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 y
e.g.:
  dat <- vroom(...)
  problems(dat)
Rows: 88875 Columns: 11
-- Column specification --------------------------------
Delimiter: ","
chr (10): datetime, city, state, country, shape, duration (
dbl  (1): duration (seconds)

i Use `spec()` to retrieve the full column specification fo
```

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"                   "duratio
 [7] "duration (hours/min)"    "comments"                "date po
[10] "latitude"                "longitude"
```

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

```
 [1] "datetime"              "city"                  "state"
 [4] "country"               "shape"                 "duration_se
 [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", "c
$ 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 sec
$ comments          <chr> "This event took place in early
$ date_posted       <chr> "4/27/2004", "12/16/2005", "1/21
$ latitude          <chr> "29.8830556", "29.38421", "53.2"
$ longitude         <chr> "-97.9411111", "-98.581082", "-2
```

Exact Swaps - recode function