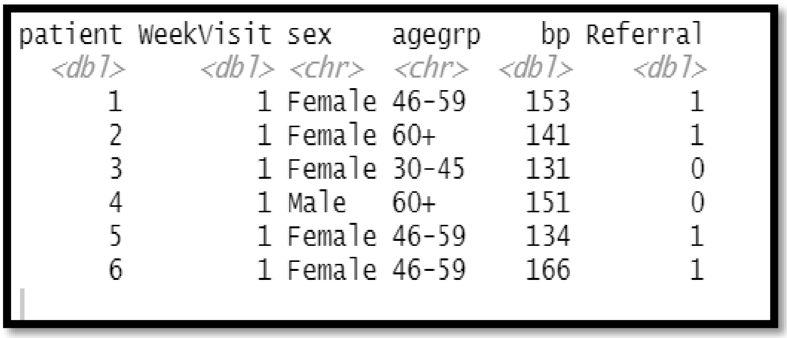
TITLE : ASSESSMENT 1

[Ans1]

head(Assessment1_dataset_1_)



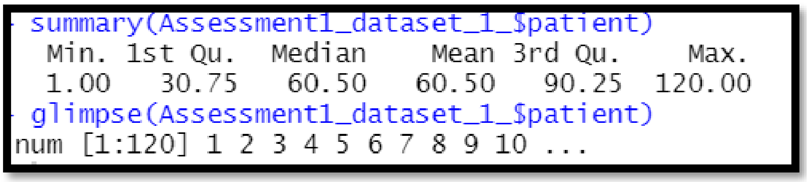
```
patient WeekVisit sex    agegrp    bp Referral
  <dbl>     <dbl> <chr>  <chr>  <dbl>    <dbl>
      1         1 Female 46-59    153        1
      2         1 Female 60+      141        1
      3         1 Female 30-45    131        0
      4         1 Male   60+      151        0
      5         1 Female 46-59    134        1
      6         1 Female 46-59    166        1
```

Comments on summaries:

- Patient

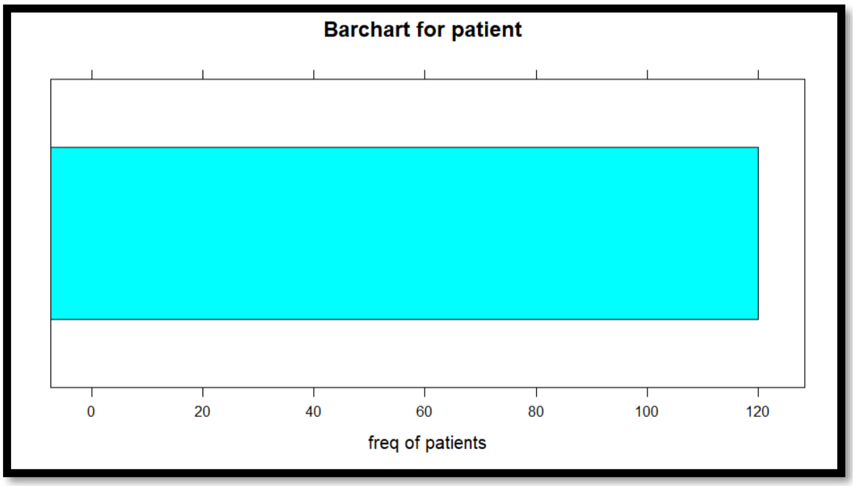
summary(Assessment1_dataset_1_$patient)

glimpse(Assessment1_dataset_1_$patient)



```
 summary(Assessment1_dataset_1_$patient)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   30.75   60.50   60.50   90.25  120.00 
 glimpse(Assessment1_dataset_1_$patient)
num [1:120] 1 2 3 4 5 6 7 8 9 10 ...
```

barchart(Assessment1_dataset_1_$patient, xlab = 'freq of patients', main = 'Barchart for patient')

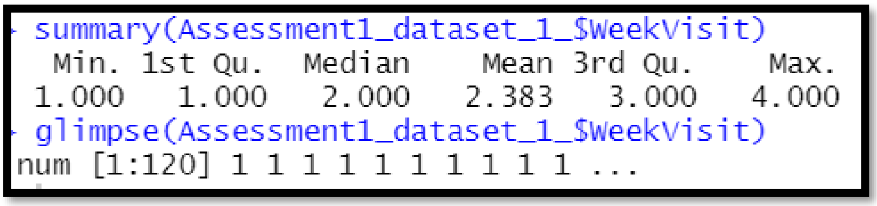


patient column consists of numerical data with distinct unique values and it a categorical variable to identify each patients.The data in this column is discrete as shown by barchart and hence will follow a discrete distribution.

- WeekVisit

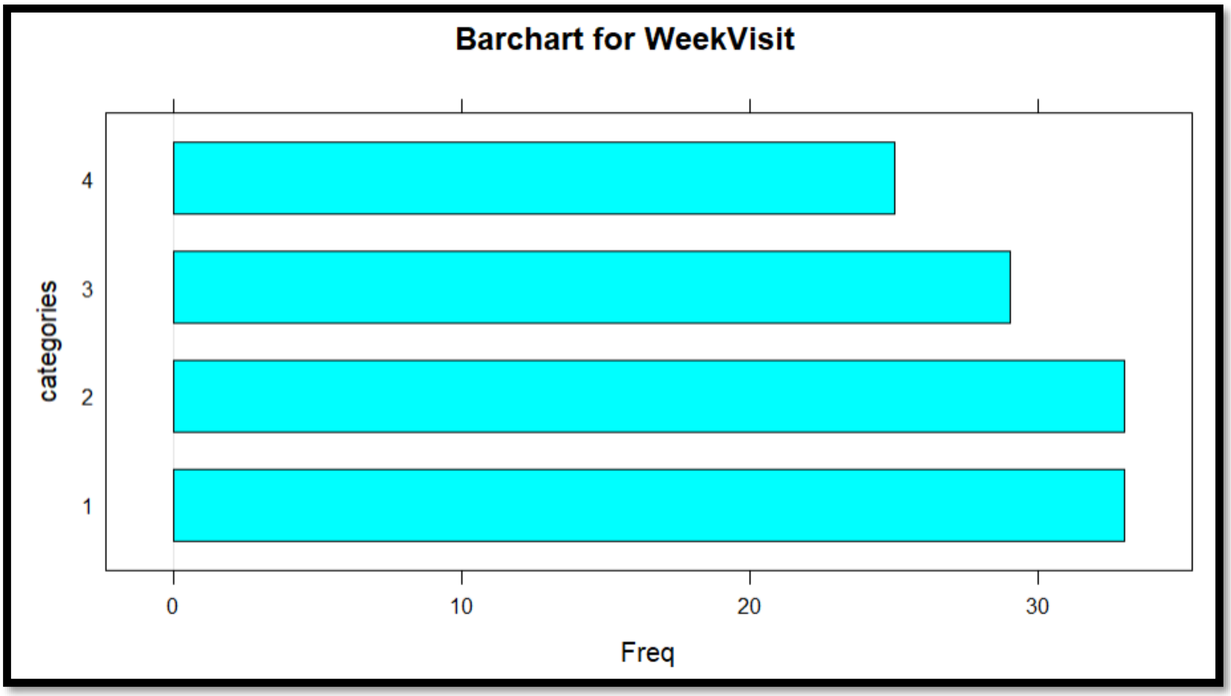
summary(Assessment1_dataset_1_$WeekVisit)

glimpse(Assessment1_dataset_1_$WeekVisit)



```
 summary(Assessment1_dataset_1_$WeekVisit)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 1.000   1.000   2.000   2.383   3.000   4.000 
 glimpse(Assessment1_dataset_1_$WeekVisit)
num [1:120] 1 1 1 1 1 1 1 1 1 1 ...
```

barchart(table(Assessment1_dataset_1_$WeekVisit),ylab = 'categories', main = 'Barchart for WeekVisit')

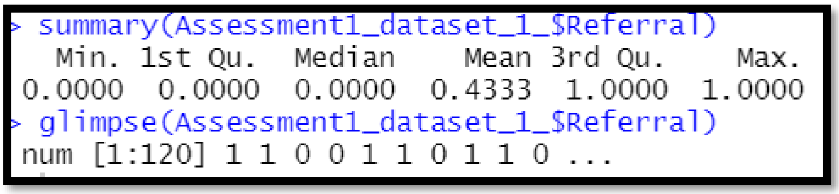


WeekVisit column consists of numerical data and it is a categorical variable to show weekly visits by each patient.The data in this column is discrete as shown by barchart and hence will follow a discrete distribution.

- Referral

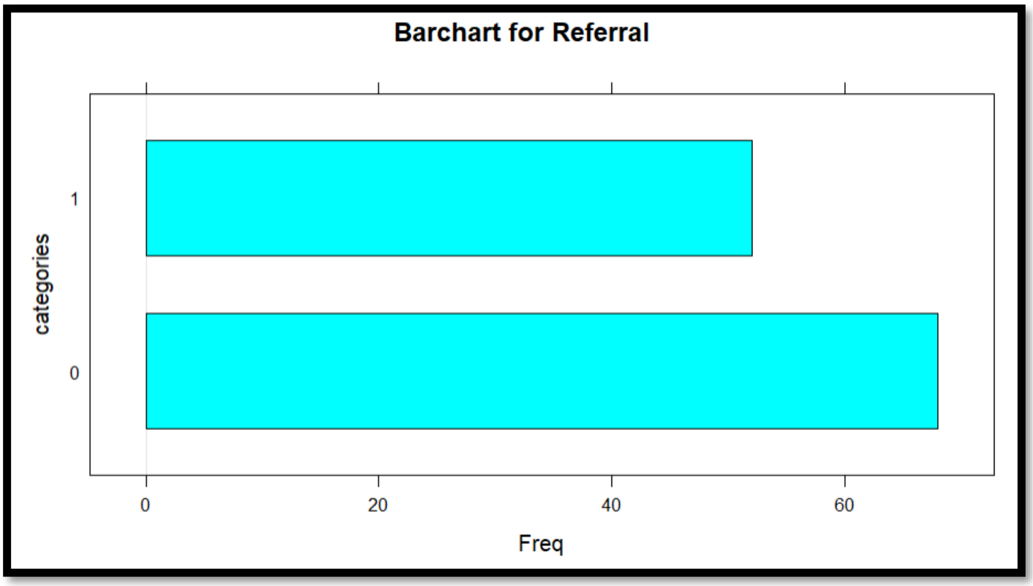
summary(Assessment1_dataset_1_$Referral)

glimpse(Assessment1_dataset_1_$Referral)



```
> summary(Assessment1_dataset_1_$Referral)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.4333  1.0000  1.0000 
> glimpse(Assessment1_dataset_1_$Referral)
 num [1:120] 1 1 0 0 1 1 0 1 1 0 ...
```

barchart(table(Assessment1_dataset_1_$Referral),ylab = 'categories', main = 'Barchart for Referral')

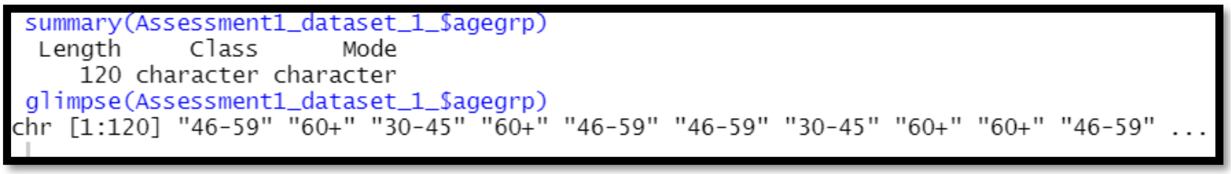


Referral column consists of numerical data and it is a categorical variable to show each patient got referral or not.The data in this column is discrete as shown by barchart and hence will follow a discrete distribution.

- Agegrp

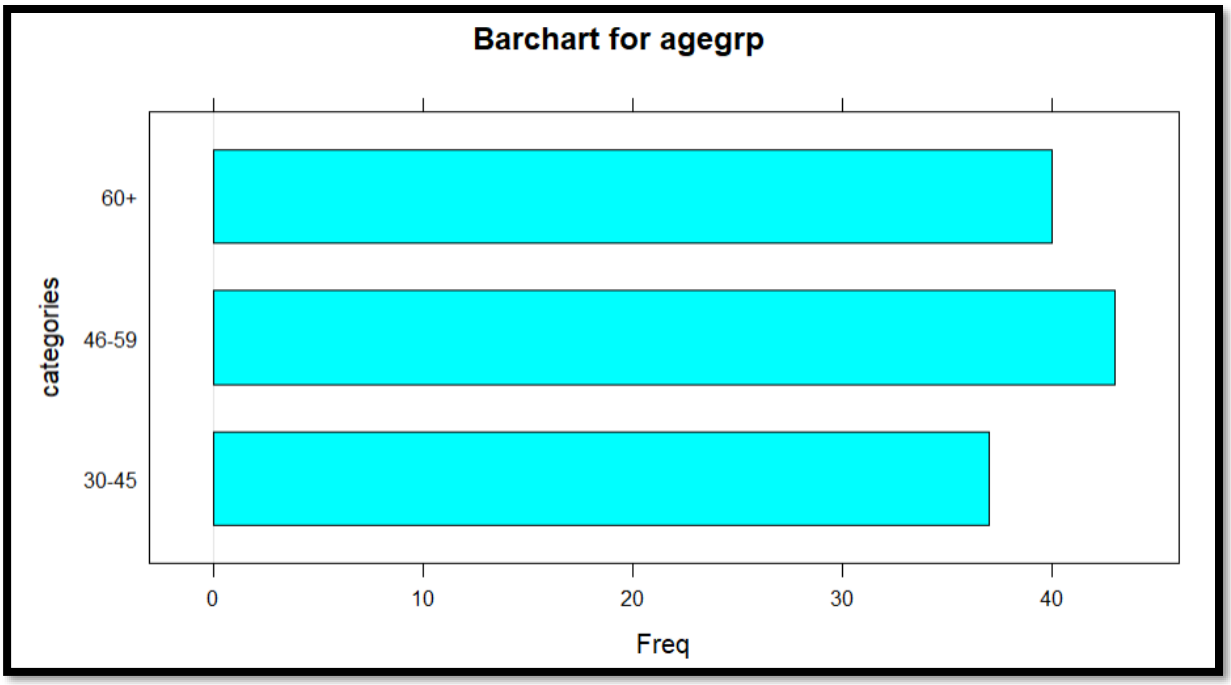
summary(Assessment1_dataset_1_$agegrp)

glimpse(Assessment1_dataset_1_$agegrp)



```
summary(Assessment1_dataset_1_$agegrp)
  Length     Class      Mode 
     120 character character 
glimpse(Assessment1_dataset_1_$agegrp)
chr [1:120] "46-59" "60+" "30-45" "60+" "46-59" "46-59" "30-45" "60+" "60+" "46-59" ...
```

barchart(table(Assessment1_dataset_1_$agegrp),ylab = 'categories', main = 'Barchart for agegrp')

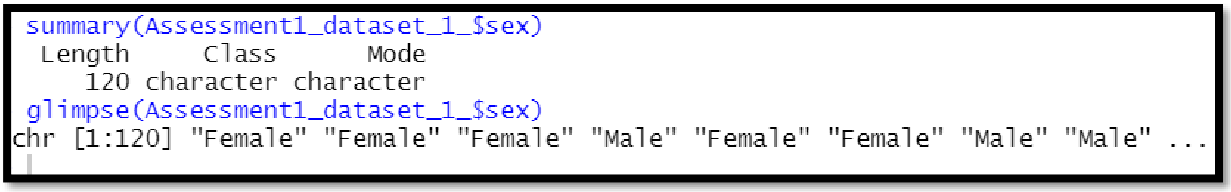


agegrp column is a categorical variable.The data in this column is discrete as shown by barchart and hence will follow a discrete distribution.

- Sex

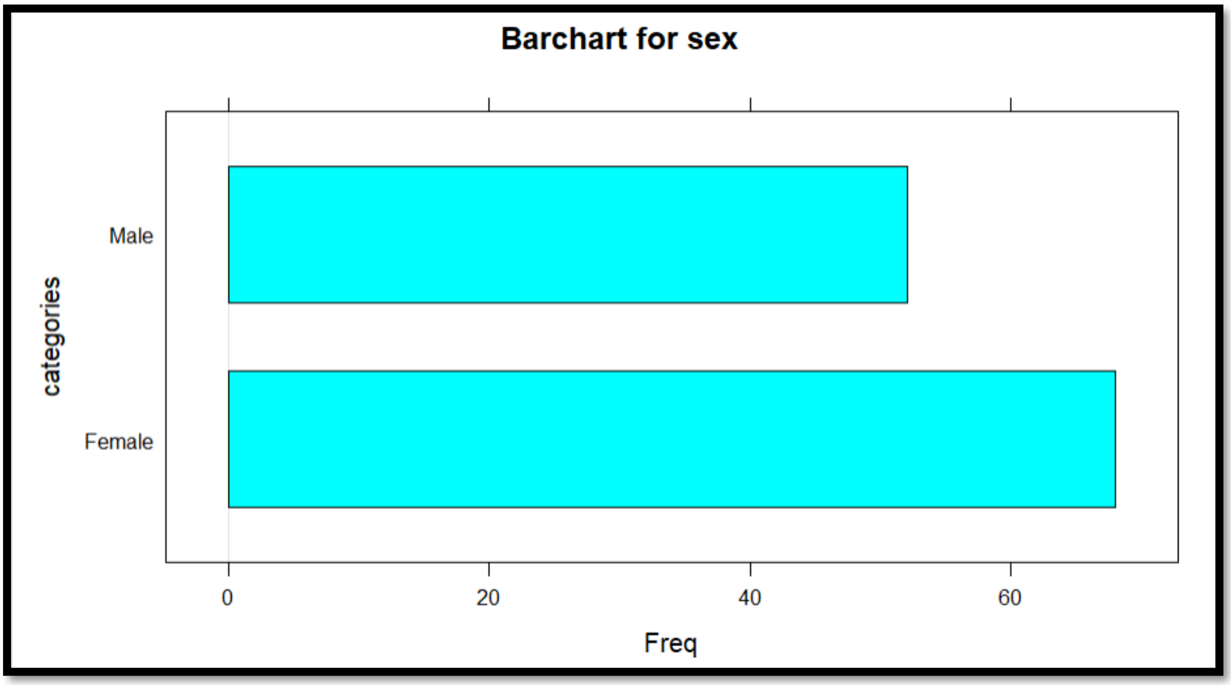
```
summary(Assessment1_dataset_1_$sex)
```

```
glimpse(Assessment1_dataset_1_$sex)
```



```
 summary(Assessment1_dataset_1_$sex)
  Length     Class      Mode 
     120 character character 
 glimpse(Assessment1_dataset_1_$sex)
chr [1:120] "Female" "Female" "Female" "Male" "Female" "Female" "Male" "Male" ...
```

```
barchart(table(Assessment1_dataset_1_$ex),ylab = 'categories', main = 
'Barchart for sex' )
```

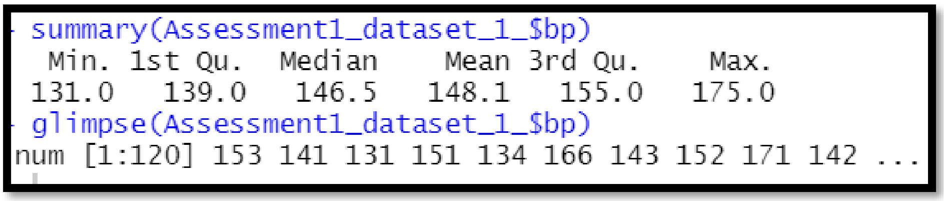


sex column is a categorical variable.The data in this column is discrete as shown by barchart and hence will follow a discrete distribution.

- Bp

summary(Assessment1_dataset_1_$bp)

glimpse(Assessment1_dataset_1_$bp)



```
 summary(Assessment1_dataset_1_$bp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  131.0   139.0   146.5   148.1   155.0   175.0
 glimpse(Assessment1_dataset_1_$bp)
num [1:120] 153 141 131 151 134 166 143 152 171 142 ...
```

Blood pressure(bp) cannot be counted hence it is a continuous variable. The mean and median values differ slightly which indicates there is skewness in the data.

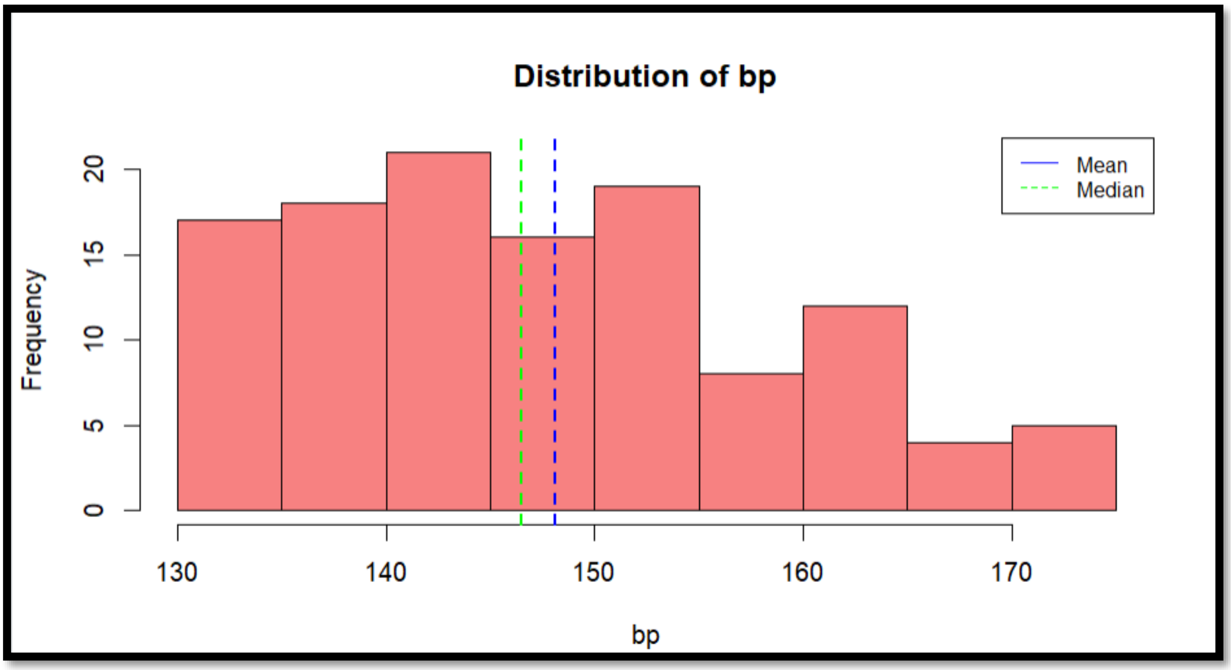
Plot to show difference between mean and median :

hist(Assessment1_dataset_1_$bp,breaks= 10, col="#F78181", xlab="bp", main="Distribution of bp")

```
abline(v = mean(Assessment1_dataset_1_$bp), col = "blue", lwd = 2, lty="dashed")
```

```
abline(v = median(Assessment1_dataset_1_$bp), col = "green", lwd = 2, lty="dashed")
```

```
legend('topright', legend=c("Mean", "Median"),col=c("blue", "green"), lty=1:2, cex=0.8)
```

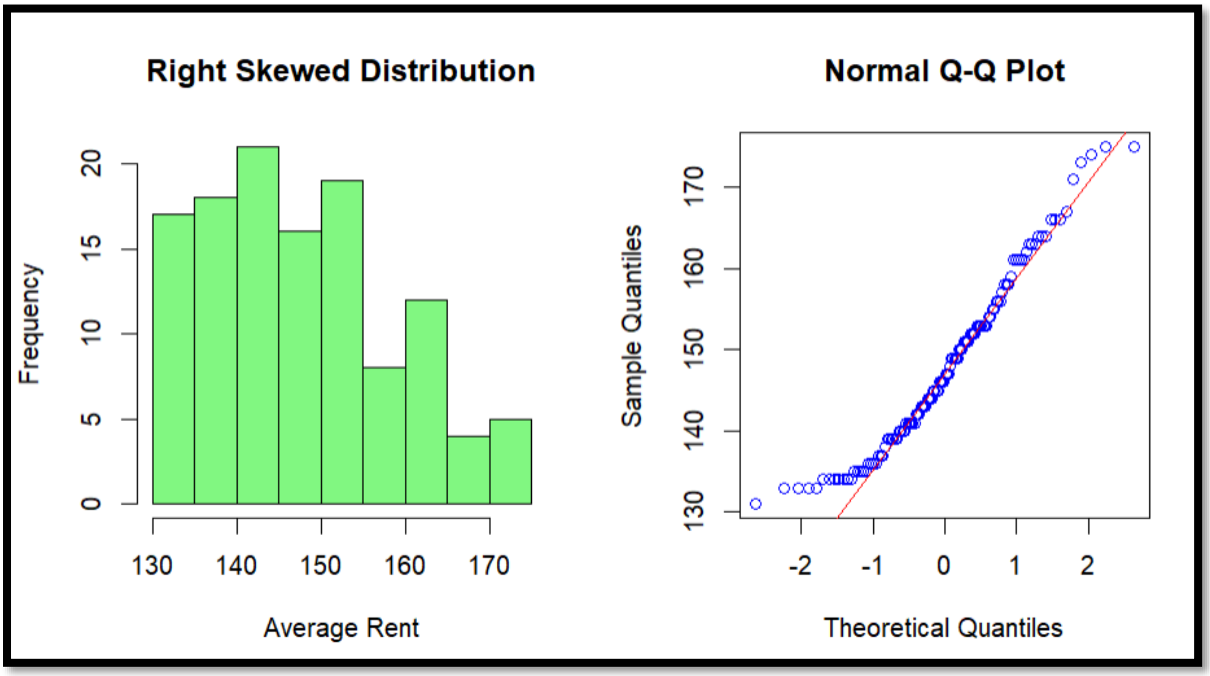


Checking the normality of bp data :

```
hist(Assessment1_dataset_1_$bp,breaks =10, main="Right Skewed Distribution", xlab="Average Rent", col="#81F781")
```

```
qqnorm(Assessment1_dataset_1_$bp, col="blue")
```

```
qqline(Assessment1_dataset_1_$bp, col="red")
```



Histogram shows bp has a right skewed distribution so most of the data is concentrated to the left side of the distribution, its mean is located to the right of median which indicates that majority of population have lower blood pressure.

We can confirm this further

```
length(which(Assessment1_dataset_1_$bp < 148.1))
```

```
[1] 64
```

Thus 64 out off 120 patients have blood pressure lower than mean value.

The histogram doesn’t look symmetric or have a bell shape. A proportion of data points do not lie on the straight line in the normal Q-Q plot so the data doesn't seem to follow a normal distribution.

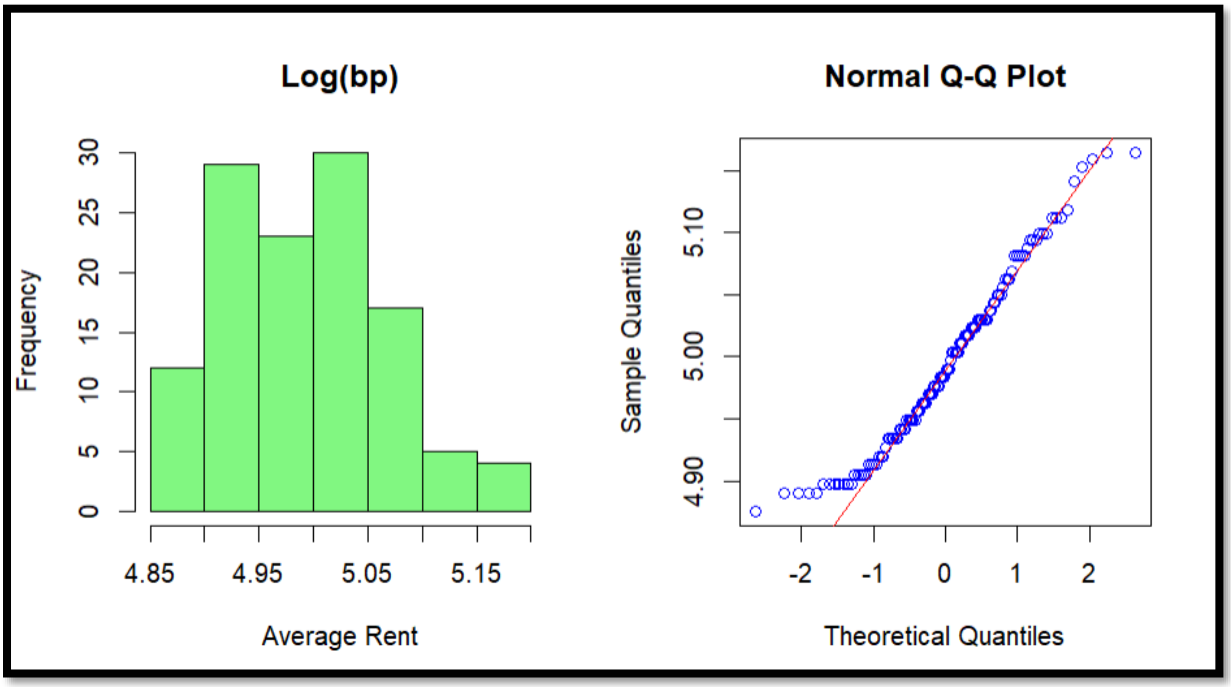
Considering taking log of the data of check normality :

```
hist(log(Assessment1_dataset_1_$bp),breaks =10, main="Log(bp)",
xlab="Average Rent", col="#81F781")
```

```
qqnorm(log(Assessment1_dataset_1_$bp), col="blue")
```

```
qqline(log(Assessment1_dataset_1_$bp), col="red")
```

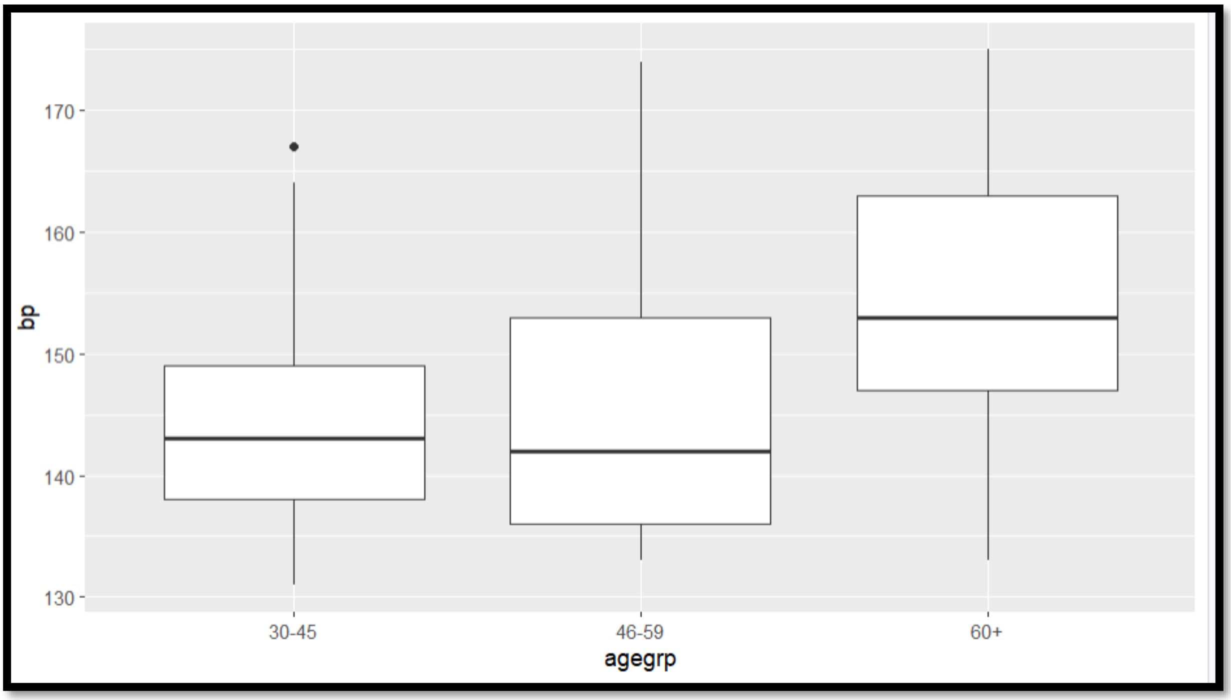
```
par(mfrow=c(1,2))
```



The histogram still doesn't look symmetric or have bell shape and few propotion of data points do not lie on the straight line in the normal Q-Q plot so it still fails to show normal distribution.

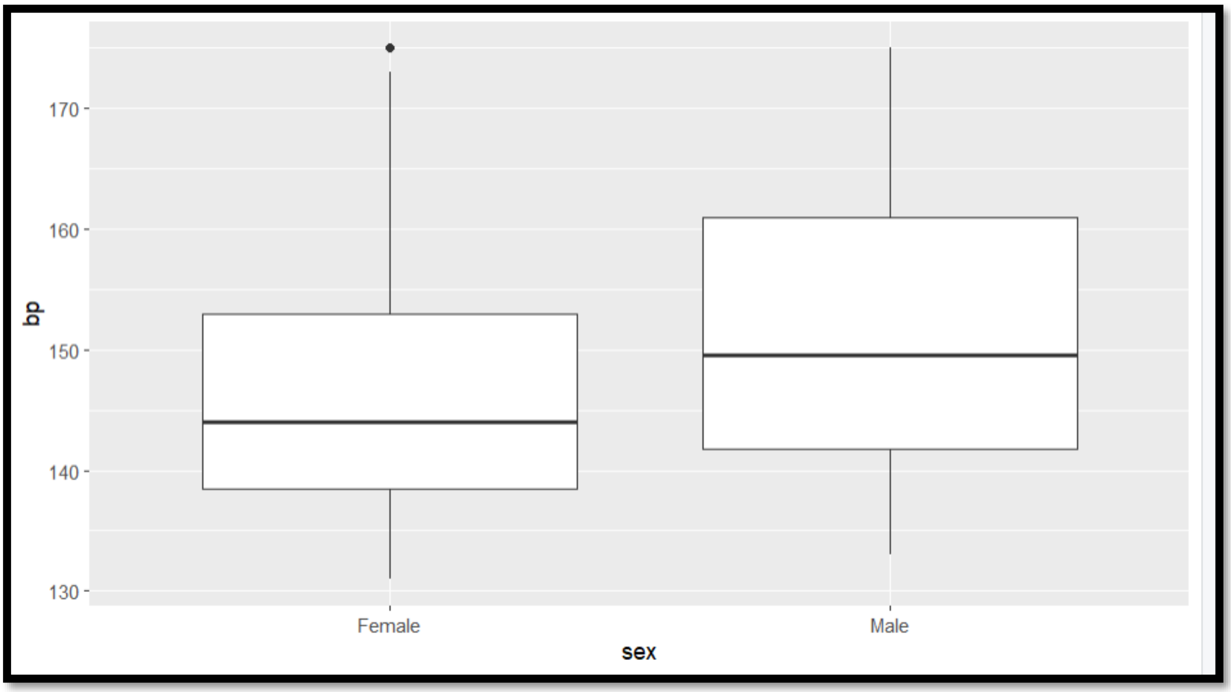
Relationship between bp and other variables for better analysis of the dataset :

```
ggplot(Assessment1_dataset_1_, aes(x=agegrp, y=bp))+ geom_boxplot()
```



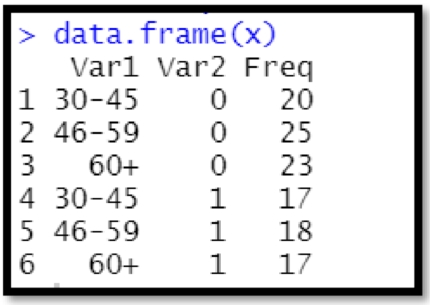
From the boxplot we understand '60+' agegroup have higher levels of blood pressure and also indicates the skewness in the data.

```
ggplot(Assessment1_dataset_1_, aes(x=sex, y=bp)) + geom_boxplot()
```



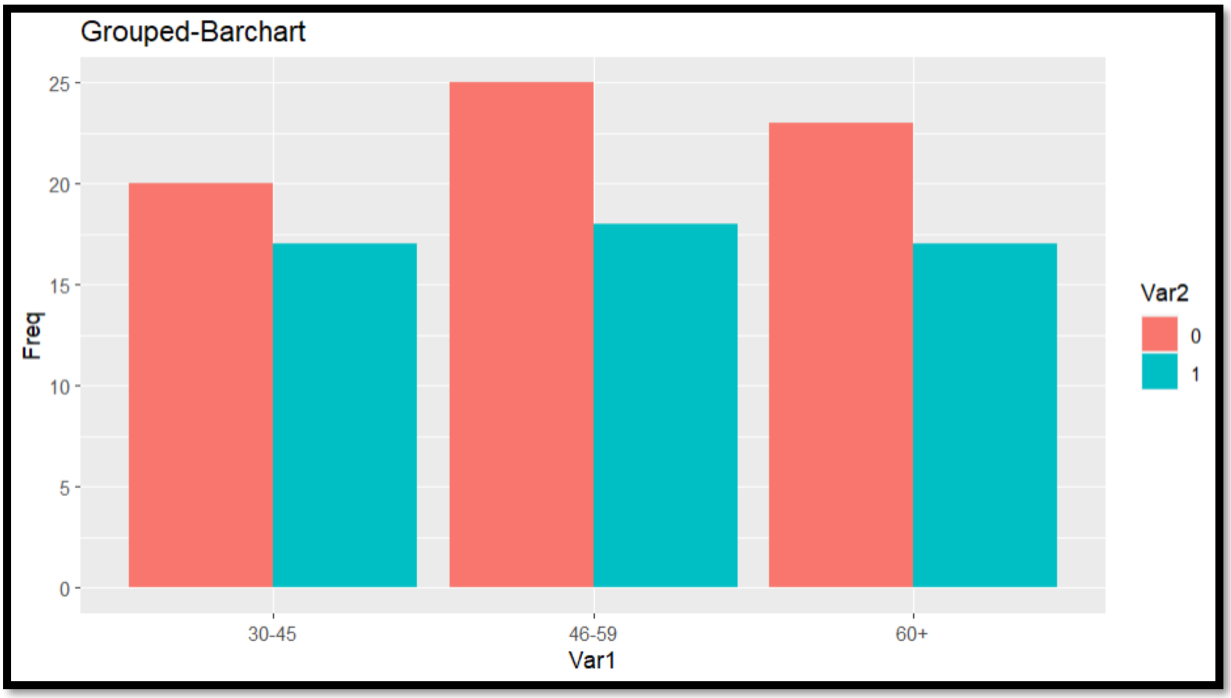
We see that male have higher levels of blood pressure compared to women and indicates skewness in the data.

```
x= table(Assessment1_dataset_1_$agegrp,Assessment1_dataset_1_$Referral)
```



```
> data.frame(x)
   Var1 Var2 Freq
1 30-45    0   20
2 46-59    0   25
3   60+    0   23
4 30-45    1   17
5 46-59    1   18
6   60+    1   17
```

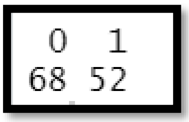
```
ggplot(data.frame(x), aes(fill=Var2, x=Var1, y=Freq)) +
geom_bar(position="dodge", stat="identity")+ labs(title='Grouped-Barchart')
# used position = 'dodge' to get grouped barchart
```



From this barchart we understand that maximum population to get referral as well as to not get referral belong to 46-59 age group

[Ans 2]

```
table(Assessment1_dataset_1_$Referral)
```



```
 0  1
68 52
```

finding the probability of patients who got referral

```
prob = 52/120
prob
[1] 0.4333333
```

plotting pmf of X from Bin(120,0.433) :

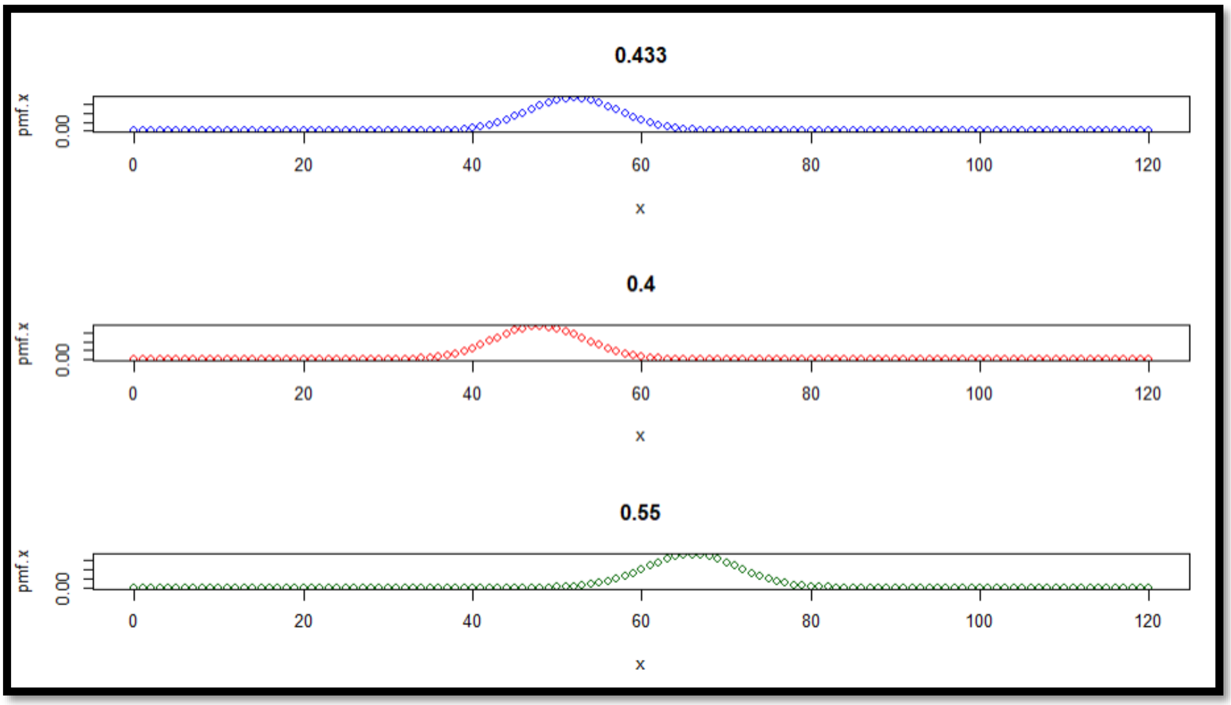
```
n <- 120
p<- 0.433
x <- 0:n
pmf.x <- dbinom(x, n, p)       # binomial distribution formula
plot(x, pmf.x, main = '0.433',col='blue')
```

plotting pmf of X from Bin(120,0.4) :

```
n <- 120
p<- 0.4
x <- 0:n
pmf.x <- dbinom(x, n, p)
plot(x, pmf.x, main = '0.4',col='red')
```

plotting pmf of X from Bin(120,0.55) :

```
n <- 120

p<- 0.55

x <- 0:n

pmf.x <- dbinom(x, n, p)

plot(x, pmf.x, main = '0.55',col='dark green')
```



From the three plots we observe that binomial distribution graph is not identical as the values of their probabilities are different. However by comparing with the referral variable binomial distribution graph we find that Bin(120,0.4) matches closely as their probabilities are almost equal.

Similarly for the probability of patients who didn't get referral :

```
prob = 68/120

prob

[1] 0.5666667
```

plotting pmf of X from Bin(120,0.566) :

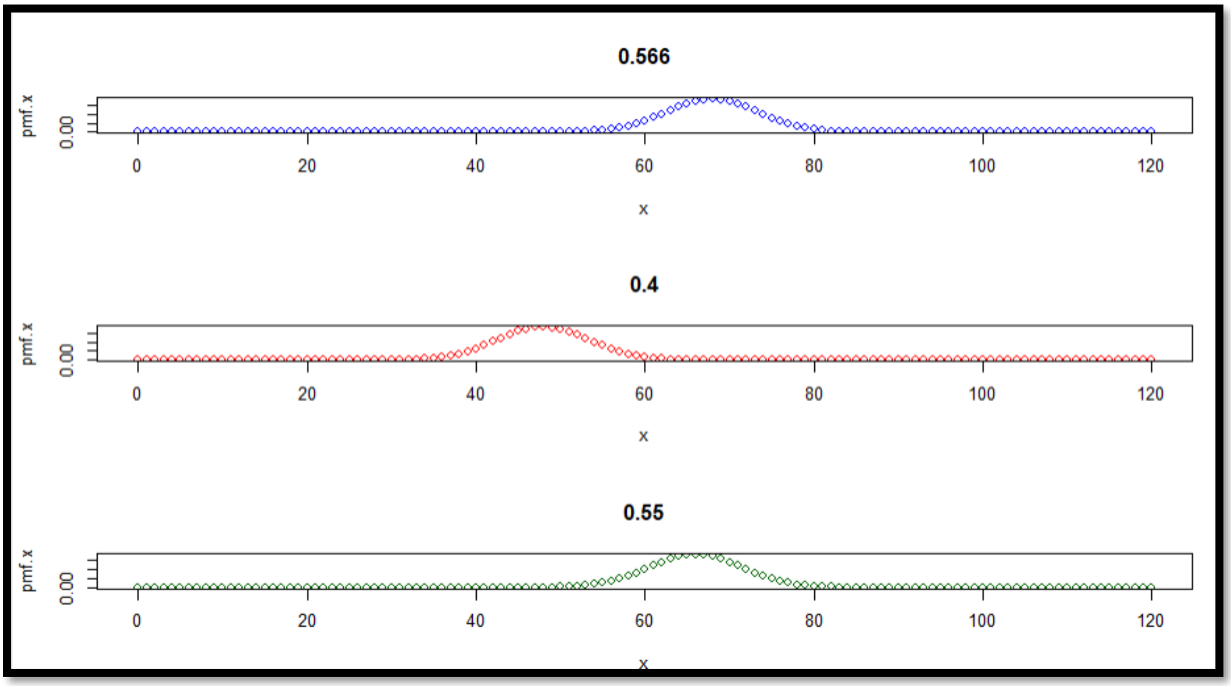
```
n <- 120
p<- 0.566
x <- 0:n
pmf.x <- dbinom(x, n, p)
plot(x, pmf.x, main = '0.566',col='blue')
```

plotting pmf of X from Bin(120,0.4) :

```
n <- 120
p<- 0.4
x <- 0:n
pmf.x <- dbinom(x, n, p)
plot(x, pmf.x, main = '0.4',col='red')
```

plotting pmf of X from Bin(120,0.55) :

```
n <- 120
p<- 0.55
x <- 0:n
pmf.x <- dbinom(x, n, p)
plot(x, pmf.x, main = '0.55',col='dark green')
```



Overhere by comparing with the referral variable binomial distribution graph we find that Bin(120,0.55) matches closely as their probabilities are almost equal. This is the result we get when x values range from 0:n

For the referral variable from the dataset we find that X take values 0 or 1

Now we get the three plots as

plotting pmf of X from Bin(120,0.433) :

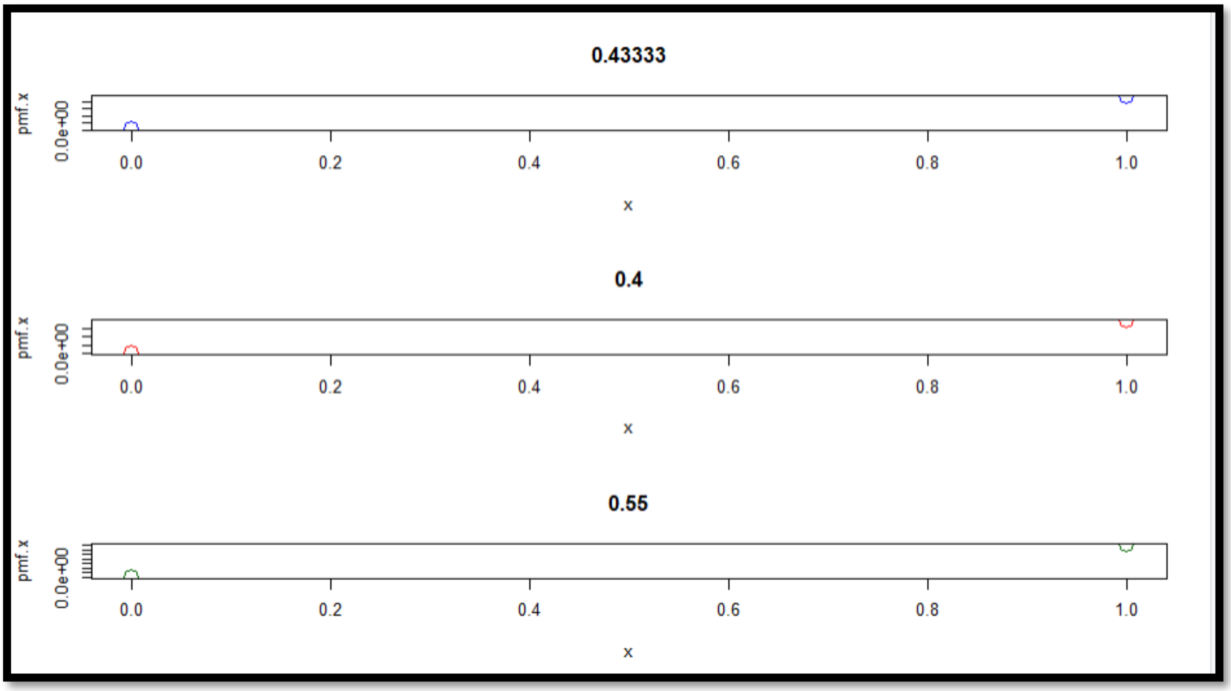
```
n <- 120
p<- 0.4333
x <- 0:1
pmf.x <- dbinom(x, n, p)
plot(x, pmf.x, main = '0.43333',col='blue', cex= 2)
```

plotting pmf of X from Bin(120,0.4) :

```
n <- 120
p<- 0.4
x <- 0:1
pmf.x <- dbinom(x, n, p)
plot(x, pmf.x, main = '0.4',col='red', cex =2 )
```

plotting pmf of X from Bin(120,0.55) :

```
n <- 120
p<- 0.55
x <- 0:1
pmf.x <- dbinom(x, n, p)
plot(x, pmf.x, main = '0.55',col='dark green', cex = 2)
# cex increases the size of data points
```



From the three plots we observe that the all three binomial distribution graphs are identical.

[Ans 3]

In answer 1 it was shown that bp variable is not exactly normally distributed. Since the sample size of the data is > 30 according to Central Limit Theorem we expect the mean of the bp variable to be normally distributed and we can perform one sample t-test comfortably.

Hypothesis testing for the true population mean bp(M) :

We formulate our hypothesis as follows:

Ho : M = 146

Ha : M != 146

- calculating standard deviation

```
 sd.BP_data1 <- sd(Assessment1_dataset_1_$bp)
sd.BP_data1
[1] 10.87283
```

- calculating standard error

```
se.BP_data1 <- sd.BP_data1/sqrt(120)
se.BP_data1
[1] 0.992549
```

- calculating mean

```
m.BP_data1 <- mean(Assessment1_dataset_1_$bp)
m.BP_data1
[1] 148.0917
```

- calculating test-statistic

```
teststat <- (m.BP_data1-146)/(se.BP_data1)
teststat
```

```
[1] 2.107369
```

- calculating p_value

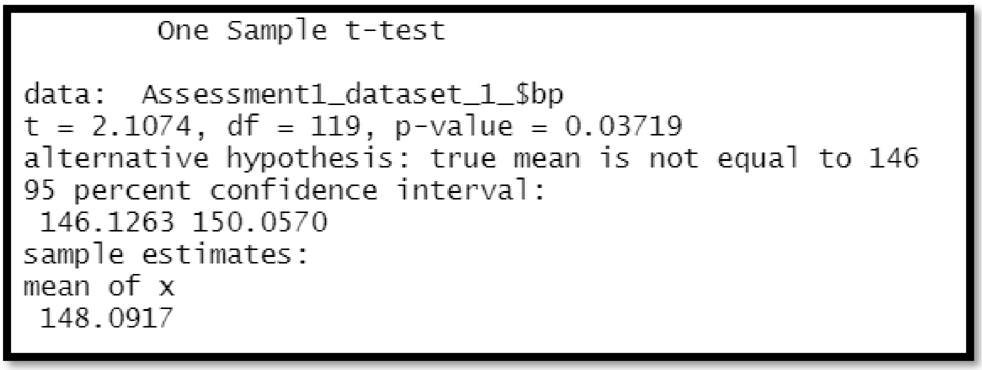
```
2*(1-pnorm(teststat))
[1] 0.03508563
```

- calculating lower and upper limit for 95% confidence interval

```
lowerlimit <- m.BP_data1 - 1.96*se.BP_data1
upperlimit <- m.BP_data1 + 1.96*se.BP_data1
lowerlimit
[1] 146.1463
Upperlimit
[1] 150.0371
```

We can confirm our above results and check whether our hypothesis holds true or not by using t.test:

```
t.test(Assessment1_dataset_1_$bp,mu=146)
```



```
        One Sample t-test

data:  Assessment1_dataset_1_$bp
t = 2.1074, df = 119, p-value = 0.03719
alternative hypothesis: true mean is not equal to 146
95 percent confidence interval:
 146.1263 150.0570
sample estimates:
mean of x
 148.0917
```

We reject our null hypothesis as p_value < 0.05

Conclusion :

There isn't sufficient evidence to prove that population mean is not equal to 146

We estimate with 95% confidence that the true mean bp of 146 is not between confidence interval of the sample mean i.e between 146.14 and 150.03.

[Ans 4]

sample size(n) = 120;

true population proportion(p) = 0.5

np = 60

n(1-p) = 60

since both np and n(1-p) > 5 we can conduct one proportion test.

Since sample size > 30 according to Central Limit Theorem we assume the mean of sex variable to be normally distributed

Hypothesis testing for the proportion(p) :

We formulate our hypothesis as follows:

Ho : p = 0.5

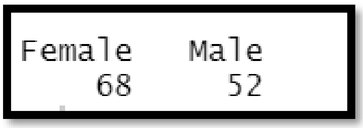
Ha : p != 0.5

As with sample means, we can apply the central limit theorem: the sampling distribution of a proportion is normal with mean equal to the true proportion p, and variance p(1-p)/p. Under the null hypothesis, therefore, the variance is by central limit theorem:

```
variance(v) = p(1-p)/n

v = 0.5*(1-0.5)/120

print(v)

[1] 0.002083333
```

```
se = sqrt(v)    # Calculating standard error

print(se)

[1] 0.04564355
```

```
table(Assessment1_dataset_1_$sex)
```



```
Female   Male
    68     52
```

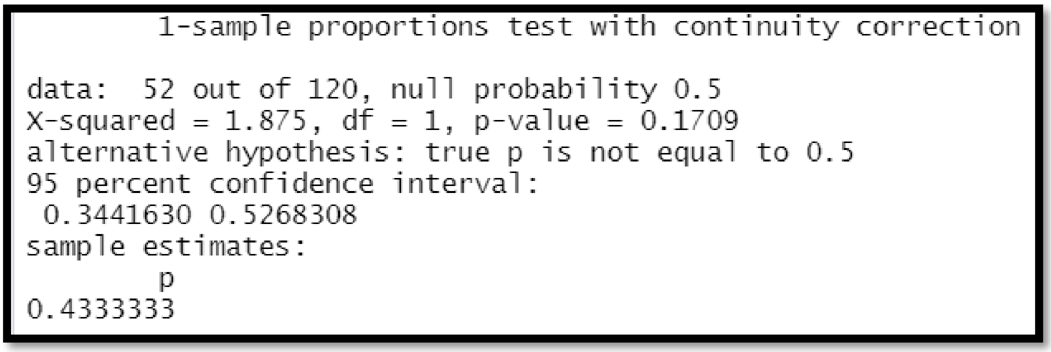
```
p1= 52/120    # Calculating the probability of male patients

print(p1)

[1] 0.4333333
```

```
test_statistic = (p1-0.5)/se   # Calculating test-statistic

print(test_statistic)

[1] -1.460593
```

```
P_value = 2*(1-pnorm(test_statistic))   # Calculating p_value

print(P_value)

[1] 1.855873
```

We can confirm our above results and check whether our hypothesis holds true or not by using prop.test:

```
prop.test(52, n= 120, p= 0.5)
```



```
        1-sample proportions test with continuity correction

data:  52 out of 120, null probability 0.5
X-squared = 1.875, df = 1, p-value = 0.1709
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3441630 0.5268308
sample estimates:
        p 
0.4333333 
```

Here too p_value > 0.05 so accept null hypothesis.(R is making a ‘continuity correction’ so we are getting difference in P-value)

Conclusion :

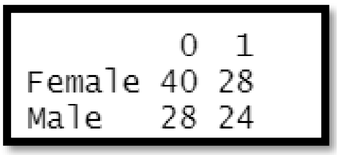
There isn't sufficient evidence to prove that population propotion of male patients is 0.5.

We estimate with 95% confidence that the population propotion of male patients is between confidence interval of the sample male proportion i.e between 0.344 and 0.53

[Ans 5]

table(Assessment1_dataset_1_$sex,Assessment1_dataset_1_$Referral)

# Sample Data



```
        0  1
Female 40 28
Male   28 24
```

For calculating the odds ratio of association between sex and referral we use the formula:

Odds_Ratio = (axd)/(bxc)

Odds_Ratio = (40*24)/(28*28)

```
[1] 1.22449
```

```
print(Odds_Ratio)
```

From the Odd_ratio value we interpret it as the odds of female not getting referral are 1.22 times the odds of male not getting referral.

Odds_ratio and its interpretation can be achieved in a easier way by using twoby2 function which is done below.

Hypothesis testing to understand the relationship between sex and referral :

We formulate our hypothesis as follows:

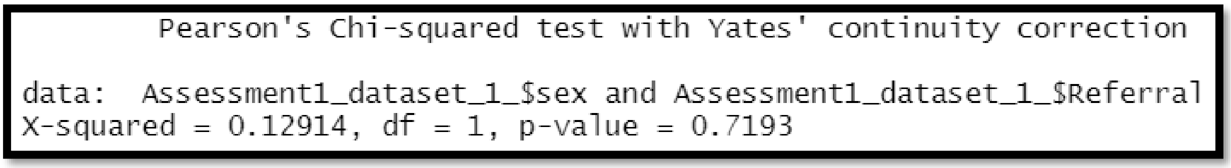
H0 : no association between referral and sex (odds ratio = 1)

HA : an association between referral and sex

For 2x2 table we use Chi-Squarred test to check whether our hypothesis holds true or not.

Using Chi-squared test :

```
chisq.test(Assessment1_dataset_1_$sex,Assessment1_dataset_1_$Referral,correct=TRUE)
```



```
        Pearson's Chi-squared test with Yates' continuity correction

data:  Assessment1_dataset_1_$sex and Assessment1_dataset_1_$Referral
X-squared = 0.12914, df = 1, p-value = 0.7193
```

p_value is > 0.5 so we accept the null hypothesis.

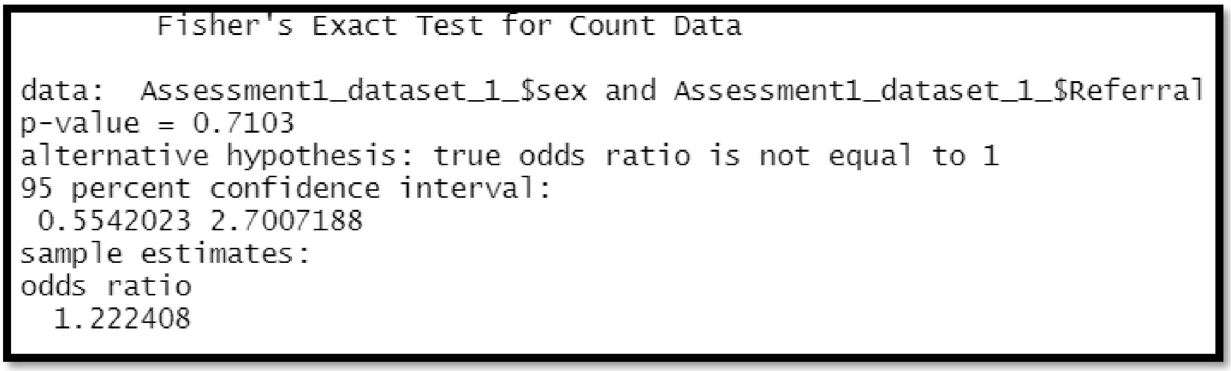
Conclusion:

There isn't sufficient evidence to prove that there is no association between referral and sex.

We can calculate the odds ratio and get its 95% confidence interval by fishers exact test and by using twoby2 function.

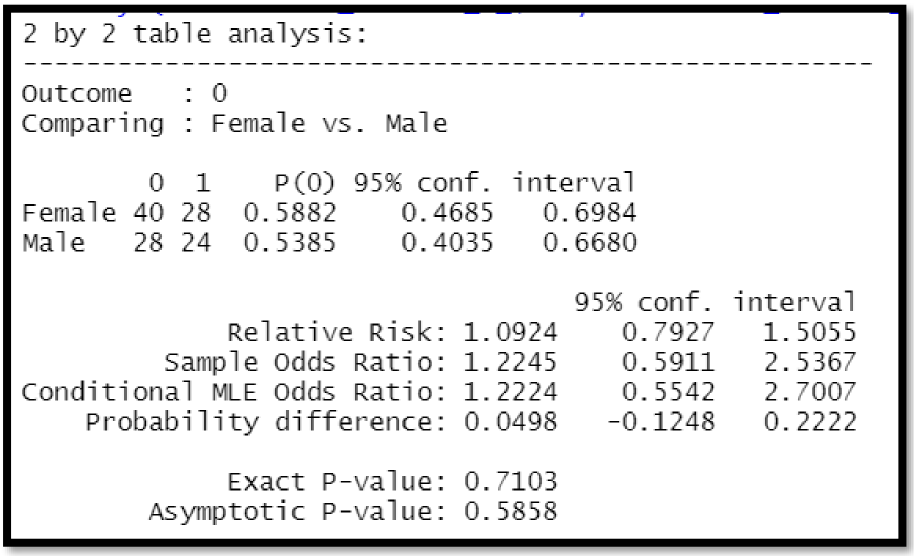
Using Fishers Exact Test:

```
fisher.test(Assessment1_dataset_1_$sex,Assessment1_dataset_1_$Referral)
```



```
        Fisher's Exact Test for Count Data

data:  Assessment1_dataset_1_$sex and Assessment1_dataset_1_$Referral
p-value = 0.7103
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.5542023 2.7007188
sample estimates:
odds ratio
  1.222408
```

```
install.packages('Epi')    # Installing ‘Epi’ package to run twoby2 function

library('Epi')

twoby2(Assessment1_dataset_1_$sex,Assessment1_dataset_1_$Referral)
```



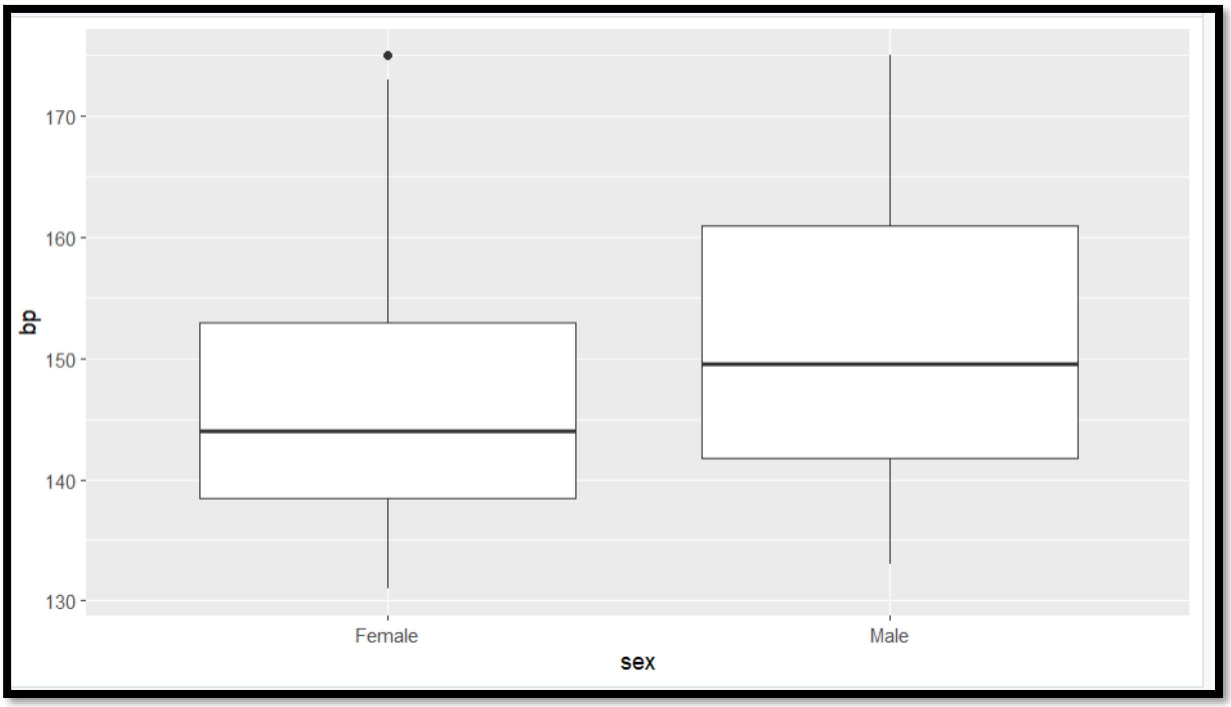
```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : 0
Comparing : Female vs. Male

        0  1    P(0) 95% conf. interval
Female 40 28  0.5882    0.4685   0.6984
Male   28 24  0.5385    0.4035   0.6680

                                   95% conf. interval
             Relative Risk: 1.0924    0.7927   1.5055
         Sample Odds Ratio: 1.2245    0.5911   2.5367
Conditional MLE Odds Ratio: 1.2224    0.5542   2.7007
    Probability difference: 0.0498   -0.1248   0.2222

             Exact P-value: 0.7103
        Asymptotic P-value: 0.5858
```

twoby2 function gives a better overall analysis of our sample data

All the above approches accepts our null hypothesis as p_value > 0.05 and thus provides the statistical evidence that there is no association between sex and referral.

[Ans 6]

```
ggplot(Assessment1_dataset_1_, aes(x=sex, y=bp)) + geom_boxplot()
par(mfrow=c(1,1)
```



The above boxplot shows the mean difference in the blood pressure levels of men and women.We can do hypothesis testing to confirm this finding.

There is skewness in both the groups which indicates that the data is not normally distributed. However, since the sample size of the dataset is > 30 according to Central Limit Theorem we can assume their means to be normally distributed and we can use two sample t-test comfortably.

Hypothesis testing for Two population means(M1 and M2) :

Let M1 and M2 be the population means of female and male groups respectively

We formulate our hypothesis as follows:

Ho: M1 = M2

Ha: M1 != M2

To perform two sample test the variances of both the groups should be approximately equal. We can test this using F-test.

let V1 and V2 be the Variances of female and male groups respectively

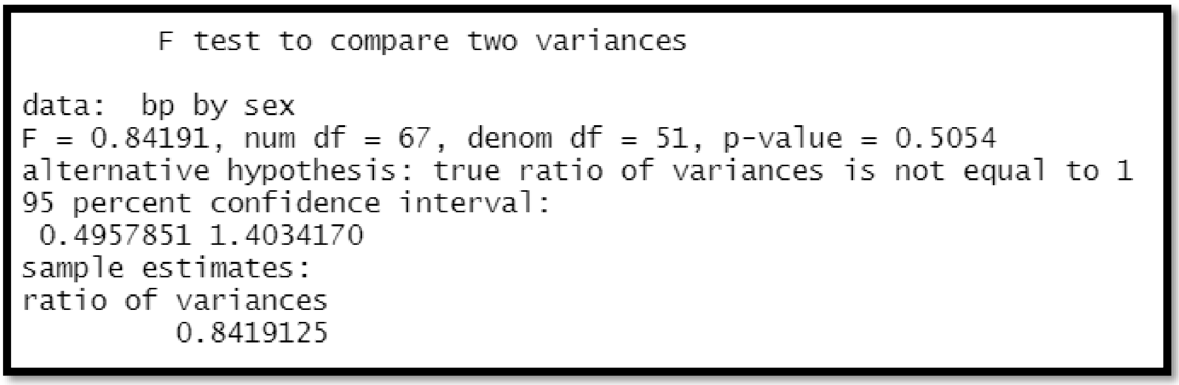
We formulate our hypothesis as follows:

Ho: V1 = V2

Ha: V1 != V2

F-test to compare two variances :

`var.test(bp~sex,data=Assessment1_dataset_1_)`



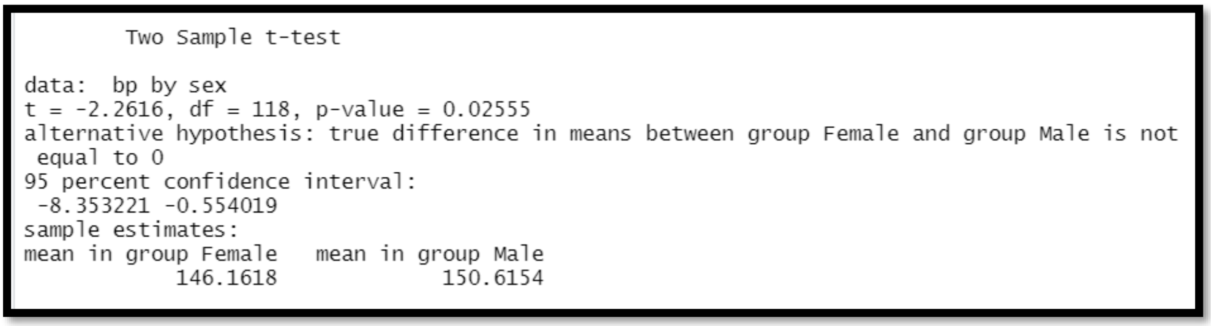
```
        F test to compare two variances

data:  bp by sex
F = 0.84191, num df = 67, denom df = 51, p-value = 0.5054
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4957851 1.4034170
sample estimates:
ratio of variances 
         0.8419125 
```

p_value > 0.05 so we can assume variances are equal

We can now do t-test since we have assumed variances of two groups to be equal

Two sample t-test :

`t.test(bp~sex,data=Assessment1_dataset_1_,var.equal=TRUE)`



```
        Two Sample t-test

data:  bp by sex
t = -2.2616, df = 118, p-value = 0.02555
alternative hypothesis: true difference in means between group Female and group Male is not
 equal to 0
95 percent confidence interval:
 -8.353221 -0.554019
sample estimates:
mean in group Female   mean in group Male 
            146.1618             150.6154 
```

p_value is less than 0.05 thus we reject null hypothesis.

Conclusion:

There isn't sufficient evidence to prove that the population means of female and male groups are different.

Thus for understanding the statistical evidence of blood pressure difference between men and women we used Two sample t-test.

[Ans 7]

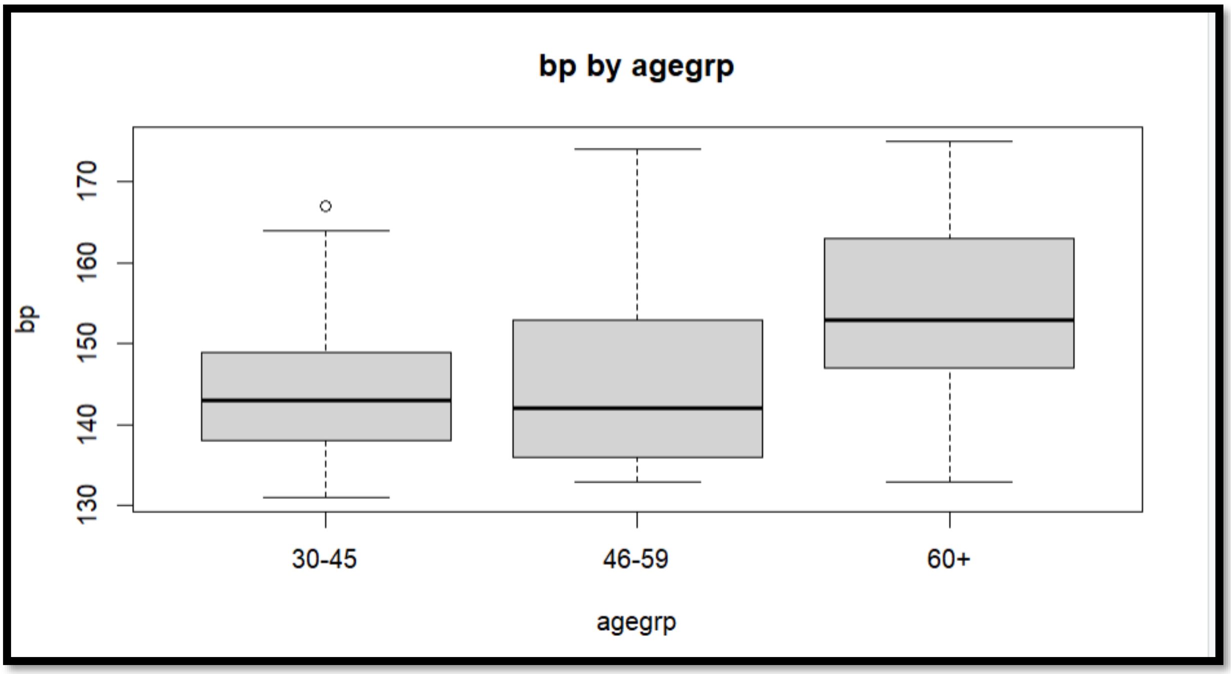
```
x1 = subset(Assessment1_dataset_1_,agegrp== '30-45' )
count(x1) ## n>30
```

```
x2 = subset(Assessment1_dataset_1_,agegrp== '46-59' )
count(x2) ##n >30
```

```
x3 = subset(Assessment1_dataset_1_,agegrp== '60+' )
count(x3) ##n > 30
```

From above we understand that the sample data collected for each group has large sample size and we can apply Central Limit Theorem on them.

```
boxplot(bp~agegrp,data=Assessment1_dataset_1_, main="bp by agegrp", xlab="agegrp", ylab="bp")
```

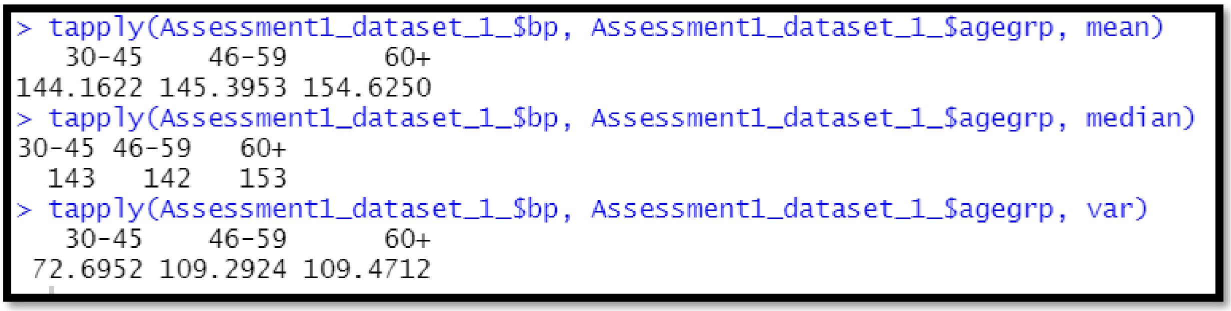


From the boxplot we understand that bp is not similarly distributed across each agegroup and The distribution appears to be skewed for all the groups and so data for each group doesnt appear to be normally distributed(unsure of statistical significance).The 60+ agegroup has the highest blood pressure values than other agegroups

tapply(Assessment1_dataset_1_$bp, Assessment1_dataset_1_$agegrp, mean)

tapply(Assessment1_dataset_1_$bp, Assessment1_dataset_1_$agegrp, median)

tapply(Assessment1_dataset_1_$bp, Assessment1_dataset_1_$agegrp, var)

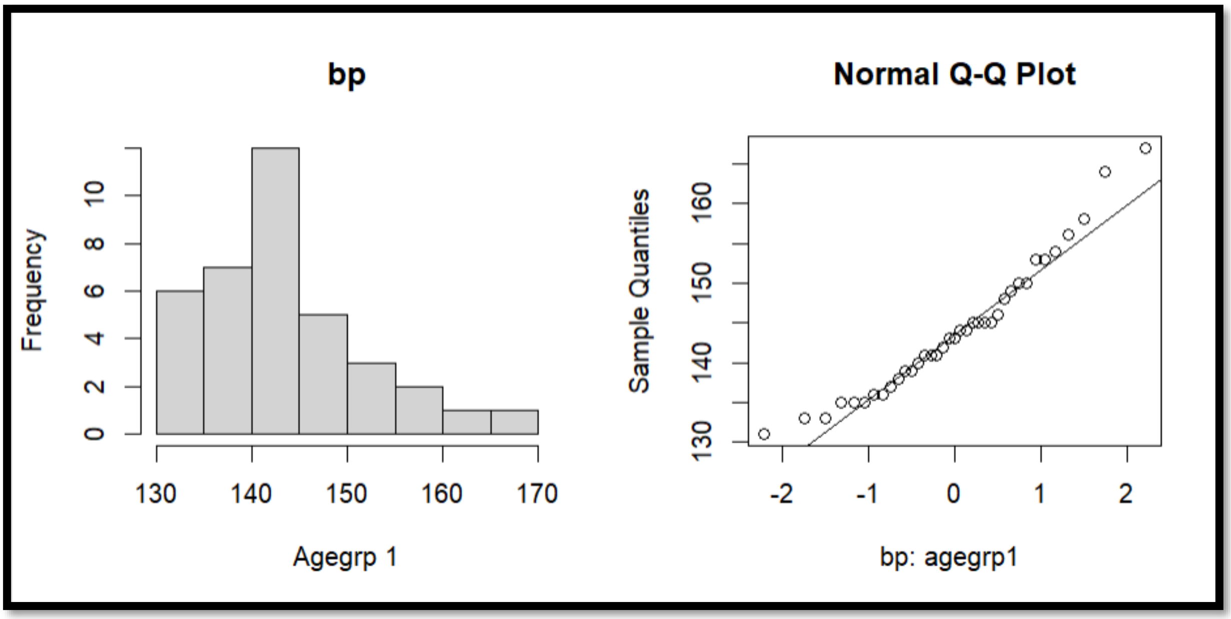


```
> tapply(Assessment1_dataset_1_$bp, Assessment1_dataset_1_$agegrp, mean)
   30-45    46-59      60+ 
144.1622 145.3953 154.6250 
> tapply(Assessment1_dataset_1_$bp, Assessment1_dataset_1_$agegrp, median)
30-45 46-59   60+ 
  143   142   153 
> tapply(Assessment1_dataset_1_$bp, Assessment1_dataset_1_$agegrp, var)
   30-45    46-59      60+ 
 72.6952 109.2924 109.4712 
```

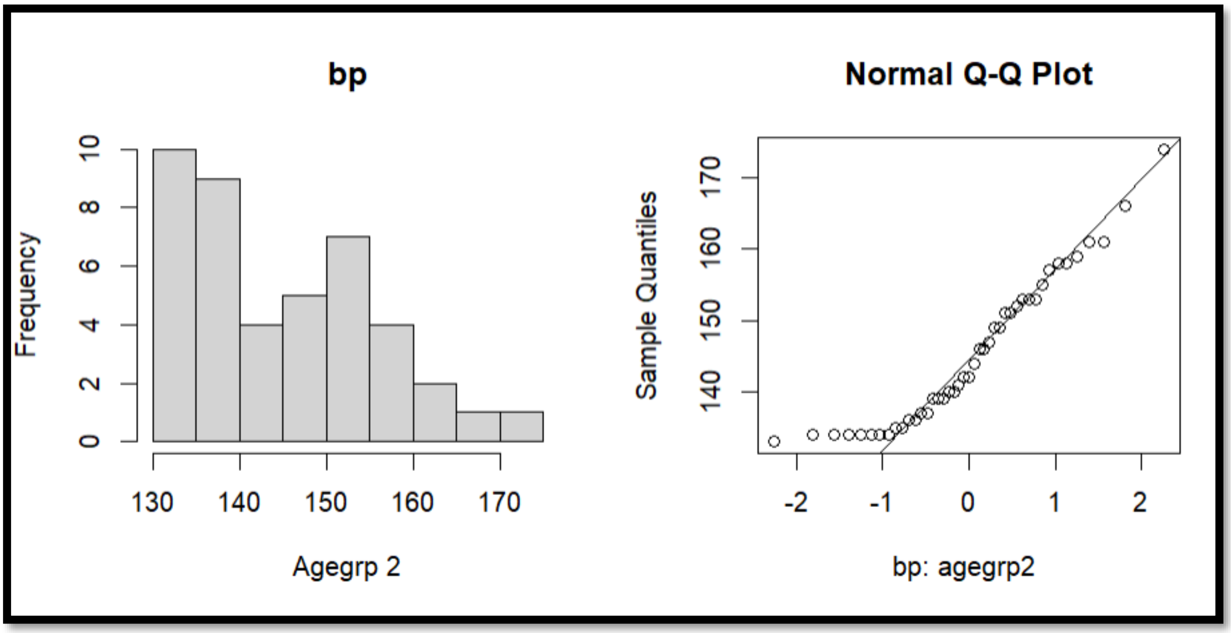
Eventhogh there is no clear clarity of the statistical signifance, the mean bp and median bp for a specific age group calculated are not exactly the same which indicates the presence of skewness in each sample data collected.

This can be confirmed further by plotting histograms and Normal q-q plots.

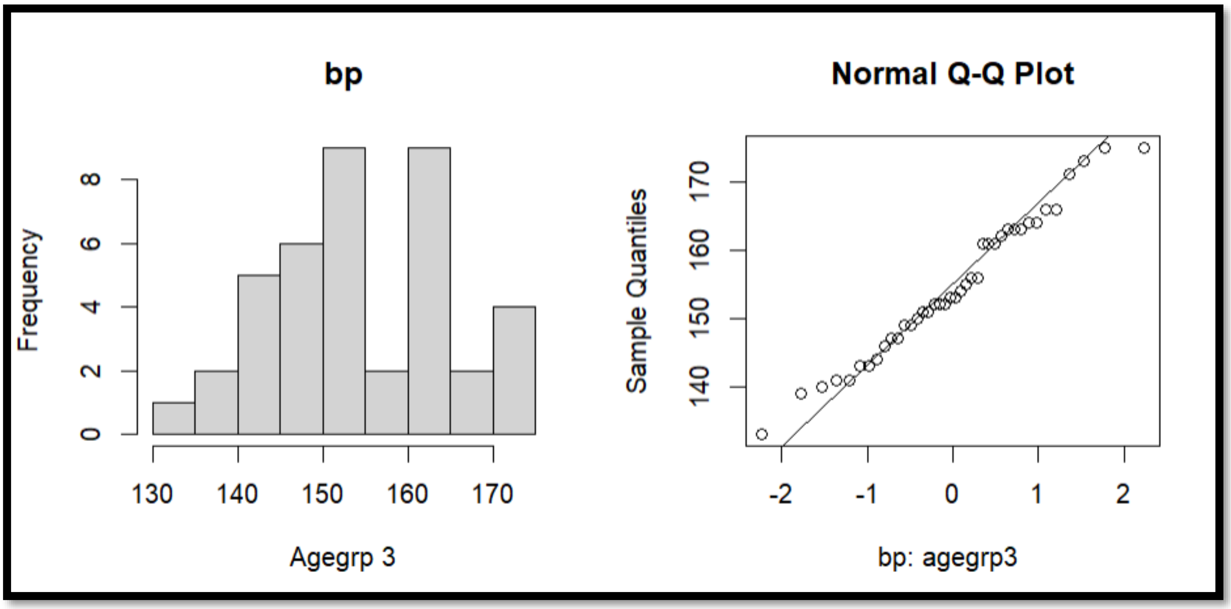
```
hist(Assessment1_dataset_1_$bp[Assessment1_dataset_1_$agegrp== '30-45'], xlab = "Agegrp 1", main="bp")

qqnorm(Assessment1_dataset_1_$bp[Assessment1_dataset_1_$agegrp== '30-45'],xlab="bp: agegrp1 ")

qqline(Assessment1_dataset_1_$bp[Assessment1_dataset_1_$agegrp== '30-45'])
```



```
hist(Assessment1_dataset_1_$bp[Assessment1_dataset_1_$agegrp== '46-59'], xlab = "Agegrp 2", main="bp")

qqnorm(Assessment1_dataset_1_$bp[Assessment1_dataset_1_$agegrp== '46-59'],xlab="bp: agegrp2 ")

qqline(Assessment1_dataset_1_$bp[Assessment1_dataset_1_$agegrp== '46-59'])
```



```
hist(Assessment1_dataset_1_$bp[Assessment1_dataset_1_$agegrp== '60+'],
xlab = "Agegrp 3", main="bp")

qqnorm(Assessment1_dataset_1_$bp[Assessment1_dataset_1_$agegrp==
'60+'],xlab="bp: agegrp3 ")

qqline(Assessment1_dataset_1_$bp[Assessment1_dataset_1_$agegrp== '60+'])
```



As we can see the sample data collected for each age-group doesnt seem to be normally distributed. But since the the sample size is large enough we assume their mean to be normally distributed by Central Limit Theorem.

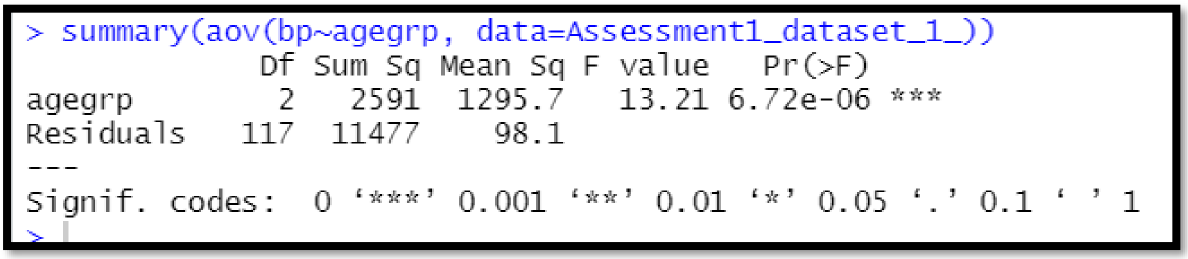
The mean values across each group appear to be different. We try to understand its statistical significance by doing the ANOVA test.

ANOVA analysis to test the null hypothesis :

Ho: M1=M2=M3

H1: M1!=M2 or M1!=M3 or M2!=M3

```
summary(aov(bp~agegrp, data=Assessment1_dataset_1_))
```



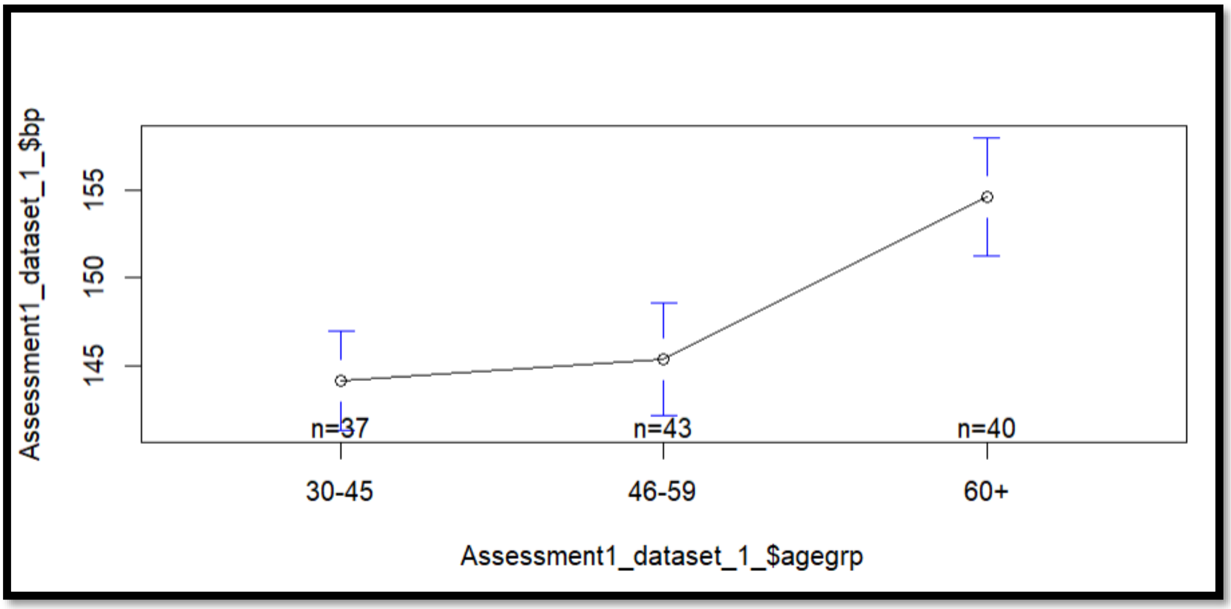
```
> summary(aov(bp~agegrp, data=Assessment1_dataset_1_))
             Df Sum Sq Mean Sq F value   Pr(>F)    
agegrp        2   2591  1295.7   13.21 6.72e-06 ***
Residuals   117  11477    98.1                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

p_value is '6.72e-06' which is very small compared to 0.05 hence we reject the null hypothesis.

Conclusion:

There isn't sufficient evidence to conclude that the mean bp is different in each age-group.

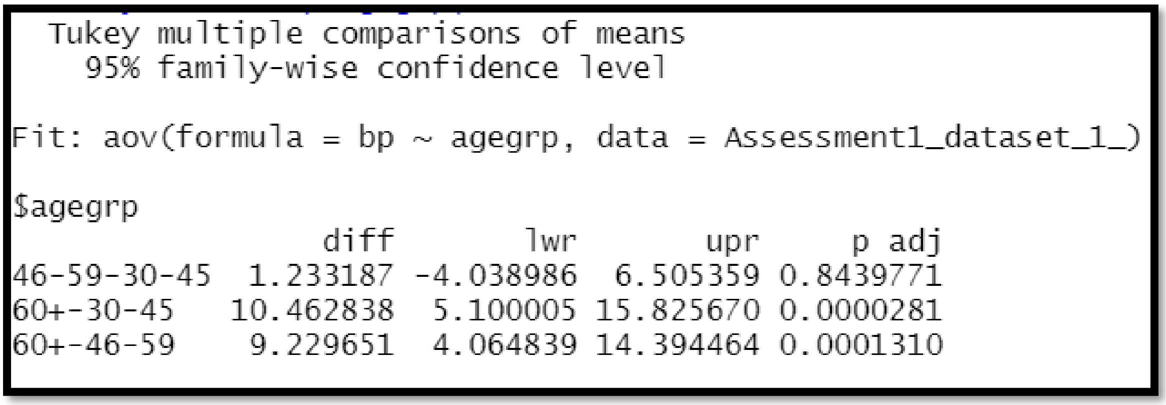
```
install.packages('gplots')
library("gplots")
plotmeans(Assessment1_dataset_1_$bp~Assessment1_dataset_1_$agegrp)
```



We get a plot of estimated bp means with 95% confidence interval for each age-group. The line joining the means of each group shows clearly that the means are different for each age-group.

This can be further strengthened by using Tukey's Adjusted Multiple Comparisons technique.

TukeyHSD(aov(bp~agegrp, data=Assessment1_dataset_1_))

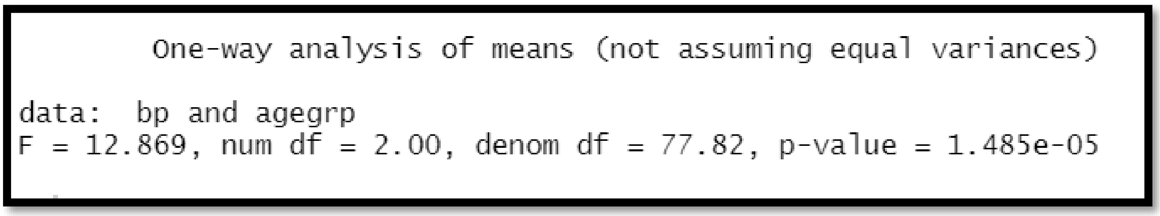


```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bp ~ agegrp, data = Assessment1_dataset_1_)

$agegrp
                 diff        lwr       upr     p adj
46-59-30-45  1.233187 -4.038986  6.505359 0.8439771
60+-30-45   10.462838  5.100005 15.825670 0.0000281
60+-46-59    9.229651  4.064839 14.394464 0.0001310
```

The 'diff' column shows the difference in mean among each age-group which further shows the statistical evidence of blood pressure difference between age groups.

We observed that variance is different in each age-group, by doing the Welch's ANOVA we can test its statistical significance.We wish to relax the assumption of equal variances.

```
oneway.test(bp~agegrp, data=Assessment1_dataset_1_)
```



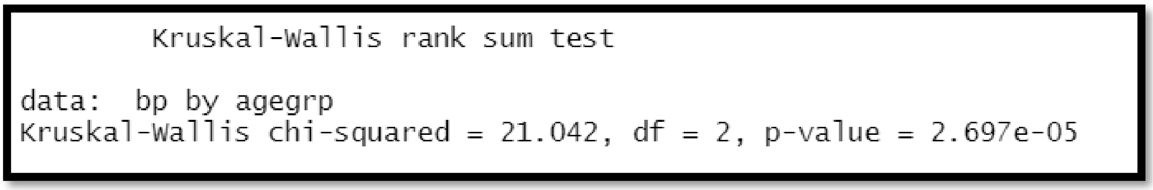
```
        One-way analysis of means (not assuming equal variances)

data:  bp and agegrp
F = 12.869, num df = 2.00, denom df = 77.82, p-value = 1.485e-05
```

The p-value is smaller than 0.05 suggesting that any differences in variance across groups are not large enough to have affected the outcome of the ANOVA analysis.

To avoid the doubt of assuming normally distributed data, we use the Kruskall-Wallis test

Testing the null hypothesis of equal medians across three groups

```
kruskal.test(bp~agegrp, data=Assessment1_dataset_1_)
```



```
        Kruskal-Wallis rank sum test

data:  bp by agegrp
Kruskal-Wallis chi-squared = 21.042, df = 2, p-value = 2.697e-05
```

The p_value is smaller than 0.05 and hence reject the null hypothesis.

All the above tests provides the statistical evidence of a blood pressure difference between age groups.