Final Project Report

## Visa Application Checker

By Jordan Lillis and Ian Mac Mathuna

28/02/2019

[Visa Application Checker](#_8zt8wkkoa9ih) 1

[**Introduction**](#_y6htia68td2i) **3**

[**Description**](#_1chslcrhc67r) **3**

[Aims](#_6gto97gbsemb) 3

[Data Cleaning](#_d6rhgbdex05d) 3

[Data Processing](#_xsdzrd4iaz7w) 6

[**Outcomes**](#_ei2dqtmpjfmb) **7**

[**Analysis Of Project Success**](#_am2eo65exnla) **7**

[**Conclusions**](#_rwh5s4j5j4c1) **8**

# Introduction

Our project was to build an application which would allow a user to predict the outcome of their American Visa application. We used Python to process our data and we used RapidMiner to run some machine learning algorithms such as Naive Bayes and Decision Trees. From undertaking this project we both learned a great deal about data science as a whole and it also helped us to improve previous skills such as Python.

# Description

## Aims

The goal of our project was to use a dataset, from Kaggle.com, which contained American Visa applications from the years 2012-2016 to predict whether or not a future Visa application would be successful.

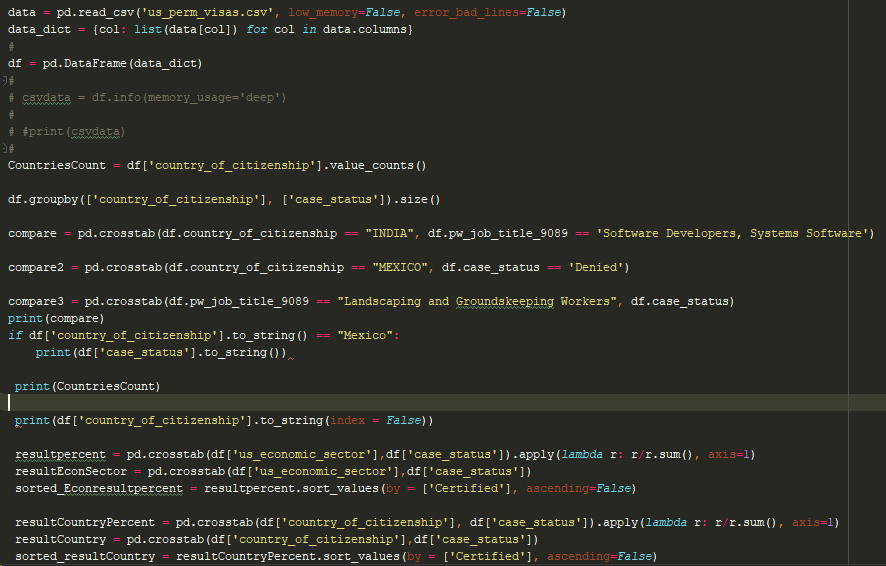
## Data Cleaning

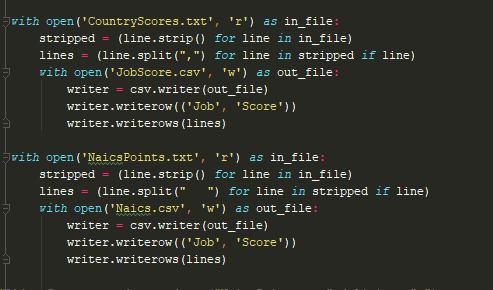
The original data contained approximately 320,000 rows of data, 120 columns for each row and had an approximate size of 250mb. However, some of this data was obsolete and had to removed or edited. We also decided not to use every column for our project as there was a lot of unnecessary information which would not of been useful to predict an applications success.

We believed that this would be a useful project as we would be working with a large dataset which would be a new challenge that neither of us had faced before. By using this data we would learn how to prepare and clean data for later processing and hopefully learn how to gain insights into large datasets. Another reason for choosing this project was that we agreed that it could have some real world use as there is currently no way to check if a visa will be accepted without directly applying for one which can take time and money.

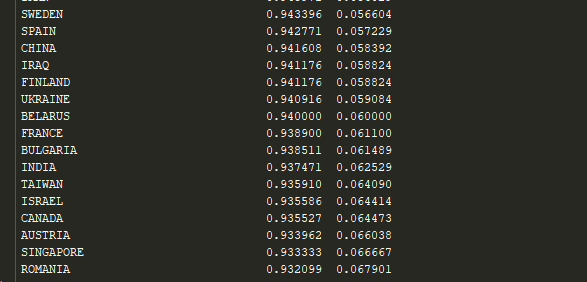
We used Python to extract the data we needed from this dataset such as country of citizenship, economic sectors and job titles of applicants. This was done using our Test.py file which we spent the first iteration of our project working on along with documentation. Python libraries such as Pandas and Matplotlib were implemented in order to help us work with this large amount of data. We first decided to clean our data by removing unnecessary columns and any duplicates or null values and we found that Python was the most efficient way of doing this. We then made separate CSV files with chunks of data such as a CSV file for each year. This allowed us to run our program quicker as it was dealing with less amount of data.

Here is an example of some of the code that was used with the pandas in order to examine the data and the effects that the individual columns had on the case status.

Here is an example of some of the code that was used to change the data which was in text files into csv files in order to make them easier to work with.



Here is some of the Countries Certified and Denied percentages:



## 

## 

## 

## 

## 

## 

## 

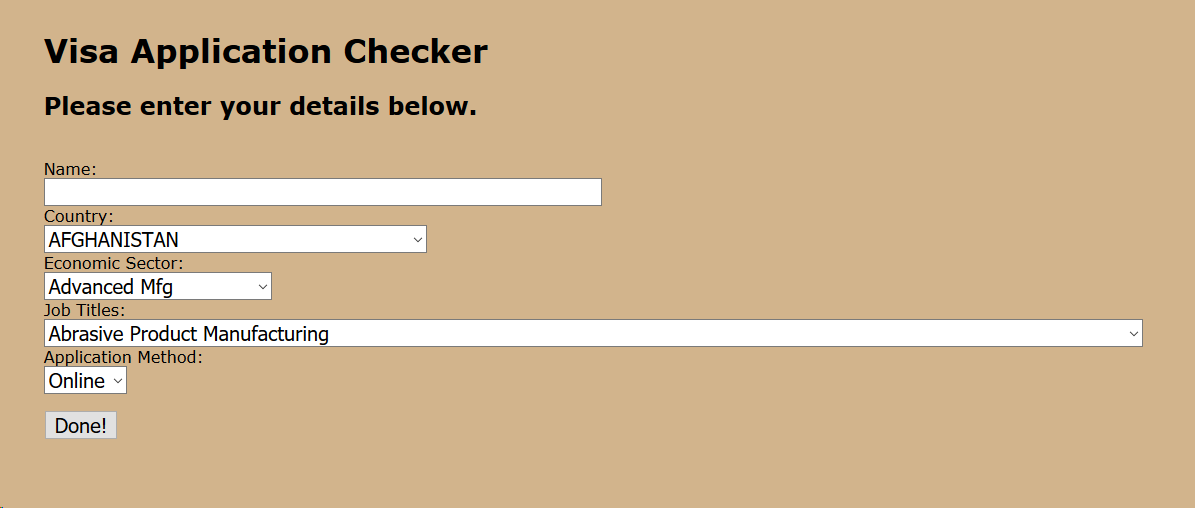
## 

## Data Processing

When our data was ready for processing we decided that we would make a points system which would determine the outcome of a application. Using the smaller Csv files we made earlier and using crosstabs in Python we gave each country, economic sector and job title a value based on the previous success rates of Visas.

In order to test our application we made a simple HTML form within Python that would take inputs from dropdown boxes populated with options from our CSV files. When this form is submitted our Python code will assign points to the inputted values and get a total. This total amount of points then determines how likely an application is to be accepted.

Here is what the form looks like:



Here is what the results screen looks like:

# 

# Outcomes

Our finished project completed the goal which had set out at the beginning of predicting whether a Visa will be accepted or denied. However, we would have liked to achieve some better accuracy with this prediction. We managed to create a simple form which takes less than five minutes to complete and gives the user an idea of the outcome of their Visa.

We also managed to clean a large set of data which was very time consuming in the early stages of the project especially when we didn’t have any previous experience with this type of work.

# 

# Analysis Of Project Success

We managed to succeed in the original goal which we had set out at the beginning of this project. Although it is a very basic implementation, it uses the points idea which we had planned to use from the beginning.

We had some slight issues with bad entries in the dataset, quite a few entries had misspelled job titles, others had nothing in some entries which hindered our progress.

The job titles entry had quite a lot of varied titles as these were actually taken from the people who typed/wrote them, this meant generalising a lot of the titles in order to make it so that when analyzing the data there weren’t too many outliers which would skew the end result.

We believe our project was somewhat successful as we managed to gain an insight into a large dataset and our project could be used by a future applicant to see if their application will be successful.

However, we would have liked to implement more machine learning into our project to gain better accuracy. We only thought of some ideas at a very late stage of our project and we didn’t have time to add these features.

One of the ideas that came to mind was that we could have used multiple countries Visa applications in order to give the user a suggestion as to where their application would have had the highest chance of being certified. This would have required visa application data from multiple different countries and would have meant that we would have required us to repeat the process for all of the data. This would have been too time consuming and probably would not have been completed in the timeframe of the project.

# Conclusions

This project, through the analysis of the dataset solidified quite a few opinions that we would have held before doing this project, such as first world countries and IT (or other high end economic sectors) applicants being more successful. However, now after using the dataset we now have concrete numbers and evidence to prove that these are not just opinions, but facts.

We also found out some interesting things through this data that we wouldn’t have known beforehand, such as India being the country with the most applicants for permanent visas into america. We found that we gain a somewhat accurate prediction of a Visa by looking at data from previous years and we discovered some trends that would not have been recognised otherwise.

While working on this project we both managed to improve our Python skills and work with libraries such as Pandas which we had not used before. We also gained some knowledge of using RapidMiner and understanding data. If we were to be assigned a data science project in the future, I believe that now we would have a good foundation to start from.