GR5702 EDAV Homework 1

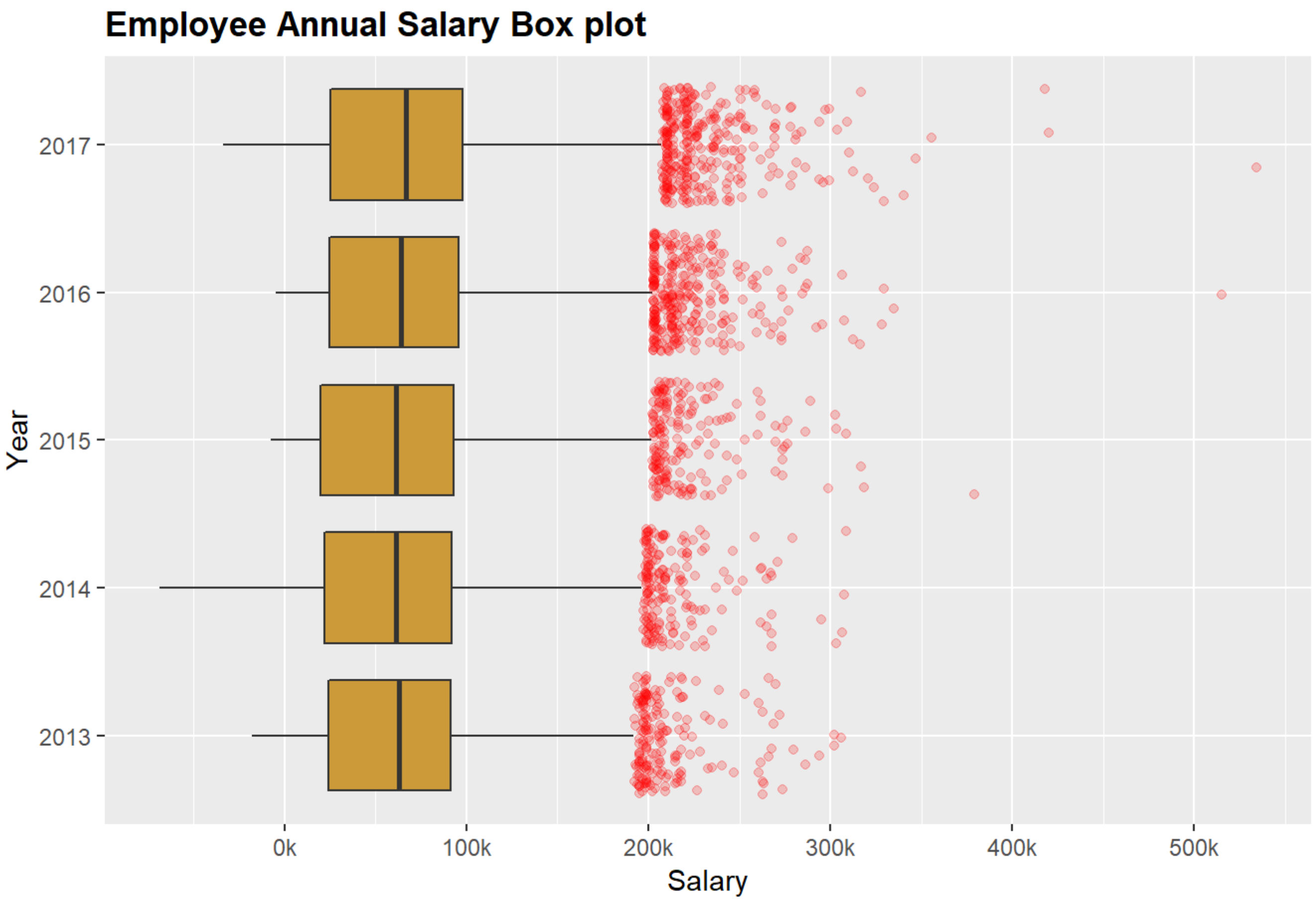
Po-Chieh Liu (pl2441)

2018-09-24

1. Salary

a. Multiple boxplots were generated by the following R codes. Based on the plots, the interquartile ranges of all boxes are visually similar from year 2013 to 2017. Slight differences in median salaries are observed. The first, second and third quartiles of 2016 and 2017 are slightly higher than the rest of the three years. The minimum values vary from year to year. Negative salary records are observed, which might be incorrect data. Further data verification is needed. The yearly maximum outliers slightly increase from year 2013 to 2014, and extreme outliers start showing up after 2015.

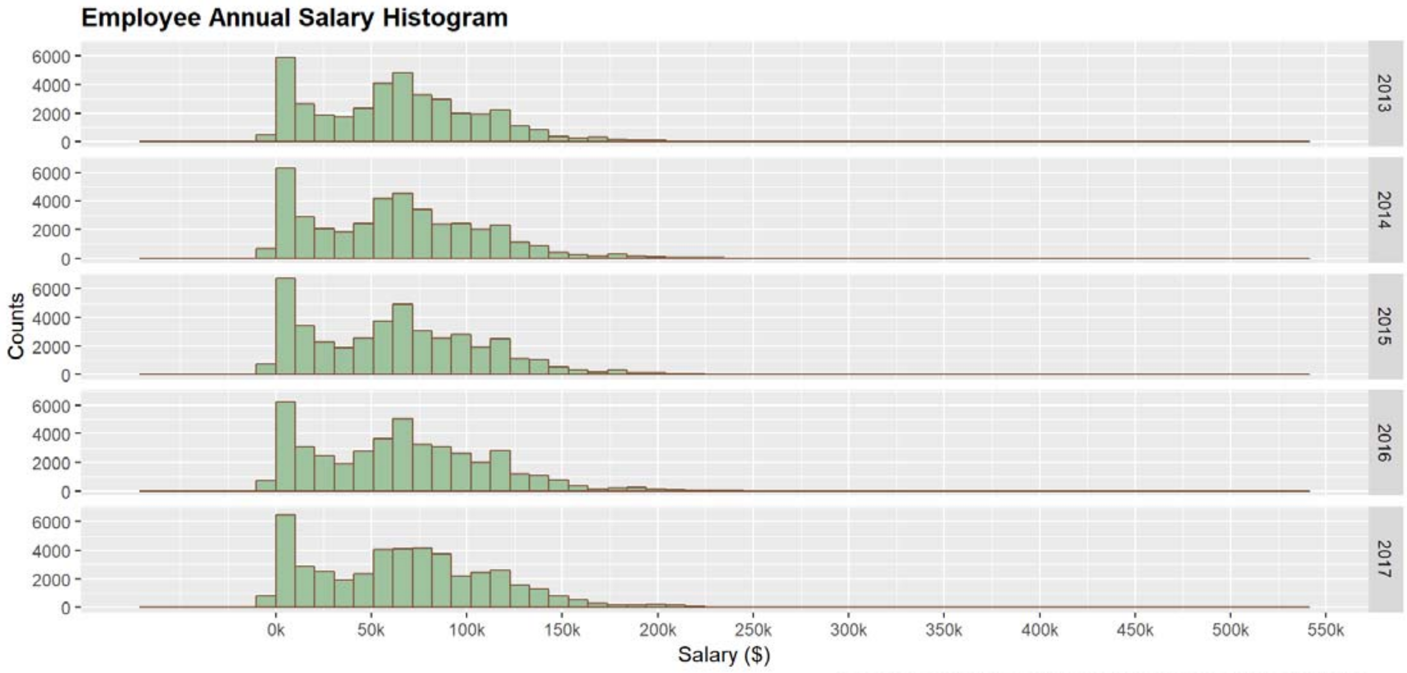
```
employee_df2 <- employee_df %>% group_by(Year) %>%
  mutate(outlier.high = Salaries > quantile(Salaries, .75) + 1.50*IQR(Salaries),
         outlier.low = Salaries < quantile(Salaries, .25) - 1.50*IQR(Salaries)) %>%
  mutate(outlier.color = case_when(outlier.high ~ "red",outlier.low ~ "steelblue")) %>% ungroup

ggplot(employee_df2 , aes(x = factor(Year), y = Salaries)) +
  geom_boxplot(fill = "#cc9a38",outlier.shape = NA) +
  scale_y_continuous(labels = c("0k","100k","200k","300k","400k","500k"),
                     breaks = seq(0, 500000, len = 6)) +
  ggtitle("Employee Annual Salary Box plot") +
  labs(x= "Year",
       y = "Salary",
       caption = "Source: https://catalog.data.gov/dataset/employee-compensation-53987") +
  theme(plot.title = element_text(face = "bold")) +
  theme(plot.caption = element_text(color = "grey68")) +
  geom_jitter(color = employee_df2$outlier.color, alpha = 0.2)+
  coord_flip()
```



Source: https://catalog.data.gov/dataset/employee-compensation-53987

b. Histogram plot of different years were generated by the following R codes. The histogram shapes of each year are similar, which agree with the previous boxplots. The plots show the salaries might potentially be bimodal. First mode is observed around 0k-30k, and second mode is observed around 60k to 70k. The majority of salary records are visually normally distributed.

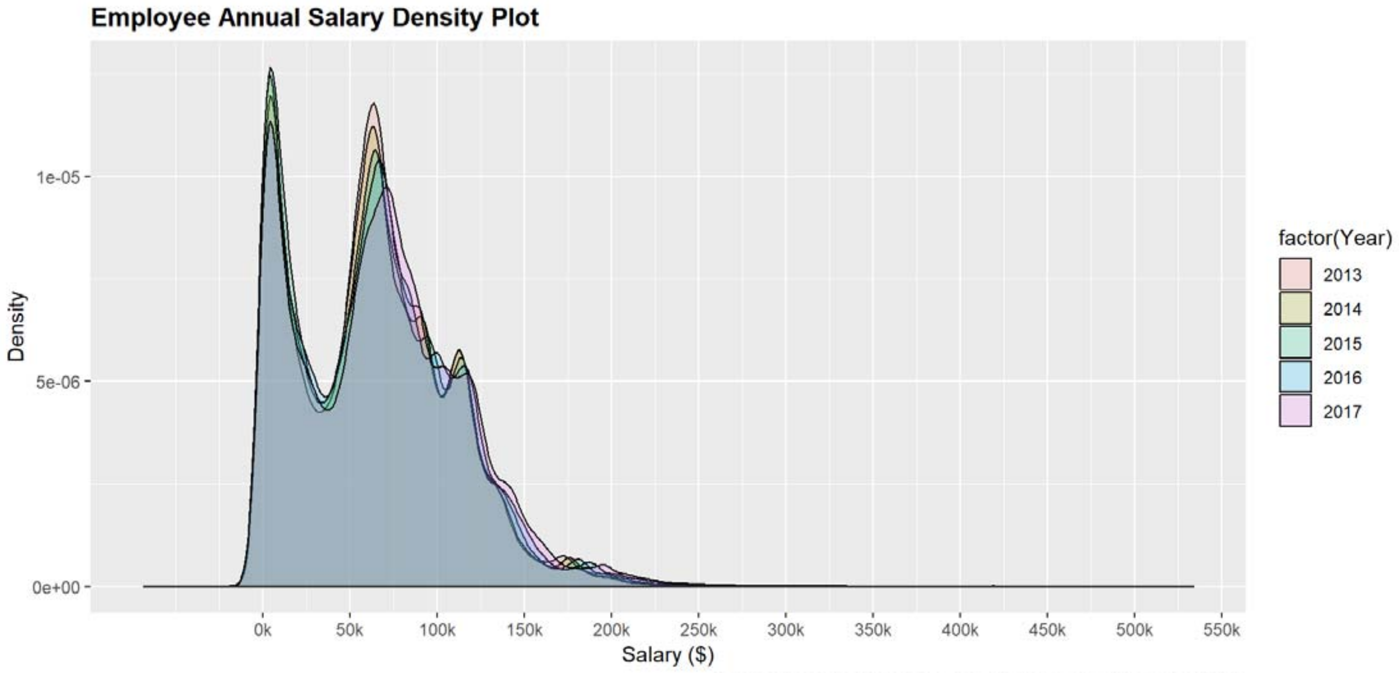
```
ggplot(employee_df, aes(x = Salaries)) +
  geom_histogram(
    bins = 60,
    color = "#80593D",
    fill = "#9FC29F",
    boundary = 0) +
  scale_x_continuous(
    labels = c("0k","50k","100k","150k","200k","250k","300k","350k","400k","450k","500k","550k"
),
    breaks = seq(0, 550000, len = 12)) +
  facet_grid(Year ~.) +
  ggtitle("Employee Annual Salary Histogram") +
  labs(x= "Salary ($)",
       y= "Counts",
       caption = "Source: https://catalog.data.gov/dataset/employee-compensation-53987") +
  theme(plot.title = element_text(face = "bold")) +
  theme(plot.caption = element_text(color = "grey68"))
```



Source: https://catalog.data.gov/dataset/employee-compensation-53987

c. Overlapping density curves were generated by the following R codes. From the density plots we can observe that most under curve areas are overlapped, which indicates the salary distributions from year 2013 to 2017 are remained without large variation. Moreover, the bimodal shape is more obvious than histograms. The density plot also shows the population with income around 50k to 75k decreases from 2013 to 2017. In contrast, the population with income above 75k increases from 2013 to 2017. Plot also shows population with income around 5k increases from 2013 and decreasing from 2016.

```
ggplot(employee_df, aes(x = Salaries)) +
  geom_density( alpha=0.2,  aes(fill = factor(Year))) +
  scale_x_continuous(
    labels = c("0k","50k","100k","150k","200k","250k","300k","350k","400k","450k","500k","550k"
),
    breaks = seq(0, 550000, len = 12)) +
  ggtitle("Employee Annual Salary Density Plot") +
  labs(x= "Salary ($)",
       y= "Density",
       caption = "Source: https://catalog.data.gov/dataset/employee-compensation-53987") +
  theme(plot.title = element_text(face = "bold")) +
  theme(plot.caption = element_text(color = "grey68"))
```



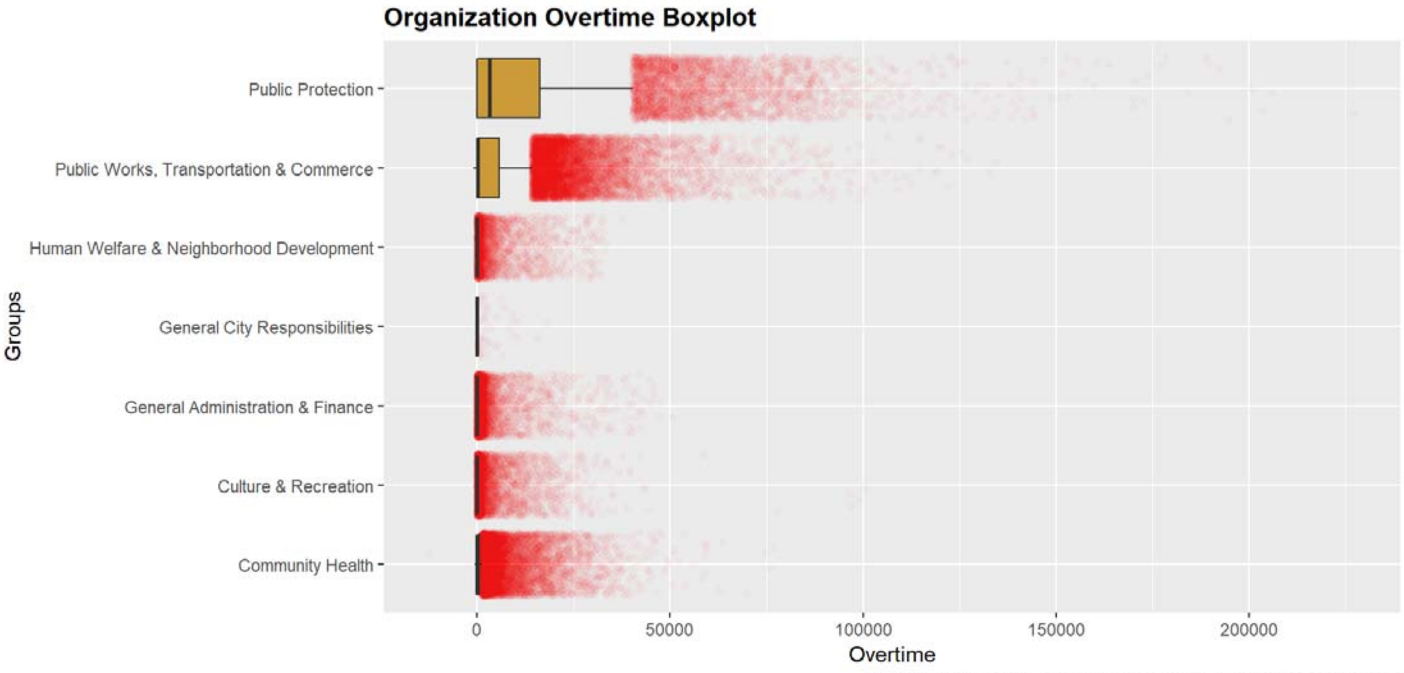
Source: https://catalog.data.gov/dataset/employee-compensation-53987

d. From boxplot of a, we can understand the majority of salary data, and the difference between differnt IQR. From histogram of b, we can observe the salary distribtuion. From density plot of c, we can understand the temporal changes of salary distribtuion from 2013 to 2017.

2. Overtime

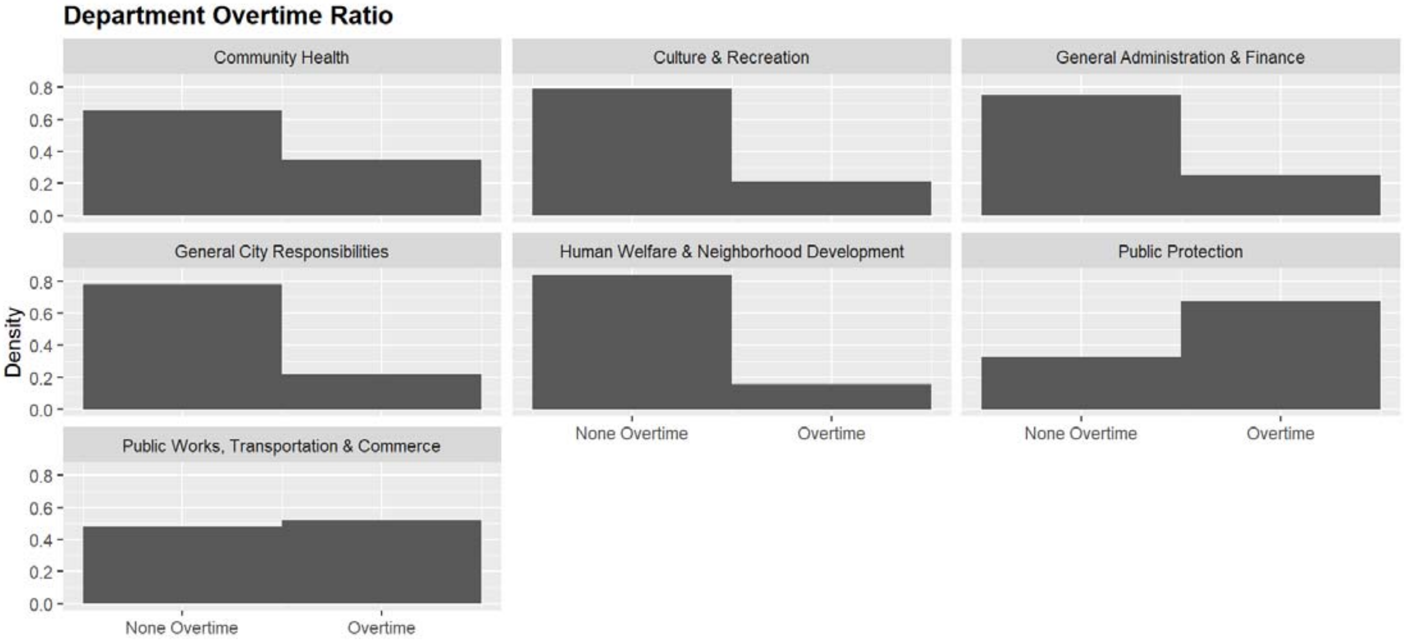
a. Horizontal boxplots of Organization groups’ overtime were generated by the following R code. We can observe majority data points are aroud 0, because for most groups the entire box is located at 0. The only interpretable information from the interquartile boxes. is that two departments have lots heavly overtime employees than other departments. Therefore, the boxplots of overtime data are not very informative majority employees. However, we can observe some information about outliers from this plot. General City Responsibilites has relative few overtime outliers. Human Welfare & Neighborhood Development, General Admisitration & Finace, and Culture & Recreation departments have similar number of outliers around 1.5IQR. Public Works and Community Health outlier ranges are larger. Public Protection shows widely spreaded outliers.

```
employee_df3 <- employee_df %>% group_by(`Organization Group`) %>%
  mutate(outlier.high = Overtime > quantile(Overtime, .75) + 1.50*IQR(Overtime),
         outlier.low = Overtime < quantile(Overtime, .25) - 1.50*IQR(Overtime)) %>%
  mutate(outlier.color = case_when(outlier.high ~ "red",outlier.low ~ "steelblue")) %>% ungroup

ggplot(employee_df3,
       aes(x = reorder(factor(`Organization Group`),Overtime,FUN=median), y = Overtime )) +
  geom_jitter(color = employee_df3$outlier.color, alpha = 0.025)+
  geom_boxplot(fill = "#cc9a38",outlier.shape = NA)+
  coord_flip() +
  ggtitle("Organization Overtime Boxplot") +
  labs(x="Groups", caption = "Source: https://catalog.data.gov/dataset/employee-compensation-539
87") +
  theme(plot.title = element_text(face = "bold")) +
  theme(plot.caption = element_text(color = "grey68"))
```

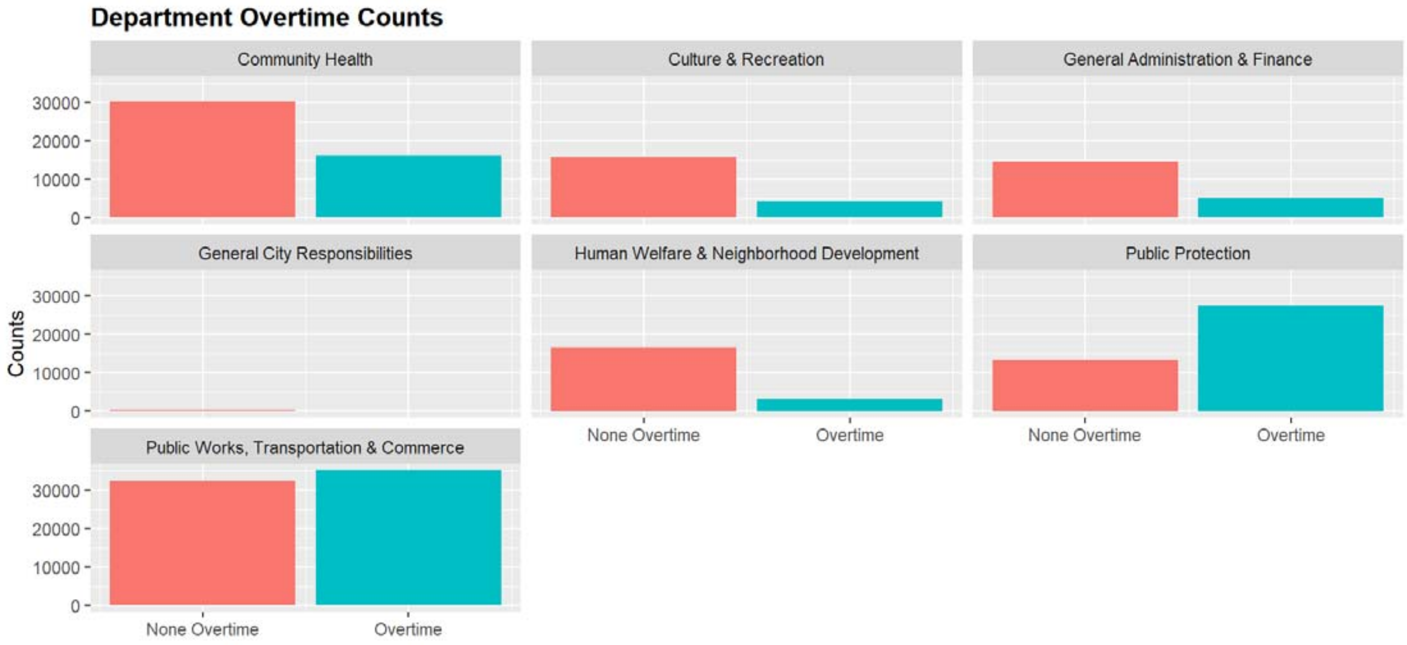


Source: https://catalog.data.gov/dataset/employee-compensation-53987

b. First, we plot the normalized histogram and bar plots of zero and nonzero overtime data points for each department. From the plot we can observe that only Public Protection has overtime employee ratio greater than 50%. The overtime and non-overtime employee ratio is similar for Public works, Transportation and Commerce department. The rest five departments share similar fashion that around 20~40% employees experienced overtime. From the bar plot, we can observe that Community Health and Public Works, Transportation & commerce departments have more employee than other departments. In contrast, General City Responsability is the smallest deparment.

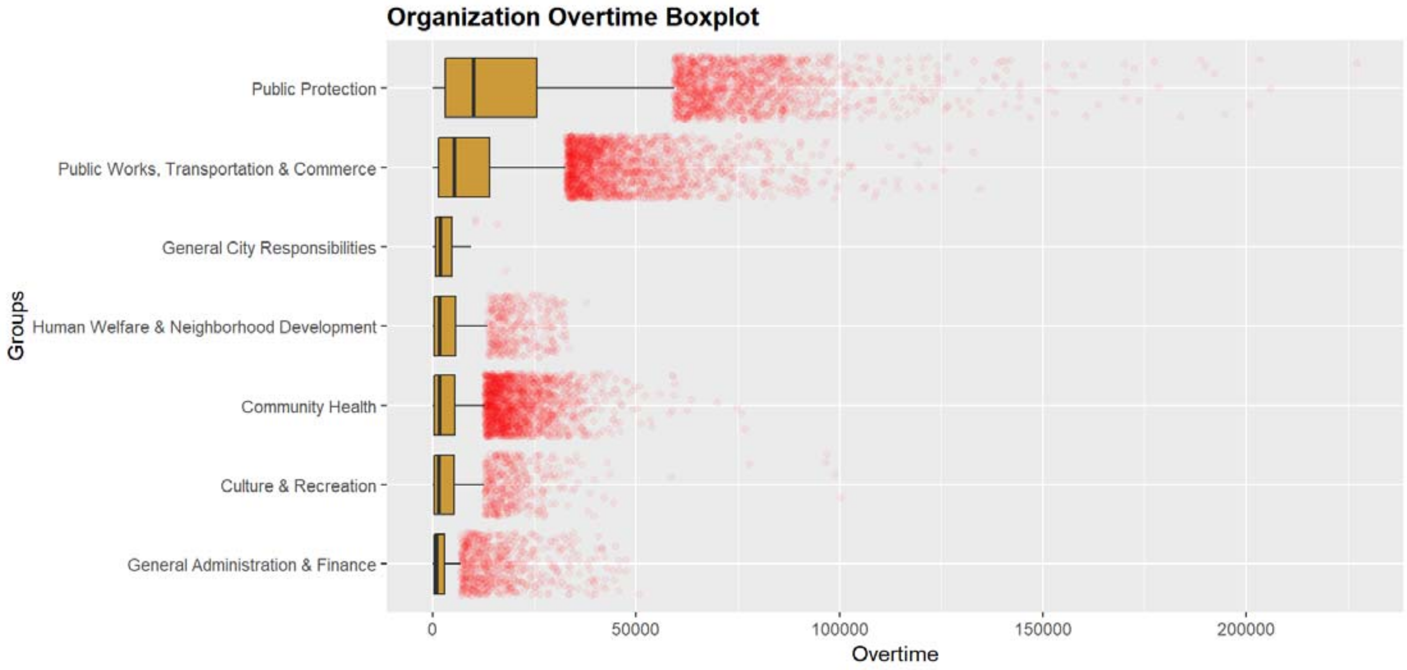


Source: https://catalog.data.gov/dataset/employee-compensation-53987



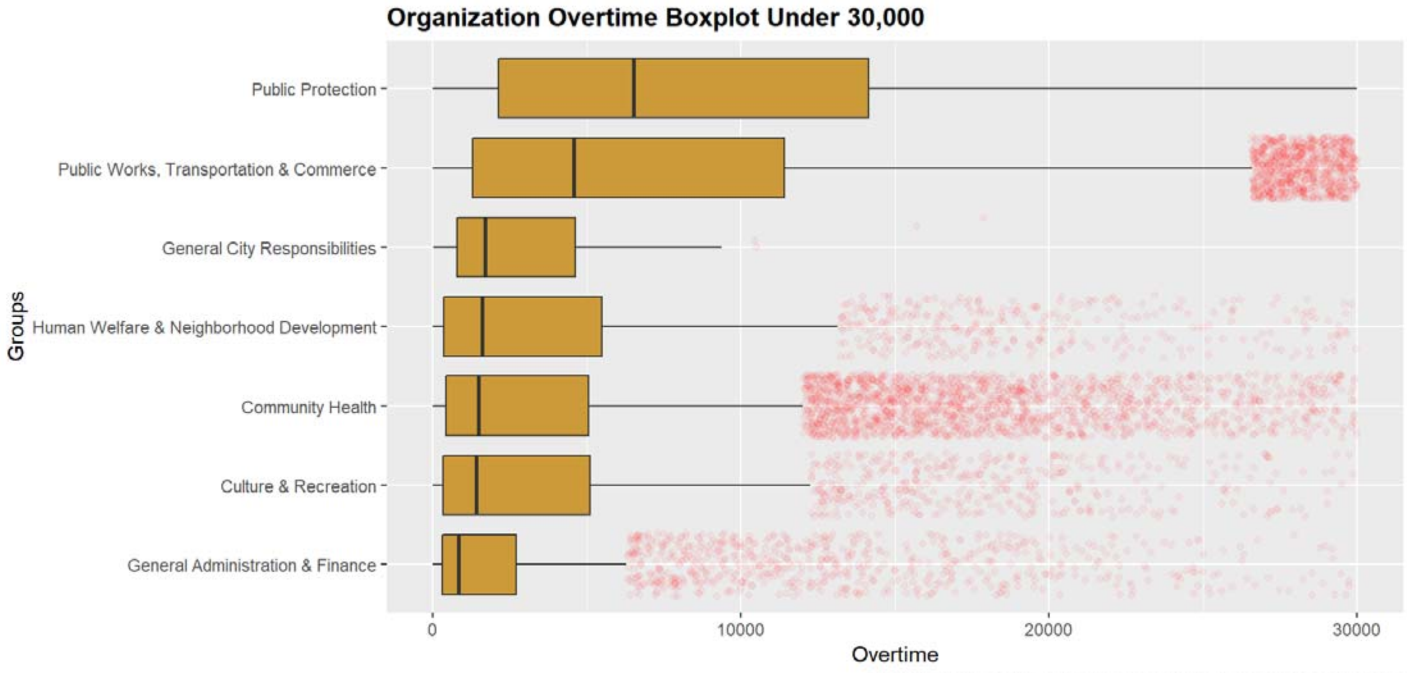
Source: https://catalog.data.gov/dataset/employee-compensation-53987

Since we have basic idea about the non-overitme employees, for example numbers and ratios of overtime employees, now we make a boxplot again excluding zero overtime data to examine the overtime data information. The boxplots provide better visibility of the data than previous boxplots, but still show that marjority data points locate in low overtime ranges. The plot is not very useful as well.



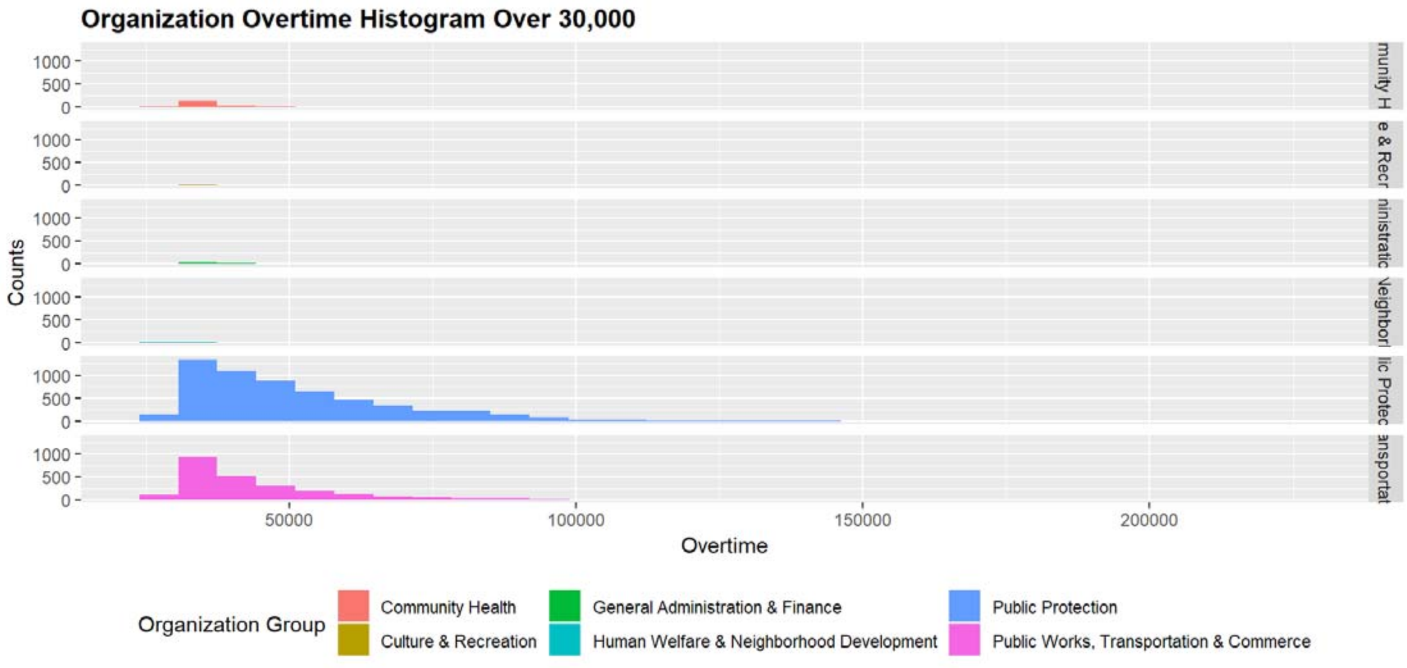
Source: https://catalog.data.gov/dataset/employee-compensation-53987

Then we make boxplots for overtime employees but exclude overtime records above 30,000. There are more inforamtion we can read from this subset plot now. From the boxplots, we can find that overitme employees in the first two deparments might have overtime loading inequality because of the longer interquartile range. For the rest of the four deparments, the overtime IQR are relatively smaller. Also for the last four groups, we can observe that Commumity health has more outliers than other three groups.



Source: https://catalog.data.gov/dataset/employee-compensation-53987

Finally, we make a histogram for overtime record greater than 30,000. The plots show the two public departments have lots employees have overtime more than 30,000. Overall, we can guess based on those plots that both two Public deparments were highly overloaded than other departments. Public Protection department might need more manpower to share the working loads to reduce the high ratio overtime employees. Public Works shows differnt trend that around half employees are overloade. Rest of the departments share same pattern with Publc Works department but with less overtime employees.



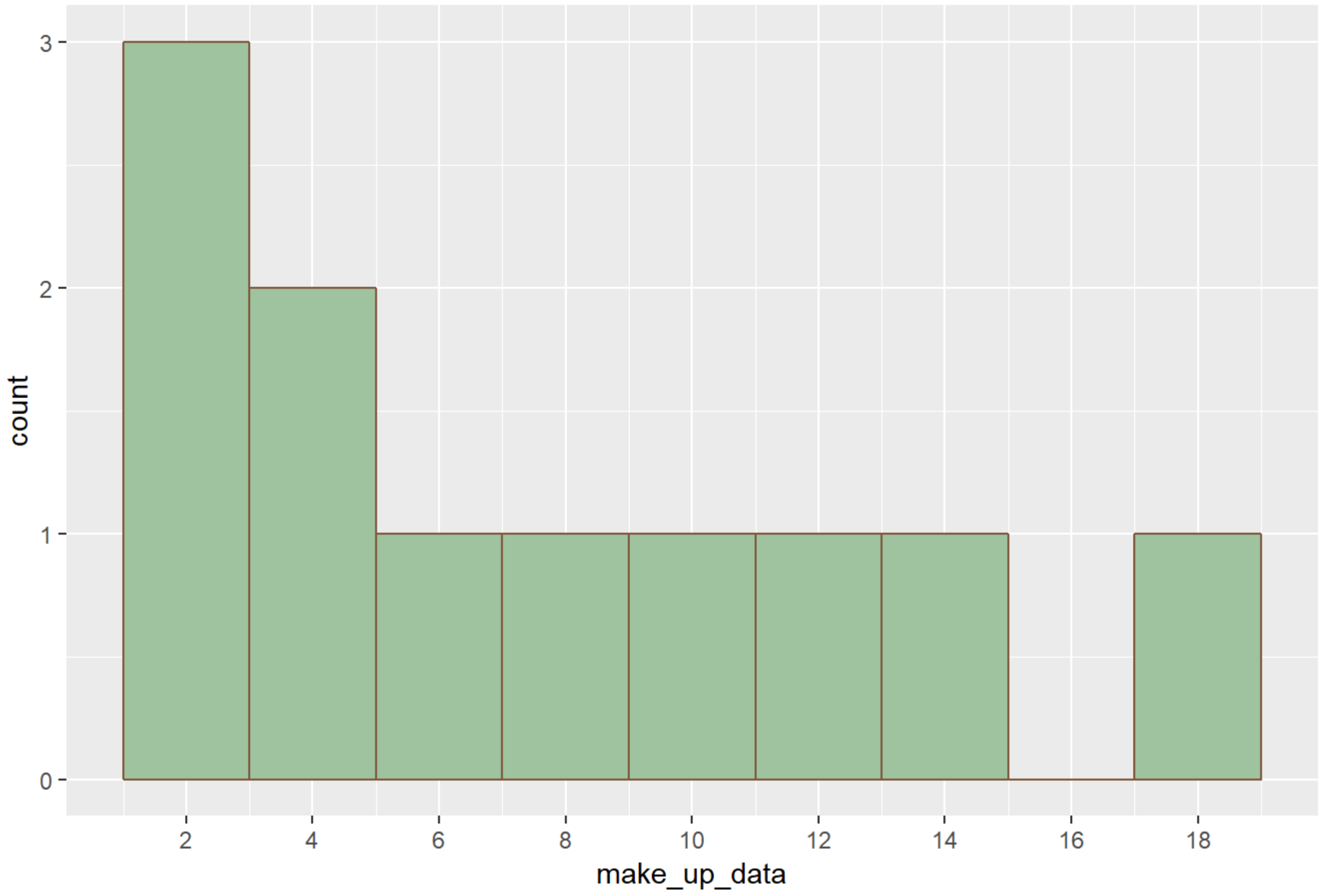
Source: https://catalog.data.gov/dataset/employee-compensation-53987

3. Boundaries

[10 points]

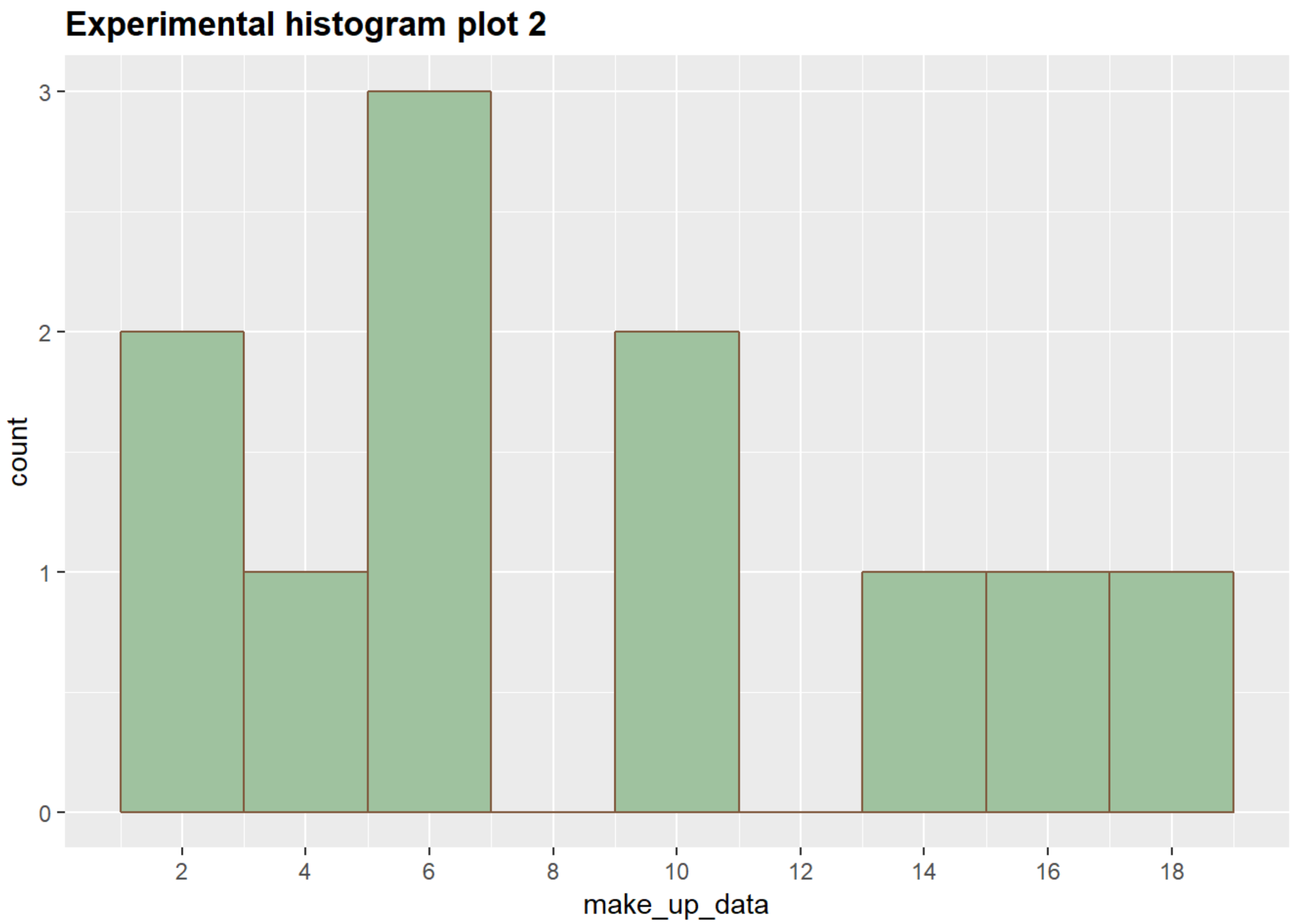
a. Manually generated integer data in range 1 and 18. Plot histogram using binwidth 2, the the left closed or right closed option will show differernt histogram results.

```
make_up_data = c(1,2,3,5,5,6,9,10,13,15,18)
ggplot(as.data.frame(make_up_data), aes(x=make_up_data))+
  geom_histogram(binwidth =  2,color = "#80593D", fill = "#9FC29F") +
  scale_x_continuous(breaks = seq(0, 18, by = 2)) +
  ggtitle("Experimental histogram plot 1") +
  theme(plot.title = element_text(face = "bold"))
```



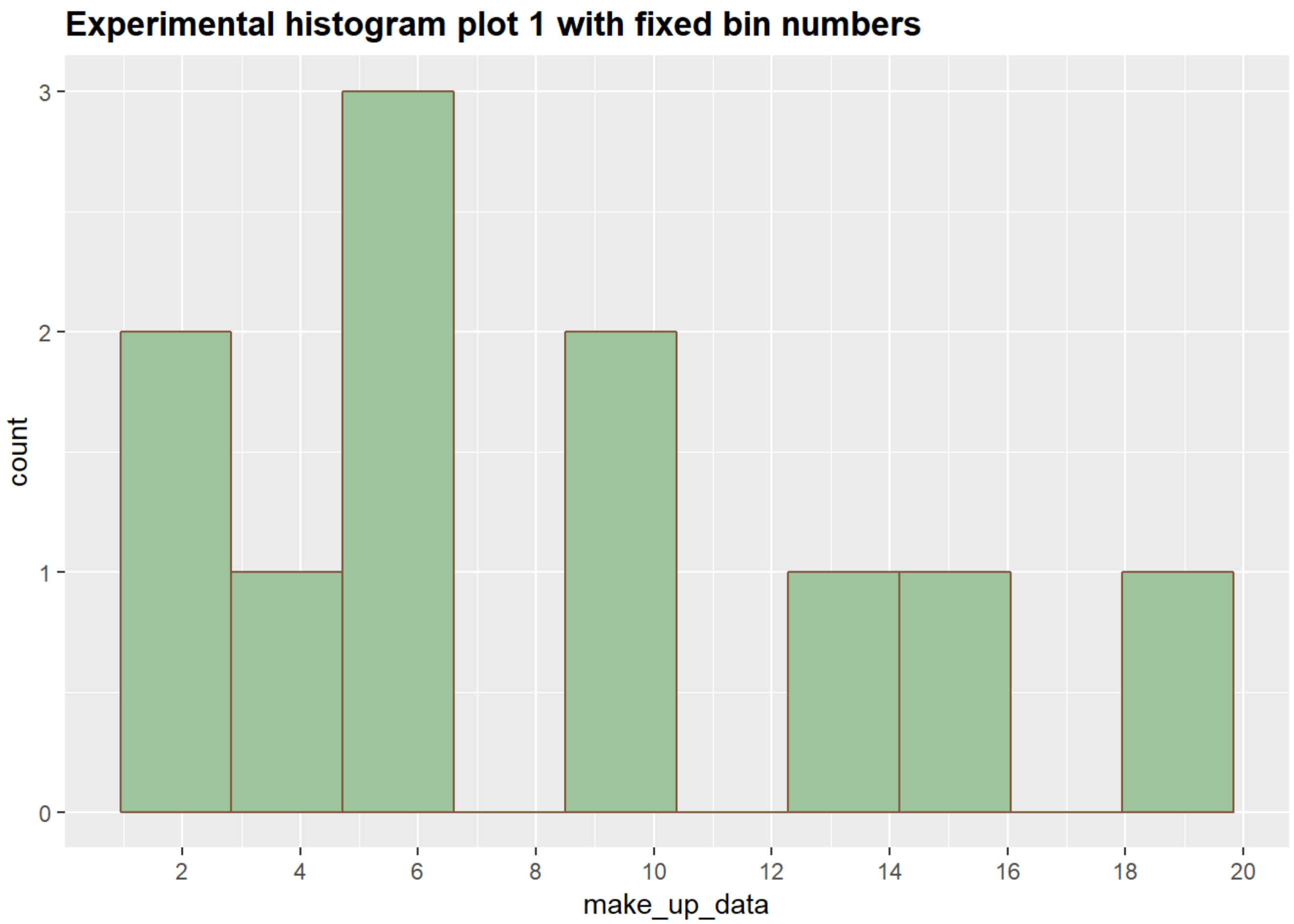
Experimental histogram plot 1

```
ggplot(as.data.frame(make_up_data), aes(x=make_up_data)) +
  geom_histogram(binwidth =  2,closed="left", color = "#80593D", fill = "#9FC29F") +
  scale_x_continuous(breaks = seq(0, 18, by = 2))+
  ggtitle("Experimental histogram plot 2") +
  theme(plot.title = element_text(face = "bold"))
```

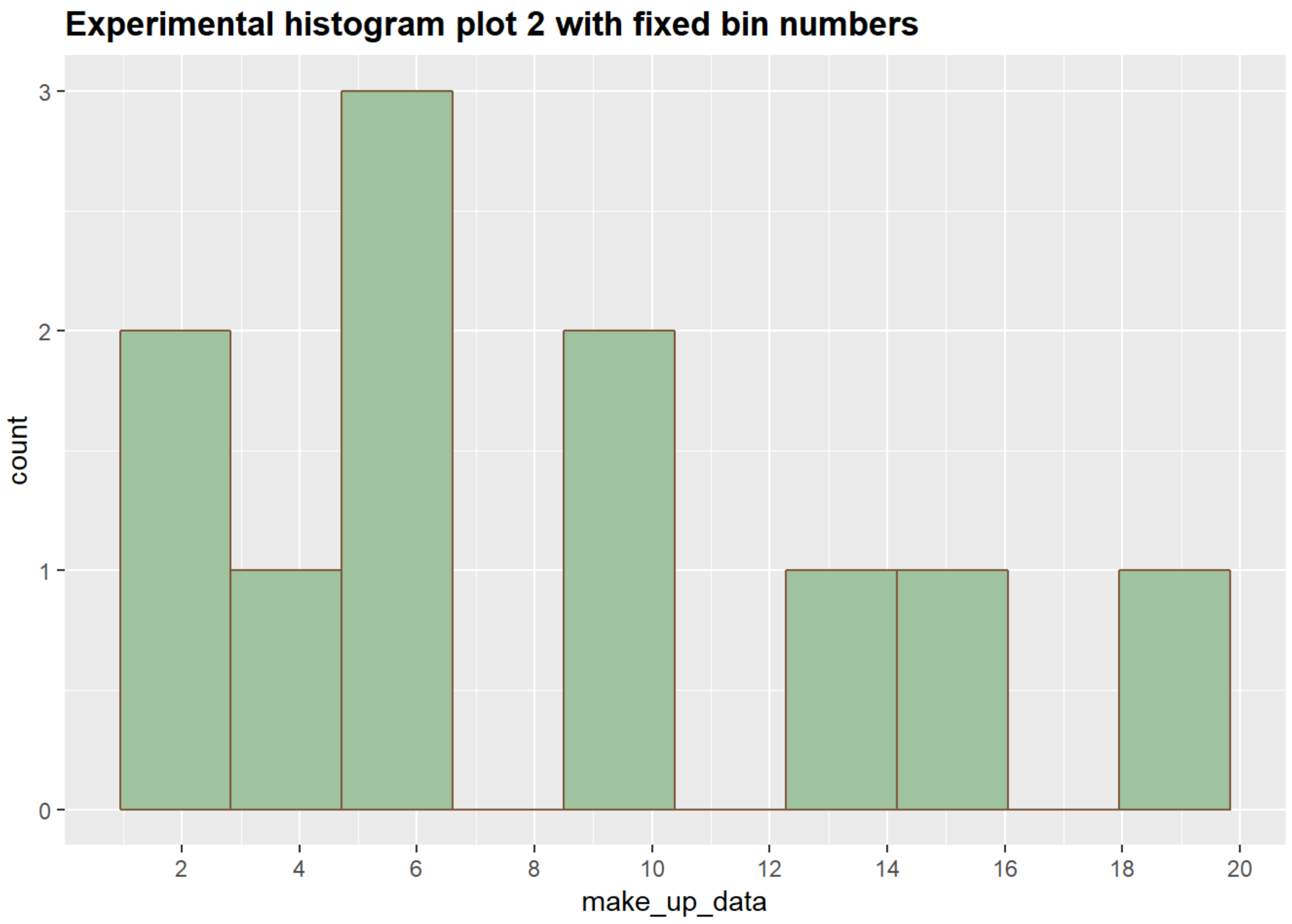


b. The differences were caused by the boundary inclusion issue. Here we add more bins (finer range for each bar) to sovle this issue. Additionaly, We can also give specific binwidth which will make the boundaries not locate on data points to avoid the conflict plots. For this specific data set, if we shift the starting points to 0.5 and keep the original bin number can fix the problem as well. However, it might not help for other data sets.

```
ggplot(as.data.frame(make_up_data), aes(x=make_up_data))+
  geom_histogram(bins = 10,color = "#80593D", fill = "#9FC29F") +
  scale_x_continuous(breaks = seq(0, 20, length.out = 11))+
  ggtitle("Experimental histogram plot 1 with fixed bin numbers") +
  theme(plot.title = element_text(face = "bold"))
```



```
ggplot(as.data.frame(make_up_data), aes(x=make_up_data))+
  geom_histogram(bins = 10,closed="left",color = "#80593D", fill = "#9FC29F") +
  scale_x_continuous(breaks = seq(0, 20, length.out = 11))+
  ggtitle("Experimental histogram plot 2 with fixed bin numbers") +
  theme(plot.title = element_text(face = "bold"))
```

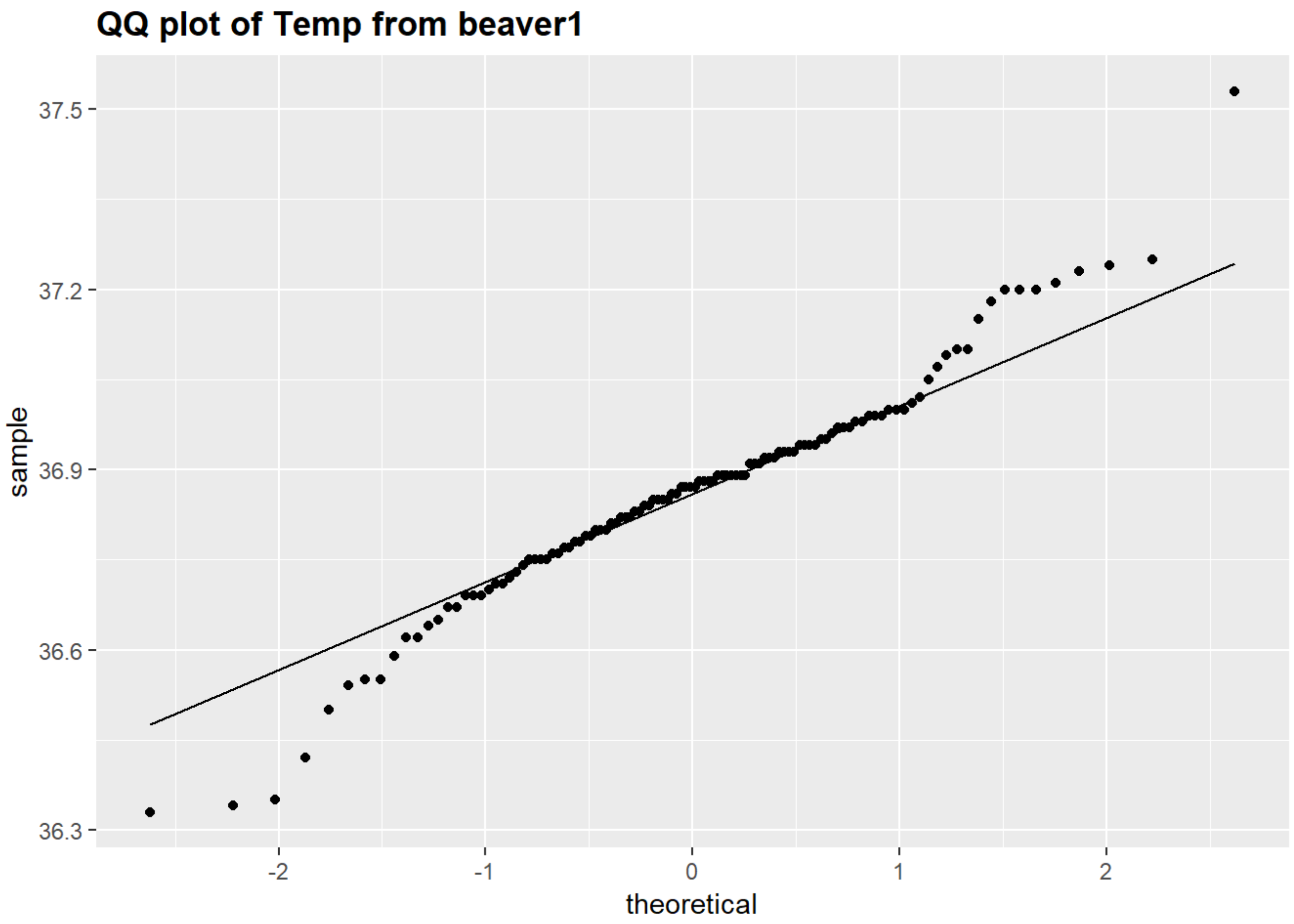


4. Beavers

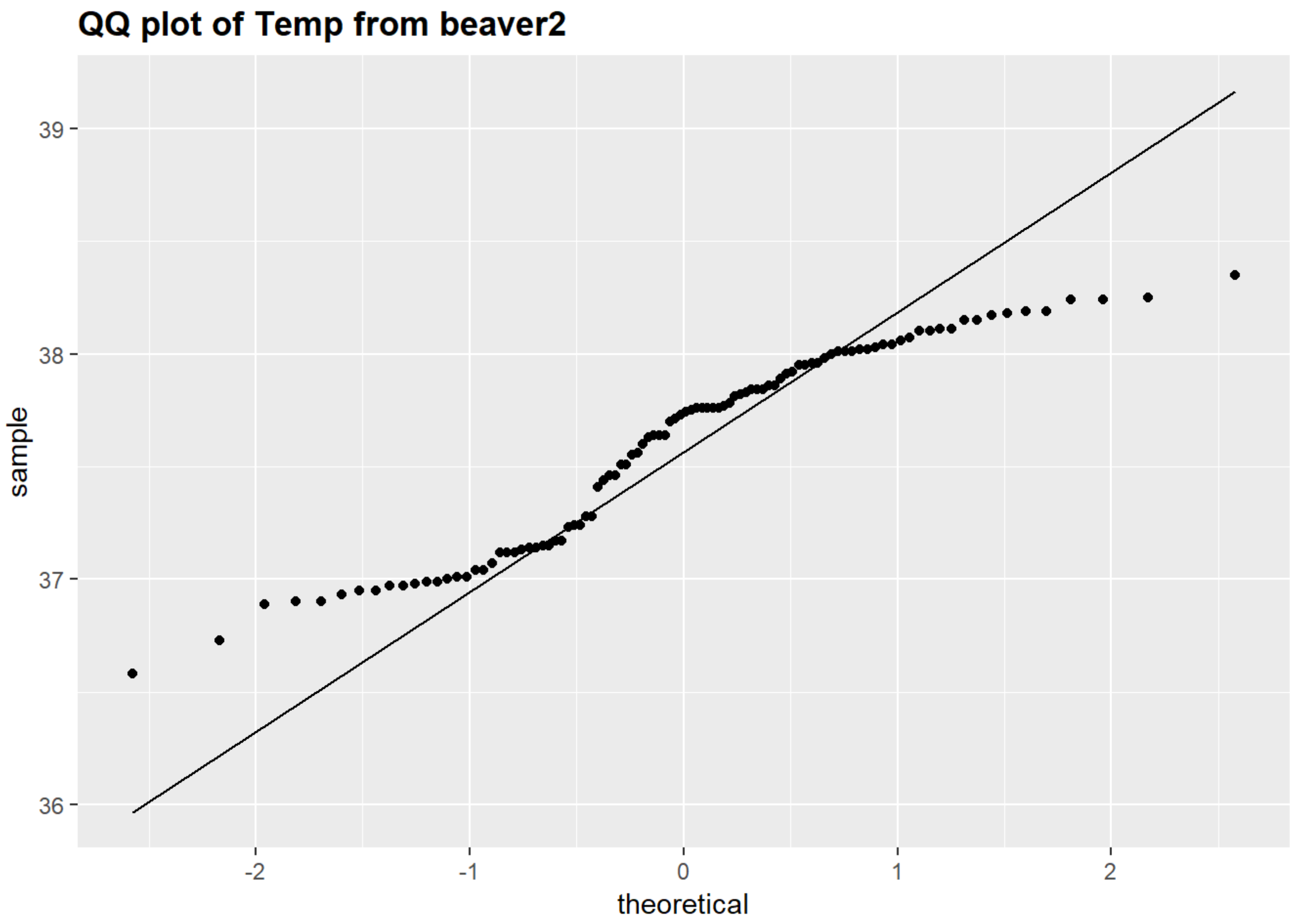
[10 points]

a. The QQ plots of beaver1 temp and beaver2 temp were generated by the following R codes. Based on the plot, both plots are deviated from the line. However, beaver1 temp data points are closer to the theoratical normal line between -1 to 1 range. In contrast, beaver2 basically is not fit the line at all, indicating the beaver1 temp data points are more normally distributed than beaver2 temp data.

```
ggplot(beaver1, aes(sample=temp)) + stat_qq() + geom_qq_line() +
  ggtitle("QQ plot of Temp from beaver1") +
  theme(plot.title = element_text(face = "bold"))
```

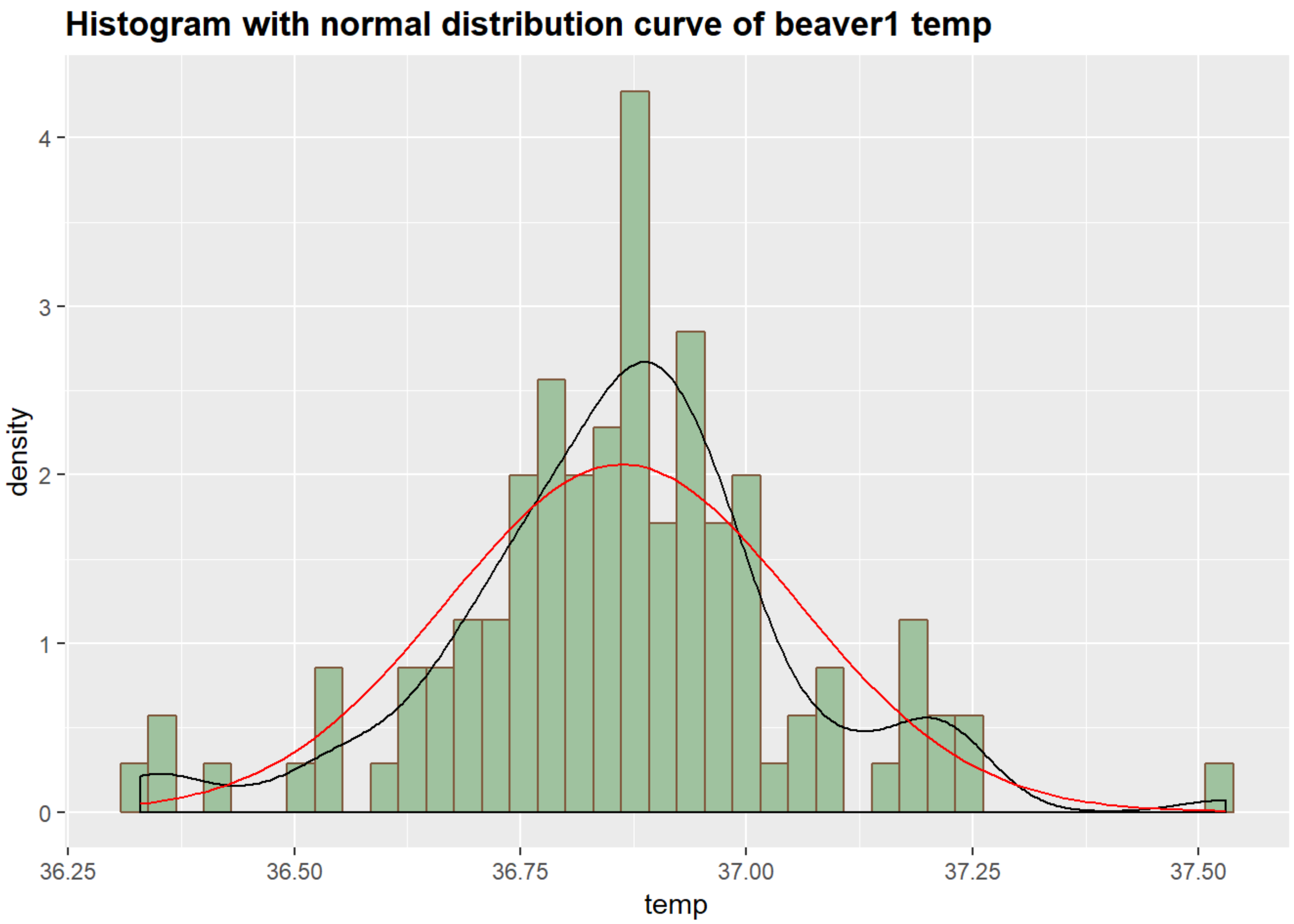


```
ggplot(beaver2, aes(sample=temp)) + stat_qq() + geom_qq_line() +
  ggtitle("QQ plot of Temp from beaver2") +
  theme(plot.title = element_text(face = "bold"))
```

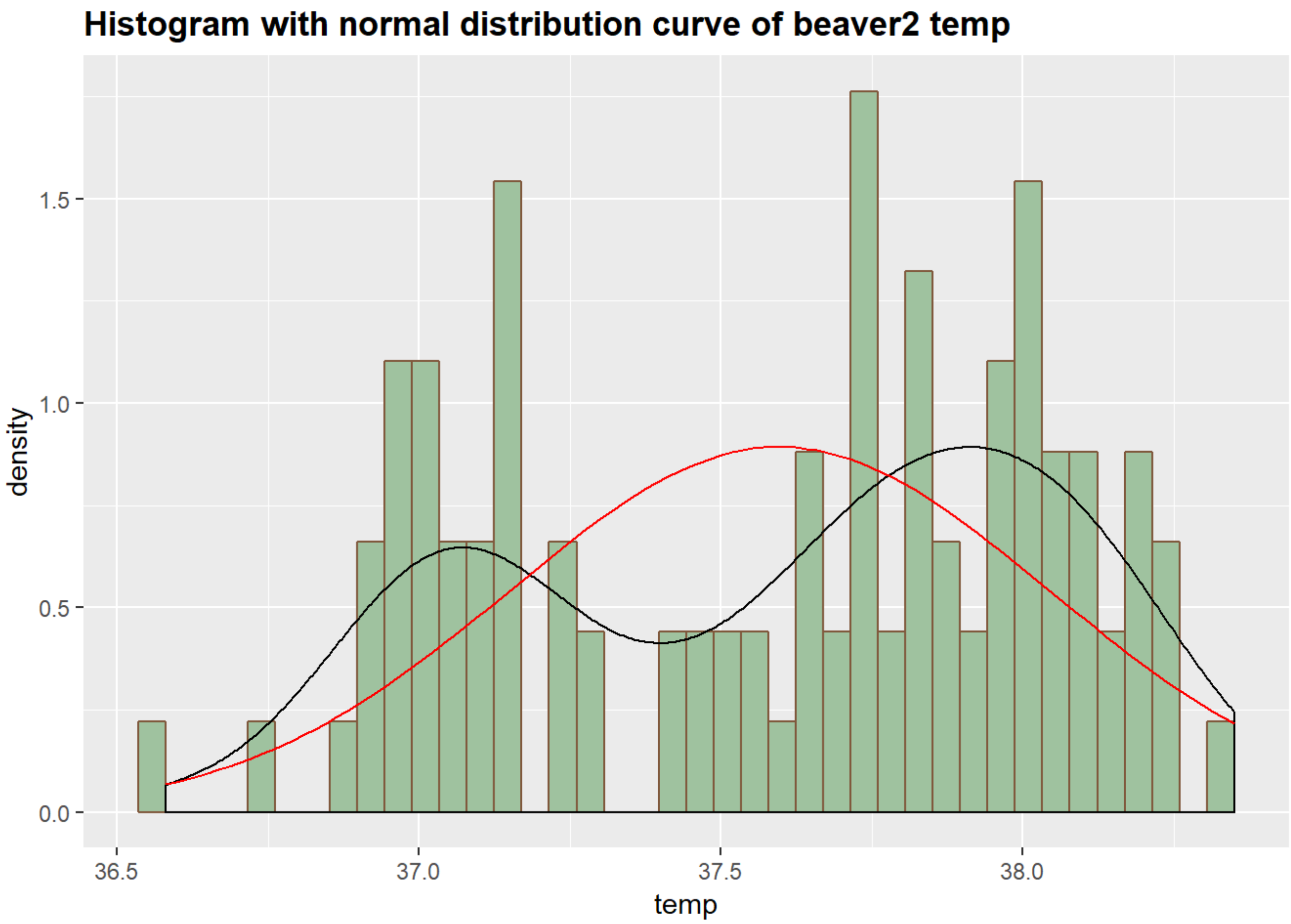


b. The density histograms with density curves were generated by the following R codes. The normal distribtion curves are aslo imposed on each plot. Based on the plots, beaver1 temp density cuvre is closer to normal distribution curve than beaver2. In contrast, beaver2 temp density curve is not similar to normal curve and also shows bimodal property. The plots show beaver1 temp more normally distrbuted than beaver2 temp.

```
ggplot(beaver1, aes( x = temp)) +
  geom_histogram(aes(y=..density..),bins = 40, colour = "#80593D", fill = "#9FC29F", boundary =
0)+
  geom_density() +
  stat_function(fun=dnorm, colour="red",
                args=list(mean=mean(beaver1$temp), sd=sd(beaver1$temp)))+
  ggtitle("Histogram with normal distribution curve of beaver1 temp")+
  theme(plot.title = element_text(face = "bold"))
```



```
ggplot(beaver2, aes( x = temp)) +
  geom_histogram(aes(y=..density..),bins = 40, colour = "#80593D", fill = "#9FC29F", boundary =
0)+
  geom_density() +
  stat_function(fun=dnorm, colour="red",
                args=list(mean=mean(beaver2$temp), sd=sd(beaver2$temp))) +
  ggtitle("Histogram with normal distribution curve of beaver2 temp") +
  theme(plot.title = element_text(face = "bold"))
```



c. The Shapiro-Wilk test for beaver1 and beaver2 temp data were conducted by following R code. The p-value for beaver1 and beaver2 temp are 0.01226 and 0.00007, respectively. Judging by the commonly used p-value threshold 0.05, we reject the hypothesis for both beaver 1 and beaver2 temp data. This results indicated that both data sets might not be normally distributed, however,the beaver1 temperature is closer to normal distribution than beaver2.

```
shapiro.test(beaver1$temp)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  beaver1$temp
## W = 0.97031, p-value = 0.01226
```

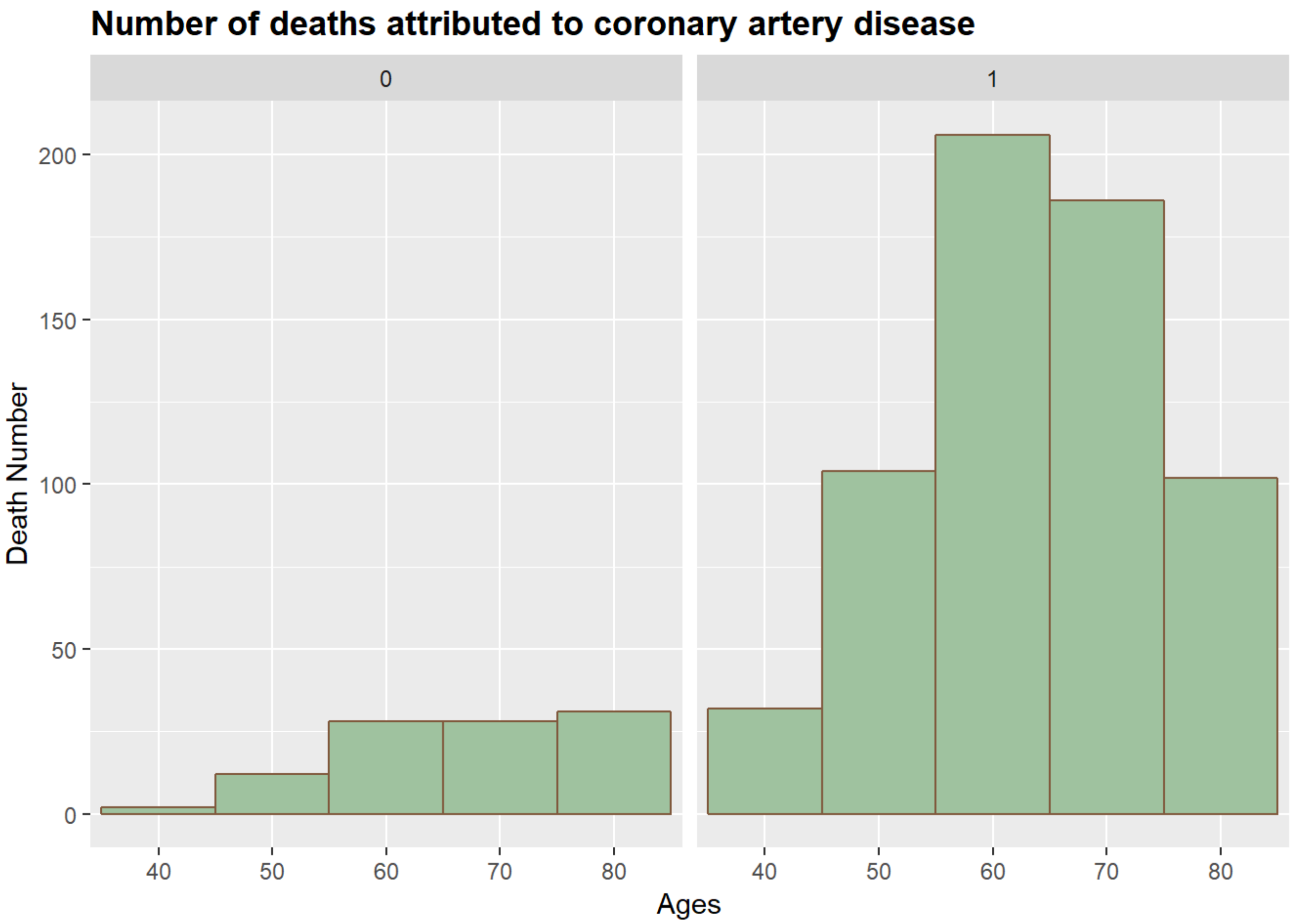
```
shapiro.test(beaver2$temp)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  beaver2$temp
## W = 0.93336, p-value = 7.764e-05
```

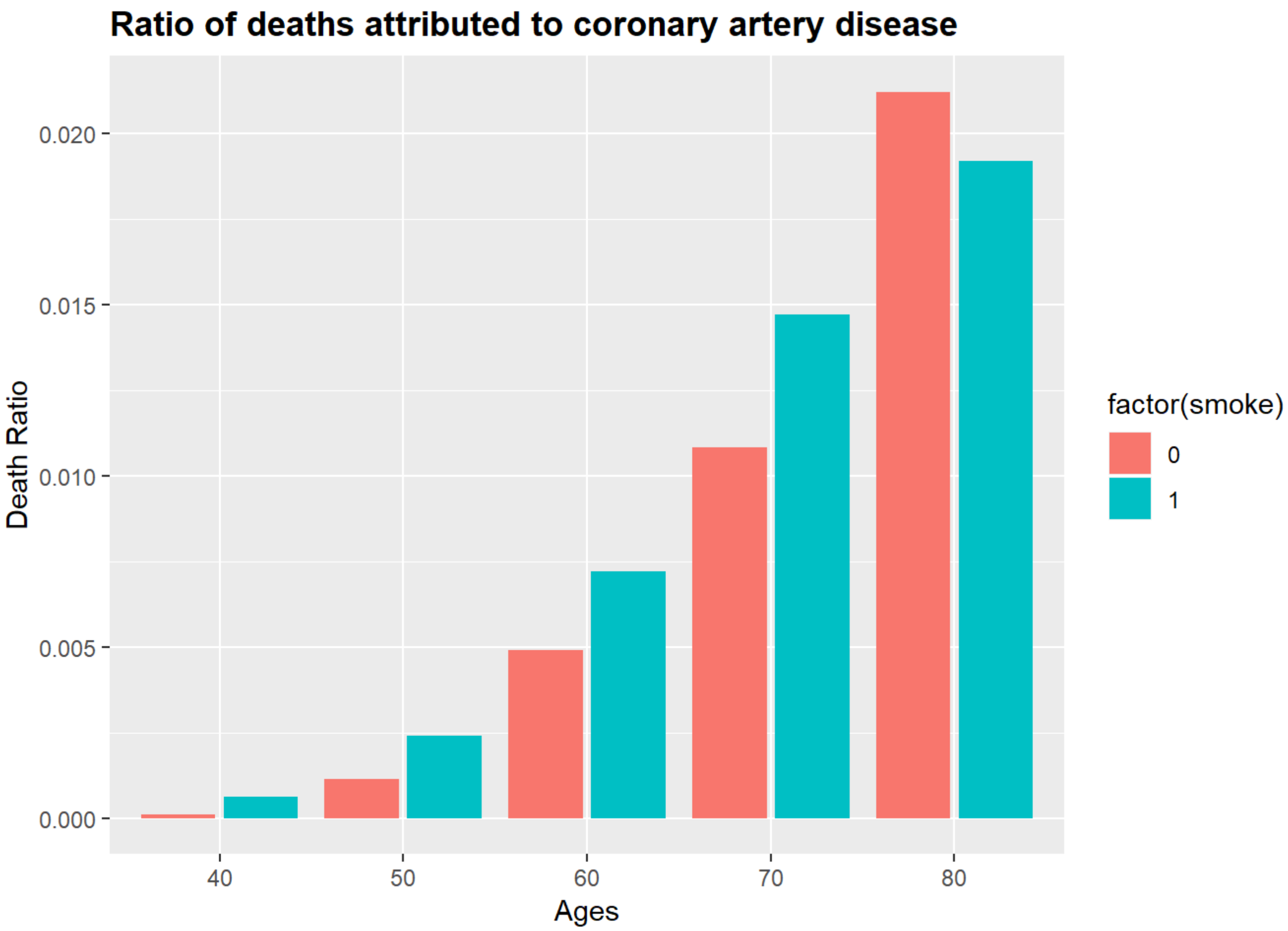
5. Doctors

The following R codes plot the historgram of deaths attributed to coronary artery disease for non-smoking (0) and smoking (1) groups. From the plot, we can find that numbers of death of each age in smoking group are higher than non-smaking group. Further we plot deaths ratio of each age group, we can observe that death ratio of each group under age 80 are higher than nonsmoking group. However, the age over 80 group shows opposite fashion. The death ratio might be more convincible than absolute death number to compare the A/B test type results. The incosist high/low ratio for above age 80 might cause by other reasons.

```
library(boot)
data("breslow")
ggplot(breslow, aes(x=age, y= y))+
  geom_bar(stat="identity",width = 1, color = "#80593D", fill = "#9FC29F") +
  facet_grid(.~smoke)+
  labs(x="Ages", y= "Death Number")+
  ggtitle("Number of deaths attributed to coronary artery disease") +
  theme(plot.title = element_text(face = "bold"))
```



```
ggplot(breslow, aes(x=age, y= y/n, fill = factor(smoke)))+
  geom_bar(stat = "identity",position = "dodge2") +
  labs(x="Ages", y= "Death Ratio")+
  ggtitle("Ratio of deaths attributed to coronary artery disease") +
  theme(plot.title = element_text(face = "bold"))
```


Ratio of deaths attributed to coronary artery disease
Death Ratio
0.020
0.015
0.010
0.005
0.000
40
50
60
70
80
Ages
factor(smoke)
0
1