Loyal Health Data Science Coding Challenge

Instructions: The following questions are designed to assess your understanding of common data science concepts with which you should be familiar. We'll have you complete some basic analysis over text reviews and their metadata from the popular music review site Pitchfork (https://pitchfork.com/). The data can be downloaded here (https://www.kaggle.com/nolanbconaway/pitchfork-data) in the form of a SQLite database. We expect this to take around 2 hours (at most 3 hours) to complete. Although the completion of the assignment will not be strictly timed, please do not go over the allotted time. If time is an issue, focus the most on problems 2, 4, and 5.

Write all of your code in the Jupyter notebook provided. When you've completed the assessment, please create a GitHub repository, and email us a link to this repository.

1) **Cursory Data Analysis:**
    a) Compute the number of albums belonging to each genre. You should notice that some albums have multiple genres listed (e.g. Folk/Country,Pop/R&B,Rock) separated by commas. Consider albums with multiple genres as belonging to each of those genres (i.e. an album with Rap,Rock as it's genres will be counted as one Rap album and one Rock album).
    b) Compute the number of albums released each year.
    c) Compute the ten artists with the highest number of albums reviewed in the data set.
    d) Compute the mean, median, standard deviation, minimum, and maximum album scores.
    e) Compute the average score by each review author and return the result in a dataframe sort in descending order.
    f) Compute the average album score per artist and return the result in a dataframe with an additional column for the number of albums they've had reviewed.
        i) Return the artists with the top 10 highest average scores
        ii) Return the artists with the top 10 lowest average scores
2) **SQL:**
    a) Merge the database tables into a dataframe containing all of the relevant metadata.
3) **Dataframe Manipulation** (Using the Dataframe from part 2) create new DataFrames based on the stipulations below):
    a) Create a new DataFrame excluding all artists with names that start with the letter "M" (either upper or lowercase).
    b) Create a new DataFrame excluding albums with a score less than 4.0.

c) Create a new DataFrame excluding albums from the label Columbia

d) Create a new DataFrame excluding albums that belong to the metal genre.

e) Create a new DataFrame excluding albums where that artist's name contains an even number of characters (including whitespace as characters)

f) Combine these DataFrames into one where each album meets the conditions required for each.

4) **Feature Engineering:**

a) Construct a Pandas DataFrame (see problem 2) containing all album reviews and metadata. Remove any rows that have null values in any column.

b) Add a column to the dataframe for each genre. The entry in this column should be a 1 if the album/row in question belongs to that genre and 0 otherwise. Remember that albums can belong to multiple genres.

c) Add an additional two columns with categorical variables for 1) the author of the review and 2) the role of the author.

d) Create a column for the number of words in the review.

e) Create a column containing the sentiment score of the review. Treat the review as a single string and take the TextBlob polarity score (https://textblob.readthedocs.io/en/dev/quickstart.html).

5) **Logistic Regression:**

You will now use the features you constructed in the previous exercise to complete a binary logistic regression task accounting for whether an album reviews Pitchfork's designation of "Best New Music." This is represented by the binary "bnm" variable in the dataset.

a) Scale all non-categorical variables as needed.

b) Perform your logistic regression model using the statsmodel library (https://www.pythonfordatascience.org/logistic-regression-python/). Treat the best new music variable as your dependent variable and use the release year, word count, sentiment, all genre binary variables, author, and author role as your independent variables.

c) Calculate the odds ratios for your independent variables

d) What features are most/least predictive of a best new music designation and why do you think that is?

e) If you were to engineer an additional feature for the regression, what would it be? Describe how you would approach constructing that feature.

6) **Data Visualization (Optional):** Using the results from your regression and data analysis create a visualization that tells a story about the data. Feel free to take personal liberties with this and be as creative as you like.