Initial analysis of “Hot\_100BBCharts\_time.csv” data pulled from BillBoard Charts Hot-100 list. This is a weekly published list since, August 4, 1958. I pulled data by weeks. I chose to pull every Wednesday from August 6th, 1958 until October 31st, 2018. Pulled data for 3144 Wednesdays to cover the 60-year period… the end result should have 314,400 rows of data plus header.

Here is a sneak peek at the DataFrame from the CSV:



The final CSV file had 311,256 rows of data plus header. Here is some information about the CSV file:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 311256 entries, 0 to 311255

Data columns (total 8 columns):

Date 311256 non-null object

Title 311256 non-null object

Artist 311256 non-null object

Peak\_Position 311256 non-null int64

Last\_Position 311256 non-null int64

Weeks\_on\_Chart 311256 non-null int64

Rank\_in\_Chart 311256 non-null int64

New\_on\_Chart 311256 non-null bool

dtypes: bool(1), int64(4), object(3)

memory usage: 16.9+ MB

Here is the report for NaN values:

Date 0

Title 0

Artist 0

Peak\_Position 0

Last\_Position 0

Weeks\_on\_Chart 0

Rank\_in\_Chart 0

New\_on\_Chart 0

dtype: int64

Conclusions:

There are 3,144 less rows than expected. Not sure if my range was incorrect; the data didn’t exist or was not entered correctly from the early days; perhaps it is really the top 99 songs and not a true 100. The missing rows account for 1% of the data and no NaN values, I feel this is a good data set to continue with.

The code is in this jupyter notebook is:

BillboardCharts\_Hot100\_Collecter.ipynb

The CSV file in in the Resources directory:

Hot\_100BBCharts\_time.csv

Feel free to run the code and play with it. FYI, it took two hours to pull the 60+ year’s worth of data in the full CSV file. I have commented out somethings, so I did not do over write the full CSV file and takes a smaller Wednesday List… but please comment and ask questions.

WooHoo, we have lots of Data now with all of Austen’s lists and this one too.