Extended Essay

Mathematics

Topic:

Statistical Analysis

Research Question

To what extent does Univariant, Bivariant analysis and Chi-Square Test help to identify the primary factors that influence the outcome of a dependant variable (salary) and, how can Logistic Regression and Confusion Matrix assist in creating an accurate model which will be able to identify the discrete binomials of the dependant variables

Word Count: - 3823

Content

1 Introduction

Rationale

As I'm commencing exploring my options towards identifying the most suitable and lucrative career path with regards to growth and good compensation, I found this case study on factors which influence salaries in various world economies quite intriguing. My curiosity pushed me to investigate the question "What are the variables that affect salary"?

For this purpose, my essay follows the research question: "To what extent does Univariant, Bivariant analysis and Chi-Square Test help to identify the primary factors that influence the outcome of a dependant variable (salary) and, how can Logistic Regression and Confusion Matrix assist in creating an accurate model which will be able to identify the discrete binomials of the dependant variables." To analyze this question, I'll use various statistical tools.

This dataset has many variables which gives an insight towards understanding the factors which influence salaries in various parts of the world and what factors have positive influence and relations to eventually determine what the worker's takes home as salary. This data set was obtained from a data repository, which is a storage space for researchers to deposit data sets associated with their research.

Initially the basic concept of Descriptive Analysis will be introduced by defining variables that I will be exploring in the dataset in relation to the dependant variable i.e., salary. The variables will be segregated as quantitative and qualitative data.
Next, I will start with Descriptive Analysis by exploring the Uni-variant analysis of the quantitative data and the Bi-variant analysis of the qualitative data which will be interpreted using various graphs.
Subsequently, we see how the variables affect each other and further analyse then using Hypothesis testing to check if a positive correlation can be found.

To sum up we will be doing a Correlational Analysis for which, I’ll be using the Chi-Square Test to understand whether the variables chosen have an influence towards the person’s salary to be above USD 50,000 or below USD 50,000.

I will be creating models for predicting and understanding the relationships between variables in terms of their influence on the dependent variable. I will create a Logistic Regression Model for the same.

Accuracy of the models created will be tested using concepts like the GLM, ROC, GINI and KS will be explored in its respective areas.

To check the exact accuracy of the model, I will use the Confusion Matrix to extract the three important aspects of Accuracy, Specificity and Sensitivity of the model.

I will conclude the essay by displaying which variables play an important role in determining if a person earns more than USD 50,000.

Aim:

The aim of this essay is in Two fold

- Identifying the primary factors influencing the outcome of the salary earned to be below USD 50,000 or above USD 50,000.
- Creating a model which most accurately will be able to identify the discrete binomials of the dependant variables.

2 Descriptive Analysis

Descriptive analytics answer the question, “What happened?”.

This type of analytics is most commonly used by customers, providing reporting and analysis centred on past events .

Descriptive Analysis would include:

> Summary of the data set which will give you statistics like the mean, median, mode, minimum, maximum various quartiles, etc.

> Missing values, outliers, and scope to customise and alter the data set by adding useful information derived by exploring the data/deleting unwanted or useless data.

> Analysing the structure of the data set for aspects like its dimensions and how many discreate & continuous variables form the data set.

> Exploratory Data Analysis (EDA) - analysing each variable individually and identifying the Dependant variable and conducting bi-variant analysis to see the influence each independent variable has on the dependent variable.

> Multicollinearity (correlation) test to see the degree of impact that all variables have on each other.

- **Structure of the data set**

The plot structure reveals the various variables of the data frame, with 15 columns and 32561 rows. This data set includes the factors – Age, work class, final weight, education, No of years of education, marital status, occupation, relationship, sex, capital gain, capital loss, working hours per week, native country, salary.

- **Exploring the dataset**

Before we start the process of univariant and bi variant analysis, it is important for us to find out the measures of central tendency that includes the Mean, Median, Mode. Measures of central tendency are an important tool for summarizing and describing data, and for making inferences about the underlying population from which the data was drawn , In our case finding the Mode of the dataset is insignificant as it does not affect our analysis in any way. Along with this, the measures of dispersion, this includes the minimum value , maximum value, 1st quartile ,3rd quartile and IQR .We do this for the same reasons because they provide information about the spread of the data, which is essential for understanding the data and making meaningful inferences. These are for the quantitative variables ; however for the qualitative variables it is important to put them in tables in a structured manner for an easier way to comprehend the large data into small sections

- **Problem Statement:**

Delta Square is a leading consultant firm with multiple locations throughout the United States. They have been careful in gathering data. They intend to use all this data to automate a critical component of Operations. Salary is a crucial component that influences their clients' purchasing power. Delta Square's Director of Operations aims to eliminate this procedure and use analytics to estimate salary.

The table below shows the summarised data of the quantitative values presented in the dataset.

3 Exploratory Analysis

Quantitative Data

Factors	MIN	MEAN	1st QUARTILE	MEDIAN	3rd QUARTILES	MAX
AGE	17	38.58	28	37	48	90
FINAL WEIGHT	12285	189778	117827	178356	237051	1484705
EDUCATION NO.OF YEARS	1	10.08	9	10	12	16
CAPITAL GAIN	0	1078	0	0	0	99999
CAPITAL LOSS	0	87.3	0	0	0	4356
WORKING HOURS PER WEEK	1	40.44	40	40	45	99

- Univariant Analysis

Univariate analysis is a statistical technique that is used to analyse and describe the distribution of a single variable. The goal of univariate analysis is to summarize the main features of the distribution of a single variable, such as its central tendency, dispersion, and shape ; this is commonly used for the analysis of quantitative data.

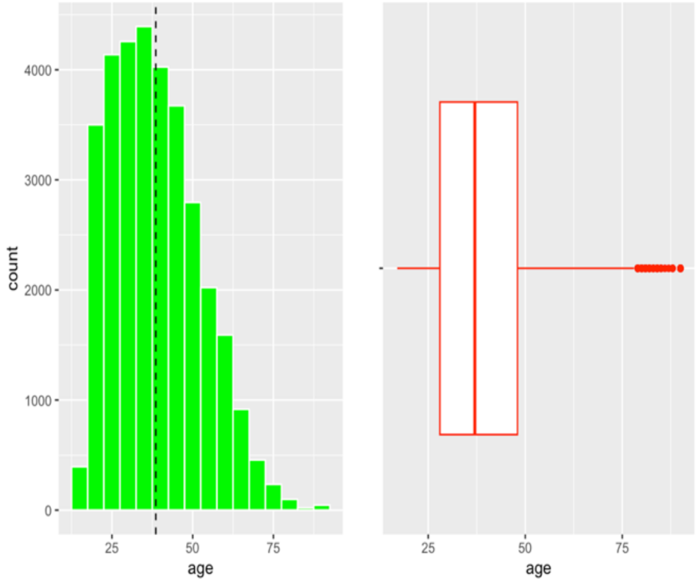
Quantitative Univariant Analysis

Now we will do the univariant analysis of the quantitative variables present in our data set:

(Number of people in the data set on: Y-axis (Count)

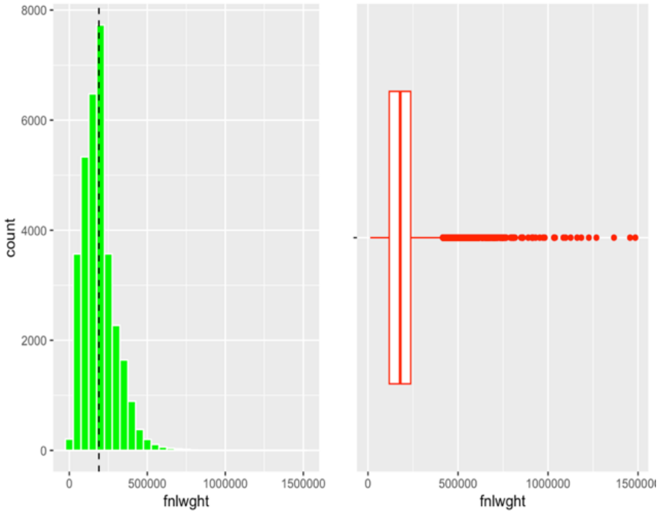
Distribution of the variable: X-axis)

Age



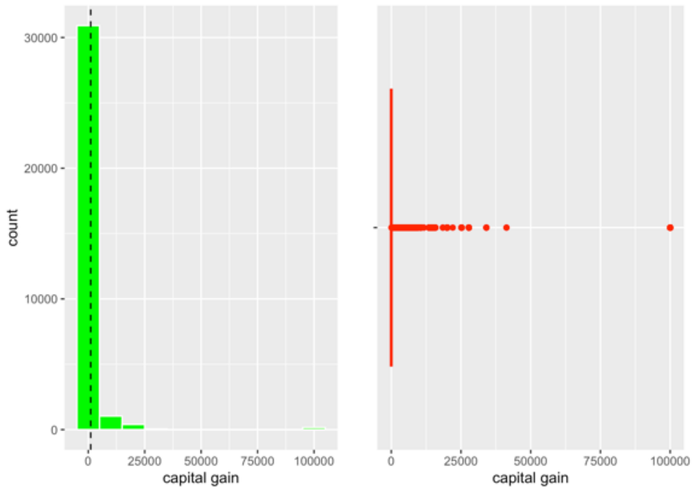
The figure shown is the distribution of the variable age. There is bar graph at the right and a box and whisker plot at the left. The variable of age seems to be right skewed, and outliers are present. The box in the box whisker plot seems to have a longer height due to the large amount of data gathered. We also see that the average age from this large Data seems to be 38 years.

Final Weight



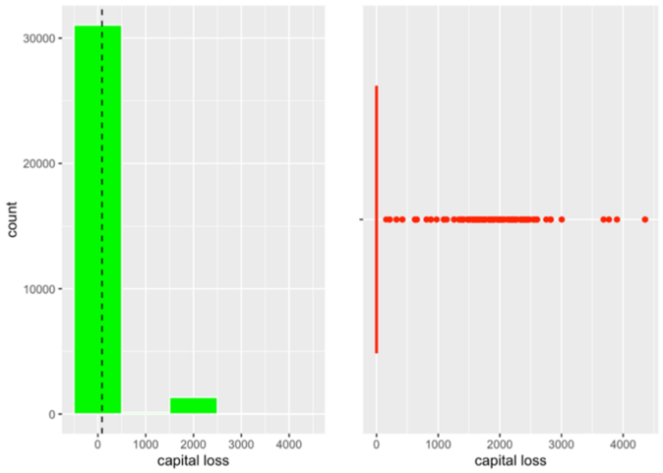
The final weight of a company refers to the significance or value placed on it compared to other companies in a particular context. Similar to the graph of age , due to the presence of a large data the height of the Box is large. The final weight graphs are rightly skewed and has outliers on the right side which we will Remove.

Capital Gain



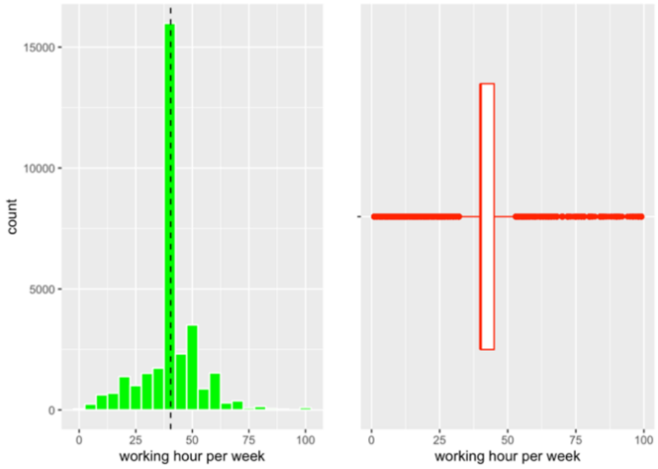
Capital gains are increases in the value of a capital asset as a result of it's own sale or exchange. From the graphs above we are able to interpret The data of the capital gain is rightly skewed with majority of the observations as 0 as capital gain and outliers are present.

Capital Loss



Capital losses are the decreases in the value of a capital asset as a result of its sale or exchange. The data presented here is similar to the capital gain with majority of the observations have 0 capital loss and the data being right skewed, due the presence of outliers on the right side of the box-whisker plot.

Working hours per week

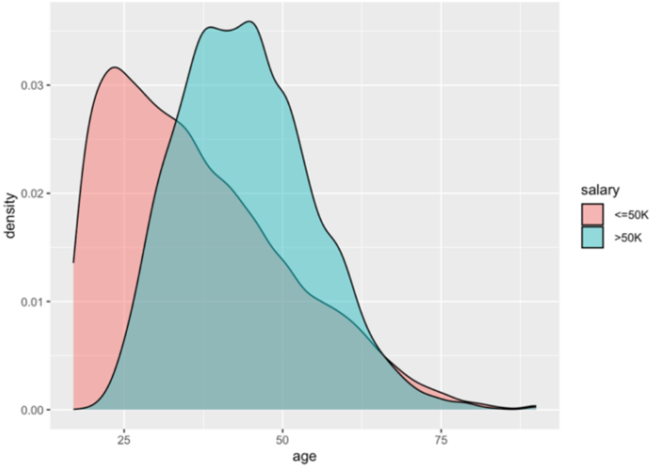


Working hours per week refers to the total number of hours an individual is expected to work during a 7-day period, usually calculated on a weekly basis . The data presented has a huge range of data that follows a normal distribution, due to the equal distribution of outliers on either side. This shows the average working hours is around 38 hours per week.

From Doing the Uni-variant analysis of the quantitative variables we learn how the data is spread and get an understand into the data and can make a link to the real life situations. To now understand the correlation of how our Dependant variable has a relation with our quantitative variables we do Bi Variant analysis

Quantitative Bi Variant Analysis

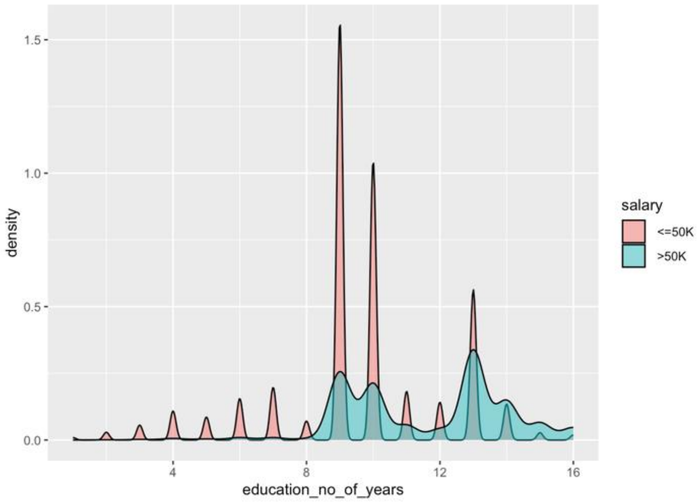
What is the distribution of age for different salary groups?



From the figure we understand that people who earn more than USD 50,000 salary are generally older having average age around 48 years. Additionally, people who have less than USD 50,000 salary have average age around 36 years. People who earn more than USD 50,000 have a rightly skewed distribution in contrast to the people that earn Below USD 50,000, which is Normally distributed.

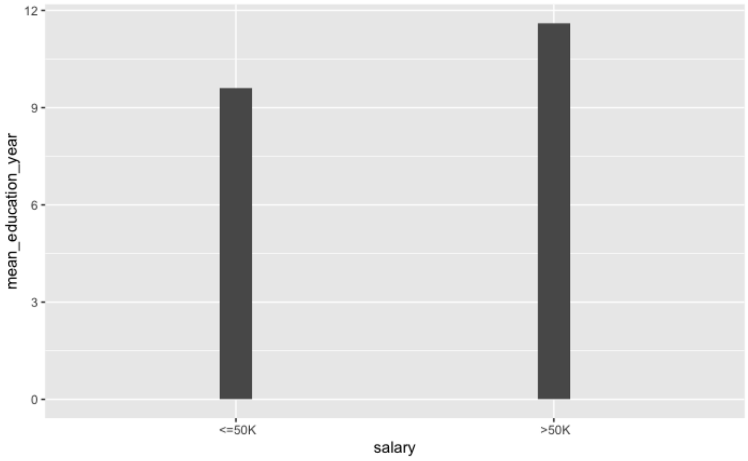
Only a few people below 25 years who have above USD 50,000 salary. Moreover, a Majority of people have having above USD 50,000 salary are of age ranging 35 to 60 years. Their working hours ranges from 36 to 60 hours per week.

How does the education years vary for different salary group?



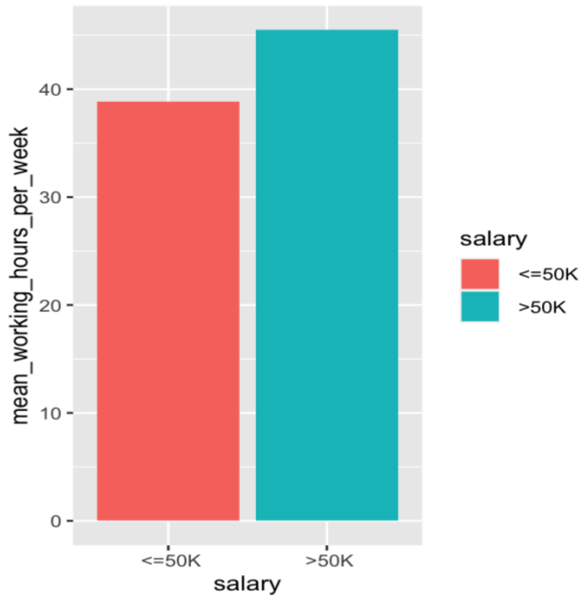
From the graph, we realize that the average education years is 10 years. Most of those people in the data having salary of above USD 50,000 have education years more than 8 years. There is peak at education years of 13 years for people having above USD 50,000 salary, indicating that large number of them have 13 years of education. There is no specific distribution here as education years can vary in a population.

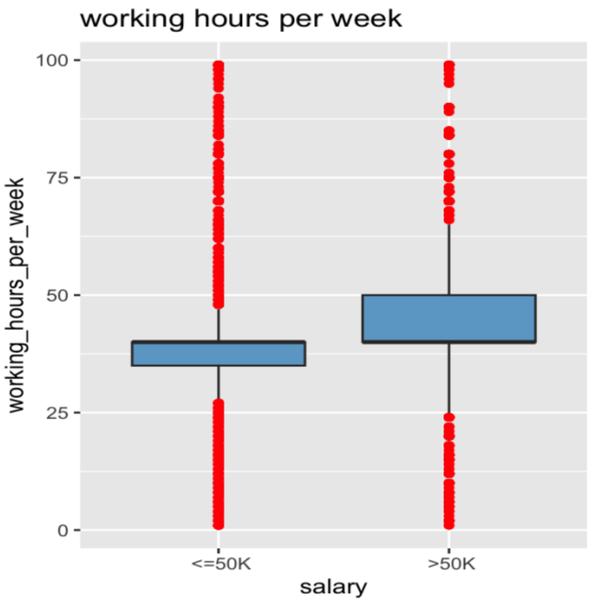
What are the average education years for different salary groups?

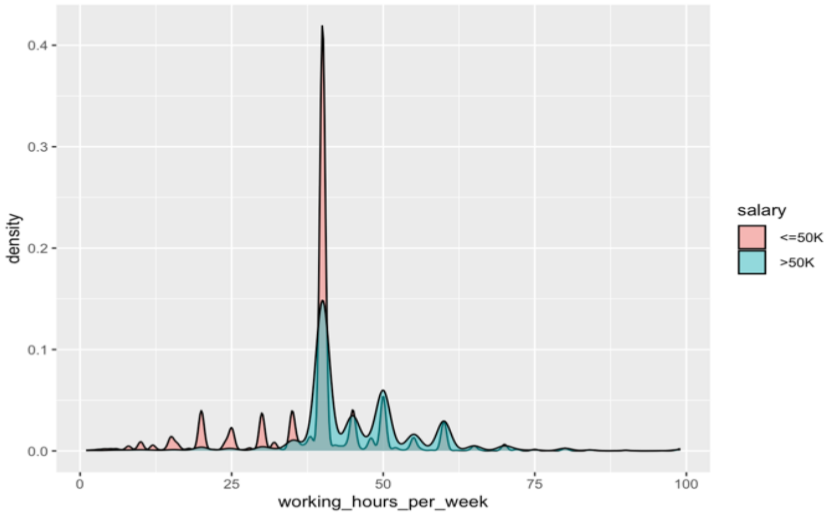


People with above USD 50,000 salary have an average of 11 years of education while the people with below USD 50,000 salary have on an average 9.5 years of education, indicating that people who are highly educated tend to have a higher salary than the rest.

How does average working hours vary for different salary group?

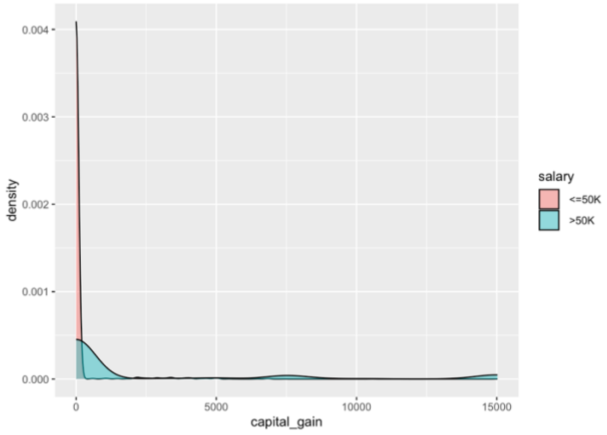






When we compare the hours worked with salary earned , people with above USD 50,000 salary on an average work 7 hours more weekly than people with less than 50K salary . Moreover , people with above USD 50,000 salary work on an average 45 hours per week. 50,000 salaries work more hours per week than majority of people with above USD 50,000 salary. Additionally, 75% of people with above USD 50,000 salary work more hours per week than 75% of people with less than USD 50,000 salary. 1st quartile of boxplot of above USD 50,000 group is higher than 3 quartiles of boxplot of below USD 50,000 group. Majority of people having above USD 50,000 salary work around 38 hours per week. This proves that more an individual works within a week the more he gets compensated.

How does distribution of capital gain vary for different salary group?



When we see the capital gain with salary, we learn that majority of people above 5000 capital gain have salary greater than USD 50,000 and majority of people with below USD 50,000 salary have less than 2500 capital gain.

- Outlier rules

There are several outliers present in the quantitative data we just analyzed. In statistics, outliers are often identified using the IQR method -

1. Calculate Interquartile range by subtracting Q1 from Q3.
2. Define the lower and upper bounds for outliers by showing :

Lower Outlier = Q1 – (1.5 x IQR)

Higher Outlier = Q3 + (1.5 x IQR)

We use the IQR method which is a useful tool for identifying outliers because it provides a simple and objective way of identifying extreme values while considering the distribution of the data. This helps to ensure that the results of statistical analysis are not unduly influenced by outliers, and provide a accurate result.

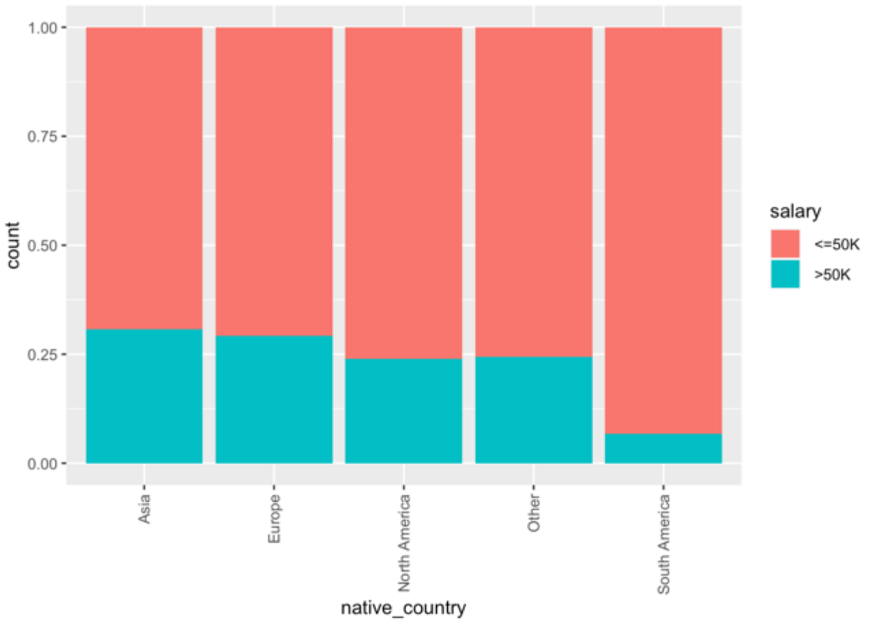
We now Analyze the Qualitative data of our data set by understanding their correlation by doing the Bi Variant analysis. Below are tables that summarize each variable it's aspects.

Qualitative Bi Variant Analysis

Qualitative Data

Native country	United States	Mexico	Unknown	Philippines	Germany	Canada	Others
	29170	643	583	198	137	121	1709

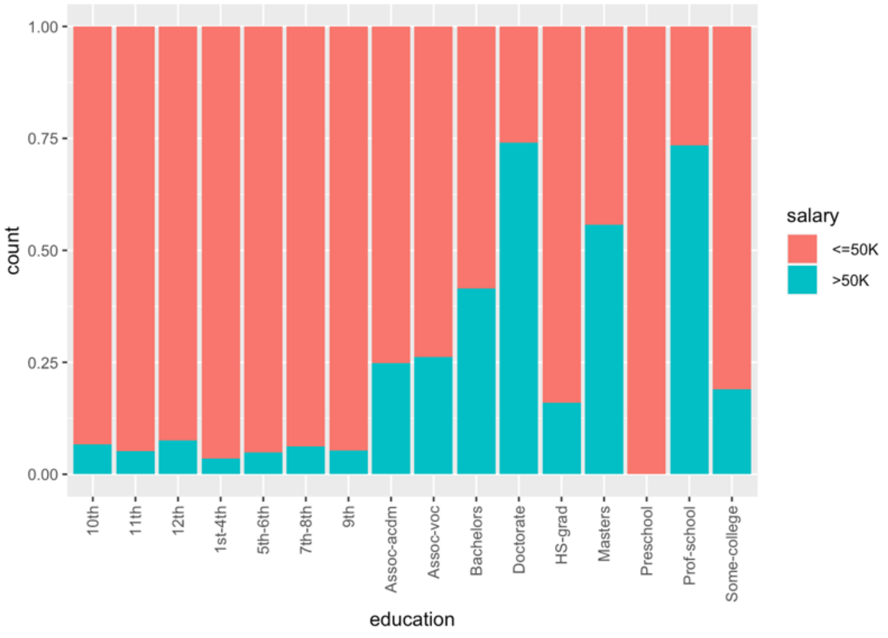
Native country in correlation to salary earned.



People from Asian and European countries have around 25% likelihood to have salary of above USD 50,000. Additionally, People from south America have a 5% chance of earning a salary of more than USD 50,000. Lastly, People from North America and other countries have around 23% chance of earning a compensation of above USD 50,000.

Education	HS-Grad	Some-College	Bachelors	Masters	Assoc-v	11th Grade	Others
	10501	7291	5355	1723	1382	1175	5134

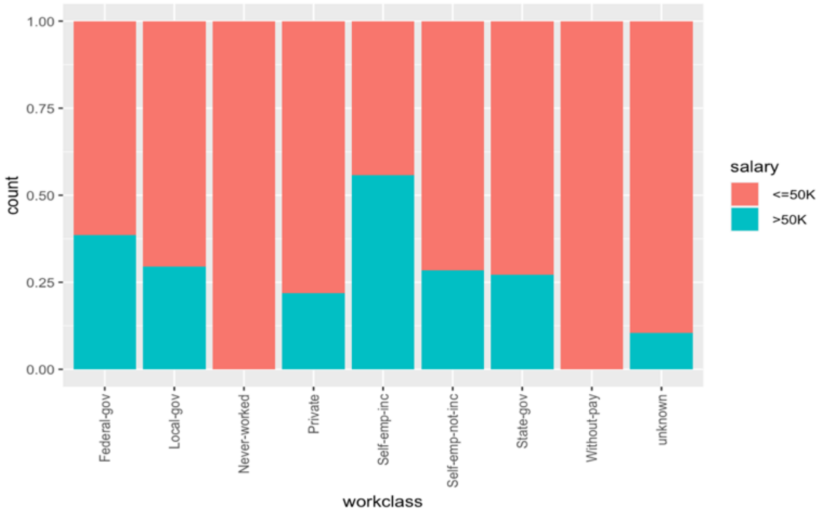
How does education affect a person’s salary?



This graph shows how a person’s educational achievement relates to the salary. From the data set we can understand that the people who are doctors, professors or having master’s degree have more than 50% likelihood of having above USD 50,000 salary. Additionally, people who have only gone to school have higher likelihood of having less than USD 50,000 salary. Majority people with above USD 50,000 salary tend to have higher education like doctorate, bachelor, and masters.

Work class	Private	Self-emp-not-inc	Local-Gov	Unknown	State-gov	Self-emp-in	Others
	22696	2541	2093	1836	1298	1116	981

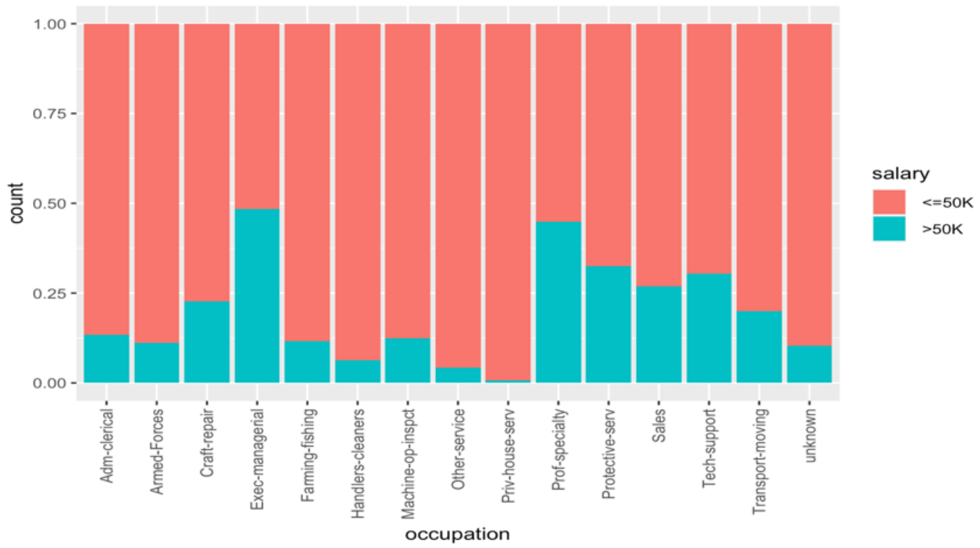
Work Class In relation to Salary earned



From the graphs above we understand in work class self-employed people have 50% chance of having above USD 50,000 salary and federal-government employee have 30% likelihood of having above 50K salary. Moreover, unemployed people who have never worked or without-pay earns less than USD 50,000. People who work in the private sector have around a 23% chance to earn more than USD 50,00 while people who work for the state government have a 25% chance to earn more than USD 50,000.

	Prof-Speciality	Craft Repair	Exec managerial	Adm - Clerical	Sales	Other services	Others
Occupation	4140	4099	4066	3770	3650	3295	9541

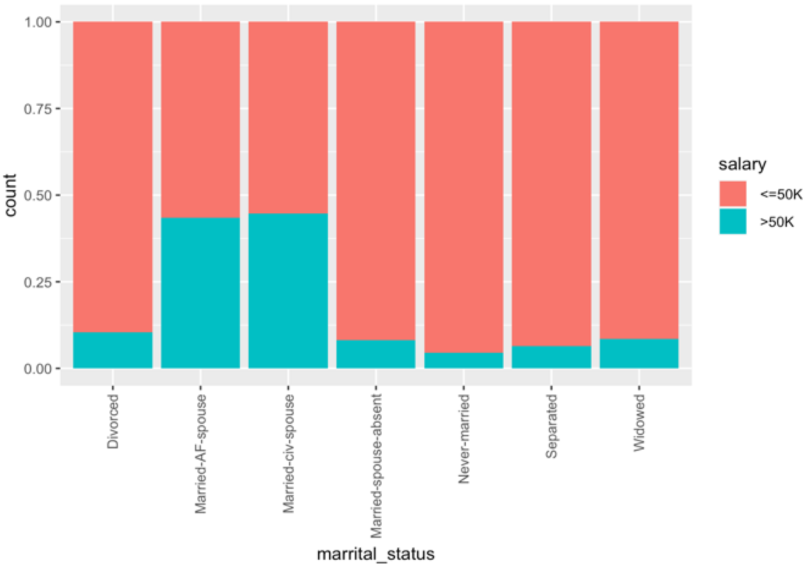
How does Salary vary across different occupation?



The occupation in which a person works in is one of the critical aspects while fixing the persons salary. From the figure above an executive manager, tech – support, Professor of a specialty have more than 40% likelihood of having a salary above 50K salary. People with occupations like private House service, handles-cleaners, farming-fishing have higher likelihood of having below USD 50,000 salary.

Marital status	Divorced	Married-AF-spouse	Married-civ-spouse	Married-spouse-absent	Never married	Separated	Widowed
	4443	23	14976	418	10638	1025	993

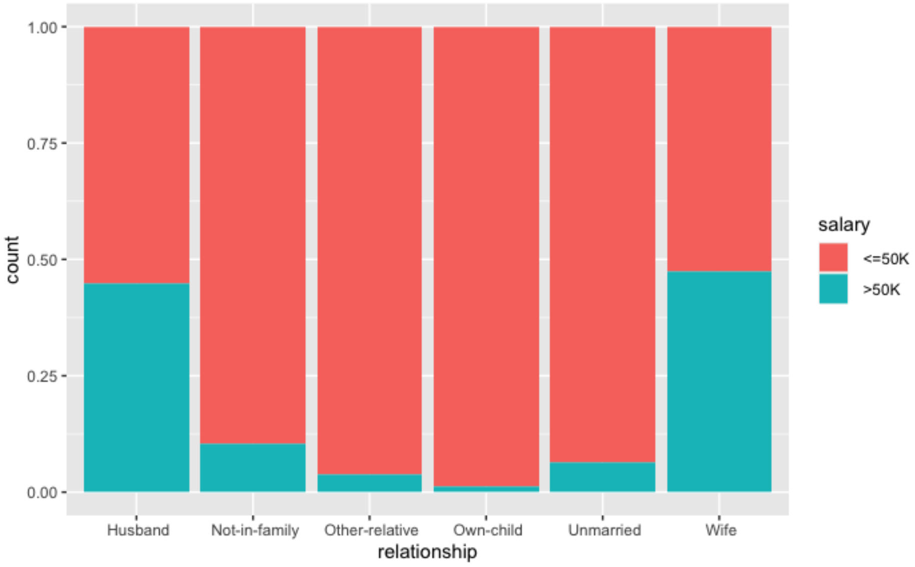
Marital status In relation to salary earned



Married people have higher likelihood of getting above USD 50,000 salary in terms of marital status in contrast to the people who have never married, who have least likelihood of getting above USD 50,000 salary. Divorced and Widowed People have a 15% chance of getting a salary above USD, unlike the Couples who are separated.

Relationship	Husband	Not-in-Family	Other-relative	Own-Child	Unmarried	Wife
	13193	8305	981	5068	3446	1568

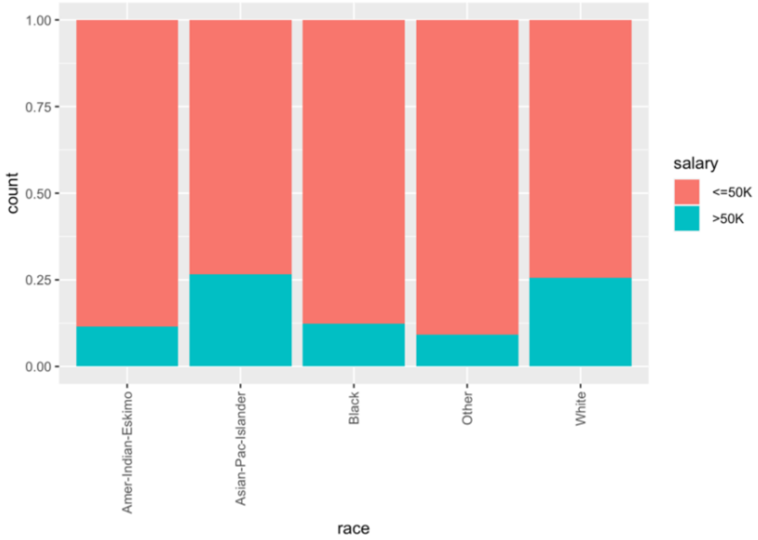
<u>Relationship in relation to salary</u>



Like what we saw in marital status, in relationship we understand that husbands and wives are almost 50% likely to earn a salary above USD 50,000. Unmarried people, people who are not in a family, and other relatives are around have around a 15% likelihood of earning a salary of above USD 50,000. Someone with an own child is the least likely to earn a high salary.

Race	Amer-Indian-Eskimo	Asian-pac-islander	Black	Other	White
	4140	4099	4066	3770	3650

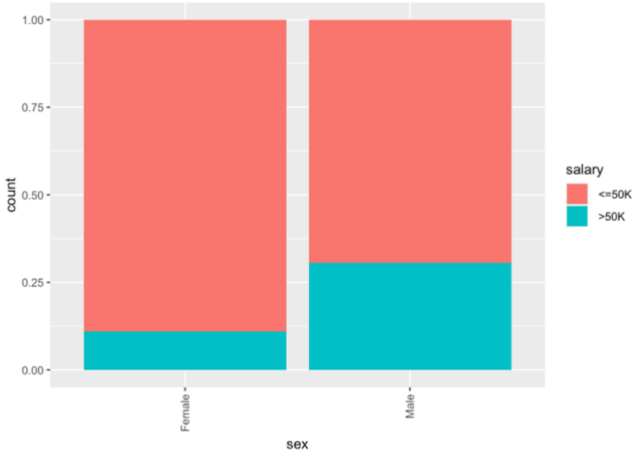
How does Race affect Salary?



Race can sometimes have an impact on an individual's salary, as studies have shown that racial discrimination and bias can result in disparities in pay and opportunities for advancement . understanding this what the simple figure above explains to us is that white and Asia – pacific people have around 25% likelihood of having more than 50K salary which is higher than another race.

Sex	Male	Female
	21790	10771

Sex in relation to salary earned



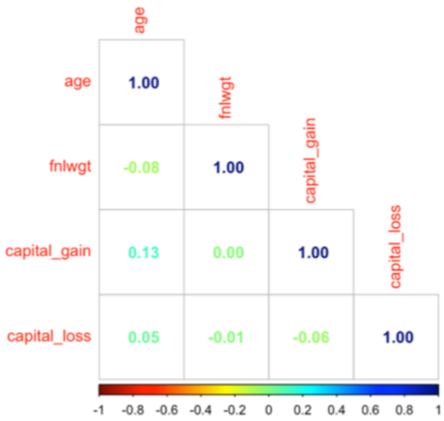
From the graph above, in terms of the sex of an individual we understand that males have a higher likelihood of 25%-37.5% to earn above USD 50,000; while females have a 15% likelihood of earning above USD 50,000.

Salary	Above or equal to 50,000	Below 50,000
	29170	643

The salary is a skewed variable; most of the people earn **less than** USD 50,000 and very few earn above USD 50,000.

Correlation

Let’s see the **correlation** between these variables



Heat maps are graphical data representations that use color-coding to depict the density of data points. Heat maps can help to identify patterns and outliers in the data. From what we observe we can clearly state that there is no high correlation found between any numerical column, as the colors shown are all cool colors indicating there is no high density found between any quantitative variables.

The Bi Variant analysis and the correlations showed us how each qualitative variable affects the DV, and the graphs helped us interpret the chance (%) of a person with certain characteristics to earn above the limit.

4 Hypothesis testing

To check the relationships of all the Independent Variables with the Dependant Variable that is Salary.

The Chi Square test is used to determine whether or not two categorical variables have a significant relationship. It is a non-parametric statistical test designed to analyse the relationship between two variables without making assumptions about the underlying data distribution. The test is based on a contingency table comparison of observed and anticipated frequencies.

The null hypothesis of the Chi Square test is that there is no association between the two variables, which implies that the distribution of observations between the categories for both variables is the same. If the calculated Chi Square statistic is greater than the critical value from the Chi Square distribution, the null hypothesis is rejected, and a significant connection between the two variables is found.

5 The Chi-square formula

The Chi Square formula was derived from the concept of maximum likelihood estimation. Maximum likelihood estimation is a technique used to estimate the parameters of a statistical model using observable data. The objective of maximum likelihood estimation is to identify the parameter values that make the observed data most probable.

The formula for chi-square is given as:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where,

- **χ^2 = Chi-Square value**
- **O_i = Observed frequency**
- **E_i = Expected frequency**

Carrying out the Chi-square test.

For this situation:

Null hypothesis: that indicates that no statistical significance exists, is that there is a relationship of all attributes to the salary.

Alternative hypothesis: that suggests there is statistical significance is that there is no relationship between any or all variables to the salary.

Row	Column	Chi-Square	Degrees of freedom	P-value
Salary	Work class	1045.7086	8	2.03 x 10^{-220}
Salary	Education	4429.6533	15	0.00 x 10^{00}
Salary	Occupation	4031.97428	14	0.00 x 10^{00}
Salary	Relationship	6699.0769	5	0.00 x 10^{00}
Salary	Race	330.92043	4	2.31 x 10^{-70}
Salary	Native country	43.34204	4	8.79 x 10^{-09}

What we understand from the table above is that the relationship between all the categorical variable with salary are significant as the p-value is less than the significance level of 0.05; and therefore, we must fail to reject the null hypothesis. Implying that there is a relationship between all the variables that attribute to the salary.

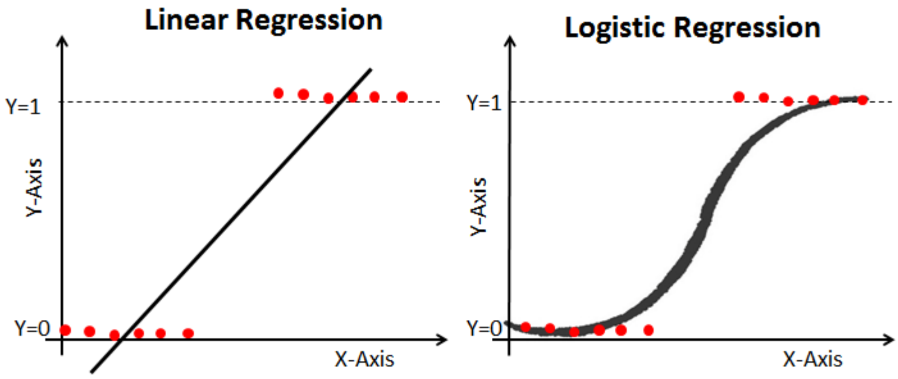
5 Model Building

- **Logistic Regression**

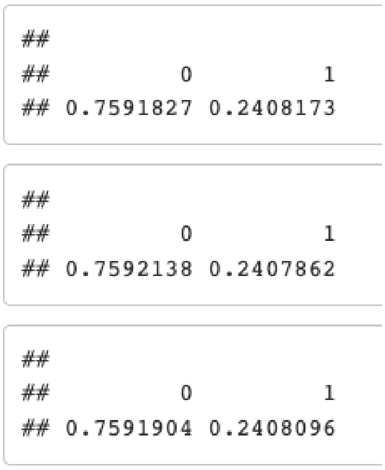
Build Logistic Regression Model to check the significant variables which influence the Salary.

For this situation, when the dependent variables are binary, logistic regression is the appropriate regression analysis. Binary data is when there are only two possible values.

Logistic regression is a statistical method for modeling the relation between a binary dependent variable and one or more independent variables. The core objective of logistic regression is to develop a mathematical model that can be used to predict the probability of the binary dependent variable based on the values of the independent variables.



To Further Move on with the building of the model we need to create a train and test data sets , a train and test dataset is a method used to assess the performance of a statistical model. The dataset is split into two parts: a training set and a testing set. The training set is used to build the model, while the testing set is used to evaluate the performance of the model on new, unseen data. A common split is to use 80% of the data for training and 20% for testing,



```
## 
##         0         1 
## 0.7591827 0.2408173 
```

```
## 
##         0         1 
## 0.7592138 0.2407862 
```

```
## 
##         0         1 
## 0.7591904 0.2408096 
```

To make the model, to see if it works on a train and test set and to see if it is accurate and to understand the model in better detail, I have used a computer software- **“R-studio”** . The outcome is seen in the image above .

- This shows Both the train and test split have 24% positive cases that is people with above USD 50,000 salary.
- The baseline accuracy is 76% because if we were to predict 0 (less than USD 50,000 salary) class, we will be correct 76% of the time.

Building the logistic regression model for all variables

- **Generalized Linear models**

A generalized linear model (GLM) is a flexible and powerful statistical method used to model the relationship between a dependent variable and one or more independent variables.
To develop the model and understand the accuracy of it , I have used a computer software known as "R-studio" . The results of the model on the software showed that the intercept value is in the **negative** (-6.682), which suggests this model is **over predicting.**

This means the model has predicted values that are larger than the actual observed values for a given set of input variables .This results in a higher residuals and can lead to a poorer fit of the model to the data .

For modifying the structures we need to add or remove variables , or predict another different model .

- Significant variables in the model are :
  - age
  - work class
  - Final weight
  - education
  - Marital status
  - occupation
  - capital gain
  - capital loss

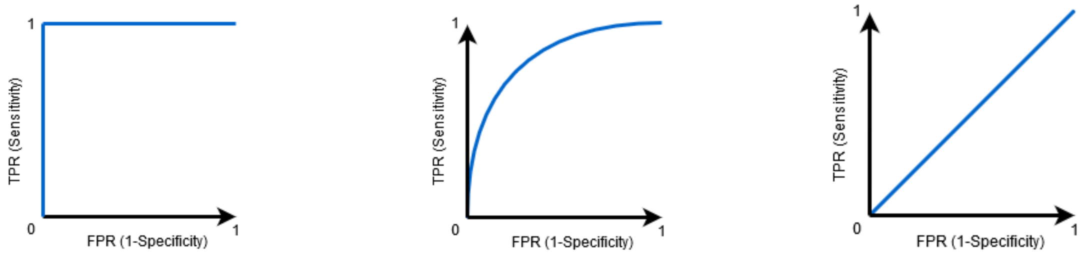
This suggests that "Race" was a variable that was insignificant that caused the model to over-predict and hence we need to get rid of it.

6 Checking various parameters to identify right model

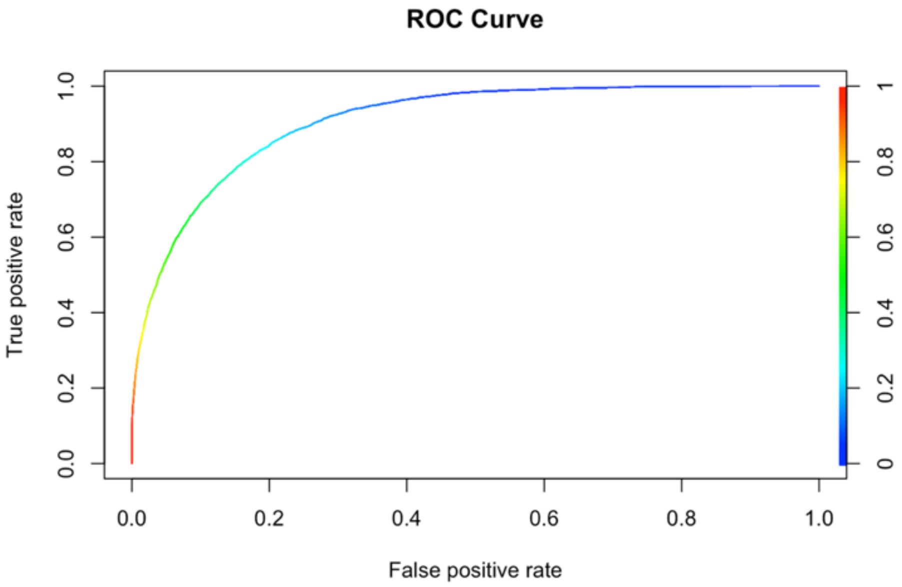
- **ROC curve, AUC, Gini and KS**

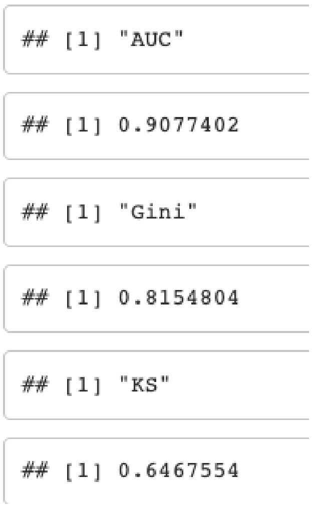
ROC and AUC

An ROC (Receiver Operating Characteristic) curve is a graphical depiction of a binary classifier system's performance while the discrimination threshold is changed. The ROC curve visualizes the trade-off between the true positive rate and the false positive rate, and allows you to choose a threshold that balances these two measurements. The Area Under the Curve (AUC) is used to describe the ROC curve and is a measure of a classifier's ability to discriminate between classes. The higher the AUC, the better the model differentiates between positive and negative categories.



When AUC = 1, the classifier is capable of effectively distinguishing between all Positive and Negative class points. When 0.5<AUC<1, there is a good possibility that the classifier will be able to distinguish between positive and negative class values. This is due to the classifier detecting more True positives and True negatives than False negatives and False positives. The greater a classifier's AUC value, the better its ability to distinguish between positive and negative classes.





```
## [1] "AUC"
```

```
## [1] 0.9077402
```

```
## [1] "Gini"
```

```
## [1] 0.8154804
```

```
## [1] "KS"
```

```
## [1] 0.6467554
```

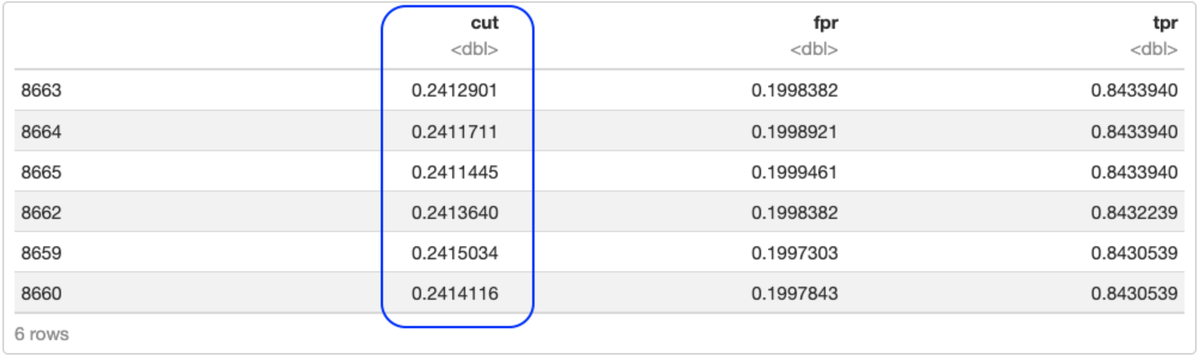
From what we see the area under the curve is 91%.

Gini is 81%

KS is 65%.

The ROC curves indicate the trade-off between sensitivity (or TPR) and specificity (1-FPR). The area under the ROC curve measures how well a parameter may distinguish between two diagnostic groups above USD 50,000 and under USD 50,000. we see from the curve we can reach ~0.8 tpr at fpr ~0.2

Checking the probability cut-off at fpr (false positive rate) threshold=0.2 to find maximum tpr (true positive rate)



	cut <dbl>	fpr <dbl>	tpr <dbl>
8663	0.2412901	0.1998382	0.8433940
8664	0.2411711	0.1998921	0.8433940
8665	0.2411445	0.1999461	0.8433940
8662	0.2413640	0.1998382	0.8432239
8659	0.2415034	0.1997303	0.8430539
8660	0.2414116	0.1997843	0.8430539

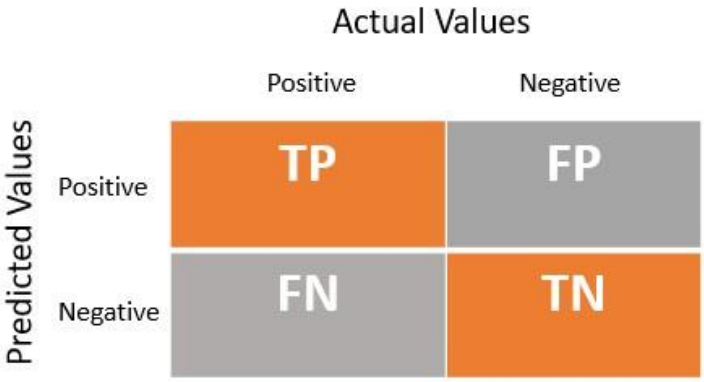
6 rows

Taking the cut off of 0.24 to have high sensitivity for logistic regression model to be used for creating a confusion matrix.

7 Confusion Matrix

A confusion matrix is a table that is used to evaluate a classifier's performance. It summarises the number of correct and incorrect predictions made by the classifier and aids in understanding the types of errors made. The matrix is made by comparing the predicted class labels to the test data's actual class labels.

The confusion matrix is used for Binary classification problem to find three important parameters namely Accuracy, Sensitivity & Specificity.



TP= True positive
TN= True Negative
FP= False Positive
FN=False Negative

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad Specificity = \frac{TN}{TN + FP} \qquad Sensitivity = \frac{TP}{TP + FN}$$

Train Data v/s Test Data

Train Data	Test Data
The sensitivity is 84% and specificity is 79.9% with cut-off of 0.24	Having a 84% sensitivity in this case means that we are able to correctly identify 84% of the people who are going to earn salary of above 50K.

- Now we can focus our efforts on a smaller group of people (95+626 = 721) and be able to take more effective measures.
- log model performs very well both on the test and train data. We will use log model as the final model.

8 Prescriptive Statistical Analysis - Conclusion

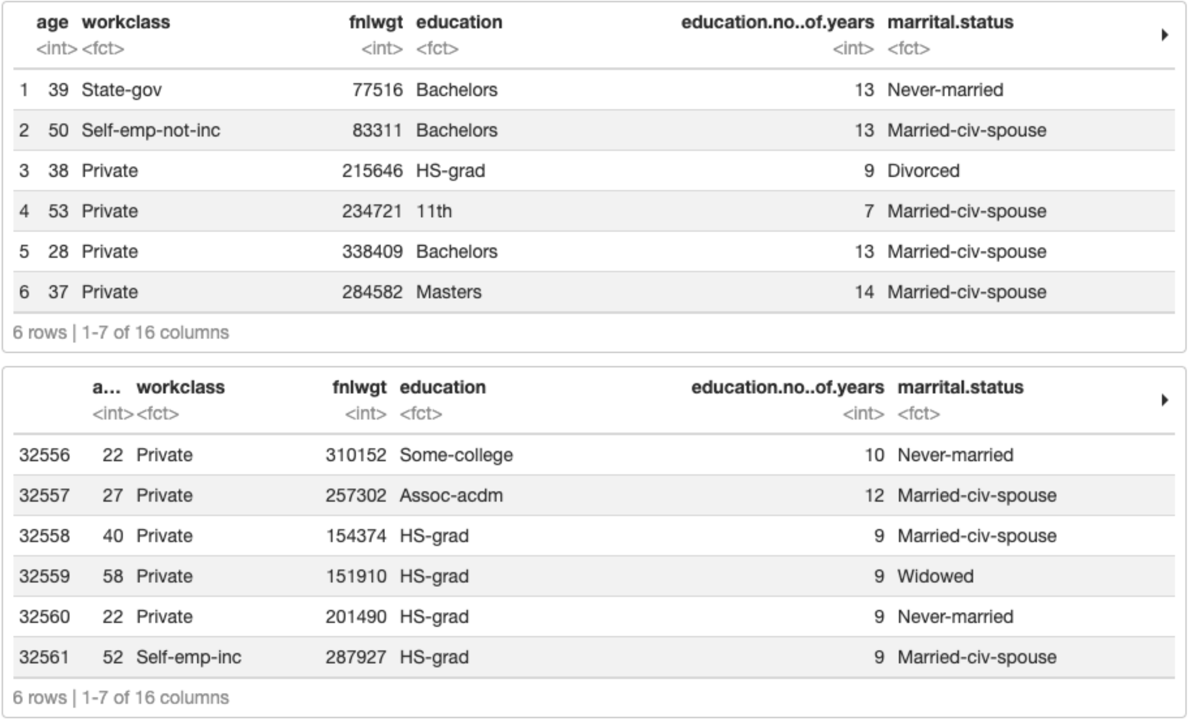
The logistic regression model has **6%** more accuracy than the baseline model . The accuracy on the training data is **85%** while the Accuracy on the test data is **82%.** The accuracy for the test data is within the **10%** accuracy which the model gets on training data set , this tells us that the model is not an overfit.

Most important variables in determining if a person will earn more than USD 50,000 include capital gain, working hours per week, capital loss, age, education (professor, Doctorate, masters).

This assignment has given me an interesting insight and a better grip into the statistical analytic world of mathematics. This investigation has given me the chances of exploring concepts of statistics , learnings of a software which I isn’t part of my school curriculum.

There were few instances where I could not find an accurate model for my study , for which I had to try and test various different models over time. I was guided by my superior to derive the required outcome . I feel that I could have added the process of Akaike Information Criterion (AIC), to further analyse to show the insignificant variables. Overall, the process of writing an Extended essay has led me to gain knowledge about mathematical concepts and has given me the opportunity to begin exploring the vast ocean of statistical studies.

9 APPENDIX



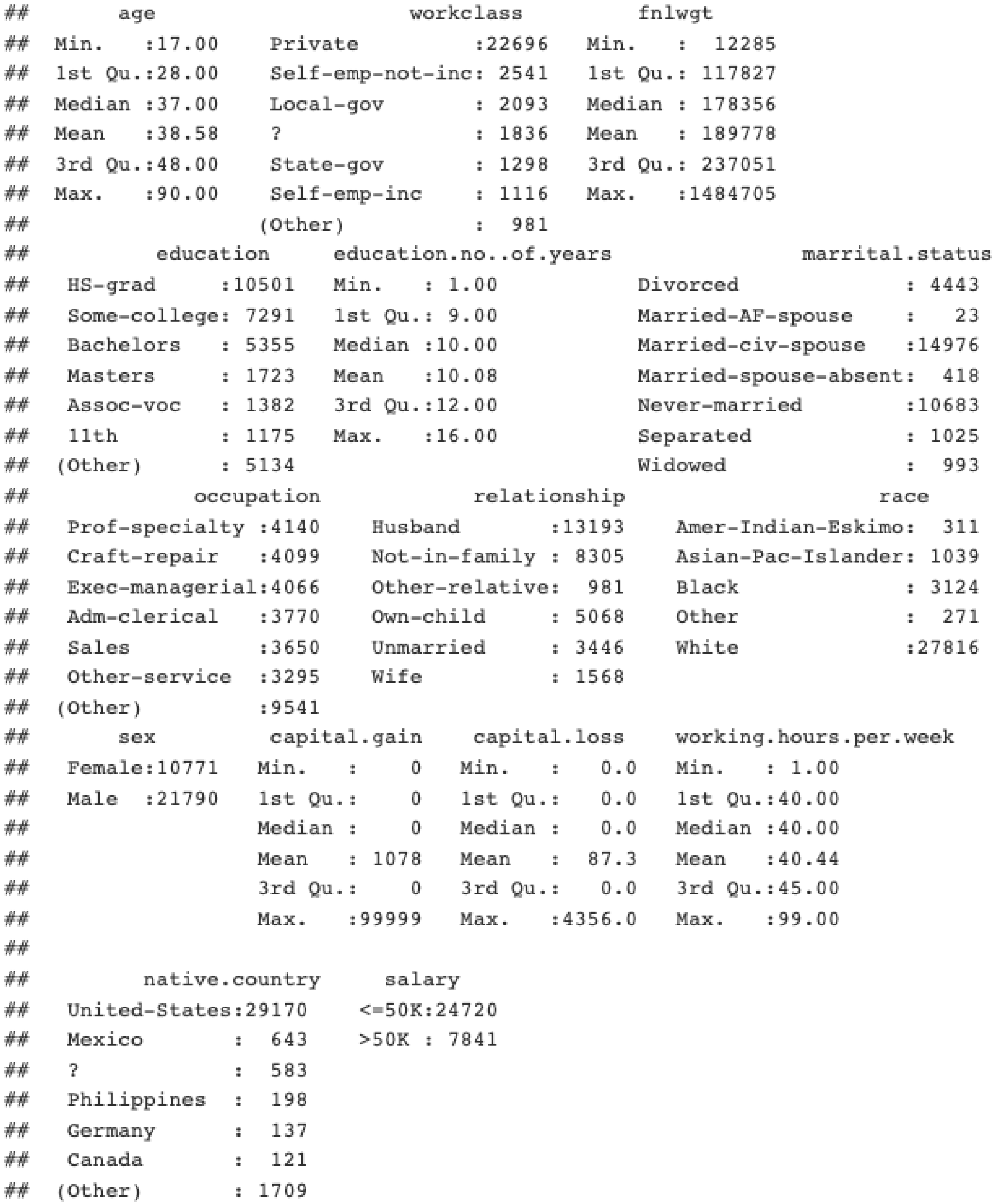
	age <int>	workclass <fct>	fnlwgt <int>	education <fct>	education.no..of.years <int>	marrital.status <fct>
1	39	State-gov	77516	Bachelors	13	Never-married
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse
3	38	Private	215646	HS-grad	9	Divorced
4	53	Private	234721	11th	7	Married-civ-spouse
5	28	Private	338409	Bachelors	13	Married-civ-spouse
6	37	Private	284582	Masters	14	Married-civ-spouse

6 rows | 1-7 of 16 columns

	a... <int>	workclass <fct>	fnlwgt <int>	education <fct>	education.no..of.years <int>	marrital.status <fct>
32556	22	Private	310152	Some-college	10	Never-married
32557	27	Private	257302	Assoc-acdm	12	Married-civ-spouse
32558	40	Private	154374	HS-grad	9	Married-civ-spouse
32559	58	Private	151910	HS-grad	9	Widowed
32560	22	Private	201490	HS-grad	9	Never-married
32561	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse

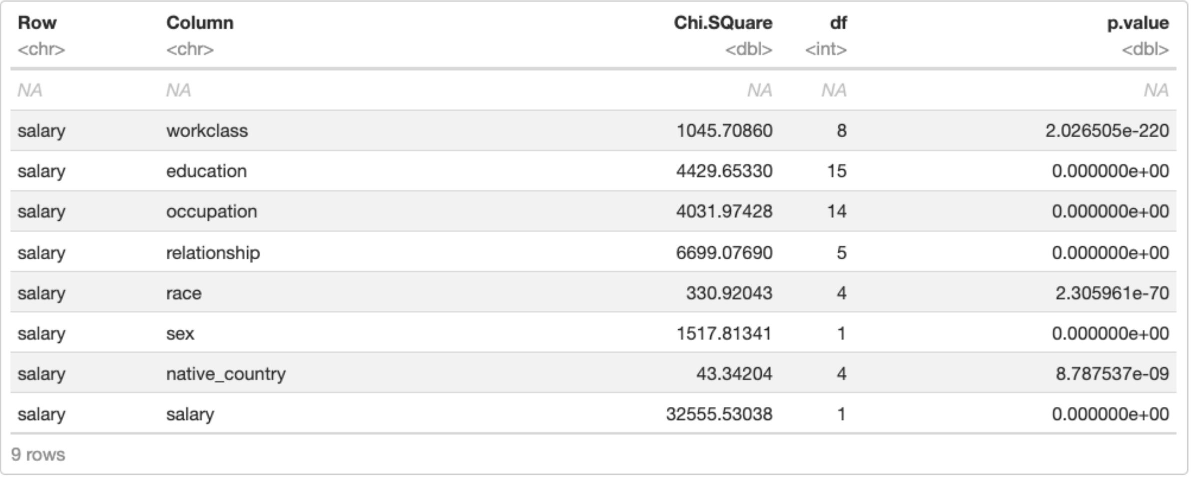
6 rows | 1-7 of 16 columns

Section of the Data set, In the computer software



```
##       age                   workclass          fnlwgt
##  Min.   :17.00    Private         :22696   Min.   :  12285
##  1st Qu.:28.00    Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00    Local-gov       : 2093   Median : 178356
##  Mean   :38.58    ?               : 1836   Mean   : 189778
##  3rd Qu.:48.00    State-gov       : 1298   3rd Qu.: 237051
##  Max.   :90.00    Self-emp-inc    : 1116   Max.   :1484705
##                  (Other)          :  981
##          education     education.no..of.years             marrital.status
##   HS-grad      :10501   Min.   : 1.00           Divorced              : 4443
##   Some-college: 7291   1st Qu.: 9.00           Married-AF-spouse     :   23
##   Bachelors    : 5355   Median :10.00           Married-civ-spouse    :14976
##   Masters      : 1723   Mean   :10.08           Married-spouse-absent:  418
##   Assoc-voc    : 1382   3rd Qu.:12.00           Never-married         :10683
##   11th         : 1175   Max.   :16.00           Separated             : 1025
##  (Other)       : 5134                           Widowed               :  993
##             occupation           relationship                   race
##   Prof-specialty :4140    Husband       :13193    Amer-Indian-Eskimo:  311
##   Craft-repair   :4099    Not-in-family : 8305    Asian-Pac-Islander: 1039
##   Exec-managerial:4066    Other-relative:  981    Black             : 3124
##   Adm-clerical   :3770    Own-child     : 5068    Other             :  271
##   Sales          :3650    Unmarried     : 3446    White             :27816
##   Other-service  :3295    Wife          : 1568
##  (Other)         :9541
##       sex          capital.gain    capital.loss    working.hours.per.week
##   Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##   Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                  Median :    0   Median :   0.0   Median :40.00
##                  Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                  3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                  Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##        native.country     salary
##   United-States:29170    <=50K:24720
##   Mexico       :  643    >50K : 7841
##   ?            :  583
##   Philippines  :  198
##   Germany      :  137
##   Canada       :  121
##  (Other)       : 1709
```

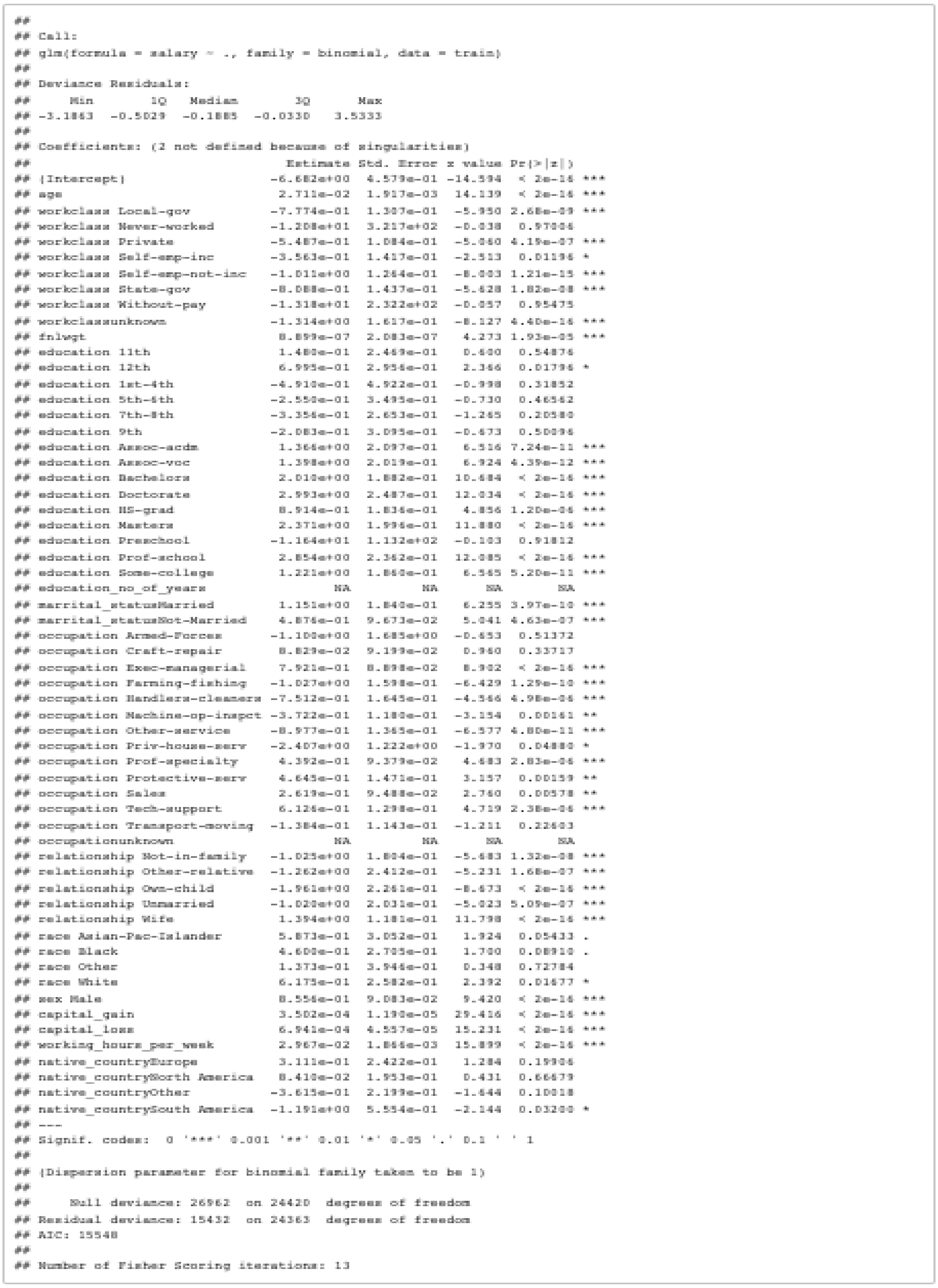
Structure of the Data set in the computer software



Row	Column	Chi.SQuare	df	p.value
<chr>	<chr>	<dbl>	<int>	<dbl>
NA	NA	NA	NA	NA
salary	workclass	1045.70860	8	2.026505e-220
salary	education	4429.65330	15	0.000000e+00
salary	occupation	4031.97428	14	0.000000e+00
salary	relationship	6699.07690	5	0.000000e+00
salary	race	330.92043	4	2.305961e-70
salary	sex	1517.81341	1	0.000000e+00
salary	native_country	43.34204	4	8.787537e-09
salary	salary	32555.53038	1	0.000000e+00

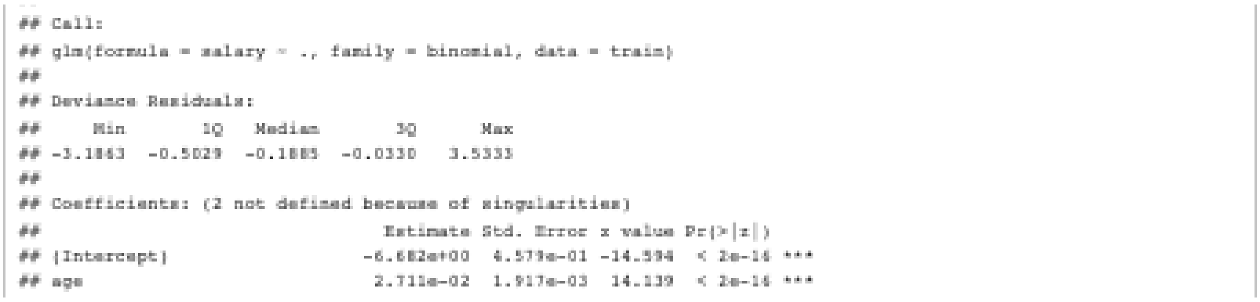
9 rows

Chi-Square result in the computer software



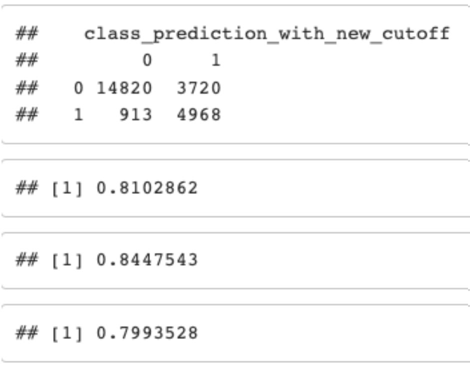
```
##
## Call:
## glm(formula = salary ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1863  -0.5029  -0.1885  -0.0330   3.5333
##
## Coefficients: (2 not defined because of singularities)
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -6.682e+00  4.579e-01 -14.594  < 2e-16 ***
## age                           2.711e-02  1.917e-03  14.139  < 2e-16 ***
## workclass Local-gov          -7.774e-01  1.307e-01  -5.950 2.68e-09 ***
## workclass Never-worked       -1.208e+01  3.217e+02  -0.038  0.97006
## workclass Private            -5.487e-01  1.084e-01  -5.060 4.19e-07 ***
## workclass Self-emp-inc       -3.563e-01  1.417e-01  -2.513  0.01196 *
## workclass Self-emp-not-inc   -1.011e+00  1.264e-01  -8.003 1.21e-15 ***
## workclass State-gov          -8.088e-01  1.437e-01  -5.628 1.82e-08 ***
## workclass Without-pay        -1.318e+01  2.322e+02  -0.057  0.95475
## workclassunknown             -1.314e+00  1.617e-01  -8.127 4.40e-16 ***
## fnlwgt                        8.899e-07  2.083e-07   4.273 1.93e-05 ***
## education 11th                1.480e-01  2.469e-01   0.600  0.54876
## education 12th                6.995e-01  2.956e-01   2.366  0.01796 *
## education 1st-4th            -4.910e-01  4.922e-01  -0.998  0.31852
## education 5th-6th            -2.550e-01  3.495e-01  -0.730  0.46562
## education 7th-8th            -3.356e-01  2.653e-01  -1.265  0.20580
## education 9th                -2.083e-01  3.095e-01  -0.673  0.50096
## education Assoc-acdm          1.366e+00  2.097e-01   6.516 7.24e-11 ***
## education Assoc-voc           1.398e+00  2.019e-01   6.924 4.39e-12 ***
## education Bachelors           2.010e+00  1.882e-01  10.684  < 2e-16 ***
## education Doctorate           2.993e+00  2.487e-01  12.034  < 2e-16 ***
## education HS-grad             8.914e-01  1.836e-01   4.856 1.20e-06 ***
## education Masters             2.371e+00  1.996e-01  11.880  < 2e-16 ***
## education Preschool          -1.164e+01  1.132e+02  -0.103  0.91812
## education Prof-school         2.854e+00  2.362e-01  12.085  < 2e-16 ***
## education Some-college        1.221e+00  1.860e-01   6.565 5.20e-11 ***
## education_no_of_years                NA         NA      NA       NA
## marrital_statusMarried        1.151e+00  1.840e-01   6.255 3.97e-10 ***
## marrital_statusNot-Married    4.876e-01  9.673e-02   5.041 4.63e-07 ***
## occupation Armed-Forces      -1.100e+00  1.685e+00  -0.653  0.51372
## occupation Craft-repair       8.829e-02  9.199e-02   0.960  0.33717
## occupation Exec-managerial    7.921e-01  8.898e-02   8.902  < 2e-16 ***
## occupation Farming-fishing   -1.027e+00  1.598e-01  -6.429 1.29e-10 ***
## occupation Handlers-cleaners -7.512e-01  1.645e-01  -4.566 4.98e-06 ***
## occupation Machine-op-inspct -3.722e-01  1.180e-01  -3.154  0.00161 **
## occupation Other-service     -8.977e-01  1.365e-01  -6.577 4.80e-11 ***
## occupation Priv-house-serv   -2.407e+00  1.222e+00  -1.970  0.04880 *
## occupation Prof-specialty     4.392e-01  9.379e-02   4.683 2.83e-06 ***
## occupation Protective-serv    4.645e-01  1.471e-01   3.157  0.00159 **
## occupation Sales              2.619e-01  9.488e-02   2.760  0.00578 **
## occupation Tech-support       6.126e-01  1.298e-01   4.719 2.38e-06 ***
## occupation Transport-moving  -1.384e-01  1.143e-01  -1.211  0.22603
## occupationunknown                    NA         NA      NA       NA
## relationship Not-in-family   -1.025e+00  1.804e-01  -5.683 1.32e-08 ***
## relationship Other-relative  -1.262e+00  2.412e-01  -5.231 1.68e-07 ***
## relationship Own-child       -1.961e+00  2.261e-01  -8.673  < 2e-16 ***
## relationship Unmarried       -1.020e+00  2.031e-01  -5.023 5.09e-07 ***
## relationship Wife             1.394e+00  1.181e-01  11.798  < 2e-16 ***
## race Asian-Pac-Islander       5.873e-01  3.052e-01   1.924  0.05433 .
## race Black                    4.600e-01  2.705e-01   1.700  0.08910 .
## race Other                    1.373e-01  3.946e-01   0.348  0.72784
## race White                    6.175e-01  2.582e-01   2.392  0.01677 *
## sex Male                      8.556e-01  9.083e-02   9.420  < 2e-16 ***
## capital_gain                  3.502e-04  1.190e-05  29.416  < 2e-16 ***
## capital_loss                  6.941e-04  4.557e-05  15.231  < 2e-16 ***
## working_hours_per_week        2.967e-02  1.866e-03  15.899  < 2e-16 ***
## native_countryEurope          3.111e-01  2.422e-01   1.284  0.19906
## native_countryNorth America   8.410e-02  1.953e-01   0.431  0.66679
## native_countryOther          -3.615e-01  2.199e-01  -1.644  0.10018
## native_countrySouth America  -1.191e+00  5.554e-01  -2.144  0.03200 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26962  on 24420  degrees of freedom
## Residual deviance: 15432  on 24363  degrees of freedom
## AIC: 15548
##
## Number of Fisher Scoring iterations: 13
```

Generalized linear model in the computer software



```
## Call:
## glm(formula = salary ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1843  -0.5029  -0.1885  -0.0330   3.5333
##
## Coefficients: (2 not defined because of singularities)
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -6.682e+00  4.579e-01 -14.594  < 2e-16 ***
## age                         2.711e-02  1.917e-03  14.139  < 2e-16 ***
```

Intercept value found in the GLM

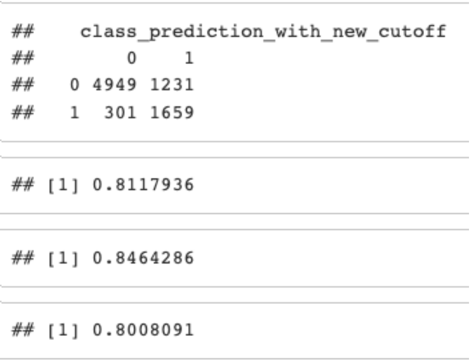


```
##    class_prediction_with_new_cutoff
##         0     1
##    0 14820  3720
##    1   913  4968
```

```
## [1] 0.8102862
```

```
## [1] 0.8447543
```

```
## [1] 0.7993528
```



```
##    class_prediction_with_new_cutoff
##        0    1
##    0 4949 1231
##    1  301 1659
```

```
## [1] 0.8117936
```

```
## [1] 0.8464286
```

```
## [1] 0.8008091
```

Train and test set for the model in the computer software

10 . CITATIONS

1 Frankenfield, Jake. “How Descriptive Analytics Work.” *Investopedia*, 24 June 2019, www.investopedia.com/terms/d/descriptive-analytics.asp.

2 Narkhede, Sarang. “Understanding Confusion Matrix.” *Medium*, Towards Data Science, 9 May 2018, towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62.

3 Pologeorgis, Nicolas A. “Employability, the Labor Force, and the Economy.” *Investopedia*, 4 Aug. 2021, www.investopedia.com/articles/economics/12/employability-labor-force-economy.asp.

4 Team, Great Learning. “Generalized Linear Models | What Does It Mean? - Great Learning.” *GreatLearning Blog: Free Resources What Matters to Shape Your Career!*, 27 Apr. 2021,www.mygreatlearning.com/blog/generalized-linear-models/.

5 Thomas, Sarah. “Calculate Outlier Formula: A Step-By-Step Guide | Outlier.” *Articles.outlier.org*, 24 Jan. 2022, articles.outlier.org/calculate-outlier-formula.

6 “Topic No. 409 Capital Gains and Losses | Internal Revenue Service.” *Www.irs.gov*,www.irs.gov/taxtopics/tc409#:~:text=You%20have%20a%20 capital%20gain.

7 Torpey, Elka. “Same Occupation, Different Pay: How Wages Vary : Career Outlook: U.S. Bureau of Labor Statistics.” *Bls.gov*, 20 May 2015, www.bls.gov/careeroutlook/2015/article/wage-differences.htm.

8 Turney, Shaun. “Chi-Square (X^2) Tests | Types, Formula & Examples.” *Scribbr*, 23 May 2022, www.scribbr.com/statistics/chi-square-tests/.

9 “Univariate, Bivariate and Multivariate Data and Its Analysis - GeeksforGeeks.” *GeeksforGeeks*, 14 Aug. 2018, www.geeksforgeeks.org/univariate-bivariate-and-multivariate-data-and-its-analysis/.

10 “What Is Logistic Regression?” *Statistics Solutions*, www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/.