1st chart

Talks about the topic and the main question that we are trying to answer. What data that we are using. Data World 2 files (<https://data.world/kcmillersean/billboard-hot-100-1958-2017> and , the web page with the current information, and the API

2nd chart

Show the 3 questions and talks about what we want to ask.

3rd chart

Talks about the data and how we massaged and cleaned it.

Data world had 2 files

Billboard-hot 100-1958-2019. It has roughly 320,000records. This represents the top 100 songs every week from 1957 to the end of 2019. This data has the information on rankings of the songs. There can be multiple entries since a song can be on the hit list multiple list. What we did is took the highest point that the song went to and used that for comparisons. While there is a peak position column, we needed to be careful as the last entry may not contain the appropriate information. We ended up using the min function to find the lowest ranking. The columns that have been highlighted are the ones that we have focused on

**url**

**weekid**

**week\_position**

**song**

**performer**

**songid**

**instance**

**previous\_week\_position**

**peak\_position**

**weeks\_on\_chart**

The second file is just over 28,000 songs with 10 spotify identifiers which we will use to see if there is any correlation to the ranking. We used the columns highlighted

**songid**

**performer**

**song**

**spotify\_genre**

**spotify\_track\_id**

**spotify\_track\_preview\_url**

**spotify\_track\_album**

**spotify\_track\_explicit**

**spotify\_track\_duration\_ms**

**spotify\_track\_popularity**

**danceability**

**energy**

**key**

**loudness**

**mode**

**speechiness**

**acousticness**

**instrumentalness**

**liveness**

**valence**

**tempo**

**time\_signature**

Once we pulled the data in and removed rows with no data. We pulled the peak position and used the SongID column to do an inner merge of the dataframes. We again cleaned up the rows if the merge brought in extra rows. Once that was done, we had a clean working dataframe.

1st question

This question was whether any of the attributes had and prediction for the ranking from a top 100 perspective. The clean dataframe was used to create scatter plots, box plots and calculate the mean and median for the different attributes of the songs. In the next few charts we will show what results we saw and conclusions accordingly.

2nd question

This was whether a certain genre was seen in the top 100. This was a little tricky from a data perspective. The genre field was not just a simple entry (R&B, rock, etc.). It was a text string which had multiple entries with unique classifications (progressive psytrance, derby indie, brill building pop, etc.). Overall there was roughly 1377 genres. To make our life easier, we pulled the list of genre from the dataframe using string manipulation and a dictionary to store the results and reduced the set to create a smaller set of genre to label the songs.

3rd question

Is around the comparison between the US and the world from a ranking perspective.