PES University, Bengaluru

UE18CS312 - Data Analytics

Session: Aug – Dec 2020

Weeks 1-2 – Code Snippets for Worksheet 1(a) (for Unit 1)

Dataset: BKB.csv
Source: Business Analytics, U. Dinesh Kumar
Libraries : *ggplot2, dplyr, plyr, corrplot ,e1071*
R Basics: The R Project for Statistical Computing: R
Relevant Courses/Content: Chapters 1-6 of the prescribed textbook
Udemy
CRAN
R Programming for Data Science Roger D Peng

Compiled by: Ms. Bharani Ujjaini Kempaiah, Mr. Ruben John and Ms. Bhavya Charan VII CSE, PES University RR Campus.

Name – B.Pravena
SRN – PES2UG19CS076
Section - B

Code Snippets

Getting started

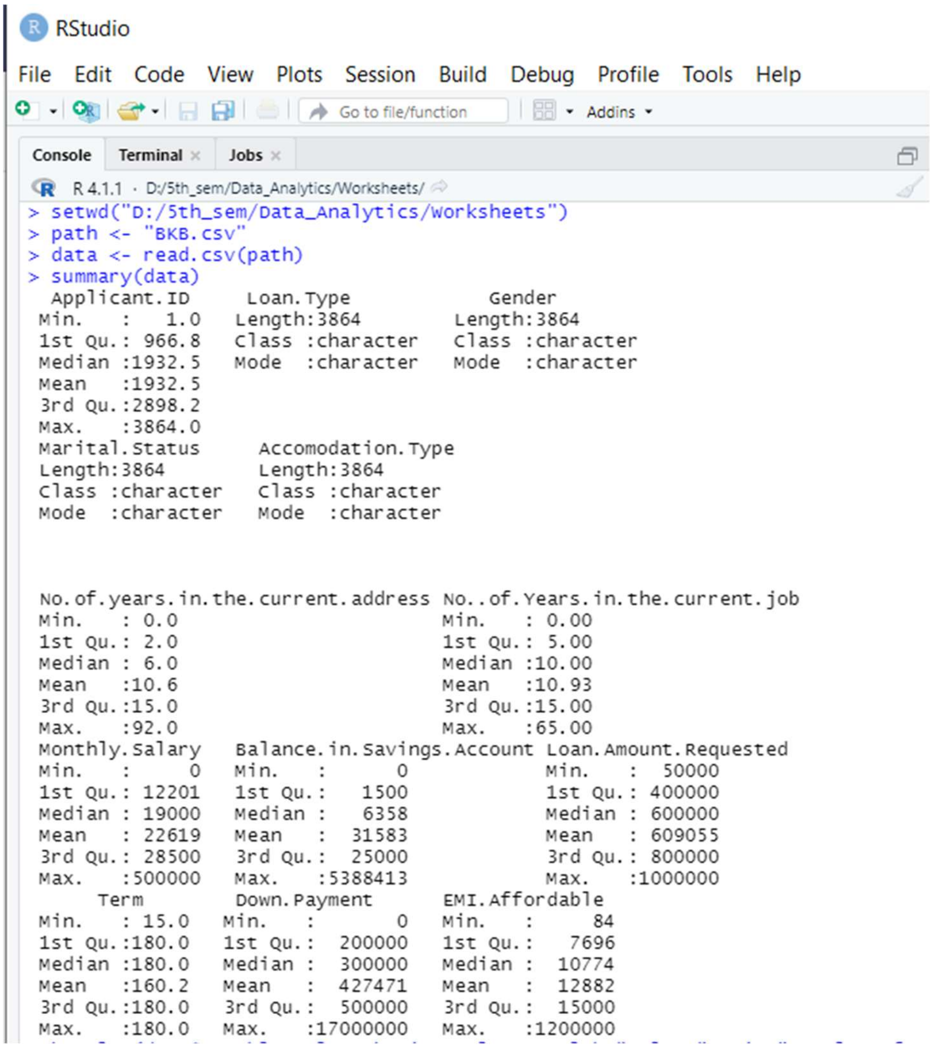
1. Read the BKB.csv dataset

```
path <- "BKB.csv"
data <-
read.csv(path)
```

2. Find a basic summary of the data

```
summary(data)
```

When applied to a data frame, the summary() function is essentially applied to each column and the results for all columns are shown together. For a continuous (numeric) variable like “Monthly.Salary”, it returns the 5-number summary. If there are any missing values (denoted by “NA”), it would also provide a count for them. In this example, there are no missing values, so there is no display for the number of NA’s .For a categorical variable like “Gender”, it returns the levels and the number of data in each level



```
> setwd("D:/5th_sem/Data_Analytics/worksheets")
> path <- "BKB.csv"
> data <- read.csv(path)
> summary(data)
  Applicant.ID     Loan.Type            Gender         
 Min.   :   1.0   Length:3864        Length:3864       
 1st Qu.: 966.8   Class :character   Class :character  
 Median :1932.5   Mode  :character   Mode  :character  
 Mean   :1932.5                                        
 3rd Qu.:2898.2                                        
 Max.   :3864.0                                        
 Marital.Status     Accomodation.Type 
 Length:3864        Length:3864       
 Class :character   Class :character  
 Mode  :character   Mode  :character  



 No.of.years.in.the.current.address No..of.Years.in.the.current.job
 Min.   : 0.0                        Min.   : 0.00                  
 1st Qu.: 2.0                        1st Qu.: 5.00                  
 Median : 6.0                        Median :10.00                  
 Mean   :10.6                        Mean   :10.93                  
 3rd Qu.:15.0                        3rd Qu.:15.00                  
 Max.   :92.0                        Max.   :65.00                  
 Monthly.Salary   Balance.in.Savings.Account Loan.Amount.Requested
 Min.   :     0   Min.   :      0            Min.   :  50000      
 1st Qu.: 12201   1st Qu.:   1500            1st Qu.: 400000      
 Median : 19000   Median :   6358            Median : 600000      
 Mean   : 22619   Mean   :  31583            Mean   : 609055      
 3rd Qu.: 28500   3rd Qu.:  25000            3rd Qu.: 800000      
 Max.   :500000   Max.   :5388413            Max.   :1000000      
      Term        Down.Payment       EMI.Affordable   
 Min.   : 15.0   Min.   :       0   Min.   :     84  
 1st Qu.:180.0   1st Qu.:  200000   1st Qu.:   7696  
 Median :180.0   Median :  300000   Median :  10774  
 Mean   :160.2   Mean   :  427471   Mean   :  12882  
 3rd Qu.:180.0   3rd Qu.:  500000   3rd Qu.:  15000  
 Max.   :180.0   Max.   :17000000   Max.   :1200000  
```

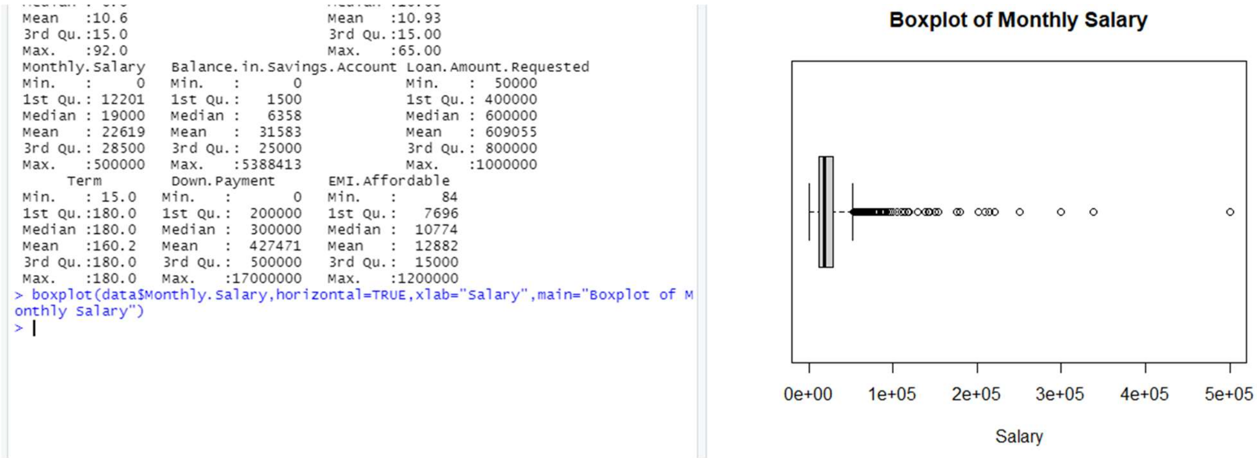
Descriptive Statistics

3. Are there any outliers in these variables? Plot a box and whisker plot to find out.

Given below is a sample for the Monthly.Salary attribute

```
boxplot(data$Monthly.Salary,horizontal=TRUE,xlab="Salary",main=
"B
oxplot of Monthly Salary")
```

<u>Output</u> -:



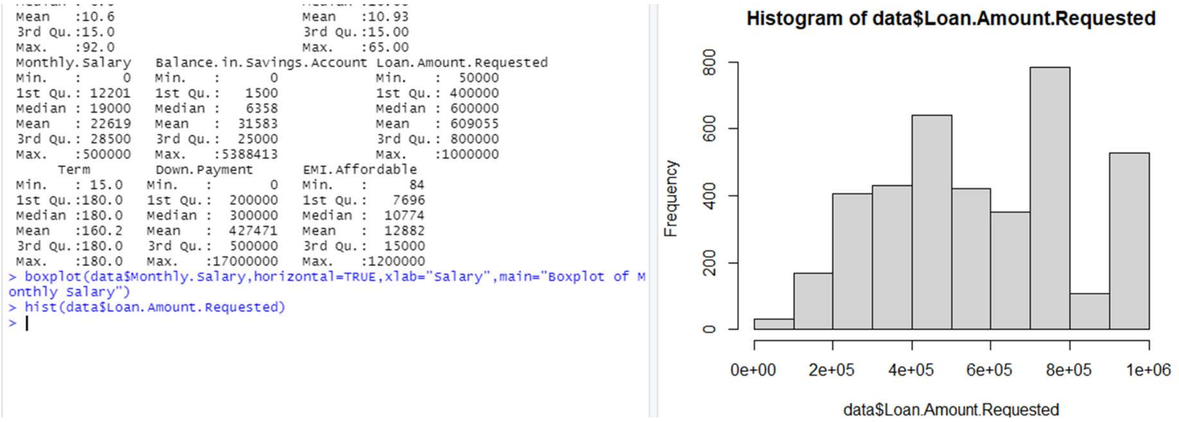
```
 Mean   :10.6                         Mean   :10.93
 3rd Qu.:15.0                         3rd Qu.:15.00
 Max.   :92.0                         Max.   :65.00
 Monthly.Salary   Balance.in.Savings.Account Loan.Amount.Requested
 Min.   :     0   Min.   :      0            Min.   :  50000
 1st Qu.: 12201   1st Qu.:   1500            1st Qu.: 400000
 Median : 19000   Median :   6358            Median : 600000
 Mean   : 22619   Mean   :  31583            Mean   : 609055
 3rd Qu.: 28500   3rd Qu.:  25000            3rd Qu.: 800000
 Max.   :500000   Max.   :5388413            Max.   :1000000
      Term          Down.Payment          EMI.Affordable
 Min.   : 15.0   Min.   :       0   Min.   :     84
 1st Qu.:180.0   1st Qu.:  200000   1st Qu.:   7696
 Median :180.0   Median :  300000   Median :  10774
 Mean   :160.2   Mean   :  427471   Mean   :  12882
 3rd Qu.:180.0   3rd Qu.:  500000   3rd Qu.:  15000
 Max.   :180.0   Max.   :17000000   Max.   :1200000
> boxplot(data$Monthly.Salary,horizontal=TRUE,xlab="Salary",main="Boxplot of M
onthly Salary")
> |
```

We can find numerous outliers from the above box plot

4. Visualise the Loan Amount attribute (Histogram is suggested, why?)

```
hist(data$Loan.Amount.Requested)
```

<u>Output</u> -:



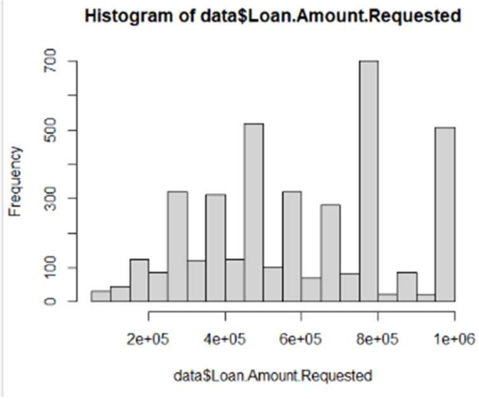
```
 Mean   :10.6                         Mean   :10.93
 3rd Qu.:15.0                         3rd Qu.:15.00
 Max.   :92.0                         Max.   :65.00
 Monthly.Salary   Balance.in.Savings.Account Loan.Amount.Requested
 Min.   :     0   Min.   :      0            Min.   :  50000
 1st Qu.: 12201   1st Qu.:   1500            1st Qu.: 400000
 Median : 19000   Median :   6358            Median : 600000
 Mean   : 22619   Mean   :  31583            Mean   : 609055
 3rd Qu.: 28500   3rd Qu.:  25000            3rd Qu.: 800000
 Max.   :500000   Max.   :5388413            Max.   :1000000
      Term          Down.Payment          EMI.Affordable
 Min.   : 15.0   Min.   :       0   Min.   :     84
 1st Qu.:180.0   1st Qu.:  200000   1st Qu.:   7696
 Median :180.0   Median :  300000   Median :  10774
 Mean   :160.2   Mean   :  427471   Mean   :  12882
 3rd Qu.:180.0   3rd Qu.:  500000   3rd Qu.:  15000
 Max.   :180.0   Max.   :17000000   Max.   :1200000
> boxplot(data$Monthly.Salary,horizontal=TRUE,xlab="Salary",main="Boxplot of M
onthly Salary")
> hist(data$Loan.Amount.Requested)
> |
```

Histogram of data$Loan.Amount.Requested
Frequency
0
200
400
600
800
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
data$Loan.Amount.Requested

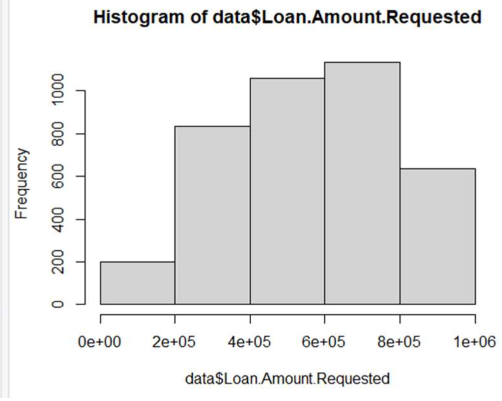
Since it a continuous variable, histogram is appropriate for initial analysis

- Try changing the bin width of the histogram by modifying the ***breaks*** attribute

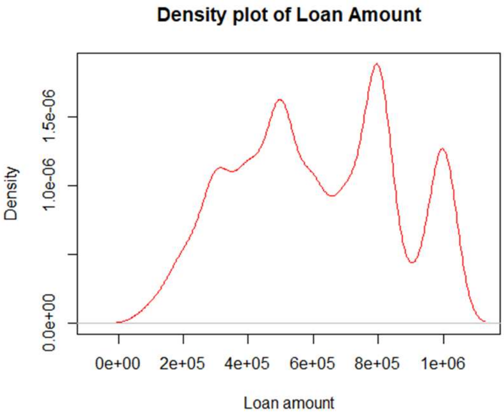
```
hist(data$Loan.Amount.Requested,breaks=15)
```



```
hist(data$Loan.Amount.Requested,breaks=c(0,200000,4000
   00,600000,800000,1000000))
```



- You can see that since the bin width influences the nature of the distribution of a histogram, in order to find the modality of the distribution, **density plots** can also be used.
  ```
  plot(density(data$Loan.Amount.Requested),col="red",main="
  De nsity plot of Loan Amount",xlab="Loan amount")
  ```



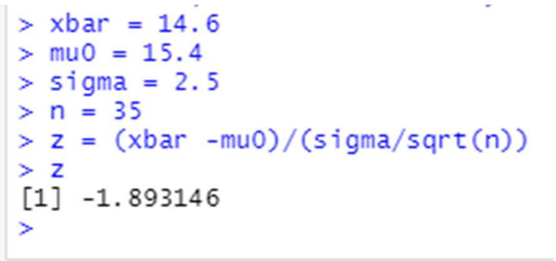
- Which other visualisation is suitable for the Loan Amount Variable?
  Other interesting alternatives could be Frequency polygon and box plot. However, there are a myriad of alternatives that you can always explore!

Confidence Interval and Hypothesis Testing

5. Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins at the same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

```
xbar = 14.6            # sample mean  mu0 = 15.4
# hypothesized value  sigma = 2.5           #
population standard deviation
n = 35                 # sample size  z = (xbar
- mu0)/(sigma/sqrt(n)) # test statistic
```

Output -:



```
> xbar = 14.6
> mu0 = 15.4
> sigma = 2.5
> n = 35
> z = (xbar -mu0)/(sigma/sqrt(n))
> z
[1] -1.893146
>
```

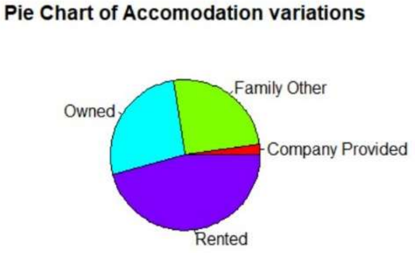
The test statistic -1.8931 lies between the critical values -1.9600 and 1.9600. Hence, at .05 significance level, we do not reject the null hypothesis that the mean penguin weight does not differ from last year

Visualizations

6. Visualize the distribution of Accomodation.Type attribute (PieChart is suggested)

```
val <- count(data, "Accomodation.Type")
     lbls = val$Accomodation.Type
     pie(val$freq, labels = val$Accomodation.Type, main="Pie Chart
     of
     Countries",col=rainbow(length(lbls)))
```

Expected Output



Since there are multiple variables (but not too many) pie chart is suitable

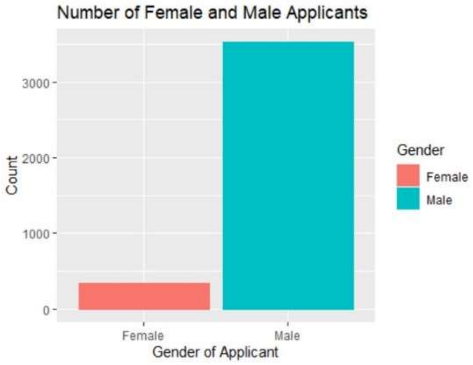
- Print and analyse the val variable
  The val variable gives a count of each of the different category values
- The basic pie function can be customised by adding in percentages to represent the sectors, gradient color scheme and many more

7. Visualize the Gender attribute (Bar Graph is suggested)

```
gender <- count(data, "Gender")
ggplot(gender, aes(x = Gender, y = freq,color=Gender,fill=
Gender )) + geom_bar(stat="identity")+ ylab("Count") +
xlab("Gender of
Applicant")+ ggtitle("Number of Female and Male Applicants")
```

Since there are only 2 categories bar graph proves to be useful

<u>Expected Output</u>



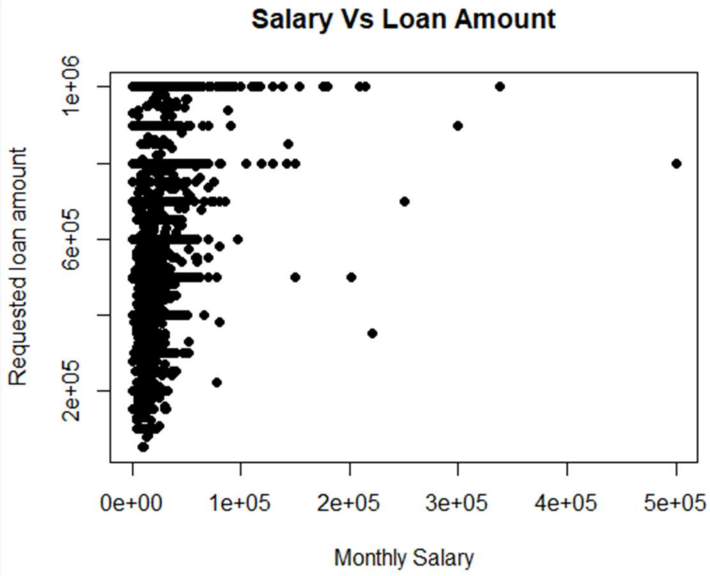
- Does it look like a biased study?

  Yes indeed! We can see the large gap indicating that the study is not representative based on gender terms.

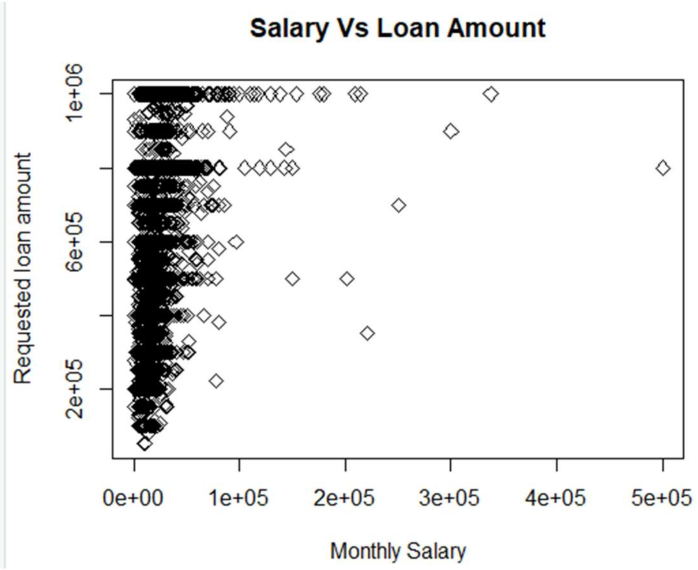
8. Find variation of Monthly Salaries with respect to EMI amount (Scatter Plot is suggested)

```
plot(data$Monthly.Salary, data$Loan.Amount.Requested,
main="Salary Vs Loan Amount", xlab="Monthly Salary",
ylab="Requested loan amount", pch=19)
```

<u>Output -:</u>



- Explore pch attribute
  Varying the value of the pch attribute changes the shape of the marker. Some options are solid circle, square, filled circle etc.



- Is there a significant trend in the plot? Does lower income imply lower loan amount requested?
  We observe no significant trend in the plot. However, there are very few individuals with a very high income and the loan amount does not seem to be strongly dependent on salary because we observe low salaried individuals also taking up higher loans.

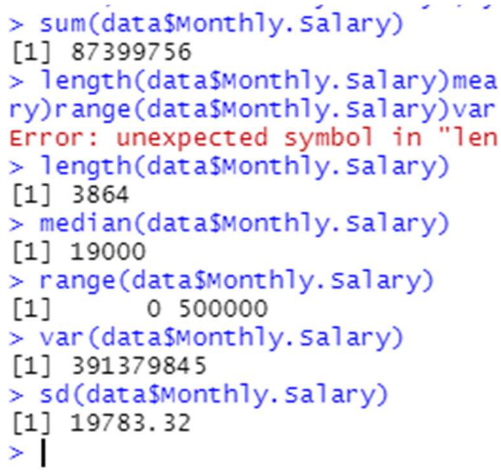
- Try plotting scatter plot matrices where you can visualize multiple variables at once

Summary Statistics and Grouping Conditions

9. Descriptive Statistics for the dataset

```
sum(data$Monthly.Salary)
length(data$Monthly.Salary)
mean(data$Monthly.Salary)
median(data$Monthly.Salary)
range(data$Monthly.Salary)
var(data$Monthly.Salary)
sd(data$Monthly.Salary)
```

<u>Output -:</u>



```
> sum(data$Monthly.Salary)
[1] 87399756
> length(data$Monthly.Salary)mea
ry)range(data$Monthly.Salary)var
Error: unexpected symbol in "len
> length(data$Monthly.Salary)
[1] 3864
> median(data$Monthly.Salary)
[1] 19000
> range(data$Monthly.Salary)
[1]      0 500000
> var(data$Monthly.Salary)
[1] 391379845
> sd(data$Monthly.Salary)
[1] 19783.32
> |
```

- Look at functions such as **seq**, **rep** to create custom sequences of numbers ● R does not have a basic function for the Mode. Try writing one by yourself.

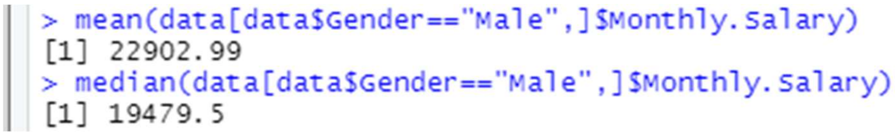
```
my_mode <- function(x) {
uniqueval <- unique(x)   tab<-
tabulate(match(x, uniqueval))
uniqueval[tab==max(tab)]
}
```

- How can we use skew and kurtosis to check whether the data is bimodal?

10. Find the mean monthly salary for females

```
mean(data[data$Gender=="Male",]$Monthly.Salary)
```

<u>Output</u> -:



```
> mean(data[data$Gender=="Male",]$Monthly.Salary)
[1] 22902.99
> median(data[data$Gender=="Male",]$Monthly.Salary)
[1] 19479.5
```

- Try finding the **median** of Monthly Salary for Males
  `median(data[data$Gender=="Male",]$Monthly.Salary)`

- What is the significance of the “,” ?
  Data frames, as they are called in R, have rows and columns just like your excel sheet has.

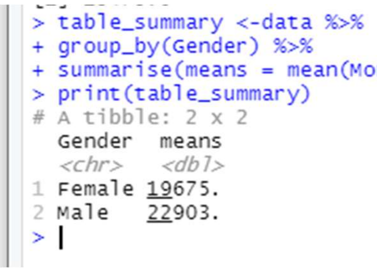
Each cell is determined by 2 numbers, its row and column number. The same applies here. So when you run data[1,2] it will return the cell formed from the intersection of the 1st row and the 2nd column.

When you run data[condition,] you're actually asking R to check and return those rows that satisfy your condition. In other words R is returning the whole row (with all cells not just a single cell; depending on how many columns you have).

11. Find the mean monthly salaries, grouped by the Gender attribute. Explore the dplyr package

```
table_summary <- data %>%
group_by(Gender) %>%
  summarise(means = mean(Monthly.Salary))
print(table_summary)
```

Output -:



```
> table_summary <-data %>%
+ group_by(Gender) %>%
+ summarise(means = mean(Monthly.Salary))
> print(table_summary)
# A tibble: 2 x 2
  Gender   means
  <chr>    <dbl>
1 Female  19675.
2 Male    22903.
> 
```

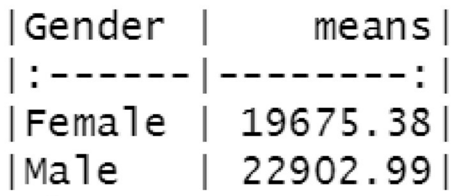
This is a much easier way than using multiple statements for each summary

- Try to get mean, median and range of salaries for both Males and Females using the group_by clause
  HINT : You'll have to use comma separated values

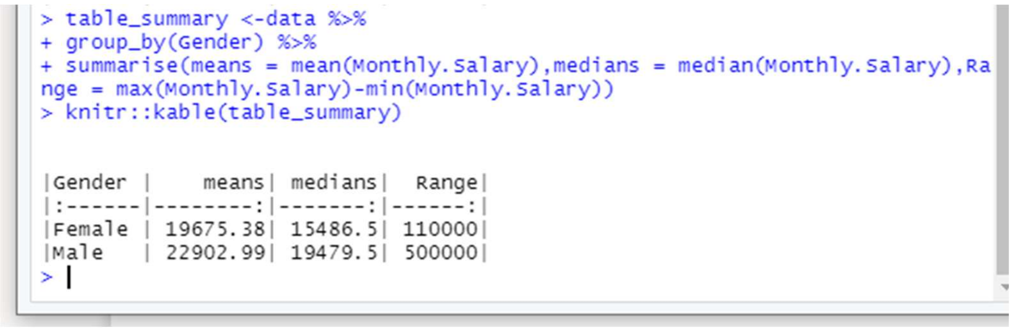
```
table_summary <- data %>%
group_by(Gender) %>%  summarise(means
= mean(Monthly.Salary),medians =
median(Monthly.Salary),Range =
max(Monthly.Salary)min(Monthly.Salary))
```

- To make this pretty you could use knitr::kable `knitr::kable(table_summary)`

Expected Output



```
|Gender |    means|
|:------|--------:|
|Female | 19675.38|
|Male   | 22902.99|
```



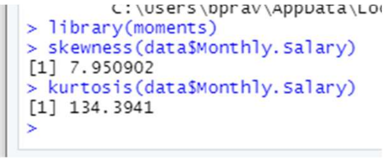
```
> table_summary <-data %>%
+ group_by(Gender) %>%
+ summarise(means = mean(Monthly.Salary),medians = median(Monthly.Salary),Ra
nge = max(Monthly.Salary)-min(Monthly.Salary))
> knitr::kable(table_summary)


|Gender |    means| medians|  Range|
|:------|--------:|-------:|------:|
|Female | 19675.38| 15486.5| 110000|
|Male   | 22902.99| 19479.5| 500000|
> 
```

12. Find the Skewness and kurtosis for the Monthly Salary attribute

```
skewness(data$Monthly.Salary)
kurtosis(data$Monthly.Salary)
```

Output -:



```
        C:\Users\bprav\AppData\Lo
> library(moments)
> skewness(data$Monthly.Salary)
[1] 7.950902
> kurtosis(data$Monthly.Salary)
[1] 134.3941
>
```

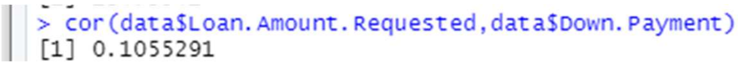
- Is the attribute left skewed?
  The positive value indicates that the monthly salary distribution is skewed towards the right

- What about it's kurtosis? platykurtic? Leptokurtic?
  Positive excess kurtosis would indicate a fat-tailed distribution, and is said to be ***leptokurtic***

Correlation and Data Reduction

13. Find the value of correlation between Loan amount and Down payment

```
cor(data$Loan.Amount.Requested,data$Down.Payment)
```

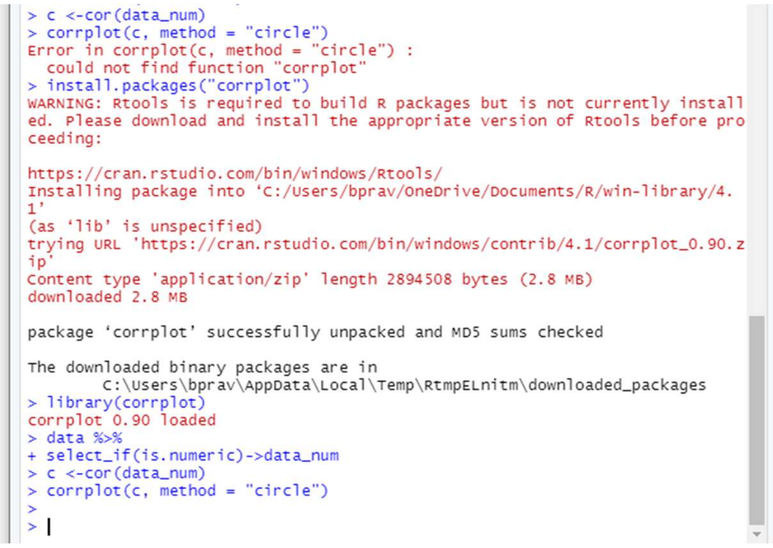


```
> cor(data$Loan.Amount.Requested,data$Down.Payment)
[1] 0.1055291
```

14. Explore the corrplot package to plot a correlogram between the various attributes

```
data %>% select_if(is.numeric)->data_num
c <- cor(data_num)
corrplot(c, method = "circle")
```

Output -:

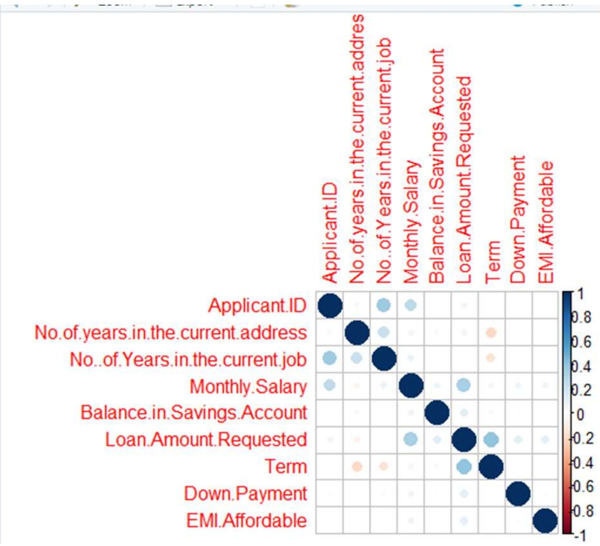


```
> c <-cor(data_num)
> corrplot(c, method = "circle")
Error in corrplot(c, method = "circle") :
  could not find function "corrplot"
> install.packages("corrplot")
WARNING: Rtools is required to build R packages but is not currently install
ed. Please download and install the appropriate version of Rtools before pro
ceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/bprav/OneDrive/Documents/R/win-library/4.
1’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/corrplot_0.90.z
ip'
Content type 'application/zip' length 2894508 bytes (2.8 MB)
downloaded 2.8 MB

package ‘corrplot’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\bprav\AppData\Local\Temp\RtmpELnitm\downloaded_packages
> library(corrplot)
corrplot 0.90 loaded
> data %>%
+ select_if(is.numeric)->data_num
> c <-cor(data_num)
> corrplot(c, method = "circle")
>
> |
```


Applicant.ID
No.of.years.in.the.current.addres
No..of.Years.in.the.current.job
Monthly.Salary
Balance.in.Savings.Account
Loan.Amount.Requested
Term
Down.Payment
EMI.Affordable
Applicant.ID
No.of.years.in.the.current.address
No..of.Years.in.the.current.job
Monthly.Salary
Balance.in.Savings.Account
Loan.Amount.Requested
Term
Down.Payment
EMI.Affordable