**Soal**

*Case*

**My Juice Business**

**My Juice Business** is the most popular and frequented juice store in the surrounding community. This time the juice owner wants to create data mining about his juice business by displaying various kinds of data visualizations and some analysis. The manager provides various data in csv format as follows:

* **Books.csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| CustomerID | Character | Customer unique ID. |
| CustomerName | Character | Customer Name. |
| CustomerPhone | Character | Customer Phone Number. |
| CustomerGender | Character | CustomerGender. |
| CustomerAddress | Character | CustomerAddress. |
| CustomerDOB | Character | Customer born date. |
| CustomerAge | Integer | Customer Age |

* **Juice.csv**

|  |  |  |
| --- | --- | --- |
| **uice** | **Data Type** | **Description** |
| JuiceID | Character | User unique ID. |
| JuiceName | Character | The place where the user lives. |
| JuicePrice | Integer | Age of the user. |
| JuiceTaste | Character | Taste rate of juice. |
| JuiceStock | Integer | Stock juice in warehouse. |
| JuiceYear | Integer | Year of juice release. |

* **MsRating.csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| CustomerID | Character | Customer unique ID. |
| JuiceID | Character | Juice unique ID. |
| Rating | Integer | Juice rating given by the customer. |

* **Header.csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| SalesID | Character | Transaction header unique ID. |
| SalesDate | Character | Date the transaction occurred. |
| StaffID | Character | Staff unique ID. |
| CustomerID | Character | Customer unique ID. |

* **Detail.csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| SalesID | Character | Transaction header unique ID. |
| JuiceID | Character | Juice unique ID. |
| Quantity | Integer | Quantity of the juice. |

As the data scientist in My Juice Business, you are asked to do some tasks below:

1. **Data Visualization**

To help the manager understand the data easier, you are asked to visualize the data in graph form. Some data that is needed to be visualized are:

* Show the **top** **7** most rated juice based on quantity number of users who gave a rating.

Chart, pie chart

Description automatically generated

**Figure 1. Top 7 Most Rated Juice**

* Show the frequency of juice released in 2000 - 2005.

Chart, bar chart

Description automatically generated

**Figure 2. Frequency of juice release in 2000-2005**

* Show the age distribution from the **Customer.csv** by taking data whose ages are 20-50. Make sure you have removed the user which has empty data of age.

Chart, bar chart, histogram

Description automatically generated

**Figure 3. Distribution customer age 20-50**

* Show the **juice model deployment** based on **JuiceTaste**. Separate the model into **3 categories** as below:
* **Best Juice**, where **JuiceTaste** is **A**.
* **Good Juice**, where **JuiceTaste** is **B**.
* **Unknown Juice**, where **JuiceTaste** is **neither A nor B**.

Make sure you must already remove any null data and remove juices that have JuiceTaste is NA.

Chart, pie chart

Description automatically generated

**Figure 4. Book model deployment chart**

1. **Frequent Pattern Analysis**

To help the manager understand and know what juices are frequently borrowed in each transaction, you are asked to do **frequent pattern analysis** to search for the frequent book. To get frequent juice data, use the **Header.csv** and **Detail.csv** file.

* **Data pre-processing**

In the pre-processing data phase, there are some data that can’t be used for further analysis. Do the following task to cleanse the data:

* **Remove** all transactions that have **missing values**.
* **Remove** all transactions that have **SalesDate** **less than or equal to** **2020**.
* **Remove** all transactions whose **JuiceName** is ‘**Radish’**.
* **Data transformation**

In this phase, you need to **change** **the** **data**, so it is suitable for use in the **apriori** **analysis**. Prepare the data in terms of the **juice title**.

* **Data mining**
* Show **frequent book title** using **apriori** algorithm with **minimum** support of **0.001** based on data that have already been pre-processed.

A picture containing table

Description automatically generated

**Figure 5. Frequent Juice Name using Apriori**

* Show the **association rules** using minimum confidence of **0.015** based on the frequent **juice title** that resulted from the step above.

**Text, letter

Description automatically generated**

**Figure 7. Association Rules**