Impact of Smoking and Drinking

Srividhya Pattabiraman, Disha Chiplonker, Shruthi Mysore Narayanaswamy

11/25/2023

INTRODUCTION

In this project, we will analyze impacts of smoking and drinking from data set collected by National Health Insurance Service in Korea.

Smoking and Drinking has always given way for numerous diseases and illnesses. It is important to study the impact of drinking and smoking on various health parameters like Cholesterol, Liver, etc.
For this very reason we chose to analyse and answer the following questions through our study:
1) Is there a relationship between smoking and drinking?
2) How Smoking affects Good Cholesterol?
3) Is there any relationship between increased waistline and elevated levels of Gamma GTP in smokers and drinkers?
4) Is there any relationship associated with Weight and Drinking?

We begin by loading the data set of 991,346 observations into the R workspace.

```
sd_data <- read.csv("smoking_driking_dataset_Ver01.csv", as.is=T)
```

Description of the data given by the insurance company:

This dataset has almost 14 attributes,

Sex - male, female

age - round up to 5 years

height - round up to 5 cm[cm]

weight - round up to 5 kg[kg]

sight_left - eyesight(left) - (0.1~2.5, eyesight $< 0.1 = 0.1$ (good), blind = 9.9)

sight_right - eyesight(right) - (0.1~2.5, eyesight $< 0.1 = 0.1$ (good), blind = 9.9)

hear_left - hearing left, 1(normal), 2(abnormal)

hear_right - hearing right, 1(normal), 2(abnormal)

SBP - Systolic blood pressure[mmHg]

DBP - Diastolic blood pressure[mmHg]

BLDS - BLDS or FSG(fasting blood glucose)[mg/dL]

tot_chole - total cholesterol[mg/dL]

HDL_chole - HDL cholesterol[mg/dL]

LDL_chole - LDL cholesterol[mg/dL]

triglyceride - triglyceride[mg/dL]

hemoglobin - hemoglobin[g/dL]

urine_protein - protein in urine, 1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4) - 1 - Normal, 2 - Borderline, >2 - Abnormal

serum_creatinine - serum(blood) creatinine[mg/dL]

SGOT_AST - SGOT(Glutamate-oxaloacetate transaminase) AST(Aspartate transaminase)[IU/L]

SGOT_ALT - ALT(Alanine transaminase)[IU/L]

gamma_GTP - y-glutamyl transpeptidase[IU/L]

SMK_stat_type_cd - Smoking state, 1(never), 2(used to smoke but quit), 3(still smoke)

DRK_YN - Drinker or Not

From the given information, we can identify the categorical and numerical columns.

Categorical (non-ordinal) - Sex, hear_left, hear_right, SMK_state_type_cd, DRK_YN

Categorical (ordinal) - unine_protein

Numerical (discrete) - age, height, weight

Numerical (continuous) - all the other attributes

```
sd_data <- sd_data %>% mutate(DRK_YN = factor(DRK_YN), SMK_stat_type_cd = factor(SMK_stat_type_cd), sex
summary(sd_data)
```

```
##      sex              age            height          weight
##  Female:464931   Min.   :20.00   Min.   :130.0   Min.   : 25.00
##  Male  :526415   1st Qu.:35.00   1st Qu.:155.0   1st Qu.: 55.00
##                  Median :45.00   Median :160.0   Median : 60.00
##                  Mean   :47.61   Mean   :162.2   Mean   : 63.28
##                  3rd Qu.:60.00   3rd Qu.:170.0   3rd Qu.: 70.00
##                  Max.   :85.00   Max.   :190.0   Max.   :140.00
##    waistline        sight_left      sight_right     hear_left  hear_right
##  Min.   :  8.00   Min.   :0.1000   Min.   :0.1000   1:960124   1:961134
##  1st Qu.: 74.10   1st Qu.:0.7000   1st Qu.:0.7000   2: 31222   2: 30212
##  Median : 81.00   Median :1.0000   Median :1.0000
##  Mean   : 81.23   Mean   :0.9808   Mean   :0.9784
##  3rd Qu.: 87.80   3rd Qu.:1.2000   3rd Qu.:1.2000
##  Max.   :999.00   Max.   :9.9000   Max.   :9.9000
##       SBP              DBP              BLDS         tot_chole
##  Min.   : 67.0   Min.   : 32.00   Min.   : 25.0   Min.   :  30.0
##  1st Qu.:112.0   1st Qu.: 70.00   1st Qu.: 88.0   1st Qu.: 169.0
##  Median :120.0   Median : 76.00   Median : 96.0   Median : 193.0
##  Mean   :122.4   Mean   : 76.05   Mean   :100.4   Mean   : 195.6
##  3rd Qu.:131.0   3rd Qu.: 82.00   3rd Qu.:105.0   3rd Qu.: 219.0
##  Max.   :273.0   Max.   :185.00   Max.   :852.0   Max.   :2344.0
##    HDL_chole         LDL_chole     triglyceride      hemoglobin
##  Min.   :   1.00   Min.   :   1   Min.   :   1.0   Min.   : 1.00
##  1st Qu.:  46.00   1st Qu.:  89   1st Qu.:  73.0   1st Qu.:13.20
##  Median :  55.00   Median : 111   Median : 106.0   Median :14.30
##  Mean   :  56.94   Mean   : 113   Mean   : 132.1   Mean   :14.23
##  3rd Qu.:  66.00   3rd Qu.: 135   3rd Qu.: 159.0   3rd Qu.:15.40
##  Max.   :8110.00   Max.   :5119   Max.   :9490.0   Max.   :25.00
```

```
##  urine_protein serum_creatinine     SGOT_AST          SGOT_ALT
##  1:935175      Min.   : 0.1000   Min.   :   1.00   Min.   :   1.00
##  2: 30850      1st Qu.: 0.7000   1st Qu.:  19.00   1st Qu.:  15.00
##  3: 16405      Median : 0.8000   Median :  23.00   Median :  20.00
##  4:  6427      Mean   : 0.8605   Mean   :  25.99   Mean   :  25.75
##  5:  1977      3rd Qu.: 1.0000   3rd Qu.:  28.00   3rd Qu.:  29.00
##  6:   512      Max.   :98.0000   Max.   :9999.00   Max.   :7210.00
##    gamma_GTP      SMK_stat_type_cd DRK_YN
##  Min.   :  1.00   1:602441         N:495858
##  1st Qu.: 16.00   2:174951         Y:495488
##  Median : 23.00   3:213954
##  Mean   : 37.14
##  3rd Qu.: 39.00
##  Max.   :999.00
```

Data Quality Check

We checked for any missing data and remove them. Since there was no missing data and the data quality was good. We proceeded with the analysis.

```
colSums(is.na(sd_data))
```

```
##              sex              age           height           weight
##                0                0                0                0
##        waistline       sight_left      sight_right        hear_left
##                0                0                0                0
##       hear_right              SBP              DBP             BLDS
##                0                0                0                0
##        tot_chole        HDL_chole        LDL_chole     triglyceride
##                0                0                0                0
##       hemoglobin    urine_protein serum_creatinine         SGOT_AST
##                0                0                0                0
##         SGOT_ALT        gamma_GTP SMK_stat_type_cd           DRK_YN
##                0                0                0                0
```

Correlation study

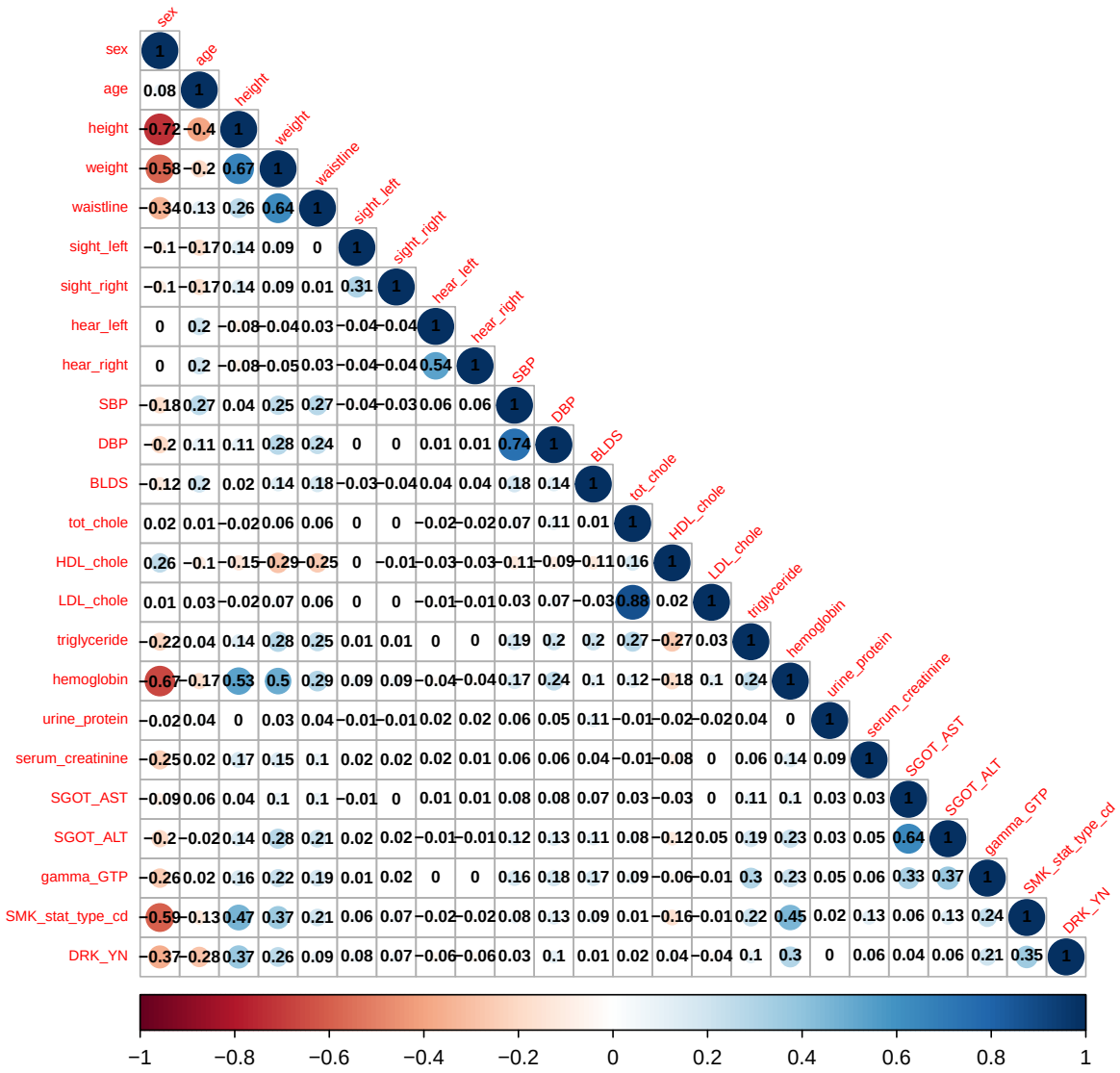
We ran a correlation study of various factors against smoking and drinking as a preliminary step of analysis.

```
sd_numeric_data <- sd_data %>%
      mutate(DRK_YN = ifelse(DRK_YN == "N",0,1), sex = ifelse(sex == "Male",0,1))
indx <- sapply(sd_numeric_data, is.factor)
sd_numeric_data[indx] <- lapply(sd_numeric_data[indx], function(x) as.numeric(as.character(x)))
summary(sd_numeric_data)
```

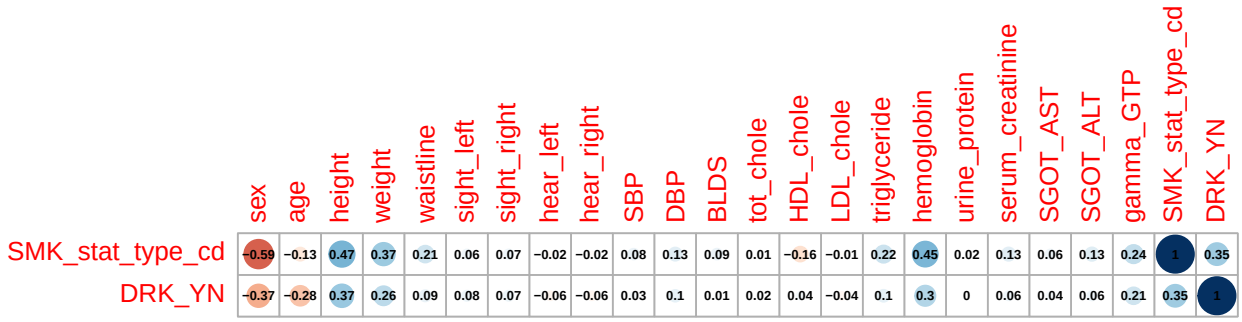
```
##       sex              age            height          weight
##  Min.   :0.000   Min.   :20.00   Min.   :130.0   Min.   : 25.00
##  1st Qu.:0.000   1st Qu.:35.00   1st Qu.:155.0   1st Qu.: 55.00
##  Median :0.000   Median :45.00   Median :160.0   Median : 60.00
##  Mean   :0.469   Mean   :47.61   Mean   :162.2   Mean   : 63.28
##  3rd Qu.:1.000   3rd Qu.:60.00   3rd Qu.:170.0   3rd Qu.: 70.00
##  Max.   :1.000   Max.   :85.00   Max.   :190.0   Max.   :140.00
##    waistline        sight_left      sight_right       hear_left
##  Min.   :  8.00   Min.   :0.1000   Min.   :0.1000   Min.   :1.000
```

```
##  1st Qu.: 74.10   1st Qu.:0.7000   1st Qu.:0.7000   1st Qu.:1.000
##  Median : 81.00   Median :1.0000   Median :1.0000   Median :1.000
##  Mean   : 81.23   Mean   :0.9808   Mean   :0.9784   Mean   :1.031
##  3rd Qu.: 87.80   3rd Qu.:1.2000   3rd Qu.:1.2000   3rd Qu.:1.000
##  Max.   :999.00   Max.   :9.9000   Max.   :9.9000   Max.   :2.000
##    hear_right        SBP              DBP              BLDS
##  Min.   :1.00   Min.   : 67.0   Min.   : 32.00   Min.   : 25.0
##  1st Qu.:1.00   1st Qu.:112.0   1st Qu.: 70.00   1st Qu.: 88.0
##  Median :1.00   Median :120.0   Median : 76.00   Median : 96.0
##  Mean   :1.03   Mean   :122.4   Mean   : 76.05   Mean   :100.4
##  3rd Qu.:1.00   3rd Qu.:131.0   3rd Qu.: 82.00   3rd Qu.:105.0
##  Max.   :2.00   Max.   :273.0   Max.   :185.00   Max.   :852.0
##    tot_chole        HDL_chole         LDL_chole     triglyceride
##  Min.   :  30.0   Min.   :   1.00   Min.   :   1   Min.   :   1.0
##  1st Qu.: 169.0   1st Qu.:  46.00   1st Qu.:  89   1st Qu.:  73.0
##  Median : 193.0   Median :  55.00   Median : 111   Median : 106.0
##  Mean   : 195.6   Mean   :  56.94   Mean   : 113   Mean   : 132.1
##  3rd Qu.: 219.0   3rd Qu.:  66.00   3rd Qu.: 135   3rd Qu.: 159.0
##  Max.   :2344.0   Max.   :8110.00   Max.   :5119   Max.   :9490.0
##    hemoglobin    urine_protein   serum_creatinine     SGOT_AST
##  Min.   : 1.00   Min.   :1.000   Min.   : 0.1000   Min.   :   1.00
##  1st Qu.:13.20   1st Qu.:1.000   1st Qu.: 0.7000   1st Qu.:  19.00
##  Median :14.30   Median :1.000   Median : 0.8000   Median :  23.00
##  Mean   :14.23   Mean   :1.094   Mean   : 0.8605   Mean   :  25.99
##  3rd Qu.:15.40   3rd Qu.:1.000   3rd Qu.: 1.0000   3rd Qu.:  28.00
##  Max.   :25.00   Max.   :6.000   Max.   :98.0000   Max.   :9999.00
##     SGOT_ALT         gamma_GTP      SMK_stat_type_cd     DRK_YN
##  Min.   :   1.00   Min.   :  1.00   Min.   :1.000    Min.   :0.0000
##  1st Qu.:  15.00   1st Qu.: 16.00   1st Qu.:1.000    1st Qu.:0.0000
##  Median :  20.00   Median : 23.00   Median :1.000    Median :0.0000
##  Mean   :  25.75   Mean   : 37.14   Mean   :1.608    Mean   :0.4998
##  3rd Qu.:  29.00   3rd Qu.: 39.00   3rd Qu.:2.000    3rd Qu.:1.0000
##  Max.   :7210.00   Max.   :999.00   Max.   :3.000    Max.   :1.0000
```

```
corrplot(cor(sd_numeric_data), method = 'circle', addCoef.col ='black', number.cex = 0.7, type = 'lower
```




```r
corr <- cor(sd_numeric_data[tail(names(sd_numeric_data), 2)], sd_numeric_data)
corrplot(corr, method = 'circle', addCoef.col ='black', number.cex = 0.5, cl.pos = 'n')
```




```
smk_correlated <- colnames(corr)[abs(corr[1, ]) >= 0.2]
drk_correlated <- colnames(corr)[abs(corr[2, ]) >= 0.2]

cat("If we consider correlation threshold as 0.2 then factors that are correlated to, \n")
```

```
## If we consider correlation threshold as 0.2 then factors that are correlated to,
```

```
cat("Smoking:", smk_correlated, "\n")
```

```
## Smoking: sex height weight waistline triglyceride hemoglobin gamma_GTP SMK_stat_type_cd DRK_YN
```

```r
cat("Drinking:", drk_correlated, "\n\n")
```

```
## Drinking: sex age height weight hemoglobin gamma_GTP SMK_stat_type_cd DRK_YN
```

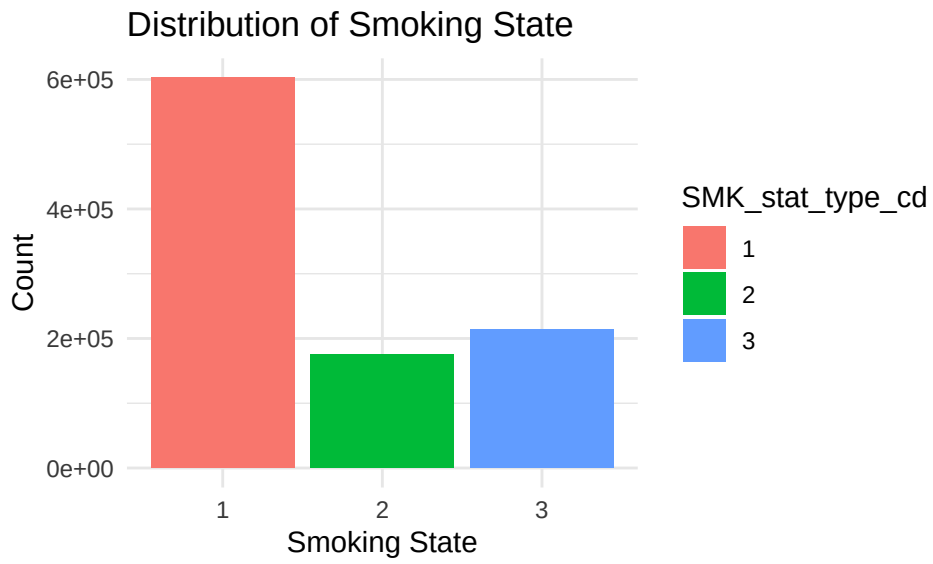
Is there a relationship between smoking and drinking?
From the correlation test we can see that smoking and drinking has a high correlation factor. Let's further test it with Chi-Square test.

In order to run the test, below are the assumptions:
1) Run on categorical data - Both SMK_stat_type_cd and DRK_YN are categorical.
2) Observations are independent - It is indicated that observations are independent.
3) Data is from a random sample - Data is drawn from random people from the population.
4) Large Sample size - Sample size is large enough (>30).

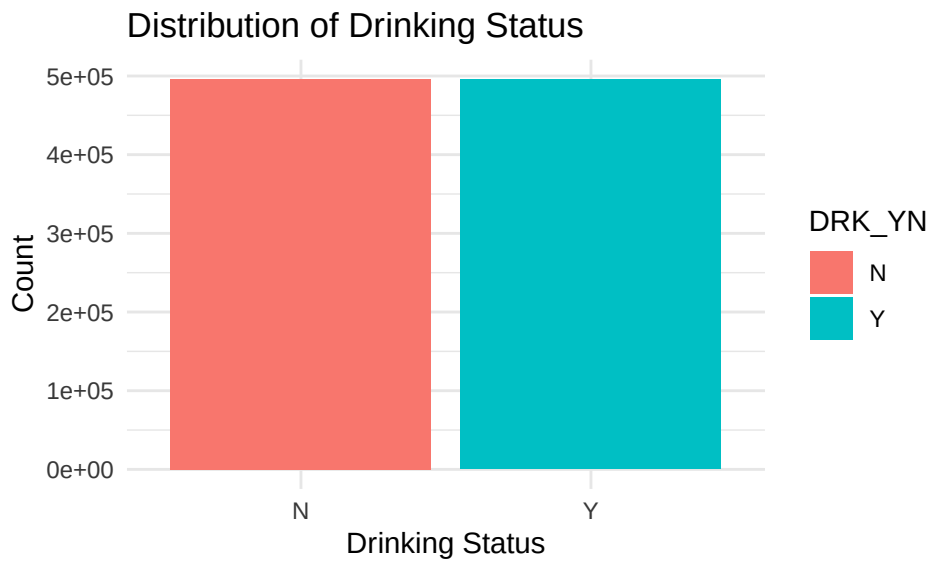
Hypothesis:

Null Hypothesis H_0: There is no association between Smoking State and Drinking Status.
Alternative Hypothesis H_a: There is a significant association between Smoking State and Drinking Status.

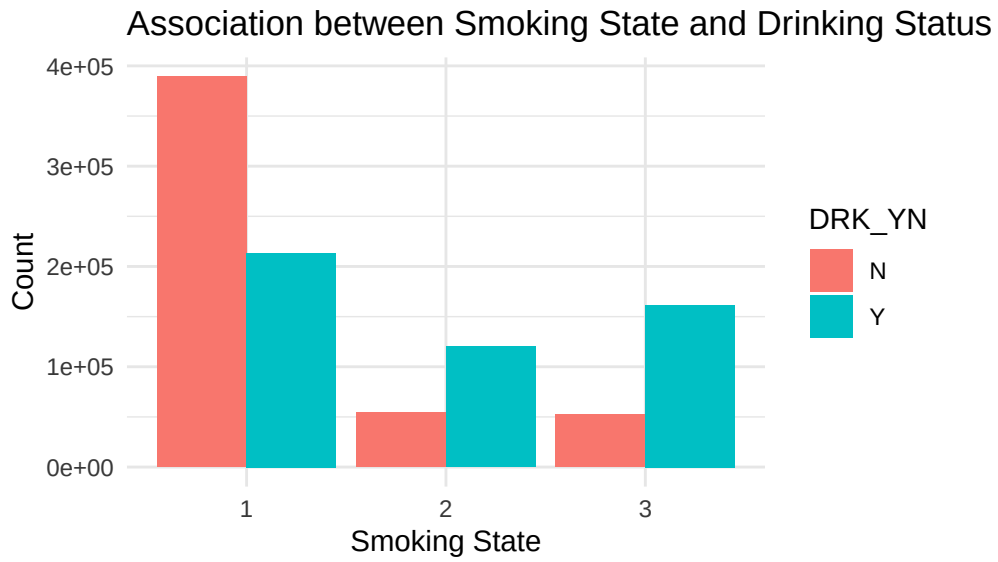
```r
# Bar chart for Smoking State
ggplot(sd_data, aes(x = SMK_stat_type_cd, fill = SMK_stat_type_cd)) +
  geom_bar() +
  ggtitle("Distribution of Smoking State") +
  xlab("Smoking State") +
  ylab("Count") +
  theme_minimal()
```




```r
# Bar chart for Drinking Status
ggplot(sd_data, aes(x = DRK_YN, fill = DRK_YN)) +
  geom_bar() +
  ggtitle("Distribution of Drinking Status") +
  xlab("Drinking Status") +
  ylab("Count") +
  theme_minimal()
```




```r
# Chi-Square Test
chisq_result <- chisq.test(sd_data$SMK_stat_type_cd, sd_data$DRK_YN)

ggplot(sd_data, aes(x = SMK_stat_type_cd, fill = DRK_YN)) +
  geom_bar(position = "dodge") +
  ggtitle("Association between Smoking State and Drinking Status") +
  xlab("Smoking State") +
  ylab("Count") +
  theme_minimal()
```




```r
# Display Chi-Square test results
print(chisq_result)
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  sd_data$SMK_stat_type_cd and sd_data$DRK_YN
## X-squared = 131811, df = 2, p-value < 2.2e-16
```

Since the **p-value is less than the commonly used significance level of 0.05**, we would reject the null hypothesis. This indicates that there is strong evidence to suggest that the distribution of drinking status is significantly dependent on different smoking states.

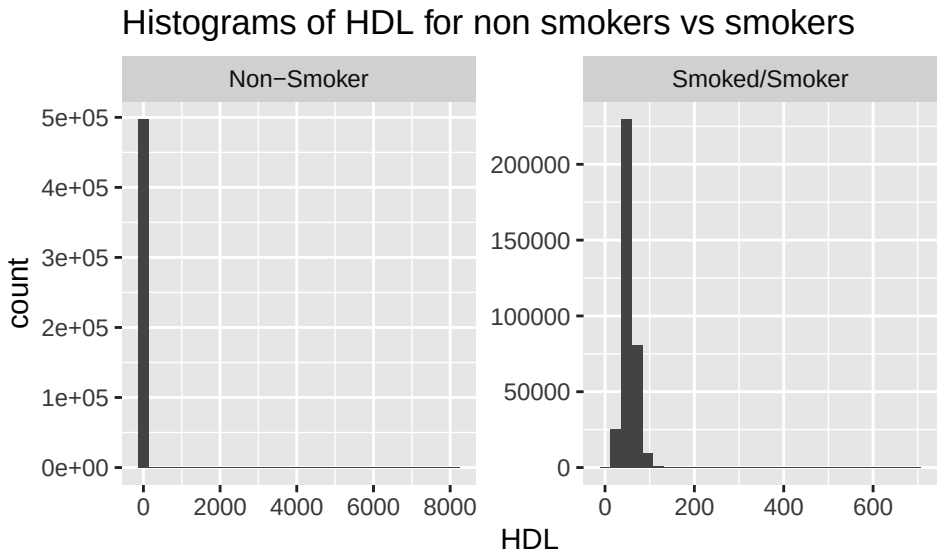
As we can see from the bar graphs, significant number of people who smoke also tend to drink.

How Smoking affects Good Cholesterol (HDL)?

Let's select age 20-60 where smoking and drinking seems to be more prevalent. Let us also combine people who have previously smoked but have given up also under Smoking category for better comparison.

```
primeage_and_smoke <- sd_data %>%
  filter(age >= 20 & age <= 60 & (SMK_stat_type_cd == 2 | SMK_stat_type_cd == 3))
primeage_and_nosmoke <- sd_data %>%
  filter(age >= 20 & age <= 60 & SMK_stat_type_cd == 1)

# Combine the filtered data frames for plotting
combined_data <- rbind(
  mutate(primeage_and_smoke, Smoking = "Smoked/Smoker"),
  mutate(primeage_and_nosmoke, Smoking = "Non-Smoker")
)

ggplot(combined_data, aes(x=HDL_chole)) +
geom_histogram() +
facet_wrap(~Smoking, scales="free")+
labs(title = "Histograms of HDL for non smokers vs smokers",
       x = "HDL",
       y = "count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

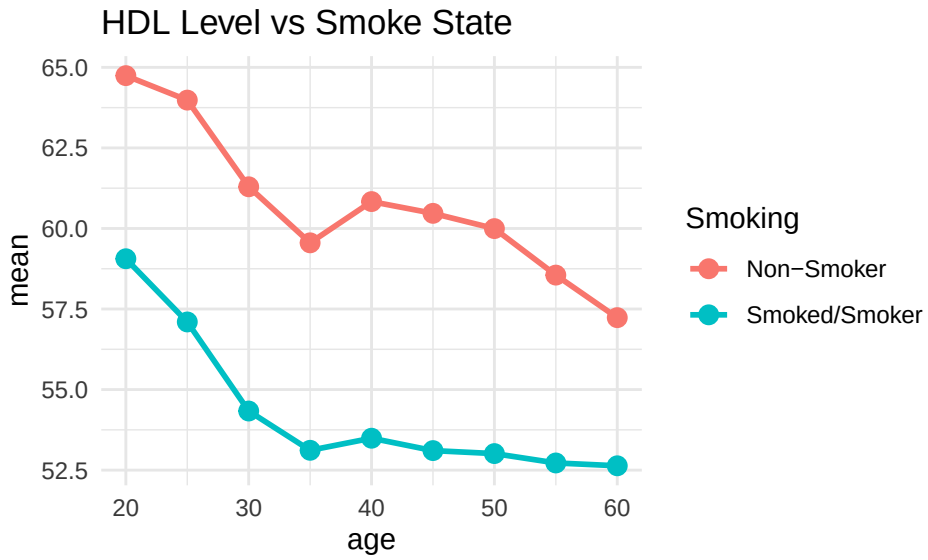


```r
plotdata_Cholestrol <- combined_data %>%
  group_by(age, Smoking) %>%
  summarize(n = count(HDL_chole),
            mean = mean(HDL_chole))
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

```r
# plot the means and standard errors by age
ggplot(plotdata_Cholestrol, aes(x = age,
                      y = mean,
                      group=Smoking,
                      color=Smoking)) +
  geom_point(size = 3) +
  geom_line(size = 1) +
  ggtitle("HDL Level vs Smoke State") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Sample means and **Standard Deviation** of HDL cholesterol for Smoked/Smoker and Non-smokers is shown below:

```r
smoker <- combined_data %>% filter(Smoking == "Smoked/Smoker")
non_smoker <- combined_data %>% filter(Smoking == "Non-Smoker")

tapply(combined_data$HDL_chole, combined_data$Smoking, mean)
```

```
##     Non-Smoker Smoked/Smoker
##       60.17699      53.53361
```

```
tapply(combined_data$HDL_chole, combined_data$Smoking, sd)
```

```
##    Non-Smoker Smoked/Smoker
##      19.23414      14.21395
```

The above result show that the HDL Cholesterol of non-smokers on average is higher than that of smoked/smoker which is expected.

The number of samples for non-smokers and smoked/smoker is as below:

```
table(select(combined_data, Smoking))
```

```
## Smoking
##    Non-Smoker Smoked/Smoker
##        497611        346516
```

Assumptions for smoked/smoker data:
1. Samples are random and independent.
2. The number of records is 346516, since sample size is greater than 30, by Central Limit Theorem we say that the sampling mean distribution is normal with mean = 53.53361 and standard deviation = 14.21395.

Assumptions for non-smoker data:
1. Samples are random and independent.
2. The number of records is 497611, since sample size is greater than 30, by Central Limit Theorem we say that the sampling mean distribution is normal with mean = 60.17699 and standard deviation = 19.23414.

The 95% confidence interval for smoked/smoker is:

```
confint(lm(smoker$HDL_chole~1),level=0.95)
```

```
##                2.5 %   97.5 %
## (Intercept) 53.48628 53.58094
```

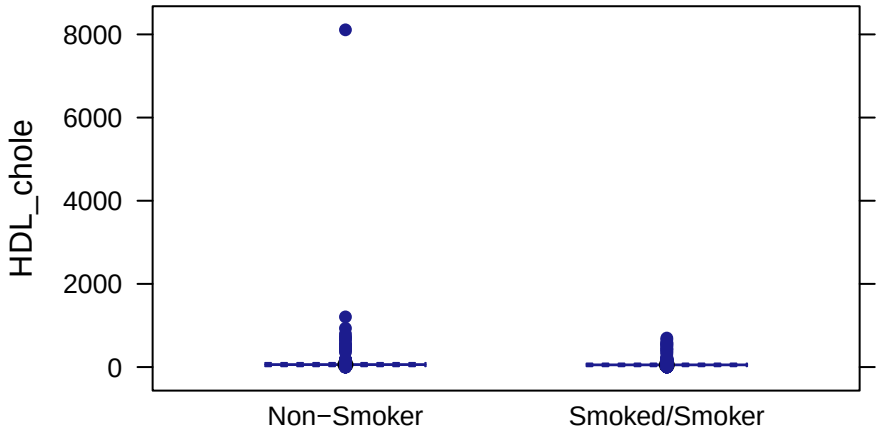
The 95% confidence interval for non-smoker is:

```
confint(lm(non_smoker$HDL_chole~1),level=0.95)
```

```
##                2.5 %   97.5 %
## (Intercept) 60.12355 60.23043
```

Box and Whisker Plot:

```
bwplot(HDL_chole ~ Smoking, data = combined_data)
```



There is outlier in HDL Cholesterol of non-smoker sample, but not much.

Hypothesis Test:

Null Hypothesis H_0**:** True difference between the means of non-smokers and smoked/smoker is 0.
Alternate Hypothesis H_a**:**True difference between the means of non-smokers and smoked/smoker is not equal to 0.

Test - Welch Two Sample t-test

Test statistic - t

```
t.test(HDL_chole ~ Smoking, data = combined_data)
```

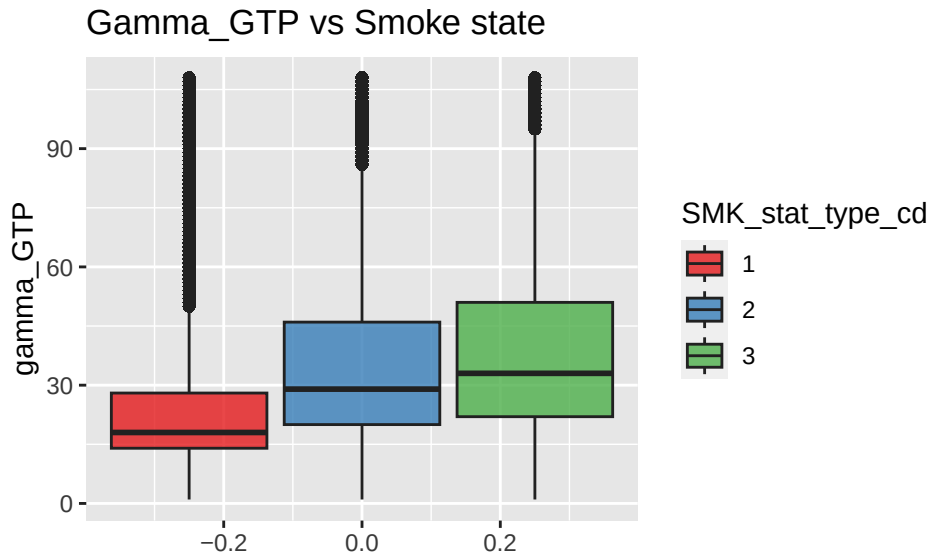
```
##
##  Welch Two Sample t-test
##
## data:  HDL_chole by Smoking
## t = 182.4, df = 841194, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Non-Smoker and group Smoked/Smoker is
## 95 percent confidence interval:
##  6.571998 6.714767
## sample estimates:
##    mean in group Non-Smoker mean in group Smoked/Smoker
##                    60.17699                    53.53361
```

Confidence Interval Approach
Since 0 is not present in the 95% confidence interval of (6.571998, 6.714767), we reject the null hypothesis at alpha = 0.05. We can say that there is difference in HDL cholesterol for smokers and non-smokers.

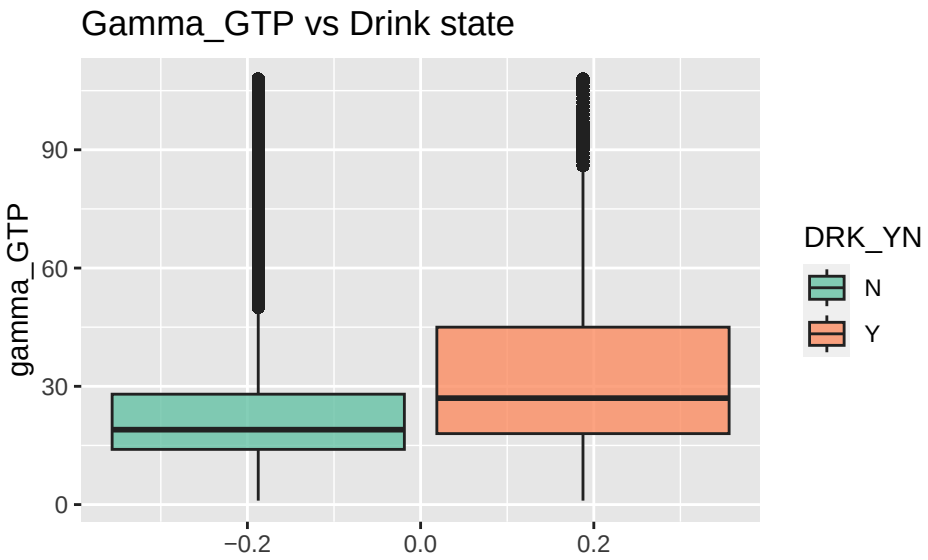
p-value approach
p-value < 2.2e-16 i.e. p-value is less than 2.2e-16 which is less than 0.05, reject the null hypothesis at alpha = 0.05. There is difference in HDL cholesterol for smokers and non-smokers.

How Smoking and Drinking affects Liver (Gamma GTP)?

```r
# Calculate the interquartile range (IQR) for triglyceride
Q1 <- quantile(sd_data$gamma_GTP, 0.25)
Q3 <- quantile(sd_data$gamma_GTP, 0.75)
IQR_value <- Q3 - Q1

# Define the lower and upper bounds for outliers
lower_outlier <- Q1 - 3 * IQR_value
upper_outlier <- Q3 + 3 * IQR_value

# Filter out rows with triglyceride values outside the bounds
filtered_gamma_GTP <- sd_data %>%
  filter(gamma_GTP >= lower_outlier, gamma_GTP <= upper_outlier)

ggplot(filtered_gamma_GTP, aes(y = gamma_GTP, fill = SMK_stat_type_cd)) +
  geom_boxplot(alpha = 0.8) +
  labs(title = "Gamma_GTP vs Smoke state") +
  scale_fill_brewer(palette = "Set1")
```




```r
ggplot(filtered_gamma_GTP, aes(y = gamma_GTP, fill = DRK_YN)) +
  geom_boxplot(alpha = 0.8) +
  labs(title = "Gamma_GTP vs Drink state") +
  scale_fill_brewer(palette = "Set2")
```



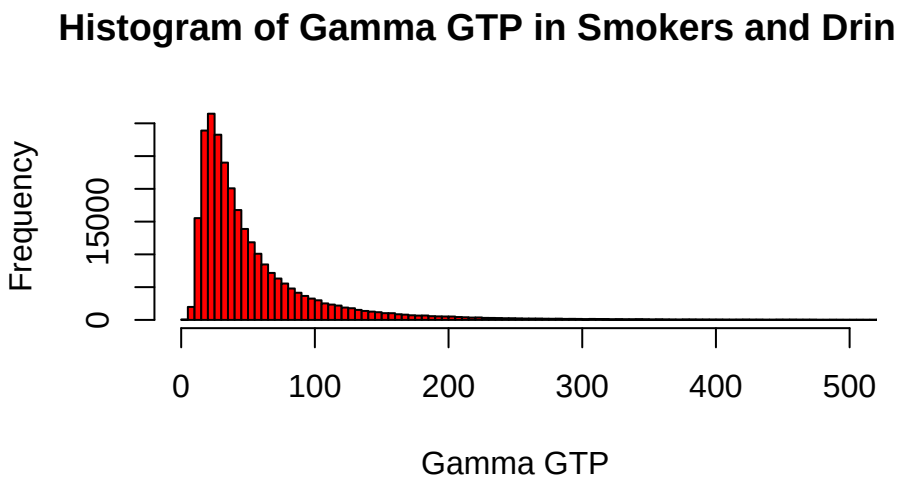
The box plot above confirms that we can find elevated levels of Gamma GTP in case of Smokers and Drinkers. As per a regular assumption, waistline is a contributing factor for the elevated levels of Gamma GTP that affects the functioning of liver. To confirm the assumption, we have the following claim:

Null Hypothesis H_0: Waistline affects increased levels of Gamma GTP in case of Smokers and Drinkers.

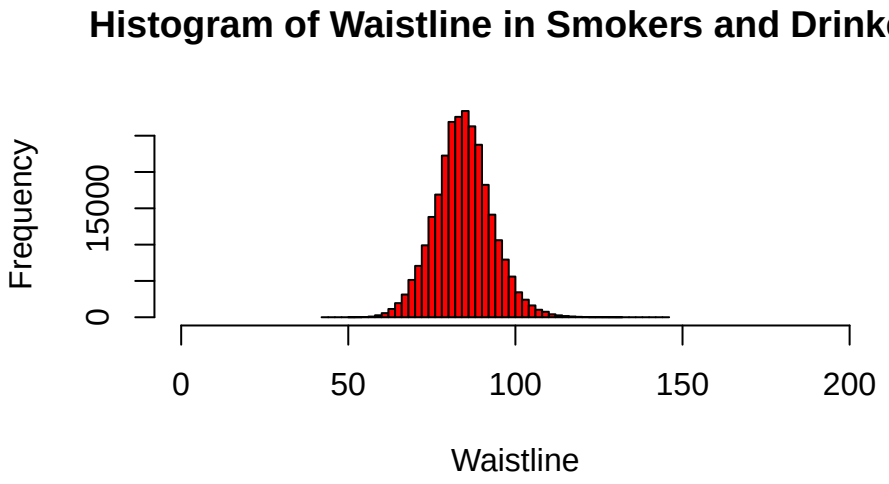
Alternate Hypothesis H_a: Waistline has no effect on increased levels of Gamma GTP in case of Smokers and Drinkers.

Histogram of Gamma GTP in Smokers and Drinkers:

```
smoke_drink_data <- sd_data %>%
  filter((SMK_stat_type_cd == 2 | SMK_stat_type_cd == 3) & DRK_YN == "Y")
hist(smoke_drink_data$gamma_GTP,
     main = "Histogram of Gamma GTP in Smokers and Drinkers",
     xlab = "Gamma GTP",
     ylab = "Frequency",
     xlim = c(0,500),
     breaks = 200,
     col = "red")
```



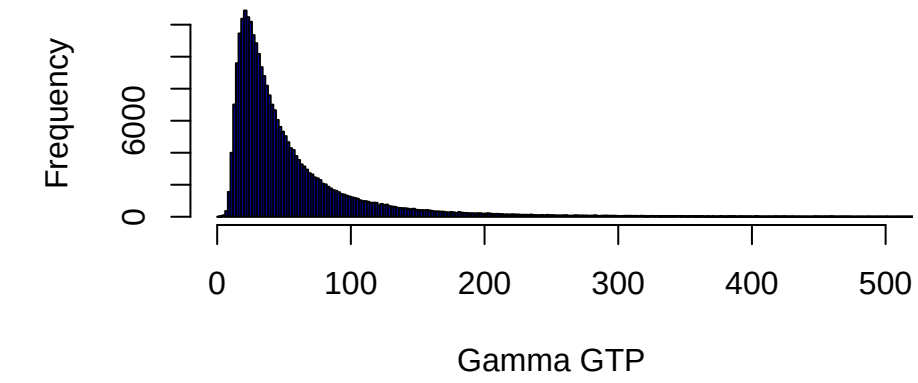

```
hist(smoke_drink_data$waistline,
     main = "Histogram of Waistline in Smokers and Drinkers",
     xlab = "Waistline",
     ylab = "Frequency",
     xlim = c(0,200),
     breaks = 50,
     col = "red")
```



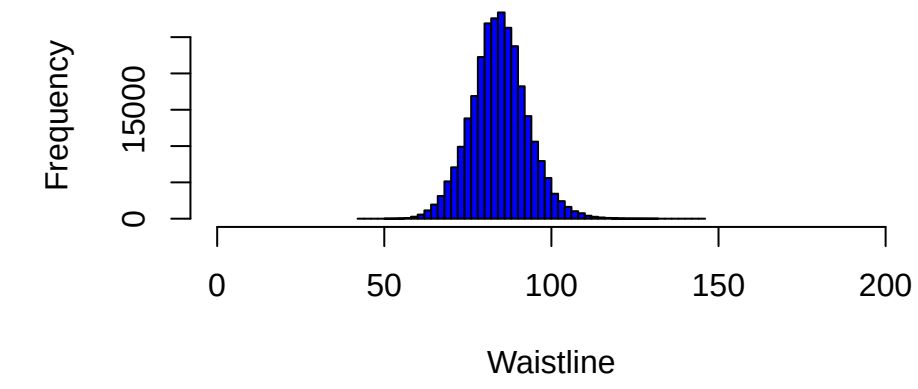
Histogram of Gamma GTP in non-Smokers and non-Drinkers:

```
nsmoke_ndrink_data <- sd_data %>%
  filter((SMK_stat_type_cd == 1) & DRK_YN == "N")
hist(smoke_drink_data$gamma_GTP,
     main = "Histogram of Gamma GTP in non-Smokers and non-Drinkers",
     xlab = "Gamma GTP",
     ylab = "Frequency",
```

```
     xlim = c(0,500),
     breaks = 500,
     col = "blue")
```

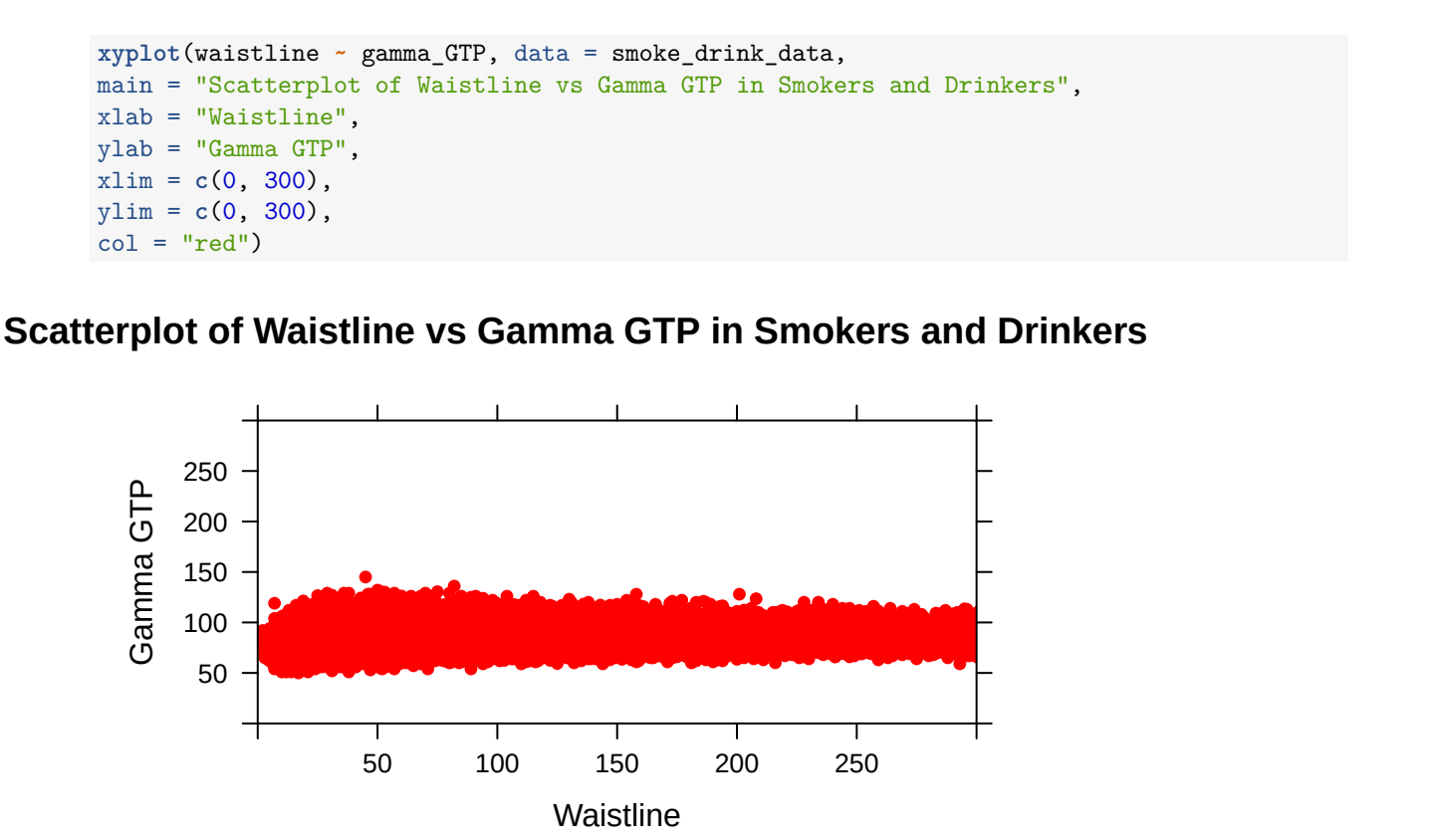



```
hist(smoke_drink_data$waistline,
     main = "Histogram of Waistline in non-Smokers and non-Drinkers",
     xlab = "Waistline",
     ylab = "Frequency",
     xlim = c(0,200),
     breaks = 50,
     col = "blue")
```



Scatterplot of Waistline in Smokers and Drinkers

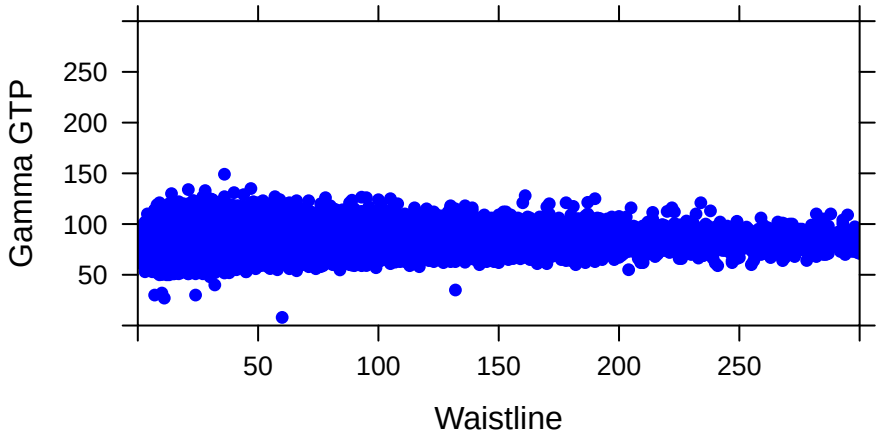
```
xyplot(waistline ~ gamma_GTP, data = smoke_drink_data,
main = "Scatterplot of Waistline vs Gamma GTP in Smokers and Drinkers",
xlab = "Waistline",
ylab = "Gamma GTP",
xlim = c(0, 300),
ylim = c(0, 300),
col = "red")
```



Scatterplot of Waistline in non-Smokers and non-Drinkers

```
xyplot(waistline ~ gamma_GTP, data = nsmoke_ndrink_data,
main = "Scatterplot of Waistline vs Gamma GTP in non-Smokers and non-Drinkers",
xlab = "Waistline",
ylab = "Gamma GTP",
xlim = c(0, 300),
ylim = c(0, 300),
col = "blue")
```

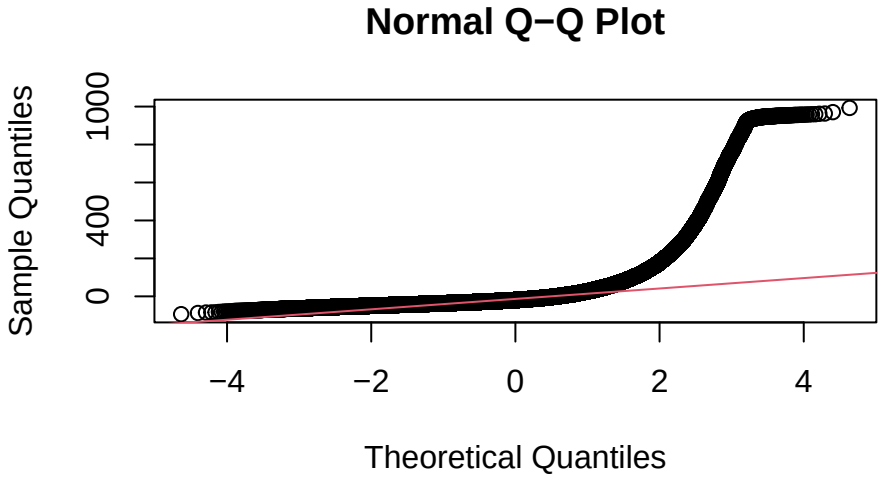
Scatterplot of Waistline vs Gamma GTP in non−Smokers and non−Drinkers



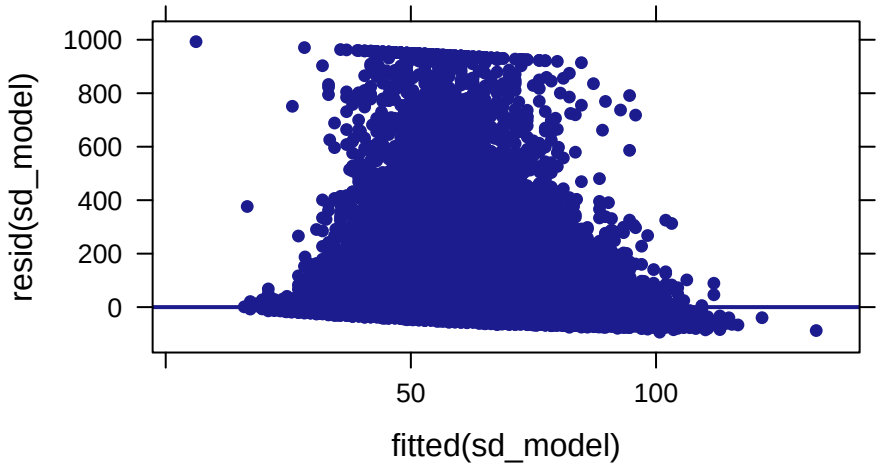
Fitting a linear model for waistline effects in Gamma GTP for Smokers and Drinkers:

```
sd_model <- lm(gamma_GTP ~ waistline, data = smoke_drink_data)
```

```
sd_residuals <- resid(sd_model)
qqnorm(sd_residuals)
qqline(sd_residuals, col = 2)
```




```
xyplot(resid(sd_model) ~ fitted(sd_model), data=smoke_drink_data, type=c("p", "r"))
```

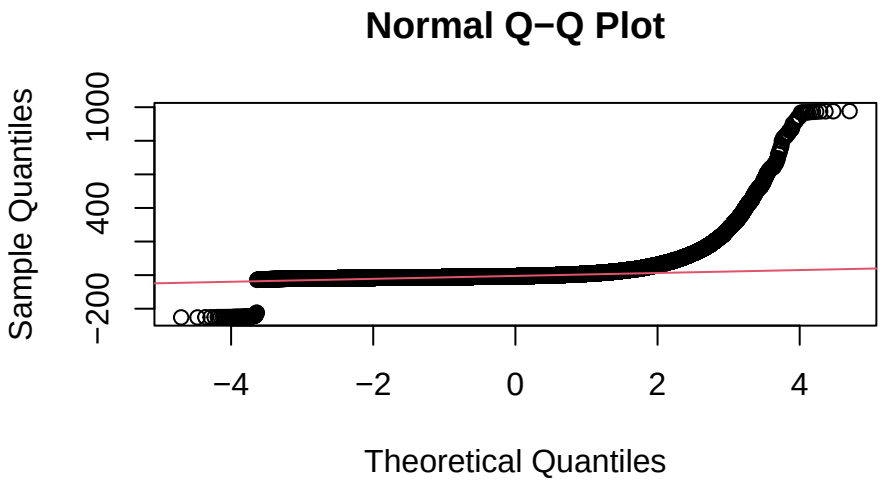


Assumptions for Linear model:
1. We assume that as per our interest of claim and scatter plot, the data follows a linear model.
2. The sample selected is a random sample and all considered records are independent of each other.
3. Sample size is 282,057. Since the sample size n is larger than 30, by Central Limit Theorem, the sample follows a normal distribution. 4. As per the QQ plot, the residuals follow a normal distribution and the expected variance of residuals is almost zero.

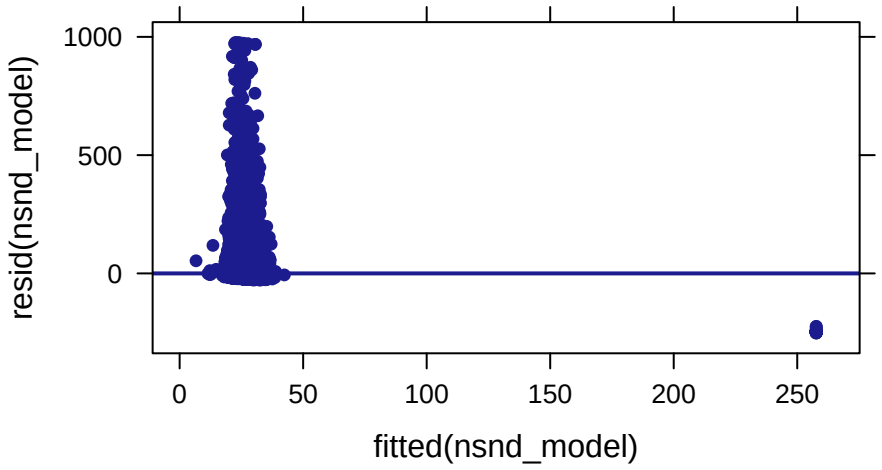
Fitting a linear model for waistline effects in Gamma GTP for non-Smokers and non-Drinkers:

```
nsnd_model <- lm(gamma_GTP ~ waistline, data = nsmoke_ndrink_data)
```

```
nsnd_residuals <- resid(nsnd_model)
qqnorm(nsnd_residuals)
qqline(nsnd_residuals, col = 2)
```



```
xyplot(resid(nsnd_model) ~ fitted(nsnd_model), data=nsmoke_ndrink_data, type=c("p", "r"))
```



Assumptions for Linear model:
1. We assume that as per our interest of claim and scatter plot, the data follows a linear model.
2. The sample selected is a random sample and all considered records are independent of each other.
3. Sample size is 389,010. Since the sample size n is larger than 30, by Central Limit Theorem, the sample follows a normal distribution.
4. As per the QQ plot, the residuals follow a normal distribution and the expected variance of residuals is almost zero.

Performing R^2 test on the linearly fitted model.

1. Waistline effects in Gamma GTP for non-Smokers and non-Drinkers:

```
summary(sd_model)
```

```
##
## Call:
## lm(formula = gamma_GTP ~ waistline, data = smoke_drink_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -93.74 -31.85 -19.39   5.15 992.81
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -45.38992    1.34642  -33.71   <2e-16 ***
## waistline     1.22801    0.01579   77.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.02 on 282055 degrees of freedom
## Multiple R-squared:  0.02099,    Adjusted R-squared:  0.02098
## F-statistic:  6046 on 1 and 282055 DF,  p-value: < 2.2e-16
```

The R^2 value of 0.02099 indicates that the linear regression model explains the variation of waistline and Gamma GTP only by 2.1%. This is not a very high explanation of variance as it might not consider all factors that are leading to the increase in Gamma GTP with respect to the waistline. The value of R^2 here is very low which suggests the model does not clearly explain the effects of increase or decrease in waistline as the only parameter which affects Gamma GTP in Smokers and Drinkers.

```
summary(nsnd_model)
```

```
##
## Call:
## lm(formula = gamma_GTP ~ waistline, data = nsmoke_ndrink_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -251.69  -10.60   -6.38    1.15  976.40
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.61181    0.24563   18.78   <2e-16 ***
## waistline    0.25333    0.00306   82.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.3 on 389008 degrees of freedom
## Multiple R-squared:  0.01731,    Adjusted R-squared:  0.01731
## F-statistic:  6853 on 1 and 389008 DF,  p-value: < 2.2e-16
```

The R^2 value of 0.01731 indicates that the linear regression model explains the variation of waistline and Gamma GTP only by 1.7%. This is not a very high explanation of variance as it might not consider all factors that are leading to the increase in Gamma GTP with respect to the waistline. The value of R^2 here is very low which suggests the model does not clearly explain the effects of increase or decrease in waistline as the only parameter which affects Gamma GTP in non-Smokers and non-Drinkers.

Conclusion of the test:. As per our assumption, the increase in size of waistline had direct effect on the increase in Gamma GTP in case of Smokers and Drinkers and vice-versa in case of non-Smokers and non-Drinkers. The R^2 test confirms that the waistline does not necessarily have any effect on the elevated levels of Gamma GTP in case of Smokers and Drinkers and normal levels of Gamma GTP in case of non-Smokers and non-Drinkers. This can be concluded because the adjusted R^2 value in both cases does not vary by a lot.
As per this test, we conclude that we reject our claim H_0 and say that, Waistline has no significant effect on increased levels of Gamma GTP in case of Smokers and Drinkers.

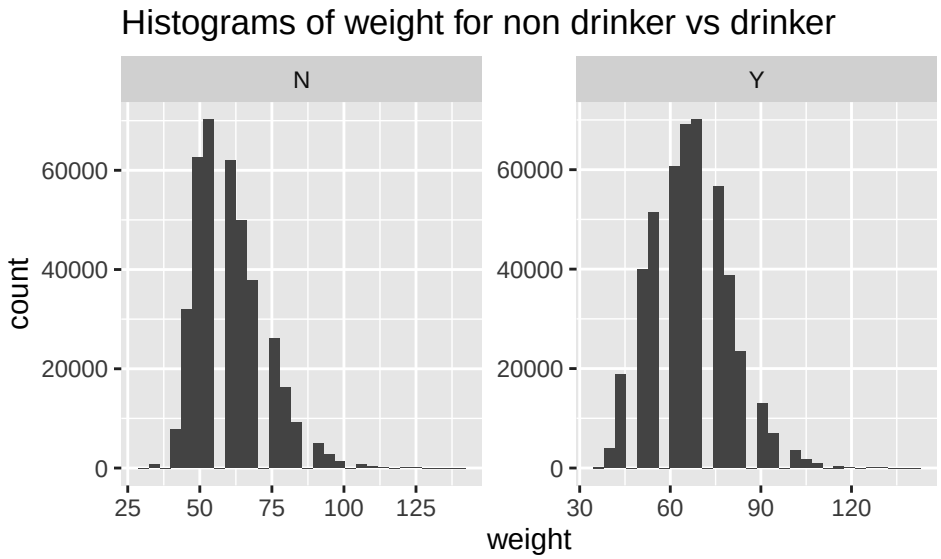
Is there any relationship associated with Weight and Drinking?

Let's select age 20-60 where smoking and drinking seems to be more prevalent.

```
primeage_and_drink <- sd_data %>%
  filter(age >= 20 & age <= 60 & (DRK_YN == 'Y'))
primeage_and_nodrink <- sd_data %>%
  filter(age >= 20 & age <= 60 & DRK_YN == 'N')

# Combine the filtered data frames for plotting
combined_data <- rbind(
  mutate(primeage_and_drink, Drink = "Y"),
  mutate(primeage_and_nodrink, Drink = "N")
```

```
)

ggplot(combined_data, aes(x=weight)) +
geom_histogram() +
facet_wrap(~Drink, scales="free")+
labs(title = "Histograms of weight for non drinker vs drinker",
       x = "weight",
       y = "count")
```

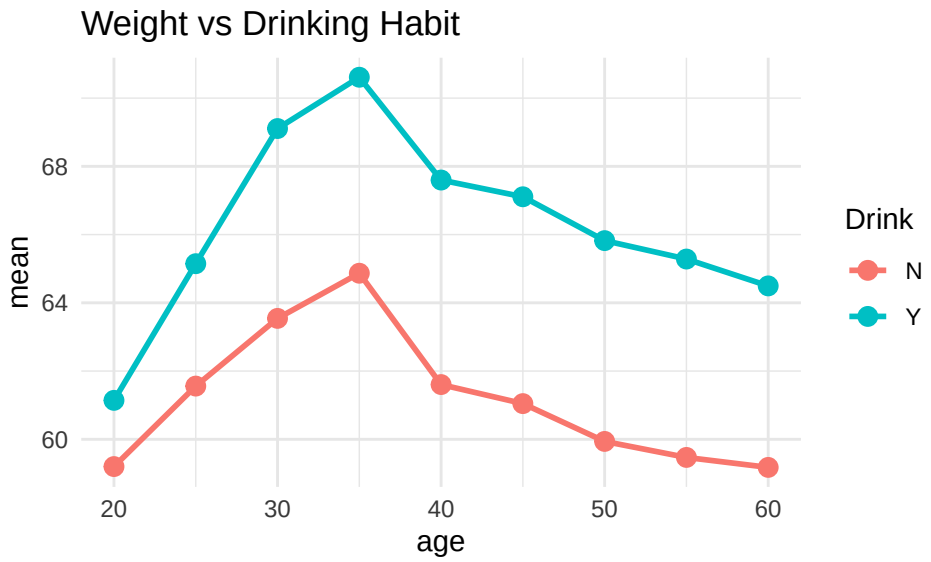
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
plotdata_weight <- combined_data %>%
  group_by(age, Drink) %>%
  summarize(n = count(weight),
            mean = mean(weight))
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

```
# plot the means and standard errors by age
ggplot(plotdata_weight, aes(x = age,
                     y = mean,
                     group=Drink,
                     color=Drink)) +
  geom_point(size = 3) +
  geom_line(size = 1) +
  ggtitle("Weight vs Drinking Habit") +
  theme_minimal()
```



Sample means and **Standard Deviation** of weight for Drinkers and Non-drinkers is shown below:

```
drinker <- combined_data %>% filter(Drink == "Y")
non_drinker <- combined_data %>% filter(Drink == "N")

tapply(combined_data$weight, combined_data$Drink, mean)
```

```
##        N        Y
## 60.78834 66.88935
```

```
tapply(combined_data$weight, combined_data$Drink, sd)
```

```
##        N        Y
## 11.92127 12.74912
```

The above result show that the Weight of drinkers on average is higher than that of non-drinkers which is expected.

The number of samples for non-smokers and smoked/smoker is as below:

```
table(select(combined_data, Drink))
```

```
## Drink
##      N      Y
## 384335 459792
```

Assumptions for Drinker data:
1. Samples are random and independent.
2. The number of records is 459792, since sample size is greater than 30, by Central Limit Theorem we say that the sampling mean distribution is normal with mean = 66.88935 and standard deviation = 12.74912.

Assumptions for non-drinker data:
1. Samples are random and independent.

2. The number of records is 384335, since sample size is greater than 30, by Central Limit Theorem we say that the sampling mean distribution is normal with mean = 60.78834 and standard deviation = 11.92127.

The 95% confidence interval for Drinker is:

```
confint(lm(drinker$weight~1),level=0.95)
```

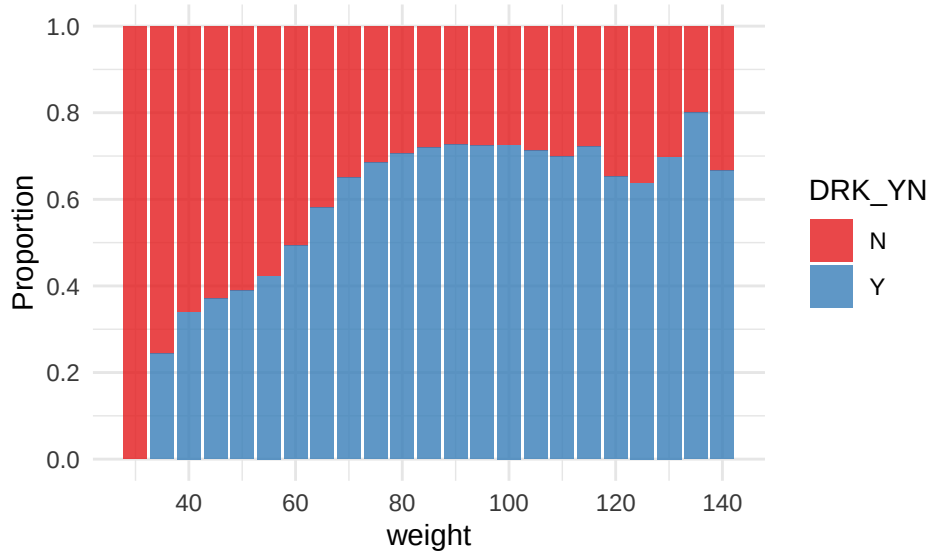
```
##                2.5 %   97.5 %
## (Intercept) 66.8525 66.92621
```

The 95% confidence interval for non-drinker is:

```
confint(lm(non_drinker$weight~1),level=0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 60.75065 60.82602
```

```
ggplot(combined_data, aes(x = weight, fill = DRK_YN)) +
  geom_bar(position = "fill", alpha = 0.8) +
  labs(y = "Proportion") +
  scale_fill_brewer(palette = "Set1") +
  scale_y_continuous(breaks = seq(0, 1, .2)) +
  scale_x_continuous(breaks = seq(0, max(combined_data$weight), by = 20), labels = seq(0, max(combined_
  theme_minimal()
```



Hypothesis Test:

Null Hypothesis H_0**:** True difference between the means of non-drinkers and drinkers is 0.
Alternate Hypothesis H_a**:**True difference between the means of non-drinkers and drinkers is not equal to 0.

Test - Welch Two Sample t-test

Test statistic - t

```r
t.test(weight ~ Drink, data = combined_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by Drink
## t = -226.86, df = 833607, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group N and group Y is not equal to 0
## 95 percent confidence interval:
##  -6.153730 -6.048308
## sample estimates:
## mean in group N mean in group Y
##        60.78834        66.88935
```

Confidence Interval Approach
Since 0 is not present in the 95% confidence interval of (6.571998, 6.714767), we reject the null hypothesis at alpha = 0.05. We can say that there is difference in weight for non-drinkers and drinkers.

p-value approach
p-value < 2.2e-16 i.e. p-value is less than 2.2e-16 which is less than 0.05, reject the null hypothesis at alpha = 0.05. There is difference in weight for non-drinkers and drinkers.