Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **923:929***

The Importance of Data Visualization in Exploratory Data Analysis

Patel Darshan Ratilal[1*], Dr. P.V Bhaskar Reddy[2]

[1,2]REVA University, Bangalore
Email: darshangorani17@gmail.com, bhskr.dwh@gmail.com

**Corresponding author's E-mail: darshangorani17@gmail.com*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 30 Nov 2023

Abstract

Data analysis or data science is the most talked about and buzz world in recent time it is also the most research area. Exploratory data analysis also popularly known as EDA is a statistical method or process which helps you to get a better understanding of the data or dataset which you are working on. Exploratory data analysis is considered an essential process in any data science project life cycle. The better you understand your data the better report you will provide or you will able to build more robust and better models. The EDA is consisting of several steps or is a process of several steps that you need to perform on your dataset. The data visualization technics help you a better representation of your data. There n-numbers of way to visualize your data. In this work, we are going to see the importance of data visualization in exploratory data analysis and the graphs you look for in any EDA. There are many paperwork and books available on exploratory data analysis and the steps involved in it. But here we will only try to focus on the different types of visualization techniques involved in the EDA. All the examples we going to see here are built by using python. There many tools available in the market to perform exploratory data analysis but in python where you write your own code to perform anything and python is widely used in the data science field. We will segregate each and every stage of EDA and see the important role plays by data visualization in order to understand the data you are working on.

Keywords: *Analytics, Exploratory data analysis, Data visualization, Python, matplotlib, seaborn.*

1. Introduction

Exploratory data analysis (EDA) is a statistical approach mainly use in the data preprocessing stage by many data scientists to get a better understanding of the data. The process of exploratory data analysis helps to investigate and summarize the main and vital points in the data. It will help them to handle data sources and to get the full benefit of them. This will make it effortless for them to come across the patterns, spot abnormality, experiment with the hypothesis, or experiment with a presupposition. EDA is primarily used to see what we can get from data beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. EDA is introduced by an American mathematician John Tukey in the year 1977, EDA techniques continue to be an evolved and most used technique in the data discovery process today field which helps in analyzing the existing data sets, to draw useful insights. These Insights are hidden information that aids in decision-making.

Data visualization refers to the graphical representation of data which is more essay understandable by humans. The Graphical representation helps you to share your thoughts more efficient way than the normal repot and it helps you to convince more people. There are many ways or different types of graphs available which you can use to represent your report or work in a more beautiful way. Data visualization is crucial in more or less each and every field. It can be used by instructors for displaying the student's test outcome, By computer scientists investigating the furtherance in artificial intelligence (AI), or by officers looking to share information with the collaborator. It will also play a crucial role in the big data sector. As businesses accumulated huge amounts of data during their first years of big data trends, they needed a way to quickly and easily access all of the data they collected from users. Visualization tools were a natural fit for them. Other importance of data visualization are:

The potential to communicate information quickly and improve details and facilitates faster commitment. A potential to communicate with the audience's interest in the information that they can easily understand. An easy lecture of information and increases the chance to share insights with everyone involved. Ability to help with immediate findings and achieve success with great speed and few mistakes.

Data visualization plays a critical role in exploratory data analysis it helps the data scientist to help the data points and the relationships between them in a very easy way and helps them to explain it to others if needed. There are different types of graphs that are used by data scientist in EDA some the Examples are:

Heatmap

Scatter plot

Bar chart

Distplot

RELATED WORK

Exploratory data analysis is an approach to know the data set fully before building any model on it.EDA engages a variety of techniques mostly graphical to get maximum insight from data, Discover the underlying facts, find out important variables. Detect outliers/anomalies and remove them. And determine optimal factor settings.

In the year 1977 John W. Tukey introduce exploratory data analysis to the world[1]. He finds out that the people are focusing more on the statistical hypothesis testing (confirmatory data analysis) rather they need to focus on the data to get the best optimal result possible. Then in the year 1997 Behrens, J. T.. publish his research named Principles and procedures of EDA[2]. Where he shows that statistical training and practice are required to increase efficiency. Yu Chong Ho in the year 1992 explains the logic behind the EDA[7].

Richard Dubes, A.K.Jain introduce the clustering method for EDA in 1980[6].

In the year 2012, Andrew Gelman proposes an approach for unifying the EDA with more ritual statistical methods based on probability models. In context to the more complex fields like psychology, medicine, and social science[3].In the year 2017 Victoria Cox writes the book on Exploratory data analysis[5]. Where he explains the EDA process in most simpler way possible

PROPOSED WORK

The proposed system is to perform the exploratory data analysis on a date and build different types of graphs and charts to understand our dataset. We will identify the graphs that are most important to the EDA process. And in the end, we will find out why data visualization empowers exploratory data analysis. We will use a sample dataset that is alive in the public domain in Kaggle. We going to using some python library in order to perform the task

A. Data set

The data set is taken from the Kaggle[8]. The dataset is a weather history dataset. Which has 96453 rows × 12 columns.the columns are Formatted Date,object Summary,object Precip Type, object Temperature (C), Apparent Temperature (C) ,Humidity, Wind Speed (km/h),Wind Bearing (degrees),Visibility (km),Loud Cover, Pressure (millibars) ,and Daily Summary.The data is recorded every 1hr from 2006 to 2016.

B. Python Libraries used

i .Pandas

Pandas in python is a Library that provides high-performance data manipulation and analysis. It uses powerful data structures. Pandas can be used to accomplish five steps in Data Analytics namely, load, prepare, manipulate, model, and analyze

ii. seaborn

Seaborn in python is a library that helps users to build highly interactive visualization with the minimum line of code.

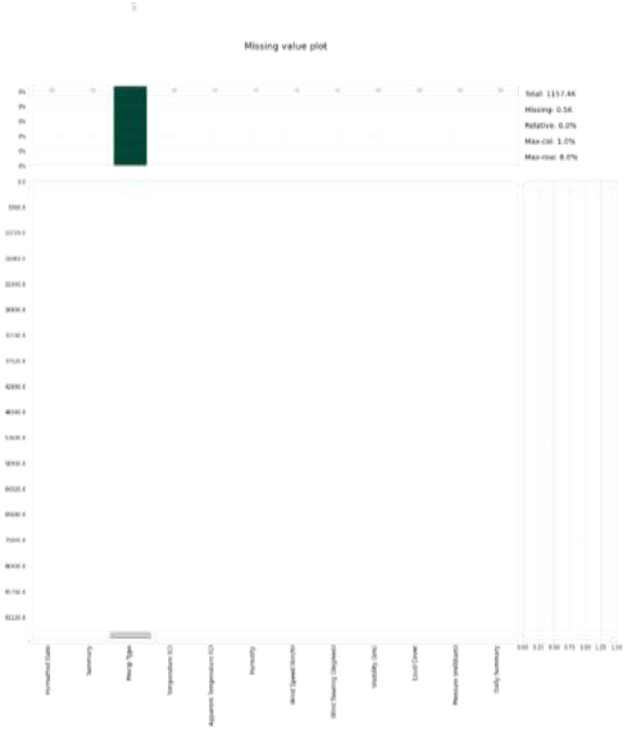
iii .klib

klib is a relatively new library in Python which aims to help the data scientist in the EDA process.

Implementation Of The Proposed Work

The implementation was done using Python visualization and data manipulation libraries. As we know the EDA process helps you to get a better understanding of the dataset but here we going to find the importance of data visualization. So let's start with one by one and discuss their importance

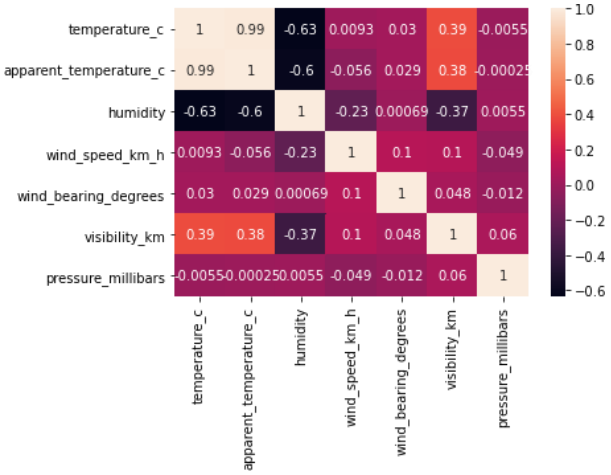
A. Missing Values Graph



(FIg. 1 Missing data plot)

Handing the missing data is an essential and challenging sept for every data science project. visualizing the missing values for every attribute is an attractive way to find out how many missing values or records are there for every attribute. The missing value plot uses the bar chart or bars plot wherein the x-axis represents the attributes name and the y-axis represented the percentage number of mission values. And right below the bar chart, you can see the link marks at the record number from where exactly the record is missing.

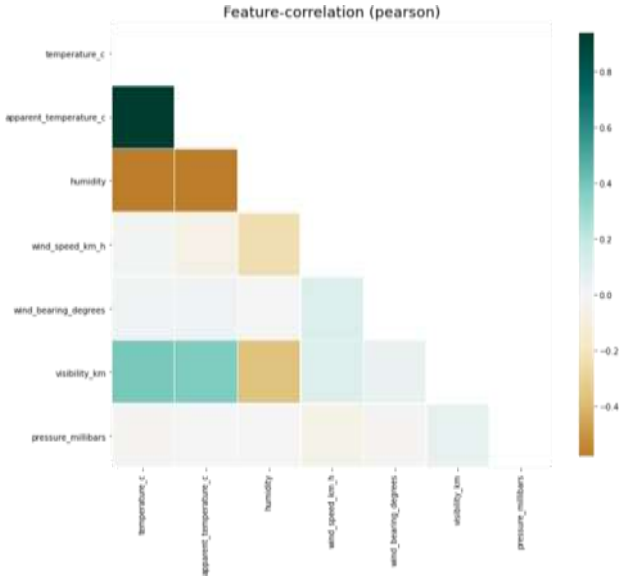
B. Correlation Graph



(Fig 2 Pearson plot using seaborn)

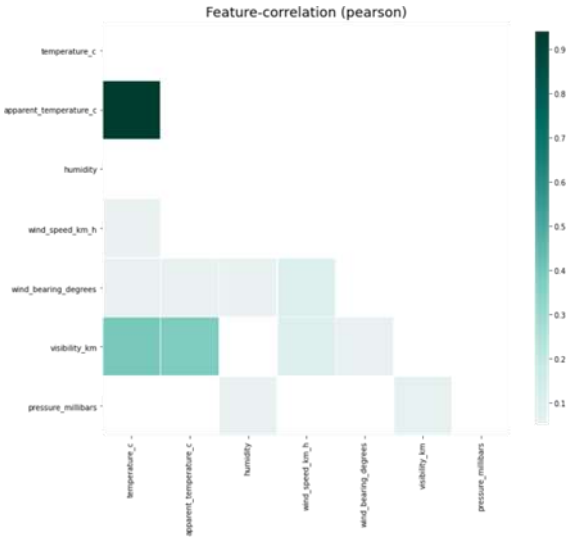
Correlation is a statistical method that helps you to find the relationship between the variable or attributes. Using the seaborn library for plotting the correlation will return a heatmap graph whit the number which represents the degree of relationship. The Pearson correlation matrix is a measure of the direct relationship between two data sets. It is a covariance of two kinds, separated by the product of

their common defects; therefore it is actually a standard measure of covariance so that the result always has a value between -1 and 1.



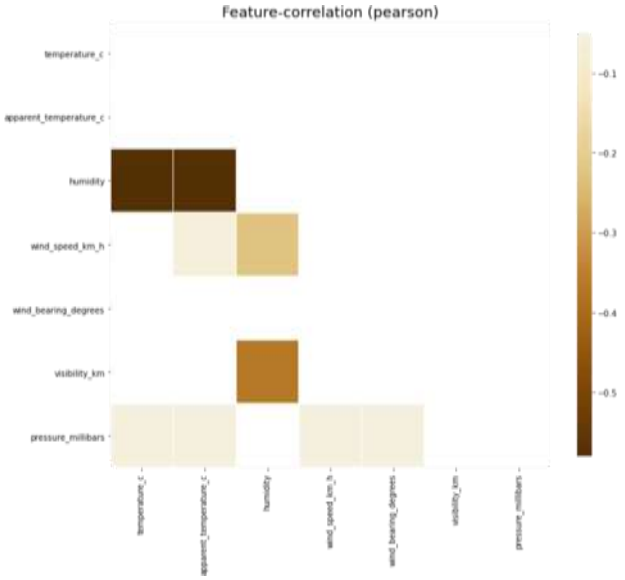
(Fig 2.1 Pearson correlation plot using Klib)

if you use the klib library to plot the correlation it will also return a heatmap but in a more attractive way. every numbers here represent a color that gives more pressure to your eyes. The klib was devolved for especially work in EDA it will provide you beautiful graphs. A heatmap is a clear representation of data where data values are represented as colors. That is, it uses color to transmit the value to the viewer.



(Fig 2.2 positive correlation)

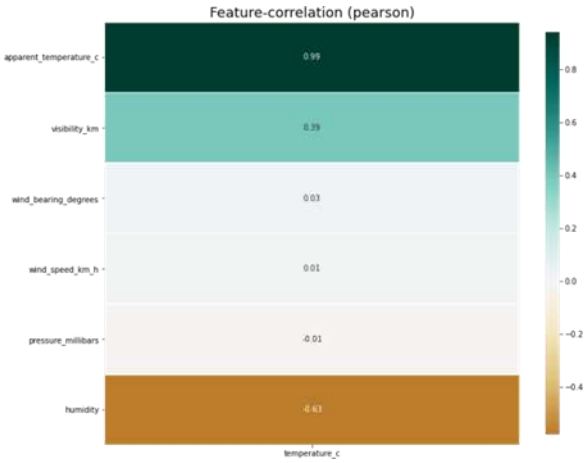
The correlation degree lies between -1 to 1 in the positive correlation plot only shows you the attributes which have a positive correlation. A positive relationship is a correspondence between two variables where the two variables move in sequence i.e., on the same side. A positive correlation exists when one variable decreases as one variance lessen or one variable escalate while the other escalates.



(Fig 2.4 Negative correlation)

The correlation degree lies between -1 to 1 in the negative correlation plot only shows you the attributes which have a negative correlation. A negative relationship is a correspondence between two variables where the two variables move in sequence i.e., on the opposite side. A negative correlation exists when one variable decreases as one variance escalate or one variable escalate while the other decreases.

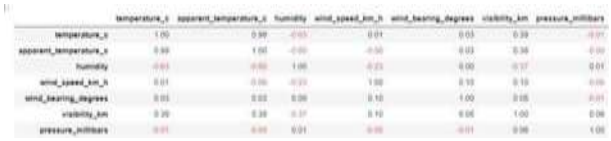
C. Correlation with respect to one variable



(Fig 3 Correlation with respect to one attribute)

The correlation coefficient gives you a fair idea of how the data are moving in your data set which will help to set the data correctly in your model. The plot correlation with respect to one attribute will help you get a better understanding of how the variable or attribute is moving with all the attributes present in the dataset. This will eventually help you to decide to set the boundary for the attribute which you want o work boundary in terms of the correlation coefficient.

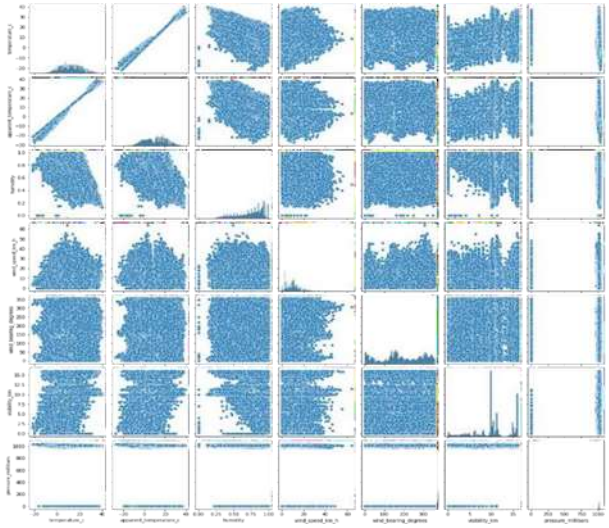
D. Correlation Table



(Fig 4 correlation Table)

The table representation is an old and less attractive way but yet this will give you more information in less time. The correlation coefficient table represents the coefficient directly with the number without any kind of graph. For this entire work, we are working with the Pearson correlation coefficient which is the most acceptable correlation coefficient by the data science community. Apart from Pearson correlation, there is Kendall rank correlation and Spearman correlation coefficients are also used by many data scientists.

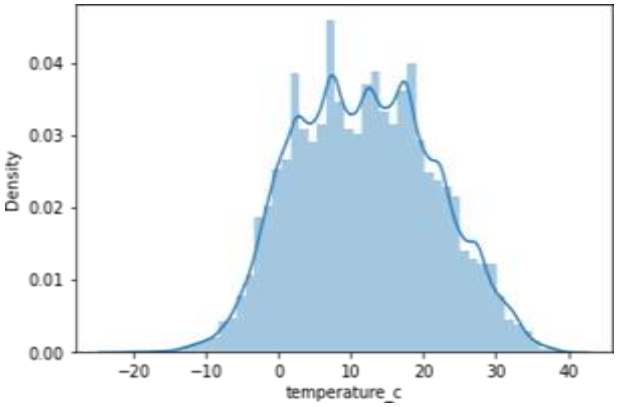
E. Pair plot



(Fig 5 Pair plot)

The Pair plot will help you to understand the range of data points how its distributed. The pair plot will return a scatter plot for every value or record present for every attribute. A scatter plot is plat in which points show the relationship between two sets of data. The pair plot gives you a rough idea of the data points distribution and the relation between them. The pair plot is a combination of scatter plots for all variables with respect to another variable. A pair plot is helpful when you want to save time other than the finding collation coefficient and data distribution.

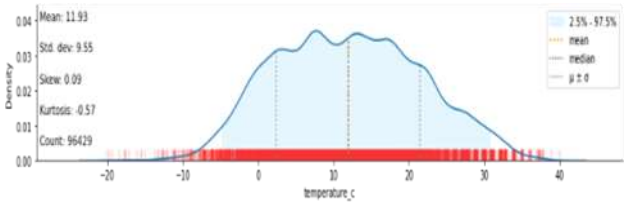
F. Frequency of occurrence



(Fig 6 Frequency plot)

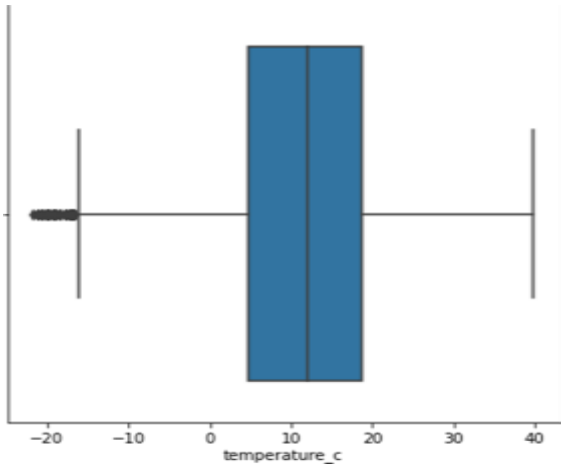
Visualizing the data points with one single target will help you to get the most occurrence data point in your data set or will help you get the rough idea of the most occurrence values. The frequency represents the number of times or the most common occurrence of something, here in our case is the data point the frequency plot show you the number of times the same data point is occurring in our dataset.

G. Distribution plot



The Distribution polt will give you the data distribution with respect to one attribute that you selected with all the necessary information like mean, median, and all other information related to the distribution. The distribution provides a partial mathematical method that can be used to compute the probability of any individual inspection from the sample space.

H. Outlier box plot



(Fig 8 Box plot)

Box structure or boxplot is a way of clearly identifying groups of numerical data by their quartiles. Box sites can have lines ranging from boxes showing differences other than the upper and lower quartiles, hence the names of the box-beard structure and the box-beard design. The box plot will help you to find out the outliers in your dataset. The other you to find the outliers is by using the clustering method but that will get lots of time and effort.

3. Results and Discussion

We have performed the exploratory data analysis in the weather dataset and used a different data visualization. Data visualization helps us to understand the data in a more efficient way. The Important visualization for any EDA process are:

- missing data plot
- correlation plot
- correlation with respect to a single attribute
- correlation table
- Frequency plot
- Data distribution plot
- Bot plot
- Pair plot

Data visualization is playing an important role in EDA to understand the dataset which we are working on its not only helps us to understand but it helps to share our information with others as well.

References:

1. John W. Tukey Exploratory Data Analysis Book 1977.
2. Behrens, John T. Principles and procedures of exploratory data analysis. Psychological Methods, 2(2), 131–160.
3. Andrew Gelman Exploratory Data Analysis for Complex Models 2012.
4. Daniel and R. Butson, "Foundations of big data and analytics in higher education," 2014.
5. Victoria Cox EDA book 2017.
6. Richard Dubes & A.K.Jain Clustering Methodologies in Exploratory Data Analysis 1980 S0065-2458(08)60034-0
7. Yu Chong Ho Abduction? Deduction? Induction? Is There a Logic of Exploratory Data Analysis? PUB DATE Apr 94 NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
8. https://www.kaggle.com/muthuj7/weather- dataset/activity
9. https://archive.ics.uci.edu/ml/datasets