CISC7201 Course Project Report
Group U
M-B9-5517-1 Susan, Ho Mei San
M-B9-5520-1 Anne, U Wai Cheng
M-B9-5562-7 Hito, Chen Haitao

University of Macau
CISC7201 INTRODUCTION TO DATA SCIENCE PROGRAMMING
Course Project, 1st semester 2019/2020

Topic: Development in Global Electrification

This report is separated into several parts. We will illustrate the background of this project, then introduce the selected dataset. Afterwards, we will discuss the data cleansing, processing and analysis, as well as findings and perform data forecast. Finally, a conclusion is made.

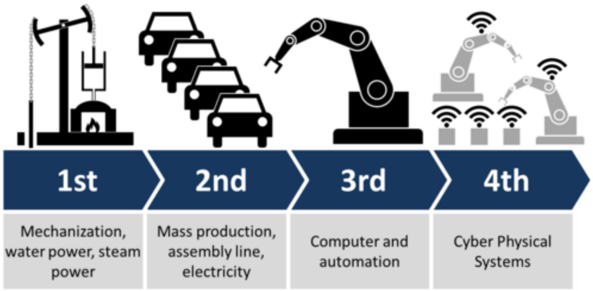
Part I. Terminology

The following terminology is mentioned throughout this project:

1. Electrification rate: The percentage of people who have access to electricity in a given area.
2. Access deficit: The number of people who do not have access to electricity in a given area.
3. Universal electricity access: The global electrification rate is 100%.

Part II. Preface

Science and technology are making huge strides every day, and it is believed that we are now in the stage of the Third Industrial Revolution[1] and moving towards the Fourth Industrial Revolution[2]. The image[3] in the right side summarized the development of Industrial Revolution. Dated back to the Second Industrial Revolution, one of the most significant technological systems, electrical power, was introduced. It is the vital element of the Third and the Fourth Industrial Revolution as well. Ever since the application of electricity in human's daily life, electricity becomes pivotal because it helps to safeguard the people's quality of life and encourage economic development. Since electricity is one of the must-have elements for national economic growth, electrification rate can be treated as a proxy of a country's wealthy and developing situation, and the level of individual access to electricity can reflect the poverty status of the country. This project tries to explore the status about the development in global electrification.


1st
2nd
3rd
4th
Mechanization, water power, steam power
Mass production, assembly line, electricity
Computer and automation
Cyber Physical Systems

Part III. Utilized Python Libraries

Different libraries and modules have been applied to this project, and they are summarized as follows:

[1] It is also called Digital Revolution.
[2] The term was firstly introduced by Klaus Schwab, the executive chairman of the World Economic Forum.
[3] Source of Picture: Christoph Roser at AllAboutLean.com.

Python Library	Purpose
Requests	Obtain data file from HTTP
Zipfile (in the Python Standard Library)	Read the compressed file
Pandas	Import data and store in DataFrame for further processing
PrettyPandas	Add number formatting to the DataFrame generated by Pandas
Numpy	Improve the efficiency of calculation
Seaborn	Create attractive and informative statistical graphs
Plotly	Create animated graphs
Matplotlib StatsModels Scikit-learn	Forecast and visualize results

Part IV. Selected Dataset

The World Bank Group collects the global development indicators[4] of the latest and accurate global development data available from different officially recognized international sources. As a well-known international organization, we believe that the collected information is reliable, so we retrieve the data to finish this project. There is a dataset called "World Development Indicators". It contains time series (from 1960 to 2019) and national data, and the information is kept updated from time to time (the latest update date is October 28, 2019). The dataset includes diversified topics, including but not limited to Economic Growth, Energy and Extractives, as well as Health, Nutrition and Population. Due to the enormous amount of data, the data of "World Development Indicators" is compressed in a zip file with the size of around 60MB, while there are six csv files in the zip file with the uncompressed sizes ranging from 43kB to 186MB (total uncompressed size is 245MB). We have written a code to directly collect the data online. Then we previewed the dataset in Python and decided to concentrate on two csv files: WDIData (size of 186MB with data of 377,784 rows × 65 columns) and WDICountry (size of 127kB with 263 rows × 31 columns) for this project. The WDIData,csv file covers Country Name[5], Country Code, Indicator Name, Indicator Code and the values of each indicator in the respective year. While the WDICountry.csv contains the country information, as a remark/complement to the WDIData,csv file.

Part V. Data Cleansing and Processing

Since the dataset includes many noisy records, we firstly remove the data not used in this project, starting with the WDIData,csv file. We found that we need to extract the data carefully. Since our focal point is the electrification rate, after previewing the indicators, we solely select the following indicators: (1) "Access to electricity (% of population)"; (2) "GDP per capita (current US$)" (we would like to study the relationship of electrification rate and economic development); and (3) "Population, total" (it

[4] The dataset is retrieved from https://datacatalog.worldbank.org/dataset/world-development-indicators.

[5] Here, "country" is just a generic term. Some regions (like Hong Kong and Macau SAR) are also included.

helps to provide a clearer picture in data visualization stage). Simultaneously, the dataset comprises value of single country and multiple countries (a group of countries under the same criteria, for example, Arab World consists of 22 countries). They must be separated in order not to mess up everything. Thus, we use the "Currency Unit" information (in the WDICountry.csv) as an indicator to filter out the records of multiple countries, because only single country contains information under that column. Also, we obtain "Region" information from the WDICountry.csv as well, which assists in performing data visualization in the later stage of this project. On the other hand, although the dataset covers the time range of 1960-2019, the data of indicator "Access to electricity (% of population)" is available from 1990 to 2017 only, so our analysis focuses on the results of 1990-2017.

In addition to drill down to the country level to perform analysis, another aspect is to have a macro-point-of-view. We select "World" in the Country Name and create the DataFrame. However, only the data from 1993 to 2017 is available, the result in this part includes the time range of 1993-2017. This will become another part of the analysis and visualization.

After removing the noisy records and restructuring the DataFrame, we can create the well-structured tables to visualize the data. Since the data structures used for different graphs are different, we try to establish least tables for data visualization. Finally, we need to set up three structured tables.

Part VI. Data Analysis, Visualization and Summary

With the created DataFrame mentioned in the previous section, we are now able to visualize the data and perform analysis. When setting up the graphs, we found the following pain points:

- The result is not reasonable if we drop all invalid electrification rates, when we count the number of countries with less than 100% electrification rate. It is because not every country has valid electrification rate during all selected years, for instance, let's assume that the earliest available data for country A is year N with less than 100% electrification rate. We can infer that the A did not have 100% electrification rate before N, so we use backfill method to fill in the previous values. In this case, A is counted as the country with less than 100% electrification rate from 1990 to N.
- By comparing electrification rate with GDP per capita, it is quite complicated to visualize the result because the range of GDP per capita is extremely huge, hence we decide to utilize logarithmic scales to establish the graphs.

We establish nine graphs in total. With those graphs, the following conclusions are made:

- Generally, the number of countries with less than 100% electrification rate was quite stable before 2011, then it dropped in recent years.
- Although the number of countries with less than 100% electrification rate was quite stable before 2011, the electrification rates for those countries were improving gradually. We have prepared an animated map showing the movement of electrification rate for those years.
- Electrification rate is a good indicator to show the status, but it does not contain the information of population. Therefore, we prepared another animated map to show the access deficit. For instance, without this animated map, we do not know the fact that even if the electrification rate of India was

improving in the past decades, there are still almost 100 million people who did not have access to electricity in 2017.

- It is proved that countries with fewer GDP per capita tend to have lower electrification rates, and many of the them are in Sub-Saharan Africa.
- Despite of the growing size of world population, access deficit is gradually decreasing over the years, with the refining electrification rate significantly.

By the way, the last section in the Jupyter Notebook is to let readers export the static images to the personal drive for their own interests.

Part VII. Data Forecast: Is it possible to achieve universal electricity access in 2030?

The United Nations set 17 major Sustainable Development Goals[6] to transform the world to achieve a better and more sustainable future, and it is expected that those goals should be accomplished by 2030. Among them, the Goal 7 is about "Affordable and Clean Energy", where one of the targets is "*By 2030, ensure universal access to affordable, reliable and modern energy services*". However, a recent report[7] specified that it is too optimistic for the world to attain universal electricity access by 2030, based on the current reforming progress. Consequently, we would like to write codes to forecast when universal electricity access will be, at least mathematically/statistically speaking. We try to forecast the global electrification rate for future years by using the following models:

1. Ordinary least squares (OLS) regression model[8]: Universal electricity access can be reached by 2046.
2. Gray forecast model[9]: Universal electricity access can be reached by 2045.
3. Least absolute shrinkage and selection operator (Lasso) regression model (L1 regularization)[10]: Universal electricity access can be reached by 2047.

All the three models provide similar outcomes, so we summarize that it is difficult to reach universal electricity access in 2030, and we may need to wait until around 2046 to have universal electricity access, if the improving progress on the electrification rate does not accelerate.

Part VIII. Summary

The project helps us to have better understanding on dealing with the real dataset with the help of Python. By cleansing and processing the dataset, we can have more ideas about the status of global electrification, while the findings are summarized above. In addition, it is still possible to realize universal electricity access by 2030, but the progress of improving electrification rates must be accelerated and rectified. There is still a long way to go.

~END~

[6] For details, please refer to https://www.un.org/sustainabledevelopment/.

[7] For details, please refer to the 2019 Tracking SDG7 Report: https://sustainabledevelopment.un.org/content/documents/2019_Tracking_SDG7_Report.pdf.

[8] It is a statistical method to estimate the relationship between independent variable(s) and a dependent variable.

[9] It is a non-statistical forecasting method to predict the behavior of non-linear time series.

[10] It is a regression analysis method to perform both variable selection and regularization to improve the accuracy and interpretability on the prediction of the produced statistical model.