Tutorial-05

Kate Saunders

Table of contents

Learning Objectives

- Practice reading in data and specifying the data type.
- Practice data wrangling using dplyr
- Learn about conventions naming variables in R

Prepartion

- Download the datasets from Moodle `swiss_exports.csv` and `apple_options.csv`
- Remember to set up and use an R project.
- Create a directory for your data within the R project
- You may also find this data wrangling cheatsheet useful

Exercise 1: Swiss exports data

The file `swiss_exports.csv` contains the export data for Switzerland. Each row represents a different date. The first column is the `Date` variable, the second column is the `Year` only and each remaining column measures exports to a different country. The country names are represented using 2 letter code.

1. Read the data into R.

```
swiss_wide <- read_csv(here('data/swiss_exports.csv'))
```

2. Get the data into long form using the `pivot_longer` function. (Hint: sometimes it's easy to say what you want don't want to pivot.)
3. Using `group_by` and `summarise` create a new data set of yearly aggregate exports to each country. Does having a long form data set help with this?
4. Now produce a scatter plot on a log-log scale of 1988 exports against 2018 exports. This is a little tricky:

- You may need to reformat your data again - think about your desired aesthetic mappings
- Look into what to do if you have a variable name that's a number
- Also see `?scale_x_log`

5. Produce the same plot but remove all countries for which exports are zero in either 1988 or 2018.

Exercise 2: Options data

The following example uses Options data from Yahoo Finance. The owner of an put option has the right (but not the obligation) to sell stocks at a predetermined price (the `Strike Price`) on some fixed date (the `Expiry date`). A call option is the same but gives the owner the right to buy stocks.

The objective of this exercise is to produce the well-known *volatility smile* result from finance. This result states that for a given `Expiry date`, a plot of `Implied Volatility` against `Strike Price` is U-shaped. Implied volatility is the volatility of a stock that is computed from stock option data assuming a specific pricing model.

The standard naming in R is snake case (variable_name), where words are separated with underscores. The names in this data set are not saved in snake_case - they have spaces between the words! To use them in R code, you need to put the name of the variable in ticks `` `variable name` ``. You can find this symbol at the top left-hand corner of your keyboard. Working with names having spaces like this is quite difficult and prone to errors. You could try modifying the column names to make them into snake case at the end of the tutorial.

1. Read the data from this csv file into R.
2. The `Implied Volatility` has been imported as a character variable. To plot this it must be converted to a numeric variable. Create this using the `mutate` function.

Hint: The following code removes the percentage sign, converts to numeric and divides by 100.

```
str_replace('25%', '%', "") |> as.numeric() / 100
```

```
[1] 0.25
```

```
str_replace('1.32%', '%', "") |> as.numeric() / 100
```

```
[1] 0.0132
```

Create the new variable.

3. The volatility smile is best observed when options with a single expiry date are used. To use as much data as possible, find the expiry date that has the most 'put' options. To do this, you might use the `n()` function, which counts the number of observations in each group.
4. Options that are very far *out of the money* (very low strike price for a put option) should be excluded from the analysis. Building on previous answers, construct a data frame that only keeps put options from the expiry date in your answer to question 3, and that have a `Strike Price` above 250.

> **Note**
>
> Note that the `filter` function could use the `&` operator as well.

5. Using the data constructed in question 4, find the median value of `Implied Volatility` for each `Strike Price`.
6. Plot `Implied Volatility` against `Strike Price` using a line plot.

Exercise 3: First Normal Form

Discuss whether the following databases satisfy first normal form.

Database A:

```
# A tibble: 3 x 2
  Name        `Social Media Username`
  <chr>       <chr>
1 Jane Smith  Facebook: jsChampion
2 Kamal Usman Twitter: kusman, LinkedIn: ku87
3 Li Xiao     WeChat: lx99
```

Database B:

```
# A tibble: 4 x 2
  Name        `Social Media Username`
  <chr>       <chr>
1 Jane Smith  Facebook: jsChampion
2 Kamal Usman Twitter: kusman
3 Kamal Usman LinkedIn: ku87
4 Li Xiao     WeChat: lx99
```

Extra: Clean variable names in options data

Install the `janitor` package and load it into R. Learn how to use the `clean_names` function to create (clean) new column names.