TDS2101
Introduction to Data Science
Trimester 2, Year 2019/20
Assignment (30%)

Tutorial Session: TT01

Lecturer: Dr. John See Su Yang

Submission Date: 19th February 2020

Prepared by:

Boo Ee Kein Ivan 1161104032

1161104032@student.mmu.edu.my

Low Seh Hong 1161104400

1161104400@student.mmu.edu.my

Table of Content

PART A

Overview

Unemployment has always been believed to be a gigantic issue for a country regardless of the development or the growth of a country. This topic is very interesting and we are intrigued by the idea as this issue is somewhat not alien to us all. We had always believed that education also plays a very big part for people to be employed in society. The myth is that if the people that do not achieve a higher education level will be deemed to be unemployed. But is that necessarily true? Besides, does the unemployment of the people hugely affect the growth of the economy of a country. Therefore, we are very intrigued to find out about the correlation regarding the level of education of the people, the unemployment rate and the economic growth of a country.

The country that our team will focus on is the United States. Historical data will be taken based on the particular country. Our team will do our very best to uncover the underlying relationship and to produce results with satisfaction.

Formulated Question

Does the level of education and unemployment rate gives impact towards the economic growth of the United States?

Aim of the Report

- To study the relationship between the level of education, the status of employment and the economic growth of United States
- To understand and to find out if a prediction can be done based on the study

Exploratory Questions

1. Does the level of education influence the rate of employment in the United States?
2. Does the level of employment correlates with the value of the GDP?
3. Are the employment status and the level of education of the people something that is very important to the country?

Predictive Questions

What are the factors that should be taken into consideration to predict the GDP of the United States?

Benefits and Imparts

Government

- Helps to predict the economic growth rate
- Helps the government to plan ahead for the sake of their country's economy
- Gives a guideline to all sectors throughout the country

Education sectors

- Able to understand the importance of education system in the country
- Able to incorporate the results and to plan a series of suitable education courses.

Individuals

- Enables individuals to know the trend of the country's economic growth rate
- Enables individuals to plan and to revise their education and career path

Target Audience

For the people that wanted to understand and to know about the economic growth trends of the country based on the correlations of employment rate and level of education.

Part B - DATA SCIENCE PIPELINE IN ACTION

1.0 Questions to Answer

The economy of a country is very important to a country as it is one of the main factors that can define a country's power. The level of the economy of a country can be measured by **Gross Domestic Product** (**GDP**) of a country. **Gross Domestic Product** (**GDP**) is the monetary value of all finished goods and services made within a country during a specific period. **GDP** provides an economic snapshot of a country, used to estimate the size of an economy and growth rate.

Besides, the importance of the workplace and employment compel some groups to monitor and change the employment rate. The U.S. government and Federal Reserve manage the employment rate by observing economic indicators, adjusting the interest rate and monitoring GDP.

And so, in this part, we limit the context to that of the government's perspective. As an organization such as the Government, severals questions could to be answered through the data science pipeline:

- Exploratory Questions
    1. Does the level of education influence the rate of employment in the United States?
    2. Does the level of employment correlates with the value of the GDP?
    3. Are the employment status and the level of education of the people something that is very important to the country?

- Predictive Questions
    1. What are the factors that should be taken into consideration to predict the GDP of the United States?

In the following sections, we shall explore the data, and build a predictive model that predicts the value of GDP for a year.

2.0 Dataset in Use

2.1 Data Sources

We acquired our dataset from two sources which is from The World Bank (https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=US) and also the U.S. Bureau of Labor Statistics (https://www.bls.gov/bls/infohome.htm).

2.2 Describing Data Set

The World Bank provides World Bank national accounts data and OECD National Accounts data files which includes the dataset of GDP growth rate of all the countries around the world from 1961 to 2018, named GDPcountries.csv and also GDP value of all the countries around the world from 1961 to 2018, named GDP value. The data site has a Creative Commons Attribution 4.0 International license which allowed the datasets to be downloaded as an open data and to be used freely.

The U.S Bureau of Labor Statistics measures labor market activity, working conditions, price changes, and productivity in the U.S. economy to support public and private decision making. The datasets such as the employment and unemployment level based on the attainment of education in U.S. can be easily extracted using their data extracting tool. We have extracted eight datasets in comma separated value form (.csv), which are two main categories: Employment Level and Unemployment Level from 1992 to 2020, where these two main categories all hold four separate files that are with four different attainment of education, which is education lower than highschool, education level at high school, education level at college and education level at degree and above.

1. GDP growth rate of all the countries around the world from 1961 to 2018 (GDPcountries.csv)

Variable	Continuous/Discrete	Comments
Country Name	Discrete	Name of the country.
Country Code	Discrete	Abbreviations of the name of the country.
Indicator Name	Discrete	An indicator of the type of data that is collected.

Indicator Code	Discrete	The code for the indicator name.
Year from 1960 to 2019 *All the years are in a separated column	Continuous	The GDP growth rate for a country

2. GDP value of all the countries around the world from 1961 to 2018 (GDPvalue.csv)

Variable	Continuous/Discrete	Comments
Country Name	Discrete	Name of the country.
Country Code	Discrete	Abbreviations of the name of the country.
Indicator Name	Discrete	An indicator of the type of data that is collected.
Indicator Code	Discrete	The code for the indicator name.
Year from 1960 to 2019 *All the years are in a separated column	Continuous	The GDP value for a country

3. Employment level of education lower than highschool (Employment_level_less_highschool.csv) , education level at high school (Employment_level_highschool.csv) , education level at college (Employment_level_college.csv) and education level at degree and above (Employment_level_degree_above.csv).

Variable	Continuous/Discrete	Comments
Year	Discrete	Years the data has been collected
Jan	Continuous	
Feb	Continuous	
Mar	Continuous	
Apr	Continuous	

May	Continuous	Employment Level [Numbers in thousands]
Jun	Continuous	
Jul	Continuous	
Aug	Continuous	
Sep	Continuous	
Oct	Continuous	
Nov	Continuous	
Dec	Continuous	

4. Unemployment level of education lower than highschool (Unemployment_level_less_highschool.csv) , education level at high school (Unemployment_level_highschool.csv) , education level at college (Unemployment_level_college.csv) and education level at degree and above (Unemployment_level_degree_above.csv).

Variable	Continuous/Discrete	Comments
Year	Discrete	Years the data has been collected
Jan	Continuous	Unemployment Level [Numbers in thousands]
Feb	Continuous	
Mar	Continuous	
Apr	Continuous	
May	Continuous	
Jun	Continuous	
Jul	Continuous	
Aug	Continuous	
Sep	Continuous	
Oct	Continuous	
Nov	Continuous	

Dec	Continuous	

3.0 Data Cleaning and Data Integration

3.1 Data Importation

3.1.1 Importing Datasets

Firstly, we will read the acquired datasets into dataframes.

```
countryGDP_percent = pd.read_csv('GDPcountries.csv')
countryGDP_value = pd.read_csv('GDPvalue.csv')
empt_lvl_less_highschool = pd.read_csv('Employment_level_less_highschool.csv')
empt_lvl_highschool = pd.read_csv('Employment_level_highschool.csv')
empt_lvl_college = pd.read_csv('Employment_level_college.csv')
empt_lvl_degree_above = pd.read_csv('Employment_level_degree_above.csv')
unempt_lvl_less_highschool = pd.read_csv('Unemployment_level_less_highschool.csv')
unempt_lvl_highschool = pd.read_csv('Unemployment_level_highschool.csv')
unempt_lvl_college = pd.read_csv('Unemployment_level_college.csv')
unempt_lvl_degree_above = pd.read_csv('Unemployment_level_degree_above.csv')
```

3.1.2 Data Cleaning Process

1. Data Cleaning Process for GDPCountries.csv

   First, we will make the index of countryGDP to follow the names of the Countries. Then, we will take the row which contains the data of the United States and read them into separate dataframe.

```
countryGDP_percent.set_index('Country Name',inplace = True)
usaGDP_percent = countryGDP_percent.loc[['United States']]
print(usaGDP_percent.head())
```

```
             Country Code         Indicator Name     Indicator Code  1960  \
Country Name
United States          USA  GDP growth (annual %)  NY.GDP.MKTP.KD.ZG   NaN

              1961  1962  1963  1964  1965  1966  ...      2010      2011  \
Country Name                                      ...
United States   2.3   6.1   4.4   5.8   6.4   6.5  ...  2.563767  1.550836

                 2012      2013      2014     2015      2016     2017  \
Country Name
United States  2.249546  1.842081  2.451973  2.88091  1.567215  2.21701

                 2018  2019
Country Name
United States  2.927323   NaN

[1 rows x 63 columns]
```

As we can see there's 2 NaN values which are in 1960 and 2019. To make sure that the 2 years are the only NaN values, we will perform the sum of total NaN values of the whole dataframe.

```
usaGDP_percent.isnull().sum()
```

```
Country Code      0
Indicator Name    0
Indicator Code    0
1960              1
1961              0
                 ..
2015              0
2016              0
2017              0
2018              0
2019              1
Length: 63, dtype: int64
```

So it is true that the NaN values are only in 1960 and 2019. And so, we will remove both of the columns to prevent misleading data analysing. Besides, we will also remove the unused columns of the dataset and reshuffle the dataframe to have the year as row values and index.

```
usaGDP_percent.drop(['Country Code','Indicator Name','Indicator Code','1960','2019'],axis = 1,
                    inplace = True)
usaGDP_percent = pd.melt(usaGDP_percent,var_name = 'Year',value_name = 'GDP_percent')
usaGDP_percent.set_index('Year', inplace = True)
usaGDP_percent.index = usaGDP_percent.index.astype('int64')
print(usaGDP_percent.head())
print(usaGDP_percent.tail())
```

```
      GDP_percent
Year
1961          2.3
1962          6.1
1963          4.4
1964          5.8
1965          6.4
      GDP_percent
Year
2014     2.451973
2015     2.880910
2016     1.567215
2017     2.217010
2018     2.927323
```

As the dataset of GDPvalue.csv has the same structure as the GDPcountries.csv, we will perform a similar cleaning and restructuring operation on it.

2. Data Cleaning Process for GDPvalue.csv

```
countryGDP_value.set_index('Country Name',inplace = True)
usaGDP_value = countryGDP_value.loc[['United States']]
usaGDP_value.drop(['Country Code','Indicator Name','Indicator Code','1960','2019'],axis = 1,
                  inplace = True)
usaGDP_value = pd.melt(usaGDP_value,var_name = 'Year',value_name = 'GDP_value')
usaGDP_value.set_index('Year', inplace = True)
usaGDP_value.index = usaGDP_value.index.astype('int64')
```

3. Data Cleaning Process for Employment_level_less_highschool.csv

The structure of the employment and unemployment level for different attainment of education level is shown below. All the files show the same structure.

```
empt_lvl_less_highschool.tail()
```

	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
24	2016	10101	9976.0	10059.0	10021.0	9710.0	9771.0	9899.0	9912.0	9829.0	9871.0	9755.0	9684.0
25	2017	9678	9501.0	9521.0	9439.0	9707.0	9658.0	9821.0	9867.0	9859.0	9732.0	9698.0	9508.0
26	2018	9710	9673.0	9676.0	9717.0	9719.0	9925.0	9682.0	9699.0	9620.0	9676.0	9647.0	9654.0
27	2019	9468	9721.0	9532.0	9453.0	9383.0	9482.0	9458.0	9478.0	9440.0	9240.0	9245.0	9379.0
28	2020	9090	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

We will want to check the sum of NaN value in the dataset.

```
empt_lvl_less_highschool.isnull().sum()
```

```
Year    0
Jan     0
Feb     1
Mar     1
Apr     1
May     1
Jun     1
Jul     1
Aug     1
Sep     1
Oct     1
Nov     1
Dec     1
dtype: int64
```

There is only one NaN value from Feb until Dec, which is the previous tail view of the data set. We know that all the NaN value confide in the row of Year 2020, which makes sense as the data is not yet available. With that, we will remove the last row and check

again.

```
empt_lvl_less_highschool.drop([28], axis = 0 , inplace = True)
```

Since it has been removed, we will check again with the sum of NaN values in the dataframe.

```
empt_lvl_less_highschool.isnull().sum()
```

```
Year    0
Jan     0
Feb     0
Mar     0
Apr     0
May     0
Jun     0
Jul     0
Aug     0
Sep     0
Oct     0
Nov     0
Dec     0
dtype: int64
```

Now, for our analysis, we want to take values for the year but not the month. So, we will find the mean value for each year. First, we will set the index to Year, then proceed with our calculation.

```
empt_lvl_less_highschool.set_index('Year',inplace = True)
empt_lvl_less_highschool['Emp_Less_Highschool'] = empt_lvl_less_highschool.mean(axis = 1)
```

We will then drop the columns of the months as we only needed the mean annually.

```
empt_lvl_less_highschool.drop(['Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'],
                              axis = 1 , inplace = True)
```

At last, we will take a quick view to our cleaned and reconstructed dataframe.

Year	Emp_Less_Highschool
1992	11844.833333
1993	11203.166667
1994	11055.416667
1995	10945.666667
1996	11314.916667

As the following datasets share the same original structures as this datasets, so the similar cleaning and reconstructing process will be applied to the datasets.

4. Data Cleaning Process for Employment_level_highschool.csv

```
empt_lvl_highschool.drop([28], axis = 0 , inplace = True)
empt_lvl_highschool.set_index('Year',inplace = True)
empt_lvl_highschool['Emp_Highschool'] = empt_lvl_highschool.mean(axis = 1)
empt_lvl_highschool.drop(['Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'],
                         axis = 1 , inplace = True)
```

5. Data Cleaning Process for Employment_level_college.csv

```
empt_lvl_college.drop([28], axis = 0 , inplace = True)
empt_lvl_college.set_index('Year',inplace = True)
empt_lvl_college['Emp_College'] = empt_lvl_college.mean(axis = 1)
empt_lvl_college.drop(['Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'],
                         axis = 1 , inplace = True)
```

6. Data Cleaning Process for Employment_level_degree_above.csv

```
empt_lvl_degree_above.drop([28], axis = 0 , inplace = True)
empt_lvl_degree_above.set_index('Year',inplace = True)
empt_lvl_degree_above['Emp_Degree_Above'] = empt_lvl_degree_above.mean(axis = 1)
empt_lvl_degree_above.drop(['Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'],
                         axis = 1 , inplace = True)
```

7. Data Cleaning Process for Unemployment_level_less_highschool.csv

```
unempt_lvl_less_highschool.drop([28], axis = 0 , inplace = True)
unempt_lvl_less_highschool.set_index('Year',inplace = True)
unempt_lvl_less_highschool['Uemp_Less_Highschool'] = unempt_lvl_less_highschool.mean(axis = 1)
unempt_lvl_less_highschool.drop(['Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'],
                         axis = 1 , inplace = True)
```

8. Data Cleaning Process for Unemployment_level_highschool.csv

```
unempt_lvl_highschool.drop([28], axis = 0 , inplace = True)
unempt_lvl_highschool.set_index('Year',inplace = True)
unempt_lvl_highschool['Uemp_Highschool'] = unempt_lvl_highschool.mean(axis = 1)
unempt_lvl_highschool.drop(['Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'],
                         axis = 1 , inplace = True)
```

9. Data Cleaning Process for Unemployment_level_college.csv

```
unempt_lvl_college.drop([28], axis = 0 , inplace = True)
unempt_lvl_college.set_index('Year',inplace = True)
unempt_lvl_college['Uemp_College'] = unempt_lvl_college.mean(axis = 1)
unempt_lvl_college.drop(['Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'],
                              axis = 1 , inplace = True)
```

10. Data Cleaning Process for Unemployment_level_degree_above.csv

```
unempt_lvl_degree_above.drop([28], axis = 0 , inplace = True)
unempt_lvl_degree_above.set_index('Year',inplace = True)
unempt_lvl_degree_above['Uemp_Degree_Above'] = unempt_lvl_degree_above.mean(axis = 1)
unempt_lvl_degree_above.drop(['Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec'],
                             axis = 1 , inplace = True)
```

3.1.3 Merging Data Frame

```
empt_merge_lesshighschool_highschool = pd.merge(empt_lvl_less_highschool,empt_lvl_highschool,
                                                on = 'Year', how= 'left')
empt_merge_college_degree_above = pd.merge(empt_lvl_college,empt_lvl_degree_above,
                                           on = 'Year', how= 'left')
empt_merge_all = pd.merge(empt_merge_lesshighschool_highschool,empt_merge_college_degree_above,
                          on = 'Year', how= 'left')
unempt_merge_lesshighschool_highschool = pd.merge(unempt_lvl_less_highschool,unempt_lvl_highschool,
                                                  on = 'Year', how= 'left')
unempt_merge_college_degree_above = pd.merge(unempt_lvl_college,unempt_lvl_degree_above,
                                             on = 'Year', how= 'left')
umempt_merge_all = pd.merge(unempt_merge_lesshighschool_highschool,unempt_merge_college_degree_above,
                            on = 'Year', how= 'left')
merge_all_empt_umempt = pd.merge(empt_merge_all,umempt_merge_all,
                                 on = 'Year', how= 'left')
merge_GDP_percent_value = pd.merge(usaGDP_percent,usaGDP_value,
                                   on ='Year', how = 'left')
merge_GDP_all_empt_umempt = merge_all_empt_umempt.join(merge_GDP_percent_value)
```

We merge all the data that has the annual mean of employment and unemployment level based on education level attainment together with the GDP value and GDP growth percentage using ‘year’ as the key.

```
merge_GDP_all_empt_umempt.head()
```

Year	Emp_Less_Highschool	Emp_Highschool	Emp_College	Emp_Degree_Above	Uemp_Less_Highschool	Uemp_Highschool
1992	11844.833333	35305.250000	25522.750000	27273.416667	1535.583333	2589.083333
1993	11203.166667	35394.500000	26898.083333	28114.416667	1358.666667	2363.833333
1994	11055.416667	35143.833333	28689.083333	29253.250000	1196.750000	1981.416667
1995	10945.666667	35008.750000	29674.666667	30407.833333	1078.500000	1744.666667
1996	11314.916667	35311.666667	29989.583333	31456.833333	1077.083333	1722.250000

Uemp_College	Uemp_Degree_Above	GDP_percent	GDP_value
1527.666667	895.000000	3.522442	6.520330e+12
1483.250000	854.166667	2.752844	6.858560e+12
1334.750000	772.666667	4.028839	7.287240e+12
1227.750000	763.583333	2.684287	7.639750e+12
1165.500000	723.666667	3.772501	8.073120e+12

3.1.4 Remove NaN from Merged Table (merge_GDP_all_empt_umempt)

As the GDP value and growth rate contains NaN value, we will remove the data for 2019.

3.1.5 Data Quality

Throughout the data cleaning process, we found that most of the data seems to be of good quality. The NaN values that have been found are all because the new data is not available yet so it was not added into the dataset.

3.1.6 Z-Score Normalization

We will then perform Z-score normalization on employment and unemployment levels in order to make visualizations easier with the GDP_percent which is the GDP growth rate.

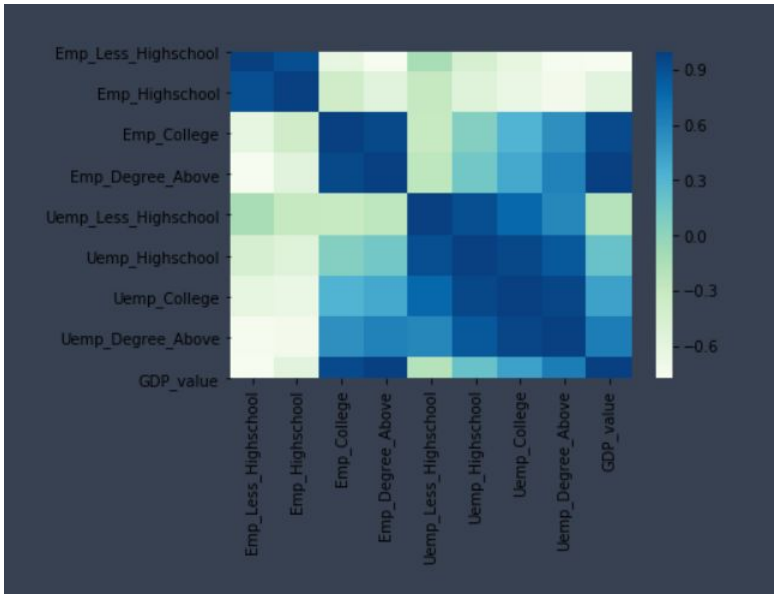
Year	Emp_Less_Highschool	Emp_Highschool	Emp_College	Emp_Degree_Above	Uemp_Less_Highschool	Uemp_Highschool
1992	1.214580	0.105672	-2.573144	-1.656498	1.452880	0.685218
1993	0.390005	0.191049	-2.082459	-1.550855	0.877734	0.364093
1994	0.200139	-0.048740	-1.443474	-1.407799	0.351353	-0.181096
1995	0.059104	-0.177961	-1.091841	-1.262764	-0.033071	-0.518616
1996	0.533610	0.111810	-0.979487	-1.130993	-0.037677	-0.550575

Uemp_College	Uemp_Degree_Above	GDP_percent	GDP_value
-0.144416	-0.576249	0.621013	6.520330e+12
-0.215442	-0.662400	0.111248	6.858560e+12
-0.452905	-0.834351	0.956438	7.287240e+12
-0.624007	-0.853515	0.065837	7.639750e+12
-0.723550	-0.937732	0.786646	8.073120e+12

Then ,we will proceed to do an exploration of our newly constructed dataframe.

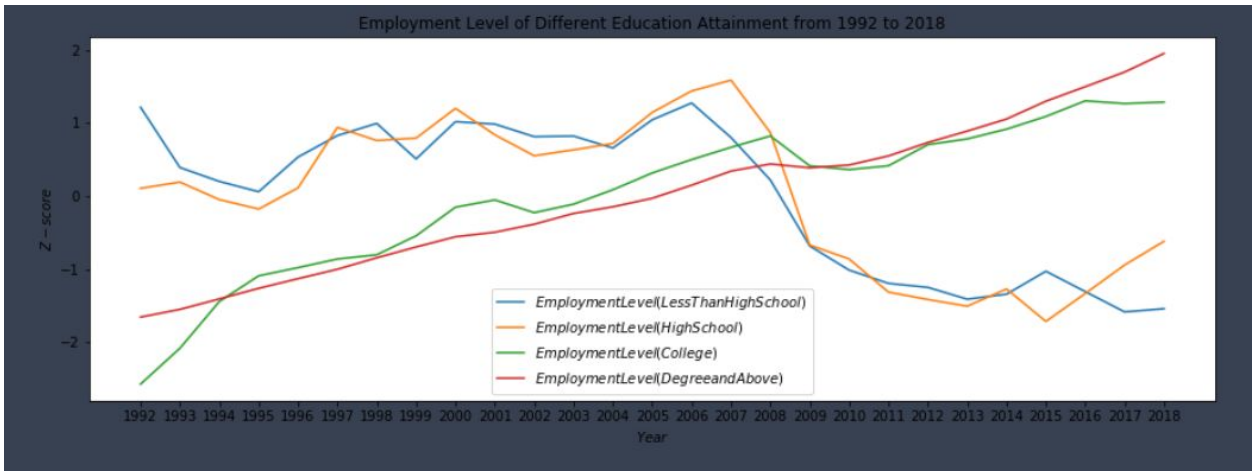
4.0 Exploratory Data Analysis

4.1 Features Correlation Heatmap

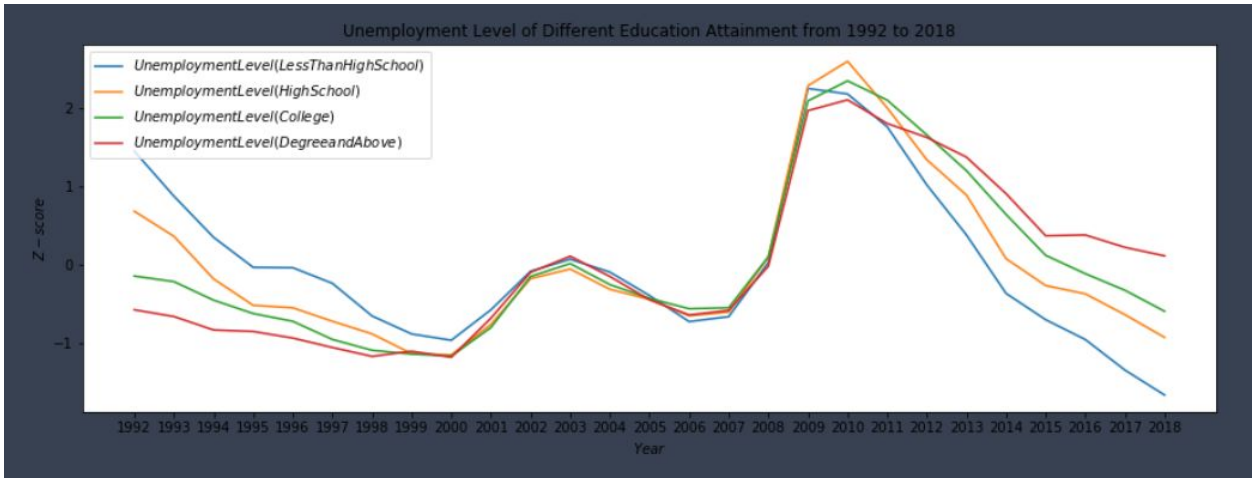


Based on the heatmap above, we can observe that some attributes in the dataset have relatively strong correlation and while some others attribute negative weak correlation. The strong positive correlations include between emp_less_highschool and emp_highschool, between emp_college and emp_degree_above. There's also a medium to high strength correlation between the attributes representing the unemployment rate; unemp_less_highschool, unemp_highschool, unemp_college, unemp_degree_above.

4.2 Employment and Unemployment Level of Different Education Attainment

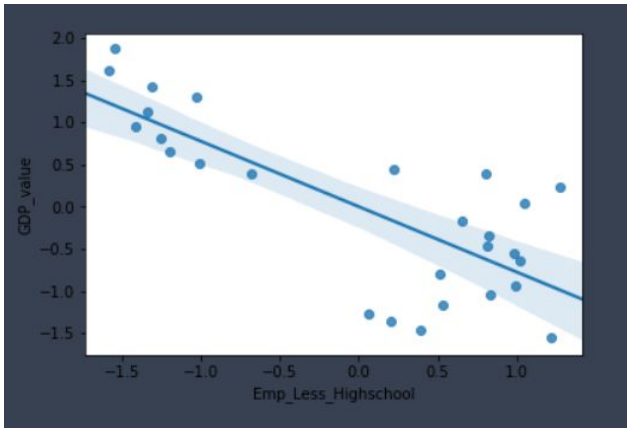


Upon studying the above line graph, we can deduce that from 1992 to 2018, the Employment level of Degree and above has been in a steady rise. For Employment level at College, it rises sharply in between the period of 1992 to 1995 and continues to rise until it has a dip in between 2008 and 2009, and continues to rise again. The employment level of education level less than high school goes up and downs, which can be considered unstable from 1992 until the period between 2006 and 2007, it begins to plummet until 2013. The employment level of education level less than high school began to plummet again in 2015 while the employment level of education level at high school rose sharply. The Employment level of college and also for degree and above plummet a little around from 2008 until 2010 and then started to rise steadily again.

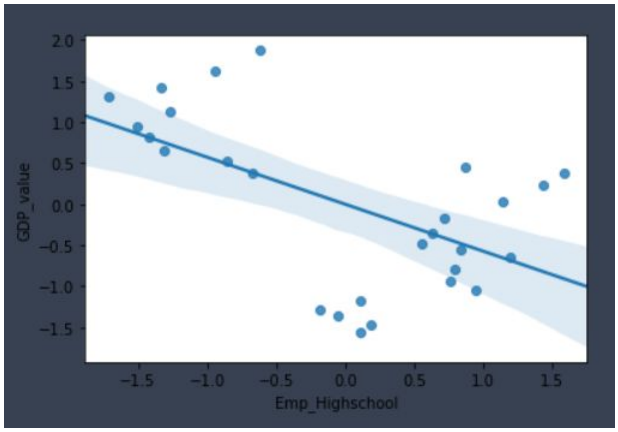


This graph shows us that the attributes follow a similar pattern in the context of unemployment rate.

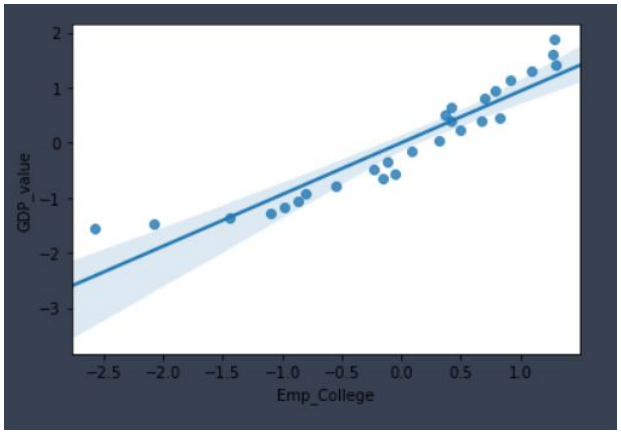
4.3 Correlation of employment level with the value of the GDP



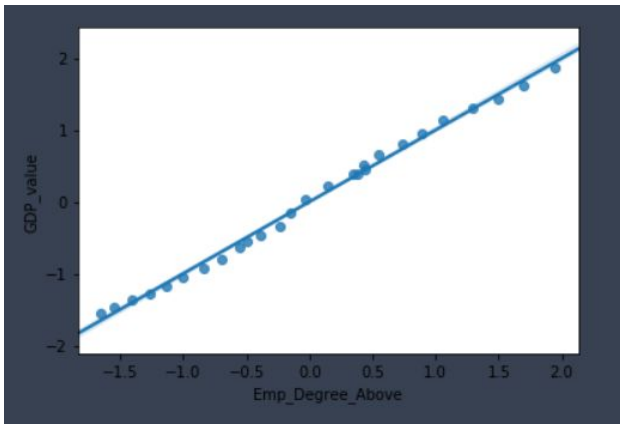
The correlation between the emp_less_highschool and GDP_value has weak to medium strength negative correlation.



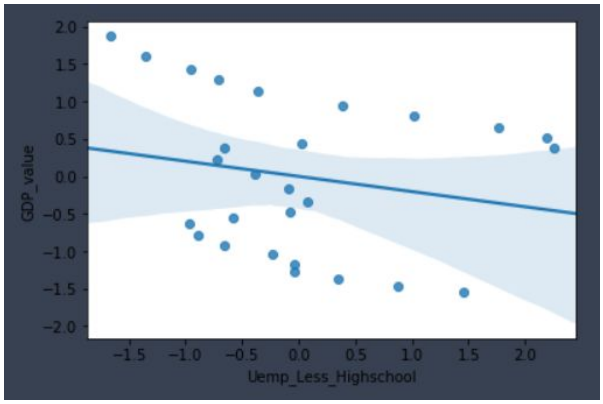
The correlation between the emp_highschool and GDP_value has weak to medium strength negative correlation.



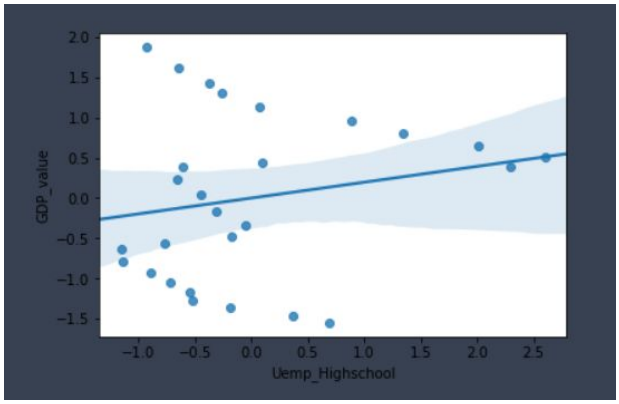
The correlation between the emp_college and GDP_value has medium to strong strength positive correlation.



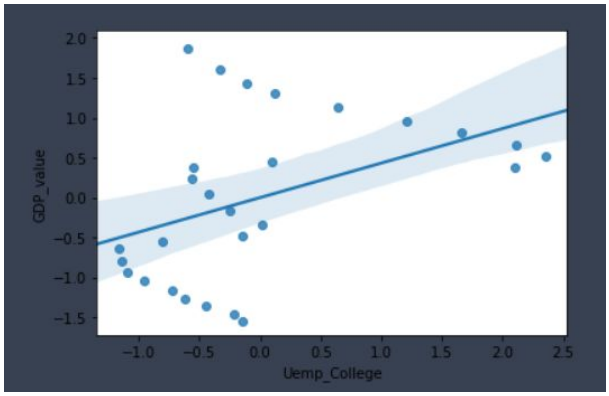
The correlation between the emp_degree_above and GDP_value has strong strength positive correlation.



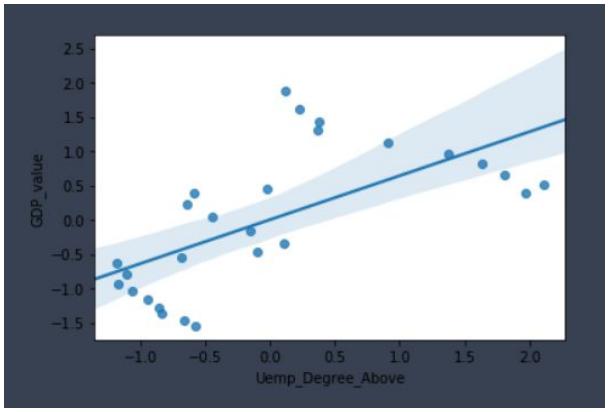
The correlation between the uemp_less_highschool and GDP_value has weak strength negative correlation.



The correlation between the uemp_highschool and GDP_value has weak strength positive correlation.

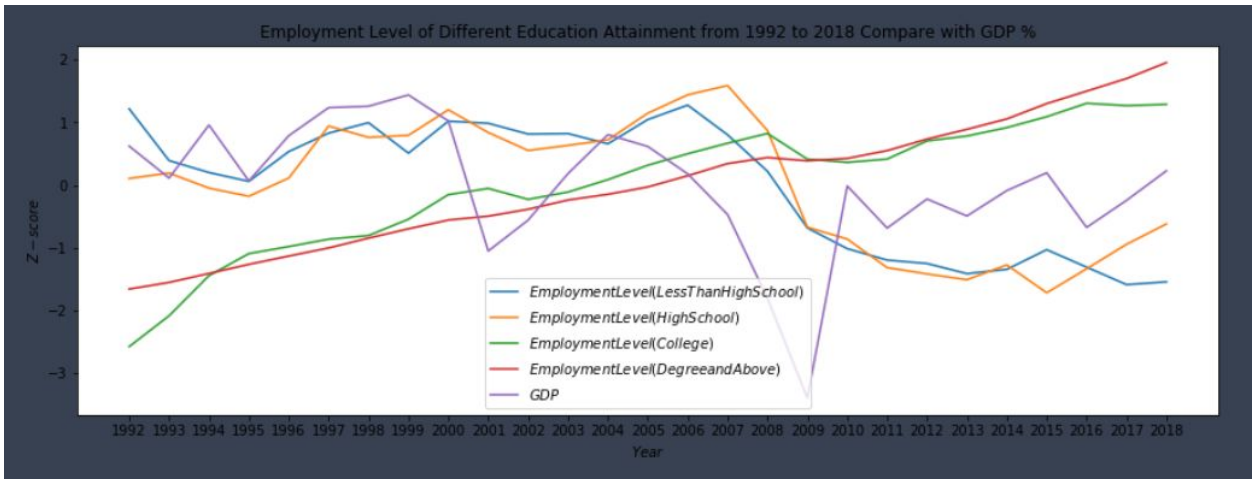


The correlation between the uemp_college and GDP_value has weak to medium strength positive correlation.

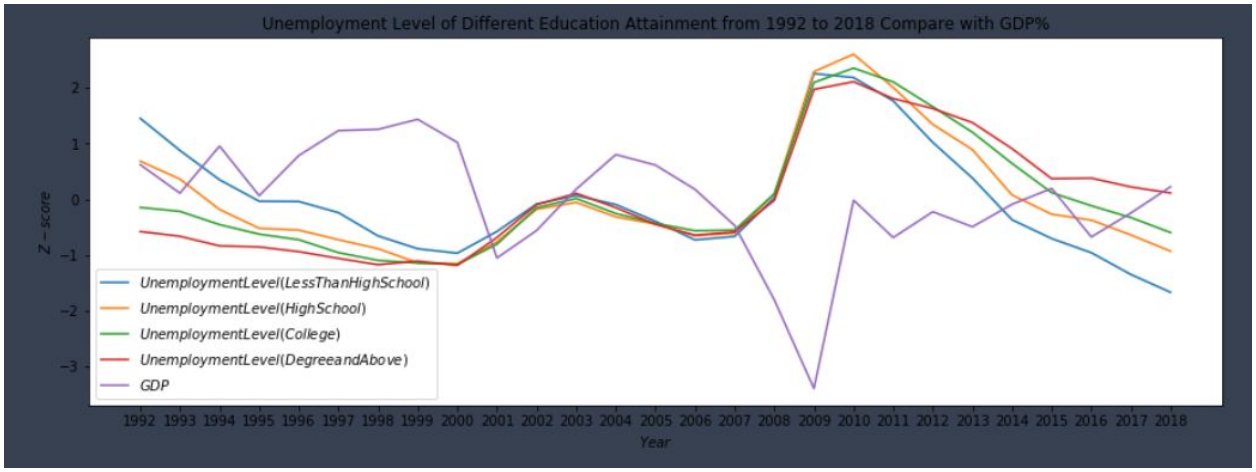


The correlation between the uemp_degree_above and GDP_value has weak to medium strength positive correlation.

4.4 Employment and Unemployment Level of Different Education Attainment Compared with GDP%



From 2000 to 2002, there was a slight plummeting of Employment Level for education less than high school, high school and college , which the GDP% plummeted badly too. For the period of 2005 until 2009, there is a big plummet for Employment Level for education less than high school and high school and slight plummet of Employment Level for education of college and degree and above, there is a big plummet in GDP% too.



In between 2000 until 2001, there's a plummet of GDP% which is a slight rise in the unemployment rate of all the education levels. Between 2007 and 2009, there was a high rise of the unemployment rate of all the education levels and there's a great plummet of GDP%.

5.0 Statistical Modelling

5.1 Dataset in Use

Our dataset (27,10) consists of the following columns:

1. Year - 1992 to 2018
2. Emp_Less_Highschool - Employment level of education level lower than highschool
3. Emp_Highschool - Employment level of education level in highschool
4. Emp_College - Employment level of education level in college
5. Emp_Degree_Above - Employment level of education level higher than degree
6. Uemp_Less_Highschool - Unemployment level of education level lower than highschool
7. Uemp_Highschool - Unemployment level of education level in highschool
8. Uemp_College - Unemployment level of education level in college
9. Uemp_Degree_Above - Unemployment level of education level higher than degree
10. GDP_value - Value of Gross Domestic Product

The attribute of GDP_percent has been removed in order for us to predict the value of the GDP value as the growth rate of GDP will not be able to be derived without the GDP value as the growth rate is derived from the GDP value.

5.2 Modelling

We have chosen linear regression as the algorithm to create the predictive model. We use the LinearRegression library from Scikit Learn to create the model. We will first split the data intMo training sets and testing sets.

```
X = merge_GDP_all_empt_umempt.drop(['GDP_value','GDP_percent'], axis = 1 )
y = merge_GDP_all_empt_umempt['GDP_value'].copy()

from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 1)
```

Then, we get the score of how fit the model is with our data.

```
reg = LinearRegression().fit(X_train, y_train)

reg.score(X_test, y_test)
```

```
0.9976879084630039
```

5.3 Model Validation

After training the model using the training data, we will validate our data using the test data.

```
some_data = X_test.iloc[:5]    # take first 5 entries from X_test
predicted_gdp = reg.predict(some_data)
predicted_gdp
```

```
array([1.45996136e+13, 1.36760569e+13, 1.48534408e+13, 7.59486797e+12,
       1.84287812e+13])
```

We use the y_test data as the 'ground truth' to see what is the true data. From what we can see, the true values and predicted values are quite similar.

```
actual_gdp = y_test.iloc[:5].values
actual_gdp
```

```
array([1.44489e+13, 1.38146e+13, 1.49921e+13, 7.63975e+12, 1.82193e+13])
```

Even though the results are quite similar, we will still apply the Mean Absolute Error and Mean Squared Error as the performance metrics to evaluate our model's performance.

5.3.1 Mean Absolute Error

```
from sklearn.metrics import mean_absolute_error

y_prediction = reg.predict(X_test)
model_mae = mean_absolute_error(y_test, y_prediction)
print(model_mae)
```

```
139148255282.06625
```

Value: 139148255282.06625

The value suggests that on average, the prediction is off by ±139148255282.06625.

5.3.2 Mean Squared Error

```
from sklearn.metrics import mean_squared_error
model_mse = mean_squared_error(y_test, y_prediction)
print(model_mse)
```

```
2.4194187518025832e+22
```

Value: 2.4194187518025832e+22

6.0 Closing notes

6.1 Answering the Question

Undoubtedly, the insights generated through our exploratory data analysis has helped answer the questions mentioned earlier. We have shown that, for example, that different education levels tend to have different patterns while comparing with each other.

Meanwhile, we have also developed a predictive model that can help predict the GDP value given the employment and unemployment levels with different levels of attainment of education.

6.2 Challenges We Faced

Throughout the assignment, we have learned more than we have known previously about Data Science and Machine Learning. The process of the data science pipeline has already proved to be a challenging task. The dataset used in this case is very challenging to find as there are not many similar datasets that are usable for our case.

Besides, another challenge posed is the amount of data that is available for us to find. For example, the dataset for the employment and unemployment level based on the education levels contains only data starting from the year 1992. As there is only a small number of the data available, our predictive model might not be a good model as there is too little data to be trained with.

Lastly and most importantly, as we are relatively new in this field and thus we are greatly inexperienced; we might not have the best method or way to approach or to “play” with the data, and this has greatly limited us in the scope of what we are able to accomplish.

6.3 Way forward for Future Insights

With increase in per capita income, the incidence of poverty often goes up. GDP value can yet have a very close relationship with the income of the people of the country. For example, India has achieved a satisfactory growth rate in recent years, but the planners have failed to alleviate poverty in 56 years. Data on the income of the people in a country can be mined to better understand the behaviour of the GDP.

7.0 Conclusion

In part A, it has been a very hard time for us to come up with an idea on what type of sector to dive into and what kind of question should we formulate as we are much inexperienced in the field of Data Science.

In part B, we spend a considerable amount of time to search for the suitable datasets and were forced to change our questions a few times as we could not gauge what type of questions are appropriate to be asked as we cannot find the relevant public datasets, or our skill were not sufficient to derive further data from the existing datasets.

All in all, it is a great experience to be able to have an assignment like this as we have learned a lot of practical skills from the process, for example: the data science pipeline. Before the assignment, we have only known the pipeline process theoretically, but with the hands-on assignment, we finally realised the importance of the pipeline process as it acts as a guide for use to explore and modelling data.

Reference

1. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=US
   GDPcountries.csv

2. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=US
   GDPvalue.csv

3. https://www.bls.gov/webapps/legacy/cpsatab4.htm
   Employment_level_less_highschool.csv
   Employment_level_highschool.csv
   Employment_level_college.csv
   Employment_level_degree_above.csv
   Unemployment_level_less_highschool.csv
   Unemployment_level_highschool.csv
   Unemployment_level_college.csv
   Unemployment_level_degree_above.csv

4. https://www.youtube.com/watch?v=bs2q0oFfxX4
   Z-Score Normalization Tutorial

5. https://bizfluent.com/info-8296076-importance-employment-workplace-society.html
   The Importance of Employment & Workplace in the Society

6. http://www.economicsdiscussion.net/gdp/6-main-factors-affecting-gdp/15344
   6 Main Factors Affecting GDP