Hexiu Ye

research on Income and Class of Work

```
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
```

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

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

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

```
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
```

```
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
```

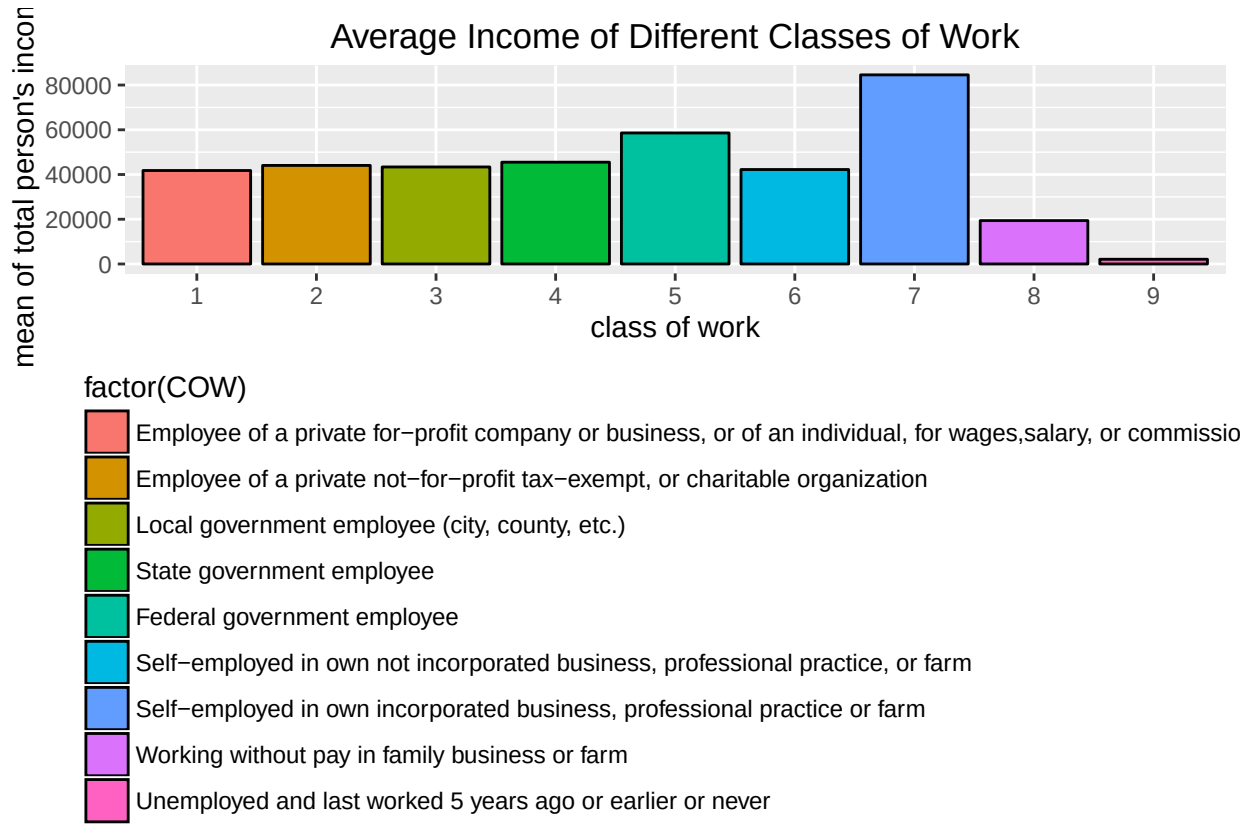
```
## Source: local data frame [9 x 2]
##
##     COW     mean
##   (int)    (dbl)
## 1     7 84561.77
## 2     5 58572.12
## 3     4 45508.14
## 4     2 44068.76
## 5     3 43373.26
## 6     6 42231.65
## 7     1 41789.49
## 8     8 19399.57
## 9     9  2115.10
```

```
#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 organizat
                               "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
```