DataEng S24: Data Transformation In-Class Assignment

Submit: Make a copy of this document and use it to record your results. Store a PDF copy of the document in your git repository along with any needed code. Submit the in-class activity submission form by Friday at 10:00 pm.

A. [MUST] Initial Discussion Questions

Discuss the following questions among your working group members at the beginning of the week and place your own response into this space. If desired, also include responses from your group members.

1. In the lecture we mentioned the benefits of Data Transformation, but can you think of any problems that might arise with Data Transformation?

- **Loss of Data**: Sometimes transformation techniques can unintentionally remove valuable data.
- **Bias Introduction**: If the transformation isn't carefully designed, it can introduce biases that skew analysis results.
- **Complexity**: Overly complex transformations can make the data hard to work with and understand.
- **Cost and Time**: These processes can be resource-intensive, taking up valuable time and computing power.
- **Misinterpretation**: Without proper documentation, transformed data might be misunderstood by those who use it later.

2. Should data transformation occur before data validation in your data pipeline or after?

- **Identify Errors Early and Often**: Validating first, and then again, seems wise. You want to catch any inaccuracies or anomalies in the data as early as possible. Validating first helps ensure that the transformations are applied to accurate and relevant data. Then, validate after transformation afterwards. A validation sandwich! Validation is cheap and should happen as often as possible!
- **Avoid Wasting Resources**: Transforming flawed data can waste computational resources and might require you to repeat both validation and transformation steps after errors are discovered.

B. [MUST] Small Sample of TriMet data

Here is sample data for one trip of one TriMet bus on one day (February 15, 2023): bc_trip259172515_230215.csv It's in .csv format not json format, but otherwise, the data is a typical subset of the data that you are using for your class project.

We recommend that you use google Colab or a Jupyter notebook for this assignment, though any python environment should suffice.

Use the pandas.read_csv() method to read the data into a DataFrame.

```
import pandas as pd
import sys, os
import matplotlib.pyplot as plt
from tqdm.auto import tqdm # For progress bar

pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

# downloads the .csv file from google drive only if it's not already in
directory
file_path = "/content/bc_trip259172515_230215.csv"
if os.path.isfile(file_path) == False:
 !gdown 1G4MUbdUFeACgjGcg6dcwykBZn21zg1jO

df = pd.read_csv(file_path, low_memory=False)
df.head(10)
```

C. [MUST] Filtering

Some of the columns in our TriMet data are not generally useful for our class project. For example, our contact at TriMet told us that the EVENT_NO_STOP column is not used and can be safely eliminated for any type of analysis of the data.

Use pandas.DataFrame.drop() to filter the EVENT_NO_STOP column.

For this in-class assignment we won't need the GPS_SATELLITES or GPS_HDOP columns, so drop them as well.

```
unneeded_cols = ['EVENT_NO_STOP', 'GPS_SATELLITES', 'GPS_HDOP']
needed_cols = list(set(df.columns) - set(unneeded_cols))
df = df.drop(unneeded_cols, axis=1)
```

Next, start over and this time try filtering these same columns using the usecols parameter of the read_csv() method.

```
df = pd.read_csv(file_path, usecols=needed_cols, low_memory=False)
```

Why might we want to filter columns this way instead of using drop()?

- **Less Cumbersome Code** - It just looks much cleaner loading in the data initially with "usecols" instead of taking in everything and then dropping columns. I haven't tested this but I'm also under the impression that it's likely a lot faster to use "usecols".

D. [MUST] Decoding

Notice that the timestamp for each bread crumb record is encoded in an odd way that might make analysis difficult. The breadcrumb timestamps are represented by two columns, OPD_DATE and ACT_TIME. OPD_DATE merely represents the date on which the bus ran, and it should be constant, unchanging for all breadcrumb records for a single day. The ACT_TIME field indicates an offset, specifically the number of seconds elapsed since midnight on that day.

We're not sure why TriMet represents the breadcrumb timestamps this way. We do know that this encoding of the timestamps makes automated analysis difficult. So your job is to decode TriMet's representation and create a new "TIMESTAMP" column containing a pandas.Timestamp value for each breadcrumb.

Suggestions:

- Use DataFrame.apply() to apply a function to all rows of your DataFrame
- The applied function should input the two to-be-decoded columns, then it should:
    - create a datetime value from the OPD_DATE input using datetime.strptime()
    - create a timedelta value from the ACT_TIME
    - add the timedelta value to the datetime value to produce the resulting timestamp

```
def Get_Timestamp(tdf):
 df = tdf.copy(deep=True)

 # convert 'OPD_DATE' column to pandas datetime object
 df['OPD_DATE'] = pd.to_datetime(df['OPD_DATE'], format='%d%b%Y:%H:%M:%S')

 # create a new 'TIMESTAMP' column by adding 'ACT_TIME' as a timedelta
object
 df['TIMESTAMP'] = df['OPD_DATE'] + pd.to_timedelta(df['ACT_TIME'],
unit='s')

 return df
```

E. [MUST] More Filtering

Now that you have decoded the timestamp you no longer need the OPD_DATE and ACT_TIME columns. Delete them from the DataFrame.

F. [MUST] Enhance

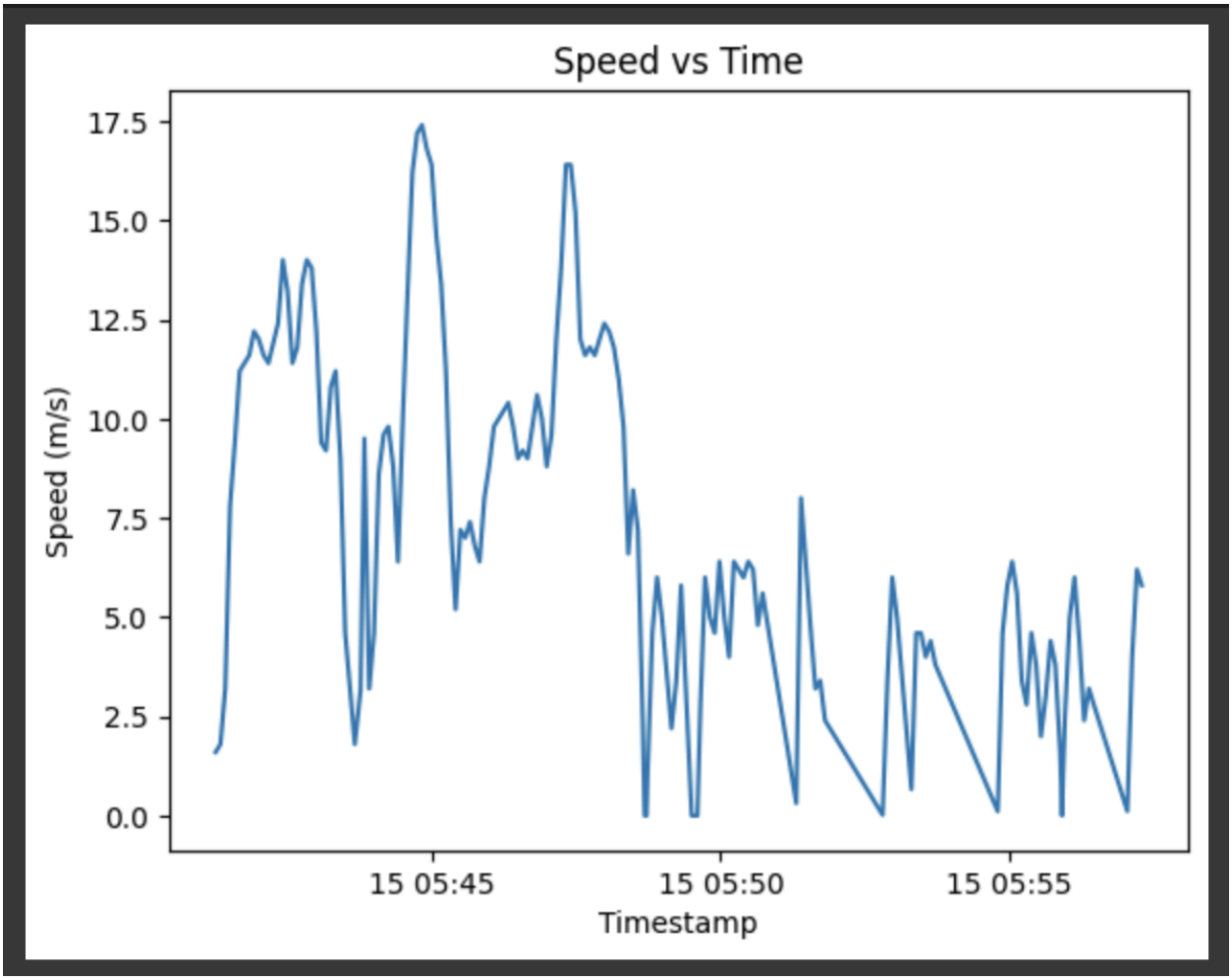
Create a new column, called SPEED, that is a calculation of meters traveled per second. Calculate SPEED for each breadcrumb using the breadcrumb's METERS and TIMESTAMP values along with the METERS and TIMESTAMP values for the immediately preceding breadcrumb record.

Utilize the pandas.DataFrame.diff() method for this calculation. diff() allows you to calculate the difference between a cell value and the preceding row's value for that same column. Use diff() to create a new dMETERS column and then again to create a new dTIMESTAMP column. Then use apply() (with a lambda function) to calculate SPEED = dMETERS / dTIMESTAMP. Finally, drop the unneeded dMETERS And dTIMESTAMP columns.

	EVENT_NO_TRIP	VEHICLE_ID	METERS	GPS_LONGITUDE	GPS_LATITUDE	TIMESTAMP	SPEED
0	259172515	4223	40	-122.648137	45.493082	2023-02-15 05:41:09	NaN
1	259172515	4223	48	-122.648240	45.493070	2023-02-15 05:41:14	1.6
2	259172515	4223	57	-122.648352	45.493123	2023-02-15 05:41:19	1.8
3	259172515	4223	73	-122.648385	45.493262	2023-02-15 05:41:24	3.2
4	259172515	4223	112	-122.648347	45.493582	2023-02-15 05:41:29	7.8
5	259172515	4223	159	-122.648357	45.494003	2023-02-15 05:41:34	9.4
6	259172515	4223	215	-122.648383	45.494510	2023-02-15 05:41:39	11.2
7	259172515	4223	272	-122.648375	45.495023	2023-02-15 05:41:44	11.4
8	259172515	4223	330	-122.648330	45.495555	2023-02-15 05:41:49	11.6
9	259172515	4223	391	-122.648213	45.496103	2023-02-15 05:41:54	12.2

Question: What is the minimum, maximum and average speed for this bus on this trip? (Suggestion: use the Dataframe.describe() method to find these statistics)

- Min = 0.0 m/s
- Max = 17.4 m/s
- Avg = 7.2 m/s



	EVENT_NO_TRIP	VEHICLE_ID	METERS	GPS_LONGITUDE	GPS_LATITUDE	TIMESTAMP	SPEED
count	161.0	161.0	161.000000	161.000000	161.000000	161	160.000000
mean	259172515.0	4223.0	3589.186335	-122.666642	45.514643	2023-02-15 05:48:21.670807552	7.227206
min	259172515.0	4223.0	40.000000	-122.677585	45.493070	2023-02-15 05:41:09	0.000000
25%	259172515.0	4223.0	1933.000000	-122.675990	45.503155	2023-02-15 05:44:34	3.800000
50%	259172515.0	4223.0	4165.000000	-122.669730	45.517902	2023-02-15 05:47:54	6.400000
75%	259172515.0	4223.0	5208.000000	-122.660792	45.523190	2023-02-15 05:51:34	10.850000
max	259172515.0	4223.0	5918.000000	-122.647997	45.528065	2023-02-15 05:57:19	17.400000
std	0.0	0.0	1810.681478	0.010093	0.010566	NaN	4.420604

G. [SHOULD] Larger Data Set

Here is breadcrumb data for the same bus TriMet for the entire day (February 15, 2023): [bc_veh4223_230215.csv]

Do the same transformations (parts C through F) for this larger data set. Be careful, you might need to treat each trip separately. For example, you might need to find all of the unique values for the EVENT_NO_TRIP column and then do the transformations separately on each trip.

Questions:
What was the maximum speed for vehicle #4223 on February 15, 2023?

- Max = 17.4 m/s

Where and when did this maximum speed occur?

- 2023-02-15

What was the median speed for this vehicle on this day?

- Avd = 7.14 m/s

H. [ASPIRE] Full Data Set

Here is breadcrumb data for all TriMet vehicles for the entire day (February 15, 2023):
bc_230215.csv

Do the same transformations (parts C through F) for the entire data set. Again, beware that simple transformations developed in parts C through F probably will need to be modified for the full data set which contains interleaved breadcrumbs from many vehicles.

Questions:
What was the maximum speed for any vehicle on February 15, 2023?

Where and when did this maximum speed occur?

Which vehicle had the fastest mean speed for any single trip on this day? Which vehicle and which trip achieved this fastest average speed?