Take Home Assigment 2025

Nedo Lazic
Study Program: DBM
Student ID: 1027495

Furkan Yildiz
Study Program: DBM
Student ID: 1042537

June 26, 2025

1 Task 1- Getting started

For this project we worked in Rstudio with different Tools to optimize our workflow , these tools are Gitlab , Nginux , Docker Compose and VIM .

We created a Gitlab repository which you can clone from this link GitLab Repository

Our repository has following folder structure : 00 docs , which has subfolder literature 01 data which has subfolder raw where raw datat is stored , 02 code , which has 4 subfolders matlab ,phyton , shell , R - the folder R has another subfolder functions where are all functions that were used in R scripts stored 03 report , with subfolder table and graphs .

The official assignment description file, *thas2025.pdf*, was saved inside the 00 docs folder so it can be tracked and referenced as needed. This Overleaf report can be found in 03 report folder.

2 Task 2

2.1 2.1)

AIS (Automatic Identification System) data are collected through a network of terrestrial and satellite-based receivers. Each vessel broadcasts information about its identity, position, speed, and status at regular intervals. This data is crucial for monitoring maritime traffic, collision avoidance, and analyzing global shipping activity. AIS messages can be divided into two types: static data (e.g., vessel name, length, type, and destination) and dynamic data (e.g., position updates and real-time movements).

2.2 2.2)

To obtain the required information, we used PostgREST `GET` requests to access various API endpoints of the AIS database. PostgREST allows SQL-like syntax to filter and query the data we want to retrieve.

The data is organized into two main tables:

- **ais_static:** Contains information about static vessel attributes.
- **ais_dynamic:** Contains time-stamped dynamic data such as position, speed, and course.

For the static table, we used the following request to retrieve a small sample of the data:

```
GET /ais_static?limit=5
```

This gave us a quick overview of the structure and content.

For the dynamic table, due to the large volume of data, we filtered it using a timestamp condition:

```
GET /ais_dynamic?select=*&msg_timestamp=gte.2024-01-24T00:00:00Z&limit=1000
```

This request retrieved the first 1000 records starting from January 24th, 2024 at midnight.

Additional endpoints were used later in the project to access specific data required for analysis, which will be discussed in the following sections.

2.3 2.3) a) The *ais static*

table includes the following columns for each vessel ***mmsi***, ***imo***, ***name***, ***call sign***, ***flag***, ***draught***, ***ship type code***, ***shiptype***, ***length***, ***width***, ***eta***, **and** ***destination*** **,static update at.** **MMSI** stands for Maritime Mobile Service Identity number. It is a unique 9-digit number assigned to a ship, boat, or maritime radio station An MMSI is used for identifying the ship in AIS System, aswell for digital radio communication via DSC (Digital Selective Calling) .It is also very important for Ship Tracking Systems and Emergency Calls . mmsi is present in both **ais static** and **ais dynamic** tabel as **Primary** and **Foreign key** which allows us to connect the data and information from both tables . **Imo - International Maritime Organization Number** is a 7 digit number which is issued by International Maritime Organization , once determined the number never changes even after being sold , renamed or reflaged . The number is mostly used as indetification of the ship at ports and documents such as insurance. **Call sign** short for Radio call sign is The vessel's official international radio identifier. It's assigned by the flag state's maritime authority and used in radio communication, AIS, radar systems, and distress calls**.** **Draught** refers to the ***depth at which a boat is immersed in the water***, i.e. the vertical distance between the waterline and the lowest part of the hull. This measurement is crucial in determining the carrying capacity, stability and seaworthiness of the vessel. eta - **Estimated Time of Arrival** The ship's self-reported expected arrival time at its next port. Other fields such as **name , ship type, ship type code , static update at and destination** are also included and self explainitory .

2.4 The *ais_dynamic* table includes:

created_at, *msg_timestamp*, *position_updated_at*, *mmsi*, *latitude*, *longitude*, *speed*, *course*, *heading*, *status*, *maneuver*, *accuracy*, *rot*, and *collection_type.*, **msg_timestamp -** The original AIS timestamp sent by the ship for this position report. This is the actual time the ship transmitted the message. **latitude -** The geographic latitude of the ship's position, in decimal degrees **longitude -** The geographic longitude of the ship's position, in decimal degrees speed - the ship's Speed Over Ground , in **knots** ,nautical miles per hour. course - The ship's Course Over Ground – the direction of travel in degrees , relative to true north. heading - he heading of the vessel (direction the bow is pointing), in degrees. status - The ship's navigational status

2.5 2.3) b)

Each vessel corresponds to one row in the static table. We found a total of 220,685 entries in *ais_static* and 40,268,416 records in *ais_dynamic* for the given timestamp of 2024-01-24T00:00:00Z .

2.6 2.3) c)

There are 226 unique flags in the static dataset. The most common flag is **CN** (China), which occurred 13489 times.

2.7 2.3) d) Overview of ais_static

We performed summary statistics for key numeric columns of the ais static table. Mmsi has been recorded 220685 times . Most of the time it is a 9 digit number, but in some cases it is just a 8 digit number. The imo is a 7 digit number. There are some cases where it is only a 1 digit number. This indicates that the data has not been inserted correctly. The amount of the imo's are 112961. The length column had a **sum** of 12,867,786, a **maximum** of 1022, a **minimum** of 0, and an **average** of 61.896 meters. There are 2078932 entires for length. The width column had a sum of 2,371,019, a maximum of 126, a minimum of 0, and an average of 11.405 meters. The total amount of entires for width is 207183. The **number** of **unique destinations** was 111,718. However, the presence of 0 values in length and width indicates likely missing or erroneous entries. Vessels cannot have a physical length or width of zero, which suggests that 0 in this context may represent missing data. To address this problem, we created an R script to exclude non positive values when calculating the true minimum. Additionally, values such as length = 3 or width = 1 were identified as implausible. According to maritime standards, even the smallest fishing vessels typically have a minimum length of 5 meters and width of 3 meters. These anomalies likely arise from input errors, test entries, or unreported data.

The unique ship types identified were: Tug, Tanker, Cargo, Other, Fishing Vessel, Pleasure Craft, NA, and High-Speed Craft. The length variates between 0 and 336 , 0 beaing

There are 111718 destinations in total .

Radio call signs have these following patterns : BRDC@@@" "9V8372" "0000000" "A6E2358" "T8A4265" "021727@" "@@@@@@@" "975660@" "V7BN9"

Here are some unique ship names: ZHOU GANG TUO 11@@@@", "PHOENIX VANTAGE", "YI DE HUI HUANG", "BASS ANN", "MY FUTURE", "MS2HO", "JIANG TAI YOU 0068", ", "ZHEJIAXINGGONG8"

As you can see, there are some cases, where the name only has placeholders. (The placeholder is being displayed as @)

Such names often originate from call signs or internal IDs of vessels without a formally registered name. Based on names, you can find out the origin of the ship or the country of the owner.

2.8 2.3) e)

We iterated over 5 minute intervals of Januar 24, 2024 and took around 1000 observations of the data from ais dynamic tabel. On this data we applied methods of descriptive statistics . The summary of the data looks like this :

Variable	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
latitude	-38.145	3.181	29.388	23.721	40.977	78.149
longitude	-157.950	-17.220	12.270	23.230	103.500	153.140
speed	0.000	0.000	0.100	2.792	2.075	25.300
course	0.0	120.1	230.5	215.0	324.6	360.0
heading	0.0	199.2	425.0	346.0	511.0	511.0

Table 1: Descriptive statistics of raw (uncleaned) numeric AIS variables, incl. NA and zero values

As you can see, **speed**, **course**, and **heading** have a minimum of 0, which is unusual since the table represents dynamic (moving) vessels.

For this reason, we created two R functions: `impute_zero_with_mean`, which replaces 0 values with the column mean, and `impute_na_with_mean`, which replaces NA values with the mean of non-NA values. Both functions are stored in the `functions` folder.

After cleaning the data, the updated table looks like this:

Variable	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
latitude	-38.15	3.18	29.39	23.72	40.98	78.15
longitude	-157.95	-17.22	12.27	23.23	103.50	153.14
speed	0.10	2.79	4.81	4.81	4.81	25.30
course	5.40	142.90	215.0	221.7	299.1	360.0
heading	11.0	235.2	346.0	353.0	511.0	511.0

Table 2: Descriptive statistics of cleaned numeric AIS variables

As we can see, the minimum of **speed** is now 0.10, **course** is 5.40, and **heading** is 11.00. Consequently, the **1st Quartile**, **Median**, **Mean**, and **3rd Quartile** also changed. The maximum values remained the same.

The cleaned dataset is saved in the folder `01_data`, while the raw dataset can be found in the subfolder `raw`.

2.9 e)vi) - Sample points on displayed on map

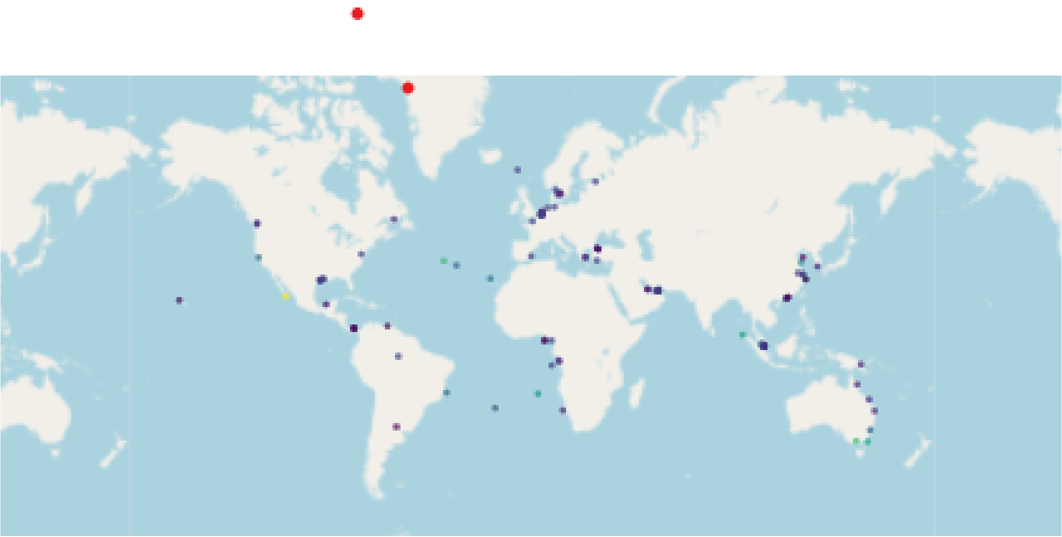

Figure 1: leaflet sample points

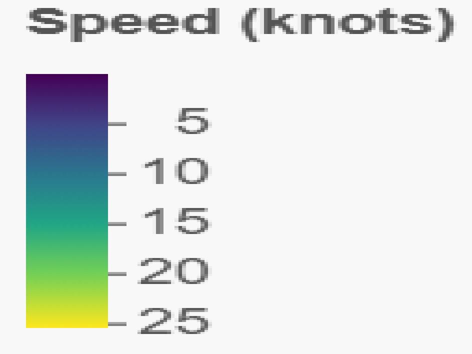

Figure 2: legend

When decided to visualize the sampled points using Leaflet , colored by vessel speed, a clear pattern can be seen . vessels tend to travel more slowly near coastal areas and move significantly faster in deep water. Near the coast vessels have to slow down when reaching the or leaving the ports . The ports have very strict traffic regulations and are more crowded for this reason the slower speed is required for safety reasons . On the other hand , vessels traveling in deep water are able to cover a longer distance without having problems with hight traffic or speed constraints . As a result , average speed is notably higher than in costal area . This map shows us how same variable speed and its different variations have different contextual meaning .

2.10 e)vii) - Sample points

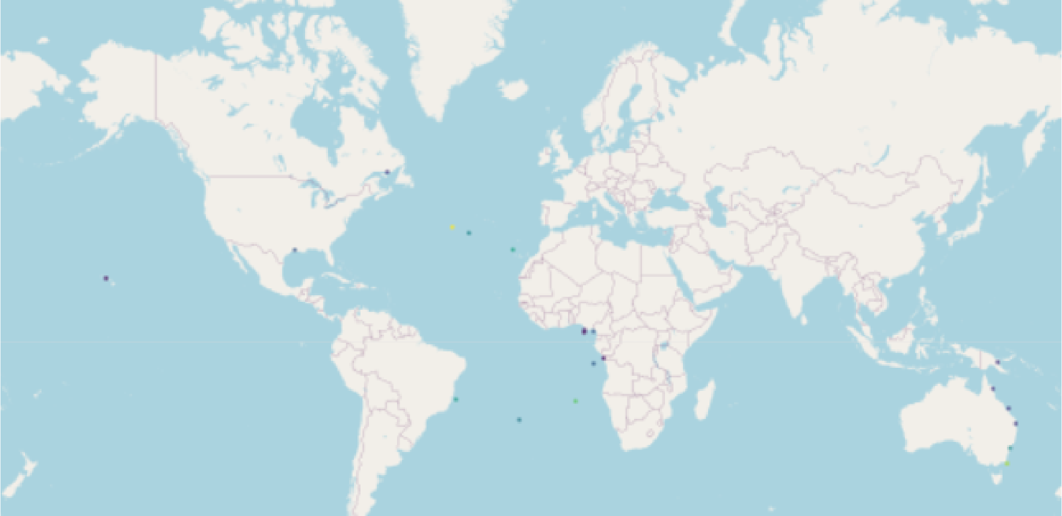

Figure 3: leaflet sample points

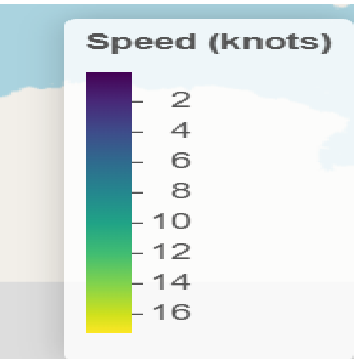

Figure 4: legend

When restricting the visualization to only AIS data with collection = 'satellite', the overall picture of vessel distribution changes significantly. Instead of showing many points near the coast as it is typical with land-based data,the satellite-collected data predominantly displays vessels located in open ocean areas. This shift occurs because satellite-based AIS reception covers remote parts of the world's oceans, where terrestrial AIS stations have no coverage. Also way less vessels are shown on the map

Terrestrial AIS receivers, by contrast, are land-based and typically have a limited range of 40 to 50 kilometers offshore. As a result, they are well-suited for capturing detailed vessel movements near shorelines and ports. Satellite receivers, on the other hand, provide global coverage but often with lower temporal resolution and sometimes lower data density, particularly in areas with high traffic, where AIS messages may collide or be dropped due to limited satellite bandwidth.

The change in the map highlights the influence of the data collection method on the spatial distribution of observed vessels. Satellite data allows us to observe global maritime traffic patterns, such as transoceanic shipping routes, but at the cost of losing some of the near coast data.

2.11 Ais dynamic individual Path

We created a R-Script which analyses static and dynamic tabel for vessels indetified under ***mmsi 2579999 and 412420898.***

The vessel under ***mmsi 2579999*** has many data quality issues. Following collumns have only Na values : speed,course ,heading status, maneuver, rot, imo, name, call sign, draught, ship type code ,ship type, length. Online Reasearch showed me that Navigation Aid is the type of this ship . On the website https://www.myshiptracking.com/ , i found out that this ship was for the last time spotted at 2025-06-04 01:23 at port in Hamburg . No further information has been published . This may indicate dummy AIS data, a misconfigured transponder, or a test device. Such cases highlight common data quality issues in AIS datasets . When visualizing this particular data with Leaflet using "clusterOptions = markerClusterOptions()" option , there are 4 green clusters with low amount of data which are located at north and west africa and 2 yellow clusters with larger amount of data which are located in UK , Norway and Germany and the other cluster being located in Australia . There is high possibility that there are some tracking problem at these are, due to high traffic .

The vessel indentified under ***mmsi 412420898*** named "ZHOU YUAN YU 2601" reveals several quality issues. When zooming in on the map you san see that for different time intervalls the same data has been written . For example speed , course and heading remain the same but the time stamp changes . This could possibly mean that the data wasnt measured the right way and was just transfered to different time stamp from last one . When visualizing vessel 412420898, no map background appears. This is most likely due to suspicious latitude/longitude values, either outside valid geographic bounds or too clustered in a narrow range. Although the vessel's speed and timestamps seem consistent, its location data may be corrupted or improperly encoded. This suggests a data quality issue, potentially caused by faulty GPS reporting or data preprocessing errors.

2.12 Forecasting and predicting Position and Error Evaluation

In order to forecast the future positions of vessels, we implemented two models based on the AIS dynamic data for a specific vessel (MMSI = 412420898):

Physically-based Model

This method uses basic kinematic equations to project the ship's position forward in time based on current speed and course. The projected displacement in nautical miles is calculated by:

$$\Delta x = v_n \cdot \Delta t \cdot \sin(\theta_n), \quad \Delta y = v_n \cdot \Delta t \cdot \cos(\theta_n)$$

where:

- v_n is the speed in knots,
- θ_n is the course in degrees from north,
- Δt is the time step in minutes.

The resulting displacement is converted back into geographic coordinates as follows:

$$\phi_{n+1} = \phi_n + \frac{\Delta y}{60}, \quad \lambda_{n+1} = \lambda_n + \frac{\Delta x}{60 \cdot \cos(\phi_n)}$$

Naive Model

This simpler model assumes that the vessel remains at its last known position, i.e., no change in latitude or longitude over time.

Accuracy Assessment: MSPE

To evaluate prediction accuracy, we computed the Mean Squared Prediction Error (MSPE) based on the great-circle distance between predicted and actual positions using the haversine formula:

$$\text{MSPE} = \frac{1}{N} \sum_{i=1}^{N} d\left((\phi_i, \lambda_i), (\hat{\phi}_i, \hat{\lambda}_i)\right)^2$$

with the distance function defined as:

$$d\left((\phi_i, \lambda_i), (\hat{\phi}_i, \hat{\lambda}_i)\right) = 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\hat{\phi}_i - \phi_i}{2}\right) + \cos(\phi_i)\cos(\hat{\phi}_i)\sin^2\left(\frac{\hat{\lambda}_i - \lambda_i}{2}\right)}\right)$$

Here, $R = 3440$ nautical miles represents Earth's radius.

Results for Different Time Steps

We applied both models using two different time steps:

- **MSPE (Physical, $\Delta t = 10$):** 3271.91 NM^2
- **MSPE (Naive, $\Delta t = 10$):** 0.0003 NM^2
- **MSPE (Physical, $\Delta t = 1$):** 32.66 NM^2
- **MSPE (Naive, $\Delta t = 1$):** 0.0003 NM^2
- **1 minute:** Surprisingly, the naive model performed better than the physically-based model, even at short horizons. This is likely because the vessel remained nearly stationary during this period, making the assumption of no movement quite accurate.
- **10 minutes:** The physically-based model's error increased drastically due to compounding errors in speed and course estimation. Again, the naive model yielded a lower MSPE by assuming position stability.

These findings show that in situations where vessels move slowly or remain stationary, the naive model can outperform even more complex prediction methods. For more dynamic movements, however, the physically-based model may perform better if calibrated more precisely.

3 Task3

Leaflet Map Based on Cleaned AIS Data

In this task, we build an interactive Leaflet map based on the cleaned AIS data from the dynamic table, which was processed in the previous task. The resulting map is saved in HTML format and can be accessed via the following URL:

https://193.197.230.65/sample_points.html

The HTML file is served through the root of the NGINX server, which is configured in the GitLab repository shiny underscore thas2025

4 Task 4

Gitlab repo shiny thas provided 3 R files , "server.R", "ui.R" and "global.R" . We modified these files with our own PostgRest URL and prediction model we created in Task 2 . The files are connected to Nginx server and can be found on

http://193.197.230.65/mmsi-search.

http://193.197.230.65/mmsi-search.

(The nginx and docker-compose setup can be found on remote server under ids student2 in folder Task3)

References