Week 7 Assignment

PA 434

2/23/2020

Get the Data

Read the data with tidytuesdayR package

You can install the package from CRAN via: install.packages("tidytuesdayR") and then install the transit cost datasets.

```
tuesdata <- tidytuesdayR::tt_load('2021-01-05')

transit_cost <- tuesdata$transit_cost
```

Variables

Here is the full list of variables in the dataset.

variable	class	description
e	double	ID
country	character	Country Code - can be joined against countrycode via ecb or iso2c
city	character	City where transit tunnel is being created
line	character	Line name or path
start_year	character	Year started
end_year	character	Year ended (predicted or actual)
rr	double	I think this is Railroad (0 or 1), where 1 == Railroad?
length	double	Length of proposed line in km
tunnel_per	character	Percent of line length completed
tunnel	double	Tunnel length of line completed in km (can take this divided by length to get tunnel_per)
stations	double	Number of stations where passengers can board/leave
source1	character	Where was data sourced
cost	double	Cost in millions of local currency
currency	character	Currency type
year	double	Midpoint year of construction
ppp_rate	double	purchasing power parity (PPP), based on the midpoint of construction
real_cost	character	Real cost in Millions of USD
cost_km_millions	double	Cost/km in millions of USD
source2	character	Where was data sourced for cost
reference	character	Reference URL for source

Your task

The dataset is already tidy - each row is a transit project. There can be multiple projects in the same city and same country.

Your goal is to produce a nice plot that could be potentially used for explanatory purposes with your readers.

You can draw inspiration (but not plagiarize) from the TidyTuesday website - just select the Transit Cost Project from the dropdown menu and scroll the gallery. I have indicated some graphs that mostly utilize what we learned so far in the following section.

Submit your final code (with annotations) and your plot as an image

Where to start

Step 1

The starting point of any good data viz is a question! But... where do you get one?

- Some people are naturally curious and just by hearing the topic "transit project costs" have a milion questions in their mind. If you are one of those, pick 2-3 questions that you are thinking and check if you have variables available to answer them.
- Other persons work better with data in their hands... if you are one of those, start by looking at the dataset and focus your attention on one or two variables that are of interest to you. What questions would they allow you to explore? What type of information can you extract from there?

Once you have a question, write it down!

Step 2

Identify the variables that you need to answer your question - at this stage you probably need 1, 2, or 3 variables at most.

Check if these variables are already in the right format in the dataset... do you need to summarise your data by country? Do you need to create a percentage?

Write down the variable that you intend to use and what operations you need to perform to clean them up.

Step 3

Think about your plot. Start by going through the tools available to you at this stage (e.g., scatterplots, histograms, bar plots...). Decide which one is the most appropriate. Sketch out on a piece of paper how your graph would look like. Focus your attention on key elements only: what do the axes represent? What variables are plotted?

Step 4

Start working in R. Go back to step 2 and clean your data. Once your data are ready, move to step 3 and realize a simple exploratory plot to see how your data look like.

You might be satisfied by your plot and decide to move forward. It is also possible that your plot disappoints you and you will have to start again the process!

Step 5

Start improving your plot by doing basic adjustements: fix your axes, put labels, insert your title and subtitle, check out axes labels and ticks...

Step 6

Before working on colors and themes, I would encourage you to look at the aes parameters. Is there anything else that you could visualize in your plot? E.g., change the size or shape of your dots in a scatterplot; split your bars in a bar plot; and so on.

You should also think about vlines and hlines that might help you tell a story to your audience - what is the main findings that you want your audience to see right away?

Try to think about the data once more and the story that you want to tell before making your graph prettier.

Step 7

Pick color for your graph. Adjust borders and fonts. Add a theme that you like and that highlights your findigs.

Step 8

Save your plot. Go back to it a few hours later. Can you tell the main message right away? Show it to someone else: can they clearly draw the same information from the plot?

Adjust your plot as necessary based on this feedback.

Some ideas

Here are some ideas of this assignment. You should feel free to explore one of them or come up with other idea of your own. All these ideas can be applied at different levels of analysis (e.g., project, country, city, continent. . .)

1. *What are the most and least costly transit infrastructure projects around the world?* Plot the correlation between cost of the project and lenght of the project (pay attention that the two unit matches). Example here. Another example here
    - Group your observations by country or continent OR
    - Draw a regression line OR
    - Highlight only observations from the United States while making the others less visible OR
    - Highlight the most costly projects OR
    - Change the size of your dots so that they are representative of the number of stations for that project OR
    - Change the size of your dots to represent the status of advancement of the project
2. *Pick a city with multiple transit project - What transit projects are going on in your city?* Visualize the length, costs, or duration of each project
    - Create a graph representing the status of advancement of each project (e.g., percent of completed lines). Example here
    - Visualize the duration of each transit project (for this one, have a look at geom_segment). Example here
    - Correlate the number of stations with the lenght of each line. You could map the cost of each project as the dot size. Example here.
3. *What are the most efficient projects?* Plot the correlation between the yearly expected costs and the expect progress per year (i.e., line km realized per year).
    - Highlight the most efficient projects
    - Dot size could be representative of the cost
    - Group them by continent