Read-Me File (Progress)

Introduction

This file serves to guide how to use the Jupter notebook for analysis.

The purpose of the notebooks is to create a predictive to gain insight into the revenue prediction for Anime. Notebooks consist of data cleaning, exploratory data analysis and data modeling with insight. Raw data was imported from various sources. The data is cleaned and organized into one data set for Exploratory Data Analysis (EDA). Exploratory Data Analysis combs the data by feature and begins to quantify all variables. EDA also serves to look for insight into the data to make informed decision about transformation that will be helpful in modeling. Modeling segement will take quantify data apply a Natural Language Processing on synopsis and set up various regression models. Models will be fitted and scored to determine best model. After model is selected, model will dissected to analyze trends for revenue prediction.

Note: all data will be present in file. Once notebook are set to run, the files may be overwritten.

Files

Data Sources

The following dataset were used:

- Kaggle Anime Dataset from Animelist: data/AnimeList.csv
- Kaggle Anime Dataset 2 from Animelist: data/anime.csv'
- Kaggle Anime Dataset 3 from Animelist: data/anime.csv.zst/anime.csv.zst
    - Required additional function read_zstd provided by Kaggle (shown Data Cleaning Notebook)
- Anime Revenue for TV shows from someanithing.com: Data/someanithing.com_rev.xlsx
- Anime Revenue for films from someanithing.com: Data/someanithing.com_film.xlsx
- Webs scrapping: Web scraping processed in supplement notebook
    - Box office for anime movies from eiren.com: importdata/eiren.csv
    - Box office for anime movies from numbers.com: importdata/numbers.csv
    - Box office for anime movies from wikipedia.com: importdata/wikipedia.csv

Cleaned Data

The data cleaned per Capstone – Data Cleaning for EDA or additional Review:

- Data set for EDA (second Juptyer Notebook): revised_data/revised_capstone_data.csv
- Data set for review (only for review. No action required for these files):
    - revised_data/leftover_data.csv

    - revised_data/leftover_data_tv_.csv
    - revised_data/leftover_data_film.csv
    - revised_data/leftover_data_film2.csv

Model data

The data cleaned and quantified per Capstone – EDA for modeling:

- Data set for modeling (Third Juptyer Notebook): data_for_model/final_df.csv

Notebooks

- Capstone – Data Cleaning : First Jupyter Notebook. Purpose of note book is to take data per *Data Source* and create output for EDA and additional files to review.
- Capstone – EDA: Second Juptyer Notebook. Purpose of notebook is to take in *revised_capstone_data.csv* and perform EDA on data. EDA includes breakdown of the features and quantify features for modeling.
- Capstone – Modeling Final: Third Juptyer Notebook. Purpose of notebook is to take in *final_df.csv* and use the data to create models. As well as deliver insight to revenue prediction.

Supplement files

- Get data1 – Dataframe for webscraping per *Data Source*. This not needed to run as data was collected.
- Get data2 – Dataframe for webscraping per *Data Source*. This not needed to run as data was collected.