Cleaning Cylistic Data 2023-02

2023-07-31

Import data

```
data_01 <- read.csv(file="dataset/202302-divvy-tripdata.csv")
```

Check data 01

Check the data type for each meta

```
str(data_01)
```

```
## 'data.frame':    190445 obs. of  13 variables:
##  $ ride_id           : chr  "CBCD0D7777F0E45F" "F3EC5FCE5FF39DE9" "E54C1F27FA9354FF" "3D561E04F739CC
##  $ rideable_type     : chr  "classic_bike" "electric_bike" "classic_bike" "electric_bike" ...
##  $ started_at        : chr  "2023-02-14 11:59:42" "2023-02-15 13:53:48" "2023-02-19 11:10:57" "2023-
##  $ ended_at          : chr  "2023-02-14 12:13:38" "2023-02-15 13:59:08" "2023-02-19 11:35:01" "2023-
##  $ start_station_name: chr  "Southport Ave & Clybourn Ave" "Clarendon Ave & Gordon Ter" "Southport Av
##  $ start_station_id  : chr  "TA1309000030" "13379" "TA1309000030" "TA1309000030" ...
##  $ end_station_name  : chr  "Clark St & Schiller St" "Sheridan Rd & Lawrence Ave" "Aberdeen St & Mon
##  $ end_station_id    : chr  "TA1309000024" "TA1309000041" "13156" "TA1309000008" ...
##  $ start_lat         : num  41.9 42 41.9 41.9 41.8 ...
##  $ start_lng         : num  -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 42 41.9 41.9 41.8 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "member" "member" ...
```

```
summary(data_01)
```

```
##    ride_id          rideable_type       started_at          ended_at
##  Length:190445      Length:190445      Length:190445      Length:190445
##  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:190445      Length:190445      Length:190445      Length:190445
##  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.   :41.65   Min.   :-87.84   Min.   :41.65   Min.   :-87.90
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66
##  Median :41.89   Median :-87.64   Median :41.89   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.   :42.07   Max.   :-87.53
##                                   NA's   :116     NA's   :116
##  member_casual
##  Length:190445
##  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

Remove duplicate data

Remove Duplicate data result : No data to remove

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 25473
```

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

```
##       n
## 1 25605
```

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

```
##       n
## 1 26738
```

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

```
##       n
## 1 26879
```

```
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: [25,473]

start_station_id: [25,605]

end_station_name: [26,738]

end_station_id: [26,879]

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
## 1 116
```

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

```
##     n
## 1 116
```

Missing value checking result :

start latitude and langitude : [0] end latitude and langitude : [116]

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
```

```
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

```
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 :  41.89807 42.07 41.6485NULL
```

```
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 :  -87.64778 -87.52841 -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 :  41.89835 42.07 41.6485NULL
```

```
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 :  -87.64791 -87.52823 -87.9NULL
```

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 CBCD0D7777F0E45F  classic_bike 2023-02-14 11:59:42 2023-02-14 12:13:38
## 2 F3EC5FCE5FF39DE9 electric_bike 2023-02-15 13:53:48 2023-02-15 13:59:08
## 3 E54C1F27FA9354FF  classic_bike 2023-02-19 11:10:57 2023-02-19 11:35:01
## 4 3D561E04F739CC45 electric_bike 2023-02-26 16:12:05 2023-02-26 16:39:55
## 5 0CB4B4D53B2DBE05 electric_bike 2023-02-20 11:55:23 2023-02-20 12:05:48
## 6 C67EB62172C472EB  classic_bike 2023-02-24 18:50:16 2023-02-24 18:56:40
##   start_station_id end_station_id start_lat start_lng  end_lat   end_lng
## 1     TA1309000030   TA1309000024  41.92077 -87.66371 41.90799 -87.63150
## 2            13379   TA1309000041  41.95788 -87.64958 41.96952 -87.65469
## 3     TA1309000030          13156  41.92077 -87.66371 41.88042 -87.65552
## 4     TA1309000030   TA1309000008  41.92087 -87.66373 41.87943 -87.63550
## 5     TA1307000160   KA1503000054  41.79483 -87.61879 41.78053 -87.60597
## 6     TA1308000050   TA1307000115  41.91213 -87.63466 41.90461 -87.64055
##   member_casual
## 1        casual
## 2        casual
## 3        member
## 4        member
## 5        member
## 6        member
```

```r
str(data_01)
```

```
## 'data.frame':    190329 obs. of  11 variables:
##  $ ride_id         : chr  "CBCD0D7777F0E45F" "F3EC5FCE5FF39DE9" "E54C1F27FA9354FF" "3D561E04F739CC45" ...
##  $ rideable_type   : chr  "classic_bike" "electric_bike" "classic_bike" "electric_bike" ...
##  $ started_at      : chr  "2023-02-14 11:59:42" "2023-02-15 13:53:48" "2023-02-19 11:10:57" "2023-02-26 16:12:05" ...
##  $ ended_at        : chr  "2023-02-14 12:13:38" "2023-02-15 13:59:08" "2023-02-19 11:35:01" "2023-02-26 16:39:55" ...
##  $ start_station_id: chr  "TA1309000030" "13379" "TA1309000030" "TA1309000030" ...
##  $ end_station_id  : chr  "TA1309000024" "TA1309000041" "13156" "TA1309000008" ...
##  $ start_lat       : num  41.9 42 41.9 41.9 41.8 ...
##  $ start_lng       : num  -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ end_lat         : num  41.9 42 41.9 41.9 41.8 ...
##  $ end_lng         : num  -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual   : chr  "casual" "casual" "member" "member" ...
```

Export clean data into csv

```r
# write.csv(data_01, "dataclean/202302-clean.csv", row.names = FALSE)
```