CS171 Midterm 1

Personal Info

Name: Justin Gonzalez

E-mail: justingonzalez@college.harvard.edu

Link to Google doc:
https://docs.google.com/document/d/1wmuQa83dTRs2HUfHhF9iNrpkmt2QGu-Lz9TGm8apla8/edit?usp=sharing

Part 1

Task 1

1.1 Design Critique

a. The intended audience of this visual is most likely Huffington Post readers who have an interest or personal affiliation to India.

b. This visual tells the audience the percentage of people who live in the selected regions of India that are vegetarian vs non-vegetarian.

c. The color of each circle categorizes the circles into vegetarian and non-vegetarian, the area of each circle is proportional to the percentage of people **in that specific area** who follow that circle's category of diet (with each pair of red and green circles having areas adding up to 100), and the location of each pair of circles indicates the region that these percentages correspond to.

d. If you only care about the dietary breakdown of any one region in India, then I think these methods of visual encoding are effective since it adequately and quickly shows the proportions for each of the diets. However, if you want to look at India as a whole, this method is ineffective since it doesn't account for the population size of each region which makes it difficult to truthfully compare different regions and can cause you to run into Simpson's Paradox.

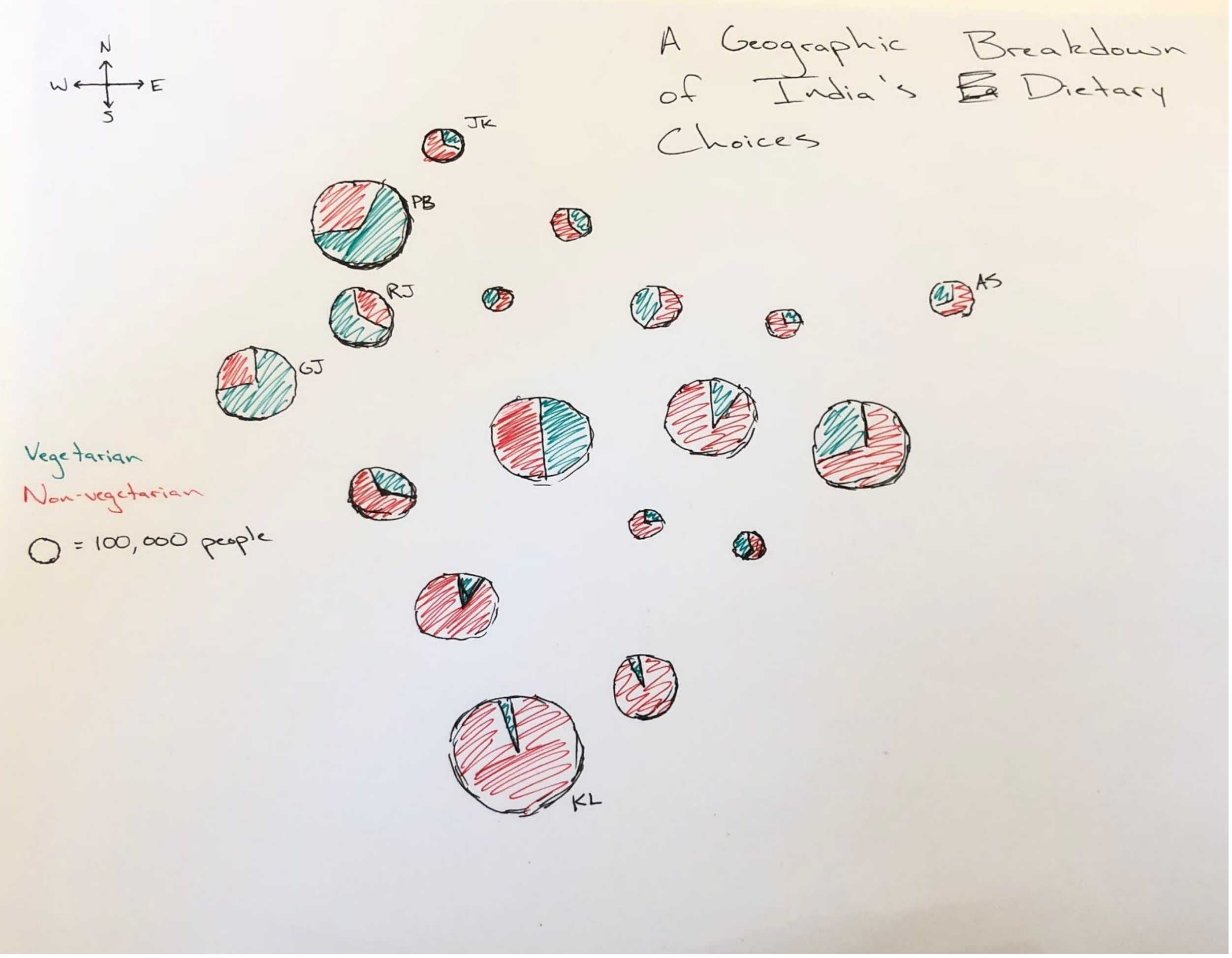
e. Three things that I found to be successful in this visual:
   - The use of contrasting colors (red and green) make it fast and easy to distinguish the two categories of diets
   - A clear breakdown of dietary choices for any one given region
   - The use of location allows you to pick up on certain trends that might be related to geographic location (for instance, it appears that northwestern regions in India have higher percentages of vegetarians **within their own region**).

   Three things that I found this visual to fail in:
   - The size of the circles should not be proportional to the percentages within their own region, but rather the percentages within all of India. This would still allow the audience to make comparisons within a region but also between regions.
   - I find the number labels to be a bit distracting and redundant
   - The stacking of the circles increases the visuals lie factor since the area of ink in the topmost circle takes away from the area of ink in the underlying circle.

f. Overall, I don't like this visual since I find it can be a bit misleading when trying to make generalizations about India as a whole, which, as the title suggests, is most likely the intended purpose of this visual.

1.2 Redesign

In my visual, I kept the components from the original visual which I thought worked. These components were the geographic location of each circle as well as the color choice to distinguish vegetarian from non-vegetarian. However, I also addressed the issue of being able to compare between regions by making the size of each pie chart proportional to the population of each region, and I minimized the lie factor by getting rid of the overlying circles for pie charts.



Task 2

2.1 Graphical Integrity

This visual uses an inconsistent time scale with a much larger gap between 1964 and 1975 than 1975 and 1990 which takes away from the graphical integrity and increases the lie factor of this visual. The use of an illustration of a doctor is also unnecessary chart junk that makes the visual much more difficult to interpret since it is an irregular shape with irregular area. This illustration also drastically reduces the data-ink ratio since the same amount of information could have been conveyed with a few lines of varying lengths.

2.2 Gestalt Principles

Connection: Each bottle of wine is physically connected to the foods it pairs well via lines. Alternatively, foods that don't pair well with wine aren't connected to any bottles at all.
Similarity: The color of each food is similarly colored to the types of wines it pairs well with.
Proximity: Similar types of wines and similar types of foods are placed near one another in the visual.

2.3 CRAP

Contrast: The parts of the visual that indicate value are clearly contrasted with the other elements through the use of color. Also, different categories are given contrasting and distinct colors.
Repetition: The colors of the different categories are repeated throughout the visual to promote cohesion.
Alignment: The items in the pinwheel are aligned next to their neighboring categories as well in the "in detail" section below. This also promotes cohesion.
Proximity: Similar categories are placed near one another in the pinwheel.

Task 3

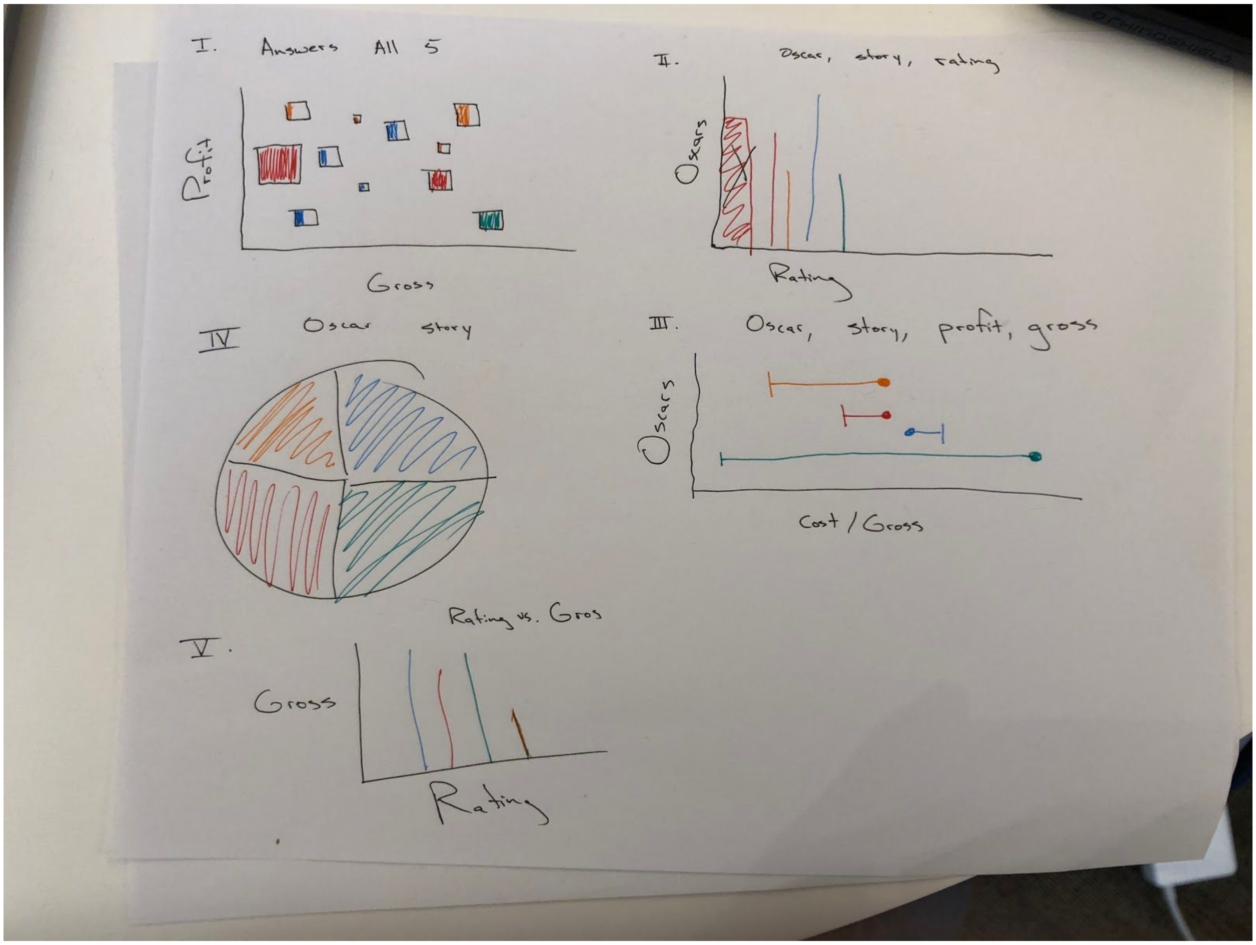
3.2 Audience/Questions

My intended audience is going to be people who are passionate about movies (film buffs). Some questions my audience might have include:

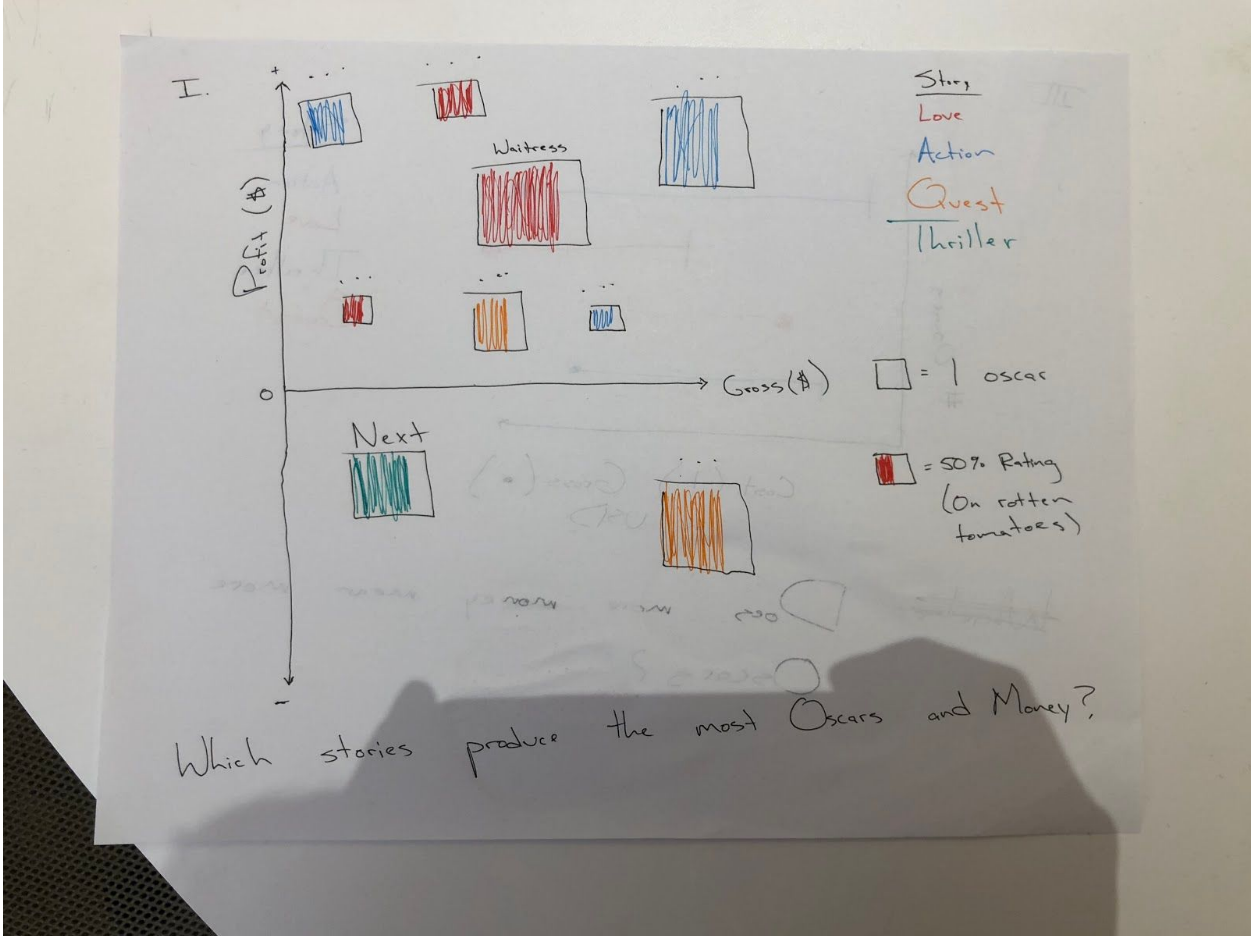
1. Which types of story are the highest grossing (domestically)? (Success)
2. Which types of story have the highest profitability? (Success)
3. Which types of story are the most highly rated (on Rotten Tomatoes)? (Success)
4. Which types of story are most likely to win an Oscar? (Success)

5. Is there a relationship between story, rating, gross, profitability, and number of Oscars?

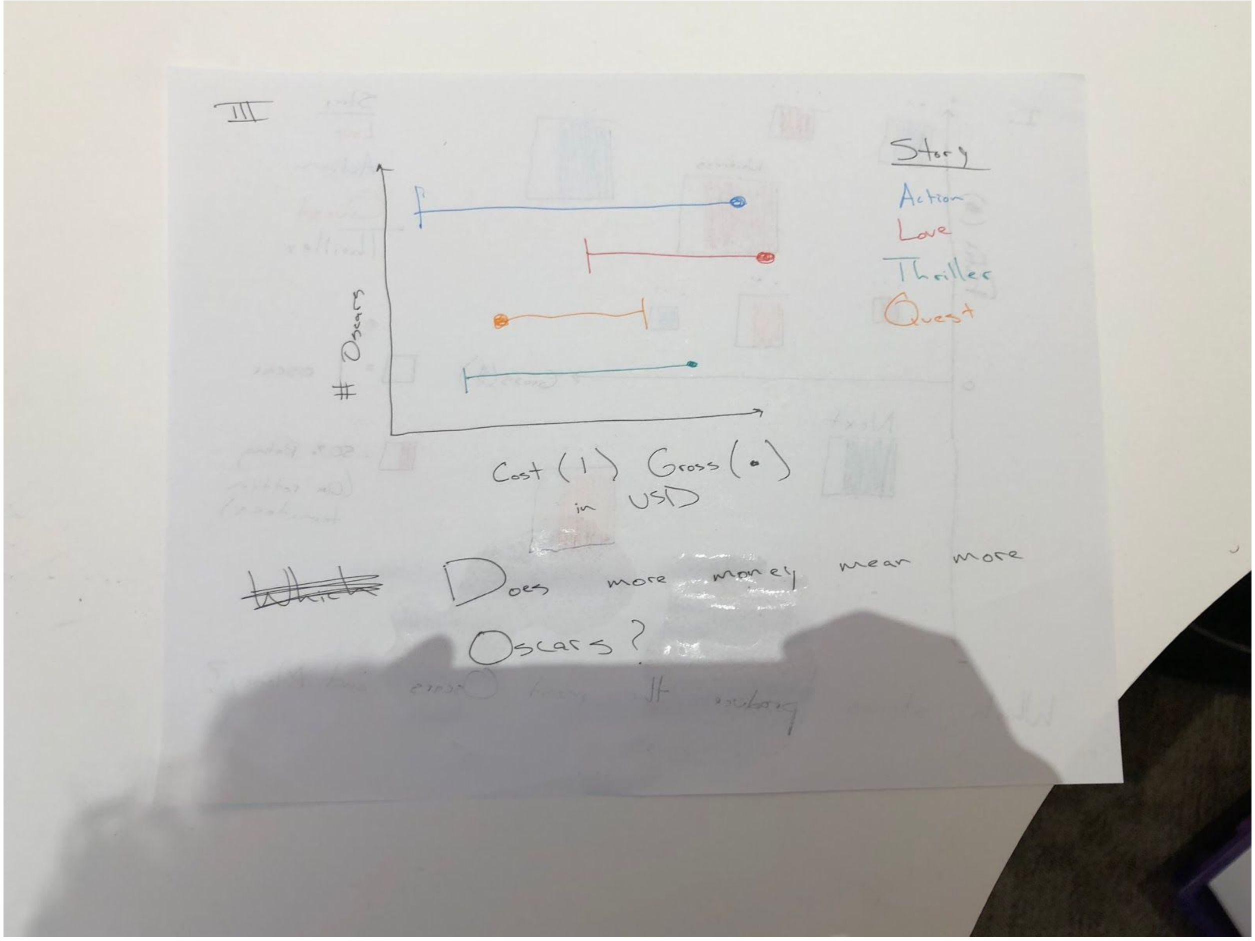
3.3 Quick Sketches



3.4 More in Depth



Which stories produce the most Oscars and Money?



~~Which~~ Does more money mean more Oscars?

High data-ink ratio

Contrast

Repetition

Marks: squares, dots lines

Channels: color

Part 2

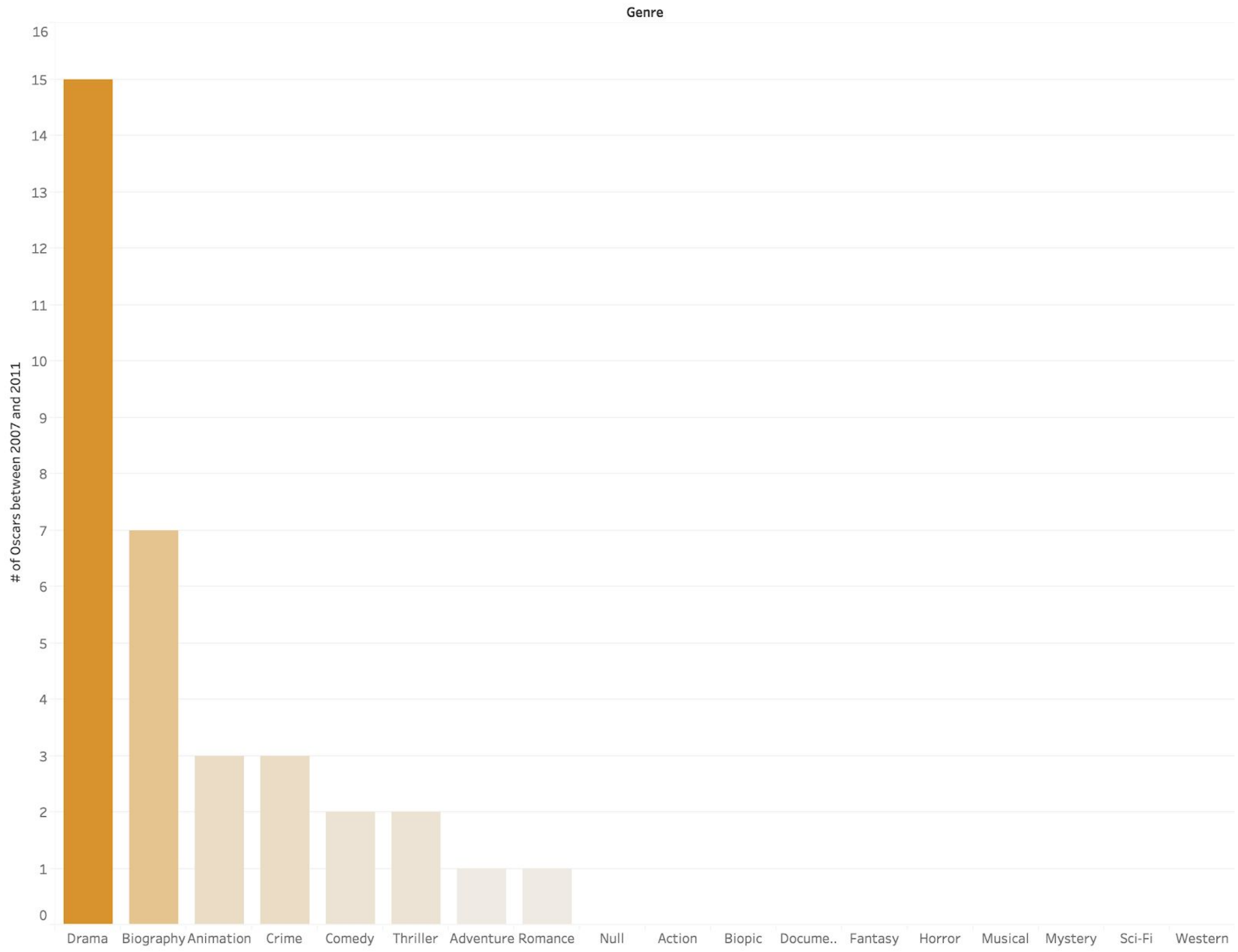
Task 1

1.1 Loading Data

Upon inspection, all of the data types made sense to me. I wanted to change the type of “Multiple Oscars” from a string to a boolean, but Tableau does not recognize Yes and No as acceptable booleans. I did not want to manipulate the data, so I decided to leave that alone.

1.2 Explore the Data

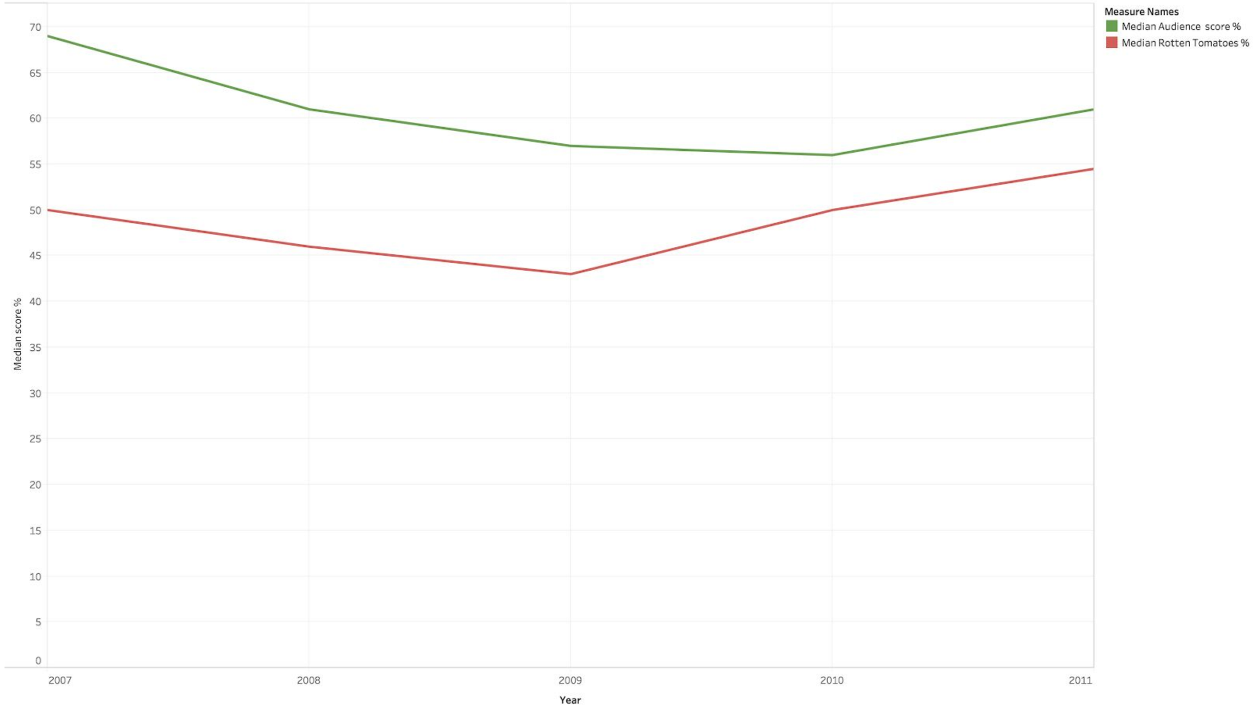
Which genre has been award the most Oscars?



Sum of # of Oscars for each Genre. Color shows sum of # of Oscars.

(Sheet 2)

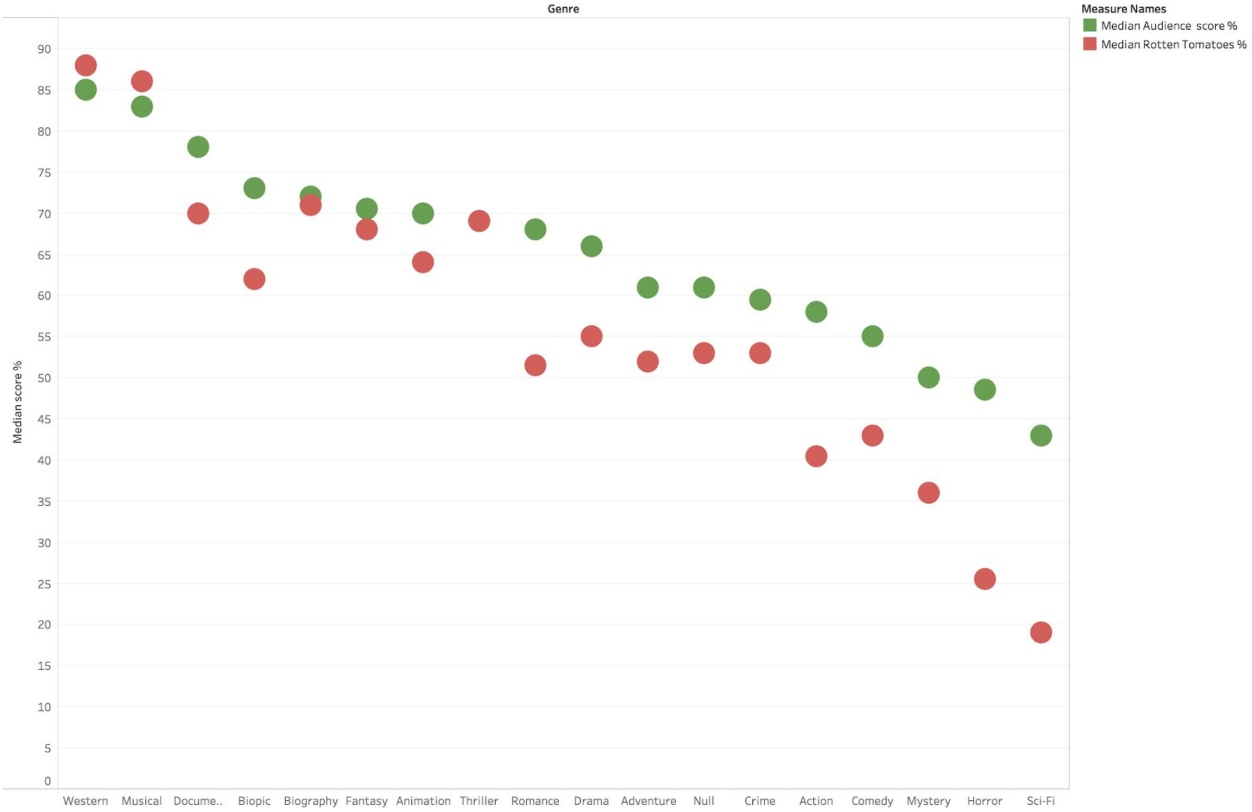
How have critic and audience scores changed over time?



The trends of Median Audience score % and Median Rotten Tomatoes % for Year. Color shows details about Median Audience score % and Median Rotten Tomatoes %.

(Sheet 3)

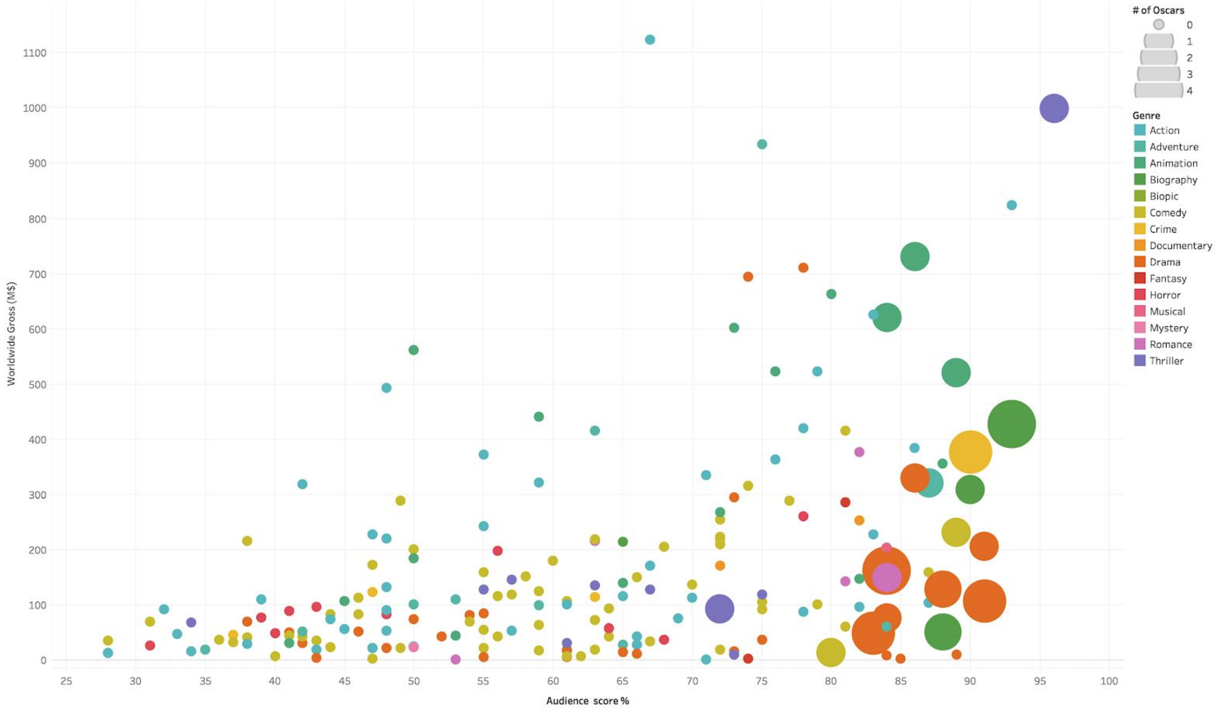
Which genres do critics and audience members tend to rate most highly?



Median Audience score % and Median Rotten Tomatoes % for each Genre. Color shows details about Median Audience score % and Median Rotten Tomatoes %.

(Sheet 9)

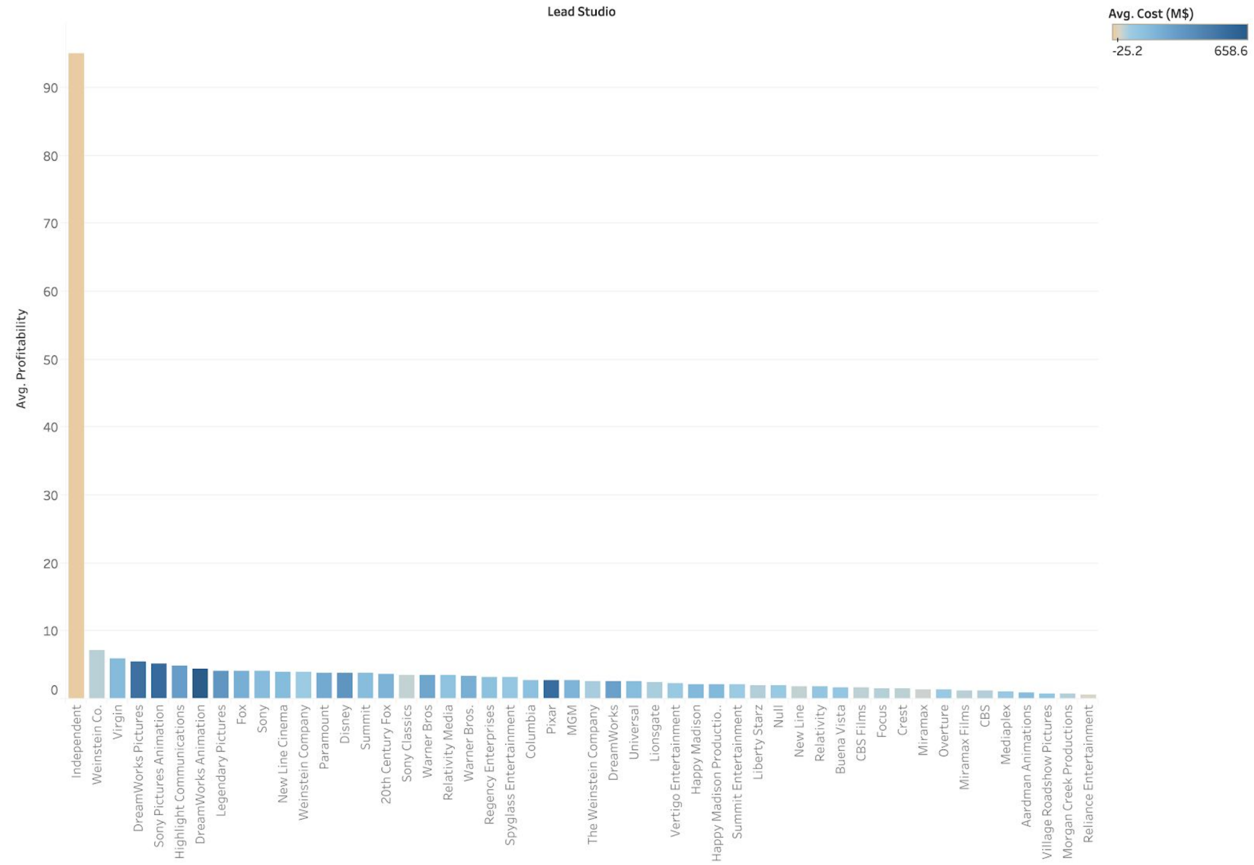
Do better audience scores mean more money and more Oscars?



Audience score % vs. Worldwide Gross (M$). Color shows details about Genre. Size shows details about # of Oscars. Details are shown for Film and Oscar. The view is filtered on Film, which keeps 200 of 669 members.

(Sheet 4)

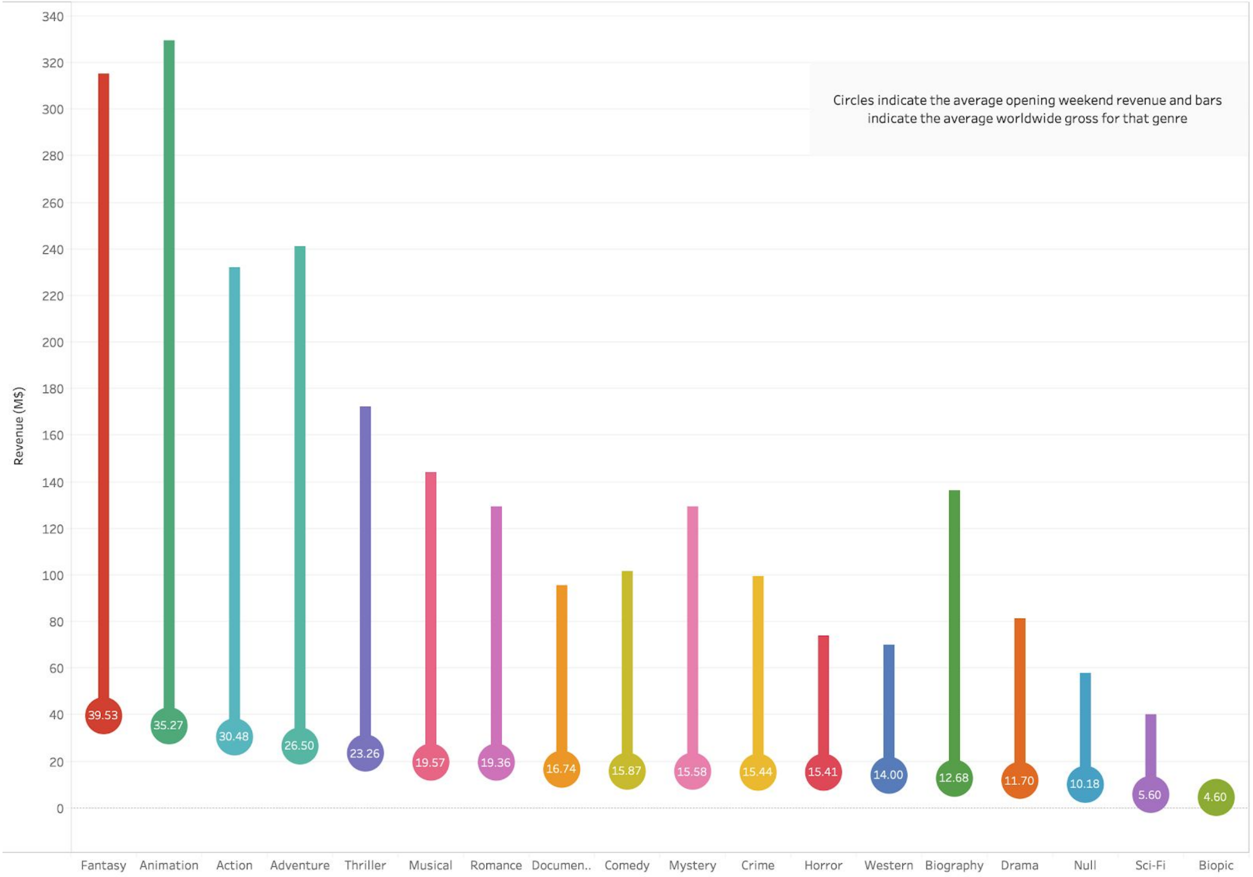
Which studios (on average) produce the most profitable films?



Average of Profitability for each Lead Studio. Color shows average of Cost (M$). The view is filtered on average of Profitability, which keeps non-Null values only.

(Sheet 11)

Which genres show the greatest increase in revenue past opening weekend?



Average of Opening Weekend (M$) and average of Opening Weekend (M$) for each Genre. Color shows details about Genre. For pane Average of Opening Weekend (M$): Size shows average of Growth after opening weekend. For pane Average of Opening Weekend (M$) (2): The marks are labeled by average of Opening Weekend (M$).

(Sheet 8)

1.3 Initial Analysis

(Sheet 2) My first visual showed that between 2007 and 2011, Drama movies were awarded the most Oscars which is what I predicted the outcome to be. I was surprised to see that animations were the third most awarded genre of film.

(Sheet 3) My second visual showed that over time, critic scores and audience scores tend to rise and fall in relatively similar fashions. It also appears that over time, the gap between critic scores and audience scores have began to narrow.

(Sheet 9) For my third visual, I was very surprised to see that Westerns and Musicals were the highest appraised genres by critics and audience members alike. Similarly, I thought it was interesting to see that these were the only two categories where the critic scores were higher than the audience scores. However, I think that these observations are due to too few films in each of these genres to form a representative sample.

(Sheet 4) From my fourth visual, I learned that the majority of Oscar winning films have an audience score falling between 80 and 95 percent. The data also appears to exhibit an exponential curve. The highest earning and most highly rated Oscar winning film in this data set was the Dark Knight, and (Disney) animations proved to be the highest earning, Oscar winning genre.

(Sheet 11) From my fifth visual, I learned that Independent films are the most profitable movies on average which makes sense since Independent films have the lowest costs of production. (To calculate cost I created a new calculated value called Cost which subtracts profit from worldwide gross. This gave me a negative value for Independent films which makes me question how this data set defined profit.)

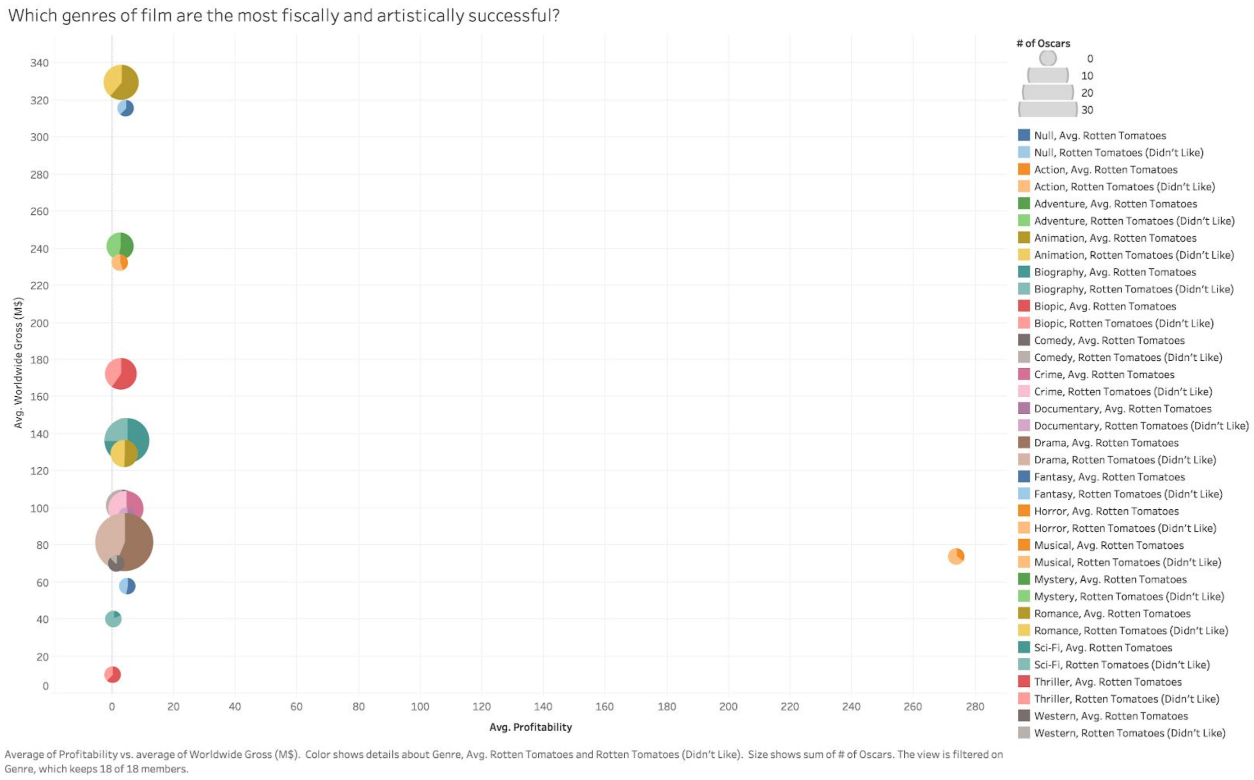
(Sheet 8) My sixth visual shows that animations have the highest average growth after opening weekend and fantasy films have the largest average opening weekends.

Task 2

2.1 Choose Questions and Sketches

After doing some exploration with the data, I find myself fascinated with success and the various capacities in which something can be successful. Some metrics of success that I explored in Task 1 of Part 2 include profitability, gross revenue, audience scores, critic scores, and number of awards (specifically Oscars). I think the visual that best indicates overall success is my first visual from Part 1 Task 3 which touches on each of these metrics. Instead of color categorizing by story, however, I think I am going to categorize by genre since it is a much more familiar and easier to define concept to the average audience member. This visual will answer the question, "Which genres of film are the most fiscally and artistically successful?" For my second visual, I want to focus more deeply on animations as I've noticed they appear to exhibit the most overall success in the box office. For this visual, I will use the third design I created in Part 1 Task 3 but forgo the color categorization by genre for a color scale that represents the number of Oscars. The axis currently labeled with number of Oscars will be changed to film titles. This question will answer "Which animated films are the most fiscally successful?"

2.2 Implement Sketches

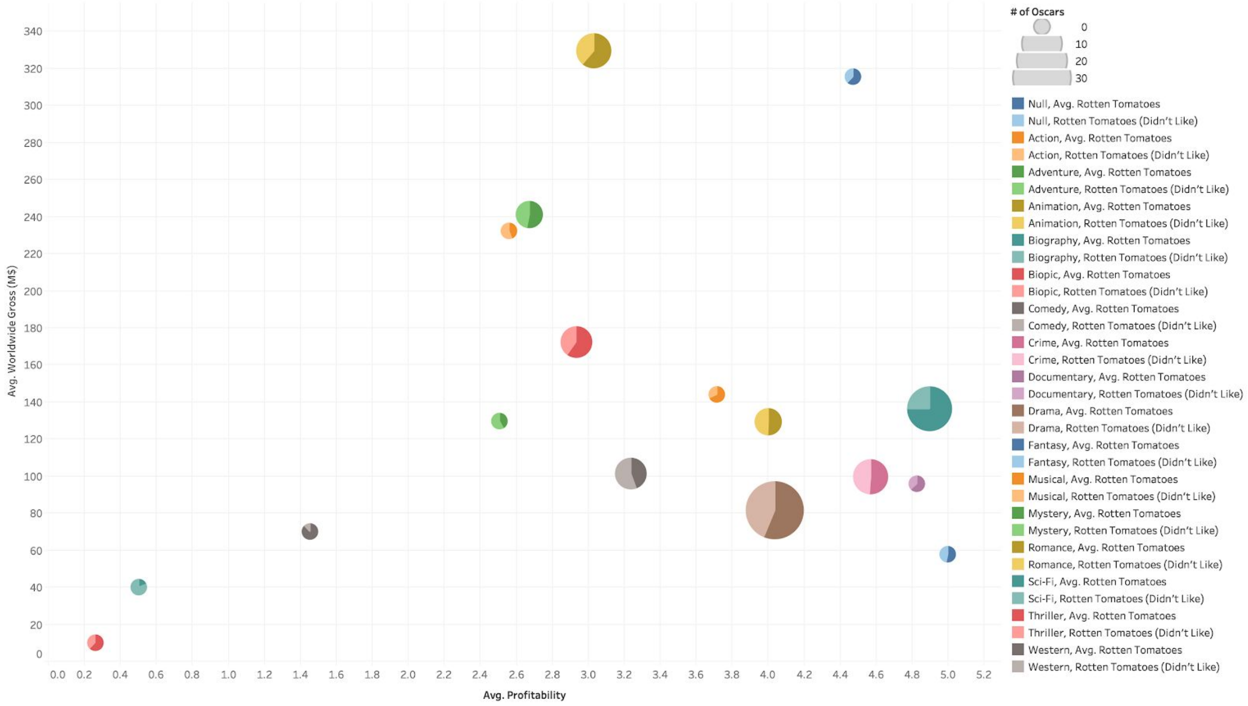


Which genres of film are the most fiscally and artistically successful?

Average of Profitability vs. average of Worldwide Gross (M$). Color shows details about Genre, Avg. Rotten Tomatoes and Rotten Tomatoes (Didn't Like). Size shows sum of # of Oscars. The view is filtered on Genre, which keeps 18 of 18 members.

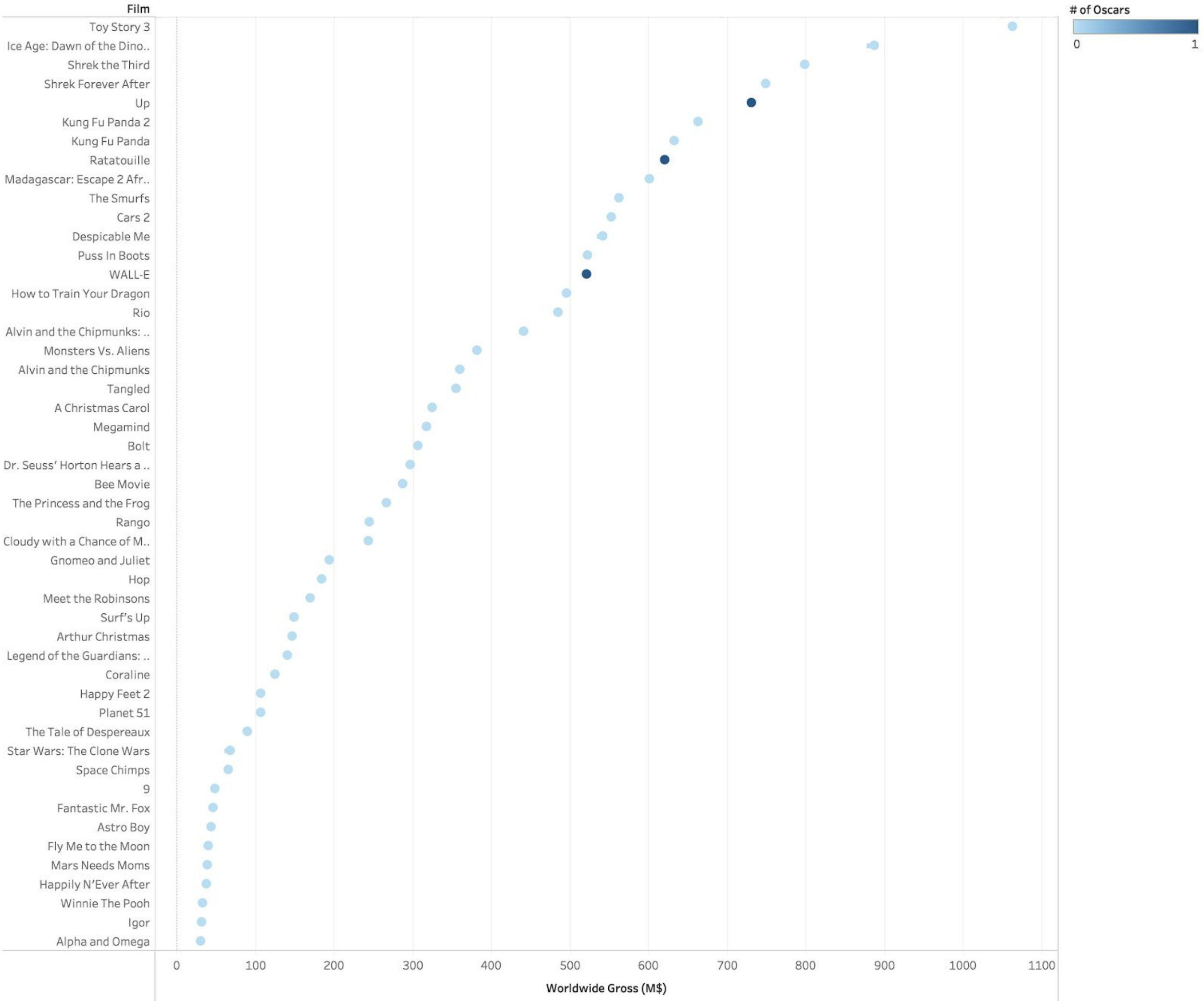
The above image was my first attempt at creating my first visual (Sheet 14). Upon creation, I realized that horror was a significant outlier. The values for this data point also do not make sense since, according to the data, horror movies had an average profit far exceeding their total worldwide gross. The following visual is the result after I removed this outlier.

Which genres of film are the most fiscally and artistically successful?



Average of Profitability vs. average of Worldwide Gross (M$). Color shows details about Genre, Avg. Rotten Tomatoes and Rotten Tomatoes (Didn't Like). Size shows sum of # of Oscars. The view is filtered on Genre, which excludes Horror.

In the end, this visual closely matches what I envisioned in my original sketch. The main difference is I had to use pie charts instead of squares since I couldn't figure out how to shade portions of a square, but I think this looks better anyways.

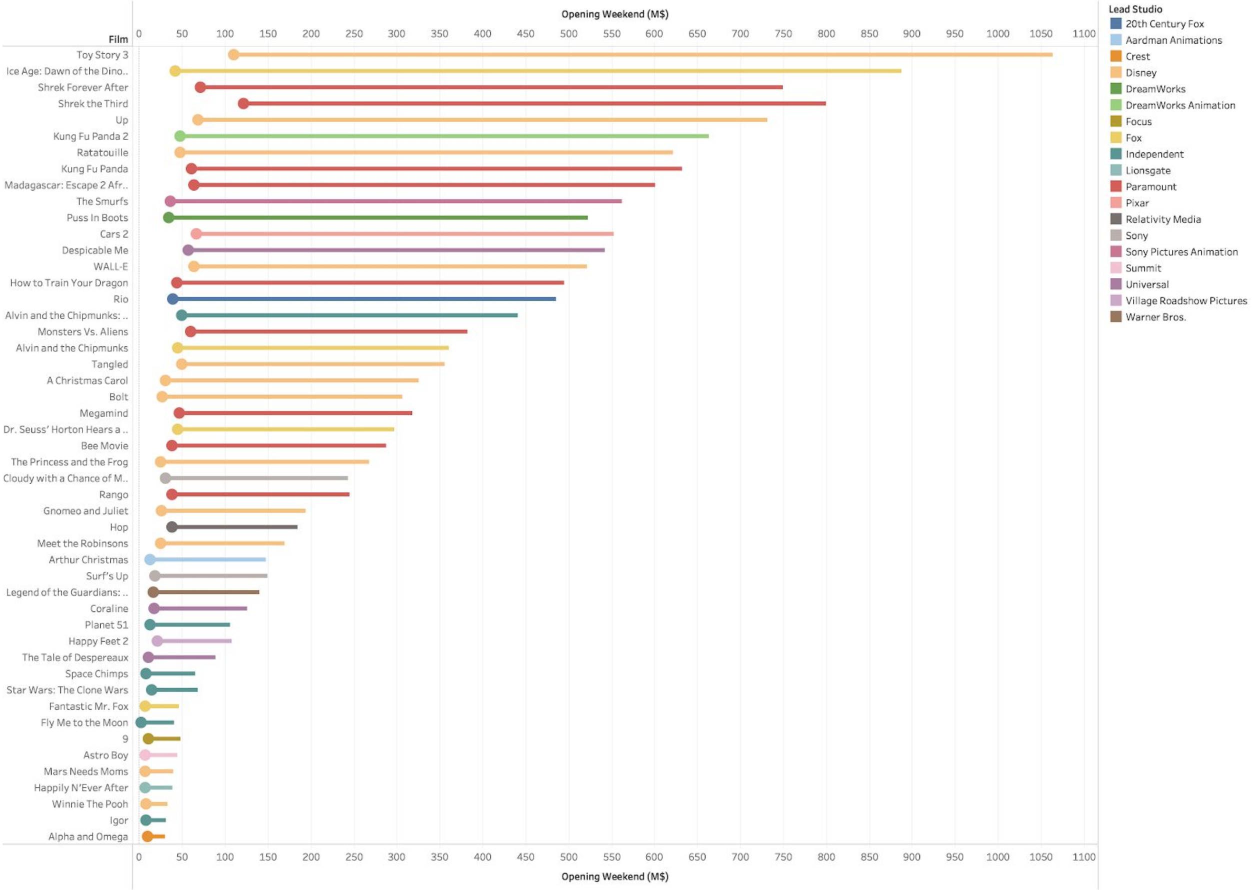


Which animated films are the most fiscally successful?

Worldwide Gross (M$) and Worldwide Gross (M$) for each Film. Color shows sum of # of Oscars. Details are shown for Lead Studio. For pane Worldwide Gross (M$): Size shows -SUM([Profitability]). The data is filtered on Genre, which keeps Animation.

Above was my first attempt at creating my second visual (Sheet 16). I tried to make a trail the size of each movie's profit, but most of the profits were so small in comparison to the worldwide gross that they were barely visible. So instead of looking at the profit, I'm going to look at how each movie's revenue grew after opening weekend similar to how I did in Sheet 8. Also, instead of using the color to represent number of Oscars (I felt as though it was a waste in the above visual since only three of the movies received one Oscar each), I will use color to represent the film's lead studio to see if certain studios produce more successful animated films.

Which animated films are the most fiscally successful?



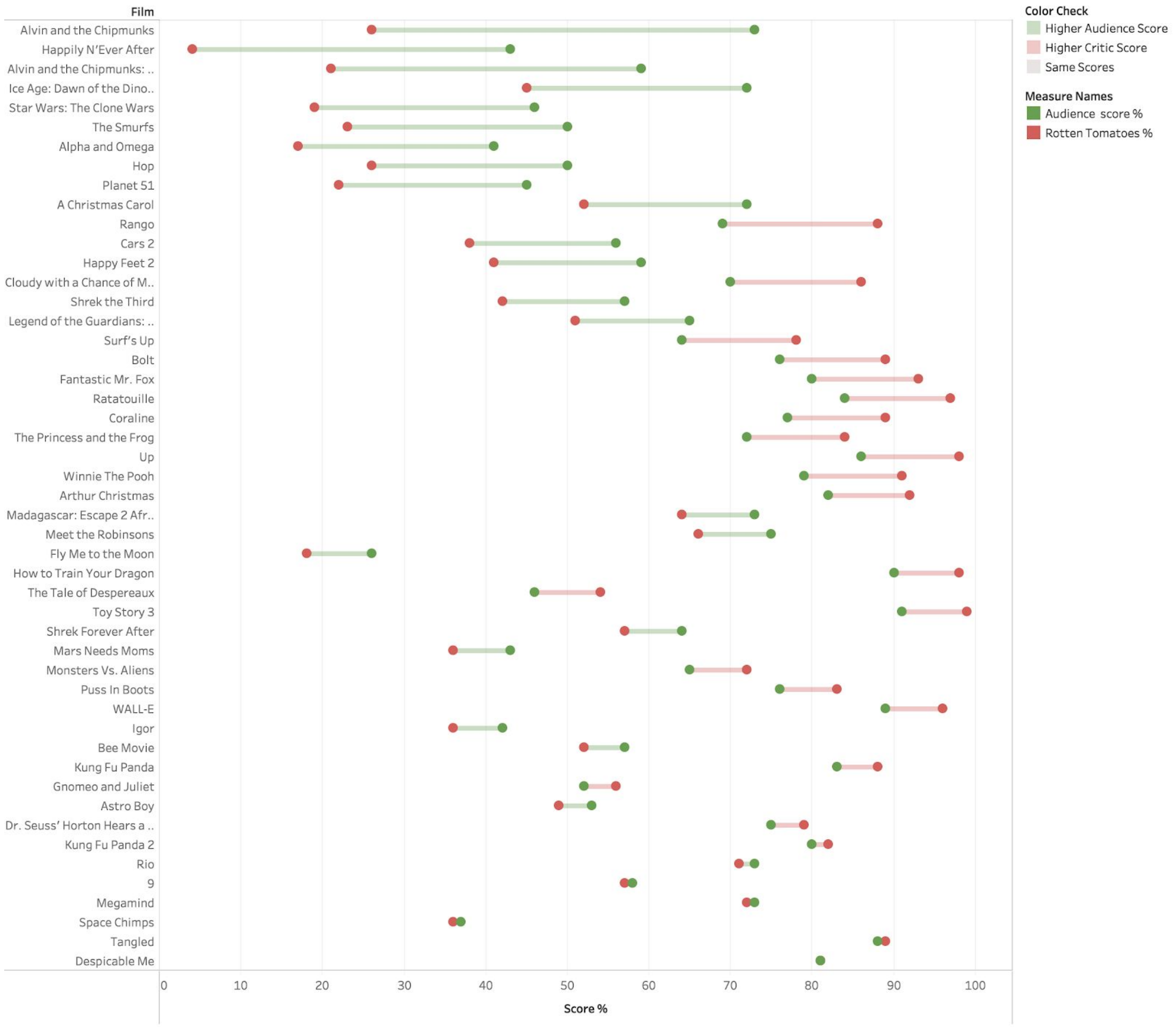
Opening Weekend (M$) and Opening Weekend (M$) for each Film. Color shows details about Lead Studio. For pane Opening Weekend (M$): Size shows details about Growth after opening weekend. The data is filtered on Genre, which keeps Animation.

This visual is much more what I pictured in my head and I believe to be much more informative than what I previously envisioned.

Task 3

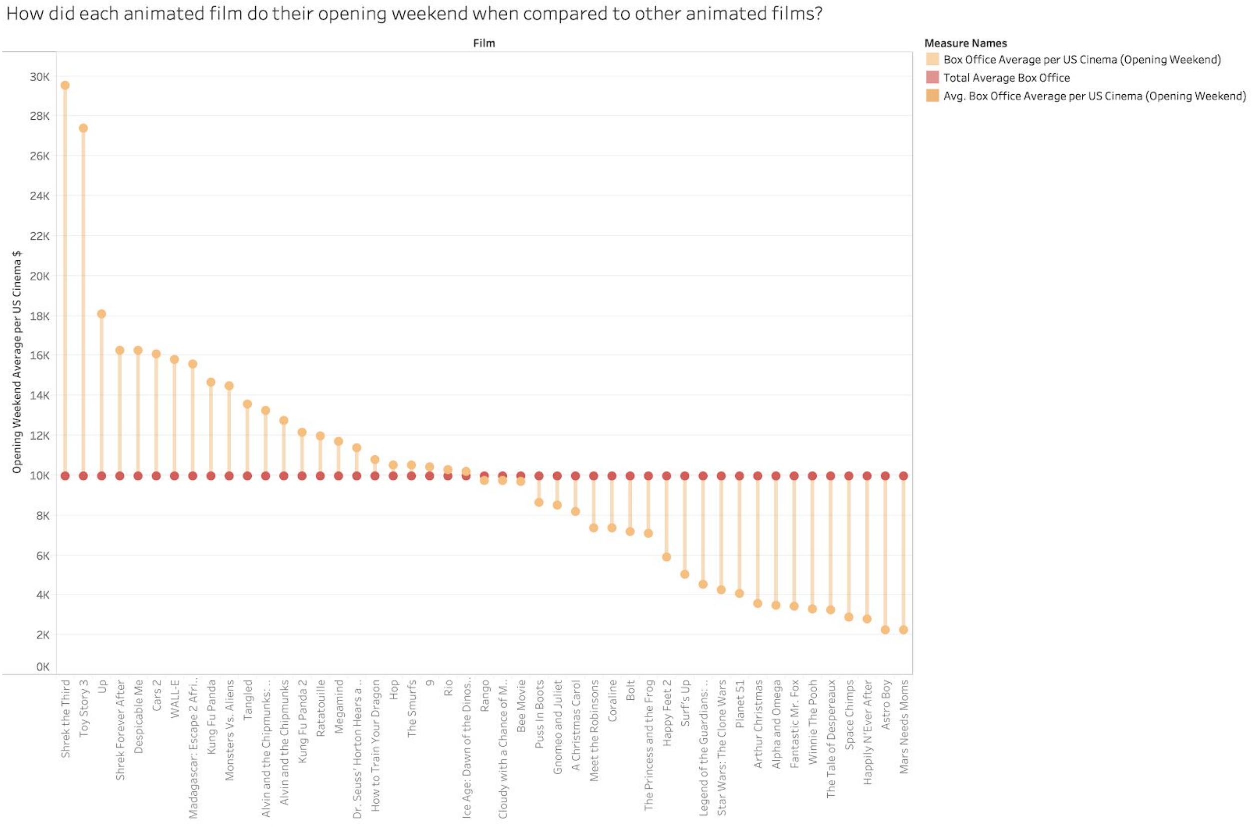
3.1 Calculated Fields

Which animated films are the most disputed between critics and the audience?



Audience score %, Rotten Tomatoes % and Rotten Tomatoes % for each Film. For pane Measure Values: Color shows details about Audience score % and Rotten Tomatoes %. For pane Sum of Rotten Tomatoes %: Color shows details about Color Check. Size shows sum of Difference between Critic and Audience Scores. The data is filtered on Genre, which keeps Animation.

For this visual (Sheet 17) I had to create three new calculated fields. One is called [Difference between Critic and Audience Scores] which subtracts the critic score from the audience score. This value is the size of the Gantt bar between the two circles. Another is the absolute value of this previous value so I can sort the films from most disputed to least disputed. The last one uses conditional statements to check whether the difference between the scores is positive or negative. (If it’s positive the audience score is higher, if it’s negative the critic score is higher.) I then use this to color the bar for a quick indication of whether the audience or critics like a movie more. These values then allow me to answer the question of which animated films are most disputed between critics and the audience.



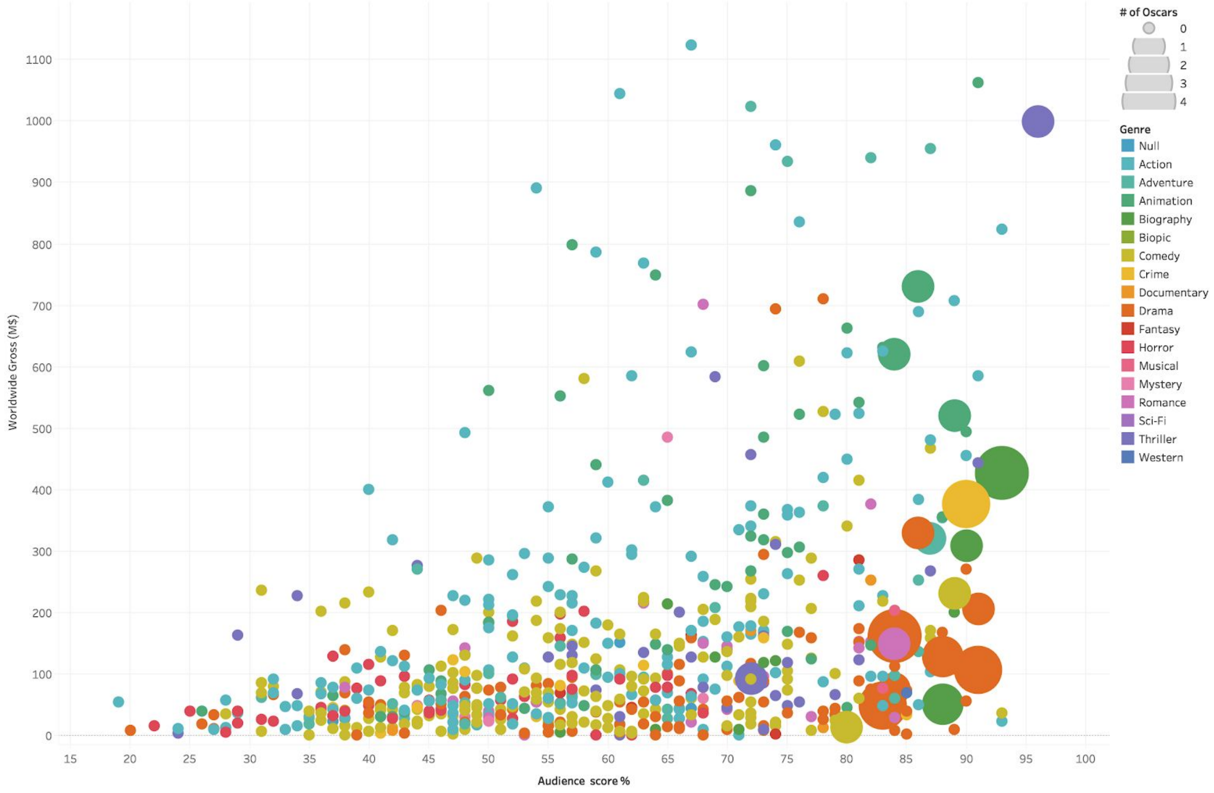
How did each animated film do their opening weekend when compared to other animated films?

Box Office Average per US Cinema (Opening Weekend), Total Average Box Office, Avg. Box Office Average per US Cinema (Opening Weekend) and Avg. Box Office Average per US Cinema (Opening Weekend) for each Film. Color shows details about Box Office Average per US Cinema (Opening Weekend), Total Average Box Office and Avg. Box Office Average per US Cinema (Opening Weekend). For pane Average of Box Office Average per US Cinema (Opening Weekend): Size shows Distance from Total Average Box Office. The data is filtered on Genre, which keeps Animation. The view is filtered on average of Box Office Average per US Cinema (Opening Weekend), which keeps non-Null values only.

For this visual (Sheet 18), I had to create two new calculated fields. The first is called [Total Average Box Office] which is represented by the red dots in the picture above. This value finds the total average of each of the average opening weekend per US cinema values (the orange dots). The other calculated field is called [Distance from Total Average Box Office] which is calculated by subtracting the [Opening Weekend Average per US Cinema] from the [Total Average Box Office] and is used to scale the Gantt bars and connect the two circles.

3.2 Add Filtering

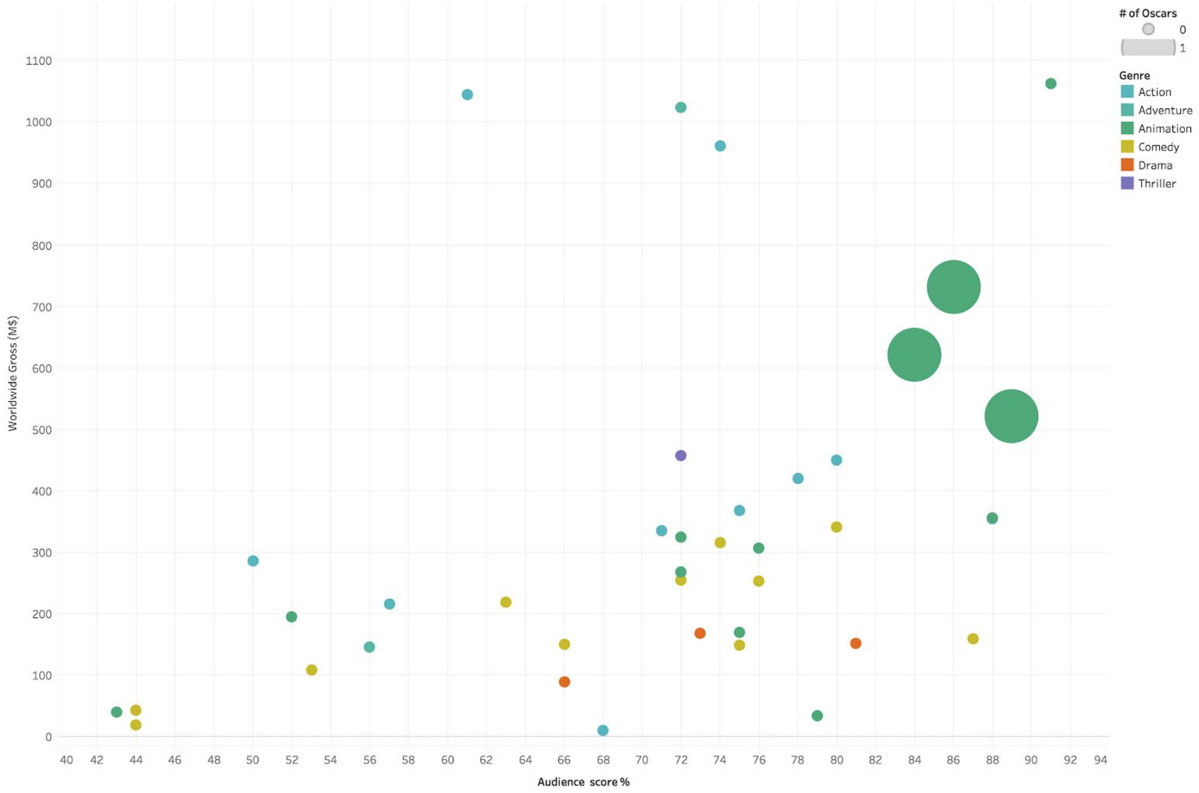
Do better audience scores mean more money and more Oscars?



Audience score % vs. Worldwide Gross (M$). Color shows details about Genre. Size shows details about # of Oscars. Details are shown for Film and Oscar. The data is filtered on Lead Studio, which keeps 51 of 51 members. The view is filtered on Film, Audience score % and Worldwide Gross (M$). The Film filter keeps 669 of 669 members. The Audience score % filter keeps non-Null values only. The Worldwide Gross (M$) filter keeps non-Null values only.

Sheet 4 is probably my messiest visualization but contains a bunch of data that is very useful. For this reason I am adding a checkbox which filters by Lead Studio and will provide even more insight about which studios produce the highest earning and most highly acclaimed movies. As an example, filtering by Disney produces the following result.

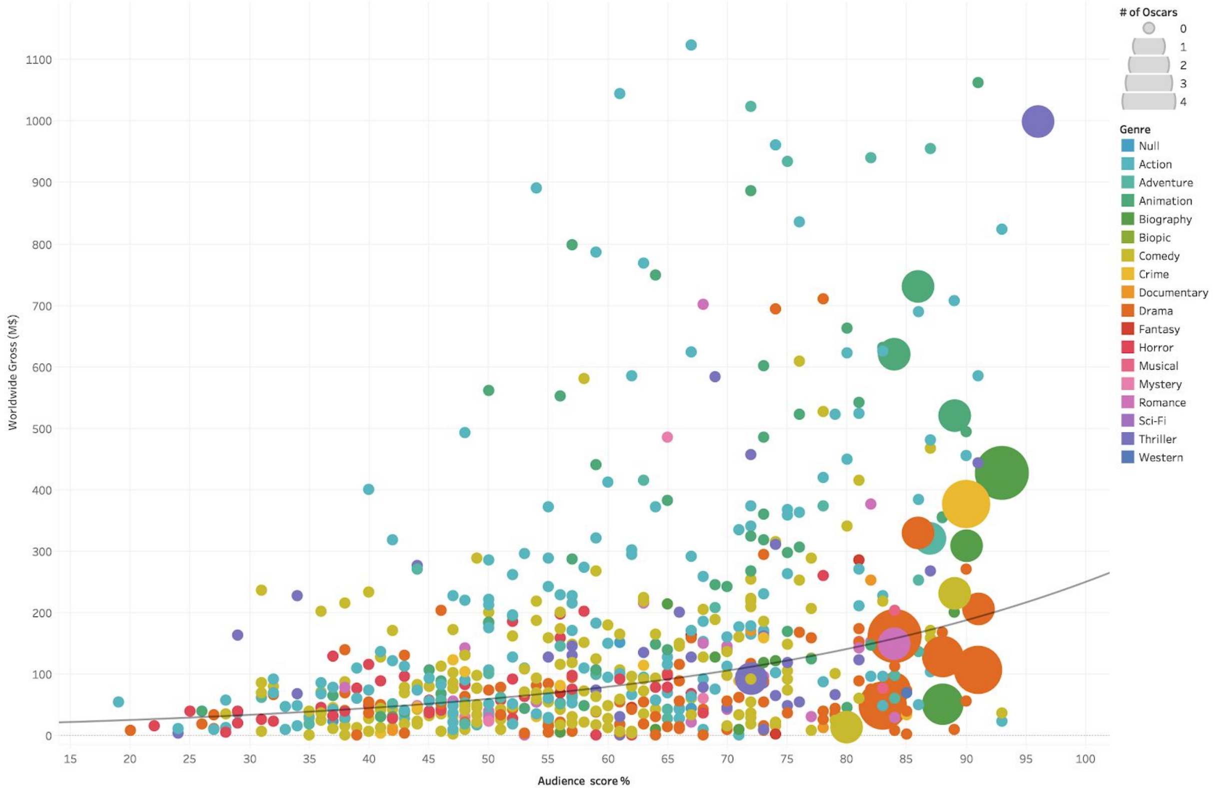
Do better audience scores mean more money and more Oscars?



Audience score % vs. Worldwide Gross (M$). Color shows details about Genre. Size shows details about # of Oscars. Details are shown for Film and Oscar. The data is filtered on Lead Studio, which keeps Disney. The view is filtered on Film, Audience score % and Worldwide Gross (M$). The Film filter keeps 669 of 669 members. The Audience score % filter keeps non-Null values only. The Worldwide Gross (M$) filter keeps non-Null values only.

3.3 Add Analytics

Do better audience scores mean more money and more Oscars?

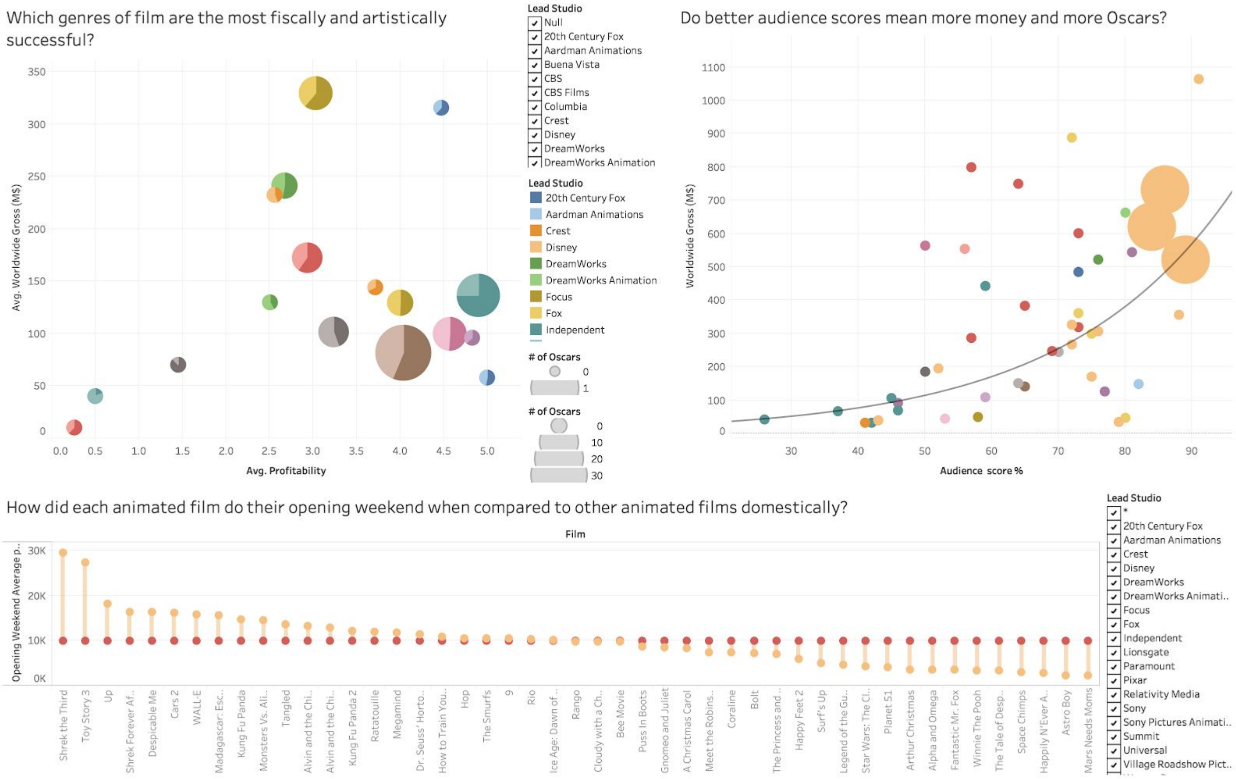


Audience score % vs. Worldwide Gross (M$). Color shows details about Genre. Size shows details about # of Oscars. Details are shown for Film and Oscar. The data is filtered on Lead Studio, which keeps 51 of 51 members. The view is filtered on Film, Audience score % and Worldwide Gross (M$). The Film filter keeps 669 of 669 members. The Audience score % filter keeps non-Null values only. The Worldwide Gross (M$) filter keeps non-Null values only.

As I previously mentioned in Part 2 Task 1.3, the data in Sheet 4 appears to be in the shape of an exponential curve. Just to confirm, I went through each option of trend line under analysis and checked the r-squared and p-values for each type of fit and found that exponential does indeed work best for this data set. I also made sure to allow the curve to be recalculated whenever you filter by lead studio or genre. What I really like about this trend line is how well it creates order amid all the chaos of Sheet 4. With this trend line, we can use a film's audience score to predict an approximate worldwide gross for a movie with a standard error of about 1.16 sigma. These predictions become even stronger once you begin to filter by genre and lead studio.

Task 4

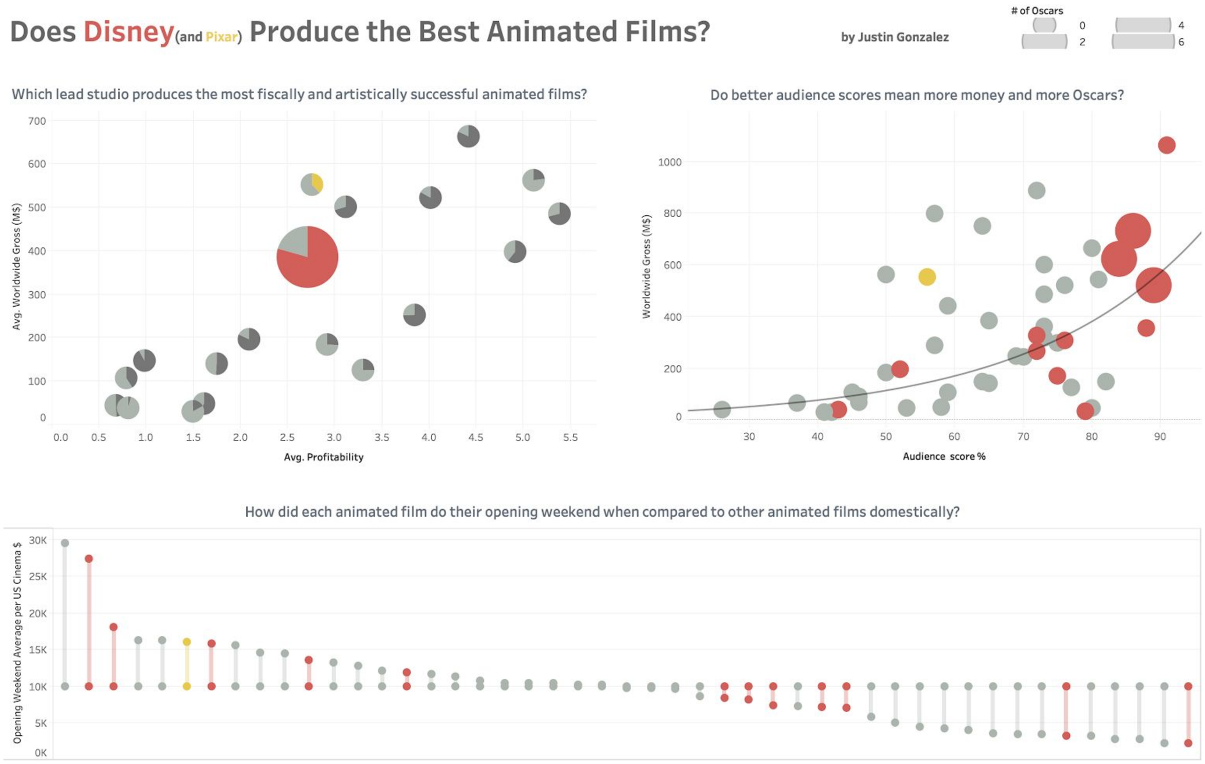
4.1 Create Dashboard


Which genres of film are the most fiscally and artistically successful?

Do better audience scores mean more money and more Oscars?

How did each animated film do their opening weekend when compared to other animated films domestically?

4.2 Improve Dashboard



For my revised dashboard, I chose to focus on Disney and Pixar Studios for simplicity, focus, and a more cohesive story in my dashboard. As far as my color choice, I chose to pay homage to Mickey Mouse's classic red and yellow outfit color scheme. I changed the color of the title to doubly act as a key to the color scheme. I made all the non-Disney and Pixar studios grey to increase contrast. The top left visual allows you to filter by studio and the top right visual allows you to filter by film. I chose to use these three visuals since they show every metric you would really want to know. The top left shows profitability, rotten tomato scores, oscars, and worldwide gross. The top right shows worldwide gross, oscars, and audience. The bottom shows opening weekend per US cinema. My visual placement was predominantly due to the the bottom being so long and the other two being relatively square. Plus, I also thought the top two were more important and deserved to be viewed first.

4.3 Hindsights

Overall, I thought this entire process went very well. If there's one thing that I've learned is that Tableau is a bit limiting if you want to make rather unique visuals. It's great though if you have more traditional designs in mind. I wish tableau allowed you to fill any shape proportional to a certain value similar to a pie chart. There are still a few parts of the data set that I am confused by. For instance, is there a difference between profitability and profit? Why did Horror movies have much more profitability than worldwide revenue? Is the budget the same as cost of production? Playing with the data, creating new calculated fields, and creating trend lines is what taught me the most about the data.