Project for ‘Cyclistic’: Data Cleaning

Igor Vysochanskyy

2024-03-31

This R Markdown template cleans the April 2023 dataset and can be applied to other 2023 datasets.

Install R packages & their libraries to enable subsequent operations.

Depending on your RStudio version, packages may be pre-installed, or you might need to install them manually.

```
library("tidyverse")
```

```
library("skimr")
```

```
library(openxlsx)
```

URL for all 2023 datasets: https://divvy-tripdata.s3.amazonaws.com/index.html (https://divvy-tripdata.s3.amazonaws.com/index.html)

If using as a template, change data frame names & directory accordingly

Upload original 202304-divvy-tripdata.csv (Only “202304…csv” is a part of GitHub repository)

```
tr304_orig <- read_csv("202304-divvy-tripdata.csv")
```

```
## Rows: 426590 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng
##
## 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.
```

Note 1: “started_at” & “ended_at” - (chr). Format to (dttm).

Creating a subset from selected 7 columns from the original dataset

```
trip304_orig <- subset(tr304_orig, select = c(ride_id, rideable_type, started_at, ended_at, memb
er_casual, start_station_name, end_station_name))
```

Detailed subset observation

```
skim_without_charts(trip304_orig)
```

Data summary

Name	trip304_orig
Number of rows	426590
Number of columns	7
________________________	
Column type frequency:	
character	7
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1.00	8	16	0	426590	0
rideable_type	0	1.00	11	13	0	3	0
started_at	0	1.00	13	15	0	38249	0
ended_at	0	1.00	13	15	0	38345	0
member_casual	0	1.00	6	6	0	2	0
start_station_name	63814	0.85	3	50	0	1069	0
end_station_name	68630	0.84	3	50	0	1071	0

Note 2: In “rideable_type” - n_unique is 3 - docked_bike is present!

Note 3: Max not = to Min in the “ride_id” column - remove duplicates!

Note 4: Missing values in the “start & end_station_name” columns.

Delete all rows with missing values in “start & end_stations_name” columns, except if rideable_type == “electric_bike”

```
trip304 <- trip304_orig %>%
  filter((!is.na(start_station_name) &  !is.na(end_station_name)) | rideable_type == "electric_b
ike")
```

Delete rows in “ride_id” if number of characters is not = 16. (Use for 2023, “03-04” datasets only)

```
bikeid304 <- trip304[nchar(trip304$ride_id) == 16, ]
```

Select 5 columns with relevant data at this point

```
trip5304 <- subset(bikeid304, select = c(ride_id, rideable_type,           started_at, ended_at, 
member_casual))
```

Format from (chr) to (dttm) “started_at” & “ended_at col”. (For 2023, “03-04” datasets only)

```
trip5304$started_at <- as.POSIXct(trip5304$started_at,
                                         format = "%m/%d/%Y %H:%M")
trip5304$ended_at <- as.POSIXct(trip5304$ended_at,
                                         format = "%m/%d/%Y %H:%M")
```

Create column “ride_length” as the trip duration in minutes

```
trip5304$ride_length <- difftime(trip5304$ended_at,
                      trip5304$started_at, units = "mins")
```

Percentiles of trip duration less or equal 2 min. for casual riders:

```
p04 <- seq(0.015, 0.04, by = 0.005)
values04 <- quantile(subset(trip5304, member_casual == "casual")$ride_length,
                    probs = p04)
```

Create a data frame with percentiles and values

```
p04_df <- data.frame(Percentile = p04, Value = values04)
```

Print the data frame for casual riders

```
print(p04_df)
```

```
##      Percentile  Value
## 1.5%      0.015 0 mins
## 2%        0.020 1 mins
## 2.5%      0.025 1 mins
## 3%        0.030 1 mins
## 3.5%      0.035 1 mins
## 4%        0.040 2 mins
```

Trim “ride_length” from outliers: keep 2 min. - 12 hour range

```
trim304 <- trip5304[(trip5304$ride_length >= 2 &
                     trip5304$ride_length <=  720), ]
```

skim_without_charts() reveals the extent of data cleanliness

```
skim_without_charts(trim304)
```

Data summary

Name	trim304
Number of rows	406644
Number of columns	6
_______________________	
Column type frequency:	
character	3
difftime	1
POSIXct	2
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1	16	16	0	406644	0
rideable_type	0	1	11	13	0	3	0
member_casual	0	1	6	6	0	2	0

Variable type: difftime

skim_variable	n_missing	complete_rate	min	max	median	n_unique
ride_length	0	1	2 mins	718 mins	9 mins	452

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
started_at	0	1	2023-04-01 00:00:00	2023-04-30 23:59:00	2023-04-15 09:54:00	38051
ended_at	0	1	2023-04-01 00:03:00	2023-05-01 08:06:00	2023-04-15 10:11:30	38084

Clean Data check list:

- Number of rows = n_unique for ride_id column - no duplicates;
- Same number of characters (16) in ride_id column;
- Whitespace and empty cells: 0 in each column;
- Completeness: complete_rate = 1 (no missing values in any column);
- Column names: Correct;
- Character length variation in columns 2-4 is normal for this data type;
- Overall, the dataset is clean and ready for analysis.

Export as .xlsx file & perform the next steps in Excel:

- Correct format in “ride_length” & add “day_of_week” columns
- Delete the “ended_at” column as not relevant anymore.
- Perform analysis by using Excel Pivot Table Charts.

```
write.xlsx(trim304, "Clean301-12\\Clean304.xlsx")
```

Final modification after analysis in Exel:

Remove rows where rideable_type = “docked_bike”. (Use for 2023, “01-08” datasets only)

```
trim_doc304 <- trim304[trim304$rideable_type != "docked_bike", ]
```

Remove “ended_at”. Keep columns with relevant data for analysis

```
data304 <- subset(trim_doc304, select = c(ride_id, rideable_type,           started_at, member_c
asual, ride_length))
```

Convert the “ride_length” column format as numeric (dbl)

```
data304$ride_length <- as.numeric(data304$ride_length)
data304$ride_length <- round(data304$ride_length, 1)
```

Add “day_of_week” column from “started_at”

```
data304$day_of_week <- format(data304$started_at, "%a")
```

Quick data observation

```
glimpse(data304)
```

```
## Rows: 398,149
## Columns: 6
## $ ride_id       <chr> "8FE8F7D9C10E88C7", "34E4ED3ADF1D821B", "5296BF07A2F77CB…
## $ rideable_type <chr> "electric_bike", "electric_bike", "electric_bike", "elec…
## $ started_at    <dttm> 2023-04-02 08:37:00, 2023-04-19 11:29:00, 2023-04-19 08…
## $ member_casual <chr> "member", "member", "member", "member", "member", "membe…
## $ ride_length   <dbl> 4, 23, 2, 4, 5, 4, 10, 5, 11, 20, 2, 11, 4, 4, 22, 4, 6,…
## $ day_of_week   <chr> "Sun", "Wed", "Wed", "Wed", "Wed", "Wed", "Wed", "Tue", …
```

Save modified and cleaned dataset as .RData object

```
save(data304, file = "RData_CleanBike\\CleanBike304.RData")
```