

Data Cleaning Part 1

Data Wrangling in R

Data Cleaning

In general, data cleaning is a process of investigating your data for inaccuracies, or recoding it in a way that makes it more manageable.

MOST IMPORTANT RULE - LOOK AT YOUR DATA!

Read in the UFO dataset

Read in data or download from: http://sisbid.github.io/Data-Wrangling/data/ufo/ufo_data_complete.csv.gz

```
ufo <- read_delim("https://sisbid.github.io/Data-Wrangling/
```

Warning: One or more parsing issues, call `problems()` on your data.
e.g.:

```
dat <- vroom(...)  
problems(dat)
```

Rows: 88875 Columns: 11

-- Column specification -----

Delimiter: ","

chr (10): datetime, city, state, country, shape, duration (seconds)

dbl (1): duration (seconds)

i Use `spec()` to retrieve the full column specification for this data.

Checking

```
nrow(ufo)-nrow(ufo_clean)
```

```
[1] 196
```

Clean names with the `clean_names()` function from the `janitor` package

```
colnames(ufo_clean)
```

```
[1] "datetime"          "city"              "state"
[4] "country"           "shape"             "duration"
[7] "duration (hours/min)" "comments"          "date posted"
[10] "latitude"          "longitude"
```

```
ufo_clean <- clean_names(ufo_clean)
```

```
colnames(ufo_clean)
```

```
[1] "datetime"          "city"              "state"
[4] "country"           "shape"             "duration_seconds"
[7] "duration_hours_min" "comments"          "date_posted"
[10] "latitude"          "longitude"
```

Recoding Variables

Exact Swaps - recode function

```
ufo_clean %>% mutate(country = recode(country, gb = "Great
```

Rows: 88,679

Columns: 11

```
$ datetime      <chr> "10/10/1949 20:30", "10/10/1949
```

```
$ city      <chr> "san marcos", "lackland afb", "dallas"
```

```
$ state      <chr> "tx", "tx", NA, "tx", "hi", "tn"
```

```
$ country      <chr> "us", NA, "Great Britain", "us"
```

```
$ shape      <chr> "cylinder", "light", "circle", "
```

```
$ duration_seconds    <chr> "2700", "7200", "20", "20", "900"
```

```
$ duration_hours_min <chr> "45 minutes", "1-2 hrs", "20 se
```

```
$ comments      <chr> "This event took place in early
```

```
$ date_posted      <chr> "4/27/2004", "12/16/2005", "1/2/2006"
```

```
$ latitude      <chr> "29.8830556", "29.38421", "53.2"
```

```
$ longitude      <chr> "-97.9411111", "-98.581082", "-99.121082", "-99.661082", "-100.201082", "-100.741082", "-101.281082", "-101.821082", "-102.361082", "-102.901082", "-103.441082", "-103.981082", "-104.521082", "-105.061082", "-105.601082", "-106.141082", "-106.681082", "-107.221082", "-107.761082", "-108.301082", "-108.841082", "-109.381082", "-109.921082", "-110.461082", "-111.001082", "-111.541082", "-112.081082", "-112.621082", "-113.161082", "-113.701082", "-114.241082", "-114.781082", "-115.321082", "-115.861082", "-116.401082", "-116.941082", "-117.481082", "-118.021082", "-118.561082", "-119.101082", "-119.641082", "-120.181082", "-120.721082", "-121.261082", "-121.801082", "-122.341082", "-122.881082", "-123.421082", "-123.961082", "-124.501082", "-125.041082", "-125.581082", "-126.121082", "-126.661082", "-127.201082", "-127.741082", "-128.281082", "-128.821082", "-129.361082", "-129.901082", "-130.441082", "-130.981082", "-131.521082", "-132.061082", "-132.601082", "-133.141082", "-133.681082", "-134.221082", "-134.761082", "-135.301082", "-135.841082", "-136.381082", "-136.921082", "-137.461082", "-137.981082", "-138.521082", "-139.041082", "-139.581082", "-140.121082", "-140.661082", "-141.201082", "-141.741082", "-142.281082", "-142.821082", "-143.361082", "-143.901082", "-144.441082", "-144.981082", "-145.521082", "-146.061082", "-146.601082", "-147.141082", "-147.681082", "-148.221082", "-148.761082", "-149.301082", "-149.841082", "-150.381082", "-150.921082", "-151.461082", "-151.981082", "-152.521082", "-153.041082", "-153.581082", "-154.121082", "-154.661082", "-155.201082", "-155.741082", "-156.281082", "-156.821082", "-157.361082", "-157.901082", "-158.441082", "-158.981082", "-159.521082", "-160.041082", "-160.581082", "-161.121082", "-161.661082", "-162.201082", "-162.741082", "-163.281082", "-163.821082", "-164.361082", "-164.901082", "-165.441082", "-165.981082", "-166.521082", "-167.041082", "-167.581082", "-168.121082", "-168.661082", "-169.201082", "-169.741082", "-170.281082", "-170.821082", "-171.361082", "-171.901082", "-172.441082", "-172.981082", "-173.521082", "-174.041082", "-174.581082", "-175.121082", "-175.661082", "-176.201082", "-176.741082", "-177.281082", "-177.821082", "-178.361082", "-178.901082", "-179.441082", "-179.981082", "-180.521082", "-181.041082", "-181.581082", "-182.121082", "-182.661082", "-183.201082", "-183.741082", "-184.281082", "-184.821082", "-185.361082", "-185.901082", "-186.441082", "-186.981082", "-187.521082", "-188.041082", "-188.581082", "-189.121082", "-189.661082", "-190.201082", "-190.741082", "-191.281082", "-191.821082", "-192.361082", "-192.901082", "-193.441082", "-193.981082", "-194.521082", "-195.041082", "-195.581082", "-196.121082", "-196.661082", "-197.201082", "-197.741082", "-198.281082", "-198.821082", "-199.361082", "-199.901082", "-200.441082", "-200.981082", "-201.521082", "-202.041082", "-202.581082", "-203.121082", "-203.661082", "-204.201082", "-204.741082", "-205.281082", "-205.821082", "-206.361082", "-206.901082", "-207.441082", "-207.981082", "-208.521082", "-209.041082", "-209.581082", "-210.121082", "-210.661082", "-211.201082", "-211.741082", "-212.281082", "-212.821082", "-213.361082", "-213.901082", "-214.441082", "-214.981082", "-215.521082", "-216.041082", "-216.581082", "-217.121082", "-217.661082", "-218.201082", "-218.741082", "-219.281082", "-219.821082", "-220.361082", "-220.901082", "-221.441082", "-221.981082", "-222.521082", "-223.041082", "-223.581082", "-224.121082", "-224.661082", "-225.201082", "-225.741082", "-226.281082", "-226.821082", "-227.361082", "-227.901082", "-228.441082", "-228.981082", "-229.521082", "-230.041082", "-230.581082", "-231.121082", "-231.661082", "-232.201082", "-232.741082", "-233.281082", "-233.821082", "-234.361082", "-234.901082", "-235.441082", "-235.981082", "-236.521082", "-237.041082", "-237.581082", "-238.121082", "-238.661082", "-239.201082", "-239.741082", "-240.281082", "-240.821082", "-241.361082", "-241.901082", "-242.441082", "-242.981082", "-243.521082", "-244.041082", "-244.581082", "-245.121082", "-245.661082", "-246.201082", "-246.741082", "-247.281082", "-247.821082", "-248.361082", "-248.901082", "-249.441082", "-249.981082", "-250.521082", "-251.041082", "-251.581082", "-252.121082", "-252.661082", "-253.201082", "-253.741082", "-254.281082", "-254.821082", "-255.361082", "-255.901082", "-256.441082", "-256.981082", "-257.521082", "-258.041082", "-258.581082", "-259.121082", "-259.661082", "-260.201082", "-260.741082", "-261.281082", "-261.821082", "-262.361082", "-262.901082", "-263.441082", "-263.981082", "-264.521082", "-265.041082", "-265.581082", "-266.121082", "-266.661082", "-267.201082", "-267.741082", "-268.281082", "-268.821082", "-269.361082", "-269.901082", "-270.441082", "-270.981082", "-271.521082", "-272.041082", "-272.581082", "-273.121082", "-273.661082", "-274.201082", "-274.741082", "-275.281082", "-275.821082", "-276.361082", "-276.901082", "-277.441082", "-277.981082", "-278.521082", "-279.041082", "-279.581082", "-280.121082", "-280.661082", "-281.2010
```

Exact Swaps - recode function