Lecture 10: Tidy Data Principles

Problem Set for Lecture 10

These questions will assist in bolstering your understanding of the material in Lecture 10. Emphasis should be placed on having the correct code/output as well as communication. The deliverable should be a knitted RMarkdown document to pdf without any code running off the page. You will be able to present these in office hours as an oral assessment interview along with the problem set for Lecture 8 and Lecture 9.

***** First, make sure you have read the Lecture Material's writeup *****

1. Using the `course_scores` dataset in the `MSMU` library, tidy the data into a form that allows you to analyze student performance across subject and grade level. Then answer the following questions.

   (a) Compute the average numeric score for each subject across all grade levels.

   (b) Compute the average numeric score for each grade level across all subjects.

   (c) Compute the average numeric score for each subject-grade combination.

   (d) Determine the distribution of letter grades for each subject.

   (e) Compute a GPA for each student using the scale F=0, D=1, C=2, B=3, A=4 across all of their letter-grade observations.

   (f) If a student needs at least a 2.0 GPA to graduate, how many students would graduate? (51)

   (g) Identify the top 5 students with the largest absolute difference between their average English score and average Math score. (Student 39, 14, 44, 18, 55)

   (h) Create the visualization (found below) that compares numeric scores across grade levels and subjects. Write 2–3 sentences describing what the graph suggests.

2. Using the `county_complete` dataset (only looking at the `state`, `name`, and any column containing `"pop"`) in the `usdata` library, tidy the population data so that year is treated as a variable rather than being embedded in the column names. Then answer the following questions.

   (a) Determine the total U.S. county-level population in each available year.

   (b) For each county, determine the absolute population change from 2000 to 2019 and identify the 10 counties with the largest increase. (Maricopa County)

   (c) For each county, determine the percent population change from 2000 to 2017 and identify the 10 counties with the largest percent increase. (Pinal County– 141%)

   (d) How many counties had population decline from 2010 to 2019?

(e) For each state, determine how many of its counties increased in population from 2010 to 2019 and how many decreased.

(f) Create the visualization (found below) using your tidied population data to show how the counties in Maryland have changed over time. Write 2–3 sentences describing what the graph suggests.

3. Using the `billboard` dataset in the `tidyr` library, tidy the data so that week number is stored as a variable and chart position is stored as a value. Also note that a low `position` is better (considered higher on the chart). Then answer the following questions.

(a) For each song, determine the best chart position it ever reached.

(b) Which artist had the most songs appear in the dataset? (Jay-Z with 5)

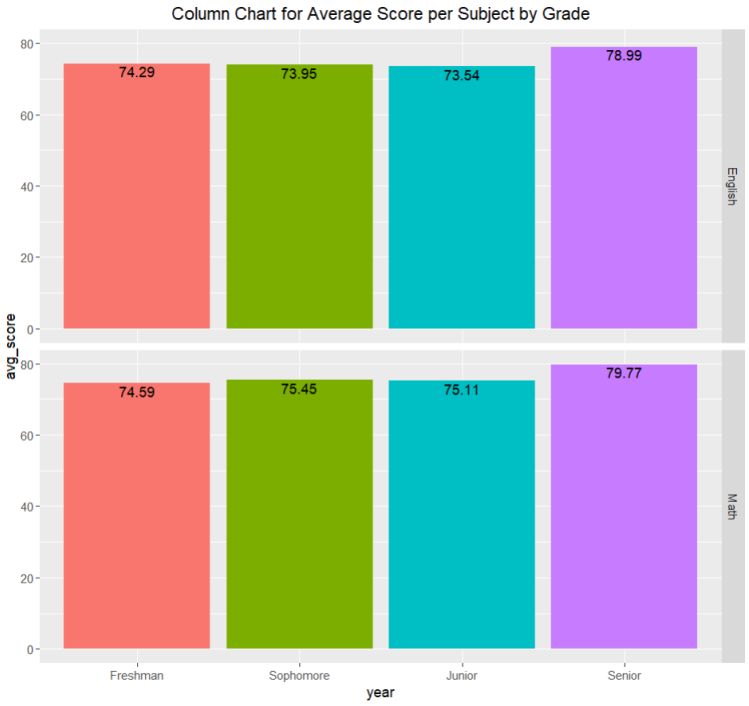
(c) Determine how many weeks each song remained on the chart. (ie was not NA)

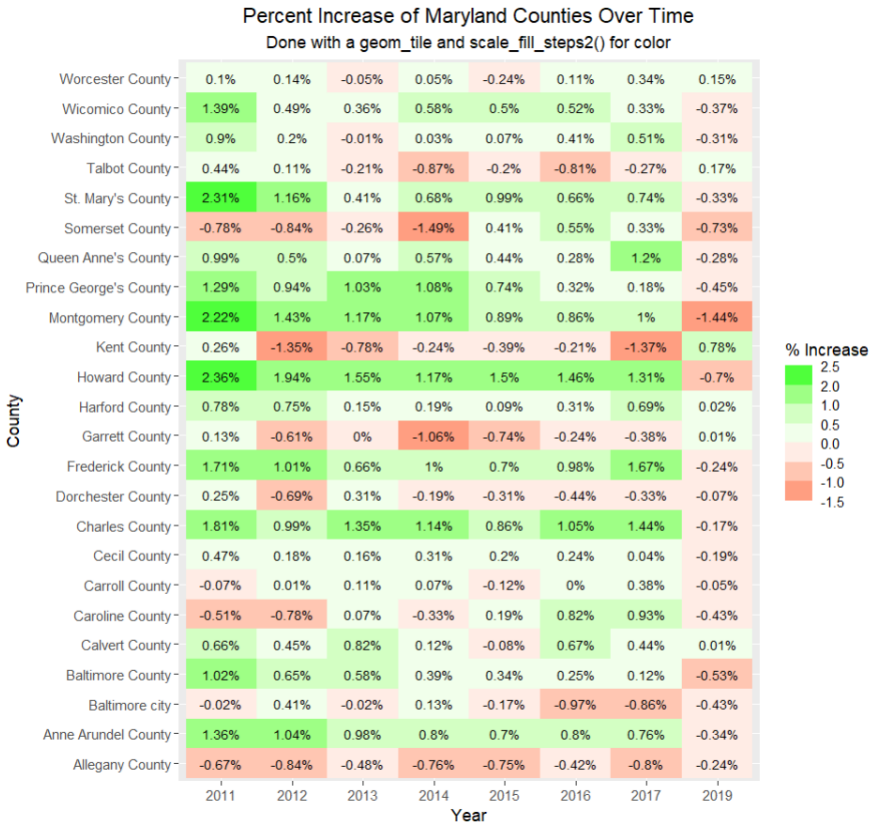
(d) Among songs that reached the top 10 at some point, which song stayed on the chart the longest?

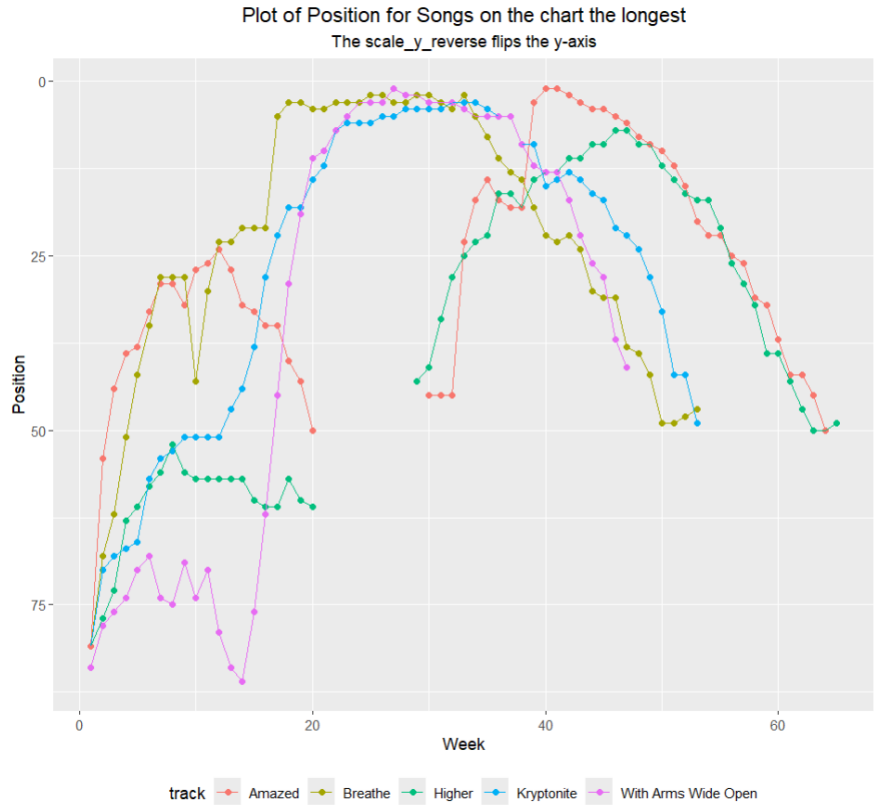
(e) For the artists with multiple songs on the list, determine which one has on average the best chart position across all of their songs. (Christina Aguilera and Destiny's Child with average top spot of 1.67)

(f) Determine which song had the largest improvement from its first recorded chart position to its best chart position. (He Loves U Not)

(g) Create the visualization (found below) using your tidied `billboard` dataset that shows how chart position changes from week to week for the 5 songs that stayed on the chart the longest. Write 2–3 sentences describing what your graph suggests.




Percent Increase of Maryland Counties Over Time
Done with a geom_tile and scale_fill_steps2() for color
Worcester County 0.1% 0.14% -0.05% 0.05% -0.24% 0.11% 0.34% 0.15%
Wicomico County 1.39% 0.49% 0.36% 0.58% 0.5% 0.52% 0.33% -0.37%
Washington County 0.9% 0.2% -0.01% 0.03% 0.07% 0.41% 0.51% -0.31%
Talbot County 0.44% 0.11% -0.21% -0.87% -0.2% -0.81% -0.27% 0.17%
St. Mary's County 2.31% 1.16% 0.41% 0.68% 0.99% 0.66% 0.74% -0.33%
Somerset County -0.78% -0.84% -0.26% -1.49% 0.41% 0.55% 0.33% -0.73%
Queen Anne's County 0.99% 0.5% 0.07% 0.57% 0.44% 0.28% 1.2% -0.28%
Prince George's County 1.29% 0.94% 1.03% 1.08% 0.74% 0.32% 0.18% -0.45%
Montgomery County 2.22% 1.43% 1.17% 1.07% 0.89% 0.86% 1% -1.44%
Kent County 0.26% -1.35% -0.78% -0.24% -0.39% -0.21% -1.37% 0.78%
Howard County 2.36% 1.94% 1.55% 1.17% 1.5% 1.46% 1.31% -0.7%
Harford County 0.78% 0.75% 0.15% 0.19% 0.09% 0.31% 0.69% 0.02%
Garrett County 0.13% -0.61% 0% -1.06% -0.74% -0.24% -0.38% 0.01%
Frederick County 1.71% 1.01% 0.66% 1% 0.7% 0.98% 1.67% -0.24%
Dorchester County 0.25% -0.69% 0.31% -0.19% -0.31% -0.44% -0.33% -0.07%
Charles County 1.81% 0.99% 1.35% 1.14% 0.86% 1.05% 1.44% -0.17%
Cecil County 0.47% 0.18% 0.16% 0.31% 0.2% 0.24% 0.04% -0.19%
Carroll County -0.07% 0.01% 0.11% 0.07% -0.12% 0% 0.38% -0.05%
Caroline County -0.51% -0.78% 0.07% -0.33% 0.19% 0.82% 0.93% -0.43%
Calvert County 0.66% 0.45% 0.82% 0.12% -0.08% 0.67% 0.44% 0.01%
Baltimore County 1.02% 0.65% 0.58% 0.39% 0.34% 0.25% 0.12% -0.53%
Baltimore city -0.02% 0.41% -0.02% 0.13% -0.17% -0.97% -0.86% -0.43%
Anne Arundel County 1.36% 1.04% 0.98% 0.8% 0.7% 0.8% 0.76% -0.34%
Allegany County -0.67% -0.84% -0.48% -0.76% -0.75% -0.42% -0.8% -0.24%
2011 2012 2013 2014 2015 2016 2017 2019
Year
County
% Increase
2.5
2.0
1.0
0.5
0.0
-0.5
-1.0
-1.5


Plot of Position for Songs on the chart the longest
The scale_y_reverse flips the y-axis
0
25
50
75
Position
0
20
40
60
Week
track
Amazed
Breathe
Higher
Kryptonite
With Arms Wide Open