	1 Fail	2 insufficient	3 Sufficient	4 Good	5 Excellent	Weight	Grade and comments
Explain the data set	< 2	Majority of the features described incorrectly	Approximately more than half features are described correctly	Majority of the features are described correctly	All features are described correctly | or (4) but specific unique insight		4
Use of tidyverse		Mostly base R	More tidyverse than base R	Good use of tidyverse	Unique use of tidyverse		3
Cleanliness and readibility of code	< 2	Understandable with a lot of effort / No use of code conventions	Understandable with moderate effort / many breaches of code conventions	Understandable with little effort / some breaches of code conventions	Perfectly understandable and conventional code		3
Summary statistics	< 2	Highly incomplete summary statistics	Incomplete summary statistics with incomplete discussion	Complete summary statistics with limited discussion	Complete summary statistics with complete discussion		4
Descriptive plots	< 2	Majority of the features described incorrectly or not at all	Approximately more than half features are described correctly	Majority of the features are described correctly	All features are described correctly | or (4) but specific unique insight		4
Explanation of exploratory findings and the analytical process	< 2	Answers usually lack theoretical interpretations and are not phrased using full sentences	Answers are in own words, but not always using full sentences and occasionally lack theoretical interpretations of the results.	Answers are in own words, and in general include theoretical interpretations of the results.	Answers are in own words, using full sentences, and include theoretical interpretations beyond the results.		4
Code and output reproducibility		Unable to reproduce the same output file from the markup archive	Able to reproduce with minor adjustments to the markup archive		Able to reproduce the with no adjustments to the markup archive		3
Layout	< 2	Unclear layout, various elements missing (i.e., headings, output, original questions)	Clear headings in the file and output largely complete	Clear headings in the file and output is complete	Clear layout, front-page included, clear headings in the file and only relevant output		2
Additional remarks, strengths and/or weaknesses	At times I have trouble following your train of thought. You present a lot of code, and only after displaying (parts of) the output, you discuss the premise of the analysis. Try to make your document much more structured. And don't forget to print the output to the markup document, otherwise any related investigations and mutations are redundant.					+ or -	

$\frac{27}{8} = 3\frac{3}{8} = 6.75$

Assignment 1

Kevin Patyk

9/27/2021

Loading the packages that will be used

```r
library(tidyverse)
library(magrittr)
library(DT)
```

Data import and preliminary examination

```r
sleep_df <- read_csv("Dataset/Subjective_Sleep_Quality.csv") #GV does not run
```

Checking the number of observations and the number of variables.

```r
nrow(sleep_df)
```

```
## [1] 546
```

```r
ncol(sleep_df)
```

```
## [1] 21
```

dim() or str()

Examining the first 6 rows of the data. Why?

```r
sleep_df %>%
  head()
```

Show 10 entries Search:

	University	Case_Number	Age	Sex	Year_of_Study	Department_Name	MEQ	Trait_Anxiety	Sta
1	1	1	19	2	2	Psychology	15	30	
2	1	2	18	2	1	Psychology	13	50	
3	1	3	19	2	2	Psychology	21	32	
4	1	4	19	2	1	Psychology	9	53	
5	1	5	19	2	2	Psychology	17	39	
6	1	6	18	2	1	Psychology	5	60	

Showing 1 to 6 of 6 entries Previous 1 Next

Examining the last 6 rows of the data. Why?

```r
sleep_df %>%
  tail()
```

Show 10 entries Search:

	University	Case_Number	Age	Sex	Year_of_Study	Department_Name	MEQ	Trait_Anxiety	Sta
1	1	541	21	2	2	Psychology	14	36	
2	1	542	21	1	2	Psychology	9	49	
3	1	543	18	2	1	Psychology	8	37	
4	1	544	22	2	2	Psychology	13	46	
5	1	545	19	2	2	Psychology	14	37	
6	1	546	19	2	1	Psychology	14	54	

Showing 1 to 6 of 6 entries Previous 1 Next

Examining the structure of the data and data types. not printed? why hide the results?

```r
str(sleep_df)
```

The variables `Department_Name`, `Year_of_Study`, `Sex`, and the other dichotomous variables need to be converted to a factors.

```r
sleep_df %<>%
  mutate(Department_Name = as.factor(Department_Name),
         Year_of_Study = as.factor(Year_of_Study),
         Sex = as.factor(Sex),
         coffee_dichotomous = as.factor(coffee_dichotomous),
         Cigarettes_dichotomous = as.factor(Cigarettes_dichotomous),
         Alcohol_dichotomous = as.factor(Alcohol_dichotomous),
         Caffeine_dichotomous = as.factor(Caffeine_dichotomous),
         coffee_dichotomous = as.factor(coffee_dichotomous))
```

Why not label to make your life easier?

Obtaining descriptive statistics and frequencies for categorical/dichotomous variables

Obtaining the mean and standard deviation for age.

```r
sleep_df %$%
  mean(Age)
```

```
## [1] 20.40842
```

```r
sleep_df %$%
  sd(Age)
```

```
## [1] 4.306155
```

Getting a table of how many students are there per each department.

```r
sleep_df %$%
  table(Department_Name)
```

Show 10 entries Search:

	Department_Name	Freq
1	Psychology	486
2	Other	20
3	Life Sciences	17
4	Social Sciences	15
5	Business School	3
6	English and Cre	2
7	Dance	1
8	Drama	1
9	Media, Culture	1

Showing 1 to 9 of 9 entries Previous 1 Next

Getting a table of sex.

```r
#1 = male, 2 = female
sleep_df %$%
  table(Sex)
```

```
## Sex
##   1   2
##  85 461
```

Getting a table of how many students are in each year of study.

```r
sleep_df %$%
  table(Year_of_Study)
```

```
## Year_of_Study
##   1   2   3
## 291 225  30
```

Getting a table of tobacco consumption.

```r
#1 = smoker, 2 = non-smoker
sleep_df %$%
  table(Cigarettes_dichotomous)
```

```
## Cigarettes_dichotomous
##   1   2
## 466  80
```

Why not use summary() or psych::describe()?

Getting a table of caffeine consumption.

```r
#1 = consumes caffeine, 2 = does not consume caffeine
sleep_df %$%
  table(Caffeine_dichotomous)
```

```
## Caffeine_dichotomous
##   1   2
## 327 219
```

Getting a table of alcohol consumption.

```r
#1 = consumes alcohol, 2 = does not consume alcohol
sleep_df %$%
  table(Alcohol_dichotomous)
```

```
## Alcohol_dichotomous
##   1   2
## 260 286
```

Getting a table of coffee consumption.

```r
#1 = consumes coffe, 2 = does not consume coffee
sleep_df %$%
  table(coffee_dichotomous)
```

```
## coffee_dichotomous
##   1   2
##  65 481
```

Describing the basics of the dataset prior to more in-depth exploratory data analysis

Perhaps this is a good section to start with.

This dataset was collected by Norbury and Evans (2018) for a study which was examining sleep quality in university students. A total of 546 university students with ages ranging from 18 to 55 were included. The students came from two universities which were located in South-East England. Participants completed an online self-report survey that asked questions about age, gender, year of study, and field of study. Furthermore, data was collected on sleep quality, anxiety, preferred start time for classes, and chronotype. Lastly, information on coffee, caffeine, alcohol, and tobacco use was collected; these variables were coded dichotomously (1 = yes, 2 = no). There was no missing data.

The mean age of participants in the study was 20.41 (SD = 4.31). A majority of the students (89%) came from the Psychology department, followed by Other, Life Sciences, and Social Sciences. Only a total of 5 students came from Business School, English and Creative Writing, Dance, Drama, and Media and Culture. There were 291 first years, 225 second years, and 30 third years. A majority of the students smoked (85%) and consumed caffeine (60%). Almost half of the students consumed alcohol (48%) and only a small minority drank coffee (12%). Lastly, a majority of the participants were female (84%).

Exploratory data analysis part 1: Exploring general sleeping trends

Getting descriptive statistics for sleep duration.

```r
#need to make a function to obtain means and standard deviations quickly
DescriptiveStatistics <- function(x) {
  Mean <- round(mean(x), digits = 2)
  Standard_Deviation <- round(sd(x), digits = 2)
  return(c(Mean = Mean, Standard_Deviation = Standard_Deviation))
}
```

better to use dplyr::summarise_all()

```r
#running the function over our sleep duration variables of interest
desc_stats_sleep <- sapply(X = sleep_df[, c(12, 13, 14)], FUN = DescriptiveStatistics)
```

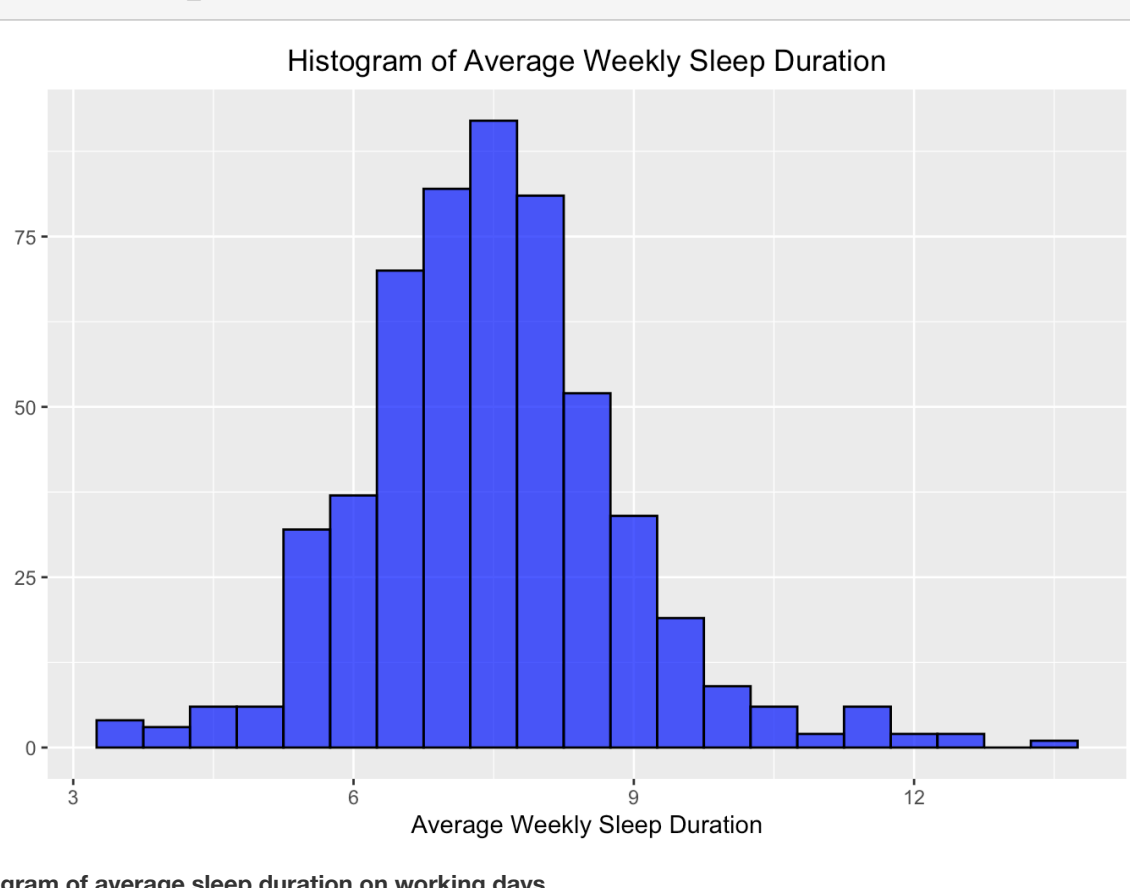
Show 10 entries Search:

	Mean	Standard_Deviation
Avg_Weekly_Sleep_Duration	7.49	1.39
Avg_Sleep_Working_days	6.55	1.5
Avg_sleep_free_days	8.8	1.61

Showing 1 to 3 of 3 entries Previous 1 Next

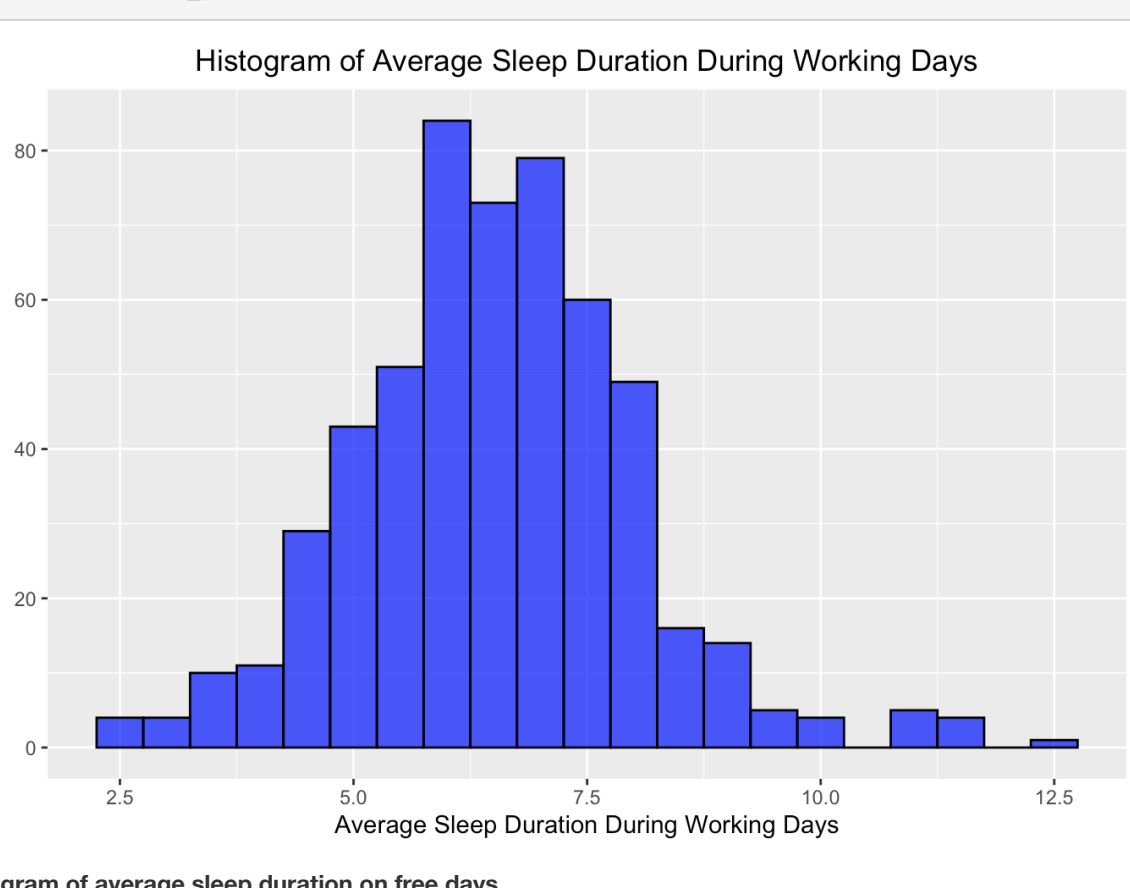
Examining a histogram of average weekly sleep duration.

```r
sleep_df %>%
  ggplot(aes(x = Avg_Weekly_Sleep_Duration)) +
  geom_histogram(binwidth = .5, color = "black", fill = "blue", alpha = .75) +
  xlab(label = "Average Weekly Sleep Duration") +
  ggtitle(label = "Histogram of Average Weekly Sleep Duration") +
  theme(axis.title.y = element_blank()) +
  theme(plot.title = element_text(hjust = 0.5))
```



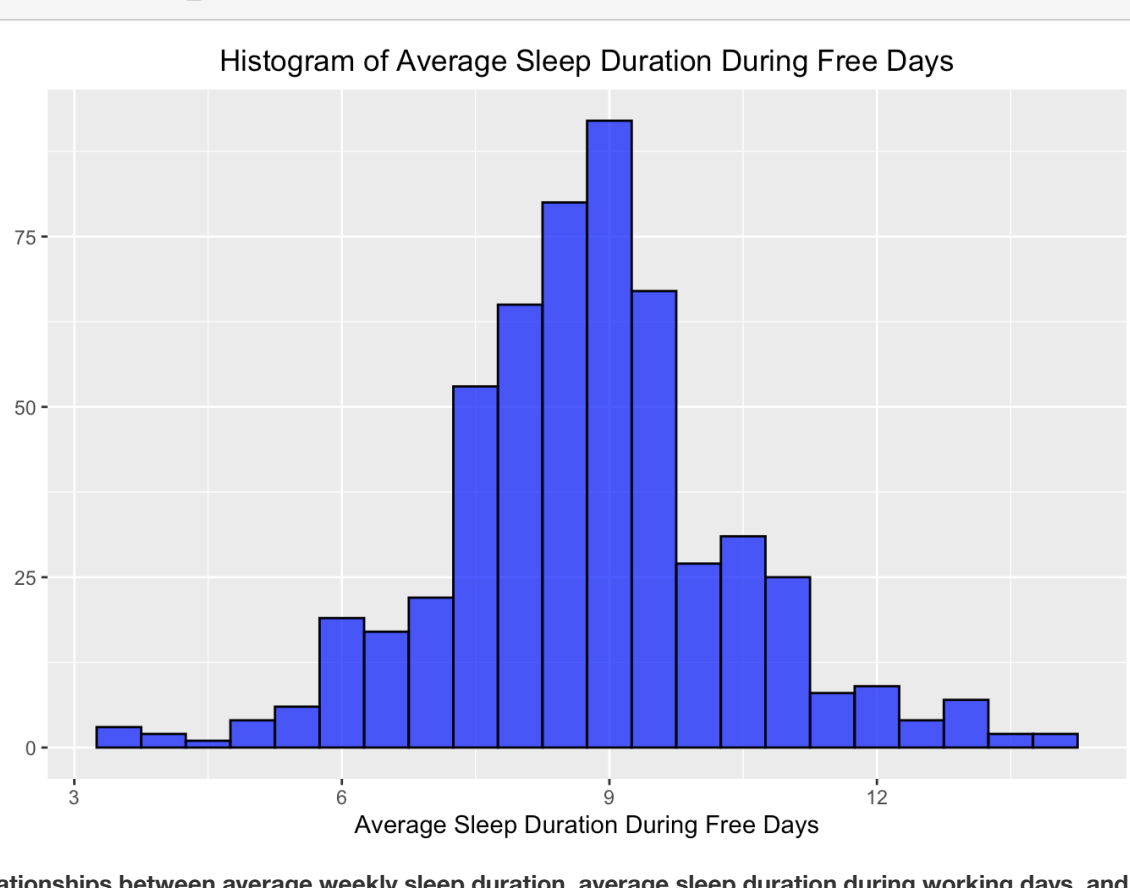
Examining a histogram of average sleep duration on working days.

```r
sleep_df %>%
  ggplot(aes(x = Avg_Sleep_Working_days)) +
  geom_histogram(binwidth = .5, color = "black", fill = "blue", alpha = .75) +
  xlab(label = "Average Sleep Duration During Working Days") +
  ggtitle(label = "Histogram of Average Sleep Duration During Working Days") +
  theme(axis.title.y = element_blank()) +
  theme(plot.title = element_text(hjust = 0.5))
```



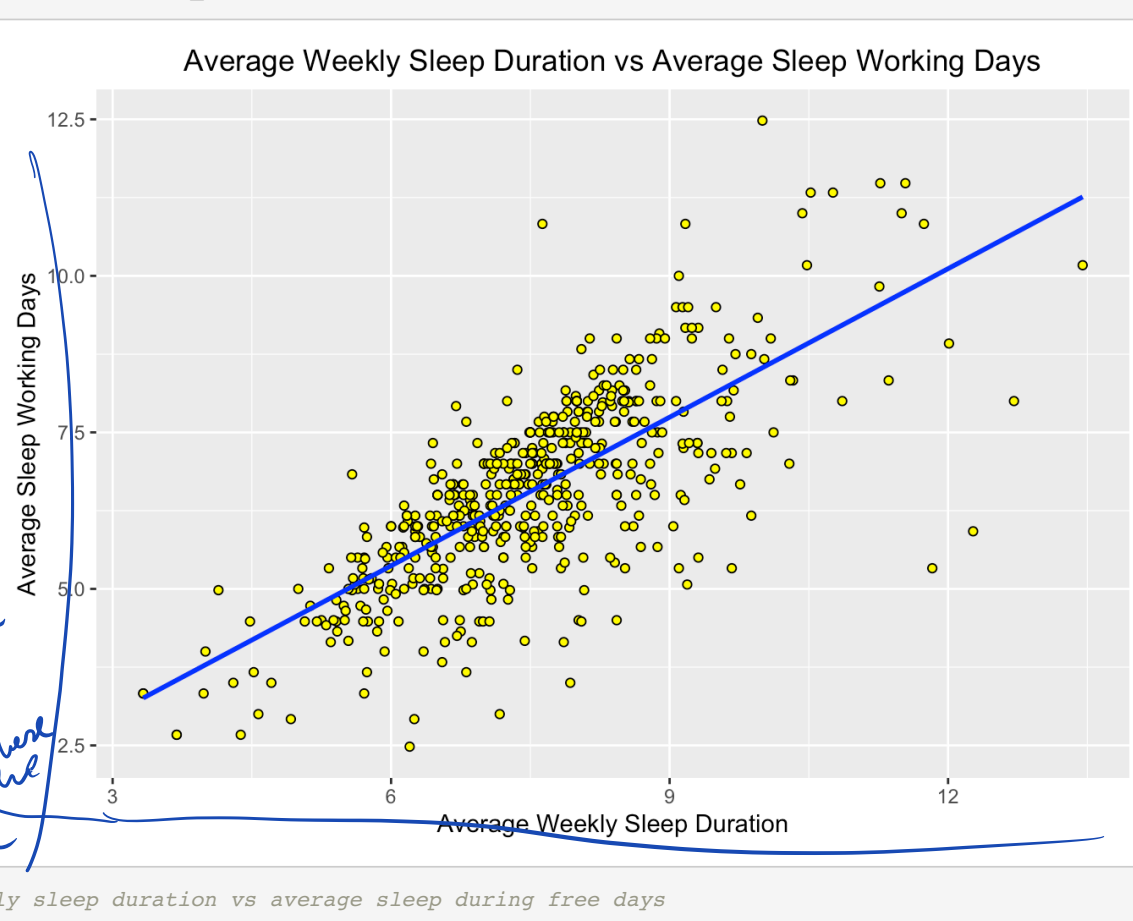
Examining a histogram of average sleep duration on free days.

```r
sleep_df %>%
  ggplot(aes(x = Avg_sleep_free_days)) +
  geom_histogram(binwidth = .5, color = "black", fill = "blue", alpha = .75) +
  xlab(label = "Average Sleep Duration During Free Days") +
  ggtitle(label = "Histogram of Average Sleep Duration During Free Days") +
  theme(axis.title.y = element_blank()) +
  theme(plot.title = element_text(hjust = 0.5))
```



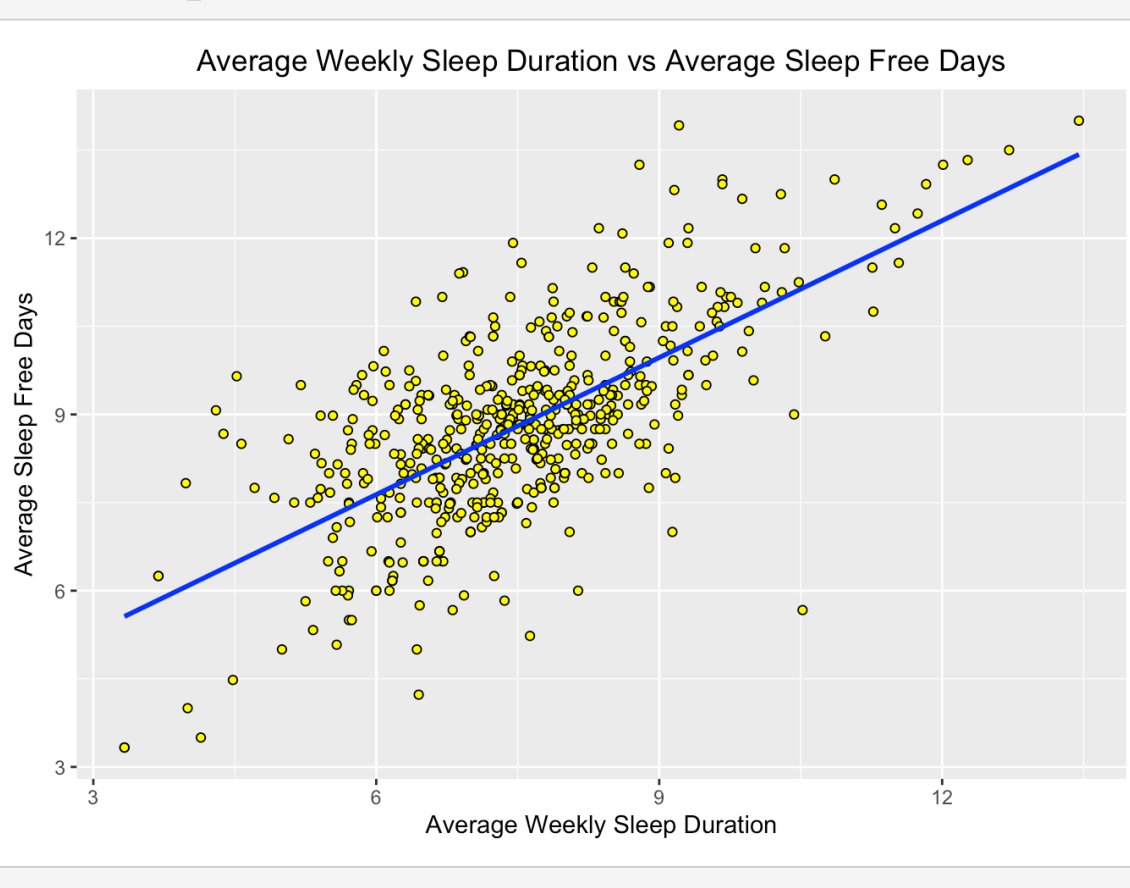
Examining the relationships between average weekly sleep duration, average sleep duration during working days, and average sleep duration during free days.

```r
#average weekly sleep duration vs average sleep working days
sleep_df %>%
  ggplot(aes(x = Avg_Weekly_Sleep_Duration, y = Avg_Sleep_Working_days)) +
  geom_point(color = "black", shape = 21, fill = "yellow") +
  geom_smooth(method = "lm", fill = NA, color = "blue") +
  xlab("Average Weekly Sleep Duration") +
  ylab("Average Sleep Working Days") +
  ggtitle("Average Weekly Sleep Duration vs Average Sleep Working Days") +
  theme(plot.title = element_text(hjust = 0.5))
```



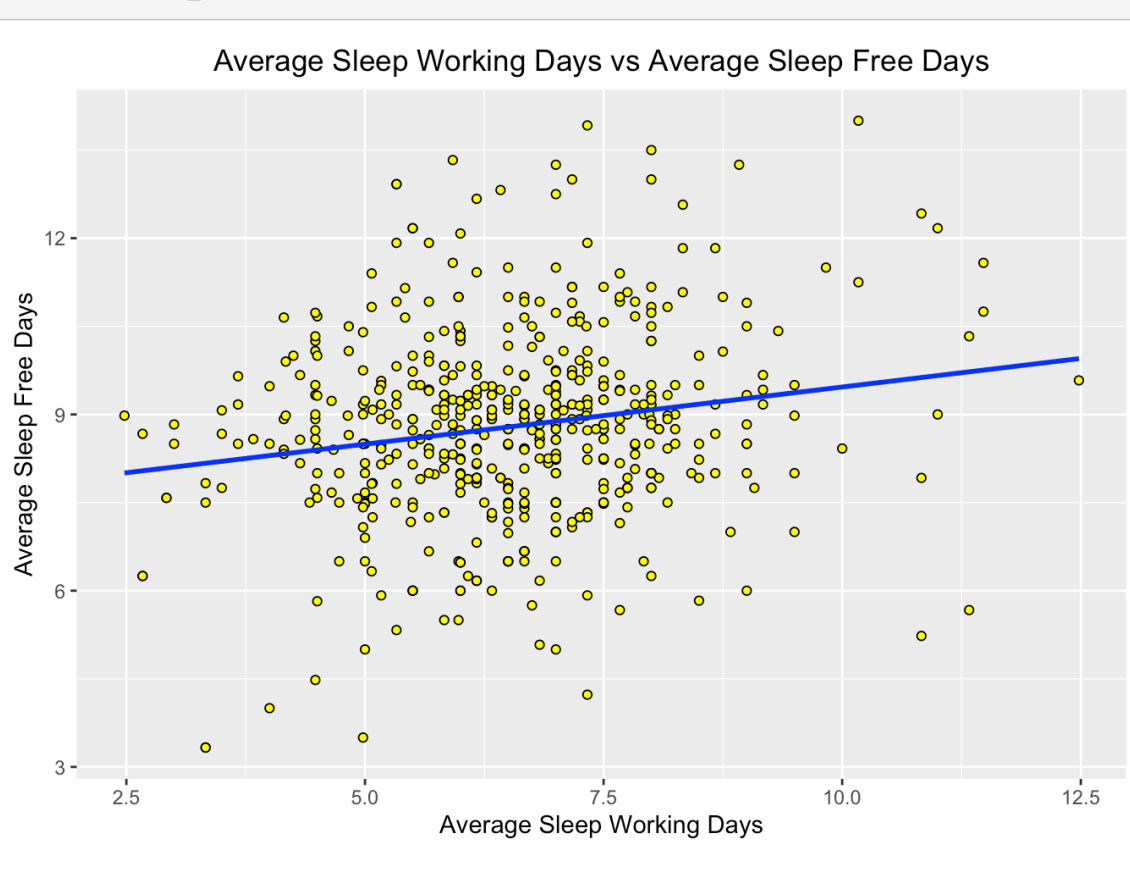
since you're highlighting the diagonal it would be nice to keep the axis the same

```r
#average weekly sleep duration vs average sleep during free days
sleep_df %>%
  ggplot(aes(x = Avg_Weekly_Sleep_Duration, y = Avg_sleep_free_days)) +
  geom_point(color = "black", shape = 21, fill = "yellow") +
  geom_smooth(method = "lm", fill = NA, color = "blue") +
  xlab("Average Weekly Sleep Duration") +
  ylab("Average Sleep Free Days") +
  ggtitle("Average Weekly Sleep Duration vs Average Sleep Free Days") +
  theme(plot.title = element_text(hjust = 0.5))
```



like this

```r
#average sleep duration working days vs average sleep during free days
sleep_df %>%
  ggplot(aes(x = Avg_Sleep_Working_days, y = Avg_sleep_free_days)) +
  geom_point(color = "black", shape = 21, fill = "yellow") +
  geom_smooth(method = "lm", fill = NA, color = "blue") +
  xlab("Average Sleep Working Days") +
  ylab("Average Sleep Free Days") +
  ggtitle("Average Sleep Working Days vs Average Sleep Free Days") +
  theme(plot.title = element_text(hjust = 0.5))
```



Conclusions from this section:

I wanted to get an idea of the average sleep duration that students within the study are getting on average per week, during working days, and during free days. More specifically, I was curious about the differences during working and free days, since, for students, the difference between the two, I thought, would be substantial. The mean average weekly sleep duration across all students is 7.49 (SD = 1.39). The mean average sleep duration during working days is 6.55 (SD = 1.5). The mean average sleep duration during free days is 8.8 (SD = 1.61). This makes sense, as students throughout the week need to be up for classes and can spend less time sleeping than on off days. I also made histograms for these variables of interest because I wanted to check the distributions and if there were any outliers. Some outliers were found, with some students having an average weekly sleep duration of 12 of hours more. For average sleep during working days, some students slept, on average, 10 hours with one student sleeping 12. For average sleep duration during free days, I excpected the histogram to shift right, as people would be sleeping more then; there were still outliers, with some students sleeping, on average, over 12 hours on free days.

Lastly, I examined the relationships between the average weekly sleep, average sleep duration during work days, and average sleep duration during free days. The scatterplots showed that average weekly sleep had a positive relationship with average sleep during work and free days, which makes sense, since the more you sleep on both types of days, the higher your average sleep. However, I expected that the relationship between average sleep work days and average sleep free days would be negative, since if a student slept less on the work days, then they would sleep more during free days. However, this was not the case as the relationship was still positive.

Exploratory data analysis part 2: Exploring sleeping trends, substance use, and anxiety

Checking the weekly average sleep duration based on cigarette use.

```r
sleep_df %>%
  group_by(Cigarettes_dichotomous) %>%
  summarise(Avg_Weekly_Sleep_Duration = round(mean(Avg_Weekly_Sleep_Duration), digits = 2)) %>%
  arrange(-Avg_Weekly_Sleep_Duration)
```

```
## # A tibble: 2 × 2
##   Cigarettes_dichotomous Avg_Weekly_Sleep_Duration
##   <fct>                                      <dbl>
## 1 2                                           7.71
## 2 1                                           7.45
```

which represents?

Checking the weekly average sleep duration based on alcohol use.

```r
sleep_df %>%
  group_by(Alcohol_dichotomous) %>%
  summarise(Avg_Weekly_Sleep_Duration = round(mean(Avg_Weekly_Sleep_Duration), digits = 2)) %>%
  arrange(-Avg_Weekly_Sleep_Duration)
```

```
## # A tibble: 2 × 2
##   Alcohol_dichotomous Avg_Weekly_Sleep_Duration
##   <fct>                                   <dbl>
## 1 1                                        7.66
## 2 2                                        7.34
```

which represents?

Checking the weekly average sleep duration based on caffeine use.

```r
sleep_df %>%
  group_by(Caffeine_dichotomous) %>%
  summarise(Avg_Weekly_Sleep_Duration = round(mean(Avg_Weekly_Sleep_Duration), digits = 2)) %>%
  arrange(-Avg_Weekly_Sleep_Duration)
```

```
## # A tibble: 2 × 2
##   Caffeine_dichotomous Avg_Weekly_Sleep_Duration
##   <fct>                                    <dbl>
## 1 2                                         7.62
## 2 1                                         7.4
```

idem

Checking the weekly average sleep duration based on coffee use.

```r
sleep_df %>%
  group_by(coffee_dichotomous) %>%
  summarise(Avg_Weekly_Sleep_Duration = round(mean(Avg_Weekly_Sleep_Duration), digits = 2)) %>%
  arrange(-Avg_Weekly_Sleep_Duration)
```

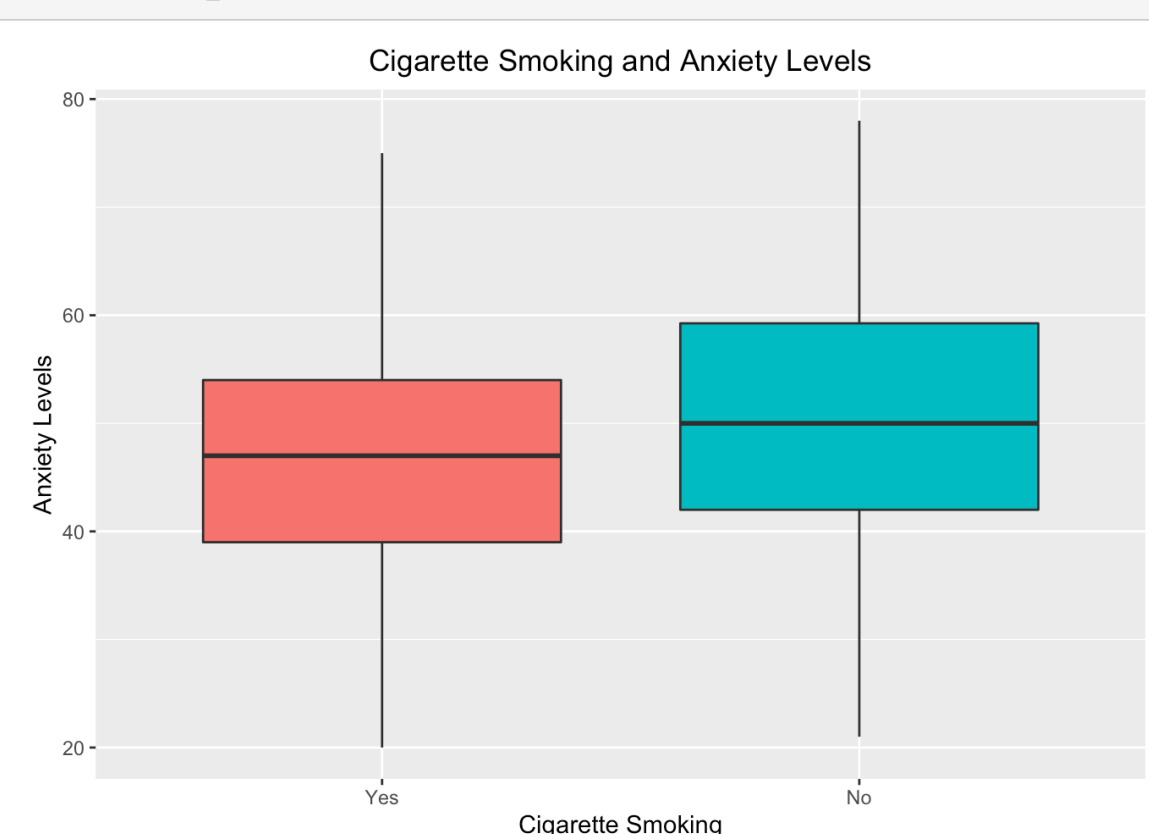
```
## # A tibble: 2 × 2
##   coffee_dichotomous Avg_Weekly_Sleep_Duration
##   <fct>                                  <dbl>
## 1 1                                       7.58
## 2 2                                       7.48
```

idem

Examining cigarette smoking and anxiety levels via boxplot.

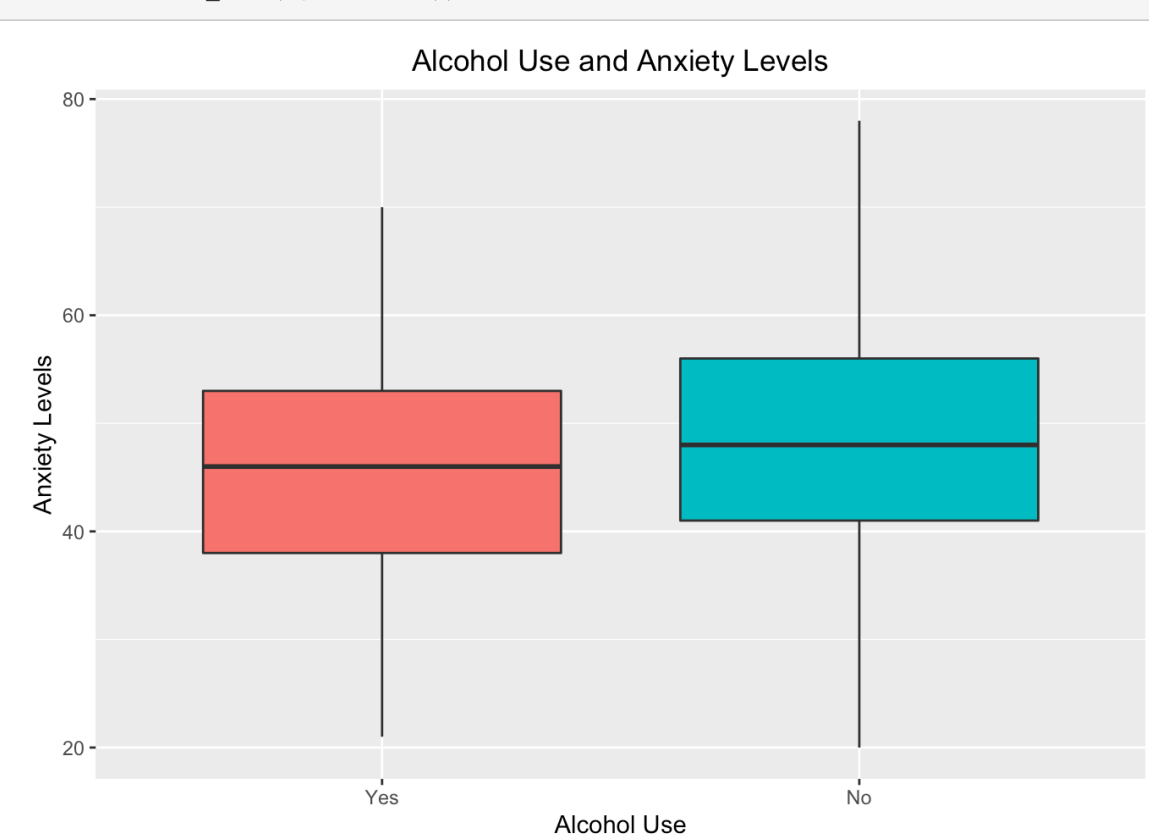
```r
sleep_df %>%
  ggplot(aes(x = Cigarettes_dichotomous, y = Trait_Anxiety)) +
  geom_boxplot(outlier.colour = NA, aes(fill = Cigarettes_dichotomous), show.legend = F) +
  ylab("Anxiety Levels") +
  xlab("Cigarette Smoking") +
  scale_x_discrete(labels = c("Yes", "No")) +
  ggtitle("Cigarette Smoking and Anxiety Levels") +
  theme(plot.title = element_text(hjust = 0.5))
```

because of the omitted labeling, you now need to manually specify this - every time



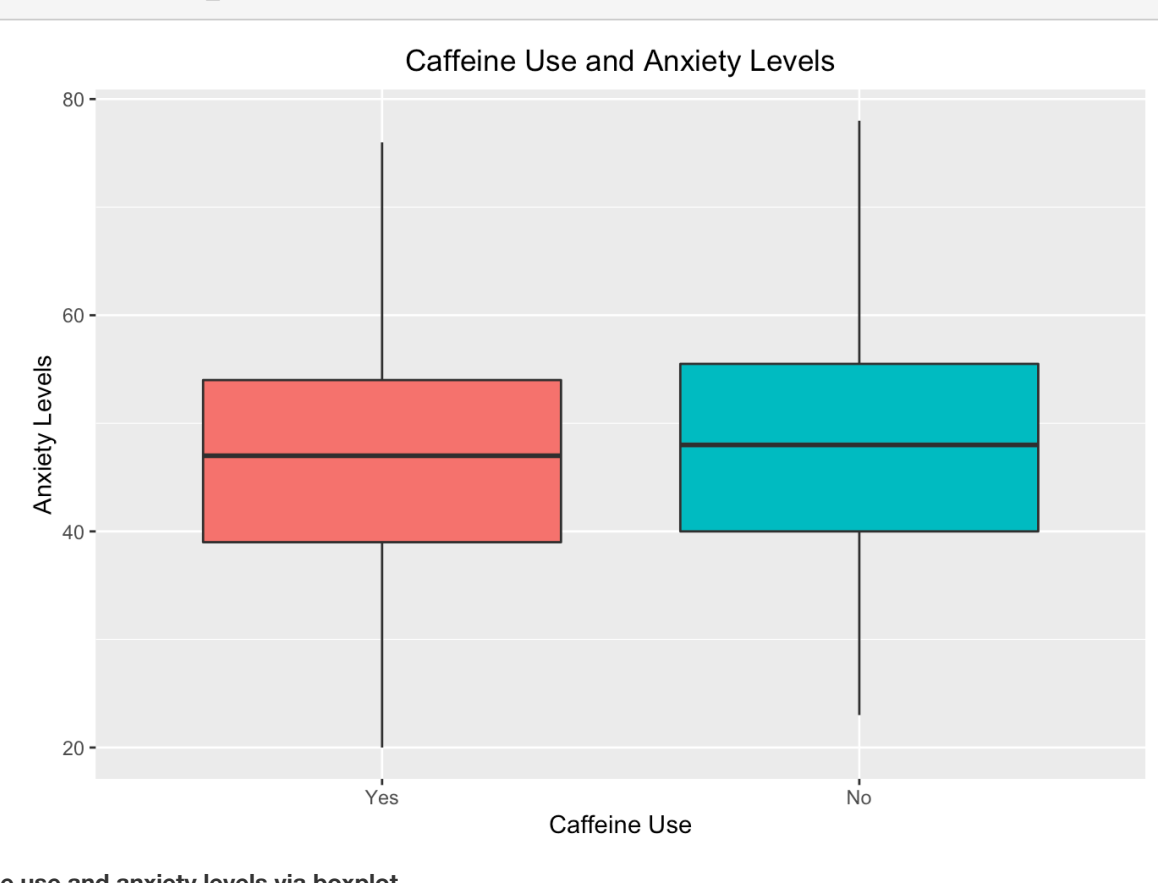
Examining alcohol use and anxiety levels via boxplot.

```r
sleep_df %>%
  ggplot(aes(x = Alcohol_dichotomous, y = Trait_Anxiety)) +
  geom_boxplot(outlier.colour = NA, aes(fill = Alcohol_dichotomous), show.legend = F) +
  ylab("Anxiety Levels") +
  xlab("Alcohol Use") +
  scale_x_discrete(labels = c("Yes", "No")) +
  ggtitle("Alcohol Use and Anxiety Levels") +
  theme(plot.title = element_text(hjust = 0.5))
```



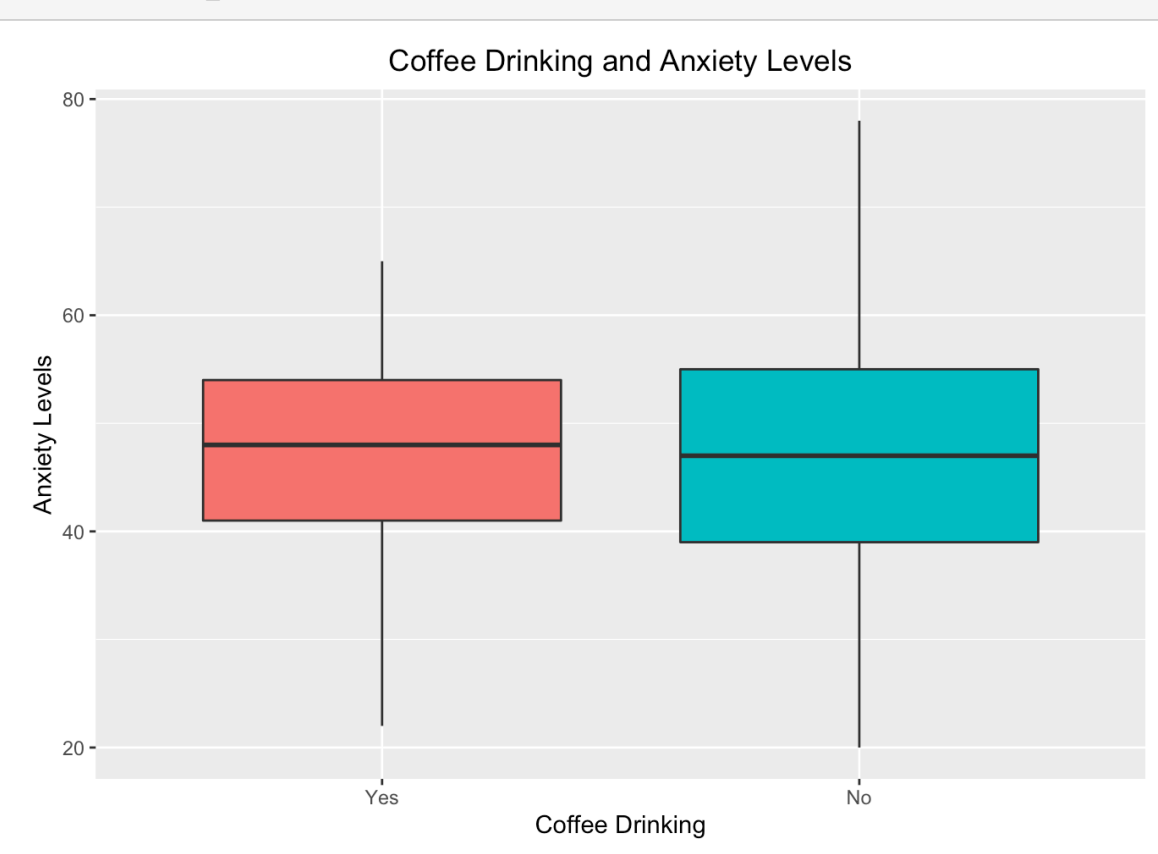
Examining caffeine use and anxiety levels via boxplot.

```r
sleep_df %>%
  ggplot(aes(x = Caffeine_dichotomous, y = Trait_Anxiety)) +
  geom_boxplot(outlier.colour = NA, aes(fill = Caffeine_dichotomous), show.legend = F) +
  ylab("Anxiety Levels") +
  xlab("Caffeine Use") +
  scale_x_discrete(labels = c("Yes", "No")) +
  ggtitle("Caffeine Use and Anxiety Levels") +
  theme(plot.title = element_text(hjust = 0.5))
```



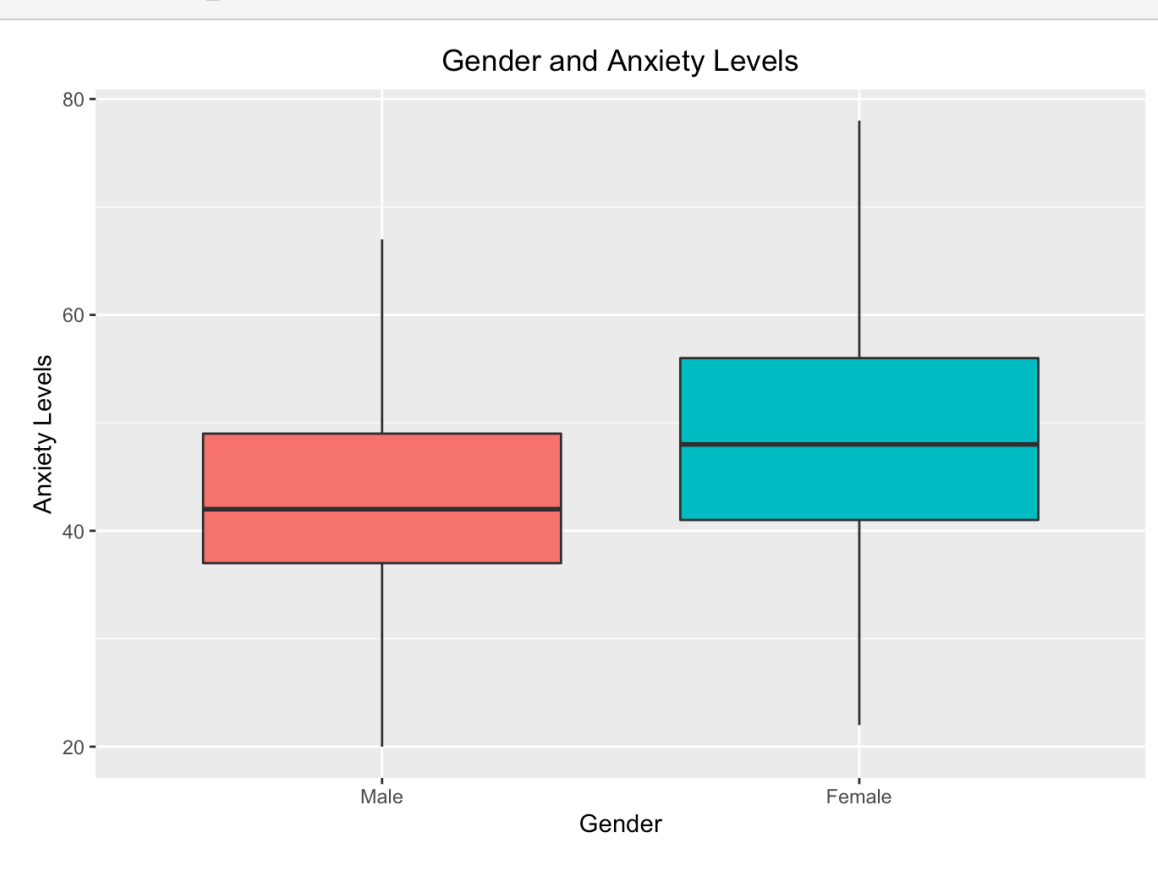
Examining coffee use and anxiety levels via boxplot.

```r
sleep_df %>%
  ggplot(aes(x = coffee_dichotomous, y = Trait_Anxiety)) +
  geom_boxplot(outlier.colour = NA, aes(fill = coffee_dichotomous), show.legend = F) +
  ylab("Anxiety Levels") +
  xlab("Coffee Drinking") +
  scale_x_discrete(labels = c("Yes", "No")) +
  ggtitle("Coffee Drinking and Anxiety Levels") +
  theme(plot.title = element_text(hjust = 0.5))
```



Examining gender and anxiety levels via boxplot.

```r
sleep_df %>%
  ggplot(aes(x = Sex, y = Trait_Anxiety)) +
  geom_boxplot(outlier.colour = NA, aes(fill = Sex), show.legend = F) +
  ylab("Anxiety Levels") +
  xlab("Gender") +
  scale_x_discrete(labels = c("Male", "Female")) +
  ggtitle("Gender and Anxiety Levels") +
  theme(plot.title = element_text(hjust = 0.5))
```



Conclusions from this section:

In this section, I wanted to get more of an idea about average weekly sleep duration, substance use, and anxiety. This is because substance use typically does have an affect on length and quality of sleep and because it can also affect anxiety levels. Lastly, I wanted to look at the difference in anxiety levels between genders because gender can have an affect on anxiety levels. The results show that those who do not smoke cigarettes have a higher average weekly sleep duration, alcohol users have a higher average weekly sleep duration, caffeine users have a lower average weekly sleep duration, and that coffee drinkers have a higher average sleep duration, but not by much. Regarding anxiety levels and substance use, smokers appear to have less anxiety than non-smokers, those who consume alcohol have less anxiety than those who do not, caffeine users and caffeine non-users have almost the same amount of anxiety, and coffee drinkers have a little bit more anxiety than those who do not consume coffee. Lastly, females tend to have far more anxiety than males.

Exploratory data analysis part 3: Exploring sleeping trends and anxiety within departments and across years

Checking the average weekly sleep duration per department.

```r
sleep_df %>%
  group_by(Department_Name) %>%
  summarise(Avg_Wk_Sleep_Dep = round(mean(Avg_Weekly_Sleep_Duration), digits = 2)) %>%
  arrange(-Avg_Wk_Sleep_Dep)
```

Show 10 entries Search:

	Department_Name	Avg_Wk_Sleep_Dep
1	Business School	7.99
2	Dance	7.67
3	Psychology	7.52
4	Life Sciences	7.51
5	English and Cre	7.44
6	Drama	7.4
7	Social Sciences	7.24
8	Other	6.97
9	Media, Culture	6.81

Showing 1 to 9 of 9 entries Previous 1 Next

Checking the average sleep duration during working days per department.

```
sleep_df %>%
  group_by(Department_Name) %>%
  summarise(Avg_Sleep_Work_Dep = round(mean(Avg_Sleep_Working_days), digits = 2)) %>%
  arrange(-Avg_Sleep_Work_Dep)
```

Show 10 entries Search:

	Department_Name	Avg_Sleep_Work_Dep
1	Dance	7
2	Drama	6.83
3	Business School	6.7
4	English and Cre	6.66
5	Psychology	6.57
6	Social Sciences	6.54
7	Life Sciences	6.37
8	Other	6.26
9	Media, Culture	5

Showing 1 to 9 of 9 entries Previous 1 Next

Checking the average sleep duration during free days per department.

```
sleep_df %>%
  group_by(Department_Name) %>%
  summarise(Avg_Sleep_Free_Dep = round(mean(Avg_sleep_free_days), digits = 2)) %>%
  arrange(-Avg_Sleep_Free_Dep)
```

Show 10 entries Search:

	Department_Name	Avg_Sleep_Free_Dep
1	Dance	9.33
2	Media, Culture	9.23
3	Life Sciences	9.16
4	English and Cre	8.91
5	Business School	8.86
6	Other	8.86
7	Drama	8.83
8	Psychology	8.8
9	Social Sciences	8.14

Showing 1 to 9 of 9 entries Previous 1 Next

Checking the average weekly sleep duration for year of study.

```
sleep_df %>%
  group_by(Year_of_Study) %>%
  summarise(Avg_Seep_Week_Yr = round(mean(Avg_Weekly_Sleep_Duration), digits = 2)) %>%
  arrange(-Avg_Seep_Week_Yr)
```

Show 10 entries Search:

	Year_of_Study	Avg_Wk_Sleep_Yr
1	2	7.55
2	1	7.47
3	3	7.22

Showing 1 to 3 of 3 entries Previous 1 Next

Checking the average sleep duration during work days for year of study.

```
sleep_df %>%
  group_by(Year_of_Study) %>%
  summarise(Avg_Sleep_Work_Yr = round(mean(Avg_Sleep_Working_days), digits = 2)) %>%
  arrange(-Avg_Sleep_Work_Yr)
```

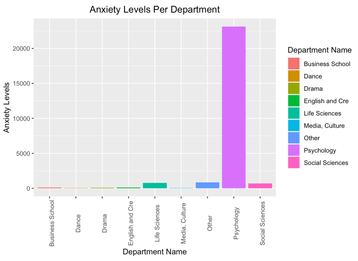
Show 10 entries Search:

	Year_of_Study	Avg_Sleep_Work_Yr
1	2	6.75
2	3	6.53
3	1	6.4

Showing 1 to 3 of 3 entries Previous 1 Next

Checking the anxiety levels between students in different years of study.

```
sleep_df %>%
  ggplot(aes(x = Department_Name, y = Trait_Anxiety)) +
  geom_bar(stat = "identity", aes(fill = Department_Name)) +
  ggtitle("Anxiety Levels Per Department") +
  xlab("Department Name") +
  ylab("Anxiety Levels") +
  guides(fill=guide_legend(title="Department Name")) +
  theme(axis.text.x = element_text(angle = 90)) +
  theme(plot.title = element_text(hjust = 0.5))
```



Checking the anxiety levels between students in different years of study.

```
sleep_df %>%
  ggplot(aes(x = Year_of_Study, y = Trait_Anxiety)) +
  geom_bar(stat = "identity", aes(fill = Year_of_Study)) +
  ggtitle("Anxiety Levels Per Year of Study") +
  xlab("Year of Study") +
  ylab("Anxiety Levels") +
  guides(fill=guide_legend(title="Year of Study")) +
  theme(plot.title = element_text(hjust = 0.5))
```

Conclusions from this section:

I wanted to get an idea of how sleep duration and anxiety changes across department and year of study. This is because some departments are seen as more lax than others, which can result in students have more/less sleep and anxiety. From the exploratory work, individuals in the Business School have the highest average weekly sleep duration, with Media and Culture being the lowest. During work days, students in the Dance department slept the most, with Media and Culture being the lowest here too. Finally, during free days, students in Dance department slept the most, with the Social Sciences being the lowest. For year of study, students in the second year had the highest average weekly sleep duration and the highest average sleep duration during work days. This made sense to me as, in the first year of study students are still adjusting to university-level education and third year students are writing their theses, so they may have to work more and sleep less. Finally, our first bar chart shows that students in the Psychology department experience substantially higher amounts of stress than students in other departments. The second bar chart shows that individuals in the first year have the highest anxiety levels, followed by the second year, and then the third. This could be that students in the first year are less experienced with education at such a level, but slowly adjust by the time they reach their final year.

References

Norbury, R., & Evans, S. (2019). *Time to think: Subjective sleep quality, trait anxiety and university start time.* Psychiatry research, 271, 214-219.