dplyr - 2

Lecture 5

Federica Fusi

MSCA, UIC

Updated: 2021-02-09

MISSING VALUES

Missing values

What is a missing value (or missing data)?

It's a value that we don't know - it's not a 'not applicable' (despite the name), it's not a negative answer nor a zero. It's an unknown value.

Why do we have missing values in a dataset?

- A survey respondent didn't provide an answer (e.g., skipped a question)
- It's was too expensive / too difficult to collect data
- We cannot retrieve information...

In sum: there has been issues with data collection or managment. It's totally normal to have missing values.

The key question is *what do you do with missing values in your dataset?*

Missing values in R

Missing values are marked as NA in R.

R assumes that if a value is missing, the result of any operation will be a missing value.

```
sum(c(9, 4, 6, NA))

5 == NA

NA + 10

NA == NA

mean(c(10, 14, 67, NA))
```

This is easy to understand if you substitute NA with "I don't know" in your mind: 10 + "I dont know" = I dont know!

> Note that NA is not a character value, in fact: NA is different than "NA".

Finding missing values

How do you find missing values?

The function is.na() identifies missing values

```
# Filter observation with a missing value in the arr_delay columnn
filter(data, is.na(dep_delay))

# Filter observation WITHOUT a missing value in the arr_delay column
filter(data, !is.na(arr_delay))

# If you want a more straightforward answer

data %>%
summarize(miss_value_length = sum(is.na(arr_delay)))

# You can look up multiple varialbes
data %>%
summarize(miss_value_arrdelay = sum(is.na(arr_delay)),
          miss_value_depdelay = sum(is.na(dep_delay)))
```

Quick note

Check that missing values are correctly coded as NA - some softwares use different symbols (.), sometimes survey data uses other annotations (999) and so on.

When your data look "weird" - i.e., they do not meet your expectations - you should always investigate more.

Missing value patterns'

If you took a statistical class, you might have heard some of these expressions:

MCAR, Missing Completely At Random: No relationship between missing values and observed or unobserved data

MAR, Missing At Random: Relationship between missingness and observed data, but not unobserved data

MNAR, Missing Not At Random: Relationship between missingness and data

Missing value patterns

MCAR, Missing Completely At Random: No relationship between missing values and observed or unobserved data- Researcher accidentally doesn't record some of the values

- Some respondents skip some questions as they don't have time

Since they are randomly missing, you can safely remove them

Missing value patterns

MAR, Missing At Random: Relationship between missingness and observed data, but not unobserved data

- Men are less likely to complete surveys on depression but not because of depression (e.g., maleness makes them less likely to do so)
- Women are less likely than men to report their weight

Quite safely ignore, but make sure to analyze some of these issues! (e.g., look at differences between male and female in the case above or, if you are doing regression, you can control for sex)

Missing value patterns

MNAR, Missing Not At Random: Relationship between missingness and data

- Depressed individuals are less likely to fill out a survey about depression because of their depression
- Individuals with high income are less likely to answer questions about their income than individuals with higher income

Do not ignore!!

- Explain why it is missing and incorporate your observations in the analysis
- If running a regression: ways to impute missing values

Diagnostic

It's very difficult to establish whether missing values are MCAR, MAR, MNAR. Some ideas:

1. Substantive literature on the topic
2. Check your data (e.g., t-test to see if male or female respondents have a greater number of missing values)
3. If you are doing more refined analysis, there are a few R packages that can help you A quick guide is available here

Missing values in practices

For the work we do in this class (and what you most likely will do in a policy / management job):

- Do not ignore them by default: they might influence your analysis. You want to be aware of missing values in your dataset and consciously decide what to do with them. This is why R will always warn you about missing values.
- Do some thinking to see if you recognize any patterns
    - What if all missing values are associated with women? Or homeless populations?
    - Consider if you can re-code them
- In most cases, you will be able to safely ignore them
    - Consider reporting them in your report / policy analysis memo

Looking for patterns

[Example from ModernDive]: Say a doctor is studying the effect of smoking on lung cancer for a large number of patients who have records measured at five-year intervals. She notices that a large number of patients have missing data points because the patient has died, so she chooses to ignore these patients in her analysis. What is wrong with this doctor's approach?

Looking for patterns

When I first looked at the data for Assignment #4, there were several missing values in the region datasets.

I got curious and I started looking them up one by one (e.g., googling the country whose region was missing):

- Quickly realized that most of them were territories (e.g., American Samoa, Greenland)
- Some countries were not internationally recognized (e.g., Palestine)
- Some others have special status (e.g., Vatican City)

They aren't real missing values, so I decided to re-code them as "Territories/Others".

When you re-code data *make sure to leave a clear comment in your code and report this information in your methodological note*.

On why you have to pay attention

We have been using this dataset from the UN website - a pretty reliable source! And yet:

- Population numbers were tricky (were expressed in thousands)
- NAs in region were actually referring to a specific groups of countries (territories); they were not actually NAs! Just missign from the UN classfication
- Some other issues to be discovered in the assignment!

Ignoring missing values

In general, missing values are easily ignore in R:

- You can remove them from the entire dataset
- You can remove them from the operation peformed (na.rm = T)
- ggplot (to make graphs) eliminates rows with missing values
- Regression analysis (e.g., lm) eliminates observations with missing values by default

Ignoring missing values

The most common way to solve this problem is to tell R to ignore missing values

We use the option na.rm (you can memorize this as "remove NAs") and set it to TRUE

```
sum(c(9, 4, 6, NA), na.rm = T)
```

```
## [1] 19
```

```
mean(c(10, 14, 67, NA), na.rm = T)
```

```
## [1] 30.33333
```

Note what na.rm does: it doesn't include the missing value in its calculation. It does NOT substitute it with another number; it's entirely removed.

```
(10 + 14 + 67)/3
```

```
## [1] 30.33333
```

Other functions

If you are using pivot_longer , you can use values_drop_na = T to remove missing values from the new columns that you are creating

If you want to remove all rows with missing values from the dataset, you can use na.omit.

fill() replaces missing values with the closest non-missing value

complete finds unique combinations in a set of columnns so it ensures that the dataset contains a