

Southeast Asia Happiness Report in 2020 Using Exploratory Data Analysis

Prismahardi Aji Riyantoko¹

Department of Data Science
UPN “Veteran” Jawa Timur
Surabaya, Indonesia
prismahardi.aji.ds@upnjatim.ac.id

Kartika Maulida Hindrayani³
Department of Information System
UPN “Veteran” Jawa Timur
line 3-City, Country
kartika.ds@upnjatim.ac.id

Tresna Maulana Fahrudin²

Department of Data Science
UPN “Veteran” Jawa Timur
Surabaya, Indonesia
tresna.maulana.ds@upnjatim.ac.id

Eristya Maya Safitri⁴
Department of Information System
UPN “Veteran” Jawa Timur
line 3-City, Country
maya.si@upnjatim.ac.id

Abstract— *The happiness index to be one of part to presents that each country has indicator which affect each other. Many countries have a basic indicator to determine that happiness score, there are economy sector, social support, trust to the government, generosity, and measure life satisfactions. The indicator is presents in the dataset, it means need to explore, analysis, and visualization to give knowledge to the other people. Data science is one knowledge field to determine data. Exploratory data analysis (EDA) is part of data science process. In this works, we present happiness report in the Southeast Asia region using the dataset World Happiness Report 2020. The results, we describe and discuss the dataset using table with column and value or score, the other we using bar-plot, correlation bar-plot, bar-plot analysis, and map-plotting visualization. Output of EDA is only recommendation to next parts in the Data Science process, minimum has knowledge to reducing data, merge data, cleansing data, visualization data to be based of knowledge to build data modelling.*

Keywords—*EDA, Happines Reports, Data Science, Visualization*

I. INTRODUCTION

About 4.2 billion people, more than half of the world's population (55.3 per cent), are living in urban areas today. By 2045, this figure is estimated to increase by 1.5 times, to more than six billion [1]. There were 371 cities with more than one million inhabitants at the turn of the century in 2000. In 2018, there were 548, and in 2030, a projected 706 cities will have at least one million inhabitants. During the same time, the number of so-called mega cities – cities that have more than ten million inhabitants, most of which are located in the Global South – is expected to increase from 33 to 43, with the fastest growth in Asia and Africa. Today, Tokyo (37.4 million), New Delhi (28.5 million), and Shanghai (25.6 million inhabitants) are the most populous cities worldwide [2].

The happiness index is a comprehensive survey instrument the assesses happiness, well-being, and aspects of sustainability and resilience [3]. The happiness index has many indicator there are the feeling of happiness, measures life satisfactions, gross domestic product (GDP), social support or family, generosity, trust (government corruptions), and the other happiness domains.

Exploratory Data Analysis (EDA) is part of the data science process. It's very important before performing feature engineering and modeling the data because at this stage we have to understand the typical data first. Data science process have many steps, there are data preparation, data cleansing, exploratory data analysis, feature engineering, data modeling, data evaluation, and deployment. In this case, probably EDA was used to knowing the contents of data from distribution data, frequency, correlation, and the others. In the experiment, curiosity is more important in this case, to understanding the main idea of the data need to check, because it will be can answer the deep of the problem.

In this works, after the data is collected, then for the next step is data is analyzed. Analysis data is the process of arranging the sequences of data, organize them into a pattern, category, and basic description units. Based on the main concept, data analysis work includes three steps namely preparation, tabulation, and application of the data accordingly with a research approach that is processing data by using formulas or rules existing according to the research approach. Hence, we can conclude that data analysis is a process organize data, perform synthesis, processing data using formulas or the rules according to the approach and draw conclusions.

Before the data is processed into data modelling, the data should be analyzed using exploratory data analysis. Certainly, in this research the data is obtained from Kaggle using data

World Happiness Report. The contains of the data there are about the country, regional indicator, social support, GDP, and et all. In the other hand, for this research will be focus in the data analysis at Southeast Asia.

II. LITERATURE REVIEW

A. Happines Index

The happiness index is a tool for the use of researchers, community organizers, and policy makers seeking to understand and enhance individual happiness, community well-being, social justice, economic equality, and environmental sustainability [3]. Community organizers, students, researchers, educators, government and the others that's is the index of happiness to promote social change by making the survey instrument. The index is unique in that it is the only widely comprehensive index available for free online that survey takers to access their own scores in comparison the entire data set, while also allowing users to customize the tool for a target population, add their own questions to the survey instrument, and readily access data for their own sample as well as comparison data to the entire data set. [3]

Historically, government have used GDP as a primary indicator of national well-being and growth [4]. For example, GDP does not take into account the hidden costs of economic development, such as inflation and unemployment, and an overemphasis of GDP diminishes the value of important well-being factors such as natural capital, knowledge, health, and social capital [4]. Measuring happiness, therefore, should not only consider observable objective well-being measures (e.g., health and socioeconomic status), but also subjective well-being measures, such as domain satisfaction and quality of life.

B. Exploratory Data Analysis

EDA is data analysis approach. It's has three popular data analysis approach as follow classical, exploratory (EDA), and Bayesian. These three approaches are similar in that they all start with a general science/engineering problem and all yield conclusions. The difference is the sequences and focus of the intermediate steps [5]. For the first classical analysis is begin from *problem, collecting data, data modelling, data analysis, and conclusion*. The second EDA has sequences, begin from *problem, collecting data, data analysis, data modelling, and conclusions*. The last approach is Bayesian was started from *problems, collecting data, data modelling, prior distribution, data analysis, and conclusions*.

The exploratory data analysis approach does not impose deterministic or probabilistic models on the data. On the contrary, the EDA approach allows the data to suggest admissible models that best fit the data. The primary goal of EDA is to maximize the analyst insight into a data set and into underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as a good fitting, a list of outliers, a sense of robustness of conclusions, estimate for parameters, uncertainties for those estimates, a ranked list of important

factors, conclusions as to whether individual factors are statistically significant, and optimal settings [5].

III. METHODOLOGY

In this paper using the exploratory data analysis (EDA) method, where EDA is an approach to extract the information enfolded in the data and summarize the main characteristic of the data.

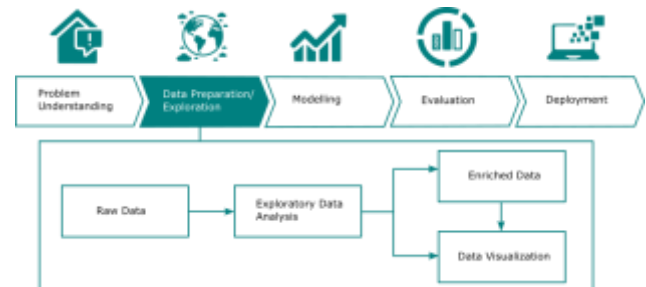


Fig.1 Exploratory Data Analysis Method

In Fig 1 above, the assume exploration data analysis start from a problem. We use dataset *World Happiness 2020 (csv)* from Kaggle, and the content there are country, regional indicator, ladder score, standard error of ladder score, upper-whisker, lower-whisker, logged GDP per capita, social support, freedom to make life choices, and et all. Because the researchers where lived in the Southeast Asia, hence this works will be focus in the country of Southeast Asia.

In this section, we will describe the steps by steps to explore the data, from get the data until visualization. *Part I*, we set the dataset as csv file and import into jupyter notebook. Furthermore, we set many library to analyzed and visualization the data, as follow using *numpy, pandas, matplotlib, seaborn, plotly-express, plotly-graph*. After set up the data, we change the name in many columns, why we change those, because the data need to generalize perception, so the others reader can understand of the purpose each column in the data. In the other hand, the data is sorted list based on rank world happiness report

Part II, the columns was no need to analysis is removed by us, so we leave a few columns to analyzed, there are *Country, Regions, Happiness Score, GDP, Family, Health, Freedom, Trust (Government Corruption), and Generosity*. Before to the next step, in this part we make sure the data is completed, means no null value in each column. *Part III*, we used the query data to classified the country only in Southeast Asia. In this data is used by us, only has nine country in the Southeast Asia. The details data is showed in the Table I below.

TABLE I. LIST OF COUNTRY IN THE SOUTHEAST ASIA

Country	Regions	Happiness Score	...	Generosity
Singapore	Southeast Asia	6.3771	...	0.029879
Philippines	Southeast Asia	6.0060	...	-0.105463
Thailand	Southeast Asia	5.9980	...	0.268685
Malaysia	Southeast Asia	5.3843	...	0.114727
Vietnam	Southeast Asia	5.3535	...	-0.094533
Indonesia	Southeast Asia	5.2856	...	0.519587
Laos	Southeast Asia	4.8886	...	0.111371
Cambodia	Southeast Asia	4.8484	...	0.051911
Myanmar	Southeast Asia	4.3080	...	0.560664

To visualization the data is used the correlation was plot by seaborn library. The correlation was use compare between each value of column with others. For example, the Happiness Score compared with Family, in this case we should the value between impact the happiness score against the family. The other visualization, we used bar chart to bring out the distribution of all numerical data. Hence, we can analyze each part of columns. For the last case, we describe each column and visualize using table, vertical bar plot, and map plot.

IV. RESULTS

In this section, we describe and discuss about the results of the exploratory data analysis in the Southeast Asia, about the Happiness Score, Health (Life Expectancy), Economy (GDP per Capita), Family, Freedom, Generosity, Trust (Government Corruption).

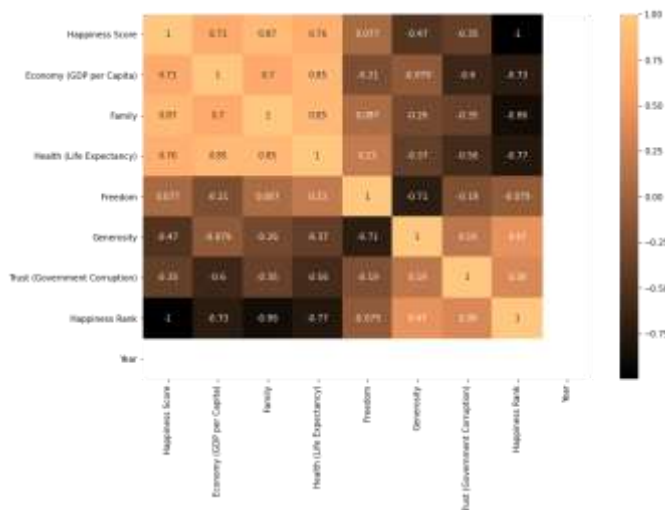


Fig.2 Correlation Data Value in Southeast Asia

Fig 2 show the correlation about one entity with the other entity in Southeast Asia. The happiness index is affected with many entity, especially economy (GDP per capita). It's describe

the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period. Actually, the economy condition in each country has big impact for the happiness index.

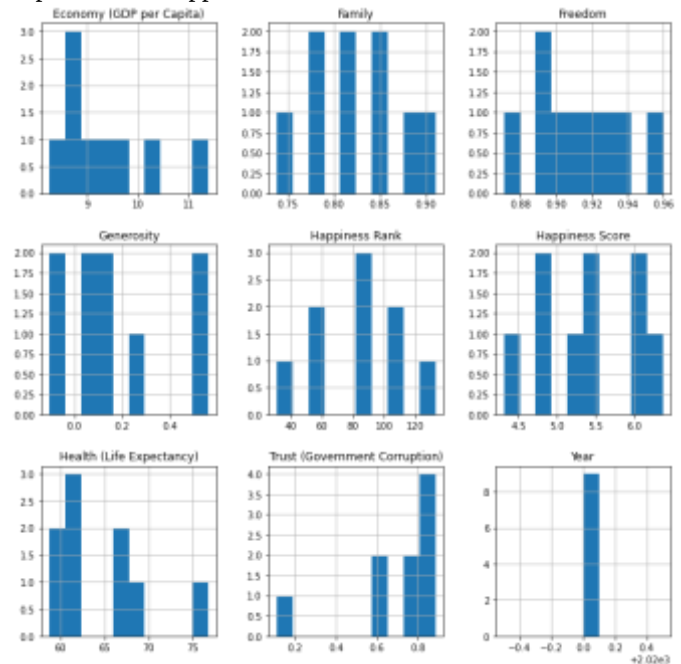


Fig.3 Distribution of All Numerical Data in Southeast Asia

Fig 3 describe about all numerical distribution data in Southeast Asia. Many country in Southeast Asia has a good indicator for happiness rank. Especially, Singapore can reach in 30th rank in the World. This country has a good indicator, there are in economy (GDP per capita), freedom, and trust (government corruption).

A. Happiness Score

TABLE II. LIST OF HAPPINESS COUNTRIES

Country	Happiness Score	Rank
Singapore	6.3771	1
Philippines	6.0060	2
Thailand	5.9980	3
Malaysia	5.3843	4
Vietnam	5.3535	5
Indonesia	5.2856	6
Laos	4.8886	7
Cambodia	4.8484	8
Myanmar	4.3080	9

Table 2 describe about happiness score in the Southeast Asia. Singapore has high score compared with the other countries. But, the value of each country does not differ much, hence we can know that each countries in this regional has a good happiness index.

Fig 3 describe the value of happiness score using bar plot. The interesting of bar plot, it's presents category data with rectangular bars with height or length proportional. In the other hand, Indonesia is stand in 6th rank at Southeast Asia. Those

country should be increased the happiness index with the other manner of the government.

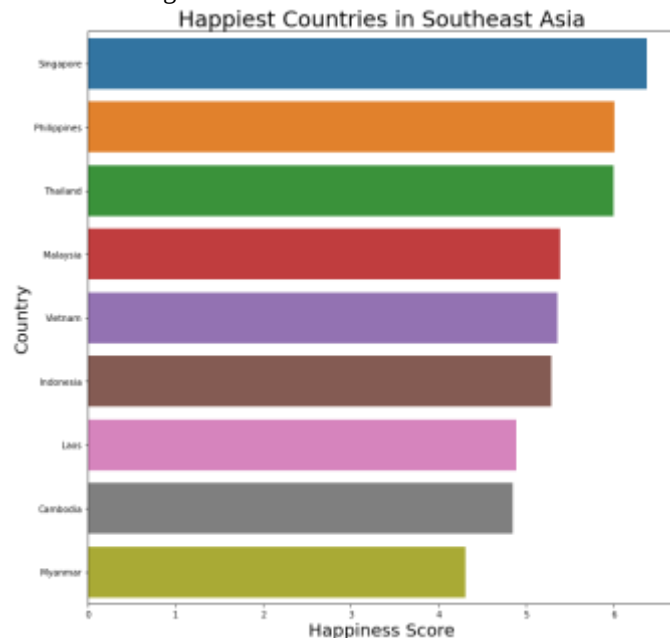


Fig.3 Bar Plot Happiest Countries in Southeast Asia

Fig 4 describe the value of happiness score using the procedure of mapping with Pandas Data-frame. Geo-pandas

B. Rank Countries Based on Indicator

TABLE III. LIST OF HELTHIEST COUNTRIES

Rank/ Countries	Indicator						
	Happiness Score	Health	Economy	Family	Freedom	Trust	Generosity
Cambodia	8	7	9	8	1	4	6
Indonesia	6	5	4	6	9	2	2
Laos	7	9	7	9	5	8	5
Malaysia	4	4	2	5	8	3	4
Myanmar	9	8	8	7	7	7	1
Philippines	2	6	5	4	4	6	9
Singapore	1	1	1	1	3	9	7
Thailand	3	3	3	2	6	1	3
Vietnam	5	2	6	3	2	5	8

In the previous section we already describe the happiest countries in the Southeast Asia. In the Table 3, we presents the rank based on the indicator was affected the happiness score.

Singapore has four best indicator better than the other countries in the Southeast Asia. This country is got 1st rank in the happiness score, health (life expectancy), economy (GDP per capita) and Family (Support Social). Apparently, Singapore indeed has a good government and the country in the middle of world trade path. Hence, the government can increased the economy condition to be better. Furthermore, Myanmar to be a country has lowest happiness score in the Southeast Asia. Laos presents a bad healthy and social support or family condition in this regional, probably the government not yet give an a good rules to their citizenship. For further, Cambodia has a bad economy sector.

makes working easier with geo-spatial data. Because, it combines the capabilities of Pandas and shapely by operating a much more compact code. It's one of the best ways to get started with making choropleth maps.



Fig.4 Map Plot Happiest Countries in Southeast Asia

Well, the happiest of country in Southeast Asia is present using orange indicator in the maps. Hence, this the other interesting of visualizations.

For the freedom indicator, Cambodia to be a good freedom country. Probably, that means the citizenship has a good democracy, policy, and has high thinking and express to freedom of speech. Surprisingly, Indonesia to be a bad freedom in the country. Apparently, in this government era, freedom of speech will be sentenced for criticizing to the government. Maybe the Reformasi era back to the Orde Baru era, because the government staff mostly get from the old era, hence their old system is doesn't works.

Trust (Government Corruption) is one of indicator has a big impact to the happiness score. Thailand become a most country with high corruption problem in their country. In the other surprisingly, Indonesia stand in the 2nd corrupt of country. But, Singapore to be a good country to handle the governor or the others staff to didn't get opportunity to corruption.

In the last indicator, Myanmar has a highest generosity people. It's mean the people happiest for now, though in the last years, the Queen has a problem and showing the dictator power with deportation minority clan, it's called Rohingya out from their country. So far, the other details data presents different.

V. CONCLUSION

In this section, we will describe about conclusion. Well, the research only focus on exploratory data analysis using World Happiness Report and simplified to focus in the Southeast Asia. For the simplified focus discussion we using EDA with get dataset, reduces the columns, sorting list the happiness score in the world, and visualization. We make easier to understand the other people or reader, we use bar plot correlation, bar plot vertical, table, and map plotting.

For the next works, we interest to encourage the research in the one country with many years data before. Especially Indonesia, so we can using algorithm, machine learning to predicting each indicator for the next years.

REFERENCES

- [1] Adler, M. D., Dolan, P. and Kavetsos, G. (2017) "Would you choose to be happy? Tradeoffs between happiness and the other dimensions of life in a large population survey," *Journal of Economic Behavior & Organization*, 139, 60-73.
- [2] Ambrey, C. L., Fleming, C. M. and Chan, A. Y.-C. (2014a) "Estimating the cost of air pollution in South East Queensland: An application of the life satisfaction non-market valuation approach," *Ecological Economics*, 97, 172-1
- [3] Musikansi, L., Cloutier, S., & et all (2017). "Happiness Index Methodology". *Journal of Social Change*, Volume 9, Issue 1, Pages 4-31.
- [4] Ovaska, T & Takashima, R. (2006). Economic policy and the level of self-perceived well-being: An international comparison. *The Journal of Socio-Economics*, 35, 308–325. doi:10.1016/j.socec.2005.11.054
- [5] <https://www.itl.nist.gov/div898/handbook/toolaids/pff/eda.pdf>. 24 Nov

