ORGANISING AND CONDENSING DATA - HOME- WORK

Ogada Joseph Ridge - 166895

```
# check if library already installed and install if does not exist then import
installedStatus <- any(grepl("ggplot2", installed.packages()))
if(!installedStatus){
install.packages("ggplot2")  # Install ggplot2 package if not already installed
}
library(ggplot2)  # Load the ggplot2 package
```

Installing of libraries -> ggplot2

1. Every week the number of missed calls was recorded in a certain company switchboard. This was done for 50 weeks and the observations were tabulated as shown below;

Number of missed calls	**Number of weeks**
11-15	2
16-20	3
21-25	4
26-30	11
31-35	14
36-40	6
41-45	7
46-50	2
51-55	1

sln.

```
# data
numberOfMissedCalls <- c("11-15", "16-20", "21-25", "26-30", "31-35", "36-40", "41-45", "46-50", "51-55
numberOfWeeks <- c(2, 3, 4, 11, 14, 6, 7, 2, 1)
cummulativeFrequency <- cumsum(numberOfWeeks)# Calculate the cumulative frequency

# display inform of a table -> data-frame
data <- data.frame(
  numberOfMissedCall = numberOfMissedCalls,
  numberOfWeeks = numberOfWeeks,
  cumulativeFrequency =cummulativeFrequency
)

data
```

```
##   numberOfMissedCall numberOfWeeks cumulativeFrequency
## 1              11-15             2                   2
## 2              16-20             3                   5
## 3              21-25             4                   9
## 4              26-30            11                  20
## 5              31-35            14                  34
## 6              36-40             6                  40
## 7              41-45             7                  47
## 8              46-50             2                  49
## 9              51-55             1                  50
```

```
# Plot the ogive graph using ggplot2
ggplot(data, aes(x = numberOfMissedCalls, y = cummulativeFrequency,group=1)) +
  geom_line() +
  geom_point() +
  labs(x = "Number of Missed Calls", y = "Cumulative Frequency" )
```

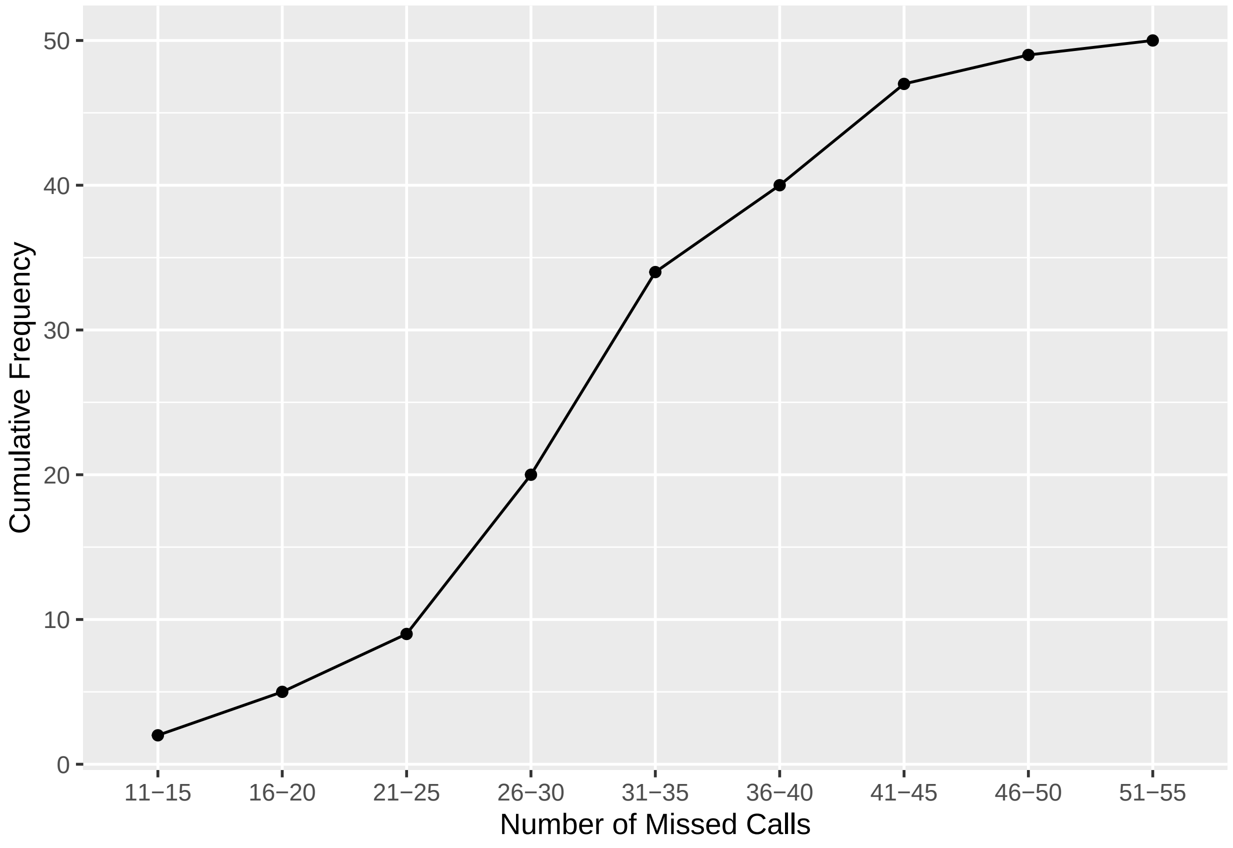


Figure 1. Number of Missed Calls against frequency - Orgive Graph

2. The examination marks in a statistics test, for 72 students were as follows:

Data : 36 62 15 28 60 30 25 35 75 14 16 33 58 72 80 92 44 57 55 56 70 34 40 18 15 28 32 60 57 68 83 30 32 40 38 45 48 52 58 62 65 75 84 63 58 47 55 60 38 25 18 29 15 32 35 38 45 43 46 53 58 64 68 54 59 60 61 64 65 70 78 90

goal: Using a class interval of 10, that is 10 – 19; 20 – 29… make a frequency distribution table.

sln.

```
# create vector for the students marks
data <- c(36, 62, 15, 28, 60, 30, 25, 35, 75, 14, 16, 33, 58, 72, 80, 92, 44, 57, 55, 56, 70, 34, 40, 1

# Step 1: sort data in ascending order
data <- sort(data)
# get length of vector and assign to N
N <- length(data)

# Step 2: unit of measurement - u (get unique  value then sort it based on unique values  difference be
diffLargeNsmall <- diff(sort(unique(data)))
# get smallest difference
smallDiff <- min(diffLargeNsmall)
u<- smallDiff

# Step 3: obtain range (R) = MaxValue - MinValue
R <- max(data) - min(data)

#step 4:  Determine the number of classes k = 1 + 3.322*log10(N)
k <- 1 + 3.322*(log10(N))
# rounded up so as to have enough number of classes
k<- ceiling(k)

# Step 5: Specify class width, according to the instructions we set w as 10
w = R/k
w <- 10
# Step 5: set the LCL
LCL <- data[1]

# Step 6: Get the other classes, adding the width to the LCL until we get to k iterations
smallestVal <- data[1] # this is taken as the LCL
largetsVal <- data[N]

# Create a sequence by adding w we will include the last value of the sequence,  breaks returns the fir
breaks <- seq(smallestVal, largetsVal+w , w)
#breaks
#breaks[-length(breaks)] # excludes the last/largest value since it is in order
#breaks[-1] - 1 #excludes  first element in position 1 then prints our sequence having reduce each by 1

# Create the labels using paste0 for concatenation -> these are inclusive grouped class limits
classLabels <- paste0("", breaks[-length(breaks)], "-", breaks[-1] - 1, "")


# creating frequency distribution table, note: right=false makes sure we include the last value in the 
frequency <- table(cut(data ,breaks,right = FALSE)) # cut -> converts numeric to a
cummulativeFreq =  cumsum(as.vector(frequency))
relativeFreq = as.vector((frequency/N)*100)
frequency <- as.vector(frequency)
#creating a data frame
dataFrame <- data.frame(
  class= classLabels,
  f =  frequency,
  cf = cummulativeFreq,
  rf   = relativeFreq
)
```

Table 1. Examination Marks in a Statistics Test

```
dataFrame
```

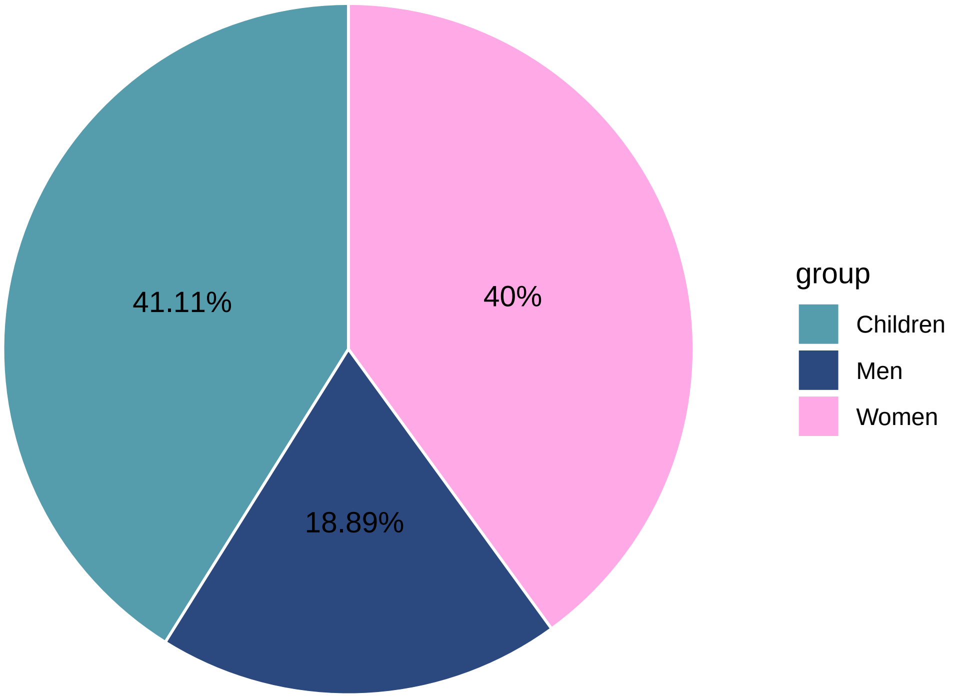
```
##   class  f cf        rf
## 1 14-23  7  7  9.722222
## 2 24-33 11 18 15.277778
## 3 34-43 10 28 13.888889
## 4 44-53  8 36 11.111111
## 5 54-63 19 55 26.388889
## 6 64-73  9 64 12.500000
## 7 74-83  5 69  6.944444
## 8 84-93  3 72  4.166667
```

3. The number of people who attended an agricultural show in one day was 510 men, 1080 women and some children. When the information was represented on a pie chart, the combined angle for the men and children was 216. Find the angle representing the children.

sln.

```
numberOfMen <-510
numberOFWomen <- 1080
totalDegrees <- 360
degreeOfChildrenNmen <- 216

# pie Chart % rep of Men + Children = 216
degreeOfWomen <- 360 - 216 # meaning women rep 144 degrees
#pie Allocation = (number/total Population) * 360
totalPopulationAttended <- numberOFWomen *360/degreeOfWomen
numberOfChildren <- (totalPopulationAttended - (numberOfMen + numberOFWomen))
degreeOfChildren <- numberOfChildren * 360/ totalPopulationAttended

degreeOfMen <- degreeOfChildrenNmen - degreeOfChildren

degrees <- data.frame(
  children = degreeOfChildren,
  women = degreeOfWomen,
  men = degreeOfMen)

group <- c("Women","Children", "Men")
deg <- c(degreeOfWomen, degreeOfChildren,degreeOfMen)

df <- data.frame(deg = deg,
                 group = group)
# convert to percentage
childrenPercentage <- signif(((numberOfChildren/totalPopulationAttended)*100),4)
childrenPercentage <- paste0(childrenPercentage,"%")
```

```
menPercentage <- signif(((numberOfMen/totalPopulationAttended)*100),4)
menPercentage<- paste0(menPercentage,"%")
womenPercentage <- signif(((numberOFWomen/totalPopulationAttended)*100),4)
womenPercentage <- paste0(womenPercentage,"%")

percentages <- c(womenPercentage,childrenPercentage,menPercentage)

#plot graph
ggplot(df, aes(x = "", y = deg, fill = group)) +
  geom_col(color = "white") +
  geom_text(aes(label = percentages),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("#559CAD", "#2C497F", "#FFA9E7"))+
  theme_void()
```



```
# angle representing children
paste0("Children are represented by ",degreeOfChildren,"°, in the pie-chart")
```

```
## [1] "Children are represented by 148°, in the pie-chart"
```

Figure 1. Population Attended An Agricultural Show

4. You have conducted a market survey with a sample size of 50 regarding the acceptability of a new product that your company wants to launch. The scores of the respondents on the appropriate scale are as follows:

Data : 40 45 41 45 45 30 39 8 48 25 26 9 23 24 26 29 8 40 41 42 39 35 18 25 35 40 42 43 44 36 27 32 28 27 25 26 38 37 36 35 32 28 40 41 43 44 45 40 39 41

Goal: Prepare a Stem – and – leaf plot of the distribution.

sln.

```
#create a vector
data <- c(40, 45, 41, 45, 45, 30, 39, 8, 48, 25, 26, 9, 23, 24, 26, 29, 8, 40, 41, 42, 39, 35, 18, 25,
data <- sort(data)
```

```
stem(data)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 889
##   1 |
##   1 | 8
##   2 | 34
##   2 | 55566677889
##   3 | 022
##   3 | 5556678999
##   4 | 000001111223344
##   4 | 55558
```

Figure 2. Product Market Acceptance Survey

5. Plot a frequency polygon for the data below.

Marks	F
5-9	2
10-14	5
15-19	3
20-24	7
25-29	12
30-34	10
35-39	6
40-44	4

sln.

```
classInterval <- c("5-9", "10-14", "15-19", "20-24", "25-29", "30-34", "35-39", "40-44")
marks <- c(2, 5, 3, 7, 12, 10, 6, 4)
cumulativeFrequency <- cumsum(marks)  # Calculate the cumulative frequency

dataFrame <- data.frame(
  class = classInterval,
  f = marks,
  cf = cumulativeFrequency
)
 # ggplot operating in layers, we first set the x n y aesthetics or data then we add the points, and fi
ggplot(dataFrame, aes(x = classInterval, y = marks, group = 1)) +
    geom_point()+
    geom_line()+
   labs(x = "marks", y = "frequency")
```

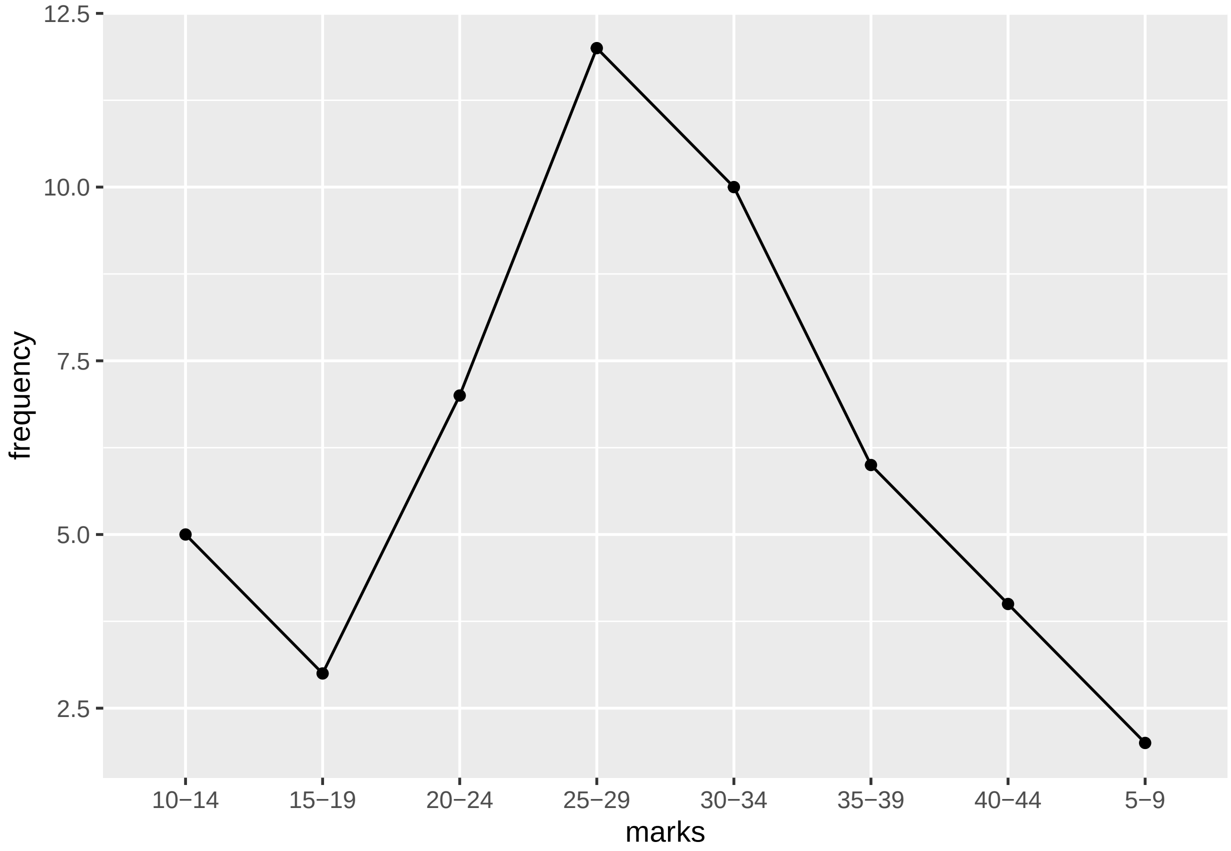


Figure 3. Frequency Polygon based on Marks achieved

6. A sample consists of 34 observations recorded correct to the nearest integer, ranging in value from 201 to 337. If it is decided to use seven classes of width 20 integers and to begin the first class limit at 199.5, find the class limits and the class mid – points of the seven classes.

Goal: - Find Class limits - Class Mid-points

sln

```
# total number of observations
N <-34
# lower class boundary
LCB <-199.5
#smallest value
```

```
smallVal <- LCB
#largest value
largeVal <- 337
# width (stated from above instructions)
w <- 20
# number of classes(stated from above instructions)
k <- 7

# attain the lower class boundaries
breaks <- seq(smallVal, largeVal+w , w)

# concatenate into classes
classLabels <- paste0("", breaks[-length(breaks)], "-", breaks[-1] - 1, "")
classLabels <- as.vector(classLabels)

#mid points => (LCB +UCB)/2
midPoints<-paste0(breaks[-length(breaks)])
midPoints <- as.vector((breaks[-length(breaks)] +(breaks[-1]-1))/2)

dataFrame <- data.frame(
  class= classLabels,
  midPoints = midPoints
)
```

Table 2. Observations with their respective mid-points

```
dataFrame
```

```
##         class midPoints
## 1 199.5-218.5       209
## 2 219.5-238.5       229
## 3 239.5-258.5       249
## 4 259.5-278.5       269
## 5 279.5-298.5       289
## 6 299.5-318.5       309
## 7 319.5-338.5       329
```

7. In a sample study about coffee habit in two towns, the following information was received:

Town A: Females were 40%; Total coffee drinkers were 45% and male non – coffee drinkers were 20%. Town B: Males were 55%; Male non – coffee drinkers were 30% and female coffee drinkers were 15%. Represent the above data in tabular form.

sln.

```
# Town A and B - observations are in percentage

data <- data.frame(
  Town = c("A", "A", "B", "B"),
  Gender = c("Female", "Male", "Female", "Male"),
  coffeeDrinkers = c(0.40, 0.05, 0.15, 0.30),
  nonCoffeeDrinkers = c(0.60, 0.95, 0.85, 0.70)
)
```

Table 3. Combination of Coffee drinking Habits in Town A and Town B based

```
data
```

```
##   Town Gender coffeeDrinkers nonCoffeeDrinkers
## 1    A Female           0.40              0.60
## 2    A   Male           0.05              0.95
## 3    B Female           0.15              0.85
## 4    B   Male           0.30              0.70
```

8. Represent the following data by means of a histogram

Weekly wages(in 00Ksh)	No. of workers
10-15	7
15-20	19
20-25	27
25-30	15
30-40	12
40-60	12
60-80	8

sln.

```
suppressWarnings({
# create class vector ->weekly wages
class <- c("10-15", "15-20","20-25", "25-30", "30-40","40-60", "60-80")
numOfWorker <- c(7,19,27,15,12,12,8)

dataFrame <- data.frame(
  class = class,
  numOfWorker = numOfWorker
)
#using ggplot we layer till we plot our histogram
ggplot(dataFrame, aes(x=class, y=numOfWorker  )) +
 geom_histogram(stat="identity",fill = "#559CAD")+
  labs(x = "Weekly Wages ( in OOKsh. )", y = "No. of Workers")
})
```

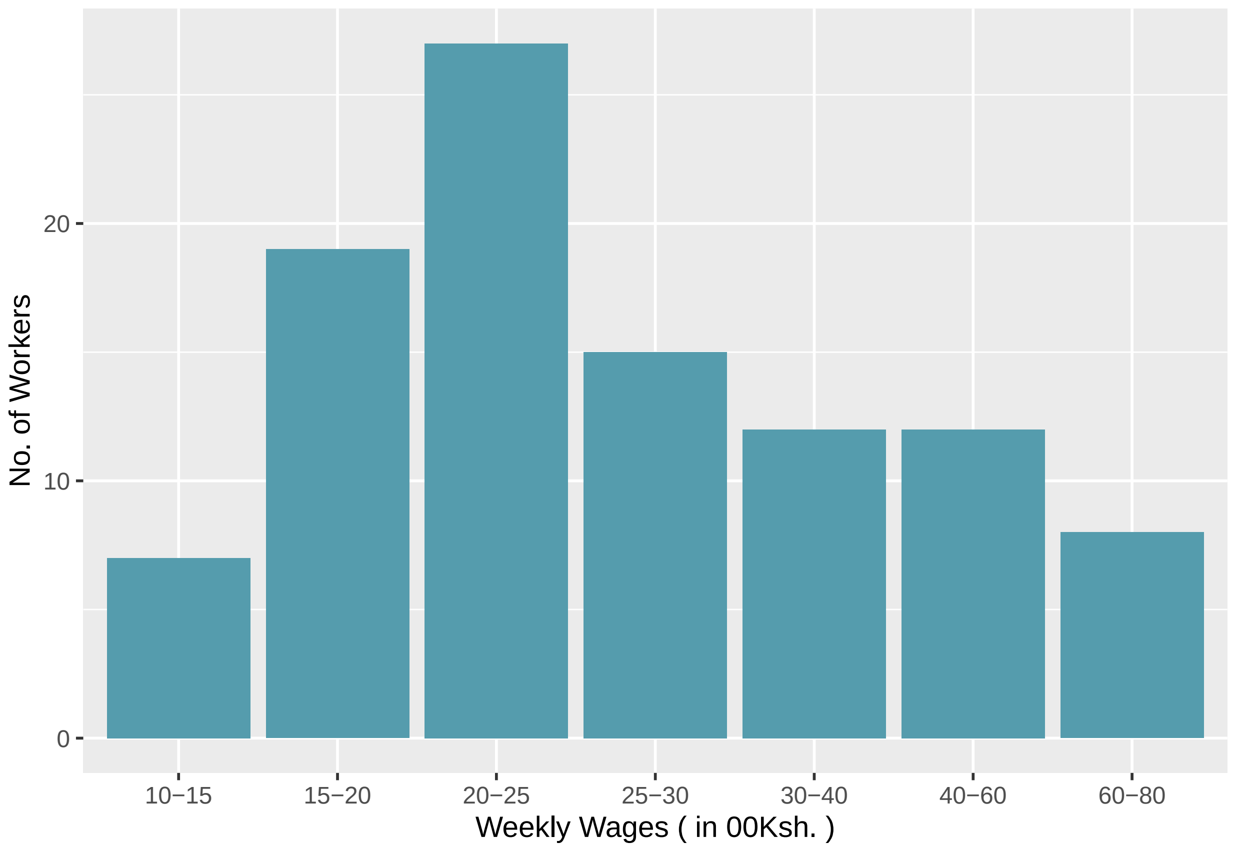


Figure 4. Weekly Wages against No. Of Workers