ML TEAM - WEEK 1

Hey guys, Welcome to the TechOdessey ML event conducted by team DCS. Here you will be
provided with a series of tasks that you need to complete within the given time. Ready for
dipping your toes in ml? Let's get started !!

Task 1

The first task is applying EDA(exploratory data analysis) to the dataset and answering the following questions. You are free to use any platform for python coding such as jupyter notebook, google colab, kaggle, vscode etc. Every question is provided with hints and reference materials, you are also free to use other resources to solve the questions.

Happy coding ;D

Questions:

1. i) Import the libraries (numpy, python, math)

   ii) Read the csv file

   (Hint: Use the .read_csv method.)

   Websites for reference: [Python | Read csv using pandas.read_csv() - GeeksforGeeks]

   iii) Do feature selection

(Hint: Use the .drop() the unwanted columns. Use the .drop method.)

Websites for reference:

Python | Delete rows/columns from DataFrame using Pandas.drop() - GeeksforGeeks)

iv) Remove null values

(Hint : Use .dropna() method)

Websites for reference:

Python | Pandas DataFrame.dropna() - GeeksforGeeks)

v) Set index for dataframe

(Hint: Use the .set_index() method)

Websites for reference:

Python | Pandas DataFrame.set_index() - GeeksforGeeks)

vi) Remove all the zeroes

(Hint: Use .replace() and .dropna() methods)

Websites for reference:

Python String | replace() - GeeksforGeeks)

vii) Remove the outliers

(Hint : How to Remove Outliers for Machine Learning? | by Anuganti Suresh | AnalyticsVidhya | Medium)

2. **Which are the movies with the third-lowest and third-highest budget**

(Hint: Use nsmallest and nhighest functions.)

Websites for reference:

Get n-largest values from a particular column in Pandas DataFrame - GeeksforGeeks

Get n-smallest values from a particular column in Pandas DataFrame - GeeksforGeeks

TECH DYSSEY

3. Which are the movies with the most and least earned revenue?

(Hint: Use the same methods mentioned above and use the results to find the solution.)

Websites for reference:

Get n-largest values from a particular column in Pandas DataFrame - GeeksforGeeks

4. What is the average runtime of movies in the year 2006?

(Hint: Use .mean() after extracting the movies that came out in 2006.)

Websites for reference:

Python | Numpy matrix.mean() - GeeksforGeeks

5. What is the average number of words in movie titles between the years 2000-2005?

(Hint: Filter out the movies between the years 2000-2005 first. On the title column,

use.split() method and find the sum of the words, and take the average.Use the

math.trunc method to give the solution in integer form.)

Websites for reference:

Python String | split() - GeeksforGeek

trunc() in Python - GeeksforGeeks

THANK YOU!