National College of Ireland

Project Submission Sheet –2020

School of Computing

Student no	Name
19183461	Pallavi Kale

Programme:	Higher Diploma in Science in Data Analytics	**Year:**	2020
Module:	Programming for Big Data		
Lecturer:	Symeon Charalabides		
Submission Due Date:	26th April 2020		
Project Title:	Data Analysis Project		
Word Count:	~		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

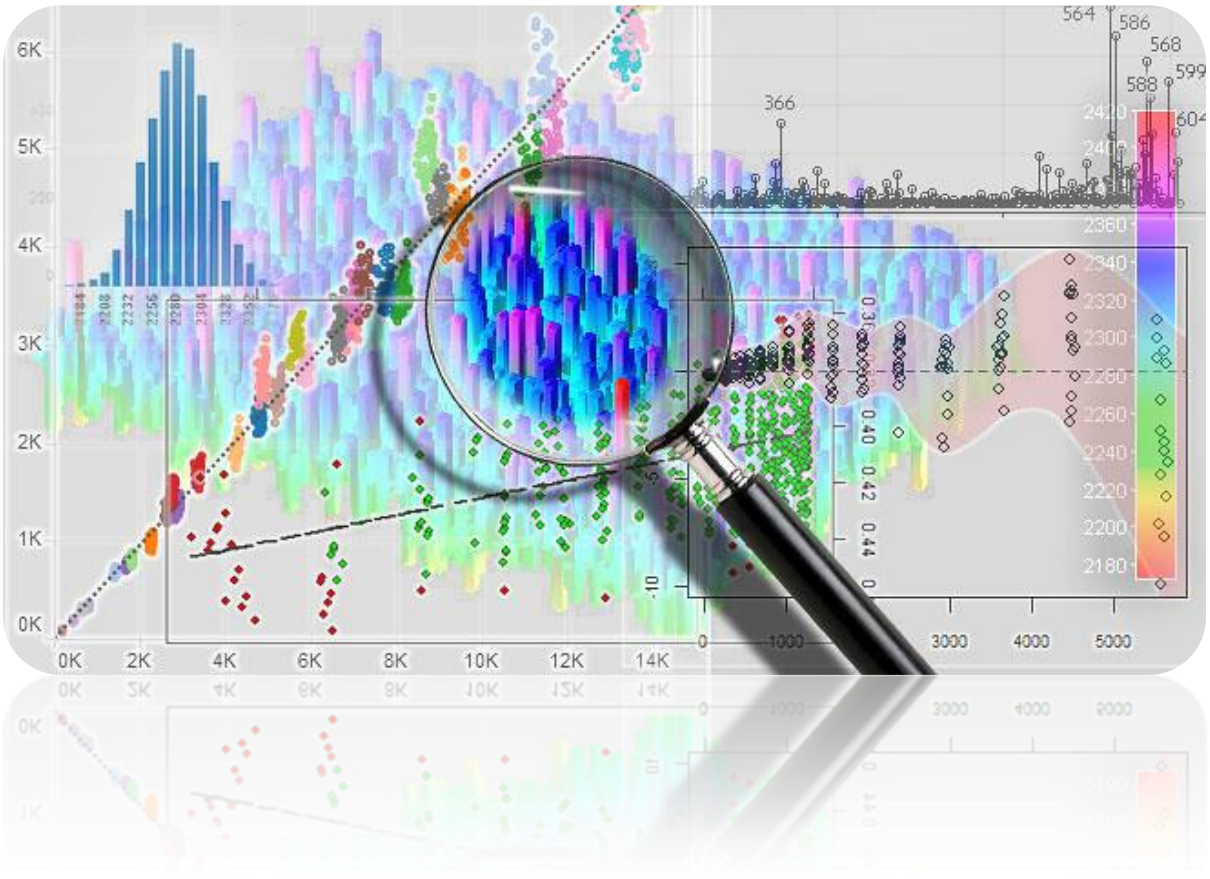
Signature: *Pallavi Kale*

Date: *3rd April 2020*

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	



Project Report

1) Correlation Analysis of GDP and Number of flights at Irish Airports

2) Correlation between variability of rainfall and production of crops.

Assignment CA 2

Table of Contents

List of Abbreviations

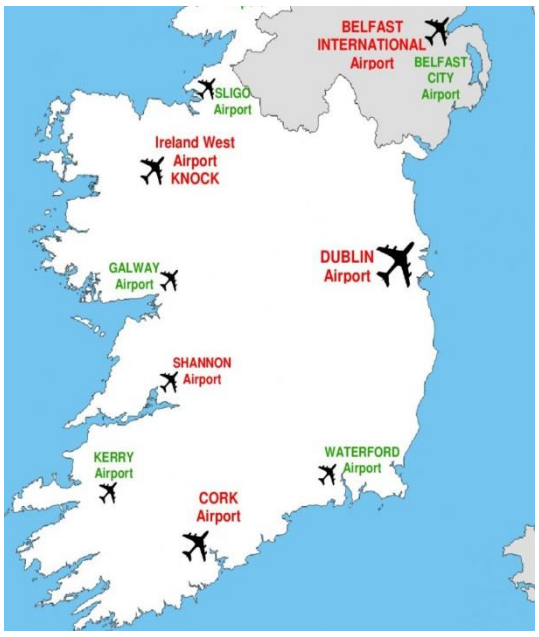
R	Programming language used for statistical computing and graphics.
IT	Information Technology
BI	Business Intelligence
MS Excel	Microsoft Excel
OECD	Organization for Economic Co-operation and Development
CSO	Central Statistics office
FAO	UN Food and Agriculture Organization
IATA	The International Air Transport Association
CSV	Comma Separated Values
GDP	Gross domestic product
Cor	Correlation
H_0	Null hypothesis
H_a	Alternative hypothesis

1. Introduction

There has been a major shift in travelling patterns across world. With the rise in GDP (Gross Domestic Product) across continents more and more people are travelling across the Globe. GDP is the key indicator of countries economic health. It is used to measure the value of services and finished goods within the domestic market.

Increase in person's income or country as a whole bring new opportunities to travel for business or pleasure. Travelling through flights can be an indicator of the aspirations and increase in monetary value in a country.

1.1. Objectives



The main objective of the study is to find if there is any correlation between rise in GDP and increase in flights to Ireland. We will base our hypothesis on the fact that there is a correlation. To achieve this objective we will conduct series of statistical analysis on the model we will develop and produce inferences based on analysis.

Since the coronavirus pandemic however there has been a major shift in travel movements. But this is not part of our study. We aim to see if there is rise in GDP is there an increase in number of flights as well.

Irish airports have seen an increase in passengers and flights. This can be because of Ireland's tourist promotions and open market for IT companies . As per Trip advisor Ireland is also gaining popularity in being a tourist destination. Dublin and Kerry being most popular destinations.

Before we start our analysis it is important to do a domain research on the subject at hand. Airports and flights are an indicator of countries progress and development. Numerous studies have been conducted all over the world which analyse this in great depth and detail. We will try to find out specific studies in relation to GDP and Flights.

1.2. Literature Review

GDP Growth is a key Driver of Air Travel Demand

The author discuss about the increase in GDP tends to increase disposable income and people purchase more tickets. Also how strong is correlation between GDP growth and air travel? It discusses a previously done research by International Air Transport Association, where there is extreme relation between countries like Uganda less than 1% population flies once a year whereas Norway and Switzerland average above two flights per capita per year. This means that every individual flies at least 2 times in a year.
A graph of air travel among all countries show that higher income group countries travel more in relation to GDP whereas developing countries like Brazil ,China or Russia have 0.5 Flights per Capita. A projection by IATA expects it to Double by 2035.It further discusses Germany as growth of passengers matches GDP per capita.
Similar trends were found for UK or Poland. For China an increase in air passengers by 105% in the year 2010 while GDP grew by 111%.
There were counter examples where number of passengers lag behind by economic growth may be a result of lack of infrastructure. It also discuss about limitation in the assumptions as rise of GDP will not suffice to increase passengers. There are other external forces and the chosen dataset for the period may change the results significantly. (Ciesluk K.(2019))

Flash analysis - The correlation between GDP growth and the increase in airline passengers 2001-2016

It discusses GDP growth being a driving force behind air passenger traffic. Also states that number of passengers around world has grown 1.5 times that of world GDP. For a period of 2001-15 there is 3.9% annual growth of passengers to 3.8% GDP growth, which is 1.0X for the correlation between GDP and passengers. It further calculates GDP multiplier and taking Regions and PAX growth. For Asia passenger growth increased by a factor of 09.x , while Africa with rising GDP multiplier of 0.8x had less passengers due to political unrest.
Europe had increase of 2.0x because of low cost carriers like Ryanair and Easyjet and increase in passengers from Asian countries.
It discuss various regions, there multipliers and growth factors. Also National passenger growth increase with increase in domestic structure.UK grew at 1.5x faster than the GDP and Finland grew by 3.5x.Finland new International hub was also accounted for growth in

passengers. It concludes that passenger growth has remained stable even after weakening economies. But also discuss that there are distinct differences as a result of National policies, infrastructure and political unrest. (KFW IPEX Bank (2017))

1.3. Dataset description

Central Statistics office database for Aviation data on transport was selected .The data provides information on Commercial Flights arrived in Ireland from 2013 till 2019 yearly. Following website was acessed , here data is free to download for public study purposes: https://statbank.cso.ie/px/pxeirestat/Database/eirestat/Aviation%20Statistics/Aviation%20Statistics_statbank.asp?SP=Aviation%20Statistics&Planguage=0

The Dataset has many three statistical Indicators:
Passengers, Freight and Commercial Flights. Passengers and Commercial Flights were chosen. Under direction category Arrival flights are chosen as that is objective of the study. Under Flight type only Scheduled flights were selected.
The below table shows the variables of the datasets.

DestinationCity	SourceCountries	Category	2013	2014	2015	2016	2017	2018	2019
All main airports	Europe(2)	Passengers (Thousand)	10504.4	11318.6	12704	14076.9	14590	15273.3	16046.3
All main airports	Europe(2)	Commercial Flights (Thousand)	89.7	96	100.5	109.8	111.7	116.2	120.4

Table 1 – Yearly Arrival of Flights and Passengers to Ireland Dataset

Yearly Passenger and Flights arrival dataset contained following columns.
There were 10 Columns in the dataset.

- First column contains cities in Ireland where the Airport is located. It's a character column. It contains for OECD group of countries
- Source Countries are the countries from where flights and passengers are arriving. It's a character column.
- Category column contains two factors Passengers (Thousand) and Commercial Flights (Thousand).
- Rest of the columns are years and number of passenger flights each year at every airport. Only scheduled flights were selected. All the columns are used in analysis. There are about 709 rows in this dataset.

Variable like DestinationCity , SourceCountries and Category Column are factors

Whereas all the years in numbers.

Columns like DestinationCity, SourceCountries have been added because the dataset did lack proper segregation between countries and cities to use in R meaningfully. This was done on Excel sheet so that data can be accessed properly.

Home Statistics Databases Methods About Us

You are here › Home / StatBank / Aviation Statistics / TAA03 / Select from table TAA03

TAA03: Passengers, Freight and Commercial Flights by Airports in Ireland, Country, Direction, Flight Type, Year and Statistic

Unit : Thousand

Download file as... Comma Separated (*.csv) | **Edit table** Pivot | **Graphics** Line Chart | Sort table | Print

☐ Codes in separate columns

☑ Incl. Footnotes

Passengers, Freight and Commercial Flights by Flight Type, Direction, Airports in Ireland, Country, Statistical Indicator and Year

	2013	2014	2015	2016	2017	2018	2019
Scheduled							
Arrival							
All main airports							
All Countries							
Passengers (Thousand)	11,900.9	12,920.3	14,544.4	16,103.6	16,948.7	17,994.3	18,768.5
Commercial Flights (Thousand)	99.1	106.6	112.1	121.9	125.6	131.8	135.0
Ireland (domestic)							
Passengers (Thousand)	46.2	50.8	60.8	68.0	70.2	80.6	79.4
Commercial Flights (Thousand)	3.1	3.1	2.8	2.7	2.7	2.7	2.7
Europe (2)							
Passengers (Thousand)	10,504.4	11,318.6	12,704.0	14,076.9	14,590.0	15,273.3	16,046.3
Commercial Flights (Thousand)	89.7	96.0	100.5	109.8	111.7	116.2	120.4
EU28 excluding Ireland							
Passengers (Thousand)	10,211.7	11,008.4	12,341.7	13,670.0	14,134.1	14,727.2	15,457.8
Commercial Flights (Thousand)	87.4	93.4	97.4	106.6	108.1	112.0	115.8
Austria							
Passengers (Thousand)	50.6	47.2	48.5	53.9	55.5	73.5	118.7
Commercial Flights (Thousand)	0.4	0.4	0.4	0.4	0.4	0.5	0.8
Belgium							
Passengers (Thousand)	210.8	209.9	269.9	251.9	253.0	254.5	253.5
Commercial Flights (Thousand)	1.9	1.8	2.2	2.1	1.9	1.9	1.9
Bulgaria							
Passengers (Thousand)	11.1	10.9	12.1	17.9	43.9	47.3	48.4

Figure 1 - Flight Arrival Datasets

For GDP data, Organization for Economic Co-operation and Development (OECD) Website was used which provide GDP for OECD group of countries. Over the period of time number of countries participating in OECD has reduced from 53 to 39 countries.
https://data.oecd.org/gdp/gross-domestic-product-gdp.htm#indicator-chart

LOCATION	INDICATOR	SUBJECT	MEASURE	FREQUENCY	TIME	Value
AUS	GDP	TOT	MLN_USD	A	1960	25034.74
AUS	GDP	TOT	MLN_USD	A	1961	25326.38

Table 2 – GDP Datasets Description

This dataset contains seven variables. There are 4391 rows. All the columns are used except frequency and subject fields.

- Location Column is the list of names of countries for which GDP is given. This is ISO3 country code .This was later matched with country codes to get a full name of countries. It's a factor.
- Indicator column is a factor variable with GDP as the name.
- Subject Field is a factor with total GDP.
- Measure Field is the factor with GDP measurement in Million USD and USD per capita.
- Frequency Field is a variable is a factor with A.
- Time Field has a value in number with years starting from 1960 till 2019.

- Value field is a numerical column with GDP in Millions.

As the CSO website gave only ISO3 country codes, another dataset for Full name of Countries was referred. This was downloaded at https://github.com/datasets/country-codes

official_name_en	ISO3166-1-Alpha-3
Tiwan	TWN

Table 3 – Country Code Datasets Description

Contains two columns official_name_en and ISO3166-1-Alpha-3 both are factors one is full country name and other is ISO3 three alphabet code.

1.4. Analysis approach

The aim of this study is to a find a correlation between GDP and flights arrivals to Ireland. To achieve this we will need to formulate a methodology which will help in reaching our desired objectives.

Cross-industry process for data mining is the correct methodology to achieve our desired goals.

IN CRISP-DM model the whole set of process is divided in six stages. Business Understanding, Data Understanding, Data Preparation, Modelling Evaluation and Deployment.

As the sequence of phases is not rigid it allows us to

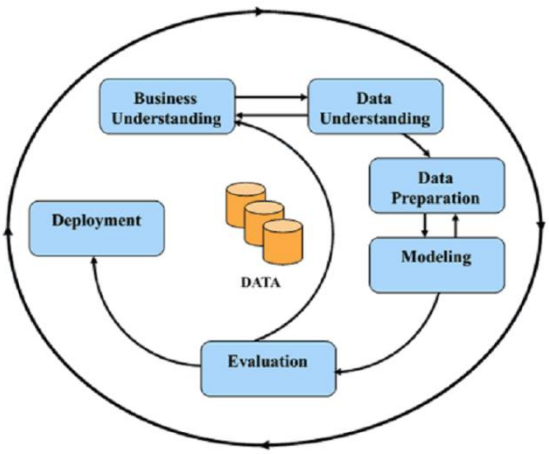
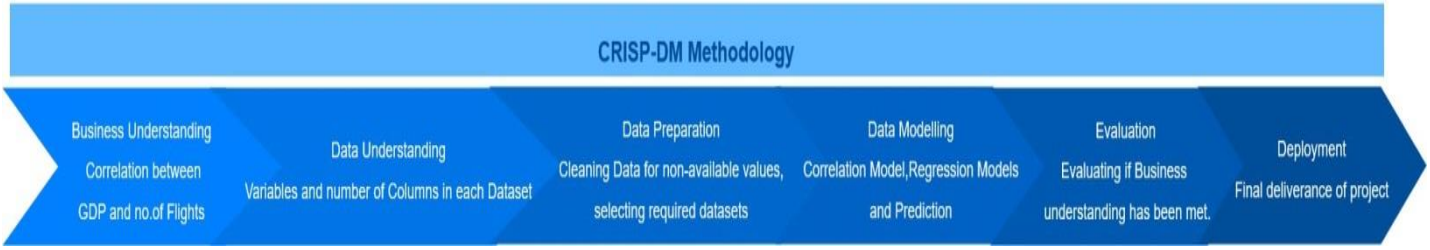


Figure 2 - CRISP-DM

go and do modifications if required at any stage.

In our analysis we have achieved the stage as follows:

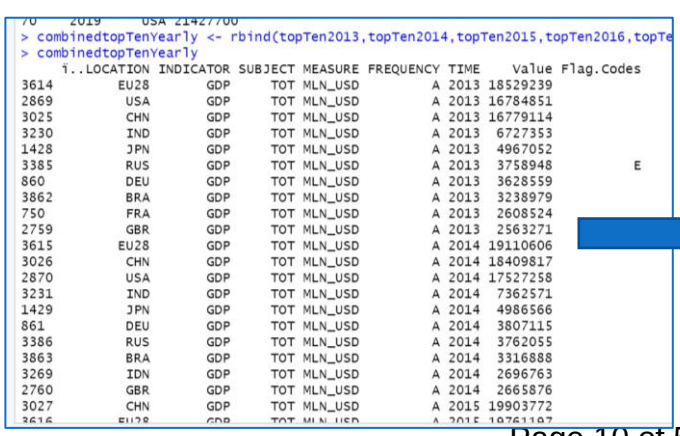
1. **Business understanding –** This stage focuses on objectives and if there are any



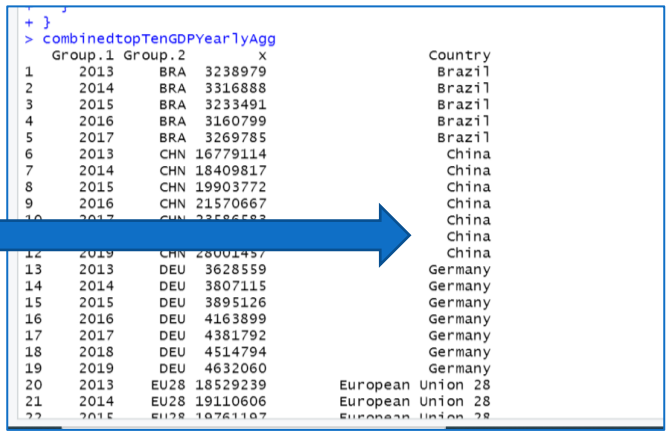
business requirements. We have fixed our set of objectives that GDP and number of Flights to Ireland are correlated and this will form the basis of our hypothesis.

2. **Data Understanding –**Second stage deals gathering right data to meet our business requirements .Wrong data would not give us desired objectives. Gaining information on the datasets. The accurate data is needed to perform statistical analysis. We sourced our data from three different websites. CSO,OECD and GIthub. We have to see if data can help us accomplish the desired objectives, we will need countries their GDP and Flight arrivals to Ireland. After performing basic analysis the data we have a raw data to do computing with.

3. **Data preparation –** This stage involves preparing data which is ready to be modelled. The dataset gathered from CSO website needed some manual corrections as few columns were ambiguous. After processing the sheet the thigh GDP performing ten countries were selected from each year. These top economies were then aggregated to for each year to see if they were constantly performing countries. Since it only contained ISO 3 codes it was further processed to get full country name from a second file to generate a model worth of merging with Flight arrival data.



```
> combinedtopTenYearly <- rbind(topTen2013,topTen2014,topTen2015,topTen2016,topTe
> combinedtopTenYearly
     i..LOCATION INDICATOR SUBJECT MEASURE FREQUENCY TIME    Value Flag.Codes
3614        EU28       GDP     TOT MLN_USD         A 2013 18529239
2869         USA       GDP     TOT MLN_USD         A 2013 16784851
3025         CHN       GDP     TOT MLN_USD         A 2013 16779114
3230         IND       GDP     TOT MLN_USD         A 2013  6727353
1428         JPN       GDP     TOT MLN_USD         A 2013  4967052
3385         RUS       GDP     TOT MLN_USD         A 2013  3758948          E
860          DEU       GDP     TOT MLN_USD         A 2013  3628559
3862         BRA       GDP     TOT MLN_USD         A 2013  3238979
750          FRA       GDP     TOT MLN_USD         A 2013  2608524
2759         GBR       GDP     TOT MLN_USD         A 2013  2563271
3615        EU28       GDP     TOT MLN_USD         A 2014 19110606
3026         CHN       GDP     TOT MLN_USD         A 2014 18409817
2870         USA       GDP     TOT MLN_USD         A 2014 17527258
3231         IND       GDP     TOT MLN_USD         A 2014  7362571
1429         JPN       GDP     TOT MLN_USD         A 2014  4986566
861          DEU       GDP     TOT MLN_USD         A 2014  3807115
3386         RUS       GDP     TOT MLN_USD         A 2014  3762055
3863         BRA       GDP     TOT MLN_USD         A 2014  3316888
3269         IDN       GDP     TOT MLN_USD         A 2014  2696763
2760         GBR       GDP     TOT MLN_USD         A 2014  2665876
3027         CHN       GDP     TOT MLN_USD         A 2015 19903772
```



```
> combinedtopTenGDPYearlyAgg
   Group.1 Group.2        x                Country
1     2013     BRA  3238979                 Brazil
2     2014     BRA  3316888                 Brazil
3     2015     BRA  3233491                 Brazil
4     2016     BRA  3160799                 Brazil
5     2017     BRA  3269785                 Brazil
6     2013     CHN 16779114                  China
7     2014     CHN 18409817                  China
8     2015     CHN 19903772                  China
9     2016     CHN 21570667                  China
13    2013     DEU  3628559                Germany
14    2014     DEU  3807115                Germany
15    2015     DEU  3895126                Germany
16    2016     DEU  4163899                Germany
17    2017     DEU  4381792                Germany
18    2018     DEU  4514794                Germany
19    2019     DEU  4632060                Germany
20    2013    EU28 18529239 European Union 28
21    2014    EU28 19110606 European Union 28
```

```
> yearlyArrivalFlightsComm
     DestinationCity          SourceCountries                   Category X2013 X2014 X2015 X2016 X2017 X2018 X2019
2   All main airports           All Countries Commercial Flights (Thousand)  99.1 106.6 112.1 121.9 125.6 131.8 135.0
4   All main airports        Ireland(domestic) Commercial Flights (Thousand)   3.1   3.1   2.8   2.7   2.7   2.7   2.7
6   All main airports               Europe(2) Commercial Flights (Thousand)  89.7  96.0 100.5 109.8 111.7 116.2 120.4
8   All main airports        European Union 28 Commercial Flights (Thousand)  87.4  93.4  97.4 106.6 108.1 112.0 115.8
10  All main airports                Austria Commercial Flights (Thousand)   0.4   0.4   0.4   0.4   0.4   0.5   0.8
12  All main airports                Belgium Commercial Flights (Thousand)   1.9   1.8   2.2   2.1   1.9   1.9   1.9
14  All main airports               Bulgaria Commercial Flights (Thousand)   0.1   0.1   0.1   0.1   0.3   0.3   0.3
16  All main airports                Croatia Commercial Flights (Thousand)   0.2   0.2   0.3   0.3   0.4   0.5   0.8
18  All main airports                 Cyprus Commercial Flights (Thousand)   0.0   0.0   0.0   0.0   0.1   0.2   0.1
20  All main airports          CzechRepublic Commercial Flights (Thousand)   0.4   0.6   0.7   0.6   0.6   0.6   0.7
22  All main airports                Denmark Commercial Flights (Thousand)   1.1   1.0   1.2   0.9   0.9   0.9   0.9
24  All main airports                Estonia Commercial Flights (Thousand)   0.1   0.0   0.0   0.1   0.1   0.1   0.1
26  All main airports                Finland Commercial Flights (Thousand)   0.1   0.1   0.3   0.4   0.5   0.5   0.5
28  All main airports                 France Commercial Flights (Thousand)   6.3   7.0   7.4   7.5   7.1   7.3   7.6
30  All main airports                Germany Commercial Flights (Thousand)   5.6   6.2   6.6   7.0   7.7   9.2   9.2
32  All main airports                 Greece Commercial Flights (Thousand)   0.1   0.2   0.2   0.3   0.4   0.4   0.5
34  All main airports                Hungary Commercial Flights (Thousand)   0.5   0.6   0.6   0.7   0.7   0.7   0.8
36  All main airports                  Italy Commercial Flights (Thousand)   3.0   3.2   3.8   3.9   4.0   4.5   5.0
```

Data from Flight arrival is cleaned and is aggregated for each Year based on country of residence and Arrival Airport.

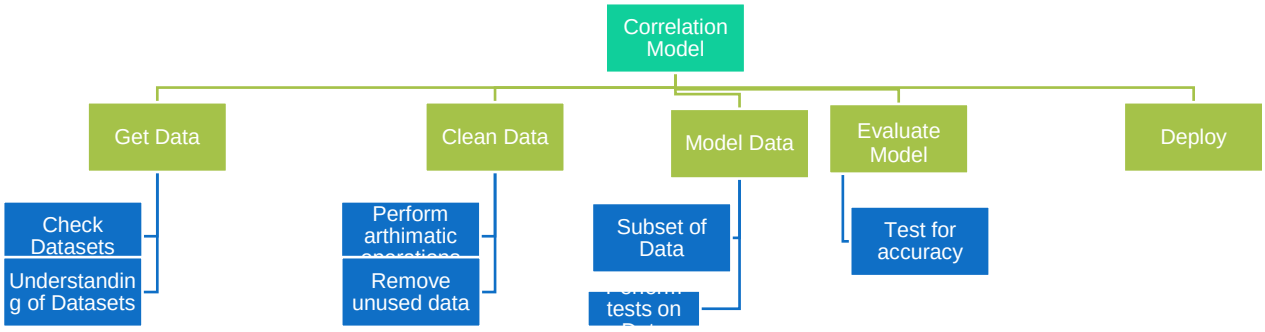
4. ***Modelling-***Two separate models were created one for passengers and other for flights and each one was studied differently. Top ten GDP countries was merged with Flight data for final Modelling.This was further broken down to countries which had available data for all the years.This was because certain countries did not form part of OECD in last 2 or 3 years.This model help us gain insights in to the data.

```
70    2019     USA 21427700 United States of America             8.7
> combinedModelGDPFlight$GDPMillion<-combinedModelGDPFlight$x/1000000
> combinedModelGDPFlight
   Group.1 Group.2        x                Country numberOfFlights GDPMillion
13    2013     DEU  3628559                Germany             5.6   3.628559
14    2014     DEU  3807115                Germany             6.2   3.807115
15    2015     DEU  3895126                Germany             6.6   3.895126
16    2016     DEU  4163899                Germany             7.0   4.163899
17    2017     DEU  4381792                Germany             7.7   4.381792
18    2018     DEU  4514794                Germany             9.2   4.514794
19    2019     DEU  4632060                Germany             9.2   4.632060
20    2013    EU28 18529239 European Union 28             87.4  18.529239
21    2014    EU28 19110606 European Union 28             93.4  19.110606
22    2015    EU28 19761197 European Union 28             97.4  19.761197
23    2016    EU28 20968620 European Union 28            106.6  20.968620
24    2017    EU28 22128645 European Union 28            108.1  22.128645
25    2018    EU28 22941793 European Union 28            112.0  22.941793
```

Before looking for Correlation Model was tested for Shapiro-Wilk normality test for number Of Flights data and GDP data .Later the correlation model was tested. A linear regression model was developed to see if they are correlated and some predictions were done to check if they are correct.

5. Evaluation is done to check if desired objective was met and if did find indication that they are correlated.

6. Deployment In this case a deployment would be the deployment of project which does the study on correlation.

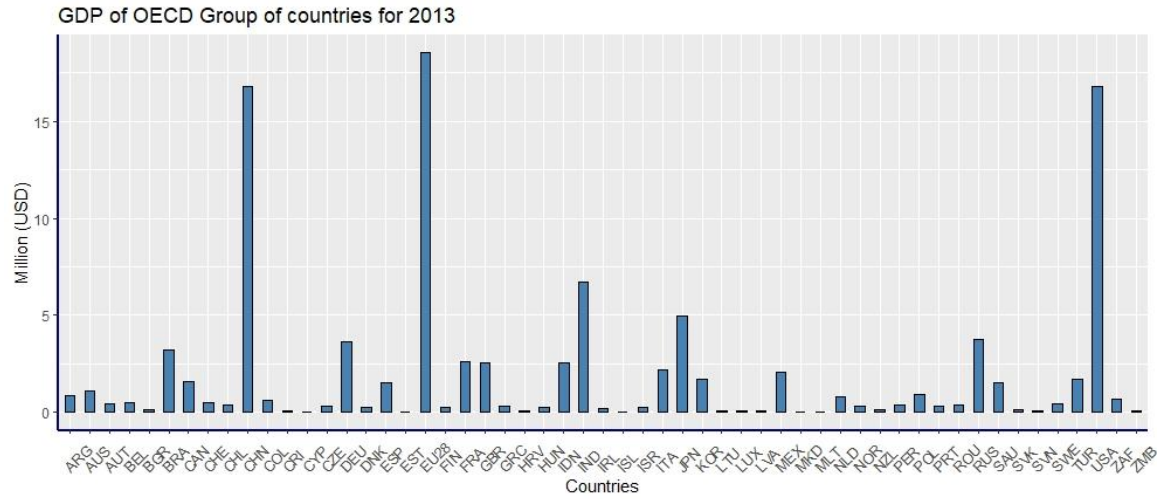


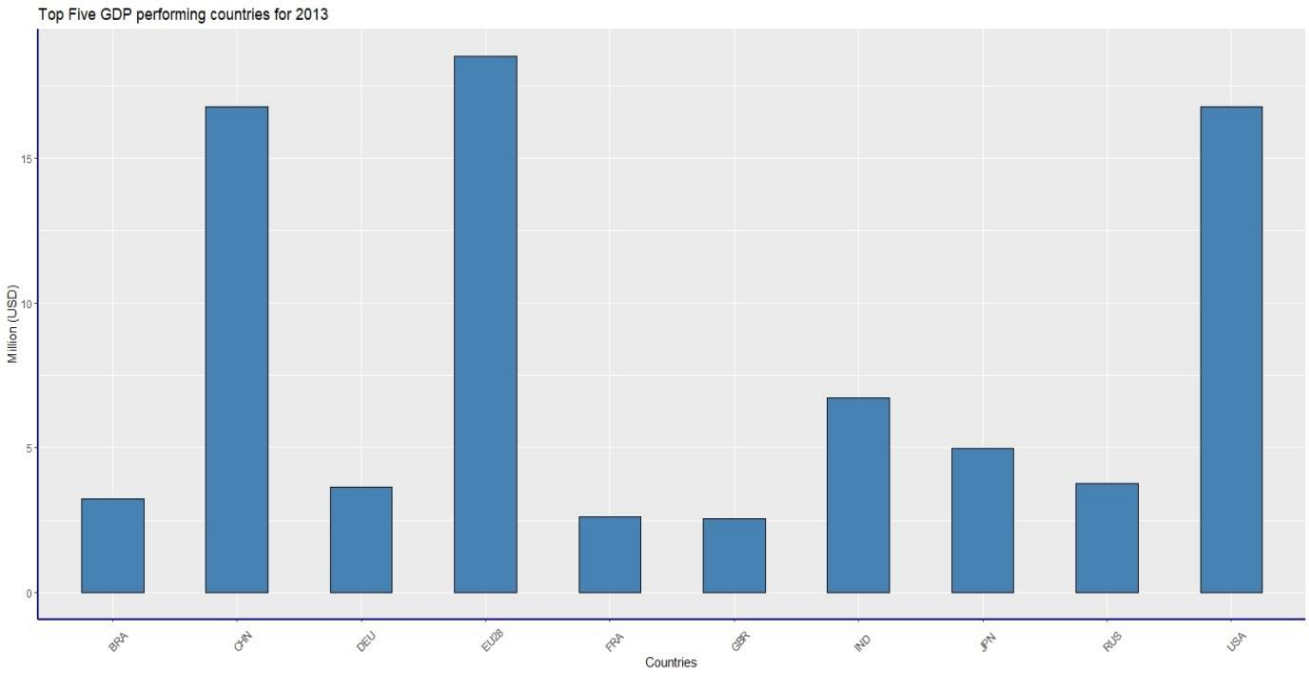
1.5. Analysis results and presentation

As a part of the task it is required to present four unique insights in to the datasets.

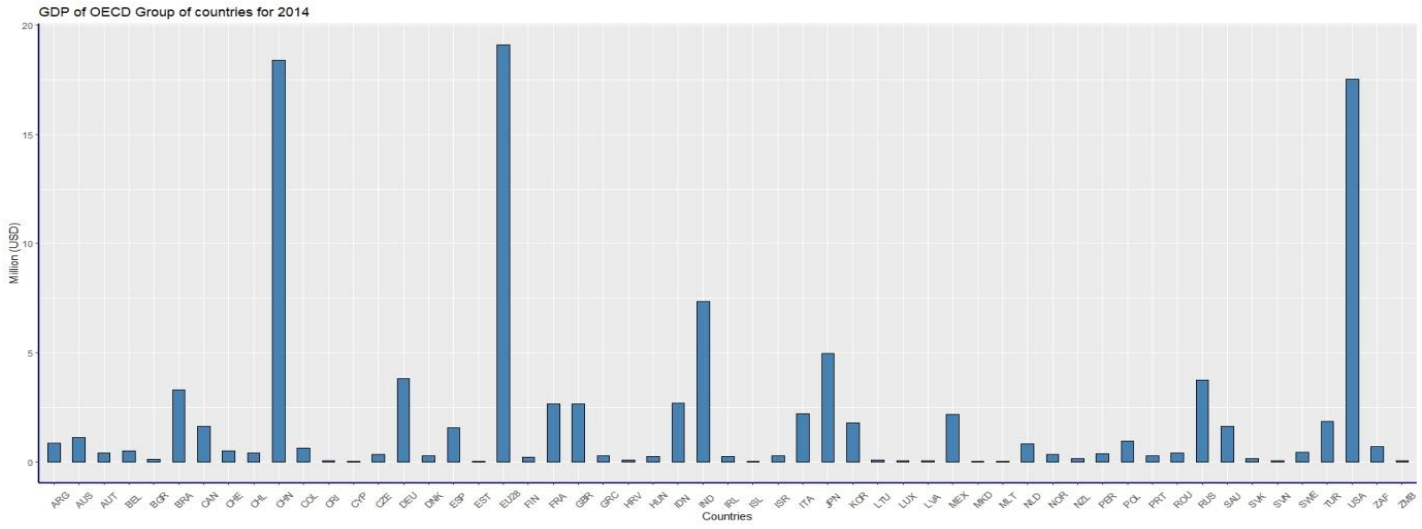
After the drop in economy in 2008 there was a slowdown in all over the countries. However this is not part of our discussion. We have datasets available from 2013 till 2019 on Flights arrival. We will do a series of discussion from the analysis we have done on the datasets.

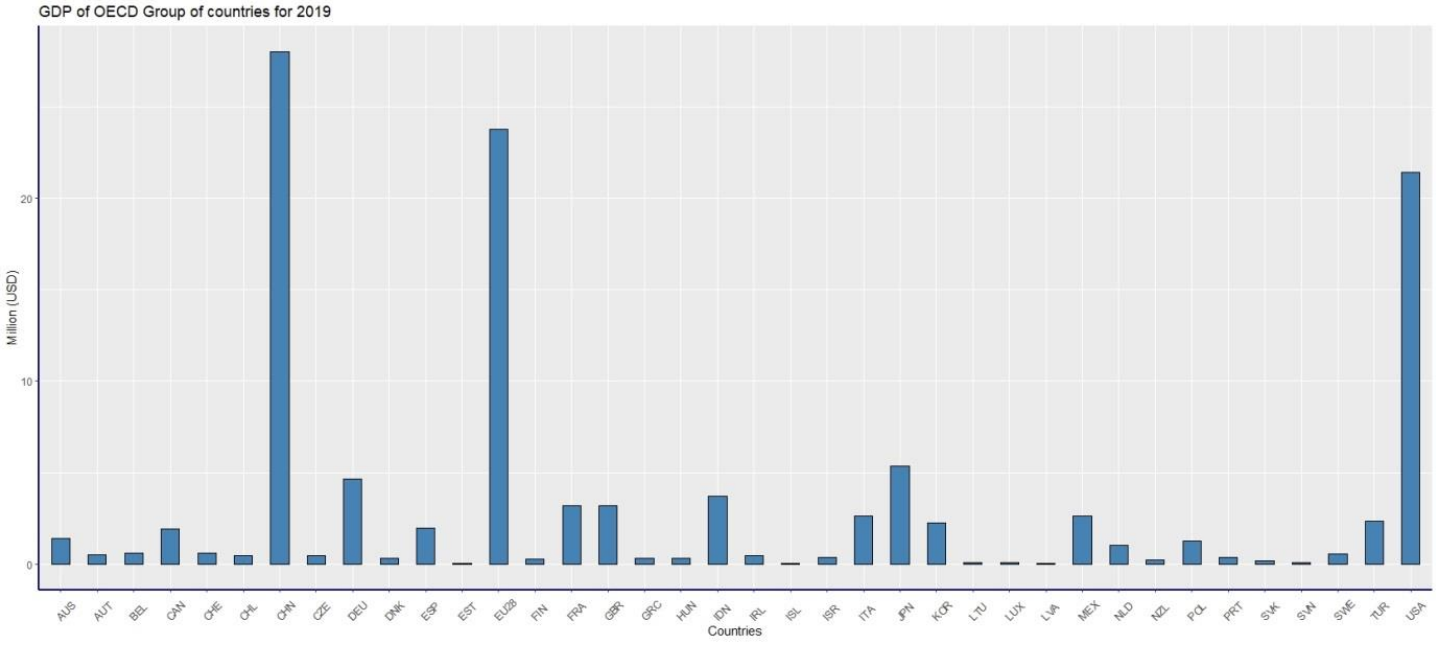
By doing a basic analysis on group of countries it can be seen that China,EU28, USA, Russia, India are top performing countries

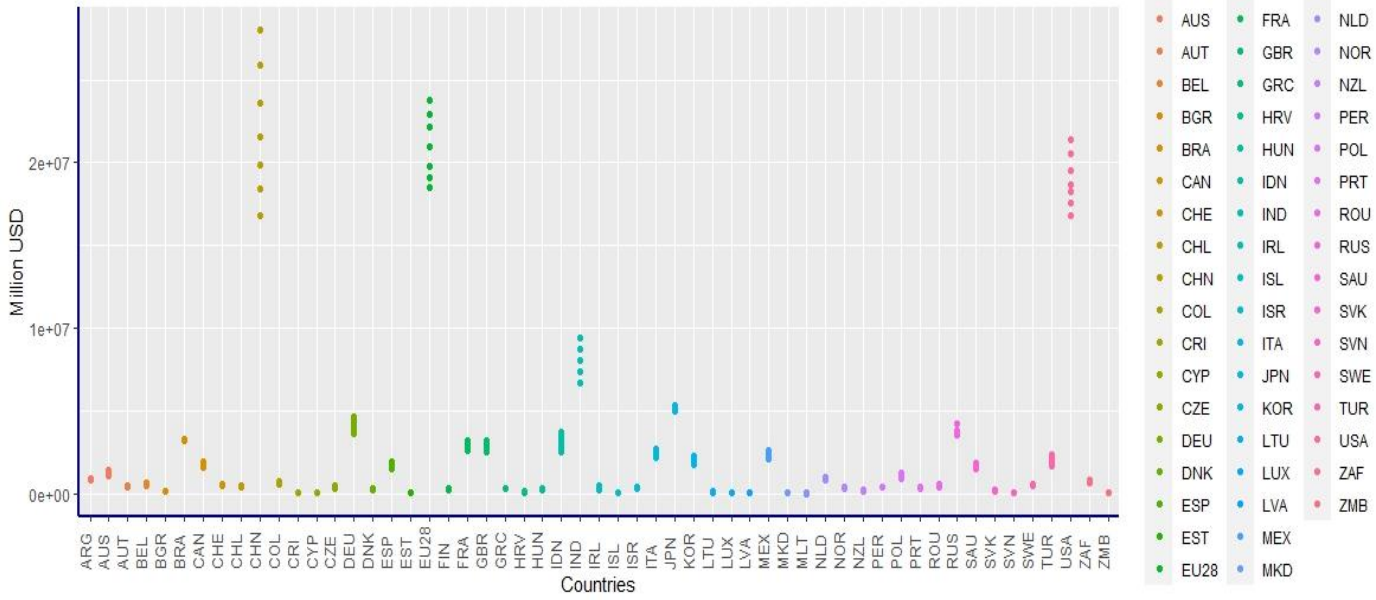


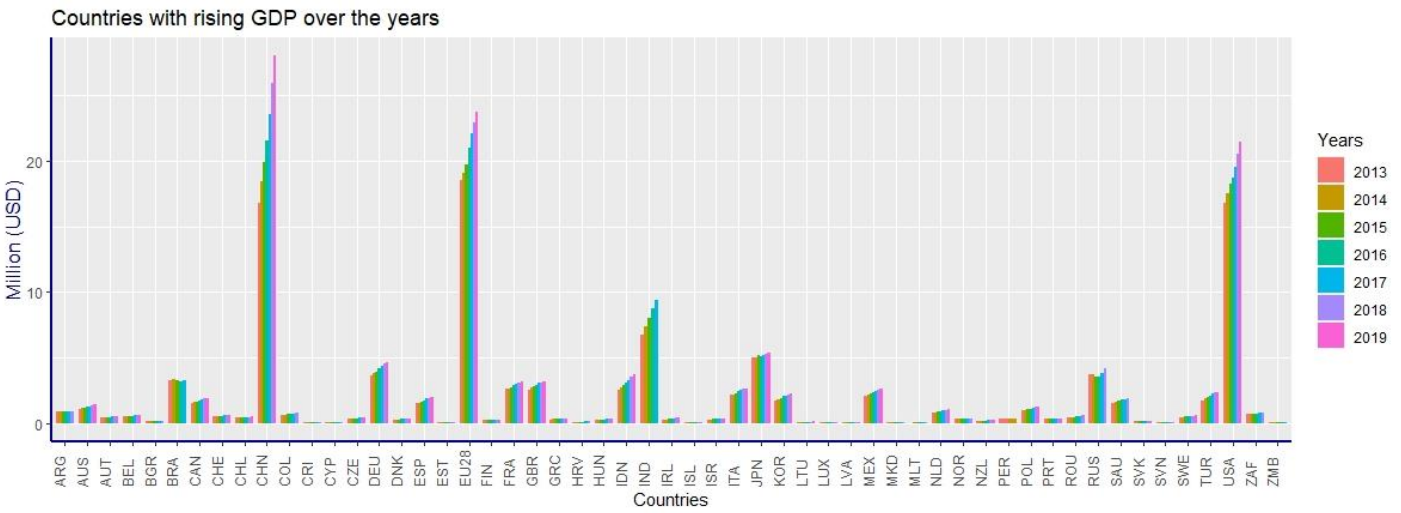


Similar trend can be seen for 2014 till 2019.

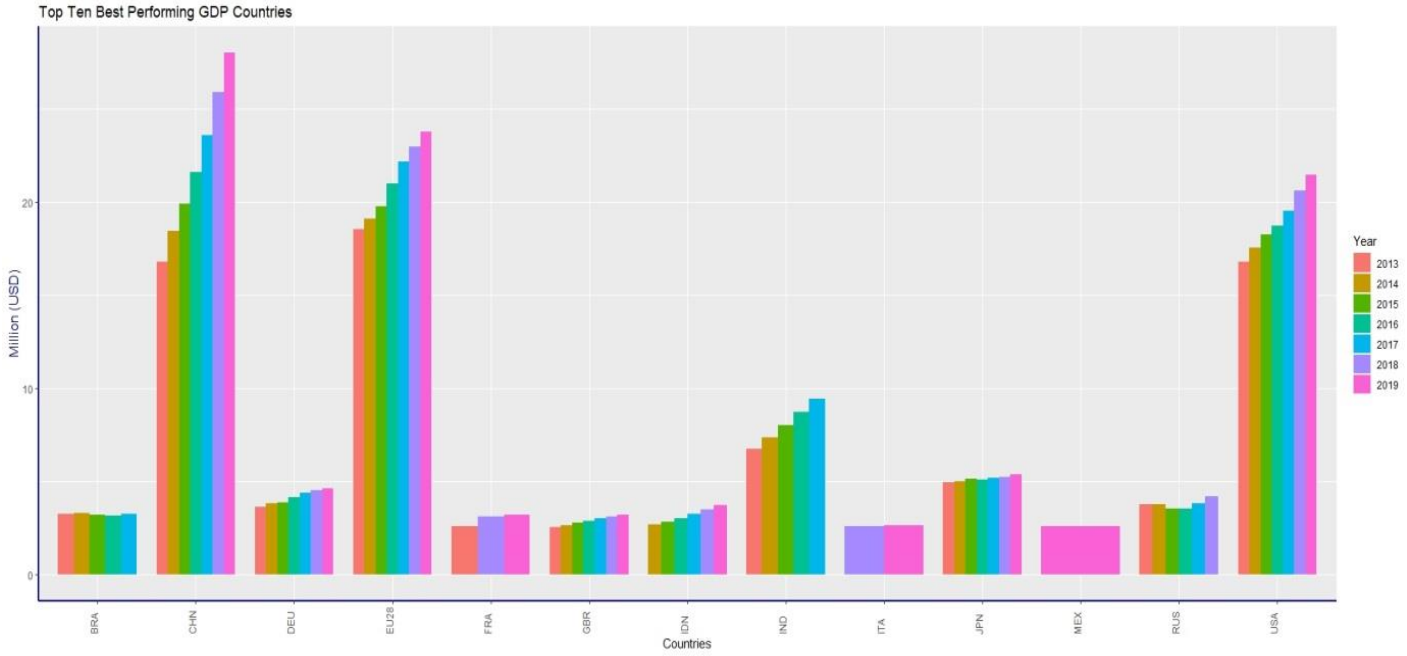

GDP of OECD Group of countries for 2014
Million (USD)
Countries


GDP of OECD Group of countries for 2019
Million (USD)
Countries

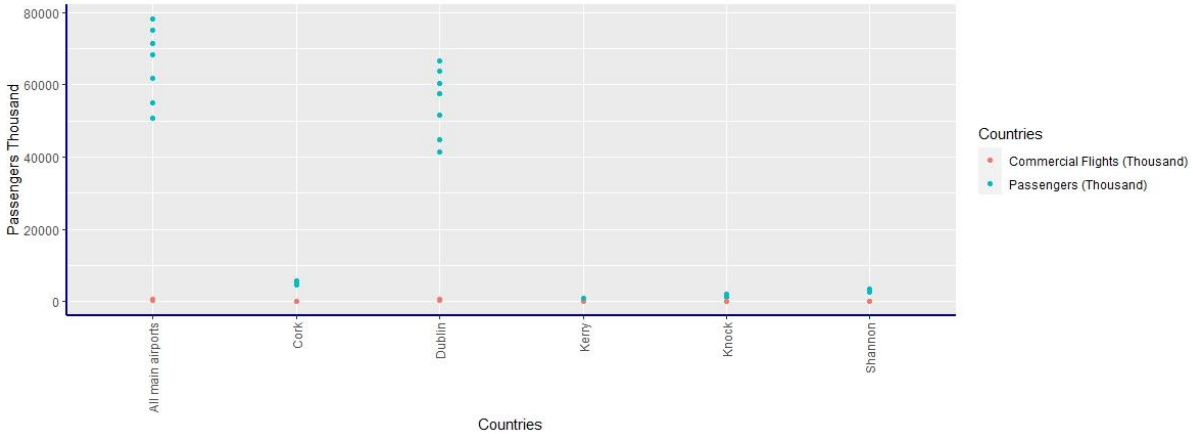

Million USD
Countries
AUS
AUT
BEL
BGR
BRA
CAN
CHE
CHL
CHN
COL
CRI
CYP
CZE
DEU
DNK
ESP
EST
EU28
FIN
FRA
GBR
GRC
HRV
HUN
IDN
IND
IRL
ISL
ISR
ITA
JPN
KOR
LTU
LUX
LVA
MEX
MKD
NLD
NOR
NZL
PER
POL
PRT
ROU
RUS
SAU
SVK
SVN
SWE
TUR
USA
ZAF
ZMB



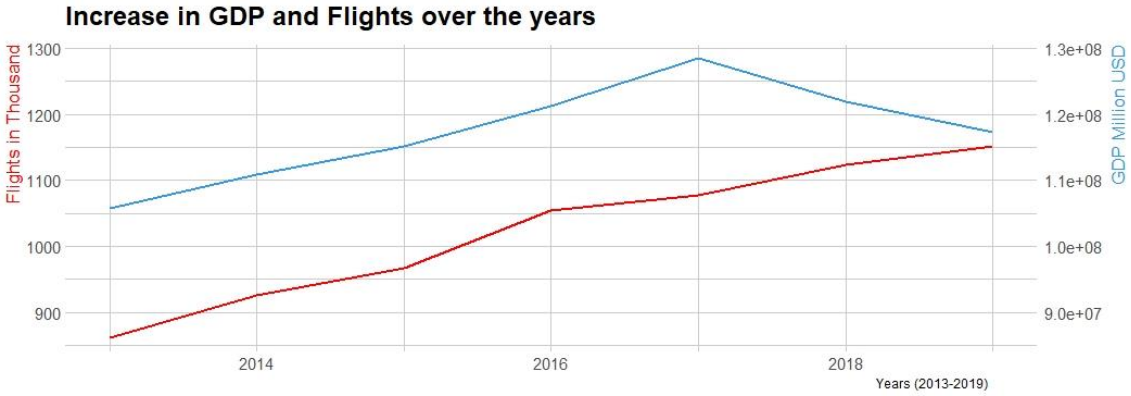
A model for seven years top performing GDPs.

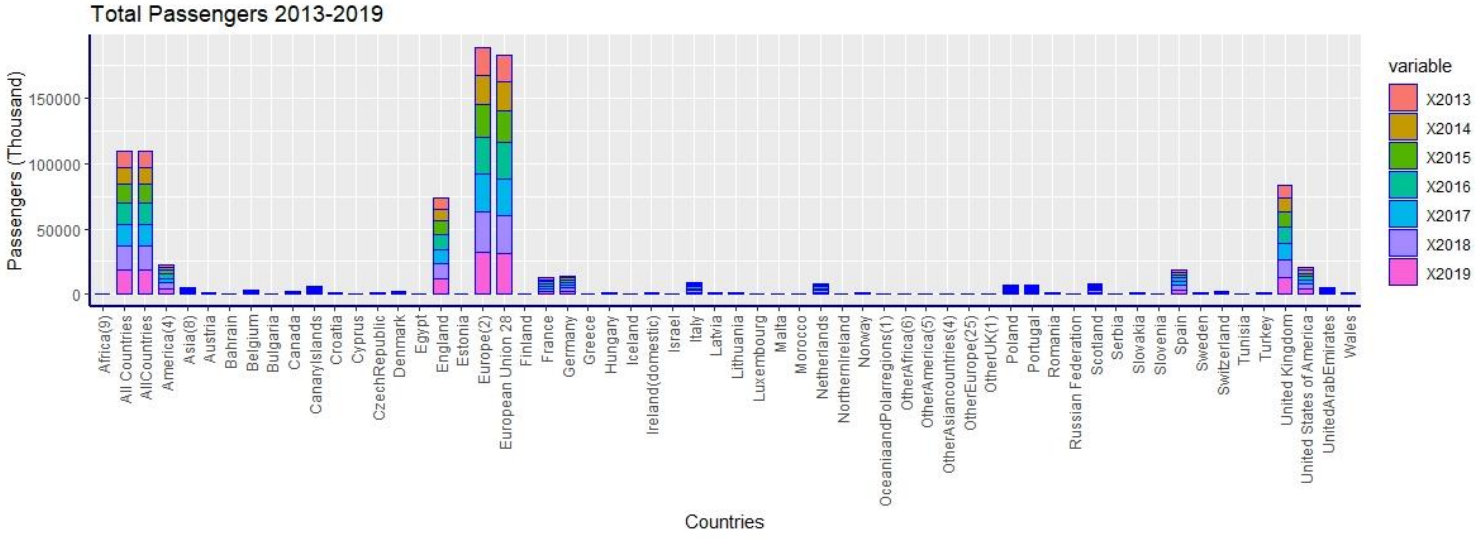


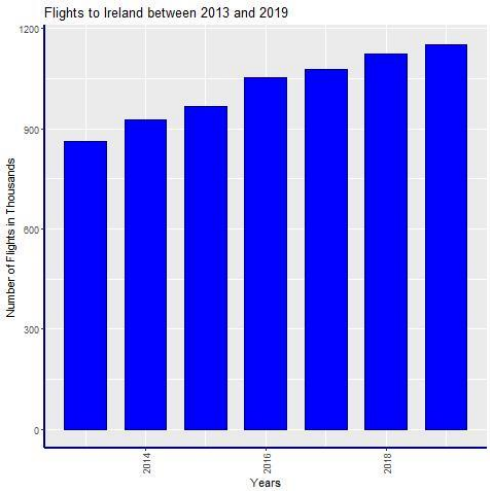
An analysis into Passenger data shows that Dublin receives highest amount of passengers around 7 million passengers between period of 2013-2019.



An analysis was done to see if GDP and number of flights have any association by looking at trends without manipulating data. Both the graphs are increasing with rise in economy.

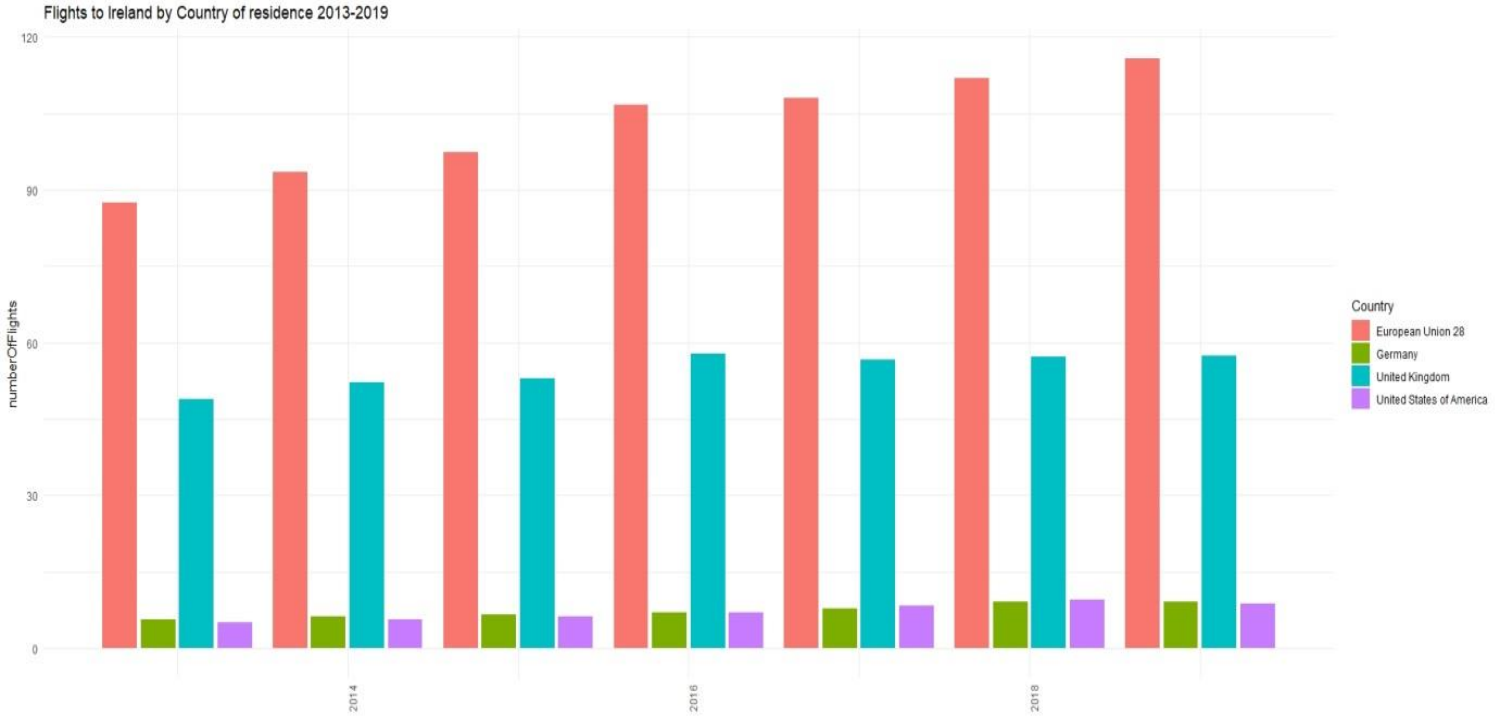






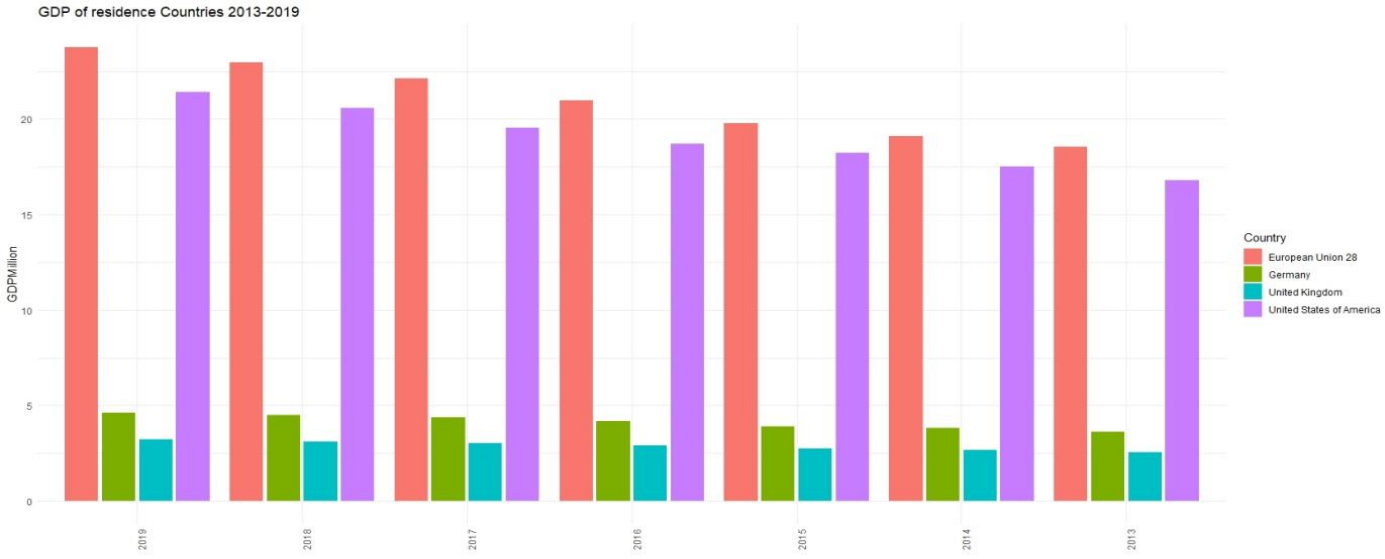
An upward trend in flights is seen by summing data from all the Airports. For the year 2013 flight arrivals have grown from 900000 to 1200000.

After a model is extracted the top four countries are studied for Flight arrival and increase in GDP. We had a limitation as Ireland does not receive direct flights from all countries hence they do not form part of our model.EU28 has most arrivals into the country year 2013 from 90,000 passengers to 110,000 passengers in year 2019. Second is UK which has consistent arrival of flights into Ireland. Similar patterns can be seen for Germany

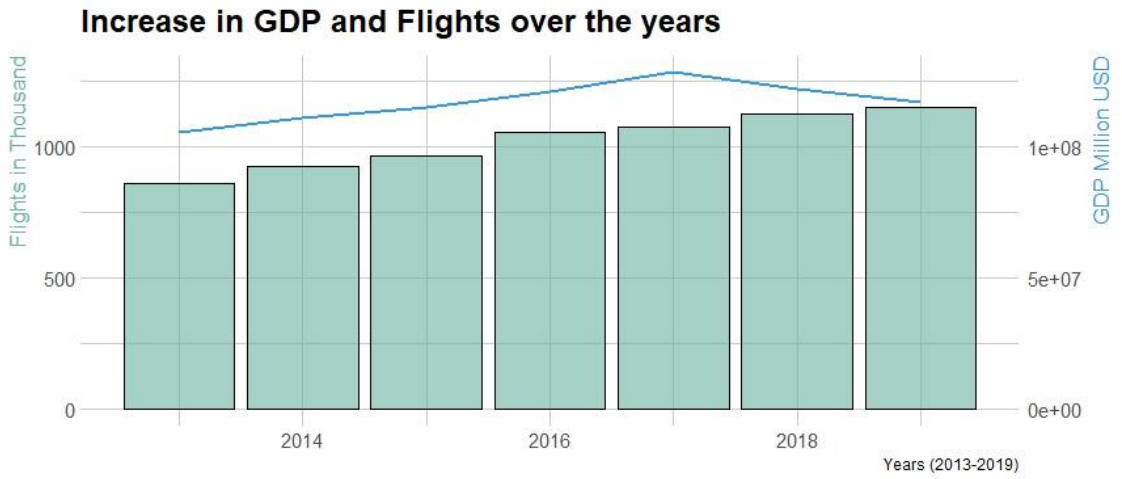


and USA.

EU28 has highest GDP over 25 million in year 2019.Seconded by USA with 23 Million USD in year 2019.



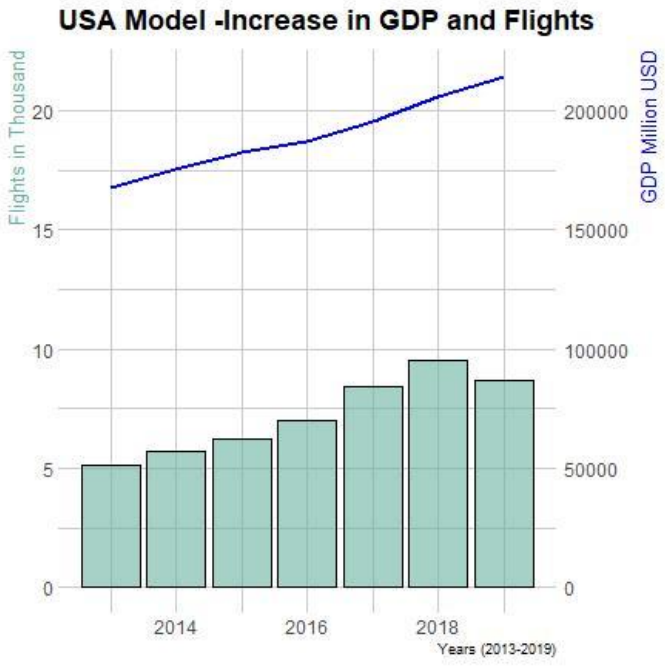
A compiled model of GDP and flight pattern can be seen on this chart. For top four countries. An upward trend can be seen.

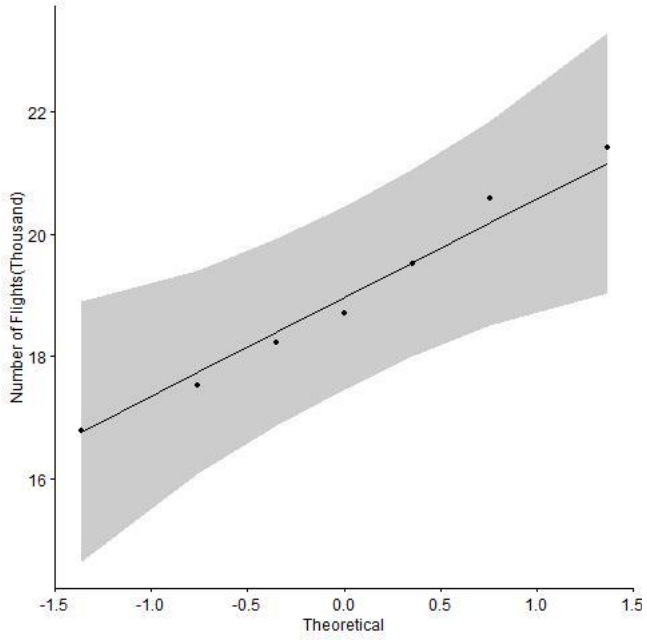


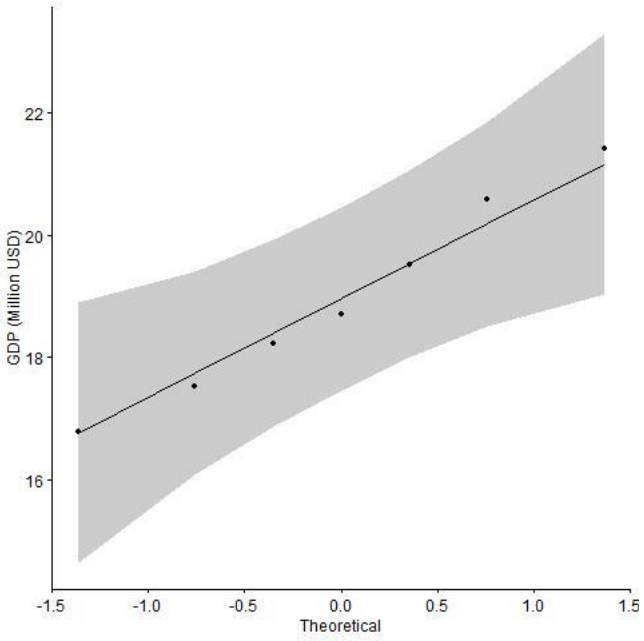
Correlation Test

From exploratory analysis we can see that there is a trend or pattern in rise in GDP and number of Flights to Ireland. But to prove that the results we have received are true and not by chance. We will conduct a correlation model to check our datasets are related.

Before this we will conduct Shapiro-Wilk normality test for number Of Flights and GDP for USA model.







Alpha value of 0.05 is taken here p-value is

P=0.6481which means data are not significantly different from the normal distribution.

Alpha value of 0.05 is taken here for GDP, p-value is p-value=0.9338 which means data are not significantly different from the normal distribution.

```
W = 0.94104, p-value = 0.6481

> shapiro.test(combinedModelGDPFlight_USA$GDPMillion) # => p = 0.9338

        Shapiro-Wilk normality test

data:  combinedModelGDPFlight_USA$GDPMillion
W = 0.97531, p-value = 0.9338

>
```

```
> shapiro.test(combinedModelGDPFlight_USA$numberOfFlights)

        Shapiro-Wilk normality test

data:  combinedModelGDPFlight_USA$numberOfFlights
W = 0.94104, p-value = 0.6481
```

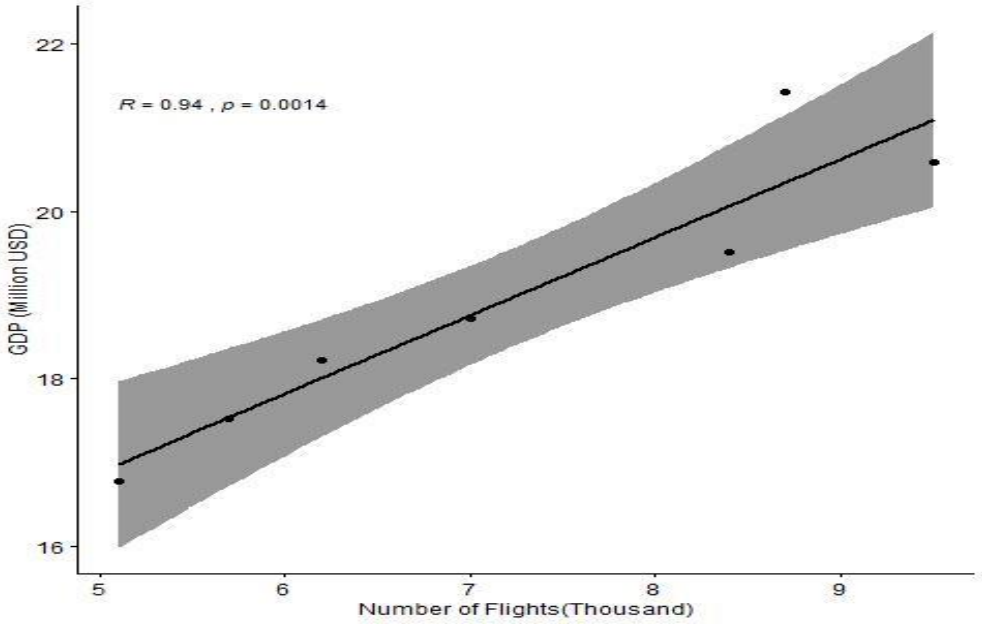
A t-test is performed t-test statistics being t=6.334. Degrees of freedom =5
And correlation coefficient is 0.9429 whereas p-value=0.0014 which is less than alpha=0.05.Hence a positive correlation is found between GDP and flights for USA model.

```
        Pearson's product-moment correlation

data:  combinedModelGDPFlight_USA$numberOfFlights and combinedModelGDPFlight_USA$GDPMillion
t = 6.3334, df = 5, p-value = 0.001447
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6550461 0.9917627
sample estimates:
      cor
0.9429557

>
```

We set an initial hypothesis that there is correlation between GDP growth and number of flights. So we can say that

H0: There is no correlation between GDP and flight data.
Ha : There is correlation between GDP and flight data.

The p-value we got is 0.001447 which is less than alpha value 0.05 .So we can reject null hypothesis for USA model.A positive upward trend can be seen for USA model.



Forecasting:

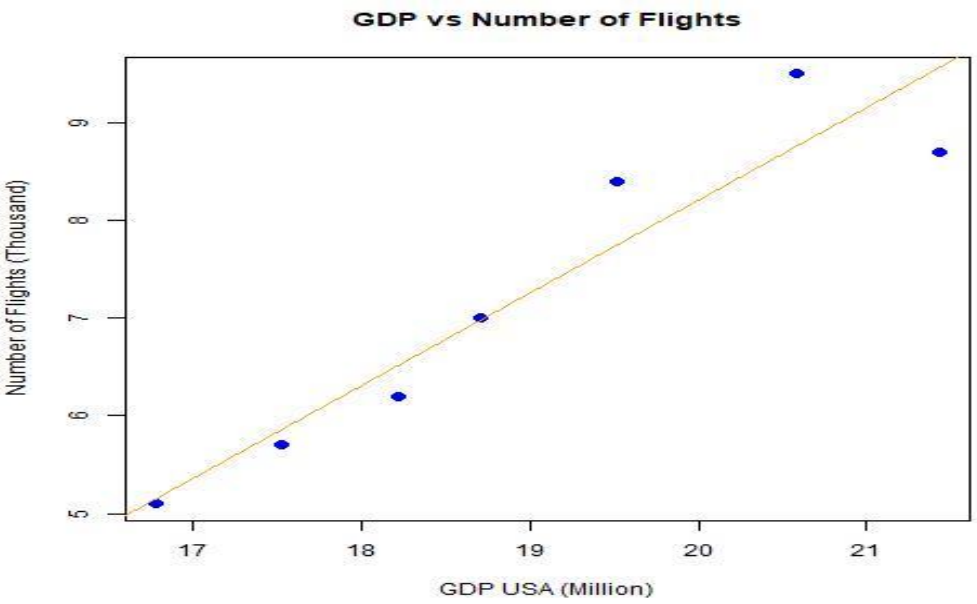
A linear regression model is created for every country in the model.

$$y = a + bx$$

Y is dependent variable which is flights in our case **a** is the intercept of Y and b is the coefficient of independent variable which is GDP.

```
> models <- by(combinedModelGDPFlight, combinedModelGDPFlight$Country, function(df) lm(
> models
combinedModelGDPFlight$Country: European Union 28

Call:
lm(formula = numberOfFlights ~ GDPMillion, data = df)

Coefficients:
(Intercept)   GDPMillion
     -4.645        5.117

------------------------------------------------------------------------------------------
combinedModelGDPFlight$Country: Germany

Call:
lm(formula = numberOfFlights ~ GDPMillion, data = df)

Coefficients:
(Intercept)   GDPMillion
     -7.549        3.595

------------------------------------------------------------------------------------------
combinedModelGDPFlight$Country: United Kingdom

Call:
lm(formula = numberOfFlights ~ GDPMillion, data = df)

Coefficients:
(Intercept)   GDPMillion
      17.96        12.69

------------------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------------------
combinedModelGDPFlight$Country: Germany

Call:
lm(formula = numberOfFlights ~ GDPMillion, data = df)

Coefficients:
(Intercept)   GDPMillion
     -7.549        3.595

------------------------------------------------------------------------------------------
combinedModelGDPFlight$Country: United Kingdom

Call:
lm(formula = numberOfFlights ~ GDPMillion, data = df)

Coefficients:
(Intercept)   GDPMillion
      17.96        12.69

------------------------------------------------------------------------------------------
combinedModelGDPFlight$Country: United States of America

Call:
lm(formula = numberOfFlights ~ GDPMillion, data = df)

Coefficients:
(Intercept)   GDPMillion
   -10.8004       0.9505

>
```



A positive linear upwards linear model is seen for USA model.

Some predictions are done for a GDP of 36 million ,numbers of flights could be 13.7 thousand to Ireland for European Union. Prediction for USA gave number of flights from 5.15 thousand flights to 9.56 for next seven years.

```
GDPMillion    0.5646973  1.336242

> y1=-4.645+5.117*3.6 # x(independent variable = 3.6 assumed)
> y1
[1] 13.7762
> y1=-4.645+5.117*3.6 # x(independent variable = 3.6 assumed)
> y1
[1] 13.7762
>
> ### Germany Intercept and Slope
> #y2=a+bx
> y2=-7.549+3.595 *3.6 # x(independent variable = 3.6 assumed)
> y2
[1] 5.393
>
> ### United Kingdom Intercept and Slope
> y3=17.96+12.69 *3.6 # x(independent variable = 3.6 assumed)
> y3
[1] 63.644
>
>
> ###United States of America
> y4=-10.8004+0.9505 *3.6 # x(independent variable = 3.6 assumed)
> y4
[1] -7.3786
>
```

2. Challenges Faced

Although the datasets had no missing values.,but there were lot of challenges faced during the creation of model.

- Flight arrival data was reading the columns as x2013 or x2014
  This was cleaned using function:

  as.integer(gsub('[a-zA-Z]', '', toFlightsGroupCountryTransposed$variable))

- Before making data available for model all the data for each country with flights all over the Ireland was aggregated.

- Series of extraction was required to form a dataframe for top ten GDP performing economies each year. This was then combined together to get a final top performing countries. But the Flight arrival data in Ireland does not show direct link of passengers and flights to Ireland from Asian countries. Hence those countries were not included as part of study.

- The GDP dataset contained only ISO3 countries whereas Flight data had country name hence a third dataset was used to extract information. A for loop was created to check ISO3 code from GDP data and matched with Country code data. A new column was initialised with no values and data was inserted to get a model which could be compared with Flight data.

- This Flight Data frame needed to be transposed so that two datasets can be merged into one mode. R does not provide proper transposing of data hence had to do research to transpose data frame. As all the years were in columns whereas GDP Years were in rows. To merge two datasets using column bind function, where equal number of records need to be aligned with flight data. For this reshape2 package was used since melt package was not performing well.

- Again a for loop was required to check if the country is present in GDP and is present in Flight data create a model.

- A model was taken were all the values were present for GDP as well as flight data.

- While plotting two variables on a different scale it was difficult to plot them together in a chart. I did research on this and plotted two different scales on same chart.

- Created function which need constant referencing.

Project Report

2) Correlation between variability of rainfall and production of crops.

Assignment CA 2

3. Introduction

India is a country which needs constant production of crops in order to feed the growing population. As population touches 140 Billion in coming years we need to find if the changes in rainfall pattern has caused effect on production of crops which need high amount of water to support their growth. As a part of study we will study four crops which need more water.

3.1. Objectives

The objective of the study is to see if there is a correlation between Rainfall and crop production. For conducting the study we will use the crops which need more water.
Based on table from Crop selection website Banana, Wheat, Cotton and Maize are selected.

Crop	Crop water need (mm/total growing period)
Banana	1200-2200
Wheat	450-650
Cotton	700-1300
Maize	500-800

We will try to find out is there any affect in crop production if rainfall pattern changes?
So we will set our Hypothesis as
Ho : There is no effect on crop production
Ha: There is effect on crop production due to lack of rainfall.

3.2. Literature Review

Impact of Rainfall Variability on Crop Production within the Worobong Ecological Area of Fanteakwa District, Ghana

Study was done on a research paper by Conrad Kyei-Mensah,1 Rosina Kyerematen . The literature conducts a study on Ghana which has suffered droughts in the eastern region district. The variability in rainfall could be a cause in crop production and food security. They studied data from 1985 till 2014.Three decades. Correlation analysis was done to assess the influence of rainfall and crops.
Studies were conducted on seasonal rainfall as they account for most rains in the country.

For each year they calculated the mean of crops which are staple to Ghana for 30 years timeframe. Minimum production and maximum production was reported. A variability in Rainfall pattern is calculated for three decades. A model for crop is developed and the

production data is plotted for all individual crops. Following this Pearson's correlation coefficient was calculated on selected crops and climate variables rainfall and temperature. Their correlation was found to be negative except for tomatoes. None of the crops showed significance with rain variability.

3.3. Dataset description

Food and Agriculture Organization of the United Nations database was used for crops data .The data provides information on the crop produced all over the world starting from year 1975 till 2016.. Following website was accessed; here data is free to download for public study purposes: http://www.fao.org/faostat/en/#data/PP

Area Code	Area	Item Code	Item	Element Code	Element	Year Code	Year	Unit	Value
2	Afghanistan	221	Almonds, with shell	5312	Area harvested	1975	1975	ha	0

There are 10 fields in this dataset.

- Area Code contains area code of the country. This is specified in numbers.
- Area is the country for which crop produce is given. This is a factor with 122 levels.
- Item.code is the code for each crop. This is given in numbers.
- Item is the name of the crop written as factor and has 180 levels
- Element code is the integer column in the dataframe.
- Elements are the column for three distinctive factors namely Area harvested, Yield and Production. Area harvested is the area of planted crops, yield refers per area harvest and production is the total harvest in tonnes per hectare.
- Year.Code and Year are integer fields and there is no difference in them.
- Unit is a factor column which has three levels hg/ha, tonnes and ha.
  For production ha (hectare) is used. For yield hg/ha(hectogram/hectare) is used and production.
- Value is the price of each item in integers.

For rainfall data World Bank data was used to look for rainfall in countries all over the world. This is Available at: https://climateknowledgeportal.worldbank.org/download-data

The rainfall data is from 1991 till year 2016.

Rainfall(MM)	Year	Statistics	Country	ISO3
64.7765	1991	Jan Average	Afghanistan	AFG

The rainfall datasets contains five fields.

- Rainfall column in MM. It's a numerical filed.
- Year as integer starting from 1991 till 2016.
- Statistics a factor field with 13 levels for each month written as Jan Average ,Feb Average.
- Country with 143 levels, this is a factor.
- ISO3 country codes with 143 levels. This too is a factor column.

3.4. Analysis approach

The aim of this study is to a find a correlation between variability in Rainfall and production of crops in India. A Cross-industry process for data mining methodology will used to achieve the desired results.
Earlier we have discussed there are six stages in the methodology.

1. Business understanding – We have set objectives for studying a correlation between rain and crop production and this will form basis for our business understanding.

2. Data Understanding –For second stage data is collected from Food and Agriculture Organization of the United Nations which is a trusted organisation and the data is available for public use. Climate rainfall data is sourced from World Bank which provide data access for performing statistical analysis. A deep analysis is done on the datasets which is discussed above most of the columns are used in developing the model.
3. ***Data preparation*** **–** The third stage involves preparing data which can be modelled. To gain fruitful insights. The dataset is selected for Crops with high water requirement. A subset of data for India with crops like banana, wheat, cotton and Maize is selected. Similarly the Rainfall for years 1996 till 2016 is selected for India. To further refine our research seasonal rains are also accounted as this gives accurate information about

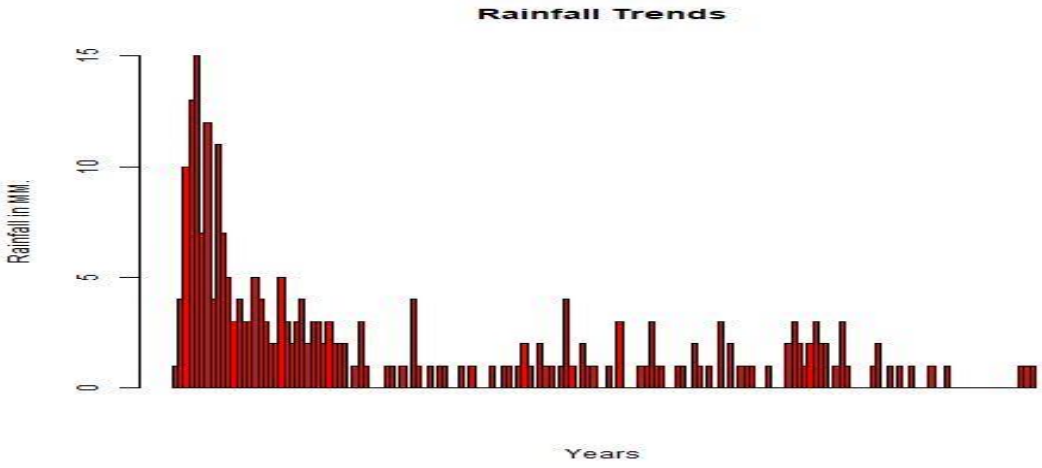
rainfall trends in the country. Seasonal rains occur during period of June till October in India. A combined model of rainfall pattern and production crops was merged to get desired model.

4.Modelling- After series of extraction a model is developed which will study rainfall trends in 10 years and crop production during this years. This modelled will be studied for correlation between the two.

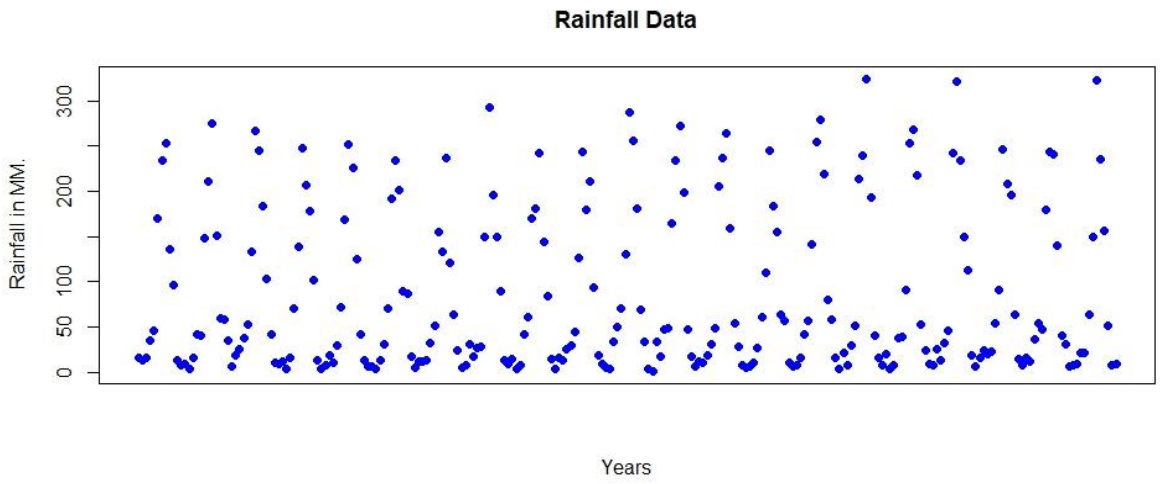
5. Evaluate We will further evaluate the results from our chosen objectives i.e. is our hypothesis was true or not.

3.5. Analysis results and presentation

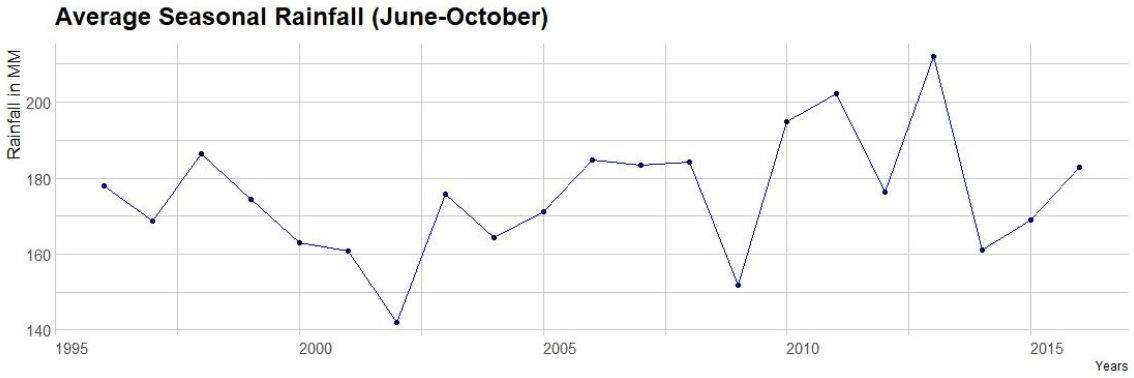
An analysis in to study of rainfall data shows that over span of ten years the amount of rainfall in the country has reduced significantly.



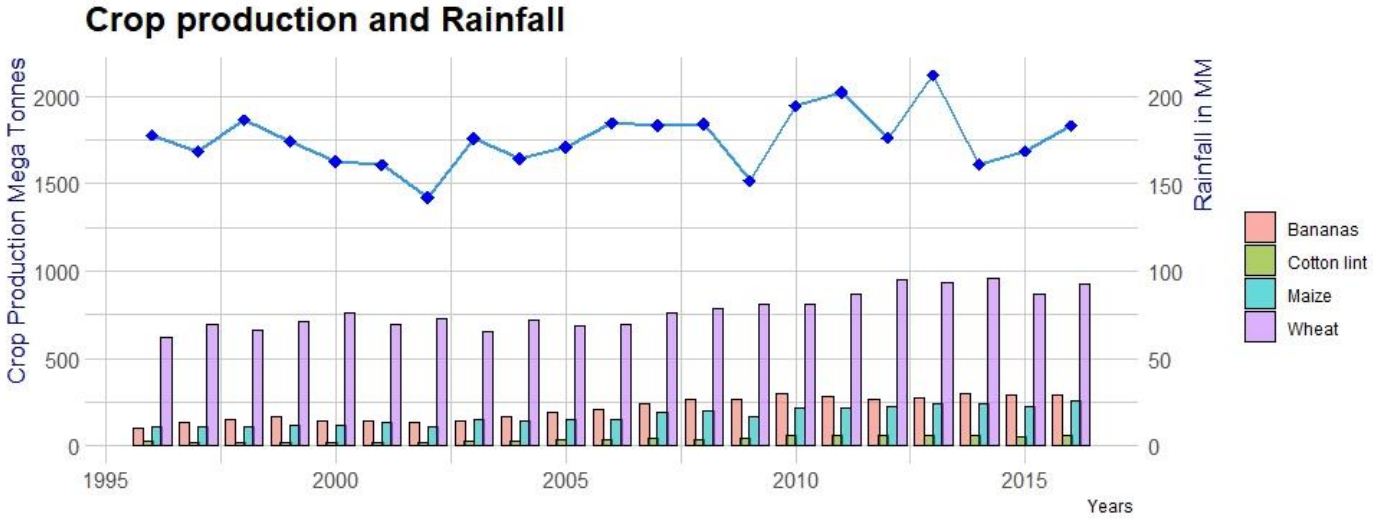
Distribution of rainfall over the years a scatter plot shows that the data is distributed all over with no consistency of rainfall patterns.

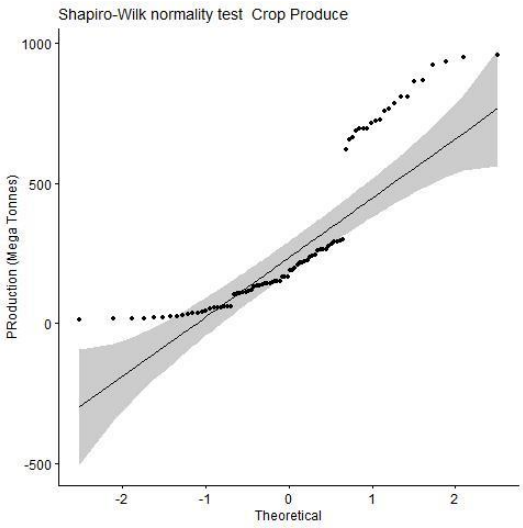


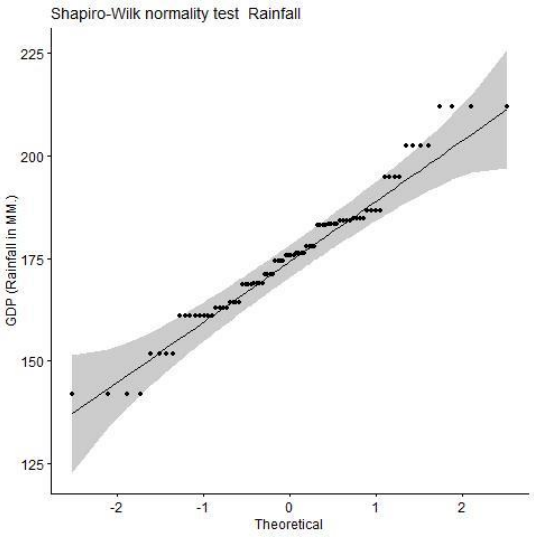
A seasonal model was created from June till October for each year this modelled shows that the average distribution of rainfall has been varying over the years. Year 1996 being lowest rainfall period with less than 180mm. 2013 being highest where seasonal rainfall average is more than 220 mm.



A combined model was created using average crop production each year against the rainfall pattern because of seasonal rains. While looking at the trends we can see that in some case we can find a relationship between crops and rainfall. Wheat is the highest producing crop whereas cotton is least producing crop. Wheat achieved maximum production in the year 2013 when rainfall was highest.







Shapiro-Wilk normality test is done for Crop produce and Rainfall data.

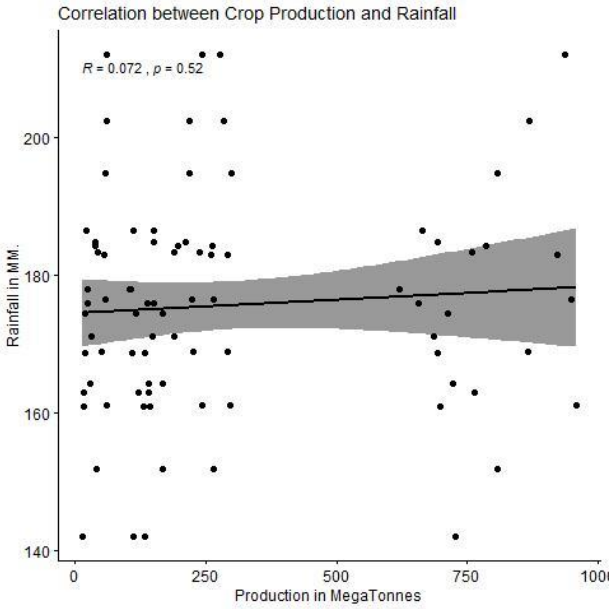
Alpha value of 0.05 is taken here p-value 2.809e-09 which means data are significantly different from normal distribution. In other words, we can't assume the normality.

For rainfall data Alpha value is 0.05 => p = 0.04603 hence rainfall data maybe significantly different from normal distribution.

```
> shapiro.test(combinedModelCropRainAgg$MegaTonnes) # => p = 2.809e-09

        Shapiro-Wilk normality test

data:  combinedModelCropRainAgg$MegaTonnes
W = 0.80141, p-value = 2.809e-09
```

```
> shapiro.test(combinedModelCropRainAgg$Group.3) # => p = 0.04603

        Shapiro-Wilk normality test

data:  combinedModelCropRainAgg$Group.3
W = 0.96987, p-value = 0.04603
```

The test statistics for the data was done. The p-value of the test is 0.5174, which is more than the significance level alpha = 0.05. We can conclude that production and rainfall are not significantly correlated with a correlation coefficient of 0.07161444 and p-value of 0.5174.



The graph plot shows a flat line showing that three is no correlation between rainfall scarcity and production of crops. Hence we cannot reject null hypothesis and our alternative hypothesis was false.

```
        Pearson's product-moment correlation

data:  combinedModelCropRainAgg$MegaTonnes and combinedModelCropRainAgg$Group.3
t = 0.65017, df = 82, p-value = 0.5174
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1450072  0.2816847
sample estimates:
       cor
0.07161444
```

4. Challenges Faced

1. Data had few NA values but after extracting required data and years the model did not need to corrected for NA values which was a relief.
2. Learning from first project helped me model this dataset fast.

5. Appendices

5.1. Appendix 1–Rise in GDP probable cause of increase in Flights

```
getwd()getwd()
install.packages("ggplot2")
library(ggplot2)
install.packages("scales")
library(scales)
install.packages("plotly")
library(plotly)
install.packages(reshape2)
library(reshape2)           #to transpose Dataframe plots
install.packages("data.table")
library(data.table)
install.packages("gapminder") #animated plots
library(gapminder)
install.packages("gganimate") #animated plots
library(gganimate)
install.packages("plotly")
library(plotly)
install.packages("dplyr")
library(dplyr)
install.packages("patchwork")
library(patchwork) # To display 2 charts together
install.packages("hrbrthemes")
library(hrbrthemes)
install.packages("impute")
library(impute)
install.packages("plotrix")
library(plotrix )## To Clip abline
install.packages("tidyverse") ## To reOrder grouped bars on X-axis Plots
install.packages("ggpubr")
library(ggpubr) ## To do scatter plot of Corelation
install.packages("tidyverse")
library(tidyverse) #tidyverse for easy data manipulation and visualization
install.packages("broom")
library(broom) #broom: creates a tidy data frame from statistical test results
install.packages("melt")
library(melt)

source('~/functionFile.R') ##Sourcing Function File into Main file

###############################################

# This Project is done with objective to study relationship of GDP and increase in flight
arrivals in Ireland
```

```
# The study try to findout that correlation between increase in GDP and number of Flights

# Data for Flights arrival  was sourced at
https://statbank.cso.ie/px/pxeirestat/Statire/SelectVarVal/saveselections.asp

### Data for GDP was sourced at https://data.oecd.org/gdp/gross-domestic-product-
gdp.htm#indicator-chart

### Data for ISO 3 Country codes was sourced at https://github.com/datasets/country-
codes

### Correlation Model was referenced at http://www.sthda.com/english/wiki/correlation-
test-between-two-variables-in-r
# https://www.r-graph-gallery.com/line-chart-dual-Y-axis-ggplot2.html

##########################################################
################################  Reading file GDP Yearly CSV file
#########################

gDPYearly<- read.csv(file="GDPYearly.csv", head=TRUE, sep=",")
gDPYearly
nrow(gDPYearly) #### number of rows =4441

gDPYearly<-subset(gDPYearly,ï..LOCATION!="EA19")
gDPYearly
nrow(gDPYearly)

# Checking for NA
gDPYearly[!complete.cases(gDPYearly),]
### There are no incomplete cases or NAs in the data
summary(gDPYearly)
str(gDPYearly)  #Structure of CSV file

############################
#Creating a subset of 2013 till 2019 GDP data as we only have flights arrival data from
this date
gDPYearly_2013_2019<-subset(gDPYearly, TIME >= 2013 & TIME <= 2019  &
MEASURE =="MLN_USD" & ï..LOCATION!= "OECD" & ï..LOCATION!="OECDE")
gDPYearly_2013_2019

#### Aggregate of GDP groupbyYEar and Location
gDPYearly_2013_2019Agg<-
aggregate(gDPYearly_2013_2019$Value,by=list(gDPYearly_2013_2019$TIME,gDPYear
ly_2013_2019$ï..LOCATION), FUN=sum)
gDPYearly_2013_2019Agg


#### Aggregate of GDP Values groupbyYear
```

```
gDPYearly_2013_2019_Year<-
aggregate(gDPYearly_2013_2019$Value,by=list(gDPYearly_2013_2019$TIME),
FUN=sum)
gDPYearly_2013_2019_Year


############################## Subsets of each Year ##########################
####### Creating a Subset of Year for study 2013 and measure as MLN_USD WHILE
TAKING OUT OECD and OECDE as it is total shown in the dataset
gDPYearly_2013<-subset(gDPYearly, TIME == 2013 & MEASURE =="MLN_USD" &
ï..LOCATION!= "OECD" & ï..LOCATION!="OECDE")
gDPYearly_2013
nrow(gDPYearly_2013)

##### Basic Statistics
## Calling Function descStats sourcing it from DifferentFile
descStats(gDPYearly_2013$Value)

## Using Function findingStats to lovcate which Countries are highest gDP and lowest
perfomeing GDPs ##########
findingStats(gDPYearly_2013$ï..LOCATION,gDPYearly_2013$Value) ## EU28 and MLT
are High and low performing Economies

# Aggregating GDP by Year and location
agg_gDP2013<-aggregate(gDPYearly_2013$Value,
by=list(gDPYearly_2013$TIME,gDPYearly_2013$ï..LOCATION), FUN=sum)
agg_gDP2013


###### Calling Function ggplot2 from source File to plot gDP in 2013#####

jpeg("OECD2013.jpg")
ggplotFunc(agg_gDP2013,agg_gDP2013$Group.2,agg_gDP2013$x,"GDP of OECD
Group of countries for 2013")
dev.off()

###### Calling function topTenGDP and  Finding top 10 performing countries for 2013
topTen2013<-topTenGDP(gDPYearly_2013,gDPYearly_2013$Value)
topTen2013

## Plot for Top Ten Countries
jpeg("TopTenGDP2013.jpg")
ggplotFunc(topTen2013,topTen2013$ï..LOCATION,topTen2013$Value,"Top Five GDP
performing countries for 2013")
dev.off()

### Assigning the result to a variable so that further into analysis we can cbind to from a
model
topTenEachYear<- topTen2013
topTenEachYear
```

```
########################################    2014
################################
####### Creating a Subset of Year for study 2014 and measure as MLN_USD WHILE
TAKING OUT OECD and OECDE as it is total shown in the dataset
gDPYearly_2014<-subset(gDPYearly, TIME == 2014 & MEASURE =="MLN_USD" &
ï..LOCATION!= "OECD" & ï..LOCATION!="OECDE")
gDPYearly_2014
nrow(gDPYearly_2014)

##### Basic Statistics
## Calling Function descStats sourcing it from Different File
descStats(gDPYearly_2014$Value)

## Using Function findingStats to lovcate which Countries are highest GDP and lowest
perfomeing GDPs ##########
findingStats(gDPYearly_2014$ï..LOCATION,gDPYearly_2014$Value)

## Aggregate GDP by group of Countries
agg_gDP2014<-aggregate(gDPYearly_2014$Value,
by=list(gDPYearly_2014$TIME,gDPYearly_2014$ï..LOCATION), FUN=sum)
agg_gDP2014

#####

jpeg("ToTenGDP2014.jpg")
ggplotFunc(agg_gDP2014,agg_gDP2014$Group.2,agg_gDP2014$x, "GDP of OECD
Group of countries for 2014")
dev.off()

###### Calling function and  Finding top 10 performing conutries for 2014
topTen2014<-topTenGDP(gDPYearly_2014,gDPYearly_2014$Value)
topTen2014

##############################

gDPYearly_2015<-subset(gDPYearly, TIME == 2015 & MEASURE =="MLN_USD" &
ï..LOCATION!= "OECD" & ï..LOCATION!="OECDE")
gDPYearly_2015
str(gDPYearly_2015)
nrow(gDPYearly_2015)


## Calling Function descStats sourcing it from Different File
descStats(gDPYearly_2015$Value)
###### Calling function and  Finding top 5 performing conutries for 2015
topTen2015<-topTenGDP(gDPYearly_2015,gDPYearly_2015$Value)
topTen2015
```

```
## Using Function findingStats to lovcate which Countries are highest GDP and lowest perfomeing GDPs ##########
findingStats(gDPYearly_2015$ï..LOCATION,gDPYearly_2015$Value)

##Aggregate by country
agg_gDP2015<-aggregate(gDPYearly_2015$Value,
by=list(gDPYearly_2015$TIME,gDPYearly_2015$ï..LOCATION), FUN=sum)
agg_gDP2015


##############################################2016

gDPYearly_2016<-subset(gDPYearly, TIME == 2016 & MEASURE =="MLN_USD" &
ï..LOCATION!= "OECD" & ï..LOCATION!="OECDE")
gDPYearly_2016
nrow(gDPYearly_2016)



## Calling Function descStats sourcing it from Different File
descStats(gDPYearly_2016$Value)

## Using Function findingStats to lovcate which Countries are highest gDP and lowest perfomeing gDPs ##########
findingStats(gDPYearly_2016$ï..LOCATION,gDPYearly_2016$Value)

agg_gDP2016<-aggregate(gDPYearly_2016$Value,
by=list(gDPYearly_2016$TIME,gDPYearly_2016$ï..LOCATION), FUN=sum)
agg_gDP2016

###### Calling function and  Finding top 10 performing conutries for 2016
topTen2016<-topTenGDP(gDPYearly_2016,gDPYearly_2016$Value)
topTen2016

##############################################

gDPYearly_2017<-subset(gDPYearly, TIME == 2017 & MEASURE =="MLN_USD" &
ï..LOCATION!= "OECD" & ï..LOCATION!="OECDE")
gDPYearly_2017
nrow(gDPYearly_2017)


## Calling Function descStats sourcing it from Different File
descStats(gDPYearly_2017$Value)

## Using Function findingStats to lovcate which Countries are highest gDP and lowest perfomeing GDPs ##########
findingStats(gDPYearly_2017$ï..LOCATION,gDPYearly_2017$Value)
```

```
agg_gDP2017<-aggregate(gDPYearly_2017$Value,
by=list(gDPYearly_2017$TIME,gDPYearly_2017$ï..LOCATION), FUN=sum)
agg_gDP2017

###### Calling function and  Finding top 5 performing conutries for 2017
topTen2017<-topTenGDP(gDPYearly_2017,gDPYearly_2017$Value)
topTen2017



##########################################################

gDPYearly_2018<-subset(gDPYearly, TIME == 2018 & MEASURE =="MLN_USD" &
ï..LOCATION!= "OECD" & ï..LOCATION!="OECDE")
gDPYearly_2018
nrow(gDPYearly_2018)



## Calling Function descStats sourcing it from Different File
descStats(gDPYearly_2018$Value)

## Using Function findingStats to lovcate which Countries are highest GDP and lowest
perfomeing GDPs ##########
findingStats(gDPYearly_2018$ï..LOCATION,gDPYearly_2018$Value)

agg_gDP2018<-aggregate(gDPYearly_2018$Value,
by=list(gDPYearly_2018$TIME,gDPYearly_2018$ï..LOCATION), FUN=sum)
agg_gDP2018

###### Calling function and  Finding top 10 performing conutries for 2018
topTen2018<-topTenGDP(gDPYearly_2018,gDPYearly_2018$Value)
topTen2018

##########################################################

gDPYearly_2019<-subset(gDPYearly, TIME == 2019 & MEASURE =="MLN_USD" &
ï..LOCATION!= "OECD" & ï..LOCATION!="OECDE"  &ï..LOCATION!="EA19")
gDPYearly_2019
nrow(gDPYearly_2019)


## Calling Function descStats sourcing it from Different File
descStats(gDPYearly_2019$Value)

## Using Function findingStats to lovcate which Countries are highest gDP and lowest
perfomeing GDPs ##########
findingStats(gDPYearly_2019$ï..LOCATION,gDPYearly_2019$Value)
```

```
agg_gDP2019<-aggregate(gDPYearly_2019$Value,
by=list(gDPYearly_2019$TIME,gDPYearly_2019$ï..LOCATION), FUN=sum)
agg_gDP2019

### GDP 2019
jpeg("TopTenGDP2019.jpg")
ggplotFunc(agg_gDP2019,agg_gDP2019$Group.2,agg_gDP2019$x, "GDP of OECD
Group of countries for 2019")
dev.off()

###### Calling function and  Finding top 10 performing conutries for 2019
topTen2019<-topTenGDP(gDPYearly_2019,gDPYearly_2019$Value)
topTen2019

#############
combinedtopTenYearly <-
rbind(topTen2013,topTen2014,topTen2015,topTen2016,topTen2017,topTen2018,topTen
2019)
combinedtopTenYearly


#### Aggregate of GDP groupbyYEar and Location
combinedtopTenYearlyAgg<-
aggregate(combinedtopTenYearly$Value,by=list(combinedtopTenYearly$TIME,combine
dtopTenYearly$ï..LOCATION), FUN=sum)
combinedtopTenYearlyAgg


######################     Plotting Descriptive Functions     #############

### Qplot showing GDP over the years each dot represent one year


### GDP 2019
jpeg("TopTenGDP2013_2019.jpg")
ggplot(gDPYearly_2013_2019Agg,aes(Group.2,x,color=Group.2))+geom_point()+
  theme(axis.line = element_line(colour = "darkblue",size = 1, linetype = "solid"),
        axis.text.x = element_text(angle =90, vjust =0 ,hjust = 0))+
  xlab("Countries") + ylab("Million USD") +
  scale_color_discrete(name = "Countries")+ guides(fill=guide_legend(title=""))
dev.off()


####Calling function from Source File to plot
jpeg("OECDGDP2013_2019.jpg")
ggplotFuncGDP(gDPYearly_2013_2019Agg,gDPYearly_2013_2019Agg$Group.2,gDPY
early_2013_2019Agg$x,
              gDPYearly_2013_2019Agg$Group.1,"Countries","Million-USD","Countries with
rising GDP over the years","Years")
dev.off()
```

```
##################################################
####Calling function from Source File to plot
jpeg("OECDTenGDP2013_2019.jpg")
ggplotFuncGDP(combinedtopTenYearlyAgg,combinedtopTenYearlyAgg$Group.2,combi
nedtopTenYearlyAgg$x,
          combinedtopTenYearlyAgg$Group.1,"Countries","Million-USD","Top Ten Best
Performing GDP Countries","Year")
dev.off()



##########################################
    ############################################################



#############################################################################
###

##### custom Functions sourced from functionFile
source('~/04_function.R')

#########################    Arrivals ##########################
yearlyArrivalsFlights<- read.csv(file="YearlyArrivalsFlights.csv", head=TRUE, sep=",")
yearlyArrivalsFlights
class(yearlyArrivalsFlights)
nrow(yearlyArrivalsFlights)

# Checking for NA
yearlyArrivalsFlights[!complete.cases(yearlyArrivalsFlights),]
yearlyArrivalsFlights<-yearlyArrivalsFlights[complete.cases(yearlyArrivalsFlights),]

#### strucrure of File
str(yearlyArrivalsFlights)


#### Aggregating Years by category and arrival Airport
yearlyArrivalsFlightsCity<-
aggregate(yearlyArrivalsFlights[,4:10],by=list(yearlyArrivalsFlights$Category,yearlyArrival
sFlights$DestinationCity), FUN=sum)
yearlyArrivalsFlightsCity
### Using melt function to melt years
yAFNew<- melt(yearlyArrivalsFlightsCity)
yAFNew


### QPlot for Airports and Passengers
jpeg("Flights and PassengersAirport.jpg")
ggplot(yAFNew,aes(Group.2,value,color=Group.1))+geom_point()+
```

```
  theme(axis.line = element_line(colour = "darkblue",size = 1, linetype = "solid"),
        axis.text.x = element_text(angle =90, vjust =0.5 ,hjust = 1))+
  xlab("Countries") + ylab("Passengers Thousand") +
  scale_color_discrete(name = "Countries")
dev.off()


#########################################################################
#############


#### Aggregating Years by Destination Countries  and arrival Airport
yearlyArrivalsFlightsCountry<-
aggregate(yearlyArrivalsFlights[,4:10],by=list(yearlyArrivalsFlights$DestinationCity,yearly
ArrivalsFlights$SourceCountries), FUN=sum)
yearlyArrivalsFlightsCountry


######################  Flights ################################
#### Subseting Arrivals as per Flights
yearlyArrivalFlightsComm<-subset(yearlyArrivalsFlights, Category == "Commercial
Flights (Thousand)")
yearlyArrivalFlightsComm


############## Total Commercial Flights in Ireland each year ##############

totalFlightsComm<-
aggregate(yearlyArrivalFlightsComm[,4:10],by=list(yearlyArrivalFlightsComm$Category),
FUN=sum)
totalFlightsComm

### Using melt function to melt years
totalFlightsCommNew<- melt(totalFlightsComm)
totalFlightsCommNew

######################### Creating Model#################
## Removing Character from the Model
totalFlightsCommNew$variable<- as.integer(gsub('[a-zA-Z]', '',
totalFlightsCommNew$variable))
totalFlightsCommNew

gDPYearly_2013_2019_Collated<-subset(gDPYearly_2013_2019_Year ,select=-
Group.1)
gDPYearly_2013_2019_Collated
#####
combinedValuesGDP_Flights <-
cbind(totalFlightsCommNew,gDPYearly_2013_2019_Collated)
combinedValuesGDP_Flights
```

```
###################### Line Plots to Show GDP and Number of Flights
############

########################### Below Code is inspired from
########################### https://www.r-graph-gallery.com/line-chart-dual-Y-axis-
ggplot2.html


jpeg("FligtVsGDP.jpg")
coeff <- 100000
valueColor <- "red"
xColor <- rgb(0.2, 0.6, 0.9, 1)
ggplot(combinedValuesGDP_Flights, aes(x=variable)) +
  geom_line( aes(y=value),size=1, color=valueColor) +
  geom_line( aes(y=x/coeff),size=1, color=xColor) + # Divide by 10 to get the same range
than the temperature
  scale_x_continuous(
   name = "Years (2013-2019)               "
  ) +#geom_point(aes(y)size=4, shape=21)
  scale_y_continuous(
    name = "Flights in Thousand",
    # Add a second axis and specify its features
    sec.axis = sec_axis(~.*coeff, name="GDP Million USD")
  )+
  theme_ipsum() +theme(
    axis.title.y = element_text(color = valueColor, size=13),
    axis.title.y.right = element_text(color = xColor, size=13)
  ) +ggtitle("Increase in GDP and Flights over the years")
dev.off()


##Aggregating All the passengers from All years
totalFlightsGroupCountry<-
aggregate(yearlyArrivalFlightsComm[,4:10],by=list(yearlyArrivalFlightsComm$SourceCou
ntries,yearlyArrivalFlightsComm$DestinationCity), FUN=sum)
totalFlightsGroupCountry

##############################
### Callling ggplot Function from source file

jpeg("TotalFlightsIreland.jpg")

ggplot(totalFlightsCommNew, aes(x=variable,y=value))+geom_bar(stat =
"identity",width=0.7,alpha = 1,colour="#CC79A7",fill="blue")+
  theme(axis.line = element_line(colour = "darkblue",size = 1, linetype = "solid"),
        axis.text.x = element_text(angle = 90, vjust =0.5 ,hjust =1))+
  xlab("Years") + ylab("Number of Flights in Thousands") +ggtitle("Flights to Ireland
between 2013 and 2019")+guides(fill=guide_legend(title=""))
```

```
dev.off()

#### Summing
totalFlightsAirport<-
aggregate(yearlyArrivalFlightsComm[,4:10],by=list(yearlyArrivalFlightsComm$SourceCou
ntries,yearlyArrivalFlightsComm$DestinationCity), FUN=sum)
totalFlightsAirport

##########################################################################
########
# Subseting Arrivals as per passengers coming in Ireland
yearlyArrivalPassenger<-subset(yearlyArrivalsFlights, Category == "Passengers
(Thousand)")
YearlyArrivalPassenger

totalPassengPerAirport<-
aggregate(yearlyArrivalPassenger[,4:10],by=list(yearlyArrivalPassenger$Category,yearly
ArrivalPassenger$SourceCountries), FUN=sum)
totalPassengPerAirport

### Using melt function to melt years
tPA_New<- melt(totalPassengPerAirport)
tPA_New

### Plotting passengers per Airport
jpeg("TotalPassengersArrivedIreland.jpg")

ggplot(tPA_New, aes(x=Group.2,y=value,fill=variable))+
  geom_bar(stat = "identity",width=0.7,alpha = 1,colour="blue")+
  theme(axis.line = element_line(colour = "darkblue",size = 1, linetype = "solid"),
        axis.text.x = element_text(angle = 90, vjust =0.5 ,hjust =1))

dev.off()


##########################################################################
##############################################

## GDP dataset contain only ISO3 country names whereas Flight contain country name
inorder to read combined model
#For Loop was created which checked ISO 3 country codes with Corresponding Country
name and added a new Column in dataframe
countryCodes<- read.csv(file="country-codes.csv", head=TRUE, sep=",")
countryCodes

### Grouping GDP by Countries
combinedtopTenGDPYearlyAgg<-
aggregate(combinedtopTenYearly$Value,by=list(combinedtopTenYearly$TIME,combine
dtopTenYearly$ï..LOCATION), FUN=sum)
as.data.frame(combinedtopTenGDPYearlyAgg)
```

```
##### We have ISO3 values for Countries to get Full name  compared it to file
CountryCodes to get full Country name
# Created new Variable Country in the file assigning it a NA
combinedtopTenGDPYearlyAgg$Country=NA
combinedtopTenGDPYearlyAgg

#Assign index1 to number of rows
indexes1 <- nrow(combinedtopTenGDPYearlyAgg)
indexes1
indexes2<-nrow(countryCodes) #Index2 to row of country codes
indexes2
#options(warn=-1)
options(warn=0)
# Iterate over each row of
for(index1 in 1:indexes1) {
  # Iterate over each row of countrycode
  for(index2 in 1:indexes2) {
    if(combinedtopTenGDPYearlyAgg[index1,"Group.2"] ==
countryCodes[index2,"ISO3166.1.Alpha.3"]){

      combinedtopTenGDPYearlyAgg[index1,"Country"]<-
as.character(countryCodes[index2,"official_name_en"])
      break
    }
  }
}
combinedtopTenGDPYearlyAgg
#combinedtopTenGDPYearlyPlot<-select(combinedtopTenYearlyAgg,-c(2))
#combinedtopTenGDPYearlyPlot

##### Creating a model from Top ten GDP and Flights from Countries
##Sourced from above code contains Flight information
totalFlightsGroupCountry

#### Transposed Data to Bind the Two different Datasets
toFlightsGroupCountryTransposed<-melt(totalFlightsGroupCountry)
toFlightsGroupCountryTransposed

### Here the Variable which is Year is a Factor as it reads x2013 and so on.
## Removing character values from Year as this is how File is read from CSV files
toFlightsGroupCountryTransposed$variable <- as.integer(gsub('[a-zA-Z]', '',
toFlightsGroupCountryTransposed$variable))
toFlightsGroupCountryTransposed
#Aggregating Flights by year and Country
toFlightsGroupCountryTransposedAgg<-
aggregate(toFlightsGroupCountryTransposed$value,by=list(toFlightsGroupCountryTrans
posed$variable,toFlightsGroupCountryTransposed$Group.1), FUN=sum)
toFlightsGroupCountryTransposedAgg
```

```
### Combined Model Creating a loop

combinedModel<-combinedtopTenGDPYearlyAgg ## created using For loop assigning it
to new variable
combinedModel

combinedModel$numberOfFlights=NA #Creating a new Column
combinedModel

indexes1 <- nrow(combinedModel)
indexes1
indexes2<-nrow(toFlightsGroupCountryTransposed)
indexes2
#options(warn=-1)
options(warn=0)
# Iterate over each row of
for(index1 in 1:indexes1) {
  # Iterate over each row of countrycode
  for(index2 in 1:indexes2) {
    if(combinedModel[index1,"Country"] ==
toFlightsGroupCountryTransposed[index2,"Group.1"] &
       combinedModel[index1,"Group.1"] ==
toFlightsGroupCountryTransposed[index2,"variable"]){

      combinedModel[index1,"numberOfFlights"]<-
as.numeric(toFlightsGroupCountryTransposed[index2,"value"])
      break
    }
  }
}
combinedModel #Model Contains all the Top ten countries GDP and Flight over 2013-19

### Selecting Countries whose Flight and GDP data is available
combinedModelGDPFlight<-combinedModel[complete.cases(combinedModel), ]
#Selecting only those who have gdp and flights to Ireland
combinedModelGDPFlight

######SELECTING COUNTRIES with Available GDP and Flight info
combinedModelGDPFlight<-combinedModelGDPFlight[c(1:14,(18:24),(33:39)),]
combinedModelGDPFlight
### Changing GDP to miilion
combinedModelGDPFlight$GDPMillion<-combinedModelGDPFlight$x/1000000
combinedModelGDPFlight

### Flights to Ireland By Country of Residence

jpeg("TopFourGDPCountry.jpg")
```

```
ggplot(combinedModelGDPFlight,aes(x=as.numeric(Group.1),y=numberOfFlights,fill=Co
untry))+
  geom_bar(stat='identity', position='dodge2')+theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) + labs(
    title = 'Flights to Ireland by Country of residence 2013-2019',
    x = '')
dev.off()

jpeg("TopFourFlightCountry.jpg")
ggplot(combinedModelGDPFlight)+geom_bar(aes(reorder(Group.1,-
GDPMillion),GDPMillion,fill=Country),
                                        stat='identity', position='dodge2')+theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) + labs(
    title = 'GDP of residence Countries 2013-2019', x = '')
dev.off()

####Subsetting USA for Linear Regression
combinedModelGDPFlight_USA<-combinedModelGDPFlight[22:28,]
combinedModelGDPFlight_USA

###### Germany Subset
combinedModelGDPFlight_Germany<-combinedModelGDPFlight[1:7,]
combinedModelGDPFlight_Germany

####EU28 Subset
combinedModelGDPFlight_EU28<-combinedModelGDPFlight[7:14,]
combinedModelGDPFlight_EU28

###Summary of GDP and Number of Flights by Countries
summaryOfCountries<-by(combinedModelGDPFlight,
combinedModelGDPFlight$Country, summary)
summaryOfCountries

###Mean of GDP by Country############
tapply(combinedModelGDPFlight$GDPMillion,combinedModelGDPFlight$Country,mean)

###Mean of Numberof flights by Country
tapply(combinedModelGDPFlight$numberOfFlights,combinedModelGDPFlight$Country,
mean)

###################### Correlation Model ##################

# Shapiro-Wilk normality test for numberOfFlights
shapiro.test(combinedModelGDPFlight_USA$numberOfFlights) # => p = 0.6481

# Shapiro-Wilk normality test for USA$GDPMillion
shapiro.test(combinedModelGDPFlight_USA$GDPMillion) # => p = 0.9338

###From the output, the two p-values are greater than the significance level 0.05
implying that the distribution of the
```

```
#data are not significantly different from normal distribution. In other words, we can
assume the normality.


###Checking normality of datasets
# GDPMillion
jpeg("NormalityNoFlights.jpg")
ggqqplot(combinedModelGDPFlight_USA$GDPMillion, ylab = "Number of
Flights(Thousand)")
dev.off()

# GDPMillion
jpeg("NormalityGDP.jpg")
ggqqplot(combinedModelGDPFlight_USA$GDPMillion, ylab = "GDP (Million USD)")
dev.off()

#The plots shows that the Flight data may be  normal distributions.

###Correlation Test for USA Data

res <- cor.test(combinedModelGDPFlight_USA$numberOfFlights,
combinedModelGDPFlight_USA$GDPMillion,
            method = "pearson")
res

jpeg("CorrelationModel.jpg")
ggscatter(combinedModelGDPFlight_USA, x = "numberOfFlights", y = "GDPMillion",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Number of Flights(Thousand)", ylab = "GDP (Million USD)")
dev.off()

### Measure correlation (r) between All selected Contries Flights and GDP
r <- by(combinedModelGDPFlight,combinedModelGDPFlight$Country, FUN = function(X)
cor(X$numberOfFlights, X$GDPMillion, method = "pearson"))
r
r2 <- r*r    # calculate r squared
r2




##########################To forecast Flights we now do a linear regression Model

##### Creating a Linear Regression Model for each Country using by function .This will
provide y=a+bx for each country
models <- by(combinedModelGDPFlight, combinedModelGDPFlight$Country,
function(df) lm(numberOfFlights~GDPMillion, data=df))
models
```

```
### Confidence Interval of Each Country
lapply(models, confint)


### European Union 28 Intercept and Slope
#y1=a+bx
y1=-4.645+5.117*3.6 # x(independent variable = 3.6 assumed)
y1

### Germany Intercept and Slope
#y2=a+bx
y2=-7.549+3.595 *3.6 # x(independent variable = 3.6 assumed)
y2

### United Kingdom Intercept and Slope
y3=17.96+12.69 *3.6 # x(independent variable = 3.6 assumed)
y3


###United States of America
y4=-10.8004+0.9505 *3.6 # x(independent variable = 3.6 assumed)
y4


####Linear Regression for USA  using  ggplot2

jpeg("LinearRegressionModel.jpg")
plot(combinedModelGDPFlight_USA$GDPMillion,
combinedModelGDPFlight_USA$numberOfFlights, pch = 16, cex = 1.3, col = "blue",
     main = "GDP vs Number of Flights", xlab = "GDP USA (Million)", ylab = "Number of
Flights (Thousand)")
ablineclip(lm(combinedModelGDPFlight_USA$numberOfFlights ~
combinedModelGDPFlight_USA$GDPMillion), col="orange",x17=1,x2=25) # show
regression line
dev.off()

## Predicted values for seven years in the dataset
predict(lm(combinedModelGDPFlight_USA$numberOfFlights ~
combinedModelGDPFlight_USA$GDPMillion))

#predicted values for Flights from USA if GDP falls because of Coronavirus Pandemic
flightsInIreland <- data.frame(GDP_Thousand<- c(3.5,4.8,5.5,6.2,6.8,7.2,7.8))
flightsInIreland
predict(lm(combinedModelGDPFlight_USA$numberOfFlights ~
combinedModelGDPFlight_USA$GDPMillion), flightsInIreland)

############## BarPlot and LinePlot  ############
########################## Below Code is inspired from
############################ https://www.r-graph-gallery.com/line-chart-dual-Y-axis-
ggplot2.html
```

```
jpeg("USALMModel.jpg")
valueColor <- "#69b3a2"
coeff <- 10000
xColor <- "blue"
ggplot(combinedModelGDPFlight_USA)+
geom_bar( aes(x=Group.1,y=numberOfFlights),stat="identity", fill=valueColor,size=.5,
color="black", alpha=.6) +
geom_line(aes(x=Group.1,y=GDPMillion),size=1, color=xColor)+
  scale_x_continuous(
    name = "Years (2013-2019)"
  ) +
  scale_y_continuous(
    name = "Flights in Thousand",
    # Add a second axis and specify its features
    sec.axis = sec_axis(~.*coeff, name="GDP Million USD")
  )+theme_ipsum() +theme(
    axis.title.y = element_text(color = valueColor, size=13),
    axis.title.y.right = element_text(color = xColor, size=13)
  ) +ggtitle("USA Model -Increase in GDP and Flights")
dev.off()

################################

jpeg("GermanyLMModel.jpg")
valueColor <- "#ffb653"
coeff <- 10000
xColor <- "blue"
ggplot(combinedModelGDPFlight_Germany)+
  geom_bar( aes(x=Group.1,y=numberOfFlights),stat="identity", fill=valueColor,size=.5,
color="black", alpha=.6) +
  geom_line(aes(x=Group.1,y=GDPMillion),size=1, color=xColor)+
  scale_x_continuous(
    name = "Years (2013-2019)"
  ) +
  scale_y_continuous(
    name = "Flights in Thousand",
    # Add a second axis and specify its features
    sec.axis = sec_axis(~.*coeff, name="GDP Million USD")
  )+theme_ipsum() +theme(
    axis.title.y = element_text(color = valueColor, size=13),
    axis.title.y.right = element_text(color = xColor, size=13)
  ) +ggtitle("Increase in GDP and Flights over the years")
dev.off()

################################
```

6. Appendix 2 – Variability in Rainfall influences on Crop Production - R programming-Codes

```
getwd()
install.packages("ggplot2")
library(ggplot2)
install.packages("scales")
library(scales)
install.packages("gapminder")
library(gapminder)
install.packages("gganimate") #For animated Plots
library(gganimate)
install.packages("plotly")
library(plotly)
install.packages(reshape2) #Used to melt Datframe instead pf melt package
library(reshape2)
install.packages("data.table")
library(data.table)
install.packages("gapminder") #animated plots
library(gapminder)
install.packages("gganimate") #animated plots
library(gganimate)
install.packages("plotly")
library(plotly)
install.packages("melt")
library(melt)
install.packages("dplyr")
library(dplyr)
install.packages("patchwork")
library(patchwork) # To display 2 charts together
install.packages("hrbrthemes")
library(hrbrthemes)
install.packages("impute")
library(impute)
install.packages("plotrix")
library(plotrix )## To Clip abline
install.packages("cvequality") ### To read Coefficient of Variation
library(cvequality)
library("ggpubr")

source('~/functionFile.R') ##Sourcing Function File into Main file

##############################################

# This Project is done with objective to study rainfall patterns in India and it's affect on Crop Production

# Through study we will try to findout if there is a correlation between amount of crop production and variability in rainfall
```

Studies on Crops which needs more water to grow

Data for Crops was sourced at http://www.fao.org/faostat/en/#data/PP

Data for Rainfall was sourced at
https://climateknowledgeportal.worldbank.org/download-data

Rainfall study was referenced through
https://www.researchgate.net/publication/270585608_Influences_of_rainfall_on_crop_production_and_suggestions_for_adaptation

Correlation Model was referenced at http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

##

reading types of crops and there production in Hectare and Production in Tonnes between Year 1996 till 2016
produceDescrip<- read.csv(file="ProductionCropsAllData.csv", head=TRUE, sep=",")
str(produceDescrip)
nrow(produceDescrip) #### number of rows =1048575
complete.cases(produceDescrip)
# Checking for NA
produceDescrip[!complete.cases(produceDescrip),]

There are incomplete cases however this will be dealt if required data is not there in selected model
produceDescrip<-subset(produceDescrip, Year >= 1996 & Year <= 2016 & Area == "India" & Element =="Production")
produceDescrip

###Subsetting Crops which requires more water
produceDescripSub<-produceDescrip[produceDescrip$Item %in% c('Wheat','Maize','Cotton lint','Bananas'),]
produceDescripSub

##Checking for incomplete rows
produceDescripSub[!complete.cases(produceDescripSub),]
There are no incomplete cases in selected model hence no need to further clean data

##Descriptive Statistics
Aggregate of Produce Finding mean Production each year
agg_Crop_MeanYear<-aggregate(produceDescripSub$Value, by=list(produceDescripSub$Item,produceDescripSub$Year), FUN=mean)
agg_Crop_MeanYear

agg_Crop_Mean<-aggregate(produceDescripSub$Value, by=list(produceDescripSub$Item), FUN=mean)

```
agg_Crop_Mean

## Aggregate of Produce Finding min Production in selected year
agg_Crop_Min<-aggregate(produceDescripSub$Value,
by=list(produceDescripSub$Item), FUN=min)
agg_Crop_Min

agg_Crop_Max<-aggregate(produceDescripSub$Value,
by=list(produceDescripSub$Item), FUN=max)
agg_Crop_Max



################################################################################
#################################


##### reading Rainfall in each country
rainfallData<- read.csv(file="RainfallCountries.csv", head=TRUE, sep=",")
rainfallData

rainfallData <- subset(rainfallData, Country==" India" & Year >= 1996 & Year <= 2016)
str(rainfallData)


nrow(rainfallData) #### number of rows =387875
complete.cases(rainfallData)
# Checking for NA
rainfallData[!complete.cases(rainfallData),]
### There are no incomplete cases or NAs in the data
summary(rainfallData)
str(rainfallData)
nrow(rainfallData)

############## Histogram of Rainfall in 10 years
jpeg("rainfallplot.jpg")
hist(rainfallData$Rainfall.MM.,120,col="red",pch=19,main="Rainfall
Trends",xlab="Years", ylab="Rainfall in MM.", xaxt='n')
dev.off()


############ Pltting scatter plot for Rainfall
jpeg("scatterplotrainfall.jpg")
plot(rainfallData$Rainfall.MM.,col="Blue",main="Rainfall Data",xlab="Years",
ylab="Rainfall in MM.", xaxt='n',pch=19)
dev.off()



##AnnualRainfall per Year
```

```
sumMean1995_2016<-sum(tapply(rainfallData$Rainfall.MM.,rainfallData$Year,mean))
sumMean1995_2016
sumStdDev1995_2016<-sum(tapply(rainfallData$Rainfall.MM.,rainfallData$Year,sd))
sumStdDev1995_2016
sumMin1995_2016<-sum(tapply(rainfallData$Rainfall.MM.,rainfallData$Year,min))
sumMin1995_2016
sumMax1995_2016<-sum(tapply(rainfallData$Rainfall.MM.,rainfallData$Year,max))
sumMax1995_2016
sumCv1995_2016<-sum(tapply(rainfallData$Rainfall.MM.,rainfallData$Year,cv))
sumCv1995_2016



#####

###Seasonal Rainfall during the month from June till October
SeasonalRains<-subset(rainfallData,  Statistics==" Jun Average" |Statistics== " Jul
Average" |
                      Statistics==" Aug Average"| Statistics==" Sep Average" | Statistics=="
Oct Average")
SeasonalRains


## Aggregate of Rainfall For each Seasonal Month Groupby Year
agg_Rainfall<-aggregate(SeasonalRains$Rainfall.MM., by=list(SeasonalRains$Year),
FUN=mean)
agg_Rainfall
d<-agg_Rainfall
n<-4
rainfallDF<-do.call("rbind", replicate(n, d, simplify = FALSE))
rainfallDF
rainfallDF<-rainfallDF[order(rainfallDF$Group.1),]
rainfallDF




#Seasonal Rainfall  Plot between 1995-2013 month from June till October
jpeg("Rainfall Pattern during Seasonal Rains(June-October).jpg")
valueColor <- "Black"
xColor <- rgb(0.2, 0.6, 0.9, 1)
ggplot(agg_Rainfall,aes(x=Group.1,y=x))+geom_line(colour="Blue")+geom_point()+
geom_point(size=1, shape=18,color="Blue")+
  theme_ipsum() +theme(
    axis.title.y = element_text(color = valueColor, size=13),
    axis.title.y.right = element_text(color = xColor, size=13),
    axis.text.x = element_text(angle = 0, vjust =0.5 ,hjust = 0)
  ) +xlab("Years") + ylab("Rainfall in MM ") +ggtitle("Average Seasonal Rainfall (June-
October)")
dev.off()
```

```
###Combining Mean Crop Production and Rainfall Data

combinedModelCropRain<- cbind(agg_Crop_MeanYear,rainfallDF)
combinedModelCropRain
##Changing Column names of dataframe
colnames(combinedModelCropRain) <- c("Country", "Year",
"Production(HA.)","Year","Rainfall Average (MM)")
combinedModelCropRain

#### Aggregate of Year and Country
combinedModelCropRainAgg<-
aggregate(combinedModelCropRain$`Production(HA.)`,by=list(combinedModelCropRain
$Country,combinedModelCropRain$Year,combinedModelCropRain$`Rainfall Average
(MM)`), FUN=sum)
combinedModelCropRainAgg<-
combinedModelCropRainAgg[order(combinedModelCropRainAgg$Group.2),]
combinedModelCropRainAgg

####Converting Tonnes to Mega Tonnes
combinedModelCropRainAgg$MegaTonnes<-combinedModelCropRainAgg$x/100000
combinedModelCropRainAgg

## Writing the input in CSV File
writeToFile (combinedModelCropRainAgg, "CombinedModelCrop&Rain.csv")
writeToFile <- function(topFiveCountry, outputFilename) {
  write.table(topFiveCountry, file=outputFilename, sep=",",append=F)
}


##### Plotting Crop Production and Rainfall Pattern during Seasonal Rains(June-
October)

jpeg("Crop Production and Rainfall Pattern during Seasonal Rains(June-October).jpg")
coeff3 <-10
ggplot(combinedModelCropRainAgg)+geom_bar( aes(x=Group.2,y=MegaTonnes,fill=
Group.1),stat="identity",size=.5, alpha=.6,position
=position_dodge(width=0.8),colour="black") +
geom_line(aes(x=Group.2,y=Group.3*coeff3,),size=1,
color=xColor)+geom_point(aes(x=Group.2,y=Group.3*coeff3),size=3,
shape=18,color="Blue")+
scale_x_continuous(name = "Years ") +scale_y_continuous(name = "Crop Production
Mega Tonnes",  # Add a second axis and specify its features
sec.axis = sec_axis(~./coeff3, name="Rainfall in MM"))+theme_ipsum()
+theme(axis.title.y = element_text(color = "darkblue", size=13),
axis.title.y.right = element_text( size=13)) +ggtitle("Crop production and Rainfall")+
guides(fill=guide_legend(title=""))
dev.off()
```

```
########################################################  Correlation Between Crop and Rainfall ########################

### Correlation Model was referenced at http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
# Shapiro-Wilk normality test for Crops

shapiro.test(combinedModelCropRainAgg$MegaTonnes) # => p = 2.809e-09

# Shapiro-Wilk normality test for Data Rainfall
shapiro.test(combinedModelCropRainAgg$Group.3) # => p = 0.04603

###From the output, the two p-values are less than the significance level 0.05 implying that the distribution of the
#data  for Crops are  significantly different from normal distribution. In other words, we can't assume the normality.


###Checking normality of datasets
# MegaTonnes
jpeg("NormalityProduction.jpg")
ggqqplot(combinedModelCropRainAgg$MegaTonnes, ylab = "PRoduction (Mega Tonnes)",title = "Shapiro-Wilk normality test  Crop Produce")
dev.off()

# Rainfall
jpeg("NormalityRainFall.jpg")
ggqqplot(combinedModelCropRainAgg$Group.3, ylab = "GDP (Rainfall in MM.)",title = "Shapiro-Wilk normality test  Rainfall")
dev.off()

#The plots shows that the data may not be  normally distributed.


###Correlation Test for Crop Data
### Correlation Model was referenced at http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
res1 <-
cor.test(combinedModelCropRainAgg$MegaTonnes,combinedModelCropRainAgg$Group.3,
             method = "pearson")
res1
##Measure of correlation between crops and
r2 <-res1*res1    # calculate r squared
r2
#The p-value of the test is 0.5174, which is more than the significance level alpha = 0.05.
```

#We can conclude that production and rainfall are not significantly correlated with a correlation coefficient of 0.07161444 and p-value of 0.5174 .

```
jpeg("Correlation between Crops and Rainfall.jpg")
ggscatter(combinedModelCropRainAgg, x = "MegaTonnes", y = "Group.3",add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",xlab = "Production in MegaTonnes", ylab = "Rainfall in MM.",title="Correlation between Crop Production and Rainfall")
dev.off()
```

7. Appendix 3 – Functions

```
##### This File is made for the functions which are repeatedly  used in Main programs

#### Finding Descriptive Statistics
descStats <- function(stats) {
  library(moments) # required for skewness
  newMatrix <- matrix (1:7, nrow = 1)
  colnames(newMatrix) <- c("Mean", "Median", "Variance", "Minimum", "Maximum", "Skewness","Sum")
  rownames(newMatrix) <- "Descriptive Statistics"
  newMatrix[1,] <- c(mean(stats),median(stats),var(stats),min(stats),max(stats),skewness(stats),sum(stats))
  return(newMatrix)
}

#### Finding Highest and Lowest Performing Countries
findingStats<-function(x,y){
  newMatrix <- matrix (1:2, nrow = 1)
  colnames(newMatrix) <- c("High Performing", "Low Performing")
  rownames(newMatrix) <- "GDP Countries"
  a<-x[which.max(as.numeric(y))]
  b<-x[which.min(as.numeric(y))]
  r1<-paste(a)
  r2<-paste(b)
  newMatrix[1,]<- c(r1,r2)
  return(newMatrix)
}


#### F1#####Function to calculate ggplot2(barplot)
ggplotFunc<-function(a1,a2,a3,a4){
     ggplot(a1, aes(x=a2,y=a3,fill =a2))+geom_bar(stat = "identity",width=.5,fill = "steelblue",alpha = 1,colour="black")+
     theme(axis.line = element_line(colour = "darkblue",size = 1, linetype = "solid"),axis.text.x = element_text(angle = 45, vjust =0 ,hjust = 0))+
```

```
     xlab("Countries") + scale_y_continuous(name = "Million (USD)",
                                            labels = function(y) y / 1000000)+
ggtitle(a4)+guides(fill=guide_legend(title=""))

}


###################F2###### Top 5 GDP####

#### Function to calculate ggplot2(barplot) for Passengers




ggplotFuncFlights<-function(a1,a2,a3,a4,a5,a6){

  ggplot(a1, aes(x=a2,y=a3,fill =a3))+geom_bar(stat = "identity",width=0.7,fill =
factor(x),alpha = 1,colour="black")+
    theme(axis.line = element_line(colour = "darkblue",size = 1, linetype = "solid"),
          axis.text.x = element_text(angle = 90, vjust =0.5 ,hjust =1))+
    xlab(a5) + ylab(a6)
+ggtitle(a4)+scale_x_discrete(labels=c("2013","2014","2015","2016","2017","2018","2019
"))
}



#### Function to calculate ggplot2 ScatterPlot
ggplotFuncScatter<-function(a1,a2,a3,a4){
        ggplot(data = a1, mapping = aes(x = a2, y = a3, color = a3)) +
          layer(geom = "a4")+ scale_color_gradient()
}
  ########################

ggplotFuncGDP<-function(a1,a2,a3,a4,a5,a6,a7,a8){
  ggplot(a1, aes(x=a2,y=a3,fill=factor(a4)))+geom_bar(stat = "identity",width=.8,alpha =
1,position =position_dodge(width=0.8))+
    theme(axis.line = element_line(colour = "darkblue",size = 1, linetype = "solid"),
          axis.text.x = element_text(angle = 90, vjust =0 ,hjust = 0))+
    xlab(a5) + scale_y_continuous(name = "Million (USD)",
                                  labels = function(y) y / 1000000)+
  ggtitle(a7)+ guides(fill=guide_legend(title=a8))+
    theme(axis.title.y = element_text(color = "darkblue", size=13),
    axis.title.y.right = element_text( size=13))

  #xlim(0,NA)
    #scale_y_discrete(limit = c("5", "10", "15","20"))

}
```

```
############## Finding five top GDP

topTenGDP<-function(GDPyear,GDPValue){
  GDPyear <- GDPyear[order(-GDPValue),]
  newFile<-GDPyear[1:10,]
  return(newFile)

}

############## Writting a csv file for List of countries with highest GDP


writeToFile <- function(topFiveCountry, outputFilename) {
  write.table(topFiveCountry, file=outputFilename, sep=",",append=F)
}



############################## Crops Project Functions ##################
```

8. Bibliography

Central Statistics Office www.statbank.cso.ie/ [online]. Available at: https://statbank.cso.ie/px/pxeirestat/Statire/SelectVarVal/saveselections.asp

OECD, www. oecd.org/ [online]. Available at: https://data.oecd.org/gdp/gross-domestic-product-gdp.htm#indicator-chart

Food and Agriculture Organization of the United Nations, www.fao.org/ [online]. Available at: http://www.fao.org/faostat/en/#data/PP

World Bank Group, climateknowledgeportal.worldbank.org [online]. Available at: https://climateknowledgeportal.worldbank.org/download-data

GitHub,www.Github.com[online]. Available at: https://github.com/datasets/country-codes

STHDA,www.sthda.com[online].Available at: http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

r-graph-gallery.com [online]. Available at: https://www.r-graph-gallery.com/line-chart-dual-Y-axis-ggplot2.html

Irish Post, Beresford J.(March,2019).”*Irelands top rated destinations on TripAdvisor have been revealed”.* [online]. Available at: https://www.irishpost.com/news/irelands-top-10-rated-destinations-tripadvisor-revealed-165565

Ciesluk K.(2019) “*GDP Growth is a key Driver of Air Travel Demand” ”.* [online]. Available at: https://simpleflying.com/gdp-driver-air-travel-demand/

KFW IPEX Bank.(2017), www.kfw-ipex-bank.de/ [online]. Available at: https://www.kfw-ipex-bank.de/International-financing/KfW-IPEX-Bank/Analyses-and-Views/Market-analyses/GDP-growth-and-airline-passengers/

Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler), CRISP-DM 1.0 ,

Crop Selection [online]. Available at https://sswm.info/sswm-solutions-bop-markets/improving-water-and-sanitation-services-provided-public-institutions-0/crop-selection

Watanabe T., Ndamani F.(2015)”*Influences of rainfall on crop production and suggestions for adaptation [online]”*. Available at :

https://www.researchgate.net/publication/270585608_Influences_of_rainfall_on_crop_production_and_suggestions_for_adaptation

Conrad Kyei-Mensah,1 Rosina Kyerematen Hindawi,(2019)"*Impact of Rainfall Variability on Crop Production within the Worobong Ecological Area of Fanteakwa District, Ghana* " [online]. Available at :https://www.hindawi.com/journals/aag/2019/7930127/