Cyclist Data Processing

Sheldon Crasta

2023-05-20

Before Processing

Beginning with the process we must first download the data from the respective **URL** https://divvy-tripdata.s3.amazonaws.com/index.html.

Once it is **downloaded**, you need to extract all the files into one folder.

Process

Following are the steps that were followed :

Step 01: Access the **tidyverse** library using the following :

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
## Warning: package 'lubridate' was built under R version 4.2.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.0     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.1     v tibble    3.1.8
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Step 02: Combining all the .csv file into one.

```
merged_data = list.files(path = "TripData", full.names = TRUE) %>% lapply(read_csv) %>% bind_rows
```

```
## Rows: 337230 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 531633 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 729595 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 822410 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 804352 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 756147 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 631226 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 359978 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 247540 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 103770 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 115609 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 284042 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Step 03: Viewing the top 6 from merged_data dataset

```
head(merged_data)
```

```
## # A tibble: 6 x 13
##   ride_id        ridea~1 started_at          ended_at            start~2 start~3
##   <chr>          <chr>   <dttm>              <dttm>              <chr>   <chr>
## 1 6C992BD37A98A~ classi~ 2021-04-12 18:25:36 2021-04-12 18:56:55 State ~ TA1307~
## 2 1E0145613A209~ docked~ 2021-04-27 17:27:11 2021-04-27 18:31:29 Dorche~ KA1503~
## 3 E498E15508A80~ docked~ 2021-04-03 12:42:45 2021-04-07 11:40:24 Loomis~ 20121
## 4 1887262AD101C~ classi~ 2021-04-17 09:17:42 2021-04-17 09:42:48 Honore~ TA1305~
## 5 C123548CAB2A3~ docked~ 2021-04-03 12:42:25 2021-04-03 14:13:42 Loomis~ 20121
## 6 097E76F3651B1~ classi~ 2021-04-25 18:43:18 2021-04-25 18:43:59 Clinto~ 15542
## # ... with 7 more variables: end_station_name <chr>, end_station_id <chr>,
## #   start_lat <dbl>, start_lng <dbl>, end_lat <dbl>, end_lng <dbl>,
## #   member_casual <chr>, and abbreviated variable names 1: rideable_type,
## #   2: start_station_name, 3: start_station_id
```

Step 04: Following activities performed on the dataset that were followed:

```
merged_data = merged_data %>% separate(started_at, into = c("started_at","started_time"), sep = " ")

merged_data = merged_data %>% separate(ended_at, into = c("ended_at","ended_time"), sep = " ")

merged_data$started_date_week = format(as.Date(merged_data$started_at), "%A")

merged_data$ended_date_week = format(as.Date(merged_data$ended_at), "%A")

merged_data = merged_data %>% relocate(started_date_week, .after = started_at)

merged_data = merged_data %>% relocate(ended_date_week, .after = ended_at)
```

Step 05: Finally saving the file:

```
write.csv(merged_data, "cleaned_cyclist_data.csv")
```

After Processing

Now we will work on the visualization using the *cleaned_cyclist_data.csv*. And this how the datasets now looks :

```
head(merged_data)
```

```
## # A tibble: 6 x 17
##   ride_id        ridea~1 start~2 start~3 start~4 ended~5 ended~6 ended~7 start~8
##   <chr>          <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
## 1 6C992BD37A98A~ classi~ 2021-0~ Monday  18:25:~ 2021-0~ Monday  18:56:~ State ~
## 2 1E0145613A209~ docked~ 2021-0~ Tuesday 17:27:~ 2021-0~ Tuesday 18:31:~ Dorche~
## 3 E498E15508A80~ docked~ 2021-0~ Saturd~ 12:42:~ 2021-0~ Wednes~ 11:40:~ Loomis~
## 4 1887262AD101C~ classi~ 2021-0~ Saturd~ 09:17:~ 2021-0~ Saturd~ 09:42:~ Honore~
## 5 C123548CAB2A3~ docked~ 2021-0~ Saturd~ 12:42:~ 2021-0~ Saturd~ 14:13:~ Loomis~
## 6 097E76F3651B1~ classi~ 2021-0~ Sunday  18:43:~ 2021-0~ Sunday  18:43:~ Clinto~
```

```
## # ... with 8 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, start_lat <dbl>, start_lng <dbl>, end_lat <dbl>,
## #   end_lng <dbl>, member_casual <chr>, and abbreviated variable names
## #   1: rideable_type, 2: started_at, 3: started_date_week, 4: started_time,
## #   5: ended_at, 6: ended_date_week, 7: ended_time, 8: start_station_name
```