Income and Education

Muiz Rahemtullah

8/5/2019

First we will set our working directory.

```
setwd("~/Desktop/Data Science Summer 19/Income and Education")
```

Now we will load the tidycensus library and enter our key.

```
library(tidycensus)
census_api_key("ab935c7b673ec9c62b7c09dc516dd36f0d8fd03a", install = TRUE, overwrite = TRUE)
```

```
## Your original .Renviron will be backed up and stored in your R HOME directory if needed.
```

```
## Your API key has been stored in your .Renviron and can be accessed by Sys.getenv("CENSUS_API_KEY").
## To use now, restart R or run `readRenviron("~/.Renviron")`
```

```
## [1] "ab935c7b673ec9c62b7c09dc516dd36f0d8fd03a"
```

```
readRenviron("~/.Renviron")
Sys.getenv("CENSUS_API_KEY")
```

```
## [1] "ab935c7b673ec9c62b7c09dc516dd36f0d8fd03a"
```

Now we will load the data in. We want the number of residents that have a bachelors degree by county in the US. We will also create a CSV of this data

```
census_data <- get_acs(geography = "county", variables = "B15003_022")
```

```
## Getting data from the 2013-2017 5-year ACS
```

```
write.csv(census_data, "Bachelors Degree Data.csv")
```

We will also obtain the total adult population by county.

```
census_data_2 <- get_acs(geography = "county", variables = "B15003_001")
```

```
## Getting data from the 2013-2017 5-year ACS
```

```
write.csv(census_data_2, "Adult Population Data.csv")
```

Finally, we will obtain the median household income by county.

```
census_data_3 <- get_acs(geography = "county", variables = "B19013_001")
```

```
## Getting data from the 2013-2017 5-year ACS
```

```
write.csv(census_data_3, "Median Income Data.csv")
```

Now we will merge the latter 2

```
census_data_2 <- merge(census_data_2, census_data_3, by = "GEOID", all.x = TRUE)
```

Now we will clean this dataset up.

```
census_data_2 <- census_data_2[, -c(3, 6, 7)]
colnames(census_data_2) <- c("GEOID", "County", "Adult Population Estimate", "Adult Population 90% Confi
```

Now we will merge this with the first dataset.

```
census_data <- merge(census_data, census_data_2, by = "GEOID", all.x = TRUE)
```

Now we will clean this dataset up.

```
census_data <- census_data[, -c(3, 6)]
colnames(census_data) <- c("GEOID", "County", "Adult Population with Bachlors Degree Estimate", "Adult
```

Now we will create a new column documenting the percentage of graduates among the entire population by county. This is the final dataset we will be working with so we will also turn this into a CSV file.

```
attach(census_data)
census_data$PCT_Degree <- `Adult Population with Bachlors Degree Estimate`/`Adult Population Estimate`
attach(census_data)
```

```
## The following objects are masked from census_data (pos = 3):
##
##     Adult Population 90% Confidence Interval, Adult Population
##     Estimate, Adult Population with Bachlors Degree 90% Confidence
##     Interval, Adult Population with Bachlors Degree Estimate,
##     County, GEOID, Median Household Income 90% Confidence
##     Interval, Median Household Income Estimate
```

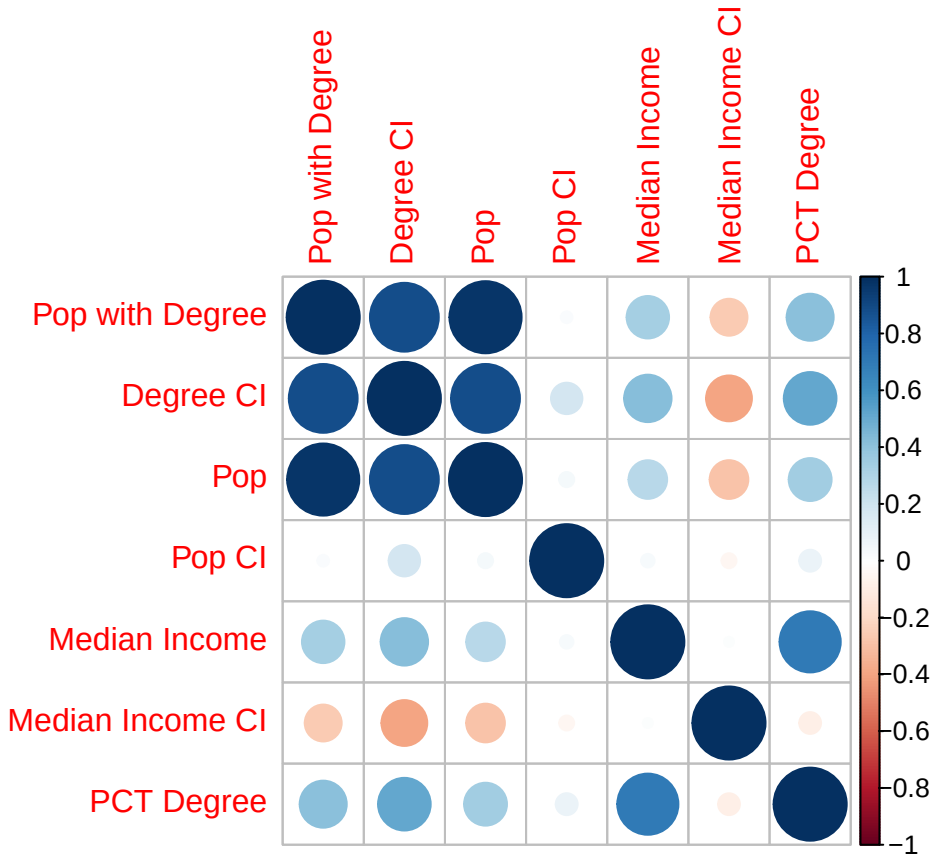
```
write.csv(census_data, "Final Dataset.csv")
```

Now we will plot a correlation graph of degree percentage and median household income. We must omit the ID and County columns and reduce the column names so we will quickly create a new dataset for this.

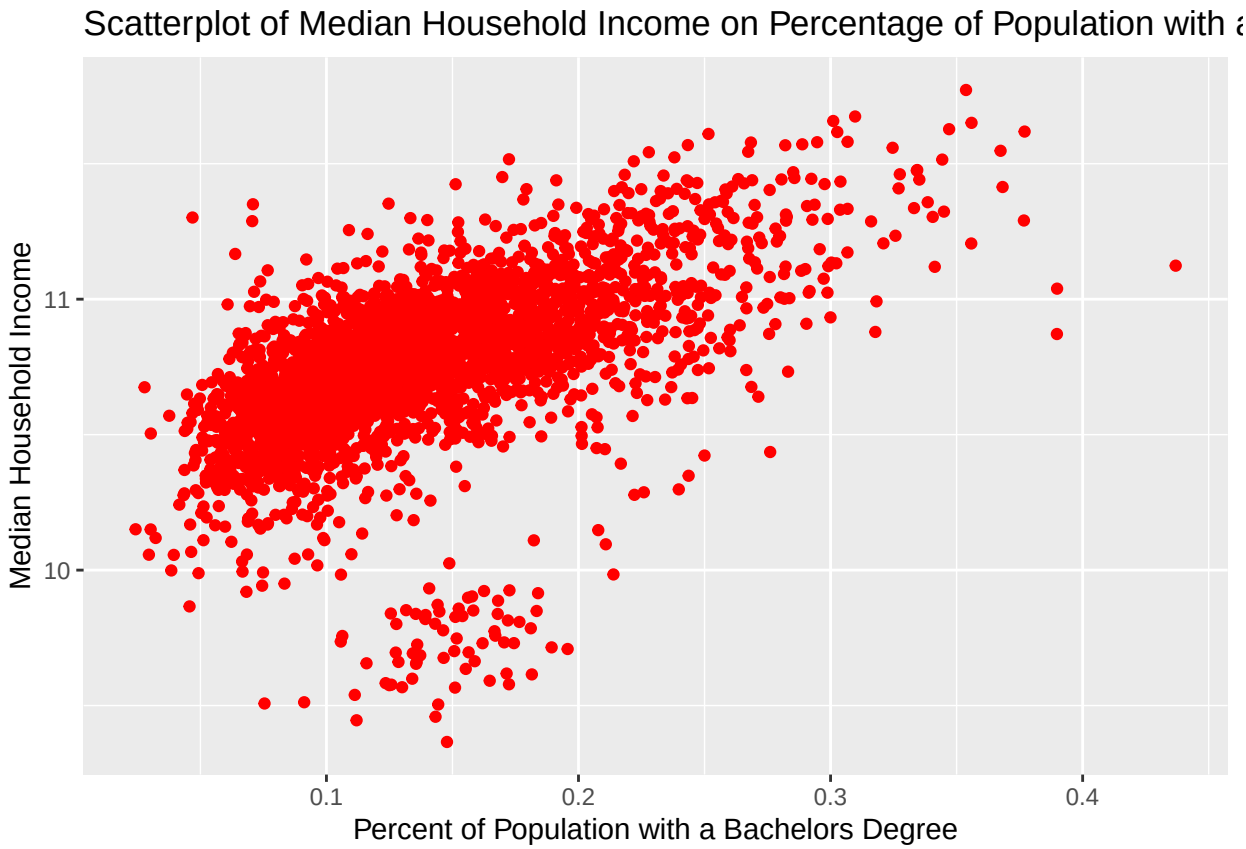
```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(ggplot2)
num_census <- census_data[,-c(1, 2)]
colnames(num_census) <- c("Pop with Degree", "Degree CI", "Pop", "Pop CI", "Median Income", "Median Inc
corrplot(cor(num_census, use="complete.obs"))
```




```
ggplot(census_data, aes(x=PCT_Degree, y=log(`Median Household Income Estimate`))) + geom_point(col = "r
```

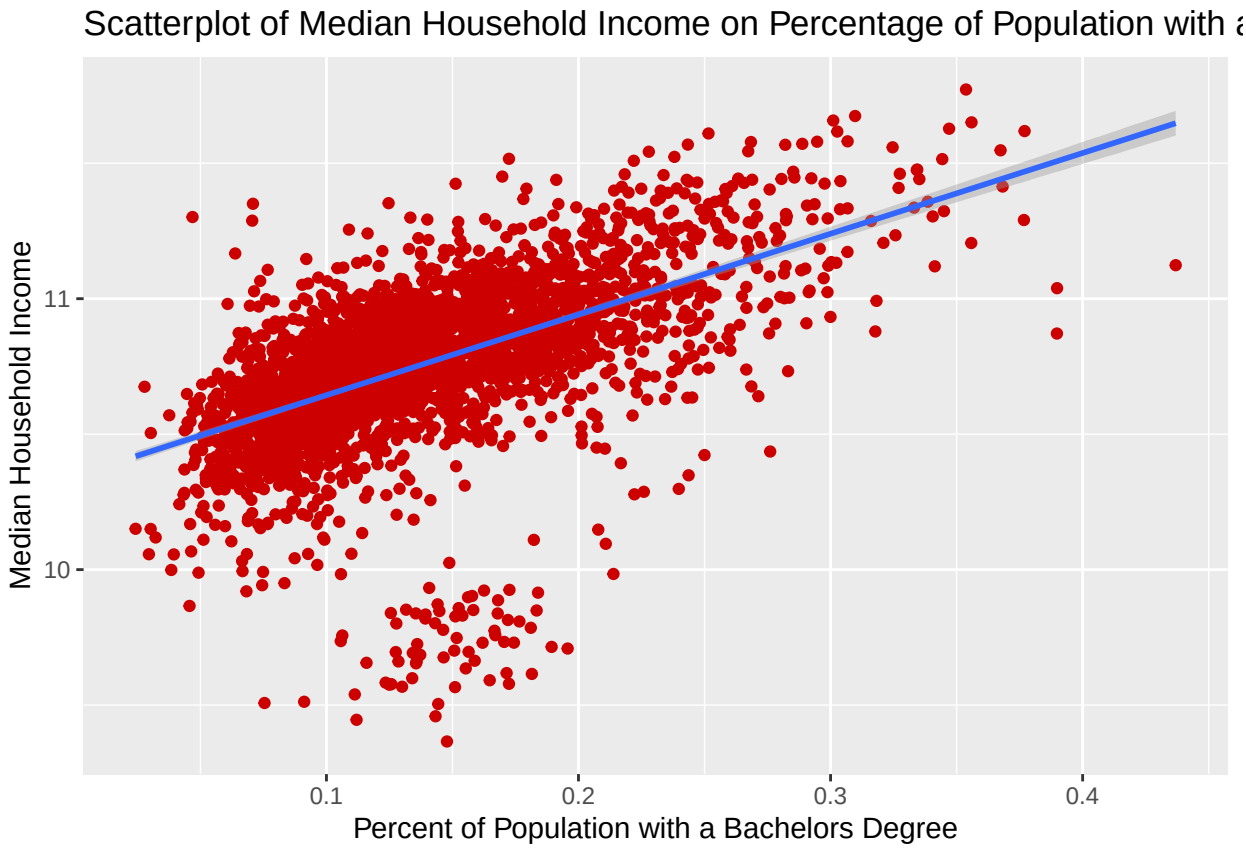


We see there is a roughly .7 correlation between Educated Population and Median household income. We will now run a regression between the two.

```
edu_inc_fit <- lm(log(`Median Household Income Estimate`)~PCT_Degree, data = census_data)
summary(edu_inc_fit)
```

```
##
## Call:
## lm(formula = log(`Median Household Income Estimate`) ~ PCT_Degree,
##     data = census_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42074 -0.09308  0.02692  0.14184  0.81584
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.34593    0.01145  903.58   <2e-16 ***
## PCT_Degree   2.97952    0.07670   38.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2416 on 3218 degrees of freedom
## Multiple R-squared:  0.3192, Adjusted R-squared:  0.319
## F-statistic:  1509 on 1 and 3218 DF,  p-value: < 2.2e-16
```

```
ggplot(census_data, aes(x=PCT_Degree, y=log(`Median Household Income Estimate`))) + geom_point(col = "r
```



We will now interpret the regression. If the county's college degree residents increase by 10 percent, the median income will increase by $10 * 2.97952 = 29.7952$ percent.

We see that there is a patch on the bottom that is stagnant and that as the percentage of people with degrees increases from .1 to .2 yet the Median Household Income remains unchanged. This is likely the outlier region mentioned. In these counties, as the percent of educated people increases, the median household income does not increase. We will highlight this region be coloring it differently.

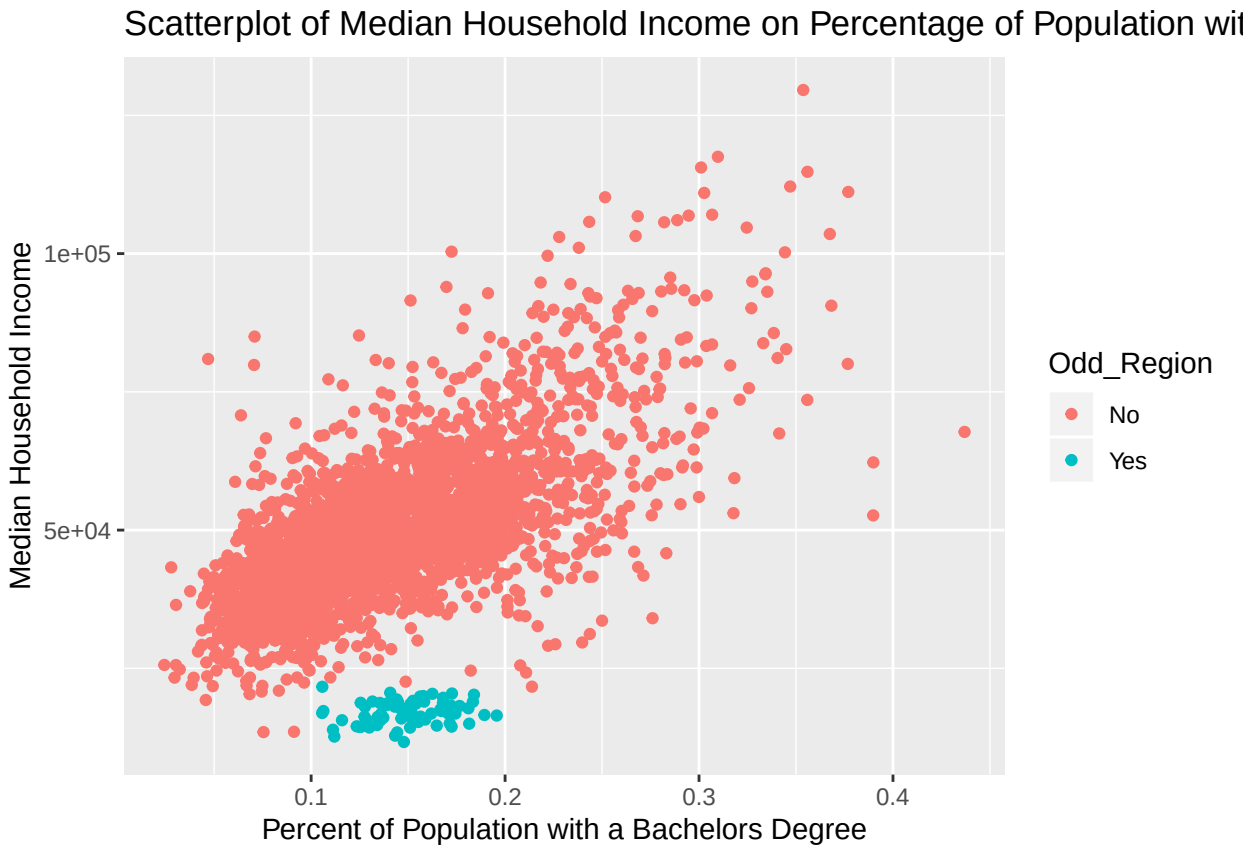
```
library(dplyr)
```

```
##
## 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
```

```
census_data %>%
  mutate(Odd_Region = ifelse(log(`Median Household Income Estimate`) < 10 & PCT_Degree > .1 & PCT_Degree
  ggplot(census_data, mapping = aes(x=PCT_Degree, y=`Median Household Income Estimate`, color = Odd_Reg
```



When we examine the dataset in detail, we can see that the Blue Region represents counties in Puerto Rico. This conclusion makes sense since Puerto Rico has suffered from natural disasters such as hurricanes recently. With their infrastructure compromised, it would be difficult to find a higher paying jobs there.