DataScience CodeonBytes Internship

Objective:

- To complete **Phase1** invloving Task1 and Task2 of Data Science Intership by CodeonBytes

Programmer	Date	Baseline
Shrishti Tiwari	20-Oct-2023	Ver 1.0

Requirements for Phase 1 Projects:

Task-1

- Call this public Api and create a csv dataset using python and pandas
- Api: - https://data.binance.com/api/v3/ticker/24hr

Solution:

- Please refer github link

Strategy:

- Coded python program
- Read json data from URL
- Loaded json data in pandas to create dataframe
- Used dataframe to write contents to csv file

Challenges:

- Using URL *"https://data.binance.com/api/v3/ticker/24hr"* didn't get the 24hr ticker data
- After doing some internet search used URL "https://api.binance.com/api/v3/ticker/24hr"

Task-2

- Clean the dataset replace missing values, remove outliers etc.

Dataset :- https://docs.google.com/spreadsheets/d/e/2PACX-1vTSS-TcErkXNk8KB0AlijhitwetxeHD2M3R0HJl2QPMAyFq0fxFX4PFKnzAWLDnratIz67DNL6GsZnV/pub?output=csv

Solution:

- Github Link

Strategy:

- Used Jupyter notebook
- Used python packages like pandas, numpy, matplotlib, seaborn
- Netflix csv dataset read into pandas dataframe
- Analyzed the dataframe data
- Cleaned data
    - Duplicate removal of title, country, director rows
    - Converting str object data to datetime field
- Missing Values
    - Identified missng values in director and country column
    - Replaced missing “Not Given” field value to empty str.
    - In Analysis filtered missing value
- Outlier processing
    - Analyzed column data using boxplot
    - Did IQR (InterQuantileRange) analysis on release_year field
    - Removed data below the IQR lower outer bound range
- Detailed Analysis
    - Dataframe expanded from 12 columns to 54 columns
    - Column listed_in (genre) a csv value field was exploded and then using crosstab and concat functions generated a wider dataframe
    - The data can thus be used for answering queries like “which director produced the most films by specific genre like PG and Action Adventure”
- EDA - Exploratary Data Analysis
    - Visualization of data done to generate various outcomes that can help get data insights and make informed decisions
    - Filter, value_counts, groupby, aggregation used to get meaningful observations
    - Data Visualizations using bar chart, pie chart and boxplots done for better understanding of data

Challenges:

- 1. Google doc URL at times was not available
    - Saved the netflix data and then analyzed it offline
- 2. Extent of data analysis:
    - Assumption made that purpose of dataset cleaning was to extract meaningful data and get insights so as to make informed business decisions.

Overall Requirements:-

- python 3.10 used
- python packages listed in <**req.txt**>. File available in github.