Data analysis and visualization on titanic and student's performance datasets-an exploratory study

Seong-Cheol Kim, Surender Reddy Salkuti, Alka Manvayalar Suresh, Madhu Sree Sankaran
Department of Railroad and Electrical Engineering, Woosong University, Daejeon, Republic of Korea

Article Info

Article history:

Received May 6, 2024
Revised Oct 6, 2024
Accepted Nov 19, 2024

Keywords:

Data analysis
Data visualization
Exploratory data analysis
Performance datasets
Visualization methods

ABSTRACT

Exploratory data analysis (EDA) is all about exploring the data in order to identify any underlying pattern before you try to use it to make a predictive model. It also plays a major role in the data discovery process as it is used to analyze data and to recapitulate their different characteristics, which is displayed efficiently with the help of data visualization methods. This paper aims to identify errors in the dataset, to understand the existing hidden structure and to identify new ones, to detect points in a dataset that deviate to a greater extent from the collected data (outliers), and also to find any relationship or intersection between the variables and constants. Two datasets are used namely 'Titanic' and 'student's performance' to perform data analysis and 'data visualization' to depict 'exploratory data analysis' which acts as an important set of tools for recognizing a qualitative understanding. The datasets were explored and hence it assisted with identifying patterns, outliers, corrupt data, and discovering the relationship between the fields in the dataset.

Corresponding Author:

Surender Reddy Salkuti
Department of Railroad and Electrical Engineering, Woosong University
Jayang-Dong, Dong-Gu, Daejeon-34606, Republic of Korea
Email: surender@wsu.ac.kr

1. INTRODUCTION

The first and foremost step in a research study is to perform exploratory data analysis (EDA). The main intent of performing EDA is to aim at obtaining insights to a maximum extent involving various methods. Precisely, these techniques are put into use before initializing a statistical model or conducting complex analysis [1]. The procedure generally describes the dataset in a visual format for easy acknowledgment and to suggest ideas in the decision-making criteria of entities in the business. Visualization of the data deals with factors of identifying and testing data, to determine tendency and to spot the interdependence between sets [2]. Productive use of visualizing supplies a deal of working with huge datasets, exploring underlying features, intersections, structures, and trends on one side, paying way to new foresight and well-organized decision plans. Hence, a package of the pliability, relativeness, and knowledge of humans, tied with the extensive capacity of storage and computational power of data management is a high score point for its acceptance [3].

Loftus [4] aims at providing the methods of EDA that convene ways of chances that supply us a base to make decisions. These analyses range focusing on the variable type that involves both categorical and quantitative data. Even further, it aids us in realizing variable relationships, especially through both terse of graphical and mathematical plots and correlations. Shreiner [5] describes different polls, analyzing the methods of fifth-grade, eighth-grade, and high school students' data visualization methods. 27 students were examined and enquired to give a thought about a question that is historical whilst usage of the textbook which involved data visualization. Outcomes from quantitative and qualitative analyses focus on the capacity for data visualizations to improve reasoning [6]. The study of a non-linear mathematical model to enquire the impact of healthy sanitation and alertness on the spread of Coronavirus disease (COVID-19) existence is analyzed in reference [6] and it makes assumptions that education on the spread of the disease and measurements for prevention promote changes in the behavior of an individual to promote healthy hygiene, paying way to reduce the number of reproduction and burden of disease. Numerical simulations are employed involving real-life data to aid the outcomes. A theoretical model for the cancellation of booking analysis on slots in intercontinental shipping services is initially displayed. Also, a case study between Asia and the US west coast on container service is then performed based on the proposal of the model. The cancellation range of a voyage depends upon changes in ways of statistics and dynamics, the booking factors, and the cases that influence the cancellation aspects are checked and verified in reference [7].

The multivariate model and techniques of statistics are detailed in reference [8] and it stresses the highlight of data science in education and commercial backgrounds. Desimoni and Po [9] presents the deep discovery of the behavior of the art of tools for linked data visualization. It focuses on providing datasets in the form of graphs or information chosen by a user based on interest, with the motive to feature their analysis. The tools of visualization are explained and correlated based on their benefits and facilities. The employment of linear and nonlinear methods to visualize data in a lower dimension in the concept of EDA, the suitable method is the one that showcases the availability of natural clusters hidden in the scatter plots presented in reference [10]. Sankaranarayanan *et al.* [11] studies the methods of visualization techniques in multiple data concerning facts such as the arrival of flights, the count of passengers, booths, time, and patterns of seasons. This paper examines the correlation between these airports corresponding to different visualization. This work can move to every airport and can further define the analytical methods to forecast waiting time. The aim of reference [12] is to contribute a summary by initializing a visualization for nominees for Academy awards of the years between 1993 and 2017. The outcome is five various dashboards are commenced to display the pattern search for the parameters and the Holt-Winters exponential application for flattening prediction. Sweetlin and Saudia [13] illustrates EDA analyzing the factors: nottingham prognostic index, the overall survival status, and relapse free status collecting the data from the metabric breast cancer to discuss the rate of survival and recurrence of the disease amidst various patients aged 5 years and 10 years. The EDA is executed employing the tools of visualization and the recorded observations are visualized enabling appropriate swarm plots and tables.

Data analysis generally deals with the inspection of data, the transformation of the dataset into a lower dimension, and modeling it to gain effective description and reasoning and to make a conclusion. It is one of the rising areas in the sector of business and technology. Data visualization works with the representation of data in the form of a graph. It converts the raw data present into an approachable way to visualize and recognize the trends. It uses visual tools like charts, graphs, maps, and other visual methods. A brief note has been provided in this section detailing the implementation of data analysis and data visualization, performed in two provided data sets from Kaggle namely “titanic” and “student’s performance”.

2. DATA ANALYSIS AND DATA VISUALIZATION

It is important for an analyst to know the business deals and problems faced by an organization to explore and draw meaningful insights from the raw data. There are four types of data analysis. Descriptive data analysis: descriptive analysis is the type of analysis that is used to obtain the summary of the set of data points to satisfy every order of the data. It is categorized as one of the prominent methods of conducting statistical data analysis [14]. The descriptive analysis draws a summary for the split of the data, helps to spot outliers, and helps to identify similarities among variables, allowing the statistical analysis to a greater extent. The descriptive uses aggregation of data and tasks related to mining to fetch results based on previous data. Diagnostic analysis: diagnostic analysis analyses and interprets the data deeply to identify the anomalies that cannot be fully explained by current understanding. It looks into the data for previous underlying patterns and determines the causal relationships between patterns that lead to anomalies. This special type of analysis helps to understand future events. Predictive analysis: the predictive model employs statistics and the solutions are forecasted based on future outcomes. It works with optimization algorithms to reach possible solutions on factors to attain finer outcomes in the future [15]. It comes up with better guidance on the chances for better outputs. Prescriptive analysis: prescriptive analytics has the top capacity today in the field of analytics that determines in development of the areas via predictive and descriptive analytics. It aids the users in providing effective answers to the existing problems and opting for the best choice among various options [16].

2.1. Data visualization techniques

It assists to handle big data, presenting comparisons, employing the required type of chart, utilizing different palettes in graphs, and also fits apt for the digital age. Visualization is a basic step in the procedure of scientific discovery, although it was not popular till the late 80's until it was fittingly accepted as a stage in the area of research [17]. Gómez-Romero *et al.* [18] highlighted the factor that visualization supports important 3 motives: (i) brief checking of models in simulation process; (ii) fast pace production of obtained results in simulation; and (iii) easy access to communicate to the end-user. In the field of Data Science, the method of visualization is considered a valuable asset in the process of exploring and presenting the different levels of the result; i.e., at the initial and the last procedure of data analysis. Visually exploring the data is specifically fruitful when source data is known less and the goals of the analysis are not provided in a detailed manner. In this idea presented, visual data exploration is looked forward to accessing an easy step to solve hypothesis-generation procedures, in which they can be accepted or neglected based on visuals, and also new ones can be introduced [19]. The visualization technique is powerful as it explores the data with presentable and interpretable results. It visualizes phenomena that cannot be observed directly. Various visualization methods include bar charts, cartograms, dot distribution maps, bullet graphs, scatter plots, area charts, and bubble clouds.

Scatterplots serve as a basic method of visualization. The available tools of flexibility lead to the usage of scatter plots in various contexts of exploration and presentation. Scatterplots illustrate every individual object in the data with a point, placed on dimensions of orthogonal and two continuous points [20]. Scatter plot methods have been employed to a greater extent to present statistical graphics and to expose hidden structures in the multivariate dataset. These plots are identified as one of the finest, polymorphic, and commonly beneficial methods for displaying pairwise axes of correlation structures and low-dimension patterns of the available data alongside the summary for a large total of data [21]. The bar chart is one of the most common visualization methods which is used to plot the x and y axes in the form of a graph for the data in the fashion of comparison of the numerical components. It can be accessed only on the elucidated dataset and is capable of detailing the basic information [22]. Histograms are among the most widely used data visualization methods. A histogram is normally referred to as a chart that displays numeric data in ranges [23]. It shows the number of times a response or range of responses occurs in a data set.

2.2. Execution requirements

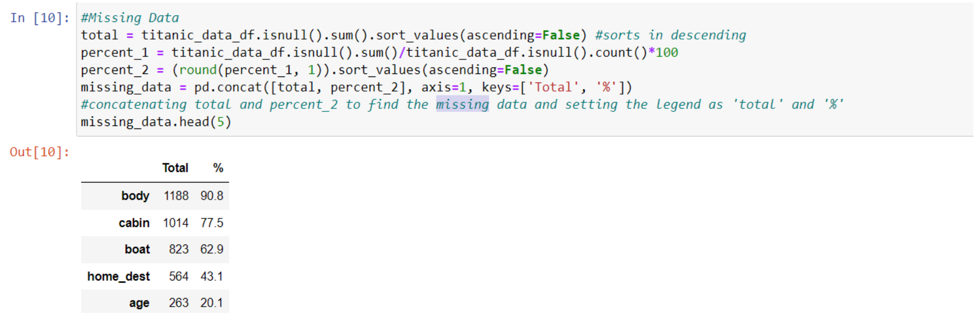
Pandas is one of the most extensively used libraries in Data Science. Pandas(pd) is an open-source Python library for enabling manipulation and analysis of data. Numpy(np) is also one of the important libraries in Python. It is used for working with arrays, linear algebra, matrices, and Fourier transforms [24], [25]. Matplotlib(plt) and seaborn(sns) are utilized for graphical representation that serves visual purposes. These libraries are accessed to visualize the data in the form of graphs and charts to present comparisons. From the mentioned libraries, there are a lot of functions associated with execution [26]. Some of them include head() which is used to represent the first five columns of the data. tail() represents the last five columns of the dataset. info() specifies the different data types used in the data set and specifies the order and name of the column, count of null and non-null values, and memory.shape gives the count of the number of rows and columns of the dataset. drop( ) drops or deletes specific columns in the dataset. size returns the size. .ndim returns the number of dimension arrays. dropna() drops the null values while fillna() fills null values. drop_duplicates() drops values that are duplicated. There are other functionalities explained along with the result obtained.

3. IMPLEMENTATION OF DATA ANALYSIS

The functions used in this cell are is.null().sum() in order to find the sum of null values, 'sort_values(ascending=False)' sorts the values present in data in descending order, 'concat()' performs concatenation along the axis, here 'axis=1' specifies the rows, 'keys' to name the columns of the output. count() is employed in order to determine how many times a particular element appears in a provided list or string of values, function head() returns the top five rows whereas tail() returns the last five rows. round() is an inbuilt function that returns round-off values to the given digit along with the floating-point number, if no digit is provided it rounds off to the nearest integer [27], [28]. Here all these functions are employed to fill missing values in Figure 1.

As missing values are filled in the above cell, 'is.null().sum()' is used to find if there are any null values present in the dataset. The output obtained is 0 which indicates there are no null values in Figure 2. In Figure 3, corr() is employed to perform a pairwise correlation of existing columns in the dataset. Any null values are automatically dropped. It only performs in numeric columns, 'method=pearson' is used to obtain standard correlation coefficient. In the Pearson method, only uses 3 numerical values where 0 is for no correlation, 1 for positive correlation in total, and -1 for the negative correlation.

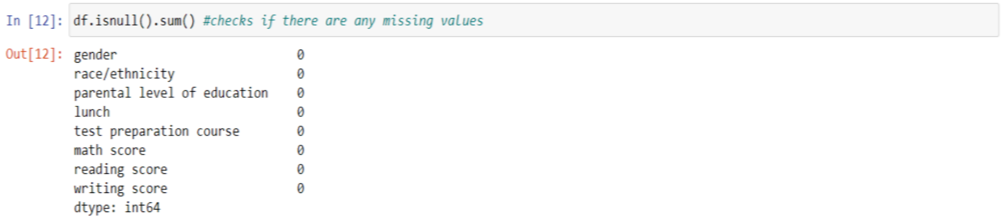
'unique()' returns the count of unique values, here the count of unique values in the sex field is returned along with its datatype. The mentioned three functions, min(), max(), and mean() are usually employed to count maximum, minimum, and average values. In Figure 4, to find the survival and death rates for the youngest and oldest, value '0' is assigned as the initial value for survival, and '1' is assigned for the death rate. The function min() is used for youngest as age is less when compared for the oldest, which is calculated using the max() function, and finally '.format' is used to handle complex strings formatting process in a more efficient way. In Figure 5, the minimum, maximum, and average age for the oldest and youngest on-board is determined to understand the patterns lying in the dataset more efficiently.



```
In [10]: #Missing Data
         total = titanic_data_df.isnull().sum().sort_values(ascending=False) #sorts in descending
         percent_1 = titanic_data_df.isnull().sum()/titanic_data_df.isnull().count()*100
         percent_2 = (round(percent_1, 1)).sort_values(ascending=False)
         missing_data = pd.concat([total, percent_2], axis=1, keys=['Total', '%'])
         #concatenating total and percent_2 to find the missing data and setting the legend as 'total' and '%'
         missing_data.head(5)
Out[10]:
```

	Total	%
body	1188	90.8
cabin	1014	77.5
boat	823	62.9
home_dest	564	43.1
age	263	20.1

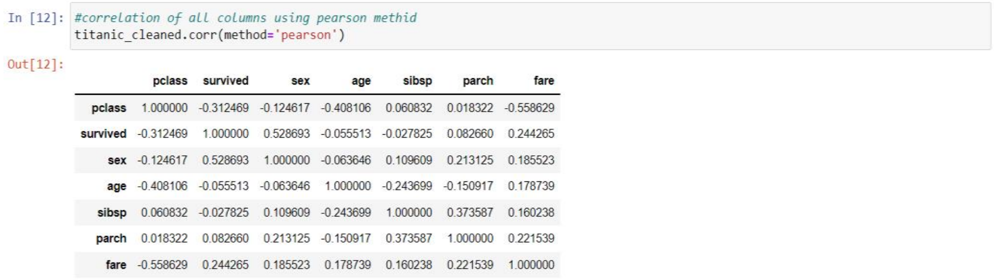
Figure 1. Fill missing values



```
In [12]: df.isnull().sum() #checks if there are any missing values

Out[12]: gender                         0
         race/ethnicity                 0
         parental level of education    0
         lunch                          0
         test preparation course        0
         math score                     0
         reading score                  0
         writing score                  0
         dtype: int64
```

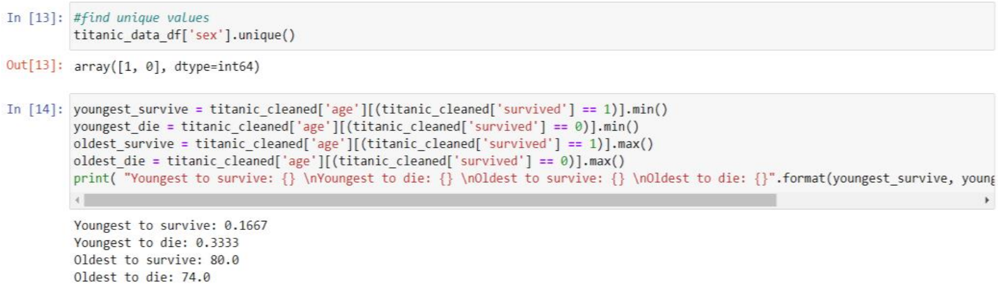
Figure 2. Sum of Null values



```
In [12]: #correlation of all columns using pearson methid
         titanic_cleaned.corr(method='pearson')
Out[12]:
```

	pclass	survived	sex	age	sibsp	parch	fare
pclass	1.000000	-0.312469	-0.124617	-0.408106	0.060832	0.018322	-0.558629
survived	-0.312469	1.000000	0.528693	-0.055513	-0.027825	0.082660	0.244265
sex	-0.124617	0.528693	1.000000	-0.063646	0.109609	0.213125	0.185523
age	-0.408106	-0.055513	-0.063646	1.000000	-0.243699	-0.150917	0.178739
sibsp	0.060832	-0.027825	0.109609	-0.243699	1.000000	0.373587	0.160238
parch	0.018322	0.082660	0.213125	-0.150917	0.373587	1.000000	0.221539
fare	-0.558629	0.244265	0.185523	0.178739	0.160238	0.221539	1.000000

Figure 3. Correlation



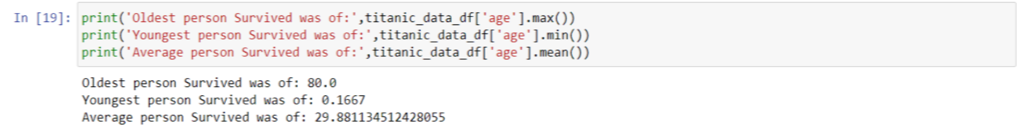
```
In [13]: #find unique values
         titanic_data_df['sex'].unique()

Out[13]: array([1, 0], dtype=int64)

In [14]: youngest_survive = titanic_cleaned['age'][(titanic_cleaned['survived'] == 1)].min()
         youngest_die = titanic_cleaned['age'][(titanic_cleaned['survived'] == 0)].min()
         oldest_survive = titanic_cleaned['age'][(titanic_cleaned['survived'] == 1)].max()
         oldest_die = titanic_cleaned['age'][(titanic_cleaned['survived'] == 0)].max()
         print( "Youngest to survive: {} \nYoungest to die: {} \nOldest to survive: {} \nOldest to die: {}".format(youngest_survive, young

         Youngest to survive: 0.1667
         Youngest to die: 0.3333
         Oldest to survive: 80.0
         Oldest to die: 74.0
```

Figure 4. Survival and death

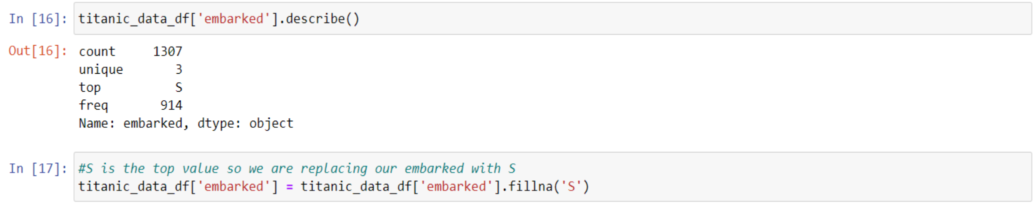


```
In [19]: print('Oldest person Survived was of:',titanic_data_df['age'].max())
         print('Youngest person Survived was of:',titanic_data_df['age'].min())
         print('Average person Survived was of:',titanic_data_df['age'].mean())

         Oldest person Survived was of: 80.0
         Youngest person Survived was of: 0.1667
         Average person Survived was of: 29.881134512428055
```

Figure 5. Minimum, maximum, and average age

'describe()' displays the number of values present in the column, count of unique values, frequency, and the top/repeated value in the column, fillna() fills null values for the field accessed. In Figure 6, describe() is used to access information about the embark column in the dataset and the result shows the most frequently used value is 's', so the same value is being employed to fill the null values in the same column. 'groupby() function is used to split the data into groups based on a given condition and 'replace()' is an inbuilt function where all occurrences of a substring are replaced with another or new string and returns a copy of it. In Figure 7, the initial column values are being replaced and 'inplace=True' is used to modify the data in place as it returns nothing and the data frame is being updated. Followed by that grouping is done based on age in an initial column where mean() is considered. 'iloc( )' provides the location of the integer based on indexing. In general, the function is used when the label of the index is not numeric or in case if the index label is not provided. In Figure 8, ':' symbolizes extracting all rows with index.

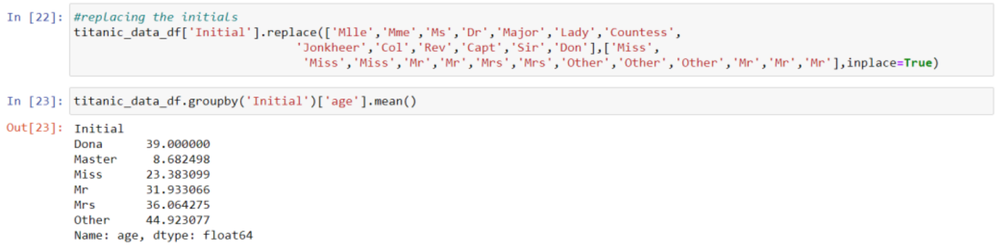


```
In [16]: titanic_data_df['embarked'].describe()

Out[16]: count     1307
         unique       3
         top          S
         freq       914
         Name: embarked, dtype: object

In [17]: #S is the top value so we are replacing our embarked with S
         titanic_data_df['embarked'] = titanic_data_df['embarked'].fillna('S')
```

Figure 6. describe(), fillna()

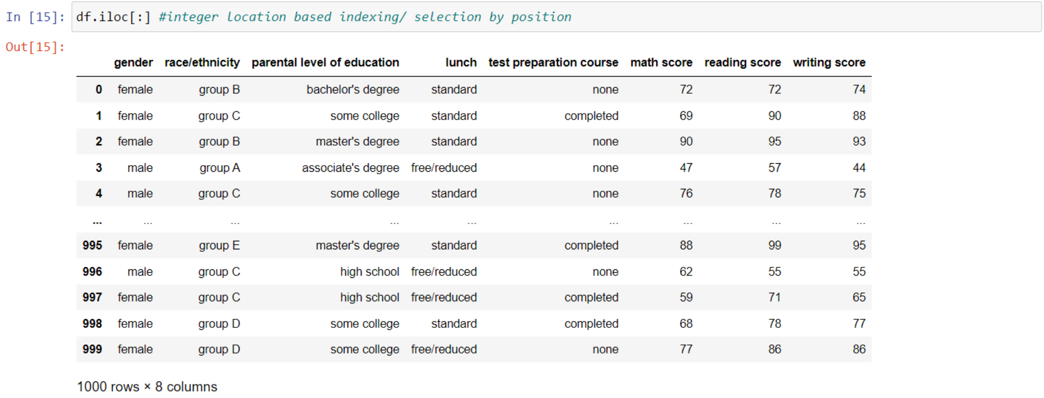


```
In [22]: #replacing the initials
         titanic_data_df['Initial'].replace(['Mlle','Mme','Ms','Dr','Major','Lady','Countess',
                                   'Jonkheer','Col','Rev','Capt','Sir','Don'],['Miss',
                                    'Miss','Miss','Mr','Mr','Mrs','Mrs','Other','Other','Other','Mr','Mr','Mr'],inplace=True)

In [23]: titanic_data_df.groupby('Initial')['age'].mean()

Out[23]: Initial
         Dona      39.000000
         Master     8.682498
         Miss      23.383099
         Mr        31.933066
         Mrs       36.064275
         Other     44.923077
         Name: age, dtype: float64
```

Figure 7. replace(), groupby()



```
In [15]: df.iloc[:] #integer location based indexing/ selection by position

Out[15]:
```

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
...	...	...	...	...	...	...	...	...
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

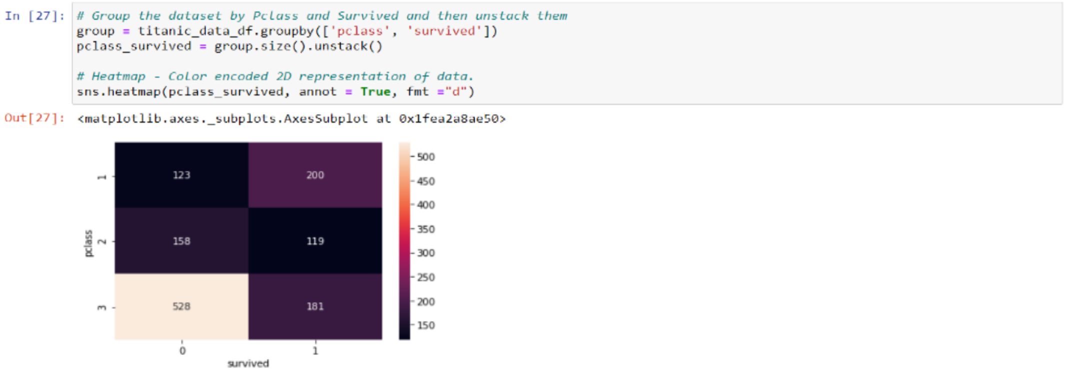
1000 rows × 8 columns

Figure 8. iloc()

4. IMPLEMENTATION OF DATA VISUALIZATION

'size()' function returns the size of the data frame whereas 'unstack()' function is used to convert the dataset into an unstacked format or in general to remove an existing stack. In 'heatmap' colors are used to represent the value in the matrix, seaborn library. The darker shade represents the common values while the lighter ones represent the least common activity. In Figure 9, the dataset is divided based on two existing columns with a return in size of the data frame in an unstacked format, followed by a heatmap plotted.

'pyplot' represents a pie-chart where the slices will be ordered and plotted counter-clockwise. In Figure 10, a pyplot is plotted for existing columns like student's gender, ethinicity, lunch, and parental level of education exhibiting different styles and variations. 'barplot' displays the central value for a numerical variable using the height of each rectangle. We use 'kwargs' with double stars and it is flexible to handle arguments. **kwargs is a built-in function that takes any number of keyword arguments. 'Keyword arguments' are declared by a name and if the value is not provided it takes a default value and it will not produce an error. **kwargs is an 'empty dictionary'. Each undefined argument is stored as key-value pair. In Figure 11, bar plots accessing different styles and enabling various functions are shown, which are plotted for different columns in the dataset provided with a legend to understand the score in a more precise fashion. In Figure 12, as already explained 'countplot()' is plotted with different variants.



```
In [27]: # Group the dataset by Pclass and Survived and then unstack them
         group = titanic_data_df.groupby(['pclass', 'survived'])
         pclass_survived = group.size().unstack()

         # Heatmap - Color encoded 2D representation of data.
         sns.heatmap(pclass_survived, annot = True, fmt ="d")

Out[27]: <matplotlib.axes._subplots.AxesSubplot at 0x1fea2a8ae50>
```

Figure 9. Heat map

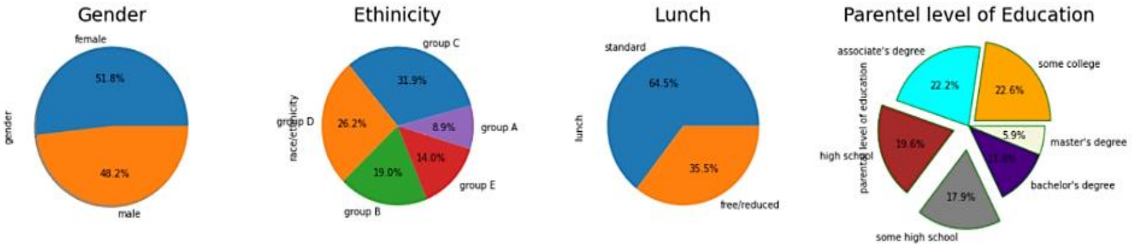


Figure 10. Pyplot

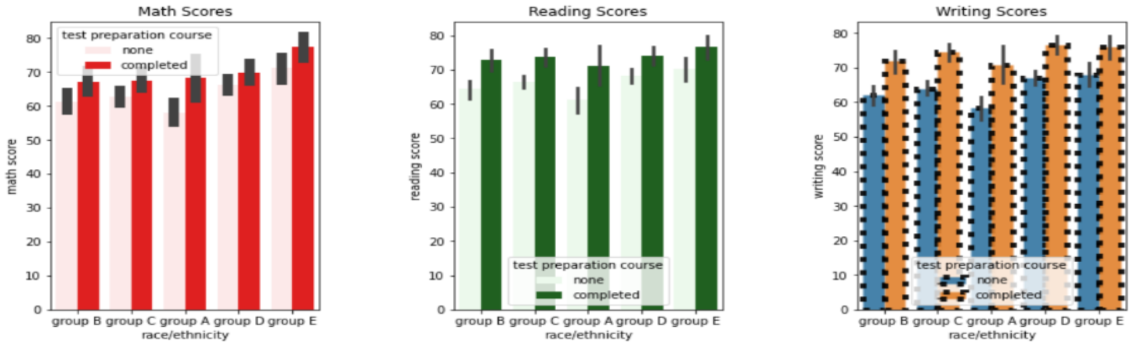


Figure 11. Bar plot

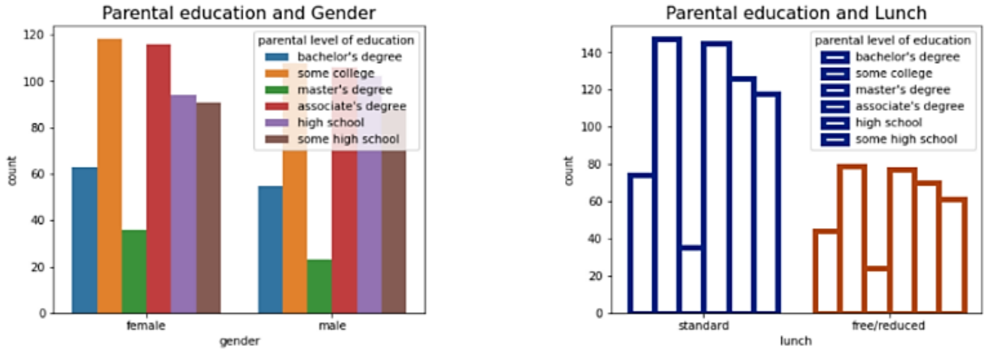


Figure 12. Countplot

Two datasets are used namely 'titanic' and 'student's performance' to perform data analysis and 'data visualization' to depict 'exploratory data analysis' which acts as an important set of tools for recognizing a qualitative understanding. The datasets were explored and it assisted with identifying patterns, outliers, corrupt data, and discovering the relationship between the fields in the dataset. Visualization reduces the tedious task of reviewing every single existing column to find the common point, highest, and lowest values and it is user-friendly as the graph demonstrates a detailed pattern with easily understandable structures. As concerning analysis, the dataset has been reduced from a higher to a lower dimension, the data is described to know about the existing data types, and information about the count of values is known, null values are filled using the repeated value, column names have been replaced the way it is recognized by the code performed, grouping and indexing are performed for easy readability. All the mentioned specifications are executed in both datasets. As concerning in titanic's data, the survival and non-survival rate depending on the status of the class is visually depicted, alongside with youngest and oldest to survive to understand the insights. Whilst, in student's data score, are plotted, sorted, and cleaned from highest to lowest or vice-versa for understanding the strongest and weakest domain among the students respectively.

5. CONCLUSIONS

This paper concludes in a way where EDA is discussed with understanding the shape of a data set, exploring correlations within the data, and determining if there's a signal for modeling an outcome based on the different features. In conclusion, EDA methods are exactly put in before the conventional method of modeling commences and the improvement of sophisticated models in statistics which helps to determine the importance of different features within a data set. Data analysis plays a crucial role as the results are acceptable, identified and detailed correctly, and applicable to the required contexts of business. Data analysis techniques have been devised to help in this conclusion. This paper summarizes the main characteristics, often employing visuals. Different techniques which either include non-graphical or graphical and univariate or multivariate are described. This way, it serves the contents required to improve a relevant model for the obstacle at the side to efficiently portray its outcomes

ACKNOWLEDGEMENTS

This research work was supported by "Woosong University's Academic Research Funding - 2025".

REFERENCES

[1] D. Lord, X. Qin, and S. R. Geedipally, "Highway safety analytics and modeling," in *Elsevier*, Elsevier, 2021, pp. 135–177.

[2] J. Dsouza and S. Senthil Velan, "Using exploratory data analysis for generating inferences on the correlation of COVID-19 cases," in *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, Jul. 2020, pp. 1–6, doi: 10.1109/ICCCNT49239.2020.9225621.

[3] O. V. Johnson, O. T. Jinadu, and O. I. Aladesote, "On experimenting large dataset for visualization using distributed learning and tree plotting techniques," *Scientific African*, vol. 8, p. e00466, Jul. 2020, doi: 10.1016/j.sciaf.2020.e00466.

[4] S. C. Loftus, "Exploratory data analyses: describing our data," in *Basic Statistics with R*, Elsevier, 2022, pp. 47–72.

[5] T. L. Shreiner, "Students′ use of data visualizations in historical reasoning: a think-aloud investigation with elementary, middle, and high school students," *Journal of Social Studies Research*, vol. 43, no. 4, pp. 389–404, Oct. 2019, doi: 10.1016/j.jssr.2018.11.001.

[6] M. O. Adeniyi *et al.*, “Dynamic model of COVID-19 disease with exploratory data analysis,” *Scientific African*, vol. 9, p. e00477, Sep. 2020, doi: 10.1016/j.sciaf.2020.e00477.

[7] H. Zhao, Q. Meng, and Y. Wang, “Exploratory data analysis for the cancellation of slot booking in intercontinental container liner shipping: a case study of Asia to US west coast service,” *Transportation Research Part C: Emerging Technologies*, vol. 106, pp. 243–263, Sep. 2019, doi: 10.1016/j.trc.2019.07.009.

[8] L. P. Fávero and P. Belfiore, “Introduction to data analysis and decision making,” in *Data Science for Business and Decision Making*, Elsevier, 2019, pp. 3–5.

[9] F. Desimoni and L. Po, “Empirical evaluation of linked data visualization tools,” *Future Generation Computer Systems*, vol. 112, pp. 258–282, Nov. 2020, doi: 10.1016/j.future.2020.05.038.

[10] A. Nasser, D. Hamad, and C. Nasr, “Visualization methods for exploratory data analysis,” in *2006 2nd International Conference on Information & Communication Technologies*, 2006, vol. 1, pp. 1379–1384, doi: 10.1109/ICTTA.2006.1684582.

[11] H. B. Sankaranarayanan, G. Agarwal, and V. Rathod, “An exploratory data analysis of airport wait times using big data visualisation techniques,” in *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Oct. 2016, pp. 324–329, doi: 10.1109/CSITSS.2016.7779379.

[12] R. H. Amier and J. Setiawan, “Visualization and prediction of film award nominations by using of visual data mining (VDM) and exploratory data analysis (EDA) method,” in *2019 5th International Conference on New Media Studies (CONMEDIA)*, Oct. 2019, pp. 84–88, doi: 10.1109/CONMEDIA46929.2019.8981822.

[13] E. J. Sweetlin and S. Saudia, “Exploratory data analysis on breast cancer dataset about survivability and recurrence,” in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, May 2021, pp. 304–308, doi: 10.1109/ICSPC51351.2021.9451811.

[14] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, “Educational data mining: predictive analysis of academic performance of public school students in the capital of Brazil,” *Journal of Business Research*, vol. 94, pp. 335–343, Jan. 2019, doi: 10.1016/j.jbusres.2018.02.012.

[15] S. Poornima and M. Pushpalatha, “A survey on various applications of prescriptive analytics,” *International Journal of Intelligent Networks*, vol. 1, pp. 76–84, 2020, doi: 10.1016/j.ijin.2020.07.001.

[16] J. Tinoco, M. Parente, A. G. Correia, P. Cortez, and D. Toll, “Predictive and prescriptive analytics in transportation geotechnics: Three case studies,” *Transportation Engineering*, vol. 5, p. 100074, Sep. 2021, doi: 10.1016/j.treng.2021.100074.

[17] “Visualization in scientific computing,” *IEEE Computer Graphics and Applications*, vol. 7, no. 10, pp. 69–69, 1987, doi: 10.1109/mcg.1987.276849.

[18] J. Gómez-Romero, M. Molina-Solana, A. Oehmichen, and Y. Guo, “Visualizing large knowledge graphs: a performance analysis,” *Future Generation Computer Systems*, vol. 89, pp. 224–238, Dec. 2018, doi: 10.1016/j.future.2018.06.015.

[19] N. Bikakis, S. Maroulis, G. Papastefanatos, and P. Vassiliadis, “In-situ visual exploration over big raw data,” *Information Systems*, vol. 95, p. 101616, Jan. 2021, doi: 10.1016/j.is.2020.101616.

[20] A. Sarikaya and M. Gleicher, “Scatterplots: tasks, data, and designs,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 402–412, Jan. 2018, doi: 10.1109/TVCG.2017.2744184.

[21] Q. V. Nguyen, N. Miller, D. Arness, W. Huang, M. L. Huang, and S. Simoff, “Evaluation on interactive visualization data with scatterplots,” *Visual Informatics*, vol. 4, no. 4, pp. 1–10, Dec. 2020, doi: 10.1016/j.visinf.2020.09.004.

[22] D. A. Keim, M. C. Hao, U. Dayal, and M. Lyons, “Value-cell bar charts for visualizing large transaction data sets,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 4, pp. 822–833, Jul. 2007, doi: 10.1109/TVCG.2007.1023.

[23] M. S. Sankaran, A. M. Suresh, and S. R. Salkuti, “Performance evaluation of time series analysis on the dataset of airline passengers: an exploratory data analysis,” in *Lecture Notes in Electrical Engineering*, vol. 1023 LNEE, 2023, pp. 613–619.

[24] S. Farzadnia and I. Raeesi Vanani, “Identification of opinion trends using sentiment analysis of airlines passengers’ reviews,” *Journal of Air Transport Management*, vol. 103, p. 102232, Aug. 2022, doi: 10.1016/j.jairtraman.2022.102232.

[25] G. L. Taboada and L. Han, “Exploratory data analysis and data envelopment analysis of urban rail transit,” *Electronics (Switzerland)*, vol. 9, no. 8, pp. 1–29, Aug. 2020, doi: 10.3390/electronics9081270.

[26] A. Jadhav and S. K. Shandilya, “Reliable machine learning models for estimating effective software development efforts: a comparative analysis,” *Journal of Engineering Research*, vol. 11, no. 4, pp. 362–376, Dec. 2023, doi: 10.1016/j.jer.2023.100150.

[27] S. R. Salkuti, “Emerging and advanced green energy technologies for sustainable and resilient future grid,” *Energies*, vol. 15, no. 18, p. 6667, Sep. 2022, doi: 10.3390/en15186667.

[28] J. A. Momoh and S. S. Reddy, “Review of optimization techniques for renewable energy resources,” in *PEMWA 2014 - 2014 IEEE Symposium on Power Electronics and Machines for Wind and Water Applications*, Jul. 2014, pp. 1–8, doi: 10.1109/PEMWA.2014.6912225.

BIOGRAPHIES OF AUTHORS

Seong-Cheol Kim received B.S., M.S. and Ph.D. degrees, in Electronic Engineering from Korea University in 1987, 1989 and 1997, respectively. He is currently serving as Head of Railroad and Electrical Engineering Department, Woosong University, Korea. His research interests are mobile communication system and pulsed power system. He can be contacted at email: kmin@wsu.ac.kr.

Surender Reddy Salkuti received Ph.D. degree in electrical engineering from the Indian Institute of Technology (IIT), New Delhi, India, in 2013. He was a Postdoctoral Researcher at Howard University, Washington, DC, USA, from 2013 to 2014. He is currently an Associate Professor at the Department of Railroad and Electrical Engineering, Woosong University, Daejeon, Republic of Korea. His current research interests include market clearing, including renewable energy sources, demand response, and smart grid development with the integration of wind and solar photovoltaic energy sources. He can be contacted at email: surender@wsu.ac.kr.

Alka Manvayalar Suresh received her degree in B.E Electrical and Electronics Engineering from Loyola ICAM College of Engineering and Technology, Chennai, Tamil Nadu, India. She had been working as data analyst to identify potential bottlenecks and adjust operations to ensure maximum efficiency. She is currently pursuing her Msc in Data Science and Artificial Intelligence at University of Liverpool, UK. She is actively in involved in projects that explore the intersection of data science, artificial intelligence, and industry application. She can be contacted at email: alkasuresh2000@gmail.com.

Madhu Sree Sankaran received B.E degree in Electrical and Electronics Engineering and Msc in Data Science at University of Southampton, UK. She serves as a Data Analyst, leveraging her expertise to support process and performance improvements within her organization. Her interests extend beyond data analysis into business analysis and project management. She recognizes the interconnectedness of these disciplines and their critical roles in driving organizational success. She can be contacted at email: madhu23rd@gmail.com.