## horizontal line



High Income Determinants

York University - CSDA 1010 - Group 1

# 

# 

# Abstract

The American Dream is a tale of financial freedom on the premise that anyone can make it. However, in a Capitalist society, this dream will never be true for everyone as it is an inherently exploitative system. As is demonstrated in the *Adult Census Income* dataset, only a few get to enjoy this dream. In this project, we set out to predict the common factors that contribute to the high income bracket. A Decision Tree, Hierarchical Clustering, K-Medoid Clustering and an OLS Linear Regression model was used to achieve this.

# Business Understanding

Background

In 1994, the average income was $23,753.53, (Social Security, 2019). Therefore, earning over $50,000 a year, would put a person into the middle class category. In as much as the average income has risen to $48, 672 in 2019, the people who make up that category have not changed much, (Bureau of Labor Statistics, 2020). Although this dataset is outdated, the data is still relevant.

According to the Pew Research Center, half of all the income in the U.S was only made by 20% of families in 2018. This number has slowly increased from 1968, where 20% of families owned 43% percent of all the income. In those 50 years, the wealth gap has actually increased as opposed to reduced. Although the Women’s movement, Civil Rights, LGBTQ and even Occupy Wall Street all occurred during this time period, the change that matters the most has yet to occur. Access to a better quality of life.

Some have referred to the COVID-19 pandemic as the great equalizer, however, we have seen that people who are in servitude roles are the ones who are most at risk (CNBC, 2020). People in the highest income brackets, with the exception of Doctors, are not in this category. Unfortunately, not even a global pandemic will bring about a change in income inequality.

Business Objective

The objective of this project is to determine the common factors that contribute to the high income bracket, (>$50,000).

In order to achieve our objective, the most optimal model will be selected after being built, analyzed and compared. The target variable is *Income*, and our subject variables are *Age, WorkClass, Education, Marital Status, Race, Gender, Sex, Hours over Week, Capital Gain and Capital Loss.*

Hypothesis & Assumption

The key assumption is that middle - aged, white males with post-secondary education make up the largest percentage of those who earn over $50k.

**Data understanding**

Ethics

In this project, we commit to continuously develop processes that allow me to understand, document and monitor bias in development and production.

After the data was cleaned, we let the results tell the story as opposed to forcing any conclusions.

Key Metrics of Success

A Correlation Test Table was used to test the efficiency of the OLS Linear Regression model. Both hierarchical clustering and K-Medoid clustering methods were validated with the use of silhouette width, which are internal measures of validation. Accuracy, balanced accuracy, recall and precision were used to test the efficiency of the decision tree.

Data Selection & Preparation

We used the “Adult Census Income” dataset from the Kaggle repository:

<https://www.kaggle.com/uciml/adult-census-income>

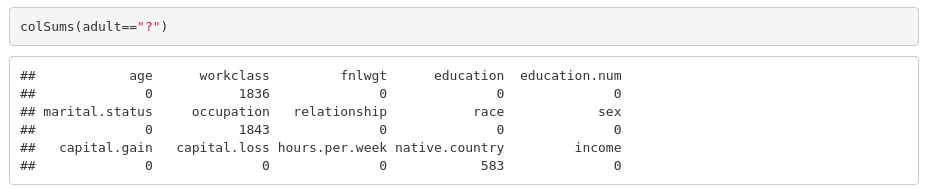
which contains data extracted from the 1994 United States Census Bureau database, and consists of 15 features and 32,561 observations. Most of the features (nine) are categorical while the remaining six are numerical, Table 1.

Table 1. Structure of investigated dataset.

|  |  |
| --- | --- |
| ***Feature*** | ***Structure*** |
| age | continuous numerical |
| workclass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked |
| fnlwgt | continuous numerical |
| education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| education-num | continuous numerical |
| marital.status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| occupation | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces. |
| relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| sex | Female, Male |
| capital.gain | continuous numerical |
| capital.loss | continuous numerical |
| hours.per.week | continuous numerical |
| native.country | United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands. |
| income | >50K, <=50K |

Most of the features are self-explanatory, and are related to various socio-economic parameters. As for the non-obvious ones, the “fnlwgt” feature (final weight) contains weighting factors for the dataset, wherein the data from each sample person is weighted by the inverse of the probability of the person being in the sample in order to provide a rough measure of the number of actual persons that the sample person represents (US Census Bureau, 2015). Education-num feature is denoting the total number of years of education, “relationship” feature relates to the different family roles, while “capital\_gain” and “capital\_loss” capture the investment income, other than salary (wages). Obviously, the “income” feature is our main target variable, consisting of two categories - low(er) income bracket (<=50K), and high(er) income category (>50K).

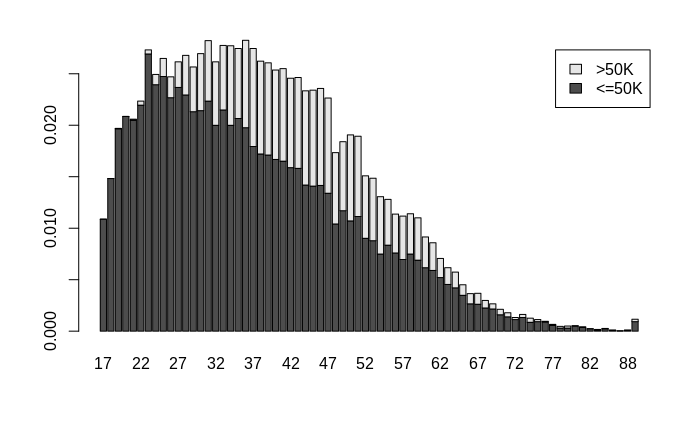
Few features, namely workclass, occupation, and native/country contain a relatively small number (compared to the total number of observations) of missing values, labeled with “?”.



**Image 1. Checking for missing data.**

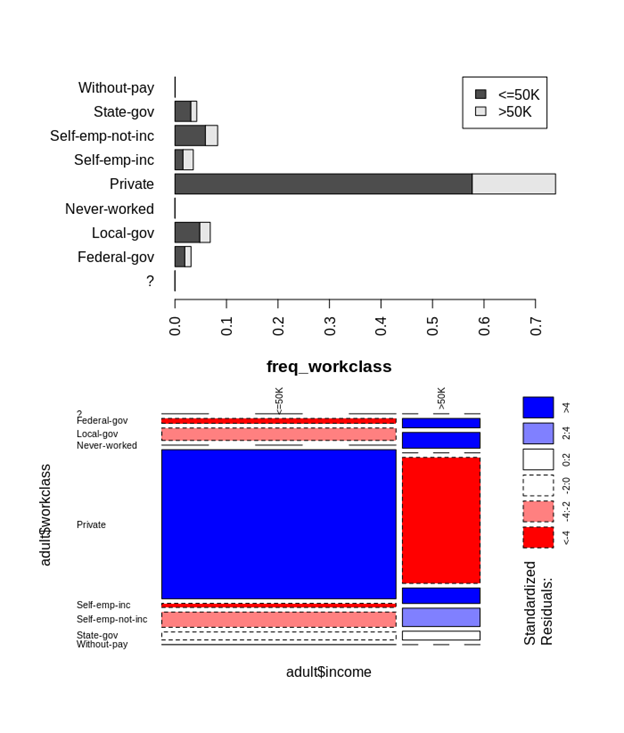
Therefore, we removed all rows which contained missing values, and used the cleaned up dataset for visualizations. Here, we will present some of the most relevant insights derived from data visualization exercise, the rest of visualizations can be found in the R markdown. We used the proportion barplots and mosaic plots with standardized residuals, as a practical and efficient way to capture the prominent trends of “income” feature, against other select features.

Fig. 1. shows the distribution of two income brackets (>50K and <=50K) as a function of age. It can be observed that the proportion of the higher income bracket is higher within the peak work productivity age range, approx. from 30 to 55 years. Younger age ranges are dominated by lower income earners, reflecting the entry level/lower pay jobs in that age bracket. Similarly, older age is also dominated by lower income individuals, reflecting the reduced productivity and dependence on more modest types of incomes, e.g. pensions.



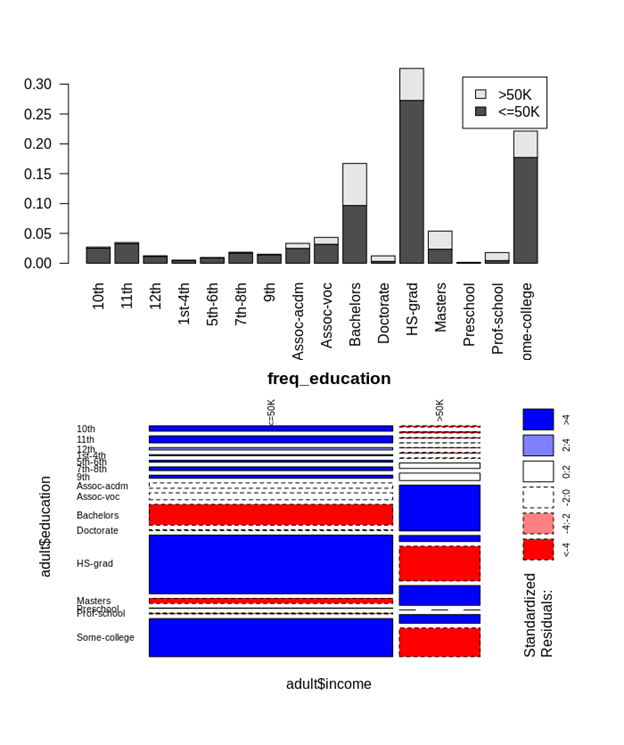
**Figure 1**. Proportion barplot of income as a function of age.

With respect to type of employment (“workclass” feature) the dataset is dominated by individuals working in the private sector; at the same time there is the higherst proportion of low income earners within that sector, Fig. 2.



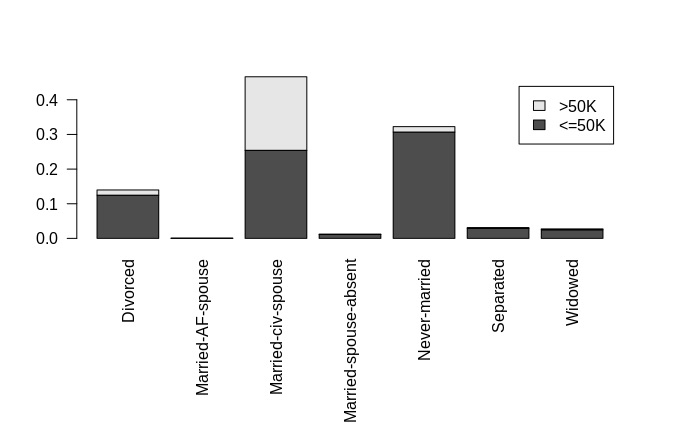
**Figure 2.** Proportion barplot and mosaic plot with standardized residuals, showing the income distributions among different types of employment.

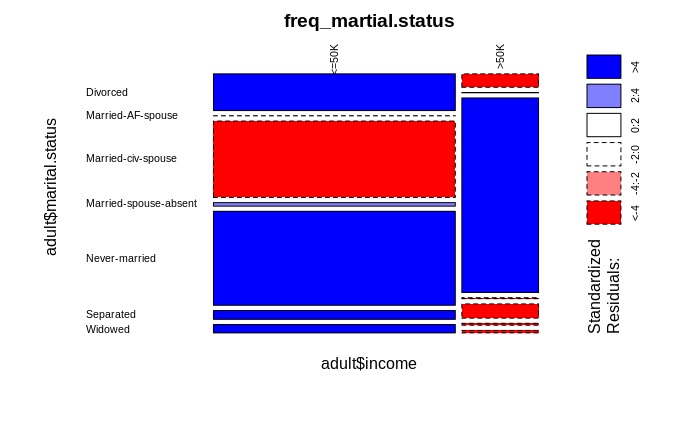
Education level also stood out as a relevant factor for income distribution, Fig. 3, we observed a clear trend of higher proportion of >50K income earners among more educated respondents, namely ones holding Bachelor, and higher degrees. On the other hand, the subjects that have high school or some college education are predominantly earning <=50K.



**Figure 3. Income distribution for different education levels**

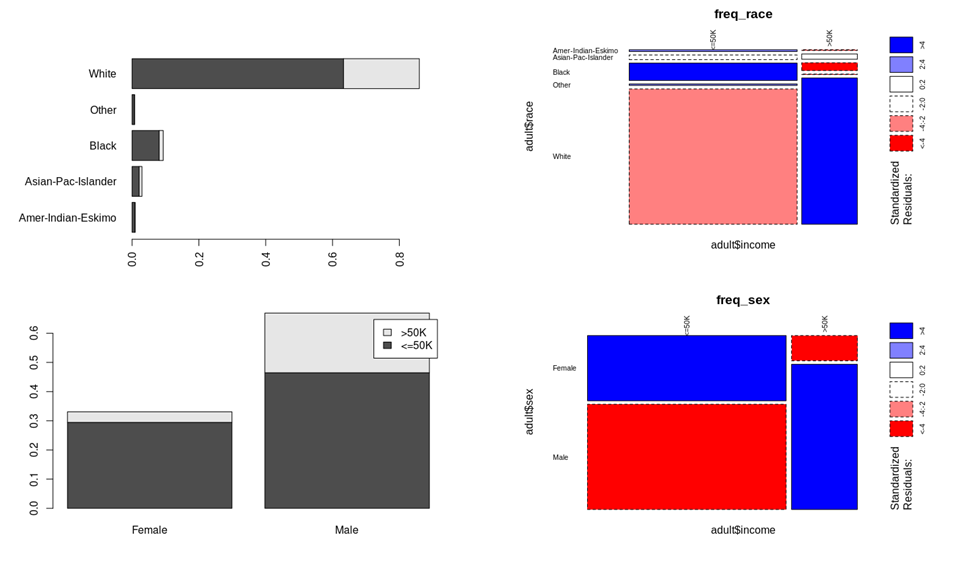
Interestingly, marital status is also showing some relation to the income, there is a clear predominance of higher income subjects among the married population, compared to single, or divorced individuals, Fig. 4.





**Figure 4. Income as a function of marital status.**

Finally, we discover an indication of racial and gender bias, as the higher income bracket was clearly dominated by white and male individuals, Fig. 5. This is likely a reflection of inequitable hiring and compensation policies; given that the dataset is based on the 1994 census, as a part of future work it would be interesting to compare this dataset with more recent data to investigate whether there was any improvement on this front in the past two decades.



**Figure 5**. The income distribution based on race and gender.

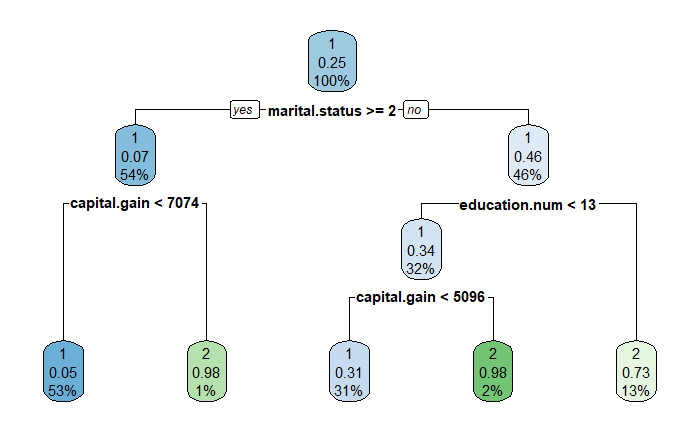
**Feature Engineering**

As we had a large dataset of over 32000 observations, sampling had to be done in order to run the clustering algorithms on our machines. To do this we plotted histograms of the full dataset and compared them with histograms of 10% of the dataset. The plots can be found in the Appendix. The distribution of the data was almost identical for each variable. This gave us confidence that models applied to 10% of the data would have similar results as applying the models to the entire dataset. In order to avoid observations being clustered based on insignificant variables, as well as to cut down on data processing time, a correlation plot was created to identify which were least correlated to income, those variables were dropped. The variables that remained were age, hours per week, education num, marital status, capital gain, capital loss, sex, income and relationship. For the hierarchical clustering, the data was converted to numeric and normalized. For the K-Medoid method of clustering, all non-numeric variables were converted to factors and then normalized.

**Modelling**

**Decision Tree**

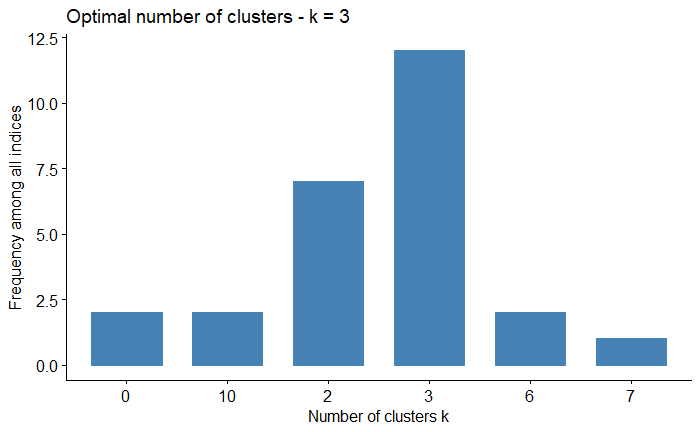
One of the models used to predict whether or not a person was a high earner of income was the decision tree. We used a weighted decision tree algorithm with the “fnlwgt” feature as the weighting factor. According to the tree there are three main variables that predict how much somebody earns: marital status, education level and capital gain. The tree below initially splits the dataset based on marital status. If the marital status is married to a civilian (1), the people went down to the right side of the tree. This was 46% of the population. All people with another marital status went down to the left side of the tree. The 54% on the left were split one more time by the capital gain variable (< 7074). The 46% that went down the right were split by the education level of ‘<13’ which means they were divided by whether or not they had a higher education than a Bachelor’s degree (level 13 in our dataset). Those with the Bachelor’s degree or below were further divided by capital gain. At the bottom of the tree, we are left with 5 end nodes totaling 100% of the dataset. The blue nodes with the value ‘1’ are those who earned <=50K, the green nodes with the value ‘2’ are the high earners. When added together, 84% of the dataset made less than or equal to $50,000 and the remaining 16% earned more than $50,000. The importance of education as a predictor of income can be seen because 13% of the population instantly are predicted to fall into the >50K group immediately after that split. Marital status is possibly the greatest determinant as it instantly takes just over half of the population and can predict they will fall into the <=50K group. Capital gain appears to be what separates the high earners from groups that are heavily populated with lower earners as it extracted and placed 1% and 2% of the population into the high-income group each time it was used to split the dataset. The decision tree had a precision of 78.8%, recall of 52.7%, accuracy of 84.7%, and a balanced accuracy of 74%.



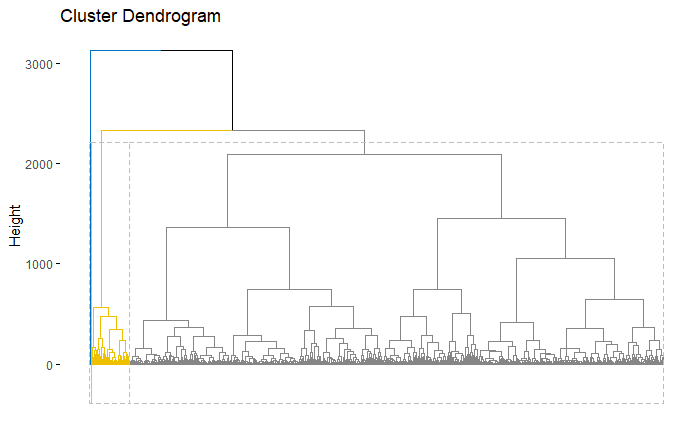
**Figure 6.** Decision Tree predicting level of income

**Hierarchical Clustering**

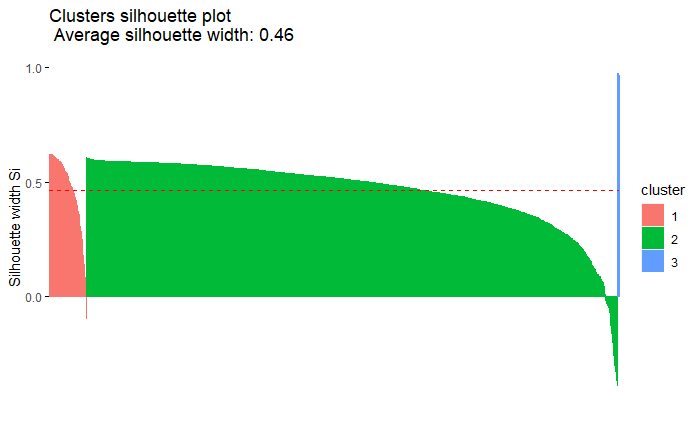
Hierarchical clustering is an unsupervised learning algorithm that is used to identify similarities in data and place them into groups. Visualization for this method is displayed in a dendrogram, which essentially is a large tree. We were hoping to use this method of clustering to identify similarities between the members of the population who earned more than $50,000. Using Euclidean distance measures, it was determined that 3 clusters would be optimal for the dendrogram. The resulting dendrogram showed one very large cluster (grey), a smaller cluster (yellow) and another one (blue) that was extremely small which indicates the data is not very well clustered.



**Figure 7.** Barplot determining optimal number of clusters



**Figure 8.** Hierarchical clustering dendrogram

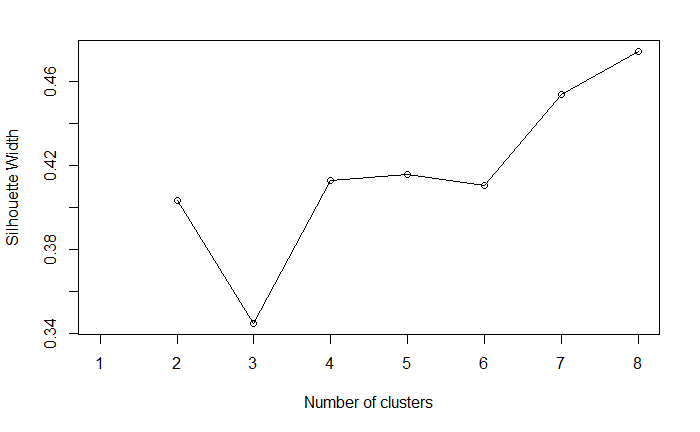


**Figure 9.** Silhouette plot for cluster validation

The silhouette plot confirms this as the average silhouette width was 0.46, which is not very high. Ideally clusters would be closer to +1 with similar silhouette shapes but the 3 clusters, as previously observed in the dendrogram, are very different in size. This makes for difficult analysis and the conclusions drawn from this method of clustering are unlikely to be valid.

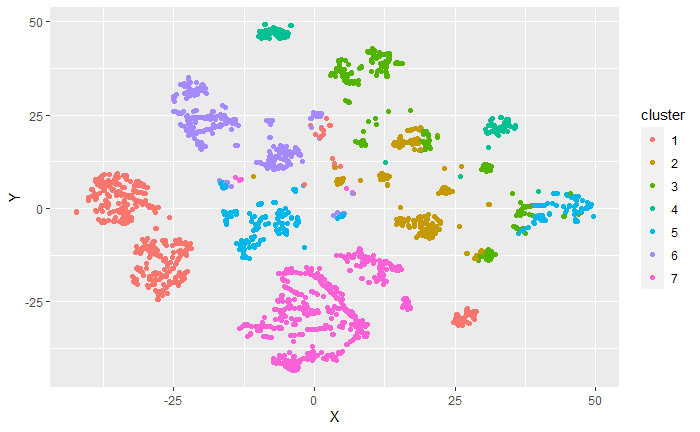
**K-Medoid Clustering**

This version of clustering involves dividing a dataset into ‘k groups.’ The number of groups is determined before implementing the algorithm, it cannot be used without knowing how many groups the data must be divided into. The process for the K-Medoid model began using Gower distance and matrix on the mixed data set. After implementing the Gower formula, we went to Partitioning Around Medoids to identify the ideal number of clusters. It was decided to use 7 clusters because it is slightly simpler than 8, and the silhouette width is only slightly less than an 8-cluster plot.



**Figure 10.** Silhouette plot for cluster analysis

The results, which can be found in the Appendix, indicate that 79.3% of the high earners (>50K) made up the entire population of Cluster 1. It is clear that this was the biggest factor in grouping these people. Using the statistical analysis provided below, it can be seen that all members in the cluster are males and 94.8% of them are married to a civilian spouse. This cluster also led the others in having the highest mean values in the following categories: education level, hours worked per week, capital gain and capital loss. The high mean of capital loss is likely due to high earners having the ability to purchase, trade and sell assets while those who earn less tend to be more conservative with their money.



**Figure 11.** Cluster plot for K-Medoid clustering method

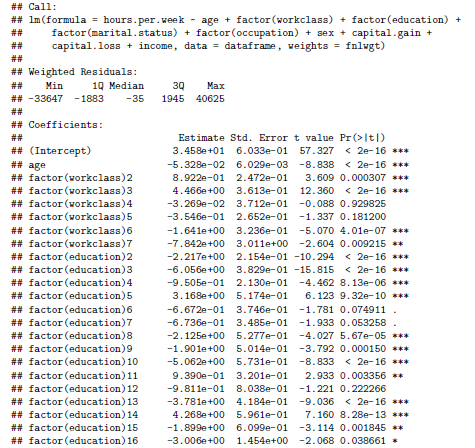
It also appears in the cluster plot itself that Cluster 1 is one of the best clustered groups, allowing us to safely identify some key characteristics shared by high earners. While the average silhouette width for both clustering techniques is about the same, the K-Medoid technique had more proportionate clusters of which the majority were well grouped. Ultimately, it appears the K-Medoid Clustering method was more accurate than the hierarchical method.

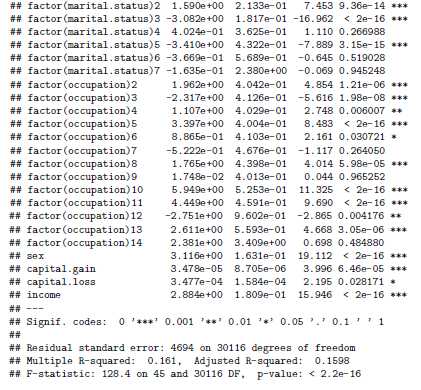
**OLS Linear Regression**

As the data set denotes the income variable as only a binary variable splitting those with under $50,000 dollars in income from those above, our group has turned to using linear regressions to analyze other potential determinants of income. The chosen variable to be used in the linear regression is the hours per week variable. This was chosen as for many workers the number of hours worked by an individual would have a direct impact on their yearly income. With this in mind the linear regression should determine what variables contribute to determining the hours individuals. The regression takes the form:

Hours per week ~ age + work class + education + marital status + occupation + sex + capital gain + capital loss + income + error

While age, capital gain and capital losses are continuous variables, the other variables are all categorical. To better capture the impact of individual categories they have been split into several series of binary variables. This would allow us to determine the precise difference in impact of specific levels of education or occupation for example differs from another rather than a flat rate. In the original design of the model, race and native country variables were included but were taken out as it had created a large number of additional binary variables, most of which were not statistically significant. With this in mind they were excluded from the calculations and had also resulted in a model with a higher adjusted R-square. When running the regression, these results are found.



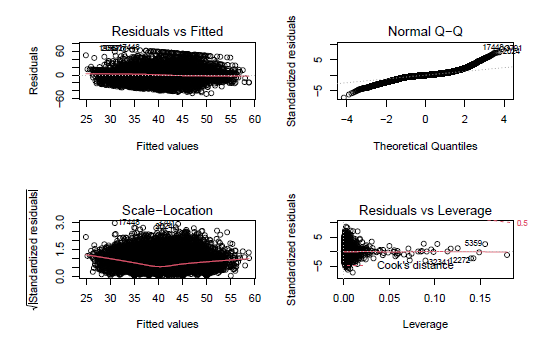
****

**Image 2.1 & 2.2.** Results of Linear Regression

Based on these results, many of the variables have a statistically significant impact in determining the number of hours worked by individuals. The first variable of note is age, which is found to have a negative impact on hours worked, suggesting that as individuals age, they work less. The results for the worker class have found that those who were self-employed (factor 2 & 3) worked more than those employed in the private sector. As well, the results suggest that those in the public sector (factor 4&5) work less hours but these results were not statistically significant. For education many of the results have a statistically significant negative impact on hours worked, these are mostly due to them at a level of education below high school, which mostly be made up of individuals with few hours worked. For those with a high school education and above, high school graduates and those with some college education work relatively less to those with bachelors. The results also show that those with graduate levels of education also work more hours.

The results also suggest that marital status also plays a role in determining the hours of work per week. In particular those who are divorced (factor 2) tend to work more and those who were never married (factor 3) work fewer hours. While the occupation types had a statistically significant impact, the exceptions were for cleaners, clerical and armed forces workers (factor 7, 9, 14). While most of the occupations had a positive impact on hours worked, other services and private house services (factor 3 & 12) had a negative impact on hours worked. Sex was also statistically significant with a negative impact on hours worked per week. As well both capital gains and losses were found to have a positive impact on hours per week worked and income was also positive.

To determine the effectiveness of the linear regression it must be determined if the linear regression is in line with the assumptions of the OLS regression model. The data does not seem to have demonstrated any multicollinearity as suggested by the correlation test table.



**Figure 12.** Data evaluation

The residual vs fitted values plot does not have any shape or funnel which would suggest that there is no heteroskedasticity in the data. The scale location also does not seem to follow any patterns which would further confirm that there is no heteroskedasticity. The normal Q-Q plot does display some level of distortion from being a straight line, which may suggest that the residuals have some non-normal distribution.

In assessing the effectiveness of the model, the regression has an adjusted R-square of 0.1598 which means much of the variation in hours per week worked remains unexplained by the model. To further test the model’s effectiveness the model has been recreated with a training data set and used to predict a test data set made from 70% and 30% of the original data respectively. When using the regression created with the train section of the data set to predict the hours worked of the test data set, there is a RMSE of 10.98665. The hours per week variable had values ranging from 1-99, which would mean that the RMSE is a relatively large number, this result along with the low adjusted R-square may suggest that the regression is a poor fit for the purposes of estimating the hours individuals work per week.

**Recommendations**

1. We recommend having more comprehensive data that covers Degree level, and Degree type which would lead to being able to analyze what degree types will generate more income, moreover the question that still stands would be “does having a higher education necessarily mean more income or individuals are going to face the common over qualification?” Moreover, it is a common belief that specific universities are more credible and degrees from them will result in better jobs. This can also be verified or denied based on the comprehensive data.
2. K-Medoid reveals most >50K fell into cluster 1: all people in the cluster were male, most were husbands, this cluster had the highest avg. education level, capital gain and loss, and highest avg overall. Hours worked suggesting these are the greatest predictors of income. Education level comes up in all the models Husbands of civilians covers a lot of individuals. More details about this cluster can improve our data findings by a significant amount. There might be a lot of factors regarding their wives that could impact the income of the husbands e.g. the civilian job,education, and level of support... etc.

**Conclusion**

In the analysis of the decision tree and K-Method, Education was viewed as a key deciding factor for income. Moreover, in the clustering the key factors were education and marital status, since people with more than $50,000 income were mostly husbands married to a civilian.

With education making the most appearance as a deciding factor for income in the different methods, it can be safely assumed that the most common factor is education. The United States had the most income inequality of all of the G7 countries (Pew Research 2020). This might be due to differences of access to education throughout the country which could be caused by: location, personal decision, community beliefs or financial situations. With the legacy tradition in Ivy League Colleges and lack of funding in inner city schools the issues in Education in the U.S are almost insurmountable. This would take a decree of dedicated programs by the governing bodies to educate the lower income families, and persons to make the right decisions as to which schools to attend and what education to pursue.

**References**

US Census Bureau, 2015, Weighting,

<https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/weighting.html>

National Average Wage Index

<https://www.ssa.gov/oact/COLA/AWI.html>

6 facts about economic inequality in the U.S.

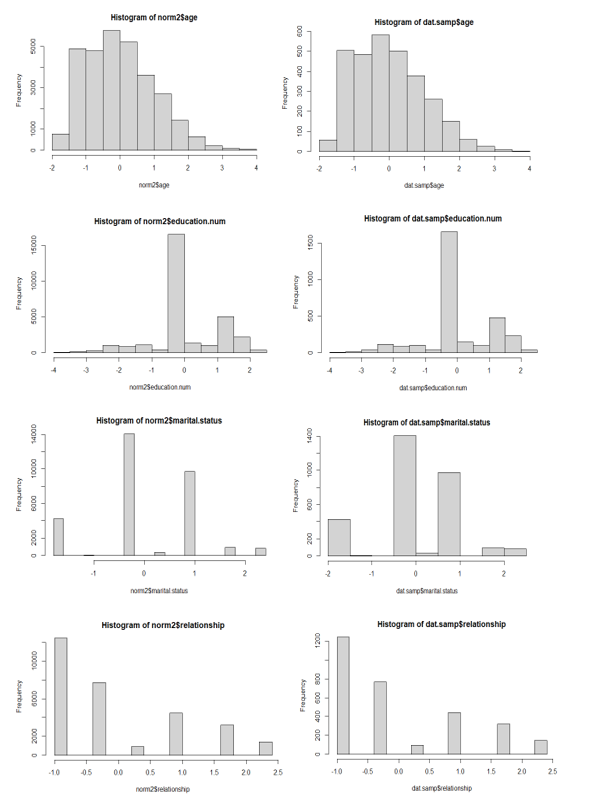
<https://www.pewresearch.org/fact-tank/2020/02/07/6-facts-about-economic-inequality-in-the-u-s/>

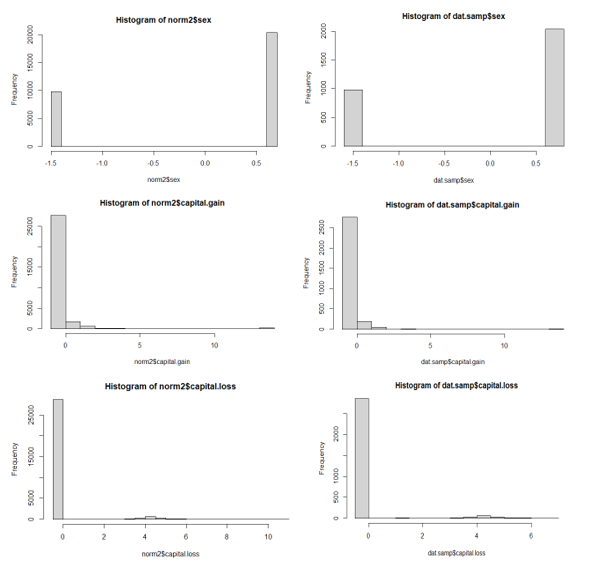
These are the most dangerous jobs in America in the age of coronavirus

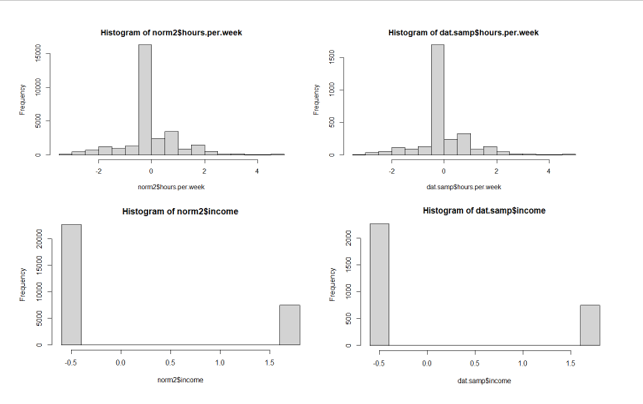
<https://www.cnbc.com/2020/05/30/here-are-the-most-dangerous-jobs-in-america-during-coronavirus.html>

**Appendix**

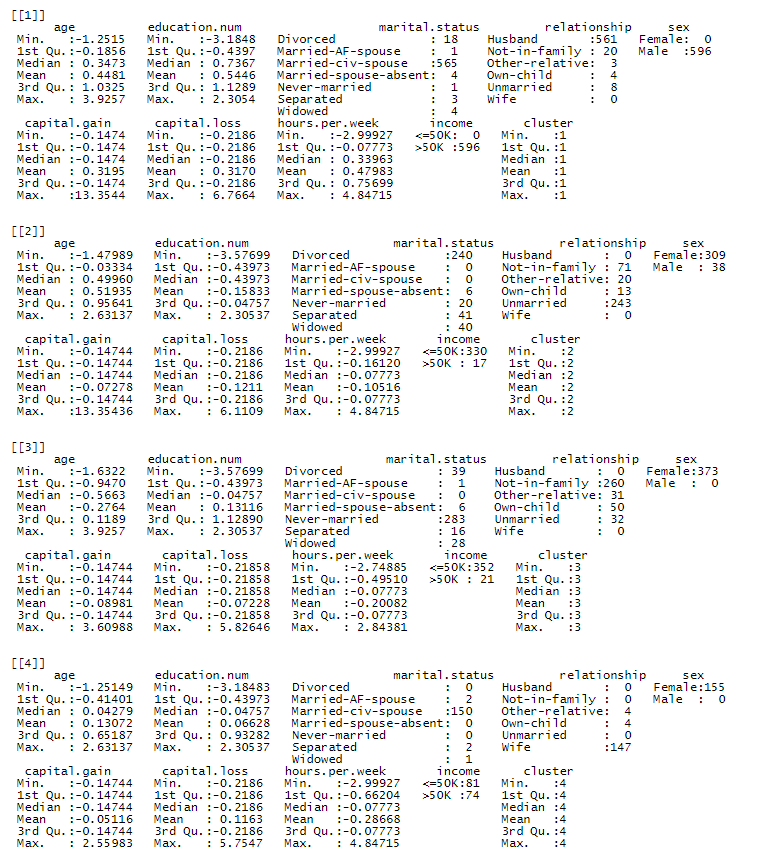
**Histograms comparing distribution between dataset (left) and 10% sample (right)**

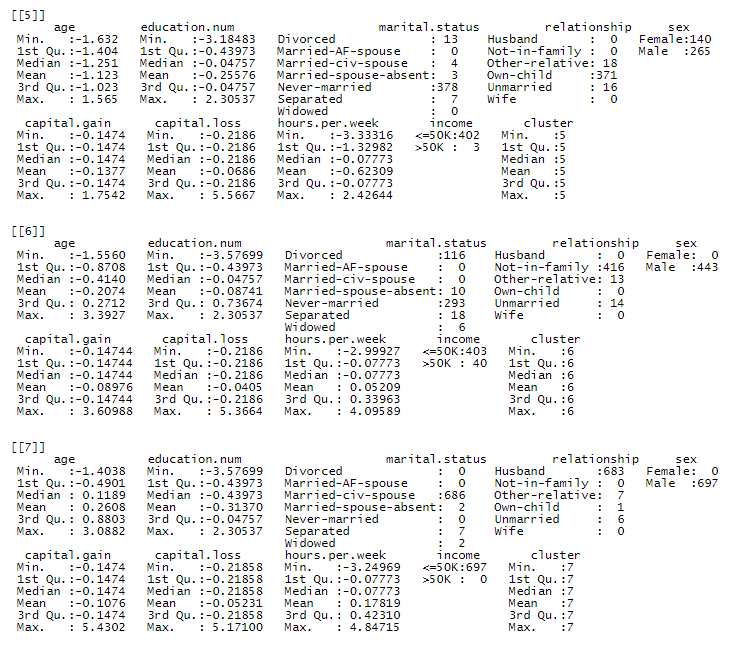
****

****

****

**Results of K-Medoid Clustering method**

****

****