ICT513 - Assignment 2

Chew Jian Yue

Contents

1 Introduction

Import the libraries into the working environment.

Import both datasets for the assignment into the R working environment.

```
setwd("A:\\OneDrive\\OneDrive - Kaplan\\1. Murdoch
↪  (2022-2023)\\!Sem3\\ICT513 Data Analytics\\!Assignment 2")

mental <- read.csv("Mental_health_hospitalisations.csv", header = TRUE)
milkp <- read_xlsx("MilkProduction.xlsx")  # No need to specify header = T?
```

2 Question 1

Based on the contextual information given by the question, hospitalisation rate (`hospitalisation.rate`) is identified as the response variable.

Using the `glimpse(..)` function from the tidyverse library, we can attempt to understand the variables in the dataset.

```
glimpse(mental)
```

```
Rows: 56
Columns: 4
$ year                 <int> 2006, 2006, 2006, 2006, 2007, 2007, 2007, 2007, 2~
$ age.group            <chr> "12 to 24", "12 to 24", "25+", "25+", "12 to 24",~
$ sex                  <chr> "Male", "Female", "Male", "Female", "Male", "Fema~
$ hospitalisation.rate <dbl> 45.8, 54.0, 59.0, 59.6, 43.7, 52.6, 57.4, 61.1, 4~
```

Names of variables contained in the dataframe:

```r
# Names of variables contained in the dataframe
names(mental)
```

```
[1] "year"                    "age.group"               "sex"
[4] "hospitalisation.rate"
```

Check if there are any missing values in the dataset. From the output, there are no missing values in the dataset.

```r
sum(is.na(mental))
```

```
[1] 0
```

Understanding the variables in the dataset, `age.group` and `sex` are most likely categorical variables and their datatype should be converted into `factor`.

```r
unique(mental$age.group)
```

```
[1] "12 to 24" "25+"
```

```r
unique(mental$sex)
```

```
[1] "Male"   "Female"
```

Converting or casting their datatype into `factor` to ensure R treats them as such.

```r
mental$age.group <- factor(mental$age.group)
mental$sex <- factor(mental$sex)
mental$year <- factor(mental$year)
```

From the code below, I understand the range of years of the study are indeed from 2006 to 2019, as described in the question. More interestingly, the data values for each of the years are equal, at 4.

```r
unique(mental$year)
```

```
 [1] 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
14 Levels: 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 ... 2019
```

```r
summary(mental$year)
```

```
2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 
   4    4    4    4    4    4    4    4    4    4    4    4    4    4 
```

2.1 Part (a) - Exploratory data analysis (EDA)

2.1.1 Descriptive statistics

```r
summary(mental)
```

```
      year          age.group       sex      hospitalisation.rate
 2006   : 4   12 to 24:28   Female:28   Min.   : 41.10      
 2007   : 4   25+     :28   Male  :28   1st Qu.: 55.95      
 2008   : 4                             Median : 60.45      
 2009   : 4                             Mean   : 62.24      
 2010   : 4                             3rd Qu.: 65.72      
 2011   : 4                             Max.   :101.20      
 (Other):32                                                 
```

The mean of hospitalisation rate is 62.24, median of hospitalisation rate is 60.45 for all sub-groups in the dataframe.

From the later analyses (e.g., histogram), the distribution of hospitalisation rate is not normal and is positively skewed. The measure of centrality, median is 60.45 and the measure of dispersion, interquartile range is 9.775.

```r
IQR(mental$hospitalisation.rate)
```

```
[1] 9.775
```

The size of the dataset is 56 rows of values, which is small. The dataset is balanced as there are equal number of values for each of the years, each of the age.group and each of the sex factor levels.

Here, we understand the data by their sub-groups.

```r
mental %>%
    group_by(sex) %>%
    summarise(`Mean hospitalisation rate` = mean(hospitalisation.rate),
        `S.D. hospitalisation rate` = sd(hospitalisation.rate))
```

sex	Mean hospitalisation rate	S.D. hospitalisation rate
Female	68.63214	13.686626
Male	55.85714	7.532961

The mean hospitalisation rate is generally higher for females than for males.

```
mental %>%
    group_by(sex) %>%
    summarise(`Median hospitalisation rate` = median(hospitalisation.rate),
        `IQR hospitalisation rate` = IQR(hospitalisation.rate))
```

sex	Median hospitalisation rate	IQR hospitalisation rate
Female	64.55	14.025
Male	57.30	12.425

```
mental %>%
    group_by(age.group) %>%
    summarise(`Mean hospitalisation rate` = mean(hospitalisation.rate),
        `S.D. hospitalisation rate` = sd(hospitalisation.rate))
```

age.group	Mean hospitalisation rate	S.D. hospitalisation rate
12 to 24	62.84286	17.698366
25+	61.64643	3.836326

The mean hospitalisation rate is generally slightly higher for age group "12 to 24" than for age group "25+". The spread of hospitalisation rate is higher for age group "12 to 24" than for age group "25+".

```
mental %>%
    group_by(year) %>%
    summarise(`Mean hospitalisation rate` = mean(hospitalisation.rate),
        `S.D. hospitalisation rate` = sd(hospitalisation.rate),
        `Range of hospitalisation rate` = max(hospitalisation.rate) -
            min(hospitalisation.rate))
```

year	Mean hospitalisation rate	S.D. hospitalisation rate	Range of hospitalisation rate
2006	54.600	6.381222	13.8
2007	53.700	7.520195	17.4
2008	53.350	7.602412	17.8
2009	53.325	8.539858	19.9
2010	55.025	6.462907	14.0
2011	57.600	7.700649	18.3
2012	60.550	10.527583	25.2
2013	60.800	10.392626	25.1
2014	64.100	11.083622	26.3
2015	68.625	11.709932	27.2
2016	71.225	13.978883	32.0
2017	72.025	14.916294	32.2
2018	73.075	16.679204	36.4
2019	73.425	18.588773	39.3

From the descriptive statistics, the mean hospitalisation rate across the years generally increased and the spread using standard deviation over the years generally increased. We can see that the spread similar increases (range of values) over the years. `year` could be considered as an index variable, but in this case, it will be included in the model for prediction. Given the year, we may be able to predict the hospitalisation rate.

```
table(mental$year)/sum(table(mental$year))
```

```
      2006       2007       2008       2009       2010       2011       2012 
0.07142857 0.07142857 0.07142857 0.07142857 0.07142857 0.07142857 0.07142857 
      2013       2014       2015       2016       2017       2018       2019 
0.07142857 0.07142857 0.07142857 0.07142857 0.07142857 0.07142857 0.07142857 
```

```
table(mental$age.group)/sum(table(mental$age.group))
```

```
12 to 24      25+ 
     0.5      0.5 
```

```
table(mental$sex)/sum(table(mental$sex))
```

```
Female   Male 
   0.5    0.5 
```

The above are the frequency tables of each of the categorical variables in the dataset. Their relative proportions are equal, this means that the frequency of data for each factor level is the same.

I will apply log transformation to the variables that are not normally distributed using `log(...)` function in R.

2.1.2 Simple visualisations

```
hist(mental$hospitalisation.rate, xlab = "Hospitalisation Rate",
    main = "Histogram of Hospitalisation Rate")
```

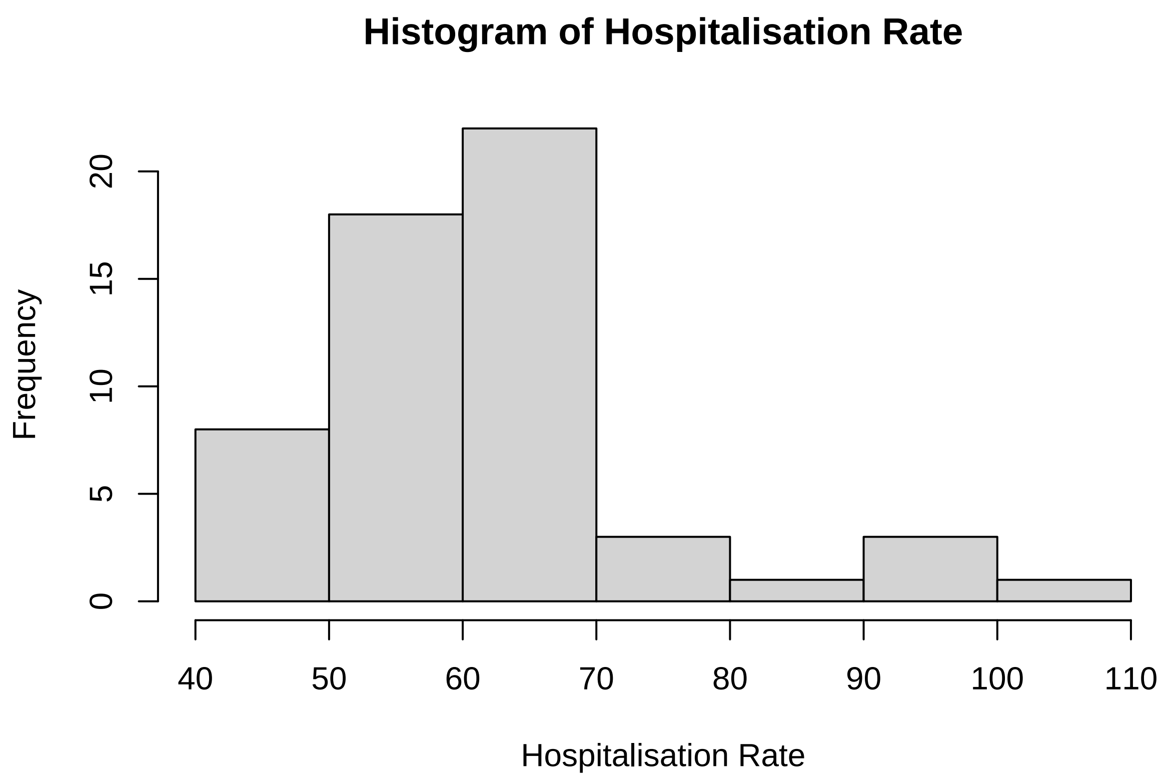
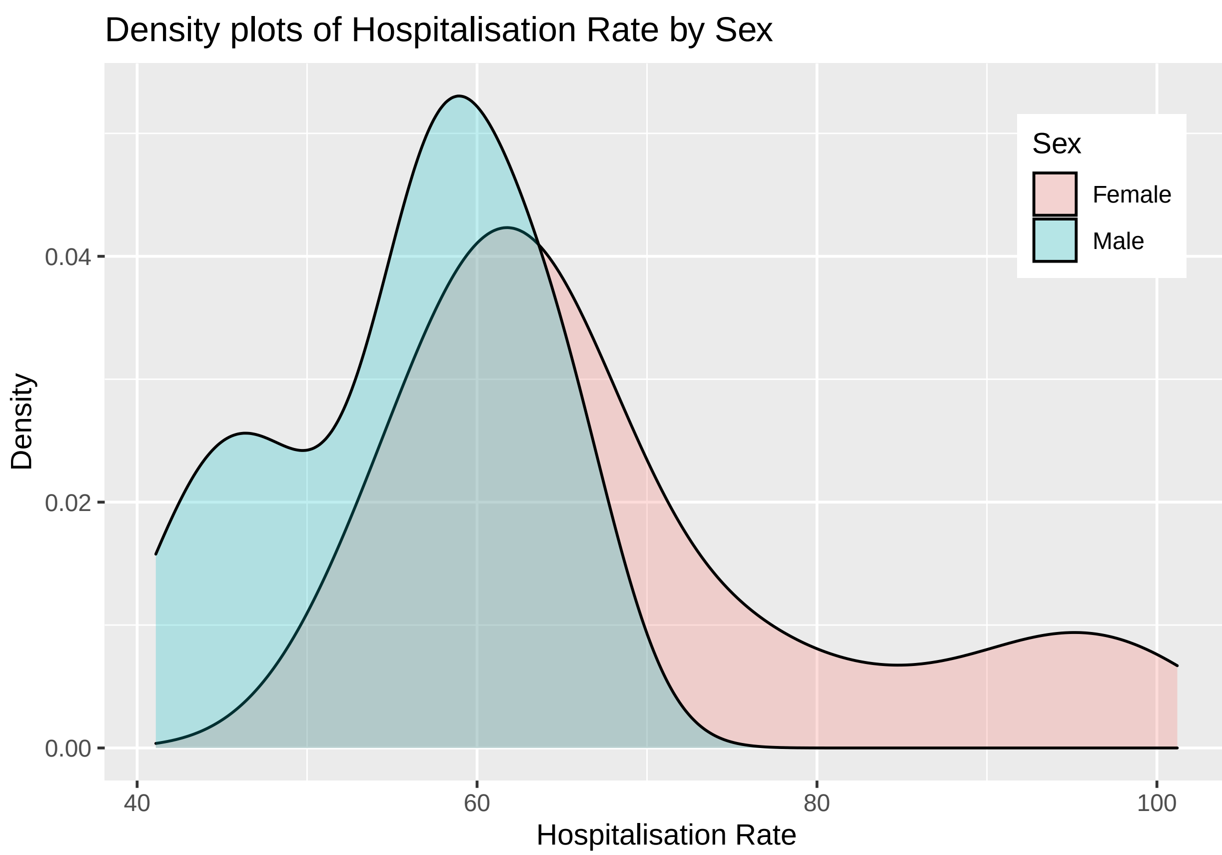


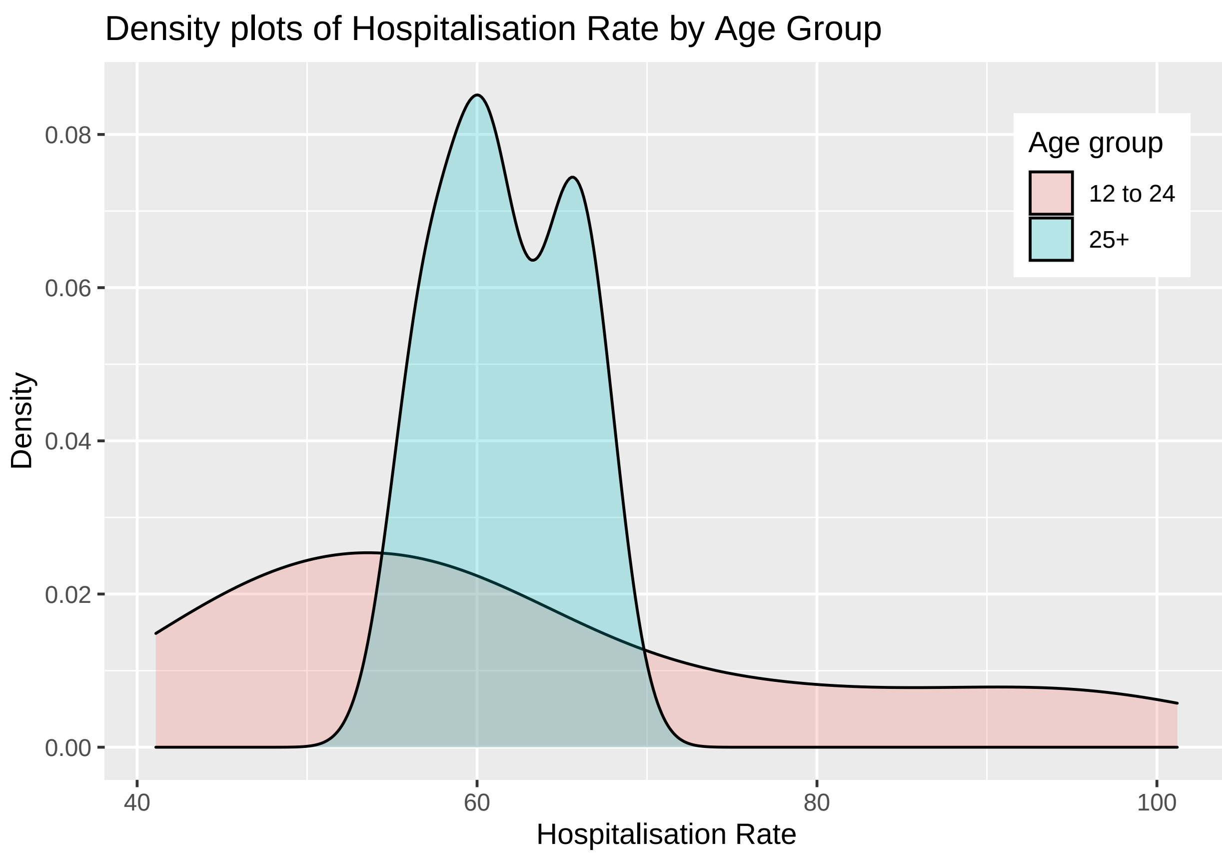
Figure 1: Histogram of hospitalisation rate.

From Figure 1, It does not seem to be normally distributed, and the distribution is positively skewed or right skewed.

```r
# Instead of colour = sex, fill = sex is used
ggplot(data = mental, aes(y = stat(density), x = hospitalisation.rate,
    fill = sex)) + geom_density(kernel = "gaussian", alpha = 0.25) +
    theme(legend.position = c(0.8875, 0.815)) + labs(title = "Density plots
     ↪  of Hospitalisation Rate by Sex",
    x = "Hospitalisation Rate", y = "Density", fill = "Sex")
```



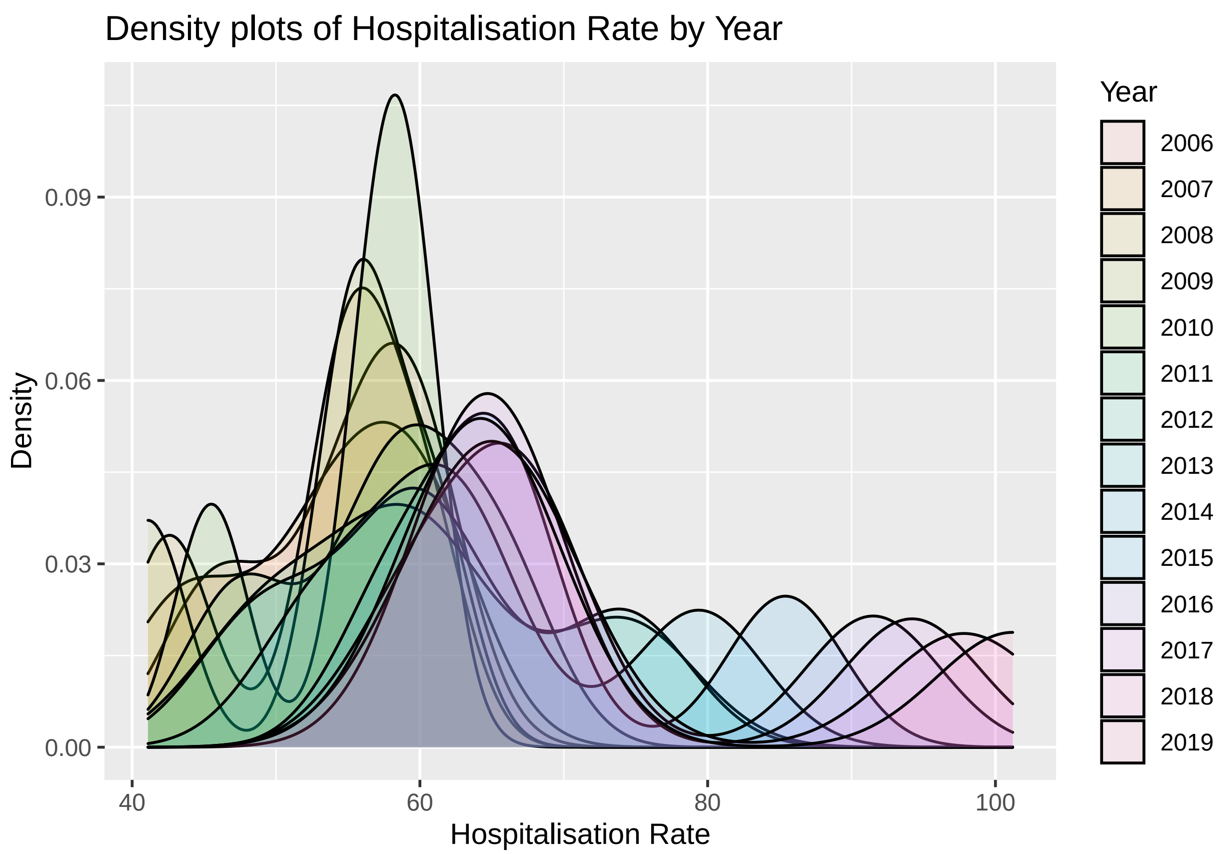

```r
ggplot(data = mental, aes(y = stat(density), x = hospitalisation.rate,
    fill = age.group)) + geom_density(kernel = "gaussian", alpha = 0.25) +
    theme(legend.position = c(0.8875, 0.815)) + labs(title = "Density plots
     ↪  of Hospitalisation Rate by Age Group",
    x = "Hospitalisation Rate", y = "Density", fill = "Age group")
```



The distribution for hospitalisation rate for age group 25+ years old patients shows a bimodal distribution, while for age group 12 to 25, the distribution is rather uniform, and can be interpreted as right-skewed.

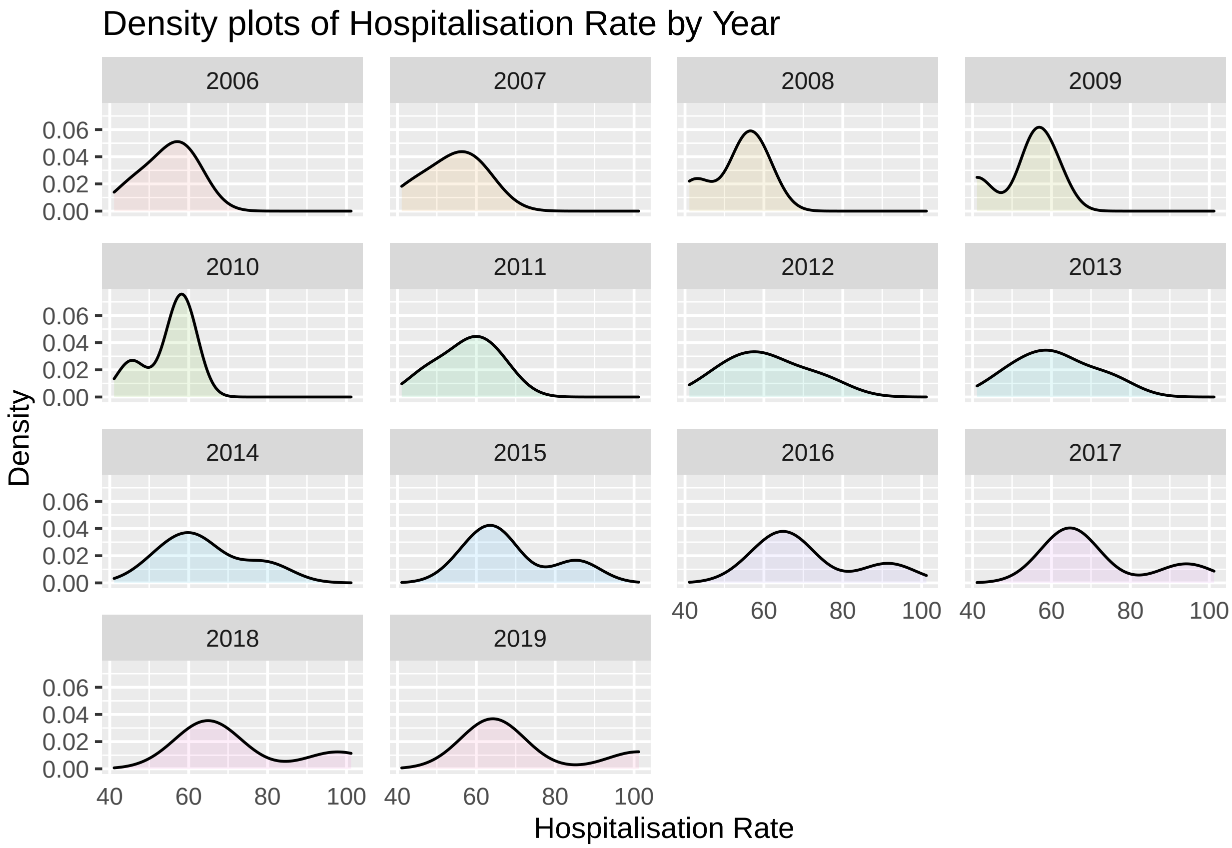
```
ggplot(data = mental, aes(y = stat(density), x = hospitalisation.rate,
    fill = year)) + geom_density(kernel = "gaussian", alpha = 0.1) +

labs(title = "Density plots of Hospitalisation Rate by Year",
    x = "Hospitalisation Rate", y = "Density", fill = "Year")
```

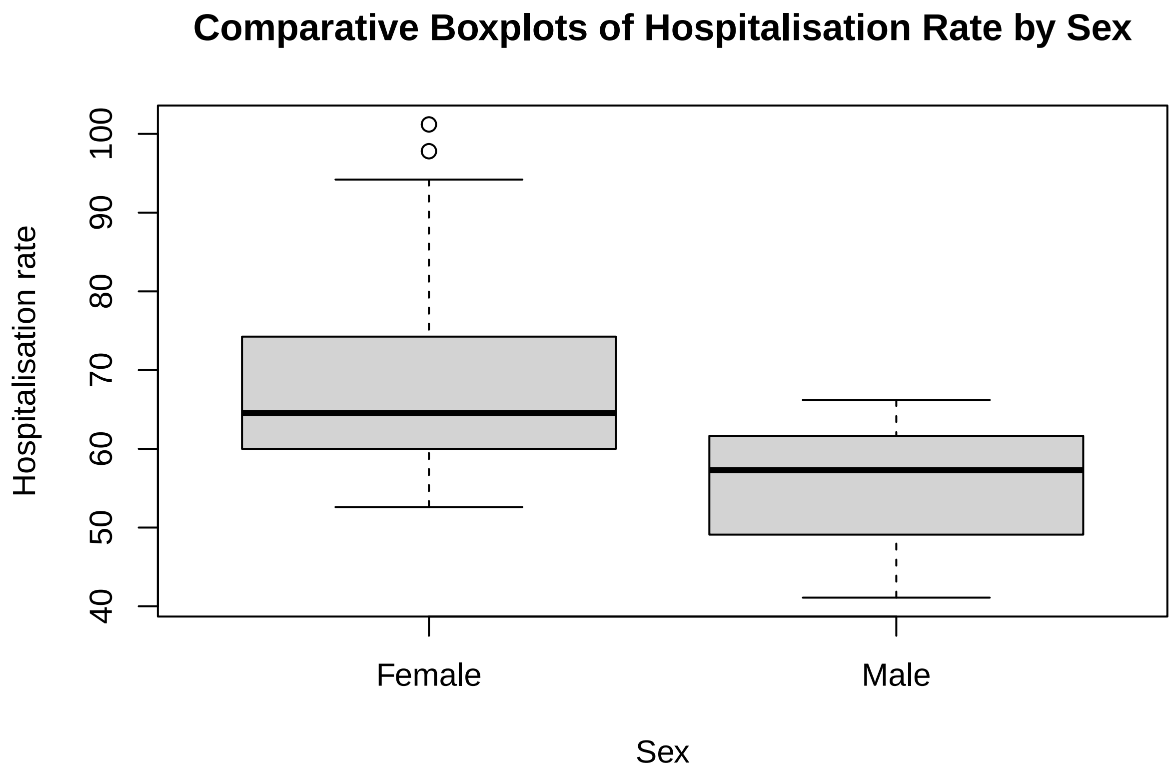


The above density plot by year is too difficult to interpret, since the factor year variable has too many levels.

```
ggplot(data = mental, aes(y = stat(density), x = hospitalisation.rate,
    fill = year)) + geom_density(kernel = "gaussian", alpha = 0.1,
    adjust = 1.5) + facet_wrap(~year) + theme(legend.position = "none",
    panel.spacing = unit(0.7, "lines"), axis.ticks.x = element_blank()) +
    labs(title = "Density plots of Hospitalisation Rate by Year",
        x = "Hospitalisation Rate", y = "Density", fill = "Year")
```




```
boxplot(mental$hospitalisation.rate ~ mental$sex, xlab = "Sex",
    ylab = "Hospitalisation rate", main = "Comparative Boxplots of
    ↪  Hospitalisation Rate by Sex")
```

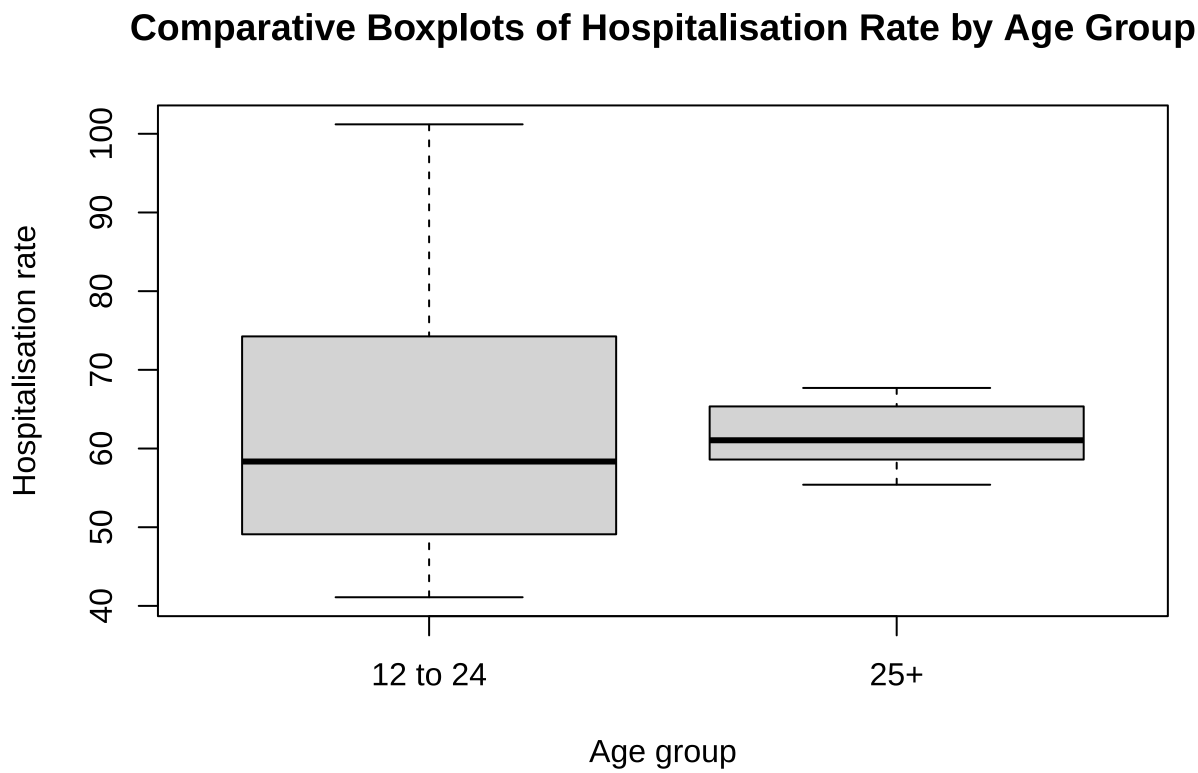



```
boxplot.stats(mental$hospitalisation.rate[mental$sex == "Female"])$out   #
↪  outliers for female
```

```
[1]  97.8 101.2
```

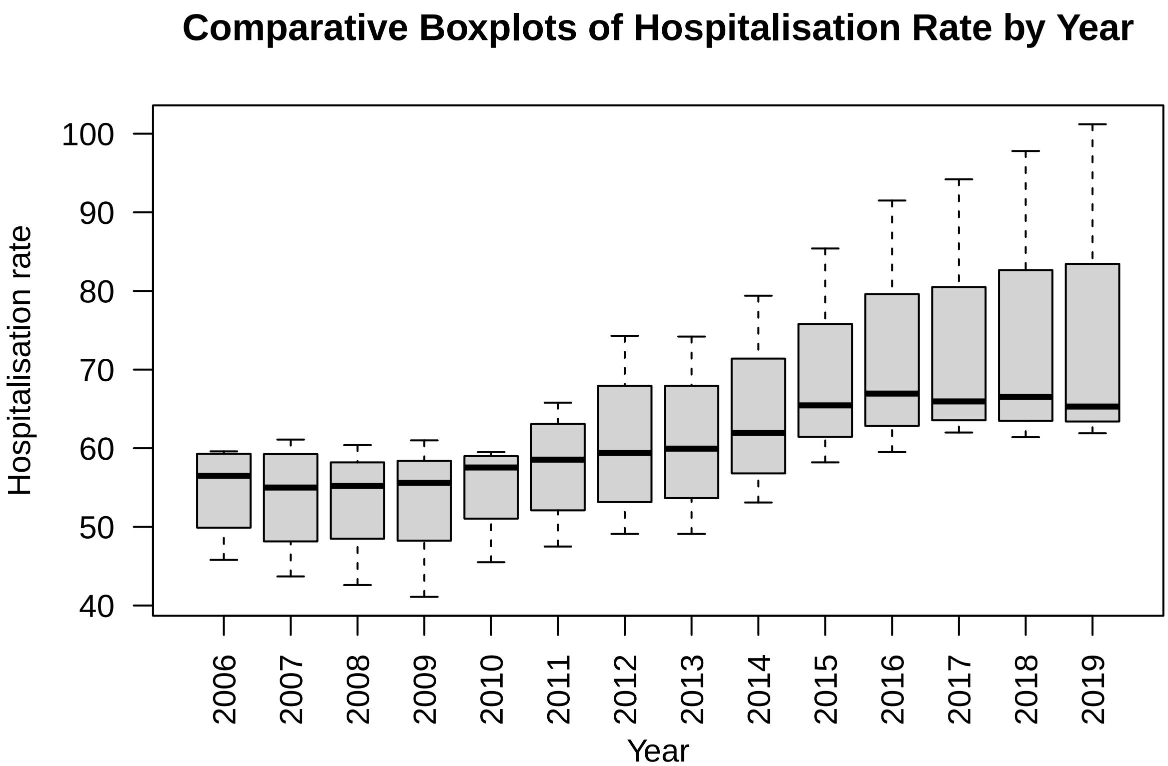
The median hospitalisation rate for males are smaller than for females. The hospitalisation rate for females have a larger variability (range), and there are two outliers with hospitalisation rate values 97.8 and 101.2.

```
boxplot(mental$hospitalisation.rate ~ mental$age.group, xlab = "Age group",
    ylab = "Hospitalisation rate", main = "Comparative Boxplots of
     ↪  Hospitalisation Rate by Age Group")
```



Visually, the spread of hospitalisation rate for age group “12 to 24” is comparatively larger than the spread of age group “25+”.

```
boxplot(mental$hospitalisation.rate ~ mental$year, xlab = "Year",
    ylab = "Hospitalisation rate", main = "Comparative Boxplots of
     ↪  Hospitalisation Rate by Year",
    las = 2)
```



Visually, the median of the hospitalisation rate increases over the years from 2006 to 2019. More interestingly, the right whisker of the box-and-whisker diagram becomes noticeably longer over the years, and between 2015 to 2019, the distribution of hospitalisation rate in each year becomes more noticeably right-skewed or positively skewed. The longer whiskers are supported by the increasing range of values over the years.

2.1.3 Size of the dataset

```
nrow(mental)
```

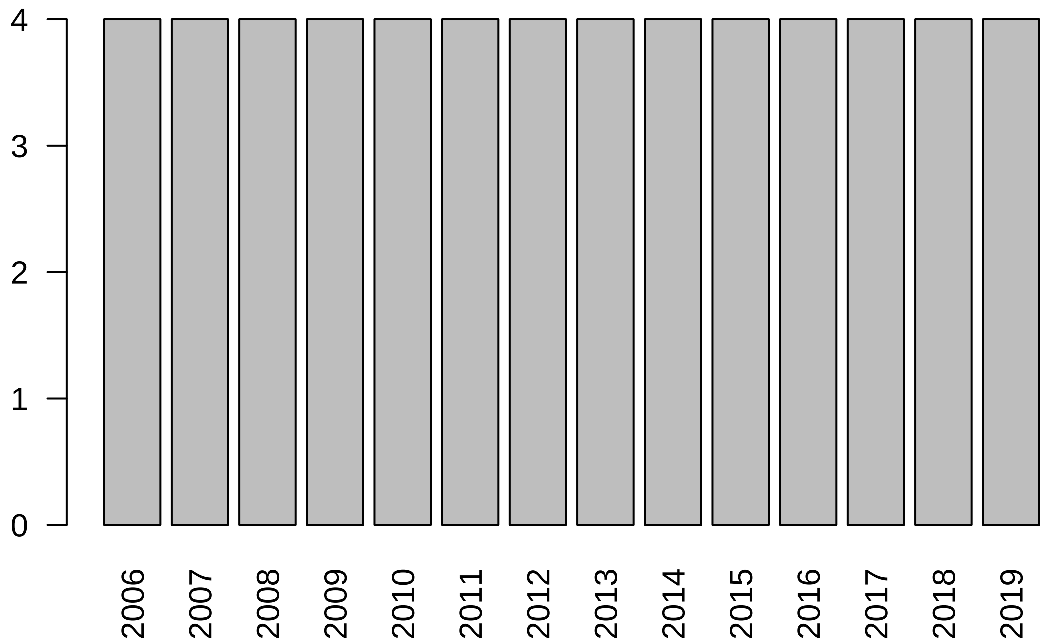
```
[1] 56
```

There are 56 rows of data in the dataset or 56 sets of values. This could be consider as a small sample size for some analyses.

2.1.4 Completeness of the dataset

From the outputs above, the dataset has no missing values, and there are equal frequency of values for each of the factor levels. Hence, I would consider that the dataset is balanced.

```r
barplot(table(mental$year), las = 2)
```



2.2 Part (b) Models considered in equation format

2.2.1 With and without *log* transformation

The variables that are categorical have already been converted into `factor` datatype in R and have been replaced in the dataframe as such.

The models would consider transformations.

The factor variable `year` is included, predictions could be made, given a year between 2006 to 2019 on a xtest dataset.

```r
models.1 <- c("hospitalisation.rate ~ factor(year) + factor(sex) +
↪  factor(age.group)",
    "log(hospitalisation.rate) ~ factor(year) + factor(sex) +
     ↪  factor(age.group)",
    "hospitalisation.rate ~ factor(sex) + factor(age.group)")
```

```r
# mental$age.group <- relevel(mental$age.group, ref =
# '25+') mental$sex <- relevel(mental$sex, ref = 'Male')
# mental$year <- relevel(mental$year, ref = '2007')
```

```r
print_eqns(models.1, dataset = mental)
```

$$\text{Model 1}: \begin{aligned} hospitalisation.rate = {} & \beta_0 + \beta_1(factor(year)_{2007}) + \\ & \beta_2(factor(year)_{2008}) + \beta_3(factor(year)_{2009}) + \\ & \beta_4(factor(year)_{2010}) + \beta_5(factor(year)_{2011}) + \\ & \beta_6(factor(year)_{2012}) + \beta_7(factor(year)_{2013}) + \\ & \beta_8(factor(year)_{2014}) + \beta_9(factor(year)_{2015}) + \\ & \beta_{10}(factor(year)_{2016}) + \beta_{11}(factor(year)_{2017}) + \\ & \beta_{12}(factor(year)_{2018}) + \beta_{13}(factor(year)_{2019}) + \\ & \beta_{14}(factor(sex)_{Male}) + \beta_{15}(factor(age.group)_{25+}) + \\ & \epsilon \end{aligned}$$

$$\text{Model 2}: \begin{aligned} log(hospitalisation.rate) = {} & \beta_0 + \beta_1(factor(year)_{2007}) + \\ & \beta_2(factor(year)_{2008}) + \beta_3(factor(year)_{2009}) + \\ & \beta_4(factor(year)_{2010}) + \beta_5(factor(year)_{2011}) + \\ & \beta_6(factor(year)_{2012}) + \beta_7(factor(year)_{2013}) + \\ & \beta_8(factor(year)_{2014}) + \beta_9(factor(year)_{2015}) + \\ & \beta_{10}(factor(year)_{2016}) + \beta_{11}(factor(year)_{2017}) + \\ & \beta_{12}(factor(year)_{2018}) + \beta_{13}(factor(year)_{2019}) + \\ & \beta_{14}(factor(sex)_{Male}) + \beta_{15}(factor(age.group)_{25+}) + \\ & \epsilon \end{aligned}$$

$$\text{Model 3}: \begin{aligned} hospitalisation.rate = {} & \beta_0 + \beta_1(factor(sex)_{Male}) + \\ & \beta_2(factor(age.group)_{25+}) + \epsilon \end{aligned}$$

I first try to fit the model without transformation.

```r
mental.model.1 <- lm(as.formula(models.1[1]), data = mental)

summary(mental.model.1)
```

```
Call:
lm(formula = as.formula(models.1[1]), data = mental)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-13.514  -5.679  -2.264   4.811  20.789

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            61.586      4.814  12.795 1.01e-15 ***
factor(year)2007       -0.900      6.368  -0.141  0.88831    
factor(year)2008       -1.250      6.368  -0.196  0.84536    
factor(year)2009       -1.275      6.368  -0.200  0.84231    
factor(year)2010        0.425      6.368   0.067  0.94712    
factor(year)2011        3.000      6.368   0.471  0.64010    
factor(year)2012        5.950      6.368   0.934  0.35570    
factor(year)2013        6.200      6.368   0.974  0.33607    
factor(year)2014        9.500      6.368   1.492  0.14356    
factor(year)2015       14.025      6.368   2.203  0.03345 *  
factor(year)2016       16.625      6.368   2.611  0.01265 *  
factor(year)2017       17.425      6.368   2.737  0.00922 ** 
factor(year)2018       18.475      6.368   2.901  0.00601 ** 
factor(year)2019       18.825      6.368   2.956  0.00520 ** 
factor(sex)Male       -12.775      2.407  -5.308 4.42e-06 ***
factor(age.group)25+   -1.196      2.407  -0.497  0.62183    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.005 on 40 degrees of freedom
Multiple R-squared:  0.6345,	Adjusted R-squared:  0.4974 
F-statistic: 4.629 on 15 and 40 DF,  p-value: 5.204e-05
```

```r
confint(mental.model.1, parm = c("factor(sex)Male", "factor(year)2015",
    "factor(year)2016"), level = 0.95)
```

```
                     2.5 %    97.5 %
factor(sex)Male  -17.639174 -7.910826
factor(year)2015   1.155606 26.894394
factor(year)2016   3.755606 29.494394
```

In the above model, we are taking year = 2006, sex = female, and age group = "12 to 24" as the reference variables. There is significantly higher hospitalisation rate for years 2015 to 2019 compared to year 2006. Males have significantly lower hospitalisation rate than females of around 12.775 (-17.64, -7.91) per 10,000 population in Australia.

```r
mental.model.2 <- lm(as.formula(models.1[2]), data = mental)

summary(mental.model.2)
```

```
Call:
lm(formula = as.formula(models.1[2]), data = mental)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.19613 -0.07991 -0.02370  0.06516  0.25000 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           4.086573   0.068605  59.567  < 2e-16 ***
factor(year)2007     -0.018960   0.090755  -0.209  0.83557    
factor(year)2008     -0.025975   0.090755  -0.286  0.77619    
factor(year)2009     -0.028806   0.090755  -0.317  0.75259    
factor(year)2010      0.007566   0.090755   0.083  0.93398    
factor(year)2011      0.051913   0.090755   0.572  0.57052    
factor(year)2012      0.097681   0.090755   1.076  0.28824    
factor(year)2013      0.102083   0.090755   1.125  0.26737    
factor(year)2014      0.155080   0.090755   1.709  0.09524 .  
factor(year)2015      0.223803   0.090755   2.466  0.01805 *  
factor(year)2016      0.257905   0.090755   2.842  0.00703 ** 
factor(year)2017      0.267930   0.090755   2.952  0.00526 ** 
factor(year)2018      0.279394   0.090755   3.079  0.00375 ** 
factor(year)2019      0.280526   0.090755   3.091  0.00362 ** 
factor(sex)Male      -0.198042   0.034302  -5.773 9.84e-07 ***
factor(age.group)25+  0.014126   0.034302   0.412  0.68268    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1283 on 40 degrees of freedom
Multiple R-squared:  0.6718,	Adjusted R-squared:  0.5488 
F-statistic: 5.459 on 15 and 40 DF,  p-value: 8.488e-06
```

As identified in section 2.3, there are no real differences between the model with transformation and the model without transformation.

Therefore, to avoid affecting the scale of hospitalisation rate, I have decided to retain and use the model without any transformation.

```r
mental.model.3 <- lm(as.formula(models.1[3]), data = mental)

summary(mental.model.3)
```

```
Call:
lm(formula = as.formula(models.1[3]), data = mental)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.630  -7.634  -0.434   5.112  31.970 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            69.230      2.577  26.865  < 2e-16 ***
factor(sex)Male       -12.775      2.976  -4.293 7.53e-05 ***
factor(age.group)25+   -1.196      2.976  -0.402    0.689    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.13 on 53 degrees of freedom
Multiple R-squared:  0.2597,	Adjusted R-squared:  0.2318 
F-statistic: 9.297 on 2 and 53 DF,  p-value: 0.0003461
```

2.2.2 With interaction effects (without transformations)

After studying the *main effects*, I would want to consider if there are any interaction effects between the independent variables.

I consider all the possible models with interaction effect, for models without any transformations.

Similarly, I have already converted the factor variables to the correct datatype (factor) in the dataframe, so I don't have to directly do it on the model equations here. The first model without interaction effect is included.

The number of possible interaction effects between the three of the categorical variables are quite numerous. This adds to the model complexity and should only be included if there is a reason to believe that interaction effects between the terms exist, and we want to determine if there is an improvement to the model compared to not having them.

The most probable interaction terms is/are:

- sex * age.group – Could sex depend on the age group that results in significant difference in the hospitalisation rate?
- year * sex – Could sex depend on the year in question resulting in significant differences in the hospitalisation rate?

```r
models.1.interaction <- c("hospitalisation.rate ~ year + sex + age.group",
    "hospitalisation.rate ~ year * sex", "hospitalisation.rate ~ year *
     ↪  age.group",
    "hospitalisation.rate ~ sex * age.group")
```

```r
mental.model.1.interaction.2 <- lm(as.formula(models.1.interaction[2]),
    data = mental)

summary(mental.model.1.interaction.2)
```

```

Call:
lm(formula = as.formula(models.1.interaction[2]), data = mental)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.750  -4.562   0.000   4.562  17.750 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        56.800      7.267   7.816 1.63e-08 ***
year2007            0.050     10.277   0.005   0.9962    
year2008            0.600     10.277   0.058   0.9539    
year2009            1.600     10.277   0.156   0.8774    
year2010            2.200     10.277   0.214   0.8320    
year2011            6.300     10.277   0.613   0.5448    
year2012           11.150     10.277   1.085   0.2872    
year2013           11.150     10.277   1.085   0.2872    
year2014           14.600     10.277   1.421   0.1665    
year2015           19.000     10.277   1.849   0.0751 .  
year2016           22.800     10.277   2.218   0.0348 *  
year2017           23.700     10.277   2.306   0.0287 *  
year2018           25.850     10.277   2.515   0.0179 *  
year2019           26.650     10.277   2.593   0.0150 *  
sexMale            -4.400     10.277  -0.428   0.6718    
year2007:sexMale   -1.900     14.534  -0.131   0.8969    
year2008:sexMale   -3.700     14.534  -0.255   0.8009    
year2009:sexMale   -5.750     14.534  -0.396   0.6954    
year2010:sexMale   -3.550     14.534  -0.244   0.8088    
year2011:sexMale   -6.600     14.534  -0.454   0.6533    
year2012:sexMale  -10.400     14.534  -0.716   0.4802    
year2013:sexMale   -9.900     14.534  -0.681   0.5014    
year2014:sexMale  -10.200     14.534  -0.702   0.4886    
```

```
year2015:sexMale   -9.950     14.534  -0.685   0.4992
year2016:sexMale  -12.350     14.534  -0.850   0.4027
year2017:sexMale  -12.550     14.534  -0.863   0.3952
year2018:sexMale  -14.750     14.534  -1.015   0.3189
year2019:sexMale  -15.650     14.534  -1.077   0.2908
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.28 on 28 degrees of freedom
Multiple R-squared:  0.6668,    Adjusted R-squared:  0.3454
F-statistic: 2.075 on 27 and 28 DF,  p-value: 0.0299
```

```
summ(mental.model.1.interaction.2)
```

```
MODEL INFO:
Observations: 56
Dependent Variable: hospitalisation.rate
Type: OLS linear regression

MODEL FIT:
F(27,28) = 2.07, p = 0.03
R² = 0.67
Adj. R² = 0.35

Standard errors: OLS
-------------------------------------------------------
                           Est.    S.E.   t val.      p
---------------------- -------- ------- -------- ------
(Intercept)               56.80    7.27     7.82   0.00
year2007                   0.05   10.28     0.00   1.00
year2008                   0.60   10.28     0.06   0.95
year2009                   1.60   10.28     0.16   0.88
year2010                   2.20   10.28     0.21   0.83
year2011                   6.30   10.28     0.61   0.54
year2012                  11.15   10.28     1.08   0.29
year2013                  11.15   10.28     1.08   0.29
year2014                  14.60   10.28     1.42   0.17
year2015                  19.00   10.28     1.85   0.08
year2016                  22.80   10.28     2.22   0.03
year2017                  23.70   10.28     2.31   0.03
year2018                  25.85   10.28     2.52   0.02
year2019                  26.65   10.28     2.59   0.01
sexMale                   -4.40   10.28    -0.43   0.67
year2007:sexMale          -1.90   14.53    -0.13   0.90
```

```
year2008:sexMale          -3.70   14.53    -0.25   0.80
year2009:sexMale          -5.75   14.53    -0.40   0.70
year2010:sexMale          -3.55   14.53    -0.24   0.81
year2011:sexMale          -6.60   14.53    -0.45   0.65
year2012:sexMale         -10.40   14.53    -0.72   0.48
year2013:sexMale          -9.90   14.53    -0.68   0.50
year2014:sexMale         -10.20   14.53    -0.70   0.49
year2015:sexMale          -9.95   14.53    -0.68   0.50
year2016:sexMale         -12.35   14.53    -0.85   0.40
year2017:sexMale         -12.55   14.53    -0.86   0.40
year2018:sexMale         -14.75   14.53    -1.01   0.32
year2019:sexMale         -15.65   14.53    -1.08   0.29
-------------------------------------------------------
```

```r
mental.model.1.interaction.3 <- lm(as.formula(models.1.interaction[3]),
    data = mental)

summary(mental.model.1.interaction.3)
```

```
Call:
lm(formula = as.formula(models.1.interaction[3]), data = mental)

Residuals:
    Min      1Q  Median      3Q     Max
-19.650  -3.125   0.000   3.125  19.650

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)              49.90       8.85   5.639 4.86e-06 ***
year2007                 -1.75      12.52  -0.140   0.8898
year2008                 -1.40      12.52  -0.112   0.9117
year2009                 -1.45      12.52  -0.116   0.9086
year2010                  2.10      12.52   0.168   0.8679
year2011                  6.75      12.52   0.539   0.5939
year2012                 11.80      12.52   0.943   0.3538
year2013                 11.75      12.52   0.939   0.3558
year2014                 16.35      12.52   1.306   0.2020
year2015                 21.90      12.52   1.750   0.0911 .
year2016                 25.60      12.52   2.046   0.0503 .
year2017                 28.20      12.52   2.253   0.0323 *
year2018                 29.70      12.52   2.373   0.0247 *
year2019                 31.65      12.52   2.529   0.0174 *
age.group25+              9.40      12.52   0.751   0.4589
```

```
year2007:age.group25+     1.70      17.70   0.096   0.9242
year2008:age.group25+     0.30      17.70   0.017   0.9866
year2009:age.group25+     0.35      17.70   0.020   0.9844
year2010:age.group25+    -3.35      17.70  -0.189   0.8512
year2011:age.group25+    -7.50      17.70  -0.424   0.6750
year2012:age.group25+   -11.70      17.70  -0.661   0.5140
year2013:age.group25+   -11.10      17.70  -0.627   0.5356
year2014:age.group25+   -13.70      17.70  -0.774   0.4454
year2015:age.group25+   -15.75      17.70  -0.890   0.3811
year2016:age.group25+   -17.95      17.70  -1.014   0.3192
year2017:age.group25+   -21.55      17.70  -1.218   0.2335
year2018:age.group25+   -22.45      17.70  -1.268   0.2151
year2019:age.group25+   -25.65      17.70  -1.449   0.1584
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.52 on 28 degrees of freedom
Multiple R-squared:  0.5058,    Adjusted R-squared:  0.02932
F-statistic: 1.062 on 27 and 28 DF,  p-value: 0.4374
```

```
summ(mental.model.1.interaction.3, confint = TRUE)
```

```
MODEL INFO:
Observations: 56
Dependent Variable: hospitalisation.rate
Type: OLS linear regression

MODEL FIT:
F(27,28) = 1.06, p = 0.44
R² = 0.51
Adj. R² = 0.03

Standard errors: OLS
---------------------------------------------------------------------
                               Est.     2.5%   97.5%   t val.       p
--------------------------- -------- -------- ------- -------- ------
(Intercept)                   49.90    31.77   68.03     5.64    0.00
year2007                      -1.75   -27.39   23.89    -0.14    0.89
year2008                      -1.40   -27.04   24.24    -0.11    0.91
year2009                      -1.45   -27.09   24.19    -0.12    0.91
year2010                       2.10   -23.54   27.74     0.17    0.87
year2011                       6.75   -18.89   32.39     0.54    0.59
year2012                      11.80   -13.84   37.44     0.94    0.35
year2013                      11.75   -13.89   37.39     0.94    0.36
```

```
year2014                        16.35    -9.29   41.99     1.31   0.20
year2015                        21.90    -3.74   47.54     1.75   0.09
year2016                        25.60    -0.04   51.24     2.05   0.05
year2017                        28.20     2.56   53.84     2.25   0.03
year2018                        29.70     4.06   55.34     2.37   0.02
year2019                        31.65     6.01   57.29     2.53   0.02
age.group25+                     9.40   -16.24   35.04     0.75   0.46
year2007:age.group25+            1.70   -34.55   37.95     0.10   0.92
year2008:age.group25+            0.30   -35.95   36.55     0.02   0.99
year2009:age.group25+            0.35   -35.90   36.60     0.02   0.98
year2010:age.group25+           -3.35   -39.60   32.90    -0.19   0.85
year2011:age.group25+           -7.50   -43.75   28.75    -0.42   0.67
year2012:age.group25+          -11.70   -47.95   24.55    -0.66   0.51
year2013:age.group25+          -11.10   -47.35   25.15    -0.63   0.54
year2014:age.group25+          -13.70   -49.95   22.55    -0.77   0.45
year2015:age.group25+          -15.75   -52.00   20.50    -0.89   0.38
year2016:age.group25+          -17.95   -54.20   18.30    -1.01   0.32
year2017:age.group25+          -21.55   -57.80   14.70    -1.22   0.23
year2018:age.group25+          -22.45   -58.70   13.80    -1.27   0.22
year2019:age.group25+          -25.65   -61.90   10.60    -1.45   0.16
---------------------------------------------------------------------
```

```r
mental.model.1.interaction.4 <- lm(as.formula(models.1.interaction[4]),
    data = mental)

summary(mental.model.1.interaction.4)
```

```
Call:
lm(formula = as.formula(models.1.interaction[4]), data = mental)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.621  -3.729  -1.393   4.705  26.979 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            74.221      2.666  27.839  < 2e-16 ***
sexMale               -22.757      3.770  -6.036 1.69e-07 ***
age.group25+          -11.179      3.770  -2.965 0.004564 ** 
sexMale:age.group25+   19.964      5.332   3.744 0.000454 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 9.976 on 52 degrees of freedom
Multiple R-squared:  0.4169,	Adjusted R-squared:  0.3833
F-statistic: 12.39 on 3 and 52 DF,  p-value: 3.136e-06
```

```r
summ(mental.model.1.interaction.4, confint = TRUE)
```

```
MODEL INFO:
Observations: 56
Dependent Variable: hospitalisation.rate
Type: OLS linear regression

MODEL FIT:
F(3,52) = 12.39, p = 0.00
R² = 0.42
Adj. R² = 0.38

Standard errors: OLS
---------------------------------------------------------------------
                              Est.     2.5%    97.5%   t val.       p
-------------------------- -------- -------- -------- -------- ------
(Intercept)                  74.22    68.87    79.57    27.84   0.00
sexMale                     -22.76   -30.32   -15.19    -6.04   0.00
age.group25+                -11.18   -18.74    -3.61    -2.96   0.00
sexMale:age.group25+         19.96     9.26    30.66     3.74   0.00
---------------------------------------------------------------------
```

The interaction effect and the main effects are statistically significant. The adjusted $R^2 = 0.38$.

2.3 Part (c) - Consideration of assumptions

2.3.1 Linearity

Definition: The relationship between the response variable `hospitalisation.rate` and each of the explanatory variables being linear. However, since the explanatory variables are all categorical, this is not applicable.

The linearity of the linear regression cannot be assessed, because the explanatory variables are all categorical. It is inappropriate to visualise the relationship between `hospitalisation.rate` and the categorical variables on a scatterplot for assessment of linearity.

2.3.2 Homoscedasticity (equal variances) and normality

Definition: Homoscedasticity assumption is violated (*heteroscedasticity*) when the size of the error term is not the same across values of the explanatory variables. Instead of the spread of data across residuals being constant (cigar-shaped), a funnel-like pattern (e.g., increasing variance of residuals) across fitted values can be seen. For normality assumption to hold, the residuals of the model should tend towards a normal distribution.

For assessment of normality and homoscedasticity of linear regression, we can use the normal Q-Q plot of residuals and scatterplot of residuals vs. fitted values for the linear model.

See figures 2 and 3 for the normal Q-Q plot and scatterplot of residuals vs. fitted values before the log transformation, and figures 4 and 5 for the plots after log transformation of the `hospitalisation.rate` variable.

With reference to figure 2, the distribution is unlikely to be normal, as there are visually noticeable deviation of the points from the normal line. Hence, the distribution is unlikely to be normal, normality assumption does not hold. Furthermore, figure 1, shows the shape of the histogram is quite positively skewed.

From figure 3, there is some funnel-like shape present in the plot of residuals vs. fitted values. Hence, the assumption of equal variances (homoscedastity) seems to be violated.

After log transformation, and taking reference from 4 and 5, there is no real difference between the scatterplot of residuals vs. fitted values. The funnel-like shape still exists, even though on smaller scales. There is only very slight to no real improvement in the normal Q-Q plot, the distribution is likely not normal and is heteroscedastic.

2.3.3 Independence of observations

Definition: Each observation is independent of other observations in the dataset (no replicates).

Independence of observations is unlikely to be violated (and assumed as not violated) as each hospitalisation rate observation is from a distinct and different year, sex and age group.

However, for the purposes of analyses, I assume that these assumptions are not violated for multiple linear regression. I would assume that the model before or after log transformation is normal and homoscedastic. Hence, for the descriptive statistics, I mostly rely on assuming that the distribution is normal.

2.4 Part (d) - Visualisation of model diagnostics with interpretation

```
qqnorm(mental.model.1$residuals, main = "Normal Q-Q Plot of Residuals")
qqline(mental.model.1$residuals, lw = 1.5, col = "red")
```

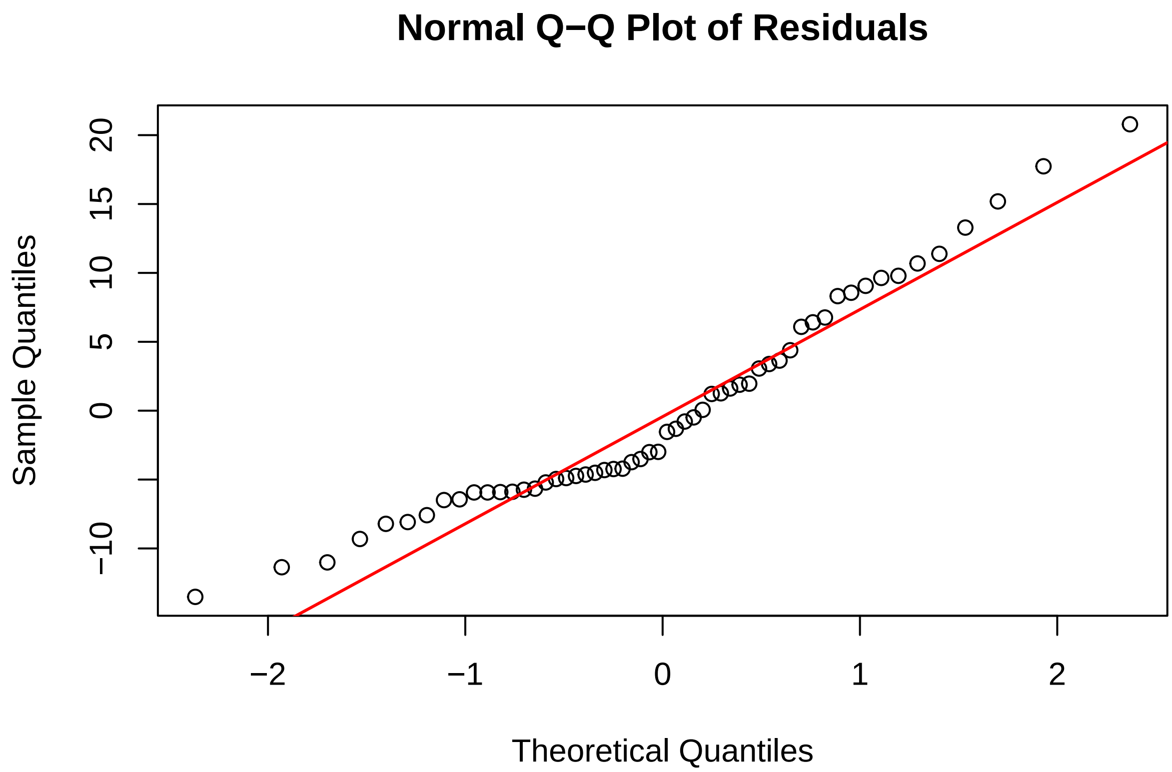


Figure 2: Normal Q-Q Plot of Residuals before any log transformation

```
plot(mental.model.1$fitted.values, mental.model.1$residuals,
    xlab = "Fitted Values", ylab = "Residuals", main = "Scatterplot of
     ↪  Residuals vs. Fitted Values")
abline(h = 0, col = "red")
```

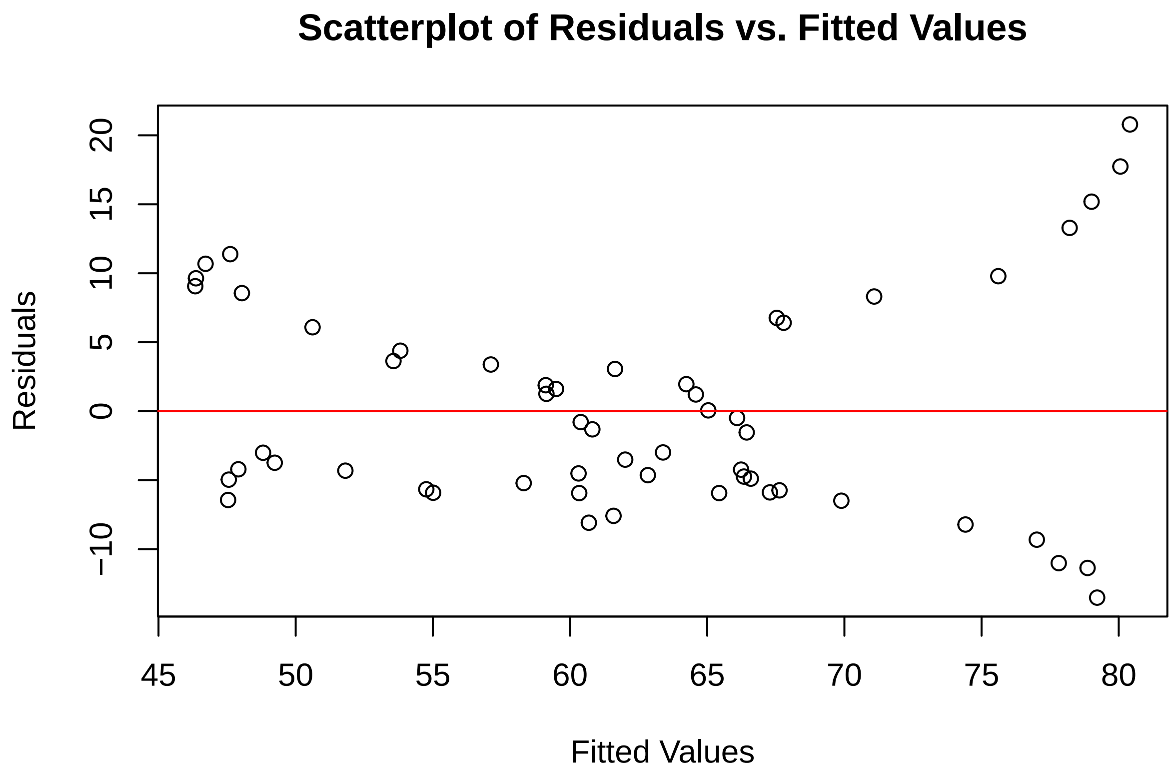


Figure 3: Scatterplot of residuals vs. fitted values before log transformation

As mentioned in section 2.3.2, there are deviation of points from the normal line on the normal QQ plot, indicating that the distribution before log transformation is not normal. There is also some funnel-like shape from the residuals of the fitted values above 65. Hence, this violates the assumptions, and makes the model unreliable.

```
qqnorm(mental.model.2$residuals, main = "Normal Q-Q Plot of Residuals")
qqline(mental.model.2$residuals, lw = 1.5, col = "red")
```

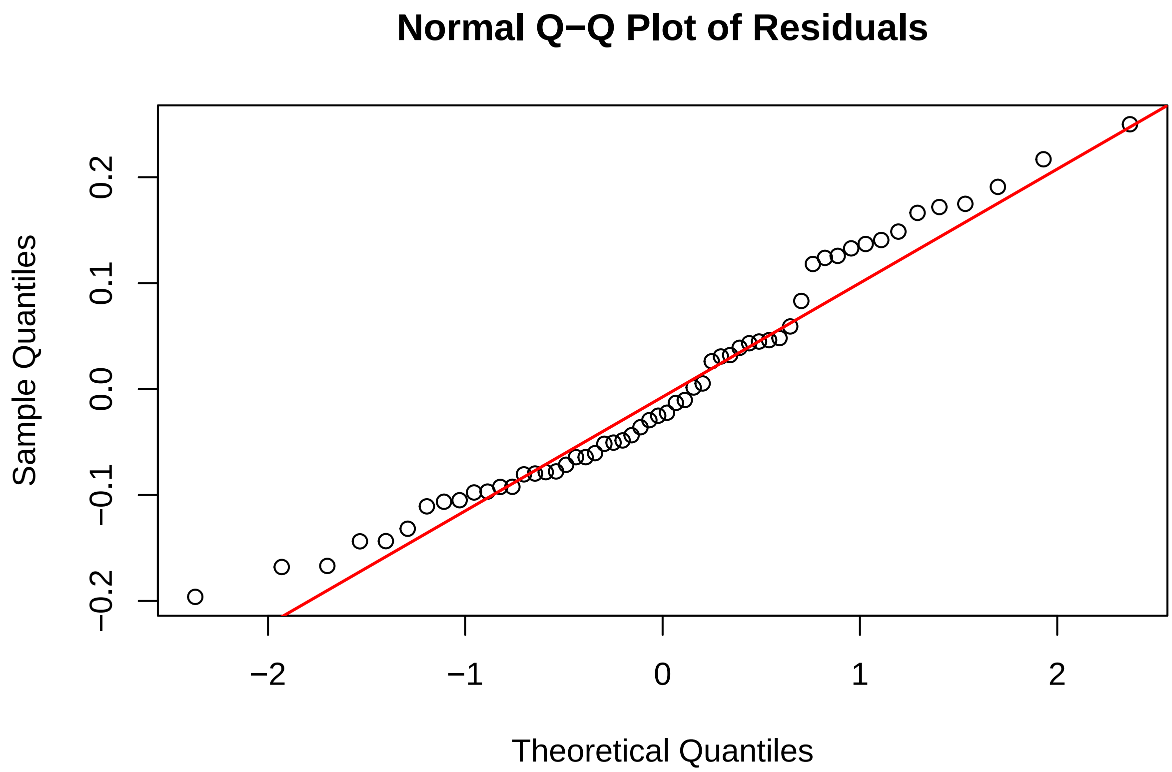


Figure 4: Normal Q-Q Plot of Residuals after log transform of `hospitalisation.rate` variable

```
plot(mental.model.2$fitted.values, mental.model.2$residuals,
    xlab = "Fitted Values", ylab = "Residuals", main = "Scatterplot of
     ↪  Residuals vs. Fitted Values")
abline(h = 0, col = "red")
```

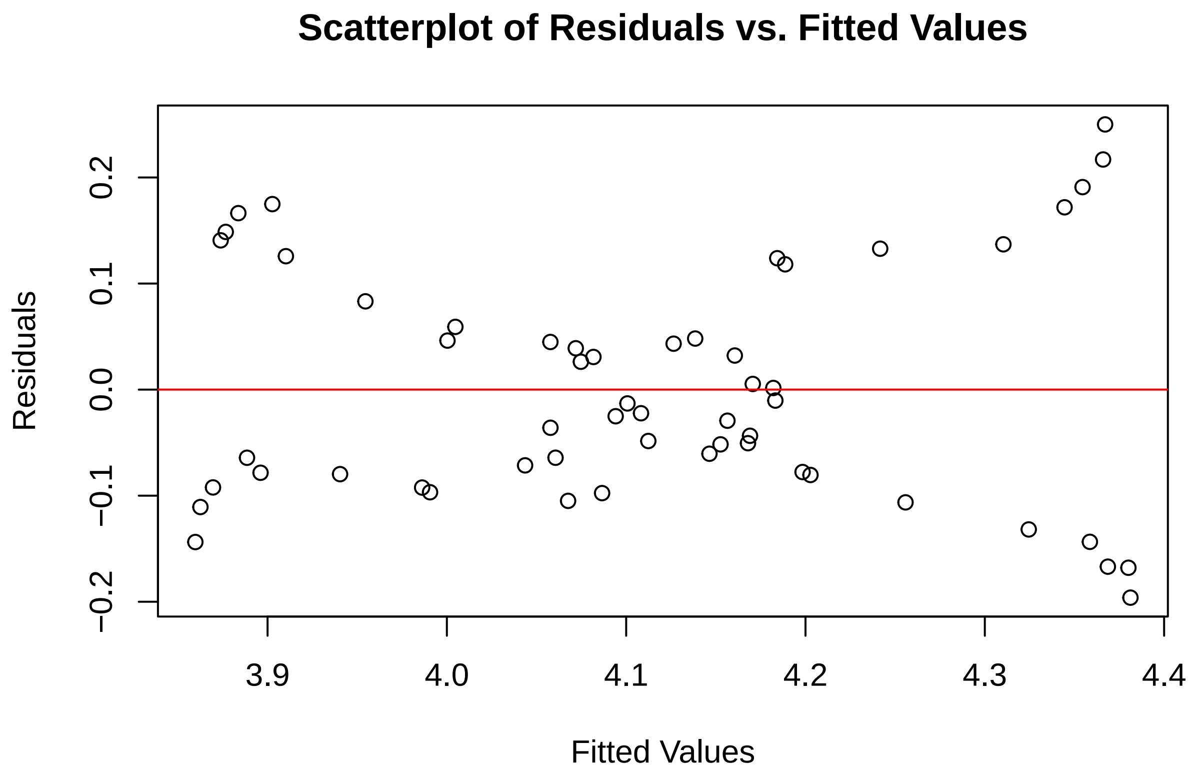
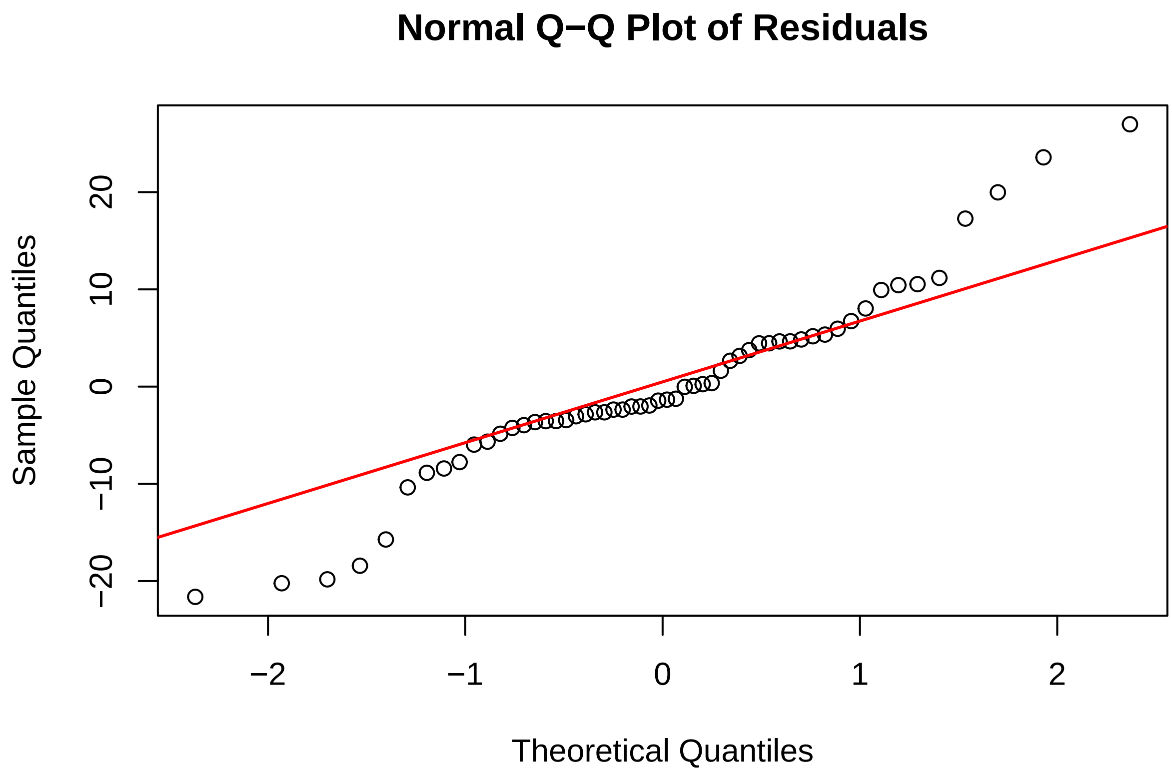


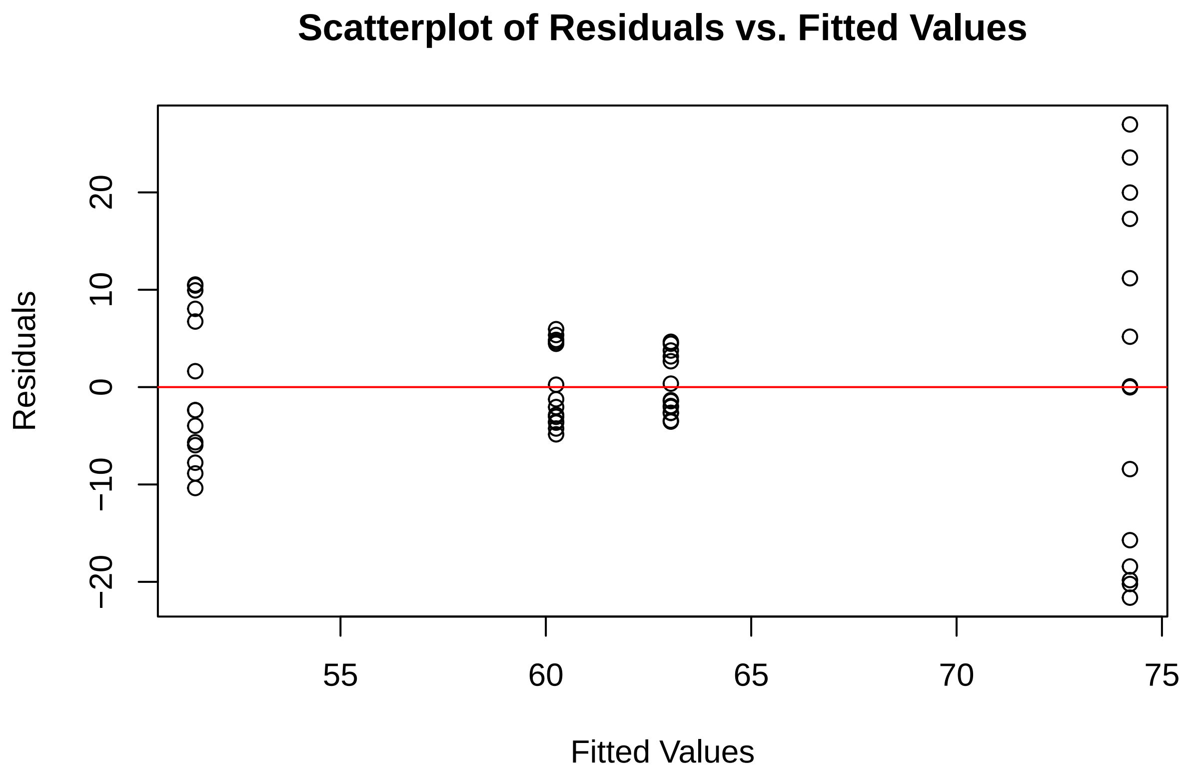
Figure 5: Scatterplot of residuals vs. fitted values after log transform of `hospitalisation.rate` variable

As mentioned in section 2.3.2, there is no real difference in the normality and homoscedasticity assumptions. They are still considered violated.

```
qqnorm(mental.model.1.interaction.4$residuals, main = "Normal Q-Q Plot of
↪  Residuals")
qqline(mental.model.1.interaction.4$residuals, lw = 1.5, col = "red")
```




```
plot(mental.model.1.interaction.4$fitted.values,
 ↪  mental.model.1.interaction.4$residuals,
    xlab = "Fitted Values", ylab = "Residuals", main = "Scatterplot of
     ↪  Residuals vs. Fitted Values")
abline(h = 0, col = "red")
```



The diagnostic plot of model interaction effect (age group * sex) is shown above. Similarly, there is funnel-like shape that violates the assumption of homoscedasticity in the scatterplot of residuals vs. fitted values and large deviation of points deviating from the normal line on the QQ plot, which means the normality assumption is violated. I am assuming that the model is reliable.

2.5 Part (e) - Table of results

Model	Estimate	Standard Errors	p-values
hospitalisation.rate ~ factor(sex) + factor(age.group)	69.23-12.775factor(sex)Male	11.13	0.000346
hospitalisation.rate ~ sex * age.group	74.221-22.757sexMale- 11.179age.group25 + 19.964sexMale:age.group25	9.976	3.14E-06

Final model chosen: `hospitalisation.rate ~ sex * age.group`

Description of the results: The residual standard error is lower. Both models are statistically significant based on the ANOVA F-test. The model with interaction effect between sex and age group accounts for some unexplained effects in the model without interaction effect (`hospitalisation.rate ~ factor(sex) + factor(age.group)`).

For the model equation chosen, there is sufficient evidence to suggest that β_1, β_2 and β_3 are significantly different from 0, p-values < 0.05. There are significantly lower males for hospitalisation rate ($\beta_1 = 22.76$). There is statistically significant relationship for males that are age group 25+, with significantly higher hospitalisation rate of 19.96. This is in contrast to the lower hospitalisation rate (11.18) for those age group 25+ (both males and females) compared to age group 12 - 24.

3 Question 2

We are attempting to predict the daily milk production using candidate linear models.

The question identified that daily milk production (*`MilkProd`* - 24 hour milk production in mL) is the response variable. The potential predictor variables are:

- Baby gender (`BabyGender`)
- Birth weight (`BabyBirthweight`)
- Maternal body mass index (`MaternalBMI`)
- Maternal health (`MaternalHealth`)

Getting a glimpse of the dataset and understanding if the variables are of the correct datatype (such as factor).

```
glimpse(milkp)  # function from tidyverse package
```

```
Rows: 160
Columns: 7
$ MotherID        <dbl> 1001, 1002, 1003, 1004, 1005, 1006, 1007, 1008, 1009, ~
$ MilkProd        <dbl> 222, 370, 167, 272, 286, 309, 353, 358, 364, 366, 376,~
$ BabyGender      <chr> "M", "F", "M", "F", "F", "M", "M", "M", "F", "M", "F",~
$ BirthCentile    <dbl> 70.2, 64.1, 87.9, 71.6, 43.6, 71.6, 21.8, 31.9, 3.8, 1~
$ BabyBirthweight <dbl> 3610, 3400, 3950, 3500, 3160, 3630, 2985, 3120, 2480, ~
$ MaternalBMI     <chr> "overweight", "overweight", "normal", "overweight", "o~
$ MaternalHealth  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

As seen from the output, `MaternalHealth` is not of the correct datatype (`double` - `dbl` instead of `factor` - `fct`). I will have to remember to cast it to the correct datatype later.

There are missing values in the dataset, and we should consider dropping the rows with missing values.

```
sum(is.na(milkp))
```

```
[1] 123
```

3.1 Part (a) - Six candidate models

The candidate models defined in the question is either one or both baby variables and either one or both maternal variables.

Firstly, I create a variable which contains a column vector of all the possible candidate models in strings.

```
models <- c("MilkProd ~ factor(BabyGender)", "MilkProd ~ BabyBirthweight",
    "MilkProd ~ factor(MaternalBMI)", "MilkProd ~ factor(MaternalHealth)",
    "MilkProd ~ factor(BabyGender) + BabyBirthweight", "MilkProd ~
    ↪  factor(MaternalBMI) + factor(MaternalHealth)")
```

Their model equations are as follows:

$$\text{Model 1}: \begin{aligned} MilkProd = \beta_0 + \beta_1(factor(BabyGender)_M) + \\ \epsilon \end{aligned}$$

$$\text{Model 2}: \begin{aligned} MilkProd = \beta_0 + \beta_1(BabyBirthweight) + \\ \epsilon \end{aligned}$$

$$\text{Model 3}: \begin{aligned} MilkProd = \beta_0 + \beta_1(factor(MaternalBMI)_{overweight}) + \\ \epsilon \end{aligned}$$

$$\text{Model 4}: \begin{aligned} MilkProd = \beta_0 + \beta_1(factor(MaternalHealth)_1) + \\ \epsilon \end{aligned}$$

$$\text{Model 5}: \begin{aligned} MilkProd = \beta_0 + \beta_1(factor(BabyGender)_M) + \\ \beta_2(BabyBirthweight) + \epsilon \end{aligned}$$

$$\text{Model 6}: \begin{aligned} MilkProd = \beta_0 + \beta_1(factor(MaternalBMI)_{overweight}) + \\ \beta_2(factor(MaternalHealth)_1) + \epsilon \end{aligned}$$

3.2 Part (b) - Bootstrap method

3.2.1 Prepare our dataset

Firstly, there are *NA* values in the dataset. Remove them. Instead of dropping all rows with NA values, there are columns that we don't require. Drop these columns first, then drop NA values, or else there could be issues with the sample size.

```
milkp.na.rm <- milkp %>%
    select(MilkProd, BabyGender, BabyBirthweight, MaternalBMI,
        MaternalHealth) %>%
    drop_na()
```

There are 40 observations left after removing NA values.

```
glimpse(milkp.na.rm)
```

```
Rows: 155
Columns: 5
$ MilkProd        <dbl> 222, 370, 167, 272, 286, 309, 353, 358, 364, 366, 376,~
$ BabyGender      <chr> "M", "F", "M", "F", "F", "M", "M", "M", "F", "M", "F",~
$ BabyBirthweight <dbl> 3610, 3400, 3950, 3500, 3160, 3630, 2985, 3120, 2480, ~
$ MaternalBMI     <chr> "overweight", "overweight", "normal", "overweight", "o~
$ MaternalHealth  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

3.2.2 Define 3 types of functions

3.2.2.1 Define a function that relates x to y

```
model.fit <- function(x, y) {

    return(lm(y ~ x - 1))
}
```

3.2.2.2 Define a function that calculates the predicted values ($\hat{Y}$)

```
predicted.values <- function(model.fit, x) {
    return(x %*% model.fit$coefficients)
}
```

3.2.2.3 Define a function that calculates the squared residuals ($(Y - \hat{Y})^2$)

```
squared.residuals <- function(Y, Y.hat) {
    return((Y - Y.hat)^2)
}
```

3.2.3 Find out the bootstrap prediction error for this model with 100 bootstrap samples

```r
for (i in 1:length(models)) {
    bootpred.results <- bootpred(model.matrix(as.formula(models[i]),
        data = milkp.na.rm), milkp.na.rm$MilkProd, nboot = 100,
        theta.fit = model.fit, theta.predict = predicted.values,
        err.meas = squared.residuals)

    print(c(bootpred.results[[3]], sqrt(bootpred.results[[3]])))

}
```

```
[1] 41946.5444   204.8086
[1] 37696.2780   194.1553
[1] 40855.7915   202.1282
[1] 33940.333   184.229
[1] 38155.3443   195.3339
[1] 33725.2177   183.6443
```

Each row of the results above indicate each model in the `models` variable. The first value of each row is the MSE and the second value of each row is the RMSE.

3.3 Part (c) - 10-fold cross-validation

3.3.1 Create a variable that contains all the candidate models in strings

I will re-use the formula above `models`.

```r
models
```

```
[1] "MilkProd ~ factor(BabyGender)"
[2] "MilkProd ~ BabyBirthweight"
[3] "MilkProd ~ factor(MaternalBMI)"
[4] "MilkProd ~ factor(MaternalHealth)"
[5] "MilkProd ~ factor(BabyGender) + BabyBirthweight"
[6] "MilkProd ~ factor(MaternalBMI) + factor(MaternalHealth)"
```

Set the number of cross validations.

```r
ncrossval <- 100
```

3.3.2 Create empty matrices to collect PRESS, MSE, and RMSE values

```r
PRESS.mat <- matrix(NA, nrow = 100, ncol = length(models))
MSE.mat <- matrix(NA, nrow = 100, ncol = length(models))
RMSE.mat <- matrix(NA, nrow = 100, ncol = length(models))
```

3.3.3 Create nested for-loops to conduct 10-fold CV for `ncrossval` repetitions

Notice, we are using the functions already defined above.

```r
for (i in 1:ncrossval) {
    for (j in 1:length(models)) {
        milkp.cv.10 <- crossval(model.matrix(as.formula(models[j]),
            data = milkp.na.rm), milkp.na.rm$MilkProd, theta.fit =
             ↪  model.fit,
            theta.predict = predicted.values, ngroup = 10)

        PRESS.mat[i, j] <- sum((milkp.na.rm$MilkProd -
 ↪  milkp.cv.10$cv.fit)^2)

        MSE.mat[i, j] <- PRESS.mat[i, j]/length(milkp.na.rm$MilkProd)

        RMSE.mat[i, j] <- sqrt(MSE.mat[i, j])
    }
}
```

Find the mean of PRESS, MSE, and RMSE values

```r
PRESS.mean.1 <- apply(PRESS.mat, MARGIN = 2, FUN = mean)  # mean PRESS
 ↪  values for each of the candidate models
MSE.mean.1 <- apply(MSE.mat, MARGIN = 2, FUN = mean)  # mean MSE values for
 ↪  each of the candidate models
RMSE.mean.1 <- apply(RMSE.mat, MARGIN = 2, FUN = mean)  # mean RMSE values
 ↪  for each of the candidate models

c(PRESS.mean.1, MSE.mean.1, RMSE.mean.1)
```

```
 [1] 6546314.6157 5870002.6870 6374636.7013 5258344.0533 5946356.4423
 [6] 5248068.2138   42234.2878   37870.9851   41126.6884   33924.8003
[11]   38363.5900   33858.5046     205.5082     194.6026     202.7952
[16]     184.1846     195.8627     184.0030
```

For the results above, the first row (first six values) represents the PRESS values, second row represents the MSE values and the third row represents the RMSE values. Each column represents each of the candidate models.

Model Number	Bootstrap		10-fold cross-validation	
	MSE	RMSE	MSE	RMSE
1	41910.4386	204.7204	42234.2878	205.5082
2	37638.1283	194.0055	37870.9851	194.6026
3	41130.7746	202.8072	41126.6884	202.7952
4	33890.0695	184.0926	33924.8003	184.1846
5	38378.0866	195.9033	38363.59	195.8627
6	33774.9362	183.7796	33858.5046	184.003

3.4 Part (d) - Table of estimators

The model number are numbered according to the order of the models defined in the variable `models` above (see section 3.3.1).

3.5 Part (e) - Best model for prediction

The sixth model, `MilkProd ~ factor(MaternalBMI) + factor(MaternalHealth)` is the best model for prediction purposes. The MSE and RMSE scores are the lowest for both bootstrap .632+ method and 10-fold cross-validation methods.

Model 4, `MilkProd ~ factor(MaternalHealth)` can also be considered as the best model for prediction. Although Model 6 has the lowest MSE and RMSE scores, Model 4 only falls slightly behind, being narrowly the second lowest MSE and RMSE scores. The one standard error rule can be utilised and Model 4 can be selected, if its lesser number of predictors lies within one standard error of the MSE. Model 4 may be considered if it is costly to obtain data of maternal BMI.

```
summary(lm(as.formula(models[4]), data = milkp.na.rm))
```

```
Call:
lm(formula = as.formula(models[4]), data = milkp.na.rm)

Residuals:
   Min     1Q Median     3Q    Max
-565.7 -105.7  -19.7  121.8  522.9

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               571.08      28.85   19.80  < 2e-16 ***
factor(MaternalHealth)1   208.63      33.49    6.23 4.31e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 182.4 on 153 degrees of freedom
Multiple R-squared:  0.2023,    Adjusted R-squared:  0.1971
F-statistic: 38.81 on 1 and 153 DF,  p-value: 4.312e-09
```

```r
summary(lm(as.formula(models[6]), data = milkp.na.rm))
```

```
Call:
lm(formula = as.formula(models[6]), data = milkp.na.rm)

Residuals:
    Min      1Q  Median      3Q     Max
-536.34 -121.40  -10.34  117.03  497.95

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     596.05      32.68  18.236  < 2e-16 ***
factor(MaternalBMI)overweight   -47.56      29.79  -1.597    0.112
factor(MaternalHealth)1         201.86      33.59   6.009 1.32e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 181.5 on 152 degrees of freedom
Multiple R-squared:  0.2155,    Adjusted R-squared:  0.2052
F-statistic: 20.88 on 2 and 152 DF,  p-value: 9.755e-09
```

```r
# knitr::purl('Assignment_2.Rmd', documentation = 0) #
# generate R script
```