Fake News - Data Cleaning

Natália Tosi

13/3/2021

Importing Data and Labels

```
data_raw_port <- read_csv(
  "BANCO_NACIONAL_FAKENEWS_2021-08-03_CLEAN - label.csv")

variable_names <- read_csv(
  "BANCO_NACIONAL_FAKENEWS_2021-08-03_CLEAN - variable_names.csv")

answers_labels <- read_csv(
  "BANCO_NACIONAL_FAKENEWS_2021-08-03_CLEAN - answers_translated.csv")
```

Translating Document

```
data_clean <- as_tibble(data_raw_port)
colnames(data_clean) <- as_vector(variable_names[,2])

data_eng <- data_clean
data_eng[-c(1,2,3,7)] <- lapply(data_clean[-c(1,2,3,7)],
      function(x) answers_labels$answers_eng[match(x,
                                        answers_labels$answers_port)])
```

Grouping and Cleaning

```r
data_eng <- data_eng %>%
  mutate(
    evaluation = case_when(
    P1 %in% c("Excellent", "Good") ~ "Excellent/Good",
    P1 %in% c("Bad", "Terrible") ~ "Bad/Terrible",
    TRUE ~ P1),
    approval = case_when(
      P2 %in% c("Strongly approves", "Approves") ~ "Approves",
      P2 %in% c("Strongly disapproves", "Disapproves") ~ "Disapproves",
      TRUE ~ P2))
```

```r
data_fake_news_dem <- data_eng %>%
  filter(P19 != "Unsure") %>%
  mutate(shared_fake_news_19 = if_else(P19 == "Yes", 1, 0),
         race = if_else(race %in% c("Black", "Pardo (brown)"), "Black/Pardo",
                        race)) %>%
  rename(c(gov_trust = P3, pol_orientation = P4, interest_politics = P5,
           interest_news = P6, frequency_news = P7,source_printed_newspaper = P8_1,
           source_online_newspaper = P8_2,source_printed_magazines = P8_3,
           source_online_magazine = P8_4, source_radio = P8_5,
           source_television = P8_6, source_alternative = P8_7, source_wpp = P8_8,
           source_family = P8_9, source_social_media = P8_10, source_podcasts = P8_11,
           source_none = P8_99, reason_source = P9, trust_newspaper = P13_1,
           trust_magazine = P13_2, trust_radio = P13_3, trust_television = P13_4,
           trust_websites = P13_5, trust_blogs = P13_6, trust_social_media = P13_7,
           format_preference = P14, same_ideology_news = P15,
           oposite_ideology_news = P16, trust_traditional_press = P18,
           frequency_fake_news = P20, reaction_fake_news = P22,
           resp_population = P23_1, resp_gov = P23_2, resp_politicians = P23_3,
           resp_press = P23_4, resp_social_media = P23_5, severity_fake_news = P24,
           impact_newspaper = P25_1, impact_magazines = P25_2, impact_radio = P25_3,
           impact_television = P25_4, impact_cinema = P25_5, impact_websites = P25_6,
           impact_blogs = P25_7, impact_social_media = P25_8,
           impact2_facebook = P26_1, impact2_youtube = P26_2,
           impact2_instagram = P26_3, impact2_twitter = P26_4,
           impact2_tiktok = P26_5, impact2_wpp = P26_6,
           fake_news_source = P27, fact_checking = P28,
           trust_agencies = P29, share_news = P30,
           pand1_fakenews_facebook = P32_1, pand1_fakenews_youtube = P32_2,
           pand1_fakenews_instagram = P32_3, pand1_fakenews_twitter = P32_4,
           pand1_fakenews_tiktok = P32_5, pand1_fakenews_wpp = P32_6,
           pand2_worse_perception_media = P33, pand3_trust_vaccine = P34_1,
           pand3_seek_science = P34_2, pand3_preventive_treat = P34_3,
           pand3_masks = P34_4, pand4_source_info = P35,
           pand5_increased_interest_science = P36, vote1_trust_ballot = P37,
           vote2_eletronic_best_option = P38, vote3_worried_hacker = P39_1,
           vote3_worried_politics = P39_2, vote3_worried_transparency = P39_3,
           vote3_worried_tech = P39_4, vote3_worried_tse = P39_5)) %>%
select(-c(P10_1, P10_2,P10_3,P10_4,P10_5,P10_6,P10_7,P11_1,P11_2,P11_3,P11_4,
          P11_5,P11_6,P11_7,P12_1,P12_2,P12_3,P12_4,P12_5,P12_6,P12_7,P12A,
          P17_1,P17_2,P17_3,P17_4,P17_5,P17_6,P17_7,P17_8,P31_1,P31_2,P31_3,
          P31_4,P31_5,P31_6,P31_7,P31_8,P31_9,P31_10,P31_11,P31_12,P31_13,
```

```
          P40_1, P40_2))


data_fake_news_dem <- data_fake_news_dem %>%
  mutate(sex = factor(sex, levels = c("Men", "Women")),
         region = factor(region, levels = c("North", "Northeast", "Center-West",
                                    "Southeast", "South")),
         type = factor(type, levels = c("Capital", "Metropolitan region",
                                        "Countryside")),
         evaluation = factor(evaluation, levels = c("Excellent/Good", "Regular",
                                        "Bad/Terrible", "Unsure")),
         approval = factor(approval, levels = c("Approves",
                                     "Neither approves nor disapproves",
                                      "Disapproves", "Unsure")),
         pol_orientation = factor(pol_orientation, levels = c("Right/Center-Right",
                                      "Center", "Left/Center-Left",
                            "I no longer have a defined political orientation",
                            "I never had a political orientation", "Unsure")),
         race = factor(race, levels = c("White", "Black/Pardo","Indigenous",
                                        "Yellow", "Other")),
         education = factor(education, levels = c("No education", "Elementary School",
                                       "High School", "Higher Education")),
         income = factor(income, levels = c("Up to 1 MW", "1 to 3 MWs", "3 to 6 MWs",
                                     "More than 6 MWs", "Did not answer")),
         class = factor(class, levels = c("A/B", "C", "D/E", "DN/DA")),
         religion = factor(religion, levels = c("Catholic", "Evangelicals",
                                      "Other religion", "No religion")),
         frequency_fake_news = factor(frequency_fake_news, levels = c("Often",
                                "Sometimes", "Hardly ever", "Never", "Unsure")))

data_fake_news_dem <- data_fake_news_dem %>%
  mutate(race_adj = fct_collapse(race,
                                 White = c("White"),
                                 Black = c("Black/Pardo"),
                                 Other = c("Indigenous", "Yellow", "Other")))
```

Dummies

Demographics in the `data_fake_news_dem` variable

```
data_fake_news_dem <- data_fake_news_dem %>%
  mutate(sex_men = if_else(sex == "Men", 1, 0)) %>%
  dummy_cols(select_columns = c("region", "religion", "age_60","evaluation",
                                "P21","reaction_fake_news")) %>%
  mutate(capital_metrop = if_else(type %in% c("Capital",
                                              "Metropolitan region"), 1, 0),
         approves_gov = if_else(approval == "Approves", 1, 0),
         pol_orientation_right = if_else(pol_orientation == "Right/Center-Right",
```

```
                                        1, 0),
        pol_orientation_center = if_else(pol_orientation == "Center", 1, 0),
        pol_orientation_left = if_else(pol_orientation == "Left/Center-Left",
                                       1, 0),
        pol_orientation_none = if_else(pol_orientation %in% c(
          "I no longer have a defined   political orientation",
          "I never had a political orientation", "Unsure"), 1, 0),
        race_is_white = if_else(race == "White", 1, 0),
        education_high = if_else(education %in% c("High School", "Higher Education"),
                                 1, 0),
        income_low = if_else(income %in% c("Up to 1 MW", "1 to 3 MWs",
                                           "Did not answer"),
                             1, 0),
        class_ab = if_else(class == "A/B", 1, 0),
        class_c = if_else(class == "C", 1, 0),
        class_de = if_else(class %in% c("D/E", "DN/DA"), 1, 0),
        has_religion = if_else(religion != "No religion", 1, 0))
```

Codifying answers as binary to track changes in new variable `data_code`

```
data_code <- data_fake_news_dem %>%
  mutate(gov_trust = if_else(gov_trust %in% c("Great deal", "Fair amount"), 1, 0),
         interest_politics = if_else(interest_politics %in% c("Extremely interested",
                             "Quite interested", "Mildly interested"), 1, 0),
         interest_news = if_else(interest_news %in% c("Extremely interested",
                             "Quite interested", "Mildly interested"), 1, 0),
         frequency_news = if_else(frequency_news %in% c("More than once a day",
                             "Once a day", "a few times a week"), 1, 0),
         same_ideology_news = if_else(same_ideology_news ==
                           "News from sources who share your point of view", 1, 0),
         trust_traditional_press = if_else(trust_traditional_press %in% c(
                           "Yes, a fair amount", "Yes, a great deal"), 1, 0),
         frequency_fake_news = if_else(frequency_fake_news %in% c("Often",
                             "Sometimes"), 1, 0),
         severity_fake_news = if_else(severity_fake_news %in% c("Yes, a great deal",
                             "Yes, a fair amount"), 1, 0),
         fact_checking = if_else(fact_checking %in% c("Yes, always",
                             "Yes, occasionally"), 1, 0),
         trust_agencies = if_else(trust_agencies == "Yes", 1, 0),
         share_news = if_else(share_news %in% c("Yes, always",
                             "Yes, occasionally"), 1, 0),
         pand5_increased_interest_science = if_else(
           pand5_increased_interest_science %in% c("Increased a great deal",
                             "Increased a fair amount"), 1, 0),
         pand2_worse_perception_media = if_else(pand2_worse_perception_media ==
                             "Yes, for worse", 1, 0),
         vote1_trust_ballot = if_else(vote1_trust_ballot == "Great deal", 1, 0),
         vote2_eletronic_best_option = if_else(vote2_eletronic_best_option ==
                             "Electronic ballot", 1, 0))
```

```r
trust_variables <- c("trust_newspaper", "trust_magazine", "trust_radio",
                     "trust_television","trust_websites", "trust_blogs",
                     "trust_social_media")

data_code[trust_variables] <-
  lapply(data_code[trust_variables], function(x) {
    ifelse(x == "Great deal", 1, 0)})

resp_variables <- grepl("^resp_", names(data_code))

data_code[resp_variables] <-
  lapply(data_code[resp_variables], function(x) {
    ifelse(x %in% c("A great deal of responsibility",
                    "A fair amount of responsibility"), 1, 0)})

impact_variables <- grepl("^impact", names(data_code))

data_code[impact_variables] <-
  lapply(data_code[impact_variables], function(x) {
    ifelse(x %in% c("Major", "Moderate"), 1, 0)})

source_variables <- grepl("^source_", names(data_code))

data_code[source_variables] <-
  lapply(data_code[source_variables], function(x) {
    ifelse(!is.na(x), 1, 0)})

pand1_variables <- grepl("^pand1_", names(data_code))

data_code[pand1_variables] <-
  lapply(data_code[pand1_variables], function(x) {
    ifelse(x == "Yes", 1, 0)})

pand3_variables <- grepl("^pand3_", names(data_code))

data_code[pand3_variables] <-
  lapply(data_code[pand3_variables], function(x) {
    ifelse(x == "Agree", 1, 0)})

vote3_variables <- grepl("^vote3_", names(data_code))

data_code[vote3_variables] <-
  lapply(data_code[vote3_variables], function(x) {
    ifelse(x %in% c("Very worried", "Slightly worried"), 1, 0)})
```

Dummies Reference

- sex_men: 1 Men, 0 Women
- region: 5 levels

- capital_metrop: 1 Capital and Metropolitan region, 0 Countryside
- approves_gov: 1 Approves, 0 Neither approves nor disapproves, Disapproves, Unsure
- pol_orientation: Right/Center-Right, Center, Left/Center-Left, No orientation (I no longer have a defined political orientation, I never had a political orientation, Unsure)
- race_is_white: 1 White, 0 Black, Pardo (brown), Indigenous, Yellow, Other
- education_high: 1 High School and Higher Education, 0 No education and Elementary School
- class: 3 levels: A/B, C, D/E and DN/DA
- religion: 4 levels: Catholic, Evangelicals, Other religion, No religion
- has_religion: 1 Catholic, Evangelicals, Other religion; 0 No religion
- gov_trust (1 if "Great deal", "Fair amount"; 0 otherwise)
- interest_politics (1 if "Extremely interested", "Quite interested", "Mildly interested"; 0 otherwise)
- interest_news (1 if "Extremely interested", "Quite interested", "Mildly interested"; 0 otherwise)
- frequency_news (1 if "More than once a day", "Once a day", "a few times a week"; 0 otherwise)
- If starts with trust (except trust_traditional_press and trust_agencies) (1 if "Great deal"; 0 otherwise)
- same_ideology_news (1 if "News from sources who share your point of view"; 0 otherwise)
- trust_traditional_press (1 if "Yes, a fair amount", "Yes, a great deal"; 0 otherwise)
- frequency_fake_news (1 if "Often", "Sometimes"; 0 otherwise)
- resp_... (1 if "A great deal of responsibility", "A fair amount of responsibility"; 0 otherwise)
- severity_fake_news (1 if "Yes, a great deal", "Yes, a fair amount"; 0 otherwise)
- impact_... (1 if "Major", "Moderate"; 0 otherwise)
- fact_checking (1 if "Yes, always", "Yes, occasionally"; 0 otherwise)
- trust_agencies (1 if "Yes"; 0 otherwise)
- share_news (1 if "Yes, always", "Yes, occasionally"; 0 otherwise)
- pand5_... (1 if "Increased a great deal", "Increased a fair amount"; 0 otherwise)
- if starts with "source" (1 if !is.na(); 0 otherwise)
- pand1_ (1 if "Yes"; 0 otherwise)
- pand2_ did get worse? (1 if "Yes, for worse"; 0 otherwise)
- pand3_ (1 if "Agree"; 0 otherwise)
- vote1_ (1 if "Great deal"; 0 otherwise)
- vote2_ (1 if "Electronic ballot"; 0 otherwise)
- all with vote3_ (1 if "Slightly worried","Very worried"; 0 otherwise)

Save as new CSV file

```
write.csv(data_code,'fake_news_data_code.csv')
```