Movie Data Analysis project Proposal

Zane J. Brown

Zbrown2@bellarmine.edu

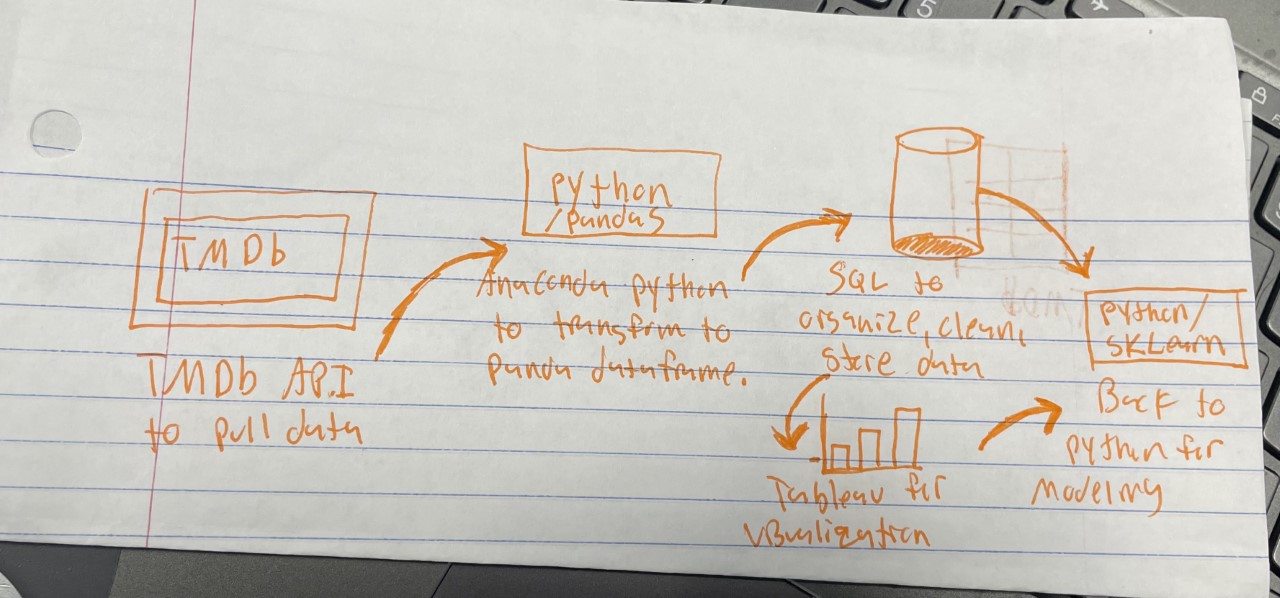
January 15, 2023

This proposal is going to go over the topic of the project, as well as how and where I am going to be getting the data from. While I am going to state the things that I originally intend on doing, things may not work out exactly how I plan as I move forward. Therefore, the data I use could be altered, I might decide to use a different data store, or I may end up changing the type of modeling. The database I am using contains data from movies as well as tv shows but I am only going to focus on the data pertaining to movies. In the proposal I am going to explain why I chose to do this; I will also explain the mapping for how I plan on performing the data analysis and implementing a model. The current potential problems and concerns I have about the project will also be briefly discussed.

For the project we have to assemble a dataset that we can do an exploratory analysis on, as well as create some type of model. To do so, we have to identify what type of data we are going to use as well as where to get the actual data from. The data used for this project has to be regularly updated, at least once a week. For my project I have decided to do my analysis using data that contains information about current movies.

I have decided to do my project with movie data because I am a big movie buff so it would be easy and fun for me to find data to work with. Also, there are a lot of different aspects that I could focus on when it comes to movies and there is a lot of different data out there for this topic. I originally wanted to do my project on ticketing data for different venues using the Ticketmaster API but I found that a lot of the data I wanted to use was not available to the public. I ultimately decided to go a different route and explore what I could find with movie sales/popularity data because there was a lot of data out there that could be used by the public.

The data for this project is going to come from The Movie Database API (TMDb). This public database contains all kinds of data on movies and tv shows, from when they came out to who was in the cast, the budget of the movie. It also includes a popularity score and an “average vote count”, the votes are coming from users that can vote for a movie or show if they like said movie or show. Streaming movies has become so popular in recent years that a lot of people don’t go to the theatres as much anymore if at all and some may not even know what movies are in theaters. People also tend to not know if a movie is good or “supposed” to be good. TMdb’s main website gives a percentage score on current and new movies in theaters and tv shows, 100 percent being really good and 0 percent being really bad. This could help people better gauge if they would want to go out and see a particular movie in theaters. At the time of this proposal, I could not figure out how to properly download the IMDb rating data or I would be also including that metric in the project. The parameters I am going to pull from the TMDb API are the following, popularity score, previous days score, average voter count, title, movie id budget, number of views, runtime, as well as genre.

All of my data collection for this project will come from TMDb API., I am going to use Anaconda Python to pull the specified data from the database. Then I am going to convert the data from Json format to Pandas so I can manipulate the data from TMDb to only pull specific lines from the data to build my own data frame. Once I have all the data I want to use in a data frame in Python, I plan on exporting it to a SQL server for my data store. This allows us to better organize and filter the data to take a closer look at smaller pieces of the data, as well as clean the data up and make sure there are no null values present in the dataset. I also want to put the data into Tableau to try to get a better visualization of the data analysis. The predictive model will be implemented using the SKLearn package from Anaconda Python once the data has been cleaned through SQL.

There are several different types of models you could create using the data in this project, I could come up with a recommendation system specifically for movies that are in theaters, I could make a model that would try to predict the popularity of movie or tv show. I am going to take a different route and make a predictive model that would try to predict the budget of a movie based on the variables mentioned before. I wanted to take this route because I have always found it intriguing how much money is poured into the production of movies and want to see if I can accurately predict the budget based on how popular the movie is and how long the movie is. I am optimistic that the model will be accurate because it is common knowledge that the longer a movie is the more it will cost to produce. What I am more worried about is that not every movie that has a high production budget is good and therefore won’t have a high popularity score. This could present some problems in the accuracy of the model.

In conclusion, this project will attempt to use an API to pull data that describes movies and measures the popularity of the movie. Using this data, I am going to do an exploratory data analysis and more specifically see if I can find anything interesting by filtering the movies by genre. The biggest obstacle for me in this project is extracting the data I need from TMDb and putting it all together into a data frame that I can work with and visualize. Once this is completed I will be able to implement a predictive analysis that is going to try to predict the budget of a movie. As I mentioned before, there will be some challenges in getting the model to be relatively accurate. I may end up having to search for other variables to include in the model to attempt to make the model more accurate if the original model does not turn out well. Overall, I am excited to see how it turns out

References:

* *API overview*. The Movie Database (TMDB). (n.d.). Retrieved January 17, 2023, from https://www.themoviedb.org/documentation/api
* IMDb.com. (n.d.). IMDb. Retrieved January 17, 2023, from https://www.imdb.com/interfaces/