STA_Wrangling

2024-08-18

```
## Wrangling the Billboard Top 100
# Part A

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
# read in file
billboard = read.csv('/Users/teamccormack/Downloads/billboard.csv')

# number of weeks a song and performer are in the top 10
# group by performer and song to get unique combinations
# summarise gets number of unique weeks
# I asked Chat GPT "how to get the top 10 values from a data frame when grouped by 2 variables"
count <- billboard %>%
  group_by(performer, song) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
```

```
## `summarise()` has grouped output by 'performer'. You can override using the
## `.groups` argument.
```

```
# Print the number of times a song is in the billboard top 100
# uses whole data frame
#print(count)

# gets only the top 10
top_10 <- head(count, 10)
top_10 # displays top 10
```

```
## # A tibble: 10 x 3
## # Groups:   performer [10]
##    performer                                song                        count
##    <chr>                                    <chr>                       <int>
##  1 Imagine Dragons                          Radioactive                    87
##  2 AWOLNATION                               Sail                           79
##  3 Jason Mraz                               I'm Yours                      76
##  4 The Weeknd                               Blinding Lights                76
```

```
##  5 LeAnn Rimes                                How Do I Live                  69
##  6 LMFAO Featuring Lauren Bennett & GoonRock Party Rock Anthem              68
##  7 OneRepublic                                Counting Stars                 68
##  8 Adele                                      Rolling In The Deep            65
##  9 Jewel                                      Foolish Games/You Were Meant~  65
## 10 Carrie Underwood                           Before He Cheats               64
```

```
print('The table shows that the top 10 song and performer combinations spent between 64 and 87 weeks on
```

```
## [1] "The table shows that the top 10 song and performer combinations spent between 64 and 87 weeks o
```

The table shows the top 10 most popular songs from 1958 to 2021. The data includes perfomer, song, and count, and differentiates different popular song names based on the artist that performed them. The top performer and song combination is Radioactive by Imagine Dragons.

Part A creates a table of the 10 most popular songs since 1958, and includes performer, song, and count. Performer and song were grouped together in order to get the count that the song and performer appeared on the top 100. The top 10 are then sorted, starting with which song was the most popular.

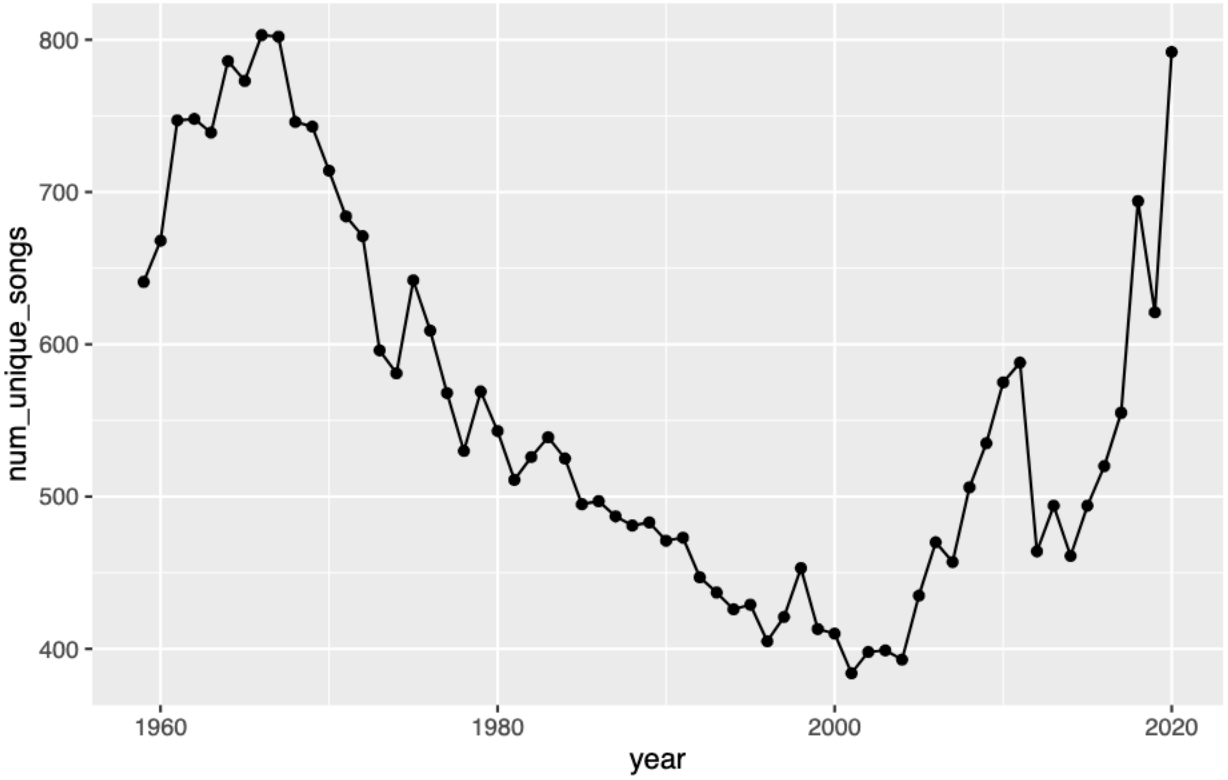
```
# Part B

# exclude year 1958 and 2021
remove_years <- billboard %>%
  filter(year != 1958 & year != 2021)

# groups by year and gets number of unique songs
# uses remove years data frame so 1958 and 2021 are not included
musical_diversity <- remove_years %>%
  group_by(year) %>%
  summarise(num_unique_songs = n_distinct(song)) %>%
  arrange(year) # sort by year
musical_diversity
```

```
## # A tibble: 62 x 2
##     year num_unique_songs
##    <int>            <int>
##  1  1959              641
##  2  1960              668
##  3  1961              747
##  4  1962              748
##  5  1963              739
##  6  1964              786
##  7  1965              773
##  8  1966              803
##  9  1967              802
## 10  1968              746
## # i 52 more rows
```

```
# I asked Chat GPT how to add a caption to a line plot and was told to use labs(caption = )
# plot the musical diversity results
ggplot(musical_diversity, aes(x = year, y = num_unique_songs)) +
  geom_line() +
  geom_point() +
  labs(caption = 'The plot shows the number of unique songs for each year from 1958 to 2020.
       1958 and 2021 were excluded becuase all data for both of those years had not been collected.')
```



The plot shows the number of unique songs for each year from 1958 to 2020.
1958 and 2021 were excluded becuase all data for both of those years had not been collected.

For years with complete data (not 1958 and 2021), the number of unique songs that appeared on the top 100 was counted. When looking at the graph, you can see that there is a distinct drop in the number of unique songs between 1980 and 2010.

```
# Part C

# get artists that have songs for at least 10 weeks
min_10_weeks <- billboard %>%
  group_by(performer, song) %>%
  summarise(weeks_on_chart = n()) %>%
  filter(weeks_on_chart >= 10)
```

```
## `summarise()` has grouped output by 'performer'. You can override using the
## `.groups` argument.
```

```
min_10_weeks
```

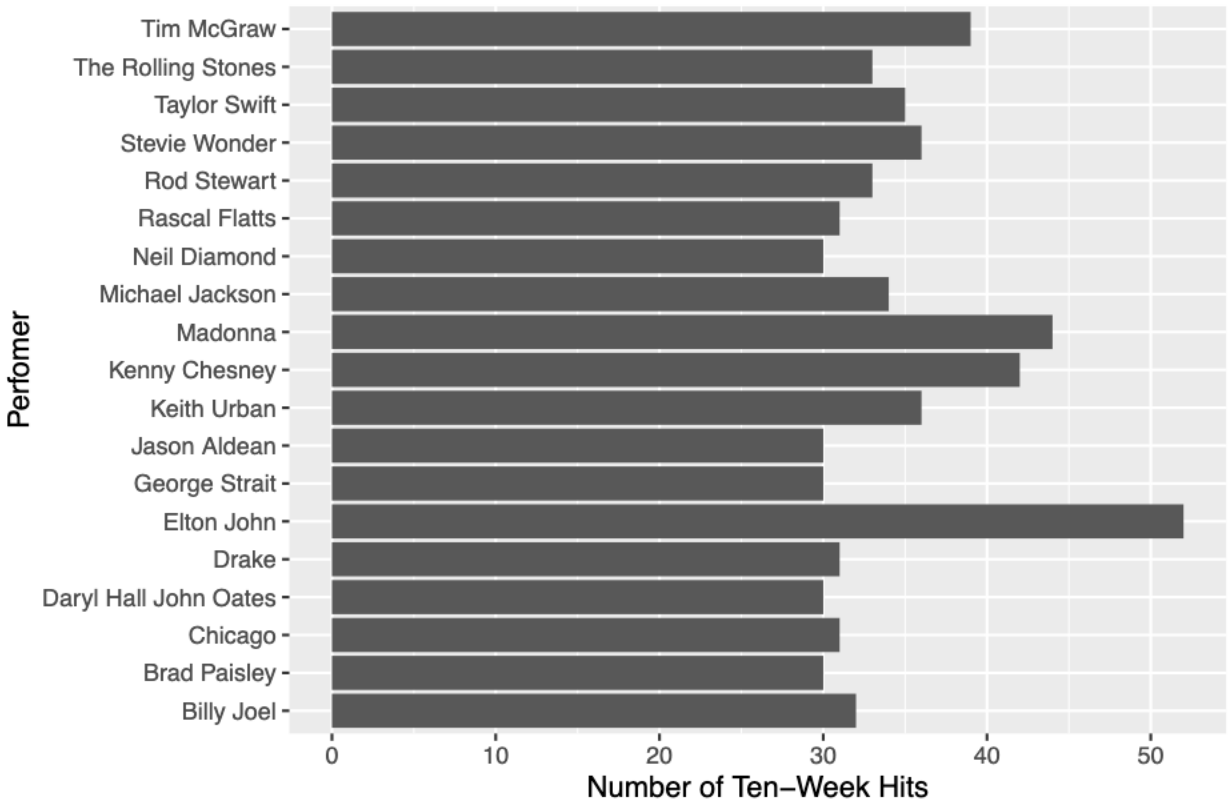
```
## # A tibble: 14,807 x 3
## # Groups:   performer [6,126]
##    performer                  song                                  weeks_on_chart
##    <chr>                      <chr>                                          <int>
##  1 "\"Groove\" Holmes"        Misty                                             11
##  2 "\"Little\" Jimmy Dickens" May The Bird Of Paradise Fly Up Yo~               10
##  3 "\"Weird Al\" Yankovic"    Amish Paradise                                    16
##  4 "\"Weird Al\" Yankovic"    Eat It                                            12
##  5 "\"Weird Al\" Yankovic"    Smells Like Nirvana                               11
##  6 "\"Weird Al\" Yankovic"    White & Nerdy                                     20
##  7 "'N Sync"                  (God Must Have Spent) A Little Mor~               22
##  8 "'N Sync"                  Bye Bye Bye                                       23
```

```
##  9 "'N Sync"                  Gone                                          24
## 10 "'N Sync"                  I Drive Myself Crazy                          12
## # i 14,797 more rows
```

```
# number of 10 week hits for each artist
# 19 artist have at least 30 songs
artist_10_weeks <- min_10_weeks %>%
  group_by(performer) %>%
  summarise(num_hit_greater10 = n()) %>%
  filter(num_hit_greater10 >= 30)
artist_10_weeks
```

```
## # A tibble: 19 x 2
##    performer             num_hit_greater10
##    <chr>                             <int>
##  1 Billy Joel                           32
##  2 Brad Paisley                         30
##  3 Chicago                              31
##  4 Daryl Hall John Oates                30
##  5 Drake                                31
##  6 Elton John                           52
##  7 George Strait                        30
##  8 Jason Aldean                         30
##  9 Keith Urban                          36
## 10 Kenny Chesney                        42
## 11 Madonna                              44
## 12 Michael Jackson                      34
## 13 Neil Diamond                         30
## 14 Rascal Flatts                        31
## 15 Rod Stewart                          33
## 16 Stevie Wonder                        36
## 17 Taylor Swift                         35
## 18 The Rolling Stones                   33
## 19 Tim McGraw                           39
```

```
# bar plot showing how many 10 week hits per artist
ggplot(artist_10_weeks, aes(x = performer, y = num_hit_greater10)) +
  geom_bar(stat = "identity") +
  coord_flip() +  # flip coordinates
  labs(
    x = "Perfomer",
    y = "Number of Ten-Week Hits",
    caption = "The bar graph shows 19 artists that have had at least 30 songs that were 10 week hits."
  )
```



The bar graph shows 19 artists that have had at least 30 songs that were 10 week hits. I began by getting all of the artists that had a song on the billboard chart for at least 10 weeks. Next, I took that data and grouped by performer, and made sure that a performer's count was greater than 30. That resulted in 19 artists having at least 30 hits. The bar plot shows the amount of ten week hits that an artist has vs. the performer.