Assignment 3

Why Not Watch Case Study

Thaddeus Lee, s3933533

Last updated: 20 June, 2022

```
# Read CSV file
wnw_df <- read.csv("streaming_data.csv")
wnw1_df <- wnw_df
# Check structure
str(wnw_df)
```

```
## 'data.frame':    1000 obs. of  8 variables:
##  $ date             : chr  "1-Jul" "1-Jul" "1-Jul" "1-Jul" ...
##  $ gender           : chr  "F" "F" "F" "M" ...
##  $ age              : int  28 32 39 52 25 51 53 42 41 20 ...
##  $ social_metric    : int  5 7 4 10 1 0 5 6 8 7 ...
##  $ time_since_signup: num  19.3 11.5 4.3 9.5 19.5 22.6 4.2 8.5 16.9 23 ...
##  $ demographic      : int  1 1 3 4 2 4 3 4 4 2 ...
##  $ group            : chr  "A" "A" "A" "A" ...
##  $ hours_watched    : num  4.08 2.99 5.74 4.13 4.68 3.4 3.07 2.77 2.24 5.39 ...
```

```
summary(wnw_df)
```

```
##      date              gender               age        social_metric
##  Length:1000        Length:1000        Min.   :18.00   Min.   : 0.000
##  Class :character   Class :character   1st Qu.:28.00   1st Qu.: 2.000
##  Mode  :character   Mode  :character   Median :36.00   Median : 5.000
##                                        Mean   :36.49   Mean   : 4.911
##                                        3rd Qu.:46.00   3rd Qu.: 8.000
##                                        Max.   :55.00   Max.   :10.000
##  time_since_signup  demographic       group           hours_watched
##  Min.   : 0.00     Min.   :1.000   Length:1000        Min.   :0.500
##  1st Qu.: 5.70     1st Qu.:2.000   Class :character   1st Qu.:3.530
##  Median :11.80     Median :3.000   Mode  :character   Median :4.415
##  Mean   :11.97     Mean   :2.603                      Mean   :4.393
##  3rd Qu.:18.70     3rd Qu.:4.000                      3rd Qu.:5.322
##  Max.   :24.00     Max.   :4.000                      Max.   :8.300
```

```
# Changing data types:-
wnw1_df$gender <- as.factor(wnw1_df$gender)
wnw1_df$social_metric <- as.factor(wnw1_df$social_metric)
wnw1_df$demographic <- as.factor(wnw1_df$demographic)
wnw1_df$date <- as.Date(wnw1_df$date, format = "%d-%b")
wnw1_df$group <- as.factor(wnw1_df$group)



# Filter Data set:-
#wnw1_df$change <- ifelse(wnw1_df$date > "2022-07-17", TRUE, FALSE)



wnw_pre_change <- wnw1_df %>%
  filter(date <= "2022-07-17")


wnw_post_change <- wnw1_df %>%
  filter(date > "2022-07-17")
```

Code Chunk: Gender and Demographics - counts and Proportions.

```
################################################################################
#######       GENDER: Pre-Change
################################################################################
# Sum of women and men in data set before change
Male_count_pre <- sum(wnw_pre_change$gender == "M") # 311
Female_count_pre <- sum(wnw_pre_change$gender == "F") # 237
Total_pre <- nrow(wnw_pre_change)
```

```
#Gender Proportion:- Before changes is algorithm
Male_prop_pre <- Male_count_pre/Total_pre
Female_prop_pre <- Female_count_pre/Total_pre

# Male Proprtion - Pre-Trial = 0.5675
# Female Proprtion - Pre-Trial = 0.4325
################################################################################
#######       GENDER: Post-Change Group B Data Frame
################################################################################

wnw_post_B <- wnw_post_change %>% filter(group == "B")

# Number of Male and Female in Group B
Male_count_post_B <- wnw_post_B %>% filter(gender == "M") %>% nrow()
Female_count_post_B <- wnw_post_B %>% filter(gender == "F") %>% nrow()

# Proportion of Male and Female in Group B
Total_post_B <- wnw_post_B %>% nrow()
Male_prop_post_B <- Male_count_post_B/Total_post_B
Female_prop_post_B <- Female_count_post_B/Total_post_B

# Male Proportion - Group B = 0.7583
# Female Proportion - Group B = 0.2417
################################################################################
#######       GENDER: Post-Change Group A Data Frame
################################################################################
wnw_post_A <- wnw_post_change %>% filter(group == "A")

# Count of Male and female in Group A
Male_count_post_A <- wnw_post_A %>% filter(gender == "M") %>% nrow()
Female_count_post_A <- wnw_post_A %>% filter(gender == "F") %>% nrow()


# Proportion of Male and Female in Group A
Total_post_A <- wnw_post_A %>% nrow()
Male_prop_post_A <- Male_count_post_A/Total_post_A
Female_prop_post_A <- Female_count_post_A/Total_post_A

# Male Proportion - Group A = 0.5090
# Female Proportion - Group A = 0.4910
################################################################################
######                   DEMOGRAPHICS - Pre-Change                        #####
################################################################################
#           Number in each demographic in data set before change               #
################################################################################
Pre_Demo1 <- sum(wnw_pre_change$demographic == "1") # 129 women aged <= 35
Pre_Demo2 <- sum(wnw_pre_change$demographic == "2") # 140 Men aged <= 35
Pre_Demo3 <- sum(wnw_pre_change$demographic == "3") # 108 Women aged > 35
Pre_Demo4 <- sum(wnw_pre_change$demographic == "4") # 171 Men Aged > 35


################################################################################
#           Demographics Proportions:- Before changes is algorithm             #
################################################################################
Total_pre <- nrow(wnw_pre_change) #Counts total number of rows


Pre_Demo1_prop <- Pre_Demo1/Total_pre # 0.235
Pre_Demo2_prop <- Pre_Demo2/Total_pre # 0.255
Pre_Demo3_prop <- Pre_Demo3/Total_pre # 0.197
Pre_Demo4_prop <- Pre_Demo4/Total_pre # 0.312




################################################################################
######                   DEMOGRAPHICS - Post-Change                       #####
################################################################################
################################################################################
```

```
# GROUP A:- Number in each demographic in data set After change                #
################################################################################
PostA_Demo1 <- sum(wnw_post_A$demographic == "1") # 74 women aged <= 35
PostA_Demo2 <- sum(wnw_post_A$demographic == "2") # 96 Men aged <= 35
PostA_Demo3 <- sum(wnw_post_A$demographic == "3") # 89 Women aged > 35
PostA_Demo4 <- sum(wnw_post_A$demographic == "4") # 73 Men Aged > 35


################################################################################
# GROUP A:- Demographics Proportions:- After changes is algorithm              #
################################################################################
Total_post_A <- nrow(wnw_post_A) #Counts total number of rows


PostA_Demo1_prop <- PostA_Demo1/Total_post_A # 0.223
PostA_Demo2_prop <- PostA_Demo2/Total_post_A # 0.289
PostA_Demo3_prop <- PostA_Demo3/Total_post_A # 0.268
PostA_Demo4_prop <- PostA_Demo4/Total_post_A # 0.22




################################################################################
# GROUP B:- Number in each demographic in data set After change                #
################################################################################
PostB_Demo1 <- sum(wnw_post_B$demographic == "1") # 13 women aged <= 35
PostB_Demo2 <- sum(wnw_post_B$demographic == "2") # 32 Men aged <= 35
PostB_Demo3 <- sum(wnw_post_B$demographic == "3") # 16 Women aged > 35
PostB_Demo4 <- sum(wnw_post_B$demographic == "4") # 49 Men Aged > 35


################################################################################
# GROUP B:- Demographics Proportions:- After changes is algorithm              #
################################################################################
Total_post_B <- nrow(wnw_post_B) #Counts total number of rows


PostB_Demo1_prop <- PostB_Demo1/Total_post_B # 0.108
PostB_Demo2_prop <- PostB_Demo2/Total_post_B # 0.267
PostB_Demo3_prop <- PostB_Demo3/Total_post_B # 0.133
PostB_Demo4_prop <- PostB_Demo4/Total_post_B # 0.492


################################################################################
# GROUP A:- SUMMARY STATISTICS - DEMOGRAPHICS
################################################################################
wnw_post_A %<>% group_by(demographic) %>%
                          mutate(mean_hours_demographic = mean(hours_watched),
                                 sd_hours_demographic = sd(hours_watched)) %<>% ungroup()

STATS_wnw_post_A <- wnw_post_A %>% select(demographic, mean_hours_demographic, sd_hours_demographic)

STATS_wnw_post_A <- distinct(STATS_wnw_post_A)

# Adding - Post-weights
STATS_wnw_post_A$post_weight[STATS_wnw_post_A$demographic == "1"] <- Pre_Demo1_prop
STATS_wnw_post_A$post_weight[STATS_wnw_post_A$demographic == "2"] <- Pre_Demo2_prop
STATS_wnw_post_A$post_weight[STATS_wnw_post_A$demographic == "3"] <- Pre_Demo3_prop
STATS_wnw_post_A$post_weight[STATS_wnw_post_A$demographic == "4"] <- Pre_Demo4_prop


GroupA_stats <- summarise(wnw_post_A,
                          mean = mean(hours_watched, na.rm = TRUE),
                          sd = sd(hours_watched, na.rm = TRUE),
                          n = Total_post_A)


################################################################################
# GROUP B:- SUMMARY STATISTICS - DEMOGRAPHICS
```

```
################################################################################
wnw_post_B %<>% group_by(demographic) %>%
                          mutate(mean_hours_demographic = mean(hours_watched),
                                 sd_hours_demographic = sd(hours_watched)) %<>% ungroup()

STATS_wnw_post_B <- wnw_post_B %>% select(demographic, mean_hours_demographic, sd_hours_demographic)

STATS_wnw_post_B <- distinct(STATS_wnw_post_B)

# Adding - Post-weights
STATS_wnw_post_B$post_weight[STATS_wnw_post_B$demographic == "1"] <- Pre_Demo1_prop
STATS_wnw_post_B$post_weight[STATS_wnw_post_B$demographic == "2"] <- Pre_Demo2_prop
STATS_wnw_post_B$post_weight[STATS_wnw_post_B$demographic == "3"] <- Pre_Demo3_prop
STATS_wnw_post_B$post_weight[STATS_wnw_post_B$demographic == "4"] <- Pre_Demo4_prop


GroupB_stats <- summarise(wnw_post_B,
                          mean = mean(hours_watched, na.rm = TRUE),
                          sd = sd(hours_watched, na.rm = TRUE),
                          n = Total_post_B)
```

Code Chunk: Bias Comparisons.

```
################################################################################
# Bias comparison
################################################################################
#Demographic
check_demo_bias <-
  data.frame(Demographic = c("1", "2", "3", "4"),
             WNW_Before = c(Pre_Demo1_prop, Pre_Demo2_prop, Pre_Demo3_prop, Pre_Demo4_prop),
             Group_A_Prop = c(PostA_Demo1_prop, PostA_Demo2_prop, PostA_Demo3_prop, PostA_Demo4_prop),
             Group_B_Prop = c(PostB_Demo1_prop, PostB_Demo2_prop, PostB_Demo3_prop, PostB_Demo4_prop))
check_demo_bias %<>% mutate(Difference = Group_B_Prop - Group_A_Prop)
check_demo_bias
```

Demographic <chr>	WNW_B... <dbl>	Group_A_Pr... <dbl>	Group_B_Prop <dbl>	Difference <dbl>
1	0.2354015	0.2228916	0.1083333	-0.11455823
2	0.2554745	0.2891566	0.2666667	-0.02248996
3	0.1970803	0.2680723	0.1333333	-0.13473896
4	0.3120438	0.2198795	0.4916667	0.27178715

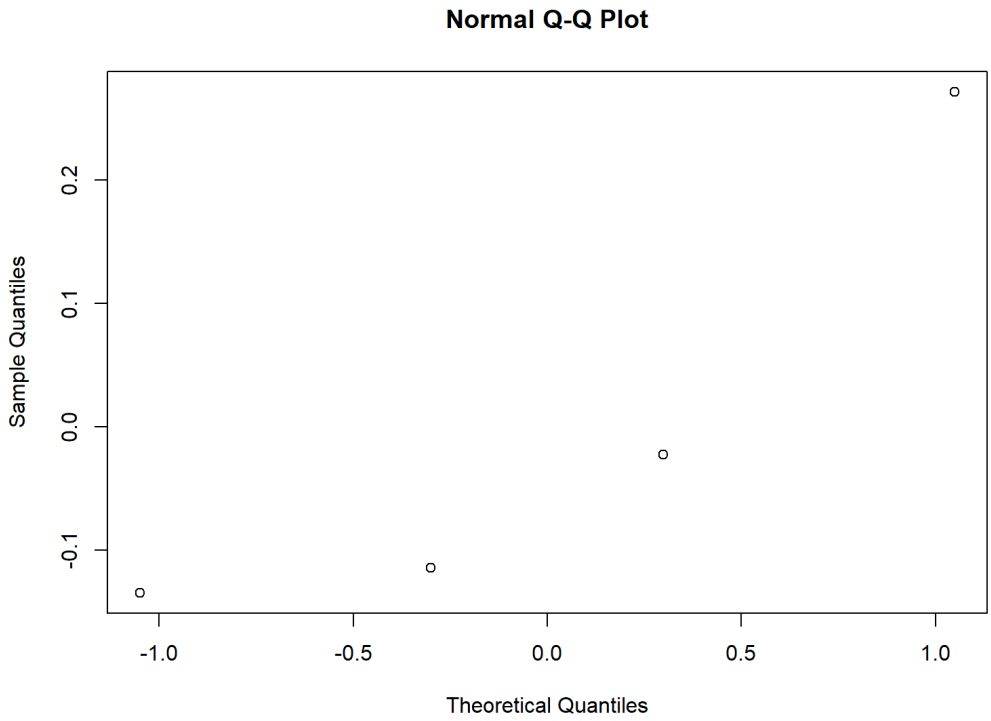
4 rows

```
knitr::kable(check_demo_bias, caption ="Demographic - Proportion Comparison")
```

Demographic - Proportion Comparison

Demographic	WNW_Before	Group_A_Prop	Group_B_Prop	Difference
1	0.2354015	0.2228916	0.1083333	-0.1145582
2	0.2554745	0.2891566	0.2666667	-0.0224900
3	0.1970803	0.2680723	0.1333333	-0.1347390
4	0.3120438	0.2198795	0.4916667	0.2717871

```
qqnorm(check_demo_bias$Difference)
```




```
################################################################################
# Gender
Check_Gender_Bias <-
  data.frame(WNW_Before = c(Female_prop_pre, Male_prop_pre),
             Group_A_Prop = c(Female_prop_post_A, Male_prop_post_A),
             Group_B_Prop = c(Female_prop_post_B, Male_prop_post_B))
Check_Gender_Bias %<>% mutate(Difference = Group_B_Prop - Group_A_Prop)

Check_Gender_Bias
```

WNW_Before <dbl>	Group_A_Prop <dbl>	Group_B_Prop <dbl>	Difference <dbl>
0.4324818	0.4909639	0.2416667	-0.2492972
0.5675182	0.5090361	0.7583333	0.2492972

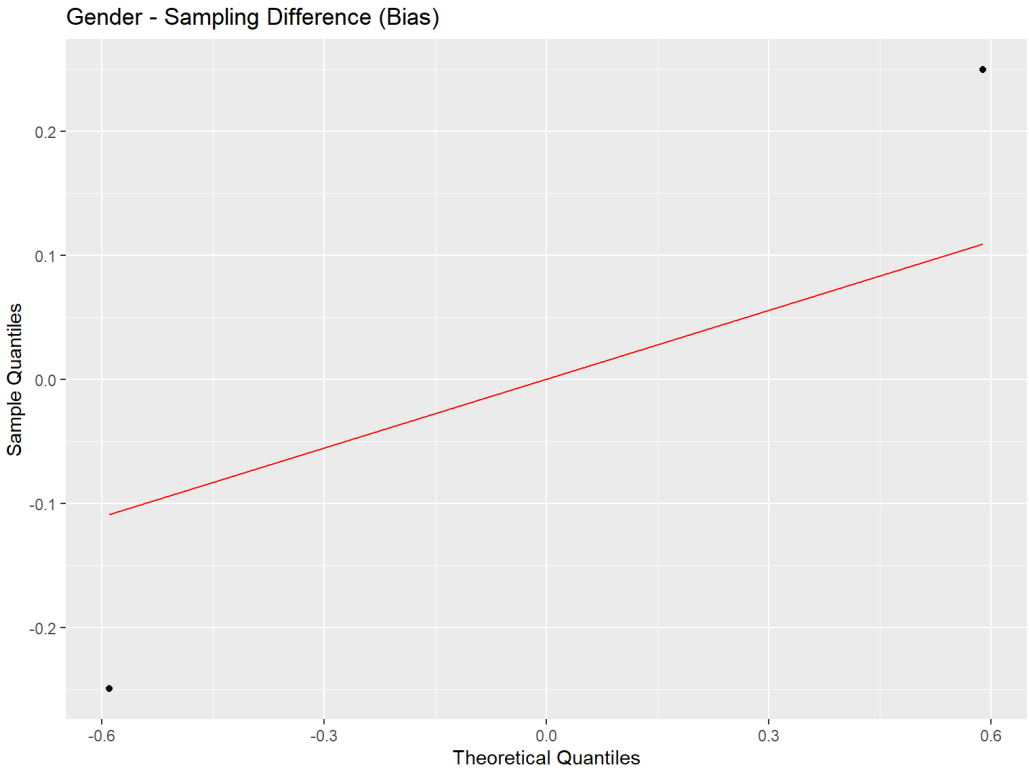
2 rows

```
knitr::kable(Check_Gender_Bias, caption ="Gender - Proportion Comparison")
```

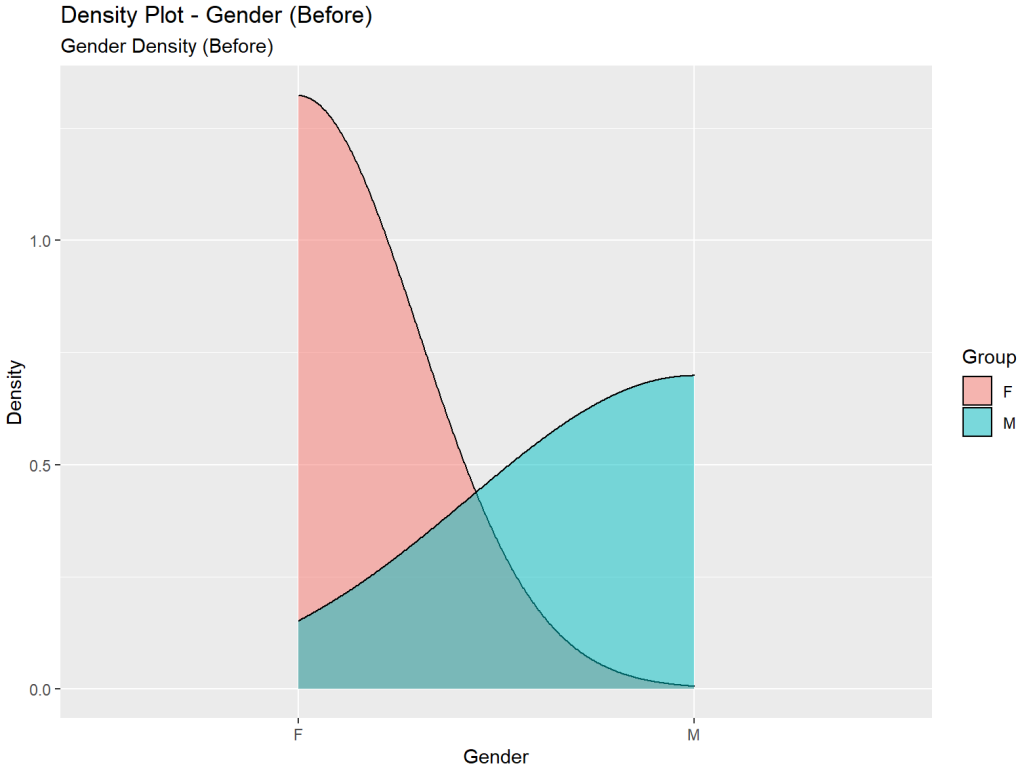
Gender - Proportion Comparison

WNW_Before	Group_A_Prop	Group_B_Prop	Difference
0.4324818	0.4909639	0.2416667	-0.2492972
0.5675182	0.5090361	0.7583333	0.2492972

```
gg_gender_plot <- ggplot(Check_Gender_Bias, aes(sample = Difference)) +
  stat_qq(aes(colour = Difference))+ stat_qq_line(colour = "red") +
  labs(title = "Gender - Sampling Difference (Bias)", x="Theoretical Quantiles", y="Sample Quantiles")
gg_gender_plot
```

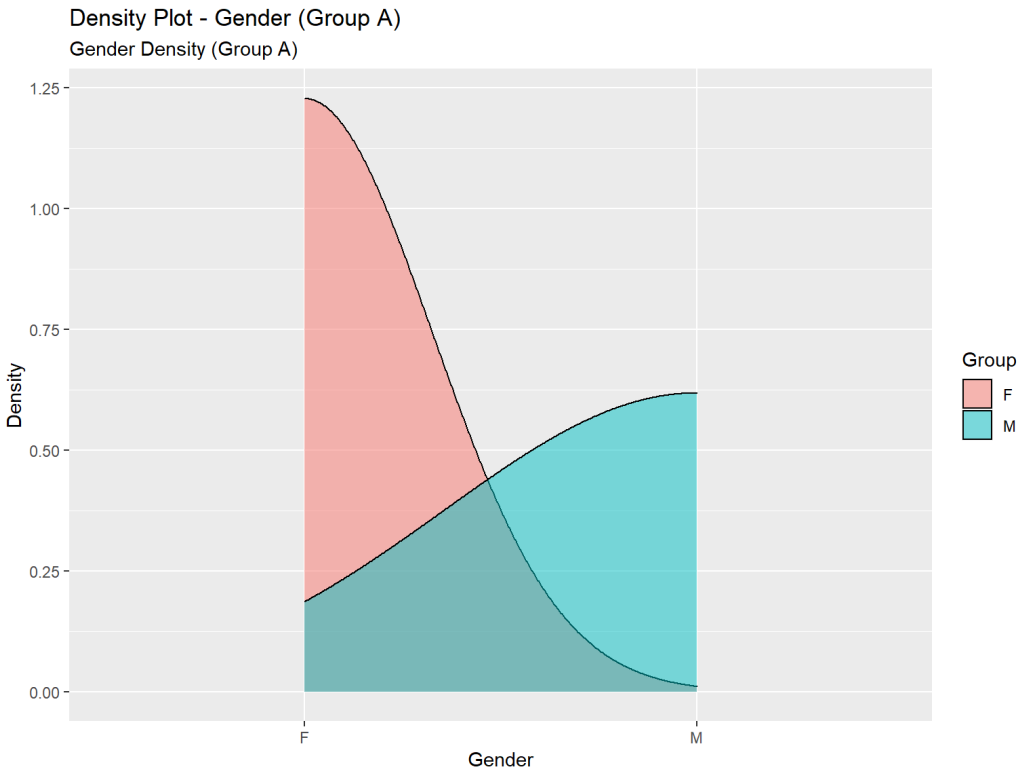



```
# Gender Plot - Density Plots
# Before trial
GG_gender_wnw <- ggplot(wnw_pre_change, aes(gender))
GG_gender_wnw <- GG_gender_wnw + geom_density(aes(fill = factor(gender)), alpha = 0.5) +
  labs(title = "Density Plot - Gender (Before)",
       subtitle = "Gender Density (Before)",
       x= "Gender", y = "Density",
       fill = "Group")
GG_gender_wnw
```

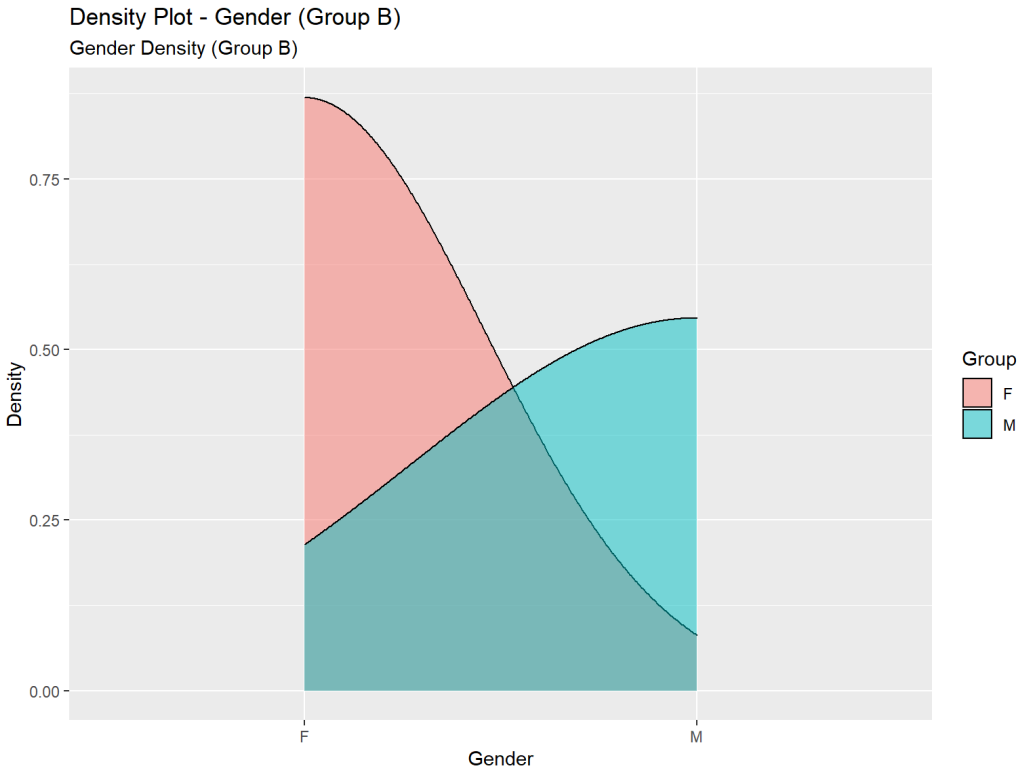



```
# Group A
GG_gender_A <- ggplot(wnw_post_A, aes(gender))
GG_gender_A <- GG_gender_A + geom_density(aes(fill = factor(gender)), alpha = 0.5) +
  labs(title = "Density Plot - Gender (Group A)",
```

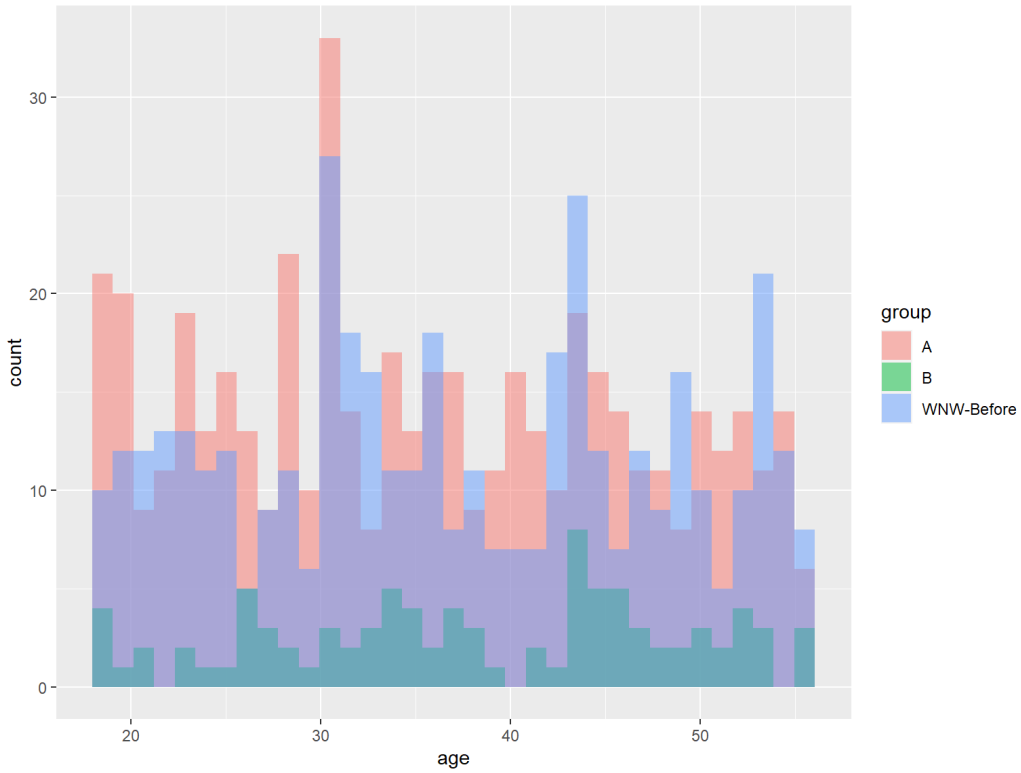
```
       subtitle = "Gender Density (Group A)",
       x= "Gender", y = "Density",
       fill = "Group")
GG_gender_A
```




```
# Group B
GG_gender_B <- ggplot(wnw_post_B, aes(gender))
GG_gender_B <- GG_gender_B + geom_density(aes(fill = factor(gender)), alpha = 0.5) +
  labs(title = "Density Plot - Gender (Group B)",
       subtitle = "Gender Density (Group B)",
       x= "Gender", y = "Density",
       fill = "Group")
GG_gender_B
```

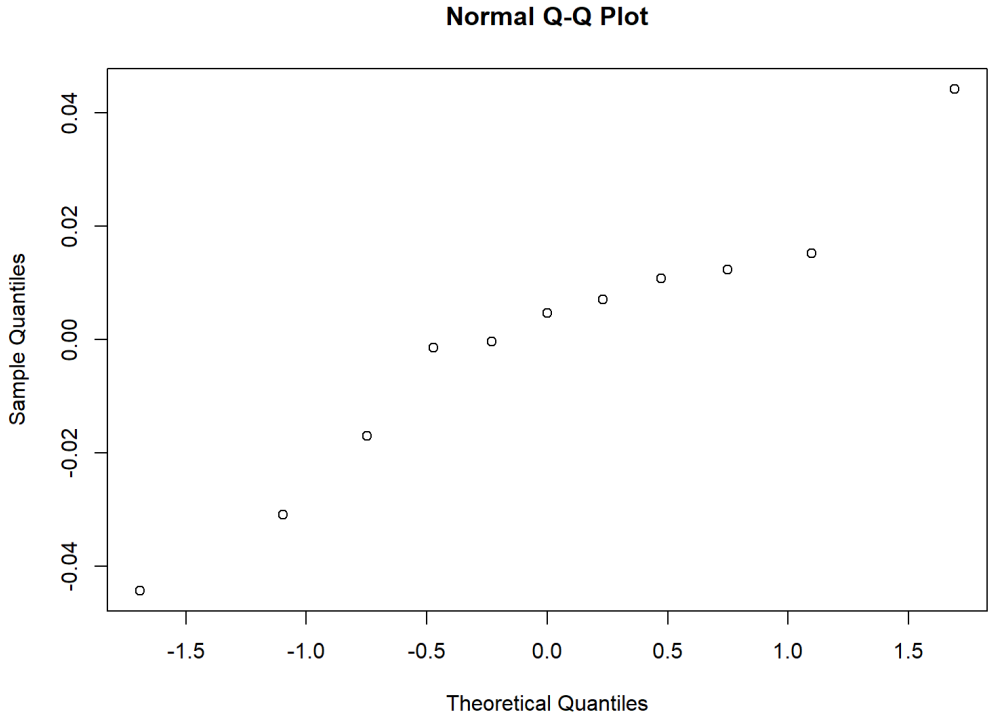


```
################################################################################
#AGE
WNW_graph_work <- wnw_df
WNW_graph_work$group[WNW_graph_work$date < "2022-07-18"] <- "WNW-Before"
WNW_graph_work$group <- as.factor(WNW_graph_work$group)

gg_Age_plot <- ggplot(WNW_graph_work, aes(x = age, fill = group)) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 35)
gg_Age_plot
```

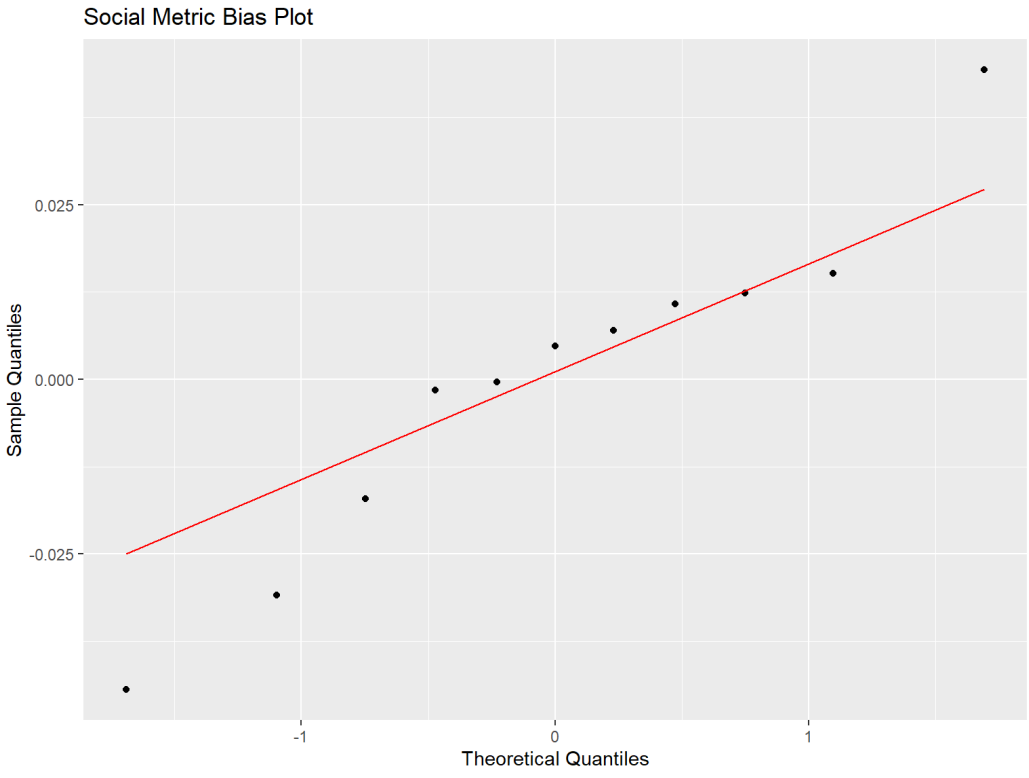



```
################################################################################
# Social Metric

# count the numbers in each demographic category based on the A/B group
check_a_social <- wnw_post_change %>%
  filter(group == 'A') %>%
  select(social_metric) %>%
  group_by(social_metric) %>%
  mutate(n_a = n()) %>%
  distinct()

check_b_social <- wnw_post_change %>%
  filter(group == 'B') %>%
  select(social_metric) %>%
  group_by(social_metric) %>%
  mutate(n_b = n()) %>%
  distinct()

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

# proportions in each demographic
check_a_social$p_a <- check_a_social$n_a / n_total_a
check_b_social$p_b <- check_b_social$n_b / n_total_b

# join on demo categories
social_bias_df <- inner_join(check_a_social, check_b_social)

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

```
# if there is no bias aside from sampling noise then the difference
# should be small and normally distributed
qqnorm(y = social_bias_df$difference)
```




```
gg_social_plot <- ggplot(social_bias_df, aes(sample = difference)) +
  stat_qq(aes(colour = difference))+ stat_qq_line(colour = "red") +
  labs(title = "Social Metric Bias Plot", x="Theoretical Quantiles", y="Sample Quantiles")
gg_social_plot
```



Hypothesis Testing

```
################################################################################
# Hypothesis Testing
################################################################################

Z_stat <- (GroupB_stats$mean - GroupA_stats$mean)/
  sqrt(((GroupB_stats$sd^2)/GroupB_stats$n)+((GroupA_stats$sd^2)/GroupA_stats$n))

Z_stat <- round(Z_stat, 3)

#p_value
p_value <- pnorm(abs(Z_stat), 0, 1, lower.tail = FALSE)
p_value
```

```
## [1] 0.002118205
```

```
################################################################################
# Effect Size
################################################################################
# Group A stats
mean_A <- GroupA_stats$mean
sd_A <- GroupA_stats$sd
n_A <- GroupA_stats$n

#Group B Stats
mean_B <- GroupB_stats$mean
n_B <- GroupB_stats$n
sd_B <- GroupB_stats$sd

# calculating pooled standard deviation:
# https://www.statology.org/pooled-standard-deviation-in-r/

pooled_sd <-sqrt(((n_A - 1)*sd_A^2 + (n_B-1)*sd_B^2)/(n_A+n_B-2))


#Calculation of d for 2 independent means.

d_hat<- round((mean_B -mean_A)/pooled_sd, 3)
# 0.3 effect size

################################################################################
# Sample size calculation - one sided test
################################################################################
alpha <- 0.05
z_alpha <- qnorm(alpha, mean = 0, sd = 1, lower.tail = FALSE)

sigma <- pooled_sd
effect <- d_hat
n_min = (z_alpha*sigma/effect)^2
ceiling(n_min)
```

```
## [1] 57
```

```
#57 people
################################################################################
# Alternative sample size calculation
power_t_test <- pwr.t.test(d = d_hat, power = 0.8, sig.level = 0.05, type = "two.sample", alternative = "greater")

sample_size2 <- ceiling(power_t_test$n)
# 139 people with power level of 0.8
```

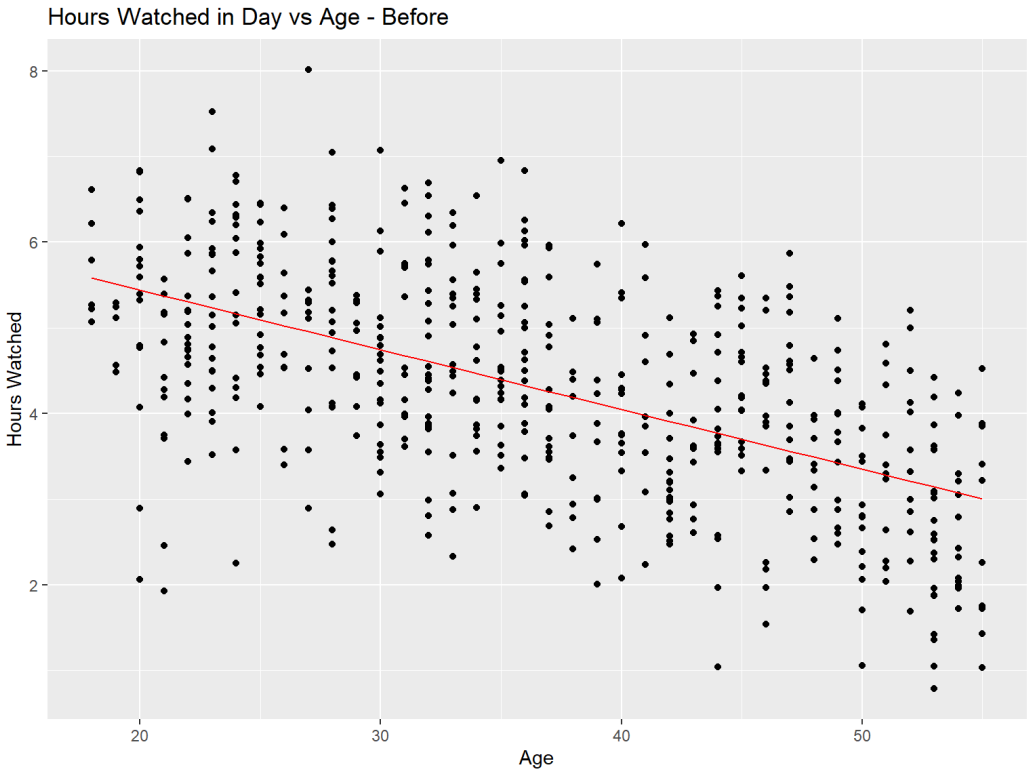
Code Chunk: Simple Linear Regression

```
################################################################################
# Regression Analysis
################################################################################
```

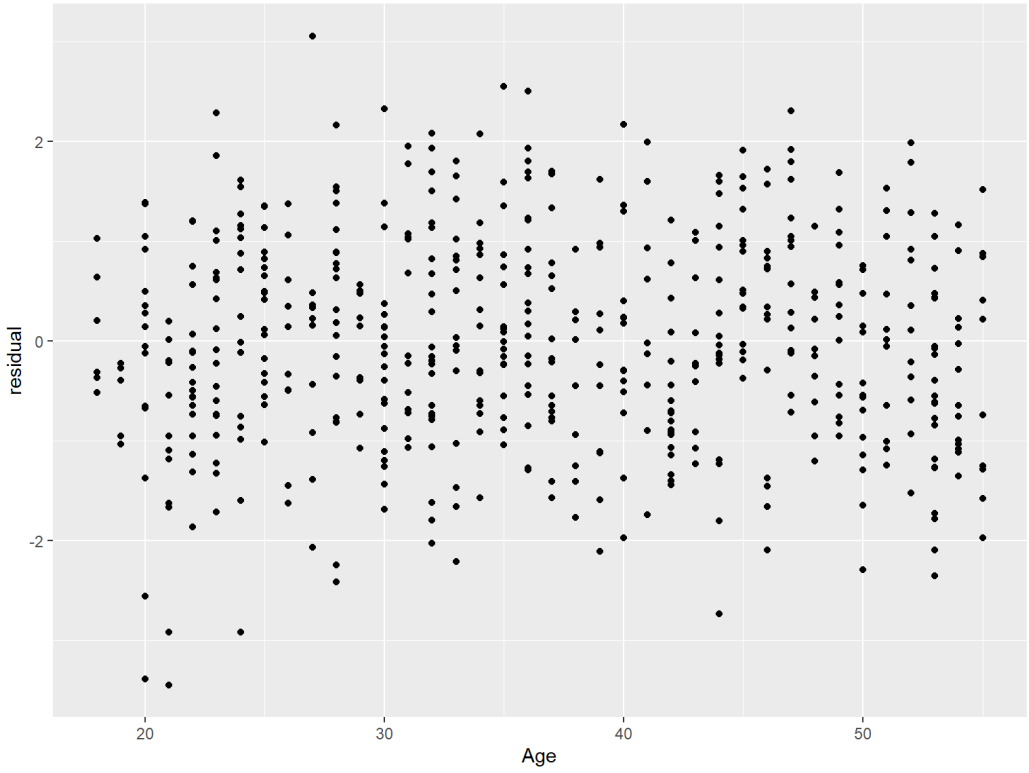
```
# Simple Linear Regression - Before
################################################################################

# create a model called "slr" for simple linear regression
slr = lm(hours_watched ~ age, data = wnw_pre_change)

# contains coeff [1]=intercept, [2]=slope; find with coef(fit)[1]
summary(slr)
```

```
##
## Call:
## lm(formula = hours_watched ~ age, data = wnw_pre_change)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4434 -0.7344 -0.0434  0.7761  3.0549
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.837467   0.164444   41.58   <2e-16 ***
## age         -0.069716   0.004336  -16.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.064 on 546 degrees of freedom
## Multiple R-squared:  0.3214, Adjusted R-squared:  0.3201
## F-statistic: 258.5 on 1 and 546 DF,  p-value: < 2.2e-16
```

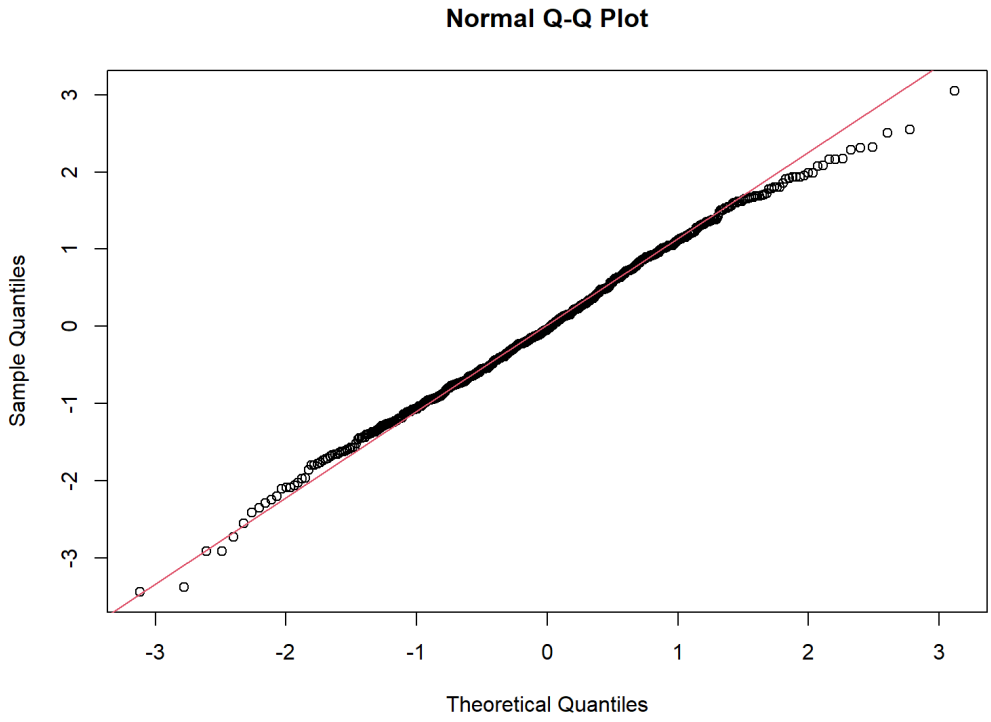
```
# Extracting co-efficients
a0 <- coef(slr)[1]
a1 <- coef(slr)[2]


# setup x variable with range of interest
x_slr <- seq(min(wnw_pre_change$age), max(wnw_pre_change$age), 1)

# calculate the fitting function yfit based on the coefficients from the model
y_slr <- a0 + a1 * x_slr

gg_linear1 <- ggplot()
gg_linear1 <- gg_linear1 + geom_point(aes(x = wnw_pre_change$age, y = wnw_pre_change$hours_watched))
gg_linear1 <- gg_linear1 + geom_line(aes(x = x_slr, y = y_slr), colour = 'red')
gg_linear1 <- gg_linear1 + labs(x = 'Age', y = 'Hours Watched', title = 'Hours Watched in Day vs Age - Before')
gg_linear1
```




```
# fit to each value of x
wnw_pre_change$hours_watched_hat <- a0 + a1 * wnw_pre_change$age

# calculate the residual
wnw_pre_change$error <- wnw_pre_change$hours_watched - wnw_pre_change$hours_watched_hat

gg <- ggplot()
gg <- gg + geom_point(aes(x = wnw_pre_change$age, y = wnw_pre_change$error))
gg <- gg + labs(x = 'Age', y = 'residual')
gg
```




```
# Check that errors are normally distributed
qqnorm(wnw_pre_change$error)
qqline(wnw_pre_change$error, col = 2)
```




```
shapiro.test(wnw_pre_change$error)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wnw_pre_change$error
## W = 0.99679, p-value = 0.3508
```

```
# p-value = 0.3508, thus can not reject null hypothesis.



# total number of degrees of freedom
n = length(wnw_pre_change$age)

df_SST <- n- 1

# One independent variable in the model, therefore df associated with SSR is 1
df_SSR <- 1

# degrees of freedom for error term
df_SSE <- df_SST - df_SSR

# calculate the mean of the observed data
y_mean_hours_watched <- mean(wnw_pre_change$hours_watched)

# calculate the sum of squares terms
SSR <- sum( (wnw_pre_change$hours_watched_hat - y_mean_hours_watched)^2 )
SSE <- sum( (wnw_pre_change$hours_watched - wnw_pre_change$hours_watched_hat)^2 )

# calculate the F statistic
f_score_slr <- (SSR/df_SSR) / (SSE/df_SSE)

# convert to p value
p_value_slr <- pf(f_score_slr, df_SSR, df_SSE, lower.tail = FALSE)
print(paste('F =', f_score_slr))
```

```
## [1] "F = 258.549455866175"
```

```
print(paste('p-value =', p_value_slr))
```

```
## [1] "p-value = 6.5415934591528e-48"
```

```
# compare to the in-built function in R
summary(slr)
```

```
##
## Call:
## lm(formula = hours_watched ~ age, data = wnw_pre_change)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4434 -0.7344 -0.0434  0.7761  3.0549
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.837467   0.164444   41.58   <2e-16 ***
## age         -0.069716   0.004336  -16.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.064 on 546 degrees of freedom
## Multiple R-squared:  0.3214, Adjusted R-squared:  0.3201
## F-statistic: 258.5 on 1 and 546 DF,  p-value: < 2.2e-16
```

```
# Read off Multiple R-squared in the summary or calculate as follows
SST <- SSR + SSE
R_squared <- SSR/SST
R_squared
```

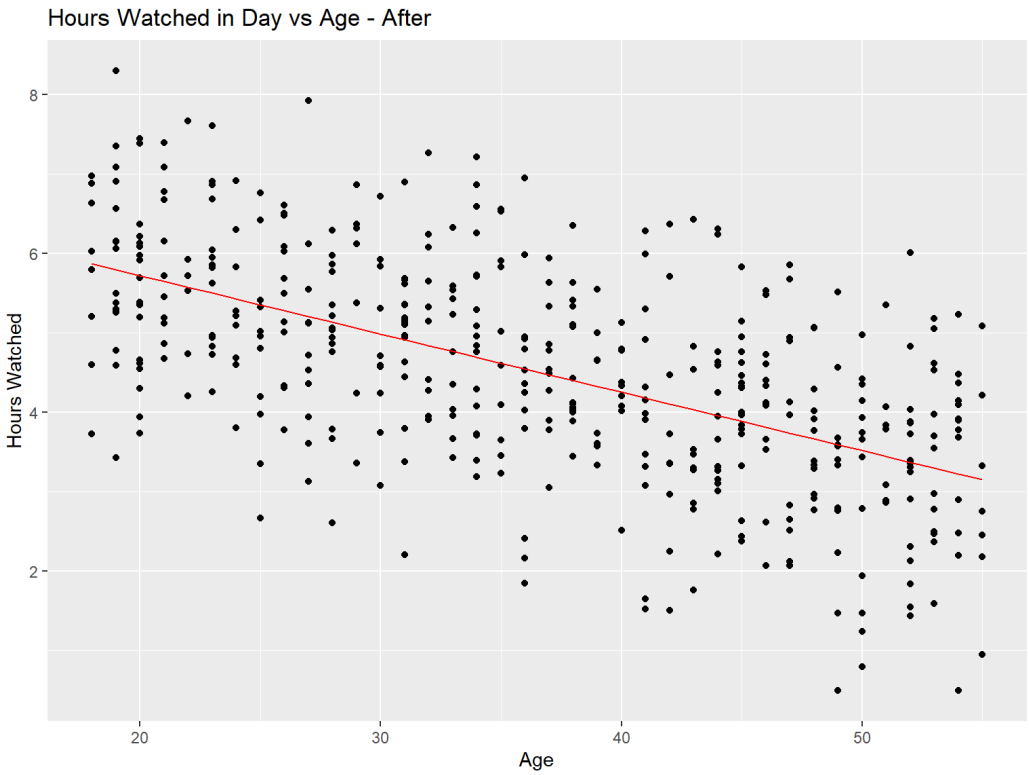
```
## [1] 0.3213593
```

```
################################################################################
# Simple Linear Regression - After
################################################################################

# create a model called "slr" for simple linear regression
slr2 = lm(hours_watched ~ age, data = wnw_post_change)

# contains coeff [1]=intercept, [2]=slope; find with coef(fit)[1]
summary(slr2)
```

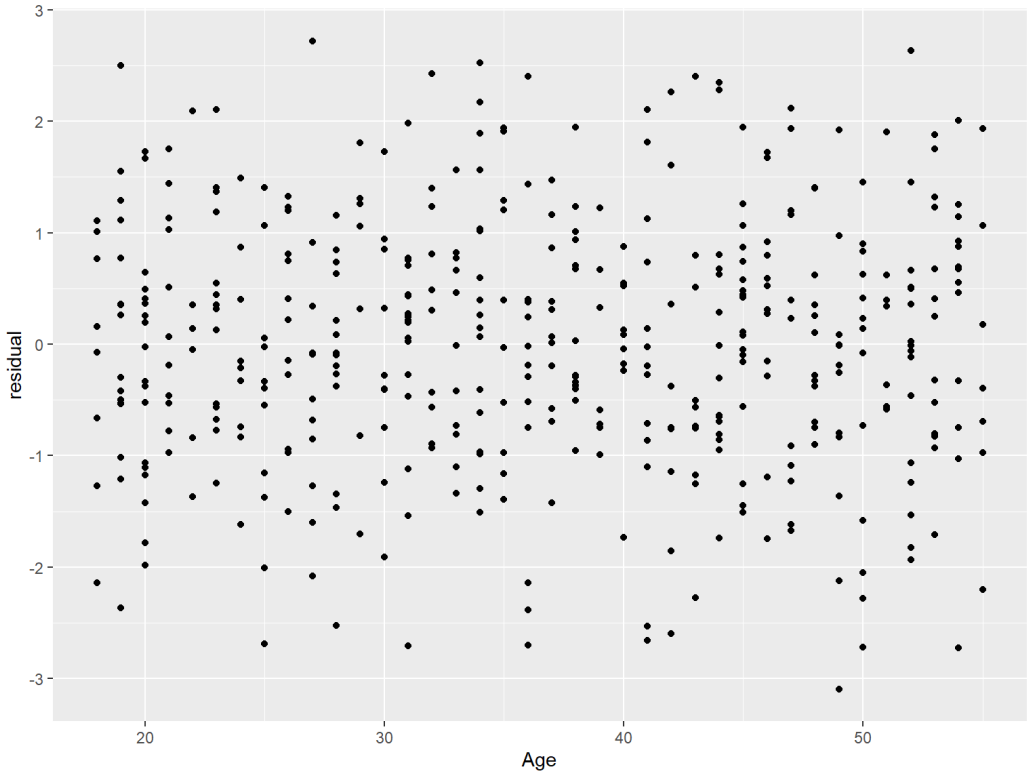
```
##
## Call:
## lm(formula = hours_watched ~ age, data = wnw_post_change)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.09423 -0.74791  0.01871  0.74290  2.71942
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.194277   0.183594   39.19   <2e-16 ***
## age         -0.073470   0.004815  -15.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.118 on 450 degrees of freedom
## Multiple R-squared:  0.3409, Adjusted R-squared:  0.3395
## F-statistic: 232.8 on 1 and 450 DF,  p-value: < 2.2e-16
```

```
# Extracting co-efficients
a0_2 <- coef(slr2)[1]
a1_2 <- coef(slr2)[2]
```

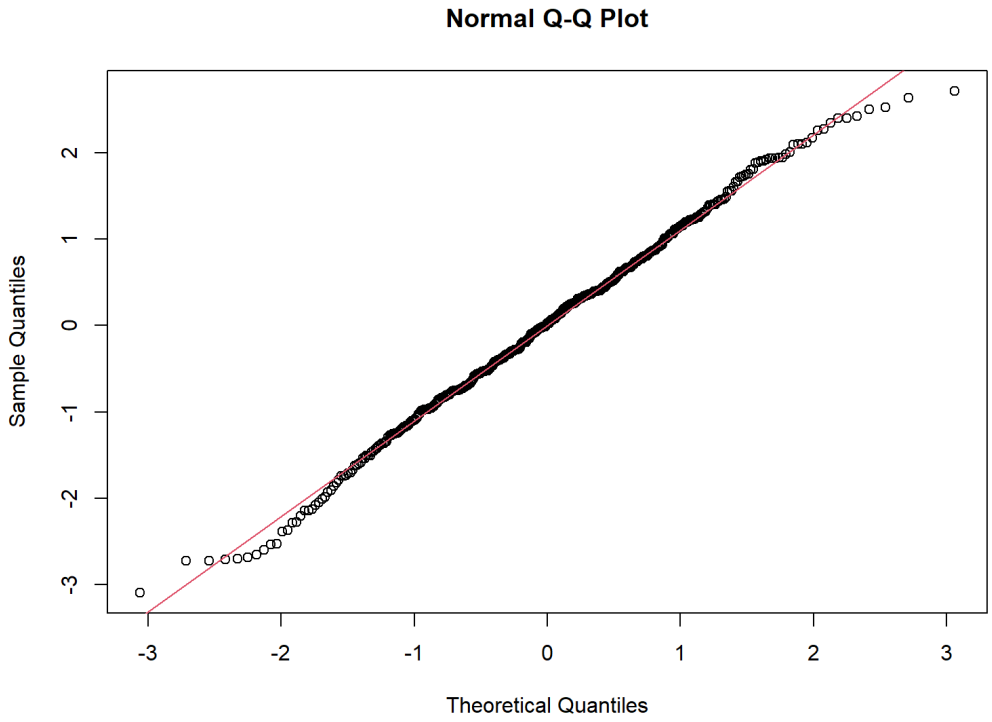
```
# setup x variable with range of interest
x_slr2 <- seq(min(wnw_post_change$age), max(wnw_post_change$age), 1)

# calculate the fitting function y fit based on the coefficients from the model
y_slr2 <- a0_2 + a1_2 * x_slr2

gg_linear <- ggplot()
gg_linear <- gg_linear + geom_point(aes(x = wnw_post_change$age, y = wnw_post_change$hours_watched))
gg_linear <- gg_linear + geom_line(aes(x = x_slr2, y = y_slr2), colour = 'red')
gg_linear <- gg_linear + labs(x = 'Age', y = 'Hours Watched', title = 'Hours Watched in Day vs Age - After')
gg_linear
```




```
# fit to each value of x
wnw_post_change$hours_watched_hat <- a0_2 + a1_2 * wnw_post_change$age

# calculate the residual
wnw_post_change$error <- wnw_post_change$hours_watched - wnw_post_change$hours_watched_hat

gg_residual <- ggplot()
gg_residual <- gg_residual + geom_point(aes(x = wnw_post_change$age, y = wnw_post_change$error))
gg_residual <- gg_residual + labs(x = 'Age', y = 'residual')
gg_residual
```




```
# Check that errors are normally distributed
qqnorm(wnw_post_change$error)
qqline(wnw_post_change$error, col = 2)
```




```
shapiro.test(wnw_post_change$error)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wnw_post_change$error
## W = 0.99578, p-value = 0.2667
```

```
# p-value = 0.3508, thus can not reject null hypothesis.



# total number of degrees of freedom
n2 = length(wnw_post_change$age)

df_SST2 <- n2- 1

# One independent variable in the model, therefore df associated with SSR is 1
df_SSR2 <- 1

# degrees of freedom for error term
df_SSE2 <- df_SST2 - df_SSR2

# calculate the mean of the observed data
y_mean_hours_watched2 <- mean(wnw_post_change$hours_watched)

# calculate the sum of squares terms
SSR2 <- sum( (wnw_post_change$hours_watched_hat - y_mean_hours_watched2)^2 )
SSE2 <- sum( (wnw_post_change$hours_watched - wnw_post_change$hours_watched_hat)^2 )

# calculate the F statistic
f_score_slr2 <- (SSR2/df_SSR2) / (SSE2/df_SSE2)

# convert to p value
p_value_slr2 <- pf(f_score_slr2, df_SSR2, df_SSE2, lower.tail = FALSE)
print(paste('F =', f_score_slr2))
```

```
## [1] "F = 232.787010658601"
```

```
print(paste('p-value =', p_value_slr2))
```

```
## [1] "p-value = 1.16292577969047e-42"
```

```
# compare to the in-built function in R
summary(slr2)
```

```
##
## Call:
## lm(formula = hours_watched ~ age, data = wnw_post_change)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.09423 -0.74791  0.01871  0.74290  2.71942
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.194277   0.183594   39.19   <2e-16 ***
## age         -0.073470   0.004815  -15.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.118 on 450 degrees of freedom
## Multiple R-squared:  0.3409, Adjusted R-squared:  0.3395
## F-statistic: 232.8 on 1 and 450 DF,  p-value: < 2.2e-16
```

```
# Read off Multiple R-squared in the summary or calculate as follows
SST2 <- SSR2 + SSE2
R_squared2 <- SSR2/SST2
R_squared2
```

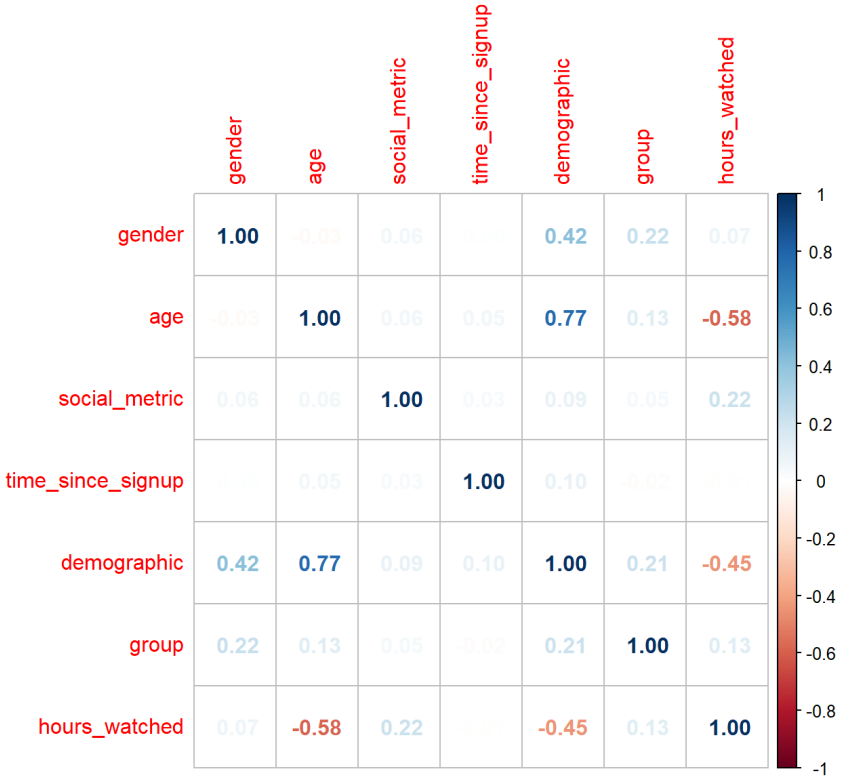
```
## [1] 0.3409365
```

Code Chunk: Multiple Regression

```
################################################################################
#Multiple Regression
################################################################################
# Dataframe set up

wnw2_df <- wnw_df
wnw2_df$date <- as.Date(wnw2_df$date, format = "%d-%b")



wnw2_df$gender[wnw2_df$gender == "F"] <- 0
wnw2_df$gender[wnw2_df$gender == "M"] <- 1
wnw2_df$gender <- as.integer(wnw2_df$gender)

wnw2_df$group[wnw2_df$group == "A"] <- 0
wnw2_df$group[wnw2_df$group == "B"] <- 1
wnw2_df$group <- as.integer(wnw2_df$group)


# New data frame after recommendation change for regression analysis.
#Post Trial Data Frame
wnw_post_2 <- wnw2_df %>%
  filter(date > "2022-07-17")

wnw_post_2 <- select(wnw_post_2, gender, age, social_metric, time_since_signup,demographic, group, hours_watched)
str(wnw_post_2)
```

```
## 'data.frame':    452 obs. of  7 variables:
##  $ gender           : int  0 1 0 1 0 0 0 0 1 0 ...
##  $ age              : int  39 45 50 39 18 52 47 32 21 20 ...
##  $ social_metric    : int  5 7 2 1 9 2 6 8 3 10 ...
##  $ time_since_signup: num  14.8 2.4 6.6 18.7 10.5 5.3 13.8 18.7 3.7 20.3 ...
##  $ demographic      : int  3 4 3 4 1 3 3 1 2 1 ...
##  $ group            : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ hours_watched    : num  3.74 4 3.66 3.58 6.64 3.36 2.83 5.15 4.87 6.37 ...
```

```
M = cor(wnw_post_2)
corrplot(M, method = "number")


# Pre Trial Dataframe

wnw_pre_2<- wnw2_df %>%
  filter(date < "2022-07-18")

wnw_pre_2 <- select(wnw_pre_2, gender, age, social_metric, time_since_signup,demographic,  hours_watched)

M2 = cor(wnw_pre_2)
corrplot(M, method = "number")
```




```
################################################################################
# Multi-variate Regresssion - Pre-Trial
################################################################################
# Pre-Trial - Model 1 - all variables are included
################################################################################

wnw_pre_model1 <- lm(hours_watched ~ gender + age + social_metric + time_since_signup + demographic, data = wnw_pre_2)
summary(wnw_pre_model1)
```

```
##
## Call:
## lm(formula = hours_watched ~ gender + age + social_metric + time_since_signup +
##     demographic, data = wnw_pre_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4934 -0.6581 -0.0302  0.7531  2.8232
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.272872   0.222084  28.245  < 2e-16 ***
## gender             0.049151   0.132060   0.372    0.710
## age               -0.064105   0.008232  -7.788 3.50e-14 ***
## social_metric      0.084755   0.015044   5.634 2.83e-08 ***
## time_since_signup  0.002400   0.006122   0.392    0.695
## demographic       -0.041685   0.086393  -0.483    0.630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.038 on 542 degrees of freedom
## Multiple R-squared:  0.3591, Adjusted R-squared:  0.3532
## F-statistic: 60.75 on 5 and 542 DF,  p-value: < 2.2e-16
```

```
#Coefficients:
#                   Estimate Std. Error t value Pr(>|t|)
#(Intercept)        6.272872   0.222084  28.245  < 2e-16 ***
#gender             0.049151   0.132060   0.372    0.710
#age               -0.064105   0.008232  -7.788 3.50e-14 ***
#social_metric      0.084755   0.015044   5.634 2.83e-08 ***
#time_since_signup  0.002400   0.006122   0.392    0.695
#demographic       -0.041685   0.086393  -0.483    0.630
```

```
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 1.038 on 542 degrees of freedom
#Multiple R-squared:  0.3591,   Adjusted R-squared:  0.3532
#F-statistic: 60.75 on 5 and 542 DF,  p-value: < 2.2e-16

################################################################################
# Pre-Trial - Model 2: Remove Time since Sign-up from model
################################################################################
wnw_pre_model2 <- lm(hours_watched ~ gender + age + social_metric + demographic, data = wnw_pre_2)
summary(wnw_pre_model2)
```

```
##
## Call:
## lm(formula = hours_watched ~ gender + age + social_metric + demographic,
##     data = wnw_pre_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5189 -0.6615 -0.0392  0.7425  2.7956
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.302204   0.208941  30.163  < 2e-16 ***
## gender         0.047669   0.131903   0.361    0.718
## age           -0.064175   0.008223  -7.804 3.10e-14 ***
## social_metric  0.084754   0.015032   5.638 2.76e-08 ***
## demographic   -0.040367   0.086260  -0.468    0.640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.037 on 543 degrees of freedom
## Multiple R-squared:  0.359,  Adjusted R-squared:  0.3542
## F-statistic: 76.02 on 4 and 543 DF,  p-value: < 2.2e-16
```

```
#Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
#(Intercept)    6.302204   0.208941  30.163  < 2e-16 ***
#gender         0.047669   0.131903   0.361    0.718
#age           -0.064175   0.008223  -7.804 3.10e-14 ***
#social_metric  0.084754   0.015032   5.638 2.76e-08 ***
#demographic   -0.040367   0.086260  -0.468    0.640
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 1.037 on 543 degrees of freedom
#Multiple R-squared:  0.359,    Adjusted R-squared:  0.3542
#F-statistic: 76.02 on 4 and 543 DF,  p-value: < 2.2e-16


################################################################################
#Model 3 - Remove Gender as not adding value
################################################################################
wnw_pre_model3 <- lm(hours_watched ~ age + social_metric + demographic, data = wnw_pre_2)
summary(wnw_pre_model3)
```

```
##
## Call:
## lm(formula = hours_watched ~ age + social_metric + demographic,
##     data = wnw_pre_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5147 -0.6768 -0.0356  0.7512  2.8126
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.337195   0.185006  34.254  < 2e-16 ***
```

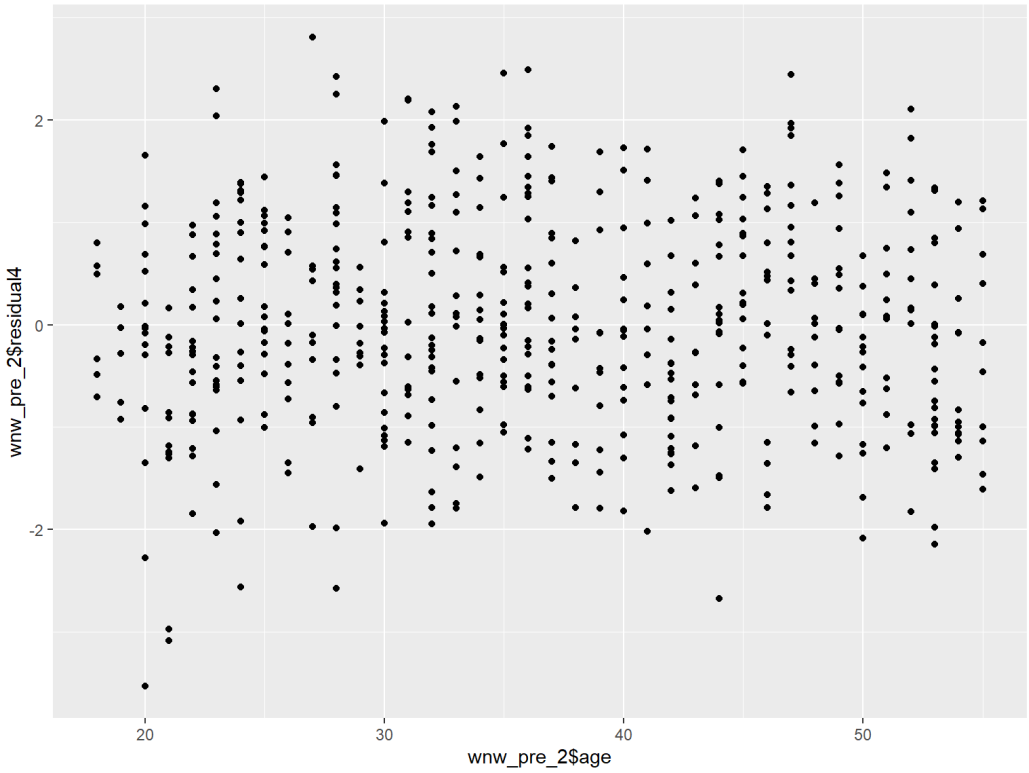
```
## age           -0.066004   0.006477 -10.190  < 2e-16 ***
## social_metric  0.084650   0.015017   5.637 2.78e-08 ***
## demographic   -0.017465   0.058475  -0.299    0.765
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.036 on 544 degrees of freedom
## Multiple R-squared:  0.3588, Adjusted R-squared:  0.3553
## F-statistic: 101.5 on 3 and 544 DF,  p-value: < 2.2e-16
```

```
#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)    6.337195   0.185006  34.254  < 2e-16 ***
#age           -0.066004   0.006477 -10.190  < 2e-16 ***
#social_metric  0.084650   0.015017   5.637 2.78e-08 ***
#demographic   -0.017465   0.058475  -0.299    0.765
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 1.036 on 544 degrees of freedom
#Multiple R-squared:  0.3588,   Adjusted R-squared:  0.3553
#F-statistic: 101.5 on 3 and 544 DF,  p-value: < 2.2e-16



## Gender has a large p-value and we can observe that it does not add anything.

################################################################################
# Pre- Trial - Model 4 :
# Remove demographic - doesn't add anything significant and is correlated with age
################################################################################

wnw_pre_model4 <- lm(hours_watched ~ age + social_metric, data = wnw_pre_2)
summary(wnw_pre_model4)
```

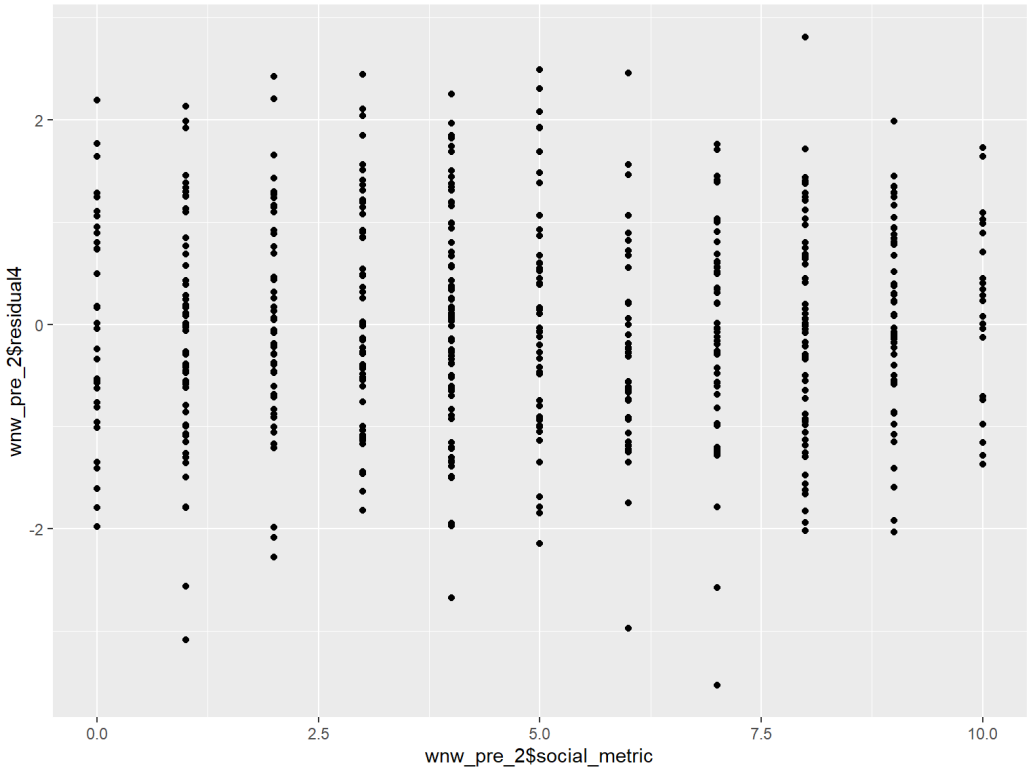
```
##
## Call:
## lm(formula = hours_watched ~ age + social_metric, data = wnw_pre_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5280 -0.6728 -0.0386  0.7521  2.8098
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.346532   0.182193  34.834  < 2e-16 ***
## age           -0.067466   0.004237 -15.921  < 2e-16 ***
## social_metric  0.084403   0.014982   5.634 2.83e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.035 on 545 degrees of freedom
## Multiple R-squared:  0.3587, Adjusted R-squared:  0.3564
## F-statistic: 152.4 on 2 and 545 DF,  p-value: < 2.2e-16
```

```
#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)    6.346532   0.182193  34.834  < 2e-16 ***
#age           -0.067466   0.004237 -15.921  < 2e-16 ***
#social_metric  0.084403   0.014982   5.634 2.83e-08 ***
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 1.035 on 545 degrees of freedom
#Multiple R-squared:  0.3587,   Adjusted R-squared:  0.3564
#F-statistic: 152.4 on 2 and 545 DF,  p-value: < 2.2e-16
```

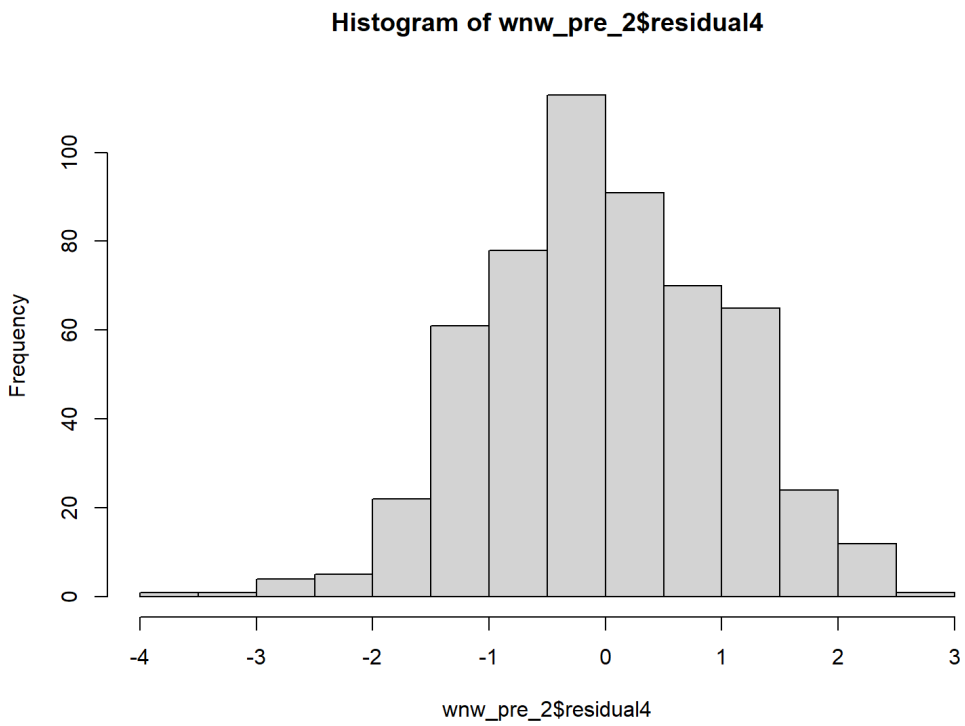
```
# Examine Residuals

wnw_pre_2$model4 <- wnw_pre_model4$coefficients[1] +
             wnw_pre_model4$coefficients[2]*wnw_pre_2$age +
             wnw_pre_model4$coefficients[3]*wnw_pre_2$social_metric


wnw_pre_2$residual4 <- wnw_pre_2$hours_watched - wnw_pre_2$model4

qplot(wnw_pre_2$age,wnw_pre_2$residual4)
```



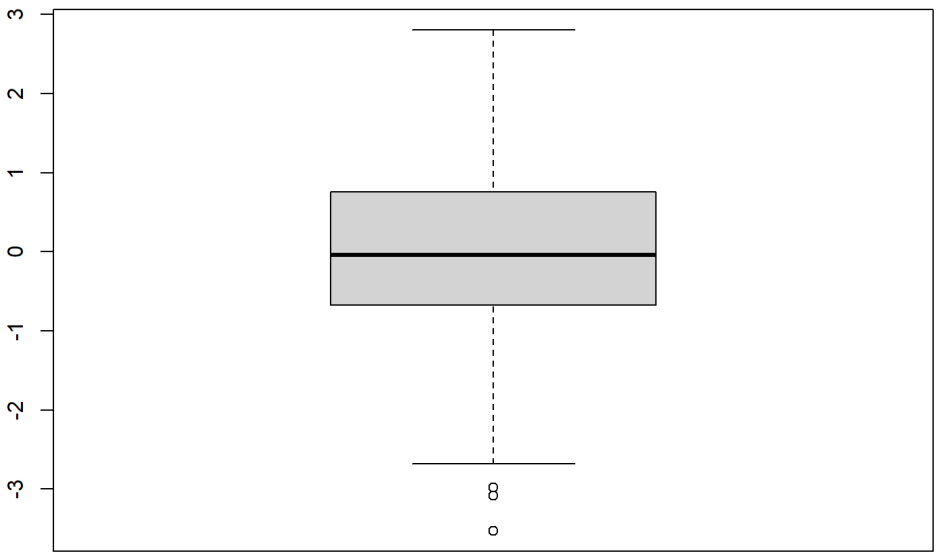

```
qplot(wnw_pre_2$social_metric, wnw_pre_2$residual4)
```



```
hist(wnw_pre_2$residual4)
```




```
wnw_pre_boxplot <- boxplot(wnw_pre_2$residual4)
```




```
boxplot.stats(wnw_pre_2$residual4)
```

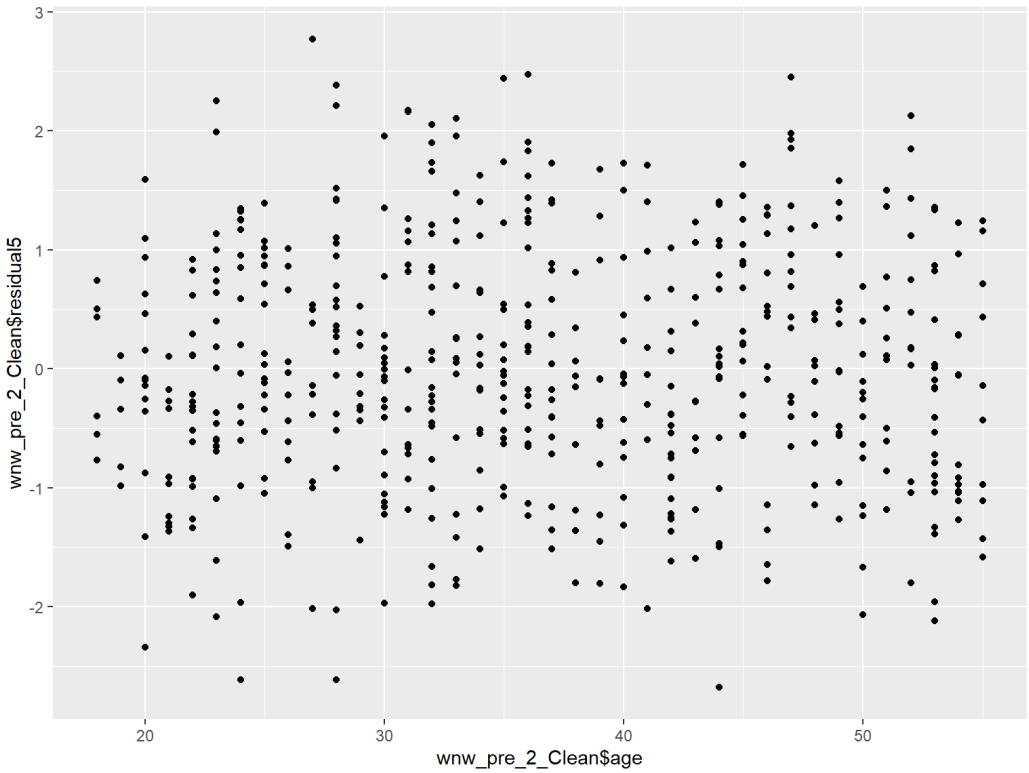
```
## $stats
## [1] -2.67562628 -0.67662370 -0.03862317  0.75517271  2.80983395
##
## $n
## [1] 548
##
## $conf
```

```
## [1] -0.13526120  0.05801486
##
## $out
## [1] -3.084142 -2.976158 -3.528027
```

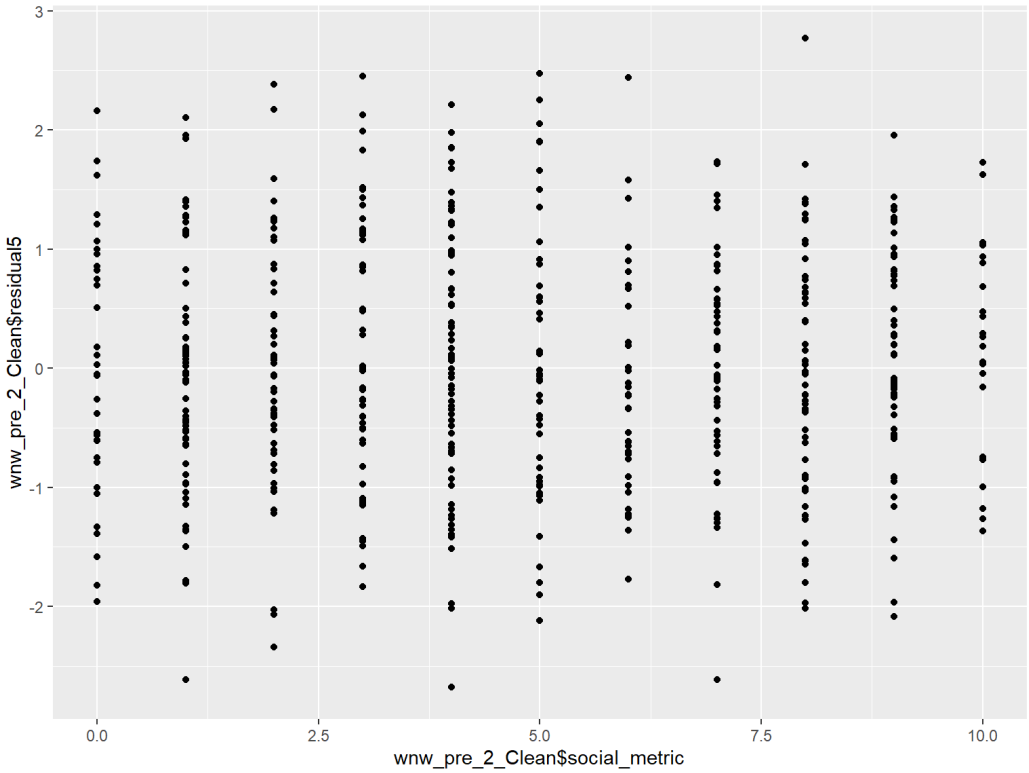
```
#We have 3 outliers present



################################################################################
#                Removing outliers from data
################################################################################

## Removed outliers
outliers_pre <- wnw_pre_boxplot$out

#Create clean data frame without outliers.
wnw_pre_2_Clean <- wnw_pre_2
wnw_pre_2_Clean <- wnw_pre_2_Clean[-which(wnw_pre_2_Clean$residual4 %in% outliers_pre),]


wnw_pre_model5 <- lm(hours_watched ~ age + social_metric, data = wnw_pre_2_Clean)
summary(wnw_pre_model5)
```

```
##
## Call:
## lm(formula = hours_watched ~ age + social_metric, data = wnw_pre_2_Clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67475 -0.66553 -0.05537  0.73719  2.77079
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.46395    0.17908  36.096  < 2e-16 ***
## age           -0.07006    0.00416 -16.842  < 2e-16 ***
## social_metric  0.08336    0.01465   5.691 2.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.01 on 542 degrees of freedom
## Multiple R-squared:  0.3837, Adjusted R-squared:  0.3814
## F-statistic: 168.7 on 2 and 542 DF,  p-value: < 2.2e-16
```

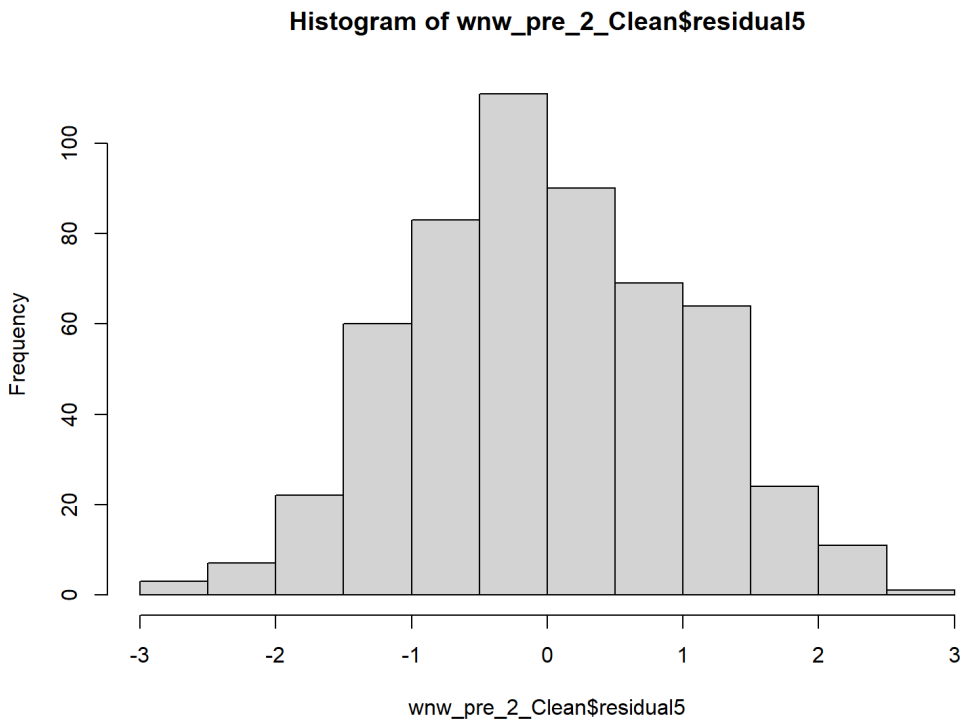
```
################################################################################
# Pre - Trial - MODEL 5 RESIDUALS
################################################################################
# Examine Residuals

wnw_pre_2_Clean$model5 <- wnw_pre_model5$coefficients[1] +
            wnw_pre_model5$coefficients[2]*wnw_pre_2_Clean$age +
            wnw_pre_model5$coefficients[3]*wnw_pre_2_Clean$social_metric


wnw_pre_2_Clean$residual5 <- wnw_pre_2_Clean$hours_watched - wnw_pre_2_Clean$model5

qplot(wnw_pre_2_Clean$age,wnw_pre_2_Clean$residual5)
```



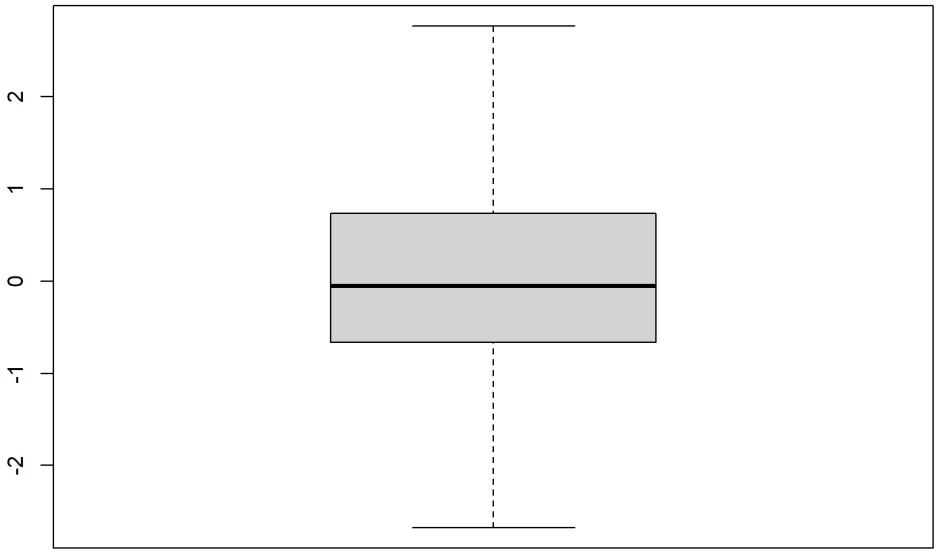

```
qplot(wnw_pre_2_Clean$social_metric, wnw_pre_2_Clean$residual5)
```




```
hist(wnw_pre_2_Clean$residual5)
```




```
wnw_Pre_boxplot2 <- boxplot(wnw_pre_2_Clean$residual5)
```




```
boxplot.stats(wnw_pre_2_Clean$residual5)
```

```
## $stats
## [1] -2.67474654 -0.66552800 -0.05536559  0.73719437  2.77079416
##
## $n
## [1] 545
##
## $conf
## [1] -0.15030151  0.03957033
##
```

```
## $out
## numeric(0)
```

```
#############################################################################
# Multi-variate Regresssion - After
#############################################################################
# Model 1 - all variables are included
#############################################################################

wnw_post2_model1 <- lm(hours_watched ~ gender + age + social_metric + time_since_signup + demographic + group, data =
        wnw_post_2)
summary(wnw_post2_model1)
```

```
##
## Call:
## lm(formula = hours_watched ~ gender + age + social_metric + time_since_signup +
##     demographic + group, data = wnw_post_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75596 -0.60838 -0.00599  0.66732  2.76992
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.468445   0.223421  28.952  < 2e-16 ***
## gender             0.150426   0.137564   1.093   0.2748
## age               -0.065597   0.008866  -7.399 6.90e-13 ***
## social_metric      0.110930   0.016091   6.894 1.87e-11 ***
## time_since_signup  0.005748   0.006793   0.846   0.3979
## demographic       -0.167337   0.097407  -1.718   0.0865 .
## group              0.640662   0.113736   5.633 3.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.029 on 445 degrees of freedom
## Multiple R-squared:  0.4476, Adjusted R-squared:  0.4401
## F-statistic: 60.09 on 6 and 445 DF,  p-value: < 2.2e-16
```

```
#Coefficients:
#                   Estimate Std. Error t value Pr(>|t|)
#(Intercept)        6.468445   0.223421  28.952  < 2e-16 ***
#gender             0.150426   0.137564   1.093   0.2748
#age               -0.065597   0.008866  -7.399 6.90e-13 ***
#social_metric      0.110930   0.016091   6.894 1.87e-11 ***
#time_since_signup  0.005748   0.006793   0.846   0.3979
#demographic       -0.167337   0.097407  -1.718   0.0865 .
#group              0.640662   0.113736   5.633 3.15e-08 ***
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 1.029 on 445 degrees of freedom
#Multiple R-squared:  0.4476,   Adjusted R-squared:  0.4401
#F-statistic: 60.09 on 6 and 445 DF,  p-value: < 2.2e-16



# Time since sign-up don't add anything valuable.


#############################################################################

wnw_post2_model1 <- lm(hours_watched ~ gender + age + social_metric + demographic + group, data = wnw_post_2)
summary(wnw_post2_model1)
```

```
##
## Call:
## lm(formula = hours_watched ~ gender + age + social_metric + demographic +
##     group, data = wnw_post_2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7885 -0.6167  0.0016  0.6678  2.8036
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.534291   0.209364  31.210  < 2e-16 ***
## gender         0.141641   0.137128   1.033    0.302
## age           -0.066152   0.008839  -7.484 3.87e-13 ***
## social_metric  0.111253   0.016081   6.918 1.59e-11 ***
## demographic   -0.157588   0.096693  -1.630    0.104
## group          0.638008   0.113657   5.613 3.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.029 on 446 degrees of freedom
## Multiple R-squared:  0.4467, Adjusted R-squared:  0.4405
## F-statistic: 72.02 on 5 and 446 DF,  p-value: < 2.2e-16
```

```
# Time since sign-up don't add anything valuable.


################################################################################
#Model 2 - remove time since sign up, Gender
################################################################################
wnw_post2_model2 <- lm(hours_watched ~ age + social_metric + demographic + group, data = wnw_post_2)
summary(wnw_post2_model2)
```

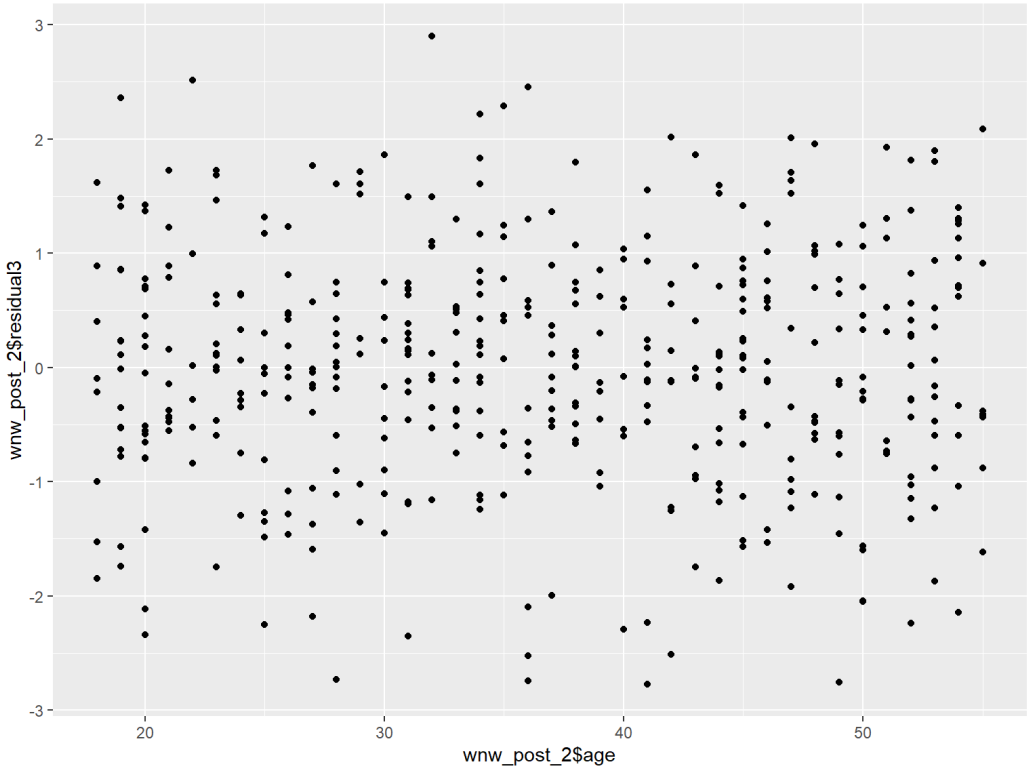
```
##
## Call:
## lm(formula = hours_watched ~ age + social_metric + demographic +
##     group, data = wnw_post_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.79815 -0.60440 -0.01698  0.65963  2.88553
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.636093   0.184729  35.923  < 2e-16 ***
## age           -0.071699   0.007021 -10.212  < 2e-16 ***
## social_metric  0.111192   0.016082   6.914 1.63e-11 ***
## demographic   -0.089828   0.071041  -1.264    0.207
## group          0.656094   0.112308   5.842 9.95e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.029 on 447 degrees of freedom
## Multiple R-squared:  0.4454, Adjusted R-squared:  0.4404
## F-statistic: 89.74 on 4 and 447 DF,  p-value: < 2.2e-16
```

```
#Residuals:
#    Min      1Q  Median      3Q     Max
#-2.7885 -0.6167  0.0016  0.6678  2.8036
#
#Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
#(Intercept)    6.534291   0.209364  31.210  < 2e-16 ***
#gender         0.141641   0.137128   1.033    0.302
#age           -0.066152   0.008839  -7.484 3.87e-13 ***
#social_metric  0.111253   0.016081   6.918 1.59e-11 ***
#demographic   -0.157588   0.096693  -1.630    0.104
#group          0.638008   0.113657   5.613 3.49e-08 ***
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 1.029 on 446 degrees of freedom
#Multiple R-squared:  0.4467,   Adjusted R-squared:  0.4405
#F-statistic: 72.02 on 5 and 446 DF,  p-value: < 2.2e-16
```

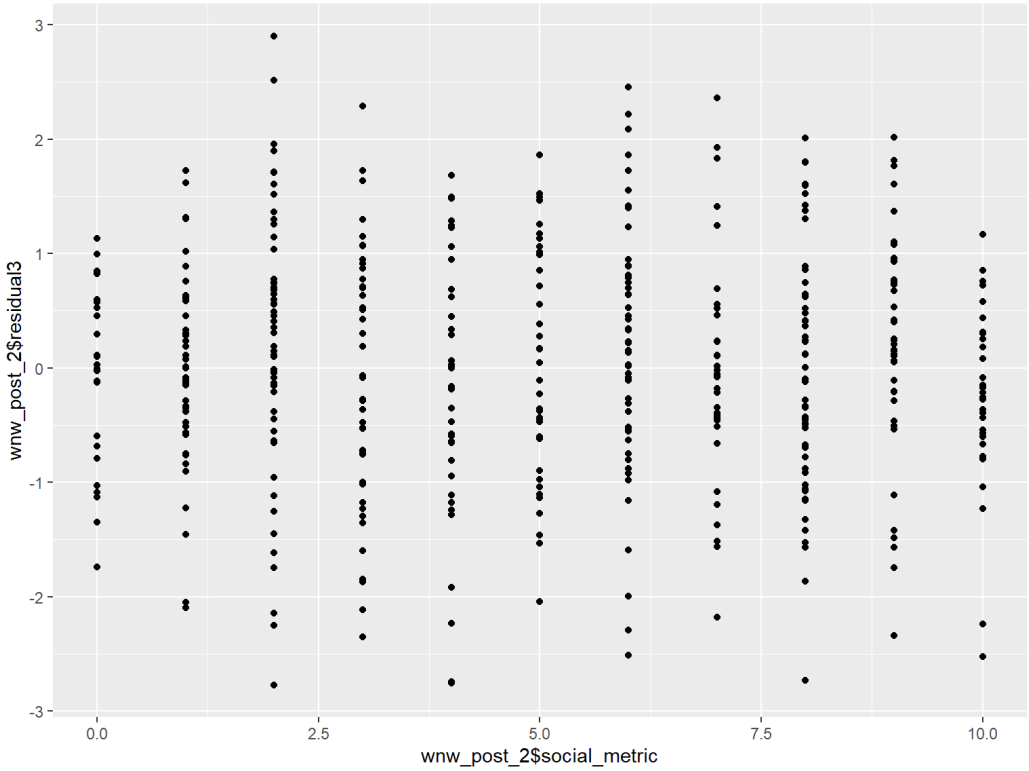
```
## Gender has a large p-value and we can observe that it does not add anything.

#############################################################################
# Model 3 :
# Remove demographic - doesn't add anything significant and is correlated with age
#############################################################################

wnw_post2_model3 <- lm(hours_watched ~ age + social_metric + group, data = wnw_post_2)
summary(wnw_post2_model3)
```

```
##
## Call:
## lm(formula = hours_watched ~ age + social_metric + group, data = wnw_post_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.77062 -0.59759  0.00196  0.68910  2.89958
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.66430    0.18350  36.318  < 2e-16 ***
## age           -0.07854    0.00448 -17.529  < 2e-16 ***
## social_metric  0.10964    0.01605   6.833 2.73e-11 ***
## group          0.63204    0.11076   5.706 2.10e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.03 on 448 degrees of freedom
## Multiple R-squared:  0.4434, Adjusted R-squared:  0.4397
## F-statistic:   119 on 3 and 448 DF,  p-value: < 2.2e-16
```

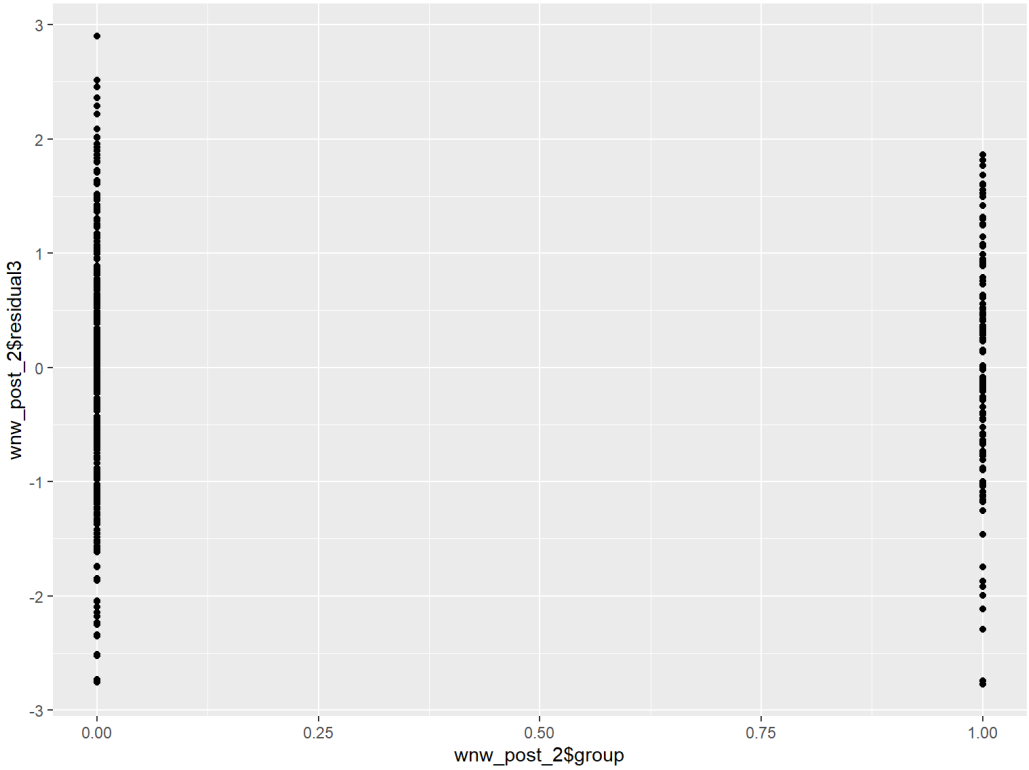
```
# Examine Residuals

wnw_post_2$model3 <- wnw_post2_model3$coefficients[1] +
             wnw_post2_model3$coefficients[2]*wnw_post_2$age +
             wnw_post2_model3$coefficients[3]*wnw_post_2$social_metric +
             wnw_post2_model3$coefficients[4]*wnw_post_2$group

wnw_post_2$residual3 <- wnw_post_2$hours_watched - wnw_post_2$model3

qplot(wnw_post_2$age,wnw_post_2$residual3)
```



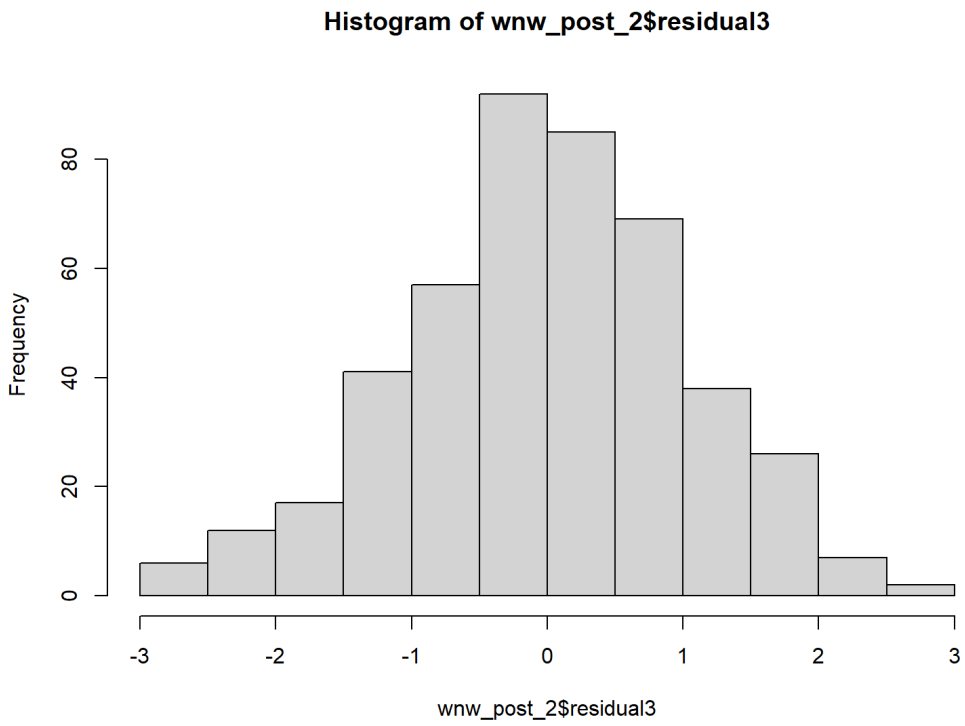

```
qplot(wnw_post_2$social_metric, wnw_post_2$residual3)
```



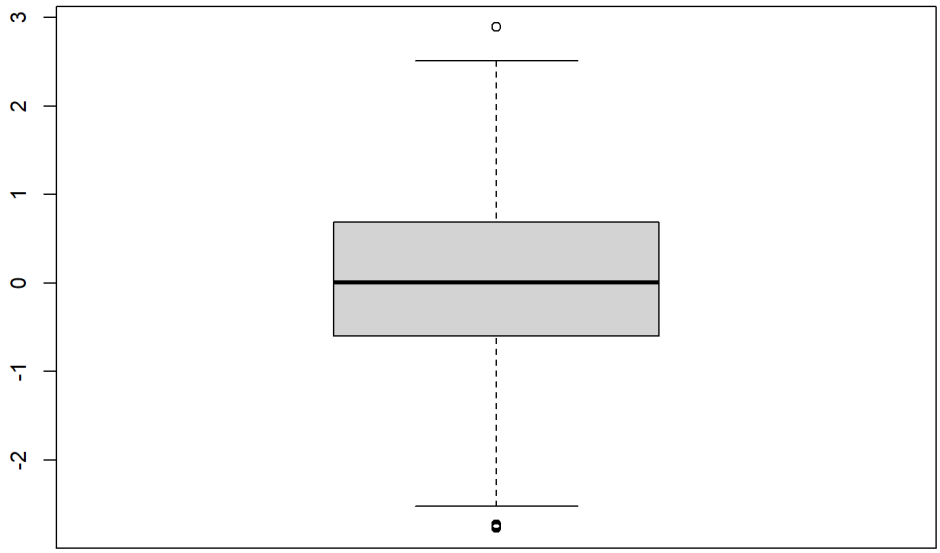

```
qplot(wnw_post_2$group, wnw_post_2$residual3)
```




```
hist(wnw_post_2$residual3)
```




```
wnw_boxplot <- boxplot(wnw_post_2$residual3)
```




```
boxplot.stats(wnw_post_2$residual3)
```

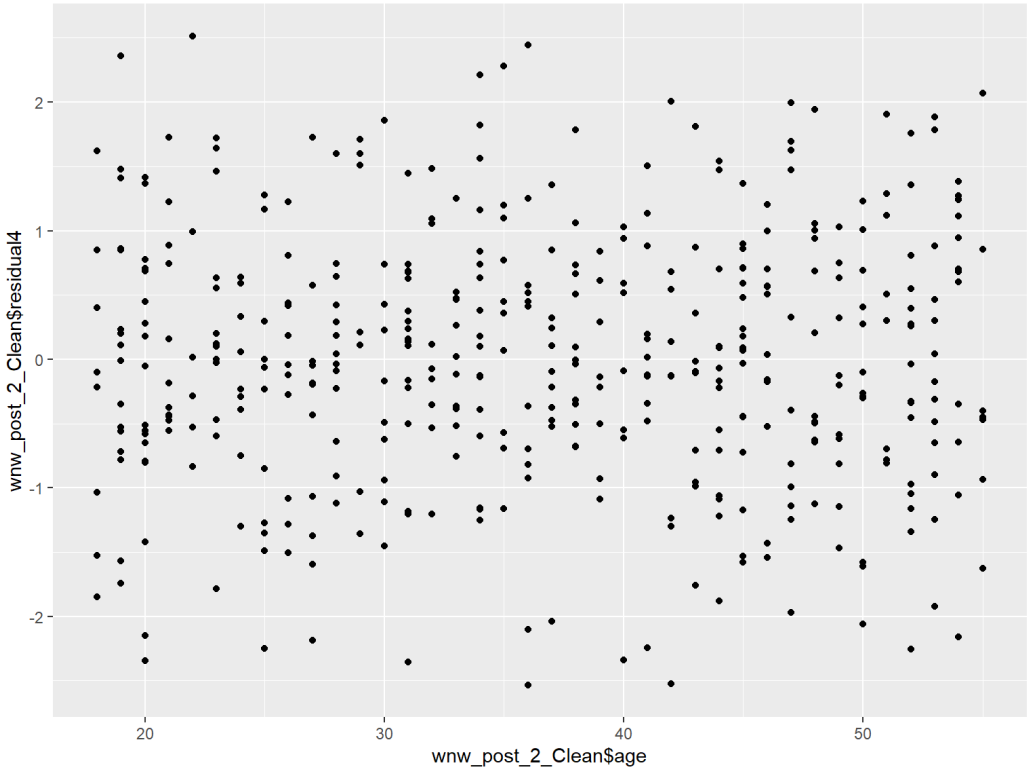
```
## $stats
## [1] -2.523410700 -0.597715238  0.001957336  0.690347081  2.514218067
##
## $n
## [1] 452
##
## $conf
## [1] -0.09376752  0.09768220
##
## $out
## [1] -2.754581  2.899583 -2.732418 -2.742591 -2.770623
```

```
#We have 5 outliers present

############## RESULTS MODEL 3

#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)    6.66430    0.18350  36.318  < 2e-16 ***
#age           -0.07854    0.00448 -17.529  < 2e-16 ***
#social_metric  0.10964    0.01605   6.833 2.73e-11 ***
#group          0.63204    0.11076   5.706 2.10e-08 ***
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 1.03 on 448 degrees of freedom
#Multiple R-squared:  0.4434,   Adjusted R-squared:  0.4397
#F-statistic:   119 on 3 and 448 DF,  p-value: < 2.2e-16


################################################################################
#                 Removing outliers from data
################################################################################

## Removed outliers
outliers <- wnw_boxplot$out

#Create clean data frame without outliers.
wnw_post_2_Clean <- wnw_post_2
wnw_post_2_Clean <- wnw_post_2_Clean[-which(wnw_post_2_Clean$residual3 %in% outliers),]
```

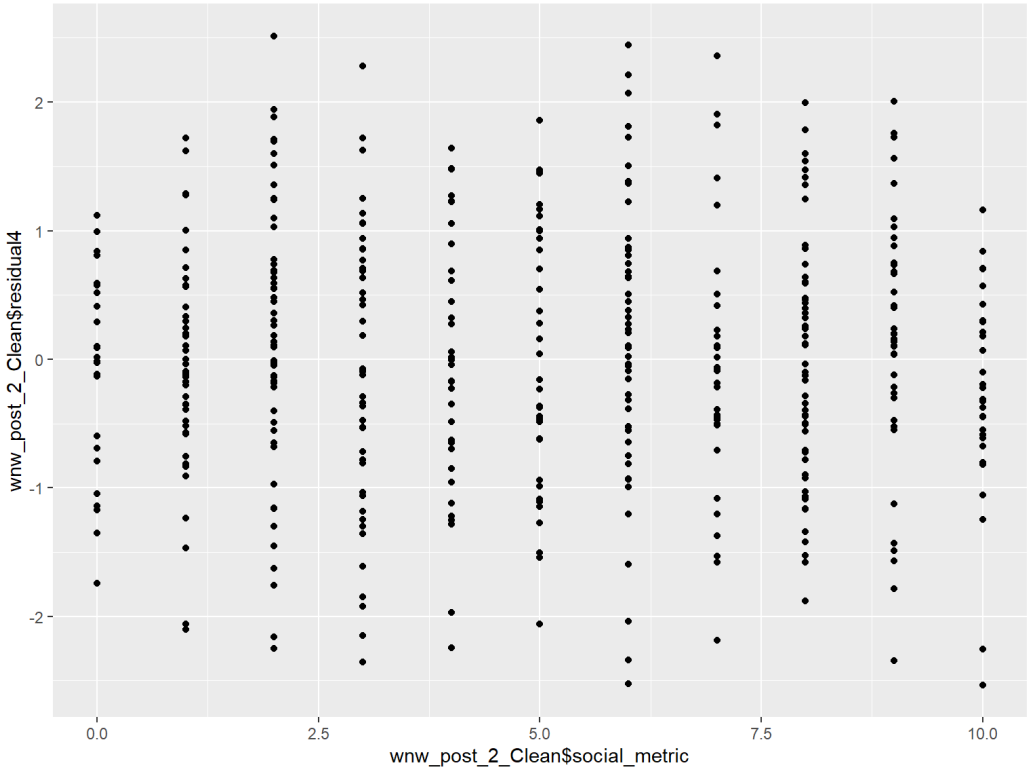
```
wnw_post2_model4 <- lm(hours_watched ~ age + social_metric + group, data = wnw_post_2_Clean)
summary(wnw_post2_model4)
```

```
##
## Call:
## lm(formula = hours_watched ~ age + social_metric + group, data = wnw_post_2_Clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.53334 -0.60517 -0.00707  0.68160  2.51360
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.654254   0.177180  37.557  < 2e-16 ***
## age           -0.078086   0.004328 -18.042  < 2e-16 ***
## social_metric  0.110017   0.015521   7.088 5.39e-12 ***
## group          0.669274   0.107529   6.224 1.12e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9923 on 443 degrees of freedom
## Multiple R-squared:  0.4627, Adjusted R-squared:  0.459
## F-statistic: 127.1 on 3 and 443 DF,  p-value: < 2.2e-16
```

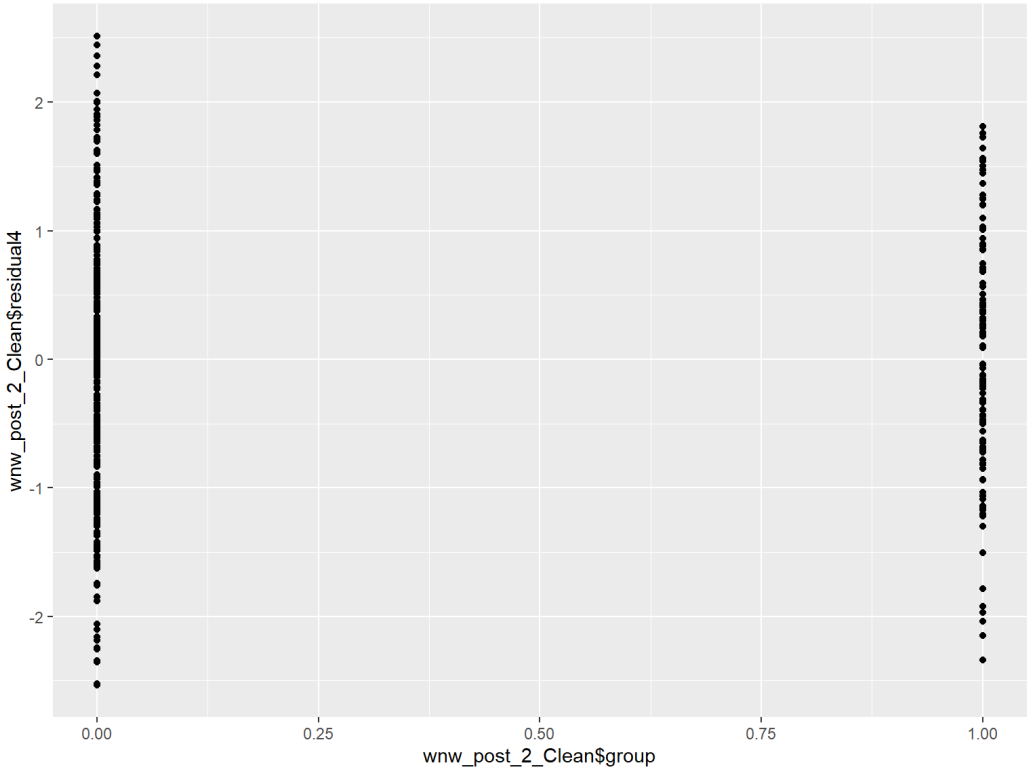
```
################################################################################
############### RESULTS Model 4


#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)    6.654254   0.177180  37.557  < 2e-16 ***
#age           -0.078086   0.004328 -18.042  < 2e-16 ***
#social_metric  0.110017   0.015521   7.088 5.39e-12 ***
#group          0.669274   0.107529   6.224 1.12e-09 ***
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#Residual standard error: 0.9923 on 443 degrees of freedom
#Multiple R-squared:  0.4627,   Adjusted R-squared:  0.459
#F-statistic: 127.1 on 3 and 443 DF,  p-value: < 2.2e-16

################################################################################
# MODEL 4 RESIDUALS
################################################################################
# Examine Residuals

wnw_post_2_Clean$model4 <- wnw_post2_model4$coefficients[1] +
            wnw_post2_model4$coefficients[2]*wnw_post_2_Clean$age +
            wnw_post2_model4$coefficients[3]*wnw_post_2_Clean$social_metric +
            wnw_post2_model4$coefficients[4]*wnw_post_2_Clean$group

wnw_post_2_Clean$residual4 <- wnw_post_2_Clean$hours_watched - wnw_post_2_Clean$model4

qplot(wnw_post_2_Clean$age,wnw_post_2_Clean$residual4)
```



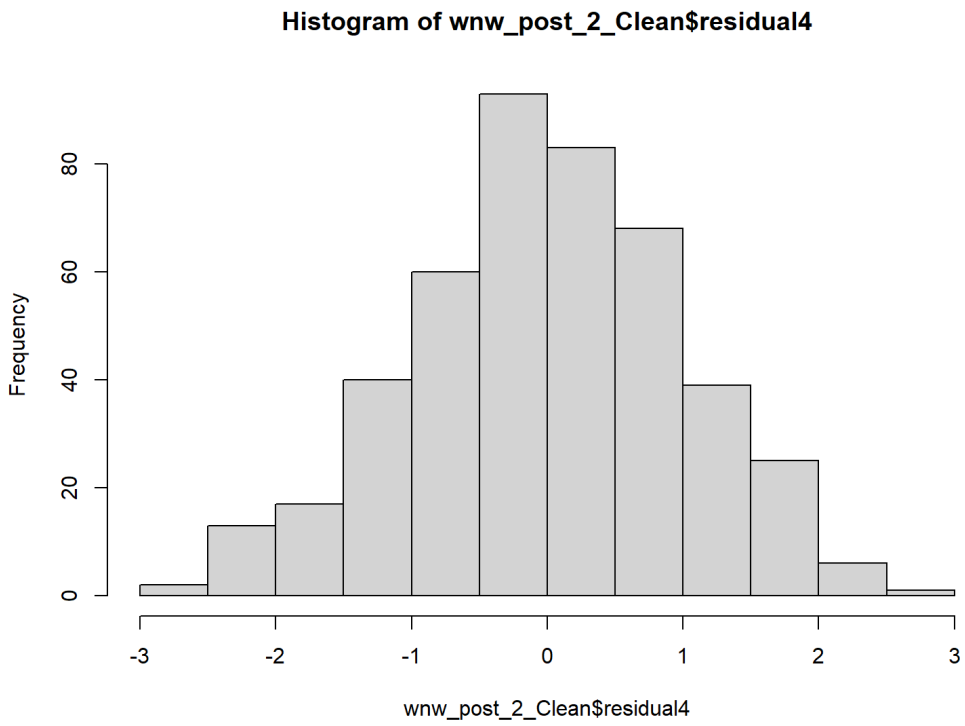

```
qplot(wnw_post_2_Clean$social_metric, wnw_post_2_Clean$residual4)
```



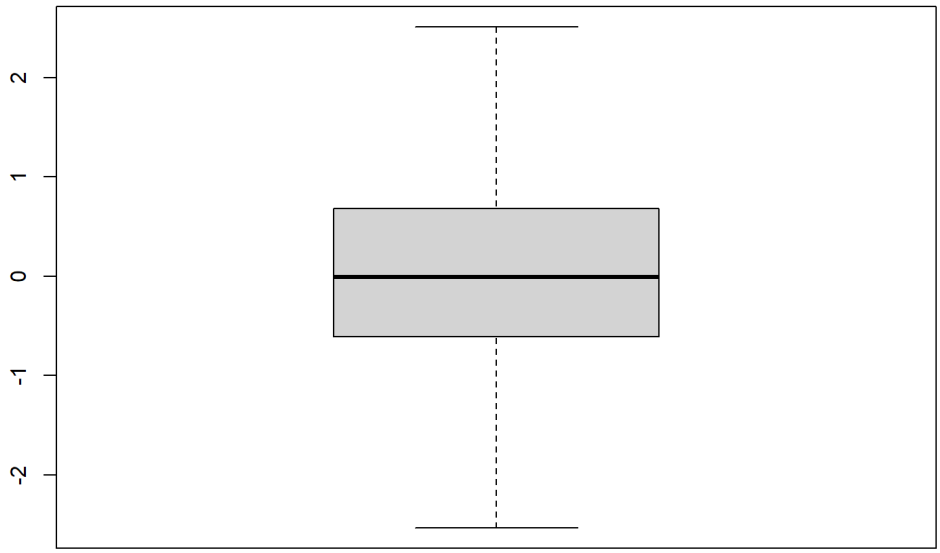

```
qplot(wnw_post_2_Clean$group, wnw_post_2_Clean$residual4)
```




```
hist(wnw_post_2_Clean$residual4)
```




```
wnw_boxplot2 <- boxplot(wnw_post_2_Clean$residual4)
```

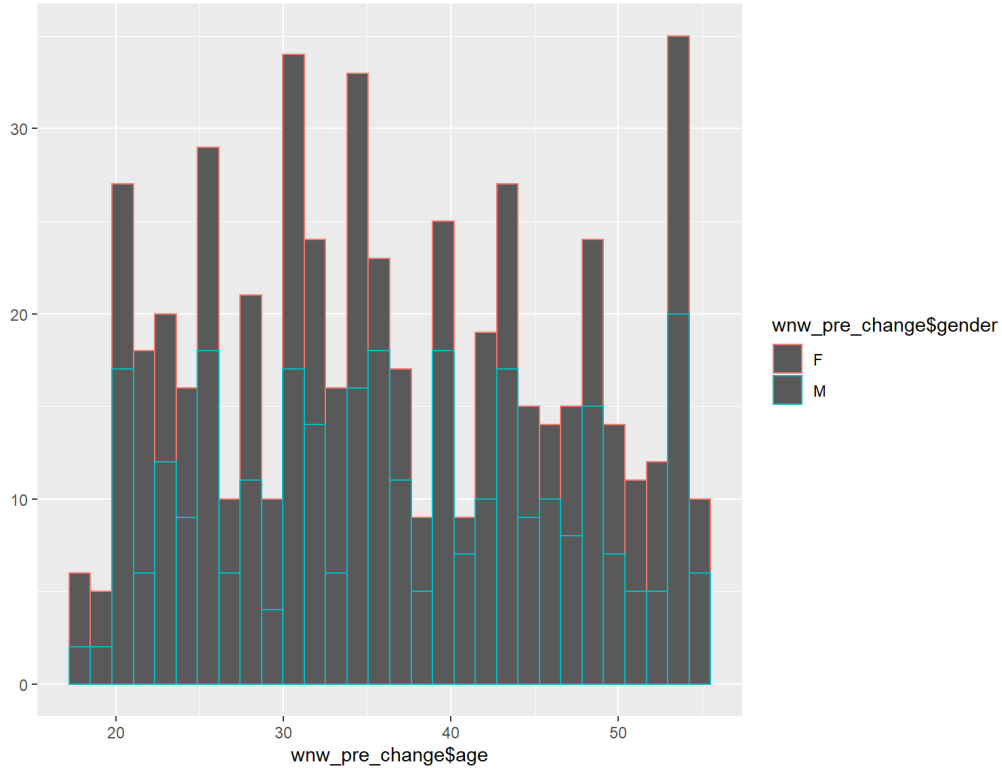



```
boxplot.stats(wnw_post_2_Clean$residual4)
```

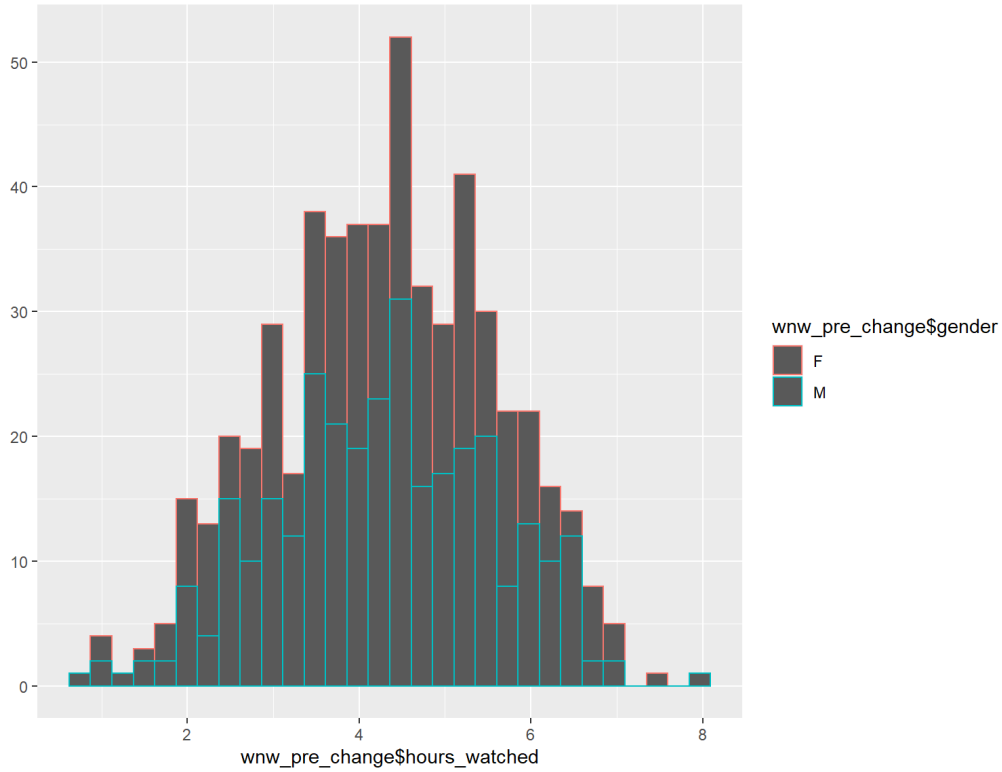
```
## $stats
## [1] -2.533336876 -0.605168630 -0.007066483  0.681595114  2.513596608
##
## $n
## [1] 447
##
## $conf
## [1] -0.10322819  0.08909522
##
## $out
## numeric(0)
```

Graphs

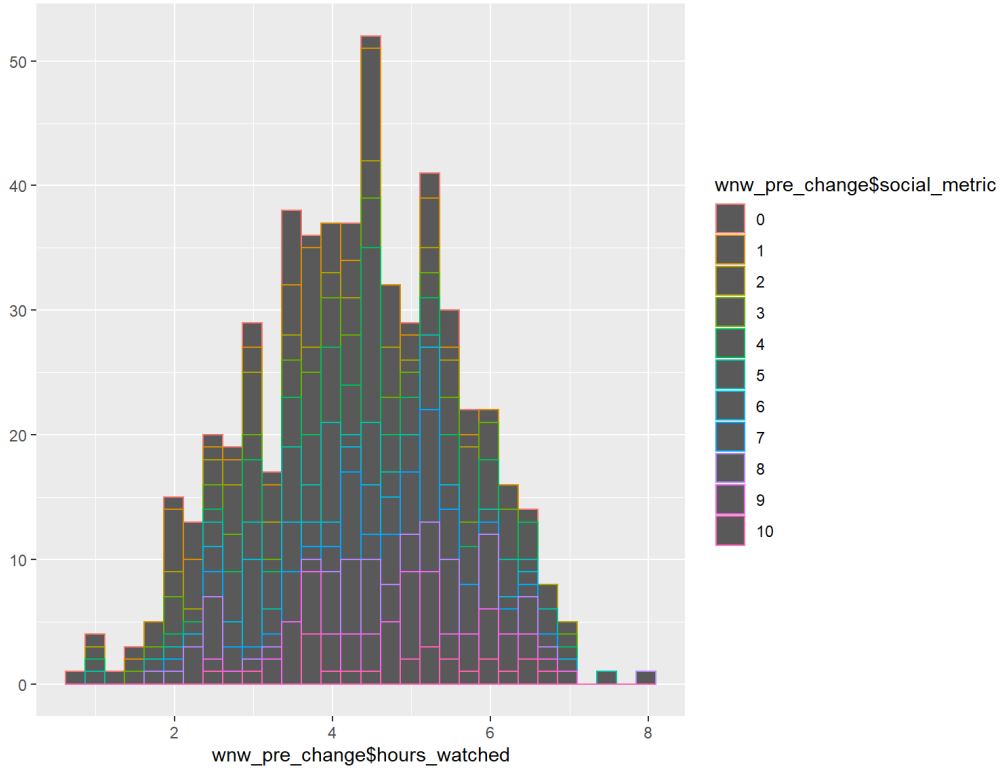
```
# Age distribution - before change
qplot(wnw_pre_change$age, colour = wnw_pre_change$gender)
```



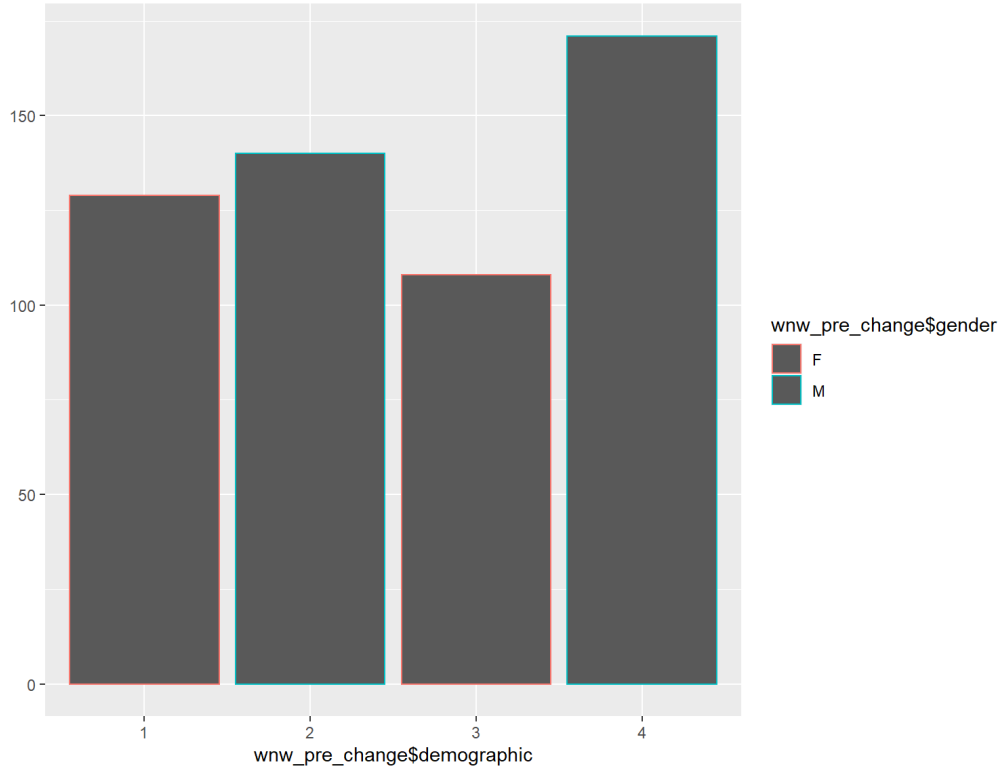

```
# Hours watched distribution - before change - with gender coloured.
# We can observe that women higher amount of hours watched than men.
qplot(wnw_pre_change$hours_watched, colour = wnw_pre_change$gender)
```



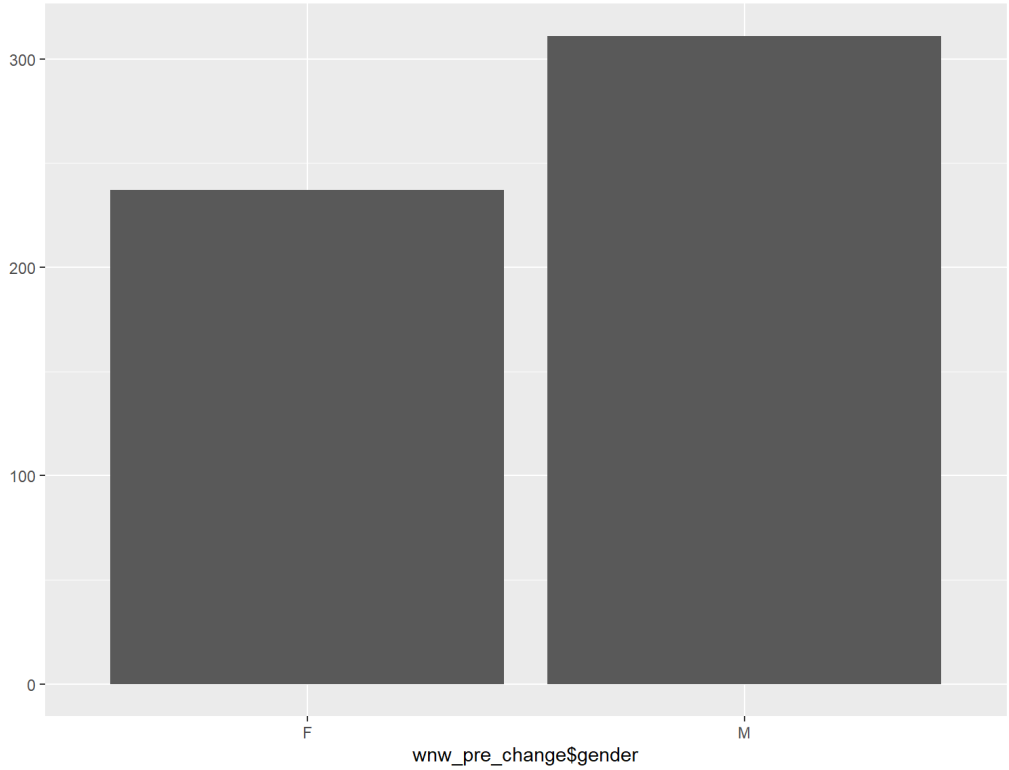

```
# Hours watched distribution - before change - social metrics colour
qplot(wnw_pre_change$hours_watched, colour = wnw_pre_change$social_metric)
```



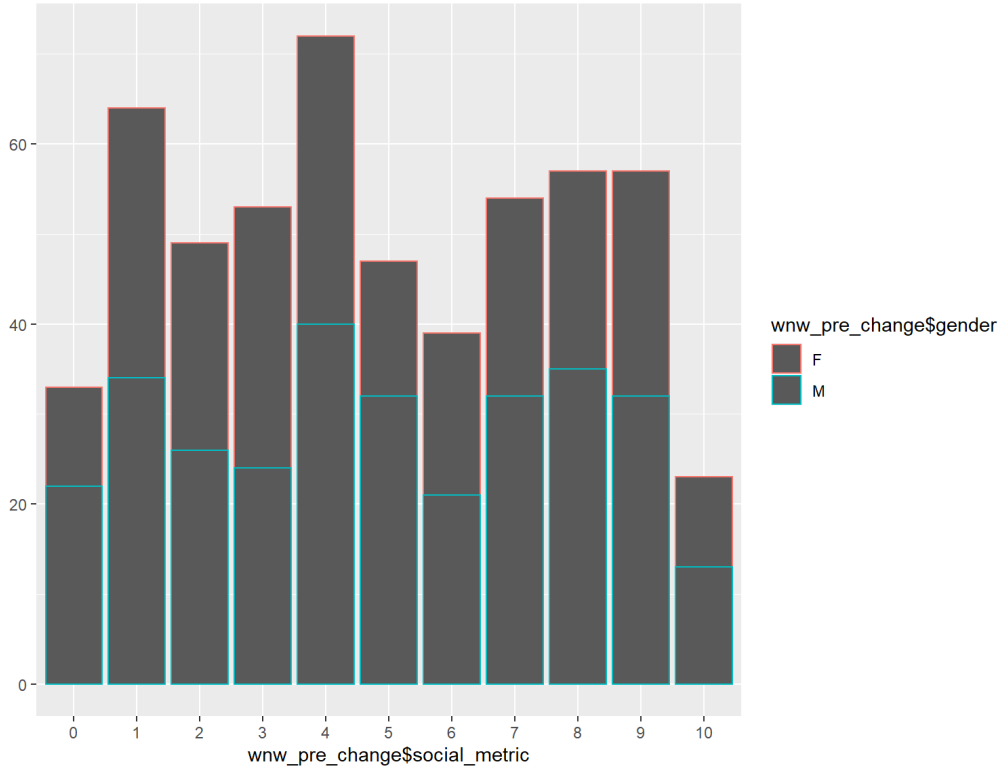

```
# Plot of social metric distribution -
# across social metrics, there are a higher amount of younger people demographics from groups 1 and 2, than the older groups
        3 and 4.
# Group 1 (Young Female) more socially engaged than group 2 (Young Males)
#qplot(wnw_pre_change$social_metric, colour = wnw_pre_change$demographic)


# Demographics split and frequency: - Groups 1 & 3 are Female. Groups 2& 4 are Male
qplot(wnw_pre_change$demographic, colour = wnw_pre_change$gender)
```



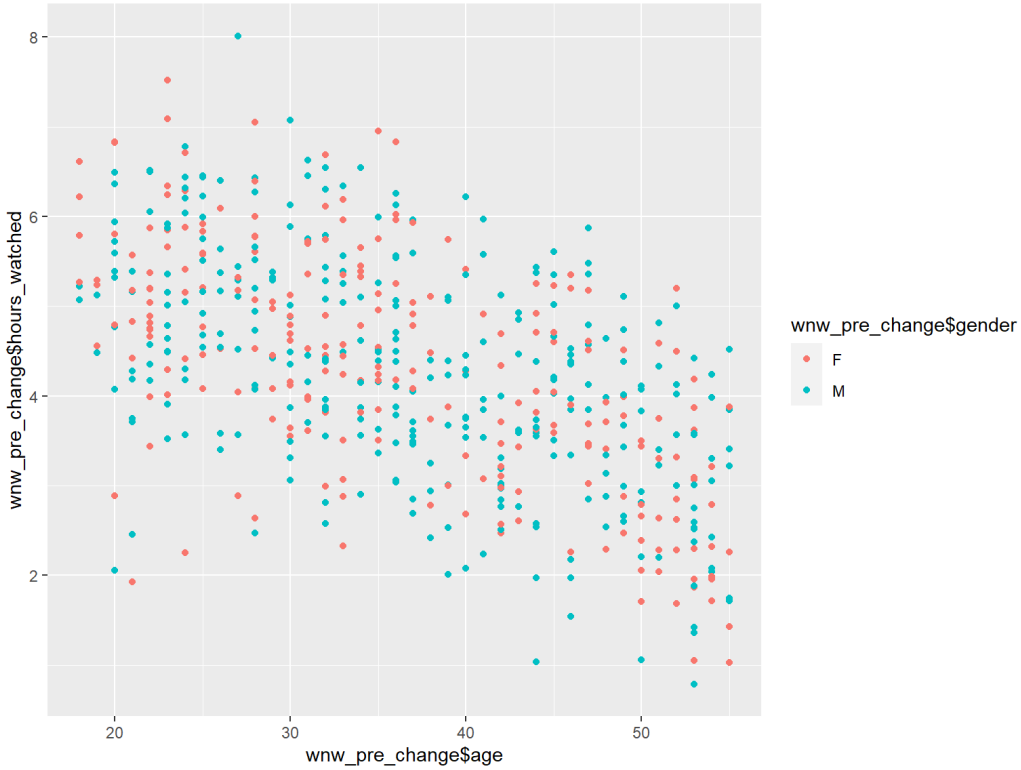

```
# Break up of gender in data set before changes
qplot(wnw_pre_change$gender)
```



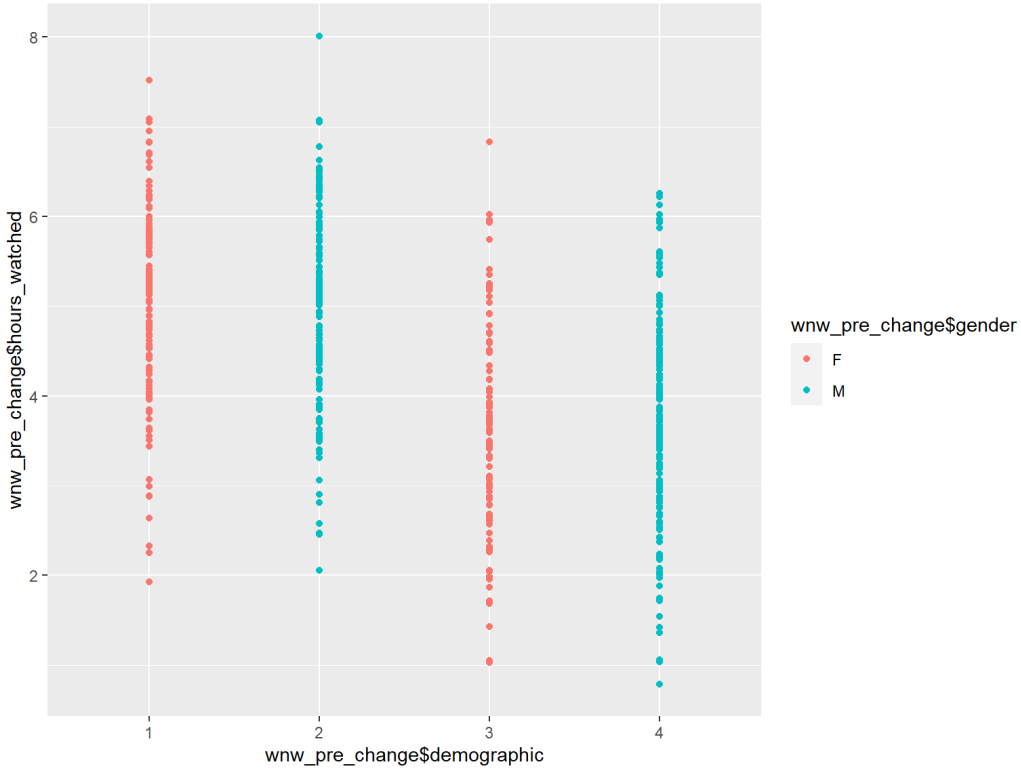

```
#hist(wnw_pre_change$time_since_signup)


# Social metrics with gender colour filter. Female have higher proportion
qplot(wnw_pre_change$social_metric, colour = wnw_pre_change$gender)
```



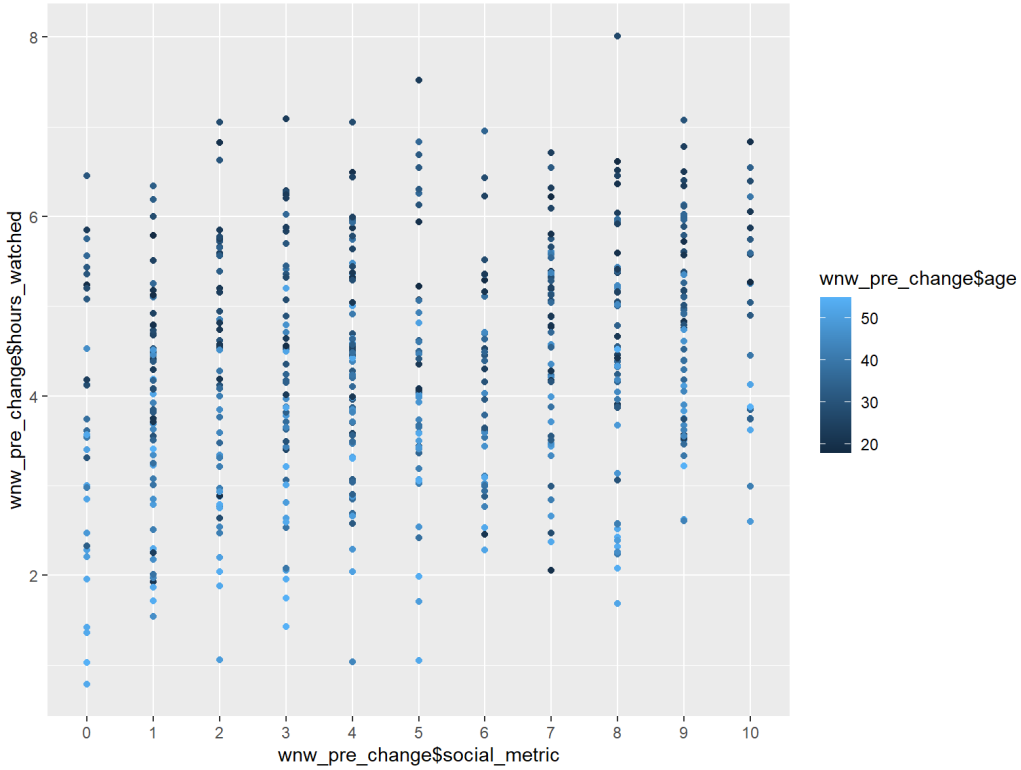

```
# Age vs Hours watched :-
# We can observe that younger customers watch more than older customers.
qplot(wnw_pre_change$age, wnw_pre_change$hours_watched, colour = wnw_pre_change$gender)# Keep
```



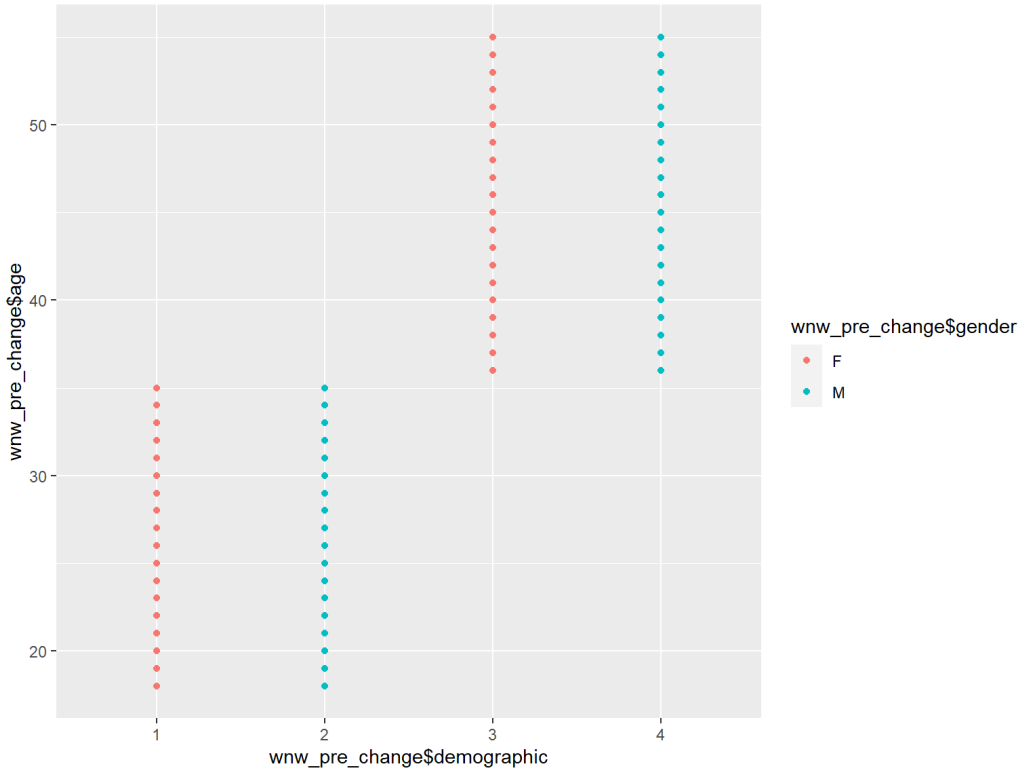

```
# Gender vs Age, coloured by age. We can observe that women watch more on average than men, older customers on average have
        less hours watched.
#qplot(wnw_pre_change$gender, wnw_pre_change$hours_watched, colour = wnw_pre_change$age)



# Demographic vs hours watched. Demographics 3 & 4 watch less than Demographics 1 & 2 which are younger.
qplot(wnw_pre_change$demographic, wnw_pre_change$hours_watched, colour = wnw_pre_change$gender) #Keep
```



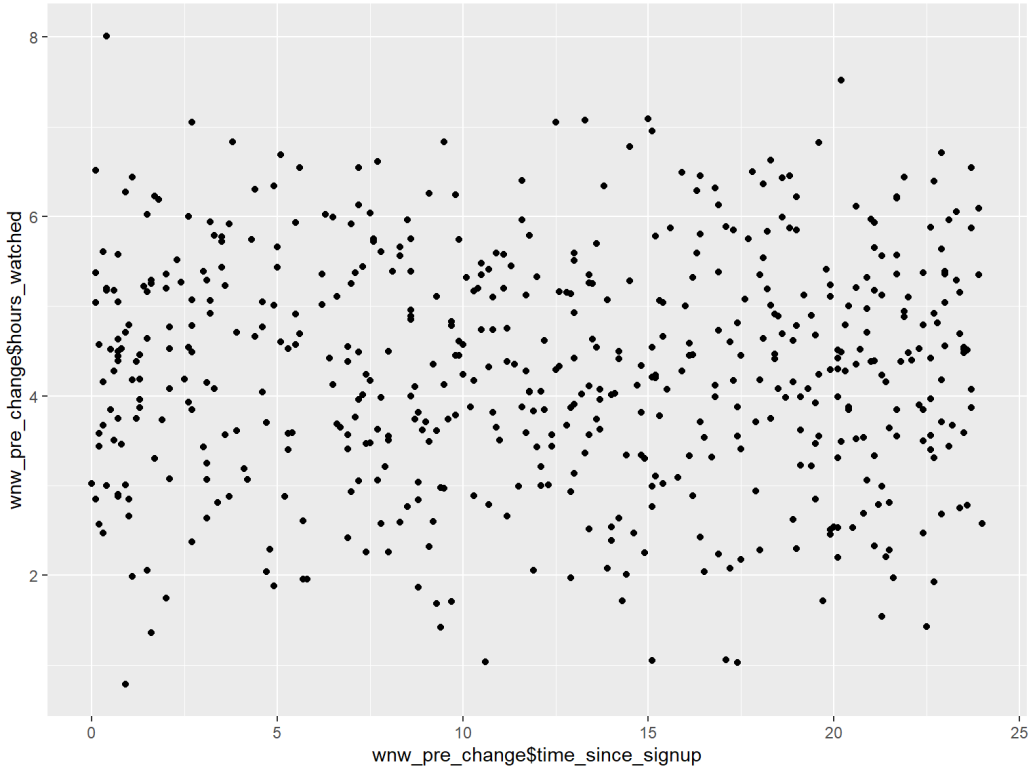

```
# Social Metric vs Hours Watched
# Observe that the higher the rating for social metric, ages of customers tend to be younger.
qplot(wnw_pre_change$social_metric, wnw_pre_change$hours_watched, colour = wnw_pre_change$age)
```



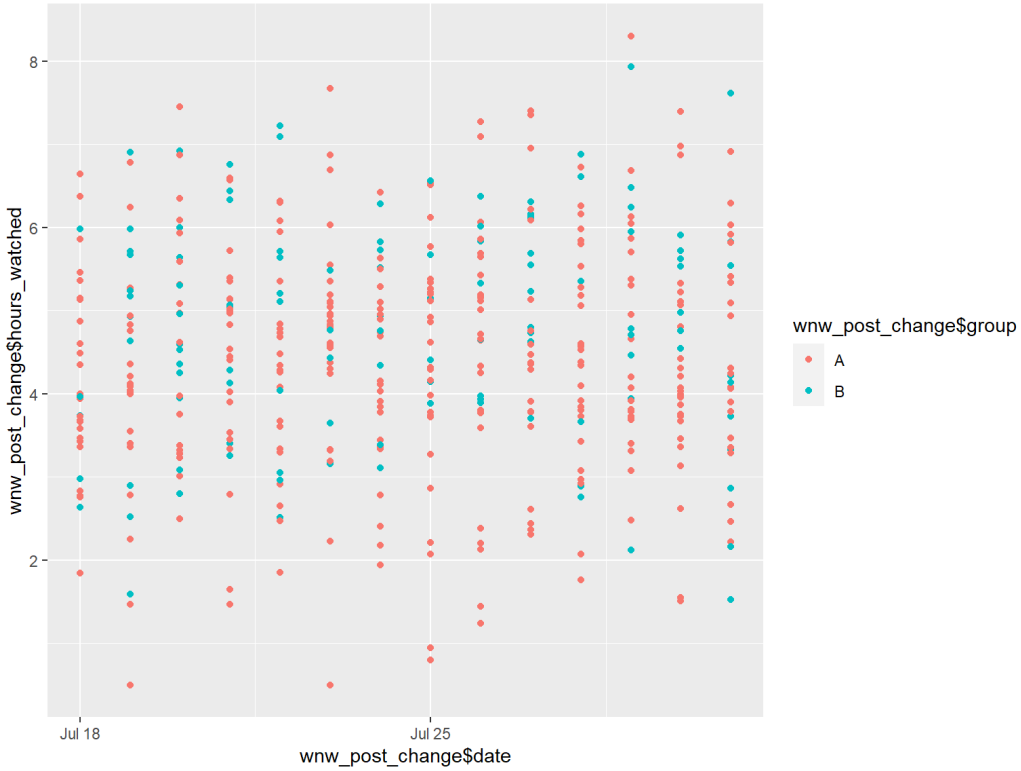

```
# We can see demo groups 1 and 2 are female and male respectively. ages 35 below
# We can see demo groups 3 and 4 are female and male respectively. ages above 35
qplot(wnw_pre_change$demographic, wnw_pre_change$age, colour = wnw_pre_change$gender)
```



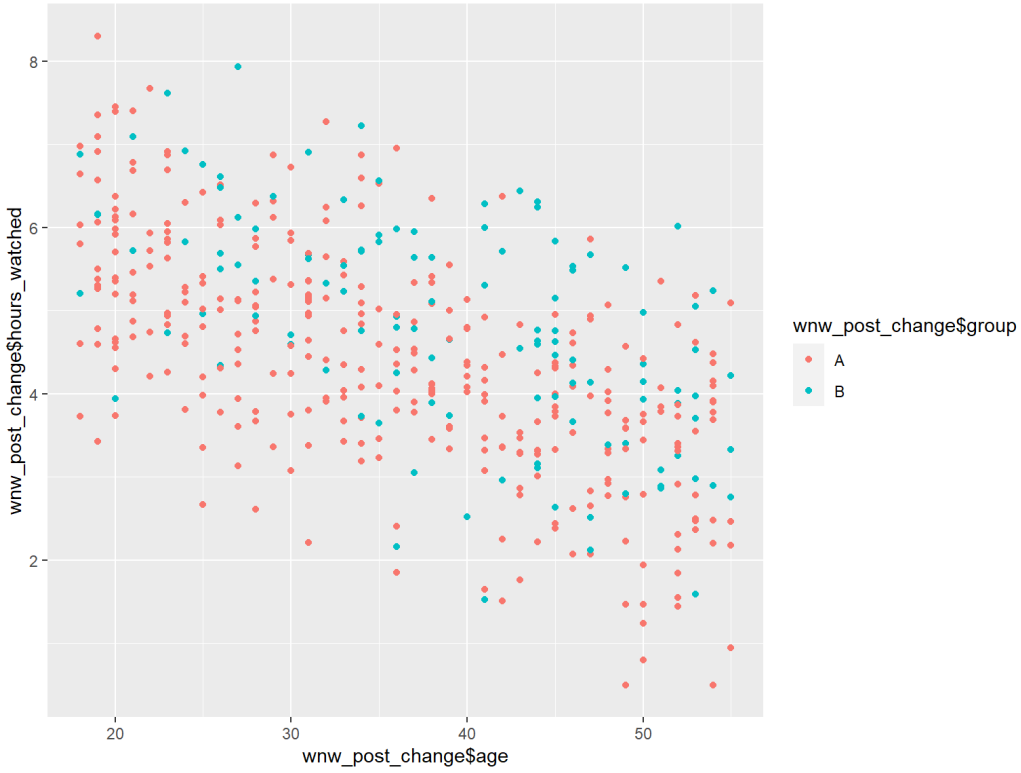

```
# Time since sign up vs hours watched - not very informative.
qplot(wnw_pre_change$time_since_signup, wnw_pre_change$hours_watched)
```



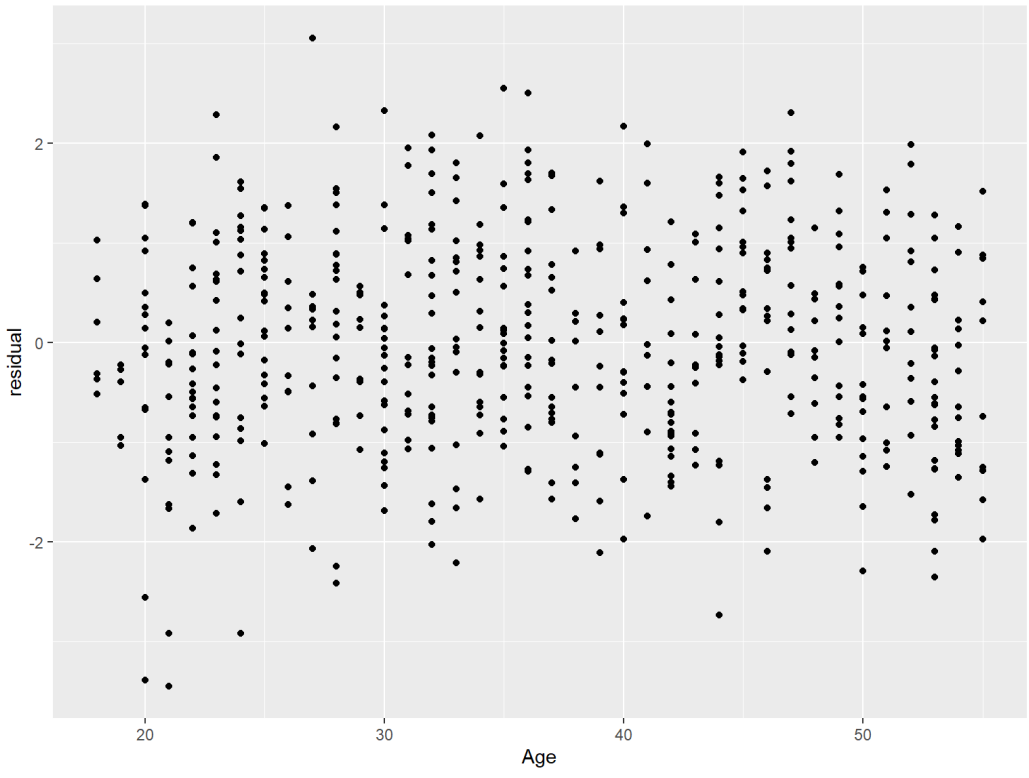

```
qplot(wnw_post_change$date, wnw_post_change$hours_watched, colour = wnw_post_change$group)
```




```
qplot(wnw_post_change$age, wnw_post_change$hours_watched, colour = wnw_post_change$group)
```




```
####################

# Hours watched / Time since sign up. Notable drop in ours watched after a couple of months.
#qplot(wnw_pre_change$hours_watched/wnw_pre_change$time_since_signup, colour = wnw_pre_change$gender)

#qplot(wnw_pre_change$hours_watched/wnw_pre_change$time_since_signup, colour = wnw_pre_change$demographic)

#qplot(wnw_pre_change$hours_watched/wnw_pre_change$time_since_signup, colour = wnw_pre_change$social_metric)



gg
```



Assignment 3

Table of Contents

- Problem Statement
- Objective
- Analysis
- Discussion
- Conclusion

Introduction

- Why Not Watch (WNW) faces increasing competition in the streaming industry. Due to Covid, the industry has seen rapid changes With increased competition from traditional film studios and production companies launching their own platforms.
- With the increased market competition, customer engagement and subscription renewals are expected to decline.
- Hours watched per day, a key metric used to price ads for third-party marketing companies, is also expected to fall, further affecting revenue.
- Improvements to our recommendation engine is critical in increasing customer engagement, maintaining subscribers and maximising profits.
- Proposed changes should be carefully tested with well designed experiments and evaluated with statistical analysis before deployment.

Objective

-This presentation aims to present the findings on the results of a test conducted on the effectiveness of an update to our platform's recommendation engine. - Statistical analysis has been performed on the data to assess the effectiveness of the proposed engine. - Present findings of analysis - Provide recommendations on future tests - A recommendation as to whether this engine should be rolled out to all subscribers will then be given based on findings.

Data Set Break Down

Data set consists of 1000 observations and 8 variables.

- **Date:** date in format yyyy-mm-dd format
- Dates 2022-07-01 to 2022-07-17 take place before trial of recommendation engine.
- Dates 2022-07-18 to 2022-07-31 take place during test of new recommendation engine.
- **Gender:** Factor with two levels noting customer gender.
- **Age:** Integer data noting customer age.
- Min age = 18, Max age = 55
- **Social Metric**: Numeric factor with 11 levels from 0 to 10. The rating is based on customer's previous viewing habits.
- **Time Since Sign-up:** Numeric value noting number of months (to 2 dec places) since customer subscribed.
- Min = 0.00, Max = 24.00

Data Set Break Down - Continued

Demographic: Numeric Factor with 4 levels (from 1 to 4). **-Demographic 1:** Women aged to 35 **-Demographic 2:** Men ages to 35 **-Demographic 3:** Women aged 36 and above **-Demographic 4:** Men aged 36 and above

Group: Two level factor. For duration of engine test. - Group A: is control group. - Group B: Treatment group for new recommendation engine

Hours Watched: Numeric data (to 3 dec places). The number of hours watched that day.

- Min = 0.500, Max = 8.300

Data Set Break Down - Continued

Data Pre-Processing: -Date: Converted to date format for easier manipulation and subsetting between testing and non testing periods.

Converted variables following variables to factors for easier data frame manipulation. -**Gender:** -**Social Metric:** -**Demographic -Group:**

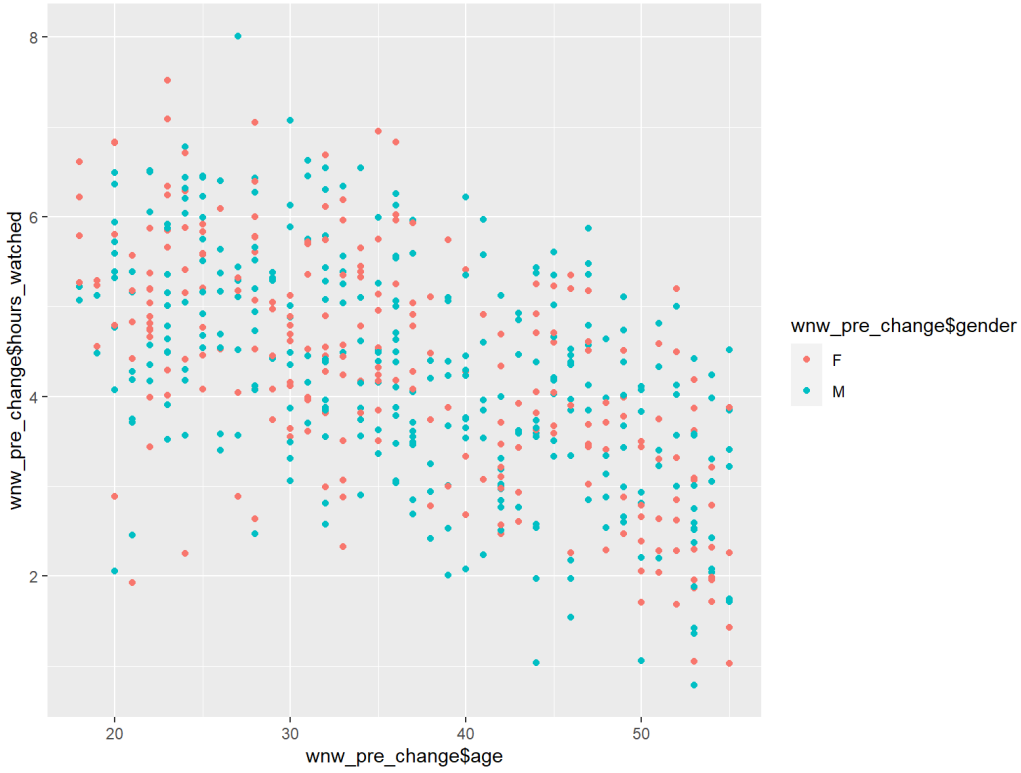
The data set is split into two by date range: - First data set is data before new recommendation is tested. - Second data set is data during test of new Recommendation engine - Second data set also allows us to analyse results between Control Group (A) and Treatment Group (B).

Analysis

Relationship Analysis: - Looking at data before the trial is conducted. - To see which variables drive our target variable of **Hours Watched**, We analyse the other variables in our data for any relationships and check if they are statistically significant with **Hours Watched** - ($p < 0.05$).

Analysis - Cont'd

Analysis of the relationship of **Hours Watched** with **Age** is performed and a relationship can be observed. Hours watched decreased with age. Women have higher hours watched than men.



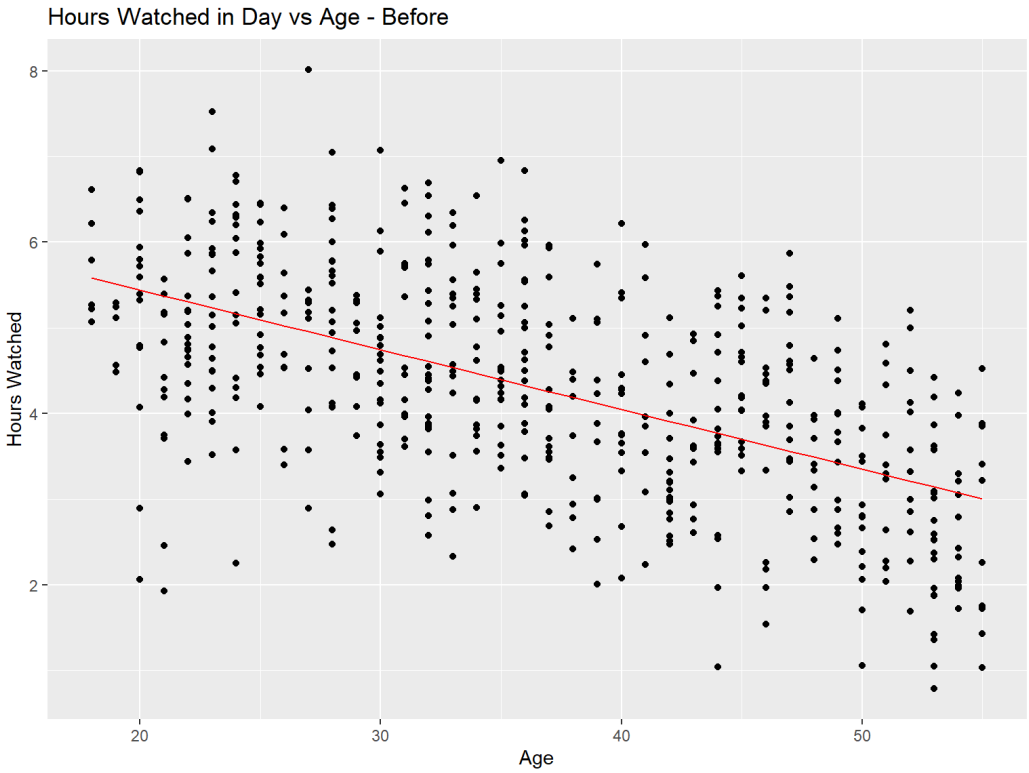
Analysis - Cont'd

- Performing a Linear Regression analysis shows **AGE** is significant and gives us R^2 score of 0.3214.
- R^2 is a "goodness of fit". It tells us the proportion of variance in the dependent variable that can be explained by the independent variable. Generally, the closer to 1, the better the fit.
- A Multiple Regression Analysis is performed to determine which variables are useful drivers to keep. We find that **Social_Metric** is an important driver of **Hours Watched**
- Including **Age** and **Social Metrics** in our model gives us an adjusted R^2 score of 0.3814.

Analysis

Linear Regression.

Plot below further demonstrates correlation of **Age** to **Hours Watched** We can observe that older audience watch less than hours than younger audiences.

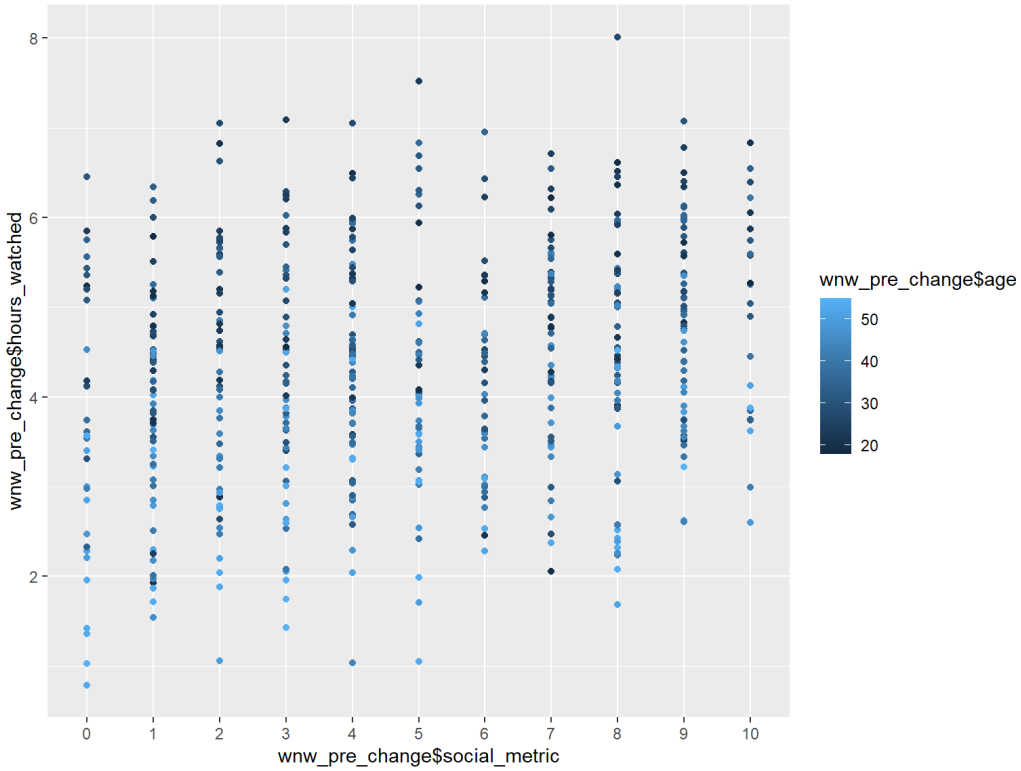


Analysis

Below plot shows relationship of **Social Metric** to **Hours Watched**

We can observe that those rated on a higher social metric tend to have more hours watchedthan those with lower Social Metric Score.

Colour of plots shows that younger audiences have higher hours watched than older audiences.

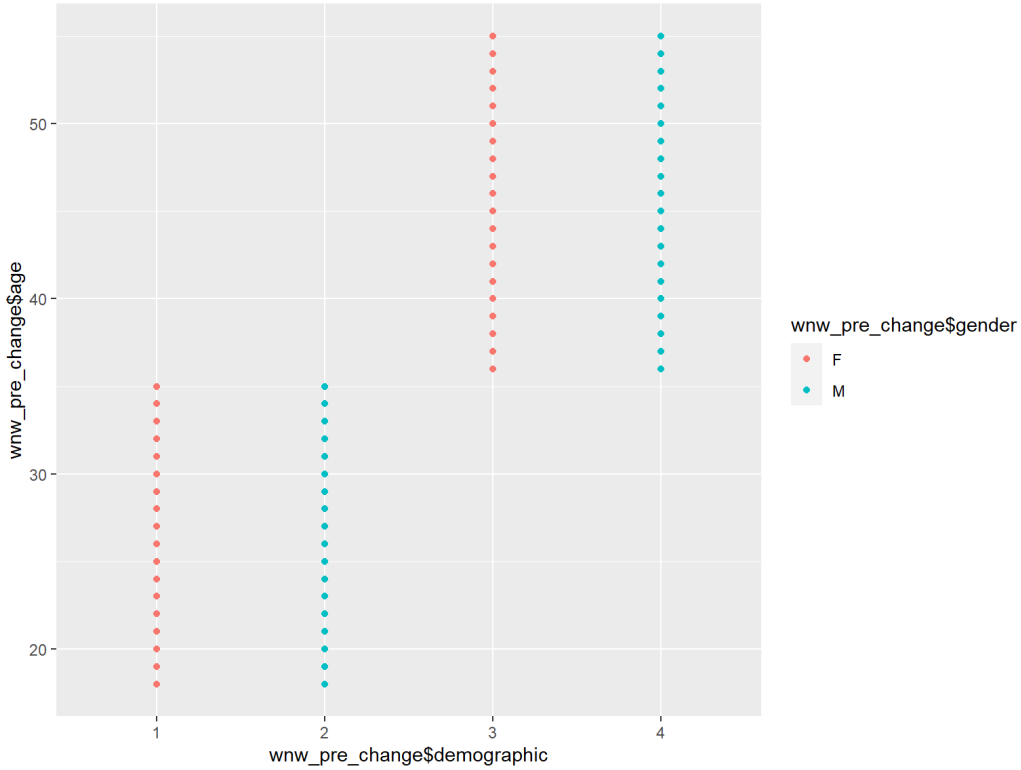


Analysis Cont'd

Demographics

Our Regression Analysis, did not see Demographics as a valuable driver to Hours Watched. However, the below graph provides a visual break up of demographics.

-Demographic 1: Women aged to 35 **-Demographic 2:** Men ages to 35 **-Demographic 3:** Women aged 36 and above **-Demographic 4:** Men aged 36 and above



Analysis - Hypothesis Test.

Test the **Means** of **Hours Watched** between Group A and Group B **Group A** using current engine. **Group B** using new engine

Use the Independent samples test of means: Test chosen as following conditions are met: -We can assume both datasets have been chosen based on simple random sampling -Both samples are large ($n > 30$)

Samples are independent: - No connection between Groups A and Groups B - Each group has different participants.

Analysis - Hypothesis Test Cont'd

Significance level: $\alpha = 0.05$$

Null Hypothesis - No difference in means. $H_0 : \mu_2 = \mu_1$

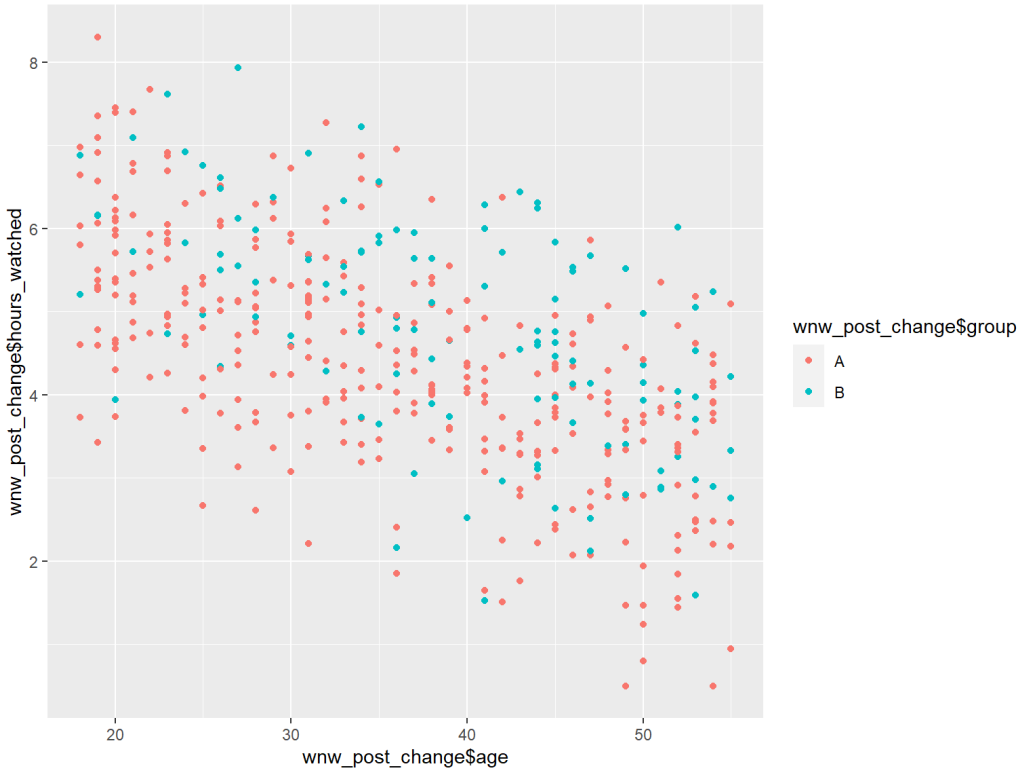
Alternate Hypothesis: mean hours for Group B > Group A $H_A : \mu_2 > \mu_1$

A $p - value < 0.05$ means we can reject the Null Hypothesis.

Analysis - Results

Test returns a $p - value$ of 0.002. This means we can reject the null hypothesis and we can assume that difference between Group A and B is statistically significant.

We can also observe **Group A** Average Hours Watched: 4.40 (2 decimal places) **Group B** Average Hours Watched: 4.80 (2 decimal places)



Discussion - Findings/ Sampling Bias

Hypothesis test found changes in engine to be statistically significant.

Analysis of data shows that neither Group A and Group B are probabilistically similar to each other. Test groups are not likely to be similar with population. **Age** People in Group A are older than Group B **Gender** The proportion of men to women in group B is much higher than Group A **Demographics** The proportions of demographics between groups are different **Social Metrics** Distribution of social metrics is different between groups.

Unrepresentative sample data means that results are not indicative of full deployment.

Bias Correction: -Possible to Post-weight statistics to when samples are not matched. -As test samples are not representative of the population across several characteristics, this would be difficult, it is recommended a new test be conducted.

Discussion - Findings/ Sampling Bias

Demographic - Proportion Comparison

Demographic WNW_Before Group_A_Prop Group_B_Prop Difference

Demographic	WNW_Before	Group_A_Prop	Group_B_Prop	Difference
1	0.2354015	0.2228916	0.1083333	-0.1145582
2	0.2554745	0.2891566	0.2666667	-0.0224900
3	0.1970803	0.2680723	0.1333333	-0.1347390
4	0.3120438	0.2198795	0.4916667	0.2717871

Gender - Proportion Comparison

WNW_Before	Group_A_Prop	Group_B_Prop	Difference
0.4324818	0.4909639	0.2416667	-0.2492972
0.5675182	0.5090361	0.7583333	0.2492972

Discussions - Recommendations

Recommendations for future investigations. Enhance effect and validity of future tests by: - Having access to population data. - Random sampling of people into test groups to avoid bias and ensuring they are, probablistically similar in all respects with the population.

Calculate and choose correct sample size with respect to desired: -Significance level -Test Power - High test power reduces probability of incorrectly failing to reject null hypothesis -Effect Size

Discussions - Recommendations Cont'd

Given our data: -Calculated Effect Size: = $\hat{d} = (\mu_2 - \mu_1)/\sigma_{pooled}$ = 0.3 - Assuming all else remaining equal, a test with power 0f 0.80 (Common test power) - Minimum sample size for Groups A and B should be 139. - Test size for Group B is too small - Larger sample means increased effect size.

Dividing data into more sub categories for more detailed insights. -For example, Demographics is currently split into gender and age groups below 35 and above 35. - We can divide demographics into more age groups.

Discussion - Conclusion

Though a Hypothesis test has shown, on the surface, that a roll out of an upgrade of WNW's recommendation engine would yield favourable results.

Our analysis has found issues with sample size, and sample selection bias which mean that the differences between the test samples is not attributable to the performance of the new engine. Therefore, it is unlikely that a full roll-out based on statistical findings is unlikely to yield similar results.

In conclusion, further testing designed to mitigate sampling bias and error is recommended before the new engine is deployed.

References

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.37.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Stefan Milton Bache and Hadley Wickham (2022). magrittr: A Forward-Pipe Operator for R. R package version 2.0.2. https://CRAN.R-project.org/package=magrittr

Taiyun Wei and Viliam Simko (2021). R package 'corrplot':Visualization of a Correlation Matrix (Version 0.92). Available from https://github.com/taiyun/corrplot

Stephane Champely (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr

Zach 2022, *How to Calculate Pooled Standard Deviation in R* statology.com, viewed 12 June 2022, https://www.statology.org/pooled-standard-deviation-in-r/

Thomas Lin Pedersen (2020). patchwork: The Composer of Plots. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork