535 Project Draft:

Intro:

CLARIFY WHO THE AUDIENCE IS (COMIC NERD OR STATISTICIAN)

Motivate problem:

Justify Variables:

* See notes below
* Ask about how to balance out what is included this section and in the data section

Describe Dataset:

* Years of movies
* how many of each movie

Describe impact of results:

* Provides an objective quantitative approach to supplement what people think about movies
* (this depends on the results that we end up getting)
* Not going for inference as much, we are just looking to see what variables might have an affect

\*\*INCLUDE THIS…?

* This paper proceeds as follows (then have an outline of the paper)

There is a constant debate between which comic book company, DC or Marvel, produces the best line of superheroes. The raging debate that takes place primarily in the Twittersphere/ blogosphere is full of opinion-based clashes of comic book fans. Rather than add to the abundance of opinions that are already out there, we will use data about Marvel and DC movies to see if we can more objectively determine which of the comic book companies produces more popular superhero characters. We will be looking at how a movie’s studio, associated comic book company, production budget, and (Rotten Tomato critic ratings relate to its fan-popularity as measured by IMDb scores. Other data will be looked at more in depth in the following sections. The intent of this study is not focused on swaying people from one side of the debate to another, but rather to provide an objective baseline with which people can supplement their opinions. Other audiences, such as the movie making industry may also be interested in seeing what variables have an impact on the popularity of a movie. The following sections will discuss setup, and results of our linear model based approach to addressing this issue.

* (There is a continuing debate, and that debate continues into the movies. )
* Include more details about the data (number of movies, how many of each comic, time frame,etc)

Source of Data for all Variables:

There are 64 observations. Each observation is a movie, and there are 4 variables for each observation. The variables are: Studio, Comic, Tomato, Budget, and IMDb. IMDb is the response variable, which measures Fan Popularity, and is recorded on a scale of 1-10.

IMDb - Fan popularity; measured on a scale of 1-10;

Budget - Production Budget; measured in millions of dollars; determines amount and quality of visual effect.

Tomato – Professional Critics Rating; measured on a scale of 1-100 (professional critics => more likely to be objective and not biased)

Studio – (Categorical); there are 8 different studios in the dataset

Comic – (Categorical); Declares

The data used in this study were either common knowledge (e.g. which movies are associated with which comic book companies), or gathered from internet sources. The variables contained in the final dataset are as follows:

Response Variable:

* IMDb – A continuous variable of fan-popularity measured on a scale of 1-10

Explanatory Variables:

* Budget – A continuous variable to indicate the production budget of each film listed in millions of dollars
* Tomato – A continuous variable of Professional Critics Rating from Rotten Tomates measured on a scale of 1-100
* Studio – A categorical variable of which studio produced the movie. There are 8 different studios in the dataset including: Warner Brothers, 20th Century Fox, Buena Vista, Sony, Lionsgate, Paramount, New Line, and Universal.
* Comic – A categorical variable of which comic book company the movie characters are from.

(WHY WE CHOSE THESE VARIABLES)

Since we are interested in the popularity of each company’s superhero characters, we figured that IMDb scores, which are given by fans) would be a good measure of popularity. For the explanatory variables, we chose to include the production budget because that dictates the amount of quality action scenes and special effects that can be in a movie, which would likely influence how it is received by fans. We also chose to include Rotten Tomato ratings (which are provided by professionals) since they are less prone to bias than fan base ratings. These professional ratings provide an objective quantification of the individual movies. As for the two categorical variables that we included, the comic book company is essential since it is foundational in our question of interest, and we figured that since in some cases the two comic book companies have non-exclusive rights with different movie studios we should have a variable that will allow us to adjust for the difference between studios. By referencing superheronation.com, the-numbers.com, and imdb.com, we gathered and compiled the data as described above for a total of 64 superhero movies, using the statistical software R, with the help of the XML package.

(describe the variables as what they control)

Data Diagnostics:

(look at good and bad influential points)

(model assumptions)

(summarize don’t list what we do)

(maybe add excess tables to an appendix at the end.)

Summary Statistics:

> summary(dat$IMDb)

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.300 5.775 6.950 6.623 7.500 9.000

> summary(dat$Tomato)

Min. 1st Qu. Median Mean 3rd Qu. Max.

8.00 30.00 66.00 59.09 84.25 94.00

> summary(dat$Budget)

Min. 1st Qu. Median Mean 3rd Qu. Max.

17.0 72.5 137.8 136.0 185.2 330.6

> summary(dat$Studio)

BV Fox Lions NL Par. Sony Uni. WB

10 14 3 3 4 9 3 18

> summary(dat$Comic)

DC Marvel

21 43

Data Procedures used in provided code:

SOURCES:

XML Package

R software

superheronation.com

the-numbers.com

imdb.com