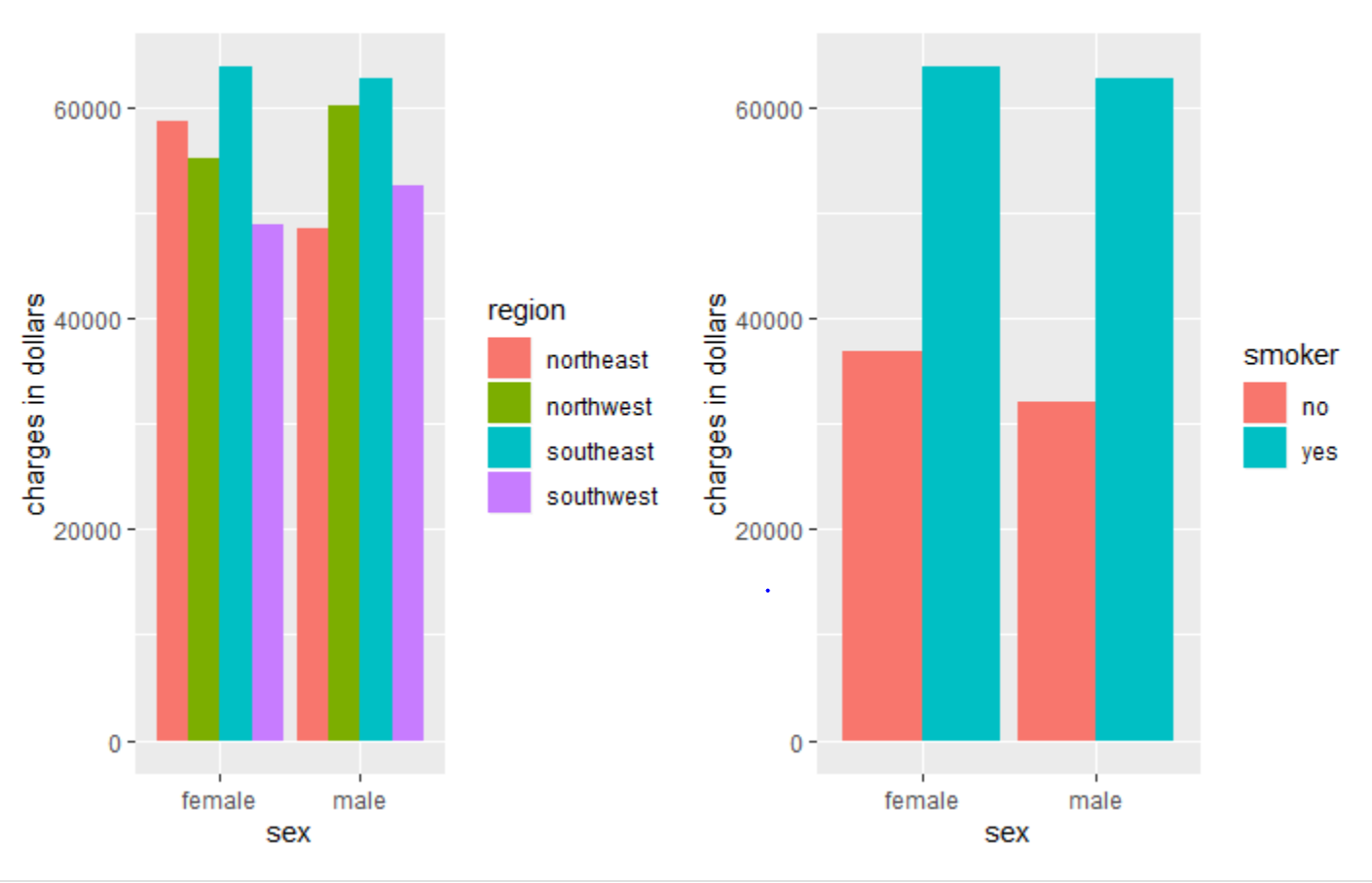
CASE STUDY REPORT

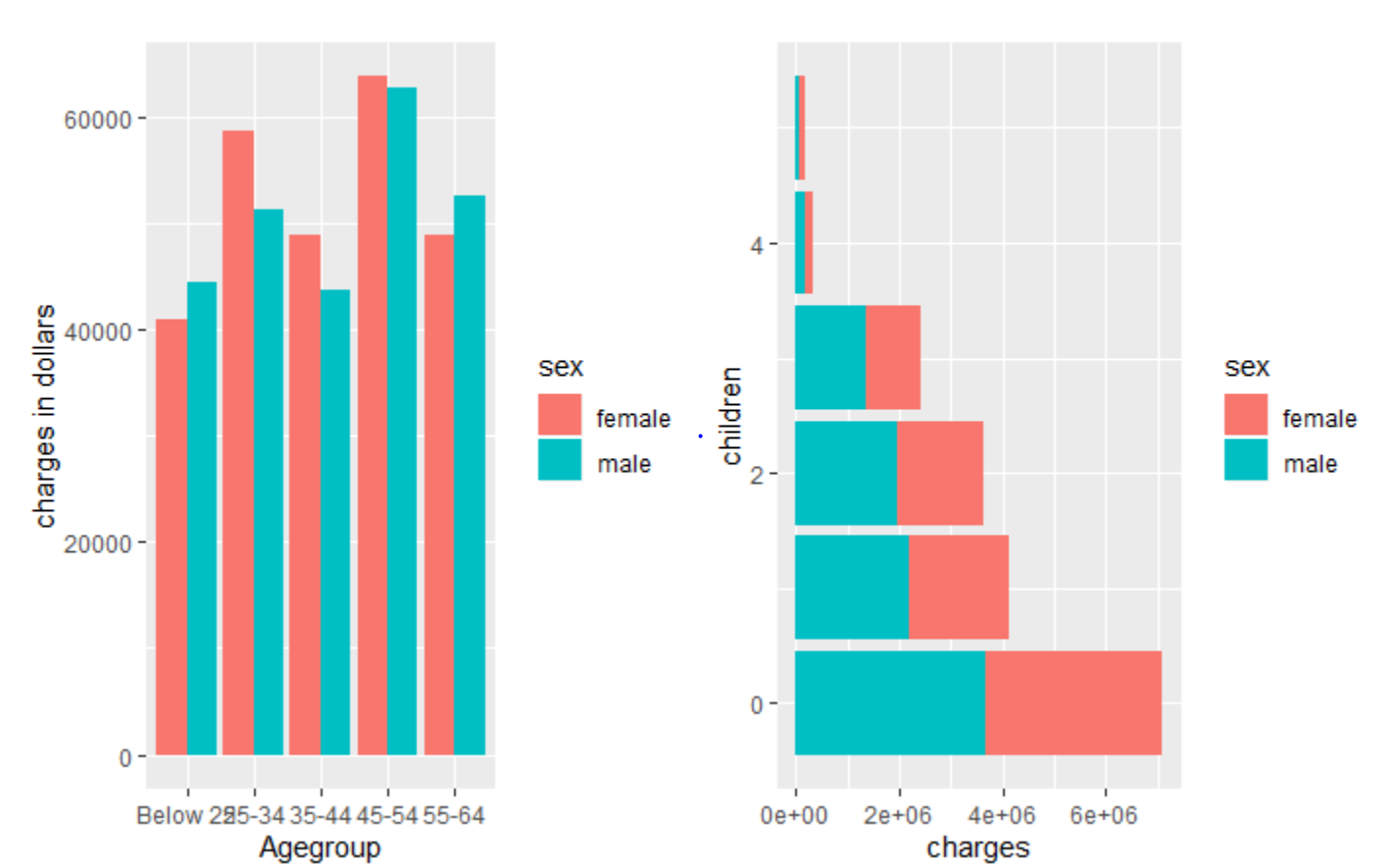
About:

The venerable insurance industry is no stranger to data driven decision making. Yet in today's rapidly transforming digital landscape, Insurance is struggling to adapt and benefit from new technologies compared to other industries, even within the Banking, financial services and insurance (BFSI) sphere (compared to the Banking sector for example). Despite this, emergent technologies like AI and Block Chain have brought a radical change in Insurance, and Data Analytics sits at the core of this transformation.

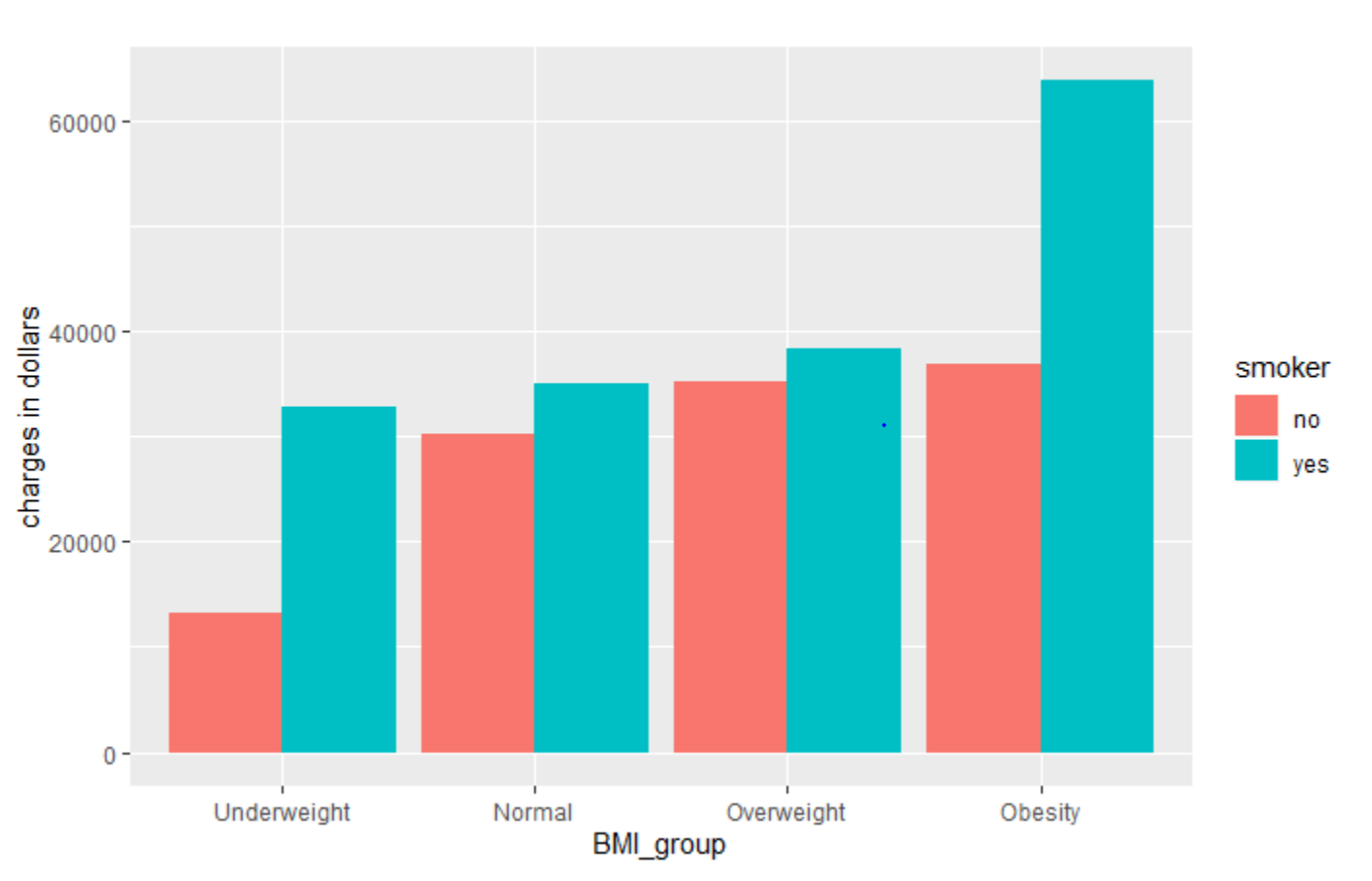
Task 1:

Perform graphical analysis on the dataset and generate 3 different graphs to gather meaningful conclusions. These graphs can be based on one attribute, multiple factors, or a portion of data.

It can be inferred from the above graph that the healthcare charges are highest in the southeast and very high if the individual is a smoker irrespective of the gender.



As apparent in the graphs above the individuals below 25 have the least premium charges of all ages and the individuals who have more kids happen to have lesser charges.



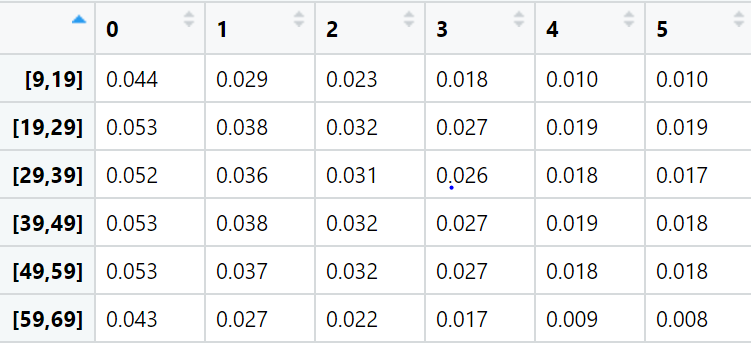
It can be seen in the above graph that smokers belonging to higher BMI groups end up paying more premium than everyone else.

TASK :2

Construct a joint distribution function for Age (with intervals of 10 yrs. old, i.e [0,9], [9,19], …) vs Number of children.

The joint probability distribution table for age intervals vs number of children can be seen in the below table.

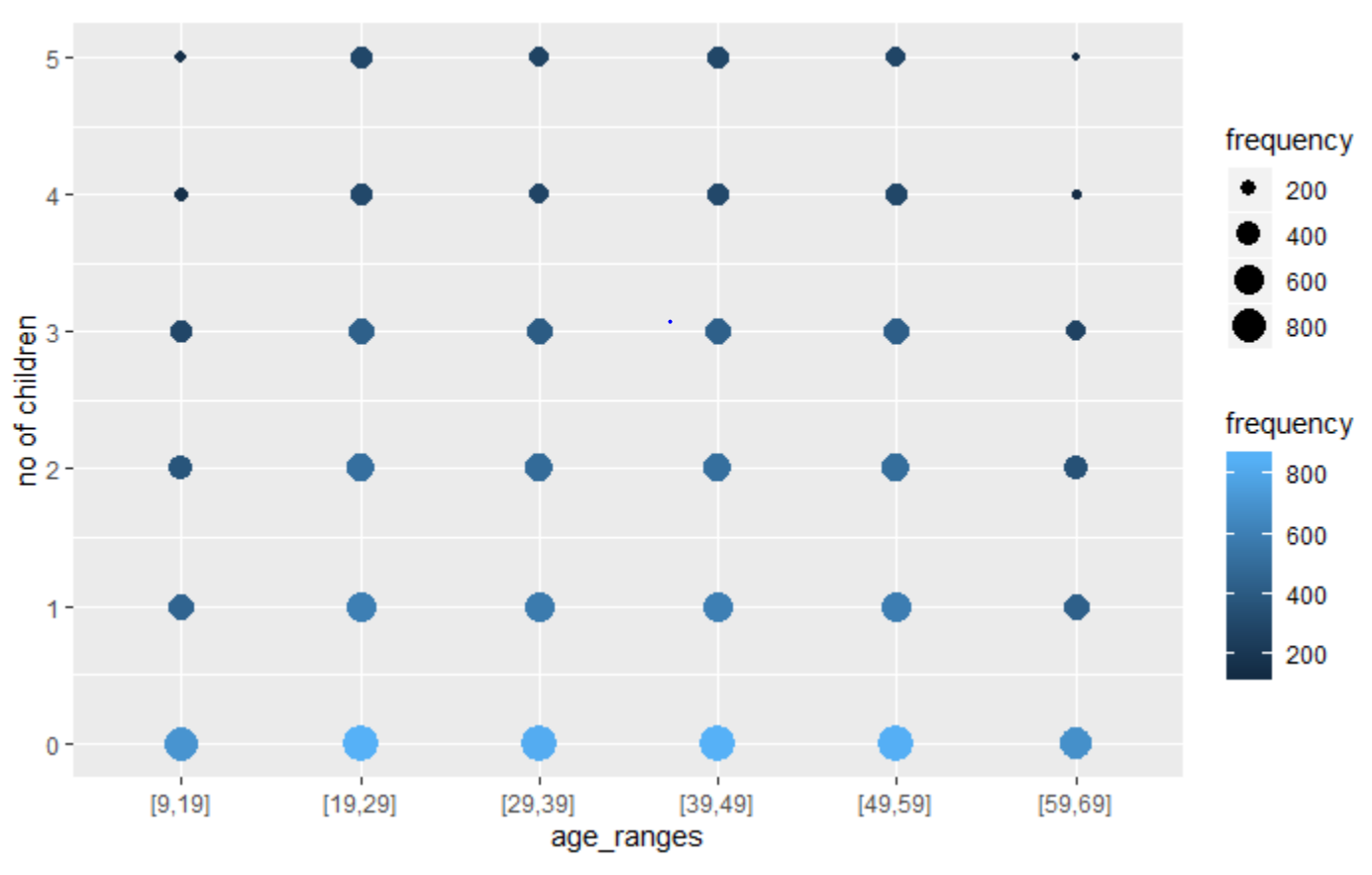
The rows denote intervals of 10 yrs. old, i.e [9,19], …[59,69] and the colums denote children i.e 0,1,…5



Task 3:

Perform correlation analysis for based on your analysis is question 2. Are age and Num of Children independent? Support your analysis by using proper graphical representation.

The below plot explains the correlation between the number of children and age groups of individuals



**Correlation Co efficient**

cor(age\_df$count, children\_df2$count)

#ans :-0.1840299

The low correlation co-efficient indicates that the two variables have a weak negative linear relationship

Task 4 :

Conduct a Test of Hypothesis (TOH) for the difference between the average charge of individuals with and without children in a particular region of your choice. Do you get the same result from its confidence interval? Add a box plot to support your analysis

Performed the Ztest on the samples obtained after filtering with respect to **with or without** children parameters.

X1 ≡ R.V. of charges from first sample

X2 ≡ R.V. of charges from second sample

TOH:

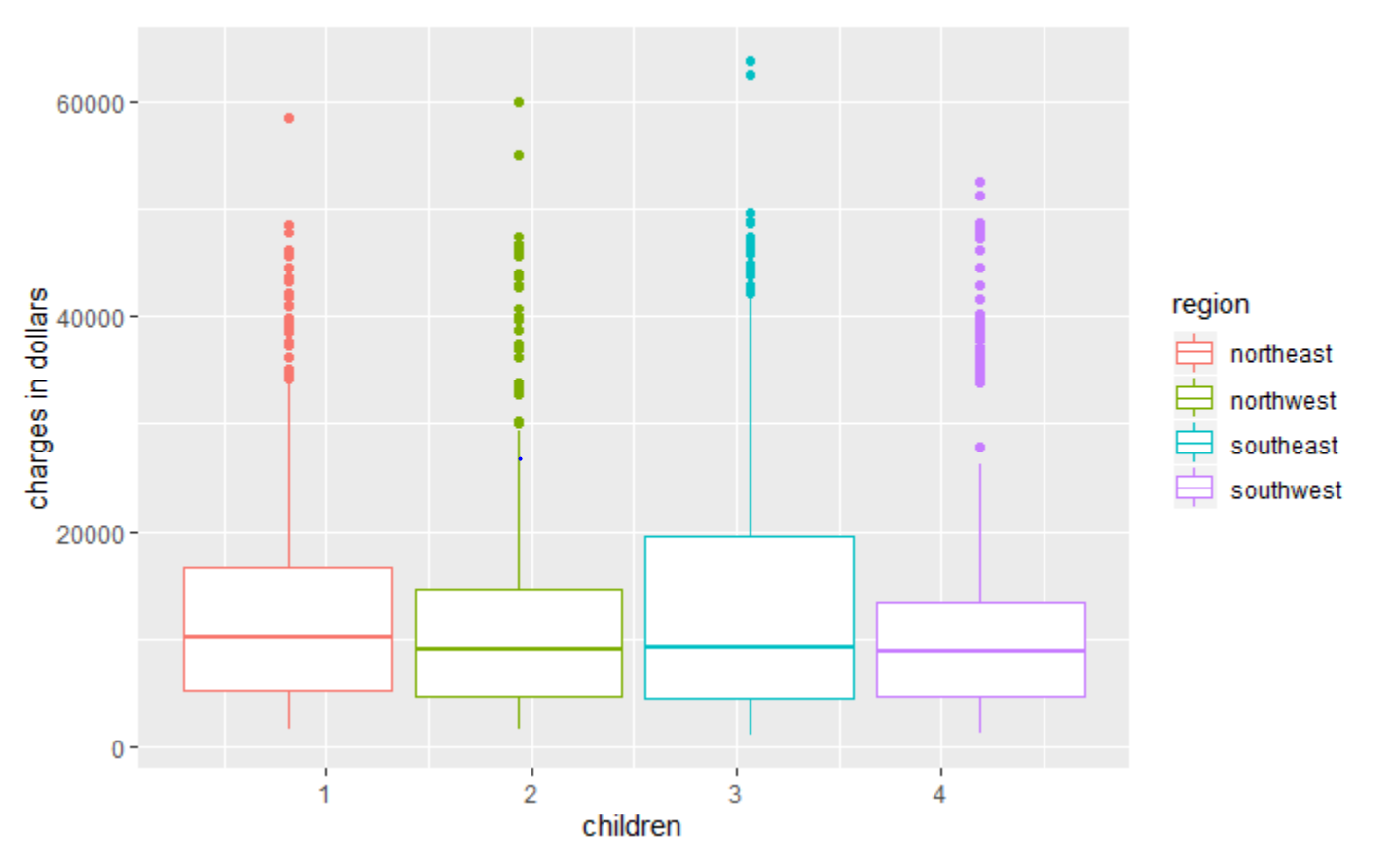
Null Hypothesis: mean 1 - mean 2 = 0

alternate Hypothesis mean 1 - mean 2 !=0

Z-test Value: 1.139751

Thus, it has been established that for a significance level of α = 0.05, we fail to reject the null hypothesis since the z-value lies within the acceptable range i.e [−1.96, 1.96] and conclude that there is no significant difference between the mean charges of two samples.

The box plot below illustrates the relationship between the number of children and the healthcare premium region wise.



Task 5 :

Conduct a Test of Hypothesis (TOH) to compare the ratio of the people who smoke in 2 different regions (of your choice).

Two samples of the regions namely “northwest” and “northeast” have been taken to conduct a test of hypothesis.

X1 = R.v of no of smokers in northwest region

X2 = R.v of no of smokers in northeast region

Hypothesis: (where p1 is the ratio of smokers in northwest and p2is the ratio of the smokers in northeast)

Null Hypothesis: p1=p2

Alternate Hypothesis: p1!=p2

Results:

z\_calc=-0.3666794

Pvalue= 0.3569291

Conclusion : Since P\_value > alpha, we fail to reject Null Hypothesis. Therefore, the ratio of smokers in northwest is not equal to the ratio of smokers in northeast.

Task 6:

Test for ratio of variances across all the regions for beneficiaries who are younger than 50 years and older than 50 years.

Two samples of the age groups namely “above\_50” and “younger\_than\_50” have been taken to conduct a test for ratio of variances.

X1 = R.v of charges of people above 50

X2 = R.v of charges of people who are younger than 50

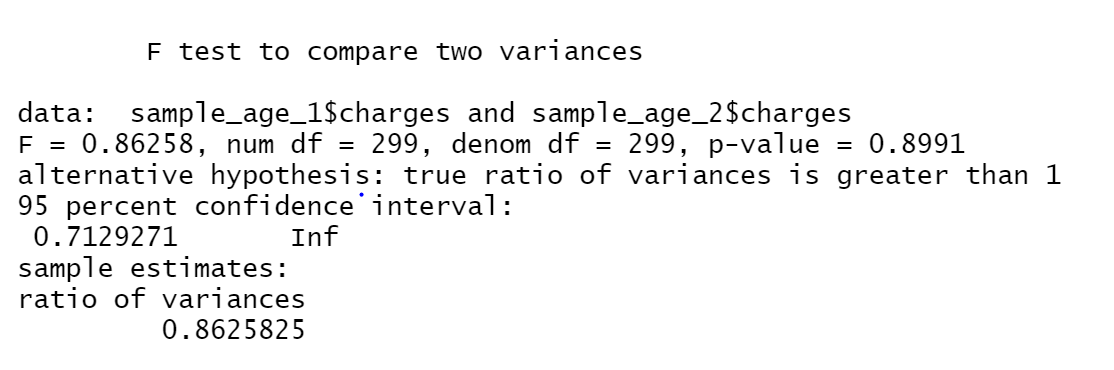
Hypothesis:

Null Hypothesis: ratio of variances for people above 50 and younger than 50 is <= 1

Alternate Hypothesis ratio of variances for people above 50 and younger than 50 is > 1

Results :

The p-value of F-test is p = 0.8991



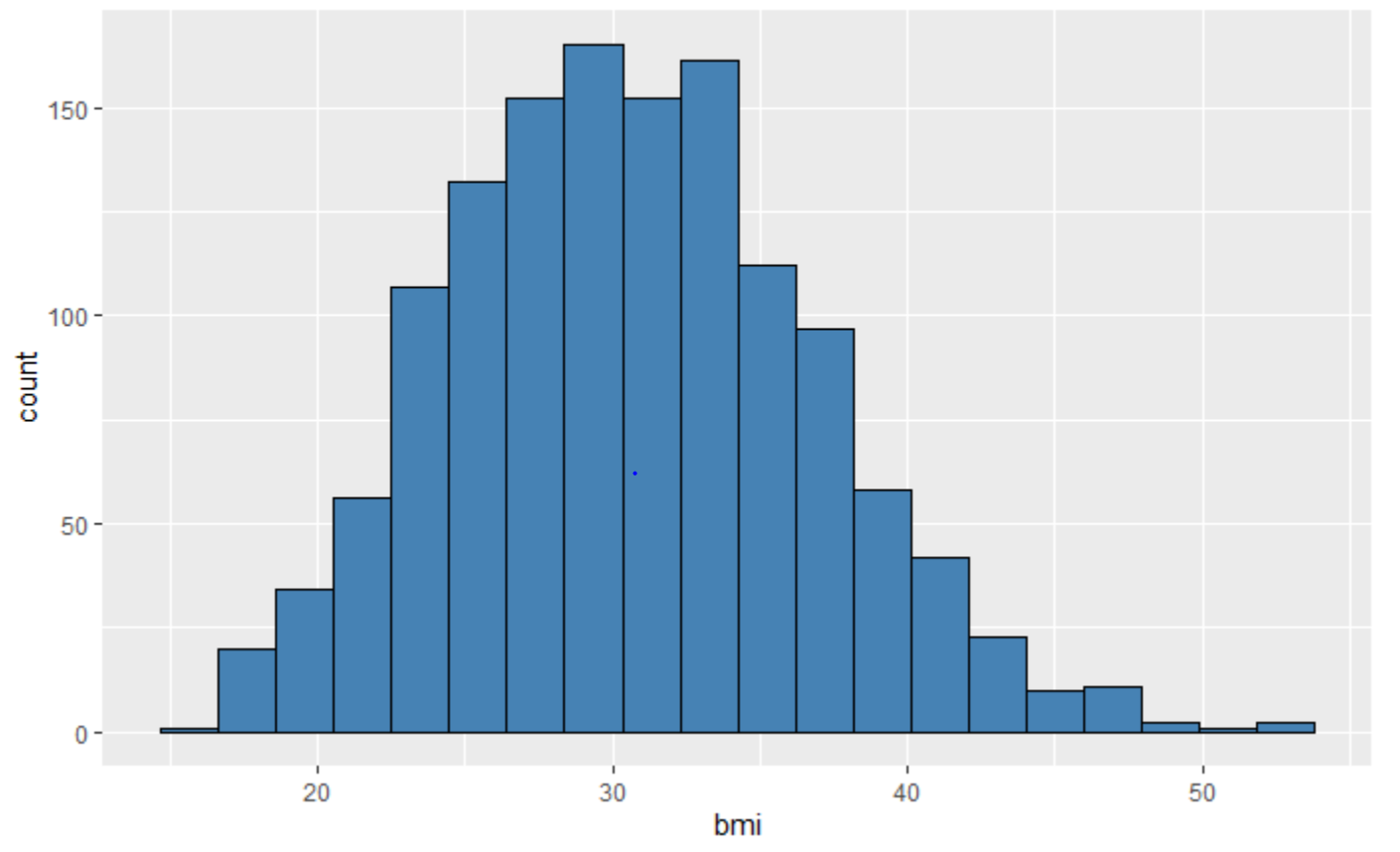
Conclusions:

As p-value of F-test is p = 0.8991 which is greater than the significance level 0.05, we fail to reject the null Hypothesis. To conclude, there is no significant difference between the two variances.

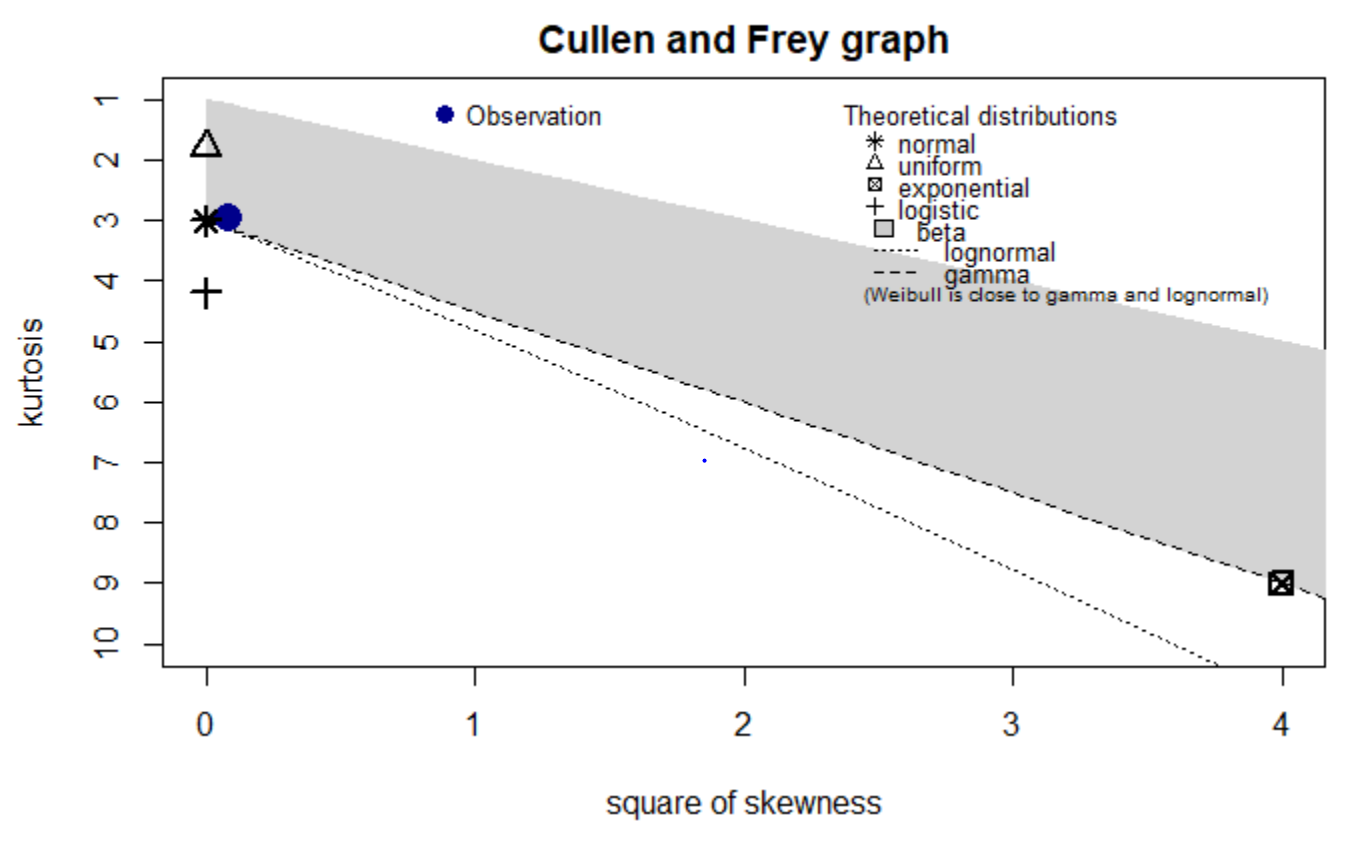
Task 7:

Perform distribution fitting on BMI and number of children and comment about their skewness, kurtosis, log likelihood, AIC and BIC values and add graphs to support your claim.

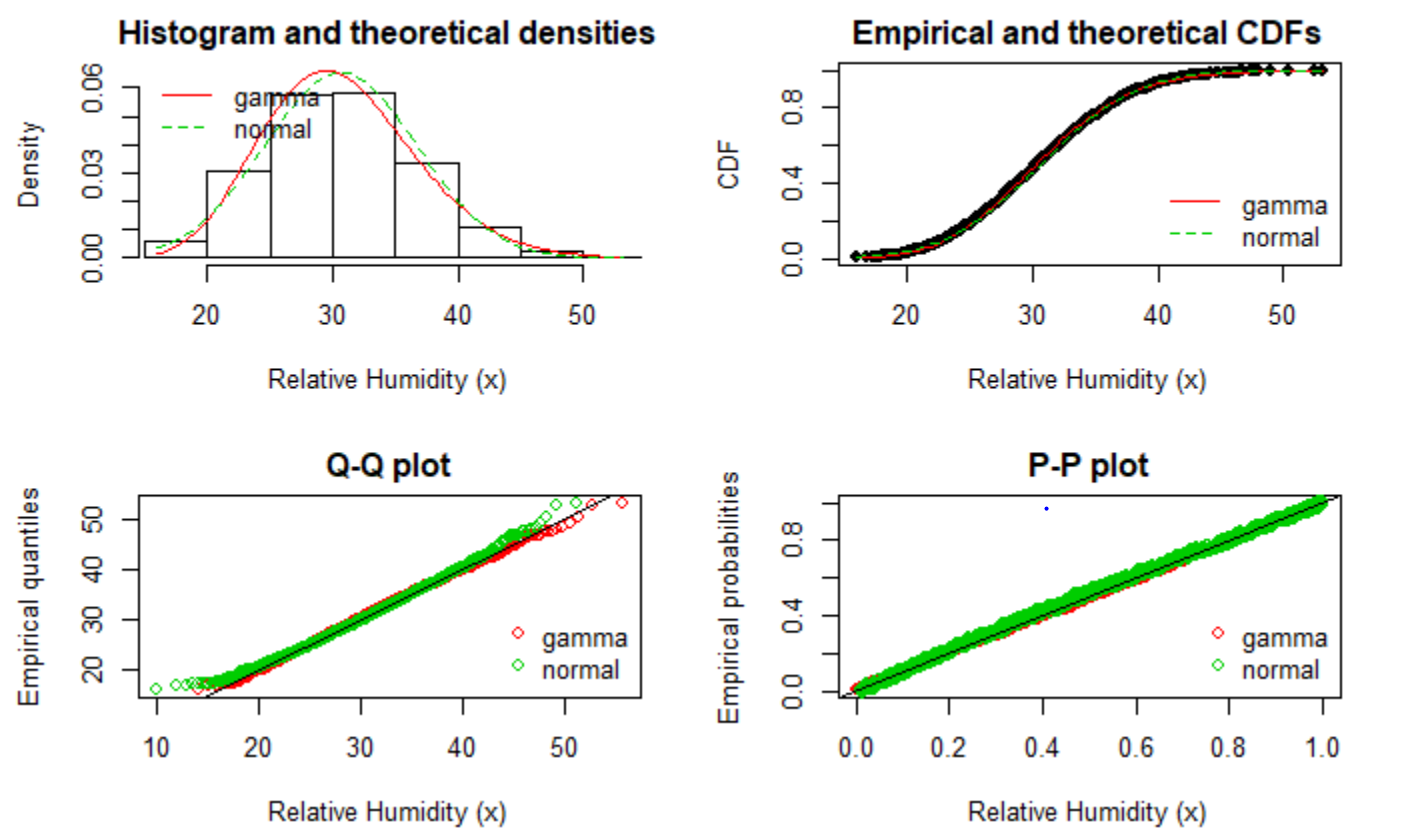
The plot below shows the visualization of BMI which is a continuous variable.



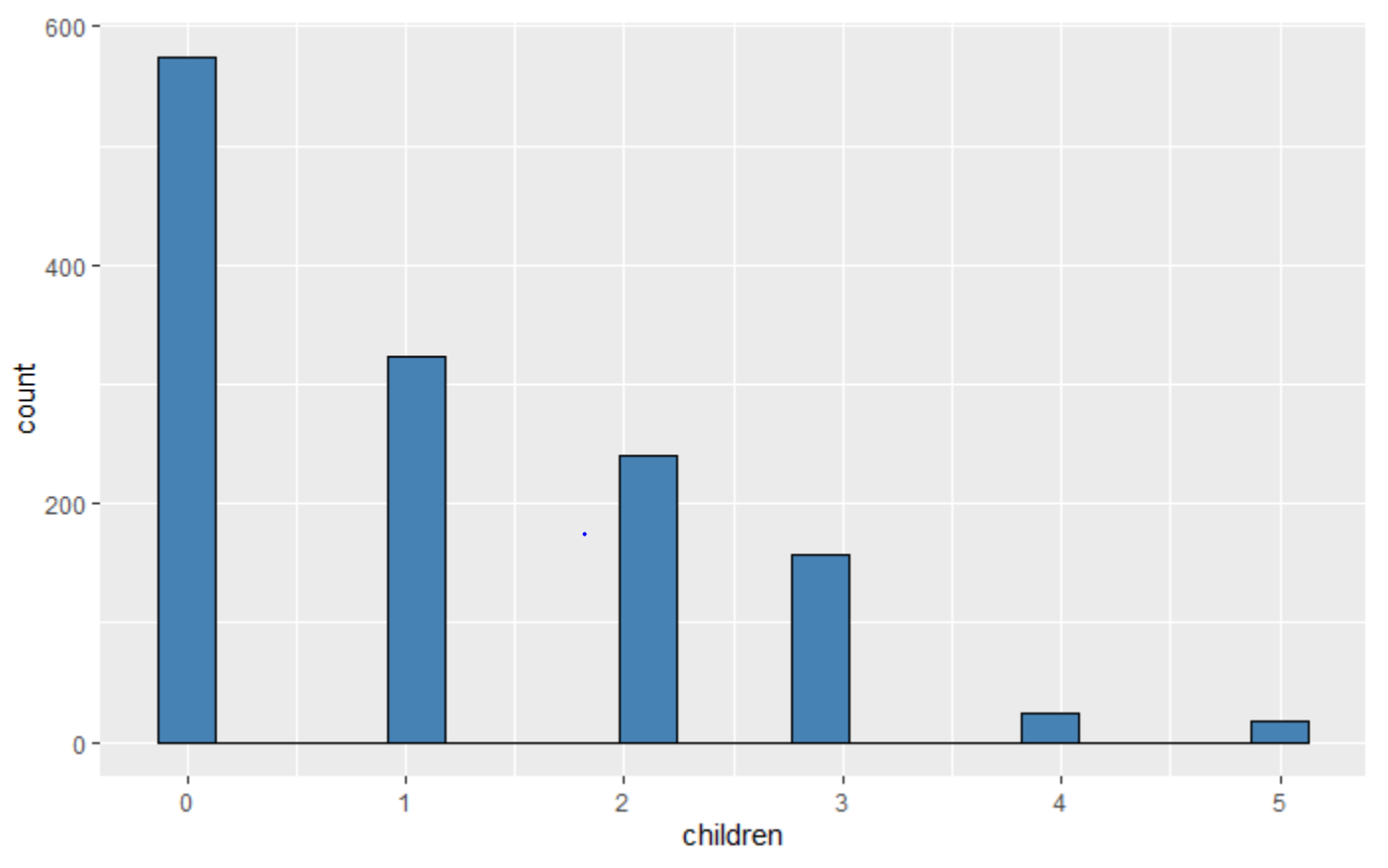
The Cullen frey graph shows that the observation is close to normal and gamma distributions.



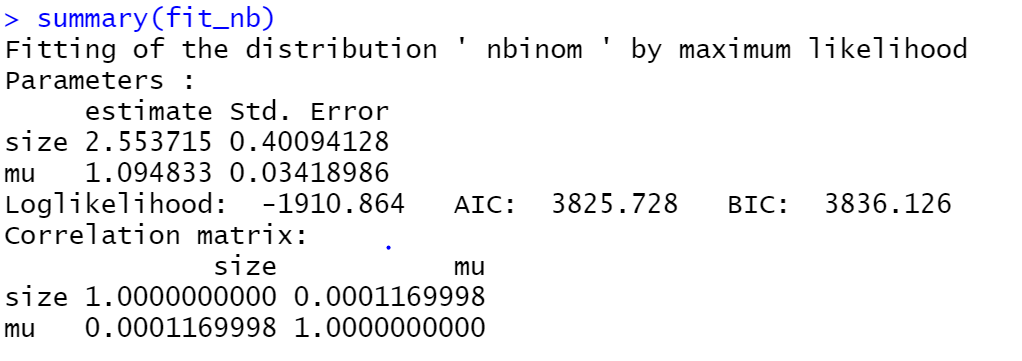
The goodness of fit plots helped tracing the distributions along side the visual data

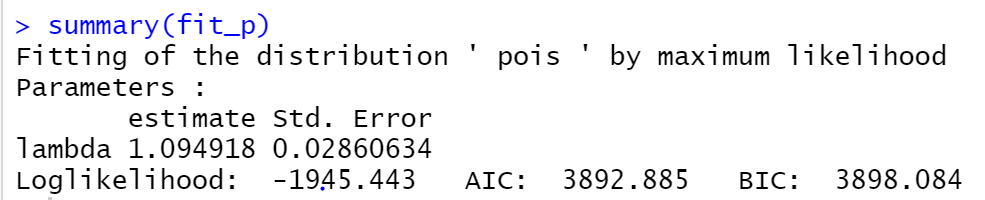


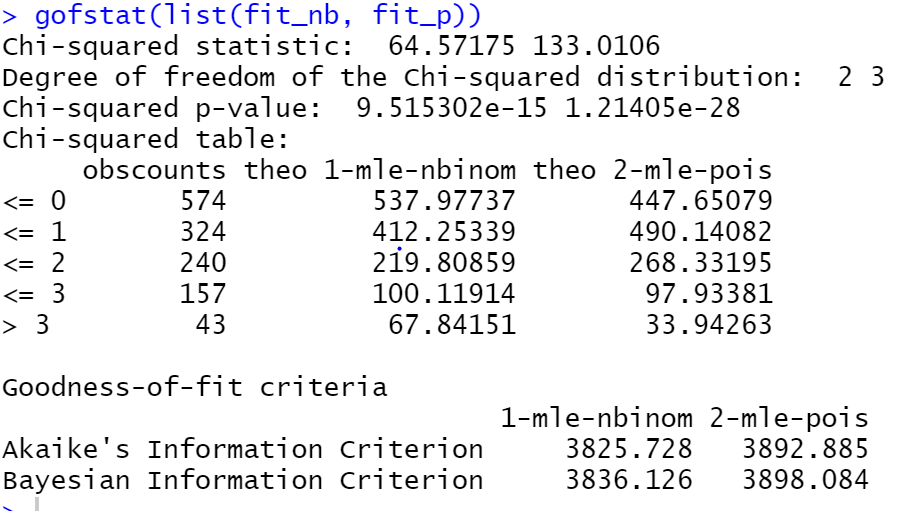
The plot below shows the visualization of no of children which is a discrete variable.



The fit distribution parameter summary of negative binomial and poisson indicate that both fit the data well but negative binomial seems to fit the data better.







The Chi-squared statistic apparently reinforces the same conclusion.

TASK 8:

One extra analysis based on your preference.

Analysis: Hypothesis testing to compare the ratio of the people who smoke for different BMI group

sample 1- Overweight individuals who smoke

sample 2- Obese individuals who smoke

The Hypothesis

p1 == proportion of Overweight individuals who smoke

p2 == proportion of Obese individuals who smoke

Null hypothesis H0 : p1 = p2

Alternative hypothesis H1 : p1 != p2

Results:

Z\_calc = -0.5282746

p-value=1.402691

Conclusions:

As the p-value is greater than alpha(0.05) , we fail to reject the null hypothesis and the proportion of Overweight individuals who smoke is equal to the proportion of Obese individuals who smoke.

Graph: indicates the number of smokers for a specific BMI catogery

