|  |
| --- |
| 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 3.14) *Since education is best viewed through factors (there isn't a numerical relationship between education levels), we'll keep education*

**Education**

**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**

**Test for Fit**

**Final Model**

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

Below are the interpretation of the parameters from our final model that determined the pay\_response.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Parameter** | **Value** | **Interpretation with CI** |
|  |  |  |  |
|  |  |  |  |

**5. Objective 2**

***Problem Statement***

XXX

***Original Model***

**Model 2 – Complex Model**

**Model 2 – Decision Tree with KNN**

***Model Results***

XXX

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

|  |  |  |
| --- | --- | --- |
| Predictive Models Test Statistics | AIC | BIC |
| Original Model |  |  |
| Model 2 |  |  |
| Model 3 |  |  |

Source: Appendix 5.2 and 5.5

***Takeaways from the Models***

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

***Model 2*** – XXX

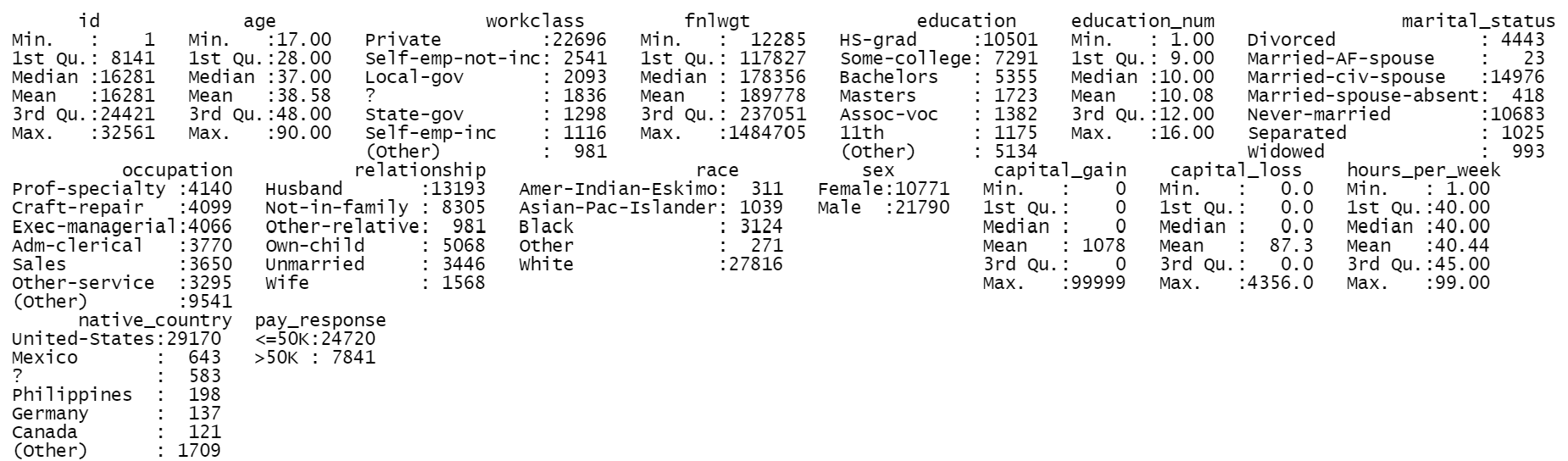
***Model 3*** – XXX

**6. Final Summary**

To be finalized once data is completed

**7. Appendix**

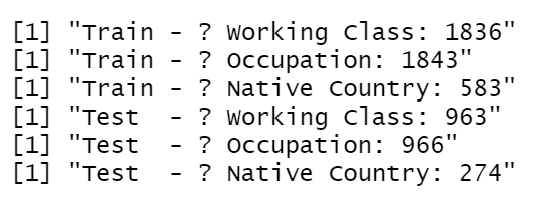
**Appendix 2.1 – Initial Summary Statistics**

****

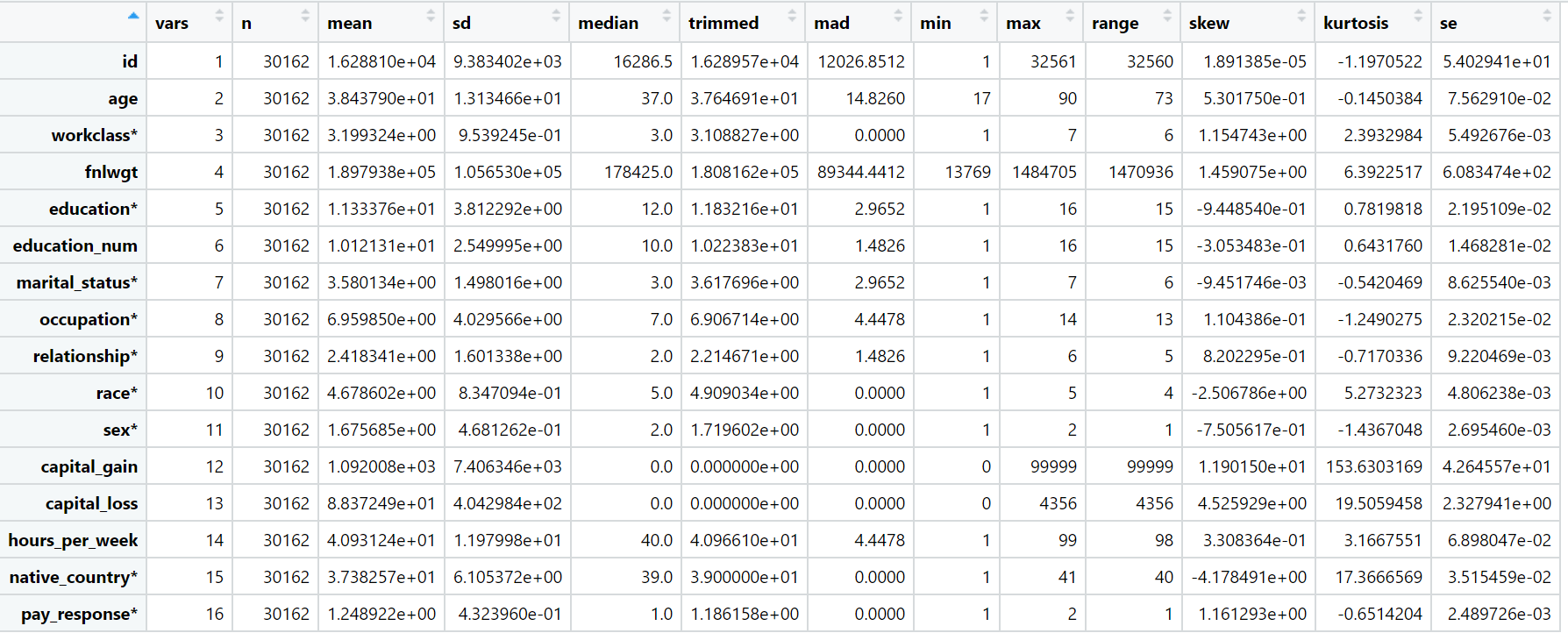
**Appendix 2.2 – Number of levels by Variable**

****

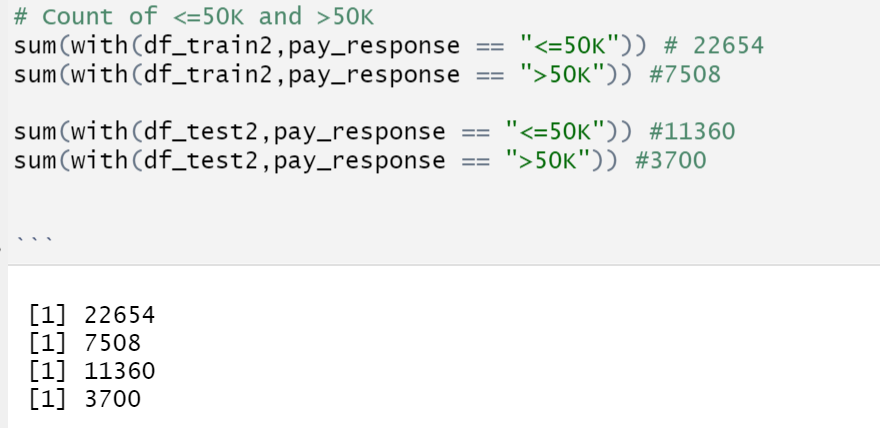
**Appendix 3.1 R output of number of unknown variables coded “?”**

****

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

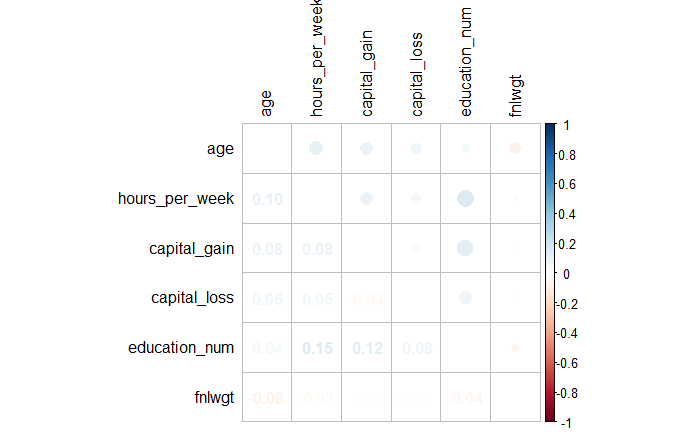
****

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

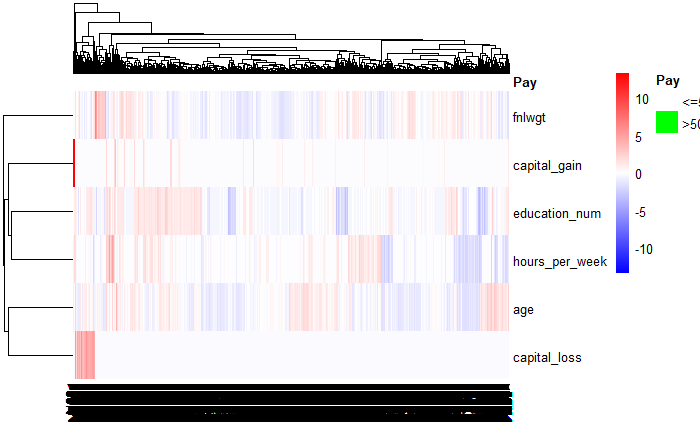
****



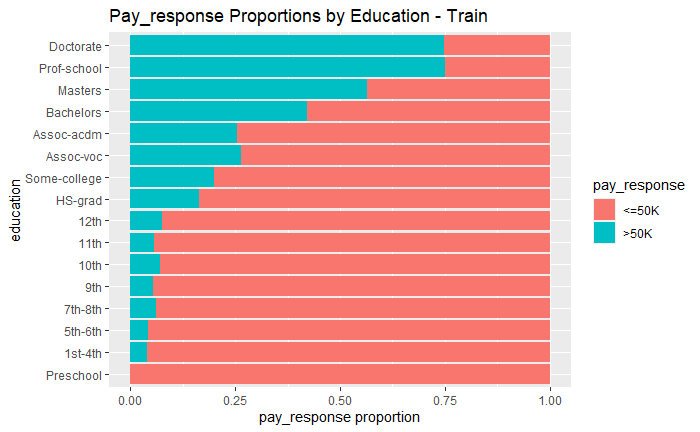
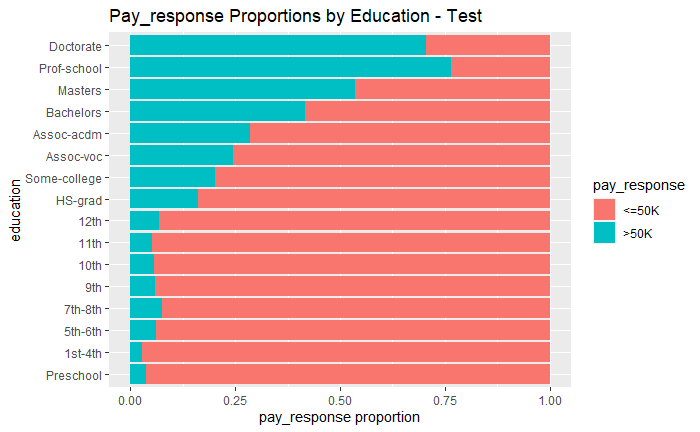
**Appendix 3.4 Correlation plot of continous variables.**



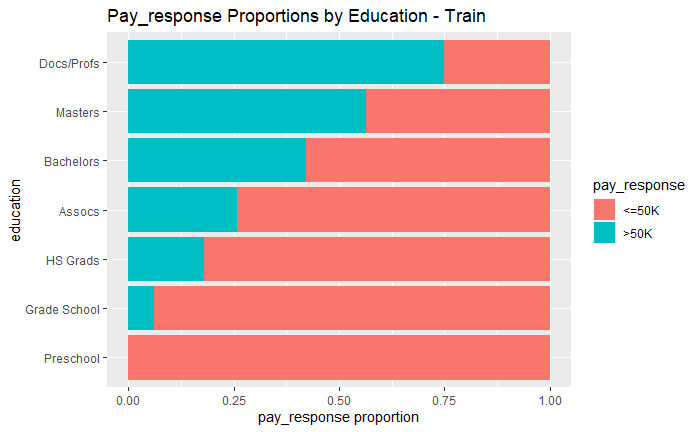
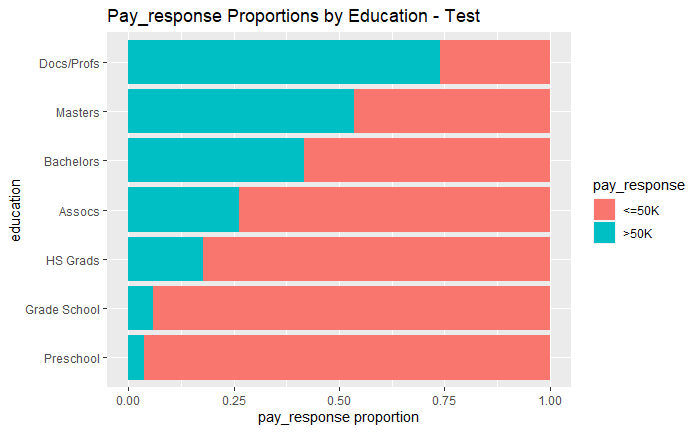
**Appendix 3.5 Cluster Heat map to determine the impact to pay response**



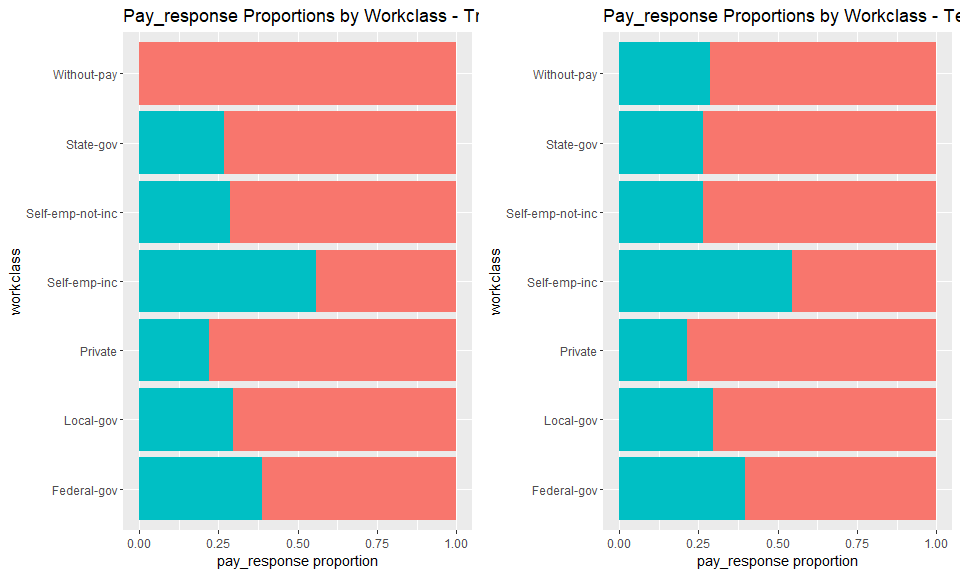
**Appendix 3.6 Pay response for Education pre transformation (Training and Test Sets)**

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

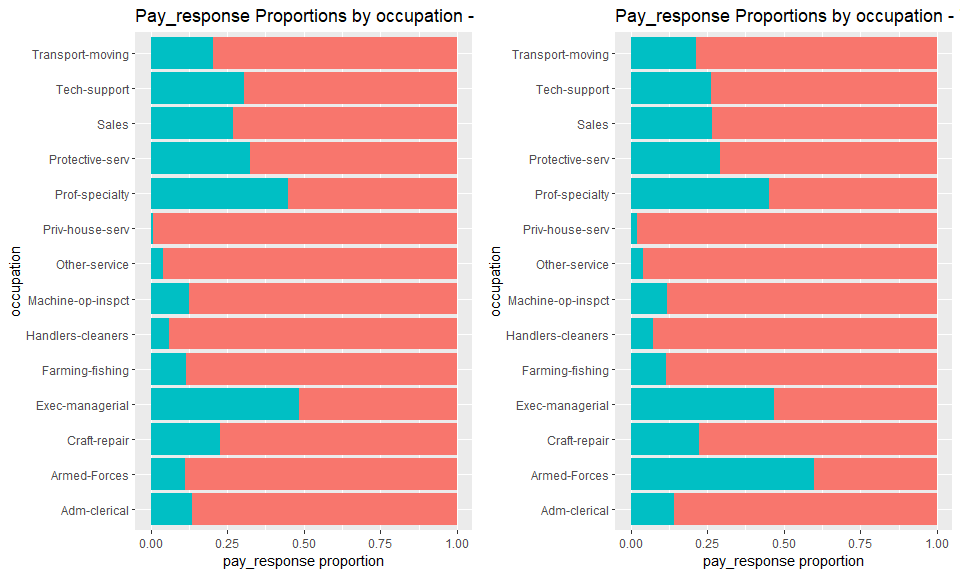
**Appendix 3.8 Pay response for Workclass pre transformation (Training and Test Sets)**



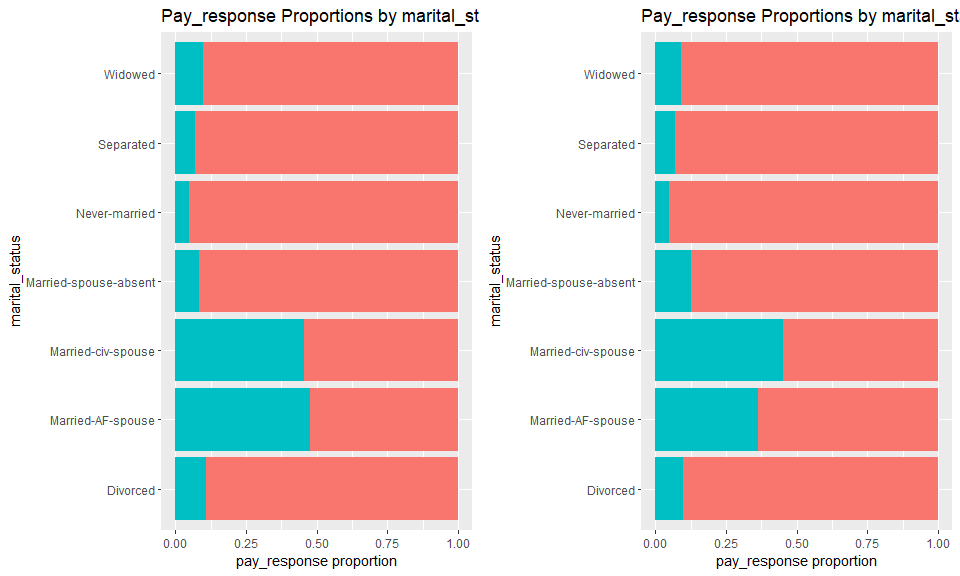
**Appendix 3.9 Pay response for Workclass post transformation (Training and Test Sets)**



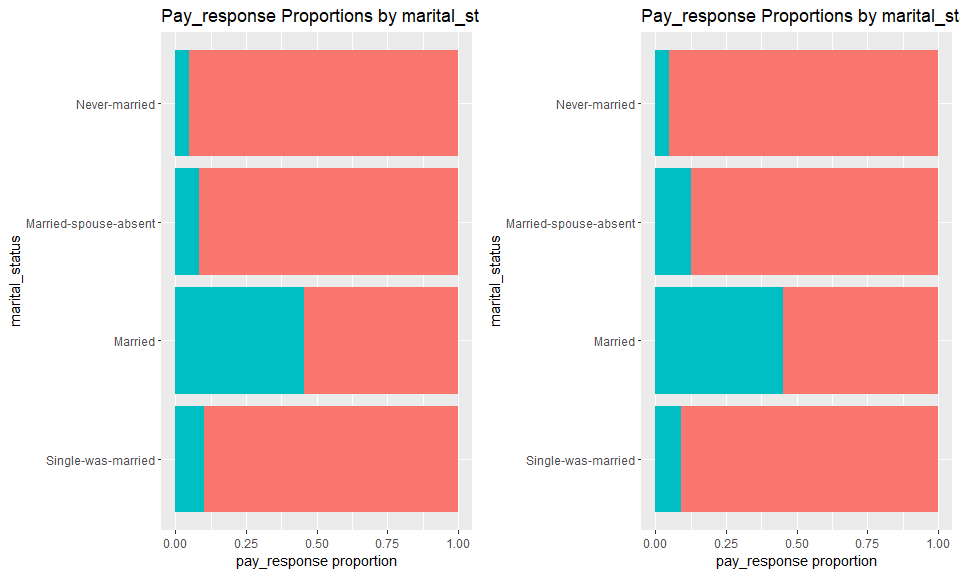
**Appendix 3.10 Pay response for Occupation**



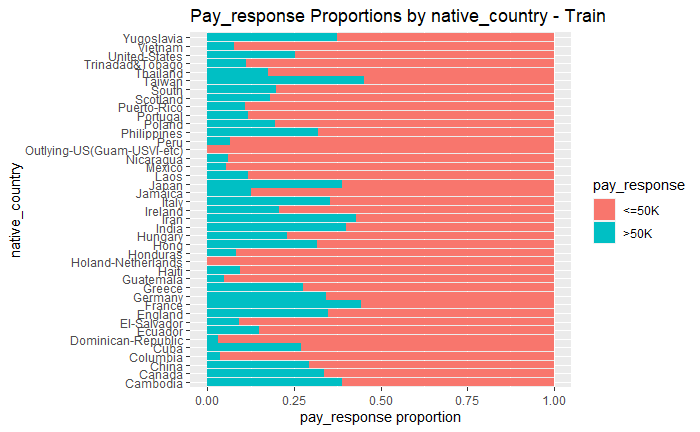
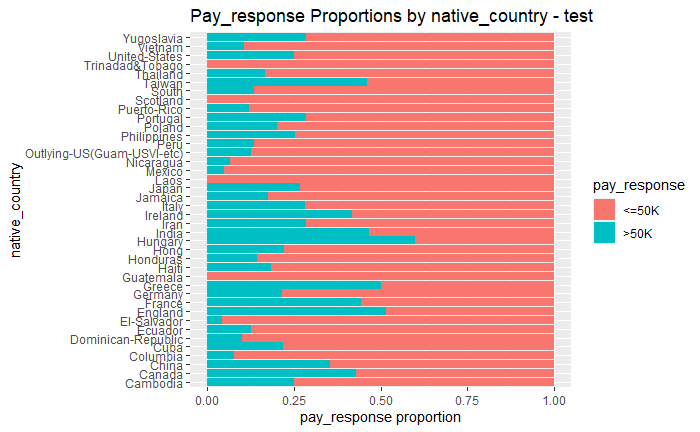
**Appendix 3.11 Pay response for Marital status pre transformation (Training and Test Sets)**



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

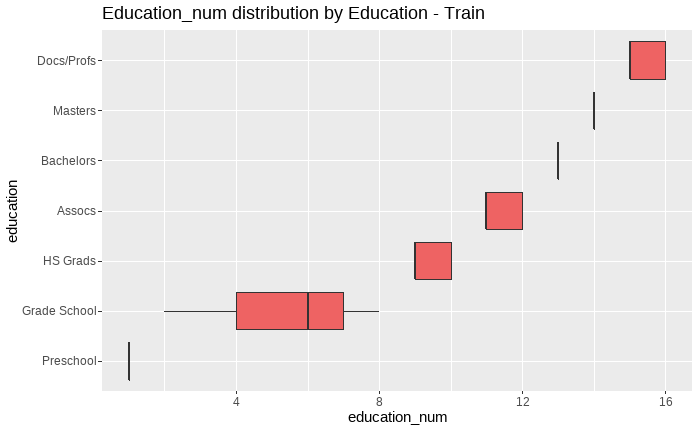


**Appendix 3.13 Pay response for Native Country pre transformation (Training and Test Sets)**

**Appendix 3.12 Pay response for Native Country post transformation (Training and Test Sets)**

**Appendix 3.14 Box plot graphs of the redundancy of education and education\_num variables**



**Appendix 4.x Question 1 Plots**

**Appendix 5.x Question 2 Plots**

**R-Mark Down Code**

**To be added when code is finished**