Topic ideas

STA 210 - Project

Soy nuggets - Madeleine Jones, Felix Hu, Isabel Siebrecht, Jason Yang

```
Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
had status 1
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
v ggplot2 3.3.5     v purrr   0.3.4
v tibble  3.1.6     v dplyr   1.0.8
v tidyr   1.2.0     v stringr 1.4.0
v readr   2.1.2     v forcats 0.5.1
```

```
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
Registered S3 method overwritten by 'tune':
  method                   from
  required_pkgs.model_spec parsnip
```

```
-- Attaching packages ---------------------------------------- tidymodels 0.1.4 --
```

```
v broom        0.7.12     v rsample      0.1.1
v dials        0.1.0      v tune         0.1.6
v infer        1.0.0      v workflows    0.2.4
v modeldata    0.1.1      v workflowsets 0.1.0
v parsnip      0.1.7      v yardstick    0.0.9
v recipes      0.1.17
```

```
-- Conflicts --------------------------------------- tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Dig deeper into tidy modeling with R at https://www.tmwr.org
```

Project idea 1: Lemurs

Introduction and data

- This dataset comes from Tidy Tuesday, and was collected by the Duke Lemur Center.
- The data were originally collected by the Duke Lemur Center (and cleaned by Jesse Mostipak) between 1966 and this past year (2021).
- The dataset includes variables that the Duke Lemur Center has decided affect the health, reproduction, and social dynamics of the lemurs the most. These variables include extensive data on a lemur's birth-place, various weights throughout their life, their offspring, taxonomic species, etc. There are both categorical (ex: a lemur's current or birth location) and numerical (ex: number of offspring) data, which are used together to comprehensively understand the factors impacting lemurs.

```
Rows: 82609 Columns: 54
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (19): taxon, dlc_id, hybrid, sex, name, current_resident, stud_book, es...
dbl  (27): birth_month, litter_size, expected_gestation, concep_month, dam_a...
date  (8): dob, estimated_concep, dam_dob, sire_dob, dod, weight_date, conce...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Research question

- How do certain characteristics of the lemurs (weight, taxonomy, age, etc) impact their offspring, overall health (as measured by lifespan, gestation, weight, etc)? What characteristics of their surroundings (birth location, current location) impact their physical attributes?

Glimpse of data

```
glimpse(lemurs)
```

```
Rows: 82,609
Columns: 54
$ taxon                   <chr> "OGG", "OGG", "OGG", "OGG", "OGG", "OGG"~
$ dlc_id                  <chr> "0005", "0005", "0006", "0006", "0009", ~
$ hybrid                  <chr> "N", "N", "N", "N", "N", "N", "N", "N", ~
$ sex                     <chr> "M", "M", "F", "F", "M", "M", "M", "M", ~
$ name                    <chr> "KANGA", "KANGA", "ROO", "ROO", "POOH BE~
$ current_resident        <chr> "N", "N", "N", "N", "N", "N", "N", "N", ~
$ stud_book               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ dob                     <date> 1961-08-25, 1961-08-25, 1961-03-17, 196~
$ birth_month             <dbl> 8, 8, 3, 3, 9, 9, 9, 5, 5, 10, 10, 6, 6,~
$ estimated_dob           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ birth_type              <chr> "CB", "CB", "CB", "CB", "CB", "CB", "CB"~
$ birth_institution       <chr> "Duke Lemur Center", "Duke Lemur Center"~
$ litter_size             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ expected_gestation      <dbl> 129, 129, 129, 129, 129, 129, 129, 129, ~
$ estimated_concep        <date> 1961-04-18, 1961-04-18, 1960-11-08, 196~
$ concep_month            <dbl> 4, 4, 11, 11, 5, 5, 5, 1, 1, 6, 6, 1, 1,~
$ dam_id                  <chr> "0001", "0001", "0001", "0001", "0001", ~
$ dam_name                <chr> "WHITE-TAIL", "WHITE-TAIL", "WHITE-TAIL"~
$ dam_taxon               <chr> "OGG", "OGG", "OGG", "OGG", "OGG", "OGG"~
$ dam_dob                 <date> 1959-01-28, 1959-01-28, 1959-01-28, 195~
$ dam_age_at_concep_y     <dbl> 2.22, 2.22, 1.78, 1.78, 4.32, 4.32, 4.32~
$ sire_id                 <chr> "0002", "0002", "0002", "0002", "0007", ~
$ sire_name               <chr> "BRUISER", "BRUISER", "BRUISER", "BRUISE~
$ sire_taxon              <chr> "OGG", "OGG", "OGG", "OGG", "OGG", "OGG"~
$ sire_dob                <date> 1959-01-28, 1959-01-28, 1959-01-28, 195~
$ sire_age_at_concep_y    <dbl> 2.22, 2.22, 1.78, 1.78, 4.32, 4.32, 4.32~
$ dod                     <date> 1977-02-07, 1977-02-07, 1974-10-15, 197~
$ age_at_death_y          <dbl> 15.47, 15.47, 13.59, 13.59, 10.38, 10.38~
$ age_of_living_y         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ age_last_verified_y     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, 14.1~
$ age_max_live_or_dead_y  <dbl> 15.47, 15.47, 13.59, 13.59, 10.38, 10.38~
$ n_known_offspring       <dbl> 7, 7, 9, 9, 1, 1, 1, 7, 7, 5, 5, 4, 4, 1~
$ dob_estimated           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ weight_g                <dbl> 1086, 1190, 947, 1174, 899, 917, 910, 11~
$ weight_date             <date> 1972-02-16, 1972-06-20, 1972-02-16, 197~
```

```
$ month_of_weight               <dbl> 2, 6, 2, 6, 2, 2, 6, 2, 6, 2, 6, 2, 6, 2~
$ age_at_wt_d                   <dbl> 3827, 3952, 3988, 4119, 3061, 3074, 3188~
$ age_at_wt_wk                  <dbl> 546.71, 564.57, 569.71, 588.43, 437.29, ~
$ age_at_wt_mo                  <dbl> 125.82, 129.93, 131.11, 135.42, 100.64, ~
$ age_at_wt_mo_no_dec           <dbl> 125, 129, 131, 135, 100, 101, 104, 92, 9~
$ age_at_wt_y                   <dbl> 10.48, 10.83, 10.93, 11.28, 8.39, 8.42, ~
$ change_since_prev_wt_g        <dbl> NA, 104, NA, 227, NA, 18, -7, NA, 51, NA~
$ days_since_prev_wt            <dbl> NA, 125, NA, 131, NA, 13, 114, NA, 125, ~
$ avg_daily_wt_change_g         <dbl> NA, 0.83, NA, 1.73, NA, 1.38, -0.06, NA,~
$ days_before_death             <dbl> 1818, 1693, 972, 841, 728, 715, 601, 208~
$ r_min_dam_age_at_concep_y     <dbl> 0.59, 0.59, 0.59, 0.59, 0.59, 0.59, 0.59~
$ age_category                  <chr> "adult", "adult", "adult", "adult", "adu~
$ preg_status                   <chr> "NP", "NP", "NP", "NP", "NP", "NP", "NP"~
$ expected_gestation_d          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ concep_date_if_preg           <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ infant_dob_if_preg            <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ days_before_inf_birth_if_preg <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ pct_preg_remain_if_preg       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ infant_lit_sz_if_preg         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

Project idea 2

Introduction and data

- This dataset comes from Tidy Tuesday, and was sourced from Billboard.com, Spotify.
- The data were originally collected for Data.World by Sean Miller using the Spotify Web API and Weekly Hot 100 Billboard chart in 2018.
- Aside from a song's performer, genre, duration, this dataset includes features about a song's popularity (number of weeks on the chart, peak position on chart) as well as Spotify song metadata (valence, loudness, danceability, etc.). Thus we see a good mix of categorical and numerical features.

```
Rows: 327895 Columns: 10
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (5): url, week_id, song, performer, song_id
dbl (5): week_position, instance, previous_week_position, peak_position, wee...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 29503 Columns: 22
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (7): song_id, performer, song, spotify_genre, spotify_track_id, spotify...
dbl (14): spotify_track_duration_ms, danceability, energy, key, loudness, mo...
lgl  (1): spotify_track_explicit

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Research question

- How do song characteristics such as energy, key, danceability affect its popularity? What are some characteristics of popular songs? Are there any trends in popular song characteristics over time?

Glimpse of data

```
glimpse(songs_joined)
```

```
Rows: 152,750
Columns: 29
$ url                          <chr> "http://www.billboard.com/charts/hot-100/197~
$ week_id                      <chr> "5/1/1971", "5/8/1971", "5/15/1971", "5/22/1~
$ week_position                <dbl> 61, 41, 32, 29, 22, 20, 17, 14, 13, 15, 26, ~
$ song                         <chr> "Don't Knock My Love - Pt. 1", "Don't Knock ~
$ performer                    <chr> "Wilson Pickett", "Wilson Pickett", "Wilson ~
$ song_id                      <chr> "Don't Knock My Love - Pt. 1Wilson Pickett",~
$ instance                     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ previous_week_position       <dbl> 96, 61, 41, 32, 29, 22, 20, 17, 14, 13, 15, ~
$ peak_position                <dbl> 61, 41, 32, 29, 22, 20, 17, 14, 13, 13, 13, ~
$ weeks_on_chart               <dbl> 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 2, 3, 4,~
$ spotify_genre                <chr> "['brill building pop', 'classic rock', 'cla~
$ spotify_track_id             <chr> "7cyLwgSVf3AnKXetNRWiTa", "7cyLwgSVf3AnKXetN~
$ spotify_track_preview_url    <chr> "https://p.scdn.co/mp3-preview/5d3332b4ae616~
$ spotify_track_duration_ms    <dbl> 136400, 136400, 136400, 136400, 136400, 1364~
$ spotify_track_explicit       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA~
$ spotify_track_album          <chr> "The Very Best Of Wilson Pickett", "The Very~
$ danceability                 <dbl> 0.731, 0.731, 0.731, 0.731, 0.731, 0.731, 0.~
$ energy                       <dbl> 0.701, 0.701, 0.701, 0.701, 0.701, 0.701, 0.~
$ key                          <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 11, 11, 11,~
$ loudness                     <dbl> -8.722, -8.722, -8.722, -8.722, -8.722, -8.7~
$ mode                         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1,~
$ speechiness                  <dbl> 0.0287, 0.0287, 0.0287, 0.0287, 0.0287, 0.02~
$ acousticness                 <dbl> 0.157, 0.157, 0.157, 0.157, 0.157, 0.157, 0.~
$ instrumentalness             <dbl> 6.75e-06, 6.75e-06, 6.75e-06, 6.75e-06, 6.75~
$ liveness                     <dbl> 0.0595, 0.0595, 0.0595, 0.0595, 0.0595, 0.05~
$ valence                      <dbl> 0.961, 0.961, 0.961, 0.961, 0.961, 0.961, 0.~
$ tempo                        <dbl> 107.521, 107.521, 107.521, 107.521, 107.521,~
$ time_signature               <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,~
$ spotify_track_popularity     <dbl> 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, ~
```

Project idea 3

Introduction and data

- The Bechdel dataset is from Tidy Tuesday, which got it from FiveThirtyEight, which sourced the data from the Bechdeltest.com API and the-numbers.com.
- Part of the data was collected by Bechdeltest.com. There, users submit movie entries. The other part was collected from the-numbers.com that stores movie business data. The two datasets' intersection spans 1990 to 2013.
- The dataset includes general movie data including budget, rating, genre, actors. There is also a Bechdel score attached showing how many of the Bechdel thresholds the movie meets. There is an even mix of numerical and categorical variables.

```
Rows: 8839 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): imdb_id, title
dbl (3): year, id, rating

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 1794 Columns: 34
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (24): imdb, title, test, clean_test, binary, domgross, intgross, code, d...
dbl  (7): year, budget, budget_2013, period_code, decade_code, metascore, im...
lgl  (2): response, error

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Research question

- How does a combination of budget, genre, runtime, and gender representation among other factors influence move ratings?

Glimpse of data

```
glimpse(movies)
```

```
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;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, ~
```