Cinema-Income Correlation in the United States

TEAM NAME: Cinema Cinammons
TEAM MEMBERS: Albert P, Manuel C, Qiyue W, Shannon C, Angelo S

DATA SCIENCE QUESTION(S) & HYPOTHESIS:

General Question
How has income affected the box office and ratings of specific movie genres?

Focused Question
How has per-home income in the United States affected the box office earnings and the ratings of movies of specific genres?

RESEARCH QUESTION:

The fluctuations of economy seems to affect every aspect of our lives. Such as, where we live, where we eat, or even what movies we see. When median household incomes are not thriving, we tend to be more cautious of our money. Individuals become less willing to spend money on non-necessities in order to get by. We believe that during economic troubles most of an individual's household income is used to pay for necessities such as mortgage payments and possibly to save up for an emergency that can arise such as being unemployed. Therefore, individuals will be more hesitant to spend money on things such as watching movies. We believe that movie's revenue value during times of economic troubles will not be outstanding due to the fact that individuals are not watching movies in a constant basis.

In addition, if and when individuals decide to go see a movie, those individuals will be more selective in the movie they see. They might decide to watch a movie based on the actors in the movie or simply the genre of the movie. We believe that during economic troubles, certain genres of movies will prosper over others because people will tend to be more selective and narrow down the movies they decide to watch. We will use data from GDP and from movie collected data to test our beliefs.

BACKGROUND:

Since the year 2000, the domestic economy in the United States has experienced many ups and downs. Historically, people have been forced to become more innovative to improve their quality of life in times of scarcity, such as during the Great Depression or the 2008 recession. In that regard, we hypothesize that people who watch movies in theaters when the economy is not doing well will be looking for an escape from their daily lives and therefore give movies better ratings than in other years. In addition, we might find a correlation with box office values with median household incomes. We hypothesize that box office values will be negatively correlated with median household incomes.

Initially, we hypothesized that during years when the economy is bad and household incomes are lower than usual, moviegoers will rate movies being played in theaters higher than in other years. However, we can take this hypothesis a step further by also asking whether moviegoers will prefer to see specific genres of movies during years when household incomes are low in comparison to genres preferred in years when household incomes are higher. During economic falls, individuals usually will have a lower household median income for their expenses and leisure activities. During such times individuals tend to spend less money on leisure activities, such as going to the theaters. We assume that individuals will be more selective with the movies they watch during economic troubles. This claim is an unsupported one and possibly biased.

The aim of our project is finding trends in ratings and box office earnings of movies and movie genres. Therefore, we need data with information about household incomes; this would be easiest to accomplish with data from the United States using the US Census Bureau. We also need to collect information about genres of movies and box office revenues of movies in and the year they aired.

Hypothesis

When the economy is not doing well (people have relatively low income), we predict that box office/ratings for movies, in general, will be higher. In addition, when in years where incomes are relatively low compared to other years, moviegoers will gravitate toward specific genres of movies. On the other hand, when the economy is doing well, moviegoers will prefer other genres of movies.

DATA:

1. https://www.kaggle.com/PromptCloudHQ/imdb-data/version/1

This link includes a dataset collected from *IMDB*. It includes a list of popular movies from 2006 to 2016 with their respective revenue, a ranking, runtime of the movie, name of actors in the movie, etc. I think this data set can help us find correlations on what makes a movie "popular." With this dataset, we can make correlation tests such as revenue and runtime of the movie or name of actors in the movie with ranking. We can also then make correlations of the years in which selected movies were released and compare them with a different dataset including median household income.

2. https://www.kaggle.com/muonneutrino/us-census-demographic-data/version/3

This link includes data collected through the American community survey data, an ongoing survey of the US Census Bureau. This dataset includes the demographics of several counties throughout the country. It includes total population count, race percentages, occupation information, etc. Most importantly, it contains median household income which will help our group make connections with ratings of movies and median household income. Note: Link includes American Community Service data of years 2015 and 2017.

3. Table H-8 from https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html

Table H-8 from this link is a xls format table of median household income by state in the unit of 2017 adjusted dollars. It includes data from years 1984 to 2017. As indicated by the URL, this data is directly from the US Census Bureau.

4. https://query.data.world/s/h4ksbq4lniukcgbvh25o5m2akczf3v

This dataset includes more than 5000 movies from as early as the 1910s. Not all of them are American movies. It also includes 27 attributes of the movies, including genres, years, IMDB score, and gross revenues.

5. https://query.data.world/s/gw6ftdnaltg55ldak7f5y45bj4hytx

Seasonally adjusted GDP from 1947 to 2016

Data Wrangling:

We set out to find a dataset containing movie-data from 1984 to 2014. We believe that this timeframe encompasses a variety of economic fluctuation such as the period of Reaganomics and the 2008 recession that will allow us to find a distinction between movies watched during outstanding economic times and economic struggles. More specifically, it will help us distinguish genres, revenue values, and possibly ratings of movies. Because of the criteria we set, we had to find a dataset containing average house-hold incomes ranging from 1984 to 2014.

Data Cleaning:

(PERTAINING TO MOVIE DATA ONLY).
Previously, we had placed upon ourselves a range of years of movies we will be analyzing, 1984-2014. The data-set we found encompassed a wide variety of dates starting from the 1920s. In order for us to adequately address our research question, we eliminated all rows for which movies did not fall within the 1984-2014 range. In addition, of those movies that had any empty cell, we decided to drop those cells as they will provide no value to our analysis. We encountered a couple of missing rows, but not several in which it might be problematic to our findings. A reason why we decided to stop our analysis at 2014 and not later because later years had more empty rows. We looked for outliers in our movie dataset both manually and programmatically and found our values to be accurate with what we expected.

VIOLATION OF ACADEMIC INTEGRITY [CLICK HERE]
http://rstudio-pubs-static.s3.amazonaws.com/342210_7c8d57cfdd784cf58dc077d3eb7a2ca3.html

ETHICAL CONSIDERATIONS:

Data Collection

The data being collected involves the income of individuals in America as well as public information about films' performance in the public eye. There is no abuse of consent regarding the income data collected was via a survey of the US Census Bureau. Given that the information was collected via surveying, the data may have biases based on voluntary input and location which we hope to dispel by choosing a wide yet focused group of individuals with which we hope to answer our question.

Data Storage

Most of the data we gathered so far all come from the site called Kaggle, which is built for people to do their data science projects. It fits our purpose perfectly. The movie dataset we found has information about 1000 movies. Most of the movies are from the year 2014 and after, but from the year 2007 and later, each year has data of at least 51 movies, which we think is enough for our analysis. Our household income data all come directly or indirectly from the US Census Bureau, and they cover the year range of the movie data we get. The data we will use throughout the development of this project has been publicly collected data so we are not concerned with security issues such as disclosing private information nor are we planning to obscure any data after reaching our conclusions.

Analysis

As already mentioned, the census data we have found may introduce bias into our exploratory analysis. We hope to avoid reaching incomplete conclusions by sampling from varied information (ex: different locations, income, ethnicities, etc) from the large dataset.

Luckily, the current datasets that we have found appear to have no missing data so that shouldn't manifest any misrepresentation issues. However, if we find datasets with missing data then we will undoubtedly clean up the dataset and scrape off the properties with missing values so as to have a (mostly) complete set to work with.

Using these strategies, we hope to produce a complete and honest answer to our question(s).

Other Key Points

Is it a well-posed question?

It is generally a well-posed question because we are looking at data over time. If we are only comparing data from one year, the taste in movies from households with different income might differ only because of their economic status in society. Comparing income change overtime let us focus on the overall economic situation of the US and how it affects people's preference for movies.

Could the information you will gain and/or the tools you are building be co-opted for nefarious purposes?

The nature of our question does not immediately invite the possibility of misuse of our data and our potential conclusions. In the case that we find a positive correlation between income and film ratings, the only real "nefarious" use of our conclusion(s) would be to use the findings to plan a successful film release which - as far as we know - is very unlikely in our group