 ISSN:2302-8149
e-ISSN:2540-9719

Exploratory Data Analysis (EDA): A Study of Olympic Medallist

Analisis Data Eksplorasi (EDA): Studi Peraih Medali Olimpiade

[1]Noviyanti T M Sagala*, [2] Fonggi Yudi Aryatama
[1]Statistics Department, School of Computer Science, Bina Nusantara
[2]Mathematics Department, School of Computer Science, Bina Nusantara
[1,2]Jl. Raya Kebon Jeruk No.27, Jakarta Barat, Indonesia
*email: noviyanti.sagala@binus.edu

(***received***: 9 Januari 2022, ***revised***: 18 Juli 2022, ***accepted***: 26 Agustus 2022)

Abstract

Olympic games are one of the most popular international sports events in the world where thousands of athletes participate in different types of sports categories. The winner has rewarded a medal (Bronze, Silver, Gold) according to the rank. An analysis can be carried out on the Olympic data to understand the changes in medalists over time. Furthermore, it helps to determine the progress of participating countries and strategies that can be used in the future. Exploratory Data Analysis (EDA) is a method for analyzing and summarizing the properties of data, either in graphical or non-graphical, to get insights from the dataset being studied. The approaches can be classified as univariate, bivariate, or multivariate. EDA is widely used in various domains including sport. The main purpose of this study is to analyze the changes of Olympic Medalist data throughout the provided years in the form of univariate, bivariate, and multivariate analysis. This analysis provides detailed, statistical, and interesting information about the changes in medal winners from time to time.

Keywords: *Exploratory Data Analysis, Graphical, Medallist, Olympic, Python*.

Abstrak

Pertandingan Olimpiade adalah salah satu acara olahraga internasional paling populer di dunia di mana ribuan atlet berpartisipasi dalam berbagai jenis kategori olahraga. Pemenang mendapatkan hadiah berupa medali (Perunggu, Perak, Emas) sesuai dengan peringkatnya. Analisis dapat dilakukan pada data Olimpiade untuk memahami perubahan peraih medali dari waktu ke waktu. Selain itu, membantu untuk menentukan kemajuan negara peserta dan strategi yang dapat digunakan di masa depan. Explorasi dan Analisis Data (EDA) adalah suatu metode untuk menganalisis dan meringkas karakteristik data, baik dalam bentuk grafis maupun non-grafis untuk mendapatkan wawasan dari kumpulan data yang sedang dipelajari. Pendekatan dapat diklasifikasikan sebagai univariat, bivariat, atau multivariat. EDA banyak digunakan di berbagai domain termasuk olahraga. Tujuan utama dari penelitian ini adalah untuk menganalisis perubahan data peraih medali Olimpiade sepanjang tahun yang diberikan dalam bentuk analisis univariat, bivariat, dan multivariat. Analisis ini memberikan informasi rinci, statistik, dan menarik tentang perubahan peraih medali dari waktu ke waktu.

Kata Kunci: *Exploratory Data Analysis*, Grafik, Peraih Medal, Olympic, Python

1 Introduction

Data are considered as oil in the 21st Century; an immensely, untapped asset. In these modern times, data are changing the face of our world. It might be part of a study helping to cure a disease, boost a company's revenue, make a building more efficient targeted ad, etc. In general, data are simply another word for information, but that information may not be discovered easily. Data have no any qualities whether important, surprising, or funny without some kind of hypothesis or theory and for that reason there is a need for hypothesis or theory to be made to make good use of the data [1]. Data are raw facts and figures with no proper information hence need to be processed to get the desired information. Data Analysis is the analysis of various data likes cleaning the data, transforming it into understandable form,

 ISSN:2302-8149
e-ISSN:2540-9719

and then modeling data to extract some useful information for business use or an organizational use [2]. The analyzed data are more insightful for identifying and improving extremely critical insights across the Organization therefore make an Data Analysis an important factor in business development [3].

Exploratory Data Analysis is an approach that has been used to analyse and summarize the main characteristics of data by using statistical graphic and other data visualization, to increase the understanding of dataset that being learned. EDA helps the analyst to be alert to unexpected patterns, relationships, and extraordinary cases [4]. EDA has been developing and is widely used in various fields and uses, with the goal to solve various problems that occur whether in the respective field or on some others. EDA helps to analyze the data sets to summarize their statistical characteristics focusing on four key aspects, like, measures of central tendency (comprising of the mean, the mode and the median), measures of spread (comprising of standard deviation and variance), the shape of the distribution and the existence of outliers [5]. There are different kind of techniques of EDA and most of them are graphical in nature with some quantitative techniques with their different objective [6]. EDA is repetitive analysis to find any information inside the data. The analysis can be done in any means without any terms that must be obeyed [7]. EDA is inherently subjective, subject matter knowledge and experience are needed to get the expected result. This includes understanding the data pedigree [4].

Olympic games are leading international sporting events which is divided into two parts, Summer and Winter Olympic with their respective sports, The Olympic games are considered the world's foremost sports competition with more than 200 nations participating, 33 Sports for Summer Olympic and 7 Sports in Winter Olympic with each different disciplines and events in the last Olympic. The evolution of Olympic in 20th and 21st centuries has resulted in changes to the Olympic Games, including increasing the number of participants, new rules, new events and much more.

The main objective of this study is to analyse and discover many interesting facts and insight from the Olympic Medallist over the years, and to investigate relationships between variables in data. In this paper the analysis will be conducted by three different categories; Univariate, Bivariate and Multivariate analysis to provide clearer insight and limit the scope of analysis being conducted. The analysis includes the visualization and the explanation of the changes that will be discovered over the years which will be used to assist in the maintenance of future Olympics, as the Olympic Games are one of the most important sporting event across the world, each country and athletes will try to give their best performances and it is hoped that with the analysis, they can improve their performance in their participation in the next Olympic games.

2 Literature Review

The characteristics of Crowdfunding Network on Twitter is analysed by identifying who the users in the crowdfunding network are, what they share, and how they are connected to each other, based on the dataset of 2,7 million tweets from January 2014 to December 2014, resulting communities with the biggest topologies in crowdfunding, the influencers, activity and visibility of the twitter and so on [8].

Regarding COVID-19 pandemic, a study conducted Exploratory Data Analysis to generate a correlation between variables regarding COVID-19 the variables are Recovered, Deaths, Total Hospitalized Patient in specific region, Total Positive cases and etc., to give insight if there are correlations between those variables [9]. Another study focusing on COVID-19 pandemic is conducted in India. The goal is to describe and study the COVID-19 spread in India from the day of the outbreak, the reason why prevention that has been done by National and local authorities is hard, comparison of the spreading to other countries, the common symptoms and pattern of the spreading which is expected to help predict the spread of the virus in the future [10].

The exploratory data analysis is performed on Groundwater Quality data. It is studied because in developing countries such as India and China groundwater plays critical job, particularly, in parched and semi-dry locales, where lacking surface water, a large portion of the district populace relies upon groundwater for everyday needs, especially drinking purposes. The analysis includes checking some attributes like mineral contents, Total Dissolved Solids (TDS), acidity, longitude and latitude of the

ISSN:2302-8149
e-ISSN:2540-9719

district by using heat map, pair plot, and box plot to compare between each attributes and outlier analysis and can be used for future work in Groundwater Quality [11].

It is useful to apply EDA on Customer Segmentation for Smartphones data. The main goal is to identify the correct category or class to which a new data will fall under later. In this paper the data used are specification of the smartphone to be used to find the pattern and correlations between attributes and "price ranged" being used as a target variable for the segmentation, with different methods of segmentation this paper compare the result of segmentation by using Random Forest Model and KNN Classification Model [12]. In terms of realm of crime, EDA gives a better insight with various visualization regarding the crimes that being commit, the distribution by region, dates and if use correctly can be used to prevent and reduces the crime rates [13]. Furthermore, EDA helps in solving Cyber Security problem by understanding the correlation between attributes and differentiate the attack type and their respective characteristic. The result might be used to give an estimation of the attack priority and to propose a security policies or countermeasures for cybersecurity attacks to come [14].

The analysis of the evolution of Olympic Games by year has been conducted. The difference with our study is the visualization tools. Our work presents the visualization in the form of Histogram, Bar graph, Box Plot, and Scatter plot to give different insightful and interesting information, such as the number of participating countries in Olympics, basis age of the athletes that has been participating in the Olympic, and the height spread of athletes differentiate by gender [15].

3 Research Method

The approach is very important when doing things so that the process can be more focused and easier to understand and give the best result as intended. In addition, using a clear approach step by step will make the progress clearer, concise, and reduce the possibility for misinterpretation. This study aims are to analyse the data of all the Medallist from The Olympic Games and discover interesting facts and insight from the Olympic Medallist over the years and to understand the relations between variables. The analysis will be based on a flow chart that can be seen in "Figure 1" to achieve the goals of this paper, the steps listed will be explained one by one in more detail in this paper.

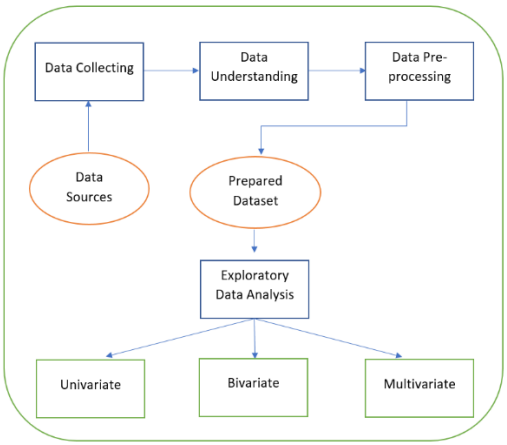

Figure 1. Steps of Analysis

3.1 Data Collection

The data that being used are data regarding the Olympic championship from 1896 – 2014 obtained from public dataset.

3.2 Data Understanding

There are 3 data folders containing different Olympic data.

a. The first data are named *dictionary.csv* which contains the country name data (Country), country code (Code), total population (Population) of that country, and GDP per capita of the country. [201 rows]

 ISSN:2302-8149
e-ISSN:2540-9719

b. The second data are named *summer.csv* which contains data on countries that won medals in Summer Olympics for the period 1896 – 2014, in the *summer.csv* file containing the year data the implementation of the Olympics (Year), the city of the implementation of the Olympics (City), the sport contested (Sport), the type of sport that is competed in each branch sport(Discipline), the name of the athlete who received the medal (Athlete), the country of origin of the athlete in the form of country code (Country), athlete's gender (Gender), type of sport from each type of sport being contested (Event), and the type of medal obtained by athlete(Medal). [31165 rows]
c. The last data are named *winter.csv* which contains data on countries that won medals in the winter Olympics for the period 1924-2014, the *winter.csv* file contains the same data types as the *summer.csv* file. [5770 rows]

3.3 Data Pre-processing

Data pre-processing refers to manipulation or dropping of data before it is used to ensure or enhance the performance of the data usage. Data pre-processing includes cleaning, instance selection, normalization, transformation, feature extraction, etc. For the research purposes, data cleansing and data transformation are performed on the data set. The data cleansing is used to delete missing values to get understandable visualization while data transformation is used to transform datasets into a form that may provide some specific and useful information includes top ten countries based on the number of the medallist, sport with the most medals, progress in the number of medallists per year, etc.

3.4 Exploratory Data Analysis (EDA)

Exploratory data analysis is an approach of analysing data sets to summarize their main characteristics, generally using statistical graphics and other data visualization methods[4]. The objectives of EDA are to suggest hypotheses about the causes of observed phenomena, assess assumptions on which statistical inference will be based, support the selection of appropriate statistical tools and techniques, provide a basis for further data collection through surveys or experiments, and in this paper will focus on helping to look at the data before making any assumptions and to discover interesting insight and relations from the data. Exploratory data analysis is divided into 3 categories:

a. Univariate analysis is the simplest form of analysing data. “Uni” which means “one”, means the data has only one variable. It does not deal with causes or relationships and its major purpose is to describe. Summarizes the data and finds patterns in the data.
b. Bivariate analysis, it involves the analysis of two variables, for the purpose of determining the empirical relationship between them.
c. Multivariate analysis is a set of techniques used for analysis of data sets that contain more than one variable, and the techniques are especially valuable when working with correlated variables.

The tools that are utilized to discover specific information and to provide a better, simple, and insightful visualization are Bar Chart, Pie Chart, Line Plot, and Pair Plot.

4 Results and Analysis

4.1 Experimental Setup

There are many programming languages provided to help specific data role such as data analyst to access, organize, and perform data analysis and the necessary calculations. Some of the most popular languages are Python R, Java, JavaScript, Scala, and SQL.

For this study, Python is used for Exploratory Data Analyst and Visual Studio Code as the platform. Python is a multipurpose programming language that is compact and easy to read. The benefit of using Python for data analysis is that the language can be used for many other projects and purposes and can be flexible in other area. Python was first release in 1991 growing rapidly in various aspects with time and now has become one of the most popular programming languages that being used, due to various advantages compared to other programming languages, namely, easy to learn, free and open source, flexible and has a very broad library for various purposes, one of which is in the field of data analysis.

 ISSN:2302-8149
e-ISSN:2540-9719

There are some of python libraries that commonly used for data analysis and in this paper the libraries that used to produce the expected analysis are:

a. **NumPy -** NumPy is a Python library focused on scientific computing. NumPy provides a convenient and efficient way to handle the vast amount of data, also very convenient with Matrix multiplication and data reshaping also fast which makes it reasonable to work with a large set of data.
b. **Pandas** - pandas is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series,
c. **Plotly -** is an interactive, open-source plotting python library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases. Mainly used for makes interactive, publication-quality graphs.
d. **Seaborn** - Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas' data structures in Python.
e. **Matplotlib** - Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
f. **SciPy -** SciPy is a free and open-source Python library used for scientific computing and technical computing.

4.2 Result Analysis

The data had been analyzed and had given some insights and stories, various visualization with graphs and plots which represent the changes of Medalists and Olympic Games over the years. Some findings are shown in figure 2 - figure 4.

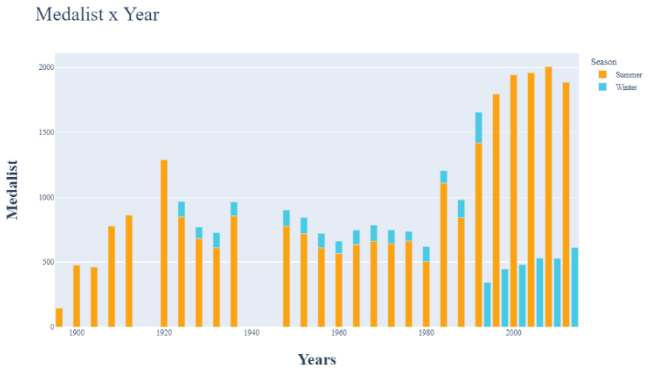

Figure 2. Total Number of Medalist In Olympic Games Over Time Using Bar Chart

Figure 2 represents the amount of medalist from each Olympic Games each year, yellow bar represents Summer and blue bar represent Winter medalist. From the chart there are big gaps between bars, indicates that the Olympic were not held for 4 years each time. The first gap is the cancelation of the 1916 Olympic because of World War 1, and the second gap was cancelation of 1940 and 1944 Olympics cause of World War 2, and in 1994 the Summer and Winter Olympic first being held separately with 2 years gap after each other.

 ISSN:2302-8149
e-ISSN:2540-9719

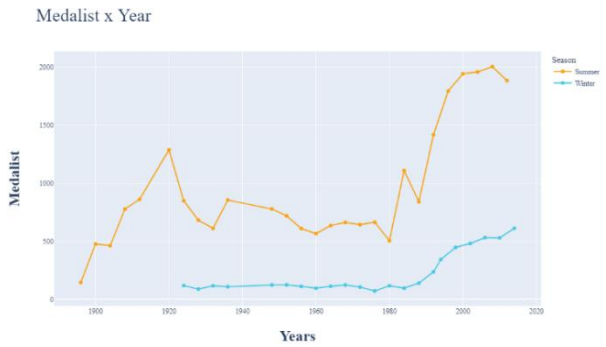

Figure 3. Total Number of Medalist In Olympic Games Over Time Using Line Plot.

Figure 3 represents the amount of medalist from each Olympic. Summer represents by yellow line and Winter by blue line. The number of winter medallist starts from 1924, indicating that Winter Olympics only started in 1924 which was held in Charmonix, France in 1924 Olympic.

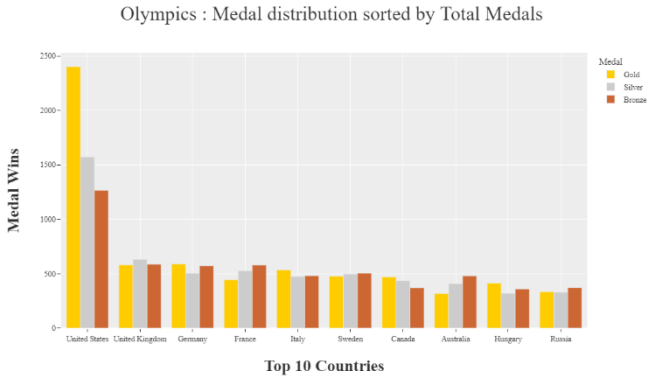

Figure 4. Top 10 Countries with the Most Medalist Achievement.

Figure 4 displays top 10 countries with the most medals from 1896-2014 according to the type of medals. The top is United States; 2402 Gold, 1571 Silver, 1265 Bronze with total of 5238 medals, followed by United Kingdom in second, and Germany in third position.

Univariate Analysis

Univariate analysis is an analysis performed on a variable with the aim of knowing and identifying the characteristics of that variable. Gender variable is chosen to identify the composition of the medalists in the Olympic games. The result is shown in Figure 5.

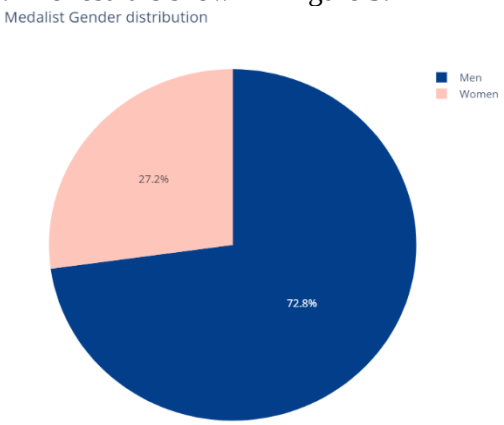

Figure 5. Composition of Medalist in Olympic Games by Gender using Pie chart

Figure 5 represents the gender distribution of medallist. It is interesting to note that Men athletes have more historical accomplishment compared to women athletes, with 22264 and 8304, respectively.

 ISSN:2302-8149
e-ISSN:2540-9719

Bivariate Analysis

Bivariate analysis is an analysis conducted to determine the relationship between 2 variables. For the work, the relationship between Medal (Categorical) and Year (Categorical) variables are identified. The relationship between these variables will be analysed using Chi-Square Test of Independence. Then, analysis is separated in two categories, namely Summer and Winter Games. The hypotheses used as follows:

a. H_o: (null hypothesis), two variables are independent.
b. H_1: (alternative hypothesis), two variables are not independent.

Summer Olympic had been held 27 times and each year 3 type of medals had been given to the accomplished athletes: bronze, silver, and gold medals. Chi-square test is performed to find the relations among them. From the test, p-value is less than 0.05; 2.46e-13, means that the year and medal variables are not independent. The result is displayed in Figure 6.

```
Chi-Square Test Statistic : 162.7200843467204
p-value :  2.463830110369268e-13
Degrees of freedom:  52
Null Hypothesis is rejected.
```

Figure 6. Result of Chi-Square Test of Independence for Summer Olympic

Medal	Bronze	Gold	Silver
Year			
1896.0	38	62	45
1900.0	133	160	183
1904.0	123	185	154
1908.0	204	295	278
1912.0	272	296	293
1920.0	345	497	446
1924.0	257	298	294
1928.0	230	226	226
1932.0	195	212	204
1936.0	282	294	279
1948.0	265	262	251
1952.0	268	248	203
1956.0	209	211	190
1960.0	208	177	181
1964.0	220	238	177
1968.0	228	221	213
1972.0	227	187	229
1976.0	261	188	215
1980.0	205	118	181
1984.0	378	365	366
1988.0	312	239	290
1992.0	513	454	451
1996.0	606	601	588
2000.0	659	624	660
2004.0	668	636	655
2008.0	680	664	661
2012.0	651	630	604

	Bronze	Silver	Gold
1896.0	48.650649	48.374641	47.974711
1900.0	159.708337	158.802269	157.489395
1904.0	155.011033	154.131614	152.857354
1908.0	260.700373	259.221350	257.078277
1912.0	288.884197	287.245280	284.870523
1920.0	432.151970	429.700256	426.147774
1924.0	284.857936	283.241862	280.900202
1928.0	228.825810	227.527620	225.646570
1932.0	205.003768	203.840727	202.155505
1936.0	286.871067	285.243571	282.885362
1948.0	261.035895	259.554969	257.409137
1952.0	241.240113	239.871494	237.888393
1956.0	204.668246	203.507109	201.824645
1960.0	189.905291	188.827908	187.266801
1964.0	213.056289	211.847564	210.096146
1968.0	222.115376	220.855256	219.029368
1972.0	215.740463	214.516510	212.743027
1976.0	222.786419	221.522492	219.691088
1980.0	169.102945	168.143579	166.753477
1984.0	372.093582	369.982597	366.923821
1988.0	282.173763	280.572916	278.253321
1992.0	475.769793	473.070624	469.159584
1996.0	602.261479	598.844690	593.893831
2000.0	651.918693	648.220185	642.861122
2004.0	657.287041	653.558076	648.154883
2008.0	672.721040	668.904514	663.374446
2012.0	632.458434	628.870329	623.671238

Figure 7. The difference values between real data and chi square test

Winter Olympic has been held 22 times with 3 types of medals given to the accomplished athletes. Chi-square test is performed to find the relations among them. From the test, p-value is less than 0.05; 0.0004375, means that the year and medal variables are not independent. The difference values between real data and chi square test are displayed in Figure 7. The result is displayed in Figure 8. Figure 9 shows "Year", "Medal" and the expected value by using Chi-Square algorithm for Winter Olympic.

```
Chi-Square Test Statistic : 79.33673872519334
p-value :  0.0004375394726699669
Degrees of freedom:  42
Null Hypothesis is rejected.
```

Figure 8. Result of Chi-Square Test of Independence for Winter Olympic

 ISSN:2302-8149
e-ISSN:2540-9719

Medal	Bronze	Gold	Silver
Year			
1924.0	41	38	39
1928.0	30	30	28
1932.0	39	33	44
1936.0	35	36	37
1948.0	46	46	31
1952.0	45	37	42
1956.0	43	24	44
1960.0	20	39	36
1964.0	29	28	55
1968.0	48	35	40
1972.0	28	24	53
1976.0	22	18	32
1980.0	46	38	32
1984.0	47	29	20
1988.0	60	33	46
1992.0	65	72	100
1994.0	114	114	115
1998.0	151	149	147
2002.0	161	163	157
2006.0	177	176	178
2010.0	176	176	177
2014.0	204	206	202

	Bronze	Silver	Gold
1924.0	39.781600	37.752176	40.466225
1928.0	29.667634	28.154165	30.178201
1932.0	39.107335	37.112308	39.780356
1936.0	36.410278	34.552839	37.036884
1948.0	41.467261	39.351844	42.180895
1952.0	41.804393	39.671778	42.523829
1956.0	37.421674	35.512640	38.065686
1960.0	32.027559	30.393701	32.578740
1964.0	37.758806	35.832574	38.408620
1968.0	41.467261	39.351844	42.180895
1972.0	35.398881	33.593038	36.008081
1976.0	24.273518	23.035226	24.691256
1980.0	39.107335	37.112308	39.780356
1984.0	32.364691	30.713634	32.921674
1988.0	46.861376	44.470783	47.667841
1992.0	79.900332	75.824285	81.275383
1994.0	115.636345	109.737257	117.626399
1998.0	150.698094	143.010361	153.291546
2002.0	162.160588	153.888106	164.951305
2006.0	179.017199	169.884791	182.098011
2010.0	178.342934	169.244923	181.412143
2014.0	206.324907	195.799420	209.875673

Figure 9. "Year" and "Medal" (Left) and the expected value by using Chi-Square algorithm (Right) for Winter Olympic

4.2.1 Multivariate Analysis

Multivariate analysis is a statistical procedure for analysis of data involving more than one type of measurement or observation. For the work purposes, the analysis that being conduct are to see the correlation between some variables in the data, The correlations are from the data base on variety of the Country, Summer and Winter medals that been achieved, Men and Women Medalist, the type of Medals count, Population, GDP per Capita and Sport type that had been won.

The algorithm used to count the correlations between each variable is Pearson's Correlation. The Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample. It is the normalization of the covariance between the two variables to give an interpretable score, resulting between -1 and 1 that represents the limit of correlation from a full negative to a full positive correlation. There are 3 types of correlations:

a. Positive Correlation: both variables change in the same direction.
b. Neutral Correlation: No relationship in the change of the variables.
c. Negative Correlation: variables change in opposite directions.

The correlation between variables visualization can be seen in Figure 12 and Figure 13.

	Summer	Winter	Men	Women	Bronze	Gold	Silver	Population	GDP per Capita	Sport
Summer	1.000000	0.675909	0.977929	0.934553	0.960908	0.974976	0.983436	0.219906	0.400978	0.762086
Winter	0.675909	1.000000	0.772886	0.711145	0.776860	0.737372	0.777643	0.083321	0.495621	0.658367
Men	0.977929	0.772886	1.000000	0.881625	0.974310	0.962395	0.986615	0.154770	0.457574	0.780627
Women	0.934553	0.711145	0.881625	1.000000	0.896883	0.939662	0.931769	0.306577	0.362295	0.719363
Bronze	0.960908	0.776860	0.974310	0.896883	1.000000	0.915746	0.979503	0.177195	0.474258	0.850573
Gold	0.974976	0.737372	0.962395	0.939662	0.915746	1.000000	0.962168	0.219721	0.380185	0.674808
Silver	0.983436	0.777643	0.986615	0.931769	0.979503	0.962168	1.000000	0.200915	0.463493	0.816304
Population	0.219906	0.083321	0.154770	0.306577	0.177195	0.219721	0.200915	1.000000	-0.090703	0.220674
GDP per Capita	0.400978	0.495621	0.457574	0.362295	0.474258	0.380185	0.463493	-0.090703	1.000000	0.483355
Sport	0.762086	0.658367	0.780627	0.719363	0.850573	0.674808	0.816304	0.220674	0.483355	1.000000

Figure 12. Result of Pearson's Correlation from different Countries Olympic 1896-2014

ISSN:2302-8149
e-ISSN:2540-9719

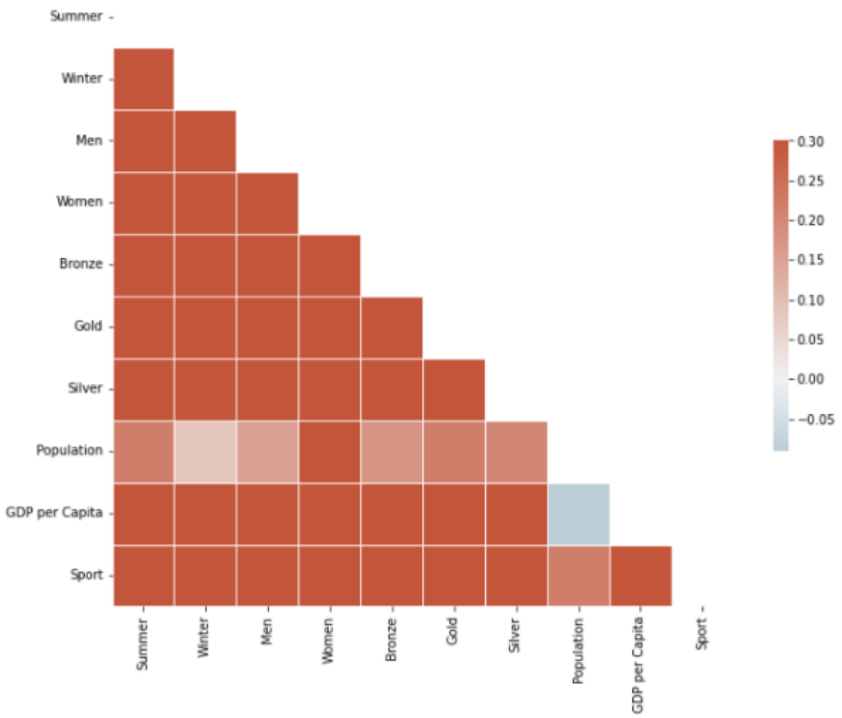

Figure 13. Pair-plot Pearson's Correlation from different Countries Olympic 1896-2014

From the result of Pearson's Correlation, we can conclude that bronze, gold, silver medals and Sport Won has a positive correlation with each other, which all the variables change in the same direction with each other.

5 Conclusion

The main objectives of this study are to find interesting insight from the data of Olympics' medalists from 1896 to 2014 and to find the relationships and correlation between variables of the data. Python are used to perform exploratory data analysis and Visual Studio Code as the Platform. It is chosen because it has various libraries that are easy-to-use and provide statistical computation and correlations between variables. The forms of analysis are presented in Univariate, Bivariate and Multivariate; produce interesting insights packaged in various forms of visualization. The result of univariate analysis is most of the medalist from 1896-2014 is Male athletes. From bivariate analysis, there is no relationship between Year and Medal variables in Olympics games. In other word, they are not independent each other. At last, Pearson correlation shows that most variables have a positive correlation with each other. The analysis result presented in this work might be used for future purposes. It is suggested to visualize the information in other forms such as map distributions, scatter plot, and boxplot.

Reference

[1] T. Felin, J. Koenderink, J. I. Krueger, D. Noble, and G. F. R. Ellis, "Data bias," *Genome Biol.*, vol. 22, no. 1, pp. 2–5, 2021, doi: 10.1186/s13059-021-02278-2.

[2] K. Nongthombam, "Data Analysis using Python - Sales Analysis," vol. 10, no. 07, pp. 463–468, 2021, [Online]. Available: https://www.storiesondata.com/post/data-analysis-using-python-sales-analysis.

[3] A. S. Rao, B. V. Vardhan, and H. Shaik, "Role of Exploratory Data Analysis in Data Science," *Proc. 6th Int. Conf. Commun. Electron. Syst. ICCES 2021*, no. July, pp. 1457–1461, 2021, doi: 10.1109/ICCES51350.2021.9488986.

[4] South Dakota State University, "Using exploratory data analysis (bivariate)," no. September, 2020, [Online]. Available: http://bioinformatics.sdstate.edu/users/gex/index/indexfiles/ch2.pdf.

[5] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory Data Analysis using Python," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 4727–4735, 2019, doi:

ISSN:2302-8149
e-ISSN:2540-9719

10.35940/ijitee.L3591.1081219.

[6] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain, "Secondary Analysis of Electronic Health Records," *Second. Anal. Electron. Heal. Rec.*, September, pp. 1–427, 2016, doi: 10.1007/978-3-319-43742-2.

[7] F. Paper, M. Rahmany, A. M. Zin, and A. Elankovan, "Comparing Tools Provided By Python," no. i, pp. 131–142, 2020.

[8] T. Lynn, P. Rosati, B. Nair, and C. M. an Bhaird, "An Exploratory Data Analysis of the #Crowdfunding Network On Twitter," *J. Open Innov. Technol. Mark. Complex.*, vol. 6, no. 3, Sep. 2020, doi: 10.3390/JOITMC6030080.

[9] J. Dsouza and S. Senthil Velan, "Using Exploratory Data Analysis for Generating Inferences on the Correlation of COVID-19 cases," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, pp. 2–7, 2020, doi: 10.1109/ICCCNT49239.2020.9225621.

[10] Sarvam Mittal, "An Exploratory Data Analysis of COVID-19 in India," *Int. J. Eng. Res.*, vol. V9, no. 04, Apr. 2020, doi: 10.17577/IJERTV9IS040550.

[11] R. Arunkumar, "An Exploratory Data Analysis Process on Groundwater Quality Data."

[12] R. Singh, "Exploratory Data Analysis and Customer Segmentation for Smartphones Analysis and Simulation of COVID-19 View project," 2021. [Online]. Available: https://www.researchgate.net/publication/351351474.

[13] I. Setiawan and S. Suprihanto, "Exploratory Data Analysis Of Crime Report," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 11, no. 2, pp. 71–80, 2021, doi: 10.31940/matrix.v11i2.2449.

[14] J. D. Miranda-Calle, V. Reddy C, P. Dhawan, and P. Churi, "Exploratory Data Analysis for Cybersecurity," *World J. Eng.*, vol. 18, no. 5, pp. 734–749, 2021, doi: 10.1108/WJE-11-2020-0560.

[15] R. Pradhan, K. Agrawal, and A. Nag, "Analyzing Evolution of the Olympics by Exploratory Data Analysis using R," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1099, no. 1, p. 012058, Mar. 2021, doi: 10.1088/1757-899x/1099/1/012058.