JJohnson_hw2

Jen

9/24/2017

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

Question 1

```
data %>%
  select(Sex, Gender) %>%
  group_by(Gender) %>%
  summarize(mean.encounters = mean(Sex))
```

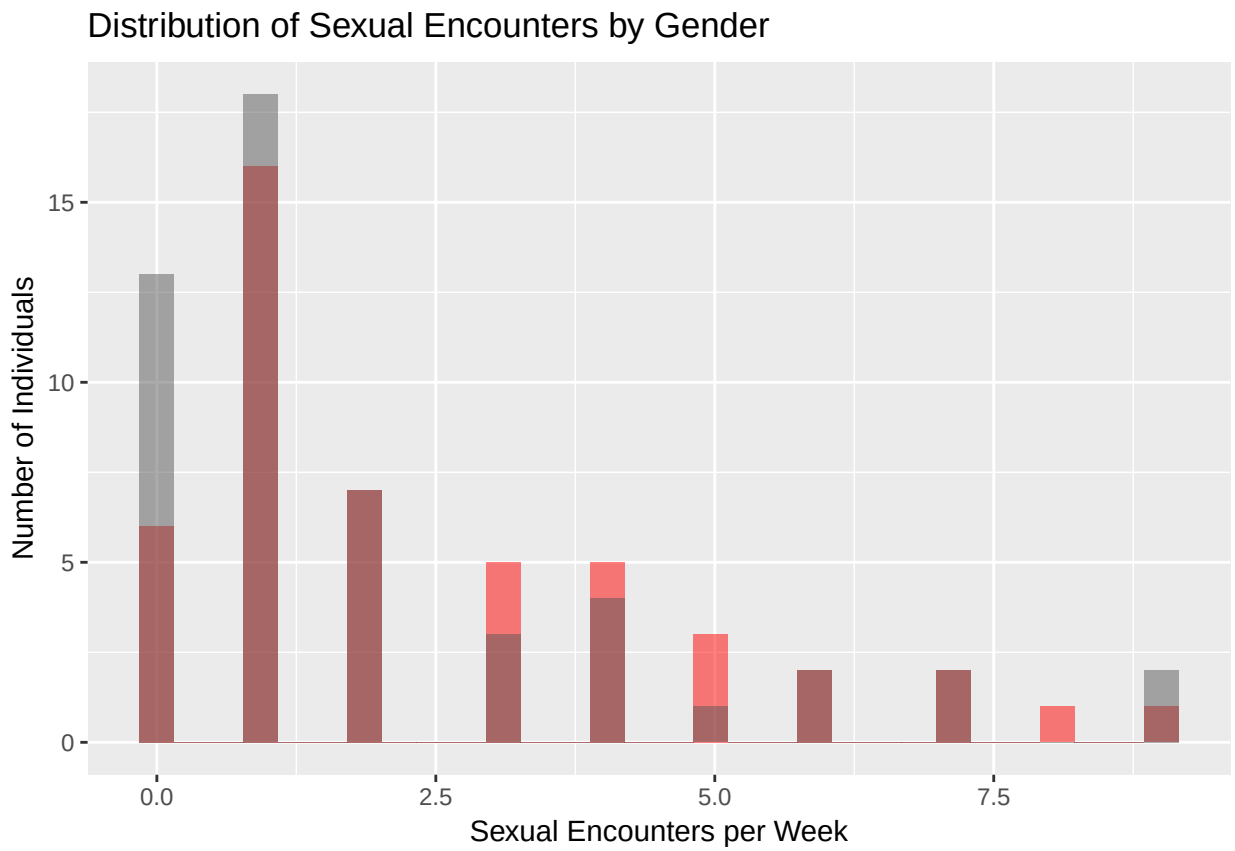
```
## # A tibble: 2 x 2
##   Gender mean.encounters
##   <fctr>           <dbl>
## 1      F        2.562500
## 2      M        2.038462
```

The mean number of sexual encounters per week for women is greater than for men.

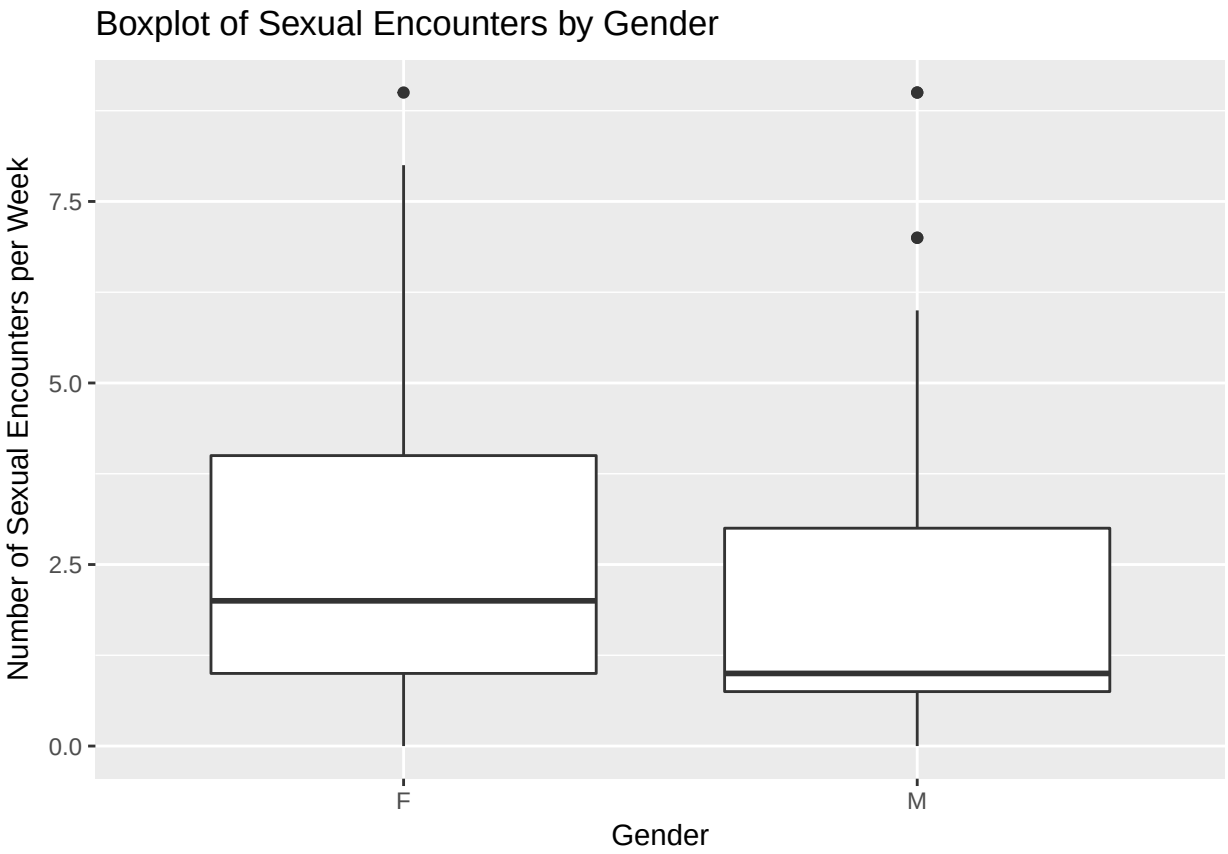
Question 2

```
female <- data %>%
  filter(Gender == "F")
male <- data %>%
  filter(Gender == "M")

ggplot() + geom_histogram(data = female, aes(Sex), fill = "red", alpha = 0.5) + geom_histogram(data = ma
```

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



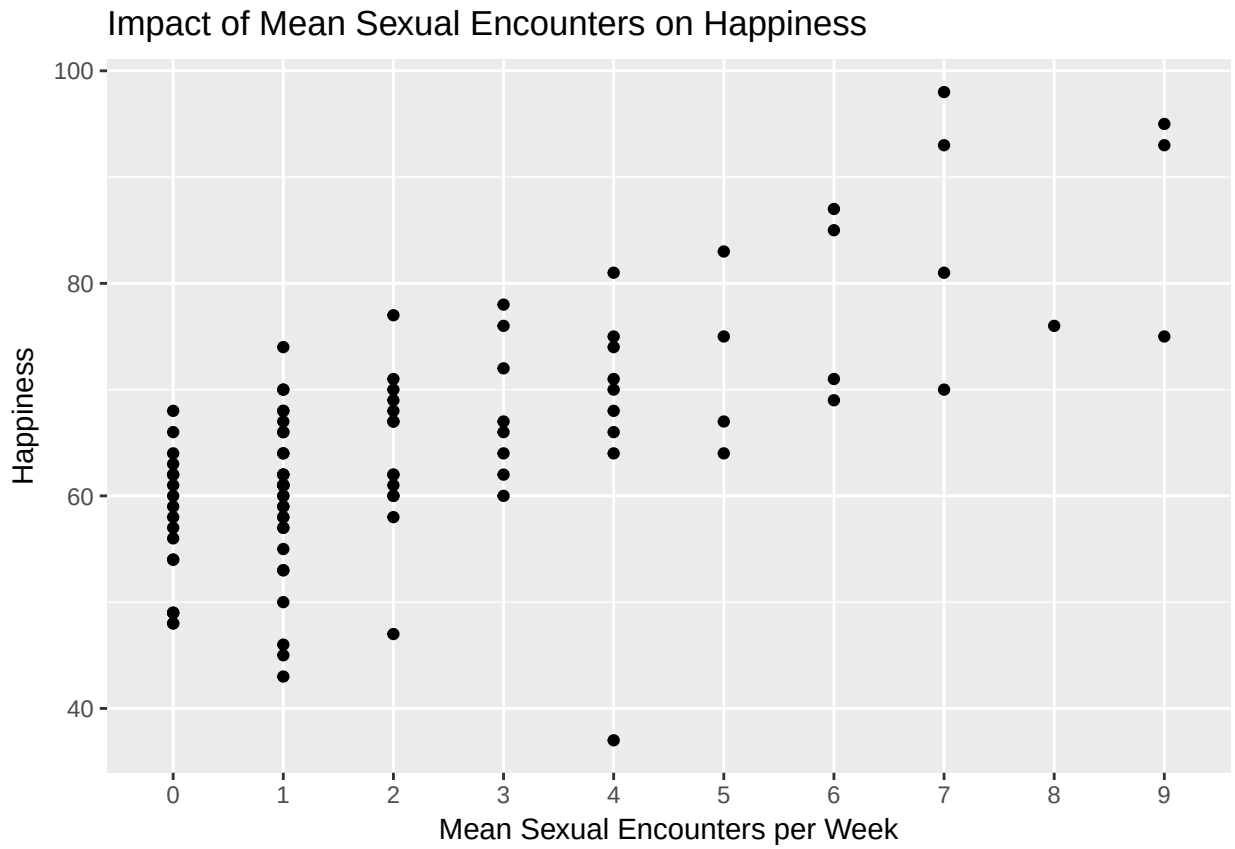

```
ggplot(data, aes(x = Gender, y = Sex)) + geom_boxplot(aes(group = Gender)) + xlab("Gender") + ylab("Numb
```



From the layered histogram, most of the people claimed to have a low number of sexual encounters than high number. More men than women claimed to have low numbers. From the boxplot, there is a wider distribution of the middle 50% of women than men. The median of the male box is very close to the lower boundary.

Question 3

```
ggplot(data, mapping = aes(x = factor(Sex), y = Happiness)) + geom_point(aes(group = Sex)) + ggtitle("In
```

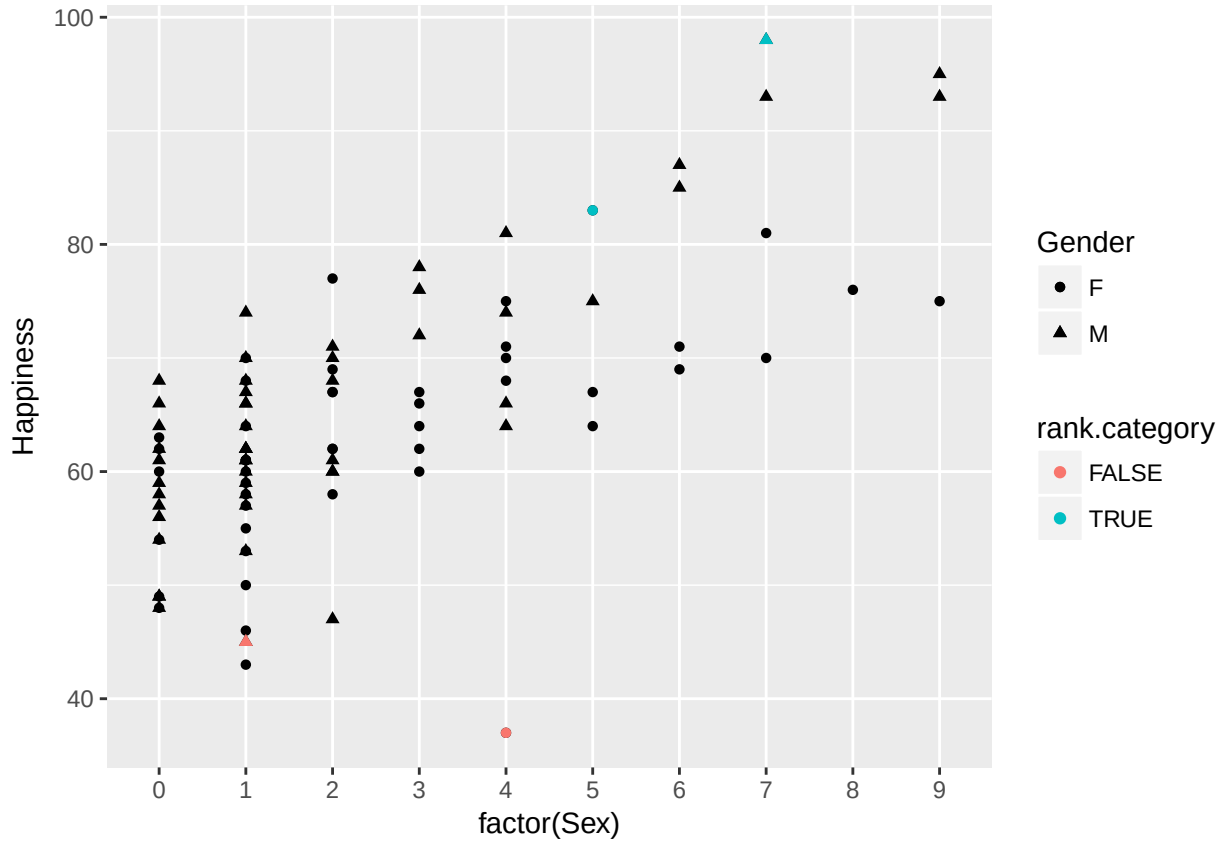


The most happy people claim to have 7-9 sexual encounters a week. The distribution of points is skewed, so more participants claimed to have <=4 encounters a week than >4. This observation is consistent with the findings of Q2. There is a slight relationship between number of sexual encounters and happiness, as shown by the upward trend. However, I do not think this is significant due to due the skewed distribution and the presence of outliers.

Question 4

```
#separate
male_min <- data %>%
  filter(Gender == "M") %>%
  arrange(Happiness) %>%
  mutate(rank.column = rank(Happiness)) %>%
  filter(rank.column == 1)

male_max <- data %>%
  filter(Gender == "M") %>%
  arrange(Happiness) %>%
  mutate(rank.column = rank(Happiness)) %>%
  filter(rank.column == nrow(male))

female_min <- data %>%
  filter(Gender == "F") %>%
  arrange(Happiness) %>%
  mutate(rank.column = rank(Happiness)) %>%
```

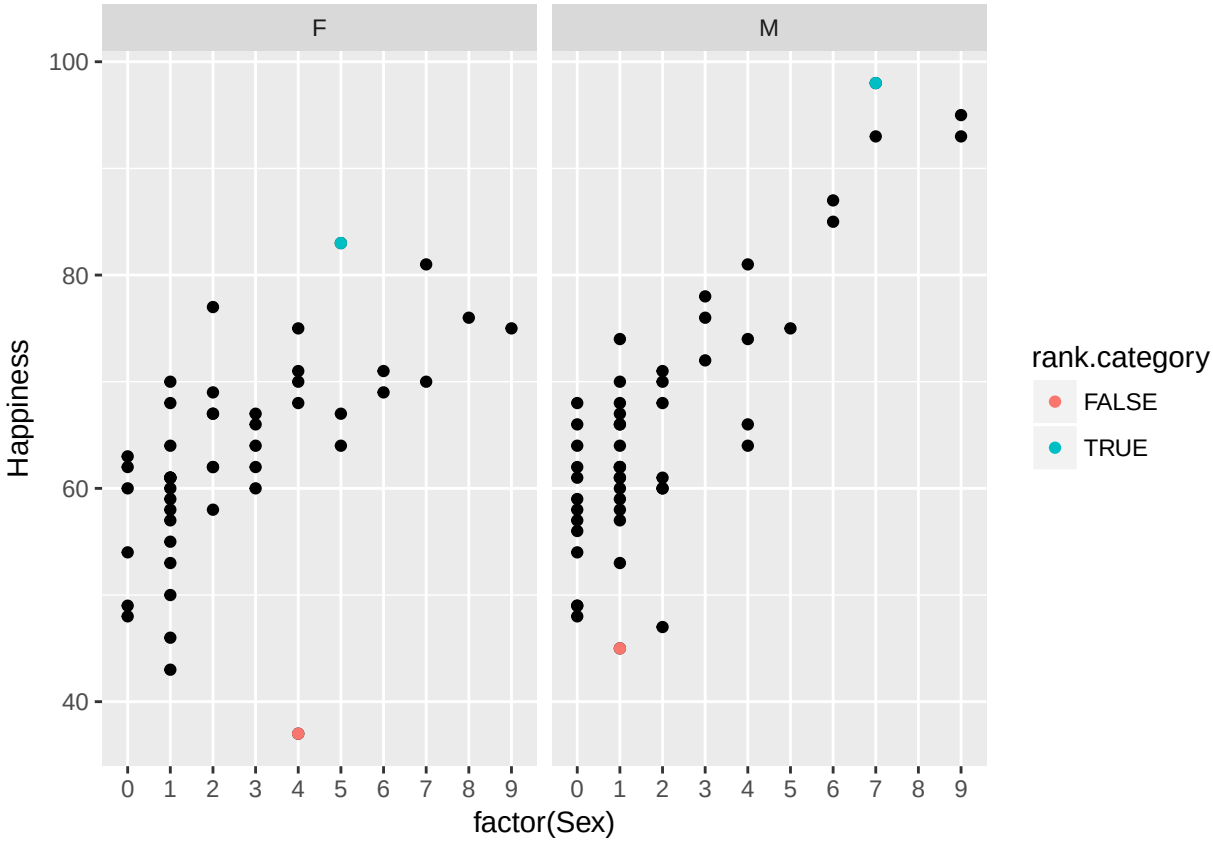
```
  filter(rank.column == 1)

female_max <- data %>%
  filter(Gender == "F") %>%
  arrange(Happiness) %>%
  mutate(rank.column = rank(Happiness)) %>%
  filter(rank.column == nrow(female))

# combine
mins_and_maxes <- rbind(male_min, male_max, female_min, female_max)
df.mins_and_maxes <- as.data.frame(mins_and_maxes)
# convert into category for coloring
df.mins_and_maxes$rank.category <- df.mins_and_maxes$rank.column !=1

#plot
ggplot(data, mapping = aes(x = factor(Sex), y = Happiness)) + geom_point(aes(group = Sex, shape = Gende
```



The happiest male was almost 20 "happiness points" away from the happiest female. The difference between the saddest people was ~ 10 happiness points. The extremes for females had a medium number of sexual encounters. The extremes for males had more "extreme" numbers of sexual encounters. Therefore, it could be said that the correlation between sexual encounters for males is more representative than that for females.

Question 5

```
ggplot(data, mapping = aes(x = factor(Sex), y = Happiness)) + geom_point() + geom_point(data = df.mins_
```



I like this graph better. In the Q4 graph, many of the dots overlap each other and it is difficult to distinguish between the male and female dots, even when they are grouped by shape. I can still compare the difference between genders easily in this graph.

Question 6

```
data %>%
  select(Gender, Marijuana) %>%
  group_by(Gender, Marijuana) %>%
  summarize(n())
```

```
## # A tibble: 4 x 3
## # Groups:   Gender [?]
##   Gender Marijuana `n()`
##   <fctr>     <int> <int>
## 1      F         0    34
## 2      F         1    14
## 3      M         0    37
## 4      M         1    15
```

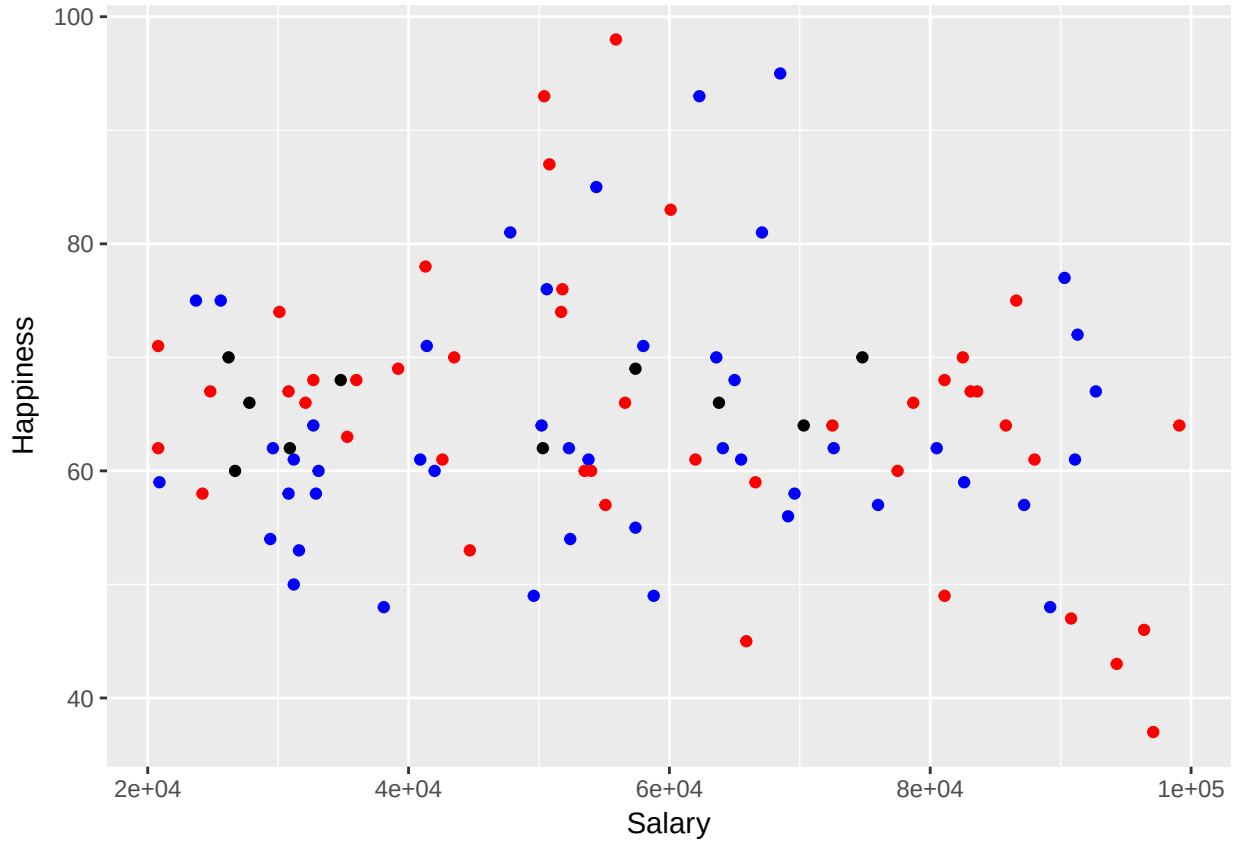
```
data %>%
  select(Gender, Cigarette) %>%
  group_by(Gender, Cigarette) %>%
  summarize(n())
```

```
## # A tibble: 4 x 3
## # Groups:   Gender [?]
##   Gender Cigarette `n()`
##   <fctr>     <int> <int>
## 1      F         0    22
## 2      F         1    26
## 3      M         0    28
## 4      M         1    24
```

Fewer people of both genders smoke Marijuana than do not smoke it. A higher proportion of women smoke than do not smoke, but for men it is the opposite.

Question 7

```
alcohol.salary.happiness <- select(data, Alcohol, Salary, Happiness)
no.alcohol <- alcohol.salary.happiness %>% filter(Alcohol < 1)
some.alcohol <- alcohol.salary.happiness %>% filter(between(Alcohol, 1, 3))
yes.alcohol <- alcohol.salary.happiness %>% filter(Alcohol >3)

ggplot(data = data, aes(x = Salary, y = Happiness)) + geom_point(data = no.alcohol) + geom_point(data =
```



There are not very many people who never drink, and they have generally lower salaries and medium levels of happiness. The people who have a few drinks have a wide range of both salaries and happiness. The people who have the most drinks have an even wider range for happiness.

Question 8

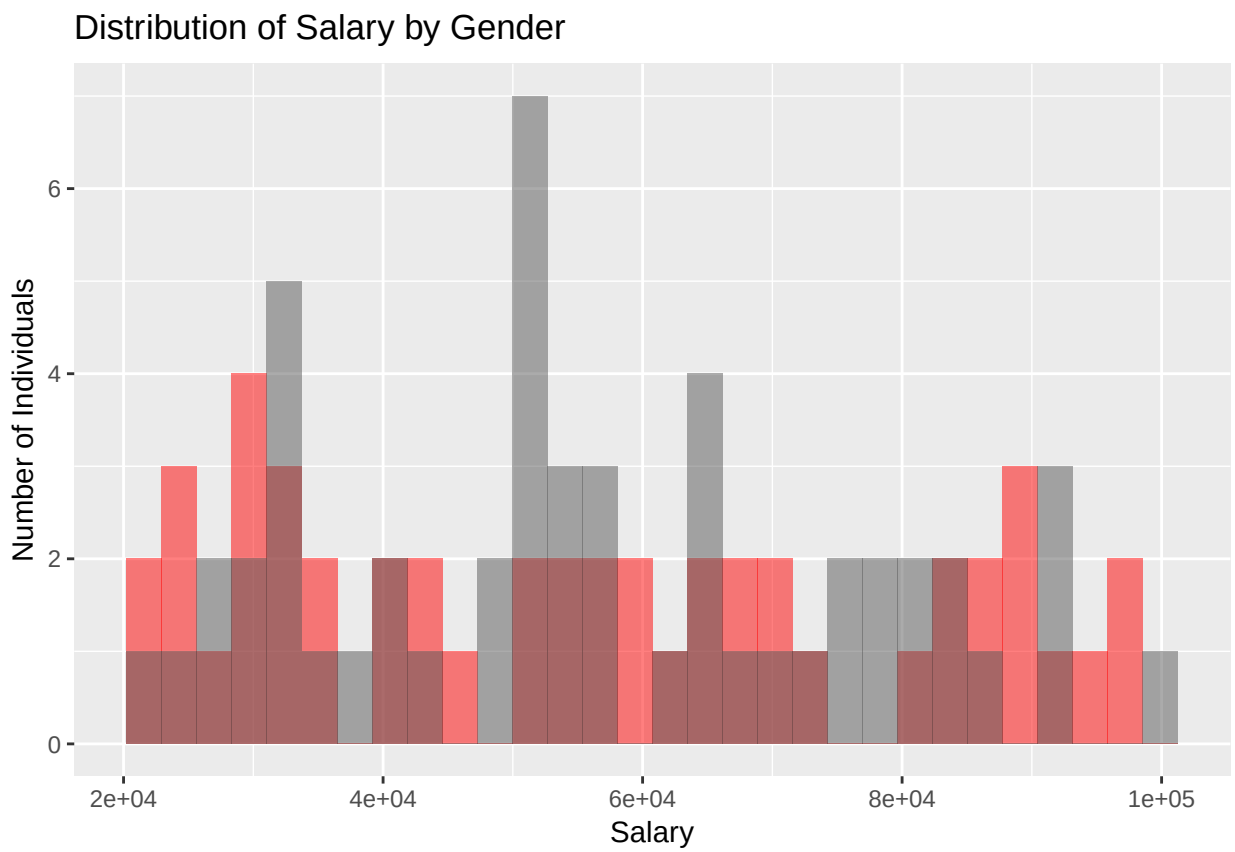
There do not seem to be significant differences between "taboo" factors for the Genders (number of sexual encounters, use of marijuana/cigarettes). The graphs in 3-5 show that they may be a slight correlation between male alcohol consumption and happiness, but I do not think this is significant. This is because the distribution of the colored dots in Q7 do not show any pattern. There is not a relationship between salary and happiness either.

```
data %>%
  select(Salary, Gender) %>%
  group_by(Gender) %>%
  summarize(mean.encounters = mean(Salary))
```

```
## # A tibble: 2 x 2
##   Gender mean.encounters
##   <fctr>           <dbl>
## 1      F        56075.00
## 2      M        56376.92
```

```
ggplot() + geom_histogram(data = female, aes(Salary), fill = "red", alpha = 0.5) + geom_histogram(data =
```

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



The mean salary is slightly higher for men than for women, but the difference is negligible. Men have a much higher number of individuals of median salary (5.5e04). There are a larger number of women who have salaries $<$ 4e04 than men. Above 8e04, there are about the same number of individuals.