Final Report - Shaima Shoukat

Introduction

The Public Use Micro Data Sample file(PUMS) contains around 5% of the actual survey data collected by the American Community Survey over a five year period(2012-2016). The variables present, cover the questions asked in the survey. Some of the variables are coded to preserve privacy of those participating in the survey. Each record is a weighted as some records were selected more often than others. The dataset used in this report contains variables from the population characteristics of the PUMS.It consists of 26 variables and 15681927 weighted records records.Some of the variables include Age, Race, Gender, Income, Earnings, Education Attainment, State, Occupation, Citizenship Status,etc.The appropriate pre processing techniques have been applied Appendix 1 and now we see some of the relationships in the data.

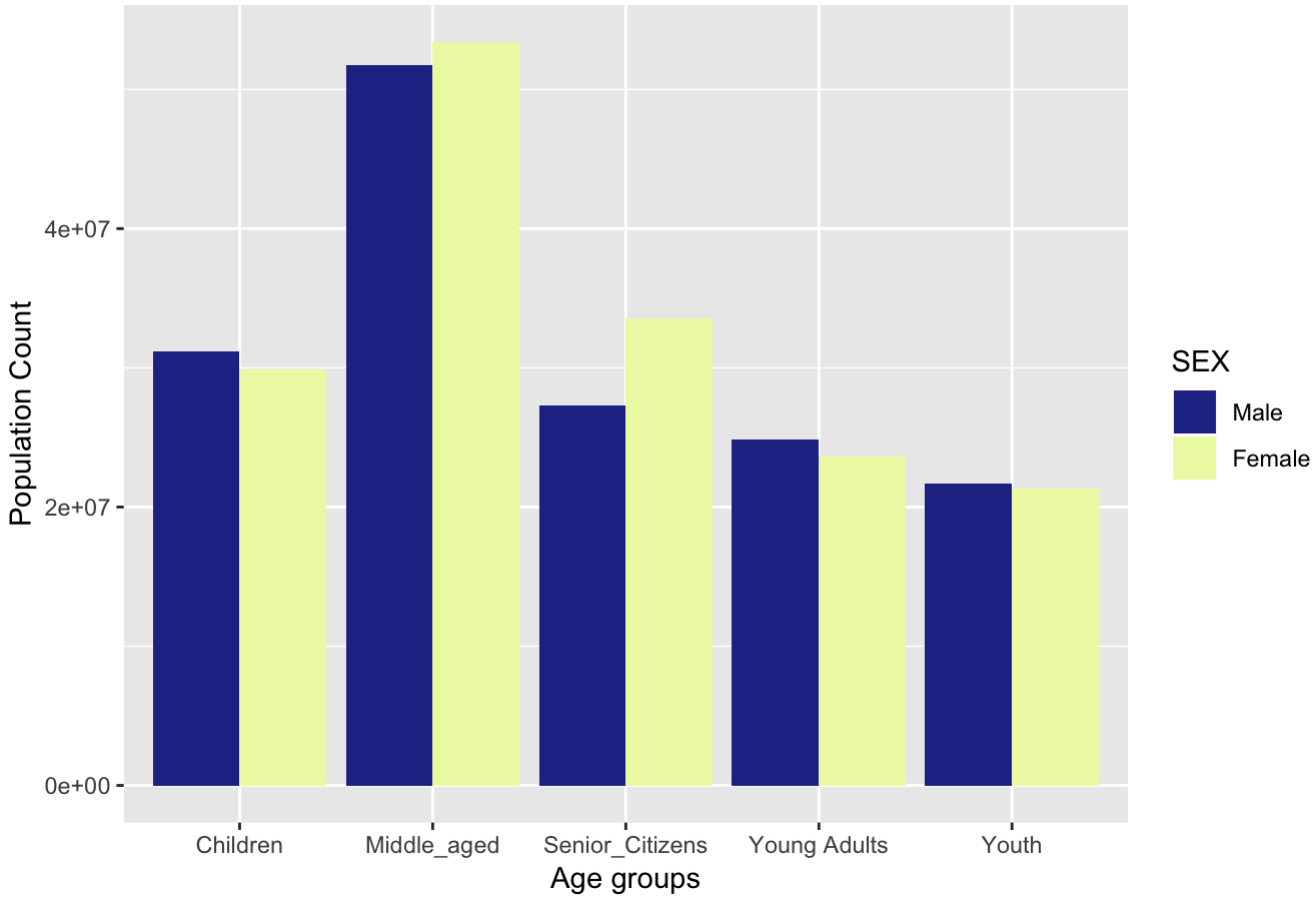
The proportion of males to females in the population is almost the same at 50%, given by the following Table 1.(Appendix [3])

Table1 : Proportion of Males and Females

Gender	Proportion
Male	0.4920909
Female	0.5079091

This distribution of males to females holds true across various age groups as well as can be seen from the the graph below(Fig1). It can also be observed that a greater share of the population belongs to the 35 to 45 age group(Middle age). Among the senior citizens, it seems like women tend to outlive men as the difference in bar heights seems significant.(Appendix [4])

Fig1: Population Distribution by age group



Next, we will explore how the work force is distributed across different industries.(Appendix [5])

Fig2 : Population Distribution by various Industries

Fig3 : Population Distribution by various Industries under the poverty line


```
## Warning in wordcloud(SOCP, count, max.words = 50, colors = brewer.pal(8, :
## OfficeWorkers could not be fit on page. It will not be plotted.
```

Fig4 : Population Distribution by various Industries over the poverty line

When the general population is considered, most people are working as Managers(7%), in Sales(8%) or as Office workers(10%) (Fig2). This distribution changes when we look for occupation of people living below the poverty line in the working age group (Fig3). A lot of the people below the poverty line are working as Cashiers in Sales industry;or as servers, bartenders, chefs in the food industry; or in the Transport Services. For people above the poverty line(Fig 4), the industry with the most number of employed people is Education, IT and Sales. Most people work as teachers, librarians, Computer Programmers, analysts,etc.

```
## Selecting by count
```

Table2 :Proportion of Population by Skilled Industry

SkilledIndustry	Proportion
	0.5729220
OfficeWorkers	0.1091328
Sales	0.0893012
Managers	0.0758514
TransportServices	0.0529528
Food	0.0509653
Education	0.0488745

One interesting thing that was found during exploration analysis how the mean earnings significantly varied for the general population, below the poverty line and above it. Shockingly, the average earning for people living in extreme poverty was a meagre 7496 dollars per annum (Table 3). On the other hand, for the proportion of people above the poverty line, the annual earnings was nearly 7 times as much as those below it(49000). The national average earnings is around 40000 dollars.(Appendix [7])

Table3 :Mean Earnings for people above and below poverty vs the national avergae

GeneralPop	BelowPoverty	AbovePoverty
39683.66	7496.618	48077.3

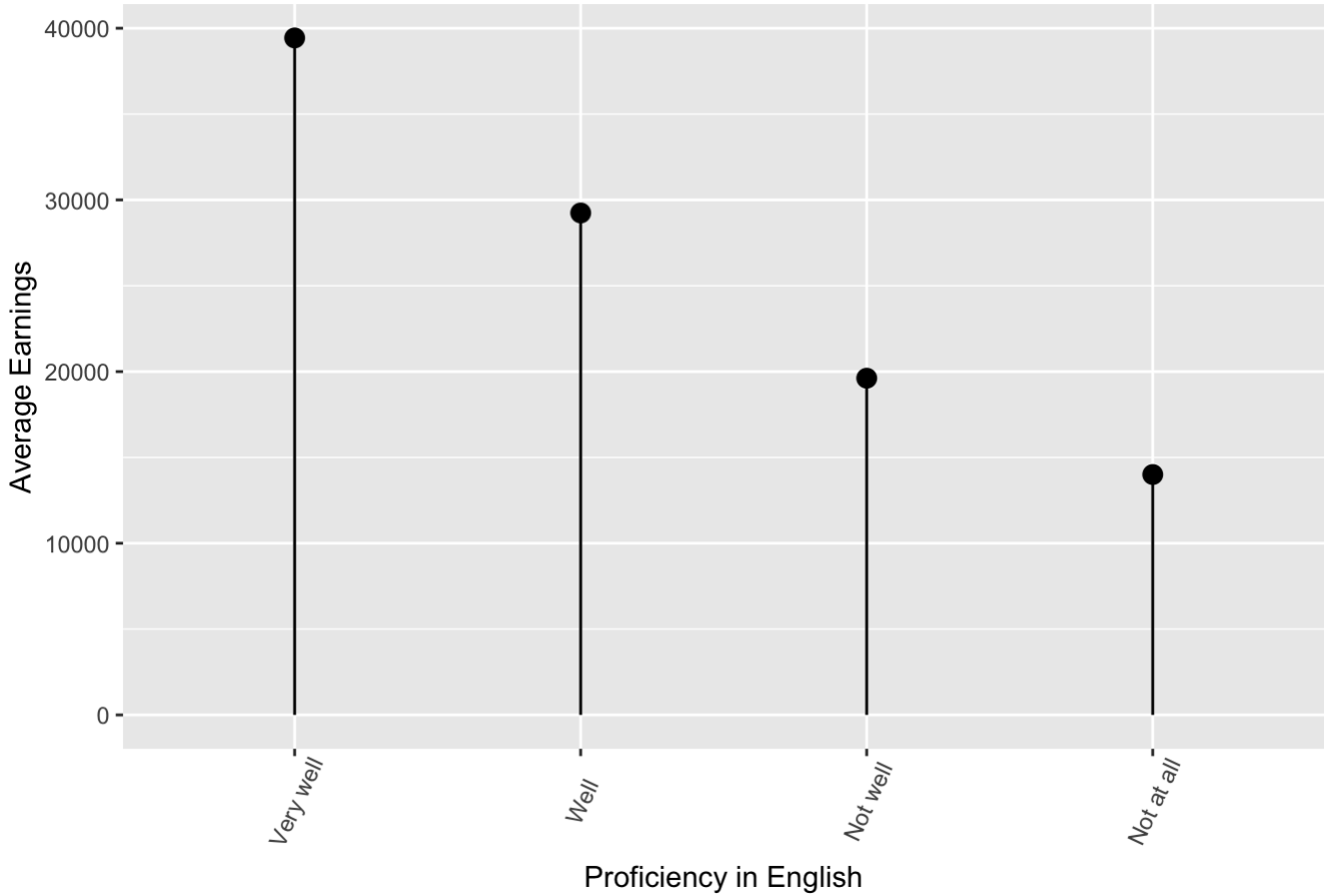
This indicated the presence of a relationship between earnings and poverty.However, the weighted correlation between yearly earnings and poverty was not as high as expected at around only 0.4 given in the table below:

Methodology

After plotting some graphs and going through the survey questions, one aspect I wanted to explore was if the level of proficiency in English effected earnings at all while living in the United States. I proceeded to plot the average earnings at various levels of proficiency from Not being able to speak at all to being very proficient(Fig 5). Before getting into the analysis, the PINCP of the yearly earnings needed to be adjusted for inflation. I did this by multiplying each PINCP value with the adjustment factor, ADJINC, and dividing by 1000000,ie, Earnings = PINCP * (ADJINC/1000000) It is important to adjust for inflation because the value of dollar is not constant and therfore income earned in one year could value less or more the next year. This would skew the results of analysis. Also, because the PUMS data is a weighted sample, each record

represents multiple records of the personal variables. In order to take this into account, each count or mean is computed as a weighted sum to account for the fact that some records were picked more often than others. Not taking weights into account would mean that each record is being considered as a single value. This significantly effects the analysis as exact value of the sample is not obtained and therefore any calculations involving the total number of observations would give incorrect results. In my analysis, the count values would be significantly less than when used with weights.

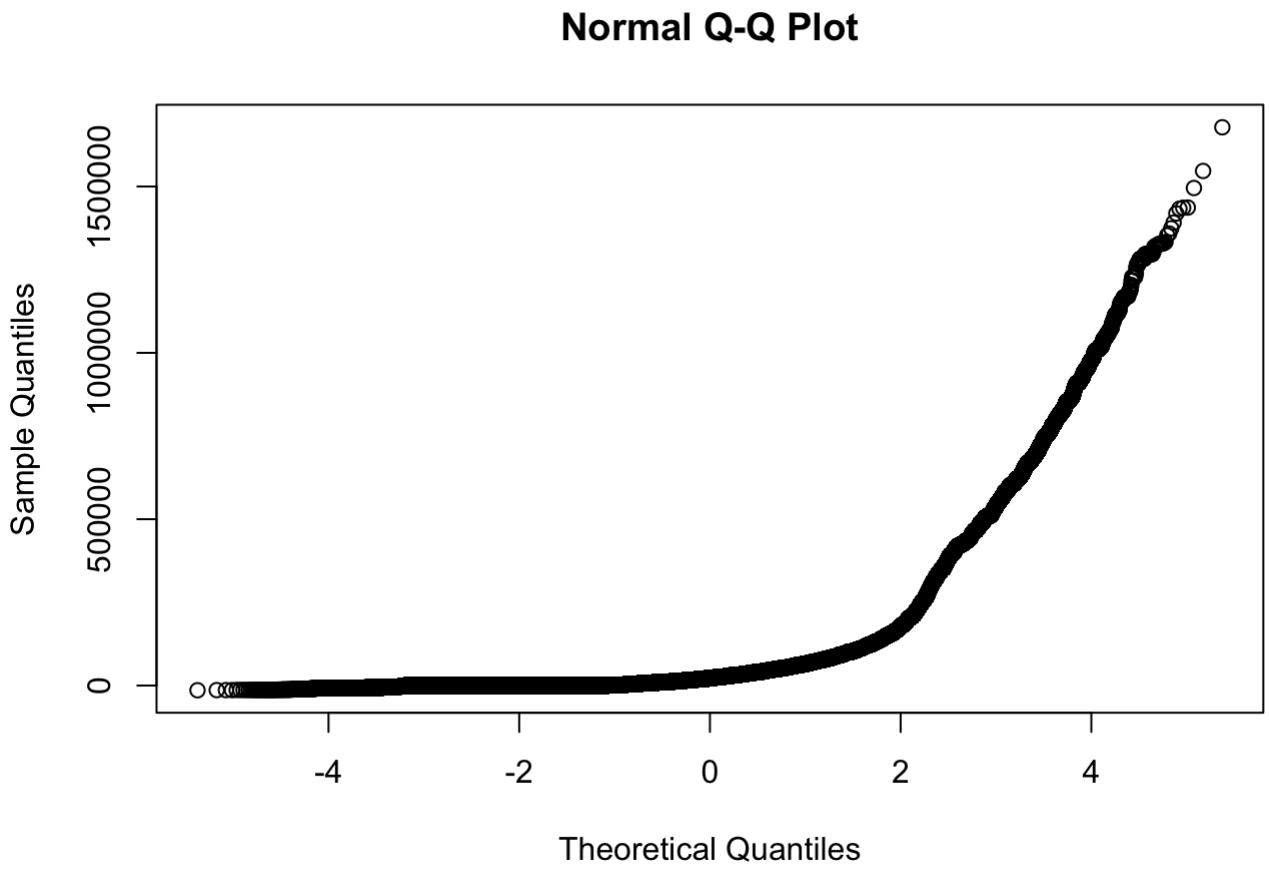
Fig5 : Relationship between Proficiency in English and its effect on Earnings



The Fig 5 illustrates that as the proficiency increased, the average earnings increased. For people not very proficient in English, the average was around 15000 dollars per annum. It increased to 20000 dollars for those who spoke better than the previous group. It was the highest at 40000 dollars for those who were very proficient. This relationship seemed to be exhibiting a strong linear trend.

The next step was to see if there was indeed a statistically significant difference in the mean earnings across various groups. I decided to run an Analysis of Covariance (ANCOVA) test for this. I wanted to see if this variation in earnings across various proficiency groups could be explained after controlling for covariates. Here, there seem to be two kinds of covariates. Age is the confounding covariate, which is not of interest and highest level of education attainment is the independent covariate. These included as independent variables to see if they improve the model's power and explain the variance. Before running the regression model, I collapsed the education attainment level into three a factor with 3 levels (No High school Diploma, High school diploma and Bachelors or more). Next, assumptions for Ancova were evaluated: 1) We assume the data is independent and there is no relationship in means of various groups.

2. The independent variable needs to be categorical and the dependent variable is continuous. This assumption holds as English proficiency is categorical and Earnings are continuous variable.
3. Normality: We draw a qqplot to check for normality.



The data do not seem to follow a straight line as required.But since, the assumption of normality is not very strict we can still proceed.

4. Homogeneity of variances : The variation of earnings for all four groups is within two times of the other except for the “Very Well” group. But, this is not too drastic (Table 4). So, we will assume equality of variances.

Table 4: Variation of income across the four groups

StandardDeviations	ENGgroups
18415.50	Not at all
26659.24	Not well
56198.92	Very well
39793.45	Well

5. There is at least one continuous covariate. This condition holds as age is the continuous covariate.
6. Homogeneity of slopes: The slopes for each group in the independent variable should be parallel and not interacting.
7. Linearity: The relationship between Age and Earnings is assumed to be linear.

The regression model is then run with Earnings as the dependent variable and age, Proficiency in English, Highest education attained as the independent and covariates, respectively.

Next, I wanted to explore the races where the proficiency in English was lowest. Because the results of this were interesting, I extracted the ancestry of the races with the highest number of people who did not soeak English well. In both of the these analyses the NA values were ignored. I decided not to remove NAs as in a lot of the cases NAs do not stand for missing values. Further, I selected the top four races common in ancestry of people with low proficiency in English. Then, I wanted to know if these people were citizen by birth or were naturalized. In this case, the NAs were not filtered as they stood for people who are not naturalized yet. This was particularly interesting because a huge chunk of people who could not speak English well were in fact not yet citizens of United States. For each of the top four ancestries I chose, I wanted to find the year of naturalization that was most popular. However, this did not give interesting results.]B] So, a new factor column was created with all people naturalized before 2000 and those after the year 2000, and those that have yet to be naturalised. With this, I stopped my analysis with the English proficiency variable.

My second analysis was based on the Poverty varibale POVPIP. I began by finding the the states with the most number of people living below the poverty line,ie, where the income to poverty ratio was less than 125.This led me to ask whether or not the people living in extreme poverty were working full time every week and what was the poorest ethnic group in each state. I proceeded to collapse the educational attainment variable,

SCHL, to a factor with three groups: 1) Those without a high school diploma 2) Those with a high school diploma 3) Those with a degree or more. I could now see the most common educational qualifications of the poor. I also wanted to know the proportion of ethnic groups at each level of education that did not have health insurance. Lastly, I compared how the working hours differed among those above and below poverty across each state.

Findings

English proficiency

After running a regression model for Earnings across English proficiency levels, with age and education as covariates, the following results were obtained(Appendix 11)

```
##
## Call:
## lm(formula = Earnings ~ AGEP + ENG + highest_edu_level, data = SSData2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -78357  -18849   -8104    9781 1644635
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   43289.063     95.712   452.3   <2e-16 ***
## AGEP                            357.461      1.739   205.5   <2e-16 ***
## ENGWell                       -7039.914     80.826   -87.1   <2e-16 ***
## ENGNot well                  -11707.321     95.132  -123.1   <2e-16 ***
## ENGNot at all                -16094.463    134.786  -119.4   <2e-16 ***
## highest_edu_levelHigh_schl   -31698.243     75.301  -421.0   <2e-16 ***
## highest_edu_levelno_high_schl -37601.236     85.920  -437.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47450 on 2347912 degrees of freedom
##   (13334008 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.1417
## F-statistic: 6.461e+04 on 6 and 2347912 DF,  p-value: < 2.2e-16
```

The Estimate coefficient: The first row of this coefficient is the intercept which tells us that the mean earnings for all groups of proficiency is `27194.600` . The remaining rows are slopes for each level of the independent variable. For example, AGEP slope of `357.461` tells that for every increase in age by one year, the earnings go up by `357.461` dollars per annum. Or, For an increase in proficiency from “Not at all” to “Not Well”, the earnings go up by 4387.142 dollars. Or, a decreasee in education level by one from “Degree or more” to “high School” would decrease the earnings by `31698.243` dollars annually.

The Std.Error Estimate: This estimate tells by how much the difference between the actual earnings at each level of proficiency and the predicted earnings vary. For the above case, where the increase in proficiency from “Not at all” to “Not well”, the earnings increase by 4387.142, there is a chance that the predicted value can vary by 144 dollars on either side of 4387.142.

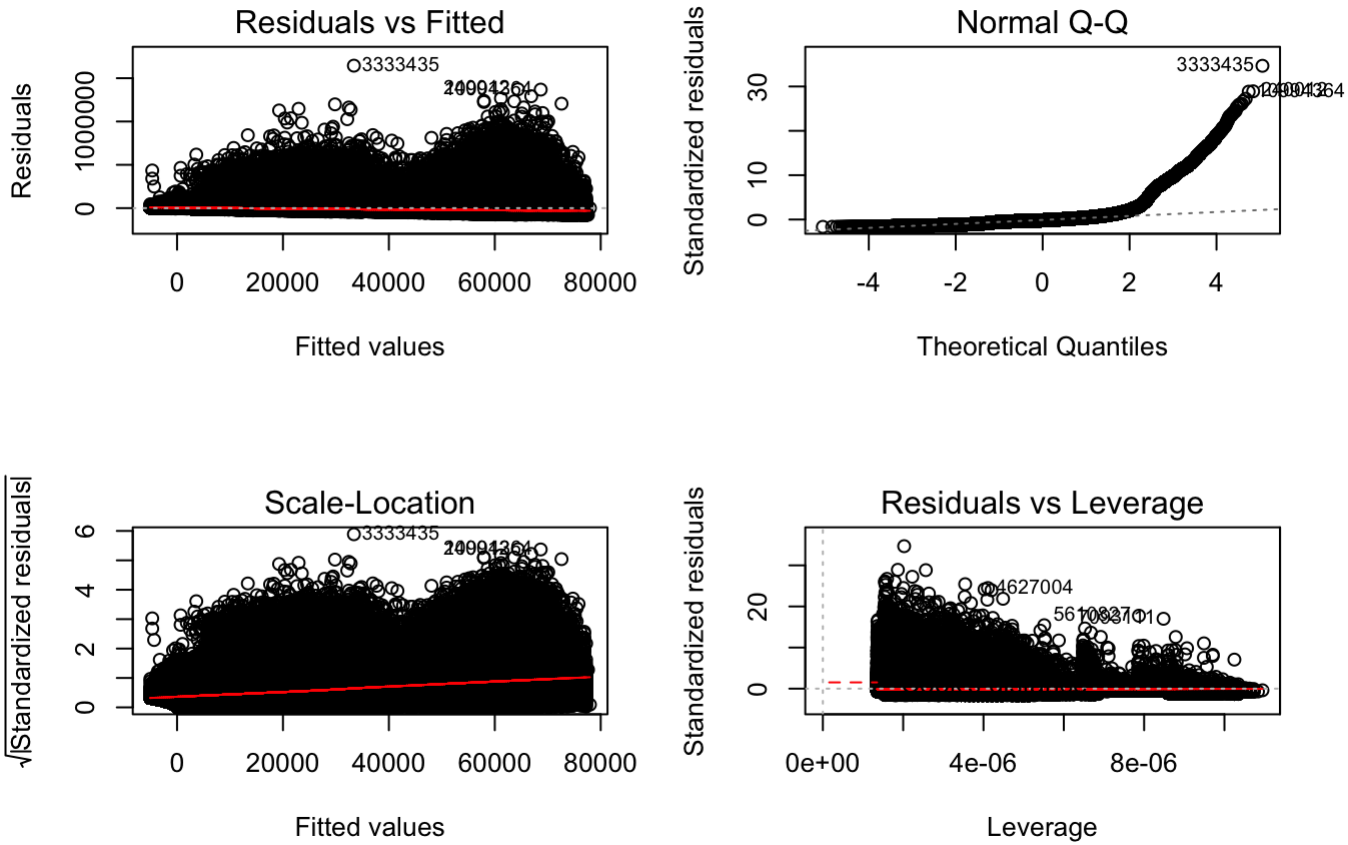
The t value coefficient: This is the gives us the number of standard devaitions, our coefficient estimate is far from 0. The bigger the number, the further it is away from zero, and the bigger the p is obtained.

The p coefficient: For the given predictor, Proficiency in English,since the p values of `2e-16` are less the significant level of 0.05.This tells us that there is a significant association between the predictor(proficiency in English) and the response variable (Earnings per annum). The p values for age and education level are less than 0.05 as well. This means that they are significant in the model as well. Therfore, it was justified to keep them in the model.

Correlation coefficient : This is the Adjusted R-squared value, 0.1417.This tells us the predicting power of the model. In this case, only 14 % of the variation in Earnings is accounted by English Proficiency, after controlling for age and education. Although this value increased from a mere 5%, when covariates were not considered to 14%, when covariates were controlled for, this is still not a great model.

The overall, p value is `2.2e-16` which is less than 0.05 significant level. Therefore, we conclude that there is a statistically significant increase in Earnings as the the abilty to speak English improves.

```
par(mfrow=c(2,2))
plot(lm.model)
```



In the Residuals vs Fitted plot, the residual movement follows the fit line. The residuals seem to be randomly distributed as no trends can be observed.

The normal QQ plot shows that the residual points are not normally distributed as there is a significant devaition from the dotted line. But since, normality is not strict in regression, we can still use the model.

The scale location plot is used to tell if the assumption of equal variance holds true. The red line is not exactly horizontal, this indicates the residual spread is not exactly uniform across the predictor range but not too drastic.

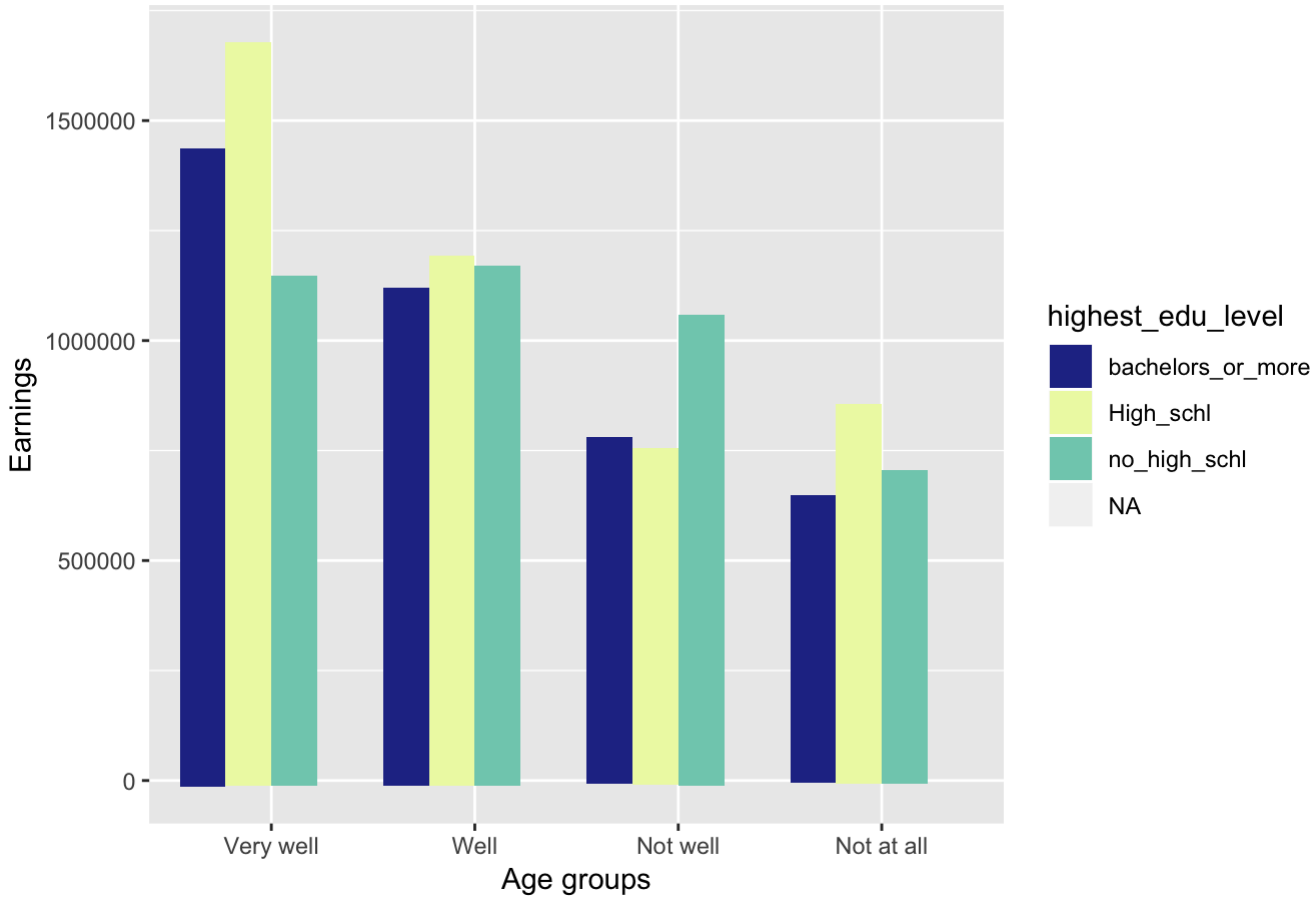
The residuals vs leverage plot has no points beyond the cooks distance. So, none of the points seem to be outliers.

English Proficiency vs Earnings across Education levels

```
## Warning: Ignoring unknown aesthetics: weight
```

```
## Warning: Removed 368162 rows containing missing values (geom_bar).
```

Fig 6: English Proficiency vs Earnings across Education levels

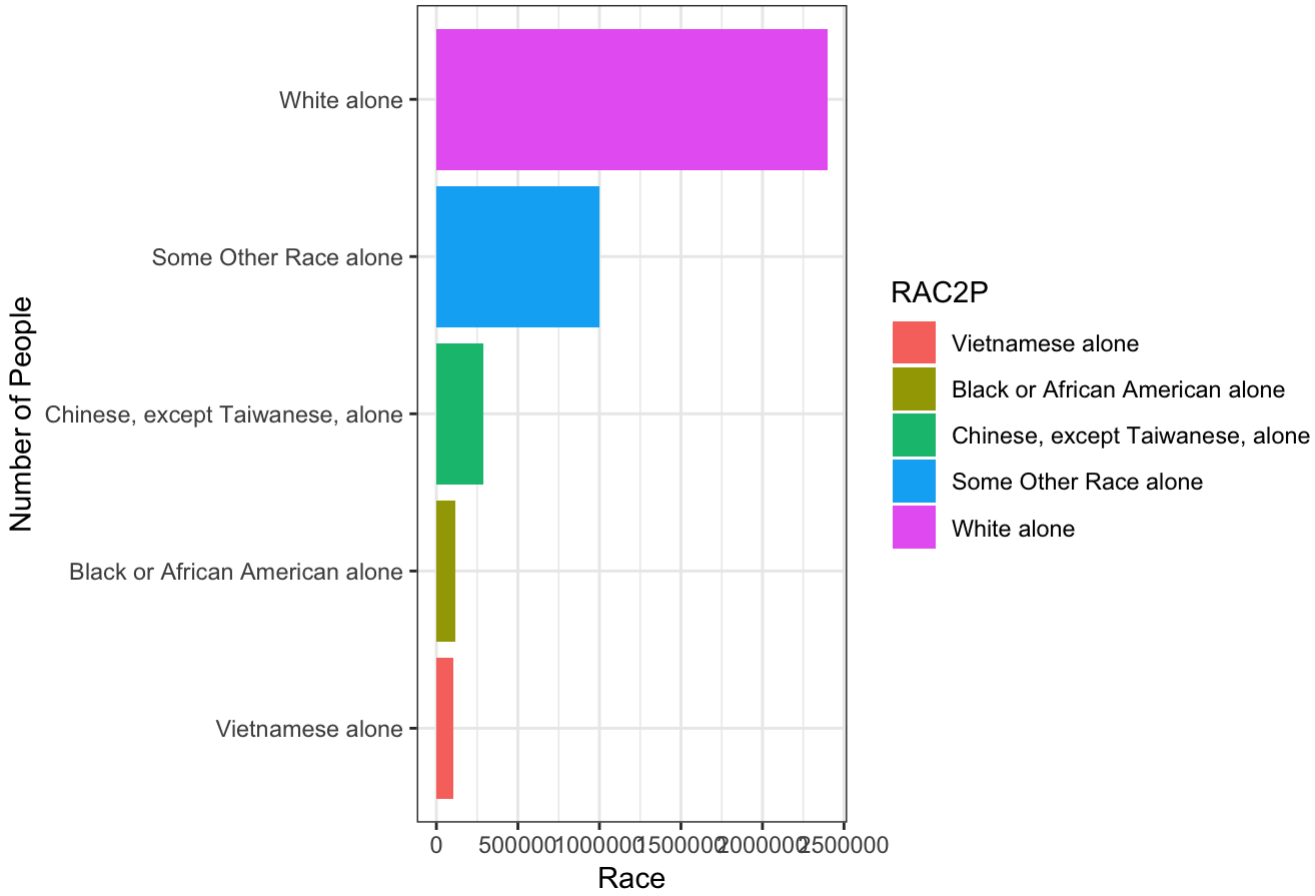


This plot(Fig 6) illustrates the relationship between English Proficiency and Earnings. As seen in the regression, it shows that earnings are highest for people having a Bachelors degree or more. However, this increase is minor from “Well” to Very Well“.

English Proficiency vs Race

```
## Selecting by number
```

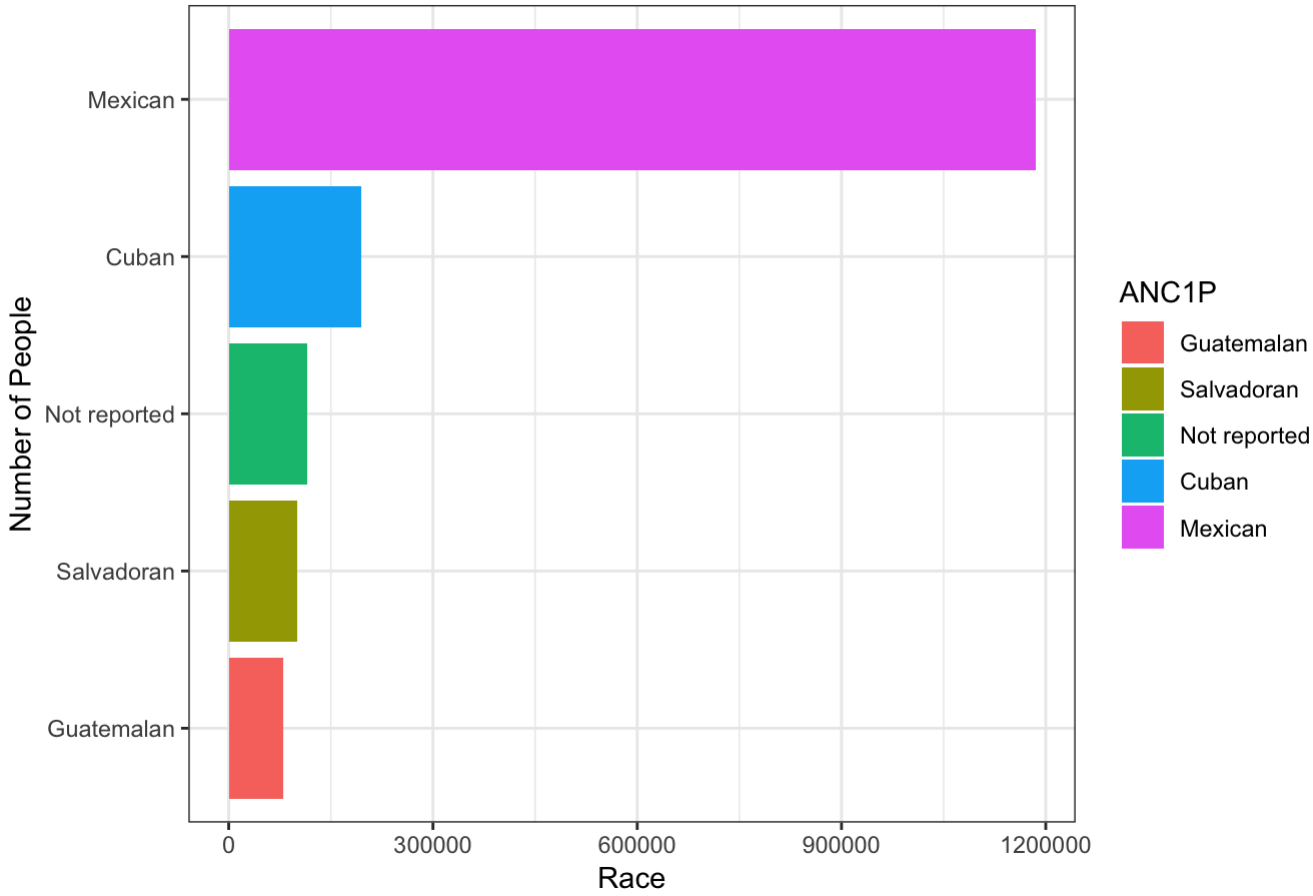
Fig 7 : Race with the most people with the lowest proficiency in English



The Fig6 illustrates that there are over 2500000 people who do not speak English well and belong to the white race. Of these white people, those of the Mexican heritage are more likely to not speak English well, followed by Cubans, Salvadorans, etc (Fig 7).

```
## Selecting by number
```

Fig 8 : Common ancestry with the most people with the lowest proficiency in English.




```
##      CitizenshipStatus                   Mexican
## [1,] "Not a citizen of the U.S"          "0.851394414575488"
## [2,] "U.S. citizen by naturalization"    "0.103131893039028"
## [3,] "Born in US"                        "0.039163646708351"
## [4,] "Born abroad of American parent(s)" "0.00631004567713332"
##      Cuban                Salvadoran
## [1,] "0.632263000867837"  "0.889734557181841"
## [2,] "0.355772143968532"  "0.0920962540312577"
## [3,] "0.00764621001658647" "0.0151426445050856"
## [4,] "0.00431864514704447" "0.00302654428181593"
```

Table 5:Citizenship status of four White Races with the highest number of people that dont speak English very well

CitizenshipStatus	Mexican	Cuban	Salvadoran
Not a citizen of the U.S	0.851394414575488	0.632263000867837	0.889734557181841
U.S. citizen by naturalization	0.103131893039028	0.355772143968532	0.0920962540312577
Born in US	0.039163646708351	0.00764621001658647	0.0151426445050856
Born abroad of American parent(s)	0.00631004567713332	0.00431864514704447	0.00302654428181593

From the above Table 5, we can see that a striking 85% of Mexicans,64% of Cubans and 89% of Salvadorans that do not speak English well are infact, not citizens of United States yet. Only a small percentage of them have been naturalized and an even smaller percentage of them , despite being born in the Us, do not speak English well (Appendix [15])

Table 6:Number of People Naturalized before and after 2000 that were not able to speak English

Year	Mexican	Cuban	Salvadoran
After 2000	101012	42955	12071
before 2000	75777	30641	5742
NA	1546906	137306	166937

From the table 7, it is interesting to note that not a lot of people from these South American countries were not granted citizenship after the year 2000. The Na here refers to people that are waiting to be naturalized.

```
## Warning: Ignoring unknown aesthetics: weight
```

```
## Warning: Removed 368162 rows containing missing values (geom_bar).
```

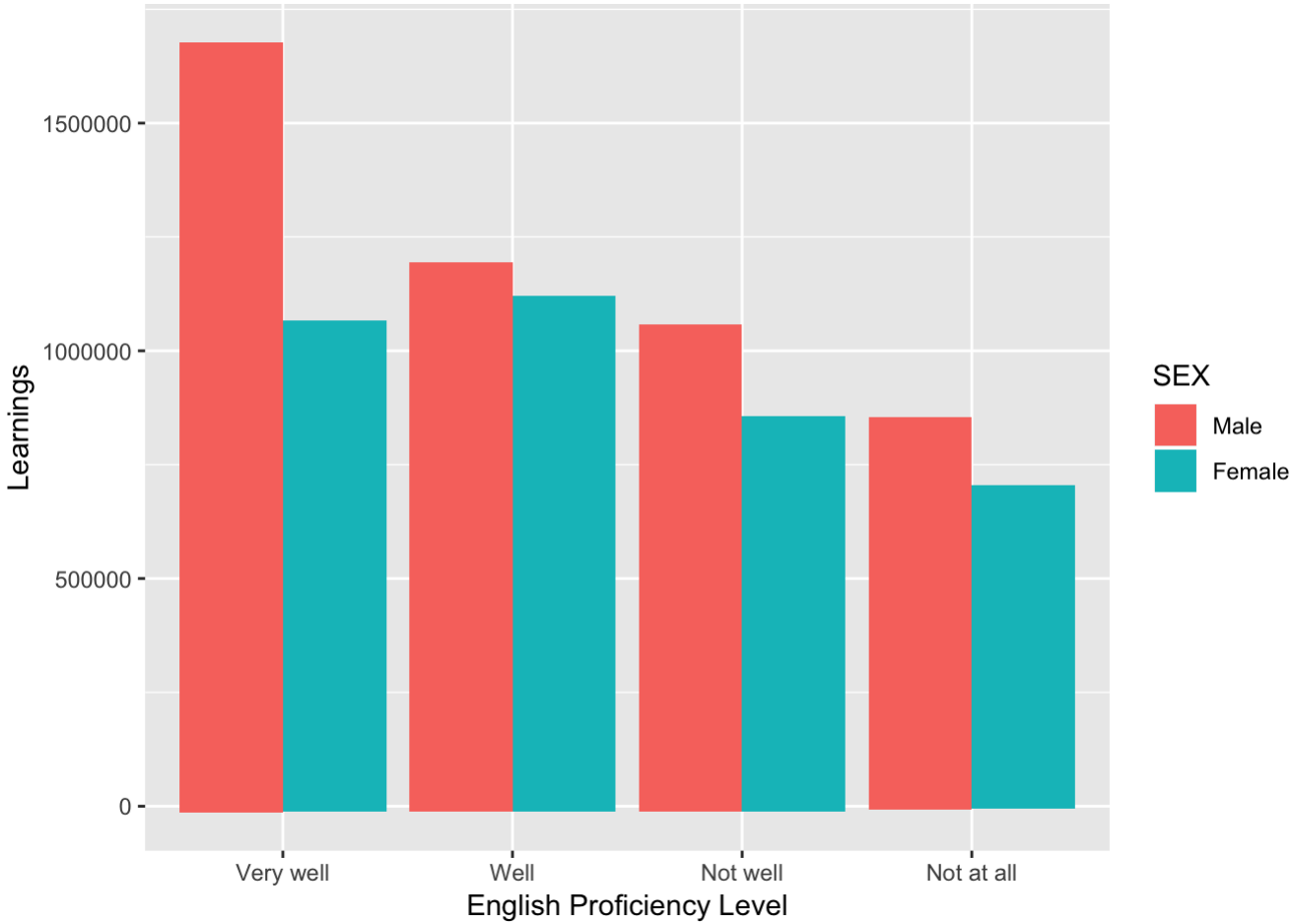
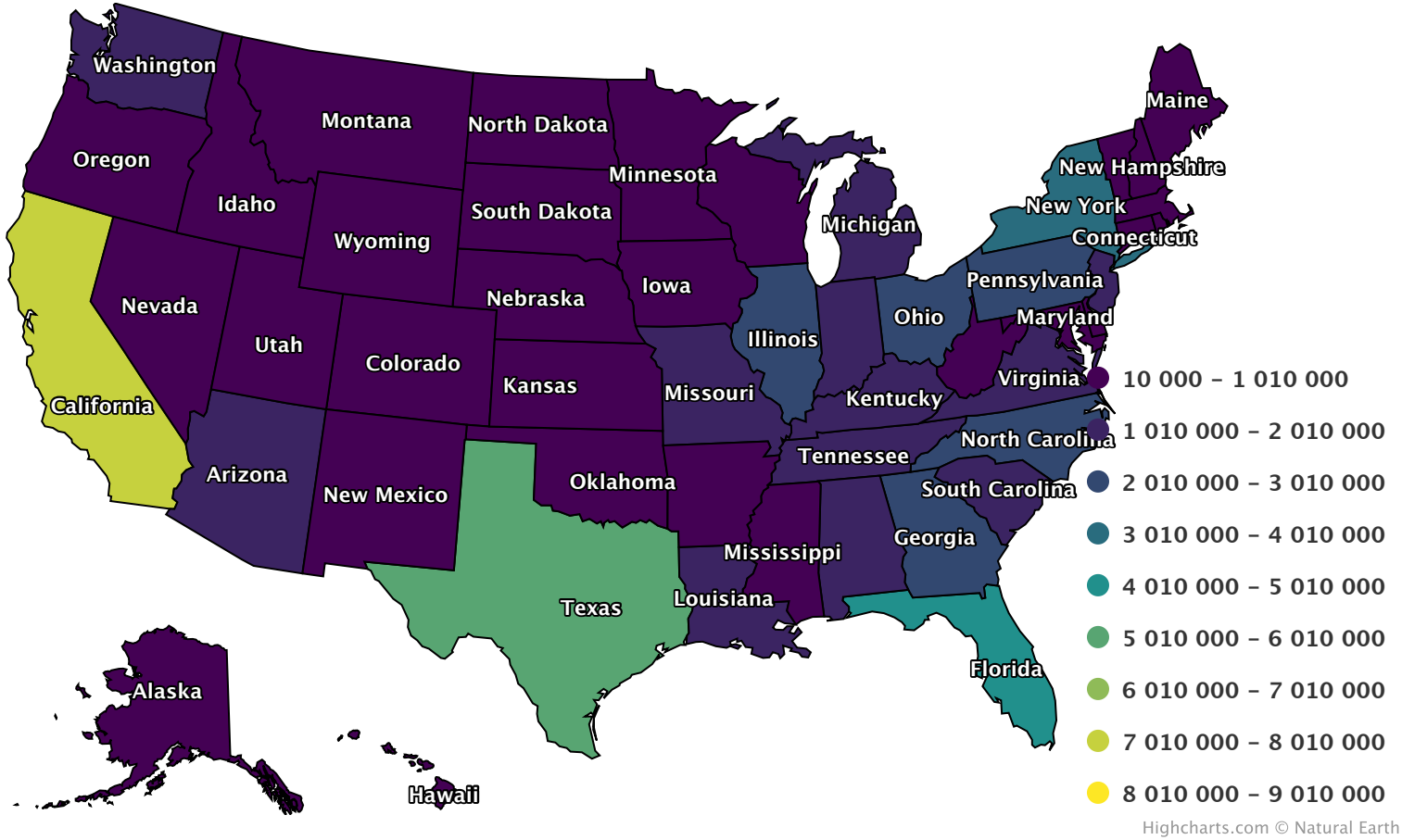


Figure 9 shows that earnings at levels “Not at all” and “Well”, the Earing among men and women is the same. However, at levels “Not well” and “Very well”, earnings of men are much higher than women.

Poverty

The following maps gives the statewise distribution of People below the poverty line.

Fig 10 :Distribution of Population below the poverty line



At 8 million, the state of California has a staggering number of people below the poverty line, followed by Texas and Florida.Around 50% of these poor are Whites. however, since the population of whites in general in much greater than other races, a proportional comparison reveals that Alaskan natives are the poorest racial groups in the three states, followed by Black or African American(Table 7)

Table 7:Proportion of Population by Race in California, Texas and Florida below poverty

Race	California	Texas	Florida
White alone	0.5650027	0.7111520	0.6848783

Race	California	Texas	Florida
Black or African American alone	0.0595742	0.1336348	0.2172956
American Indian #alone	0.0084326	0.0044199	0.0024582
Alaska Native alone	0.0000680	0.0000279	0.0000212
American Indian	0.0030250	0.0028115	0.0010611
Asian alone	0.1048220	0.0329247	0.0253641
Native #Hawaiian	0.0036755	0.0009212	0.0008686
Some Other Race	0.2188136	0.0917988	0.0437107
Two or More Races	0.0365864	0.0223092	0.0243422

Now, let's look at the Educational Qualifications of the poor in these three states(Table 8):

Table 8:Proportion of Population by Education in California, Texas and Florida below poverty

Education	Proportion_
bachelors_or_more	0.1189764
High_schl	0.3780418
no_high_schl	0.5029818

We can see from the table that, over 50% of the poor have not completed high school and nearly 40% have just a high school diploma. This table also shows that there are people who have a bachelors degree or more but are not able to make ends meet.I also able to find that over 65% of those in extreme poverty without access to basic health insurance were whites, irrespective of the education attained in these three states(Table 9).The next groups that did not have access to health insurance while having a degree, were Blacks and Asians.

Table 9:Racial breakdown of people in extreme poverty with no health insurance at different education levels in the states of CA, FL, TX

Race	No_high_schl	High_schl	Degree_or_more
White alone	0.6568464	0.6503934	0.6410209
Black or African American alone	0.0811692	0.1560478	0.1228150
American Indian #alone	0.0046449	0.0052747	0.0034468
Alaska Native alone	0.0000230	0.0000513	0.0000651
American Indian	0.0027252	0.0025954	0.0016065
Asian alone	0.0298496	0.0446898	0.1302535
Native #Hawaiian	0.0013169	0.0021539	0.0016776
Some Other Race	0.2056947	0.1143556	0.0667678
Two or More Races	0.0177301	0.0244381	0.0323468

Around 30% of these people that do not have acces to health insurance work in the Construction industry and around 35% of them work in Transport and Production servies. They seem to be working around 36 hours a week on average. This shows that they are not working full time (Table 10)

```
## Selecting by WorkingHours
```

Table 10:Percentage of people below poverty, with healthcare along with their average number of hours worked per week

Occupation	PplBelowPoverty	WorkingHours
Construction	32.9899473	36.49020
TransportServices	21.1093553	36.10937
ProductionServices	16.0347109	37.05902
FishingFarmingAgri	11.6601797	40.64302
RepairServices	8.0167247	37.18835

Occupation	PplBelowPoverty	WorkingHours
Managers	6.5662693	38.12966
PublicRescueLawInforce	2.6862324	34.12923
Engineering	0.5686914	35.42851
MineralExtraction	0.3553891	45.83543
Military	0.0125000	57.36697

These working hours vary greatly from those above and below the poverty line. For people above the poverty line, the average hours worked per week is from 40-42. On the other hand, it drops to an average of 30 hours a week for those living below poverty.

Fig 11:Working hours across states for people above poverty line

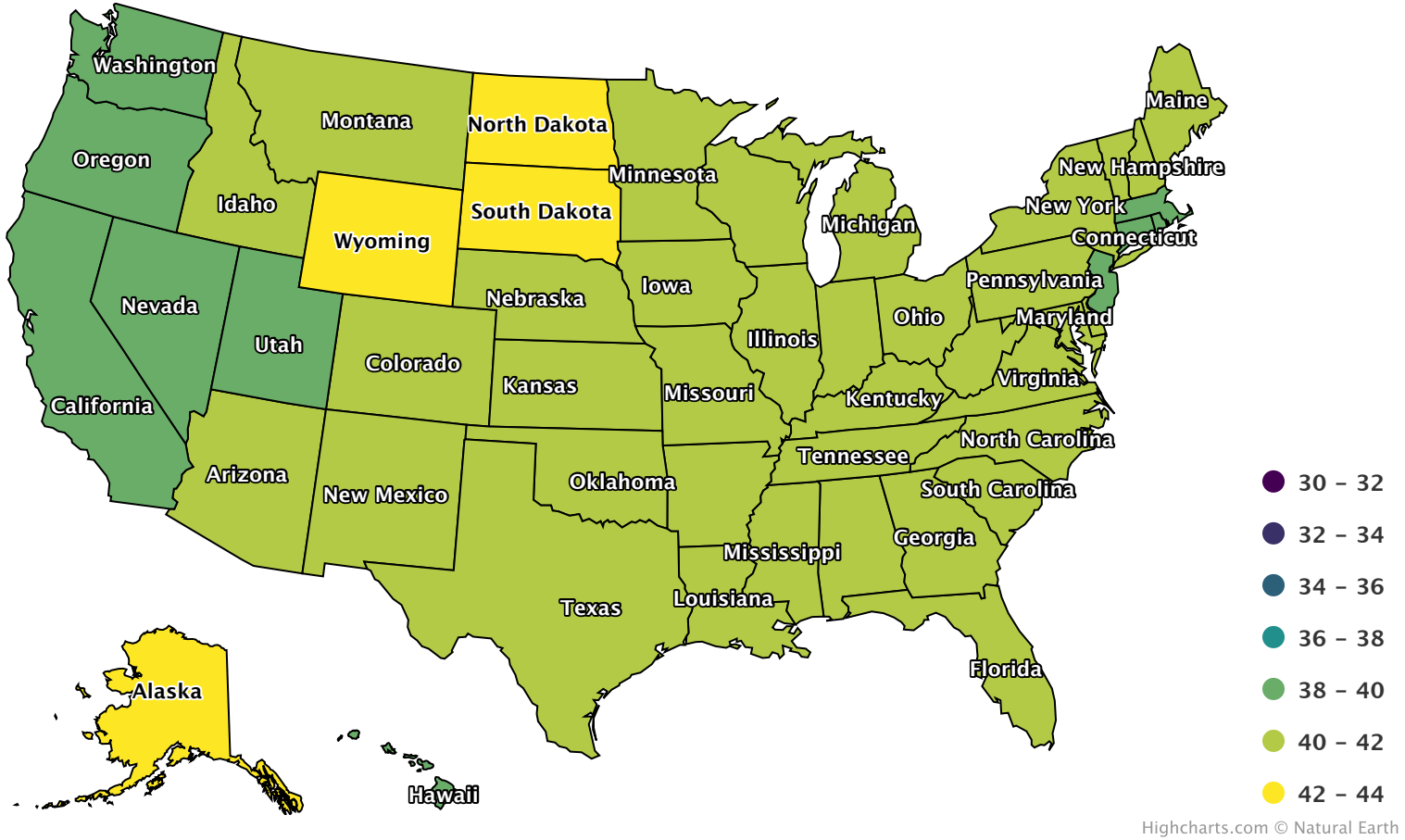
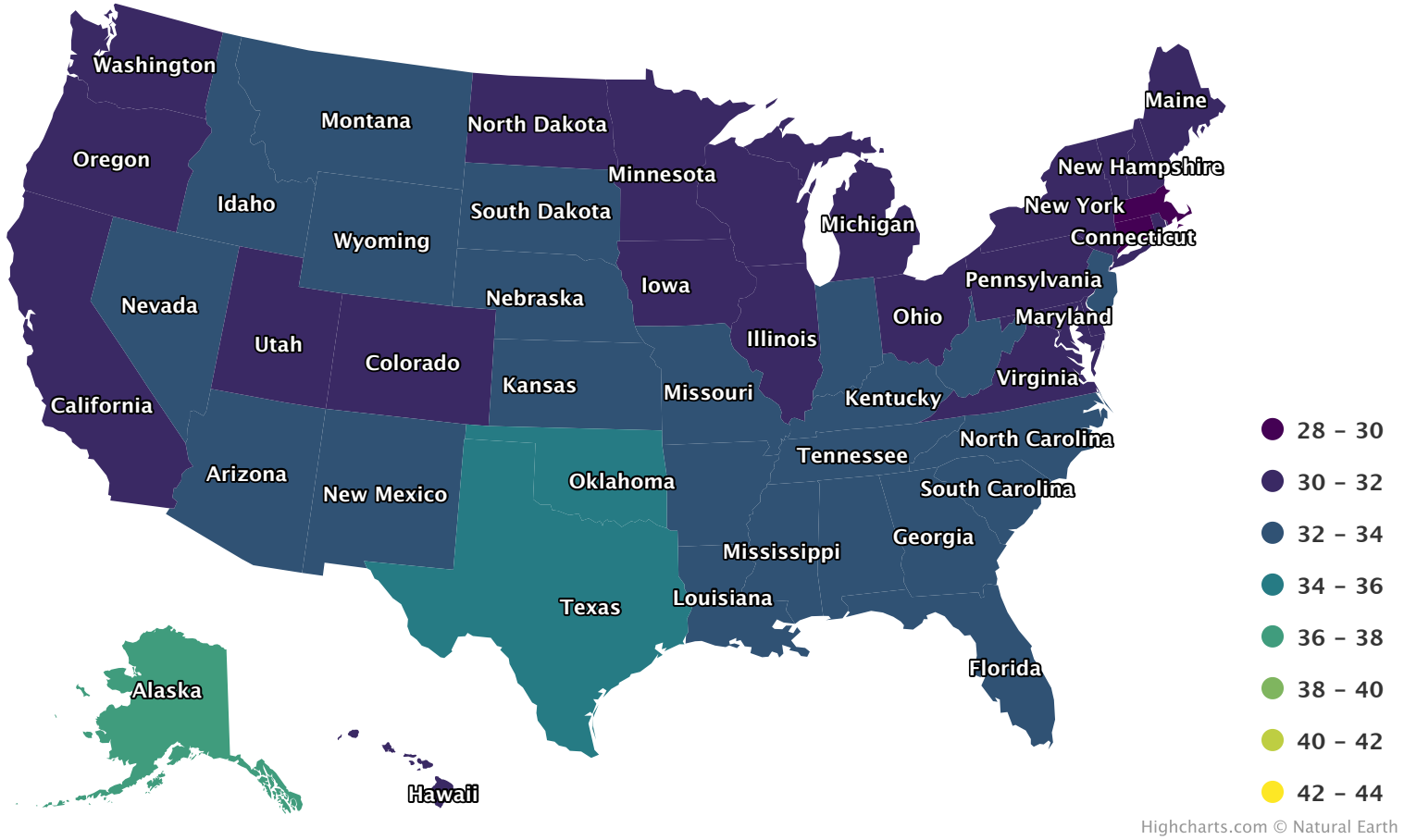


Fig 12:Working hours across states for people below poverty line



Discussion

By running a regression model I was able to conclude that there is a relationship between how well you speak English and yearly Earnings. However, covariates age and education were able to explain only some of the variation in the model as the correlation coefficient was 0.14. This means that the model has a poor predictive power and some other covariates need to be measured and accounted for.This means that the government can takes steps to improve English among those of Mexican and Cuban heritage to improve their income. If my use of weights and the adjustment factors is right, then I believe my approach towards analysis of proficiency on earnings is proper.The analysis on the poverty

variable showed that California, Texas and Florida had the highest number of poor with those of Alaskan descent most likely to be below poverty. Whites as race were the most likely to not have medical insurance when comapred with other races. This analysis can help ppolicy makers to divert funding to these states and to races which are more affected than others.

```
library(lattice)

y<- SSData2$POVPIP
x<- SSData2$Earnings
z<- SSData2$RAC1P
bwplot(POVPIP~Earnings|RAC1P,SSData2,layout = c(3,3),xlab="Earnings", ylab= "English Proficiency level",main = "F
ig6 :Earnings per year breakdown among Gender across various Racial Groups.")
```

APPENDIX 1

Data PreProcessing [1]

```
#Code chunk 1
#Loading the data in the rscript
df1 <- c("AGEP", "ENG", "CIT","CITWP", "LANX", "SCHL", "SEX", "PERNP","PINCP", "PWGTP", "ADJINC","RAC2P","LANP",
"ANC1P","POVPIP","SOCP","WKHP","RAC1P","ST","ESR","HICOV")

ssd <- fread("/Users/ShaimaShoukat/Downloads/csv_pus/ss16pusd.csv",na.strings = c("NA","N/A",""),stringsAsFactors
=FALSE,
             data.table = FALSE,select = df1 )
ssa <- fread("/Users/ShaimaShoukat/Downloads/csv_pus/ss16pusa.csv",na.strings = c("NA","N/A",""),stringsAsFactors
=FALSE,
             data.table = FALSE,select = df1 )
ssb <- fread("/Users/ShaimaShoukat/Downloads/csv_pus/ss16pusb.csv",na.strings = c("NA","N/A",""),stringsAsFactors
=FALSE,
             data.table = FALSE,select = df1 )
ssc<- fread("/Users/ShaimaShoukat/Downloads/csv_pus/ss16pusc.csv",na.strings = c("NA","N/A",""),stringsAsFactors=
FALSE,
            data.table = FALSE,select = df1 )

SS <- rbind(ssa,ssb,ssc,ssd)

fwrite(SS, file = "SS.csv")
```

[2]

```
#Code chunk 2
SSData2 <-fread(file = "SS.csv")

#Earning adjustments
SSData2<-SSData2%>%mutate(Earnings=PINCP*(ADJINC/1000000))
SSData2<-SSData2%>%mutate(Income=PERNP*(ADJINC/1000000))

#Factoring Citizenship
SSData2$CIT<-factor(SSData2$CIT, levels=c(1,2,3,4,5), labels=c("Born in US", "Born in Puerto Rico", "Born abroad
 of American parent(s)", "U.S. citizen by naturalization","Not a citizen of the U.S"))

#Factoring Gender
SSData2$SEX<-factor(SSData2$SEX, levels=c(1,2), labels=c("Male", "Female"))

#Factoring on English proficiency
SSData2$ENG<-factor(SSData2$ENG, levels=c(1,2,3,4), labels=c("Very well", "Well", "Not well", "Not at all"))

#Factors for educational attainment variable
SSData2$SCHL<-factor(SSData2$SCHL, levels =c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24), lab
els = c("No schooling completed","Nursery school, preschool","Kindergarten","Grade 1", "Grade 2","Grade 3","Grade
4","Grade 5","Grade 6","Grade 7","Grade 8","Grade 9","Grade 10","Grade 11","12th grade - no diploma", "Regular hi
gh school diploma","GED or alternative credential","Some college, but less than 1 year","1 or more years of colle
ge credit, no degree","Associate's degree","Bachelor's degree","Master's degree","Professional degree beyond a ba
chelor's degree","Doctorate degree"))


#Factoring on Race2
SSData2$RAC2P<-factor(SSData2$RAC2P,levels=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26
,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,6
4,65,66,67,68),labels=c("White alone","Black or African American alone","Apache alone","Blackfeet alone","Cheroke
e alone","Cheyenne alone","Chickasaw alone","Chippewa alone","Choctaw alone","Comanche alone","Creek alone","Crow
alone","Hopi alone","Iroquois alone","Lumbee alone","Mexican American Indian alone","Navajo alone","Pima alone",
"Potawatomi alone","Pueblo alone","Puget Sound Salish alone","Seminole alone","Sioux alone","South American India
n alone","Tohono O'Odham alone","Yaqui alone","Other specified American Indian tribes alone","All other specified
American Indian tribe combinations","American Indian, tribe not specified","Alaskan Athabascan alone","Tlingit-Ha
ida alone","Inupiat alone","Yup'ik alone","Aleut alone","Other Alaska Native","Other American Indian and Alaska N
ative specified","American Indian and Alaska Native, not specified","Asian Indian alone","Bangladeshi alone","Bhu
tanese alone","Burmese alone","Cambodian alone","Chinese, except Taiwanese, alone","Taiwanese alone","Filipino al
one","Hmong alone","Indonesian alone","Japanese alone","Korean alone","Laotian alone","Malaysian alone","Mongolia
```

```
n alone","Nepalese alone","Pakistani alone","Sri Lankan alone","Thai alone","Vietnamese alone","Other Asian alon
e","All combinations of Asian races only","Native Hawaiian alone","Samoan alone","Tongan alone","Guamanian or Cha
morro alone","Marshallese alone","Fijian alone","Other Native Hawaiian and Other Pacific Islander","Some Other Ra
ce alone","Two or More Races"))

#Factoring Ancestry1
SSData2$ANC1P<-factor(SSData2$ANC1P,levels=c(001,003,005,008,009,011,012,020,021,022,024,026,032,040,046,049,050,
051,068,077,078,082,084,087,088,089,091,094,097,098,099,100,102,103,109,111,112,114,115,122,124,125,128,129,130,1
31,142,144,146,148,152,153,154,168,169,170,171,176,177,178,179,181,183,185,187,190,194,195,200,210,211,212,213,21
5,218,219,221,222,223,224,225,226,227,231,232,233,234,235,236,237,238,239,249,250,251,252,261,271,275,290,291,295
,300,301,302,308,310,314,322,325,329,330,331,335,336,359,360,370,400,402,406,411,416,417,419,421,425,427,429,431,
434,435,442,465,483,484,490,495,496,499,508,510,515,522,523,529,534,541,553,564,566,568,570,576,587,588,598,599,6
00,603,607,609,615,618,620,650,680,690,700,703,706,707,712,714,720,730,740,748,750,765,768,770,776,782,785,793,79
5,799,800,803,808,811,814,815,820,821,822,825,841,850,899,900,901,902,903,904,907,913,914,917,918,919,920,921,922
,924,925,927,929,931,935,937,939,940,983,994,995,996,997,998,999) ,labels=c("Alsatian","Austrian","Basque","Belgi
an","Flemish","British","British Isles","Danish","Dutch","English","Finnish","French","German","Prussian","Greek"
,"Icelander","Irish","Italian","Sicilian","Luxemburger","Maltese","Norwegian","Portuguese","Scotch Irish","Scotti
sh","Swedish","Swiss","Irish Scotch","Welsh","Scandinavian","Celtic","Albanian","Belorussian","Bulgarian","Croati
an","Czech","Bohemian","Czechoslovakian","Estonian","German Russian","Rom","Hungarian","Latvian","Lithuanian","Ma
cedonian","Montenegrin","Polish","Romanian","Moldavian","Russian","Serbian","Slovak","Slovene","Turkestani","Uzbe
g","Georgia CIS","Ukrainian","Yugoslavian","Bosnian and Herzegovinian","Slavic","Slavonian","Central European","N
orthern European","Southern European","Western European","Eastern European","Germanic","European","Spaniard","Mex
ican","Mexican American","Mexicano","Chicano","Mexican American Indian","Mexican State","Mexican Indian","Costa R
ican","Guatemalan","Honduran","Nicaraguan","Panamanian","Salvadoran","Central American","Argentinean","Bolivian",
"Chilean","Colombian","Ecuadorian","Paraguayan","Peruvian","Uruguayan","Venezuelan","South American","Latin Ameri
can","Latin","Latino","Puerto Rican","Cuban","Dominican","Hispanic","Spanish","Spanish American","Bahamian","Barb
adian","Belizean","Jamaican","Dutch West Indian","Trinidadian Tobagonian","British West Indian","Antigua and Barb
uda","Grenadian","Vincent-Grenadine Islander","St Lucia Islander","West Indian","Haitian","Other West Indian","Br
azilian","Guyanese","Algerian","Egyptian","Moroccan","North African","Iranian","Iraqi","Israeli","Jordanian","Leb
anese","Saudi Arabian","Syrian","Armenian","Turkish","Yemeni","Kurdish","Palestinian","Assyrian","Chaldean","Mide
ast","Arab","Arabic","Other Arab","Cameroonian","Cape Verdean","Congolese","Ethiopian","Eritrean","Ghanaian","Ken
yan","Liberian","Nigerian","Senegalese","Sierra Leonean","Somali","South African","Sudanese","Other Subsaharan Af
rican","Ugandan","Western African","African","Afghan","Bangladeshi","Bhutanese","Nepali","Asian Indian","Bengali"
,"East Indian","Punjabi","Pakistani","Sri Lankan","Burmese","Cambodian","Chinese","Cantonese","Mongolian","Tibeta
n","Filipino","Indonesian","Japanese","Okinawan","Korean","Laotian","Hmong","Malaysian","Thai","Taiwanese","Vietn
amese","Eurasian","Asian","Other Asian","Australian","New Zealander","Polynesian","Hawaiian","Samoan","Tongan","M
icronesian","Guamanian","Chamorro","Marshallese","Fijian","Pacific Islander","Other Pacific","Afro American","Afr
o","African American","Black","Negro","Creole","Central American Indian","South American Indian","Native America
n","Indian","Cherokee","American Indian","Aleut","Eskimo","White","Anglo","Appalachian","Pennsylvania German","Ca
nadian","French Canadian","Cajun","American","United States","Texas","North American","Mixture","Uncodable entrie
```

```
s","Other groups","Other responses","Not reported"))

#Factoring Ancestry2
#SSData2$ANC2P<-factor(SSData2$ANC2P,levels= c(001 ,003 ,005 ,008 ,009 ,011 ,012 ,020 ,021 ,022 ,024 ,026 ,032 ,0
40 ,046 ,049 ,050 ,051 ,068 ,077 ,078 ,082 ,084 ,087 ,088 ,089 ,091 ,094 ,097 ,098 ,099 ,100 ,102 ,103 ,109 ,111
 ,112 ,114 ,115 ,122 ,124 ,125 ,128 ,129 ,130 ,131 ,142 ,144 ,146 ,148 ,152 ,153 ,154 ,168 ,169 ,170 ,171 ,176 ,1
77 ,178 ,179 ,181 ,183 ,185 ,187 ,190 ,194 ,195 ,200 ,210 ,211 ,212 ,213 ,215 ,218 ,219 ,221 ,222 ,223 ,224 ,225
 ,226 ,227 ,231 ,232 ,233 ,234 ,235 ,236 ,237 ,238 ,239 ,249 ,250 ,251 ,252 ,261 ,271 ,275 ,290 ,291 ,295 ,300 ,3
01 ,302 ,308 ,310 ,314 ,322 ,325 ,329 ,330 ,331 ,335 ,336 ,359 ,360 ,370 ,400 ,402 ,406 ,411 ,416 ,417 ,419 ,421
 ,425 ,427 ,429 ,431 ,434 ,435 ,442 ,465 ,483 ,484 ,490 ,495 ,496 ,499 ,508 ,510 ,515 ,522 ,523 ,529 ,534 ,541 ,5
53 ,564 ,566 ,568 ,570 ,576 ,587 ,588 ,598 ,599 ,600 ,603 ,607 ,609 ,615 ,618 ,620 ,650 ,680 ,690 ,700 ,703 ,706
 ,707 ,712 ,714 ,720 ,730 ,740 ,748 ,750 ,765 ,768 ,770 ,776 ,782 ,785 ,793 ,795 ,799 ,800 ,803 ,808 ,811 ,814 ,8
15 ,820 ,821 ,822 ,825 ,841 ,850 ,899 ,900 ,901 ,902 ,903 ,904 ,907 ,913 ,914 ,917 ,918 ,919 ,920 ,921 ,922 ,924
 ,925 ,927 ,929 ,931 ,935 ,937 ,939 ,940 ,983 ,994 ,995 ,996 ,997 ,998 ,999),labels= c("Alsatian","Austrian","Bas
que","Belgian","Flemish","British","British Isles","Danish","Dutch","English","Finnish","French","German","Prussi
an","Greek","Icelander","Irish","Italian","Sicilian","Luxemburger","Maltese","Norwegian","Portuguese","Scotch Iri
sh","Scottish","Swedish","Swiss","Irish Scotch","Welsh","Scandinavian","Celtic","Albanian","Belorussian","Bulgari
an","Croatian","Czech","Bohemian","Czechoslovakian","Estonian","German Russian","Rom","Hungarian","Latvian","Lith
uanian","Macedonian","Montenegrin","Polish","Romanian","Moldavian","Russian","Serbian","Slovak","Slovene","Turkes
tani","Uzbeg","Georgia CIS","Ukrainian","Yugoslavian","Bosnian and Herzegovinian","Slavic","Slavonian","Central E
uropean","Northern European","Southern European","Western European","Eastern European","Germanic","European","Spa
niard","Mexican","Mexican American","Mexicano","Chicano","Mexican American Indian","Mexican State","Mexican India
n","Costa Rican","Guatemalan","Honduran","Nicaraguan","Panamanian","Salvadoran","Central American","Argentinea
n","Bolivian","Chilean","Colombian","Ecuadorian","Paraguayan","Peruvian","Uruguayan","Venezuelan","South America
n","Latin American","Latin","Latino","Puerto Rican","Cuban","Dominican","Hispanic","Spanish","Spanish America
n","Bahamian","Barbadian","Belizean","Jamaican","Dutch West Indian","Trinidadian Tobagonian","British West India
n","Antigua and Barbuda","Grenadian","Vincent-Grenadine Islander","St Lucia Islander","West Indian","Haitian","Ot
her West Indian","Brazilian","Guyanese","Algerian","Egyptian","Moroccan","North African","Iranian","Iraqi","Israe
li","Jordanian","Lebanese","Saudi Arabian","Syrian","Armenian","Turkish","Yemeni","Kurdish","Palestinian","Assyri
an","Chaldean","Mideast","Arab","Arabic","Other Arab","Cameroonian","Cape Verdean","Congolese","Ethiopian","Eritr
ean","Ghanaian","Kenyan","Liberian","Nigerian","Senegalese","Sierra Leonean","Somali","South African","Sudanes
e","Other Subsaharan African","Ugandan","Western African","African","Afghan","Bangladeshi","Bhutanese","Nepal
i","Asian Indian","Bengali","East Indian","Punjabi","Pakistani","Sri Lankan","Burmese","Cambodian","Chinese","Can
tonese","Mongolian","Tibetan","Filipino","Indonesian","Japanese","Okinawan","Korean","Laotian","Hmong","Malaysia
n","Thai","Taiwanese","Vietnamese","Eurasian","Asian","Other Asian","Australian","New Zealander","Polynesian","Ha
waiian","Samoan","Tongan","Micronesian","Guamanian","Chamorro","Marshallese","Fijian","Pacific Islander","Other P
acific","Afro American","Afro","African American","Black","Negro","Creole","Central American Indian","South Ameri
can Indian","Native American","Indian","Cherokee","American Indian","Aleut","Eskimo","White","Anglo","Appalachia
n","Pennsylvania German","Canadian","French Canadian","Cajun","American","United States","Texas","North America
n","Mixture","Uncodable entries","Other groups","Other responses","Not reported"))
```

```
SSData2$LANP<-factor(SSData2$LANP,levels=c(1000 ,1025 ,1055 ,1069 ,1110 ,1120 ,1125 ,1130 ,1132 ,1134 ,1140 ,1141
,1142 ,1155 ,1170 ,1175 ,1200 ,1210 ,1220 ,1231 ,1235 ,1242 ,1250 ,1260 ,1262 ,1263 ,1270 ,1273 ,1274 ,1275 ,1276
,1277 ,1278 ,1281 ,1283 ,1288 ,1290 ,1292 ,1315 ,1327 ,1340 ,1350 ,1360 ,1380 ,1420 ,1435 ,1440 ,1450 ,1500 ,1525
,1530 ,1540 ,1564 ,1565 ,1582 ,1675 ,1690 ,1730 ,1737 ,1750 ,1765 ,1900 ,1960 ,1970 ,2000 ,2030 ,2050 ,2100 ,2160
,2270 ,2350 ,2430 ,2475 ,2525 ,2535 ,2560 ,2575 ,2715 ,2770 ,2850 ,2910 ,2920 ,2950 ,3150 ,3190 ,3220 ,3270 ,3350
,3420 ,3500 ,3570 ,3600 ,4500 ,4545 ,4560 ,4565 ,4590 ,4640 ,4830 ,4840 ,4880 ,4900 ,5150 ,5345 ,5525 ,5645 ,5845
,5900 ,5940 ,5950 ,6120 ,6205 ,6230 ,6290 ,6300 ,6370 ,6500 ,6795 ,6800 ,6839 ,6930 ,6933 ,6936 ,7019 ,7032 ,7039
,7050 ,7059 ,7060 ,7124 ,7300 ,9999), labels = c("Jamaican Creole English","Other English-based Creole languages"
,"Haitian","Kabuverdianu","German","Swiss German","Pennsylvania German","Yiddish","Dutch","Afrikaans","Swedish",
"Danish","Norwegian","Italian","French","Cajun French","Spanish","Portuguese","Romanian","Irish","Greek","Albania
n","Russian","Ukrainian","Czech","Slovak","Polish","Bulgarian","Macedonian","Serbocroatian","Bosnian","Croatian",
"Serbian","Lithuanian","Latvian","Armenian","Farsi","Dari","Kurdish","Pashto","India N.E.C.","Hindi","Urdu","Beng
ali","Punjabi","Konkani","Marathi","Gujarati","Nepali","Pakistan N.E.C.","Sinhala","Other Indo-Iranian languages"
,"Other Indo-European languages","Finnish","Hungarian","Turkish","Mongolian","Telugu","Kannada","Malayalam","Tami
l","Khmer","Vietnamese","Chinese","Mandarin","Min Nan Chinese","Cantonese","Tibetan","Burmese","Chin languages",
"Karen languages","Thai","Lao","IuMien","Hmong","Japanese","Korean","Malay","Indonesian","Other languages of Asi
a","Filipino","Tagalog","Cebuano","Ilocano","Other Philippine languages","Chamorro","Marshallese","Chuukese","Sam
oan","Tongan","Hawaiian","Other Eastern Malayo-Polynesian languages","Arabic","Hebrew","Assyrian Neo-Aramaic","Ch
aldean Neo-Aramaic","Amharic","Tigrinya","Oromo","Somali","Other Afro-Asiatic languages","Nilo-Saharan languages"
,"Swahili","Ganda","Shona","Other Bantu languages","Manding languages","Other Mande languages","Fulah","Wolof","A
kan (incl. Twi)","Ga","Gbe languages","Yoruba","Edoid languages","Igbo","Other Niger-Congo languages","Other lang
uages of Africa","Aleut languages","Ojibwa","Apache languages","Navajo","Kiowa-Tanoan languages","Dakota language
s","Muskogean languages","Keres","Cherokee","Zuni","Uto-Aztecan languages","Other Native North American language
s","Other Central and South American languages","Other and unspecified languages"))

#Factoring Healthcare
SSData2$HICOV<-factor(SSData2$HICOV, levels = c(1,2), labels = c("With health insurance coverage","No health insu
rance coverage"))

#Factoring on Race1
SSData2$RAC1P<-factor(SSData2$RAC1P, levels = c(1,2,3,4,5,6,7,8,9), labels = c("White alone","Black or African Am
erican alone","American Indian #alone","Alaska Native alone","American Indian","Asian alone","Native #Hawaiian",
"Some Other Race","Two or More Races"))

#Factoring States
SSData2$ST<-factor(SSData2$ST,levels=c(1,2,4,5,6,8,9,10,11,12,13,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,
31,32,33,34,35,36,37,38,39,40,41,42,44,45,46,47,48,49,50,51,53,54,55,56,72),
labels=c("AL","AK","AZ","AR","CA","CO","CT","DE","DC","FL","GA","HI","ID","IL","IN","IA","KS","KY","LA","ME","MD"
,"MA","MI","MN","MS","MO","MT","NE","NV","NH","NJ","NM","NY","NC","ND","OH","OK","OR","PA","RI","SC","SD","TN","T
```

```
X","UT","VT","VA","WA","WV","WI","WY","PR"))

#Factoring Employment
SSData2$ESR<-factor(SSData2$ESR,levels=c(1,2,3,4,5,6),
labels=c("Civilian employed, at work","Civilian employed, with a job but not at work","Unemployed","Armed forces,
at work","Armed forces, with a job but not at work","Not in labor force"))

# The Skilled Ocuupation variable has over 400 levels.
#Hence,Collapsing the Skilled Occupation variable into broad categories.
SSData2$SOCP <- fct_collapse(factor(SSData2$SOCP), Managers = c("1110XX","111021","112011","112020","112031","113
011","113021","113031","113111","113121","113131","113051","113061","113071","119013","119021","119030","119041",
"119051","119071","119081","119111","119121","119141","119151","119161","119XXX"),
                             Business = c("131011","131021","131022","131023","131030","131041","131051","131070",
"131081","131111","131121","131131","131141","131151","131161","131199"), "Finance" = c("132011","132021","13203
1","132041","132051","132052","132053","132061","132070","132081","132082","132099"),
                             ComputerIT = c("151111","151121","151122","151131","15113X","151134","151141","15114
2","151143","151150","151199","152011","152031","1520XX"),
                             Engineering = c("171010","171020","172011","1720XX","172041","172051","172061","17207
0","172081","172110","172121","172131","172141","1721XX","1721YY","173010","173020","173031"), Sciences = c("1910
10","191020","191030","1910XX","192010","192021","192030","192040","192099","193011","193030","193051","1930XX",
"194011","194021","194031","1940XX","1940YY"),
                             CommunityService = c("211010","211020","211092","211093","21109X","212011","212021",
"212099"),
                             Law = c("2310XX","231012","232011","232090"),
                             Education = c("251000","252010","252020","252030","252050","253000","254010","254021"
,"254031","259041","2590XX"), Entertainment = c("271010","271020","272011","272012","272020","272030","272040","2
72099","273010","273020","273031","273041","273042","273043","273090","2740XX","274021","274030"),
                             MedicalProfessionals = c("291011","291020","291031","291041","291051","291060","29107
1","291081","291122","291123","291124","291125","291126","291127","29112X","291131","291141","291151","291181","2
91199","2911XX","292010","292021","292030","292041","292050","292061","292071","292081","292090","299000"),
                             HealthCare= c("311010","312010","312020","319011","319091","319092","319094","319095"
,"319096","319097","31909X"), PublicRescueLawInforce = c("331011","331012","331021","331099","332011","332020","3
33010","333021","3330XX","333050","339011","339021","339030","339091","339093","33909X"),
                             Food = c("351011","351012","352010","352021","353011","353021","353022","353031","353
041","3590XX","359021","359031"), CleaningMaintenance = c("371011","371012","37201X","372012","372021","373010"),
                             UtilityServicesPRS = c("391010","391021","392011","392021","393010","393021","393031"
,"393090","394031","3940XX","395011","395012","395090","396010","397010","399011","399021","399030","399041","399
099"),
                             Sales = c("411011","411012","412010","412021","412022","412031","413011","413021","41
3031","413041","413099","414010","419010","419020","419031","419041","419091","419099"),
```

```
                            OfficeWorkers = c("431011","432011","432021","432099","433011","433021","433031","433
041","433051","433061","433071","433099","434011","434031","434041","434051","434061","434071","434081","434111",
"434121","434131","434141","434XXX","434161","434171","434181","434199","435011","435021","435030","435041","4350
51","435052","435053","435061","435071","435081","435111","436010","439011","439021","439022","439041","439051",
"439061","439071","439081","439111","439XXX"),
                            FishingFarmingAgri = c("451011","452011","452041","4520XX","453000","454011","454020"
), Construction = c("471011","472011","472031","472040","472050","472061","472071","47207X","472080","472111","47
2121","472130","472140","472150","472161","472181","472211","472221","472XXX","473010","474011","474021","474031"
,"474041","474051","474061","47XXXX"),
                            MineralExtraction = c("4750YY","475021","475031","475040","4750XX"),
                            RepairServices = c("491011","492011","492020","492091","492092","49209X","492096","49
2097","492098","493011","493021","493022","493023","493031","493040","493050","493090","499010","499021","499031"
,"49904X","499043","499044","499051","499052","499060","499071","499091","499094","499096","499098","4990XX"),
                            ProductionServices= c("511011","512011","512020","512031","512041","512090","513011",
"513020","513091","513092","513093","513099","514010","514021","514022","514023","514030","514041","514050","5140
XX","514111","514120","514XXX","515111","515112","515113","516011","516021","516031","516040","516050","51606X",
"516063","516064","516093","51609X","517011","517021","517041","517042","5170XX","518010","518021","518031","5180
90","519010","519020","519030","519041","519051","519061","519071","519080","519111","519120","519151","519191",
"519194","519195","519196","519197","519198","5191XX"),
                            TransportServices = c("531000","532010","532020","532031","533011","533020","533030",
"533041","533099","534010","534031","5340XX","5350XX","535020","536021","536031","536051","536061","5360XX","5370
21","537030","5370XX","537051","537061","537062","537063","537064","537070","537081","5371XX"), Military = c("551
010","552010","553010","559830"), Unemployed = c("999920"))

write.csv(SSData2, file = "SSdata2.csv")
```

[3]

```
#code cunk 3
# Proportion of males and females
EX2 <- SSData2 %>% filter(!is.na(SEX)) %>% group_by(SEX) %>% summarise(PeopleCount = sum(PWGTP))
Gender <- EX2$SEX
Proportion <- prop.table(EX2$PeopleCount)
EX22 <- data.frame(Gender , Proportion)
kable(EX22,caption = "Table1 : Proportion of Males and Females") %>% kable_styling(bootstrap_options = c("stripe
d", "hover", "condensed"),full_width = FALSE)
```

[4]

```
#Population distribution by age gap
(SSData2<-SSData2%>%filter(!is.na(AGEP))%>%mutate(age_grps = factor(case_when(AGEP > 60 ~ "Senior_Citizens",AGEP
 > 45 & AGEP <= 60 ~ "Middle_aged",AGEP >35 & AGEP <= 45 ~ "Middle_aged",AGEP >25 & AGEP <= 35 ~ "Youth",AGEP >=1
5 & AGEP <= 25 ~ "Young Adults",AGEP <=14 ~ "Children"))))

ggplot(data = SSData2) +
  geom_bar(mapping = aes(x = age_grps, fill = SEX, group=SEX, weight =PWGTP ), position = "dodge",stat="count") +
labs(subtitle="Fig1 : Population Distribution by age group ", x = "Age groups", y = "Population Count")+ scale_fi
ll_manual(values=c("#253494", "#edf8b1"))
```

[5]

```
#Distribution of the working population across different industries.
layout(matrix(c(1, 2), nrow=2), heights=c(1, 4))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Fig2 : Population Distribution by various Industries")
EX3a <- SSData2 %>% filter(!is.na(SOCP), AGEP > 18 & AGEP <= 70) %>% group_by(SOCP) %>% summarise(count = sum(PWG
TP)) %>% with(wordcloud(SOCP, count, max.words = 50,colors = brewer.pal(8, "Dark2")))

layout(matrix(c(1, 2), nrow=2), heights=c(1, 4))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Fig3 : Population Distribution by various Industries under\n the poverty line")
EX3b <- SSData2 %>% filter(!is.na(SOCP),POVPIP <125,AGEP > 18 & AGEP <= 70) %>% group_by(SOCP) %>% summarise(coun
t = sum(PWGTP)) %>% with(wordcloud(SOCP, count, max.words = 50,colors = brewer.pal(8, "Dark2")))

layout(matrix(c(1, 2), nrow=2), heights=c(1, 4))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Fig4 : Population Distribution by various Industries over \nthe poverty line")
EX3c <- SSData2 %>% filter(!is.na(SOCP),POVPIP >125,AGEP > 18 & AGEP <= 70) %>% group_by(SOCP) %>% summarise(coun
t = sum(PWGTP)) %>% with(wordcloud(SOCP, count, max.words = 50,colors = brewer.pal(8, "Dark2")))
```

[6]

```
## Selecting by count
```

Table2 :Proportion of Population by Skilled Industry

SkilledIndustry	Proportion
	0.5729220
OfficeWorkers	0.1091328
Sales	0.0893012
Managers	0.0758514
TransportServices	0.0529528
Food	0.0509653
Education	0.0488745

[7]

```
# Mean Earnings for people above and below poverty vs the national average
m1 <- SSData2 %>%filter((AGEP > 18 & AGEP <= 60)) %>% summarise(mean_earnings = weighted.mean(Earnings,PWGTP))
m2 <- SSData2 %>%filter((AGEP > 18 & AGEP <= 60),POVPIP <125) %>% summarise(mean_earnings = weighted.mean(Earning
s,PWGTP))
m3 <- SSData2 %>%filter((AGEP > 18 & AGEP <= 60),POVPIP >=125) %>% summarise(mean_earnings = weighted.mean(Earnin
gs,PWGTP))
GeneralPop <- m1$mean_earnings
BelowPoverty <- m2$mean_earnings
AbovePoverty <- m3$mean_earnings
df <- data.frame(GeneralPop,BelowPoverty,AbovePoverty)
kable(df,caption = "Table3 :Mean Earnings for people above and below poverty vs the national average") %>% kable_
styling(bootstrap_options = c("striped", "hover", "condensed"),full_width = FALSE)
```

[8]

```
#Relationship between Proficiency in English and average Earnings
df_SS <- SSData2
df_SS %>% filter((AGEP>18 & AGEP<61 & !is.na(ENG))) %>% group_by(ENG) %>% summarise(Avg_Earnings=weighted.mean(Ea
rnings,PWGTP)) %>%  ggplot(aes(x=ENG, y=Avg_Earnings)) +
  geom_point(size=3) +
  geom_segment(aes(x=ENG,
                   xend=ENG,
                   y=0,
                   yend=Avg_Earnings)) +
  labs(subtitle="Fig5 : Relationship between Proficiency in English and its effect on Earnings ", x = "Proficienc
y in English", y = "Average Earnings") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```

[9]

```
wtd.cor(SSData2$POVPIP, y=SSData2$Income, weight=SSData2$PWGTP, mean1=FALSE, collapse=TRUE) %>% kable(caption =
"Correlation between Earnings and Poverty")
```

[10]

```
#Calculating the means and standard deviations
Means <- lapply(split(SSData2, SSData2$ENG), function(SSData2) weighted.mean(SSData2$Earnings, SSData2$PWGTP,na.r
m = TRUE))
Variances <- lapply(split(SSData2, SSData2$ENG), function(SSData2) wtd.var(SSData2$Earnings, SSData2$PWGTP,na.rm
 = TRUE))
v1 <- sqrt(339130470)
v2 <- sqrt(710715316)
v3 <- sqrt(3158319029)
v4 <- sqrt(1583518875)
StandardDeviations <- c(v1,v2,v3,v4)
ENGgroups <- c("Not at all","Not well", "Very well","Well")
df <- data.frame(StandardDeviations,ENGgroups)
kable(df,caption = "Variation of income across the four groups") %>% kable_styling(bootstrap_options = c("stripe
d", "hover", "condensed"),full_width = FALSE)
```

[11]

```
#Multiple regression
#creating a new variable from education attainment
SSData2 <- SSData2 %>% mutate(highest_edu_level =factor(case_when((SCHL == "No schooling completed"|SCHL =="Nurse
ry school"|SCHL ==" preschool"|SCHL =="Kindergarten"|SCHL =="Grade 1"|SCHL =="Grade 2"|SCHL =="Grade 3"|SCHL =="G
rade 4"|SCHL =="Grade 5"|SCHL =="Grade 6"|SCHL =="Grade 7"|SCHL =="Grade 8"|SCHL =="Grade 9"|SCHL =="Grade 10"|SC
HL =="Grade 11"|SCHL =="12th grade - no diploma") ~ "no_high_schl",
                                                                (SCHL== "Regular high school diploma"|SCHL=="GED or
 alternative credential"|SCHL=="Some college, but less than 1 year"|SCHL=="1 or more years of college credit, no
 degree") ~ "High_schl",
                                                                (SCHL == "Associate's degree"|SCHL =="Bachelor's deg
ree"|SCHL =="Master's degree"|SCHL =="Professional degree beyond a bachelor's degree"|SCHL =="Doctorate degree" ~
"bachelors_or_more"))))

lm.model <- lm(Earnings ~ AGEP + ENG + highest_edu_level, data = SSData2)
summary(lm.model)

par(mfrow=c(2,2))
plot(lm.model)
```

[12]

```
#Mutating the age column and plotting Earnings vs Education levels
ENGvsEarn <- SSData2 %>%filter(!is.na(AGEP),!is.na(SCHL),!is.na(ENG))%>% mutate(age_grps = factor(case_when(AGEP
 >= 65 ~ "Senior_Citizens",
                                                    AGEP >= 25 & AGEP <= 64 ~ "Middle_aged",
                                                    AGEP >=15 & AGEP <= 24 ~ "Youth",
                                                    AGEP <=14 ~ "Children")),
                          highest_edu_level =factor(case_when((SCHL == "No schooling completed"|SCHL =="Nursery sc
hool"|SCHL ==" preschool"|SCHL =="Kindergarten"|SCHL =="Grade 1"|SCHL =="Grade 2"|SCHL =="Grade 3"|SCHL =="Grade
 4"|SCHL =="Grade 5"|SCHL =="Grade 6"|SCHL =="Grade 7"|SCHL =="Grade 8"|SCHL =="Grade 9"|SCHL =="Grade 10"|SCHL =
="Grade 11"|SCHL =="12th grade - no diploma") ~ "no_high_schl",
                                                                (SCHL== "Regular high school diploma"|SCHL=="GED or
 alternative credential"|SCHL=="Some college, but less than 1 year"|SCHL=="1 or more years of college credit, no
 degree") ~ "High_schl",
                                                                (SCHL == "Associate's degree"|SCHL =="Bachelor's deg
ree"|SCHL =="Master's degree"|SCHL =="Professional degree beyond a bachelor's degree"|SCHL =="Doctorate degree" ~
"bachelors_or_more"))))
ggplot(data = ENGvsEarn) +
  geom_bar(mapping = aes(x = ENG, y = Income, fill = highest_edu_level,weight = PWGTP), position = "dodge",stat=
"identity") + labs(subtitle="Fig : English Proficiency vs Earnings across Education levels ", x = "Age groups", y
= "Earnings")+ scale_fill_manual(values=c("#253494", "#edf8b1","#7fcdbb","#2c7fb8"))
```

[13]

```
# Which Race has the most people with the lowest proficiency in English.
SSData2 %>% select(ENG,RAC2P,PWGTP)%>% filter(ENG=="Not at all") %>%group_by(RAC2P) %>%summarise(number=sum(PWGT
P)) %>% mutate(RAC2P=reorder(RAC2P,number)) %>% top_n(5) %>%
ggplot(mapping = aes(x = RAC2P, y=number,fill=RAC2P)) +geom_bar(stat="identity") +labs(x = "Number of People", y
 = "Race", subtitle = "Fig 6 : Race with the most people with the lowest proficiency in English.") +theme_bw() +
 guides(guide=guide_legend("my title"))+coord_flip()
```

[14]

```
#What is the ancestry of white people with lowest proficiency in English.
SSData2 %>% filter(RAC2P == "White alone",ENG =="Not at all" )%>%group_by(ANC1P) %>% summarise(number=sum(PWGTP))
%>%  mutate(ANC1P=reorder(ANC1P,number)) %>% top_n(5) %>%
ggplot(mapping = aes(x = ANC1P, y=number,fill=ANC1P)) +geom_bar(stat="identity") + labs(x = "Number of People", y
= "Race", subtitle = "Fig 7 : Common ancestry with the most people with the lowest proficiency in English.") +the
me_bw() + guides(guide=guide_legend("my title"))+coord_flip()
```

[15]

```
#What is the ancestry of white people with lowest proficiency in English.
#ancestry <-  SSData2 %>% select(RAC2P,ENG,ANC1P,PWGTP)%>%filter(RAC2P == "White alone",ENG =="Not at all")%>%gro
up_by(ANC2P)%>% summarise(NumberOfPeople=sum(PWGTP)) %>% arrange(desc(NumberOfPeople))
```

```
(industry <-  SSData2 %>% select(RAC2P,ENG,SOCP,PWGTP)%>%filter(RAC2P == "White alone",ENG =="Not at all")%>%grou
p_by(SOCP)%>% summarise(NumberOfPeople=sum(PWGTP)) %>% arrange(desc(NumberOfPeople)))
```

```
#citizenship status
Cstatus11 <-  SSData2 %>% select(RAC2P,ENG,CIT,ANC1P,PWGTP)%>%group_by(CIT)%>% filter(RAC2P == "White alone",ENG
 =="Not at all",(ANC1P=="Mexican")) %>% summarise(count=sum(PWGTP)) %>% arrange(desc(count))

Cstatus12 <-  SSData2 %>% select(RAC2P,ENG,CIT,ANC1P,PWGTP)%>%group_by(CIT)%>% filter(RAC2P == "White alone",ENG
 =="Not at all",(ANC1P=="Cuban")) %>% summarise(count=sum(PWGTP)) %>% arrange(desc(count))

Cstatus13 <-  SSData2 %>% select(RAC2P,ENG,CIT,ANC1P,PWGTP)%>%group_by(CIT)%>% filter(RAC2P == "White alone",ENG
 =="Not at all",(ANC1P=="Salvadoran")) %>% summarise(count=sum(PWGTP)) %>% arrange(desc(count))
Cstatus14 <-  SSData2 %>% select(RAC2P,ENG,CIT,ANC1P,PWGTP)%>%group_by(CIT)%>% filter(RAC2P == "White alone",ENG
 =="Not at all",(ANC1P=="Guatemalen")) %>% summarise(count=sum(PWGTP)) %>% arrange(desc(count))

CitizenshipStatus <- c("Not a citizen of the U.S","U.S. citizen by naturalization","Born in US","Born abroad of A
merican parent(s)")
Mexican <-prop.table(c(1008578,122172,46394,7475))
Cuban <- prop.table(c(123125,69282,1489,841))
Salvadoran <- prop.table(c(89663,9281,1526,305))
df <- cbind(CitizenshipStatus, Mexican , Cuban, Salvadoran)
df
kable(df,caption = "Table:Citizenship status of four White Races with the highest number of people that dont spea
k English very well") %>% kable_styling(bootstrap_options = c("striped", "hover", "condensed"),full_width = FALSE
)
```

[16]

```
# mutating the year of naturalization column
df1 <-  SSData2 %>%filter(ANC1P=="Mexican",ENG =="Not at all")%>% mutate(Year = ifelse(CITWP >= 2000 ,"After 200
0", "before 2000"))%>% group_by(Year) %>% summarise(n=sum(PWGTP))
df2 <-  SSData2 %>%filter(ANC1P=="Cuban",ENG =="Not at all")%>% mutate(Year = ifelse(CITWP >= 2000 ,"After 2000",
"before 2000"))%>% group_by(Year) %>% summarise(n=sum(PWGTP))
df3 <-  SSData2 %>%filter(ANC1P=="Salvadoran",ENG =="Not at all")%>% mutate(Year = ifelse(CITWP >= 2000 ,"After 2
000", "before 2000"))%>% group_by(Year) %>% summarise(n=sum(PWGTP))
df4 <-  SSData2 %>%filter(ANC1P=="Guatemalen",ENG =="Not at all")%>% mutate(Year = ifelse(CITWP >= 2000 ,"After 2
000", "before 2000"))%>% group_by(Year) %>% summarise(n=sum(PWGTP))

Year <-df1$Year
Mexican <- df1$n
Cuban <- df2$n
Salvadoran <- df3$n
Guatemalen <- df4$n
df <- cbind(Year, Mexican , Cuban, Salvadoran, Guatemalen)
kable(df,caption = "Number of People Naturalized before and after 2000 that were not able to speak English") %>%
 kable_styling(bootstrap_options = c("striped", "hover", "condensed"),full_width = FALSE)
```

[17]

```
#Graph for Proficiency vs earnings among males and females.
ENGvsEarn <- SSData2 %>%filter(!is.na(AGEP),!is.na(SCHL),!is.na(ENG))%>% mutate(age_grps = factor(case_when(AGEP
 >= 65 ~ "Senior_Citizens",
                                                    AGEP >= 25 & AGEP <= 64 ~ "Middle_aged",
                                                    AGEP >=15 & AGEP <= 24 ~ "Youth",
                                                    AGEP <=14 ~ "Children")),
                          highest_edu_level =factor(case_when((SCHL == "No schooling completed"|SCHL =="Nursery sc
hool"|SCHL ==" preschool"|SCHL =="Kindergarten"|SCHL =="Grade 1"|SCHL =="Grade 2"|SCHL =="Grade 3"|SCHL =="Grade
 4"|SCHL =="Grade 5"|SCHL =="Grade 6"|SCHL =="Grade 7"|SCHL =="Grade 8"|SCHL =="Grade 9"|SCHL =="Grade 10"|SCHL =
="Grade 11"|SCHL =="12th grade - no diploma") ~ "no_high_schl",
                                                                (SCHL== "Regular high school diploma"|SCHL=="GED or
 alternative credential"|SCHL=="Some college, but less than 1 year"|SCHL=="1 or more years of college credit, no
 degree") ~ "High_schl",
                                                                (SCHL == "Associate's degree"|SCHL =="Bachelor's deg
ree"|SCHL =="Master's degree"|SCHL =="Professional degree beyond a bachelor's degree"|SCHL =="Doctorate degree" ~
"bachelors_or_more"))))

ggplot(data = ENGvsEarn) +
  geom_bar(mapping = aes(x = ENG, y = Income, fill = SEX,weight = PWGTP), position = "dodge",stat="identity")
```

[18]

```
#States with the highest poverty
DF <- SSData2%>% filter(POVPIP <125) %>% group_by(ST) %>% summarise(count = sum(PWGTP)) %>% arrange(desc(count))

hcmap("countries/us/us-all", data = DF, name = c("Poverty"), value = "count", joinBy = c("hc-a2", "ST"), borderCo
lor = "black",dataLabels = list(enabled = TRUE, format = '{point.name}'))%>%
hc_colorAxis(dataClasses = color_classes(c(seq(10000, 10000000, by = 1000000))))%>%
hc_legend(layout = "vertical", align = "right",floating = TRUE, valueDecimals = 0)
```

[19]

```
#Racial Profile of poor in California, Florida and Texas
Cal_pov1 <- SSData2 %>% filter(POVPIP <125,!is.na(WKHP),ST == "CA")  %>% group_by(RAC1P) %>% summarise(count = su
m(WKHP,PWGTP,na.rm = TRUE))
Cal_pov2 <- SSData2 %>% filter(POVPIP <125,!is.na(WKHP),ST == "TX")  %>% group_by(RAC1P) %>% summarise(count = su
m(WKHP,PWGTP,na.rm = TRUE))
Cal_pov3 <- SSData2 %>% filter(POVPIP <125,!is.na(WKHP),ST == "FL")  %>% group_by(RAC1P) %>% summarise(count = su
m(WKHP,PWGTP,na.rm = TRUE))

California <- prop.table(Cal_pov1$count)
Texas <- prop.table(Cal_pov2$count)
Florida <- prop.table(Cal_pov3$count)
Race <- Cal_pov$RAC1P
df <- data.frame(Race, California,Texas,Florida)
kable(df,caption = "Table :Proportion of Population by Race in California, Texas and FLorida below poverty") %>%
 kable_styling(bootstrap_options = c("striped", "hover", "condensed"),full_width = FALSE)
```

[20]

```
SSData2 <- SSData2 %>% mutate( highest_edu_level =factor(case_when((SCHL == "No schooling completed"|SCHL =="Nurs
ery school"|SCHL ==" preschool"|SCHL =="Kindergarten"|SCHL =="Grade 1"|SCHL =="Grade 2"|SCHL =="Grade 3"|SCHL ==
"Grade 4"|SCHL =="Grade 5"|SCHL =="Grade 6"|SCHL =="Grade 7"|SCHL =="Grade 8"|SCHL =="Grade 9"|SCHL =="Grade 10"|
SCHL =="Grade 11"|SCHL =="12th grade - no diploma") ~ "no_high_schl",(SCHL== "Regular high school diploma"|SCHL==
"GED or alternative credential"|SCHL=="Some college, but less than 1 year"|SCHL=="1 or more years of college cred
it, no degree") ~ "High_schl",(SCHL == "Associate's degree"|SCHL =="Bachelor's degree"|SCHL =="Master's degree"|S
CHL =="Professional degree beyond a bachelor's degree"|SCHL =="Doctorate degree" ~ "bachelors_or_more"))))


# breakdown of the poor along with their education qualifications in CA,TX and FL.
q <- SSData2 %>%filter(POVPIP <= 125,ST =="CA"|ST =="TX"|ST=="FL", !is.na(highest_edu_level))  %>% group_by(highe
st_edu_level ) %>% summarise(Ppl_below_pv= sum(PWGTP))

Proportion_ <- prop.table(q$Ppl_below_pv)
Education <- q$highest_edu_level
df <- data.frame(Education, Proportion_)
kable(df,caption = "Table :Proportion of Population by Education in California, Texas and Florida below poverty")
%>% kable_styling(bootstrap_options = c("striped", "hover", "condensed"),full_width = FALSE)
```

[21]

```
#the most impoversihed group by race below poverty, with no health care and edu attainment = no high school diplo
ma.
s <- SSData2 %>%filter(POVPIP <= 125, HICOV == "No health insurance coverage",highest_edu_level == "no_high_schl"
,ST =="CA"|ST =="TX"|ST=="FL")  %>% group_by(RAC1P) %>% summarise(Ppl_below_pv= sum(PWGTP))
#the most impoversihed group by race below poverty, with no health care and edu attainment = high school diploma.
t <- SSData2 %>%filter(POVPIP <= 125, HICOV == "No health insurance coverage",highest_edu_level == "High_schl",ST
=="CA"|ST =="TX"|ST=="FL")%>% group_by(RAC1P)  %>%summarise(Ppl_below_pv= sum(PWGTP))
#the most impoversihed group by race below poverty, with no health care and edu attainment = degree or more
u <- SSData2 %>%filter(POVPIP <= 125, HICOV == "No health insurance coverage",highest_edu_level == "bachelors_or_
more",ST =="CA"|ST =="TX"|ST=="FL") %>% group_by(RAC1P)  %>%summarise(Ppl_below_pv= sum(PWGTP))
#Weighted Proportion of Races in extreme poverty with no health care and different education levels
#proportion_table<-wpct(SSData2$RAC1P, weight=SSData2$PWGTP)
#df<-data.frame(Race=names(proportion_table),Proportion=as.numeric(proportion_table))
#kable(df)
No_high_schl<- prop.table(s$Ppl_below_pv)
High_schl<- prop.table(t$Ppl_below_pv)
Degree_or_more<- prop.table(u$Ppl_below_pv)
Race <- s$RAC1P
df1<-data.frame(Race,No_high_schl,High_schl,Degree_or_more)
kable(df1,caption = "Racial breakdown of people in extreme poverty with no health insurance at different educatio
n levels in the states of CA, FL, TX") %>% kable_styling(bootstrap_options = c("striped", "hover", "condensed"),f
ull_width = FALSE)
#[1]
```

[22]

```
# Jobs of people under poverty line,with no medicare along with their average number of hours worked per week.
s4 <- SSData2 %>%filter(POVPIP <= 125,(AGEP>18 & AGEP <=60), HICOV =="No health insurance coverage",!is.na(SOC
P),!is.na(WKHP),ST =="CA"|ST =="TX"|ST=="FL")  %>% group_by(SOCP) %>% summarise(PplBelowPoverty= sum(PWGTP),Worki
ngHours = weighted.mean(WKHP,PWGTP,na.rm = TRUE)) %>% arrange(desc(PplBelowPoverty)) %>% top_n(10)
PplBelowPoverty<- prop.table(s4$PplBelowPoverty) *100
WorkingHours<- (s4$WorkingHours)

Occupation <- s4$SOCP
df1<-data.frame(Occupation,PplBelowPoverty,WorkingHours)
kable(df1,caption = "Table :Percentage of people below poverty, with healthcare along with their average number o
f hours worked per week") %>% kable_styling(bootstrap_options = c("striped", "hover", "condensed"),full_width = F
ALSE)
```

[23]

```
# How does the working hour differ across different states for those above poverty
s4 <- SSData2 %>%filter(POVPIP > 125,(AGEP>18 & AGEP <=60),!is.na(WKHP))  %>% group_by(ST) %>% summarise(WorkingH
ours = weighted.mean(WKHP,PWGTP,na.rm = TRUE))
# How does the working hour differ across different states for the poor
s5 <- SSData2 %>%filter(POVPIP <= 125,(AGEP>18 & AGEP <=60),!is.na(WKHP))  %>% group_by(ST) %>% summarise(Working
Hours = weighted.mean(WKHP,PWGTP,na.rm = TRUE))

hcmap("countries/us/us-all", data = s4, name = c("WorkingHours"), value = "WorkingHours", joinBy = c("hc-a2", "S
T"), borderColor = "black",dataLabels = list(enabled = TRUE, format = '{point.name}'))%>%
hc_colorAxis(dataClasses = color_classes(c(seq(30, 44, by = 2))))%>%
hc_legend(layout = "vertical", align = "right",floating = TRUE, valueDecimals = 0)

hcmap("countries/us/us-all", data = s5, name = "WorkingHours", value = "WorkingHours", joinBy = c("hc-a2", "ST"),
borderColor = "transparent",dataLabels = list(enabled = TRUE, format = '{point.name}'))%>%
hc_colorAxis(dataClasses = color_classes(c(seq(28, 44, by = 2))))%>%
hc_legend(layout = "vertical", align = "right",floating = TRUE, valueDecimals = 0)
#[2]
```

References

[1] https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html (https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html) [2]http://jkunst.com/highcharter/highmaps.html (http://jkunst.com/highcharter/highmaps.html)