Project Proposal

Melanie Neller, David Khella, Alex dePillis-Lindheim, Daniel Rea

Wednesday, October 12, 11:59pm

Contents


```
library(tidyverse)
library("stringr")
```

```
# reading in data

chocolate <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/

broadway <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/1

colony <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/202

stressor <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/1
```

Proposal 1

Source of the data: It is from Tidyverse Tuesday who got the data from Playbill. Each observation represents a musical that was performed and includes information like where it was performed, in what year, and even in what theatre, etc..

Research Question: What have been the most successful Broadway musicals from 1985 to 2020? We should first decide how to define success, or explore different definitions of success (i.e. seats sold, money generated, or how long it ran for).

Hypothesis: The plays that are most successful given any one of the above definitions of success will also be most successful following the other definitions.

```
glimpse(broadway)
```

```
## Rows: 47,524
## Columns: 14
## $ week_ending          <date> 1985-06-09, 1985-06-09, 1985-06-09, 1985-06-09, ~
## $ week_number          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ weekly_gross_overall <dbl> 3915937, 3915937, 3915937, 3915937, 3915937, 3915~
## $ show                 <chr> "42nd Street", "A Chorus Line", "Aren't We All?",~
## $ theatre              <chr> "St. James Theatre", "Sam S. Shubert Theatre", "B~
## $ weekly_gross         <dbl> 282368, 222584, 249272, 95688, 61059, 255386, 306~
```

```
## $ potential_gross      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ avg_ticket_price     <dbl> 30.42, 27.25, 33.75, 20.87, 20.78, 31.96, 28.33, ~
## $ top_ticket_price     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ seats_sold           <dbl> 9281, 8167, 7386, 4586, 2938, 7992, 10831, 5672, ~
## $ seats_in_theatre     <dbl> 1655, 1472, 1088, 682, 684, 1018, 1336, 1368, 148~
## $ pct_capacity         <dbl> 0.7010, 0.6935, 0.8486, 0.8405, 0.5369, 0.9813, 1~
## $ performances         <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 0, 8, 8~
## $ previews             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8, 0, 0~
```

Proposal 2

Source of the data: The data comes from the USDA, hat tip to Georgios Karamanis. https://usda.library.cornell.edu/concern/publications/rn301137d?locale=en.

Research question: What is the relationship between stressors and the decline of bees colonies? Each observation in this data represents bee colonies in different states at different times of the year.

Hypothesis: Colonies with more stressors will tend to experience more decline than other colonies.

```
bees <- colony %>%
  left_join(stressor)

glimpse(bees)
```

```
## Rows: 7,332
## Columns: 12
## $ year            <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, ~
## $ months          <chr> "January-March", "January-March", "January-March", "Ja~
## $ state           <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Alabama",~
## $ colony_n        <dbl> 7000, 7000, 7000, 7000, 7000, 7000, 35000, 35000, 3500~
## $ colony_max      <dbl> 7000, 7000, 7000, 7000, 7000, 7000, 35000, 35000, 3500~
## $ colony_lost     <dbl> 1800, 1800, 1800, 1800, 1800, 1800, 4600, 4600, 4600, ~
## $ colony_lost_pct <dbl> 26, 26, 26, 26, 26, 26, 13, 13, 13, 13, 13, 13, 11, 11~
## $ colony_added    <dbl> 2800, 2800, 2800, 2800, 2800, 2800, 3400, 3400, 3400, ~
## $ colony_reno     <dbl> 250, 250, 250, 250, 250, 250, 2100, 2100, 2100, 2100, ~
## $ colony_reno_pct <dbl> 4, 4, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 1, 1, 1, 1, 1, 1, ~
## $ stressor        <chr> "Varroa mites", "Other pests/parasites", "Disesases", ~
## $ stress_pct      <dbl> 10.0, 5.4, NA, 2.2, 9.1, 9.4, 26.9, 20.5, 0.1, NA, 1.8~
```

Proposal 3

The source of the data is Flavors of Cocao; it's a website with information about different kinds of chocolate.

Research question: What is the correlation between the percentage of cocoa and the most memorable characteristics of the chocolate? Each observation represents a different chocolate from different sources and companies.

Hypothesis: Chocolates with a higher percent cocoa are more likely to be described as being rich.

```
glimpse(chocolate)
```

```
## Rows: 2,530
## Columns: 10
## $ ref                          <dbl> 2454, 2458, 2454, 2542, 2546, 2546, 2~
## $ company_manufacturer         <chr> "5150", "5150", "5150", "5150", "5150~
## $ company_location             <chr> "U.S.A.", "U.S.A.", "U.S.A.", "U.S.A.~
```

```
## $ review_date                      <dbl> 2019, 2019, 2019, 2021, 2021, 2021, 2~
## $ country_of_bean_origin           <chr> "Tanzania", "Dominican Republic", "Ma~
## $ specific_bean_origin_or_bar_name <chr> "Kokoa Kamili, batch 1", "Zorzal, bat~
## $ cocoa_percent                    <chr> "76%", "76%", "76%", "68%", "72%", "8~
## $ ingredients                      <chr> "3- B,S,C", "3- B,S,C", "3- B,S,C", "~
## $ most_memorable_characteristics   <chr> "rich cocoa, fatty, bready", "cocoa, ~
## $ rating                           <dbl> 3.25, 3.50, 3.75, 3.00, 3.00, 3.25, 3~
```