Case Study - Bike Share: Data Cleaning & Transformation

Jannatul Ashpia

```
#Load libraries
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.1     ✓ tibble    3.2.0
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors
```

```
library(lubridate)
library(ggplot2)
```

```
glimpse(april_Data)
```

```
## Rows: 426,590
## Columns: 13
## $ ride_id            <chr> "8FE8F7D9C10E88C7", "34E4ED3ADF1D821B",
"5296BF07A2…
## $ rideable_type      <chr> "electric_bike", "electric_bike",
"electric_bike", …
## $ started_at         <dttm> 2023-04-02 08:37:28, 2023-04-19 11:29:02,
2023-04-…
## $ ended_at           <dttm> 2023-04-02 08:41:37, 2023-04-19 11:52:12,
2023-04-…
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ start_lat          <dbl> 41.80, 41.87, 41.93, 41.92, 41.91, 41.91,
41.93, 42…
## $ start_lng          <dbl> -87.60, -87.65, -87.66, -87.65, -87.65, -87.63,
```

```
-87…
## $ end_lat            <dbl> 41.79, 41.93, 41.93, 41.91, 41.91, 41.92,
41.91, 41…
## $ end_lng            <dbl> -87.60, -87.68, -87.66, -87.65, -87.63, -87.65,
-87…
## $ member_casual      <chr> "member", "member", "member", "member",
"member", "…
```

```
glimpse(may_Data)
```

```
## Rows: 604,827
## Columns: 13
## $ ride_id            <chr> "0D9FA920C3062031", "92485E5FB5888ACD",
"FB144B3FC8…
## $ rideable_type      <chr> "electric_bike", "electric_bike",
"electric_bike", …
## $ started_at         <dttm> 2023-05-07 19:53:48, 2023-05-06 18:54:08,
2023-05-…
## $ ended_at           <dttm> 2023-05-07 19:58:32, 2023-05-06 19:03:35,
2023-05-…
## $ start_station_name <chr> "Southport Ave & Belmont Ave", "Southport Ave &
Bel…
## $ start_station_id   <chr> "13229", "13229", "13162", "13196",
"TA1308000047",…
## $ end_station_name   <chr> NA, NA, NA, "Damen Ave & Cortland St",
"Southport A…
## $ end_station_id     <chr> NA, NA, NA, "13133", "13229", "TA1306000029",
"1343…
## $ start_lat          <dbl> 41.93941, 41.93948, 41.85379, 41.89456,
41.95708, 4…
## $ start_lng          <dbl> -87.66383, -87.66385, -87.64672, -87.65345, -
87.664…
## $ end_lat            <dbl> 41.93000, 41.94000, 41.86000, 41.91598,
41.93948, 4…
## $ end_lng            <dbl> -87.65000, -87.69000, -87.65000, -87.67733, -
87.663…
## $ member_casual      <chr> "member", "member", "member", "member",
"member", "…
```

```
glimpse(june_Data)
```

```
## Rows: 719,618
## Columns: 13
## $ ride_id            <chr> "6F1682AC40EB6F71", "622A1686D64948EB",
"3C88859D92…
## $ rideable_type      <chr> "electric_bike", "electric_bike",
"electric_bike", …
## $ started_at         <dttm> 2023-06-05 13:34:12, 2023-06-05 01:30:22,
2023-06-…
## $ ended_at           <dttm> 2023-06-05 14:31:56, 2023-06-05 01:33:06,
2023-06-…
```

```
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ start_lat          <dbl> 41.91, 41.94, 41.95, 41.99, 41.98, 41.99,
41.88, 41…
## $ start_lng          <dbl> -87.69, -87.65, -87.68, -87.65, -87.66, -87.68,
-87…
## $ end_lat            <dbl> 41.91, 41.94, 41.92, 41.98, 41.99, 41.94,
41.88, 41…
## $ end_lng            <dbl> -87.70, -87.65, -87.63, -87.66, -87.65, -87.65,
-87…
## $ member_casual      <chr> "member", "member", "member", "member",
"member", "…
```

Merge the 3 data frames

```
merged_Df <- bind_rows(april_Data, may_Data, june_Data)
```

Clean up

Information Regarding longitude and Latitude is not necessary for our case, thus it can be removed

```
ride_data <- merged_Df %>%
  select(1:8,13)

colnames(ride_data)

## [1] "ride_id"            "rideable_type"      "started_at"
## [4] "ended_at"           "start_station_name" "start_station_id"
## [7] "end_station_name"   "end_station_id"     "member_casual"

summary(ride_data)

##    ride_id          rideable_type        started_at
##  Length:1751035     Length:1751035     Min.   :2023-04-01 00:00:02.00
##  Class :character   Class :character   1st Qu.:2023-05-02 12:11:47.00
##  Mode  :character   Mode  :character   Median :2023-05-25 07:27:46.00
##                                        Mean   :2023-05-22 00:56:35.10
##                                        3rd Qu.:2023-06-12 10:08:07.50
##                                        Max.   :2023-06-30 23:59:56.00
##     ended_at                       start_station_name start_station_id
##  Min.   :2023-04-01 00:03:10.00   Length:1751035     Length:1751035
##  1st Qu.:2023-05-02 12:24:17.00   Class :character   Class :character
##  Median :2023-05-25 07:39:58.00   Mode  :character   Mode  :character
##  Mean   :2023-05-22 01:15:33.91
```

```
##  3rd Qu.:2023-06-12 10:27:57.50
##  Max.   :2023-07-10 20:26:44.00
##  end_station_name   end_station_id     member_casual
##  Length:1751035     Length:1751035     Length:1751035
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Changes to apply to the Dataset

- Add a *ride_length* column calculating the duration for trips
- Remove rows for column *ride_length* if it is less than 0
- Split the *started_at* column further into *month, day, days of week*

```
#add ride_length column
ride_data <- ride_data %>%
  mutate(ride_length = ended_at - started_at)
```

This results in ride_length to appear as **seconds**

```
#convert it to numeric
ride_data$ride_length <- as.numeric(as.character(ride_data$ride_length))

#check for negative values
ride_data%>%
  select(ride_length) %>%
  filter(ride_length < 0)
```

```
## # A tibble: 21 × 1
##    ride_length
##          <dbl>
##  1          -3
##  2          -4
##  3          -5
##  4        -536
##  5         -12
##  6          -7
##  7         -90
##  8         -36
##  9         -11
## 10          -3
## # … with 11 more rows
```

```
#add columns for month, date, day
ride_data <- ride_data %>%
  mutate(month = format(as.Date(started_at), "%m"),
         day = format(as.Date(started_at), "%d"),
         day_of_week = format(as.Date(started_at), "%A"),
         hour = hour(started_at))
```

```
# create a new data frame excluding negative ride_length
trip_data <- ride_data[!(ride_data$ride_length<0),]
```

Descriptive Analysis

Note: ride_length is calculated in *seconds*

```
#analyzing ride_length
summary(trip_data$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     337     602    1139    1079 1767958
```

Casual vs Member Statistics on ride_length

```
trip_data %>%
  group_by(member_casual) %>%
  summarise(avg = mean(ride_length),
            min = min(ride_length),
            median = median(ride_length),
            max = max(ride_length))
```

```
## # A tibble: 2 × 5
##   member_casual   avg   min median     max
##   <chr>         <dbl> <dbl>  <dbl>   <dbl>
## 1 casual        1724.     0    751 1767958
## 2 member         765.     0    529   90031
```

Weekdays ride_length comparision for the member types

```
trip_data %>%
  group_by(member_casual, day_of_week) %>%
  summarise(avg = mean(ride_length),
            min = min(ride_length),
            median = median(ride_length),
            max = max(ride_length)) %>%
  arrange(member_casual, day_of_week)
```

```
## # A tibble: 14 × 6
## # Groups:   member_casual [2]
##    member_casual day_of_week   avg   min median     max
##    <chr>         <ord>       <dbl> <dbl>  <dbl>   <dbl>
##  1 casual        Sunday      1956.     0    874 1677336
##  2 casual        Monday      1694.     0    727 1199858
##  3 casual        Tuesday     1528.     0    656 1374550
##  4 casual        Wednesday   1476.     0    650 1054424
##  5 casual        Thursday    1496.     0    674 1002844
##  6 casual        Friday      1681.     0    734 1752631
##  7 casual        Saturday    2003.     0    886 1767958
##  8 member        Sunday       842.     0    565   90031
##  9 member        Monday       717.     0    497   89996
## 10 member        Tuesday      739.     0    509   89996
## 11 member        Wednesday    727.     0    515   89996
```

```
## 12 member        Thursday     736.     0    525   89996
## 13 member        Friday       756.     0    525   89996
## 14 member        Saturday     869.     0    590   89996
```

Total weekdays ride_length among member types

```
trip_data %>%
  group_by(member_casual, day_of_week) %>%
  summarise(total_ride = n(),
            avg_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week)
```

```
## # A tibble: 14 × 4
## # Groups:   member_casual [2]
##    member_casual day_of_week total_ride avg_duration
##    <chr>         <ord>            <int>        <dbl>
##  1 casual        Sunday          102512        1956.
##  2 casual        Monday           73928        1694.
##  3 casual        Tuesday          74162        1528.
##  4 casual        Wednesday        83511        1476.
##  5 casual        Thursday         94694        1496.
##  6 casual        Friday          111054        1681.
##  7 casual        Saturday        142827        2003.
##  8 member        Sunday          113242         842.
##  9 member        Monday          135834         717.
## 10 member        Tuesday         159655         739.
## 11 member        Wednesday       171681         727.
## 12 member        Thursday        178275         736.
## 13 member        Friday          164617         756.
## 14 member        Saturday        145022         869.
```