IMDB Movie Data
Final Project Report
CIS 9655

Tanaya Nandanwar
Raphael Okochu
Archana Kishore
Alex Dorsinville

Preface

When brainstorming ideas and looking around for datasets to use for our project, we wanted to make sure we hit several key points. We wanted to base the project on a topic or subject matter that would be easy to understand for any audience at face value, without having to spend time on any extra explanation. We also wanted to make sure that the data set for whichever subject we choose would be easy to read, comprehend, and work with. To meet these two points we decided to work with film data from IMDB. Almost everyone watches movies or is at least somewhat acquainted with them. This makes film an easy topic to discuss and it holds some relevance with just about any audience. Additionally, movie data, and our data in particular, is was very easy to go through and understand. This made the process of constructing meaningful questions and building different visualizations more straightforward.

Data and Data Preparation

Data Attribute	Type
Rank	Interval
Title	Nominal
Genre	Nominal
Description	Nominal
Director	Nominal
Actors	Nominal
Year	Interval
Runtime (Minutes)	Ratio
Rating	Interval
Votes	Ratio
Revenue (Millions)	Ratio
Metascore	Interval

The movie data is in csv format and pulled the top 1000 films from IMDB between the years of 2006 and 2016. The vast majority of the attributes are simple to understand. Rank refers to the films placement according to IMDB's own ranking system. Rating refers to the out of 10 score given to a film by users and votes count the amount of individual scores given by users to that film. Finally, metascore is the out of 100 aggregate score that the film holds on metacritic. It is essentially a gauge of the critic score.

When working with the data we ran into several challenges. First the genre and actor fields were non atomic. Most movies belong to more than one genre and all of them have more than one

actor. Those had to be split up for analysis. Some fields such as metascore were also blank for some films, and those were filled using python.

Goals and Questions

After looking over the data and making sure it was cleaned and primed for use, we pinpointed several questions of interest to use as a guide as we began to figure out what visualizations to create. Right from the beginning when looking at all of the attributes of our data, we knew it would be important to visualize any possible correlations between movie critic rating (metascore), user rating (rating), revenue, rank, and length. Showing which attributes had the most pull on both user and critic scores was crucial. On top on this we also wanted to make use of some of the more categorical data and show which which movie genres and directors have had the most influence over the span of the data. Finally we also brought in a second source of data relating to words spoken per character. Using this we planned to analyze the participation of different actors/characters in different films.

Our five primary questions consisted of the following:

Does movie length have any effect on popularity or critic score (rating or metascore)?

What genres have the most frequently produced movies and how many movies were made each year for a genre?

Do fans generally rate movies similarly to critics (rating vs metascore)?

What was the average revenue generated for each genre in a year?

Who are the top revenue generating Directors?

Design and Key Visualizations

Question 1. *Who directed the movies with the most revenue (2006 - 2016)*

The film industry over a span of ten years from 2006 -2016 has successfully generated revenue of over $72.337 Billion. To understand the trends in revenue generation and to provide the user with deeper dive into the trends of revenue generation, we chose to calculate the revenue of

those top ten directors who have made most number of movies, from our data source of 1000 Movies.

Rationale behind chosen visualization A lot of devising went into choosing the most appropriate visualization to show this minimalistic data. Though chart junk and minimal data ink were a few parameters considered to drive the decision, bar chat was the best choice, as it showed clear trends in the difference of each of the top ten directors (unlike the triangle and circle charts which were being considered.)

Design This chart shows Directors (Categorical Data) against the Revenue (Continuous Data) generated by each of them. Calculation of the Revenue involved spanning of various rows which belonged to each of the Directors listed on the X-Axis and performing a running total operation for each of these directors using the numpy library operators.

Shneiderman's key information seeking tasks achieved

Zoom The wheel Zoom has been implemented here, as it is the most appropriate zoom options which would work with a bar chart, where comfortable zoom in and out can be done depending upon the height of each bar.

Extract An option to save a copy of the visualization(image format) for further reference has also been provided.

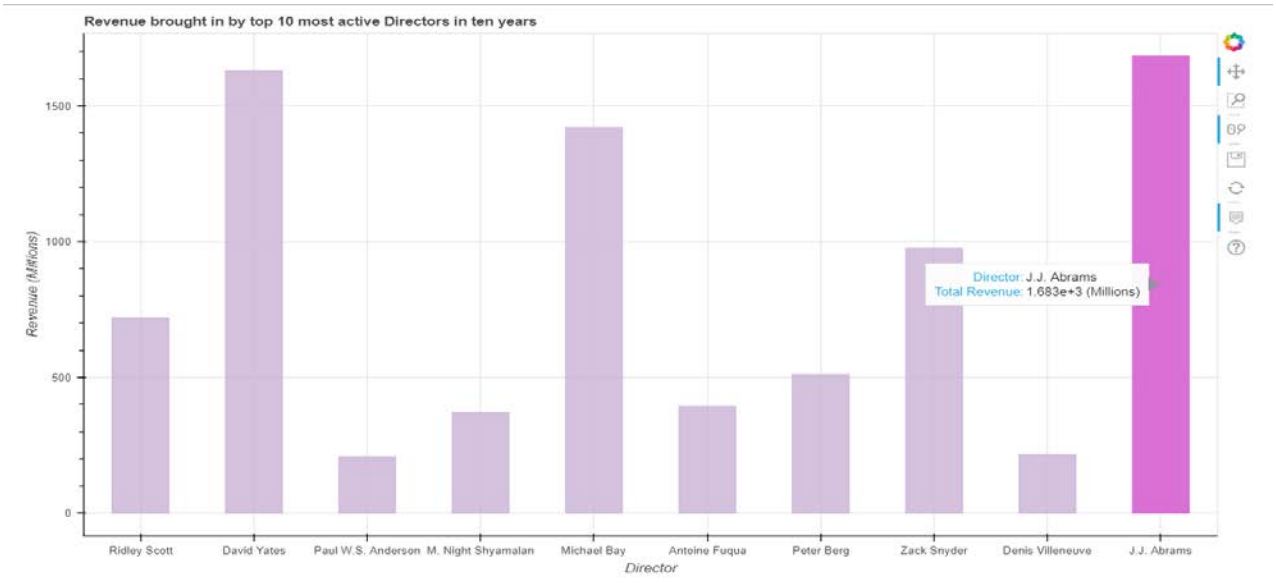

Figure 1)

Question 2 *Top ten ranked movies – Linked rectangle, circle and triangle chart.*

The dataset contains an attribute Rank which is correlated to most of the other attributes like metascore and ratings. The charts below show a comparison metascore, rating and revenue against the rank given to each of the top 10 movies.

Rationale behind chosen visualization Our focus was to provide the user with an easy access to a few important attributes and their comparative positioning with rank. So, we decided on utilizing the power of linking and brushing with bokeh, which facilitates selected comparison between visualizations.

This chart shows Rank (Categorical Data) against the Rating, Revenue and Metascore (Continuous Raw Data) generated by each of them.

Linked Panning Helps user to link pan or zooming actions across ratings, metascore and revenue.

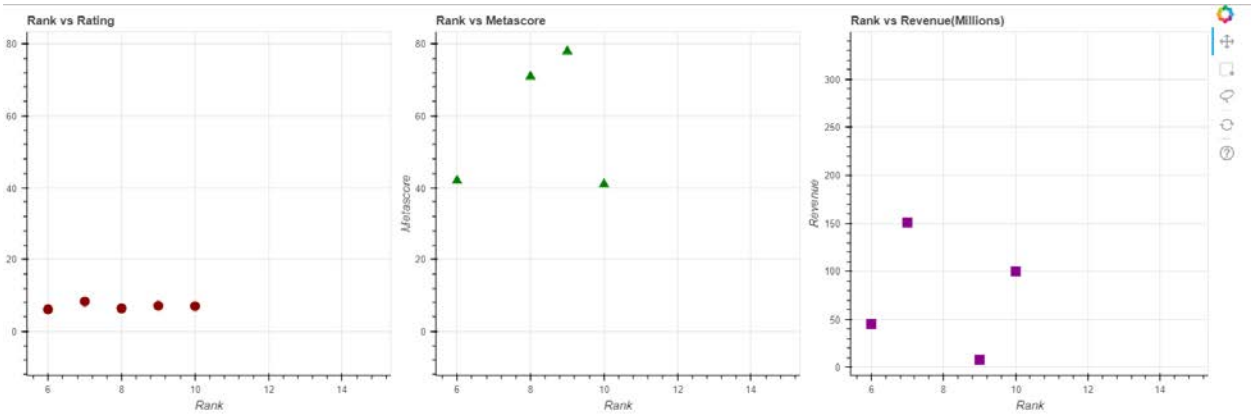

Figure 2(i)

Linked Brushing User can choose specific ranks and view the details of just those on demand across all the plots.

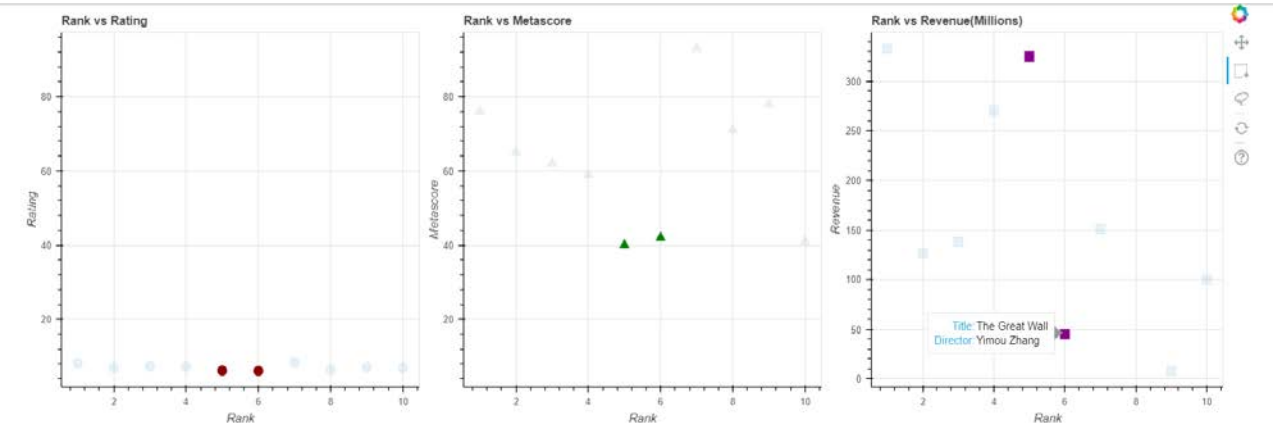

Figure 2(ii)

Question 3. *What genres have the most frequently produced movies and how many movies were made each year for a genre?*

To understand which genre movies were most frequently produced and to see if this trend has remained constant or changed over years, we decided to look into an overview of data for different genres over the span of 11 years from 2006-2016.

Rationale behind chosen visualization A stacked bar graph, is a graph that is used to break down and compare parts of a whole. Here we want an overview of the number of movies produced for different genres over 11 years span as well as we wish to distinguish between this count in each year for different genres so that a comparison over the years and genres is possible. A stacked bar chart serves our purpose.

Design To answer this question we plotted a horizontally stacked bar chart using 2 categorical data (Genre, Year) and 1 numerical data obtained by aggregating their counts from the data source. The number of movies produced for each genre, in each year is displayed. We can easily compare the number of movies produced in different years for different genres. Using the box-selection, we can chose to display the data only for the years we want. There is an option to use 'Box Zoom' tool to zoom into the details of chart which aren't clearly visible(e.g. Data for 'War' genre). The 'Save' tool can be used to save the graph for future reference. 'Pan' tool is also a useful utility. The 'Reset' tool can help bring the figure to default position.

Shneiderman's key information seeking tasks achieved

Overview, Zoom, Filter, Details-on-Demand, History, Extract

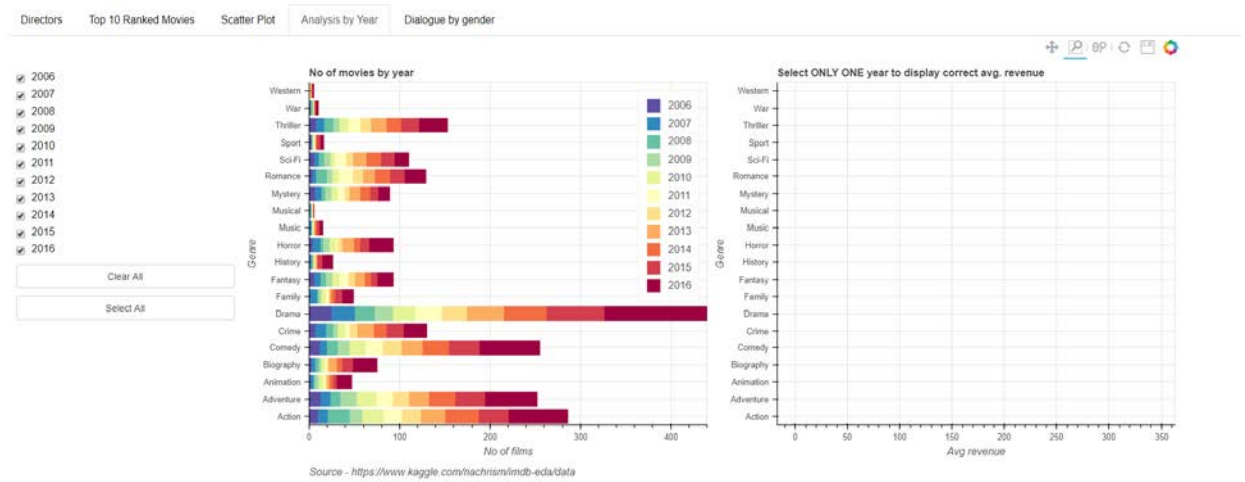

Figure 2

Question 4. *What was the average revenue generated for each genre in a year?*

The accompanying box selector as seen in Figure 3(i), makes it easy to select a year, and filter out the data for that particular year. We can easily compare the data for no. of movies produced in that selected year vs the average revenue generated for different genres in that year. One of the most important observation over the years is that, *very less number of 'Animation' movies are produced each year, but they generate highest average revenues.* The legend is muted for the years that are not currently selected for display. Please refer below diagram for the output, after selecting '2006' year from the box-selection:

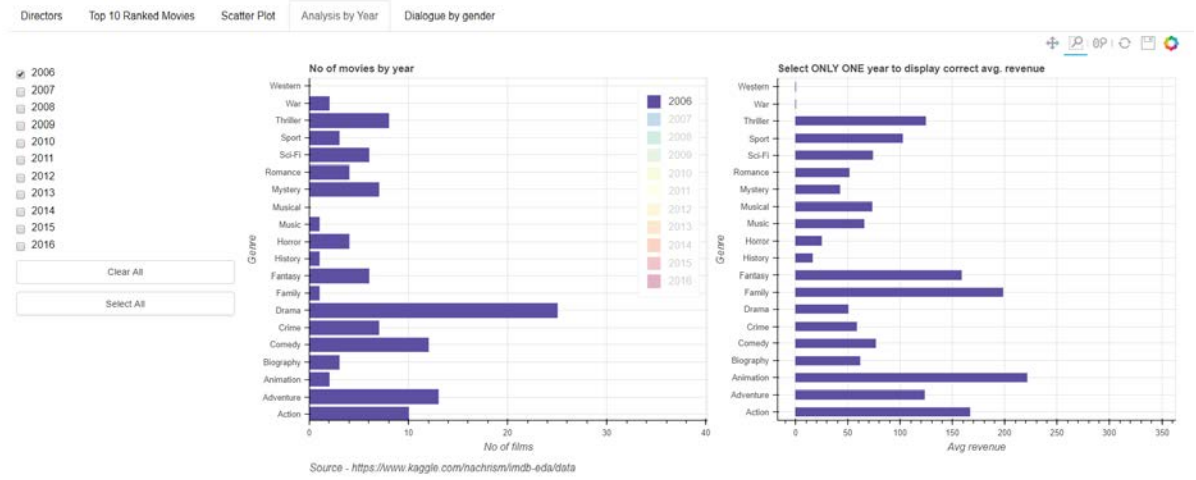

Figure 3(i)

'Select All' button in the widgetbox, makes it easier to select all the checkboxes and display the chart for all the years accordingly. The figure would look like Figure 2(i).

Similarly, the **'Clear All'** button in the widget box, makes it easier to clear all the checkboxes and remove all the data from the graphs. This graphs looks as follows:

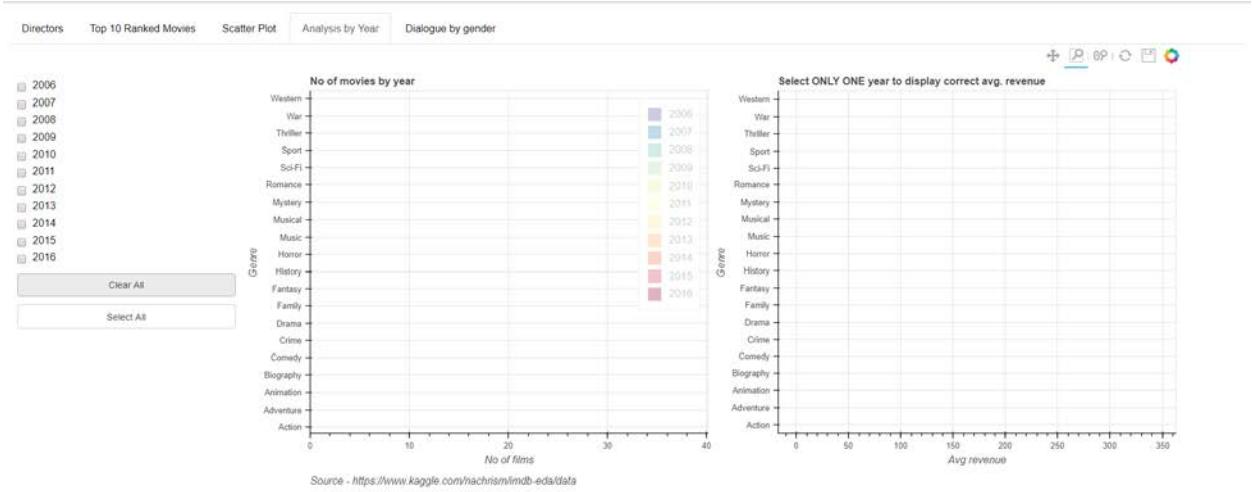

Figure 3(ii)

Shneiderman's key information seeking tasks achieved

Zoom, *Filter*, *Details-on-Demand, History, Extract*

Question 5: What is the distribution of dialogue by gender?

In 2016, several new movie releases featured a lead female character. Does that indicate males and females now share equal roles and dialogue in films?

We got a second dataset that focuses on dialogue data from the top 10 grossing movies. Luckily, dedicated movie fans often transcribe a movie's dialogue and make it freely available online.

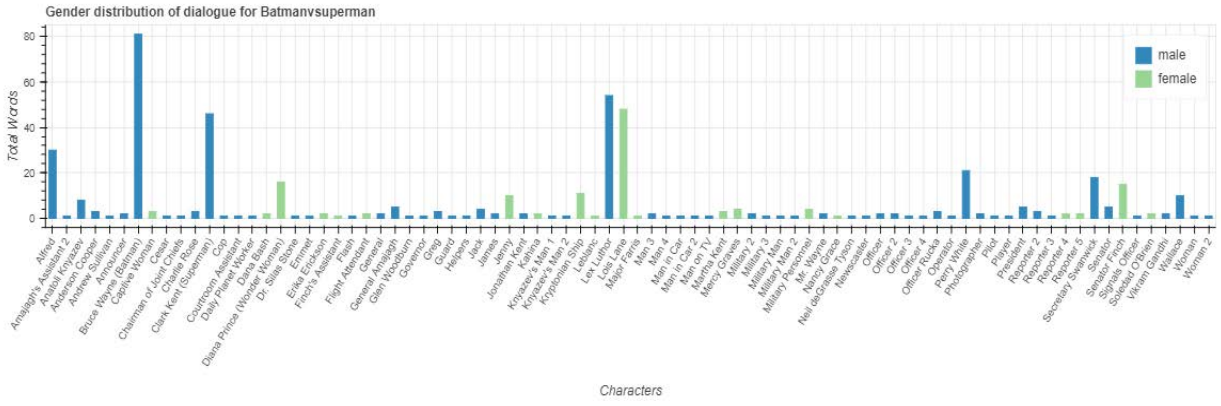

Figure 4

Design

Every bar represents a single speaking character. The height of the bar is scaled based on the number of words spoken by that character. Blue bars represent males, while green bars represent females. Figure 4 shows a visualization for “Batman vs Superman”. We see clearly that only a handful of characters largely dominated the dialogue and very few of those were female.

Rationale behind chosen visualization

We wanted to make comparisons between the total number of words spoken by each character. Using column bar charts was an excellent choice. We were also able to separate the genders which is categorical using colors in the same graph.

Shneiderman’s key information seeking tasks achieved

Zoom, *Details-on-Demand, Extract*

Question 6: Is there a correlation between Rating and Metascore?

The IMDb website registered users can cast a vote (from 1 to 10) on every released movie title in the database. Individual votes are then aggregated and summarized as a single IMDb rating. While the metascore is an aggregated average of movie critic scores. Values are between 0 and 100. Higher scores represent positive reviews.

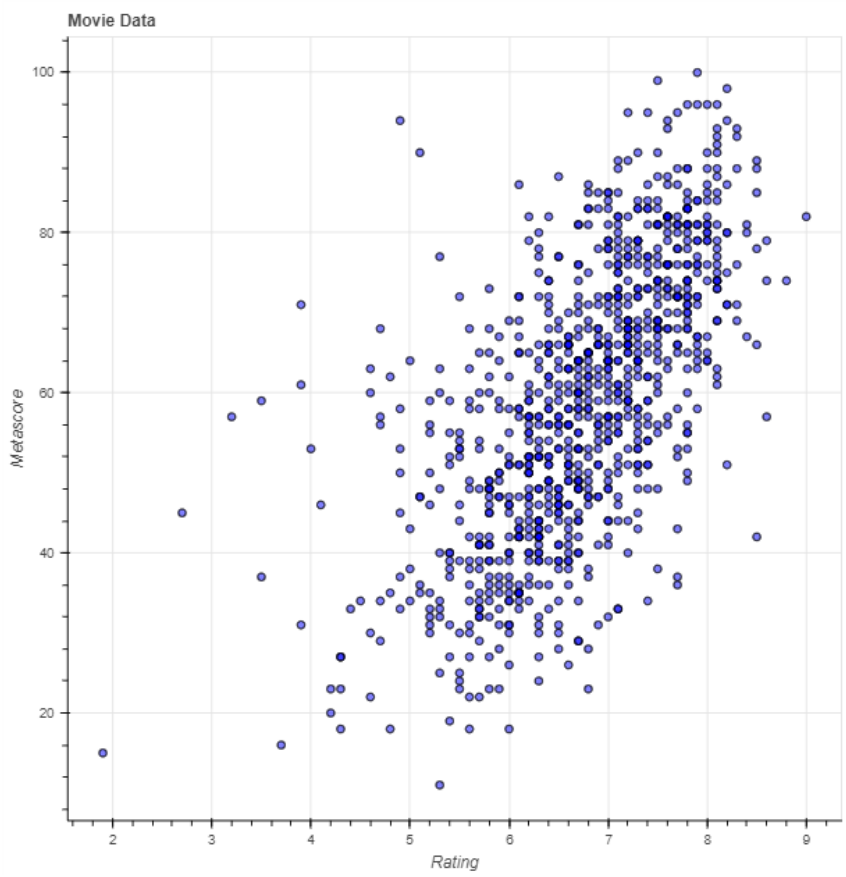

Figure 5

Design

Using a scatterplot easily reveals any such relationships. We made the visualization flexible by adding dropdowns to select the variables for each axis. We also added a date range selector using the `bokeh.models.RangeSlider` class to filter out movies by release date.

Rationale behind chosen visualization

We wanted to be able to discover any relationships that exists between any of the variables in the data. The variables compared here are; rank, runtime, rating, votes, revenue (Millions), and metascore

Shneiderman's key information seeking tasks achieved

Zoom, *Filter*, *Details-on-Demand, History, Extract*

Design Highlights

- We have tried to incorporated all the seven Shneiderman's key information seeking tasks of: *Overview, Zoom, Filter, Details-on-Demand, Relate, History, Extract,* in different plots generated for this project.
- The 'Top 10 Ranked Movies' plot depicts *brushing and linking* functionalities.
- Basic visual symbols or marks, can be arranged in a particular way to represent information. These representations are known as visual variables. We have used following visual variables in our charts, for information visualization:
    - Size
    - Color
    - Position
- Bivariate data is the data in which analysis are based on two variables per observation simultaneously. For our project, the director vs revenue graph, scatter plot, Rank vs Rating plot, Rank vs Metascore plot, Rank vs Revenue plot, represent bivariate data.
- Multivariate data is the data in which analysis are based on more than two variables per observation. Hence, 'Analysis by Year' graph which represents the number of movies produced each year by genre, and the average revenue for each year by genre represents trivariate data. Similarly, the dialogue distribution by gender of each character would fall in this category.
- We have made use of colors for representing categorical data(e.g. Year, gender). This helps in selection, association and adjacent distinction between different attributes of a variable.
- We have tried to remove all the chart junk and tried to optimize data ink ratio, through all of the charts.
- Source of the data is identified so that the audience knows about data gathering practices and legitimacy of the source.

Insights and Conclusion

- Drama, Action, Adventure, Comedy and Thriller were the most frequently produced generes.
- One of the most important observation over the years is that, least number of 'Animation' movies are produced each year, but they generate highest average revenues.
- More than 10%($8B) of the total revenue generated ($72B) comes from only 10 (1.5%) of the top movie directors.
- Males dominate in the dialogue of most of the top movies, however females are starting to take on leading roles.
- There is a high level of correlation between movie ratings and metascore