**Test for Pole Star KTP associate**

You are require to analyse two sets of data

1. Port visit data

2. Position data

The description of the data and the questions are presented in following pages and you are required to present your answers, graphs, discussion and codes in a single document, either a Microsoft word file or a pdf file.

You’re given 24 hours to work on the data, please submit your file before the deadline

**Port visit data**

This dataset (“port\_calls\_prod\_60\_8.csv”) contains port visit information for several ships around the world (total about 3 million records). **hours\_in port** is computed from **entered** and **departed** timestamps. It has following columns (fields):

ship\_name,

Imo\_number (unique for each ship),

mmsi,

ship\_status (In service),

ship\_type (some common ones are: *'Container Ship (Fully Cellular)', 'General Cargo Ship', 'Bulk Carrier', 'Passenger Ship', 'Products Tanker', 'Chemical/Products Tanker', 'Fishing Vessel', 'Chemical Tanker', 'Ro-Ro Cargo Ship', 'Crude Oil Tanker'* )

Flag\_name (country name: China, Panama, etc)

entered (port entry timestamp in UTC)

departed (port departure timestamp in UTC)

latitude,

longitude,

speed,

port\_name,

port\_country\_name,

port\_code,

status (moored, At anchor, could be empty/null)

**hours\_in\_port,**

gross\_tonnage,

length\_overall\_loa,

draught,

breadth

## Data Exploration

Some coding in dataframe package like panda (python) may be required to read data from the CSV dataset

* Compute mean **hours\_in\_port** for all dataset
* Which port is most visited in the dataset
* Compute mean, max and min **hours\_in\_port** for all ships entered US ports (port\_country\_name=’United States of America’)
* How can we remove outliers from the dataset?
* How can we visualize some port entry on a world map? (Bonus)
* How can we group port calls by ship\_type

## Prediction

* How can we extract train and test dataset from the provided dataset?
* Can we train a ML model to predict **hours\_in\_port** from the provided dataset?
* Which features could be useful in this prediction?

## Clustering

How we can cluster/group ships ports in following bins:

1. Hours\_in\_port < 6 hours
2. Hours\_in\_port in between 6 and 12 hours
3. Hours\_in\_port in between 12 and 24 hours
4. Hours\_in\_port > 24 hours

**Position Data**

“Ship9794850\_p1.csv” is a dataset that contains the positions of one ship at 45 timestamps. It has following columns (fields):

timestamp,

mmsi,

latitude,

longitude,

speed,

course,

heading,

draught,

destination,

status,

collection\_type

## Prediction

Use the data set, predict the latitude and longitude of the ship at the following time points:

2022-10-10 19:30:05 UTC

2022-10-10 19:33:06 UTC

2022-10-10 20:30:08 UTC

Explain how you work on the question and how you get the result. Please also include the codes that you used to implement the work.

If you can have more information what information do you think is useful to improve the prediction?