Complete these works showing your codes and outputs from R studio:

Work 1: See slide 25 of session 2 slide deck and provide answers here.

Work 2: See slide 26-30 of session 2 slide deck and provide answers here. Data is attached.

Work 3: See slide 31 of session 2 slide deck and provide answers here. Data is attached.

Work 4: See slide 32 of session 2 slide deck and provide answers here. Data is attached.

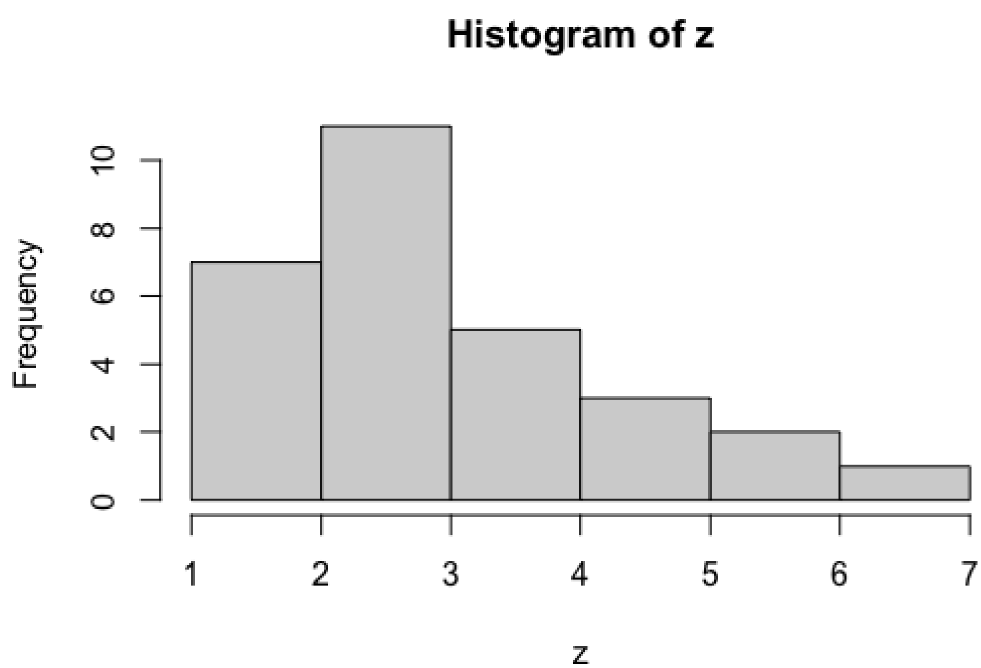
—---

Work 1: Work/Assignment 1:

1.1 Show the histogram of the z variable and interpret it carefully.

```
z<- c(1,1,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,5,5,5,6,6,7)
hist(z)
```

This will create a histogram with 7 bins. Corresponding to the 7 unique values in the z vector as shown below:



Interpretation:
The histogram shows the distribution of values in the z vector. We can see that the most frequent value is 3, which appears 11 times in the vector. Values 2 and 4 are the next most frequent, each appearing 5 times, 5 appearing 3 times, followed by values 1 and 6 which each appear 2 times. Finally,7 appear once.

The histogram also shows that the distribution is skewed with a longer tail to the right. This indicates that values 6 and 7 are outliers, as they are far from the most frequently occurring values.

1.2 Get a summary of this variable and decide which measure of central tendency and measure of dispersion must be used for this variable?

summary(z)

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   3.000   3.000   3.414   4.000   7.000
```

From the summary, we can see that the minimum value is 1, the maximum value is 7, the median is 3, the mean is 3.414 , and the first and third quartiles are 3 and 4, respectively.

For this variable, both the mean and median can be used as measures of central tendency. However, since the distribution is skewed to the right (the mean is greater than the median), the median may be a more appropriate measure of central tendency.

For the measure of dispersion, the range, interquartile range, and standard deviation are commonly used. Since the data is skewed and has some outliers, the interquartile range would be a more appropriate measures of dispersion than the range or standard deviation.

1.3 Get the five number summary of this variable and interpret them carefully.

fivenum(z)
[1] 1 3 3 4 7

The five-number summary consists of the

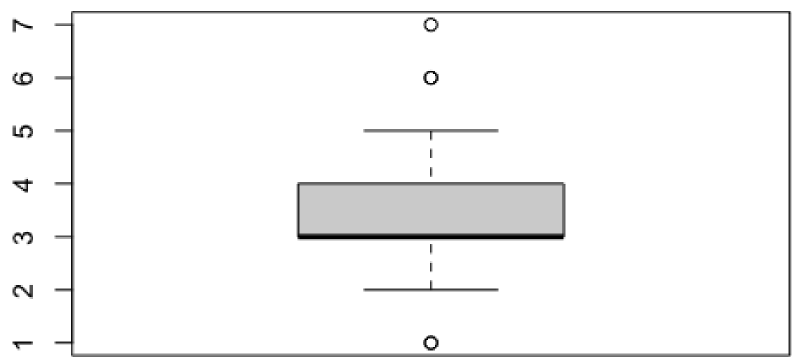
1. Minimum
2. First quartile (Q1)
3. Median
4. Third quartile (Q3)
5. Maximum.

The minimum value of 1 indicates that the smallest value in the dataset is 1.
The first quartile (Q1) of 3 means that 25% of the values in the dataset are less than or equal to 3.
The median of 3 indicates that 50% of the values in the dataset are less than or equal to 3.
The third quartile (Q3) of 4 means that 75% of the values in the dataset are less than or equal to 4.
The maximum value of 7 indicates that the largest value in the dataset is 7.

The five-number summary can be used to construct a box plot. The box represents the interquartile range (IQR), which is the distance between the first and third quartiles. The median is represented by a line inside the box, and the whiskers extend to the minimum and maximum values in the dataset, unless there are any outliers.

1.4 Create a boxplot of this variable and interpret it carefully.

boxplot(z)



The box plot consists of several elements that provide information about the distribution of the data:
The box represents the middle 50% of the data, with the bottom of the box being the 25th percentile (Q1) and the top of the box being the 75th percentile (Q3).
The line inside the box represents the median, which is the middle value of the dataset. Here in this given data we got Q1 and Median as same value i.e. 3
The whiskers extend from the top and bottom of the box to the highest and lowest data points within 1.5 times the interquartile range (IQR) from the box. Data points beyond the whiskers are considered as outliers also called extreme values and are plotted as individual points.

1.5 Do you get an outlier for this variable in the box plot? Why?

According to the box plot, there are outliers in the data because some of the data points do not fall within the whiskers.

In Summary we got
Min. = 1
Q1 = 3
Median = 3
Mean = 3.14
Q3 = 4
Max. = 7

we can confirm this by calculating the interquartile range (IQR)
IQR = Q3 - Q1 = 4 - 3 = 1

Any data point that falls below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR is considered an outlier.
In this case,
Q1 - 1.5 * IQR = 1.5
and
Q3 + 1.5 * IQR = 5.5.

Since there are some data points like 1,6,7 which are outside this range, we can conclude that 1,6 and 7 are the outliers in the data.

Here we get the outlier data from Box plot in R by:

```
boxplot(z)$out
```

Output:
1 1 6 6 7

—--

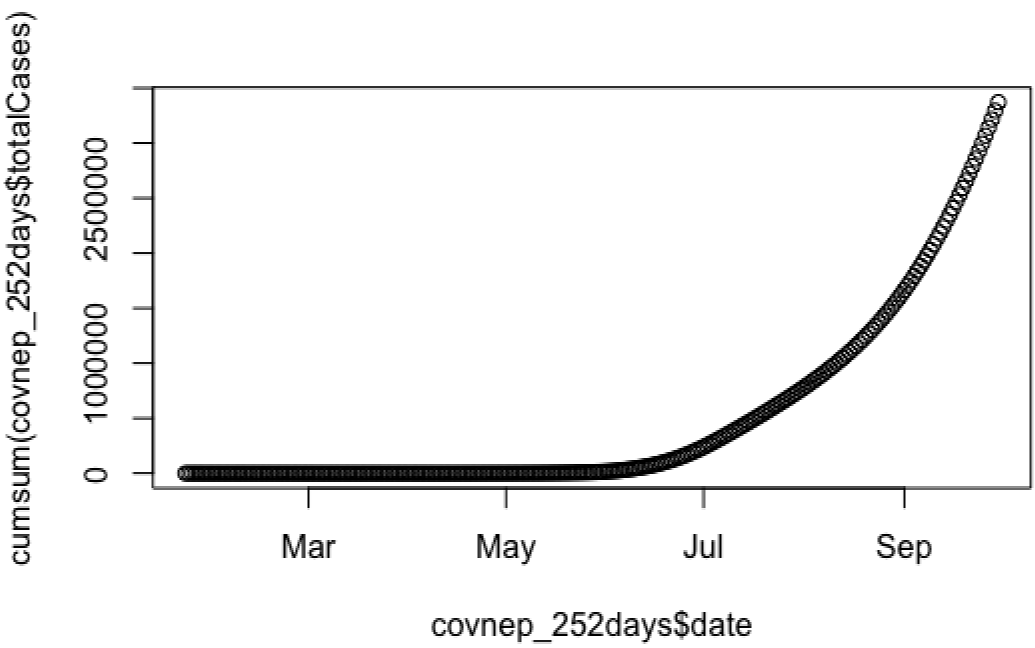
Work 2: Working with covnep_252days.csv file

#Importing the file
library(readr)
covnep_252days <- read_csv("/Users/arpan/desktop/mds/r/lab/Arpan Sapkota - covnep_252days.csv")

#Date is in character data type so need to change into date
class(covnep_252days$date)
covnep_252days$date <- as.Date(covnep_252days$date, format = "%m/%d/%Y")

#Plotting Date Vs Total Case (Cumulative Sum of total cases)
plot(covnep_252days$date,cumsum(covnep_252days$totalCases))

Cumulative COVID-19 cases in Nepal: 2020-01-23 to 2020-09-30



#Get summary of totalCases variable:
summary(covnep_252days$totalCases)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	2	963	13376	19340	77816

What is the problem with this result?
=>Summary result with Outlier

To identify potential outliers, we use the rule that any data point that falls more than 1.5 times the IQR below Q1 or above Q3 is considered an outlier. Specifically, a data point is an outlier if it satisfies either of the following conditions:

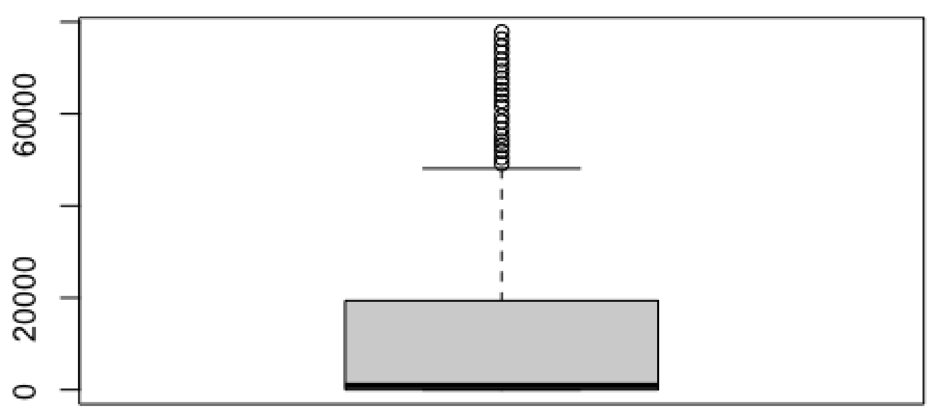
Data point < Q1 - 1.5*IQR
Data point > Q3 + 1.5*IQR

Using the summary, we can calculate the IQR and use it to identify any outliers:
Q1 = 2
Q3 = 19340
IQR = Q3 - Q1 = 19338

Lower bound = Q1 - 1.5*IQR = -28906
Upper bound = Q3 + 1.5*IQR = 38648

Potential outliers: any data point < -28906 or > 38648

Based on this analysis, any data point below -28906 or above 38648 should be considered a potential outlier. However, this is just a rule of thumb, and the presence of outliers may depend on the specific context and distribution of the data. It's also worth visualizing the data using a box plot to get a better sense of the distribution and identify any potential outliers.



Here, the box plot also shows the presence of an outlier.

Now, we fix the problem by removing the outlier and get the clean data to get the summary again.

```
#Removing the outliers from the data
bp<-boxplot.stats(covnep_252days$totalCases)
outliers <- bp$out
clean_data <- subset(covnep_252days$totalCases, !covnep_252days$totalCases %in% outliers)
```

The boxplot.stats function calculates the statistics needed to create a box plot of the data, including the lower and upper fences that define outliers. The out component of the output contains the values that are considered outliers according to the box plot.

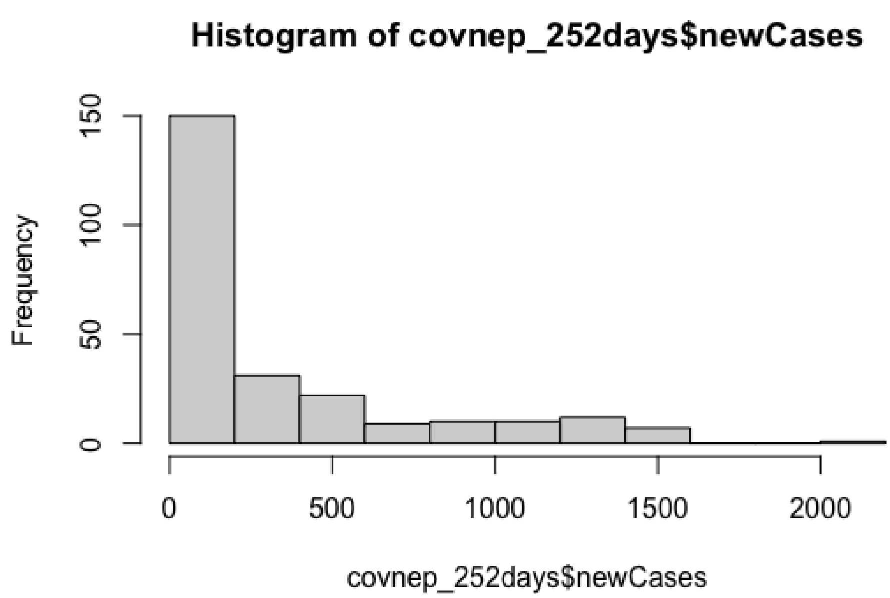
The subset function is then used to remove the outliers from the original dataset. The !covnep_252days$totalCases %in% outliers statement specifies that any values in the data vector that are not in the outliers vector should be retained.

Now we get the summary of clean data as:

```
summary(clean_data)
Min.  1st Qu.  Median    Mean   3rd Qu.    Max.
  0      0          287      8608     16908   48137
```

#Get histogram of newCases

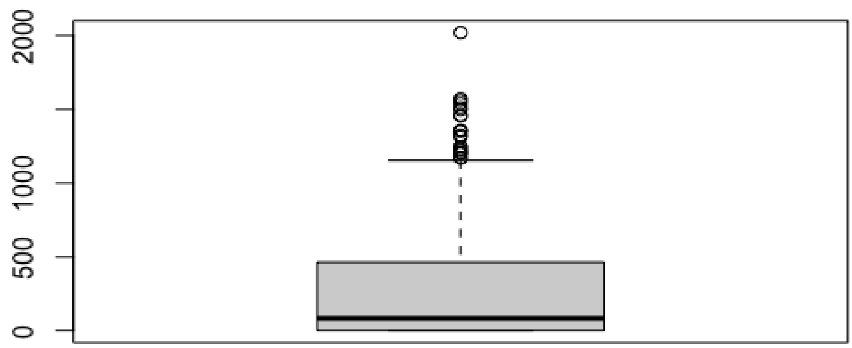
```
hist(covnep_252days$newCases)
```



#Get summary of newCases
summary(covnep_252days$newCases)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	0.0	82.5	308.8	463.2	2020.0

The box plot for newCases is shown below:



With a minimum value of 0 and a maximum value of 2020. The first quartile (25th percentile) of the data is also 0, which suggests that a significant proportion of the data is clustered around this value. The median (50th percentile) is 82.5, which indicates that the data is positively skewed or skewed to the right. The mean of 308.8 is higher than the median, which further supports the idea of positive skewness. The third quartile (75th percentile) of the data is 463.2, indicating that the majority of the data falls below this value.
This distribution appears heavily skewed to the right, with a large spread of values above the median. It is also worth noting that there may be some outliers or extreme values in the data, as suggested by the significant difference between the median and the maximum value. Which is also clearly shown in the above box plot.

—---

Work 3: Working with SAQ8.sav file

Needs a foreign library to work with this SPSS data file.

Here, the replication of the table can be done in different ways. However, the function “sjp.frq()” from the library “sjPlot” is not available in my installed version of R (4.4.2) which was the easiest way to replicate the table. When I tried to use this function by re-installing the package several times it shows a “could not find function "sjp.frq" “ error. In order to solve this problem I have used a similar functionality package called “epiDisplay”.

```
(.packages()) #list packages
#install.packages("foreign")

library(foreign)
Arpan_Sapkota_SAQ8 <- read.spss("Desktop/MDS/01 MDS I-I/MDS 503 - Statistical Computing with R/Lab/Arpan Sapkota - SAQ8.sav")

# View the current column names
names(Arpan_Sapkota_SAQ8)

# Rename the columns with their labeled attribute name
names(Arpan_Sapkota_SAQ8) <- c("Statistics makes me cry", "My friend will think I'm stupid for not being able to cope with SPSS", "Standard deviations excite me", "I dream that Pearson is attacking me with correlation coefficients", "I don't understand statistics", "I have little experience of computers", "All computers hate me", "I have never been good at mathematics")


#install.packages('epiDisplay')
library(epiDisplay)

tab1(Arpan_Sapkota_SAQ8$`Statistics makes me cry`)
```

```
> tab1(Arpan_Sapkota_SAQ8$`Statistics makes me cry`)
Arpan_Sapkota_SAQ8$`Statistics makes me cry` :
                  Frequency Percent Cum. percent
Strongly agree          270    10.5         10.5
Agree                  1338    52.0         62.5
Neither                 735    28.6         91.1
Disagree                187     7.3         98.4
Strongly disagree        41     1.6        100.0
Not answered              0     0.0        100.0
  Total                2571   100.0        100.0
> 
```

tab1(Arpan_Sapkota_SAQ8$`Standard deviations excite me`)

```
> tab1(Arpan_Sapkota_SAQ8$`Standard deviations excite me`)
Arpan_Sapkota_SAQ8$`Standard deviations excite me` :
                  Frequency Percent Cum. percent
Strongly agree          497    19.3         19.3
Agree                   672    26.1         45.5
Neither                 878    34.2         79.6
Disagree                448    17.4         97.0
Strongly disagree        76     3.0        100.0
  Total                2571   100.0        100.0
>
```

tab1(Arpan_Sapkota_SAQ8$`I have little experience of computers`)

```
> tab1(Arpan_Sapkota_SAQ8$`I have little experience of computers`)
Arpan_Sapkota_SAQ8$`I have little experience of computers` :
                  Frequency Percent Cum. percent
Strongly agree          702    27.3         27.3
Agree                  1127    43.8         71.1
Neither                 344    13.4         84.5
Disagree                252     9.8         94.3
Strongly disagree       146     5.7        100.0
  Total                2571   100.0        100.0
>
```

tab1(Arpan_Sapkota_SAQ8$`I have never been good at mathematics`)

```
> tab1(Arpan_Sapkota_SAQ8$`I have never been good at mathematics`)
Arpan_Sapkota_SAQ8$`I have never been good at mathematics` :
                  Frequency Percent Cum. percent
Strongly agree          383    14.9         14.9
Agree                  1487    57.8         72.7
Neither                 482    18.7         91.5
Disagree                147     5.7         97.2
Strongly disagree        72     2.8        100.0
  Total                2571   100.0        100.0
>
```

—---

Work 4: Working with MR_drugs.xls file

Needs a readxl library to work with this xlsx data file.

Not able to find a single function that replicates the multiple response frequencies, So, followed the manual process to replicate the table.

```
# Calculate the multiple response frequencies for all the income columns

#install.packages("readxl")
library(readxl)

#Loading the xls file
Arpan_Sapkota_MR_Drugs <- read_excel("Desktop/MDS/01 MDS I-I/MDS 503 - Statistical Computing with R/Lab/Arpan Sapkota - MR_Drugs.xlsx")

#install.packages("readxl")
library(readxl)

head(Arpan_Sapkota_MR_Drugs)
#names(Arpan_Sapkota_MR_Drugs[4:10])

# Calculate the counts and percentages for each income variable
incomes <- c("inco1", "inco2", "inco3", "inco4", "inco5", "inco6", "inco7")
counts <- sapply(Arpan_Sapkota_MR_Drugs[incomes], sum)
percentages <- round(counts/sum(counts) * 100, 1)


# Create a data frame with the counts and percentages
income_freq <- data.frame(
  Income = incomes,
  Frequencies = counts,
  Percent = percentages,
  `Percent.of.Cases` = round(percentages / 100 * 182.9, 1)
)


# Add a row for the total count and percentage
income_freq <- rbind(
  income_freq,
  c("Total", sum(counts), 100, 182.9)
)
```

Print the table
Income_freq

```
> # Print the table
> income_freq
      Income Frequencies Percent Percent.of.Cases
inco1  inco1         226    12.8             23.4
inco2  inco2         607    34.5             63.1
inco3  inco3         293    16.6             30.4
inco4  inco4          50     2.8              5.1
inco5  inco5          82     4.7              8.6
inco6  inco6         151     8.6             15.7
inco7  inco7         352      20             36.6
8      Total        1761     100            182.9
> 
```

—---

Please Check My GitHub repository link below for the full R code compiled from R Studi:

https://github.com/arpansapkota/Statistical-Computing-with-R/blob/main/03_Variable_and_Data_Exploration.R