Ocean Test Task

Introduction

Imagine that you are involved in the development of a report on Ocean AIS dwell behavior; this report will be used as the basis to develop an *Insight* (i.e., dashboard-like web app with critical information to assist in decision making). While the main goal of this task is for you the answers the items in the "Requirements" section below, please keep in mind that your responses should have the *Insight* as the "greater" goal in mind , with the following priorities:

- describe the data
- understand the behavior of a dwell (i.e., time spent by a vessel being static, or close to being static)
- leverage the findings thus far to generate a satisfactory model to predict dwell times

Requirements

NOTE: Items 3 to 8 are Must-Haves. Items 9 to 13 are Nice-To-Haves.

1. We need to see your own code.
2. The data can be downloaded from the following URLs:
    a. Dataset in JSON format:
        i. https://tv-datascience.s3.amazonaws.com/public/data/test_tasks/ocean_ais/json/json.zip
    b. Dataset in Parquet format:
        i. https://tv-datascience.s3.amazonaws.com/public/data/test_tasks/ocean_ais/parquet/parquet.zip
    c. Both JSON and Parquet datasets are identical in contents, but you must choose to use one over the other. Please provide your justification for your choice of dataset.
3. What is(are) the main time period(s) in the data?
4. Which are the top three most sparse variables?
5. What region(s) of the world and ocean port(s) does this data represent? Provide evidence to justify your answer.
6. Provide a frequency tabulation of the various Navigation Codes & Descriptions (i.e., navCode & NavDesc). Optionally, provide any additional statistics you find interesting.
7. For MMSI = 205792000, provide the following report:
    a. Limit the data to only the TOP 5 Navigation Codes based from the response to question 6
    b. Provide the final state for each series of contiguous events with the same Navigation Code; series may be interrupted by other series, but each contiguous series must be its own record.
    c. Final report should include at least the following fields/columns:
        i. mmsi = the MMSI of the vessel
        ii. timestamp = the timestamp of the last event in that contiguous series
        iii. Navigation Code = the navigation code (i.e., navigation.navCode)
        iv. Navigation Description = the navigation code description (i.e., navigation.navDesc)
        v. lead time (in Milliseconds) = the time difference in milliseconds between the last and first timestamp of that particular series of the same contiguous navigation codes
8. For MMSI = 413970021, provide the same report as number 7
    a. Do you agree with the Navigation Code(s) and Description(s) for this particular vessel?
        i. If you do agree, provide an explanation why you agree.
        ii. If you do not agree, provide an explanation why do disagree. Additionally, if you do not agree, what would you change it to and why?
9. For each of the time period(s) from item three, provide a tabulation of the top 10 series of vessel navigation code/description ordered states.
10. Using the results from item 9, compare the volume of each vessel navigation code/description ordered states for each time period(s) from item three.
    a. Which increased/decreased?
11. For each of the time period(s) from item three and using only the "At Anchor" and "Moored" navigation descriptions, quantify the average "dwell"
12. Describe or show how you would quantify if the difference(s) in "dwells" between the time-period(s) is(are) significant.
13. Describe or show how you would create a Machine Learning Model to predict "dwell" times for the region.
    a. Bonus Points: Provide the code and performance results for your machine learning model on an OOB sample.
    b. In respect of time, performance of the ML model is not important and the number of features can be minimal. What mattes is that the code works and you can explain your thought process if asked.
14. We value code readability and consistency, and usage of modern community best practices and architectural approaches, as well, as functionality correctness. So pay attention to code quality.

15. Target completion time is about 4 hours. We would rather see what you were able to do in 4 hours than a full-blown algorithm you've spent days implementing. Note that in addition to quality, time used is also factored into scoring the task.

Expected Deliverables

1. Source code.
2. Readme, with instructions, how to build and run.