**Introduction**

Data preparation is the act of transforming raw data into a format that can be analyzed and used to train machine learning models (Hardin et al., 2015). This entails a number of distinct tasks, such as data collection, formatting, cleaning, and consolidation. While the appellation ‘descriptive analysis’ in Machine Learning is all about gaining a new perspective on the data and its various patterns (Stieglitz et al., 2018). It is a part of four different types of data analysis concepts. It assists individuals in the right comprehension of particular circumstances and their outcomes. As a result, they will be able to make better business decisions (Cao, 2017). Data preparation and descriptive analysis will be performed on the Covid-19 dataset in Malaysia. There are various datasets on various folders, including but isn’t limited to epidemics, and vaccination. Datasets related to epidemics are concerned with the widespread of the pandemic in Malaysia, and includes cases in various states and deaths. These datasets also provides information regarding the number of Covid-19 patients in hospitals. On the other hand, the vaccination datasets provides information related to adverse event following immunization and Covid-19 boosters. This paper aims to carry out descriptive analysis (statistics) with Python on the datasets and present findings which lead to better management of the pandemic Covid19 situation.

**Data Preparation Processes**

Before entering the data into the machine learning model, this is the most important step. The reason for this is that the data set must be unique and specific to the model, thus we must identify the data's required characteristics (Qiu et al., 2016). The data preparation process provides a mechanism for preparing data for project definition as well as project evaluation of machine learning algorithms (Syafrudin et al., 2018). There are a variety of predicting machine learning models available, each with its own method. However, some processes are common to all models, and they allow us to identify the underlying business problem and its solutions (Jia & Ma, 2017).

1. **Data Collection**

This is the first step of data preparation. Operational systems, data warehouses, data lakes, and other data sources are combed for relevant information (Roh et al., 2017). In this case, the data for are drawn from the following github link <https://github.com/MoH-Malaysia/covid19-public>. These data entails a lot of information, including but isn’t limited to Covid-19 pandemic, and vaccination.

1. **Data Discovery and Profiling**

The next step is to look over the collected information to see what it contains and what needs to be done to get it ready for its intended purpose. Data profiling aids with this by identifying patterns, correlations, and other qualities in the data, as well as inconsistencies, abnormalities, missing values, and other errors, which may then be resolved (Jo & Gebru, 2020). In this step, the missing values in all the available datasets have been identified by utilizing the following function dataset\_name.isnull().sum(). The above, mentioned function, identifies all the sum of missing values in every column of a dataset. For instance, the results shows that cluster\_import column of cases\_malaysia dataset has 342 missing values. Other columns with missing values include cluster\_community, cluster\_highRisk, cluster\_education, just to mention a few. Mode imputation has been utilized to replace missing values. For instance, the missing values for cluster\_community are replaced as follows; cases\_malaysia['cluster\_community'] = cases\_malaysia['cluster\_community'].fillna(cases\_malaysia['cluster\_community'].mode()[0]). Some of the possible ways of handling missing values include mean and median imputation (García et al., 2015). In this scenario, mean imputation was used to fill the missing values. The concept of missing values is critical to grasp in order to manage data effectively. Failure to handle missing numbers properly may result in incorrect data conclusions, which will have a severe impact on the modeling phase. It's an important issue in data analysis because it affects the results (Pratama et al., 2016). When you know that numerous elements are missing data, it's tough to have complete faith in the findings. It has the potential to weaken research's statistical power and lead to erroneous results due to skewed estimates.

1. **Data cleaning**

One of the most critical steps that should never be overlooked is data cleaning. The accuracy of your model will be jeopardized if the data is not fully cleansed (Chu et al., 2016). Due to poor data quality, outcomes are skewed, with low accuracy and significant error percentages. As a result, fully cleaning the data before fitting a model to it is critical. As a data scientist, it's critical to recognize that not all of the data we're given is helpful, and we need to know how to deal with it (Krishnan et al., 2016). The given datasets have been cleaned using various methods, including getting rid of null values, dropping unnecessary columns, renaming columns, and handling missing values. Some of the columns dropped include week column in cases\_age dataset using the ‘drop’ function. Del command is another technique used to drop commands in the given datasets.

1. **Data structuring**

This step involves the process of modelling and arranging date in order to suit the analytics requirements (de Oliveira et al., 2021). To make data accessible to BI and analytics tools, data saved in comma-separated values (CSV) files or other file formats must be translated into tables. However, this step hasn’t been considered, since descriptive analysis is done on comma-separated values (CSV) files using python programming language.

1. **Data transformation**

The data must usually be translated into a consistent and useful format in addition to being structured (Kanter & Veeramachaneni, 2015). Data transformation may, for example, entail the creation of new fields or columns that aggregate values from existing ones. Through strategies such as enhancing and adding data, data enrichment improves and optimizes data sets as needed.

1. **Data validation and publishing**

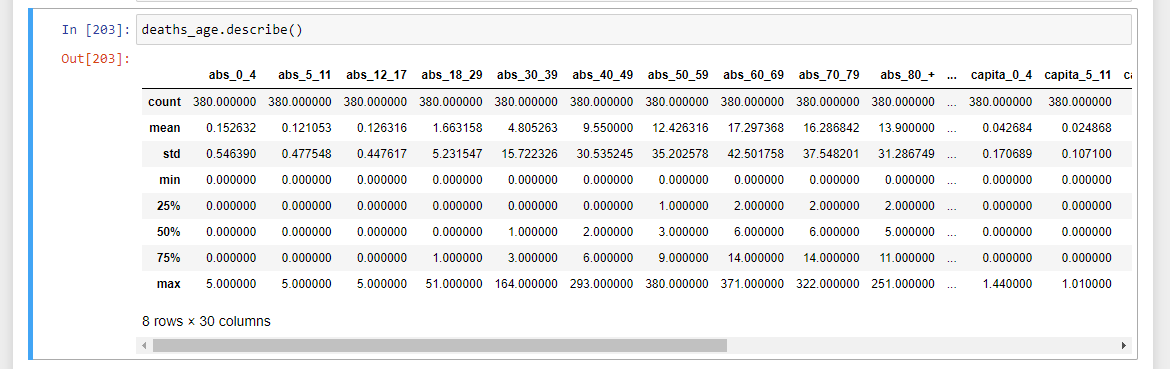
Automated algorithms are run on the data in this final step to ensure its consistency, completeness, and accuracy (George et al., 2016). The data is subsequently stored in a data warehouse, a data lake, or similar repository, where it can be accessed directly by the person who generated it or made available to other users. In this case, the notebooks and other data are stored in Github repository, which is a platform for sharing and collaborating on coding projects. It is, however, a potentially valuable resource for researchers who wish to make their data publicly available. You can use it to do the following things: data storage and monitoring alterations (Olson et al., 2016).

**Descriptive Analysis**

Descriptive analytics is a branch of statistics concerned with gathering and summarizing raw data such that it may be easily analyzed (Kemp et al., 2018). In general, descriptive analytics focuses on historical data, providing the context needed to comprehend data and numbers. The field is utilized in a variety of businesses and for a variety of objectives, ranging from inventory management to yearly revenue and sales benchmarking. The field is typically used as a first phase in the business intelligence process, laying the groundwork for subsequent analysis and comprehension (Bhatnagar et al., 2021). In this section, we shall have a look on descriptive analysis of some of the datasets available from the github link provided:

**Deaths Age Dataset**

This dataset shows the deaths of Covid-19 patients based on their ages, ranging from zero to above eighty years. The following figure 1.0 shows the descriptive analysis performed on the dataset under consideration using ‘describe’ function.

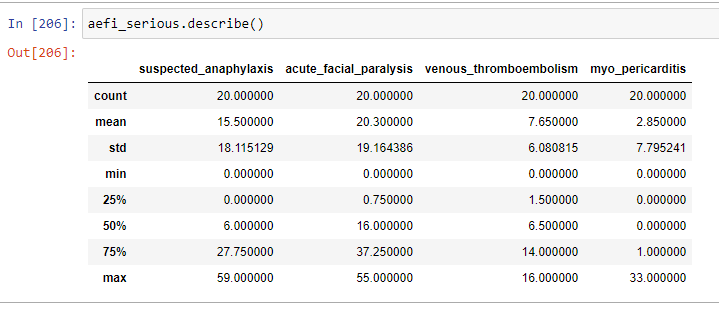


*Figure 1.0 – Descriptive analysis of deaths age dataset*

In the above figure, we can see the mean, and standard deviation of various Covid -19 patients who die in several states, such as Malaysia and Johor based upon their age groups; for instance, 0-4 years have mean of 0.15. Comparing the mean values, the people aged between 60-69 years have the highest mean value of 17.30, which indicates that approximately 17 people aged between 60-69 years die in a week. We can also say that the maximum number of deaths per week is associated with the individuals aged between 50 to 59 years. Considering the standard deviation, individuals aged 17.30 have the highest standard deviation, implying that the values in the deaths age dataset are far away from the mean. On the contrary, children aged between 12 to 17 years have the lowest standard deviation, indicating that values in deaths age dataset are close to the mean of the dataset.

**AEFI Serious Dataset**

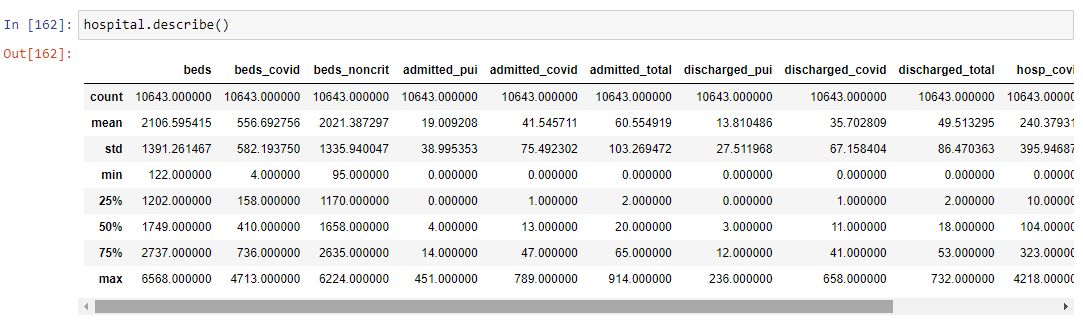
AEFI stands for adverse events following immunization. Various Covid-19 vaccines, including but isn’t limited to astrazeneca and sinovac, impacts negatively on different individuals. The figure below shows a descriptive analysis of the effects of several types of vaccines on individuals.



In the figure above, we can obtain several information regarding the dataset including mean, standard deviation and maximum value of the dataset. The acute\_facial\_paralysis has the highest mean of 20.30, while the myo\_pericarditis has the lowest mean of 2.85. This implies that the acute facial analysis is higher than myocardial inflammation. The above information can be utilized by medical practitioners to come up with effective ways of lessening the complete weakness of face muscles. The analysis can further give information regarding a particular vaccine that have serious effects to the individuals. From the above figure, we see that the lion’s share of individuals have suspected anaphylaxis on 5th September, 2021. The acute facial paralysis has the greatest standard deviation, implying that the values in the aefi serious dataset are far away from the mean, while venous thromboembolism has the least standard deviation, indicating that the values in the aefi serious dataset are close to the mean.

**Hospital Dataset**

This dataset entails information about the number of beds available in a given state, number of beds occupied by Covid-19 patients and people who are being discharged at a given date. The following figure shows the descriptive analysis performed on the dataset under consideration.



In the above figure, we can see that the beds has a mean number of 2106, meaning that a certain state has approximately 2106 beds on a certain date. We can also see the mean of the number of beds occupied by Covid-19 patients, and discharged Covid-19 patients. This information is enable the state to determine the need of increasing the number of beds at the given point in time.

**Summary**

This section provides the summary of what has been discussed in this paper, and the summary of results obtained from the descriptive analysis of the given dataset. The given dataset has information regarding the Covid-19 pandemic and vaccine. Regarding the case age dataset, the mean of aged individuals between 0-4 years is 444.09. The highest mean is associated with those individuals whose age ranged from 18 to 29 years. While those aged above 80 years have the lowest mean of 72.27. This implies that at a particular weeks, there is high chances of people ranging from 18 to 29 years to test Covid-19 positive. Regarding the deaths age dataset, the results shows that people aged between 60 to 69 years have the highest mean. While people aged between 5 to 11 years have the lowest mean of 0.21. Therefore, the results shows that young people can contract the Covid-19 easily as compared to the adults, and many adults die of covid-19 as compared to the young people. The results also shows that various types of Covid-19 vaccines have mild to adverse effects towards people. Acute facial paralysis is leading by the mean of 20.30, meaning that this is a common effect to individuals who have been vaccinated with different types of vaccine.

**Recommendation**

The individuals should consider other types of Covid-19 vaccine, in a scenario where one vaccine type has serious effects on an individual’s health. Besides, the medical researchers should undertake a study to determine why less number of young people die of Covid-19, while many elderly people die of the Covid-19 pandemic.

**Conclusion**

In conclusion, this paper aimed to carry out descriptive analysis (statistics) with Python on the datasets and present findings which lead to better management of the pandemic Covid19 situation. The data preparation has been conducted by following a series of steps, ranging from data collection to data validation and publishing. The data is drawn from the Covid-19 datasets in Malaysia, which have information pertaining the Covid-19 pandemic and vaccination. During the data preparation, some of the things done include removal of missing values, and at some point replacing the missing values using mode computation technique. The ‘describe’ function has played significant role in performing descriptive analysis on the available datasets. This function provides standard deviation, mean, minimum and maximum values of the various dataset features. Descriptive analysis is used for better management of the Covid-19 pandemic; for instance, information about the number of beds in a given state can help it to make a decision whether there is need to increase the number of beds, based on the Covid-19 patients occupying the available beds.

References

Bhatnagar, V., Poonia, R. C., Nagar, P., Kumar, S., Singh, V., Raja, L., & Dass, P. (2021).

Descriptive analysis of COVID-19 patients in the context of India. *Journal of Interdisciplinary Mathematics*, *24*(3), 489-504.

Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys*

*(CSUR)*, *50*(3), 1-42.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and

emerging challenges. In *Proceedings of the 2016 international conference on management of data* (pp. 2201-2206).

de Oliveira, J. M., da Costa, C. A., & Antunes, R. S. (2021). Data structuring of electronic health

records: a systematic review. *Health and Technology*, *11*(6), 1219-1235.

García, S., Luengo, J., & Herrera, F. (2015). Dealing with missing values. In *Data preprocessing*

*in data mining* (pp. 59-105). Springer, Cham.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods

for management research. *Academy of Management Journal*, *59*(5), 1493-1507.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., ... & Ward, M. D.

(2015). Data science in statistics curricula: Preparing students to “think with data”. *The American Statistician*, *69*(4), 343-353.

Jia, Y., & Ma, J. (2017). What can machine learning do for seismic data processing? An

interpolation application. *Geophysics*, *82*(3), V163-V177.

Jo, E. S., & Gebru, T. (2020, January). Lessons from archives: Strategies for collecting

sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 306-316).

Kanter, J. M., & Veeramachaneni, K. (2015, October). Deep feature synthesis: Towards

automating data science endeavors. In *2015 IEEE international conference on data science and advanced analytics (DSAA)* (pp. 1-10). IEEE.

Kemp, S. E., Ng, M., Hollowood, T., & Hort, J. (2018). Introduction to descriptive

analysis. *Descriptive analysis in sensory evaluation*, *1*.

Krishnan, S., Wang, J., Wu, E., Franklin, M. J., & Goldberg, K. (2016). Activeclean: Interactive

data cleaning for statistical modeling. *Proceedings of the VLDB Endowment*, *9*(12), 948-959.

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016, July). Evaluation of a tree-

based pipeline optimization tool for automating data science. In *Proceedings of the genetic and evolutionary computation conference 2016* (pp. 485-492).

Pratama, I., Permanasari, A. E., Ardiyanto, I., & Indrayani, R. (2016, October). A review of

missing values handling methods on time-series data. In *2016 International Conference on Information Technology Systems and Innovation (ICITSI)* (pp. 1-6). IEEE.

Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: a

big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1328-1347.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–

Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, *39*, 156-168.

Syafrudin, M., Alfian, G., Fitriyani, N. L., & Rhee, J. (2018). Performance analysis of IoT-based

sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing. *Sensors*, *18*(9), 2946.

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data

processing. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 1-16.