STAT40800 Midterm Assignment

Exploratory data analysis of the Billboard Hot 100 Chart

Due Oct 29 at 6pm.

For this assignment you will perform an exploratory data analysis (EDA) for historic data from Billboard Hot 100 Chart. The Billboard Hot 100 is the music industry standard record chart in the United States for songs, published weekly by Billboard magazine. Chart rankings are based on sales, radio play, and online streaming in the United States.

Every week, Billboard releases *The Hot 100* chart of songs that were trending on sales and airplay for that week. The dataset you will analyse is a collection of all *The Hot 100* charts released since its inception in 1958. The song with rank 1 is the most popular song that week.

Marks will be awarded for complete and correct answers to all 10 questions. For full marks, you must show clearly all steps and computations in your answers. **You must submit your completed Jupyter notebook in .ipynb and PDF format**. To create a PDF of your notebook select *File* → *Download as* → *PDF via LaTeX (.pdf)*.

[ ]:
```
# Load in necessary packages
```

1. Load the `charts.csv` dataset into Python as a pandas DataFrame. Include the arguments `parse_dates=True` in the read_csv function to convert the date strings into datetime objects. Describe the data. Comment on the size of the dataset and the type of data in each column.

[ ]:
```

```

2. Write code to answer the following questions:
    1. How many different dates are included in the data set?
    - How many different artists featured in the charts since its inception in 1958?
    - Determine the top 10 artists that feature in the charts the most amount of times?
    - What is the maximum length of time a song featured in the charts? What was this song and who was the artist?

[ ]:
```

```

3. Determing how many missing values there are in each column of the dataset. Can you think of a reason why why these values are missing? Replace any missing values in the `last-week` column with the integer 101. Is this a reasonable value to replace the missing values with?

[ ]:
```

```

4. For each entry compute the difference between its rank last week and its current rank. Add the change in ranking as an additional column to the DataFrame. Create a numerical summary of change in ranking (mean, standard deviation, minimum, maximum, etc.) and interpret your results.

Note: If the song is higher up the charts this week, the change in ranking should be positive. While if the song falls down the charts, the change in ranking should be negative.

[ ]:

5. Create a scatter plot of the current rank versus the change in ranking from the previous week. Describe and interpret the plot. Compute the correlation coefficient between the current rank and the change in ranking. Does is agree with your what you found in your plot?

[ ]:

6. Songs will typically appear in the charts for a number of weeks, as such, there are a number of duplicate entries. Delete all duplicate entries, keeping only the most recent entries of a particular song by a particular artist in the charts. How many unique songs (with unique artists) have appeared in the charts since its inception in 1958? Repeat Q2C for the new DataFrame, with unique entries only. Are the top 10 entries the same? If not, discuss why they may have changed.

   You may find the pandas method `drop_duplicates` useful. See the pandas documentation for details on how to apply this method to a DataFrame.

[ ]:

7. Recompute the summary statistics for the change in ranking from the previous week and recreate the scatter plot of the current ratings versus the change in ranking for the new DataFrame. Discuss and interpret any changes in these statistics and the plot.

[ ]:

8. Using the new DataFrame, with unique entries only, create a histogram for the length of time each entry spent in the charts. Describe and interpret your plot. Compute the summary statistics for the length of time a song spends in the charts. Discuss these statistics in relation to your histogram.

[ ]:

9. Using the new DataFrame, with unique entries only, create a histogram for the peak rank. Describe and interpret your plot. Compute the summary statistics for peak rank. Discuss these statistics in relation to your histogram.

[ ]:

10. Write a short summary on what you have learnt about the Billboard Hot 100 Chart from this analysis. (5-10 sentences)

[ ]: