Money, Money, Money!

Ziyue Jin, Hexiu Ye, Yimin Zhang, Juan Campos

February 2, 2016

Ziyue Jin

First Part

```
library(data.table)
library(ggplot2)
library(dplyr)
library(choroplethr)
library(choroplethrMaps)

setwd("~/Documents/cycle1-2/data")
control <- 1

if(control == 0) {
  # Select variables: state, income, occupation, race, language spoken at home
  var <- c("ST", "WAGP","WKL", "OCCP", "INDP")
  dA <- fread("~/Documents/cycle1-2/data/csv_pus/ss13pusa.csv", select = var)
  dB <- fread("~/Documents/cycle1-2/data/csv_pus/ss13pusb.csv", select = var)
  data <- rbind(dA, dB)
  setwd("~/Documents/cycle1-2/data")
  save(data, file="4249Data.RData")
  rm(dA, dB)
} else {
  load(file="4249Data.RData")
}

# Add a new column to test whether a person is computer related job or not
computerCode <- c(0110, 1005, 1006, 1010, 1050, 1105, 1106, 1107, 1400, 5800,
                  5900, 7010, 7900)
infoCode <- c(6470 ,6480 ,6490 ,6570 ,6590 ,6670 ,6672 ,6680 ,6690 ,6695 ,6770 ,6780)

finanCode <- c(0120, 0800, 0810, 0820, 0830, 0840, 0850, 0860, 0900, 0910, 0930, 0940, 0950)
indFCode <- c(6870, 6880, 6890, 6970, 6990, 7070, 7080, 7170, 7180, 7190)

techOrNot <- function(x, y) {
  ifelse(x %in% computerCode | y %in% infoCode, 1, 0)
}
finOrNot <- function(x, y) {
  ifelse(x %in% finanCode | y %in% indFCode, 1, 0)
}

techLabel <- mutate(data, computerCheck = techOrNot(OCCP, INDP))
finLabel <- mutate(data, finanCheck = finOrNot(OCCP, INDP))

# Separate data by their state
gfinLabel <- group_by(finLabel, ST)
```

```r
gtechLabel <- group_by(techLabel, ST)
# Calculate total number of computer related workers
techByState <- summarize(gtechLabel, value=sum(computerCheck, na.rm = T))
# Total number of financial related workers
finByState <- summarize(gfinLabel, value=sum(finanCheck, na.rm = T))
# Get the proportion of related workers in that state
total <- sum(techByState$value)
ftotal <- sum(finByState$value)
techByState <- mutate(techByState, value = value/total)
finByState <- mutate(finByState, value = value/ftotal)
# Transfer state code to state name
state_list =list("1"="alabama", "2"="alaska", "4"="arizona", "5" = "arkansas",
                 "6" = "california", "8" = "colorado", "9" = "connecticut",
                 "10" = "delaware", "11" = "district of columbia", "12" = "florida",
                 "13" = "georgia", "15" = "hawaii",  "16" = "idaho",
                 "17" = "illinois", "18" = "indiana", "19" = "iowa", "20" = "kansas",
                 "21" = "kentucky", "22" = "louisiana",  "23" = "maine",
                 "24" = "maryland", "25" = "massachusetts", "26" = "michigan",
                 "27" = "minnesota",  "28" = "mississippi", "29" = "missouri",
                 "30" = "montana", "31" = "nebraska", "32" = "nevada",
                 "33" = "new hampshire", "34" = "new jersey", "35" = "new mexico",
                 "36" = "new york", "37" = "north carolina", "38" = "north dakota",
                 "39" = "ohio",    "40" = "oklahoma", "41" = "oregon",  "42" = "pennsylvania",
                 "44" = "rhode island", "45" = "south carolina", "46" = "south dakota",
                 "47" = "tennessee",  "48" = "texas",   "49" = "utah",    "50" = "vermont",
                 "51" = "virginia", "53" = "washington", "54" = "west virginia", "55" = "wisconsin",
                 "56" = "wyoming", "72" = "puerto rico")
regionTransfer <- function(x){return ( state_list[[as.character(x)]]) }
plotData <- techByState %>% mutate(region = vapply(ST, regionTransfer, "")) %>%
  select(region, value)
plotfData <- finByState %>% mutate(region = vapply(ST, regionTransfer, "")) %>%
  select(region, value)
state_choropleth(plotData, title = "Fraction of Technology Employment among States",
                 legend = "Fraction", num_colors = 9)
```

```r
state_choropleth(plotfData, title = "Fraction of Financial Employment among States",
                 legend = "Fraction", num_colors = 9)
```

```r
plotData[which.max(plotData$value),]$region
```

```
## [1] "california"
```

```r
plotfData[which.max(plotfData$value),]$region
```

```
## [1] "california"
```

Second Part

```r
finLabel$finanCheck <- finLabel$finanCheck*2
combine <- mutate(techLabel, check=computerCheck+finLabel$finanCheck)
groupData <- combine %>%
  filter(ST %in% c(6,12,17,36,48)) %>%
  na.omit() %>%
  filter(WAGP!=000000) %>% #exclude no salary person
  filter(WAGP!='bbbbbb') %>% # exclude N/A
  filter(WKL==1) %>% #only person who last worked in 12 month
  group_by(ST, check)
sumData <- summarize(groupData, avgSalary = mean(WAGP))
sumData$state <- vapply(sumData$ST, regionTransfer, "")
occupationList <- c("0"="others", "1"="IT", "2"="Finance", "3"="cross")
occupationTransfer <- function(x){return ( occupationList[[as.character(x)]]) }
sumData$occ <- vapply(sumData$check, occupationTransfer, "")
ggplot(sumData, aes(x=state , y=avgSalary, fill=factor(occ))) +
  geom_bar(stat="identity",position="dodge") + scale_fill_hue(l=40)
```

90000
60000
30000
0
avgSalary
california
florida
illinois
new york
texas
state
factor(occ)
cross
Finance
IT
others

Hexiu Ye

research on Income and Class of Work

```r
setwd("~/GitHub/project1")
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
```

```r
library(ggplot2)
library(data.table)
```

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

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

```r
library(RColorBrewer)

colsToKeep <- c("ST","PINCP","OCCP","COW")

#load data from set A and B
populDataA <- fread("ss13pusa.csv",select=colsToKeep)
```

```
##
Read 0.0% of 1613672 rows
Read 17.4% of 1613672 rows
Read 35.9% of 1613672 rows
Read 55.2% of 1613672 rows
Read 74.4% of 1613672 rows
Read 93.0% of 1613672 rows
Read 1613672 rows and 4 (of 283) columns from 1.416 GB file in 00:00:08
```

```r
populDataB <- fread("ss13pusb.csv",select=colsToKeep)
```

```
##
Read 0.0% of 1519123 rows
Read 20.4% of 1519123 rows
Read 40.8% of 1519123 rows
Read 61.2% of 1519123 rows
Read 81.0% of 1519123 rows
Read 1519123 rows and 4 (of 283) columns from 1.333 GB file in 00:00:08
```

```
#concat data to one
populData <- rbind(populDataA, populDataB)

populData <- tbl_df(populData)
ds <-  populData %>%
  na.omit() %>%
  #filter(populData,PINCP!='bbbbbb') %>% #exclude no income person or N/A
  group_by(COW) #group by class of work
ds<-filter(ds,PINCP!='bbbbbb') #exclude no income N/A

mean_cow<-summarise(ds,mean=mean(PINCP))
mean_cow<-arrange(mean_cow, desc(mean))
mean_cow
```

```
## Source: local data frame [9 x 2]
##
##     COW     mean
##   (int)    (dbl)
## 1     7 84561.77
## 2     5 58572.12
## 3     4 45508.14
## 4     2 44068.76
## 5     3 43373.26
## 6     6 42231.65
## 7     1 41789.49
## 8     8 19399.57
## 9     9  2115.10
```

```
#boxplot(mean_cow$mean~mean_cow$COW,outline=TRUE)


ggplot(data=mean_cow, aes( x=factor(COW), y=mean,fill=factor(COW))) +
  geom_bar(colour="black",stat="identity")+
  xlab("class of work") + ylab("mean of total person's income ") +
  ggtitle("Average Income of Different Classes of Work")+
  scale_fill_hue(c=40, l=75)+
  scale_fill_discrete(
                      breaks=c("1", "2", "3","4","5","6","7","8","9"),
                      labels=c("Employee of a private for-profit company or business, or of an individua
                               "Employee of a private not-for-profit tax-exempt, or charitable organiza
                                "Local government employee (city, county, etc.)",
                                "State government employee",
                                "Federal government employee",
                                "Self-employed in own not incorporated business, professional practice,
                                "Self-employed in own incorporated business, professional practice or fa
```

```
                                  "Working without pay in family business or farm",
                                   "Unemployed and last worked 5 years ago or earlier or never")
                        )+
  theme(legend.position="bottom",legend.direction = "vertical")
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill',
## which will replace the existing scale.
```

```
ds5<-filter(populData,COW==5)%>%
    na.omit()%>%
    filter(PINCP!='bbbbbb')%>%
    group_by(OCCP)%>%
    summarise(mean=mean(PINCP))%>%
    arrange(desc(mean))

ds5head<-head(ds5)
ds5head
```

```
## Source: local data frame [6 x 2]
##
##    OCCP     mean
##   (int)    (dbl)
## 1  3060 180823.9
## 2   360 140396.1
```

```
## 3  3010 140227.3
## 4  3256 139156.2
## 5  1800 126614.3
## 6  2100 122488.7
```

```
ds5tail<-tail(ds5)
ds5tail
```

```
## Source: local data frame [6 x 2]
##
##    OCCP  mean
##   (int) (dbl)
## 1  8510  3000
## 2  4150  1480
## 3  7840  1000
## 4  7260   140
## 5  4410     0
## 6  6240     0
```

```
ds7<-filter(populData,COW==7)%>%
  na.omit()%>%
  filter(PINCP!='bbbbbb')%>%
  group_by(OCCP)%>%
  summarise(mean=mean(PINCP))%>%
  arrange(desc(mean))

ds7head<-head(ds7)
ds7tail<-tail(ds7)
ds7head
```

```
## Source: local data frame [6 x 2]
##
##    OCCP     mean
##   (int)    (dbl)
## 1  3200 429000.0
## 2  1930 404000.0
## 3  3060 270687.0
## 4  1800 260700.0
## 5  1200 232429.2
## 6  9050 227000.0
```

```
ds5tail
```

```
## Source: local data frame [6 x 2]
##
##    OCCP  mean
##   (int) (dbl)
## 1  8510  3000
## 2  4150  1480
## 3  7840  1000
## 4  7260   140
## 5  4410     0
## 6  6240     0
```

The poorest 5%: who they are?

Juan Campos


```
##
Read 0.0% of 1613672 rows
Read 8.7% of 1613672 rows
Read 16.7% of 1613672 rows
Read 26.0% of 1613672 rows
Read 34.7% of 1613672 rows
Read 42.8% of 1613672 rows
Read 50.8% of 1613672 rows
Read 58.9% of 1613672 rows
Read 66.9% of 1613672 rows
Read 75.6% of 1613672 rows
Read 84.9% of 1613672 rows
Read 93.0% of 1613672 rows
Read 1613672 rows and 15 (of 283) columns from 1.416 GB file in 00:00:20
```

```
##
Read 0.0% of 1519123 rows
Read 8.6% of 1519123 rows
Read 16.5% of 1519123 rows
Read 25.0% of 1519123 rows
Read 33.6% of 1519123 rows
Read 42.8% of 1519123 rows
Read 52.7% of 1519123 rows
Read 61.9% of 1519123 rows
Read 70.4% of 1519123 rows
Read 79.0% of 1519123 rows
Read 87.6% of 1519123 rows
Read 96.1% of 1519123 rows
Read 1519123 rows and 15 (of 283) columns from 1.333 GB file in 00:00:19
```

The 5% poorest people live with **$2000**.

Income distribution by age

The age distribution of the poor is drastically different from the rest of the population with the share of young adults being very high.

Age distribution among the masses
Count
200000
150000
100000
50000
0
25
50
75
100
Age

Age distribution among the Poor
Count
30000
20000
10000
0
25
50
75
100
Age

The states with the highest populations of the poor are **California**, **Texas** and **New York**. However this plot may be misleading as California is state with a high share of the population. Actually When the richest 5% is ploted instead, a similar plot is obtained.

```
tb.poor <- table (poor$ST)
tb <- tb.poor / sum (tb.poor)
plot.map (tb, "Distribution of the poor across states", "Share of total")
```

The following plot shows the share per state. Alaska and Utah have the highest concentation of poor people:

```
tb.mass <- table (mass$ST)
tb <- tb.poor /  (tb.poor + tb.mass)
plot.map (tb, "Share of the poor by state", "Share per state")
```

Race

The proportion of poor **Whites** is lower than the proportion of the rest of Whites. All other races are overrepresented, this is especially true of **African Americans**. **Native Americans** are clearly overrepresented within the poor.

Largest races within income groups
Share
0.8
0.6
0.4
0.2
0.0
White
African American
Asian
Other
Multiple
Group
Masses
Poor

Native Americans within income groups
Share
0.009
0.006
0.003
0.000
American Indian
Alaska Native
Unspecified Native
Hawaiian Native
Group
Masses
Poor

yimin zhang

February 2, 2016

```r
# 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
```

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

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

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

```r
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)
```

```r
  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")
```

```r
# 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