Assignment 1.3

Saurabh Biswas

## Warning: package 'ggplot2' was built under R version 3.6.2

state\_M2019\_dl <- read\_excel("/Users/saurabhbiswas/Desktop/DSC630\_R/state\_M2019\_dl.xlsx")  
nebraska\_2019 <- subset(state\_M2019\_dl, state\_M2019\_dl$area\_title == "Nebraska")

## 1. Import, Plot, Summarize, and Save Data

### **tot\_emp:**

summary(as.numeric(nebraska\_2019$tot\_emp))

## Warning in summary(as.numeric(nebraska\_2019$tot\_emp)): NAs introduced by  
## coercion

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 30 170 485 4529 1690 982040 30

### **h\_mean:** Mean hourly wage

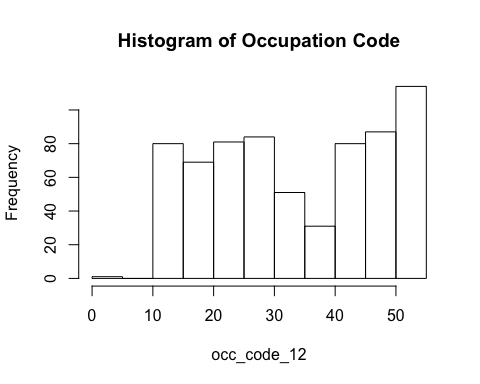
summary(as.numeric(nebraska\_2019$a\_mean))

## Warning in summary(as.numeric(nebraska\_2019$a\_mean)): NAs introduced by coercion

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 22100 36980 48240 56100 67485 262570 15

### **Histogram of occ\_code(1-2)**.

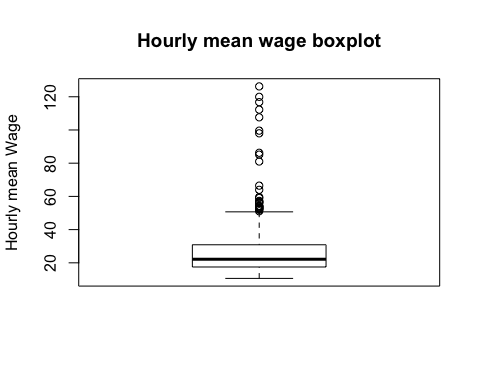
occ\_code\_12 <- substr(nebraska\_2019$occ\_code,1,2)  
occ\_code\_12 <- as.numeric(occ\_code\_12)  
hist(occ\_code\_12, main ="Histogram of Occupation Code")



### **Box plot of hourly mean wage**

boxplot(as.numeric(nebraska\_2019$h\_mean), main='Hourly mean wage boxplot', ylab="Hourly mean Wage")

## Warning in boxplot(as.numeric(nebraska\_2019$h\_mean), main = "Hourly mean wage  
## boxplot", : NAs introduced by coercion



### **Density plot of hourly mean wage**

ggplot(nebraska\_2019,   
 aes(x = as.numeric(nebraska\_2019$h\_mean),   
 fill = substr(nebraska\_2019$occ\_code,1,2))) +  
 geom\_density(alpha = 0.4) +  
 labs(title = 'Salary distribution') +  
 xlab("Hourly Mean Salary")

## Warning in FUN(X[[i]], ...): NAs introduced by coercion  
  
## Warning in FUN(X[[i]], ...): NAs introduced by coercion

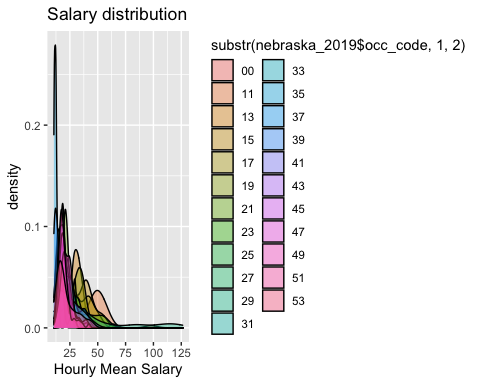
## Warning: Use of `nebraska\_2019$h\_mean` is discouraged. Use `h\_mean` instead.

## Warning: Use of `nebraska\_2019$occ\_code` is discouraged. Use `occ\_code` instead.

## Warning: Removed 65 rows containing non-finite values (stat\_density).

## Warning: Groups with fewer than two data points have been dropped.

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -  
## Inf



### **Density plot of hourly median wage**

ggplot(nebraska\_2019,   
 aes(x = as.numeric(nebraska\_2019$h\_mean),   
 fill = o\_group)) +  
 geom\_density(alpha = 0.4) +  
 labs(title = 'Salary distribution') +  
 xlab("Hourly Median Salary")

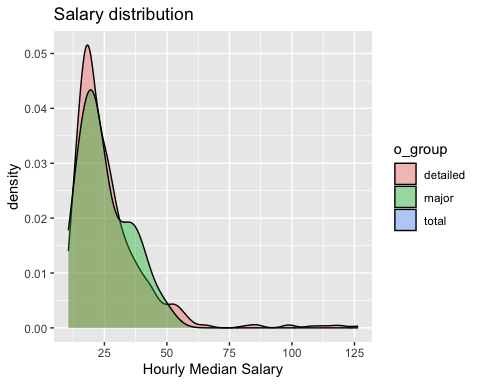
## Warning in FUN(X[[i]], ...): NAs introduced by coercion  
  
## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning: Use of `nebraska\_2019$h\_mean` is discouraged. Use `h\_mean` instead.

## Warning: Removed 65 rows containing non-finite values (stat\_density).

## Warning: Groups with fewer than two data points have been dropped.

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -  
## Inf



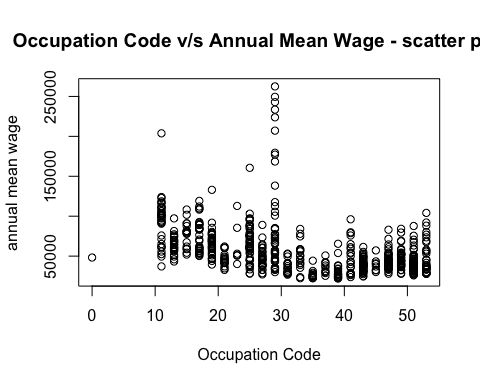
##### **Write into CSV file**

write.csv(nebraska\_2019, "nebraska\_2019.csv")

## 2. Explore Some Bivariate Relations

### **Scatter plot between annual mean wage and occupation code**

plot(as.numeric(substr(nebraska\_2019$occ\_code,1,2)), as.numeric(nebraska\_2019$a\_mean), main="Occupation Code v/s Annual Mean Wage - scatter plot",   
 ylab="annual mean wage", xlab="Occupation Code")

 \*\*\*

## 3. Organize a Data Report

### **Structure and Type of variables**

str(nebraska\_2019)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 678 obs. of 30 variables:  
## $ area : chr "31" "31" "31" "31" ...  
## $ area\_title : chr "Nebraska" "Nebraska" "Nebraska" "Nebraska" ...  
## $ area\_type : chr "2" "2" "2" "2" ...  
## $ naics : chr "000000" "000000" "000000" "000000" ...  
## $ naics\_title : chr "Cross-industry" "Cross-industry" "Cross-industry" "Cross-industry" ...  
## $ i\_group : chr "cross-industry" "cross-industry" "cross-industry" "cross-industry" ...  
## $ own\_code : chr "1235" "1235" "1235" "1235" ...  
## $ occ\_code : chr "00-0000" "11-0000" "11-1011" "11-1021" ...  
## $ occ\_title : chr "All Occupations" "Management Occupations" "Chief Executives" "General and Operations Managers" ...  
## $ o\_group : chr "total" "major" "detailed" "detailed" ...  
## $ tot\_emp : chr "982040" "52180" "1380" "15200" ...  
## $ emp\_prse : chr "0.8" "1.5" "6.1" "2.7" ...  
## $ jobs\_1000 : chr "1000" "53.131999999999998" "1.407" "15.483000000000001" ...  
## $ loc\_quotient: chr "1" "0.97" "1" "0.95" ...  
## $ pct\_total : logi NA NA NA NA NA NA ...  
## $ h\_mean : chr "23.2" "48.46" "97.99" "47.84" ...  
## $ a\_mean : chr "48250" "100800" "203810" "99520" ...  
## $ mean\_prse : chr "0.9" "0.9" "3.2" "1.2" ...  
## $ h\_pct10 : chr "10.33" "20.29" "32.46" "18.920000000000002" ...  
## $ h\_pct25 : chr "13.07" "29.29" "57.4" "27.41" ...  
## $ h\_median : chr "18.46" "42.72" "97.54" "39.92" ...  
## $ h\_pct75 : chr "28.16" "59.7" "#" "59.62" ...  
## $ h\_pct90 : chr "40.450000000000003" "82.46" "#" "88.34" ...  
## $ a\_pct10 : chr "21480" "42210" "67510" "39360" ...  
## $ a\_pct25 : chr "27190" "60930" "119390" "57010" ...  
## $ a\_median : chr "38390" "88850" "202890" "83040" ...  
## $ a\_pct75 : chr "58580" "124170" "#" "124010" ...  
## $ a\_pct90 : chr "84140" "171520" "#" "183740" ...  
## $ annual : chr NA NA NA NA ...  
## $ hourly : chr NA NA NA NA ...

### 1. **occ\_code** is a chr field. First two positions represent occupation types and next four positions repersent sub-types.

### 2. **tot\_emp** is a nueric field. It represents total number of employess for each occupation codes.

### 3. **h\_mean** is a float field. It represents hourly mean wage for each occupation codes.

### 4. **h=a\_mean** is a float field. It represents annual mean wage for each occupation codes.