Gender wage gap: Exploratory Analysis

STAT515 Final Project

Why this data set?

A persistent wage gap exists between men and women even in this progressive century. The proportion of women entering professional fields of education and work have increased in the last few decades. However, the wage gap continues to be stubborn. This is an interesting topic as even though women's contributions have significantly increased, there are some factors due to which this wage gap continues to exist, and I wanted to explore these features like level of education, work force number, gender inequality, etc. to find out the root cause of it and to understand the severity of the problem. This would help in creating awareness, finding a solution and hence eventually lead to reforms for a better inclusive society.

Dataset description:

I have used US Census data. Some questions I intend to answer using this dataset are:

1. Is there a gender wage gap and if it exists, how severe is that gap?
2. Does any specific age group exhibit a higher wage gap?
3. What impact does the level of education have on the wage gap?
4. Apart from education, are there other factors like social stigma, lesser participation in the workforce or in a particular field of occupation, gender inequalities, etc. that has an impact on wage gap.

The dataset contains fields like:

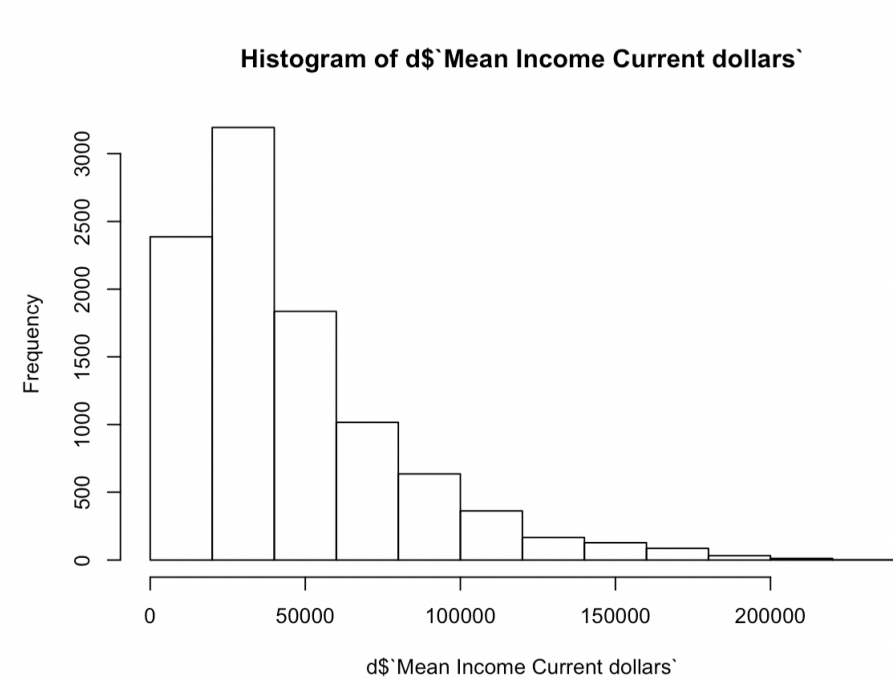
- Number in thousands for male and female working population
- Year
- Age group
- Income in Dollars
- Level of education attained
- Number in thousands for male and female in particular education level
- Occupation
- Gender, etc.

This dataset has been obtained from US Census Data in the form excel files by downloading each feature file from the year 1990 to 2020. Most of the data had multi-level index excel files. I transformed the data into an appropriate format, making it convenient to start with data exploratory analysis. Some steps followed post the above procedure for pre-processing are:

1. Checked for missing or NA values and replaced accordingly with mean values.
2. Checked for the structure of the data and performed data type conversion wherever required.
3. Converted columns to rows for Gender column as the data had separate columns and values for both Male and Female using gather function.
4. Finally, selected columns of interest for further data analysis.

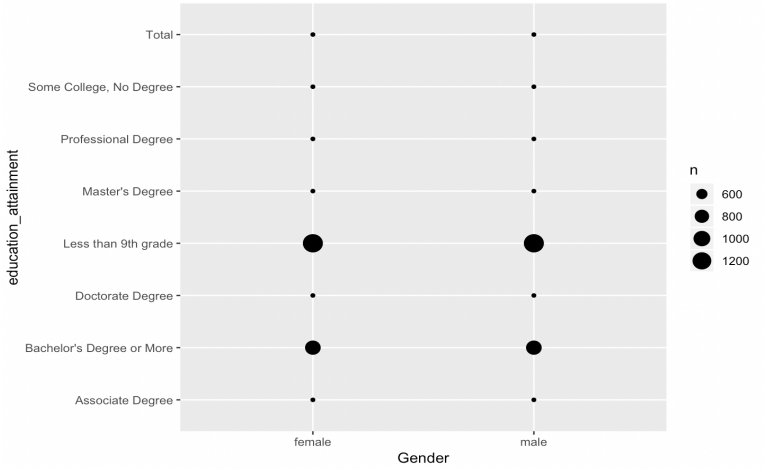
• What are some of the variables that will be addressed?

A continuous variable 'Mean income in dollars' exhibits skewness towards right



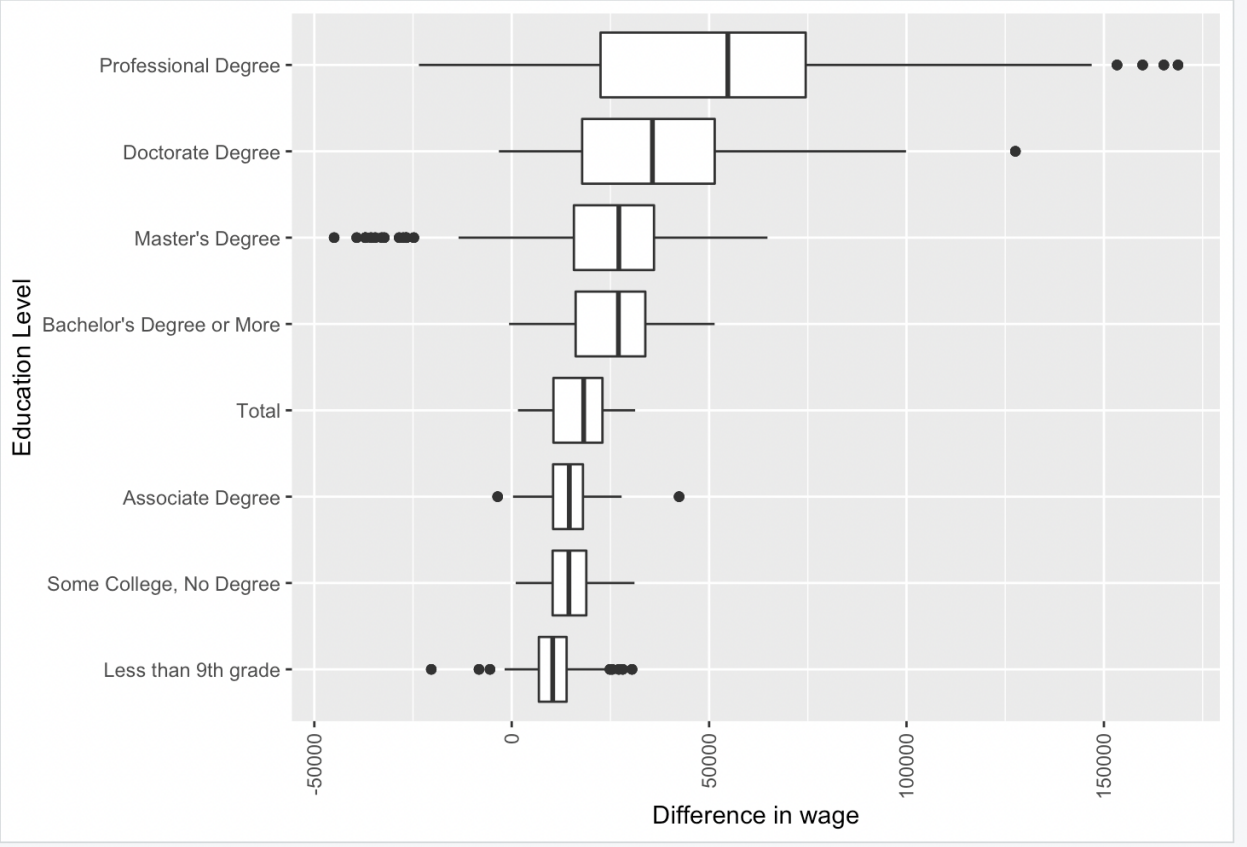
Further, checked skewness and kurtosis to see if it lies within limits.

Relation between two categorical variables show that there is a larger population of men and women that fall under the category of "lesser than 9th grade" level of education.



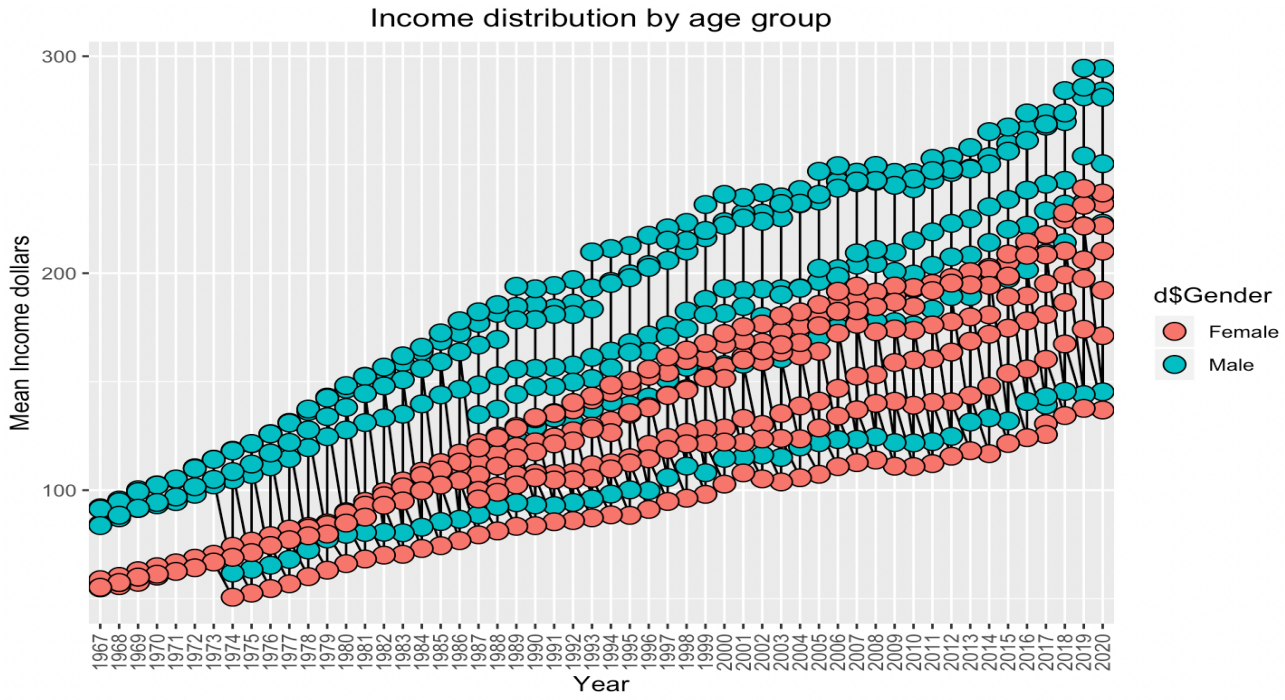
Covariate analysis:

This graph shows a descending order of difference in income between men and women over various levels of education. The median wage gap value is the highest for a Professional Degree and lowest for a person with less than 9th grade of education. It also exhibits outliers in this data. On checking for outliers, the population with professional degree exhibit a higher income level and it is difficult to judge if that value is correct or not from the dataset.

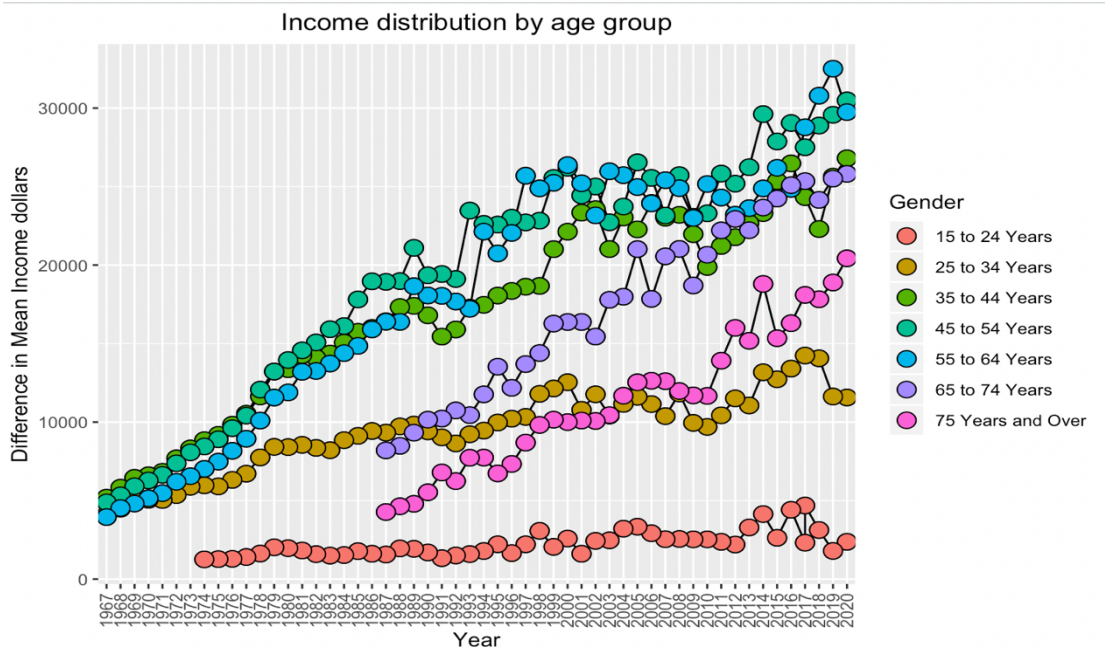


Below is a box plot of mean income vs gender. Men exhibit a higher median in mean income as compared to women.

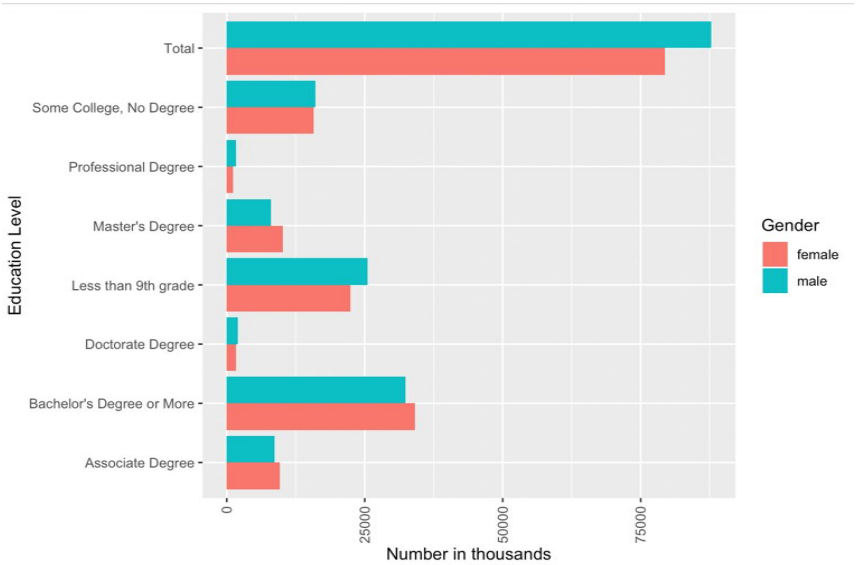
Below is a graph of mean income by year. Men exhibit a higher wage than women and the trend seems to follow over the years.



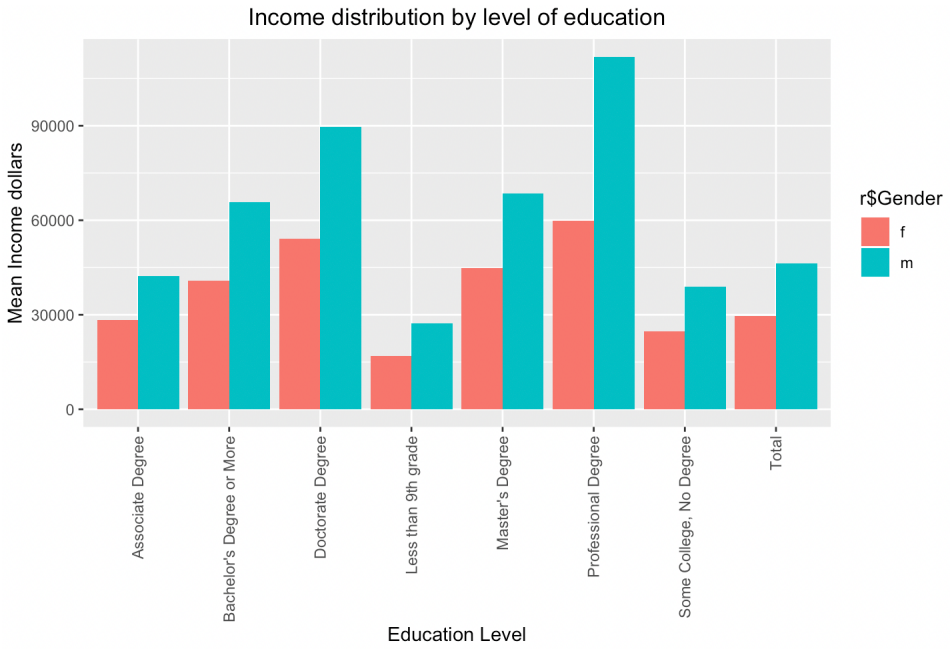
Below is a graph of difference mean income vs year over different age groups. The age groups 35 to 64 exhibit the highest wage gap difference.



Number of Men and women enrolled in each level of education. This shows higher number of men have enrolled in each field of education as compared to women. However, the difference is marginal.

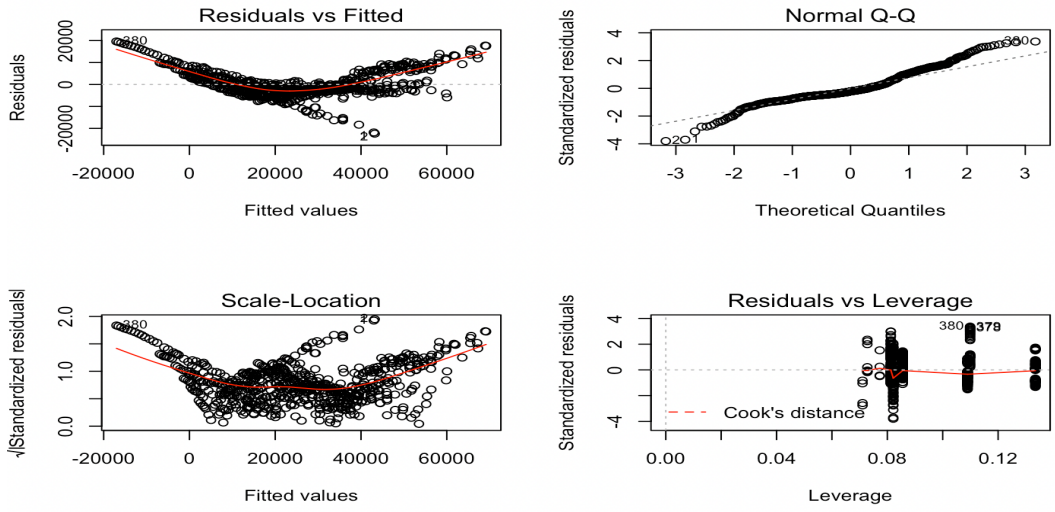


The graph below shows income distribution per level of education per for men and women.



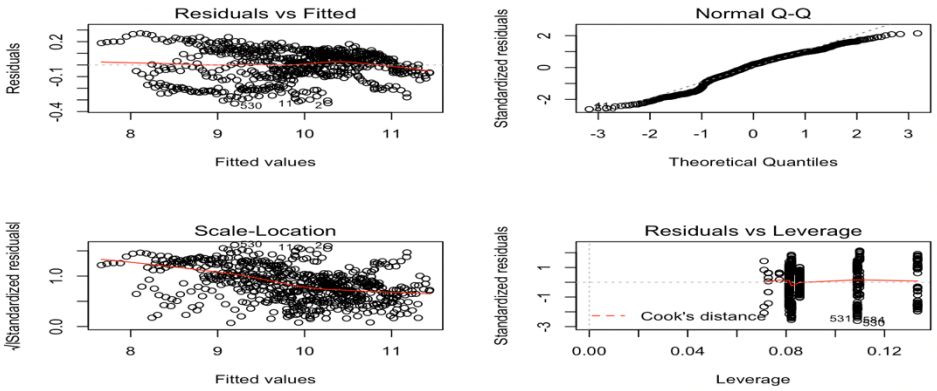
Model building:

Linear regression model to understand relation between mean income and all other features. Below are the diagnostic plots to understand the four properties Linearity, normality in error distribution, homoscedasticity, and to understand if outliers or leverage points exist, respectively.



Adjusted R-squared: 88.26%

After performing log transformation:



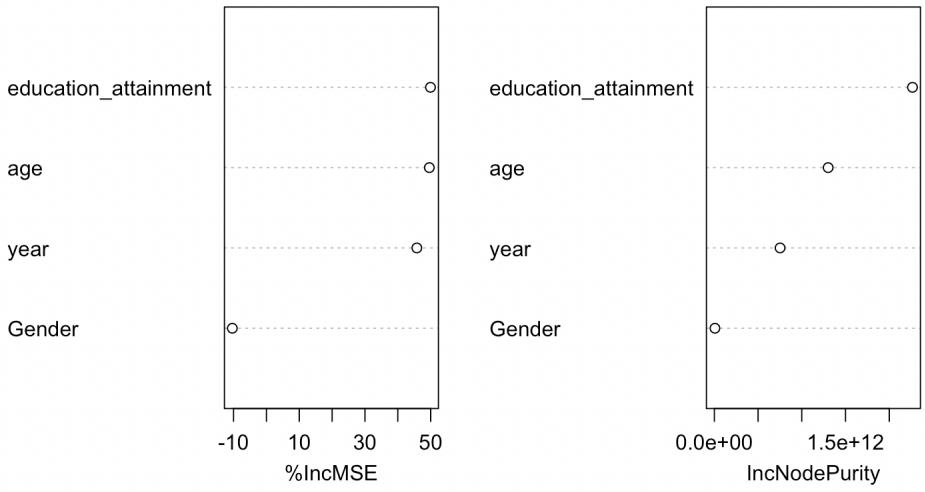
Adjusted R-squared: 97.08%

The Adjusted R-squared: 88.26%

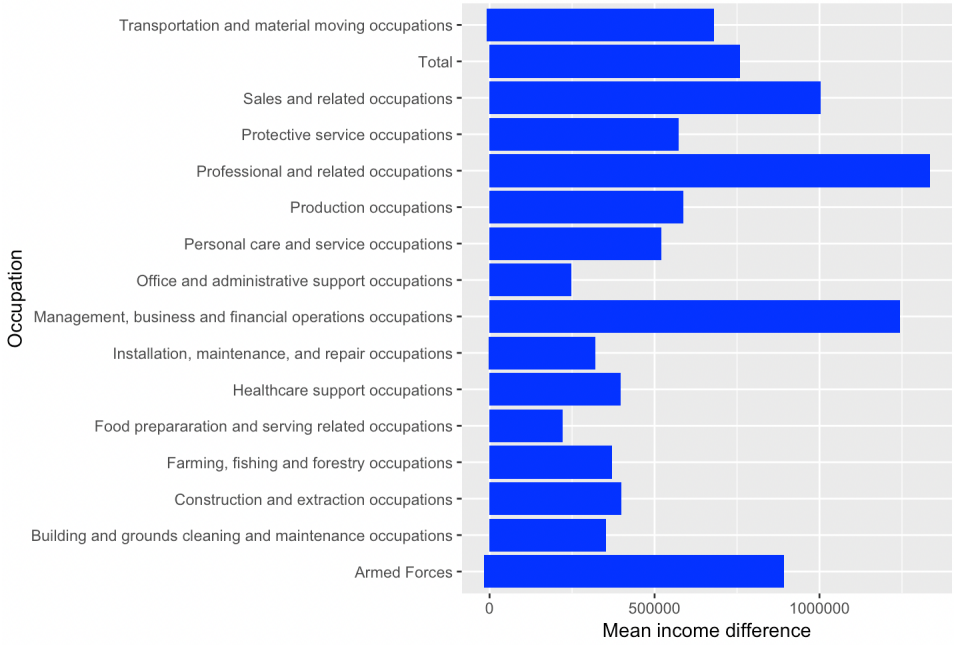
On performing log transformation, the accuracy of the model improved to 97.08%

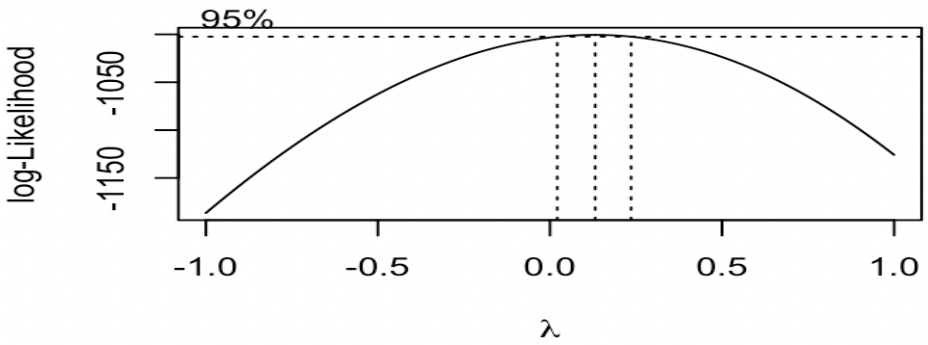
Random forest model for regression for subset selection: The below figure indicates the features education level, age, year, and gender in their order of importance, with gender having the least impact on accuracy of the model of removing

rf.Income



Occupation





After boxcox transformation with lambda value as 0.02, accuracy increases from 84.2% to 85.6%.

Conclusion:

1. Men exhibit a higher wage than women over the years from 1990 to 2020. Gender based Wage gap exists and with 2013 and 2017 years having highest gap.
2. The age groups 35 to 64 exhibit the highest wage gap difference.
3. Even though the number of men and women completed education in each level are similar, there is a huge gap in income earned between men and women, highest in professional degree.
4. Professional level of education highlighted outliers, but it is difficult to judge if that income is incorrect or not, just using this dataset.

Future scope:

This dataset does not include social stigma parameters and other socio-economic parameters which could further explain the wage gap in detail. Understanding these factors could help us identify a solution and become aware of the problem of inequality that exists in society and ensure a better future for the coming generation.

References:

[1]. US Census Bureau. (2021, November 9). Historical Income Tables: People. Census.Gov. Retrieved December 3, 2021, from https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-people.html