Second Wave of RCons Data Cleaning

Luke Sonnet

2019-05-07

```
library(haven)
library(tidyverse)
f28n <- read_dta("source_data/rcons_data/Form_28_NA_List.dta")
f28p <- read_dta("source_data/rcons_data/Form_28_PROVINCIAL_List.dta")
f45 <- read_dta("source_data/rcons_data/Form_45_Male_Female_Turnout.dta")
f48 <- read_dta("source_data/rcons_data/Form_48_ResultForm.dta")
f49 <- read_dta("source_data/rcons_data/Form_49_Candidate_List.dta")
```

Form 28s

First, I will rowbind (append) the form 28s for the PAs and the form 28s for the NAs.

```
f28 <- bind_rows(
  f28n %>% mutate(assembly = "National"),
  f28p %>%
    mutate(assembly = "Provincial") %>%
    select(-assembly_type)
) %>%
  mutate(constituency_ps_id = paste0(constituency_id, "_", ps_id))
```

Inconsistent names in form 28s

RCONS FIX: There are some polling stations that have two names per polling station ID. Below I only print the few examples, but all of the errors are in "f28_ps_name_error.csv". That file and the below data has the combinations of `constituency_id` and `ps_id` that have more than one `ps_name` within the polling station ID, which should be impossible. Please check the entries for these polling stations to see what is wrong with the names. Some are clearly just typos, other's appear to be different names altogether.

```
ps_name_error <- f28 %>%
  group_by(constituency_id, ps_id, ps_name) %>%
  summarize() %>%
  group_by(constituency_id, ps_id) %>%
  filter(n() > 1)
ps_name_error
```

```
## # A tibble: 284 x 3
## # Groups:   constituency_id, ps_id [142]
##    constituency_id ps_id ps_name
##    <chr>           <dbl> <chr>
##  1 NA1                21 Govt Middle School (GMS)
##  2 NA1                21 Jinjirate Koh
##  3 NA1                34 Regional Institute for Teachers
##  4 NA1                34 Training (RITT) Drosh
##  5 NA1                58 Govt Girls Community Modle
##  6 NA1                58 School (GGCMS) Sahan Payeen
##  7 NA107              102 GOVT. REFORMER GIRLS HIGH SCHOOL, CHAK NO. 220/RB~
##  8 NA107              102 GOVT. REFORMER GIRLS HIGH SCHOOL, CHAK NO. 220/RB~
##  9 NA12               232 GovernmentPrimary School
## 10 NA12               232 Toba Pashto
## # ... with 274 more rows
```

```
write_csv(ps_name_error, "f28_ps_name_error.csv")
```

RCONS FIX: There are some `constituency_id` files that have different `constituency_area` names. Some of these again are just typos, but others show bigger underlying problems with the data entry. Please check thoroughly and correct. This is all 10 of the errors (all the others are correct).

```
electoral_name_error <- f28 %>%
  group_by(constituency_id, constituency_area) %>%
  summarize() %>%
  group_by(constituency_id) %>%
  filter(n() > 1)
electoral_name_error
```

```
## # A tibble: 10 x 2
## # Groups:   constituency_id [5]
##    constituency_id constituency_area
##    <chr>           <chr>
##  1 NA58            RAWALPIDNI-II
##  2 NA58            RAWALPINDI-II
##  3 PP8             RAWALPINDI-III
##  4 PP8             RAWALPINDI-VII
##  5 PP99            BHAKKAR-III
##  6 PP99            FAISALABAD-III
##  7 PS48            MIRPURKHAS-II
##  8 PS48            MIRPURKHAS-II39603010
##  9 PS67            (HYDERABAD-VI)
## 10 PS67            TANDO MUHAMMAD KHAN-I
```

Mismatch between PA and NA polling stations

RCONS CHECK/FIX: There are some PA polling stations that correspond to multiple NA polling stations. Is this possible? Are there some PA polling stations that are used as multiple NA polling stations? This seems unlikely, so please check and report back. The file "f28_ps_id_mismatch.csv" has all of the PS polling stations which correspond to multiple NA polling station IDs, and I only print some examples below.

```
ps_id_mismatch <- f28 %>%
  group_by(constituency_id, ps_id, ps_id_NA) %>%
  summarize() %>%
  group_by(constituency_id, ps_id) %>%
  filter(n() > 1)
ps_id_mismatch
```

```
## # A tibble: 688 x 3
## # Groups:   constituency_id, ps_id [341]
##    constituency_id ps_id ps_id_NA
##    <chr>           <dbl>    <dbl>
##  1 PB1               133      322
##  2 PB1               133      323
##  3 PB15               25       57
##  4 PB15               25      212
##  5 PB17               10      207
##  6 PB17               10      208
##  7 PB17               11      118
##  8 PB17               11      209
##  9 PB17               28      226
## 10 PB17               28      227
## # ... with 678 more rows
```

```
write_csv(ps_id_mismatch, "f28_ps_id_mismatch.csv")
```

Form 45

First, I create a combo of `constituency_id` and `ps_id` to create a unique id for the polling station constituency combo.

```
f45_clean <- f45 %>%
  mutate(constituency_ps_id = paste0(constituency_id, "_", ps_id))
```

Extra, missing, polling station

Then I show that there is one of these combos that isn't present in the form 28 data.

RCONS FIX: The below polling station appears to be missing in the form 28 data. When we look at the name, I think it should just be changed to 903 from 904 but please check and correct.

```
# There's an extra constituency_ps_id in form 45s
setdiff(f45_clean$constituency_ps_id, f28$constituency_ps_id)
```

```
## [1] "NA21_904"
```

```
f45_clean %>%
  filter(constituency_ps_id == "NA21_904") %>%
  as.data.frame()
```

```
##   province     assembly_type constituency_id ps_id
## 1      KPK National Assembly            NA21   904
##                                                    ps_name total_votes
## 1 GOVT. ELEMENTARY PRIMARY SCHOOL, SHAMILAT (P) COMBINED         145
##   total_male_turnout total_female_turnout total_turnout comments
## 1                669                  301           970
##   constituency_ps_id
## 1           NA21_904
```

Mismatched totals

RCONS CHECK: There are many instances with mismatched total turnout numbers, but most of these are commented by your staff as numbers that appeared incorrect in the original data. Only the two below entries have no entries for comment. Please check the source file and report back.

```
f45 %>%
  mutate_if(is.numeric, list(~ifelse(. < 0, NA, .))) %>%
  mutate(total_turnout_sum_check = total_male_turnout + total_female_turnout) %>%
  filter(total_turnout_sum_check != total_turnout, comments == "") %>%
  as.data.frame()
```

```
##   province     assembly_type constituency_id ps_id
## 1   PUNJAB National Assembly            NA79   187
## 2   PUNJAB Provincial Punjab           PP162   115
##                                                               ps_name
## 1            GOVERNMENT BOYS PRIMARY SCHOOL PANDORI KALAN (COMBINED)
## 2 GOVT. GIRLS HIGH SCHOOL NISHAT COLONY LAHORE CANTT. (P) (COMBINED)
##   total_votes total_male_turnout total_female_turnout total_turnout
## 1        1265                675                  570          1265
## 2         974                559                  353           952
##   comments total_turnout_sum_check
## 1                             1245
## 2                              912
```

Form 48

Inconsistent data within polling stations

Again, as with the Form 28s, there are some inconsistent data within polling stations.

RCONS FIX: The following polling stations have different candidate names per candidate ID in the form 48 data. These are simple typos, but please correct them.

```
f48 %>%
  group_by(constituency_id) %>%
  select(constituency_id, contains("can_name")) %>%
  gather(variable, name, -constituency_id) %>%
  group_by(constituency_id, variable, name) %>%
  summarize() %>%
  group_by(constituency_id, variable) %>%
  filter(n() > 1)
```

```
## # A tibble: 6 x 3
## # Groups:   constituency_id, variable [3]
##   constituency_id variable   name
##   <chr>           <chr>      <chr>
## 1 NA17            can_name_5 RIZWAN
## 2 NA17            can_name_5 RIZWAN SAEED MUGHAL
## 3 NA40            can_name_4 SYED AKHUN SADA CHATTAN
## 4 NA40            can_name_4 SYED AKHUN ZADA CHATTAN
## 5 PP188           can_name_6 HAFIZ SHABAN AHMAD
## 6 PP188           can_name_6 HAFIZ SHABAN AHMED
```

RCONS FIX: there also appears to be a problem with the variables for the total number of PS for NA 83.

```
f48 %>%
  group_by(constituency_id) %>%
  select(constituency_id, starts_with("total_number"), starts_with("registered_voters"), constituency_na
  gather(variable, name, -constituency_id) %>%
  group_by(constituency_id, variable, name) %>%
  summarize() %>%
  group_by(constituency_id, variable) %>%
  filter(n() > 1)
```

```
## # A tibble: 4 x 3
## # Groups:   constituency_id, variable [2]
##   constituency_id variable                 name 
##   <chr>           <chr>                    <chr>
## 1 NA83            total_number_ps_combined -77  
## 2 NA83            total_number_ps_combined -88  
## 3 NA83            total_number_ps_total    75   
## 4 NA83            total_number_ps_total    86
```

Conclusion

Once the above are fixed, it should be easy for me to merge the data confidently and to check mismatches in vote totals and names across the different datasets. Please let us know when you can check the above! Thank you.