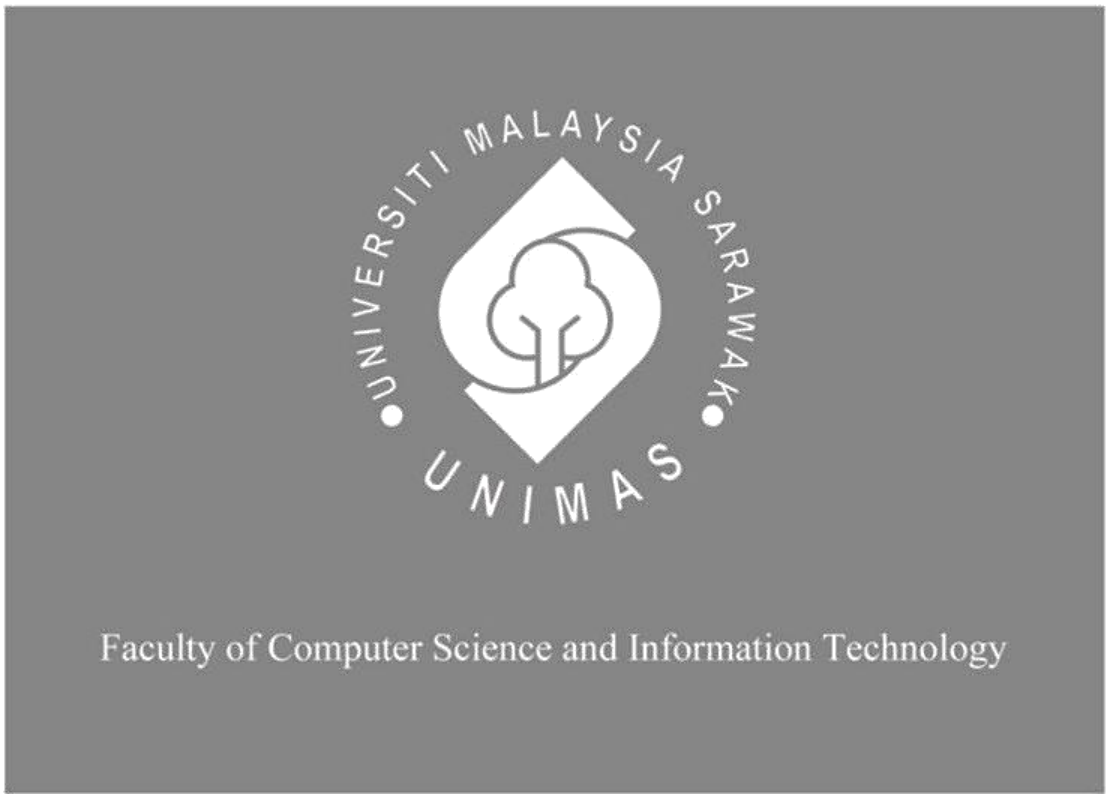


A DATA-DRIVEN ANALYSIS ON THE GLOBAL VACCINATION COVERAGE

IVY SYZLYNDA NASHION (51976)

Bachelor of Computer Science with Honours (Computational Science)

2019

A DATA-DRIVEN ANALYSIS ON THE GLOBAL VACCINATION COVERAGE

IVY SYZLYNDA NASHION (51976)

This project is submitted in partial fulfillment of the

requirements for the degree of Bachelor of Computer Science

with Honours

Faculty of Computer Science and information Technology

UNIVERSITI MALAYSIA SARAWAK 2019

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE A DATA DRIVEN ANALYSIS ON THE GLOBAL VACCINATION COVERAGE

ACADEMIC SESSION: 18/19

I, IVY SYZLYNDA NASHION (51976)

(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

☐	CONFIDENTIAL	(Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)
☐	RESTRICTED	(Contains restricted information as dictated by the body or organization where the research was conducted)
☑	UNRESTRICTED	

Validated by

(AUTHOR'S SIGNATURE)

(SUPERVISOR'S SIGNATURE)

Permanent Address

K.G. NARAWANG,
89308 RANAU,
SABAH

Date: 13/05/2019

Date: 15/5/19

Note * Thesis refers to PhD, Master, and Bachelor Degree
** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

UNIVERSITI MALAYSIA SARAWAK

Grade: _____

Please tick (√)

Final Year Project Report	☑
Masters	☐
PhD	☐

DECLARATION OF ORIGINAL WORK

This declaration is made on the .....MAY.....day of.....15..... year .....2019.....

Student's Declaration:

I IVY SYZLYNDA NASHION, 51976, FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

(PLEASE INDICATE NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled, A DATA-DRIVEN ANALYSIS ON GLOBAL VACCINATION COVERAGE is my original work. I have not copied from any other students' work or from any other sources with the exception where due reference or acknowledgement is made explicitly in the text, nor has any part of the work been written for me by another person.

15/5/2019
Date submitted

IVY SYZLYNDA NASHION(51976)
Name of the student (Matric No.)

Supervisor's Declaration:

I, PHANG PIAU (SUPERVISOR'S NAME), hereby certify that the work entitled, A DATA-DRIVEN ANALYSIS ON THE GLOBAL VACCINATION COVERAGE (TITLE) was prepared by the aforementioned or above mentioned student, and was submitted to the "FACULTY" as a * partial/full fulfillment for the conferment of BACHELOR OF COMPUTER SCIENCE WITH HONOURS(COMPUTATIONAL SCIENCE) (PLEASE INDICATE THE DEGREE TITLE), and the aforementioned work, to the best of my knowledge, is the said student's work

Received for examination by: _____ Date: 15/5/19
(Name of the supervisor)

.....
Dr Phang Piau
Lecturer
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak

I declare this Project/Thesis is classified as (Please tick (√)):

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*
☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*
☑ **OPEN ACCESS**

I declare this Project/Thesis is to be submitted to the Centre for Academic Information Services (CAIS) and uploaded into UNIMAS Institutional Repository (UNIMAS IR) (Please tick (√)):

☑ **YES**
☐ **NO**

Validation of Project/Thesis

I hereby duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic and research purposes only and not for other purposes.
- The Centre for Academic Information Services has the lawful right to digitize the content to be uploaded into Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis if required for use by other parties for academic purposes or by other Higher Learning Institutes.
- No dispute or any claim shall arise from the student himself / herself neither a third party on this Project/Thesis once it becomes the sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student himself/herself without first obtaining approval from UNIMAS.

Student's signature _____ (Date) 15/5/2019

Supervisor's signature: _____ (Date) 15/5/19

Current Address:
K.G. NARAWANG, 89308 RANAU, SABAH.

Notes: * If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the date of restriction indicated, and the reasons for the confidentiality and restriction.

[The instrument was prepared by The Centre for Academic Information Services]

Acknowledgement

I would like to deliver my heartiest gratitude to all of person who involve directly or indirectly in my Final Year Project. First of all, special thanks to my Final Year supervisor Dr. Phang Piau for all his assistance, guidance and suggestion supervision throughout my Final Year Project. Not forgetting my examiner, Sir Terrin Lim for all the information and knowledge shared during every presentation.

Next appreciation is goes to my family member, my parent and my older sister, Ivena Nashion. I managed to continue my studies with their support and encouragement during hard times. Furthermore, all my needs during my studies in UNIMAS are being provided by both my parent and my older sister.

Finally, I would like to show my appreciation to all of my friends who was always there with me during finishing our Final Year Project and all the subjects in our last semester.

Abstract

Many of the existing data visualization tools are using PHP or other plug-ins to generate a graphical data representation. The purpose of data visualization is to help data analyst to understand better on what the behavior of the datasets .Therefore, a free statistical language R as a platform is proposed. The objective of data visualization for the vaccination coverage is to provide a better insight on how the vaccination coverage growth. The data visualization help data analyst to easily discover the changes in behavior of datasets and how the parameters related to each other. This data visualization process is using the powerful R statistical to generate graphical output such as line plotting, scatter plot and median quartile range plotting. To carry out this research, a research-based methodology was used. It consist of five phases that is data collection, problem identification, solution implementation, analysis of solution and problem reporting. The solution will show the variations of vaccination coverage across countries by using data visualization tools in R, the association between socioeconomic/demographic factors and vaccination coverage using clustering techniques and an explanatory analysis of the global immunization programs.

Abstrak

Terdapat pelbagai jenis perisian yang digunakan untuk tujuan data visualisasi pada masa kini dengan menggunakan *PHP* atau ***plugin.*** Perisian tersebut mempamerkan data dalam bentuk gambar rajah. Tujuan data visualisasi adalah bagi memudahkan para penganalisis data untuk memahami data mereka dengan lebih jelas. Objektif data visualisasi bagi liputan vaksin adalah untuk memberi penjelasan yang lebih mendalam mengenai perkembangan program vaksinasi. Melalui data visualisasi juga, para penganalisis data dapat mengkaji perubahan yang berlaku dalam data dan setiap perubahan dalam parameter yang berkaitan antara satu dengan yang lain. Data visualisasi ini menggunakan perisian R yang melibatkan banyak algoritma statistik dan menghasilkan gambar rajah yang menarik, contohnya, plot garisan, plot berselerak dan plot kuartil median.Kajian ini dijalankan berdasarkan satu kaedah yang berasaskan penyelidikan . Kaedah ini terdiri daripada lima fasa iaitu pengenalan masalah, pengumpulan data, pelaksanaan penyelesaian, analisis penyelesaian dan laporan berdasarkan analisis yang telah dilakukan. Kajian ini akan menghasilkan variasi liputan vaksin di seluruh negara dengan menggunakan visualisasi data dalam perisian R, hubungkait antara faktor sosioekonomi/ demografik dan liputan vaksinasi menggunakan teknik pengelompokan dan analisis mengenai program imunisasi global.

Table of Contents

CHAPTER 1 Project Proposal

1.0 Title

A Data-Driven Analysis On The Global Vaccination Coverage

1.1 Introduction

Tyron, 1939 was the first man to use clustering analysis technique that involved a number of different algorithms and methods where the object of similar type are grouped together and distributed into specific categories. A common question facing researchers in various fields of research is how to re-arrange the observed data into more organize and useful structure, that is, to produce scientific classifications. Cluster analysis is an exploratory data analysis mechanism which targeting at categorizing different objects into categories in a way that the degree of association among two objects either is maximal if they belong to the same group or in different group if its minima. Hence, cluster analysis can be used to explore the structures in data without an explanation or interpretation. Cluster analysis simply explore structures in data without revealing why they exist.

Vaccination is the administration of a vaccine to stimulate an individual's immune system to develop adaptive immunity to a bacterium. This vaccination is important to a human as everyone are exposed to a various type of harmful bacteria. Socioeconomic is a measure of the combined economic and social status. In addition, socioeconomic also tends to be positively associated with better health. For example, Beard et al. found that double of the amount of children with a recorded vaccination objection and either no or at least vaccinated once was recorded in 2013. Those children were living in the top 10% of postcodes ranked by economic resources, in contrast with those who were living in the bottom 10% (1.9% vs 1.1%) for the Victorian. This information can be clustered and visualized in order to make the researcher understand their trend easily. For instance, model-based clustering can be used to present the

correlation between two different parameters. The data will be clustered and visualized using R. Data visualization is another way to explain the statistical data analysis which is more convenient.

1.2 Problem Statement

Incomplete immunisation coverage might cause severe public health issue in both developing and developed countries. This is simply because the vaccine-preventable childhood diseases should otherwise be eradicated if the vaccination coverage is able to provide herd immunity to the society. Some of the socioeconomic factors, for instance, high-income populations, parental education level, employment status, and workplace and age can influence global vaccination coverage. Countries that have low daily income mostly have low rate of vaccination. This is because, peoples who live in such an area cannot afford the cost that needed for vaccination especially when they have a lot of children in a family. Hence, identifying the factors that might modulate vaccination coverage could increase the effectiveness of immunization programmes and help the policy makers in their decision making.

World Health Organization's (WHO) is targeting that one-third of countries worldwide reaching an immunization rate of 90% by 2015) for the Global Vaccine Action Plan (GVAP). However, many countries had not yet reach this immunization rate even though the prevalence of vaccine-preventable diseases was decreasing uniformly. Therefore, this project focuses on investigating the association between socioeconomic factors and global vaccination coverage by carrying out a data-driven analysis in R.

1.3 Objectives

There are three primary aims of this study:

a) To examine the variations of vaccination coverage across countries by using data visualization tools in R.

b) To investigate the association between socioeconomic factors and vaccination coverage using clustering technique.

c) To present an explanatory analysis of the global immunization programs.

1.4 Research Methodologies

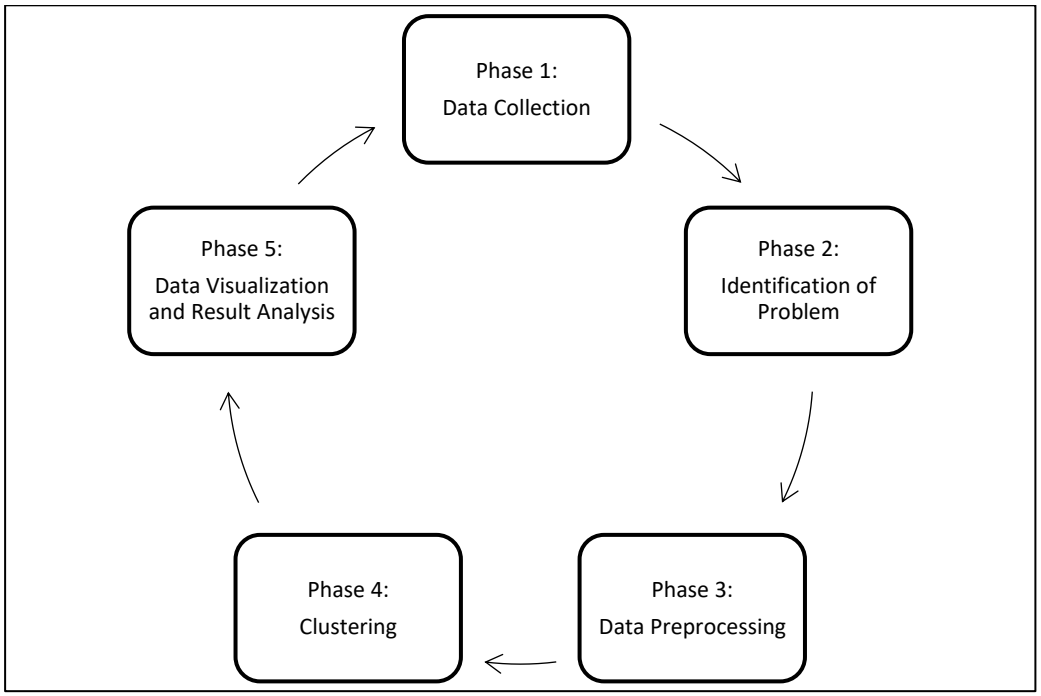


Figure 1.1 The Research Methodology phases

This project is a research-based methodology. Therefore, the methodology which is presented in this project will follow the research methodology. The research starts by conducting a literature survey on the related topic. The data was obtained from the Gapminder (Phase 1).

After doing some literature review, the identification of the problem is conducted. All the data will be collected and analysed using the specific technique. The main problem is identified. Next is a discussion of the simulation platform that will be used in this project (Phase 2).

The pre-processing data was conducted in order to transform the raw data into a meaningful and understandable format. As most of the real data contain incomplete or inconsistent or lacking certain behaviour or trends, it is important to conduct the pre-processing for further process. (Phase 3)

In this phase the clustering technique is used to examine the correlation between the vaccination coverage and the socioeconomic factors. For choosing among lots of socioeconomic factors, the linear regression is used to see which factors that has strong impact to the vaccination coverage (Phase 4).

The visualized data will be analysed. The changes and the differences obtained from the data-driven which has been visualized will be taken into consideration. Finally, the visualized data and the final result will be presented for reporting. The result is shown (Phase 5).

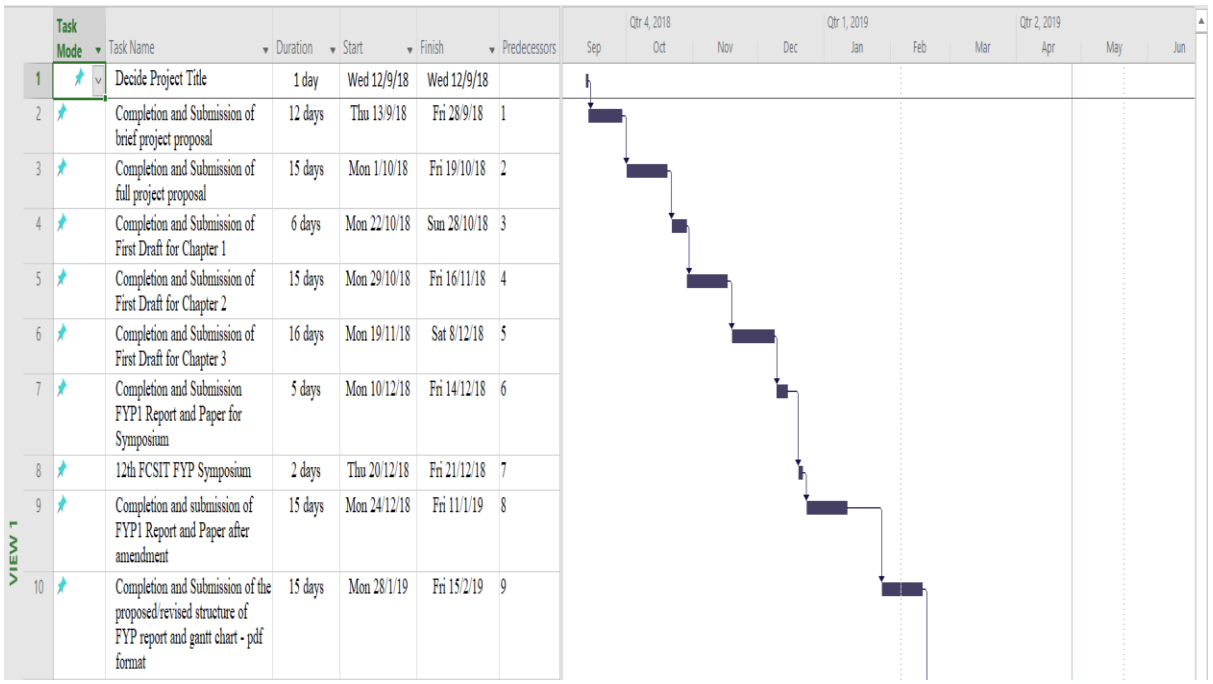
1.5 Scope

The scope of this project is the data visualization of percentages of DTP3 coverage immunized of one year olds children over the world at year 2011. In addition this project also producing a data-driven analysis of the association between the global vaccination coverage rate and the socioeconomic and demographic factors using clustering in R.

1.6 Significant of Project

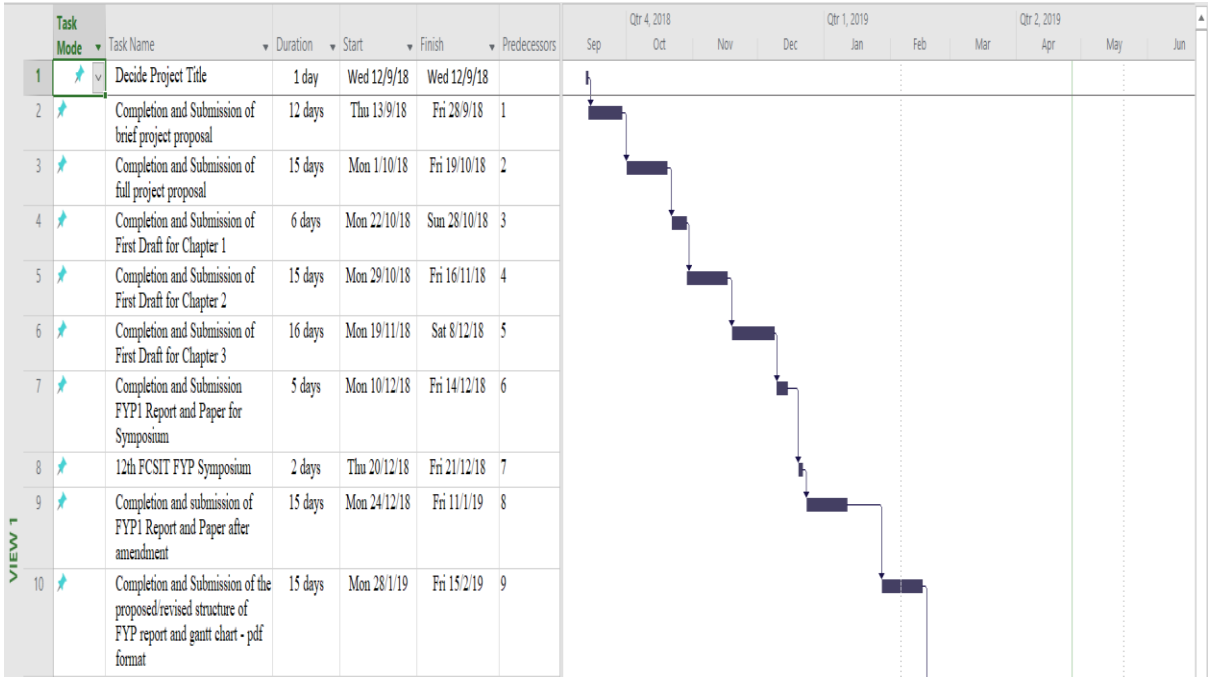
This project aims at performing a data-driven analysis on the global vaccination coverage datasets, obtained from Gapminder. As the datasets are huge and contain a lot of missing values, some data wrangling procedures have to be carried out using tidy verse and zoo packages in R. Besides, as the datasets contain vaccination coverage across different countries, it is a multivariable data which needed to be analysed by applying clustering techniques. Apart from that, the association between the vaccination coverage and socioeconomic factors can be obtained using the statistical and visualisation tools in R.

1.7 Project Schedule



	Task Mode	Task Name	Duration	Start	Finish	Predecessors
1		Decide Project Title	1 day	Wed 12/9/18	Wed 12/9/18	
2		Completion and Submission of brief project proposal	12 days	Thu 13/9/18	Fri 28/9/18	1
3		Completion and Submission of full project proposal	15 days	Mon 1/10/18	Fri 19/10/18	2
4		Completion and Submission of First Draft for Chapter 1	6 days	Mon 22/10/18	Sun 28/10/18	3
5		Completion and Submission of First Draft for Chapter 2	15 days	Mon 29/10/18	Fri 16/11/18	4
6		Completion and Submission of First Draft for Chapter 3	16 days	Mon 19/11/18	Sat 8/12/18	5
7		Completion and Submission FYP1 Report and Paper for Symposium	5 days	Mon 10/12/18	Fri 14/12/18	6
8		12th FCSIT FYP Symposium	2 days	Thu 20/12/18	Fri 21/12/18	7
9		Completion and submission of FYP1 Report and Paper after amendment	15 days	Mon 24/12/18	Fri 11/1/19	8
10		Completion and Submission of the proposed/revised structure of FYP report and gantt chart - pdf format	15 days	Mon 28/1/19	Fri 15/2/19	9

Figure 1.2 Project Schedule



	Task Mode	Task Name	Duration	Start	Finish	Predecessors
1		Decide Project Title	1 day	Wed 12/9/18	Wed 12/9/18	
2		Completion and Submission of brief project proposal	12 days	Thu 13/9/18	Fri 28/9/18	1
3		Completion and Submission of full project proposal	15 days	Mon 1/10/18	Fri 19/10/18	2
4		Completion and Submission of First Draft for Chapter 1	6 days	Mon 22/10/18	Sun 28/10/18	3
5		Completion and Submission of First Draft for Chapter 2	15 days	Mon 29/10/18	Fri 16/11/18	4
6		Completion and Submission of First Draft for Chapter 3	16 days	Mon 19/11/18	Sat 8/12/18	5
7		Completion and Submission FYP1 Report and Paper for Symposium	5 days	Mon 10/12/18	Fri 14/12/18	6
8		12th FCSIT FYP Symposium	2 days	Thu 20/12/18	Fri 21/12/18	7
9		Completion and submission of FYP1 Report and Paper after amendment	15 days	Mon 24/12/18	Fri 11/1/19	8
10		Completion and Submission of the proposed/revised structure of FYP report and gantt chart - pdf format	15 days	Mon 28/1/19	Fri 15/2/19	9

Figure 1.3 Project Schedule (continue)

1.8 Expected Outcome

A data visualization that created using R that represents the correlation between socioeconomic factors and vaccination coverage. The data visualization will show how many people that have been vaccinated according they're socioeconomic in various area. The visualized data will be used as an explanatory analysis of the global immunization programs.

1.9 Project Outline

Chapter 1

The purpose of this project is to provide explanatory analysis of the global immunization programs. By using the statistical tool in R, the variations of vaccination coverage across countries will be visualize. In order to achieve this, the existing technologies to create a number of data visualization type will be reviewed. For instance, the type of data visualization. This project adopt research-based methodology that will further discussed in Chapter three.

Chapter 2

Chapter 2 discusses about the problem that arise concerning to the vaccination coverage and the correlation between the vaccination coverage and the socioeconomic factors that cause the incomplete immunization done in certain country. In addition, this chapter also reviewing about the review done on existing techniques and platform used for solution similar to the proposed project. The overall study is done based on articles, journals and conference papers. The investigation that has been done to see the association between socioeconomic factors and vaccination coverage from existing research. At the end of the chapter, a brief description on the technology tools utilized for the execution of this project.

Chapter 3

Chapter 3 portrays about the methodology utilized for the development of this entire project. The research-based methodologies will be used as a model to provide the visualized data-driven analysis. This chapter explain about step-by-step taken in order to visualize the data correctly. Hence, the explanatory data can be presented correctly.

Chapter 4

This chapter focuses on the results and analysis. In addition, chapter 4 also will discuss on how to visualize the data correctly. The correlation between socioeconomic factors and the vaccination coverage also will be visualized to show the trend in an area over the world.

Chapter 5

This chapter summarizes the result explanatory with visualization of data-driven analysis on the global vaccination coverage. This chapter concludes this project report by revisiting the objectives and list down possible future works. A conclusion for this report is wrapped up in the end of this chapter, hence concluding the project.

Chapter 2 Literature Review

2.0 Introduction

This chapter provide an insight into various methods used to examine the variations of vaccination coverage across countries by using data visualization tools. Firstly, this chapter will look into the method used to analyze the data sets and how existing technology provides graphical representation of any problems. To create a graphical data representation, analysis of data sets and how existing technology tools interpret these into a graphical data visualization is conducted in this chapter

This chapter will also explain the data visualization tools that have been used by the data analyst to represent their finding or data in graphical form. A study by Schiano and Tversky (1989) showed that charts or maps are considered to be more symmetrical than they actually were [1]. Hence, the correlation between the socioeconomic factor and the vaccination coverage will also be visualized with a graphical visualization as a data-driven explanatory analysis. A summary of a review on existing technologies features is added as a conclusion at the end of this chapter.

2.1 Vaccination coverage

The aim of the World Health Organization's (WHO) Global Vaccine Action Plan (GVAP) is that one-third of countries worldwide have yet to reach an immunization rate of 90% by 2015, as the prevalence of vaccine-preventable diseases was decreasing uniformly [2]. There are several socioeconomic factors that are associated with growth in vaccination coverage. For instance, the parent's age, education levels, poverty, religions, children gender and the distance from the medical facilities [3]. Socioeconomic wealth index, for example, the ownership of assets as furniture and household characteristics [4] are some of the factors that correlate with