# Economic Forecasting: Predicting World Economic Health

# CMPT 732 – Big Data Programming – I

## Contents

**1. Introduction……………………………………………………………………………………………………………………………1**

**2. Methodology…………………………………………………………………………………………………………………………1**

**2.1. Data Processing………………………………………………………………………………………………………2**

**2.2. Data Analysis…………………………………………………………………………………………………………2**

**2.3 Web Interface and Android App ……………………………………………………………………………3**

**3. Problems & Challenges……………………………………………………………………………………………………………3**

**3.1. Data Processing………………………………………………………………………………………………………3**

**3.2. Data Analysis…………………………………………………………………………………………………………3**

**3.3. User Interface and Visualization……………………………………………………………………………3**

**4. Results……………………………………………………………………………………………………………………………………4**

**APPENDIX – Project Summary……………………………………………………………………………………………………6**

# 1. Introduction

Economic Forecasting deals with trying to predict the future state of economies of countries around the world and the world as a whole using a formulated approach involving hundreds and hundreds of different parameters. These kinds of forecasts generally cater to economists who use these models to predict chief economic parameters like Growth Rate, Gross Domestic Product, Poverty Rates, Development, Human Development Index etc.

The problem with such forecasting is that they are generally meant for economists and are hard to understand in layman terms. We've attempted to resolve that by predicting several indicators using both conventional and unconventional parameters in a way that shall be understandable to most people.

We have obtained our data from myriad sources including but not limited to Web scraping, Big Query accesses, datasets from several international forums and economic websites like United Nations, IMF, World Bank etc. The data at 600MBs contains information of 193 countries and 10 regional groupings for around 100+ features from as far back as 1800s to 2017. The enormous amount of information has been scaled to our use and filtered extensively. We have faced extreme challenges of missing and incorrect data. After an extensive ETL process, the data was made suitable for our use.

We have divided the project into 3 modules:

1. **Global Economic Health Prediction**: Prediction of Growth rates of countries around the world and classification of these rates into 4 categories:
   a. Booming Economies (Growth Rate > 5%) (India, China, Philippines)
   b. Healthy Economies (Growth Rate > 3%) (Ireland, Egypt, Turkey)
   c. Stagnant Economies (Growth Rate > 0%) (Greece, France, Germany)
   d. Declining Economies (Growth Rate < 0%) (Venezuela, North Korea)
2. **GDP Prediction**: The GDP Prediction forecasts for you the global GDP of countries using a host of parameters. The high accuracy that we obtained using GBT Regressor makes our results comparable to that of the World Bank and IMF.
3. **Big Mac Index**: The Big Mac Index is where we present to you a wacky indicator that can be used to state the Global disparity and the strength of certain currencies using mundane day to day items like in this case - A Big Mac Burger

# 2. Methodology

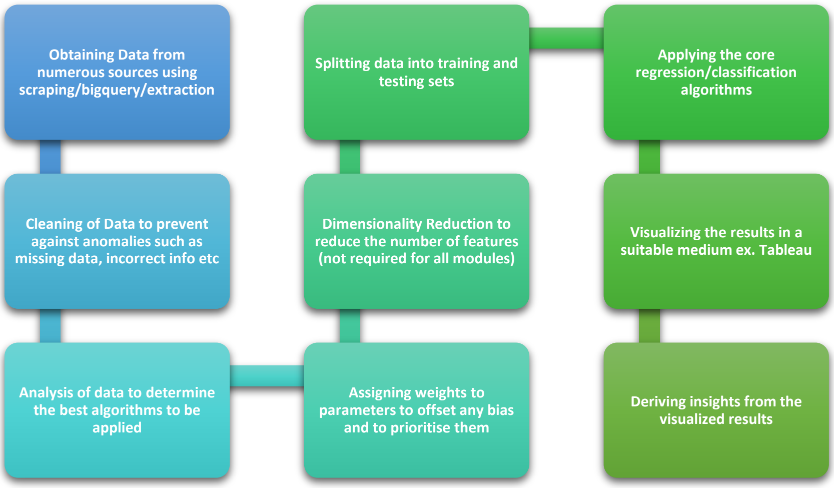
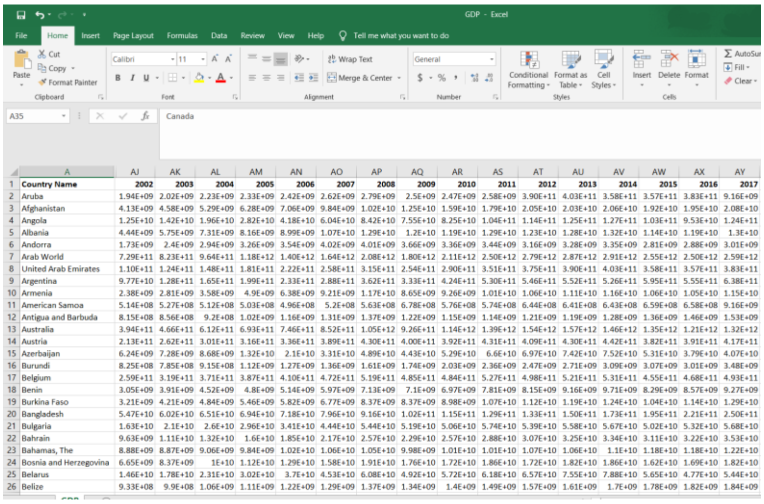

Obtaining Data from numerous sources using scraping/bigquery/extraction
Cleaning of Data to prevent against anomalies such as missing data, incorrect info etc
Analysis of data to determine the best algorithms to be applied
Assigning weights to parameters to offset any bias and to prioritise them
Dimensionality Reduction to reduce the number of features (not required for all modules)
Splitting data into training and testing sets
Applying the core regression/classification algorithms
Visualizing the results in a suitable medium ex. Tableau
Deriving insights from the visualized results

*Fig 2.1 - End-to-End Solution*

## 2.1 Data Processing

The net data that we collected was approximately 600+ MBs. This data ranged from csv files, json files, txt files and other unclassified information. The data was carefully crafted into csv files which were pivoted and edited to obtain a single uniform format of data. The final chosen format was csv files due to their versatility and ease of access using Spark, Python and R.



| | A | AJ | AK | AL | AM | AN | AO | AP | AQ | AR | AS | AT | AU | AV | AW | AX | AY |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Country Name | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
| 2 | Aruba | 1.94E+09 | 2.02E+09 | 2.23E+09 | 2.33E+09 | 2.42E+09 | 2.62E+09 | 2.79E+09 | 2.5E+09 | 2.47E+09 | 2.58E+09 | 3.90E+11 | 4.03E+11 | 3.58E+11 | 3.57E+11 | 3.83E+11 | 9.16E+09 |
| 3 | Afghanistan | 4.13E+09 | 4.58E+09 | 5.29E+09 | 6.28E+09 | 7.06E+09 | 9.84E+09 | 1.02E+10 | 1.25E+10 | 1.59E+10 | 1.79E+10 | 2.05E+10 | 2.03E+10 | 2.06E+10 | 1.92E+10 | 1.95E+10 | 2.08E+10 |
| 4 | Angola | 1.25E+10 | 1.42E+10 | 1.96E+10 | 2.82E+10 | 4.18E+10 | 6.04E+10 | 8.42E+10 | 7.55E+10 | 8.25E+10 | 1.04E+11 | 1.14E+11 | 1.25E+11 | 1.27E+11 | 1.03E+11 | 9.53E+10 | 1.24E+11 |
| 5 | Albania | 4.44E+09 | 5.75E+09 | 7.31E+09 | 8.16E+09 | 8.99E+09 | 1.07E+10 | 1.29E+10 | 1.2E+10 | 1.19E+10 | 1.29E+10 | 1.23E+10 | 1.28E+10 | 1.32E+10 | 1.14E+10 | 1.19E+10 | 1.3E+10 |
| 6 | Andorra | 1.73E+09 | 2.4E+09 | 2.94E+09 | 3.26E+09 | 3.54E+09 | 4.02E+09 | 4.01E+09 | 3.66E+09 | 3.36E+09 | 3.44E+09 | 3.16E+09 | 3.28E+09 | 3.35E+09 | 2.81E+09 | 2.88E+09 | 3.01E+09 |
| 7 | Arab World | 7.29E+11 | 8.23E+11 | 9.64E+11 | 1.18E+12 | 1.40E+12 | 1.64E+12 | 2.08E+12 | 1.80E+12 | 2.11E+12 | 2.50E+12 | 2.79E+12 | 2.87E+12 | 2.91E+12 | 2.55E+12 | 2.50E+12 | 2.59E+12 |
| 8 | United Arab Emirates | 1.10E+11 | 1.24E+11 | 1.48E+11 | 1.81E+11 | 2.22E+11 | 2.58E+11 | 3.15E+11 | 2.54E+11 | 2.90E+11 | 3.51E+11 | 3.75E+11 | 3.90E+11 | 4.03E+11 | 3.58E+11 | 3.57E+11 | 3.83E+11 |
| 9 | Argentina | 9.77E+10 | 1.28E+11 | 1.65E+11 | 1.99E+11 | 2.33E+11 | 2.88E+11 | 3.62E+11 | 3.33E+11 | 4.24E+11 | 5.30E+11 | 5.46E+11 | 5.52E+11 | 5.26E+11 | 5.95E+11 | 5.55E+11 | 6.38E+11 |
| 10 | Armenia | 2.38E+09 | 2.81E+09 | 3.58E+09 | 4.9E+09 | 6.38E+09 | 9.21E+09 | 1.17E+10 | 8.65E+09 | 9.26E+09 | 1.01E+10 | 1.06E+10 | 1.11E+10 | 1.16E+10 | 1.06E+10 | 1.05E+10 | 1.15E+10 |
| 11 | American Samoa | 5.14E+08 | 5.27E+08 | 5.12E+08 | 5.03E+08 | 4.96E+08 | 5.2E+08 | 5.63E+08 | 6.78E+08 | 5.76E+08 | 5.74E+08 | 6.44E+08 | 6.41E+08 | 6.43E+08 | 6.59E+08 | 6.58E+08 | 9.16E+09 |
| 12 | Antigua and Barbuda | 8.15E+08 | 8.56E+08 | 9.2E+08 | 1.02E+09 | 1.16E+09 | 1.31E+09 | 1.37E+09 | 1.22E+09 | 1.15E+09 | 1.14E+09 | 1.21E+09 | 1.19E+09 | 1.28E+09 | 1.36E+09 | 1.46E+09 | 1.53E+09 |
| 13 | Australia | 3.94E+11 | 4.66E+11 | 6.12E+11 | 6.93E+11 | 7.46E+11 | 8.52E+11 | 1.05E+12 | 9.26E+11 | 1.14E+12 | 1.39E+12 | 1.54E+12 | 1.57E+12 | 1.46E+12 | 1.35E+12 | 1.21E+12 | 1.32E+12 |
| 14 | Austria | 2.13E+11 | 2.62E+11 | 3.01E+11 | 3.16E+11 | 3.36E+11 | 3.89E+11 | 4.30E+11 | 4.00E+11 | 3.92E+11 | 4.31E+11 | 4.09E+11 | 4.30E+11 | 4.42E+11 | 3.82E+11 | 3.91E+11 | 4.17E+11 |
| 15 | Azerbaijan | 6.24E+09 | 7.28E+09 | 8.68E+09 | 1.32E+10 | 2.1E+10 | 3.31E+10 | 4.89E+10 | 4.43E+10 | 5.29E+10 | 6.6E+10 | 6.97E+10 | 7.42E+10 | 7.52E+10 | 5.31E+10 | 3.79E+10 | 4.07E+10 |
| 16 | Burundi | 8.25E+08 | 7.85E+08 | 9.15E+08 | 1.12E+09 | 1.27E+09 | 1.36E+09 | 1.61E+09 | 1.74E+09 | 2.03E+09 | 2.36E+09 | 2.47E+09 | 2.71E+09 | 3.09E+09 | 3.07E+09 | 3.01E+09 | 3.48E+09 |
| 17 | Belgium | 2.59E+11 | 3.19E+11 | 3.71E+11 | 3.87E+11 | 4.10E+11 | 4.72E+11 | 5.19E+11 | 4.85E+11 | 4.84E+11 | 5.27E+11 | 4.98E+11 | 5.21E+11 | 5.31E+11 | 4.55E+11 | 4.68E+11 | 4.93E+11 |
| 18 | Benin | 3.05E+09 | 3.91E+09 | 4.52E+09 | 4.8E+09 | 5.14E+09 | 5.97E+09 | 7.13E+09 | 7.1E+09 | 6.97E+09 | 7.81E+09 | 8.15E+09 | 9.16E+09 | 9.71E+09 | 8.29E+09 | 8.57E+09 | 9.27E+09 |
| 19 | Burkina Faso | 3.21E+09 | 4.21E+09 | 4.84E+09 | 5.46E+09 | 5.82E+09 | 6.77E+09 | 8.37E+09 | 8.37E+09 | 8.98E+09 | 1.07E+10 | 1.12E+10 | 1.19E+10 | 1.24E+10 | 1.04E+10 | 1.14E+10 | 1.29E+10 |
| 20 | Bangladesh | 5.47E+10 | 6.02E+10 | 6.51E+10 | 6.94E+10 | 7.18E+10 | 7.96E+10 | 9.16E+10 | 1.02E+11 | 1.15E+11 | 1.29E+11 | 1.33E+11 | 1.50E+11 | 1.73E+11 | 1.95E+11 | 2.21E+11 | 2.50E+11 |
| 21 | Bulgaria | 1.63E+10 | 2.1E+10 | 2.6E+10 | 2.96E+10 | 3.41E+10 | 4.44E+10 | 5.44E+10 | 5.19E+10 | 5.06E+10 | 5.74E+10 | 5.39E+10 | 5.58E+10 | 5.67E+10 | 5.02E+10 | 5.32E+10 | 5.68E+10 |
| 22 | Bahrain | 9.63E+09 | 1.11E+10 | 1.32E+10 | 1.6E+10 | 1.85E+10 | 2.17E+10 | 2.57E+10 | 2.29E+10 | 2.57E+10 | 2.88E+10 | 3.07E+10 | 3.25E+10 | 3.34E+10 | 3.11E+10 | 3.22E+10 | 3.53E+10 |
| 23 | Bahamas, The | 8.88E+09 | 8.87E+09 | 9.06E+09 | 9.84E+09 | 1.02E+10 | 1.06E+10 | 1.05E+10 | 9.98E+09 | 1.01E+10 | 1.01E+10 | 1.07E+10 | 1.06E+10 | 1.1E+10 | 1.18E+10 | 1.18E+10 | 1.22E+10 |
| 24 | Bosnia and Herzegovina | 6.65E+09 | 8.37E+09 | 1E+10 | 1.12E+10 | 1.29E+10 | 1.58E+10 | 1.91E+10 | 1.76E+10 | 1.72E+10 | 1.86E+10 | 1.72E+10 | 1.82E+10 | 1.86E+10 | 1.62E+10 | 1.69E+10 | 1.82E+10 |
| 25 | Belarus | 1.46E+10 | 1.78E+10 | 2.31E+10 | 3.02E+10 | 3.7E+10 | 4.53E+10 | 6.08E+10 | 4.92E+10 | 5.72E+10 | 6.18E+10 | 6.57E+10 | 7.55E+10 | 7.88E+10 | 5.65E+10 | 4.77E+10 | 5.44E+10 |
| 26 | Belize | 9.33E+08 | 9.9E+08 | 1.06E+09 | 1.11E+09 | 1.22E+09 | 1.29E+09 | 1.37E+09 | 1.34E+09 | 1.4E+09 | 1.49E+09 | 1.57E+09 | 1.61E+09 | 1.7E+09 | 1.78E+09 | 1.82E+09 | 1.84E+09 |

*Fig 2.2 - Sample File*

By the end of cleaning process, we were left with 20 datasets of features from the years 1960-2017 for all countries. These datasets were then used for analysis. In certain cases (like the LifeExpectancy v/s IncomePerCapita graph) we traced back separate datasets to data since 1800s.

## 2.2 Data Analysis

Our Data Analysis has been performed in a scalable and efficient manner. We have used popular algorithms for analysis for all 3 modules. Primarily we have performed classification, regression, dimensionality reduction and several optimization algorithms. We have used R, Python and Spark machine learning libraries. Spark excels at iterative computation, enabling MLlib to run fast. At the same time, we care about algorithmic performance: MLlib contains high-quality algorithms that leverage iteration, and can yield better results.

For the Global Economy Health prediction, the dataset was scraped and obtained from different sources in a largely incoherent format. This data was cropped to obtain relevant parameters and converted to suitable datatypes. Using Spark with casting made this possible. Later the data was imported to Python module where it was run through Pearson Correlation Coefficient to obtain parameters with clear cut correlation to facilitate their removal and choosing appropriate parameters. Next dimensionality reduction using Principal Component Analysis was performed. The PCA analysis rendered the number of significant components at 3. These parameters were then passed though the Train_test_split module of ScikitLearn's model selection module. The distribution was split among Training, Validation and Test data. The data was then trained through Logistic Regression and Decision Trees. For Logisitic Regression we provided a binary classification and for Decision Trees a 4 way classification. The greater accuracy obtained for Decision Trees (65%) made them the modus operandi for choice. With the Growth Rate predictions done for every year(2010-17), we classified the results into categories and displayed them accordingly in a Tableau representation.

The Big Mac dataset was obtained from 'The Economist' which publishes the Big Mac Index every year. This data was filtered according to our needs and preprocessed accordingly. The base currencies selected were the US Dollar, Euro, British Pound, Japanese Yen and Chinese Yuan. The index was calculated by dividing the local price by the price of the base currency. Next, Linear Regression of GDP and Big Mac Index was taken in order to calculated which currency was undervalued and overvalued.

The GDP forecast analysis was done with data obtained from similar sources as the Growth Rate datasets. The data was also converted into a coherent format using typecasting and conversion of datatypes and substituting missing values. The data of the 5 base parameters was then loaded into spark through read_csv function and a common table constituting relevant factors was constructed using Spark SQL. For the prediction module, we constructed a schema. The data was loaded and taken through this schema and passed through a vector assembler which was imported using Spark Machine Learning library i.e. Spark.ml.feature and the module Vector Assembler along with StringIndexer and SQL Transformer. For the input to the regressor, we converted the country names into string indexer. The converted data which held the string indexed country names and features was then passed to the GBT Regressor.

The above GDP prediction was put into a pipeline and the model was trained upon the training data was again obtained using train_test_split with a ratio of 3:1. The data was then transformed. The output obtained was the GDP prediction with an accuracy of 87%.

### 2.3 Web Interface and Android App

We created a website (www.ecoforecast.ml) to provide an interactive interface for our project. The website contains Tableau dashboards embedded in it. These visualizations provided a dynamic outlook to our website and helps the user analyze the results in layman terms. We also created an Android application to better represent our data for the users who are more phone-savvy. The application displays the latest statistics for 47 countries along with predicted GDPs and other statistics.

## 3. Problems & Challenges

### 3.1 Data Processing

As mentioned above, extraction of data was done with great efforts since several underdeveloped countries do not publish their data annually. Moreover obtaining historical data was another cause of concern for most countries. The data quality in several cases was abysmal and had to be thoroughly revamped. Methods like scraping often fail when such varied data must be collected from hundreds of different websites.

### 3.2 Data Analysis

The data is highly diverse. Parameters like slum populations and high-tech exports have no correlation with each other but significantly affect the growth rate. Due to this, the right algorithm to predict the data is very difficult to find and accuracy obtained is seldom high. The same goes for the Decision Trees we used as their presented accuracy goes only till 60% even after using the parameters causing the greatest impact on the growth rate.

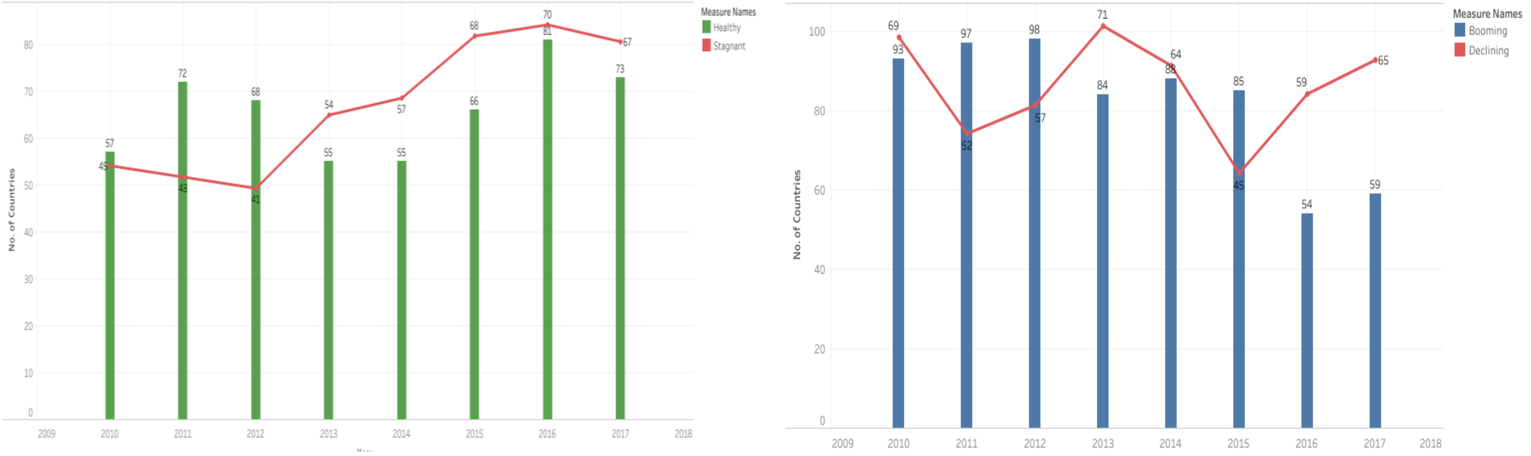
### 3.3 UI & Visualization

Android Studio does not possess Big Data libraries and hence we it becomes exceedingly difficult to embed interactive large scale visualizations onto the app. The Neural network libraries that are often not found compatible with popular platforms. Moreover handling large scale data using Tableau is efficient however heavy visualizations make websites heavier requiring excessive usage of already limited resources.

# 4. Results

The results obtained were displayed graphically as shown below:

1. Health of Global Economy: The graphs below represent how many countries were booming / Declining / Stagnant / Healthy economies. These graphs correspond to the actual growth estimates over the years for ex. The 2011 stock market crash.



*(a)*

*(b)*

*Fig 4.1 – (a) No. of Countries under Healthy & stagnant economy (b) No. of Countries under Booming & Declining economy*

Interactive animated graph made from parameters of Global Economic Health.

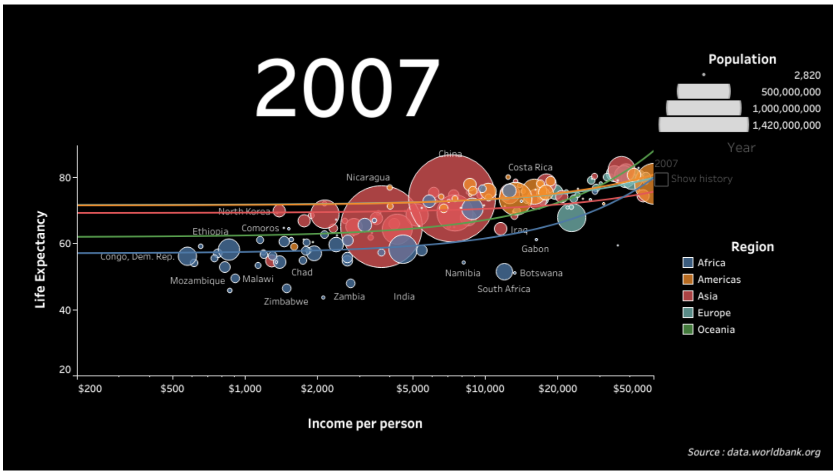
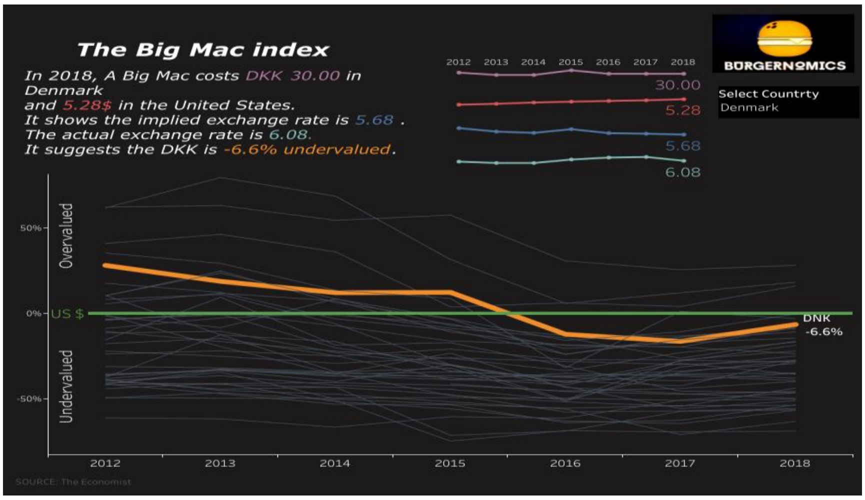


*Fig 4.2 – Income per person v/s Life expectancy*

2. The Big Mac Index: The below graph shows the currency valuations against dollar as the base rate. The graph also shows how these currencies have fared over the years.


The Big Mac index

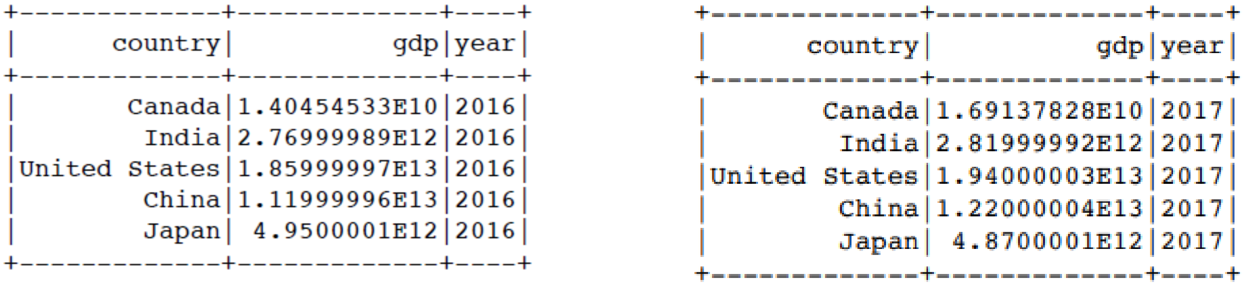
In 2018, A Big Mac costs DKK 30.00 in Denmark
and 5.28$ in the United States.
It shows the implied exchange rate is 5.68 .
The actual exchange rate is 6.08.
It suggests the DKK is -6.6% undervalued.

Select Countrty
Denmark

*Fig 4.3 – Currency valuations against dollar*

In the visualization above, we can see how Denmark's currency was valued above dollar till mid of 2015 following which it dropped and was since undervalued.

3. GDP forecast: The GDP forecast was calculated in a tabular format and fed to Tableau's interactive and animated visualization feature to calculate the variance over the years.



```
+-------------+-------------+----+
|      country|          gdp|year|
+-------------+-------------+----+
|       Canada|1.40454533E10|2016|
|        India|2.76999989E12|2016|
|United States|1.85999997E13|2016|
|        China|1.11999996E13|2016|
|        Japan| 4.9500001E12|2016|
+-------------+-------------+----+
```

```
+-------------+-------------+----+
|      country|          gdp|year|
+-------------+-------------+----+
|       Canada|1.69137828E10|2017|
|        India|2.81999992E12|2017|
|United States|1.94000003E13|2017|
|        China|1.22000004E13|2017|
|        Japan| 4.8700001E12|2017|
+-------------+-------------+----+
```

*Fig 4.4 – GDP prediction for 2016 & 2017*

The animated graph shows the increase in GDP gradually over several decades the last of which have been predicted and displayed.

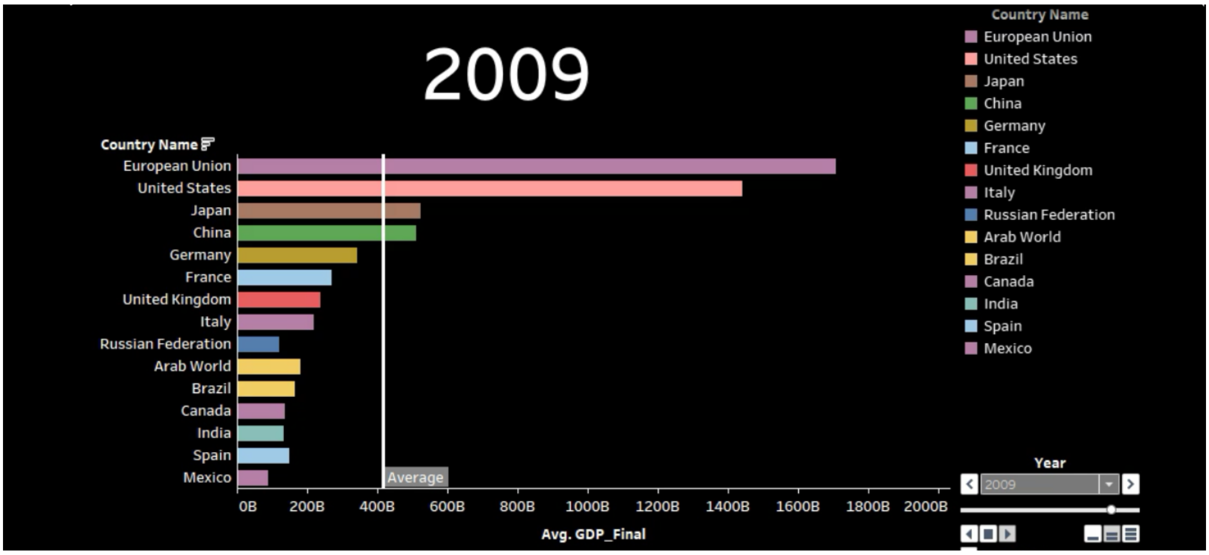


*Fig 4.5 – GDP over the years*

## APPENDIX – Project Summary

| Category | Points |
|---|---|
| **Getting the data**: Scraping data from various websites, using BigQuery to fetch data from Google cloud, datasets from UN,IMF,WB | 2 |
| **ETL:** **C**leaning the dataset and perform Extract-Transform-Load using Apache Spark and R. | 2 |
| **Algorithmic Work:** MLLib, Scikit to perform analysis and predict GDP. | 5 |
| **Bigness/Parallelization:** 600+ MB of data for 200 countries and groupings with 150+ features and 300,000+ values | 3 |
| **UI:** Developed an interactive website to display the results and embedded it with Tableau dashboards. | 2 |
| **Visualization:** Interactive and animated dashboards created for three features in Tableau and Bar Charts for classification of health of the economy. | 4 |
| **New Technology:** Using Android Studio developed an app to create an interactive medium for the user to see the results. | 2 |
| **Total** | 20 |