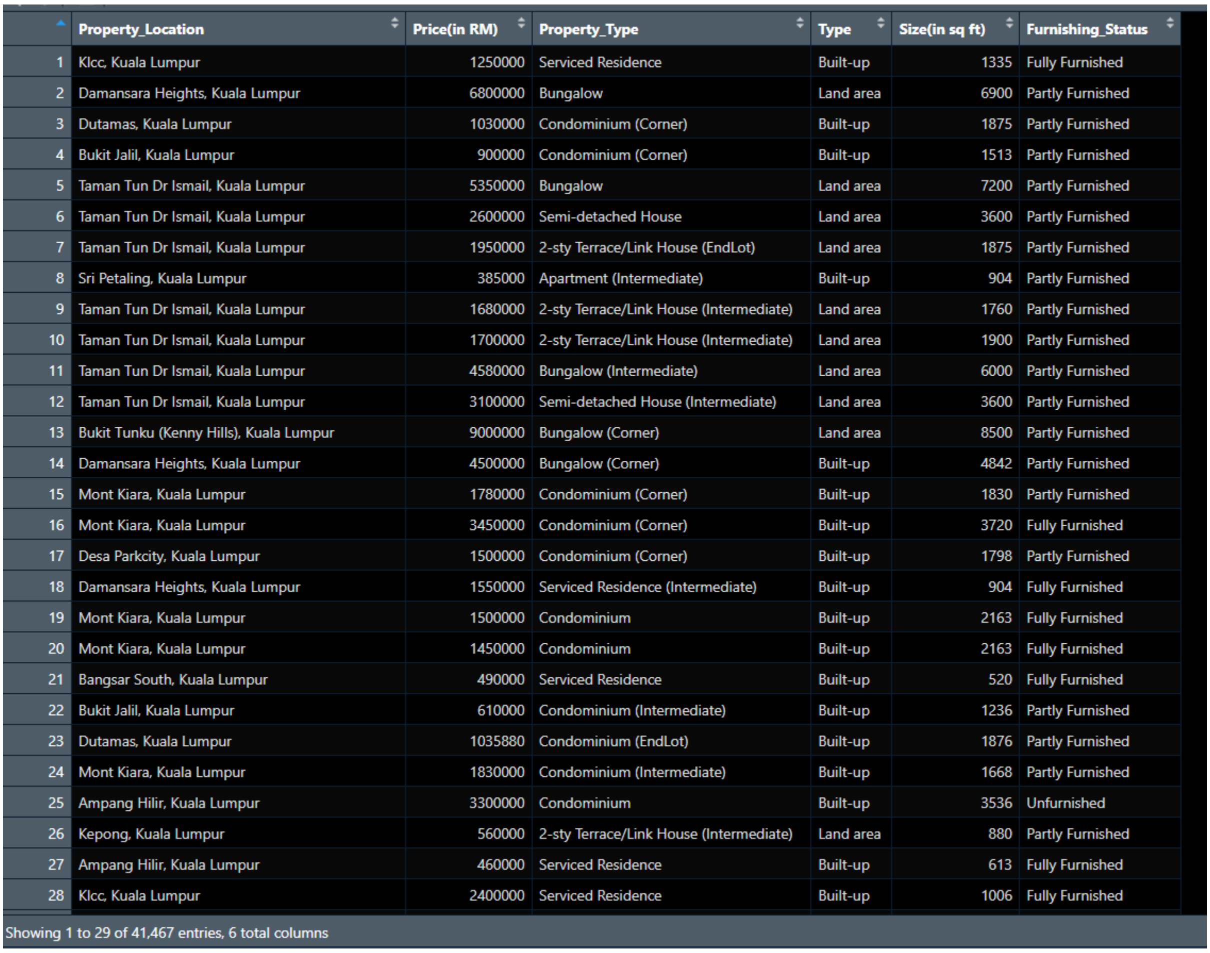
Lastly, postings that are less than 0.00 square feet are eliminated. In order to ensure the quality and dependability of the dataset, this stage attempts to remove erroneous data items that did not offer correct size information.



```
# Apply the function to the Size(in sq ft) column
filter_df$`Size(in sq ft)` <- sapply(filter_df$`Size(in sq ft)`, convert_to_sqft)

#Round the Size_in_sqft column to two decimal places
filter_df$`Size(in sq ft)` <- round(filter_df$`Size(in sq ft)`, 2)

#Removing values of 0.00 in Size(in sq ft) column
filter_df <- filter_df %>%
  filter(`Size(in sq ft)` != 0.00)
```

Figure 1.11: Code Snippet of Applying Function and Setting Data Type of Column

Together with R's basic features, these packages provide a complete toolkit for handling the different data preparation and cleaning issues this project presented. These tools were selected because they are widely used in the data science community, have a wealth of documentation, and can be used for the kinds of data manipulation tasks needed to meet the project's goals.



	Property_Location	Price(in RM)	Property_Type	Type	Size(in sq ft)	Furnishing_Status
1	Klcc, Kuala Lumpur	1250000	Serviced Residence	Built-up	1335	Fully Furnished
2	Damansara Heights, Kuala Lumpur	6800000	Bungalow	Land area	6900	Partly Furnished
3	Dutamas, Kuala Lumpur	1030000	Condominium (Corner)	Built-up	1875	Partly Furnished
4	Bukit Jalil, Kuala Lumpur	900000	Condominium (Corner)	Built-up	1513	Partly Furnished
5	Taman Tun Dr Ismail, Kuala Lumpur	5350000	Bungalow	Land area	7200	Partly Furnished
6	Taman Tun Dr Ismail, Kuala Lumpur	2600000	Semi-detached House	Land area	3600	Partly Furnished
7	Taman Tun Dr Ismail, Kuala Lumpur	1950000	2-sty Terrace/Link House (EndLot)	Land area	1875	Partly Furnished
8	Sri Petaling, Kuala Lumpur	385000	Apartment (Intermediate)	Built-up	904	Partly Furnished
9	Taman Tun Dr Ismail, Kuala Lumpur	1680000	2-sty Terrace/Link House (Intermediate)	Land area	1760	Partly Furnished
10	Taman Tun Dr Ismail, Kuala Lumpur	1700000	2-sty Terrace/Link House (Intermediate)	Land area	1900	Partly Furnished
11	Taman Tun Dr Ismail, Kuala Lumpur	4580000	Bungalow (Intermediate)	Land area	6000	Partly Furnished
12	Taman Tun Dr Ismail, Kuala Lumpur	3100000	Semi-detached House (Intermediate)	Land area	3600	Partly Furnished
13	Bukit Tunku (Kenny Hills), Kuala Lumpur	9000000	Bungalow (Corner)	Land area	8500	Partly Furnished
14	Damansara Heights, Kuala Lumpur	4500000	Bungalow (Corner)	Built-up	4842	Partly Furnished
15	Mont Kiara, Kuala Lumpur	1780000	Condominium (Corner)	Built-up	1830	Partly Furnished
16	Mont Kiara, Kuala Lumpur	3450000	Condominium (Corner)	Built-up	3720	Fully Furnished
17	Desa Parkcity, Kuala Lumpur	1500000	Condominium (Corner)	Built-up	1798	Partly Furnished
18	Damansara Heights, Kuala Lumpur	1550000	Serviced Residence (Intermediate)	Built-up	904	Fully Furnished
19	Mont Kiara, Kuala Lumpur	1500000	Condominium	Built-up	2163	Fully Furnished
20	Mont Kiara, Kuala Lumpur	1450000	Condominium	Built-up	2163	Fully Furnished
21	Bangsar South, Kuala Lumpur	490000	Serviced Residence	Built-up	520	Fully Furnished
22	Bukit Jalil, Kuala Lumpur	610000	Condominium (Intermediate)	Built-up	1236	Partly Furnished
23	Dutamas, Kuala Lumpur	1035880	Condominium (EndLot)	Built-up	1876	Partly Furnished
24	Mont Kiara, Kuala Lumpur	1830000	Condominium (Intermediate)	Built-up	1668	Partly Furnished
25	Ampang Hilir, Kuala Lumpur	3300000	Condominium	Built-up	3536	Unfurnished
26	Kepong, Kuala Lumpur	560000	2-sty Terrace/Link House (Intermediate)	Land area	880	Partly Furnished
27	Ampang Hilir, Kuala Lumpur	460000	Serviced Residence	Built-up	613	Fully Furnished
28	Klcc, Kuala Lumpur	2400000	Serviced Residence	Built-up	1006	Fully Furnished

Showing 1 to 29 of 41,467 entries, 6 total columns

Figure 1.12: The Cleaned Table

3.0 DATA ANALYSIS

3.1 OBJECTIVE 1: TO INVESTIAGE THE IMPACT OF LOCATION ON PRICE.

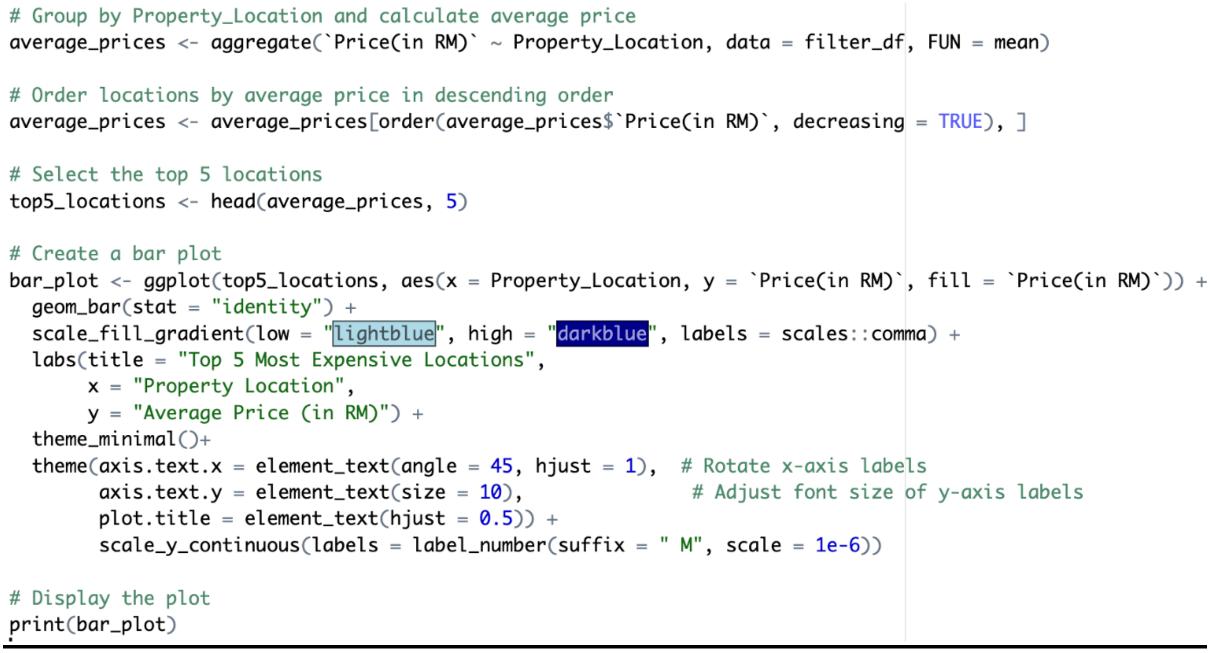
Student Name: Abdul Muhaimin Aman

TP Number: TP069510

Specialism: Data Analysis

Analysis 3.1.1: What are the top five most expensive locations on average?

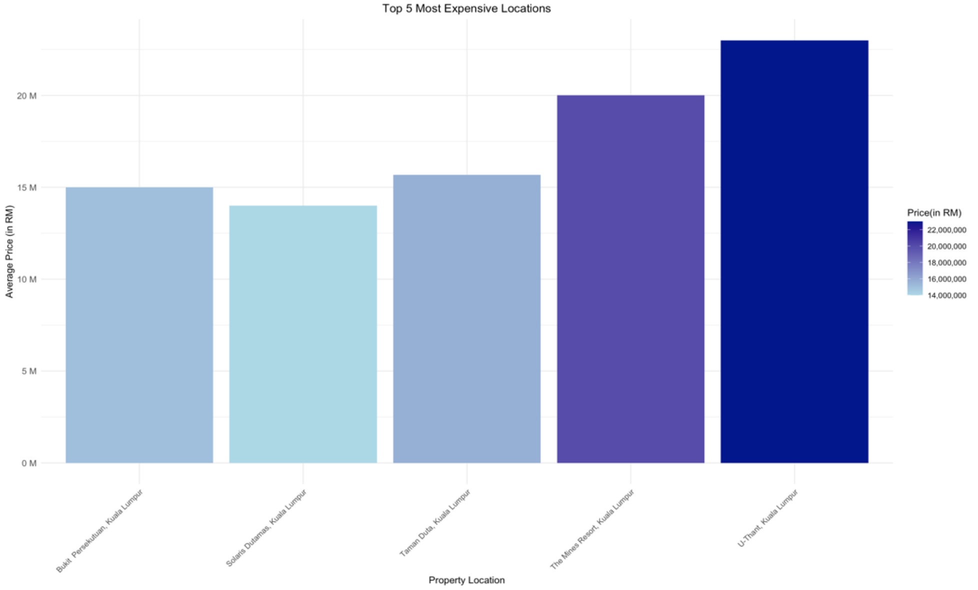
Screenshot of Code:



```
# Group by Property_Location and calculate average price
average_prices <- aggregate(`Price(in RM)` ~ Property_Location, data = filter_df, FUN = mean)

# Order locations by average price in descending order
average_prices <- average_prices[order(average_prices$`Price(in RM)`, decreasing = TRUE), ]

# Select the top 5 locations
top5_locations <- head(average_prices, 5)

# Create a bar plot
bar_plot <- ggplot(top5_locations, aes(x = Property_Location, y = `Price(in RM)`, fill = `Price(in RM)`)) +
  geom_bar(stat = "identity") +
  scale_fill_gradient(low = "lightblue", high = "darkblue", labels = scales::comma) +
  labs(title = "Top 5 Most Expensive Locations",
       x = "Property Location",
       y = "Average Price (in RM)") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels
        axis.text.y = element_text(size = 10),              # Adjust font size of y-axis labels
        plot.title = element_text(hjust = 0.5)) +
        scale_y_continuous(labels = label_number(suffix = " M", scale = 1e-6))

# Display the plot
print(bar_plot)
```

The code supplied uses the R ggplot2 package to produce a clear bar plot. The objective is to use the provided dataset ({filter_df}) to visualise and list the top 5 locations with the highest average property prices. The height of each bar in the plot denotes the average property price in that particular area, and each bar represents a distinct place. A visual depiction of the different pricing points is added by the colour gradient that goes from light blue to dark blue. This layout makes it easy to find and contrast the most expensive real estate places.

Screenshot of Output:

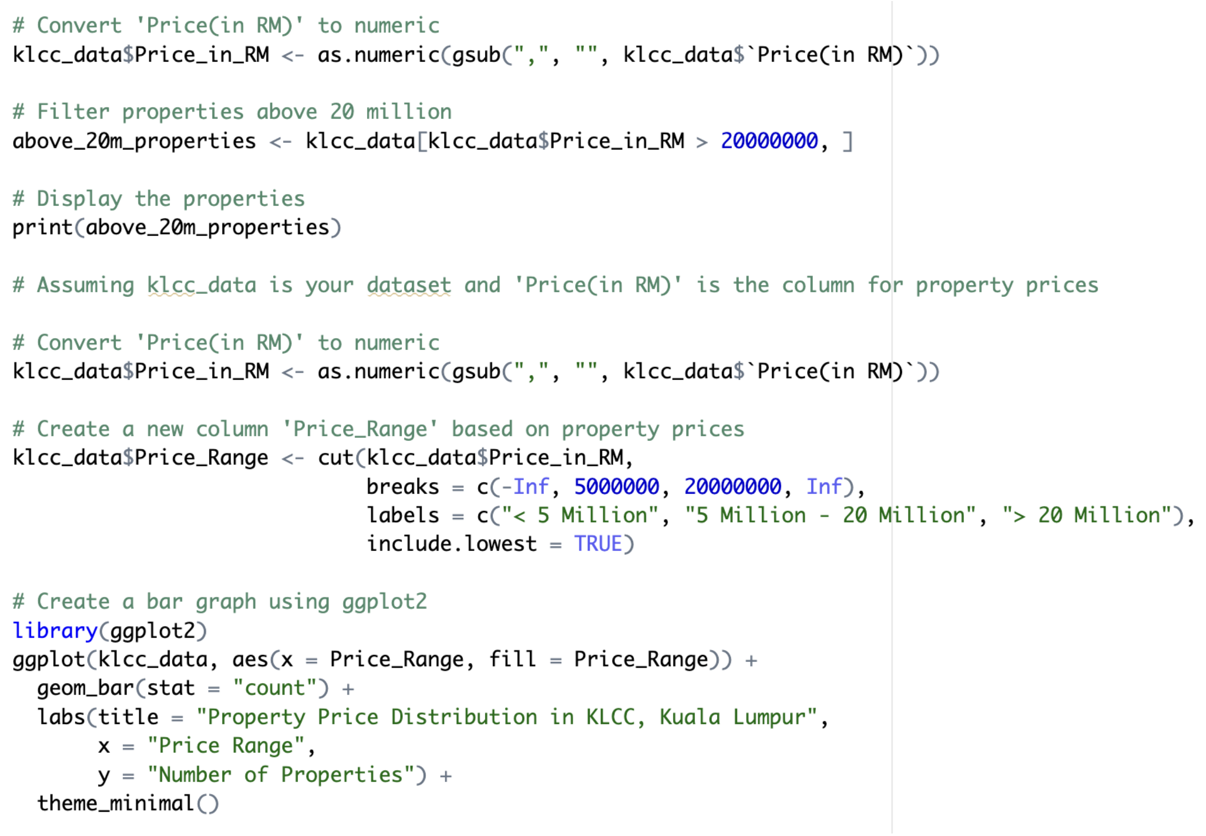


Analysis Justification:

The distribution of the dataset's property categories is briefly summarised in this bar graph, which offers insightful information about recurring patterns. This bar graph helps us to declare the fact that the area KLCC does not seem to have the highest average price for properties, at least not in the top five. Thus making the analysis useful.

Analysis 3.1.2: Property Price Distribution in KLCC, Kuala Lumpur.

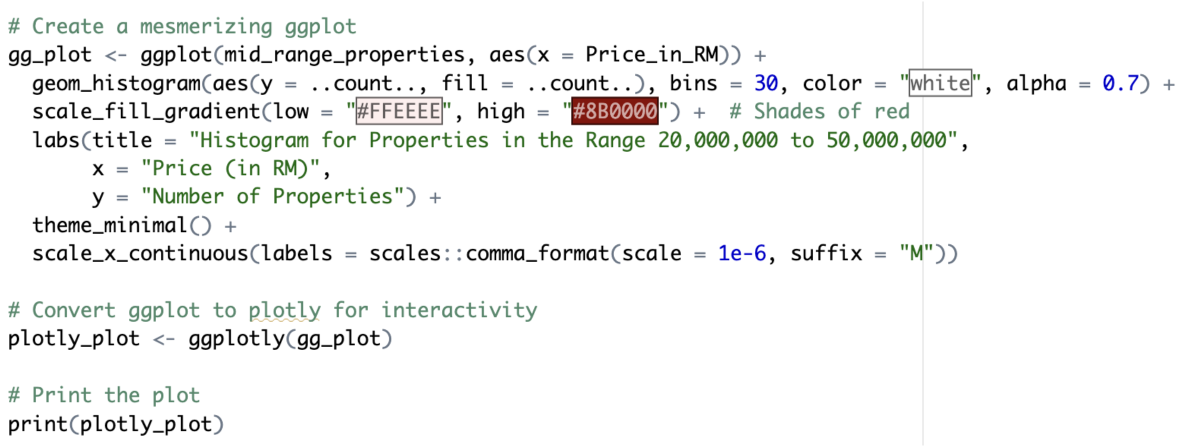
Screenshot of Code:



```
# Convert 'Price(in RM)' to numeric
klcc_data$Price_in_RM <- as.numeric(gsub(",", "", klcc_data$`Price(in RM)`))

# Filter properties above 20 million
above_20m_properties <- klcc_data[klcc_data$Price_in_RM > 20000000, ]

# Display the properties
print(above_20m_properties)

# Assuming klcc_data is your dataset and 'Price(in RM)' is the column for property prices

# Convert 'Price(in RM)' to numeric
klcc_data$Price_in_RM <- as.numeric(gsub(",", "", klcc_data$`Price(in RM)`))

# Create a new column 'Price_Range' based on property prices
klcc_data$Price_Range <- cut(klcc_data$Price_in_RM,
                             breaks = c(-Inf, 5000000, 20000000, Inf),
                             labels = c("< 5 Million", "5 Million - 20 Million", "> 20 Million"),
                             include.lowest = TRUE)

# Create a bar graph using ggplot2
library(ggplot2)
ggplot(klcc_data, aes(x = Price_Range, fill = Price_Range)) +
  geom_bar(stat = "count") +
  labs(title = "Property Price Distribution in KLCC, Kuala Lumpur",
       x = "Price Range",
       y = "Number of Properties") +
  theme_minimal()
```

First snippet.

Several activities relating to property price analysis in the dataset are carried out by the code snippet that is supplied. The ‘Price(in RM)’ field is first converted to numerical format, with any commas eliminated for uniformity. It then refines and shows homes that cost more than $20 million. In the second section of the code, ‘Price(in RM)’ is assumed to be the price column. A new categorical column ‘Price_Range’ is created based on predetermined price intervals (< 5 Million, 5 Million – 20 Million, > 20 Million). Lastly, it uses ggplot2 to create a bar graph that shows how properties are distributed across various price points in KLCC, Kuala Lumpur. To help with the comprehension of price trends, the graph shows the number of properties within each designated price range.

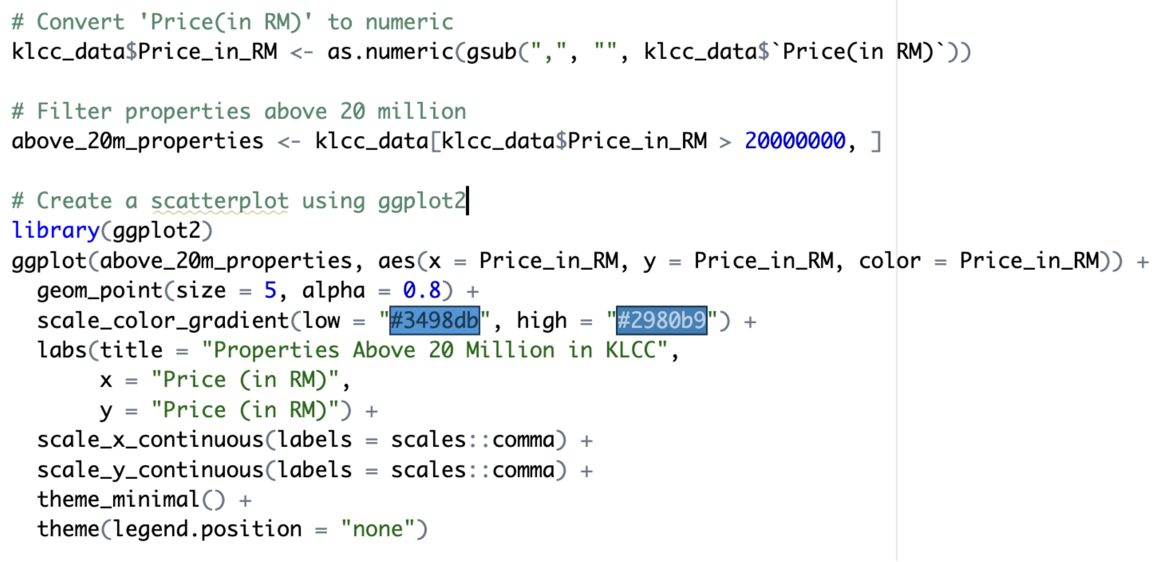


```r
# Create a mesmerizing ggplot
gg_plot <- ggplot(mid_range_properties, aes(x = Price_in_RM)) +
  geom_histogram(aes(y = ..count.., fill = ..count..), bins = 30, color = "white", alpha = 0.7) +
  scale_fill_gradient(low = "#FFEEEE", high = "#8B0000") +  # Shades of red
  labs(title = "Histogram for Properties in the Range 20,000,000 to 50,000,000",
       x = "Price (in RM)",
       y = "Number of Properties") +
  theme_minimal() +
  scale_x_continuous(labels = scales::comma_format(scale = 1e-6, suffix = "M"))

# Convert ggplot to plotly for interactivity
plotly_plot <- ggplotly(gg_plot)

# Print the plot
print(plotly_plot)
```

Second snippet.

This snippet of code aims to visualise the dataset 'klcc_data''s distribution of property prices between 20,000,000 and 50,000,000. The 'Price(in RM)' column is first converted to numerical format, then commas are removed to ensure uniformity. Filtering homes that come inside the designated price range is the next step. The x-axis shows property prices, while the y-axis shows the number of properties. The code then uses ggplot2 to produce an interesting histogram. The histogram has a visually pleasing depiction thanks to its gradient colour scheme, which goes from a light shade of red to a deep crimson. Because ggplot was converted to plotly, the final plot is interactive, enabling users to examine and evaluate the distribution of real estate values in the designated

Additional coding was done to analyse the properties between the range of 20 million RM to 50 million RM to depict that there is only one outlier which goes extremely out of range. So there are 10 properties within this range.

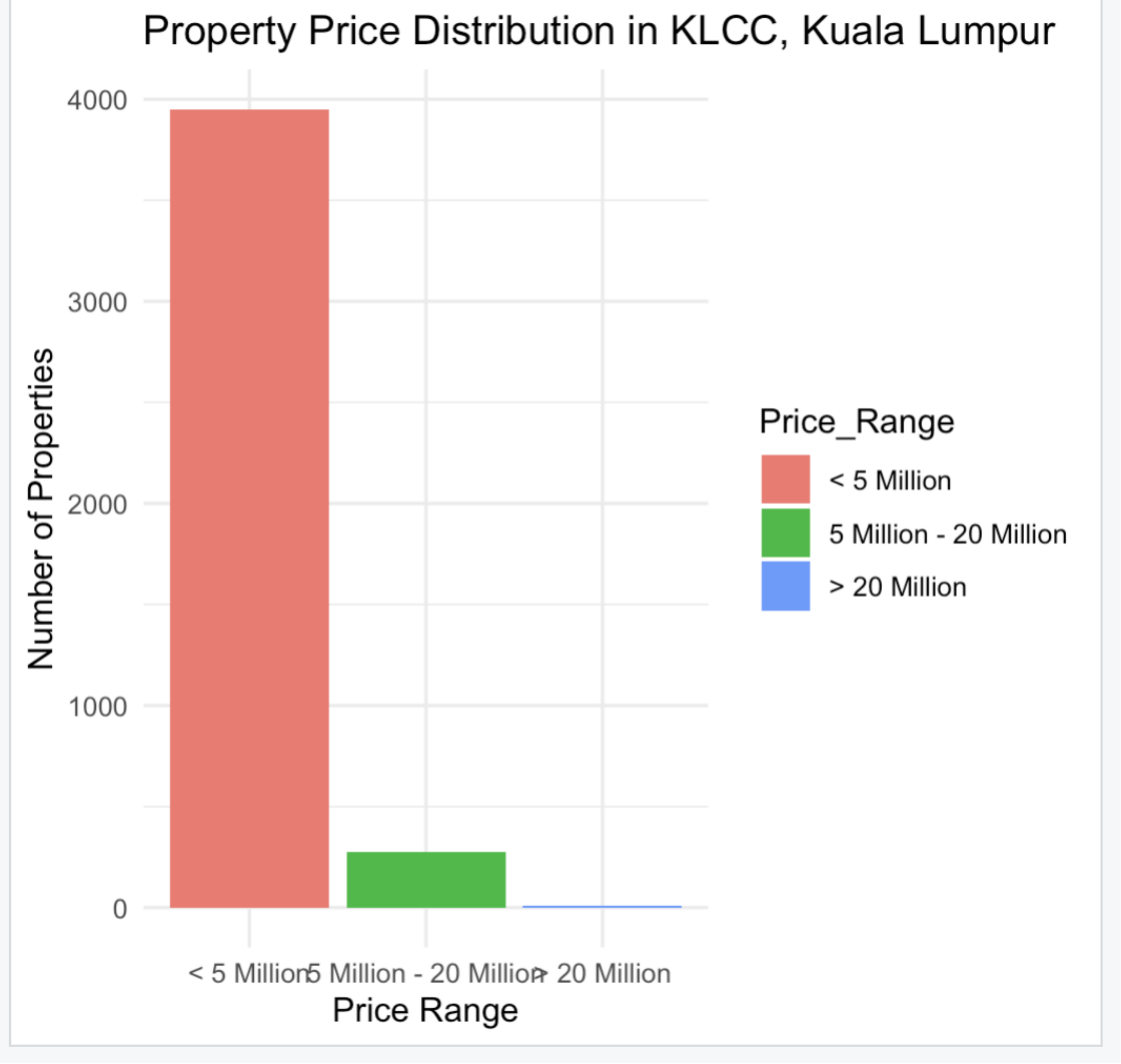


```r
# Convert 'Price(in RM)' to numeric
klcc_data$Price_in_RM <- as.numeric(gsub(",", "", klcc_data$`Price(in RM)`))

# Filter properties above 20 million
above_20m_properties <- klcc_data[klcc_data$Price_in_RM > 20000000, ]

# Create a scatterplot using ggplot2
library(ggplot2)
ggplot(above_20m_properties, aes(x = Price_in_RM, y = Price_in_RM, color = Price_in_RM)) +
  geom_point(size = 5, alpha = 0.8) +
  scale_color_gradient(low = "#3498db", high = "#2980b9") +
  labs(title = "Properties Above 20 Million in KLCC",
       x = "Price (in RM)",
       y = "Price (in RM)") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(legend.position = "none")
```

Third snippet.

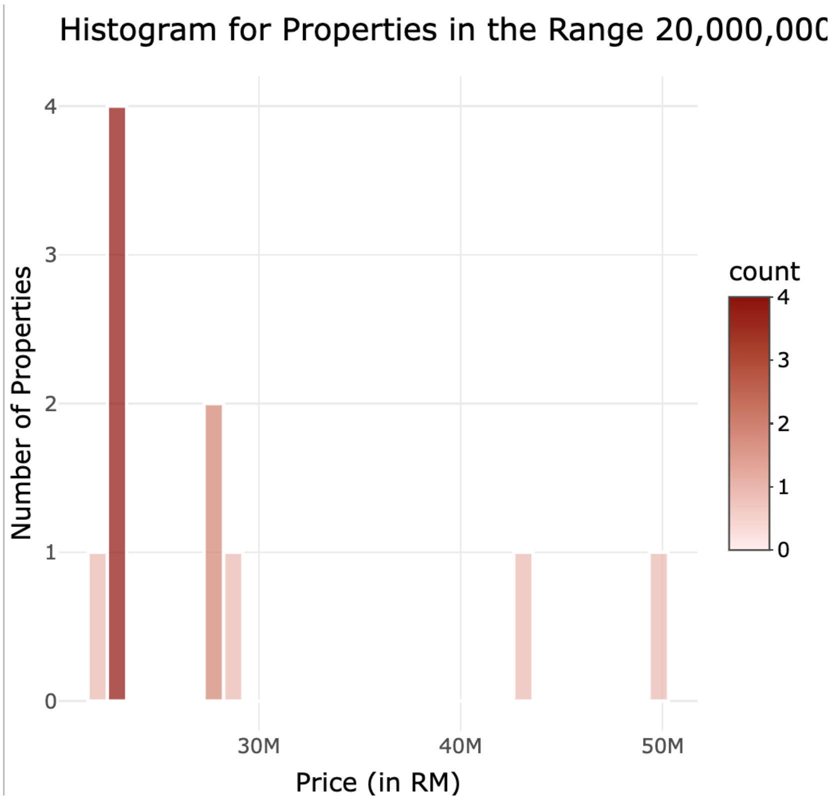
This code snippet uses ggplot2 to create a scatterplot that shows properties in KLCC that cost more than 20 million. First, the required ggplot2 library is loaded, and the 'Price(in RM)' column is converted to a numeric format for consistency. Next, property prices are plotted on the x and y axes to form a scatterplot, where each point is color-coded according to its price. The plot becomes more visually appealing using a gradient colour scale that goes from a light blue to a darker shade. For easier reading, commas are used in the formatting of the x and y axes. The resulting plot makes it easy to quickly analyse the price trends of high-value properties in KLCC by offering a clear and meaningful summary of their distribution. The legend has been purposefully left out for a cleaner and more convenient outlook.

Screenshot of Output



Output for the first code snippet.

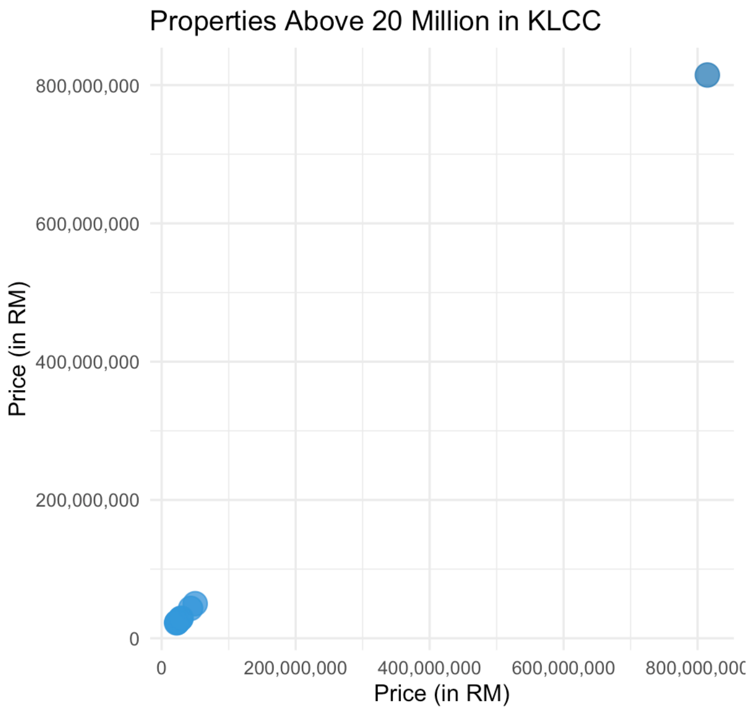
As you can see there are a lot of properties below the range of 5 million RM, 3950 properties to be exact and 276 properties between the range of 5 million RM to 20 million RM. Whereas the amount of properties above the range of 20 million is only 11.



Output for the second code snippet.

As you can see there are ten properties within the range of 20 million RM to 50 million RM.

Also there was a reason I decided to analyse on the properties above 20 million even further, It was to display an outlier through a scatter plot.



Output for the third snippet.

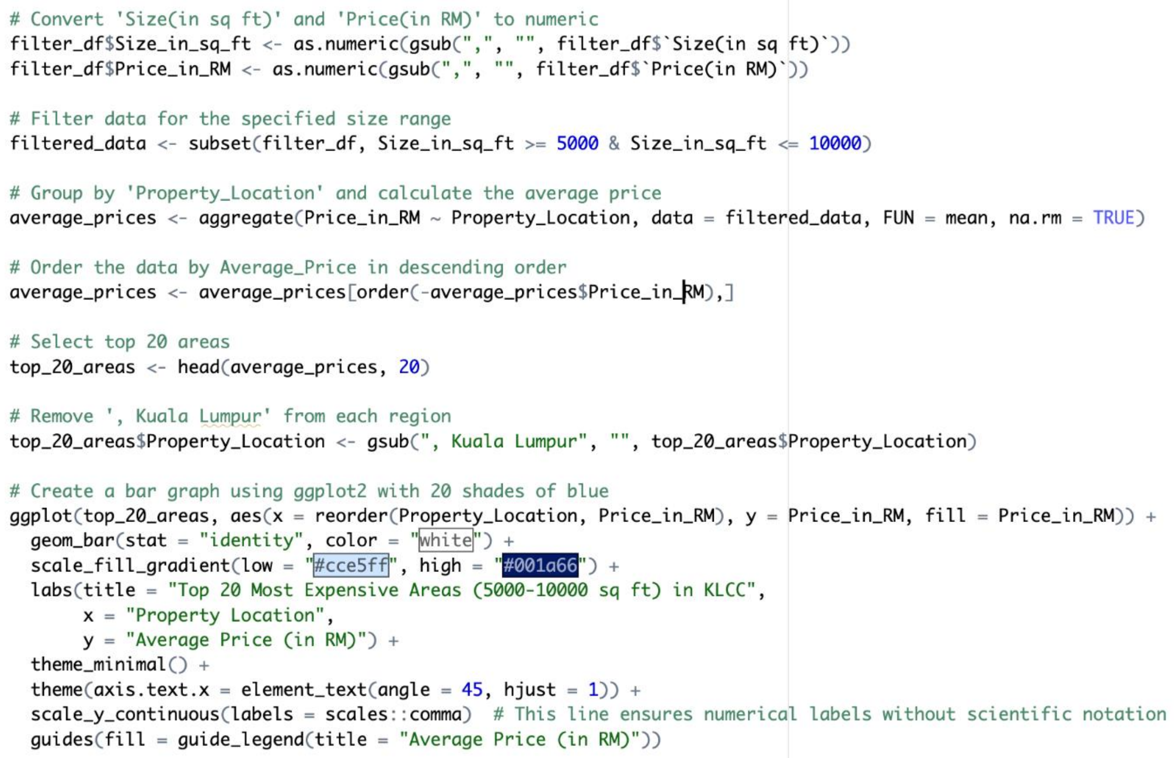
This code snippet uses ggplot2 to create a scatterplot that shows properties in KLCC that cost more than 20 million. First, the required ggplot2 library is loaded, and the 'Price(in RM)' column is converted to a numeric format for consistency. Next, property prices are plotted on the x and y axes to form a scatterplot, where each point is color-coded according to its price. The plot becomes more visually appealing using a gradient colour scale that goes from a light blue to a darker shade. For easier reading, commas are used in the formatting of the x and y axes. The resulting plot makes it easy to quickly analyse the price trends of high-value properties in KLCC by offering a clear and meaningful summary of their distribution. The legend has been purposefully left out for a cleaner and more convenient outlook.

Analysis Justification:

In analysis 3.1.2 the code is done to give us a range of property prices only in KLCC. This analysis displays a calculation and depiction of prices within the range of 5 million to 20 million and below 5 million, also above 20 million. This wide range was chosen to display the affordability in the bustling downtown area of KLCC. As you can see from the data and the graphs provided above the area KLCC is not as expensive as it may seem to be especially compared to other places as most properties are below the range of 5 million RM. And further analysis was done in R to depict the outlier through a scatterplot where range of price was chosen above 20 Million RM. Hence the justification can be made that the data was helpful to decide the fact that KLCC has most properties below the range of 5 million RM.

Analysis 3.1.3: To display the top 20 most expensive areas on average based on the sq. ft range of 5000-10000.

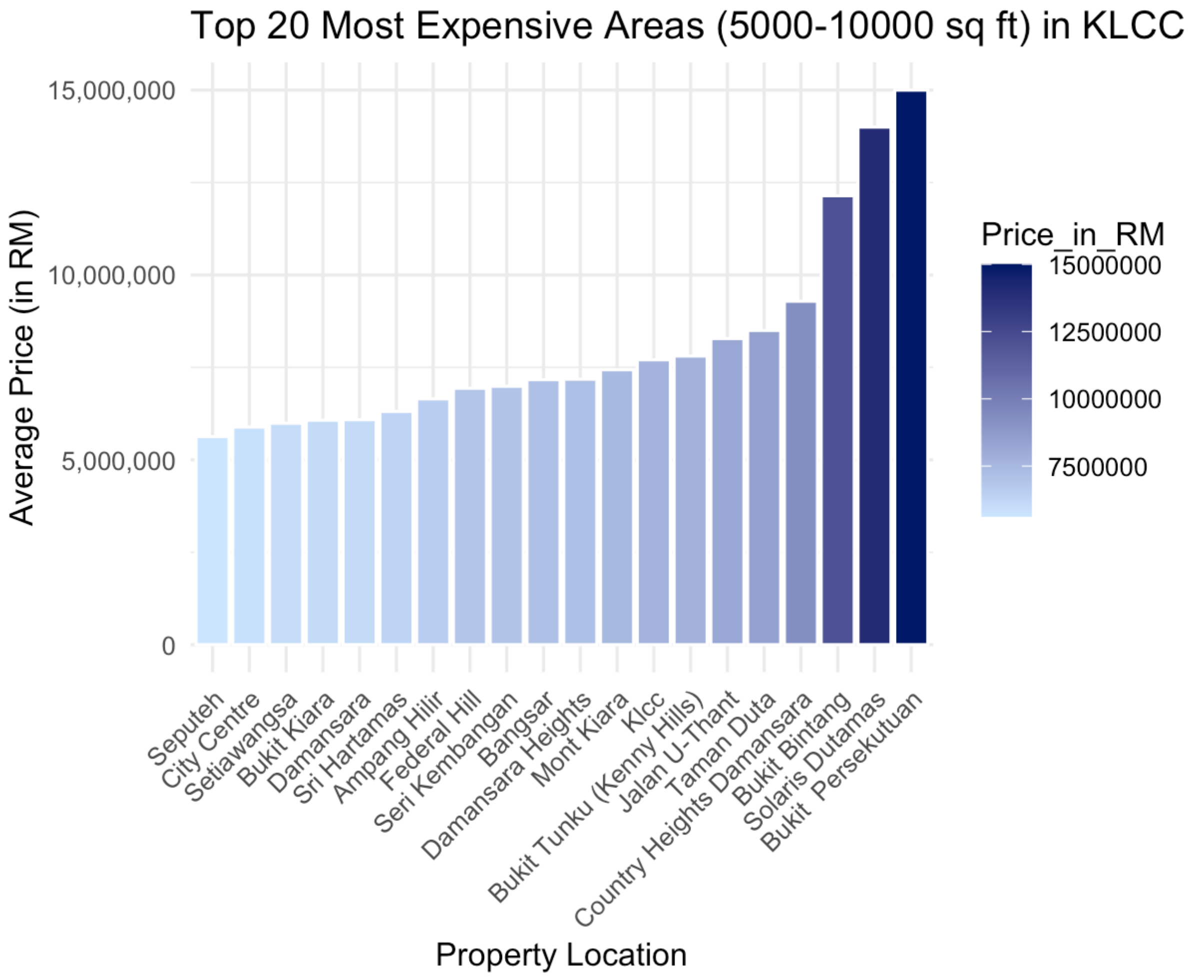
Screenshot of Code:



```
# Convert 'Size(in sq ft)' and 'Price(in RM)' to numeric
filter_df$Size_in_sq_ft <- as.numeric(gsub(",", "", filter_df$`Size(in sq ft)`))
filter_df$Price_in_RM <- as.numeric(gsub(",", "", filter_df$`Price(in RM)`))

# Filter data for the specified size range
filtered_data <- subset(filter_df, Size_in_sq_ft >= 5000 & Size_in_sq_ft <= 10000)

# Group by 'Property_Location' and calculate the average price
average_prices <- aggregate(Price_in_RM ~ Property_Location, data = filtered_data, FUN = mean, na.rm = TRUE)

# Order the data by Average_Price in descending order
average_prices <- average_prices[order(-average_prices$Price_in_RM),]

# Select top 20 areas
top_20_areas <- head(average_prices, 20)

# Remove ', Kuala Lumpur' from each region
top_20_areas$Property_Location <- gsub(", Kuala Lumpur", "", top_20_areas$Property_Location)

# Create a bar graph using ggplot2 with 20 shades of blue
ggplot(top_20_areas, aes(x = reorder(Property_Location, Price_in_RM), y = Price_in_RM, fill = Price_in_RM)) +
  geom_bar(stat = "identity", color = "white") +
  scale_fill_gradient(low = "#cce5ff", high = "#001a66") +
  labs(title = "Top 20 Most Expensive Areas (5000-10000 sq ft) in KLCC",
       x = "Property Location",
       y = "Average Price (in RM)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = scales::comma)  # This line ensures numerical labels without scientific notation
  guides(fill = guide_legend(title = "Average Price (in RM)"))
```

This code sample shows the average property prices in the top 20 KLCC neighbourhoods in Kuala Lumpur within a certain size range of 5000 to 10000 square feet. It does this by using the ggplot2 library. The average costs for every distinct "Property_Location" are computed and arranged in decreasing order after the dataset has been filtered according to the size range. After choosing the top 20, the region names are purified by eliminating the ', Kuala Lumpur' suffix. The average price of real estate in various places is shown by a bar graph that uses twenty colours of blue, making for an eye-catching presentation. In addition, a legend is provided to give context for the colour and the y-axis labels are structured with commas for numerical clarity.

Screenshot of Output:



This bar graph shows us the top 20 most expensive areas on average based on sq. ft of 5000-10000 ft. The Y axis depicts the Average Price in RM and the X axis depicts the property location. The range applied here goes up till 15 million in RM. Again this is only the average not the most expensive pricing of each area.

As you can see KLCC does not appear on this graph so even after picking out the condition of sq. ft to give a more detailed concept of the pricing of properties in Kuala Lumpur, there seems to be no evidence of KLCC belonging in the expensive category.

Analysis Justification:

With an emphasis on a particular size range of 5000 to 10000 square feet, the offered code and analysis seek to provide insights into the typical property values throughout the top 20 neighbourhoods in KLCC, Kuala Lumpur. The data visualisation gives a clear picture of the average property prices in various locations, shown as a bar graph with 20 hues of blue. It's crucial to remember that averages, not the highest prices in each location, are the basis for this research.

The algorithm efficiently filters and analyses the dataset analytically, displaying the average property price distribution. The lack of KLCC in the graph indicates that, on average, KLCC could not be among the most costly regions within the given size range. This might be the result of a number of things, such the effect of adjacent neighbourhoods with higher average prices in the selected size range or the availability of larger-sized houses in KLCC.

The data highlights how crucial it is to take certain factors into account when assessing home values, such as size range. Because averages might not fully convey the range of pricing variances, it also emphasises the importance of having a detailed grasp of real estate dynamics. While this code offers a useful beginning point for examining and visualising trends in property prices, it also stimulates more research into the variables impacting these patterns.

Conclusion:

It is clear from the thorough examination of KLCC property prices, which includes data on overall prices as well as specific size ranges, that KLCC properties—with or without size considerations—do not consistently show a pattern of being more expensive than other regions of Kuala Lumpur by more than 20%. The careful examination, which takes into account a number of variables including neighbourhood, size, and price range, offers a comprehensive view of the real estate market. This emphasises how crucial it is to look at particular parameters rather than assuming broad generalisations about real estate values. The dynamic real estate market in KLCC and its comparison to other places would be better understood with further investigation and study into localised patterns and affecting variables.

Additional Feature(s)

- Gradient on bar graph: allows for better visualization of data and range applied
- Grouping of bars: allows for better comparisons between two factors
- Tilt in x-axis unit: provides a cleaner and readable form

3.2 OBJECTIVE 2:

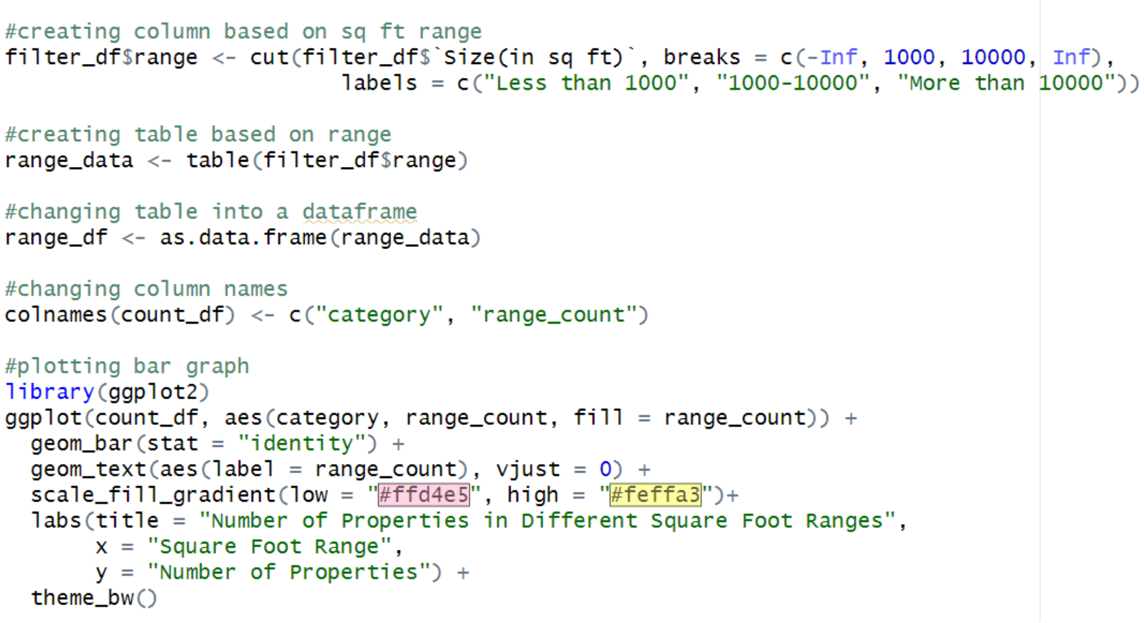
Student Name: Ameera Farooq

TP Number: TP071807

Specialism: Data Analysis

Analysis 3.2.1: Which square foot range has the highest number of properties

Screenshot of Code:



```
#creating column based on sq ft range
filter_df$range <- cut(filter_df$`Size(in sq ft)`, breaks = c(-Inf, 1000, 10000, Inf),
                          labels = c("Less than 1000", "1000-10000", "More than 10000"))

#creating table based on range
range_data <- table(filter_df$range)

#changing table into a dataframe
range_df <- as.data.frame(range_data)

#changing column names
colnames(count_df) <- c("category", "range_count")

#plotting bar graph
library(ggplot2)
ggplot(count_df, aes(category, range_count, fill = range_count)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = range_count), vjust = 0) +
  scale_fill_gradient(low = "#ffd4e5", high = "#feffa3")+
  labs(title = "Number of Properties in Different Square Foot Ranges",
       x = "Square Foot Range",
       y = "Number of Properties") +
  theme_bw()
```

Figure 1: code snippet for number of properties in square feet range

The provided code initiates the first analysis by categorizing the dataset into three ranges: "Less than 1000", "1000-10000", and "More than 10000". Subsequently, the data within each range is extracted and transformed into a separate dataframe, accompanied by adjustments to column names.

Finally, utilizing the ggplot2 library, a customized bar graph is generated using the newly created dataframe. The graph incorporates distinctive colors and labels to enhance visual representation.

Screenshot of Output:

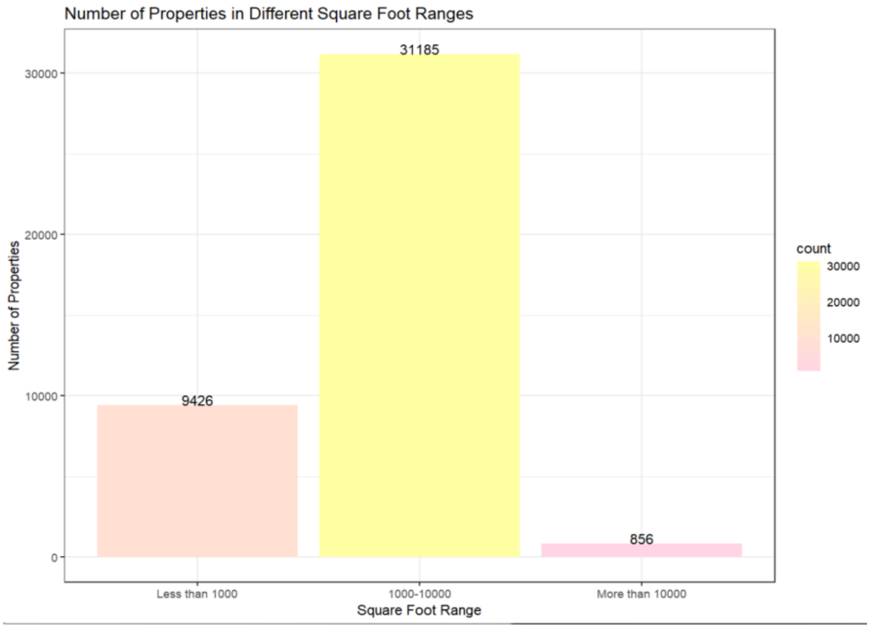


Figure 2: bar graph depicting number of properties with specified square feet range

The illustration provided above showcases the visual representation derived from the analysis, presenting a graph that encapsulates the distribution of properties across predefined ranges

Analysis Justification:

The bar graph shows that a significant number of properties fall within the 1000-10000 sq ft range, indicating that many properties have moderate to larger sizes. This variety caters to a diverse group of people looking for different property sizes.

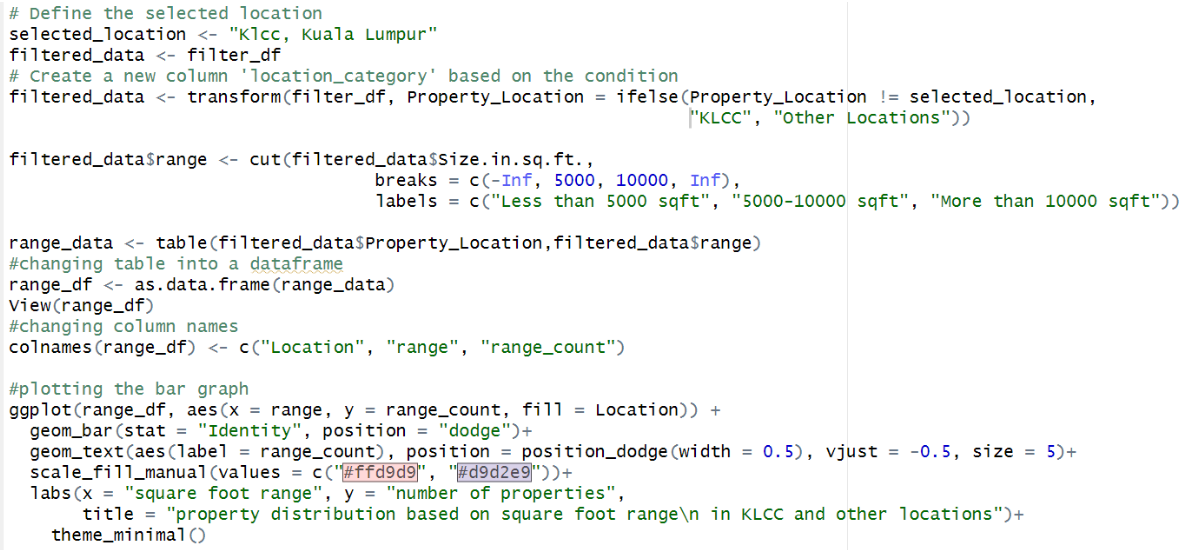
There's also a considerable number of properties (9,426) below 1000 sq ft. This highlights the presence of smaller-sized properties, like apartments or studios, which might be attractive to individuals or families wanting more compact living spaces.

On the other hand, the group of properties above 10000 sq ft is smaller (856). This suggests that very large properties are less common in the dataset, indicating that homes with extremely large square footage are not as frequently featured or sought after in the current real estate market.

These insights provide valuable information about the distribution of property sizes within the dataset, enabling a better understanding of the market dynamics. The dataset's diversity in size ranges, with a concentration in the moderate to larger sizes, offers insights into the preferences and offerings within the real estate market represented by the dataset. The presence of both smaller and larger properties contributes to a comprehensive view of the available housing options.

Analysis 3.2.2: what is the relationship between number of properties in KLCC and other locations based on square feet

Screenshot of Code:



```
# Define the selected location
selected_location <- "Klcc, Kuala Lumpur"
filtered_data <- filter_df
# Create a new column 'location_category' based on the condition
filtered_data <- transform(filter_df, Property_Location = ifelse(Property_Location != selected_location,
                                                                 "KLCC", "Other Locations"))

filtered_data$range <- cut(filtered_data$Size.in.sq.ft.,
                                   breaks = c(-Inf, 5000, 10000, Inf),
                                   labels = c("Less than 5000 sqft", "5000-10000 sqft", "More than 10000 sqft"))

range_data <- table(filtered_data$Property_Location,filtered_data$range)
#changing table into a dataframe
range_df <- as.data.frame(range_data)
View(range_df)
#changing column names
colnames(range_df) <- c("Location", "range", "range_count")

#plotting the bar graph
ggplot(range_df, aes(x = range, y = range_count, fill = Location)) +
  geom_bar(stat = "Identity", position = "dodge")+
  geom_text(aes(label = range_count), position = position_dodge(width = 0.5), vjust = -0.5, size = 5)+
  scale_fill_manual(values = c("#ffd9d9", "#d9d2e9"))+
  labs(x = "square foot range", y = "number of properties",
       title = "property distribution based on square foot range\n in KLCC and other locations")+
    theme_minimal()
```

Figure 3: code snippet for square feet ranges in klcc and other locations

The R code in Figure 3 does a property analysis and visualization, concentrating on square footage distribution in KLCC (Kuala Lumpur City Centre) and other locations. The data is initially filtered to include all properties, and an updated column 'Property_Location' is established to categorize each property as being in KLCC or another area based on a predefined requirement. Subsequently, the square footage values are divided into three categories: less than 1000 square feet, 1000-10000 square feet, and more than 10000 square feet. The resultant tabular data, 'range_df,' is turned into a data frame and shown using ggplot2's bar graph. The graph depicts the distribution of properties across various square footage ranges in KLCC and other locations, with specific colors used to symbolize each site.

Screenshot of Output:

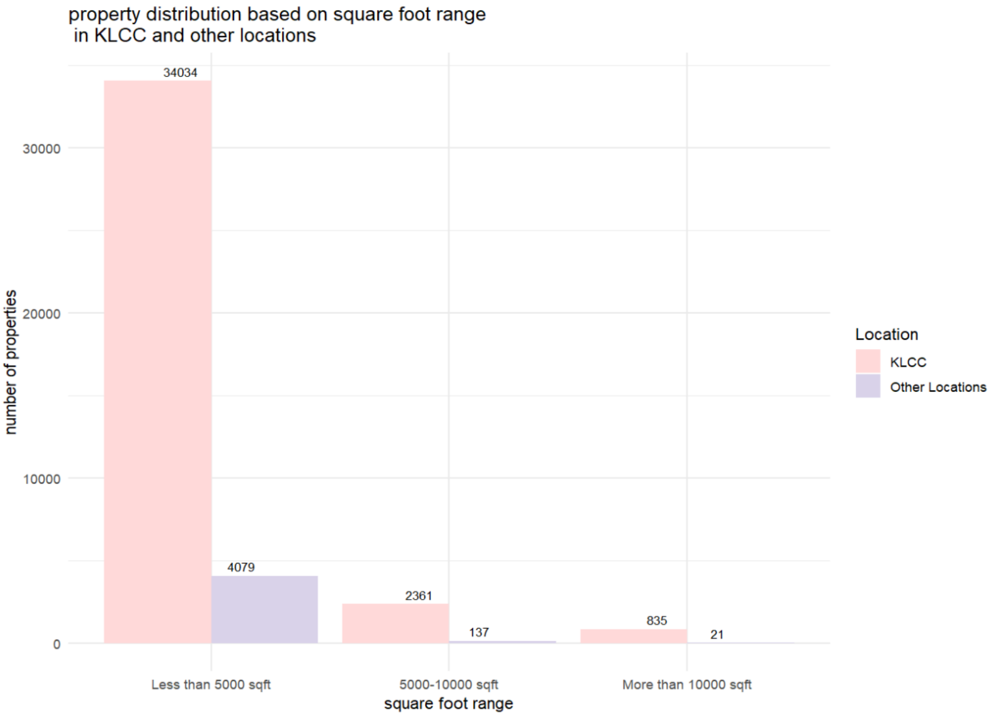


Figure 4: graph depicting the square feet ranges in klcc and other locations

The bar graph in Figure 4 shows the varying ranges of properties located in KLCC and other locations

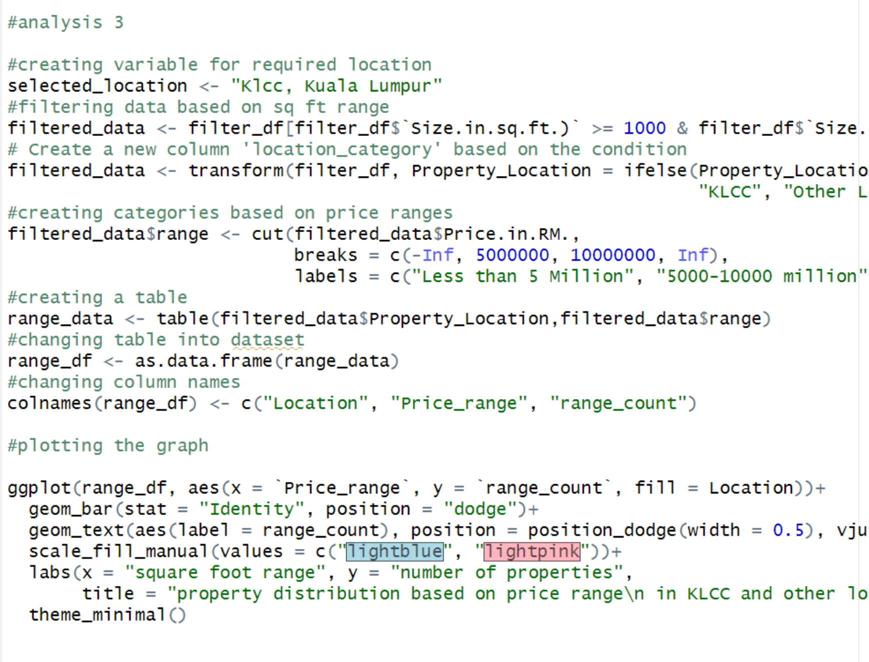
Analysis Justification:

The research of property distribution in KLCC and other areas demonstrates significant trends across various size categories. KLCC has a higher number of properties under 5000 square feet, with 34,034 compared to 4,079 in other areas. In the 5000-10000 sq ft category, both districts have a significant presence of mid-sized houses, but KLCC has a larger count at 2,361. KLCC leads with 835 properties over 10,000 square feet, while other sites have a lesser count of 21, demonstrating a concentration of larger-sized properties in KLCC.

Percentage disparities reflect KLCC's strong presence in smaller-sized properties, whilst the 5000-10000 sq ft and greater than 10000 sq ft categories show a more fair distribution. Calculating the average property size in each place would provide information about typical sizes in KLCC and other locations. Analyzing the relationship between property size and price, along with geographic mapping, could provide useful insights into pricing dynamics and spatial patterns, thereby improving our understanding of the real estate landscape.

Analysis 3.2.3: What is the relationship between KLCC and other locations based on square feet range and price

Screenshot of Code:



```r
#analysis 3

#creating variable for required location
selected_location <- "Klcc, Kuala Lumpur"
#filtering data based on sq ft range
filtered_data <- filter_df[filter_df$`Size.in.sq.ft.)` >= 1000 & filter_df$`Size.in.sq.ft.` <= 10000]
# Create a new column 'location_category' based on the condition
filtered_data <- transform(filter_df, Property_Location = ifelse(Property_Location != selected_location,
                                                                 "KLCC", "Other Locations"))
#creating categories based on price ranges
filtered_data$range <- cut(filtered_data$Price.in.RM.,
                           breaks = c(-Inf, 5000000, 10000000, Inf),
                           labels = c("Less than 5 Million", "5000-10000 million", "More than 10000000"))
#creating a table
range_data <- table(filtered_data$Property_Location,filtered_data$range)
#changing table into dataset
range_df <- as.data.frame(range_data)
#changing column names
colnames(range_df) <- c("Location", "Price_range", "range_count")

#plotting the graph

ggplot(range_df, aes(x = `Price_range`, y = `range_count`, fill = Location))+
  geom_bar(stat = "Identity", position = "dodge")+
  geom_text(aes(label = range_count), position = position_dodge(width = 0.5), vjust = -0.5, size = 3)+
  scale_fill_manual(values = c("lightblue", "lightpink"))+
  labs(x = "square foot range", y = "number of properties",
       title = "property distribution based on price range\n in KLCC and other locations")+
  theme_minimal()
```

Figure 5: code snippet for prices ranges in klcc and other locations based on square feet

The code snippet displayed in Figure 4 analyzes and visualizes property distribution in KLCC (Kuala Lumpur City Centre) and other areas according to price ranges. It limits the data to properties with square footages ranging from 1000 to 10,000 square feet. The 'Property_Location' category was designed to categorize properties as KLCC or other areas. Prices are then divided into three categories: less than 5 million, 5-10 million, and more than 10 million Malaysian Ringgit. The resultant tabular data, called 'range_df,' is utilized to create a bar graph with ggplot2. This graph depicts the property distribution across various price ranges in KLCC and other areas, with distinct colors identifying each location. The plot contains text labels for property counts and is customized with a simple style.

Screenshot of Output:

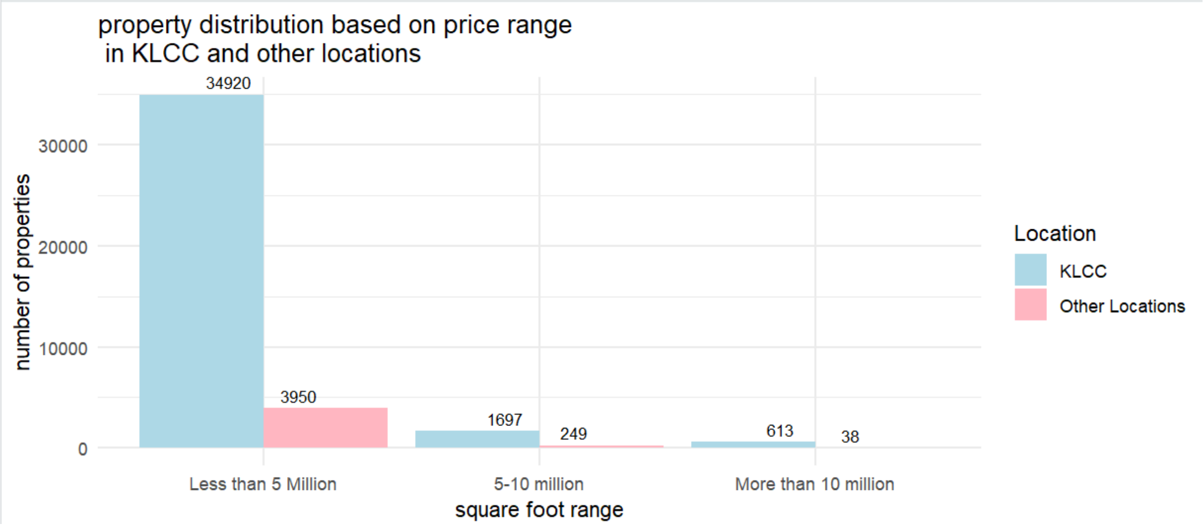


Figure 6: bar graph showing price ranges in klcc and other locations based on square feet

The bar graph in Figure 5 shows the prices ranging from less than 5 million, 5 million to 10 million and more than 10 million respectively.

<u>Analysis Justification:</u>

The data shown shows a clear pattern in the distribution of properties between KLCC and other locations across various price levels. KLCC stands out as a real estate hub, with much more properties than other regions.

Looking at the general distribution, KLCC has a significant concentration of properties in all three price ranges: less than 5 million, 5-10 million, and more than 10 million. This suggests a varied real estate market in KLCC, including both low- and high-end properties.

An in-depth investigation of the less than 5 million price range finds a significant presence of 34,920 homes in KLCC, reinforcing its standing as a popular location for more inexpensive housing options. This concentration highlights KLCC's appeal to a wide range of property.

In contrast, other localities have lower counts across all price levels, indicating a smaller concentration of homes in these categories. Other locations, despite their varying price ranges, do not have the same robust property density as KLCC.

Furthermore, the data demonstrates KLCC's prominent position in the high-end real estate market, with 613 houses priced over $10 million. In comparison, other locales have a small representation in this category, with only 38 properties topping the 10 million level.

Conclusion:

The dataset analysis highlights KLCC's dominance in the real estate market, which includes a varied variety of assets ranging from 1000 to 10,000 square feet. This variation appeals to a wide range of property buyers, with possibilities for both modest and large homes.

Other locations have fewer property counts across all price groups, indicating a more concentrated market in KLCC. The report underscores KLCC's appeal as a real estate hub, particularly in the sub-5 million price category, where it stands out for its large number of inexpensive housing options.

To summarize, KLCC's strong representation in the dataset across various sizes and price points emphasizes its importance in the real estate scene. The report provides significant insights for stakeholders and demonstrates KLCC's appeal to a varied variety of property seekers.

Additional Feature(s)

- Group bar graph: a visual representation of categorical data where multiple bars for each category are grouped together, facilitating comparisons within and across subcategories.
- Figure count on bars: an accurate count of the figures for the bar chart on each category has been provided in order to enhance understanding of the chart

3.3 OBJECTIVE 3: Impact of Furnishing Status on Property Prices in KLCC and Other Regions

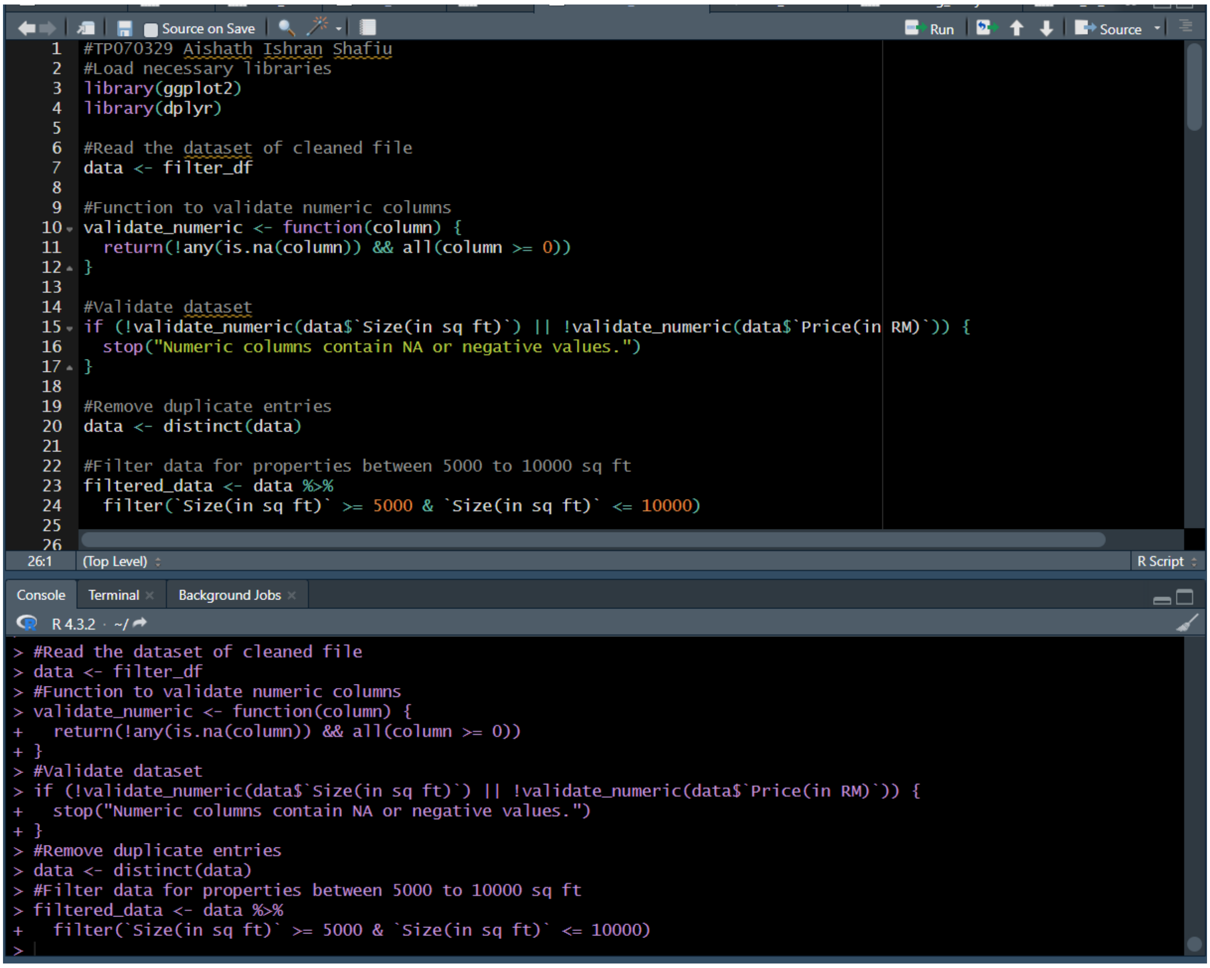
Student Name: Aishath Ishran Shaifu

TP Number: TP070329

Specialism: Data Analysis

Analysis 3.3.1: What is the impact of Furnishing Status on Property Prices?

<u>Screenshot of Code:</u>

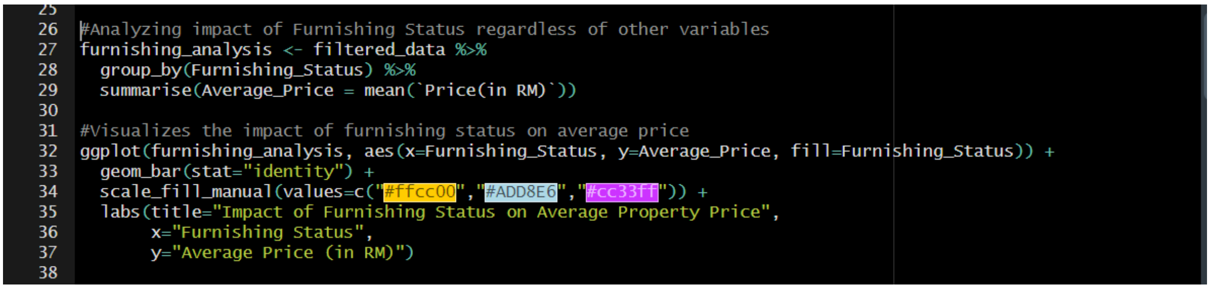


```
#TP070329 Aishath Ishran Shafiu
#Load necessary libraries
library(ggplot2)
library(dplyr)

#Read the dataset of cleaned file
data <- filter_df

#Function to validate numeric columns
validate_numeric <- function(column) {
  return(!any(is.na(column)) && all(column >= 0))
}

#Validate dataset
if (!validate_numeric(data$`Size(in sq ft)`) || !validate_numeric(data$`Price(in RM)`)) {
  stop("Numeric columns contain NA or negative values.")
}

#Remove duplicate entries
data <- distinct(data)

#Filter data for properties between 5000 to 10000 sq ft
filtered_data <- data %>%
  filter(`Size(in sq ft)` >= 5000 & `Size(in sq ft)` <= 10000)

```

```
> #Read the dataset of cleaned file
> data <- filter_df
> #Function to validate numeric columns
> validate_numeric <- function(column) {
+   return(!any(is.na(column)) && all(column >= 0))
+ }
> #Validate dataset
> if (!validate_numeric(data$`Size(in sq ft)`) || !validate_numeric(data$`Price(in RM)`)) {
+   stop("Numeric columns contain NA or negative values.")
+ }
> #Remove duplicate entries
> data <- distinct(data)
> #Filter data for properties between 5000 to 10000 sq ft
> filtered_data <- data %>%
+   filter(`Size(in sq ft)` >= 5000 & `Size(in sq ft)` <= 10000)
> 
```

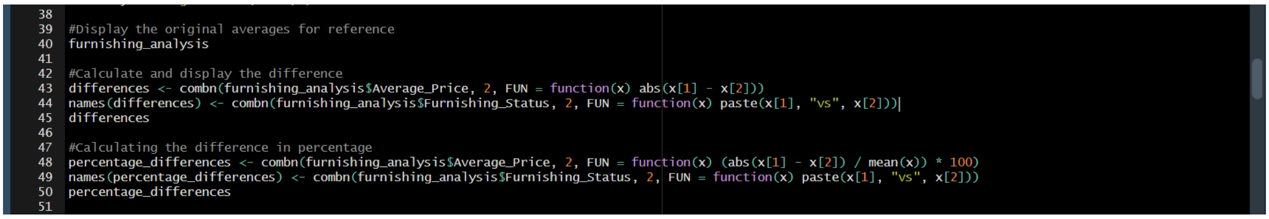
Figure 4.0 Snippet of Code: Data Validation

The cleaned dataset, known as "filter_df", was first loaded into a data frame, named "data", to undergo further data validation before beginning the analysis. "filtered_data" contains all the properties between 5000 to 10000 sq feet.



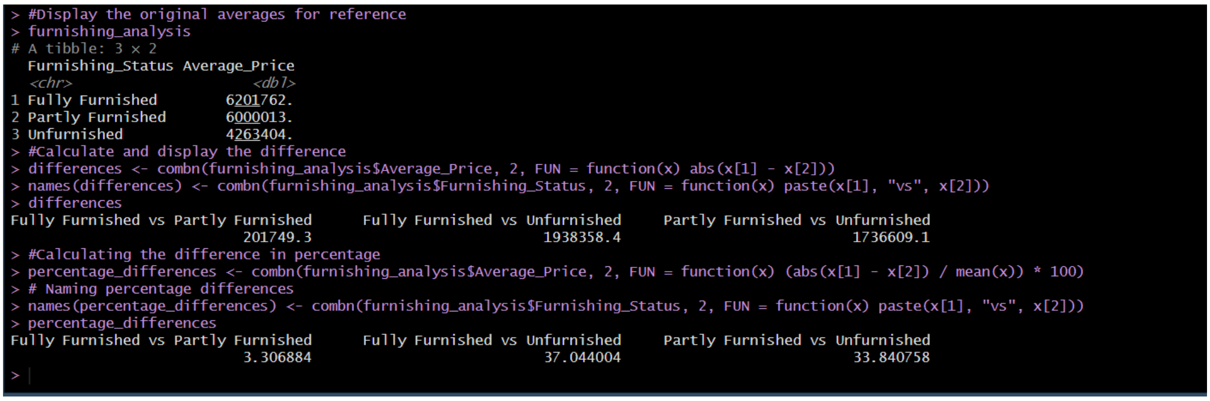
```
#Analyzing impact of Furnishing Status regardless of other variables
furnishing_analysis <- filtered_data %>%
  group_by(Furnishing_Status) %>%
  summarise(Average_Price = mean(`Price(in RM)`))

#Visualizes the impact of furnishing status on average price
ggplot(furnishing_analysis, aes(x=Furnishing_Status, y=Average_Price, fill=Furnishing_Status)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values=c("#ffcc00","#ADD8E6","#cc33ff")) +
  labs(title="Impact of Furnishing Status on Average Property Price",
       x="Furnishing Status",
       y="Average Price (in RM)")

```

Figure 4.1 Code for Bar Graph



```

#Display the original averages for reference
furnishing_analysis

#Calculate and display the difference
differences <- combn(furnishing_analysis$Average_Price, 2, FUN = function(x) abs(x[1] - x[2]))
names(differences) <- combn(furnishing_analysis$Furnishing_Status, 2, FUN = function(x) paste(x[1], "vs", x[2]))
differences

#Calculating the difference in percentage
percentage_differences <- combn(furnishing_analysis$Average_Price, 2, FUN = function(x) (abs(x[1] - x[2]) / mean(x)) * 100)
names(percentage_differences) <- combn(furnishing_analysis$Furnishing_Status, 2, FUN = function(x) paste(x[1], "vs", x[2]))
percentage_differences

```

Figure 7.2 Code for Finding Differences Between Mean Prices



```
> #Display the original averages for reference
> furnishing_analysis
# A tibble: 3 × 2
  Furnishing_Status Average_Price
  <chr>                     <dbl>
1 Fully Furnished         6201762.
2 Partly Furnished        6000013.
3 Unfurnished             4263404.
> #Calculate and display the difference
> differences <- combn(furnishing_analysis$Average_Price, 2, FUN = function(x) abs(x[1] - x[2]))
> names(differences) <- combn(furnishing_analysis$Furnishing_Status, 2, FUN = function(x) paste(x[1], "vs", x[2]))
> differences
Fully Furnished vs Partly Furnished      Fully Furnished vs Unfurnished     Partly Furnished vs Unfurnished 
                           201749.3                           1938358.4                           1736609.1 
> #Calculating the difference in percentage
> percentage_differences <- combn(furnishing_analysis$Average_Price, 2, FUN = function(x) (abs(x[1] - x[2]) / mean(x)) * 100)
> # Naming percentage differences
> names(percentage_differences) <- combn(furnishing_analysis$Furnishing_Status, 2, FUN = function(x) paste(x[1], "vs", x[2]))
> percentage_differences
Fully Furnished vs Partly Furnished      Fully Furnished vs Unfurnished     Partly Furnished vs Unfurnished 
                           3.306884                           37.044004                           33.840758 
> 
```

Figure 4.3 Output for Figure 4.2

<u>Screenshot of Output:</u>

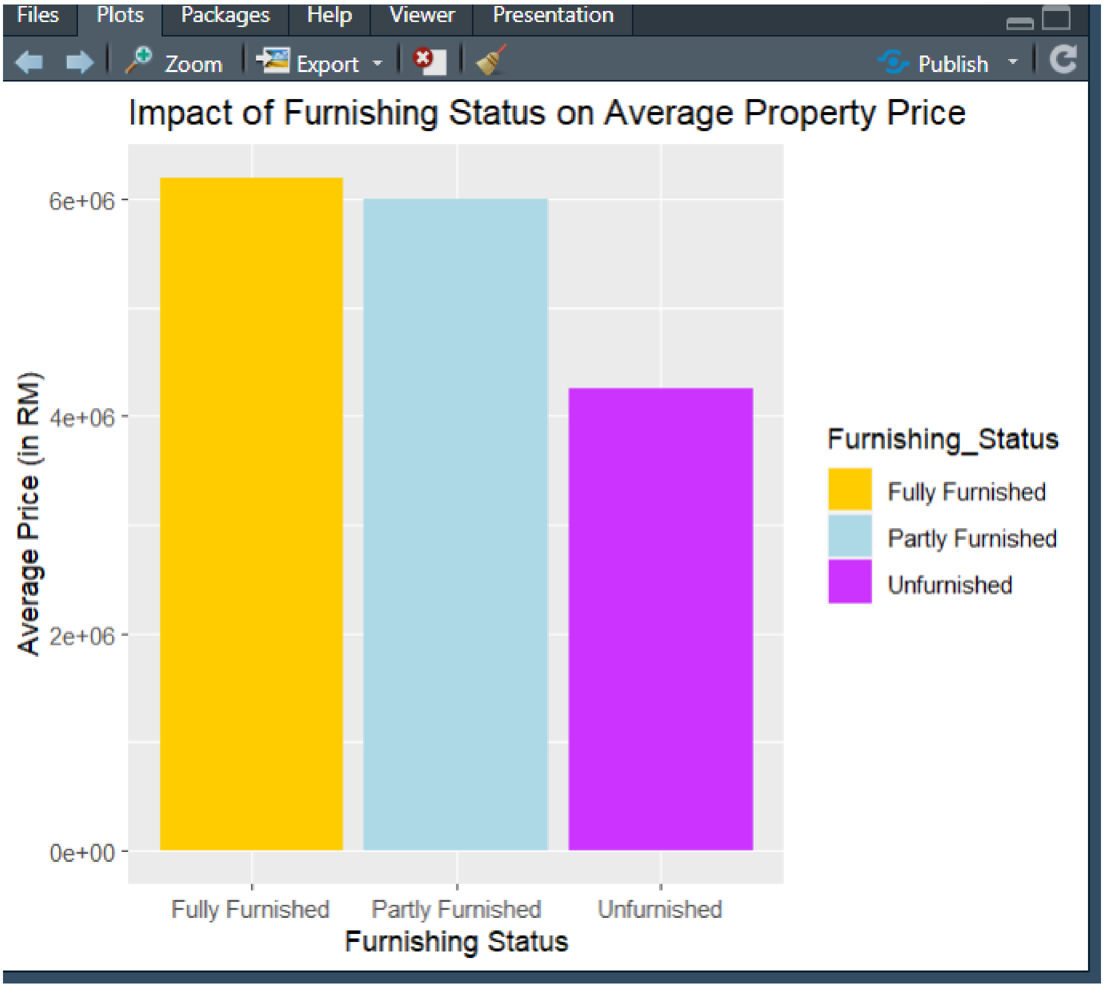


Figure 4.4 Result of Bar Graph

Analysis Justification:

There are three states of being furnished: Fully Furnished, Partly Furnished and Unfurnished. The properties were grouped by their respective states and the average price for each of the groups were made. The bar graph generated in this section illustrates the average property prices categorized by furnishing status.

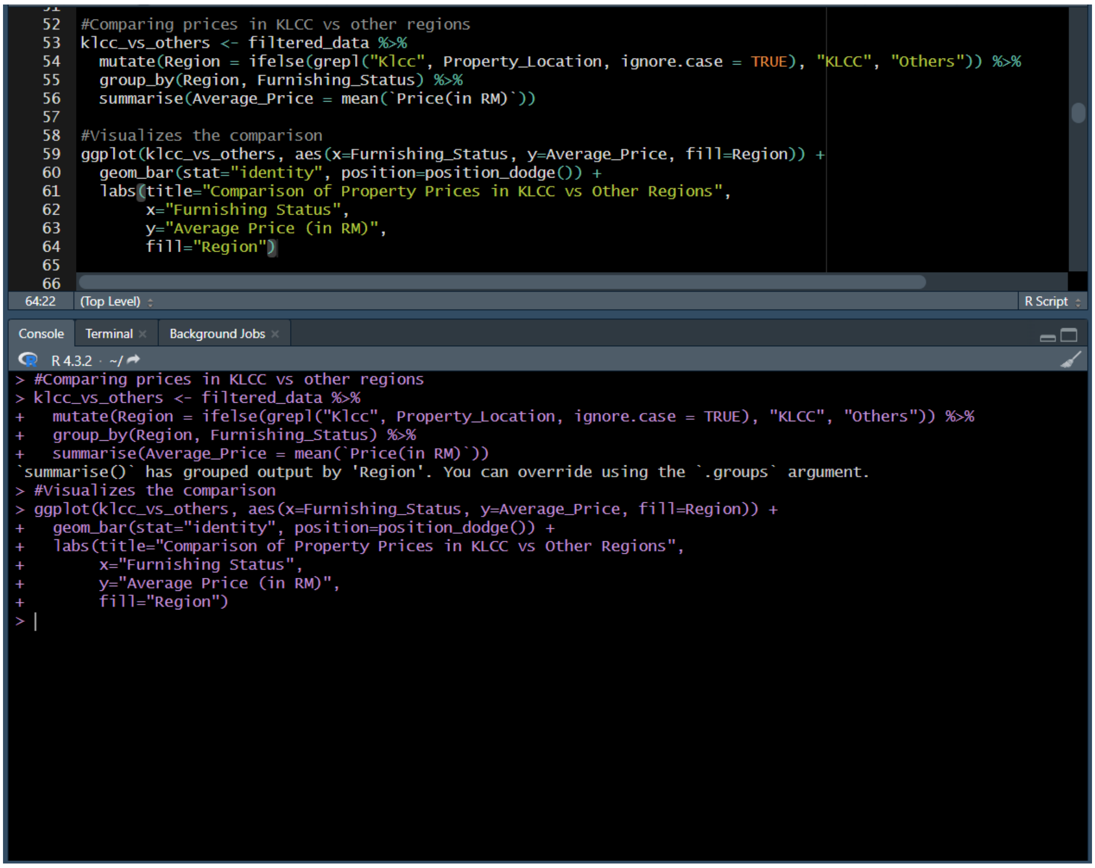
The graph in Figure 4.4 does indeed show that there is a correlation between furnishing status and price.

For a more comprehensive evaluation, additional calculations have been made to explain the results of the graph. The difference between the mean prices for Fully Furnished and Partly Furnished is only 3.31%. The difference between Unfurnished with both the other furnishing states is a whopping 37% and 33.8% (Fully Furnished and Partly Furnished respectively).

In conclusion, the furnishing status plays a role on the price. From Unfurnished compared to other furnishing states the impact is greater. There is not a considerable difference in mean price when it comes to being Fully or Partly furnished.

Analysis 3.3.2: What is the Price Comparison Between KLCC and Other Regions?

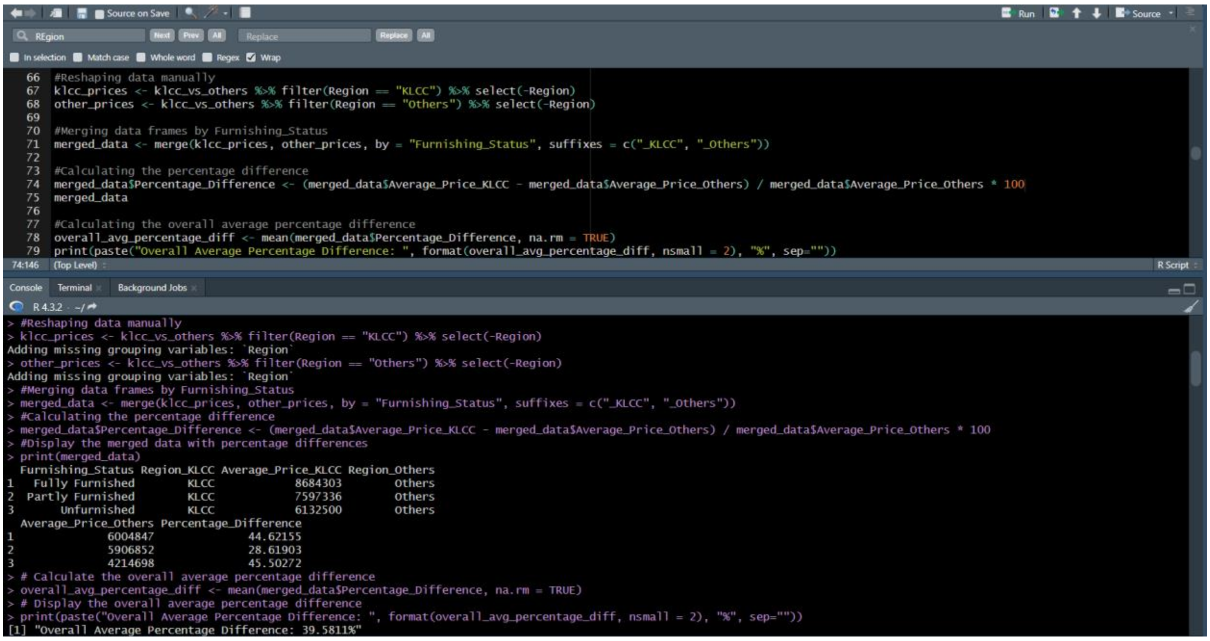
Screenshot of Code:



```
#Comparing prices in KLCC vs other regions
klcc_vs_others <- filtered_data %>%
  mutate(Region = ifelse(grepl("Klcc", Property_Location, ignore.case = TRUE), "KLCC", "Others")) %>%
  group_by(Region, Furnishing_Status) %>%
  summarise(Average_Price = mean(`Price(in RM)`))

#Visualizes the comparison
ggplot(klcc_vs_others, aes(x=Furnishing_Status, y=Average_Price, fill=Region)) +
  geom_bar(stat="identity", position=position_dodge()) +
  labs(title="Comparison of Property Prices in KLCC vs Other Regions",
       x="Furnishing Status",
       y="Average Price (in RM)",
       fill="Region")
```

```
> #Comparing prices in KLCC vs other regions
> klcc_vs_others <- filtered_data %>%
+   mutate(Region = ifelse(grepl("Klcc", Property_Location, ignore.case = TRUE), "KLCC", "Others")) %>%
+   group_by(Region, Furnishing_Status) %>%
+   summarise(Average_Price = mean(`Price(in RM)`))
`summarise()` has grouped output by 'Region'. You can override using the `.groups` argument.
> #Visualizes the comparison
> ggplot(klcc_vs_others, aes(x=Furnishing_Status, y=Average_Price, fill=Region)) +
+   geom_bar(stat="identity", position=position_dodge()) +
+   labs(title="Comparison of Property Prices in KLCC vs Other Regions",
+        x="Furnishing Status",
+        y="Average Price (in RM)",
+        fill="Region")
> |
```

Figure 4.5



```
#Reshaping data manually
klcc_prices <- klcc_vs_others %>% filter(Region == "KLCC") %>% select(-Region)
other_prices <- klcc_vs_others %>% filter(Region == "Others") %>% select(-Region)

#Merging data frames by Furnishing_Status
merged_data <- merge(klcc_prices, other_prices, by = "Furnishing_Status", suffixes = c("_KLCC", "_Others"))

#Calculating the percentage difference
merged_data$Percentage_Difference <- (merged_data$Average_Price_KLCC - merged_data$Average_Price_Others) / merged_data$Average_Price_Others * 100
merged_data

#Calculating the overall average percentage difference
overall_avg_percentage_diff <- mean(merged_data$Percentage_Difference, na.rm = TRUE)
print(paste("Overall Average Percentage Difference: ", format(overall_avg_percentage_diff, nsmall = 2), "%", sep=""))
```

```
> #Reshaping data manually
> klcc_prices <- klcc_vs_others %>% filter(Region == "KLCC") %>% select(-Region)
Adding missing grouping variables: `Region`
> other_prices <- klcc_vs_others %>% filter(Region == "Others") %>% select(-Region)
Adding missing grouping variables: `Region`
> #Merging data frames by Furnishing_Status
> merged_data <- merge(klcc_prices, other_prices, by = "Furnishing_Status", suffixes = c("_KLCC", "_Others"))
> #Calculating the percentage difference
> merged_data$Percentage_Difference <- (merged_data$Average_Price_KLCC - merged_data$Average_Price_Others) / merged_data$Average_Price_Others * 100
> #Display the merged data with percentage differences
> print(merged_data)
  Furnishing_Status Region_KLCC Average_Price_KLCC Region_Others
1   Fully Furnished        KLCC            8684303        Others
2  Partly Furnished        KLCC            7597336        Others
3       Unfurnished        KLCC            6132500        Others
  Average_Price_Others Percentage_Difference
1              6004847              44.62155
2              5906852              28.61903
3              4214698              45.50272
> # Calculate the overall average percentage difference
> overall_avg_percentage_diff <- mean(merged_data$Percentage_Difference, na.rm = TRUE)
> # Display the overall average percentage difference
> print(paste("Overall Average Percentage Difference: ", format(overall_avg_percentage_diff, nsmall = 2), "%", sep=""))
[1] "Overall Average Percentage Difference: 39.5811%"
```

Figure 4.6 Calculating the Overall Difference between KLCC and Other Regions

<u>Screenshot of Output:</u>

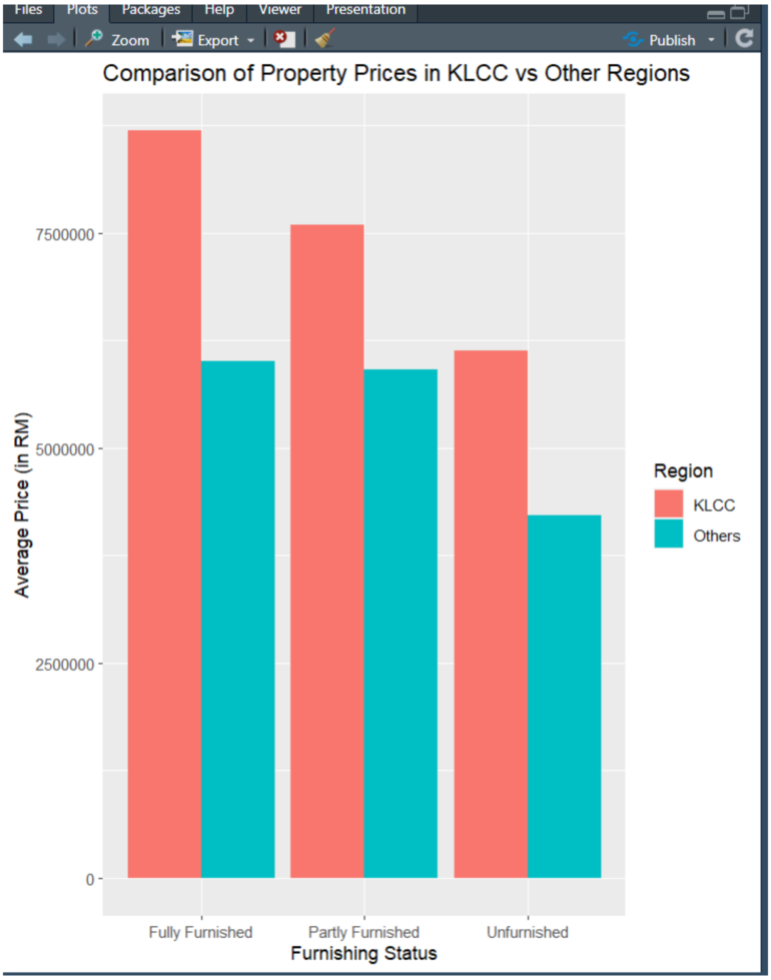


Figure 4.7 Bar Graph: KLCC vs Other Regions

Analysis Justification:

This comparison is visualized through a bar plot showing average prices in KLCC against other regions, segregated by furnishing status. The expected observation would be that properties in KLCC are hypothesized to be priced 20% higher than those in other regions, given similar size and furnishing status.

The above bar graph in Figure 4.7 depicts the mean price for each furnishing status in KLCC and other regions. Looking at the graph there is a significant price difference between the prices. In all furnishing states the properties in KLCC are more expensive compared to other regions. More calculations shall be conducted to explain the results.

To begin, the price difference between KLCC and Others for each furnishing status was calculated. The prices for each region are separated into two data frames known as klcc_prices and other_prices. Then they are merged in accordance to Furnishing_Status. Now that the prices that are to be compared with each other are all in the same row the difference between them can be calculated. The overall difference between KLCC and Other regions are as such:

Fully Furnished: 44.6%

Partly Furnished: 28.61%

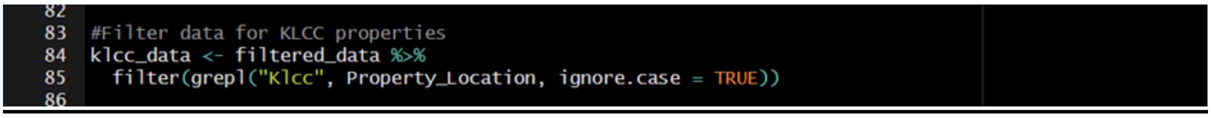
Unfurnished: 45.50%

Overall percentage difference: 39.6%

Compared to the expected observation, the calculated result is 19.6% more.

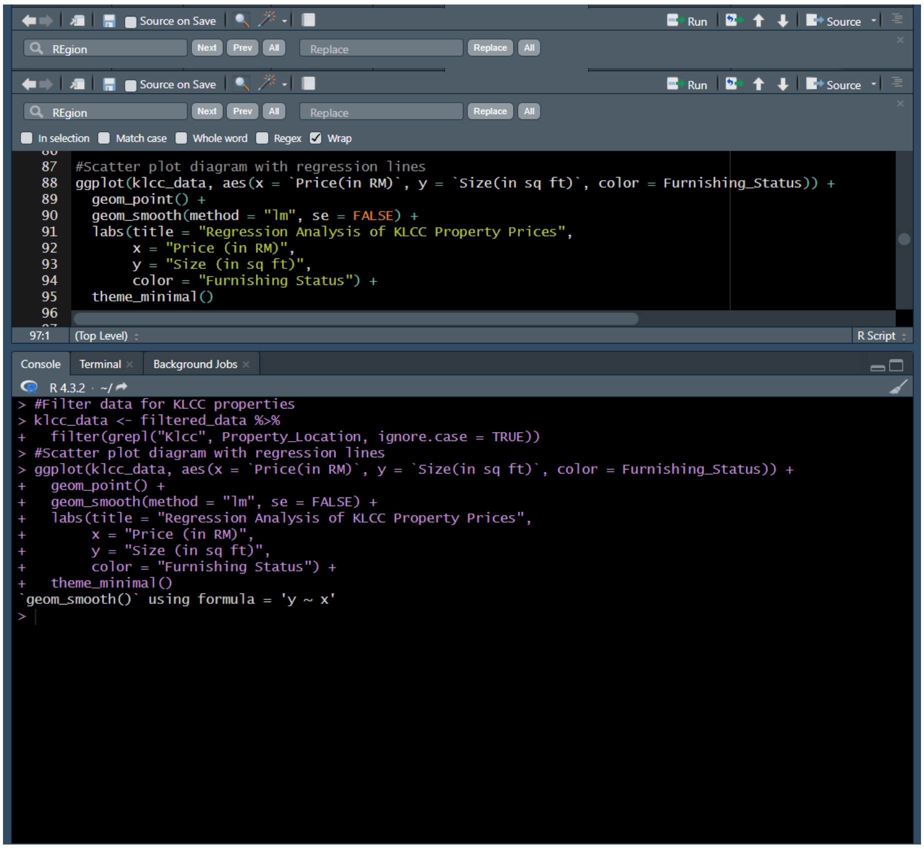
Analysis 3.3.3: Is there a Significant Impact of Furnishing Status on Property Price?

Screenshot of Code:



```
#Filter data for KLCC properties
klcc_data <- filtered_data %>%
  filter(grepl("Klcc", Property_Location, ignore.case = TRUE))
```

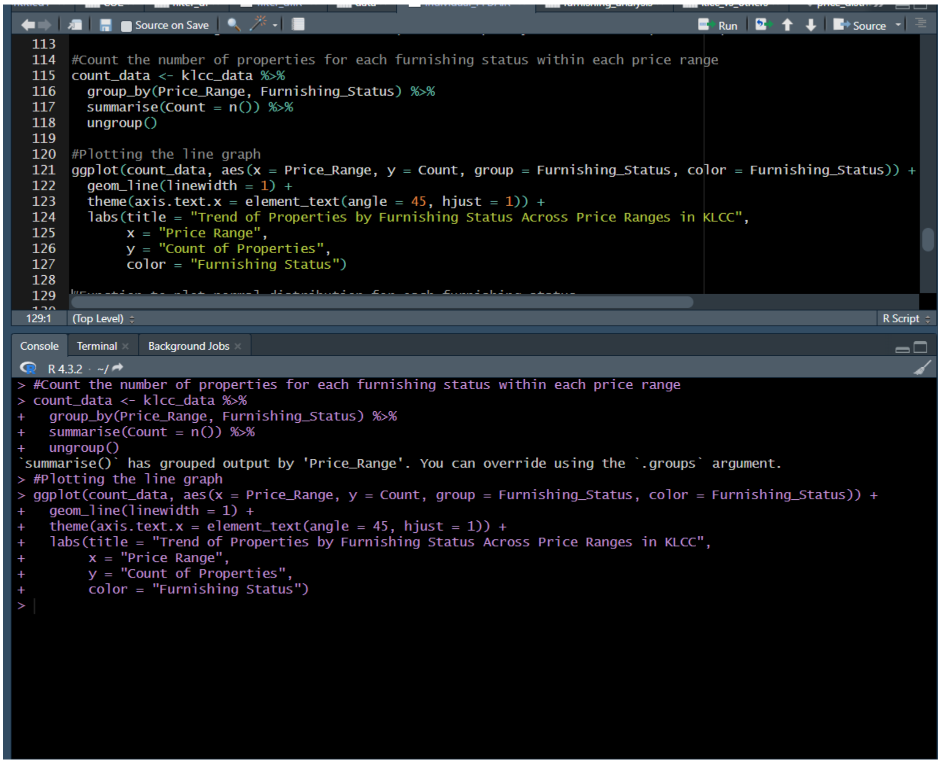
Figure 4.8 Code Snippet for klcc_data



```
#Scatter plot diagram with regression lines
ggplot(klcc_data, aes(x = `Price(in RM)`, y = `Size(in sq ft)`, color = Furnishing_Status)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Regression Analysis of KLCC Property Prices",
       x = "Price (in RM)",
       y = "Size (in sq ft)",
       color = "Furnishing Status") +
  theme_minimal()
```

```
> #Filter data for KLCC properties
> klcc_data <- filtered_data %>%
+   filter(grepl("Klcc", Property_Location, ignore.case = TRUE))
> #Scatter plot diagram with regression lines
> ggplot(klcc_data, aes(x = `Price(in RM)`, y = `Size(in sq ft)`, color = Furnishing_Status)) +
+   geom_point() +
+   geom_smooth(method = "lm", se = FALSE) +
+   labs(title = "Regression Analysis of KLCC Property Prices",
+        x = "Price (in RM)",
+        y = "Size (in sq ft)",
+        color = "Furnishing Status") +
+   theme_minimal()
`geom_smooth()` using formula = 'y ~ x'
>
```

Figure 4.9 Code Snippet for Scatter Plot Graph



```r

#Count the number of properties for each furnishing status within each price range
count_data <- klcc_data %>%
  group_by(Price_Range, Furnishing_Status) %>%
  summarise(Count = n()) %>%
  ungroup()

#Plotting the line graph
ggplot(count_data, aes(x = Price_Range, y = Count, group = Furnishing_Status, color = Furnishing_Status)) +
  geom_line(linewidth = 1) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Trend of Properties by Furnishing Status Across Price Ranges in KLCC",
       x = "Price Range",
       y = "Count of Properties",
       color = "Furnishing Status")

#Function to plot normal distribution for each furnishing status
```

```
> #Count the number of properties for each furnishing status within each price range
> count_data <- klcc_data %>%
+   group_by(Price_Range, Furnishing_Status) %>%
+   summarise(Count = n()) %>%
+   ungroup()
`summarise()` has grouped output by 'Price_Range'. You can override using the `.groups` argument.
> #Plotting the line graph
> ggplot(count_data, aes(x = Price_Range, y = Count, group = Furnishing_Status, color = Furnishing_Status)) +
+   geom_line(linewidth = 1) +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
+   labs(title = "Trend of Properties by Furnishing Status Across Price Ranges in KLCC",
+        x = "Price Range",
+        y = "Count of Properties",
+        color = "Furnishing Status")
> 
```

Figure 4.10 Code Snippet for Line Graph

Screenshot of Output:

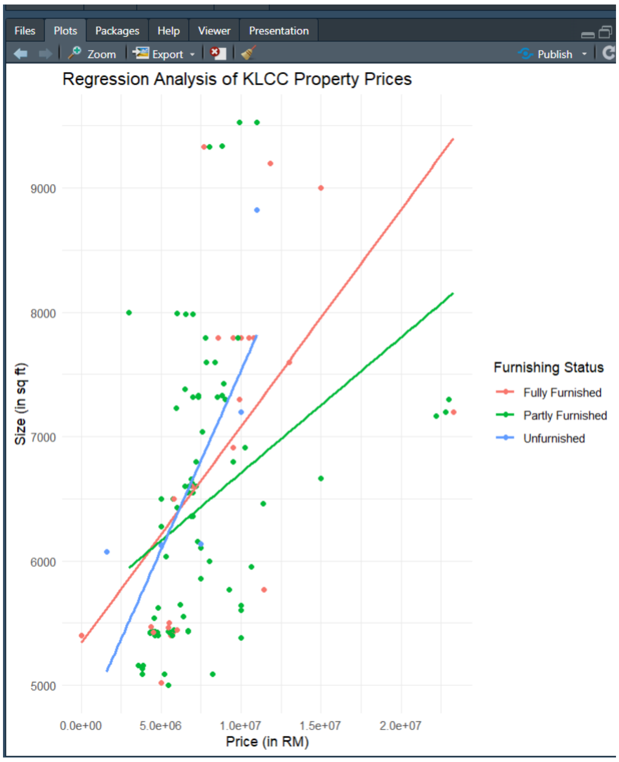


Figure 4.11 Scatter Plot Graph

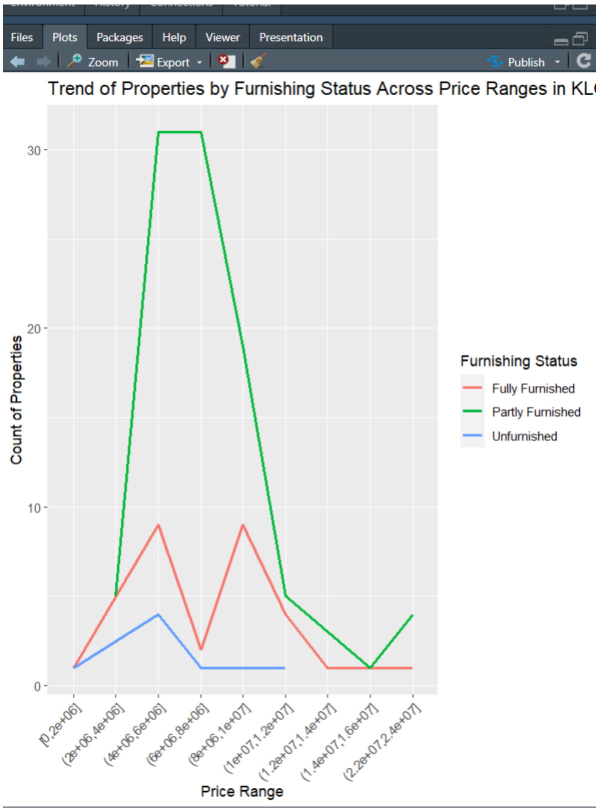


Figure 4.12 Line Graph

Analysis Justification:

To show how impactful that furnishing status truly is, additional graphs shall be illustrated. From this point on another data set is made, called klcc_data. This data set contains properties that are in between 5000 to 10000 sq feet and is located in the region KLCC.

The scatter plot in Figure 1.11 shows how the size of the property and furnishing status affects the price. A line of best fit is drawn to make better inferences from. There are several conclusions that can be made. Firstly, across all lines a consistent trend is seen. As the value of the y-axis (Size in sq ft) increases so does the value of the x-axis (Price in RM). This indicates regardless of furnishing status the size is a factor for price. Secondly, the steepest line is the Unfurnished line. That means the size of an unfurnished property plays a significant role in its selling price comparative to other furnishing status. The size of partly furnished properties matters the least while a moderate impact of size can be seen in fully furnished properties. Lastly, many points deviate from the line of best fit, despite there being an indication of a positive correlation to be seen. The correlation between size and price does not seem to be as significant as furnishing status.

The line graph in Figure 1.12 shows several insights provided. Firstly, the value of adding furnishing, where the fully and partly furnished properties have a wide range of price brackets while the unfurnished properties occupy the lower bracket of the price range. Secondly, the high activity in 4 million (4e+06) to 8 million (8e+06) range: There are a large number of properties within this price range. Most noticeably, the partly furnished properties have a high concentration within this region. This indicates that there is a robust market for mid- to high-end properties. Lastly, the lower demand for unfurnished properties. There are low numbers of unfurnished properties across all price ranges. This indicates that the market for unfurnished places is unattractive to most buyers.

Conclusion:

The data demonstrates a correlation between furnishing status and property values, with Unfurnished houses exhibiting a 37% to 33.8% difference from Fully Furnished and Partly Furnished. The bar graph in KLCC shows a substantial price differential across all furnishing states, with KLCC residences priced 44.6% more for Fully Furnished, 28.61% for Partly Furnished, and 45.50% for Unfurnished than in other regions. However, this is greater than the projected 20%, demonstrating a significant impact of location on costs.

Further research inside KLCC, concentrating on houses ranging from 5000 to 10000 sq ft, reveals that size constantly effects pricing, with unfurnished properties having the greatest impact. The scatter plot shows a positive relationship between size and price across all furnishing statuses, with unfurnished properties having the strongest correlation. The line graph illustrates the importance of adding furnishings, with Fully and Partly Furnished houses having a broader price range. Notably, demand for unfurnished properties is decreased across all price groups, indicating a less appealing market for such properties.

In conclusion, furnishing status has a significant impact on property prices, notably in KLCC, where the observed variances outweigh the expected impact. Furthermore, in KLCC, the size of unfurnished houses has a considerable impact on cost. Market demand also changes, with furnished properties seeing more activity, emphasising the relevance of furniture in property appeal.

Additional Feature(s)

4.0 CONCLUSION

4.1 DISCUSSION ON OBJECTIVE FINDINGS

The report emphasizes KLCC's critical role in the real estate market, highlighting its broad and extensive portfolio of properties that appeal to a wide variety of potential buyers. Unlike the widespread impression that KLCC is only connected with high-end real estate, the research reveals its dominance over a wide range of property sizes and price points. This calls into question the concept that KLCC is only available in the most costly areas of Kuala Lumpur, instead promoting it as a dynamic and varied real estate market.

The relationship between furnishing status and property values adds dimension to the investigation. The substantial price disparities between unfurnished, fully furnished, and partially furnished residences highlight the importance of the furnishing component in determining property values. Particularly in KLCC, where pricing differentials exceed the previously estimated impact, this demonstrates the powerful influence of location on property costs. This insight stresses the complex interplay of elements such as location, furnishing status, and pricing dynamics, which helps to provide a more comprehensive picture of the KLCC real estate scene.

In conclusion, KLCC's importance goes beyond being seen solely as an exclusive and expensive destination. Its adaptability in supplying a varied variety of properties, as well as the observed impact of location and furnishing status on pricing, highlight KLCC's vital role in developing Kuala Lumpur's real estate sector.

4.2 RECOMMENDATIONS

Given the clear impact of furnishing status and location on property prices, stakeholders, particularly real estate developers and investors, are urged to improve their data analytic skills. Advanced analytical tools, such as predictive modeling and machine learning algorithms, may provide more detailed insights into the intricate interactions between the different elements that influence property values. Using data analytics technologies, stakeholders can discover complex patterns and trends in the dataset, allowing for more informed decision-making.

Furthermore, applying geospatial research could provide a spatial perspective on property values by considering proximity to major landmarks, public amenities, and infrastructure. This geospatial insight can help to provide a more complete view of market dynamics in KLCC and other places.

In conclusion, adopting advanced data analysis methodologies and investigating new dimensions of data collecting can enable stakeholders to get deeper insights, modify plans, and make more accurate predictions in the ever-changing real estate sector.

4.3 LIMITATIONS AND FUTURE DIRECTION

Future data analysis efforts should focus on identifying the specific factors that explain KLCC's supremacy in the real estate market. Exploring buyer preferences, as well as examining the impact of characteristics like as amenities, property age, and proximity to significant landmarks, can improve our understanding of market dynamics. The goal is to perform a thorough analysis of these elements to determine their impact on property values. To remain competitive in Kuala Lumpur's ever-changing real estate ecosystem, players must continuously evaluate and adapt their tactics to new market trends.

WORKLOAD MATRIX

Name	TP Number	Task(s)
Abdul Muhaimin Aman	TP069510	• Introduction • Data description • Setting up the document • Individual component (Analysis 3.1) • Conclusion
Ameera Farooq	TP071807	• Compilation of document • Individual component (Analysis 3.2) • Cleaning dataset • Hypothesis and objective framing
Aishath Ishran Shaifu	TP070329	• Data preparation • Cleaning dataset • Individual component (Analysis 3.3) • Hypothesis and objective framing