# Exploratory Data Analysis

*Kailen Shantz*

Last updated: 2019-02-17

In this notebook, I'll be exploring the data that has been scraped from the web and cleaned to get a sense of what features might do well at predicting a season's winner, how the data should be modeled, and what types of feature transforms might be useful.

```
# load libraries
library(here)
library(tidyverse)

# load data
rpdr <- readr::read_delim(here("data/processed/rpdr_df.txt"), delim = "\t")
```

To start, we'll take a look at the structure of the data and some summary statistics.

```
str(rpdr)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    129 obs. of  15 variables:
##  $ season             : num  1 1 1 1 1 1 1 1 1 2 ...
##  $ contestant         : chr  "Akashia" "BeBe Zahara Benet" "Jade" "Nina Flowers" ...
##  $ name               : chr  "Eric Flint" "Nea Marshall Kudi Ngwa" "David Sotomayor" "Jorge Luis Flo
##  $ age                : num  24 28 25 34 26 26 29 29 39 29 ...
##  $ hometown           : chr  "Cleveland, Ohio" "Minneapolis, Minnesota" "Chicago, Illinois" "Bayamón
##  $ was_winner         : logi  FALSE TRUE FALSE FALSE FALSE FALSE ...
##  $ n_episodes         : num  6 6 6 6 6 6 6 6 6 9 ...
##  $ n_contestants      : num  9 9 9 9 9 9 9 9 9 12 ...
##  $ n_appearances      : num  3 6 4 6 5 6 6 2 1 7 ...
##  $ n_lipsync          : num  3 1 1 0 1 2 2 1 1 1 ...
##  $ n_in_top           : num  0 3 0 5 4 2 2 0 0 4 ...
##  $ n_in_bottom        : num  4 1 3 1 2 3 4 2 2 2 ...
##  $ n_safe             : num  0 2 2 0 0 1 1 1 0 2 ...
##  $ n_wins             : num  0 2 0 1 2 1 0 0 0 1 ...
##  $ mini_challenge_wins: num  1 0 1 1 1 1 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   season = col_double(),
##   ..   contestant = col_character(),
##   ..   name = col_character(),
##   ..   age = col_double(),
##   ..   hometown = col_character(),
##   ..   was_winner = col_logical(),
##   ..   n_episodes = col_double(),
##   ..   n_contestants = col_double(),
##   ..   n_appearances = col_double(),
##   ..   n_lipsync = col_double(),
##   ..   n_in_top = col_double(),
##   ..   n_in_bottom = col_double(),
##   ..   n_safe = col_double(),
##   ..   n_wins = col_double(),
##   ..   mini_challenge_wins = col_double()
##   .. )
```

Unsurprisingly, `season` as been read in as a numeric variable. This shouldn't matter, but just to be safe let's change this to a character variable.

```
rpdr$season <- as.character(rpdr$season)
```

```
summary(rpdr)
```

```
##     season           contestant            name                age
##  Length:129         Length:129         Length:129         Min.   :21.00
##  Class :character   Class :character   Class :character   1st Qu.:25.00
##  Mode  :character   Mode  :character   Mode  :character   Median :28.00
##                                                           Mean   :28.95
##                                                           3rd Qu.:32.00
##                                                           Max.   :52.00
##    hometown         was_winner        n_episodes    n_contestants
##  Length:129         Mode :logical   Min.   : 6.00   Min.   : 9.00
##  Class :character   FALSE:119       1st Qu.: 9.00   1st Qu.:12.00
##  Mode  :character   TRUE :10        Median :12.00   Median :14.00
##                                     Mean   :11.02   Mean   :13.08
##                                     3rd Qu.:12.00   3rd Qu.:14.00
##                                     Max.   :12.00   Max.   :14.00
##  n_appearances      n_lipsync        n_in_top      n_in_bottom
##  Min.   : 1.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 4.000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:2.000
##  Median : 7.000   Median :1.000   Median :2.000   Median :3.000
##  Mean   : 6.938   Mean   :1.574   Mean   :2.341   Mean   :3.039
##  3rd Qu.:10.000   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :12.000   Max.   :4.000   Max.   :8.000   Max.   :8.000
##      n_safe          n_wins       mini_challenge_wins
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :2.000   Median :0.0000   Median :0.0000
##  Mean   :2.264   Mean   :0.8605   Mean   :0.6822
##  3rd Qu.:4.000   3rd Qu.:2.0000   3rd Qu.:1.0000
##  Max.   :6.000   Max.   :4.0000   Max.   :5.0000
```

## Exploring the raw data