Data Cleaning and Transformation

by Godswill Enaohwo | godswilleo@gmail.com | +2347035600213

Introduction

This Portfolio project touches on cleaning of a dataset and transforming it to a state were it can be easily used for analysis and visualizaion.

Portfolio Folder

The folder is contains of six files. Below are the description of the files

1. README.pdf - This is the readme file mearnt to give clarity to this portfolio project
2. data_cleaning.ipynb - This is the jupyter notebook file which contains the code of all the operations carried out in this project
3. tmbdb-movies.csv - This is the original csv file containing the raw data used for the project
4. genre_cleaned.csv – Generated from the transformation process and it is mearnt to be used for analysis involving genres
5. prodname_cleaned.csv – Generated from the transformation process and it is mearnt to be used for analysis involving production company
6. genre_production_cleaned.csv – Generated from the transformation process and it is mearnt to be used for analysis involving both genres and production companies

About the Dataset

The dataset used (tmdb_movies.csv) was downloaded from kaggle.com. The dataset has 20 columns. Below are the column titles and their properties

1. imdb_id 10856 non-null object
2. popularity 10866 non-null float64
3. budget 10866 non-null int64
4. revenue 10866 non-null int64
5. original_title 10866 non-null object
6. cast 10790 non-null object
7. homepage 2936 non-null object
8. director 10822 non-null object
9. tagline 8042 non-null object
10. keywords 9373 non-null object
11. overview 10862 non-null object
12. runtime 10866 non-null int64
13. genres 10843 non-null object
14. production_companies 9836 non-null object
15. release_date 10866 non-null object
16. vote_count 10866 non-null int64
17. vote_average 10866 non-null float64
18. release_year 10866 non-null int64
19. budget_adj 10866 non-null float64
20. revenue_adj 10866 non-null float64

Questions to Answer

The questions which were mearnt to be answered after analyzing the data plays an important role in deciding how and what kind of clean up is carried out on the dataset. In this case below are the questions to be answered after this clean up

1. What genre of movie is most produced in the last ten years?
2. What are the first two most profitable movie genres in the last ten years?
3. What range of movie runtime is more prefered by movie goers?
4. Which production company made the largest profit within the time frame covered by the dataset?
5. Which production company made more films within the time frame covered by the dataset?
6. Which genre of movie is most produced by the company with the highest profit within the time frame covered by the dataset

Findings after investigation

After going through the dataset the following issues were noted

1. Missing data in some columns
2. Remove duplicate record
3. There is no column for profit
4. Genres are moduled up together in one column seperated by "|"
5. Production_companies are also moduled up together in one column seperated by "|"

Cleaning and Transformation operation

The follow operations were carried out on the dataset in other to make it clean and ready to be analyzed so as to answer the questions stated above

1. **Missing data in some columns:** This was handled in two operations first the fields like homepage,tagline and keywords which contains over 1000 empty data were dropped from the dataset. This as well also took care of the fact that they columns were not necessary for the analysis to be carried out. On the other hand the missing data in the remaining text fields where filled with dummy data.

2. **Remove duplicate record:** Just one duplicate record was discovered and it was deleted

3. **Create a column for profit:** A "profit" column was added to the dataset since questions on profit were going to be answered.

4. **Genres are moduled up together in one column seperated by "|":** In other to accurately analyze the genres there is need to unbundle the respective genres under which each movie falls under. This was done and a new csv file was created named ***genre_cleaned.csv***. This is was done to be able to answer questions 1 and 2.

5. **Production_companies are also moduled up together in one column seperated by "|":** Similarly, there was a need unbundle all of the production_companies which partook in the creation of each film in other to be able to effectively carry out analysis and answer questions which has to do with production_companies. A new csv ***prodname_cleaned.csv*** file was created from the original dataset. This is was done to be able to answer questions 4 and 5.

6. **Genre and Production_companies analysis**: A third csv file was created from the genre file. This file is specifically created in other to be able to answer question 6 which will require the combine analsis of borth genre and production_companies.