Assignment 3
CE 787A: Computational Tools for Transportation Engineering
Total Marks: 55

Instructions: Use Jupyter notebook to work on this assignment using Python language. All cell outputs and visualizations (if any) should be visible and necessary comments should be put to make the code readable. Once the code is ready, convert it to HTML (File > Download as HTML), save the html as pdf, and submit the pdf. Also, submit the *.ipynb file* of the jupyter notebook.

Q1. In this question, the task will be to preprocess a dataset for detecting transportation mode detection (TMD). TMD involves finding out the specific mode of transport (i.e., bus/car/train/walking/running etc.) the user is involved. Smartphone sensors have become an important data source for transportation mode detection. A sample paper details is given below to get further details on transportation mode detection using smartphones:

Hemminki, S., Nurmi, P., & Tarkoma, S. (2013). Accelerometer-based transportation mode detection on smartphones. *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems - SenSys '13*. 11th ACM Conference. https://doi.org/10.1145/2517351.2517367

Dataset:

The dataset file to be used for this question is named "cleaned.zip", which contains the csv file named "cleaned.csv". Unzip the folder to get the main csv data file.

The schema of the dataset is as follows: ***user, timestamp, x, y, z, class***

user: unique user id given to each participant
timestamp: timestamp of data recorded
x: Accelerometer reading X-axis obtained from sensor
y: Accelerometer reading Y-axis obtained from sensor
z: Accelerometer reading Z-axis obtained from sensor
class: labeled class i.e., transportation mode

There are 6 different transportation mode labels present in this dataset, namely bike, bus, car, e-bike, train, walk. Note, each user can have multiple sequences as you can see in one of the sample plot from the data in the Figure 1, given below where the user shown in the row 1 of the figure had 4 distinct sequences: two of the label "car" and two of the label "walk".

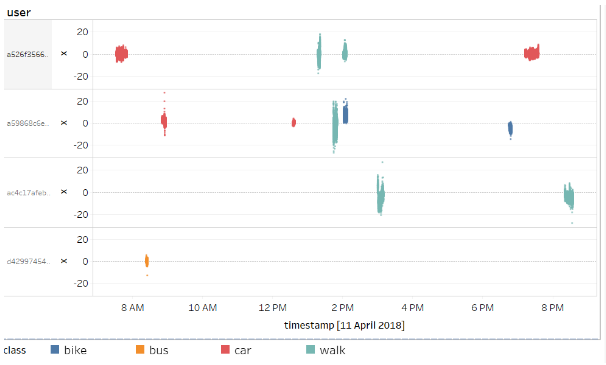


Figure 1. Sample plot from the dataset

The data is obtained from https://github.com/QROWD/QROWD_TMD. Refer to this github page if you want further details on the dataset. The detailed questions are provided next along with the marks allotted for each question. The specific outputs that need to be provided are mentioned and highlighted in italics. To better explain your code, you can output any other relevant details.

1. **Read the dataset “cleaned.csv”**. You can use python library pandas for reading the dataset. 1+1

a) Output the number of rows, columns, and columns names of the input data.
b) Find out the number of unique users present in the dataset and output the value (Column “user”)

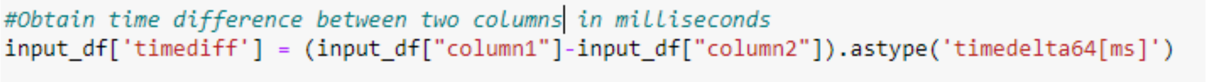
2. ***Determine the number of unique sequences*** of different transportation modes present in the dataset. For example, as stated before, the particular user shown in the row 1 of Figure 1 had 4 distinct sequences on 11th April 2018: two sequences of the class “car” and the remaining two of the class “walk”. If there is any discontinuity of greater than 10 seconds, assume it to be a separate sequence i.e., the minimum time gap between two unique sequences is assumed to be 10 seconds. 8+2

a) Output using a table the number of unique sequences present for each transportation mode for each user.

b) Output the time taken for your code to run this particular section of code (determining number of unique sequences).

You can use “timeit” function in Python to obtain the time taken to run a particular section of your code. Note, try to avoid “for loops” for determining the sequences since it can take a lot of time to process such a large file sequentially.

Hint: To find the time difference between two timestamps in two columns in **milliseconds**, you can use following command:



```
#Obtain time difference between two columns in milliseconds
input_df['timediff'] = (input_df["column1"]-input_df["column2"]).astype('timedelta64[ms]')
```

3. ***Extract features for each sequence***. Feature extraction is the main component for classifying sequence. For this assignment, extract features such as minimum, maximum, average, and standard deviation for each sequence. Therefore, when you extract these 4 features for each of 3 acceleration variables (x, y, and z), you will have 12 features in total (x_{min}, y_{min}, z_{min}, x_{max}, y_{max}, z_{max}, …). 8

Q2. The file “track_data.csv” contains pixel coordinates of vehicle trajectories for 3 different videos (represented as “video_id”).

The file schema is:

video_id, *veh_id*, *frame*, *bb_top*, *bb_left*, *bb_height*, *bb_width*

video_id : ID number of video.
veh_id: Tracking ID of vehicle. Each unique vehicle tracked in a video is provided with a unique id
frame: frame number of the corresponding video_id
bb_top: vehicle bounding box coordinate (in pixels) from top edge (see Figure 2)
bb_left: vehicle bounding box coordinate (in pixels) from left edge (see Figure 2)
bb_width: vehicle bounding box width (in pixels) (see Figure 2)
bb_height: vehicle bounding box height (in pixels) (see Figure 2)

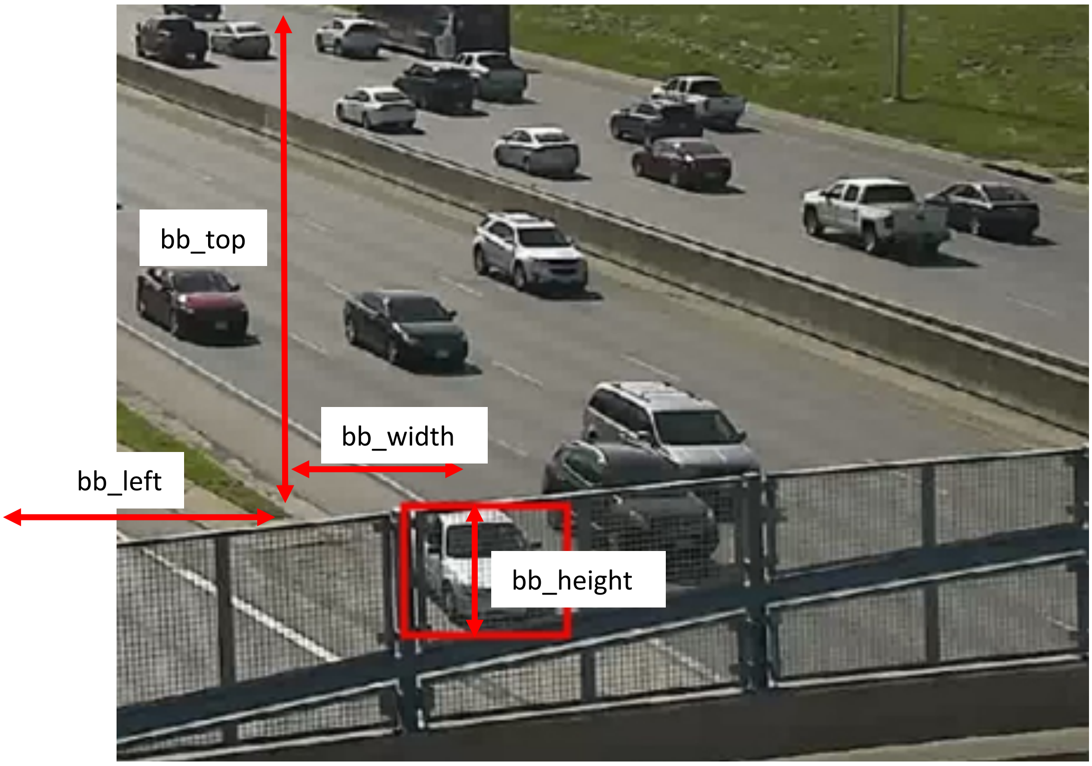


Figure 2

With this required information, perform the following tasks:

a) Read the file. Find the number of unique vehicle IDs in each *video_id* 2

b) Find the maximum frame number for each *video_id*. Assuming frame rate is 30 frames per second, what is the video duration for each *video_id* (For example, if maximum frame is 120, then video duration will be 120/30 = 4 seconds). 2

c) Remove all rows where bounding box height (*bb_height*) is less than 30 pixels. 1

d) For each *veh_id* and *video_id*, find out the standard deviation of bb_left and the range of frame count (difference between maximum and minimum frame count). 3

e) A vehicle can be said to be a stalled vehicle if it's position doesn't shift significantly for a specific period of time (say, at least 10 seconds). Therefore, for each *video_id*, find out the 3 *veh_ids* (total 9 *veh_ids* for 3 videos) having the lowest standard deviation of *bb_left* and their corresponding range of frame count. (Note, standard deviation of *bb_left* of a given veh_id can be an estimator of whether the vehicle is moving or not). Comment on whether any of them can be classified as a stalled vehicle. 5+2

Q3. The *wave_data.csv* and *inrix_data.csv* contain speed data obtained from two different sensors, named Inrix and Wavetronix. Inrix data is collected at 1 minute interval, while Wavetronix data is collected at 20-seconds interval.

The schema of *inrix_data.csv* and *wave_data.csv* is: *Code*, *Time*, *Speed*

Code: Unique sensor ID
Time: Timestamp in the format yyyy-mm-dd HH:MM:SS (For example, 2016-10-05 17:30:00)
Speed: Speed observed for the given code at the given timestamp

Each Inrix *Code* is associated with a specific Wavetronix *Code*. The correspondence between Inrix *Code* and Wavetronix *Code* is given below:

Inrix *Code*	Wavetronix *Code*
5033374	I-80-EB at WEST MIX-EB
5033375	I-235 EB to VALLEY WEST-EB
5033347	I-235 EB from Vly West Dr-EB
5032600	I-235 EB EAST OF 63RD-EB

With the above information, perform the following tasks:

1. Read both the data files *wave_data.csv* and *inrix_data.csv*. Use the Time column as the index column and parse the Time column in the datetime format. 3

2. Inrix data is collected at 1-minute interval. Resample the Inrix data with any interpolation method to obtain Inrix data at 20-seconds interval. 4

3. Merge the resampled Inrix data and Wavetronix data based on the timestamp and *code*. using the correspondence between Wavetronix and inrix code. 4

4. Determine the average Inrix and Wavetronix speed for 15-minute interval of each date (e.g., 2016-10-05 17:00, 2016-10-05 17:15, 2016-10-05 17:30, 2016-10-05 17:45, etc.) 4

5. We want to determine if the difference of speed between inrix and wavetronix is higher during any specific time period of the day. Therefore, determine the 15-minute intervals for each inrix-wavetronix pair for each date where the difference in Inrix and Wavetronix speed are maximum. Can you comment if you can find any pattern when the speed difference is higher between Inrix and Wavetronix? 5