

--	--

Dataset Name	Date Created
divvy bike	2023/8/10

Date	Description
2023/8/13	add column: ride distance
2023/8/13	add column: ride_length
2023/8/13	add column: day_of_week
2023/8/14	add column: work_day
2023/8/13	add column: is_holiday
2023/8/14	add column: speed
2023/8/14	add column: rush_hour

Step	Date
1	2023/8/10
2	2023/8/10

3	2023/8/10
4	2023/8/11-12

Date	Strategy
2023/8/11-12	Fill the missing value with current matched values

Date	Transformation performed
2023/8/14	transform ride_length into four categories: very short trip, short trip. medium trip, long trip

Changelog

Data information

Last Updated

2023/8/14

Changes made

Change Details

calculate the distance between start and end station based on gps

calculate the length of the ride

determine the day of week

determine if it's work day
1: work day 0: not

determine if it's holiday
1: holiday 0: not

calculate the speed

determine if it's rush hour
1: rush hour 0: not
rush hour: 7-9 or 16-18 on work day

Data cleaning steps

Description

combine last 12 months data

basic data validation

advanced data validation

handle missing value

Missing data handling

Detail

For missing latitude and longitude I created a dictionary of each station and their location, if matched then fill the missing values.

For missing station names I used the same dictionary but with location as key and station as value to find matches to fill the missing values.

Data transformation

Reason for transformation

assign to 4 groups make analysis of user behavior easier

og

tion

Data Source

<https://divvy-tripdata.s3.amazonaws.com/index.html>

de

Reason for change

may need for further analysis

may need for further analysis

may need for further analysis

may need for further analysis

may need for further analysis

may need for further analysis

may need for further analysis

steps

Detail

check the type of every column

check if there are two stations with same id but different names and change the id if they are indeed different stations by calculating their distance $> 0.5\text{km}$

See Missing data handling

ndling

Note

ation

Note