Data Cleaning Report

Veena Muralidharan

26/04/2020

Contents

Data Understanding

Data Cleaning

The first step to building a data science project includes tidying up of data. For this purpose, one must first understand the data in order to hone in onto the errors present in the data.

- **Getting the feel of the Data**

Understanding the dataframe, its dimensions, the variables and building a summary of the data

Step 1: Basic Information of the data

```
who_report_tbl <- read_rds("../../project/00_Data/02_data_wrangled/who_report.rds")
dim(who_report_tbl)
```

```
## [1] 7758   25
```

```
names(who_report_tbl)
```

```
##  [1] "reported_date"
##  [2] "reporting_country_territory"
##  [3] "confirmed_cases"
##  [4] "new_confirmed_cases"
##  [5] "total_deaths"
##  [6] "new_total_deaths"
##  [7] "transmission_classification"
##  [8] "total_cases_with_travel_history_to_china"
##  [9] "new_cases_with_travel_history_to_china"
## [10] "total_cases_with_possible_or_confirmed_transmission_outside_china"
## [11] "new_cases_with_possible_or_confirmed_transmission_outside_china"
## [12] "total_cases_with_site_of_transmission_under_investigation"
## [13] "new_cases_with_site_of_transmission_under_investigation"
## [14] "place_of_exposure_in_china_cases"
## [15] "place_of_exposure_in_china_cases_new"
## [16] "place_of_exposure_outside_reporting_country_and_china_cases"
## [17] "place_of_exposure_outside_reporting_country_and_china_cases_new"
## [18] "place_of_exposure_in_reporting_country_cases"
## [19] "place_of_exposure_in_reporting_country_cases_new"
## [20] "suspected_cases"
## [21] "daily_suspected_cases"
## [22] "daily_lab_confirmed_cases"
## [23] "cumulative_lab_confirmed_cases"
## [24] "days_since_last_reported"
## [25] "report_url"
```

Step 2: Basic Summary of the Data

```
summary(who_report_tbl)
```

```
##  reported_date        reporting_country_territory confirmed_cases
##  Min.   :2020-01-21   Length:7758                 Min.   :0.000e+00
##  1st Qu.:2020-02-29   Class :character            1st Qu.:5.000e+00
##  Median :2020-03-15   Mode  :character            Median :4.400e+01
##  Mean   :2020-03-11                               Mean   :5.478e+05
##  3rd Qu.:2020-03-27                               3rd Qu.:2.850e+02
##  Max.   :2020-04-05                               Max.   :1.106e+09
##                                                   NA's   :8
##  new_confirmed_cases  total_deaths      new_total_deaths
##  Min.   :    0.0     Min.   :    0.00   Min.   :   0.000
##  1st Qu.:    0.0     1st Qu.:    0.00   1st Qu.:   0.000
##  Median :    1.0     Median :    0.00   Median :   0.000
##  Mean   :  172.3     Mean   :   76.43   Mean   :   7.469
##  3rd Qu.:   19.0     3rd Qu.:    3.00   3rd Qu.:   0.000
##  Max.   :56249.0     Max.   :14681.00   Max.   :2003.000
##  NA's   :715         NA's   :635        NA's   :764
##  transmission_classification total_cases_with_travel_history_to_china
##  Length:7758                 Mode :logical
##  Class :character            FALSE:33
##  Mode  :character            TRUE :106
##                              NA's :7619
##
##
##
##  new_cases_with_travel_history_to_china
##  Mode :logical
##  FALSE:302
##  TRUE :30
##  NA's :7426
##
##
##
##  total_cases_with_possible_or_confirmed_transmission_outside_china
##  Mode :logical
##  FALSE:206
##  TRUE :33
##  NA's :7519
##
##
##
##  new_cases_with_possible_or_confirmed_transmission_outside_china
##  Mode :logical
##  FALSE:287
##  TRUE :24
##  NA's :7447
##
##
##
##  total_cases_with_site_of_transmission_under_investigation
##  Mode :logical
##  FALSE:94
##  TRUE :13
##  NA's :7651
##
##
##
##  new_cases_with_site_of_transmission_under_investigation
##  Mode :logical
##  FALSE:44
##  TRUE :5
##  NA's :7709
##
##
##
##  place_of_exposure_in_china_cases place_of_exposure_in_china_cases_new
##  Mode :logical                    Mode :logical
##  FALSE:79                         FALSE:281
##  TRUE :55                         TRUE :5
##  NA's :7624                       NA's :7472
##
##
##
##  place_of_exposure_outside_reporting_country_and_china_cases
##  Mode :logical
##  FALSE:170
##  TRUE :37
##  NA's :7551
##
##
##
##  place_of_exposure_outside_reporting_country_and_china_cases_new
##  Mode :logical
##  FALSE:235
##  TRUE :22
##  NA's :7501
##
##
##
##  place_of_exposure_in_reporting_country_cases
##  Mode :logical
##  FALSE:152
##  TRUE :23
##  NA's :7583
##
##
##
##  place_of_exposure_in_reporting_country_cases_new suspected_cases
##  Mode :logical                                    Mode :logical
##  FALSE:257                                        FALSE:598
##  TRUE :6                                          TRUE :64
##  NA's :7495                                       NA's :7096
##
##
##
##  daily_suspected_cases daily_lab_confirmed_cases cumulative_lab_confirmed_cases
##  Mode :logical         Mode :logical             Mode:logical
##  FALSE:16              FALSE:19                  TRUE:3
##  TRUE :2               TRUE :10                  NA's:7755
##  NA's :7740            NA's :7729
##
##
##
##  days_since_last_reported  report_url
##  Min.   : 0.000           Length:7758
##  1st Qu.: 0.000           Class :character
##  Median : 0.000           Mode  :character
##  Mean   : 1.411
##  3rd Qu.: 1.000
##  Max.   :59.000
##  NA's   :2413
```