1. **Query + Visualization**

**GameThings** is an e-commerce platform designed for managing game-related product transactions. The system uses a relational database structure to efficiently handle customers, products, games, and purchase transactions. To get insightful information regarding the best-selling games for the developer to focus on, you are tasked with some data visualization tasks. Below are the **ERD** and the **tasks** you are required to do.

**A diagram of a diagram

AI-generated content may be incorrect.**

1. Show the **total quantity** sold **per game** where **quantity > 1**. Use **bar chart** to visualize the data.

A graph of a bar chart

AI-generated content may be incorrect.

1. Show the **total revenue** **per product** where **price > 1000** and **quantity >= 2**. Use **pie chart** to visualize the data.

A pie chart with different colored circles

AI-generated content may be incorrect.

1. Show the **total revenue** and **quantity** **per game** where **price > 500**. Use **line plot** to visualize the data.

A graph with a line

AI-generated content may be incorrect.

1. **Classification**

A doctor got tired of manually diagnosing each of his patients, so he made a model to predict them based on their medical checkup results. The results are split into 2 files named **Train.csv** and **Test.csv**. You are tasked to build a predictive model to predict whether someone has heart disease or not. Here is the description of the columns **Train.csv** and **Test.csv**:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Age | Age of the patient in years. |
| Sex | Biological sex of the patient. |
| ChestPainType | Type of chest pain experienced by the patient. Categorical feature with possible values such as: **TA**: Typical Angina. **ATA**: Atypical Angina. **NAP**: Non-Anginal Pain. ASY: Asymptomatic. |
| RestingBP | Resting blood pressure (in mm Hg) taken when the patient is at rest. A value above **120** mm Hg may indicate hypertension. |
| Cholesterol | Serum cholesterol in mg/dL. A total cholesterol level above **200** mg/dL is generally considered high. |
| FastingBS | Fasting blood sugar status (after fasting for >8 hours): **1** if fasting blood sugar > 120 mg/dL (abnormal). **0** if ≤ 120 mg/dL (normal) |
| RestingECG | Resting electrocardiogram results. Categorical feature with values such as: **Normal**. **ST** (having ST-T wave abnormality. **LVH** (showing probable or definite left ventricular hypertrophy) |
| MaxHR | Maximum heart rate achieved during exercise. |
| ExerciseAngina | Indicates whether the patient experienced angina (chest pain) during exercise: **Y** = Yes. **N** = No |
| Oldpeak | ST depression induced by exercise relative to rest. This measures the abnormality in the ST segment. A higher value may suggest ischemia. |
| ST\_Slope | Slope of the peak exercise ST segment. Categorical values typically include: **Up**: Upsloping. **Flat. Down**: Downsloping. |
| HeartDisease | Target variable indicating the presence of heart disease: **1** = Presence of heart disease **0** = Absence of heart disease |

Below are the **steps** you are required to take to generate the model:

1. **Load Data**

Given the files “**Train.csv**” and “**Test.csv**”, you are asked to load the data using **SparkSession**.

1. **Select Features**

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

1. **Data Preprocessing**

In this step, please **remove any** **missing values** in the data.

1. **Transform Data**

In this step, **transform** the raw data so that it is **suitable for training**. For example, **encode** **ChestPainType** value to be either **0, 1, 2, or so on**.

1. **Generate Model**

Next, you are required to **generate** a **model** from the data. Use the **LogisticRegression** package to generate the model with **1000** **as the** **max iteration**.

1. **Model Testing and Evaluation**

After the model is generated, you can **test** the model to predict whether the status will be canceled or not. Use the **BinaryClassificationEvaluator** package to print the accuracy of your model. Get a good enough model so patients won’t be mistreated. Their **life** depends on **you**.

1. **Clustering**

A hotel chain maintains annual records of reservation data and cancellation patterns. Based on this historical booking information, they require a **predictive model** to **classify reservations that are likely to be cancelled**. You will be given **Train.csv** and **Test.csv**, and your task is to make a **classification model** and testit to **Test.csv**. Here is the description of the columns **Train.csv** and **Test.csv**:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| ID | Unique identifier for each customer. |
| Year\_Birth | Birth year of the customer. |
| Education | Highest level of education attained by the customer. |
| Marital\_Status | Marital status of the customer. |
| Income | Annual income of the customer in the local currency. |
| Kidhome | Number of children in the household aged 12 years or younger. |
| Teenhome | Number of teenagers in the household aged between 13 and 19 years. |
| Recency | Number of days since the customer’s last purchase. Lower values indicate more recent activity. |
| MntWines | Amount spent on wine products in the last 2 years. |
| MntFruits | Amount spent on fruit products in the last 2 years. |
| MntMeatProducts | Amount spent on meat products in the last 2 years. |
| MntFishProducts | Amount spent on fish products in the last 2 years. |
| MntSweetProducts | Amount spent on sweet products in the last 2 years. |
| MntGoldProds | Amount spent on gold (luxury) products in the last 2 years. |
| NumDealsPurchases | Number of purchases made using a discount or promotional offer. |
| NumWebPurchases | Number of purchases made through the company’s website. |
| NumCatalogPurchases | Number of purchases made using a catalog (typically mail order). |
| NumStorePurchases | Number of purchases made directly in physical stores. |
| NumWebVisitsMonth | Number of visits to the company’s website in the last month. |

Below are the **steps** you are required to take to generate the model:

1. **Load Data**

Given the file “**AirBnD.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.

1. **Data Preprocessing**

In this step, please **remove any** **missing values** in the data.

1. **Transform Data**

In this step, analyze the data and **transform** the raw data so that it is **suitable for training**. For example, **convert text** data to **numerical** data to be normalized later.

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 cluster the model into **three clusters.**

1. **Visualization**

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