**A**

**Seminar**

**Report on**

**“Exploratory data analysis techniques”**

**Submitted For**

**Partial Fulfilment of the Requirement for the degree of**

**Master of Computer Application**

**(Under Engineering)**

**Submitted By**

**Ms. Siddhi Pramod Wani (56)**

**Under The Guidance Of**

**Prof. Dhanashree Kolape**



**Department of Master of Computer Application**

**G. S. Moze COLLEGE OF ENGINEERING,**

**BALEWADI, PUNE**

**Savitribai Phule Pune University, Pune**

**2023-2024**

**G. S. Moze College Of Engineering, Balewadi Pune**

# CANDIDATE’S DECLARATION

I hereby certify that the work which is being presented in the Seminar entitled

“**Exploratory data analysis techniques”** by

“**Wani Siddhi Pramod**” in partial fulfillment of requirements for the award of the degree of MCA (Engineering) submitted in the Department of

MCA at G.S.Moze College of Engineering, Balewadi Pune under Savitribai Phule Pune University an authentic record of my own work carried out during a period from 16/01/2024 to 25/03/2024. The matter presented in this seminar has not been submitted by me or anybody else in any other University for the award of an M.C.A Degree.

Name & Signature Of The Candidate   
Siddhi Pramod Wani

**ACKNOWLEDGEMENT**

I feel great pleasure in submitting this **seminar** report on **“Exploratory data analysis techniques”**. I would like to express profound gratitude to **Prof. Dhanashree Kolape** for her valuable support, encouragement, supervision, and useful suggestions throughout this report work. Her moral support and continuous guidance enable me to complete my work successfully.

I am thankful to my Head of Department**, Prof. Mukta Deshpande** for his continuous motivation and support. I take this opportunity to thank our Hon. Principal, **Prof. Dr. Ratnaraja**

**Kumar Jambi** for setting high benchmarks and allowing us for the seminar.

**Project Guide:** **Head Of Dept:**  **Principal** Prof.Dhanashree Kolape Prof .Mukta Deshpande Dr.RatnarajaKuma Jambi

# ABSTRACT

Exploratory Data Analysis (EDA) serves as the cornerstone of data analysis, enabling practitioners to gain insights, detect patterns, and identify anomalies within datasets. This abstract delves into the essential techniques of EDA, highlighting its significance in uncovering actionable insights from raw data. Starting with understanding the data's structure and characteristics, EDA progresses through univariate, bivariate, and multivariate analyses, allowing analysts to explore relationships and dependencies among variables. Furthermore, data preprocessing techniques are discussed, emphasizing the importance of preparing data for downstream analysis. Through real-world case studies and examples, this abstract demonstrates how EDA empowers analysts to make informed decisions and drive business outcomes. Finally, best practices and tips are provided to guide practitioners in conducting effective EDA and extracting valuable insights from diverse datasets.

# 

**INDEX**

|  |  |  |
| --- | --- | --- |
| **SR.NO** | **TOPICS** | **PAGE NO.** |
| **1** | **INTRODUCTION** | **6** |
| **2** |  |  |
| **3** |  |  |
| **4** |  |  |
| **5** |  |  |
| **6** |  |  |
| **7** |  |  |
| **8** | **CONCLUSION** |  |
| **9** | **REFERENCE** |  |
| **10** |  |  |

# INTRODUCTION

**Definition of Exploratory Data Analysis:**

Data scientists implement exploratory data analysis tools and techniques to investigate, analyse, and summarize the main characteristics of datasets, often utilizing data visualization methodologies.

Exploratory Data Analysis is a process of examining or understanding the data and extracting insights dataset to identify patterns or main characteristics of the data. EDA is generally classified into two methods, i.e. graphical analysis, and non-graphical analysis.

EDA techniques allow for effective manipulation of data sources, enabling data scientists to find the answers they need by discovering data patterns, spotting anomalies, checking assumptions, or testing a hypothesis.

Data specialists primarily use exploratory data analysis to discern what datasets can reveal further beyond formal modeling of data or hypothesis testing tasks. This enables them to gain in-depth knowledge of the variables in datasets and their relationships.

Exploratory data analysis can help detect obvious errors, identify outliers in datasets, understand relationships, unearth important factors, find patterns within data, and provide new insights.

Developed in the 1970s by American statistician John Tukey - famed for his box plot techniques and the Fast Fourier Transform algorithm - EDA continues to find relevance even today in the field of statistical analysis. It allows data professionals to produce relevant and valid results that drive desired business goals.

Exploratory Data Analysis is a process of examining or understanding the data and extracting insights dataset to identify patterns or main characteristics of the data. EDA is generally classified into two methods, i.e. graphical analysis and non-graphical analysis

# Exploratory Data Analysis Techniques

Table suggest a few EDA techniques depending on the type of data and the objective of the analysis.

|  |  |
| --- | --- |
| Type of data | Suggested EDA techniques |
| Categorical | Descriptive statistics |
| Univariate continuous | Line plot, Histograms |
| Bivariate continuous | 2D scatter plots |
| 2D arrays | Heatmap |
| Multivariate: trivariate | 3D scatter plot or 2D scatter plot with a 3rd variable represented in different color, shape or size |
| Multiple groups | Side-by-side boxplot |

## Non-graphical EDA

These non-graphical methods will provide insight into the characteristics and the distribution of the variable(s) of interest.

1. **Univariate non-graphical** :

This is the simplest type of EDA, where data has a single variable. Since there is only one variable, data professionals do not have to deal with relationships.

A simple univariate non-graphical EDA method for categorical variables is to build a table containing the count and the fraction (or frequency) of data of each category

Sample statistics express the characteristics of a sample using a limited set of parameters. They are generally seen as estimates of the corresponding population parameters from which the sample comes from. These characteristics can express the central tendency of the data (arithmetic mean, median, mode), its spread (variance, standard deviation, interquartile range, maximum and minimum value) or some features of its distribution (skewness, kurtosis). Many of those characteristics can easily be seen qualitatively on a histogram.

A diagram of a normal distribution

Description automatically generated

Fig. Symmetrical versus asymmetrical (skewed) distribution, showing mode, mean and median.

Central tendency parameters

The arithmetic mean, or simply called the mean is the sum of all data divided by the number of values. The median is the middle value in a list containing all the values sorted. Because the median is affected little by extreme values and outliers, it is said to be more “robust” than the mean.

Variance

When calculated on the entirety of the data of a population (which rarely occurs), the variance r2 is obtained by dividing the sum of squares by n, the size of the population.

The sample formula for the variance of observed data conventionally has n-1 in the denominator instead of n to achieve the property of “unbiasedness”, which roughly means that when calculated for many different random samples from the same population, the average should match the corresponding population quantity (here r2). s2 is an unbiased estimator of the population variance r2.

A mathematical equation with numbers and symbols

Description automatically generated

The standard deviation is simply the square root of the variance. Therefore it has the same units as the original data, which helps make it more interpretable.

The sample standard deviation is usually represented by the symbol s. For a theoretical Gaussian distribution, mean plus or minus 1, 2 or 3 standard deviations holds 68.3, 95.4 and 99.7 % of the probability density, respectively.

Interquartile range (IQR)

The IQR is calculated using the boundaries of data situated between the 1st and the 3rd quartiles. Please refer to the Chap. [13](http://dx.doi.org/10.1007/978-3-319-43742-2_13) “Noise versus Outliers” for further detail about the IQR.

IQR ¼ Q3 — Q1

In the same way that the median is more robust than the mean, the IQR is a more robust measure of spread than variance and standard deviation and should therefore be preferred for small or asymmetrical distributions.

Important rule:

* Symmetrical distribution (not necessarily normal) and N > 30: express results as mean ± standard deviation.
* Asymmetrical distribution or N < 30 or evidence for outliers: use

median ± IQR, which are more robust.

Skewness/kurtosis

Skewness is a measure of a distribution’s asymmetry. Kurtosis is a summary statistic communicating information about the tails (the smallest and largest values) of the distribution. Both quantities can be used as a means to communicate infor- mation about the distribution of the data when graphical methods cannot be used.

1. **Multivariate Non-graphical EDA**

Multivariate data consists of several variables. Non-graphic multivariate EDA methods illustrate relationships between 2 or more data variables using statistics or cross-tabulation.

*Cross-Tabulation*

Cross-tabulation represents the basic bivariate non-graphical EDA technique. It is an extension of tabulation that works for categorical data and quantitative data with only a few variables. For two variables, build a two-way table with column headings matching the levels of one variable and row headings matching the levels of the other variable, then *ﬁ*ll in the counts of all subjects that share a pair of levels. The two variables may be both exposure, both outcome variables, or one of each.

*Covariance and Correlation*

Covariance and correlation measure the degree of the relationship between two random variables and express how much they change together. The covariance is computed as follows:

A math equation with numbers and symbols

Description automatically generated

where *x* and *y* are the variables, *n* the number of data points in the sample, ¯*x* the mean of the variable x and ¯*y* the mean of the variable y.

A positive covariance means the variables are positively related (they move

together in the same direction), while a negative covariance means the variables are inversely related. A problem with covariance is that its value depends on the scale of the values of the random variables. The larger the values of x and y, the larger the

A diagram of a diagram of a number of black dots

Description automatically generated

Fig: Examples of covariance for three different data sets

covariance. It makes it impossible for example to compare covariances from data sets with different scales (e.g. pounds and inches). This issue can be *ﬁ*xed by dividing the covariance by the product of the standard deviation of each random variable, which gives Pearson’s correlation coef*ﬁ*cient.

Correlation is therefore a scaled version of covariance, used to assess the linear relationship between two variables and is calculated using the formula below.



where Cov (x, y) is the covariance between x and y and  are the sample standard deviations of x and y. The signi*ﬁ*cance of the correlation coef*ﬁ*cient between two normally distributed variables can be evaluated using Fisher’s z transformation Other tests exist for measuring the non-parametric rela- tionship between two variables, such as Spearman’s rho or Kendall’s tau.

## Graphical EDA

1. **Univariate Graphical EDA:**

*Histograms*

Histograms are among the most useful EDA techniques, and allow you to gain insight into your data, including distribution, central tendency, spread, modality and outliers.

Histograms are bar plots of counts versus subgroups of an exposure variable. Each bar represents the frequency (count) or proportion (count divided by total count) of cases for a range of values. The range of data for each bar is called a bin. Histograms give an immediate impression of the shape of the distribution (symmetrical, uni/multimodal, skewed, outliers…). The number of bins heavily influences the *ﬁ*nal aspect of the histogram; a good practice is to try different values, generally from 10 to

50. Some examples of histograms are shown below.

Histograms enable to con*ﬁ*rm that an operation on data was successful. For example, if you need to log-transform a data set, it is interesting to plot the his- togram of the distribution of the data before and after the operation.

Histograms are interesting for *ﬁ*nding outliers. For example, pulse oximetry can be expressed in fractions (range between 0 and 1) or percentage, in medical records.

A graph of a graph with Willis Tower in the background

Description automatically generated

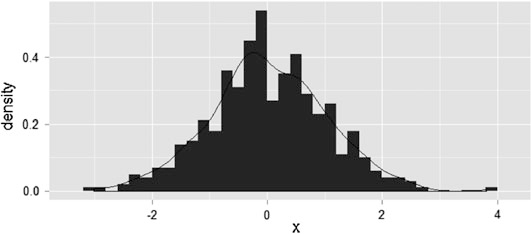
Fig. Example of histogram

Fig. Example of histogram with density estimate

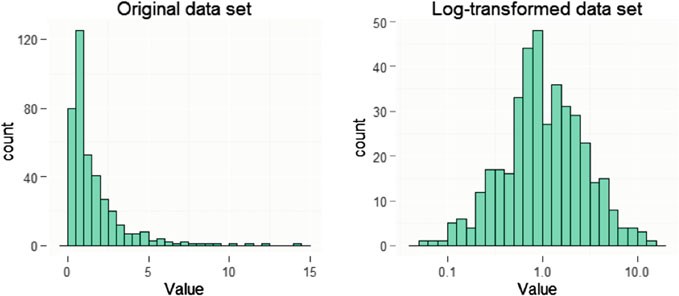


Fig. Example of the effect of a log transformation on the distribution of the dataset

Stem Plots

Stem and leaf plots (also called stem plots) are a simple substitution for histograms. They show all data values and the shape of the distribution.

A graph of a function

Description automatically generated

Fig. Example of stem plot

*Boxplots*

Boxplots are interesting for representing information about the central tendency, symmetry, skew and outliers, but they can hide some aspects of the data such as multimodality. Boxplots are an excellent EDA technique because they rely on robust statistics like median and IQR.

Below Figure shows an annotated boxplot which explains how it is constructed. The central rectangle is limited by Q1 and Q3, with the middle line representing the median of the data. The whiskers are drawn, in each direction, to the most extreme point that is less than 1.5 IQR beyond the corresponding hinge. Values beyond 1.5 IQR are considered outliers.

The “outliers” identi*ﬁ*ed by a boxplot, which could be called “boxplot outliers” are de*ﬁ*ned as any points more than 1.5 IQRs above Q3 or more than 1.5 IQRs below Q1. This does not by itself indicate a problem with those data points. Boxplots are an exploratory technique, and you should consider designation as a boxplot outlier as just a suggestion that the points might be mistakes or otherwise unusual. Also, points not designated as boxplot outliers may also be mistakes. It is also important to realize that the number of boxplot outliers depends strongly on the size of the sample. In fact, for data that is perfectly normally distributed, we expect

0.70 % (about 1 in 140 cases) to be “boxplot outliers”, with approximately half in either direction.

A diagram of a bar

Description automatically generated

Fig. Example of boxplot with annotations

A graph showing a line graph

Description automatically generated2D Line Plot

Fig. Example of 2D line plot

2D line plots represent graphically the values of an array on the y-axis, at regular intervals on the x-axis .

*Probability Plots (Quantile-Normal Plot/QN Plot, Quantile-Quantile Plot/QQ Plot)*

Probability plots are a graphical test for assessing if some data follows a particular distribution. They are most often used for testing the normality of a data set, as many statistical tests have the assumption that the exposure variables are approx- imately normally distributed. These plots are also used to examine residuals in models that rely on the assumption of normality of the residuals (ANOVA or regression analysis for example).

The interpretation of a QN plot is visual either the points fall randomly around the line (data set normally distributed) or they follow a curved pattern instead of following the line (non-normality). QN plots are also useful to identify skewness, kurtosis, fat tails, outliers, bimodality etc.

A graph of a normal q-q plot

Description automatically generated

Fig. Example of QQ plot

Besides the probability plots, there are many quantitative statistical tests (not graphical) for testing for normality, such as Pearson Chi2, Shapiro-Wilk, and Kolmogorov-Smirnov.

Deviation of the observed distribution from normal makes many powerful statistical tools useless. Note that some data sets can be transformed to a more normal distribution, in particular with log-transformation and square-root trans- formations. If a data set is severely skewed, another option is to discretize its values into a *ﬁ*nite set.

1. **Multivariate Graphical EDA:**

*Side-by-Side Boxplots*

This EDA technique makes use of graphics to show relationships between 2 or more datasets. The widely-used multivariate graphics include bar chart, bar plot, heat map, bubble chart, run chart, multivariate chart, and scatter plot.

A diagram of a number of levels of p

Description automatically generated Representing several boxplots side by side allows easy comparison of the characteristics of several groups of data. An example of such boxplot is shown in the case study.

Fig. Side-by-side boxplot showing the cardiac index for ﬁve levels of Positive end-expiratory pressure (PEEP)

*Scatterplots*

A graph of a number of blue dots

Description automatically generated with medium confidenceScatterplots are built using two continuous, ordinal or discrete quantitative variables Each data point’s coordinate corresponds to a variable. They can be complexi*ﬁ*ed to up to *ﬁ*ve dimensions using other variables by differentiating the data points’ size, shape or color.

Fig. Scatterpolot showing an example of actual mortality per rate of predicted mortality

A graph showing a number of dots

Description automatically generatedScatterplots can also be used to represent high-dimensional data in 2 or 3D , using T-distributed stochastic neighbor embedding (t-SNE) or prin- cipal component analysis (PCA). t-SNE and PCA are dimension reduction features used to reduce complex data set in two (t-SNE) or more (PCA) dimensions.

Fig. 3D representation of the *ﬁ*rst three dimension of a PCA

*Curve Fitting*

A graph with a line and numbers

Description automatically generatedCurve *ﬁ*tting is one way to quantify the relationship between two variables or the change in values over time. The most common method for curve *ﬁ*tting relies on minimizing the sum of squared errors (SSE) between the data and the

Fig. Example of linear regression

*ﬁ*tted function. Please refer to the “Linear Fit” function to create linear regression slopes in R.

More Complicated Relationships

Many real life phenomena are not adequately explained by a straight-line relationship. An always increasing set of methods and algorithms exist to deal with that issue. Among the most common:

* Adding transformed explanatory variables, for example, adding x2 or x3 to the model.
* Using other algorithms to handle more complex relationships between variables (e.g., generalized additive models, spline regression, support vector machines, etc.).

Heat Maps and 3D Surface Plots

Heat maps are simply a 2D grid built from a 2D array, whose color depends on the value of each cell. The data set must correspond to a 2D array whose cells contain the values of the outcome variable. This technique is useful when you want to represent the change of an outcome variable (e.g. length of stay) as a function of two other variables (e.g. age and SOFA score).

A graph of a graph of a graph

Description automatically generated with medium confidenceThe color mapping can be customized (e.g. rainbow or grayscale). Interestingly, the Matlab function *imagesc* scales the data to the full colormap range. Their 3D equivalent is mesh plots or surface plots.

Fig. Heat map (*left*) and surface plot (*right*)

# Conclusion

In summary, EDA is an essential step in many types of research but is of particular use when analyzing electronic health care records. The tools described in this chapter should allow the researcher to better understand the features of a dataset and also to generate novel hypotheses.

Take Home Messages

1. Always start by exploring a dataset with an open mind for discovery.
2. EDA allows to better apprehend the features and possible issues of a dataset.
3. EDA is a key step in generating research hypothesis.

## References

## 

## PLAGIARISM REPORT