Analysis of Immigrants in U.S

Prateek Rajput

18/12/2019

Data Summary: The data used in this project is from 2017 American Community Survey(ACS). Two types of data was made available, Household data and Population data. The one which I chose is the Population data. I selected 19 columns, but all of them were not used in my project. The one's that felt useful to me were:

ST: State code.
AGEP: Age of a person.
CIT: Citizenship status.
COW: Class of worker.
NWLK: Looking for work.
SCHL: Educational Attainment.
SCHG: Grade Level Attending.
SEX: Gender of a person.
WKL: When last worked.
ESR: Employment status recode.
FOD1P: Field of degree.
PINCP: Total person's income.
POBP: Place of birth of a person.
RAC3P: Recorded detailed race code.
ADJINC: Adjustment factor for income.

The main aim of this project is to glance over the numbers about immigrants in the States of U.S.A. Therefore, I did and exploratory analysis of immigrants in the U.S.

Pre Processing of the data: I created a string of countries' code and name to use it later in showing some of the data of the countries on world map.

```
person_1<- read_csv("psam_pusa.csv",col_types = cols_only(RT=col_character(),SERIALNO=col_character(),AGEP=col_integer(),CIT=col_character(),ST=col_character(),COW=col_character(),NWAB=col_character(),NWAV=col_character(),NWLA=col_character(),NWLK=col_character(),SEX=col_character(),WKL=col_character(),ANC1P=col_character(),ANC2P=col_character(),FOD1P=col_character(),FOD2P=col_character(),NATIVITY=col_character(),POBP=col_character(),RAC3P=col_character(),SCHL=col_character(),ESR=col_character(),YOEP=col_character(),PERNP=col_character()))
person_2<- read_csv("psam_pusb.csv",col_types = cols_only(RT=col_character(),SERIALNO=col_character(),AGEP=col_integer(),CIT=col_character(),ST=col_character(),COW=col_character(),NWAB=col_character(),NWAV=col_character(),NWLA=col_character(),NWLK=col_character(),SEX=col_character(),WKL=col_character(),ANC1P=col_character(),ANC2P=col_character(),FOD1P=col_character(),FOD2P=col_character(),NATIVITY=col_character(),POBP=col_character(),RAC3P=col_character(),SCHL=col_character(),ESR=col_character(),YOEP=col_character(),PERNP=col_character()))
person_3<- read_csv("psam_pusc.csv",col_types = cols_only(RT=col_character(),SERIALNO=col_character(),AGEP=col_integer(),CIT=col_character(),ST=col_character(),COW=col_character(),NWAB=col_character(),NWAV=col_character(),NWLA=col_character(),NWLK=col_character(),SEX=col_character(),WKL=col_character(),ANC1P=col_character(),ANC2P=col_character(),FOD1P=col_character(),FOD2P=col_character(),NATIVITY=col_character(),POBP=col_character(),RAC3P=col_character(),SCHL=col_character(),ESR=col_character(),YOEP=col_character(),PERNP=col_character()))
person_4<- read_csv("psam_pusd.csv",col_types = cols_only(RT=col_character(),SERIALNO=col_character(),AGEP=col_integer(),CIT=col_character(),ST=col_character(),COW=col_character(),NWAB=col_character(),NWAV=col_character(),NWLA=col_character(),NWLK=col_character(),SEX=col_character(),WKL=col_character(),ANC1P=col_character(),ANC2P=col_character(),FOD1P=col_character(),FOD2P=col_character(),NATIVITY=col_character(),POBP=col_character(),RAC3P=col_character(),SCHL=col_character(),ESR=col_character(),YOEP=col_character(),PERNP=col_character()))

#binding all the 4 data frames.
full_person<- rbind(person_1,person_2,person_3,person_4)

#making temporary variables to add some new columns.
t1<- read_csv("psam_pusa.csv",col_types = cols_only(SCHG=col_guess(),ADJINC=col_guess()))
t2<- read_csv("psam_pusb.csv",col_types = cols_only(SCHG=col_guess(),ADJINC=col_guess()))
t3<- read_csv("psam_pusc.csv",col_types = cols_only(SCHG=col_guess(),ADJINC=col_guess()))
t4<- read_csv("psam_pusd.csv",col_types = cols_only(SCHG=col_guess(),ADJINC=col_guess()))
t<- rbind(t1,t2,t3,t4)

#column binding the temporary variables to the original dataframe.
full_person<- cbind(t,full_person)
```

The full file containing all the data of person and all the columns is read in the following line. The file is in rds format.

```
#reading the whole file in RDS format.
full_person<- readRDS("full_person.rds")
```

All the libraries needed:

```
library(readr)
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)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
library(forcats)
library(usmap)
library(tidyverse)
```

```
## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2
```

```
## ── Attaching packages ──────────────────────────────────────────── tidyverse 1.2.1 ──
```

```
## ✓ tibble  2.1.1     ✓ purrr   0.3.2
## ✓ tidyr   0.8.3     ✓ stringr 1.4.0
```

```
## ── Conflicts ────────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(tools)
```

Methodology: The approcah I used to do exploratory analysis on the data of immigrants is: The first thing that I did is to find all the immigrants in the Person data. To do so, I chose all the entries with CIT==5(i.e. Not a citizen of U.S). **I am assuming that all the non-citizens of the U.S are immigrants.**

```
#non-citizens/immigrants
non_citizen <- full_person %>%
  filter(CIT==5)
```

I adjusted the income using ADJINC(adjusting factor for income).

```
#adjusting income.
non_citizen$PINCP<- as.integer(as.numeric(non_citizen$ADJINC) * as.numeric(non_citizen$PINC
P)/1000000)
```

I also needed the names of the states in U.S. So I installed a library called “usmap”. It has a data frame by the name “statepop” that contains the data of the states(code, abbreviation of the name, full name). I then joined the statepop and the non-citizen dataframes to add columns of the abbreviation and the full name of the states.

```
library(usmap)
statepop<-statepop %>% rename(ST=fips)
states_noncitizen<- inner_join(statepop,non_citizen,by="ST")
```

After doing all the things above,I started to explore interesting things about the data on immigrants. In this project, I mainly foused on statewise numbers of the immigrants to see if there is a difference or similarity in the immigrants in states. Therefore, I wanted to find below things:

1) Number of immigrants in each state.

2) The place of birth of the immigrants(i.e to see people from which country come to U.S).

3) Number of males and females from each country.

4) Average age of the immigrants in each state.

5) Per Capita Income of immigrants in each state.

6) If income of Male and Female immigrants were different.

7) Top Fields of degree based on income.

8) See if there are any jobless immigrants.

9) Immigrants currently pursuing bacheor’s or higher level degree(countrywise to see people from which country visit U.S to study.)

10) See if people have a bachelor’s degree but are jobless / looking for a job and see this nationwise.

11) People from which country have the highest income.

12) Check if there is a difference in the income of females and males in the Field of Degree that I will find the most earning(i.e. question 7).

13) Check if literacy affects income of graduates in a state.

14) Try to predict income using diferrent variables and also see if the number of immigrants and number of jobless immigrants affect the income of immigrants in a state.

Dealing with NAs: I removed the NAs in PINCP while doing any analysis on the income.

Dealing with income adjusted values: I adjusted the PINCP column using the ADJINC column.

Uninteresting and failed analyses:
I tried to predict the total income of immigrants in a state using total number of immigrants and jobless immigrants in that state. The r-squared value was 0.9934 which is really great, but the residual standard error was too big. I think that as the data was less(i.e 51 states), the model was overfitting the data.
Next I again tried to predict the income in a state using literacy and number of graduates in that state. I thought that more literate people earn more and the more the number of graduates in a state, more will be the total income of that state. But the r-squared was just 27%.
I also wanted to see Races of the immigrants in U.S , but I didn’t find anything interesting about that.
While exploring the data, I found that there were immigrants in armed forces of U.S, which I found very interesting, but that information was a bit irrelevant to me.

##Findings:
1) California has the most amount of immigrants.

```
#number of immigrants in each state.

states_df <- read.csv("https://raw.githubusercontent.com/plotly/datasets/master/2011_us_ag_ex
ports.csv")

statewise_immigrants<- non_citizen %>%
  inner_join(statepop,by="ST") %>%
  group_by(abbr) %>%
  summarise(immigrants=n()) %>%
  arrange(desc(immigrants)) %>%
  rename(state=abbr)

statewise_immigrants<-inner_join(statewise_immigrants,states_df , by=c('state'='code')) %>% s
elect(state,immigrants)

statewise_immigrants
```

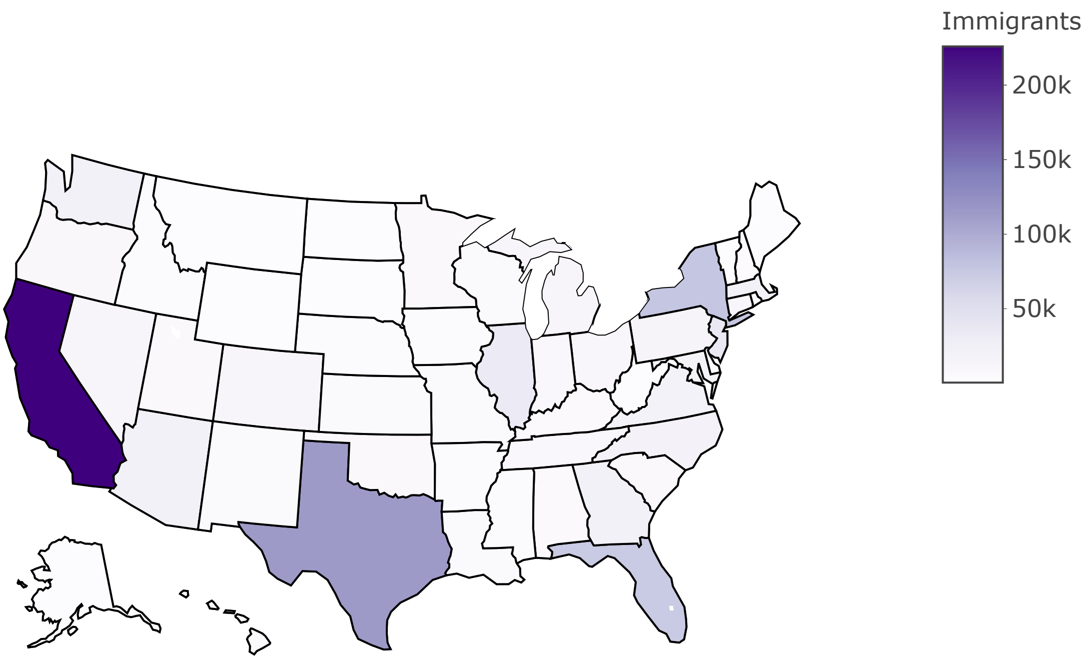
```
## # A tibble: 50 x 2
##    state immigrants
##    <chr>      <int>
##  1 CA        225964
##  2 TX        113504
##  3 NY         76376
##  4 FL         72484
##  5 NJ         35270
##  6 IL         32593
##  7 AZ         22053
##  8 GA         21944
##  9 MA         21848
## 10 WA         20448
## # … with 40 more rows
```

```
#plotting the data on U.S map.
statewise_immigrants$hover <- with(statewise_immigrants, paste(state, '<br>', "Immigrants ",
 immigrants))

l <- list(color = toRGB("white"), width = 2)

g <- list(
  scope = 'usa',
  projection = list(type = 'albers usa'),
  showlakes = TRUE,
  lakecolor = toRGB('white')
)
plot_geo(statewise_immigrants, locationmode = 'USA-states') %>%
  add_trace(
    z = ~immigrants, text = ~hover, locations = ~state,
    color = ~immigrants, colors = 'Purples'
  ) %>%
  colorbar(title = "Immigrants") %>%
  layout(
    title = 'Number of Immigrants in each state',
    geo = g
  )
```

Number of Immigrants in each state



By looking at the table and the map, we can see that California is the state that recieves most of the immigrants with over 225k. While Texas is the second state. Montana is the state that has the least number of immigrants with just 451 immigrants in 5 years.

2) Mexicans are the one’s that migrate to U.S the most, compared to all the people from other countries.

```
country_df <- read.csv('https://raw.githubusercontent.com/plotly/datasets/master/2014_world_g
dp_with_codes.csv') %>% rename(Country=COUNTRY)
countries <- "code,Country\n001,alabama\n002,alaska\n004,arizona\n005,arkansas\n006,californi
a\n008,colorado\n009,connecticut\n010,delaware\n011,district of columbia\n012,florida\n013,ge
orgia\n015,hawaii\n016,idaho\n017,illinois\n018,indiana\n019,iowa\n020,kansas\n021,kentucky\n
022,louisiana\n023,maine\n024,maryland\n025,massachusetts\n026,michigan\n027,minnesota\n028,m
ississippi\n029,missouri\n030,montana\n031,nebraska\n032,nevada\n033,new hampshire\n034,new j
ersey\n035,new mexico\n036,new york\n037,north carolina\n038,north dakota\n039,ohio\n040,okla
homa\n041,oregon\n042,pennsylvania\n044,rhode island\n045,south carolina\n046,south dakota\n0
47,tennessee\n048,texas\n049,utah\n050,vermont\n051,virginia\n053,washington\n054,west virgin
ia\n055,wisconsin\n056,wyoming\n060,american samoa\n066,guam\n069,commonwealth of the norther
n mariana islands\n072,puerto rico\n078,us virgin islands\n100,albania\n102,austria\n103,belg
ium\n104,bulgaria\n105,czechoslovakia\n106,denmark\n108,finland\n109,france\n110,germany\n11
6,greece\n117,hungary\n118,iceland\n119,ireland\n120,italy\n126,netherlands\n127,norway\n128,
poland\n129,portugal\n130,azores islands\n132,romania\n134,spain\n136,sweden\n137,switzerland
\n138,\"united kingdom, not specified\"\n139,england\n140,scotland\n147,yugoslavia\n148,czech
republic\n149,slovakia\n150,bosnia and herzegovina\n151,croatia\n152,macedonia\n154,serbia\n1
56,latvia\n157,lithuania\n158,armenia\n159,azerbaijan\n160,belarus\n161,georgia\n162,moldova
\n163,russia\n164,ukraine\n165,ussr\n168,montenegro\n169,\"other europe, not specified\"\n20
0,afghanistan\n202,bangladesh\n203,bhutan\n205,myanmar\n206,cambodia\n207,china\n208,cyprus\n
209,hong kong\n210,india\n211,indonesia\n212,iran\n213,iraq\n214,israel\n215,japan\n216,jorda
n\n217,korea\n218,kazakhstan\n222,kuwait\n223,laos\n224,lebanon\n226,malaysia\n229,nepal\n23
1,pakistan\n233,philippines\n235,saudi arabia\n236,singapore\n238,sri lanka\n239,syria\n240,t
aiwan\n242,thailand\n243,turkey\n245,united arab emirates\n246,uzbekistan\n247,vietnam\n248,y
emen\n249,asia\n253,\"south central asia, not specified\"\n254,\"other asia, not specified
\"\n300,bermuda\n301,canada\n303,mexico\n310,belize\n311,costa rica\n312,el salvador\n313,gua
temala\n314,honduras\n315,nicaragua\n316,panama\n321,antigua & barbuda\n323,bahamas\n324,barb
ados\n327,cuba\n328,dominica\n329,dominican republic\n330,grenada\n332,haiti\n333,jamaica\n33
9,st. lucia\n340,st. vincent & the grenadines\n341,trinidad & tobago\n343,west indies\n34
4,\"caribbean, not specified\"\n360,argentina\n361,bolivia\n362,brazil\n363,chile\n364,colomb
ia\n365,ecuador\n368,guyana\n369,paraguay\n370,peru\n372,uruguay\n373,venezuela\n374,south am
erica\n399,\"americas, not specified\"\n400,algeria\n407,cameroon\n408,cabo verde\n412,congo
\n414,egypt\n416,ethiopia\n417,eritrea\n420,gambia\n421,ghana\n423,guinea\n427,kenya\n429,lib
eria\n430,libya\n436,morocco\n440,nigeria\n444,senegal\n447,sierra leone\n448,somalia\n449,so
uth africa\n451,sudan\n453,tanzania\n454,togo\n457,uganda\n459,democratic republic of congo
 (zaire)\n460,zambia\n461,zimbabwe\n462,africa\n463,\"eastern africa, not specified\"\n46
4,\"northern africa, not specified\"\n467,\"western africa, not specified\"\n468,\"other afri
ca, not specified\"\n501,australia\n508,fiji\n511,marshall islands\n512,micronesia\n515,new z
ealand\n523,tonga\n527,samoa\n554,\"other us island areas, oceania, not specified, or at sea
\"\n"
country<- read_csv(countries) %>% rename(POBP=code)
country$Country<- tools::toTitleCase(country$Country)

world_map_countries<- inner_join(country,country_df,by="Country")
pob_of_immigrants<- non_citizen %>%
  inner_join(country,by="POBP") %>%
  group_by(Country) %>%
  summarise(Population=n()) %>%
  arrange(desc(Population))
pob_of_immigrants
```

```
## # A tibble: 159 x 2
##    Country            Population
##    <chr>                   <int>
##  1 Mexico                 305425
##  2 India                   51350
##  3 China                   46473
##  4 El Salvador             30963
##  5 Philippines             27679
##  6 Canada                  23167
##  7 Guatemala               22537
##  8 Korea                   18070
##  9 Cuba                    17813
## 10 Dominican Republic      16325
## # … with 149 more rows
```

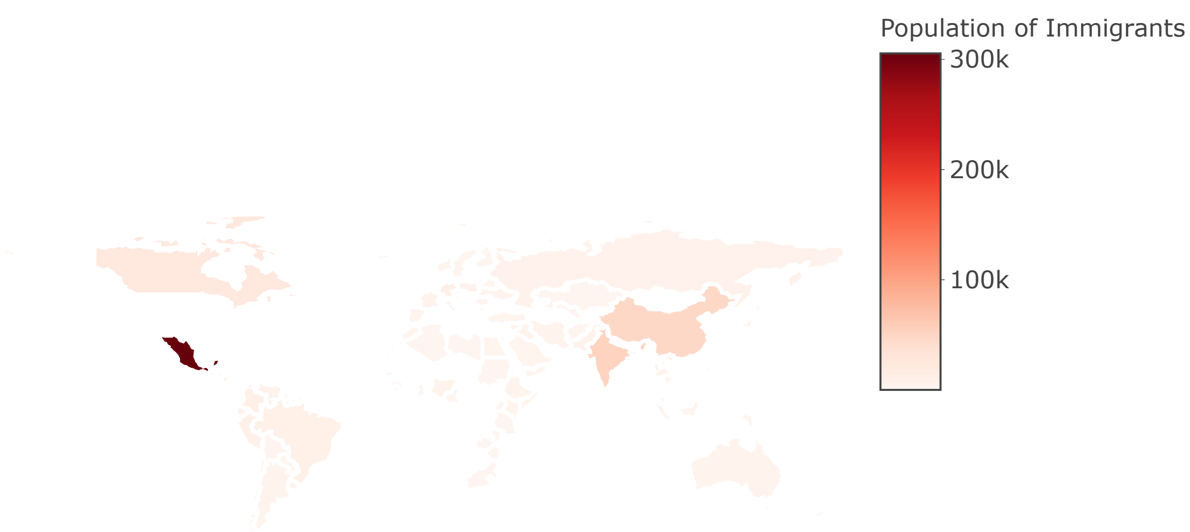
```
pob_of_immigrants<- inner_join(world_map_countries,pob_of_immigrants)
pob_of_immigrants<- pob_of_immigrants %>%  arrange(desc(Population))

l1 <- list(color = toRGB("grey"), width = 0.5)

# specify map projection/options
g1 <- list(
  showframe = FALSE,
  showcoastlines = FALSE,
  projection = list(type = 'Mercator')
)

plot_geo(pob_of_immigrants) %>%
  add_trace(
    z = ~Population, color = ~Population, colors = 'Reds',
    text = ~Country, locations = ~CODE, marker = list(line = l)
  ) %>%
  colorbar(title = 'Population of Immigrants') %>%
  layout(
    title = "Number of Immigrants in U.S from each Country",
    geo = g1
  )
```

Number of Immigrants in U.S from each Country



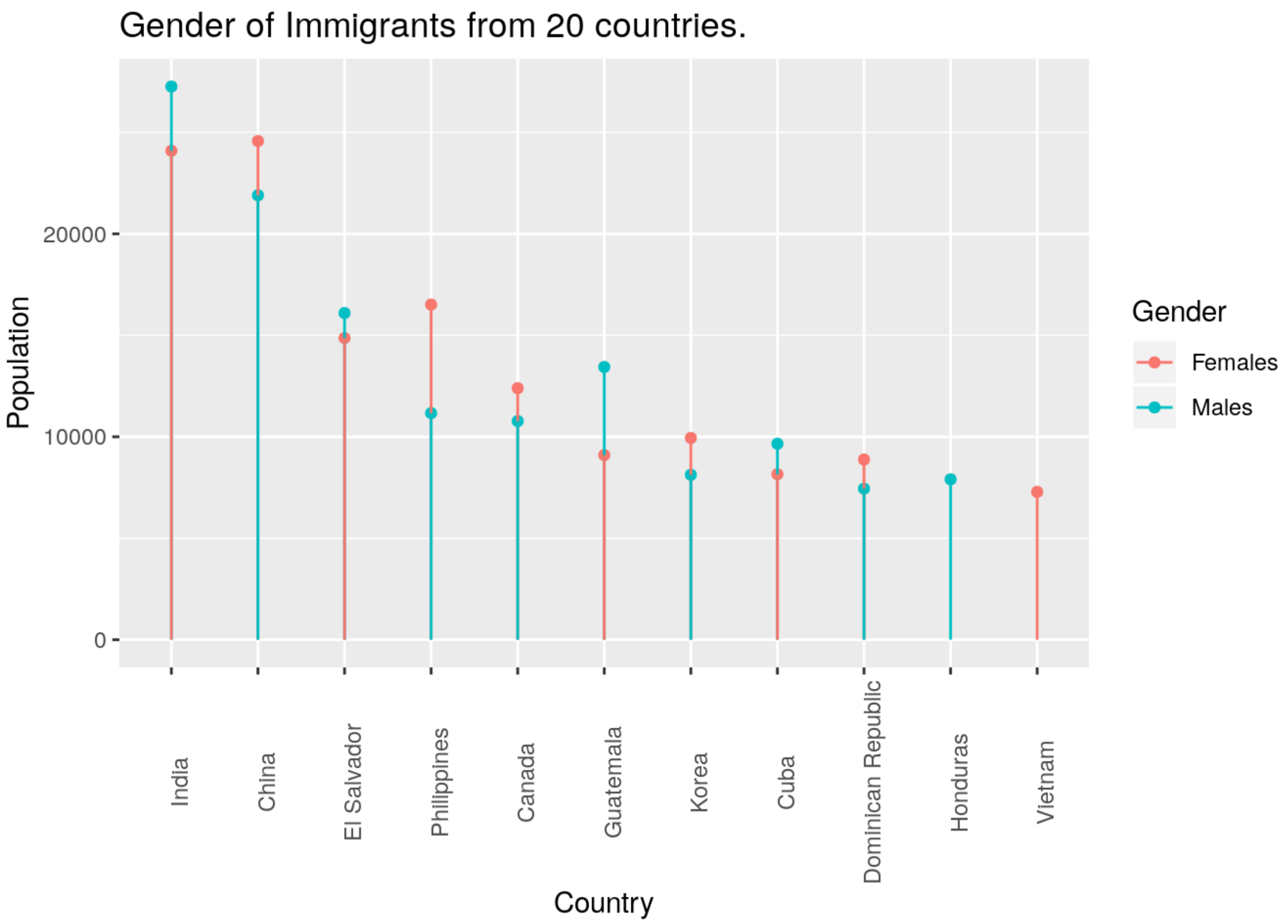
Mexicans migrate the most to U.S. There is a huge difference in the numbers. Even in Mexico and India, there is a difference of approximately 250k, in other words, there are 6x Mexicans in U.S compared to Indians.

3) More females migrates from China, Phillipines and Korea.

```
#number of male and female from each country.
gender_of_immigrants<- states_noncitizen %>%
  inner_join(country,by="POBP") %>%
  filter(!Country=="Mexico") %>%
  mutate(Gender=case_when(SEX == 1 ~ "Males",
                          SEX == 2 ~ "Females")) %>%
  group_by(Country,Gender) %>%
  summarise(n=n()) %>%
  arrange(desc(n)) %>%
  rename(Count=n) %>%
  head(20)
gender_of_immigrants
```

```
## # A tibble: 20 x 3
## # Groups:   Country [11]
##    Country            Gender  Count
##    <chr>              <chr>   <int>
##  1 India              Males   27259
##  2 China              Females 24575
##  3 India              Females 24091
##  4 China              Males   21898
##  5 Philippines        Females 16514
##  6 El Salvador        Males   16097
##  7 El Salvador        Females 14866
##  8 Guatemala          Males   13438
##  9 Canada             Females 12394
## 10 Philippines        Males   11165
## 11 Canada             Males   10773
## 12 Korea              Females  9942
## 13 Cuba               Males    9658
## 14 Guatemala          Females  9099
## 15 Dominican Republic Females  8880
## 16 Cuba               Females  8155
## 17 Korea              Males    8128
## 18 Honduras           Males    7903
## 19 Dominican Republic Males    7445
## 20 Vietnam            Females  7286
```

```
#plotting.
ggplot(gender_of_immigrants, aes(x=fct_reorder(Country,desc(Count)),Count, color=Gender)) +
  geom_point() +
  geom_segment( aes(x=Country, xend=Country, y=0, yend=Count)) +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Country") + ylab("Population") + labs(title = "Gender of Immigrants from 20 countrie
s.")
```



In this code, I removed the number of Mexican immigrants because there was a huge difference and the plot was not easily interpretable. From the above plot we can see, that number of females are more in the countries China,Phillipines,Korea and Dominican Republic. An interesting thing to notice is that, from Honduras, there is unsimilarity in the number of males and females. There are less than 7k females from Honduras that migrate to U.s. And the same is for the males from Vietnam.

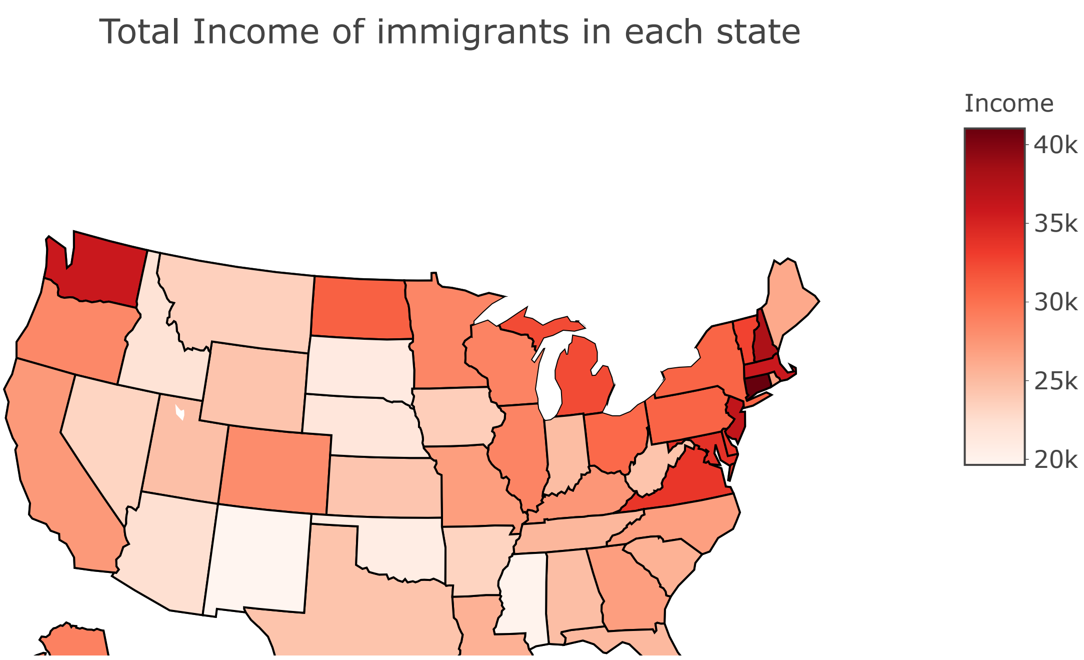
4)District of Columbia(DC) and Connecticut(CT) are the states with highest Per Capita Income of immigrants.

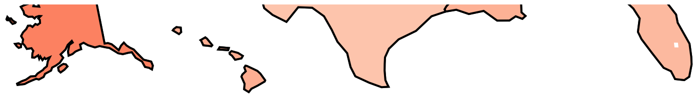
```
#calculating Per capita income of immigrants in each state.
statewise_income<- states_noncitizen %>%
  filter(!is.na(PINCP))  %>%
  group_by(abbr) %>%
  summarise(count=n(),Per_capita_income=sum(as.numeric(PINCP))/count) %>%
  arrange(desc(Per_capita_income)) %>%
  rename(state=abbr)
statewise_income
```

```
## # A tibble: 51 x 3
##    state count Per_capita_income
##    <chr> <int>             <dbl>
##  1 DC     2158            49081.
##  2 CT     9423            41032.
##  3 NH     1372            37829.
##  4 NJ    32931            36459.
##  5 WA    19134            35824.
##  6 MA    20478            35819.
##  7 MD    15936            33876.
##  8 DE     1571            33497.
##  9 VA    17350            33478.
## 10 VT      471            32652.
## # … with 41 more rows
```

```
statewise_income<-inner_join(statewise_income,states_df,by=c("state"="code")) %>% select(stat
e,Per_capita_income)

#plotting on the U.S map
statewise_income$hover <- with(statewise_income, paste(state, '<br>', "Income", Per_capita_in
come))
l3 <- list(color = toRGB("white"), width = 2)
g3 <- list(
  scope = 'usa',
  projection = list(type = 'albers usa'),
  showlakes = TRUE,
  lakecolor = toRGB('white')
)
plot_geo(statewise_income, locationmode = 'USA-states') %>%
  add_trace(
    z = ~Per_capita_income, text = ~hover, locations = ~state,
    color = ~Per_capita_income, colors = 'Reds'
  ) %>%
  colorbar(title = "Income") %>%
  layout(
    title = 'Total Income of immigrants in each state',
    geo = g3
  )
```





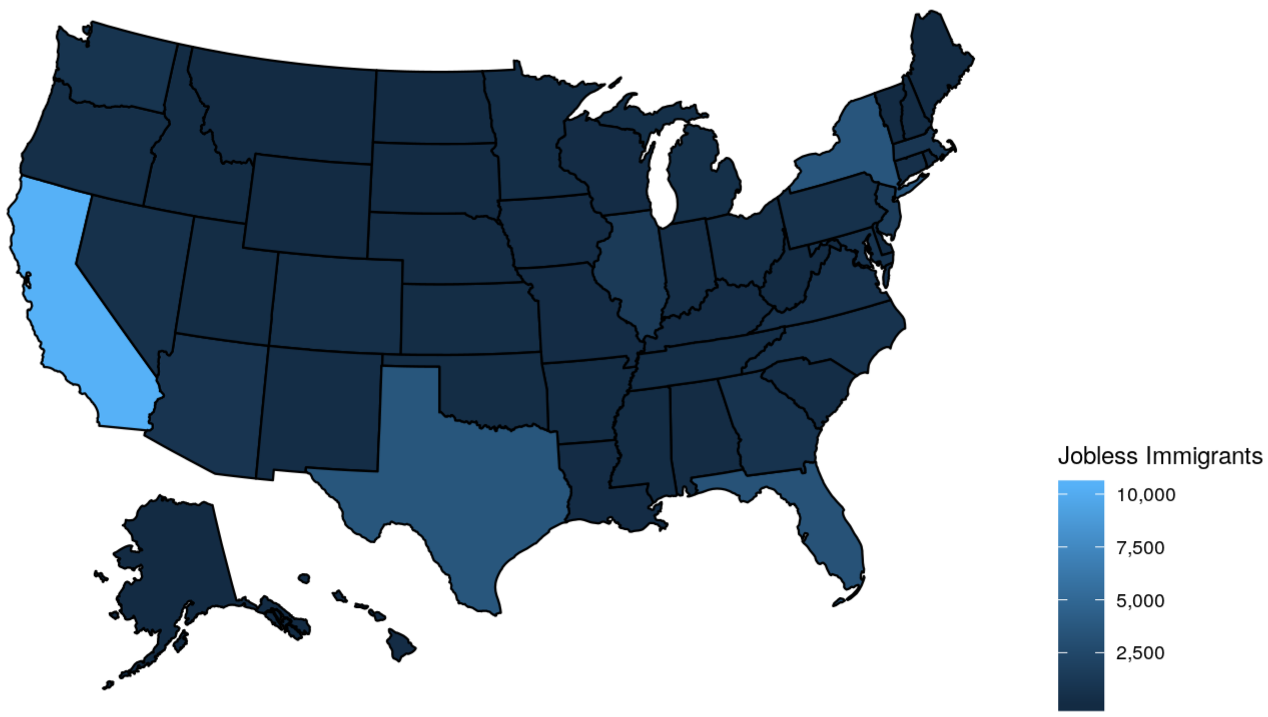
In the table we can see that immigrants in District of Columbia(DC) have the highest income. But there is no information of DC in the map. This is because in “states_df” there is no data of DC. Therefore, in the map, Connecticut(CT) is shown as the state with highest income.

5) California has most amount of immigrants, but it has the most amount of Jobless Immigrants too!

```
#jobless immigrants in each state.
jobless_immigrants <- non_citizen %>%
  inner_join(statepop,by="ST") %>%
  filter(ESR==3) %>%
  group_by(abbr) %>%
  summarise(jobless=n()) %>%
  arrange(desc(jobless)) %>%
  rename(state=abbr)
jobless_immigrants
```

```
## # A tibble: 51 x 2
##    state jobless
##    <chr>   <int>
##  1 CA      10373
##  2 TX       3581
##  3 NY       3554
##  4 FL       3279
##  5 NJ       1482
##  6 IL       1297
##  7 MA        961
##  8 AZ        825
##  9 WA        782
## 10 GA        723
## # … with 41 more rows
```

```
#plotting it on a static U.S map
plot_usmap(data = jobless_immigrants, values = "jobless", color = "black") +
  scale_fill_continuous(name = "Jobless Immigrants", label = scales::comma) +
  theme(legend.position = "right")
```



In this U.S map, it can be seen that California(CA) has the most amount of Jobless Immigrants. I defined people as Jobless by using “ESR==3” (Employment Status Recode). The value “3” means Unemloyed.. Texas also has a large amount of jobless people(around 3.5k).The least amount of jobless immigrants are in Alaska(AK), but that is because the total number of immigrants are very less too.

6) People from China migrate the most to U.S to pursue a Bachelor’s or higher level degree.

```
#immigrants currently pursuing bachelor's or higher level degree from U.S.
immigrants_studying<- non_citizen %>%
  inner_join(country,by="POBP") %>%
  filter(SCHG>=15) %>%
  group_by(Country) %>%
  summarise(Population=n()) %>%
  arrange(desc(Population))
immigrants_studying
```

```
## # A tibble: 159 x 2
##    Country      Population
##    <chr>             <int>
##  1 China             12220
##  2 Mexico            10701
##  3 India              5228
##  4 Korea              3946
##  5 Philippines        2186
##  6 Canada             1874
##  7 Vietnam            1857
##  8 Brazil             1558
##  9 Saudi Arabia       1371
## 10 Colombia           1311
## # … with 149 more rows
```

```
immigrants_studying<- inner_join(world_map_countries,immigrants_studying)
```

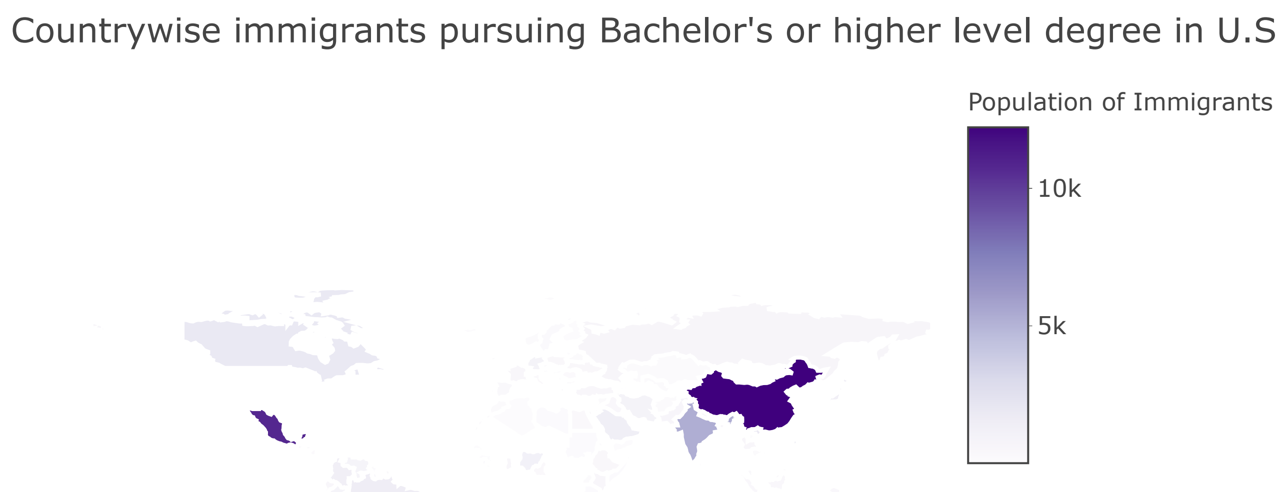
```
## Joining, by = "Country"
```

```
immigrants_studying<- immigrants_studying %>%  arrange(desc(Population)) %>% select(-GDP..BIL
LIONS.)

l4 <- list(color = toRGB("grey"), width = 0.5)

# specify map projection/options
g4 <- list(
  showframe = FALSE,
  showcoastlines = FALSE,
  projection = list(type = 'Mercator')
)

plot_geo(immigrants_studying) %>%
  add_trace(
    z = ~Population, color = ~Population, colors = 'Purples',
    text = ~Country, locations = ~CODE, marker = list(line = l)
  ) %>%
  colorbar(title = 'Population of Immigrants') %>%
  layout(
    title = "Countrywise immigrants pursuing Bachelor's or higher level degree in U.S",
    geo = g4
  )
```

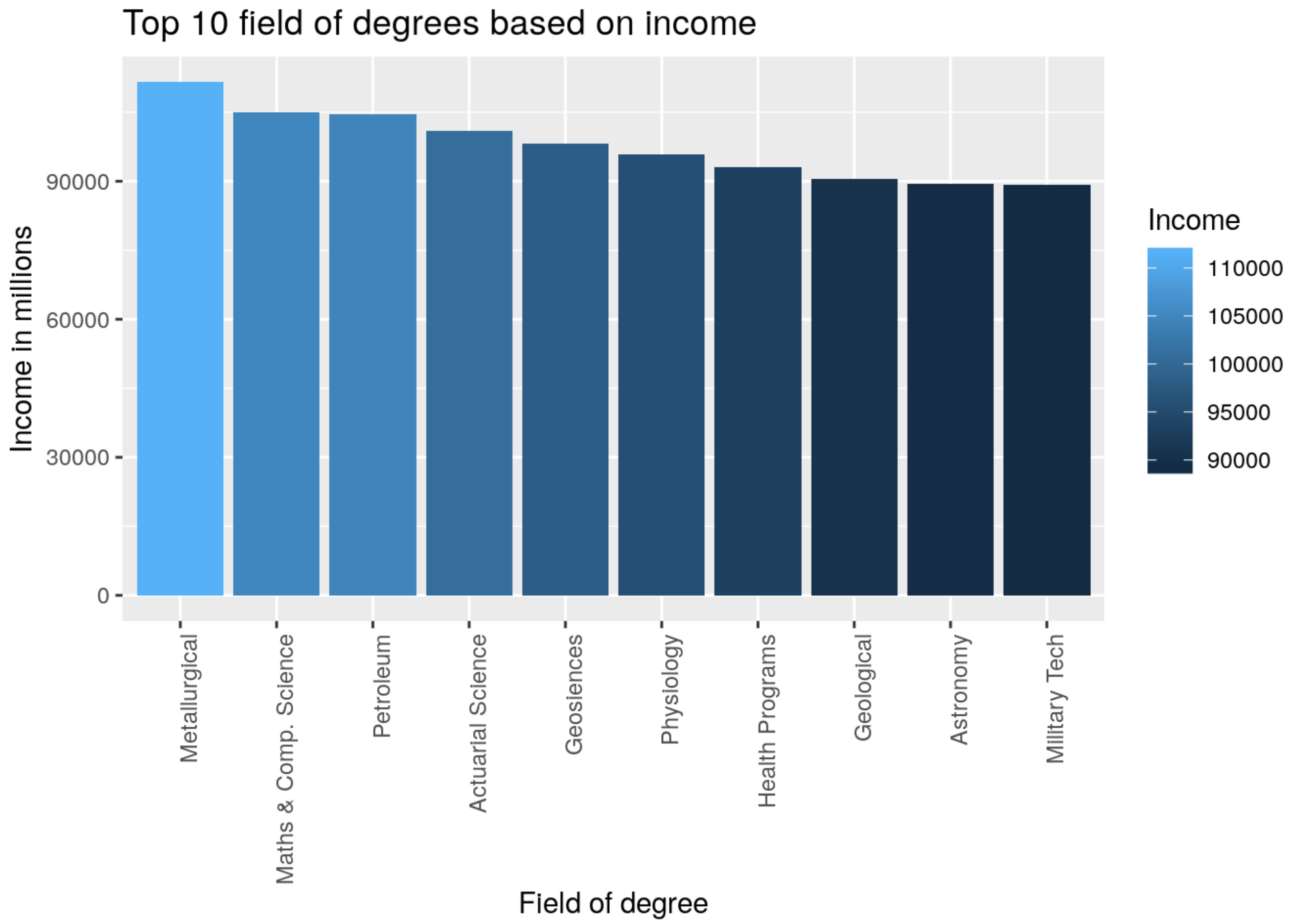


The most number of immigrants that migrate to U.S are from Mexico. So it can be said that most number of students must also be from Mexico. But it’s not true. China has the highest number of students that migrate to U.S.

7)People who have a degree in Mettalurgical Engineering have the highest income.

```
#top field of degree based on income.
fodwise_income<-states_noncitizen %>%
  filter(!is.na(FOD1P)) %>%
  group_by(FOD1P) %>%
  summarise(Number_of_People=n(), per_capita_income= sum(as.numeric(PINCP))/Number_of_People) %>%
  arrange(desc(per_capita_income)) %>%
  head(10)

#plotting the data.
ggplot(fodwise_income,aes(fct_reorder(FOD1P,desc(per_capita_income)),per_capita_income, fill= per_capita_income)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90,hjust = 1)) +
  xlab("Field of degree") +
  ylab("Income in millions") +
  scale_x_discrete(labels=c("Metallurgical","Maths & Comp. Science","Petroleum","Actuarial Science","Geosiences","Physiology","Health Programs","Geological","Astronomy","Military Tech")) +
  labs(title="Top 10 field of degrees based on income",fill="Income")
```

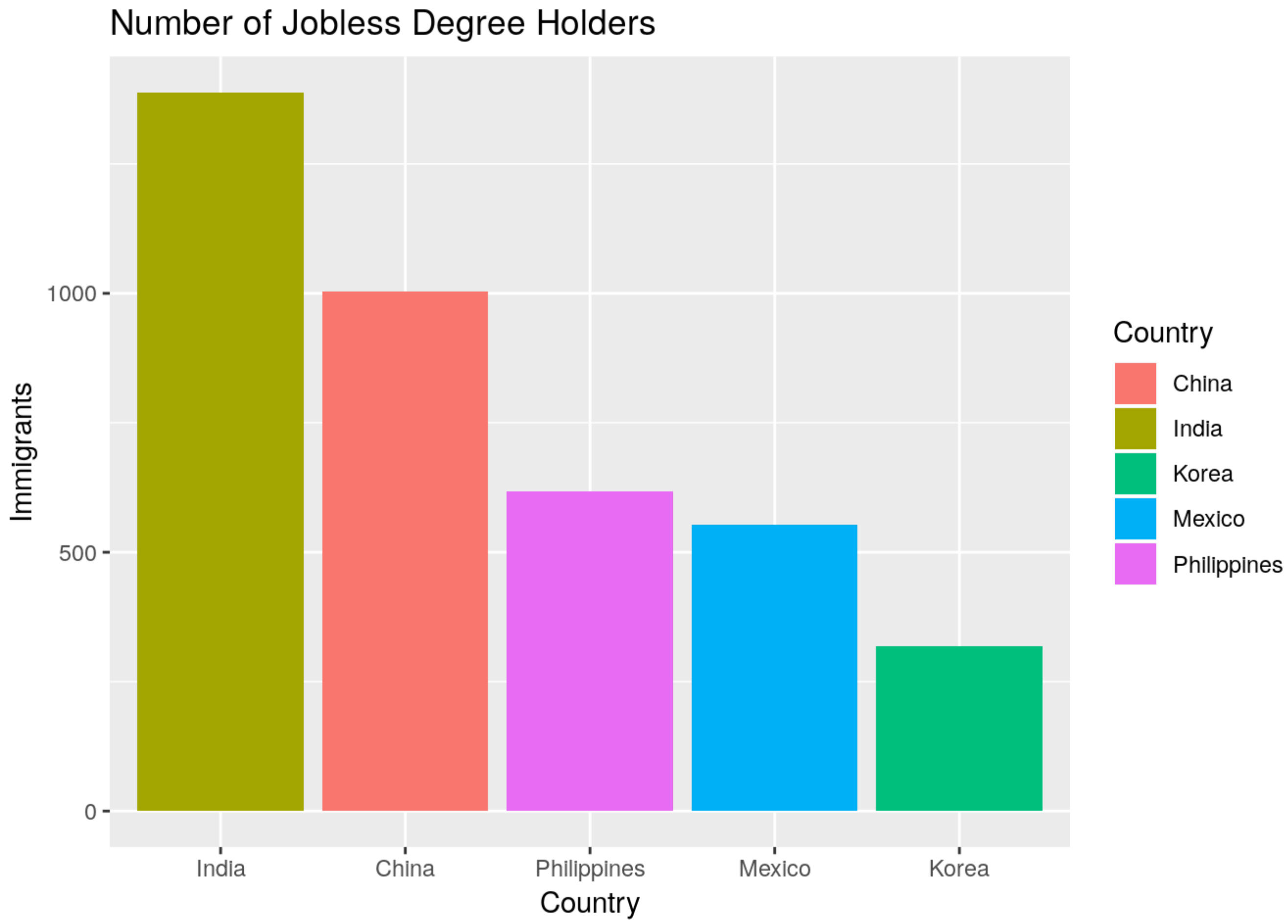


The above table shows that top 10 fields of degree based on the income by that degree holders. Metallurgical Engineering is field with highest income.

8) Indian students are the students who have a Degree and are looking for a job in U.S.

```
#nationality wise top 5: immigrants who have a Bachelor's or higher level degree and are look
ing for a job
jobless_degreeholder_immigrants <- states_noncitizen %>%
  inner_join(country,by="POBP") %>%
  filter(SCHL>=21 & NWLK==1) %>%
  group_by(Country) %>%
  summarise(Number_of_Immigrants=n()) %>%
  arrange(desc(Number_of_Immigrants)) %>%
  head(5)

#plotting
ggplot(jobless_degreeholder_immigrants,aes(fct_reorder(Country,desc(Number_of_Immigrants)),Nu
mber_of_Immigrants,fill=Country)) +
  geom_bar(stat="identity") + xlab("Country") + ylab("Immigrants") + labs(title="Number of Jo
bless Degree Holders")
```



The most number of jobless degree holders immigrants are from India. The number is not very large as compared to total immigrants from India.

9. Australian immigrants have the highest income in U.S compared to other immigrants.

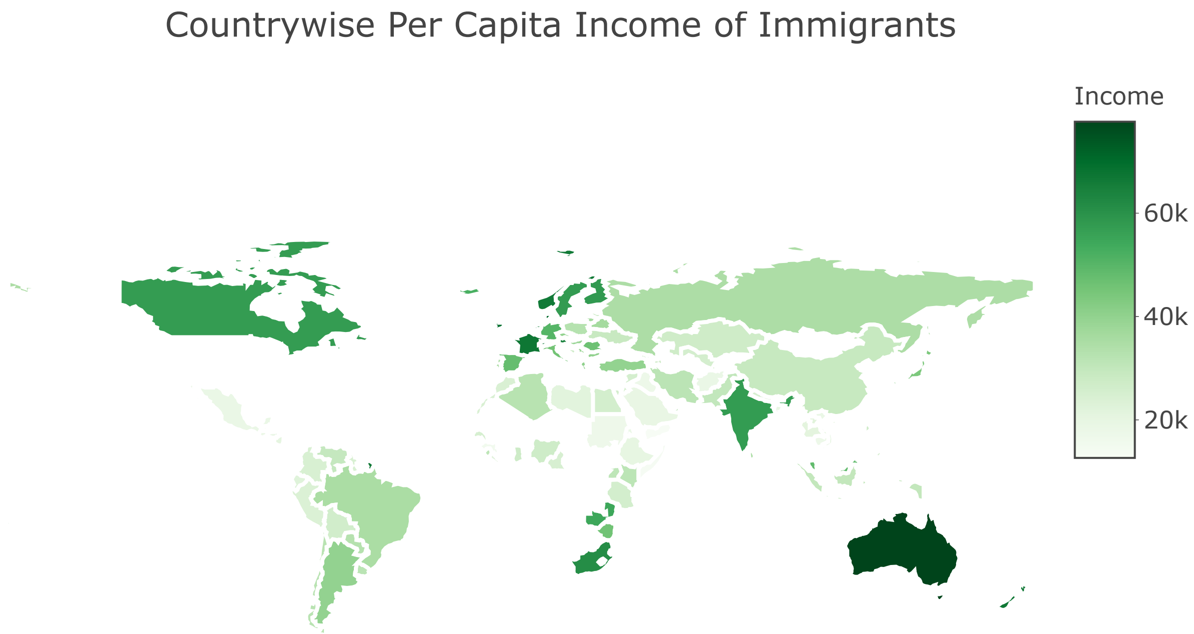
```
#income of immigrants.
pob_income<- non_citizen %>%
  filter(!is.na(PINCP)) %>%
  inner_join(country,by="POBP") %>%
  group_by(Country) %>%
  summarise(count=n(),Per_capita_income=sum(as.numeric(PINCP))/count) %>%
  arrange(desc(Per_capita_income))
pob_income <- inner_join(world_map_countries,pob_income) %>% select(-GDP..BILLIONS.) %>% arra
nge(desc(Per_capita_income)) %>% select(-c(POBP,count))
```

```
## Joining, by = "Country"
```

```
pob_income
```

```
## # A tibble: 128 x 3
##    Country     CODE  Per_capita_income
##    <chr>       <fct>             <dbl>
##  1 Australia   AUS              77700.
##  2 Belgium     BEL              71384.
##  3 Denmark     DNK              70552.
##  4 Switzerland CHE              69355.
##  5 Ireland     IRL              69223.
##  6 France      FRA              66881.
##  7 New Zealand NZL              66485.
##  8 Norway      NOR              66087.
##  9 Netherlands NLD              65353.
## 10 Cyprus      CYP              64679.
## # … with 118 more rows
```

```
#plotting the data on world map.
l2 <- list(color = toRGB("grey"), width = 0.5)
g2 <- list(
  showframe = FALSE,
  showcoastlines = FALSE,
  projection = list(type = 'Mercator')
)
plot_geo(pob_income) %>%
  add_trace(
    z = ~Per_capita_income, color = ~Per_capita_income, colors = 'Greens',
    text = ~Country, locations = ~CODE, marker = list(line = l)
  ) %>%
  colorbar(title = 'Income') %>%
  layout(
    title = "Countrywise Per Capita Income of Immigrants",
    geo = g2
  )
```



Australians have the highest income in U.S and people from european countries earn a lot too. Like Belgians, Danish, Swiss,Irish ,French, Norwegians etc.

10) The females with a degree in Metallurgical Engineering earn a lot less than males.

```
options(scipen = 999)
#income of males with a degree in Metallurgical Engineering.
male_income<- non_citizen %>%
  filter(!is.na(PINCP) & SEX==1 & FOD1P==2415) %>%
  select(PINCP,SEX,FOD1P)

#income of females with a degree in Metallurgical Engineering.
female_income<- non_citizen %>%
  filter(!is.na(PINCP) & SEX==2 & FOD1P==2415) %>%
  select(PINCP,SEX,FOD1P)

#testing the incomes of males and females.
t.test(male_income$PINCP,female_income$PINCP)
```

```
##
##  Welch Two Sample t-test
##
## data:  male_income$PINCP and female_income$PINCP
## t = 6.013, df = 55.718, p-value = 0.0000001471
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   51869.41 103704.50
## sample estimates:
## mean of x mean of y
## 119888.01  42101.06
```

I ran a t-test to check the means of incomes of Females and Males with Metallurgical engineering. As the variance of both of them were different, therefore a Welch Test was run by default.The p-value that I got is way lower than the significant level(0.05). Therefore the null hypothesis that the means are equal, is rejected.

Dicussion:

The main aim of my project was to do an exploratory analysis in immigrants in U.S. So, I selected the colummns that I felt would contribute to find something interesting about the trends of immigrants in U.S. The first thing I found was that most of the immigrants from all around the migrated to California state. And the least went to Montana state.
Next, Mexicans migrate the most to U.S and second to Mexicans are the immigrants from India .
I checked the gender of immigrants from all the countries and there were some interesting thihngs about the data.I found out that females from China, Canada, Korea , Phillipines and some other countries migrate to U.S more than the males of the country.
Although the number of immigrants are the highest in California, still the per capita income of immigrants is the highest in District of Columbia(DC).
After looking at the income, I thought I should look that people with which degree earn the most. So I found out

that Metallurgical Engineering degree holders have the highest income in United States of America.
Then I found out that the highest amount of jobess imimigrants are in California state.
I wanted to check that from which country, most students go to U.S to pursue a Bachelor's or a higher level degree. I found a very interesting thing that even though the total number of immigrants in U.S were the highest from Mexico, still the highest number of students immigrated from India. Then I found out that, Indians are the one's who have a degree and are still finding a job in U.S. This doesn't necessarily means that they pursued the degree from U.S itself. There may be a possibility that the students have completed their degree from India and are looking for job opportunities in U.S.
Then as I went on the exploratory analysis, I thought of finding immigrants from which country earn the most in U.s. Australians are the one's who earn the most. Except them, immigrants from europe also earn a lot, as compared to immigrants from other countries.
When I was finding the highest earning field of degree,a thought came to my mind that I should check whether there is some bias in the income based on the gender of a person.And I found out that males earn more than double the amount females earn.
I am quite confident that the analysis that I have done is accurate and I believe most of my conclusions.Although,I am not confident enough in my linear models, because the results were not as expected by me.
The limitations in my project is that I couldn't predict the income of immigrants based on various variables.

Appendix:

1) Predicting income based on various variables:

```
#filtering graduates from immigrants data.
grads<- states_noncitizen %>%
  filter(SCHL>=21) %>%
  group_by(abbr)

#income of grads in each state.
grads_income<- grads %>% filter(!is.na(PINCP)) %>% summarise(number_of_grads=n(),literacy=sum
(as.numeric(SCHL)),per_capita_income = sum(as.numeric(PINCP))/number_of_grads)%>% arrange(des
c(literacy)) %>% rename(state=abbr)
grads_income
```

```
## # A tibble: 51 x 4
##    state number_of_grads literacy per_capita_income
##    <chr>           <int>    <dbl>             <dbl>
##  1 CA              44357   958606            62509.
##  2 TX              19111   413432            55622.
##  3 NY              18678   404367            63256.
##  4 FL              15828   341306            47209.
##  5 NJ              11595   250611            65470.
##  6 IL               8129   176138            53228.
##  7 MA               7672   168507            62769.
##  8 VA               6325   137175            56710.
##  9 WA               5739   124265            70304.
## 10 MD               5454   119295            56700.
## # … with 41 more rows
```

Calculating literacy rates in each state.

```
#Literacy in each state
literacy_in_state<- grads_income %>% select(state,literacy)
literacy_in_state
```

```
## # A tibble: 51 x 2
##    state literacy
##    <chr>    <dbl>
##  1 CA      958606
##  2 TX      413432
##  3 NY      404367
##  4 FL      341306
##  5 NJ      250611
##  6 IL      176138
##  7 MA      168507
##  8 VA      137175
##  9 WA      124265
## 10 MD      119295
## # … with 41 more rows
```

Predicting income using literacy.

```
income_literacy_lm <- lm(per_capita_income~literacy,grads_income)
summary(income_literacy_lm)
```

```
##
## Call:
## lm(formula = per_capita_income ~ literacy, data = grads_income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12398.1  -7266.6   -357.5   4803.2  26541.4
##
## Coefficients:
##                 Estimate   Std. Error t value             Pr(>|t|)
## (Intercept) 48903.165415  1375.964344  35.541 < 0.0000000000000002 ***
## literacy        0.023181     0.007737   2.996              0.00428 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8640 on 49 degrees of freedom
## Multiple R-squared:  0.1548, Adjusted R-squared:  0.1376
## F-statistic: 8.976 on 1 and 49 DF,  p-value: 0.004283
```

The adjusted r-squared is just 0.1376. This means that the model is just 13% confident.

```
income_numberofgrads_lm<- lm(per_capita_income~number_of_grads,grads_income)
summary(income_numberofgrads_lm)
```

```
##
## Call:
## lm(formula = per_capita_income ~ number_of_grads, data = grads_income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12408.8  -7275.2   -363.5   4823.0  26543.1
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     48914.3409  1375.8687  35.552 < 0.0000000000000002 ***
## number_of_grads     0.4996     0.1673   2.985              0.00441 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8645 on 49 degrees of freedom
## Multiple R-squared:  0.1539, Adjusted R-squared:  0.1366
## F-statistic: 8.913 on 1 and 49 DF,  p-value: 0.00441
```

While predicting using the number of graduates , the model was still just 13% confident.

```
income_literacy_grads_lm<- lm(per_capita_income~literacy+number_of_grads,grads_income)
summary(income_literacy_grads_lm)
```

```
##
## Call:
## lm(formula = per_capita_income ~ literacy + number_of_grads,
##     data = grads_income)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10769  -5585  -1267   3275  24630
##
## Coefficients:
##                  Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     47164.109   1370.312  34.419 < 0.0000000000000002 ***
## literacy            7.600      2.346   3.239              0.00218 **
## number_of_grads  -163.788     50.718  -3.229              0.00224 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7912 on 48 degrees of freedom
## Multiple R-squared:  0.3057, Adjusted R-squared:  0.2767
## F-statistic: 10.57 on 2 and 48 DF,  p-value: 0.0001576
```

When predicting the income using two variables, i.e. literacy and total number of grads in a state, the model was 27% confident.

This code gave something interesting results, because I didn’t think that a country would take immigrants in their armed forces.This is initeresting, but irrevelant to my project.

```
#number of people in armed forces in each state ,immigrants in armed forces?? \(0_o)/
statewise_armed_forces<- states_noncitizen %>%
  group_by(abbr) %>%
  summarise(armed_forces=n()) %>%
  arrange(desc(armed_forces)) %>%
  rename(state=abbr)
statewise_armed_forces
```

```
## # A tibble: 51 x 2
##    state armed_forces
##    <chr>        <int>
##  1 CA          225964
##  2 TX          113504
##  3 NY           76376
##  4 FL           72484
##  5 NJ           35270
##  6 IL           32593
##  7 AZ           22053
##  8 GA           21944
##  9 MA           21848
## 10 WA           20448
## # … with 41 more rows
```