Script – Draft 03

**INTRO**

(Maggie - slides 1-5)

* (opening slide)
  + We believe everyone loves movies! Or at least everyone has an opinion about which movies are the best!
* (advance to slide 2 – popcorn icon)
  + We picked movies as a topic since they have been our go-to escape from the ever-present covid topic.
  + Since we know streaming services use data modeling to help us pick what to watch next, we wondered who’s using data to determine what entertainment content to make next.
  + Introducing Movie MARK, a startup movie data company seeking funding.
* (advance to slide 3 - white page on orange background)
  + The Movie MARK team is comprised of Maggie, Andrew, Rose, and Kathy and we chose to answer the question- what makes a movie successful?
* (advance to slide 4 - black slide with white/orange text)
  + But first we must define movie success
* (advance to slide 5 - oscar award)
  + 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)

* (advance to slide 6 - Ellen Ripley)
  + 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 a streaming service and wants to make sure she’s delivering what her audiences want and maximizing her investment capital.
* (advance to slide 7 - Sam)
  + Meet Sam. He got a few promotions and is now an up-and-coming movie exec ready to launch the next blockbuster hit. He thinks Movie MARK® can give his studio a competitive edge.
* (advance to slide 8 - Steel)
  + Last but not least, meet Steel. His lovely new bundle of joy means his time is limited.
  + Being a big movie enthusiast, he wants maximize his movie watching time without wasting hours searching for what to watch next.
* (advance to slide 9 - spoiler alert!)
  + Spoiler alert: our proof-of-concept machine learning model was only able to evaluate our numerical variables around ratings.
  + The complexity of the data prevented us from predicting a movie’s success.
  + Let’s talk about why.

**DATA EXPLORATION**

(Rose - slides 10-13)

* (advance to slide 10 – orange worldwide 1970+)
  + To help our cast of 3 movie enthusiast stakeholders, we pulled movie data from two primary sources – ‘IMDb movies.csv’ and ‘movies\_metatdata.csv’ both from Kaggle.
  + The files included: categorical variables like ‘Directors’, ‘Actors’, ‘Genres’, ‘Producers’, fields with plot keywords and then there was numeric data like ‘budget’ and ‘revenue’, as well as popularity data, ranging from proprietary rating systems (metascore, for example) to user and critic – ‘ratings’ and ‘counts’ and finally there was dates and times such fields as ‘release dates’ and ‘duration’ or ‘length of film’.
  + While we initially explored using a third additional file, the ‘film awards IMDb.csv’, which listed various global awards, their award’s categories and the associated winners, but because the initial data exploration of all three files identified two common but difficult issue when working with big data; ‘variety’ and ‘veracity’ and these issues created such a large amount of scope creep, we had to make the difficult decision to eliminate the use of a third file, as cleaning our data in preparation for analysis already took us a solid 2 weeks even with the more robust IMDb and TMDB data-sets.
* (advance to slide 11 – list of data)
  + We ended up eliminating some of the fields, such as actors, writers, directors and producers, because the shear breadth of the data (the variety issue) one field could be so vast and unmanageable.
* + Additionally, some of the fields were not well populated (veracity) such as (and surprisingly) budget and revenue which were missing 55-80% over-all proving to be an inadequate field for our modeling.
* In order to make the files readily available to the public, we stored our cleaned files ‘imdb\_main.csv’ and ‘tmdb\_main.csv’ in a D3 bucket in AWS.
* We then used Jupyter Notebook to access the files on AWS and retrieve the files and do some additional manipulation of the files before we write them to a relational database hosted on Amazon web services in postgres (using pgAdmin 4).
* We merged database using sqlalchemy in jupyter notebook resulting in 47 potential features in total.

**DATA EXPLORATION** continued

(Andrew - slides 17-21 + dashboard)

* (advance to slide 17 – full list of technologies)
  + Real quick recap of other tech we used.
  + Jupyter Notebooks, obviously significant.
  + Call out: ProfileReport dependency.
  + Data hosted in a PostgreSQL on AWS.
  + Chose Tableau for versatility/plug ‘n play.
* (advance to slide 18 – release year heat map)
  + We had priors about what could predict movie success. And a few of those features jumped out at us when we began poking around the raw and clean versions of the data.
  + For instance, there seemed to be a fairly clear relationship between release year/engagement with a film, as depicted by the number of user reviews (internet, recency bias, expansion of IMDb/TMDB datasets? All plausible contributors)
* (advance to slide 19 – first scatter plot)
  + Another one – average IMDb scores and TMDB scores of films. These are weighted user scores of each film.
* (advance to slide 20 – second scatter plot)
  + That IMDb average also seemed to correspond with Metascore, a measure generated by critics’ input (similar lines as RT?)
* (advance to slide 21 – confusion matrix)
  + Similar story told by correlation matrix.
  + Not all of our variables were represented here, and you already know we didn’t end up keeping many of them.
  + But perhaps just as interesting as positive correlations were weak/negative correlations.
  + Release year – more engagement, more negativity with more new projects?
  + Fewer high-quality projects?
  + Nostalgia for old movies?
* (DASHBOARD)
* \*You can get a more granular feel for some of the data in our dashboard. 1st viz – average vote by duration/worldwide box office numbers.
  + Spider-Man search: -- differing opinions, similar stratosphere at the box (dedicated fans of the franchise?)
  + Fast & Furious (type: furious) – nice correspondence between well-liked/willingness to pay to see. Word-of-mouth factor?
  + Both franchises: similar duration/runtime – with some exceptions, around the 120-130 sweet-spot
* Genres
  + Biggest categories – drama and comedy. Interestingly, some of our modeling work suggested drama and animation were the best predictors for success. Drama/biggest piece of the pie. Animation/one of the smallest (not nearly the overlap in animation)
* Average Vote by Release year:
  + Just quickly circle back to that potential, negative correlation between release year and scores could be due to drop-off in number of recent projects – data is not updated/pandemic releases held back, etc.)
* Director Success
  + Should we mention? Ask Kathy for reminder of what these scores referred to
  + TAKEAWAY:
* There were plenty of interesting pieces to the data. But just because something was interesting didn’t necessarily mean it was viable for our prediction task.

**DATA MODEL**

**(**Kathy slides 22-32)

* (advance to slide 22 – list of data issues)
* we tested 3 to 4 models, then all performed similarly resulting in accuracy of 90%
* More bullets from Kathy on the models coming Sunday

**SUMMARY**

(Maggie slides 33 - 39)

* (advance to slide 33 – black with white/orange text)
* So in summary, having a ton of data to start doesn’t mean it’s useful.
* (advance to slide 34 – black “where to go from here”)
* Despite our data challenges, we learned a lot.
* While there’s limitless possibilities for further work on this topic, we identified a few possibilities here.
* Potential next steps include getting more data.
* We’d deep dive into the categorical variables to extract meaning in a systematic numeric way.
* Also, we thought natural language processing of reviews could identify new features to consider.
* Ultimately, our goal would be to build a model for predicting future success.
* The current approach classifies winning movies but it’s not designed to predict future success.
* That said, there’s not much we’d do differently.
* We worked great together, had a good division of labor and did the best we could with the data we had.
* Only thing we might have done differently is cloned this clever crew to double our capacity and maybe try a different topic.
* (advance to slide 35 – 3 characters)
* So for our 3 characters, our proof of concept does not yet help Ellen launch her streaming service with a differentiated pricing model for successful movies.
* (advance to slide 36 – Sam is sad)
* Also, movie Exec Sam will have to wait to use machine learning to guarantee blockbuster after blockbuster.
* (advance to slide 37 – but Steel is thrilled)
* However, we can proudly let Steel know to target dramas and animated movies rated 7-8 on IMDb if he wants to maximize his time watching movies, rather than searching for them.
* (advance to slide 38 – fin)
* Fin