Script – Draft 02

**INTRO**

(Maggie - slides 1-5)

* We believe everyone loves movies! Or at least everyone has an opinion about which movies are the best!
* While avoiding the ever-present covid topic, we thought movies were one of our prime escapes and with streaming services disrupting the industry using data on what entertainment to deliver to you, we wondered who’s using data to determine what content is made in the first place. Introducing Movie MARK, a startup movie data company seeking funding.
* The Movie MARK team is comprised of Maggie, Andrew, Rose, and Kathy and we chose to answer the question- what makes a movie successful?
* But first we must define movie success
* For our purposes, we consider a successful movie as one that is popular, is profitable, and perhaps even won notoriety through industry awards.

**CHARACTERS**

(Maggie - slides 6-9)

* Before we dive into our data, let’s meet our stakeholder/potential customers who will benefit most from our work
* Meet Ellen. Ellen is launching her own streaming service and is looking for us to help make sure she’s got the right content at the right price to draw audiences and maximize her investment capital.
* Meet Sam. He got a few promotions and is now an up-and-coming movie exec looking to launch the next blockbuster hit with the perfect mix of director, actors, and movie genre. He thinks Movie MARK® can give his studio a competitive edge.
* Last but not least, meet Steel. His lovely new bundle of joy means his time for relaxing is limited so being the movie enthusiast that he is, he’ll want to maximize time spent watching the good stuff without wasting hours searching for what to watch next.
* So, what did we learn? Spoiler alert: our proof-of-concept machine learning model was only able to evaluate our numerical variables.
* While we made an incredible effort at evaluating categorical variables, the complexity of the data prevented us from predicting a movie’s success. Let’s talk about why.

**DATA EXPLORATION**

(Rose - slides 11-13)

* To help our cast of 3 movie enthusiast stakeholders, we pulled movie data from two primary sources- IMDb and TMDB including categorical variables like Directors, Actors, Genres, numeric data like budget and revenue figures, as well as popularity data ranging from proprietary rating systems to user and critic scores out of a possible range of 10.
* We initially explored using Kaggle and Wikipedia data as well, but the initial data exploration identified too many issues with the data. Cleaning our data in preparation for analysis already took us a solid 2 weeks with the more robust IMDb and TMDB sets.
* We used a relational database hosted on Amazon web services for our database in postgres merged database with 55 categorical variables, 13 numeric variables, 2 Boolean variables and 70 potential features in total.
* Here are the technologies we used.

**DATA DATABASE**

(Rose slides)

**DATA EXPLORATION** continued

(Kathy - slides 14-16)

* Although we started out with an enormous amount of data, once we finished cleaning and evaluating the data we were left with a relatively small set of variables as features including genres, directors, duration of the movie, and release year (Can we add how many rows we started with and how many we ended up with for the slide?)
* As we cleaned and combined our data sets, we found about 30% of the data was missing including financial data around budget, revenue, and gross income
* Also, although we started out with three different types of metrics for measuring success again after cleaning and evaluation, we were left with the IMDb score which is based on a proprietary algorithm and user ratings. (budget/revenue figures are out, awards are out, ratings are all we had left)

**Data Model**

**(**Kathy slide 17)

* we tested 3 to 4 models, then all performed similarly resulting in accuracy of 90%
* More bullets from Kathy on the models coming Sunday

**DATA ANALYSIS**

(Andrew slides 18-18)

* Features included meta score and usability, genre, engagement variables (user and critic reviews, popularity, votes, database scores) (point from Andrew)
* Some of the meta score data used to determine popularity was missing but we were able to find a strong correlation with average vote scores.

**DASHBOARD**

(pivot to dashboard for interactive walk through Andrew)

(dashboard in Tableau hosted on git page)

* We chose to create our dashboard using Tableau because of its versatility, intuitiveness and style.

**SUMMARY**

(Maggie

(slides 20-24)

* Despite our data challenges, we learned a lot and were able to validate a few expected results. So where do we go next?
* Improvements would include setting up samples to build a model for predicting future success. The current approach classifies winning movies well but it’s not designed to predict future success
* While there are limitless possibilities for further research on this topic, we identified a few here, including really digging into categorical data for Directors and Actors, potentially pulling in A-list celebrity data.
* Also, we thought natural language processing of reviews could identify new features to consider.
* So for our cast of 3 characters, our proof of concept does not yet help Ellen launch her streaming service with a way to differentiate pricing for successful movies.
* Also, Movie Exec Sam will have to wait for the next iteration to predict the perfect mix of genre, Director, and stars to make a blockbuster.
* Our analysis did find, however, that Steel can focus on movies rated 7-8 to maximize his time relaxing with a great flick without spending a lot of time finding the ever-elusive top-rated films.
* Fin