yimin zhang

February 2, 2016

```
# setup and load library
setwd("~/Desktop/studying/w4249 applied data models")
library("plyr")
library("dplyr")
```

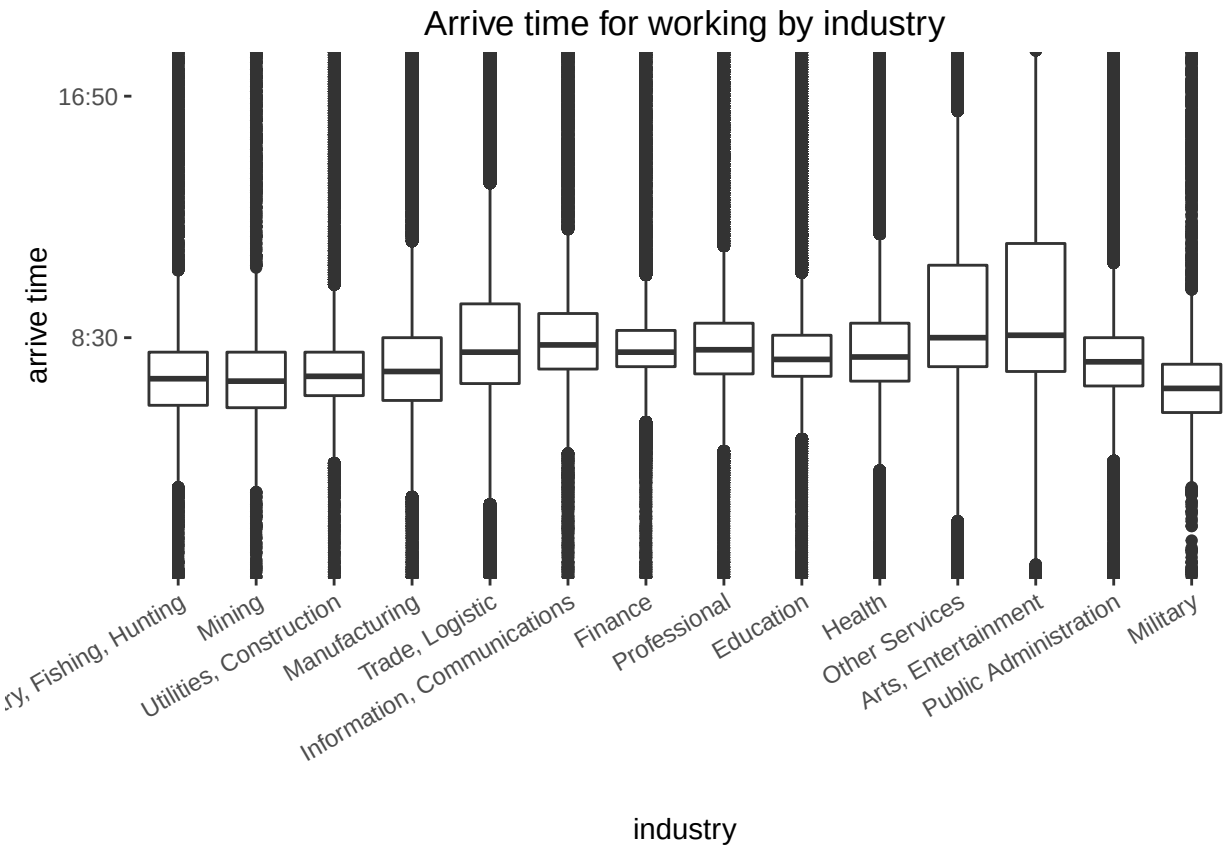
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

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

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

```
library("data.table")
```

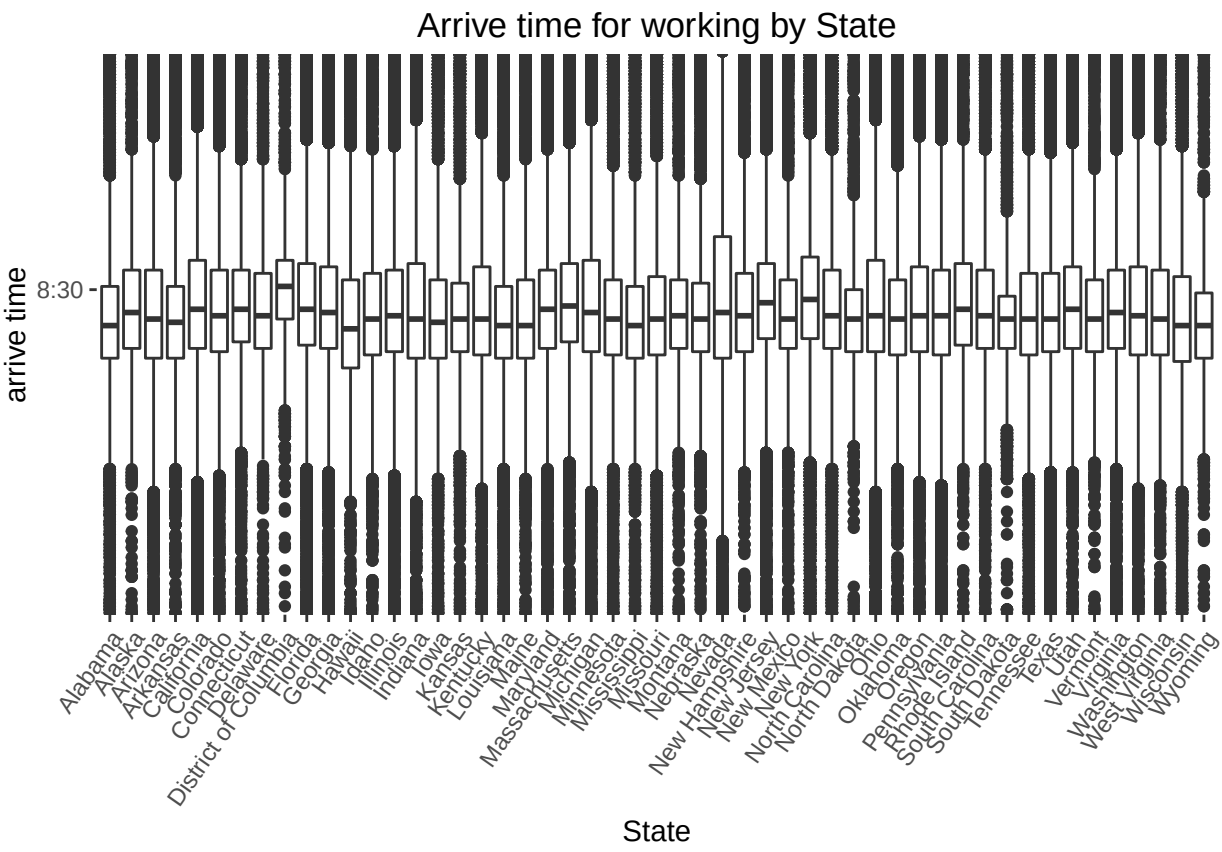
```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, last
```

```
library("ggplot2")

# read in data and save as RData
reread <- 1
if ( reread == 0 ){
  colsToKeep <- c("JWAP", "PERNP","INDP", "ST" )
  # JWAP- arriving time, PERNP-total income, INDP-industry, ST-state
  popDataA <- fread("2013-american-community-survey/pums/ss13pusa.csv", select=colsToKeep )
  popDataB <- fread("2013-american-community-survey/pums/ss13pusb.csv", select=colsToKeep )
  populData <- rbind(popDataA, popDataB)
  rm(popDataA, popDataB)

  # combine the industry with same 3 letter code
  populData$INDP <- ifelse(populData$INDP>=170 & populData$INDP<=290, 170, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=370 & populData$INDP<=490, 370, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=570 & populData$INDP<=770, 570, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=1070 & populData$INDP<=3990, 1070, populData$INDP)
```

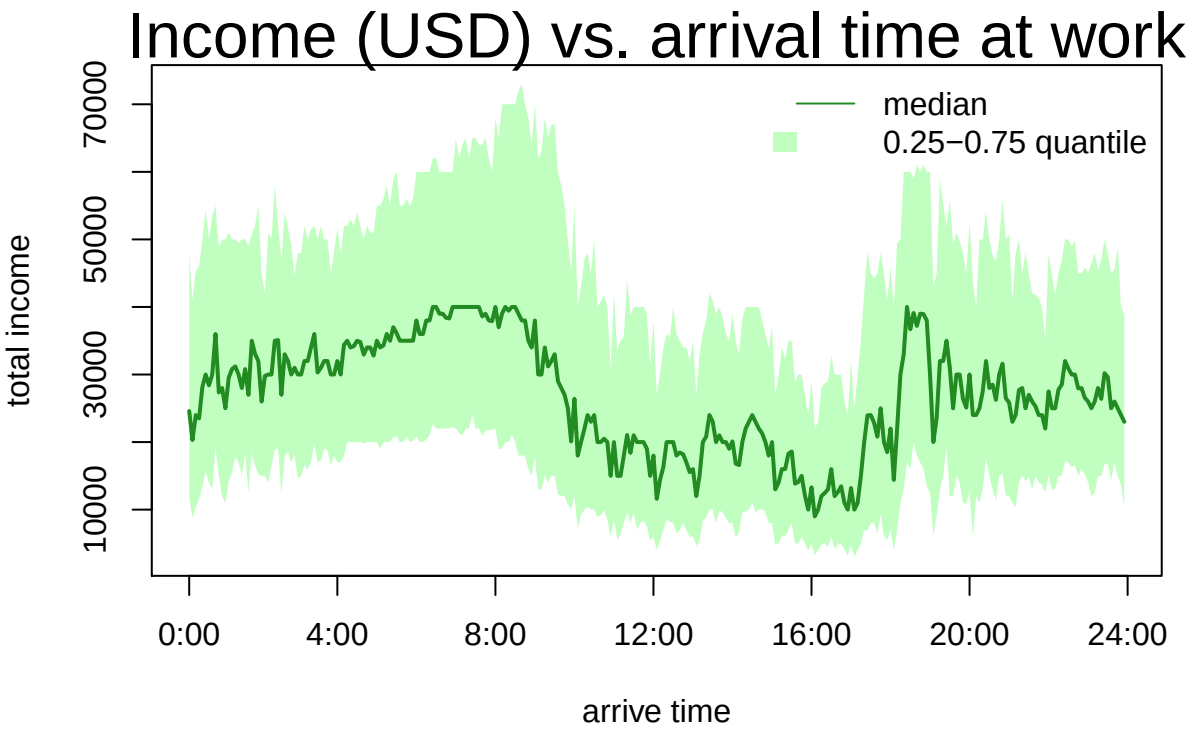
```
  populData$INDP <- ifelse(populData$INDP>=4070 & populData$INDP<=6390, 4070, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=6470 & populData$INDP<=6780, 6470, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=6870 & populData$INDP<=7190, 6870, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=7270 & populData$INDP<=7790, 7270, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=7860 & populData$INDP<=7890, 7860, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=7970 & populData$INDP<=8290, 7970, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=8370 & populData$INDP<=8470, 8370, populData$INDP)
  populData$INDP <- ifelse(populData$INDP %in% c(8660, 8680, 8690), 8370, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>= 8770& populData$INDP <= 9290, 8370, populData$INDP)
  populData$INDP <- ifelse(populData$INDP %in% c(8560,8570,8580,8590,8670), 8560, populData$INDP)
  populData$INDP <- ifelse(populData$INDP>=9370& populData$INDP <= 9590, 9370, populData$INDP)
  populData$INDP <- ifelse(populData$INDP >= 9670& populData$INDP <= 9870, 9670, populData$INDP)
  populData$INDP <- ifelse(populData$INDP >= 9920, 9920, populData$INDP)
  populData$INDP <- factor(populData$INDP)
  levels(populData$INDP) <- c("Agriculture, Forestry, Fishing, Hunting", "Mining",
                              "Utilities, Construction", "Manufacturing", "Trade, Logistic",
                              "Information, Communications", "Finance", "Professional",
                              "Education", "Health", "Other Services", "Arts, Entertainment",
                              "Public Administration", "Military", "Unemployed")

  populData$ST <- as.factor(populData$ST)
  levels(populData$ST) <- c("Alabama", "Alaska", "Arizona", "Arkansas", "California",
       "Colorado", "Connecticut", "Delaware", "District of Columbia", "Florida", "Georgia",
       "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa", "Kansas", "Kentucky", "Louisiana",
       "Maine", "Maryland", "Massachusetts", "Michigan", "Minnesota", "Mississippi", "Missouri",
       "Montana", "Nebraska", "Nevada", "New Hampshire", "New Jersey", "New Mexico", "New York",
       "North Carolina", "North Dakota", "Ohio", "Oklahoma", "Oregon", "Pennsylvania",
       "Rhode Island", "South Carolina", "South Dakota", "Tennessee", "Texas", "Utah", "Vermont",
       "Virginia", "Washington", "West Virginia", "Wisconsin", "Wyoming", "Puerto Rico")

  populData = na.omit(populData)
  save(populData, file="populData.RData")
}else{
  load("populData.RData")
}


# arrive time with industry boxplot
ggplot(populData, aes(x= populData$INDP , y=populData$JWAP, fill= populData$JWAP)) +
  geom_boxplot() +
  scale_y_discrete(breaks=c("0", "100", "200"), labels=c("0:00", "8:30", "16:50")) +
  ggtitle("Arrive time for working by industry") +
  xlab("industry") +
  theme(axis.text.x = element_text(angle = 30, hjust = 1),
        panel.background = element_rect(fill = 'white' )) +
  ylab("arrive time")
```




```
# arrive time by state boxplot
ggplot(populData, aes(x= populData$ST , y=populData$JWAP, fill= populData$JWAP)) +
  geom_boxplot() +
  scale_y_discrete(breaks=c("0", "100", "200"), labels=c("0:00", "8:30", "16:50")) +
  ggtitle("Arrive time for working by State") +
  xlab("State") +
  theme(axis.text.x = element_text(angle = 55, hjust = 1),
        panel.background = element_rect(fill = 'white' )) +
  ylab("arrive time")
```




```r
# arrive time plot(0.25,0.5 and 0.75 quantile ) v.s. income
findQ <- function(x) quantile(x, probs = c(0.25, 0.5, 0.75))
quarts <- ddply(populData, .(JWAP), summarize,
                q1 = findQ(PERNP)[1], q2 = findQ(PERNP)[2], q3 = findQ(PERNP)[3])
# polygon function
drawPoly <- function(x, y1, y2) {
  polygon(c(x, rev(x)), c(y1,  rev(y2)), border = NA, col = "darkseagreen1")}
# plot
with(quarts, {
  plot(q2 ~ JWAP, ylim = range(c(q1, q3)), type = "n",
       xaxt = "n",yaxt = "n",xlab = "arrive time",ylab = "total income")
  drawPoly(JWAP, q1, q3)
  lines(q2 ~ JWAP, col = "forestgreen", lwd = 2)
})
axis(2)
axis(1, c(1,seq(46,286,48)), paste0(seq(0, 24, by = 4), ":00"),col.axis="black")
mtext("Income (USD) vs. arrival time at work", cex = 2)
legend("topright", c("median", "0.25-0.75 quantile"),
                 fill = c(NA, "darkseagreen1"), lty = c(1, NA), border = NA,
                 col = c("forestgreen", NA), text.col = "black", bty = "n")
```


Income (USD) vs. arrival time at work
median
0.25−0.75 quantile
total income
10000
30000
50000
70000
0:00
4:00
8:00
12:00
16:00
20:00
24:00
arrive time