**SQL Capstone Project Report**

**Introduction**

Aim behind this amazon sales data analysis is to understand performance of three branches situated in three different cities and factors affecting their performance which includes the line of products they are offering and customer’s behaviour in its regard.

**Analysis Required**

This sales data required analysis on three major domains:

1. Product Analysis
2. Sales Analysis
3. Customer Analysis

**Dataset Used**

We have used **Amazon** dataset that shows sale performance of three stores for the month of January, February, and March in the cities of **Naypytaw, Yangon** and **Mandalay**. This dataset has **17 columns** and **1000 rows**.

**Tools Used**

* ***MySQL Workbench***
* ***Microsoft Excel***
* ***Microsoft Word***

**Steps we have Followed:**

1. **Data Wrangling**

Data wrangling is the process of transforming raw data into a more usable format for analytics and other purposes. Data wrangling is important because it ensures that data is reliable and complete before it is used to create insights.

* First, we have created a table in MySQL Workbench
* Then we have imported the CSV file into the table
* Lastly, we checked for any NULL values if any present in the dataset.

1. **Feature Engineering**

After Data Wrangling, the next step was Feature Engineering. Feature engineering is a data analytics technique that involves manipulating raw data to create variables that can be used to train and predict machine learning models.

In this step we have altered the table and added three new columns as follows:

* To provide information on sales in the morning, afternoon, and evening, a new column called **timeofday** was added. This will assist in determining the time of day when the greatest sales occur.
* The extracted days of the week on which the specified transaction occurred are contained in a new column called **dayname**. This will enable us to determine the busiest day of the week for each branch.
* The months of the year that the specified transaction occurred are extracted and stored in a new column called **monthname**. Determine which month of the year generates the highest profit and sales.

**Product Analysis**

* There are 6 distinct product lines namely Health and Beauty, Electronic Accessories, Home and Lifestyles, Sports and Travel, Food and Beverages and Fashion Accessories.
* The product line which has recorded highest sale and generated highest revenue is ***Food and Beverages***.
* The product line that has incurred highest Value Added Tax is also ***Food and Beverages***.
* For the Product line ***Health and Beauty***, the sales performance was **below average** as compared to other products whereas the remaining product lines have fared well.
* ***Food and Beverages*** have earned the **highest rating** among all with an average rating of **7.1** whereas ***Home and Lifestyle*** segment has the lowest rating with average rating of **6.8.**

**Sales Analysis**

* The most preferred payment method is **E-Wallet** out of three payment methods.
* Month of **January** has generated ***highest revenue*** with **February** being the ***lowest*** among the first three months of the year.
* Also, in January the Cost of Goods Sold has reached its peak.
* Among the three cities, **Naypytaw** has generated the ***highest revenue*** and **Mandalay** being the ***lowest***.
* It has been observed that **most preferred shopping time** of the day is ***Afternoon*** with Evening being the least preferred.
* Among all the cities **Yangon** has the ***highest VAT Percentage.***

**Customer Analysis**

* There are two distinct Customer types: Members and Normal.
* **Fashion Accessories** is the most associated product line with ***female*** customers whereas among the ***men*** **Health and Beauty** products were the most preferred ones.
* **Members** has the highest purchase frequency and they have also contributed to the highest overall revenue and as well as highest VAT Payment.
* Between the two genders, **Female** customers have ***higher shopping streak*** than their male counterparts.
* Branch C has recorded highest female customers whereas Branch A has recorded highest male customers.
* Most of the customers have preferred Afternoon for rating which shows high customer traffic during that portion of the day.
* **Monday** has recorded the ***highest average rating*** among all the days in the week.
* Branches A and C have recorded their highest average rating on Friday whereas for Branch B it was during Monday.

**Recommendations**

* All three branches need **better advertising strategy** for the Product Line ***Health and Beauty.***
* Ensure **adequate inventory** for the product line ***Food and Beverages*** to ensure there is no inadequacy during time of high demand.
* Introspection needed about the quality of products being offered under the product line Home and Lifestyle to improve ratings.
* Mandalay’s branch needs to maintain more inventory and provide offers to boost their sales performance.
* Branches need to know about which product lines most preferred by customers when they shop in the evening, so that they can provide end-of-the-day offers. A customer feedback or survey can help in this regard.
* Branches need to focus upon how they can bring more male customers by advertising their most preferred product line.
* Stores also need to focus on the normal customer types as their revenue share is quite less as compared to the members.