Applied Analytics Assignment 3

Lynne Joanne Mercer S3613002

16/10/2020

```
library(tibble)
library(ggplot2)
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```
library(readxl)
#Imported  data set
Assignment_3_WNW_Data <- read.csv("C:/Users/mercer/Documents/RMIT/APPLIED ANALYTICS/Assignment 3/Assign
ment_3_ WNW_ Data.csv")
```

#Introduction #Summarise, visualise and analysis of data #Assignment_3_WNW_Data is the whole data set #Data meets tidy principles. #no N/A missing values,no white spaces, each variable has its own column and each observation has its own row. # Original data set consists of 8 Columns or variables and 1000 observations or rows. #Variables consist of Date, gender, age, social metric, time since sign up, demographic, group and hours watched. #Date variable was changed to include the year 2020 #The Group data set is spilt by A (control group) and B (customers unknowingly using new recommendation engine) #Data was provided by WNW executives for analysis of newly launched algorithm on 17th July 12.01am, to assess if the new recommendations algorithm should be rolled out to all users. #The brief is to understand if the roll out of the new algorithm to Group B should be rolled out to all users.

```
#summary and descriptions  of the whole data set
summary(Assignment_3_WNW_Data)
```

```
##          date     gender       age        social_metric    time_since_signup
##  12/07/2020: 33   F:429   Min.   :18.00   Min.   : 0.000   Min.   : 0.00
##  16/07/2020: 33   M:571   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 5.70
##  20/07/2020: 33           Median :36.00   Median : 5.000   Median :11.80
##  24/07/2020: 33           Mean   :36.49   Mean   : 4.911   Mean   :11.97
##  28/07/2020: 33           3rd Qu.:46.00   3rd Qu.: 8.000   3rd Qu.:18.70
##  31/07/2020: 33           Max.   :55.00   Max.   :10.000   Max.   :24.00
##  (Other)   :802
##   demographic    group   hours_watched
##  Min.   :1.000   A:880   Min.   :0.500
##  1st Qu.:2.000   B:120   1st Qu.:3.530
##  Median :3.000           Median :4.415
##  Mean   :2.603           Mean   :4.393
##  3rd Qu.:4.000           3rd Qu.:5.322
##  Max.   :4.000           Max.   :8.300
##
```

```
describe(Assignment_3_WNW_Data)
```

```
## Assignment_3_WNW_Data
##
##  8  Variables      1000  Observations
## --------------------------------------------------------------------------------
## date
##        n  missing distinct
##     1000        0       31
##
## lowest : 1/07/2020  10/07/2020 11/07/2020 12/07/2020 13/07/2020
## highest: 5/07/2020  6/07/2020  7/07/2020  8/07/2020  9/07/2020
## --------------------------------------------------------------------------------
## gender
##        n  missing distinct
##     1000        0        2
##
## Value          F     M
## Frequency    429   571
## Proportion 0.429 0.571
## --------------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       38    0.999    36.49    12.34       20       22
##      .25      .50      .75      .90      .95
##       28       36       46       52       53
##
## lowest : 18 19 20 21 22, highest: 51 52 53 54 55
## --------------------------------------------------------------------------------
## social_metric
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       11    0.991    4.911    3.442        0        1
##      .25      .50      .75      .90      .95
##        2        5        8        9       10
##
## lowest :  0  1  2  3  4, highest:  6  7  8  9 10
##
## Value          0     1     2     3     4     5     6     7     8     9    10
## Frequency     59   109   103    96   110    85    89    86   113    97    53
## Proportion 0.059 0.109 0.103 0.096 0.110 0.085 0.089 0.086 0.113 0.097 0.053
## --------------------------------------------------------------------------------
## time_since_signup
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0      239        1    11.97     8.35     0.70     1.70
##      .25      .50      .75      .90      .95
##     5.70    11.80    18.70    21.81    22.90
##
## lowest :  0.0  0.1  0.2  0.3  0.4, highest: 23.6 23.7 23.8 23.9 24.0
## --------------------------------------------------------------------------------
## demographic
##        n  missing distinct     Info     Mean      Gmd
##     1000        0        4    0.933    2.603    1.262
##
## Value          1     2     3     4
## Frequency    216   268   213   303
## Proportion 0.216 0.268 0.213 0.303
## --------------------------------------------------------------------------------
## group
##        n  missing distinct
##     1000        0        2
##
## Value         A    B
## Frequency   880  120
## Proportion 0.88 0.12
## --------------------------------------------------------------------------------
## hours_watched
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0      501        1    4.393    1.512    2.163    2.609
##      .25      .50      .75      .90      .95
##    3.530    4.415    5.322    6.120    6.530
##
## lowest : 0.50 0.79 0.80 0.95 1.03, highest: 7.61 7.67 7.93 8.01 8.30
## --------------------------------------------------------------------------------
```

#In order to fully understand the data we need to split the data set into 2 groups Group A and Group B where the new algorithm was applied from the 18th July.

```
#Imported Split data  set into groups A & B
Group_A <- read_xlsx('C:/Users/mercer/Documents/RMIT/APPLIED ANALYTICS/Assignment 3/Group A.xlsx')
Group_B <- read_xlsx('C:/Users/mercer/Documents/RMIT/APPLIED ANALYTICS/Assignment 3/Group B.xlsx')
summary(Group_A)
```

```
##       date                        gender               age
##  Min.   :2020-07-01 00:00:00   Length:880         Min.   :18.00
##  1st Qu.:2020-07-07 00:00:00   Class :character   1st Qu.:27.00
##  Median :2020-07-14 00:00:00   Mode  :character   Median :36.00
##  Mean   :2020-07-14 20:45:16                      Mean   :36.15
##  3rd Qu.:2020-07-22 00:00:00                      3rd Qu.:45.00
##  Max.   :2020-07-31 00:00:00                      Max.   :55.00
##  social_metric    time_since_signup  demographic       group
##  Min.   : 0.000   Min.   : 0.000    Min.   :1.000   Length:880
##  1st Qu.: 2.000   1st Qu.: 5.875    1st Qu.:2.000   Class :character
##  Median : 5.000   Median :11.900    Median :3.000   Mode  :character
##  Mean   : 4.868   Mean   :12.034    Mean   :2.548
##  3rd Qu.: 8.000   3rd Qu.:18.700    3rd Qu.:4.000
##  Max.   :10.000   Max.   :24.000    Max.   :4.000
##  hours_watched
##  Min.   :0.500
##  1st Qu.:3.487
##  Median :4.355
##  Mean   :4.336
##  3rd Qu.:5.250
##  Max.   :8.300
```

```
describe(Group_A)
```

```
## Group_A
##
##  8  Variables      880  Observations
## --------------------------------------------------------------------------------
## date
##          n    missing   distinct       Info       Mean        Gmd        .05
##        880          0         31      0.999 2020-07-15     877515 2020-07-02
##        .10        .25        .50        .75        .90        .95
## 2020-07-03 2020-07-07 2020-07-14 2020-07-22 2020-07-28 2020-07-30
##
## lowest : 2020-07-01 2020-07-02 2020-07-03 2020-07-04 2020-07-05
## highest: 2020-07-27 2020-07-28 2020-07-29 2020-07-30 2020-07-31
## --------------------------------------------------------------------------------
## gender
##        n  missing distinct
##      880        0        2
##
## Value          F     M
## Frequency    400   480
## Proportion 0.455 0.545
## --------------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      880        0       38    0.999    36.15    12.38       20       22
##      .25      .50      .75      .90      .95
##       27       36       45       52       53
##
## lowest : 18 19 20 21 22, highest: 51 52 53 54 55
## --------------------------------------------------------------------------------
## social_metric
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      880        0       11    0.991    4.868    3.462        0        1
##      .25      .50      .75      .90      .95
##        2        5        8        9       10
##
## lowest :  0  1  2  3  4, highest:  6  7  8  9 10
##
## Value          0     1     2     3     4     5     6     7     8     9    10
## Frequency     56    98    90    85    96    76    73    76    98    87    45
## Proportion 0.064 0.111 0.102 0.097 0.109 0.086 0.083 0.086 0.111 0.099 0.051
## --------------------------------------------------------------------------------
## time_since_signup
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      880        0      238        1    12.03    8.342    0.700    1.800
##      .25      .50      .75      .90      .95
##    5.875   11.900   18.700   21.900   23.000
##
## lowest :  0.0  0.1  0.2  0.3  0.4, highest: 23.6 23.7 23.8 23.9 24.0
## --------------------------------------------------------------------------------
## demographic
##        n  missing distinct     Info     Mean      Gmd
##      880        0        4    0.936    2.548    1.257
##
## Value          1     2     3     4
## Frequency    203   236   197   244
## Proportion 0.231 0.268 0.224 0.277
## --------------------------------------------------------------------------------
## group
##        n  missing distinct    value
##      880        0        1        A
##
```

```
## Value        A
## Frequency  880
## Proportion   1
## --------------------------------------------------------------------------------
## hours_watched 
##        n  missing distinct     Info     Mean      Gmd      .05      .10 
##      880        0      423        1    4.336      1.5    2.079    2.540 
##      .25      .50      .75      .90      .95 
##    3.488    4.355    5.250    6.050    6.490 
## 
## lowest : 0.50 0.79 0.80 0.95 1.03, highest: 7.45 7.52 7.67 8.01 8.30
## --------------------------------------------------------------------------------
```

```
#Group A  is the control group
```

```
summary(Group_B)
```

```
##       date                          gender                age       
##  Min.   :2020-07-18 00:00:00   Length:120         Min.   :18.00  
##  1st Qu.:2020-07-21 00:00:00   Class :character   1st Qu.:31.00  
##  Median :2020-07-24 00:00:00   Mode  :character   Median :39.50  
##  Mean   :2020-07-24 10:48:00                      Mean   :38.94  
##  3rd Qu.:2020-07-28 00:00:00                      3rd Qu.:47.00  
##  Max.   :2020-07-31 00:00:00                      Max.   :55.00  
##  social_metric    time_since_signup  demographic       group          
##  Min.   : 0.000   Min.   : 0.00     Min.   :1.000   Length:120        
##  1st Qu.: 3.000   1st Qu.: 5.15     1st Qu.:2.000   Class :character  
##  Median : 5.000   Median :11.35     Median :3.000   Mode  :character  
##  Mean   : 5.225   Mean   :11.47     Mean   :3.008                     
##  3rd Qu.: 8.000   3rd Qu.:18.98     3rd Qu.:4.000                     
##  Max.   :10.000   Max.   :23.70     Max.   :4.000                     
##  hours_watched  
##  Min.   :1.525  
##  1st Qu.:3.939  
##  Median :4.860  
##  Mean   :4.811  
##  3rd Qu.:5.723  
##  Max.   :7.930
```

```
describe(Group_B)
```

```
## Group_B
##
##  8  Variables      120  Observations
## --------------------------------------------------------------------------------
## date
##          n    missing   distinct       Info       Mean        Gmd        .05
##        120          0         14      0.994 2020-07-24     404652 2020-07-19
##        .10        .25        .50        .75        .90        .95
## 2020-07-19 2020-07-21 2020-07-24 2020-07-28 2020-07-30 2020-07-31
##
## lowest : 2020-07-18 2020-07-19 2020-07-20 2020-07-21 2020-07-22
## highest: 2020-07-27 2020-07-28 2020-07-29 2020-07-30 2020-07-31
## --------------------------------------------------------------------------------
## gender
##        n  missing distinct
##      120        0        2
##
## Value          F     M
## Frequency     29    91
## Proportion 0.242 0.758
## --------------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      120        0       37    0.999    38.94    11.62     21.0     25.0
##      .25      .50      .75      .90      .95
##     31.0     39.5     47.0     52.0     53.0
##
## lowest : 18 19 20 21 23, highest: 51 52 53 54 55
## --------------------------------------------------------------------------------
## social_metric
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      120        0       11    0.989    5.225    3.293        1        1
##      .25      .50      .75      .90      .95
##        3        5        8        9       10
##
## lowest :  0  1  2  3  4, highest:  6  7  8  9 10
##
## Value          0     1     2     3     4     5     6     7     8     9    10
## Frequency      3    11    13    11    14     9    16    10    15    10     8
## Proportion 0.025 0.092 0.108 0.092 0.117 0.075 0.133 0.083 0.125 0.083 0.067
## --------------------------------------------------------------------------------
## time_since_signup
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      120        0       92        1    11.47    8.437     0.70     1.58
##      .25      .50      .75      .90      .95
##     5.15    11.35    18.97    20.71    22.30
##
## lowest :  0.0  0.1  0.3  0.6  0.7, highest: 22.3 22.4 22.7 23.0 23.7
## --------------------------------------------------------------------------------
## demographic
##        n  missing distinct     Info     Mean      Gmd
##      120        0        4    0.859    3.008    1.172
##
## Value          1     2     3     4
## Frequency     13    32    16    59
## Proportion 0.108 0.267 0.133 0.492
## --------------------------------------------------------------------------------
## group
##        n  missing distinct    value
##      120        0        1        B
##
```

```
## Value         B
## Frequency  120
## Proportion   1
## ---------------------------------------------------------------------------------
## hours_watched
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      120        0      116        1    4.811    1.516    2.629    2.958
##      .25      .50      .75      .90      .95
##    3.939    4.860    5.723    6.377    6.881
##
## lowest : 1.525 1.590 2.125 2.165 2.515, highest: 6.920 7.090 7.220 7.610 7.930
## ---------------------------------------------------------------------------------
```

```
#Group B is the Test group that was the measure of the effectiveness of the change to the recommendatio
n engine
```

#upon review of the data sets descriptions and summaries it is identified that the demographic is a key variable and represents #1 F age group 18-35 #2 M age group 18-35 #3 F Age group 36-55 #4 M Age group 36-55 #Comparisons of Group A and B reveal the size of each group are considerably different 120 in group B compared to 880 in Group A, 7.3 times larger than B.The proportions of male and female in each group are very different,with a higher proportion of males in Group B.Not only is there a higher proportion of males but a much higher proportion of older males from demographic 4 as per the summary and descriptions for each group.

#Group B F M #Frequency 29 91 #Proportion 0.242 0.758 #Percentage 24.2% 75.8%

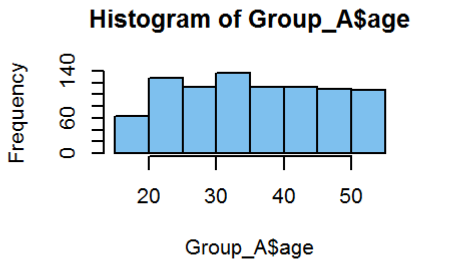
#Group A F M #Frequency 400 480 #Proportion 0.455 0.545 #Percentage 45.5% 54.5%

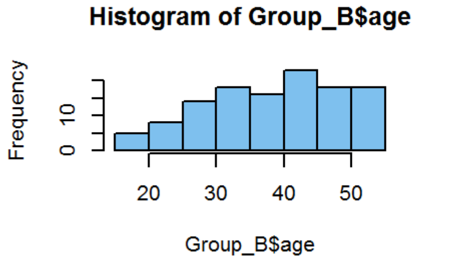
#Group B 1 2 3 4 #Frequency 13 32 16 59 #Proportion 0.108 0.267 0.133 0.492 #Percentage 10.8% 26.7% 13.3% 49.2%

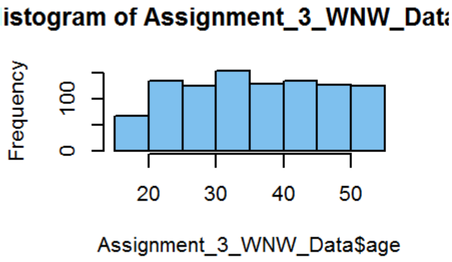
#Group A 1 2 3 4 #Frequency 203 236 197 244 #Proportion 0.231 0.268 0.224 0.277 #Percentage 23.1% 26.8% 22.4% 27.7%

#The treatment group B is significantly biased in proportion of the number of people, this alone would not be an issue however the demographics of the group are not a true sample size the number of males to females plus the number of older males in demographic 4. I would not try to correct the bias as this will impact all the results. #The Key variables are age, demographic & hours watched

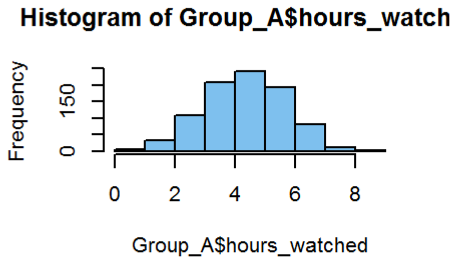
```
#Visualisation of variables for comparisons
par(mfcol=c(3,3))
hist(Group_A$age ,col= 'skyblue2')
hist(Group_B$age ,col= 'skyblue2')
hist(Assignment_3_WNW_Data$age, col = 'skyblue2')
par(mfcol=c(3,3))
```

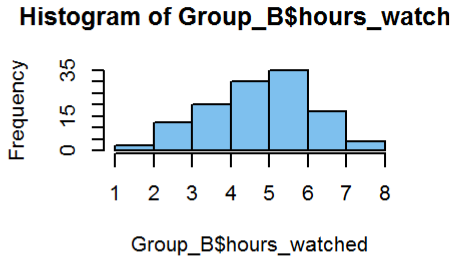


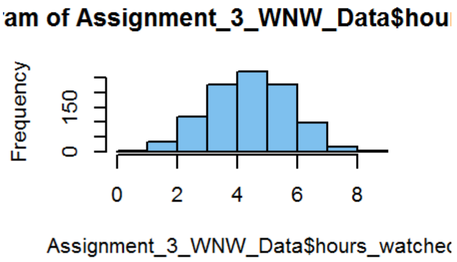




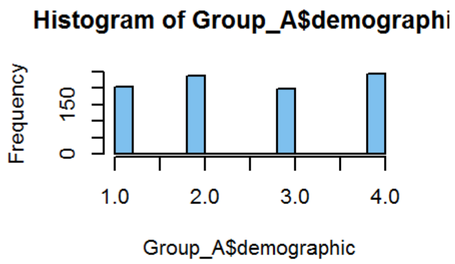
```
hist(Group_A$hours_watched ,col= 'skyblue2')
hist(Group_B$hours_watched ,col= 'skyblue2')
hist(Assignment_3_WNW_Data$hours_watched, col = 'skyblue2')
par(mfcol=c(3,3))
```

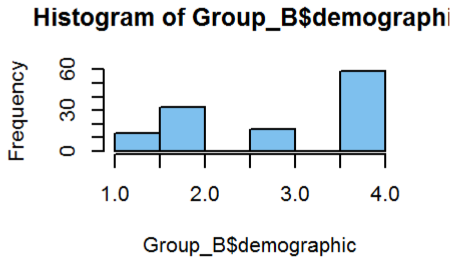


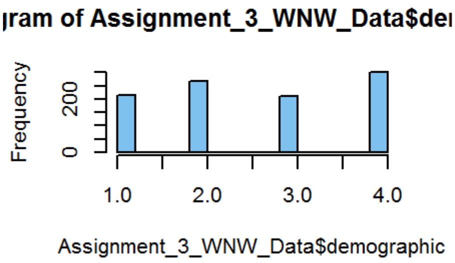




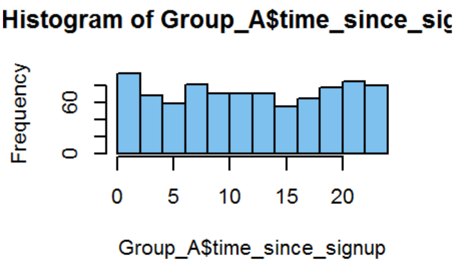
```
hist(Group_A$demographic  ,col= 'skyblue2')
hist(Group_B$demographic  ,col= 'skyblue2')
hist(Assignment_3_WNW_Data$demographic , col = 'skyblue2')
par(mfcol=c(3,3))
```

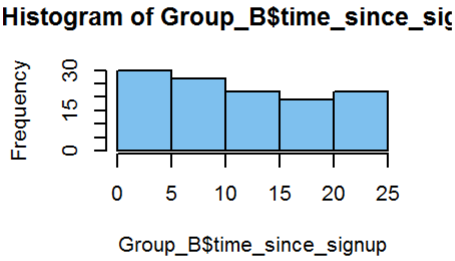


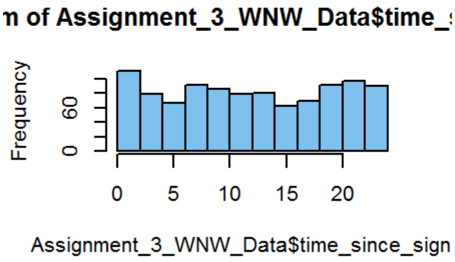




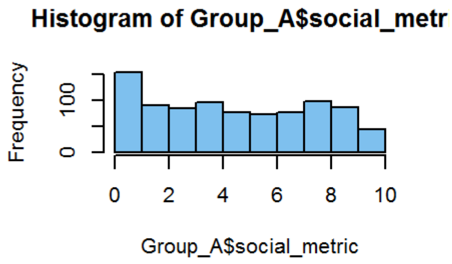
```
hist(Group_A$time_since_signup  ,col= 'skyblue2')
hist(Group_B$time_since_signup  ,col= 'skyblue2')
hist(Assignment_3_WNW_Data$time_since_signup , col = 'skyblue2')
par(mfcol=c(3,3))
```

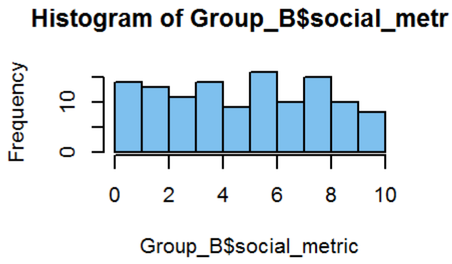




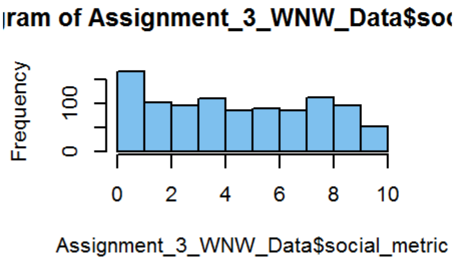


```
hist(Group_A$social_metric  ,col= 'skyblue2')
hist(Group_B$social_metric  ,col= 'skyblue2')
hist(Assignment_3_WNW_Data$social_metric , col = 'skyblue2')
```





#Group B



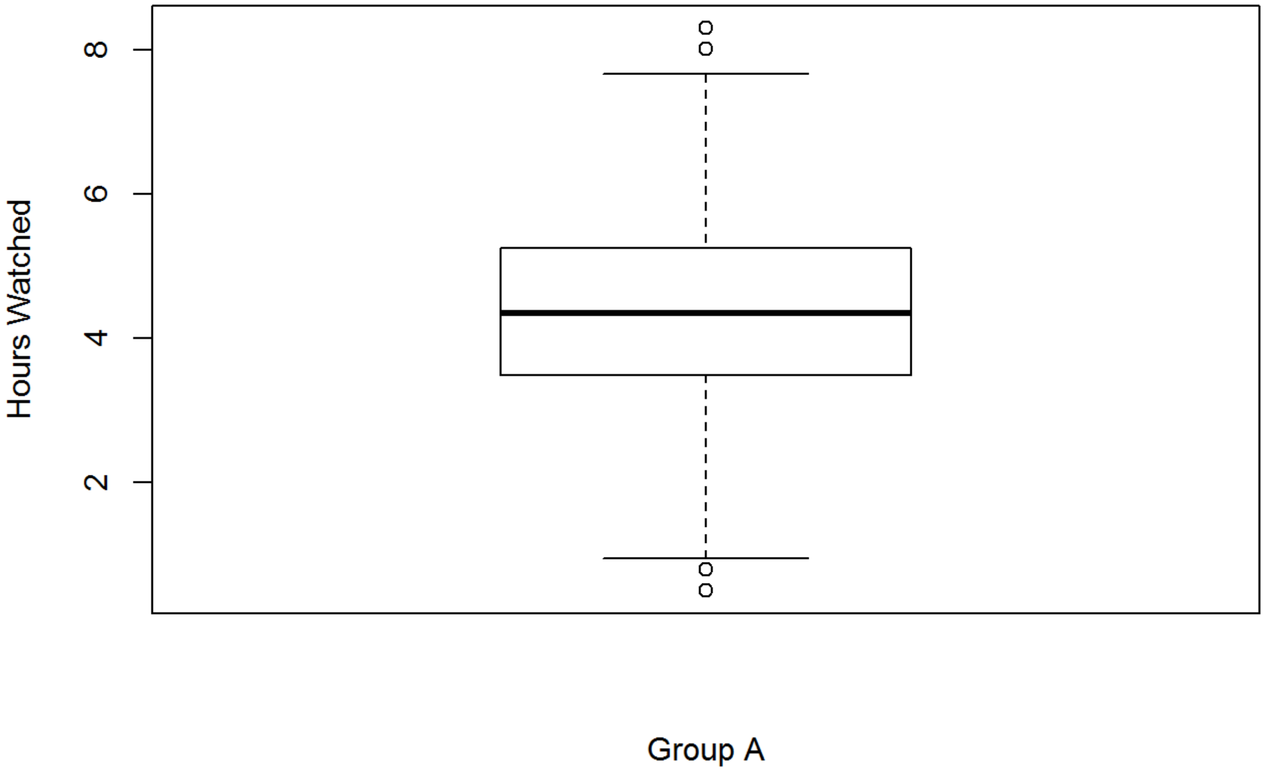
age data is left skewed compared to Group A and the whole data set. Hours watched is also left skewed for Group B # Group B has a disproportionate number of male in the demographic 4 which is male 36-55 years old.Group A and the whole data set have a somewhat similar proprtion of males / females and age groups #There appears to be no correllation for Social metric to the rest of the data

#Outliers are identified from the boxplots below and in the stats, I would not remove any of these as this will impact future results.It’s important we are aware of the outliers and how these impact the results of the mean.

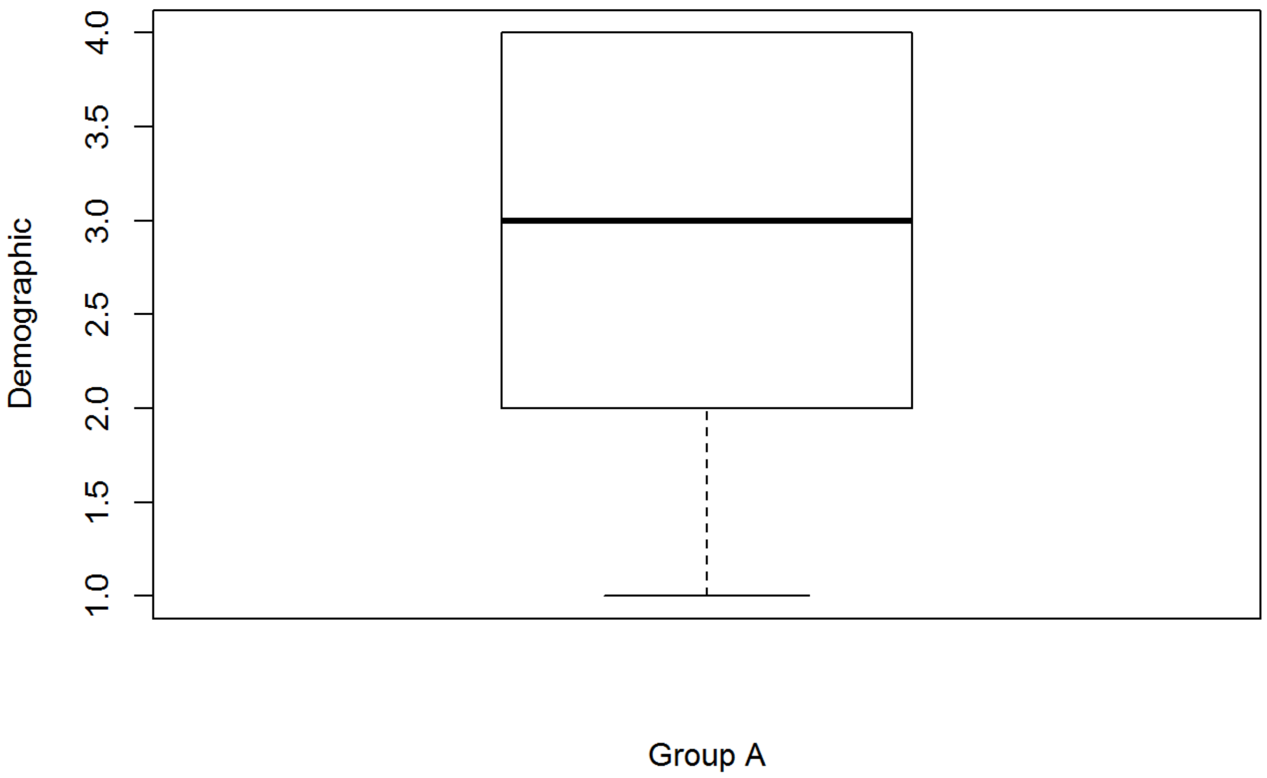
```
boxplot.stats(Group_A$hours_watched)
```

```
## $stats
## [1] 0.950 3.485 4.355 5.250 7.670
##
## $n
## [1] 880
##
## $conf
## [1] 4.260993 4.449007
##
## $out
## [1] 8.30 8.01 0.50 0.50 0.80 0.79
```

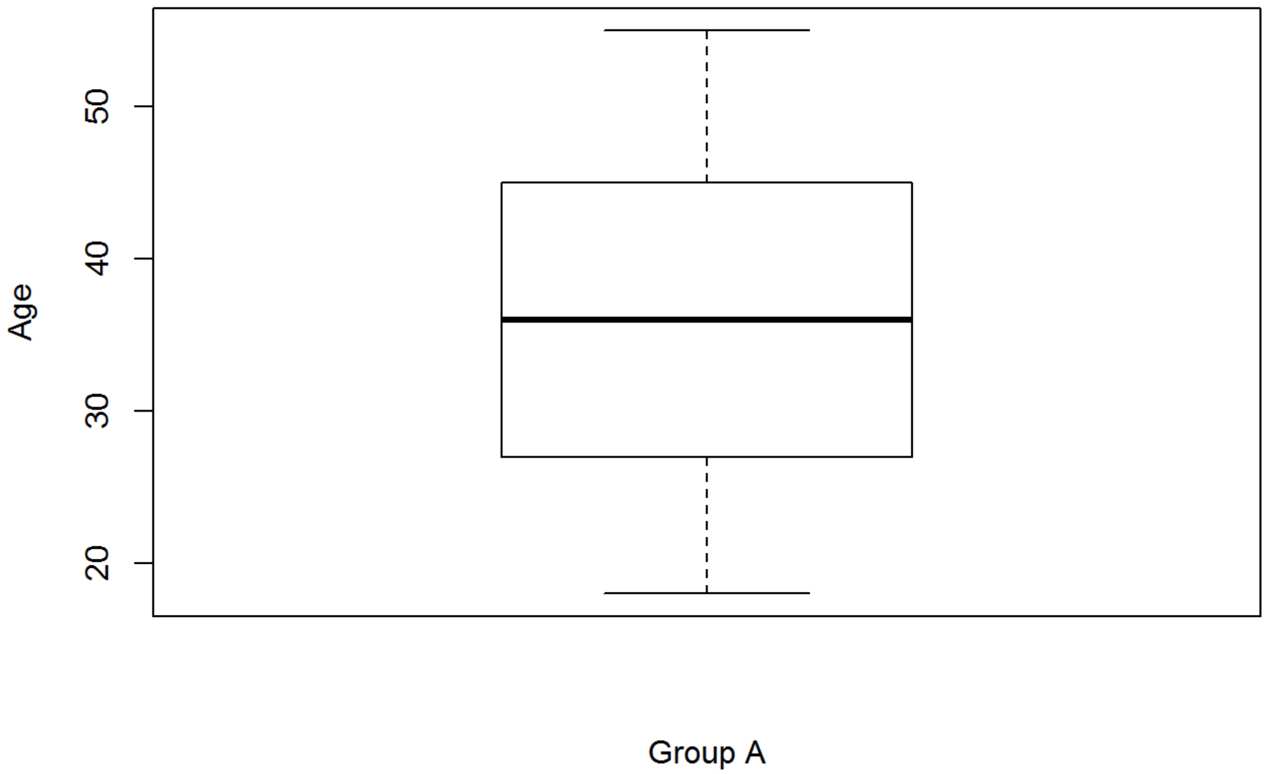
```
boxplot(Group_A$hours_watched,xlab="Group A", ylab="Hours Watched")
```



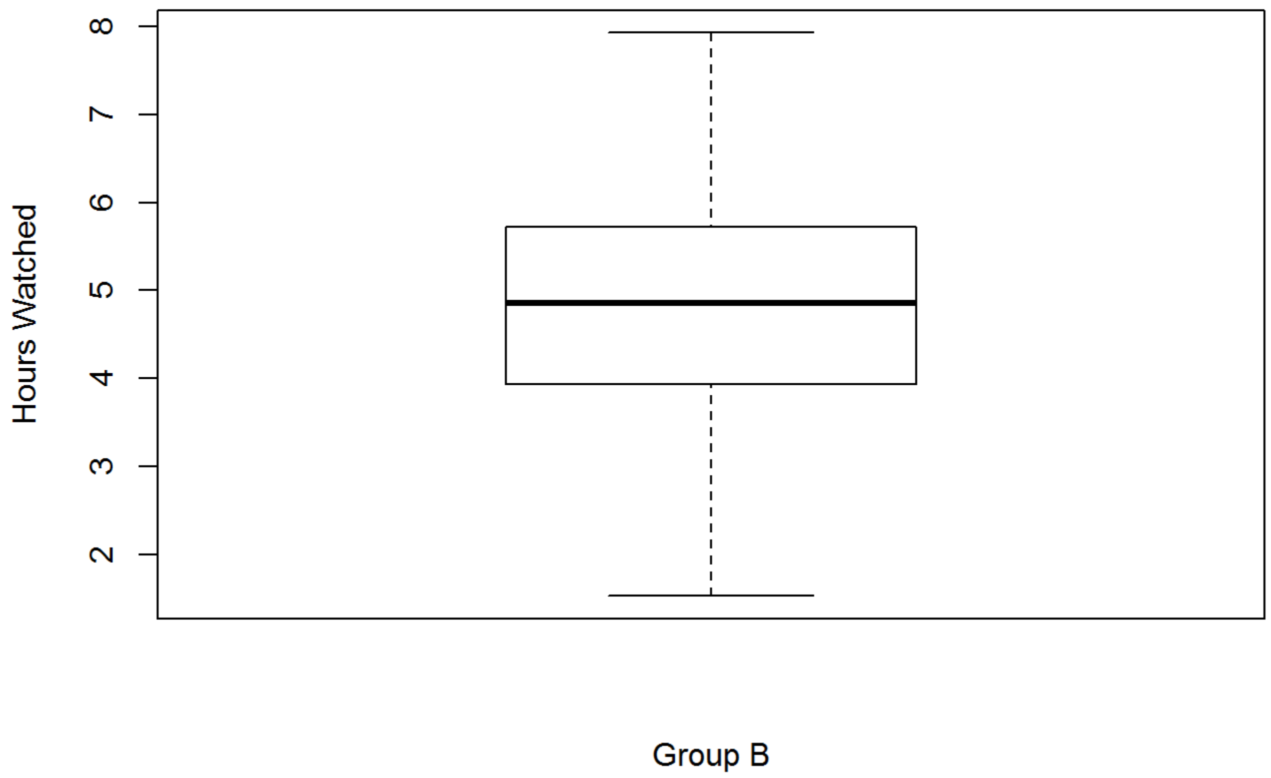
```
boxplot(Group_A$demographic,xlab="Group A", ylab="Demographic")
```



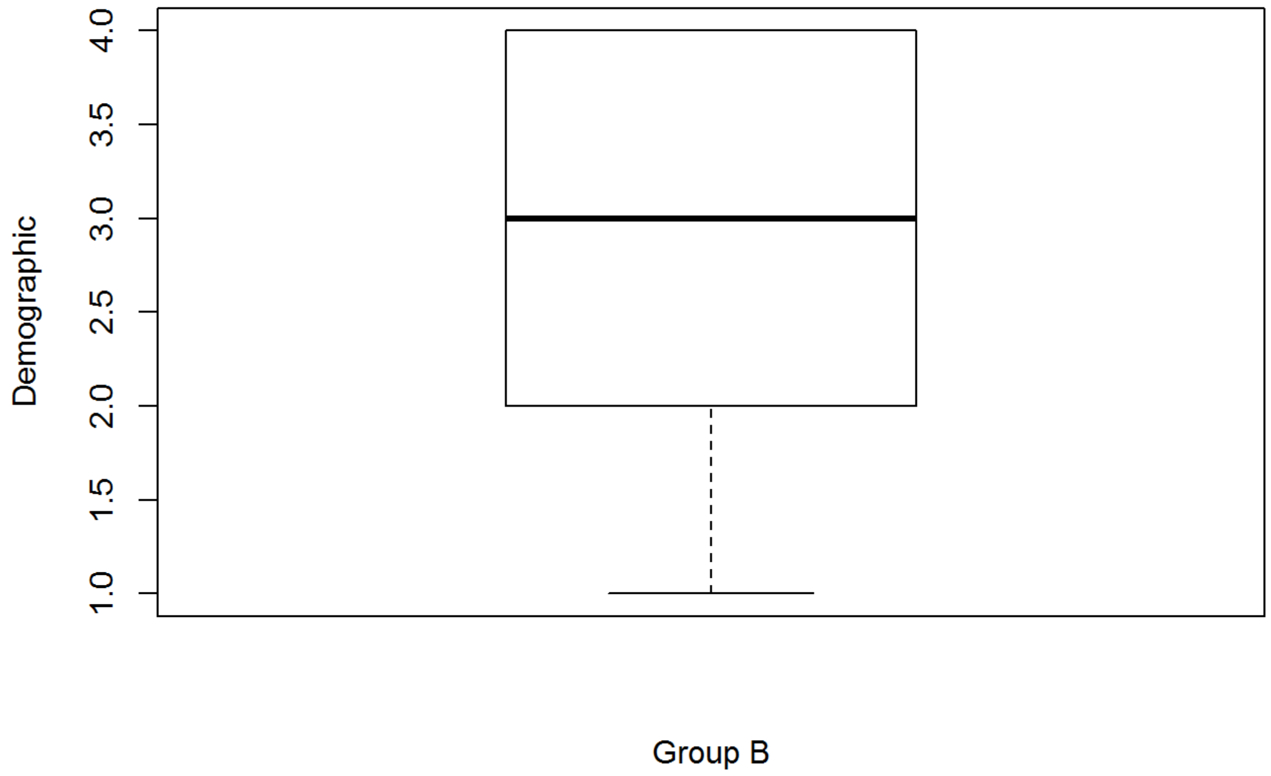
```
boxplot(Group_A$age,xlab="Group A", ylab="Age")
```



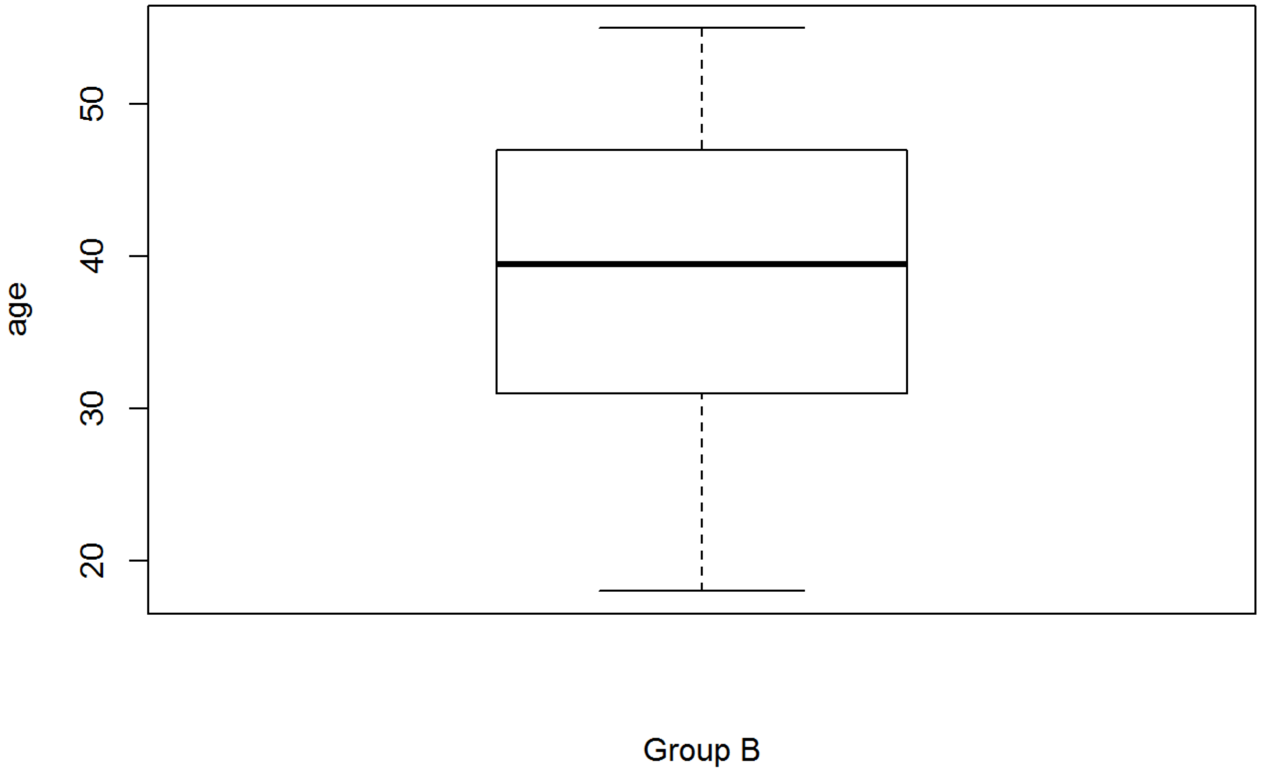
```
boxplot(Group_B$hours_watched,xlab="Group B", ylab="Hours Watched")
```



```
boxplot(Group_B$demographic,xlab="Group B", ylab="Demographic")
```



```
boxplot(Group_B$age,xlab="Group B", ylab="age")
```



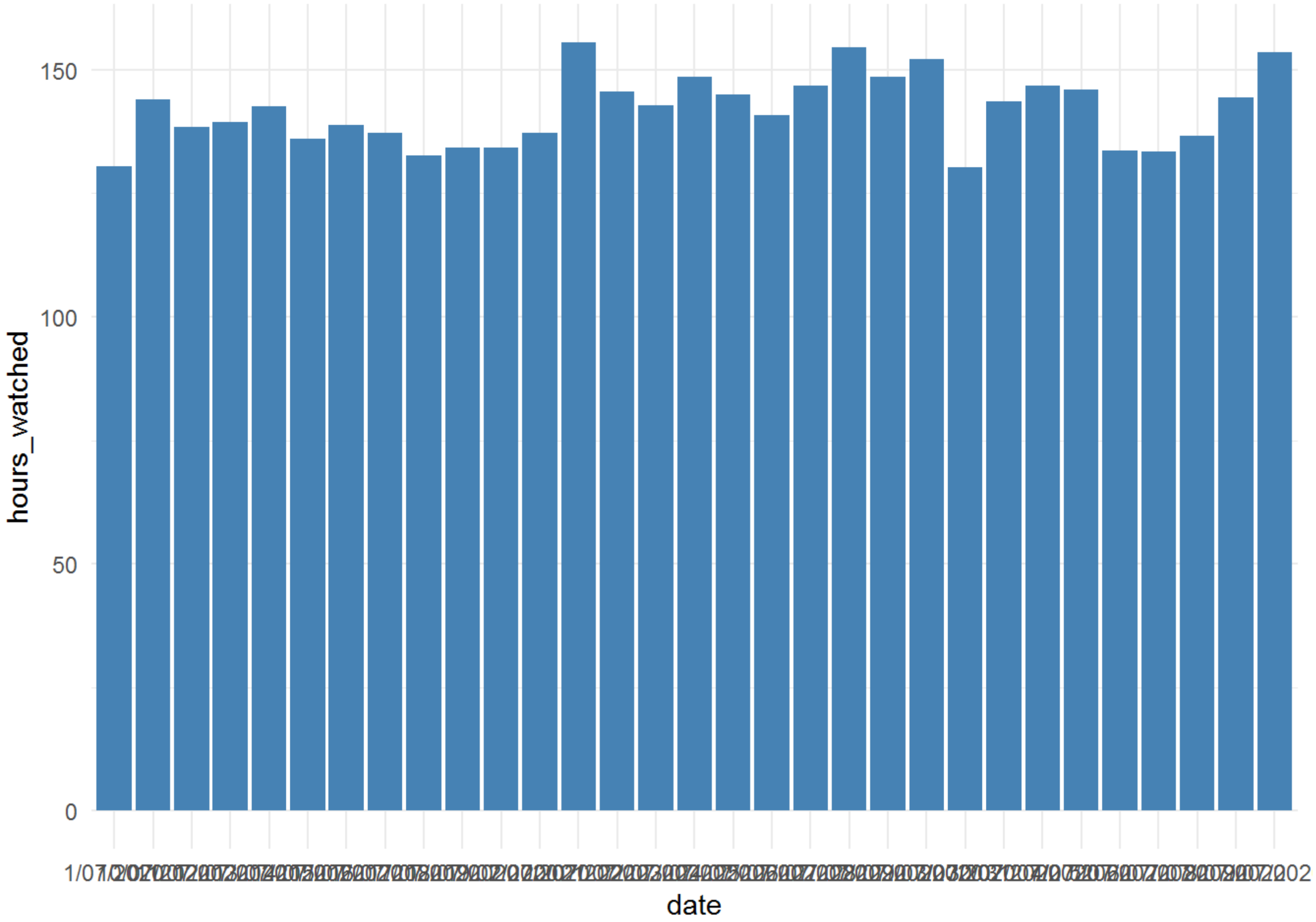
```
boxplot.stats(Group_B$demographic)
```

```
## $stats
## [1] 1 2 3 4 4
##
## $n
## [1] 120
##
## $conf
## [1] 2.711533 3.288467
##
## $out
## numeric(0)
```

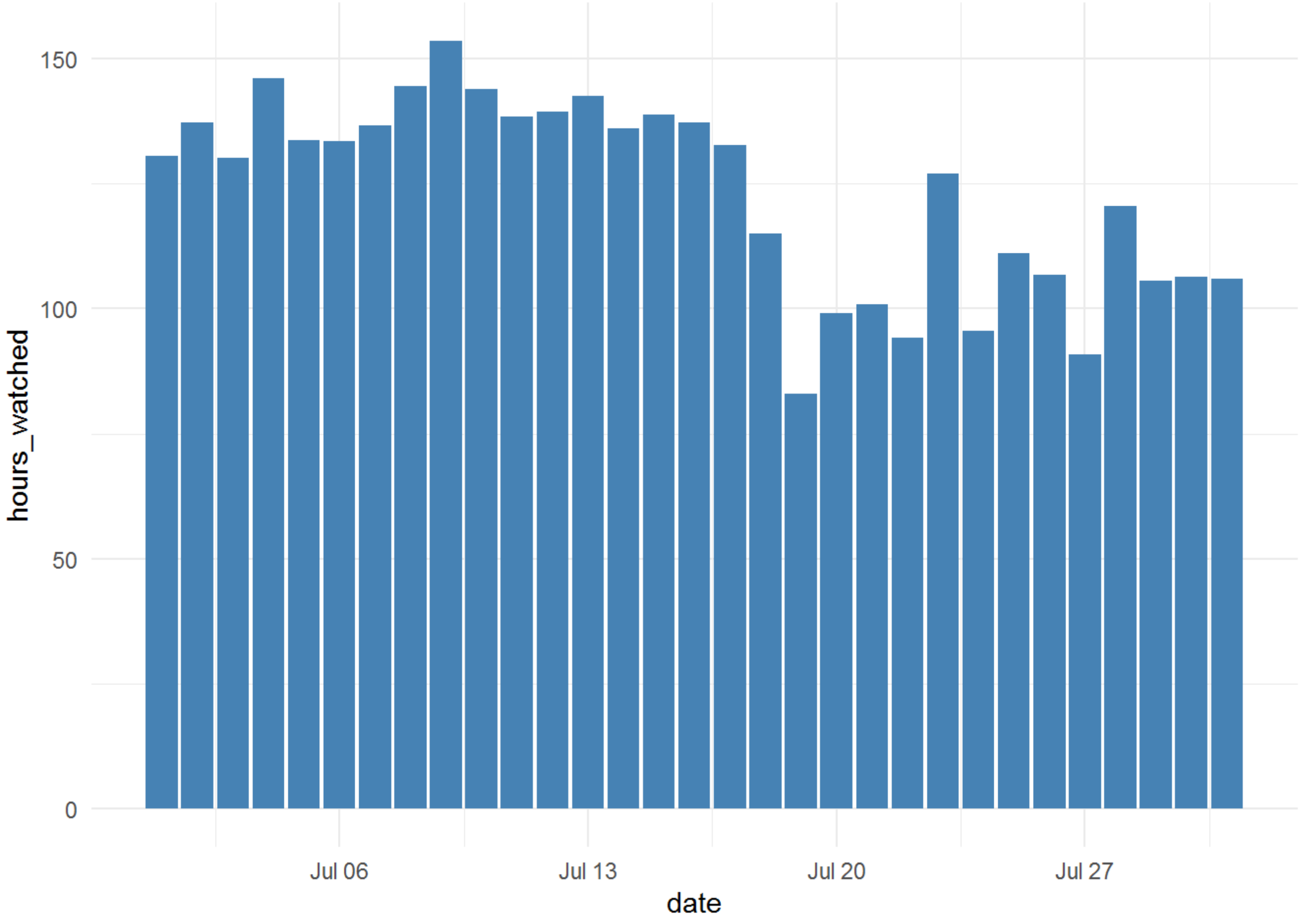
#demographic in Group B clearly identifies as significant in the upper age group of demographic 4

#Data Visualisations of hrs watched and the date of all data sets to identify if there is an increase in hours watched across the 3 data sets post 17th July.

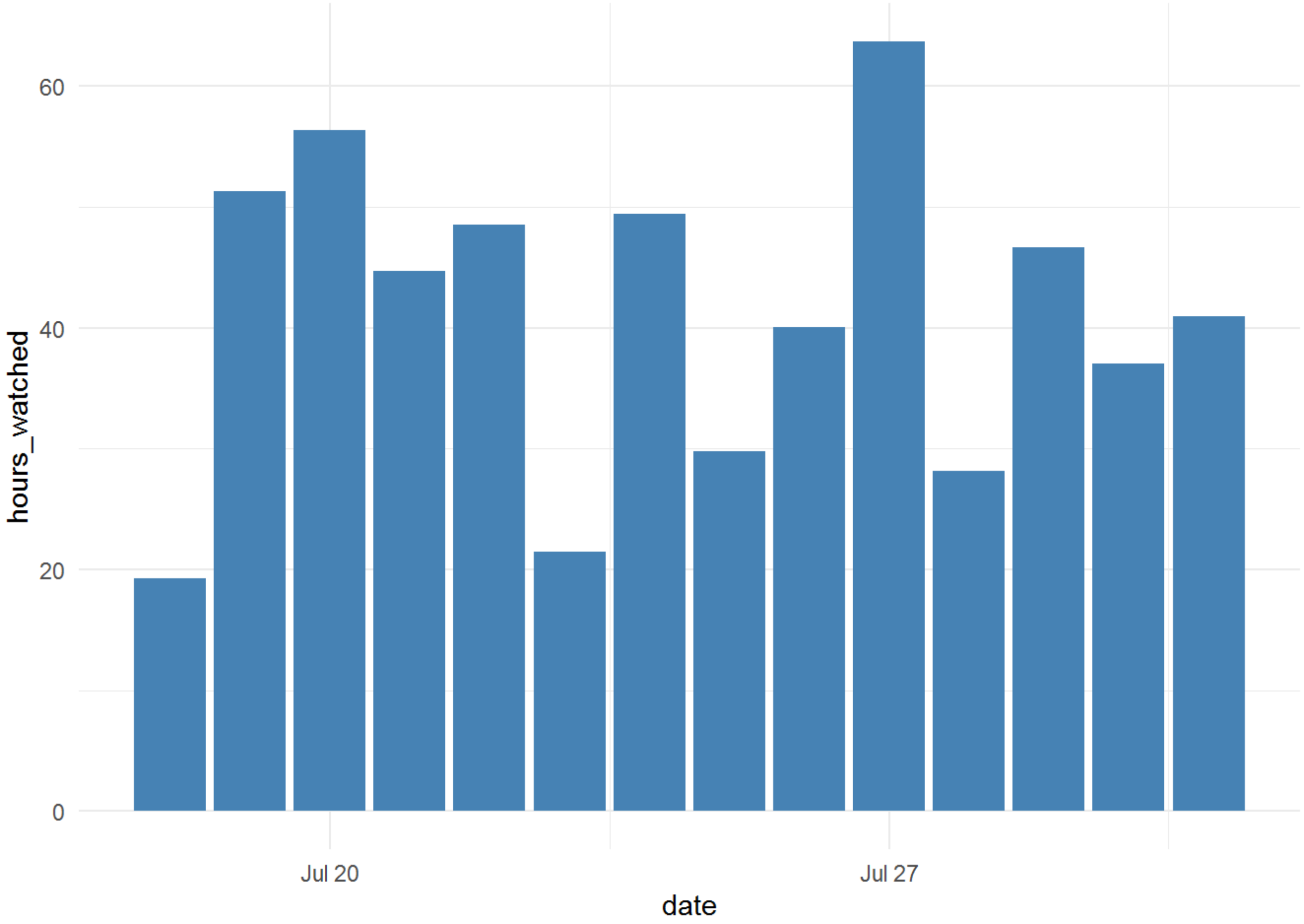
```
par(mfcol=c(3,3))
p<-ggplot(data=Assignment_3_WNW_Data, aes(x=date, y=hours_watched)) +
  geom_bar(stat="identity", fill="steelblue")+
  theme_minimal()
p
```



```
p<-ggplot(data=Group_A, aes(x=date, y=hours_watched)) +
  geom_bar(stat="identity", fill="steelblue")+
  theme_minimal()
p
```



```
p<-ggplot(data=Group_B, aes(x=date, y=hours_watched)) +
  geom_bar(stat="identity", fill="steelblue")+
  theme_minimal()
p
```



#Note the reduction in Group A is reduced from the 18th as I have split the data set. No real indication of hours watched increasing from the visualisations

```
#Ztest
#Since this is > 30 we can perform the z-test
mean(Group_A$hours_watched)
```

```
## [1] 4.336125
```

```
mean(Group_B$hours_watched)
```

```
## [1] 4.810875
```

```
mean(Group_A$hours_watched)-mean(Group_B$hours_watched)
```

```
## [1] -0.47475
```

```
#mean of the treatment - mean of the control group
sd(Group_A$hours_watched)
```

```
## [1] 1.324221
```

```
sd(Group_B$hours_watched)
```

```
## [1] 1.32919
```

```
Group_A %>%
mutate(zscore = (hours_watched - mean(hours_watched))/sd(hours_watched))
```

```
## # A tibble: 880 x 9
##    date                gender   age social_metric time_since_sign~ demographic
##    <dttm>              <chr>  <dbl>         <dbl>            <dbl>       <dbl>
##  1 2020-07-01 00:00:00 F         28             5             19.3           1
##  2 2020-07-01 00:00:00 F         32             7             11.5           1
##  3 2020-07-01 00:00:00 F         25             5              3.3           1
##  4 2020-07-01 00:00:00 F         32            10             19.4           1
##  5 2020-07-01 00:00:00 F         32             1             17.5           1
##  6 2020-07-01 00:00:00 F         26             8              2.6           1
##  7 2020-07-01 00:00:00 F         32             8              6.9           1
##  8 2020-07-02 00:00:00 F         25            10             11.1           1
##  9 2020-07-02 00:00:00 F         22             5              0.2           1
## 10 2020-07-02 00:00:00 F         19             6              3.1           1
## # ... with 870 more rows, and 3 more variables: group <chr>,
## #   hours_watched <dbl>, zscore <dbl>
```

```
Group_B %>%
mutate(zscore = (hours_watched - mean(hours_watched))/sd(hours_watched))
```

```
## # A tibble: 120 x 9
##    date                gender   age social_metric time_since_sign~ demographic
##    <dttm>              <chr>  <dbl>         <dbl>            <dbl>       <dbl>
##  1 2020-07-18 00:00:00 F         39             5             14.8           3
##  2 2020-07-18 00:00:00 M         45             0              2.2           4
##  3 2020-07-18 00:00:00 F         28             8              1.4           1
##  4 2020-07-18 00:00:00 M         53             4              8.2           4
##  5 2020-07-18 00:00:00 M         45             8              9.1           4
##  6 2020-07-19 00:00:00 F         31             5              0.6           1
##  7 2020-07-19 00:00:00 M         42             9              1.3           4
##  8 2020-07-19 00:00:00 F         40             6              0.1           3
##  9 2020-07-19 00:00:00 F         54             4              8.6           3
## 10 2020-07-19 00:00:00 M         44             6             20.8           4
## # ... with 110 more rows, and 3 more variables: group <chr>,
## #   hours_watched <dbl>, zscore <dbl>
```

```
sigma <- 1.32919
mu0 <- 4.810875
# first calculate the standard error
SE <- 1.32919/sqrt(120)
print(SE)
```

```
## [1] 0.1213379
```

```
# calculate the z score
z_score <- (mean(Group_B$hours_watched) - 4.810875)/0.1013498
print(z_score)
```

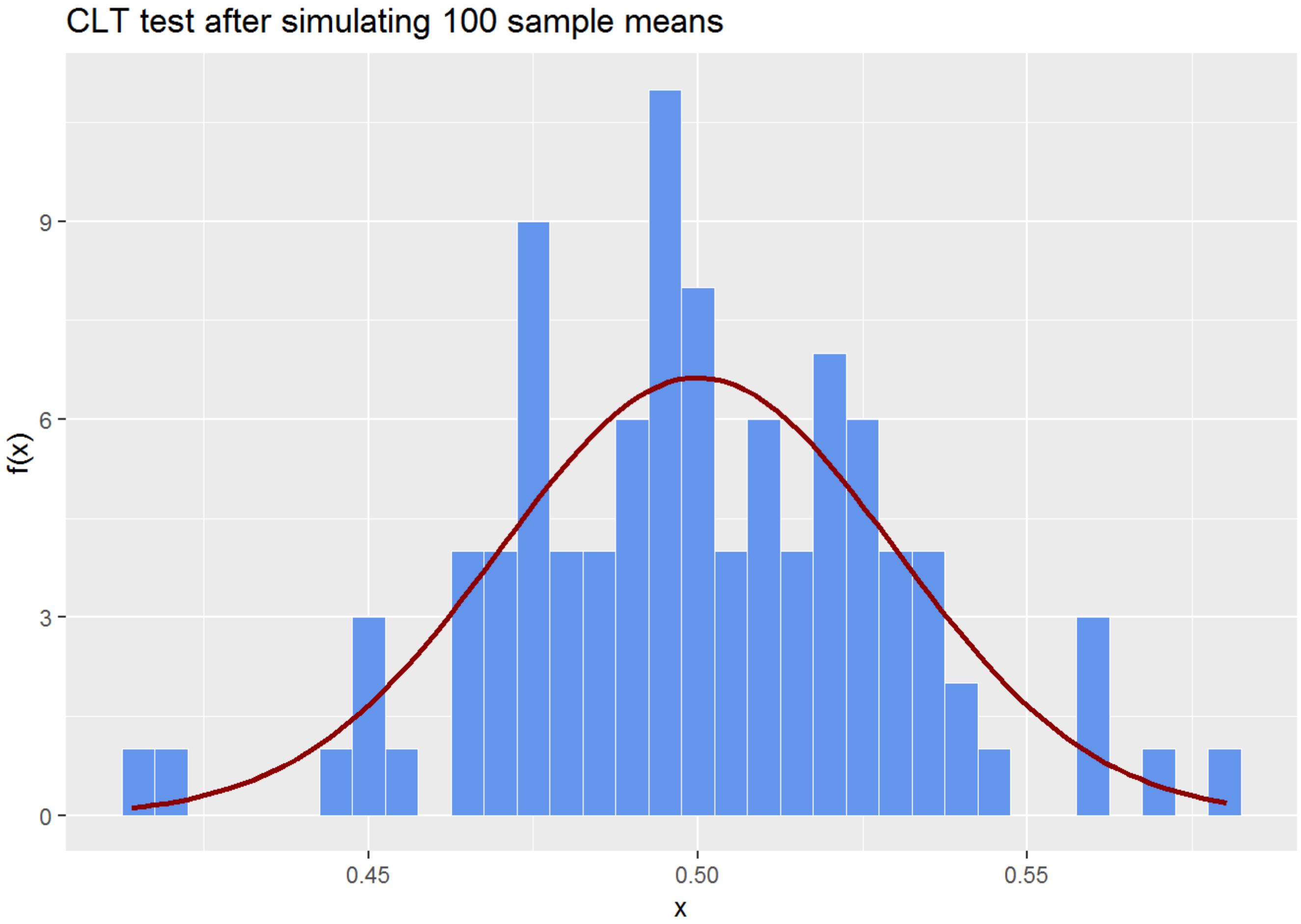
```
## [1] 0
```

```
pnorm(0,)
```

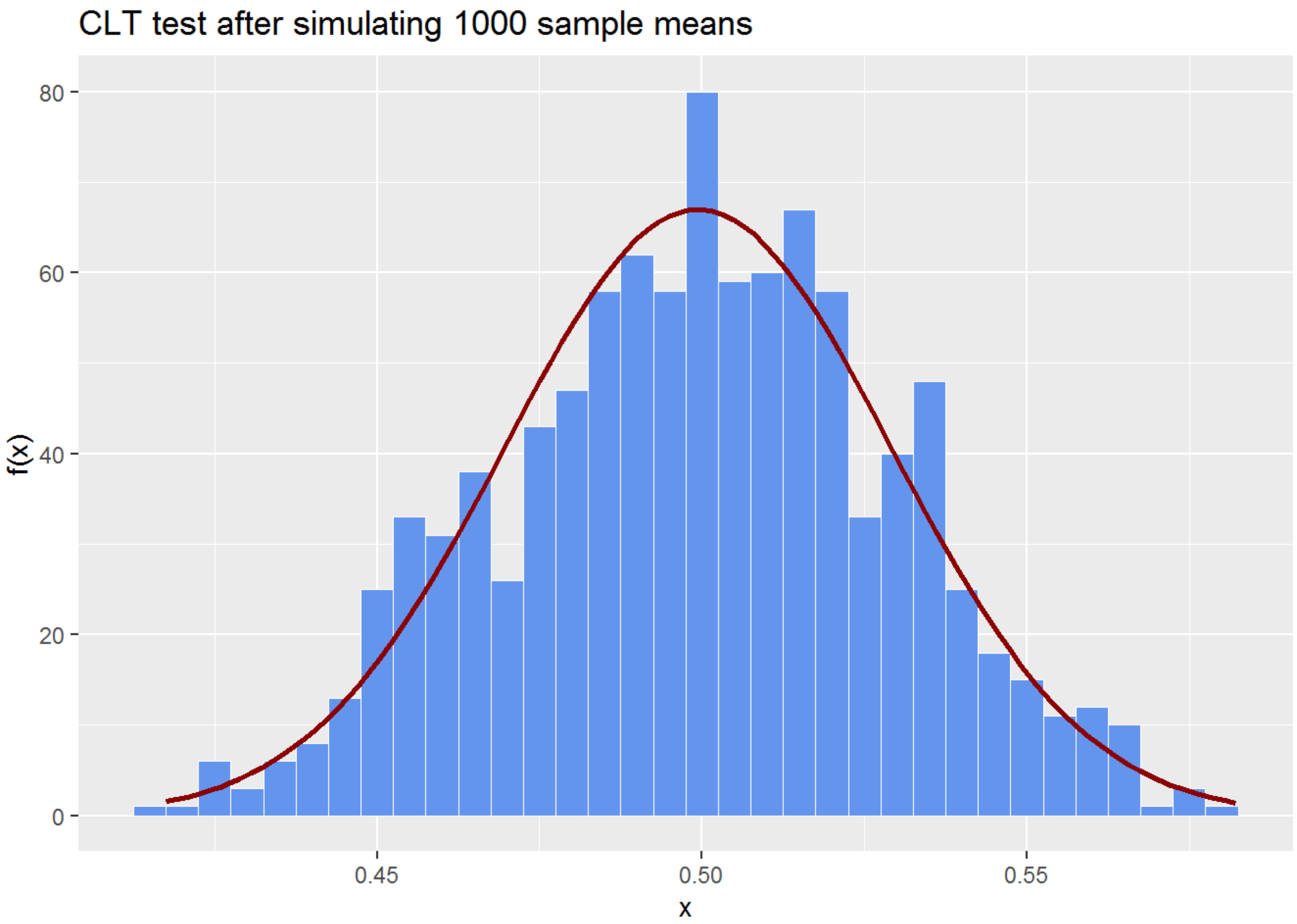
```
## [1] 0.5
```

```
#Hypothesis
#The null Hypothesis is the new algorithm does not have  an impact on Hours watched.
#The alternative Hypothesis  the new algorithm does have an impact on Hours watched
#With a .5 p value we can reject the null hypothesis given that the null hypothesis is true
```

```
#set sample size
n_sample <- 100

# set number of times the samples will be summed
n_sum <- 100

hours_watched_means <- rep(0, n_sum)
for (i in seq(1, n_sum)){
  # create a list of random numbers between 0 and 1
  x_rand <- runif(n_sample, min=0, max=1.)

  # add these to the sum
  hours_watched_means[i] <- mean(x_rand)
}


# add a normal distribution curve for comparison
mean <- mean(hours_watched_means)
sd <- sd(hours_watched_means)
binwidth <- 0.005

title_txt <- sprintf('CLT test after simulating %d sample means', n_sum)

gg <- ggplot()
gg <- gg + geom_histogram(aes(x = hours_watched_means),
                          binwidth=binwidth,
                          colour = "white",
                          fill = "cornflowerblue",
                          size = 0.1)

gg <- gg + geom_line(stat = "function",
                     aes(x = hours_watched_means),
                     fun = function(x) dnorm(x, mean = mean, sd = sd) * n_sum * binwidth,
                     color = "darkred",
                     size = 1)

gg <- gg + labs(x = 'x', y = 'f(x)', title=title_txt)
# ggsave('clt_example2.png', width = 6, height = 5)
gg
```


CLT test after simulating 100 sample means
f(x)
9
6
3
0
0.45
0.50
0.55
x

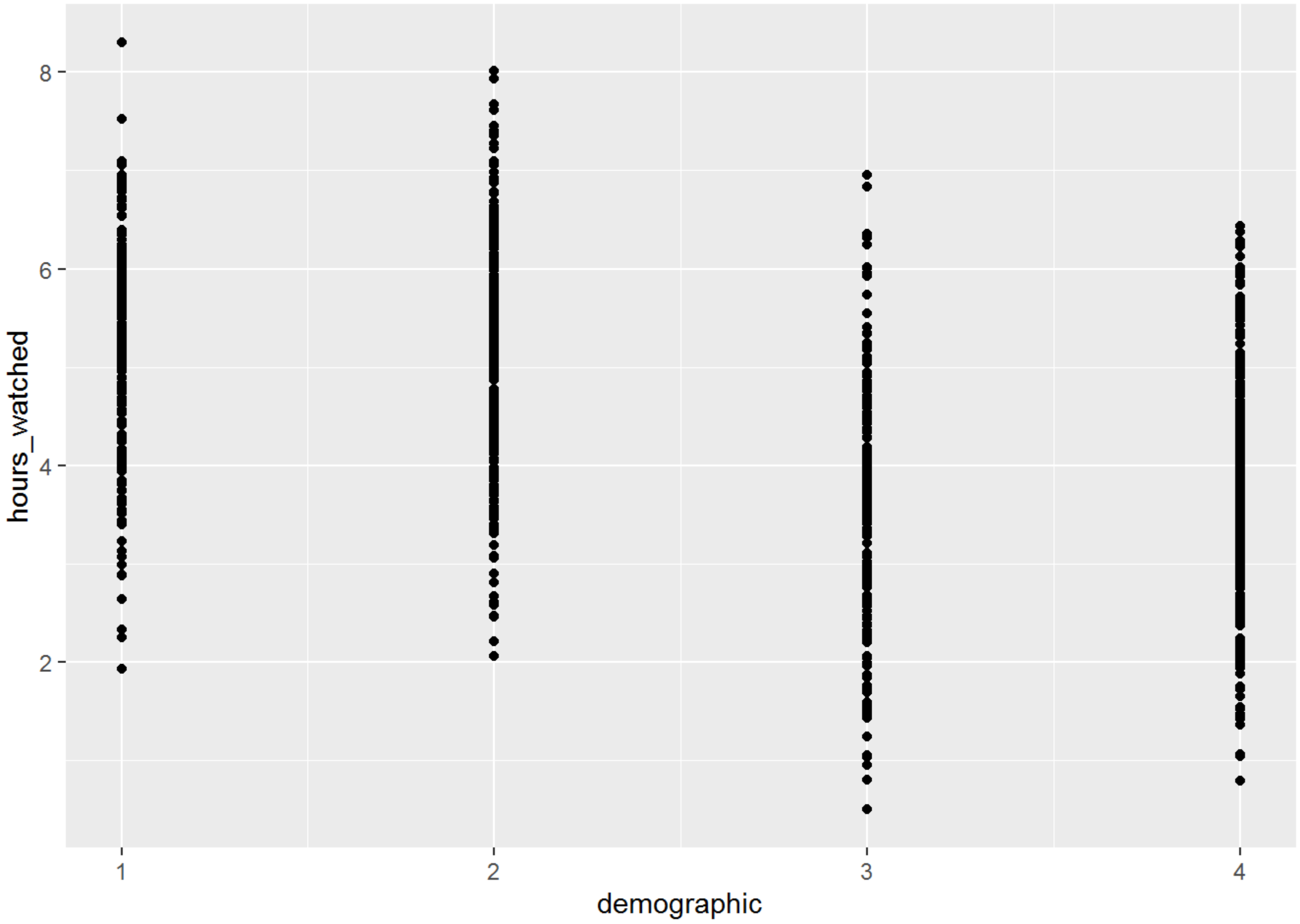
```r
# set sample size
n_sample <- 100

# set number of times the samples will be summed
n_sum <- 1000

hours_watched_means <- rep(0, n_sum)
for (i in seq(1, n_sum)){
  # create a list of random numbers between 0 and 1
  x_rand <- runif(n_sample, min=0, max=1.)

  # add these to the sum
  hours_watched_means[i] <- mean(x_rand)
}


# add a normal distribution curve for comparison
mean <- mean(hours_watched_means)
sd <- sd(hours_watched_means)
binwidth <- 0.005

title_txt <- sprintf('CLT test after simulating %d sample means', n_sum)

gg <- ggplot()
gg <- gg + geom_histogram(aes(x = hours_watched_means),
                          binwidth=binwidth,
                          colour = "white",
                          fill = "cornflowerblue",
                          size = 0.1)

gg <- gg + geom_line(stat = "function",
                     aes(x = hours_watched_means),
                     fun = function(x) dnorm(x, mean = mean, sd = sd) * n_sum * binwidth,
                     color = "darkred",
                     size = 1)

gg <- gg + labs(x = 'x', y = 'f(x)', title=title_txt)
# ggsave('clt_example2.png', width = 6, height = 5)
gg
```



```
ggplot(Assignment_3_WNW_Data, aes(x = demographic, y = hours_watched)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.
```



```
ggplot(Group_A, aes(x = demographic, y = hours_watched)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 4.015
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2.015
```

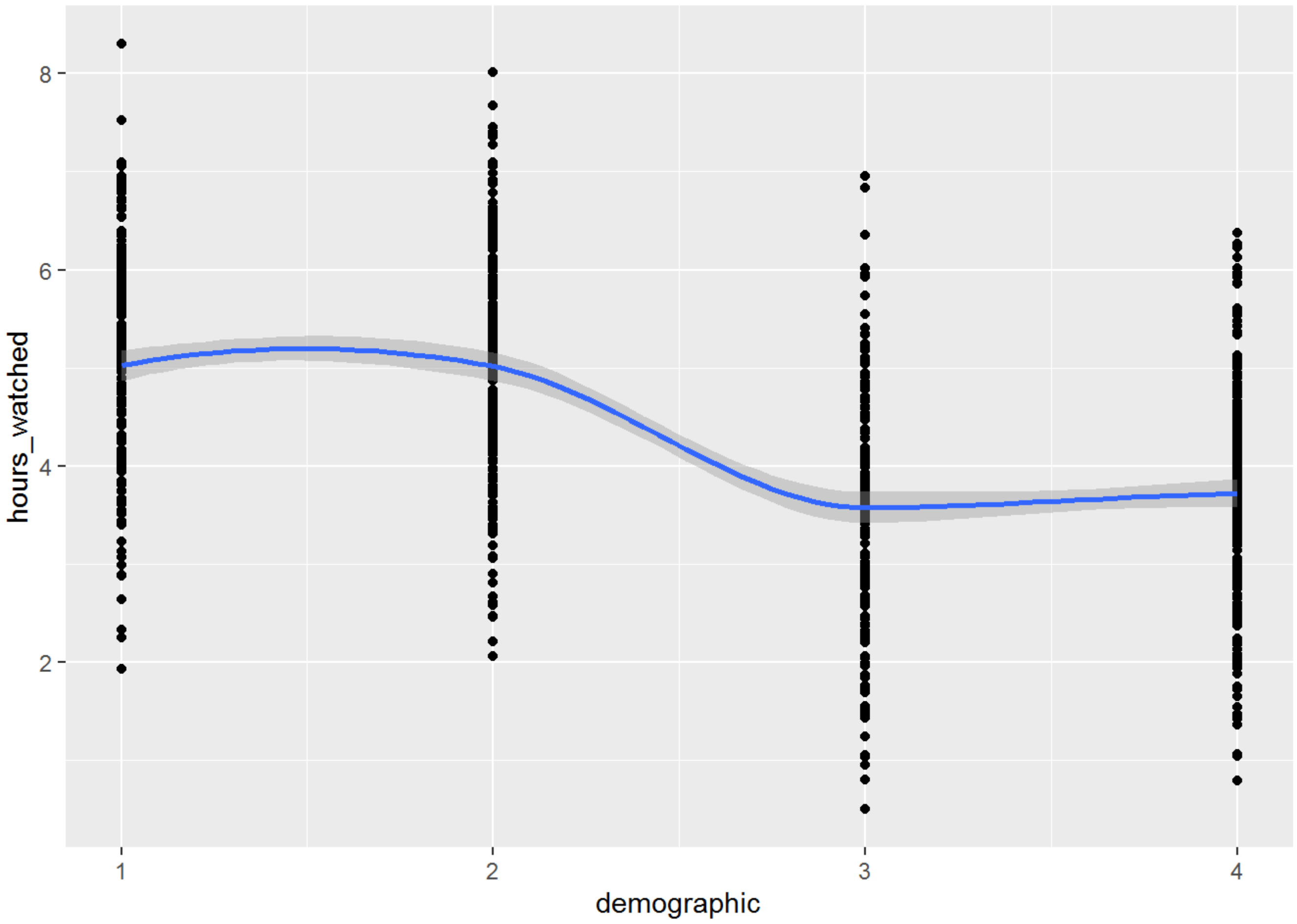
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1.056e-016
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 4.015
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 2.015
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 1.056e-016
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1
```



```
ggplot(Group_B, aes(x = demographic, y = hours_watched)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 4.015
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2.015
```

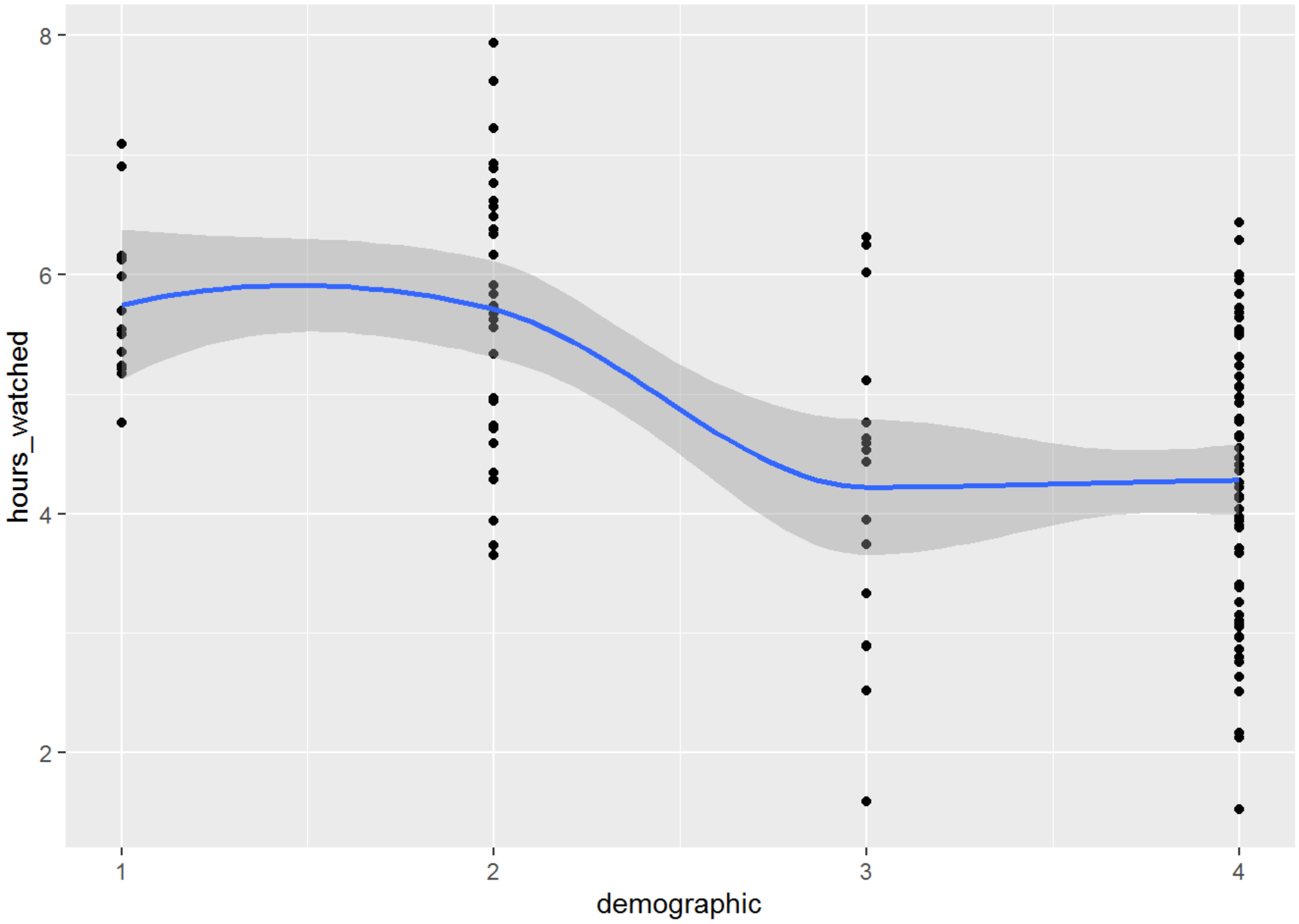
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 4.015
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 2.015
```

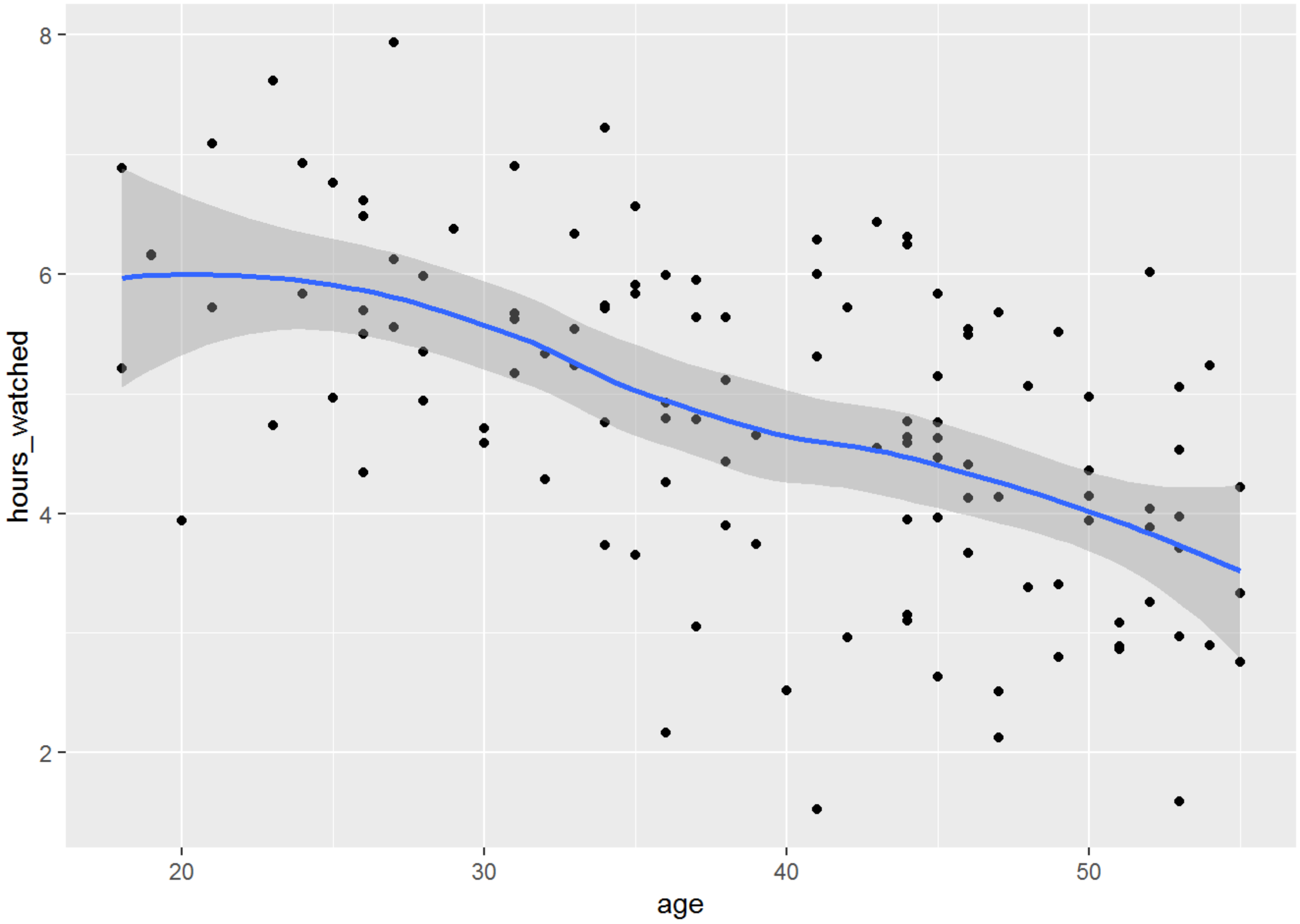
```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1
```



```
ggplot(Group_B, aes(x = age, y = hours_watched)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



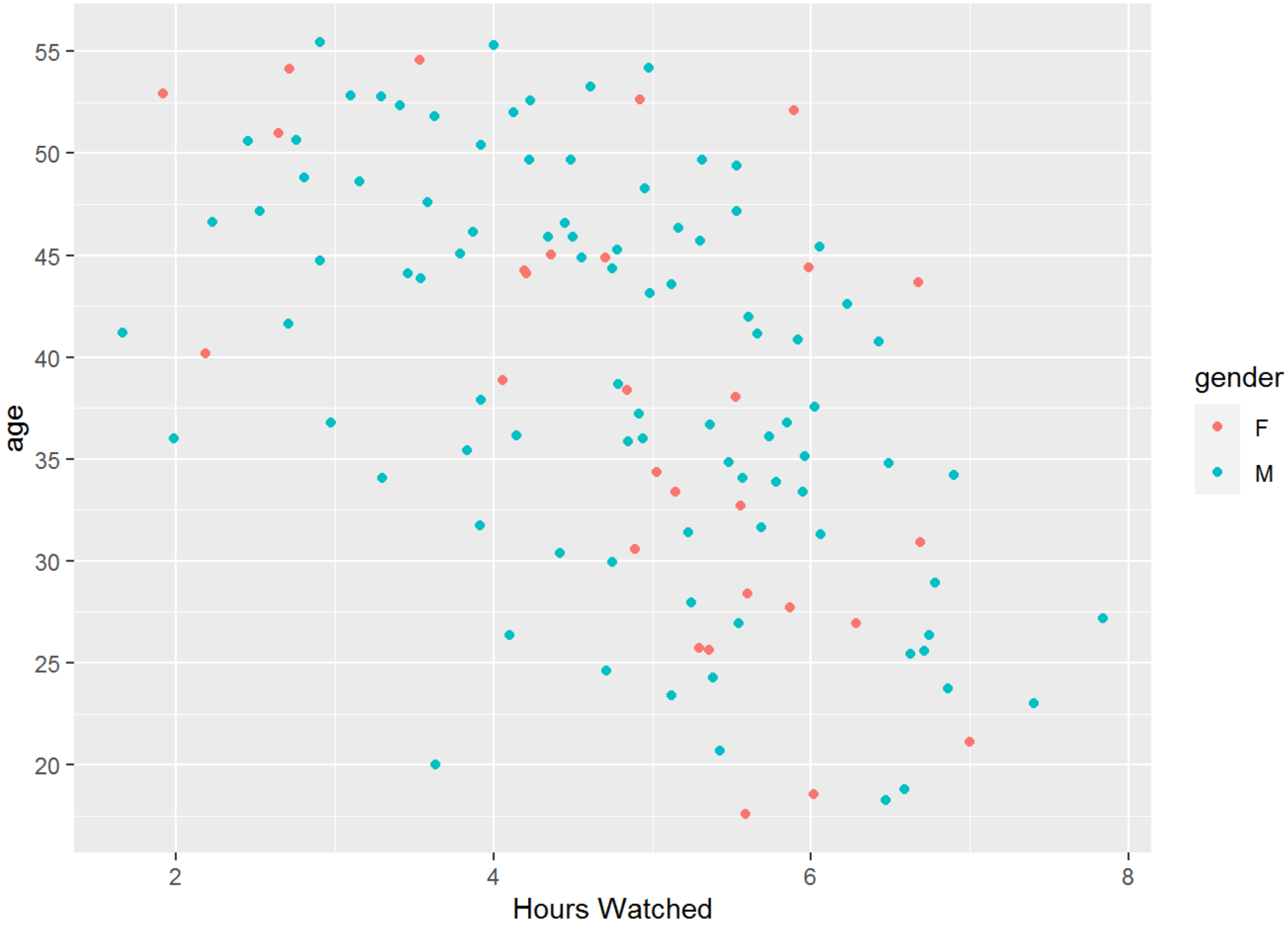
```
z_alpha <- 1.96
effect_est <- 0.03365
sd_est <- -0.50983
n_ss <- ceiling((z_alpha * sd_est / effect_est)^2)
(paste('Min sample size', n_ss))
```

```
## [1] "Min sample size 882"
```

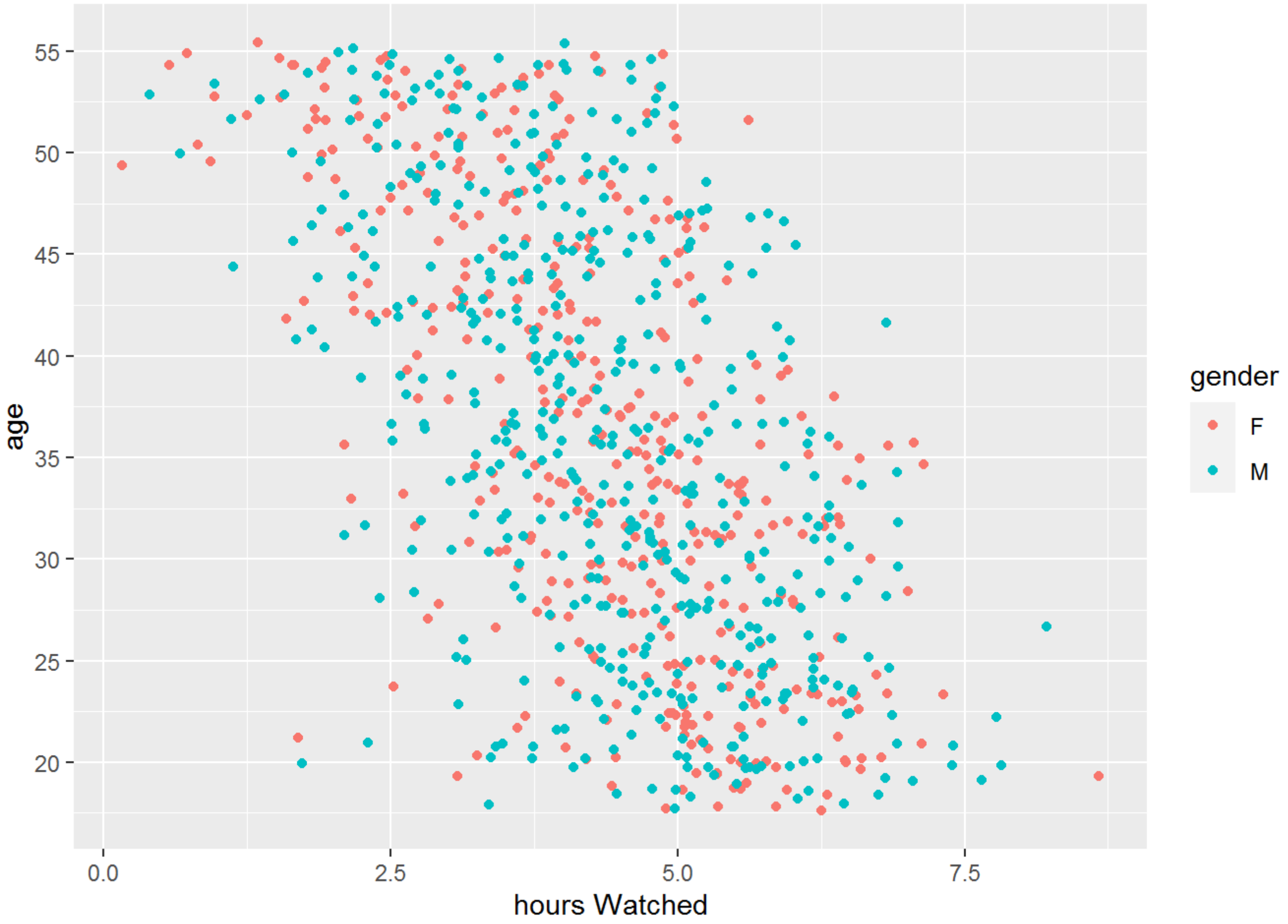
```
#The minimum sample size in which we could obtain statistically significant results
```

#Let’s see if the algorithm had any effect at all on the 18th July onwards on the customers targeted.We should see an increase in the percentage of group B with outcome = 1 compared to group A.

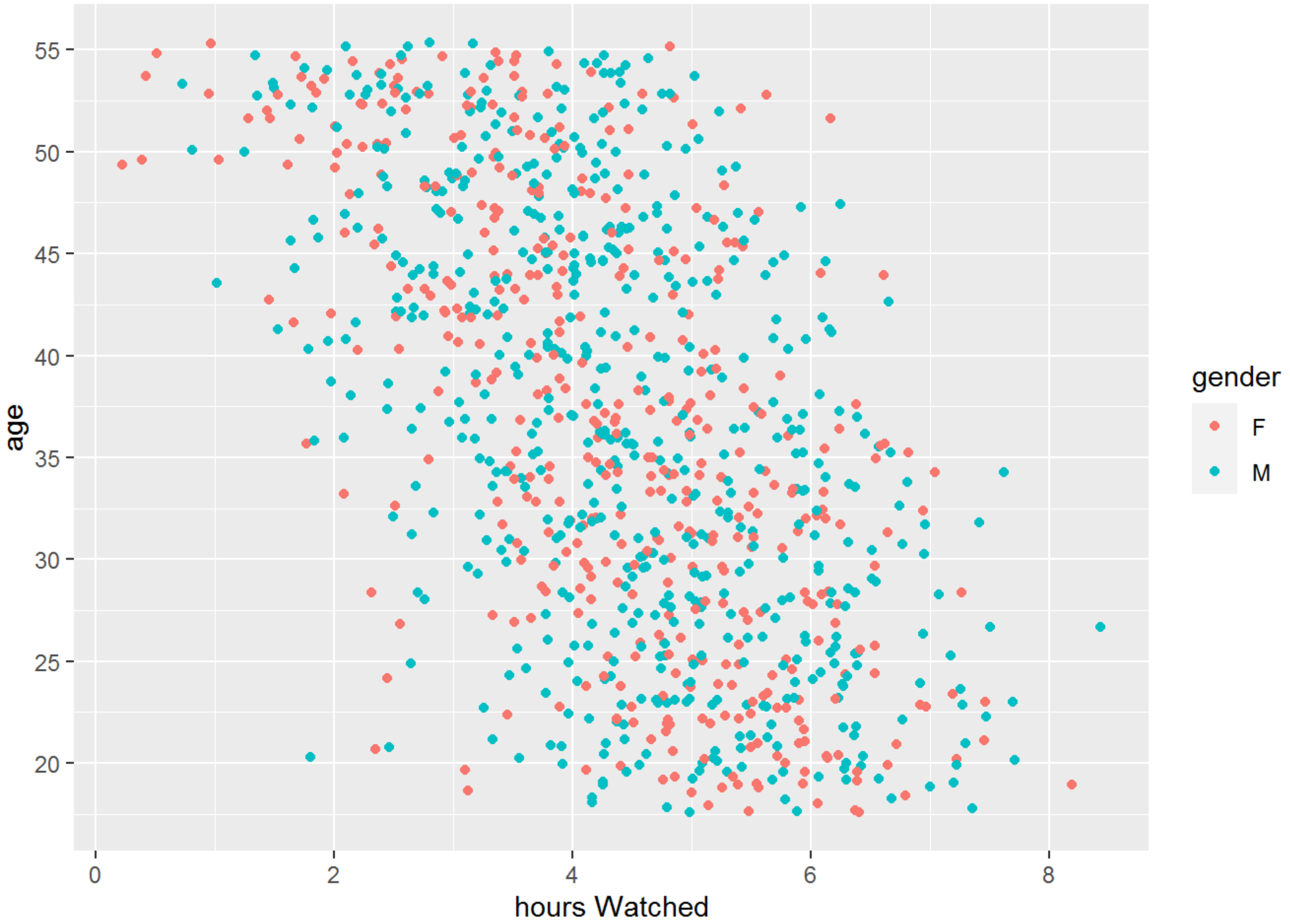
```
#Group B data treatment group
int_breaks_rounded <- function(x, n = 10) pretty(x, n)[round(pretty(x, n),1)%% 1 == 0]
gg <- ggplot()
gg <- gg + geom_point(position = position_jitter(width = 0.45,
                                                 height = 0.45),
                      aes(x=Group_B $hours_watched,
                          y=Group_B $age,
                          colour=factor(Group_B$gender)))
gg <- gg + scale_y_continuous(breaks= int_breaks_rounded )
gg <- gg + labs(x='Hours Watched', y='age',
                colour='gender')
gg
```



```
#Group A Control group
int_breaks_rounded <- function(x, n = 10) pretty(x, n)[round(pretty(x, n),1)%% 1 == 0]
gg <- ggplot()
gg <- gg + geom_point(position = position_jitter(width = 0.45,
                                                 height = 0.45),
                      aes(x=Group_A $hours_watched,
                          y=Group_A $age,
                          colour=factor(Group_A$gender)))
gg <- gg + scale_y_continuous(breaks= int_breaks_rounded )
gg <- gg + labs(x='hours Watched', y='age',
                colour='gender')
gg
```



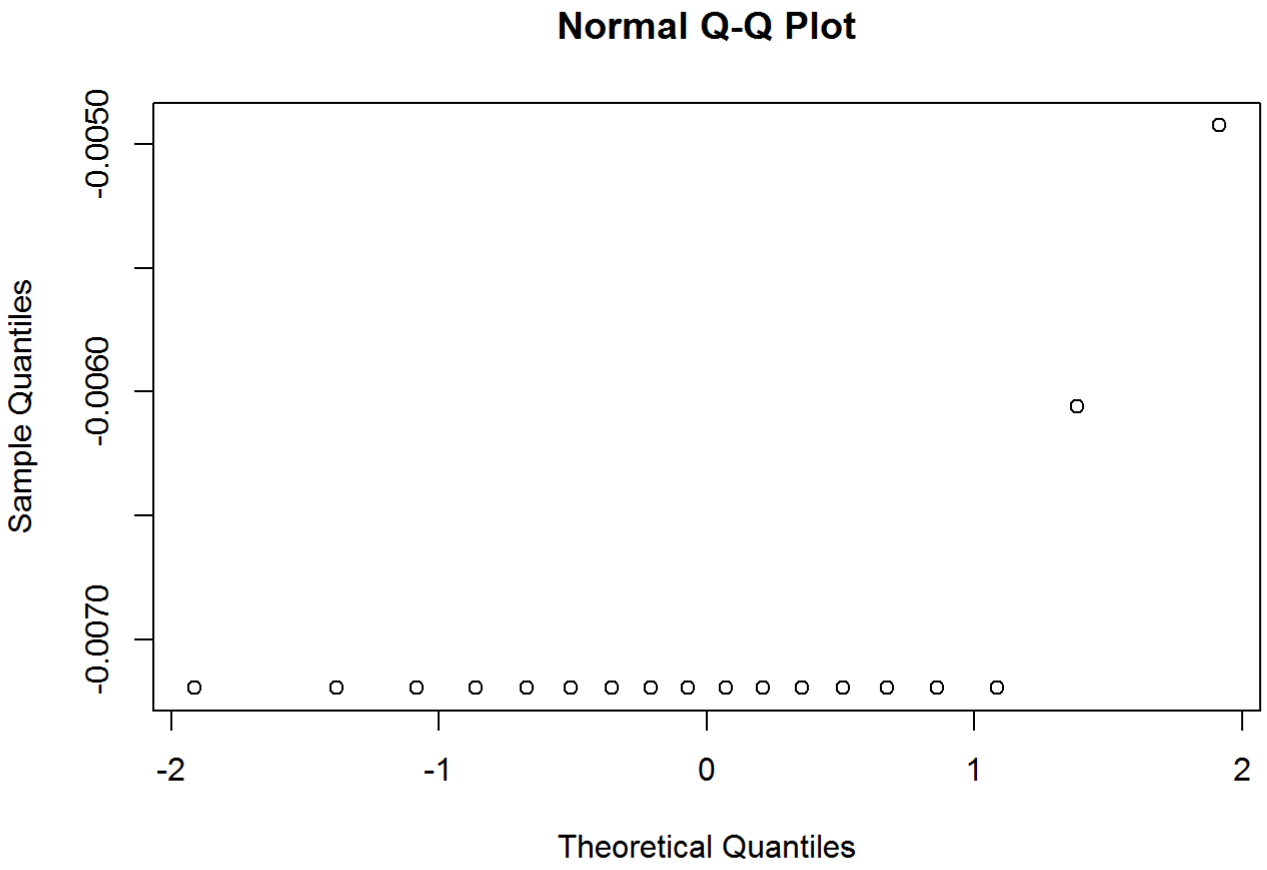
```
# Whole data set
int_breaks_rounded <- function(x, n = 10) pretty(x, n)[round(pretty(x, n),1)%% 1 == 0]
gg <- ggplot()
gg <- gg + geom_point(position = position_jitter(width = 0.45,
                                                 height = 0.45),
                      aes(x=Assignment_3_WNW_Data $hours_watched,
                          y=Assignment_3_WNW_Data $age,
                          colour=factor(Assignment_3_WNW_Data$gender)))
gg <- gg + scale_y_continuous(breaks= int_breaks_rounded )
gg <- gg + labs(x='hours Watched', y='age',
                colour='gender')
gg
```



```
# count the numbers in each demographic category based on the A/B group
check_a_df <- Group_A %>% filter(group=='A') %>% select(hours_watched,gender,demographic) %>%
  group_by(hours_watched,gender,demographic) %>% mutate(n_a=n()) %>% distinct()

check_b_df <- Group_B %>% filter(group=='B') %>% select(hours_watched,gender,demographic) %>%
  group_by(hours_watched,gender,demographic) %>% mutate(n_b=n()) %>% distinct()

# total numbers in each group
n_total_a <- sum(Group_A$group=='A')
n_total_b <- sum(Group_B$group=='B')

# proportions in each demographic
check_a_df$p_a <- check_a_df$n_a / n_total_a
check_b_df$p_b <- check_b_df$n_b / n_total_b

# join on demo categories
check_df <- inner_join(check_a_df, check_b_df)
```

```
## Joining, by = c("hours_watched", "gender", "demographic")
```

```
# calculate the difference in proportions
check_df$diff <- check_df$p_a - check_df$p_b

# if there is no bias aside from sampling noise then the difference should be small and normally distri
buted not the case in this distribution the samples are not drawn from the same population.
qqnorm(y=check_df$diff)
```



Let's see if the experiment had any effect at all on proportion of hours watched. If the new recommendations were successful then there should be an increase in the percentage of group B with outcome = 1 compared to group A.

```
print('Outcome breakdown:')
```

```
## [1] "Outcome breakdown:"
```

```
cond_A <- Group_A$group == 'A'
print(paste('A:', sum(Group_A$hours_watched[cond_A])/sum(cond_A)))
```

```
## [1] "A: 4.336125"
```

```
cond_B <- Group_B$group == 'B'
print(paste('B:', sum(Group_B$hours_watched[cond_B])/sum(cond_A)))
```

```
## [1] "B: 0.656028409090909"
```

```
# prepare data for group A
g_a <- Group_A %>% filter(group=='A') %>% ungroup() %>%
  select(hours_watched,gender,demographic) %>% group_by(hours_watched,gender,demographic) %>%
mutate(n_a=n(), n_a_o=sum(hours_watched)) %>% select(hours_watched,gender,demographic, n_a, n_a_o) %>%
distinct()
g_a$p_a <- g_a$n_a_o/g_a$n_a

g_b <- Group_B %>% filter(group=='B') %>% ungroup() %>%
  select(hours_watched,gender,demographic) %>% group_by(hours_watched,gender,demographic) %>%
  mutate(n_b=n(), n_b_o=sum(hours_watched)) %>% select(hours_watched,gender,demographic, n_b, n_b_o) %
>%
  distinct()
g_b$p_b <- g_b$n_b_o/g_b$n_b

# effect comparison: join on all common column names
effect_comp_df <- inner_join(g_a, g_b)
```

```
## Joining, by = c("hours_watched", "gender", "demographic")
```

```
effect_comp_df$effect <- effect_comp_df$p_b - effect_comp_df$p_a

pop_sd <- 0.02

z_alpha <- 1.96
effect_comp_df$n_ss <- (z_alpha * pop_sd / effect_comp_df$effect)^2

effect_comp_df$significant <- effect_comp_df$n_a > effect_comp_df$n_ss

effect_comp_df
```

```
## # A tibble: 18 x 12
## # Groups:   hours_watched, gender, demographic [18]
##    hours_watched gender demographic   n_a n_a_o   p_a   n_b n_b_o   p_b effect
##            <dbl> <chr>        <dbl> <int> <dbl> <dbl> <int> <dbl> <dbl>  <dbl>
##  1          5.35 F                1     1  5.35  5.35     1  5.35  5.35      0
##  2          7.09 F                1     1  7.09  7.09     1  7.09  7.09      0
##  3          5.21 F                1     1  5.21  5.21     1  5.21  5.21      0
##  4          4.76 F                1     1  4.76  4.76     1  4.76  4.76      0
##  5          5.69 F                1     1  5.69  5.69     1  5.69  5.69      0
##  6          5.98 F                1     1  5.98  5.98     1  5.98  5.98      0
##  7          4.94 M                2     3 14.8    4.94     1  4.94  4.94      0
##  8          5.72 M                2     1  5.72  5.72     1  5.72  5.72      0
##  9          4.28 M                2     1  4.28  4.28     1  4.28  4.28      0
## 10          4.73 M                2     1  4.73  4.73     1  4.73  4.73      0
## 11          4.96 M                2     1  4.96  4.96     1  4.96  4.96      0
## 12          4.59 M                2     2  9.18  4.59     1  4.59  4.59      0
## 13          3.73 M                2     1  3.73  3.73     1  3.73  3.73      0
## 14          5.33 M                2     1  5.33  5.33     1  5.33  5.33      0
## 15          3.33 F                3     1  3.33  3.33     1  3.33  3.33      0
## 16          3.74 F                3     1  3.74  3.74     1  3.74  3.74      0
## 17          5.11 F                3     1  5.11  5.11     1  5.11  5.11      0
## 18          4.59 F                3     1  4.59  4.59     1  4.59  4.59      0
## # ... with 2 more variables: n_ss <dbl>, significant <lgl>
```

#From this AB test we would conclude that the new recommendation engine did not increase the hours watched and was only due to a biased sample. With no impact in the effect column.

#This is not an actionable insight for the business with the current sample.Improvements to the sample by extracting a population sample would then become an actionable insight that would help the business both in the short term regarding an increasing hours watched and in the long term regarding better targeting of demographics as per Demographic 4.