**Soal**

*Case*

**DezzDryth Online**

**DezzDryth** **Online** is an online game which is currently being popular in South East Asia. Since the popularity are going high, you are hired to make some analysis on the data they have. Below is the **DezzDryth** **Online ERD** that you are to analyze:

A screenshot of a social media post

Description automatically generated

**Figure 1. DezzDryth Online ERD**

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

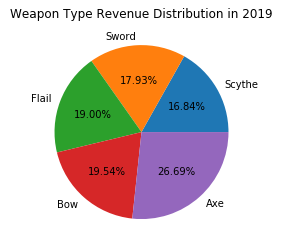
1. **Load Data from CSV to Spark**

Using **SparkSession**, **read** the following files (“MsNPC.csv”, “MsPlayer.csv”, “MsWeapon.csv”, “MsWeaponType.csv”, “TransactionHeader.csv”, “TransactionDetail.csv”).

1. **Query Analysis and Visualization**

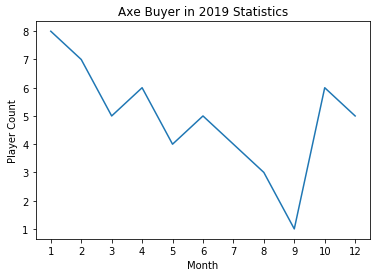
You are asked to gain some sales insight about the data. Below are some statements you need to answer. Use **SparkSQL** to answer the question and **pyplot** package to **visualize** the answer.

1. Show the **percentage of weapon type sold** for **weapon type** that is **sold** **in 2019** using **pie plot**.



**Figure 2. Weapon Type Revenue in 2019 Figure**

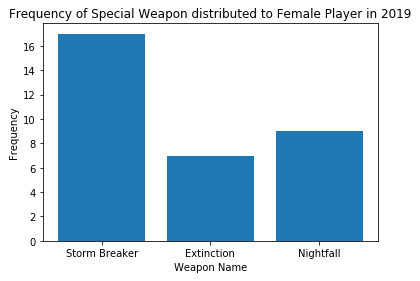
1. Show the **total number of players** who bought **axe weapon type per month in 2019** using **line plot**. **Multiple transaction** within the **same month** will be **counted as 1**.



**Figure 3. Axe Buyer in 2019 Figure**

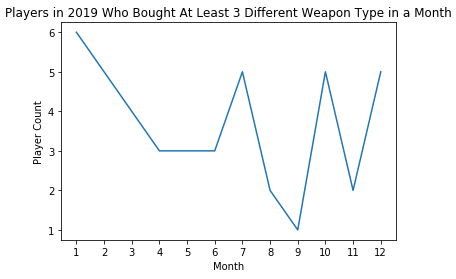
1. Show the **frequency** **of special weapon** being distributed to **female players in 2019** using **bar plot**. The player will receive special weapon if their **monthly spending** reached this following condition:

|  |  |
| --- | --- |
| **Special Weapon** | **Monthly Spending** |
| Storm Breaker | 1000000 - 1499999 |
| Nightfall | 1500000 – 1999999 |
| Extinction | > 2000000 |



**Figure 4. Frequency of Special Weapon Distribution Figure**

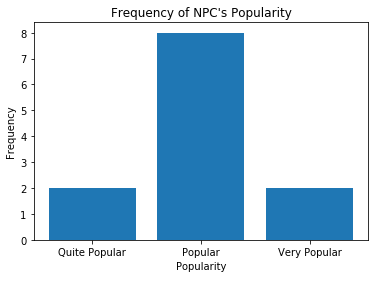
1. Show **total number of players per month in 2019** who bought **at least 3 different weapon type** **in a month** using **line plot**.



**Figure 5. Players in 2019 Statistics Figure**

1. Show the **frequency of NPC's popularity** who has done **at least transaction with 5 different player** using **bar plot**. The **popularity** is determined based on the **total income** made by the NPC as follows:

|  |  |
| --- | --- |
| **Popularity** | **Total Income** |
| Very Popular | >= 15,000,000 |
| Popular | >= 7,500,000 – 14,999,999 |
| Quite Popular | >= 2,500,000 – 7,499,999 |
| Not Popular | < 2,500,000 |



**Figure 6. Frequency of NPC's Popularity Figure**

**Grizzer**

**Grizzer** is a wild animal hunting organization that is in Indonesia. They also want to prove myths about grizzly bears, because some people argue that grizzly bears are brown in fur, while some people argue that it is not. Due to that, they decided that they want to make a **clustering model** that will have 2 cluster, **grizzly bear cluster** and **non-grizzly bear cluster**. You will be given “**BearTraining.csv”** and “**BearTesting.csv”** and here is the description of the columns:

|  |  |
| --- | --- |
| Column Name | Description |
| Weight | The bear’s weight in pounds. |
| Front Claws | The length of bear’s front claws.  Different type of bear uses their claws for  survival such as to dig roots and excavating prey. |
| Pupillary Distance | Pupillary distance measured in cm.  Pupillary distance is measured based on the distance  between the centers of eyes’ pupil. |
| Hair Color | The color of bear’s fur.  Some bears may have different type of hair color. |
| Ear Shape | The ear shape of the bear (Pointed, Broad, Rounded).  The shape may look different based on the angle  you are seeing. |
| Length | The bear’s length in feet. |

**Figure 8. BearTraining.csv**

|  |  |
| --- | --- |
| Column Name | Description |
| Weight | The bear’s weight in pounds. |
| Front Claws | The length of bear’s front claws.  Different type of bear uses their claws for  survival such as to dig roots and excavating prey. |
| Pupillary Distance | Pupillary distance measured in cm.  Pupillary distance is measured based on the distance  between the centers of eyes’ pupil. |
| Hair Color | The color of bear’s fur.  Some bears may have different type of hair color. |
| Ear Shape | The ear shape of the bear (Pointed, Broad, Rounded).  The shape may look different based on the angle  you are seeing. |
| Length | The bear’s length in feet. |
| Grizzly | Whether the bear is a grizzly bear or not (Yes, No). |

**Figure 9. BearTesting.csv**

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

1. **Load Data**

Given the file “**BearTraining.csv**” and “**BearTesting.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** in the data.

1. **Transform Data**

In this step, transform the raw data so that it is suitable for training. For example, **recode** the **‘Ear Shape’** column value to be either 0, 1, or 2.

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 cluster**.

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.

1. **Model Testing and Evaluation**

Then, you can **test** the model to check predict whether the data will be in **grizzly bear cluster** or in the **non-grizzly bear cluster**. Print the accuracy of your model and get the **model** with **minimum accuracy 80% or higher**.

**Good Luck 😊**