Milestone 2

Data Science Report

CS3753 – Data Analysis

Fall 2020

Aaron Perez

Department of Computer Science

University of Texas at San Antonio

One UTSA Circle, San Antonio, TX 78249

Email: [tle728@my.utsa.edu](mailto:tle728@my.utsa.edu)

Mykel Roa

Department of Computer Science

University of Texas at San Antonio

One UTSA Circle, San Antonio, TX 78249

Email: [bva733@my.utsa.edu](mailto:bva733@my.utsa.edu)

David Levy

Department of Computer Science

University of Texas at San Antonio

One UTSA Circle, San Antonio, TX 78249

Email: [mca880@my.utsa.edu](mailto:mca880@my.utsa.edu)

1. Describe the data you have.

a. Where did you get the data?

A: We are using the dataset on https://www.kaggle.com/c/lyft-motion-prediction-autonomous-vehicles/overview

b. Did/will you do any preprocessing and if yes, what did/will you do and why?

A: Yes, we’re building a LocalDataManager object. This will resolve relative paths from the config using the L5KIT\_DATA\_FOLDER env variable we have just set. We will work with sample.zarr for developing intuition regarding the data. Please use train.zarr, validate.zarr and test.zarr for actual model training, validation and prediction

c. What is the format of the data that you will be using in your analysis? A: The format we are using is ZARR

d. Display a small sample of your data (before/after pre-processing if applicable).

A: We have 320124624 agents, 16265 scenes, 4039527 frames and 38735988 traffic light faces in train.zarr.

We have 312617887 agents, 16220 scenes, 4030296 frames and 29277930 traffic light faces in validation.zarr.

We have 88594921 agents, 11314 scenes, 1131400 frames and 7854144 traffic light faces in test.zarr.

e. What is the size of your data? (e.g, number of rows / columns if in table format, or number of records to be processed, etc.)

A: 40.5 GB (43,534,770,176 bytes)

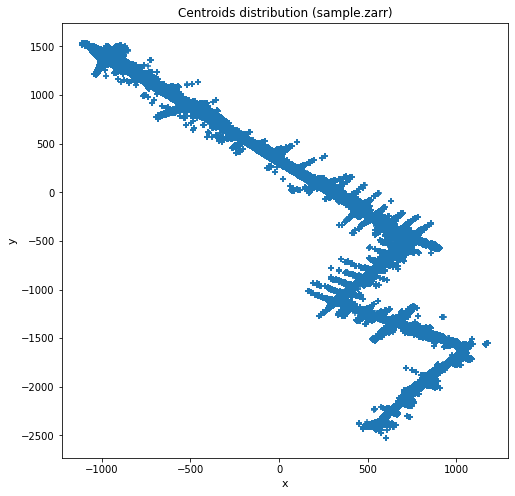
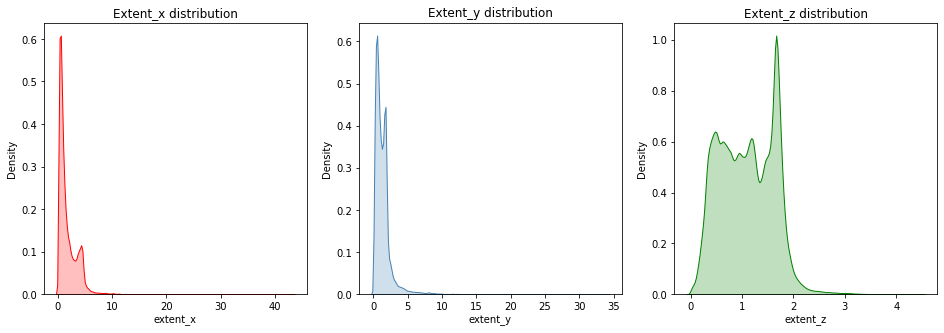
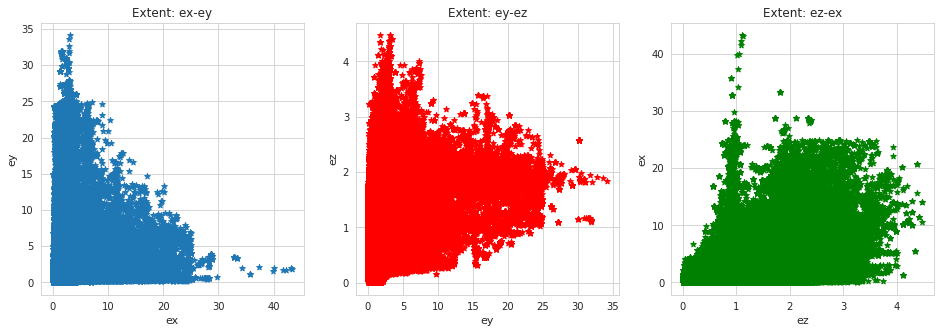
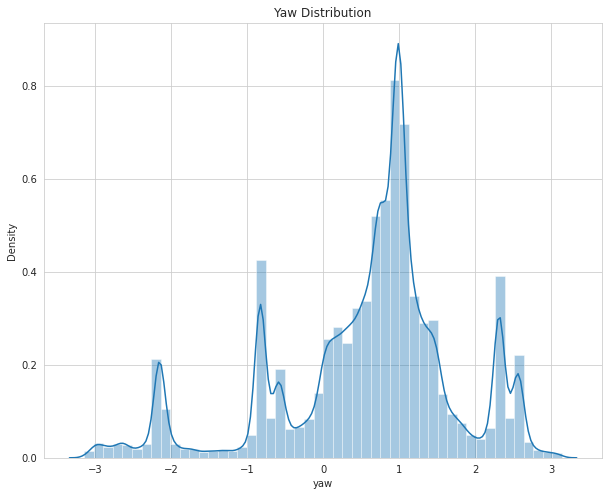
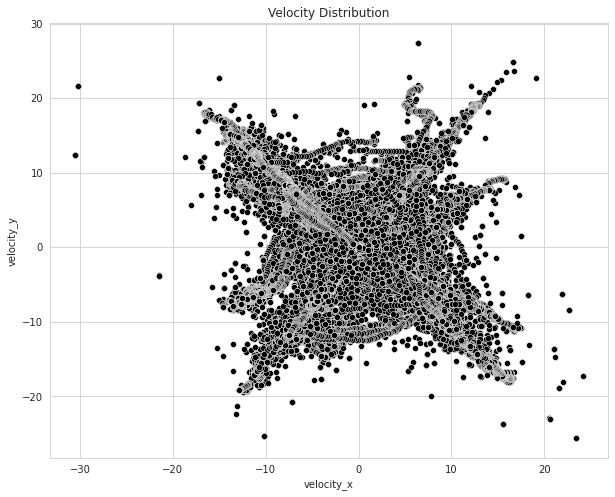
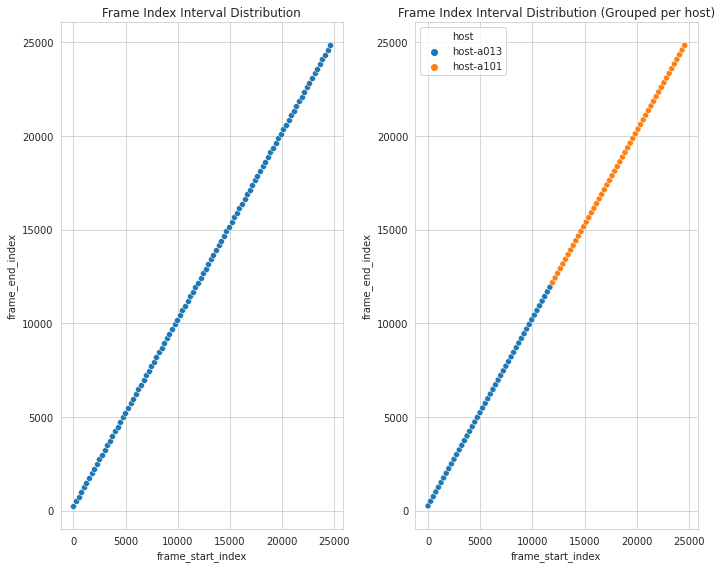
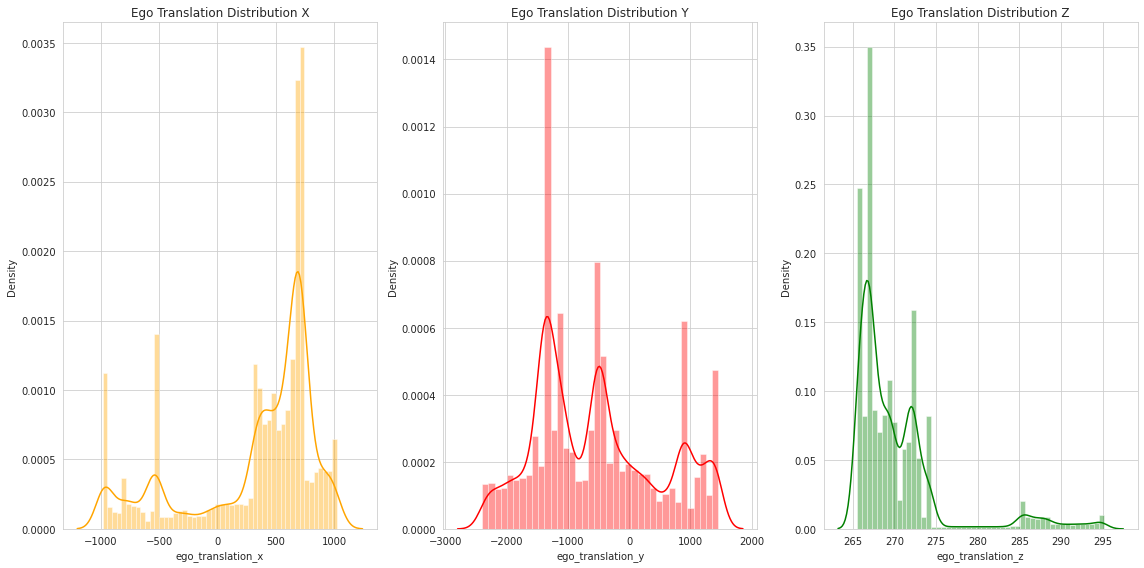
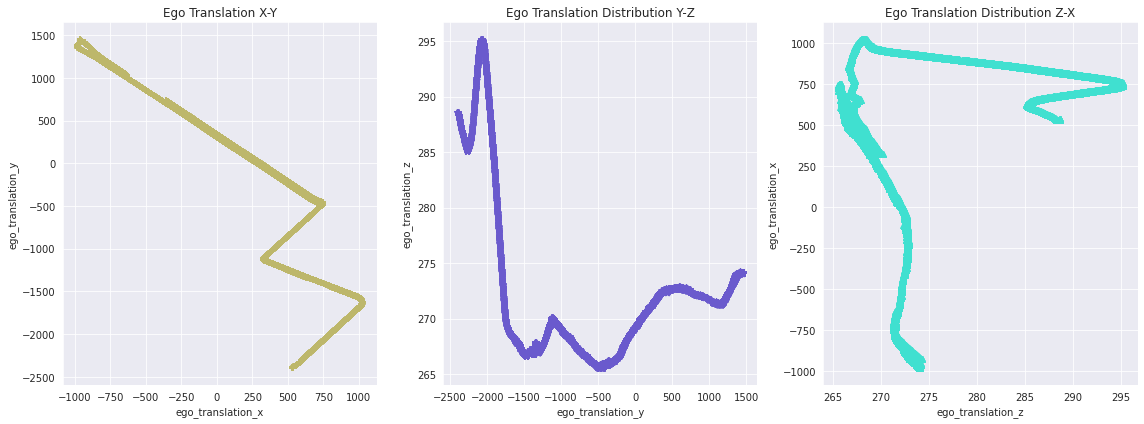
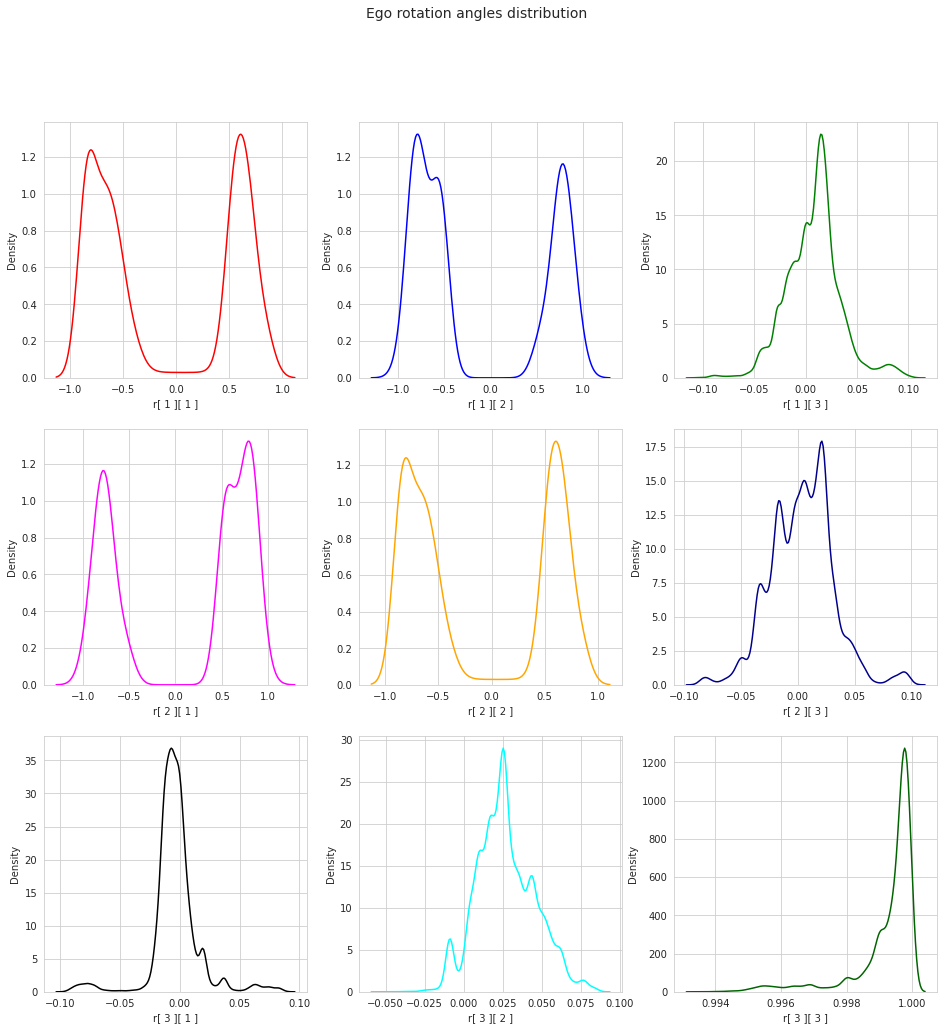
f. Do you have all the data that you will need, or you are planning to get more data to complete the project (if yes, discuss your plan.)?

A: We have all the data we need; we will pump the data into a frame and begin using the keys in each dataset to train using PyTorch for predictions. Once done, we will make sense of the data and provide an overview.

2. Show and discuss your preliminary analysis. For example, this could include:

a. What each feature in your data means and what are their distributions?

A:

* Centroid – Position of the agent. We can use this to pinpoint where the agent is. The positions show that there is a correlation of a road in the graph.
  + 
* Extents x, y, and z – The agent’s dimensions. These can be used to find the scale and area of the agent. These are right skewed, there are long tails in the positive direction.
  + 
* Extents – The agent’s dimensions compared to other dimensions. Again, there are long tails in the positive direction.
  + 
* Yaw – Rotation of an agent with respect to the vertical axis. This is important because we can find the direction that is pointing from the change in rotation. The yaw is generally between -1 and 3 but has the most amount in 1.
  + 
* Velocity – The speed of the agent. This graph is more skewed towards the middle, implying that there are lower or even stopped speeds that make up the data.
  + 
* Frame Index Interval – The differences in time between frame indexes. This gives an accurate time and should not have major changes. The graph shown is constant.
  + 
* Ego Translation – Position of the host car. This also shows the different ago translations relating to another dimension.
  + 
  + 
* Ego Rotation Angles – Rotation of the host car, which is collected by ego sensors. All of the graphs have at least one spike but some graphs have 2, avoiding the 0 in the middle.
  + 

b. Did/will you do any data transformation/normalization in order to achieve your goals?

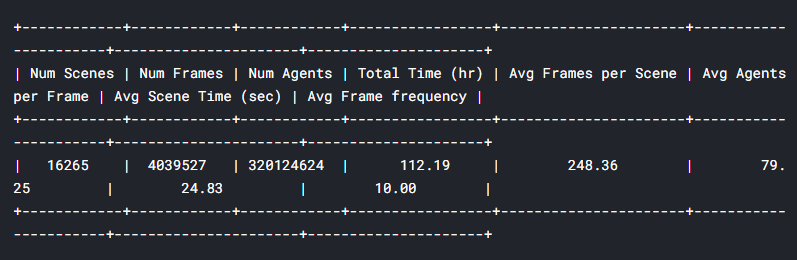
A: We have not manipulated the data in any way, nor do we plan to.

c. Did you have any hypothesis? If yes, how did you test it and did the test result support your hypothesis and why or why not?

A: TBD

d. Which machine learning / statistical methods did you use and what results do you have?

A: We’re using PyTorch to train our data, the result is the following:



e. What else do you plan to try and why?

A: N/A

f. How did / will you measure the performance of your model and how do you plan to improve?

A: The goal of this competition is to predict the trajectories of other traffic participants. Once done, we must yield a time sensitive run that trains the data in a valid fashion. More detail on this here https://github.com/lyft/l5kit/blob/master/competition.md