

Pol NERISSON
Gauthier GLOANEC
Paul LE GOFF

A3 - G1
A3 - G1
A3 - G1

Rapport de projet Big DATA



Projet réalisé du 02/06 au 06/06/2025.

Préambule :

Au cours de la première semaine de travail, du 2 au 6 juin 2025, nous avons posé les fondations de notre projet en Big Data dédié à l'analyse de trajectoires maritimes à partir des données AIS. L'objectif principal était de mettre en place une chaîne de traitement complète, depuis l'exploration et le nettoyage des données (plus de 420 000 points issus de 150 navires) jusqu'au développement de cinq grandes fonctionnalités : exploration des données, visualisations interactives, cartographie des routes et des ports fréquentés, analyse de corrélations et modélisation prédictive. Une application web Shiny a été conçue pour intégrer ces fonctionnalités, incluant une détection automatique des ports par DBSCAN et une interpolation temporelle des trajectoires.

Sommaire :

P.02 *Fonctionnalité 1 : Description et exploration des données*

P.07 *Fonctionnalité 2 : Graphiques Explicatifs*

P.12 *Fonctionnalité 3 : Visualisation sur une carte*

P.14 *Fonctionnalité 4 : Étude des corrélations entre variables*

P.18 *Fonctionnalité 5 : Prédiction / régression*

P.20 Conclusion

Fonctionnalité 1 : Description et exploration des données

A - Description des données :

Identification :

Nom du navire : nom officiel déclaré sur la licence radio maritime du navire.

Numéro MMSI (Identité dans le service mobile maritime): identification **unique, tous les bateaux en ont un.**

Numéro IMO (Organisation Maritime Internationale): navires de commerce de plus de 100 tonnes.

Call sign (Indicatif d'appel): numéro d'identification de la radio (donnés aux navires, sémaphores, ...).

→ Le numéro MMSI est le meilleur identifiant pour distinguer les navires, car il est attribué à tous les bateaux et il est unique à chacun.

Position et horodatage:

Latitude et longitude en degrés

Date et heure de référence

Données de navigation :

Vitesse par rapport à la surface terrestre (vitesse réelle) en noeuds

Direction (Heading, là où pointe le bateau) en degré, de 0 à 359.

Cap (cap suivi sur le fond, de 0 à 359,9°)

Remarque : Un navire ne se déplace pas toujours dans la direction où il pointe. Le cap sur le fond (COG) indique la trajectoire réelle du navire, qui peut différer de son orientation à cause du vent, des courants marins...

Caractéristiques du navire :

Type (vessel type correspondant au code AIS)

Longueur, largeur et tirant d'eau (hauteur immergé du bateau) en mètre

Type de cargaison (correspondant au code NAIS)

Statut :

État de navigation selon les règles COLREGS (0 - En route sous moteur, 1 - Au mouillage, ..., 15 - Indéfini / inconnu / valeur par défaut si non renseignée)

Classe d'équipement :

Type de transpondeur AIS (A ou B)

Classe A: Navires commerciaux, grands navires

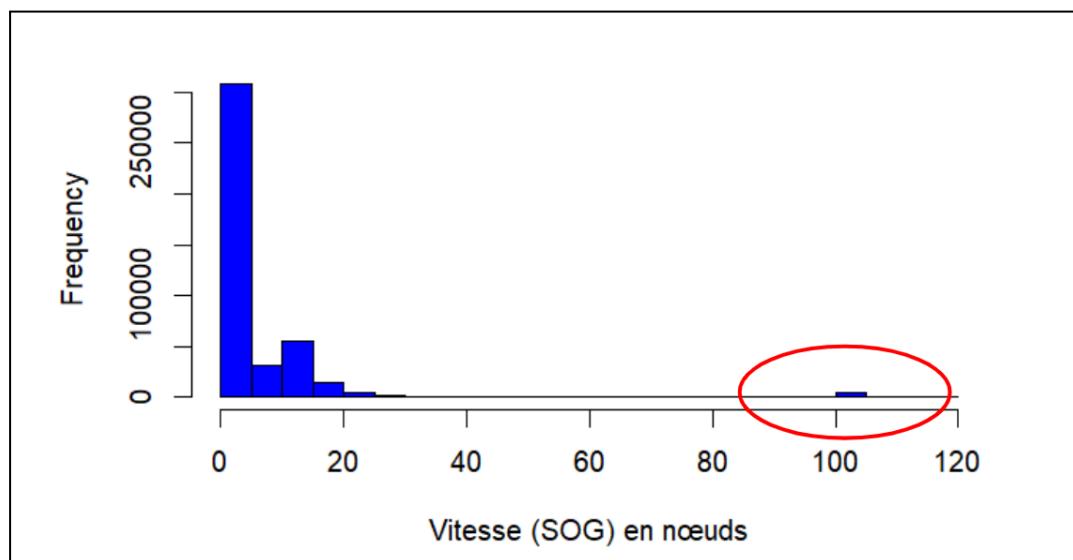
Classe B: Plaisanciers, pêcheurs côtiers, ...

B - Observation des données brutes :

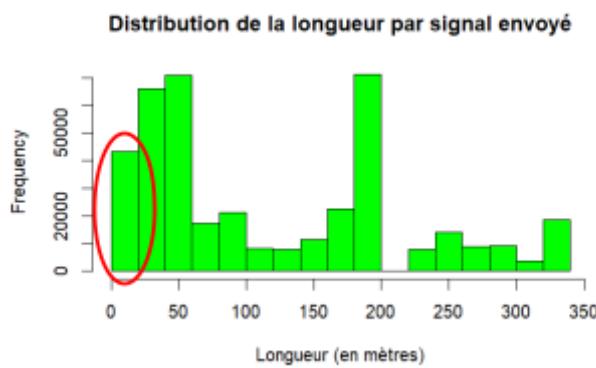
Nous disposons de 420 217 enregistrements AIS, correspondant à 150 bateaux, collectés entre le 25 mai et le 31 mai 2023.

Valeurs physiquement impossibles:

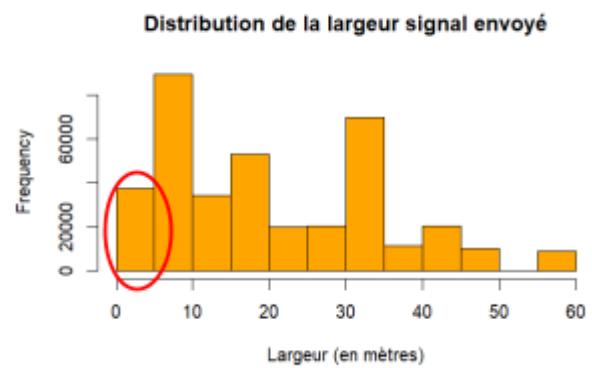
Vitesses aberrantes : Des bateaux affichent des vitesses jusqu'à 102 nœuds (189 km/h), ce qui est physiquement impossible pour la plupart des navires civils dans le Golfe du Mexique



Dimensions nulles : Certains bateaux ont une longueur ou largeur de 0 mètre, ce qui est techniquement impossible.



Length
Min. : 0.0

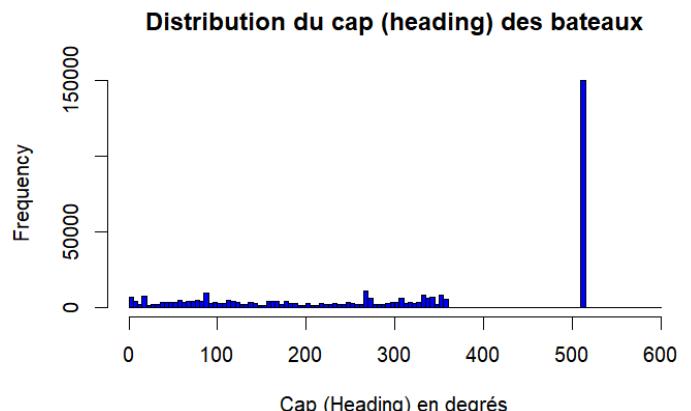
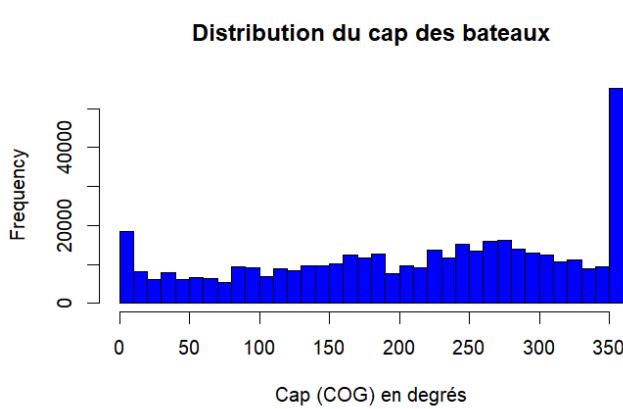


width
Min. : 0.00

Valeurs hors limites

Cap (Heading) aberrant : Des valeurs de cap supérieures à 359° alors que la plage valide est de 0° à 359°.

COG (Course Over Ground) : Certaines valeurs dépassent 360°, sortant de la plage normale 0-359,9°.



Positions hors zone d'étude : Certaines coordonnées placent les navires en dehors du Golfe du Mexique.

Ces incohérences rendent les données inexploitables en l'état car elles peuvent :

- Biaiser les modèles de prédiction
- Compromettre la fiabilité des statistiques de navigation

Feuille de route pour le tri :

On supprime les doublons.

Supprimer toute la ligne si valeur manquante :

MMSI, LAT, LON, SOG, BaseDateTime.

→ Ces informations sont cruciales. Sans l'une de ces informations, la donnée est inexploitable. On supprime la ligne.

Supprimer toute la ligne si valeur aberrante :

Si la vitesse est supérieure à 40 nœuds, on la considère comme aberrante.

Si LON et LAT hors Golf du Mexique

Supprimer uniquement la valeur : (la remplacer par NA)

COG: la plage de valeur est de 0 à 359,9°. Si la valeur est de 360° ou plus, on la supprime.

Heading: la plage de valeur est de 0 à 359. Si la valeur est de 360° ou plus, on la supprime.

Remplacer si valeur manquante :

Status : S'il n'y a pas de valeur, il faut mettre le bateau en code 15, qui correspond à Indéfini / inconnu : valeur par défaut si non renseignée

Inchangé :

VesselName, IMO et CallSign. COG, VesselType, Type de transpondeur, AIS (A ou B), Base de temps, Cargo, Draft.

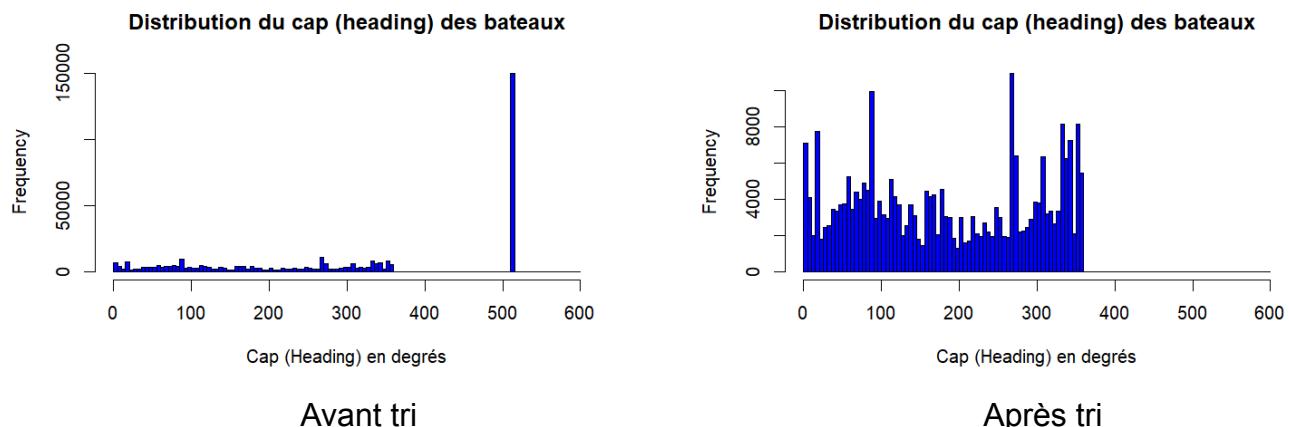
C - Tri / Nettoyage

En suivant la fiche de route précédemment établie, nous avons supprimé / remplacé et nettoyé l'ensemble des données problématiques en prenant en compte deux scénarios possibles.

Dans le premier scénario, on supprime les valeurs vides relative aux longueurs et largeurs du bateaux afin d'obtenir des prédictions plus précises, ainsi qu'une matrice de corrélation correcte.

Enfin dans le second, nous prenons en compte la possibilité que lors du projets d'Intelligence artificiel de la seconde semaine de projets il y ai la possibilité que l'on puisse prédire la largeur et la longueur à partir des autres données (notamment à partir du draft qui est très corrélé à ces deux valeurs).

Exemple de tri avant/après du Cap (Heading) :



Fonctionnalité 2 : Graphiques explicatifs

Pour mettre en image les données, nous avons fait plusieurs graphiques, ces graphiques trient surtout les différents types de bateaux en fonction du 420 217.

Passenger	60	60	Passenger, all ships of this type
Passenger	61	61	Passenger, hazardous category A
Passenger	62	62	Passenger, hazardous category B
Passenger	63	63	Passenger, hazardous category C
Passenger	64	64	Passenger, hazardous category D
Passenger	65	65	Passenger, reserved for future use
Passenger	66	66	Passenger, reserved for future use
Passenger	67	67	Passenger, reserved for future use
Passenger	68	68	Passenger, reserved for future use
Passenger	69	69	Passenger, no additional information
Cargo	70	70	Cargo, all ships of this type
Cargo	71	71	Cargo, hazardous category A
Cargo	72	72	Cargo, hazardous category B
Cargo	73	73	Cargo, hazardous category C
Cargo	74	74	Cargo, hazardous category D
Cargo	75	75	Cargo, reserved for future use
Cargo	76	76	Cargo, reserved for future use
Cargo	77	77	Cargo, reserved for future use
Cargo	78	78	Cargo, reserved for future use
Cargo	79	79	Cargo, no additional information
Tanker	80	80	Tanker, all ships of this type
Tanker	81	81	Tanker, hazardous category A
Tanker	82	82	Tanker, hazardous category B
Tanker	83	83	Tanker, hazardous category C
Tanker	84	84	Tanker, hazardous category D
Tanker	85	85	Tanker, reserved for future use
Tanker	86	86	Tanker, reserved for future use
Tanker	87	87	Tanker, reserved for future use
Tanker	88	88	Tanker, reserved for future use
Tanker	89	89	Tanker, no additional information

De plus :

Hazardous category A (explosifs, gaz hautement inflammables)

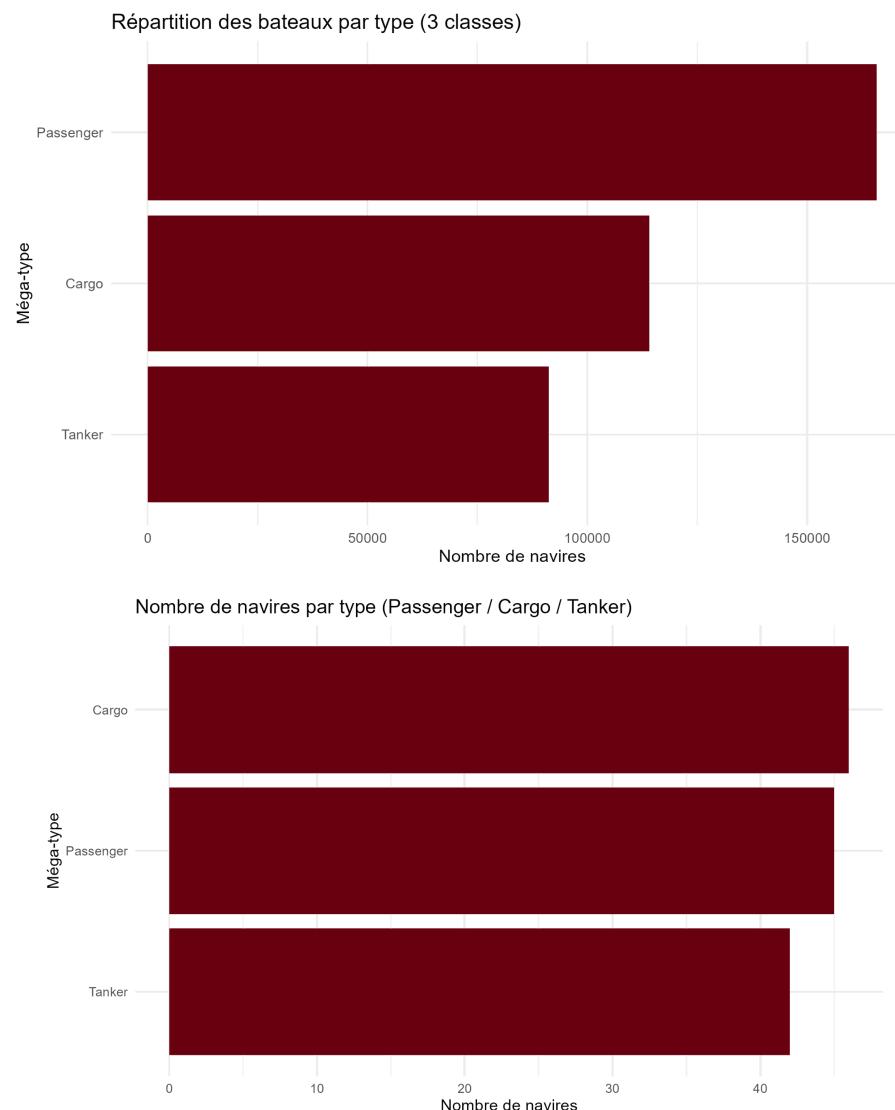
Hazardous category B (liquides inflammables)

Hazardous category C (matières toxiques)

Hazardous category D (matières dangereuses diverses)

Même les navires passagers qui sont réservés au transport de personnes peuvent embarquer du cargo dangereux comme du gaz par exemple.

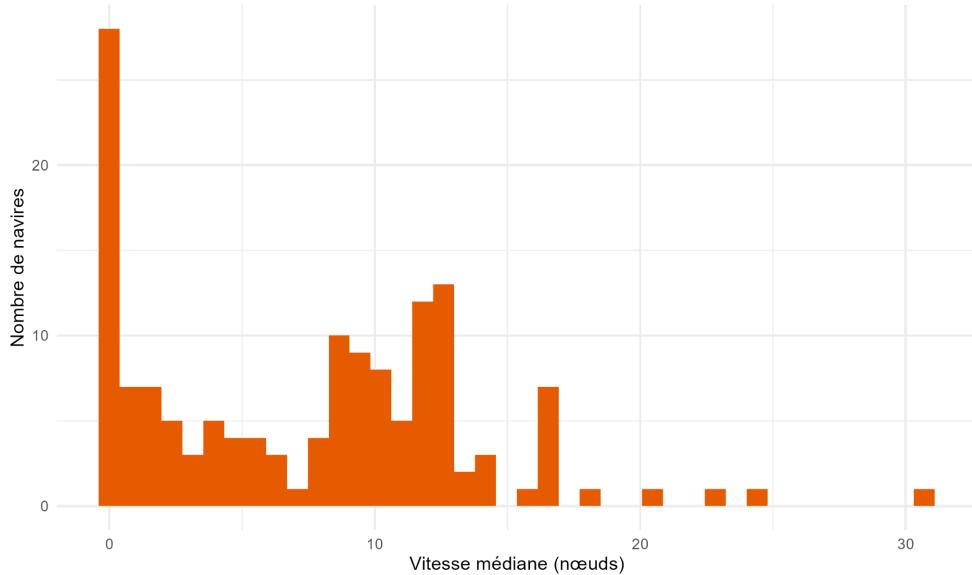
A : Représentation des navires



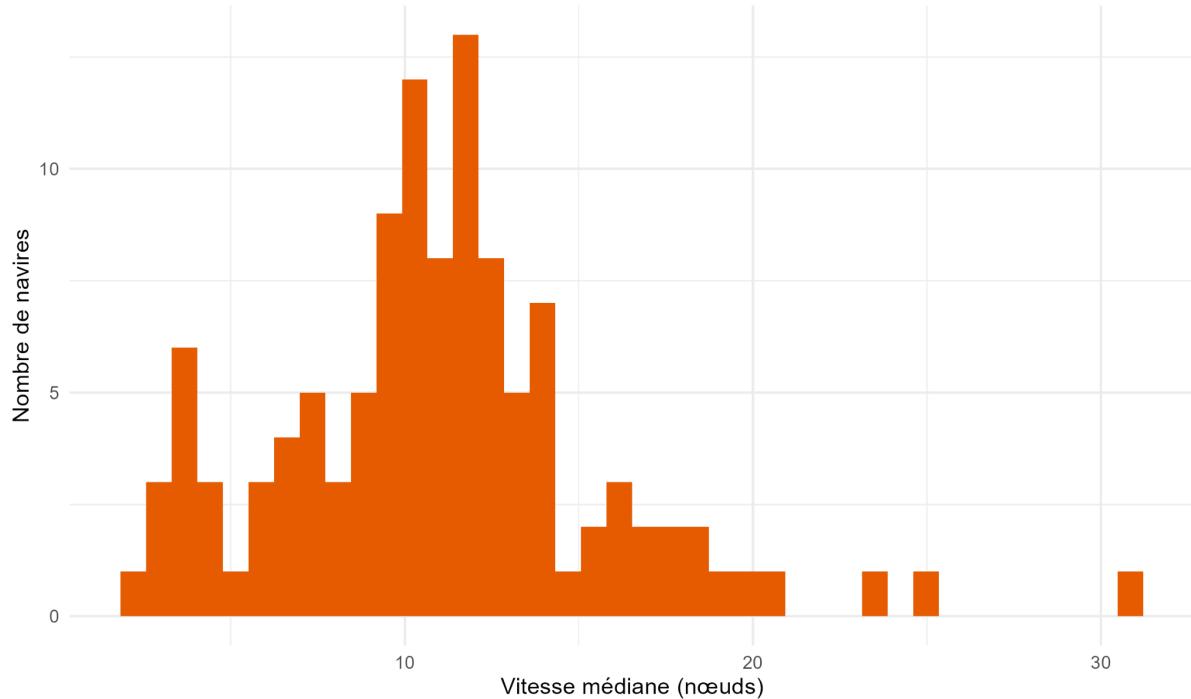
Pour le premier graphique, nous avons un dénombrement des types de bateaux, les trois premiers sont des bateaux de diverses classes, qui ne transportent pas de matériaux dangereux, avec la première catégorie étant les navires passagers, c'est-à-dire des navires qui majoritairement des personnes.

B : Médiane de vitesse

Vitesse médiane par navire

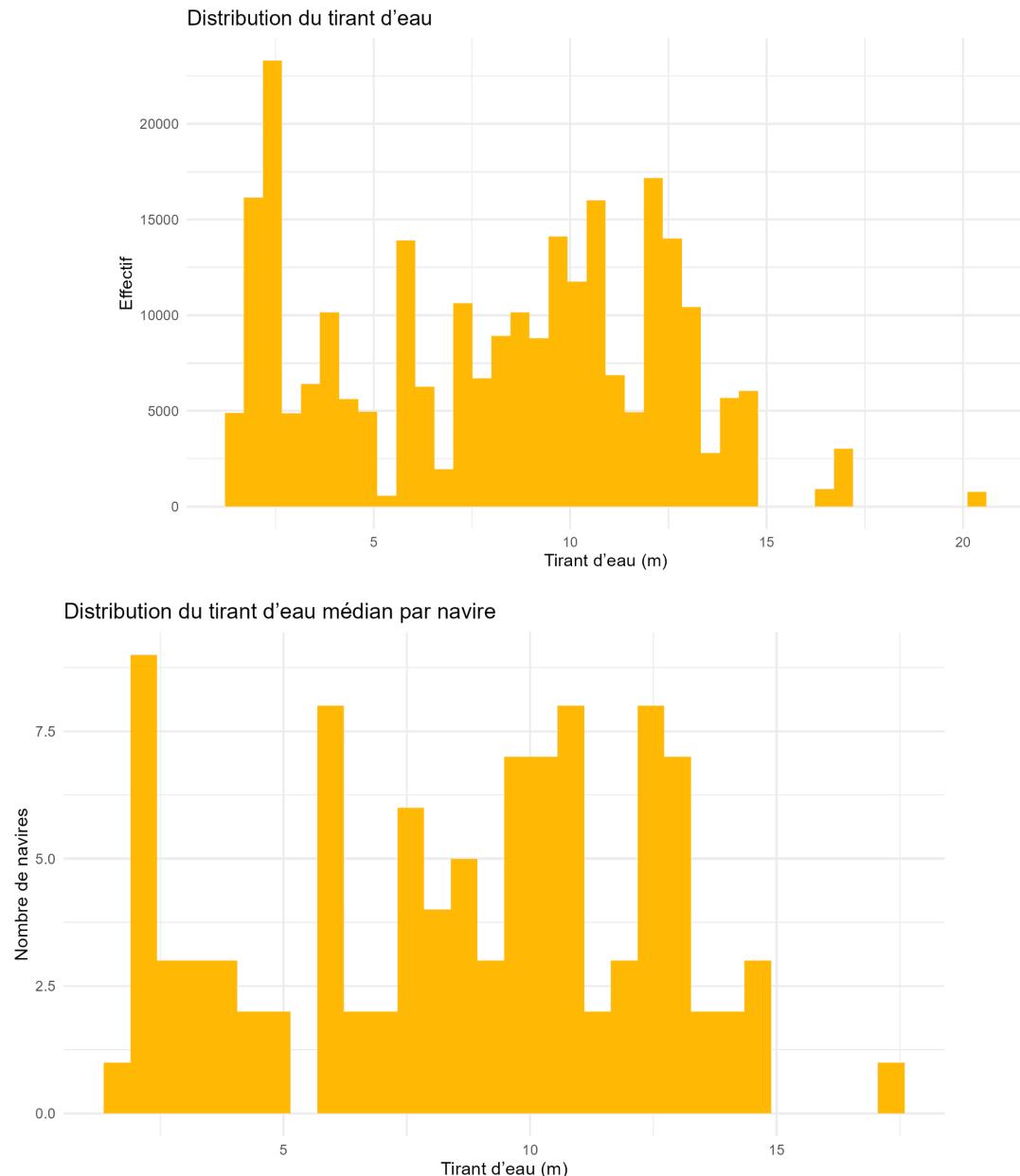


Vitesse médiane par navire



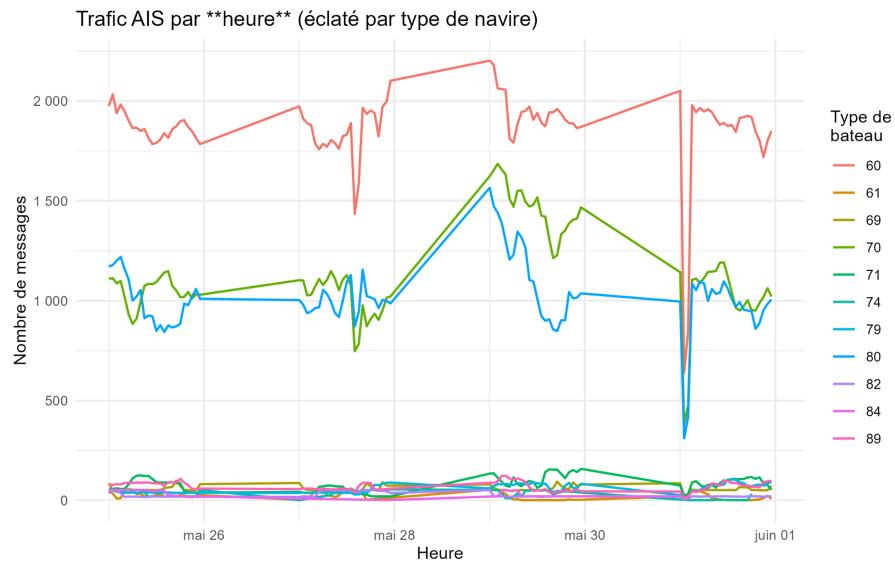
Voici un graphique corroborant l'effectif de bateaux et leur vitesse, nous pouvons voir deux choses la première la majorité des messages navires sont à quai et donc immobile, ce n'est pas dû à une très faible utilisation, cela montre simplement que plus les bateaux sont proches des balises de réception plus le message a des chances d'être correctement réceptionné, ce qui indique également que la plupart des messages sont perdus, mais pour que quelques-uns arrivent à destination, ils doivent en envoyer beaucoup.

C : Distribution de tirant d'eau

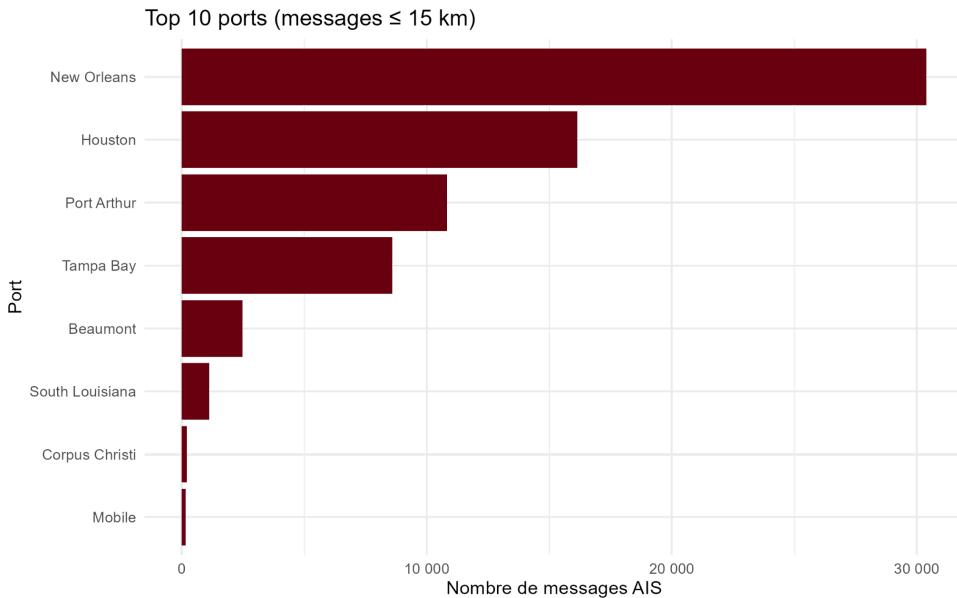


Voici un histogramme représentant l'effectif du tirant d'eau des navires sachant que le tirant d'eau moyen est de 8.12 m de tirant d'eau ce qui est à la fois plutôt grand et logique, d'une part la majorité des navires ne sont pas que des navires de plaisance, mais de transport de passager. Il y a également une différence entre les deux navires, qui est expliquée par la présence de cargo. La majorité des bateaux de plaisance, on peut de tirant d'eau, car il n'y a pas besoin de beaucoup de cargo, mais plus on augmente en tirant d'eau plus on charge le bateau.

D : Représentation au cours du temps des messages



Et pour le dernier graphique, le nombre de messages à l'heure, indiquant toutes les informations des navires, nous pouvons voir des pics qui peuvent être expliqués par des bugs dans la prise de donnée et des plats qui peuvent être expliqués par des blancs ou de la mise hors service du matériel de réception.



Et finalement un graphique sur l'ordre de préférence des ports, quels sont les ports les plus visités cependant il se peut que ce graphique relatif aux positions connues, comme nous ne connaissons pas assez la région, il se peut que nous n'ayons pas mentionné un port ou deux à l'image des ports de Floride.

Fonctionnalité 3 : Visualisation sur une carte

Trajectoires bateaux

Choisir un bateau :

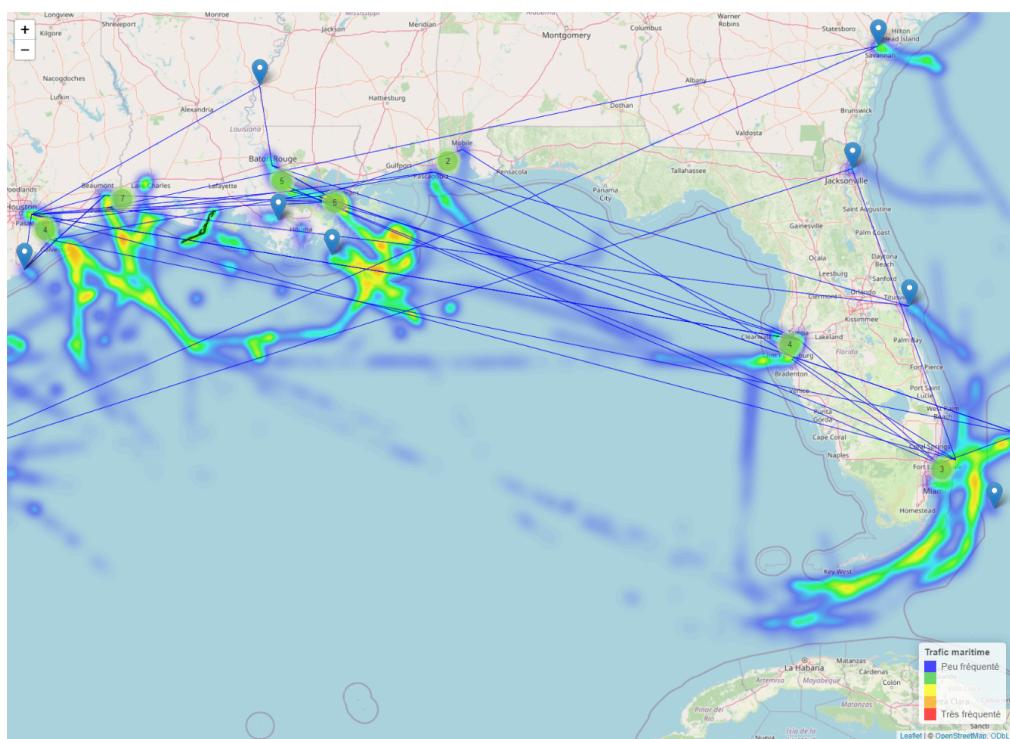
ADRI LAB

Afficher les routes fréquentées

Afficher les ports détectés

Afficher les liens entre les ports

MMSI	368112000
IMO	IMO8739023
CallSign	WDF2836
VesselType	60
Length	49
Width	9
Draft	NA
Cargo	60
TransceiverClass	A



L'objectif de cette fonctionnalité est double : permettre une exploration dynamique des trajectoires d'un navire, et détecter automatiquement les zones portuaires à partir des positions géographiques. Le tout est intégré dans une application web Shiny, interactive et accessible à l'utilisateur final.

A-Données et préparation

Les données utilisées proviennent d'un fichier CSV contenant des millions de points issus du système AIS : identifiants des navires, position GPS, heure, type de cargaison, statut, etc. Une fois chargées avec `readr`, ces données sont nettoyées et ordonnées avec `dplyr`. Les dates sont converties avec `lubridate` pour faciliter l'analyse temporelle.

Pour chaque navire sélectionné, les points GPS sont triés chronologiquement. Si un écart temporel important est détecté ($> 1\text{h}$), une interpolation linéaire ajoute des points intermédiaires toutes les 15 minutes afin d'obtenir une trajectoire fluide et continue.

B-Affichage des trajectoires

L'interface permet à l'utilisateur de choisir un bateau dans une liste. Une fois le choix fait, la trajectoire est affichée sur une carte `leaflet`. Celle-ci est automatiquement centrée sur la zone concernée. Un tableau résume aussi les informations principales du bateau : taille, type, identifiants, tirant d'eau, etc.

C-Analyse du trafic maritime

En option, l'utilisateur peut afficher une carte de chaleur représentant les zones les plus fréquentées. Cette heatmap repose sur `leaflet.extras` et met en évidence les couloirs maritimes et les zones de mouillage. Un dégradé de couleurs, avec légende, facilite l'interprétation.

D-Détection des ports

Une autre fonctionnalité clé est la détection des zones portuaires. Les points où le navire est à l'arrêt (`Status == 5`) sont isolés, puis groupés spatialement avec l'algorithme DBSCAN (librairie `dbscan`). Cela permet d'identifier automatiquement les ports à partir des données, sans liste préexistante.

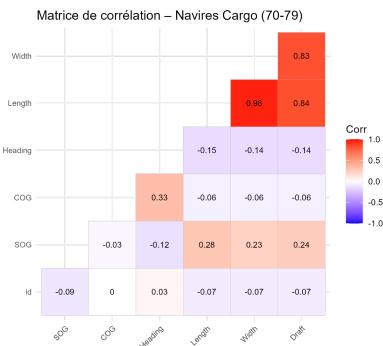
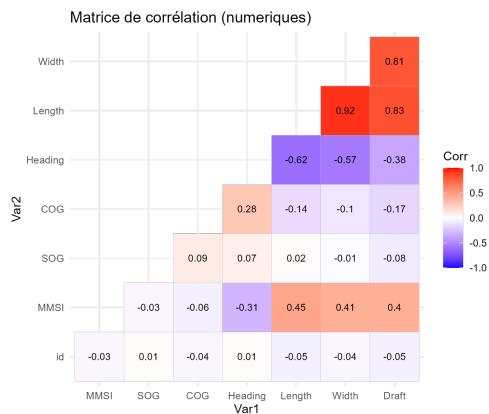
Chaque port détecté est représenté par un marqueur contenant le nombre de navires l'ayant fréquenté et la cargaison dominante. Cette approche permet d'enrichir la visualisation sans intervention manuelle.

E-Visualisation des flux portuaires

Enfin, on calcule les liaisons principales entre les ports. En retracant les mouvements des navires entre clusters DBSCAN successifs, on met en évidence les flux les plus courants. Seule la liaison la plus fréquente est affichée pour chaque port de départ, afin d'éviter la surcharge visuelle.

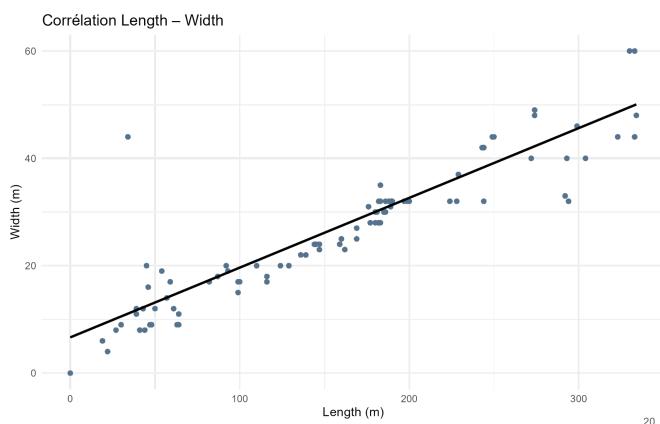
Fonctionnalité 4 : Étude des corrélations entre variables

A : Matrice de corrélation

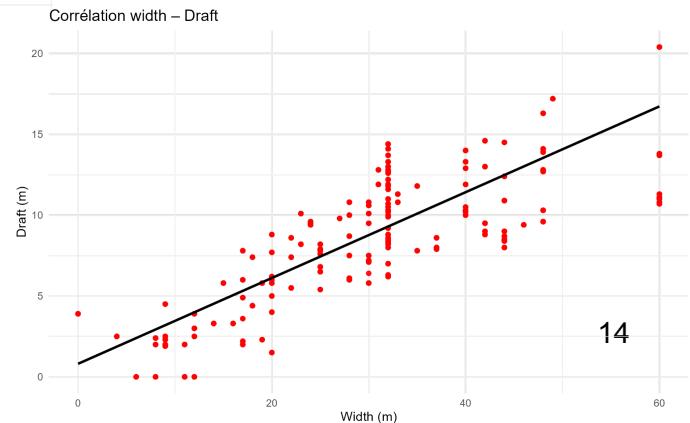


La matrice de corrélation représente la corrélation entre plusieurs données par exemple plus un bateau est long plus, il sera large, la corrélation physique a d'ailleurs beaucoup de relation, avec le tirant d'eau et la longueur ou largeur, et logiquement, il y a de faibles corrélations entre deux statistiques comme la direction et la longueur du bateau.

À titre de comparaison, Voici une matrice de corrélation sur un seul type ; les cargos



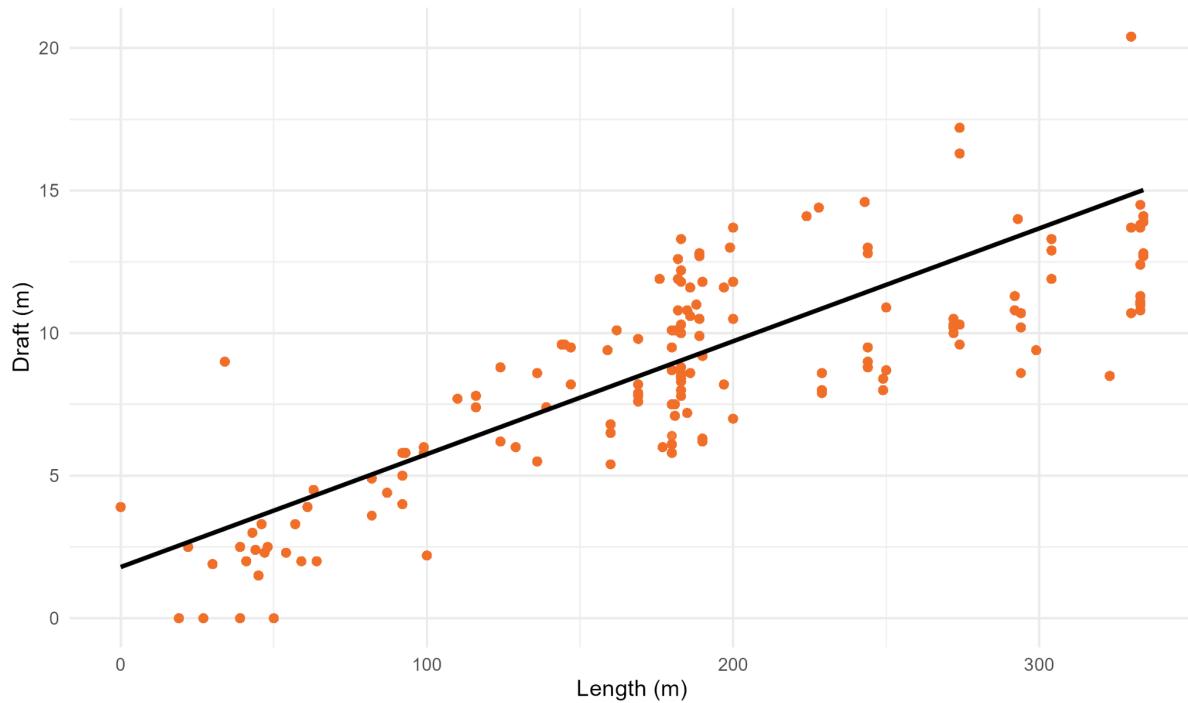
La forte corrélation est due à la forme des bateaux en eux-mêmes, la plupart ont besoin de respecter un ratio de 1:8 à 1:7.



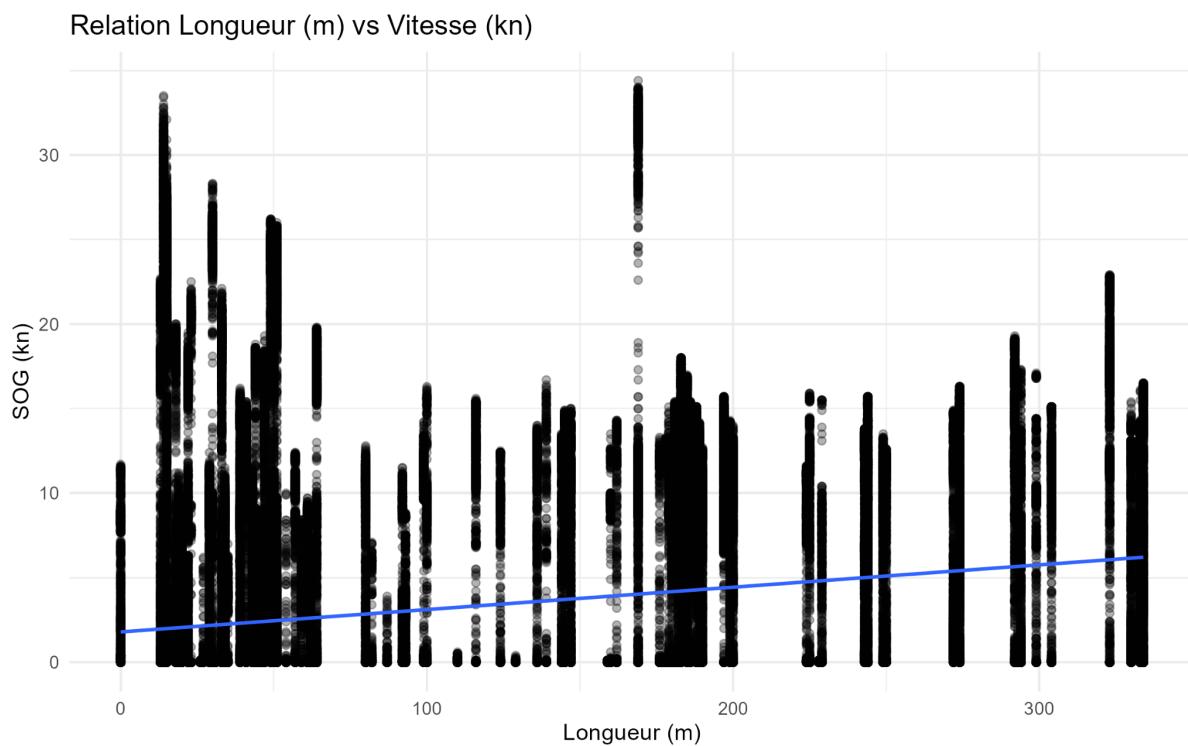
Le tirant d'eau croît avec la largeur du navire, mais l'étalement n'est pas très

étiré pour les cargos et les tankers, car certains voyagent avec ballasts et d'autres sans.

Corrélation Length – Draft

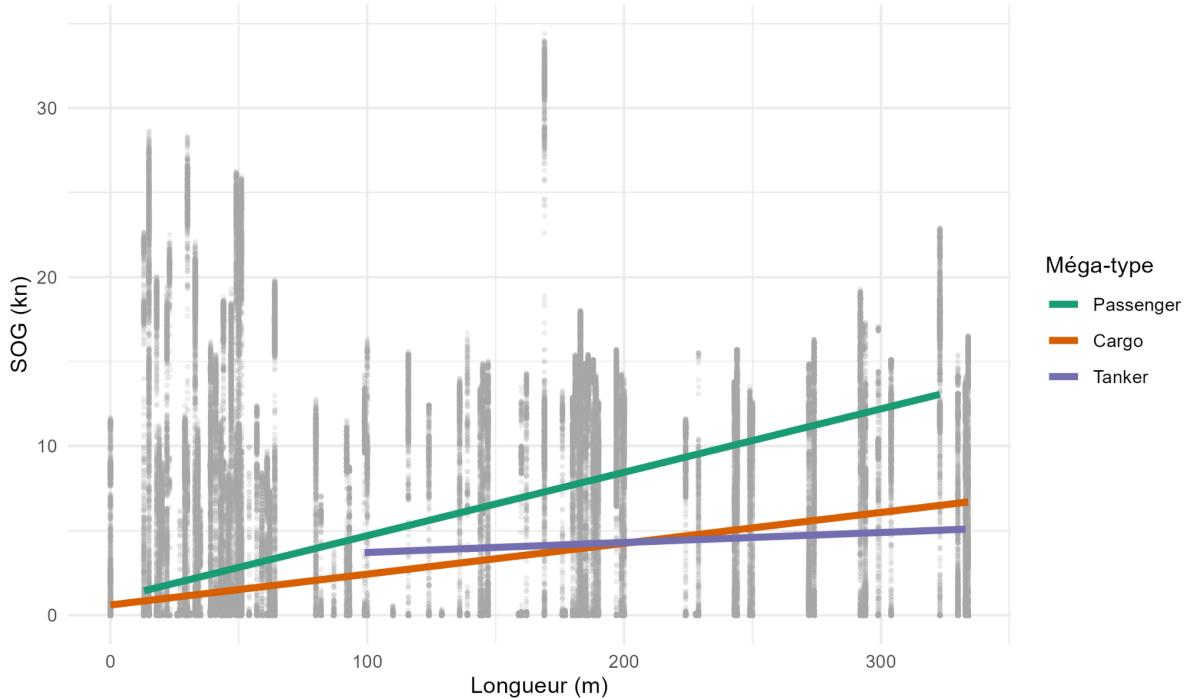


C : Graphique représentatif de la relation longueur - vitesse



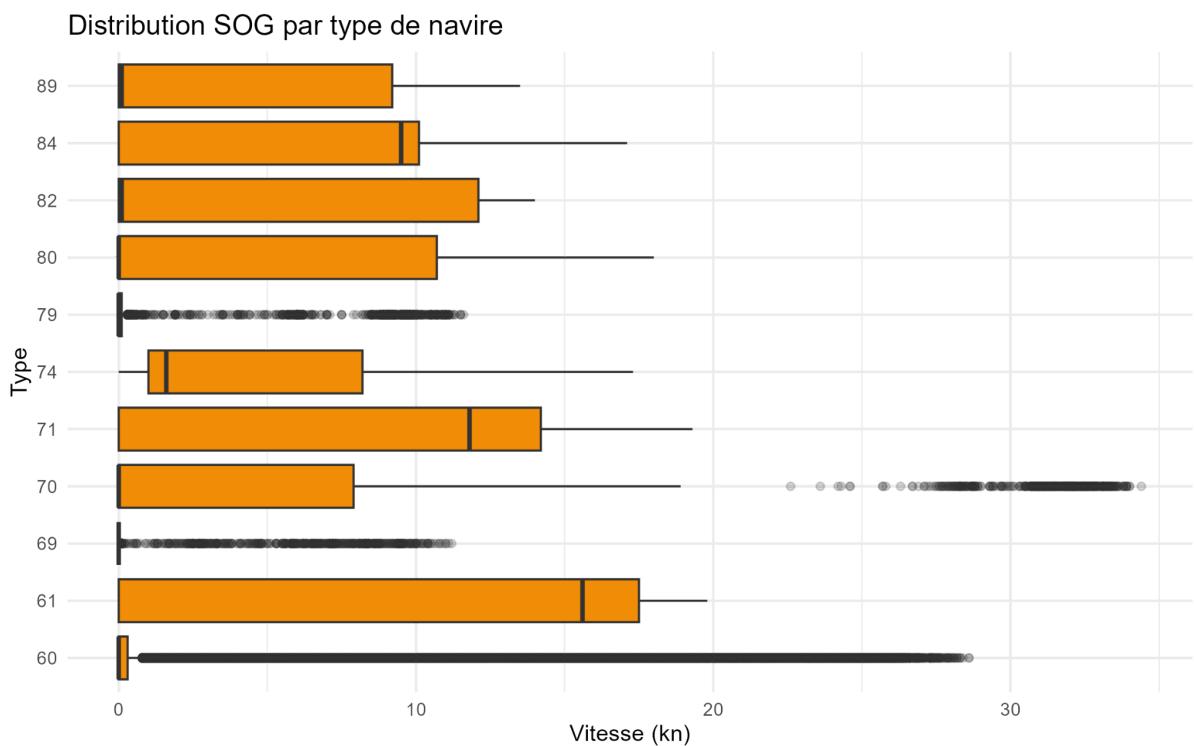
De manière générale, il ne semble pas y avoir de grandes différences entre la longueur des bateaux et, et leur vitesse (malgré quelques valeurs aberrantes).

Vitesse ~ Longueur : régression par méga-type (60-69 / 70-79 / 80-89)



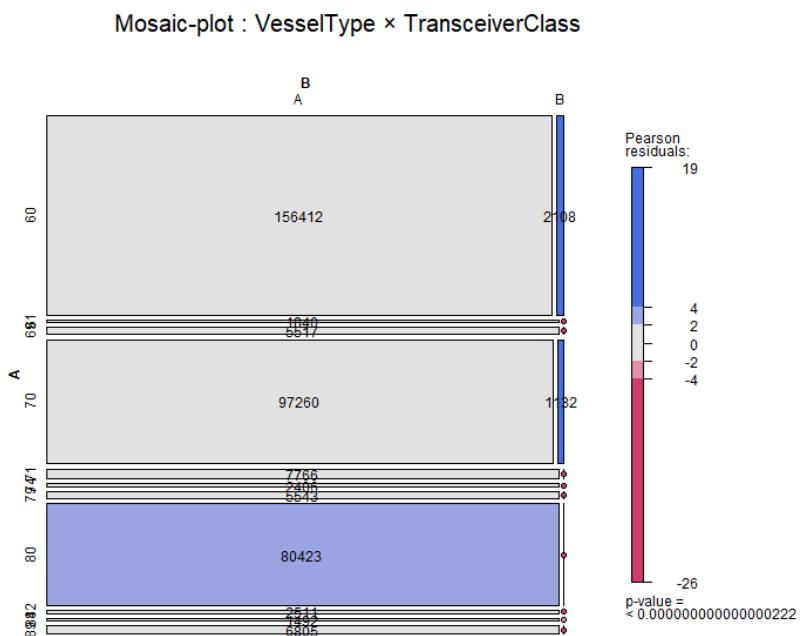
Afin d'y voir plus clair, nous avons décidé de traiter ce graphique en séparant les types. les navires passagers et cargos suivent une courbe simple plus un navire est gros plus il doit arriver à quai vite

Les navires tankers eux, suivent plutôt une ligne d'efficacité à 12 nœuds pour économiser le carburant.



Ce box-plot représente la disparité entre les types et les vitesses, les rectangles oranges sont les quartiles de vitesse, et le trait noir représente le deuxième quartile Q2 soit 50 % et les lignes noires représentent et sont les bornes maximales sans être de valeurs aberrantes les points gris représentent des vitesses anormales, par exemple dans la plage choisie la plupart des navires passagers sont à quai et donc ont une vitesse normale plutôt basse.

D : Mosaic plot ou l'étude de l'indépendance



Finalement le “mosaic plot”, d'une part chaque rectangle est un type de navire avec comme AIS de A ou B, plus un rectangle a de la hauteur plus des messages ont été envoyés par ces types de navires.

Le bleu ou le rouge est la différence par rapport à l'indépendance, est-ce qu'on attend ou pas à avoir cette statistique, si un rectangle est bleu l'effectif observé est bien plus grand que ce qu'on aurait sous indépendance, s'il est en rouge l'effectif est plus petit que l'on attendait.

La pvalue indique également que l'hypothèse où les deux variables sont indépendantes est formellement rejetée.

Fonctionnalité 5: comparaison de vitesse et prédictions

A : Prédictions vis à vis des caractéristique physiques, vitesse position ou cargo

```
#split 70 / 30
idx <- createDataPartition(ais_clf$vesselType, p = .7, list = FALSE)
train <- ais_clf[idx, ]
test <- ais_clf[-idx, ]
```

Pour cette partie, après avoir trié dans l'ordre nos données : nous allons mettre 70 % des données dans une liste pour prédire la variable vesseltype et le reste des 30 % pour tester nos capacités à prédire vesseltype

```
#Modèle multinomial
fit_mn <- nnet::multinom(
  VesselType ~ Length + Width + Draft + SOG + Slenderness + BlockCoeff,
  data = train, trace = FALSE, maxit = 400
)
```

Nous construisons notre modèle de prédiction avec la longueur, la largeur, la vitesse au sol, la minceur du navire (longueur sur largeur) et le coef block (tirant d'eau/longueur).

```
> print(conf$table)      # matrice de confusion
  Reference
Prediction   60    70    80
  60 15116  2341     0
  70   534 18567  8984
  80     0  7288 15142
```

Voici la matrice de confusion, en soi combien de navire notre modèle a confondu avec un autre, a environ ~90 % notre modèle à trouver les navires passagers, qui sont plus petits, mais les cargos et les tankers sont confondus à ~63 % chacun, cela est dû au fait qu'ils sont tous deux grand et long, cependant la minceur et le blockcoef aide à différencier les deux.

Ce modèle n'est donc pas parfait, nous pourrions plus simplement prendre la cargaison pour mieux les déterminer et les catégories de dangers.

```
> print(conf$overall)  # Accuracy, Kappa
  Accuracy       Kappa  AccuracyLower  AccuracyUpper  AccuracyNull AccuracyPValue McNemarPValue
  0.7183105    0.5681894    0.7149122    0.7216901    0.4148179    0.0000000          NaN
```

En conséquence, 70 % de nos prédictions sont correctes, et selon le kappa, nous dépassons de manière évidente le hasard, mais ce n'est pas parfait encore, notamment, car nous ne dépassons pas le 0.6.

```
| + VesselType ~ LAT + LON + Cargo,
```

```

> print(conf$table)      # matrice de confusion
   Reference
Prediction Cargo Passenger Tanker
Cargo     14472      1158      0
Passenger  247       4868      0
Tanker    2448       0      14025
> print(conf$overall)  # Accuracy, Kappa, etc.
  Accuracy      Kappa AccuracyLower AccuracyUpper AccuracyNull AccuracyPValue McNemarPValue
  0.8964748  0.8322801  0.8933354  0.8995523  0.4612553  0.0000000           NaN
>

```

Ici, nous prenons en compte la latitude, la longitude et le cargo d'un navire pour entraîner le modèle et nous pouvons voir qu'il reconnaît à 89 % les navires.

+ **VesselType ~ Length + width + Draft + SOG + LAT + LON + Cargo,**

```

> print(conf$table)      # matrice de confusion
   Reference
Prediction Cargo Passenger Tanker
Cargo     15529      326      512
Passenger  800       5700      0
Tanker    838       0      13513
> print(conf$overall)  # Accuracy, Kappa, etc.
  Accuracy      Kappa AccuracyLower AccuracyUpper AccuracyNull AccuracyPValue McNemarPValue
  0.9334731  0.8933139  0.9308939  0.9359843  0.4612553  0.0000000           NaN
>

```

Ici, nous avons un modèle de régression linéaire qui est entraîné avec les données précédentes, même si nous n'avons pas beaucoup plus de précision, nous en gagnons un peu plus.

B : Calcul de vitesse

```

+++ Tanker Speed Model ***
RMSE : 5.83 kn
MAE : 5.35 kn

```

Nous avons prédit l'écart moyen de vitesse des tankers, cependant elle n'est pas à notre goût ; car elle atteint 50 % de la vitesse normale des tankers soit 12 noeuds, donc notre modèle linéaire n'explique qu'une fraction du comportement.

Pour expliquer cette erreur, nous pouvons l'expliquer avec le tirant d'eau, le tirant d'eau indique la profondeur immersion, mais pas la charge, de plus certain tanker navigue à ballast pour réduire le tirant d'eau ce qui fausse nos calculs.

MMSI	time_start	time_end	SOG_reported	speed_calc	diff_kn	
<chr>	<dttm>	<dttm>	<dbl>	<dbl>	<dbl>	
1	228075700	2023-05-29 12:59:55	2023-05-29 13:00:05	13.2	13.2	0.04
2	305291000	2023-05-29 20:59:55	2023-05-29 21:00:05	7.1	7.69	0.59
3	367518920	2023-05-27 13:59:56	2023-05-27 14:00:01	14.9	14.8	-0.08
4	369390000	2023-05-31 16:59:57	2023-05-31 17:00:02	15.2	18.8	3.57
5	371823000	2023-05-31 16:59:51	2023-05-31 17:00:02	11.7	11.7	0.03
6	477948700	2023-05-29 00:59:55	2023-05-29 01:00:06	9.3	8.39	-0.91
7	538008309	2023-05-25 07:59:55	2023-05-25 08:00:04	10	10.1	0.09
8	636017514	2023-05-25 01:59:54	2023-05-25 02:00:03	1.6	1.7	0.1
9	636021151	2023-05-29 02:59:48	2023-05-29 03:00:08	13	13.0	-0.02
10	636022111	2023-05-25 15:59:56	2023-05-25 16:00:05	12.2	13.6	1.45

Et finalement voici 10 navires pris au hasard, auquel nous avons calculer une vitesse théorique en prenant la position entre deux messages et leur date pour calculer leur vitesse, nous avons ensuite comparé cette vitesse a la vitesse SOG pour connaître la fiabilité de cette méthode

Conclusion

Cette première semaine a permis de structurer efficacement le projet et de livrer une première version fonctionnelle de l'application. Les traitements de base sont opérationnels : exploration intuitive des données, détection des zones portuaires, visualisation des mouvements, et déploiement de premiers modèles prédictifs. Ces derniers affichent une précision globalement satisfaisante, avoisinant 70 %, avec des pics allant jusqu'à 93 % pour certaines catégories de navires. Néanmoins, des marges d'amélioration subsistent, notamment dans la différenciation fine entre cargos et tankers, ainsi que dans la prise en compte de facteurs exogènes. Les prochaines semaines seront consacrées à l'enrichissement des modèles, à l'intégration de nouvelles variables (ex. : conditions météo, état de charge), et à l'amélioration de l'interface utilisateur pour un usage fluide et robuste.