Ryan James Calabio

Prof. Smith

D214 – Data Analytics Capstone

5 May 2025

ANOVA Analysis of Length of Music Before & After Spotify's Release

The problem for this analysis is answering the question of whether or not songs before and released before and after the release of Spotify, have statistically significantly changed in average length. The hypothesis I have created for this is that the length of music has on average been reduced. The null hypothesis for this analysis is that the mean of duration for songs before and after have not statistically significantly changed since the release of Spotify. If I get a p-value of over .05 from my ANOVA analysis, then I have proven my null hypothesis as false, and my original hypothesis is then given support. Then I measure on a graph to see if there is a reduction in average length before and after the release of Spotify. If it goes down, and the ANOVA shows that the difference is statistically significant, then my hypothesis that songs have on average from before and after Spotify's release data been significantly reduced is further supported. This does not however, prove causation. It only shows that there is a correlation. The scope of this analysis is also only limited to songs within the Spotify platform, so it does not prove correlation for all songs that are released, just songs that are on Spotify.

For my data analysis, I got my dataset from Kaggle. This dataset provides metadata for music on over 1.3 million songs within Spotify. To start, I looked at the older songs and noticed that some of them were not set to the right year. I removed them until the older songs were the proper year.

```
df.loc[df['year'] == 1900]
```

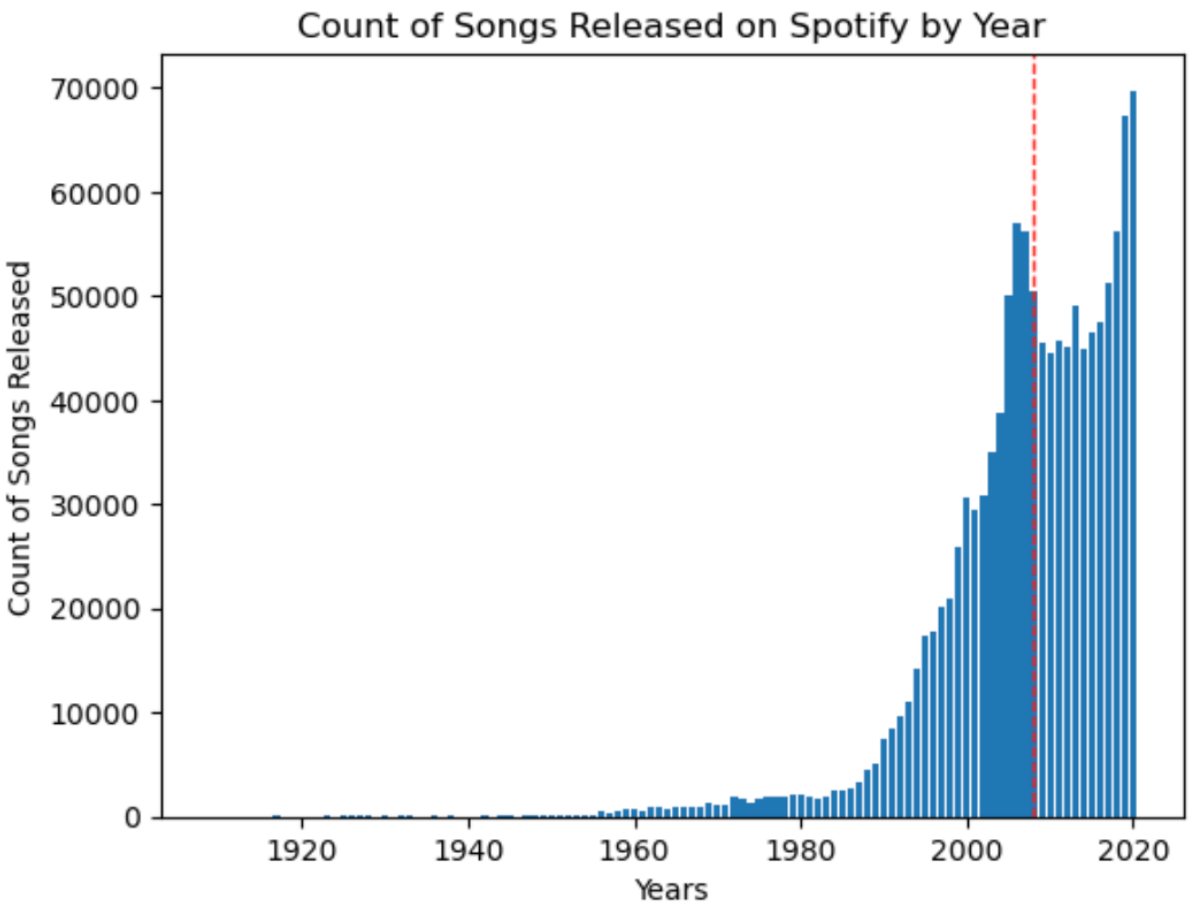
	name	album	artists	duration_ms	year	release_date
450071	Arabian Waltz	Arabian Waltz	['Rabih Abou-Khalil']	493867	1900	1900-01-01
450072	Dreams Of A Dying City	Arabian Waltz	['Rabih Abou-Khalil']	730667	1900	1900-01-01

I also checked for nulls within the dataset and removed those.

```
df.loc[df['year'] == 0]
```

	name	album	artists	duration_ms	year	release_date
815351	Jimmy Neutron	Optimism 2	['iCizzle']	183000	0	0000
815352	I Luv You	Optimism 2	['iCizzle']	145161	0	0000
815353	My Heart	Optimism 2	['iCizzle']	176561	0	0000

Then once the data cleaning was complete, I visualized the count of songs within the dataset by year. I also put in a red line to signify the year that Spotify was released.



Then I went through and found the oldest and newest songs within the dataset.

```
Oldest:  1909
Newest:  2020
```

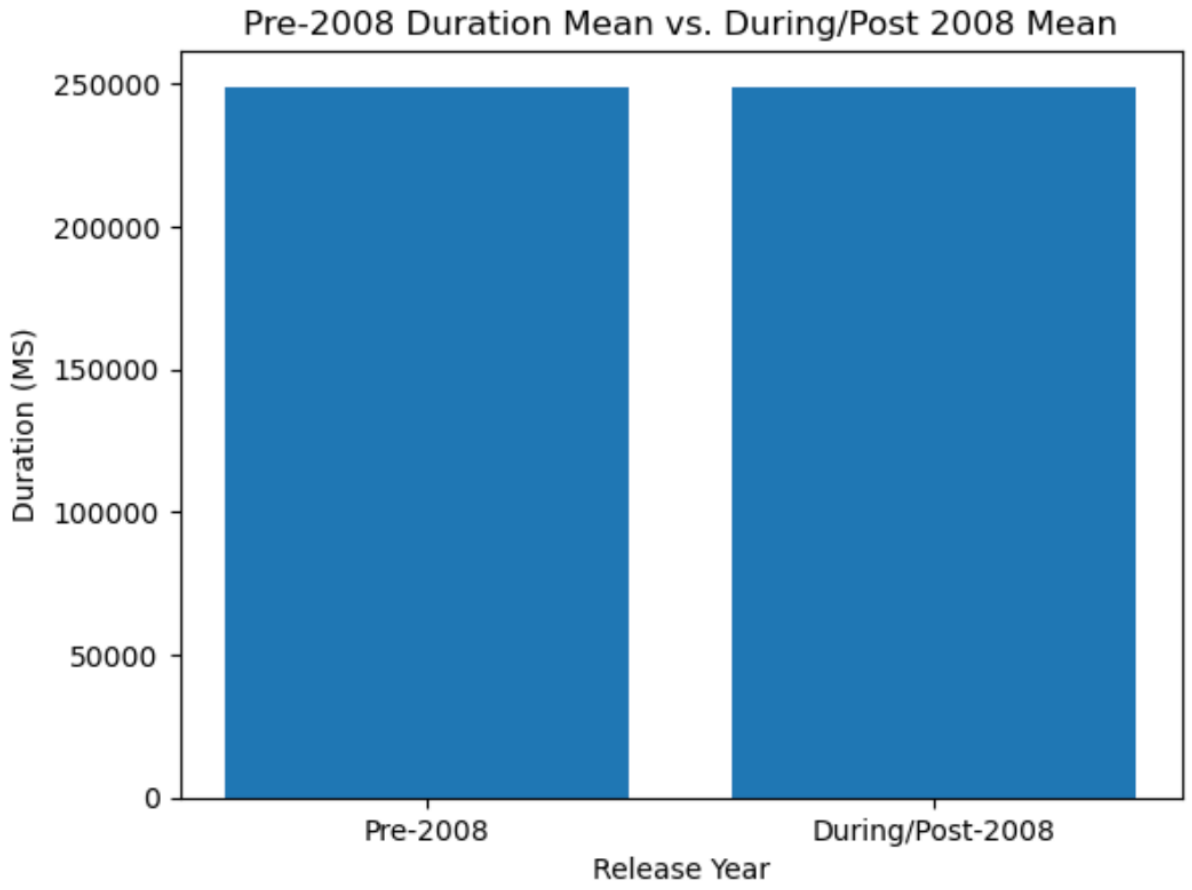
The last metric I look into for my exploratory data analysis was the count of songs before 2008, and the count of songs released during and after 2008.

```
Pre-2008 Count:  540129
2008 & After Count:  663853
```

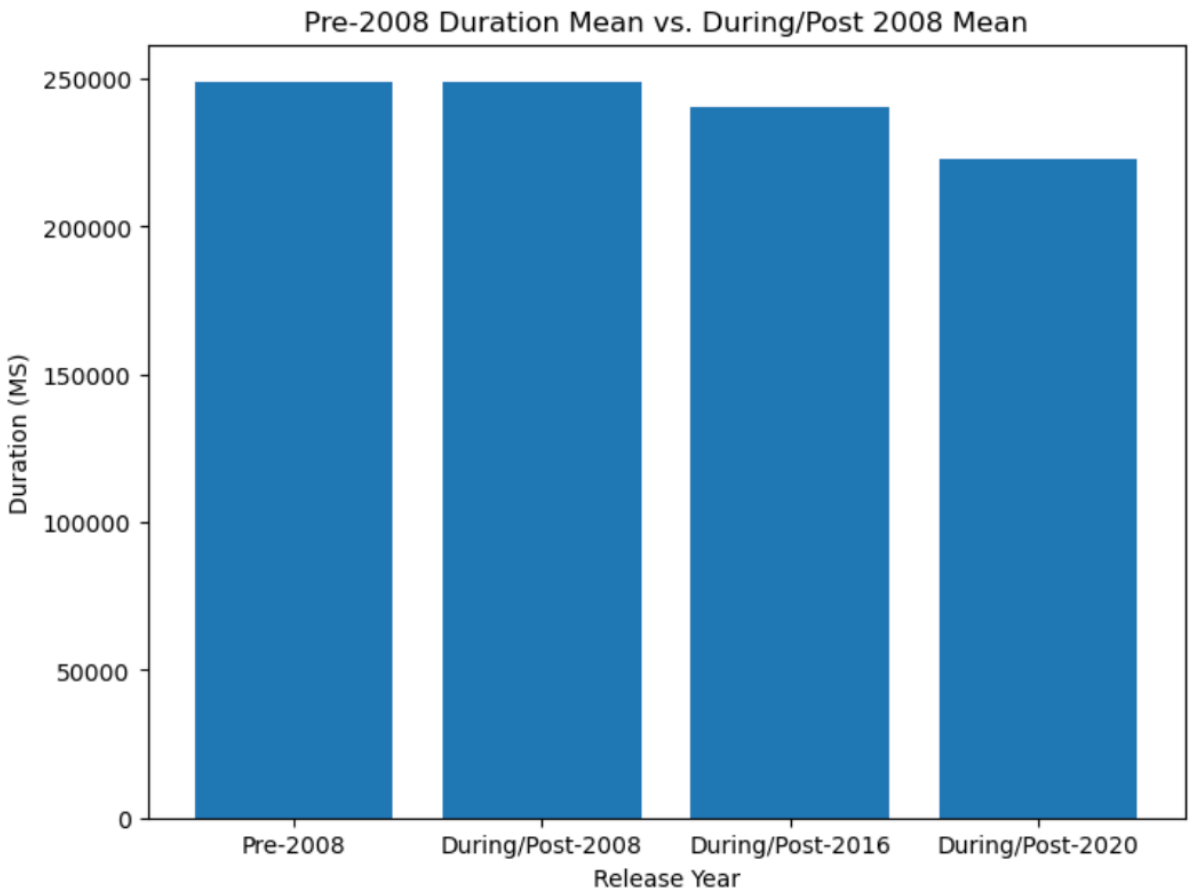
After that, I split the two timeframes into two separate data frames. I inputted this into an ANOVA function from Scipy to identify if the null hypothesis which said that the means were the same had statistical significance. The results of the ANOVA analysis were the following:

```
ANOVA Results:
 F_onewayResult(statistic=1.9157774994603458, pvalue=0.16632315098
948963)
```

The p-value for the ANOVA analysis is 16.63%. This is much higher than the 5% limit for the p-value required to say that the mean is the same with statistical significance. After showing that the means are different, I created a graph to show the difference in the means:



It is a small difference, but we have proved it is statistically significant using the ANOVA analysis. Therefore we can say that with statistical significance, the average song length on Spotify has reduced from the year spotify was released. Another interesting part of the analysis I wanted to add on was to see if the trend continued on past 2008. This graph shows during and after 2016 and during and after 2020:



We can see from the graphs here, that the trend of average duration of songs reducing in the duration seems to continue as you zoom into more and more recent timeframes.

When it comes to limitations of the techniques and tools, there are few important things to note. For instance, we should note that this does not prove correlation for songs outside of the Spotify environment. This dataset only includes music that is contained within the Spotify environment. It is important also to note that no causality was proven in this analysis, only correlation. From the last graph that was shown, one can see that the trend of songs reducing in duration was happening irrespective of Spotify's release date. Whether or not Spotify's release date accelerated that trend could be up for debate and further analysis. The tools I used for this were pretty simple: matplotlib for graphs, scipy for ANOVA, and pandas for data frame manipulation.

The proposed actions coming from this analysis depends on the position of the person reading the analysis. For a musician that produces music for Spotify, this finding could imply that music consumers' desire to consume short music is growing, and as a result the musicians are responding in kind. This could help to justify the creation of shorter and shorter music. It could also be used by a musician to make a statement about going against the trend, since creating a song with a longer than average song would make them stand out. When it comes to companies like music production companies and labels, the trend could indicate the consumers are looking for shorter music, and that they should respond in kind by assisting artists in creating shorter duration music. For data analysts, further action could be taken by further divvying up the time frames into smaller sections and then showing the trend from there. Looking at other metrics like genre and the change in song duration by genre could provide more specific analysis when it comes to trends in different sections of the music industry.

The benefits of this analysis will differ based upon who the reader is. For musicians, it provides some insight into the shifting landscape of the digital music industry. Something that would allow them to pivot their strategy based upon the trend of both music creators and music consumers. For music companies, this analysis provides benefit in showing where things are trending so they can position themselves to be in a more profitable position. For data analysts, this creates a foundation for deeper analysis to be done on the trend of shortening content in the entertainment industry, specifically within the music section of the entertainment industry.