|  |
| --- |
| If real wages aren't rising, how is household income going up? |
| DS6372 Project 2 Logistic Regression Adult Income  April 10, 2022 |
| |  | | --- | | Eric Laigaie, Rayon Morris & Douglas Yip | |

**1. Introduction**

This project will focus on Logistic Regression where we will be analyzing the response (pay\_reponse) of an individual making either greater or less than $50,000.

The following report will contain a detailed analysis and conclusions on the following;

* Initial (Exploratory Data Analysis) EDA
* Building a Logistic Regression Model to predict the binary pay\_response
* Comparing and compiling different regression models, where at least one contains complex variables and at least one that is non-parametric.
* Conclusion and determination of our best model that can predict a binary outcome of in an individual make greater or less than 50,000.

***2. Data description***

For this project, we downloaded two data sets one for training and one for test. We will explore mainly in train and transform test based on the learnings from our training test EDA.

The training data set contains 32,561 individuals data with 16 different attributes (Table 2.1) that our analysis will pertain to predicting whether an individual will make more than >$50,000.. Further changes of the Data Set will be addressed in our Exploratory Data Analysis. Below is a summary of the original file.

***Table 2.1. R output of the car data set that contains the 16 different variables.***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable Name | Data Type | Levels | Min | Max | Mean | Median |
| Id | Int | 32,561 | 1 | 32,561 | 16,281 | 16,281 |
| Age | Int | 73 | 17 | 90 | 38.58 | 37 |
| Workclass | Factor | 9 | -- | -- | -- | -- |
| fnlwgt | Int | 21648 |  |  |  |  |
| Education | Factor | 16 | -- | -- | -- | -- |
| Education\_num | Factor | 16 | -- | -- | -- | -- |
| Martial\_status | Factor | 7 | -- | -- | -- | -- |
| Occupation | Factor | 15 | -- | -- | -- | -- |
| Relationship | Factor | 6 | -- | -- | -- | -- |
| Race | Factor | 5 | -- | -- | -- | -- |
| Sex | Factor | 72 | -- | -- | -- | -- |
| Capital\_gain | Int | 119 | 0 | 99999 | 1082 | 0 |
| Capital\_loss | Int | 92 | 0 | 3770 | 87.9 | 0 |
| Hours\_per\_week | Int | 94 | 1 | 99 | 40.39 | 40 |
| Native\_country | Factor | 42 | -- | -- | -- | -- |
| Pay\_Response | Factor | 2 | -- | -- | -- | -- |

Source: Appendix 2.1 and 2.2**3. Exploratory Data Analysis**

**Missing Values**

Identifying missing values was necessary to obtain accurate summary statistics. We first evaluated whether there were any NA variables or blanks within both data sets. Based on our results, there were no NA. However, based on the summary out put, we noticed that workingclass, occupation and native country contained “?”. We ran a count to see how many rows contain “?” for each column and identify <2000 rows that contain at least one “?” (Appendix 3.1). Given that we have a train dataset of 32,561 and that there was no logical method to imput the data, in our analysis, *we removed all rows from both train and test data to complete this study.*

**Unbalanced datasets**

As a result of this study being a logistical regression, we checked the response value to see if we have a balanced dataset for both training/test data sets. Based on our count (Appendix 3.3), we identified approximately 25% of the results showing a pay\_response greater than $50,000. This would suggest that we are dealing with an unbalanced data set. As such, *we will identify the optimal cut off to maximize accuracy in our logistic regression model and prediction*.

***Continuous variables collinearity check***

The correlation grid (Appendix 3.4) for the continuous variables provides no evidence that any of the variables are correlated. *No action was taken to the continuous variables as a result of the correlation grid*.

***Effects of continuous variables on pay response***

The cluster heat map (Appendix 3.5) to response was evaluated and we see an effect of capital gain and capital loss to the response variables. All other variables were difficult to determine to see if there was any separation of the response. *No action was taken to the continuous variable as a result of the heat map.*

**Education categorical variable**

A histogram of pay response was performed on the education variable where we identified 16 levels of education in the data (Appendix 3.6). *We grouped education based on response and reduce the levels from 16 to 7 levels* based on the following groupings;

* Preschool
* Grade School (grade 1-12 of original data)
* HS Grads (HS Grads + Some college of original data)
* Assocs (Assoc-voc + Assoc-acdm of original data)
* Bachelors
* Masters
* Docs/Profs (Prof-school + Doctorate of original data)

A histogram of pay response was redone (Appendix 3.7) for education and the results were much clearer with the reduction.

**Workclass categorical variable**

A histogram of pay response was performed on the workclass variable where we identified 7 levels in the data (Appendix 3.8). The pay response for the government classified workers were similar in both data sets and *we grouped government (Local, State and Federal) workclass and reduce the levels from 7 to 5 levels*. A histogram of pay response was redone (Appendix 3.10) for workclass and the results were much clearer with the reduction.

**Occupation categorical variable**

A histogram of pay response was performed on the occupation variable as we identified 15 levels in the data. (Appendix 3.10). *We didn’t see any similarities of responses and left the levels as is*.

**Marital Status categorical variable**

A histogram of pay response was performed on the marital status variable where we identified 7 levels in the data (Appendix 3.11). The pay response for the married and formerly married responses were similar in both data sets and *we grouped and created “married” and “single was married” marital status and reduce the levels from 7 to 4 levels*. A histogram of pay response was redone (Appendix 3.12) for marital status and the results were much clearer with the reduction.

**Native Country categorical variable**

A histogram of pay response was performed on the marital status variable where we identified 7 levels in the data (Appendix 3.13). W*e grouped countries into the specific continental regions and reduced the levels from 42 to 7 levels*. A histogram of pay response was redone (Appendix 3.14) for native continent and the results were much clearer with the reduction.

**Redundancy in education and education \_num**

A box plot graph was made between the variable education and education\_num and we observe that this metrics are redundant and exhibit a positive correlation. (Appendix 3.15) *Education is best viewed through factors (there isn't a numerical relationship in education levels), we'll keep education in our final database.*

**Capital Gain and Loss continuous variables**

Capital Gain and Loss are inputs to the Net capital gains formula. *As such, a new variable called Capital Net was created. The formula to his new variable is Captial Gain – Captial Loss.* This in the reduction of a column from the data set.

**Mosaic plots to check multicollinearity for categorical variable**

Mosaic plots were made to check for multicollinearity for categorical variables. Four of the graphs (appendix 3.16) workclass & education; marital status & education; race & education; and race & marital status exhibit little to no evidence of correlation. However, relationship & marital status (appendix 3.17) should strong evidence of correlation. As a result, *the relationship column was removed in favor for marital status*.

**Race Categorial Variable**

We observed that the data set is high skewed with the white race level in the dataset. No additional changes were to the database (Appendix 3.18)

**Removal of categorical ID and flgwt variables**

Two variables were removed from dataset prior to any modeling as the variables provided do not support the prediction of income based on our research and knowledge. ID is an identifier and US census treats flgwt as a portion calculation of the population that is not understood. *We removed the two variables from the data*.

**4. Objective 1**

Build a logistic regression model to complete the analysis; 1) hypothesis test whether to see if we have any significant variables that could predict an individual either have income greater or less than $50,000 2) Determine the model and variables used for the analysis, which include the interpretation of each variable and confidence intervals for each parameter of the model.

**Model Selection Methodology**

Logistical Regression was selected for our model given that the response is a binary variable of <$50K and =>$50K. After our EDA, a total of 9 Variables were used in step.wise logistical regression model. The following explanatory variables were inputted into the model; age, workclass, education, marital status, occupation, race, sex hours per week, native country and capital net.

We ran the step wise logistical regression and result from the model yielded a high number of significant predictors (P >0.01). There were 20 predictors that would make interpretability extremely difficult. Majority of the predictors came from occupation, and we reran the model without occupation to see if can reduce the number of levels without accuracy will suffering. After running the model without occupation we observed not only the number of predictors drop that accuracy did not suffer much. (Table 4.1). Practically we felt that without occupation the our model is sufficient and more interpretable.

**Table 4.1 Interpretation of Parameters for Final Simple Model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **# of betas at P=<0.01** | **Training Accuracy** | **Training**  **Specificity** | **Training**  **Sensitivity** |
| ***All predictor*** | ***20*** | ***83.9%*** | ***92.5%*** | ***57.6%*** |
| ***Occupation Removed*** | ***9*** | ***83.1%*** | ***92.9%*** | ***53.2%*** |

*Source: Appendix (4.1 and 4.2)*

**Test for Fit**

Given the training sample for our train is greater than +30,000 observation, we did not perform the Hosmer-Lemeshow tests to determine goodness of fit as it is not robust for large datasets. The combination of the training accuracy and the cross validation with the test set and the plot for the ROC curve (Appendix 4.3), we determine that our model is a good fit. In addition, a sensitivity of the cutoff was reviewed (Appendix 4.4) and we determine that the default 0.5 was a suitable cutoff to use to maximize our model’s accuracy.

**Final Model**

|  |
| --- |
| **Final Simple Model** |
|  |

Below is the interpretation of the parameters from our final model that determined the pay\_response for >50k.

**Table 4.1 Interpretation of Parameters for Final Simple Model**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Odds Ratio** | | |  |
| **Variable (P<0.01)** | **Parameter** | **Value** | **Odds** | **CI 2.5%** | **CI**  **97.5%** | **Interpretation with CI** |
|  |  | -24.43 | 0.00 | 0.00 | 0.00 | Odds of a person pay >50k are even when all is zero and after accounting for all categorical variables. |
| Age |  | 0.0268 | 1.027 | 1.024 | 1.03 | The odds of a person having >50k is 1.027 times higher than a person 1 year younger holding all other variables fixed. 95% CI (1.024,1.031). |
| WorkClass Self-emp-inc |  | 0.3395 | 1.404 | 1.175 | 1.679 | The odds of Self-employed with income workers having >50k is 1.404 times more than government worker holding all other variables fixed. 95% CI (1.17,1.68) |
| WorkClass Self-emp-not-inc |  | -0.621 | 0.537 | 0.467 | 0.618 | The odds of Self-employed with no income workers having >50k is 1.89 (1/0.5237) times less than a government worker holding all other variables fixed. 95% CI (1.61,2.14) |
| Marital Status Married |  | 2.136 | 8.461 | 7.533 | 9.505 | The odds of Married individual having >50k is 8.46 times more than a previously married individual holding all other variables fixed. 95% CI (7.53,9.505) |
| Marital Status  Never Married |  | -0.531 | 0.587 | 0.506 | 0.682 | The odds of Never married individual having >50k is 1.7 times less than a previously married individual holding all other variables fixed. 95% CI (1.97,1.47) |
| Race White |  | 0.0642 | 1.900 | 1.240 | 2.912 | The odds of White individual having >50k is 2.0 times more than Indian American individual holding all other variables fixed. 95% CI (1.24,2.912) |
| Sex Male |  | 0.0135 | 1.144 | 1.039 | 1.261 | The odds of males having >50k is 1.1 times more than females holding all other variables fixed. 95% CI (1.039,1.261) |
| Hours per week |  | 0.03 | 1.030 | 1.027 | 1.033 | The odds of a person having >50k is 1.030 times higher than a person working an additional hour holding all other variables fixed. (95% CI (1.027,1.033) |
| Capital Net |  | 0.0003 | 1.000 | 1.000 | 1.000 | The odds of a person having >50k is even to a person making $1 younger holding all other variables fixed. (95% CI (1.000,1.0000)) |

**Conclusion**

As expected based on the EDA, the counts where we had more >50k responses for each variable reflects the odds to be higher for levels that had few >50k individuals identified. We observe that individuals that are married, white, older and make income have great odds to secure income greater than $50,000.

**5. Objective 2**

***Problem Statement***

In objective 1, our goal was to create a simple, highly interpretable model. For this objective, we will be using that model as a baseline and use various methods in an attempt to improve model performance. First, we well keep the variables produced in the stepwise model but add complexity through interaction terms. Then, we will use LDA and QDA to …. Lastly, these models will be contrasted with the stepwise model to compare performance and determine the most optimal solution.

**Complex Model**

Our method for creating this model was to run through many prospective models with varying terms and viewing how these interactions turned out in the model summary. If we found certain pairings of variables were often insignificant, we moved away from them and explored new interactions. To better understand any roadblocks for this model, we used the cooksd() function to provide us with any outlying and highly influential points. Similar to objective 1, we did not find any points returned here.

After reaching a final model, we investigated every cutoff value between 10 and 90 percent with an interval of .1%. Our code found that a classification cutoff of 48% provided the maximum accuracy of 82.71%. In Appendices 5.1-5.4, you can find this model’s summary, confusion matrix, accuracy vs. cutoff chart, and roc curve.

**LDA/QDA**

**Random Forest**

***Model Results***

Accuracy and Sensitivity, Specificity

Table 5.1 Summary of statics for the models predicting test and validation data.

|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Models Test Statistics | Accuracy | Sensitivity | Specificity |
| Objective 1 Model | 83.12% | 92.88% | 53.16% |
| Complex Model | 82.71% | 90.87% | 57.65% |
| LDA/QDA |  |  |  |
| Random Forest |  |  |  |

***ROC graph***

***Takeaways from the Models***

***Original Model*** (Intuition/Forward)–

***Model 2*** – XXX

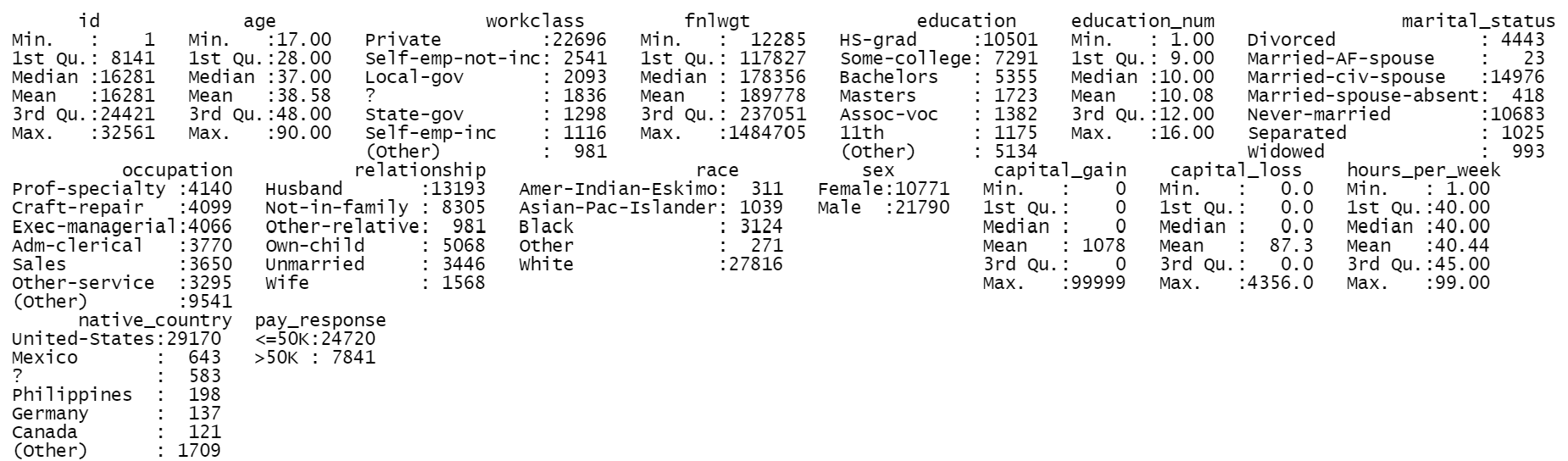
***Model 3*** – XXX

**6. Final Summary**

Through both objectives, we built a menu of models to be evaluated both alone and together. Although we attempted to build high-accuracy models, our results lead us to believe that this dataset leaves a lot to be desired. With only three numeric features, the regression models we built have struggled to achieve a strong accuracy. Future steps to be taken could include exploration into related data that could bolster the robustness of the dataset and exploring more complicated models, such as random forests, k-nearest neighbors, and neural networks.

**7. Appendix**

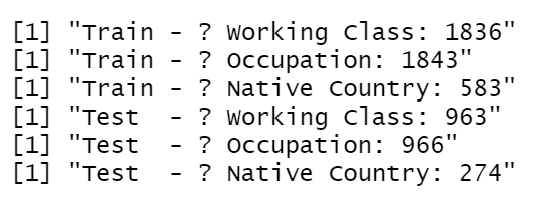
**Appendix 2.1 – Initial Summary Statistics**

****

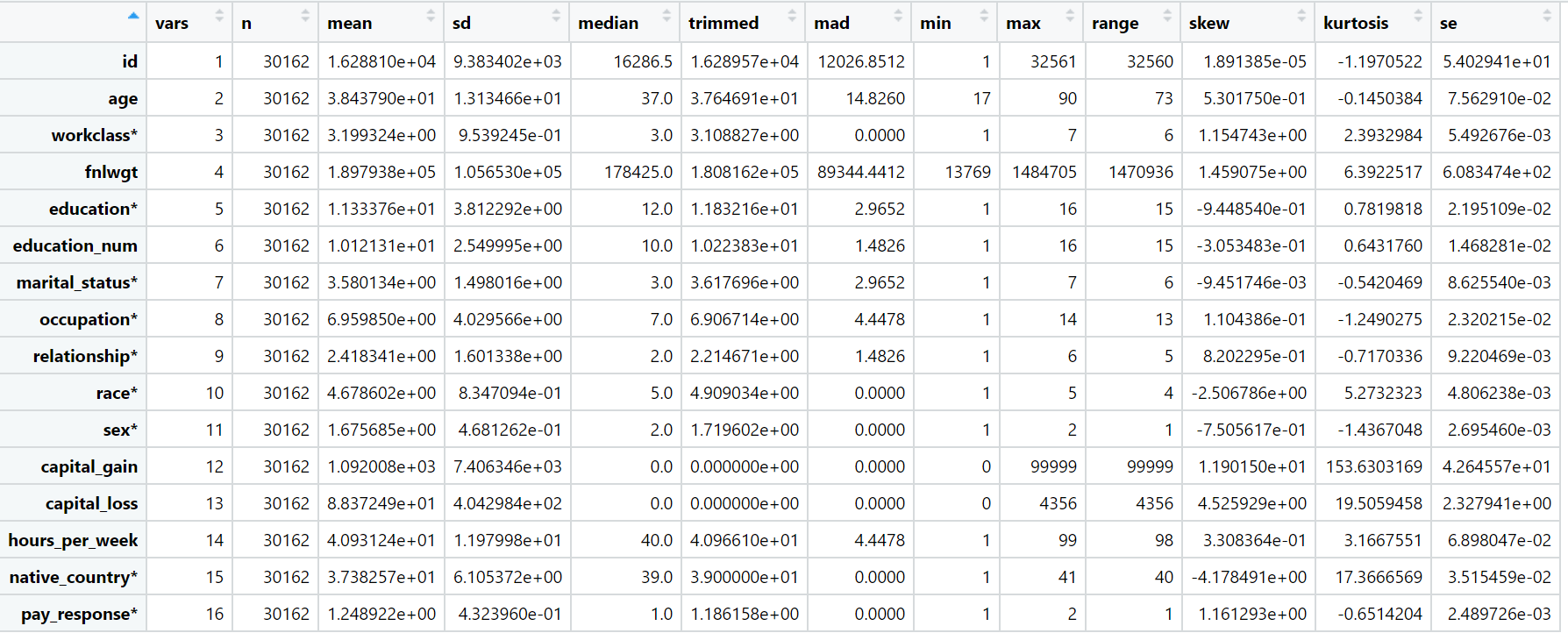
**Appendix 2.2 – Number of levels by Variable**

****

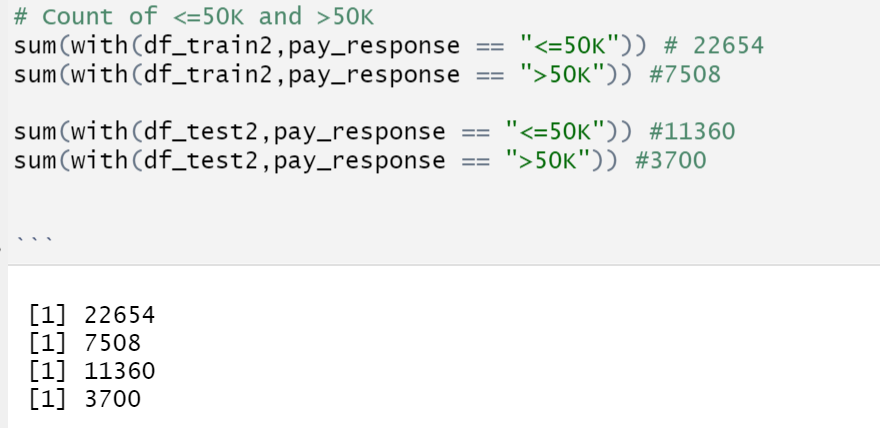
**Appendix 3.1 R output of number of unknown variables coded “?”**

****

**Appendix 3.2 Summary of data set after removal of “?”**

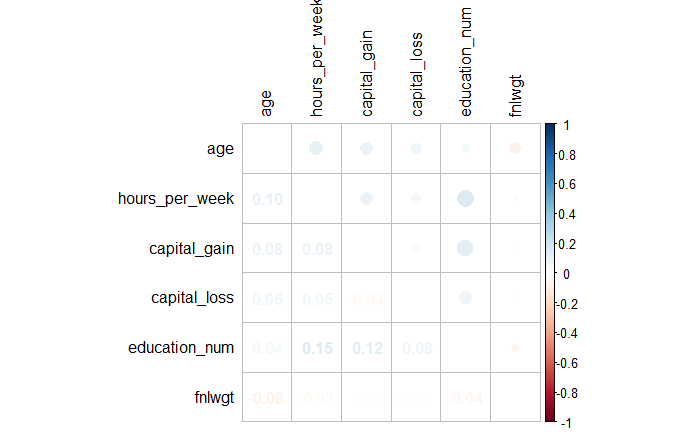
****

**Appendix 3.3 R output of Unbalanced response for both the train and test data sets**

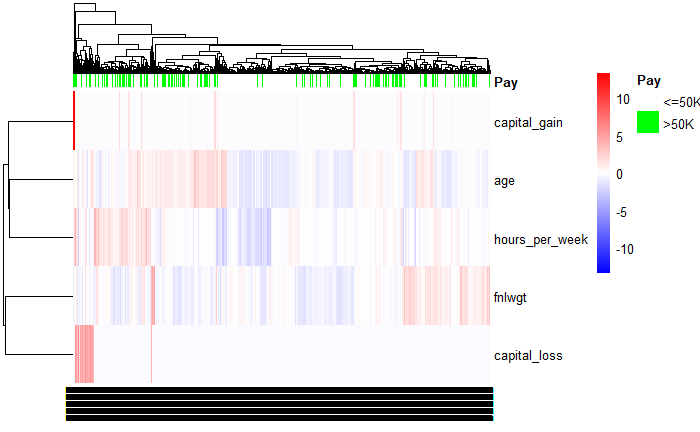
****



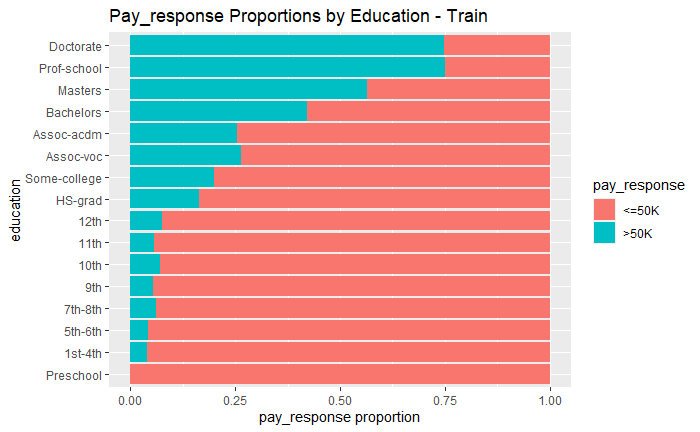
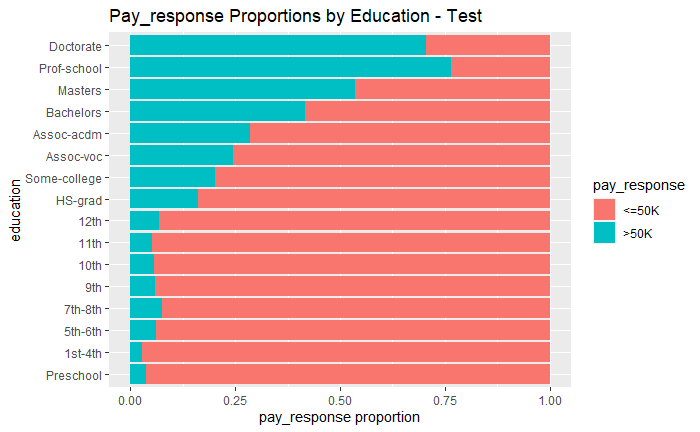
**Appendix 3.4 Correlation plot of continous variables.**



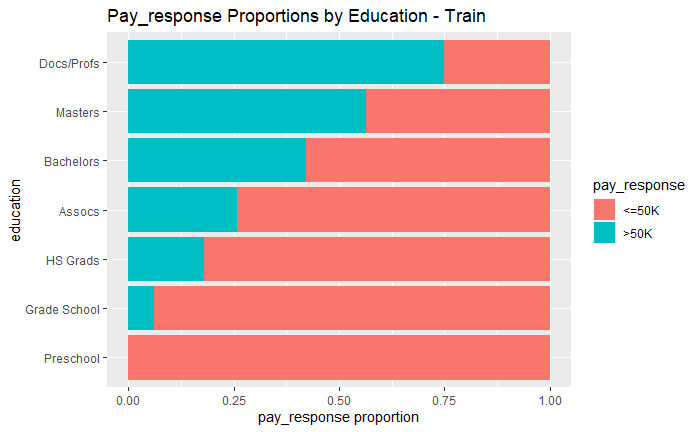
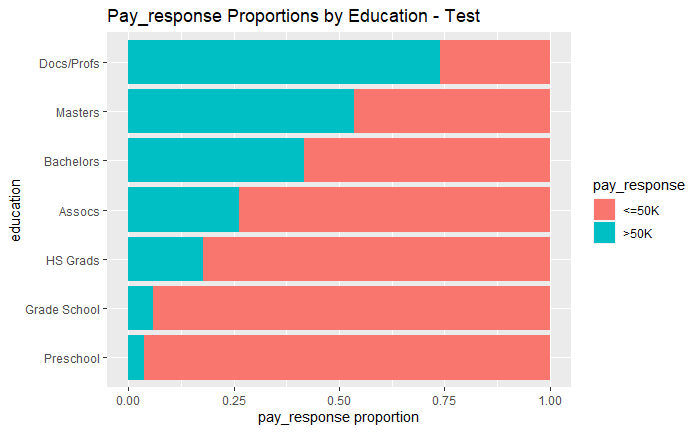
**Appendix 3.5 Cluster Heat map to determine the impact to pay response**



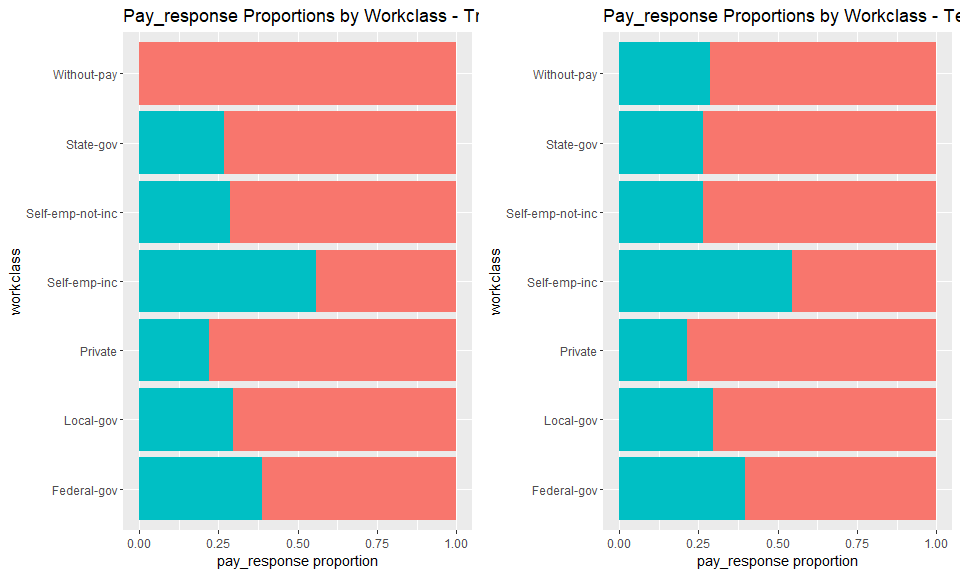
**Appendix 3.6 Pay response for Education pre transformation (Training and Test Sets)**

**Appendix 3.7 Pay response for Education post transformation (Training and Test Sets)**

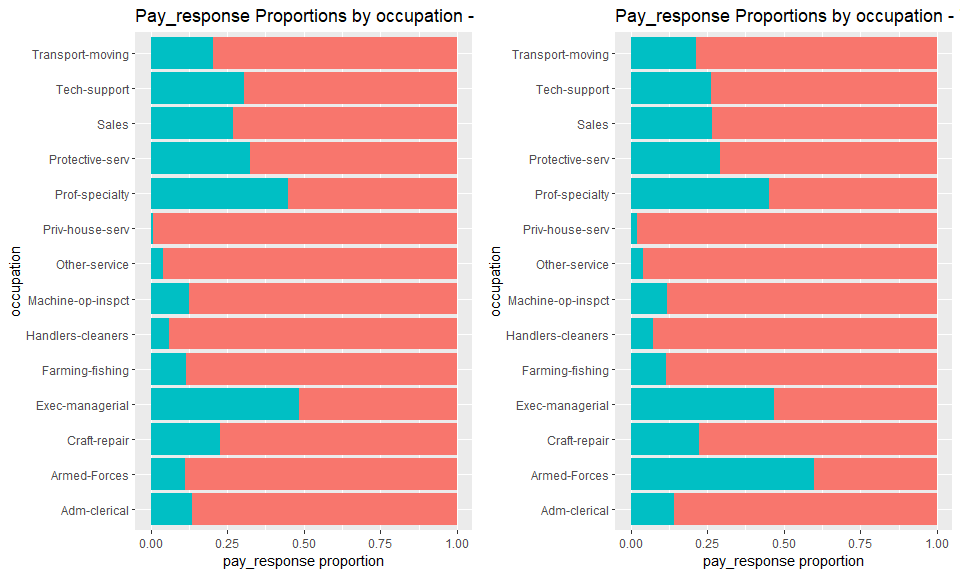
**Appendix 3.8 Pay response for Workclass pre transformation (Training and Test Sets)**



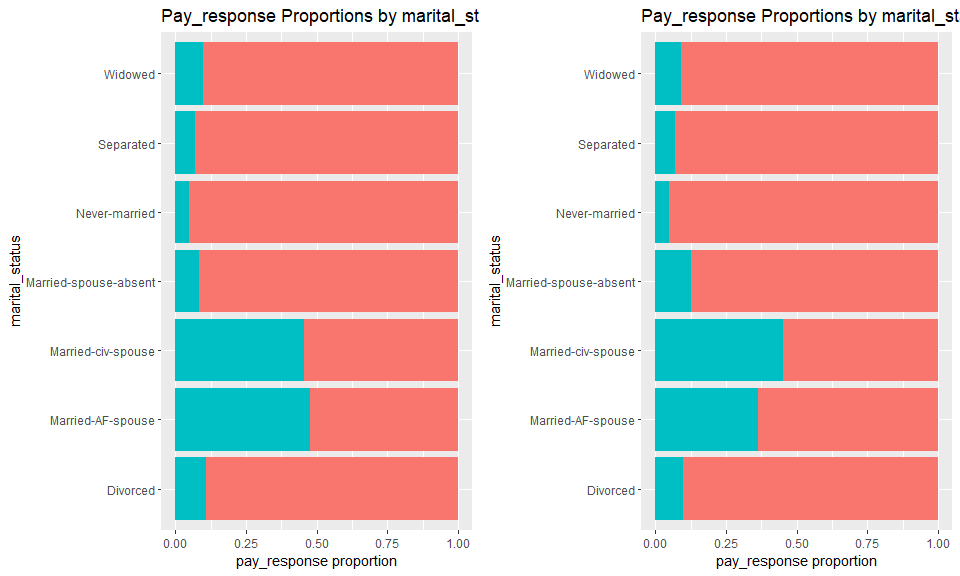
**Appendix 3.9 Pay response for Workclass post transformation (Training and Test Sets)**



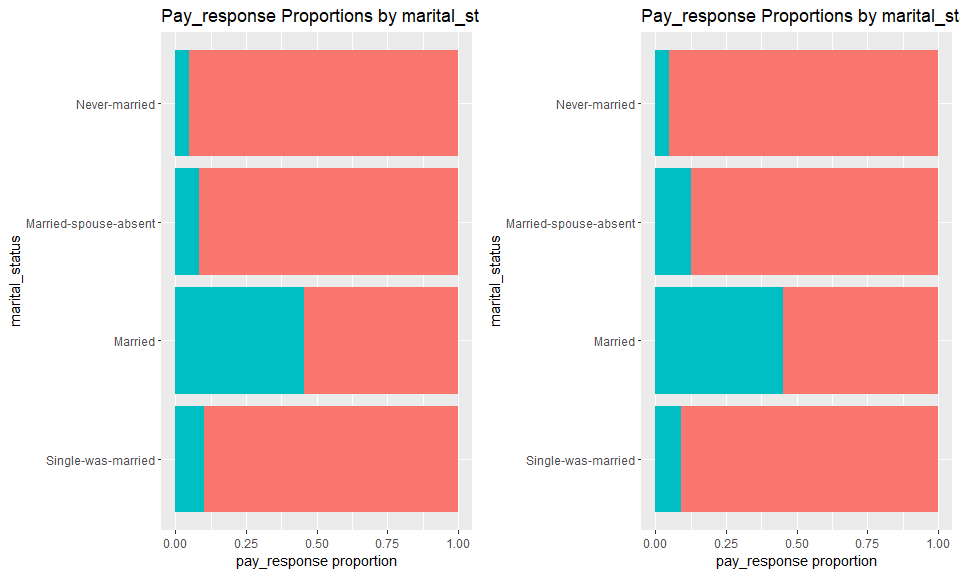
**Appendix 3.10 Pay response for Occupation**



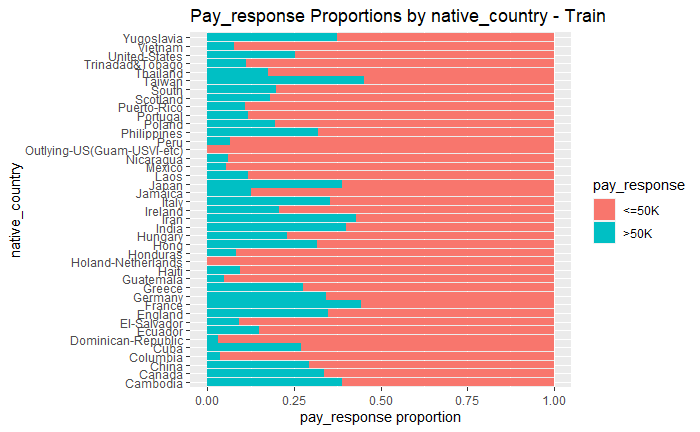
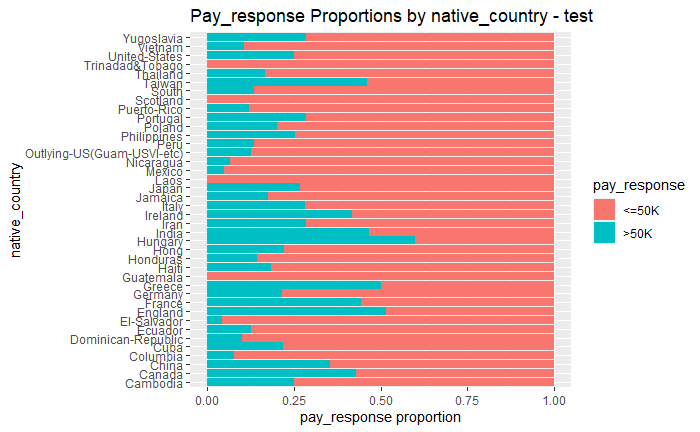
**Appendix 3.11 Pay response for Marital status pre transformation (Training and Test Sets)**



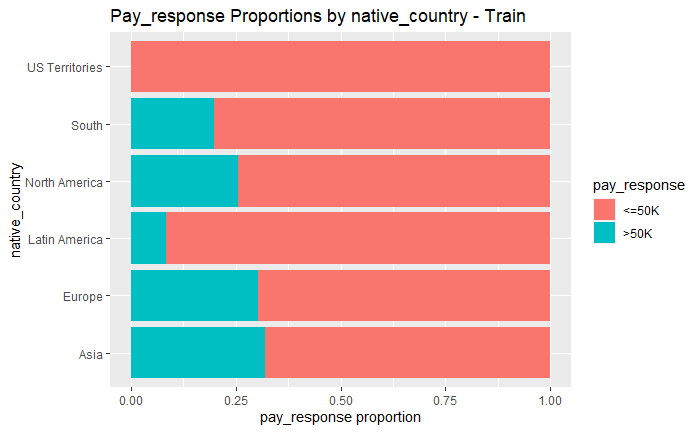
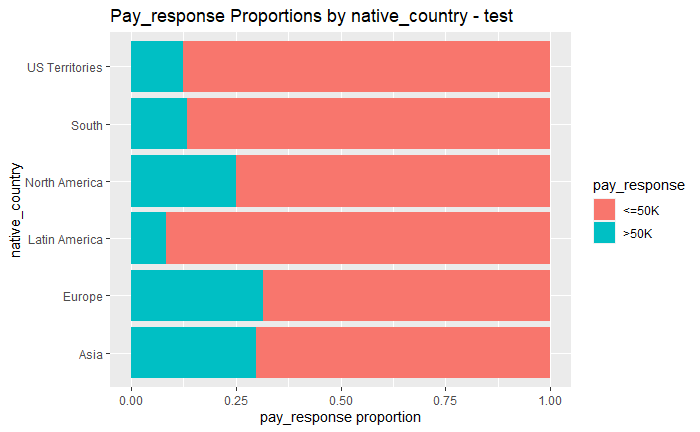
**Appendix 3.12 Pay response for Marital status post transformation (Training and Test Sets)**



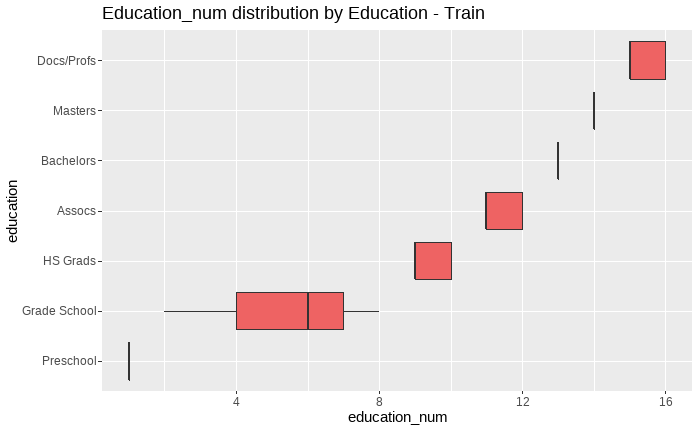
**Appendix 3.13 Pay response for Native Country pre transformation (Training and Test Sets)**

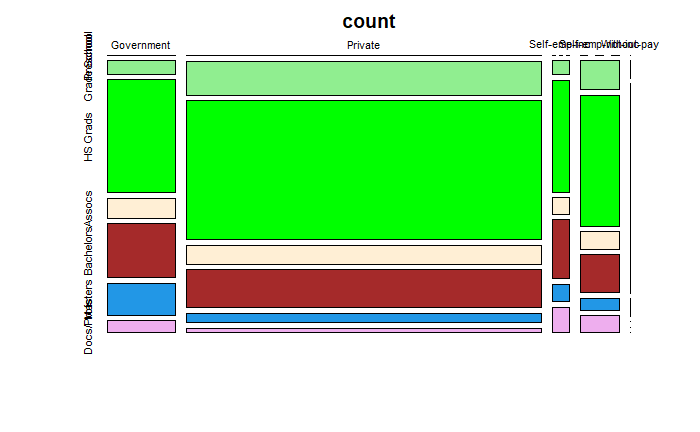
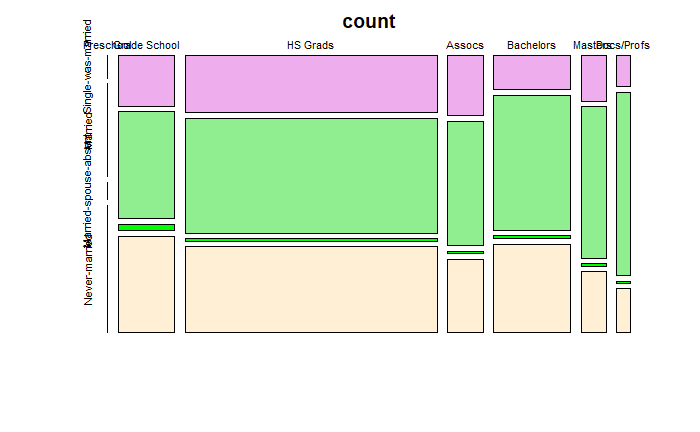
**Appendix 3.14 Pay response for Native Country post transformation (Training and Test Sets)**

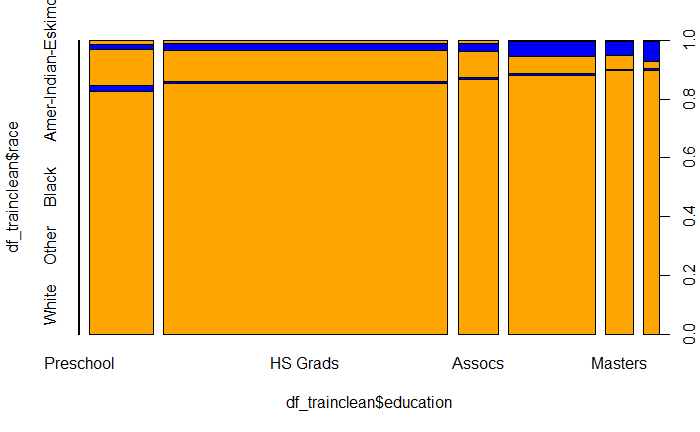
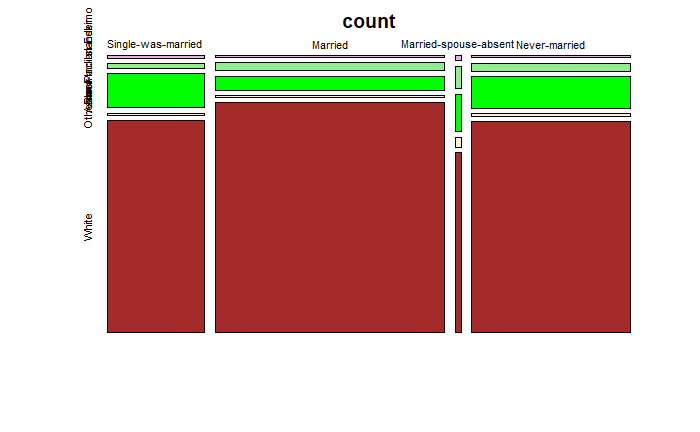
 

**Appendix 3.15 Box plot graphs of the redundancy of education and education\_num variables**

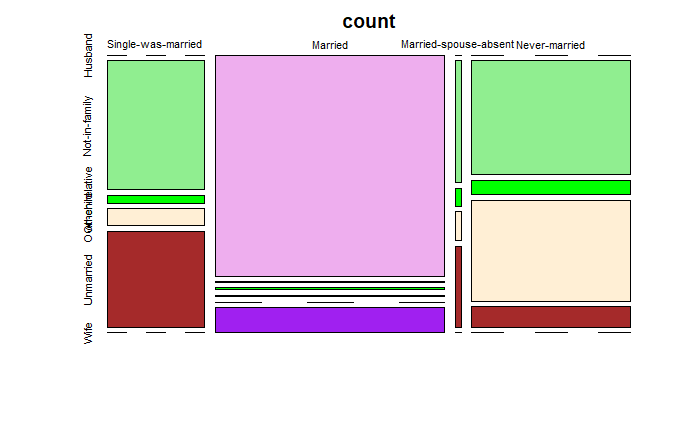


**Appendix 3.16 Mosaic plots to check for collinearity among categorial variables (minimal to no evidence)**

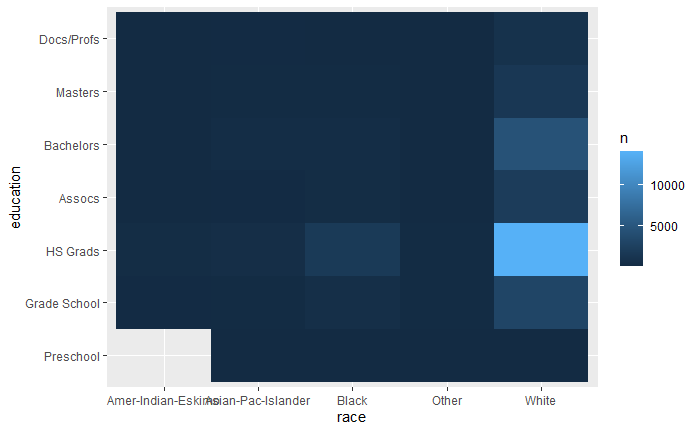
 

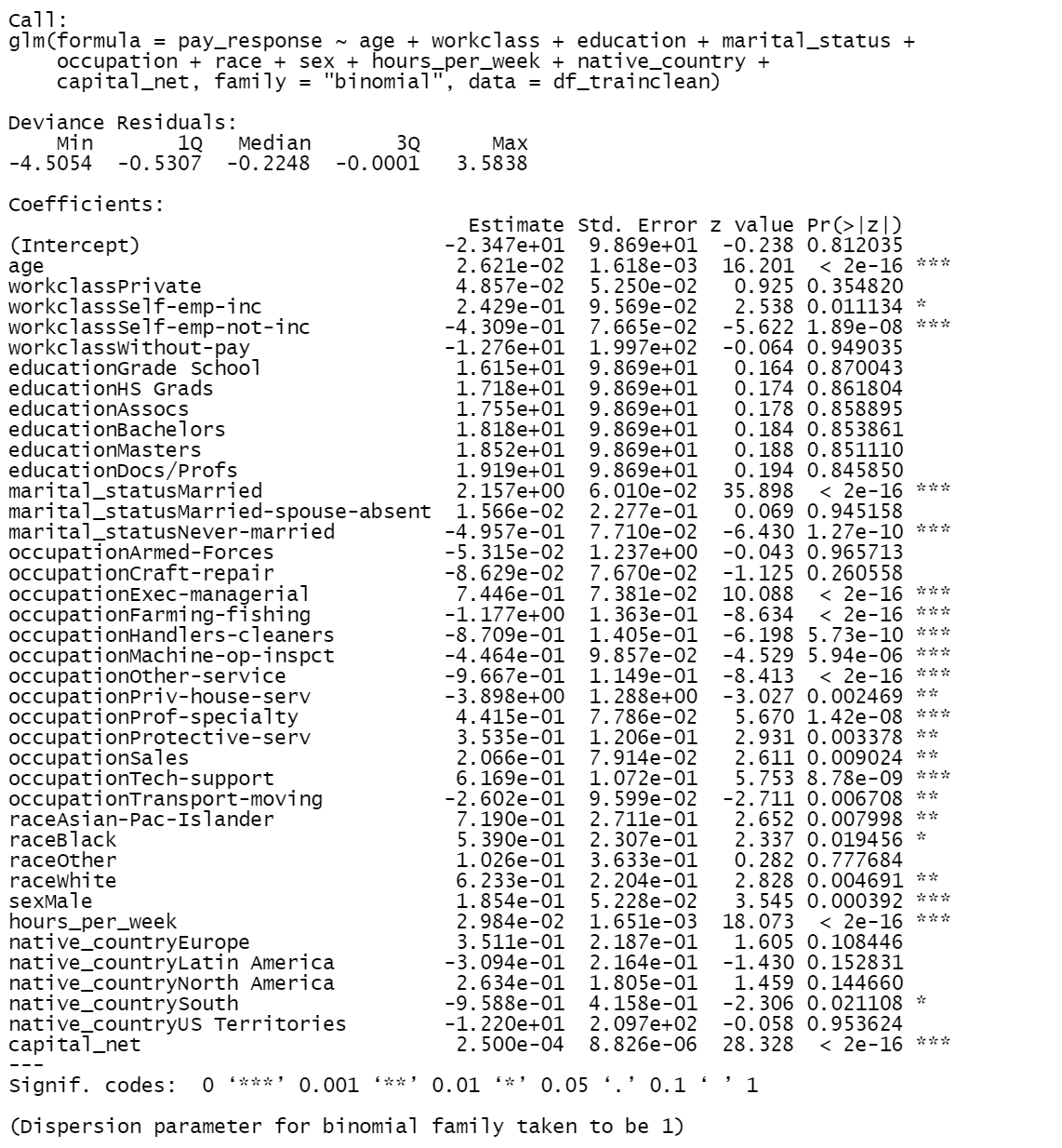
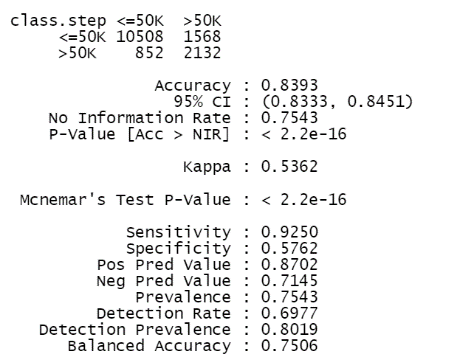
**Appendix 3.17 Mosaic plots to check for collinearity among categorial variables (Strong evidence)**



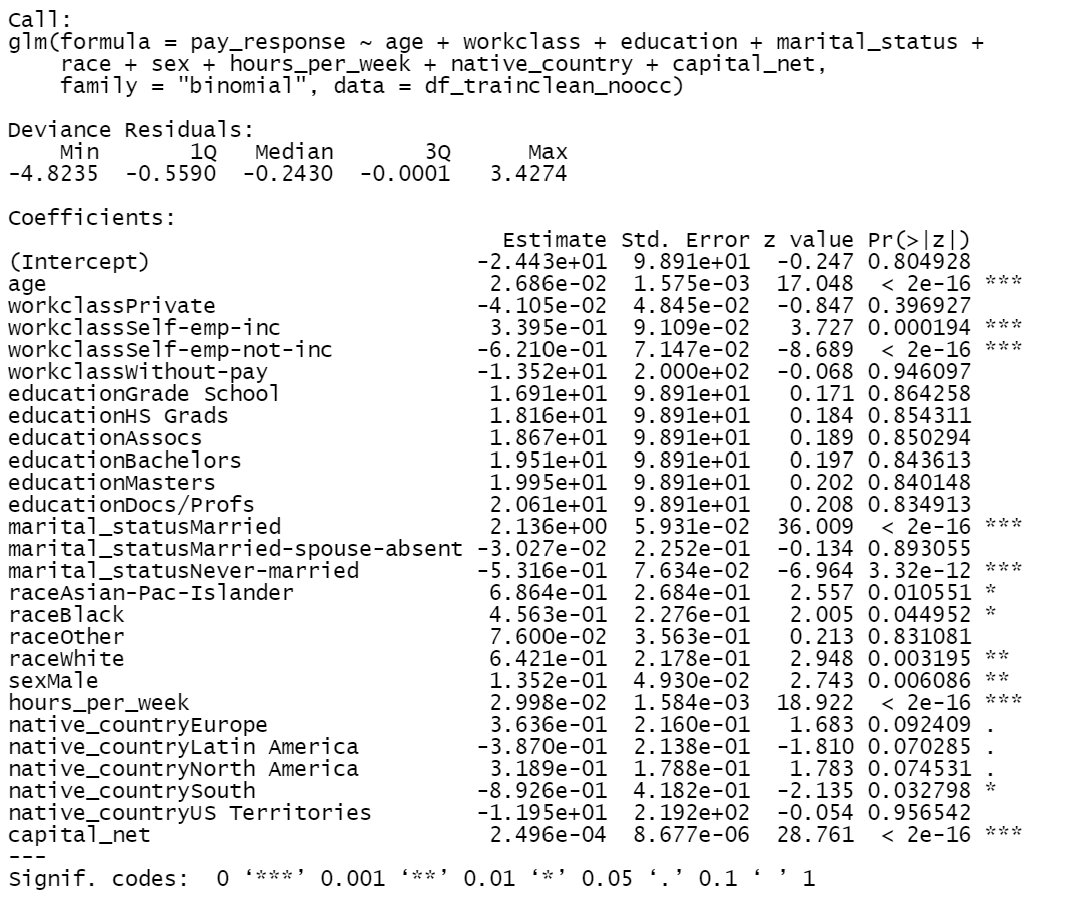
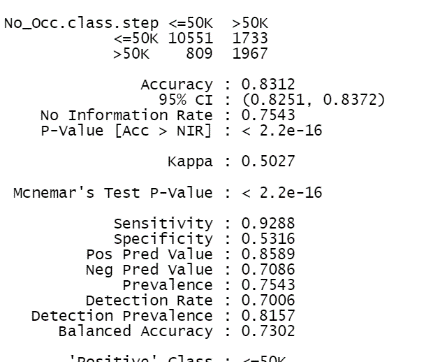
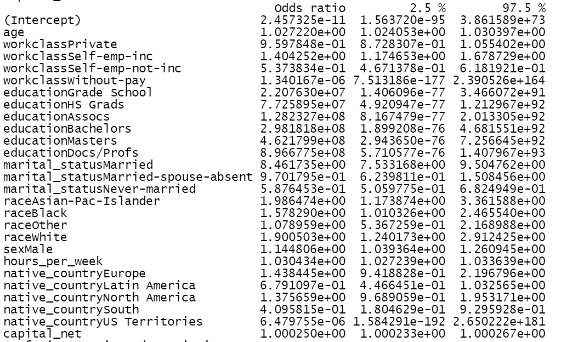
**Appendix 3.18 Race by education heat map count**



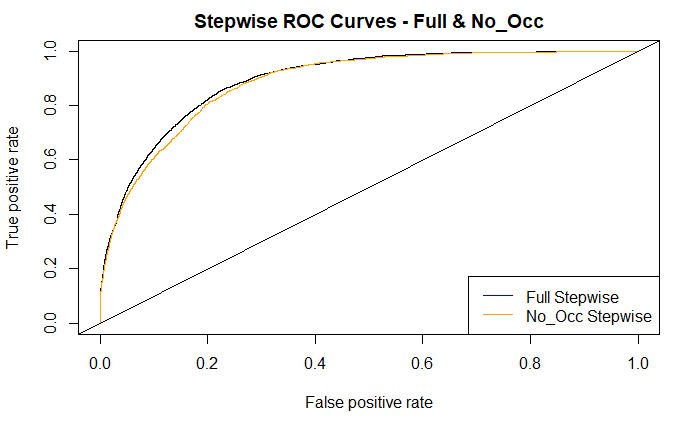
**Appendix 4.1 Step Wise Logistic Regression Output**

****

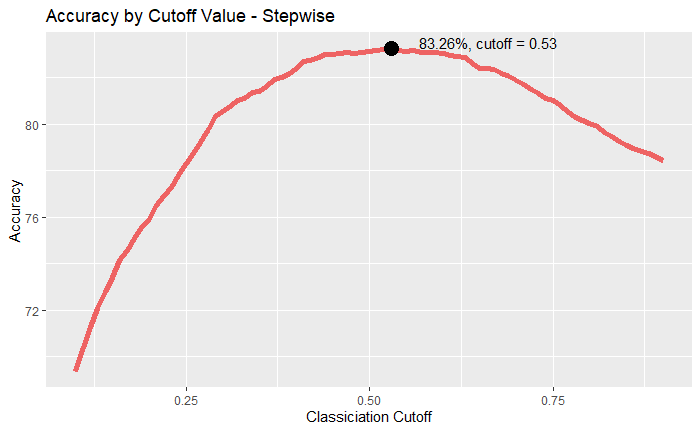
**Appendix 4.2 Step Wise Logistic Regression Output without occupation**

****

**Appendix 4.3 ROC for both stepwise test (one with occupation and one without)**



**Appendix 4.4 Cut off value of stepwise model excluding occupation**



**Appendix 5.1 Interaction Regression Output**

**Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated**

**Appendix 5.2 Interaction Regression Confusion Matrix**

Graphical user interface, text

Description automatically generated

**Appendix 5.3 Interaction Regression Accuracy / Cutoff Chart**

Chart, line chart

Description automatically generated

**Appendix 5.4 Interaction Regression ROC Curve**

Chart, scatter chart

Description automatically generated

**R-Mark Down Code**

**To be added when code is finished**