A Comparative Study of Data Visualization between Python Libraries and Business Intelligence Tools

# Authors

Nitis Monburinon, mo.nitisst@tni.ac.th, Sasithron Sakuntaviphard, sa.sasithronst@tni.ac.th, Yaninee Saekok, sa.yanineest@tni.ac.th, Phumipat Saengpradab, sa.phumipatst@tni.ac.th, Prajak Chertchom, prajak@tni.ac.th, Nattagit Jiteurtragool, nattagit@tni.ac.th,

Thai-Nichi Institute of Technology, Bangkok, Thailand

# Abstract

This research is a comparative study between Python's data visualization libraries and business intelligence tools, including Microsoft's Power BI and Tableau Public. In this study, we conduct a series of experiments to evaluate the data visualization capability of each chosen tool by performing related tasks on the targeted dataset. Each iteration of the experiment involves exploratory data analysis, feature engineering, and visualizing the dataset in useful formats.

The objective is to provide reliable information to help researchers, analysts, and business executives decide which tool is more suitable for their purposes.

# Keywords (5 maximum)

Python Libraries, Business Intelligence Tools, Data Visualization, Sustainable Development Goals (SDGs)

Abstract Submission <http://icbir.tni.ac.th/submission/>

The abstract consists of up to 200 words and 5 keywords. The abstract can be submitted under ICBIR online submission system as the instruction below.

Submission: <http://icbir.tni.ac.th/openconf/author/submit.php>

# 1. Intro

Over the years, the growth of internet usage and information technology have resulted in a massive amount of data available on the internet. In particular, the rapid development of information technology, mobile technology, and web technology have contributed to the considerable increases in the amount of data. <r> Experts estimate that in 2020, there will be 40 zettabytes of data in the digital universe with machine-generated data as a critical driver of the growth. <r> Another study revealed that only 0.5% of global data was analyzed in 2012 <r>.

A considerable amount of data is generated every second from the web and mobile applications, services, smart products, the Internet of Things, as well as users. These data are often collected in raw form and required pre-processing to be more useful. Collected data are immensely valuable for creating insightful products as they can be analyzed for useful business insights. For some instances, data are used as training resources for machine learning applications. In most business cases, though, raw data are engineered, analyzed, processed, and visualized to create reports or dashboards for decision-makers and business executives.

The practice known as data visualization, which involves representing data in understandable formats such as graphs and diagrams, is a crucial step in every data analysis process. <r> Data visualization helps analysts convey the most important messages of the data and explain the insight hidden in the raw dataset.

In the modern days, analysts and researchers make use of specialized programs to create meaningful and insightful visualizations. These specialized programs are often called data visualization or business intelligence tools. Data visualization tools are designed and developed to make required data visualization tasks such as importing data, data preparation, exploratory data analysis, and visualizations more streamlined and convenient.

In this research, we conduct a comparative study between Python programing language with its data analysis libraries and frameworks, Microsoft's PowerBI, and Tableau Public. Python is an open-source general-purpose programing language with significant data science and analytic ecosystem, while PowerBI and Tableau are business intelligence tools commonly used among professionals.

Each of the tools, as mentioned earlier, has advantages over the others and disadvantages compared to the others. In this paper, we conduct a series of experiments by using Python, PowerBI, and Tableau Public to perform exploratory data analysis, feature engineering, and data visualization.

This study considers performance, ease of usage, pricing, and various other features. Ultimately, our objective is to provide an up-to-date and reliable source of information that can help researchers, analysts, and business executives to decide which tool is more suitable for their purposes.

# 2. Literature review

TBA

# 3. Research methodology

The comparative study of data visualization tools is carried out using Python, Microsoft's PowerBI, and Tableau to perform data visualization tasks on sample datasets in order to determine their performance and providing a complete analysis of their functionality.

While a specific dataset is not particularly important in this type of comparative study, it is advisable to evaluate each data visualization tool against the same dataset. For this research, we use *Olympic Sports and Medals (1896-2014) - Summer* dataset. The dataset was provided by the *IOC Research and Reference Service*, published by *The Guardian's Datablog*, and also available on <https://www.kaggle.com>. The *summer.csv* file contains 31,165 records. All of which are about the Olympians who have been awarded medals at the Olympics summer game since 1896 as well as along with other related information with nine total columns. The features of this dataset are described in Table.

|  |  |
| --- | --- |
| columns | description |
| Year | The year in which the game took place |
| City | The city where the game took place |
| Sport | The type of sport competition |
| Discipline | The discipline that was participated |
| Athlete | The name of the particular olympian |
| Country | The olympian's country |
| Gender | The olympian's gender |
| Event | The particular event the olympian participated in |
| Medal | The type of medal that the olympian won |

## 3.1 Python

Python is a general-purpose programming language known for its ease of usage, flexibility, and readability. Python is considered a high-level language and has many implementations, most notably in C (CPython). It is interpreted rather than compiled make it not the most suitable for high-performance applications. However, when it comes to data science, data analytics, and visualization, Python is one of the most utilized and seen widespread usage across the industry. This is primarily due to its vast ecosystem consisting of many libraries and frameworks explicitly built for data science. For visualization, there are libraries such as Matplotlib and seaborn, amongst others, that are heavily utilized by professionals, academic, as well as data enthusiastic around the world.

* Python was released by Guido van Rossum in 1991.
* The current release version of Python is 3.8.2.
* Python puts a significant emphasis on code readability by using white spaces and tabs for organizing purposes.
* Python supports many paradigms, such as procedural, functional, object-oriented, and more, which makes it versatile.
* Python has many implementations.

The default and the most popular is CPython, which is written in C and Python itself.

* Python is dynamically typed, which means data types are dynamically assigned to variables at creation based on the values.

Data types can be modified on the fly by assigning new values.

* Python is a programming language that supports garbage collection.
* Python has an open-source license and is supported by Python Software Foundation.
* Python license indicates that individuals and organizations can use it free of charges.

In this research, we utilize Python's data analytic and visualization ecosystem to import, explore, prepare, and visualize data. Jupyter notebook, which is a web-based interactive computational environment that can provide feedback and support graph displays, is a particularly useful application that we use extensively for visualizing data using Python <r>.

### 3.1.1 Importing dataset

With Python, the dataset can be imported using *pandas* which is a Python library for data manipulation and analysis. For this particular dataset, we will import it from a commas-separate-values file. To do so, we must first import pandas library, often aliased as *pd* and read data from a file named 'summer.csv' using the following source code.

import pandas as pd df = pd.readcsv('summer.csv')

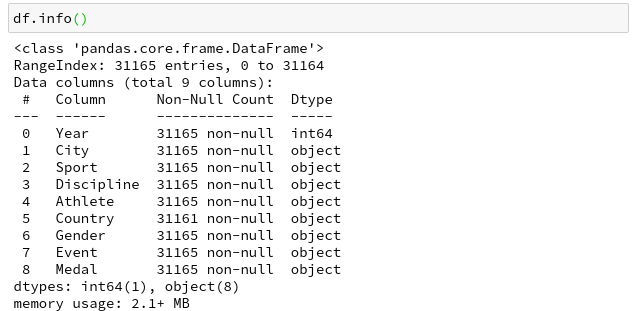
Data from the source will be stored in *df* variable, which is an object belong to *pd.DataFrame* class.

### 3.1.2 Data preparation

In this study, methods and properties belong to *pandas.DataFrame* class is used to perform mandatory exploratory data analysis. Some of which are described in Table.

|  |  |
| --- | --- |
| command | description |
| df.head() | Show the first *n* rows of the DataFrame |
| df.shape | Show the dimensionality of the DataFrame |
| df.columns | Show the column labels |
| df.isnull().sum() | Count rows that contain missing data |
| df.describe() | Generate descriptive statistics |
| df.info() | Print a concise summary of a DataFrame |
| df['Medal'].valuecounts() | Count number of samples from a specific field |

Data preparation and feature engineering, in general, is handled using libraries such as pandas, NumPy, and Python standard libraries.



### 3.1.3 Visualization

Visualizations can be created using various libraries and frameworks. Among Python visualization libraries, Matplotlib and Seaborn are particularly popular.

### 3.1.4 Performance

Python offered excellent performance compared to other data visualization tools we used in this comparative study, especially when handling a large dataset. For example, using PySpark library on Apache Spark makes manipulating a dataset containing millions of records relatively simple.

Python, with its ecosystem, offers superior flexibility over PowerBI and Tableau. For most data visualization tasks, the only limits to Python's capabilities are the abilities of developers.

### 3.1.5 Ease of usage

Python offers excellent performance, a wide range of data analysis and visualization libraries, as well as the flexibility to adapt to tasks with complex structures. However, being a full-fledged programing language, it requires analysts who are knowledgeable in programming and have technical expertise. People who can use Python comfortably for data visualization and can utilize its ability to a full extent are often developers and engineers.

Python is often viewed as a more straightforward language to learn compared to other programming languages. Nevertheless, it has more learning curves than graphical user interface tools. As such, training is necessary in case, decision-makers decide to employ Python as a data visualization in the work environment.

### 3.1.6 Pricing

Python, as well as most of its ecosystem, are free for commercial use. The cost of using Python as a data visualization tool lies in the salary of Python programmers. As of 2020, the average Python developer annual salary in the US is $120,365. This can be perceived as high compared to the salary of other analysts.

## 3.2 Microsoft's PowerBI

TBA

## 3.3 Tableau

TBA

# 4. Research result

TBA

# 5. Conclusion

TBA