Capstone Project 1 - Data Wrangling

Spotify Audio Features

The data for this project was originally obtained from Spotify's Web API and made available for use through on kaggle as a CSV file (link: https://www.kaggle.com/tomigelo/spotify-audio-features). The data, with 116,372 rows (one song per row) and 17 columns, was read into Python from a CSV file. As expected from a kaggle dataset, this data was relatively clean. The following steps were taken to ensure creation of an analyses ready dataset:

1. The data was then tested for tidiness. In this dataset, each attribute is measured in a separate variable, each observation (i.e. a song) is in a separate row and there are no row duplicates in the data (tested using drop_duplicates() method). So the dataset is considered tidy.

2. The data was then examined for missingness in any column using the .info() attribute (as well as .describe()). There were no missing values in any column. If there was any missing values in the 'variable' variable, we could fill it using the syntax:

data['variable'] = data['variable'].fillna(some_value)

The some_value value could be any value such as 0, a summary statistic (e.g. mean, median) or a string

3. I then tried to gain an understanding of each variable, how it is created and the possible range of values from this document: https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/. This would help identify outliers, if any.

Histograms of all numeric variables were examined (the non-numeric ones are explained below). These variables were the following:

Acousticness, danceability, duration_ms, energy, instrumentalnes, liveness, loudness, speechiness, tempo, valence, popularity

Out of the above list, only duration_ms seemed to have extremely large values. These were tracks with 20,30 or even 90 minutes length. A sample of these observations were printed out and the track was looked at on Spotify's application. The duration values in all the examined tracks were correct. Some examples are:

- Artist name: John, Tracks: Whatever, 5,040,048 ms (~84 mins) and Ever, 5,610,020 ms (~93.5 mins)

- Artist name: Gentle Whispering, Track: Fluffy Sleepy Whispers, 2,097,245 (~35 mins)

4. The variables Mode, Key and time_signature (although originally numeric) need to be used in the analyses as categorical. This is because these 3 variables do not have an ordinal nature. (e.g. for variable, the difference between C major, D minor and A minor keys should all be one unit). So these variables were converted to ‘category’.