SEMESTER I, ACADEMIC SESSION 2021/2022

CPC351 - ASSIGNMENT 2

DATA EXPLORATION AND VISUALIZATION

Lecturer: Dr. Wong Li Pei

Name	Matric Number	Faculty
PRAVEEN NAIR A/L SANGARA NARAYANAN	149311	COMPUTER SCIENCE
DIVENESH A/L SHAMUGAM	146311	COMPUTER SCIENCE
OMSYARAN A/L CHANDRAN	148869	COMPUTER SCIENCE

Question 1a

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

The dataset is read, and the datatype of the variable is displayed. It can be noticed that all the variables listed from the dataset is classified in inaccurate data type. Firstly, there are only two types of school.stage which is Primary_school and Secondary_school so it is more likely to classify them into a categorical datatype which is factor. Next there are 16 states but initially the sate variable is classified as character, and it will be more suitable to make them into a category that will 16 different categories. The District.Education.office variable is initially in character but it can be examined that there are several entries of data for the same district. It is more suitable to store this variable as a categorical datatype. There is only two type of data entry for variable year and sex which is 2017, 2018 and male, female respectively so it shall be converted to categorical datatype. The Number.of.pupils and Number.of.teachers contains a continuous variable of numerical data, but it is initially stored in a datatype of character. Thus, this variable must changed to a numerical datatype which is more suitable to hold a wide range of numbers of continuous type of data.

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

The screenshot above shows the variables after they have been converted to a suitable data type.

Question 1b

```
>   # get the total number of missing values
>   summary(school_pupils_copy[is.na(school_pupils_copy$Number.of.pupils),
ber.of.teachers", "Number.of.pupils")])
 Number.of.teachers Number.of.pupils
 Min.   :6.00       Min.   : NA
 1st Qu.:6.00       1st Qu.: NA
 Median :6.50       Median : NA
 Mean   :6.75       Mean   :NaN
 3rd Qu.:7.25       3rd Qu.: NA
 Max.   :8.00       Max.   : NA
 NA's   :342        NA's   :346
>   # get the sum of the missing values
>   cat ("The sum of the missing value:" , sum(is.na(school_pupils_copy)))
The sum of the missing value: 688
```

```
> #Identify the number of rows before removing the miss
> cat("The number of rows before dropping missing value
The number of rows before dropping missing value: 1756
> # remove the missing  values
> school_pupils <- na.omit(school_pupils)
> #Identify the number of rows after removing the missi
> cat("The number of rows after dropping missing value:
The number of rows after dropping missing value: 1410
```

There are missing values in the variable ‘Number.of.teachers’ where 342 data is missing and for ‘Number.of.pupils’ 346 data is missing. The sum of missing values is 688. All the missing values are then removed from the dataset. To ensure the missing data is dropped the number of rows is counted before and after dropping the data.

Question 1c

Filter

	School.stage	State	District.Education.Office	Year	School.type	Sex	Number.of.pupils	Number.of.teachers
1	Primary school	Johor	Batu Pahat	2017	Academic	Female	18324	2530
2	Primary school	Johor	Batu Pahat	2017	Academic	Male	19592	1102

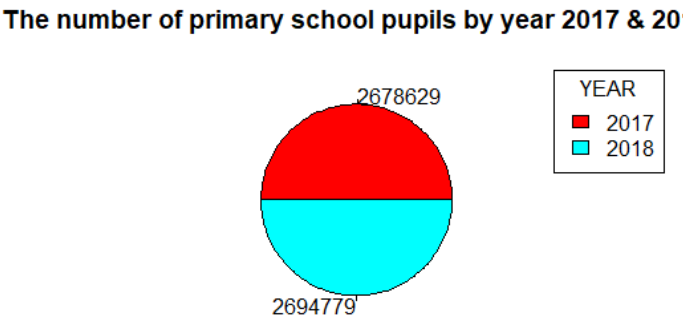
The diagram above depicts the initial name of each variable as how it is stored in the dataset. The variable names are renamed to “school_stage", "state", "district", "year", "school_type", "gender", "number_of_pupils", "number_of_teachers".

Filter

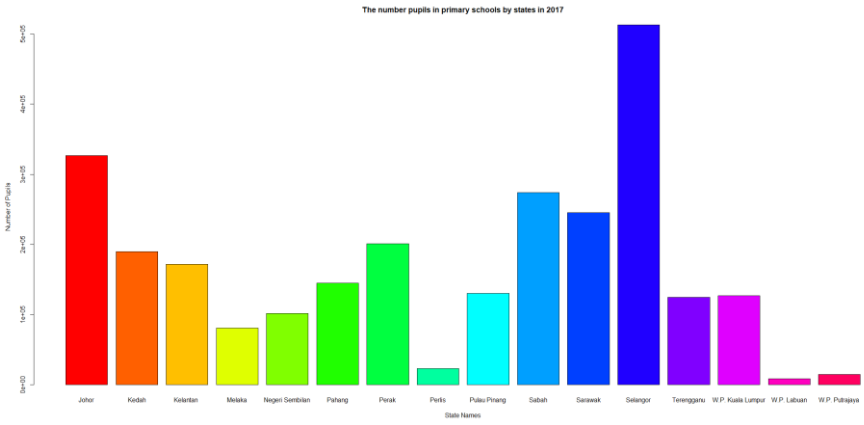
	school_stage	state	district	year	school_type	gender	number_of_pupils	number_of_teachers
1	Primary school	Johor	Batu Pahat	2017	Academic	Female	18324	2530
2	Primary school	Johor	Batu Pahat	2017	Academic	Male	19592	1102

The diagram above portrays the variables names after it has been changed.

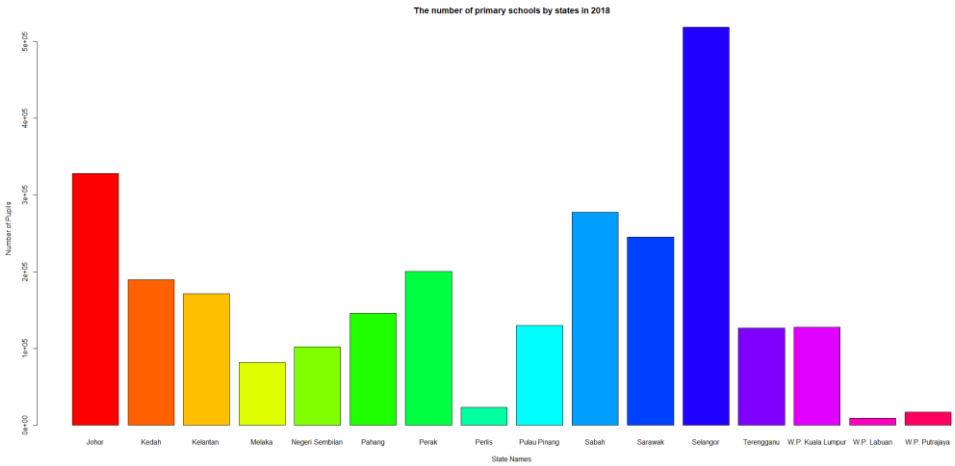
Question 2



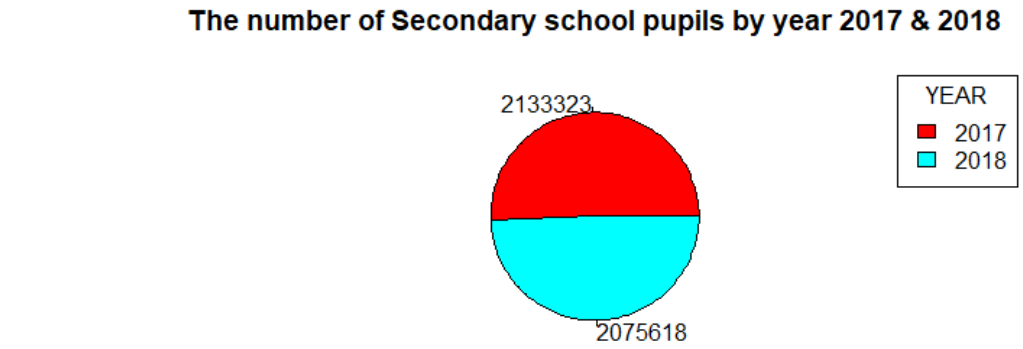
It can be portrayed that there are 2678629 primary school pupils in year 2017 while there are 2694779 primary school pupils in year 2018. There are more primary school pupils in 2018 than 2017 but the difference in percentage is minute to be noticed in the pie graph.



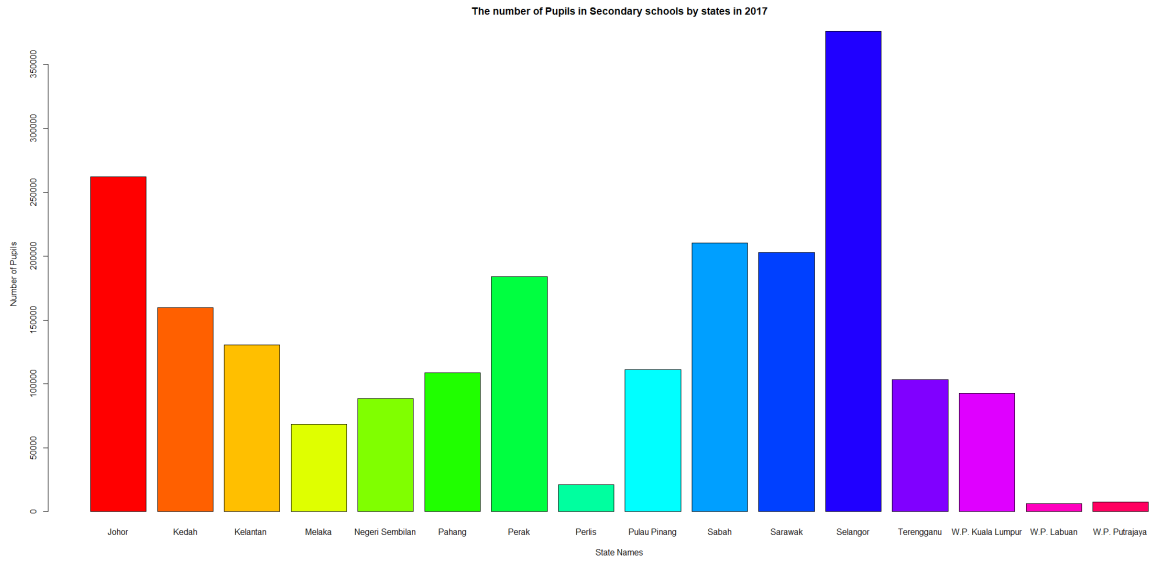
The graph above displays the number of people in primary school by states in year 2017.



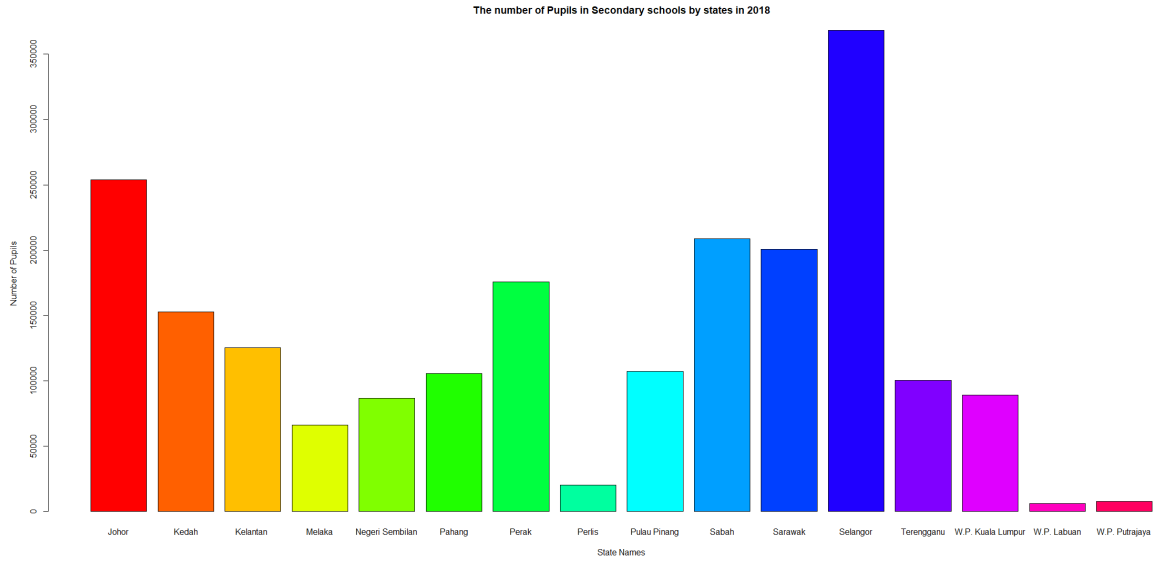
The graph above displays the number of people in primary school by states in year 2018.



It can be portrayed that there are 2133323 primary school pupils in year 2017 while there are 2075618 primary school pupils in year 2018. There are more primary school pupils in 2017 than 2018 so the red side of the pie has a larger area than the blue side of the pie.

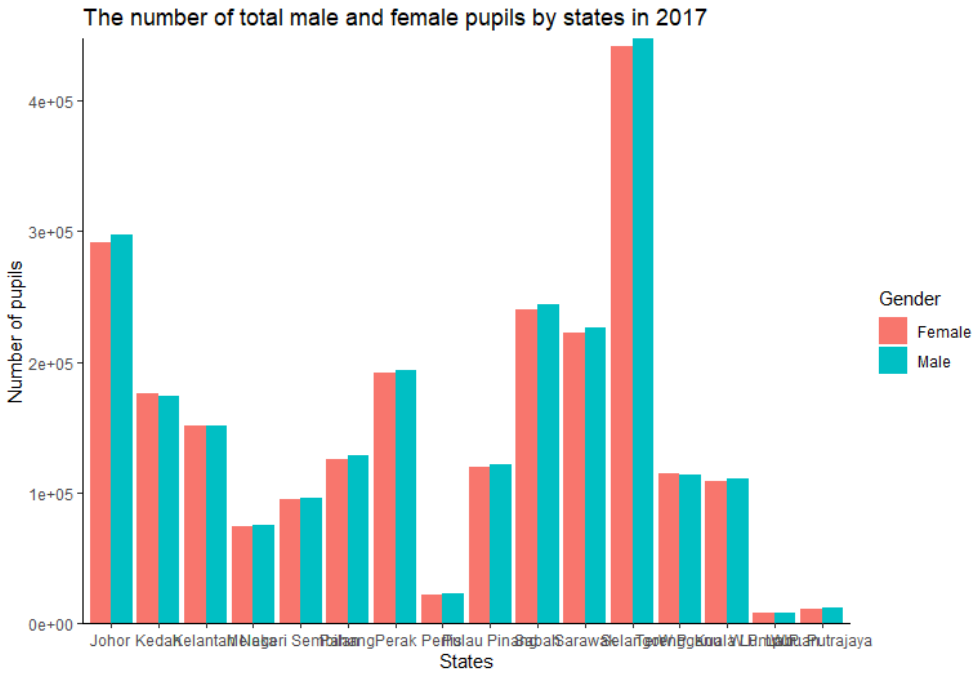


The graph above displays the number of people in secondary school by states in year 2017.

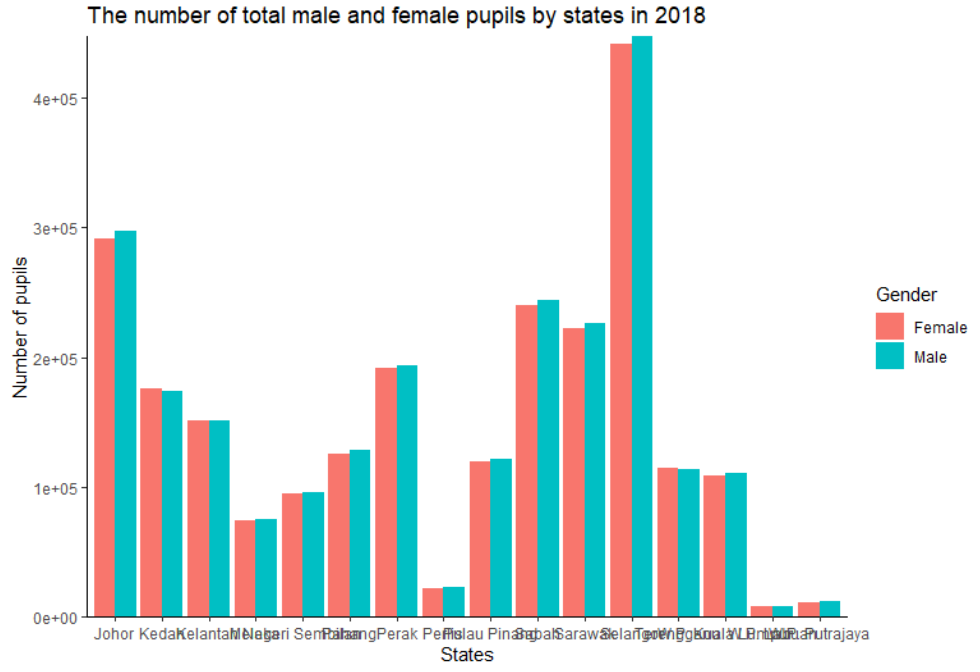


The graph above displays the number of people in secondary school by states in year 2018.

Question 3

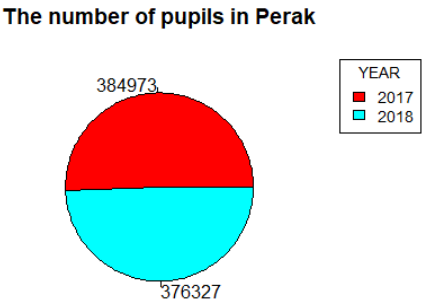


Based on the grouped bar chart above, Kedah, Kelantan and Terengganu are the only states that has higher number of female pupils compared to male pupils in 2017. Even though Selangor has the highest number of female pupils compared to other states, the male pupils in Selangor is much higher than that.



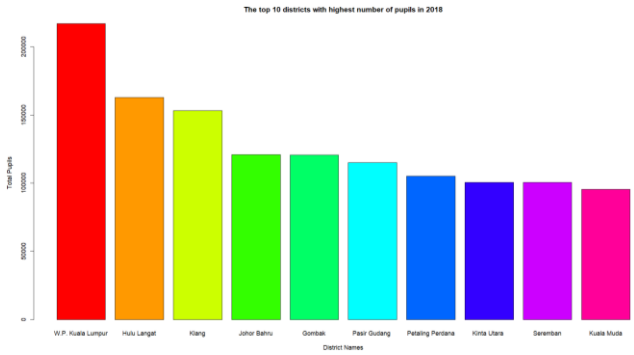
According to the grouped bar chart above, Kedah, Kelantan and Terengganu are the only states that has higher number of female pupils compared to male pupils in 2018. Yet again, even though Selangor has the highest number of female pupils compared to other states, the male pupils in Selangor is much higher when compared.

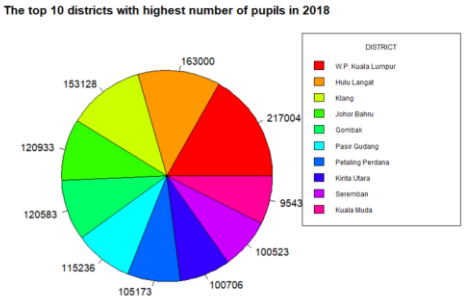
Question 4



The number of pupils in Perak did not experience any sort of drastic changes in 2-year time. There is no increase in number of female pupils from year 2017 to 2018 in Perak. There is a slight drop that can be seen in the bar chart above.

Question 5





According to the graph, the district W.P. Kuala Lumpur is recorded to be the district with the highest number of pupils. Followed by Hulu Langat, Klang, Johor Bahru, Gombak, Pasir Gudang, Petaling Perdana, Kinta Utara, Seremban, and Kuala Muda. Most of the districts are from Selangor since Selangor is one of the states that is dense in terms of population. The number of pupils were added up together regardless of the gender of the pupils to get the total number of pupils within a district. The aggregate function is used to achieve this.

Question 6a

```
> str(stroke_copy) # get the data types for each variables in the data frame
'data.frame':   5110 obs. of  11 variables:
 $ gender           : chr  "Male" "Female" "Male" "Female" ...
 $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
 $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
 $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
 $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
 $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
 $ avg_glucose_level: num  229 202 106 171 174 ...
 $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
 $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
> summary(stroke_copy)
   gender          age        hypertension heart_disease ever_married        work_type    Residence_type
Female:2994   Min.   : 0.00   0:4612       0:4834        No :1757     children     : 687   Rural:2514
Male  :2115   1st Qu.:25.00   1: 498       1: 276        Yes:3353     Govt_job     : 657   Urban:2596
Other :   1   Median :45.00                                           Never_worked :  22
              Mean   :43.22                                           Private      :2925
              3rd Qu.:61.00                                           Self-employed: 819
              Max.   :82.00

avg_glucose_level      bmi             smoking_status stroke
Min.   : 55.12    Min.   :10.30   formerly smoked: 885   0:4861
1st Qu.: 77.25    1st Qu.:23.50   never smoked   :1892   1: 249
Median : 91.89    Median :28.10   smokes         : 789
Mean   :106.15    Mean   :28.89   Unknown        :1544
3rd Qu.:114.09    3rd Qu.:33.10
Max.   :271.74    Max.   :97.60
                  NA's   :201
```

Firstly, a new data frame called ‘stroke_copy’ is created and the variable ‘id’ will be removed. Then, the str () function is used to get the data type of each variable in the ‘stroke_copy’ data frame. The data type for each variable are as such where gender , ever_married , work_type , Resindendce_type , bmi and smoking_status is in character. The variable age , avg_glucose_level is in numeric while hypertension , heart_disease and stroke are in integer. Later, the data type for the variables are changed by using the ‘as.factor()’,’as.numeric()’ and ‘as.integer()’ functions.

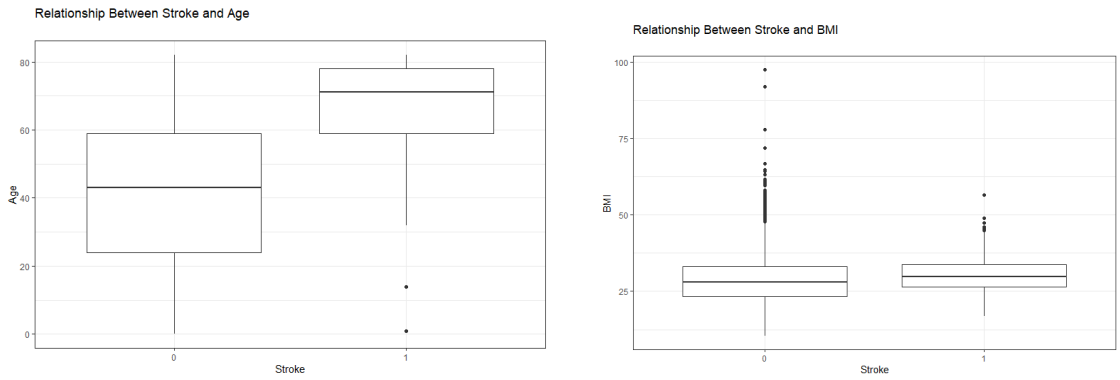
Question 6b

The variables gender and BMI can affect the analysis process. For BMI, we can notice that there are 201 null values that exists within this variable. These null values which are labelled as NA in the dataset can affect the analysis since it indicates that the dataset is incomplete. In certain cases, the NA might indicate some sort of meaning which depends on the context of the variable. But since this is BMI, having a NA does not convey any other meaning except for null. Hence, any sort of conclusion or hypothesis that is being made at end will be inaccurate and faulty due to this missing values. The relationship between BMI and stroke cannot be established correctly. For variable gender, it was assumed to only have 2 types of genders for this dataset which is ‘male’ and ‘female’. However, another input known as ‘other’ can be identified in this dataset during the pre-processing step. Hence, we can say that there is presence of noisy data in this dataset which can also be a problem during the visualization process.

Question 6c

However, the problems can be solved. For the variable BMI, it can be solved by removing the instances of the dataset that has null values for BMI. This step can be done when the number of null values in BMI is relatively less compared to the overall dataset. Another method would be by just replacing the null values in the BMI with the mean (average) value of the BMI variable. This step is taken when the number of null values is large, and the removal of all those null values might cause a significant decrease in the overall dataset. For the variable gender, we can simply just remove or delete the noisy data from the dataset.

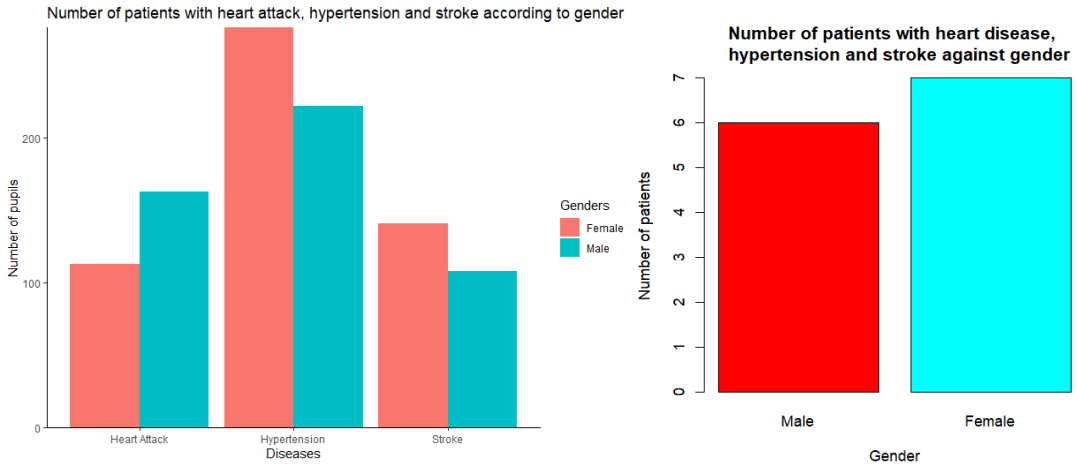
Question 7



The diagram above portrays two set of box plot where the y variables are the factors that are being tested which is age and BMI while the x variable is stroke where 0 indicate no stroke and 1 indicates stroke. Based on the boxplot diagram on the left, it can be said that people with age in between 60 to 80 are much likely to get stroke. Few outliers are found where people age less than 30 are getting stroke. People with the age in between 20 and 60 are not likely to get stroke.

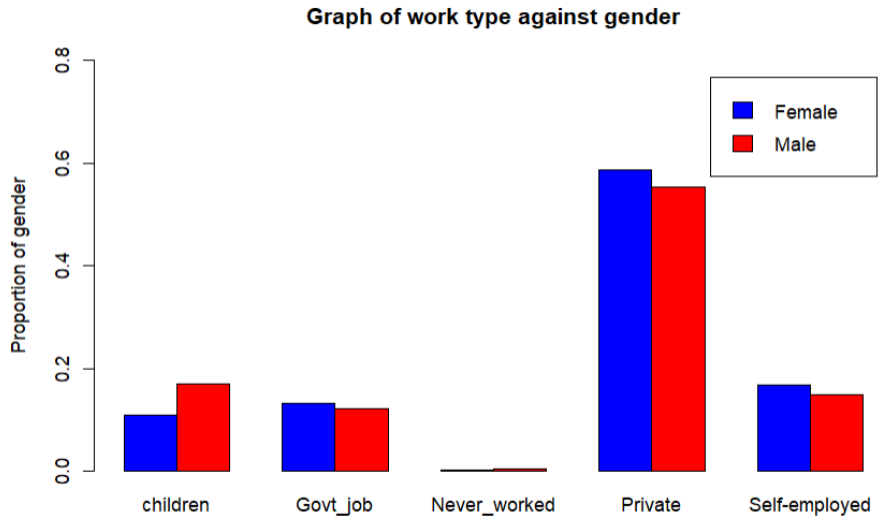
No outliers are spotted. According to the boxplot picture on the right, the relationship between BMI and stroke is imprecise since the range of BMI of people who got stroke and didn't is quite similar. However, there are many outliers where it does indicate that higher BMI value does not mean that the person will get stroke. However, the likeliness of getting stroke and not getting stroke is much similar, but there are chances of getting stroke when the BMI value is higher.

Question 8



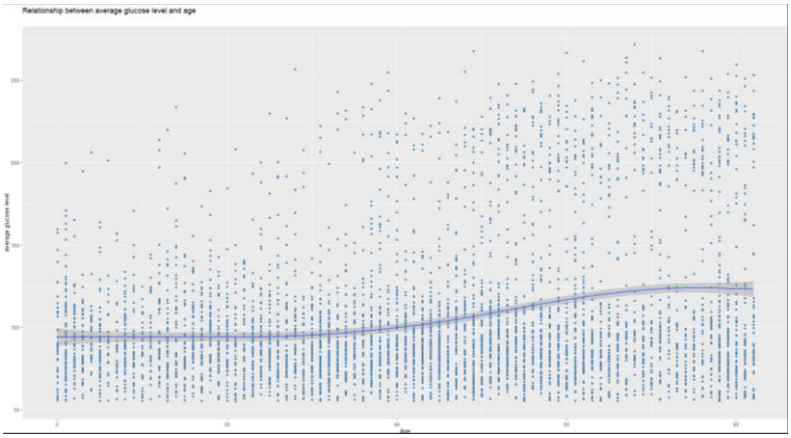
For the first graph, we can say that there is a lot of pupils that has hypertension compared to heart attack and stroke. The females are recorded to be the highest with the value of 276 and male with 222. Heart attack is much more prominent on males compared to females where more males are recorded for having heart attack (163) compared to females (113). But it is the exact opposite for stroke where stroke is much more prominent in females (141) compared to males (108). Hence, we can say that more females are getting diseases compared to males with hypertension and stroke being the dominant disease for them according to this graph. For the second graph, it shows the total number of patients that has all 3 diseases, and they are grouped according to their gender. Yet again, the number of females outnumbers the number of males by 1.

Question 9



Based on the grouped bar chart above, the work type that has more males than females will be the children related types. The same applies for the never worked type. Hence, there is 2 type which is children and never worked type that has more males than females. Other jobs such as government job, private jobs and self-employed type has more female than males.

Question 10



The diagram above portrays a scatter plot with a smoothing curve for average glucose level against age. In general, plotted based on the visualisation above the average glucose level increases as the age increases from the age 30 to 70. The average glucose level remains constant after the age 70 and before the age 30. The data is denser in the middle of the graph and slightly less dense at the ends of the graph which stresses the fact there is more data concerning in the middle of the age range.

References

1) Data visualization with R - github pages. (n.d.). Retrieved January 9, 2022, from https://rkabacoff.github.io/datavis/Multivariate.html

2) Erik Marsja, & *, N. (2021, August 18). How to rename column (or columns) in R with dplyr. Erik Marsja. Retrieved January 9, 2022, from https://www.marsja.se/how-to-rename-column-or-columns-in-r-with-dplyr/#:~:text=To%20rename%20a%20column%20in,A

3) Data cleanup: Remove NA rows in R. ProgrammingR. (n.d.). Retrieved January 9, 2022, from https://www.programmingr.com/examples/remove-na-rows-in-r/