Character variables

Richard Careaga

2022-02-13

Purpose of this note

Study data has been collated from responses into an Excel spreadsheet.[1] The data are to be imported into an `R` data frame, processed, and exported to an SQL database. The appropriate type of each column must be mapped from `R` syntax to `SQL` syntax.

The plan is to use a first SQL data base as an archival record and a second SQL data base as a static data store for analysis.

In a Word document[2], **SB** identified 60 variables that should have been imported to `R` as type character, but found only 30. There are 318 variables in total from the spreadsheet. The Word document also identified variables that should be treated as factors.

Resolution

The variables identified by SB are

var	factor
dem_urn	FALSE
dem_gender	TRUE
dem_ethnicity	TRUE
dem_ethnicity_other	FALSE
dem_nationality	FALSE
dem_location	FALSE
dem_location_clean	FALSE
dem_sexuality	TRUE
dem_sexuality_other	FALSE
dem_relationship	TRUE
dem_relationship_other	FALSE
dem_kids	TRUE
Dem_access_kids	TRUE
dem_access_kids_other	FALSE
Dem_accomodation	TRUE
dem_accomodation_other	FALSE
Dem_education	TRUE
Dem_employment	TRUE
Dem_financial	TRUE
Dem_mh_diagnosis	TRUE
dem_mh_diagnosis_what	FALSE
dem_covid_me	TRUE

[1] **SB** provided **RC** with a small sample, consisting of seven rows, 2_DATA_Sample.xlsx, created 10/10/2021, 21:10:32 and modified 02/08/2022, 10:20:16

[2] Character Variables.docx, undated

var	factor
dem_covid_others	TRUE
dem_covid_impact	FALSE
dem_covid_1	TRUE
dem_covid_1	TRUE
dem_covid_1	TRUE
child_ctq_intro	FALSE
child_dce_intro	FALSE
child_mpe_f_intro	FALSE
child_mpe_m_intro	FALSE
child_open	FALSE
emot_depress_intro	FALSE
emot_masc_intro	FALSE
emot_express_intro	FALSE
pain_domains_intro	FALSE
pain_open	FALSE
pain_talk_who	FALSE
pain_talk_barriers	FALSE
pain_entrap_intro	FALSE
pain_defeat_intro	FALSE
pain_flood_numb_intro	FALSE
si_freq	FALSE
si_domains_intro	FALSE
si_open	FALSE
sa_freq	FALSE
sa_rfl_intro	FALSE
sa_support	FALSE
sa_open	FALSE
sa_helpseek_open	FALSE
self_se_intro	FALSE
self_slsc_intro	FALSE
self_satisf_intro	FALSE
social_supp_intro	FALSE
social_lone_intro	FALSE
social_matt_intro	FALSE
survey_open	FALSE
start_date	FALSE
complete_date	FALSE
suicide_flag	TRUE

There are 2 non-unique entries, reducing the number of character variables desired to 58.

Of these, 17 are `factor` variables, which are a type of numeric/character hybrid. They are recorded as numeric values that are labelled by characters. As they are not recorded in the spreadsheet as character values, but as numeric, these values will need to be matched with the corresponding character labels (or *levels*) separately.

Factors
dem_consent
dem_age
dem_location
dem_sexuality
dem_relationship

Factors
dem_relationship_other
dem_no_kids
dem_access_kids_other
dem_accomodation
dem_accomodation_other
dem_education
dem_financial
dem_mh_diagnosis
dem_covid_me
dem_covid_others
dem_covid_impact
child_dce_intro

Omitting the remaining number of character variables that are factors reduces the number of variables to be accounted for to 41.

Missing
Dem_access_kids
Dem_accomodation
Dem_education
Dem_employment
Dem_financial
Dem_mh_diagnosis
dem_covid_1
social_matt_intro

Differences in capitalization account for 6 of the missing variables and 2 are not present in the data. One of these social_matt_intro is mispelt social_matt_into, and the other, dem_covid_1 is not among the variables with similar names in the data.

Covid
dem_covid_me
dem_covid_others
dem_covid_impact
dem_covid_impact_1
dem_covid_impact_2
dem_covid_impact_3
pain_covid
si_covid

The variable name social_matt_into will be corrected in the process of creating the working SQL table. **SB** *should check* to see if dem_covid_1 was omitted from the data. *Otherwise*, **SB** and **RC** should compare versions of the data. If it has, **SB** should prepare a supplementary file with the columns dem_urn and dem_covid_1 from the source data. Accounting for the 2 missing variables and capitalized variables 33 variable remain to be reconciled.

First, however, the capitalized variables should be lower cased and checked against the variable names and types in the data. After doing so, only the two missing variables remain.

Missing
dem_covid_1
social_matt_intro

The differences between the character variables in the data and those in the running list are

Missing
si_age
social_close_vol
social_close_men
social_close_wom
social_close_nb

	Type
si_age	character
social_close_vol	character
social_close_men	character
social_close_wom	character
social_close_nb	character

These have been properly imported as type character. When those 5 variables are added to the 25 variables imported, the result, 30, agrees with **SB**'s result of character values found.

Recap

- **SB** specified character variables: 60
- Less duplicated entries: 2
- Less factor variables: `factrs`
- Less capitalized variables: 6
- Less mispelt in data: 1
- Less missing in data: 1
- Remaining: 33

In addition, **RC** sees only 25 character variables.

Found
dem_urn
dem_nationality
dem_location
dem_location_clean
dem_sexuality_other
dem_access_kids_other
dem_mh_diagnosis_what
child_open
pain_open
pain_talk_who
pain_talk_barriers
si_freq
si_age

Found
si_open
sa_freq
sa_open
sa_helpseek_open
social_close_vol
social_close_men
social_close_wom
social_close_nb
survey_open
start_date
complete_date
suicide_flag

To discuss: What additional character variables does **SB** expect that are still missing?