project

Ella van Engen and Connor Wilchusky

3-21-21

```
set1 <- read.csv("~/R/project/data/set1.csv")
set2 <- read.csv("~/R/project/data/set2.csv")
set3 <- read.csv("~/R/project/data/set3.csv")
```

```
glimpse(set1)
```

```
#> Rows: 939
#> Columns: 17
#> $ X                          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13...
#> $ rank                       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13...
#> $ name                       <chr> "Auburn University", "University of Alaba...
#> $ state_name                 <chr> "Alabama", "Alabama", "Alabama", "Alabama...
#> $ early_career_pay           <int> 54400, 57500, 52300, 54500, 48400, 46600,...
#> $ mid_career_pay             <dbl> 104500, 103900, 97400, 93500, 90500, 8910...
#> $ make_world_better_percent <int> 51, 59, 50, 61, 52, 53, 48, 57, 56, 58, 6...
#> $ stem_percent               <int> 31, 45, 15, 30, 3, 12, 27, 17, 17, 20, 8,...
#> $ state                      <chr> "Alabama", "Alabama", NA, "Alabama", "Ala...
#> $ state_code                 <chr> "AL", "AL", NA, "AL", "AL", "AL", NA, "AL...
#> $ type                       <chr> "Public", "Public", NA, "Private", "Priva...
#> $ degree_length              <chr> "4 Year", "4 Year", NA, "4 Year", "4 Year...
#> $ room_and_board             <int> 13332, 10094, NA, 9650, 10550, 13462, NA,...
#> $ in_state_tuition           <int> 11276, 10714, NA, 22170, 31650, 39464, NA...
#> $ in_state_total             <int> 24608, 20808, NA, 31820, 42200, 52926, NA...
#> $ out_of_state_tuition       <int> 30524, 22362, NA, 22170, 31650, 39464, NA...
#> $ out_of_state_total         <int> 43856, 32456, NA, 31820, 42200, 52926, NA...
```

```
glimpse(set2)
```

```
#> Rows: 544
#> Columns: 21
#> $ X                <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
#> $ e                <int> 7136, 7137, 7138, 7139, 7144, 7145, 7146, 7147, 71...
#> $ country          <chr> "CA", "CA", "CA", "CA", "CA", "NL", "CA", "US", "U...
#> $ city             <chr> "Vancouver", "Toronto", "Toronto", "Toronto", "Tor...
#> $ line             <chr> "Broadway", "Vaughan", "Scarborough", "Ontario", "...
#> $ start_year       <chr> "2020", "2009", "2020", "2020", "2020", "2003", "2...
#> $ end_year         <chr> "2025", "2017", "2030", "2030", "2030", "2018", "2...
#> $ rr               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
#> $ length           <dbl> 5.7, 8.6, 7.8, 15.5, 7.4, 9.7, 5.8, 5.1, 4.2, 4.2,...
#> $ tunnel_per       <chr> "87.72%", "100.00%", "100.00%", "57.00%", "100.00%...
#> $ tunnel           <dbl> 5.0, 8.6, 7.8, 8.8, 7.4, 7.1, 5.8, 5.1, 4.2, 4.2, ...
#> $ stations         <int> 6, 6, 3, 15, 6, 8, 5, 2, 2, 2, 3, 3, 4, 7, 13, 4, ...
#> $ source1          <chr> "Plan", "Media", "Wiki", "Plan", "Plan", "Wiki", "...
#> $ cost             <dbl> 2830, 3200, 5500, 8573, 5600, 3100, 4500, 1756, 36...
```

```
#> $ currency         <chr> "CAD", "CAD", "CAD", "CAD", "CAD", "EUR", "CAD", "...
#> $ year             <int> 2018, 2013, 2018, 2019, 2020, 2009, 2018, 2012, 20...
#> $ ppp_rate         <dbl> 0.840, 0.810, 0.840, 0.840, 0.840, 1.300, 0.840, 1...
#> $ real_cost        <chr> "2377.2", "2592", "4620", "7201.32", "4704", "4030...
#> $ cost_km_millions <dbl> 417.05263, 301.39535, 592.30769, 464.60129, 635.67...
#> $ source2          <chr> "Media", "Media", "Media", "Plan", "Media", "Media...
#> $ reference        <chr> "https://www.translink.ca/Plans-and-Projects/Rapid...
```

```
glimpse(set3)
```

```
#> Rows: 32,833
#> Columns: 24
#> $ X                        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...
#> $ track_id                 <chr> "6f807x0ima9a1j3VPbc7VN", "0r7CVbZTWZgbTCY...
#> $ track_name               <chr> "I Don't Care (with Justin Bieber) - Loud ...
#> $ track_artist             <chr> "Ed Sheeran", "Maroon 5", "Zara Larsson", ...
#> $ track_popularity         <int> 66, 67, 70, 60, 69, 67, 62, 69, 68, 67, 58...
#> $ track_album_id           <chr> "2oCs0DGTsRO98Gh5ZSl2Cx", "63rPSO264uRjW1X...
#> $ track_album_name         <chr> "I Don't Care (with Justin Bieber) [Loud L...
#> $ track_album_release_date <chr> "2019-06-14", "2019-12-13", "2019-07-05", ...
#> $ playlist_name            <chr> "Pop Remix", "Pop Remix", "Pop Remix", "Po...
#> $ playlist_id              <chr> "37i9dQZF1DXcZDD7cfEKhW", "37i9dQZF1DXcZDD...
#> $ playlist_genre           <chr> "pop", "pop", "pop", "pop", "pop", "pop", ...
#> $ playlist_subgenre        <chr> "dance pop", "dance pop", "dance pop", "da...
#> $ danceability             <dbl> 0.748, 0.726, 0.675, 0.718, 0.650, 0.675, ...
#> $ energy                   <dbl> 0.916, 0.815, 0.931, 0.930, 0.833, 0.919, ...
#> $ key                      <int> 6, 11, 1, 7, 1, 8, 5, 4, 8, 2, 6, 8, 1, 5,...
#> $ loudness                 <dbl> -2.634, -4.969, -3.432, -3.778, -4.672, -5...
#> $ mode                     <int> 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, ...
#> $ speechiness              <dbl> 0.0583, 0.0373, 0.0742, 0.1020, 0.0359, 0....
#> $ acousticness             <dbl> 0.10200, 0.07240, 0.07940, 0.02870, 0.0803...
#> $ instrumentalness         <dbl> 0.00e+00, 4.21e-03, 2.33e-05, 9.43e-06, 0....
#> $ liveness                 <dbl> 0.0653, 0.3570, 0.1100, 0.2040, 0.0833, 0....
#> $ valence                  <dbl> 0.518, 0.693, 0.613, 0.277, 0.725, 0.585, ...
#> $ tempo                    <dbl> 122.036, 99.972, 124.008, 121.956, 123.976...
#> $ duration_ms              <dbl> 194754, 162600, 176616, 169093, 189052, 16...
```

Research Question for Dataset 1 (Colleges): Is there any correlation between mid-career salary and the percentage of of alumni who think they are making the world a better place across all North Carolina colleges? Hypothesis for Dataset 1: There will be no correlation between mid-career salary and the percentage of alumni who believe they are making the world a better place across North Carolina colleges.

Research Question for Dataset 2 (Transit): Is the cost per km of urban rail more dependent on the time length of the project or the total rail length of the project? Hypothesis for Dataset 2: Cost per km of urban rail is more dependent on the total rail length of the project.

Research Question for Dataset 3 (Spotify): For music tracks (tracks with a speechiness < 0.33), what is the correlation between energy and popularity? What is the mean energy for tracks that go #1? Hypothesis for Dataset 3: Songs with higher energy scores (> 0.8) are more likely to be more popular (< 25) than songs with low energy scores (0.2). The mean energy for tracks that go #1 is 0.8 or greater.