CST8502 Project Dataset Selection

Select datasets from any of the following links:

https://gis-fdot.opendata.arcgis.com/
https://data.cityofnewyork.us/browse?q=public+safety

Choice 1

Dataset name with link: Bus Breakdown and Delays,
https://data.cityofnewyork.us/Transportation/Bus-Breakdown-and-Delays/ez4e-fazm

Goal (question in mind – 1 line):
With this dataset we could predict time how long a bus will be delayed by its attributes.

Description (in 2-3 lines, the tasks to find the answer to the question):
Since time delayed is a numeric value, we can use linear regression with relevant attributes to predict the amount of time a bus is delayed. Split into training and test data.

Number of attributes: 528K
Number of instances: 21

Choice 2

Dataset name with link: 2015 green taxi trip data ,
https://data.cityofnewyork.us/Transportation/2015-Green-Taxi-Trip-Data/gi8d-wdg5

Goal (question in mind – 1 line):
We would predict cab fare price.

Description (in 2-3 lines, the tasks to find the answer to the question):
Using relevant attributes, we could perform linear regression on a sample the data to be able generate a model. Then test the model on test data. The class will be the continuous value of cab fare.

Number of attributes: 21
Number of instances: 19.2M

Choice 3

Dataset name with link: https://data.cityofnewyork.us/Education/2014-2015-School-Closure-Discharge-Reporting-GPA/qd93-w582

Goal (question in mind – 1 line):
With this dataset we could predict the GPA of an individual

Description (in 2-3 lines, the tasks to find the answer to the question):
We could predict a GPA range for an individual using decision trees.
Number of attributes: 8
Number of instances: 1,323