**Phase 2 Submission**

**TEAM NAME: Data Digits**

Sravani Ganipisetti - [sgani2@unh.newhaven.edu](mailto:sgani2@unh.newhaven.edu)

Divya Astalapuram – [dasta1@unh.newhaven.edu](mailto:dasta1@unh.newhaven.edu)

Bindu Priya Basa- [bbasa4@unh.newhaven.edu](mailto:bbasa4@unh.newhaven.edu)

1. **Research Question:**

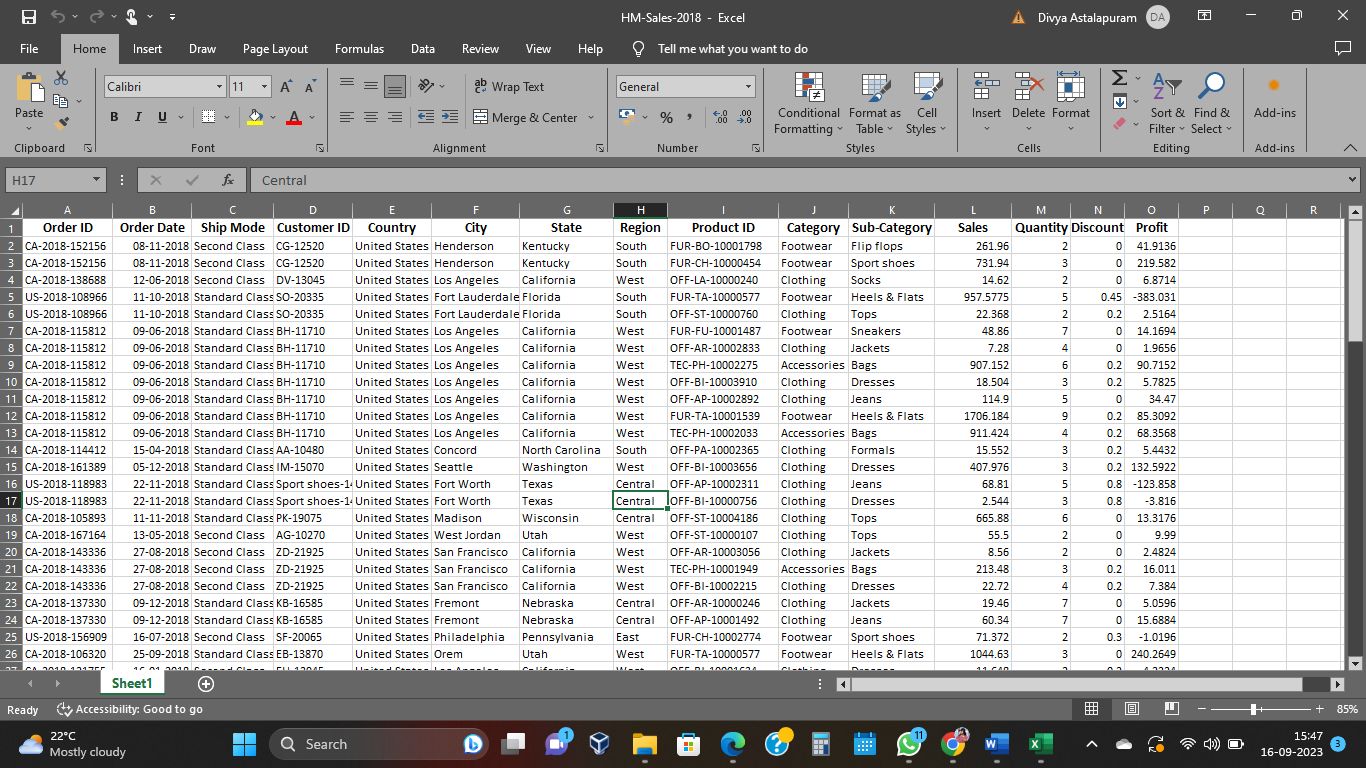
This research project digs into a thorough examination of the key factors that have had an important effect on H&M's sales growth in the modern fashion retail sector. We intend to analyze and evaluate these elements in order to get insights about the brand's growth in present fashion retail market .

1. **Support the merit of answering your research question using literature review or other valid references (min 2, max 5):**

* [**https://www.scirp.org/pdf/iim\_2023050814520088.pdf**](https://www.scirp.org/pdf/iim_2023050814520088.pdf)
* [**https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/9946478**](https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/9946478)
* [**https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/9706766**](https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/9706766)
* [**https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/7583967**](https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/7583967)

1. **In a separate section of your report describe the selected dataset that you want to work with. Your descriptions should also include the accessibility of the selected dataset, data collection methods, and information about the datatype, etc.**

The dataset we are working with is “H&M sales 2018 data”. The dataset is taken from Kaggle website (<https://www.kaggle.com/datasets/tulasiram574/hm-sales-data> ). This dataset contains 15 attributes and 100 rows. This dataset is about different products purchased from H&M company in United States and about the profits.



**Detailed description of data set**:

|  |  |  |
| --- | --- | --- |
| S.no | Column Name | Data type |
| 1 | Order Date | object |
| 2 | Ship Mode | object |
| 3 | Customer ID | object |
| 4 | Country | object |
| 5 | City | object |
| 6 | State | object |
| 7 | Region | object |
| 8 | Product ID | object |
| 9 | Category | object |
| 10 | Sub- category | object |
| 11 | Sales | float64 |
| 12 | Quantity | int64 |
| 13 | Discount | float64 |
| 14 | Profit | float64 |

1. **Please clearly describe your plan [steps, data mining model/algorithm] for solving the research question using the selected dataset.**

**Data Preparation**

We have gathered a sales data set of H&M from 2018. Which include various factors such as Category (the type of item), Quantity, Discount and Profit.

* **Regression Analysis**

Here we choose data relevantly by considering the appropriate factors in our data set which enhances the growth in sales. We divide the data into two sets one will be training set and the other would be testing set. We may use methods like Linear regression, Time series and Random Forests. Linear regression can be useful to form a relationship among the dependent variables such as category, quantity with the independent variables such as time of the categorized sales. Where both are directly proportioned. Time Series Regression depends on the time frame for which the sales raised. here we can forecast the future data depending on historical sales data we have. Whereas while using Random forests method we can combine the prediction of different decision trees and provide potential prediction. Then we will assess the models using the MSE that is Mean Square Error Method for accuracy score. the least the error value will be the best fit one.

* **Cluster Analysis**

Here we preprocess the data by considering different factors such as Store location and the categorized type of clothing has been sale and the profits according to it. In this method we can form the clusters of data by the historical datasets and predict according to it. We take data into two sets testing and training data. And chose a method like K-mean, or hierarchical clustering methods. K-mean can be useful where we consider “k” as number of sales patterns which are called as clusters. it is dependent on the historical sales pattern based on the store location and the time frame. Whereas Hierarchical pattern designs a tree like structure based on the different clusters and identifies the relationship between store location time frame and helps in predicting the sales. Using any of these methods we can identify the which region and stores exhibit similar sales behavior.

* **Time Series Analysis**

Here we identify the patterns using historical data and factors like time frame, region, the categorized type of clothes sale and the profit according to it. In this analysis everything depends on the time frame into consideration. It basically takes a particular time period for example if we consider Winter in New York city the type of clothes mostly got sale are sweatshirts and thermals so like this it categorizes the most sale items in particular time frame and predicts the future profits according to it. in this we can use a method called as LSTM Long-Short Term memory or ARIMA Auto Regressive Integrated Moving Average methods to retrieve data future sales data depending on the past sales figures and trends. We will experiment with the Mean Squared Error method to select the most effective method.

* **Association Rules**

Here we identify the Frequent patterns and relationships among datasets. we analyze how frequently a particular combination (categorized clothes) occurs in the data sets and calculate it. We also calculate i.e.; we predict that how many times can those particular combinations can be repeated at each specific time. This can be done using Mining algorithm such as Apriori algorithm to the taken dataset. It Determines the occurrence of frequent combinations in dataset. Which helps us to make the recommendations for the customers in future.