Data Wrangling Report: Capstone 1

Data Cleaning

There was a fair amount of collection and organization needed for this data. I followed two pathways to process this data.
I began by downloading 9 different .txt files that I then converted to .csv format and uploaded into google drive to remove the separators, <SEP> within the files, so that these columns could be easily loaded into pandas. It was only after I worked with several files that I implemented / found the code which would automatically recognize <SEP> as the delimiter in this case (it varies away from the normal comma found in .csv files). Once I figured out how to parse out the <SEP> delimiters with the sep function, I was able to load two data files this way without having to go through google docs first, to remove the duplicate separators. After I uploaded the .csv files into my Jupyter Notebook, I was able to open a new Python file and load these files into pandas.

Pathway 1

1. Found and downloaded 9 .txt files from Columbia website.
2. Removed the <SEP> separator in the .txt files using Google Docs.
3. Converted 9 .txt files to .csv files.
4. Uploaded these to Jupyter Notebook.
5. Loaded the csv data into pandas.
6. Labelled all of the columns. (ex: EchoNestArtistID, tag, Artist.)
7. Pushed rows to begin with 1 instead of 0.
8. df.info() to understand a bit more of the data.

Pathway 2

1. Found and downloaded 3 SQL files with DB Browser for SQL Lite.
2. Exported the relevant data from DB Browser to .csv files.
3. Uploaded all .csv files to Jupyter Notebook.
4. Loaded the csv data into pandas using a sep argument that parsed out <SEP> for me, instead of using Google Docs.
5. Labelled all of the columns. (Duration, Artist Hotttnesss.)
6. Pushed rows to begin with 1 instead of 0.
7. df.info() to understand a bit more of the data.

Missing Values

The main missing values I faced with this data were the labels for the data, or rather, what the data meant. To deal with this, I cross referenced a descriptor file to differentiate between MusicBrainz data and EchoNest data.

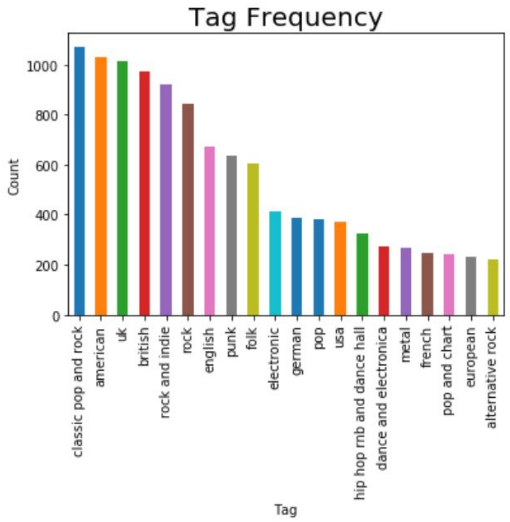
Outliers

I've gotten pretty lucky because so far to my knowledge I have yet to come across any outliers.

Data Wrangling

Bar Plot (Top 20 Tags By Frequency)

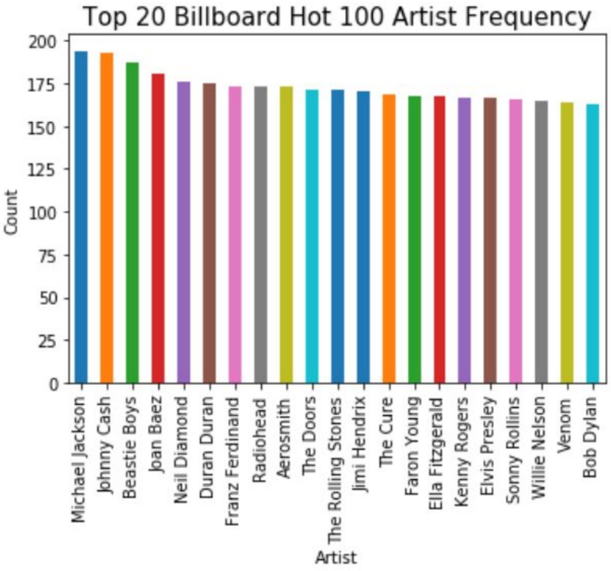
Visualizing the tags that most top songs fall under helps give some context for the kinds of songs that consistently rank in the The Billboard Hot 100 year over year. For example, we can tell from this chart that at least 8 of these tags are geographically oriented, which might possibly have implications for the language an artist should use to have reliable chances of making it into the Hot 100.



Most of the tags here fall into the classic pop and rock category, with the least amount of tags being alternative rock. This could indicate that studios already seek to publish classic pop and classic rock songs more often than alternative rock songs, or that there is simply a larger market for classic pop and classic rock songs.

Bar Plot (Top 20 Artists By Frequency)

Visualizing the frequency that the top 20 artists on the chart occur within helps give some insight into possible trends of the chart. For example, we can tell from this chart that the top 20 artists are predominantly male. This could possibly have implications for studio bias and their clients.


Top 20 Billboard Hot 100 Artist Frequency
Count
Michael Jackson
Johnny Cash
Beastie Boys
Joan Baez
Neil Diamond
Duran Duran
Franz Ferdinand
Radiohead
Aerosmith
The Doors
The Rolling Stones
Jimi Hendrix
The Cure
Faron Young
Ella Fitzgerald
Kenny Rogers
Elvis Presley
Sonny Rollins
Willie Nelson
Venom
Bob Dylan
Artist