milestone-documentation-0x001

Required

Explore the central tendency and dispersion of your topic's dataset:

- Import the necessary libraries and dataset
- Analyze the dataset and identify the measures of central tendency (mean,median,mode)
- Dispersion (range, quartiles, variance, standard deviation).

Explain how these measures provide insights into the dataset and what they reveal about the distribution of the data.

Execution

The first dataset contains the following features:-

- movie_id - A unique identifier for each movie.
- cast - The name of lead and supporting actors.
- crew - The name of Director, Editor, Composer, Writer etc.

The second dataset has the following features:-

- budget - The budget in which the movie was made.
- genre - The genre of the movie, Action, Comedy ,Thriller etc.
- homepage - A link to the homepage of the movie.
- id - This is infact the movie_id as in the first dataset.
- keywords - The keywords or tags related to the movie.
- original_language - The language in which the movie was made.
- original_title - The title of the movie before translation or adaptation.
- overview - A brief description of the movie.
- popularity - A numeric quantity specifying the movie popularity.
- production_companies - The production house of the movie.
- production_countries - The country in which it was produced.
- release_date - The date on which it was released.
- revenue - The worldwide revenue generated by the movie.
- runtime - The running time of the movie in minutes.
- status - "Released" or "Rumored".
- tagline - Movie's tagline.

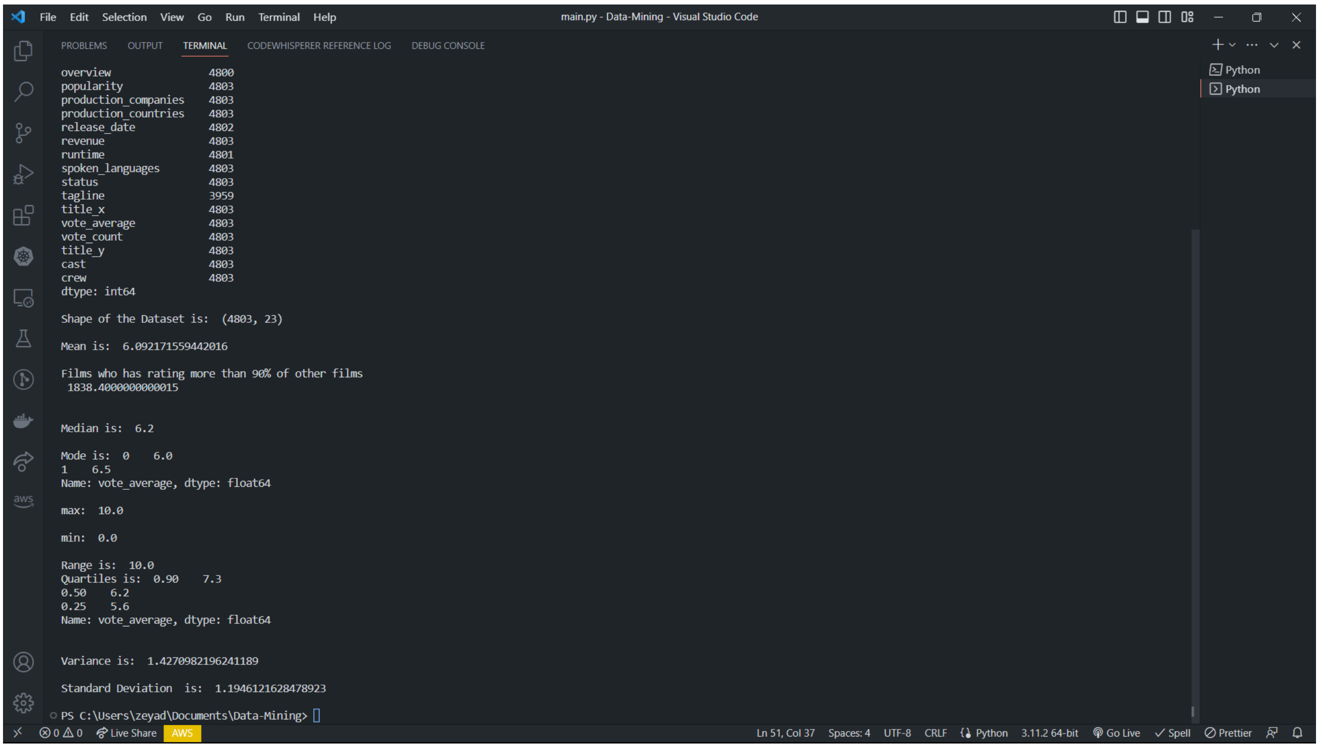
- title - Title of the movie.
- vote_average - average ratings the movie recieved.
- vote_count - the count of votes recieved.

Code

```
import pandas as pd

df1 = pd.read_csv("./DataSets/tmdb_5000_credits.csv", low_memory=False)

df2 = pd.read_csv("./DataSets/tmdb_5000_movies.csv", low_memory=False)




# merge the 2 data sets on the id column

df1.columns = ['id', 'title', 'cast', 'crew']

df2 = df2.merge(df1, on='id')

print(df2.head())




print('Count of the Dataset is: ', df2.count(), "\n")

# Shape of the Dataset is:  (4803, 23)

print('Shape of the Dataset is: ', df2.shape, "\n")

mean = df2['vote_average'].mean()

print("Mean is: ", mean, '\n')  # Mean is:  6.092171559442016




# Choosing the films that have a rating more than 90% of the others

m = df2['vote_count'].quantile(0.9)

# 1838.4000000000015
```

```
print('Films who has rating more than 90% of other films\n', m)
print("\n")


median = df2['vote_average'].median()
print("Median is: ", median, '\n')  # Median is: 6.2


mode = df2['vote_average'].mode()
# Mode is: 0    6.0    1    6.5     Name: vote_average, dtype: float64
print("Mode is: ", mode, '\n')


MaximumValue = df2['vote_average'].agg(max)
print("max: ", MaximumValue, '\n')  # max:  10.0
MinimumValue = df2['vote_average'].agg(min)
print("min: ",  MinimumValue, '\n')  # min:  0.0


range = MaximumValue - MinimumValue
print("Range is: ", range)  # Range is:  10.0



m = df2['vote_average'].quantile([0.9, 0.5, 0.25])
print('Quartiles is: ', m)
print("\n")



# Quartiles is:  0.90    7.3

# 0.50    6.2

# 0.25    5.6

# Name: vote_average, dtype: float64




variance = df2['vote_average'].var()

print('Variance is: ', variance, "\n")  # Variance is:  1.4270982196241189



std = df2['vote_average'].std()

# Standard Deviation  is:  1.1946121628478923

print('Standard Deviation  is: ', std, "\n")
```

Screenshots



```
overview                4800
popularity              4803
production_companies    4803
production_countries    4803
release_date            4802
revenue                 4803
runtime                 4801
spoken_languages        4803
status                  4803
tagline                 3959
title_x                 4803
vote_average            4803
vote_count              4803
title_y                 4803
cast                    4803
crew                    4803
dtype: int64

Shape of the Dataset is:  (4803, 23)

Mean is:  6.092171559442016

Films who has rating more than 90% of other films
 1838.4000000000015


Median is:  6.2

Mode is:  0    6.0
1    6.5
Name: vote_average, dtype: float64

max:  10.0

min:  0.0

Range is:  10.0
Quartiles is:  0.90    7.3
0.50    6.2
0.25    5.6
Name: vote_average, dtype: float64


Variance is:  1.4270982196241189

Standard Deviation  is:  1.1946121628478923

PS C:\Users\zeyad\Documents\Data-Mining> 
```