Sardar Patel Institute of Technology,Mumbai
Department of Computer Science and Engineering
B.E. Sem-VII- PE-IV (2024-2025)

UID	2022701010
Name	Hawaiza Siddiqui
Class and Batch	BE CSE DS - Batch K

Experiment no 5

Aim :
Create advanced charts using R programming language on the dataset - Housing data

Dataset:
https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market

R Script :-

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("plotly")
install.packages("RColorBrewer")
library(ggplot2) library(dplyr)
library(plotly)
library(RColorBrewer)

data <- read.csv("C:/Users/manth/Downloads/MELBOURNE_HOUSE_PRICES_LESS.csv")

summary(data) head(data)


# Word Cloud
```

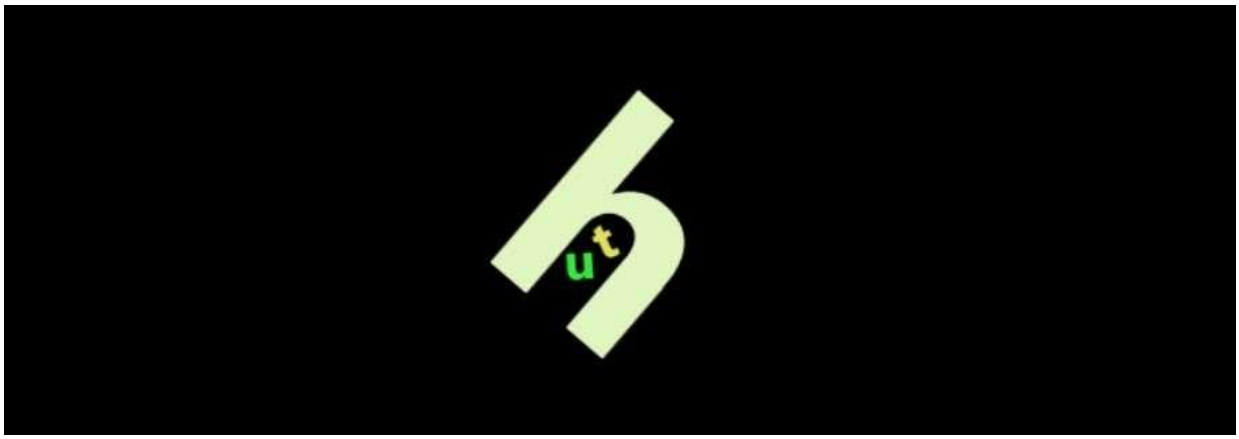
```
# Install and load wordcloud2 library
install.packages("wordcloud2") library(wordcloud2)

# Prepare data for the word cloud (example using 'Suburb') word_freq <- data
%>% count(Suburb)

# Generate word cloud wordcloud2(word_freq, size = 1,

color = 'random-dark') property_types <- c("h", "h", "h", "h",

"u", "t", "h", "h")

# Create word cloud freq_table <- data.frame(property_types =
names(table(property_types)), freq = as.numeric(table(property_types)))

# Create word cloud
wordcloud2(data = freq_table, size = 1, color = "random-light", backgroundColor = "black")

# Box and Whisker
# plot of Price vs Suburb
top_10_prices <- data %>%
arrange(desc(Price)) %>%
  slice(1:10) # Select the top 10 rows by Price

# Create a boxplot for the top 10 suburbs based on Price
ggplot(top_10_prices, aes(x = reorder(Suburb, Price), y = Price)) +
geom_boxplot(fill = "lightblue") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) + labs(title =
  "Box and Whisker Plot of Top 10 Prices by Suburb", x = "Suburb", y
  = "Price")


top_5_regions <- data %>%
  group_by(Regionname) %>% summarise(avg_price =
  mean(Price, na.rm = TRUE)) %>% arrange(desc(avg_price))
  %>% slice(1:5) %>% pull(Regionname)

# Filter the data to only include the top 5 regions top_5_data <-
data %>%
  filter(Regionname %in% top_5_regions)

# Create a boxplot for the top 5 regions by Price
ggplot(top_5_data, aes(x = Regionname, y = Price)) +
geom_boxplot(fill = "lightblue") +
  labs(title = "Boxplot of Price by Top 5 Regions", x = "Region", y = "Price")
```

```
# Violin plot of Price vs Regionname ggplot(data, aes(x = factor(Rooms), y
= Price, fill = factor(Rooms))) +
  geom_violin(trim = TRUE) + # trim = TRUE to show only relevant distribution labs(title = "Violin
  Plot of Property Prices by Number of Rooms", x = "Number of Rooms", y =
"Price") + theme_minimal() + # Apply minimal theme for a clean look
  scale_fill_brewer(palette = "Set1") # Apply a different color palette

ggplot(data, aes(x = Regionname, y = Price, fill = Regionname)) +
  geom_violin(trim = FALSE) + # trim = FALSE to show the full distribution labs(title = "Violin Plot of
  Property Prices by Region", x = "Region", y = "Price") + theme(axis.text.x = element_text(angle = 45,
  hjust = 1)) + # Rotate x-axis labels for readability scale_fill_brewer(palette = "Set3") # Add color
  palette for fill

# Linear regression
# plot between Rooms and Price ggplot(data,
aes(x = Rooms, y = Price)) + geom_point() +
  geom_smooth(method = "lm", col = "blue") + labs(title =
  "Linear Regression: Rooms vs Price")

# Non-linear regression plot (using LOESS)
ggplot(data, aes(x = Rooms, y = Price)) +
geom_point() +
  geom_smooth(method = "loess", col = "red") + labs(title = "Non-
  linear Regression (LOESS): Rooms vs Price") # Linear regression
  of Price vs Distance ggplot(data, aes(x = Distance, y = Price)) +
  geom_point(alpha = 0.6, color = "darkgreen") + # Add transparency to the points
  geom_smooth(method = "lm", se = FALSE, color = "red") + # Fit linear model labs(title =
  "Linear Regression of Price vs Distance", x = "Distance", y = "Price") + theme_minimal() #
  Use a minimal theme for better aesthetics



# Nonlinear regression plot: Price vs Distance
# Polynomial regression (degree 2) of Price vs Distance ggplot(data, aes(x
= Distance, y = Price)) +
  geom_point(alpha = 0.6, color = "purple") + # Add scatter plot points geom_smooth(method = "lm",
  formula = y ~ poly(x, 2), se = FALSE, color = "orange") + # Fit
quadratic regression
  labs(title = "Polynomial Regression (Degree 2) of Price vs Distance", x = "Distance", y =
"Price") + theme_light() # Apply a light theme for a clean
  look
```

```
# 3D scatter plot plot_ly(data, x = ~Rooms, y = ~Price, z = ~Distance, color = ~Regionname, type =
'scatter3d', mode = 'markers') %>% layout(scene = list(
    xaxis = list(title = 'Rooms'),
    yaxis = list(title = 'Price'),
    zaxis = list(title = 'Distance')
  ),
  title = "3D Scatter Plot of Rooms, Price, and Distance")


# Jitter plot for Rooms vs Price ggplot(data, aes(x = Rooms, y = Price)) +
geom_jitter(width = 0.2, height = 0.2, color = "blue", alpha = 0.5) +
labs(title = "Jitter Plot of Rooms vs Price")

top_10_suburbs <- data %>%
  group_by(Suburb) %>%
  summarize(MedianPrice = median(Price)) %>%
  top_n(10, MedianPrice) %>% pull(Suburb)

# Filter the original data for only these top 10 suburbs
filtered_data <- data %>%
  filter(Suburb %in% top_10_suburbs)

# Create the plot ggplot(filtered_data, aes(x = reorder(Suburb, Price, FUN = median), y
= Price)) +
  geom_jitter(aes(color = as.factor(Rooms)), width = 0.2, height = 0, size = 3) +
  labs(title = "Jitter Plot of Price vs Top 10 Suburbs", x = "Suburb", y = "Price", color
  = "Rooms") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Visualization -

Word Cloud -

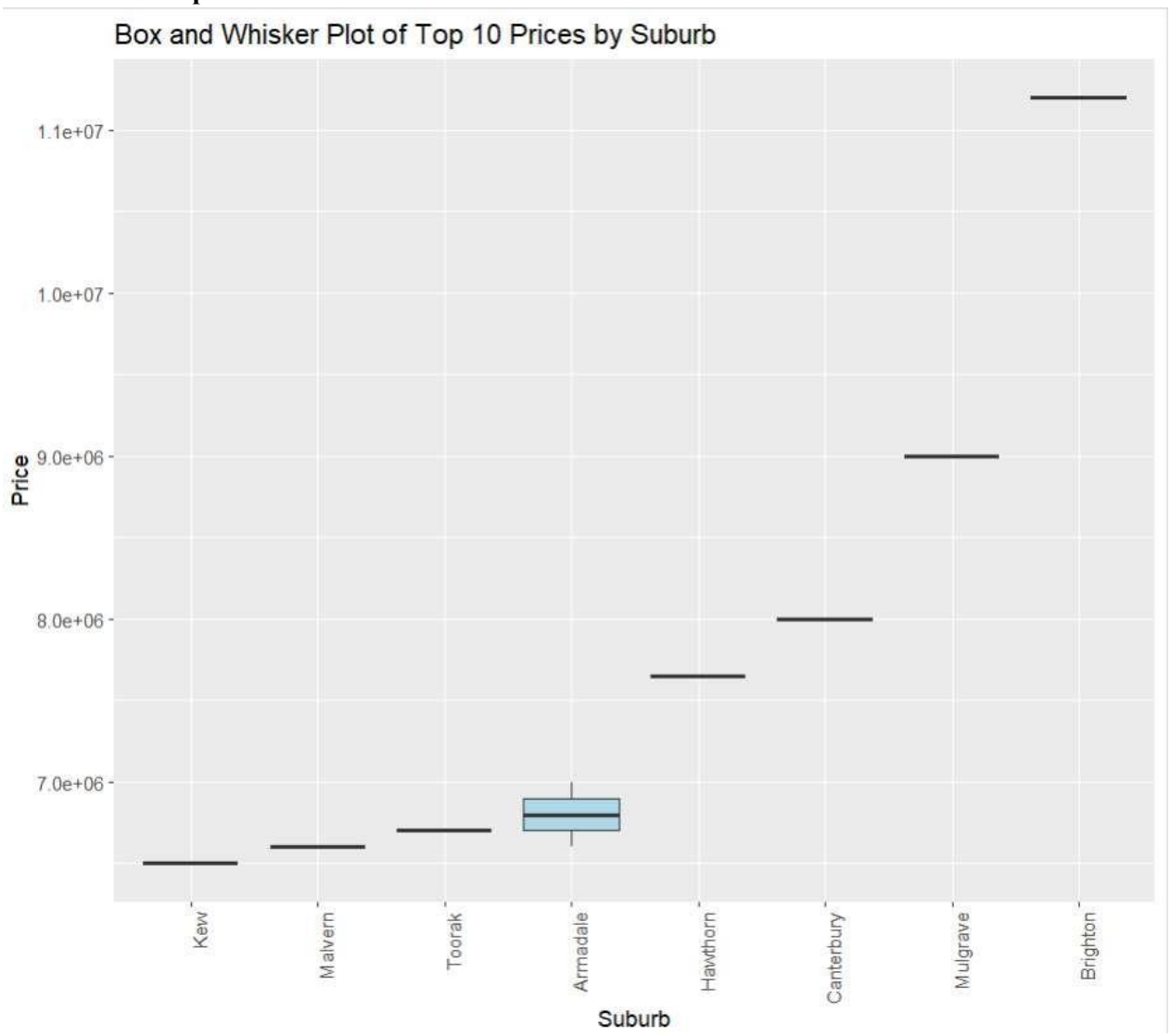


Observation :-

- Prominence of certain suburbs: The most prominent suburbs in the word cloud, indicated by their larger font sizes, appear to be:
    1. Ascot Vale
    2. Balwyn
    3. Blackburn
    4. Armadale
    5. Airport West
- Variety of suburbs: There's a wide range of suburbs represented, indicating a diverse dataset covering many areas.
- Geographic spread: The suburbs seem to cover various regions of Melbourne, including northern (e.g., Blackburn North), eastern (e.g., Box Hill), and western (e.g., Airport West) areas.
- Naming patterns: Many suburb names include directional indicators (e.g., Carlton North, Blackburn North) or descriptive terms (e.g., Meadow Heights, Avondale Heights).
- Color coding: The use of different colors doesn't seem to carry specific meaning but helps in visually distinguishing between different suburbs.

- Nested names: Some suburbs appear to be part of larger areas, like "Blackburn" and "Blackburn North" being separately represented.
- Less frequent suburbs: Smaller font sizes indicate suburbs that appear less frequently in the dataset, such as Jacana, Kingsville, and Aberfeldie.
- Compound names: Many suburb names are compound words (e.g., Ascot Vale, Clifton Hill), which is typical for Australian suburb naming conventions.

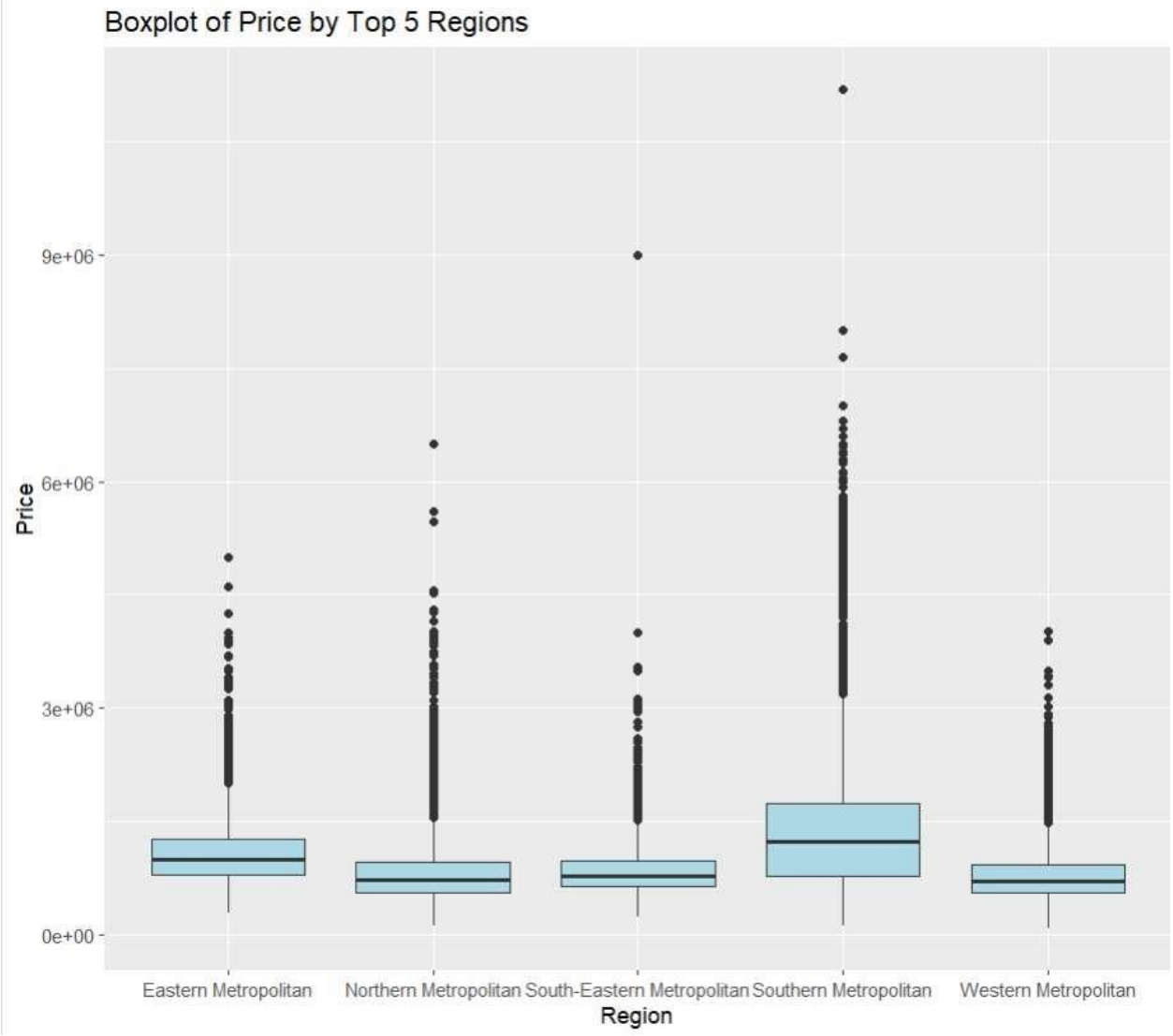
Box and whisker plot -



Observation :-

- Price Range: The prices range from about 6.5 million to over 11 million, indicating these are high-value properties.
- Top Suburb: Brighton appears to have the highest-priced property, with an outlier reaching above 11 million.
- Median Prices: The median prices (represented by the horizontal line in each box) vary across suburbs, with Brighton, Toorak, and Canterbury showing higher median prices compared to others.

- Price Spread: Some suburbs like Armadale show a larger spread of prices (larger box), indicating more variability in property values within that suburb.
- Outliers: Several suburbs, particularly Brighton, Toorak, and Canterbury, have notable outliers at the high end, representing exceptionally expensive properties.
- Clustering: There seems to be a cluster of suburbs with similar price ranges (Kew, Malvern, Toorak, Armadale) in the middle of the chart.



Observations:-

Price Range: The overall price range across all regions is approximately from 0 to 10 million dollars, with some outliers exceeding this range.

Median Prices:

- Southern Metropolitan appears to have the highest median price.
- Western Metropolitan seems to have the lowest median price among these top 5 regions.
- Eastern, Northern, and South-Eastern Metropolitan areas have similar median prices, falling between Southern and Western.

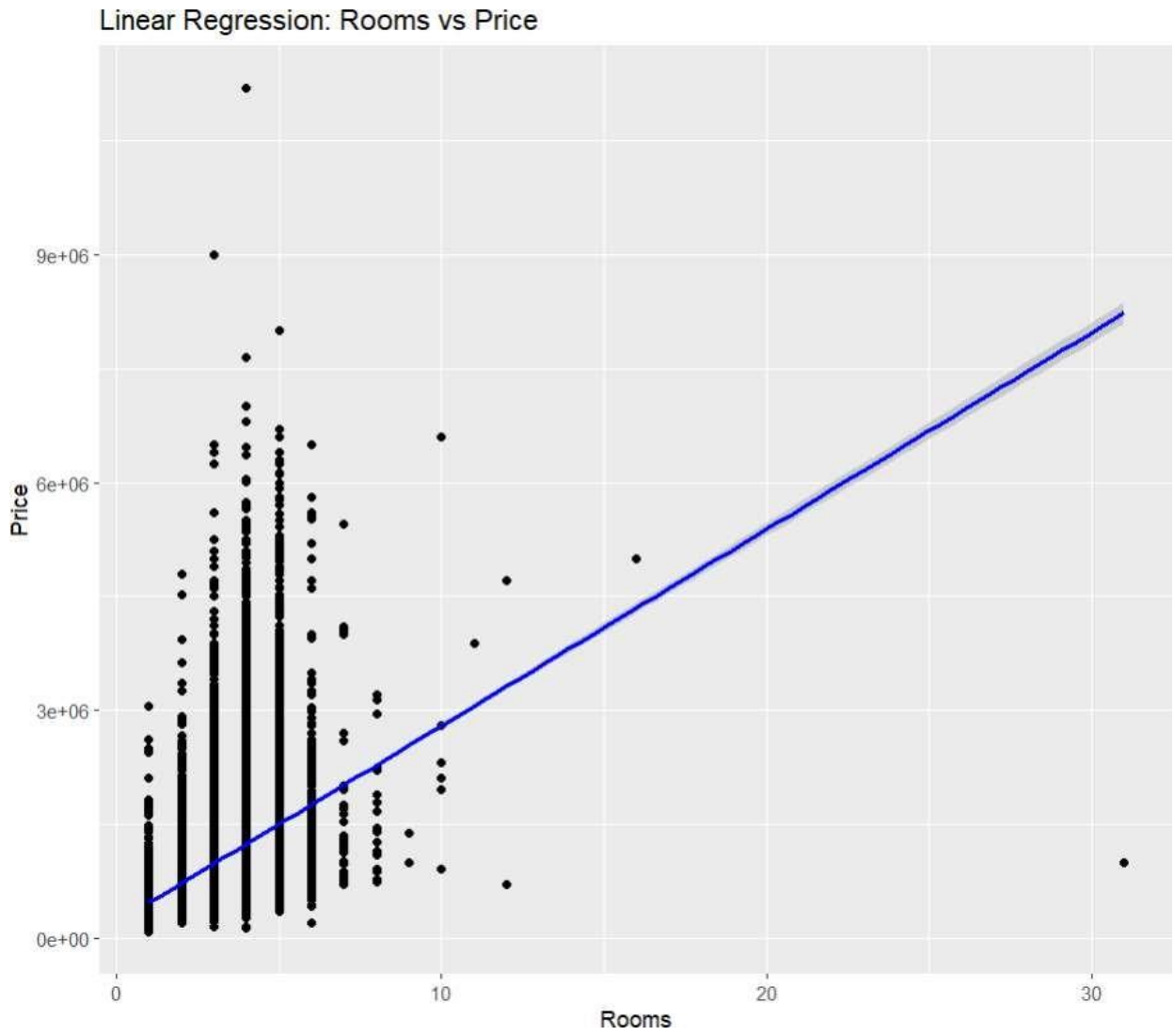
Price Spread:

- Southern Metropolitan shows the largest interquartile range (box size), indicating greater price variability.
- Western Metropolitan has the smallest interquartile range, suggesting more consistent pricing.

Outliers:

- All regions have numerous high-price outliers, represented by dots above the whiskers.
- Southern Metropolitan has the most extreme outliers, with some properties priced above 10 million dollars.
- Western Metropolitan has fewer high-value outliers compared to other regions.

Regression plot (linear and nonlinear) –



Observation :- **Overall**

Trend:

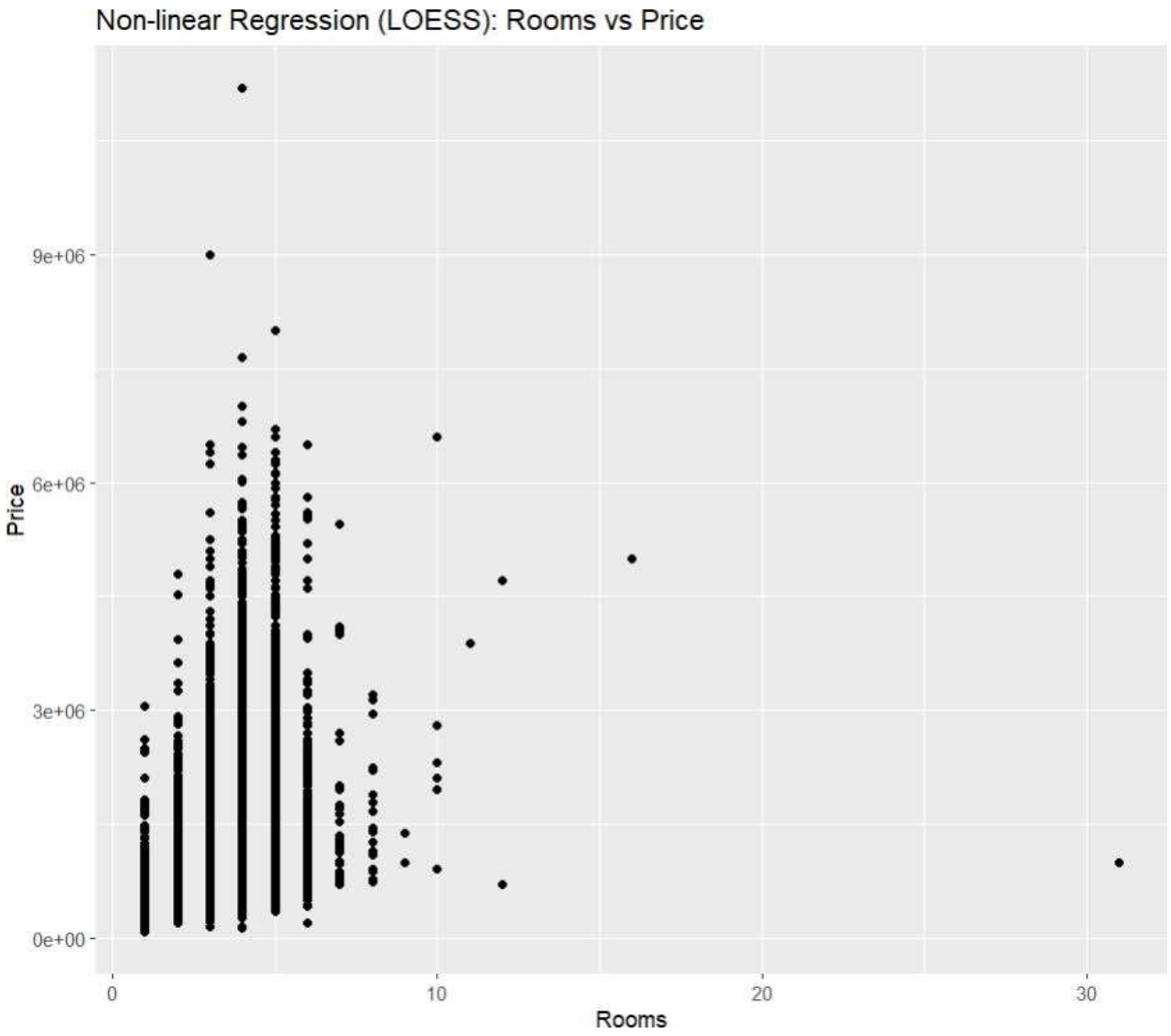
- There appears to be a positive linear relationship between the number of rooms and the price. This means that as the number of rooms in a property increases, the price tends to increase as well.

Scatter Plot:

- The scatter plot shows a clustering of data points around the lower end of the x-axis (number of rooms). This suggests that a majority of the properties analyzed have a relatively small number of rooms.
- There are a few outliers with a high number of rooms and price, which could be due to factors such as location, amenities, or unique features.

Regression Line:

- The blue regression line represents the best-fit linear relationship between the two variables. It shows the general trend of the data and can be used to make predictions.
- The slope of the line indicates the rate at which the price changes with respect to the number of rooms. A steeper slope would mean that price increases more rapidly with each additional room



Observation :- **Overall**

Trend:

- Similar to the linear regression plot, there appears to be a positive relationship between the number of rooms and the price. However, the LOESS curve suggests a non-linear relationship, indicating that the rate of increase in price with respect to the number of rooms is not constant.

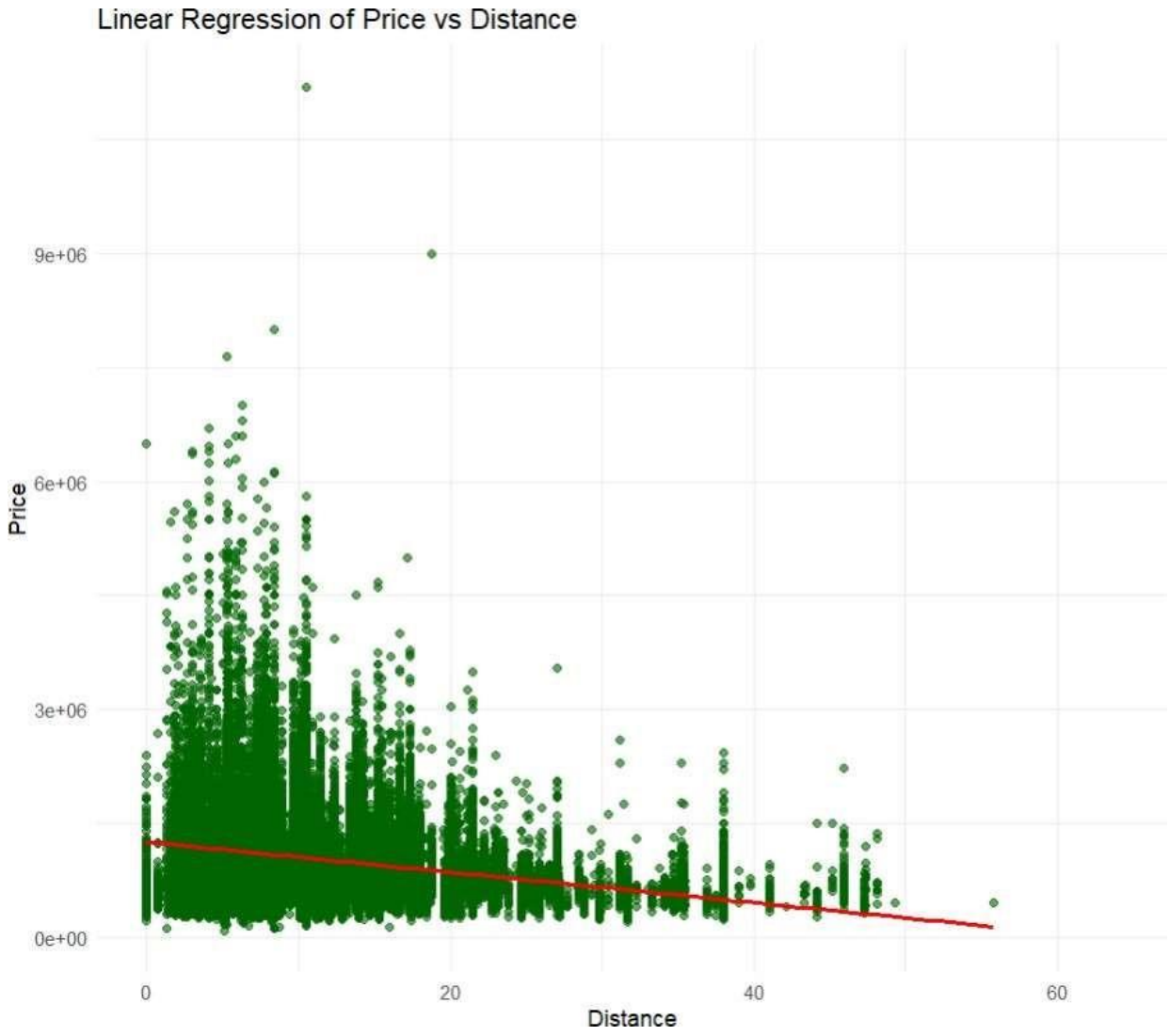
Scatter Plot:

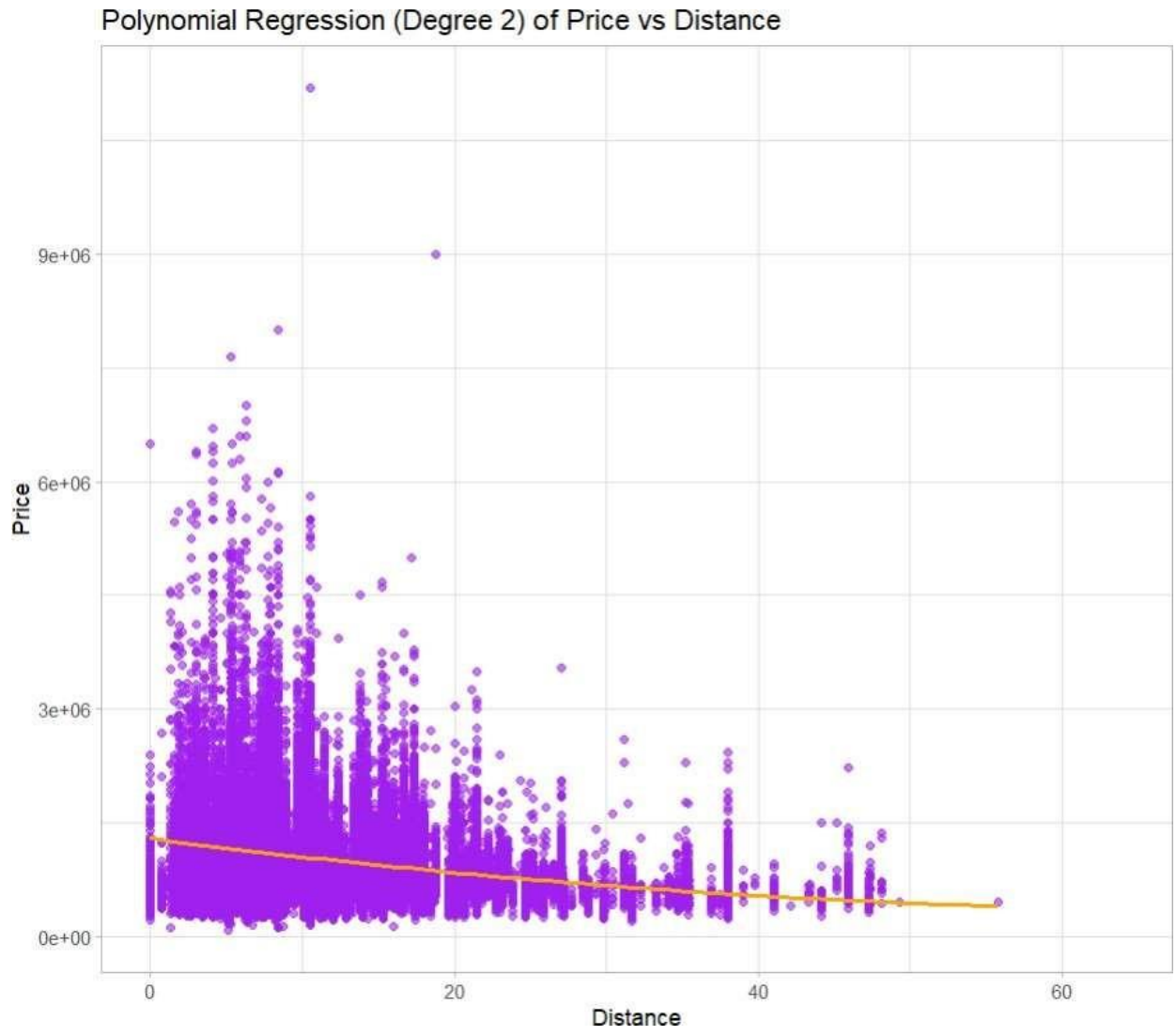
- The scatter plot remains unchanged, showing the same clustering of data points around the lower end of the x-axis and a few outliers.

LOESS Curve:

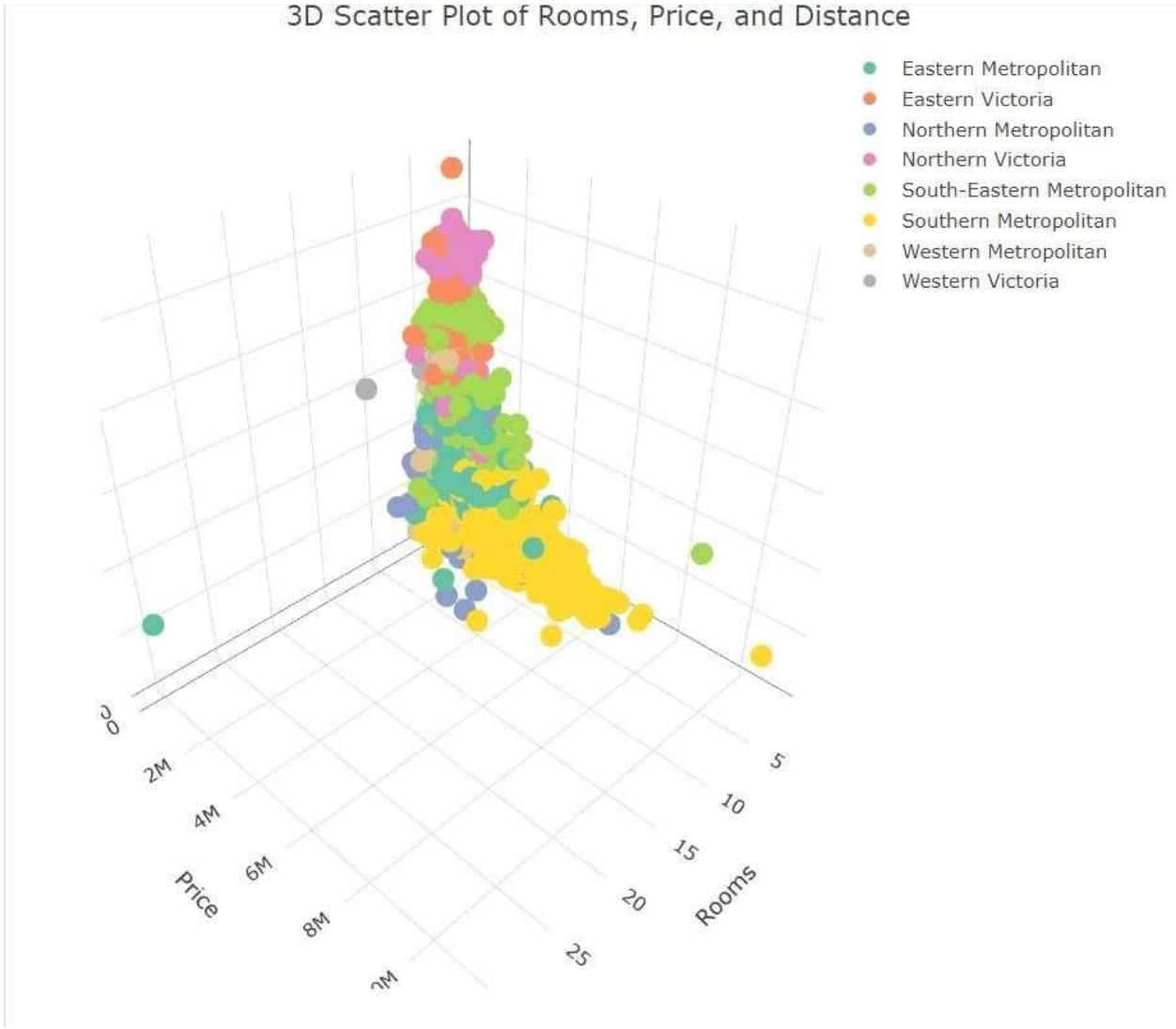
- The red LOESS curve captures the non-linear pattern in the data. It shows that the price increases at a faster rate for properties with fewer rooms compared to properties with a larger number of rooms.

- This suggests that there might be a diminishing returns effect, where the additional value of each extra room decreases as the total number of rooms increases.




Polynomial Regression (Degree 2) of Price vs Distance
Price
9e+06
6e+06
3e+06
0e+00
0
20
40
60
Distance

3D chart -



Observation :-

Overall Relationship:

- Te plot visually represents the relationship between the number of rooms, price, and distance for properties in different regions.
- There appears to be a general trend where properties with more rooms and higher prices tend to be located further away from the city center. However, there are also exceptions to this trend.

Regional Clusters:

- The different colors represent properties from different regions. This allows for visual identification of regional clusters and patterns.
- Some regions might have a higher concentration of properties in specific areas of the 3D space, indicating that factors like distance and room count might have a stronger influence on prices in those regions.

Outliers and Anomalies:

- The plot highlights outliers, which are properties that deviate significantly from the general trend. These could be due to factors such as unique features, historical significance, or other factors not captured in the data.

Specific Observations:

- Rooms vs. Price: There is a general positive correlation between the number of rooms and price, meaning that properties with more rooms tend to have higher prices. However, this relationship is not perfectly linear, and there are variations within each region.
- Distance vs. Price: The relationship between distance and price is more complex. Some properties located further away might have higher prices due to factors like proximity to natural amenities or transportation hubs. However, in other cases, properties located closer to the city center might have higher prices due to demand and accessibility.
- Regional Differences: The plot reveals that the relationship between these variables can vary across different regions. For example, properties in some regions might have a stronger correlation between rooms and price, while in others, distance might be a more significant factor.

Conclusion :- From this experiment, I learned to plot advance visualization like 3d chart , box and whisker plot , violin plot etc in r studio using r language.