**TECHNICAL REPORT SPEA-V506**

2023-05-03

**Name**       :  KIRAN RANGWANI (kirarang@iu.edu)

      :  LOGAN GUNNING([lsgunnin@indiana.edu](mailto:lsgunnin@indiana.edu))

      : VASU SHARMA([vasushar@iu.edu](mailto:vasushar@iu.edu))

**Semester** :  Spring 2023

**Faculty**     :   Joanna Woronkowicz

Table of Contents

1. [**Abstract**](#Abstract)………………………………………………………………………………………………….
2. [**Introduction**](#Introduction)………………………………………………………………………..…………………..
3. [**Data**](#Data)…………………………………………………………………………………………………………

3.1 [Data Preprocessing and Cleaning](#DataPreprocessingandCleaning)…………………………………………………………

3.2 [Exploratory Data Analysis](#ExploratoryDataAnalysis)…………………………………………………………………….

4[**. Multiple Linear Regression Analysis**](#MultipleLinearRegression)…………………………………………………………

4.1 [MODEL 1 : Model with all variables](#MODEL1) ……………………………………………………..

4.2 [MODEL 2 : Model without English variable](#MODEL2)…………………………………………….

* 1. [MODEL 3 : Model without race](#MODEL3) ………………………………………………………………
  2. [MODEL 4 : Model without occupation](#MODEL4)………………………………………………………
  3. [MODEL 5 : Model without marriage](#MODEL5)………………………………………………………….
  4. [MODEL 6 : Model without education](#MODEL6)…………………………………………….
  5. [MODEL 7 : Model with only English](#MODEL7)……………………………………………..
  6. [MODEL 8 : Model with only few occupations](#MODEL8)…………………………………….
  7. [MODEL 9 : Model with sample east coast states](#MODEL9)…………………………………..

5**.**  [**Conclusions**](#Conclusion) …………………………………………………………………………………………………….

6. [**Bibliography**](#Bibliography)…………………………………………………………………………………………………….

**Abstract**:

This study employs a linear regression model to examine the relationship between English language proficiency and annual income outcomes in the United States, while controlling for other demographic and social variables, such as educational attainment, sex, profession, race, marital status, and age. The analysis is based on data obtained from the U.S. Census Bureau, and attempts to supplement existing research that has found a positive correlation between these variables. The results show that English language proficiency has a significant positive effect on annual income outcomes, expressed as median income, with an R-squared value of 0.3426 indicating that the model can explain 34.26% of the variation in the dependent variable. The regression coefficients in the model are statistically significant, with p-values smaller than 2.2e-16 and an F-statistic of 1.851e+04, suggesting an adequate fit of the models to the data. The other models, which exclude one or more predictor variables, yield lower R-squared values, implying less variation in the dependent variable. Though, an analysis of a subset of sample states yielded a higher R-squared value of 0.3544. Ultimately, the study concludes that language proficiency is a predictor of economic assimilation, supporting the need for programs that promote and enhance English language skills among non-native speakers to enhance economic opportunities and social mobility.

**Introduction**:

Based on Census Bureau data, nearly ten percent of individuals residing in the United States speak English less than “very well” (U.S. Census Bureau, 2021). In a primarily monolingual nation whose business, services, and economy at large is dominated by English, the effect of language fluency on earnings outcomes is thus an interesting and worthwhile avenue of research as it may relate to economic opportunity and social mobility. In fact, there is already a robust set of literature that sets out to establish the relationship between English language proficiency and economic assimilation, defined as earnings. Chaswick and Miller’s study, “Language in the Labor Market: The Immigrant Experience in Canada and the United States” (1990), for example, found that “fluency in the dominant language has a large positive effect on earnings, independent of other personal characteristics and country of origin” (p. 4). Cadena et al. (2015) came to this same conclusion as well, and found age of immigration to be a determinant of English proficiency, which minimizes earning deficits with native-born counterparts. Gill and Ahmad (2019) quantified the relationship, finding that “the results for English-language proficiency show that individuals with no proficiency earn about 26% less than individuals with native-level proficiency” (p. 40). Generally, the literature indicates a positive correlation of English language proficiency and earnings, suggesting economic assimilation to be facilitated linguistically. However, there is still a need to examine this relationship further.

The following regression analysis has endeavored to offer additional support to this body of research, addressing the question as to whether English language proficiency, holding other social and demographic variables constant, impacts an individual’s annual income in the United States. The linear model created to answer this research question uses data accessed from the U.S. Census Bureau website, and compares annual income outcomes, expressed in terms of median income, with the variable of interest: language proficiency in English, as well as educational attainment, sex, profession, race, marital status, and age. A sample analysis of the East Coast was conducted to offer a point of comparison. This particular research question offers policy implications as well. If it is found that income outcomes are at least partially mediated by English proficiency, this research may offer support to the establishment, funding, or maintenance of programs intending to benefit the economic well-being of individuals for whom English is not a first language. Thus, this study aims to further contribute to the existing research regarding language proficiency and economic outcomes, and inform evidence-based policy decisions with respect to English second language education.

**Data:**

Our research data set is based on Census data that contains a wide range of demographic, social, and economic variables such as age, sex, race, education, English proficiency, marital status, and occupation. This data is available from the data.usa.gov website, which provides access to data from various federal agencies. The data set covers all 50 states in the US and pertains to the year 2021, with income measured in dollars and other variables being unitless.

To prepare our data set, we performed data preprocessing and cleaning, including adding dummy variables for multiple linear regression analysis. Dummy variables are created for discrete variables. To avoid having too many dummy variables, we grouped some of the columns that had numerous categories by combining similar categories together. For instance, the dummy variable "graduate" contained 1 for individuals who had a bachelor's degree or higher and 0 having education level less than bachelors degree. This strategy helped us to use dummy variables more efficiently. Our analysis majorly focuses on the East Coast states, which represent approximately 36% of the US population. However, we acknowledge that our sample may differ from the overall population due to demographic, economic, political, and social variations, although the East Coast is known to be relatively diverse.

The states that are included in the sample are as follows:

1. Maine

2. New Hampshire

3. Massachusetts

4. Rhode Island

5. Connecticut

6. New York

7. New Jersey

8. Delaware

9. Maryland

10.       Virginia

11. North Carolina

12. South Carolina

13. Georgia

14. Florida

Our sample comprises respondents from above mentioned 14 states, with income as the response variable. We have included all individuals aged 16 and over, and removed negative values as they represent a loss for individuals or negative income for those below 15 years of age. To normalize the data and reduce the impact of extreme values, we applied a log transformation to the income variable, with income ranges from $1 to $4209995. We created a transformed income column by adding 1 to each row in the income column in order to avoid undefined values resulting from the logarithm of 0.

The income column is measured in US dollars, while the age column is measured in years. Other columns in the data set do not have units and are considered dimensionless.

**Data Preprocessing and Cleaning:**

The initial stage in the data cleaning and preprocessing phase was to remove any superfluous data that would not add to the research. This was done to enable a more concentrated analysis and higher results accuracy. After that, exploratory data analysis (EDA) was performed to acquire a better understanding of the data. Dummy variables were constructed to aid in regression analysis. This procedure was carried out in order to prepare the data for future analysis and to draw relevant inferences from the results.

##Reading data

##For analysis Perspective, better to go ahead with shorter Names only. # colnames(dataframe) <- c(“Income”, “Age”,“White recode”, # “Black or African American recode”,“Marital status”,“SEX”,“Educational attainment”, # “State”,“English Speaking Ability”,“Pacific Islander recode”,“Native Hawaiian recode”, # “American Indiana Alaska Native”, # “other race recode”,“Asian recode”,“Standard Occupational Classification”)

dataframe = read.csv("RAW-DATA.csv")

dim(dataframe)

## [1] 3252599 17

dataframe <- subset(dataframe, select = -X)  
colnames(dataframe)

## [1] "PWGTP" "PINCP" "AGEP" "RACWHT" "RACBLK" "MAR" "SEX"   
## [8] "SCHL" "POWSP" "ENG" "RACPI" "RACNH" "RACAIAN" "RACSOR"   
## [15] "RACASN" "SOCP"

###Data cleaning Making the data more efficient for our analysis

dataframe<-subset(dataframe, POWSP!="N")  
dim(dataframe)

## [1] 1448821 16

dataframe <- subset(dataframe, POWSP < "072")  
dim(dataframe)

## [1] 1448268 16

dataframe<-subset(dataframe, SOCP!="N")  
dim(dataframe)

## [1] 1448268 16

dataframe <- subset(dataframe, PINCP >= 0)

####RECODING ###Recoding Marital Status to dummy variables Married = 1 or else 0 Unmarried : rest of the categories as 1 and married as 0

dataframe$married <- as.factor(ifelse(dataframe$MAR==1, 1, 0))  
  
dataframe$unmarried <- as.factor(ifelse(dataframe$MAR %in% c(2,3,4,5), 1, 0))  
  
colnames(dataframe)

## [1] "PWGTP" "PINCP" "AGEP" "RACWHT" "RACBLK" "MAR"   
## [7] "SEX" "SCHL" "POWSP" "ENG" "RACPI" "RACNH"   
## [13] "RACAIAN" "RACSOR" "RACASN" "SOCP" "married" "unmarried"

###We recode ABILITY TO SPEAK ENGLISH to dummy variables as it's our variable of interest  
$OnlyEnglish = 0  
Very\_well = 1  
Well = 2  
Not\_well = 3  
Not\_at\_all = 4

dataframe$English <- as.factor(ifelse(dataframe$ENG == 0, "Only English",  
 ifelse(dataframe$ENG == 1, "Very well",  
 ifelse(dataframe$ENG == 2, "Well",  
 ifelse(dataframe$ENG == 3, "Not Well",  
 ifelse(dataframe$ENG == 4, " Not at all",  
 "Not at all"))))))

#Recoding Gender 1: male, 2:female

dataframe$Gender <- as.factor(ifelse(dataframe$SEX == 1, "Male",  
 ifelse(dataframe$SEX == 2, "Female","no")))

#Recoding EDUCATIONAL ATTAINMENT ##Considering values from 12th grade or higher, not considering values before 12th grade ##Here we choose Bachelors and above as one category

# Create a new data frame with values from 16 to 24 in the SCHL column  
dataframe <- subset(dataframe, SCHL >= 15 & SCHL <= 24 & SCHL!=17)  
dataframe$NoBachelors <- as.factor(ifelse(dataframe$SCHL %in% c(15, 16,18,19,20),1,0))  
  
dataframe$Bachelors <-as.factor(ifelse(dataframe$SCHL %in% c(21,22,23,24),1,0))  
  
colnames(dataframe)

## [1] "PWGTP" "PINCP" "AGEP" "RACWHT" "RACBLK"   
## [6] "MAR" "SEX" "SCHL" "POWSP" "ENG"   
## [11] "RACPI" "RACNH" "RACAIAN" "RACSOR" "RACASN"   
## [16] "SOCP" "married" "unmarried" "English" "Gender"   
## [21] "NoBachelors" "Bachelors"

##Occupation recode to ddummy variables ##We group records that fall under same unique code Some columns are not obvious so here’s what they mean :

PRT : Protective Services EATERY : Food Service CLN : Maintenance PRS : Personal Care Official : Administrative FFF : Agricultural CON : Construction EXT : Extraction RPR : Repairs TRN : Transport MIL : Military

dataframe$Managers= ifelse(grepl("^11", dataframe$SOCP),1,0)  
dataframe$Business= ifelse(grepl("^131", dataframe$SOCP),1,0)  
dataframe$Financial= as.factor(ifelse(grepl("^132", dataframe$SOCP),1,0))  
dataframe$IT= as.factor(ifelse(grepl("^15", dataframe$SOCP),1,0))  
dataframe$Engineering= as.factor(ifelse(grepl("^17", dataframe$SOCP), 1,0))  
dataframe$Science= as.factor(ifelse(grepl("^19", dataframe$SOCP), 1,0))  
dataframe$Coucelling= as.factor(ifelse(grepl("^21", dataframe$SOCP), 1,0))  
dataframe$Legal= as.factor(ifelse(grepl("^23", dataframe$SOCP), 1,0))  
dataframe$Education= as.factor(ifelse(grepl("^25", dataframe$SOCP), 1,0))  
dataframe$Entertainment= as.factor(ifelse(grepl("^27", dataframe$SOCP), 1,0))  
dataframe$Medical= as.factor(ifelse(grepl("^29", dataframe$SOCP), 1,0))  
dataframe$HealthCare= as.factor(ifelse(grepl("^31", dataframe$SOCP), 1,0))  
dataframe$PRT= as.factor(ifelse(grepl("^33", dataframe$SOCP), 1,0))  
dataframe$EATERY= as.factor(ifelse(grepl("^35", dataframe$SOCP), 1,0))  
dataframe$CLN= as.factor(ifelse(grepl("^37", dataframe$SOCP), 1,0))  
dataframe$PRS= as.factor(ifelse(grepl("^39", dataframe$SOCP), 1,0))  
dataframe$Sales= as.factor(ifelse(grepl("^41", dataframe$SOCP), 1,0))  
dataframe$Official= as.factor(ifelse(grepl("^43", dataframe$SOCP), 1,0))  
dataframe$FFF= as.factor(ifelse(grepl("^45", dataframe$SOCP), 1,0))  
dataframe$CON= as.factor(ifelse(grepl("^474", dataframe$SOCP),1,0))  
dataframe$EXT= as.factor(ifelse(grepl("^475", dataframe$SOCP),1,0))  
dataframe$RPR= as.factor(ifelse(grepl("^49", dataframe$SOCP), 1,0))  
dataframe$Production= as.factor(ifelse(grepl("^51", dataframe$SOCP), 1,0))  
dataframe$TRN= as.factor(ifelse(grepl("^53", dataframe$SOCP), 1,0))  
dataframe$MIL= as.factor(ifelse(grepl("^55", dataframe$SOCP), 1,0))  
dim(dataframe)

## [1] 1314196 47

colnames(dataframe)

## [1] "PWGTP" "PINCP" "AGEP" "RACWHT"   
## [5] "RACBLK" "MAR" "SEX" "SCHL"   
## [9] "POWSP" "ENG" "RACPI" "RACNH"   
## [13] "RACAIAN" "RACSOR" "RACASN" "SOCP"   
## [17] "married" "unmarried" "English" "Gender"   
## [21] "NoBachelors" "Bachelors" "Managers" "Business"   
## [25] "Financial" "IT" "Engineering" "Science"   
## [29] "Coucelling" "Legal" "Education" "Entertainment"  
## [33] "Medical" "HealthCare" "PRT" "EATERY"   
## [37] "CLN" "PRS" "Sales" "Official"   
## [41] "FFF" "CON" "EXT" "RPR"   
## [45] "Production" "TRN" "MIL"

class(dataframe$MIL)

## [1] "factor"

##As for state variable, we will only retain 50 states of the United States ##State ##072,166,251,254,301,303,399,555 are other countries hence we do not select those

us\_states <- sprintf("%03d", 001:056)  
  
dataframe <- subset(dataframe, POWSP %in% us\_states)  
colnames(dataframe)

## [1] "PWGTP" "PINCP" "AGEP" "RACWHT"   
## [5] "RACBLK" "MAR" "SEX" "SCHL"   
## [9] "POWSP" "ENG" "RACPI" "RACNH"   
## [13] "RACAIAN" "RACSOR" "RACASN" "SOCP"   
## [17] "married" "unmarried" "English" "Gender"   
## [21] "NoBachelors" "Bachelors" "Managers" "Business"   
## [25] "Financial" "IT" "Engineering" "Science"   
## [29] "Coucelling" "Legal" "Education" "Entertainment"  
## [33] "Medical" "HealthCare" "PRT" "EATERY"   
## [37] "CLN" "PRS" "Sales" "Official"   
## [41] "FFF" "CON" "EXT" "RPR"   
## [45] "Production" "TRN" "MIL"

##This will form our population. As for our sample we will select East code states and create dummy variables for those. #We will perform this part of recoding further down the project timeline

## Converting integer to discrete values

dataframe$ENG <- as.factor(dataframe$ENG)  
dataframe$SEX <- as.factor(dataframe$SEX)  
dataframe$Bachelors <- as.factor(dataframe$Bachelors)

##Dependent variable log transformation Log transformation performed because taking the logarithm of income can help to reduce the impact of these extreme values and make the distribution of income more symmetrical and easier to work with.

colnames(dataframe)

## [1] "PWGTP" "PINCP" "AGEP" "RACWHT"   
## [5] "RACBLK" "MAR" "SEX" "SCHL"   
## [9] "POWSP" "ENG" "RACPI" "RACNH"   
## [13] "RACAIAN" "RACSOR" "RACASN" "SOCP"   
## [17] "married" "unmarried" "English" "Gender"   
## [21] "NoBachelors" "Bachelors" "Managers" "Business"   
## [25] "Financial" "IT" "Engineering" "Science"   
## [29] "Coucelling" "Legal" "Education" "Entertainment"  
## [33] "Medical" "HealthCare" "PRT" "EATERY"   
## [37] "CLN" "PRS" "Sales" "Official"   
## [41] "FFF" "CON" "EXT" "RPR"   
## [45] "Production" "TRN" "MIL"

hist(dataframe$PINCP,main="Before log Transformation",col = "RED")

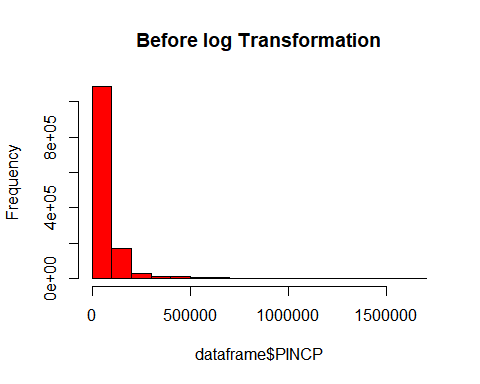


Fig 1 . Income value Before Log transformation done

dataframe$transformed\_income<-log(dataframe$PINCP+1)  
hist(dataframe$transformed\_income,main="After log Transformation",col = "BLUE")

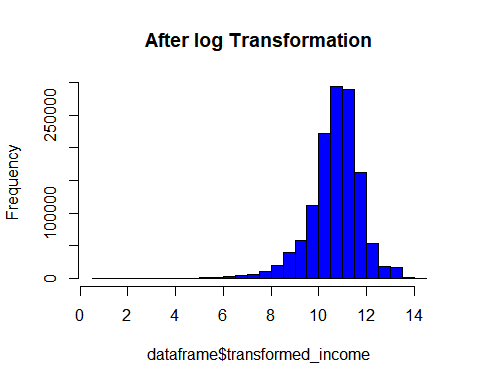


Fig 2 . Income value After Log transformation done

write.csv(dataframe,file = "tranformed\_v1.csv",row.names = FALSE)

dataframe = read.csv("tranformed\_v1.csv")

colnames(dataframe)

## [1] "PWGTP" "PINCP" "AGEP"   
## [4] "RACWHT" "RACBLK" "MAR"   
## [7] "SEX" "SCHL" "POWSP"   
## [10] "ENG" "RACPI" "RACNH"   
## [13] "RACAIAN" "RACSOR" "RACASN"   
## [16] "SOCP" "married" "unmarried"   
## [19] "English" "Gender" "NoBachelors"   
## [22] "Bachelors" "Managers" "Business"   
## [25] "Financial" "IT" "Engineering"   
## [28] "Science" "Coucelling" "Legal"   
## [31] "Education" "Entertainment" "Medical"   
## [34] "HealthCare" "PRT" "EATERY"   
## [37] "CLN" "PRS" "Sales"   
## [40] "Official" "FFF" "CON"   
## [43] "EXT" "RPR" "Production"   
## [46] "TRN" "MIL" "transformed\_income"

##Checking for NA values

sum(is.na(dataframe))

## [1] 0

dataframe <- na.omit(dataframe)  
nrow(dataframe)

## [1] 1314196

**Exploratory Data Analysis:**

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.2

library(reshape2)

## Warning: package 'reshape2' was built under R version 4.2.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.2

##   
## 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

#creating a subset has columns income and english proficiency for building plot below

income\_eng <- dataframe[, c("PINCP", "English")]

income\_eng\_mean\_median <- income\_eng %>%  
 group\_by(English) %>%  
 summarize(mean\_income = mean(PINCP), median\_income = median(PINCP))  
income\_eng\_mean\_median

## # A tibble: 5 × 3  
## English mean\_income median\_income  
## <chr> <dbl> <dbl>  
## 1 " Not at all" 35257. 26400  
## 2 "Not Well" 42080. 30000  
## 3 "Only English" 71313. 50000  
## 4 "Very well" 72458. 49000  
## 5 "Well" 54976. 38800

ggplot(income\_eng\_mean\_median, aes(x = English, y = median\_income, fill = English)) +  
 labs(x = "Ability to Speak English", y = "Median Income ($)") +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_discrete() +  
 scale\_x\_discrete(labels = c("Not at all", "Not Well", "Only English", "Very Well", "Well")) +  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text.y = element\_text(angle = 0, hjust = 1, size = 8),   
 axis.text.x = element\_text(angle = 90,size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 )+  
 labs(title = "Median Income by English Proficiency")+  
 labs(x = "Ability to Speak English", y = "Median Income ($)") +  
 labs(fill = "English Proficiency")

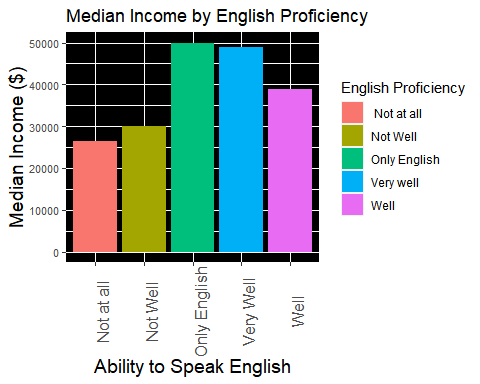


Fig 3 . Bar plot of Median Income vs. English Proficiency

ggplot(income\_eng\_mean\_median, aes(x = English, y = mean\_income, fill = English)) +  
 labs(x = "Ability to Speak English", y = "Mean Income ($)") +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_discrete() +  
 scale\_x\_discrete(labels = c("Not at all", "Not Well", "Only English", "Very Well", "Well")) +  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text.y = element\_text(angle = 0, hjust = 1, size = 8),  
 axis.text.x = element\_text(angle = 90,size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 )+  
 labs(title = "Mean Income by English Proficiency")+  
 labs(x = "Ability to Speak English", y = "Mean Income ($)") +  
 labs(fill = "English Proficiency")

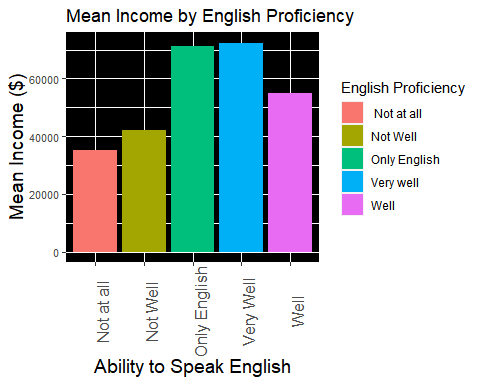


Fig 4 . Bar plot  of Mean Income vs English Proficiency

###creating a subset that has columns income and education : Bachelors for building plot below

income\_grad <- dataframe[, c("PINCP", "Bachelors")]   
  
income\_graduation <- income\_grad %>%  
 group\_by(Bachelors) %>%  
 summarize(mean\_income = mean(PINCP), median\_income = median(PINCP))  
income\_graduation

## # A tibble: 2 × 3  
## Bachelors mean\_income median\_income  
## <int> <dbl> <int>  
## 1 0 47150. 37000  
## 2 1 99754. 70580

income\_graduation <- income\_grad %>%  
 group\_by(Bachelors) %>%  
 summarize(mean\_income = mean(PINCP), median\_income = median(PINCP))  
  
income\_graduation$Bachelors <- factor(income\_graduation$Bachelors, levels = c("0", "1"), labels = c("Less than Bachelor", "Bachelor's or more"))

ggplot(income\_graduation, aes(x = Bachelors, y = median\_income, fill = Bachelors)) +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_discrete(name = "Education Level")+  
 scale\_x\_discrete(labels = c("No Bachelor's", "Bachelor's or more"))+  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text.y = element\_text(angle = 0, hjust = 1, size = 8),   
 axis.text.x = element\_text(angle = 90,size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 ) +  
 labs(title = "Median Income by Educational Attainment",  
 fill = "Graduate")+  
 xlab("Degree Holding Status")+  
 ylab("Median Income ($)")

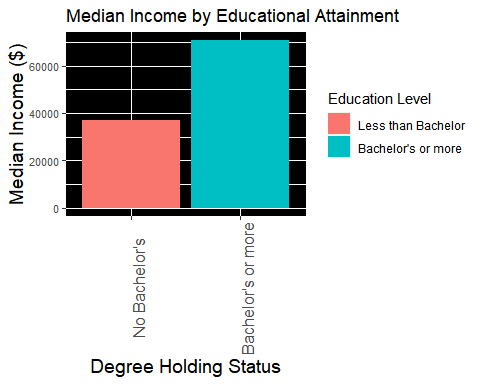


Fig 5 . Bar plot of Median Income vs Educaitonal Attainment

ggplot(income\_graduation, aes(x = Bachelors, y = mean\_income, fill = Bachelors)) +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_discrete(name = "Education Level") +  
 scale\_x\_discrete(labels = c("No Bachelor's", "Bachelor's or more"))+  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text.y = element\_text(angle = 0, hjust = 1, size = 8), # change the size of y-axis text  
 axis.text.x = element\_text(angle = 90,size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 )+  
 labs(title = "Mean Income by Educational Attainment",  
 fill = "Bachelors")+  
 xlab("Degree Holding Status")+  
 ylab("Mean Income ($)")

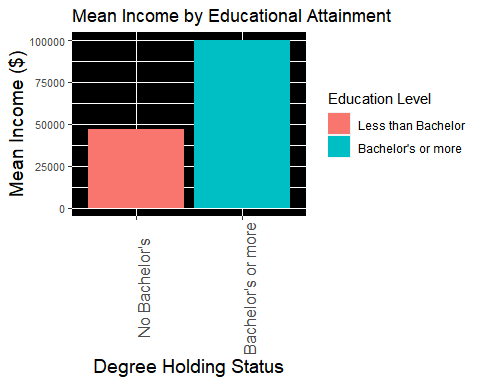


Fig 6 . Bar plot of Mean Income vs Educaitonal Attainment

###creating a subset that has columns income and Marital status : married for building plot below

income\_marital\_status <- dataframe[, c("PINCP", "married")]   
  
income\_marital\_status <- income\_marital\_status %>%  
 group\_by(married) %>%  
 summarize(mean\_income = mean(PINCP), median\_income = median(PINCP))  
income\_marital\_status$married <- factor(income\_marital\_status$married, levels = c("0", "1"), labels = c("Single", "Married"))  
income\_marital\_status

## # A tibble: 2 × 3  
## married mean\_income median\_income  
## <fct> <dbl> <int>  
## 1 Single 50506. 36900  
## 2 Married 85573. 60000

ggplot(income\_marital\_status, aes(x = married, y = median\_income, fill = married)) +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_discrete(name = "Marital Status") +  
 scale\_x\_discrete(labels = c("Single", "Married")) +  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text = element\_text(size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 )+  
 labs(title = "Median Income by Marital Status")+  
 xlab("Marital Status")+  
 ylab("Median Income ($)")

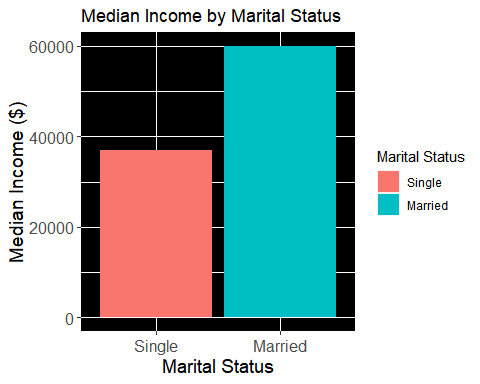


Fig 7 . Bar plot of Median Income vs Marital Status

ggplot(income\_marital\_status, aes(x = married, y = mean\_income, fill = married)) +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_discrete(name = "Marital Status") +  
 scale\_x\_discrete(labels = c("Single", "Married")) +  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text = element\_text(size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 )+  
 labs(title = "Mean Income by Marital Status")+  
 xlab("Marital Status")+  
 ylab("Mean Income ($)")

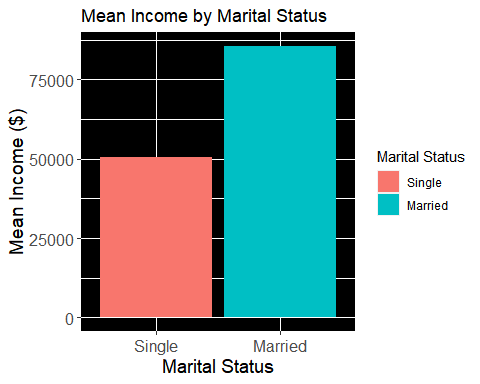


Fig 8 . Bar plot of Mean Income vs Marital Status

###creating a subset that has columns income and Gender for building plot below

income\_gender <- dataframe[, c("PINCP", "Gender")]   
  
income\_graduation <- income\_gender %>%  
 group\_by(Gender) %>%  
 summarize(mean\_income = mean(PINCP), median\_income = median(PINCP))  
income\_graduation

## # A tibble: 2 × 3  
## Gender mean\_income median\_income  
## <chr> <dbl> <int>  
## 1 Female 56224. 41000  
## 2 Male 83507. 57600

ggplot(income\_graduation, aes(x = Gender, y = median\_income, fill = Gender)) +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_discrete() +  
 scale\_x\_discrete(labels = c("Female", "Male")) +  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text = element\_text(size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 )+  
 labs(title = "Median Income by Gender")+  
 xlab("Gender")+  
 ylab("Median Income ($)")

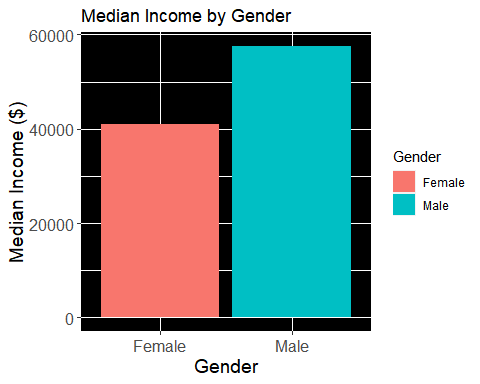


Fig 9 . Bar plot of Median Income vs Gender

ggplot(income\_graduation, aes(x = Gender, y = mean\_income, fill = Gender)) +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_discrete() +  
 scale\_x\_discrete(labels = c("Female", "Male")) +  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text = element\_text(size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 )+  
 labs(title = "Mean Income by Gender")+  
 xlab("Gender")+  
 ylab("Mean Income ($)")

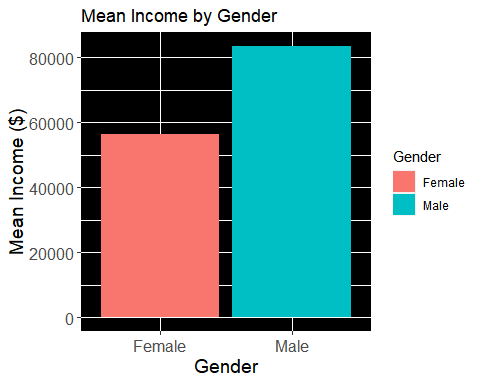


Fig 10 . Bar plot of Mean Income vs Gender

###creating a plot income and age

ggplot(dataframe, aes(x = dataframe$AGE, y = dataframe$PINCP, color = dataframe$AGE)) +   
 geom\_point() +   
 labs(title = "Age vs. Income", x = "Age", y = "Income")

## Warning: Use of `dataframe$AGE` is discouraged.  
## ℹ Use `AGE` instead.

## Warning: Use of `dataframe$PINCP` is discouraged.  
## ℹ Use `PINCP` instead.

## Warning: Use of `dataframe$AGE` is discouraged.  
## ℹ Use `AGE` instead.

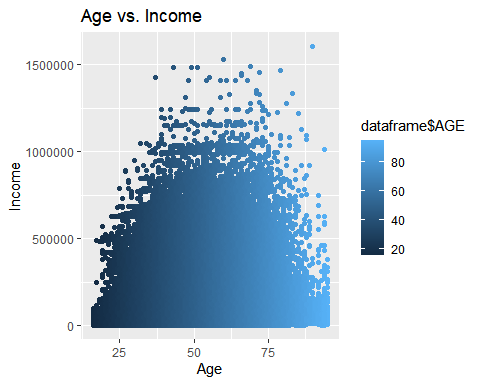


Fig 11 Scatterplot plot of Income vs Age

###creating a subset that has columns occupations and income

subset\_data\_profession <- dataframe[, c("PINCP", "Managers", "Business", "Financial", "IT", "Engineering", "Science", "Coucelling", "Legal", "Education", "Entertainment", "Medical", "HealthCare", "PRT", "EATERY", "CLN", "PRS", "Sales", "Official", "FFF", "CON", "EXT", "RPR", "Production", "TRN", "MIL")]  
  
subset\_melted\_data <- melt(subset\_data\_profession,id.vars = "PINCP", variable.name = "profession")  
subset\_melted\_data <-subset(subset\_melted\_data,value == 1)  
subset\_melted\_data <- subset\_melted\_data %>% rename(income = PINCP)  
income\_by\_profession <- aggregate(income ~ profession, subset\_melted\_data, median)

ggplot(income\_by\_profession, aes(x = profession, y = income, fill = profession)) +  
 geom\_bar(stat = "identity") +  
 scale\_fill\_discrete(name = "Profession") +  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text.y = element\_text(angle = 0, hjust = 1, size = 8), # change the size of y-axis text  
 axis.text.x = element\_text(angle = 90,size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 ) +  
 coord\_flip() +  
 labs(title = "Median Income by Professions", x = "Professions", y = "Median Income($)")

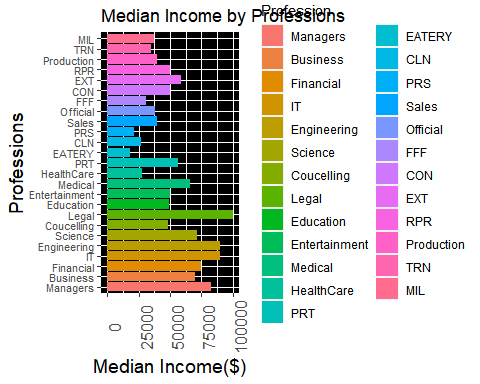


Fig 12. Bar plot of Professions vs Median Income

###creating a subset that has columns race and income

subset\_data\_race <- dataframe[, c("PINCP", "RACWHT", "RACBLK", "RACPI", "RACNH","RACAIAN","RACSOR","RACASN")]  
subset\_melted\_data\_race <- melt(subset\_data\_race,id.vars = "PINCP", variable.name = "race")  
subset\_melted\_data\_race <-subset(subset\_melted\_data\_race,value == 1)  
#income\_by\_race <- aggregate(PINCP ~ race, subset\_melted\_data\_race, mean)  
subset\_melted\_data\_race <- subset\_melted\_data\_race %>% rename(income = PINCP)  
  
income\_by\_race <- aggregate(income ~ race, subset\_melted\_data\_race, median)

ggplot(income\_by\_race, aes(x = race, y = income, fill = race)) +  
 geom\_bar(stat = "identity") +  
 scale\_x\_discrete(labels = c("White", "Black or African American","Pacific Islanders","Native Hawaiian","Native Alaskan","Other","Asian" )) +  
 scale\_fill\_discrete(name = "Race",labels = c("White", "Black or African American","Pacific Islanders","Native Hawaiian","Native Alaskan","Other","Asian" )) +  
 theme(  
 panel.background = element\_rect(fill = "black"),  
 axis.text.x = element\_text(angle = 90, hjust = 1, size = 12),  
 axis.text.y = element\_text(size = 12),  
 axis.title = element\_text(size = 14),  
 legend.position = "right"  
 )+  
 labs(title = "Median Income by Race")+  
 xlab("Race")+  
 ylab("Median Income ($)")

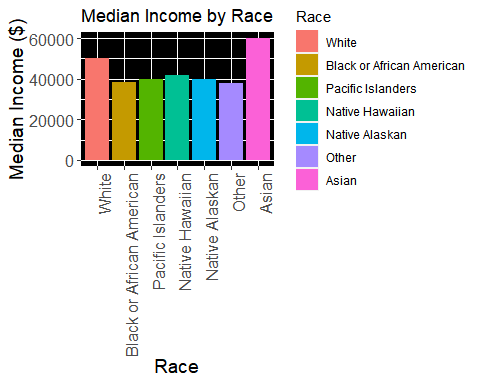


Fig 13. Bar plot of Median Income vs Race

**Multiple Linear Regression:**

The goal of this study is to look into the relationship between income and a variety of independent variables such as age, gender, occupation, education, English proficiency, race, and marital status. English proficiency is the variable of interest in this study, and the hypothesis is that those with better English proficiency have a higher median income than those with lower proficiency.

In this project, multiple linear regression is utilized to describe the connection between the dependent variable (income) and the seven independent factors. The linear regression method is suited for this investigation because it allows us to examine the relationship between a continuous dependent variable and one or more independent factors.

We performed eight multiple linear regression models in this project to examine the impact of various independent factors on income.

The models created are as follows :

**MODEL 1 :- MODEL WITH ALL VARIABLES CONSIDERED**

In addition to the dependent variable (income), this model contains all seven independent variables (age, gender, occupation, education, English proficiency, race, and marriage). The goal of this model is to determine the overall effect of all independent factors on the dependent variable and which variables have the largest impact on income.

**R Code Implementation** :

Removing married and NoBachelors as we don’t need those{Unmarried and Bachelors cover all the values}

model <- lm(transformed\_income ~ ENG+AGEP + SEX + RACWHT + RACAIAN + RACSOR + RACASN + RACBLK + RACNH + POWSP +Bachelors+unmarried+ Managers + Business + Financial + IT + Engineering + Science + Coucelling + Legal + Education + Entertainment + Medical + HealthCare + PRT + EATERY + CLN + PRS + Sales + Official + FFF + CON + EXT + RPR + Production + TRN + MIL, data = dataframe,na.action=na.exclude)

summary(model)

##   
## Call:  
## lm(formula = transformed\_income ~ ENG + AGEP + SEX + RACWHT +   
## RACAIAN + RACSOR + RACASN + RACBLK + RACNH + POWSP + Bachelors +   
## unmarried + Managers + Business + Financial + IT + Engineering +   
## Science + Coucelling + Legal + Education + Entertainment +   
## Medical + HealthCare + PRT + EATERY + CLN + PRS + Sales +   
## Official + FFF + CON + EXT + RPR + Production + TRN + MIL,   
## data = dataframe, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.3396 -0.3589 0.1089 0.5077 4.1045   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.031e+01 6.337e-03 1626.444 < 2e-16 \*\*\*  
## ENG -7.404e-02 1.457e-03 -50.813 < 2e-16 \*\*\*  
## AGEP 1.731e-02 5.517e-05 313.805 < 2e-16 \*\*\*  
## SEX -3.144e-01 1.713e-03 -183.514 < 2e-16 \*\*\*  
## RACWHT 4.100e-03 3.300e-03 1.242 0.21411   
## RACAIAN -4.508e-02 4.736e-03 -9.520 < 2e-16 \*\*\*  
## RACSOR 1.255e-03 2.963e-03 0.424 0.67186   
## RACASN 2.708e-02 4.072e-03 6.650 2.92e-11 \*\*\*  
## RACBLK -9.330e-02 3.898e-03 -23.933 < 2e-16 \*\*\*  
## RACNH 7.461e-03 1.704e-02 0.438 0.66149   
## POWSP -1.205e-03 4.801e-05 -25.095 < 2e-16 \*\*\*  
## Bachelors 4.921e-01 1.822e-03 270.103 < 2e-16 \*\*\*  
## unmarried -2.344e-01 1.662e-03 -141.059 < 2e-16 \*\*\*  
## Managers 4.235e-01 4.798e-03 88.266 < 2e-16 \*\*\*  
## Business 3.260e-01 5.747e-03 56.726 < 2e-16 \*\*\*  
## Financial 3.708e-01 6.470e-03 57.314 < 2e-16 \*\*\*  
## IT 4.712e-01 5.718e-03 82.419 < 2e-16 \*\*\*  
## Engineering 3.876e-01 6.377e-03 60.782 < 2e-16 \*\*\*  
## Science 2.363e-01 7.815e-03 30.237 < 2e-16 \*\*\*  
## Coucelling -1.569e-01 6.956e-03 -22.561 < 2e-16 \*\*\*  
## Legal 5.343e-01 7.676e-03 69.605 < 2e-16 \*\*\*  
## Education -2.257e-01 5.307e-03 -42.528 < 2e-16 \*\*\*  
## Entertainment -1.926e-01 6.668e-03 -28.889 < 2e-16 \*\*\*  
## Medical 3.705e-01 5.271e-03 70.282 < 2e-16 \*\*\*  
## HealthCare -2.818e-01 6.221e-03 -45.295 < 2e-16 \*\*\*  
## PRT 1.515e-01 6.673e-03 22.700 < 2e-16 \*\*\*  
## EATERY -6.561e-01 5.821e-03 -112.709 < 2e-16 \*\*\*  
## CLN -4.989e-01 6.344e-03 -78.645 < 2e-16 \*\*\*  
## PRS -6.285e-01 6.675e-03 -94.156 < 2e-16 \*\*\*  
## Sales -9.197e-02 4.934e-03 -18.641 < 2e-16 \*\*\*  
## Official -1.038e-01 4.880e-03 -21.273 < 2e-16 \*\*\*  
## FFF -3.263e-01 1.220e-02 -26.741 < 2e-16 \*\*\*  
## CON 4.666e-02 1.707e-02 2.733 0.00628 \*\*   
## EXT 2.989e-01 2.500e-02 11.957 < 2e-16 \*\*\*  
## RPR 1.051e-01 6.007e-03 17.503 < 2e-16 \*\*\*  
## Production -3.011e-02 5.396e-03 -5.579 2.42e-08 \*\*\*  
## TRN -2.268e-01 5.098e-03 -44.490 < 2e-16 \*\*\*  
## MIL 1.087e-01 1.173e-02 9.264 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8722 on 1314158 degrees of freedom  
## Multiple R-squared: 0.3426, Adjusted R-squared: 0.3426   
## F-statistic: 1.851e+04 on 37 and 1314158 DF, p-value: < 2.2e-16

confint(model)

## 2.5 % 97.5 %  
## (Intercept) 10.294172032 10.319012196  
## ENG -0.076893297 -0.071181773  
## AGEP 0.017204319 0.017420579  
## SEX -0.317766713 -0.311050816  
## RACWHT -0.002368024 0.010567243  
## RACAIAN -0.054366203 -0.035802657  
## RACSOR -0.004551733 0.007061696  
## RACASN 0.019099238 0.035061022  
## RACBLK -0.100938020 -0.085656852  
## RACNH -0.025934556 0.040855651  
## POWSP -0.001298942 -0.001110743  
## Bachelors 0.488506645 0.495648037  
## unmarried -0.237683145 -0.231168620  
## Managers 0.414082544 0.432889838  
## Business 0.314743790 0.337271777  
## Financial 0.358133137 0.383494359  
## IT 0.460040588 0.482453576  
## Engineering 0.375100759 0.400097507  
## Science 0.220983422 0.251617477  
## Coucelling -0.170560119 -0.143294760  
## Legal 0.519235974 0.549324932  
## Education -0.236086240 -0.215283991  
## Entertainment -0.205698701 -0.179561070  
## Medical 0.360137111 0.380799609  
## HealthCare -0.293967222 -0.269581836  
## PRT 0.138397341 0.164554418  
## EATERY -0.667525607 -0.644706416  
## CLN -0.511378831 -0.486509903  
## PRS -0.641586790 -0.615420655  
## Sales -0.101637489 -0.082297889  
## Official -0.113387474 -0.094256337  
## FFF -0.350186913 -0.302358673  
## CON 0.013192561 0.080121554  
## EXT 0.249938133 0.347945667  
## RPR 0.093374453 0.116922684  
## Production -0.040683542 -0.019530471  
## TRN -0.236802418 -0.216818652  
## MIL 0.085671915 0.131650213

anova(model)

## Analysis of Variance Table  
##   
## Response: transformed\_income  
## Df Sum Sq Mean Sq F value Pr(>F)   
## ENG 1 7575 7575 9.9585e+03 < 2.2e-16 \*\*\*  
## AGEP 1 151198 151198 1.9877e+05 < 2.2e-16 \*\*\*  
## SEX 1 44311 44311 5.8254e+04 < 2.2e-16 \*\*\*  
## RACWHT 1 1463 1463 1.9228e+03 < 2.2e-16 \*\*\*  
## RACAIAN 1 775 775 1.0195e+03 < 2.2e-16 \*\*\*  
## RACSOR 1 1855 1855 2.4382e+03 < 2.2e-16 \*\*\*  
## RACASN 1 10808 10808 1.4209e+04 < 2.2e-16 \*\*\*  
## RACBLK 1 2220 2220 2.9191e+03 < 2.2e-16 \*\*\*  
## RACNH 1 120 120 1.5723e+02 < 2.2e-16 \*\*\*  
## POWSP 1 729 729 9.5810e+02 < 2.2e-16 \*\*\*  
## Bachelors 1 159055 159055 2.0910e+05 < 2.2e-16 \*\*\*  
## unmarried 1 24036 24036 3.1599e+04 < 2.2e-16 \*\*\*  
## Managers 1 22888 22888 3.0090e+04 < 2.2e-16 \*\*\*  
## Business 1 5472 5472 7.1937e+03 < 2.2e-16 \*\*\*  
## Financial 1 4377 4377 5.7545e+03 < 2.2e-16 \*\*\*  
## IT 1 12690 12690 1.6683e+04 < 2.2e-16 \*\*\*  
## Engineering 1 6337 6337 8.3306e+03 < 2.2e-16 \*\*\*  
## Science 1 1625 1625 2.1369e+03 < 2.2e-16 \*\*\*  
## Coucelling 1 195 195 2.5597e+02 < 2.2e-16 \*\*\*  
## Legal 1 7049 7049 9.2673e+03 < 2.2e-16 \*\*\*  
## Education 1 2086 2086 2.7427e+03 < 2.2e-16 \*\*\*  
## Entertainment 1 451 451 5.9294e+02 < 2.2e-16 \*\*\*  
## Medical 1 22121 22121 2.9082e+04 < 2.2e-16 \*\*\*  
## HealthCare 1 331 331 4.3534e+02 < 2.2e-16 \*\*\*  
## PRT 1 2669 2669 3.5085e+03 < 2.2e-16 \*\*\*  
## EATERY 1 11432 11432 1.5030e+04 < 2.2e-16 \*\*\*  
## CLN 1 4616 4616 6.0686e+03 < 2.2e-16 \*\*\*  
## PRS 1 7943 7943 1.0443e+04 < 2.2e-16 \*\*\*  
## Sales 1 5 5 6.1142e+00 0.01341 \*   
## Official 1 66 66 8.6167e+01 < 2.2e-16 \*\*\*  
## FFF 1 383 383 5.0359e+02 < 2.2e-16 \*\*\*  
## CON 1 34 34 4.5052e+01 1.919e-11 \*\*\*  
## EXT 1 166 166 2.1805e+02 < 2.2e-16 \*\*\*  
## RPR 1 1418 1418 1.8644e+03 < 2.2e-16 \*\*\*  
## Production 1 496 496 6.5158e+02 < 2.2e-16 \*\*\*  
## TRN 1 1840 1840 2.4184e+03 < 2.2e-16 \*\*\*  
## MIL 1 65 65 8.5822e+01 < 2.2e-16 \*\*\*  
## Residuals 1314158 999616 1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

par(mfrow=c(2,2))  
plot(model)

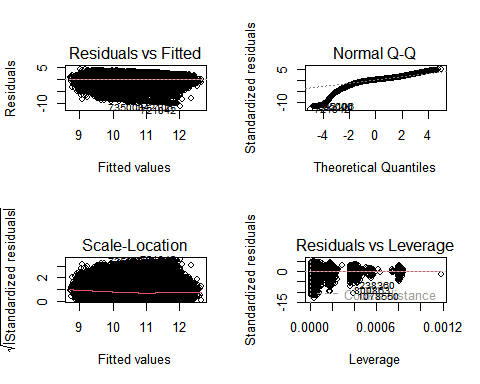


Fig 13. Various plots created for the above model showing how good the model is fitting the data

plot(model, which = 1)

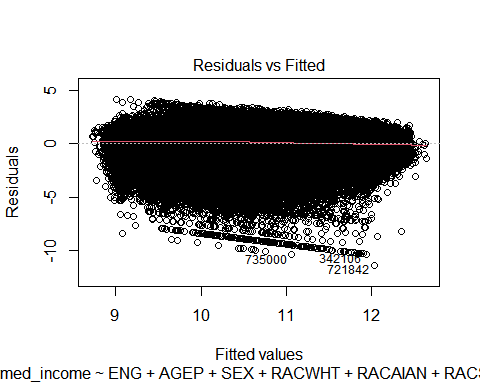


Fig 14. Residual vs Fitted values for MODEL 1 with all variables

install.packages("ggplot2", repos = "http://cran.us.r-project.org")

## Warning: package 'ggplot2' is in use and will not be installed

install.packages("coefplot", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/vasuv/AppData/Local/R/win-library/4.2'  
## (as 'lib' is unspecified)

## package 'coefplot' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\vasuv\AppData\Local\Temp\RtmpkNMUcQ\downloaded\_packages

library(ggplot2)  
library(coefplot)

## Warning: package 'coefplot' was built under R version 4.2.3

coefplot(model,color= )

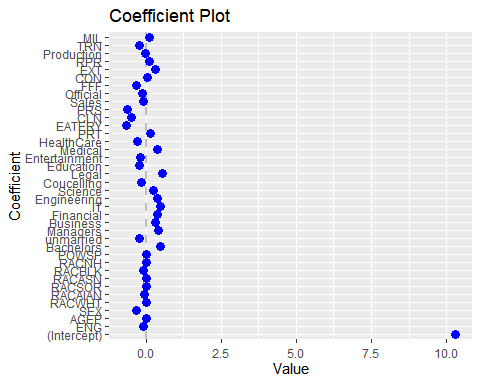


Fig 15. Coefficient plot for MODEL 1 with all variables

**Result :**

The R-squared value of the model is 0.3426, suggesting that the model explains 34.26% of the variation in the transformed income. The adjusted R-squared value is 0.3426, which is the same as the R-squared value and takes into account the number of predictors. The F-statistic has a p-value less than 2.2e-16, indicating that the model is significant overall.

Looking at the coefficients, we can see that all of the variables have statistically significant coefficients, with the exception of RACWHT(WHITE), RACSOR(OTHER), and RACNH(NATIVE HAWAIAN), which all have p-values greater than 0.05. The intercept coefficient is 10.31, which means that when all of the independent variables are zero, the average transformed income is 10.31.

In terms of interpretation, the coefficient for ENG is -0.07404, which means that for every unit increase in ENG, the income drops by 0.07404 while all other variables remain constant. The AGEP coefficient is 0.01731, which means that for each additional year of age, the transformed income increases by 0.01731 while all other variables remain constant. The SEX coefficient is -0.3144, showing that women have lower income than men on average. RACAIAN and RACBLK have negative effects on income, while RACASN has a favorable effect, according to the coefficients for the various racial categories.

Bachelors, Managers, Business, Financial, Information Technology, Engineering, Science, Legal, Medical, and PRT have favorable benefits on income, while unmarried, Counseling, Education, Entertainment, and HealthCare have negative effects.

The total F-statistic for the model with all variables is significant, and the R-squared value of 0.3426 indicates that the model explains 34.26% of the variance in income. Some individual coefficients, however, are not significant, and some variables have a negative effect on income.

**MODEL 2 :- MODEL WITHOUT ENGLISH VARIABLE :**

This model is similar to the first, but it lacks the English proficiency variable. The goal of this model is to determine how much English proficiency affects income in comparison to the other independent variables. If the other factors in this model have the same influence on income as the previous model, it can be inferred that English proficiency is not a significant contribution to income.

**R Code Implemetation** :

model\_without\_eng <- lm(transformed\_income ~ AGEP + SEX + RACWHT + RACAIAN + RACSOR + RACASN + RACBLK + RACNH + POWSP + unmarried + Bachelors + Managers + Business + Financial + IT + Engineering + Science + Coucelling + Legal + Education + Entertainment + Medical + HealthCare + PRT + EATERY + CLN + PRS + Sales + Official + FFF + CON + EXT + RPR + Production + TRN + MIL, data = dataframe,na.action=na.exclude)

summary(model\_without\_eng)

##   
## Call:  
## lm(formula = transformed\_income ~ AGEP + SEX + RACWHT + RACAIAN +   
## RACSOR + RACASN + RACBLK + RACNH + POWSP + unmarried + Bachelors +   
## Managers + Business + Financial + IT + Engineering + Science +   
## Coucelling + Legal + Education + Entertainment + Medical +   
## HealthCare + PRT + EATERY + CLN + PRS + Sales + Official +   
## FFF + CON + EXT + RPR + Production + TRN + MIL, data = dataframe,   
## na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.2678 -0.3596 0.1094 0.5086 4.1114   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.027e+01 6.292e-03 1631.603 < 2e-16 \*\*\*  
## AGEP 1.723e-02 5.520e-05 312.163 < 2e-16 \*\*\*  
## SEX -3.152e-01 1.715e-03 -183.811 < 2e-16 \*\*\*  
## RACWHT 2.987e-02 3.264e-03 9.152 < 2e-16 \*\*\*  
## RACAIAN -4.580e-02 4.740e-03 -9.662 < 2e-16 \*\*\*  
## RACSOR -5.148e-02 2.778e-03 -18.533 < 2e-16 \*\*\*  
## RACASN -1.369e-02 3.996e-03 -3.425 0.000615 \*\*\*  
## RACBLK -7.314e-02 3.882e-03 -18.842 < 2e-16 \*\*\*  
## RACNH 3.912e-02 1.704e-02 2.295 0.021717 \*   
## POWSP -1.120e-03 4.803e-05 -23.329 < 2e-16 \*\*\*  
## unmarried -2.305e-01 1.662e-03 -138.732 < 2e-16 \*\*\*  
## Bachelors 4.922e-01 1.824e-03 269.922 < 2e-16 \*\*\*  
## Managers 4.359e-01 4.796e-03 90.881 < 2e-16 \*\*\*  
## Business 3.396e-01 5.746e-03 59.103 < 2e-16 \*\*\*  
## Financial 3.831e-01 6.472e-03 59.191 < 2e-16 \*\*\*  
## IT 4.828e-01 5.719e-03 84.426 < 2e-16 \*\*\*  
## Engineering 3.984e-01 6.380e-03 62.456 < 2e-16 \*\*\*  
## Science 2.460e-01 7.820e-03 31.454 < 2e-16 \*\*\*  
## Coucelling -1.451e-01 6.958e-03 -20.848 < 2e-16 \*\*\*  
## Legal 5.490e-01 7.678e-03 71.507 < 2e-16 \*\*\*  
## Education -2.143e-01 5.307e-03 -40.370 < 2e-16 \*\*\*  
## Entertainment -1.806e-01 6.670e-03 -27.070 < 2e-16 \*\*\*  
## Medical 3.830e-01 5.271e-03 72.676 < 2e-16 \*\*\*  
## HealthCare -2.795e-01 6.227e-03 -44.894 < 2e-16 \*\*\*  
## PRT 1.641e-01 6.675e-03 24.591 < 2e-16 \*\*\*  
## EATERY -6.527e-01 5.827e-03 -112.022 < 2e-16 \*\*\*  
## CLN -5.036e-01 6.350e-03 -79.306 < 2e-16 \*\*\*  
## PRS -6.238e-01 6.681e-03 -93.366 < 2e-16 \*\*\*  
## Sales -8.207e-02 4.935e-03 -16.631 < 2e-16 \*\*\*  
## Official -9.217e-02 4.880e-03 -18.888 < 2e-16 \*\*\*  
## FFF -3.330e-01 1.221e-02 -27.266 < 2e-16 \*\*\*  
## CON 5.736e-02 1.709e-02 3.357 0.000789 \*\*\*  
## EXT 3.117e-01 2.503e-02 12.456 < 2e-16 \*\*\*  
## RPR 1.143e-01 6.010e-03 19.025 < 2e-16 \*\*\*  
## Production -2.612e-02 5.401e-03 -4.837 1.32e-06 \*\*\*  
## TRN -2.217e-01 5.102e-03 -43.451 < 2e-16 \*\*\*  
## MIL 1.211e-01 1.174e-02 10.319 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.873 on 1314159 degrees of freedom  
## Multiple R-squared: 0.3413, Adjusted R-squared: 0.3413   
## F-statistic: 1.891e+04 on 36 and 1314159 DF, p-value: < 2.2e-16

**Result** :

Looking at the output, the model explains a considerable percentage of the variability in the response variable, as evidenced by the F-statistic's very low p-value (less than 2e-16). After adjusting for the number of predictors, the adjusted R-squared value is 0.3413, suggesting that this model explains approximately 34% of the variation in the response variable.

Although there are some minor differences, the coefficients of the variables in the model without English variables are generally similar in magnitude to the coefficients in the model with all variables. This shows that, when compared to the other variables in the model, the English language proficiency variables may not have a large influence on the response variable. Overall, the model without English variables fits the data well and provides a plausible explanation for the link between the independent factors and the response variable.

**MODEL 3 :- WITHOUT RACE VARIABLE :**

The goal of this model is to determine how much race variable affects income in comparison to the other independent variables. A model that excludes race is critical since race is a sensitive and contentious issue that might add bias into the analysis. Race is a sociological, not a biological construct that is frequently used to stereotype and discriminate against particular groups of people. We can avoid spreading damaging assumptions and biases by omitting race as an independent variable in the model and focusing on other characteristics that are more directly related to wealth.

When race is included as an independent variable in a multiple linear regression model, it might cause multicollinearity, which happens when two or more independent variables are significantly associated with one another. This can result in unstable and unreliable regression coefficient estimates, and in some situations, false inferences regarding the relationship between the independent factors and the dependent variable.

**R Code Implementation**:

model\_without\_RACE <- lm(transformed\_income ~ ENG+AGEP + SEX + POWSP + unmarried+ Bachelors + Managers + Business + Financial + IT + Engineering + Science + Coucelling + Legal + Education + Entertainment + Medical + HealthCare + PRT + EATERY + CLN + PRS + Sales + Official + FFF + CON + EXT + RPR + Production + TRN + MIL, data = dataframe,na.action=na.exclude)

summary(model\_without\_RACE)

##   
## Call:  
## lm(formula = transformed\_income ~ ENG + AGEP + SEX + POWSP +   
## unmarried + Bachelors + Managers + Business + Financial +   
## IT + Engineering + Science + Coucelling + Legal + Education +   
## Entertainment + Medical + HealthCare + PRT + EATERY + CLN +   
## PRS + Sales + Official + FFF + CON + EXT + RPR + Production +   
## TRN + MIL, data = dataframe, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.3161 -0.3597 0.1090 0.5081 4.1165   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.031e+01 5.424e-03 1900.455 < 2e-16 \*\*\*  
## ENG -6.887e-02 1.221e-03 -56.422 < 2e-16 \*\*\*  
## AGEP 1.728e-02 5.489e-05 314.772 < 2e-16 \*\*\*  
## SEX -3.161e-01 1.714e-03 -184.455 < 2e-16 \*\*\*  
## POWSP -1.206e-03 4.780e-05 -25.234 < 2e-16 \*\*\*  
## unmarried -2.412e-01 1.652e-03 -145.967 < 2e-16 \*\*\*  
## Bachelors 4.962e-01 1.811e-03 273.968 < 2e-16 \*\*\*  
## Managers 4.233e-01 4.797e-03 88.254 < 2e-16 \*\*\*  
## Business 3.248e-01 5.746e-03 56.537 < 2e-16 \*\*\*  
## Financial 3.712e-01 6.468e-03 57.388 < 2e-16 \*\*\*  
## IT 4.735e-01 5.696e-03 83.131 < 2e-16 \*\*\*  
## Engineering 3.904e-01 6.372e-03 61.280 < 2e-16 \*\*\*  
## Science 2.386e-01 7.811e-03 30.543 < 2e-16 \*\*\*  
## Coucelling -1.650e-01 6.955e-03 -23.723 < 2e-16 \*\*\*  
## Legal 5.345e-01 7.679e-03 69.608 < 2e-16 \*\*\*  
## Education -2.274e-01 5.308e-03 -42.843 < 2e-16 \*\*\*  
## Entertainment -1.912e-01 6.669e-03 -28.678 < 2e-16 \*\*\*  
## Medical 3.704e-01 5.266e-03 70.348 < 2e-16 \*\*\*  
## HealthCare -2.914e-01 6.211e-03 -46.912 < 2e-16 \*\*\*  
## PRT 1.427e-01 6.670e-03 21.390 < 2e-16 \*\*\*  
## EATERY -6.581e-01 5.820e-03 -113.087 < 2e-16 \*\*\*  
## CLN -5.052e-01 6.345e-03 -79.616 < 2e-16 \*\*\*  
## PRS -6.292e-01 6.672e-03 -94.306 < 2e-16 \*\*\*  
## Sales -9.190e-02 4.933e-03 -18.632 < 2e-16 \*\*\*  
## Official -1.064e-01 4.878e-03 -21.807 < 2e-16 \*\*\*  
## FFF -3.256e-01 1.221e-02 -26.670 < 2e-16 \*\*\*  
## CON 4.567e-02 1.708e-02 2.673 0.00751 \*\*   
## EXT 2.992e-01 2.502e-02 11.960 < 2e-16 \*\*\*  
## RPR 1.053e-01 6.009e-03 17.519 < 2e-16 \*\*\*  
## Production -3.279e-02 5.395e-03 -6.078 1.22e-09 \*\*\*  
## TRN -2.347e-01 5.093e-03 -46.077 < 2e-16 \*\*\*  
## MIL 1.031e-01 1.173e-02 8.785 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8727 on 1314164 degrees of freedom  
## Multiple R-squared: 0.3418, Adjusted R-squared: 0.3418   
## F-statistic: 2.202e+04 on 31 and 1314164 DF, p-value: < 2.2e-16

**Result** :

When the model without racial variables is compared to the model with all variables, the former has an R-squared value of 0.3426 and an adjusted R-squared value of 0.3426. These figures are lower than the latter model's R-squared value of 0.3418 and adjusted R-squared value of 0.3418. This suggests that the model that includes all variables explains more of the variation in the target variable (income) than the model that excludes race variables.

Without racial factors, the p-values for all variables in the model are less than 0.05, indicating that all variables are statistically significant in predicting the target variable.

The model's F-statistic without racial variables is 2.202e+04  with 31 degrees of freedom, indicating that it is statistically significant. The p-value for the F-statistic is less than 2.2e-16, which is substantially lower than 0.05, indicating that the model is statistically significant.

Looking at the model coefficients without race variables, we can observe that all variables have negative coefficients except AGEP (age), which has a positive coefficient. This suggests that when the predictor variables' values grow, the anticipated income drops, with the exception of age, which results in an increase in income

**MODEL 4 :-MODEL WITHOUT OCCUPATION :**

The goal of this model is to determine how much occupation variables affect income in comparison to the other independent variables. For numerous reasons, the model without occupation must be performed as part of the analysis: Occupation and income are frequently significantly associated, implying that occupation may be a good predictor of income. Inclusion of both variables in the same model, however, may result in collinearity concerns, in which the two variables are too tightly related and make it impossible for the model to evaluate their individual effects on income. We may examine whether other variables, such as education, English proficiency, or age, have a greater impact on income by removing the employment variable from the model.

Generalizability: By removing employment from the model, we may test the model's generalizability to individuals with different sorts of jobs. Incorporating occupation into the model may result in a model that only applies to certain occupations and not others. By deleting this variable, we may test if the model applies to a broader group of people.

Discrimination: A sensitive variable that can lead to discrimination is occupation. may avoid any potential difficulties associated with utilizing occupation as a predictor variable by omitting it from the model.

**R Code Implementation**:

model\_without\_occup <- lm(transformed\_income ~ ENG+AGEP + SEX + POWSP +RACWHT + RACAIAN + RACSOR + RACASN + RACBLK + RACNH +unmarried + Bachelors, data = dataframe,na.action=na.exclude)

summary(model\_without\_occup)

##   
## Call:  
## lm(formula = transformed\_income ~ ENG + AGEP + SEX + POWSP +   
## RACWHT + RACAIAN + RACSOR + RACASN + RACBLK + RACNH + unmarried +   
## Bachelors, data = dataframe, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.1431 -0.3976 0.1197 0.5501 3.8537   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.034e+01 5.396e-03 1915.609 < 2e-16 \*\*\*  
## ENG -1.099e-01 1.531e-03 -71.785 < 2e-16 \*\*\*  
## AGEP 1.840e-02 5.766e-05 319.091 < 2e-16 \*\*\*  
## SEX -3.804e-01 1.615e-03 -235.539 < 2e-16 \*\*\*  
## POWSP -1.338e-03 5.071e-05 -26.377 < 2e-16 \*\*\*  
## RACWHT 1.326e-02 3.486e-03 3.803 0.000143 \*\*\*  
## RACAIAN -5.238e-02 5.004e-03 -10.467 < 2e-16 \*\*\*  
## RACSOR -7.517e-03 3.129e-03 -2.402 0.016297 \*   
## RACASN 8.774e-02 4.286e-03 20.474 < 2e-16 \*\*\*  
## RACBLK -1.205e-01 4.113e-03 -29.289 < 2e-16 \*\*\*  
## RACNH -4.394e-02 1.800e-02 -2.440 0.014668 \*   
## unmarried -2.935e-01 1.745e-03 -168.211 < 2e-16 \*\*\*  
## Bachelors 6.799e-01 1.661e-03 409.265 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9217 on 1314183 degrees of freedom  
## Multiple R-squared: 0.2658, Adjusted R-squared: 0.2658   
## F-statistic: 3.965e+04 on 12 and 1314183 DF, p-value: < 2.2e-16

**Result** :

R-squared multiple: 0.2658

R-squared adjusted: 0.2658

p-value: 2.2e-16, F-statistic: 3.965e+04 on 12 and 1314183 DF

All of the coefficients are significant at the p-value 0.001 level.

The R-squared and adjusted R-squared values are lower when compared to the model with all variables. It implies that the model's performance is influenced by the excluded variable (occupation). The F-statistic is likewise lower, showing that the model without occupation variables does not explain the variation in the response variable as well as the model with all factors.

However, all of the remaining variables' coefficients are significant, indicating that these variables are strongly related to the response variable. The interpretation of the coefficients stays unchanged: the estimates represent the change in the response variable for a unit change in the related predictor while holding other predictors fixed.

In summary, the model with no occupation variables outperforms the model with all variables. The remaining variables in the model, however, remain significant predictors of the response variable.

**MODEL 5 :- WITHOUT MARRIAGE:**

The goal of this model is to determine how much marriage variables affect income in comparison to the other independent variables.

Marriage may have an effect on income, and if it is accounted for in the model, it may be difficult to determine if any apparent association between income and English proficiency is due to marriage or English proficiency. As a result, we can better isolate the influence of English proficiency on income by running a model without marriage.

Furthermore, this model can assist us in determining whether the relationship between English proficiency and income is consistent across marital statuses. For example, if the relationship is stronger for unmarried people, it may imply that there are specific career paths or job opportunities that demand superior English proficiency and are more accessible to unmarried people.

**R code implementation** :

model\_without\_mar <- lm(transformed\_income ~ AGEP + SEX + RACWHT + RACAIAN + RACSOR + RACASN + RACBLK + RACNH + POWSP + Bachelors + Managers + Business + Financial + IT + Engineering + Science + Coucelling + Legal + Education + Entertainment + Medical + HealthCare + PRT + EATERY + CLN + PRS + Sales + Official + FFF + CON + EXT + RPR + Production + TRN + MIL+ENG, data = dataframe,na.action=na.exclude)

summary(model\_without\_mar)

##   
## Call:  
## lm(formula = transformed\_income ~ AGEP + SEX + RACWHT + RACAIAN +   
## RACSOR + RACASN + RACBLK + RACNH + POWSP + Bachelors + Managers +   
## Business + Financial + IT + Engineering + Science + Coucelling +   
## Legal + Education + Entertainment + Medical + HealthCare +   
## PRT + EATERY + CLN + PRS + Sales + Official + FFF + CON +   
## EXT + RPR + Production + TRN + MIL + ENG, data = dataframe,   
## na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.3116 -0.3632 0.1117 0.5128 4.0811   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.010e+01 6.220e-03 1624.561 < 2e-16 \*\*\*  
## AGEP 1.966e-02 5.300e-05 370.838 < 2e-16 \*\*\*  
## SEX -3.268e-01 1.724e-03 -189.559 < 2e-16 \*\*\*  
## RACWHT 1.047e-02 3.324e-03 3.149 0.001636 \*\*   
## RACAIAN -5.363e-02 4.771e-03 -11.241 < 2e-16 \*\*\*  
## RACSOR -1.034e-02 2.984e-03 -3.466 0.000529 \*\*\*  
## RACASN 3.235e-02 4.102e-03 7.885 3.16e-15 \*\*\*  
## RACBLK -1.277e-01 3.920e-03 -32.575 < 2e-16 \*\*\*  
## RACNH 2.012e-03 1.717e-02 0.117 0.906689   
## POWSP -1.009e-03 4.835e-05 -20.859 < 2e-16 \*\*\*  
## Bachelors 5.099e-01 1.831e-03 278.433 < 2e-16 \*\*\*  
## Managers 4.403e-01 4.833e-03 91.117 < 2e-16 \*\*\*  
## Business 3.307e-01 5.790e-03 57.115 < 2e-16 \*\*\*  
## Financial 3.758e-01 6.519e-03 57.650 < 2e-16 \*\*\*  
## IT 4.748e-01 5.761e-03 82.423 < 2e-16 \*\*\*  
## Engineering 3.951e-01 6.425e-03 61.499 < 2e-16 \*\*\*  
## Science 2.329e-01 7.874e-03 29.574 < 2e-16 \*\*\*  
## Coucelling -1.584e-01 7.008e-03 -22.602 < 2e-16 \*\*\*  
## Legal 5.395e-01 7.734e-03 69.758 < 2e-16 \*\*\*  
## Education -2.170e-01 5.346e-03 -40.588 < 2e-16 \*\*\*  
## Entertainment -2.015e-01 6.718e-03 -29.994 < 2e-16 \*\*\*  
## Medical 3.842e-01 5.310e-03 72.357 < 2e-16 \*\*\*  
## HealthCare -3.008e-01 6.266e-03 -48.008 < 2e-16 \*\*\*  
## PRT 1.580e-01 6.723e-03 23.501 < 2e-16 \*\*\*  
## EATERY -6.954e-01 5.859e-03 -118.690 < 2e-16 \*\*\*  
## CLN -5.183e-01 6.391e-03 -81.098 < 2e-16 \*\*\*  
## PRS -6.411e-01 6.725e-03 -95.327 < 2e-16 \*\*\*  
## Sales -1.014e-01 4.970e-03 -20.393 < 2e-16 \*\*\*  
## Official -1.104e-01 4.917e-03 -22.460 < 2e-16 \*\*\*  
## FFF -3.413e-01 1.229e-02 -27.765 < 2e-16 \*\*\*  
## CON 5.836e-02 1.720e-02 3.393 0.000692 \*\*\*  
## EXT 3.185e-01 2.519e-02 12.645 < 2e-16 \*\*\*  
## RPR 1.104e-01 6.052e-03 18.244 < 2e-16 \*\*\*  
## Production -3.607e-02 5.437e-03 -6.635 3.25e-11 \*\*\*  
## TRN -2.438e-01 5.135e-03 -47.485 < 2e-16 \*\*\*  
## MIL 1.161e-01 1.182e-02 9.823 < 2e-16 \*\*\*  
## ENG -6.458e-02 1.466e-03 -44.038 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8787 on 1314159 degrees of freedom  
## Multiple R-squared: 0.3326, Adjusted R-squared: 0.3326   
## F-statistic: 1.819e+04 on 36 and 1314159 DF, p-value: < 2.2e-16

**Result** :

The R-squared value in the output is 0.3326, suggesting that the model explains 33.26% of the variation in income. R-squared is modified to 0.3326.

p-value:  The lower the p-value, the more evidence there is against the null hypothesis that the coefficient is zero (no effect). All predictor variables in this model have p-values less than 0.05, indicating that they are all statistically related to the outcome variable

**MODEL 6 :- MODEL WITHOUT EDUCATION:**

The goal of this model is to determine how much education variables affect income in comparison to the other independent variables. The model without education is significant in this project since it helps to determine how much education influences the association between English competence and income. If the model without education reveals a significant weakening of the association between English proficiency and income, then demonstrates that education is a confounding variable in this relationship.

Education is frequently related to higher income, and it may influence both English proficiency and income. As a result, it is critical to evaluate whether schooling accounts for the observed association between English proficiency and income. If schooling is considered a confounding variable, the association between English proficiency and income may be exaggerated, and the model may be inaccurate. In such a circumstance, controlling for schooling may be important to provide a more accurate assessment of the association between English competence and income. As a result, the model without education is significant because it allows us to analyze the impact of education on the relationship between English proficiency and income and draws more accurate conclusions regarding the importance of English proficiency for income.

**R Code implementation** :

model\_without\_edu <- lm(transformed\_income ~ AGEP + SEX + RACWHT + RACAIAN + RACSOR + RACASN + RACBLK + RACNH + POWSP + unmarried + Managers + Business + Financial + IT + Engineering + Science + Coucelling + Legal + Education + Entertainment + Medical + HealthCare + PRT + EATERY + CLN + PRS + Sales + Official + FFF + CON + EXT + RPR + Production + TRN + MIL+ENG, data = dataframe,na.action=na.exclude)

summary(model\_without\_edu)

##   
## Call:  
## lm(formula = transformed\_income ~ AGEP + SEX + RACWHT + RACAIAN +   
## RACSOR + RACASN + RACBLK + RACNH + POWSP + unmarried + Managers +   
## Business + Financial + IT + Engineering + Science + Coucelling +   
## Legal + Education + Entertainment + Medical + HealthCare +   
## PRT + EATERY + CLN + PRS + Sales + Official + FFF + CON +   
## EXT + RPR + Production + TRN + MIL + ENG, data = dataframe,   
## na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.2679 -0.3831 0.1062 0.5234 4.0821   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.038e+01 6.504e-03 1595.952 < 2e-16 \*\*\*  
## AGEP 1.717e-02 5.668e-05 302.930 < 2e-16 \*\*\*  
## SEX -3.262e-01 1.760e-03 -185.382 < 2e-16 \*\*\*  
## RACWHT 1.889e-02 3.390e-03 5.572 2.52e-08 \*\*\*  
## RACAIAN -7.923e-02 4.864e-03 -16.291 < 2e-16 \*\*\*  
## RACSOR -2.298e-02 3.042e-03 -7.554 4.24e-14 \*\*\*  
## RACASN 1.063e-01 4.173e-03 25.483 < 2e-16 \*\*\*  
## RACBLK -1.026e-01 4.005e-03 -25.625 < 2e-16 \*\*\*  
## RACNH -6.511e-02 1.750e-02 -3.720 0.000199 \*\*\*  
## POWSP -1.360e-03 4.932e-05 -27.564 < 2e-16 \*\*\*  
## unmarried -2.655e-01 1.703e-03 -155.860 < 2e-16 \*\*\*  
## Managers 6.712e-01 4.838e-03 138.721 < 2e-16 \*\*\*  
## Business 6.061e-01 5.807e-03 104.371 < 2e-16 \*\*\*  
## Financial 7.097e-01 6.521e-03 108.838 < 2e-16 \*\*\*  
## IT 7.667e-01 5.766e-03 132.982 < 2e-16 \*\*\*  
## Engineering 6.852e-01 6.453e-03 106.179 < 2e-16 \*\*\*  
## Science 6.015e-01 7.908e-03 76.066 < 2e-16 \*\*\*  
## Coucelling 1.917e-01 7.022e-03 27.294 < 2e-16 \*\*\*  
## Legal 8.944e-01 7.766e-03 115.166 < 2e-16 \*\*\*  
## Education 1.203e-01 5.291e-03 22.746 < 2e-16 \*\*\*  
## Entertainment 8.505e-02 6.769e-03 12.565 < 2e-16 \*\*\*  
## Medical 6.452e-01 5.314e-03 121.418 < 2e-16 \*\*\*  
## HealthCare -2.292e-01 6.388e-03 -35.877 < 2e-16 \*\*\*  
## PRT 2.602e-01 6.843e-03 38.025 < 2e-16 \*\*\*  
## EATERY -6.325e-01 5.980e-03 -105.768 < 2e-16 \*\*\*  
## CLN -4.812e-01 6.518e-03 -73.831 < 2e-16 \*\*\*  
## PRS -5.637e-01 6.853e-03 -82.243 < 2e-16 \*\*\*  
## Sales 3.459e-02 5.046e-03 6.854 7.17e-12 \*\*\*  
## Official -8.149e-03 5.001e-03 -1.630 0.103193   
## FFF -2.899e-01 1.253e-02 -23.128 < 2e-16 \*\*\*  
## CON 8.206e-02 1.754e-02 4.678 2.89e-06 \*\*\*  
## EXT 2.838e-01 2.569e-02 11.047 < 2e-16 \*\*\*  
## RPR 1.037e-01 6.172e-03 16.804 < 2e-16 \*\*\*  
## Production -1.268e-02 5.544e-03 -2.288 0.022158 \*   
## TRN -2.033e-01 5.237e-03 -38.819 < 2e-16 \*\*\*  
## MIL 2.042e-01 1.205e-02 16.950 < 2e-16 \*\*\*  
## ENG -7.468e-02 1.497e-03 -49.891 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.896 on 1314159 degrees of freedom  
## Multiple R-squared: 0.3061, Adjusted R-squared: 0.3061   
## F-statistic: 1.61e+04 on 36 and 1314159 DF, p-value: < 2.2e-16

**Result** :

The R-squared value of the model without education factors is 0.3061, indicating that the model explains 30.61% of the variation in the response variable. 0.3061 is the adjusted R-squared value.

The F-statistic has a p-value of 2.2e-16, which indicates that the model is significant and that at least one of the variables is relevant in predicting income. Because their p-values are less than 0.05, the coefficient estimates for all variables are statistically significant at a 5% significance level.

When compared to the model with all variables, the R-squared and adjusted R-squared values in this model are lower because it excludes education variables, which are known to be significant in predicting income. Furthermore, the F-statistic and p-values in this model are lower, indicating that it is less significant than the model with all variables. Overall, this model is a simplified version of the model with all variables, although it predicts income less accurately.

**MODEL 7 :-MODEL WITH ONLY ENGLISH :**

The model with English proficiency as the sole independent variable is significant because it allows us to isolate the influence of English competence on income. The project's hypothesis is that people with higher English proficiency have higher median incomes than people with lower English proficiency. We can see the influence of English proficiency on income without the potential confounding effects of other variables such as occupation, education, race, and marriage by designing a model with only English proficiency as the independent variable. Furthermore, English proficiency is frequently regarded as a critical factor in many aspects of life, such as education, job opportunities, and social mobility. As a result, even after adjusting for other characteristics, it is realistic to expect English competence to have a considerable effect on income. We can obtain useful insights into the relationship between English proficiency and income by including this variable in our study and examining its independent effect on income.

**R code implementation**:

model\_only\_eng <- lm(transformed\_income ~ENG , data = dataframe,na.action=na.exclude)  
summary(model\_only\_eng)

##   
## Call:  
## lm(formula = transformed\_income ~ ENG, data = dataframe, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.0252 -0.4717 0.1015 0.6321 3.7325   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.718338 0.001013 10576.07 <2e-16 \*\*\*  
## ENG -0.119968 0.001479 -81.12 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.073 on 1314194 degrees of freedom  
## Multiple R-squared: 0.004982, Adjusted R-squared: 0.004981   
## F-statistic: 6580 on 1 and 1314194 DF, p-value: < 2.2e-16

**Result** :

This model predicts transformed\_income using only the English language variable. The model contains an intercept of 10.718 and a coefficient of -0.11 for the ENG variable, which means that for every unit rise in ENG, transformed\_income drops by 0.11 units while all other variables are held constant.

The multiple R-squared value is 0.0049, indicating that the English language variable alone can explain just 0.49% of the variation in income. The adjusted R-squared value is equal to the multiple R-squared value, indicating that adding more variables to the model does not improve its goodness of fit.

The F-statistic is 6580, and the p-value is less than 2.2e-16, suggesting that the model is statistically significant, and that the ENG variable has a substantial effect on income.

The confidence interval for the ENG variable is between -0.1228672 and -0.1170697, indicating that we are 95% confident that the true coefficient for the ENG variable is within this range.

This model has a lower R-squared value than the entire model, indicating that the other factors in the full model explain some of the variation in income that the English language variable alone cannot account for. This model, however, is statistically significant and can provide light on the relationship between language and income.

**MODEL 8 :- WITH ONLY FEW OCCUPATIONS CONSIDERED:**

In terms of the English proficiency variable, the model with only a few occupations is crucial since it can help us better understand the relationship between English proficiency and income for specific occupations. This also simplifies our analysis by reducing the number of variables.

Different occupations may have varied English proficiency requirements, and so the influence of English proficiency on income may vary depending on the occupation. A customer service representative or a language translator, for example, may have a larger association between English competence and income than another occupation where communication skills are less crucial.

We can isolate the influence of English proficiency on income for those specific occupations by running a model with only a few occupations, thereby gaining insights into the importance of English proficiency in those industries. This data might be valuable for politicians, educators, and job seekers who want to focus on developing English proficiency in specific occupations in order to boost their earning potential.

**R Code implementation** :

model\_only\_few\_occu <- lm(transformed\_income ~ ENG+AGEP + SEX + RACWHT + RACAIAN + RACSOR + RACASN + RACBLK + RACNH + POWSP + unmarried + Bachelors + Managers + Business + Financial + IT + Engineering + Science + Coucelling + Legal + Education + Entertainment + Medical + HealthCare + Sales , data = dataframe,na.action=na.exclude)  
  
summary(model\_only\_few\_occu)

##   
## Call:  
## lm(formula = transformed\_income ~ ENG + AGEP + SEX + RACWHT +   
## RACAIAN + RACSOR + RACASN + RACBLK + RACNH + POWSP + unmarried +   
## Bachelors + Managers + Business + Financial + IT + Engineering +   
## Science + Coucelling + Legal + Education + Entertainment +   
## Medical + HealthCare + Sales, data = dataframe, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.3696 -0.3702 0.1149 0.5245 3.8655   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.019e+01 5.236e-03 1945.188 < 2e-16 \*\*\*  
## ENG -8.624e-02 1.474e-03 -58.502 < 2e-16 \*\*\*  
## AGEP 1.777e-02 5.554e-05 319.914 < 2e-16 \*\*\*  
## SEX -3.601e-01 1.618e-03 -222.556 < 2e-16 \*\*\*  
## RACWHT 5.415e-03 3.349e-03 1.617 0.10594   
## RACAIAN -4.302e-02 4.807e-03 -8.949 < 2e-16 \*\*\*  
## RACSOR 3.104e-03 3.007e-03 1.032 0.30186   
## RACASN 2.027e-02 4.129e-03 4.908 9.22e-07 \*\*\*  
## RACBLK -9.561e-02 3.954e-03 -24.182 < 2e-16 \*\*\*  
## RACNH 1.195e-02 1.730e-02 0.691 0.48944   
## POWSP -1.190e-03 4.872e-05 -24.415 < 2e-16 \*\*\*  
## unmarried -2.551e-01 1.683e-03 -151.598 < 2e-16 \*\*\*  
## Bachelors 4.994e-01 1.837e-03 271.910 < 2e-16 \*\*\*  
## Managers 5.913e-01 2.633e-03 224.589 < 2e-16 \*\*\*  
## Business 5.024e-01 4.106e-03 122.353 < 2e-16 \*\*\*  
## Financial 5.447e-01 5.091e-03 107.004 < 2e-16 \*\*\*  
## IT 6.376e-01 4.138e-03 154.095 < 2e-16 \*\*\*  
## Engineering 5.467e-01 5.060e-03 108.043 < 2e-16 \*\*\*  
## Science 4.109e-01 6.768e-03 60.713 < 2e-16 \*\*\*  
## Coucelling 2.290e-02 5.685e-03 4.028 5.63e-05 \*\*\*  
## Legal 7.043e-01 6.596e-03 106.776 < 2e-16 \*\*\*  
## Education -4.256e-02 3.368e-03 -12.637 < 2e-16 \*\*\*  
## Entertainment -1.619e-02 5.363e-03 -3.020 0.00253 \*\*   
## Medical 5.560e-01 3.312e-03 167.889 < 2e-16 \*\*\*  
## HealthCare -8.070e-02 4.721e-03 -17.094 < 2e-16 \*\*\*  
## Sales 8.645e-02 2.887e-03 29.947 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8853 on 1314170 degrees of freedom  
## Multiple R-squared: 0.3226, Adjusted R-squared: 0.3226   
## F-statistic: 2.503e+04 on 25 and 1314170 DF, p-value: < 2.2e-16

**Result :**

The corrected R-squared of the model is 0.3226, indicating that the independent variables explain 32.26% of the variance in transformed income. Because it takes into account the number of independent variables in the model, adjusted R-squared is preferred over R-squared.

The F-statistic is 2.503e+04, and the p-value is less than 2.2e-16, indicating that the model as a whole is statistically significant, which means that at least one of the independent variables has a substantial effect on the transformed income.

We would anticipate the adjusted R-squared to be greater when comparing these results to the entire model with all variables, indicating that the whole model is a stronger predictor of income. Furthermore, the p-value for all variables in the full model should be lower than that of the model with only a few occupation variables, indicating that the full model contains more statistically significant predictors of income.

**[MODEL 9](#MODEL9) :- WITH ONLY EAST COAST STATE SUBSET CONSIDERED :**

This is our sample dataset, which consists of 14 states out of the 50 states of the United States. The East Coast states may be significant to include in a model since they may have different economic or demographic characteristics than other regions. East Coast states such as New York, Massachusetts, and Maryland, for example, are recognized for having strong concentrations of technology and financial services companies, which may have an impact on median income levels in these locations. Furthermore, East Coast states may have higher living costs, which may have an impact on income levels.

By incorporating East Coast states into a distinct model, we will be able to better understand the particular characteristics that contribute to income levels in these regions. This could be useful for designing targeted policies or initiatives in these locations to address income inequality or other relevant issues. It may also aid in identifying any regional income gaps that could be rectified through policy changes or other measures.

**R Code implementation** :

state\_subset\_df<- dataframe[dataframe$POWSP %in% c("23", "33", "25", "44", "9","36","34","10","24","51","37","45", "13","12"), ]  
  
state\_subset\_df <- na.omit(state\_subset\_df)  
  
colnames(state\_subset\_df)

## [1] "PWGTP" "PINCP" "AGEP"   
## [4] "RACWHT" "RACBLK" "MAR"   
## [7] "SEX" "SCHL" "POWSP"   
## [10] "ENG" "RACPI" "RACNH"   
## [13] "RACAIAN" "RACSOR" "RACASN"   
## [16] "SOCP" "married" "unmarried"   
## [19] "English" "Gender" "NoBachelors"   
## [22] "Bachelors" "Managers" "Business"   
## [25] "Financial" "IT" "Engineering"   
## [28] "Science" "Coucelling" "Legal"   
## [31] "Education" "Entertainment" "Medical"   
## [34] "HealthCare" "PRT" "EATERY"   
## [37] "CLN" "PRS" "Sales"   
## [40] "Official" "FFF" "CON"   
## [43] "EXT" "RPR" "Production"   
## [46] "TRN" "MIL" "transformed\_income"

nrow(state\_subset\_df)

## [1] 433358

model\_state\_sample <- lm(transformed\_income ~ ENG+AGEP + SEX + RACWHT + RACAIAN + RACSOR + RACASN + RACBLK + RACNH + POWSP+unmarried+Bachelors+ Managers + Business + Financial + IT + Engineering + Science + Coucelling + Legal + Education + Entertainment + Medical + HealthCare + PRT + EATERY + CLN + PRS + Sales + Official + FFF + CON + EXT + RPR + Production + TRN + MIL, data = state\_subset\_df)

summary(model\_state\_sample)

##   
## Call:  
## lm(formula = transformed\_income ~ ENG + AGEP + SEX + RACWHT +   
## RACAIAN + RACSOR + RACASN + RACBLK + RACNH + POWSP + unmarried +   
## Bachelors + Managers + Business + Financial + IT + Engineering +   
## Science + Coucelling + Legal + Education + Entertainment +   
## Medical + HealthCare + PRT + EATERY + CLN + PRS + Sales +   
## Official + FFF + CON + EXT + RPR + Production + TRN + MIL,   
## data = state\_subset\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.4292 -0.3641 0.1093 0.5124 4.0456   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.015e+01 1.176e-02 863.449 < 2e-16 \*\*\*  
## ENG -7.957e-02 2.432e-03 -32.715 < 2e-16 \*\*\*  
## AGEP 1.790e-02 9.704e-05 184.459 < 2e-16 \*\*\*  
## SEX -3.081e-01 2.990e-03 -103.051 < 2e-16 \*\*\*  
## RACWHT 2.469e-02 6.114e-03 4.038 5.39e-05 \*\*\*  
## RACAIAN -6.640e-02 9.810e-03 -6.768 1.31e-11 \*\*\*  
## RACSOR -1.274e-02 5.338e-03 -2.387 0.01698 \*   
## RACASN -4.466e-04 7.502e-03 -0.060 0.95253   
## RACBLK -7.967e-02 6.697e-03 -11.896 < 2e-16 \*\*\*  
## RACNH -1.890e-01 5.803e-02 -3.257 0.00113 \*\*   
## POWSP 9.717e-04 1.035e-04 9.390 < 2e-16 \*\*\*  
## unmarried -2.382e-01 2.931e-03 -81.282 < 2e-16 \*\*\*  
## Bachelors 5.067e-01 3.176e-03 159.521 < 2e-16 \*\*\*  
## Managers 5.044e-01 8.694e-03 58.013 < 2e-16 \*\*\*  
## Business 3.960e-01 1.016e-02 38.980 < 2e-16 \*\*\*  
## Financial 4.755e-01 1.131e-02 42.026 < 2e-16 \*\*\*  
## IT 5.135e-01 1.007e-02 50.975 < 2e-16 \*\*\*  
## Engineering 3.977e-01 1.164e-02 34.164 < 2e-16 \*\*\*  
## Science 3.163e-01 1.349e-02 23.445 < 2e-16 \*\*\*  
## Coucelling -8.372e-02 1.242e-02 -6.742 1.56e-11 \*\*\*  
## Legal 5.976e-01 1.283e-02 46.576 < 2e-16 \*\*\*  
## Education -1.468e-01 9.491e-03 -15.462 < 2e-16 \*\*\*  
## Entertainment -1.167e-01 1.169e-02 -9.983 < 2e-16 \*\*\*  
## Medical 4.301e-01 9.489e-03 45.327 < 2e-16 \*\*\*  
## HealthCare -1.931e-01 1.134e-02 -17.025 < 2e-16 \*\*\*  
## PRT 2.122e-01 1.167e-02 18.181 < 2e-16 \*\*\*  
## EATERY -6.032e-01 1.054e-02 -57.234 < 2e-16 \*\*\*  
## CLN -4.429e-01 1.135e-02 -39.031 < 2e-16 \*\*\*  
## PRS -5.783e-01 1.188e-02 -48.698 < 2e-16 \*\*\*  
## Sales -4.578e-02 8.925e-03 -5.129 2.92e-07 \*\*\*  
## Official -5.643e-02 8.871e-03 -6.362 2.00e-10 \*\*\*  
## FFF -2.916e-01 2.726e-02 -10.699 < 2e-16 \*\*\*  
## CON 9.866e-02 3.139e-02 3.143 0.00167 \*\*   
## EXT 1.459e-01 6.576e-02 2.219 0.02648 \*   
## RPR 1.457e-01 1.106e-02 13.182 < 2e-16 \*\*\*  
## Production -2.592e-02 1.031e-02 -2.515 0.01191 \*   
## TRN -2.194e-01 9.426e-03 -23.280 < 2e-16 \*\*\*  
## MIL 1.846e-01 1.923e-02 9.600 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8812 on 433320 degrees of freedom  
## Multiple R-squared: 0.3544, Adjusted R-squared: 0.3544   
## F-statistic: 6429 on 37 and 433320 DF, p-value: < 2.2e-16

**Result** :

The model with all variables except the East Coast state subset has an R-squared of 0.3544 and an adjusted R-squared of 0.3544, which suggests that the independent variables in the model explain 35.44% of the variance in transformed\_income. The p-value for the F-statistic is relatively low (2.2e-16), indicating that at least one of the independent variables is significantly associated to income.

Overall, the model without the East Coast state subset has the slightly higher R-squared and adjusted R-squared as the model with all variables, indicating that the removed data have very little meaningful effect on the model's overall fit. It’s probably due to the fact that East coast states have higher income levels and the higher values overall for all other variables.

The values for the Results drawn for each model can be summarized in the following table as below :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL Number | MODEL | R-SQUARE | P-VALUE | F-STATISTIC |
| Model 1 | Model with all variables | 0.3426 | < 2.2e-16 | 1.851e+04 |
| Model 2 | Model without English variable | 0.3413 | < 2.2e-16 | 1.891e+04 |
| Model 3 | Model without race | 0.3418 | < 2.2e-16 | 2.202e+04 |
| Model 4 | Model without occupation | 0.2667 | < 2.2e-16 | 3.676e+04 |
| Model 5 | Model without marriage | 0.3333 | < 2.2e-16 | 1.776e+04 |
| Model 6 | Model without education | 0.3069 | < 2.2e-16 | 1.573e+04 |
| Model 7 | Model with only English | 0.004981 | <2.2e-16 | 6580 |
| Model 8 | Model with only few occupations | 0.3226 | < 2.2e-16 | 2.503e+04 |
| Model 9 | MODEL WITH SAMPLE EAST COAST STATES | 0.3544 | < 2.2e-16 | 6429 |

**CONCLUSION:**

In conclusion, various regression models were created to investigate the relationship between the independent variables and the response variable, income. The performance of the models was evaluated using the R-squared value, adjusted R-squared value, and F-statistic. When the models were examined, it was discovered that some had higher R-squared values, suggesting superior performance in explaining variation in the response variable. The coefficients of the variables in the models were also examined, and it was discovered that certain factors had substantial influence on the response variable while others did not.

In the context of the given data, the English variable of interest relates to the impact of English language ability on income. The results indicate that the model with no English variables performs similarly to the model with all variables, implying that the English language competence variables may not have a major influence on income when compared to the other variables in the model.

In the model with all variables, the coefficient for ENG is negative (-0.07404), showing that as English proficiency grows, income drops while other variables remain constant. However, in the model without English variables, this coefficient is not statistically significant, indicating that the impact of English proficiency on income may be negligible or less important than other factors in predicting income. As a result of the findings, it indicates that English proficiency may not have a meaningful effect on income beyond the effects of the model's other predictor factors.

**Bibliography:**

Cadena, B. C., Duncan, B., & Trejo, S. J. (2015). Chapter 22 - The Labor Market Integration and Impacts of US Immigrants. In B. R. Chiswick & P. W. Miller (Eds.), Handbook of the Economics of International Migration 1197–1259. [https://doi:10.1016/B978-0-444-53768-3.00022-9](about:blank)

Chiswick B. R., & Miller P. W. (1990). Language in the Labor Market: The Immigrant Experience in Canada and the United States, Working Paper 784, Economics Department, Queen's University.

Gill, F., & Ahmad, W. (2019). The Earnings Disadvantage of 21st Century Immigrants in the United States. The American Economist, 64(1), 31–44.<https://www.jstor.org/stable/26725792>

U.S. Census Bureau. (2021, October 8). People that speak English less than "very well" in the United States. Census.gov.<https://www.census.gov/library/visualizations/interactive/people-that-speak-english-less-than-very-well.html>