capstone_project_mod.R

ABHISHEK

2021-09-15

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(skimr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
#collecting data

Trips_sep20 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202009-divvy-tripdata.csv')
Trips_oct20 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202010-divvy-tripdata.csv')
Trips_nov20 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202011-divvy-tripdata.csv')
Trips_dec20 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202012-divvy-tripdata.csv')
Trips_jan21 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202101-divvy-tripdata.csv')
Trips_feb21 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202102-divvy-tripdata.csv')
Trips_mar21 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202103-divvy-tripdata.csv')
Trips_apr21 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202104-divvy-tripdata.csv')
Trips_may21 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202105-divvy-tripdata.csv')
Trips_jun21 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202106-divvy-tripdata.csv')
Trips_jul21 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202107-divvy-tripdata.csv')
Trips_aug21 <- read.csv('C:\\Users\\ayush\\Desktop\\Google Capstone Project\\data\\202108-divvy-tripdata.csv')

#compare all columns datatype

compare_df_cols(Trips_sep20, Trips_oct20, Trips_nov20, Trips_dec20, Trips_jan21, Trips_feb21, Trips_mar21, Trips_
apr21, Trips_may21, Trips_jun21, Trips_jul21, Trips_aug21, return = "mismatch")
```

```
##        column_name Trips_sep20 Trips_oct20 Trips_nov20 Trips_dec20 Trips_jan21
## 1   end_station_id     integer     integer     integer   character   character
## 2 start_station_id     integer     integer     integer   character   character
##   Trips_feb21 Trips_mar21 Trips_apr21 Trips_may21 Trips_jun21 Trips_jul21
## 1   character   character   character   character   character   character
## 2   character   character   character   character   character   character
##   Trips_aug21
## 1   character
## 2   character
```

```
#start_station_id and end_station_id for trips in sept, oct and nov are in integer format
#converting them to character format

Trips_sep20 <- mutate(Trips_sep20, end_station_id = as.character(end_station_id), start_station_id = as.character
(start_station_id))
Trips_oct20 <- mutate(Trips_oct20, end_station_id = as.character(end_station_id), start_station_id = as.character
(start_station_id))
Trips_nov20 <- mutate(Trips_nov20, end_station_id = as.character(end_station_id), start_station_id = as.character
(start_station_id))

#combining all individual data frames into a single one

all_trips <- bind_rows(Trips_sep20, Trips_oct20, Trips_nov20, Trips_dec20, Trips_jan21, Trips_feb21, Trips_mar21,
Trips_apr21, Trips_may21, Trips_jun21, Trips_jul21, Trips_aug21)

#removing unused columns (lat and lan)
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng))

#rename columns
all_trips <- all_trips %>%  rename(trip_id = ride_id,
                                   ride_type = rideable_type,
                                   start_time = started_at,
                                   end_time = ended_at,
                                   usertype = member_casual)

dim(all_trips)
```

```
## [1] 4913072       9
```

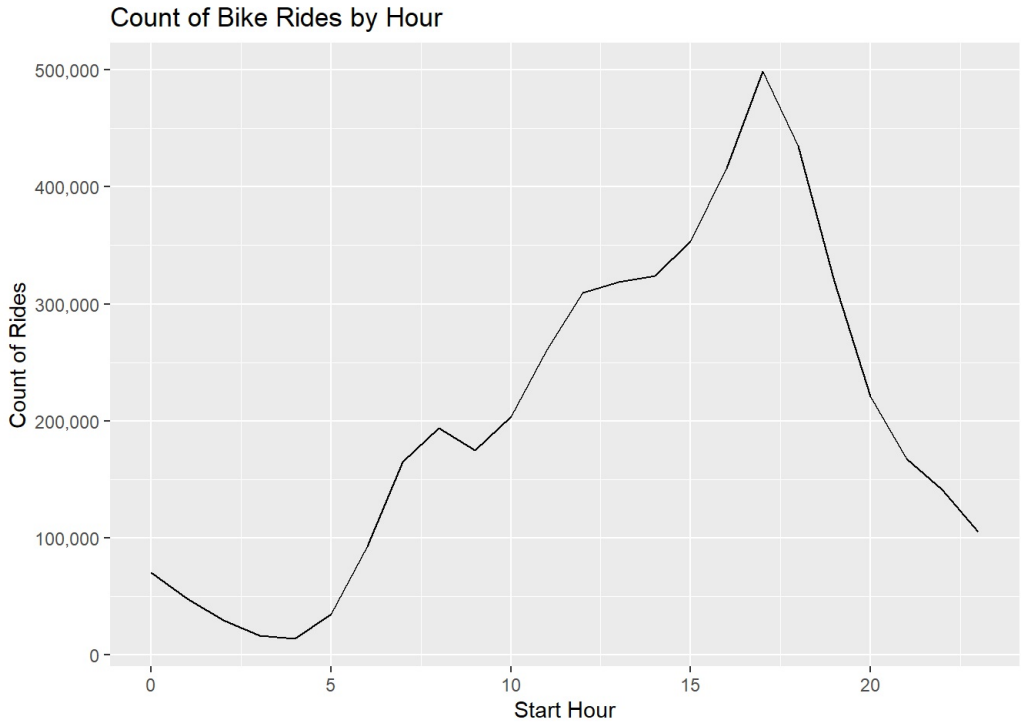
```
#remove all blank rows and columns


all_trips <- janitor::remove_empty(all_trips, which = c("rows", "cols"), quiet = TRUE)
dim(all_trips)
```

```
## [1] 4913072       9
```

```
#convert start_time and end_time in timestamps

all_trips$start_time <- lubridate::ymd_hms(all_trips$start_time)
all_trips$end_time <- lubridate::ymd_hms(all_trips$end_time)

#creating hour field
all_trips$start_hour <- lubridate::hour(all_trips$start_time)
all_trips$end_hour <- lubridate::hour(all_trips$end_time)

#creating date field
all_trips$start_date <- lubridate::date(all_trips$start_time)
all_trips$end_date <- lubridate::date(all_trips$end_time)

#counting no of rides each hour in a day
all_trips %>% count(start_hour, sort = T)
```

```
##    start_hour      n
## 1          17 499035
## 2          18 435037
## 3          16 415752
## 4          15 353319
## 5          14 323942
## 6          19 319336
## 7          13 318543
## 8          12 309405
## 9          11 260028
## 10         20 220792
## 11         10 203606
## 12          8 193682
## 13          9 174757
## 14         21 167851
## 15          7 165048
## 16         22 141681
## 17         23 105094
## 18          6  92918
## 19          0  70121
## 20          1  48314
## 21          5  34634
## 22          2  29544
## 23          3  16397
## 24          4  14236
```

```
#plotting graph between hours and no of corresponding trips

all_trips %>% count(start_hour, sort = T) %>%
  ggplot() + geom_line(aes(x=start_hour, y=n)) + scale_y_continuous(labels = comma) +
  labs(title = "Count of Bike Rides by Hour", x= "Start Hour", y= "Count of Rides")
```



```
all_trips$date <- as.Date(all_trips$start_time)
all_trips$month <- format(as.Date(all_trips$date), "%m")
all_trips$day <- format(as.Date(all_trips$date), "%d")
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")

#calculating trip duration
all_trips$trip_duration <- difftime(all_trips$end_time, all_trips$start_time, units = "hour")

#converting trip_duration datatype to numeric for further calculation
is.factor(all_trips$trip_duration)
```

```
## [1] FALSE
```

```
all_trips$trip_duration <- as.numeric(as.character(all_trips$trip_duration))
is.numeric(all_trips$trip_duration)
```

```
## [1] TRUE
```

```
all_trips <- all_trips[!(all_trips$trip_duration<0),]
skim(all_trips)
```

Data summary

Name	all_trips
Number of rows	4907672
Number of columns	18
_____	
Column type frequency:	
character	10
Date	3
numeric	3
POSIXct	2
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
trip_id	0	1.00	16	16	0	4907672	0
ride_type	0	1.00	11	13	0	3	0
start_station_name	0	1.00	0	53	449991	758	0
start_station_id	75735	0.98	0	36	374782	1294	0
end_station_name	0	1.00	0	53	491299	757	0
end_station_id	86114	0.98	0	36	405569	1294	0
usertype	0	1.00	6	6	0	2	0
month	0	1.00	2	2	0	12	0
day	0	1.00	2	2	0	31	0
day_of_week	0	1.00	6	9	0	7	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
start_date	0	1	2020-09-01	2021-08-31	2021-05-26	365
end_date	0	1	2020-09-01	2021-09-01	2021-05-26	366
date	0	1	2020-09-01	2021-08-31	2021-05-26	365

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
start_hour	0	1	14.31	4.94	0	11.00	15.00	18.00	23.0	
end_hour	0	1	14.48	5.04	0	11.00	15.00	18.00	23.0	
trip_duration	0	1	0.39	3.25	0	0.12	0.21	0.39	932.4	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
start_time	0	1	2020-09-01 00:00:07	2021-08-31 23:59:35	2021-05-26 17:55:52	4134134
end_time	0	1	2020-09-01 00:04:43	2021-09-01 17:37:35	2021-05-26 18:17:33	4120661

```
quantile(all_trips$trip_duration, probs = seq(.99, 1.0, by= .001))
```

```
##        99%      99.1%      99.2%      99.3%      99.4%      99.5%      99.6%
##   2.267500   2.377222   2.499722   2.645278   2.818889   3.051944   3.390556
##      99.7%      99.8%      99.9%       100%
##   4.077219   6.520176  17.116016 932.402500
```

```
#performing winsorization , defaulting the high outliers to a specified value of 30 hrs

high_pct <- 30

all_trips$trip_duration[all_trips$trip_duration > high_pct] <- high_pct
skim(all_trips)
```

Data summary

Name	all_trips
Number of rows	4907672
Number of columns	18
_____	
Column type frequency:	
character	10
Date	3
numeric	3
POSIXct	2
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
trip_id	0	1.00	16	16	0	4907672	0
ride_type	0	1.00	11	13	0	3	0
start_station_name	0	1.00	0	53	449991	758	0
start_station_id	75735	0.98	0	36	374782	1294	0
end_station_name	0	1.00	0	53	491299	757	0
end_station_id	86114	0.98	0	36	405569	1294	0
usertype	0	1.00	6	6	0	2	0
month	0	1.00	2	2	0	12	0
day	0	1.00	2	2	0	31	0
day_of_week	0	1.00	6	9	0	7	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique

start_date	0	1	2020-09-01	2021-08-31	2021-05-26	365
end_date	0	1	2020-09-01	2021-09-01	2021-05-26	366
date	0	1	2020-09-01	2021-08-31	2021-05-26	365

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
start_hour	0	1	14.31	4.94	0	11.00	15.00	18.00	23	
end_hour	0	1	14.48	5.04	0	11.00	15.00	18.00	23	
trip_duration	0	1	0.36	0.94	0	0.12	0.21	0.39	30	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
start_time	0	1	2020-09-01 00:00:07	2021-08-31 23:59:35	2021-05-26 17:55:52	4134134
end_time	0	1	2020-09-01 00:04:43	2021-09-01 17:37:35	2021-05-26 18:17:33	4120661

```
summary(all_trips$trip_duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.1200  0.2136  0.3646  0.3881 30.0000 
```