DS311 - R Lab Assignment

Ana Paula Felix de Queiroz

2023-04-14

## R Assignment 1

* In this assignment, we are going to apply some of the build in data set in R for descriptive statistics analysis.
* To earn full grade in this assignment, students need to complete the coding tasks for each question to get the result.
* After finished all the questions, knit the document into HTML format for submission.

### Question 1

Using the **mtcars** data set in R, please answer the following questions.

# Loading the data  
data(mtcars)  
  
# Head of the data set  
head(mtcars)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

1. Report the number of variables and observations in the data set.

# Enter your code here!  
dim(mtcars)

## [1] 32 11

# Answer:  
print("There are total of 11 variables and 32 observations in this data set.")

## [1] "There are total of 11 variables and 32 observations in this data set."

1. Print the summary statistics of the data set and report how many discrete and continuous variables are in the data set.

# Enter your code here!  
  
summary(mtcars)

## mpg cyl disp hp   
## Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0   
## 1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5   
## Median :19.20 Median :6.000 Median :196.3 Median :123.0   
## Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7   
## 3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0   
## Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0   
## drat wt qsec vs   
## Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000   
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000   
## Median :3.695 Median :3.325 Median :17.71 Median :0.0000   
## Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375   
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000   
## Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000   
## am gear carb   
## Min. :0.0000 Min. :3.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000   
## Median :0.0000 Median :4.000 Median :2.000   
## Mean :0.4062 Mean :3.688 Mean :2.812   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :1.0000 Max. :5.000 Max. :8.000

continuous <- sum(sapply(mtcars, is.double))  
discrete <- ncol(mtcars) - continuous  
continuous

## [1] 11

discrete

## [1] 0

# Answer:  
print("There are 0 discrete variables and 11 continuous variables in this data set.")

## [1] "There are 0 discrete variables and 11 continuous variables in this data set."

1. Calculate the mean, variance, and standard deviation for the variable **mpg** and assign them into variable names m, v, and s. Report the results in the print statement.

# Enter your code here!  
attach(mtcars)  
m=mean(mpg)  
v=sd(mpg)^2  
s = sd(mpg)  
  
print(paste("The average of Mile Per Gallon from this data set is ",m , " with variance ",v , " and standard deviation", s , "."))

## [1] "The average of Mile Per Gallon from this data set is 20.090625 with variance 36.3241028225806 and standard deviation 6.0269480520891 ."

1. Create two tables to summarize 1) average mpg for each cylinder class and 2) the standard deviation of mpg for each gear class.

# Enter your code here!  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# Create a table of average mpg for each cylinder class  
mpg\_by\_cyl <- mtcars %>%   
 group\_by(cyl) %>%   
 summarise(avg\_mpg = mean(mpg))  
  
mpg\_by\_cyl

## # A tibble: 3 × 2  
## cyl avg\_mpg  
## <dbl> <dbl>  
## 1 4 26.7  
## 2 6 19.7  
## 3 8 15.1

# Create a table of standard deviation of mpg for each gear class  
mpg\_by\_gear <- mtcars %>%   
 group\_by(gear) %>%   
 summarise(sd\_mpg = sd(mpg))  
  
# Print the table  
print(mpg\_by\_gear)

## # A tibble: 3 × 2  
## gear sd\_mpg  
## <dbl> <dbl>  
## 1 3 3.37  
## 2 4 5.28  
## 3 5 6.66

1. Create a crosstab that shows the number of observations belong to each cylinder and gear class combinations. The table should show how many observations given the car has 4 cylinders with 3 gears, 4 cylinders with 4 gears, etc. Report which combination is recorded in this data set and how many observations for this type of car.

# Enter your code here!  
crosstab = table(mtcars$cyl,mtcars$gear,  
 dnn = c("Cyl","gears"))  
crosstab

## gears  
## Cyl 3 4 5  
## 4 1 8 2  
## 6 2 4 1  
## 8 12 0 2

print("The most common car type in this data set is car with 8 cylinders and 3 gears. There are total of 12 cars belong to this specification in the data set.")

## [1] "The most common car type in this data set is car with 8 cylinders and 3 gears. There are total of 12 cars belong to this specification in the data set."

### Question 2

Use different visualization tools to summarize the data sets in this question.

1. Using the **PlantGrowth** data set, visualize and compare the weight of the plant in the three separated group. Give labels to the title, x-axis, and y-axis on the graph. Write a paragraph to summarize your findings.

# Load the data set  
data("PlantGrowth")  
  
# Head of the data set  
head(PlantGrowth)

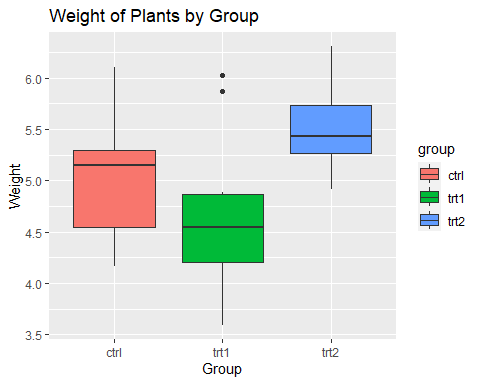
## weight group  
## 1 4.17 ctrl  
## 2 5.58 ctrl  
## 3 5.18 ctrl  
## 4 6.11 ctrl  
## 5 4.50 ctrl  
## 6 4.61 ctrl

# Enter your code here!  
  
library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'mtcars':  
##   
## mpg

ggplot(PlantGrowth, aes(x=group, y=weight, fill=group)) +   
 geom\_boxplot() + labs(title = "Weight of Plants by Group", x = "Group", y = "Weight")



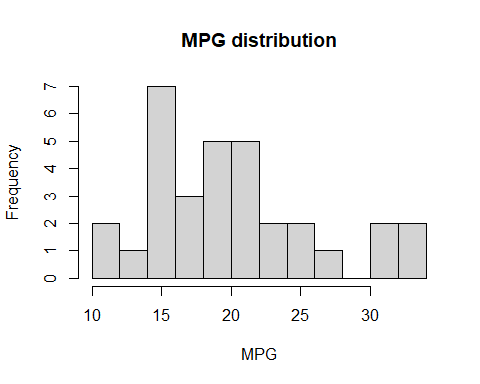
Result:

=> Report a paragraph to summarize your findings from the plot!

The box plot shows that weights of plants for group TRT2 is higher than the other two with a median at 5.4. While the group Trt1 has the lower median around 4.6. Group CTRL seems to have the largest spread of the 50% of the data. The distribution of group ctrl seem to be skewed to the write while group trt2 seem to be skewed to the left.We are also able to identify that trt1 has outliers.

1. Using the **mtcars** data set, plot the histogram for the column **mpg** with 10 breaks. Give labels to the title, x-axis, and y-axis on the graph. Report the most observed mpg class from the data set.

hist(mtcars$mpg, breaks = 10,  
 main = "MPG distribution",  
 xlab = "MPG",  
 ylab = "Frequency")



print("Most of the cars in this data set are in the class of 10 mile per gallon.")

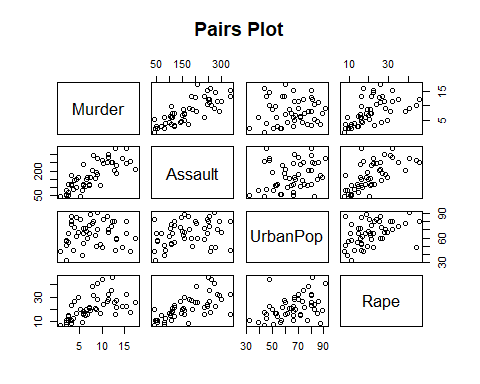
## [1] "Most of the cars in this data set are in the class of 10 mile per gallon."

1. Using the **USArrests** data set, create a pairs plot to display the correlations between the variables in the data set. Plot the scatter plot with **Murder** and **Assault**. Give labels to the title, x-axis, and y-axis on the graph. Write a paragraph to summarize your results from both plots.

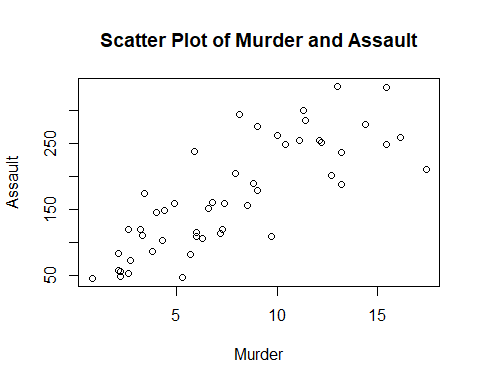
# Load the data set  
data("USArrests")  
  
# Head of the data set  
head(USArrests)

## Murder Assault UrbanPop Rape  
## Alabama 13.2 236 58 21.2  
## Alaska 10.0 263 48 44.5  
## Arizona 8.1 294 80 31.0  
## Arkansas 8.8 190 50 19.5  
## California 9.0 276 91 40.6  
## Colorado 7.9 204 78 38.7

# Enter your code here!  
  
  
pairs(USArrests, main = "Pairs Plot")



plot(USArrests$Murder, USArrests$Assault,  
 main = "Scatter Plot of Murder and Assault",  
 xlab = "Murder", ylab = "Assault")



Result:

=> Report a paragraph to summarize your findings from the plot! It seems like murder and Assault have the highest positive correlation. We could say that there is a moderate positive correlation between Rape and murder as well. It seems like Urban Population is the variable with the lowest correlation with Assault and murder.

### Question 3

Download the housing data set from www.jaredlander.com and find out what explains the housing prices in New York City.

Note: Check your working directory to make sure that you can download the data into the data folder.

## Warning in dir.create("data"): 'data' already exists

1. Create your own descriptive statistics and aggregation tables to summarize the data set and find any meaningful results between different variables in the data set.

# Head of the cleaned data set  
head(housingData)

## Neighborhood Market.Value.per.SqFt Boro Year.Built  
## 1 FINANCIAL 200.00 Manhattan 1920  
## 2 FINANCIAL 242.76 Manhattan 1985  
## 4 FINANCIAL 271.23 Manhattan 1930  
## 5 TRIBECA 247.48 Manhattan 1985  
## 6 TRIBECA 191.37 Manhattan 1986  
## 7 TRIBECA 211.53 Manhattan 1985

unique(housingData$Boro)

## [1] "Manhattan" "Brooklyn" "Queens" "Bronx"   
## [5] "Staten Island"

summary(housingData)

## Neighborhood Market.Value.per.SqFt Boro Year.Built   
## Length:2530 Min. : 10.66 Length:2530 Min. :1825   
## Class :character 1st Qu.: 75.10 Class :character 1st Qu.:1926   
## Mode :character Median :114.89 Mode :character Median :1986   
## Mean :133.17 Mean :1967   
## 3rd Qu.:189.91 3rd Qu.:2005   
## Max. :399.38 Max. :2010

# Enter your code here!  
#table discribing mean and median grouped by value by var Boro  
library(dplyr)  
housingData %>%   
 group\_by(Boro) %>%   
 summarize(mean\_value = mean(Market.Value.per.SqFt),   
 median\_value = median(Market.Value.per.SqFt))

## # A tibble: 5 × 3  
## Boro mean\_value median\_value  
## <chr> <dbl> <dbl>  
## 1 Bronx 47.9 47.4  
## 2 Brooklyn 80.1 81.6  
## 3 Manhattan 181. 184.   
## 4 Queens 77.4 66.9  
## 5 Staten Island 41.3 41.0

housingData %>%  
 group\_by(Neighborhood) %>%  
 summarize(mean\_value = mean(Market.Value.per.SqFt),  
 median\_value = median(Market.Value.per.SqFt)) %>%  
 arrange(desc(mean\_value))

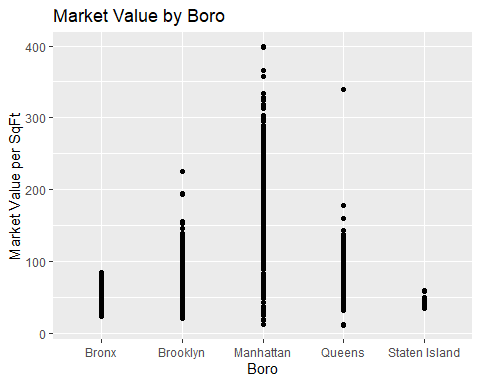
## # A tibble: 148 × 3  
## Neighborhood mean\_value median\_value  
## <chr> <dbl> <dbl>  
## 1 MIDTOWN CBD 234. 227.  
## 2 FLATIRON 223. 230.  
## 3 MIDTOWN WEST 222. 223.  
## 4 UPPER EAST SIDE (59-79) 217. 218.  
## 5 CHELSEA 216. 214.  
## 6 MIDTOWN EAST 211. 220.  
## 7 EAST VILLAGE 207. 200.  
## 8 MURRAY HILL 206. 209.  
## 9 UPPER EAST SIDE (79-96) 202. 210.  
## 10 GREENWICH VILLAGE-WEST 202. 214.  
## # … with 138 more rows

summary(housingData$Market.Value.per.SqFt)

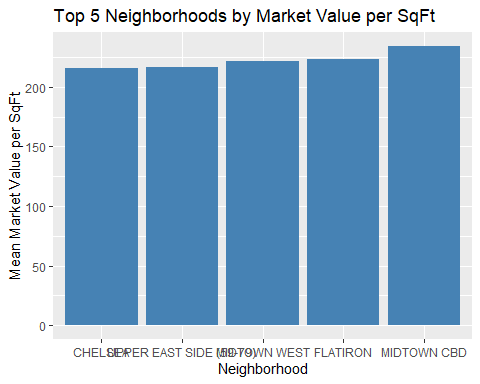
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.66 75.10 114.89 133.17 189.91 399.38

1. Create multiple plots to demonstrates the correlations between different variables. Remember to label all axes and give title to each graph.

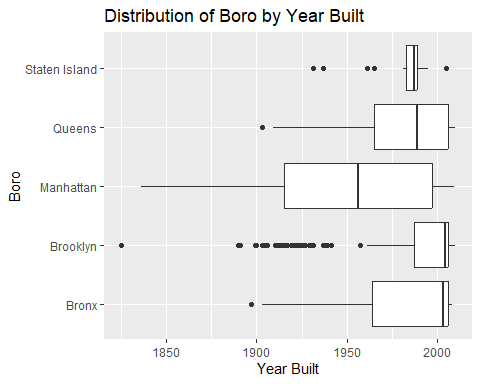
# Enter your code here!  
library(ggplot2)  
ggplot(housingData, aes(x = Boro, y = Market.Value.per.SqFt)) +  
 geom\_point() +  
 labs(title = "Market Value by Boro",  
 x = "Boro",  
 y = "Market Value per SqFt")



top10\_neighborhoods <- housingData %>%  
 group\_by(Neighborhood) %>%  
 summarize(mean\_value = mean(Market.Value.per.SqFt)) %>%  
 arrange(desc(mean\_value)) %>%  
 top\_n(5, mean\_value)  
  
  
ggplot(top10\_neighborhoods, aes(x = reorder(Neighborhood, mean\_value), y = mean\_value)) +  
 geom\_bar(stat = "identity", fill = "steelblue") +  
 labs(title = "Top 5 Neighborhoods by Market Value per SqFt",  
 x = "Neighborhood", y = "Mean Market Value per SqFt")



ggplot(housingData, aes(x = Year.Built, y = Boro)) +  
 geom\_boxplot() +  
 labs(title = "Distribution of Boro by Year Built",  
 x = "Year Built",  
 y = "Boro")



1. Write a summary about your findings from this exercise.

If we look at the variables Boro and market value we can see that Mahattan has the highest market value per square ft, queens and brooklyn seem to have a similar Market price exept for outliers. its followed by Bronx and then Staten Island.

Among all the Neighboors Midtown, Flatiron, Midtown West , uper East Side and Chelsea hank on the top highest mean for Market value per square ft

We see that broklyn has one observation as the oldest building, and manhatan has the largest distribution of buildings age with the median around 1960.