Capstone Project 1 Proposal:
The Next Superhero Box Office Smash

Jean-Luc Dinello
June 13, 2019

Problem Description:
Movies adapted from comics are some of the highest grossing films out now. Out of the top 25 Highest grossing films, nearly 40% of them are comic book adaptations from Marvel or DC. Although most of the more-well known heroes have been adapted, Marvel and DC have nearly a century of IP and most of their characters have not yet made their big screen debut. Even smaller films in this genre are very expensive to make and a flop could cost a studio millions of dollars.[1]

Value to Client:
My Client is a producer that is offered to get involved in countless projects. My analysis will help to inform his decision on which projects have a higher likelihood for him to see a return on his investment. This insight into lesser-known characters could help my client to work on the next Guardians of the Galaxy, a huge success, instead of the next Jonah Hex, a financial disaster.

Datasets:
For this analysis I will be using two datasets

1. The first dataset comes from kaggle and contains metadata on over 5,000 movies. This dataset contains information about the cast, budget, revenue, and many more aspects of all of the major films about comic book characters.
2. The second dataset also comes from kaggle. This dataset is FiveThirtyEight's comic book character dataset that contains both the Marvel and DC Wiki. Each record in the dataset contains information about a specific character such as sex, gender, number of appearances, date of first appearance, and much more.

[1] https://bombreport.com/yearly-breakdowns/2010-2/jonah-hex/

Solution Methodology:

I plan on using a Randomforrest classifier to create a machine learning model that will target a high level of profit for a potential movie character.

Data Wrangling:
I will start by isolating the movies which have already been adapted from Marvel and DC characters. I then will create a profit feature for the movies based on their revenues and budget. From there I plan to drop any features that are not important to my findings such as the overview, status, and taglines. For the comic book dataset, I plan to merge the two wikis into a single dataset and also drop irrelevant features.

Data Cleaning:
Because both of these datasets come from kaggle, they are relatively clean to start out with but nonetheless I will check for any duplicates, inspect missing values, fill in some of the blanks depending on their relevance. I will also need to adjust movies revenues and budgets for inflation.

Exploration and Analysis:
I will explore the relationships between features for characters between their comic and movie adaptations such as whether the comic characters have the same physical attributes as their movie counterparts. Look at which features are most directly correlated with profits, such as critical ratings, popularity of cast members, studio etc.
I also plan to explore the statistical figures of what are the most common attributes of the characters, which film types are the most successful.

Deliverables:

Jupyter Notebook: This will show my annotated code along with visualizations that show the relationships of features and statistics regarding the datasets

Report: This will be a write up on my findings to show what insight has been achieved through my analysis. This will detail how I came to the conclusion that I have arrived at.

Tableau Data Story: This will act a presentation method to show the transformation of the data through visualizations at each stage of the project.