Use of ggplot & RMarkdown

Eleni Domzaridou

Oct 2018

## Introduction

The xls document of financial data can be found on this [link](https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwi3x__kuaveAhXF6qQKHe8OBggQFjAAegQICRAC&url=http%3A%2F%2Fwww3.wabash.edu%2Feconometrics%2Feconometricsbook%2FChapters%2FCh03PivotTables%2FExcelFiles%2FEastNorthCentralFTWorkers.xls&usg=AOvVaw2rpiXMtxewvLn7R1nvur1l). The data came from full-time workers in East North Central USA from March 1999 CPS. We use the data to work with rmarkdown, ggplot and regression

# Pre-processing

# read the file  
myfd <- read.csv("D:/\_MSc/myMSc/BOOKS & ARTICLES/MATHS & Bio Modelling/examples/findata.csv")  
  
# remove unneeded columns  
myfd <- myfd[,-c(10, 15, 16)]  
  
# change columns' names  
names(myfd) <- c("hours", "eduyears", "earnyear", "race", "sex", "earnweek", "state", "fumonths", "n\_earnweek", "age", "edu", "n\_race", "n\_sex", "n\_earnyear")

Start with data cleaning and general checkings:

calculate the missingness for either earnweek (character) using grep or by counting zeros in weekly earnings variable. Here I show both ways:

sum(myfd$n\_earnweek == 0)

## [1] 4960

length(grep("\\b0\\b", myfd$earnweek))

## [1] 0

# (starts-ends with 0: use "\\b \_pattern\_ \\b")

The ~77% of weekly salary observations is equal to zero and moreover we don’t really care as we can calculate this salary based on annual salary. For annual earnings we have 316 values equal to zero (~0.05%), which means, according to the dataset code file, that these are not be applicable values.

sum(myfd$n\_earnyear == 0)

## [1] 316

note: We can exclude these values from the dataset before creating working hours subcategories:

myfd <- myfd[myfd$hours != 0,]

define categories for working hours: 1: [35-40] 2: [41-50] etc

myfd$hour\_cat <- 0  
myfd$hour\_cat <- ifelse(myfd$hours >= 35 & myfd$hours <= 40, 1, myfd$hour\_cat)  
myfd$hour\_cat <- ifelse(myfd$hours > 40 & myfd$hours <= 50, 2, myfd$hour\_cat)  
myfd$hour\_cat <- ifelse(myfd$hours > 50 & myfd$hours <= 60, 3, myfd$hour\_cat)  
myfd$hour\_cat <- ifelse(myfd$hours > 60, 4, myfd$hour\_cat)  
myfd$hour\_cat <- as.factor(myfd$hour\_cat)  
table(myfd$hour\_cat)

##   
## 1 2 3 4   
## 3827 1384 564 252

# devide the wage with 1,000 to produce better visualisations:  
myfd$wage\_in\_thous <- myfd$n\_earnyear/1000  
  
# check the variable age:  
summary(myfd$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 31.00 40.00 39.97 48.00 90.00

# define 4 age categories:  
myfd$age\_cat <- 0  
myfd$age\_cat <- ifelse(myfd$age < 25, 1, myfd$age\_cat)  
myfd$age\_cat <- ifelse(myfd$age >= 25 & myfd$age < 35, 2, myfd$age\_cat)  
myfd$age\_cat <- ifelse(myfd$age >= 35 & myfd$age < 45, 3, myfd$age\_cat)  
myfd$age\_cat <- ifelse(myfd$age >= 45 & myfd$age < 55, 4, myfd$age\_cat)  
myfd$age\_cat <- ifelse(myfd$age >= 55, 5, myfd$age\_cat)  
myfd$age\_cat <- as.factor(myfd$age\_cat)  
table(myfd$age\_cat)

##   
## 1 2 3 4 5   
## 548 1502 1823 1523 631

edu\_vector <- c("Less than 1st grade", "1st,2nd,3rd,or 4th grade", "5th or 6th grade", "7th and 8th grade",  
 "9th grade", "10th grade", "11th grade", "12th grade no diploma", "High school graduate-high school diploma",  
 "Some college but no degree", "Assc degree-occupation/vocation", "Assc degree-academic program",  
 "Bachelor's degree (BA,AB,BS)", "Master's degree (MA,MS,MENG,MED,MSW,MBA)",  
 "Professional school degree (MD,DDS,DVM)", "Doctorate degree (PHD,EDD)")  
  
  
# #check the vector with the definition  
# cbind(edu\_vector, 31:46)  
#   
# myfd$edu\_cat <- 0  
#   
# for(k in 1:length(myfd$edu)){  
# for(i in 31:46){  
# myfd$edu\_cat <- ifelse(myfd$edu == i, edu\_vector[i-30], myfd$edu\_cat)  
# }  
# }  
#   
# myfd$edu\_cat <- as.factor(myfd$edu\_cat)  
#   
# # final check  
# table(myfd$edu\_cat, myfd$edu)

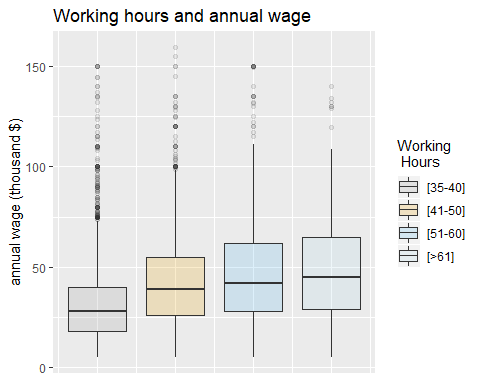
We excluded 281 people with annual wage smaller than 100 we can take more comprehensive boxplots that represent better the variability between races and genders, without sacrificing power. We also keep only those who earn more than 4500$ annualy:

#temp <- myfd[myfd$wage\_in\_thous<200 & myfd$n\_earnyear > 100,]  
  
#temp <- temp[temp$n\_earnyear>4500,]   
  
# save this new dataset:  
#write.csv(temp, file = "mydata.csv")  
  
temp <- read.csv("D:/\_MSc/myMSc/BOOKS & ARTICLES/MATHS & Bio Modelling/examples/mydata.csv")

## Create some initial visualisations

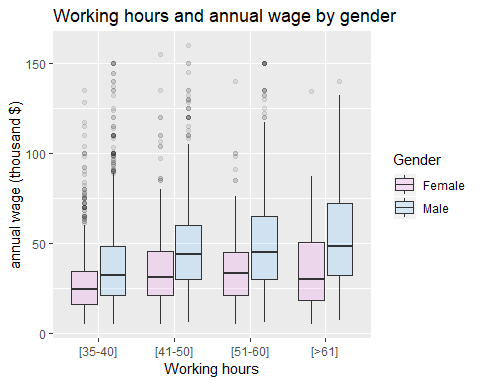
The following boxplot shows ……..

library(ggplot2)  
  
  
p <- ggplot(data = temp, aes(x=hour\_cat, y=wage\_in\_thous, fill=factor(hour\_cat))) +   
 geom\_boxplot(outlier.alpha = 0.1, alpha=0.2) +   
 theme(axis.title.x=element\_blank(),  
 axis.text.x=element\_blank(),  
 axis.ticks.x=element\_blank())  
  
p <- p + scale\_fill\_manual(values=c("#999999", "#E69F00", "#56B4E9", "lightblue", "pink"),   
 name="Working \n Hours",  
 breaks=c("1", "2", "3", "4"),  
 labels=c("[35-40]", "[41-50]", "[51-60]", "[>61]"))  
  
p <- p + labs(  
 title = "Working hours and annual wage",  
 x = "Working hours",  
 y = "annual wage (thousand $)")  
  
p



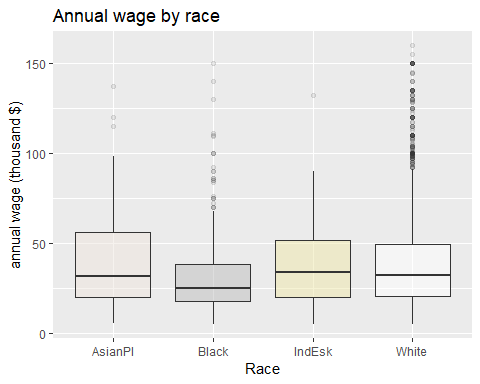
Here we can observe that women are better paid than men:

p1 <- ggplot(data = temp, aes(x=factor(hour\_cat), y=wage\_in\_thous, fill=sex)) +  
 geom\_boxplot(show.legend = TRUE, outlier.alpha = 0.1, alpha=0.2) +  
 scale\_x\_discrete(labels = c("1"="[35-40]", "2"="[41-50]", "3"="[51-60]", "4"="[>61]"))  
  
p1 <- p1 + scale\_fill\_manual(values=rep(c("violet", "steelblue1"), 4),   
 name="Gender")  
  
p1 <- p1 + labs(  
 title = "Working hours and annual wage by gender",  
 x = "Working hours",  
 y = "annual wage (thousand $)")  
  
p1

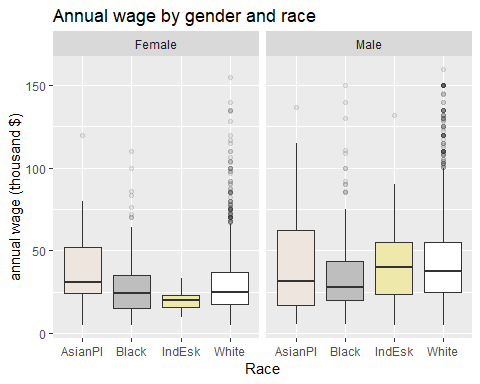


The boxplot that follows shows …… :

p2 <- ggplot(data = temp, aes(x=race, y=wage\_in\_thous, fill=race)) +  
 geom\_boxplot(show.legend = FALSE, outlier.alpha = 0.1, alpha=0.5)  
  
p2 <- p2 + scale\_fill\_manual(values=rep(c("seashell2", "grey", "palegoldenrod", "white"), 4))  
  
p2 <- p2 + labs(  
 title = "Annual wage by race",  
 x = "Race",  
 y = "annual wage (thousand $)")  
  
p2

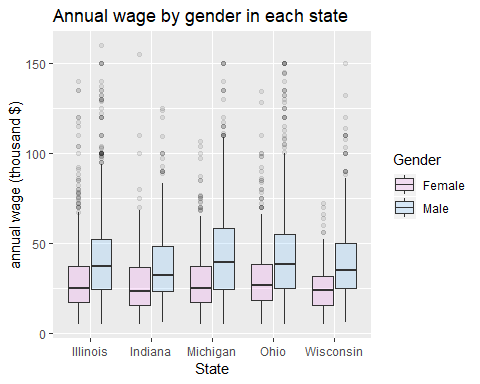


p3 <- ggplot(data = temp, aes(x=race, y=wage\_in\_thous, fill=factor(race))) +  
 geom\_boxplot(show.legend = FALSE, outlier.alpha = 0.1) +  
 facet\_wrap(~sex)  
  
p3 <- p3 + scale\_fill\_manual(values=c("seashell2", "grey", "palegoldenrod", "white"))  
  
p3 <- p3 + labs(  
 title = "Annual wage by gender and race",  
 x = "Race",  
 y = "annual wage (thousand $)")  
  
p3

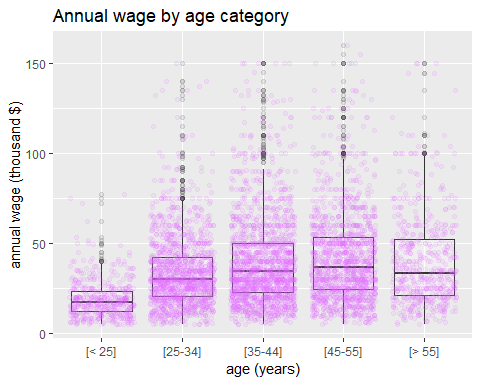


….

p5 <- ggplot(data = temp, aes(x=state, y=wage\_in\_thous, fill=sex)) +  
 geom\_boxplot(show.legend = TRUE, outlier.alpha = 0.1, alpha=0.2)  
  
p5 <- p5 + scale\_fill\_manual(values=rep(c("violet", "steelblue1"), 4),   
 name="Gender")  
  
p5 <- p5 + labs(  
 title = "Annual wage by gender in each state",  
 x = "State",  
 y = "annual wage (thousand $)")  
  
p5

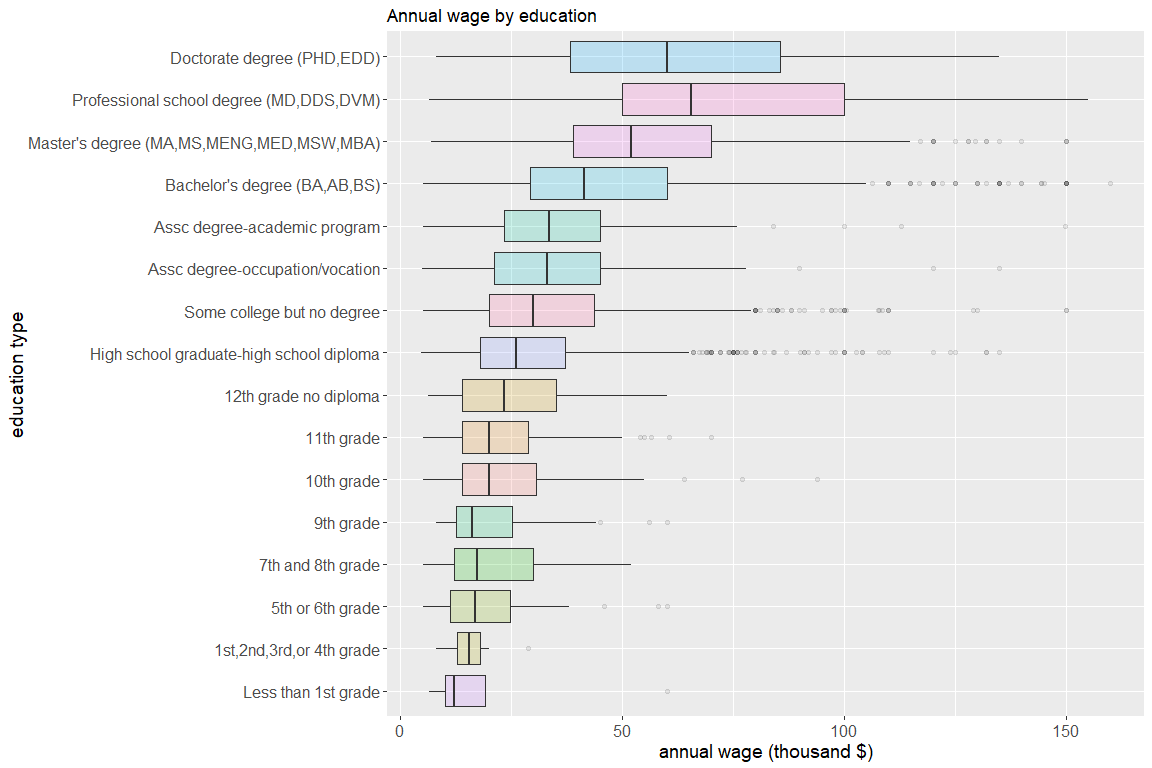


p6 <- ggplot(data = temp, aes(x=factor(age\_cat), y=wage\_in\_thous)) +  
 geom\_boxplot(show.legend = FALSE, outlier.alpha = 0.1, alpha=0.2) +  
 geom\_jitter(alpha=0.1, color="mediumorchid1") +  
 scale\_x\_discrete(labels = c("1"="[< 25]", "2"="[25-34]", "3"="[35-44]", "4"="[45-55]", "5"="[> 55]"))  
   
p6 <- p6 + labs(  
 title = "Annual wage by age category",  
 x = "age (years)",  
 y = "annual wage (thousand $)")  
  
p6

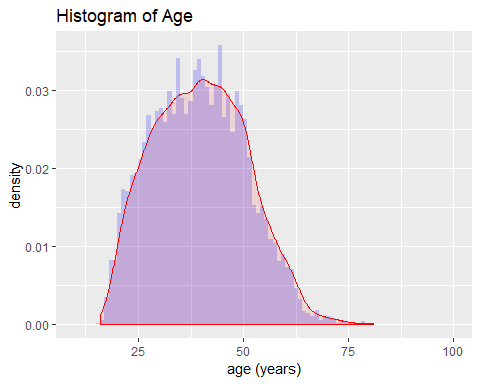


Create a horizontal boxplot to show the variability between education and annual earnings:

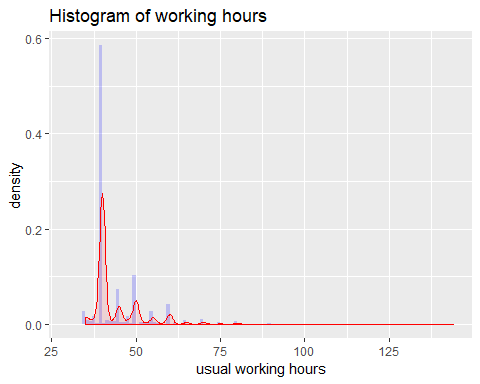
p7 <- ggplot(data = temp, aes(x=edu\_cat, y=wage\_in\_thous, fill=factor(edu\_cat))) +  
 geom\_boxplot(show.legend = FALSE, outlier.alpha = 0.1, alpha=0.2) +  
 #geom\_jitter(alpha=0.1, color="mediumorchid1", show.legend = NA) +  
 scale\_x\_discrete(limits=edu\_vector[1:16]) +  
 theme(axis.text=element\_text(size=12) ,axis.title=element\_text(size=14)) +  
 coord\_flip()  
   
p7 <- p7 + labs(  
 title = "Annual wage by education",  
 x = "education type",  
 y = "annual wage (thousand $)")  
  
p7



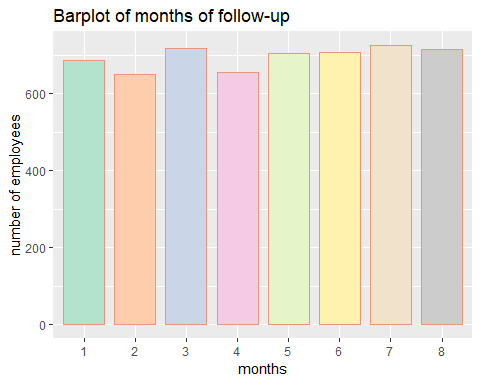
p8 <- ggplot(data=temp, aes(age)) +   
 geom\_histogram(aes(y = ..density..),   
 breaks=seq(10, 100, by = 1),   
 #col="purple",   
 fill="blue",   
 alpha = .2) +   
 geom\_density(col = 2, fill = "red", alpha = 0.1) +   
 labs(title = "Histogram of Age") +  
 labs(x = "age (years)", y = "density")  
  
p8



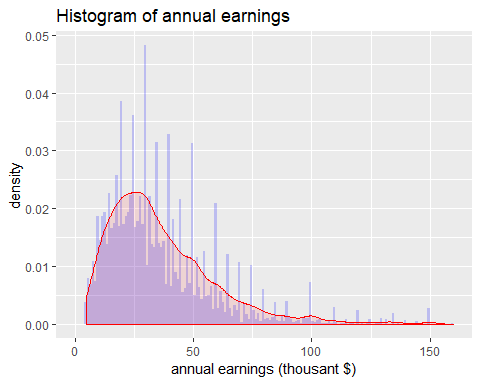
p9 <- ggplot(data=temp, aes(hours)) +   
 geom\_histogram(aes(y = ..density..),   
 breaks=seq(30, 144, by = 1),   
 fill="blue",   
 alpha = .2) +   
 geom\_density(col = 2, fill = "red", alpha = 0.1) +   
 labs(title = "Histogram of working hours") +  
 labs(x = "usual working hours", y = "density")  
  
p9



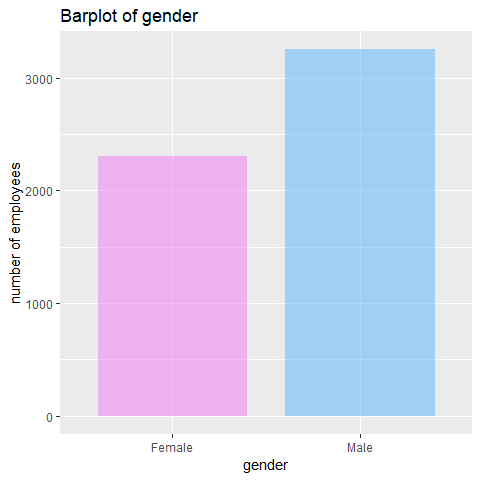
p9.1 <- ggplot(data=temp, aes(x=factor(fumonths), fill=factor(fumonths))) +   
 geom\_bar(stat = "count", show.legend= FALSE, width = 0.8, colour="darksalmon") +   
 scale\_fill\_brewer(palette = "Pastel2") +  
 labs(title = "Barplot of months of follow-up") +  
 labs(x = "months", y = "number of employees")  
  
p9.1

 ###### Note 1: If you do not convert fmonths into a factor, you cannot use different colors in the scale\_fill\_brewer ###### Note 2: colour=“darksalmon” inside geom\_bar defines the bars’ border colour as darksalmon

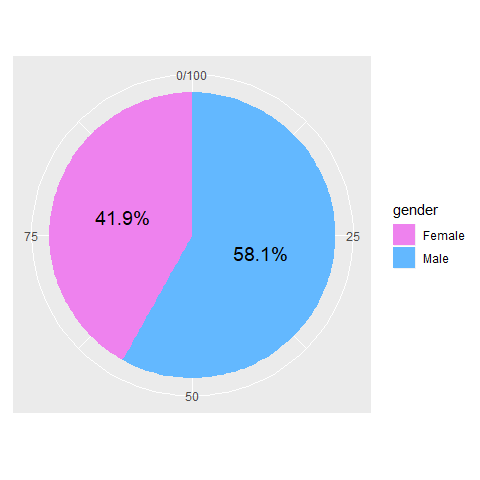
p10 <- ggplot(data=temp, aes(wage\_in\_thous)) +   
 geom\_histogram(aes(y = ..density..),   
 breaks=seq(0, 160, by = 1),   
 fill="blue",   
 alpha = .2) +   
 geom\_density(col = 2, fill = "red", alpha = 0.1) +   
 labs(title = "Histogram of annual earnings") +  
 labs(x = "annual earnings (thousant $)", y = "density")  
  
p10



p11 <- ggplot(data=temp, aes(x=sex, fill=sex, alpha=0.2)) +   
 geom\_bar(stat = "count", show.legend= FALSE, width = 0.8) +   
 scale\_fill\_manual(values=c("violet", "steelblue1")) +  
 labs(title = "Barplot of gender") +  
 labs(x = "gender", y = "number of employees")  
  
p11

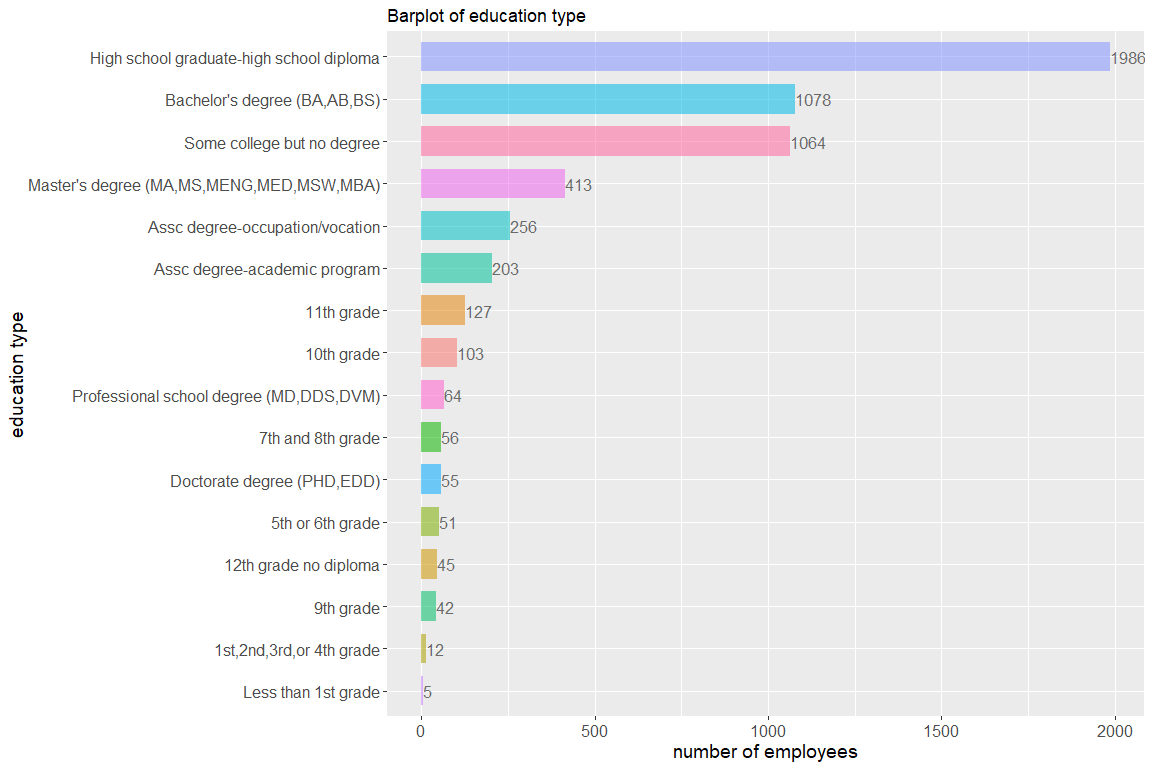
 Alternatively we can also use a pie:

p11df <- data.frame(  
 gender = c("Male", "Female"),  
 percent = c(58.1, 41.9))  
  
  
p11.2 <- ggplot(data=p11df, aes(x="", y=percent, fill=gender)) +   
 geom\_bar(stat = "identity") +  
 scale\_fill\_manual(values=c("violet", "steelblue1")) +  
 coord\_polar("y", start=0)  
  
blank\_theme <- #theme\_minimal()+  
 theme(  
 axis.title.x = element\_blank(),  
 axis.title.y = element\_blank(),  
 #panel.border = element\_blank(),  
 #panel.grid=element\_blank(),  
 axis.ticks = element\_blank(),  
 plot.title=element\_text(size=12, face="bold"))  
  
pie <- p11.2 + blank\_theme + geom\_text(aes(y = p11df$percent),   
 label = paste0(p11df$percent, sep="%"),   
 size=5,   
 position = position\_stack(vjust = 0.5))  
pie



Here, we can define the p12df data.frame that keeps the count of edu\_cat, otherwhise we will not be able to sort the barplot according to count:

p12df <- dplyr::count(temp, edu\_cat)  
   
p12 <- ggplot(data=p12df, aes(x=reorder(edu\_cat, n), y=n, fill=edu\_cat, alpha=0.2)) +   
 geom\_bar(stat = "identity", show.legend= FALSE, width = 0.7) +  
 geom\_text(aes(label=n, size=3), show.legend = FALSE, hjust=0) +   
 labs(title = "Barplot of education type") +  
 labs(x = "education type", y = "number of employees") +  
 theme(axis.text=element\_text(size=12) ,axis.title=element\_text(size=14)) +  
 coord\_flip()  
  
p12



Use of tableone package to produce summary statistics stratified by gender:

library(tableone)  
  
#dput(names(temp))  
  
myvars <- c("hours", "eduyears", "race", "state", "fumonths",   
 "n\_earnweek", "age", "edu", "n\_earnyear", "hour\_cat",   
 "wage\_in\_thous", "age\_cat", "edu\_cat")  
  
mytableOne <- CreateTableOne(vars = myvars, strata = c("sex"), data = temp)  
#mytableOne