Development of Automated Environmental Data Collection System and Environment Statistics Dashboard[*]

Dede Yoga Paramartha[1], Ana Lailatul Fitriyani[1], and Setia Pramana[1,2‡]

[1]Directorate of Analysis and Statistics Development, BPS Statistics Indonesia, Jakarta, 10710, Indonesia.
[2]Associate Professor, Politeknik Statistika STIS, Jakarta, 13330, Indonesia.
[‡]corresponding author: setia.pramana@stis.ac.id

Abstract

Environmental data such as pollutants, temperature, and humidity are data that have a role in the agricultural sector in predicting rainfall conditions. In fact, pollutant data is common to be used as a proxy to see the density of industry and transportation. With this need, it is necessary to have automated data from outside websites that are able to provide data faster than satellite confirmation. Data sourced from IQair, can be used as a benchmark or confirmative data for weather and environmental statistics in Indonesia. Data is taken by scraping method on the website. Scraping is done on the API available on the website. Scraping is divided into 2 stages, the first is to determine the location in Indonesia, the second is to collect statistics such as temperature, humidity, and pollutant data (AQI). The module used in python is the scrapy module, where the crawling is effective starting from May 2020. The data is recorded every three hours for all regions of Indonesia and directly displayed by the Power BI-based dashboard. We also illustrated that AQI data can be used as a proxy for socio-economic activity and also as an indicator in monitoring green growth in Indonesia.

Keywords: big data, environment statistics, pollutant, socio-economic, web scraping

---

[*] Received: Jan 2021; Reviewed: Jan 2021; Published: Jun 2021

1. Introduction

Indonesia, as a country that is still based on agriculture, will certainly be very sensitive to the importance of temperature and humidity data. The reason is based on research from the National laboratory of the U.S. Department of Energy (National laboratory of the U.S. Department of Energy, 2011), temperature and humidity are indicators that show the health and resilience of living things. In addition, temperature and humidity are able to affect the level of rainfall, respectively by 50 to 70 percent (Mawonike & Mandonga, 2018). This will certainly be very useful in making a long-term analysis in Indonesia regarding agriculture.

In addition, pollutant data (measured by Air Quality Index) is very useful for showing the health level of an area, given that pollutants have a high chance of developing diseases, such as shortness of breath or even mild to severe allergies. However, recently AQI data can also be used as a means of monitoring socio-economic conditions. Pollutants can erode traffic congestion and also industrial conditions. With this perspective, it can be said that AQI is a new monitoring tool for regions in Indonesia from a socio-economic perspective. In addition, the AQI can also show the impact of COVID-19 pandemic on the environment (Pramana et al., 2020). In his discussion, air quality can be a parameter in determining how environmental conditions are.

The problem that is currently happening is regarding public data that is easily accessible by data users. It is not uncommon to encounter difficulties in accessing environmental data such as temperature, humidity, and AQI which have been recapitulated from time to time. This makes it difficult to interpret the data as well as the insights contained in it because it only depends on a certain time, making environmental monitoring difficult. People find it difficult to see whether today in their environment there is a change in environmental quality or not. With the need for data for long-term analysis and monitoring, it will be closely related to the need for data that has a mandated volume and its presentation is also fast.

Current access to technology also supports to solve the problems, given the very large data lake of existing web sites. Access to technology is also developing in the direction of obtaining these data, such as the use of web-scrapping. With these conditions, it will be very easy to meet the needs that exist today. In this study, the development of Environmental Data Collection System was carried out by collecting data on real-time weather website based on sources from big data using the web scraping method. Several studies have been carried out regarding data collection using the web scraping method as a comparison or as complementary data for official statistical data. This is not new, considering that many previous studies have applied the same thing as in research on 'The potential of big data form official statistics in Indonesia by utilizing several data sources and for different purposes'. (Pramana et al., 2017).

Other research has also been applied to the BPS (Statistics Indonesia) regarding the use of big data. The use of pollutant data has been echoed by BPS in observing the conditions of mobility that occurred during the COVID-19 pandemic. Its purpose is none other than monitoring with the association between pollutants and motorized

vehicle activity(BPS Statistics Indonesia, 2020). This research is also developing to be more specific to big data analysis in social studies. According to BPS, the use of big data during a pandemic is a means of forming statistical studies with high efficiency. This study also has a role in confronting the statistics produced by BPS through publications and other releases. In its development, pollutant analysis in big data publications in social studies is to be linked to indicators such as population density and mobility(BPS Statistics Indonesia, 2020).

Not only is data collection being developed, the need for visualization that fast and easy to understand are also very important at this time. This will facilitate statistical literacy and dissemination of statistics. On this basis, current technology such as dashboards is an opportunity that can be developed. Thus, it can be compared between data retrieval which is classified as fast with a presentation that can be processed directly in the dashboard. Digital dashboards not only help you to drill down into the insights that matter most to your business, but they also offer an interactive visual representation that assists in swifter, more informed decision-making as well as the discovery of priceless new insights (Durcevic S, 2020).

The dashboard that is presented also has other benefits that can be developed. One of them is to perform quality control on certain official statistics. An example is BPS which uses phenomenon analysis which is used to strengthen some of the data released. This opportunity would certainly be very good to take advantage of. And developed, adapting to big data and also technological developments in supporting statistical processes based on data needs.

The aims of the study are to (1) build an automatics system of environmental data collection using web-scraping technology along with a real-time and informative environment statistics dashboard, and (2) analysis the association between Air Quality Index and the socio-economic activity in all provinces in Indonesia. This goal is based on the potential that big data development can do. Referring to the previous references, this research is a combination of data processing, processing and visualization. This causes this research to be more complex and comprehensive.

2. Methodology

In this study, data collection was carried out by taking environmental information data in all regions and weather statistics in Indonesia from the IQAir website (www.airvisual.com). The selection of this site is based on the completeness and credibility. In addition, this site has collaborated with several institutions and has been verified by the Indonesia Meteorological, Climatological, and Geophysical Agency (BMKG). The IQAir website also performs real-time updates on 1800 points in Indonesia. In this study, the workflow that is compiled is by defining the data on the site, Determining Web Scraping Technology, Flight Data Collection, Daily Scraping, Data Cleaning, and Build the Dashboard.

2.1. Variables Description

The Air Quality Index (AQI) is a pollutant state condition of an area which is calculated based on measurements of particulates (PM2.5 and PM10), Ozone (O_3),

Nitrogen Dioxide (NO2), Sulphur Dioxide (SO2) and Carbon Monoxide (CO) emissions. Based on the standards of the United States Environmental Protection Agency (EPA) the air quality index scale is divided into six categories, namely:

Table 1**:** AQI Categories

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Source: EPA.gov

Air temperature is the temperature of hot/cold air in a place at a certain time.
- Average temperature is the value of the average air temperature within a certain period, whether daily, monthly, or within a certain period of time.
- Air humidity is the concentration of water vapor in the air (atmosphere).
- Average humidity is the average humidity value in a certain period, whether daily, monthly, or over a certain period of time.
- Air pressure is the power that works to move the air mass within a certain unit area. Air pressure will be inversely proportional to the height of a place so that the higher the place above sea level, the lower the air pressure.
- Average - average air pressure is the average air pressure value over a certain period, whether daily, monthly, or over a certain period of time.
- Wind speed is a fundamental atmospheric quantity. Wind speed is caused by the movement of the wind from high pressure to low pressure, usually due to changes in temperature.
- Average wind speed is the average wind speed value over a certain period, either daily, monthly, or within a certain time span.

2.2. Determining Web Scraping Technology
The first thing to consider is understanding the data structure on the website. Thus, it can be determined the logic in retrieving the required data. After an inspection of the elements in the website on the https://www.iqair.com/id/indonesia website, it is known that before obtaining all Indonesian data, it is necessary to know the regional ID number. This is because the data can be accessed when it is entered specifically in the target area. This can be seen from the element structure in Figure 1.

Figure 1 : Elements to go to all place (cities and station) of Indonesia

Figure 2: API Form in IQAir Website

After heading to an address with a certain province, an API can be obtained which contains the required data. This can be seen in Figure 2. Thus, it is necessary to collect identity codes from regions in Indonesia. To get this ID, it can be accessed through the IQAir site-map on the https://www.airvisual.com/sitemap places-en-1.xml page.

2.3. Environmental Data Scraping
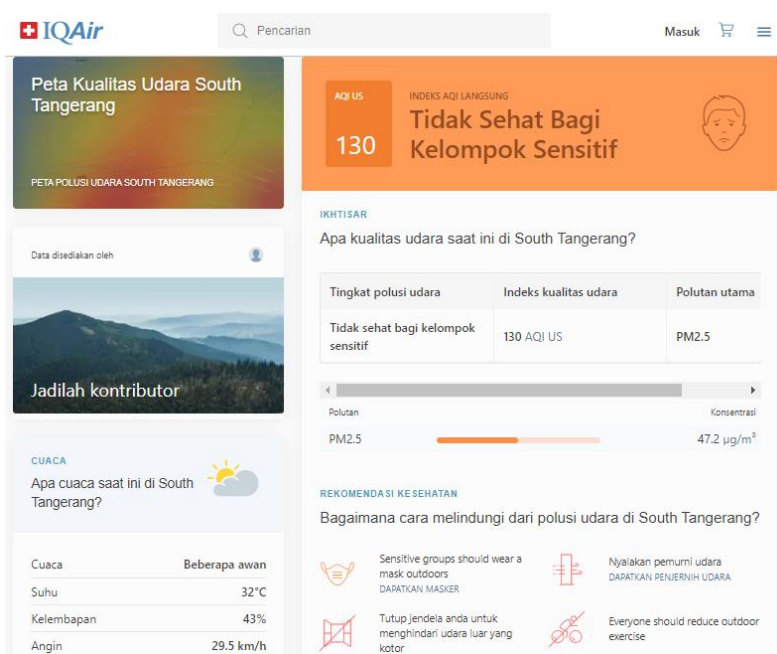
Figure 3: Example of the page that contains detailed information about Environmental Data

In this study, we use the Scrapy framework to do the job (Chaves A, 2020). We use Scrapy because it can run asynchronously, so the weather information gathering process can be done faster rather than using only the requests module in python. With the Scrapy module, any information in the image will be extracted by determining the location of the element in the HTML structure of the page. For example the city's code element (ID * 5b7a94b96a15c1ae2d8a9a31) as shown in the Figure 3. The location code has been obtained in the previous process on the available sitemap. In addition, weather stations are also collected so that the data provided is more complete.

2.4. Daily Scraping

The data collection process is scheduled to run every 3 hours. The collection is carried out with this range because it is necessary to carry out a process in taking the entire area to completion. If we experience unforeseen problems, we can take historical data but only for the daily average.

2.5. Data Cleaning

The next process is checking the completeness of the collected data. Completeness that needs to be considered is the city, the weather station, and the variables. In addition, checking was carried out in case of missing data, and resolved with scraper

repair and also periodic website checks to anticipate structural changes. After that, the variables that will be used in the dashboard are selected.

2.6. Build the Dashboard

Creating a dashboard is done using the Power BI application. This application can compile data directly in .csv form quickly. After that, the matching is done with the map data and also the selection of the variables to be displayed. In addition, a presentation format was chosen from the statistics obtained, as shown in Figure 4.
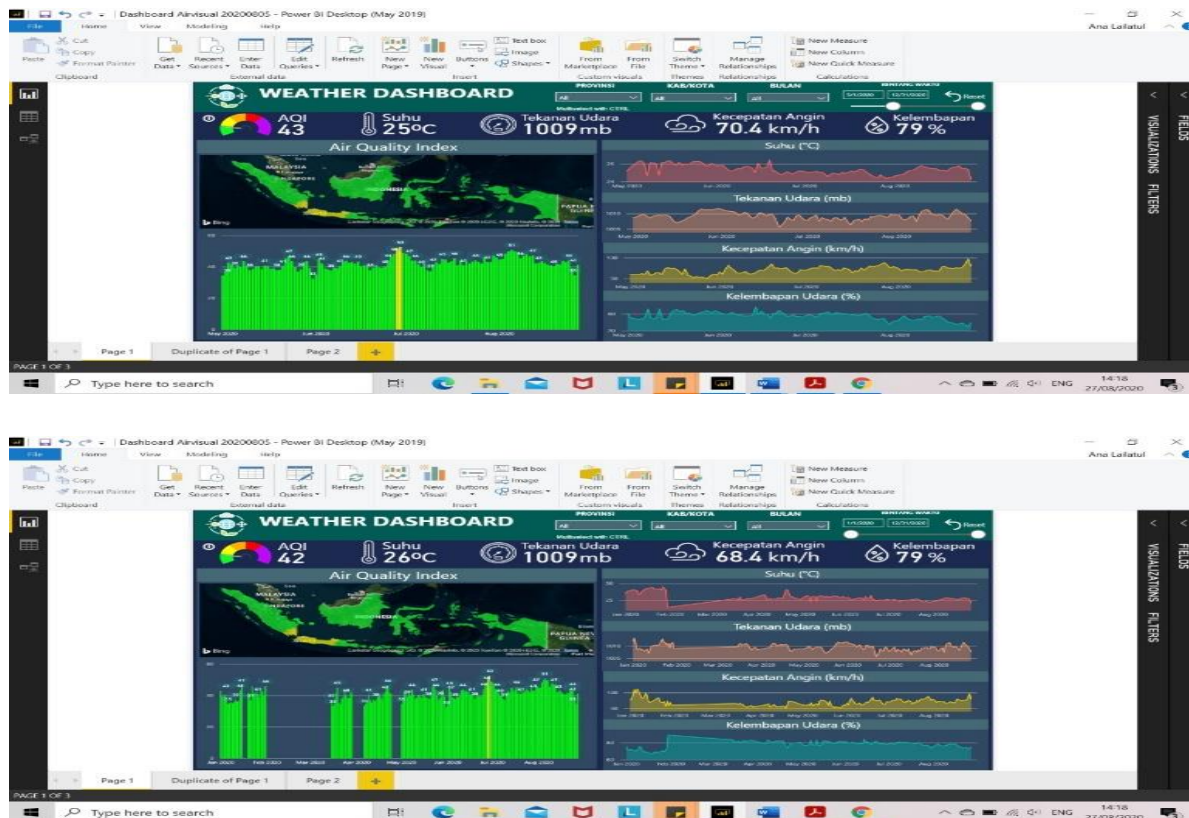
Figure 4: Environment Statistics Dashboard

2.7. Related Work

There are several previous studies that also highlight the making of monitoring for existing pollutant data. One of the related studies is from (Hananto & Putra, 2018), with a research entitled A Dashboard System for Monitoring Air Pollution in Surabaya based on PM2.5. This research aims to raise the awareness of the community and decision makers in order to solve the air-pollution problem, the level of the PM2.5 index should be monitored.

In addition, there is also another study by (Bogdan, 2015), entitled Temperature and Humidity Measurement System. One of the objectives to be achieved in this research is to achieve a functional system in terms of hardware and software that allows measuring and monitoring temperature, humidity and the time. With the monitoring carried out, it is hoped that many advantages will be gained from the data that has been collected.

3. Results and Findings

Based on the results that have been done, the analysis can be divided into three things. The first is about the role of the scrapper and the efficiency it provides, then about the dashboard, and the last is about the findings found from the use of environmental data.

3.1. Scrapper Efficiency

With the scrapping technology, the need for fast data is one that is easy to fulfil. The daily web scrapping the data about the environment can be obtained quickly. Thus, it is easier to provide the required statistics. Besides that, real-time data movement can be used as a basis for conducting certain analyzes. The data obtained is daily data with a time span every 3 hours. Every 3 hours, the dashboard can be updated to get the latest developments from environmental data. In addition, manual data retrieval from each region can be done more quickly. This large volume of data, in turn, will become a good asset to environmental statistics. From the efficiency measure carried out, it will refer to the efficiency of data collection costs, the speed of collection, and a little on the cost.

The obstacles that need to be considered in the future are regarding the storage and also long-term analysis. In terms of storage, it will be very draining if it is done for a long period of time. The long timeframe is also a challenge regarding the possibility of the analysis being carried out.

3.2. Dashboard

This dashboard provides information on air quality index data and weather data from the air quality information provider website www.airvisual.com. The data visualization displayed consists of air quality index data, temperature, air pressure, and wind speed and air humidity. The visualization of the numbers displayed in the top row is the average number based on the selected time frame, while the more detailed daily average figures can be seen in the graphic visualization below.

In the visualization of the air quality index data, a barometer is also displayed showing the category of the air quality index based on the color category, namely green indicates good air quality, yellow indicates moderate air quality, i.e. air quality is still acceptable but risky for those who are sensitive to air pollution. Orange indicates unhealthy air quality for sensitive groups, red indicates unhealthy air quality and everyone is starting to experience health effects, and purple indicates very unhealthy air quality. In addition, it can also be seen the distribution map of the air quality index category per region.

However, this distribution map can only display the distribution of air quality index categories to the provincial level only. Then for a more detailed daily quality index average figure, see the bar chart below, which can also be seen the category of the daily air quality index based on the color that appears on the bar chart. Meanwhile, to see the development of data on average daily temperature, daily average air pressure, daily average wind speed and daily average air humidity are displayed in the form of a line graph.

Presentation of data visualization in a dashboard can display data visualization in a comprehensive and interactive manner so that users can select variable categories according to their needs. In addition, this data visualization dashboard can also make it easier for users to see real-time data developments. The filter feature by region and time can make it easier for data users to choose the visualization they want to display according to user needs. The dashboard filter defaults that are displayed are data for the entire region and time. Regional filters are available up to the district/city level, while the time filters are available in two options, namely filter by certain months or filter by certain time ranges. Users can select more than one filter category and can return the visualization display back to the initial view by pressing the reset button located in the upper right corner.

3.3.   Association between Air Quality Index and Socioeconomic Activity

Based on the data that has been collected, there are several analyses that can be developed. In addition, it can also be a finding in observing existing phenomena. This is an advantage that can be juxtaposed with official statistics.
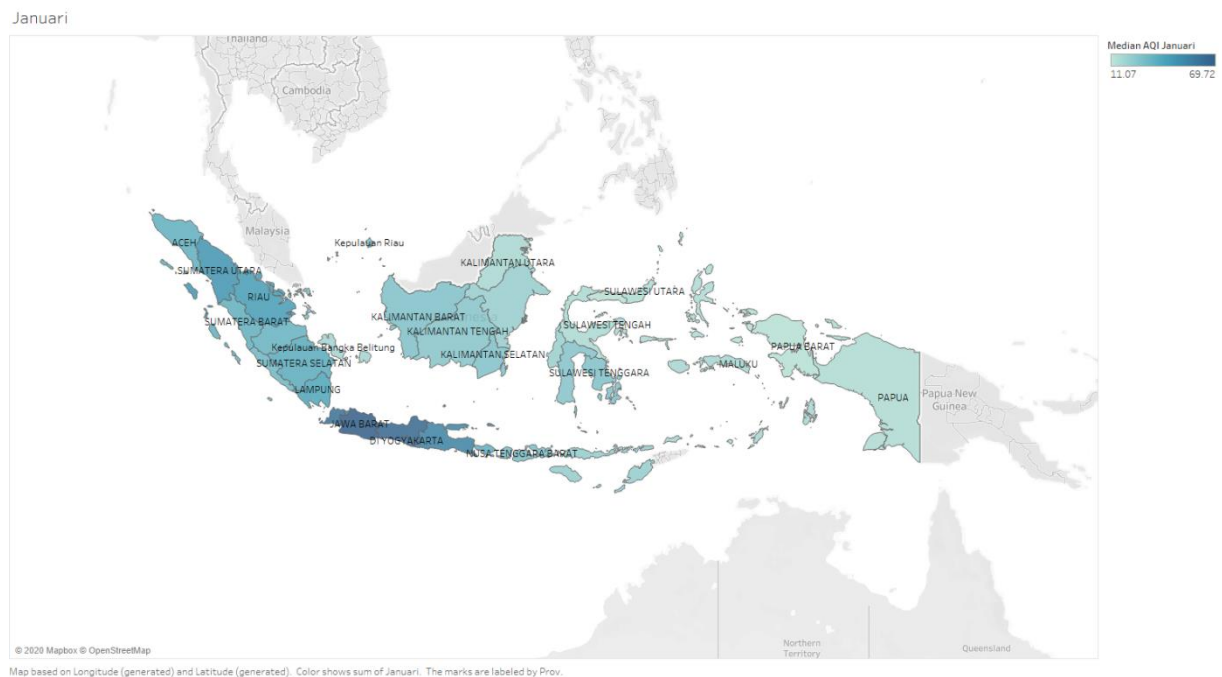
Figure 5: Distribution of Indonesia's Province Air Quality Index

AQI as a measure of pollutant density is a very useful indicator of the environment. Not only for the health aspect, this data is also a proxy for indicators of social and economic activity. If you look at the distribution of pollutant concentrations in Indonesia (Figure 5), it can be seen that the western region is an area with quite poor air quality compared to eastern Indonesia. The island with the highest density is Java, with an average area having moderate to unhealthy pollutant characteristics for sensitive population groups. Java Island is the island with the highest socio-economic activity compared to other islands in Indonesia. In other words, it is true that socio-economic activities will have an impact on the environment, however, if it still has a high effect, then indirectly Indonesia has not been able to optimize green economic growth and is

still struggling with the steam and electricity industry which is less environmentally friendly.

Apart from being used as a marker of high economic activity, pollutant level data can also illustrate the opposite. The Eastern Region is an area with very low pollutant levels. However, this also indicates that the development is experiencing imbalances between the western and eastern regions of Indonesia. With the character of Industry in Indonesia, it appears that the development has not been massive in the eastern region, especially in the Maluku-Papua island. This does not mean that the same development as in the western region must be carried out, but it is necessary to make use of the potential based on sustainable green growth.

Based on the explanation above, indirectly, air density can be a marker of what regional and sector bases are dominant or even need to be developed and improved so that in the future the resulting economic growth will be better. With fast data collection, this data can become an explicit monitoring of air quality and implicit socio-economic activities. Thus the means of providing a perspective on sustainable growth will be more numerous and of course fast.

In terms of phenomena, the dashboard created can illustrate how COVID-19 has little association with environmental conditions. This can be seen from the condition of the pollutants which have decreased several times during March to April.

Apart from the social aspect, the level of air pollution can also be a proxy for the size of the economy, such as regional income. This is very possible in Indonesia, given that the structure of Indonesia's GDP is dominated by the Industrial sector. If you look at its distribution, Java as the central economic island has the highest economic value compared to other regions. If we relate it to the level of air pollution, then the island of Java as an island that has a high level of pollutants is currently experiencing economic growth accompanied by environmental degradation.
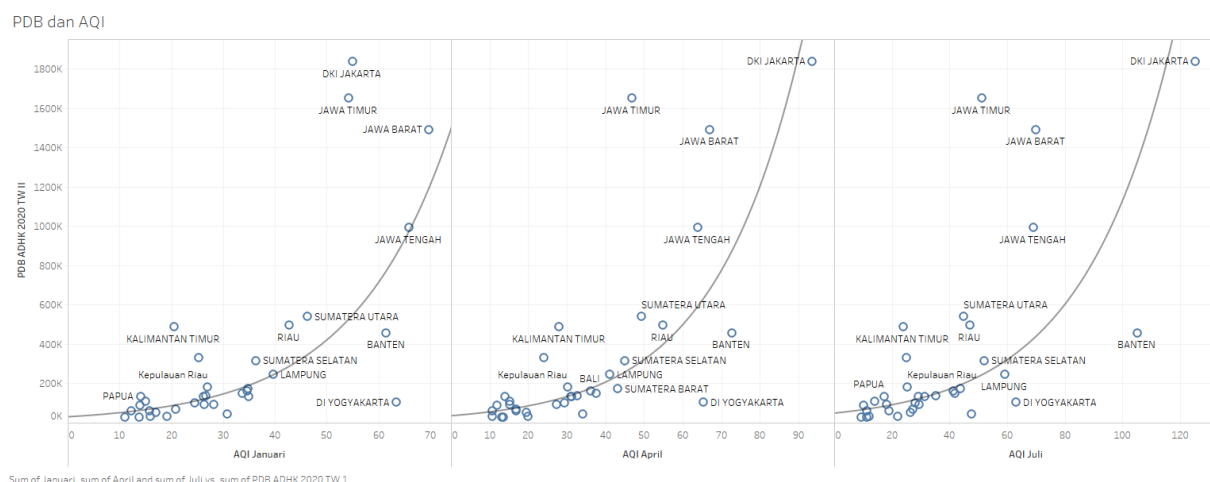

Figure 6: Relationship Between Economic Activity (GDP) with Air Quality Index

If you look at the relationship with pollutant conditions at the beginning of each quarter, the level of pollution and also the regional economic value (GRDP) appears to have an exponentially positive relationship (Figure 6). This is certainly not the desired

economic growth. According to Todaro (2003: 521), environmental degradation can reduce the rate of economic development in the future and the level of productivity of natural resources as well as the emergence of various kinds of health problems and disturbances in the comfort of life. So the utilization of AQI data that is taken regularly can also be a parameter in guiding the economy towards a sustainable economy (green growth).

In terms of developing analysis, this data collection can be used as

- Long-term analysis of rainfall on climate change conditions in supporting agricultural productivity in Indonesia.
- Supporting mobility analysis with the dominance of pollutants from motorized vehicles in Indonesia
- Confirmative means for environmental statistics produced by BPS, so that there is a form of support or even quality control.

4.   Conclusion

From the study results, the use of web scraping on environmental data is a new use that can be developed in order to present data more quickly. In addition, the use of technology in presenting developing visualizations is a good first step in supporting statistical literacy. Widespread literacy is the beginning of a quality society through the data held. However, this research is also a first step for the development of other things. One of them is in overcoming possible environmental problems.

AQI data can be used as a proxy for socio-economic data. From a social perspective, it can be used for monitoring the condition of the population and their mobility, as well as monitoring for population control policies. From an economic perspective, pollutant level data can be used as a proxy for the productivity of the Indonesian industrial sector. AQI data apart from being an indicator of environmental degradation can also be an indicator in monitoring green growth in Indonesia.

References

Bogdan, M. (2015). *Temperature and Humidity Measurement System*.

BPS Statistics Indonesia. (2020a). *Big Data Review of The Impact of COVID-19*.

BPS Statistics Indonesia. (2020b). *Study Big Data as Complete Social Statistic Data and Information*.

Chaves A. (2020). *Scrapy 2.3 documentation. (Python Software Foundation)*. Retrieved from https://docs.scrapy.org/en/latest/

Durcevic S. (2020). *Utilize The Potential Of Digital Dashboards In A Business Environment*. Retrieved from https://www.datapine.com/blog/digital-dashboard-definition-and-examples/#:
:text=Digital%20dashboards%20not%20only%20help,discovery%20of%20priceless%20new%20insights

Hananto, V. R., & Putra, I. (2018). A Dashboard System for Monitoring Air Pollution in Surabaya based on PM2. 5. *Journal of Information Systems Engineering and Business Intelligence*, *4*(2): 139–147.

Mawonike, R., & Mandonga, G. (2018). The effect of temperature and relative humidity on rainfall in Gokwe region, Zimbabwe: A factorial design perspective. *British View*, *3*(2).

National laboratory of the U.S. Department of Energy. (2011). *Indoor Temperature and Humidity Data Collection and Analysis*.

Pramana, S., Paramartha, D. Y., Adhinugroho, Y., & Nurmalasari, M. (2020). Air Pollution Changes of Jakarta, Banten, and West Java, Indonesia During the First Month of COVID-19 Pandemic. *The? Journal of Business, Economics, and Environmental Studies (JBEES)*, *10*(4): 15–19.

Pramana, S., Yuniarto, B., Kurniawan, R., Yordani, R., Lee, J., Amin, I., … Indriani, R. (2017). Big data for government policy: Potential implementations of bigdata for official statistics in Indonesia. *2017 International Workshop on Big Data and Information Security (IWBIS)*, 17–21. IEEE.