DataEng S24: Data Transformation In-Class Assignment

A. Initial Discussion Questions

1. In the lecture we mentioned the benefits of Data Transformation, but can you think of any problems that might arise with Data Transformation?
Response:
One problem with data transformation is that we might lose important information. For example, if we remove some columns or change the data in a way that makes it simpler, we could miss important details that might be useful later.When we interpolate or fill missing values, there is a risk of introducing inaccuracies, especially if the filled values do not accurately represent the original data.These discrepancies can lead to incorrect analysis.

2. Should data transformation occur before data validation in your data pipeline or after?
Response:
Data transformation can happen both before and after data validation. When data is transformed before validation, it can help prepare the data for analysis and ensure that it meets the necessary requirements. Sometimes data validation can also lead to the identification of transformation needs.Data transformation can occur both before and after data validation to ensure that the data is accurate and suitable for analysis.

C. Filtering

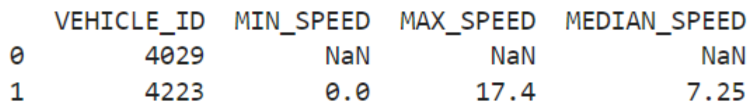
Why might we want to filter columns this way instead of using drop()?
Response:
When using the usecols parameter, we select the columns we want to keep in the DataFrame. On the other hand, with the drop parameter, we specify the columns we want to remove.With usecols, the specified columns are directly included in the DataFrame, replacing the original columns. However, with drop, the specified columns are removed from the DataFrame, but the original DataFrame remains unchanged. The drop operation returns a new DataFrame without the specified columns, leaving the original DataFrame as such.

F. Enhance

What is the minimum, maximum and average speed for this bus on this trip?
Minimum speed: 0.0
Maximum speed: 17.4
Average speed: 7.227205815018314

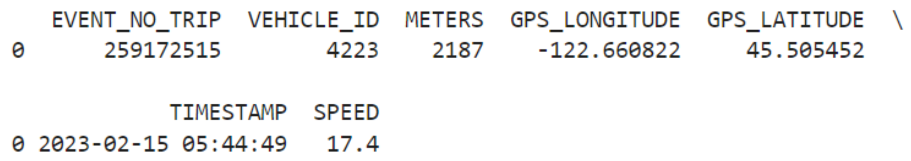
G. Larger Data Set

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



```
   VEHICLE_ID  MIN_SPEED  MAX_SPEED  MEDIAN_SPEED
0        4029        NaN        NaN           NaN
1        4223        0.0       17.4          7.25
```

Where and when did this maximum speed occur?



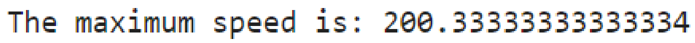
```
   EVENT_NO_TRIP  VEHICLE_ID  METERS  GPS_LONGITUDE  GPS_LATITUDE  \
0      259172515        4223    2187    -122.660822     45.505452

            TIMESTAMP  SPEED
0 2023-02-15 05:44:49   17.4
```

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

The median speed is 7.25

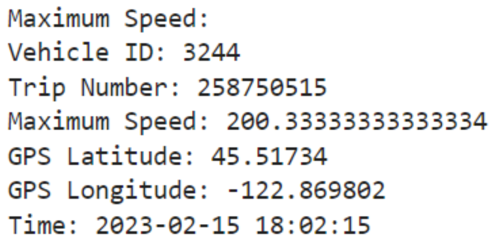
H. Full Data Set

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



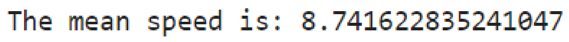
```
The maximum speed is: 200.33333333333334
```

Where and when did this maximum speed occur?

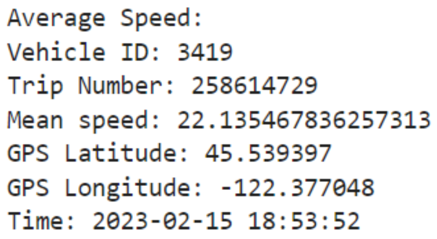


```
Maximum Speed:
Vehicle ID: 3244
Trip Number: 258750515
Maximum Speed: 200.33333333333334
GPS Latitude: 45.51734
GPS Longitude: -122.869802
Time: 2023-02-15 18:02:15
```

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



```
The mean speed is: 8.741622835241047
```



```
Average Speed:
Vehicle ID: 3419
Trip Number: 258614729
Mean speed: 22.135467836257313
GPS Latitude: 45.539397
GPS Longitude: -122.377048
Time: 2023-02-15 18:53:52
```