Data Wizards (Group 4) Project 1

Di Chen Mai Castellano Tyler Kussee Spencer (Hutchison) Yang

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?

Obtain/Scrub the data

In the case of the ACS dataset, we'll pull it via an API. In this case we'll want to pull our Age (AGEP), Education Level (SCHL), sex (SEX), Race (RAC1P), Occupation (OCCP), Tenure (TEN), Marital Status (MAR), and Total Person's Income (PINCP) while filtering to the year 2022 and the State of OR.

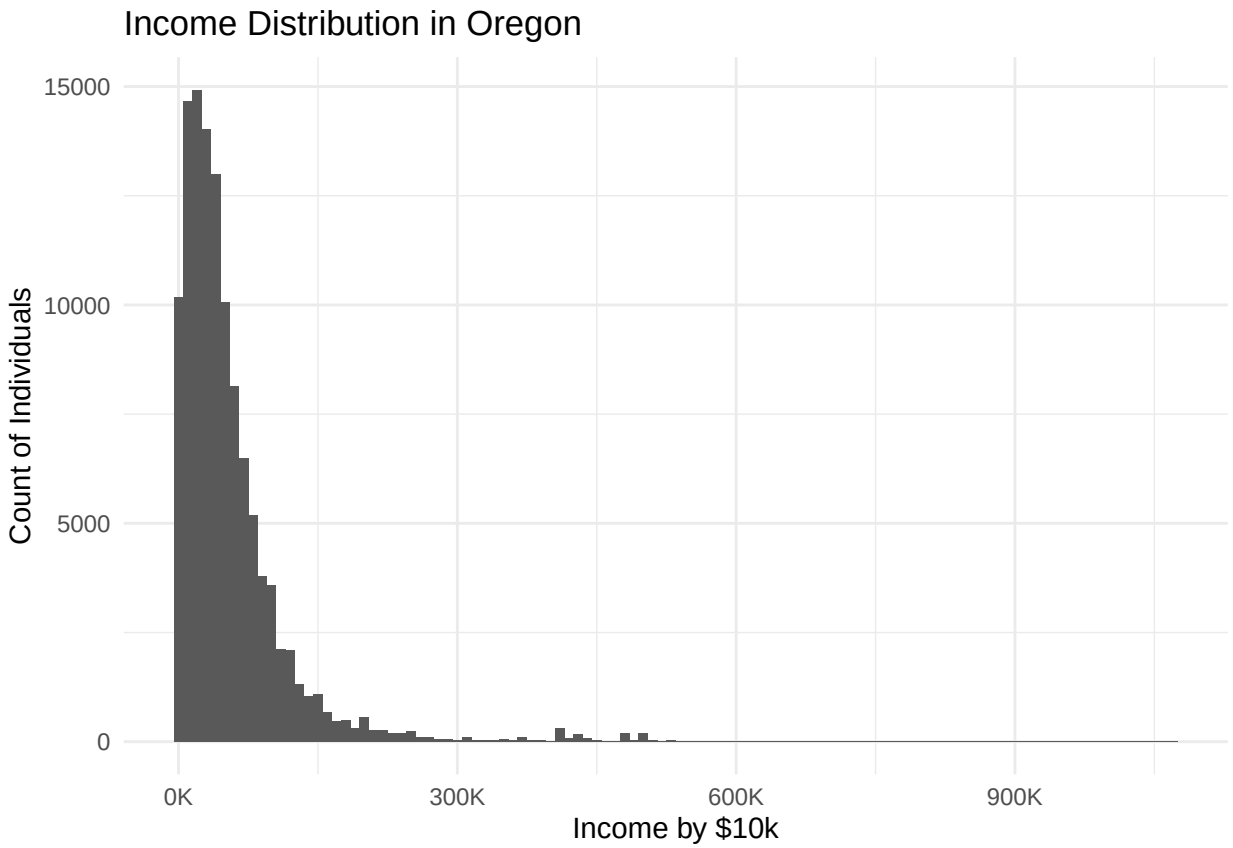
In the data cleaning process, we began by assigning more descriptive column names to enhance clarity. Columns such as "ST", "Serial Number", and "housing weight" were removed as they were deemed unnecessary for our analysis, while "PersonNumber" was retained for potential future use. Subsequently, we converted instances of "bb" in the Education level, "bbbb"/"000N" in the Occupation, and "b" in Homeownership columns to 0, followed by transformation of all 3 columns into numeric values. Groupings were then created based on education levels, Occupation values, Race, Marital Status, and HouseAcreage. Various columns received descriptive categories like Homeownership, then 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. Additional filtering was included on our occupation groupings and genders for data quality assurance. This systematic approach ensures that the data is cleaned and structured appropriately for subsequent analysis.

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 (50.99 and 51 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 (50,432 and 33,700 respectively).

Next, we explored the income distribution using a graphical representation, removing individuals with negative income values to gain a clearer understanding of the data. The resulting graph revealed a highly skewed distribution with a long tail, indicating a concentration of individuals around the mean income of approximately $21,000. We also examined the occupation groups variable in relation to income, but found it challenging to interpret due to the presence of three broad categories - Management, Office/Admin, and Sales - which collectively accounted for a significant portion of the working population. Despite this limitation, by leveraging the other available variables, we can develop a model that effectively incorporates the occupation groups and provides valuable insights into their relationship with income.

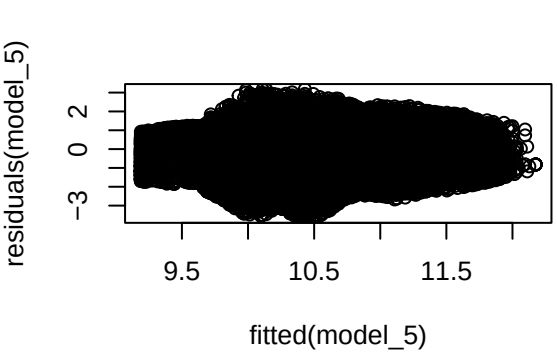


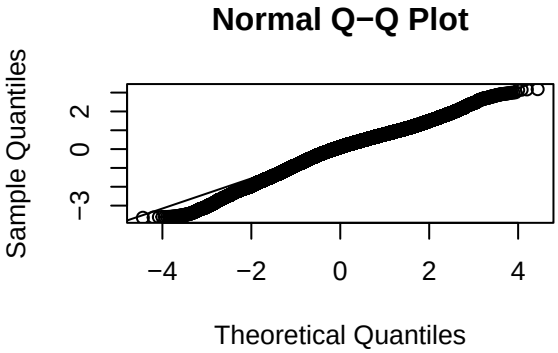
Model the data

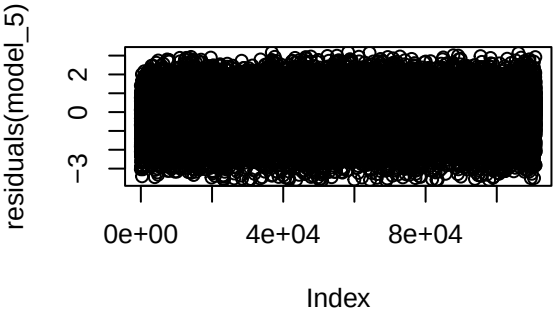
Question 1

We compared two regression models: the first model included EducationGroup, Sex, and Age as explanatory variables, while the second model extended to incorporate an interaction between Sex and Age. ANOVA comparisons between the full and nested models provided compelling evidence that the latter, which accounted for the interaction term, was a better fit (Analysis of variance, $p < 0.00001$). However, the examination of model diagnostics showed a violation of the linearity assumption. We then transformed the response variable using a logarithmic function.

Upon visualizing the model diagnostics, we found that, while the plot of fitted values versus residuals generally centered around zero, there was a change in clustering pattern around a certain fitted value and QQ plot shows deviation from normality. Furthermore, Cook's distance indicated the presence of outliers and identified as influential. By fitting another model without these influential observations, we observed an improvement in the regression model, with the adjusted R-squared increasing from 0.1758 to 0.2313, while the coefficients remained significant. The model without outliers demonstrated improved normality, as evidenced by the normal QQ plot. The residual versus fitted plot displayed a cloud of observations with a constant gap between each cluster, though they were symmetrically distributed around the y=0 line. Therefore, the assumptions of linearity, constant variance, and normality were satisfied.



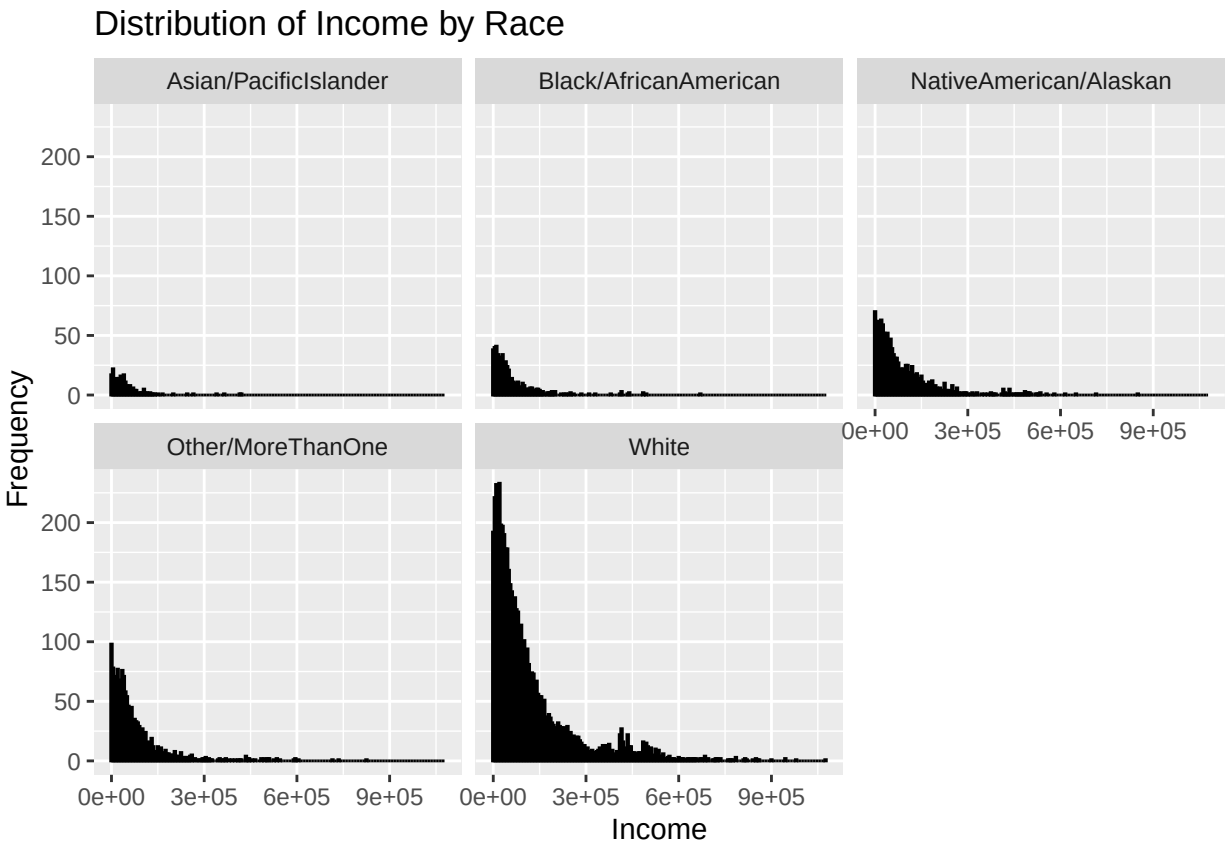




Further examination of the Variance Inflation Factor (VIF) values revealed multicollinearity issues, particularly with the Sex and Sex:Age variables, which exhibited VIF values of 8.46 and 9.14, respectively. We employed Ridge regression to address the issue. Using 10-fold cross-validation, we determined the optimal lambda value to be $\lambda = 0.04258514$. The resulting regression model is provided below:

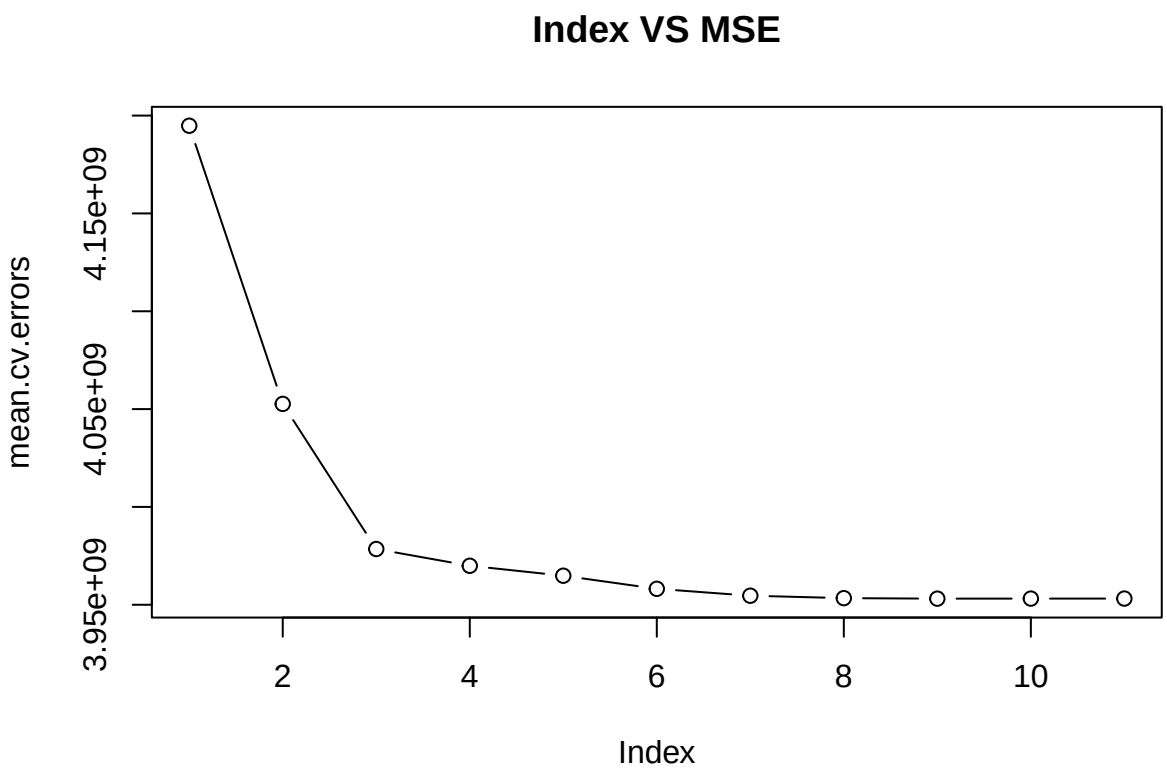
$$log(Income) = 9.436 + 0.379(HS, GED, or Associates Degree) +$$
$$1(Bachelors Degree) + 1.323(Masters Degree or higher) +$$
$$0.232(Male) + 0.002(Age) + 0.003(Male : Age)$$

Question 2



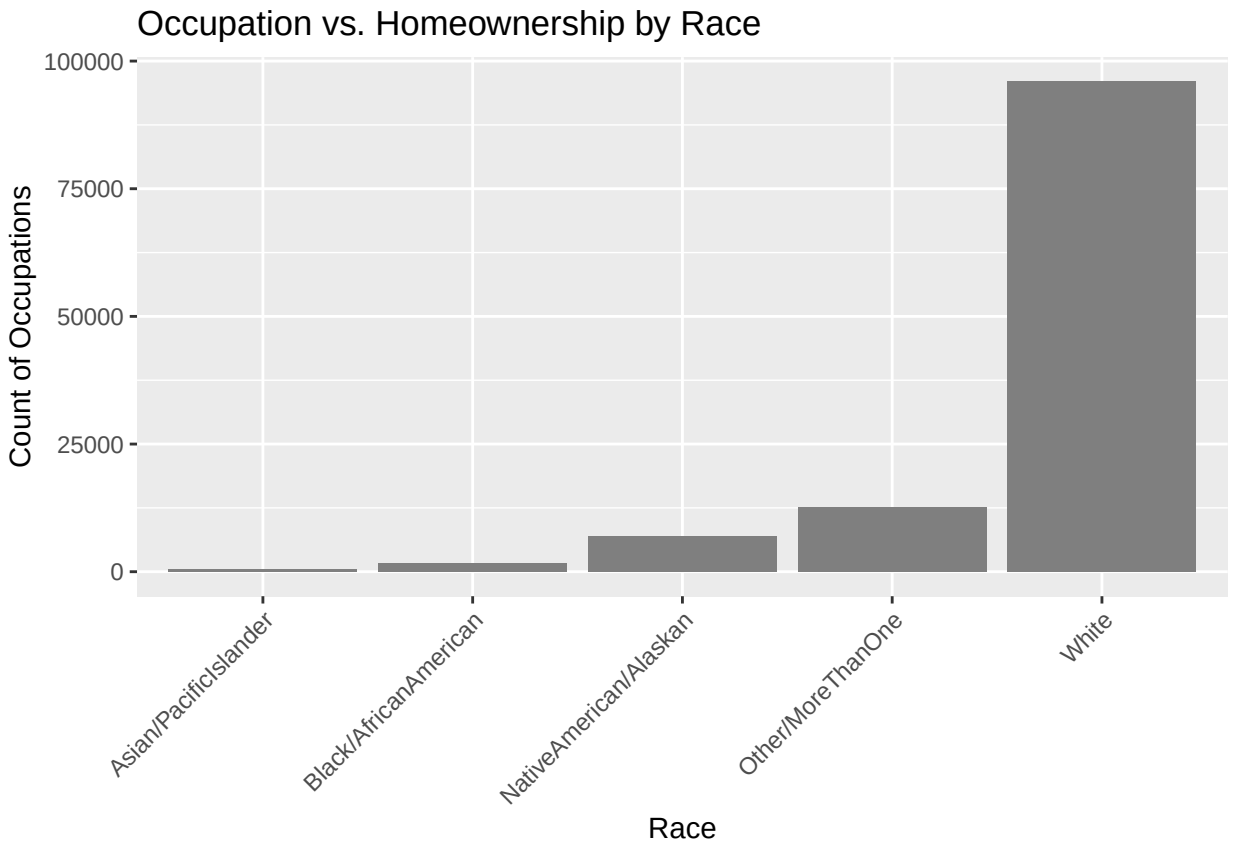
Can we predict household income based on education, occupation, and race/ethnicity? We explored the predictive power of EducationGroup, OccupationGroup, and Race in forecasting household income through the Validation-Set Approach and Cross-Validation. Initially, we separated our dataset into training and testing sets. Employing best subset selection exclusively on the training data, we determined the most effective model for each model size, assessing the validation set error accordingly. Consequently, the model that includes nine variables identified as the most optimal. This process was then replicated on the entire dataset, confirming that the best nine-variable model from the training data are identical to the full dataset.

Furthermore, we utilized cross-validation to choose models of varying sizes. 10-fold cross-validation was deployed and resulted in the selection of an eleven-variable model, which are all variables present in the dataset. Therefore, we performed best subset selection on the full data set in order to obtain the 11-variable model, we obtained all variables. The plot below shows that the model with 11 variables has the least mean square error.



Question 3

Is there a relationship between Occupation, home ownership, race/ethnicity and Marial Status, OccupationGroup, income?



From the model fitting, Model 1 that includes MaritalStatus, OccupationGroup, Homeownership, Race, and Income has a lower residual sum of squares (RSS) compared to Model 2, where Race is excluded, indicating that it fits the data better. The F-statistic for the comparison between Model 1 and Model 2 is 5.4446, witha very low p-value (6.915e-07), indicating that the additional predictors inModel 1 significantly improve the fit of the model.

Based on this information, we can conclude that Model 1 performs better than Model 2. The additional predictors in Model 1 (Race) significantly improve the model's ability to explain the variation in the occupation variable compared to Model 2.

Based on our findings, Model 3, which includes the interaction between MaritalStatus and OccupationGroup demonstrated a better fit to the data compared to Model 4 that includes the interaction between OccupationGroup and Homeownership, as evidenced by its lower residual sum of squares (RSS). Additionally, when comparing Model 1 and Model 2, Model 1 exhibited a lower RSS and the same level of degrees of freedom, indicating a better model fit in this comparison. However, given that Model 3 outperformed Model 4 in terms of RSS and considering the relative performance of Model 1 and Model 2 from our findings, we would conclude that Model 3 is the preferred choice based on current data.

Interpret the data

Question 1

To answer our question, does higher education mean higher income accounting for sex and age, we utilized the fitted model to compare income levels across four educational categories for 50-year-old males and females in Oregon (mean and median). The results showed a clear positive relationship between educational attainment and income for both sexes.

For 50-year-old males in Oregon, educational attainment significantly affects median income. Less than high school: $21,273.80. High school, GED, or associate degree: $31,089.90. Bachelor's degree: $57,829.09. Master's degree or higher: $79,867.54. Similarly, for females: less than high school: $14,136.47. High school, GED, or associate degree: $20,659.28. Bachelor's degree: $38,427.51. Master's degree or higher: $53,072.1. These results show a clear positive

correlation between education and income for both genders, with significant increases seen at bachelor's degree level and above.

Question 2

To determine if we can predict household income based on education, occupation, and race/ethnicity, we conducted validation and cross-validation tests and analyses on our dataset. our current findings suggest that Education, Race, and OccupationGroup are significant predictors for household income. These findings shed light on the key factors influencing income levels and could guide future research and policy efforts aimed at reducing income disparities.

Question 3

To investigate the relationship between race/ethnicity and home ownership, we employed a linear regression model that incorporated MaritalStatus, OccupationGroup, Homeownership, and Race against MaritalStatus and OccupationGroup to delve into the various insights we can gather from the model's coefficients, along with various diagnostic plotting. From there we can gather a meaningful conclusion on the strength of these relationships.

Based off what we saw via the summary of the linear regression model, we get the following formula to use:

$$\begin{aligned} Occupation = {} & 3.779623 \times Intercept - 5.682494 \times Married + 5.002508 \times NotMarried/Under15 \\ & -69.46579 \times Separated - 23.86628 \times Widowed + 5092.785 \times Manufacturing/Transportation \\ & +5967.028 \times Military - 2301.276 \times Professional/Technical + 1500.935 \times WhiteCollar/BlueCollar \\ & -86.83271 \times Occupiedwithoutpaymentofrent + 31.75971 \times Ownedfreeandclear + 13.18805 \times Ownedwithmortgageorloan \\ & +10.9471 \times Rented + 41.31494 \times Black/AfricanAmerican + 91.19606 \times NativeAmerican/Alaskan \\ & +113.0133 \times Other/MoreThanOne + 112.7199 \times White - 0.0008919726 \times Income \end{aligned}$$

This formula allows us to see how each variable contributes to our home ownership values. We employed a cross-validation approach using 10 folds (the default) to assess the model's performance. We began with splitting our data into a training and testing set, and then trained and evaluated the results. since we employed cross-validation, we repeated this 10 time times, and found the average Root Mean Square Error of 896.4357 and Mean Absolute Error of 739.1021 for the folds.

We also considered an interaction term-focused model between marital status and occupation group, which resulted in a average RMSE of 895.5156 and a MAE of 737.0633. This slightly lower RMSE and MAE suggests that the interaction between marital status and occupation group provides additional value to predicting home ownership.

Obstacles

Git collaboration

At the beginning, team members made changes directly to the main branch without checking others' work. Later, we switched to personal branches for validation before merging into the main. Conflicts arose when pushing changes simultaneously, but was resolved using Git commands.

Future work/improvement

For our team data, in the categorical variable, there are too many categories. It causes our R notebook compute time to be too long and hard to interpret the result when we have so many categories. In the future, when we do data cleaning, there should be fewer categories in each variable.

Conclusion

The analysis for the first question revealed a strong positive correlation between educational attainment and income for both males and females aged 50 in Oregon. Individuals with higher levels of education, particularly those holding bachelor's degrees or above, were found to have significantly higher median incomes compared to those with lower educational attainment.

Furthermore, the validation set and cross-validation analyses confirmed that EducationGroup, Race, and Occupation-Group are significant predictors of household income. These results help underscore the importance of these factors in shaping income levels and highlight potential areas for future research and policy interventions aimed at reducing income disparities.

Lastly, the study explored the complex relationships between Occupation, home ownership, race/ethnicity, Marital Status, OccupationGroup, and income. The selected models provided valuable insights into how these variables interact and influence occupation and income outcomes.

Therefore, the Data Wizards have conducted a comprehensive analysis of the American Community Survey data to find from this study valuable insights into the factors influencing income levels and disparities within the state of Oregon.

Appendix

```
##An if statement for checking if a package is installed

if (!require(tidycensus)) {
  install.packages("tidycensus")
}
if (!require(tidyverse)) {
  install.packages("tidyverse")
}
if (!require(dplyr)) {
  install.packages("dplyr")
}
if (!require(ggplot2)) {
  install.packages("ggplot2")
}
if (!require(faraway)) {
  install.packages("faraway")
}
if (!require(car)) {
  install.packages("car")
}
if (!require(gridExtra)) {
  install.packages("gridExtra")
}
if (!require(glmnet)) {
  install.packages("glmnet")
}
if (!require(ISLR2)) {
  install.packages("ISLR2")
}
if (!require(leaps)) {
  install.packages("leaps")
}

#Load libraries
library(tidycensus)
```

```r
library(tidyverse)
library(dplyr)
library(ggplot2)
library(faraway)
library(car)
library(gridExtra)
library(glmnet)
library(ISLR2)
library(leaps)

census_api_key("2547c95ce33b1ed0eec3aafd0fd8526a5bb9a22e")

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

#Rename the columns
puma_data <- puma_data %>% dplyr::rename(Age = AGEP,
                                  EducationLevel = SCHL,
                                  Income = PINCP,
                                  Sex = SEX,
                                  PersonNumber = SPORDER,
                                  Race = RAC1P,
                                  Homeownership = TEN,
                                  Occupation = OCCP,
                                  MaritalStatus = MAR,
                                  HouseAcreage = ACR
                                  )
head(puma_data[, 2:8])

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

#Convert education levels
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)

#Convert Occupation to numerical
puma_data <- puma_data %>%
  mutate(
    Occupation = case_when(
      Occupation == "bbbb" ~ "0000",
      Occupation == "000N" ~ "0000",
      TRUE ~ Occupation
    ),
```

```
    Occupation = as.numeric(Occupation)
  )

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

#Convert homeownership
puma_data <- puma_data %>%
  mutate(
    Homeownership = case_when(
      Homeownership == "b" ~ "0",
      TRUE ~ Homeownership
    ),
    Homeownership = as.numeric(Homeownership)
  )

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

#Convert HouseAcreage
puma_data <- puma_data %>%
  mutate(
    HouseAcreage = case_when(
      HouseAcreage == "b" ~ "0",
      TRUE ~ HouseAcreage
    ),
    HouseAcreage = as.numeric(HouseAcreage)
  )

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

##Group education level
puma_data <- puma_data %>%
  mutate(
     EducationGroup = cut(EducationLevel, breaks = c(0, 16, 21, 22, Inf),
                          labels = c("Less than HS", "HS, GED, or Associates Degree",
                                     "Bachelors Degree", "Masters Degree or higher"),
                          right = FALSE)
  )
unique(puma_data$EducationGroup)
puma_data[1:50,]

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


#Drop income that are less than 1 from the data set
#Drop observation with 0 and negative income
puma_data <- puma_data %>% filter(Income >= 1)
head(puma_data)
sum(puma_data$Income<1)

#relabeling race
puma_data$Race <- as.numeric(puma_data$Race)
```

```r
puma_data <- puma_data %>%
  mutate(
    Race = case_when(
      Race == 1 ~ "White",
      Race == 2 ~ "Black/AfricanAmerican",
      between(as.numeric(Race), 3, 6) ~ "NativeAmerican/Alaskan",
      between(as.numeric(Race), 6, 7) ~ "Asian/PacificIslander",
      between(as.numeric(Race), 8, 9) ~ "Other/MoreThanOne",
      TRUE ~ as.character(Race)
    )
)
puma_data$Race[1:100]

#relabeling MaritalStatus
puma_data <- puma_data %>%
  mutate(
    MaritalStatus = case_when(
      MaritalStatus == "1" ~ "Married",
      MaritalStatus == "2" ~ "Widowed",
      MaritalStatus == "3" ~ "Divorced",
      MaritalStatus == "4" ~ "Separated",
      MaritalStatus == "5" ~ "Not Married/Under 15",
      TRUE ~ as.character(MaritalStatus)
    )
)

#Checking labels are assigned.
unique(puma_data$MaritalStatus)
puma_data$MaritalStatus[11200:11300]

#relabeling Acreage
puma_data <- puma_data %>%
  mutate(
    HouseAcreage = case_when(
      HouseAcreage == 0 ~ "Not a one-family home",
      HouseAcreage == 1 ~ "< 1 Acre",
      HouseAcreage == 2 ~ "1 - 10 Acres",
      HouseAcreage == 3 ~ "> 10 Acres",
      TRUE ~ as.character(HouseAcreage)
    )
)

#Checking labels are assigned.
unique(puma_data$HouseAcreage)
puma_data$HouseAcreage[1500:1600]

#relabeling tenure
puma_data <- puma_data %>%
  mutate(
    Homeownership = case_when(
      Homeownership == 0 ~ "N/A",
      Homeownership == 1 ~ "Owned with mortgage or loan",
      Homeownership == 2 ~ "Owned free and clear",
      Homeownership == 3 ~ "Rented",
      Homeownership == 4 ~ "Occupied without payment of rent",
      TRUE ~ as.character(Homeownership)
```

```
    )
)

#Checking labels are assigned.
unique(puma_data$Homeownership)
puma_data$Homeownership[11200:11300]

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

#Checking labels are assigned.
unique(puma_data$Sex)
head(puma_data$Sex)

#Occupation Grouping
puma_data <- puma_data %>%
  filter(as.numeric(Occupation) >= 0010 & as.numeric(Occupation) < 9920) %>%
  mutate(
    OccupationGroup = case_when(
      between(as.numeric(Occupation), 0010, 3550) ~ "Professional/Technical",
      between(as.numeric(Occupation), 3600, 4160) ~ "Healthcare/FoodServices",
      between(as.numeric(Occupation), 4200, 7640) ~ "WhiteCollar/BlueCollar",
      between(as.numeric(Occupation), 7700, 9760) ~ "Manufacturing/Transportation",
      between(as.numeric(Occupation), 9800, 9830) ~ "Military",
      TRUE ~ "Not Classified"
    )
  )

#Checking labels are assigned.
unique(puma_data$OccupationGroup)
puma_data$OccupationGroup[2020:2120]
puma_data$Occupation[puma_data$OccupationGroup == "Not Classified"]

puma_data_raw <- puma_data
head(puma_data_raw)

#Rearrange the columns
puma_data <- puma_data %>%
  select(EducationGroup, EducationLevel, OccupationGroup, Occupation, Income, Age, Sex, Race, Homeownership,
head(puma_data)

#Summarizing the ranges of numerical
summary(puma_data$Age) #Range of Age
summary(puma_data$Income) #Range of Age
summary(puma_data$EducationLevel) #Range of Education
table(puma_data$EducationGroup) #Count of each level of education
table(puma_data$Race) #Count of each race
table(puma_data$Homeownership) #Count of each Homeownership
table(puma_data$OccupationGroup) #Count of each Occupationgroup
```

```r
#Plot income 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
# 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)

#Compare model 1 and 2
anova(model_1, model_2)

#The full model provides a significantly better fit.

#Linearity and Homoscedasticity assumption
plot(fitted(model_2), residuals(model_2))
```

```r
#Normality assumption
qqnorm(residuals(model_2))
qqline(residuals(model_2))
```

```r
#Independence assumptions
plot(residuals(model_2))
```

```r
#Fitted vs residuals plot exhibiting pattern

#Logged models
model_3 <- lm(log(Income) ~ EducationGroup + Sex + Age, data = puma_data, na.action = na.exclude)
model_4 <- lm(log(Income) ~ EducationGroup + Sex * Age, data = puma_data, na.action = na.exclude)

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

#Model 4 is the best from 2 models

#Linear regression assumption
plot(fitted(model_4), residuals(model_4))
```

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

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

```r
#Add observation ID
puma_data$ID <- 1:nrow(puma_data)

## Attaching the Case Influence Statistics with the Data
Data <- fortify(model_4, puma_data)
```

```r
## Plot the Case Influence Statistics for each observation (subject)
par(mfrow=c(1,3))
qplot(ID,.hat, data = Data)
```

```r
qplot(ID,.stdresid, data = Data)
```

```r
qplot(ID,.cooksd, data = Data)
```

```r
#There is evidence of outliers

#Test for influential observations
#Cut off point is 4/(n-k-2)
cutoff <- 4/((nrow(puma_data)-length(model_4$coefficients)-2))

#Calculate Cook's distance
cook.d <- cooks.distance(model_4)

#Plot to identify observation with Cook's distance higher than cutoff
plot(cook.d, pch=".", cex=2, main="Influential Obs by Cooks distance")  # plot cook's distance
abline(h = cutoff, col="red")  # add cutoff line
text(x=1:length(cook.d)+1, y=cook.d, labels=ifelse(cook.d>cutoff, names(cook.d),""), col="red")  # add labe

# Removing Outliers
# influential row numbers
influential <- as.numeric(names(cook.d)[(cook.d > cutoff)])

#Create another data frame without influential observations
puma_data_2 <- puma_data[-influential, ]

#Fit a regression without influential
model_5 <- lm(log(Income) ~ EducationGroup + Sex * Age, data = puma_data_2, na.action = na.exclude)

#Compare the summary before and after removing outliers
summary(model_4)
summary(model_5)

par(mfrow=c(2,2))
```

```r
#Linear regression assumption
plot(fitted(model_5), residuals(model_5))

#Normality assumption
qqnorm(residuals(model_5))
qqline(residuals(model_5))

#Independence assumptions
plot(residuals(model_5))


#Multi-colinearlity
options(warn=-1)
vif(model_5)
options(warn=0)

#Fit ridge regression
x <- model.matrix(log(Income) ~ EducationGroup + Sex * Age, puma_data_2)[, -1]
```

```r
y <- log(puma_data_2$Income)

ridge_model <- glmnet(x, y, alpha = 0)


summary(ridge_model)

#perform 10 fold cross-validation to find optimal lambda value
cv_model <- cv.glmnet(x, y, alpha = 0)

#find optimal lambda value that minimizes test MSE
best_lambda <- cv_model$lambda.min
best_lambda

#produce plot of test MSE by lambda value
plot(cv_model)
```

```r
#The lambda value that minimizes the test MSE turns out to be 0.04258514

#find coefficients of model with best lambda
model_6 <- glmnet(x, y, alpha = 0, lambda = best_lambda)
coef(model_6)

#produce Ridge trace plot
plot(ridge_model, xvar = "lambda")

puma_data_q2 <- puma_data |>
  select(EducationGroup, Income, Race, OccupationGroup)
head(puma_data_q2)

summary(puma_data_q2)

#Data visualization
puma_data_q2_race_income <- puma_data_q2 |>
  group_by(Race, Income) |>
  summarise(n = n())

ggplot(puma_data_q2_race_income, aes(x = Income)) +
  geom_histogram(binwidth = 5000, fill = "skyblue", color = "black") +
  facet_wrap(~Race) +
  labs(title = "Distribution of Income by Race", x = "Income", y = "Frequency")

#Data visualization
puma_data_q2_occ_income <- puma_data_q2 |>
  group_by(OccupationGroup, Income) |>
  summarise(n = n())

ggplot(puma_data_q2_occ_income, aes(x = Income)) +
  geom_histogram(binwidth = 5000, fill = "skyblue", color = "black") +
  facet_wrap(~OccupationGroup) +
  labs(title = "Distribution of Income by Occupation Group", x = "Income", y = "Frequency")


#Validation set and CV
#Separate data into 2 groups, training and testing data
set.seed(538)
train <- sample(c(TRUE, FALSE), nrow(puma_data_q2), replace = TRUE)
```

```
test <- (!train)


#apply regsubsets() to the training set in order to perform best subset selection
regfit.best <- regsubsets(Income ~ ., data = puma_data_q2[train, ], nvmax = 11)

#make a model matrix from the test data
test.mat <- model.matrix(Income ~ ., data = puma_data_q2[test, ])

#run a loop, and for each size i
val.errors <- rep(NA, 11)

for (i in 1:11) {
coefi <- coef(regfit.best, id = i)
pred  <- test.mat[, names(coefi)] %*% coefi
val.errors[i] <- mean((puma_data_q2$Income[test] - pred)^2)
}

#Show MSE
val.errors
which.min(val.errors)
coef(regfit.best, 9)

#perform best subset selection on the full data set, and select the best 9 variable model
regfit.best <- regsubsets(Income ~ ., data = puma_data_q2, nvmax =11)
coef(regfit.best, 9)

k <- 10
n <- nrow(puma_data_q2)
set.seed(1)
folds <- sample(rep(1:k, length = n))
cv.errors <- matrix(NA, k, 11,
                    dimnames = list(NULL, paste(1:11)))


#Create predict function
predict.regsubsets <- function(object, newdata, id, ...) {
form <- as.formula(object$call[[2]])
mat <- model.matrix(form, newdata)
coefi <- coef(object, id = id)
xvars <- names(coefi)
mat[, xvars] %*% coefi
}


for (j in 1:k) {
best.fit <- regsubsets(Income ~ .,
data = puma_data_q2[folds != j, ],
nvmax = 11)
for (i in 1:11) {
pred <- predict(best.fit, puma_data_q2[folds == j, ], id = i)
cv.errors[j, i] <- mean((puma_data_q2$Income[folds == j] - pred)^2) }
}


mean.cv.errors <- apply(cv.errors, 2, mean)
mean.cv.errors
```

```r
par(mfrow = c(1, 1))
```

```r
plot(mean.cv.errors, type = "b", main = "Index VS MSE")
```

```r
#Cross-validation selects a 11-variable model
#perform best subset selection on the full data set in order to obtain the 11-variable model
#11 variables includes all variables in the dataset
reg.best <- regsubsets(Income ~ ., data = puma_data_q2, nvmax = 11)
coef(reg.best, 11)

head(puma_data_raw)

puma_data_q3 <- puma_data_raw |>
  select(Homeownership, Race, MaritalStatus, OccupationGroup, Occupation, Income)
head(puma_data_q3)

puma_data_filtered <- puma_data_q3 |>
  filter(Occupation != 0)

par(mfrow=c(2,2))

ggplot(puma_data_filtered, aes(x = Occupation)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribution of Occupations", x = "Occupation", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
ggplot(puma_data_q3, aes(x = OccupationGroup, y = Occupation)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "The relationship between OccupationGroup and Occupation Size", x = "OccupationGroup", y = "O
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
ggplot(puma_data_q3, aes(x = Homeownership, y = Occupation)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "The relationship between Homeownership and Occupation Size", x = "Homeownership", y = "Occup
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
ggplot(puma_data_q3, aes(x = Race, fill = factor(Occupation))) +
  geom_bar() +
  labs(title = "Occupation vs. Homeownership by Race", x = "Race", y = "Count of Occupations") +
  scale_fill_manual(values = c("White" = "blue", "Black" = "red")) +  # Customize colors for each race
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
model_1_q3 <- lm(Occupation ~ as.factor(MaritalStatus) + as.factor(OccupationGroup)
      + as.factor(Homeownership) + as.factor(Race) + Income, data = puma_data_q3)

summary(model_1_q3)

model_2_q3 <- lm(Occupation ~ as.factor(MaritalStatus) + as.factor(OccupationGroup)
      + as.factor(Homeownership) + Income, data = puma_data_q3)

summary(model_2_q3)

#Compare models 1 and model 2
anova(model_1_q3, model_2_q3)
```

```r
plot(fitted(model_1_q3), residuals(model_1_q3))


qqnorm(residuals(model_1_q3))
qqline(residuals(model_1_q3))


plot(residuals(model_1_q3))

#Fitted vs residuals plot exhibiting pattern

model_3_q3 <- lm(Occupation ~ as.factor(MaritalStatus) * as.factor(OccupationGroup)
      + as.factor(Homeownership) + as.factor(Race) + Income, data = puma_data_q3)

model_4_q3 <- lm(Occupation ~ as.factor(MaritalStatus) + as.factor(OccupationGroup)
      * as.factor(Homeownership) + as.factor(Race) + Income, data = puma_data_q3)

summary(model_3_q3)

summary(model_4_q3)

anova(model_3_q3, model_4_q3)

par(mfrow=c(1,3))
```

```r
plot(fitted(model_3_q3), residuals(model_3_q3))


qqnorm(residuals(model_3_q3))
qqline(residuals(model_3_q3))


plot(residuals(model_3_q3))
```

```r
#Fitted vs residuals plot exhibiting pattern

#Calculate median income for 50 YO males
coef <- coef(model_6)
lw_HS_m  <- coef[1, ] + coef[5, ] + (50*coef[6, ]) + (50*1*coef[7, ]) #lower than HS male
GED_m    <- coef[1, ] + coef[2, ] + coef[5, ] + (50*coef[6, ]) + (50*1*coef[7, ]) #HS, GED, Associate degree
bach_m   <- coef[1, ] + coef[3, ] + coef[5, ] + (50*coef[6, ]) + (50*1*coef[7, ]) #Bachelor's degree male
master_m <- coef[1, ] + coef[4, ] + coef[5, ] + (50*coef[6, ]) + (50*1*coef[7, ]) #master or higher male

exp(lw_HS_m)
exp(GED_m)
exp(bach_m)
exp(master_m)

#Calculate median income for 50 YO females

lw_HS_f  <- coef[1, ] + (50*coef[6, ]) #lower than HS female
GED_f    <- coef[1, ] + coef[2, ] + (50*coef[6, ]) #HS, GED, Associate degree female
bach_f   <- coef[1, ] + coef[3, ] + (50*coef[6, ]) #Bachelor's degree female
master_f <- coef[1, ] + coef[4, ] + (50*coef[6, ]) #master or higher female

exp(lw_HS_f)
```

```r
exp(GED_f)
exp(bach_f)
exp(master_f)

par(mfrow=c(2,2))
summary(model_1_q3)$coefficients
summary(model_1_q3)$r.squared
plot(model_1_q3)
```

```r
train_index <- sample(1:nrow(puma_data_q3), 0.8 * nrow(puma_data_q3))  # 80% for training
train_data <- puma_data_q3[train_index, ]
test_data <- puma_data_q3[-train_index, ]


num_folds <- 10

model_1_q3 <- lm(Occupation ~ as.factor(MaritalStatus) + as.factor(OccupationGroup)
      + as.factor(Homeownership) + as.factor(Race) + Income, data = puma_data_q3)

test_rmse <- numeric(num_folds)
test_mae <- numeric(num_folds)


for (i in 1:num_folds) {

  train_index <- sample(1:nrow(puma_data_q3), 0.8 * nrow(puma_data_q3))  # 80% for training
  train_data <- puma_data_q3[train_index, ]
  test_data <- puma_data_q3[-train_index, ]


  model_1 <- lm(Occupation ~ as.factor(MaritalStatus) + as.factor(OccupationGroup)
      + as.factor(Homeownership) + as.factor(Race) + Income, data = train_data)


  predictions <- predict(model_1, newdata = test_data)


  test_rmse[i] <- sqrt(mean((test_data$Occupation - predictions)^2))
  test_mae[i] <- mean(abs(test_data$Occupation - predictions))
}


avg_test_rmse <- mean(test_rmse)
avg_test_mae <- mean(test_mae)

# Print average test set performance
cat("Average Test RMSE:", avg_test_rmse, "\n")
cat("Average Test MAE:", avg_test_mae, "\n")



test_rmse <- numeric(num_folds)
test_mae <- numeric(num_folds)


for (i in 1:num_folds) {
```

```
  train_index <- sample(1:nrow(puma_data_q3), 0.8 * nrow(puma_data_q3))  # 80% for training
  train_data <- puma_data_q3[train_index, ]
  test_data <- puma_data_q3[-train_index, ]


  model_3 <- lm(Occupation ~ as.factor(MaritalStatus) * as.factor(OccupationGroup)
      + as.factor(Homeownership) + as.factor(Race) + Income, data = puma_data_q3)


  predictions <- predict(model_3, newdata = test_data)


  test_rmse[i] <- sqrt(mean((test_data$Occupation - predictions)^2))
  test_mae[i] <- mean(abs(test_data$Occupation - predictions))
}


avg_test_rmse <- mean(test_rmse)
avg_test_mae <- mean(test_mae)

# Print average test set performance
cat("Average Test RMSE:", avg_test_rmse, "\n")
cat("Average Test MAE:", avg_test_mae, "\n")
```