Cyclistic Process markdown - 2022-06-08 - v00

Luis Alejandro

2022-06-08

Load needed libraries:

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✔ ggplot2 3.3.6     ✔ purrr   0.3.4
## ✔ tibble  3.1.7     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

Load the dataframe:

This is the previous DF (data frame) in which we combined all the 12 months worth of data.

```
Cyclisticdf <- read.csv("C:/Users/Luis Alejandro/Documents/Google Data Analytics Professional Certificate/CASE STUDY 1 - CYCLISTIC/12 months worth of data/Data/All_the_data.csv")
```

Checking for repeated ride_id values:

```
ride_id_occure <- data.frame(table(Cyclisticdf$ride_id))
ride_id_occure[ride_id_occure$Freq > 1]

## data frame with 0 columns and 5757551 rows
```

Observations: It does not have repeated values since the list where Freq > 1 is empty.

Cleaning the started_at and ended_at columns:

Removing the extra T's and Z's at the begging and end of each cell.

```
Cyclisticdf$started_at <- gsub("T", " ",
as.character(Cyclisticdf$started_at)) #removes the T from the started_at column
Cyclisticdf$started_at <- gsub("Z", "", as.character(Cyclisticdf$started_at))
#removes the Z from the started_at column
Cyclisticdf$ended_at <- gsub("T", " ", as.character(Cyclisticdf$ended_at))
#removes the T from the ended_at column
Cyclisticdf$ended_at <- gsub("Z", "", as.character(Cyclisticdf$ended_at))
#removes the Z from the ended_at column
```

Adding the ride_length and Week_day columns to the dataframe:

These columns are needed for subsequent calculations.

```
Cyclisticdf$ride_length_in_minutes <- difftime(Cyclisticdf$ended_at,
Cyclisticdf$started_at, units= "mins")
Cyclisticdf$week_day <- Cyclisticdf$started_at %>% wday(label= TRUE, abbr= FALSE)
```

Correcting the ended_at and started_at columns where they are switched:

Note: there are observations who have negative ride_length_in_minutes, to solve this, we need to switch the ended_at and started_at columns at the negatives.

```
Cyclisticdf$new_started_at <- ifelse(Cyclisticdf$ride_length_in_minutes < 0,
Cyclisticdf$ended_at, Cyclisticdf$started_at)
Cyclisticdf$new_ended_at <- ifelse(Cyclisticdf$ride_length_in_minutes < 0,
Cyclisticdf$started_at, Cyclisticdf$ended_at)
Cyclisticdf$new_ride_length_in_minutes <- difftime(Cyclisticdf$new_ended_at,
Cyclisticdf$new_started_at, units= "mins")
```

Rearrange columns:

This is simply to prepare the data frame to drop the old started_at and ended_at columns.

```
Cyclisticdf <- Cyclisticdf %>% relocate(new_started_at, new_ended_at,
.before= started_at)
Cyclisticdf <- Cyclisticdf %>% relocate(new_ride_length_in_minutes, .before=
ride_length_in_minutes)
```

Drop old started_at, ended_at and ride_length_in_minutes columns:

```
Cyclisticdf <- Cyclisticdf[,!(names(Cyclisticdf) %in% c("started_at",
"ended_at", "ride_length_in_minutes"))]
```

Rename new columns to "started_at", "ended_at" and "ride_length_in_minutes":

```
names(Cyclisticdf) <- c("ride_id", "rideable_type", "started_at", "ended_at",
"start_station_name", "start_station_id", "end_station_name",
"end_station_id", "start_lat", "start_lng", "end_lat", "end_lng",
"member_casual", "ride_length_in_minutes", "week_day1")
```

Create a week_day in english and drop the week_day1 column:

This is only bc my system is installed in spanish so the wday() function will deliver weekdays in spanish.

```
Cyclisticdf$week_day <- case_when(
  Cyclisticdf$week_day1 == "domingo" ~ "sunday",
  Cyclisticdf$week_day1 == "lunes" ~ "monday",
  Cyclisticdf$week_day1 == "martes" ~ "tuesday",
  Cyclisticdf$week_day1 == "miércoles" ~ "wednesday",
  Cyclisticdf$week_day1 == "jueves" ~ "thursday",
  Cyclisticdf$week_day1 == "viernes" ~ "friday",
  Cyclisticdf$week_day1 == "sábado" ~ "saturday"
)
```

Drop week_day1:

We wont need this column anymore

```
Cyclisticdf <- Cyclisticdf[,!(names(Cyclisticdf) %in% c("week_day1"))]
```

Verify unique id's for start station name:

```
distinct_start_station_id <-
  data.frame(Cyclisticdf %>%
               group_by(start_station_id) %>%
               summarise(unique_names= n_distinct(start_station_name)))
#creates a df with unique count names grouped by start_station_id
```

Filters distinct_start_station_id looking for id's with more than 1 distinct string

```
distinct_start_station_id %>%
  filter(distinct_start_station_id$unique_names > 1)# filters
distinct_start_station_id looking for id's with more than 1 distinct string
```

```
##    start_station_id unique_names
## 1             13099            2
## 2             13221            2
## 3             13300            2
## 4             20215            2
## 5               351            2
## 6               444            2
## 7            LF-005            2
## 8      TA1305000039            2
## 9      TA1306000029            3
## 10     TA1307000041            2
## 11     TA1309000039            2
## 12     TA1309000049            2
## 13           WL-008            2
```

Cleaning the start_station_name to fix distinct values for the same start_station_id

```
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
13221, "Wood St & Milwaukee Ave", Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
13099, "Halsted St & 18th St", Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
13300, "DuSable Lake Shore Dr & Monroe St", Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
20215, "Hegewisch Metra Station", Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id== 351,
"Mulligan Ave & Wellington Ave", Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id== 444,
"N Shore Channel Trail & Argyle St", Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
"LF-055", "DuSable Lake Shore Dr & North Blvd",
Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
"TA1305000039", "Elston Ave & Cortland St", Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
"LF-005", "DuSable Lake Shore Dr & North Blvd",
Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
"TA1306000029", "DuSable Lake Shore Dr & Ohio St",
Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
"TA1307000041", "DuSable Lake Shore Dr & Wellington Ave",
Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
"TA1309000039", "DuSable Lake Shore Dr & Diversey Pkwy",
Cyclisticdf$start_station_name)
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
"TA1309000049", "DuSable Lake Shore Dr & Belmont Ave",
Cyclisticdf$start_station_name)
```

```
Cyclisticdf$start_station_name <- ifelse(Cyclisticdf$start_station_id==
"WL-008", "Clinton St & Roosevelt Rd", Cyclisticdf$start_station_name)
```

Saving this cleaned version of the DataFrame:

This is done only to save the process we have already covered:

```
write_csv(Cyclisticdf, "C:/Users/Luis Alejandro/Documents/Google Data
Analytics Professional Certificate/CASE STUDY 1 - CYCLISTIC/12 months worth
of data/Data/Cleaned_data-All_the_data-2022-05-22_V01.csv")
```