`

**TDS2101 - INTRODUCTION TO DATA SCIENCE**

**Lecture:**

TC1V

**Project Title:**

Financial Crimes Enforcement Network (FinCEN)

|  |  |  |
| --- | --- | --- |
| **Name** | **Student ID** | **Email** |
| Tan Kuang Chen | 1171103939 | 1171103939@student.mmu.edu.my |
| Ng Rui Wen | 1171103253 | 1171103253@student.mmu.edu.my |
| Sashrreik Menon | 1171103922 | 1171103922@student.mmu.edu.my |

# 

Table of contents

**Part A**

[**1.0 Introduction**](#_aplfrshrhmzv) **2**

[**2.0 Motivation**](#_sycetmh1c28x) **2**

[**3.0 Problem Description**](#_i0ddqtn92bok) **2**

[**4.0 Questions**](#_q15rvtu3rtj7) **3**

[**5.0 Dataset**](#_3znutorzo2du) **3**

[**6.0 Benefits of this study**](#_n1pks8r1d9ja) **4**

**Part B**

[**7.0 Data Science Pipeline**](#_qk90as4f7uh) **5**

[**8.0 Question Description**](#_7ex7tr9vp9ni) **5**

[**9.0 Data Collection**](#_ljyraet4dz19) **5**

[9.1 Transaction map](#_mc2xgq27axnc) 6

[9.2 Bank connection](#_9e62sp40e7i9) 7

[9.3 Annual gdp](#_k4sgkk57b5o7) 7

[9.4 Tax revenue share of gdp](#_yqprqyvyto27) 8

[**10.0 Data Preprocessing**](#_i18q0ayeaijn) **9**

[**11.0 Data Mining**](#_567kl54tfl1e) **14**

[11.1 Data Clustering](#_vo0iz6sxax6n) 14

[**12.0 Data Visualization**](#_97u6vzsm2ear) **15**

[12.1 Descriptive Question 1](#_r3v40lql5044) 15

[12.2 Descriptive Question 2](#_xrc1zc3hzpef) 17

[12.3 Exploratory Question](#_5n4551gpll56) 20

[12.4 Causal Question](#_twfx1druu6sl) 24

[**13.0 Challenges Encountered**](#_o763liihlo5) **26**

[**14.0 References**](#_8mvgjt73tme9) **27**

# 

# **Part A**

# 1.0 Introduction

The dataset we have chosen is FinCEN which is also known as the Financial Crimes Enforcement Network. It is a bureau of the United States Department of the Treasury that collects and analyzes information about financial transactions. FinCEN’s mission is to safeguard the financial system from illicit use, combat money laundering and promote national security.

# 2.0 Motivation

In current times, the use of online banking has become much more dominant than traditional ways. Businesses are slowly converting from the traditional banking methods to online transactions. Due to the increase of usage of online banking, we can see that e-banking is playing a huge role. People nowadays are relying on online shopping platforms such as Shopee, Lazada etc. which utilise online banking as their primary payment method. However there are downsides to online banking such as illegal transactions that may be harder to trace as there are millions of online transactions happening on a daily basis. Many large corporations use this as an opportunity to perform money laundering.

# 3.0 Problem Description

**Title: The effect of money laundering to the economy**

Money laundering damages financial sector institutions that are critical for economic growth, promoting crime and corruption that slow economic growth, reducing efficiency in the real sector of the economy. We believe that this is caused by money laundering activities that are performed by corrupt politicians. Among its other negative socioeconomic effects, money laundering transfers economic power from the market, government, and citizens to criminals. The purpose of this study is to identify the effects of money laundering on the economy of different countries. The data used are transactions from the year 2000 to 2017 that were flagged by financial institutions as suspicious to the authorities of the United States of America.

# 4.0 Questions

We have come up with a number of questions that we intend to answer by going through the data science process. The questions are listed below:

Descriptive:

1. Which country has the most transactions flagged as suspicious around the world?
2. What is the total number of suspicious transactions discovered per year?

Exploratory:

1. What is the relationship between financial crime and the gdp of a country?

Causal:

1. Does the money generated by financial crimes decrease tax revenues of a country?

# 5.0 Dataset

|  |  |
| --- | --- |
| **Dataset Name** | **Resources** |
| transaction map | Record of transaction from originator to beneficiary country |
| bank connection | Relationship between filer and entity bank |
| annual gdp | Annual gdp of countries |
| tax revenue share of gdp | Total tax revenue as a percentage of gdp |

Datasets that were given to us:

1. FinCEN - [Explore the FinCEN Files data](https://www.icij.org/investigations/fincen-files/download-fincen-files-transaction-data/)

This link includes both the transaction map and bank connection datasets.

Supplementary datasets:

1. Annual gdp - [The World Bank GDP Data](https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2018&most_recent_year_desc=false&start=2000)
2. Tax revenue share of gdp - [Global data on taxes](https://ourworldindata.org/the-new-oecd-global-revenue-statistics-database)

# 6.0 Benefits of this study

By studying the various effects of money laundering to the economy of different countries, the root cause of this can be identified and methods to overcome this problem can be found. This may heavily influence the development of several countries that are trying to achieve a well balanced economic system.

# **Part B**

# 7.0 Data Science Pipeline

The data science pipeline consists of five steps which are known as data collection, data preprocessing, data analysis, data mining and data visualization. We are going to solve the problems and questions we have regarding the datasets given to us using the data science pipeline.

# 8.0 Question Description

For the descriptive questions, we want to find out which country has the most illegal transactions around the world to find out if a financial crime would affect the economic system of that certain country. The second descriptive question, we try to figure out the total number of suspicious transactions discovered every year is to see if there is an increasing pattern for the coming years. For the exploratory question, we want to find the relationship between financial crime and gdp of a country to see if they are related or not. For the predictive question, we want to be able to estimate the gdp of a country with the continuation of financial crimes and how it will affect the gdp. For the causal question, we want to find out if the tax revenue of a country decreases when there are more illegal transactions happening in the country.

# 9.0 Data Collection

We were given 2 datasets which are related to FinCEN, the data contains information on more than $35 billion in transactions dated from 2000-2017 that were flagged by financial institutions as suspicious to United States authorities. The first dataset given to us named transaction map contained many vital details of the flagged transactions such as the names and countries of the originator and beneficiary banks, dates and even transaction amounts. The second dataset named bank connections gives us the details of the connections of the banks. It has the data of the filer organizations and the details of the banks that were involved in the flagged transactions. Besides that, we also have two supplementary datasets. The first supplementary dataset we have is called annual gdp. This dataset consists of the annual gdp of countries ranging from the year 1960-2019. The second supplementary dataset we have is called tax revenue share of gdp. This dataset contains the total tax revenue that is collected by the government as a percentage of gdp. This dataset consists of data from the year 1990-2016.

|  |  |
| --- | --- |
| **Dataset Name** | **Resources** |
| transaction map | Record of transaction from originator to beneficiary country |
| bank connection | Relationship between filer and entity bank |
| annual gdp | Annual gdp of countries |
| tax revenue share of gdp | Total tax revenue as a percentage of gdp |

## 9.1 Transaction map

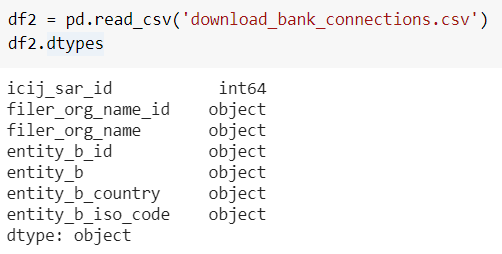
Transaction map dataset contains attributes which have the records of flagged transactions.



filer\_org is the organisation that performed the transaction. The dates of each transaction have also been recorded to know when the transaction was made. originator\_bank is the bank the organisation used to perform the transaction. beneficiary\_bank is the bank that was supposed to be credited the money from the transaction. The number and amount of the transactions are also recorded in the dataset.

## 9.2 Bank connection

Bank connection dataset contains attributes which records the filer organisation and the banks.

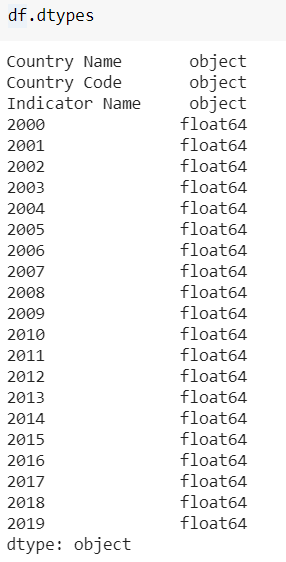


filer\_org is as mentioned in the above, the organisation that performed the transaction. entity\_b is the bank that is associated with said organisation for the flagged transaction.

However, we did not use this dataset as it was not needed for our questions.

## 9.3 Annual gdp

Annual gdp dataset contains attributes of gdp from many countries from the year 2000-2019.



## 9.4 Tax revenue share of gdp

Tax revenue share of gdp dataset contains the percentage of a country’s gdp that is generated by tax revenues.



The entity and code are to identify the countries.

# 10.0 Data Preprocessing

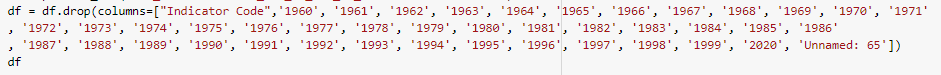
First of all, we used Google Colab Notebook to read all the datasets that were used.



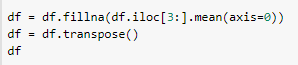




The data cleaning methods that were used were dropping columns. The dataset contained the gdp of countries from the year 1960 to 2020. However, we only needed the data from the year 2000 to 2017 which is why we dropped the remaining years for the annual gdp supplementary dataset.



We also dropped the indicator code as the name of the countries are sufficient for our understanding. Besides that, we also filled the empty spaces with the mean of the gdp. We transposed the dataset because it is easier to manipulate the data once the dataset was transposed.



For the tax revenue dataset, we only used the data after the year 2000 since the FINCEN dataset only has information from 2000 to 2017.



We also splitted the date in the FINCEN transaction map dataset to several attributes such as day, month and year and also converted the attribute data type to numeric which will make it easier to process the data later.



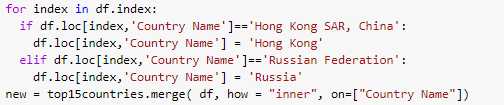
After splitting the date to a new column, we renamed the column values to a suitable way in order to arrange the data clearly.



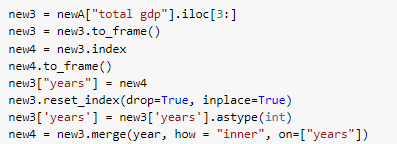
Due to having a large number of countries, we selected the top 15 countries with the highest amount of suspicious transactions flagged to find the solutions to our questions.



We also renamed a few attribute values in order to merge the dataset which made it easier to discover the relationship between datasets.



The total gdp of the gdp dataset of each year has been calculated and merged into another dataset in order to do a comparison and find the answer of the questions later on.

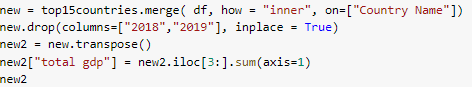


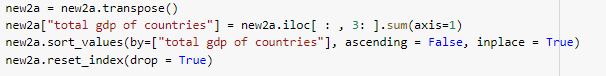
As we have said in the previous, after we finished cleaning the dataset we merged the dataset to find out the answers to the questions that we have asked.







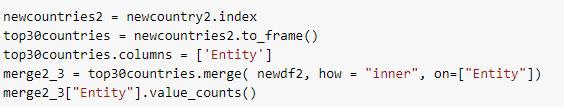


By transposing a dataset, it can manipulate the dataset which could possibly make the dataset easier to merge and perform data processing faster. Sorting also makes the dataset easier to be read.****

By using the FINCEN transaction map dataset, we managed to get the total amount of suspicious transactions of the countries in our sample and appended it into a new dataframe which contained only the attributes needed to solve the questions later. 

We also do the same for tax revenue and financial crime by merging a subset of financial crime that has the top 30 countries to tax revenue dataset. We choose 30 instead because some of the countries in the financial crime dataset is not included inside the tax revenue dataset. After merging, we counted the total countries and found only 15 are included.





# 11.0 Data Mining

## 11.1 Data Clustering

For data mining, we used a clustering method in order to find out the relationship of the different clusters between the 15 countries and the total amount of transactions in different countries from 2000 to 2017. We applied KMeans function to scale the dataset based on certain attributes and it gives a kmeans attribute which can cluster the dataset to 2 different groups.





# 12.0 Data Visualization

## 12.1 Descriptive Question 1

The first description question asked which country has the most transactions flagged as suspicious around the world which we answered by using the graph as shown below. We have chosen the top 15 countries with the highest amount of flagged transactions in order to form a bar chart. This is done to be able to see the number of flagged transactions for every country clearly. As we can see from the graph below, Latvia is the country that has sent the most transactions that have been flagged as suspicious.

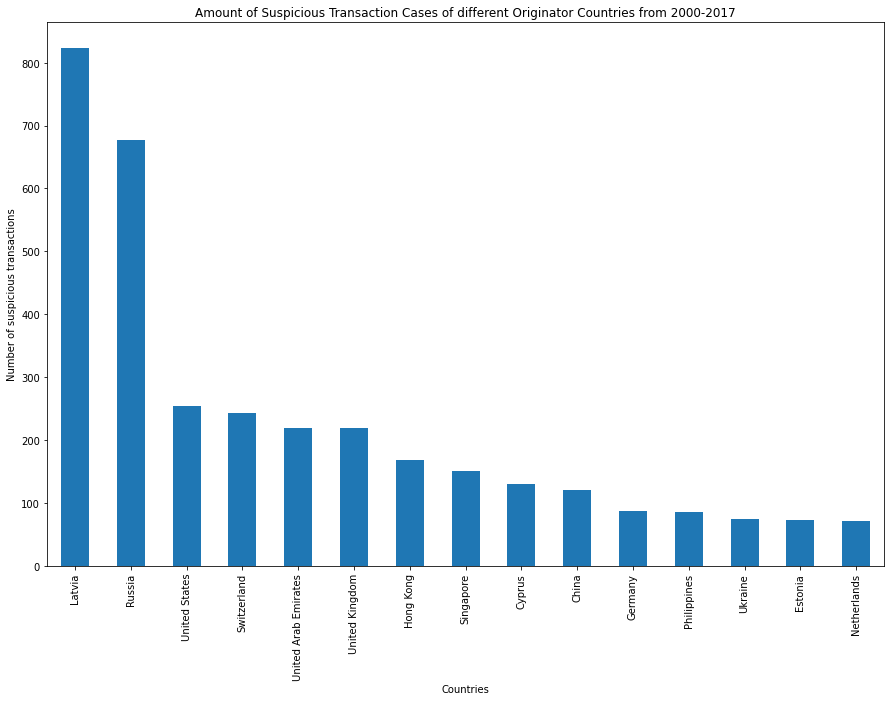
****

Figure 1: Amount of Suspicious Transaction Cases of different Originator Countries from 2000-2017

Latvia is also found to be the country that has received the most transactions that have been flagged as suspicious.

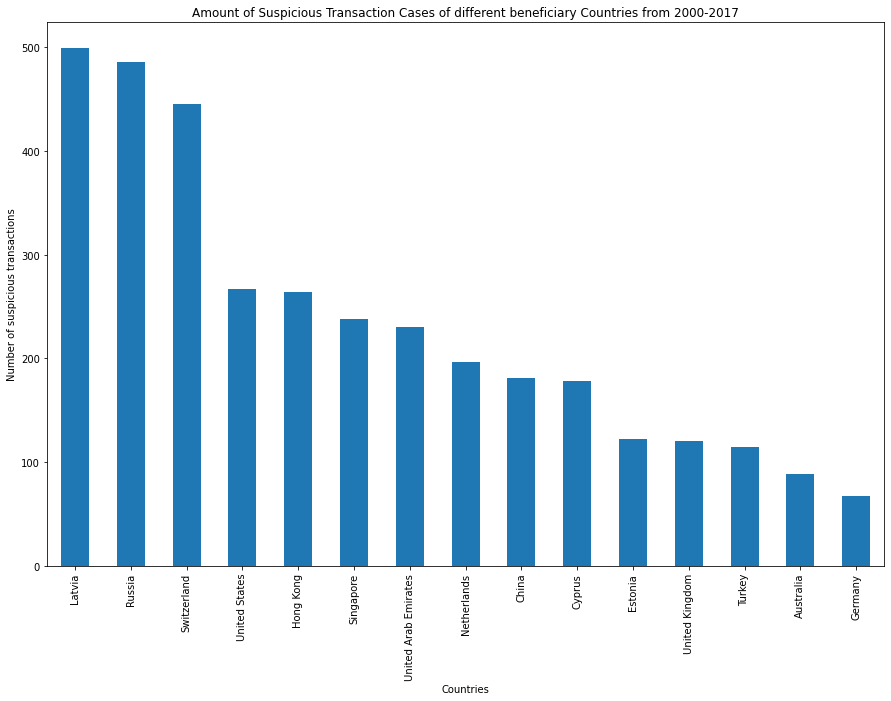
****

Figure 2: Amount of Suspicious Transaction Cases of different beneficiary Countries from 2000-2017

## 12.2 Descriptive Question 2

For the second descriptive question, we counted the transactions flagged as suspicious per year and represented it through a bar chart because the comparison through the bar chart is clear and easy to understand. As we can see from the graph below, 2014 has the highest amount of suspicious transactions discovered and we can see that the trend of it has been constantly increasing since 2007 but it seems like the suspicious transactions flagged during 2017 has significantly reduced. We also calculated the total amount transacted through suspicious transactions per year since 2000 and we can see that both the graphs are having similar patterns and the total amount also began to increase since 2007.

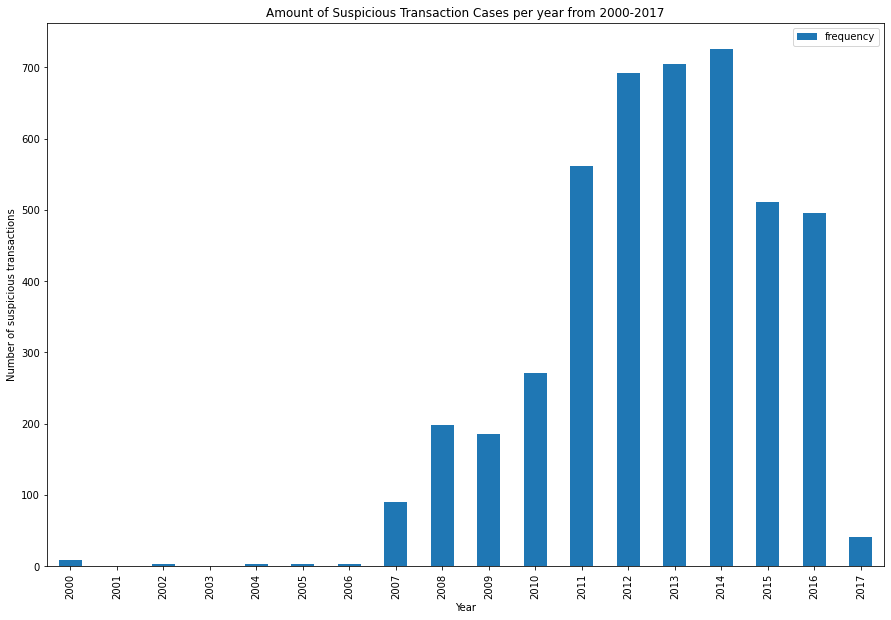
****

Figure 3: Amount of Suspicious Transaction Cases per year from 2000 - 2017

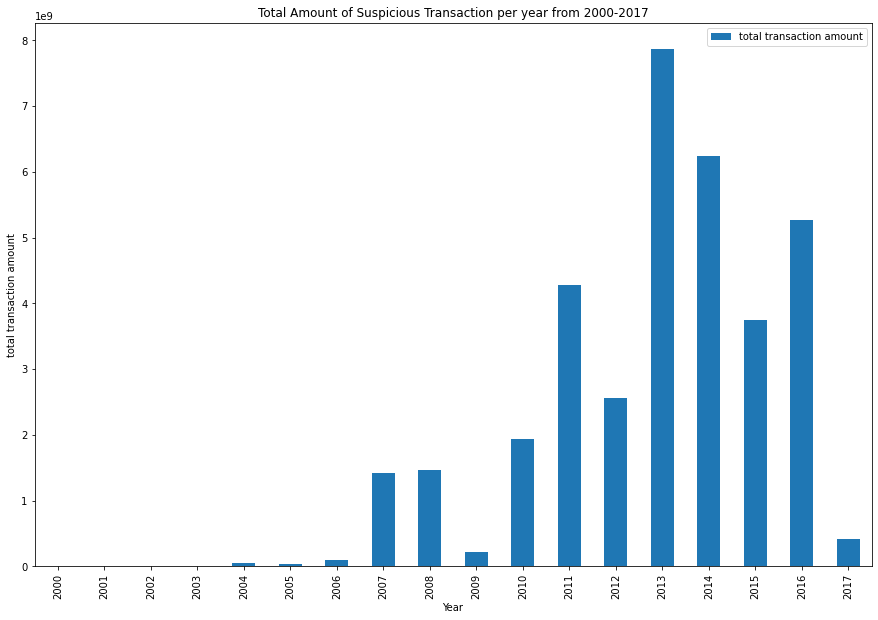
****

Figure 4: Total Amount of Suspicious Transactions per year from 2000-2017

We have plotted a heatmap to study the correlation between the suspicious cases discovered, years and total transaction amount. From the heatmap we can see that the suspicious cases discovered and total transaction amounts are having a really strong correlation which is approximately 0.9. The correlation value between year and amount of suspicious cases, and year with total transaction amounts are sitting above 0.7 which also means that there is somewhat of a strong relationship between them.

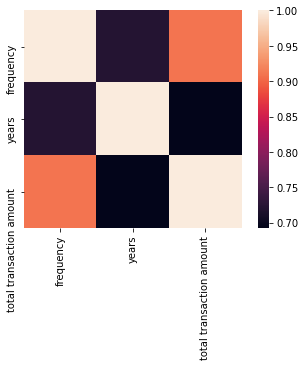
****

Figure 5: Heatmap of correlation between amount of suspicious cases, year and total amount of suspicious transactions

## 12.3 Exploratory Question

The scatter plot used the total gdp of the countries that were involved in financial crimes and made a comparison with the frequency of flagged transactions. But we have found out that the correlation between both the attributes did not turn out as we were expecting.

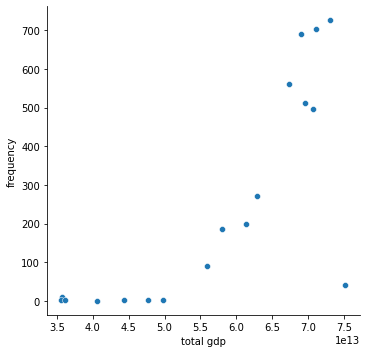
****

Figure 6: Scatter plot of total gdp of all countries against the frequency of suspicious transactions cases

The heatmap below has shown the result of the scatter plot and the correlation between the different attributes. The heatmap has indicated that the total flagged transactions discovered has brought no impact towards gdp and we can see that both the attributes are forming a positive relationship which is not the ideal answer that we were looking for.

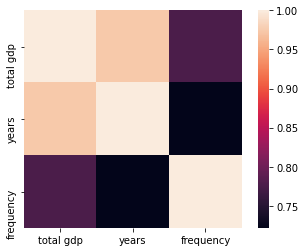
****

Figure 7: Heatmap of correlation between gdp of countries, year and total amount of suspicious transactions

We figured out that the amount of suspicious transaction cases discovered is not an appropriate attribute to be compared with the total gdp of a country. Hence, we break down the suspicious transaction cases discovered into the amount of total transactions based on the countries we’ve chosen from 2000 to 2017. After we merged the datasets, we implemented a data mining rule which is clustering to separate the countries into two different clusters. By employing a clustering method, we successfully plotted a new scatter plot which has the cluster of the dataset and it has shown the relationship between each country's total amount of suspicious transactions and the total gdp of the specific country. The K-Means Clustering scatter plot below has shown the result of the employment.

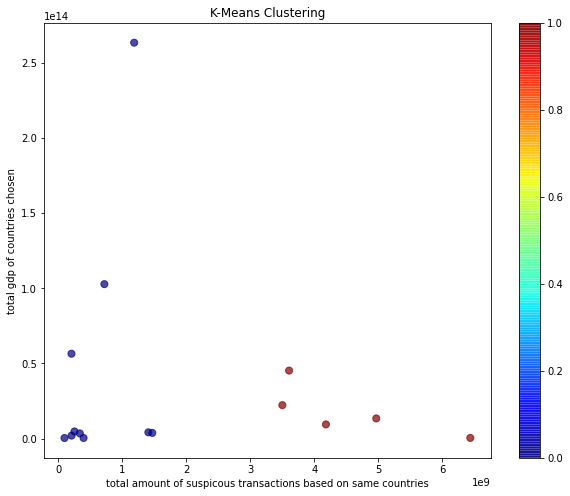
****

Figure 8: K-Mean clustering plot of the total gdp of countries with total amount of suspicious transactions

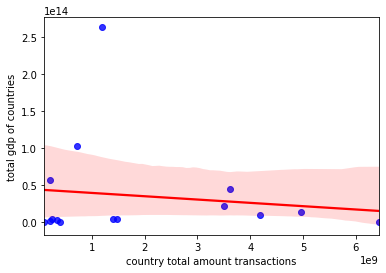
****

Figure 9: Regression plot of correlation between the total gdp of countries with total amount of suspicious transactions

As we can see from the K-mean plot above, it clearly shows that the countries which are represented by red colour are having a slightly lower gdp compared to some of the countries which are represented by blue colours. But we can see that most of the countries with lower amount of suspicious transactions have an average higher gdp compared to countries with high total amount of suspicious transactions. We can see that from the second regression plot, both the attributes are having a weak negative linear relationship which we can answer the question by assuming that the total amount of suspicious transactions will bring a minor impact on a country’s gdp. Suspicious transactions include money laundering which damages the financial sector of a country which is vital in the economic growth of the country. The economic growth of a country also depends on various other factors. According to K. Amadeo(2020), the four main factors that affect a country’s gdp are personal consumption, business investment, government spending, and net exports. This is why the total amount of suspicious transactions only show a minor impact on the gdp of a country.

## 12.4 Causal Question

From the financial crime dataset, we chose the top 15 countries that have the highest financial crime to explore their relationship towards tax revenue. Then, we merge that subset dataset to the tax revenue dataset to have those values together for better comparison. We used regplot to find the relationship between tax revenue and financial crime but as seen in the graph those points are spread quite far away which form a weak relationship. However, according to J. McDowell, there is a toll taken on tax revenues as they are diminished when there are a higher number of financial crimes in the country. This would automatically cause for higher tax rates which would burden honest taxpayers into paying more than they should.

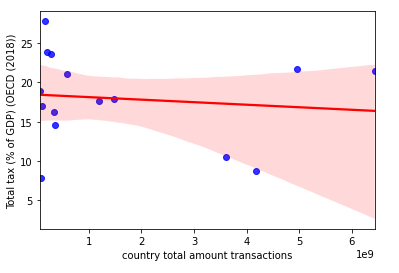


Figure 10: Regression plot of correlation between the tax revenue of different countries with total amount of suspicious transactions

The heatmap below has shown the result of the scatter plot and the correlation between tax revenue and financial crime. The heatmap has shown that the correlation between the attributes are negative but it can be seen that the relation of them are weak thus we can’t conclude that financial crimes decrease tax revenues of a country heavily.

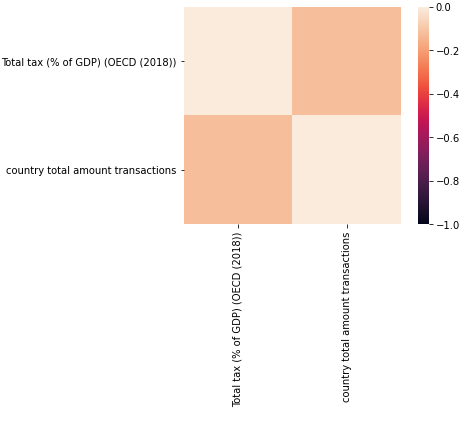


Figure 11: Heatmap of correlation between the tax revenue of different countries with total amount of suspicious transactions

# 13.0 Challenges Encountered

As students pursuing the field of IT, none of us had any knowledge regarding the economic status of countries. We had to gain knowledge regarding the economic sectors and factors that weigh in them. We also had difficulties with our questions as some of the answers did not turn out as we expected it to be. When coding, we experienced difficulties when we had to merge data from different datasets to be able to plot the graphs that we desired.

# 14.0 References

Amadeo, K. (2020, June 26). Four critical components of America's economic growth. Retrieved from https://www.thebalance.com/components-of-gdp-explanation-formula-and-chart-3306015

Negative effects of money laundering on the economy. (n.d.). Retrieved from https://sanctionscanner.com/blog/negative-effects-of-money-laundering-on-the-economy-132

McDowell, J. and Novis, G. (2001, May 2). The Consequences of Money Laundering and Financial Crime. Retrieved from <https://www.hsdl.org/?view&did=3549>