Aaron Wechsler - Data exploration
Keirnen Dossey - sketches, bug fixes, visualization creation

OPEN PROJECT - Group Project 3

Introduction:

Our group set out to create compelling visualizations based on an extremely large music related dataset that had over 1,000,000 data points. Each had several variables linked to it including general information such as song year and genre down to more specific information like acousticness, danceability, and valence. The goal here was to find a relative correlation between some of the categories that have shifted between 1950 to 2019. While the shift in types of music and music production has contributed to a large amount of this change over these 70 years. With the introduction of software based music production, the types and attributes of different music genres and song lengths can be seen correlated to trends and shifts in these categories that we had focused on, like acousticness, danceability, and valence. Our goal for this project is to accurately display these trends and shifts over this time frame.

Data exploration:

Finding data set:

For our final group project we were assigned to create a story by visualizing data using data sets regarding music. Initially our first data set consisted of over a million songs with information about each one from the million song dataset website. Working with a 280 gb data set was not an option for us so instead we decided to use a sample of 10000 songs that was auto generated from the website. After a bit of tweaking and aggregating data we realized that this dataset for reasons unknown would not work with Altair. Despite debugging and adding new

column names nothing seemed to work so we had to switch. Soon after we quickly moved on to finding another data set which we found on Kaggle. This data set was the top songs from 1950 to 2019 based on genre. There were multiple songs per year per genre along with over 20 other variables per song. Variables like danceability, loudness, energy, violence, obscenity, e.t.c were all ranked from 0-1. 0 being very low on the scale and 1 being high, for example an electronic song can have very high energy close to 1 but a very low acoustic score of 0.

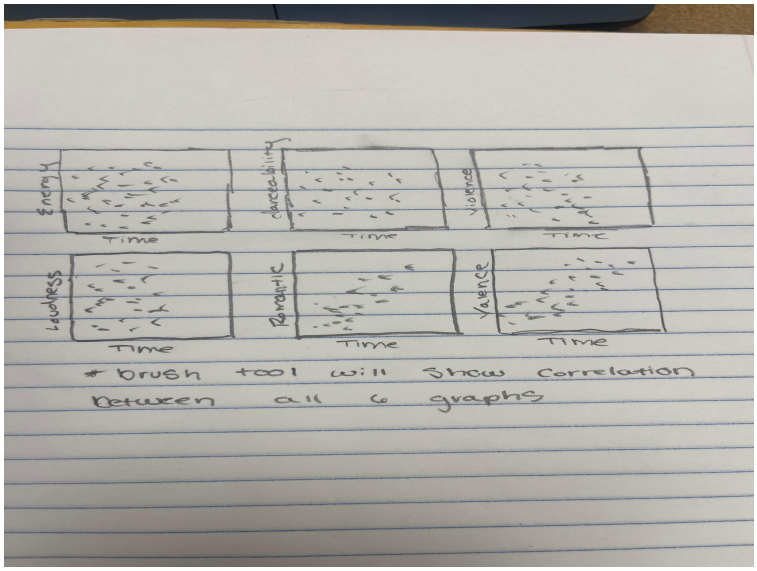
Data aggregation/filtering:

This data set was very large with over 25000 rows, at first we wanted to only select a few genres to show to make the data easier to work with and visualize. However even then there were way too many data points being shown on our graphs mixed with slow compile times made us go a different direction. Instead of only having a few selected Genres we used all of them which allowed us to draw more conclusions about trends. Using a lambda and sample function we selected 3 random songs per year per genre to make it much easier to work with and still have enough variety. Using a randomized selection means each time the code is run different songs are run however every time our group ran it our trend lines were still relatively the same. Having a randomized new data set also removes all biases to our data selection and visualization.

Sketching Process:

Before we started writing code to produce visualizations we began with a sketch, done on paper, to give us a rough idea of what we wanted our final product to look like. We came up with the idea to have the majority of our visualization be made up of smaller scatter plots that showed

the relationship of a bunch of different musical attributes (i.e. danceability, loudness, energy, etc) and their change over time (1950 - 2019). We knew that we wanted all of these scatter plots to be connected by a brush tool, so when you selected an area on one chart, it would highlight the corresponding data on the other scatterplots. We figured that this would be a good start to showing our data and answering our question. The feedback we got on our sketch recommended that we add more supporting visualizations outside of the scatter plots. We hadn't decided which kinds of graphs to use at this point but started creating our visualizations for the scatterplots.

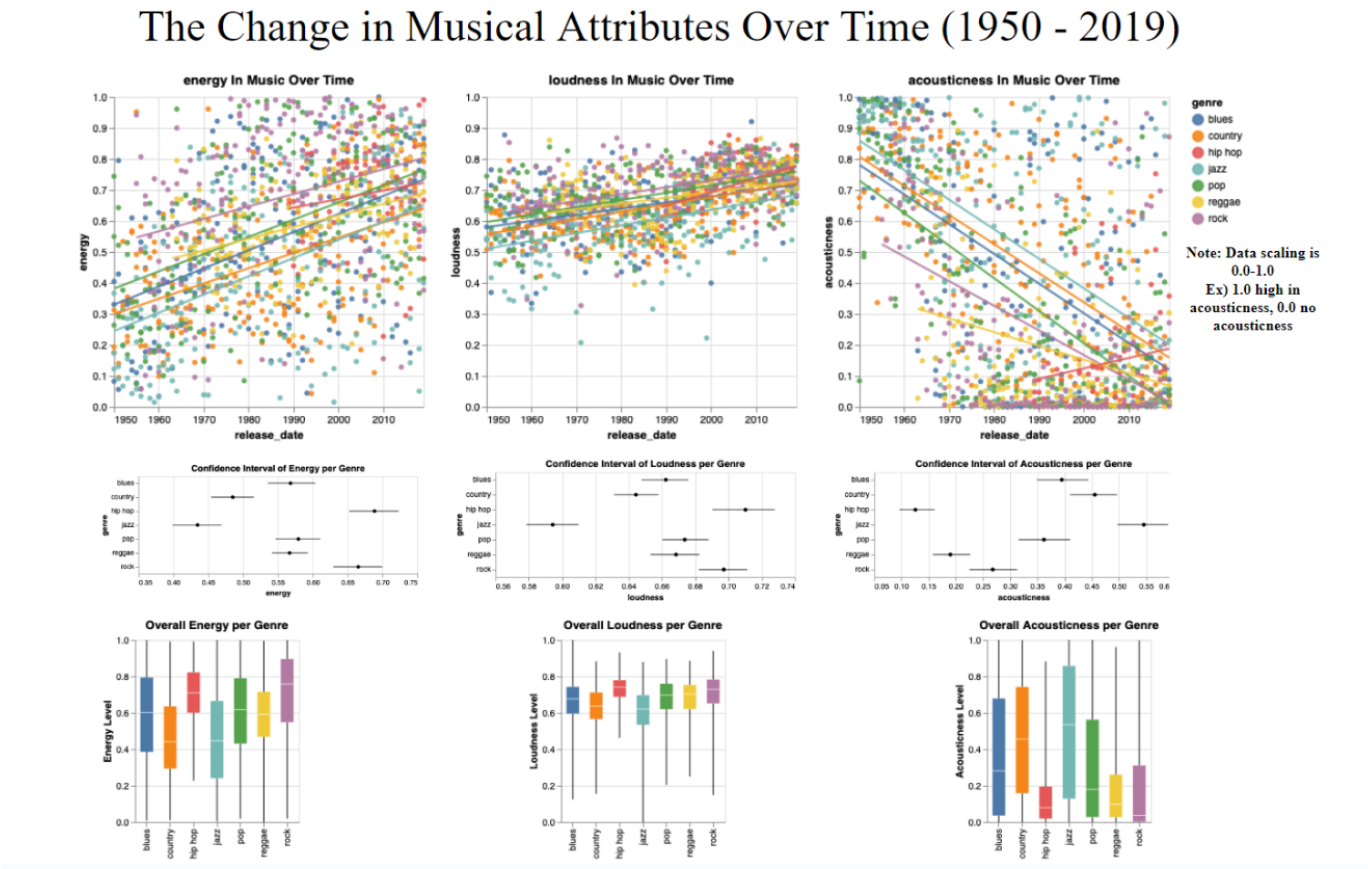


Visual Creation Process:

The first step to the visualization creating process involved preparing the dataset. This phase was more time-consuming than anticipated, as it required cleaning, normalizing, and structuring the data to suit our visualization needs. The first problem we encountered while doing this was the amount of blank data in the data set. There were a bunch of “NaN” values in the data set that would interfere with our ability to get an accurate analysis of the data. We solved this problem by writing a function that goes through the csv file and removes any missing values

from the data set. The second major problem we found was the sheer size of the data set. Altair will encounter problems when it tries to run data frames larger than a certain threshold. To help mitigate this problem we decided to take 3 data points, per genre, for each year and average them and plot those. This helped lower the amount of data that Altair had to handle. After cleaning the data and writing a function to take the average of each genre per year, we started off creating the scatter plots. The biggest error/problem we encountered while making the scatterplots was that the regression lines had trouble lining up with the drop down menu. It took a couple of hours but we managed to make it work. We also realized during the process of producing these scatterplots that a lot of these musical attributes didn't have any correlation with each other. We managed to find 3 variables that had a reasonable correlation and chose those 3 for our final visualization. After creating the scatter plots we experimented with different types of graphs to support our visualization. We tried heatmaps but found them to not display any new or supporting information efficiently. We tried line graphs but found those to also not be efficient at supporting our main visualization. We settled on a confidence interval graph using a box and whisker plot. We thought this would be a good way to show the statistical connection between these variables. We also made the other box and whisker plots showing the average values for each of the three musical attributes. After combining all of these visualizations together we finally had our slide to present at demo day.

Graphs and Demo Day:



Here is our demo day slide deck. In addition to these visualizations we also had an interactive graph. We have three columns of three graphs that are each relating to a different variable of the dataset: acousticness, loudness, and energy. They are graphing data about each of the 7 genres which are blues, country, hiphop, jazz, pop, reggae, and rock. Our first set of graphs were scatter plots that had the variable on the x axis and the release date of songs on the y axis. There was an extremely large amount of data points that made it hard to understand. On top of the fact that we color coded each of the genres and also cut down our data points by a significant margin. Even still it was hard to understand so lastly we topped them off with some regression lines to follow the trends of the data. Although these were very clustered in parts I think they were terrific initial representations of the data that made it quite easy to understand the other two

sets of graphs. The next set of graphs had to do with confidence intervals. This is the relation that each topic had to the selected variable. Although this graph was quite straightforward it also felt necessary for us to include. Each of the variables selected had a number value associated with each song that varied between 0 and 1 so naturally for our last static graph we plotted these energy levels for each genre. Although this was similar to the middle set of graphs, it showed the data in a different way that was also color coded which made it easier to differentiate between topics. Although these were all great visualizations, our group shines through with our interactive model. It modeled our three variables in a scatter plot but the differentiating factor was that it had a drop down menu. It made it much easier to understand our surplus of data as when a genre was selected, the rest of them had a much lower opacity in their points and lines. Once selected, as you only saw the points and regression line of your selected genre, it was extremely easy to differentiate and understand the data. For the most part our feedback was extremely positive and correlated with our own thoughts on each graph. People thought that our bottom 2 sets of static graphs were straightforward and that our static scatter plots were a bit clustered but still understandable. Our saving grace was the reception of our interactive model. It cleared up any comprehension issues we received and many people actually gave us praise. I think the inclusion of this model was, although difficult in a coding aspect, the most enjoyable and fulfilling part of the project.

Conclusion:

Taking the many data points that we were using in this data set, we were able to display the shifts and trends that occurred over this time frame of 1950-2019. We found that over the seven different genres that we had analyzed and debugged when it came to our dashboard and visualizations that there were noticeable increases in loudness and energy. As more styles and variations of each genre expanded, so did the types of music in each genre. This can be seen going back to the idea of how music is produced and the consumer desire for the music that is being released. While more of an evolution that came at the end of the data set to the current time we are living in this music world, social media has made forming an audience and desire for certain genres and music sets much more obvious. Through social media like tik-tok and instagram, a pressure for artists to release unreleased music has placed this time constraint on how much a song can be used and replayed until the desired audiences become tired or sick of the songs themselves. Which all comes back full circle to these changes in music through media, that illuminates our trends that were exemplified in the dashboard as a whole, but in particular our interactive scatter plots.