Assignment 2

Ariz Kazani

2024-07-15

Assignment 2

Name: Ariz Kazani

Student ID: 101311311

Notes

```
# Libraries

# NOTE: if you do not have any of the below libraries installed, un-comment the line and run it
#install.packages("rmarkdown")
library(rmarkdown)

#install.packages("ggplot2")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.1
```

```
#install.packages("patchwork")
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.4.1
```

```
#install.packages("dplyr")
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.1

##
## 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
```

```r
# install.packages("tidyr")
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.4.1
```

```r
# TODO:
# - Format ctrl + shift + a
# Make sure all questions are completed
# double check questions
# install.packages("ggplot2")
```

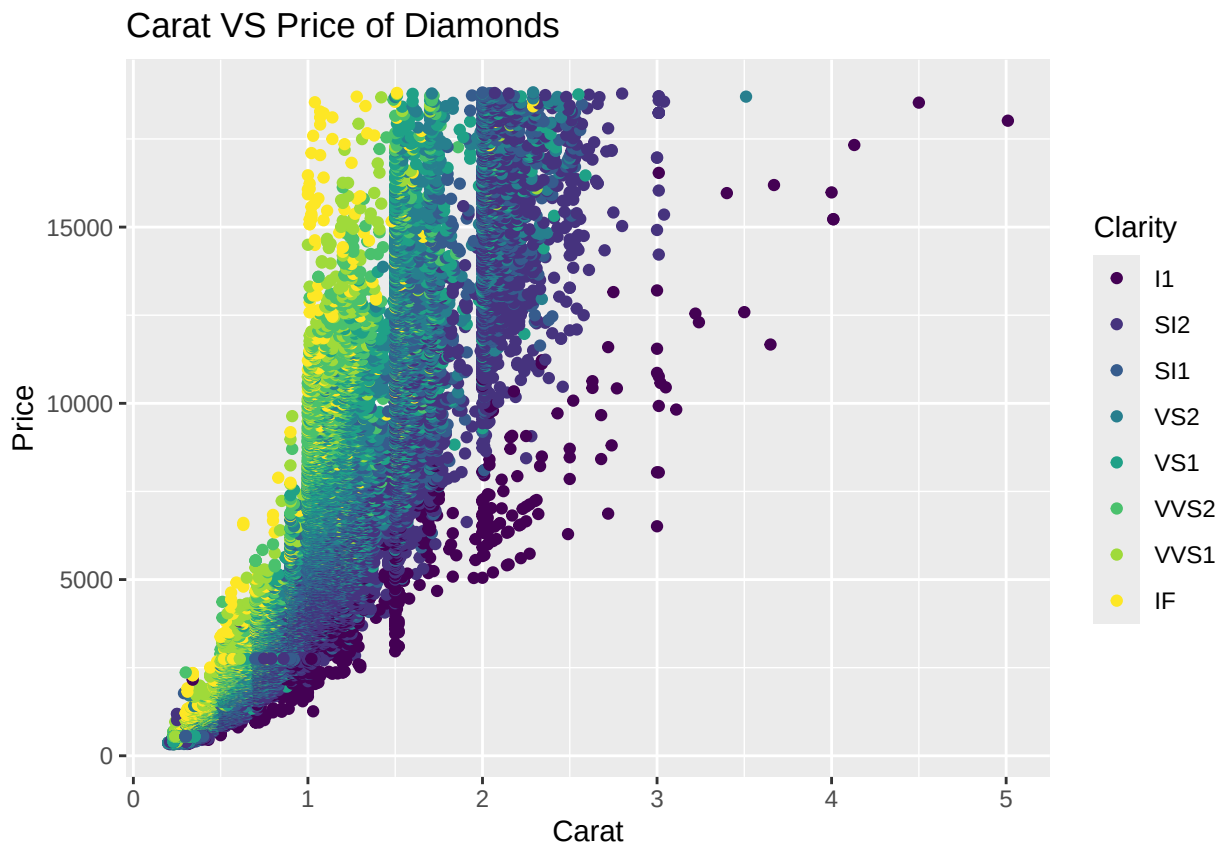
Solutions

---

1. Advanced ggplot2 Visualizations

A. Load the diamonds dataset from the ggplot2 package. Create a scatter plot of carat vs price with points colored by clarity.

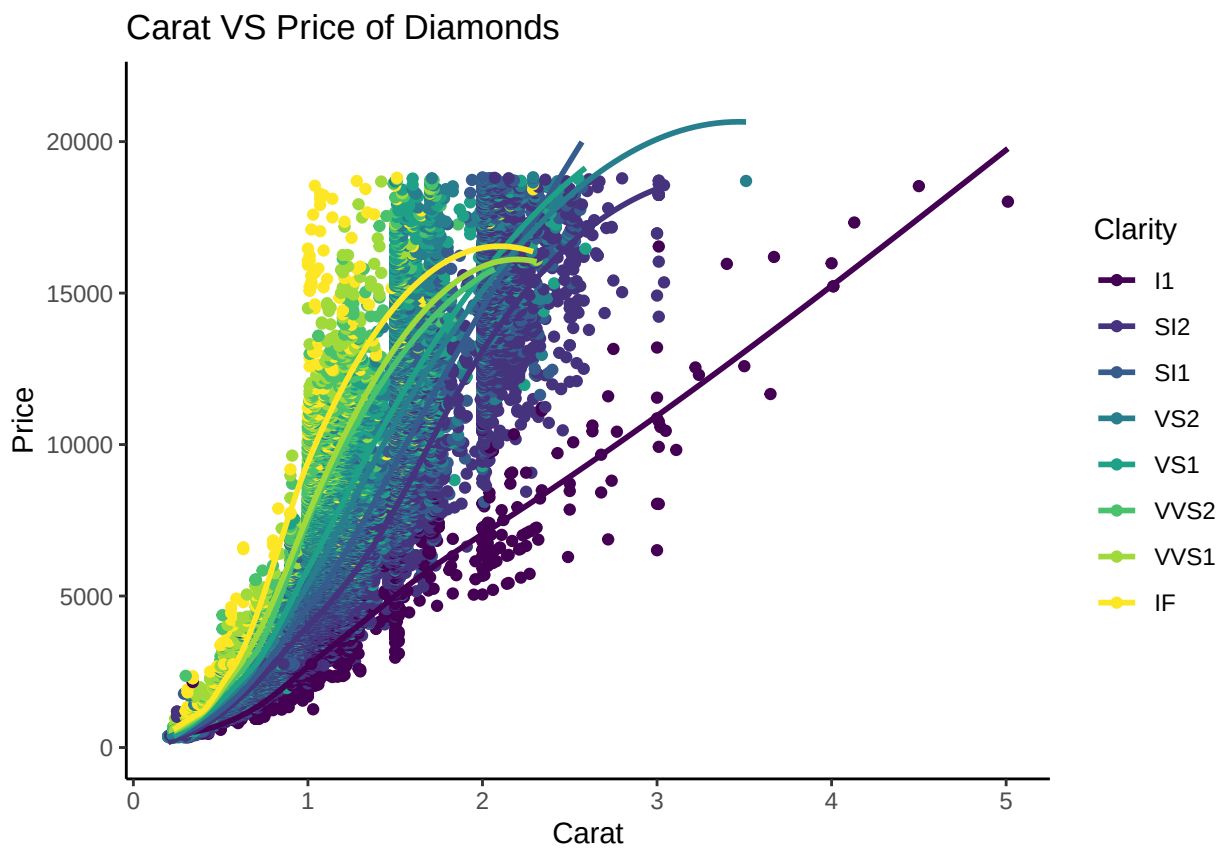
```r
data("diamonds")
scatPlot <- ggplot(diamonds, aes(x = carat, y = price, colour = clarity)) +
  geom_point() +
  labs(
    x = "Carat",
    y = "Price",
    title = "Carat VS Price of Diamonds",
    colour = "Clarity"
  )

scatPlot
```



B. Modify the scatter plot to include a smoothing line (e.g., LOESS) and customize the theme for better readability.

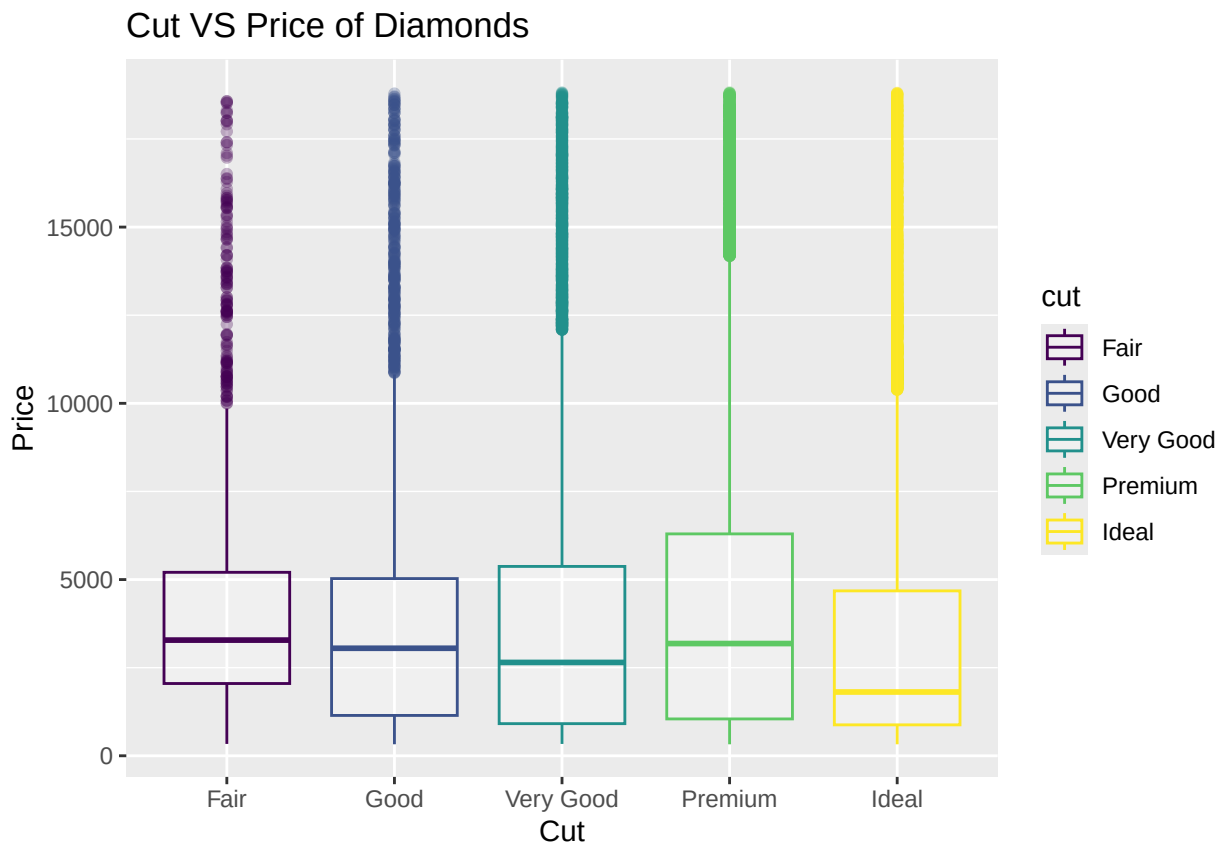
```
scatPlot <- scatPlot +
  geom_smooth(fill = NA, method = "loess") +
  theme_classic()

scatPlot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



C. Create a boxplot of price by cut, with different fill colors for each cut.

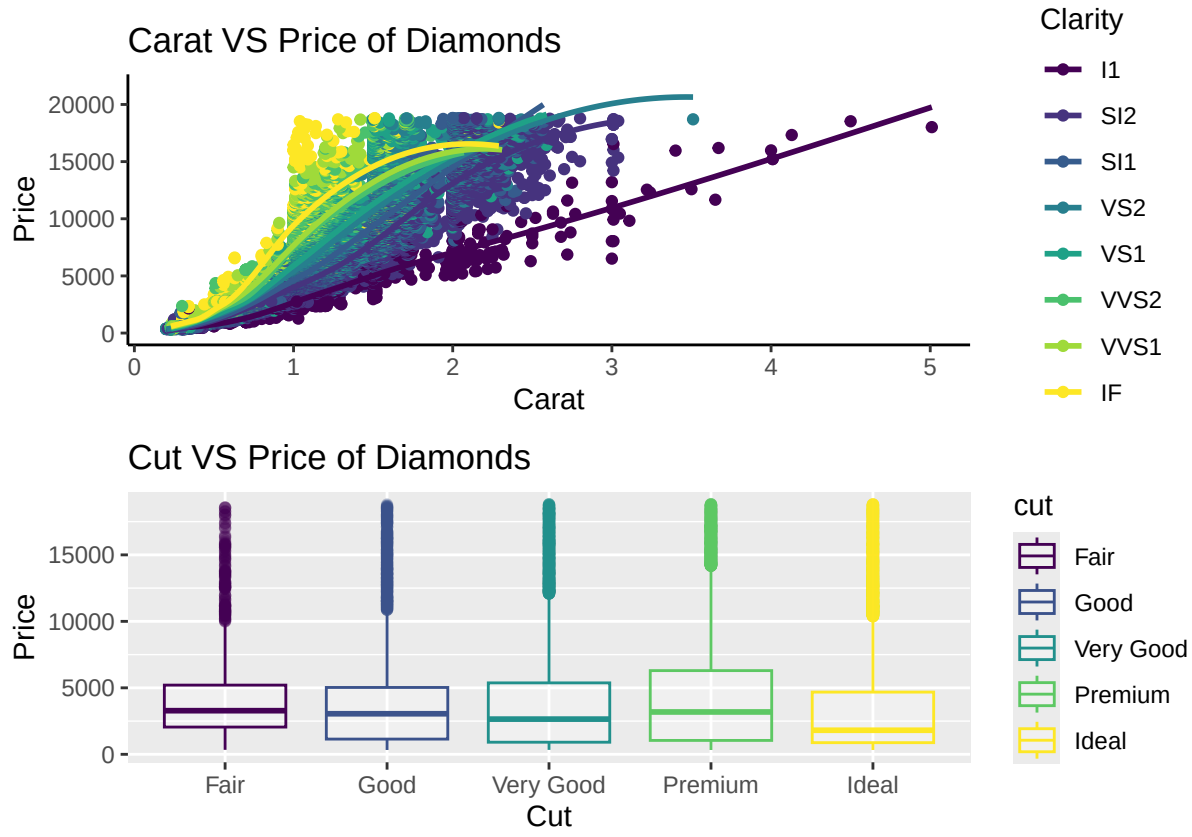
```
boxPlot <- ggplot(diamonds, aes(x = cut, y = price, colour = cut)) +
  geom_boxplot(alpha = 0.3) +
  labs(x = "Cut", y = "Price", title = "Cut VS Price of Diamonds")

boxPlot
```



D. Combine the scatter plot and boxplot into a single visualization using patchwork.

```
combinedPlot <- scatPlot / boxPlot
combinedPlot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



---

2. Advanced Group Manipulations

A. Load the mtcars dataset. Group the data by the number of cylinders and calculate the mean mpg for each group.

```
data("mtcars")

mtcarsDS <- mtcars

meanMpg <- tapply(mtcarsDS$mpg, mtcarsDS$cyl, mean)

meanMpg
```

```
##        4        6        8
## 26.66364 19.74286 15.10000
```

B. Add a column to the original dataset indicating whether each car's mpg is above or below the mean mpg of its cylinder group.

```
mtcarsDS$posMean <- ifelse(mtcarsDS$mpg > meanMpg[as.character(mtcarsDS$cyl)],
                           "above",
                           ifelse(mtcarsDS$mpg < meanMpg[as.character(mtcarsDS$cyl)], "below", "same"))

head(mtcarsDS)
```

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

C. Create a summary table showing the mean and median hp and wt for each combination of cyl and gear.

```
summaryTable <- mtcars %>%
  group_by(cyl, gear) %>%
  summarise(
    meanHp = mean(hp),
    medianHp = median(hp),
    meanWt = mean(wt),
    medianWt = median(wt),
  )
```

```
## `summarise()` has grouped output by 'cyl'. You can override using the `.groups`
## argument.
```

```
summaryTable
```

```
## # A tibble: 8 x 6
## # Groups:   cyl [3]
##     cyl  gear meanHp medianHp meanWt medianWt
##   <dbl> <dbl>  <dbl>    <dbl>  <dbl>    <dbl>
## 1     4     3    97       97    2.46     2.46
## 2     4     4    76       66    2.38     2.26
## 3     4     5   102      102    1.83     1.83
## 4     6     3   108.     108.   3.34     3.34
## 5     6     4   116.     116.   3.09     3.16
## 6     6     5   175      175    2.77     2.77
## 7     8     3   194.     180    4.10     3.81
## 8     8     5   300.     300.   3.37     3.37
```

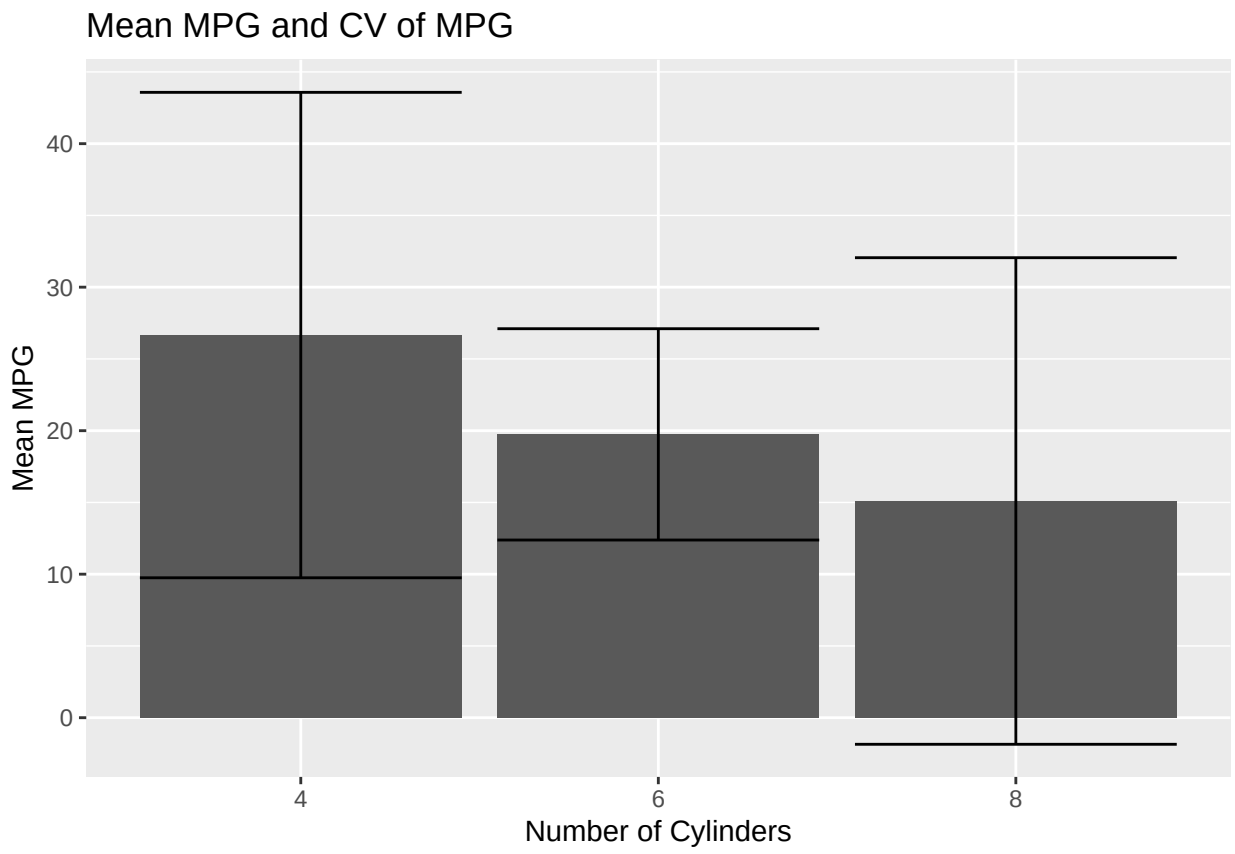
D. Write a function to calculate the coefficient of variation (CV) for a given numeric column and apply this function to mpg, hp, and wt for each cylinder group.

```r
getCV <- function(x) {
  cv <- sd(x) / mean(x) * 100
  return(cv)
}

cvs <- mtcars %>%
  group_by(cyl) %>%
  summarise(cvMpg = getCV(mpg),
            cvHp = getCV(hp),
            cvWt = getCV(wt))

cvs
```

```
## # A tibble: 3 x 4
##     cyl cvMpg  cvHp  cvWt
##   <dbl> <dbl> <dbl> <dbl>
## 1     4 16.9   25.3  24.9
## 2     6  7.36  19.8  11.4
## 3     8 17.0   24.4  19.0
```

E. Plot the mean mpg and CV of mpg for each cylinder group using a bar plot with error bars.

```r
plotData <- mtcars %>%
  group_by(cyl) %>%
  summarise(meanMpg = mean(mpg), cvMpg = getCV(mpg))

ggplot(plotData, aes(x = factor(cyl), y = meanMpg)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = meanMpg - cvMpg, ymax = meanMpg + cvMpg), ) +
  labs(x = "Number of Cylinders", y = "Mean MPG", title = "Mean MPG and CV of MPG")
```



---

3. Data Reshaping with tidyr

A. Load the airquality dataset. Reshape the dataset from wide to long format, using gather() for the measurements (Ozone, Solar.R, Wind, Temp).

```
data("airquality")
airqualityData <- airquality
airqualityData <- airqualityData %>% gather(key = "variable", value = "value", Ozone:Temp)

head(airqualityData)
```

```
##   Month Day variable value
## 1     5   1    Ozone    41
## 2     5   2    Ozone    36
## 3     5   3    Ozone    12
## 4     5   4    Ozone    18
## 5     5   5    Ozone    NA
## 6     5   6    Ozone    28
```

B. Reshape the dataset back to wide format using spread().

```
airqualityData <- airqualityData %>% spread(key = variable, value = value)

head(airqualityData)
```

```
##   Month Day Ozone Solar.R Temp Wind
## 1     5   1    41     190   67  7.4
## 2     5   2    36     118   72  8.0
## 3     5   3    12     149   74 12.6
## 4     5   4    18     313   62 11.5
## 5     5   5    NA      NA   56 14.3
## 6     5   6    28      NA   66 14.9
```

C. Use separate() to split the Month column into Month and Day columns (if it were combined), and then recombine them using unite().

```
# since its not combine the following code is commented out
# airqualityData <- airqualityData %>% separate(Date, c("Month", "Day"), sep="-")

airqualityData <- airqualityData %>% unite("Date", Month, Day, sep = "-")

head(airqualityData)
```

```
##   Date Ozone Solar.R Temp Wind
## 1  5-1    41     190   67  7.4
## 2  5-2    36     118   72  8.0
## 3  5-3    12     149   74 12.6
## 4  5-4    18     313   62 11.5
## 5  5-5    NA      NA   56 14.3
## 6  5-6    28      NA   66 14.9
```

D. Create a summary table showing the average values for each variable by month.

```
airqualityData <- airqualityData %>% separate(Date, c("Month", "Day"), sep =
                                                "-")

aqdSummary <- airqualityData %>%
  group_by(Month) %>%
  summarise(
    meanOzone = mean(Ozone , na.rm = TRUE),
    meanSolar.R = mean(Solar.R , na.rm = TRUE),
    meanTemp = mean(Temp , na.rm = TRUE),
    meanWind = mean(Wind , na.rm = TRUE),
  )

aqdSummary
```

```
## # A tibble: 5 x 5
##   Month meanOzone meanSolar.R meanTemp meanWind
##   <chr>     <dbl>       <dbl>    <dbl>    <dbl>
## 1 5          23.6        181.     65.5    11.6
## 2 6          29.4        190.     79.1    10.3
## 3 7          59.1        216.     83.9     8.94
## 4 8          60.0        172.     84.0     8.79
## 5 9          31.4        167.     76.9    10.2
```

---

4. Introduction to Probability

A. Simulate rolling a fair six-sided die 1000 times. Calculate the empirical probability of each outcome.

```
# TODO: Finish this question
```

B. Simulate drawing a card from a standard deck of 52 cards 1000 times. Calculate the empirical probability of drawing an Ace.

```
# TODO: add information about assignment and libraries used
```

C. Use the binomial distribution to calculate the probability of getting exactly 5 heads in 10 flips of a fair coin. Repeat for getting 5 or more heads.

```
# TODO: add information about assignment and libraries used
```

D. Generate a plot showing the probability mass function (PMF) of a binomial distribution with parameters n = 10 and p = 0.5.

```
# TODO: add information about assignment and libraries used
```

---

5. Advanced Data Manipulation and Visualization

A. Load the iris dataset and create a summary table showing the mean, median, and standard deviation of each

numerical variable grouped by Species.

```
# TODO: Finish this question
```

B. Create a pairwise scatter plot matrix using the pairs() function for the iris dataset colored by Species.

```
# TODO: add information about assignment and libraries used
```

C. Use ggplot2 to create a violin plot for Petal.Length grouped by Species.

```
# TODO: add information about assignment and libraries used
```

D. Create a heatmap of the correlation matrix for the numerical variables in the iris dataset.

```
# TODO: add information about assignment and libraries used
```

E. Write a short analysis (5-7 sentences) interpreting the results from the summary table, scatter plot matrix, violin plot, and heatmap.

```
# TODO: add information about assignment and libraries used
```

---

6. Data Reshaping and Aggregation

A. Load the gapminder dataset from the gapminder package. Reshape the dataset to long format, focusing on the variables year and gdpPercap.

```
# TODO: Finish this question
```

B. Aggregate the data to calculate the average gdpPercap by continent and year.

```
# TODO: add information about assignment and libraries used
```

C. Create a line plot of the average gdpPercap over time for each continent.

```
# TODO: add information about assignment and libraries used
```

D. Create a faceted plot showing gdpPercap distributions by continent for the most recent year in the dataset.

```
# TODO: add information about assignment and libraries used
```

E. Write a detailed report (6-8 sentences) analyzing the trends and patterns observed in the plots.

```
# TODO: add information about assignment and libraries used
```

---

7. Probability

A local fraternity is conducting a raffle where 50 tickets are to be sold, one per customer. There are three prizes to be awarded. If the four organizers of the raffle each buy one ticket, what is the probability that the four organizers win

A. all of the prizes?

```
# TODO: Finish this question
```

B. exactly two of the prizes?

```
# TODO: add information about assignment and libraries used
```

C. exactly one of the prizes?

```
# TODO: add information about assignment and libraries used
```

D. none of the prizes?

```
# TODO: add information about assignment and libraries used
```