

Wrangling Lab

Abbie M Popa

Set up

You will need the following packages for this lab

```
library(dplyr)
library(tidyr)
library(magrittr)
library(ggplot2)
```

Baby Names

Install and load the `babynames` package using the following code:

```
install.packages("babynames")
```

```
library(babynames)
```

Please comment out the “`install.packages`” line of code before knitting!

- (1) Output a tibble indicating the number of births there were for each year in the dataset.
- (2) Output a tibble indicating the number of unique names there were for each year in the dataset.
- (3) Make (and save) a tibble with a column for each year/sex pair in the dataset and a column for the count of each pair
- (4) Use a `tidyr` function to separate your year,sex interaction column into a column for year and a column for sex. (hint1: make sure the interaction column is a character string!) (hint2: “.” is a special character, like `_`, so will need to have two backslashes in front of it)
- (5) Make a bar plot showing the number of births each year where the color of the bar separates the counts by sex. “Dodge” your bars so they don’t overlap. (hint: look up how we used `stat = “identity”` in class to make a bar plot)

Movies

Install and load the `movies` data base using the following code

```
install.packages("ggplot2movies")
```

```
library(ggplot2movies)
```

Be sure to comment out the “`install.packages`” line before knitting!

- (6) Look at the `movies` data frame using `str()`
- (7) Using a `dplyr` function, efficiently find all the columns named `r#` (e.g., `r1`, `r2`, `r3`, ...)
- (8) Using a `dplyr` function, find the means of all the `r#` columns
- (9) If you did the above without using pipes `%>%` from the `magrittr` package now do it with pipes. If you did it with pipes above, now do it without pipes

- (10) Using a `tidyr` function, efficiently gather all the `r#` columns into two long columns where the key is named “rater” and the value is named “rating”. You can name the new data frame “long_movies”
- (11) Make a new even longer data frame “longer_movies” where the columns Action, Animation, Comedy, Drama, Documentary, Romance, and Short have the key “genre” and the value “encoding”
- (12) Remove all rows where encoding is equal to zero in longer_movies, leaving only rows where encoding equals 1. Also remove rows where the new rating column exceeds 15
- (13) If you did 12 the traditional way, now do it with filter, if you did 12 with filter, now do it the traditional way
- (14) Make a boxplot where each box represents a genre and the values are the ratings, please color your boxes by genre as well
- (15) Add a `facet_wrap` for rater
- (16) Which rater seems to like action movies more than they like other movies?