CA682 Data management and visualisation

Name	Ashish Poigal
Student Number	17211775
Programme	M.Sc. in Computing (Data Analytics)
Module Code	CA682
Assignment Title	Data Visualisation
Submission date	December 10, 2017
Module coordinator	Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations set out in the module documentation. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found recommended in the assignment guidelines.

Name: Ashish Poigal

Date: December 10, 2017

TABLE OF CONTENTS

1 INTRODUCTION

This study aims to investigate the impact of the Gross domestic product (GDP) on the Sports Capital Programme Allocations for various counties of Ireland between 2000 to 2016. GDP is a monetary measure of all final goods and services produced in a period and comprises personal consumption, business investment, government spending and net exports. Assuming GDP to have a direct influence on the sports capital allocated, we try to analyze the allocation pattern over time and for individual counties. As the saying goes, 'A picture is worth a thousand words,' a visual representation can be used to interpret this data efficiently. Bokeh, a python library designed for large and streaming datasets producing D3.js like interactive visualization, is used on open-data made available by the Department of Public Expenditure and Reform, Ireland and the World Bank Group.

2 DATA

2.1 MONGODB ATLAS

MongoDB Atlas, a cloud-based database service, is used to store the pre-processed data for data-backup and easy retrieval for further processing. The 'studentData' database is created with a replication factor of 3. Access is generated for admin with change permissions, and for users with read-only permission. PyMongo is used as a connector for handling the database queries from python environment using the generated access details. The tables are referred as collections and the records or observation as documents.

2.2 GDP DATA

2.2.1 Data Gathering

The GDP data of Ireland is loaded into python using the World Bank Group REST based Application Programming Interfaces (API). The API gives direct access to most of its updated data, straight-off its databases, without the need for a manual download.

The URL based API consists of:

Standard URL - "**http://api.worldbank.org/v2/**"

Select Country (Ireland) - "**countries/ie/**"

Select indicator (GDP.MKTP.CD)- "**indicators/NY.GDP.MKTP.CD?**"

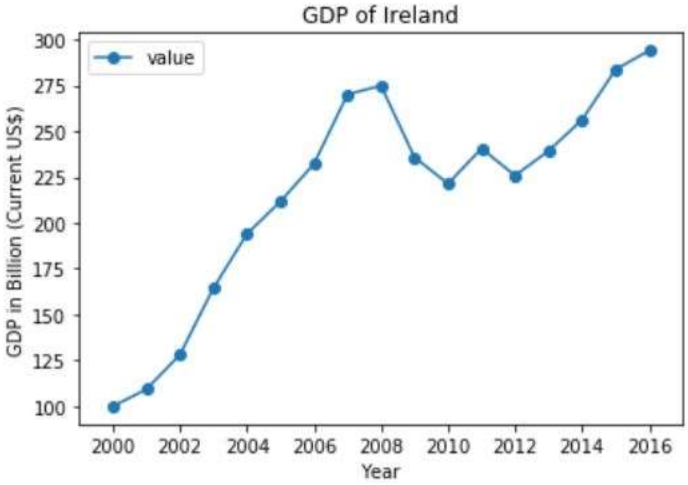
Date range - "**date=2000:2016**"

Download format - "**&format=json**"

"**http://api.worldbank.org/v2/countries/ie/indicators/NY.GDP.MKTP.CD?date=2000:2016&format=json**"

2.2.2 Data Pre-processing

The data loaded in json format is parsed using the 'json' package in python which then returns a nested list. Required fields are extracted and converted to pandas dataframe. The GDP values, in current US dollar, are converted to Billion dollars for convenience. It is sorted by years and then stored in MongoDB for future use.



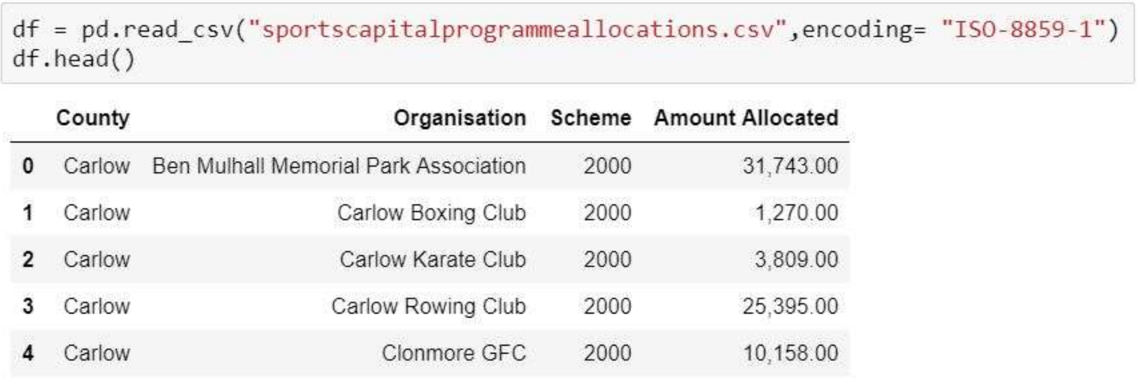
2.3 SPORTS CAPITAL PROGRAMME ALLOCATIONS DATA

2.3.1 Data Gathering

The dataset with all the grants, in Euros, allocated under the 'Sports Capital Programme' was published by Department of Transport, Tourism, and Sport on Ireland's open-data portal(data.gov.ie). The Programme aimed to promote local authorities and schools to develop high quality, safe, sustainable sports facilities and prioritize the facilities in disadvantaged areas. The portal had no option for API access and was only available in 'xls' and 'csv' formats. Here, the 'csv' file was loaded into python using pandas 'read_csv' method.

2.3.2 Data Pre-processing

The 'Amount Allocated' field of the data was read as string, since commas were used to separate thousands and a dot before decimal values.



```
df = pd.read_csv("sportscapitalprogrammeallocations.csv",encoding= "ISO-8859-1")
df.head()
```

	County	Organisation	Scheme	Amount Allocated
0	Carlow	Ben Mulhall Memorial Park Association	2000	31,743.00
1	Carlow	Carlow Boxing Club	2000	1,270.00
2	Carlow	Carlow Karate Club	2000	3,809.00
3	Carlow	Carlow Rowing Club	2000	25,395.00
4	Carlow	Clonmore GFC	2000	10,158.00

The formatting issue was dealt by declaring necessary parameters in 'read_csv' method. The 'Scheme' field was found to have some text along with the year. Therefore, a string manipulation was applied to extract only the year. Since the study focused on the amount allocated annually to individual counties, the amount was aggregated on year and county level. The pre-processed data was then moved to MongoDB.

2.4 MERGE SPORTS CAPITAL WITH GEO-JSON DATA

2.4.1 Geo-Json Data

Ireland’s GeoJSON, a widely used format for encoding geographic data, was used to define the county boundaries which contains 'id' with counties, and 'geometry' with a list of coordinates of type polygon and multi-polygon. Since bokeh is in development and provides limited options for GeoJSON plotting, the polygons are converted to separate lists of latitude and longitude. The dataset with county and its respective latitude and longitude are retained.

2.4.2 Merging Data

The dataset with counties and its respective coordinates is merged with the sports capital data. It is done iteratively by making a query to MongoDB for each year and binding it on counties with the year as its field name. The final data consisting of counties, latitude, longitude and years from 2000 to 2016 with allocated amount values are stored locally as 'json' file using pandas 'to_json' method. Also, the GDP data is queried and saved locally as a 'csv' file which will be used later for visualization.

3 VISUALIZATION

Bokeh is an open-source library, backed by Anaconda, aimed to create versatile, interactive, browser-based visualizations of streaming or Big Data from Python, R, Scala, and Julia without the need to write JavaScript. HTML5 Canvas as its primary output backend, bokeh is engineered to operate in a client/server model for the modern web.

Plot and Glyphs are the basic building blocks for plotting in bokeh. Plots are containers that hold various objects that comprise the final visualization. The Figure objects - figure() is a plot. Glyphs are the basic visual marks such as patches, line, circle, ray, etc.

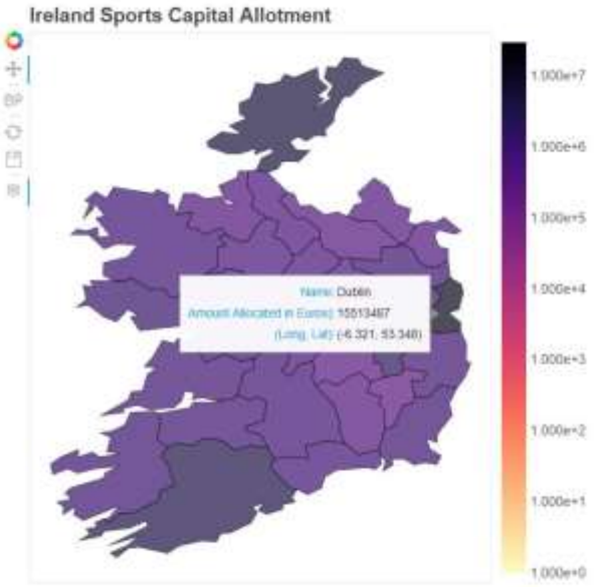
3.1 STATIC PLOTS

3.1.1 Choropleth Map

The county boundaries, from the merged dataset, was used to build a Choropleth Map using the patches glyph with shades representing Sports Capital allocated for the year 2000. Due to the large-scale and variation in the amount allocated, a LogColorMapper was applied to shade the regions. The mapper's max value was set to the maximum amount allotted and min values to 1 (log scale). A ColorBar, with a ticker, was added to describe the shade scale. A hovering tool was added to display the amount allocated to the individual county. The color palette 'Magma256' was found to be safe for color-blind by using color oracle tool.

3.1.2 Line Chart

The GDP was provided in Current US dollar whereas the sports capital in Euros. Also, a common scale cannot be used to study the pattern in both the quantities over years since GDP suppresses the pattern visible of the much smaller Capital amount. Using log-scale had no improvement. Therefore, taking GDP in 'Billion US dollar' and the sports capital in 'Million Euros' yielded a better visible patter. The GDP and total Sports Capital for each year are represented using line and circle glyph. Legends are manually added near the data to make it reader-friendly. A ray (vertical line) is attached to the year 2000, which guide the reader during interactive plots.


Ireland Sports Capital Allotment
Name: Dublin
Amount Allocated in Euros: 15513487
(Long, Lat): (-6.321, 53.348)
1.000e+7
1.000e+6
1.000e+5
1.000e+4
1.000e+3
1.000e+2
1.000e+1
1.000e+0

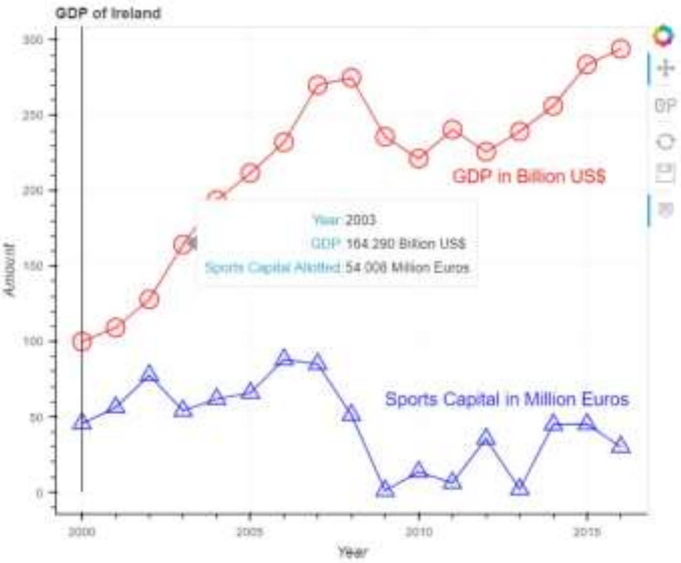


Fig. Choropleth Map

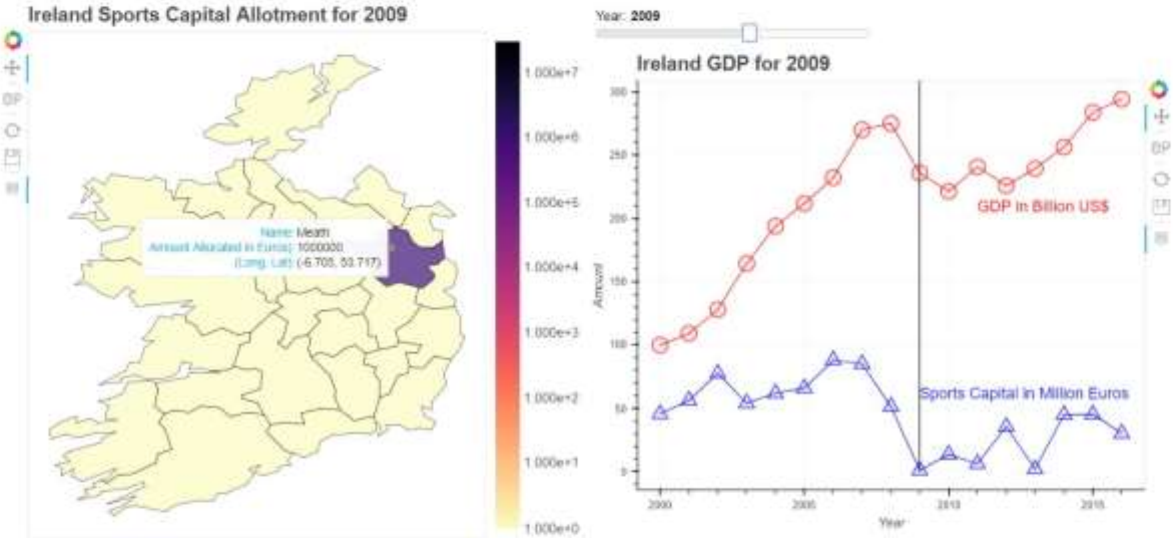
Fig. Line Graph

3.2 Interactive Plots

The Choropleth map and GDP line graph are combined, and a slider is added to introduce interactivity. The 'update_plot' method is used to update the data, based on the year selected, generating a new plot. A bokeh server is employed to generate the plot. The script 'interactivePlot.py' is passed to bokeh server using command Prompt as below

```
currentDir > bokeh serve [--port 5006] interactivePlot.py
```

The above command starts the server on port 5006 by default. A port can be assigned manually in the square bracket(optional). The plot generated can be accessed in a browser on 'localhost:5006'.



Note: The script requires a 'data' folder, on current directory, with required datasets.

4 Conclusion

A steady increase in GDP was accompanied by a similar pattern in Sports Capital till the year 2003 with a higher allotment in Dublin, Cork, Kildare, and Galway counties while varying amounts in the rest. In the year 2004, there was a reduction in sports capital although a steady increase of GDP persisted. This pattern may be attributed to the government's decision to reduce tax and hence its expenditure. The GDP continued to increase due to the property and construction bubble until 2008. Interestingly, there was a considerable drop in sports capital in the year 2007 before the economy collapsed. In 2009, during the crisis, the Sports Capital was allocated only to Meath and was the only year for excluding Dublin. After the crisis, there was a steady increase in the GDP, but the Capital remained nearly constant, excluding the drop in the year 2013. The graph can be improved further by providing options to choose individual counties and various performance metrics. Therefore, a visual data representation can help finding a hidden pattern and can guide in establishing a causal relationship. Though bokeh is an excellent library for plotting, with a smaller learning curve, D3.js provides better customization. An attempt was made to plot a treemap on D3plus library resulted in a great looking interactive plot with only a few commands. Hence, I would prefer JavaScript library to bokeh if the data demands high customization.

5 References

1 The World Bank API (https://datahelpdesk.worldbank.org/knowledgebase/articles/898581-api-basic-call-structure)

2 Open Data Portal – Ireland (https://data.gov.ie/data)

3 GeoJSON File of Ireland (https://gist.github.com/aerrity/4338818)

4 Bokeh Documentation for python (https://bokeh.pydata.org/en/0.12.13/)

5 D3plus Documentation (https://d3plus.org/docs/)