Honors Assignment: Extracting and Exporting Data Subsets

Author's name here

2022-11-01

1 Intro

Welcome!

For this Honor's assignment, you will be carrying out a new task as a data analyst: preparing data subsets for **someone else to use**. It is a hands-on approach to using the `select()` and `filter()` verbs.

The assignment should be submitted individually, but you are encouraged to brainstorm with partners.

The final due date for the assignment is Tuesday, November 28th at 23:59 PM UTC+2.

2 Get the assignment repo

To get started, you should download and look through the assignment folder.

1. First download the repo to your local computer here.

You should ideally work on your local computer, but if you would rather work on RStudio Cloud, you can upload the zip file to RStudio Cloud through the Files pane. Consult one of the instructors for guidance on this.

2. Unzip/Extract the downloaded folder.

   If you are on macOS, you can simply double-click on a file to unzip it.

   If you are on Windows and are not sure how to "unzip" a file, see this image. You need to right-click on the file and then select "extract all".

3. Once done, click on the RStudio Project file in the unzipped folder to open the project in RStudio.

4. In RStudio, navigate to the Files tab and open the "rmd" folder. The instructions for your exercise are outlined there (these are the same instructions you see here).

5. Open the "data" folder and observe its components. You will work with the "rabies_dataset.csv" file. (You can also open the "00_info_about_the_dataset" file to learn more about this dataset.)

6. In the same folder, open the metadata (data dictionary) file, "rabies_metadata.pdf". This data dictionary is necessary because the data for today is numerically encoded; that is, all variables are stored as numbers which correspond to categories. Without the variable dictionary, you wouldn't know what the numeric categories correspond to!

3 Load and clean the data

Now that you understand the structure of the repo, you can load in and clean your dataset.

In the code section below, **load in the needed packages** (hint: load {tidyverse} and {here}).

Pro tip: Use `p_load()` to load in your packages, since it both loads and installs packages as needed.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

Now, **read the dataset into R**. The data frame you import should have 1466 rows and 23 columns. Remember to use the `here()` function to allow your Rmd to use project-relative paths.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

Next, perform the following two cleaning tasks on the imported dataset, then store the cleaned dataset in a new object.

- **Bring the respondent ID to the front**. The respondent ID is the 23rd column of your dataset. It should be more visible so that someone who opens up the CSV knows immediately that each row corresponds to a respondent. Move it to the first position in the data frame.

- **Remove the `Education` variable, which has not been properly encoded.** (To notice the encoding issues, look at both the data frame and the metadata file). We are considering this variable unusable, and hence removing it.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

4 Create and export data subsets

In each “Data subset” section below, you should

- decide whether to use the `filter()` or `select()` function, then apply that function to create the required extract of the dataset;
- export the data subset into an appropriately-named CSV file in the “data_exports” folder.
- NOTE: For all data subsets, always keep the respondent ID variable! Without it, you cannot link back your data to the original dataset and you lose crucial information.

4.1 Data Subset 1: Extract demographic information

Create and export a data subset of respondents’ demographic information—their age, gender and geographic background.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

Don’t forget that you were asked to include the respondent ID variable in all subsets.

Remember to use the `here()` function to allow your Rmd to use project-relartive paths.

4.2 Data Subset 2: Extract all male adults from the dataset

Create and export a data subset containing only records for males aged over 18.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

(Hint: this is a row-filtering question, so there is no need to select or drop columns.)

4.3 Data Subset 3: Extract at-risk individuals

Create and export a subset with “at-risk” individuals. These are people who a) have a pet at home, b) have no access to health facilities, and c) consider that the rabies vaccine is not affordable for them.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

4.4 Data Subset 4: Extract the knowledge-evaluation survey question variables

Reading the metadata, you will see that some variables correspond to “knowledge-evaluation” questions about rabies. Create and export a subset that includes these variables.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

4.5 Data Subset 5: Extract respondents with “ideal” knowledge, attitudes and practices (KAPs) towards rabies

People with *ideal* KAPs are defined as people who answered that they:

- vaccinate their pets,
- know the clinical signs for rabies, and
- visit a doctor after being bitten by an animal

Create and export a data subset that includes just these individuals.

```
## [1] "WRITE_YOUR_CODE_HERE"
```

5 Submission: Upload zipped folder

Once you have finished the tasks above, you should **create a zipped file** of your project folder, containing all the work you have done, and **upload it** on the assignment page.