Grab-bag and Exploratory Data Analysis

Questions

Negative case numbers in the coronavirus table

```
> coronavirus::coronavirus %>%
      arrange(cases) %>%
      select(date, country,
             cases) %>%
      head()

         date country  cases
1 2020-04-24   Spain -10034
2 2020-04-29  France  -2512
3 2020-05-12      US  -2446
4 2020-04-22  France  -2206
5 2020-05-07 Ecuador  -1583
6 2020-05-08 Ecuador  -1480
```

https://github.com/RamiKrispin/coronavirus

You always need to check for missing data

```
nycflights13::flights
# A tibble: 336,776 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier flight tailnum
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>    <int> <chr>
 1  2013     1     1      517            515         2      830            819        11 UA        1545 N14228
 2  2013     1     1      533            529         4      850            830        20 UA        1714 N24211
 3  2013     1     1      542            540         2      923            850        33 AA        1141 N619AA
 4  2013     1     1      544            545        -1     1004           1022       -18 B6         725 N804JB
 5  2013     1     1      554            600        -6      812            837       -25 DL         461 N668DN
 6  2013     1     1      554            558        -4      740            728        12 UA        1696 N39463
 7  2013     1     1      555            600        -5      913            854        19 B6         507 N516JB
 8  2013     1     1      557            600        -3      709            723       -14 EV        5708 N829AS
 9  2013     1     1      557            600        -3      838            846        -8 B6          79 N593JB
10  2013     1     1      558            600        -2      753            745         8 AA         301 N3ALAA
# … with 336,766 more rows, and 7 more variables: origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>
```

You always need to check for missing data

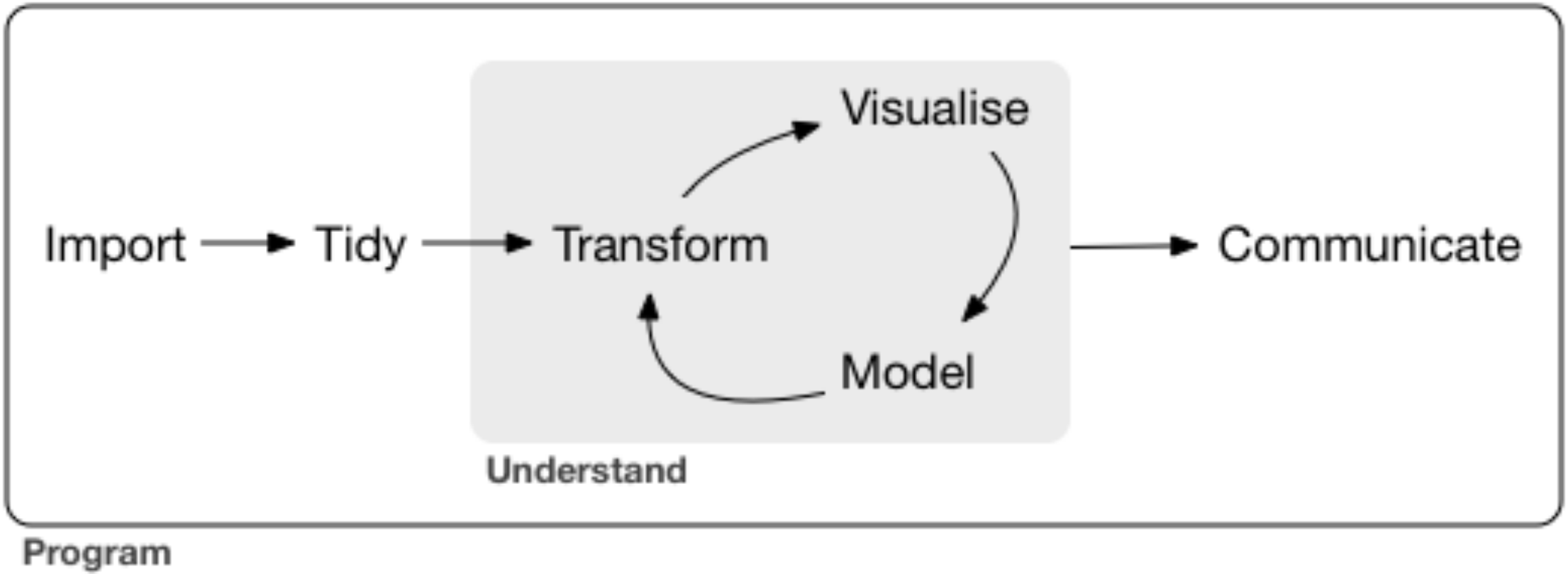
```
> nycflights13::flights %>% filter_all(any_vars(is.na(.) == T))
# A tibble: 9,430 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier flight tailnum
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>    <int> <chr>
 1  2013     1     1     1525           1530        -5     1934           1805        NA MQ        4525 N719MQ
 2  2013     1     1     1528           1459        29     2002           1647        NA EV        3806 N17108
 3  2013     1     1     1740           1745        -5     2158           2020        NA MQ        4413 N739MQ
 4  2013     1     1     1807           1738        29     2251           2103        NA UA        1228 N31412
 5  2013     1     1     1939           1840        59       29           2151        NA 9E        3325 N905XJ
 6  2013     1     1     1952           1930        22     2358           2207        NA EV        4333 N11194
 7  2013     1     1     2016           1930        46       NA           2220        NA EV        4204 N14168
 8  2013     1     1       NA           1630        NA       NA           1815        NA EV        4308 N18120
 9  2013     1     1       NA           1935        NA       NA           2240        NA AA         791 N3EHAA
10  2013     1     1       NA           1500        NA       NA           1825        NA AA        1925 N3EVAA
# … with 9,420 more rows, and 7 more variables: origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>
```

Heatmap annotations

DEMO

Exploratory Data Analysis

Exploratory Data Analysis



Assignment

1. Pick a dataset from the Tidy Tuesday datasets
2. Explore the data
3. Generate a report answering 10 questions you come up with about the data