Getting the Stations

Emily Lachtara and Shukry Zablah

2019-10-01

Setup

```
library(dplyr)
library(janitor)
library(vroom)
library(leaflet)
```

Ingestion

Vroom is a package developed by the tidyverse guys who does faster loading of files. You can also use readr.

```
June8 <- vroom("../data-raw/VB_Routes_Data_2019_06_08.csv.gz",
               col_types = cols(
                   `Route ID` = col_character(),
                   Bike = col_number(),
                   Date = col_datetime(format = ""),
                   Latitude = col_double(),
                   Longitude = col_double(),
                   `User ID` = col_character()
               ),
               skip = 2) %>%
    clean_names()
```

Take a look at the data.

```
June8 %>% head()
```

```
## # A tibble: 6 x 6
##   route_id          bike date                latitude longitude user_id
##   <chr>            <dbl> <dttm>                 <dbl>     <dbl> <chr>
## 1 route_06_2019@c~  1134 2019-06-08 04:00:00     42.2     -72.6 9ca844ff-0~
## 2 route_06_2019@c~  1134 2019-06-08 04:00:05     42.2     -72.6 9ca844ff-0~
## 3 route_06_2019@c~  1134 2019-06-08 04:00:10     42.2     -72.6 9ca844ff-0~
## 4 route_06_2019@c~  1134 2019-06-08 04:00:15     42.2     -72.6 9ca844ff-0~
## 5 route_06_2019@c~  1134 2019-06-08 04:00:20     42.2     -72.6 9ca844ff-0~
## 6 route_06_2019@c~  1134 2019-06-08 04:00:25     42.2     -72.6 9ca844ff-0~
```

Stations

We are not given information about the stations. Let's find that out by taking the start and end of every route. A recent day should have most of the stations we want to find.

```
Stations <- June8 %>%
    group_by(route_id) %>%
    filter(date == max(date) | date == min(date)) %>%
    ungroup() %>%
    mutate(lat_rounded = round(latitude, 3),
           lon_rounded = round(longitude, 3)) %>%
    group_by(lat_rounded, lon_rounded) %>%
    summarize() %>%
    ungroup() %>%
    mutate(name = paste("Station", row_number())) %>%
    select(name, latitude = lat_rounded, longitude = lon_rounded)
```

```
Stations
```

```
## # A tibble: 89 x 3
##    name       latitude longitude
##    <chr>         <dbl>     <dbl>
##  1 Station 1      42.1     -72.6
##  2 Station 2      42.1     -72.6
##  3 Station 3      42.1     -72.6
##  4 Station 4      42.1     -72.6
##  5 Station 5      42.1     -72.6
##  6 Station 6      42.1     -72.6
##  7 Station 7      42.1     -72.6
##  8 Station 8      42.1     -72.6
##  9 Station 9      42.1     -72.6
## 10 Station 10     42.1     -72.6
## # ... with 79 more rows
```

```
Stations %>%
    vroom_write("../data/stations.tsv")
```

Now we have a table of the stations. We have to keep in mind that since we used the method above to create the table then two things might have happened:

1) We might not have all the stations because they were not visited that day.
2) We might have extraneous "stations" due to the fact we used the start and end of routes.

A possible solution is to do this for more days and take locations used above x number of times.

```
leaflet(Stations) %>%
    addTiles() %>%
    addMarkers(~ longitude, ~ latitude)
```

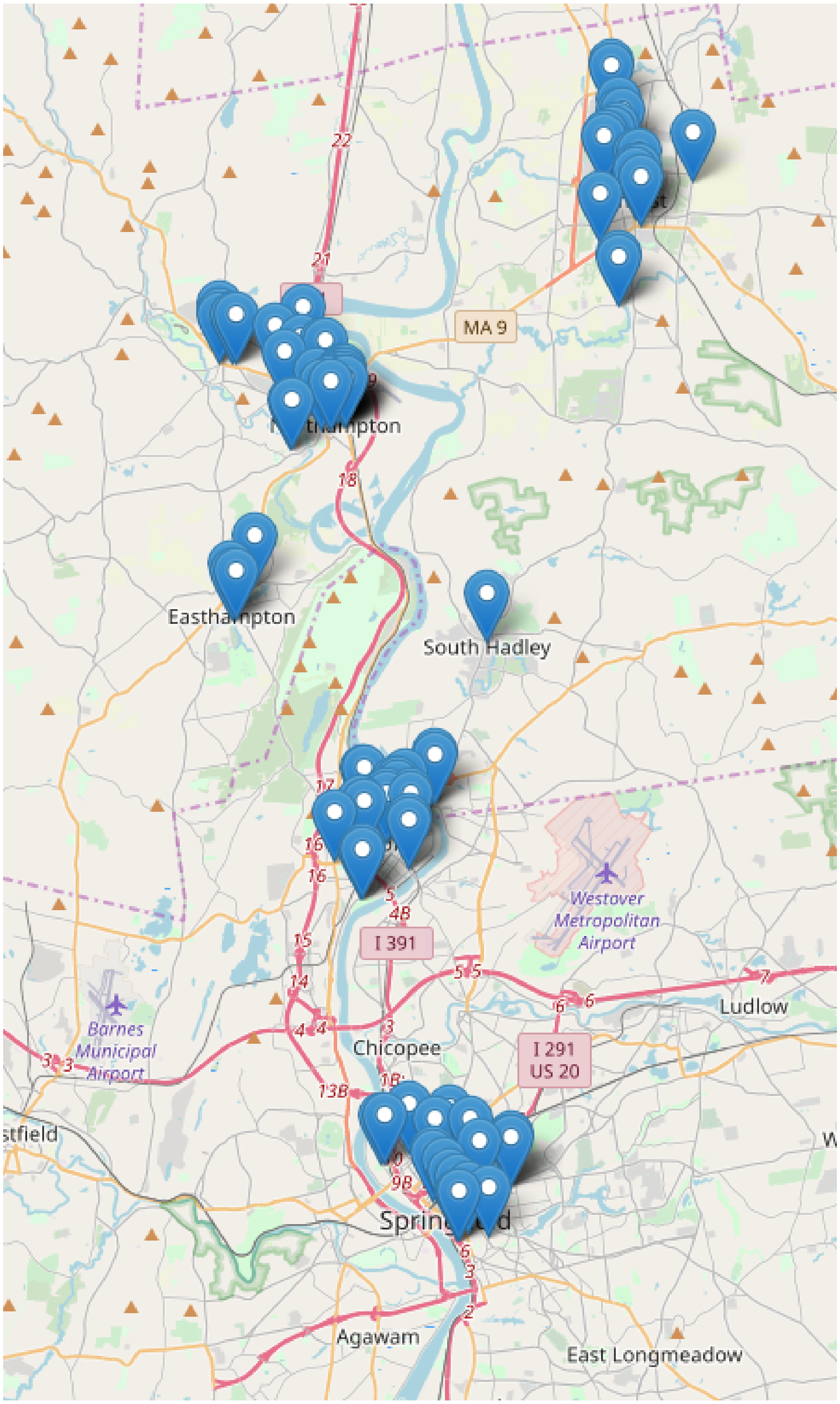

Figure 1: Map of Valley Bike stations used in June 8.