Survey Monkey - Data Transformation

Vish

24/08/2021

The following is my method of data transformation of SurveyMonkey. The methodology is from Shashank Kalanith. Please watch his video first **Day in the Life of a Data Analyst - SurveyMonkey Data Transformation (Using R)**. To get the original files please visit his github **kshashank03**.

```
library(dplyr)#data manipulating
library(tidyr)#to gather data (melt)
library(readxl)#to read xlsx files
library(openxlsx)#write to excel
```

Uploading File

```
pwd <- getwd()
pwd
```

```
## [1] "D:/Documents/Data Analyst/Data Transformation/Data Transform 2/SurveyMonkey-DataTransformation"
```

```
data_import.T <- read_excel("Data - Survey Monkey Output Edited.xlsx" , sheet = "Edited_Data")
```

```
dataset_modified.T <- data_import.T # Make a copy of the dataframe
colnames(dataset_modified.T)
```

```
##   [1] "Respondent ID"
##   [2] "Start Date"
##   [3] "End Date"
##   [4] "Email Address"
##   [5] "First Name"
##   [6] "Last Name"
##   [7] "Custom Data 1"
##   [8] "Identify which division you work in. - Response"
##   [9] "Identify which division you work in. - Other (please specify)"
##  [10] "Which of the following best describes your position level? - Response"
##  [11] "Which generation are you apart of? - Response"
##  [12] "Please select the gender in which you identify. - Response"
##  [13] "Which duration range best aligns with your tenure at your company? - Response"
##  [14] "Which of the following best describes your employment type? - Response"
##  [15] "Question 1 - Response"
##  [16] "Question 2 - Response"
##  [17] "Question 3 - Open-Ended Response"
##  [18] "Question 4 - Response"
##  [19] "Question 4 - Other (please specify)"
##  [20] "Question 5 - Response 1"
##  [21] "Question 5 - Response 2"
```

```
##  [22] "Question 5 - Response 3"
##  [23] "Question 5 - Response 4"
##  [24] "Question 5 - Response 5"
##  [25] "Question 5 - Response 6"
##  [26] "Question 6 - Response 1"
##  [27] "Question 6 - Response 2"
##  [28] "Question 6 - Response 3"
##  [29] "Question 6 - Response 4"
##  [30] "Question 6 - Response 5"
##  [31] "Question 6 - Response 6"
##  [32] "Question 7 - Response 1"
##  [33] "Question 7 - Unscheduled"
##  [34] "Question 8 - Response 1"
##  [35] "Question 8 - Response 2"
##  [36] "Question 8 - Response 3"
##  [37] "Question 8 - Response 4"
##  [38] "Question 9 - Response 1"
##  [39] "Question 9 - Response 2"
##  [40] "Question 9 - Response 3"
##  [41] "Question 9 - Response 4"
##  [42] "Question 10 - Response 1"
##  [43] "Question 10 - Response 2"
##  [44] "Question 10 - Response 3"
##  [45] "Question 10 - Response 4"
##  [46] "Question 10 - Response 5"
##  [47] "Question 11 - Reponse 1"
##  [48] "Question 11 - Response 2"
##  [49] "Question 12 - Response"
##  [50] "Question 13 - Response"
##  [51] "Question 14 - Response"
##  [52] "Question 15 - Response"
##  [53] "Question 16 - Response"
##  [54] "Question 17 - Response"
##  [55] "Question 18 - Response"
##  [56] "Question 19 - Response"
##  [57] "Question 19 - Other (please specify)"
##  [58] "Question 20 - Response"
##  [59] "Question 21 - Response"
##  [60] "Question 22 - Reponse 1"
##  [61] "Question 22 - Reponse 2"
##  [62] "Question 23 - Response"
##  [63] "Question 24 - Response 1"
##  [64] "Question 24 - Response 2"
##  [65] "Question 24 - Response 3"
##  [66] "Question 24 - Response 4"
##  [67] "Question 24 - Response 5"
##  [68] "Question 25 - Response 1"
##  [69] "Question 25 - Response 2"
##  [70] "Question 25 - Response 3"
##  [71] "Question 25 - Response 4"
##  [72] "Question 25 - Response 5"
##  [73] "Question 25 - Response 6"
##  [74] "Question 25 - Response 7"
##  [75] "Question 25 - Response 8"
```

```
##  [76] "Question 25 - Response 9"
##  [77] "Question 26 - Response 1"
##  [78] "Question 26 - Response 2"
##  [79] "Question 26 - Response 3"
##  [80] "Question 26 - Response 4"
##  [81] "Question 27 - Response 1"
##  [82] "Question 27 - Response 2"
##  [83] "Question 28 - Response"
##  [84] "Question 29 - Response 1"
##  [85] "Question 29 - Response 2"
##  [86] "Question 29 - Response 3"
##  [87] "Question 29 - Response 4"
##  [88] "Question 29 - Response 5"
##  [89] "Question 29 - Response 6"
##  [90] "Question 29 - Response 7"
##  [91] "Question 29 - Response 8"
##  [92] "Question 29 - Response 9"
##  [93] "Question 29 - Response 10"
##  [94] "Question 29 - Response 11"
##  [95] "Question 29 - Response 12"
##  [96] "Question 29 - Response 13"
##  [97] "Question 29 - Response 14"
##  [98] "Question 30 - Response 1"
##  [99] "Question 30 - Response 2"
## [100] "Question 30 - Response 3"
```

Remove Unwanted Columns

```
columns_to_drop.T <- c('Start Date', 'End Date', 'Email Address', 'First Name', 'Last Name', 'Custom Da
columns_to_drop.T
```

```
## [1] "Start Date"    "End Date"      "Email Address" "First Name"
## [5] "Last Name"     "Custom Data 1"
```

```
dataset_modified.T <- select(dataset_modified.T,-(columns_to_drop.T))
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(columns_to_drop.T)‘ instead of ‘columns_to_drop.T‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
dim(dataset_modified.T)
```

```
## [1] 198  94
```

```
id_vars.T <- colnames(dataset_modified.T)[1:8]
id_vars.T
```

```
## [1] "Respondent ID"
## [2] "Identify which division you work in. - Response"
## [3] "Identify which division you work in. - Other (please specify)"
## [4] "Which of the following best describes your position level? - Response"
## [5] "Which generation are you apart of? - Response"
## [6] "Please select the gender in which you identify. - Response"
## [7] "Which duration range best aligns with your tenure at your company? - Response"
## [8] "Which of the following best describes your employment type? - Response"
```

Transpose Columns

Melts all columns after the first 8 columns. Using `gather` from tidyr.

```
dataset_melted.T <- dataset_modified.T %>%
  gather("Question.+.Subquestion", "Answer", -id_vars.T)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(id_vars.T)‘ instead of ‘id_vars.T‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
dim(dataset_melted.T)
```

```
## [1] 17028    10
```

```
questions_import.T <- read_excel("Data - Survey Monkey Output Edited.xlsx", sheet="Question")
```

```
questions.T <- questions_import.T
dim(questions.T)
```

```
## [1] 100   5
```

```
str(questions.T)
```

```
## tibble [100 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Raw Question          : chr [1:100] "Respondent ID" "Start Date" "End Date" "Email Address" ...
##  $ Raw Subquestion       : chr [1:100] NA NA NA NA ...
##  $ Question              : chr [1:100] "Respondent ID" "Start Date" "End Date" "Email Address" ...
##  $ Subquestion           : chr [1:100] NA NA NA NA ...
##  $ Question + Subquestion: chr [1:100] "Respondent ID" "Start Date" "End Date" "Email Address" ...
```

Joining Tables

Join two datasets tables using `left_join` from dplyr library.

```
dataset_merged.T <- left_join(dataset_melted.T,questions.T,
                              by= c("Question.+.Subquestion" = "Question + Subquestion"))

dim(dataset_merged.T)
```

```
## [1] 17028    14
```

Find Number of Unique Respondents

Filter all `na` Answers.

```
respondents.T <- dataset_merged.T %>%
  filter(!is.na(Answer))

dim(respondents.T)
```

```
## [1] 9664   14
```

To find the unique respondents for each question we use `group_by` followed by `n_distinct` on the Respondent ID

```
respondents.T1 <- respondents.T %>%
    group_by(Question) %>%
    summarise(number_of_distinct_answers = n_distinct(`Respondent ID`))
```

```
dim(respondents.T1)
```

```
## [1] 30  2
```

```
str(respondents.T1)
```

```
## tibble [30 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Question                  : chr [1:30] "Question 1" "Question 10" "Question 11" "Question 12" ...
##  $ number_of_distinct_answers: int [1:30] 119 198 164 114 108 105 114 117 135 109 ...
```

We now merge the two tables so the number of unique respondents are shown along side the question.

```
dataset_merged_two.T <- left_join(dataset_merged.T, respondents.T1,
                                  by =c("Question"= "Question"))

dim(dataset_merged_two.T)
```

```
## [1] 17028    15
```

Find Number of Same Answers

```
same_answer.T <- dataset_merged.T %>%
  filter(!is.na(Answer))
dim(same_answer.T)
```

```
## [1] 9664   14
```

To find the same answers for each question we use `group_by` on both Question+Subquestion and Answer followed by `n_distinct` on the Respondent ID

```
same_answer.T1 <- same_answer.T %>%
 group_by(`Question.+.Subquestion`, Answer) %>%
  summarise(number_of_same_answer = n_distinct(`Respondent ID`))
```

```
## `summarise()` has grouped output by 'Question.+.Subquestion'. You can override using the `.groups` a
```

```
dim(same_answer.T1)
```

```
## [1] 688   3
```

Now merge the same answer table with the dataset. We use the columns 'Question.+.Subquestion' and 'Answer' as matching columns.

```
dataset_merged_three.T <- left_join(dataset_merged_two.T, same_answer.T1,
                                  by=c('Question.+.Subquestion', 'Answer'))

dim(dataset_merged_three.T)
```

```
## [1] 17028    16
```

```
colnames(dataset_merged_three.T)
```

```
##  [1] "Respondent ID"
##  [2] "Identify which division you work in. - Response"
##  [3] "Identify which division you work in. - Other (please specify)"
##  [4] "Which of the following best describes your position level? - Response"
##  [5] "Which generation are you apart of? - Response"
##  [6] "Please select the gender in which you identify. - Response"
##  [7] "Which duration range best aligns with your tenure at your company? - Response"
```

```
##  [8] "Which of the following best describes your employment type? - Response"
##  [9] "Question.+.Subquestion"
## [10] "Answer"
## [11] "Raw Question"
## [12] "Raw Subquestion"
## [13] "Question"
## [14] "Subquestion"
## [15] "number_of_distinct_answers"
## [16] "number_of_same_answer"
```

Rename Columns and Export File

```
dataset_merged_three.T <- dataset_merged_three.T %>%
  rename("Division" =`Identify which division you work in. - Response` ,
         "Division Other"=`Identify which division you work in. - Other (please specify)`,
         "Position"=`Which of the following best describes your position level? - Response` ,
         "Generation"=`Which generation are you apart of? - Response`,
         "Gender"=`Please select the gender in which you identify. - Response`,
         "Tenure"=`Which duration range best aligns with your tenure at your company? - Response`,
         "EmploymentType"=`Which of the following best describes your employment type? - Response`,
         "Respondents"=number_of_distinct_answers ,
         "SameAnswer"=number_of_same_answer)
```

Exporting transformed data.

```
write.xlsx(dataset_merged_three.T, paste(pwd , "/Final_Output_R_T1.xlsx", sep=""))
```