Topic ideas

The Kable Guys: Jerry Hou, Arjun Prabhakar, Nathan Huang, Ryan Mitchell

October 7, 2021

Data Set 1

Introduction and Data

The `movies.csv` dataset is a part of the Tidy Tuesday repository. The dataset comes from a FiveThirtyEight article, which relied on data from BechdelTest.com and The-Numbers.com. FiveThirtyEight used the former to determine if films passed the Bechdel Test and the latter to obtain financial information on the films. The `movies.csv` dataset has 1794 observations (each representing a film) and 34 variables. Key variables in the dataset are genre, IMDB rating, domestic gross, international gross, budget, awards, and whether or not the film passes the Bechdel Test.

Research questions

Are films that pass the Bechdel Test better received by audiences and critics? Does the IMDB rating of a movie have a relationship with domestic gross or the amount of awards the movie receives?

Data Set 2

Introduction and Data

The `audio_features.csv` dataset was found on the Tidy Tuesday repository and comes from Data.world, courtesy of Sean Miller, Billboard.com and Spotify. The data was collected by Spotify's analytics of tracks uploaded and user interaction with the tracks. The csv file has 29,503 observations and 22 columns. Some notable variables are song, genre, duration in ms, and track popularity on Spotify.

Research questions

Some research questions we had were: Are songs that use major or minor melodies more popular? Does the musical genre of a song predict its popularity?

Data Set 3

Introduction and Data

The `pollution.csv` dataset was found on the Tidy Tuesday repository and comes from Break Free from Plastic courtesy of Sarah Sauve. The csv file has 13,380 observations and 14 columns. Some notable variables are country of pollution incident, parent company causing pollution incident, grand pollutant total, and pvc amount.

Research questions

How does parent company affect the amount of pollutant and does this differ for different kinds of pollutants?

Glimpse of data sets

Data set 1

```
## Rows: 1,794
## Columns: 34
## $ year          <dbl> 2013, 2012, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 20~
## $ imdb          <chr> "tt1711425", "tt1343727", "tt2024544", "tt1272878", "tt0~
## $ title         <chr> "21 &amp; Over", "Dredd 3D", "12 Years a Slave", "2 Guns~
## $ test          <chr> "notalk", "ok-disagree", "notalk-disagree", "notalk", "m~
## $ clean_test    <chr> "notalk", "ok", "notalk", "notalk", "men", "men", "notal~
## $ binary        <chr> "FAIL", "PASS", "FAIL", "FAIL", "FAIL", "FAIL", "FAIL", ~
## $ budget        <dbl> 1.30e+07, 4.50e+07, 2.00e+07, 6.10e+07, 4.00e+07, 2.25e+~
## $ domgross      <chr> "25682380", "13414714", "53107035", "75612460", "9502021~
## $ intgross      <chr> "42195766", "40868994", "158607035", "132493015", "95020~
## $ code          <chr> "2013FAIL", "2012PASS", "2013FAIL", "2013FAIL", "2013FAI~
## $ budget_2013   <dbl> 13000000, 45658735, 20000000, 61000000, 40000000, 225000~
## $ domgross_2013 <chr> "25682380", "13611086", "53107035", "75612460", "9502021~
## $ intgross_2013 <chr> "42195766", "41467257", "158607035", "132493015", "95020~
## $ period_code   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ decade_code   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ imdb_id       <chr> "1711425", "1343727", "2024544", "1272878", "0453562", "~
## $ plot          <chr> NA, NA, "In the antebellum United States, Solomon Northu~
## $ rated         <chr> NA, NA, "R", "R", "PG-13", "PG-13", "R", "R", "PG-13", "~
## $ response      <lgl> NA, NA, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ~
## $ language      <chr> NA, NA, "English", "English, Spanish", "English", "Engli~
## $ country       <chr> NA, NA, "USA, UK", "USA", "USA", "USA", "USA", "UK", "US~
## $ writer        <chr> NA, NA, "John Ridley (screenplay), Solomon Northup (base~
## $ metascore     <dbl> NA, NA, 97, 55, 62, 29, 28, 55, 48, 33, 90, 58, 52, 78, ~
## $ imdb_rating   <dbl> NA, NA, 8.3, 6.8, 7.6, 6.6, 5.4, 7.8, 5.7, 5.0, 7.5, 7.4~
## $ director      <chr> NA, NA, "Steve McQueen", "Baltasar Kormákur", "Brian Hel~
## $ released      <chr> NA, NA, "08 Nov 2013", "02 Aug 2013", "12 Apr 2013", "25~
## $ actors        <chr> NA, NA, "Chiwetel Ejiofor, Dwight Henry, Dickie Gravois,~
## $ genre         <chr> NA, NA, "Biography, Drama, History", "Action, Comedy, Cr~
## $ awards        <chr> NA, NA, "Won 3 Oscars. Another 131 wins & 137 nomination~
## $ runtime       <chr> NA, NA, "134 min", "109 min", "128 min", "118 min", "98 ~
## $ type          <chr> NA, NA, "movie", "movie", "movie", "movie", "movie", "mo~
## $ poster        <chr> NA, NA, "http://ia.media-imdb.com/images/M/MV5BMjExMTEzO~
## $ imdb_votes    <dbl> NA, NA, 143446, 87301, 43608, 25735, 123837, 85871, 1897~
## $ error         <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

Data set 2

```
## Rows: 29,503
## Columns: 22
## $ song_id                   <chr> "-twistin'-White Silver SandsBill Black's Co~
## $ performer                 <chr> "Bill Black's Combo", "Augie Rios", "Andy Wi~
## $ song                      <chr> "-twistin'-White Silver Sands", "¿Dònde Està~
## $ spotify_genre             <chr> "[]", "['novelty']", "['adult standards', 'b~
## $ spotify_track_id          <chr> NA, NA, "3tvqPPpXyIgKrm4PR9HCf0", "1fHHq3qHU~
## $ spotify_track_preview_url <chr> NA, NA, "https://p.scdn.co/mp3-preview/cef48~
## $ spotify_track_duration_ms <dbl> NA, NA, 166106, 172066, 211066, 208186, 2055~
## $ spotify_track_explicit    <lgl> NA, NA, FALSE, FALSE, FALSE, FALSE, TRUE, FA~
## $ spotify_track_album       <chr> NA, NA, "The Essential Andy Williams", "Comp~
## $ danceability              <dbl> NA, NA, 0.154, 0.588, 0.759, 0.613, NA, 0.64~
## $ energy                    <dbl> NA, NA, 0.185, 0.672, 0.699, 0.764, NA, 0.68~
## $ key                       <dbl> NA, NA, 5, 11, 0, 2, NA, 2, NA, NA, 7, NA, 1~
## $ loudness                  <dbl> NA, NA, -14.063, -17.278, -5.745, -6.509, NA~
## $ mode                      <dbl> NA, NA, 1, 0, 0, 1, NA, 0, NA, NA, 1, NA, 0,~
## $ speechiness               <dbl> NA, NA, 0.0315, 0.0361, 0.0307, 0.1360, NA, ~
## $ acousticness              <dbl> NA, NA, 0.91100, 0.00256, 0.20200, 0.05270, ~
## $ instrumentalness          <dbl> NA, NA, 2.67e-04, 7.45e-01, 1.31e-04, 0.00e+~
## $ liveness                  <dbl> NA, NA, 0.1120, 0.1450, 0.4430, 0.1970, NA, ~
## $ valence                   <dbl> NA, NA, 0.150, 0.801, 0.907, 0.417, NA, 0.95~
## $ tempo                     <dbl> NA, NA, 83.969, 121.962, 92.960, 160.015, NA~
## $ time_signature            <dbl> NA, NA, 4, 4, 4, 4, NA, 4, NA, NA, 4, NA, 4,~
## $ spotify_track_popularity  <dbl> NA, NA, 38, 11, 77, 73, 61, 40, NA, NA, 31, ~
```

Data set 3

```
## Rows: 13,380
## Columns: 14
## $ country        <chr> "Argentina", "Argentina", "Argentina", "Argentina", "Ar~
## $ year           <dbl> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2~
## $ parent_company <chr> "Grand Total", "Unbranded", "The Coca-Cola Company", "S~
## $ empty          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ hdpe           <dbl> 215, 155, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ ldpe           <dbl> 55, 50, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ o              <dbl> 607, 532, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 13, 0, 0, 0,~
## $ pet            <dbl> 1376, 848, 222, 39, 38, 22, 21, 26, 19, 14, 14, 14, 14,~
## $ pp             <dbl> 281, 122, 35, 4, 0, 7, 6, 0, 1, 4, 3, 1, 0, 0, 3, 0, 4,~
## $ ps             <dbl> 116, 114, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ pvc            <dbl> 18, 17, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ grand_total    <dbl> 2668, 1838, 257, 43, 38, 29, 27, 26, 20, 18, 17, 15, 14~
## $ num_events     <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ volunteers     <dbl> 243, 243, 243, 243, 243, 243, 243, 243, 243, 243, 243, ~
```