ACROPOLIS INSTITUTE OF TECHNOLOGY AND RESEARCH

Department of Information Technology

Synopsis

On

Exploratory Data Analysis

1.INTRODUCTION

1.1 Overview

Data are growing very faster in today's 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.

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.EDA quickly describes the data sets number of rows/columns, missing data, data types and preview. Clean corrupted data; handle missing data, invalid data types andincorrect values.EDA visualize data distributions; bar charts, histograms, box plots. Calculate and visualize correlations (relationships) between variables; heat map.

1.2 Abstract

Data need to be analyzed so as to produce good result. Using the result decision can be taken. For example recommendation system, ranking of the page, demand fore casting, prediction of purchase of the product. There are some leading companies where the review of the customer plays a great role to analyze the factor which influences the review rating. We have used exploratory data analysis (EDA) where data interpretations can be done in row and column format. We have used python for data analysis. it is object oriented ,interpreted and interactive programming language. it is open source with rich sets of libraries like pandas, MATplotlib etc. We have used different types of charts and various types of parameter to analyze data sets. We have used python programming for the data analysis.

1.3 Purpose

The purpose of this Project is to maximize insights into a dataset by mapping out its underlying structure, identifying useful variables, detecting outliers and anomalies, and testing hypotheses. Through exploratory data analysis (EDA) using Python programming and associated libraries, we aim to uncover hidden patterns, relationships, and trends within the data to make informed decisions and drive actionable insights across various domains.

2. LITERATURE SURVEY

2.1 Existing Problems and Solutions

Researchers have explored various data exploration tools and techniques, including classical data analysis versus exploratory data analysis, studies on bank lending, credit risk assessment, and loan prediction using machine learning techniques.

Few pre existing analysis includes:

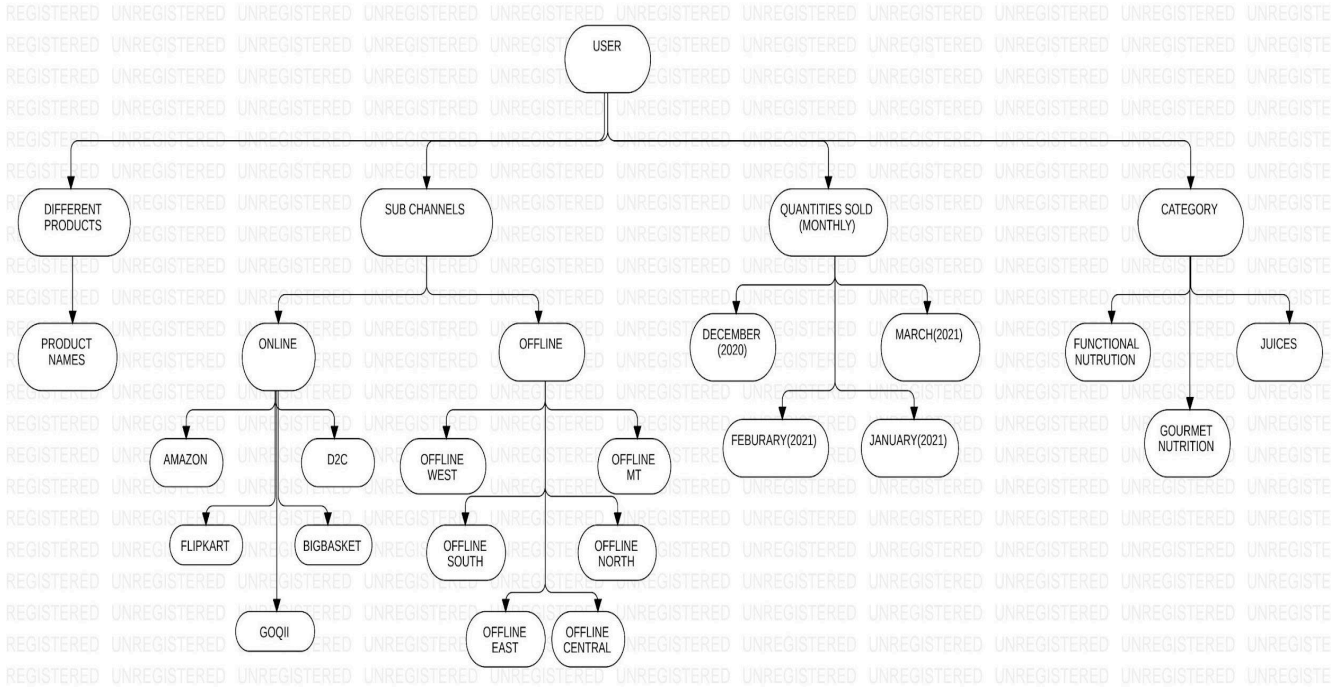
- Exploratory Data Analysis of Iris Dataset
- COVID-19 Data Analysis (Data.gov)
- Titanic: Machine Learning from Disaster
- Airbnb Price Prediction (Kaggle)
- Netflix Movies and TV Shows Analysis

2.2 Proposed Solution

The proposed solution involves conducting exploratory data analysis using Python programming, leveraging libraries like pandas and Matplotlib for visualization, and Power BI and MS Excel for visual insights and detailed analysis, respectively.By combining these tools and techniques, researchers and analysts can gain deeper insights into their datasets and make informed decisions based on the findings.

3.THEORETICAL ANALYSIS

3.1 Block Diagrams



3.2 Designing requirements

HARDWARE

The specification of the system is as follows:

- SYSTEM

Intel/Pentium/ 500 MHz Processor

- RAM

256MB

- STORAGE

500MB Hard Disk/SSD

SOFTWARE

The specification of the software is as follows:

- DEVELOPMENT

Jupyter notebook,Power bi, MS-Excel

- LANGUAGE

Python-3

- DOCUMENTATION

MS-Word, MS-PowerPoint

- OPERATING SYSTEM

Microsoft Windows XP Onwards

4.APPLICATIONS

- Exploratory Data Analysis Using Dimension Reduction
- Visualization Methods for Exploratory Data Analysis
- Exploratory Analysis of Geo-Locational Data for Accommodation Recommendation
- Exploratory Analysis of Fire Statistics using Automatic Time Series Decomposition

5.REFERENCES

- Exploratory Data Analysis Using Dimension Reduction [Tejas Nanaware , Prashant Mahajan , Ravi Chandak, Pratik Deshpande, Prof. Mahendra Patil]
- Automating Exploratory Data Analysis via Machine Learning [Tova Milo, Amit Somech]
- Visualization Methods for Exploratory Data Analysis [IEEE A.Nasser , D.Hamad , C.Sar]
- Exploratory Analysis of Geo-Locational Data - Accommodation Recommendation [M. Sumithra, A.Sai Pavithra, L.Sowmiya]
- Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means [Akhilesh Kumar Singh;Shantanu Mittal;Prashant Malhotra]
- Exploratory Data Analysis using Artificial Neural Networks by Sriram D , Kalaivani K , Ulaga Priya K , Saritha A , Sajeevram A
- Exploratory analysis of the fire statistics using automatic time series decomposition [M.M. Tatur;A.G. Ivanitskiy]

Guided By:
Prof. Varsha Choudhary.

Coordinated By:
Prof. Vishal Trivedi

Group Members:
Kashish Sharma (0827IT211058)
Manjot Singh Nagi(0827It211067)
Abhishek Pathak (0827IT211006)
Astha Jain(0827IT211020)