Work sample

Odyssey

2/9/2020

This test contains some key challenges we face daily in our work as analysts at Odyssey. We would like to see how you approach data cleaning and briefly test your analytical skills. Be as detailed as possible, comment your code and make plots if you find it adequate.

We have provided you with a data file (.sav format), census data (.xlsx format) and a questionnaire. The data set comes from a company within the online payment solution industry and our task for this client was to build a segmentation model based on needs attitudes to online shopping and payment solutions.

You are more than welcome to contact us with questions for help or tips. There are several "unusual" parts in this test which are common when working with surveys that are extra tricky for the inexperienced, and don't want this fact to stop you from trying to apply or finish the task rather contact us and discuss the issue (this is how we work in our everyday life). Either call or e-mail someone in the analytics team:

Contacts:

- Henrik Karlsson (henrik.karlsson@odyssey.se), 073-323 47 92
- Fredrik Augustsson (fredrik.augustsson@odyssey.se), 076-118 02 81
- Igor Gusev (igor.gusev@odyssey.se), 073-435 20 06

Recommended packages (but feel free to use whatever you prefer):

```
library(anesrake)
library(haven)
library(tidyverse)
```

1. Read an SPSS-file

Task:

- Read the data file

The file is .sav format which is the default format for SPSS-files. SPSS-files are still the main data format for surveys, as it enables you to store both the data and metadata of each variable, such as the question phrasing and the question alternatives.

TIP: use `haven::read_sav()` to read the file. It is usually quite hard to get a good overview of survey data files due to the large number of variables. The function `sjPlot::view_df()` gives a nice overview of the data file.

```
# Your code here
```

2. Survey weigthing

Task:

- Weight the data so it is representative to age and gender, so the sample file have the same distribution as the target population. Save and append you final weight (for each individual) in the same data frame names as "weight".

We would like you to compute weights for the sample to ensure that the data file have the correct proportions for age and gender. Please use the age groups:

- 18-25
- 26-35
- 36-45
- 46-55
- 56-65
- 65+

You'll need to aggregate the census data provided to the correct age groups and then compute the target proportion for the population. Thereafter, you'll need to compute the weight for each individual in the sample (which is done by `anesrake::anesrake()`.

Background: The purpose of our surveys is to do inference on the population, but data collection is hard and some groups might be easier than others to collect data from. Therefore, it is standard to weight the survey data before the analysis starts which is a way to adjust the sample so the proportion of the dimensions matches the target population. If the concept of survey weighting is new to you, you can get an introduction here. In short, once the weight is computed, you'll use the weight when analyzing data or making inference from the data set. This enables us to count individuals with features underrepresented in the data by a factor higher than 1 meanwhile overrepresented individuals have a factor below 1.

TIP: use the function `anesrake::anesrake()` to weight the data. The package is not the best and especially not the error messages. If you prefer another package for weighting, please use that one instead. To get `anesrake` to function properly, ensure that your weighting target is within a named list, that contains a named vector with the target proportions. Make sure that the data file contains variables whose name matches the named levels of the list and that the variable is a factor variable where each factor level is present in the target vector. The data frame must be class "data.frame" (achieved by `as.data.frame()`) and not a tibble which is the default data structure you get from `haven::read_sav()`. The parameter `caseid` in `anesrake()` should be a vector in format `df$id` and **not** a string argument with the variable name.

This task is hard, but you'll need a value in the weight variable for the next task. If you get stuck, reach out to us for help or create a variable called "weight" where each individual have value = 1.

```
# Your code here
```

3. Build a brand funnel

A brand funnel is a traditional group of questions in market research to understand the awareness and strength of a brand. It is based on three questions:

- Which of the following brands are you aware of (bf1_x)?
- Which of the following brands have you used (bf2_x)?
- Which of the following brands could you consider to use (bf3_x)?

The awareness question is given to all respondents with a set of brands as question alternatives. Then all the brands that the respondent is aware of are used as question alternatives in the second question if you've used any of the brands. The usage question is filtered as you can't be loyal to a brand if you are unaware of its existence, and therefore individuals that are unaware of a specific brand will have a missing value in the usage question. Lastly, we ask whether they consider using any brands they are aware of but do not currently use.

If you do not consider a brand, you will not use that brand. But due to how the data is collected, the consideration question (bf2_x) only contains individuals that consider a brand but do not currently use it. To make a correct brand funnel, the consideration question needs to be adjusted, so that everyone that consider **OR** use a brand should be selected (value = 1) in the consideration question.

If you didn't manage to compute the weight in the previous exercise, set the weight equal to 1 for everyone in this task.

Task:

- Recode the consideration question for each brand
- Visualize the brand funnel, either in counts (easier) or in percent (harder and preferable)

```
# Your code here:
```

4. Analysis of the needs attributes

Task:

- What are the key conclusions you can draw from data set? Give a brief description of the "needs" in the market. Is there a difference between different demographic groups or individuals prefering a specific brand?

TIP: The function `sjlabelled::get_label()` and `sjlabelled::as_label()` might be useful here.

```
# Your code here:
```