****

**Work Integrated Learning Programmes Division**

**M.Tech (Data Science and Engineering) Machine Learning DSECLZG565**

**Second Semester, 2022 -23**

**Assignment 1 – PS15 CSM (Conventional and Social Media Movies)**

|  |  |  |
| --- | --- | --- |
| **Name** | **Bits ID** | **Contribution** |
| M THEJASWINI | 2021SC04822 | 100% |
| ISHWAR CHANDRA TIWARI | 2021SC04556 | 100% |
|  |  | 100% |
|  |  | 100% |

**PART A**

**(5-marks) Research**

Select the research paper of your choice.

Attach the chosen paper along with the assignment submission.

Write a synopsis and find below pointers:

3. Paper Contribution

4. Data Pre-processing

5. Machine Learning Activity

6. Result analysis with metrics used from paper

7. Exploratory Data Analysis / Visualization

**PART B**

**(15 – marks) Dataset-based Implementation**

Refer to the dataset mapped against your group.

Use python based APIs and perform the following three classes of activities.

**EDA 1**. Perform Exploratory Data Analysis to gather insight from the dataset. Write your inference about the analysis learned from visualizations (minimum 3) [3]

**Steps followed for EDA**

1. Imported the data first as csm

2. Detailed structure of data was identified (231 rows and 14 columns)

3. Identified null values in the data by command csm.isnull().sum()

Movie 0

Year 0

Ratings 0

Genre 0

Gross 0

Budget 1

Screens 10

Sequel 0

Sentiment 0

Views 0

Likes 0

Dislikes 0

Comments 0

Aggregate\_Followers 35

dtype: int64

As per EDA only columns Budget, screens and Aggregate Followers have Null values screens and Aggregate followers may not influence the overall analysis and one null value of Budget could be removed.

4. Statistical analysis using the command csm.describe()

5. Interactive histogram and boxplot have also been used to identify the outliers

6. Created bar plot of individual columns to get the unique values

col\_category=['Year', 'Ratings', 'Genre', 'Sequel', 'Sentiment']

k=0

plt.figure(figsize=(200,250))

for col in col\_category:

k=k+1

plt.subplot(40, 30,k)

csm[col].value\_counts().plot(kind='bar');

plt.title(col)

7. Value counts of each unique value under respective column

8. Used line plot to find if any linear relationship exist between any columns

As per EDA No linear relation could be established between the columns.

9. No clear linear relationship could be identified between any columns.

10. Created scatter plot of individual columns to get the linear relationship or classification

x=csm['Ratings']

a=csm['Gross']

b=csm['Budget']

c=csm['Views']

d=csm['Likes']

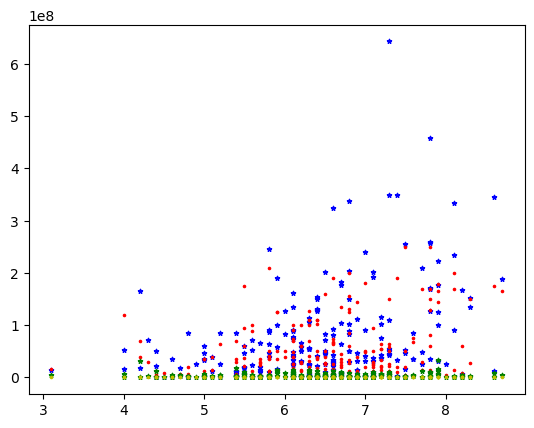
plt.scatter(x,a,marker="\*",c="b",s=10)

plt.scatter(x,b,marker=".",c="r",s=10)

plt.scatter(x,c,marker="\*",c="g",s=10)

plt.scatter(x,d,marker=".",c="y",s=10)

plt.show()



x=csm['Ratings']

a=csm['Screens']

b=csm['Sequel']

c=csm['Views']

d=csm['Likes']

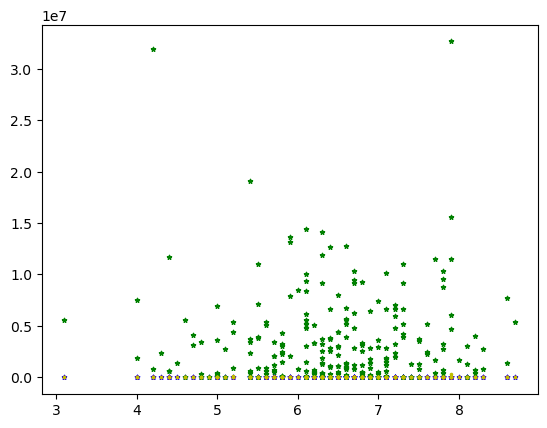
plt.scatter(x,a,marker="\*",c="b",s=10)

plt.scatter(x,b,marker=".",c="r",s=10)

plt.scatter(x,c,marker="\*",c="g",s=10)

plt.scatter(x,d,marker=".",c="y",s=10)

plt.show()



x=csm['Ratings']

a=csm['Likes']

b=csm['Dislikes']

c=csm['Comments']

d=csm['Aggregate\_Followers']

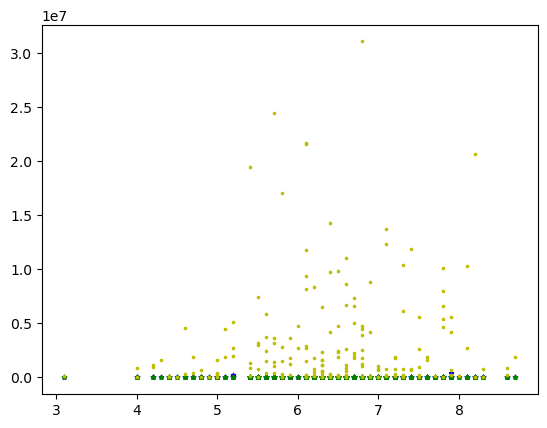
plt.scatter(x,a,marker="\*",c="b",s=10)

plt.scatter(x,b,marker=".",c="r",s=10)

plt.scatter(x,c,marker="\*",c="g",s=10)

plt.scatter(x,d,marker=".",c="y",s=10)

plt.show()



x=csm['Sentiment']

a=csm['Likes']

b=csm['Dislikes']

c=csm['Comments']

d=csm['Aggregate\_Followers']

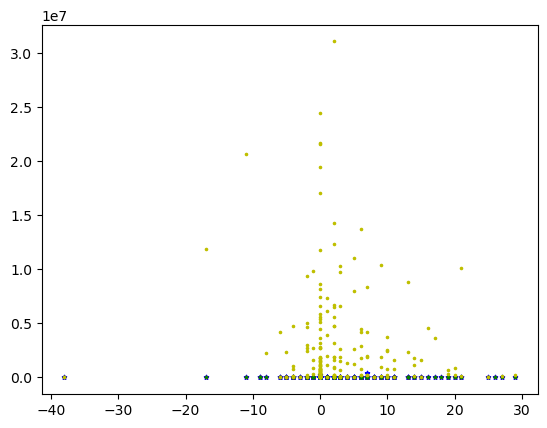
plt.scatter(x,a,marker="\*",c="b",s=10)

plt.scatter(x,b,marker=".",c="r",s=10)

plt.scatter(x,c,marker=".",c="g",s=10)

plt.scatter(x,d,marker=".",c="y",s=10)

plt.show()



x=csm['Sentiment']

a=csm['Likes']

b=csm['Dislikes']

c=csm['Ratings']

d=csm['Comments']

e=csm['Sequel']

f=csm['Genre']

g=csm['Gross']

j=csm['Budget']

plt.scatter(x,a,marker=".",c="b",s=10)

plt.scatter(x,b,marker=".",c="r",s=10)

plt.scatter(x,c,marker=".",c="g",s=10)

plt.scatter(x,d,marker=".",c="y",s=10)

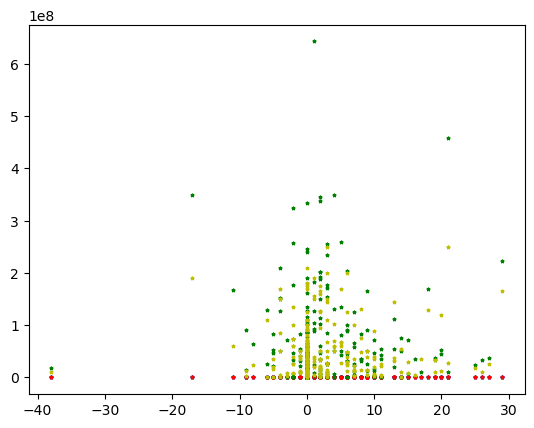
plt.scatter(x,e,marker="\*",c="b",s=5)

plt.scatter(x,f,marker="\*",c="r",s=5)

plt.scatter(x,g,marker="\*",c="g",s=5)

plt.scatter(x,j,marker="\*",c="y",s=5)

plt.show()



11. Corr analysis between various columns

|  |  |  |
| --- | --- | --- |
| Column1 | Column 2 | corr |
| Sentiment | Ratings | 0.14318226622664207 |
|  | Views | 0.06413397569902228 |
|  | Comments | 0.056372415114508315 |
|  | Sequel | -0.10676916982093186 |
|  | Budget | 0.03361570710643898 |
|  | Gross | -0.017100100951425345 |
|  | Genre | -0.017046193225347228 |
|  | Dislikes | 0.04057354790057843 |
| Budget | Gross | 0.719838644021398 |
| Ratings | Gross | 0.3422040512817663 |
| Sequel | Gross | 0.4237111617494215 |
| Likes | Gross | 0.11043168098031901 |
| Dislikes | Gross | 0.1615360528945531 |
| Views | Gross | 0.176362892508679 |

12. Based on above analysis it is decided that linear Regression for Gross analysis to be carried out

13. Classification of sentiments to be carried out based on other columns to be carried out

**Classification**. Any of the Logistic Regression / SVM / Decision Tree/ Naïve Bayes/KNN/ANN.

Justify your design choices at each step:

Write as a markdown cell in jupyter notebook at the beginning of each subsection.

**1. Perform and explain necessary pre-processing / feature engineering on this dataset [0.5]**

**2. Perform the Machine Learning activity. Explain the choice of target attribute, classification type, model selected with reason [1.5]**

**3. Quantify and explain the quality of your ML model. Explain the choice of evaluation metric [1.5]**

**4. Your observation about the results (Hint: comment on the problem statement and conclude the effectiveness of the machine learning activity) [0.5]**

**Regression**. Any of the Linear Regression (any of Gradient / Stochastic / MiniBatch)/linear basis models/KNN/Locally weighted regression/ any of the regularization techniques).

Justify your design choices at each step:

Write as a markdown cell in jupyter notebook at the beginning of each subsection.

1. Perform and explain necessary pre-processing / feature engineering on this dataset [0.5]

2. Perform the Machine Learning activity. Explain Attributes of interest, Regularization type with reason, model selected with reason [1.5]

3. Quantify and explain the quality of your ML model. Explain the choice of evaluation metric [1.5]

4. Your observation about the results (Hint: comment on the problem statement and conclude the effectiveness of the machine learning activity) [0.5]

**Ensemble ML**.

Justify your design choices at each step:

Write as a markdown cell in jupyter notebook at the beginning of each subsection.

**1. Perform and explain necessary pre-processing / feature engineering on this dataset [0.5]**

**2. Perform the Machine Learning activity. Explain Attributes of interest, base classifier chosen with reason, model selected with reason [1.5]**

**3. Quantify and explain the quality of your ML model. Explain the choice of evaluation metric [1.5]**

**4. Your observation about the results (Hint: comment on the problem statement and conclude the effectiveness of the machine learning activity) [0.5]**