DASC 1104 Project Proposal

Hayden McDonald

12/3/2020

My Blog Link

My blog is available at https://hpmcdona.netlify.app/

```r
library(here)
library(ggplot2)
library(tidyverse)
library(readxl)
knitr::opts_chunk$set(echo = FALSE, tidy = TRUE)
spotify <- read_csv(here::here("data", "tidytuesday",
                               "data","2020","2020-01-21","spotify_songs.csv"))
```

Exploring the Spotify data.

```
## Observations: 32,833
## Variables: 23
## $ track_id                 <chr> "6f807x0ima9a1j3VPbc7VN", "0r7CVbZTWZgbTCYdfa2P31", "1z1Hg7Vb0A...
## $ track_name               <chr> "I Don't Care (with Justin Bieber) - Loud Luxury Remix", "Memor...
## $ track_artist             <chr> "Ed Sheeran", "Maroon 5", "Zara Larsson", "The Chainsmokers", "...
## $ track_popularity         <dbl> 66, 67, 70, 60, 69, 67, 62, 69, 68, 67, 58, 67, 67, 68, 63, 66,...
## $ track_album_id           <chr> "2oCs0DGTsRO98Gh5ZSl2Cx", "63rPSO264uRjW1X5E6cWv6", "1HoSmj2eLc...
## $ track_album_name         <chr> "I Don't Care (with Justin Bieber) [Loud Luxury Remix]", "Memor...
## $ track_album_release_date <chr> "2019-06-14", "2019-12-13", "2019-07-05", "2019-07-19", "2019-0...
## $ playlist_name            <chr> "Pop Remix", "Pop Remix", "Pop Remix", "Pop Remix", "Pop Remix"...
## $ playlist_id              <chr> "37i9dQZF1DXcZDD7cfEKhW", "37i9dQZF1DXcZDD7cfEKhW", "37i9dQZF1D...
## $ playlist_genre           <chr> "pop", "pop", "pop", "pop", "pop", "pop", "pop", "pop", "pop", ...
## $ playlist_subgenre        <chr> "dance pop", "dance pop", "dance pop", "dance pop", "dance pop"...
## $ danceability             <dbl> 0.748, 0.726, 0.675, 0.718, 0.650, 0.675, 0.449, 0.542, 0.594, ...
## $ energy                   <dbl> 0.916, 0.815, 0.931, 0.930, 0.833, 0.919, 0.856, 0.903, 0.935, ...
## $ key                      <dbl> 6, 11, 1, 7, 1, 8, 5, 4, 8, 2, 6, 8, 1, 5, 5, 0, 1, 10, 1, 7, 6...
## $ loudness                 <dbl> -2.634, -4.969, -3.432, -3.778, -4.672, -5.385, -4.788, -2.419,...
## $ mode                     <dbl> 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, ...
## $ speechiness              <dbl> 0.0583, 0.0373, 0.0742, 0.1020, 0.0359, 0.1270, 0.0623, 0.0434,...
## $ acousticness             <dbl> 0.10200, 0.07240, 0.07940, 0.02870, 0.08030, 0.07990, 0.18700, ...
## $ instrumentalness         <dbl> 0.00e+00, 4.21e-03, 2.33e-05, 9.43e-06, 0.00e+00, 0.00e+00, 0.0...
## $ liveness                 <dbl> 0.0653, 0.3570, 0.1100, 0.2040, 0.0833, 0.1430, 0.1760, 0.1110,...
## $ valence                  <dbl> 0.518, 0.693, 0.613, 0.277, 0.725, 0.585, 0.152, 0.367, 0.366, ...
## $ tempo                    <dbl> 122.036, 99.972, 124.008, 121.956, 123.976, 124.982, 112.648, 1...
## $ duration_ms              <dbl> 194754, 162600, 176616, 169093, 189052, 163049, 187675, 207619,...
```

Spotify

For this project I am examining the Spotify Songs dataset contained in the spotify_songs.csv file on the Tidy Tuesday website. The data consists of 23 variables with 32,833 observations, the definitions for each variable were taken directly from the Tidy Tuesday Spotify website. The variables track_id(song unique ID), track_name(song name), track_artist(song artist), track_album_id(album unique ID), track_album_name(song album name), track_album_release_date(date when album released), playlist_name(name of playlist), playlist_id(unique playlist ID for each playlist), playlist_genre(playlist genre), and playlist_subgenre(playlist subgenre) are all characters. The variable track_popularity(song popularity, 0-100) where higher is better is a double. The variable danceability is a double that describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. The variable energy is a double that measures from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. The variable key is a double where Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = Csharp/Dflat, 2 = D, and so on. If no key was detected, the value is -1. The variable loudness is a double that measures the overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. The variable mode is a double that indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. The variable speechiness is a double that detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. The variable acousticness is a double 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. The variable instrumentalness is a double that predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. The variable liveness is a double that detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. The variable valence is a double that measures from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). The variable tempo is a double that represents the overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. The variable duration_ms is a double that shows the duration of the song in milliseconds

- Question 1: How does energy compare to the duration of a song? To test this I will create a chart, possibly a point or violin, to show dispersion and look for clustering at certain energy levels and times to see if there is a trend for how long high energy songs typically are and how long low energy songs are.

- Question 2: What are the 10 most popular songs? What genres show up more than once? To test this I will need to filter for genre, song name, and popularity and then arrange them in descending order. From there I will get a list of the top 10 popular songs and can see what genres show up more than once.

- Question 3: How does speechiness compare to popularity? Is there a sweetspot? To test this I will need to graph speechiness against popularity and look for a trend, I will likely start with a point graph and change from there if there is a better solution that presents itself.

Global Crop Yields

Exploring the Global Crop Yields data

```
## Observations: 13,075
## Variables: 14
## $ Entity                             <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghani...
## $ Code                               <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG...
## $ Year                               <dbl> 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969,...
## $ ‘Wheat (tonnes per hectare)‘       <dbl> 1.0220, 0.9735, 0.8317, 0.9510, 0.9723, 0.8666, 1.123...
## $ ‘Rice (tonnes per hectare)‘        <dbl> 1.5190, 1.5190, 1.5190, 1.7273, 1.7273, 1.5180, 1.922...
## $ ‘Maize (tonnes per hectare)‘       <dbl> 1.4000, 1.4000, 1.4260, 1.4257, 1.4400, 1.4400, 1.414...
## $ ‘Soybeans (tonnes per hectare)‘    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ‘Potatoes (tonnes per hectare)‘    <dbl> 8.6667, 7.6667, 8.1333, 8.6000, 8.8000, 9.0667, 9.800...
## $ ‘Beans (tonnes per hectare)‘       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ‘Peas (tonnes per hectare)‘        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ‘Cassava (tonnes per hectare)‘     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ‘Barley (tonnes per hectare)‘      <dbl> 1.0800, 1.0800, 1.0800, 1.0857, 1.0857, 1.0714, 1.129...
## $ ‘Cocoa beans (tonnes per hectare)‘ <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ‘Bananas (tonnes per hectare)‘     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
```

```
## Observations: 11,965
## Variables: 5
## $ Entity                                            <chr> "Afghanistan", "Afghanistan", "Afghani...
## $ Code                                              <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AF...
## $ Year                                              <dbl> 1961, 1962, 1963, 1964, 1965, 1966, 19...
## $ ‘Cereal yield (tonnes per hectare)‘               <dbl> 1.1151, 1.0790, 0.9858, 1.0828, 1.0989...
## $ ‘Nitrogen fertilizer use (kilograms per hectare)‘ <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
```

```
## Observations: 49,403
## Variables: 6
## $ Entity                                                  <chr> "Afghanistan", "Afghanistan", "A...
## $ Code                                                    <chr> "AFG", "AFG", "AFG", "AFG", "AFG...
## $ Year                                                    <chr> "1961", "1962", "1963", "1964", ...
## $ ‘Tractors per 100 sq km arable land‘                    <dbl> 0.1568627, 0.1948052, 0.2580645,...
## $ ‘Cereal yield (kilograms per hectare) (kg per hectare)‘ <dbl> 1115.1, 1079.0, 985.8, 1082.8, 1...
## $ ‘Total population (Gapminder)‘                          <dbl> 9169000, 9351000, 9543000, 97450...
```

```
## Observations: 49,259
## Variables: 6
## $ Entity                                                      <chr> "Afghanistan", "Afghanistan"...
## $ Code                                                        <chr> "AFG", "AFG", "AFG", "AFG", ...
## $ Year                                                        <chr> "1961", "1962", "1963", "196...
## $ ‘Cereal yield index‘                                        <dbl> 100, 97, 88, 97, 99, 91, 110...
## $ ‘Change to land area used for cereal production since 1961‘ <dbl> 100, 103, 103, 104, 104, 104...
## $ ‘Total population (Gapminder)‘                              <dbl> 9169000, 9351000, 9543000, 9...
```

```
## Observations: 11,280
## Variables: 4
## $ Entity                                                                   <chr> "Afghanistan", ...
## $ Code                                                                     <chr> "AFG", "AFG", "...
## $ Year                                                                     <dbl> 1961, 1962, 196...
## $ ‘Arable land needed to produce a fixed quantity of crops ((1.0 = 1961))‘ <dbl> 1.0000000, 0.98...
```

For this project I am examining the Global Crop Yields dataset contained in 5 different csv files on the Tidy Tuesday website. The definitions for each variable were taken from the Tidy Tuesday Global Crop Yields website. All 5 datasets have the variables Entity(character, Country or Region Name), Code (character, Country Code (note is NA for regions/continents)), and Year(double, Year). key_crop_yields(13,075 observation with 14 variables) holds the variables Wheat (tonnes per hectare) (double, Wheat yield), Rice (tonnes per hectare) (double, Rice Yield), Maize (tonnes per hectare) (double, Maize yield), Soybeans (tonnes per hectare) (double, Soybeans yield), Potatoes (tonnes per hectare) (double, Potato yield), Beans (tonnes per hectare) (double, Beans yield), Peas (tonnes per hectare) (double, Peas yield), Cassava (tonnes per hectare) (double, Cassava (yuca) yield), Barley (tonnes per hectare) (double, Barley yield), Cocoa beans (tonnes per hectare) (double, Cocoa beans yield), and Bananas (tonnes per hectare) (double, Bananas) all in tonnes per hectare. arable_land(11,280 observations with 4 variables) holds the variable Arable land needed to produce a fixed quantity of crops ((1.0 = 1961)) which is a double that shows the arable land needed to produce a fixed quantity of crops with arable land normalized to 1961. fertilizer(11,965 observations with 5 variables) contains the variables Cereal yield (tonnes per hectare) which is a double that shows the cereal yield in tonnes per hectare and Nitrogen fertilizer use (kilograms per hectare) which is a double that shows the Nitrogen fertilizer use kg per hectare. land_use(49,259 observations with 6 variables) contains the variables cereal yield index(double, cereal yield index), change to land area used for cereal production since 1961(double, change to land area use for cereal production relative since 1961), and total population (gapminder)(double, total population from gapminder data). tractors(49,403 observations with 6 variables) contains the variables Tractors per 100 sq km arable land(double, number of tractors per 100 sq km of arable land), Cereal yield (kilograms per hectare) (kg per hectare) which is a double showing the cereal yield in kg per hectare, and Total population (Gapminder)(double, total population from gapminder).

- Question 1: How does Tractors per 100 sq km arable land affect Total population over time in Ireland? Is there a clear trend that the number tractors affect population growth? To test this I will need to use the tractors dataset, filter for the years 1961-2005 since those are the years with available data, and I can go about answering the question in two ways. I can either make two line graphs with time as a variable or just one comparing the two variables against eachother. The two variables are Tractors per 100 sq km arable land and Total population (Gapminder). I will need to look at when the Tractors ratio is increasing if the population increases at a faster rate.
- Question 2: How has the output of potatoes been affected in Ireland from 1961 to 2018? What years had the highest and lowest potato production? To test this I need to use the key_crop_yields dataset and select Ireland as the country and the years past 1961. Then I need to graph the output of potatoes vs time in years as a line graph. From there I can either see what years had the highest or lowest from the graph or use max/min functions in the piping or arrange in ascending and descending order.
- Question 3: How efficient has Ireland become with producing crops on arable land since 1961? To test this I will need to use the arable_land dataset and use the Arable land needed to produce a fixed quantity of crops ((1.0 = 1961)). I will have to filter for the years between 1961 and 2005 since those are the years where data is recorded for this variable. Then I will need to make a line graph of the variable previously noted over those years and look for any trends.