R for data science notes

Tyler Nardone

2022-09-06

Two types of questions that are always useful for making discoveries in data: 1. What is the variation within my variables? 2. What is the co-variation between my variables?

Histogram can be used to view the distribution of a continuous variable.

EDA questions

Which values are most common? why?

Which values are rare and why?

Are there any unusual patterns, what might explain them?

Identify clusters. How are these observations similar to one another?

What might explain the clusters? Is the apperance of clustering misleading?

Unusual values. Outliers. repeat analyses w/ and w/o outliers to see how robust results are

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

General Form of ggplot:

```
ggplot(data = <DATA>) +
 <GEOM_FUNCTION>(
 mapping = aes(<MAPPINGS>),
 stat = <STAT>,
 position = <POSITION>
 ) +
 <COORDINATE_FUNCTION> +
 <FACET_FUNCTION>
```

dplyr basics

dplyr is a grammar of data manipulation. In this chapter you are going to learn the **five key dplyr functions** that allow you to solve the vast majority of your data manipulation challenges:

mutate() Adds new variables that are functions of existing variables select() Picks variables based on their names filter() Picks cases based on their values summarise() Reduces multiple values down to a single summary arrange() Changes the ordering of the rows

These can all be used in conjunction with group_by() which changes the scope of each function from operating on the entire dataset to operating on it group-by-group. These six functions provide the verbs for a language of data manipulation.

All verbs work similarly:

The first argument is a data frame.

The subsequent arguments describe what to do with the data frame, using the variable names (without quotes).

The result is a new data frame.

Together these properties make it easy to chain together multiple simple steps to achieve a complex result. Let's dive in and see how these verbs work.

Workflow: projects

Dont want to save R environments, want to save scripts and data files that recreate the environment when run. Set RStudio to NOT preserve workspaces between sessions. This forces you to capture everything you want to have code instructions to recreate it saved.

Tibbles

Tibbles are dataframes, with some tweaked behaviors.

```
library(tidyverse)
vignette("tibble")
```

```
## starting httpd help server ... done
```

```
as_tibble(iris)
```

```
## # A tibble: 150 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl> <fct>
##  1          5.1         3.5          1.4         0.2 setosa
##  2          4.9         3            1.4         0.2 setosa
##  3          4.7         3.2          1.3         0.2 setosa
##  4          4.6         3.1          1.5         0.2 setosa
##  5          5           3.6          1.4         0.2 setosa
##  6          5.4         3.9          1.7         0.4 setosa
##  7          4.6         3.4          1.4         0.3 setosa
##  8          5           3.4          1.5         0.2 setosa
##  9          4.4         2.9          1.4         0.2 setosa
## 10          4.9         3.1          1.5         0.1 setosa
## # ... with 140 more rows
```

data import - readr

`read_csv()` - reads comma delimited files `read_fwf()` - reads fixed width files `write_csv()` - writes csv file

feather is a package that implements a fast binary file, shareable across programming languages.

```
libary(feather)
write_feather()
read_feather()
```

Tidy Data - tidyr

Three rules which make a dataset tidy: 1. Each variable must have its own column 2. Each observation must have its own row 3. Each value must have its own cell

relational data

regular expressions

Factors

Factors are used to represent categorical variables with a known set of possible values. ## Dates and times