- W excelu zmieniłem wartości gross (usunąłem przecinki miedzy tysiącami i zmieniłem na format number, aby łatwiej było przeprowadzać pozniejsza analizę, tak samo zrobiłem z czasem filmu, usunąłem min z komórek a w tytule kolumny dodałem (min) Korzystając z formuły SUBSTITUTE.
- Przegldam wartości w kolumnach, stworzyłem tabele aby łatwiej mi było przegladac dane.
- W released year dla Apollo 13 jest błędna wartość 'PG' zmieniam ją na rok premiery (1995r. z dane ze strony imdb)
- Importowałem dane do Jupyter Notebooka stworzyłem tabelkę wykorzystując pakiet pandas i dzięki niej szybko sprawdziłem ile wystepuje błędnych(pustych) wartości w rankingu. Przy importowaniu z excela pojawiła mi się jedna pusta kolumna którą usunąłem. Sprawdziłem typy danych oraz zapisałem tą tabele(po usunięciu zbędnej kulumny) jako csv i zaimportowałem ją do PostgreSQL.
- Stworzyłem tabelkę
- Przypisałem typ varchar niektórym kolumnom, aby upewnić się, że wszystkie dane zostały poprawnie zaimportowane, unikając błędów związanych z typami danych.
- Sprawdziłem, czy dane zostały prawidłowo zaimportowane, analizując ilość pustych wartości w wybranych kolumnach.
- Zmieniłem typ danych w niektórych kolumnach, aby umożliwić przeprowadzanie na nich działań matematycznych.
- Zająłem się pustymi wartościami w kolumnach:
- Dla kolumny z zarobkami (ang. revenue) imputowałem średnią wartość, aby zachować spójność danych, co pozwoliło uniknąć przerwania ciągłości rankingu. Średnia zarobków w top 1000 filmów wynosi 68 025 381.
- Dla kolumny z oceną Meta Score imputowałem dominantę, uznając ją za najbardziej reprezentatywną wartość. Dominanta wynosi 76 i występuje 32 razy.

- Dla kolumny kategoria, w której występowały puste wartości, uzupełniłem je na podstawie dwóch wspólnych gatunków: przypisałem kategorię najczęściej występującą dla filmów, które mają takie same gatunki (np. drama, crime).
- Po imputacji pustych wartości sprawdziłem, czy występują duplikaty w tabeli, uzyskując brak wyników, co oznacza, że tabela nie zawiera duplikatów.
- Stworzyłem z tego tabele imdb_cleaned.csv i zaimportowałem do Notebooka
- Zaimportowałem bibloteke seaborn i matplotlib.pyplot
- Stworzyłem histogram ‘Rozkład ocen IMDb’, dodałem do niego linie wskazujące średnia, min i max wartoscna osi x
- Stworzyłem 2 tabele w SQL plik (imdb_aktorzy.csv, imdb_gatunki.csv )
- Stworzyłem dwa wykresy typu box, oba zostały stworzone na tej samej zasadzie jednak różnią się badanymi obiektami (reżyser, aktorzy), dodałem do nich linie średniej na osi y
- Zaimportowałem imdb_gatunki.csv i rozdzieliłem każdy z wyrazów w każdym rzędzie ,aby każdy pojedynczy gatunek trafił do własnego rzędu.
- Na podstawie tego nowego df stworzyłem wykes z gatunkami na osi y.
- Wyodrębniłem kolumny aby stowrzyć wykres macierzy korelacji.
- Na koniec stworzyłem wykres ‘Przychód filmu a ocena IMDB’ i dodałem do niego palette crest aby widać było róznice w ratingu.
- Analize przeprowadziłem w PostgreSQL plik (Analiza)