Project_1_final

Team 1

February 3, 2016

Introduction

In our first project, we are delighted to present our research on the demographics of Chinese living at the USA. As most of our audience are on the path to be a Chinese Master-degree holder, we explored further about the living conditions and salary level of Chinese master-degree holders.

Through our research, we hope to exploit the income level, standard of living, residence distribution, working condition, gender disparity and marriage. To set off the journey, we began with

1: Intall necessary packages for the project

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

```
## Warning: package 'data.table' was built under R version 3.1.3
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.1.2
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:data.table':
##
##     between, last
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```

```r
library(maps)
```

```
## Warning: package 'maps' was built under R version 3.1.1
```

```r
library(gridExtra)
```

```
## Loading required package: grid
```

```r
library(RColorBrewer)
```

```
## Warning: package 'RColorBrewer' was built under R version 3.1.2
```

2: Data Preparation

As the concentration of the project is only a subset of the total population, filters are mandatory to our research.

```r
setwd("~/Desktop/Applied Data Science/csv_pus")
pusa <- fread("~/Desktop/Applied Data Science/csv_pus/ss13pusa.csv")
```

```
##
Read 0.0% of 1613672 rows
```

```
## Warning in fread("~/Desktop/Applied Data Science/csv_pus/ss13pusa.csv"):
## Bumped column 126 to type character on data row 38, field contains
## '1721YY'. Coercing previously read values in this column from logical,
## integer or numeric back to character which may not be lossless; e.g., if
## '00' and '000' occurred before they will now be just '0', and there may
## be inconsistencies with treatment of ',,' and ',NA,' too (if they occurred
## in this column before the bump). If this matters please rerun and set
## 'colClasses' to 'character' for this column. Please note that column type
## detection uses the first 5 rows, the middle 5 rows and the last 5 rows, so
## hopefully this message should be very rare. If reporting to datatable-help,
## please rerun and include the output from verbose=TRUE.
```

```
##
Read 9.3% of 1613672 rows
Read 18.6% of 1613672 rows
Read 27.9% of 1613672 rows
Read 37.2% of 1613672 rows
Read 46.5% of 1613672 rows
Read 55.8% of 1613672 rows
Read 65.1% of 1613672 rows
Read 74.4% of 1613672 rows
Read 83.7% of 1613672 rows
Read 93.0% of 1613672 rows
Read 1613672 rows and 283 (of 283) columns from 1.416 GB file in 00:00:21
```

```r
pusb <- fread("~/Desktop/Applied Data Science/csv_pus/ss13pusb.csv")
```

```
##
Read 0.0% of 1519123 rows
Read 9.9% of 1519123 rows
```

```
Read 19.7% of 1519123 rows
Read 29.6% of 1519123 rows
Read 39.5% of 1519123 rows
Read 49.4% of 1519123 rows
Read 59.2% of 1519123 rows
Read 69.1% of 1519123 rows
Read 79.0% of 1519123 rows
Read 88.9% of 1519123 rows
Read 98.7% of 1519123 rows
Read 1519123 rows and 283 (of 283) columns from 1.333 GB file in 00:00:18
```

```
pus <- rbind(pusa, pusb)

# Here we define chinese data

chinese <- pus%>%
filter(RAC2P==43|RAC2P==44|POBP==207|POBP==209|POBP==240)

# Here we interpret our varibles

chinese$ST <- as.factor(chinese$ST)
chinese$MSP<-as.factor(chinese$MSP)
chinese$SCIENGRLP<-as.factor(chinese$SCIENGRLP)
chinese$SEX<-as.factor(chinese$SEX)
chinese$ESR<-as.factor(chinese$ESR)
levels(chinese$MSP)<-c("married&spouse present","married&spouse absent","Widowed","Divorced","Separated
levels(chinese$SCIENGRLP)<-c("Sci","Non-sci")
levels(chinese$SEX)<-c("Male","Female")
levels(chinese$ESR)<-c("empd&work","empd not work","unempd","af&work","af with job but not work","not i
chinese$MSPG <- ifelse(chinese$MSP == 1|chinese$MSP == 2, 1, 0)
chinese$MSPG <-factor(chinese$MSPG)
levels(chinese$MSPG)<-c("Now Married","Other Conditions")
chinese$MSP <- factor(chinese$MSP)
levels(chinese$MSP) <- c("Now married, spouse present", "Now married, spouse absent","Widowed", "Divorc
#Add Indicator ESRG
chinese$ESR <- factor(chinese$ESR)
levels(chinese$ESR) <- c("Employed", "Employed, not at work", "Unemployed", "Employed", "Employed, not a
chinese$ESRG <- ifelse(chinese$ESR == "Employed", 1, 0)

# Code for state name

levels(chinese$ST) <- c("Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado", "Connecti
"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")

# code for industry

chinese$INDP <- ifelse(chinese$INDP>= 170 & chinese$INDP <= 290, 170, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP>= 370 & chinese$INDP <= 490, 370, chinese$INDP)
```

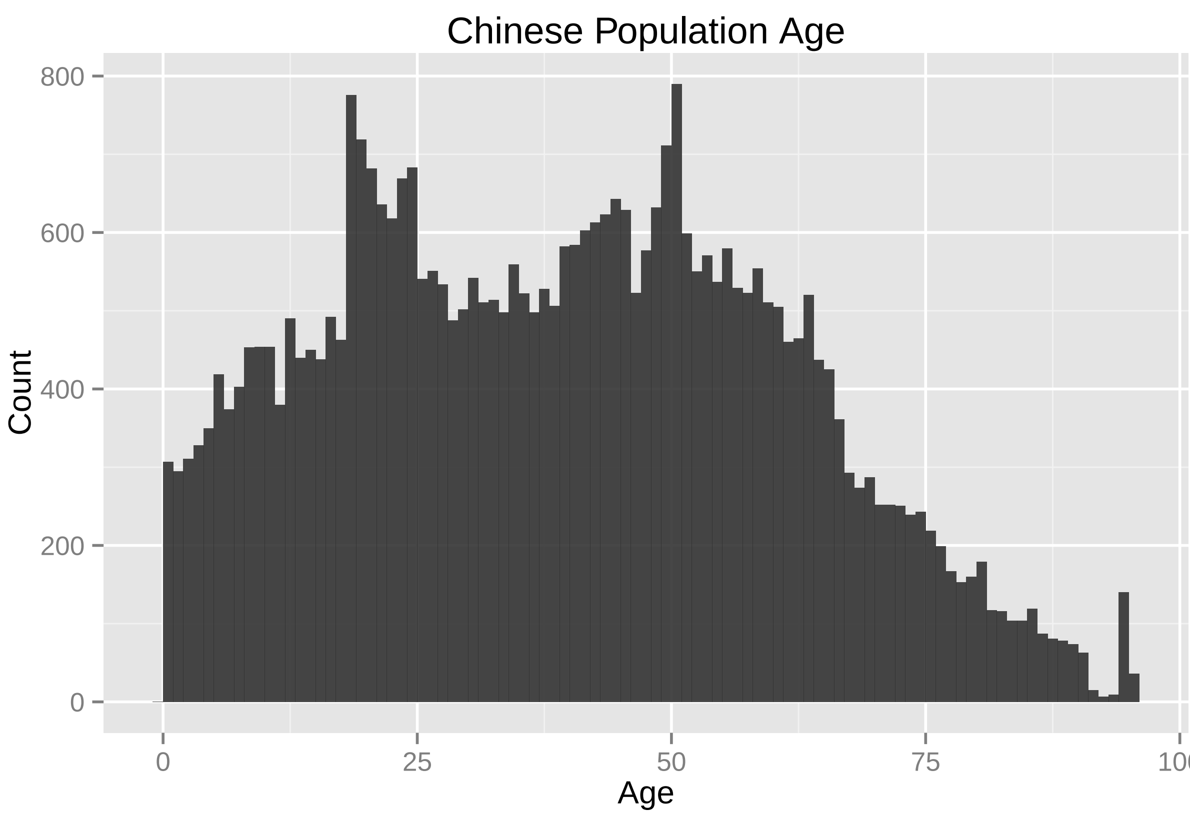
```
chinese$INDP <- ifelse(chinese$INDP >= 570 & chinese$INDP<= 770, 570, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 1070 & chinese$INDP <= 3990, 1070, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 4070 & chinese$INDP <= 6390, 4070, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 6470 & chinese$INDP <= 6780, 6470, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP>= 6870 & chinese$INDP <= 7190, 6870, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 7270 & chinese$INDP <= 7790, 7270, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 7860 & chinese$INDP<= 7890, 7860, chinese$INDP)
chinese$INDP<- ifelse(chinese$INDP >= 7970 & chinese$INDP <= 8290, 7970, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 8370 & chinese$INDP <= 8470, 8370, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP %in% c(8660, 8680, 8690), 8370, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 8770 & chinese$INDP <= 9290, 8370, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP %in% c(8560, 8570, 8580, 8590, 8670), 8560, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 9370 & chinese$INDP <= 9590, 9370, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 9670 & chinese$INDP<= 9870, 9670, chinese$INDP)
chinese$INDP <- ifelse(chinese$INDP >= 9920, 9920, chinese$INDP)
chinese$INDP <- factor(chinese$INDP)
levels(chinese$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"
            )

# code for decade

chinese$DECADE <- factor(chinese$DECADE)
levels(chinese$DECADE) <- c("~1950's", "1950's", "1960's", "1970's", "1980's", "1990's", "2000's~")
chinese$OCCP <- ifelse(chinese$OCCP >= 10 & chinese$OCCP <= 430, 10, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 500 & chinese$OCCP <= 740, 500, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 800 & chinese$OCCP <= 950, 800, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 1005 & chinese$OCCP <= 1240, 1005, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 1300 & chinese$OCCP <= 1560, 1300, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 1600 & chinese$OCCP <= 1965, 1600, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 2000 & chinese$OCCP <= 2060, 2000, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 2100 & chinese$OCCP <= 2160, 2100, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 2200 & chinese$OCCP <= 2550, 2200, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 2600 & chinese$OCCP <= 2920, 2600, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 3000 & chinese$OCCP <= 3540, 3000, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 3600 & chinese$OCCP <= 3655, 3600, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 3700 & chinese$OCCP <= 3955, 3700, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 4000 & chinese$OCCP <= 4150, 4000, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 4210 & chinese$OCCP <= 4250, 4210, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 4300 & chinese$OCCP <= 4650, 4300, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 4700 & chinese$OCCP <= 4965, 4700, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 5000 & chinese$OCCP <= 5940, 5000, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 6005 & chinese$OCCP <= 6130, 6005, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 6200 & chinese$OCCP <= 6765, 6200, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 6800 & chinese$OCCP <= 6940, 6800, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 7000 & chinese$OCCP <= 7630, 7000, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 7700 & chinese$OCCP <= 8965, 7700, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 9000 & chinese$OCCP <= 9750, 9000, chinese$OCCP)
chinese$OCCP <- ifelse(chinese$OCCP >= 9800 & chinese$OCCP <= 9830, 9800, chinese$OCCP)

tempchinese<-chinese%>%
```

```
    filter(OCCP %in% c(9920,5000,3700,800,1600,1005,10,500))
tempchinese$OCCP <- factor(tempchinese$OCCP)
levels(tempchinese$OCCP)<-c("MGR","BUS","FIN","CMM","SCI","PRT","OFF","Unemplyed(broad)")
chinese_Ed<-tempchinese[SCHL>=21]
```

3: Age distribution of Chinese

How old are most of the Chinese living in the United States now ?

```
ggplot(chinese, aes(AGEP)) +
  geom_bar( binwidth=1, alpha=0.9) +
  xlab("Age") + ylab("Count") + ggtitle("Chinese Population Age")
```



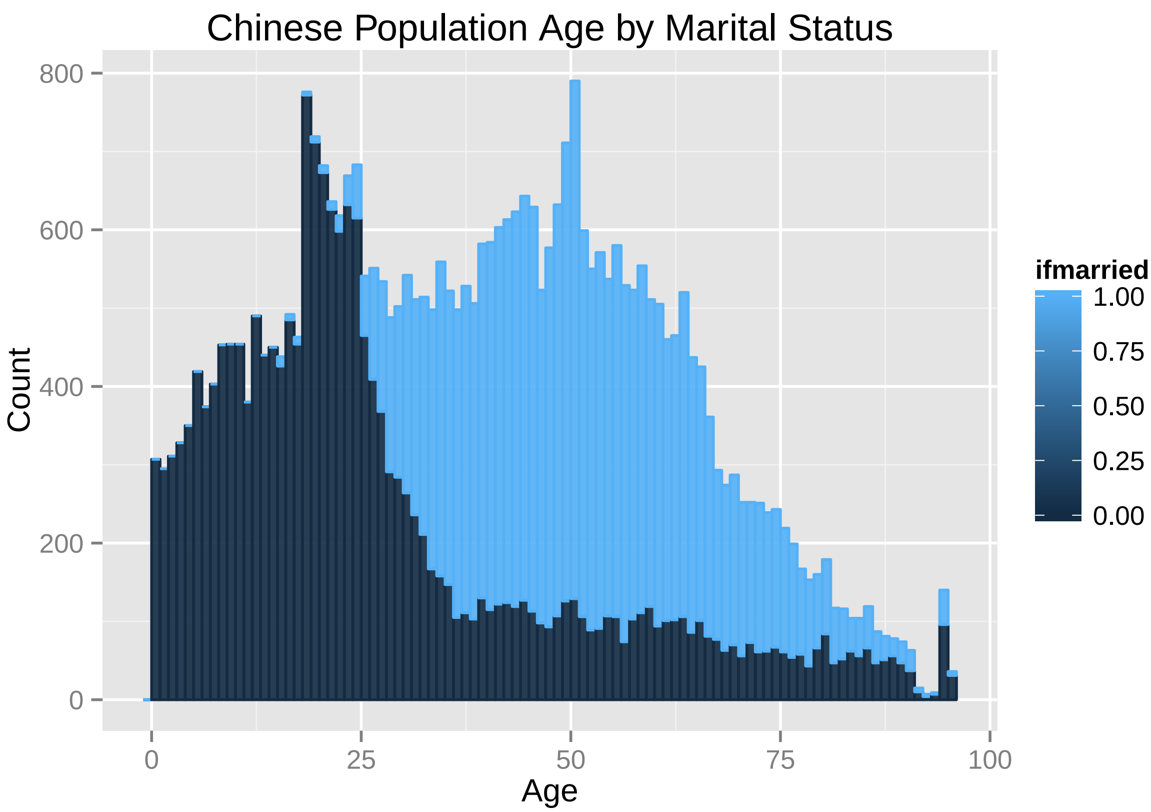
Two peaks:

4: Are they married ?

How does the pattern of marital status change with time ?

```
ifmarried <- rep(0,dim(chinese)[1])
for (i in 1:dim(chinese)[1]){
  if (chinese$MAR[i]==1){
    ifmarried[i]= 1
  }
}
```

```r
ggplot(chinese, aes(AGEP, group=ifmarried)) +
  geom_bar(aes(colour=ifmarried, fill=ifmarried), binwidth=1, alpha=0.9) +
  xlab("Age") + ylab("Count") + ggtitle("Chinese Population Age by Marital Status")
```

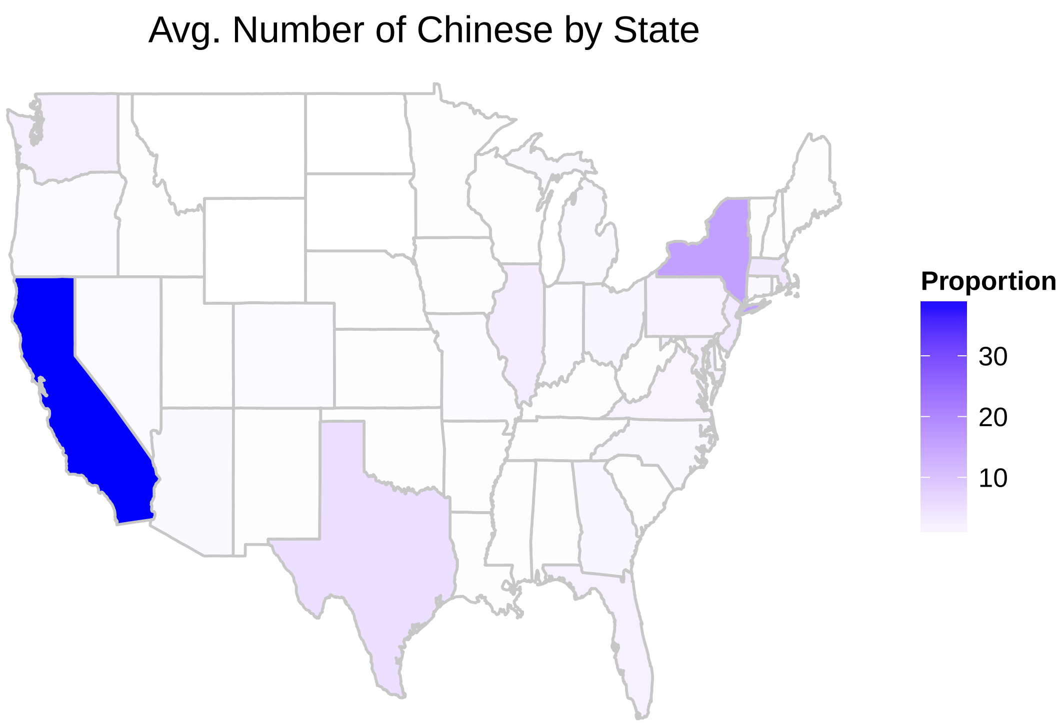


Most of the Chinese marry between the age of 25-30.

5: Where do they live ?

```r
# prepare data
all_state <- map_data("state")
data <- as.data.frame(prop.table(table(chinese$ST)))
data$state <- c(sort(tolower(c("district of columbia", state.name))),tolower("Puerto Rico"))
all_state$freq <- data$Freq[match(all_state$region, data$state)]*100

# draw map
p_1 <- ggplot(all_state, aes(x=long, y=lat, group=group)) +
      geom_polygon(aes(fill=freq), colour="gray78") +
      scale_fill_gradient(name="Proportion", low="white", high="blue")
p_1 <- p_1 + theme(strip.background = element_blank(),
               strip.text.x     = element_blank(),
               axis.text.x      = element_blank(),
               axis.text.y      = element_blank(),
               axis.ticks       = element_blank(),
               axis.line        = element_blank(),
               panel.background = element_blank(),
               panel.border     = element_blank(),
```

```
                panel.grid        = element_blank(),
                legend.position   = "right") +
           xlab("") + ylab("") + ggtitle("Avg. Number of Chinese by State")

p_1
```



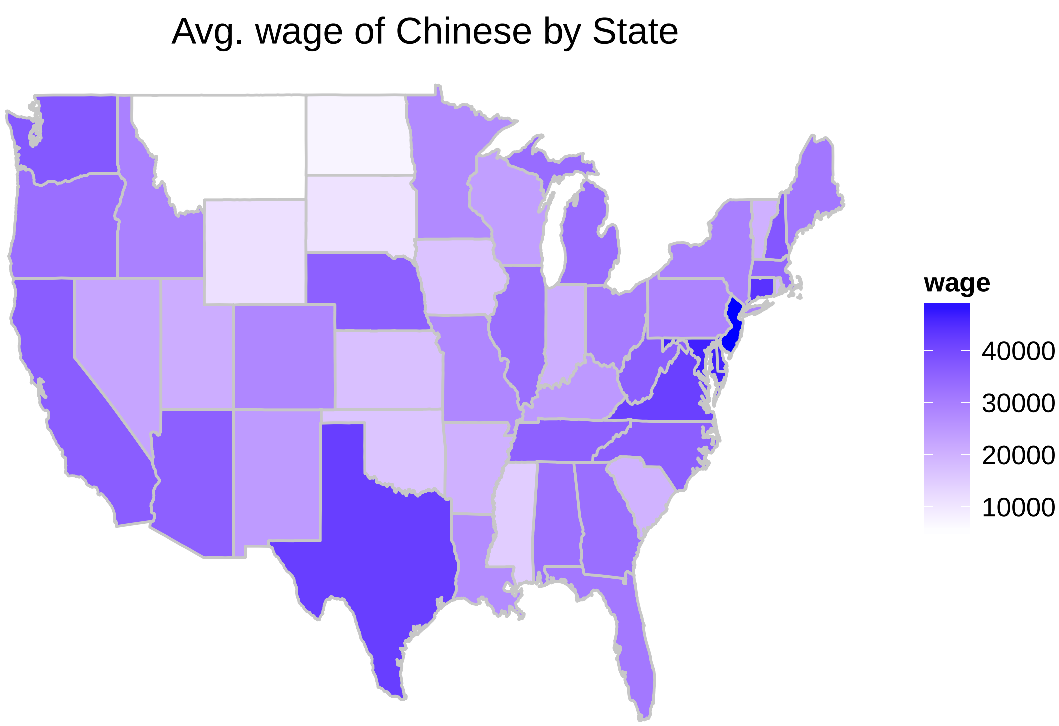
California, New York and Texas seem to have the most Chinese

6: How much do they earn ?

The average standard of living in the US is high, but what about the Chinese? Do they live well and earn a decent living ?

```
#prepare average income data

average_wage <- chinese %>%
        filter(is.na(WAGP) == F) %>%
        group_by(ST) %>%
        summarise(wage = mean(WAGP))
# we don't have Puerto Rico here
state_51 = data[-52,]
state_51$wage = average_wage$wage
all_state$wage = state_51$wage[match(all_state$region,state_51$state)]

#draw map

p_2 <- ggplot(all_state, aes(x=long, y=lat, group=group)) +
      geom_polygon(aes(fill=wage), colour="gray78") +
      scale_fill_gradient(name="wage", low="white", high="blue")
```

```r
p_2 <- p_2 + theme(strip.background = element_blank(),
               strip.text.x     = element_blank(),
               axis.text.x      = element_blank(),
               axis.text.y      = element_blank(),
               axis.ticks       = element_blank(),
               axis.line        = element_blank(),
               panel.background = element_blank(),
               panel.border     = element_blank(),
               panel.grid       = element_blank(),
               legend.position  = "right") +
          xlab("") + ylab("") + ggtitle("Avg. wage of Chinese by State")

p_2
```



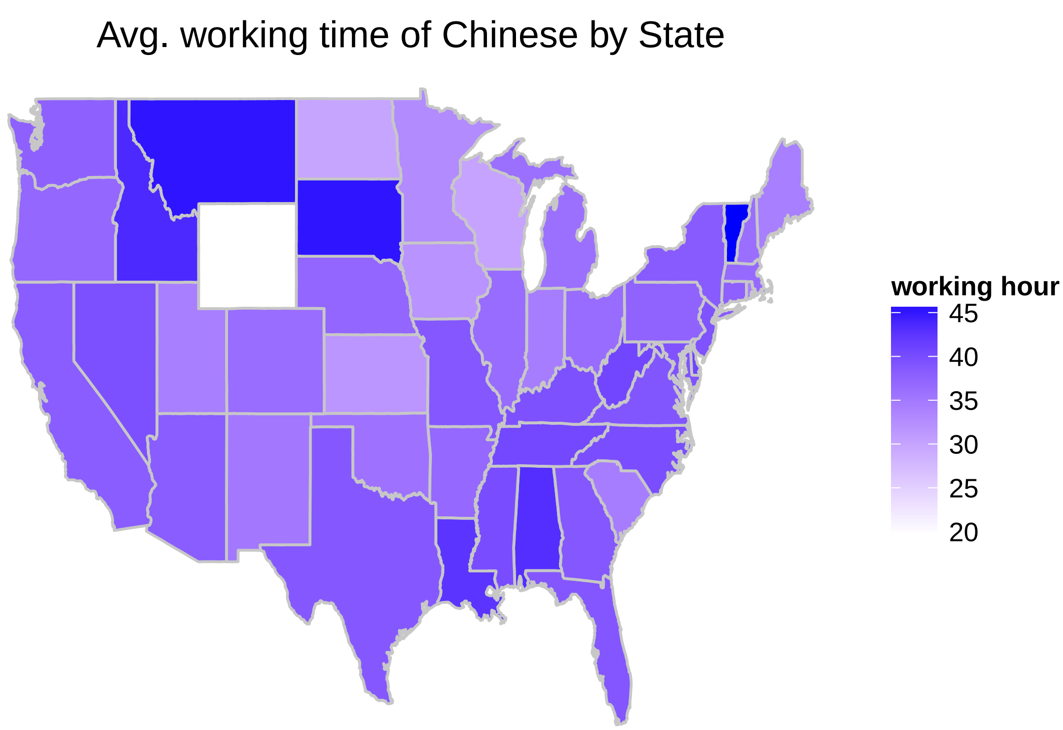
It seems to vary by states.

7: How long do they work ?

Chinese are proud to be hard-working, are there some Chinese more hard-working than others?

```r
#prep
all_state <- map_data("state")
data <- as.data.frame(prop.table(table(chinese$ST)))
data$state <- c(sort(tolower(c("district of columbia", state.name))),tolower("Puerto Rico"))
all_state$freq <- data$Freq[match(all_state$region, data$state)]*100

#work hour of chinese masters
work_hour <- chinese %>%
        filter(is.na(WKHP) == F) %>%
```

```r
        group_by(ST) %>%
        summarise(workhour = mean(WKHP))
state_51 = data[-c(52),]
state_51$workhour = work_hour$workhour
all_state$workhour = work_hour$workhour[match(all_state$region,state_51$state)]

p_work <- ggplot(all_state, aes(x=long, y=lat, group=group))+
        geom_polygon(aes(fill=workhour), color = "gray78") +
        scale_fill_gradient(name="working hour", low="white", high = "blue")
p_work <- p_work + theme(strip.background = element_blank(),
                         strip.text.x     = element_blank(),
                         axis.text.x      = element_blank(),
                         axis.text.y      = element_blank(),
                         axis.ticks       = element_blank(),
                         axis.line        = element_blank(),
                         panel.background = element_blank(),
                         panel.border     = element_blank(),
                         panel.grid       = element_blank(),
                         legend.position  = "right") +
        xlab("") + ylab("") + ggtitle("Avg. working time of Chinese by State")

p_work
```

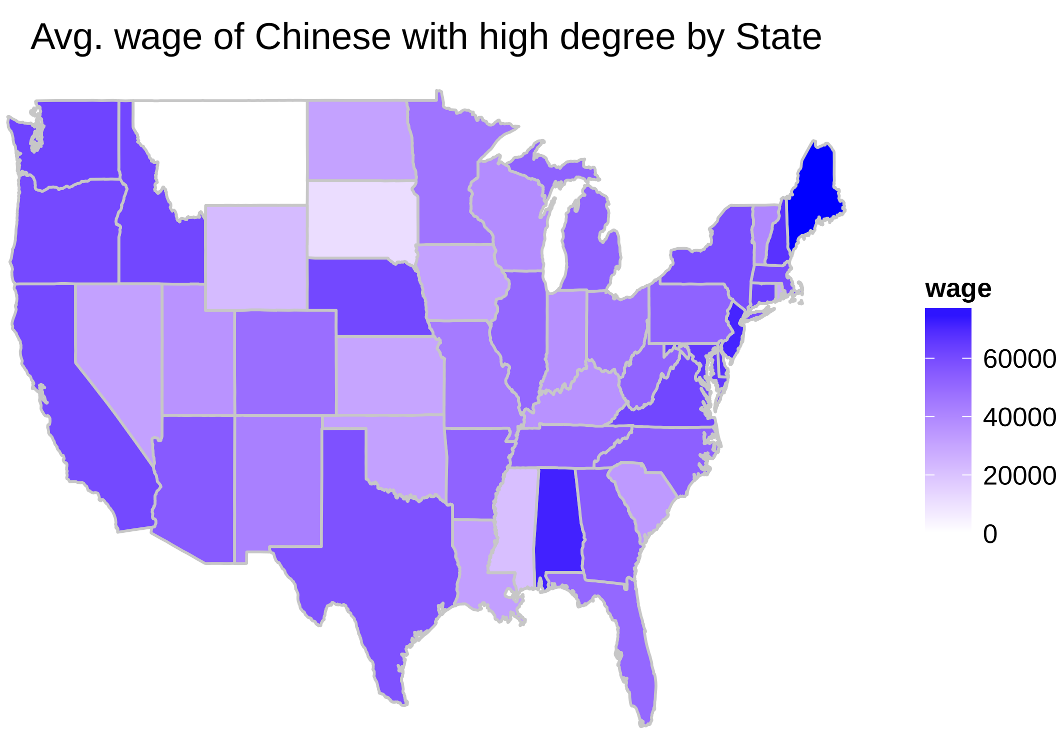


South Dakota, Idaho and Montana residents has proven themselves to be the most hard-working people!

8: Average wage for chinese immigrants with a master degree or higher by state

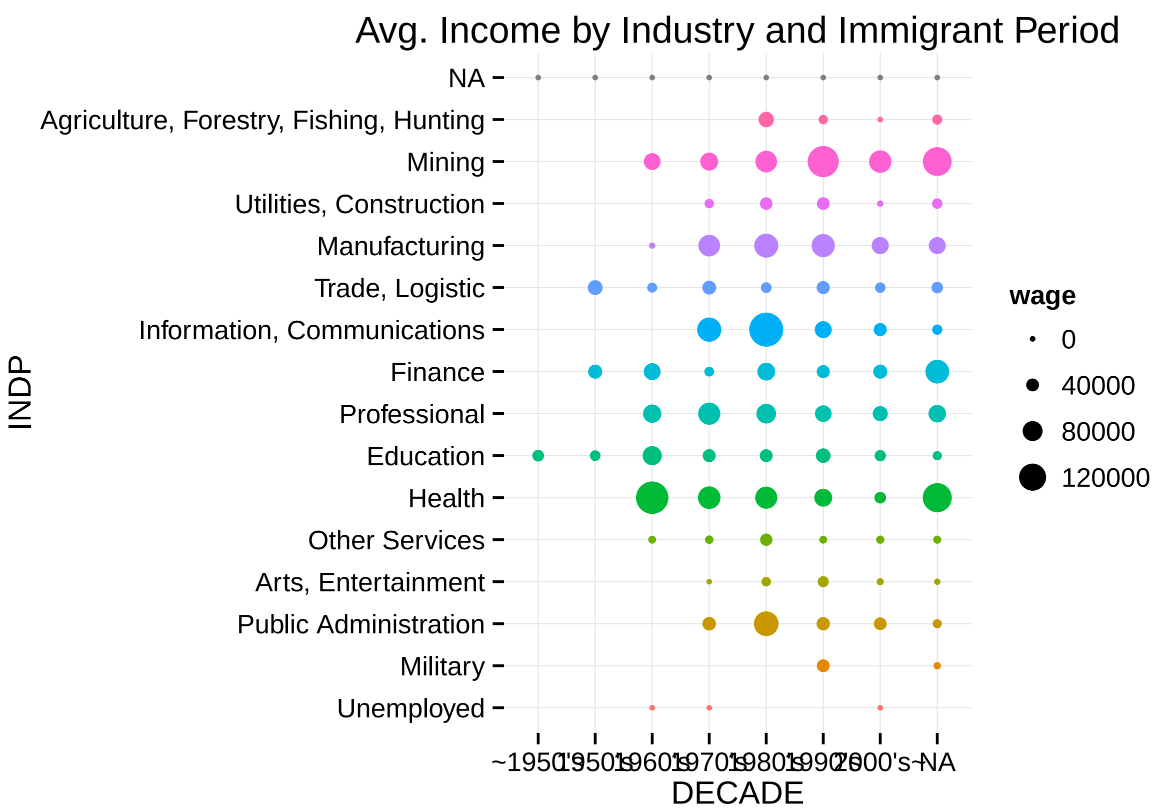
As most of the our audiences are Master-degree Chinese students, it is important to delight them with our findings in their possible income level after graduation.

```r
# prepare data
wage_degree <- chinese %>%
        filter(is.na(WAGP) == F, SCHL>=21) %>%
        group_by(ST) %>%
        summarise(wage = mean(WAGP))
state_51d = data[-52,]
# We don't have Puerto Rico, Montana and North Dakoda here
state_51d$wage = wage_degree$wage
all_state$wage = state_51d$wage[match(all_state$region,state_51d$state)]

# draw map

p_3 <- ggplot(all_state, aes(x=long, y=lat, group=group)) +
      geom_polygon(aes(fill=wage), colour="gray78") +
      scale_fill_gradient(name="wage", low="white", high="blue")
p_3 <- p_3 + theme(strip.background = element_blank(),
               strip.text.x     = element_blank(),
               axis.text.x      = element_blank(),
               axis.text.y      = element_blank(),
               axis.ticks       = element_blank(),
               axis.line        = element_blank(),
               panel.background = element_blank(),
               panel.border     = element_blank(),
               panel.grid       = element_blank(),
               legend.position  = "right") +
          xlab("") + ylab("") + ggtitle("Avg. wage of Chinese with high degree by State")

p_3
```



9: The wage structure of chinese immigrants in Texas

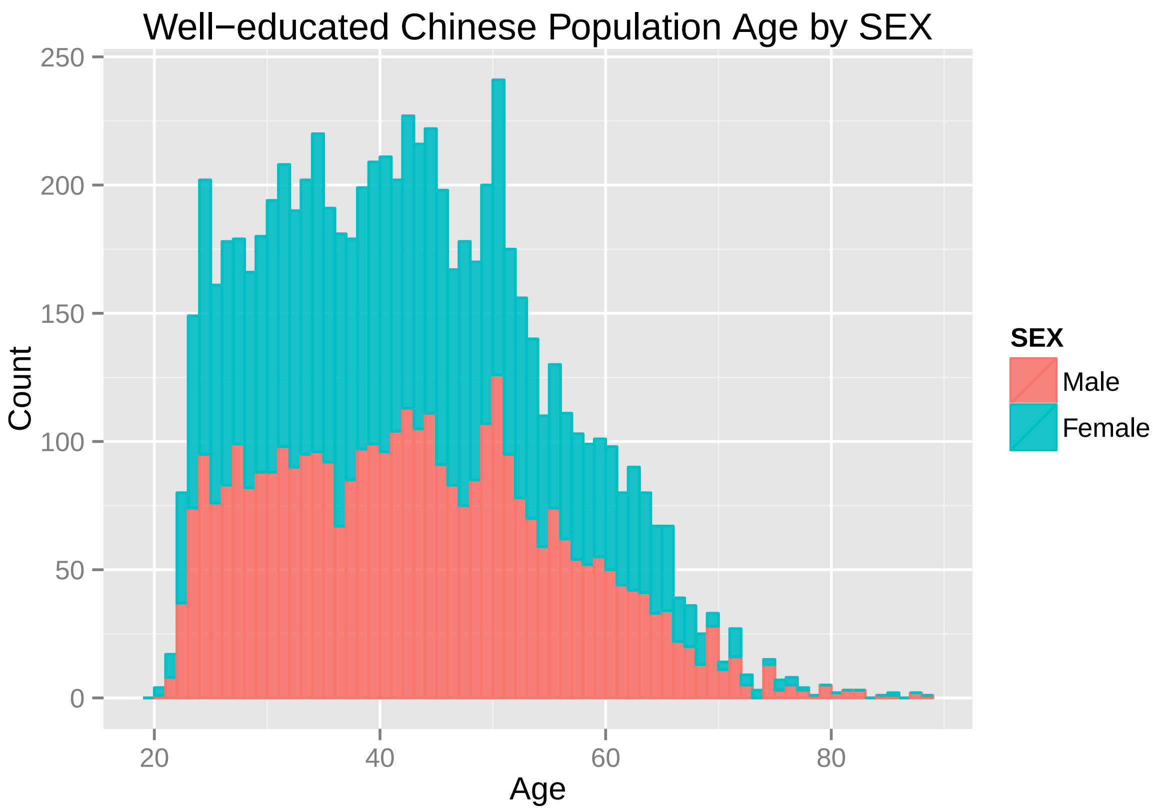
```
# prepare the data
TX_structure = chinese %>%
        filter(ST == "Texas",is.na(WAGP) == F) %>%
        group_by(DECADE, INDP) %>%
        summarise(wage = mean(WAGP))
TX_structure$INDP <- factor(TX_structure$INDP, levels = levels(TX_structure$INDP)[length(levels(TX_stru
#Plot the Taxes income structure
ggplot(TX_structure, aes(x=DECADE)) + geom_point(aes(y=INDP, size=wage, colour=INDP)) +
  ggtitle("Avg. Income by Industry and Immigrant Period") +
  guides(colour=FALSE) + theme_minimal()
```



From this graph, we can see that Mining and are a job with a comparetively good salary in each generation; Information and Comunication dominates the job market in 1980's. Health industry would be a place to start from ground up and build your exprience in.

10: Well-educated Chinese Imigrants' gender distribution on different ages

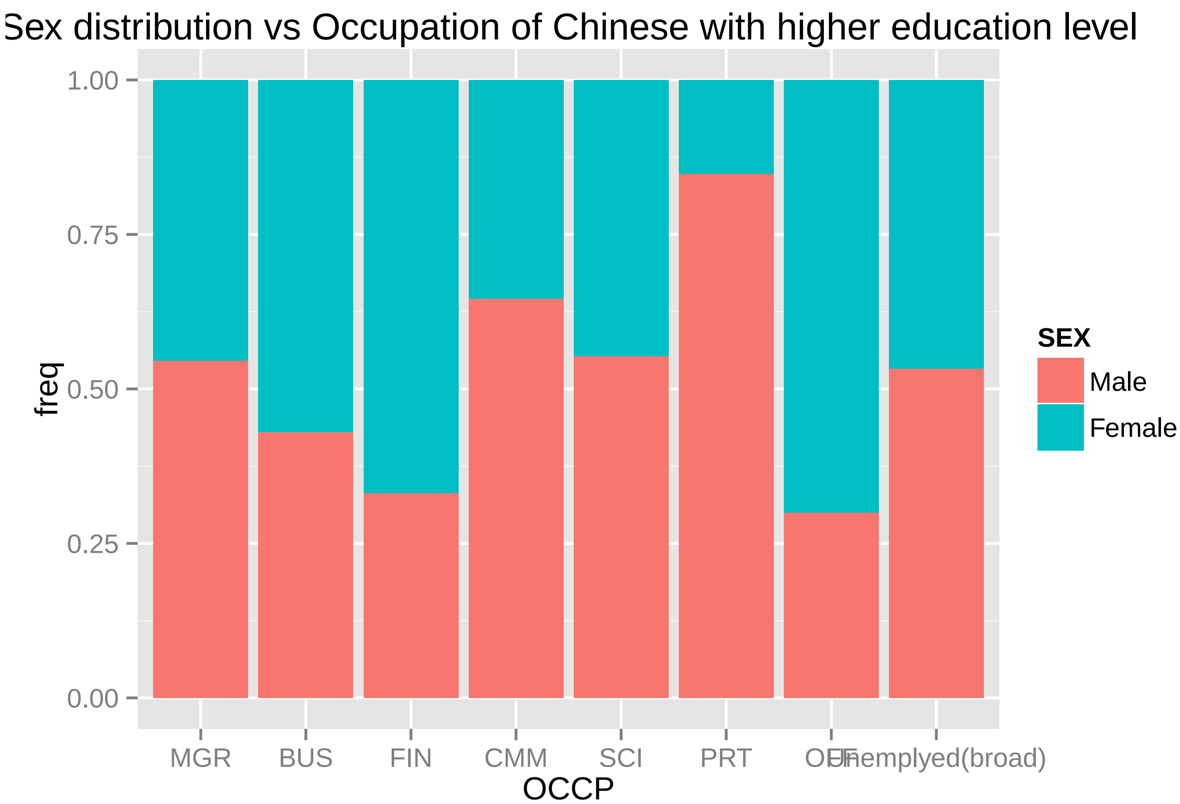
```
ggplot(chinese_Ed, aes(AGEP, group=SEX)) +
  geom_bar(aes(colour=SEX, fill=SEX), binwidth=1, alpha=0.9) +
  xlab("Age") + ylab("Count") + ggtitle("Well-educated Chinese Population Age by SEX")
```



11: Gender distribution in each occupation

We can see each occupation does have different gender distribution.

```
ggplot(chinese_Ed, aes(x=OCCP)) +
    geom_bar(aes(fill=SEX), position="fill") +ylab("freq")+
    ggtitle("Sex distribution vs Occupation of Chinese with higher education level")
```



Is there gender discrimination ?

```
saldif.sci=
  chinese_Ed%>%
  filter(SCIENGP=="1")%>%
  group_by(SEX)%>%
  summarise(
    avgsalary=mean(PINCP,na.rm=T)
  )
saldif.sci$SEX<-as.factor(saldif.sci$SEX)
levels(saldif.sci$SEX)<-c("Male","Female")
saldif.nonsci=
  chinese_Ed%>%
  filter(SCIENGP=="2")%>%
  group_by(SEX)%>%
  summarise(
    avgsalary=mean(PINCP,na.rm=T)
  )
saldif.nonsci$SEX<-as.factor(saldif.nonsci$SEX)
levels(saldif.nonsci$SEX)<-c("Male","Female")

gender = c("male","male","female","female")
salary = c(100125.58,76621.28,76621.28,64367.85)
industry = c("Science","Non-Science","Science","Non-Science")
df = cbind(gender,industry,salary)
df = data.frame(df)
```

```
df$salary = salary

ggplot (data = df,aes(industry,salary,fill = gender)) +
                geom_bar(stat = "identity",position="dodge")+
                xlab("Gender") + ylab("Salary") +
                ggtitle("Salary for Science Related Work by Gender")
```

As we can see, there is a high possibilty of gender discrimination. For people under the same education level, working hour and industry, there's still much difference in salary between males and females.