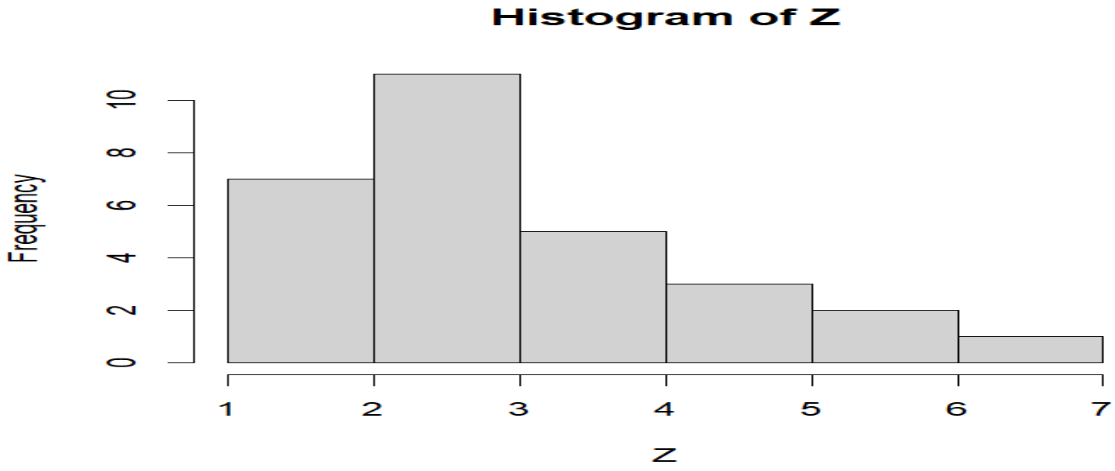
Complete these works showing your codes and outputs from R studio:

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

```
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)
```



```
> summary(Z)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   3.000   3.000   3.414   4.000   7.000 
```

Measure of central tendency:

After Seeing the histogram and summary, the median would be a more appropriate measure of central tendency than the mean(i.e. only we used the mean when the histogram shape of data is bell shaped curve). Median separates the data into two equal halves and In this case, its average accuracy is good. To find the median we can also use the **median()** function separately.

median(Z)

console:

```
> median(Z)
[1] 3
```

It indicates that half of the values are below median 3 and half are above. Therefore, the median is the good choice to measure the central tendency for this data.

Measure of dispersion:

To find the dispersion for this data , the interquartile range (IQR) is best for this data.To find the Find the Interquartile range we use the IQR**()** function , where we pass variables as arguments to this function to find the dispersion.

Ex:

IQR(Z)

```
> IQR(Z)
[1] 1
```

It indicates that data is spread over a range of 1 unit.

Five number summary

```
> summary(Z)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   3.000   3.000   3.414   4.000   7.000 
```

The five-number summary provides us with a quick overview of the central tendency and spread of the data. We can see that the median value is 3.0, indicating that half of the values in the dataset are below 3.0 and half are above. The mean value is slightly higher than the median, which suggests that the distribution is slightly skewed to the right due to the presence of a few large values.

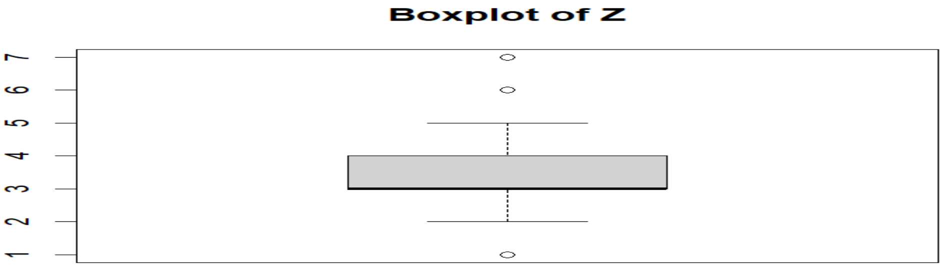
Here, the output shows the minimum value of Z is 1, the first quartile (25th percentile) is 3, the median (50th percentile) is also 3, the mean is 3.414, the third quartile (75th percentile) is 4, and the maximum value is 7.This suggests that the values of Z are positively skewed, as the mean is greater than the median.

Boxplot:

To create a box plot, we use the boxplot() function.

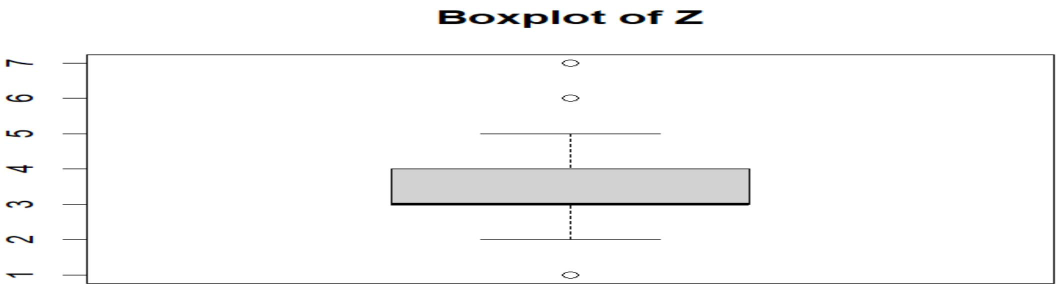
Ex:

boxplot(Z)



The boxplot is a graphical representation of the five-number summary (minimum, first quartile, median, third quartile, and maximum) and any outliers in the data.

Outlier:



In the box plot, there are two points beyond the whiskers (the vertical lines extending from the boxes) on the right-hand side of the plot, indicating values that are significantly higher than the majority of the data points. These points are commonly referred to as outliers.

outlier->Median + 1.5 * IQR

= 3 + 1.5 * 1 = 4.5.

Therefore, any value greater than 4.5 is considered an outlier.

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

To import the csv file to r studio from computer, following code should be execute

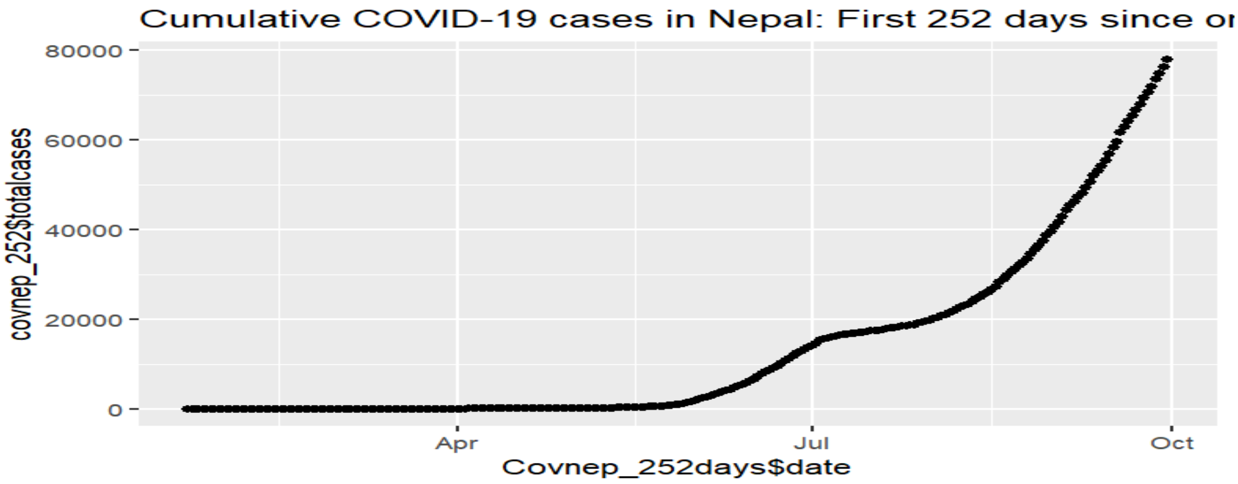
CODE:

```
 data <-read.csv(file.choose(), header = TRUE, sep = ",")
 data1<-data.frame(data)
```

Then to get the chart in R studio, I installed lubridate and ggplot2. Lubridate is for parsing date-time strings, extracting parts of dates and times, manipulating time zones, and performing arithmetic operations on dates and times. Ggplot2 is used for creating a wide variety of visualizations, including scatterplots, bar charts, histograms, and more.

CODE:

```
library(lubridate)
data1$date<-mdy(data1$date)
data1
library(ggplot2)
data1
class(data1$date)
ggplot(data1, aes(x = date, y = totalCases)) +
  geom_point() +
  labs(x = "Covnep_252days$date", y = "covnep_252$totalcases", title = "Cumulative COVID-19 cases in Nepal: First 252 days since onset at 23/01/2021")
```



```
summary(data1$totalCases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0       2     963   13376   19341   77816 
```

What is the problem with this data?

The issue that I found here is that the range between minimum and maximum is large.The majority of the data falls below the mean, with a relatively small number of outliers at the end of the distribution. It's important to note that the mean is heavily influenced by these outliers, so the median may be a better measure in this case.

Fix the summary and get this again.

Ans: In the total cases column of records, have seen some zero after seeing some cases before. So, to fix the problem for getting the accurate result , first we need to add number 1 in the column of total cases from 2 to 60   and then get the summary again .

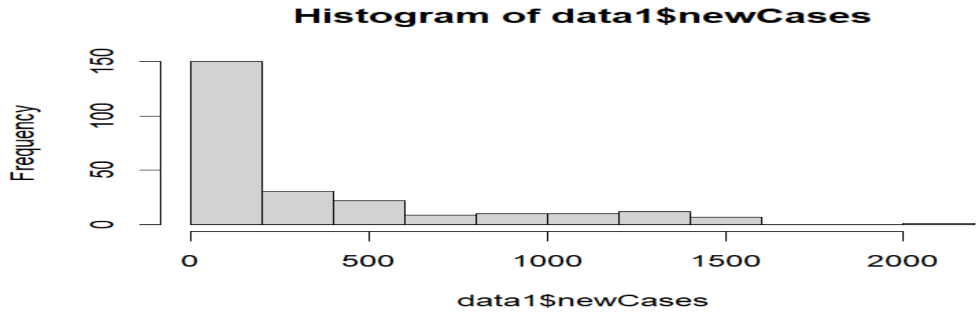
```
data$totalCases[2:60]<-1
head(data)
summary(data$totalCases)
```

```
> summary(data$totalCases)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      1       2     963   13377   19341   77816 
```

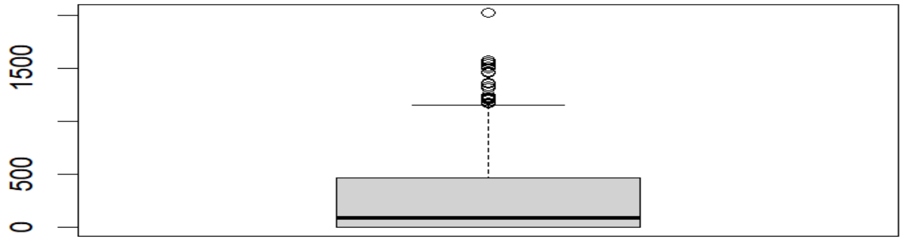
Then for the histogram chart and summary of newcases:

In this datasheet , it is found that some rows of all columns contain only zero values .Therefore, it is necessary to remove those rows to get a more accurate result.The following code can be used for this problem.

```
data2 <- data1[rowSums(data1[, 2:ncol(data)]) > 0, ]
hist(data1$newCases)
```



```
> summary(data1$newCases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0     0.0    82.5   308.8   463.2  2020.0 
boxplot(data1$newCases) #code to see outlier in new cases
```



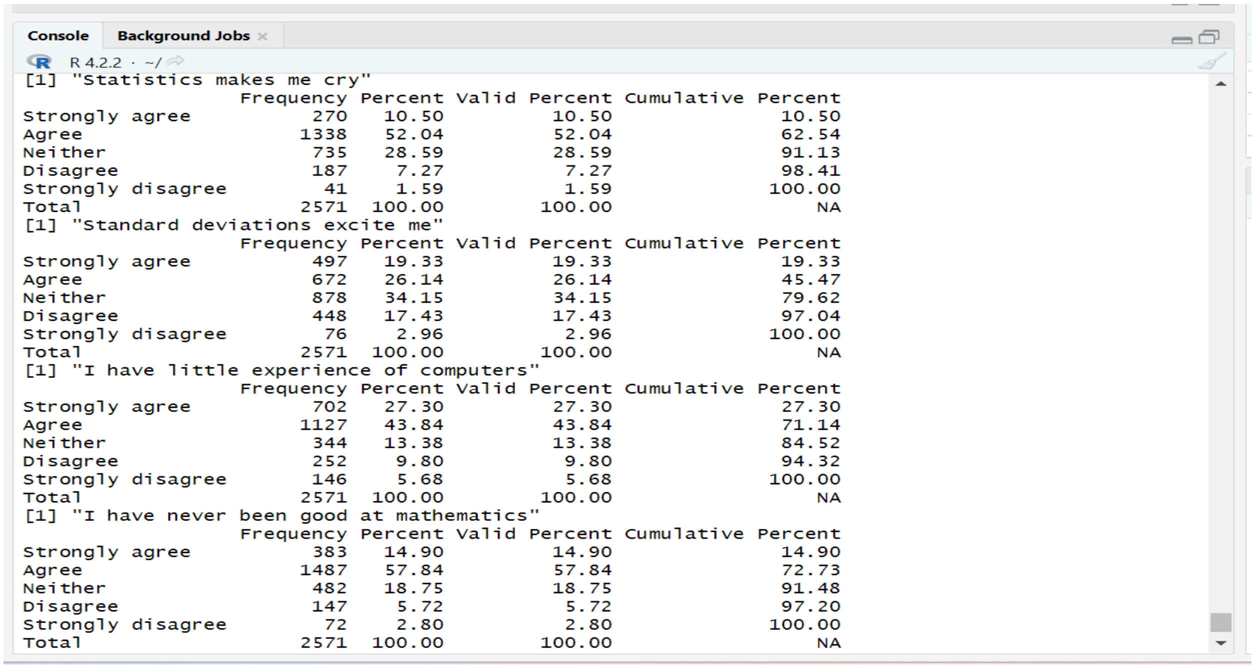
Work 3: See slide 31 of session 2 slide deck and provide answers here. Data is attached.

Code:

```
library("haven")
data.frame <- read_sav(file.choose())

freq <- table(data.frame["q01"])
percent <- freq / sum(freq) * 100
valid_per <- freq / sum(!is.na(data.frame["q01"])) * 100
cum_per <- cumsum(percent)
```

```
result_df <- cbind(freq, percent, valid_per, cum_per)
total <- c(colSums(result_df[,-4]), NA)
result_df <- round(result_df, 2)
result_df <- rbind(result_df, total)
rownames(result_df) <- c("Strongly agree", "Agree", "Neither", "Disagree",
                        "Strongly disagree", "Total")
colnames(result_df) <- c("Frequency", "Percent", "Valid Percent",
                        "Cumulative Percent")
features <- list("q01", "q03", "q06", "q08")
for (feature_name in features) {
  print(attributes(data.frame[[feature_name]])$label)
  freq <- table(data.frame[feature_name])
  percent <- freq / sum(freq) * 100
  valid_per <- freq / sum(!is.na(data.frame[feature_name])) * 100
  cum_per <- cumsum(percent)
  result_df <- cbind(freq, percent, valid_per, cum_per)
  total <- c(colSums(result_df[,-4]), NA)
  result_df <- round(result_df, 2)
  result_df <- rbind(result_df, total)
  rownames(result_df) <- c("Strongly agree", "Agree", "Neither", "Disagree",
                          "Strongly disagree", "Total")
  colnames(result_df) <- c("Frequency", "Percent", "Valid Percent",
                       "Cumulative Percent")
  print(result_df)
}
```

Output

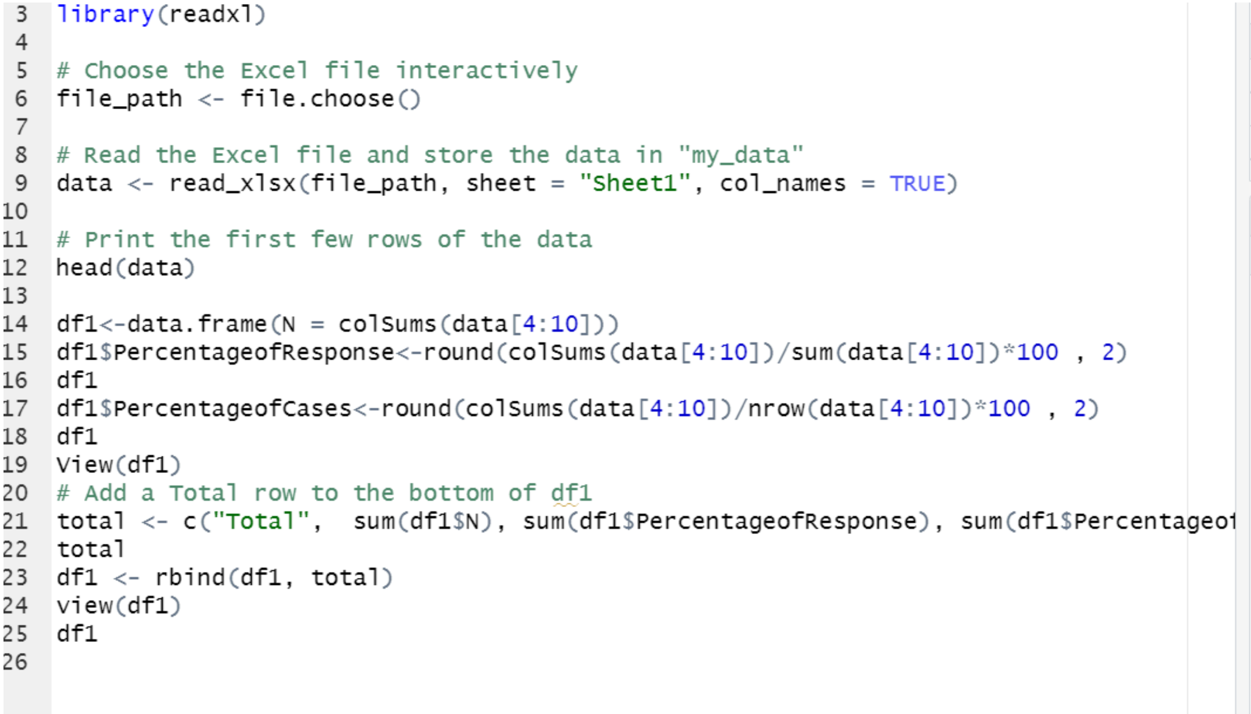


```
[1] "Statistics makes me cry"
                  Frequency Percent Valid Percent Cumulative Percent
Strongly agree          270   10.50         10.50              10.50
Agree                  1338   52.04         52.04              62.54
Neither                 735   28.59         28.59              91.13
Disagree                187    7.27          7.27              98.41
Strongly disagree        41    1.59          1.59             100.00
Total                  2571  100.00        100.00                 NA
[1] "Standard deviations excite me"
                  Frequency Percent Valid Percent Cumulative Percent
Strongly agree          497   19.33         19.33              19.33
Agree                   672   26.14         26.14              45.47
Neither                 878   34.15         34.15              79.62
Disagree                448   17.43         17.43              97.04
Strongly disagree        76    2.96          2.96             100.00
Total                  2571  100.00        100.00                 NA
[1] "I have little experience of computers"
                  Frequency Percent Valid Percent Cumulative Percent
Strongly agree          702   27.30         27.30              27.30
Agree                  1127   43.84         43.84              71.14
Neither                 344   13.38         13.38              84.52
Disagree                252    9.80          9.80              94.32
Strongly disagree       146    5.68          5.68             100.00
Total                  2571  100.00        100.00                 NA
[1] "I have never been good at mathematics"
                  Frequency Percent Valid Percent Cumulative Percent
Strongly agree          383   14.90         14.90              14.90
Agree                  1487   57.84         57.84              72.73
Neither                 482   18.75         18.75              91.48
Disagree                147    5.72          5.72              97.20
Strongly disagree        72    2.80          2.80             100.00
Total                  2571  100.00        100.00                 NA
```

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

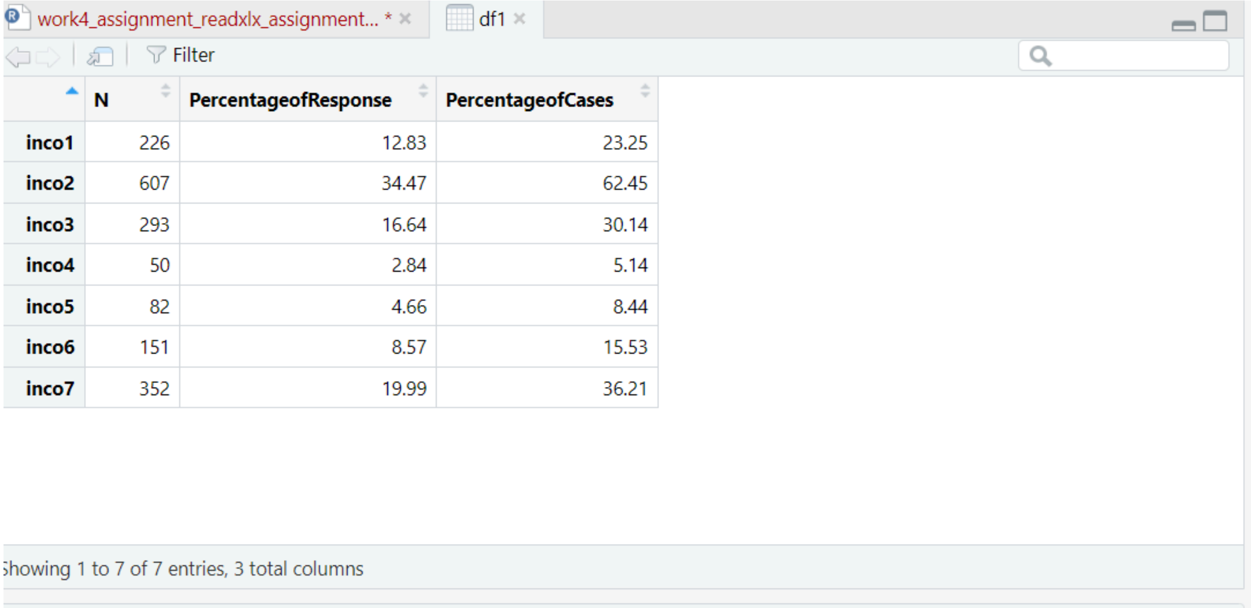
Code:

```
# Load the "readxl" package
library(readxl)
# Choose the Excel file interactively
file_path <- file.choose()
# Read the Excel file and store the data in "my_data"
data <- read_xlsx(file_path, sheet = "Sheet1", col_names = TRUE)
# Print the first few rows of the data
head(data)
df1<-data.frame(N = colSums(data[4:10]))
df1
df1$PercentageofResponse<-round(colSums(data[4:10])/sum(data[4:10])*100 , 2)
df1
df1$PercentageofCases<-round(colSums(data[4:10])/nrow(data[4:10])*100 , 2)
# Add a Total row to the bottom of df1
total <- c("Total",  sum(df1$N), sum(df1$PercentageofResponse), sum(df1$PercentageofCases)
)
total
df1 <- rbind(df1, total)
View(df1)
```



```r
library(readxl)

# Choose the Excel file interactively
file_path <- file.choose()

# Read the Excel file and store the data in "my_data"
data <- read_xlsx(file_path, sheet = "Sheet1", col_names = TRUE)

# Print the first few rows of the data
head(data)

df1<-data.frame(N = colSums(data[4:10]))
df1$PercentageofResponse<-round(colSums(data[4:10])/sum(data[4:10])*100 , 2)
df1
df1$PercentageofCases<-round(colSums(data[4:10])/nrow(data[4:10])*100 , 2)
df1
View(df1)
# Add a Total row to the bottom of df1
total <- c("Total",  sum(df1$N), sum(df1$PercentageofResponse), sum(df1$Percentageof
total
df1 <- rbind(df1, total)
view(df1)
df1

```

OutPut:



	N	PercentageofResponse	PercentageofCases
inco1	226	12.83	23.25
inco2	607	34.47	62.45
inco3	293	16.64	30.14
inco4	50	2.84	5.14
inco5	82	4.66	8.44
inco6	151	8.57	15.53
inco7	352	19.99	36.21

Showing 1 to 7 of 7 entries, 3 total columns