Machine Learning For Data Analytics

Independent Report on Findings

by

Mechak Holondo

Table of Contents

[1 Introduction 3](#_Toc39787313)

[1.1 Business Understanding 3](#_Toc39787314)

[2 Data Understanding, Data Pre-processing, and Exploratory Data Analysis 4](#_Toc39787315)

[2.1 Feature encoding 5](#_Toc39787316)

[2.2 Describing the characteristics 5](#_Toc39787317)

[2.2.1 Scope of the data 5](#_Toc39787318)

[2.3 Data cleansing 6](#_Toc39787319)

[2.3.1 Missing values 6](#_Toc39787320)

[2.3.2 Outlier detection 7](#_Toc39787321)

[2.4 Finding the min/max/mean 8](#_Toc39787322)

[2.5 Data Visualisation 9](#_Toc39787323)

[2.5.1 Histograms 9](#_Toc39787324)

[2.5.2 Count Plots 12](#_Toc39787325)

[2.5.3 Examining Feature-Target Relationships 15](#_Toc39787326)

[3 Machine Learning Methods and their Implementation 20](#_Toc39787327)

[3.1 Predictive modelling in machine learning 20](#_Toc39787328)

[3.1.1 Pre-processing 20](#_Toc39787329)

[3.1.2 Machine Learning 20](#_Toc39787330)

[3.1.3 Evaluation 20](#_Toc39787331)

[3.1.4 Prediction 20](#_Toc39787332)

[3.2 Problem-solving methods for machine learning 21](#_Toc39787333)

[1.1.1 Decision Tree 21](#_Toc39787334)

[1.1.2 K Nearest Neighbour 21](#_Toc39787335)

[1.1.3 Logistic Regression 21](#_Toc39787336)

[1.1.4 Random Forest 22](#_Toc39787337)

[4 Evaluation Machine Learning Models 22](#_Toc39787338)

[5 Discussions 23](#_Toc39787339)

[Bibliography 23](#_Toc39787340)

[Appendix 24](#_Toc39787341)

# Introduction

My task as a data analyst is to take the “Census Income” dataset and gain insight on its characteristics and patterns using Python programming to cluster through and find any patterns in high- and low-income records. Another goal is to predict whether income exceeds $50,000 per annum (p/a) using classification through the same Python programming means. As the data set is quite large it will be required to be mined and therefore the CRISP-DM methodology will be useful here as it will help provide a structured approach when planning a data mining project (Simmons, 2014). The CRISP-DM will be applied to my problem in the business understanding to begin with to properly outline the data mining goal before proceeding to the data understanding and other stages of the methodology, which will be addressed later in this report.

The cross-industry process for data mining (more commonly known as CRISP-DM) methodology is described as a “hierarchical process model” split into four levels of abstraction ranging from general to specific levels respectively: *phases*, *generic tasks*, *specialised tasks****,*** and *process instances*.

At the **first level** the process is split into phases, with each phase containing several generic tasks as a **second level** which is designed to cover all possible situations when mining data.

In the **third level** you see the specialised tasks where how actions in generic tasks are described to show how they should be carried out (i.e. the actions that are specific for a problem and its chosen data mining tool).

The **fourth level** is the final level, the process instance level, in which we record any of the actions, decisions, and results found during the real data mining process. Instead of describing what generally happens this level documents what *actually* happened in the specific data mining experience being modelled. (Wirth and Hipp, 2000).

## Business Understanding

To understand the business is to determine the project objectives and requirements from a business perspective to first establish the aim of this data mining engagement. As the specification states, there are three primary goals:

* “To understand the characteristics and patterns of the data set;”
* “to understand the patterns in both high-income (>$50K) and low-income (<$50k) records;”
* “and to predict whether income exceeds $50K per year based on census data.” (He 2020)”

In this coursework I am applying the CRISP-DM methodology following the six steps outlined by Pro Global Business Solutions in their article. First is understanding the business: census income data. The current situation is that the dataset contains a vast number of incomes with no real order or categorisation beyond high- and low-income. Therefore, the data must be clustered and classified to make sense of it and gain any insight to identify patterns.

Next is understanding the data which is partially done as figures have been collected from the sources and compiled in the UC Irvine dataset that is the subject of this assignment. All that is left to do for this stage is to verify the quality of the information by questioning the completeness and accuracy of the material. This will be explored in much more detail in Section 2: Data Understanding, Data Pre-processing, and Exploratory Data Analysis

Plotting

Once that has been verified, I will move onto the preparation of the data where the selection, cleansing, constructing, and formatting of the data will be done before the modelling can take place on a complete dataset.

The modelling is the first step after these stages of preparation are complete as it will generate test scenarios for validating the model quality and a couple of models will be created to assess against the goals set out in the business understanding stage.

The penultimate stage will evaluate the data using the results of the models in which new information may be found in the patterns discovered during modelling, but this step will include referring back to the models iteratively until satisfied enough to move onto the final step.

The final step will be in the form of this report’s presentation as my findings will be set out in a presentation that is accessible for both specialists and non-specialists.

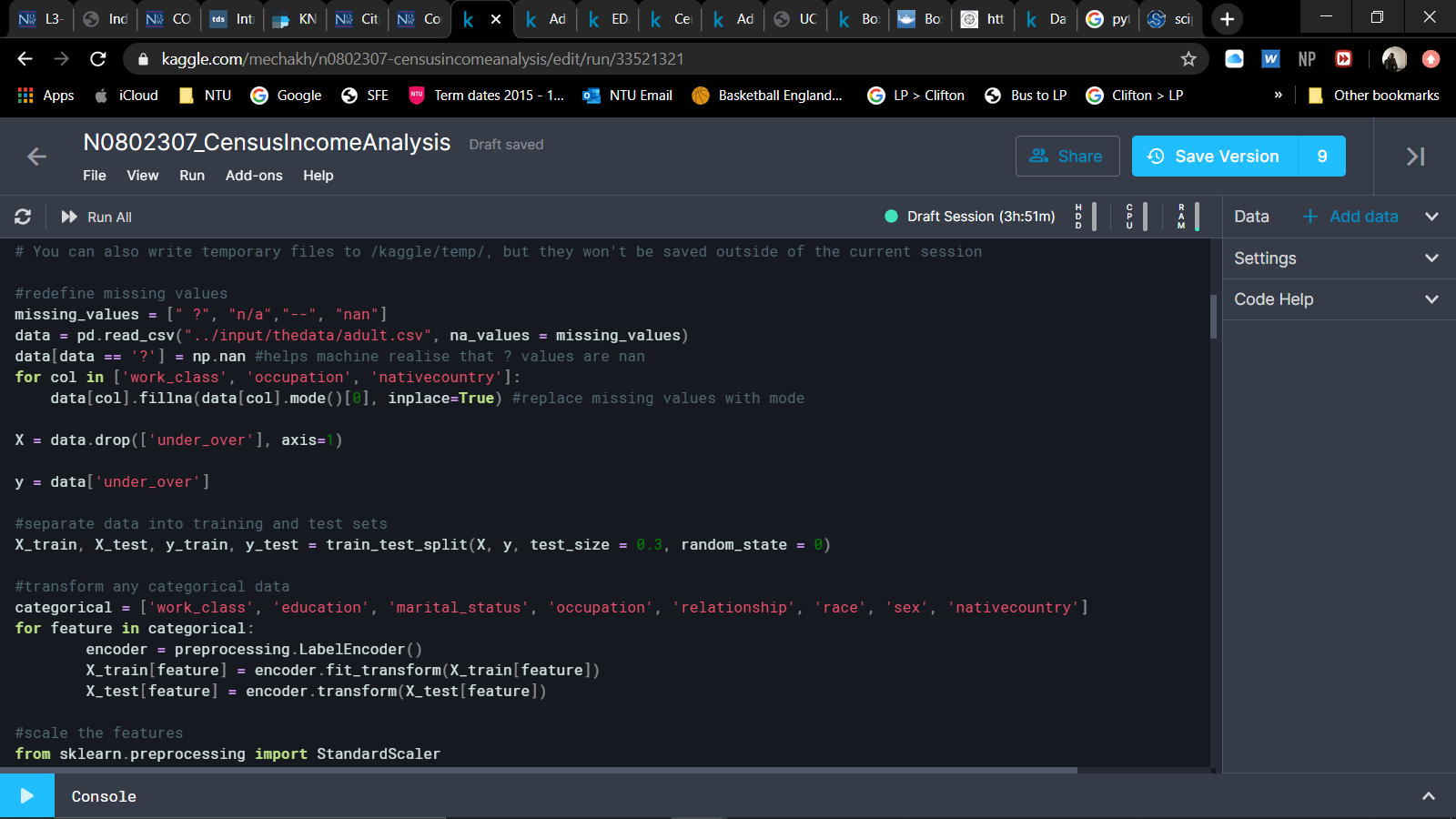
# Data Understanding, Data Pre-processing, and Exploratory Data Analysis

The target for this model is the “Census Income” adult dataset available for download from UC Irvine Machine Learning (1996), its initial purpose was to “predict whether income exceeds $50k/yr based on census data” and was submitted by donors Ronny Kohavi and Barry Becker in Data Mining and Visualization at Silicon Graphics.

The data represents the Census Income data from 1994, taken from the Census database to generate a reasonably clean extract of the records for use in data analysis.

## Feature encoding

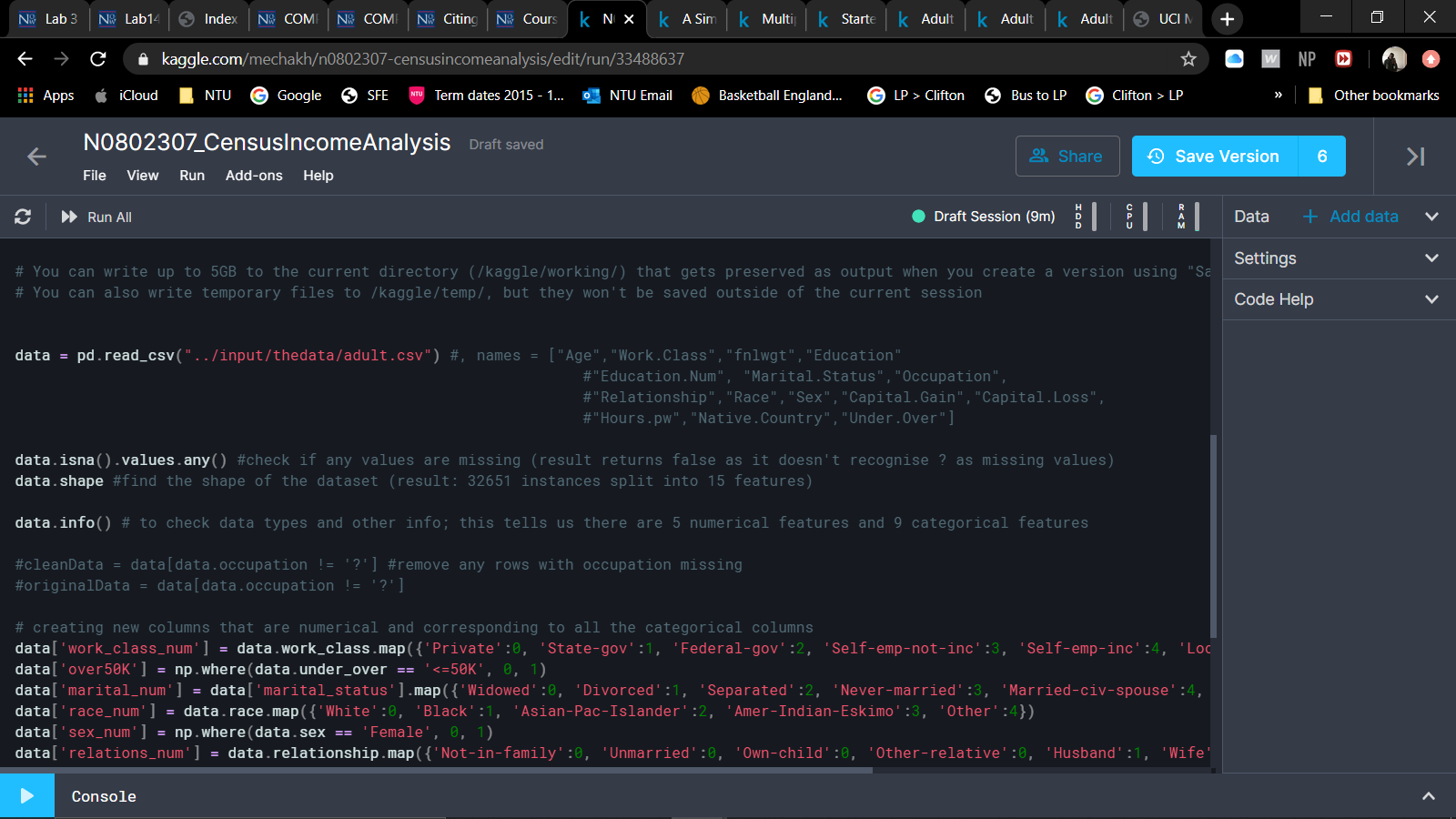
To begin with I resolved the missing values (more information on that later in the report) and split the data into its training and test set.



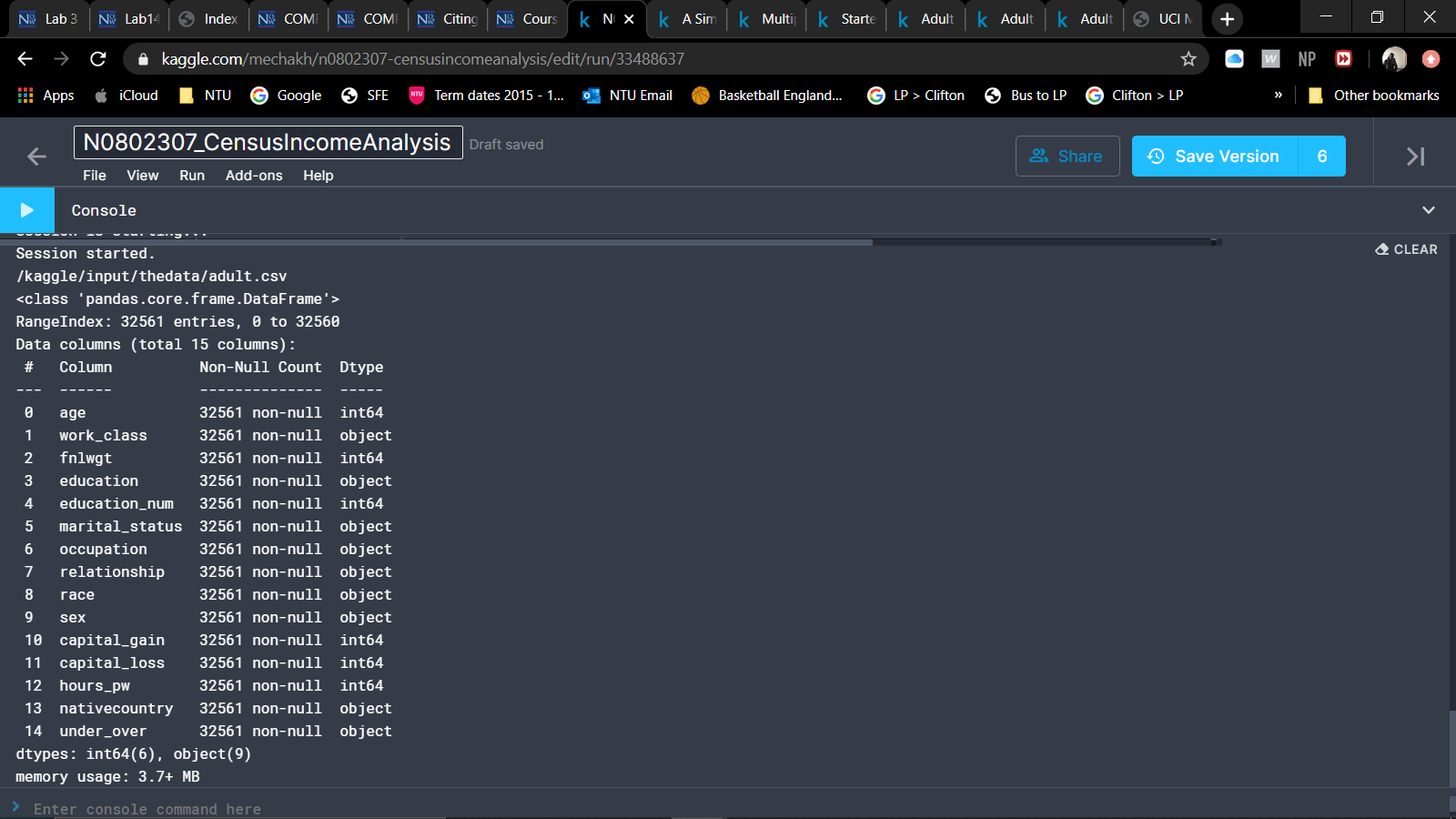
## Describing the characteristics

### Scope of the data

I determined the scope of the data by first finding its size and how many instances and features it held. This was done using small lines of Python code:



The resulting output:



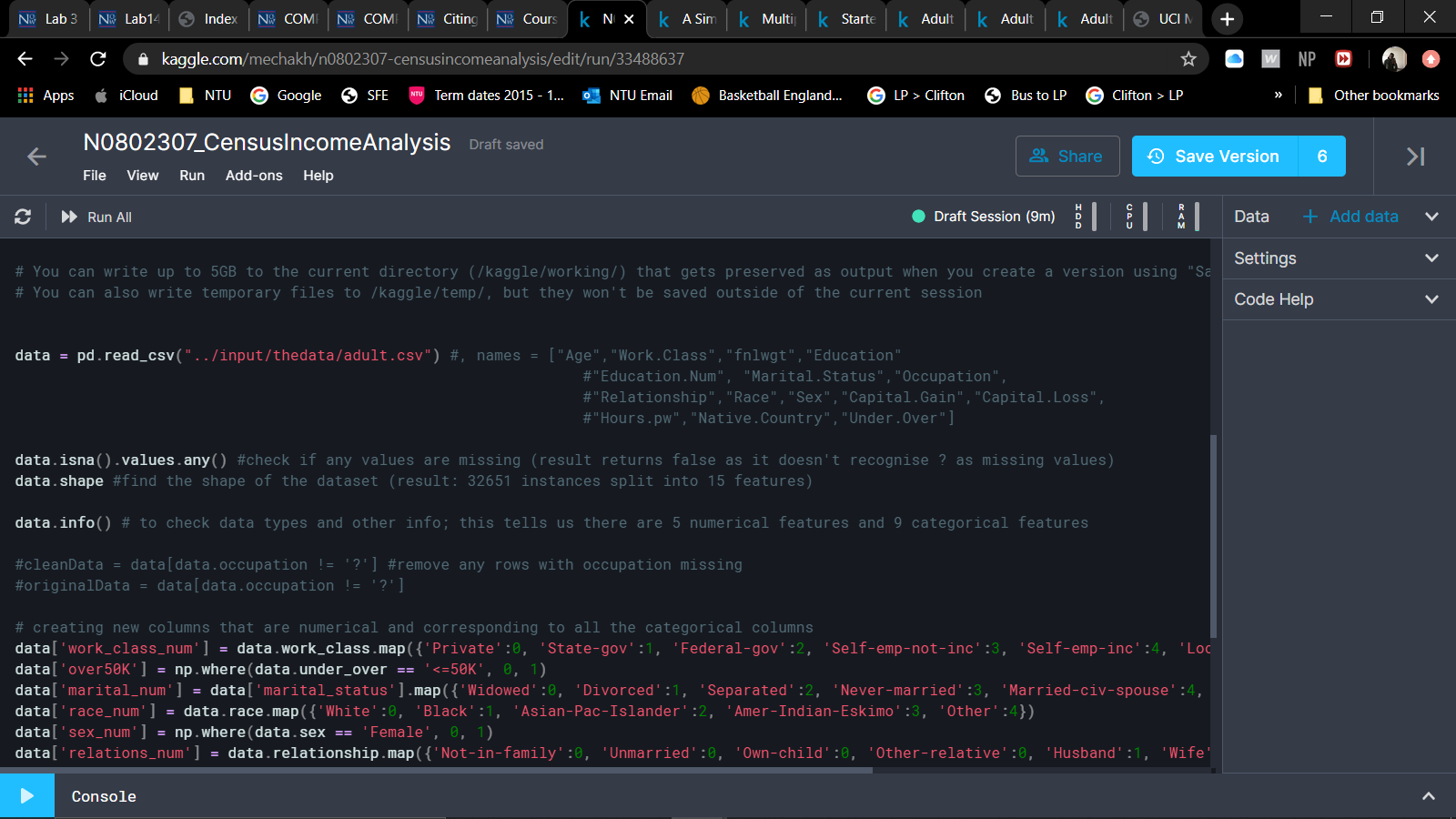
This tells us that the dataset consists of 15 features or ‘columns’, which are indexed so they number from 0 to 14.

It also shows that the data contains 32,561 instances of which six (6) columns of them are numerically represented data and nine (9) columns of them are categorically represented data.

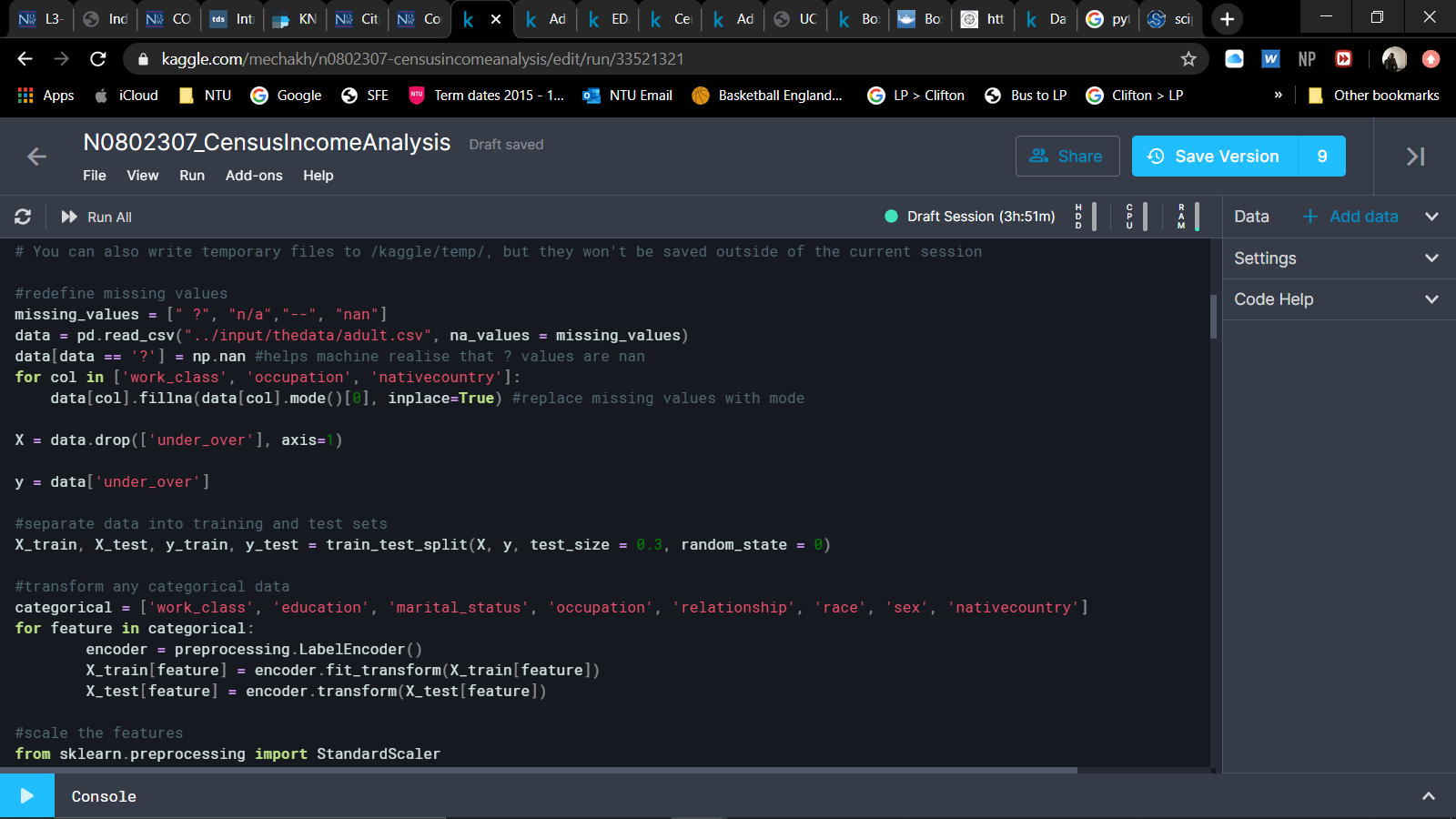
## Data cleansing

### Missing values

Before I set to describe the data after determining its size, I checked it for any missing values or outliers.



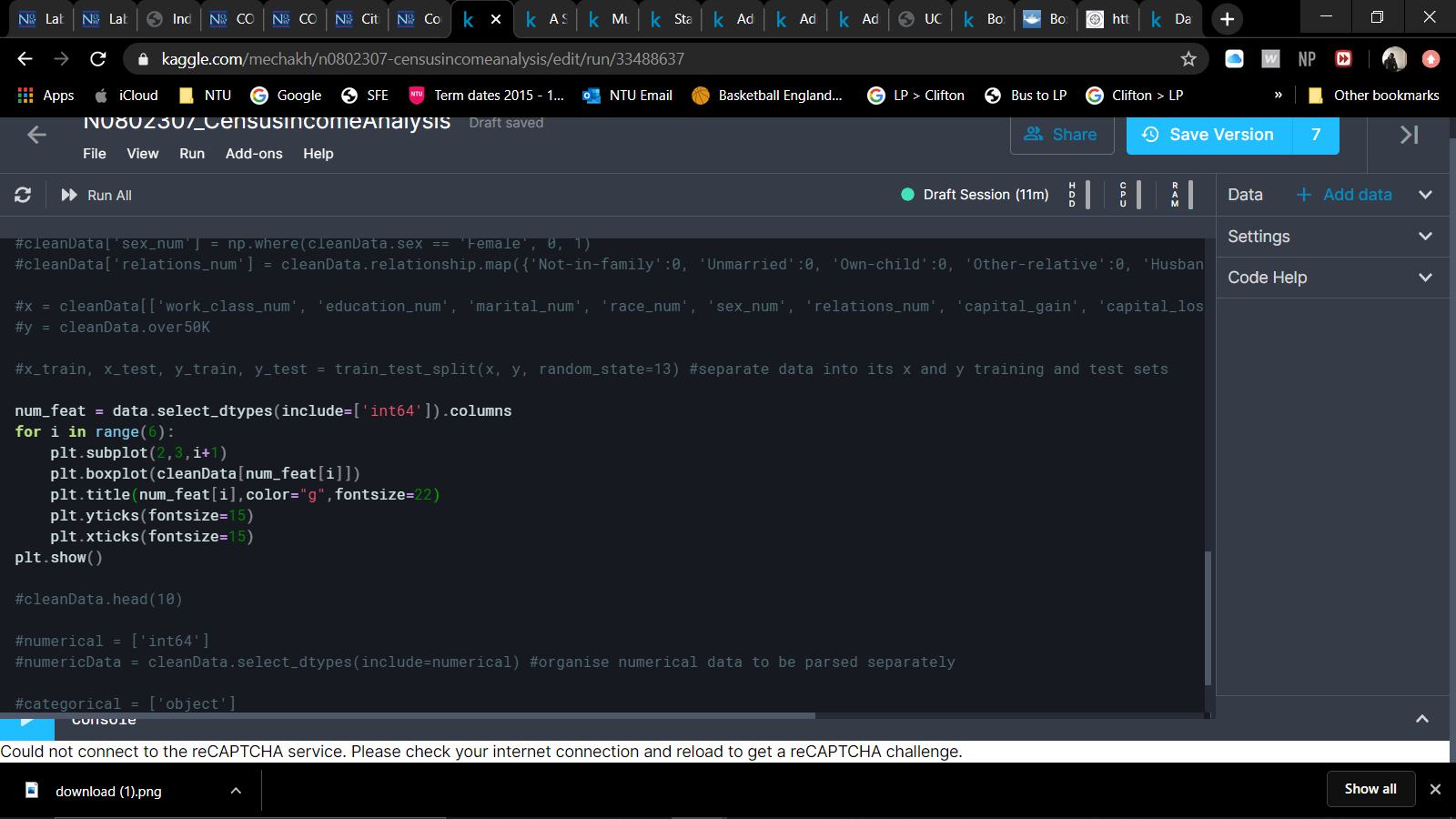
This code returned false as a result as it did not recognise the “?” values in the occupation column as being missing values. I then chose to replace the rows with the most common instance for the column to clean the data in the hopes of providing more accurate results later.



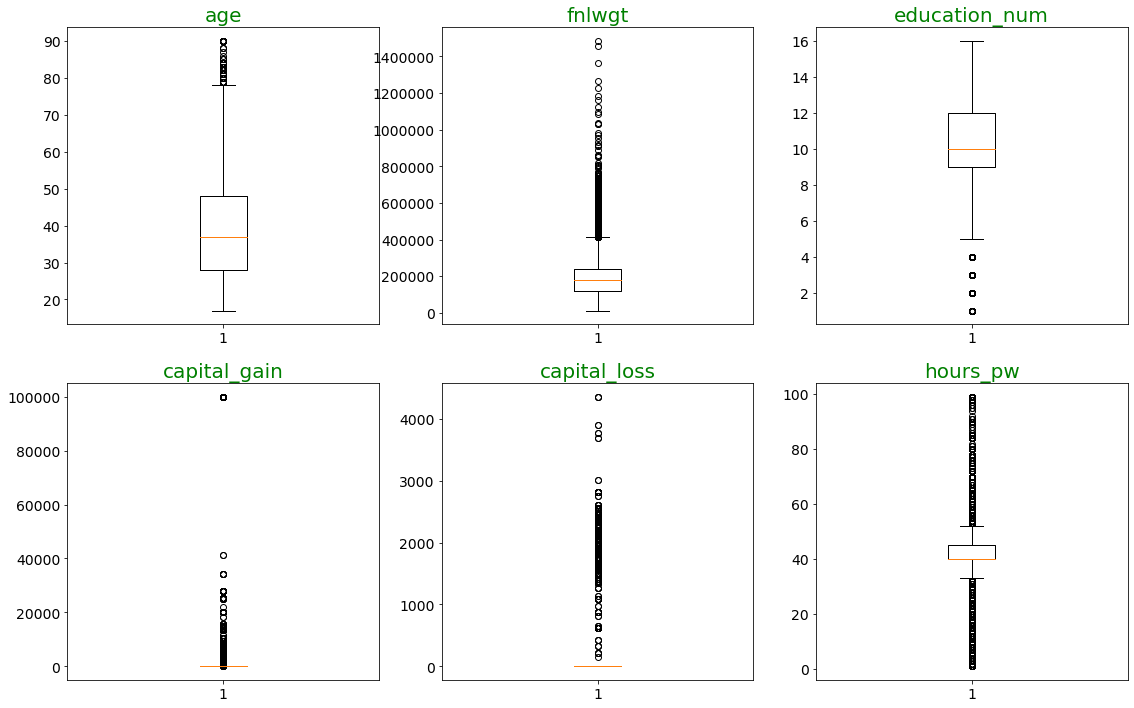
The above code is to find instances in which there are missing values and to replace them with its respective column’s most common instance.

### Outlier detection

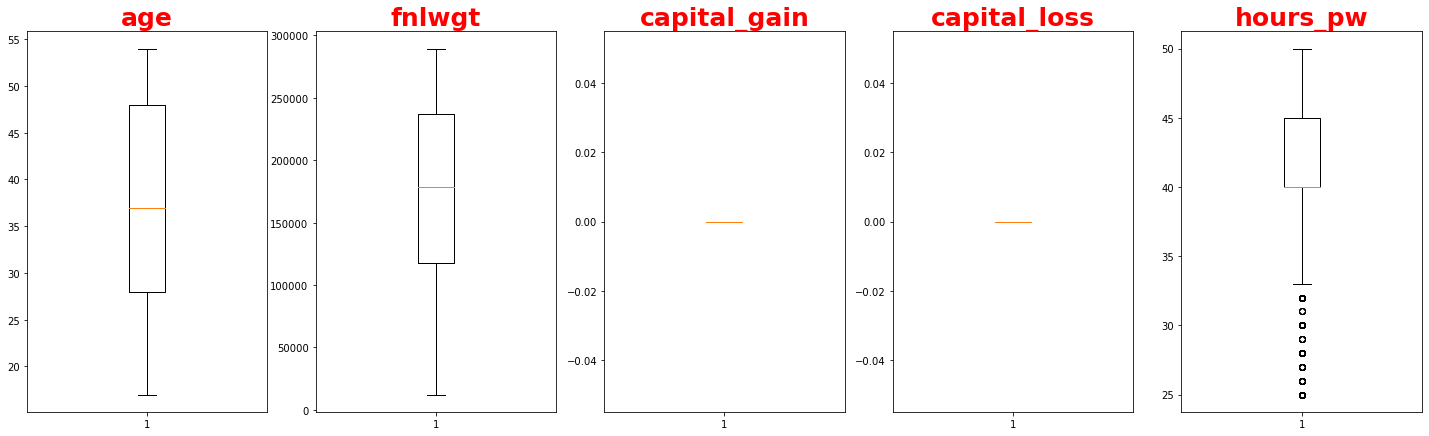
Outliers are defined as being “a data point that lies outside the overall pattern in a distribution”. If left in the data, it may skew the results of the data mining process, so I am going to check to make sure there are no outliers.



This code uses the mathplotlib library to plot box plots using the min/max and quartiles of the dataset. Anything that is outside of the minimum and maximum values is an outlier.

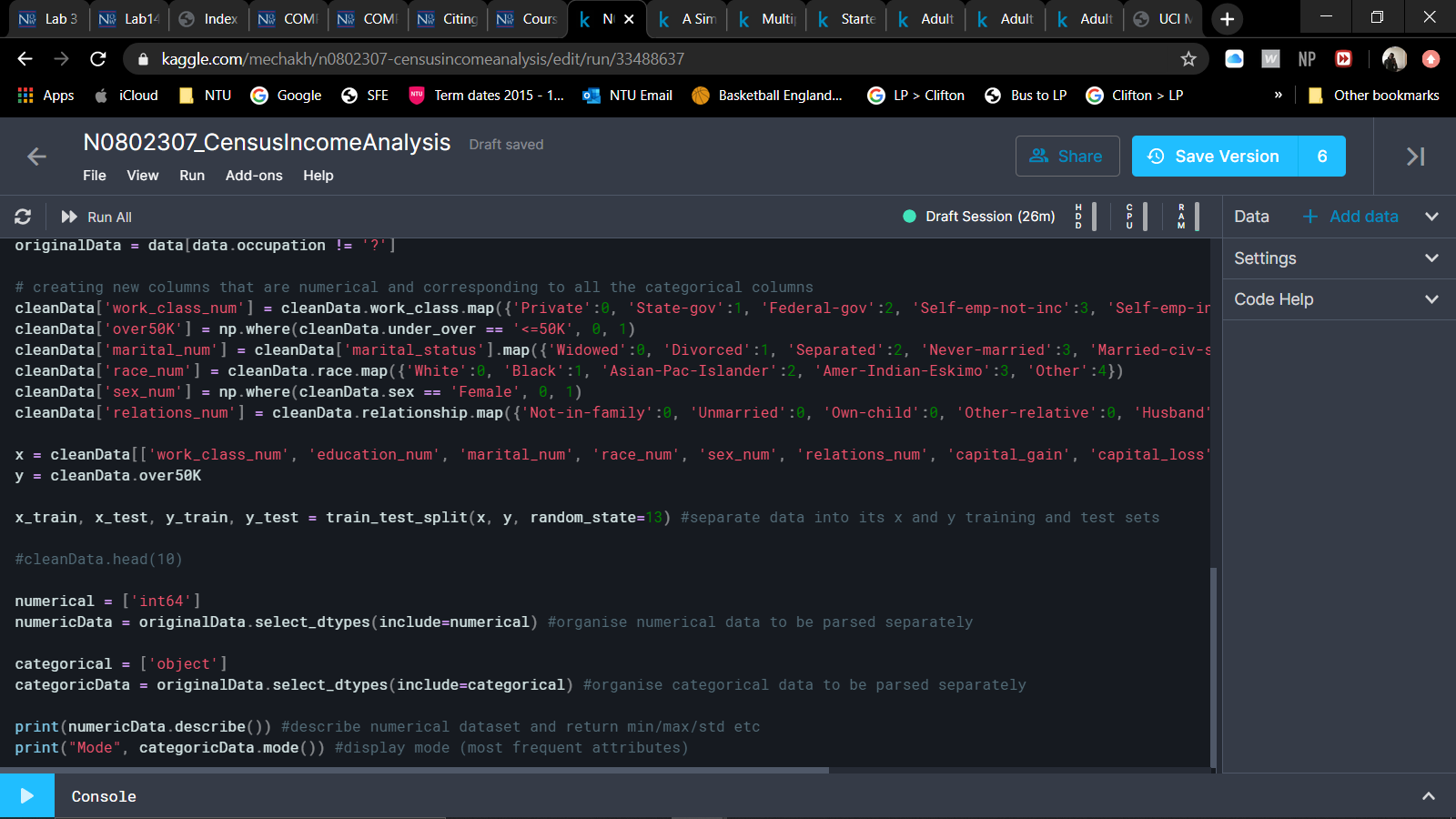


These outliers need to be removed so to do this I researched different methods and found one called ‘winsorize’ using the scipy library:

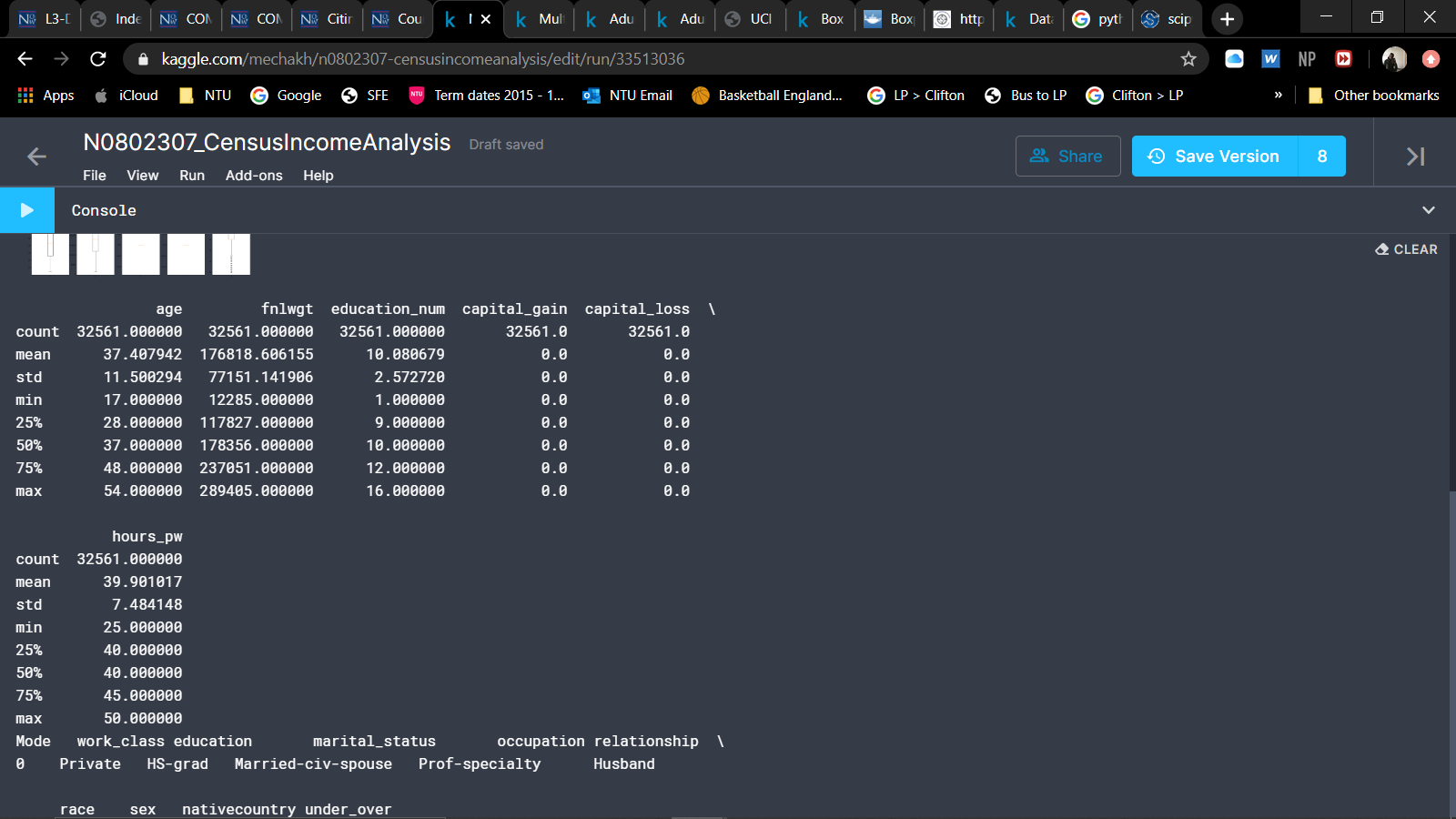


## Finding the min/max/mean

Using this data, I can find the minimum, average, and maximum values along with the standard deviation from the training and test sets:



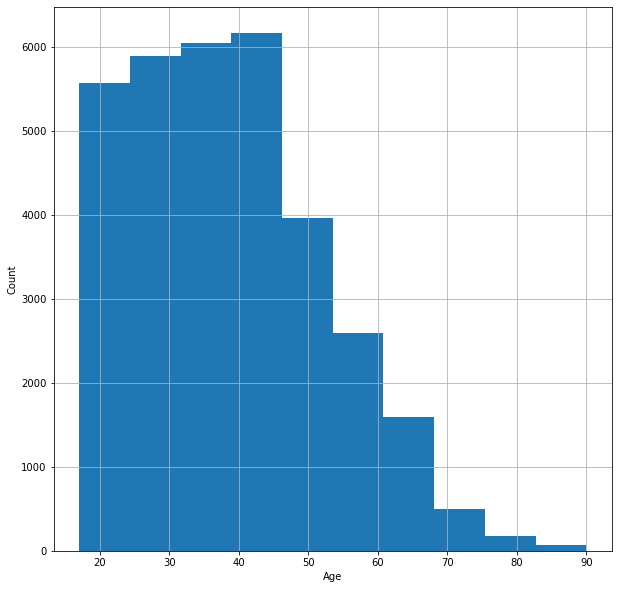
For this I used the modified dataset to ensure that the results were as accurate as possible, excluding any outliers/erroneous values.



## Data Visualisation

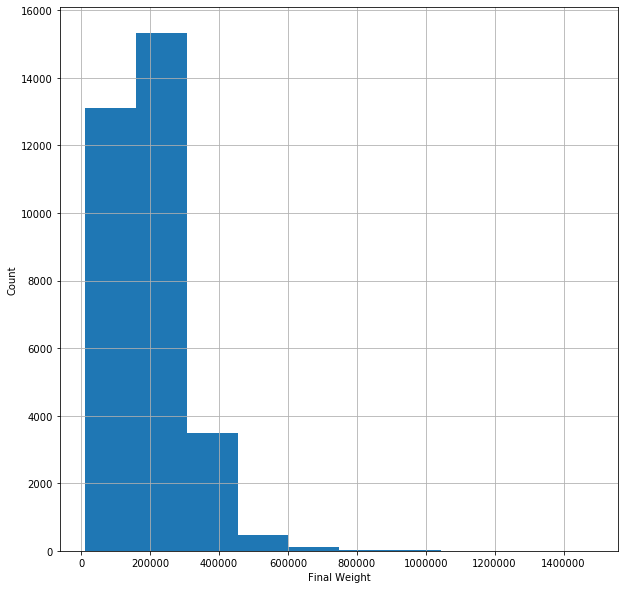
### Histograms

Age



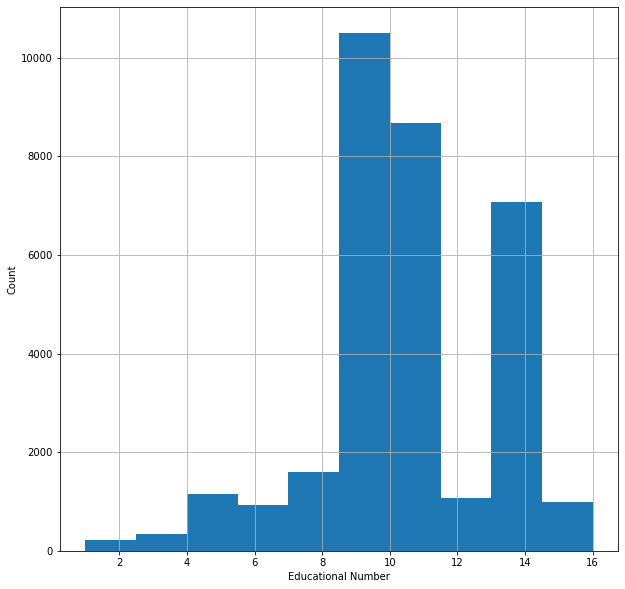
The takeaway from this histogram is that the most common ages ranged between 40-50 years old and the number of younger people outnumbered the older people who took part in the census.

Final Weight



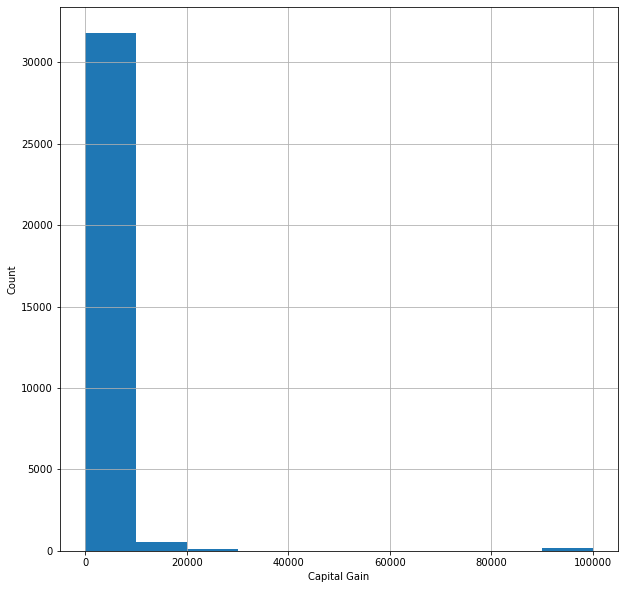
The mean outweighed the medium as seen in the graph above as the sides are unbalanced with the majority being on the left-hand side.

Educational Number

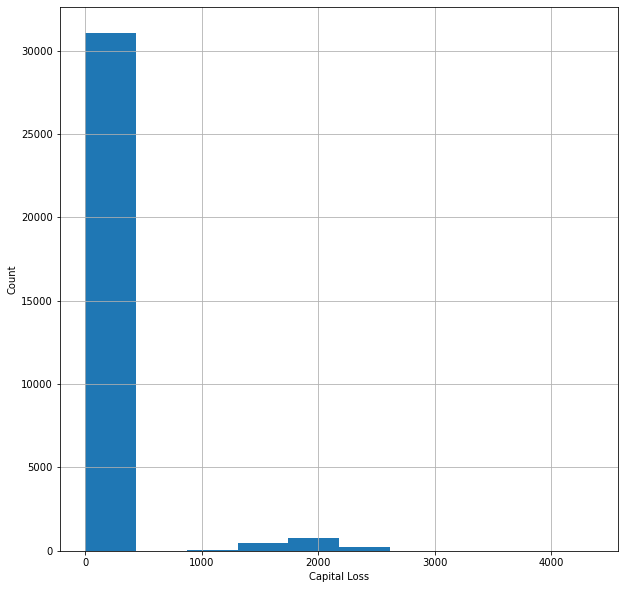


This histogram demonstrated that the educational numbers that were most frequent were between 8 and 10 whilst the least frequent was 2.

Capital Gain

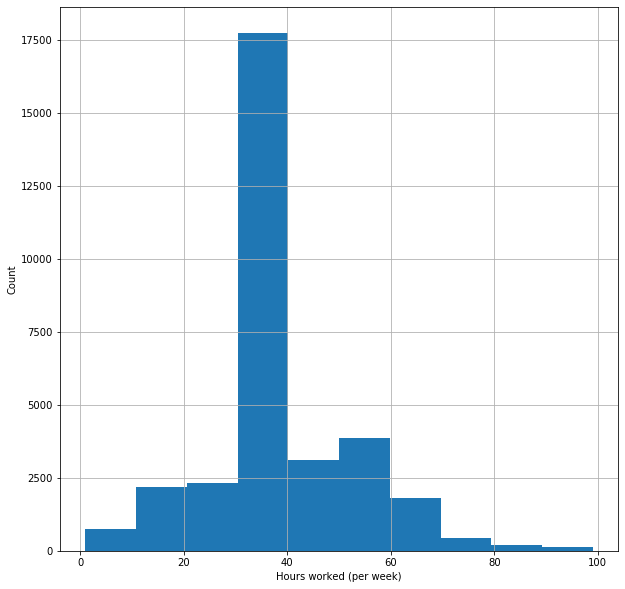


Capital Loss



The capital gain histogram showed that the majority of people’s capital gain was 0, which was already present in the dataset when sampled but this helped confirm.

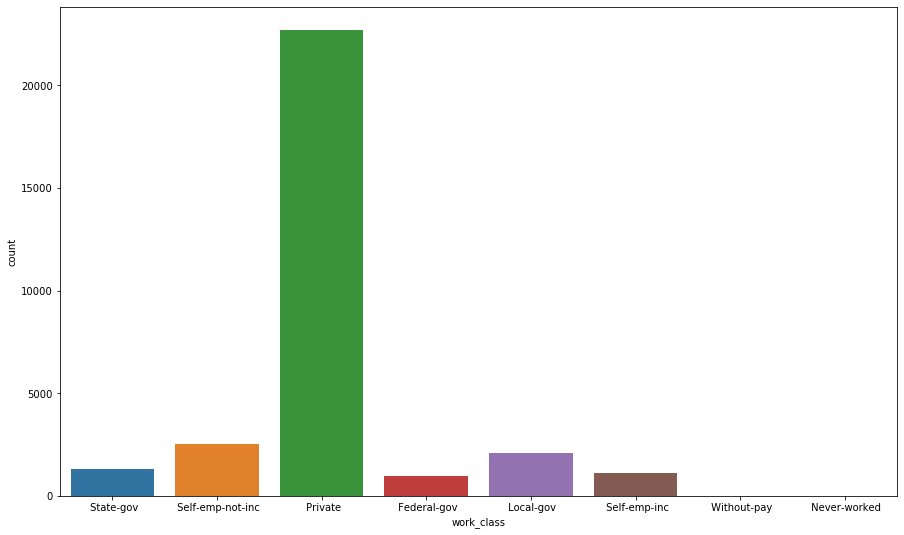
Hours worked (per week)



Similarly, the capital loss histogram showed that the majority of people’s capital gain was 0, which was already present in the dataset when sampled but this helped confirm.

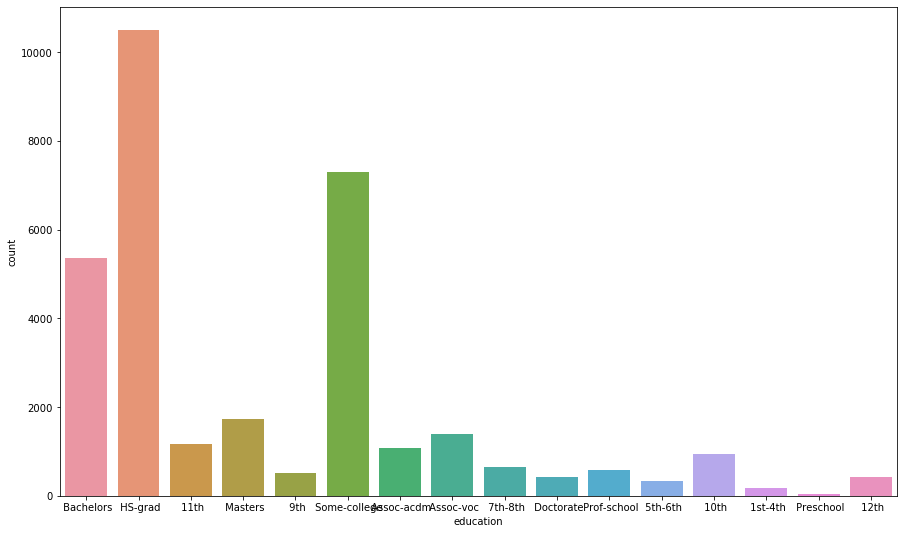
### Count Plots

Work Class



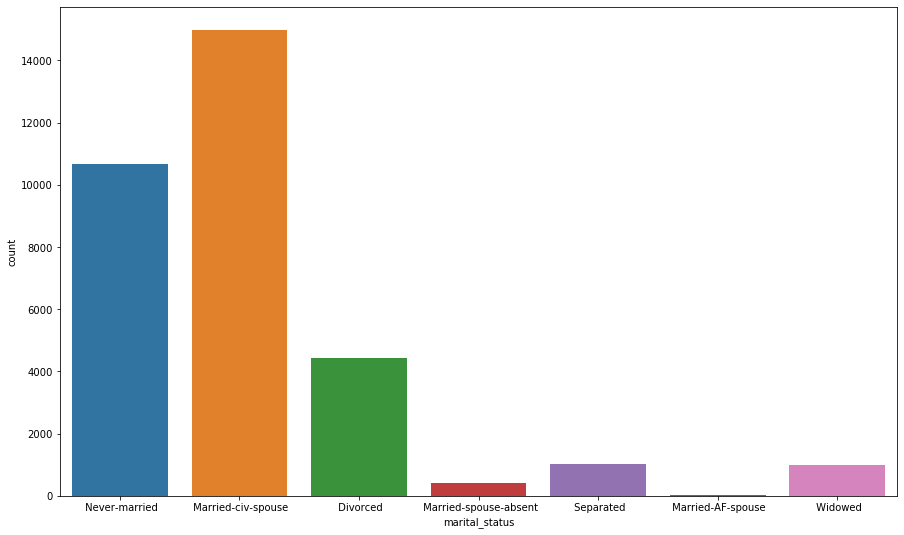
The work class count plot illustrates that 75% of the census takers worked in private jobs out of the eight (8) categories.

Education



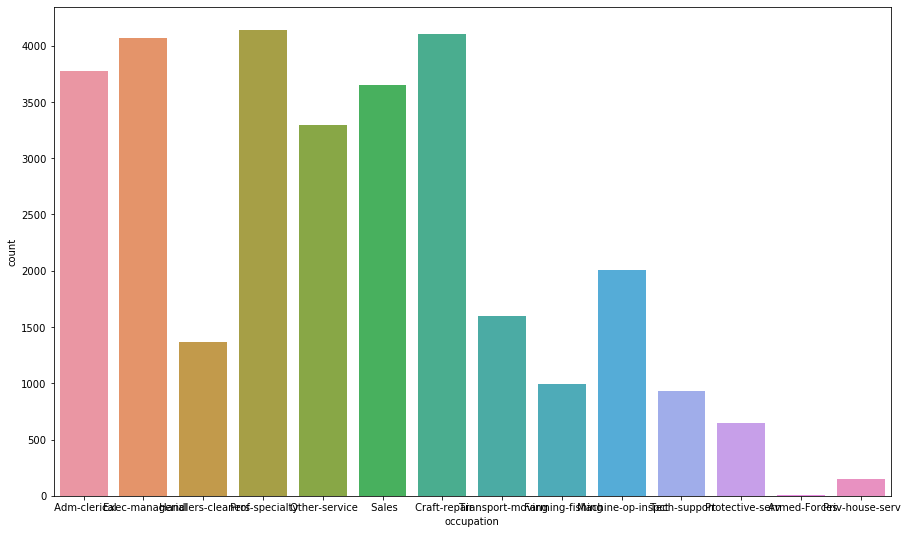
Out of the sixteen (16) educational categories: roughly 32% of these people were high school graduates; roughly 22% were college educated and 16% had a bachelors. Pre-school education was the lowest percentage category with only 0.16% which is to be expected.

Marital Status



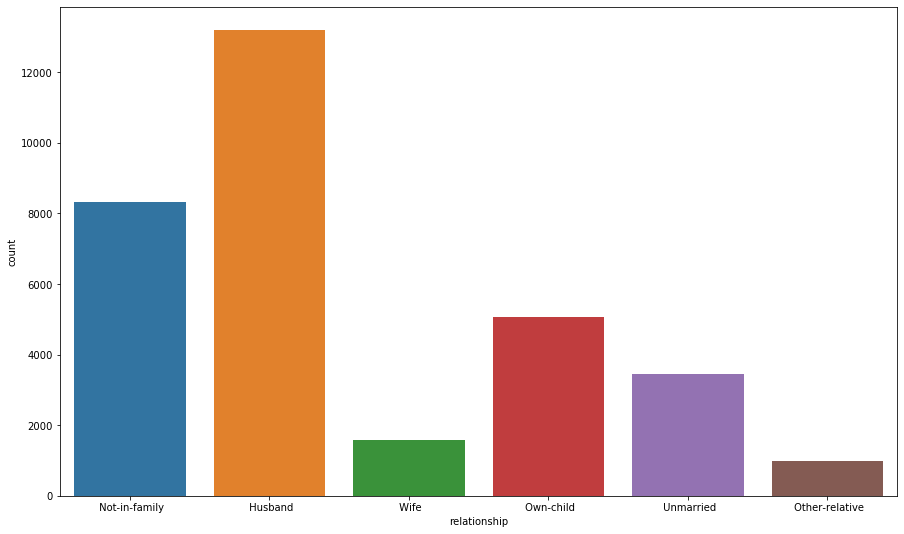
Most of the people who took part were married with a civil spouse, which is to be expected with when the census took place and the general ideals at the time as well as the ages of the census takers.

Occupation



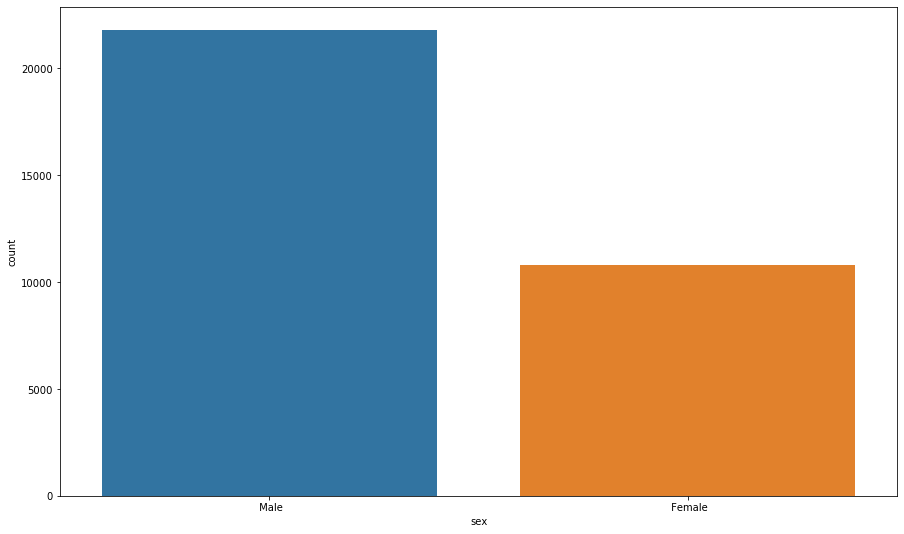
Across the 14 categories, prof-specialty had the highest percentage with 18%, followed by Adm-clerical, Exec-manager, Other-Service, Sales and Craft-Repair with rough percentages of 12%, 13%, 10%, 11% and 13% respectively.

Relationship



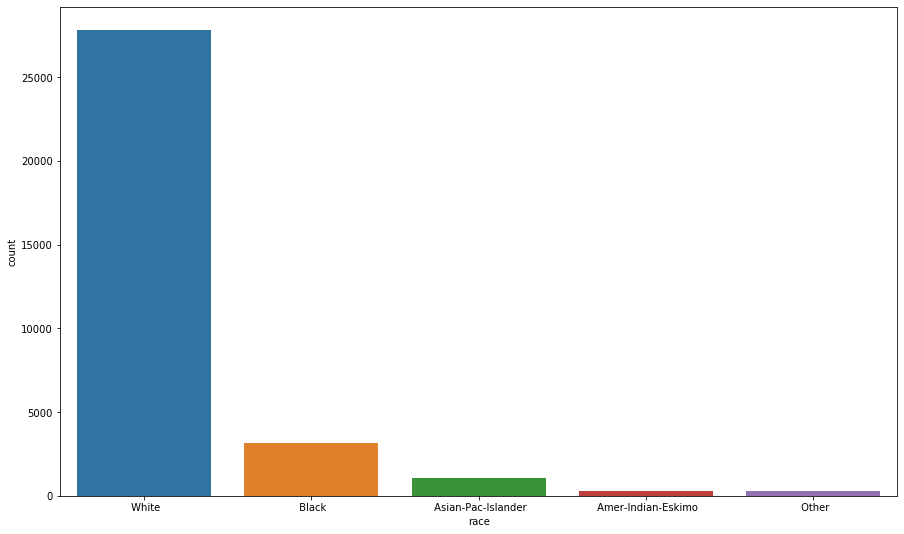
From the six (6) categories almost half of the census takers were husbands, which is not surprising when seen against the amount of men and their ages who partook in the census.

Sex



Most of the people who took part in the census were men.

Race



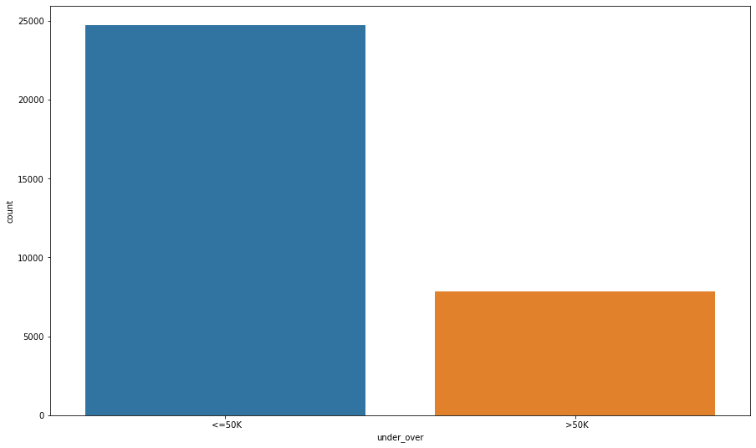
The majority of census takers where white Americans with them taking up 85% of the data.

Native Country



The native country was again mostly Americans, which is to be expected as the census data is from an America census.

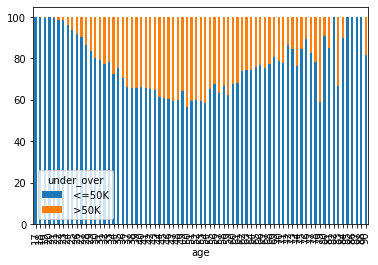
Under/Over



The number of people earning below $50,000 was almost double those earning above that threshold.

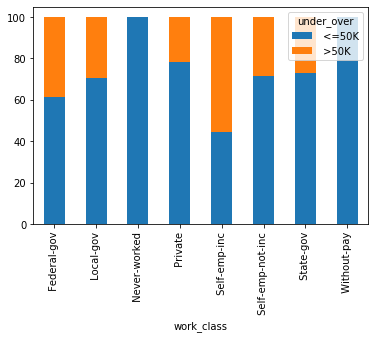
### Examining Feature-Target Relationships

Age and Under/Over

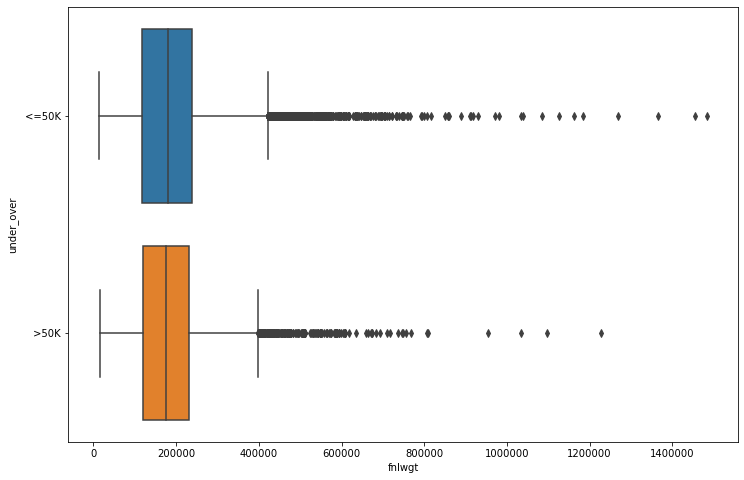


Work Class and Under/Over

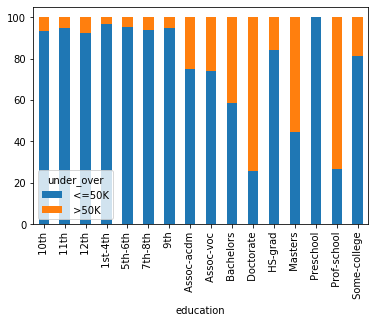
\*Under/Over = whether they earn under or over $50k



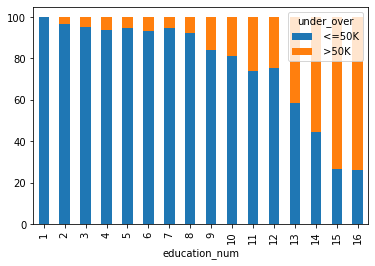
Final Weight and Under/Over



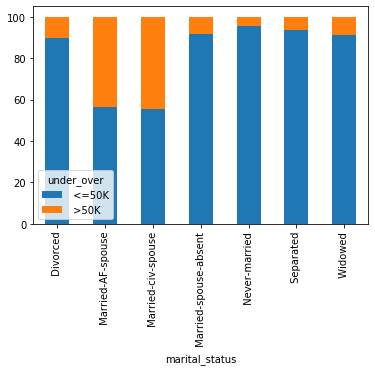
Education and Under/Over



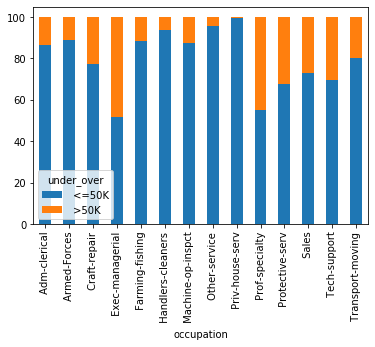
Educational Number and Under/Over



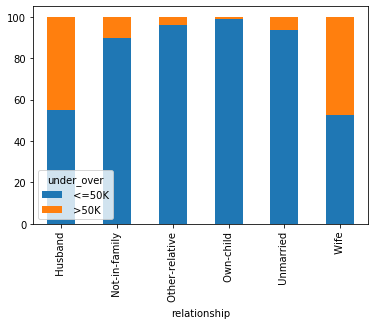
Marital Status and Under/Over



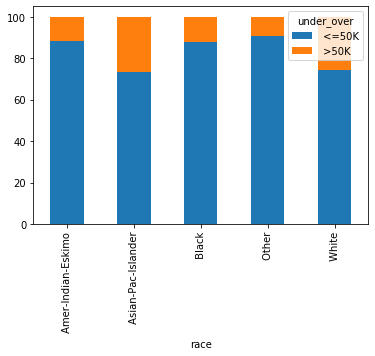
Occupation and Under/Over



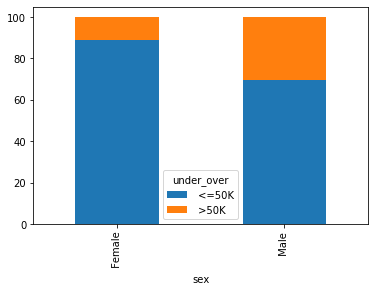
Relationship Status and Under/Over



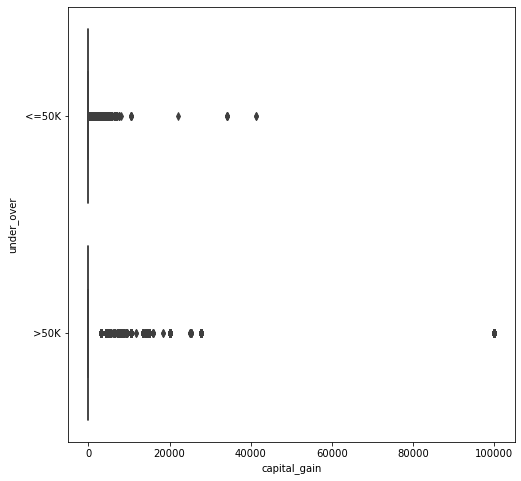
Race and Under/Over



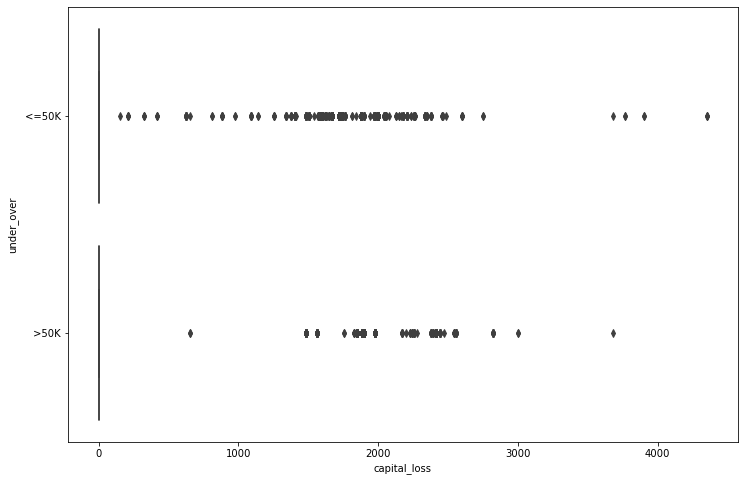
Sex and Under/Over



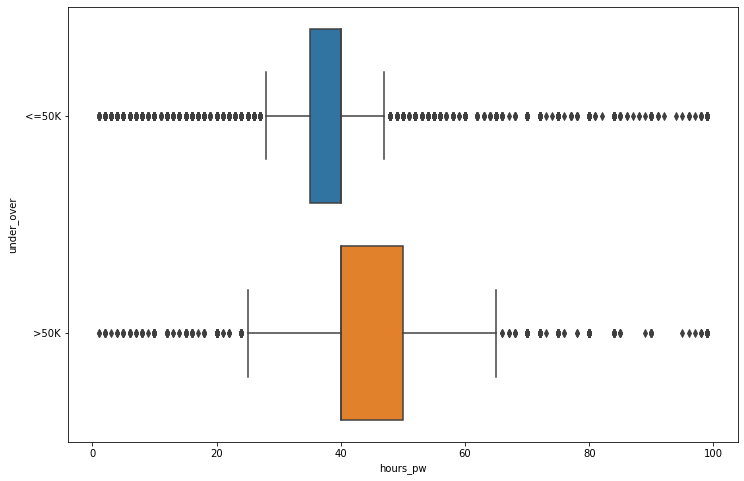
Capital Gain and Under/Over



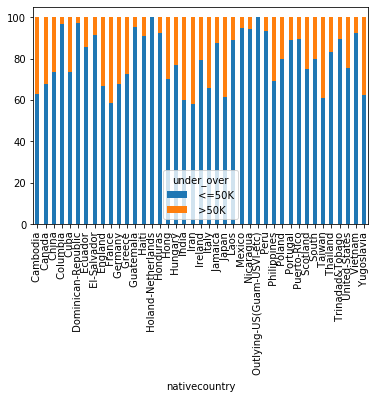
Capital Loss and Under/Over



Hours (per week) and Under/Over



Native Country and Under/Over



# Machine Learning Methods and their Implementation

## Predictive modelling in machine learning

### Pre-processing

Pre-processing handles the raw data from various sources therefore at this stage the modelling cannot proceed as the data is incompatible, which is why the analyst needs to clean it to turn real world data into a compatible format.

### Machine Learning

This newly compatibly data is then used against cross validation where several machine learning models are evaluated to determine an appropriate model to use for the dataset. Following this, grid searching is used to optimise the hyperparameter.

### Evaluation

The evaluation phase is where the several machine learning models I mentioned earlier are cut down until one optimal model remains at which point the process attempts to predict its future performance.

### Prediction

Prediction is the last step where the model can then use the data and its chosen model to make the prediction.

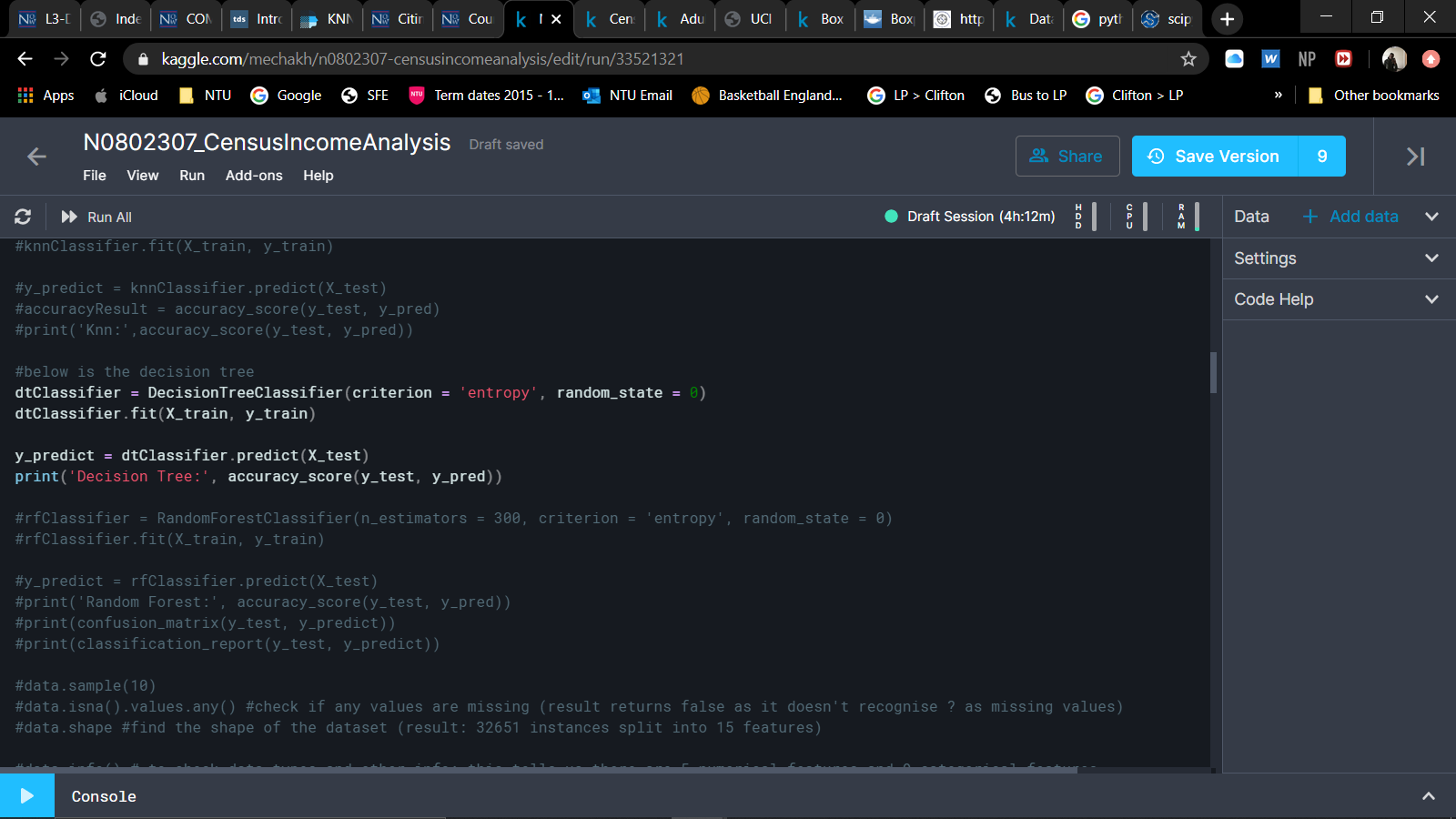
## Problem-solving methods for machine learning

The chosen algorithms for solving the problems are:

* Decision Tree
* K Nearest Neighbour
* Logistic Regression
* Random Forest

### Decision Tree

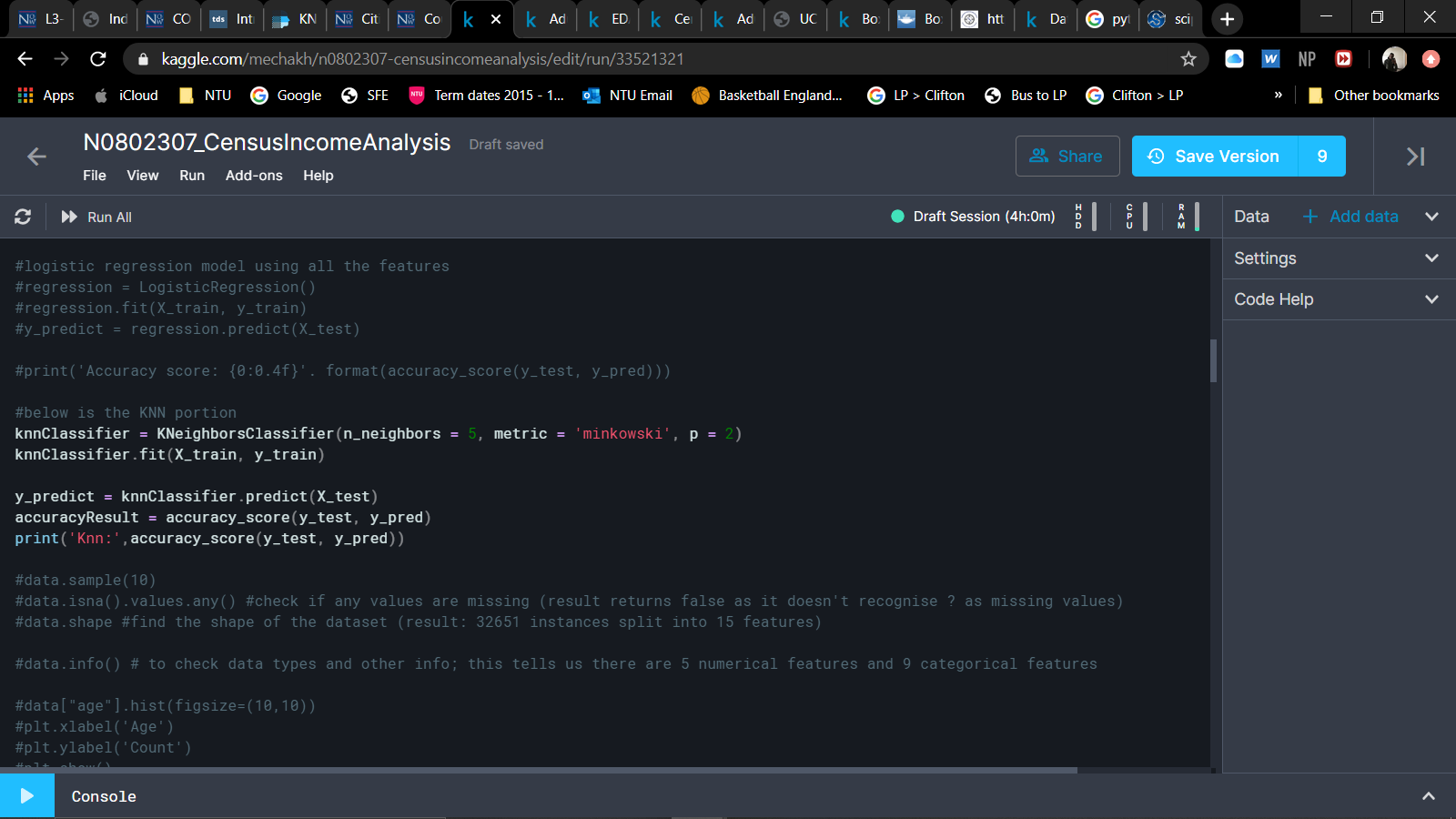
When using a Decision Tree, the goal is to create a training model to predict class or value of a target value by learning simple decision rules from training data. (Chauhan 2019)



The result of the Decision Tree was 0.8240352134302386

### K Nearest Neighbour

The K Nearest Neighbour algorithm uses available cases to classify new cases based on a similarity function (Sayad S).



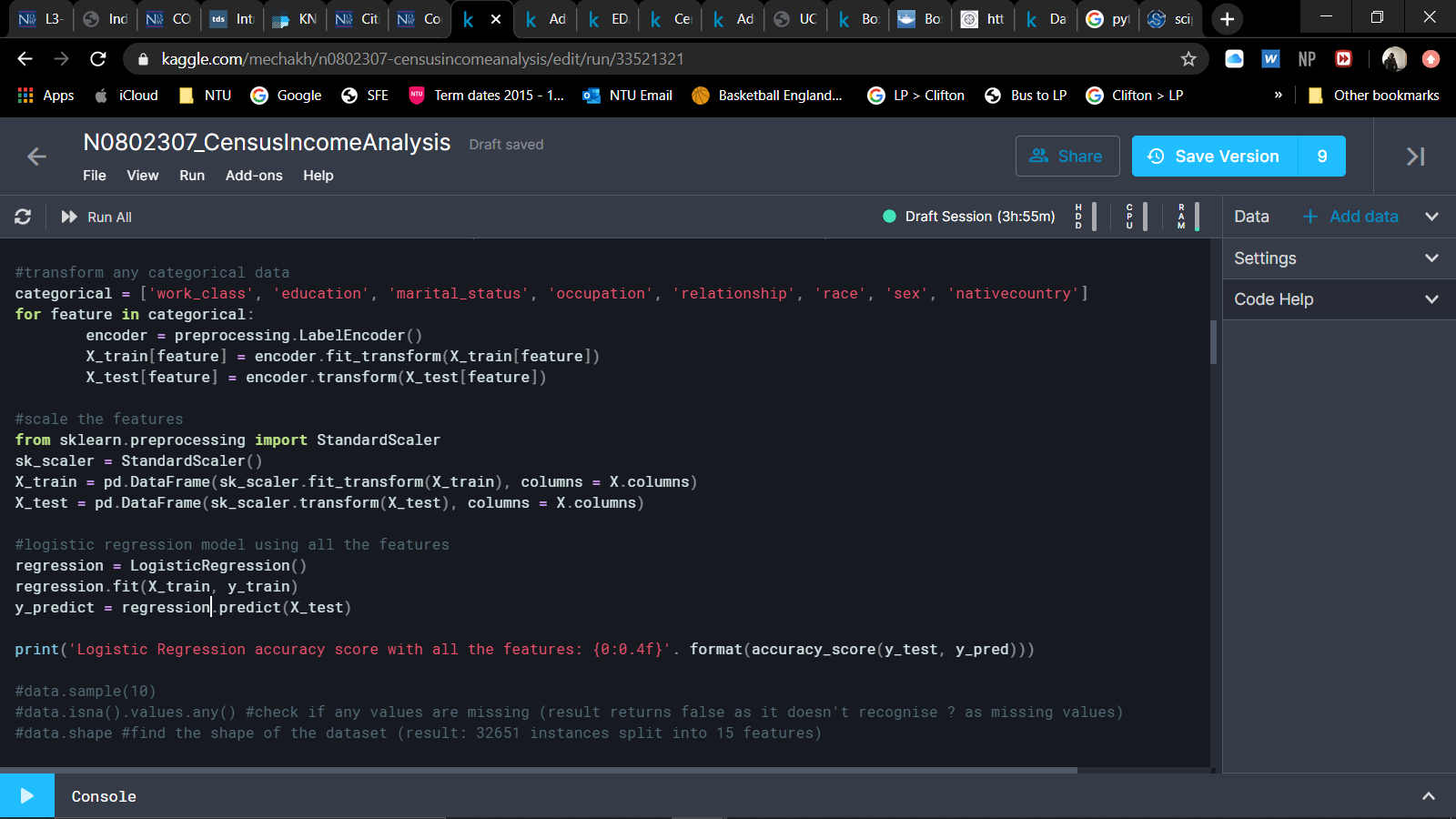
The result of the KNN algorithm returned as 0.8240352134302386

### Logistic Regression

Logistic regression is used to assign observations to a discrete set of classes, it is commonly used by emailing services to separate relevant emails from junk and phishing scams. (Pant 2019)

Logistic regression has another model known as Linear Regression which differs as when done linearly, variables are assumed to be in a relationship whereas logistic regression assumes a nonlinear relationship between the predictor and response.

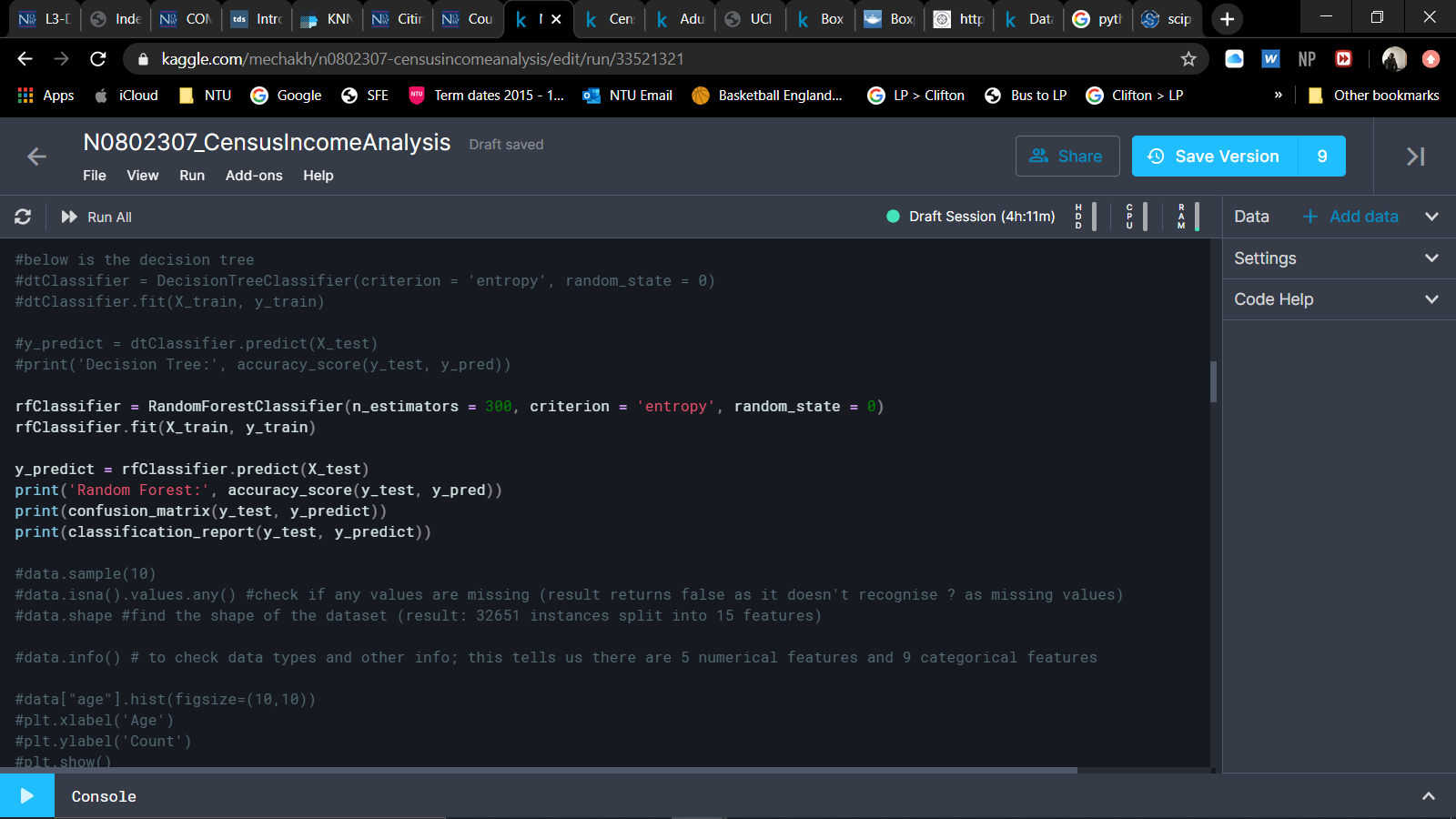
To prepare for the logistic regression I transformed the categorical data and scaled the features so that I could run a logistic regression and determine its accuracy.



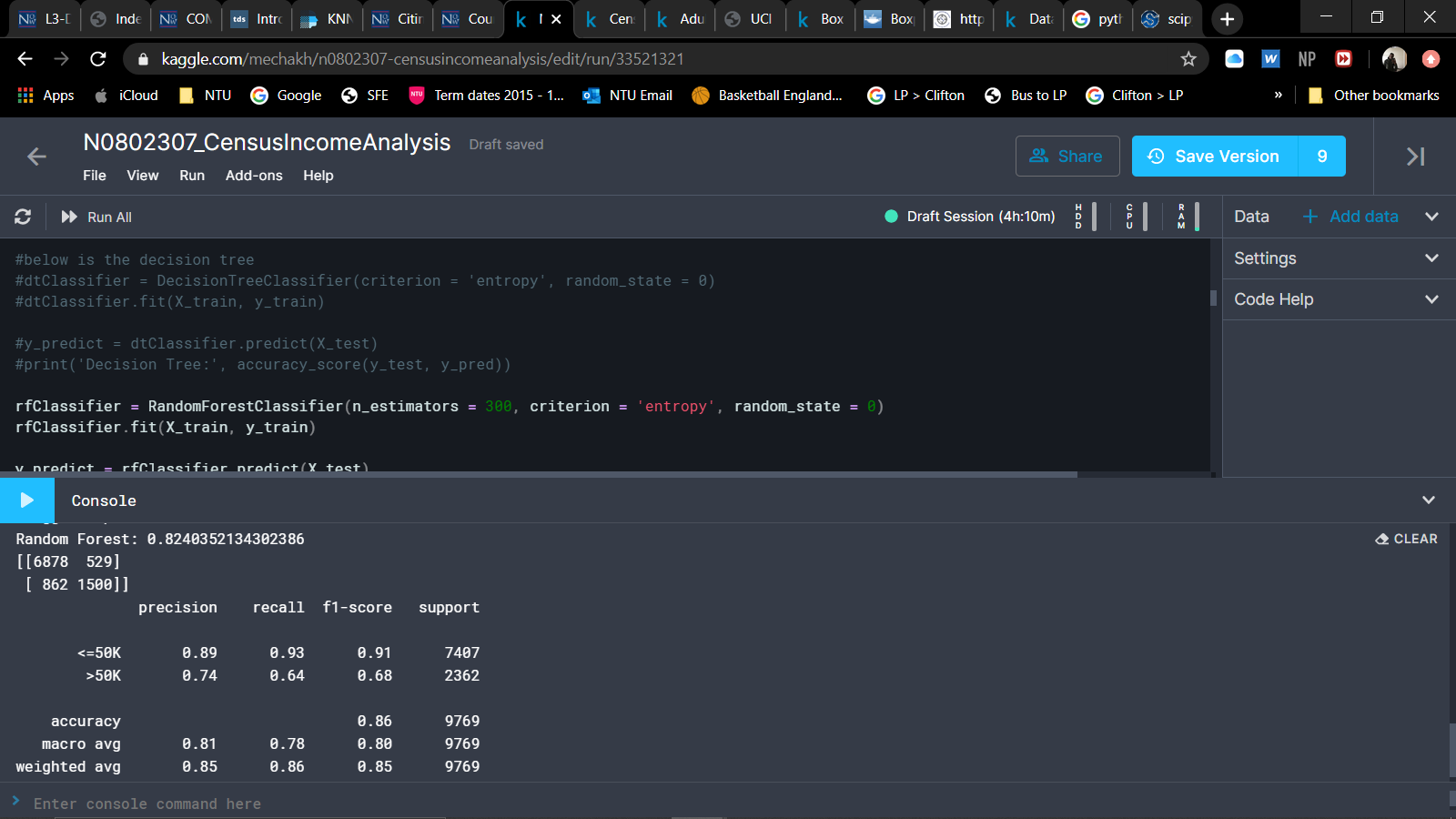
Using all the features it also returned a logistic regression of 0.8240352134302386

### Random Forest

As the name implies, the Random Forest algorithm uses many decision trees working together as an ensemble with each tree producing class predictions and the class with the most votes are used as the final model prediction (Yiu 2019).



The result of the Random Forest:



# Evaluation Machine Learning Models

To evaluate the models, I plotted a graph to compare the accuracy scores and determine which would be optimal for use in my data processing.



# Discussions

Through my work on this assignment in the module I learned to understand the importance of data integrity and the large role that machine learning plays in systems we use daily. I was also able to gain new knowledge on cleansing and normalising data; classification; data visualisation and much more.

I have grown more confident in my ability in regards to machine learning and I believe that to the best of my abilities at this current level I have demonstrated my knowledge so far in the report and achieved what I believed to be the best parameter in the Random Forest Classifier which returned an accuracy score of 0.8240.

# Bibliography

He, J. 2020. Comp20121 Coursework Assessment Specification (UG) [online] Nottingham Trent University. Available at: <<https://now.ntu.ac.uk/d2l/le/content/631075/viewContent/3414283/View> > [Accessed 07 April 2020]

Simmons, B., 2014. *What is the CRISP-DM methodology?* [online] Smart Vision Europe. Available at: <<https://www.sv-europe.com/crisp-dm-methodology/>> [Accessed 06 April 2020]

UCI Machine Learning Repository, 1996. *Adult Data Set* [online] Available at: <<https://archive.ics.uci.edu/ml/datasets/adult>> [Accessed 21 Jan 2020]

Chapman, P. *The CRISP-DM User Guide* [online] NCR Systems Engineering Copenhagen. Available at: <<https://s2.smu.edu/~mhd/8331f03/crisp.pdf>> [Accessed 06 April 2020]

Pro Global Business Solutions. Six steps in CRISP-DM the standard data mining process [online] Pro Global Solutions. Available at: <<http://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process/>> [Accessed 06 April 2020]

Chauhan, N. 2019. Decision Tree Algorithm – Explained [online] Towards Data Science. Available at <<https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4>> [Accessed 07 April 2020]

Sayad, S. K Nearest Neighbors – Classification [online] Available at <<https://www.saedsayad.com/k_nearest_neighbors.htm>> [Accessed 07 April 2020]

Pant, A. 2019. Introduction to Logistic Regression [online] Towards Data Science. Available at <<https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148>> [Accessed 07 April 2020]

Yiu, T. 2019. Understanding Random Forest [online] Towards Data Science. Available at <<https://towardsdatascience.com/understanding-random-forest-58381e0602d2>> [Accessed 07 April 2020]