FINAL PROJECT

Hung Khuu

I. Data preparation

“How one’s incarnation affects his/her earning ability?”
“Does how one feel at school have an impact on furure income?”
“Are you destined to make more money than other people if you are born on a particular month or year?”
“What impact does the degree earned brings to the income gap between men and women?”
“How marrital status affects one’s income?”

And finally,

“Does your SAT performance or the number of jobs you held increase your income potential?”

These are the questions that immediately comes to my mind when I was examining the data descriptions. They acted as a guideline for me to pick my variable of interest, which are:

- Total number of incarceration `ttl.incarc`
- Age at first incarceration `age.first.incarc`
- Sentiment toward school `school.sentiment`
- Birth month and year `brth.mth` & `brth.yr`
- Whether the surveyee has a special physical/emotional condition `special.needs`
- Highest degree earned `degree.earned`
- Marrital status `marrital.stat`
- SAT math and verbal scores `SAT.math` & `SAT.verbal`
- Number of jobs worked as adult `adlthood.numjobs`

And the three main variables in the data set:

- Gender `GENDER`
- Race `RACE`
- Income `INCOME` , the dependent variable

```
##
## Attaching package: 'reshape'
```

```
## The following objects are masked from 'package:plyr':
##
##     rename, round_any
```

```
## Loading required package: lattice
```

There are 5302 non-missing `INCOME` observations. We will only work on cases that have income information, as it is the dependent variable that we are trying to describe with our model.

Dealing with missing value. The data cleaning task continue with the missing values marked as -1, -2, -3, -4, and -5 in the data set. From the Bureau of Labor Statistics site, we know that missing values coded as -3, -4, -5 represent either the question is irrelevant in the surveyee case, or it was given to the wrong target, thus being removed by the surveyer (-3 invalid skip). The only potentialy meaningful missing value is -1, meaning the surveyees refused to answer a particular question, for one reason or another. That is the reason why I decided to recode all of the -1 missing values to `no answer` . All other values of missing value will be recoded as `NA` . As the unique values are shown below, only `school.sentiment` and `special.needs` have -1 missing values.

```
##  [1] -4 23 18 19 24 26 16 29 20 22 27 21 25 28 15 12 14 17 30 13 11
```

```
## [1] 0 1 2 3 4 5 6 7 9
```

```
## [1]  1  2  3  4 -4 -2 -1
```

```
## [1] 2 1
```

```
##  [1]  9  7  2 10  4  6  1 11 12  5  3  8
```

```
## [1] 1981 1982 1983 1984 1980
```

```
## [1] 4 2 1 3
```

```
## [1]  4  2  1  5  3 -3  7  0  6
```

```
## [1]  0 -4  1 -2 -1
```

```
## [1]  0  1  2  4  3 -3
```

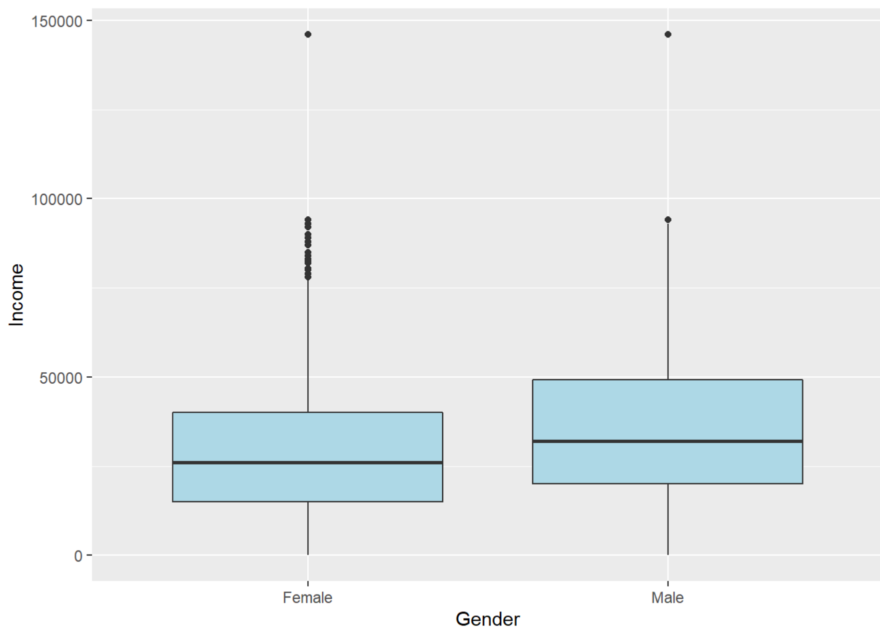
```
## [1]  4 -4  2 -3  5  3  6  1
```

```
## [1]  3  4 -4  6 -3  5  2  1
```

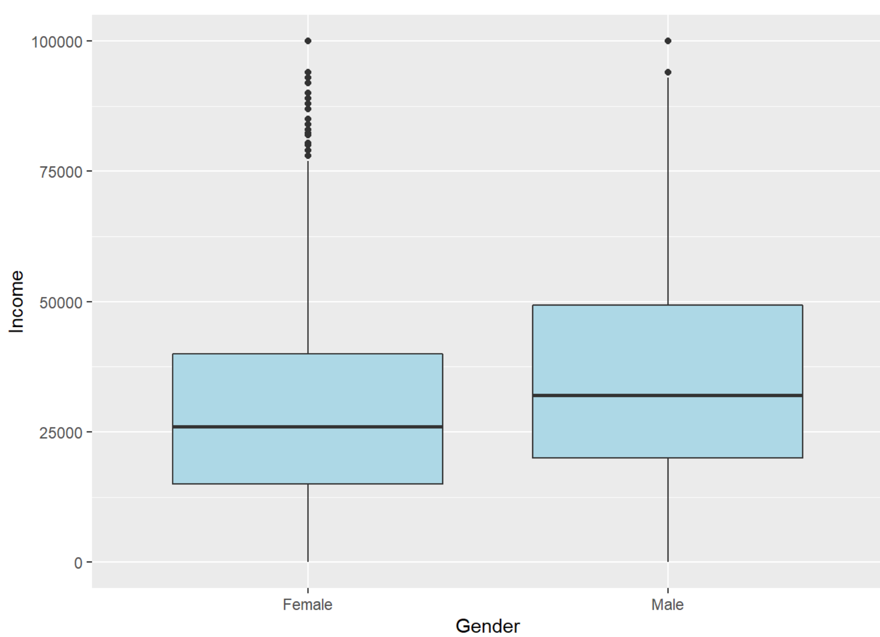
```
##  [1]  4  5  7  6  2  8  3 10 12  1  0 -3 11 36 18 13  9 14 16 20 15 17 22 19 21
## [26] 23 27 25 31
```

```
## The following `from` values were not present in `x`: 0
## The following `from` values were not present in `x`: 0
```

Dealing with topcoded `INCOME`. From the description, we know that the top earners in the data was topcoded by the mean of their income, with the value of $ 146002. Below is the plot of `INCOME` by `GENDER` for the untreated data:



The topcoding caused the data set to have many big outliers. After looking into different methods to solve the problem, including trying to recode the value with random normal distributed value around the mean, I decided to recoded all topcoded values to 100000. Doing that enable me to 1) keep 120 observations with topcoded values and improve the stability of the model, and 2) reduce the impact of outliers on the model. Let's have a look at the income data after the treatment:



After the processes above, the data set of interest is now ready to be examined further

II. Data exploration

First, we will have a general look at the structure as well as a summary of our data

```
## 'data.frame':	5302 obs. of  14 variables:
##  $ ttl.incarc      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age.first.incarc: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ school.sentiment: Factor w/ 5 levels "no answer","safe",..: 4 2 3 2 2 4 4 2 2 2 ...
##  $ GENDER          : Factor w/ 2 levels "Female","Male": 1 2 1 2 2 1 2 1 2 1 ...
##  $ brth.mth        : Factor w/ 12 levels "APR","AUG","DEC",..: 12 6 4 11 1 7 11 7 11 5 ...
##  $ brth.yr         : int  1981 1982 1981 1982 1983 1981 1982 1982 1981 1983 ...
##  $ special.needs   : Factor w/ 3 levels "no","no answer",..: 1 NA 3 1 1 1 1 1 1 NA ...
##  $ RACE            : Factor w/ 4 levels "BLACK","HISPANIC",..: 4 2 2 2 2 4 4 2 2 2 ...
##  $ degree.earned   : Factor w/ 8 levels "ASSOCIATE","BACHELOR",..: 2 4 4 4 3 5 5 4 3 4 ...
##  $ marrital.stat   : Factor w/ 5 levels "divorced","married",..: 3 3 3 2 2 3 3 3 3 3 ...
##  $ INCOME          : num  50000 81000 51000 68000 0 65000 30000 17000 68000 12000 ...
##  $ SAT.math        : Factor w/ 6 levels "200-300","301-400",..: 4 4 NA 2 NA 5 4 NA NA NA ...
##  $ SAT.verbal      : Factor w/ 6 levels "200-300","301-400",..: 3 4 NA 6 NA 5 2 NA NA NA ...
##  $ adlthood.numjobs: int  4 5 7 5 6 6 6 2 8 5 ...
```

ttl.incarc	age.first.incarc	school.sentiment	GENDER	brth.mth	brth.yr	special.needs	RACE	degree.earned	marrital.stat
Min. :0.0000	Min. :11.00	no answer : 1	Female:2511	SEP : 504	Min. :1980	no :4435	BLACK :1192	HS.DIPLOMA:2329	divorced : 333
1st Qu.:0.0000	1st Qu.:19.00	safe :2889	Male :2791	AUG : 463	1st Qu.:1981	no answer: 1	HISPANIC:1148	BACHELOR :1278	married :2055
Median :0.0000	Median :21.00	unsafe : 544	NA	JAN : 463	Median :1982	yes : 286	MIXED : 51	GED : 520	never.married
Mean :0.1113	Mean :21.81	very.safe :1725	NA	OCT : 455	Mean :1982	NA's : 580	OTHERS :2911	ASSOCIATE : 418	separeated : 7
3rd Qu.:0.0000	3rd Qu.:24.00	very.unsafe: 135	NA	MAR : 446	3rd Qu.:1983	NA	NA	NONE : 347	widowed : 7
Max. :9.0000	Max. :30.00	NA's : 8	NA	JUL : 443	Max. :1984	NA	NA	(Other) : 373	NA's : 10
NA	NA's :4981	NA	NA	(Other):2528	NA	NA	NA	NA's : 37	NA

Diving a little deeper, we will examine the relationships between variables through the spectrum of `GENDER`, `RACE`, and `INCOME`. The table below shows **the average age of first incarceration among different groups of total number of incarceration.**

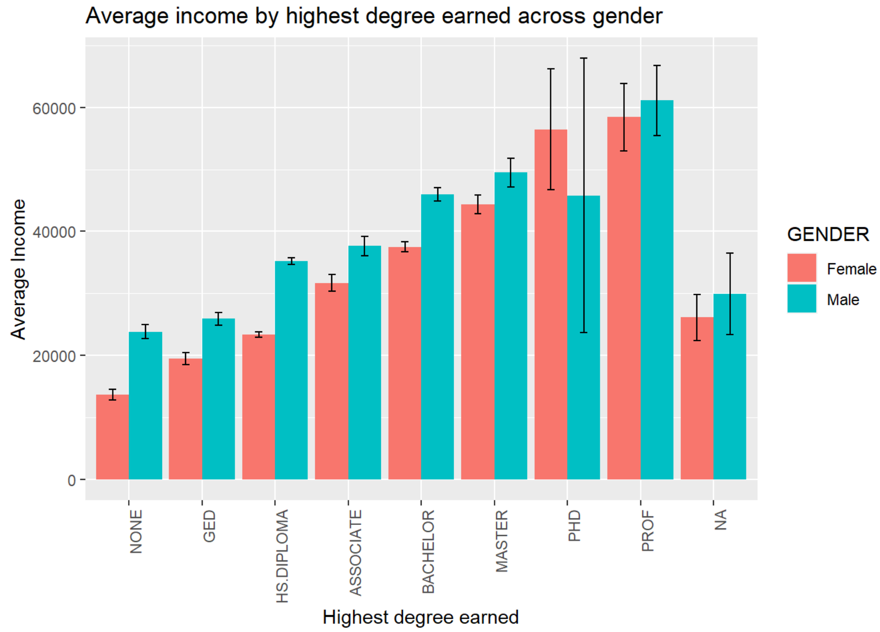
	Female	Male
0	NaN	NaN
1	23.146	22.662
2	19.500	21.028
3	21.444	20.821
4	19.000	19.273
5	NA	16.750
6	17.000	19.667
7	NA	24.000
9	NA	20.000

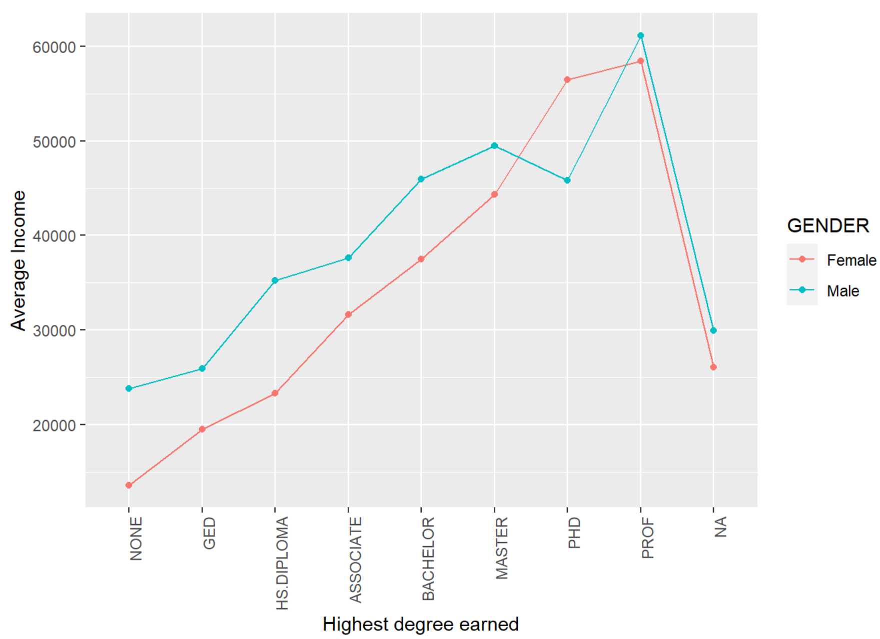
It is interesting to see that in the groups of 2 and 6 incarcerations, the average age of first incarceration of female is lower than that of male. One explanation for this can be the violation that led to the incarceration is minor, and the term of incarceration is shorter for women than men.

Next, we look at the relationship between the highest degree earned and income across gender.

	Female	Male
ASSOCIATE	31666.32	37646.69
BACHELOR	37500.50	45970.41
GED	19509.62	25929.18
HS.DIPLOMA	23343.63	35216.61
MASTER	44348.72	49481.71

	Female	Male
NONE	13626.59	23835.63
PHD	56444.44	45808.33
PROF	58421.21	61116.46





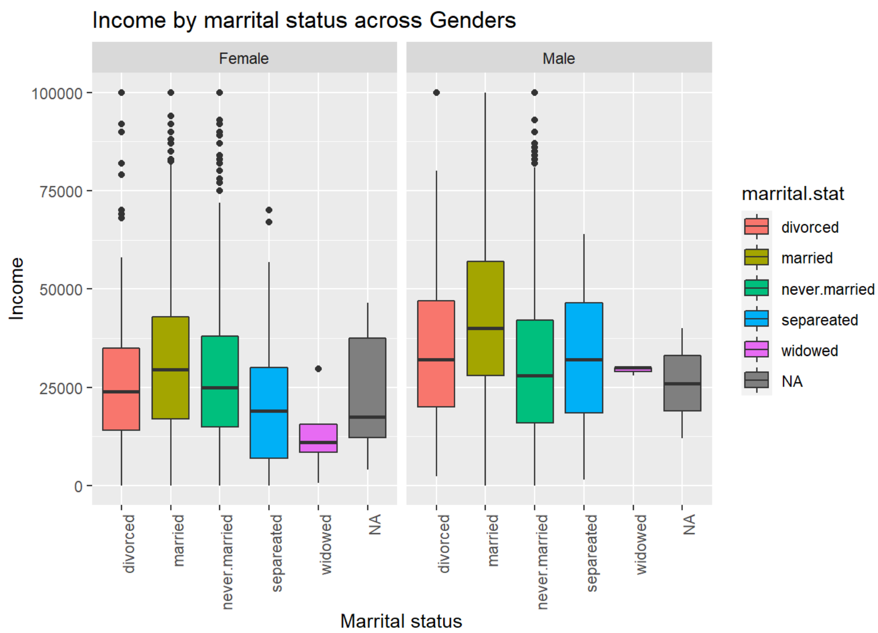
Looking at the table, we can easily observe that apart from Phd, regardless of the highest degree you have, men always make more money than women, even though the gap shrinks by the small amount when moving to higher degree.

The result in the bar chart is consistent with the law of return: the more effort you put in your education, the higher your income potential will be. With that being said, there's a larger variance in income for PhD and Professional degree holder. It is also interesting to see that the missing values seems to make more money annually compare to those without any degrees.

Below we will continue to look at how marrital status interact with other variables

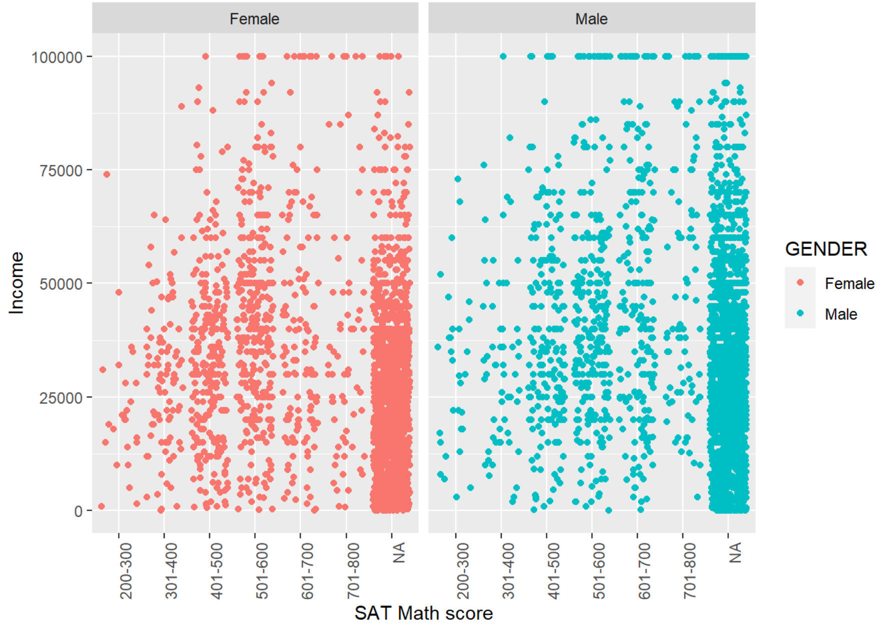
	BLACK	HISPANIC	MIXED	OTHERS
divorced	4.54	6.73	7.84	6.81
married	24.54	38.81	35.29	44.75
never.married	69.50	52.10	52.94	47.27
separeated	1.18	2.19	3.92	1.10
widowed	0.25	0.17	0.00	0.07

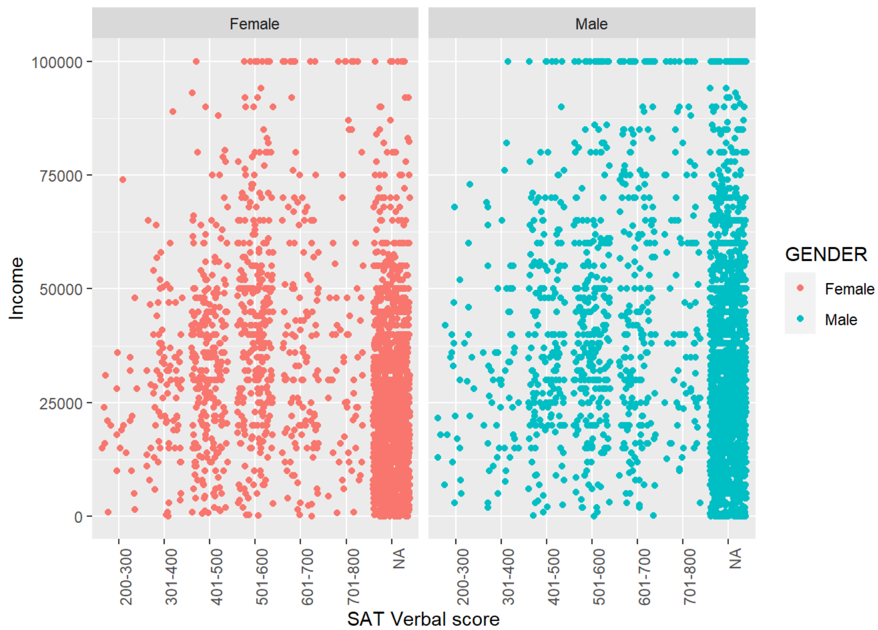
The table above shows the percentage of each marrital status by race groups. People belongs to race groups other than black, hispanic or mixed has the highest percentage of married status, but they also has the second highest number of divorced percentage. The group with the highest divorced percentage, and also seperated percentage, is non-hispanic mixed. The following graph will reveal how marrital status can affect income potential



It turned out being married can improve your income potential, while a divorce can make you earn less annually. Another noticeable trend is that men of all marrital status earn more than women, another evidence of the income inequality among the gender line. Interestingly, men who are never married earn approximately the same as their female counterpart.

Lastly, we look at how **SAT scores** correlate with income

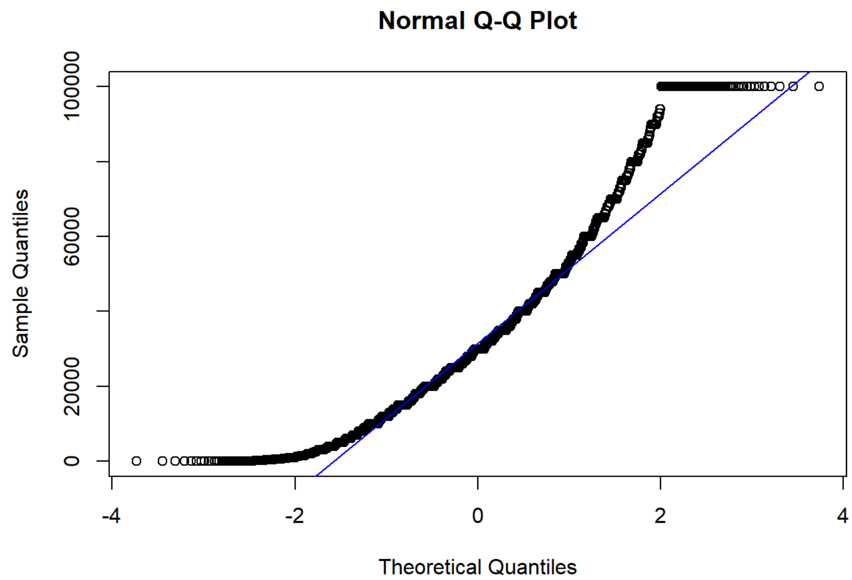




There is no clear trend in both graphs to support that the higher SAT scores might give a hint on how much one will be able to earn later on in life. However, people in the lowest group of SAT score is much more unlikely to be able to have an income above $60000.

III. Building the model

Before setting out to find a model that can best describe the data set, we need to check the normality of the dependent variable



Our `INCOME` data seems to have a right skew. This can be explain by the relative large number of topcoded values, even though treatment has been applied to reduce this effect.

Next, a general linear regression model will be run, from which each variable will be taken out to answer the questions stated at the beginning

General Model

```
##
## Call:
## lm(formula = INCOME ~ ., data = nlsy.1)
##
## Residuals:
## ALL 30 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (10 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -3878501.9         NA      NA       NA
## ttl.incarc                     17615.4         NA      NA       NA
## age.first.incarc                3788.5         NA      NA       NA
## school.sentimentunsafe        -71557.7         NA      NA       NA
## school.sentimentvery.safe     -25230.8         NA      NA       NA
## school.sentimentvery.unsafe   -83917.3         NA      NA       NA
## GENDERMale                     -1740.4         NA      NA       NA
## brth.mthAUG                   -64584.6         NA      NA       NA
## brth.mthDEC                    19942.3         NA      NA       NA
## brth.mthFEB                    29048.1         NA      NA       NA
## brth.mthJAN                   -53711.5         NA      NA       NA
## brth.mthJUL                   -84632.7         NA      NA       NA
## brth.mthJUN                   -91430.8         NA      NA       NA
## brth.mthMAR                     -394.2         NA      NA       NA
## brth.mthMAY                   -66790.4         NA      NA       NA
## brth.mthNOV                   -12980.8         NA      NA       NA
## brth.mthOCT                   -12432.7         NA      NA       NA
## brth.mthSEP                   -40144.2         NA      NA       NA
## brth.yr                         1951.9         NA      NA       NA
## special.needsyes               27461.5         NA      NA       NA
## RACEHISPANIC                   20634.6         NA      NA       NA
## RACEMIXED                       2884.6         NA      NA       NA
## RACEOTHERS                     45076.9         NA      NA       NA
## degree.earnedBACHELOR        -120036.5         NA      NA       NA
## degree.earnedGED               -9240.4         NA      NA       NA
## degree.earnedHS.DIPLOMA       -80950.0         NA      NA       NA
## degree.earnedNONE             -94900.0         NA      NA       NA
## marrital.statmarried           66201.9         NA      NA       NA
## marrital.statnever.married          NA         NA      NA       NA
## SAT.math301-400                76219.2         NA      NA       NA
## SAT.math401-500                26442.3         NA      NA       NA
## SAT.math501-600                     NA         NA      NA       NA
## SAT.math601-700                     NA         NA      NA       NA
## SAT.math701-800                     NA         NA      NA       NA
## SAT.verbal301-400                   NA         NA      NA       NA
## SAT.verbal401-500                   NA         NA      NA       NA
## SAT.verbal501-600                   NA         NA      NA       NA
## SAT.verbal601-700                   NA         NA      NA       NA
## SAT.verbal701-800                   NA         NA      NA       NA
## adlthood.numjobs                    NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
##   (5272 observations deleted due to missingness)
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 29 and 0 DF,  p-value: NA
```

This general model resulted in most of values being `NA` 's. Therefore, our next step will try to troubleshoot the model to see what is causing the problem

Model 1: No SAT score

```
##
## Call:
## lm(formula = INCOME ~ ttl.incarc + age.first.incarc + school.sentiment +
##     GENDER + brth.mth + brth.yr + special.needs + RACE + degree.earned +
##     marrital.stat + adlthood.numjobs, data = nlsy.1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29908  -8668  -1754   6735  47658
##
## Coefficients:
##                              Estimate Std. Error t value    Pr(>|t|)
## (Intercept)                 1772969.9  1360949.5   1.303     0.19404
## ttl.incarc                     -802.0      964.8  -0.831     0.40671
## age.first.incarc               -447.2      286.4  -1.562     0.11984
## school.sentimentunsafe         2613.6     3036.5   0.861     0.39035
## school.sentimentvery.safe     -1890.8     2213.5  -0.854     0.39392
## school.sentimentvery.unsafe    1267.8     5304.8   0.239     0.81133
## GENDERMale                     8152.4     2546.5   3.201     0.00157 **
## brth.mthAUG                    6071.3     4693.7   1.293     0.19721
## brth.mthDEC                   -2801.8     5176.2  -0.541     0.58887
## brth.mthFEB                    4607.2     5102.1   0.903     0.36752
## brth.mthJAN                    4623.3     5107.4   0.905     0.36636
## brth.mthJUL                   -6894.5     5165.2  -1.335     0.18334
## brth.mthJUN                   -4837.5     5276.2  -0.917     0.36024
## brth.mthMAR                    3766.3     5006.3   0.752     0.45268
## brth.mthMAY                   -3398.2     5130.9  -0.662     0.50847
## brth.mthNOV                    1562.2     5449.1   0.287     0.77462
## brth.mthOCT                   -1925.9     4866.6  -0.396     0.69269
## brth.mthSEP                     525.8     4870.8   0.108     0.91413
## brth.yr                        -879.2      687.1  -1.280     0.20205
## special.needsyes              -8217.2     3184.7  -2.580     0.01053 *
## RACEHISPANIC                  12162.2     2670.7   4.554 0.000008773 ***
## RACEMIXED                      9453.6     8768.9   1.078     0.28220
## RACEOTHERS                    12611.8     2362.4   5.339 0.000000236 ***
## degree.earnedBACHELOR         12045.2     6885.0   1.749     0.08162 .
## degree.earnedGED              -4119.5     5519.7  -0.746     0.45628
## degree.earnedHS.DIPLOMA       -1482.2     5428.6  -0.273     0.78509
## degree.earnedMASTER             972.7    15761.7   0.062     0.95085
## degree.earnedNONE             -7372.7     5739.7  -1.285     0.20034
## marrital.statmarried          -1938.2     4905.2  -0.395     0.69313
## marrital.statnever.married    -6646.0     4671.3  -1.423     0.15626
## marrital.statsepareated      -10785.4     6675.8  -1.616     0.10763
## adlthood.numjobs               -576.5      221.3  -2.606     0.00980 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14090 on 217 degrees of freedom
##   (5053 observations deleted due to missingness)
## Multiple R-squared:  0.3886, Adjusted R-squared:  0.3013
## F-statistic: 4.449 on 31 and 217 DF,  p-value: 2.412e-11
```

As soon as the SAT scores were removed, the model works again. This might be because of the large number of `NA` within these variable. `SAT.math` has 3604 missing values, whereas the berbal score has 3604 missing data points.

Additionally, the summary report showed that birth month and year does NOT have any significance in determining one's `INCOME` . This helped debunked the myth about a certain birth month will make a person more properous than another, which has been popular in Asian countries.

Model 2: No SAT score & Birth date

```
##
## Call:
## lm(formula = INCOME ~ ttl.incarc + age.first.incarc + school.sentiment +
##     GENDER + special.needs + RACE + degree.earned + marrital.stat +
##     adlthood.numjobs, data = nlsy.1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25733  -8895  -2839   6463  50186
##
## Coefficients:
##                             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                  31864.2    11038.9   2.887    0.00427 **
## ttl.incarc                   -1023.8      961.6  -1.065    0.28815
## age.first.incarc              -498.5      286.6  -1.739    0.08332 .
## school.sentimentunsafe        2575.2     3010.9   0.855    0.39327
## school.sentimentvery.safe    -2421.7     2184.6  -1.109    0.26880
## school.sentimentvery.unsafe   2550.5     5150.7   0.495    0.62096
## GENDERMale                    8594.8     2526.5   3.402    0.00079 ***
## special.needsyes             -7476.2     3170.2  -2.358    0.01920 *
## RACEHISPANIC                 12152.1     2701.3   4.499 0.00001088 ***
## RACEMIXED                    10636.0     8726.4   1.219    0.22417
## RACEOTHERS                   11396.1     2352.7   4.844 0.00000235 ***
## degree.earnedBACHELOR        15113.2     6758.9   2.236    0.02631 *
## degree.earnedGED             -3204.5     5507.6  -0.582    0.56125
## degree.earnedHS.DIPLOMA       -440.1     5386.9  -0.082    0.93496
## degree.earnedMASTER           9258.2    15682.3   0.590    0.55553
## degree.earnedNONE            -7011.7     5683.2  -1.234    0.21856
## marrital.statmarried         -1651.3     4836.0  -0.341    0.73307
## marrital.statnever.married   -6718.6     4511.3  -1.489    0.13779
## marrital.statsepareated      -9276.4     6562.6  -1.414    0.15885
## adlthood.numjobs              -612.8      219.1  -2.797    0.00559 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14370 on 229 degrees of freedom
##   (5053 observations deleted due to missingness)
## Multiple R-squared:  0.3289, Adjusted R-squared:  0.2732
## F-statistic: 5.906 on 19 and 229 DF,  p-value: 5.077e-12
```

This second model has a smaller adjusted RSquare value of 0.2731869, compared to the value of 0.301262 for the first model. Let us test to see whether the removal of birth date makes a significance difference

```
## Analysis of Variance Table
##
## Model 1: INCOME ~ ttl.incarc + age.first.incarc + school.sentiment + GENDER +
##     brth.mth + brth.yr + special.needs + RACE + degree.earned +
##     marrital.stat + adlthood.numjobs
## Model 2: INCOME ~ ttl.incarc + age.first.incarc + school.sentiment + GENDER +
##     special.needs + RACE + degree.earned + marrital.stat + adlthood.numjobs
##   Res.Df         RSS  Df   Sum of Sq Pr(>Chi)
## 1    217 43092349516
## 2    229 47302525380 -12 -4210175863  0.04751 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Res.Df"    "RSS"       "Df"        "Sum of Sq" "Pr(>Chi)"
```

The p-value of NA, 0.0475111 implying some significant different between the 2 model. This can be explained by the fact that the older one gets, the more money one earn, and removing birth year is the cause of the problem. To prove the point, we will add `brth.yr` back in the model

```
## Analysis of Variance Table
##
## Model 1: INCOME ~ ttl.incarc + age.first.incarc + school.sentiment + GENDER +
##     special.needs + RACE + degree.earned + marrital.stat + adlthood.numjobs
## Model 2: INCOME ~ ttl.incarc + age.first.incarc + school.sentiment + GENDER +
##     brth.yr + special.needs + RACE + degree.earned + marrital.stat +
##     adlthood.numjobs
##   Res.Df         RSS Df Sum of Sq Pr(>Chi)
## 1    229 47302525380
## 2    228 46935232586  1 367292794   0.1816
```

Immediately, we see that there is no longer any significant difference between the model with and without `brth.mth` . So we will amend the second model to exclude only `brth.mth` , apart from the exclusions in model 1

Model 3: Marrital status impact

In this step, we will investigate the impact of `marrital.stat` on `INCOME` . Let's run the new model and compare with our previous ones.

```
##
## Call:
## lm(formula = INCOME ~ ttl.incarc + age.first.incarc + school.sentiment +
##     GENDER + brth.yr + special.needs + RACE + degree.earned +
##     adlthood.numjobs, data = nlsy.1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28256  -9099  -2187   7562  53401
##
## Coefficients:
##                              Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                 2172268.3  1325676.4   1.639   0.102650
## ttl.incarc                    -1338.4      953.0  -1.404   0.161520
## age.first.incarc               -596.3      278.2  -2.144   0.033097 *
## school.sentimentunsafe         2472.4     3025.5   0.817   0.414648
## school.sentimentvery.safe     -2409.1     2192.9  -1.099   0.273089
## school.sentimentvery.unsafe    1620.5     5159.7   0.314   0.753746
## GENDERMale                     9478.8     2416.9   3.922   0.000116 ***
## brth.yr                       -1081.8      668.5  -1.618   0.106962
## special.needsyes              -7916.0     3160.3  -2.505   0.012937 *
## RACEHISPANIC                  12267.5     2687.0   4.565 0.00000809 ***
## RACEMIXED                     11885.9     8738.7   1.360   0.175102
## RACEOTHERS                    11651.1     2324.6   5.012 0.00000107 ***
## degree.earnedBACHELOR         16475.8     6494.4   2.537   0.011840 *
## degree.earnedGED              -2602.2     5259.8  -0.495   0.621257
## degree.earnedHS.DIPLOMA         363.4     5105.8   0.071   0.943323
## degree.earnedMASTER           12548.3    15497.7   0.810   0.418950
## degree.earnedNONE             -6029.6     5438.5  -1.109   0.268710
## adlthood.numjobs               -668.6      221.7  -3.015   0.002852 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14440 on 232 degrees of freedom
##   (5052 observations deleted due to missingness)
## Multiple R-squared:  0.315,  Adjusted R-squared:  0.2648
## F-statistic: 6.275 on 17 and 232 DF,  p-value: 4.948e-12
```

Removing `marrital.stat` resulted in lower adjusted RSquare value. Moreover, it made comparison between models impossible, as it adding more data points to the model (due to less number of `NA`). Thus, on the basis of adjusted RSquare, marrital status should be included back into the model. Instead, we will remove `ttl.incarc` and `school.sentiment` to examine the effect.

```
##
## Call:
## lm(formula = INCOME ~ age.first.incarc + GENDER + brth.yr + special.needs +
##     RACE + degree.earned + marrital.stat + adlthood.numjobs,
##     data = nlsy.1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28338  -8902  -2530   7341  52971
##
## Coefficients:
##                              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                2274683.94 1321531.39   1.721  0.086533 .
## age.first.incarc              -374.12     267.64  -1.398  0.163495
## GENDERMale                    8733.33    2515.01   3.472  0.000615 ***
## brth.yr                      -1134.37     666.78  -1.701  0.090227 .
## special.needsyes             -7779.14    3162.46  -2.460  0.014627 *
## RACEHISPANIC                 11809.19    2686.48   4.396 0.0000168 ***
## RACEMIXED                    11009.12    8680.73   1.268  0.205983
## RACEOTHERS                   11780.36    2344.19   5.025 0.0000010 ***
## degree.earnedBACHELOR        14218.97    6745.62   2.108  0.036110 *
## degree.earnedGED             -3380.66    5477.20  -0.617  0.537689
## degree.earnedHS.DIPLOMA       -784.99    5368.22  -0.146  0.883867
## degree.earnedMASTER           5912.93   15660.89   0.378  0.706101
## degree.earnedNONE            -6493.31    5602.09  -1.159  0.247607
## marrital.statmarried           -52.79    4830.61  -0.011  0.991290
## marrital.statnever.married   -5794.80    4560.04  -1.271  0.205074
## marrital.statsepareated      -8247.75    6553.91  -1.258  0.209490
## adlthood.numjobs              -661.40     221.59  -2.985  0.003141 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14370 on 233 degrees of freedom
##   (5052 observations deleted due to missingness)
## Multiple R-squared:  0.3248, Adjusted R-squared:  0.2785
## F-statistic: 7.006 on 16 and 233 DF,  p-value: 4.253e-13
```

Comparison between this new model and the previous ones was also impossible due to data size difference. Nevertheless, there is a slight improvement in adjusted RSquare. This model also showed that, despite earlier analysis of the bar chart, the highest degree earned does not seems to have significant impact on `INCOME`. The next model will try to look deeper into this matter.

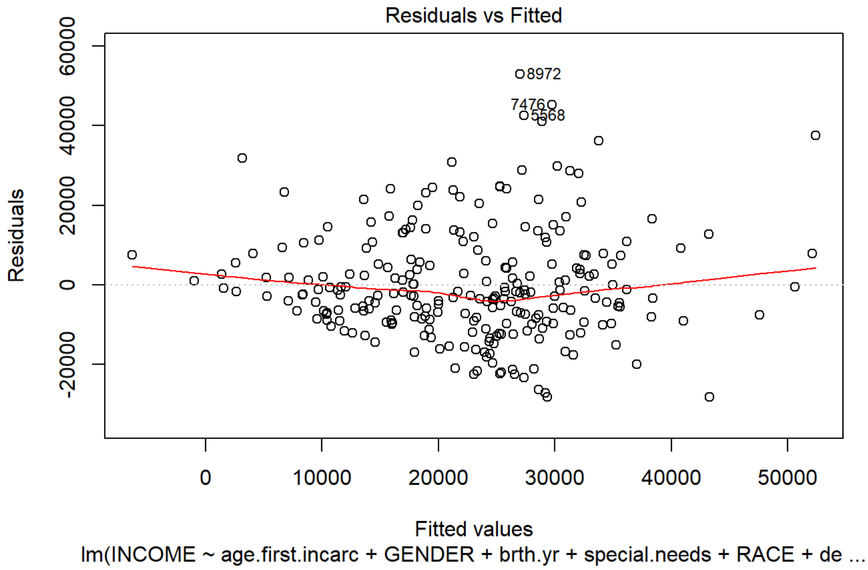
Model 4: Highest degree earned impact

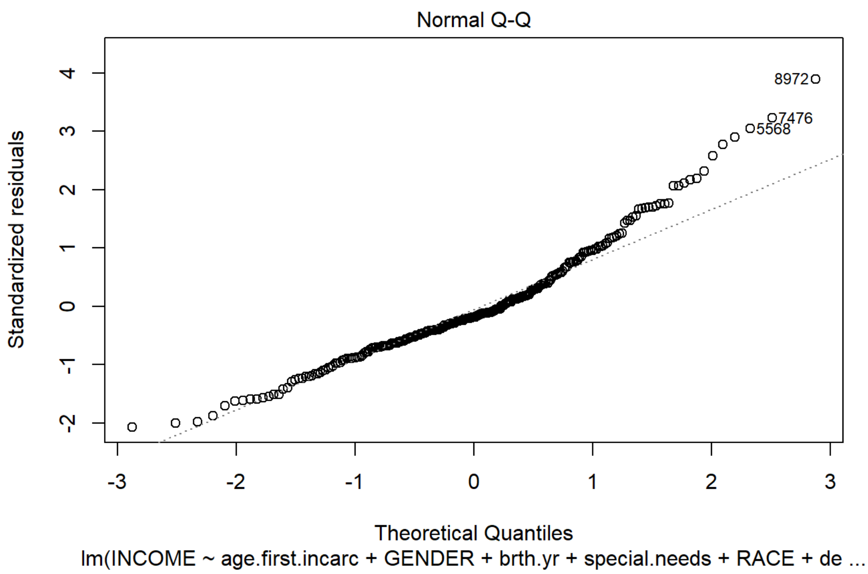
```
##
## Call:
## lm(formula = INCOME ~ age.first.incarc + GENDER + brth.yr + special.needs +
##     RACE + marrital.stat + adlthood.numjobs, data = nlsy.1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28089  -9188  -2090   7283  54147
##
## Coefficients:
##                             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                2141587.8  1343149.3   1.594  0.112134
## age.first.incarc              -255.7      268.8  -0.951  0.342493
## GENDERMale                    8286.9     2481.7   3.339  0.000972 ***
## brth.yr                      -1071.0      677.8  -1.580  0.115355
## special.needsyes             -9033.1     3246.7  -2.782  0.005822 **
## RACEHISPANIC                 11658.8     2726.5   4.276 0.0000273 ***
## RACEMIXED                    15924.5     8857.0   1.798  0.073425 .
## RACEOTHERS                   12581.2     2387.1   5.271 0.0000003 ***
## marrital.statmarried          3810.2     4754.2   0.801  0.423655
## marrital.statnever.married   -3151.5     4505.9  -0.699  0.484953
## marrital.statsepareated      -6382.5     6612.8  -0.965  0.335412
## adlthood.numjobs              -636.9      225.8  -2.820  0.005198 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14830 on 243 degrees of freedom
##   (5047 observations deleted due to missingness)
## Multiple R-squared:   0.26,  Adjusted R-squared:  0.2265
## F-statistic: 7.762 on 11 and 243 DF,  p-value: 1.698e-11
```

The adjusted RSquare for this model falls to 0.2265214, a significant drop from earlier value of 0.2784631 of the third model. Thus, it can be concluded that dropping `degree.earned` will make the model perform worse.

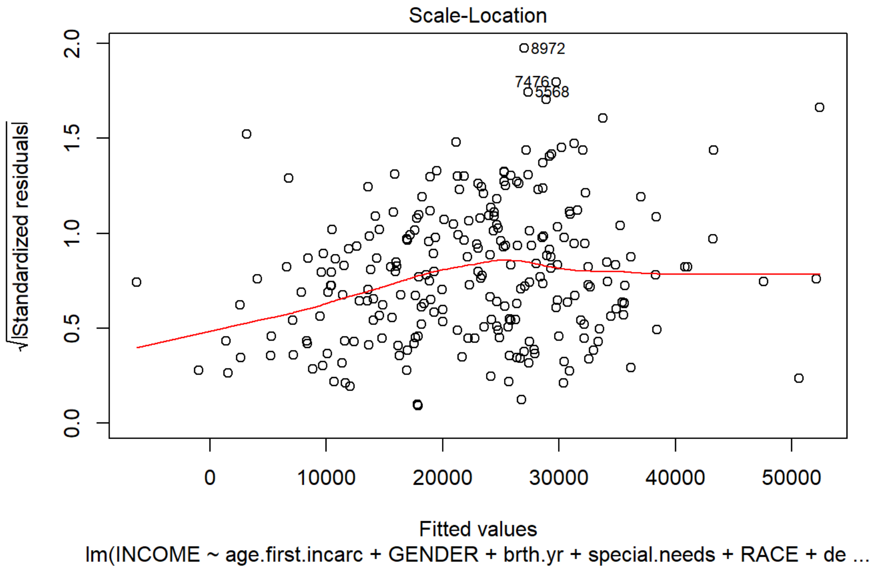
Model selection. Through running and comparing different models, I concluded that the third model, the one containing `age.first.incarc`, `GENDER`, `brth.yr`, `special.needs`, `RACE`, `degree.earned`, `marrital.stat`, and `adlthood.numjobs`, best accounts for one's income potential. below is the performance plots of that model

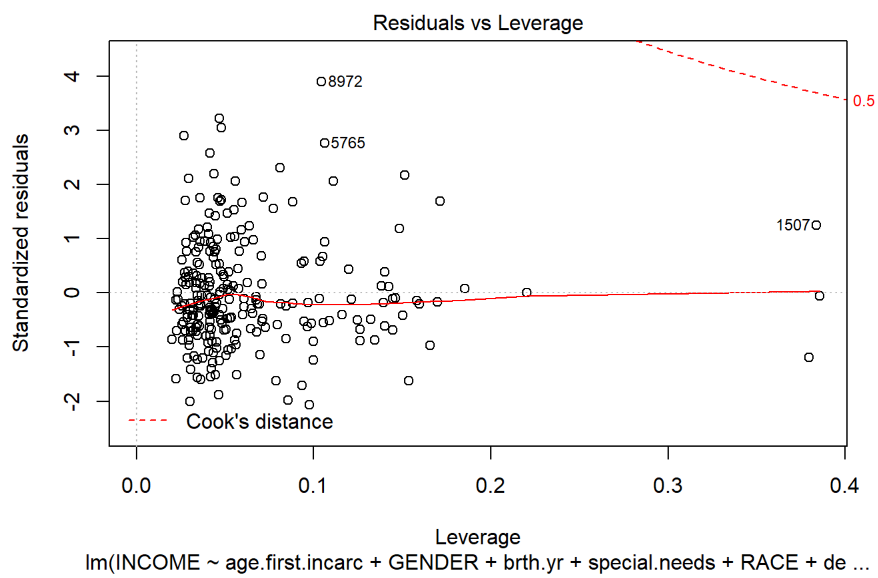
```
## Warning: not plotting observations with leverage one:
##   174
```






```
## Warning: not plotting observations with leverage one:
##   174
```





The Residuals vs. Fitted plot does not have any clear pattern, even though the line has a small dip at midrange of the fitted values. Normal QQ plot shows that the residuals have a normal distribution for most of the range of data. However, the Residuals vs. Leverage plot has a slightly funnel shape.

IV. Conclusion

Go back to the questions when we start out, there are some that has been answered along the way when the model analysis was performed. It can be confirmed that birth month has no impact on how much one is entitle to earned later in life. Whether one has been incarcerated or not also have no predictive power in predicting that person’s future income, and the same applies to whether one feels safe at school or not. On the contrary, if one have a special condition in needs of assistance, the person will likely earn -7779.14 than those who don’t have any condition.

It is also confirmed by the model that the higher the degree you earn, the higher you can get paid. The evidence for this is that earning a bachelor degree can add 14218.97 with p-value of 0.04 to the income compare to the professional degree, while holding the GED will make one’s income worse off by -3380.66. Marrital status can also makes an impact on your earning, albeit may not be clear and vary from case to case. Married people seems to earn more compare to never married or seperated groups.

Most importantly, and sadly, your gender have a very high chance of determining you will earn more or less base on the data set. Keeping everything else constant, men earn on average 8733.33 than women at 95% confidence.