# Analysis of Vietnamese E-commerce datasets using python programming language

## Sub- Programming for AI

Team members (4) – 1. Devender Reddy Polapalli (23275766),

2. Nagalakshmi Guntumadugu (23280409),

3. Yasaswini Kasturi (23281294),

4. Sandhya Kukka (23342480).

Task1- selection of datasets by Devender Reddy Polapalli and Naga Lakshmi Guntumadugu,

Reason for selection- In real world mostly dealing with the business type entities such as E-Commerce data, Customer contact information and public data.

From these we have chosen Vietnamese fashion E-Commerce data to gain practical knowledge such as finding similar patterns and visualizing effectively, to perform data cleaning operations and transforming according to effective visualization.

In this project we have selected overall 6 datasets 4 datasets contain more than 1000 records, and remaining two were little amount of data less than 1000 records to perform more effectively we have chosen from Kaggle website the link is provided in the report as well as program code.

**Challenges while selecting datasets:**

- Identify appropriate media that fits your mission dream and include adequate information.
- Achieve a balance between data length and complexity for true document evaluation and visualization.
- Maintain a diverse record while specializing in Vietnamese e-commerce.

1-Vietnamese tiki products fashion accessories and backpack suit cases datasets (2) assigned for Devender Reddy Polapalli

2- Vietnamese tiki products men bags and men shoes (2) datasets were assigned for Nagalakshmi Guntumadugu

3- Vietnamese tiki products women bags dataset assigned for Yasaswini Kasturi

4- Vietnamese tiki products women shoes dataset assigned for Sandhya Kukka

Task -2 Related works done by Yasaswini Kasturi and Devender Reddy

Such as choosing the data base MongoDB or PostgreSQL

Findings- we have chosen MongoDB data base for this project it is easy to setup and it is efficient and fast for storing the data and easy for retrieving data when its needed we can transform the data when its required.

**Challenges while selecting database:**

- Choose between MongoDB and PostgreSQL based on your project needs.
- Balancing the need for flexibility (NoSQL) with a relational database structure Assessing the dissatisfaction, deliverability, and suitability of semi structured e-commerce data.

Task 3- comparison of NoSQL and Distributed databases work done by Nagalakshmi Guntumadugu and Sandhya Kukka,

Findings – these are great for handling growth and performance.

**Challenges:**

- Analysis and understanding of overall performance exchange-offs among facts kinds Demo of specific capability for e-commerce data systems.
- Analysing those facts combinations in a Python-based totally statistics pipeline

Task 4- whole team members done work with the assigned datasets such as using python programming language data cleaning and data preprocessing used python libraries NumPy and pandas, Jupyter Environment,

Data Transformation- we have transformed the data to semi structured data according to our requirements and effective visualization work done by Devender Reddy Polapalli and Naga Lakshmi.

**Challenges:**

- Fixing missing or inconsistent results across diverse datasets
- Deleting duplicate data without losing important data
- Changes have been made to convert estimates into semi-executable plans.

Task 5- we have all combined our individual code and made a structure to visualize the data by using seaborn and matplotlib libraries work done by whole team.

**Challenges:**

- Creating powerful images that truly express ideas Combining more than one data source into an integrated credential.
- Solving technical challenges includes using libraries such as Seaborn and Matplotlib to manage large datasets.

Task 6 – Final project report done by collecting all the information and their findings during the whole project from all team members and combined the work and gained our final aim of the project – By 1- Devender Reddy Polapalli, 2- Nagalakshmi Guntumadugu, 3- Yasaswini Kasturi, 4- Sandhya Kukka

Conclusion – visualizing pricing and discount patterns gave useful insights such as optimal discount timing and price segmentation to improve e-commerce strategies,

Machine learning can be used to predict pricing trends, further research on machine learning could explore utilization of advanced clustering algorithms such as DBSCAN or hierarchical clustering to purify customer segmentation.

It was a total of 30 days to complete the whole project which includes 1-2 days of dataset collection and 3-4 days to select which database has to be used and 2-3 days for analysing and comparing databases, writing findings, and understanding their scalability. 5-7 days of Cleaning, handling missing data, removing duplicates, and performing feature engineering for all datasets and 4-6 days of creating visualizations using Python libraries like Seaborn and Matplotlib, refining visuals, and integrating insights. 3-4 days of Compiling results, visualizations, and team contributions into a cohesive document.

We have provided the two types of datasets such as structured datasets and transformed data in the git hub provided link is - https://github.com/bothacker-hub/Programming-for-AI-MSCAI1.git