Coursework 3 Report

Foundation of Data Science, 2017/18

Old Trafford Team
University of Southampton
Southampton, United Kingdom

Chaipichet Palotaitakerng*
University of Southampton
Southampton, United Kingdom
cp2n17@soton.ac.uk

Lu Yang†
University of Southampton
Southampton, United Kingdom
ly4n17@soton.ac.uk

Jiachen Li‡
University of Southampton
Southampton, United Kingdom
jl4g17@soton.ac.uk

Phisit Srirattanawong§
University of Southampton
Southampton, United Kingdom
ps4y17@soton.ac.uk

Yue Wang¶
University of Southampton
Southampton, United Kingdom
yw3y17@soton.ac.uk

ABSTRACT

This report presents a descriptive analysis about the different development of countries in our world between the year 1990 and 2015. In this project, we build a logistic regression model to analyse the factors contributing to how countries have developed differently. Some country was in upper middle class (*class B*) in 1990 before the country become high class (*class A*) in 2015. These Countries like South Korea are classified as *BA countries*. The high class comprises of countries which have high GDP per capita and life expectancy. To avoid an over-fitting problem, a model reduction is applied. Our result shows that some indicators, like percentage of the population who use the internet, are significant in a classification of the *BA countries*. Data visualisations on a web application are used to visualise the significant indicator data of countries worldwide. The data application comprises of dynamic scatter plot, Sankey diagram, Chord diagram, world map, and other data visualisations on a dashboard which allow further data exploration.

KEYWORDS

Data Science, Logistic Regression, Countries, World Development

ACM Reference Format:
Old Trafford Team, Chaipichet Palotaitakerng, Lu Yang, Jiachen Li, Phisit Srirattanawong, and Yue Wang. 2018. Coursework 3 Report: Foundation of Data Science, 2017/18. In *Proceedings of 2017/18 (Data Science).* ACM, New York, NY, USA, 6 pages. https://doi.org/

1 INTRODUCTION

In the last three decades, the economy and well-being have been enhanced in most countries. Among those countries in the world, some special instances draw the attention of us. We have noticed that during the past 50 years, South Korea has become the fast developing country in South and East Asia, whose Gross Domestic Product (GDP) and life-expectancy have growing from average level of most disadvantaged countries (in 1965), to match the level of advantaged countries such as most European countries and the USA (in 2015). According to the facts, we are interested in finding the factors that affect the development of a fast-growth country like South Korea.

We study the development of our world by using the data about exports, imports, transport, internet access, innovative capability, natural resource, population from the Gapminder, the International Labour Organisation (ILO) and the World Bank Websites.

1.1 Groups of countries

According to the World Bank, countries are classified into four groups corresponding to their income level, GDP *per capita*. In our analysis, we introduce a different group of countries based on two indicators: GDP *per capita* and life expectancy. So that we can have another kind of groups that are not only based on the income level. But, the groups are also based on the health level, life expectancy, too.

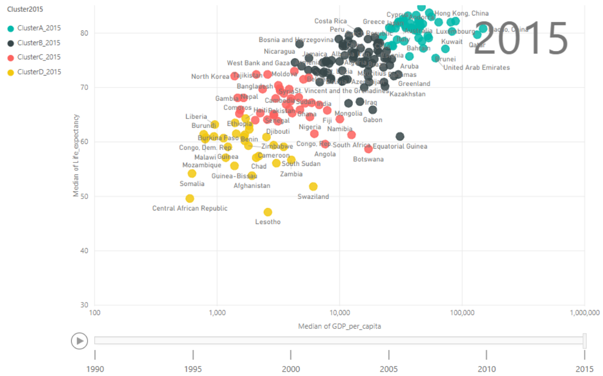


Figure 1: Scatter plot of countries that are coloured by their *Clusters in 2015*

In addition, we consider two points in time which are the year 1990 and 2015. Classification of countries into four groups has been done at these two years. Therefore, GDP *per capita* and life expectancy in 1990 are used to classify countries into four groups,

*Team Leader
†Deputy Team Leader
‡Member who analyse data by trees
§Member who visualises all cool graphs.
¶Member who prepare and analyse data.

Data Science, Group Work, Southampton, UK
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

called *Clusters in 1990*. And, countries are divided into another four groups, called *Clusters in 2015*, depending on their GDP *per capita* and life expectancy in 2015.

The four groups formed by K-mean clustering algorithm where $K = 4$ and the distance is calculated in logarithmic scale as illustrated in Figure 1 where the vertical axis and horizontal axis of the scatter plot use the logarithmic scale to present GDP *per capita* and life expectancy in 2015.

These four groups represent four different classes of countries. The first class is a class of high income and life expectancy, called *Cluster A*, or the *high class*. The second class is a class of upper middle income and life expectancy, called *Cluster B*, or the *upper middle class*. The third class is a class of lower middle income and life expectancy, called *Cluster C*, or the *lower middle class*. Finally, the fourth class is a class of low income and life expectancy, called *Cluster D*, or the *low class*. Following table presents examples of countries in each class.

Table 1: Clusters in 1990

Cluster	Count	Examples
A	49	Japan, USA, UK, etc.
B	65	South Korea, Thailand, etc
C	41	China, Egypt, India, etc.
D	38	Afghanistan, Lao, Zambia, etc.

Source: Data from Gapminder and the World Bank with our data processing.

Table 2: Clusters in 2015

Cluster	Count	Examples
A	47	South Korea, Japan, USA, UK, etc.
B	76	Egypt, China, Thailand, etc
C	40	India, North Korea, Lao, etc.
D	30	Zambia, Afghanistan, etc.

Source: Data from Gapminder and the World Bank with our data processing.

Cluster B has the highest number of countries in 1990 and 2015, 65 and 76 respectively. While Cluster D has the lowest number of countries in 1990 and 2015, 38 and 30 respectively.

1.2 Cluster Transitions

Most of *Cluster A in 2015* countries are developed countries. Some country like South Korea used to be a developing country before. The country was a country in *Cluster B in 1990*. We consider this change as a *Cluster transition* which represents a change of country from one group in 1990 to another group in 2015. In case of South Korea, this Cluster transition is called *BA transition*. Countries that have *BA transition* are classified as **BA group**. Therefore, **BA group** is a group of countries in *Cluster B in 1990* and in *Cluster A in 2015*.

Sankey diagram and Chord diagram are used to present these *Cluster transitions* as flows from one Cluster in 1990 to another Cluster in 2015. The diagrams are described in the following section of our data application. From our analysis, there are eleven flows from 1990 to 2015 as presented in the following table.

Table 3: Cluster Transition Groups

Group	Count	Examples
AA	42	United States, Japan, Sweden, etc.
AB	7	Libya, Montenego, Croatia, etc
BA	5	South Korea, Poland, Estonia, etc.
BB	57	Malaysia, Turkey, Mexico, etc.
BC	3	Honduras, Syria, Tonga, etc.
CB	11	China, Indonesia, Egypt, etc
CC	25	India, Uzbekistan, Sudan, etc.
CD	5	Cameroon, Zimbabwe, Lesotho, etc.
DB	1	Equatorial Guinea
DC	12	Rwanda, Myanmar, Bangladesh, etc.
DD	5	Afghanistan, Guinea, Uganda, etc.

Source: Data from Gapminder and the World Bank with our data processing.

In our analysis, we focus on countries that used to be a member of cluster B in 1990 and have developed differently. While many countries are still in cluster B in 2015, some countries have upgraded to cluster A in 2015. The countries of the *Cluster B in 1990* such as South Korea, Poland, and Estonia are what we interested in researching for the reason why they can make a better progress than the other countries.

1.3 Problem Statement

The analysis attempts to answer a question: *which indicators can forecast a group of countries that developed differently?* In our case, we noticed some fast-growth countries. So, we intend to analyse indicators that could reflect how these countries have developed differently than other countries in our world.

1.4 Hypothesis

In our first data exploration, we found that South Korea and United Arab Emirates (UAE) had developed rapidly in the last century. Assuming that the UAE may have benefited from natural resources like petroleum while South Korea may have benefited from exports or innovative capability.

According to the cluster transitions, we intend to classify the **BA group** of countries, the countries that were in Cluster B (*upper middle class*) in 1990 and then moved to Cluster A (*high class*) in 2015. We came up with fifty-two indicators that may possibly be a good predictor in the classification. For example, it is supposed that: Individuals using the Internet, patent applications and the number of departures in international tourism are the significant indicators reflecting why some of the countries in *Cluster B in 1990* have moved to *Cluster A in 2015*.

2 IMPLEMENTATION

2.1 Tools

Microsoft Excel and R are used for data processing. Using MATLAB and R for data analysis, Microsoft Power BI for basic data visualization and clustering, and finally using HTML for showing results. Our application stored the data in CSV files.

Our team used many tools throughout the project. Microsoft Excel, Microsoft PowerBI, MATLAB and R are used for data processing in the period of data preparation. Then the data analysis is carried out by using MATLAB and R, with the basic data visualized through Microsoft PowerBI and using D3.js data visualisations. And, our data application is developed on web technology like HTML, CSS, and Javascript.

2.2 Data Preparation

Once the topic was identified, data collection is carried out. We looked for datasets that related to our idea. Meanwhile, we cleaned the data to make the dataset ready for further analysis in R and MATLAB.

There are 193 countries which have available data about GDP *per capita* and life expectancy over the time span from 1990 to 2015, selected as a subject to be analysed. All countries are grouped into clusters by their GDP *per capita* and life expectancy using the clustering function of Microsoft Power BI.

The time-series data from 1990 to 2015 of fifty-two indicators which represent the factors that could affect the country development are acquired. These fifty-two factors are chosen because they include many aspect factors may affect a country 's development, including innovation, information access, transportation, natural resources, import and export.

To link many time-series from many data sources, data of the 193 countries are linked together by their country code. Because the name of countries varies from one data source to another. So, we create a reference data, called *Master Data*, that contains all different names together with the three-letter international codes of all countries, used by the World Bank. Then, we use the *Master Data* to refer the same country to the same country code even that country has many different names. The labels of every cluster country are also set.

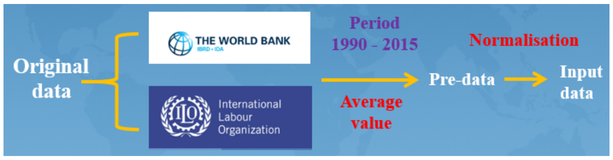


Figure 2: Data preparation process

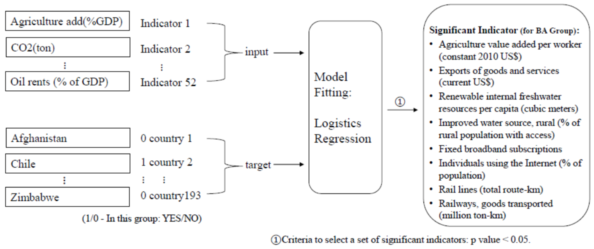
Datasets were downloaded and merged from the World Bank[6] and ILO (International Labour Organization)[5] which are reliable data sources.When processing the data of every factor, there exist some missing data. At first, we decide to use indicators from 1965 to 2015 that may affect the development of a nation, but most of the datasets we found are not able to provide data at such an early time period back to 50 years ago. In order to balance the datasets, we choose the fields of each indicator from the year of 1990 to 2015. And, the average value of each indicator data during the period, between 1990 and 2015, is used as the *Pre-data*. Finally, we normalise the *Pre-data* for the input value for classification by the logistic regression.

2.3 Data Analysis

The cluster transitions of each country are labelled. The labels are used as the target value in our analysis. Our aim is to identify indicators that affect the country development. It is important to choose a suitable method. We use logistic regression to classify the target value, the cluster transition type, from indicator values of that country. The logistic regression model is implemented in R. [1]

Before we tried using logistic regression, we tried sparse regression at first. The regression method is mainly used to predict the continuous value. This technique decreases the coefficient value of non-relevant inputs to near zero, called regularisation. Then, the sparse regression provides us with the non-zero coefficients as relevant inputs. The sparse regression is implemented using Matlab toolbox. However, we find that the results of sparse regression are very different from the target value. So, this approach is not suitable for processing our dataset.

Then, we tried the logistic regression instead. The results show that the logistic regression model can classify the target value of the BA group correctly. Therefore, logistic regression method is chosen.


Agriculture add(%GDP)
CO2(ton)
Oil rents (% of GDP)
Indicator 1
Indicator 2
Indicator 52
input
Afghanistan
Chile
Zimbabwe
0 country 1
1 country 2
0 country193
target
(1/0 - In this group: YES/NO)
Model Fitting:
Logistics Regression
①
Significant Indicator (for BA Group):
Agriculture value added per worker (constant 2010 US$)
Exports of goods and services (current US$)
Renewable internal freshwater resources per capita (cubic meters)
Improved water source, rural (% of rural population with access)
Fixed broadband subscriptions
Individuals using the Internet (% of population)
Rail lines (total route-km)
Railways, goods transported (million ton-km)

①Criteria to select a set of significant indicators: p value < 0.05.

Figure 3: Using logistic regression to identify significant indicators of the classification for *BA group* countries

Using logistic regression, we make a fitting model to the data of *BA countries*. And, the model can correctly classify countries that have changed from *B class* to *A class*. Next, we calculate a p-value for every input variable. Then, we interpret the coefficient value and p-value of the indicators. With p-value less than 0.05, some input variables are identified as significant indicators shown in Figure 3.

However, analysing the data using logistic regression with as many inputs as the 52 indicators, the model can have an over-fitting problem. In order to reduce the indicators and build the model with the useful indexes, we use the logistic regression algorithm to derive the p-value of each input variable and pick the ones with a p-value less than 0.05 (significant indicators) to construct a new model, the reduced model. If the number of significant indicators is less than the original dataset and the AIC value[2] decreased, the model would be better and not over-fitting.

The second model, the reduced model, has better AIC value than the first model. Lower AIC value indicates less over-fitting than the first model.

2.4 Application Development

First of all, we have tried many tools to develop the web application. Started with Microsoft Azure which contains a lot of service like database services. However, this service does not suit our project because it does not flexible as our project wanted. SQLite is another database that we have tried. It worked fine with the local application but if we would like to move to an online application, this database could not do the job. Then we designed to use MySQL server to store the data and using PHP as the connection between the front end and back end. Even though, by our process of preparing data, we have to change the table schema much time. It is better to move back from MySQL server to read the data from the CSV files directly. In addition, it is not a large data set so database might not be necessary for this application. Although the database is needed if we have to publish our application online, in this case, our application with a pre-processing data(not a real-time processing data) is basically used for visualisation and database service is not requisite.

To visualise data, our application used D3 and jQuery, a Javascript library, to create the front end of the application. In addition, we also used another library for D3 to create a map[4] and a Sankey diagrams[3]. The reason why our team chooses D3 as the combination of the Sankey diagram and the world map is that D3 much faster than Tableau or Microsoft Power BI in many cases. Laggardly response mains a lack of user-friendly. Although, this application still use Microsoft Power Bi for other parts which will be given a detail in the following part.

For Sankey and map diagram, the data in CSV files have been read and stored in the variables using the d3.csv function. However, it leads to a problem with Google Chrome web browser. The problem is that Google Chrome does not allow the application access to local files. We found two ways to solve this. The first is using a localhost server on the computer. Another way is using other web browsers such as Firefox and this one is our solution because it is more flexible for us to work with. Even though, it is better to use online database for everyone to access the same data to develop the application as well as to public this application online.

The Sankey diagram has been created from nodes and links which are clusters, A, B, C and D, and the transitions between the year of 1990 and 2015. Each of them contains the name of countries and ISO ALPHA-3 code, the three-letter country code, in that group. These data also link to the map which will change the map colour whenever the user moves the mouse to others transition.

For the world map, this map also connected to the radio box which is a list of most relevant indicators for the group BA transition. We have picked only eight indicators because we would like the user to focus the BA group first. Next, whenever the user changes the indicator it will change the countries colour depended on how much the value of that country compared with the maximum value. However, some of these indicators have a very high value for some countries. To solve this, the logarithm has been used in the calculation step to make a clearer colour of countries.

In this application, we also have two more visualisations from Power BI. This part is for the user who looking for in-depth information such as other indicators or other transition groups on the world fact. The reason that we used Power BI rather than Tableau is it can create a dynamic bubble chart over time. Furthermore, it is much easier to use Power BI comparing to develop this kind of visualisation by D3 or other scripts.

Finally, to decorate this web application, W3.CSS have been used to do this job. This framework likes a simplify version of Bootstrap. Comparing with Bootstrap, W3.CSS is more convenient and easy to organize so that it take less time to develop a small application like our project.

3 RESULTS

3.1 Logistic Regression

As mentioned in the data analysis subsection, logistic regression is used to make a fitting model to the data of *BA countries*, of which the countries used to be in upper-middle class in 1990 and upgraded to developed countries 25 years later. Next, we calculate a p-value for every input variable. Then, we interpret the coefficient value and p-value of the indicators. With p-value less than 0.05, some input variables are identified as significant indicators.

Eight significant variables from the fifty-two variables are identified as show in Table 4.

Table 4: Statistics of Significant indicators

Indicator Id	Coefficient	p-value
8[a]	-1.898127	0
29[b]	-1.111023	0.048562261
33[c]	1.577956	1.62187E-07
34[d]	0.127068	0.043153353
40[e]	1.353456	0.001128608
42[f]	2.809061	1.79264E-05
47[g]	2.111252	0.000120241
48[h]	-1.522498	2.66352E-05

Source: from World Bank database and International Labour Organization database.

[a] Agriculture value added per worker (constant 2010 US$)
[b] Exports of goods and services (current US$) Renewable internal freshwater resources per capita (cubic
[c] Renewable internal freshwater resources per capita (cubic meters)
[d] Improved water source, rural (% of rural population with access)
[e] Fixed broadband subscriptions Individuals using the Internet (% of population)
[f] Individuals using the Internet (% of population)
[g] Rail lines (total route-km)
[h] Railways, goods transported (million ton-km)

Since analysing the data using logistic regression with as many inputs, the model can have an over-fitting problem. We did a model reduction to get a reduced model. The second model, the reduced model, has better AIC value than the first model. Lower AIC value indicates less over-fitting than the first model as shown in Figure 5.

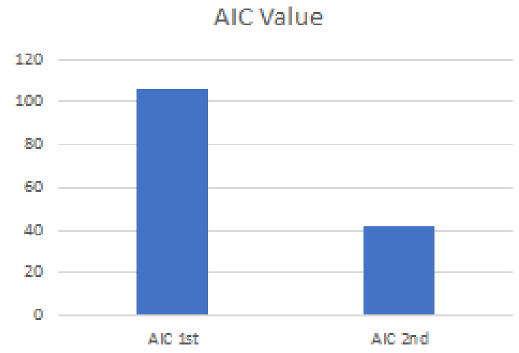


Figure 4: AIC value for the second model is lower than the first model which indicates less over-fitting

The AIC value decrease from 106.0002 in the first model to 42.20031 in the second model. The second model, reduced model, uses only the top relevant indicators as input variables. This demonstrates that the reduced model is better and less over-fitting.

Table 5: Residual Deviance value comparison

Cluster Transition	1^{st} Model	2^{nd} Model
BA group	0.000168606	24.20031

The first logistic regression model has lower residual deviance than the reduced model, the second model.

However, the residual deviance increases from 0.000168606 in the first model to 24.20031 in the second model. Hence, the second model has larger error value than the first model.

Therefore, only top significant indicators are not sufficient to classify the BA countries. To identify the group of countries with high accuracy, the top significant indicators need to be considered with some less significant indicators.

3.2 Data Application

Our data application consists of four functional parts. In the first part, scatter plot displays the clusters of 193 countries in year 1990 and 2015 into cluster A to D separately. In the second part, Sankey diagram shows the transition of the four clusters from 1990 to 2015. In the third part, eight significant indicators can be selected, and the world map will present the difference between 193 countries. The final part is for further investigation. There is a dashboard which contains many data visualisations that present the cluster transition in Chord diagram, the median value of the selected indicator for each transition types in a bar chart, and represent the value of each country by bubble size in a world map.

Scatter plot. This part contains four pages: the first page presents all transition types on a map, the second and third part present four clusters in 1990 and 2015 separately with animated transition path. Figure 5 illustrates the clusters in 2015 and transition path of South Korea clearly.

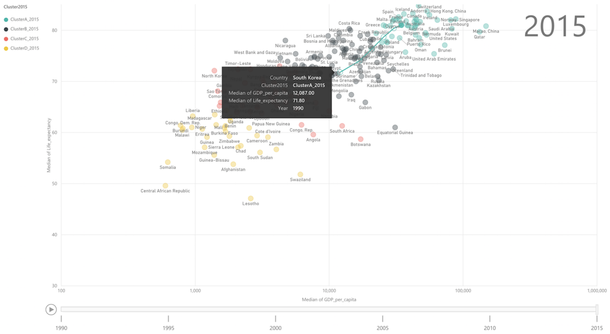


Figure 5: Transition path of South Korea from 1990 to 2015 which display an example of the cluster transition from class B to class A

Sankey diagram. A Sankey diagram is a visualization used to depict a flow from one set of values to another. For example, the cluster C in 1990 contains 41 countries. In 2015, while twenty-five countries remained in cluster C and five countries dropped to cluster D, eleven of them developed and rose to cluster B, which is shown in green line in Figure 6.

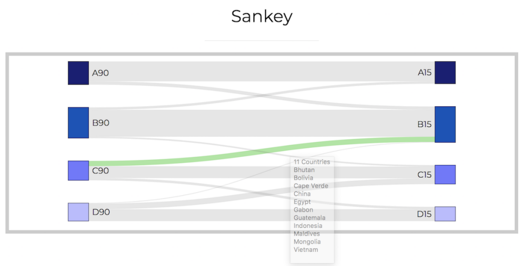
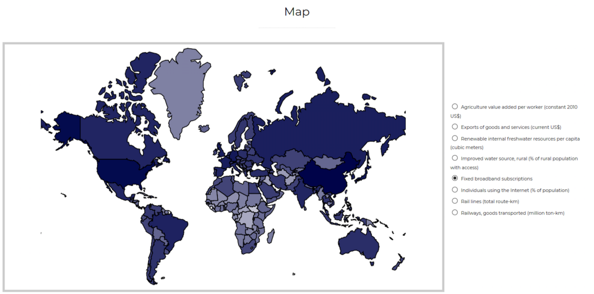


Figure 6: Transitions of four clusters from 1990 to 2015, highlighting the countries developed from cluster C to cluster B in green.

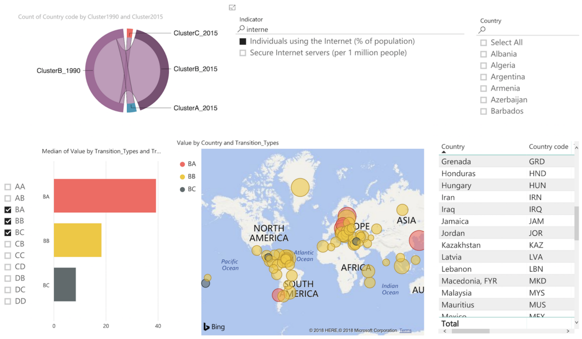
World map. The following part is an example of the world map chart with an interactive feature layer showing the areas with different colour depth. As indicator "fixed broadband subscriptions" is selected, the map shows that the countries with darker colour have better performance than one with the lighter colour in this field.



○ Agriculture value added per worker (constant 2010 US$)
○ Exports of goods and services (current US$)
○ Renewable internal freshwater resources per capita (cubic meters)
○ Improved water source, rural (% of rural population with access)
● Fixed broadband subscriptions
○ Individuals using the Internet (% of population)
○ Rail lines (total route-km)
○ Railways, goods transported (million ton-km)

Figure 7: Different performances in various countries to the indicator "fixed broadband subscriptions".

Further investigation. This part is of exploring the significance of more variable factors to different clusters. In the case of cluster B, which is divided into BA, BB, BC group shown in Chord diagram, after the indicator "individuals using the Internet" is selected, Figure 8 present the different performance of these three group distinctly in the bar chart. It can also be notified in the world map that more important factor is illustrated by a larger circle.


Count of Country code by Cluster1990 and Cluster2015
ClusterC_2015
ClusterB_1990
ClusterB_2015
ClusterA_2015
Median of Value by Transition_Types and Tr...
BA
BB
BC
0
20
40
Value by Country and Transition_Types
BA
BB
BC
NORTH AMERICA
SOUTH AMERICA
AFRICA
ASIA
OPE
AU
Pacific Ocean
Atlantic Ocean
Indian Ocean
Bing
© 2018 HERE, © 2018 Microsoft Corporation Terms

Indicator
interne
■ Individuals using the Internet (% of population)
□ Secure Internet servers (per 1 million people)

Country
□ Select All
□ Albania
□ Algeria
□ Argentina
□ Armenia
□ Azerbaijan
□ Barbados

□ AA
□ AB
☑ BA
☑ BB
☑ BC
□ CB
□ CC
□ CD
□ DB
□ DC
□ DD

Country	Country code
Grenada	GRD
Honduras	HND
Hungary	HUN
Iran	IRN
Iraq	IRQ
Jamaica	JAM
Jordan	JOR
Kazakhstan	KAZ
Latvia	LVA
Lebanon	LBN
Macedonia, FYR	MKD
Malaysia	MYS
Mauritius	MUS
Mexico	MEX
Total	

Figure 8: Indicator "individuals using the Internet" have high median value in the countries which developed from cluster B to Cluster A

4 CONCLUSIONS

Countries are divided into four classes at two points in time, the year 1990 and 2015, by GDP per capita and life expectancy. There are some countries like South Korea have changed from the upper middle class in 1990 to the high class in 2015. The high class comprises of countries which have high GDP per capita and life expectancy.

Our data application allows users to explore the development indicator data of countries worldwide. Data visualisation in the application display the classes of countries in dynamic bubble chart, the cluster transition in Sankey diagram and Chord diagram, indicator value by colour in world map, indicator value by bubble in another world map, median value of indicator for each cluster transition in bar chart, and further details in data table.

According to the result, some indicators, such as the percentage of the population who use the internet, are significant in the classification of the BA countries. However, the significant indicators are not sufficient to classify which country will be in the BA group. The significant indicators need to be considered with other indicators to correctly identify the group of countries.

5 LIMITATIONS AND FUTURE WORK

5.1 Input variable on *per capita* basis

In future work, values of some indicators could be divided by the number of population of that country before used as input variables for the classification model. Then, we can analyse the factors on *per capita* basis. So that effect of population number would be reduced.

5.2 Interpolation for missing values

In the data processing period of this project, we averaged the data over our period of interest, from 1990 to 2015. We use the average values to represent every indicator data, even there are some missing values for some country or some year. For a more precise result, these data should be fitted by an appropriate model to interpolate data on the missing values using this fitting.

5.3 Different methods for classification of the cluster transition

Since our analysis is identifying the relevant indicators that reflect the development of countries to have different cluster transition. In future work, the model could be improved using different methods such as Artificial Neural Network (ANN) instead of the logistic regression model to forecast the target value, the cluster transition.

5.4 Using server-side web technology

This application is created only for the pre-processing data because we have to process data in Matlab and R which hard to integrate with our web application within a given time-frame of this project. Moreover, some data which we have processed by the script programming should be done by the server-side technology like PHP or Node.js rather than using Javascript like what we implemented in our application. The reason is the security and the performance of the application.

REFERENCES

[1] S. Agarwal. 2016. How to use Multinomial and Ordinal Logistic Regression in R? (2016). https://www.analyticsvidhya.com/blog/2016/02/multinomial-ordinal-logistic-regression/

[2] H. Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (Dec 1974), 716–723. https://doi.org/10.1109/TAC.1974.1100705

[3] M. Bostock. 2012. Sankey Diagrams. (2012). https://bost.ocks.org/mike/sankey/

[4] M. DiMarco. 2016. DataMaps. (2016). http://datamaps.github.io/

[5] ILO. 2017. International Labour Organization. (2017). http://www.ilo.org/global/lang--en/index.htm

[6] World Bank. 2017. World Bank. (2017). http://www.worldbank.org/