# Data Wizards (Group 4) Project 1

Di Chen Mai Castellano Tyler Kussee Spencer Hutchison

## Introduction

The American Community Survey offers crucial insights about our country and its citizens each year. The data from this survey plays a pivotal role in guiding the allocation of trillions of dollars in federal funding, ensuring that resources are directed where they are needed most.

With the data from the American Community Survey, the Data Wizards have come up with the following 3 questions they would like to answer:

- Does higher education mean higher income accounting for sex and age?
- Can we predict household income based on education, occupation, and race/ethnicity?
- Is there a relationship between race/ethnicity and home ownership?

In particular, we'll be focusing on the first question to get started.

### Obtain the Data

In the case of the ACS dataset, we'll pull it via an API. We've already declared our key earlier, so now we just need to pull our dataset with the variables and filtering we'd like. In this case we'll want to pull our Age (AGEP), Education Level (SCHL), sex (SEX), and Total Person's Income (PINCP) while filtering to the year 2022 and the State of OR.

### Scrub the data

In the data cleaning process, we began by assigning more descriptive column names to enhance clarity. Columns such as "ST", "Serial Number", and "housing weight" were removed as they were deemed unnecessary for our analysis, while "PersonNumber" was retained for potential future use. Subsequently, we converted instances of "bb" in the Education level column to 0, followed by transforming the Education level into numeric values. Groupings were then created based on education levels, ranging from "No Education" to "Doctoral Degree". The column order was rearranged for better organization. As our focus is on income, individuals under the age of 14 were excluded from the dataset, aligning with Oregon's legal working age. Variable codes were converted to their corresponding descriptions, starting with Education Level. Lastly, we identified the genders present in the dataset to facilitate further analysis. This systematic approach ensures that the data is cleaned and structured appropriately for subsequent analysis.

#### First Question

```r
#Occupation Grouping
puma_data <- puma_data %>%
  mutate(
    OccupationGroup = case_when(
      between(as.numeric(Occupation), 0000, 0009) ~ "Less than 16 years old",
      between(as.numeric(Occupation), 0010, 0440) ~ "Management",
      between(as.numeric(Occupation), 0500, 0960) ~ "Business/Finance",
      between(as.numeric(Occupation), 1000, 1240) ~ "CS/Math/Statistics",
      between(as.numeric(Occupation), 1300, 1560) ~ "Engineering/Architecture",
      between(as.numeric(Occupation), 1600, 1980) ~ "Science/Economics",
      between(as.numeric(Occupation), 2000, 2060) ~ "Social/Therapy/Religious",
      between(as.numeric(Occupation), 2100, 2180) ~ "Legal",
      between(as.numeric(Occupation), 2200, 2555) ~ "Education/Library",
      between(as.numeric(Occupation), 2600, 2920) ~ "Entertainment/Arts/Media/Sports",
      between(as.numeric(Occupation), 3000, 3550) ~ "Medical",
      between(as.numeric(Occupation), 3600, 3655) ~ "Health",
      between(as.numeric(Occupation), 3700, 3960) ~ "Protective/Essential",
      between(as.numeric(Occupation), 4000, 4160) ~ "Food",
      between(as.numeric(Occupation), 4200, 4255) ~ "Sanitation/Groundskeeping",
      between(as.numeric(Occupation), 4300, 4655) ~ "Personal/Lifestyle",
      between(as.numeric(Occupation), 4700, 4965) ~ "Sales",
      between(as.numeric(Occupation), 5000, 5940) ~ "Office/Administrative",
      between(as.numeric(Occupation), 6005, 6130) ~ "Farming/Fishing/Forestry",
      between(as.numeric(Occupation), 6200, 6765) ~ "Construction",
      between(as.numeric(Occupation), 6800, 6950) ~ "BlueCollar",
      between(as.numeric(Occupation), 7000, 7640) ~ "Repairs/Mechanics",
      between(as.numeric(Occupation), 7700, 8990) ~ "Manufacturing",
      between(as.numeric(Occupation), 9005, 9760) ~ "Transportation",
      between(as.numeric(Occupation), 9800, 9830) ~ "Military",
      TRUE ~ "Not Classified"
    )
  )
```

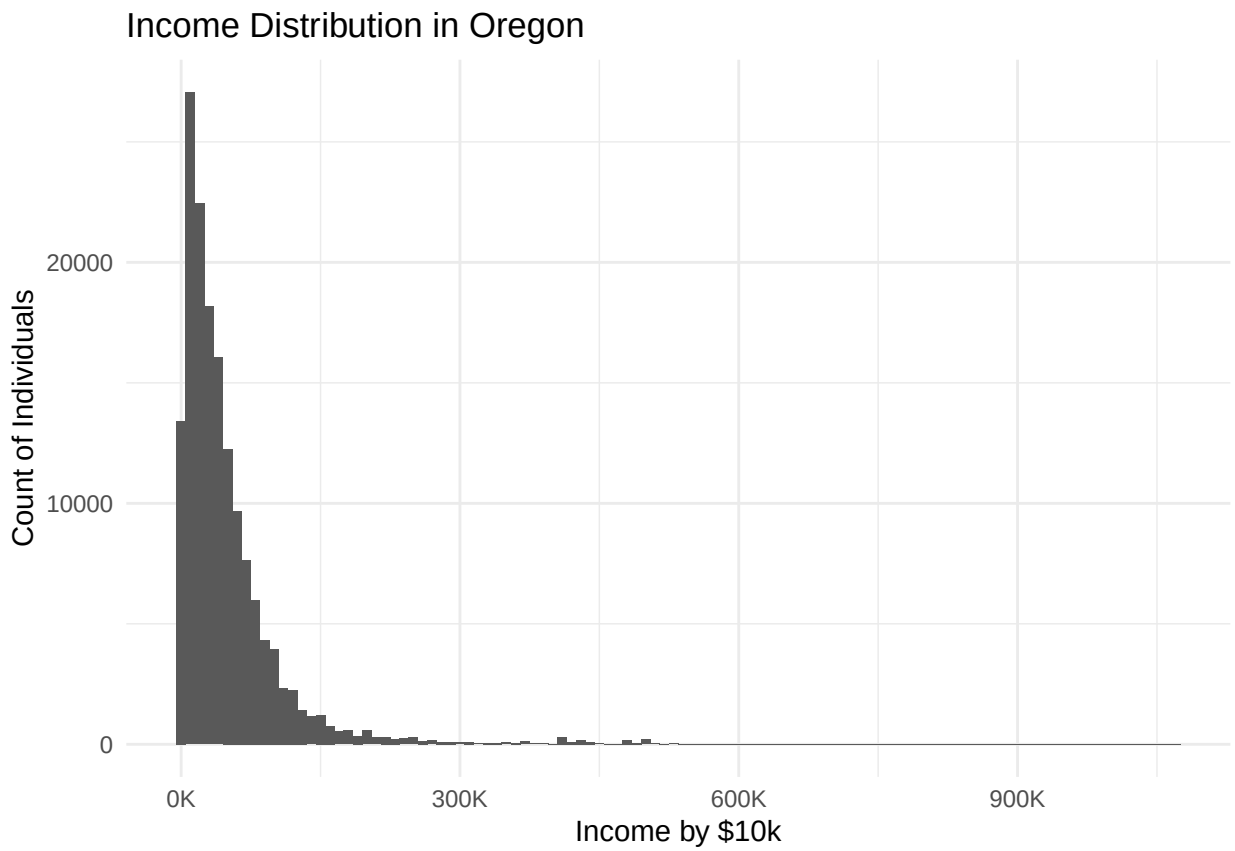
With the data mostly cleaned up, we can now explore the data.

## Explore the data

We want to get a sense for the data that we'll be working with so first we'll find the range of ages that we have in our dataset. This shows that our mean and median age is very close together (43.1 and 43 respectively) showing a good distribution of ages; with a max age of 95. Education level has a similar mean and median (16.31, and 18 respectively), and it gives us a good way to picture our distribution as well. Unlike our age distribution the income mean and median are a bit off (35,364 and 21,000 respectively). This is something we may need to look into more in a graphical view. We also review a table of the levels of education and the their counts.

For Homeownership the the distribution is very interesting with 68% of the sample owning a home either free and clear or though a mortgage. Renters make up a quarter of the sample and the rest may be cleaned up in later sections of the report. As for race 82% of this sample are White alone. The next three largest categories are Two Or More Races, Asian alone, and Some other race alone. This will make for some more difficult predictions.

Next we'll take a look into the income distribution in a graphical view. For this view we will also remove those that are below 0. This allows us to see a very long tailed view of the income distribution for our dataset. Our mean being the tallest point in the graph being around that $21,000 we saw in our exploration.

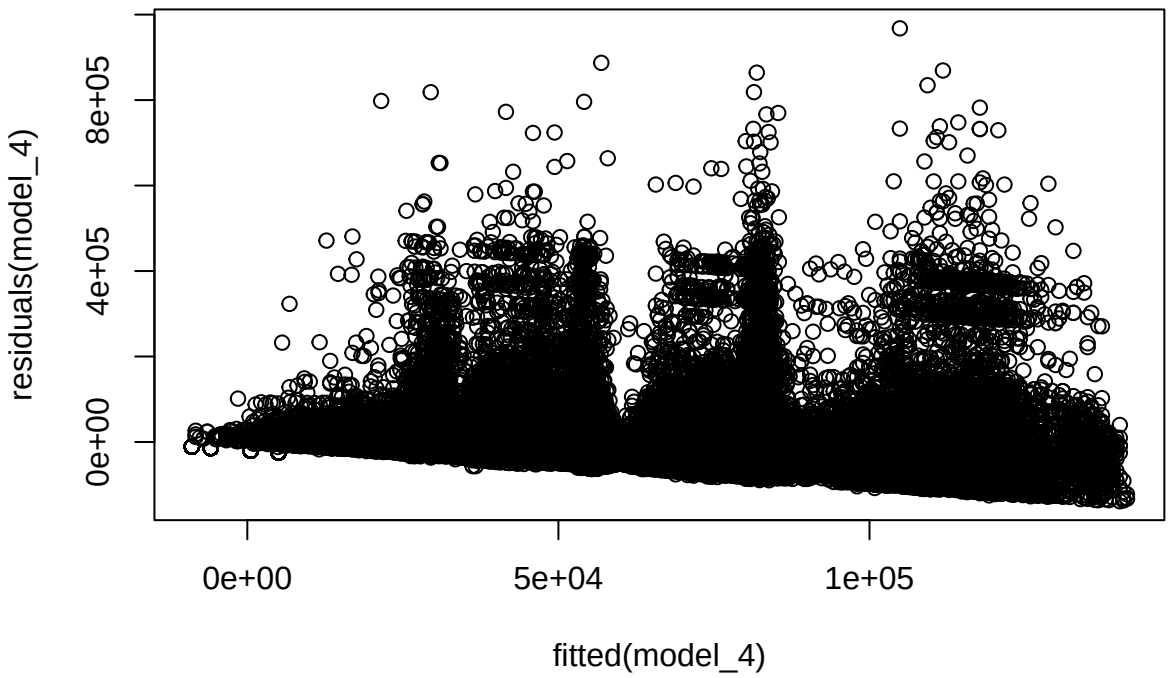


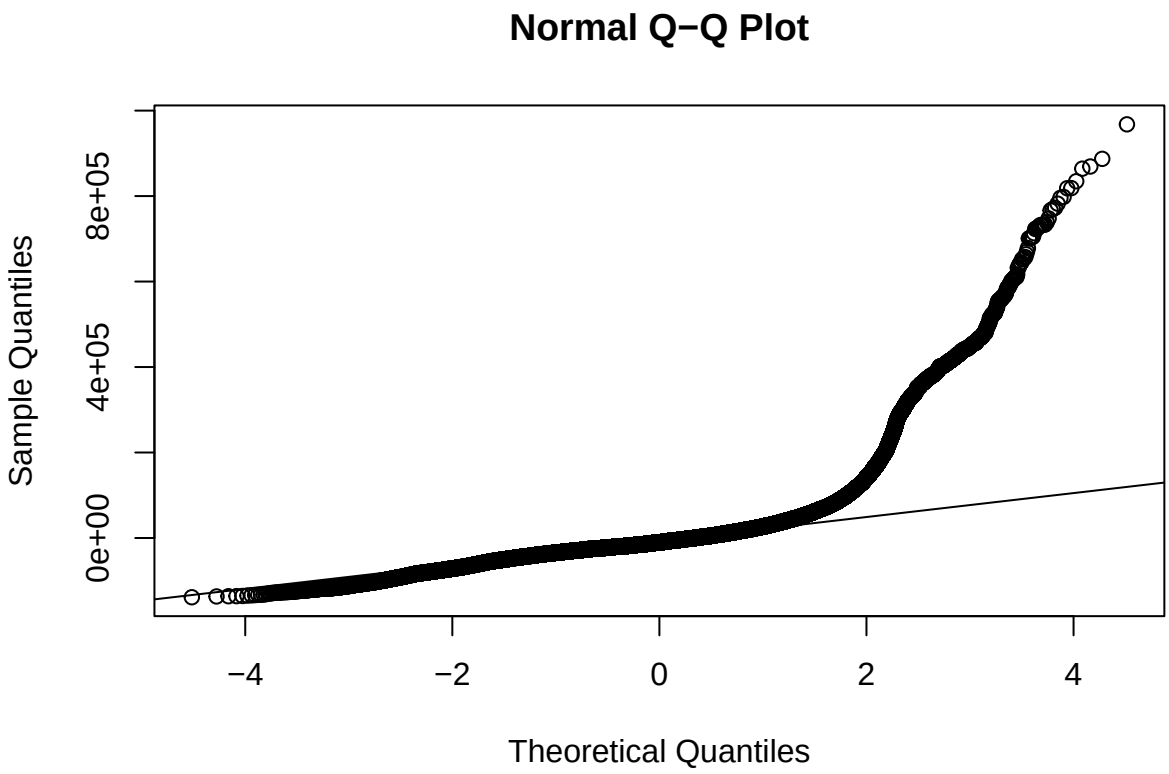

Reference on Variables for now

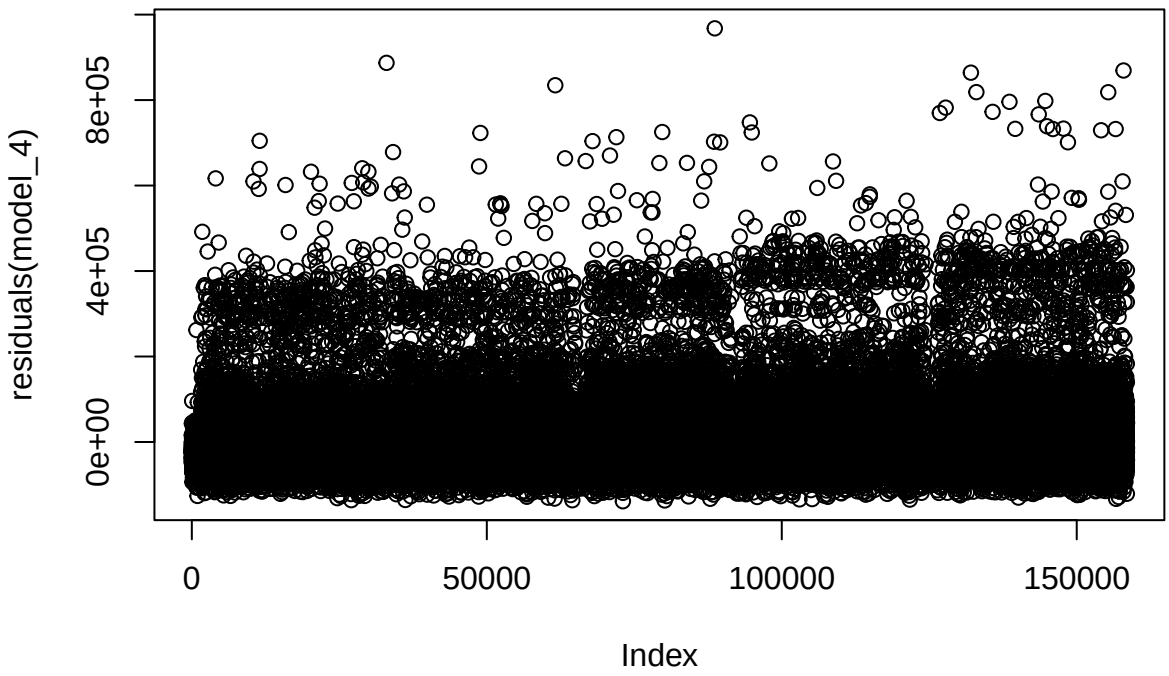
## Model the data

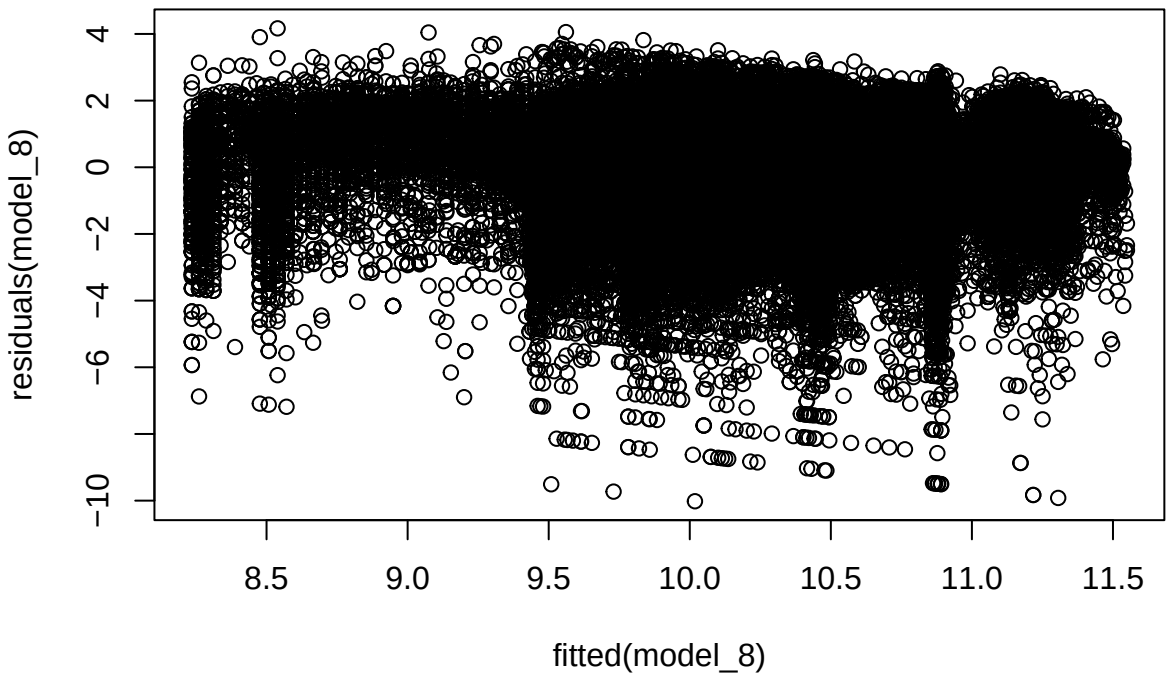
We compared four regression models, each including different combinations of EducationGroup, Sex, and Age as explanatory variables. After conducting ANOVA comparisons between the full and nested models, there was compelling evidence that the last model, which included a three-way interaction term between EducationGroup, Sex, and Age along with two interaction terms between each variable, was a better fit. However, further examination of the model diagnostics revealed a violation of the linearity assumption. To address this issue, we transformed the response variable using logarithm.

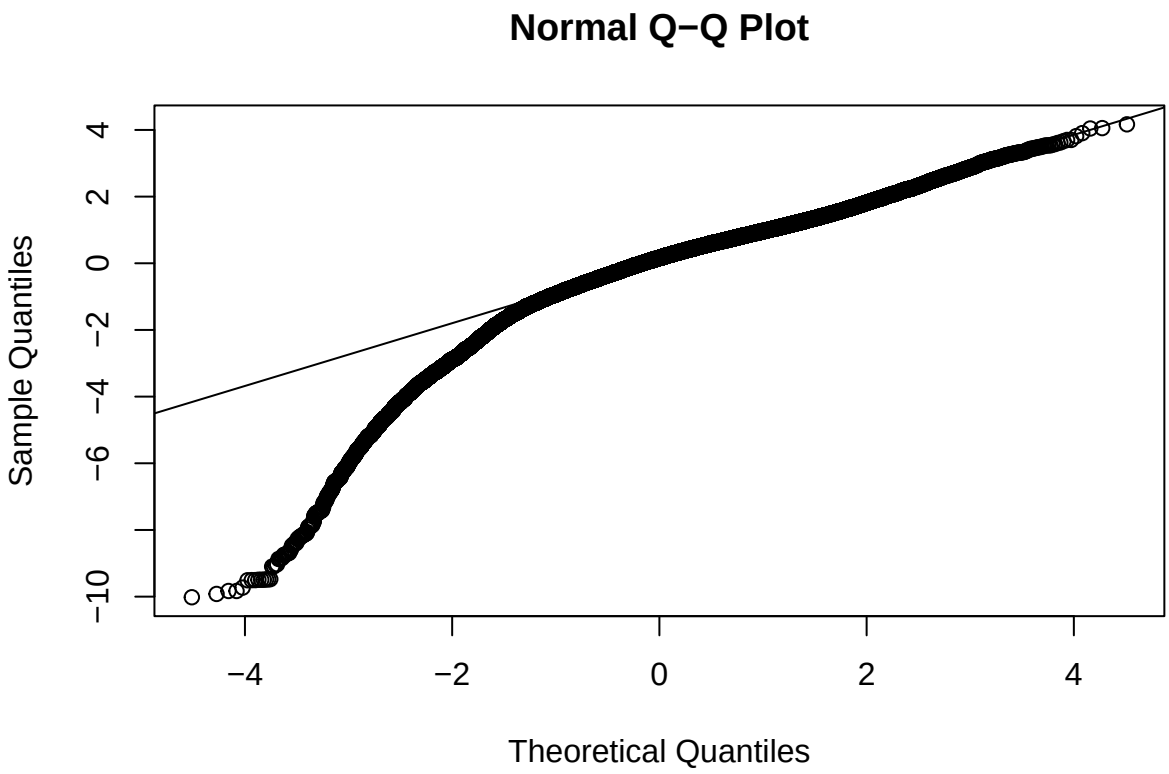
Upon visualizing the model diagnostics, the plot of fitted values versus residuals shows the residuals generally centered around zero, there was a change in clustering pattern around a certain fitted value. Despite this, there was no clear systematic pattern observed in the residuals, suggesting that the assumptions of homoscedasticity and linearity were generally met. In addition, the QQ plot showed some deviation from normality. However, linear regression tends to be robust against such deviations, particularly with large sample sizes. Overall, despite these minor issues, the model still appeared to be a suitable fit for the data, with the assumptions of homoscedasticity, linearity, and normality of residuals reasonably satisfied.


residuals(model_4)
0e+00
4e+05
8e+05
0e+00
5e+04
1e+05
fitted(model_4)


Normal Q−Q Plot
Sample Quantiles
0e+00
4e+05
8e+05
−4
−2
0
2
4
Theoretical Quantiles


residuals(model_4)
0e+00
4e+05
8e+05
0
50000
100000
150000
Index


residuals(model_8)
4
2
0
−2
−4
−6
−10
8.5
9.0
9.5
10.0
10.5
11.0
11.5
fitted(model_8)




### Interpret the data

To answer our question, does higher education mean higher income accounting for sex and age? Use the fitted model to compare the income across all 8 levels of education for a 50 years old males and females in Oregon.

The average income for 50-year-old males in Oregon varies significantly across different educational categories. For individuals with no schooling completed, their average income is $19,927.16, slightly lower than those with less than a high school education at $21,855.60. For those with only a high school education (without a diploma) experience a decrease in average income to $14,441.65. However, individuals with a high school diploma or GED see a notable increase in average income to $26,611.61, followed by even higher incomes for those with higher degrees.

For 50-year-old females in Oregon, a similar pattern was observed. For individuals with no schooling completed, their average income is $13,670.59, slightly increasing for those with less than a high school education to $14,888.97. Similarly observed in male, those with only a high school education (without a diploma) see a decrease in average income to $9,206.92. However, there is a significant increase in average income for females with a high school diploma or GED, reaching $17,306.20, and continuing to rise with higher educational attainment.

There is a clear relationship between educational attainment and average income for both 50-year-old males and females in Oregon. Individuals with higher levels of education tend to have higher average incomes, with notable increases observed at certain educational milestones, such as obtaining a high school diploma or GED and above.

## Appendix

```
## An if statement for checking if a package is installed :)
#if(!require(somepackage)){
#    install.packages("somepackage")
#    library(somepackage)
#}

if (!require(tidycensus)) {
  install.packages("tidycensus")
}
if (!require(tidyverse)) {
  install.packages("tidyverse")
}
if (!require(dplyr)) {
  install.packages("dplyr")
}
if (!require(ggplot2)) {
  install.packages("ggplot2")
}

#Load libraries
library(tidycensus)
library(tidyverse)
library(dplyr)
library(ggplot2)

census_api_key("2547c95ce33b1ed0eec3aafd0fd8526a5bb9a22e")




### Obtain the Data

#Pull data set for specific variables
puma_data <- get_pums(variables = c("AGEP","SCHL","PINCP", "SEX", "RAC1P", "TEN")
                      ,state = "OR"
                      ,year = 2022)

#Rename the columns
puma_data <- puma_data %>% rename(Age = AGEP,
                                  EducationLevel = SCHL,
                                  Income = PINCP,
                                  Sex = SEX,
                                  PersonNumber = SPORDER)
head(puma_data[, 2:5])

#Remove columns that are not needed
puma_data <- puma_data %>%
  select(-ST, -SERIALNO, -WGTP, -PWGTP)
head(puma_data)

#Convert education levels
```

```r
puma_data <- puma_data %>%
  mutate(
    EducationLevel = case_when(
      EducationLevel == "bb" ~ "00",
      TRUE ~ EducationLevel
    ),
    EducationLevel = as.numeric(EducationLevel)
  )

#Make sure it is numeric
class(puma_data$EducationLevel)

##Group education level
puma_data <- puma_data %>%
  mutate(
     EducationGroup = cut(EducationLevel, breaks = c(0, 2, 11, 16, 20, 21, 22, 24, Inf),
                          labels = c("No schooling completed", "Less than High School",
                                     "High School", "high school diploma or GED",
                                     "Associates Degree", "Bachelors Degree",
                                     "Masters Degree", "Doctorate Degree"),
                          right = FALSE)
  )
unique(puma_data$EducationGroup)
puma_data[1:50,]

#Rearrange the columns
puma_data <- puma_data %>%
  select(EducationGroup, EducationLevel, Income, Age, Sex, PersonNumber)
head(puma_data)

#Drop age below 14
puma_data <- puma_data %>%
  filter(Age >= 14)
head(puma_data[puma_data$Age < 14,])

#Identify gender
puma_data <- puma_data %>%
  mutate(
    Sex = case_when(
      Sex == 1 ~ "Male",
      Sex == 2 ~ "Female",
      TRUE ~ as.character(Sex)
    )
)
head(puma_data)




### Explore the data
#Summarizing the ranges of numerical
summary(puma_data$Age) #Range of Age
```

```r
summary(puma_data$Income) #Range of Age
summary(puma_data$EducationNumber) #Range of Education
table(puma_data$EducationLevel) #Count of each level of education

#Plot inceom distribution
ggplot(puma_data %>% filter(Income > 0), aes(Income)) +
  geom_histogram(binwidth = 10000) +
  scale_x_continuous(labels = function(x) paste0(x / 1000, "K")) +
  labs(title = "Income Distribution in Oregon"
       , x = "Income by $10k"
       , y = "Count of Individuals") +
  theme_minimal()
```

```r
##[Reference on Variables for now](https://usa.ipums.org/usa/resources/codebooks/DataDict1822.pdf)
```

```r
### Model the data

#Drop zero income from the data set
#Drop observation with 0 income (unemployed individuals)
puma_data <- puma_data[puma_data$Income != 0, ]
puma_data
head(puma_data)

# Does higher education mean higher income accounting for sex and age?
model_1 <- lm(Income ~ EducationGroup + Sex + Age, data = puma_data)
model_2 <- lm(Income ~ EducationGroup + Sex * Age, data = puma_data)
model_3 <- lm(Income ~ EducationGroup * Sex + Age, data = puma_data)
model_4 <- lm(Income ~ EducationGroup * Sex * Age, data = puma_data)

#Compare model 3 and 4
anova(model_3, model_4)

#The full model provides a significantly better fit.

#Compare model 2 and 4
anova(model_2, model_4)

#Model 4 is a better fit.

#Compare model 1 and 4
anova(model_1, model_4)


##Check the assumption

#Linearity and Homoscedasticity assumption
plot(fitted(model_4), residuals(model_4))
```

```r
#Normality assumption
qqnorm(residuals(model_4))
```

```r
qqline(residuals(model_4))
```

```r
#Independence assumptions
plot(residuals(model_4))
```

```r
#Fitted vs residuals plot exhibiting pattern

#Consider log transformation of the response variable:
#Logged models
model_5 <- lm(log(Income) ~ EducationGroup + Sex + Age, data = puma_data, na.action = na.exclude)
model_6 <- lm(log(Income) ~ EducationGroup + Sex * Age, data = puma_data, na.action = na.exclude)
model_7 <- lm(log(Income) ~ EducationGroup * Sex + Age, data = puma_data, na.action = na.exclude)
model_8 <- lm(log(Income) ~ EducationGroup * Sex * Age, data = puma_data, na.action = na.exclude)

#Compare models 7 and 8
anova(model_7, model_8)

#Compare models 6 and 8
anova(model_6, model_8)

#Compare models 5 and 8
anova(model_5, model_8)

#Model 8 is the best from 4 models


#Check the assumption
#Linear regression assumption
plot(fitted(model_8), residuals(model_8))
```

```r
#Normality assumption
qqnorm(residuals(model_8))
qqline(residuals(model_8))
```

```r
#Use predict() to fit the regression and find the average income for 50 YO males
new_data_1 <- data.frame(EducationGroup = c("No schooling completed","Less than High School", "High Scho
                         Sex = "Male",
                         Age = 50)

mu_1 <- predict(model_8, newdata = new_data_1 , type = "response")
exp(mu_1)

#Use predict() to fit the regression and find the average income for 50 YO femails
new_data_2 <- data.frame(EducationGroup = c("No schooling completed","Less than High School", "High Scho
                         Sex = "Female",
                         Age = 50)

mu_2 <- predict(model_8, newdata = new_data_2 , type = "response")
exp(mu_2)
```