Jacob Zimmer

9428 0978

Music by the Numbers

**Motivation**

In this project, I set out to explore the data behind one of my biggest interests: Music. As someone who listens to " a stupid amount of music" (direct quote from a friend after Spotify's 2017 year-end account statistics were released), I knew that this would be an area that was both compelling for me to explore, and allowed me the freedom to pursue meaningful and interesting questions. In particular, I wanted to focus on live music, as I have recently begun to attend (and plan to continue attending) a large number of live shows. In the context of this class, I felt that I could formulate a strong project by studying the intersection between concert sales and overall music sales. As I thought more about the relationship between these groups of data, I identified several questions/ideas that I was interested in exploring:

1. What is the relationship between chart success and tour success? Are top charting artists more likely to have financially successful tours?
    a. Hypothesis: Top charting artists will consistently have more financially successful tours than low charting artists.
2. Do certain genres tend to have more successful tours? Which genres are the most/least expensive for fans to attend?
    a. Hypothesis: Pop tours are the most successful and most expensive, followed by Rap/Hip-Hop.

3. Based on my hypothesis that top charting artists will have the most successful tours: Are tickets for top charting artists' shows consistently more expensive than average? Are low charting artist's tickets consistently cheaper?
    a. Hypothesis: Tickets for shows featuring top charting artists will consistently be more expensive than low charting artists.

**Data Sources**

This project featured the use of three distinct data sources: a collection of pdf files, a collection of URL's from which data was scraped, and an API. The pdf files, retrieved from concert trade publication website *Pollstar*, contained tour data for the Year-End Top 200 North American Tours for each year in the time period of 2013 - 2017. Within in these files were tables containing the following information: the Artist's name, the Overall Tour Rank, the Gross Millions of Dollars accrued by each tour, the Average Ticket Price, Average Tickets Sold per Show, Total Tickets Sold, and Average Gross per Show. Of these categories, the vital variables for my analysis were Gross Millions and Average Ticket Price. The URL's were located on the Billboard magazine website, containing a list of the names and ranks - both of which represented key variables for my project - of the Year-End Top 100 Artists for each year from 2013 - 2017. The API used in my project, the iTunes API, relied on the artist names retrieved from the Tour Data PDFs, as I made a call to the API for each artist represented on both the Tour chart and Top 100 chart for a given year in order to retrieve their associated genre. In total, this represents 66 records used/retrieved in my project (5 Pollstar PDFs, 5 Billboard URLs, and 56 calls to the iTunes API).

**Data Manipulation**

For the Pollstar PDF charts of the tour data, I started my data manipulation by feeding the PDF's into a PDF-to-CSV converter. The 2017 and 2016 conversions worked perfectly, but the 2013 - 2015 charts would not convert correctly (even across multiple conversion websites). The best I could get for these three charts was a conversion into a csv containing all the columns in the first column of an excel sheet. Fixing this issue was a long and intensive process, as I had to use guess-and-check, adding spaces between various observations in order to line them up well enough to use the "Text to Columns" feature in Excel. After this, I had 5 csv's in the proper structure to use the read_csv method of the Pandas library, reading them in as tour_20XX based on the year data that the DataFrame would contain.

For the Billboard Top Artist Rankings URL's, I compiled a list in Jupyter Notebook of each of the 5 URL's and created an empty dataframe called df_charts to feed the data into. From there, I created a for loop to loop through the list of URL's and scrape the Artist Name from the sites using BeautifulSoup. This processing step took a lot of trial-and-error in order to locate the correct parts of the html to grab (which took place in a separate python file - 330_final_project.py - located in my project zip file). After grabbing the names from each URL, I created a temporary list within the loop and populated it with these names. Since the artist's names were appended to the list in the correct order of their ranking, I then simply created a list of numbers from 1 to 100, as well as a temporary DataFrame, and fed the list of artists and numbers into the temporary DataFrame under the columns 'artist' and 'rank', along with a column for the year that they achieved that rank. Finally, I concatenated the temporary DataFrame onto the permanent one that I created prior to the for loop and reset the index. In this

process, I came across a few instances of missing data: namely, a the 2015 and 2016 Billboard lists were both missing a single artist, leading to errors in my assignment of ranks in their respective temporary dataframes. To correct this, I went back to my external python file and started printing each artist along with their expected ranking. From their, I compared artist ranks on the actual Billboard sites to the printed rankings resulting from my scraping, and eventually realized that the websites themselves are simply missing a rank - skipping from 87 to 89 in 2016 and 82 to 84 in 2015. Not knowing who truly belonged in these slots, I decided to simply assign these rows manually within the DataFrame with 'unknown' for the artist value in order to maintain correct rankings for the artists who followed them.

After the Billboard DataFrame was created, I then split it into 5 smaller ones - one for each year - and began individually exploring the 2 sets of data associated with each year before creating two combined DataFrame of all artists from each year who appeared on both the Billboard charts and the Pollstar Tour charts, both relying on grouping the rows by Artist name. In this combined DataFrame, I calculated possible statistics of interest for each artist, including: Pure Difference in Billboard and Tour ranks, the Absolute Difference in Rank, the Percent Change in Rank, and the Number of Years Appearing on Both Charts. After creating these new features, I had a the first of my two combined dataframes; the one featuring their total statistics (mainly relevant for the tour data). After finishing this DataFrame, I created another combined one featuring their average statistics (again, mostly for tour data, but divided by the number of times they featured on both lists).

In my code, the next step after creating these DataFrames was to have a for loop iterate through the averaged DataFrame, access the Artist Name column, and make a call to the iTunes

API to retrieve their associated genre before accessing the relevant part of the response and appending it to a list. Once the list was completed, it was used to populate a new column called 'Genre'. While this is the existing structure in my code, I actually did it at the bottom after performing some of my analysis; I realized after doing this analysis that finding out Artists' Genres would make for more interesting Data Exploration and Visualization. Using the exact code that is in my zipped ipynb file, I successfully performed this data manipulation before saving it to an external csv for use in Tableau. Only later did I decide to transfer this code to the same area as my other manipulations for the sake of continuity, as well as with the intention of adding it to my summed DataFrame. However, when I did this, the API code stopped working: through print statements, I determined that the API requests were working properly, but something was going wrong with the loading of the data before appending it to the list. I tried fixing this issue, but I do not understand why this is happening. Since I had already saved the DataFrame with the correct data to an external csv for use in Tableau visuals, and did not need the data for any of my visuals in the python file, I felt that this was only a small issue. I contacted Abhraneel to confirm and possibly get help on the issue. Currently, it is unfixed but not affecting my project.

**Analysis and Visualization**

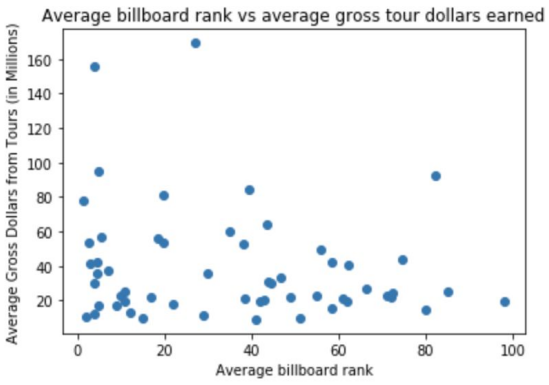
Using the separate datasets of tour and billboard data for each year, I created new DataFrames containing only those artists found in both of that year's charts. Then, I began by calculating simple statistics to try to get a feel for the data. One of my first routes of exploration was playing with the various difference in ranking stats that I had calculated, such as checking the normality of the absolute difference in ranking. However, I quickly decided to disregard most

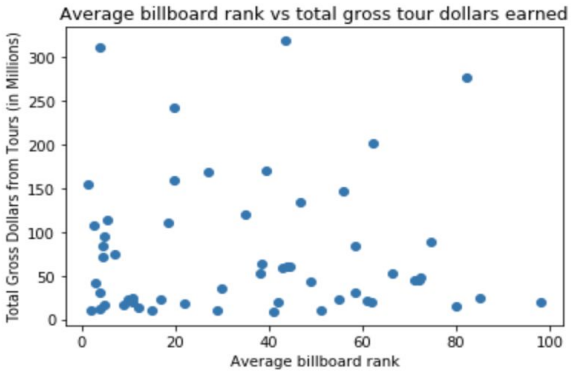
of these ranking difference stats; some observations were extremely large in both positive and negative directions, skewed by the number of appearances an artist made on the combined lists (For example, Luke Bryan's average rank difference of 31 seems ok, but Macklemore & Ryan Lewis's -86 difference is skewed by the fact that they only appeared once on the combined list).

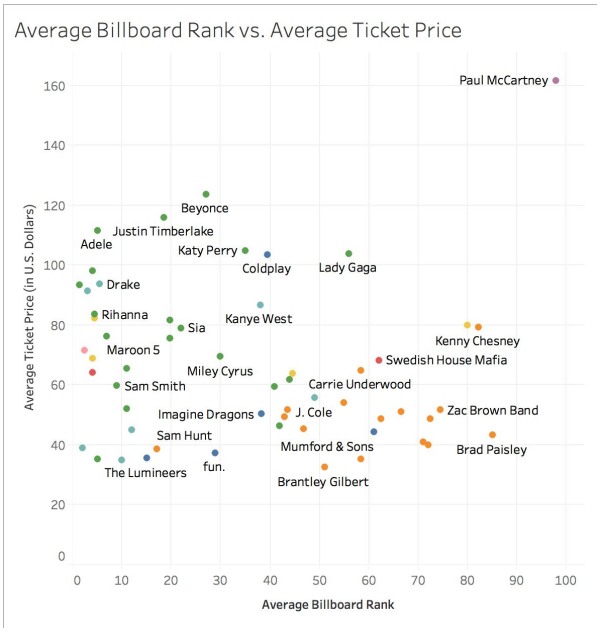
My next steps were to explore some simple statistics annual statistics to get a better understanding of my dataset. In particular, I focused on the yearly percentage of artists appearing on both lists, as well as the average billboard and tour ranks of these cross-sections of artists:

| | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|
| **Percent_artists_both** | 12.50 | 10.00 | 9.000000 | 8.500000 | 9.500000 |
| **Avg_billboard_rk** | 38.96 | 41.45 | 39.500000 | 34.705882 | 36.105263 |
| **Avg_tour_rk** | 40.08 | 44.15 | 32.611111 | 25.352941 | 40.842105 |

As it turned out, not many artists featured on both the tour and billboard charts; furthermore, artists on both lists tended to place higher on the billboard charts than the tour charts. This suggested to me that being a top touring artist was only loosely related to doing well in the overall music charts. To further explore this, I decided to plot the average billboard rank of artists against 3 indicators of tour success: average gross millions of dollars earned (per tour), total gross millions of dollars earned on tour, and average ticket price.



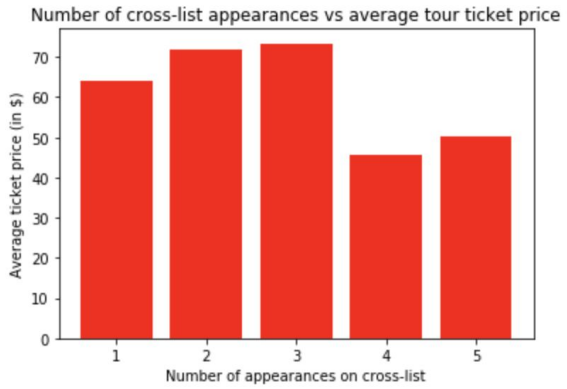



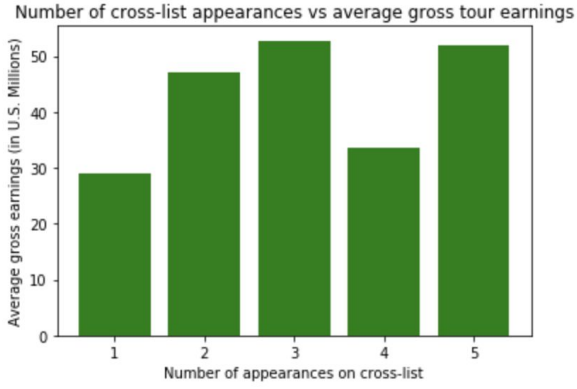



Looking at these charts, we see that there is very little coupling of billboard ranking with touring success. Surprisingly, many of the top artists in total gross tour dollars rank at 40 or lower; when adjusted for the number of appearances on both charts, top 40 billboard artists place only marginally better than sub-40 billboard artist (in average gross tour dollars). Clearly, this shows that lower-charting billboard artists are tend to go on tours placing in the top 100 for North America than their higher-charting billboard counterparts - likely implying that lower-charting billboard artists simply take part in more live performances. Yet, while the edge in average gross tour dollars certainly favors high-charting billboard artists, the results are still unexpectedly close. Further displaying this lack of separation of high and low billboard artists is the respective ticket prices of their shows, as there seems to be little-to-no correlation between billboard ranking and ticket price. Ultimately, this razor-thin difference leads me to deny my hypothesis that higher charting billboard artists have consistently more successful tours, and instead conclude that being a top-charting billboard artist has little bearing on touring success.

After discovering that top-charting billboard artists did not, in fact, have significantly more successful tours, and seemed to have fewer live performances than lower-charting artists, I set out to explore the relationship between the overall number of tours and the financial success of these tours. In Questions 1 and 3 that top billboard artists would have more successful tours

and more expensive concert tickets. While this baseline assumption was ultimately incorrect, I had some new questions to explore based on my new assumption that these same artists had fewer concerts than lower-charting artists: (1) Did the number of tours that an artist went on affect tour prices? Logically, if a fan has more opportunities to see an artist live, the less they would be willing to pay for a single opportunity. (2) If offering more concerts indeed devalued ticket prices, what would be the optimal number of tours for maximum financial success? By comparing cross-list appearances with ticket prices, I found that my first new hypothesis was

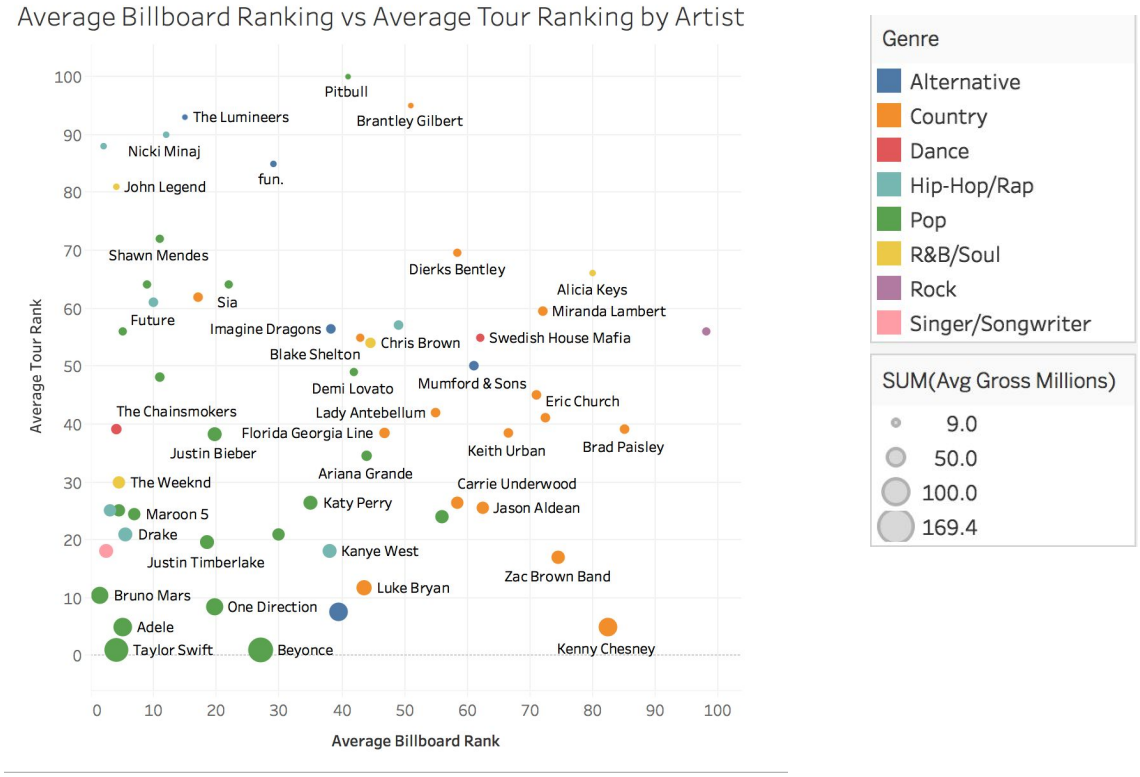


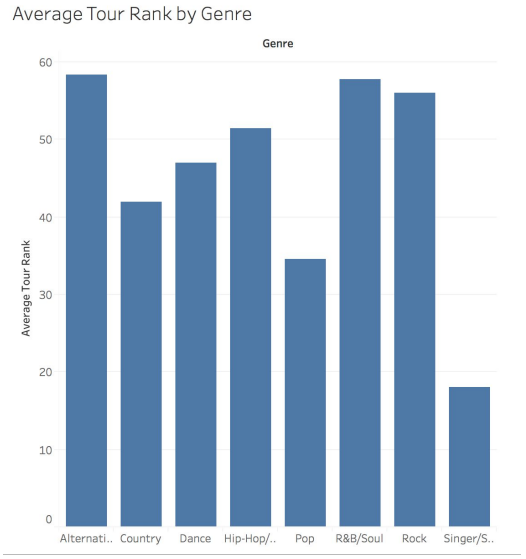

only mildly correct; ticket prices actually rose with the number of tours until the 3rd such offering, after which ticket prices fell sharply. Additionally, the comparison of cross-list appearances with average gross earnings revealed that the optimal number of tour offerings was 3, with 5 following closely behind. Taken in combination with the previous graphs comparing billboard ranking and ticket prices/tour gross, these visuals suggest that, for high-charting billboard artists, 3 tour offerings is the optimal balance for high ticket prices and maximum selling opportunities, while - for lower-charting billboard artists - 5 offerings is the optimal number of concerts (since that supplies the most opportunities for selling tickets).

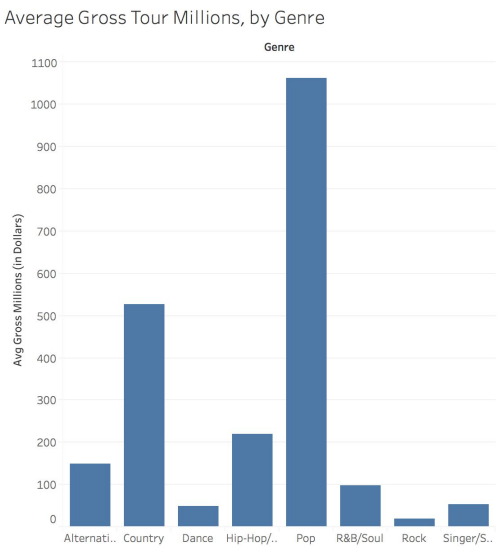
In pursuit of answering my final questions - whether certain genres have more successful tours and more expensive tickets - I decided to send my averaged cross-list DataFrame to an

external CSV for use in Tableau. To begin answering part one of the question, I created a scatter plot of the artists' billboard ranks against their tour ranks, with each point demarcated by the artist's associated genre (as catalogued by iTunes).

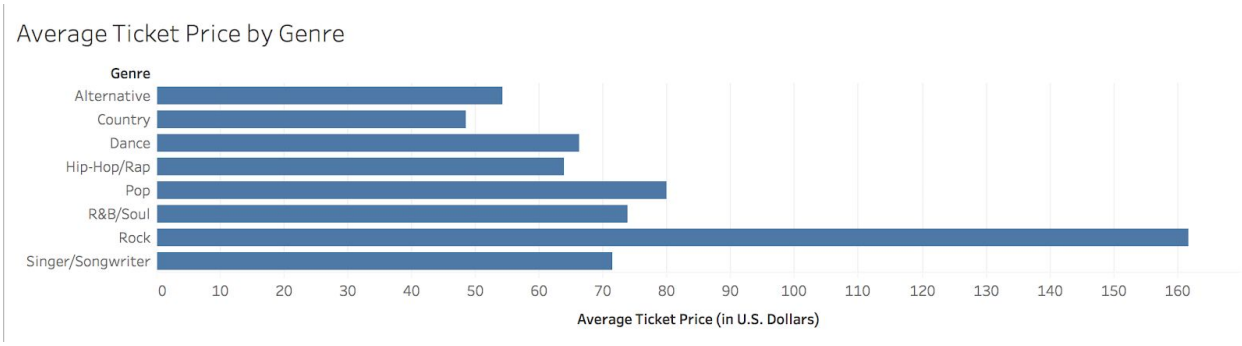



Pop artists overwhelmingly dominated other genres in all aspects; Pop artists were most inclined to chart well on both the billboard and tour ranks (excluding Singer/Songwriter,



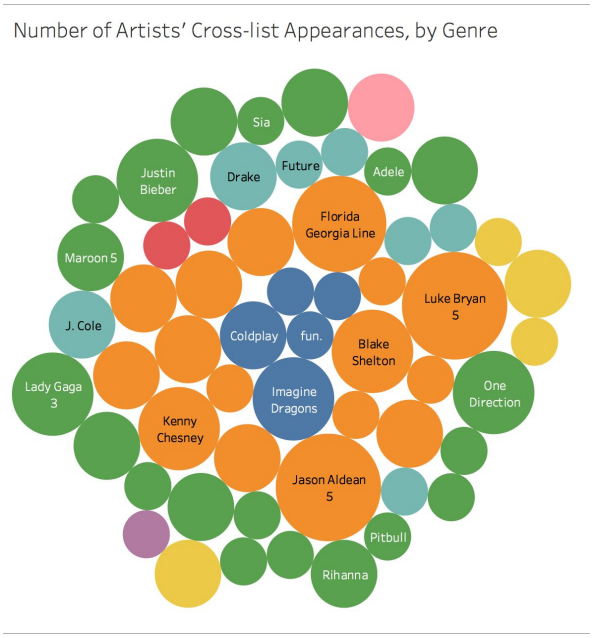



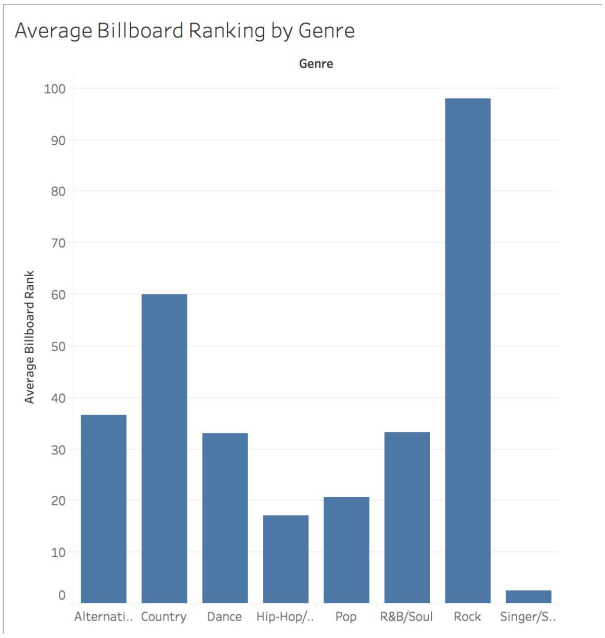
a genre populated only by artist Ed Sheeran), while simultaneously raking in millions of dollars in average gross tour dollars more than the next highest genre. The runaway second-place genre for average gross tour millions was country, despite consistently placing on the low end of the billboard ranks - further cementing the conclusion that billboard rankings is a poor indicator of touring success. With that, we can conclude that the certain genres are indeed more successful on tours than others, with pop and country taking a clear lead.

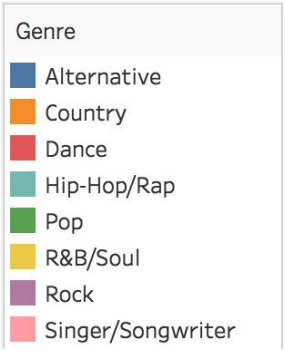



As for the second part of Research Question 2 - whether certain genres tended to have higher ticket prices, I simply graphed the average ticket price of each genre. While Rock appears to have an enormously large ticket price, this is the result of Paul McCartney being the only artist in this genre to place on both the billboard and tour charts. The highest priced genre besides Rock, Pop, is unsurprising, as it would make logical sense that the genre with the highest Billboard and Tour ranks would also have the highest ticket prices. However, Country interestly placed as the lowest average ticket price, despite being the second highest average grossing genre for tours. This finding calls back to some of our previous exploration, namely the conclusion that, for low-charting billboard artists, a high number of concerts is the optimal financial strategy.

This conclusion is further supported when considering the above graph in conjunction with the graphs of the number of total tours by artist and genre and the average billboard ranking of each genre (below).




With the exception of Rock (again, an outlier with a single data point), Country placed as the genre with the lowest average billboard ranking. Simultaneously, all of the artists with the maximum (in this dataset) number of cross-list appearances were Country artists. This result would require these artists to have gone on 5 tours in as many years. While there is a possibility that the culture of Country music could act as a confounding factor in the number of tour offerings of these artists, these findings generally support my earlier conclusion that higher-charting billboard artists are less inclined to go on tour often, while lower-charting billboard artists are more likely undertake a large number of live performances (reflecting our conclusion that the optimal number of tours for high-charting billboard artists is 3, while the optimal number for lower charting artists is 5).