CPC351/CPM351 Principles of Data Analytics

Assignment 2

Submission Date:
09/01/2022

Lecturer Name:
Dr. Wong Li Pei

Group 29

Name	Matric No.
THANISH A/L NATARAJAN	149156
SHARVIN A/L KOGILAVANAN	148056
KATHEERAVAN A/L BALASUBRAMANIAM	147744

Question 1

1a) Changing Data Type

```
> summary(schooldata)
 School.stage          State          District.Education.Office      Year
 Length:1756        Length:1756        Length:1756               Min.   :2017
 Class :character   Class :character   Class :character          1st Qu.:2017
 Mode  :character   Mode  :character   Mode  :character          Median :2018
                                                                 Mean   :2018
                                                                 3rd Qu.:2018
                                                                 Max.   :2018
 School.type            Sex            Number.of.pupils   Number.of.teachers colNames
 Length:1756        Length:1756        Length:1756        Length:1756        Mode:logical
 Class :character   Class :character   Class :character   Class :character   TRUE:1756
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Figure 1.1.1 shows summary of school data

```
> summary(schooldata)
         School.stage      State     District.Education.Office      Year                  School.type
 Primary school  : 580   Sarawak:360   Alor Gajah   :  12       Min.   :2017   Academic          :1168
 Secondary school:1176   Sabah  :288   Bachok       :  12       1st Qu.:2017   Vocational College: 588
                         Kedah  :144   Bagan Datuk  :  12       Median :2018
                         Johor  :140   Baling       :  12       Mean   :2018
                         Perak  :140   Bandar Baharu:  12       3rd Qu.:2018
                         Pahang :132   Baram        :  12       Max.   :2018
                         (Other):552   (Other)      :1684
     Sex       Number.of.pupils Number.of.teachers colNames
 Female:878   Min.   :   21    Min.   :   6.0     Mode:logical
 Male  :878   1st Qu.: 1396    1st Qu.: 129.0     TRUE:1756
              Median : 4194    Median : 372.5
              Mean   : 6796    Mean   : 583.7
              3rd Qu.: 8432    3rd Qu.: 678.5
              Max.   :64934    Max.   :6901.0
              NA's   :346      NA's   :342
```

Figure 1.1.2 shows summary of school data after data type changed

Based on figure 1.1.1, We can see that School.stage, State, Distinct.Education.Office, School.type, Sex, Number.of.pupils and Number.of.teachers are all in the character data type which is not suitable for analysis. School.stage, State, Distinct.Education.Office, School.type, and Sex are all changed to factor datatype while Number.of.pupils and Number.of.teachers are changed to numeric datatype which can be seen in figure 1.1.2.

1b) Removing Missing Values

```
> summary(schooldata)
         School.stage      State     District.Education.Office      Year                  School.type
 Primary school  :564   Sarawak :264   Alor Gajah   :  12       Min.   :2017   Academic          :1128
 Secondary school:846   Sabah   :220   Bachok       :  12       1st Qu.:2017   Vocational College: 282
                        Perak   :120   Bandar Baharu:  12       Median :2018
                        Johor   :112   Batang Padang:  12       Mean   :2018
                        Pahang  :112   Batu Pahat   :  12       3rd Qu.:2018
                        Selangor:108   Beaufort     :  12       Max.   :2018
                        (Other) :474   (Other)      :1338
     Sex       Number.of.pupils Number.of.teachers colNames
 Female:705   Min.   :   21    Min.   :  10.0     Mode:logical
 Male  :705   1st Qu.: 1396    1st Qu.: 129.0     TRUE:1410
              Median : 4194    Median : 373.5
              Mean   : 6796    Mean   : 585.4
              3rd Qu.: 8432    3rd Qu.: 684.2
              Max.   :64934    Max.   :6901.0
```

Figure 1.2 **shows summary of school data after missing values removed**

Based on figure 1.1.2 we can see that there are only two variables with missing values which are Number.of.pupils and Number.of.teachers.A total of 346 data of Number.of.pupils were missing and a total of 342 data of Number.of.teachers were missing. All the rows with these missing values were omitted which can be seen in Figure 1.2.

1c) Changing variable names

```
> summary(schooldata)
          school_stage      state              district        year                  school_type
 Primary school  :564   Sarawak :264   Alor Gajah   :  12   Min.   :2017   Academic          :1128
 Secondary school:846   Sabah   :220   Bachok       :  12   1st Qu.:2017   Vocational College: 282
                        Perak   :120   Bandar Baharu:  12   Median :2018
                        Johor   :112   Batang Padang:  12   Mean   :2018
                        Pahang  :112   Batu Pahat   :  12   3rd Qu.:2018
                        Selangor:108   Beaufort     :  12   Max.   :2018
                        (Other) :474   (Other)      :1338
   gender    number_of_pupils number_of_teachers colNames
 Female:705   Min.   :   21    Min.   :  10.0     Mode:logical
 Male  :705   1st Qu.: 1396    1st Qu.: 129.0     TRUE:1410
              Median : 4194    Median : 373.5
              Mean   : 6796    Mean   : 585.4
              3rd Qu.: 8432    3rd Qu.: 684.2
              Max.   :64934    Max.   :6901.0
```

Figure 1.3 shows the summary of the dataset after the names of the variables are changed

Based on Figure 1.2 shows the old column names while figure 1.3 shows the new column names.

Question 2

1) Primary school

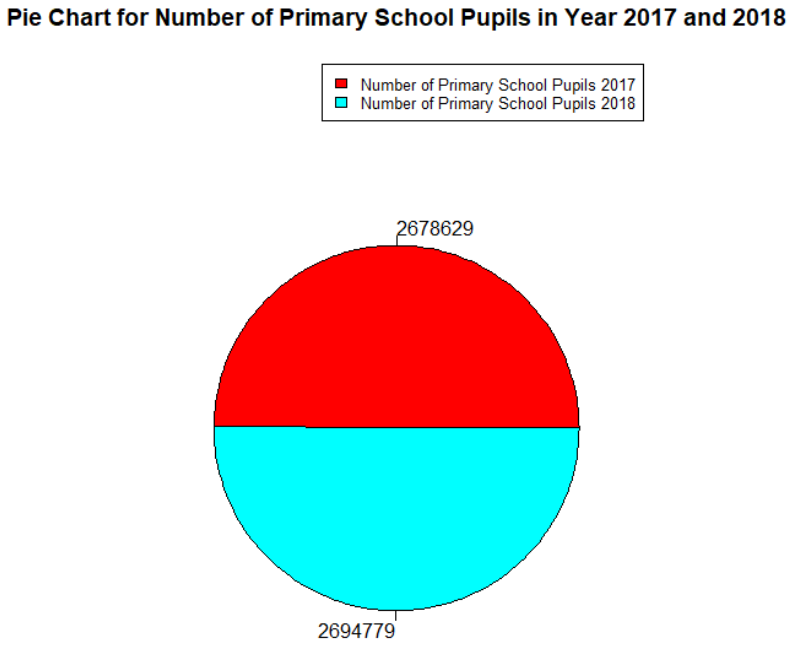

Figure 2.1.1 shows the pie chart for the number of primary pupils in 2017 and 2018

The number of primary school pupils in 2018 increased compared to the ones in 2017. This can be seen from figure 2.1.1 that the number of pupils in 2018 rose to 2694779 from 2678629 in 2017.

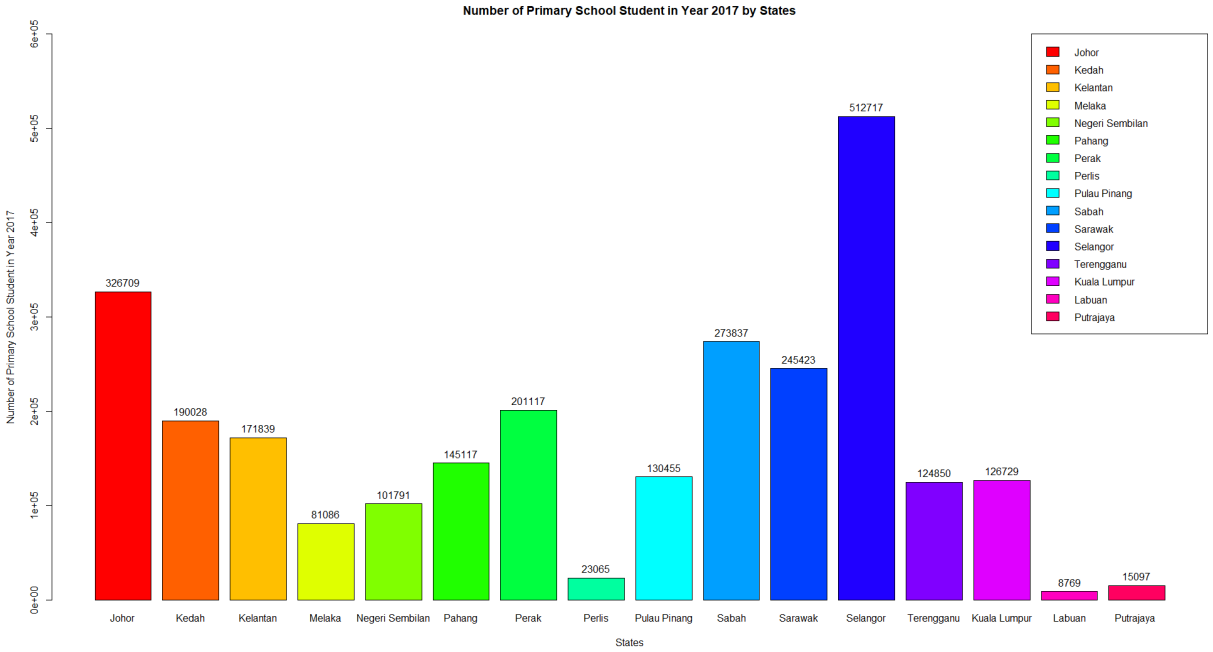

Figure 2.1.2 shows the bar chart of Number of Primary School Pupils in 2017 for each state.

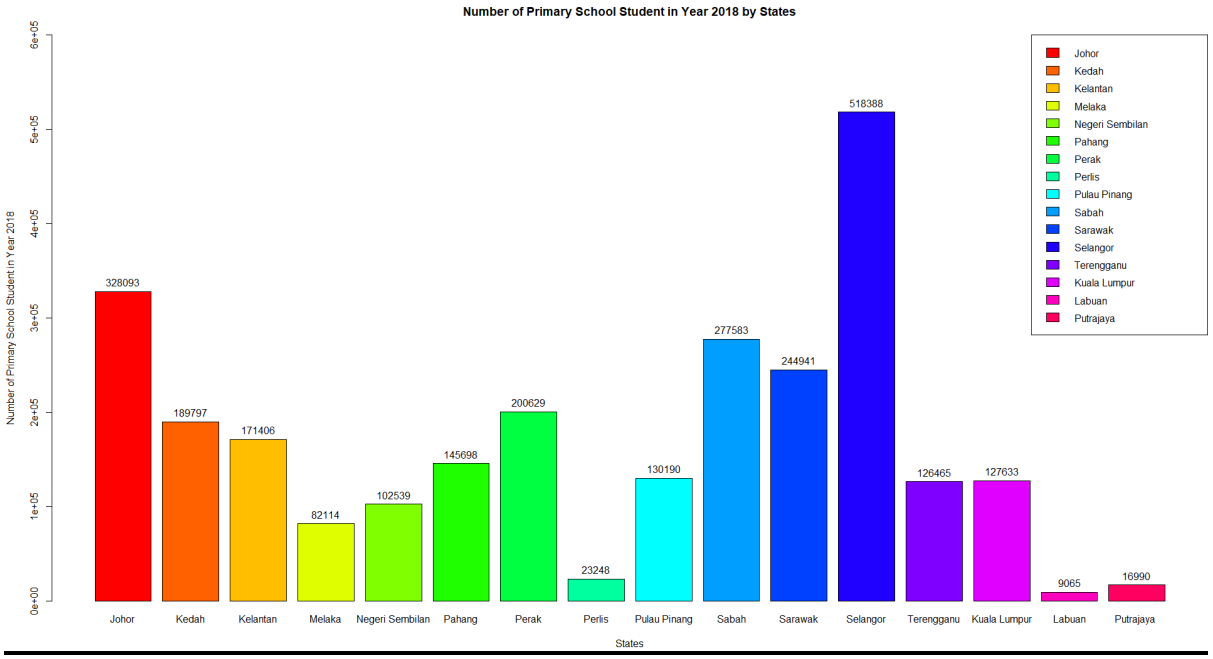

Figure 2.1.3 shows the bar chart of Number of Primary School Pupils in 2018 for each state

Based on figure 2.1.2 and figure 2.1.3, the top 3 states with the most number of primary school pupils remain the same for both years. However, the number of pupils for both years have fluctuated for these top 3 states.

2) Secondary School

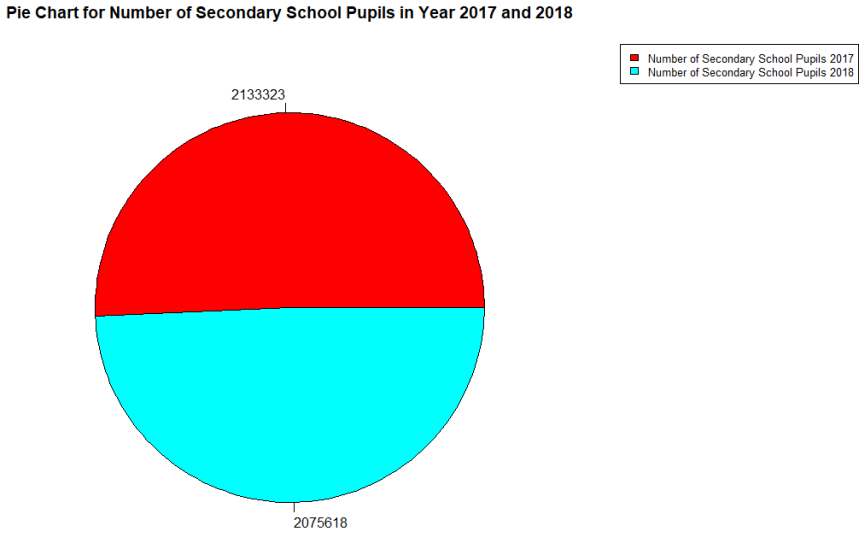

Figure 2.2.1 shows the pie chart for the number of secondary pupils in 2017 and 2018

The number of secondary school pupils in 2018 decreased compared to the ones in 2017. This can be seen from figure 2.1.1 that the number of pupils in 2018 drops to 2075618 from 2133323 in 2017.

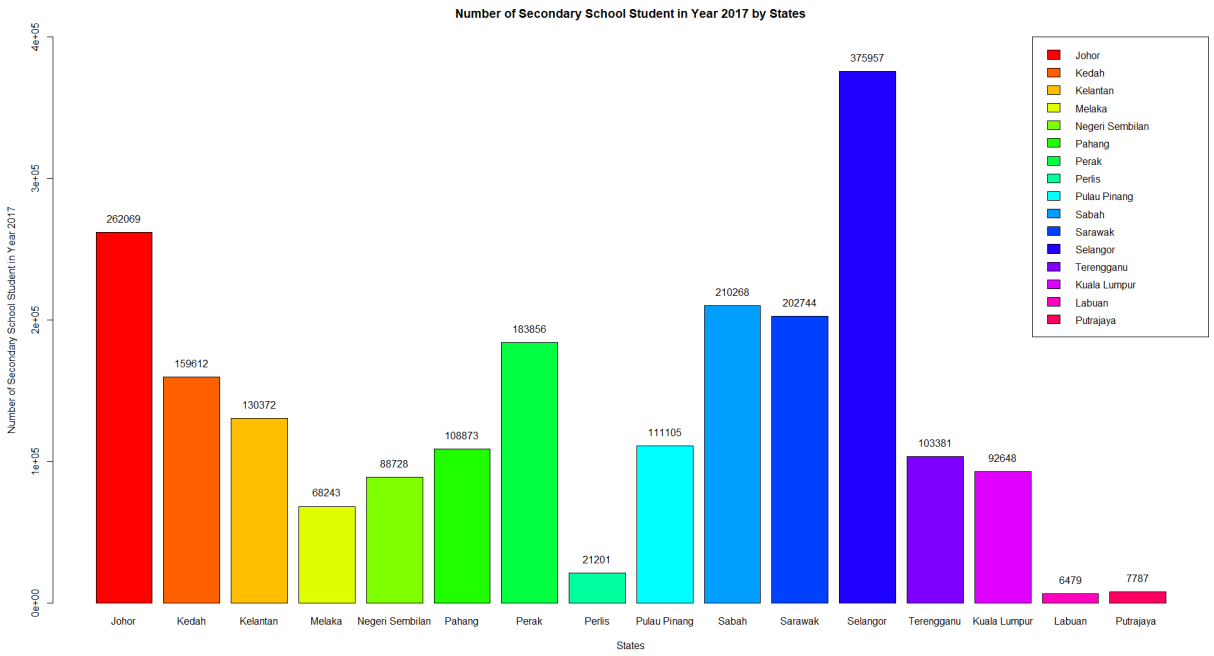

Figure 2.2.2 shows the bar chart of Number of Secondary School Pupils in 2017 for each state.

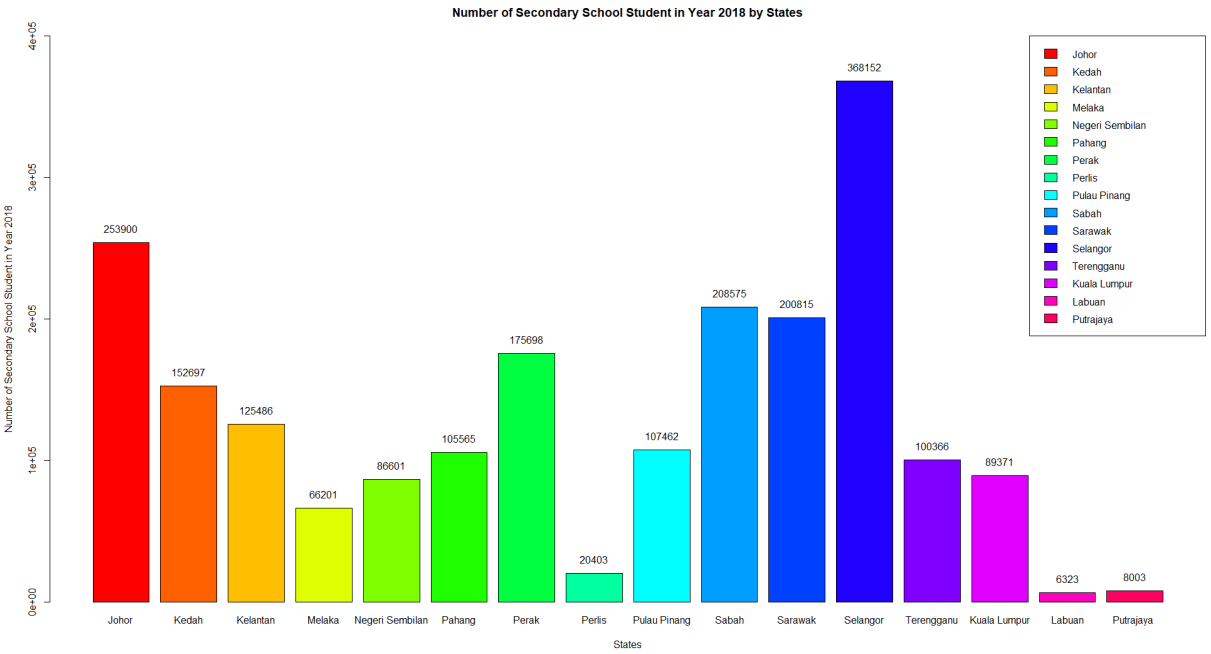

Figure 2.2.3 shows the bar chart of Number of Secondary School Pupils in 2018 for each state.

Based on figure 2.2.2 and figure 2.2.3, the top 3 states with the most number of secondary school pupils remain the same for both years. However, the number of pupils for 2018 have slightly dropped from 2017 for these top 3 states.

Question 3

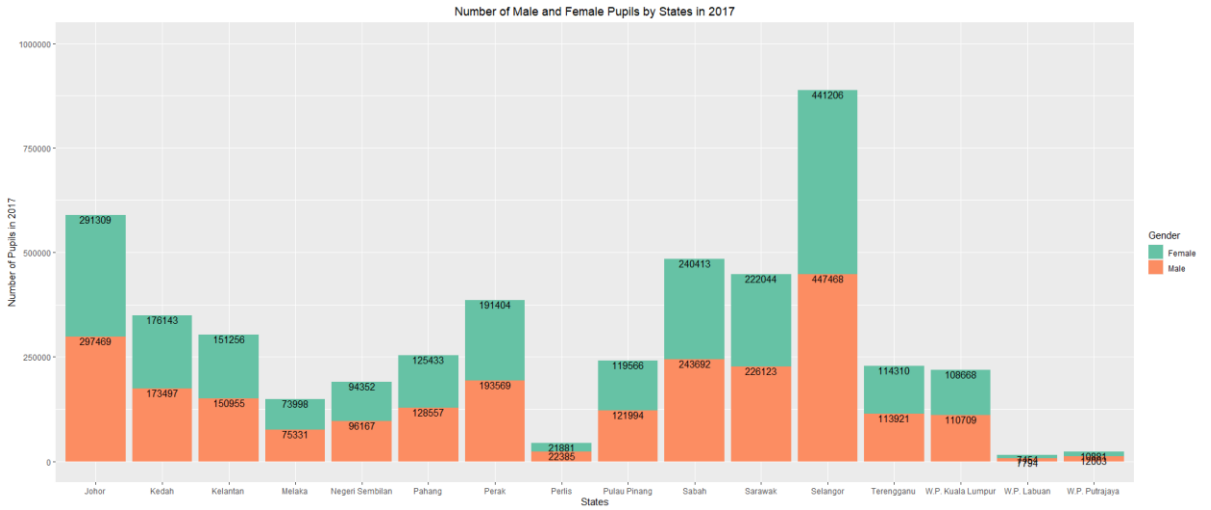

Figure 3.1 shows the bar plot of Number of Male and Female pupils in 2017

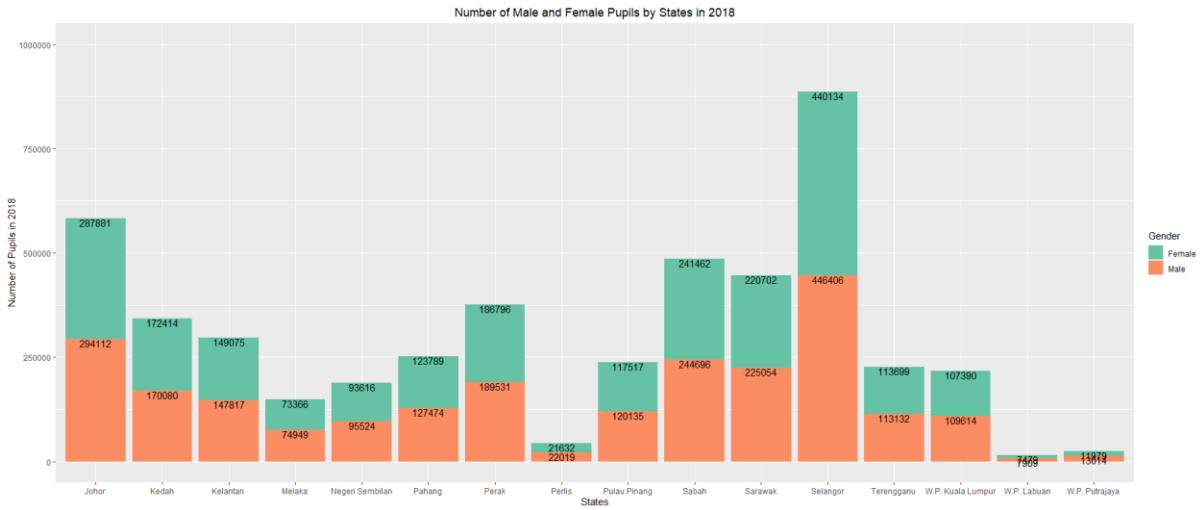

Figure 3.2 shows the bar plot of Number of Male and Female pupils in 2018

The states which have more female pupils than male pupils are Kedah, Kelantan and Terengganu for both years which can be seen in Figure 3.1 and Figure 3.2.

Question 4

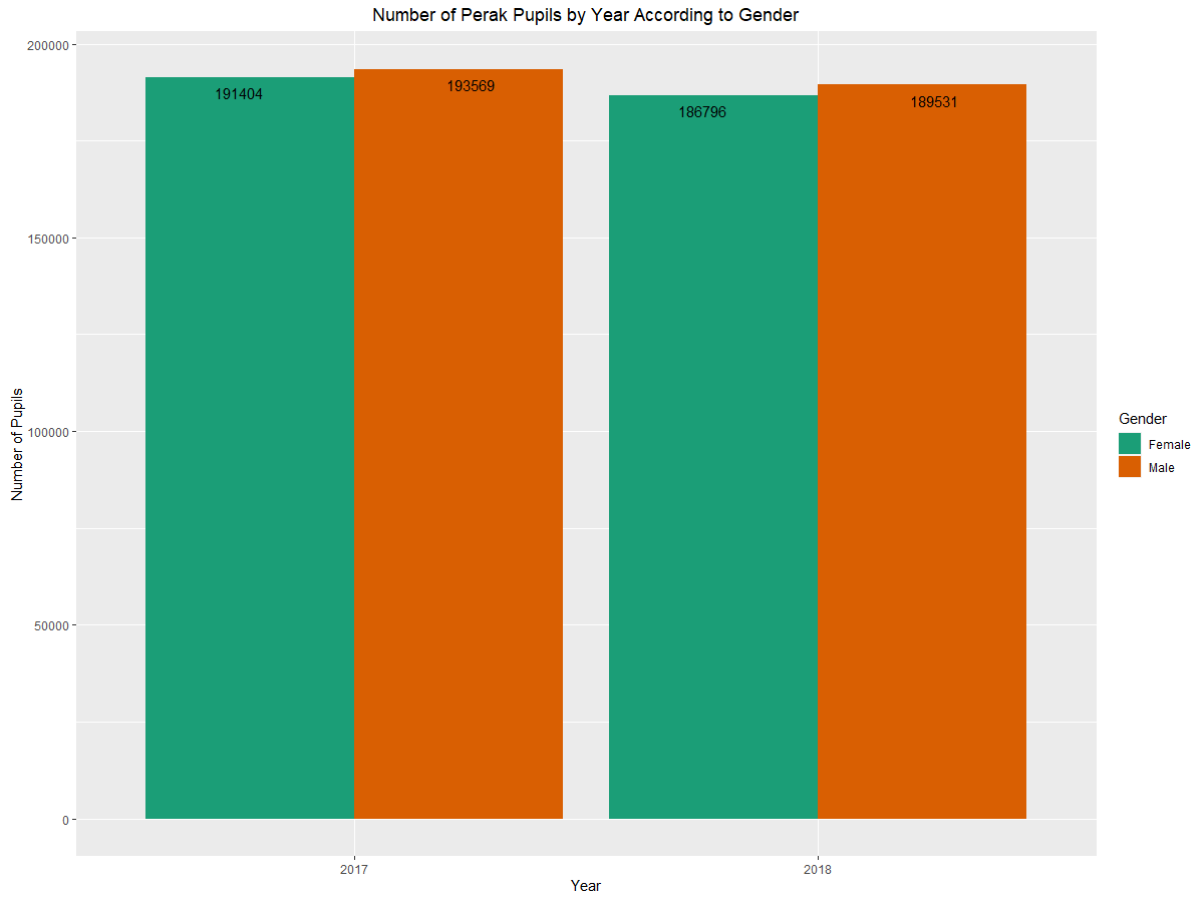

Figure 4 shows the number of pupils in Perak state by year 2017 and 2018 according to gender

Based on figure 4, the number of male pupils in both 2017 and 2018 is higher compared to female pupils. However, the number of pupils from both genders faces a slight drop from the year 2017 to 2018. Hence, there is no increase in the number of pupils for females in the year 2018 and faces a drop of 4,608 in the number of female pupils.

Question 5

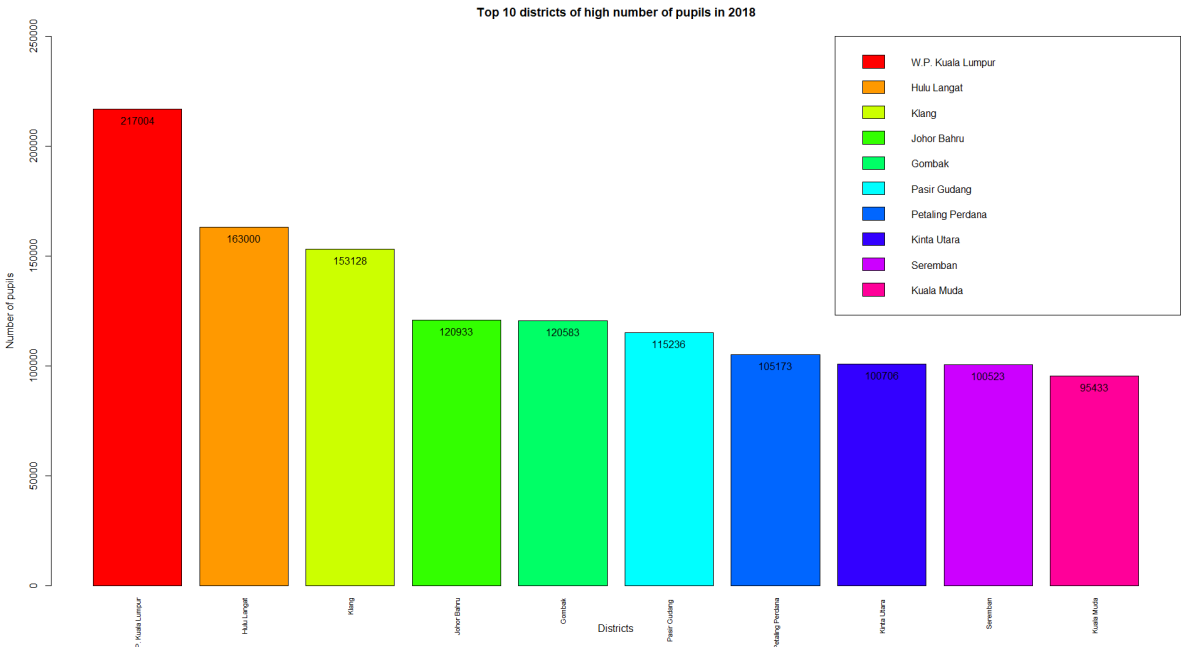

Figure 5 shows a bar chart graph that visualizes the top 10 districts that have the highest number of pupils in 2018

The most highest ranked district is W.P Kuala Lumpur with 217004 pupils while the last ranked district is Kuala Muda with 95433 pupils. The top three districts with the highest number of pupils are from the state of Selangor.

Question 6

a.) Display data type of each variable, convert to suitable data type and display summary

```
gender : character
age : numeric
hypertension : integer
heart_disease : integer
ever_married : character
work_type : character
Residence_type : character
avg_glucose_level : numeric
bmi : character
smoking_status : character
stroke : integer
```

Figure 6.0 the data type of each variable

Based on figure 6.0 it can be seen that the variables are not in the proper data type.

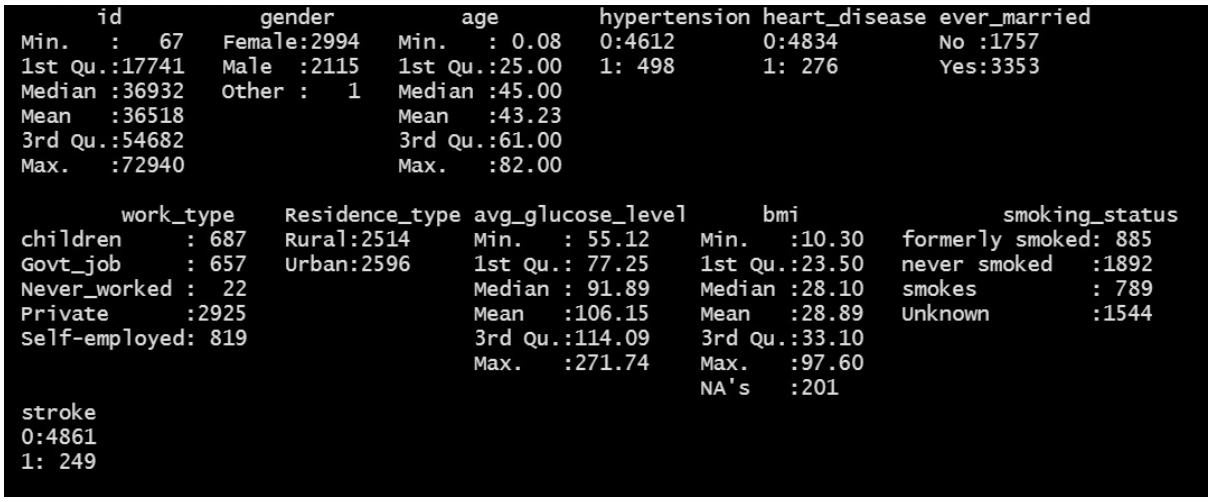

Figure 6.1 shows the summary of stroke data after changing the data types of variables

First we changed all the variables with character data type such as gender, ever_married, work_type, residence_type, and smoking_status as shown in **figure 6.0** to factors. Next variables such as hypertension, heart_disease, and stroke also convert into factors where they previously found in integers of 0s and 1s. The bmi data type was changed to numeric since it was found in characters previously which is not useful as we cannot plot the BMI values as characters as it will not produce correct outputs.

B.) Issues with variable gender and BMI

By referring to the figure 6.1 we can see that the gender has a category for other. Usually there are only two main gender categories available which are Male and Female. But here, we have something for others. The other category is not defined properly here in this context and it could be a invalid value. Since there is only one row with the gender value of the other, hence it can affect the visualization as it will introduce anomalies in the graphs. The second issue is that the BMI column has missing values. Hence leaving the missing values untreated will cause inconsistent results during analysis because missing values will be ignored by the R program.

c.) Proposed solution to overcome issues in (b)

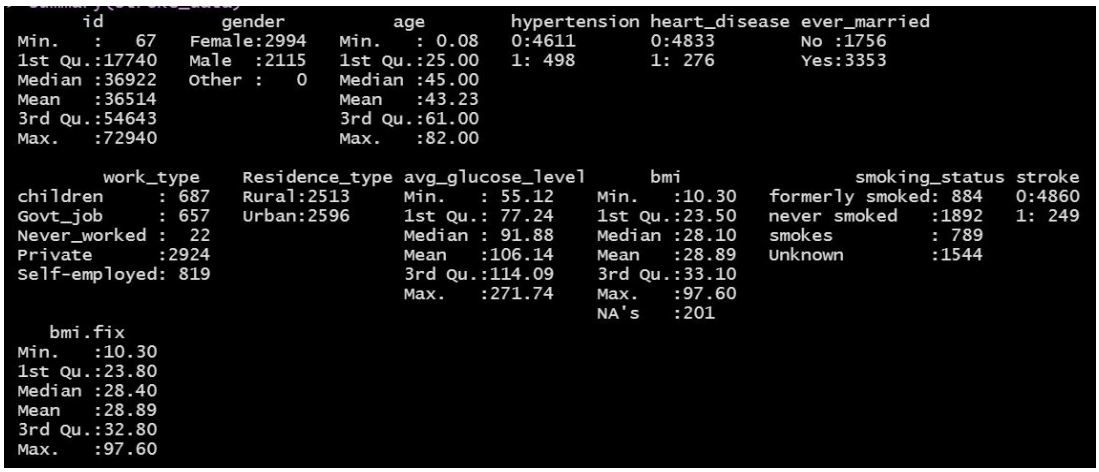

Figure 6.2 shows the summary of stroke_data after resolving the issues in 6(b)

For the issue with other value in gender, we have decided to remove it since there is only one patient with that value for the gender. It's about only 0.0002% of the overall data and it does not give a huge impact in the analysis. However, we have decided not to remove the missing values for BMI since around 200 rows are having these missing values and they could be trivial to this analysis. Hence we will replace the missing values with the mean of overall BMI.

Question 7

a.) Relationship between Age and Stroke

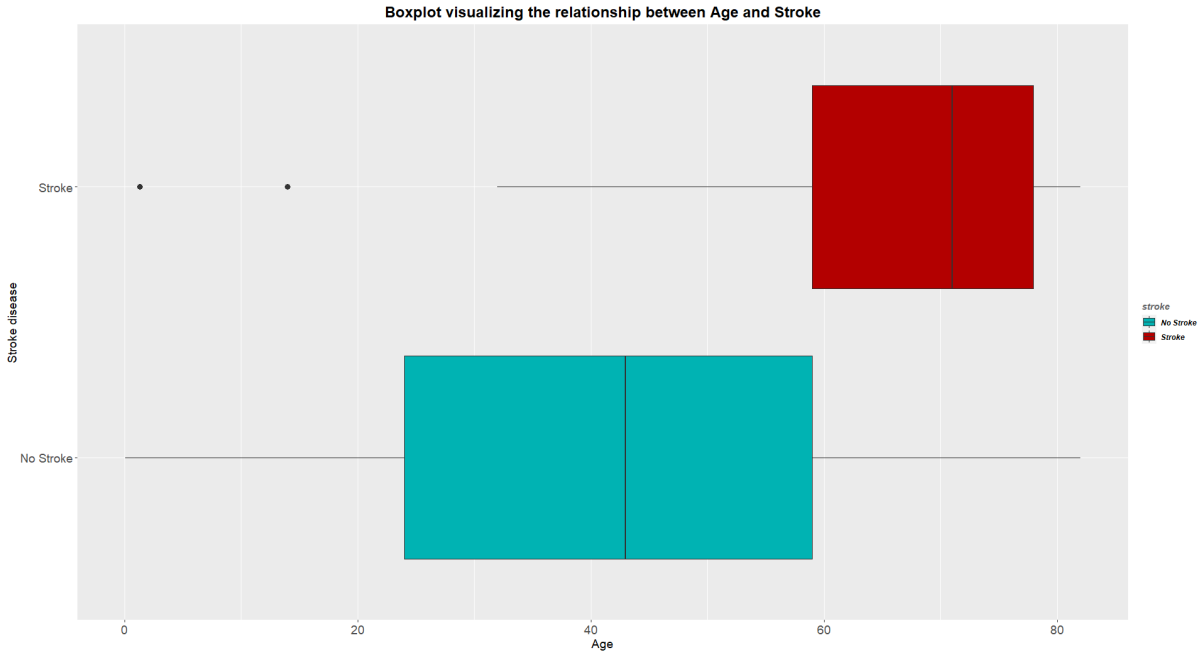

Figure 7.0 shows the boxplot visualizing the relationship between age and stroke

Based on the Figure 7.0 it is most unlikely that a patient below 30 years old will get a stroke. There are many patients between the early 20s and late 50s who do not get a stroke based on the interquartile range of patients with no stroke. However, a few patients started experiencing stroke from the late 50s up till late 70s based on the interquartile range of patients with stroke. This shows that age has positive correlation on stroke as people get older the chances for them to get stroke is very high.

b.) Relationship between BMI and Stroke

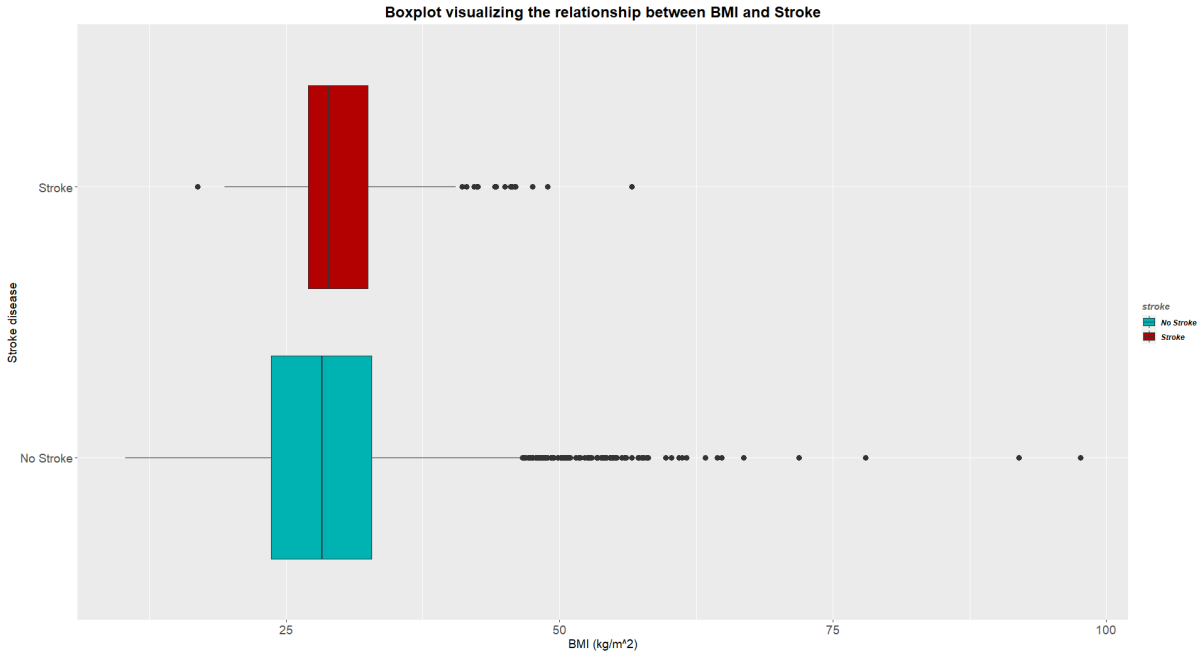

Figure 7.1 shows the boxplot visualizing the relationship between BMI and Stroke

Based on Figure 7.1 we can see that around 75% of patients who are not having strokes roughly have BMI less than 35. At the same time, around 75% of patients who had strokes roughly had BMI less than 35 as well. Moreover, the interquartile range of patients with stroke is smaller compared to the interquartile range of patients with no stroke. From this we can infer that BMI has no relation with stroke and it does directly affect any patients in getting a stroke.

Question 8

a.) Visualizing number of patients with heart disease based on gender

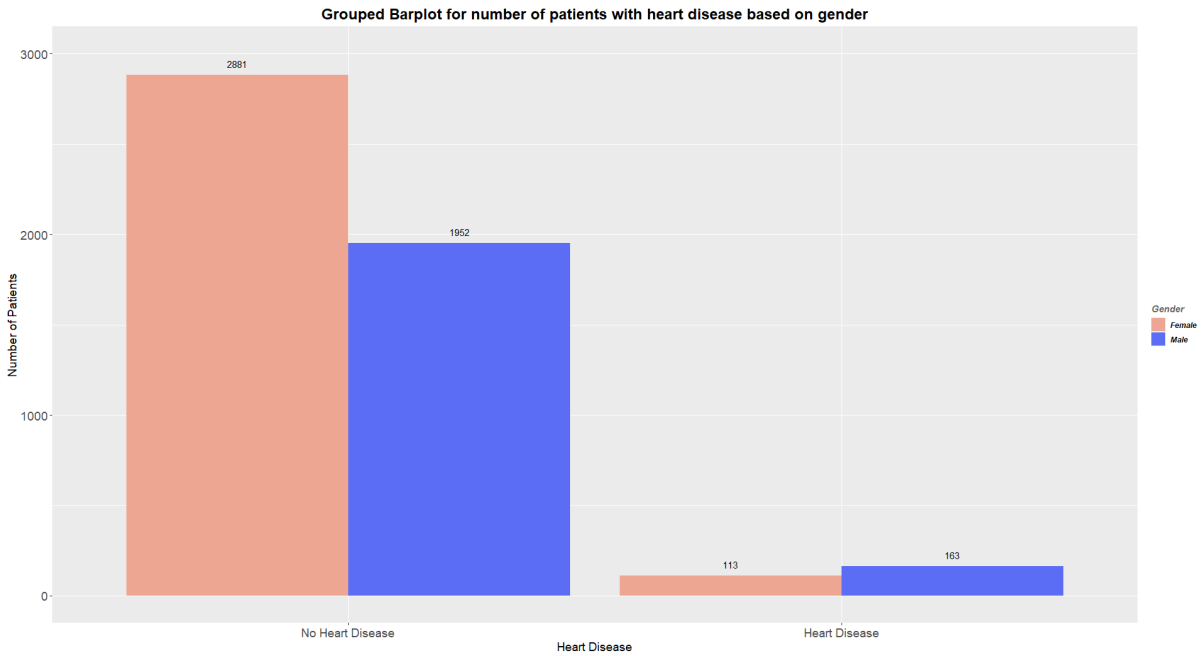

Figure 8.0 shows the grouped barplot for number of patients with heart disease based on gender

Based on figure 8.0 we can see that many male patients tend to get heart disease over the females. At the same time, there are more female patients who have no heart disease compared to males. This shows that the chance for female patients to get heart disease is very low.

b.) Visualizing number of patients with hypertension based on gender

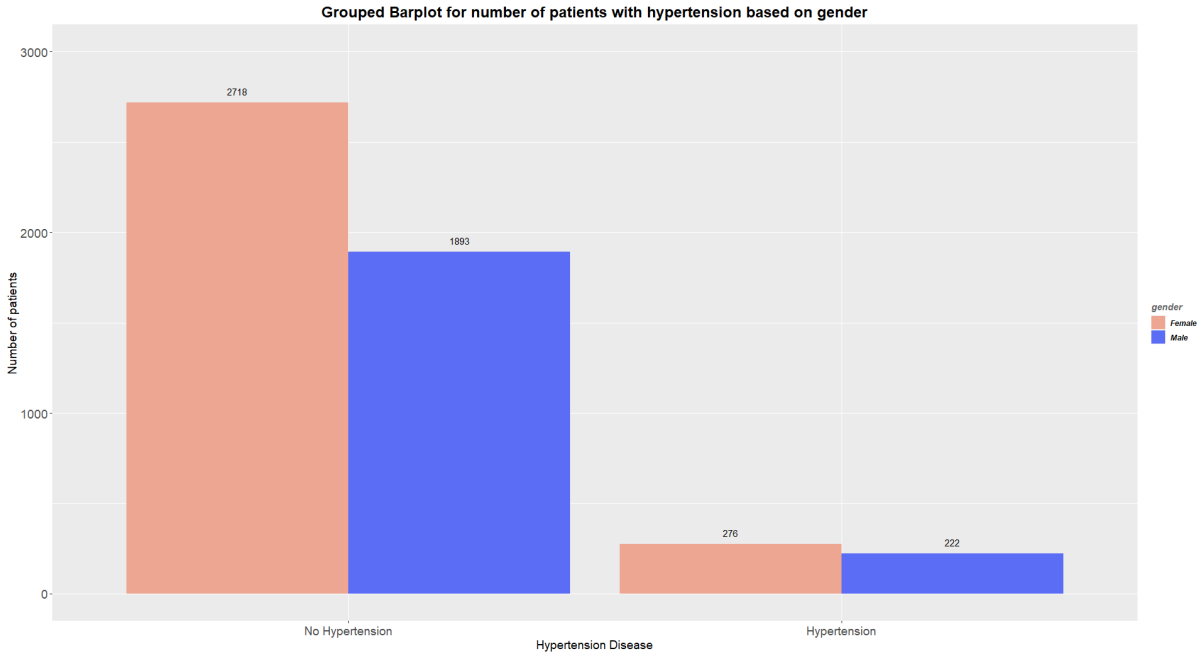

Figure 8.1 shows grouped barplot for number of patients with hypertension based on gender

Based on figure 8.1 we can see that females had to deal with a lot of pressure since there are many females having hypertension over males. However, there are also more female patients with no hypertension.

c.) Visualizing number of patients with stroke based on gender

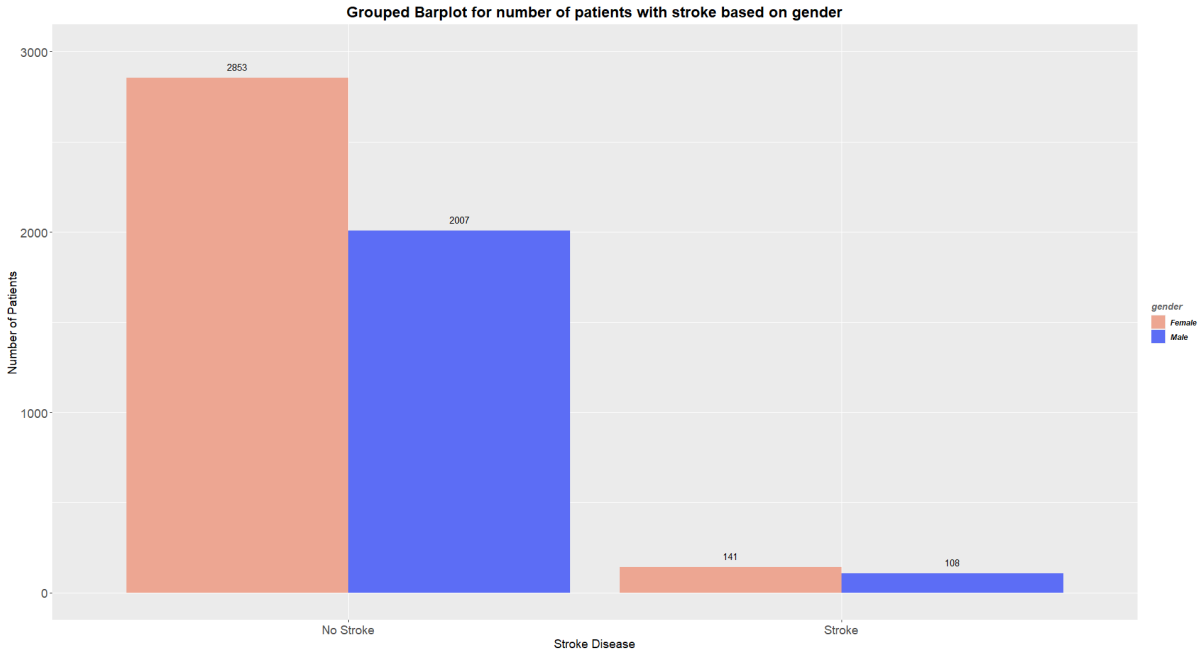

Figure 8.2 shows the grouped barplot for number of patients with stroke based on gender

Based on figure 8.2 we can see that many female patients do not get a stroke compared to males. However there is a small group of patients with stroke where the females are higher numbers.

Question 9

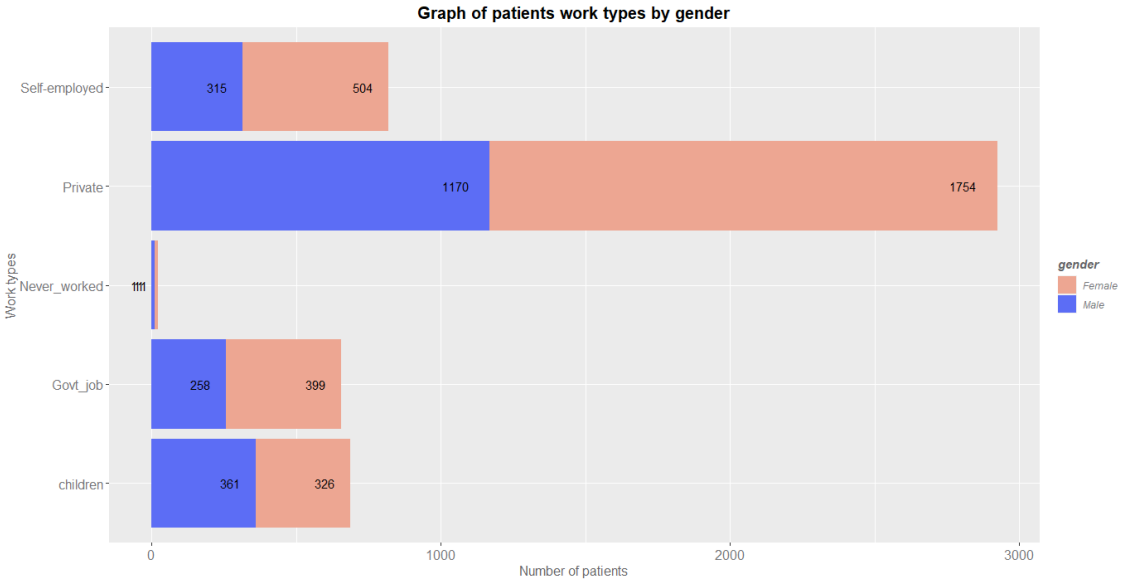

Figure 9 shows the graph of patients work types by gender

Based on figure 9, both female patients and male patients are present in all work types. Out of five work types, females are dominant in categories such as self-employed, private and government jobs. Apart from that, both female and male patients are equally present in the never worked category. However,

males were dominant only in the children category where the total number of males is higher than the number of females in this category.

Question 10

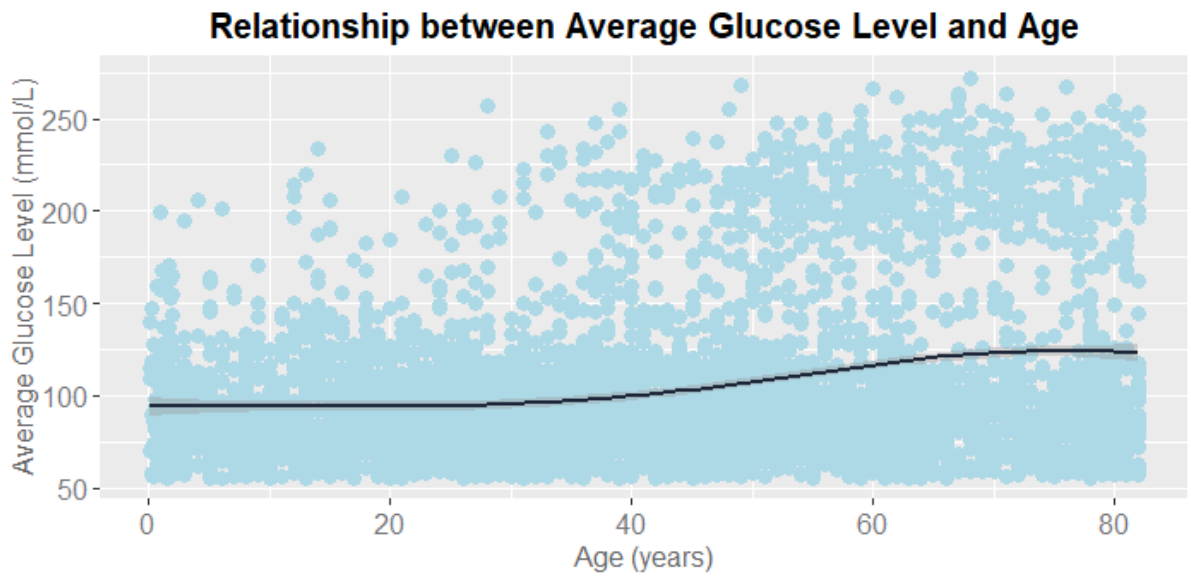

Figure 10

Based on figure 10, a scatter plot was plotted to analyse the relationship between average glucose level and age. We have plotted the scatter plot with the smooth curve since its difficult to discover the relationship patterns. The average glucose level maintains its level until the age of 40 and starts to rise up after that. The ribbon at the both ends of the smoothing curve is wider which shows the data is sparse at that certain area.