LITERATURE SURVEY

DATA ANALYSIS

I. INTRODUCTION

Data are growing very faster in today’s world. It is not so

1.*INTRODUCTION*

world. It is not so

easy to process the data manually. Data analysis and

visualization programs allow for reaching even deeper

understanding. The programming language Python, with its

English commands and easy-to-follow syntax, offers an

amazingly powerful (and free!) open-source alternative to

traditional techniques and applications.

Data analytics allow businesses to understand their

efficiency and performance, and ultimately helps the business

make more informed decisions. For example, an e-commerce

company might be interested in analyzing customer attributes

in order to display targeted ads for improving sales. Data

analysis can be applied to almost any aspect of a business if

one understands the tools available to process information.

The ecommerce companies are analyzing the reviews of

customer by using proper visualization method. Exploratory

Data Analysis (EDA) is an approach to summarize the data by

taking their main characteristics and visualize it with proper

representations. EDA focuses more narrowly on checking

assumptions required for model fitting and hypothesis testing,

and handling missing values and making transformations of

variables as needed. EDA encompasses IDA.

*Exploratory Data Analysis using Python*

It is the first stage of data analysis. Here we can know about

the content of the data set and characteristic of data set. It tells

about the size of the data. We can find the missing value of

data. We can find the possible relationship among data. Data

visualization is done by the use of tabular data and

understanding the characteristics.

*Data Cleaning* :

It is process of detecting the corrupt data, removing the

irrelevant parts of the data and replacing the correct data. The

actual process of data cleaning is to remove the error and

validating the data. Data can be cross checked to remove the

error. Issue can be resolved by validating the data.

*Model Building*

We use the statistical model or machine learning model to

describe the variable and working of the variable. Model can

be supervised or unsupervised model. We can use

classification, regression model to get the output. We can

visualize the result by the use of model. After that we have to

evaluate the model

*Present Result*

We can visualize large amount of complex data by the use

of chart, graph and tables. Human brain can process

information using chart, graphs. It is an easy way to convey

the concept. It can identify the area which needs

improvement. It can clarify the factor very well.

*Graphical EDA*

Fundamentally, graphical exploratory data analysis is nothing

but the graphical counterpart of the traditional non-graphical

EDA that analyzes the data sets to help summarize their

statistical characteristics focusing on the same four key

aspects, like, measures of central tendency, measures of

spread, the shape of the distribution and the existence of

outliers.

*Univariate Graphical EDA*

Univariate GEDA provides statistical summary for each field

in the raw data set or the summary only on one variable.

*Bivariate Graphical EDA*

Bivariate GEDA is accomplished to understand the

connections between each variable in the dataset and the

target variable of interest or using two variables and finding

connection among them. Example of these types of GEDA

includes Box plot and Violin plot

*Multivariate Graphical EDA*

Multivariate GEDA is accomplished to understand the

connections between different fields in the dataset or finding

the connections between more than two variables.

*EDA IN PYTHON*

We are using python for exploratory data analysis. It is

simple to learn. It has rich sets of libraries. Data handling

capacity are much higher. It is used as open source language.

It has the capacity to with all the third party language .it can

run on any platform. It can transfer the process from one

platform to another. It is easy to read.

*Pandas*

It is the most powerful package for data analysis. We can

clean, transform and analyze the data. Data can be stored in

CSV format in computer. Cleaning, visualizing and storing

the data can be done. It is built on the top of the NumPy

package. Plotting functions from Matplotlib and machine

learning algorithm in Scikit-learn.

*Jupiter Notebook*

It gives ability to execute the code in a particular cell. It gives

the console based approach for computing. It provides web

based application process. It includes input and output of the

computation. It gives rich media representation of the object.

*WORKING WITH THE DATA SETS*

It’s time to explore the data and find about it. The data we

are using belongs to Amazon review data set. We are going to

analyse the data with possible set of options.

1. In the first step we have imported the Pandas libraries.

numpy packages.

2. After that we have imported fairly large amazon CSV file

as a data frame df. It gives the data sets in the form of rows

and column.

*Box Plot:* We have used categorical variable which takes a

fixed number of possible values .it describes the

characteristics of a data unit. It is represented by box plot.

We have done the box plot between:

1. Manufacturers no and reviews ratings.

2. Manufacturer and reviews ratings

*Count Plot:* We have used count plot to count the no

of observations. It can be taught as a histogram across a

categorical variable

*Descriptive Statistical Analysis:* We have used descriptive

statistical analysis which is used to describe the entire data

sets with a single value or metric. The describe function

automatically computes basic statistics for all continuous

variables.

*Counts:* We have used count function that returns the

number of occurrences. It tells about how many units of

each characteristic/ variable we have. We got number of

brand value and the different categories electronic

products.

*Applications of EDA*

1. Mistakes and anomalies can be detected using EDA

2. We can gain new insight in to various types of data

3. Outliers in data can be detected

4. We can test assumption using EDA

*CONCLUSION*

In this article we have explained the detail about explorative

data analysis. We have used the language python

programming language for implementation. We have used

jupyter note book for detail analysis. We have implemented

different library packages of python.