Ethan Gutterman
Isabelle Phraner
William Smithers
Donya Behroozi
STAT 365-01

Data Proposal

Data Set

We are using the "taylor_album_songs" and "taylor_albums" datasets from TidyTuesday for the week of October 17, 2023. The dataset is accessible through TidyTuesday's GitHub page.

Data Description

The two datasets encompass Taylor Swift's entire discography to date, from her debut album release in 2006 through to her most recent work as of 2023. The dataset, "taylor_album_songs", includes all songs from Swift's albums, not including singles released separately from an album (e.g., Only the Young, Christmas Tree Farm, etc.), and non-Taylor-owned albums that have a Taylor-owned alternative. The dataset, "taylor_albums", is a smaller collection of informational data summarizing Swift's album and EP release history, including public scoring of the albums (metacritic and user scores). W. Jake Thompson provided the data on October 17, 2023, and sourced information including lyrics and audio features for each Taylor Swift song from Genius.com and the Spotify Web API. He collected the audio and lyrical data by retrieving the web API from Spotify for Developers and obtaining an access token to request artist data for musical attribute scores of the songs, and sourced informational variables like album release dates and track names via Genius.com. The "taylor_album_songs" dataset's subjects are Taylor Swift's musical outputs, more specifically, a comprehensive listing of all her songs from officially released albums. Each song is a distinct observational unit. The dataset includes all officially released tracks and albums by Taylor Swift, including songs from Swift's released albums. Spotify's API provided detailed musical attribute scores (e.g., danceability, energy, acousticness) for each song. Additional data on album releases, song credits, and commercial performance for "taylor_albums" were compiled from authoritative music industry databases and chart records (e.g. Genius.com). For this dataset, each observation is a Taylor Swift album. Spotify's API technology for analyzing musical attributes served as the primary specialized equipment used in the data collection. This tool allows for the extraction of information about song characteristics, including tempo, key, mode, and more, based on digital signal processing algorithms. Data collection was entirely digital and relied on advanced computational methods, eliminating the need for traditional physical equipment. We will actively work with the following variables: Spotify API track audio features including danceability, which describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity (value of 0.0 is least danceable and 1.0 is most danceable) and acousticness, which is a confidence measure from 0.0 to 1.0 of whether the track is acoustic (1.0 represents high confidence the track is acoustic). Additionally, we will use the following informational variables: album_release, the album release

date, metacritic_score, the metacritic score for each album, and user_score, the user score for each album.

Research Questions

1. Evolution of Musical Style Over Time:
   a. Data set: taylor_album_songs.csv
   b. Variables: Spotify attributes (e.g., danceability, valence), album release dates.
   c. Method: Time-series analysis, clustering to identify musical style shifts.
   d. Hypotheses: H_0: There is not a statistically significant relationship between danceability, acousticness, and time. H_a: There is a statistically significant relationship between danceability, acousticness, and time

2. What is the relationship between danceability and acousticness of Taylor Swift songs?
   a. Data set: taylor_album_songs.csv
   b. Variables: Acousticness (x), danceability (y)
   c. Method: (1) Create scatter plot of danceability vs. acousticness with ggplot2, (2) Fit linear regression model between the variables "danceability " and "acousticness", (3) Check assumptions of the simple linear regression, (4) Test whether there is a statistically significant relationship between danceability and acousticness scores.
   d. Hypotheses: H_0: There is not a statistically significant relationship between danceability and acousticness. H_a: There is a statistically significant relationship between danceability and acousticness.

3. Questions: As Swift's career progresses, does her music's critical reception match their popular appeal?
   a. Data set: taylor_albums.csv
   b. Variables: metacritic_score, user_score, album_release
   c. Method: (1) Create two scatter plots, mapping the critic and user scores on album release, (2) check assumptions of regression, (3) fit linear regression on each plot, (4) compare p-values.
   d. Hypothesis: H_0: There is not a statistically significant change over time between the critic and audience scores. H_a: There is a statistically significant change over time between the critic and audience scores

Group Contract

- Communication: iMessage group chat will be the primary platform for communication, along with meeting in person or via Zoom (depending on our schedules).
- Updates Check Frequency: Members are expected to check for updates and respond within 24 hours on weekdays.

- Leadership and Deadlines: Leadership roles will rotate throughout the project. Deadlines will be managed through a shared Google Calendar.
- Timeline for completion: We shared our current school/work schedules together and will rotate leadership roles/type of work done based on each of our weekly schedules and when deliverables need to be submitted. Work on assignments be equally split between all members.
- Collaboration Tools: Google Docs will be used for shared document editing and collaboration, with a GitHub repository for version control of code and data analysis.
- Submission Responsibilities: The responsibility for submitting project components will rotate among group members.
- Covering for Each Other: The group commits to supporting any member facing unexpected challenges, ensuring equitable distribution of workload.
- Conflict Resolution: Conflicts will be addressed internally through dialogue and mediation initially (if necessary, warnings will be given), with structured plans for unresolved issues, potentially involving instructor intervention as a final measure. If a member is given three warnings, we will involve the instructor for possible point reduction.