Cleaning Cylistic Data 2022-10

2023-07-31

Import data

```r
data_01 <- read.csv(file="dataset/202210-divvy-tripdata.csv")
```

Check data 01

Check the data type for each meta

```r
str(data_01)
```

```
## 'data.frame':    558685 obs. of  13 variables:
##  $ ride_id           : chr  "A50255C1E17942AB" "DB692A70BD2DD4E3" "3C02727AAF60F873" "47E653FDC2D992
##  $ rideable_type     : chr  "classic_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2022-10-14 17:13:30" "2022-10-01 16:29:26" "2022-10-19 18:55:40" "2022-
##  $ ended_at          : chr  "2022-10-14 17:19:39" "2022-10-01 16:49:06" "2022-10-19 19:03:30" "2022-
##  $ start_station_name: chr  "Noble St & Milwaukee Ave" "Damen Ave & Charleston St" "Hoyne Ave & Balm
##  $ start_station_id  : chr  "13290" "13288" "655" "KA1504000133" ...
##  $ end_station_name  : chr  "Larrabee St & Division St" "Damen Ave & Cullerton St" "Western Ave & Le
##  $ end_station_id    : chr  "KA1504000079" "13089" "TA1307000140" "620" ...
##  $ start_lat         : num  41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
```

```r
summary(data_01)
```

```
##    ride_id          rideable_type       started_at          ended_at        
##  Length:558685      Length:558685      Length:558685      Length:558685     
##  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:558685      Length:558685      Length:558685      Length:558685     
##  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.   :41.59   Min.   :-87.87
##  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   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.07   Max.   :-87.53   Max.   :42.13   Max.   :-87.52
##                                   NA’s   :475     NA’s   :475
##  member_casual
##  Length:558685
##  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

```r
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

Remove duplicate data

Remove Duplicate data result : No data to remove

Check missing value data in character data type

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

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

```r
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 91355
```

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

```
##       n
## 1 91355
```

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

```
##       n
## 1 96617
```

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

```
##       n
## 1 96617
```

```
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: [91,355] start_station_id: [91,355] end_station_name: [96,617] end_station_id: [96,617] 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

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

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

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

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

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

```
##     n
## 1 475
```

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

```
##     n
## 1 475
```

Missing value checking result :

start latitude and langitude : [0] end latitude and langitude : [475]

Remove Missing value with NA

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

```
# remove missing value data in this other data if there are also missing values
# data_01 <- data_01[!is.na(data_01$rideable_type), ]
# data_01 <- data_01[!is.na(data_01$started_at), ]
# data_01 <- data_01[!is.na(data_01$ended_at), ]
# data_01 <- data_01[!is.na(data_01$member_casual), ]

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

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

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

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

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

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

Check outliers in coordinate data

```r
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 :  41.90038 42.07 41.64NULL
```

```r
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 :  -87.64747 -87.52823 -87.84NULL
```

```r
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 :  41.90054 42.13 41.59NULL
```

```r
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 :  -87.64766 -87.52 -87.87NULL
```

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

```r
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 A50255C1E17942AB  classic_bike 2022-10-14 17:13:30 2022-10-14 17:19:39
## 2 DB692A70BD2DD4E3 electric_bike 2022-10-01 16:29:26 2022-10-01 16:49:06
## 3 3C02727AAF60F873 electric_bike 2022-10-19 18:55:40 2022-10-19 19:03:30
## 4 47E653FDC2D99236 electric_bike 2022-10-31 07:52:36 2022-10-31 07:58:49
## 5 8B5407BE535159BF  classic_bike 2022-10-13 18:41:03 2022-10-13 19:26:18
## 6 A177C92E9F021B99 electric_bike 2022-10-13 15:53:27 2022-10-13 15:59:17
##   start_station_id end_station_id start_lat start_lng  end_lat   end_lng
## 1            13290   KA1504000079  41.90068 -87.66260 41.90349 -87.64335
## 2            13288          13089  41.92004 -87.67794 41.85497 -87.67570
## 3              655   TA1307000140  41.97988 -87.68190 41.96640 -87.68870
## 4     KA1504000133            620  41.90227 -87.62769 41.89820 -87.63754
## 5            13028          13431  41.87475 -87.64981 41.86610 -87.60727
## 6            13028          13332  41.87472 -87.64983 41.87219 -87.66150
##   member_casual
## 1        member
## 2        casual
## 3        member
## 4        member
## 5        casual
## 6        casual
```

```
str(data_01)
```

```
## ’data.frame':    558210 obs. of  11 variables:
##  $ ride_id           : chr  "A50255C1E17942AB" "DB692A70BD2DD4E3" "3C02727AAF60F873" "47E653FDC2D99236
##  $ rideable_type     : chr  "classic_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2022-10-14 17:13:30" "2022-10-01 16:29:26" "2022-10-19 18:55:40" "2022-10
##  $ ended_at          : chr  "2022-10-14 17:19:39" "2022-10-01 16:49:06" "2022-10-19 19:03:30" "2022-10
##  $ start_station_id: chr  "13290" "13288" "655" "KA1504000133" ...
##  $ end_station_id  : chr  "KA1504000079" "13089" "TA1307000140" "620" ...
##  $ start_lat       : num  41.9 41.9 42 41.9 41.9 ...
##  $ start_lng       : num  -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ end_lat         : num  41.9 41.9 42 41.9 41.9 ...
##  $ end_lng         : num  -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual   : chr  "member" "casual" "member" "member" ...
```

Export clean data into csv

```
# write.csv(data_01, "dataclean/202210-clean.csv", row.names = FALSE)
```