Bush 631: Research Design Task

Working with R for data analysis

Instructions

This document provides you with instructions regarding the first research design task - working with R.

Below, you will find multiple questions in which you are asked to apply R to analyze a data set.

For this task, you may 'brainstorm' the solutions with your peers, however you must submit **your own solutions and code**.

For each question, you are asked to provide the answer as well as the code using the R Markdown template.

The deadline for submitting the task is **October 4, 2022**.

The Data

The data for this assignment is based on a research that explores the effect of individual leaders on their countries international behavior. Primarily, the researchers were interested in learning whether change in leadership due to assassination had an impact on some aspects of states international status and actions.

The data *leaders.csv* contains information about assassination attempts of leaders. The variables included in this data:

Name	Description
`country`	The name of the country
`year`	Year of assassination
`leadername`	Name of leader who was targeted
`age`	Age of the targeted leader
`politybefore`	Average polity score during the 3 year period prior to the attempt
`polityafter`	Average polity score during the 3 year period after the attempt
`civilwarbefore`	1 if country is in civil war during the 3 year period prior to the attempt, or 0
`civilwarafter`	1 if country is in civil war during the 3 year period after the attempt, or 0
`interwarbefore`	1 if country is in international war during the 3 year period prior to the attempt, or 0
`interwarafter`	1 if country is in international war during the 3 year period after the attempt, or 0
`result`	Result of the assassination attempt, one of 10 categories described below

Questions

The tasks are ordered by categories.

Prepare R, upload data

Four tasks:

1. Upload the package tidyverse.
2. Upload the data file leaders.csv.
3. What function help us learn of the size of the data? (number of rows and columns)
4. How can we see the 1st 6 observations (rows) of the data?

Explore data

Two tasks:

1. Use a function that provides basic summary information of **all the variables** in the data? What is the earliest *year* recorded in the data?
2. Apply a function that summarizes the information for the variable *age*.

Proportions in our data

Two tasks (to make it clearer, I recommend that you add labels to each column in your table):

1. Tabulate the data in a way that will present the proportions of the variables *civilwarbefore* and *civilwarafter*.
2. Tabulate the data in way that presents the proportions of *interwarafter* and the results of the assassination attempt.

Subsets and descriptive stats

You will need to create several subsets:

- **(1)**: create a subset and name it 'war_before'. It will include **only** the observations in which countries were involved in *international war* in the 3-year period *before* the assassination attempt.
- For this subset, calculate:
    1. The mean polity score in the 3-year period *before* the assassination attempt.
    2. The mean polity score in the 3-year period *after* the assassination attempt.
- **(2)**: create a subset and name it 'civil_after'. It will include **only** the observations in which countries were involved in a *civil war* in the 3-year period *after* the assassination attempt.
- Now, for this subset, calculate the **difference in means** between the polity score after and before the assassination attempt. Did the score increase or decrease? By how much?
- **(3)**: create a subset and name it 'civil_both'. It will include observations of countries that experienced a civil war **both before and after** the assassination attempt (hint: this subset is based on **two** variables).
- In the subset 'civil_both', what is the age of the youngest leader? What function did you use?
- How can we find the range of ages of the leaders in this subset?

- How can you check if the average age of leaders has the same value as the median age?

Working with NAs

The data includes a variable called *randomvar*.

- Use the function that allows you to see the 1st **10 observations** of the variable *randomvar*, and identify which observation is the missing value?
- How can we summarize the **total** number of missing values for the variable *randomvar*?
- How can we calculate the **proportion** of missing values for the variable *randomvar*?

Create variables and the tapply() function

using ifelse():

- Add to the leaders data set a variable called 'failed'. It will have the value of 1 if the result of the assassination attempt is "not wounded", otherwise the value of 'failed' will be zero (0).
- Tabulate the **proportions** of each value of the variable 'failed'. What was the percentage of successful assassination attempts in the data?
- Use the tapply() function to present the average polity score (use the variable *polityafter*) for each value/category of the *failed* variable? Was the polity score higher for failed or non-failed attempts?

Plotting

Using some of the subsets and variables you created earlier, you are asked to create the following plots (you can implement either the base R or ggplot approaches).

- Create a **Bar-plot** of the proportions of the variable *failed* in the full data. Here are the details for the plot:
  1. The label (name) of the x-axis is "Successful attempt?"
  2. The title of the plot is "How many assassination attempts failed?".
- Plot a **Histogram** of the *age* variable in the full data set. Here are the details for this plot:
  1. The label (name) of the x-axis is "Leaders Age"
  2. The title of the plot is "Leaders Age distribution".
  3. Add *a vertical line* that shows the mean value of the variable (I recommend to use a different color for the mean line).
- Create a **boxplot** that demonstrate the relationship between *interwarbefore* and leaders age variables. What is the approximate range of leaders' age for either value of the *interwarbefore* variable? (in other words, what is the IQR and how can we see it in this plot?).
- Create a **scatterplot** that shows the relationship between two variables: *politybefore* and *polityafter*. Few more details for your plot:
  1. The label (name) of the x-axis is "Polity before"
  2. The label (name) of the y-axis is "Polity After".
  3. Add a 45 degree line to your plot (use different color for this line).
  4. What type of relationship between both variables is evident in the plot?

Correlations

Three tasks:

1. What is the correlation between the *politybefore* and *polityafter* variables?
2. What is the correlation between the proportion of failed assassination attempts (the variable you created earlier) and the *polityafter* variable? How do we interpret this correlation? (i.e. "When X increase, Y increase/decrease")
3. What is the correlation between the age of leaders and the proportion of begin involved in an international war in the 3-year period *after* the assassination attempt?