Rohan Taneja & Mursal Hajiyev - 46457 & 46480

Introduction :

Our objective is to gain insight into the relationship between different factors such as employment stability, median salary, and the field of study, in order to understand which majors have the best outcomes in terms of employment stability and earning potential. This information could be useful for students considering which major to pursue, as well as for educators and policymakers looking to improve career outcomes for graduates in different fields. Additionally, the data could be used to identify patterns and trends in the job market and to inform strategies for improving career opportunities for graduates in different fields. Overall, the goal is to understand the connection between one's major and their future career opportunities and earning potential, in order to make better-informed decisions.

The analysis also helps us to understand how the majors of graduates affects their employability and the sort of salaries they can earn while putting in the median amount of effort for their degree's cohort. This will help us solve two different issues. A lot of the younger generation today do not know what they want to study and end up studying subjects they do not like with salaries that do not live up to expectations. This analysis will provide students with realistic expectations of what they can earn while putting in average amounts of effort. This will lead to happier students and graduates if not less disappointment further down the line

Our dataset :
https://www.kaggle.com/datasets/thedevastator/uncovering-insights-to-college-majors-and-their

Manipulating the dataset to fit our needs :

Our dataset contains data on the number of graduates and non-graduates from various majors, as well as their employment and salary statistics.

It has has the following variables :

1. Index - index
2. Major_code: The code associated with the major. (Integer)
3. Major_category: The category of the major. (String)
4. Total: The total number of students in the major. (Integer)
5. Employed: The number of employed graduates from the major. (Integer)
6. Unemployed: The number of unemployed graduates from the major. (Integer)
7. Unemployment_rate: The unemployment rate of graduates from the major. (Float)
8. Median: The median salary of graduates from the major. (Integer)
9. P25th: The 25th percentile salary of graduates from the major. (Integer)
10. P75th: The 75th percentile salary of graduates from the major. (Integer)
11. Rank: The rank of the major in terms of popularity. (Integer)

12. Sample_size: The sample size of graduates from the major. (Integer)
13. Men: The number of male students in the major. (Integer)
14. Women: The number of female students in the major. (Integer)
15. ShareWomen: The percentage of female students in the major. (Float)
16. Full_time: The number of graduates employed full-time. (Integer)
17. Part_time: The number of graduates employed part-time. (Integer)
18. Full_time_year_round: The number of graduates employed full-time year-round. (Integer)
19. College_jobs: The number of college jobs held by graduates from the major. (Integer)
20. Non_college_jobs: The number of non-college jobs held by graduates from the major. (Integer)
21. Low_wage_jobs: The number of low-wage jobs held by graduates from the major. (Integer)

Our data had some NA values in it so we proceeded to remove those and clean the data. We then had our perfect dataset. So we set out to understand the distribution of the most relevant variables and how they were distributed against major_category. Here are the results :

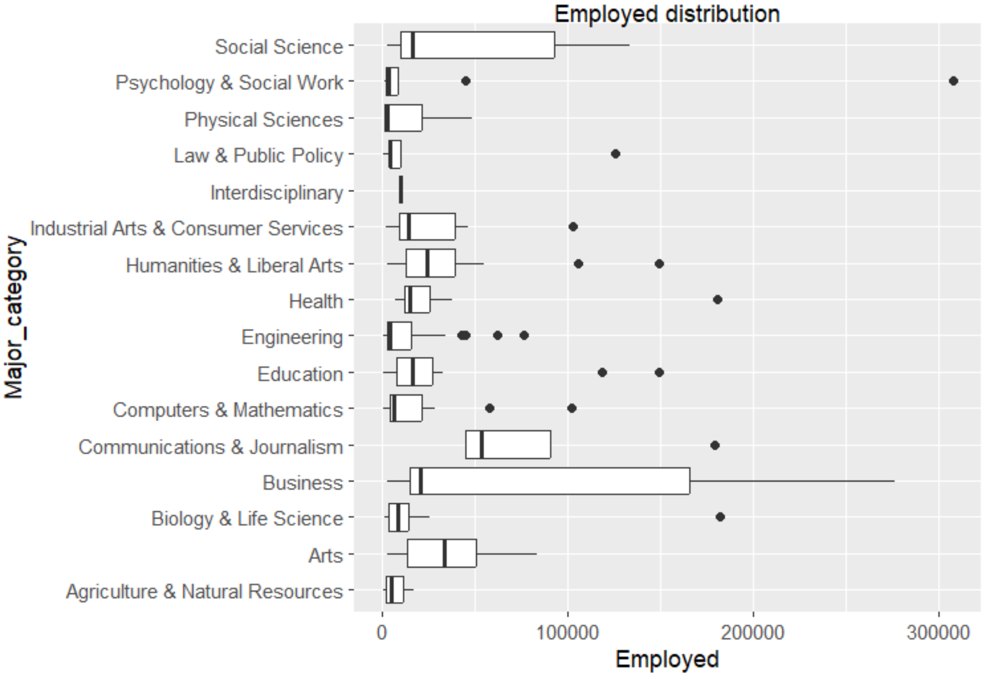
We felt like the most important and relevant variables were

1. "Employed"
2. "Full_time_year_round"
3. "Unemployed"
4. "Unemployment_rate"
5. "Median"
6. "Men"
7. "Women"
8. "College_jobs"

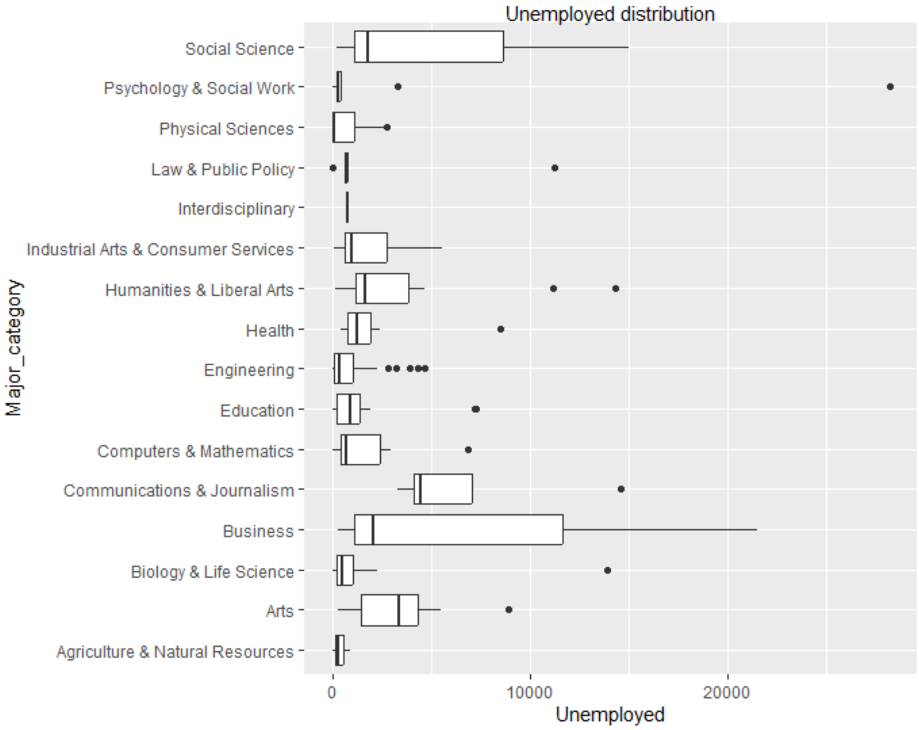
Exploratory Data Analysis

We start by creating plots of each of the important variables against major_categories.This would highlight how gender, median income & employability are affected by major categories.

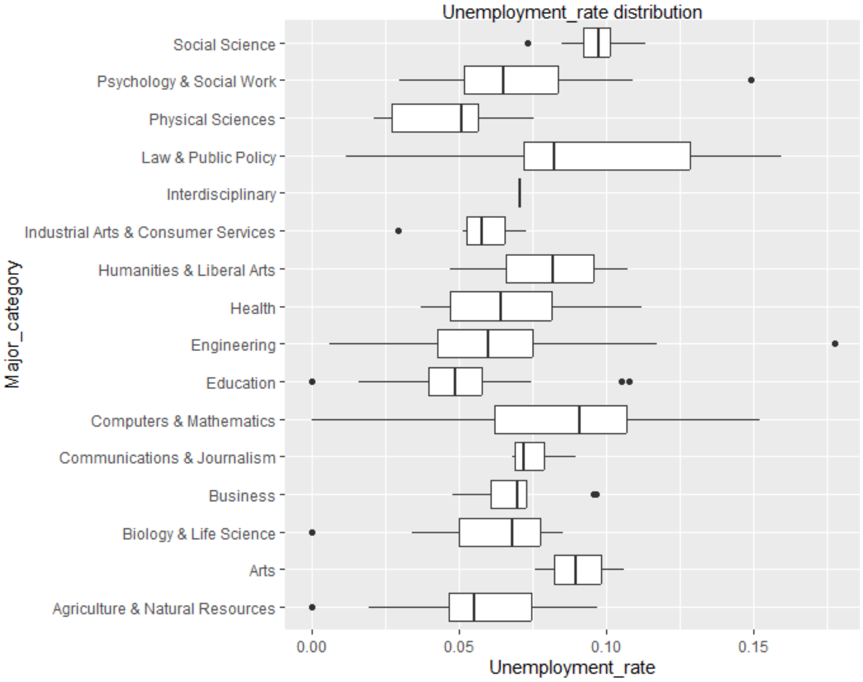
Our code generates boxplots for each column in the list ‘columns_list’ with ‘Major_category’ on the y-axis. The purpose of these boxplots is to visualize the distribution of values for each column in relation to the categories in the ‘Major_category’ column. The title of each graph is the name of the column being plotted with "distribution" appended to it.



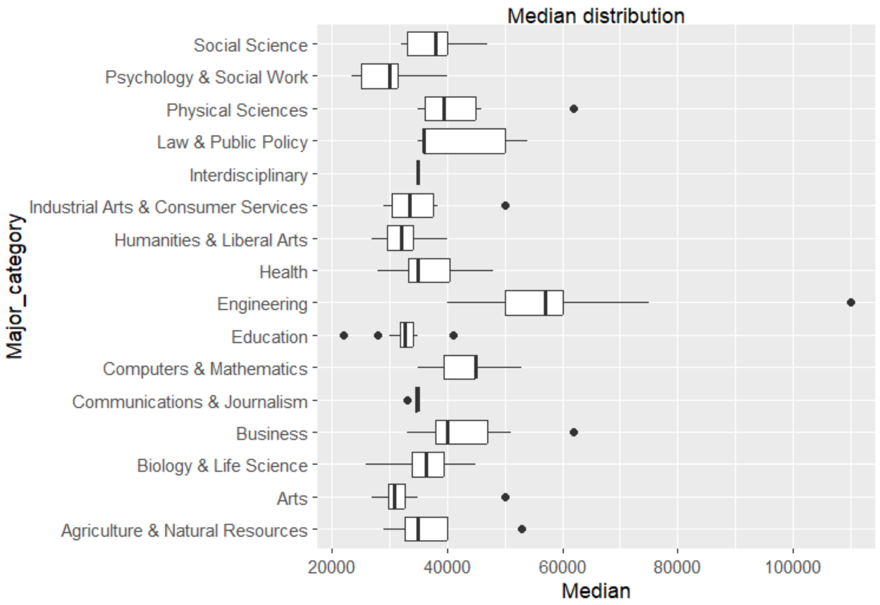
Our observations here tell us that the field of Business has varying levels of employed people per degree with the median value being in the low.

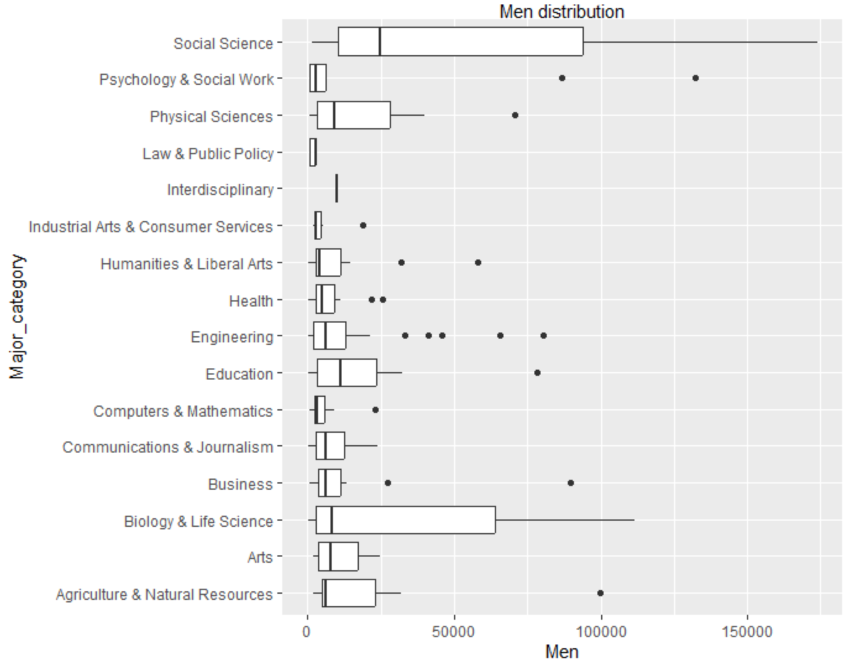


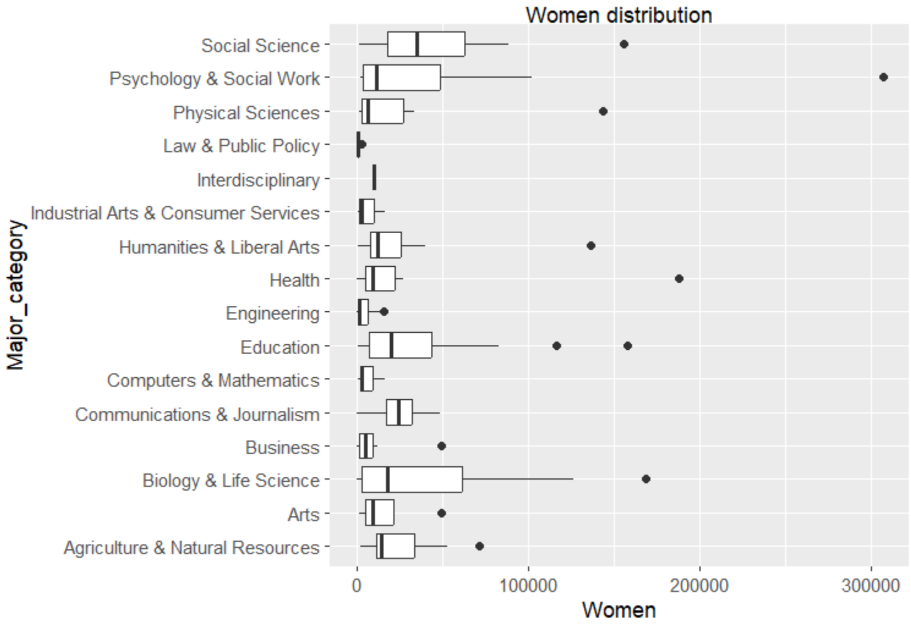
We observe that people that do Business tend to have a higher rate of Unemployment than other Major Categories.



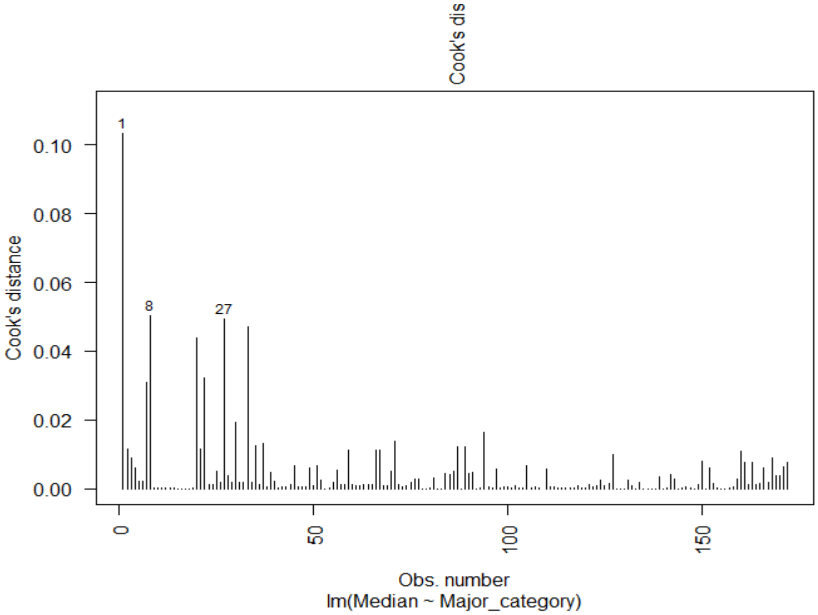
The Unemployment rate is heavily distributed between the majors.

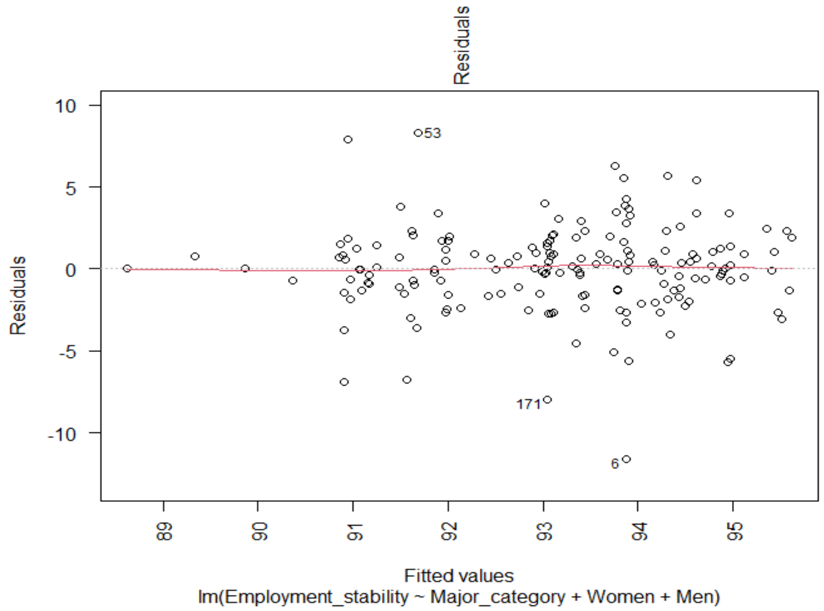

Median distribution
Social Science
Psychology & Social Work
Physical Sciences
Law & Public Policy
Interdisciplinary
Industrial Arts & Consumer Services
Humanities & Liberal Arts
Health
Engineering
Education
Computers & Mathematics
Communications & Journalism
Business
Biology & Life Science
Arts
Agriculture & Natural Resources
Major_category
20000
40000
60000
80000
100000
Median

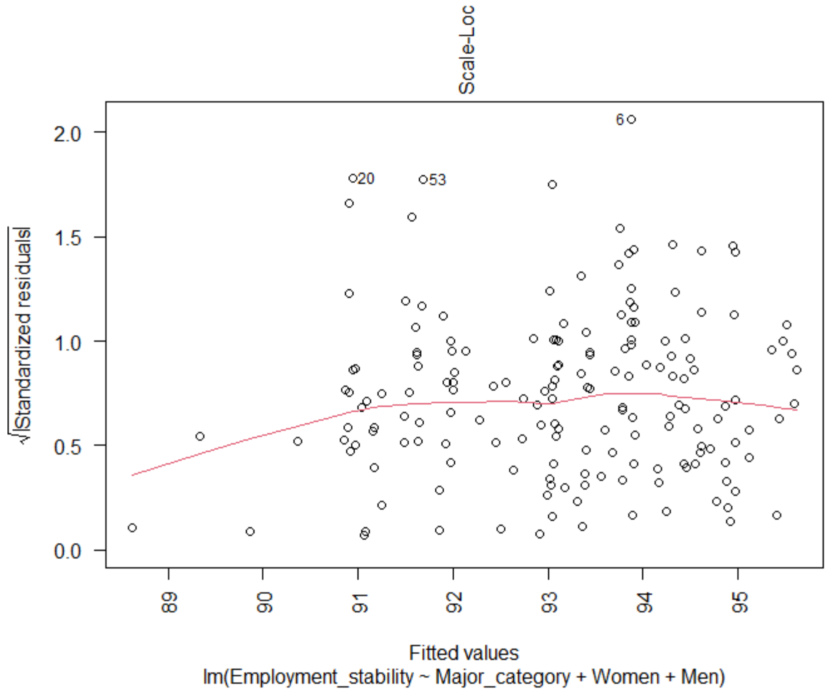

Men distribution
Social Science
Psychology & Social Work
Physical Sciences
Law & Public Policy
Interdisciplinary
Industrial Arts & Consumer Services
Humanities & Liberal Arts
Health
Engineering
Education
Computers & Mathematics
Communications & Journalism
Business
Biology & Life Science
Arts
Agriculture & Natural Resources
Major_category
0
50000
100000
150000
Men

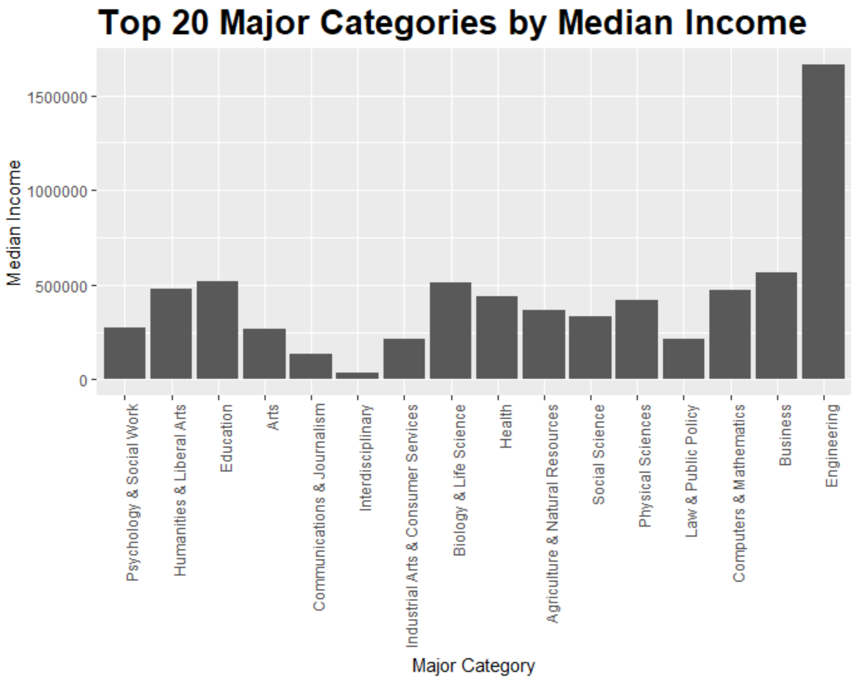


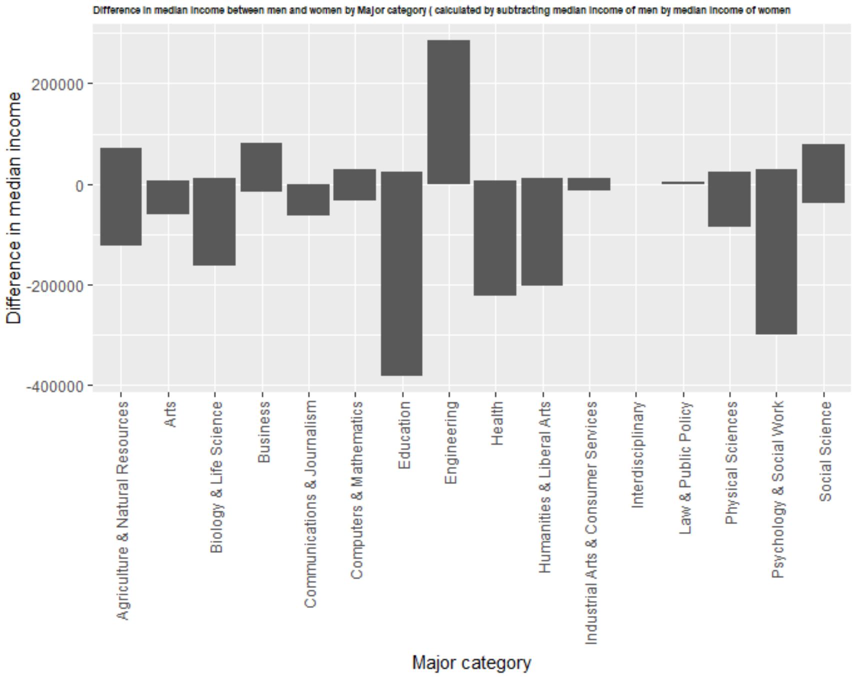
These graphs tells us that our data is relatively normalized



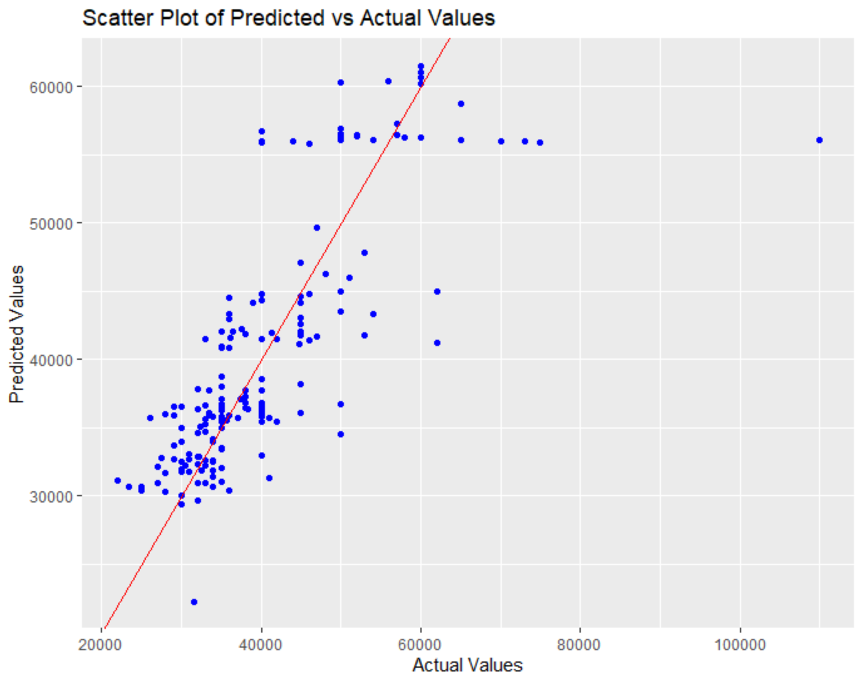

Residuals
10
5
0
-5
-10
Residuals
53
171
6
89
90
91
92
93
94
95
Fitted values
lm(Employment_stability ~ Major_category + Women + Men)

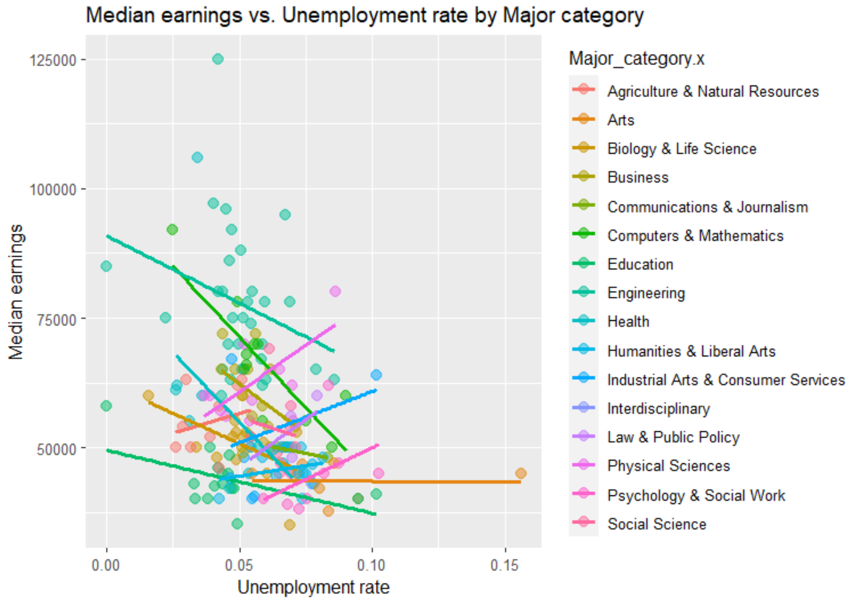

Scale-Loc
2.0
1.5
1.0
0.5
0.0
$\sqrt{|\text{Standardized residuals}|}$
6
20
53
89
90
91
92
93
94
95
Fitted values
lm(Employment_stability ~ Major_category + Women + Men)



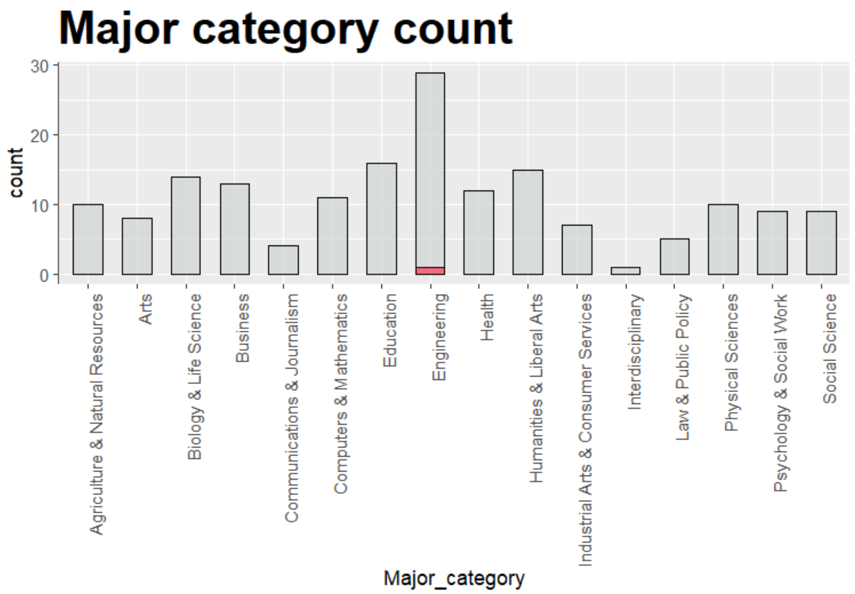


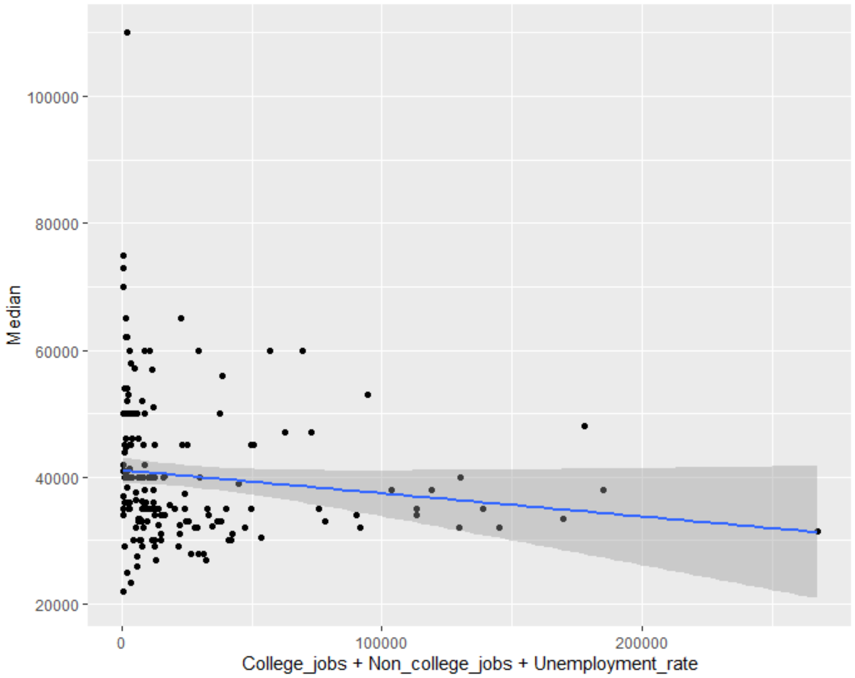
Difference in median income between men and women by Major categories





This is a scatterplot with a linear regression line that shows the relationship between Median earnings and Unemployment rate by Major category. The color aesthetic is used to represent the Major category and the points are displayed with size 3 and alpha 0.5 to show the distribution of data points for each Major category. The smooth line represents the best fit line of the relationship between Median earnings and Unemployment rate.


Major category count
count
30
20
10
0
Agriculture & Natural Resources
Arts
Biology & Life Science
Business
Communications & Journalism
Computers & Mathematics
Education
Engineering
Health
Humanities & Liberal Arts
Industrial Arts & Consumer Services
Interdisciplinary
Law & Public Policy
Physical Sciences
Psychology & Social Work
Social Science
Major_category

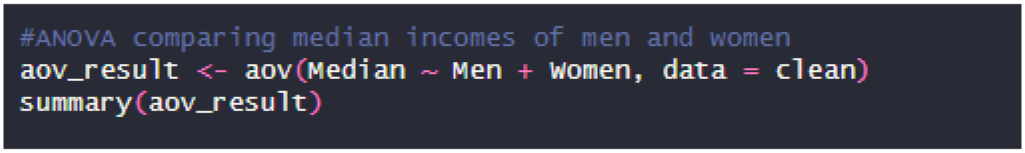

Median
100000
80000
60000
40000
20000
0
100000
200000
College_jobs + Non_college_jobs + Unemployment_rate

The purpose of our investigation was to find out what sort of effect Major_category has on median income & to find out what sort of effect gender plays on Median income:

First we used an ANOVA test to test the effect of gender on Median income:

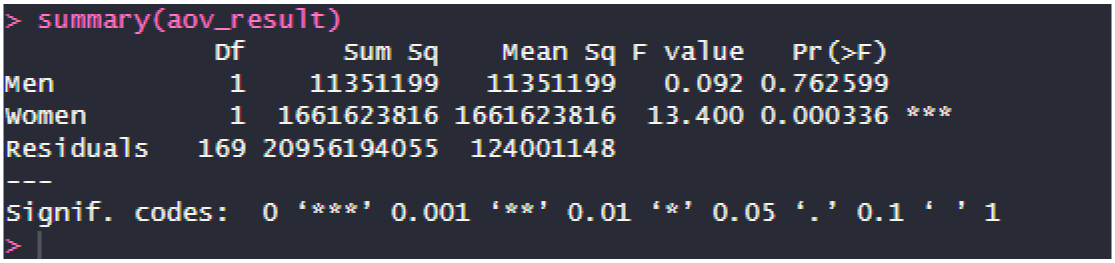
Our IV is Men and Women

Our DV is Median



```
#ANOVA comparing median incomes of men and women
aov_result <- aov(Median ~ Men + Women, data = clean)
summary(aov_result)
```

Our result:



```
> summary(aov_result)
             Df      Sum Sq    Mean Sq F value   Pr(>F)    
Men           1    11351199   11351199   0.092 0.762599    
Women         1  1661623816 1661623816  13.400 0.000336 ***
Residuals   169 20956194055  124001148                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

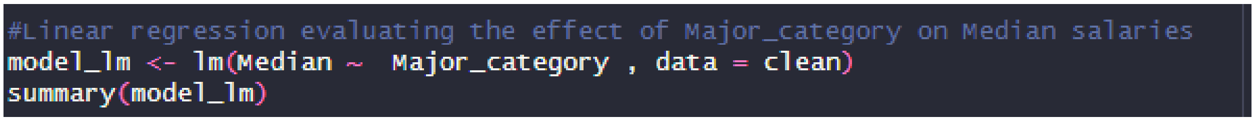
The results show that there is a significant difference in the mean of the dependent variable between Men and Women (p = 0.000336, indicated by ***), as the F-value of 13.400 is larger than the critical value, and the p-value is less than the significance level (typically 0.05).

However, the effect size is small, as indicated by the low F-value of 0.092 for the Men group comparison. So we can conclude that there is an effect , it has a lesser effect than previously thought so.

Next we evaluate the effect of Major_category on Median incomes

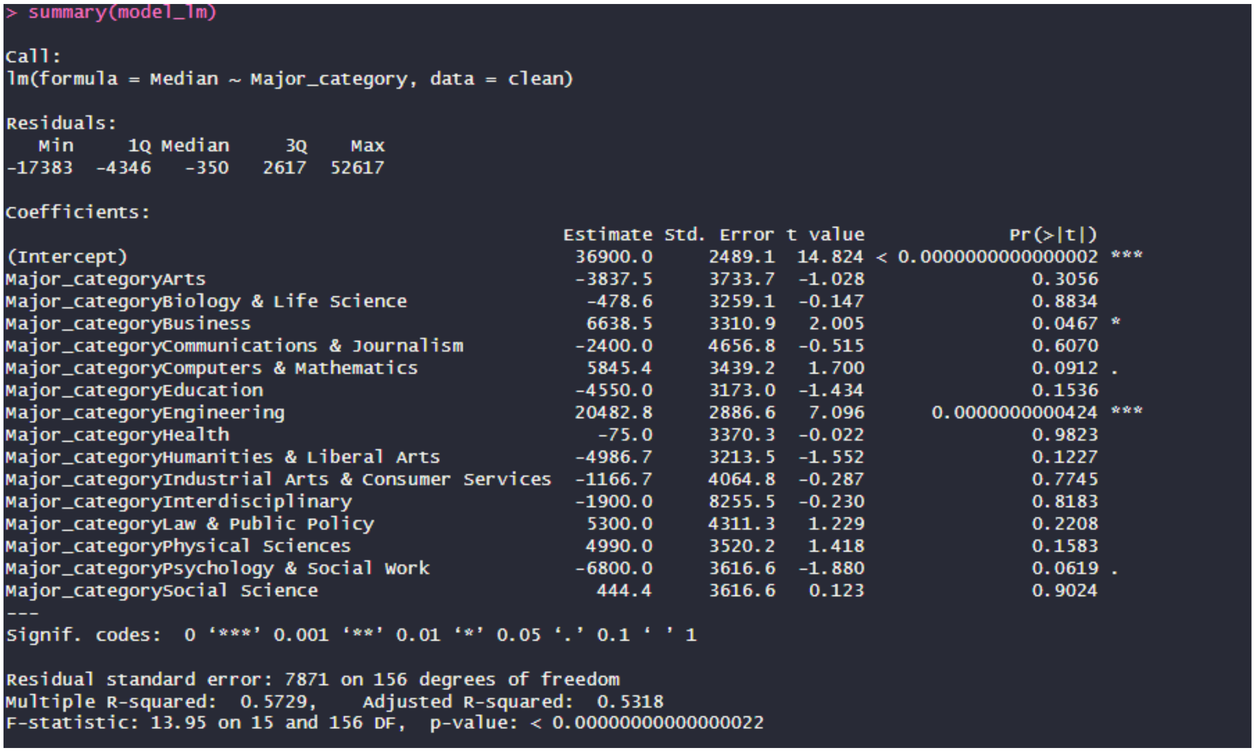
Our IV is Major_category

Our DV is Median



```
#Linear regression evaluating the effect of Major_category on Median salaries
model_lm <- lm(Median ~  Major_category , data = clean)
summary(model_lm)
```

Our results:

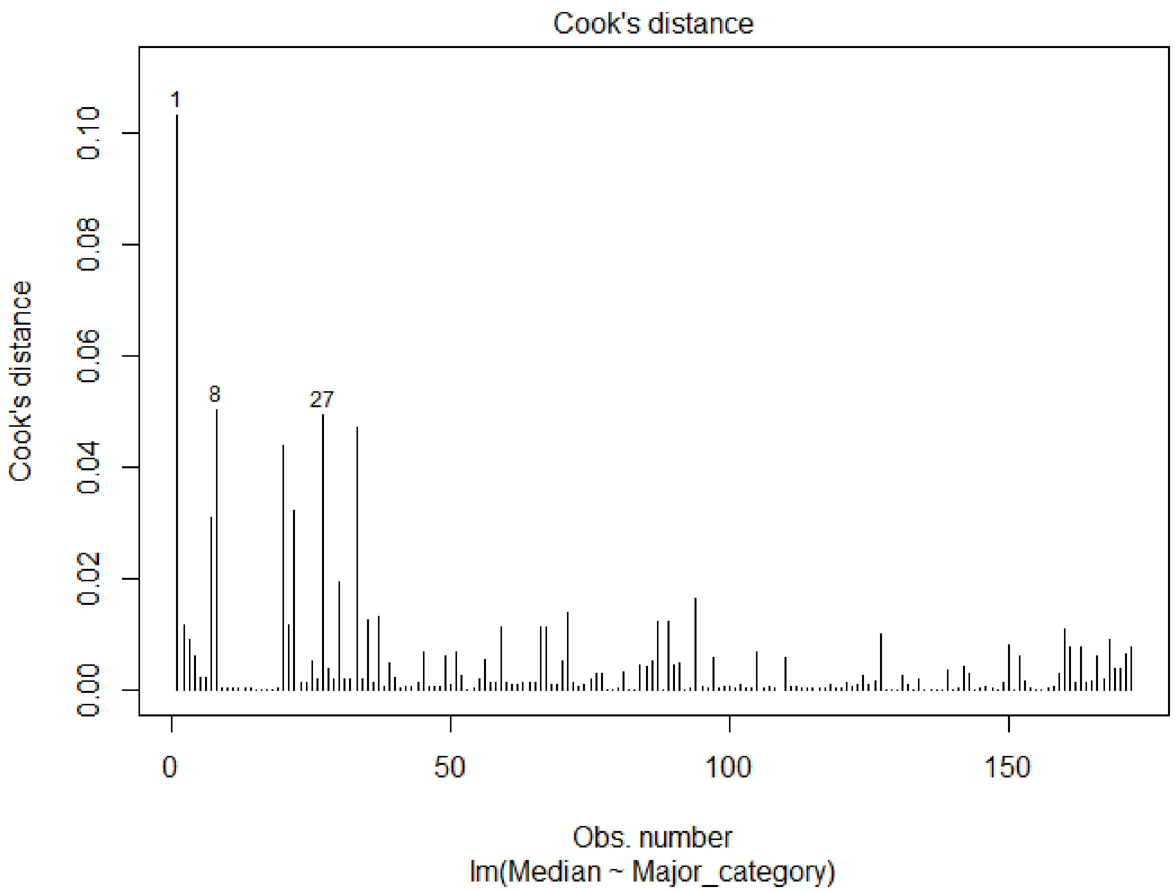


```
> summary(model_lm)

Call:
lm(formula = Median ~ Major_category, data = clean)

Residuals:
   Min     1Q Median     3Q    Max 
-17383  -4346   -350   2617  52617 

Coefficients:
                                              Estimate Std. Error t value             Pr(>|t|)    
(Intercept)                                    36900.0     2489.1  14.824 < 0.0000000000000002 ***
Major_categoryArts                             -3837.5     3733.7  -1.028               0.3056    
Major_categoryBiology & Life Science            -478.6     3259.1  -0.147               0.8834    
Major_categoryBusiness                          6638.5     3310.9   2.005               0.0467 *  
Major_categoryCommunications & Journalism      -2400.0     4656.8  -0.515               0.6070    
Major_categoryComputers & Mathematics           5845.4     3439.2   1.700               0.0912 .  
Major_categoryEducation                        -4550.0     3173.0  -1.434               0.1536    
Major_categoryEngineering                      20482.8     2886.6   7.096       0.0000000000424 ***
Major_categoryHealth                             -75.0     3370.3  -0.022               0.9823    
Major_categoryHumanities & Liberal Arts        -4986.7     3213.5  -1.552               0.1227    
Major_categoryIndustrial Arts & Consumer Services -1166.7     4064.8  -0.287               0.7745    
Major_categoryInterdisciplinary                -1900.0     8255.5  -0.230               0.8183    
Major_categoryLaw & Public Policy               5300.0     4311.3   1.229               0.2208    
Major_categoryPhysical Sciences                 4990.0     3520.2   1.418               0.1583    
Major_categoryPsychology & Social Work         -6800.0     3616.6  -1.880               0.0619 .  
Major_categorySocial Science                     444.4     3616.6   0.123               0.9024    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7871 on 156 degrees of freedom
Multiple R-squared:  0.5729,	Adjusted R-squared:  0.5318 
F-statistic: 13.95 on 15 and 156 DF,  p-value: < 0.00000000000000022
```

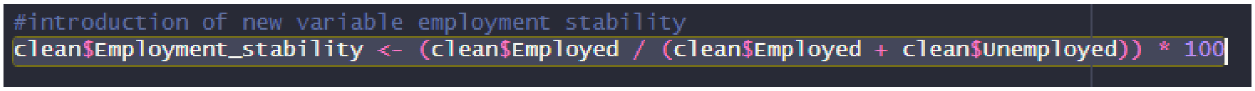
As we can see the major categories that influence median income the most are Business and Engineering. Notable mentions include Computers and Mathematics , Psychology & Social Work

a positive coefficient indicates a positive relationship between the independent and dependent variables, and a negative coefficient indicates a negative relationship. The magnitude of the coefficient indicates the strength of the relationship: a larger magnitude means a stronger relationship, and a smaller magnitude means a weaker relationship. Engineering has by the far the largest positive correlation of 20482.5 and Business has a correlation of 6638.5 . Even though the model does not think so, picking degrees in Arts, Psychology & Social Work and Humanities & Liberal Arts is seen to have a negative coefficient which means picking degrees like this actively harms our median income.

Cook’s distance for the linear model:

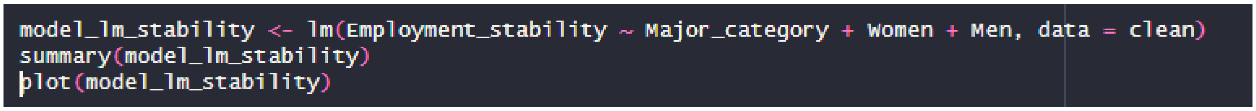


We have now answered our primary hypothesis. We move onto secondary hypotheses that include Employment stability. We introduce a new variable Employment stability:



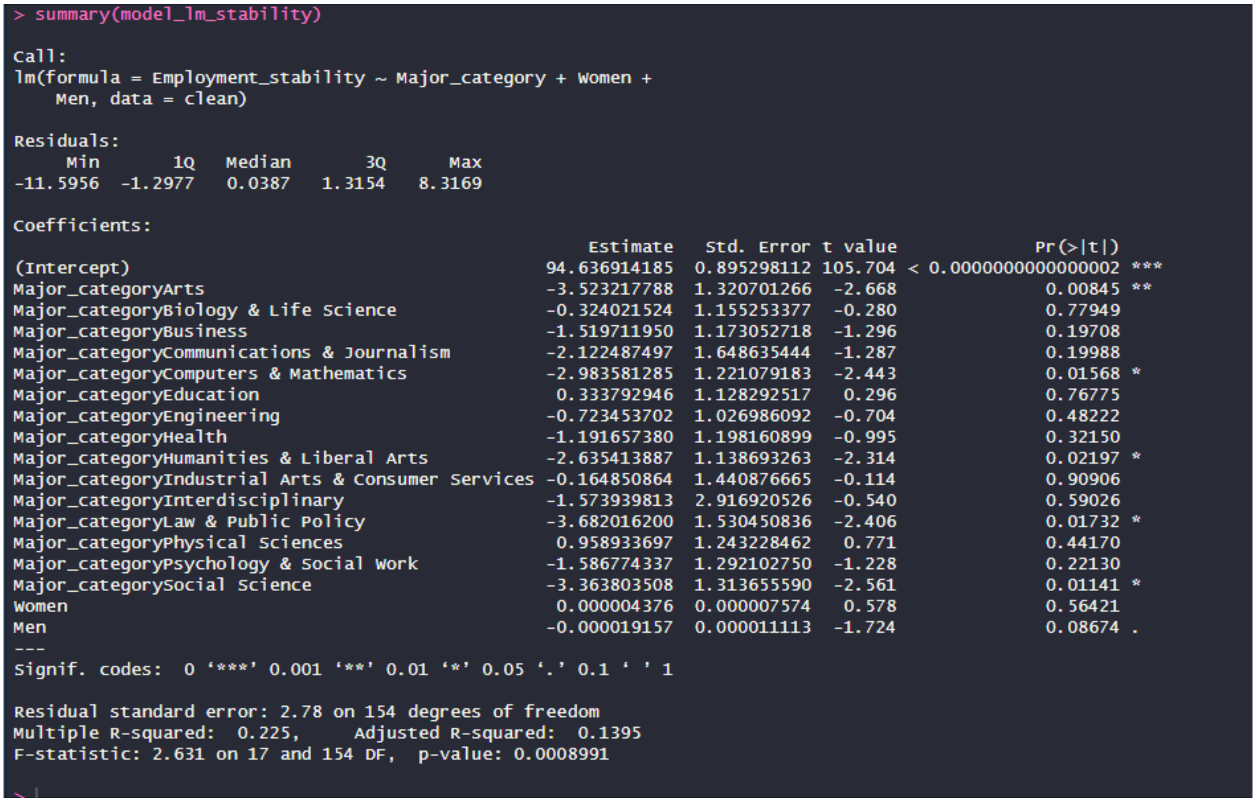
```
#introduction of new variable employment stability
clean$Employment_stability <- (clean$Employed / (clean$Employed + clean$Unemployed)) * 100
```

We then attempt to find how our new variable is affected by Major_category and Gender



```
model_lm_stability <- lm(Employment_stability ~ Major_category + Women + Men, data = clean)
summary(model_lm_stability)
plot(model_lm_stability)
```

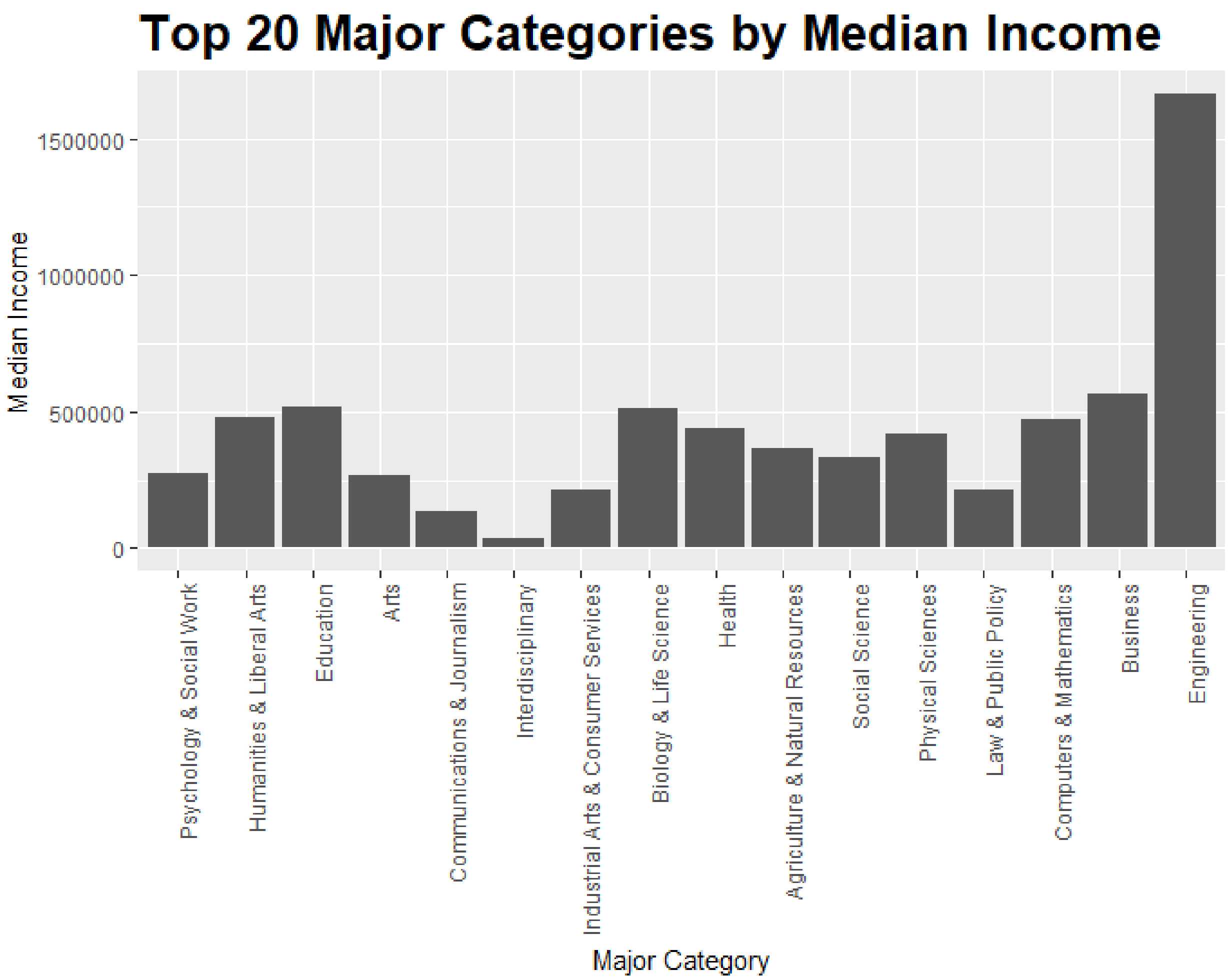
Our result



```
> summary(model_lm_stability)

Call:
lm(formula = Employment_stability ~ Major_category + Women +
    Men, data = clean)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.5956  -1.2977   0.0387   1.3154   8.3169 

Coefficients:
                                            Estimate   Std. Error t value             Pr(>|t|)    
(Intercept)                              94.636914185  0.895298112 105.704 < 0.0000000000000002 ***
Major_categoryArts                       -3.523217788  1.320701266  -2.668              0.00845 ** 
Major_categoryBiology & Life Science     -0.324021524  1.155253377  -0.280              0.77949    
Major_categoryBusiness                   -1.519711950  1.173052718  -1.296              0.19708    
Major_categoryCommunications & Journalism -2.122487497  1.648635444  -1.287              0.19988    
Major_categoryComputers & Mathematics    -2.983581285  1.221079183  -2.443              0.01568 *  
Major_categoryEducation                   0.333792946  1.128292517   0.296              0.76775    
Major_categoryEngineering                -0.723453702  1.026986092  -0.704              0.48222    
Major_categoryHealth                     -1.191657380  1.198160899  -0.995              0.32150    
Major_categoryHumanities & Liberal Arts  -2.635413887  1.138693263  -2.314              0.02197 *  
Major_categoryIndustrial Arts & Consumer Services -0.164850864  1.440876665  -0.114              0.90906    
Major_categoryInterdisciplinary          -1.573939813  2.916920526  -0.540              0.59026    
Major_categoryLaw & Public Policy        -3.682016200  1.530450836  -2.406              0.01732 *  
Major_categoryPhysical Sciences           0.958933697  1.243228462   0.771              0.44170    
Major_categoryPsychology & Social Work   -1.586774337  1.292102750  -1.228              0.22130    
Major_categorySocial Science             -3.363803508  1.313655590  -2.561              0.01141 *  
Women                                     0.000004376  0.000007574   0.578              0.56421    
Men                                      -0.000019157  0.000011113  -1.724              0.08674 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.78 on 154 degrees of freedom
Multiple R-squared:  0.225,	Adjusted R-squared:  0.1395 
F-statistic: 2.631 on 17 and 154 DF,  p-value: 0.0008991

>
```

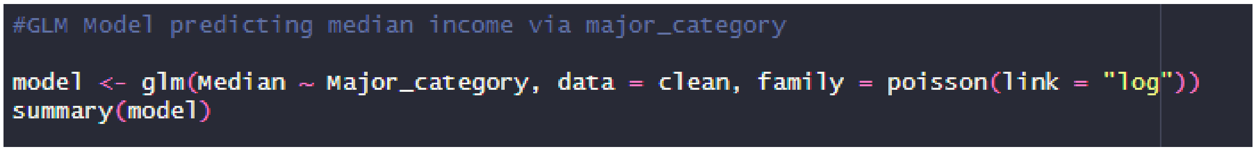
When it comes to employment stability , we can see that in the grand scale of things gender does not play as significant a role.But picking degrees in the faculties of Arts, Computers & Mathematics , Humanities & Liberal Arts , Law & Public Policy and Social Science can actively harm a potential graduates employment stability.

We then find out which major categories make the most income:



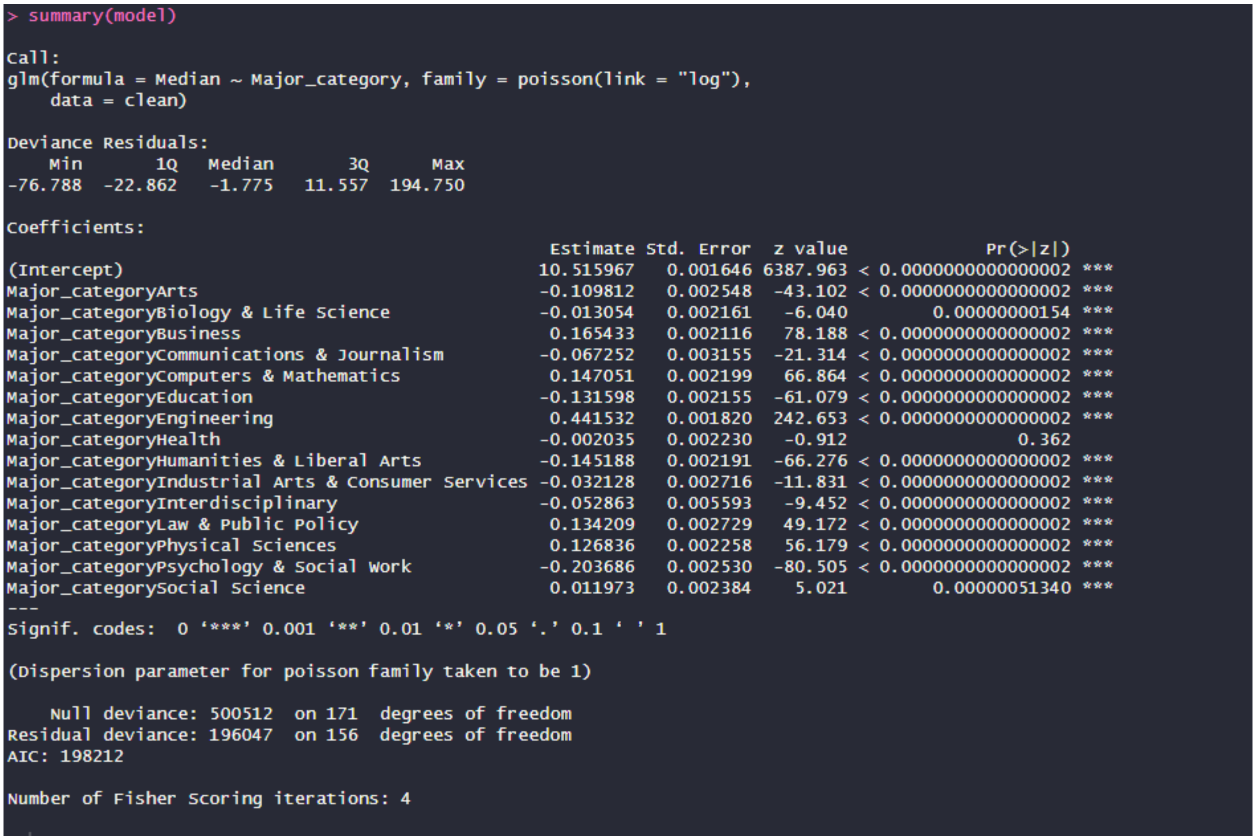
Not surprisingly Engineering makes a lot of income.

We think our prediction of median income by major category can be improved with another model. So we decide to use a GLM with poisson distribution:



```
#GLM Model predicting median income via major_category

model <- glm(Median ~ Major_category, data = clean, family = poisson(link = "log"))
summary(model)
```

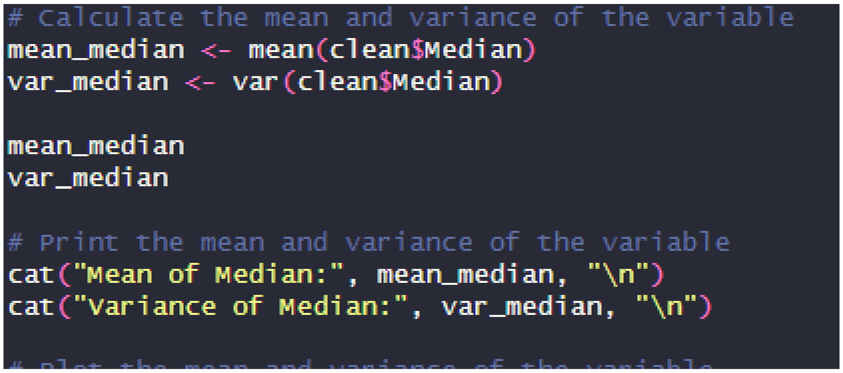
Our results:



```
> summary(model)

Call:
glm(formula = Median ~ Major_category, family = poisson(link = "log"), 
    data = clean)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-76.788  -22.862   -1.775   11.557  194.750  

Coefficients:
                                                Estimate Std. Error  z value             Pr(>|z|)    
(Intercept)                                    10.515967   0.001646 6387.963 < 0.0000000000000002 ***
Major_categoryArts                             -0.109812   0.002548  -43.102 < 0.0000000000000002 ***
Major_categoryBiology & Life Science           -0.013054   0.002161   -6.040        0.00000000154 ***
Major_categoryBusiness                          0.165433   0.002116   78.188 < 0.0000000000000002 ***
Major_categoryCommunications & Journalism      -0.067252   0.003155  -21.314 < 0.0000000000000002 ***
Major_categoryComputers & Mathematics           0.147051   0.002199   66.864 < 0.0000000000000002 ***
Major_categoryEducation                        -0.131598   0.002155  -61.079 < 0.0000000000000002 ***
Major_categoryEngineering                       0.441532   0.001820  242.653 < 0.0000000000000002 ***
Major_categoryHealth                           -0.002035   0.002230   -0.912                0.362    
Major_categoryHumanities & Liberal Arts        -0.145188   0.002191  -66.276 < 0.0000000000000002 ***
Major_categoryIndustrial Arts & Consumer Services -0.032128   0.002716  -11.831 < 0.0000000000000002 ***
Major_categoryInterdisciplinary                -0.052863   0.005593   -9.452 < 0.0000000000000002 ***
Major_categoryLaw & Public Policy               0.134209   0.002729   49.172 < 0.0000000000000002 ***
Major_categoryPhysical Sciences                 0.126836   0.002258   56.179 < 0.0000000000000002 ***
Major_categoryPsychology & Social Work         -0.203686   0.002530  -80.505 < 0.0000000000000002 ***
Major_categorySocial Science                    0.011973   0.002384    5.021        0.00000051340 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 500512  on 171  degrees of freedom
Residual deviance: 196047  on 156  degrees of freedom
AIC: 198212

Number of Fisher Scoring iterations: 4
```

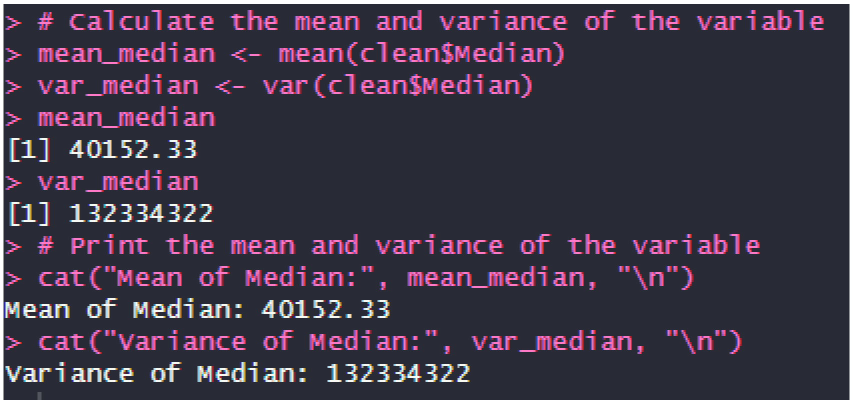
They are less than favorable and it might have something to do with our data’s normalization level

So we check the median and variance of Median



```
# Calculate the mean and variance of the variable
mean_median <- mean(clean$Median)
var_median <- var(clean$Median)

mean_median
var_median

# Print the mean and variance of the variable
cat("Mean of Median:", mean_median, "\n")
cat("Variance of Median:", var_median, "\n")

# Plot the mean and variance of the variable
```

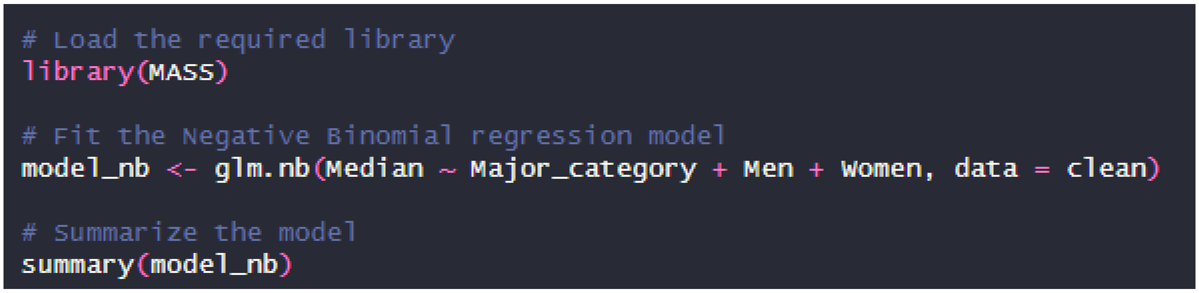
Our result:



```
> # Calculate the mean and variance of the variable
> mean_median <- mean(clean$Median)
> var_median <- var(clean$Median)
> mean_median
[1] 40152.33
> var_median
[1] 132334322
> # Print the mean and variance of the variable
> cat("Mean of Median:", mean_median, "\n")
Mean of Median: 40152.33
> cat("Variance of Median:", var_median, "\n")
Variance of Median: 132334322
```

Our mean and variance are not even close to equal which might explain why Poisson didn't work. We decide to try another method known as Negative Binomial Regression :

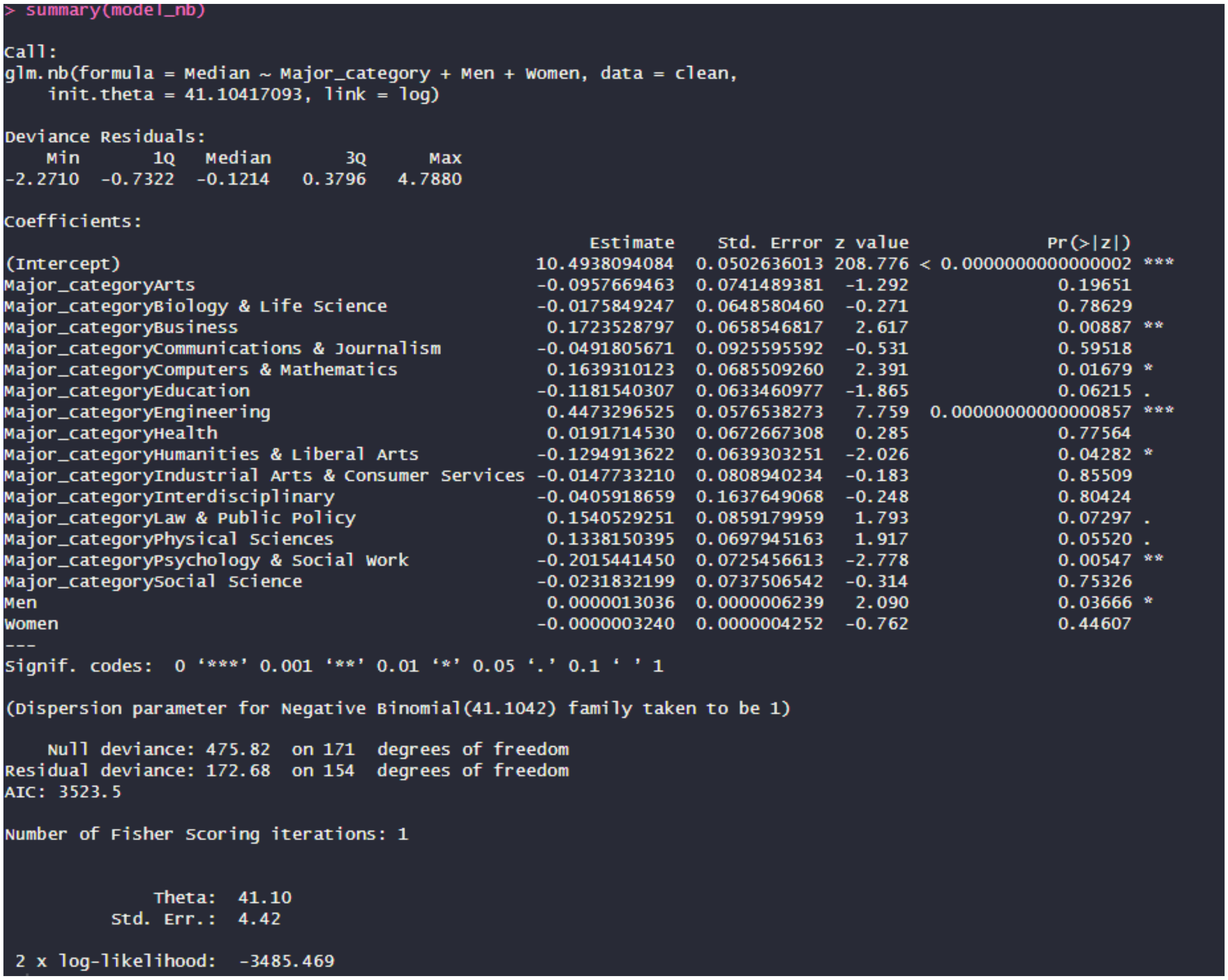
Our code looks like this



```
# Load the required library
library(MASS)

# Fit the Negative Binomial regression model
model_nb <- glm.nb(Median ~ Major_category + Men + Women, data = clean)

# Summarize the model
summary(model_nb)
```

DV - Median , IV - Major_category, Men & Women

Our result:



```
> summary(model_nb)

Call:
glm.nb(formula = Median ~ Major_category + Men + Women, data = clean,
    init.theta = 41.10417093, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2710  -0.7322  -0.1214   0.3796   4.7880

Coefficients:
                                                Estimate    Std. Error z value             Pr(>|z|)
(Intercept)                                   10.4938094084  0.0502636013 208.776 < 0.0000000000000002 ***
Major_categoryArts                            -0.0957669463  0.0741489381  -1.292              0.19651
Major_categoryBiology & Life Science          -0.0175849247  0.0648580460  -0.271              0.78629
Major_categoryBusiness                         0.1723528797  0.0658546817   2.617              0.00887 **
Major_categoryCommunications & Journalism     -0.0491805671  0.0925595592  -0.531              0.59518
Major_categoryComputers & Mathematics          0.1639310123  0.0685509260   2.391              0.01679 *
Major_categoryEducation                       -0.1181540307  0.0633460977  -1.865              0.06215 .
Major_categoryEngineering                      0.4473296525  0.0576538273   7.759  0.00000000000000857 ***
Major_categoryHealth                           0.0191714530  0.0672667308   0.285              0.77564
Major_categoryHumanities & Liberal Arts       -0.1294913622  0.0639303251  -2.026              0.04282 *
Major_categoryIndustrial Arts & Consumer Services -0.0147733210  0.0808940234  -0.183              0.85509
Major_categoryInterdisciplinary               -0.0405918659  0.1637649068  -0.248              0.80424
Major_categoryLaw & Public Policy              0.1540529251  0.0859179959   1.793              0.07297 .
Major_categoryPhysical Sciences                0.1338150395  0.0697945163   1.917              0.05520 .
Major_categoryPsychology & Social Work        -0.2015441450  0.0725456613  -2.778              0.00547 **
Major_categorySocial Science                  -0.0231832199  0.0737506542  -0.314              0.75326
Men                                            0.0000013036  0.0000006239   2.090              0.03666 *
Women                                         -0.0000003240  0.0000004252  -0.762              0.44607
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(41.1042) family taken to be 1)

    Null deviance: 475.82  on 171  degrees of freedom
Residual deviance: 172.68  on 154  degrees of freedom
AIC: 3523.5

Number of Fisher Scoring iterations: 1


              Theta:  41.10
          Std. Err.:  4.42

 2 x log-likelihood:  -3485.469
```

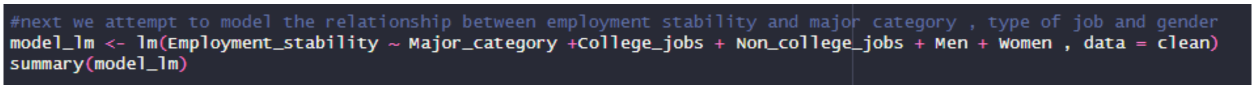
We can see from this result that it is very similar to what we had with a normal linear regression model and we can see that median income is adversely affected by major categories -

Positive correlation - Business,Computers & Mathematics,Engineering (More income)

Negative correlation - Humanities & Liberal Arts,Psychology & Social Work (Less income)

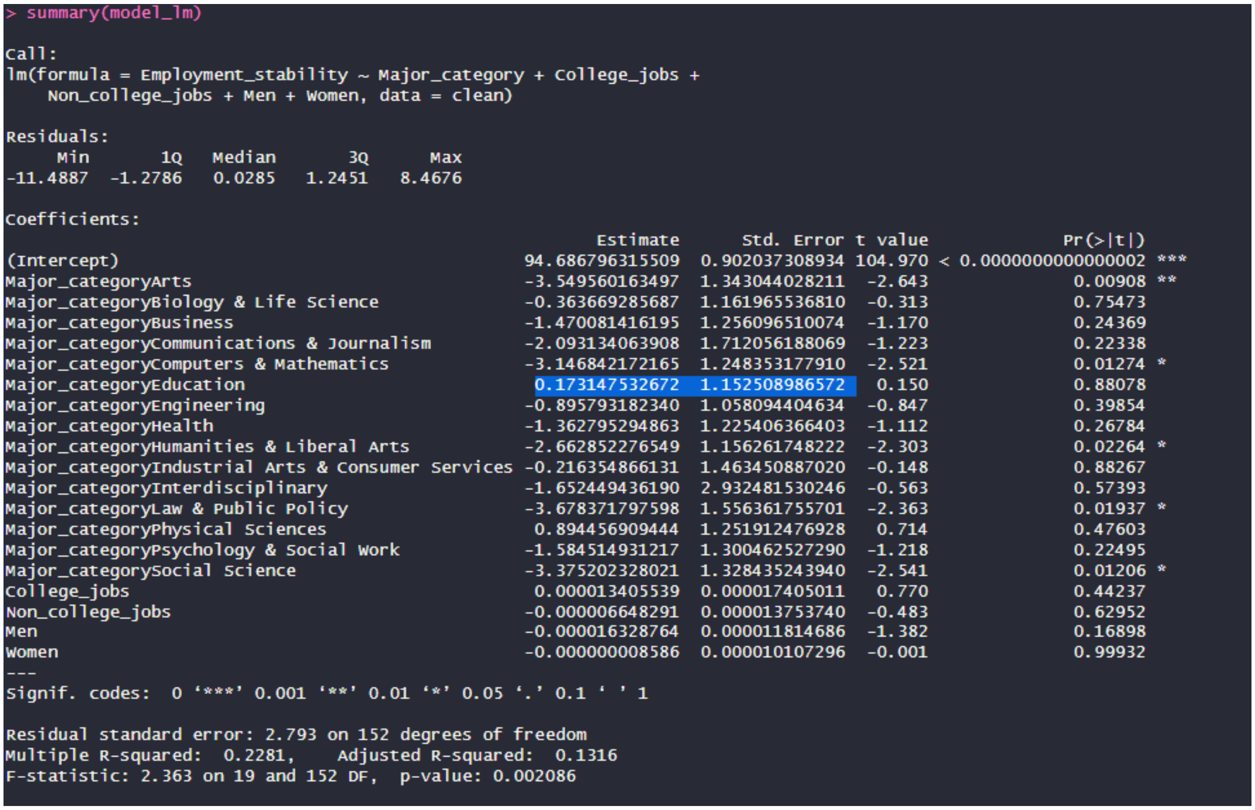
But we also see something interesting here , being classified as a man may indicated a very fractional increase in 1.3036×10^-4% in income(which is classified in thousands.

Next we attempt to model the relationship between employment stability, type of job and gender



```
#next we attempt to model the relationship between employment stability and major category , type of job and gender
model_lm <- lm(Employment_stability ~ Major_category +College_jobs + Non_college_jobs + Men + Women , data = clean)
summary(model_lm)
```

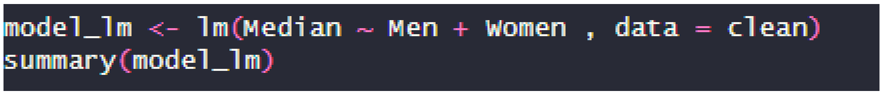
Our result:



```
> summary(model_lm)

Call:
lm(formula = Employment_stability ~ Major_category + College_jobs +
    Non_college_jobs + Men + Women, data = clean)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4887  -1.2786   0.0285   1.2451   8.4676

Coefficients:
                                                  Estimate      Std. Error t value              Pr(>|t|)
(Intercept)                                   94.686796315509  0.902037308934 104.970 < 0.0000000000000002 ***
Major_categoryArts                            -3.549560163497  1.343044028211  -2.643              0.00908 **
Major_categoryBiology & Life Science          -0.363669285687  1.161965536810  -0.313              0.75473
Major_categoryBusiness                        -1.470081416195  1.256096510074  -1.170              0.24369
Major_categoryCommunications & Journalism     -2.093134063908  1.712056188069  -1.223              0.22338
Major_categoryComputers & Mathematics         -3.146842172165  1.248353177910  -2.521              0.01274 *
Major_categoryEducation                        0.173147532672  1.152508986572   0.150              0.88078
Major_categoryEngineering                     -0.895793182340  1.058094404634  -0.847              0.39854
Major_categoryHealth                          -1.362795294863  1.225406366403  -1.112              0.26784
Major_categoryHumanities & Liberal Arts       -2.662852276549  1.156261748222  -2.303              0.02264 *
Major_categoryIndustrial Arts & Consumer Services -0.216354866131  1.463450887020  -0.148              0.88267
Major_categoryInterdisciplinary               -1.652449436190  2.932481530246  -0.563              0.57393
Major_categoryLaw & Public Policy             -3.678371797598  1.556361755701  -2.363              0.01937 *
Major_categoryPhysical Sciences                0.894456909444  1.251912476928   0.714              0.47603
Major_categoryPsychology & Social Work        -1.584514931217  1.300462527290  -1.218              0.22495
Major_categorySocial Science                  -3.375202328021  1.328435243940  -2.541              0.01206 *
College_jobs                                   0.000013405539  0.000017405011   0.770              0.44237
Non_college_jobs                              -0.000006648291  0.000013753740  -0.483              0.62952
Men                                           -0.000016328764  0.000011814686  -1.382              0.16898
Women                                         -0.000000008586  0.000010107296  -0.001              0.99932
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.793 on 152 degrees of freedom
Multiple R-squared:  0.2281,	Adjusted R-squared:  0.1316
F-statistic: 2.363 on 19 and 152 DF,  p-value: 0.002086
```

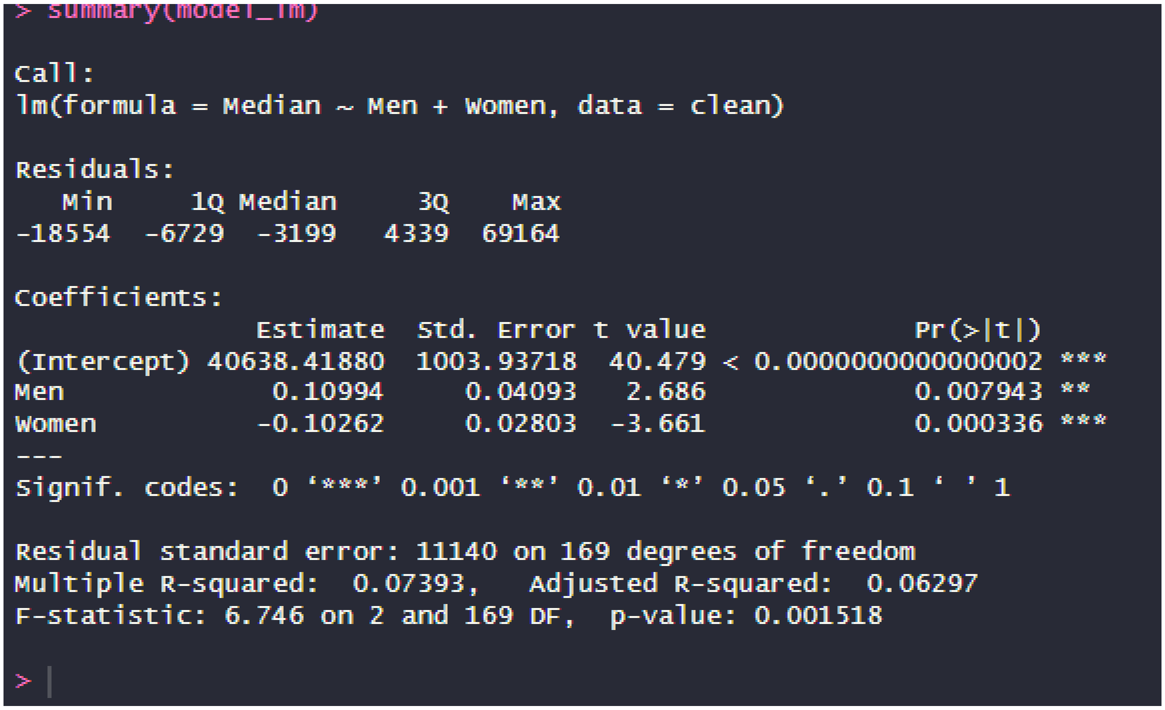
We find that when it comes to Employment stability , Arts is one of the worst when it comes to stability. Followed by Computer & Mathematics, Law & Public Policy and Social Science are bad.

We then double check our earning disparity between men and women with the following linear model:



```
model_lm <- lm(Median ~ Men + Women , data = clean)
summary(model_lm)
```

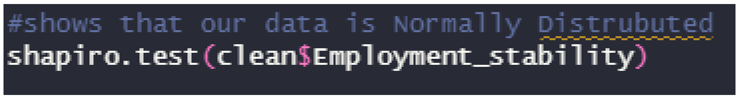
Our results are:



```
> summary(model_lm)

Call:
lm(formula = Median ~ Men + Women, data = clean)

Residuals:
   Min     1Q Median     3Q    Max 
-18554  -6729  -3199   4339  69164 

Coefficients:
               Estimate  Std. Error t value             Pr(>|t|)    
(Intercept) 40638.41880  1003.93718  40.479 < 0.0000000000000002 ***
Men             0.10994     0.04093   2.686             0.007943 ** 
Women          -0.10262     0.02803  -3.661             0.000336 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11140 on 169 degrees of freedom
Multiple R-squared:  0.07393,	Adjusted R-squared:  0.06297 
F-statistic: 6.746 on 2 and 169 DF,  p-value: 0.001518

> 
```

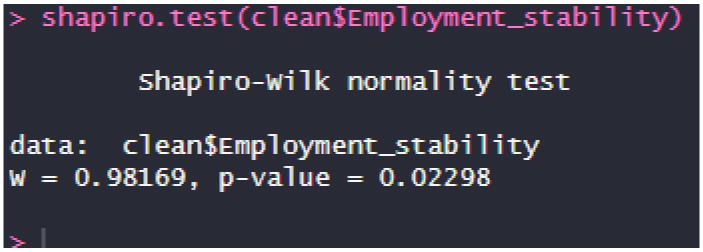
We find that women on average are likely to earn less but this is a very niche scenario and doesn't account for major_category which as we have seen before in the grand scale of things makes more of a difference.

To end , we use a shapiro wilk test to check for normality in Employment_stability



```
#shows that our data is Normally Distrubuted
shapiro.test(clean$Employment_stability)
```

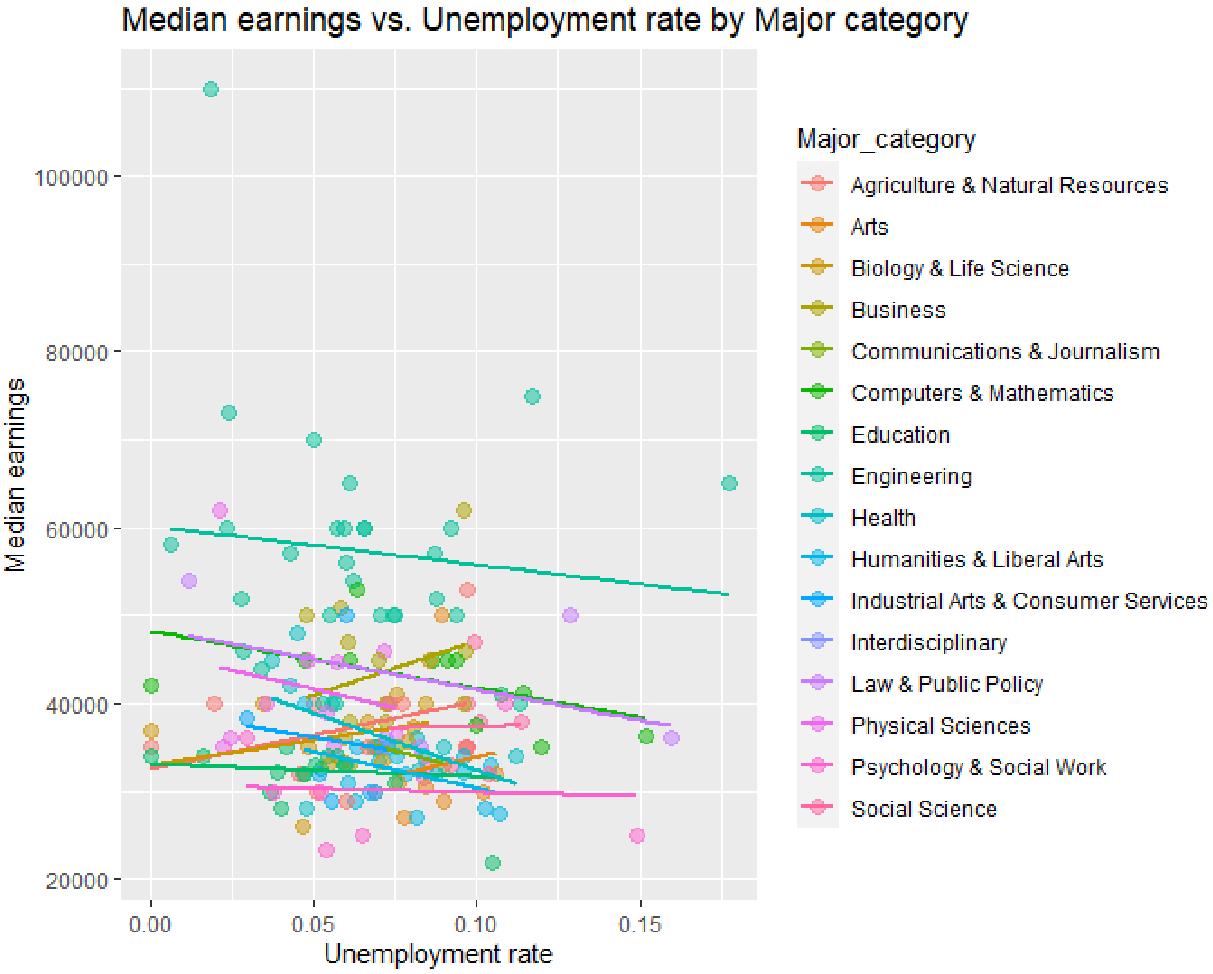
Our data is normally distributed



```
> shapiro.test(clean$Employment_stability)

	Shapiro-Wilk normality test

data:  clean$Employment_stability
W = 0.98169, p-value = 0.02298

> 
```

To conclude,

Our aim to find the most stable major categories has been achieved. We have also found the categories which have the highest median income.



And that turns out to be Engineering. We also find out that when compared with gender, Major_category is a more important variable when it comes to deciding median income. That is not to say the wage gaps do not exist, it is just that there are too many factors to get into to decide the matter. We find that when it comes to Employment stability , Arts is one of the worst when it comes to stability. Followed by Computer & Mathematics, Law & Public Policy and Social Science are slightly for the worse.

We hope this allows students to make better informed choices and pick degrees in which they are interested in. While some may be less stable than others or might even make less , it is important for students to follow their passion.