CS-489 Group Project
American Movies Analysis

Members: Meghan Nulf, Sameer Mankotia, Donald Hammer, Carson Rueber

December 7, 2023

1 Introduction

With abundant data availability, our research explores the application of various SPARQL-related query tools to extract valuable insights from Wikidata. Through structured queries, our group successfully compiled a dataset encompassing essential information, including gross earnings, release dates, titles, and genres of multiple films. By employing SPARQL to generate a diverse dataset, our project aimed to determine the optimal season for releasing a new movie and identify the genre (or subgenre) with the highest revenue potential. This approach highlights how semantic technology and effective data manipulation facilitate the seamless extraction of complex information, enabling researchers to glean valuable insights for endeavors similar to our own.

2 Workflow and Implementation

A dataset was created containing information such as movie title, genre name, gross revenue, adjusted gross, gross unit, duration, publication date, publication month, and publication year for each individual movie. Secondly, a Python query tool was employed to assist in parsing the data into the correct format for use in the data analysis segment of this project. Amidst the substantial shift from utilizing DBpedia to Wikidata and adopting a different query format, a robust and user-friendly data query was developed, facilitating seamless data grouping.

The query development for Wikidata brought to light a number of challenges, the most difficult of which to overcome being the oftentimes inaccurate format of the release dates of movies. Many of the film entries on Wikidata contain multiple release dates corresponding to the region in the world they were released. For our purposes, we wanted only the first release date, which for approximately half of the film entries is simple enough to do; however, the other approximate half of film entries contain at least one release date that only lists the year. The entries, when treated as also having a month and day, are considered to be January 1st of that year, meaning that by only querying for the

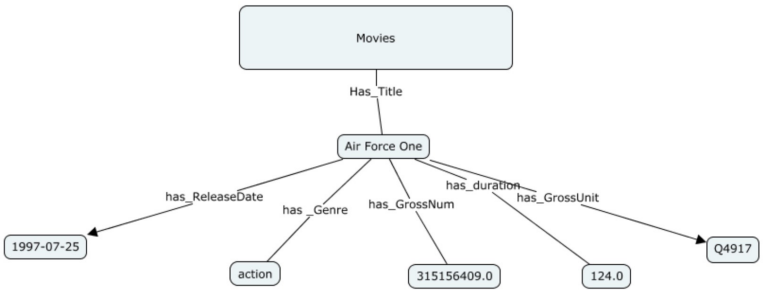


Figure 1: Query Structure Map

earliest release date, we oftentimes receive a January 1st release, which almost always inaccurate. To remove these imprecise dates from our query of the oldest release date, Wikidata provides a facility to check the precision of a date. We use this facility to filter out release date values with a precision other than '11' which, in Wikidata at least, means dates with precision to a day.

Lastly, to address the two data analysis questions independently, each analysis incorporated visualizations of the data to validate accuracy and showcase trends found in the data, aiding in predicting outcomes for these distinct inquiries. By implementing graphs such as scatter plots, pair plots, and distribution plots, patterns within the dataset were identified. Finally, through effective communication fostered in team meetings, we were able to smoothly execute each part of the group assignment.

3 Research Questions and Analysis Design

For the data research and application segment of this project, the group examined the dataset to identify information about the movies and used that information to formulate two questions. The chosen queries focused on determining which genre or combination of genres generates the highest gross revenue and identifying the season of the year (spring, summer, fall, or winter) that, when a movie is released, yields the highest revenue. To accurately achieve these goals, the use of Machine Learning models and averaging functions is used to predict these answers. For the first question, using averaging functions from pandas to predict which of these genres made the most money. This was achieved through using data preprocessing methods in pandas explodes, merge,

Then, for the second question, a random forest classifier was employed to predict, through clustering, which season would yield the highest revenue. This was achieved by using data preprocessing methods in Pandas, such as dummies, mapping, and other preprocessing techniques like TFID vectorizer and numerical mappings, to preprocess the data for each algorithmic application. The visualization for question two meets the requirements of this assignment as

it involves identifying patterns within the data to help select which methods, either encoding or weight selection, are needed to be passed into the model.

Through implementing vectorization, considering movie name relevancy is crucial for this prediction, as sequels can significantly impact revenue. Another way the data visualization for this question was relevant to the task is through using a scatter plot to reference gross revenue and month of release. Examining this scatter plot revealed that in most fall months, the revenue was higher, allowing me to statistically ensure that my model predicted correctly, as it correlates accurately back to the data visualization.

Lastly, by including two bar graphs at the end, the verification of both the predicted and actual seasons for revenue was conducted, confirming this with all the trends found in my data analysis visualizations.

4 Programming Languages/Tools/Methods

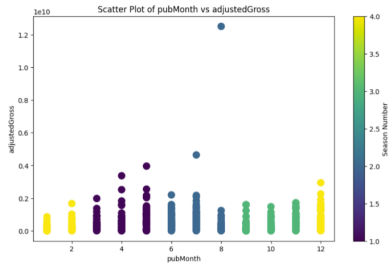
The Programming Languages, tools, and methods we used to execute this project starts with using SPARQL and used the wikidata query editor. This provided an interface have an editor with a results viewer for modifying and creating the query. Then wikidata-dl provided the ability to save all the query results into a .csv format. Lastly, for query visualization Protoge was used to make the diagram of the classes, properties, and individuals in the query. Next, for the methods and tools used for the data analyzation began with using jupyter's notebook with python. Python granted the ability to use pandas to read in the .csv file and manipulate it and its structures for the project. This resulted in the ease of data manipulation. Then, matplotlib was instituted to reveal the data in different graphical forms to help perform analysis on what method of data modeling or manipulation would be most efficient and accurate in terms of predicting.

5 Validation Steps

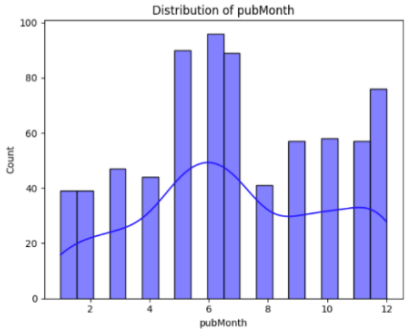
We preprocessed the data into more usable forms before doing preliminary analysis. This included graphical visualization for the overall dataset. Following this we ran more rigorous analysis to determine trends and be able to make predictions about both genre revenue and release season revenue.

6 Data Visualizations and Interpretations

The visualizations of our data were created in Matplotlib and Seaborn. A data frame named 'results' was then used to pull our data. The visualizations were created as images, each graph shows its data, the scale of its axes, and what the axes represent. In Figure 2a we show a scatter plot of publication month and gross revenue. Here each dot corresponds to a film, each dot is color-coded to a season, winter, spring, summer or fall. We can see the majority of movies

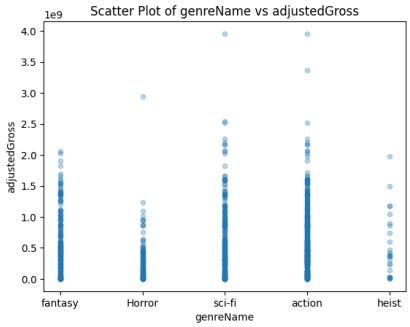


(a) Publication Month Vs Gross Revenue

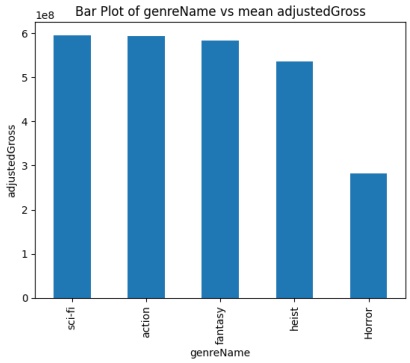


(b) Publications Per Month

Figure 2: Comparison of Figures



(a) Genre Vs Gross Revenue



(b) Average Revenue per Genre

Figure 3: Comparison of Figures

gross about the same amount, and that the movies with generally the most revenue release in spring. In Figure 2b we show the number of films released per month. This bar graph indicates that May, June, and July are the most common months to release a film, with December close behind. The months with the fewest movies released are January, February, and August. Figure 3a shows a scatter plot of genre and gross revenue. Similarly each dot corresponds to a film, but they are categorized by genre. We can see that action, fantasy, and sci-fi have many more films breaking the 1.5 billion dollar mark than heist and horror. In 3b we can see that results in them having the highest three averages for revenue.

7 Conclusion

In conclusion, our study utilized the strength of SPARQL-based query strategies to extract useful information from Wikidata with a particular emphasis on the American film business. We created an extensive dataset using structured queries that addressed important research issues on genre revenue and the best times to release movies. Our study showcased the effectiveness of semantic technologies and skilled data processing with SPARQL and Python in identifying complex trends inside extensive datasets.

The process included creating the dataset, processing the data using Python, and resolving issues with Wikidata's incorrect release dates. Research questions were answered by machine learning models and averaging functions, while visualizations such as scatter plots and pair plots helped validate accuracy and identify trends.

Seaborn, Python, pandas, matplotlib, Jupyter notebook, Wikidata-dl, and SPARQL were all part of our toolkit, which made data extraction and analysis more effective. Thorough data preprocessing, graphical displays, and in-depth analysis were all part of the validation process, which guaranteed the accuracy of our conclusions.

Visualizations created with Matplotlib and Seaborn enhanced the interpretation by showing the relationship between the month of publication and gross revenue as well as the seasonal variations in revenue. To sum up, our strategy and clear communication enabled the smooth integration of various tools, demonstrating the potential of complex data manipulation and semantic technologies for extracting useful information from large datasets in the American film industry.