**Start with an analysis of the general properties of the MyCosmetics dataset by using appropriate visualisation techniques. Explore all transaction factors and highlight any patterns or trends.**

**• Statistical summary, e.g., descriptive statistics by timeline, gender and store.**

**• Visualisations, e.g., histograms, scatter plots, box plots, column charts, etc.**

**• Suggestion of features chosen for modelling and their relevance to the analysis needed.**

**Introduction**:

Data mining is the process of sorting and cleaning the data for better understanding of data by developing a pattern for decision-making. It involves a series of steps such as data exploration, visualization, preparation, etc. These can certainly be achieved using techniques developed for data mining tasks such as clustering, regression, association rule mining, classification and predictive analysis.

**Data Preparation:**

The crucial initial steps to data mining are processing of data intended for user’s understandability of the data.

1. **Understanding of data**, provides a clearer insight of the data to the business analyst or data analyst in order to process the information and decide on the most necessary focal points to propose the most effective and efficient decisions.
2. **Data cleaning** is defined as the validating of data by checking for missing values, duplicates and format issues. Prepare the data by completing all the steps to do quality analysis suitable for the business performance.
3. **Data selection and integration,** focuses on simplifying the tasks to speed up the analysis by eliminating unwanted and irrelevant data by being very precise. This is claimed as feature selection. Data integration combines the data together using references to be readable for the machine or tool to do its process.
4. **Data transformation** is the process of converting unstacked data to stacked by fulfilling the previous steps and selecting the appropriate model to perform data modelling. This helps in pattern identification to represent the knowledge. This is called pattern evaluation.
5. **Knowledge presentation** is the final step to Data Mining, which is the deployment of data and the process of projecting the mined data in a visual form for the users. This may result in continuous monitoring and changes, correlating to the feedback in improving the results consistently.

In ‘My Cosmetic dataset’, they apply the data processing steps to identify and analyse the customer shopping behaviour patterns and associations for data predictions.

Furthermore, exploration of this dataset might give us a deeper understanding of the patterns and elevations to improve our business. Here are some data explorations methods:

**Data Visualization:**

**Simple random sampling**:

Sampling is used for analysing large sets by selecting a small sample of data from the dataset. Here, we used the same technique for ‘My cosmetics dataset’. Identifying the population to obtain samples and simple random sampling is initiated. By determining the sample size, which should be large data to provide the reliable results. Once the sampling is done, it helps us to analyse the customers interests and spending habits.

*A screenshot of a computer

AI-generated content may be incorrect.*

**Box plot**

Box plot comprises of five number summary of data. They are : minimum, first quartile, median, third quartile and maximum. Here in the box plot of the my cosmetics dataset for each and every product listed in the dataset. The vertical line goes through the box of the median. The whiskers displays the value from each of the quartile to the least or highest plot in the graph also finds the missing values.

*A screenshot of a computer

AI-generated content may be incorrect.*

**Histogram:**

Histogram helps the data scientists/ business analysts to understand the distribution of data in intervals or bins. This helps in identifying the outliers and evaluate the quality of the data by comparisons among the data. This helps in identifying and understanding the pattern of the data and helps to make the decision accurate and quick.

*A screenshot of a computer

AI-generated content may be incorrect.*

**Data Exploration:**

**Descriptive statistics:**

Descriptive statistics is defined as a particular set of methods combined to summarize and describe the data. This is usually achieved with the use of tables, charts and key descriptive values such as the size of the dataset, measures of spread, etc. Four main types of descriptive statistics include, the measure of frequency tendency, variability or dispersion, and the measure of position.

***A screenshot of a spreadsheet

AI-generated content may be incorrect.***

**Pivot table:**

Pivot table is a business tool, powerful in summarization. It is used to sort, group, analyse and summarise data by different categories. It can generate insights from raw data promptly. Pivot tables generally display dynamic updates, such as, changes in data sources can automatically alter and reflect in the pivot table. In conclusion, pivot tables assist businesses in making data-driven decisions from turning complex datasets into more user-friendly summaries. Likewise, we have analysed information such as store locations, types of beauty products, and gender of customers.

***A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.***

**Suggest a predictive model:**

1. **Regression**- is the Analysis of the relationship between a dependant variable (the target) and the independent variable. One or more variable is the predicted values or the features. Regression is often used for prediction (of the future values of the forecast) and trend analysis (helps understand trends in large datasets). Regression also specializes on optimization which improves the efficiency of decision making. Pattern Discovery is a peculiar feature offered by regression as it detects the various relationships between variables.
2. **Association rule mining/Market basket analysis-** is a method which concentrates on understanding shopping behaviours of customers. Foundation and Lipstick are bought together at a 60% rate from the dataset provided above. Therefore, the data can be utilized to identify the products that have most of the demand, which can be led to implementation of promotional offers consisting of the most demanded pairs. The offers could be furtherly beneficial with the manipulation of limited time offers by analysing the peak hours of purchases*.*

**References:**

|  |  |
| --- | --- |
| *1* | [*www.investopedia.com/terms/d/datamining.asp*](file:///C:\Users\sathy\Downloads\www.investopedia.com\terms\d\datamining.asp) |
| *2* | [*www.geeksforgeeks.org/introduction-to-data-mining/*](file:///C:\Users\sathy\Downloads\www.geeksforgeeks.org\introduction-to-data-mining\) |
| *3* | [*www.qlik.com/us/data-analytics/data-mining*](file:///C:\Users\sathy\Downloads\www.qlik.com\us\data-analytics\data-mining) |
| *4* | [*www.qlik.com/us/data-analytics/data-mining*](file:///C:\Users\sathy\Downloads\www.qlik.com\us\data-analytics\data-mining) |
| *5* | [*www.egnyte.com/guides/life-sciences/data-sampling*](file:///C:\Users\sathy\Downloads\www.egnyte.com\guides\life-sciences\data-sampling) |
| *6* | [*www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review*](http://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review) |
| *7* | [*www.mokkup.ai/blogs/importance-of-histogram/*](http://www.mokkup.ai/blogs/importance-of-histogram/) |
| *8* | [*www.investopedia.com/terms/r/regression.asp*](http://www.investopedia.com/terms/r/regression.asp) |
| *9* | [*www.mygreatlearning.com/blog/what-is-regression/*](http://www.mygreatlearning.com/blog/what-is-regression/) |
| *10* | [*www.techtarget.com/searchcustomerexperience/definition/market-basket-analysis*](http://www.techtarget.com/searchcustomerexperience/definition/market-basket-analysis) |
| *11* | *Morgan-Kaufmann series, Jiawei-Han-Micheline-Kamber-Jian-Pei,2011* |
| *12* | *Pang-Ning T., Michael S., Anuj K. And Vipin K,2018* |