**Profanity in Lyrics – Top 100 Songs in U.S. Billboard**

<https://drive.google.com/file/d/1K6XuKQmbCmlWC3OLUkMDzA7jsiU6hccE/view?usp=sharing>

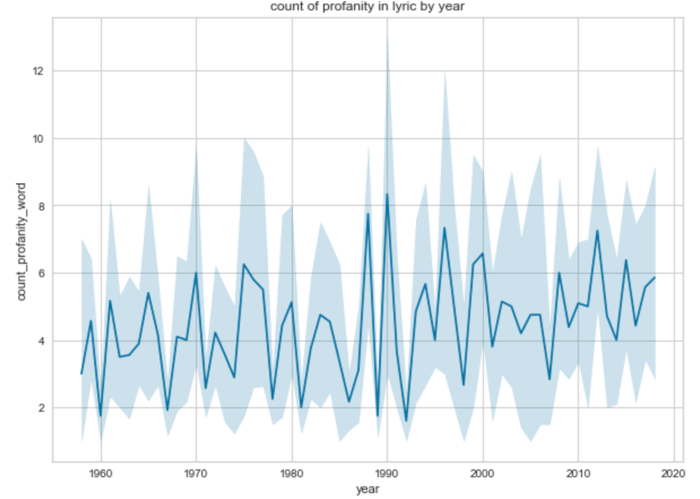
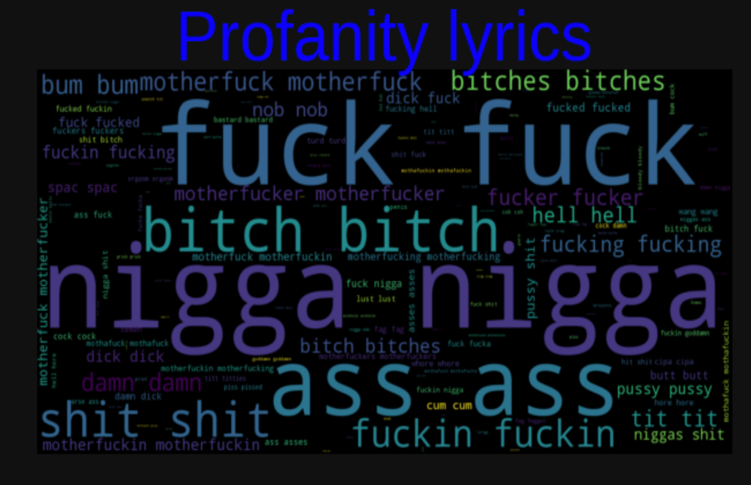
**WARNING: There might be some profanity words in the content, please do not continue read if you are “allergic” to those uncomfortable wordings in the graph.**

**Overview**

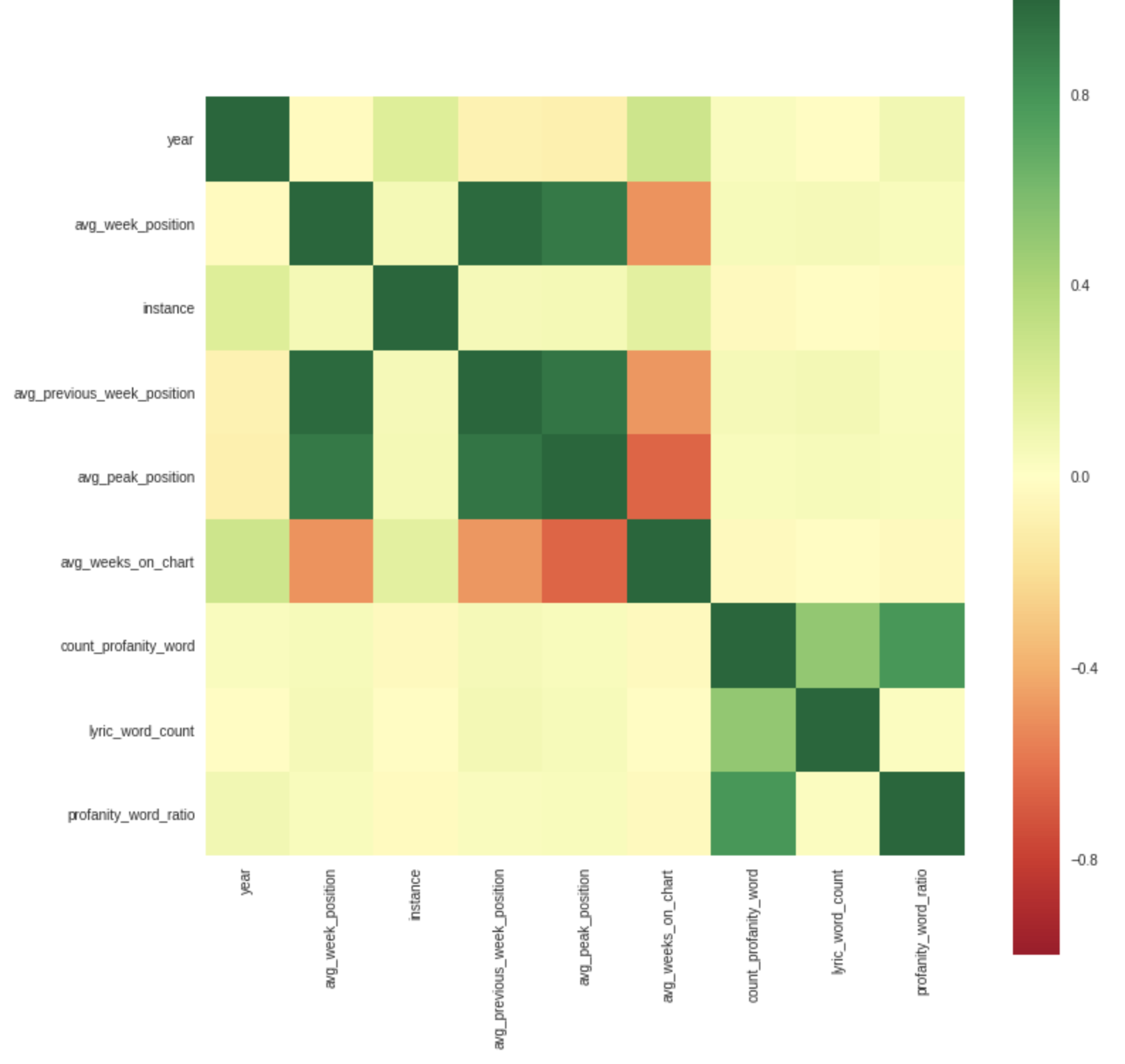
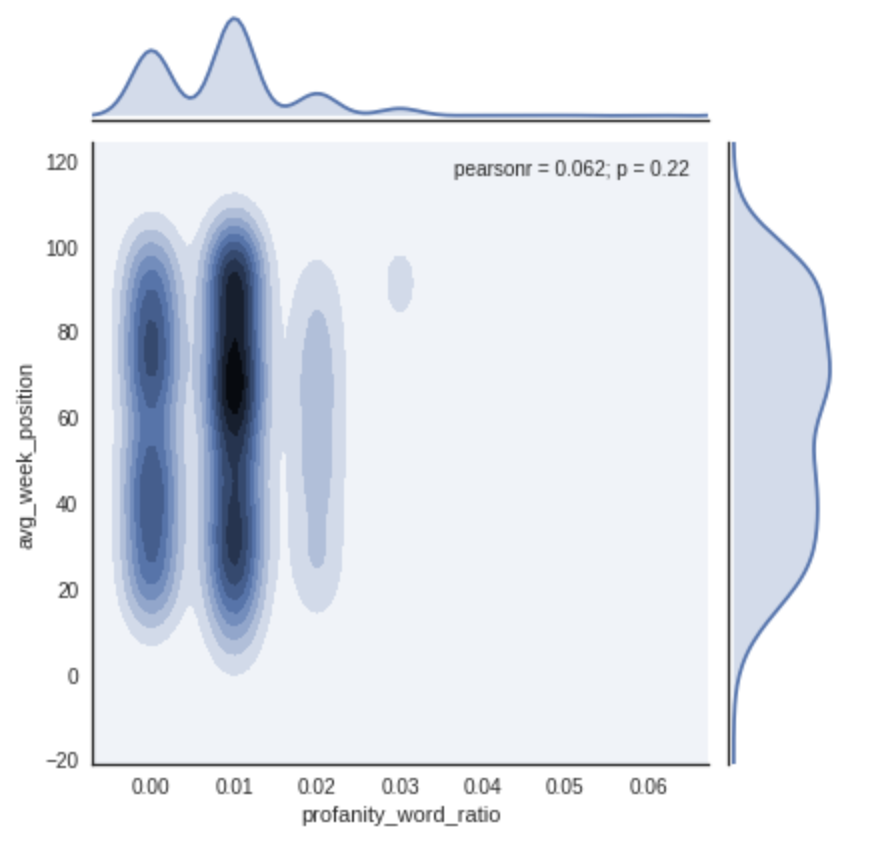
I have been a fan of U.S. Billboard for long since middle school. I noticed that many songs in the U.S. Billboard are full of profanity words. Over the years, there seems to be more and more songs that curse somehow in the lyric. That annoyed me, so I turn my attention to the U.K. Billboard. There are still profanity words in the U.K. songs but less awful than those in the U.S. Therefore, I decided to gather data on U.S. Billboard and explore how the U.S. top 100 songs on the Billboard are evolving.

**Techniques and Procedure**

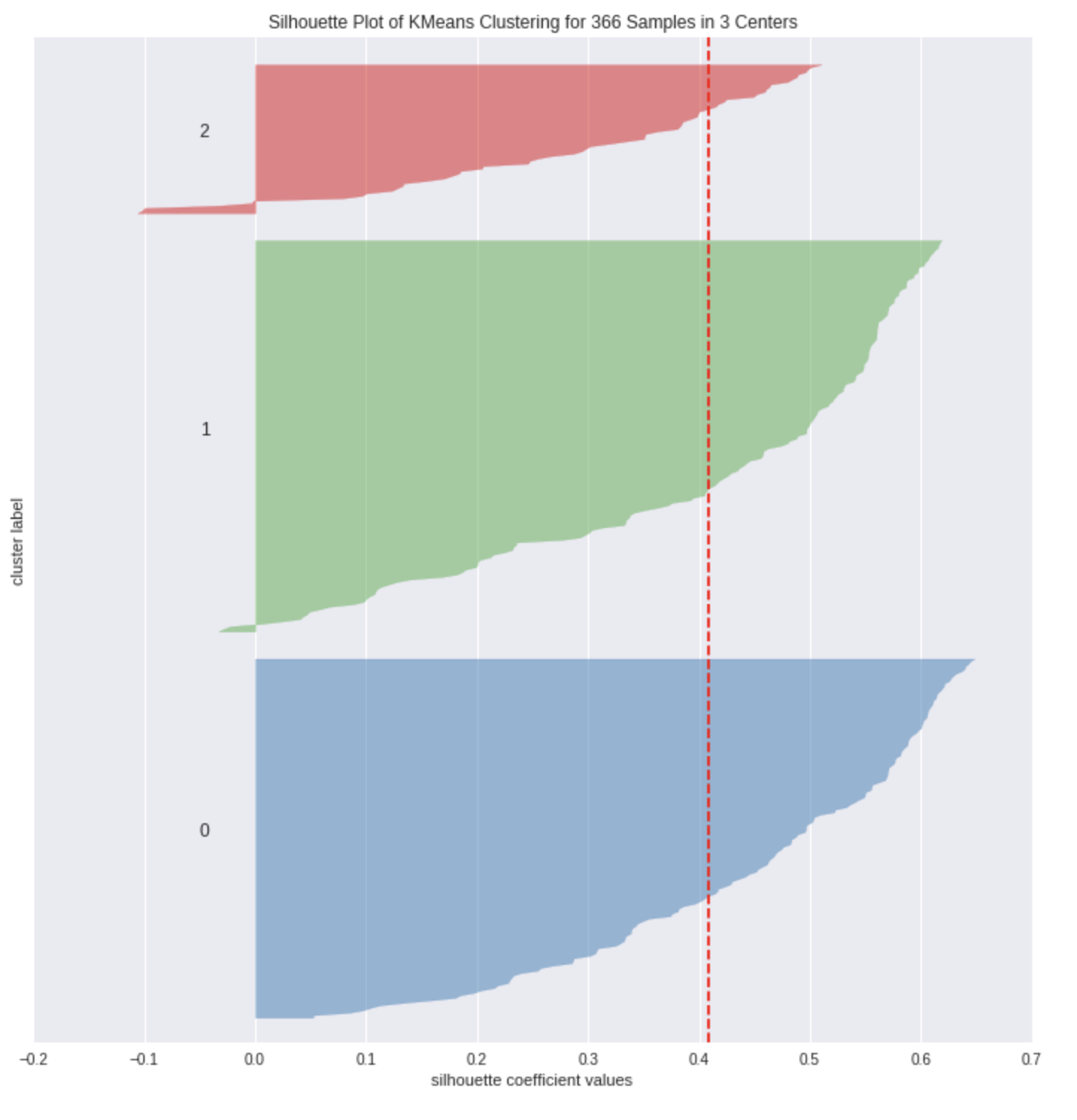
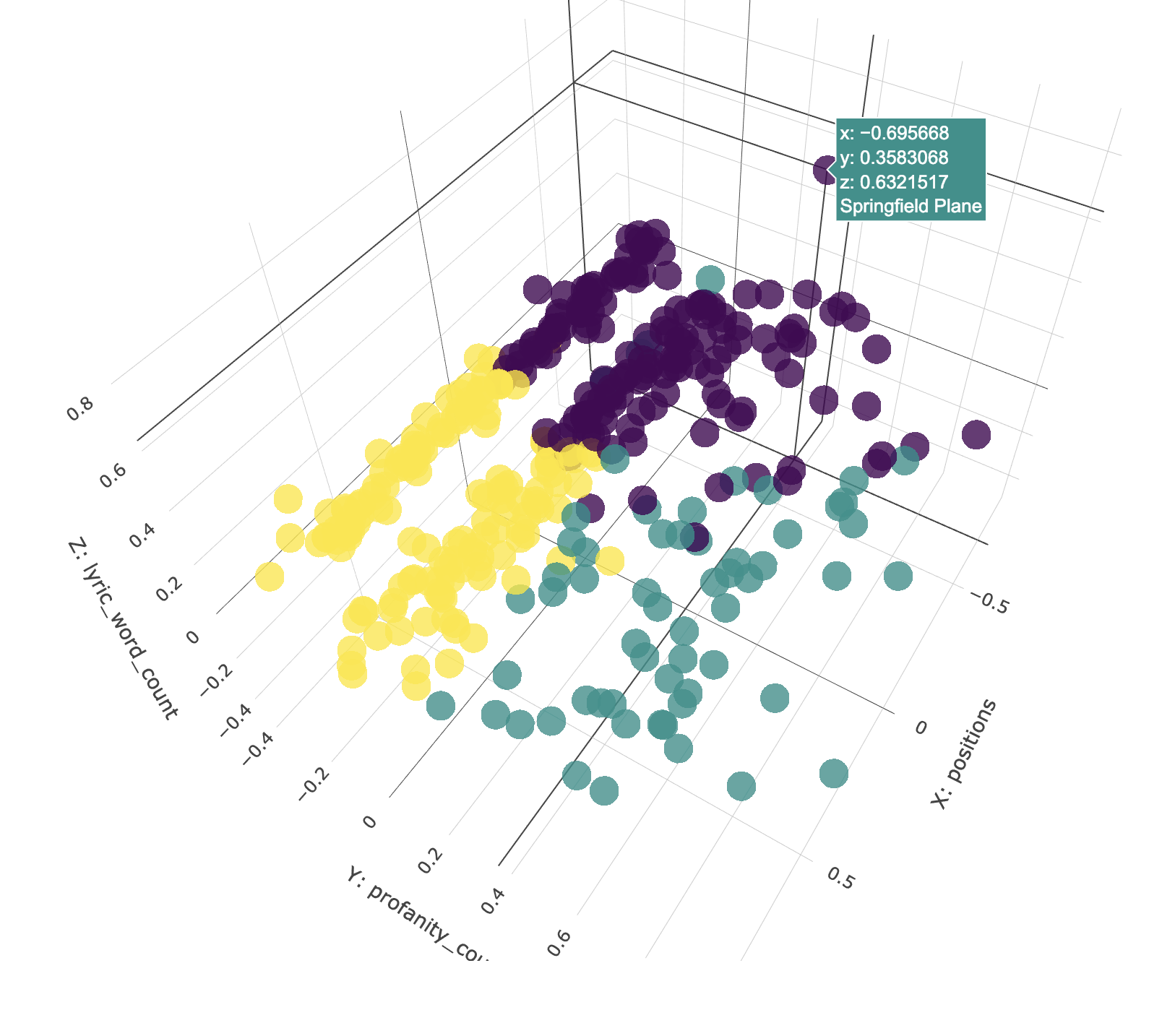
Steps

1. **Ingesting:** Collect a dataset about top 100 songs on U.S. Billboard by years and a list of profanity words. However, the dataset does not contain lyrics so I spent some time scaping lyrics through Genius API.
2. The scraping process was long and painful and experienced server timeout all the time but luckily I was able to collect lyrics for 1000 songs sampled.
3. **EDA:** I calculated the occurrence of profanity words in all the collected lyrics and the ratio of profanity word count against total word count in a lyric.
4. ****Then I visualize what are the most popular words and the most popular profanity words in the lyrics with word cloud:

1. Then I took a look at how the average count of profanity words in lyrics are by year and found out the amount of profanity words are slightly higher than a decade ago in general with an upward trend.
2. Next step is to find out whether there is a correlation between count of profanity word to the average peak weekly rank of a song. The answer is no. Therefore, I created a jointplot to find out whether I can classify songs by amount of profanity words in the lyrics. The answer is yes! There are three clear clusters.

&. Then PCA is used to reduce the amount of variables for clustering and make possible for a 3-D graph to visualize the clustering. The Silhouette scores show three clusters is acceptable and here it is: A interactive 3-D graph showing song lustering by profanity words in song lyrics.

**** ****

**Conclusion**

EDA shows some distribution of the profanity words in lyrics among top 100 US Billboard songs. With visualization tools like word cloud, we will be able to visualize those most used profanity words. PCA is very helpful to identify those highly correlated variables to shrink the number of variables in machine learning model while keeping as much as information as possible. Clustering methods like K-means can be useful to cluster.

Understanding lyrics and profanity words used in the profanity can be helpful when recommending songs to customers. Companies such as Spotify, iTunes will find this helpful to craft the user experience after understanding users preference on lyrics preferences.

Also, over time there are more profanity words occurred in the lyrics and teenagers are being exposed under this music environment. Worth introspecting.