Homework 11

Li Zhang

11/22/2021

## Create a folder, download the data file “state\_employee\_salary\_data\_2017.csv” into it, and create an Rstudio project in the folder. Read in the data file, peak the structure of your data set and clean it step by step:

# Read in the data file  
raw0 <- read.csv("state\_employee\_salary\_data\_2017.csv")   
  
# Peak the structure of the data set  
str(raw0)

## 'data.frame': 148891 obs. of 38 variables:  
## $ Pyrl.Fiscal.Yr : int 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...  
## $ Calendar.Year : int 2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...  
## $ DeptID : chr "OSC15000" "JUD95000" "JUD95000" "JUD95000" ...  
## $ EmplId.Empl.Rcd : chr "2D051257707CA0EDD16C357ED2A299FC" "923C131DD4B764C83CB34A83D97E3939" "4349244348E773367E4177066F5C3106" "6DA50D291B700281D02EABE4B926EF77" ...  
## $ First.Name : chr "Kevin" "Thomas" "John" "Bethany" ...  
## $ Middle.Initial : chr "" "G" "J" "J" ...  
## $ Last.Name : chr "Lembo" "Moukawsher" "Nazzaro" "Alvord" ...  
## $ Name.Suffix : chr "" "" "" "" ...  
## $ Check.. : int 46943743 46926862 46926866 46926751 46926774 46926802 46943784 46926842 46943785 46943818 ...  
## $ Check.Dt : chr "06/30/2017 12:00:00 AM" "06/30/2017 12:00:00 AM" "06/30/2017 12:00:00 AM" "06/30/2017 12:00:00 AM" ...  
## $ Chk.Option : chr "A" "A" "A" "A" ...  
## $ Chk.Status : chr "F" "F" "F" "F" ...  
## $ Annnual.Rate : num 110000 167634 167634 174323 167634 ...  
## $ Bi.Weekly.Comp.Rate : num 4583 6985 6985 7263 6985 ...  
## $ Other : num 0 0 0 193 492 ...  
## $ Fringe : num 3869 3904 4840 4805 4213 ...  
## $ Overtime : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Salaries...Wages : num 4583 6985 6985 7263 6985 ...  
## $ Fringe.Amt.No.Retire: num 1349 178 1114 930 487 ...  
## $ SERS.Amount : num 2520 0 0 0 0 ...  
## $ ARP.Amount : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Teachers.Amount : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Judges.Amount : num 0 3726 3726 3875 3726 ...  
## $ Tot.Gross : num 4583 6985 6985 7456 7477 ...  
## $ Age : int 53 54 58 60 51 54 51 56 55 66 ...  
## $ Job.Cd.Descr : chr "Comptroller(Elected)" "Judicial(Semimonthly)" "Judicial(Semimonthly)" "Judicial(Semimonthly)" ...  
## $ EE.Class.Descr : chr "Permanent" "Permanent" "Permanent" "Permanent" ...  
## $ Job.Indicator : chr "P" "P" "P" "P" ...  
## $ Ethnic.Grp : chr "WHITE" "" "" "NSPEC" ...  
## $ Sex : chr "M" "M" "M" "F" ...  
## $ Full.Part : chr "F" "F" "F" "F" ...  
## $ Orig.Hire : chr "12/04/1998 12:00:00 AM" "03/06/2013 12:00:00 AM" "08/21/2007 12:00:00 AM" "01/01/1901 12:00:00 AM" ...  
## $ Term.Date : chr "" "" "" "" ...  
## $ City : chr "Hartford" "Hartford" "Hartford" "Hartford" ...  
## $ Postal : chr "06106" "06106" "06106" "06106" ...  
## $ State : chr "CT" "CT" "CT" "CT" ...  
## $ Union.Descr : chr "Exempt/Elected/Appointed" "Judicial - Judges" "Judicial - Judges" "Judicial - Judges" ...  
## $ Agency : chr "State Comptroller's Office" "Judicial Branch" "Judicial Branch" "Judicial Branch" ...

## 10.1.a

Select only columns of “EmplId.Empl.Rcd,” “Bi.Weekly.Comp.Rate,” “Age,” “Ethnic.Grp,” “Sex,” “Full.Part,” “City,” get summary information for all these variables. Make sure “EmplId.Empl.Rcd,” “Sex,” “Ethnic.Grp” to be factors, “Age” and “Bi.Weekly.Comp.Rate” to be numerics.

# Select certain columns from the original dataset.  
raw1 <- raw0[, c("EmplId.Empl.Rcd","Bi.Weekly.Comp.Rate",   
 "Age", "Ethnic.Grp", "Sex", "Full.Part", "City")]  
  
# Set some variables to be factors  
raw1$EmplId.Empl.Rcd <- as.factor(raw1$EmplId.Empl.Rcd)  
raw1$Sex <- as.factor(raw1$Sex)  
raw1$Ethnic.Grp <- as.factor(raw1$Ethnic.Grp)  
  
# Set some variables to be numerics  
raw1$Age <- as.numeric(raw1$Age)  
raw1$Bi.Weekly.Comp.Rate <- as.numeric(raw1$Bi.Weekly.Comp.Rate)  
  
  
#Summary information for all these variables  
summary(raw1)

## EmplId.Empl.Rcd Bi.Weekly.Comp.Rate Age   
## 9B4E4728FAAF512EA0F4A92AB2A865EA: 30 Min. : 0.00 Min. : 16.00   
## 915EA55FFF225A4517BD558C7266BC46: 29 1st Qu.: 52.66 1st Qu.: 34.00   
## EC58C09D5571F33839F872C9FCC735C2: 29 Median : 2214.30 Median : 47.00   
## 14D7D8F93C618CD5151833ECDE1AB192: 28 Mean : 2113.66 Mean : 44.69   
## 280EC65004396342237052D2FC305A51: 28 3rd Qu.: 3133.41 3rd Qu.: 55.00   
## 29B6A5D71DE115377D2E76708159BDB6: 28 Max. :61111.11 Max. :115.00   
## (Other) :148719   
## Ethnic.Grp Sex Full.Part City   
## WHITE :81248 F:77855 Length:148891 Length:148891   
## :23870 M:65525 Class :character Class :character   
## BLACK :18853 U: 5511 Mode :character Mode :character   
## HISPA :10882   
## NSPEC : 9031   
## ASIAN : 4395   
## (Other): 612

str(raw1) # peak the structure of the raw1 data.

## 'data.frame': 148891 obs. of 7 variables:  
## $ EmplId.Empl.Rcd : Factor w/ 8500 levels "00034CF9004E4E0D7872FEB52CB5933F",..: 1490 4819 2215 3597 8038 1668 37 7738 5923 6025 ...  
## $ Bi.Weekly.Comp.Rate: num 4583 6985 6985 7263 6985 ...  
## $ Age : num 53 54 58 60 51 54 51 56 55 66 ...  
## $ Ethnic.Grp : Factor w/ 8 levels "","AMIND","ASIAN",..: 8 1 1 6 1 1 4 1 8 8 ...  
## $ Sex : Factor w/ 3 levels "F","M","U": 2 2 2 1 1 2 2 1 1 2 ...  
## $ Full.Part : chr "F" "F" "F" "F" ...  
## $ City : chr "Hartford" "Hartford" "Hartford" "Hartford" ...

## 10.1.b

Pick up full-time employees in the city “Hartford.”

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# pick full-time employees in the city Hartford by using filter() function.  
raw2 <- filter(raw1, Full.Part == "F" & City == "Hartford")  
head(raw2)

## EmplId.Empl.Rcd Bi.Weekly.Comp.Rate Age Ethnic.Grp Sex  
## 1 2D051257707CA0EDD16C357ED2A299FC 4583.34 53 WHITE M  
## 2 923C131DD4B764C83CB34A83D97E3939 6984.75 54 M  
## 3 4349244348E773367E4177066F5C3106 6984.75 58 M  
## 4 6DA50D291B700281D02EABE4B926EF77 7263.46 60 NSPEC F  
## 5 F2528A6B200A9881A4074431BFDC0B20 6984.75 51 F  
## 6 3284E7090925A24F494E3915CBDFAB6F 6984.75 54 M  
## Full.Part City  
## 1 F Hartford  
## 2 F Hartford  
## 3 F Hartford  
## 4 F Hartford  
## 5 F Hartford  
## 6 F Hartford

## 10.1.c

Select only those employees with 26 paychecks for the 2017 fiscal year(Hint: Each employee has a unique ID of “EmplId.Empl.Rcd”).

# Check the situation of employees.  
EmplId.Empl.Rcd <- as.character((raw2$EmplId.Empl.Rcd))  
counts <- data.frame(table(EmplId.Empl.Rcd))  
head(counts)

## EmplId.Empl.Rcd Freq  
## 1 00C52BE59EFB743D0572D1126F67AAA2 26  
## 2 00C669A2AEA0B28A20159A1129E269A2 26  
## 3 00DC9714C1472D1D01667624B76961E9 26  
## 4 0106896DD5677C57673354A15CEC15B1 26  
## 5 012DE02BA888BD979F4C47B3DFE8A216 1  
## 6 013395C714634DDC82AE9C3CBC3F321F 23

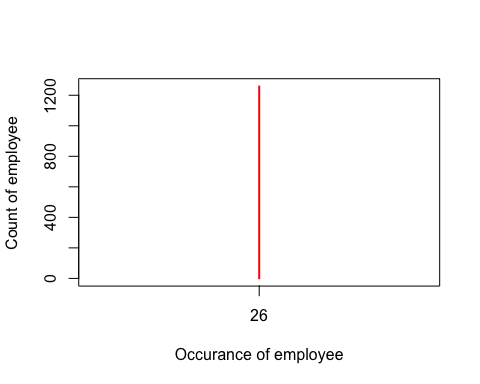
PersonOccurCount <- table(counts[, 2])  
plot(PersonOccurCount, col = rainbow(30),   
 xlab = "Occurance of employee", ylab = "Count of employee")



# The most employee-occurrence is 26.  
  
subEmpl26 <- counts[which(counts$Freq != 26),]  
len <- length(subEmpl26$EmplId.Empl.Rcd)  
for (i in 1: len) {  
 j <- which(raw2$EmplId.Empl.Rcd == as.character(subEmpl26$EmplId.Empl.Rcd)[i])  
 if (i == 1){  
 id = j  
 } else if(i>1){  
 id = c(id, j)  
 }  
}  
  
# Select only those employees with 26 paychecks   
# by removing employees with check-count unequal to 26.  
raw26 <- raw2[-id, ]   
#head(raw26)  
  
# Check the result  
EmplId.Empl.Rcd <- as.character((raw26$EmplId.Empl.Rcd))  
counts <- data.frame(table(EmplId.Empl.Rcd))  
head(counts)

## EmplId.Empl.Rcd Freq  
## 1 00C52BE59EFB743D0572D1126F67AAA2 26  
## 2 00C669A2AEA0B28A20159A1129E269A2 26  
## 3 00DC9714C1472D1D01667624B76961E9 26  
## 4 0106896DD5677C57673354A15CEC15B1 26  
## 5 0137D28FE72FFE6AA90BE4783E7ED50E 26  
## 6 013E7915383ABAC092BB503250FC3BB0 26

PersonOccurCount <- table(counts[, 2])  
plot(PersonOccurCount, col = rainbow(30),   
 xlab = "Occurance of employee", ylab = "Count of employee")



## 10.1.d

Remove sex group other than “F” and “M” if there is any. Hint: Use str() after droplevels().

# Remove sex group other than "F" and "M".  
raw.d <- raw26[raw26$Sex == "F" |raw26$Sex == "M", ] %>% droplevels()  
str(raw.d)

## 'data.frame': 32708 obs. of 7 variables:  
## $ EmplId.Empl.Rcd : Factor w/ 1258 levels "00C52BE59EFB743D0572D1126F67AAA2",..: 575 55 403 468 1248 820 486 1172 907 94 ...  
## $ Bi.Weekly.Comp.Rate: num 2629 3298 2029 1401 3570 ...  
## $ Age : num 27 41 56 32 53 57 56 42 55 29 ...  
## $ Ethnic.Grp : Factor w/ 8 levels "","AMIND","ASIAN",..: 8 1 6 1 4 8 4 1 4 8 ...  
## $ Sex : Factor w/ 2 levels "F","M": 2 1 2 2 1 1 1 1 2 2 ...  
## $ Full.Part : chr "F" "F" "F" "F" ...  
## $ City : chr "Hartford" "Hartford" "Hartford" "Hartford" ...

# Check whether sex "U" employee is still in our data.  
nrow(raw.d[which(raw.d$Sex == "U"),])

## [1] 0

#raw.d

## 10.1.e

Replace level of empty space in “Ethnic.Grp” as “unknown.”

levels(raw.d$Ethnic.Grp)[levels(raw.d$Ethnic.Grp) == ""] <- "unknown"  
head(raw.d)

## EmplId.Empl.Rcd Bi.Weekly.Comp.Rate Age Ethnic.Grp Sex  
## 18 76973A44AAC8C44B71B79CDBD1A27D84 2629.47 27 WHITE M  
## 19 0ABF4E90AFC466BC4CF0C04025200A08 3298.36 41 unknown F  
## 20 52E7C8149375F8AD17A090C66CF976C4 2028.93 56 NSPEC M  
## 21 60923808A13A3FD8A7B3E4B9CAF99369 1400.70 32 unknown M  
## 22 FE823FA6A61339B8522354C661EF9CBE 3569.85 53 BLACK F  
## 23 A68CA35AD5997101DC347BA11C23B7D0 4233.68 57 WHITE F  
## Full.Part City  
## 18 F Hartford  
## 19 F Hartford  
## 20 F Hartford  
## 21 F Hartford  
## 22 F Hartford  
## 23 F Hartford

#summary(raw.d)

## 10.1.f

Remove employees with “Bi.Weekly.Comp.Rate” equal or less than zero if there is any.

empl.f <- raw.d[which(raw.d$Bi.Weekly.Comp.Rate <= 0), ]  
raw.f <- raw.d[-which(raw.d$Bi.Weekly.Comp.Rate == empl.f$Bi.Weekly.Comp.Rate),]  
head(raw.f)

## EmplId.Empl.Rcd Bi.Weekly.Comp.Rate Age Ethnic.Grp Sex  
## 18 76973A44AAC8C44B71B79CDBD1A27D84 2629.47 27 WHITE M  
## 19 0ABF4E90AFC466BC4CF0C04025200A08 3298.36 41 unknown F  
## 20 52E7C8149375F8AD17A090C66CF976C4 2028.93 56 NSPEC M  
## 21 60923808A13A3FD8A7B3E4B9CAF99369 1400.70 32 unknown M  
## 22 FE823FA6A61339B8522354C661EF9CBE 3569.85 53 BLACK F  
## 23 A68CA35AD5997101DC347BA11C23B7D0 4233.68 57 WHITE F  
## Full.Part City  
## 18 F Hartford  
## 19 F Hartford  
## 20 F Hartford  
## 21 F Hartford  
## 22 F Hartford  
## 23 F Hartford

# Check whether employees with “Bi.Weekly.Comp.Rate” equal or less than zero is still in data  
nrow(raw.f[which(raw.f$Bi.Weekly.Comp.Rate <= 0),])

## [1] 0

## 10.1.g

Group your data set by “EmplId.Empl.Rcd,” “Ethnic.Grp,” and “Sex.” For each employee, take the mean of “Age,” and the mean of “Bi.Weekly.Comp.Rate.” Then round the means to the nearest integer and rename them as “Age” and “Bi.Weekly.Rate.”

raw.g <- raw.f %>%  
 group\_by(EmplId.Empl.Rcd, Ethnic.Grp, Sex) %>%  
 summarise(Age = round(mean(Age),0),   
 Bi.Weekly.Rate = round(mean(Bi.Weekly.Comp.Rate),0)) %>%  
 droplevels()

## `summarise()` has grouped output by 'EmplId.Empl.Rcd', 'Ethnic.Grp'. You can override using the `.groups` argument.

head(raw.g)

## # A tibble: 6 x 5  
## # Groups: EmplId.Empl.Rcd, Ethnic.Grp [6]  
## EmplId.Empl.Rcd Ethnic.Grp Sex Age Bi.Weekly.Rate  
## <fct> <fct> <fct> <dbl> <dbl>  
## 1 00C52BE59EFB743D0572D1126F67AAA2 NSPEC M 71 3031  
## 2 00C669A2AEA0B28A20159A1129E269A2 HISPA F 47 3257  
## 3 00DC9714C1472D1D01667624B76961E9 WHITE M 54 1423  
## 4 0106896DD5677C57673354A15CEC15B1 BLACK F 40 3257  
## 5 0137D28FE72FFE6AA90BE4783E7ED50E WHITE F 50 2557  
## 6 013E7915383ABAC092BB503250FC3BB0 NSPEC F 37 3249

## 10.1.h

Show structure and summary of your data set.

str(data.frame(raw.g)) # structure of the data set

## 'data.frame': 1257 obs. of 5 variables:  
## $ EmplId.Empl.Rcd: Factor w/ 1257 levels "00C52BE59EFB743D0572D1126F67AAA2",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Ethnic.Grp : Factor w/ 8 levels "unknown","AMIND",..: 6 5 8 4 8 6 8 8 6 1 ...  
## $ Sex : Factor w/ 2 levels "F","M": 2 1 2 1 1 1 2 1 2 1 ...  
## $ Age : num 71 47 54 40 50 37 41 59 45 45 ...  
## $ Bi.Weekly.Rate : num 3031 3257 1423 3257 2557 ...

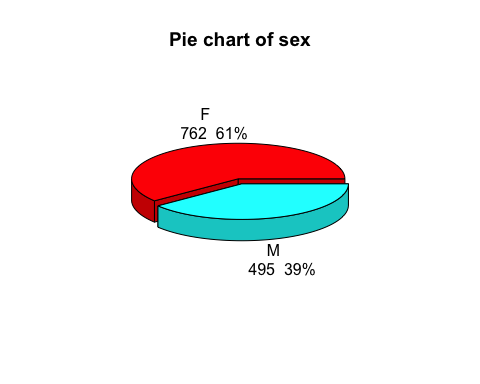
summary(raw.g) # summary of the data set

## EmplId.Empl.Rcd Ethnic.Grp Sex Age   
## 00C52BE59EFB743D0572D1126F67AAA2: 1 WHITE :607 F:762 Min. :22.00   
## 00C669A2AEA0B28A20159A1129E269A2: 1 NSPEC :221 M:495 1st Qu.:41.00   
## 00DC9714C1472D1D01667624B76961E9: 1 BLACK :198 Median :50.00   
## 0106896DD5677C57673354A15CEC15B1: 1 unknown:105 Mean :48.68   
## 0137D28FE72FFE6AA90BE4783E7ED50E: 1 HISPA : 99 3rd Qu.:57.00   
## 013E7915383ABAC092BB503250FC3BB0: 1 ASIAN : 23 Max. :81.00   
## (Other) :1251 (Other): 4   
## Bi.Weekly.Rate  
## Min. : 898   
## 1st Qu.:2372   
## Median :3111   
## Mean :3116   
## 3rd Qu.:3621   
## Max. :8812   
##

## 10.1.j

Using pie plot to show the distribution of “Sex.”

library(plotrix)  
counts <- sort(table(raw.g$Sex), decreasing = TRUE)  
p <- paste(round(counts/sum(counts),2)\*100, "%", sep = "")  
lbls <- paste(names(counts), "\n", counts, p, sep = " ")  
pie3D(counts, labels = lbls, explode = 0.05, main = "Pie chart of sex", labelcex = 1.0,  
 labelpos = c(1.8, 5.0))



There are around 22% more female than males in the city Hartford.

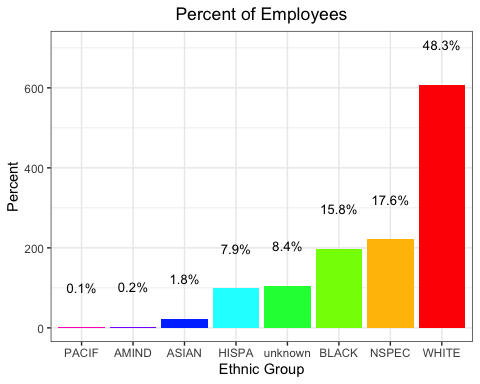
## 10.1.k

List the count number and percent of employees in each ethnic group.

library(ggplot2)  
  
# count number in each ethnic group  
counts.k<-data.frame(sort(table(raw.g$Ethnic.Grp), decreasing = TRUE))  
colnames(counts.k)<-c("Ethnic.Grp", "Count")  
counts.k # print table of counts

## Ethnic.Grp Count  
## 1 WHITE 607  
## 2 NSPEC 221  
## 3 BLACK 198  
## 4 unknown 105  
## 5 HISPA 99  
## 6 ASIAN 23  
## 7 AMIND 3  
## 8 PACIF 1

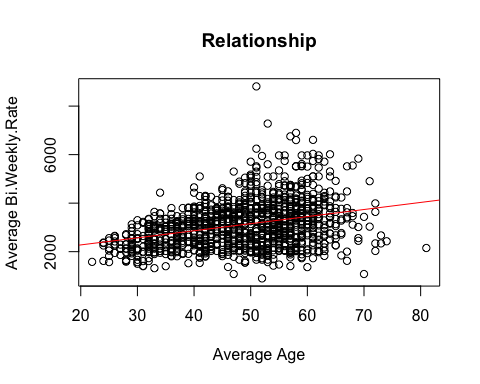
# Percent of employees in eacg ethnic group  
percent.k <-paste(round(counts.k$Count / sum(counts.k$Count),3)\*100, "%", sep = "")  
#percent.k   
ggplot(counts.k, aes(x = reorder(Ethnic.Grp, Count), y = Count)) +  
 geom\_bar(stat = 'identity', fill = rainbow(8)) +   
 geom\_text(aes(x = c(8:1), y = Count + 100 , label = percent.k), size = 3.5) +  
 labs(x = 'Ethnic Group', y = 'Percent', title = 'Percent of Employees') +   
 theme\_bw() +   
 #coord\_flip() + # Flip the plot  
 theme(plot.title = element\_text(hjust = 0.5))



## 10.1.l

Using a scatter plot to view the relationship between average “Bi.Weekly.Rate” and “Age,” add a regression line and give your interpretation.

x <- raw.g$Age  
y <- raw.g$Bi.Weekly.Rate   
  
# create a scatter plot  
plot(x, y, main = "Relationship",   
 xlab = "Average Age",   
 ylab = "Average Bi.Weekly.Rate")  
  
# Add a regression line  
abline(lm(y ~ x, data = raw.g), col = "red")

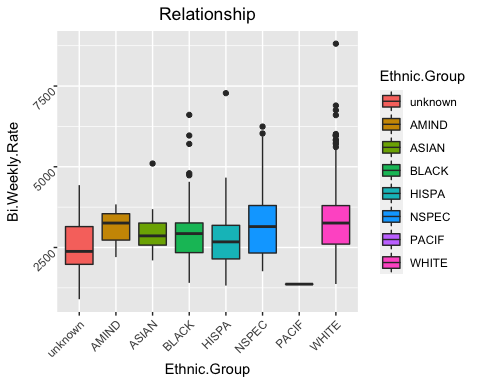


Interpretation: As Average Age increases, the average Bi.Weekly.Rate will increase but very slowly. Average Age from 50 to 60 has relatively higher average Bi.Weekly.Rate.

## 10.1.m

Using box plot to view the relationship between “Bi.Weekly.Rate” by Ethnic groups.

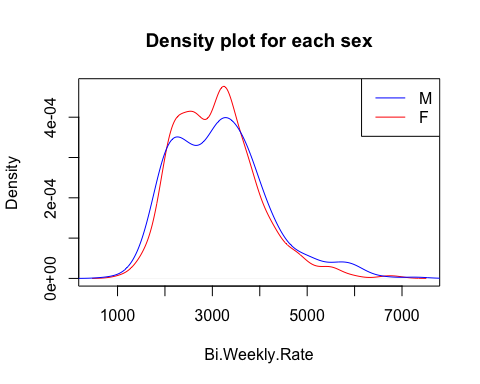
library(ggplot2)  
Ethnic.Group <- raw.g$Ethnic.Grp  
Bi.Weekly.Rate <-raw.g$Bi.Weekly.Rate  
ggplot() +  
 geom\_boxplot(data = raw.g, aes(x = Ethnic.Group, y= Bi.Weekly.Rate, fill = Ethnic.Group))+  
 labs(x= "Ethnic.Group", y = "Bi.Weekly.Rate", title = "Relationship") +  
 theme(plot.title = element\_text(hjust = 0.5))+  
 theme(axis.text = element\_text(angle = 45, vjust = 1, hjust =1))



## 10.1.n

Using a density plot similar to Figure 10.8 to view the distribution of “Bi.Weekly.Rate” for different sex. Package ggplot2 is referred.

library(agricolae)  
#par(mfrow = c(1,2))  
  
# Density plot for each sex  
df1 <- raw.g[which(raw.g$Sex == "F"), ]  
df2 <- raw.g[which(raw.g$Sex == "M"), ]  
plot(density(df1$Bi.Weekly.Rate), col = "red", xlab = "Bi.Weekly.Rate",   
 main = "Density plot for each sex")  
lines(density(df2$Bi.Weekly.Rate), col = "blue")   
# Add density line of df2 to density plot of df1  
  
legend("topright", c("M","F"), lty = c(1, 1), col = c("blue", "red")) # Add legend



# LSD comparison for sex group  
model <- aov(raw.g$Bi.Weekly.Rate~raw.g$Sex)  
res <- LSD.test(raw.g$Bi.Weekly.Rate, raw.g$Sex,   
 DFerror = model$df.residual, MSerror = anova(model)[["Mean Sq"]][2])   
rate <- round(res$groups[, "raw.g$Bi.Weekly.Rate"], 1)  
plot(res, xlab = "Sex Group", ylab = "Range between max and min",   
 main = "Sex groups and variation range")  
text(x = c(seq(from = 1, to = 10, by = 1.2)), y = 3500, rate)

