

R Programming: Worksheet 4. Tidyverse

By the end of today you should have experience using the tidyverse

You can install the tidyverse if not already installed using code like `install.packages("tidyverse")`

Most of this problem sheet will deal with the analysis of movies contained in `movies.csv`. You can download this from the github repo `ggplot2movies` through this URL <https://github.com/hadley/ggplot2movies>, specifically here <https://raw.githubusercontent.com/hadley/ggplot2movies/master/data-raw/movies.csv>. If you any trouble downloading the file, it is available on the course website as well.

1. Reading in data

- (a) Read the movie directly from the URL into R using the `readr` package and assign it to a variable named `movies`. Compare how long it takes to read the data using `readr` versus the relevant base R command on your machine.
- (b) Read the description file `movie_description.csv` into R. Have a look at the contents to understand what the columns represent.

2. Basic summaries

- (a) Use the `str` command to have a look at the structure of `movies`. How many movies (rows) are there in the dataset? How many columns are there?
- (b) Count how many times each of the different MPAA ratings are seen in the `mpaa` column. Try to do this using base R as well

3. Relationship between budget and ratings

- (a) Make a histogram using `ggplot2` of the ratings (`ratings` column, representing IMDB user ratings) of the movies
- (b) Similarly, make a histogram of the `budget` column.
- (c) Visualize the relationship between budget and rating using a 2D scatterplot. Transform the x-axis to log10 space to better visualize log budget films.
Use `scale_x_continuous(trans='log10')` to transform the x-axis
- (d) Is there a relationship between budget and rating? Confirm this by performing a linear regression base R (use R code like `lm(y ~ x, data)` to perform a regression of y on x , both of which are columns of `data`).

4. Querying the dataset

- (a) What are the two movies with the largest budget in the database?
- (b) What are the three movies with the largest budget from the 1990's?
- (c) Which movie titles are represented the most in the dataset? Try to do this using both base R and the tidyverse. Why do movie titles appear more than once in the dataset?

5. Relationship between budget and movie type

- (a) In `movie` there are 7 movie types (Action, Animation, Comedy, Drama, Documentary, Romance and Short), represented by binary variables. Each movie can have more than one type. Make a new variable called `category_count` that counts the number of categories each movie has. How many movies are only of one type?
- (b) Building on the above, among movies that are uniquely of one type (Action, Animation, Comedy, Drama, Documentary, Romance and Short), which movie type has the smallest average budget? *Hint, one option, first use `gather` to make a longer version of `movie`, and then use `group_by` and `summarize`*
- (c) Make a barplot of budget by movie type for movies that are uniquely of one type using `ggplot2`

6. Join with US presidents

- (a) Read in data about US presidents from `presidents.csv` found on the course website, and assign it to the variable `presidents`.
- (b) Augment `movies` into `movies_plus` with the information from `presidents_per_year`, using a relevant `join` command.
- (c) For which of the two political parties (Democratic or Republican) are movies rated on average higher (the `rating` column)?
- (d) During which president's time in office were movies on average the longest?
- (e) *Optional, advanced manipulation* The `presidents.csv` file was built from `presidents_raw.csv`. Here, starting from `presidents_raw.csv`, transform it into the same form as what you loaded with `presidents.csv`, with one row per year with the US president and their party. Assign a president's tenure to years by starting with the year of their inauguration, and ending with their last full year in office (for example George Washington with a tenure of 30/04/1789 4/03/1797 would get assigned to years 1789 to 1796 inclusive).