Tim Vigers

February 25, 2020

Information Visualization

*Analyzing Multivariate Movie Data with Tableau*

Part 1

**Question 1:** What are the projected median gross and budget in 2020 according to this data?

I started with this question because it’s the closest to what I do in my biostatistics consulting work. First, I checked the distribution of gross to decide whether to plot average or median. Tableau makes it really easy to produce histograms of each variable, and also to change which aggregate measure is plotted on the y axis. It’s also very easy to add a new variable, like genre, and choose the right channel and mark for it.

Next, I used the analytics tab to add various trend lines and see which model fit the data best, based on the R-squared value provided by Tableau. Among the basic options, the exponential model had the highest R-squared while still providing a reasonable projection for gross in 2020. The default polynomial model (up to the third degree) fit the data extremely well but projecting the median gross to almost triple from 2016 to 2020 didn’t look right to me. So, based solely on the look of the plot I chose what seems to be a more reasonable model. I also tried the forecast option in the analytics tab but was only able to get the projection to 2017. Based on this I expect the median gross this year to be approximately $70 million.

**Question 2:** Which countries tend to have the longest movies? What is the most common genre in each country? Which countries have the highest rated movies? Who is the most common actor in each country? What is the highest grossing movie in each country?

This is obviously a question with many parts, partially because the data map functions in Tableau are really good, so just picking one question felt too easy.

There are a lot of stereotypes about what kinds of movies different countries make, so I was interested to see if any of them were borne out in the visualization. For example, I often think of Russian films as incredibly long (based solely on *Solaris*), and assumed that American blockbusters would all be approximately 90 minutes. At first, I was struck by how long Icelandic movies appeared to be, but then realized that there were only two Icelandic films in this dataset. After limiting the visualization to countries with at least 5 films (a completely arbitrary cutoff), most of the interesting effects unfortunately went away. India generally has longer movies than the rest of the world, which isn’t a surprise considering the reputation of Bollywood films, but everywhere else had a median length of around 100 or 110 minutes.

Next I added most common genre, median IMDB score, and most common actor for each country. Aside from the data manipulation, which took a very long time due to my lack of familiarity with Tableau, updating the map with these new variables was really easy. The genre and most common actor maps were unfortunately not particularly interesting, although there were some pleasant surprises (I had no idea the Channing Tatum was in so many Irish movies, or that Switzerland and Chile produce a lot of biography films). The IMDB score map was the most interesting, particularly the fact that Italian movies tend to be so highly rated, and German and Russian films tend to be less highly rated[[1]](#footnote-1).

Finally, I tried playing around with using multiple channels to represent multiple different attributes. Some of the data maps are not the best and I think try to convey too much information, but Tableau makes it very easy to experiment with various mappings.

**Question 3:** What are the most common first words in movie titles (aside from a, the, etc.)?

I wanted my third question to involve a new plot that I don’t get to make very often, so I tried to think of questions that wouldn’t involve boxplots, timeseries, or scatterplots. At first, I wanted to make a network of all the actors who had been in movies with one another but could not figure out how to do that in Tableau (see the limitations section). So, I thought it would be interesting to make a word cloud of all the first words in movie titles, in order to see which ones tend to come up frequently.

Unfortunately, I think this question would be more interesting with a larger dataset, as there wasn’t a huge amount of repetition, which doesn’t make for a very good word cloud. However, filtering the words to those that appear more than once helped a little bit. I also excluded “a” and “the,” as those dominate the cloud when included (although this was pretty entertaining, so I included the full plot in my story). There are some words that are simply part of a popular series, like Transformers or XMen, but I was surprised how common it is to start titles with a color (red, green, blue, or black) or the words “American” or “my.”

One thing I wasn’t able to figure out was how to adjust where certain words appear in the cloud. I found that my eyes tend to be drawn to words that make grammatical sense together (e.g. “kill something Cinderella” near the center of the cloud). Also, in previous drafts there were two numbers right next to each other, which made parsing the plot difficult without being able to interact with it. Finally, now that I’ve made one, I’m not sure that word clouds are the best way to display information unless the point is to show that there are obvious outliers. It’s particularly difficult with long words, which look bigger than short words even if they are technically on the same scale.

Part 2

One of the more interesting plots that I made on accident while experimenting with Tableau was the treemap of total gross split by film language. I certainly expected that English language movies would have the biggest share of the market, but the plot shows the scale really well. English language films take up so much of the market that you can’t even see most of the other blocks. It was also sort of interesting to see the market divided by genre in a similar treemap, although I think that a bar chart would get the point across more effectively.

I also tried to make a violin plot in Tableau, which turned out to be a little bit beyond me for now. However, in preparation I made a pretty simple boxplot of film budget by content rating. I thought it was interesting that median budget was highest for G-rated movies and decreased as the age restriction increases. Also, PG movies are the most variable in terms of funding, and R-rated movies are the least.

Part 3

Overall, Tableau has a lot of excellent features, and seems particularly well-suited to those who need to analyze data but perhaps don’t have any programming or statistical background. The fact that it allows the user to drag and drop dimensions and measures, and then automatically creates and suggests well-designed plots based on those choices makes it a much more useful visualization tool than something like Excel. Also, the video tutorials and other online resources make it relatively easy to learn the basics of the program.

One of the cleverest features is the ability to assign dimensions and measures to be “columns” or “rows” like in the Polaris paper. This is a great way to start a visualization design, because it’s intuitive, and generally follows how one is taught to make a graph from early in life (i.e. first you determine what goes on your x and y axes).

Very easy to split string fields in Tableau! Calculated fields are more confusing though. Also, impossible to fit a model to categorical data.

One major limitation is that you can’t make network graphs directly in Tableau.

1. Apologies Carsten, but I’m assuming that this is due to limitations in the data, because the German movies I’ve seen are great. *Western* (2017) is an all-time favorite of mine. [↑](#footnote-ref-1)