Name: Omkar Shirwadkar
UID: 2021600062
Batch: D
Class: CSE(AI & ML)

EXPERIMENT NO: 5
Create advanced charts using R programming language on the dataset - Housing data

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

Dataset:
https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

Description:
The dataset is about houses and their pricing based on different features.

Timeline: This dataset is updated annually.

Attributes/Columns:
The dataset contains the following columns:

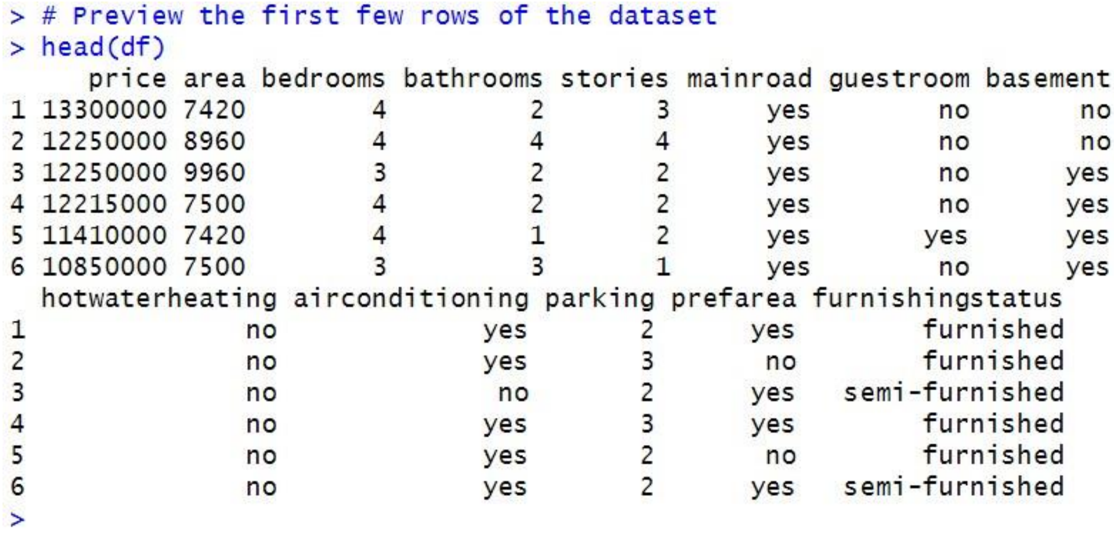
Price
Area
Bedrooms
Bathrooms
Stories
mainroad
Guestroom
Basement
Hotwaterheating
airconditioning
Parking
Prefarea
furnishingstatus

Code:

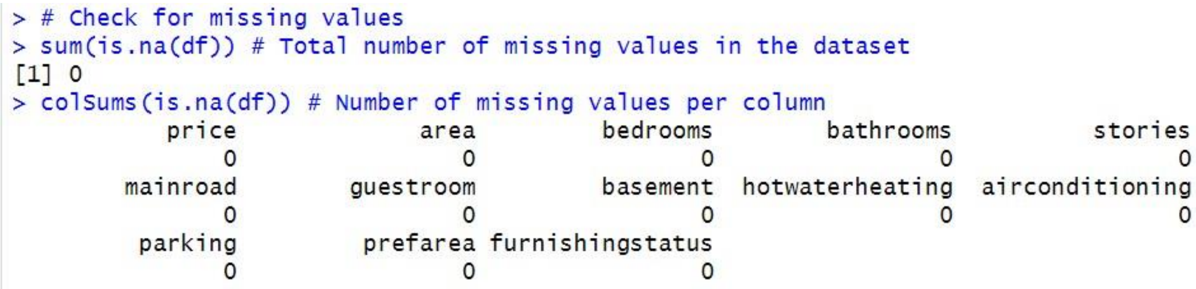
```
# Install required packages
install.packages("ggplot2")
install.packages("dplyr")
install.packages("plotly")
install.packages("ggpubr")
install.packages("GGally")
install.packages("tidyverse")
install.packages("car")
install.packages("wordcloud")

# Load the libraries library(ggplot2)
library(dplyr) library(plotly) library(ggpubr)
library(GGally) library(tidyverse)
library(car) df <-
read.csv("D:/adv/expt5/Housing.csv")
head(df)

sum(is.na(df))
colSums(is.na(df))
library(wordcloud)

# Create a word cloud for 'furnishingstatus' (for example) word_freq <-
table(df$furnishingstatus) wordcloud(words = names(word_freq), freq = as.vector(word_freq),
min.freq = 1, scale=c(3,0.5), colors=brewer.pal(8, "Dark2"))

# Boxplot for price by number of bedrooms
ggplot(df, aes(x=factor(bedrooms), y=price)) +
  geom_boxplot(fill="lightblue", color="darkblue") + labs(title="Boxplot of Price by
  Number of Bedrooms", x="Bedrooms", y="Price") + theme_minimal()

# Linear regression of price vs area
ggplot(df, aes(x=area, y=price)) +
  geom_point() + geom_smooth(method="lm", col="red") +
  labs(title="Linear Regression: Price vs Area", x="Area", y="Price")
  + theme_minimal()

# Non-linear regression of price vs area
ggplot(df, aes(x=area, y=price)) +
  geom_point() + geom_smooth(method="loess", col="blue") +
  labs(title="Nonlinear Regression: Price vs Area", x="Area", y="Price")
  + theme_minimal()
```

```
# Jitter plot for bedrooms vs price ggplot(df,
aes(x=factor(bedrooms), y=price)) +
  geom_jitter(width=0.2, color="purple", size=2) + labs(title="Jitter Plot:
  Bedrooms vs Price", x="Bedrooms", y="Price") + theme_minimal()
```

R Output:



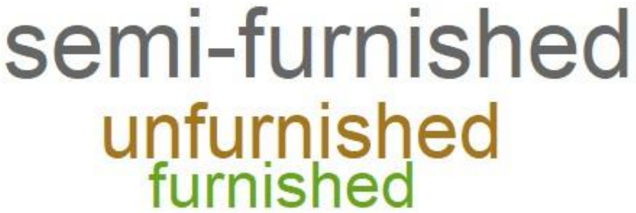
```
> # Preview the first few rows of the dataset
> head(df)
     price area bedrooms bathrooms stories mainroad guestroom basement
1 13300000 7420        4         2       3      yes        no       no
2 12250000 8960        4         4       4      yes        no       no
3 12250000 9960        3         2       2      yes        no      yes
4 12215000 7500        4         2       2      yes        no      yes
5 11410000 7420        4         1       2      yes       yes      yes
6 10850000 7500        3         3       1      yes        no      yes
  hotwaterheating airconditioning parking prefarea furnishingstatus
1              no             yes       2      yes        furnished
2              no             yes       3       no        furnished
3              no              no       2      yes   semi-furnished
4              no             yes       3      yes        furnished
5              no             yes       2       no        furnished
6              no             yes       2      yes   semi-furnished
>
```



```
> # Check for missing values
> sum(is.na(df)) # Total number of missing values in the dataset
[1] 0
> colSums(is.na(df)) # Number of missing values per column
           price             area         bedrooms        bathrooms          stories 
               0                0                0                0                0 
        mainroad        guestroom         basement  hotwaterheating  airconditioning 
               0                0                0                0                0 
         parking         prefarea furnishingstatus 
               0                0                0 
```

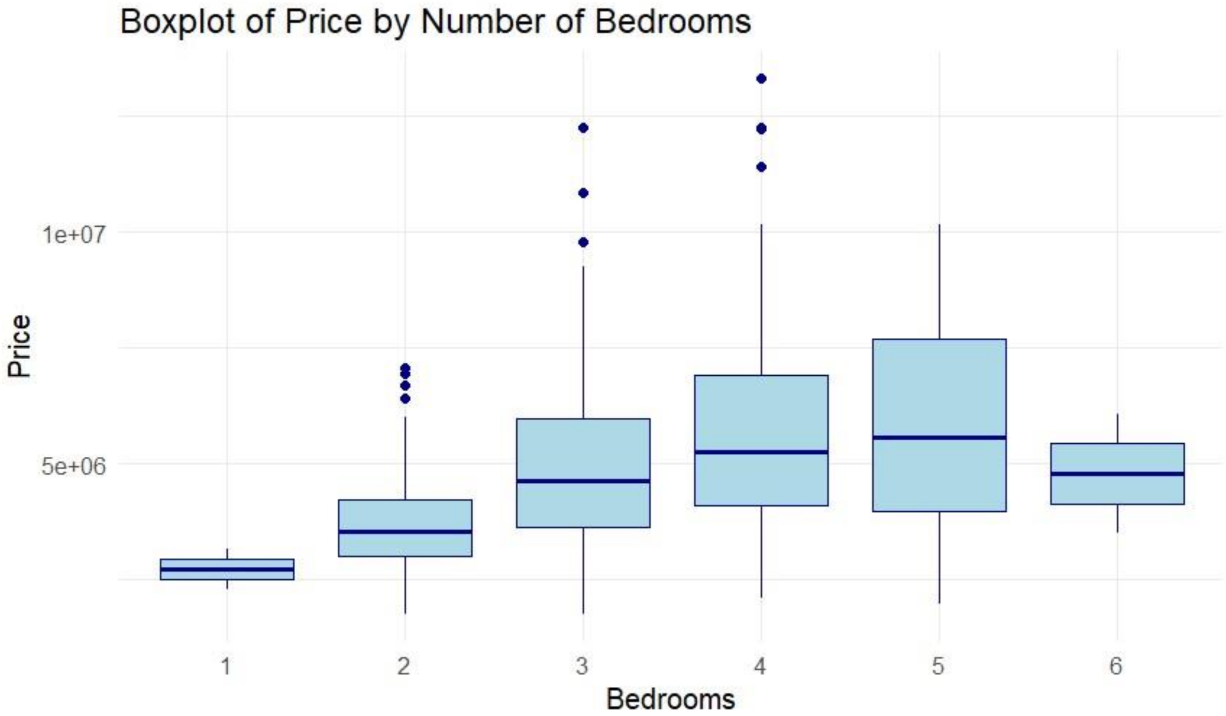
Plots:

1. **Word Cloud**



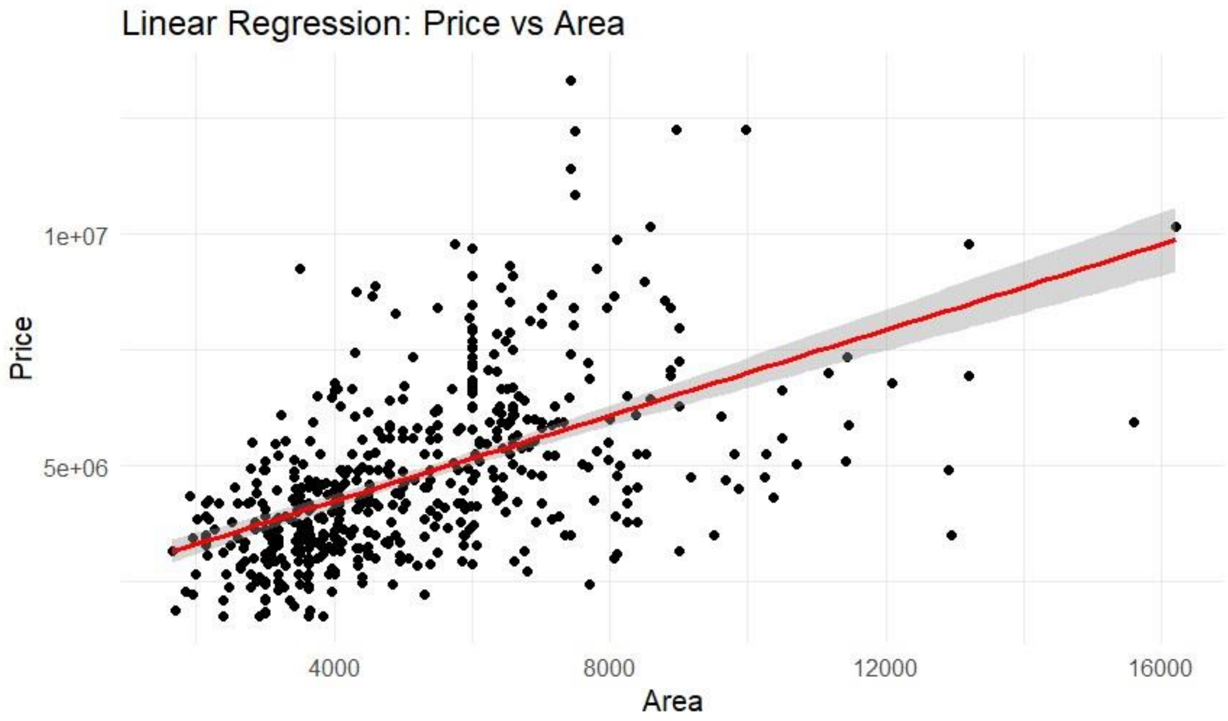
A word cloud visually represents the frequency of different categories in the furnishingstatus column of the dataset. The size of each word in the cloud corresponds to the frequency of that furnishing status in the data. Larger words indicate more frequent categories.

2. **Box Plot**



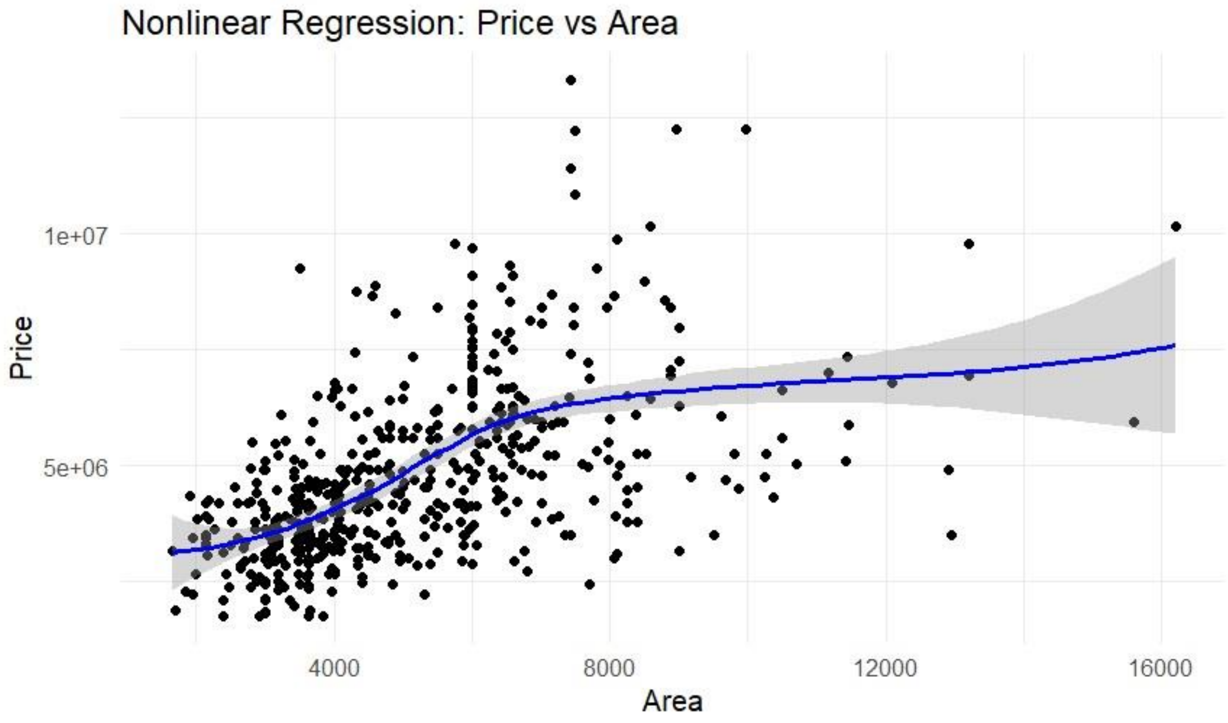
This boxplot shows the distribution of house prices across different numbers of bedrooms. Each box represents the interquartile range (IQR) of prices for a given number of bedrooms, with the line inside indicating the median price. The "whiskers" extend to capture the overall spread of prices, and any outliers are shown as individual points.

3. Linear Regression



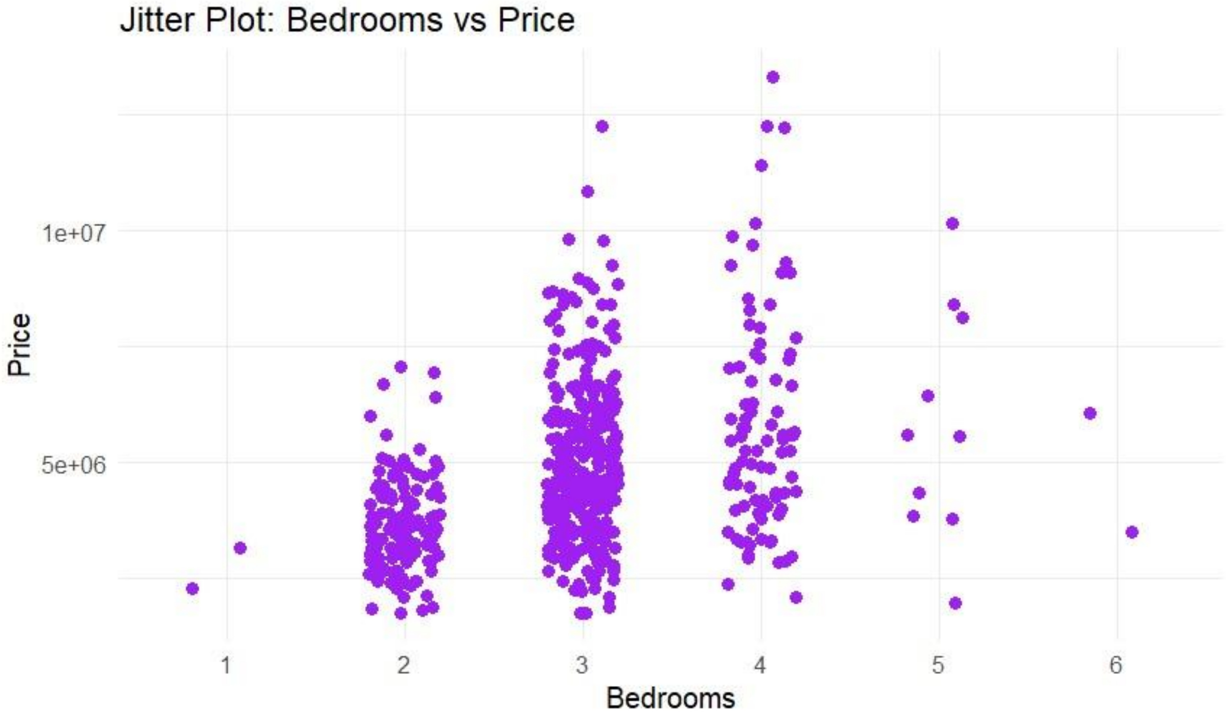
This linear regression plot shows the relationship between house prices and area. Each point represents an individual house, plotting its area against its price. The red line represents the best-fit linear regression line, indicating the overall trend in the data.

4. Non Linear Regression



This nonlinear regression plot depicts the relationship between house prices and area using a locally estimated scatterplot smoothing (LOESS) method. The blue curve represents the smoothed trend, capturing more complex patterns than a straight line.

5. **Jitter Plot**



This jitter plot displays the distribution of house prices across different numbers of bedrooms, with each point representing a house. This allows for a clearer visualization of the spread and concentration of prices within each bedroom category.

Conclusion:

In this exploration of the housing dataset, various visualizations were plotted. The linear and nonlinear regression plots revealed how area influences price. Jitter plot and word cloud offered clarity on categorical data distributions.