Final project 151

Jimmy Zhang, Yijun Shen, Wenzhe Shi, Min Jin

4/6/2021

Introduction

Nowadays, more and more people start to smoke. The cigarette production in China increases by 20% in 2020. The increase of cigarette usage is raising different problems in society. For the project, we are trying to figure out how smoking influences sleep patterns. Because a lot of people tend to use e-cigarette due to its convenience, we are also adding e-cigarette as a variable.

Reserach Questions

For the proposal, we are choosing the Questionnaire data set sleep disorders and cigarette use as our dataset, and we are trying to figure out the relationship between smoking and sleep pattern. From these data set, we are trying to figure out the following questions:

1. The relationship between start age of smoke and snore;
2. The difference of average sleeping hour of e-cigarette users(when e-cigarette times>= 2) and non-e-cigarette users;
3. The relationship between time of quitting cigarettes and number of times feeling drowsy during a day.

Data Description

All the data used in this study is provided by Centers for Disease Control and Prevention(CDC) and selected from the NHANES 2017-18 data wave. The specific dataset used in this study are **Sleep Disorders** and **Smoking - Cigarette Use** from the Questionnaire Data.

```
# import dataset
smoking_cigarette_use<-read.xport("SMQ_J.XPT")
sleeping<-read.xport("SLQ_J.XPT")

# examine the dataset
head(smoking_cigarette_use)
```

```
##    SEQN SMQ020 SMD030 SMQ040 SMQ050Q SMQ050U SMD057 SMQ078 SMD641 SMD650 SMD093
## 1 93705      1     16      3      30       4      5     NA     NA     NA     NA
## 2 93706      2     NA     NA      NA      NA     NA     NA     NA     NA     NA
## 3 93707     NA     NA     NA      NA      NA     NA     NA     NA     NA     NA
## 4 93708      2     NA     NA      NA      NA     NA     NA     NA     NA     NA
## 5 93709      1     15      1      NA      NA     NA      1     30      5      1
## 6 93711      2     NA     NA      NA      NA     NA     NA     NA     NA     NA
##        SMDUPCA   SMD100BR SMD100FL SMD100MN SMD100LN SMD100TR SMD100NI SMD100CO
```

```
## 1         <NA>       <NA>       NA       NA       NA       NA       NA       NA
## 2         <NA>       <NA>       NA       NA       NA       NA       NA       NA
## 3         <NA>       <NA>       NA       NA       NA       NA       NA       NA
## 4         <NA>       <NA>       NA       NA       NA       NA       NA       NA
## 5 * NO MATCH * WAVE GREEN        1        1        3       NA       NA       NA
## 6         <NA>       <NA>       NA       NA       NA       NA       NA       NA
##   SMQ621 SMD630 SMQ661 SMQ665A SMQ665B SMQ665C SMQ665D SMQ670 SMQ848 SMQ852Q
## 1     NA     NA     NA      NA      NA      NA      NA     NA     NA      NA
## 2     NA     NA     NA      NA      NA      NA      NA     NA     NA      NA
## 3      1     NA     NA      NA      NA      NA      NA     NA     NA      NA
## 4     NA     NA     NA      NA      NA      NA      NA     NA     NA      NA
## 5     NA     NA     NA      NA      NA      NA      NA      2     NA      NA
## 6     NA     NA     NA      NA      NA      NA      NA     NA     NA      NA
##   SMQ852U SMQ890 SMQ895 SMQ900 SMQ905 SMQ910 SMQ915 SMAQUEX2
## 1      NA      2     NA      2     NA      2     NA        1
## 2      NA      2     NA      2     NA      2     NA        1
## 3      NA     NA     NA     NA     NA     NA     NA        2
## 4      NA      2     NA      2     NA      2     NA        1
## 5      NA      1      0      1      0      1      0        1
## 6      NA      2     NA      2     NA      2     NA        1
```

```
head(sleeping)
```

```
##    SEQN SLQ300 SLQ310 SLD012 SLQ320 SLQ330 SLD013 SLQ030 SLQ040 SLQ050 SLQ120
## 1 93705  23:00  07:00    8.0  23:00  07:00    8.0      2      0      2      0
## 2 93706  23:30  10:00   10.5  00:30  12:00   11.5      1      0      2      1
## 3 93708  22:30  06:30    8.0  22:30  06:30    8.0      9      0      2      2
## 4 93709  22:30  05:30    7.0  22:30  05:00    6.5      1      0      2      1
## 5 93711  22:00  05:00    7.0  23:00  08:00    9.0      2      1      1      3
## 6 93712  23:30  07:00    7.5  01:00  10:00    9.0      1      1      2      2
```

From the first couple observations of two datasets we've already seen many NAs. Therefore, before we answer any of the research questions, let's first clean up the data and join the two dataset to create a new dataset for each question.

```
# join the two table by "SEQN"
smoking_sleeping<-full_join(smoking_cigarette_use, sleeping, by="SEQN")
```

Now let's focus on question 1: relationship between start age of smoking and snoring.

Relationship between start age of smoking and snoring

Data Description

The focus of this question uses the following two variables:

- **SLQ030**: records how often the each observation snore while sleeping in the past 12 months
  - 0: Never
  - 1: Rarely - 1-2 nights a week
  - 2: Occasionally - 3-4 nights a week

  - 3: Frequently - 5 or more nights a week
  - 7: Refused
  - 9: Don't know
  - .: missing

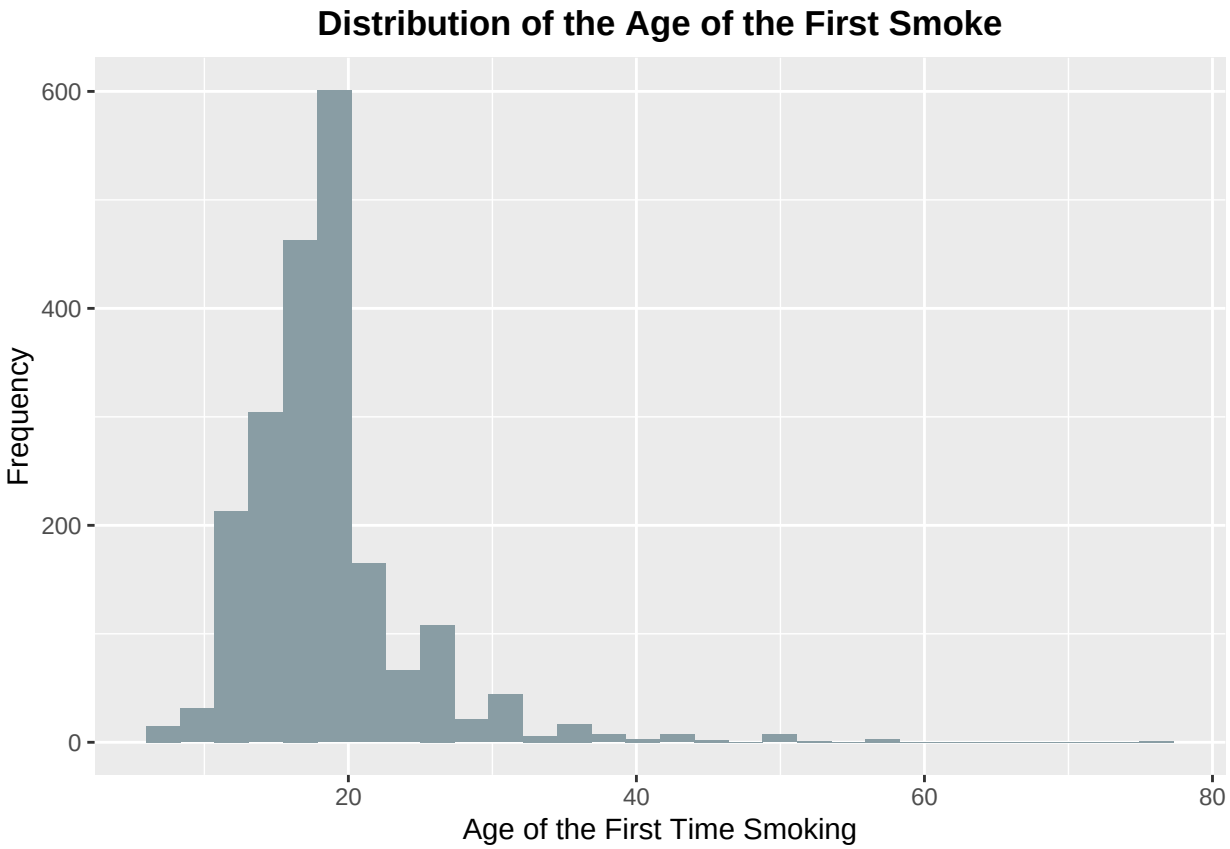
- **SMD030**: Age started smoking cigarettes regularly
  - 7 to 76: Range of Values
  - 0: Never smoked cigarettes regularly
  - 777: Refused
  - 999: Don't know
  - .: Missing

```
#clean the data
# exclude observations that are refused/don't know/missing
smoking_sleeping$SMD030[smoking_sleeping$SMD030 == 999 | smoking_sleeping$SMD030 == 0]<-NA

# select the variables needed for analyses and filter out observations that refused/don't know/missing
snore_1stsmk<-smoking_sleeping %>%
  select(SEQN, SMD030, SLQ030)%>%
  filter(SMD030 !=777,  SLQ030!= 7, SLQ030 != 9)
# remove NA from this dataset
snore_1stsmk <- na.omit(snore_1stsmk)

# change the numebric value of SLQ030 to factors of three levels
snore_1stsmk$SLQ030 <- factor(snore_1stsmk$SLQ030, labels = c("Never", "Rarely", "Occasionally", "Frequ
# check the summary statistics of the dataset
summary(snore_1stsmk)
```

```
##       SEQN            SMD030               SLQ030
##  Min.   : 93705   Min.   : 7.00   Never       :518
##  1st Qu.: 95915   1st Qu.:15.00   Rarely      :454
##  Median : 98161   Median :18.00   Occasionally:396
##  Mean   : 98258   Mean   :18.38   Frequently  :717
##  3rd Qu.:100578   3rd Qu.:20.00
##  Max.   :102956   Max.   :76.00
```

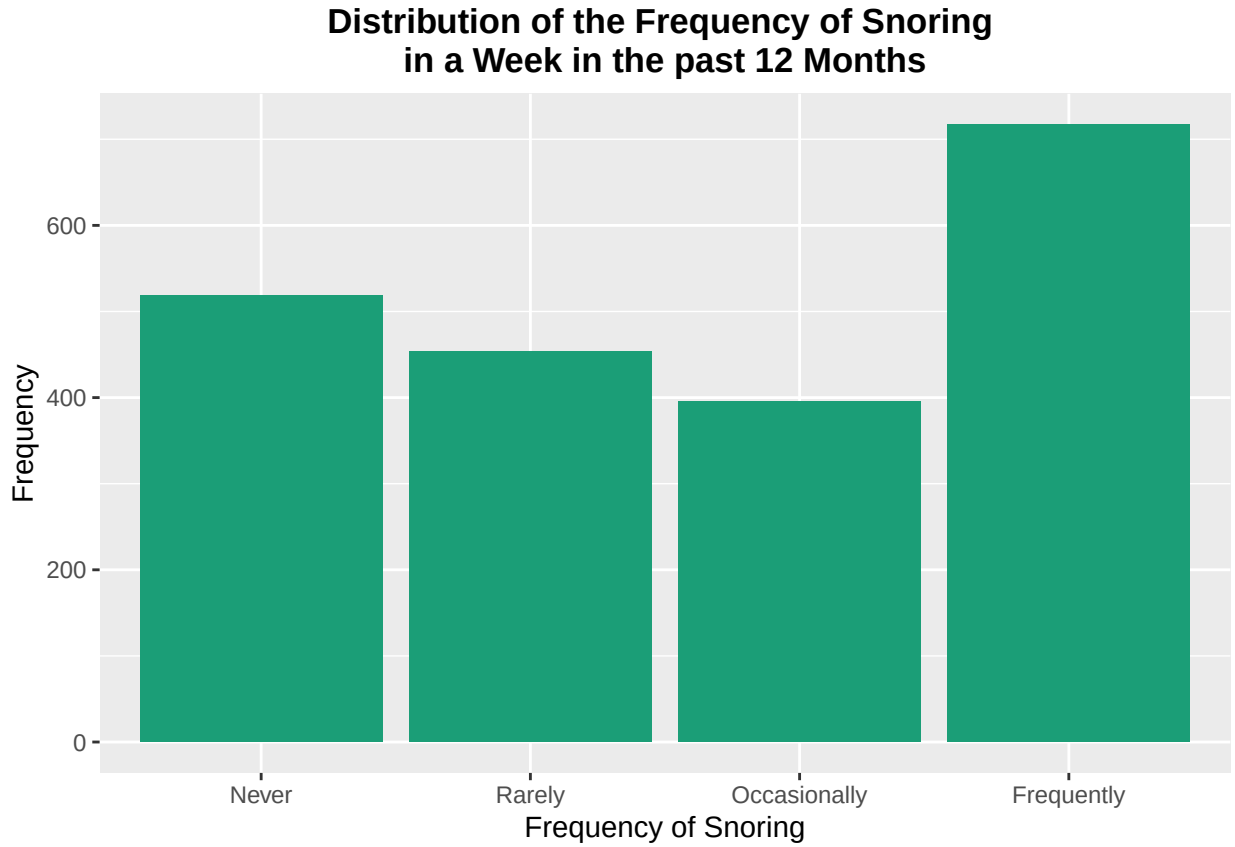
The following graphs show the distribution of each variable

```
# graph the distribution of smoke
ggplot(data = snore_1stsmk, aes(x = SMD030, fill = "yellow")) +
  geom_histogram()+
  labs(x = "Age of the First Time Smoking", y = "Frequency", title = "Distribution of the Age of the Fi
  theme(plot.title = element_text(face = "bold", hjust = 0.5),legend.position = "none")+
  scale_fill_manual(values=wes_palette(n=3, name="Royal1"))
```



From the graph, we can see that the age distribution is right skewed (mean > median) which means that most people starts smoking at around 20 years old with some extreme values at the right tail of the distribution.

```
# graph the distribution of frequency of snoring
ggplot(data = snore_1stsmk, aes(x = SLQ030, fill = "blue")) +
  geom_bar()+
  labs(x = "Frequency of Snoring", y = "Frequency", title = "Distribution of the Frequency of Snoring \n
  theme(plot.title = element_text(face = "bold", hjust = 0.5),legend.position = "none")+
  scale_fill_brewer(palette="Dark2")
```

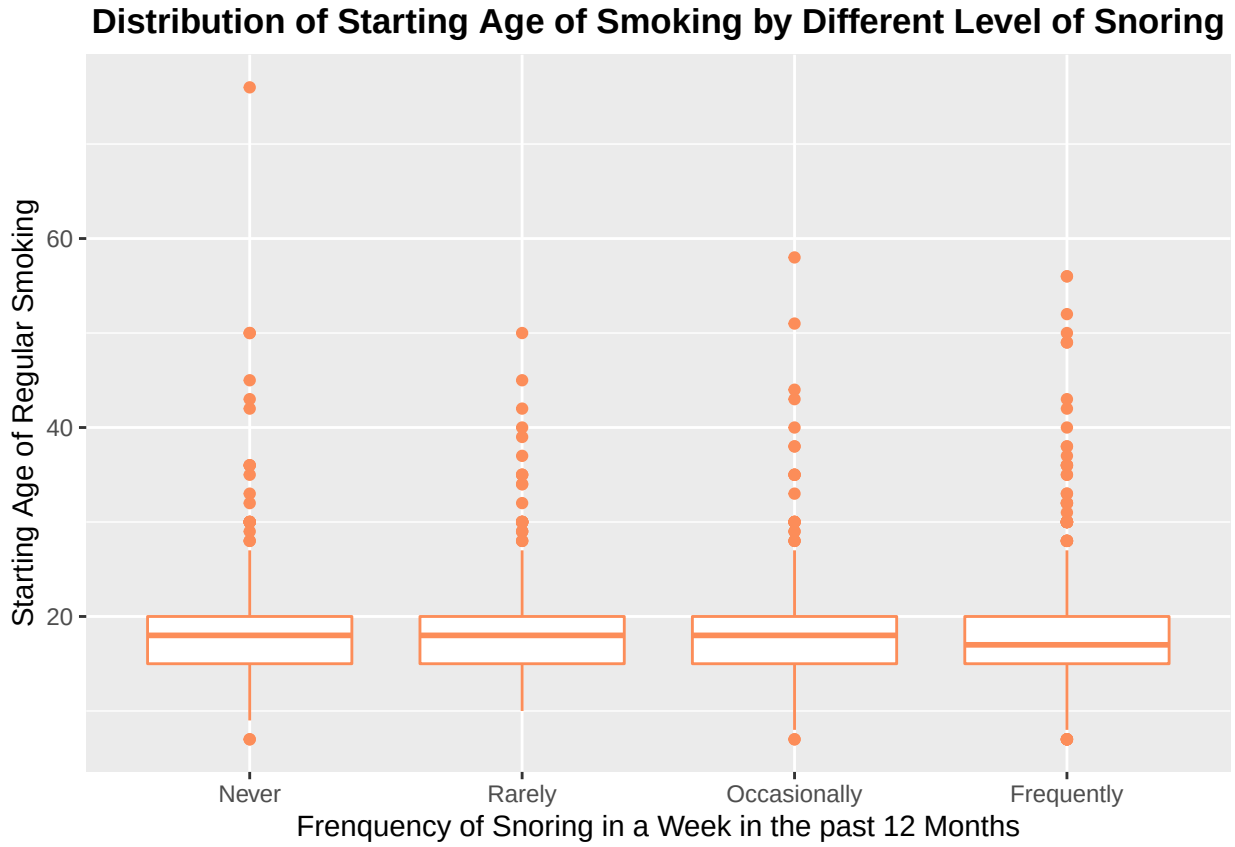


From the bar plot above, we can see that most people snore frequently in the past 12 months (about 5 or more nights a week). This is interesting because we can already see some pattern for the smokers.

Next let's put the two variables together and examine their relationship

```
#examine the summary statistics of starting age of smoking by different level of snoring
snore_1stsmk%>%
  select(SMD030, SLQ030)%>%
  group_by(SLQ030)%>%
  summarise("Min" = min(SMD030), "Mean" = mean(SMD030), "Max" = max(SMD030))%>%
  kable(digits = 2, booktabs = T,col.names = c("Frequency of Snoring", "Min","Mean","Max"))%>%
   kable_material(c("striped", "hover"))

# examine the distribution of starting age of smoking by different level of snoring using boxplot
snore_1stsmk%>%
  select(SMD030, SLQ030)%>%
  group_by(SLQ030)%>%
  ggplot(aes(x = SLQ030, y = SMD030, col = "green'"))+
  geom_boxplot()+
  labs(x = "Frenquency of Snoring in a Week in the past 12 Months", y = "Starting Age of Regular Smoking
       title = "Distribution of Starting Age of Smoking by Different Level of Snoring")+
    theme(plot.title = element_text(face = "bold", hjust = 0.5),legend.position = "none")+
  scale_color_brewer( palette="Spectral")
```

Frequency of Snoring	Min	Mean	Max
Never	7	18.46	76
Rarely	10	18.48	50
Occasionally	7	18.45	58
Frequently	7	18.22	56



From the summary table and the boxplot above, there's no clear difference or trend of the start age of smoking among people who never snore, rarely snore, occasionally snore, or frequently snore. This discovery reject our initial conjecture that young smokers who start smoking at early ages are more likely to develop sleep disorders since their respiratory systems may not be as good as a healthy or smokers who start late.

Let's add in another categorical variable SMQ670 (whether or not a smoker tries to quit smoking for one day or longer during the past 12 months) to see if quitting smoke can affect the relationship among the three variables.

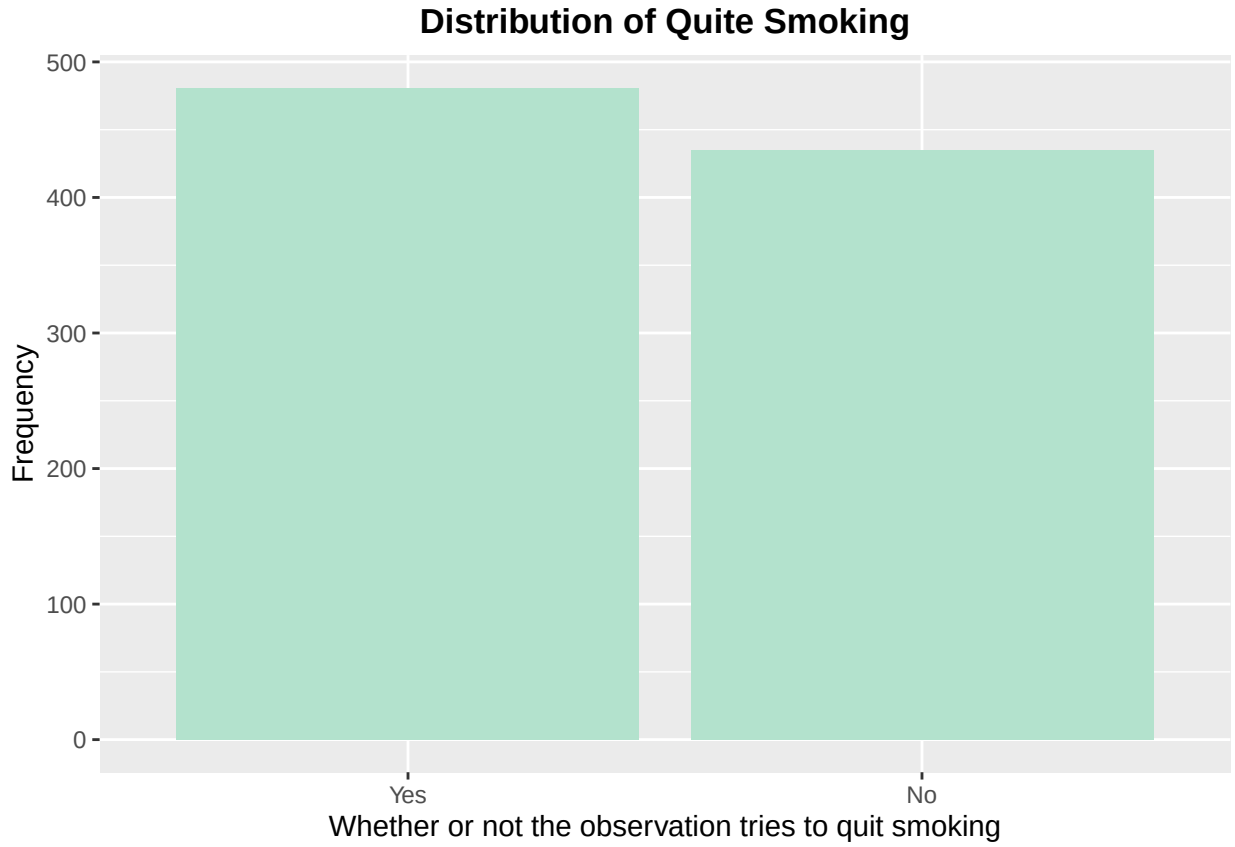
- **SMQ670**: Whether or not the observation tries to quite smoking for one day or longer in the past 12 months
  - 1: Yes
  - 2: No

```
snore_1stsmk_quit<-smoking_sleeping %>%
  select(SEQN, SMD030, SLQ030, SMQ670)%>%
  filter(SMD030 !=777,  SLQ030!= 7, SLQ030 != 9, SMQ670!= 7, SMQ670!=9, SMD030!= 999, SMD030!= 0)
```

```r
# remove NA from this dataset
snore_1stsmk_quit <- na.omit(snore_1stsmk_quit)

# change the numeric value of SLQ030 to factors of three levels
snore_1stsmk_quit$SLQ030 <- factor(snore_1stsmk_quit$SLQ030, labels = c("Never", "Rarely", "Occasionally
# change the numeric value of SMQ670 to factors of two levels
snore_1stsmk_quit$SMQ670 <- factor(snore_1stsmk_quit$SMQ670, labels = c("Yes", "No"))
```

The following graph shows the distribution of SMQ670.

```r
ggplot(data = snore_1stsmk_quit, aes(x = SMQ670, fill = "yellow")) + geom_bar()+ labs(x = "Whether or n
  theme(plot.title = element_text(face = "bold", hjust = 0.5),legend.position = "none")+
  scale_fill_brewer( palette="Pastel2")
```

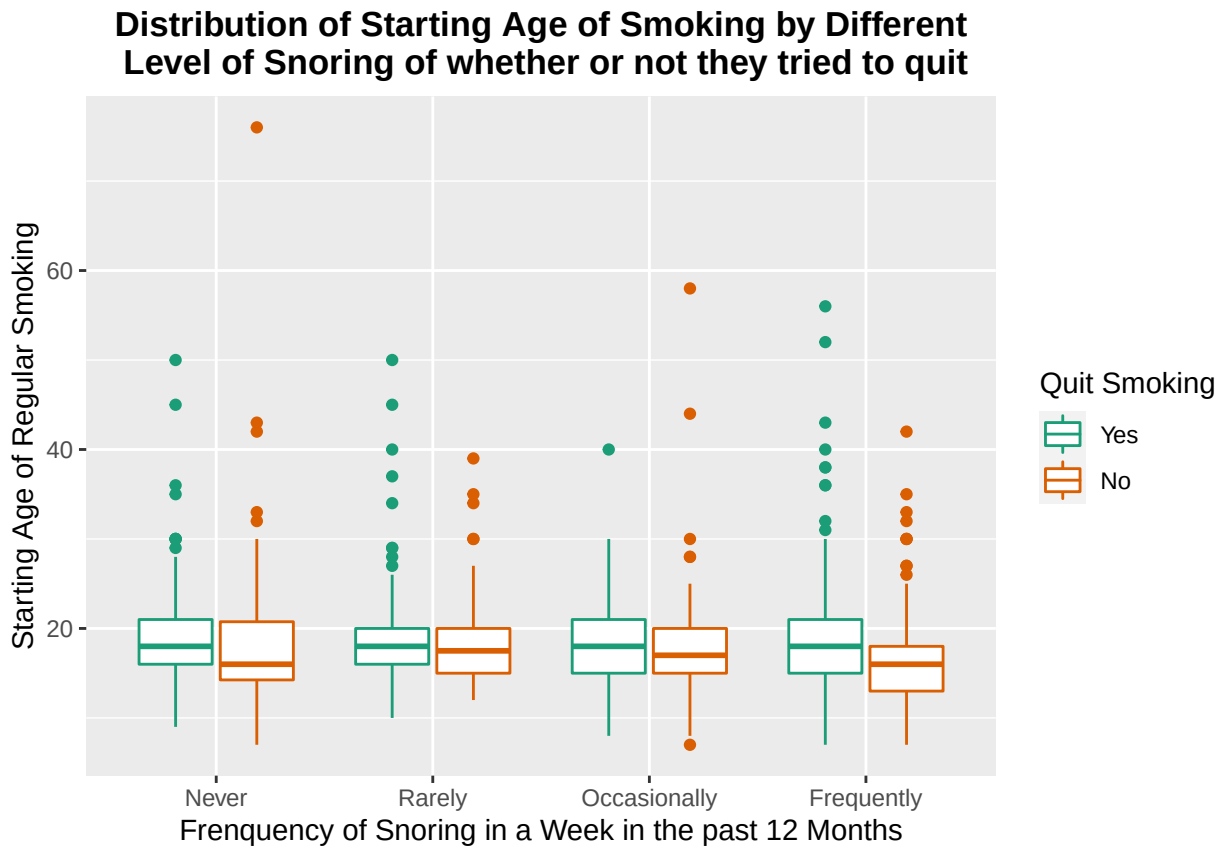


We see that the distribution of quit smoking is about equal for both groups, let's add this variable into the previous graph and summary table

```r
#examine the summary statistics of starting age of smoking by different level of snoring of whether or
snore_1stsmk_quit%>%
  select(SMD030, SLQ030,SMQ670)%>%
  group_by(SLQ030,SMQ670)%>%
  summarise("Min" = min(SMD030), "Mean" = mean(SMD030), "Max" = max(SMD030))%>%
  kable(digits = 2, booktabs = T,col.names = c("Frequency of Snoring", "Quit Smoking","Min","Mean","Max
   kable_material(c("striped", "hover"))
```

Frequency of Snoring	Quit Smoking	Min	Mean	Max
Never	Yes	9	19.24	50
Never	No	7	18.10	76
Rarely	Yes	10	18.80	50
Rarely	No	12	18.48	39
Occasionally	Yes	8	18.62	40
Occasionally	No	7	17.78	58
Frequently	Yes	7	19.41	56
Frequently	No	7	16.59	42

```
# examine the distribution of starting age of smoking by different level of snoring of whether or not t

snore_1stsmk_quit%>%
  select(SMD030, SLQ030,SMQ670)%>%
  group_by(SLQ030)%>%
  ggplot(aes(x = SLQ030, y = SMD030, col = SMQ670))+
  geom_boxplot()+
  labs(x = "Frenquency of Snoring in a Week in the past 12 Months", y = "Starting Age of Regular Smokin
       title = "Distribution of Starting Age of Smoking by Different\n Level of Snoring of whether or n
    theme(plot.title = element_text(face = "bold", hjust = 0.5))+
    guides(col=guide_legend(title="Quit Smoking"))+
  scale_color_brewer(palette="Dark2")
```



Both the graph and table demonstrates an interesting pattern comparing with the previous graph. smokers that never tried to quit smoking are most likely to be smokers that start smoking at a young age, despite the level of snoring! This can be reasonable since people who never tried to quit smoking can be people who have deep addiction to nicotine when they began to smoke at young age.

Relationship between average sleeping hour on weekdays by different cigarette brands and whether or not they report to have sleep problem

Data Description

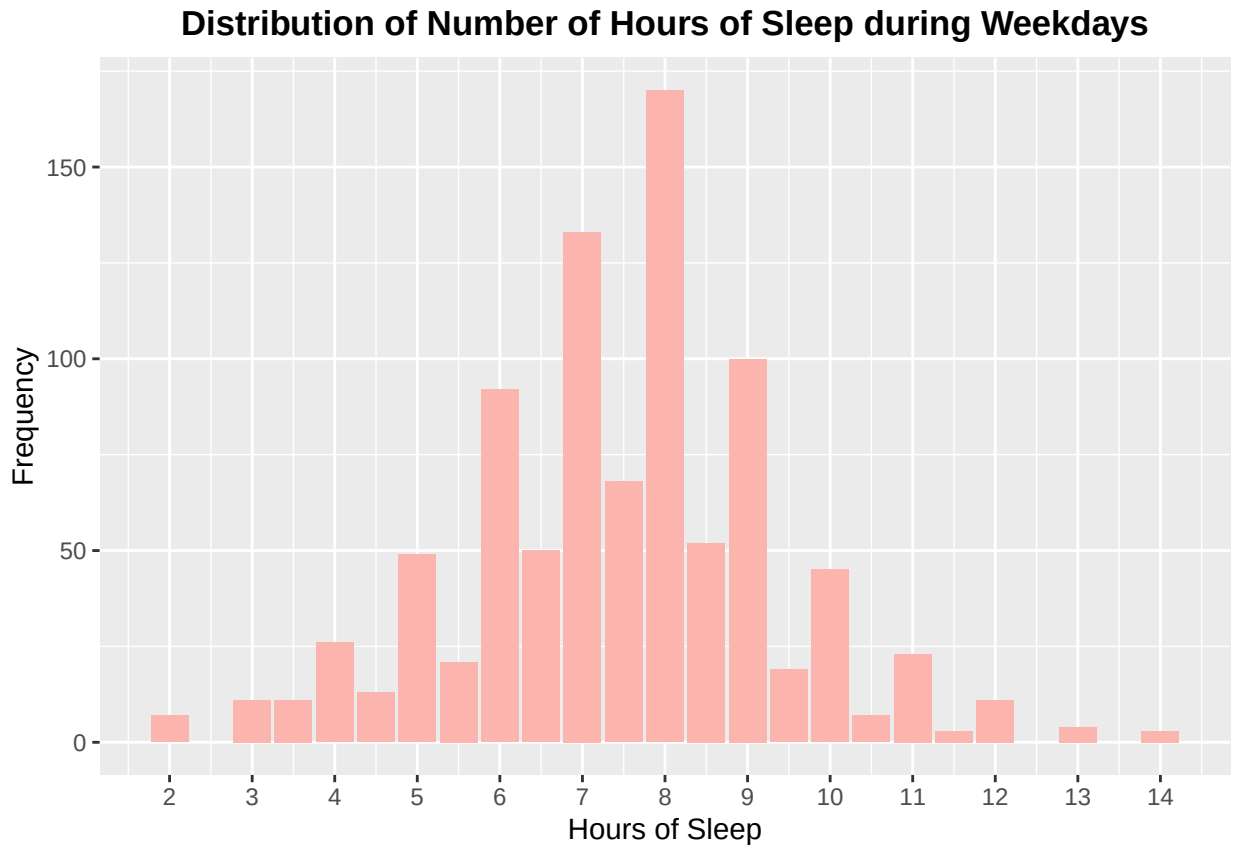
use the template I have for the first question for this part

talk about how important sleep on weekdays can relate to job performance and how cigarattes become addiction for employees in stressful working environment (something like that)

```
# create a new dataset
avghr_brands_report<-smoking_sleeping %>%
  select(SLD012 , SMD100BR, SLQ050)%>%
  filter(SLQ050 != 7,SLQ050 != 9 )
# Remove na
avghr_brands_report <- na.omit(avghr_brands_report)
```

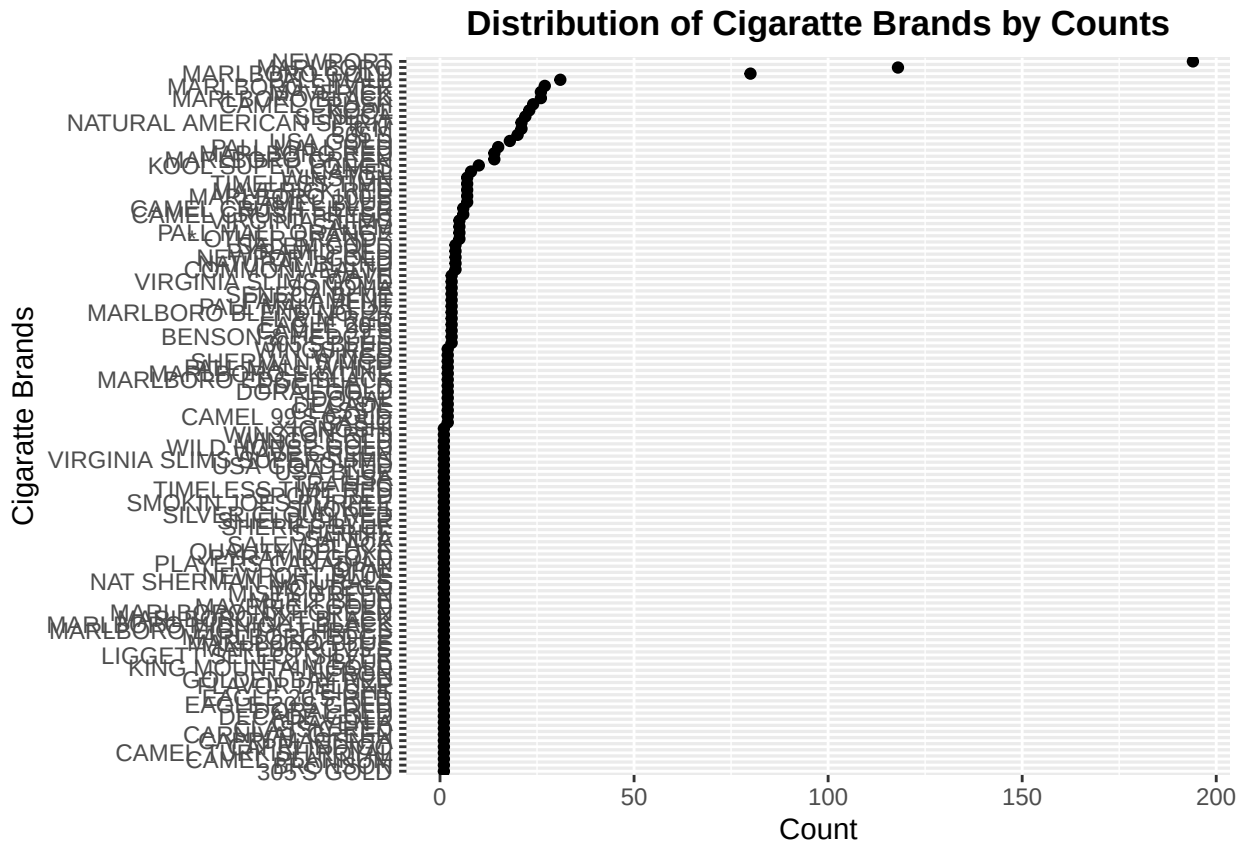
check the distribution of each variable

```
# sleep hour during weekdays
ggplot(data = avghr_brands_report, aes(x = SLD012, fill = "red")) +
  geom_bar()+
  labs(x = "Hours of Sleep", y = "Frequency", title = "Distribution of Number of Hours of Sleep during W
  theme(plot.title = element_text(face = "bold", hjust = 0.5),legend.position = "none")+
  scale_fill_brewer( palette="Pastel1")+
  scale_x_continuous(breaks = round(seq(min(avghr_brands_report$SLD012), max(avghr_brands_report$SLD012)
```



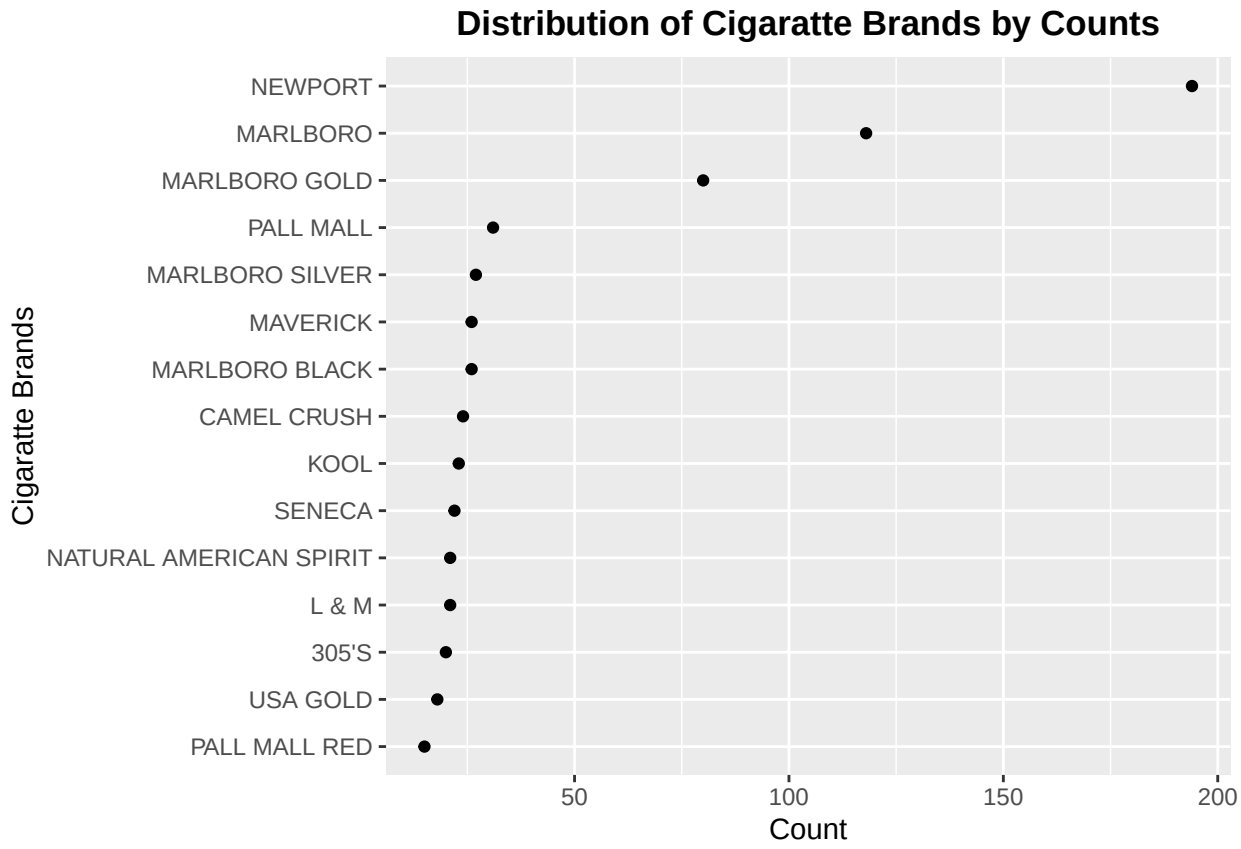
explain that this is about normally distributed

```
#distribution of cigarette brands

# first create a table of brands and their counts
avghr_brands_report%>%
  select(SMD100BR)%>%
  group_by(SMD100BR)%>%
  summarise(num = n())%>%
  #graph the count distribution
  ggplot(aes(x = num, y = fct_reorder(SMD100BR, num)%>%
               fct_relevel("* OTHER BRAND * ", after=0))) +
  geom_point()+
  labs(x = "Count", y = "Cigaratte Brands", title = "Distribution of Cigaratte Brands by Counts")+
  theme(plot.title = element_text(face = "bold", hjust = 0.5))
```



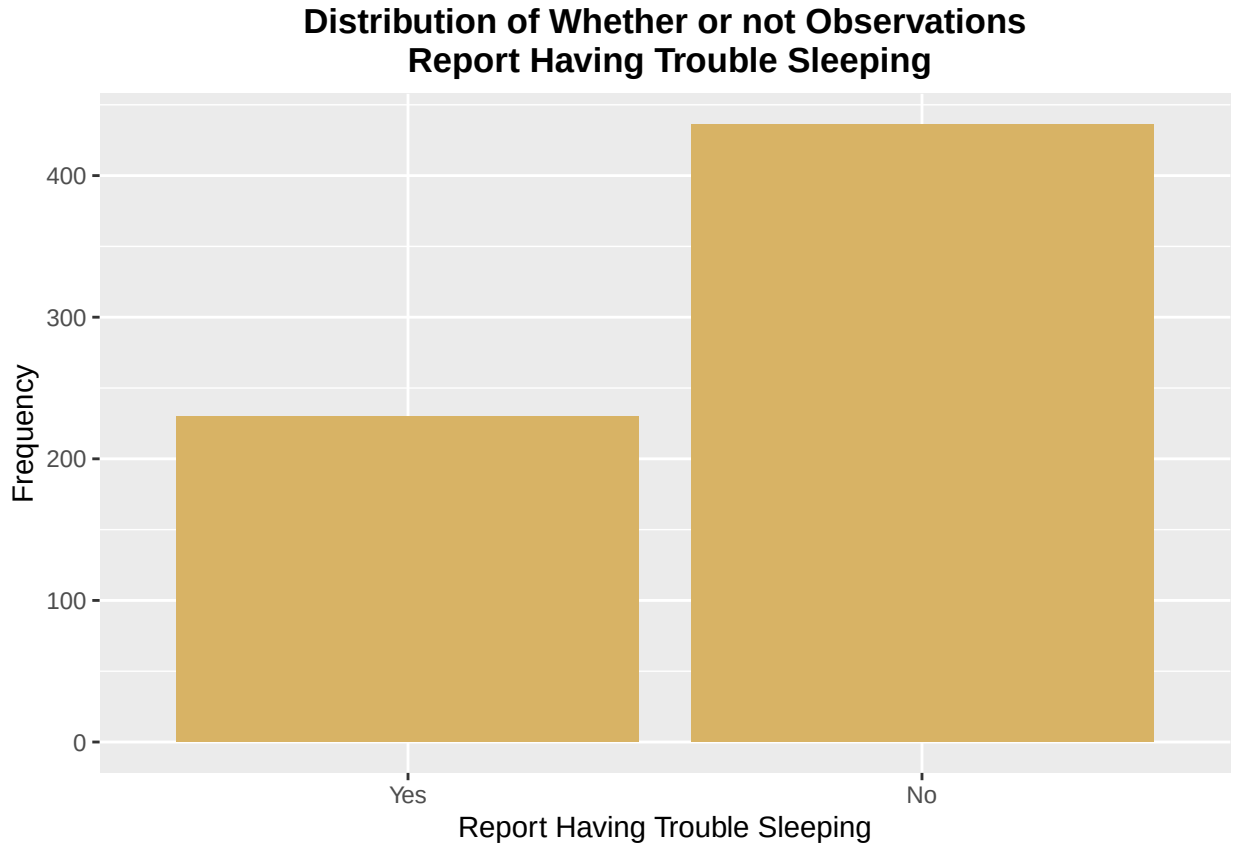
this is too messy and many only have 1 or 2 smokers, so filter out only brands that have over 15 users for the dataset

```
# filter out brands that have more than 15 users in the study
avghr_brands_report_15 <- avghr_brands_report%>%
  group_by(SMD100BR)%>%
  filter(n() >= 15)

# change the numeric value of sleeping report variable to yes or no
avghr_brands_report_15$SLQ050 <- factor(avghr_brands_report_15$SLQ050, labels = c("Yes", "No"))

# check the distribution of brands again
avghr_brands_report_15%>%
  select(SMD100BR)%>%
  group_by(SMD100BR)%>%
  summarise(num = n())%>%
  #graph the count distribution
  ggplot(aes(x = num, y = fct_reorder(SMD100BR, num)%>%
               fct_relevel("* OTHER BRAND * ", after=0))) +
  geom_point()+
  labs(x = "Count", y = "Cigaratte Brands", title = "Distribution of Cigaratte Brands by Counts")+
  theme(plot.title = element_text(face = "bold", hjust = 0.5))
```



say few things about the trend here

```
# check the distribution of report unable to sleep
ggplot(data = avghr_brands_report_15, aes(x = SLQ050, fill = "yellow")) + geom_bar()+
  labs(x = "Report Having Trouble Sleeping", y = "Frequency", title = "Distribution of Whether or not O
  theme(plot.title = element_text(face = "bold", hjust = 0.5),legend.position = "none")+
  scale_fill_brewer( palette="BrBG")
```



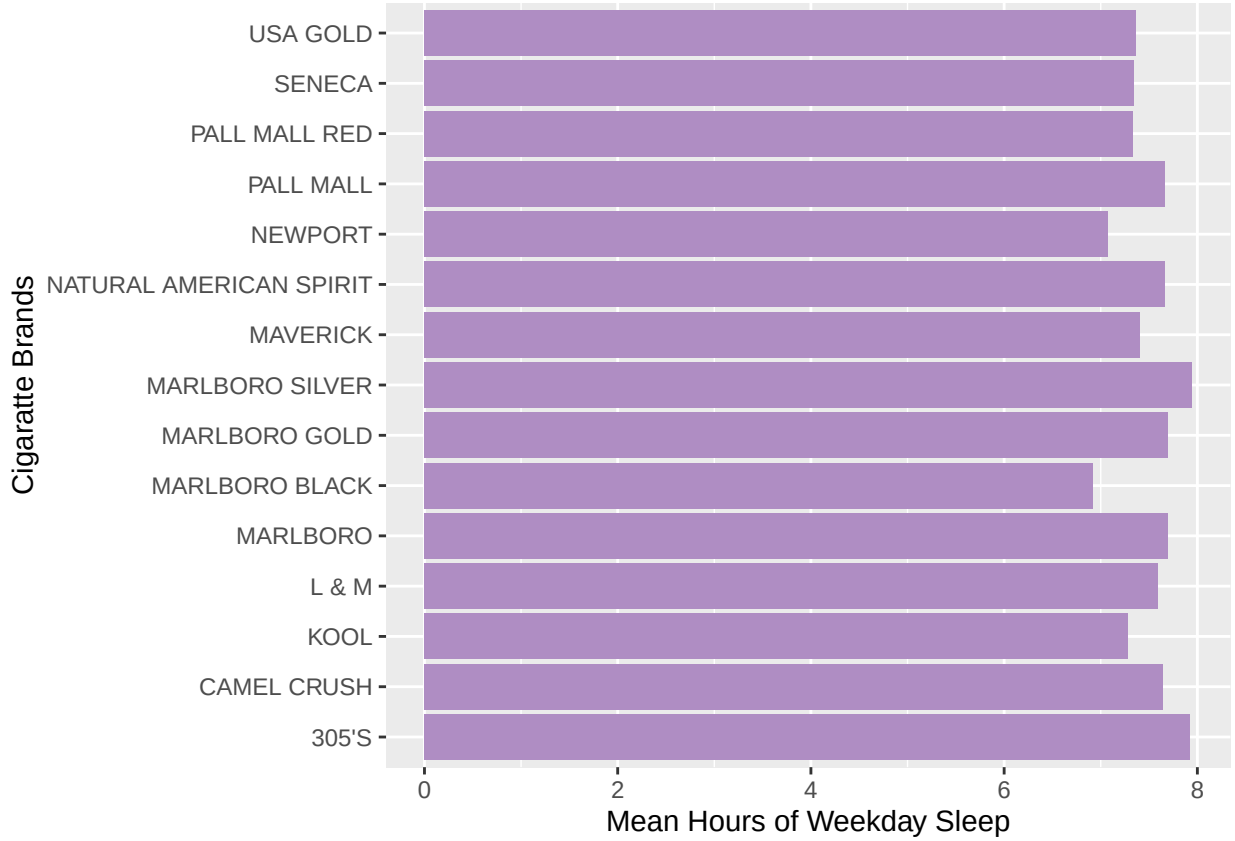
say something about the trend here (more people not reporting sleeping problem)

Now produce a graph brands by average number of sleep to see the relationship

```
# a summary table of cigarettes and sleep
avghr_brands_report_15%>%
  select(SLD012 , SMD100BR)%>%
  group_by(SMD100BR)%>%
  summarise(Mean = mean(SLD012))%>%
  kable(digits = 2, booktabs = T,col.names = c("Cigaratte Brands","Mean Hours of Weekday Sleep"))%>%
   kable_material(c("striped", "hover"))
```

```
# a graph
avghr_brands_report_15%>%
  select(SLD012 , SMD100BR)%>%
  group_by(SMD100BR)%>%
  summarise("Mean" = mean(SLD012))%>%
  ggplot(aes(x = Mean, y = SMD100BR, fill = "purple")) +
  geom_bar(stat="identity") +
  labs(x = "Mean Hours of Weekday Sleep", y = "Cigaratte Brands", main = "Distribution of Mean Hours of
  theme(plot.title = element_text(face = "bold", hjust = 0.5),legend.position = "none")+
  scale_fill_brewer( palette="PRGn")
```

Cigaratte Brands	Mean Hours of Weekday Sleep
305'S	7.92
CAMEL CRUSH	7.65
KOOL	7.28
L & M	7.60
MARLBORO	7.69
MARLBORO BLACK	6.92
MARLBORO GOLD	7.69
MARLBORO SILVER	7.94
MAVERICK	7.40
NATURAL AMERICAN SPIRIT	7.67
NEWPORT	7.07
PALL MALL	7.66
PALL MALL RED	7.33
SENECA	7.34
USA GOLD	7.36



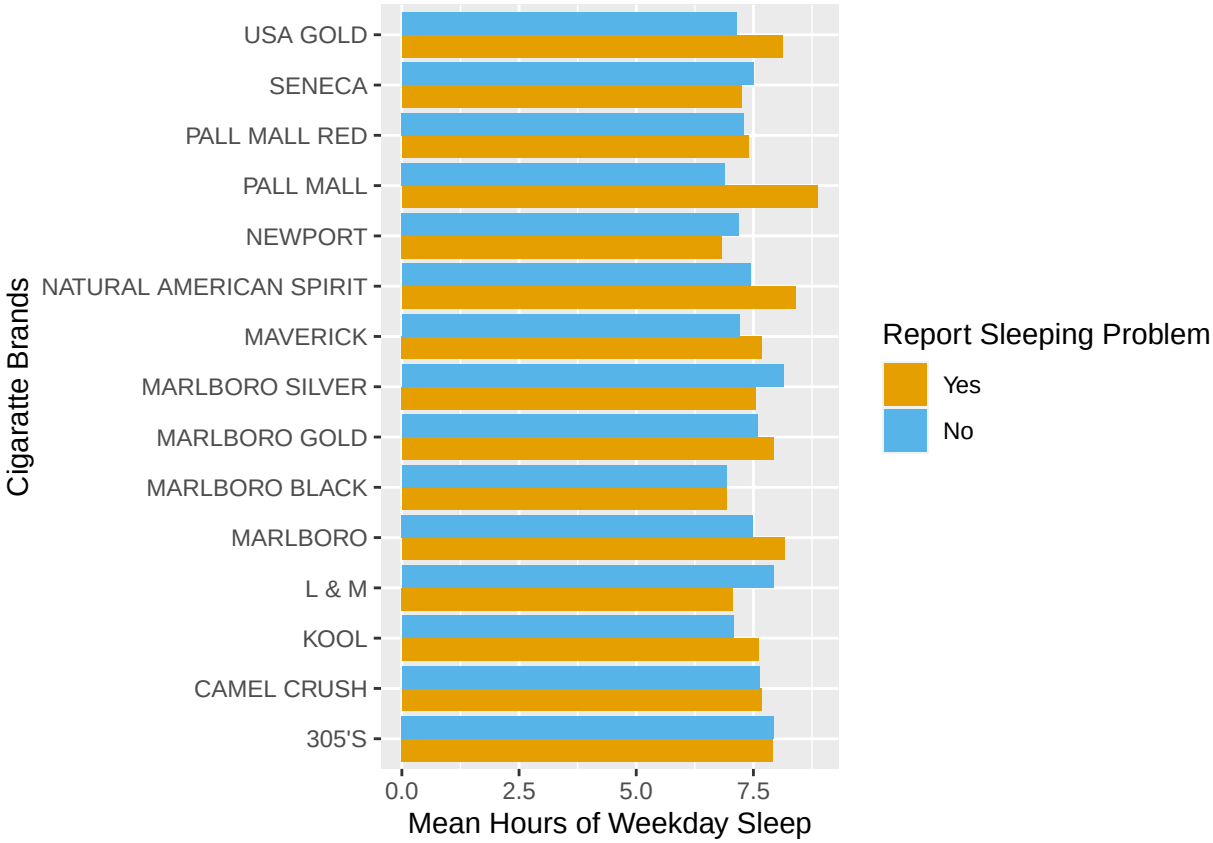
say a few sentences here

Next incorporate another variable, **whether or not the observation report having sleeping problem** to the model

Cigaratte Brands	Mean Hours Sleep(Report)	Mean Hours Sleep(not Report)
305'S	7.92	7.94
CAMEL CRUSH	7.67	7.63
KOOL	7.61	7.07
L & M	7.06	7.92
MARLBORO	8.16	7.49
MARLBORO BLACK	6.93	6.92
MARLBORO GOLD	7.92	7.58
MARLBORO SILVER	7.56	8.14
MAVERICK	7.68	7.20
NATURAL AMERICAN SPIRIT	8.40	7.44
NEWPORT	6.83	7.19
PALL MALL	8.88	6.89
PALL MALL RED	7.40	7.30
SENECA	7.25	7.50
USA GOLD	8.12	7.14

```
# a summary table of cigarettes and sleep by whether the observation report having sleeping problem or
avghr_brands_report_15%>%
  select(SLD012 , SMD100BR,SLQ050)%>%
  group_by(SLQ050,SMD100BR)%>%
  summarise(Mean = mean(SLD012))%>%
  # use pivot_wider to tidy the table
  pivot_wider(names_from = "SLQ050", values_from = "Mean")%>%
  kable(digits = 2, booktabs = T,col.names = c("Cigaratte Brands","Mean Hours Sleep(Report)","Mean Hours
  kable_material(c("striped", "hover"))
```

```
# a graph
avghr_brands_report_15%>%
  select(SLD012 , SMD100BR,SLQ050)%>%
  group_by(SMD100BR,SLQ050)%>%
  summarise(Mean = mean(SLD012))%>%
  ggplot(aes(x = Mean, y = SMD100BR, fill = SLQ050)) +
  geom_bar(stat="identity",position="dodge") +
  labs(x = "Mean Hours of Weekday Sleep", y = "Cigaratte Brands", main = "Distribution of Mean Hours of
  theme(plot.title = element_text(face = "bold", hjust = 0.5))+
  guides(fill=guide_legend(title="Report Sleeping Problem"))+
  scale_fill_manual(values=c( "#E69F00", "#56B4E9"))
```


USA GOLD
SENECA
PALL MALL RED
PALL MALL
NEWPORT
NATURAL AMERICAN SPIRIT
MAVERICK
MARLBORO SILVER
MARLBORO GOLD
MARLBORO BLACK
MARLBORO
L & M
KOOL
CAMEL CRUSH
305'S
Cigaratte Brands
0.0
2.5
5.0
7.5
Mean Hours of Weekday Sleep
Report Sleeping Problem
Yes
No