Data Visualization

Read the data

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts -------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(s20x)
```

I will use RStudio to do question 4. This has been approved by the instructor.

We will do data visualizations on two data sets. The first one is employee with 108 records, and the second one is movies, with 100 records.

```
# Read the data into R
movie = read_csv("movies.csv")
```

```
## Parsed with column specification:
## cols(
##   addr = col_character(),
##   room_num = col_double(),
##   mdate = col_date(format = ""),
##   start_time = col_time(format = ""),
##   genre = col_character(),
##   rating = col_double()
## )
```

```
employee = read_csv("employee.csv")
```

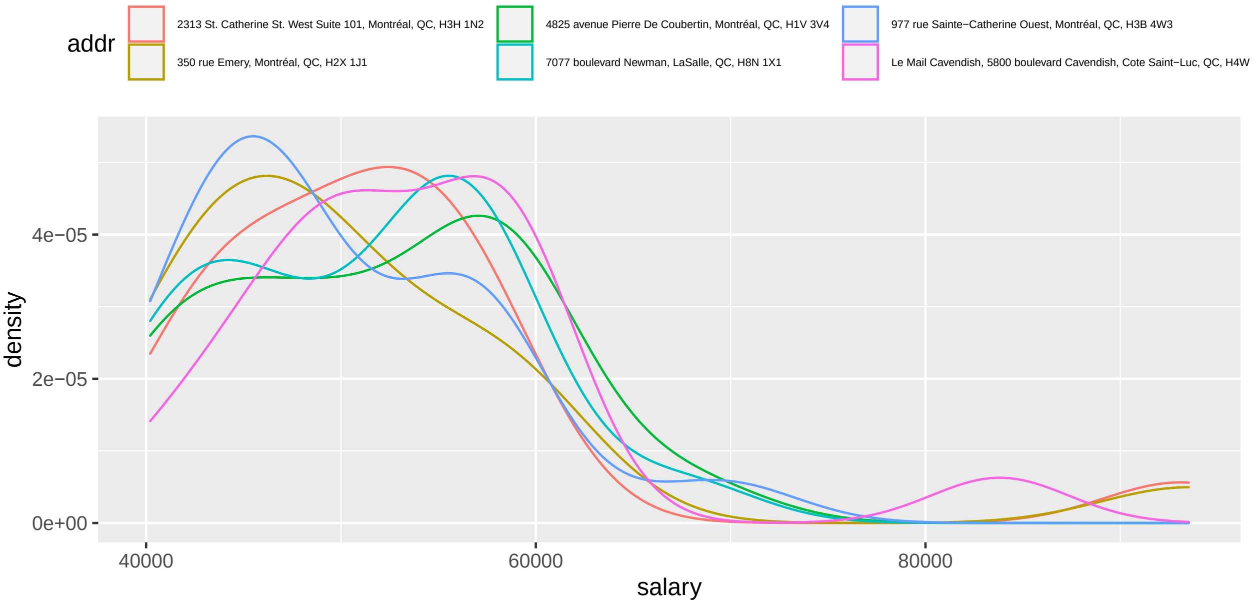
```
## Parsed with column specification:
## cols(
##   eid = col_double(),
##   addr = col_character(),
##   ename = col_character(),
##   pnum = col_double(),
##   salary = col_double()
## )
```

First Chart

For the movie dataset, we will study whether the salary depends on the locations of theatres.

To estimate the distribution of continous random variables, we should use density plots.

```
ggplot(employee,aes(x=salary,col=addr)) +
  geom_density(size=0.5) + theme(legend.position = "top", legend.text = element_text(size = 5))
```

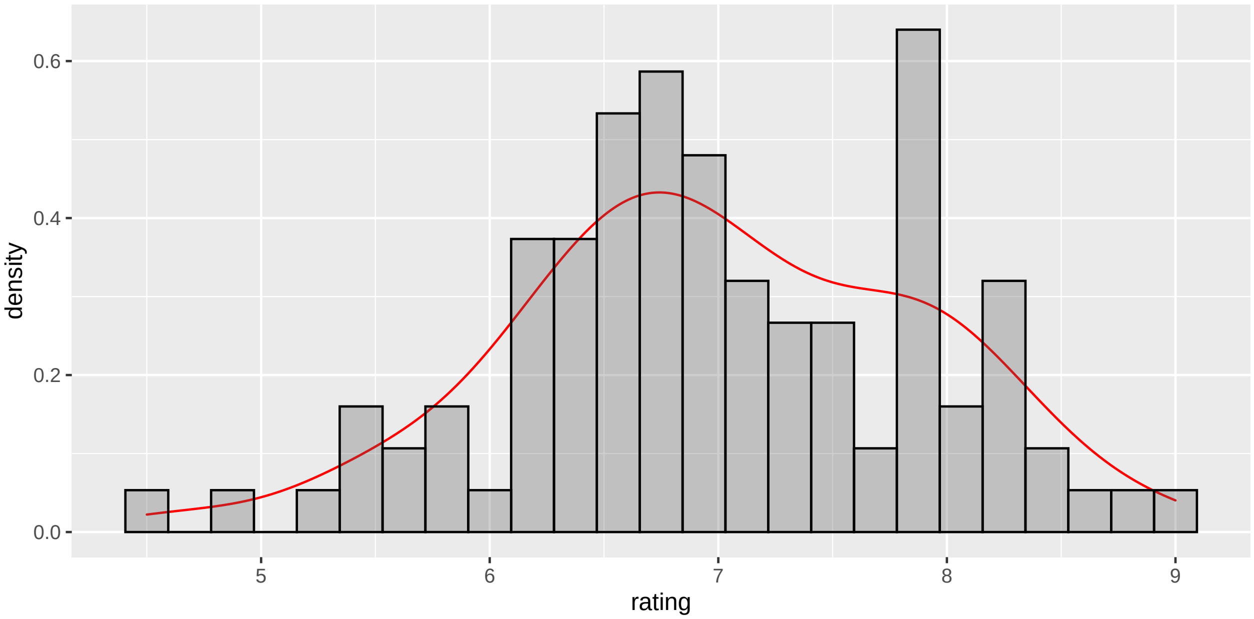


From this plot, we can see the estimated marginal distribution of salary does not vary a lot between different theatres. All the estimated distributions are skewed to the left and have long tails on the right. Thus, there is no significant evidence that the salary of employees depends on the theatres.

Second Chart

For the movie dataset, we will estimate the distribution of ratings.

```
ggplot(movie, aes(x=rating)) + geom_density(col="red")+
geom_histogram(aes(y=..density..),bins=25,col="black",alpha=0.3)
```



The distribution looks symmetrical, so the data might be normally distributed. We could verify this by Shapiro test.

```
shapiro.test(movie$rating)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  movie$rating
## W = 0.99051, p-value = 0.7062
```

From the Shapiro Test, the p-value is large and we fail to reject the null hypothesis and conclude that there is no strong evidence showing the data are not normally distributed.

```
mean(movie$rating)
```

```
## [1] 6.992
```

```
sd(movie$rating)
```

```
## [1] 0.896647
```

The estimated mean and standard deviation are respectively 6.992 and 0.896647.