Name (Last, First):

Lopez Sepulveda,
Kevin

Student ID:

U59702827

Assignment 7

METCS544A3A4_F2024

Instructions:

1. For answering programming questions, please use Adobe Acrobat to edit the pdf file in two steps **[See Appendix: Example Question and Answer]**:
    a. Copy and paste your R code as text in the box provided (so that your teaching team can run your code);
    b. Screenshot your R console outputs, save them as a .PNG image file, and paste/insert them in the box provided.
    c. Show all work—credit will not be given for code without showing it in action, including a screenshot of R console outputs.
2. To answer non-programming questions, please type or handwrite your final answers clearly in the boxes. Show all work - credit will not be given for numerical solutions that appear without explanation in the space above the boxes. **You're encouraged to use R to graph/plot the data and produce numerical summaries; please append your code and screenshot of the outputs at the end of your PDF submission.**
4. **[Total 86 pts = 33 + 50 pts + 3 Extra Credit pts]**

Grading Rubric

Each question is worth 3 points and will be graded as follows:

3 points: Correct answer with work shown

2 points: Incorrect answer but attempt shows some understanding (work shown)

1 point: Incorrect answer but an attempt was made (work shown), or **correct answer without explanation (work not shown)**

0 points: Left blank or made little to no effort/work not shown

Reflective Journal [3 pts]

(Copy and paste the link to your live Google doc in the box below)

https://drive.google.com/drive/folders/1_8qcBjQVMfZggF42UYJuHQzMoBcyAy0Q?usp=drive_link

Part I. Collecting Data: Experimental Design (11 x 3 = 33 pts)

1) A local hospital compiles data on the length of time a patient is in surgery and the length of their stay in the hospital after surgery. A newbie hospital worker notices that the length of time a patient is in surgery is highly correlated with their hospital stay after. Explain to the newbie why this correlation does not imply a causation.

Answer:

There might be other factors unaccounted for in this inference. These include the complexity of the medical condition and patient health. These underlying variables may be a cause to this. It can show that the medical treatment being undergone requires more monitoring or that the patient is consistently sick with something and comes in a lot.

2) The good folks at Apples Inc want to create bags of pre-sliced apples for kids to easily put in their lunch boxes. They want to test different mixtures of preservatives (A, B, and C) on their Honeycrisp apples. They will treat all the apples in a bushel of Honeycrisps by randomly assigning them to each preservation treatment, and then comparing how long they are able to remain in the bag before they begin to brown.

(a) Identify the experimental units, the explanatory and response variables, and the treatments.

Answer:

The experimental unit is the apples, the explanatory variable is the type of preservative and the response variable is time before browning and the treatments are preservatives A,B,C.

(b) The researchers plan to use a completely randomized design. Describe how they should assign treatments to the 126 experimental units.

Answer:

They should separate the 126 into 3 groups of 42. Then for group A randomly select 1 of the original 126 and subtract it from the first group of 42 and subtract it from the original 126. Once the first group of 42 reaches 0 then go to the next group where there should be 84 original apples left.

(c) The researchers suspect that the type of apple will influence how well the preservation method works. They want to repeat this experiment with 4 different types of apples. Describe how they should change the design of the experiment to account for this addition.

Answer:

They should assign each apple into a block and repeat the same exact experiment. They would assign randomly each preservative and see the results and determine if the type of preservative works better on some apples more than others by comparing them between the blocks and determine if a preservative is universally better among all the blocks.

3) A study was conducted to determine if taking a daily dose of aspirin reduces the chance of catching the common cold. This study was conducted using 550 volunteers who were not already on an aspirin regimen. The subjects were randomly assigned to one of two groups: a treatment group who received a low dose of aspirin daily, or a control group who received a placebo. The subjects were unaware of what group they were in. At the end of the study, the subjects were asked to meet with a doctor to discuss if they had any symptoms pertaining to the common cold.

(a) Is this study an experiment or an observational study? Explain your answer.

Answer:

This is an experiment. In an observational study, no variables are explicitly manipulated and everything is observed and recorded. In this example the study used a placebo to make the study participants unaware of the variables they were participating in. The study implemented a change and looked for a cause-and-effect relationship in the outcomes they were studying.

(b) What would be the advantage of having this study be double blind?

Answer:

This would ensure that the results aren't being swayed by participants or researchers. It allows for the participants to not exaggerate results of being on aspirin or not and it would allow the researchers to not have bias in looking for sickness in those without aspirin specifically or assuming less sick results in those with aspirin.

(c) Would blocking according to gender be worthwhile in this study? Explain your answer.
Answer:

It would not be worthwhile because we would limit the efficacy of the study unless we grab a larger sample size. It would have to be predetermined whether the aspirin has a separate effect on men and women and that would affect the study.

(d) Describe what the "placebo effect" would look like in this experiment.
Answer:
The placebo effect would show that those taking the fake aspirin will report not feeling sick at all or less symptoms largely because they believe they are being treated.

4) Type 1 diabetes is thought to be caused by an autoimmune reaction (the body attacks itself by mistake) that destroys the cells in the pancreas that make insulin. It is a disease that currently has no cure. Researchers were to conduct a clinical trial on a promising new drug that helps the body produce its own insulin.

(a) You have 400 volunteers with Type 1 diabetes. Describe a completely randomized design for this experiment.

Answer:

Step 1: Define the Experimental Units

First establish the experimental units as the 400 volunteers with Type 1 diabetes. Then use randomization to divide the 400 volunteers into two groups: 200 participants to receive the new treatment and 200 to receive regular treatment. Next use blinding so that neither the participants nor the researchers know who is receiving the new drug or the placebo. Lastly measure the outcomes of their blood sugar, and insulin levels. Finally analyze results.

(b) You have 250 children (17 and younger) with Type 1 diabetes (120 females and 130 males), as well as 300 adults (18 and older) with Type 1 diabetes (190 females and 110 males). Describe a completely randomized block design.

Answer:

A completely randomized block design would be separating the children from adults and splitting them based on that by half each on that. So the 250 children would be 125 as a control and 125 as experimental undergoing the treatment and the 300 adults would have 150 of them randomly assigned to undergo treatment and the other half would get a placebo.

(c) Explain why a matched pairs design would not work for this experiment.

Answer:

A matched pairs design would not work because of the many differences in age and gender causing a difficult exact match for each one.

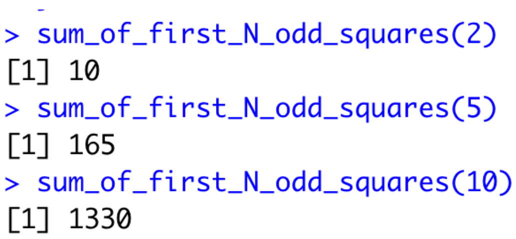
Part II. Statistical Programming (50 pts)

1. Functions (20 pts)

a) Using a **for** loop or a **while** loop, write your own **R function**,
sum_of_first_N_odd_squares (***n***),
that returns the sum of the squares of the first **n** odd numbers.

For example, if n = 5, the first five odd numbers are 1, 3, 5, 7, and 9, and the required result is
$1^2 + 3^2 + 5^2 + 7^2 + 9^2 = 165$.

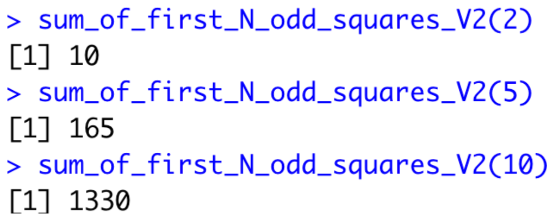
Test your function as follows:



```
> sum_of_first_N_odd_squares(2)
[1] 10
> sum_of_first_N_odd_squares(5)
[1] 165
> sum_of_first_N_odd_squares(10)
[1] 1330
```

b) Now, **without** using any loop, write your own **R function**,
sum_of_first_N_odd_squares_V2 (***n***),
that returns the sum of the squares of the first **n** odd numbers.

Test your function as follows:



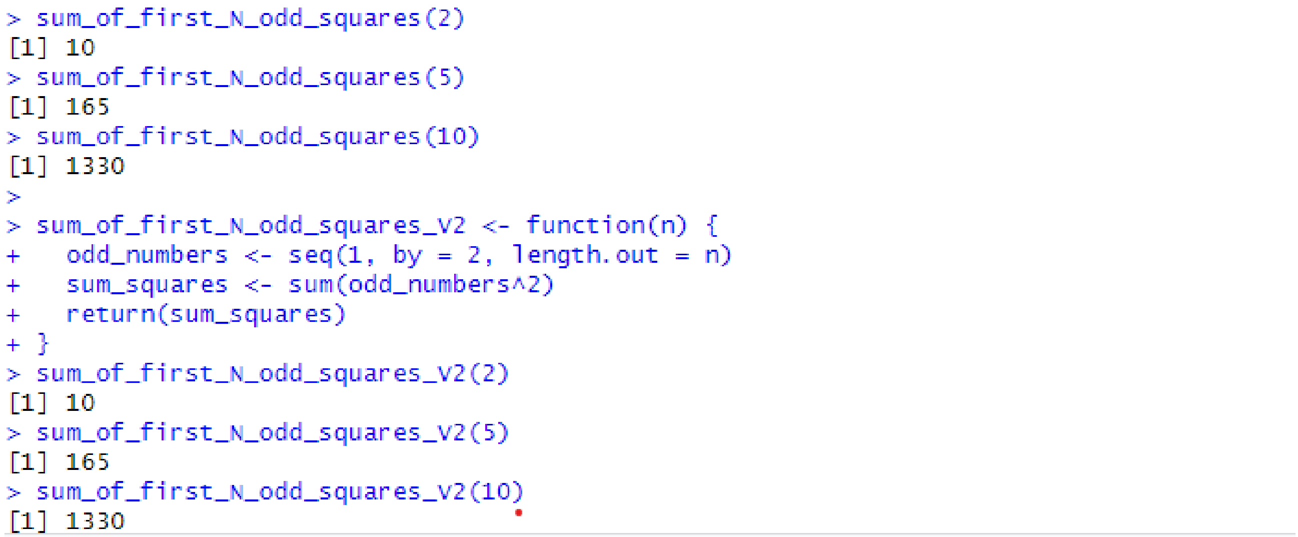
```
> sum_of_first_N_odd_squares_V2(2)
[1] 10
> sum_of_first_N_odd_squares_V2(5)
[1] 165
> sum_of_first_N_odd_squares_V2(10)
[1] 1330
```

Answer: Copy and paste your R code in the box below (not an image but the text).

```
sum_of_first_N_odd_squares <- function(n) {
  sum_squares <- 0
  for (i in 1:n) {
    odd_number <- 2 * i - 1
    sum_squares <- sum_squares + odd_number^2
  }
  return(sum_squares)
}


sum_of_first_N_odd_squares(2)
sum_of_first_N_odd_squares(5)
sum_of_first_N_odd_squares(10)


sum_of_first_N_odd_squares_V2 <- function(n) {
  odd_numbers <- seq(1, by = 2, length.out = n)
  sum_squares <- sum(odd_numbers^2)
  return(sum_squares)
}
sum_of_first_N_odd_squares_V2(2)
sum_of_first_N_odd_squares_V2(5)
sum_of_first_N_odd_squares_V2(10)
```

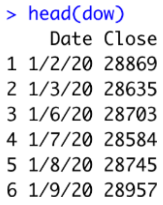


```
> sum_of_first_N_odd_squares(2)
[1] 10
> sum_of_first_N_odd_squares(5)
[1] 165
> sum_of_first_N_odd_squares(10)
[1] 1330
>
> sum_of_first_N_odd_squares_V2 <- function(n) {
+   odd_numbers <- seq(1, by = 2, length.out = n)
+   sum_squares <- sum(odd_numbers^2)
+   return(sum_squares)
+ }
> sum_of_first_N_odd_squares_V2(2)
[1] 10
> sum_of_first_N_odd_squares_V2(5)
[1] 165
> sum_of_first_N_odd_squares_V2(10)
[1] 1330
```

2. R Programming (30 pts)

Initialize the Dow Jones Industrials daily closing data, *dow*, using the read.csv function with the link: DJI_2020.csv

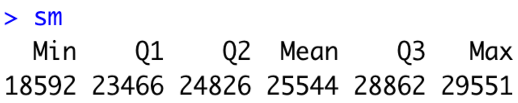
The first 6 rows of the dataset are as shown below:



```
> head(dow)
     Date Close
1 1/2/20 28869
2 1/3/20 28635
3 1/6/20 28703
4 1/7/20 28584
5 1/8/20 28745
6 1/9/20 28957
```

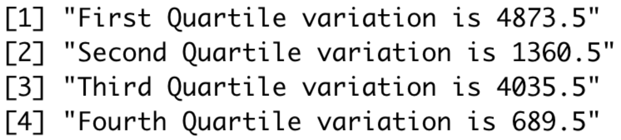
Provide the simplest R code and output for all of the following. **The code should work for any given data**.

a) Store the result of the **summary** function for the *Close* attribute as the variable *sm*. Change the *names* of this variable so that the output appears as shown below.



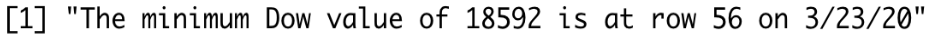
```
> sm
  Min    Q1    Q2  Mean    Q3   Max 
18592 23466 24826 25544 28862 29551 
```

Using the above data, show the quartile variations for the four quartiles as shown below. You can use **paste** or **sprintf.**



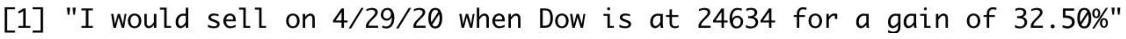
```
[1] "First Quartile variation is 4873.5"
[2] "Second Quartile variation is 1360.5"
[3] "Third Quartile variation is 4035.5"
[4] "Fourth Quartile variation is 689.5"
```

b) Produce the output for the minimum of the Dow closing value in the dataset as shown below:



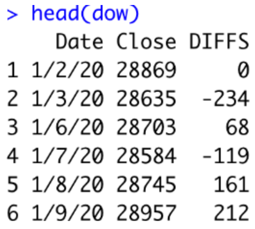
```
[1] "The minimum Dow value of 18592 is at row 56 on 3/23/20"
```

c) Suppose you have an index fund tied to the Dow closing value. If you have invested on the minimum date, what date from the dataset you would have sold to gain the maximum percentage gain. The output is as shown below. Note that the code should be generic so that it works on any such dataset.



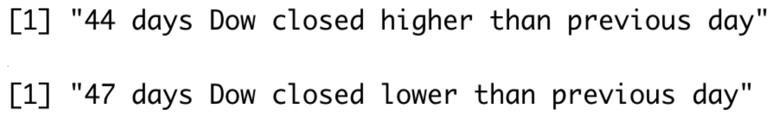
```
[1] "I would sell on 4/29/20 when Dow is at 24634 for a gain of 32.50%"
```

d) Use the **diff** function to calculate the differences between consecutive closing values in the dataset. Insert the value 0 at the beginning of these differences. Add this result as the DIFFS column of the data frame. The result is as shown below.



```
> head(dow)
    Date Close DIFFS
1 1/2/20 28869     0
2 1/3/20 28635  -234
3 1/6/20 28703    68
4 1/7/20 28584  -119
5 1/8/20 28745   161
6 1/9/20 28957   212
```

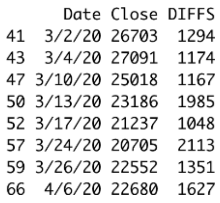
e) How many days did the Dow close higher than its previous day value? How many days did the Dow close lower than its previous day's value?



```
[1] "44 days Dow closed higher than previous day"

[1] "47 days Dow closed lower than previous day"
```

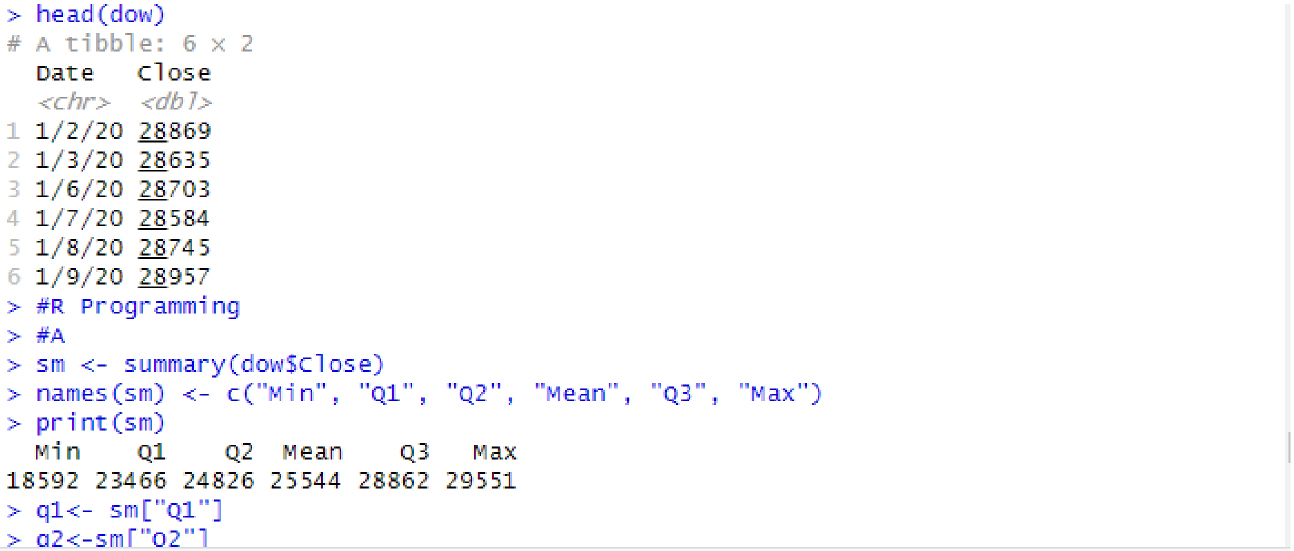
f) Show the subset of the data where there was a **gain** of at least 1000 points from its previous day value.



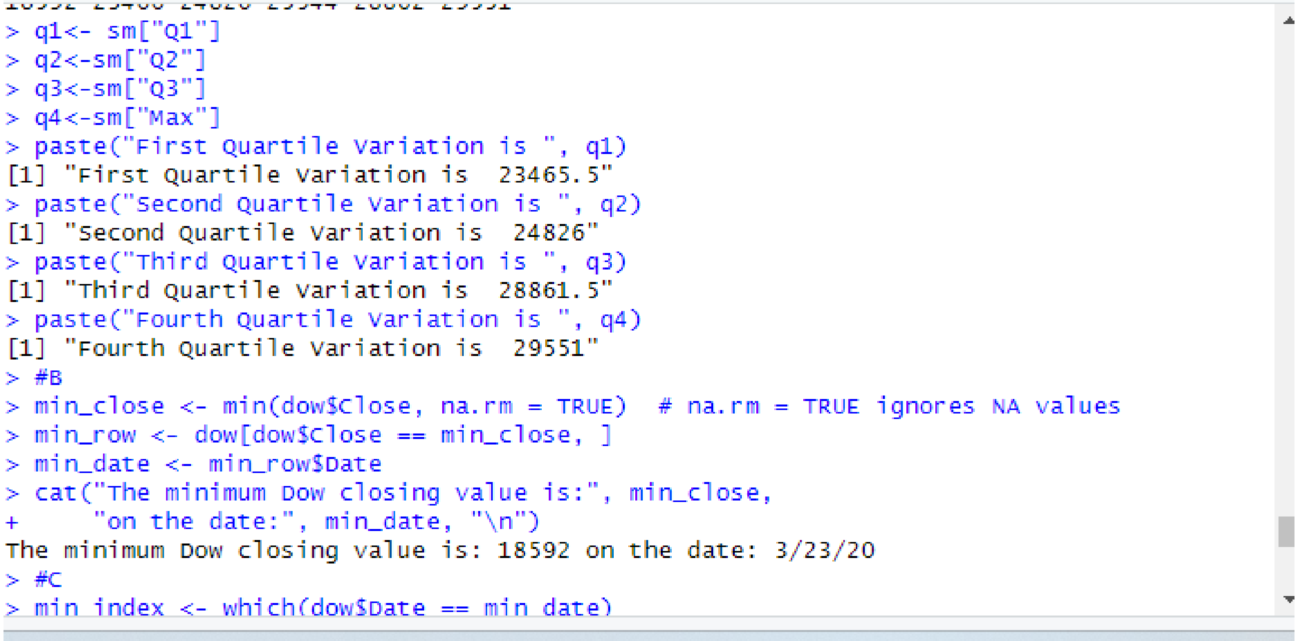
```
      Date Close DIFFS
41  3/2/20 26703  1294
43  3/4/20 27091  1174
47 3/10/20 25018  1167
50 3/13/20 23186  1985
52 3/17/20 21237  1048
57 3/24/20 20705  2113
59 3/26/20 22552  1351
66  4/6/20 22680  1627
```

Answer: Copy and paste your R code in the box below (not an image but the text).

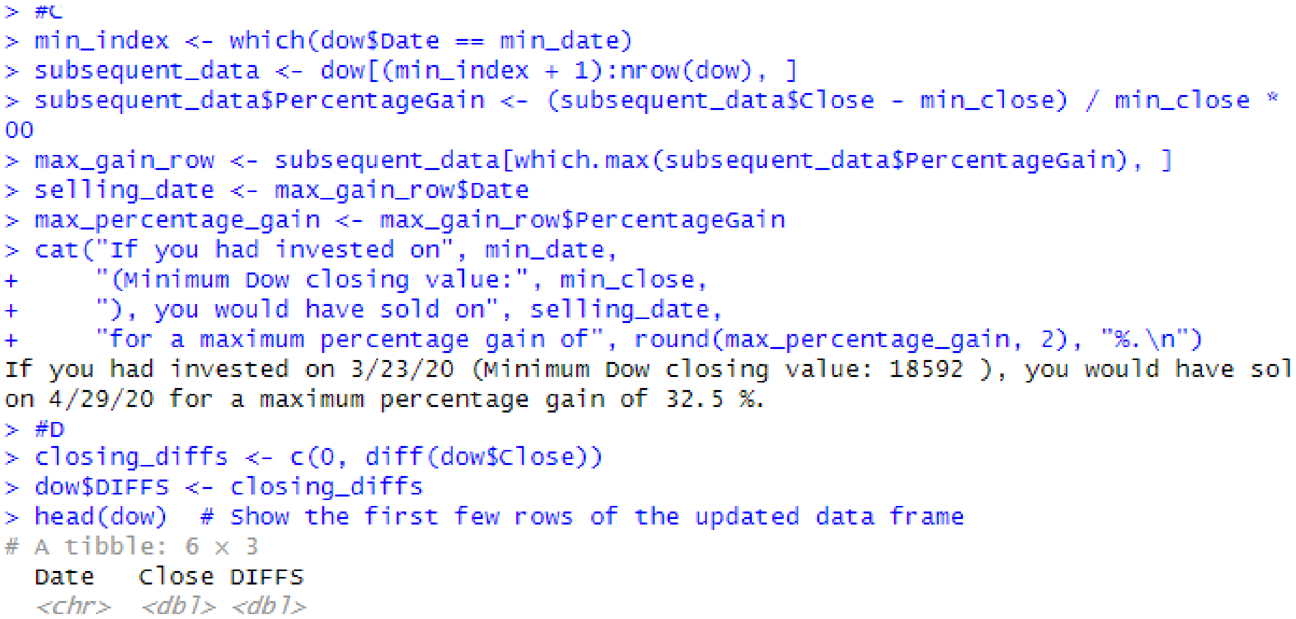
```
library(readr)
dow <- read_csv("CS544-BU/DJI_2020.csv")
head(dow)
#R Programming
#A
sm <- summary(dow$Close)
names(sm) <- c("Min", "Q1", "Q2", "Mean", "Q3", "Max")
print(sm)
q1<- sm["Q1"]
q2<-sm["Q2"]
q3<-sm["Q3"]
q4<-sm["Max"]
paste("First Quartile Variation is ", q1)
paste("Second Quartile Variation is ", q2)
paste("Third Quartile Variation is ", q3)
paste("Fourth Quartile Variation is ", q4)
#B
min_close <- min(dow$Close, na.rm = TRUE)  # na.rm = TRUE ignores NA values
min_row <- dow[dow$Close == min_close, ]
min_date <- min_row$Date
cat("The minimum Dow closing value is:", min_close,
    "on the date:", min_date, "\n")
#C
min_index <- which(dow$Date == min_date)
subsequent_data <- dow[(min_index + 1):nrow(dow), ]
subsequent_data$PercentageGain <- (subsequent_data$Close - min_close) / min_close * 100
max_gain_row <- subsequent_data[which.max(subsequent_data$PercentageGain), ]
selling_date <- max_gain_row$Date
max_percentage_gain <- max_gain_row$PercentageGain
cat("If you had invested on", min_date,
    "(Minimum Dow closing value:", min_close,
    "), you would have sold on", selling_date,
    "for a maximum percentage gain of", round(max_percentage_gain, 2), "%.\n")
#D
closing_diffs <- c(0, diff(dow$Close))
dow$DIFFS <- closing_diffs
head(dow)  # Show the first few rows of the updated data frame
#E
days_higher <- sum(dow$DIFFS > 0)
days_lower <- sum(dow$DIFFS < 0)
cat(days_higher, " Days Dow closed higher than the previous day \n")
cat(days_lower, " Days Dow closed higher than the previous day \n")
#F
gains_over_1000 <- dow[dow$DIFFS >= 1000, ]
print(gains_over_1000)
```



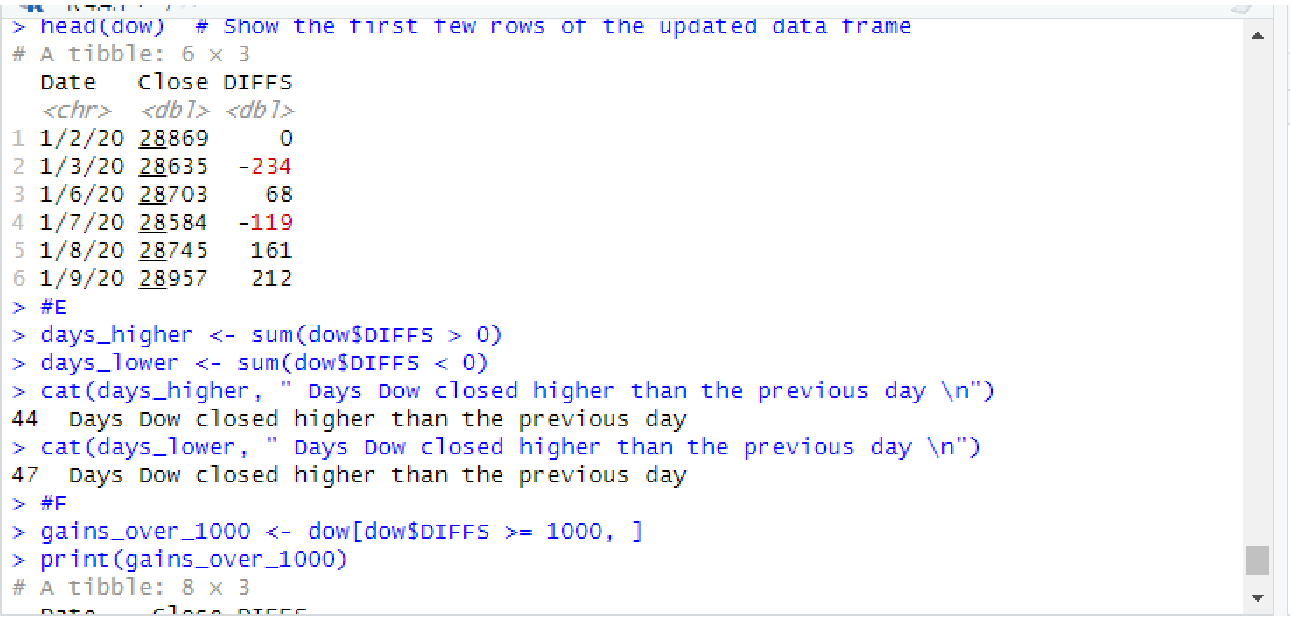
```
> head(dow)
# A tibble: 6 × 2
  Date   Close
  <chr>  <dbl>
1 1/2/20 28869
2 1/3/20 28635
3 1/6/20 28703
4 1/7/20 28584
5 1/8/20 28745
6 1/9/20 28957
> #R Programming
> #A
> sm <- summary(dow$Close)
> names(sm) <- c("Min", "Q1", "Q2", "Mean", "Q3", "Max")
> print(sm)
  Min    Q1    Q2  Mean    Q3   Max 
18592 23466 24826 25544 28862 29551 
> q1<- sm["Q1"]
> q2<-sm["Q2"]
```



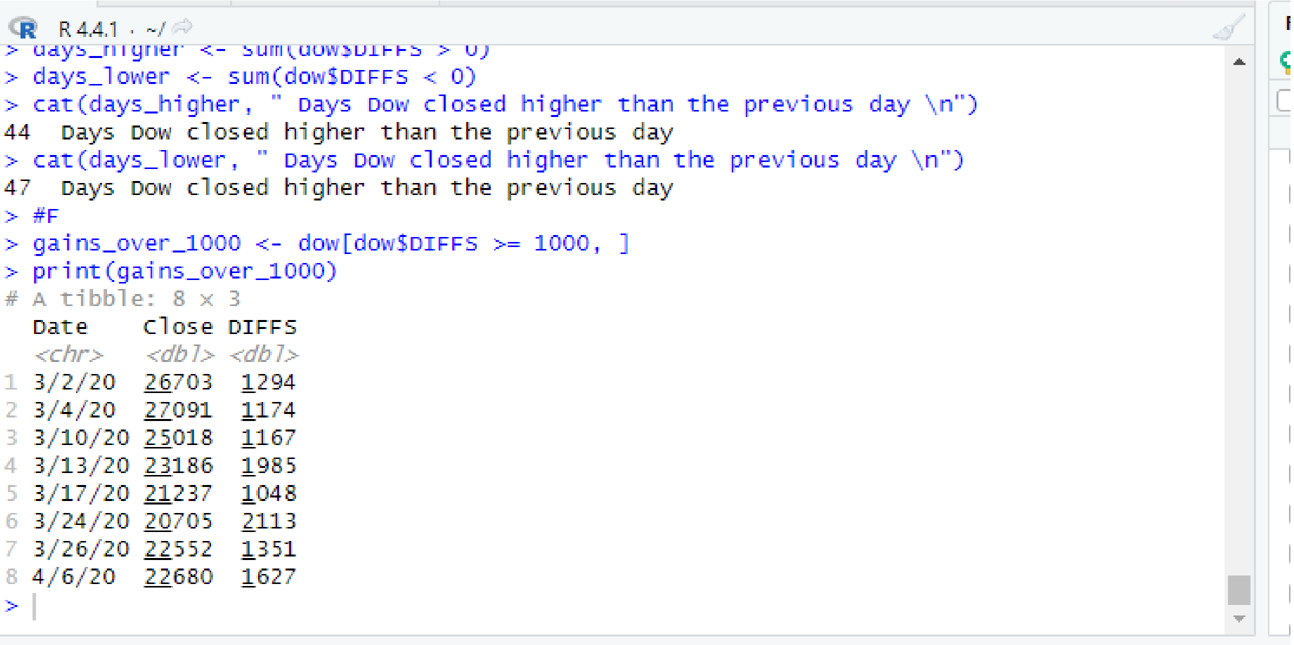
```
18592 23466 24826 25544 28862 29551 
> q1<- sm["Q1"]
> q2<-sm["Q2"]
> q3<-sm["Q3"]
> q4<-sm["Max"]
> paste("First Quartile Variation is ", q1)
[1] "First Quartile Variation is  23465.5"
> paste("Second Quartile Variation is ", q2)
[1] "Second Quartile Variation is  24826"
> paste("Third Quartile Variation is ", q3)
[1] "Third Quartile Variation is  28861.5"
> paste("Fourth Quartile Variation is ", q4)
[1] "Fourth Quartile Variation is  29551"
> #B
> min_close <- min(dow$Close, na.rm = TRUE)  # na.rm = TRUE ignores NA values
> min_row <- dow[dow$Close == min_close, ]
> min_date <- min_row$Date
> cat("The minimum Dow closing value is:", min_close,
+     "on the date:", min_date, "\n")
The minimum Dow closing value is: 18592 on the date: 3/23/20 
> #C
> min_index <- which(dow$Date == min_date)
```



```
> #C
> min_index <- which(dow$Date == min_date)
> subsequent_data <- dow[(min_index + 1):nrow(dow), ]
> subsequent_data$PercentageGain <- (subsequent_data$Close - min_close) / min_close *
00
> max_gain_row <- subsequent_data[which.max(subsequent_data$PercentageGain), ]
> selling_date <- max_gain_row$Date
> max_percentage_gain <- max_gain_row$PercentageGain
> cat("If you had invested on", min_date,
+     "(Minimum Dow closing value:", min_close,
+     "), you would have sold on", selling_date,
+     "for a maximum percentage gain of", round(max_percentage_gain, 2), "%.\n")
If you had invested on 3/23/20 (Minimum Dow closing value: 18592 ), you would have sol
on 4/29/20 for a maximum percentage gain of 32.5 %.
> #D
> closing_diffs <- c(0, diff(dow$Close))
> dow$DIFFS <- closing_diffs
> head(dow)  # Show the first few rows of the updated data frame
# A tibble: 6 × 3
  Date    Close DIFFS
  <chr>   <dbl> <dbl>
```

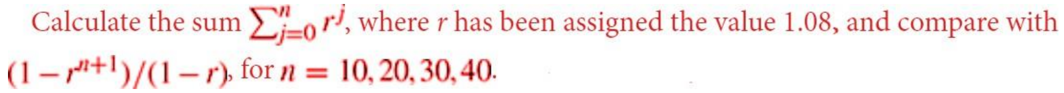


```
> head(dow)  # Show the first few rows of the updated data frame
# A tibble: 6 × 3
  Date    Close DIFFS
  <chr>   <dbl> <dbl>
1 1/2/20 28869     0
2 1/3/20 28635  -234
3 1/6/20 28703    68
4 1/7/20 28584  -119
5 1/8/20 28745   161
6 1/9/20 28957   212
> #E
> days_higher <- sum(dow$DIFFS > 0)
> days_lower <- sum(dow$DIFFS < 0)
> cat(days_higher, " Days Dow closed higher than the previous day \n")
44  Days Dow closed higher than the previous day
> cat(days_lower, " Days Dow closed higher than the previous day \n")
47  Days Dow closed higher than the previous day
> #F
> gains_over_1000 <- dow[dow$DIFFS >= 1000, ]
> print(gains_over_1000)
# A tibble: 8 × 3
  Date     Close DIFFS
```



```
> days_higher <- sum(dow$DIFFS > 0)
> days_lower <- sum(dow$DIFFS < 0)
> cat(days_higher, " Days Dow closed higher than the previous day \n")
44  Days Dow closed higher than the previous day
> cat(days_lower, " Days Dow closed higher than the previous day \n")
47  Days Dow closed higher than the previous day
> #F
> gains_over_1000 <- dow[dow$DIFFS >= 1000, ]
> print(gains_over_1000)
# A tibble: 8 × 3
  Date    Close DIFFS
  <chr>   <dbl> <dbl>
1 3/2/20  26703  1294
2 3/4/20  27091  1174
3 3/10/20 25018  1167
4 3/13/20 23186  1985
5 3/17/20 21237  1048
6 3/24/20 20705  2113
7 3/26/20 22552  1351
8 4/6/20  22680  1627
> 
```

Appendix: Example Question and Answer for R programming questions:



Calculate the sum $\sum_{j=0}^{n} r^j$, where r has been assigned the value 1.08, and compare with $(1-r^{n+1})/(1-r)$, for $n = 10, 20, 30, 40$.

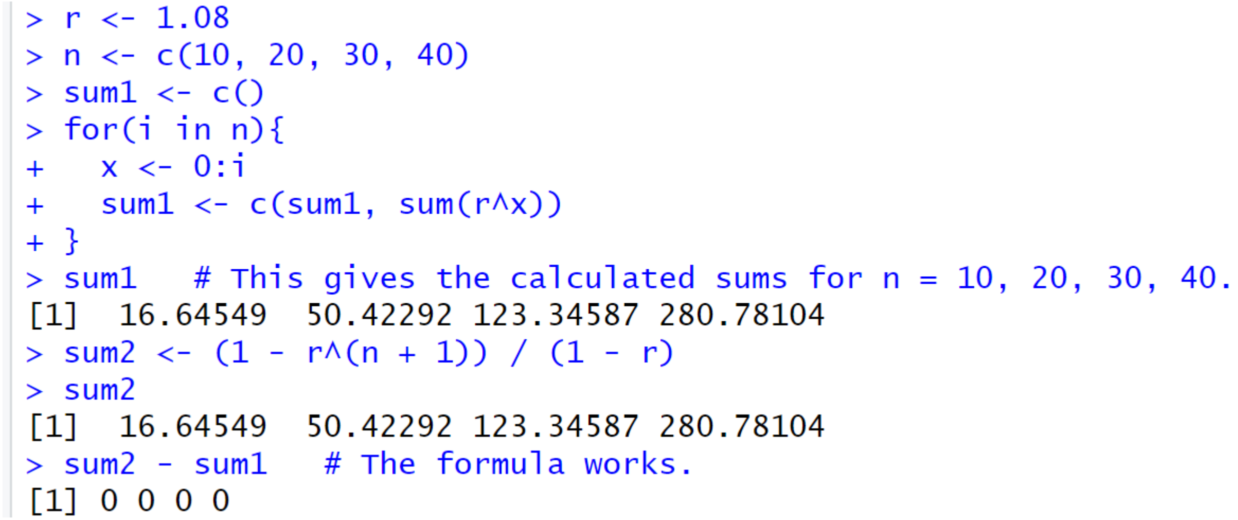
Answer: Copy and paste your R code in the box below (not an image but the text).

```
r <- 1.08
n <- c(10, 20, 30, 40)
sum1 <- c()
for(i in n){
  x <- 0:i
  sum1 <- c(sum1, sum(r^x))
}
sum1   # This gives the calculated sums for n = 10, 20, 30, 40.

sum2 <- (1 - r^(n + 1)) / (1 - r)
sum2

sum2 - sum1   # The formula works.
```

Screenshot of your R console outputs and paste the image in the box below



```
> r <- 1.08
> n <- c(10, 20, 30, 40)
> sum1 <- c()
> for(i in n){
+   x <- 0:i
+   sum1 <- c(sum1, sum(r^x))
+ }
> sum1   # This gives the calculated sums for n = 10, 20, 30, 40.
[1]  16.64549  50.42292 123.34587 280.78104
> sum2 <- (1 - r^(n + 1)) / (1 - r)
> sum2
[1]  16.64549  50.42292 123.34587 280.78104
> sum2 - sum1   # The formula works.
[1] 0 0 0 0
```

THE END