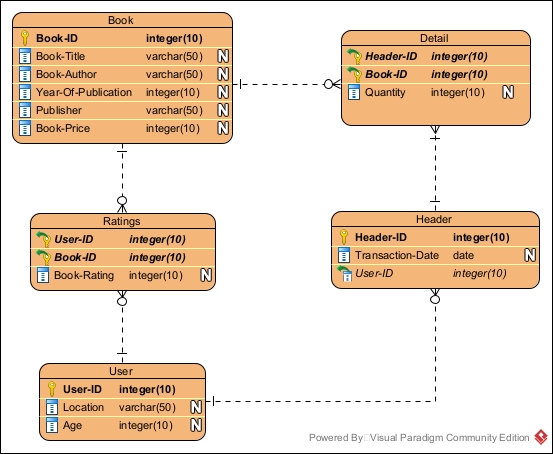
**Soal**

*Case*

**Bookstore**

**Bookstore** is a simple bookstore. It just a bookstore that has some data with ERD below:



**Figure 1. Bookstore ERD**

Below is the task you must do to analyze the data:

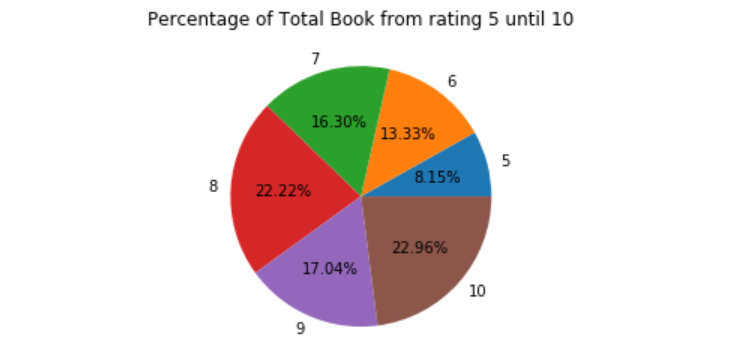
1. **Load Data from CSV to Spark**

Using **SparkSession**, **read** the following **CSV files** that have been given

1. **Query Analysis and Visualization**

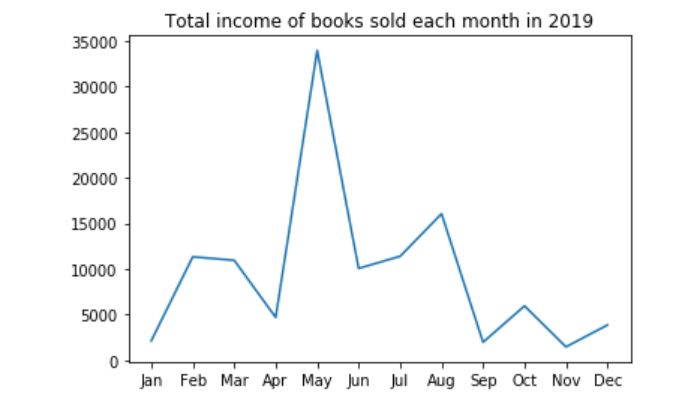
You are asked to gain some sales insight about the data. Below are some statements you need to answer. Use **SparkSQL** to answer the question and the **pyplot** package to **visualize** the answer.

1. Show the **percentage of total books** with **rating from 5 – 10.**



**Figure 2. Percentage of total books with rating from 5 - 10**

1. Show the **total income of books** **sold each month** in **2019** using a **line plot**.



**Figure 3. Total Income of books sold each month in 2019**

1. Show the **frequency of user rate** by **user category** for **every user that rate the book above 5** using a **bar plot**. The **user** **category** is determined based on the following condition.

|  |  |
| --- | --- |
| **User Category** | **Age** |
| Old | > 65 |
| Adult | 25 - 65 |
| Child | < 25 |

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

**Figure 4. Frequency of user rate**

**Classification**

Below are the steps you are required to do to generate the model.

1. **Load Data**

Given the file **BigfootTraining.csv** and **BigfootTesting.csv**, you are asked to load the data using **SparkSession**.

1. **Select Features**

After you load the data, you need to **select important features** that will be used for training. Pick **three** **important features**.

1. **Data Preprocessing**

In this step, please remove any **missing values** and **process** the data to suit the following process.

1. **Transform Data**

In this step, transform the raw data so that it is suitable for training. For example, **recode** the **gender column** value to be either **0 or** **1**.

1. **Normalization**

After data preprocessing, you are required to **normalize** the data. Use the **StandardScaler** package to normalize the data.

1. **Generate Model**

Next, you are required to **generate** a **model** from the data. Use the **LogisticRegression** package to **generate the model** with the **max iteration** of **10**.

1. **Model Testing and Evaluation**

After the model is generated, you can **test** the model to predict whether the bigfoot is prehistoric or not. Use the **BinaryClassificationEvaluator** package to print the accuracy of your model. Getthe **model** with a **minimum accuracy of ..% or higher**.

**Clustering**

Below are the steps you are required to do to generate the model:

1. **Load Data**

Given the file **Dataset.csv** and you are asked to load the data using **SparkSession**.

1. **Select Features**

After you load the data, you need to **select important features** that will be used for training.

1. **Data Preprocessing**

In this step, please remove any **missing values** and **process** the data to suit the following process.

1. **Transform Data**

In this step, transform the raw data so that it is suitable for training. For example, **recode** the **food\_colo2r** column value to be **numbered**.

1. **Normalization**

After data preprocessing, you are required to **normalize** the data. Use the **StandardScaler** package to normalize the data.

1. **Generate Model**

Next, you are required to **generate** a **model** from the data. Use the **KMeans** package to generate the model into **2 clusters**.

1. **Visualization**

After the model is generated, you can **visualize** the model using the **pyplot** package. Don't forget to add **x-label**, **y-label**, and **title** for your plot.