Project 1 - Final Report

Regression Analysis Public Use Microdata Sample (PUMS)

Mitchell Below, Kyle Livermore, Ruslan Mamedov

2022-05-30

Introduction

The question of interest for this analysis is: Does the mean income for persons with a degree in Statistics and Decisions Sciences differ across four industries (Finance, Manufacturing, Medicine, and Academics) after accounting for confounding variables such as age, sex, marital status, years of schooling, hours worked per week, weeks worked during the past 12 months, place of birth, and presence of children in the home?

H~0: The mean total income for each of these industries is equal after controlling for differences in the other proposed explanatory variables.

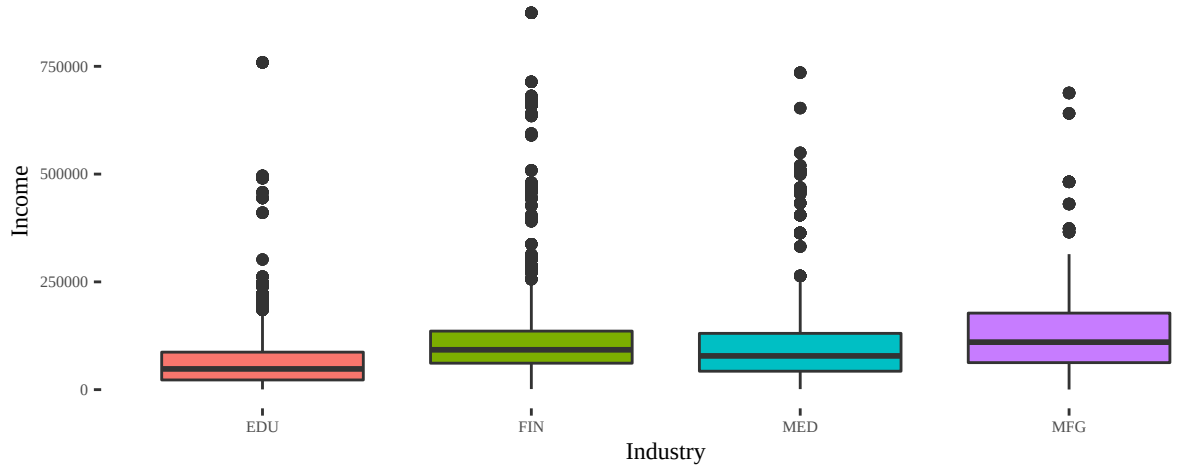
H~A: The mean total income for each of these industries is not equal after controlling for differences in the other proposed explanatory variables.

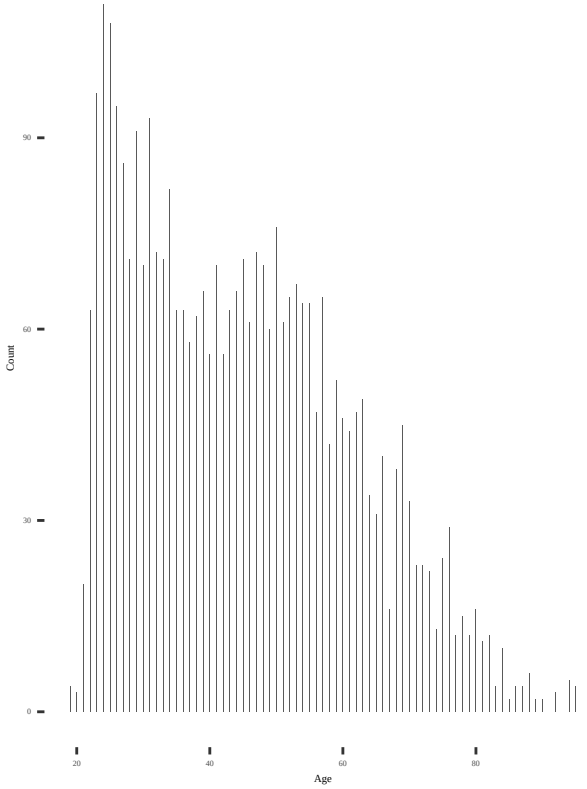
Our data source is the 5-year data from the American Community Survey Public Use Microdata Sample (PUMS) filtered on records listing field of degree as Statistics and Decision Science.

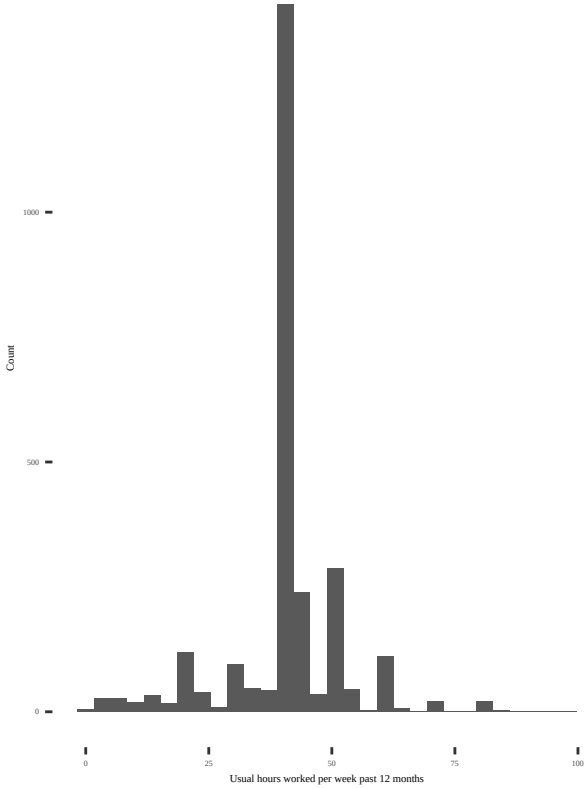
Because this is a question of statistical inference, we will fit a model to the data, test the model against several others to determine the best fit, and determine the significance of the coefficients of our proposed explanatory variables. We will determine a adjusted R-squared for our model and a p-value for the significance of any difference in means.

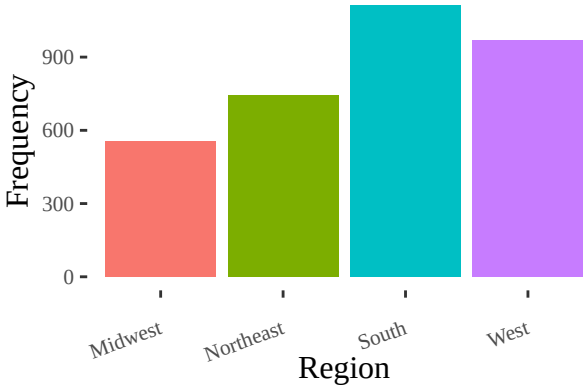
Exploratory Data Analysis

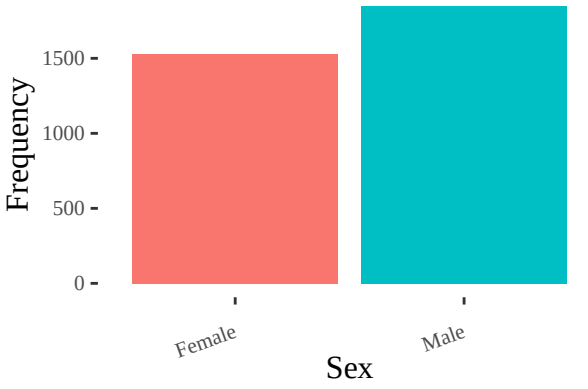
First, let's have some exploratory analysis of the data: distribution of unique values for both numerical and categorical data, summary of missing values.

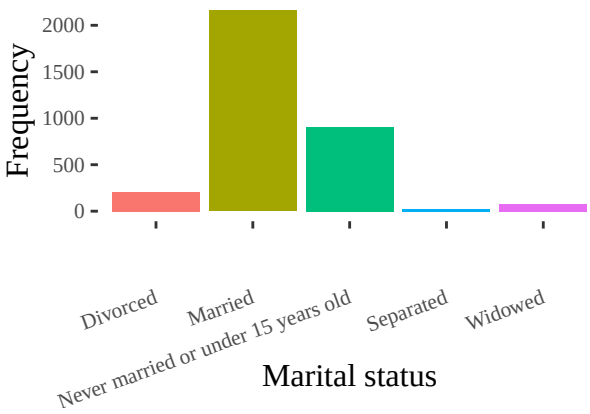


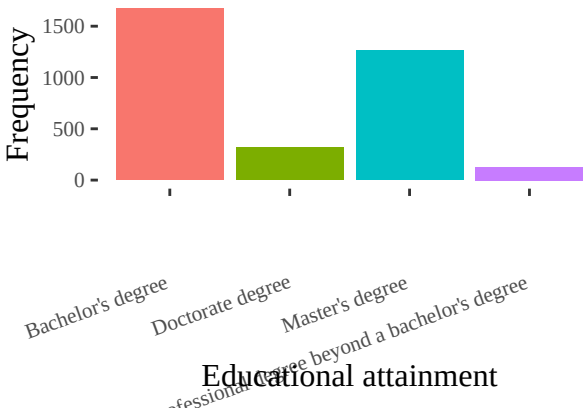

Count
Age


Count
Usual hours worked per week past 12 months

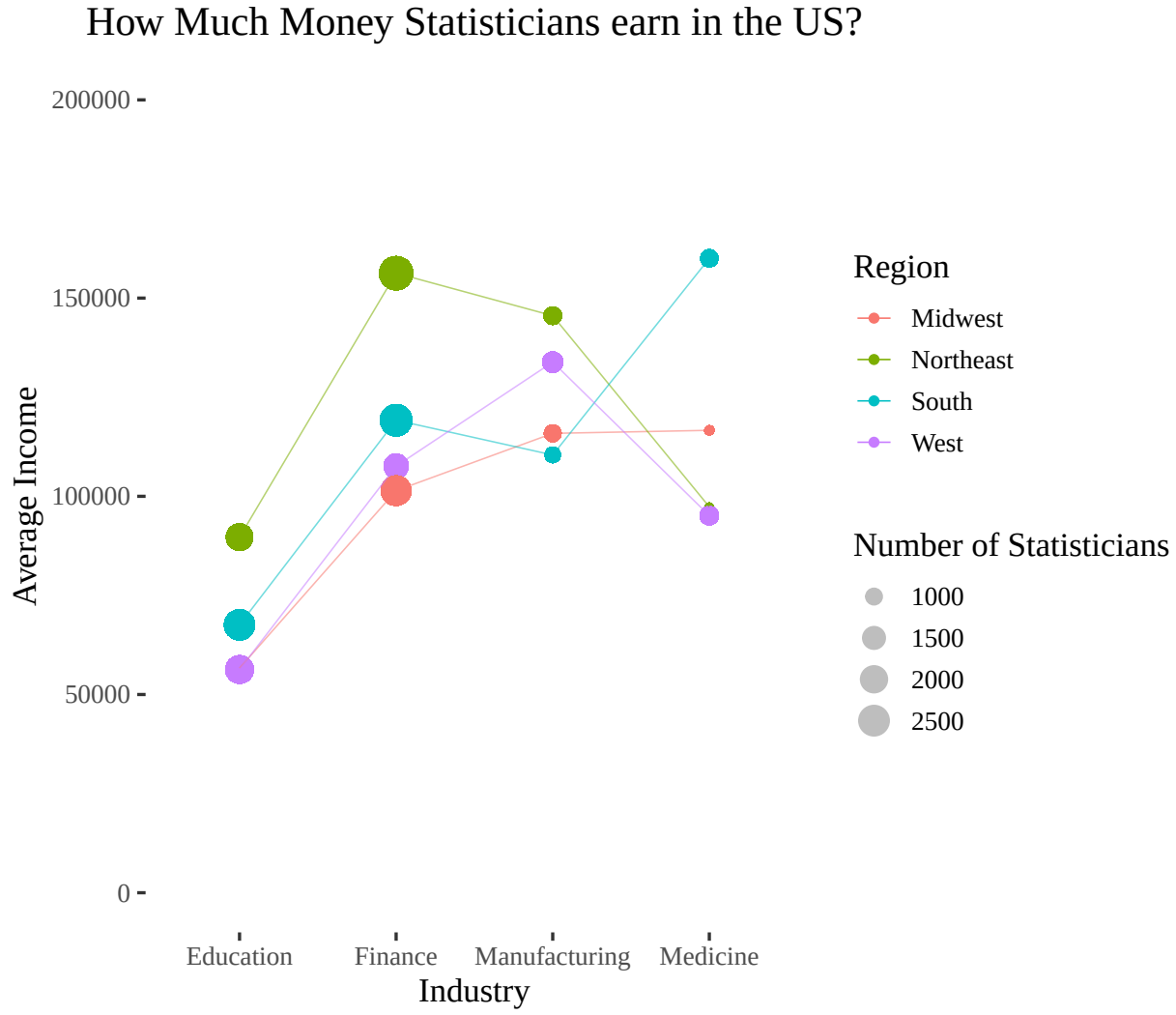

Frequency
Midwest
Northeast
South
West
Region


Frequency
Female
Male
Sex


Frequency
Divorced
Married
Never married or under 15 years old
Separated
Widowed
Marital status


Frequency
Bachelor's degree
Doctorate degree
Master's degree
Educational attainment

```
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```



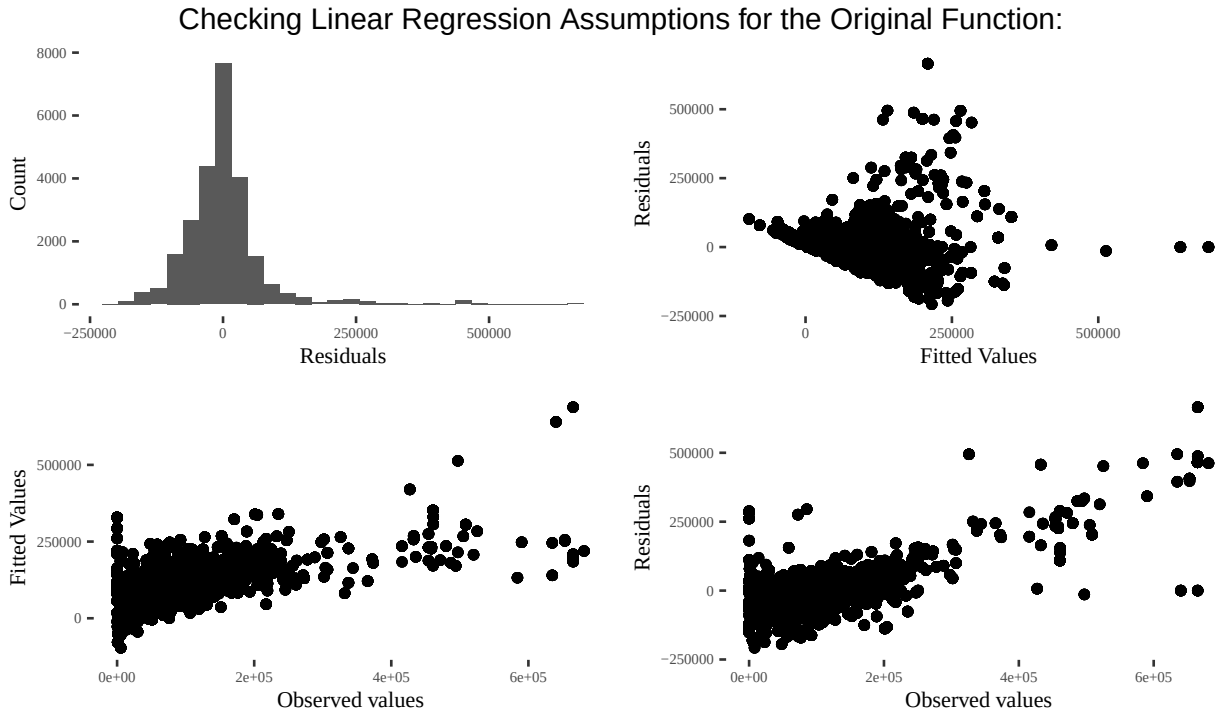
There appear to be some differences in income depending on the industry. Now let's build an inferential model to test that hypothesis and estimate the differences. We'll be adjusting for the following factors: "Region", "Age,"Class of worker", "Educational attainment", "Sex", "Usual hours worked per week past 12 months", "Weeks worked during past 12 months", "Health insurance coverage recode", "Married, spouse present/spouse absent", "Nativity", "Presence and age of own children", "Place of birth"

Preliminary model fit

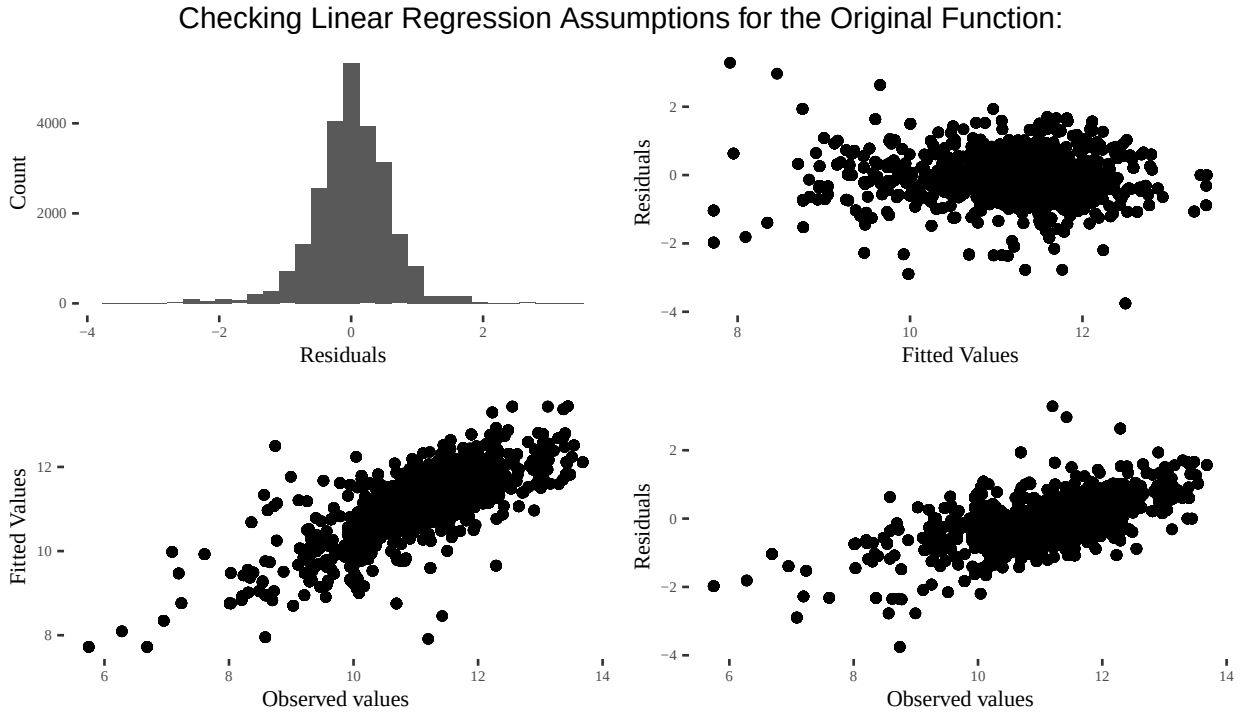
```
#lm.inference<-lm(PINCP_Adj ~ IND_Category + REGION + AGEP + COW + SCHL + SEX +
#+WKHP + WKW + HICOV + MSP + NATIVITY + PAOC + POBP, data=project1_red)
```

Model checking and selection

Next, let's perform residual diagnostics to check the fit for the model to confirm it satisfies the regression assumptions.



The histogram of the residuals is skewed to the right and that could be concerning. Let's transform the response with log operator and see if it helps with the residuals distribution:



Log-transformation of the response did help with the residuals and overall fit of the linear model. Our final step will be back-transformation of the confidence intervals and interpretation of results:

```
##                                2.5%    97.5% Mean Difference
## Industry: Finance       62.59967 72.30363        67.38134
## Industry: Medicine      61.03101 71.77709        66.31728
## Industry: Manufacturing 33.28532 41.91565        37.53280
```

Results

The linear model we chose has an adjusted R-square of 0.4641, suggesting that it explains approximately 46% of the difference in mean total income. The coefficients for the industry variables we focused on - FIN, ACA, MED, and MFG - were all significant at a p-level of $p \sim 0$. Age, level of education attained, and hours worked per week also had highly significant coefficient, along with region and sex.

Obstacles

The first obstacle encountered with this analysis was slicing out the data of interest from the entire American Community Survey 5 year PUMS data. To accomplish this a line reading python script was implemented to separate data entries where the degree was Statistics And Decision Science. It was necessary to parse the data at a line level because the US level data was too large to fit in memory. Once the data was obtained

the data had to be expanded, as it was in a compressed format where each row represented multiple people. Additionally, all of the monetary fields had to be adjusted for annual inflation to standardize across the years of collection. This was done with the provided annual adjustment factor in the the PUMS data.

Conclusion

There is strong evidence that that the mean total incomes of persons with a degree in Statistics and Decision Sciences is not equal across the industry categories of Academics, Finance, Manufacturing, and Medicine (p-value < 2.2e-16 with t-test on each of the industry indicators, df = 161). It is estimated that working in finance industry increased the mean income by 67%, in medicine - by 37%, in manufacturing - by 66%, when compared against education industry. The model was adjusted for confounding variables such as age, sex, marital status, years of schooling, hours worked per week, weeks worked during the past 12 months, place of birth, and presence of children in the home.

Future research could include examining income differences by sub-industry and occupation, or developing a model to predict income based on some subset of the explanatory variable included in this analysis. Also of interest is the impact of degree with in each industry. The model developed in this analysis showed a strong correlation between degree and increased salary and further exploration of this relationship would be interesting.

R code appendix

```
knitr::opts_chunk$set(include=FALSE)
library(dplyr)
library(ggplot2)
library(ggthemes)
library(gridExtra)
library(tidyr)
library(stringr)
#library(kableExtra)
# Read in data
project1 <- read.csv("statSubset_Categories_V1.2.csv")

# Select columns of interest
project1<-project1%>%select(DIVISION,SPORDER, PUMA, REGION, ST, ADJINC, PWGTP, AGEP, CIT, CITWP, COW, E

# Filter to people currently employed and at work at time of survey
target <- c("Armed forces  at work", "Civilian employed  at work")
project1_Mod <- filter(project1, ESR %in% target)

# Substring the Industry field to get the larger industry categories
project1_Mod$IND_Category = str_sub(project1_Mod$NAICSP,1,3)

# Filter to Large industry categories
target2 <- c("FIN","EDU","MFG","MED")
project1_Mod <- filter(project1_Mod, IND_Category %in% target2)

# Expand the table based on weights
project1_Mod <- project1_Mod %>% uncount(PWGTP)

# Adjust income fields to constant dollars by applying Income Adjustment factor
project1_Mod$WAGP_Adj <- as.numeric(project1_Mod$WAGP) * as.numeric(project1_Mod$ADJINC) / 1000000
project1_Mod$PINCP_Adj <- as.numeric(project1_Mod$PINCP) * as.numeric(project1_Mod$ADJINC) / 1000000
```

```r
# Log transform of income fields to reduce scale
project1_Mod$WAGP_Adj_log <- log(project1_Mod$WAGP_Adj)
project1_Mod$PINCP_Adj_log <- log(project1_Mod$PINCP_Adj)

# Reduce data set to final columns to be used in analysis
project1_red<-project1_Mod%>%select(REGION, AGEP, COW, SCHL, SEX, WAGP_Adj, WKHP, WKW, HICOV, MSP, NATI

# project1_red

# Modify column names
colnames(project1)<-c("Division code based on 2010 Census definitions","Person number","Public use micr

project1.int = as.data.frame(project1[,1])
project1.chr = as.data.frame(project1[,1])
for (i in 1:ncol(project1)){
  if (!is.numeric(project1[,i])){
      project1.chr<-cbind(project1.chr,project1[,i])
      colnames(project1.chr)[ncol(project1.chr)]<-colnames(project1)[i]
  }
  else{
      project1.int<-cbind(project1.int,project1[,i])
      colnames(project1.int)[ncol(project1.int)]<-colnames(project1[i])
  }
}
project1.int <- subset (project1.int, select = -c(1))
project1.chr <- subset (project1.chr, select = -c(1))
project1_red.na<-data.frame()
#NA values
for (i in 1:ncol(project1_red)){
    if (sum(is.na(project1_red[,i]))>0){
    project1_red.na<-rbind(project1_red.na,c(names(project1_red)[i],c(sum(is.na(project1_red[,i]))))))}
}
ggplot(project1_red, aes(IND_Category, PINCP_Adj, fill = IND_Category)) +
  geom_boxplot()+theme_tufte()+theme(text = element_text(size=rel(3)),legend.position = "none")+xlab("I
textSize <- 3.5

#summary for some of the numeric columns

#Age
p1<-ggplot(project1.int, aes(x=project1.int[,5])) + geom_bar(width =0.1)+xlab(colnames(project1.int[5]))

#Hours of wrok
p2<-ggplot(project1.int, aes(x=project1.int[,13])) + geom_histogram()+xlab(colnames(project1.int[13]))+

#total earnings
#p3<-ggplot(project1.int, aes(x=project1.int[,15])) + geom_histogram()+xlab(colnames(project1.int[15]))

#summary for some of non-numeric columns:

# Region
p4<-ggplot(as.data.frame(table(project1.chr[,2])), aes(Var1, Freq, fill =

# Sex
```

```
p5<-ggplot(as.data.frame(table(project1.chr[,49])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnar

# Marital Status
p6<-ggplot(as.data.frame(table(project1.chr[,12])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnar

# Educational Attainment
p7<-ggplot(as.data.frame(table(project1.chr[,25])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnar

# Employment Status
p8<-ggplot(as.data.frame(table(project1.chr[,50])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnar

grid.arrange(p1, p2, nrow=1)
grid.arrange(p4, p5, p6, p7, nrow = 2)
library(plyr)
detach("package:plyr")
library(dplyr)
project1_red<-project1_red%>%group_by(REGION, IND_Category)%>%mutate(Avg_Income=mean(PINCP_Adj))
project1_red<-project1_red%>%mutate(n_ind_region=n())%>%mutate(IND_Category = recode(IND_Category, "EDU
ggplot(project1_red,aes(IND_Category, Avg_Income,group=REGION, color=REGION, size=n_ind_region, alpha=0
#
lm.inference<-lm(PINCP_Adj ~ REGION + AGEP + COW + SCHL + SEX + WKHP + WKW + HICOV + MSP + NATIVITY + PA
summary(lm.inference)
#lm.inference<-lm(PINCP_Adj ~ IND_Category + REGION + AGEP + COW + SCHL + SEX +
#+WKHP + WKW + HICOV + MSP + NATIVITY + PAOC + POBP, data=project1_red)
data.additional<- fortify(lm.inference, data = project1_red)
p1<-ggplot(data.additional, aes(x=.resid)) + geom_histogram()+xlab("Residuals")+ylab("Count")+theme_tuf
p2<-ggplot(data.additional, aes(y=.resid, x=.fitted))+ geom_point()+xlab("Fitted Values")+ylab("Residual
p3<-ggplot(data.additional, aes(y=.fitted, x=WAGP_Adj))+ geom_point()+xlab("Observed values")+ylab("Fit
p4<-ggplot(data.additional, aes(y=.resid, x=WAGP_Adj))+ geom_point()+xlab("Observed values")+ylab("Resi
grid.arrange(p1,p2,p3,p4, nrow=3, top = 'Checking Linear Regression Assumptions for the Original Functi
lm.inference.log<-lm(PINCP_Adj_log~REGION+AGEP+COW+SCHL+SEX+WKHP+WKW+HICOV+MSP+NATIVITY+PAOC+POBP, data=
data.additional<- fortify(lm.inference.log, data = project1_red)
p1<-ggplot(data.additional, aes(x=.resid)) + geom_histogram()+xlab("Residuals")+ylab("Count")+theme_tuf
p2<-ggplot(data.additional, aes(y=.resid, x=.fitted))+ geom_point()+xlab("Fitted Values")+ylab("Residual
p3<-ggplot(data.additional, aes(y=.fitted, x=PINCP_Adj_log))+ geom_point()+xlab("Observed values")+ylab
p4<-ggplot(data.additional, aes(y=.resid, x=PINCP_Adj_log))+ geom_point()+xlab("Observed values")+ylab(
grid.arrange(p1,p2,p3,p4, nrow=3, top = 'Checking Linear Regression Assumptions for the Original Functi
lm.inference.log<-lm(PINCP_Adj_log~IND_Category+REGION+AGEP+COW+SCHL+SEX+WKHP+WKW+HICOV+MSP+NATIVITY+PA
summary(lm.inference.log)

ci<-confint(lm.inference)
conf.int<-as.data.frame(ci[2:4,])
conf.int<-cbind(conf.int,lm.inference$coefficients[2:4])
colnames(conf.int)<-c("2.5%", "97.5%", "Mean Difference")
rownames(conf.int)<-c("Industry: Finance", "Industry: Medicine", "Industry: Manufacturing" )
log.ci.convert<-function(n){(exp(n)-1) * 100}
ci<-confint(lm.inference.log)
conf.int<-log.ci.convert(as.data.frame(ci[2:4,]))
conf.int<-cbind(conf.int,log.ci.convert(lm.inference.log$coefficients[2:4]))
colnames(conf.int)<-c("2.5%", "97.5%", "Mean Difference")
rownames(conf.int)<-c("Industry: Finance", "Industry: Medicine", "Industry: Manufacturing" )
(conf.int)
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=I(60)), tidy=TRUE)
```