Cleaning Cylistic Data 2022-07

2023-07-31

Import data

```
data_01 <- read.csv(file="dataset/202207-divvy-tripdata.csv")
```

Check data 01

Check the data type for each meta

```
str(data_01)
```

```
## 'data.frame':    823488 obs. of  13 variables:
##  $ ride_id           : chr  "954144C2F67B1932" "292E027607D218B6" "57765852588AD6E0" "B5B6BE44314590
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2022-07-05 08:12:47" "2022-07-26 12:53:38" "2022-07-03 13:58:49" "2022-
##  $ ended_at          : chr  "2022-07-05 08:24:32" "2022-07-26 12:55:31" "2022-07-03 14:06:32" "2022-
##  $ start_station_name: chr  "Ashland Ave & Blackhawk St" "Buckingham Fountain (Temp)" "Buckingham Fou
##  $ start_station_id  : chr  "13224" "15541" "15541" "15541" ...
##  $ end_station_name  : chr  "Kingsbury St & Kinzie St" "Michigan Ave & 8th St" "Michigan Ave & 8th S
##  $ end_station_id    : chr  "KA1503000043" "623" "623" "TA1307000164" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "casual" "casual" ...
```

```
summary(data_01)
```

```
##    ride_id          rideable_type       started_at          ended_at
##  Length:823488      Length:823488      Length:823488      Length:823488
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  start_station_name start_station_id   end_station_name   end_station_id
##  Length:823488      Length:823488      Length:823488      Length:823488
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##    start_lat       start_lng         end_lat         end_lng
##  Min.   :41.64   Min.   :-87.84   Min.   :41.62   Min.   :-87.92
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66
##  Median :41.90   Median :-87.64   Median :41.90   Median :-87.64
##  Mean   :41.90   Mean   :-87.65   Mean   :41.91   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.07   Max.   :-87.52   Max.   :42.37   Max.   :-87.52
##                                   NA's   :947     NA's   :947
##  member_casual
##  Length:823488
##  Class :character
##  Mode  :character
##
##
##
##
```

From meta check we know that data type of column "started_at" and "end_at" should be datetime

Check duplicate data 01

```
print(data_01[duplicated(data_01), ])
```

```
##  [1] ride_id            rideable_type      started_at         ended_at
##  [5] start_station_name start_station_id   end_station_name   end_station_id
##  [9] start_lat          start_lng          end_lat            end_lng
## [13] member_casual
## <0 rows> (or 0-length row.names)
```

Duplicate data checking result : no data duplicate in data_01

Check missing value data in character data type

```
count(data_01[data_01$ride_id=="", ])
```

```
##   n
## 1 0
```

```
count(data_01[data_01$rideable_type=="", ])
```

```
##   n
## 1 0
```

```
count(data_01[data_01$started_at=="", ])
```

```
##   n
## 1 0
```

```
count(data_01[data_01$ended_at=="", ])
```

```
##   n
## 1 0
```

```
count(data_01[data_01$start_station_name=="", ])
```

```
##        n
## 1 112031
```

```
count(data_01[data_01$start_station_id=="", ])
```

```
##        n
## 1 112031
```

```
count(data_01[data_01$end_station_name=="", ])
```

```
##        n
## 1 120951
```

```
count(data_01[data_01$end_station_id=="", ])
```

```
##        n
## 1 120951
```

```
count(data_01[data_01$member_casual=="", ])
```

```
##   n
## 1 0
```

Missing value checking result :

ride_id: [0] rideable_type: [0] started_at: [0] ended_at: [0] start_station_name: [112,031] start_station_id: [112,031] end_station_name: [120,951] end_station_id: [120,951] member_casual: [0]

Fill Missing value with NA

Missing value (empty data) in start_station_name, start_station_id, end_station_name, end_station_id will be filling with NA

```
data_01 <- replace(data_01, data_01 == "", NA)
```

Fill missing value result : empty data was replace with NA

Check missing value data

```r
count(data_01[is.na(data_01$start_lat) | data_01$start_lat=="", ])
```

```
##   n
## 1 0
```

```r
count(data_01[is.na(data_01$start_lng) | data_01$start_lng=="", ])
```

```
##   n
## 1 0
```

```r
count(data_01[is.na(data_01$end_lat) | data_01$end_lat=="", ])
```

```
##     n
## 1 947
```

```r
count(data_01[is.na(data_01$end_lng) | data_01$end_lng=="", ])
```

```
##     n
## 1 947
```

Missing value checking result :

start latitude and langitude : [0] end latitude and langitude : [947]

Remove Missing value with NA

Missing value in end_lat, end_lng will be delete by remove the row

```r
data_01 <- data_01[!is.na(data_01$end_lat), ]
data_01 <- data_01[!is.na(data_01$end_lng), ]

print(data_01[is.na(data_01$end_lat) | data_01$end_lat=="", ])
```

```
##  [1] ride_id            rideable_type      started_at         ended_at
##  [5] start_station_name start_station_id   end_station_name   end_station_id
##  [9] start_lat          start_lng          end_lat            end_lng
## [13] member_casual
## <0 rows> (or 0-length row.names)
```

```r
print(data_01[is.na(data_01$end_lng) | data_01$end_lng=="", ])
```

```
##  [1] ride_id            rideable_type      started_at         ended_at
##  [5] start_station_name start_station_id   end_station_name   end_station_id
##  [9] start_lat          start_lng          end_lat            end_lng
## [13] member_casual
## <0 rows> (or 0-length row.names)
```

Remove missing value result : Row woth missing value data was removed

Check outliers in coordinate data

```
print(cat("start_lat : mean max min : ",
    mean(data_01$start_lat),
    max(data_01$start_lat),
    min(data_01$start_lat)))
```

```
## start_lat : mean max min :  41.90495 42.07 41.64NULL
```

```
print(cat("start_lng : mean max min : ",
    mean(data_01$start_lng), max(data_01$start_lng), min(data_01$start_lng)))
```

```
## start_lng : mean max min :  -87.6473 -87.52 -87.84NULL
```

```
print(cat("end_lat : mean max min : ",
    mean(data_01$end_lat), max(data_01$end_lat), min(data_01$end_lat)))
```

```
## end_lat : mean max min :  41.90517 42.37 41.62NULL
```

```
print(cat("end_lng : mean max min : ",
    mean(data_01$end_lng), max(data_01$end_lng), min(data_01$end_lng)))
```

```
## end_lng : mean max min :  -87.64742 -87.52 -87.92NULL
```

Outliers checking result : no outliers in coordinate data, max and min value for each data doesnt far from average value

Remove useless column data

Acording to the bussines task, start_station_name and end_station_name will be remove

```
data_01 <- data_01[, -which(names(data_01) == "start_station_name")]
data_01 <- data_01[, -which(names(data_01) == "end_station_name")]

head(data_01)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 954144C2F67B1932  classic_bike 2022-07-05 08:12:47 2022-07-05 08:24:32
## 2 292E027607D218B6  classic_bike 2022-07-26 12:53:38 2022-07-26 12:55:31
## 3 57765852588AD6E0  classic_bike 2022-07-03 13:58:49 2022-07-03 14:06:32
## 4 B5B6BE44314590E6  classic_bike 2022-07-31 17:44:21 2022-07-31 18:42:50
## 5 A4C331F2A00E79E0  classic_bike 2022-07-13 19:49:06 2022-07-13 20:15:24
## 6 579D73BE2ED880B3 electric_bike 2022-07-01 17:04:35 2022-07-01 17:13:18
##   start_station_id end_station_id start_lat start_lng  end_lat   end_lng
## 1            13224   KA1503000043  41.90707 -87.66725 41.88918 -87.63851
## 2            15541            623  41.86962 -87.62398 41.87277 -87.62398
## 3            15541            623  41.86962 -87.62398 41.87277 -87.62398
## 4            15541   TA1307000164  41.86962 -87.62398 41.79526 -87.59647
```

```
## 5     TA1307000117   TA1307000052  41.89147 -87.62676 41.93625 -87.65266
## 6            15535         WL-008  41.88461 -87.64456 41.86712 -87.64109
##   member_casual
## 1        member
## 2        casual
## 3        casual
## 4        casual
## 5        member
## 6        member
```

```
str(data_01)
```

```
## 'data.frame':    822541 obs. of  11 variables:
##  $ ride_id         : chr  "954144C2F67B1932" "292E027607D218B6" "57765852588AD6E0" "B5B6BE44314590E6
##  $ rideable_type   : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at      : chr  "2022-07-05 08:12:47" "2022-07-26 12:53:38" "2022-07-03 13:58:49" "2022-07
##  $ ended_at        : chr  "2022-07-05 08:24:32" "2022-07-26 12:55:31" "2022-07-03 14:06:32" "2022-07
##  $ start_station_id: chr  "13224" "15541" "15541" "15541" ...
##  $ end_station_id  : chr  "KA1503000043" "623" "623" "TA1307000164" ...
##  $ start_lat       : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng       : num  -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat         : num  41.9 41.9 41.9 41.8 41.9 ...
##  $ end_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual   : chr  "member" "casual" "casual" "casual" ...
```

Export clean data into csv

```
write.csv(data_01, "dataclean/202207-clean.csv", row.names = FALSE)
```