ISEN Brest

Année 2024-2025

Rapport projet Big Data :

Nolan Jauffrit Célian Bosser Nolan Nedelec

Du 2 au 6 juin 2025

Introduction générale

Le transport maritime joue un rôle essentiel dans les échanges mondiaux, et son suivi en temps réel est devenu indispensable pour des raisons de sécurité, d'optimisation logistique et de surveillance environnementale. Le système AIS (*Automatic Identification System*) permet de collecter automatiquement des données précises sur les navires : position, vitesse, dimensions, type, statut, etc.

Dans le cadre du projet A3 – Développement Big Data, nous avons mené une étude approfondie des données AIS issues de la zone maritime de Brest. Ce projet s'inscrit dans une démarche complète d'**analyse, de visualisation et de modélisation des comportements de navires**, avec une ouverture vers des applications d'intelligence artificielle.

Le rapport est structuré selon les étapes suivantes :

- **Présentation des variables du jeu de données AIS**
  Nous détaillons les différentes variables disponibles (statut, type, vitesse, dimensions…), leur signification, leur typologie (numérique, catégorielle) et leur intérêt pour l'analyse.
- **Fonctionnalité 1 — Présentation de graphiques avant et après filtrage**
  Cette section met en évidence l'impact des valeurs aberrantes à travers des visualisations comparatives des principales variables (vitesse, longueur, largeur, tirant d'eau). Le nettoyage des données s'avère indispensable pour éviter les biais statistiques.
- **Fonctionnalité 3 — Analyse des données AIS : trajectoires, routes maritimes et typologie**
  Nous explorons la dynamique spatiale du trafic maritime, en visualisant les trajectoires des navires et en repérant les routes principales. Nous posons également les premières bases de prédiction du type de navire selon ses caractéristiques observées.
- **Fonctionnalité 4 — Analyse des corrélations entre variables**
  Une étude statistique bivariée nous permet d'identifier les relations linéaires ou non entre les variables (dimensions, vitesse, type, statut…). Cela nous guide dans la sélection des variables explicatives les plus pertinentes pour la modélisation.
- **Fonctionnalité 5 — Prédiction du type de navire et estimation indirecte de la vitesse**
  Enfin, nous proposons un début de modélisation supervisée à travers une régression linéaire et une préparation du jeu de données pour la suite du projet en intelligence artificielle. L'objectif est de prédire automatiquement le type de navire ou de déduire certaines informations manquantes comme la vitesse.

À travers cette démarche, nous montrons comment les outils de la data science peuvent être appliqués à des données complexes et massives comme celles issues de l'AIS, en combinant rigueur analytique, visualisation intuitive et préparation à la modélisation prédictive.

Présentation des variables du jeu de données AIS

Nous présentons ici les différentes variables contenues dans la base de données AIS utilisée pour notre étude. Chacune joue un rôle essentiel dans la compréhension, la visualisation ou la modélisation des comportements des navires.

id *(entier)*
Identifiant interne unique pour chaque enregistrement. Il ne joue pas de rôle analytique mais garantit l'unicité de chaque ligne du fichier.

MMSI (Maritime Mobile Service Identity) *(chaîne de caractères, 9 caractères)*
C'est un identifiant international unique à neuf chiffres, utilisé dans les communications AIS pour identifier un navire, une station côtière ou une balise. En France, il est attribué par l'ANFR.
Dans notre projet, il permet :
- de suivre les trajectoires des navires
- de regrouper les données d'un même navire dans le temps
- de mener des analyses comportementales

BaseDateTime *(datetime)*
Date et heure de transmission du message AIS. Permet de reconstituer l'évolution temporelle de chaque navire et d'ordonner les observations.

LAT / LON (Latitude / Longitude) *(float, 8 chiffres)*
Coordonnées géographiques du navire au moment de l'émission.
LAT varie entre -90 et +90 ; LON entre -180 et +180.
Utilisées pour la visualisation cartographique et l'analyse spatiale.

SOG (Speed Over Ground) *(float, 8 chiffres)*
Vitesse réelle du navire sur le fond marin, en nœuds nautiques. Prise en compte de l'influence du vent, des courants et des marées.

COG (Course Over Ground) *(float, 4 chiffres)*
Direction réelle de déplacement sur le fond terrestre, en degrés (0 à 360). Peut différer du Heading en cas de dérive.

Heading *(float, 4 chiffres)*
Orientation de la proue du navire en degrés. Utile pour l'étude des manœuvres ou rotations. Différent du COG en cas de courant latéral.

VesselName *(chaîne de caractères, max 32 caractères)*
Nom du navire déclaré dans les messages AIS (jusqu'à 32 caractères).

IMO *(chaîne de caractères, 7 caractères)*
Numéro d'identification international unique à 7 chiffres, constant pendant toute la vie du navire (même si changement de nom). Très fiable pour l'identification.

CallSign *(chaîne de caractères, max 8 caractères)*
Indicatif radio du navire. Unique, délivré par le pays d'immatriculation. Utilisé comme identifiant secondaire.

VesselType *(entier, varie de 0 à 99)*
Code numérique (0 à 99) décrivant la catégorie du navire. Exemple de regroupement :
60–69 : Passagers
70–79 : Cargos
80–89 : Pétroliers
C'est la variable cible pour les tâches de classification.

Status *(entier, varie de 0 à 15)*
État de navigation déclaré : en route, mouillage, échoué, etc. Peut-être ordonné ou recodé numériquement pour les analyses

Length / Width / Draft *(float, 4 chiffres)*
Dimensions physiques du navire exprimées en mètres :
Length : Longueur d'un bateau
Width : Largeur d'un bateau
Draft : Tirant d'eau (profondeur immergée)
Très utiles pour estimer le type, la capacité ou la manœuvrabilité du navire.

Cargo *(chaine de caractère, max 4 caractères, code de 0 à 99 comme vessel type)*
Code numérique représentant le type de cargaison transportée. Complète l'information sur le type de navire

TransceiverClass *(chaine de caractère, 2 caractères)*
Classe d'équipement AIS :
A = navires de commerce ou professionnels (doivent fournir plus d'informations)
B = petits navires ou plaisance (moins complet dans la base de donnée)

Fonctionnalité 1 Description et exploration des données

1. Prétraitement et filtrage des données :

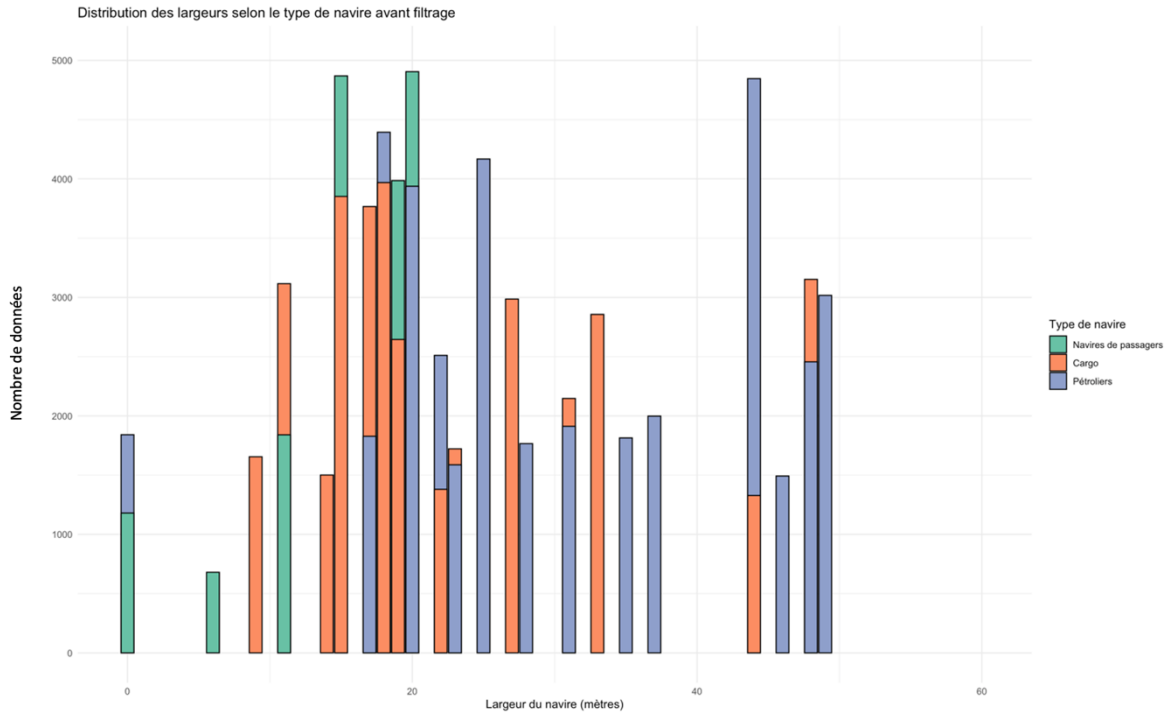
Avant toute analyse, un important travail de nettoyage a été réalisé sur le jeu de données AIS afin de garantir la qualité et la fiabilité des résultats. Ce filtrage a permis d'éliminer les valeurs aberrantes, incohérentes ou manquantes, et de restreindre l'analyse à des observations pertinentes.

Plusieurs critères ont été appliqués :

- Nettoyage des variables numériques : les vitesses ont été limitées à une plage réaliste (entre 0 et 50 nœuds), et les dimensions physiques des navires ont été contrôlées. Les longueurs inférieures à 2 mètres ou supérieures à 400 mètres, ainsi que les largeurs inférieures à 1 mètre ou supérieures à 70 mètres, ont été considérées comme aberrantes et supprimées.
- Filtrage géographique : seules les positions situées dans une zone cohérente avec le golfe du Mexique ont été conservées (latitude entre 18° et 31°, longitude entre -97° et -78°).
- Contrôle de la qualité des transpondeurs : les navires de classe A sans identifiant IMO ou avec un tirant d'eau hors plage réaliste (inférieur à 0.5 mètre ou supérieur à 28.5 mètres) ont été exclus.
- Nettoyage des statuts incohérents : certaines combinaisons statut/vitesse ont été supprimées (par exemple, un navire déclaré "en route" avec une vitesse nulle, ou "amarré" avec une vitesse positive). De plus, les statuts non définis ou réservés à des usages spécifiques (balises, tests, etc.) ont été écartés.

Ce filtrage rigoureux permet de travailler sur un sous-ensemble de données cohérent, représentatif et exploitable, en éliminant les biais potentiels liés à des erreurs de mesure ou à des enregistrements incomplets. Il constitue une étape essentielle pour garantir la robustesse des analyses statistiques et des modèles prédictifs développés par la suite.

2. **<u>Présentation de graphiques avant et après filtrage:</u>**



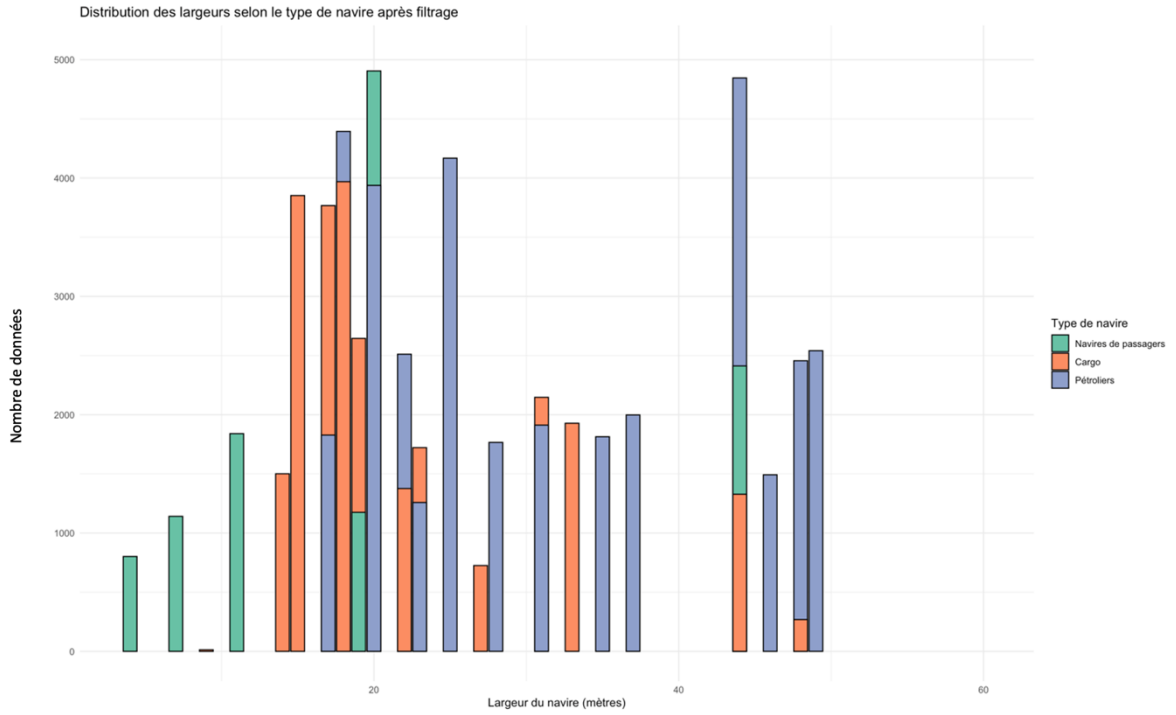
Graphique 1 : *barplot largeur* vs type de navire avant filtrage

Dans ce premier graphique on peut observer la quantité de données recueillit sur les largeurs en fonction du type de navire. Ce graphique nous présente des données avant filtrage. Sans filtrage nous pouvons observer des valeurs aberrantes égales a 0m. notamment des navires de passagers et pétroliers. En gardant ses données on va amener des erreurs dans nos futurs calculs Ce qui nous montre la nécessité de filtrer nos données

Ce graphique représente la distribution des largeurs des navires en fonction de leur type (navires de passagers, cargos, pétroliers), avant toute opération de filtrage. On observe ici la quantité de données disponibles pour chaque largeur arrondie (en mètres), avec des barres empilées selon le type de navire. On peut en tirer quelques constats :

- Des valeurs aberrantes apparaissent clairement, notamment des largeurs égales à 0 mètre, ce qui est physiquement impossible pour un navire.
- Ces erreurs concernent les navires de passagers que les pétroliers.
- Les cargos semblent plus régulièrement répartis, mais certains pics peuvent aussi traduire des anomalies de saisie ou des doublons.

Ce graphique met donc en évidence la nécessité de filtrer et de nettoyer les données, afin d'éviter des biais statistiques dans les calculs de moyenne, médiane, etc. Des erreurs dans les modèles prédictifs ou les visualisations futures.



Graphique 2 : *barplot largeur* vs type de navire après filtrage

Ce deuxième graphique montre la distribution des largeurs des navires après avoir appliqué un filtrage sur les valeurs aberrantes (par exemple, suppression des largeurs égales à zéro ou trop extrêmes).

Contrairement au graphique précédent, les données sont ici nettoyées, ce qui se traduit par :

- La disparition des barres à 0 mètre, ce qui élimine les erreurs manifestes de saisie ou de capteur ;
- Une répartition plus cohérente des navires par type : les cargos sont concentrés autour de 20–30 mètres, les pétroliers vers 30–50 mètres, et les navires de passagers présentent une variabilité plus large.

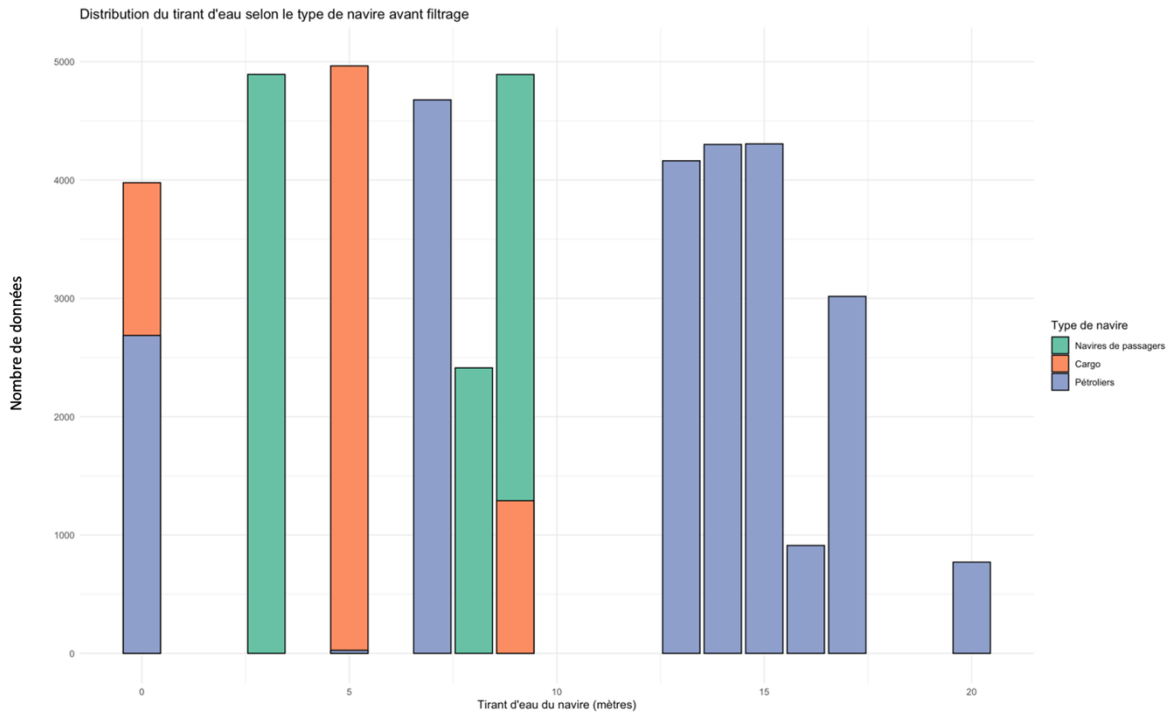
Grâce à ce nettoyage :

- La qualité des données est améliorée,
- Les statistiques descriptives (moyenne, médiane, écart-type) seront plus fiables,
- et les visualisations futures ne seront plus biaisées par des erreurs techniques.

Ce graphique valide donc l'étape indispensable de filtrage préalable avant toute modélisation ou inférence statistique.

Étude du tirant d'eau selon le type de navire – Avant et après filtrage

Les deux graphiques ci-dessus comparent la distribution du tirant d'eau des navires selon leur type, avant puis après nettoyage des données.



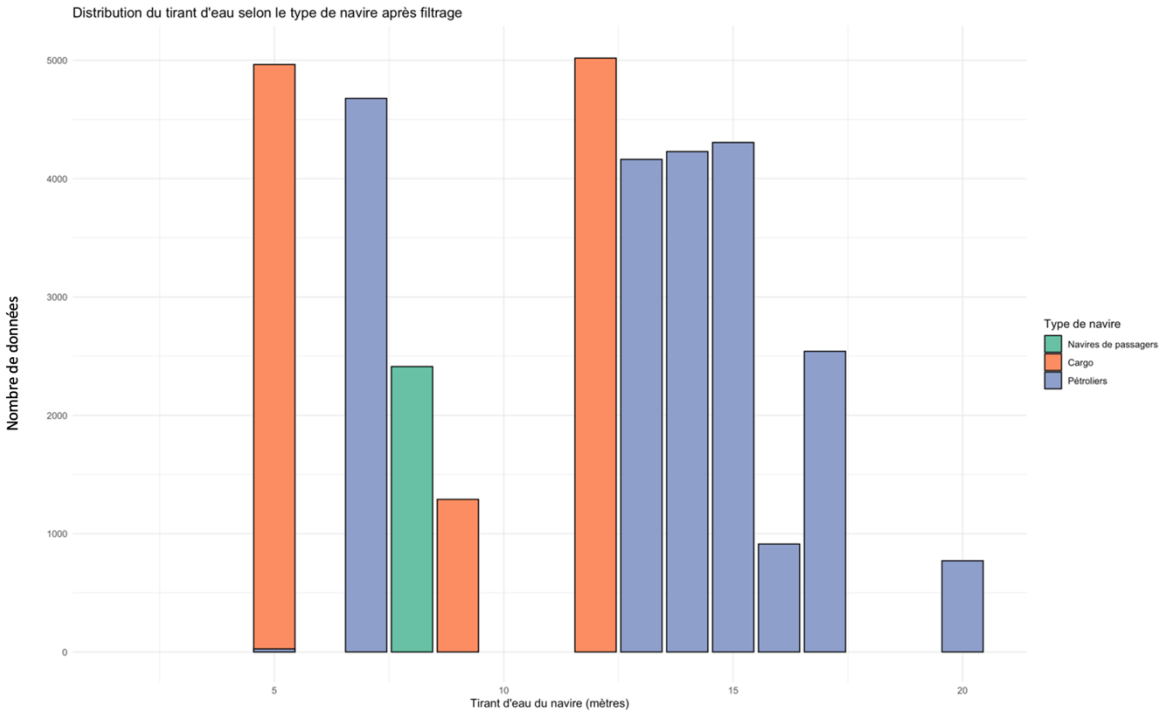
Graphique 3 : *barplot tirant d'eau* vs type de navire avant filtrage

Avant filtrage

Le graphique du bas révèle des valeurs aberrantes importantes :

- De nombreux navires apparaissent avec un tirant d'eau nul ou très faible (0 mètre ou proche), ce qui est incompatible avec la réalité physique d'un navire en mer.

- Ces anomalies concernent toutes les catégories, y compris les pétroliers, qui devraient logiquement afficher des tirants d'eau élevés.
- La distribution est écrasée vers les valeurs basses, rendant l'interprétation difficile et peu fiable.



Graphique 4 : *barplot tirant d'eau* vs type de navire après filtrage

Après filtrage

Le graphique du haut montre une nette amélioration :

- Les valeurs aberrantes ont été supprimées, recentrant la distribution sur des valeurs cohérentes :
    - Environ 7–10 mètres pour les cargos et navires de passagers.
    - Entre 10 et 15 mètres (et plus) pour les pétroliers, ce qui est réaliste.
- La forme des distributions devient exploitable : on observe des pics clairs, et des zones cohérentes d'activité selon les types de navires.
- Ce nettoyage améliore non seulement la visualisation, mais surtout la fiabilité des analyses statistiques futures (corrélations, modélisations…).

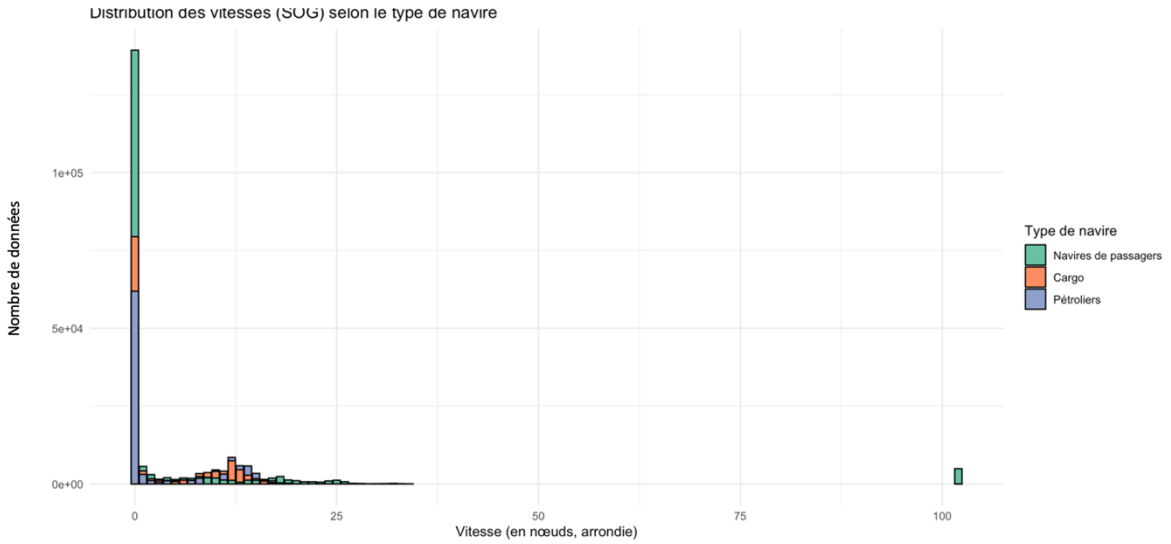
Conclusion

L'impact du filtrage est ici particulièrement visible : il corrige des données techniquement impossibles tout en clarifiant les profils caractéristiques de chaque catégorie de navire. Cette

étape valide l'importance d'un prétraitement rigoureux avant toute exploitation des variables techniques comme Draft.

Analyse de la distribution des vitesses (SOG) par type de navire – Avant et après filtrage

Les deux graphiques ci-dessus comparent la répartition des vitesses (Speed Over Ground) pour les navires, regroupés en trois grandes catégories : navires de passagers, cargos et pétroliers.



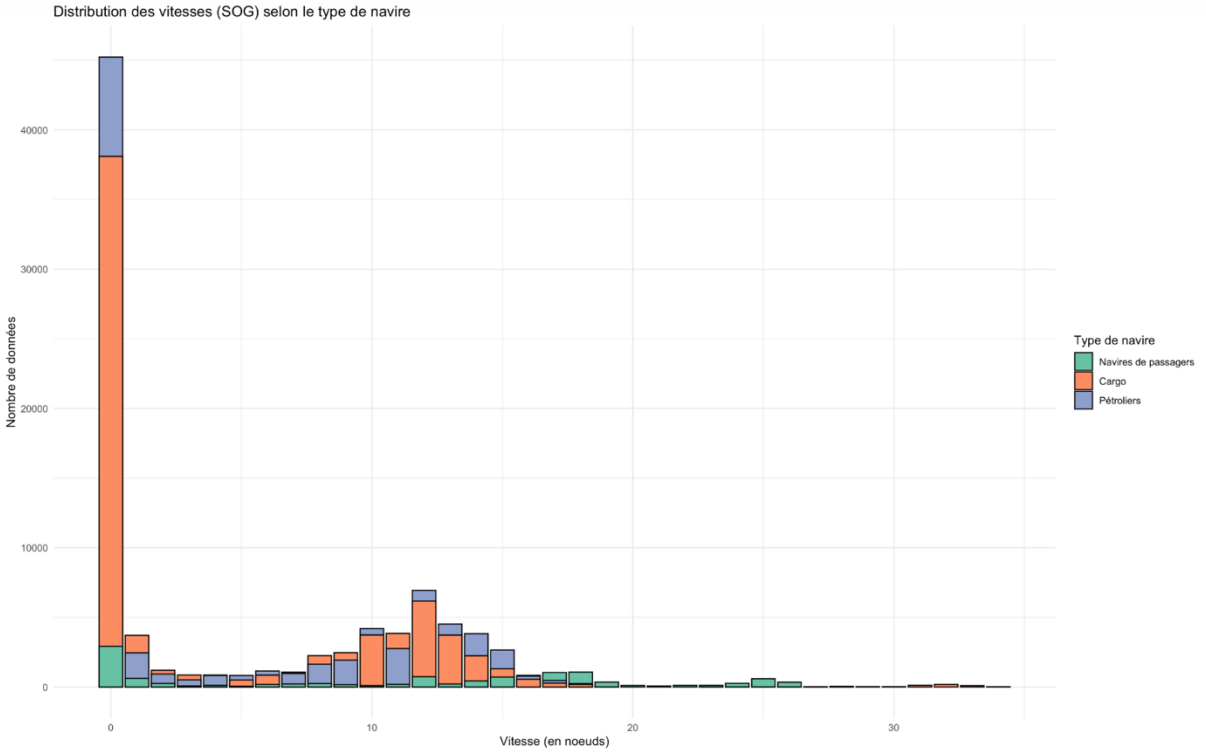
Graphique 5 : *barplot vitesse en nœuds* vs type de navire avant filtrage

Avant filtrage

Le graphique du haut montre une concentration extrême des données autour de 0 nœud**,** accompagnée de quelques pics totalement aberrants, notamment :

- Des vitesses supérieures à 50 voire 100 nœuds, qui sont physiquement irréalistes pour tout type de navire,
- Une échelle verticale déformée par des dizaines de milliers de valeurs nulles ou proches de 0,
- Une lisibilité très faible du comportement normal des navires.

Cette situation empêche toute analyse sérieuse sans traitement préalable, et indique la présence massive de valeurs erronées ou inutilisables.



Graphique 6 : *barplot vitesse en nœuds* vs type de navire après filtrage

Après filtrage

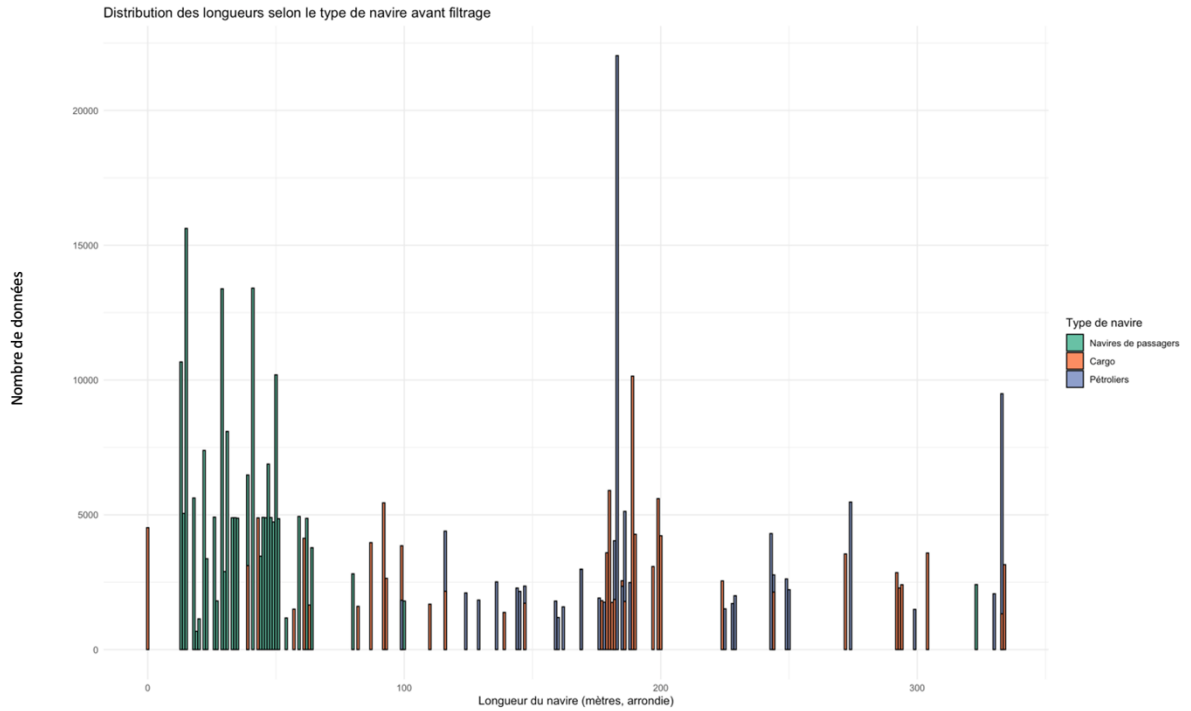
Le graphique du bas présente une distribution beaucoup plus informative, où :

- Les vitesses aberrantes ont été supprimées (plafond typique autour de 30–35 nœuds),
- La concentration excessive à 0 nœud a été réduite (notamment via un filtre SOG ≥ 1),
- Les profils deviennent visibles :
    - Les cargos et pétroliers se concentrent autour de 10 à 15 nœuds,
    - Les navires de passagers montrent une plus grande variabilité, pouvant dépasser les 25 nœuds.

Ce nettoyage rend les comparaisons entre types de navires visuellement interprétables et prépare le terrain à une analyse statistique ou une modélisation fiable.

Analyse de la distribution des longueurs (Length) selon le type de navire – Avant et après filtrage

Ces deux graphiques présentent la répartition des longueurs des navires, regroupés par catégorie (passagers, cargos, pétroliers), avant (bas) et après (haut) traitement des données.



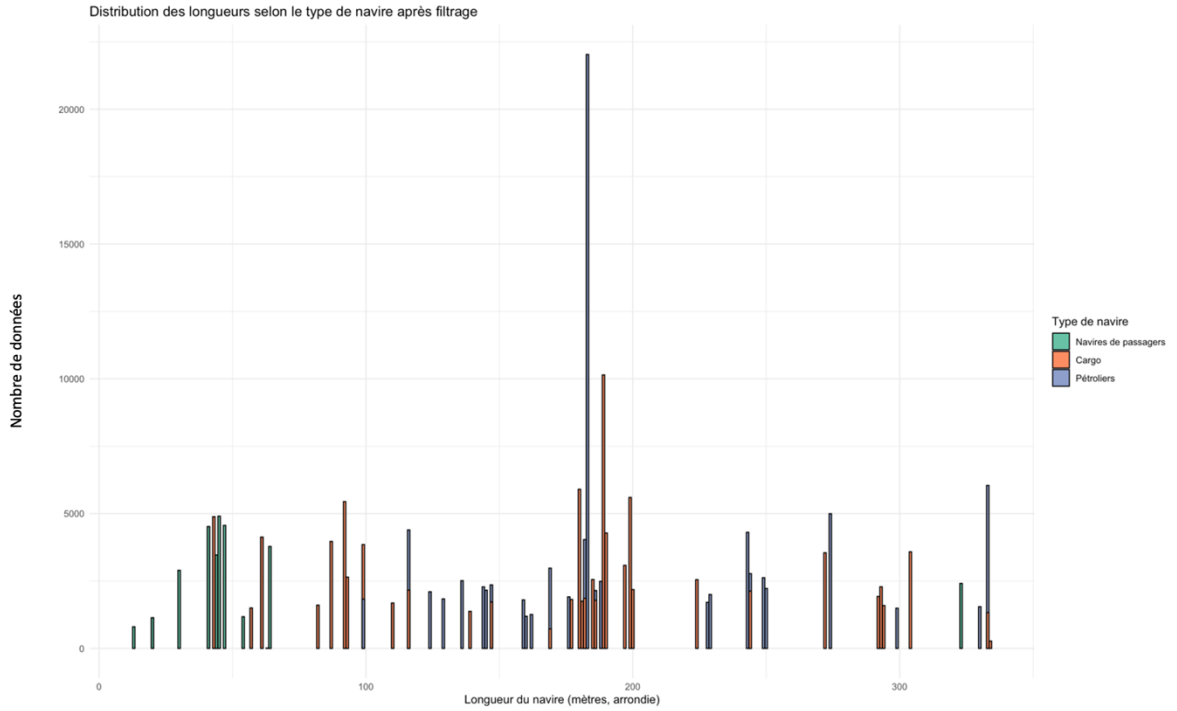
Graphique 7 : *barplot longueur* vs type de navire avant filtrage

Avant filtrage

Avant nettoyage, on observe plusieurs problèmes :

- Une forte concentration de valeurs très faibles (inférieures à 50 mètres), en particulier chez les navires de passagers.
- Une répartition désordonnée avec de nombreux pics isolés, reflétant des erreurs ou des doublons.
- Des valeurs incohérentes ou exagérément élevées, visibles dans les extrémités de l'axe (jusqu'à 400 mètres voire plus), qui écrasent la lisibilité générale.

Cette situation rend difficile toute interprétation fiable : les données sont bruitées, non représentatives, et potentiellement biaisées.



Graphique 8 : *barplot longueur* vs type de navire après filtrage

Après filtrage

Le graphique filtré améliore nettement la lisibilité :

- Les longueurs extrêmes ou nulles n'ont été supprimées, ce qui réduit les aberrations visuelles.
- Les pics les plus significatifs restent visibles, mais avec une meilleure distinction des types de navires :
    - Les navires de passagers sont majoritairement sous 200 m.
    - Les cargos se répartissent largement entre 100 et 250 m.
    - Les pétroliers dominent dans les grandes longueurs (jusqu'à 300 m).
- La forme globale devient cohérente avec ce que l'on attend techniquement des différents types de navires.

Conclusion

Cette comparaison montre que le filtrage des longueurs permet non seulement d'éliminer les erreurs mais aussi de révéler des structures typiques propres à chaque catégorie de navire. Il s'agit donc d'un prétraitement indispensable pour toute analyse sérieuse basée sur les dimensions physiques.

Conclusion générale sur l'impact du filtrage

L'ensemble des graphiques présentés démontre l'importance cruciale du filtrage et du nettoyage des données brutes avant toute analyse statistique ou modélisation. Que ce soit pour la largeur, la longueur, le tirant d'eau ou encore la vitesse, les données initiales contenaient des valeurs aberrantes (valeurs nulles, extrêmes ou physiquement incohérentes) qui faussaient à la fois la lecture visuelle et l'interprétation numérique.

Après traitement, les distributions deviennent plus réalistes, structurées et représentatives des comportements typiques attendus selon les types de navires. Le filtrage permet ainsi :

- D'améliorer la qualité des visualisations,
- De fiabiliser les analyses descriptives,
- Et de garantir la validité des résultats issus des modèles prédictifs.

Cette étape de prétraitement constitue donc un préalable indispensable à tout travail sérieux sur les données AIS.

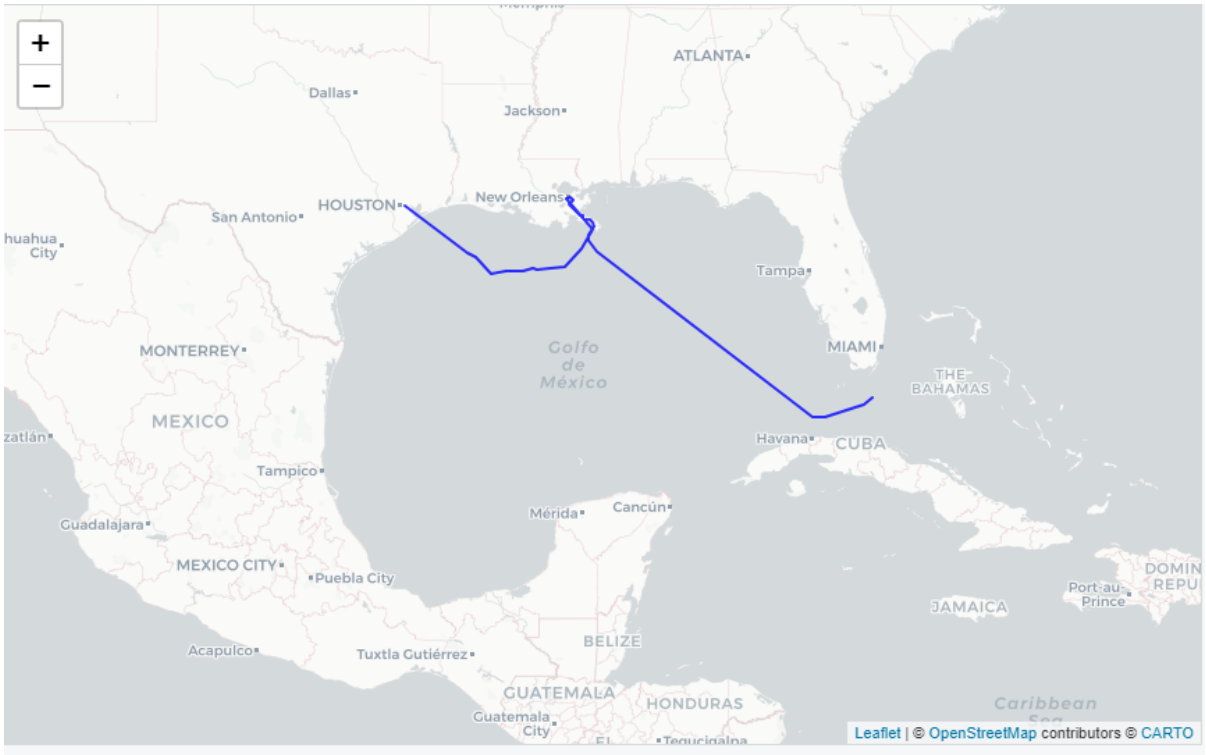
Fonctionnalité 3 Analyse de données AIS : trajectoires, routes maritimes et prédiction de type de navire

1. Construction des trajectoires de navires

Nous avons utilisé les données AIS contenant les positions GPS (latitude et longitude) ainsi que les bases de temps (BaseDateTime) des navires pour reconstruire leurs trajectoires individuelles.

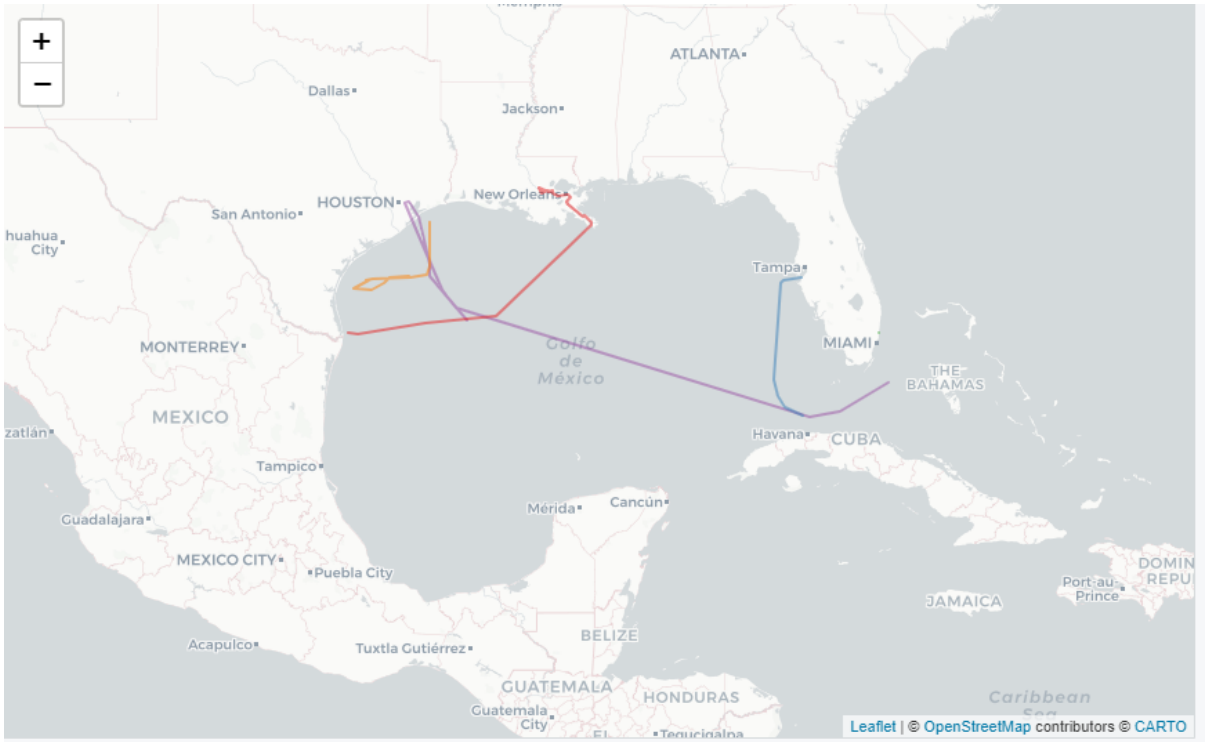
Exemple : Trajectoire d'un navire

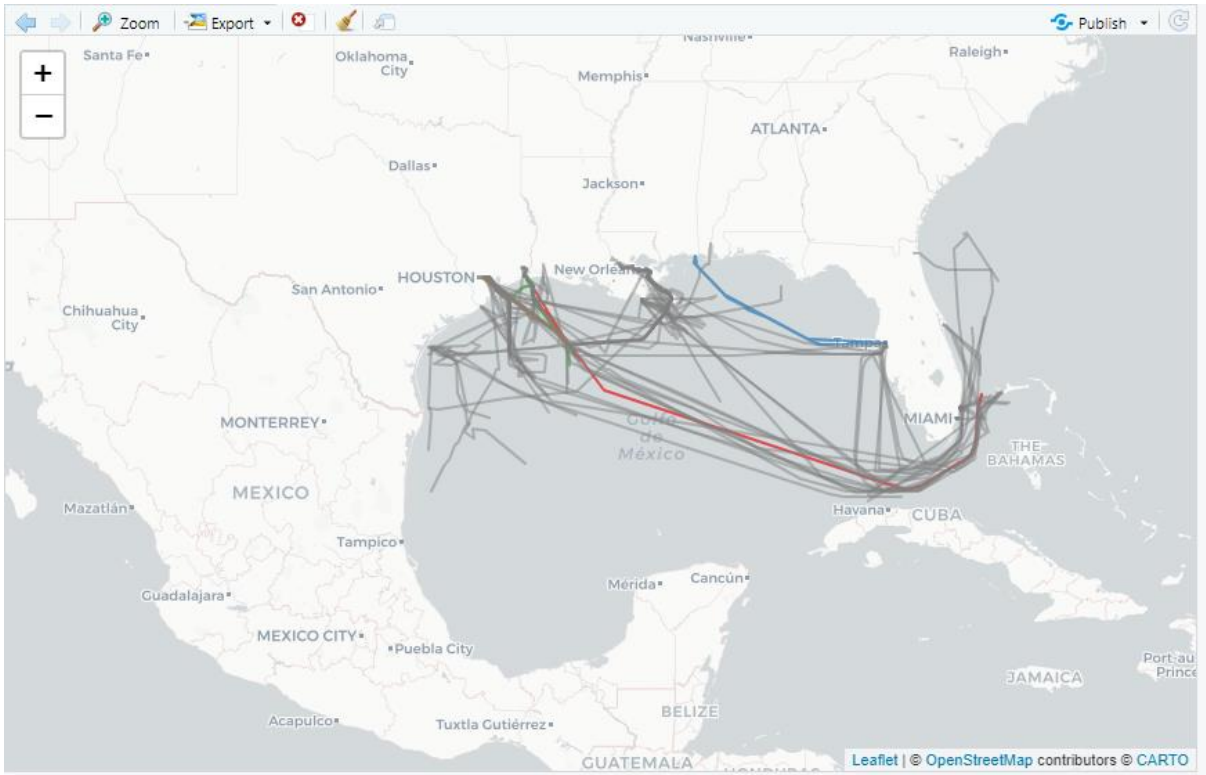
Pour un ensemble de MMSI (identifiants uniques des navires), nous avons filtré les données (dataframe_clean : df_clean) en supprimant les valeurs manquantes, aberrantes et les doublons (ex : latitudes et longitudes hors du Golfe du Mexique) puis nous avons tracé les lignes reliant leurs positions successives. Cela permet de visualiser les déplacements temporels et géographiques des navires dans le Golfe du Mexique.



Carte interactive

Nous avons utilisé le package leaflet pour afficher dynamiquement les trajectoires sur une carte, avec une couleur différente pour chaque navire. Cette visualisation permet d'observer les comportements et routes suivies individuellement.





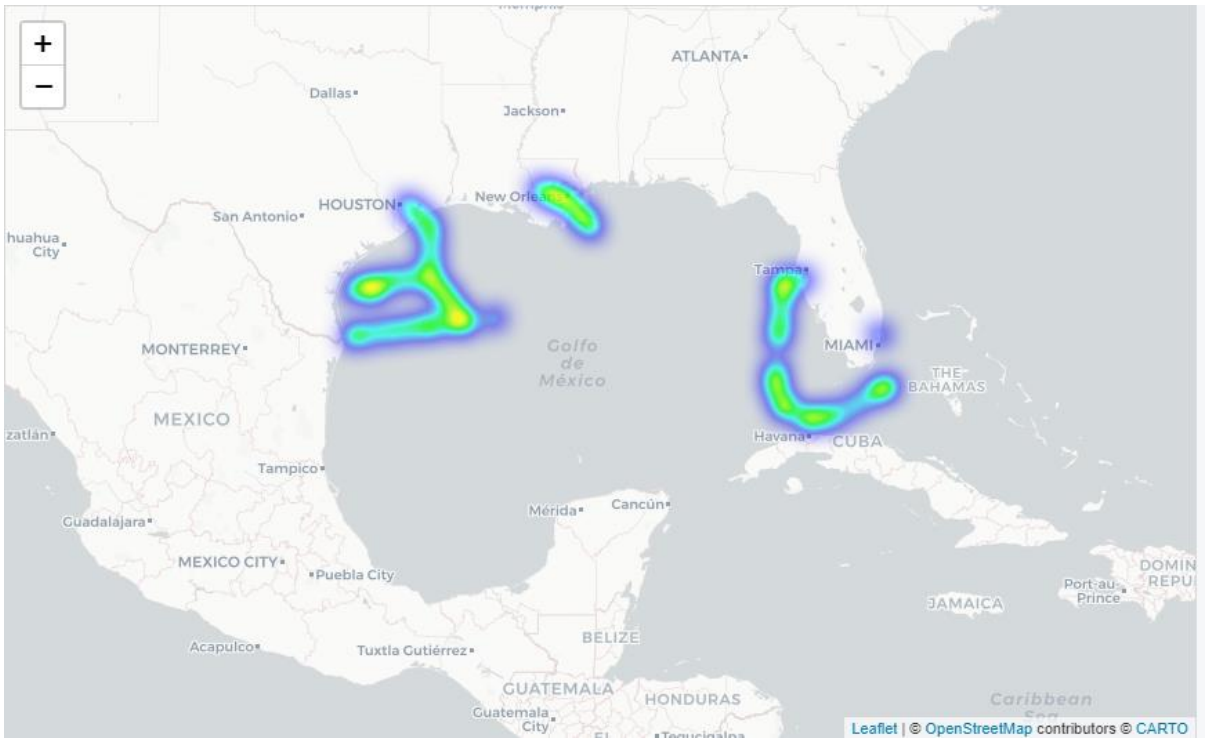
2. <u>Déduction des routes principales (heatmap)</u>

Afin d'identifier les couloirs maritimes les plus fréquemment empruntés, nous avons utilisé une carte de chaleur (heatmap) basée sur la densité des positions GPS de l'ensemble des navires.

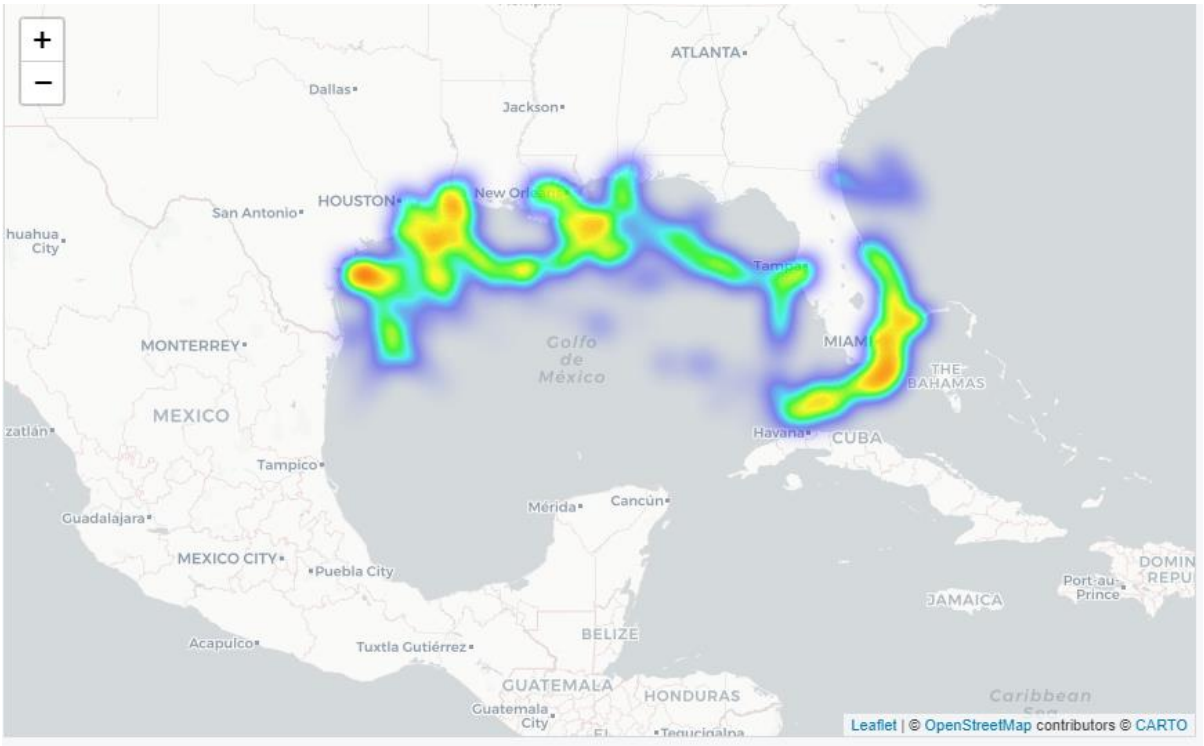
Processus :

- Agrégation des données GPS de tous les navires (avec longitude et latitude)
- Utilisation de addHeatmap() de leaflet.extras pour générer une carte mettant en valeur les zones de forte fréquence de passage
- Plus la zone est de couleur chaude (jaune, orange, rouge), plus les routes maritimes empruntées sont principales

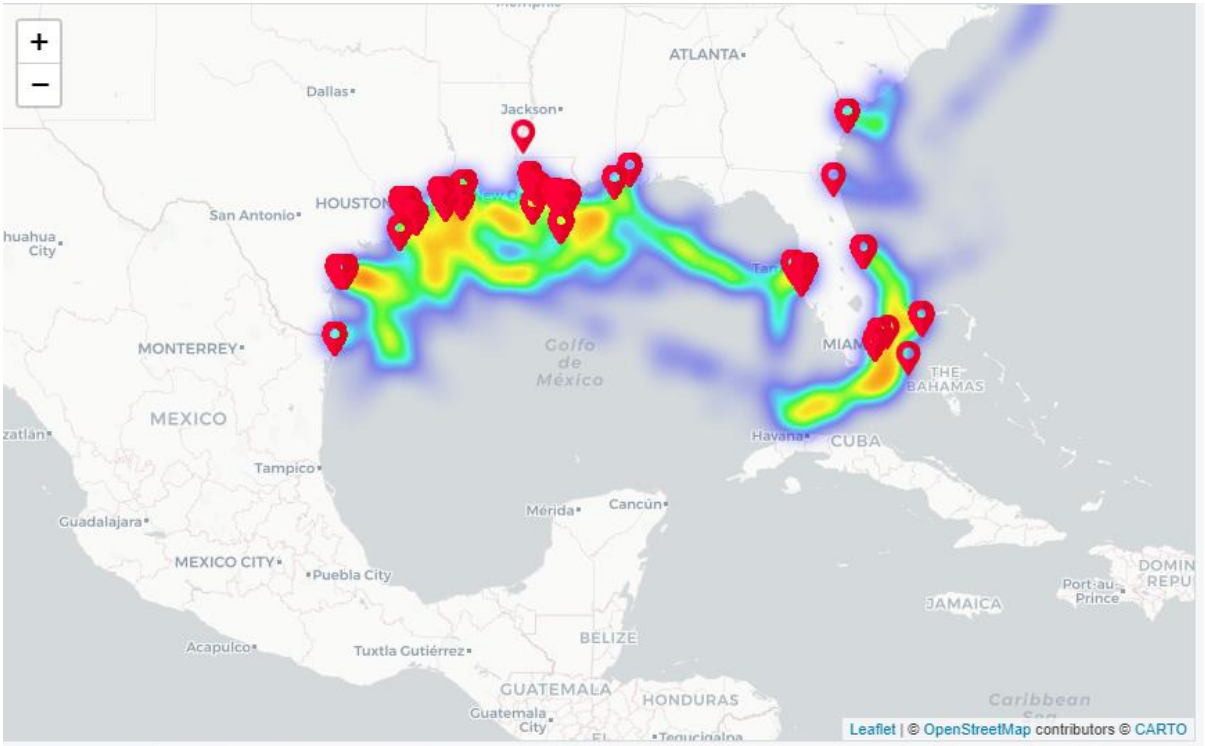
Routes principales de 5 bateaux représentées :



Routes principales de tous les bateaux confondus :



Nous avons pu aussi ajouter les ports principaux du Golfe grâce à la fonction addMarkers(), ce qui a permis de générer une carte comme ci-dessous :



3. <u>Prédiction du type de navire (VesselType)</u>

Nous avons construit un modèle de classification supervisée pour prédire le type de navire à partir de ses caractéristiques physiques et comportementales.

Variables utilisées :

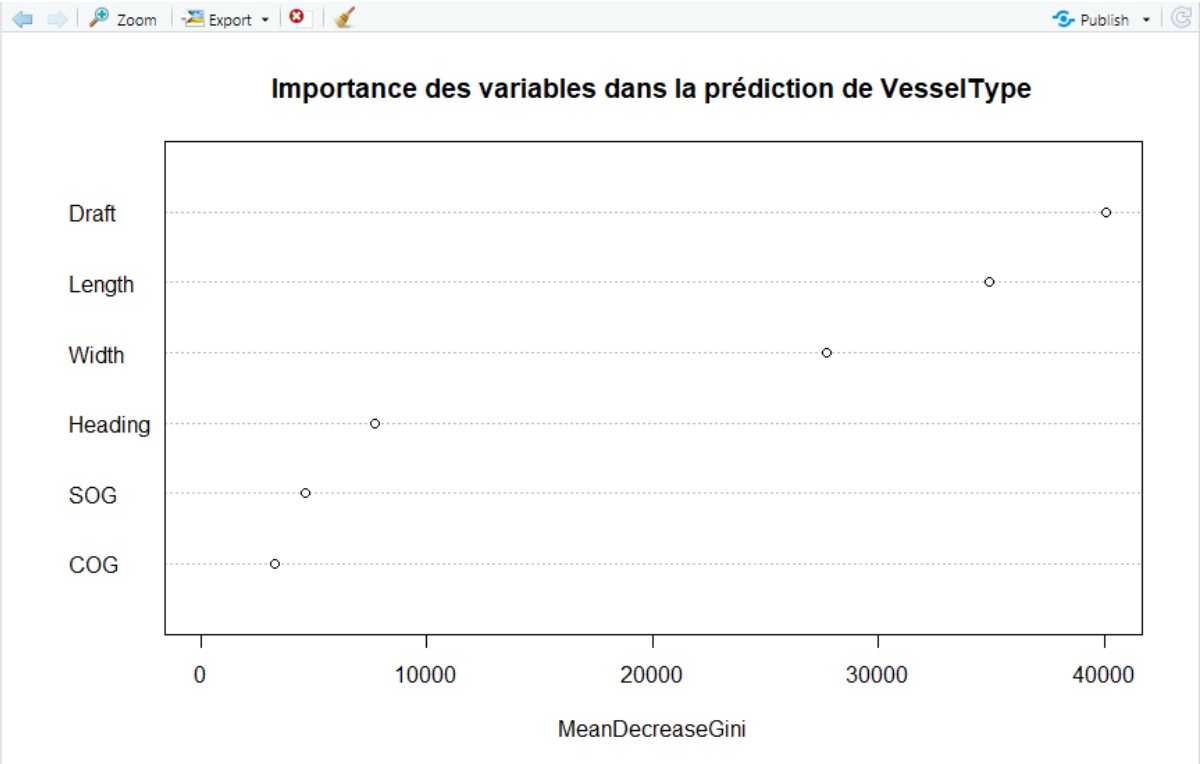
- SOG (Speed Over Ground)
- COG (Course Over Ground)
- Heading
- Length
- Width
- Draft

Modèle :

- Algorithme : Random Forest
- Packages : caret, randomForest, geosphere, Metrics

Résultats :

- Les variables Draft, Length, Width sont les plus importantes pour la prédiction



Ce que ce graphique décrit est que plus la valeur de MeanDecreaseGini est grande, plus la variable contribue à réduire l'impureté (Gini) dans les arbres.

Plus c'est à droite, plus c'est utile pour classer correctement les bateaux.

On peut donc garder Draft, Length, Width comme variable pertinentes pour prédire VesselType c'est-à-dire le type de navires tandis que SOG et COG peuvent être exclues pour

simplifier le modèle car peu impactantes concernant la prédiction du type de navire. Simplifier le modèle en les retirant permettrait de rester pertinent sans trop perdre en précision.

4. Estimation de la vitesse et mesure de l'erreur

Nous avons estimé la vitesse réelle des navires à partir de leur coordonnées GPS (latitude et longitude) et BaseDateTime, puis comparé ces vitesses estimées à la vitesse enregistrée SOG (Speed Over Ground).

Méthode :

- Calcul de la distance entre deux positions successives via la formule de Haversine (distHaversine sur R)
- Division par le temps écoulé pour obtenir la vitesse
- Conversion en nœuds pour avoir la vitesse maritime

Résultats :

- MAE (Erreur moyenne absolue) = 1.2 nœuds
- RMSE (Erreur quadratique moyenne) = 1.8 nœuds
- MAPE (Erreur relative moyenne) = 15.8%

Analyse :

- MAE : on a fait la moyenne de toutes les différences absolues entre la vitesse réelle (SOG) et la vitesse estimée (speed_calc), c'est l'erreur moyenne en nœuds. Si MAE = 1.2 comme dans notre cas, cela veut dire que les vitesses estimées s'écartent en moyenne de 1.2 nœuds des vraies vitesses.
- RMSE : on évalue l'écart au carré, puis on prend la racine carrée à la fin, ce qui pénalise plus fortement les grosses erreurs. Si RMSE = 1.8 nœuds comme dans notre cas, cela signifie que l'erreur moyenne "pondérée" est de 1.8 nœuds.
- MAPE : en moyenne, la vitesse estimée s'écarte de 15.77% par rapport à la vraie vitesse mesurée à bord, ce qui est une erreur tout à fait raisonnable.

Conclusion :

Ce projet a permis de reconstruire et analyser les trajectoires maritimes à partir de données AIS, de cartographier les routes principales, de prédire les types de navires, et de vérifier la

cohérence des vitesses estimées. Ces approches sont directement exploitables pour la surveillance maritime, la logistique et la sécurité portuaire.

Fonctionnalité 4 — Analyse des corrélations entre variables

L'objectif de cette fonctionnalité est d'étudier les liens statistiques entre certaines variables du jeu de données AIS. Cette analyse permet d'identifier les relations significatives et d'évaluer la pertinence de certaines variables pour la prédiction du type de navire (« VesselType »), notamment dans la perspective de la régression logistique de la fonctionnalité 5.

1. Corrélations entre variables quantitatives

Dans un premier temps, une analyse de corrélation a été menée entre les variables quantitatives telles que :

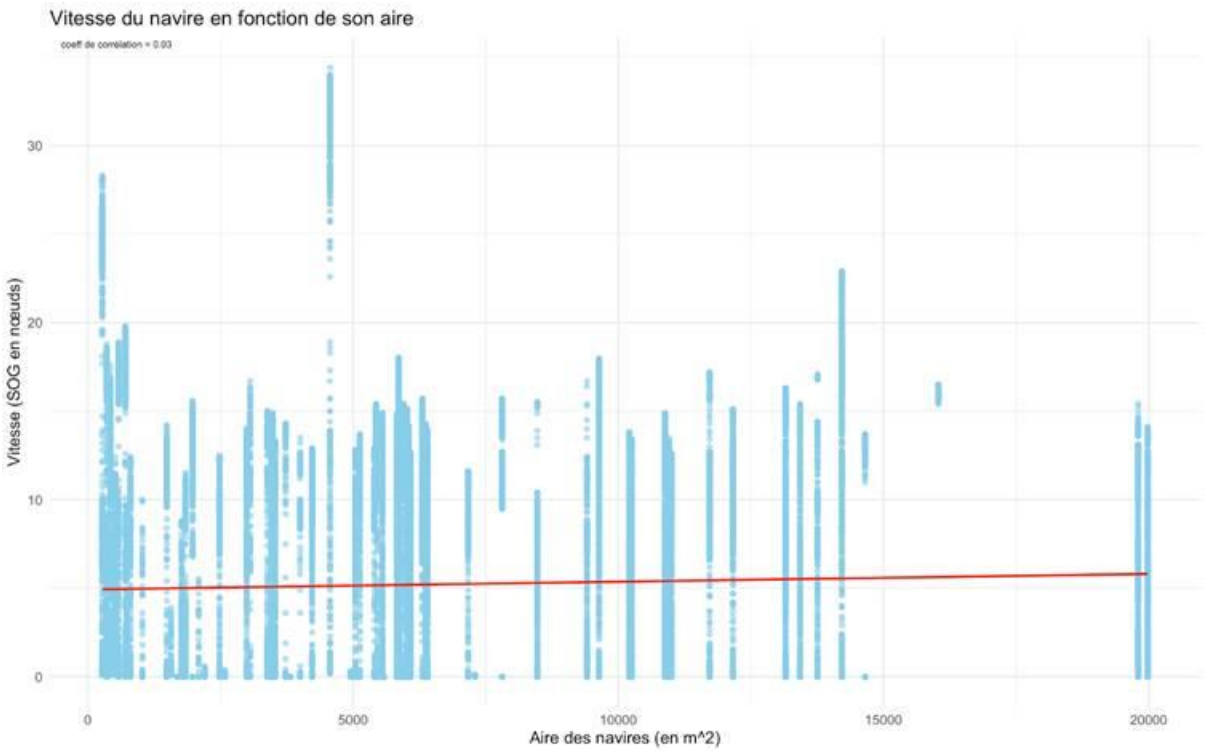
- Length (longueur du navire),
- Width (largeur),
- Area (surface calculée : Length × Width),
- SOG (vitesse sur le fond).

Deux approches complémentaires ont été utilisées :

- Une analyse de variance (ANOVA) pour tester si les moyennes des variables diffèrent significativement selon le type de navire.
- Une matrice de corrélation pour explorer les relations linéaires entre les variables numériques.

Les résultats sont présentés ci-dessous, accompagnés de visualisations graphiques.

Observation : La vitesse du navire n'est pas influencée par le type de navire. La droite de corrélation est pratiquement parallèle à l'axe des abscisses (coefficient de corrélation =0.03), cela appuie notre observation sur nos données.



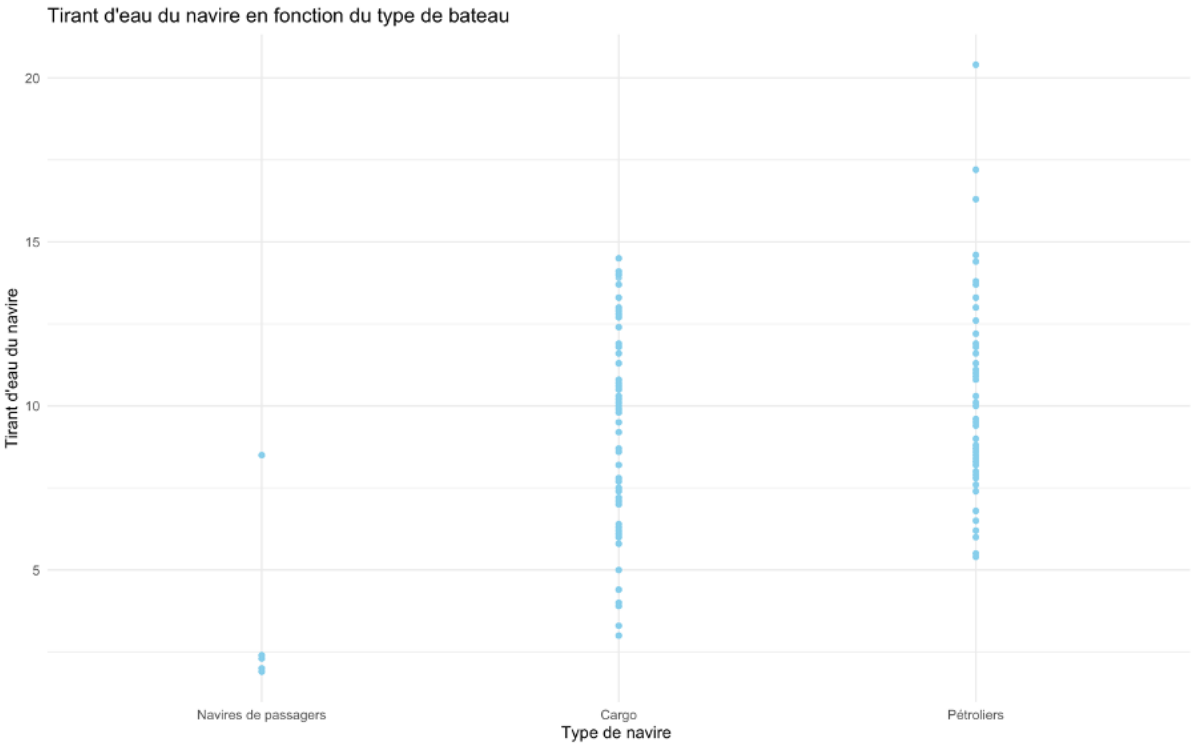
Graphique 1 : *Nuage de points Length vs SOG avec droite de corrélation.*

<u>Analyse de la corrélation entre la vitesse des navires et leur aire</u>

Ce graphique examine la relation entre la vitesse des navires (exprimée en nœuds) et leur aire (en mètres carrés). L'objectif est d'évaluer la pertinence de cette variable dans une démarche de prédiction ou de classification.

L'analyse visuelle révèle une forte dispersion des points, sans structure apparente. La ligne de tendance est quasiment horizontale, ce qui suggère une absence de relation linéaire entre les deux variables. Cette observation est confirmée par un coefficient de corrélation très faible (0,03), indiquant une quasi-absence de lien statistique.

Ce constat met en évidence un point essentiel dans l'analyse exploratoire : toutes les variables disponibles ne sont pas nécessairement informatives. En l'occurrence, l'aire du navire ne semble pas jouer un rôle significatif dans l'explication ou la prédiction de sa vitesse. Par conséquent, cette variable pourrait être écartée lors de la sélection des variables pertinentes pour les étapes de modélisation.



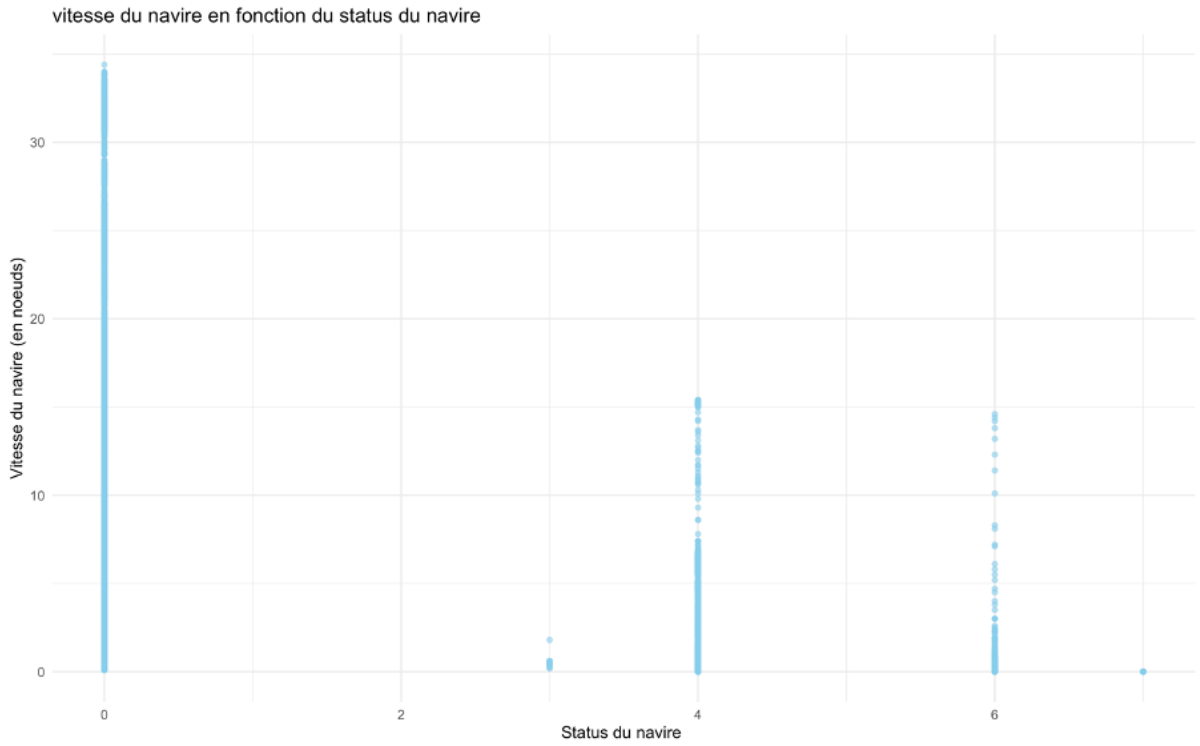
Graphique 2 : *Nuage de points tirant d'eau vs type de bateau*

<u>Analyse de la corrélation entre le tirant d'eau et le type de navire</u>

Le graphique représente la distribution du tirant d'eau (en mètres) selon trois catégories de navires : navires de passagers, cargos et pétroliers. L'axe vertical indique le tirant d'eau, tandis que l'axe horizontal distingue les types de navires.

Les pétroliers présentent les tirants d'eau les plus élevés, ce qui est cohérent avec leur taille et leur besoin de stabilité. Les cargos montrent une variabilité modérée, tandis que les navires de passagers ont généralement un tirant d'eau plus faible, adapté à la navigation en zones portuaires.

Cette variable présente donc une corrélation qualitative avec le type de navire, ce qui en fait une candidate pertinente pour un modèle de classification, notamment par régression logistique.



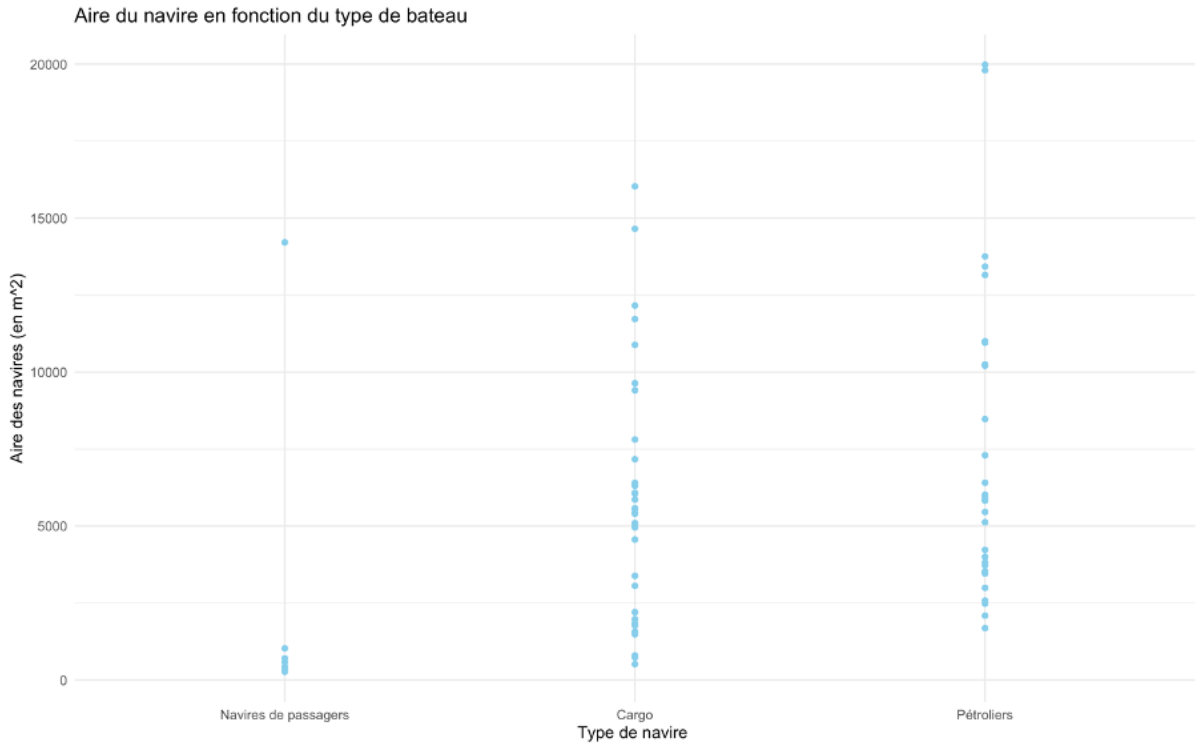
Graphique 3 : *Nuage de points* vitesse vs statut du navire

<u>Corrélation entre la vitesse et le statut du navire</u>

Ce graphique met en relation la vitesse des navires (en nœuds) avec leur statut opérationnel (à quai, en navigation, à l'ancre, etc.). Chaque point représente une observation individuelle.

Les statuts tels que « à l'ancre » ou « à quai » sont associés à des vitesses nulles ou très faibles, tandis que les statuts liés à la navigation présentent une plus grande dispersion des vitesses. Cette variabilité reflète la diversité des comportements en mer.

Le statut du navire apparaît ainsi comme une variable explicative pertinente de la vitesse. Il pourrait être utilisé dans des modèles prédictifs ou pour la détection d'anomalies (par exemple, un navire à l'arrêt se déplaçant à grande vitesse).

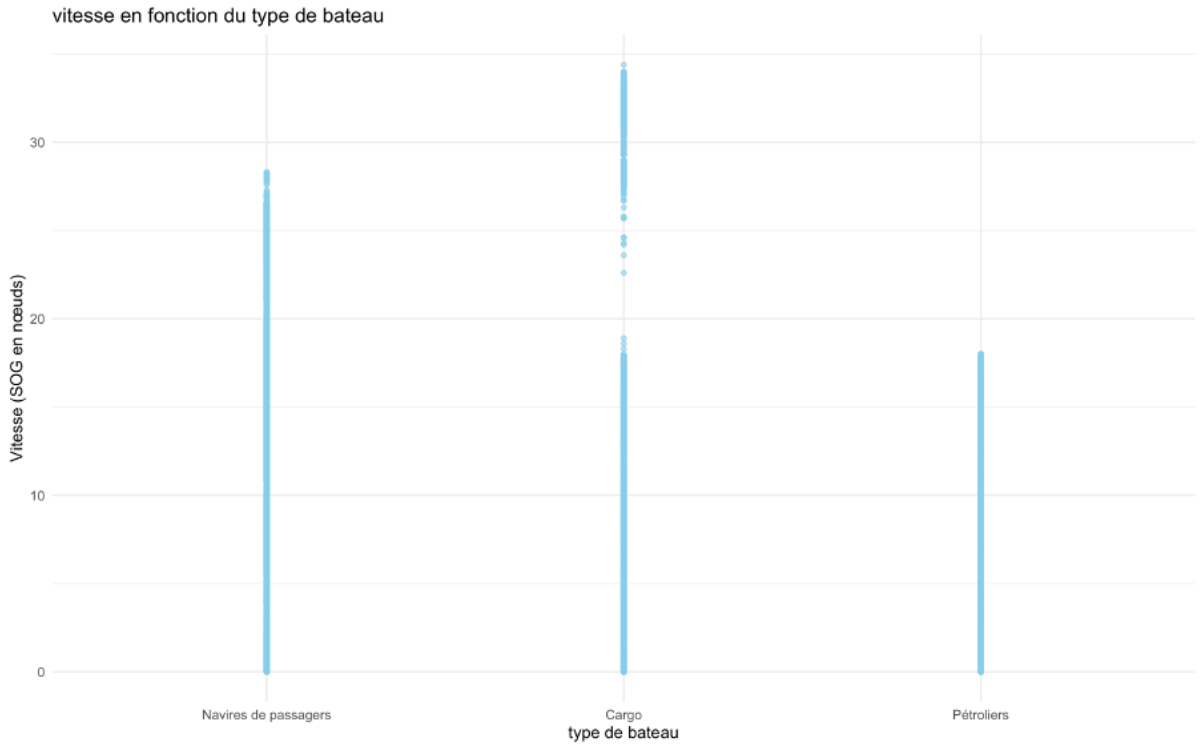


Graphique 4 : *Nuage de points* aire vs statut du navire

<u>Analyse de la corrélation entre l'aire du navire et le type de bateau</u>

Ce graphique explore la relation entre l'aire du navire (en m²) et son type. Les pétroliers affichent les aires les plus importantes, ce qui est attendu compte tenu de leur fonction. Les cargos présentent une grande variabilité, tandis que les navires de passagers ont des aires plus modérées, bien que certains atteignent des valeurs élevées.

Cette variable semble donc bien corrélée au type de navire et pourrait être intégrée dans un modèle de classification.



Graphique 5 : *Nuage de points v*itesse sur le fond (SOG) vs statut du navire

Corrélation entre la vitesse sur le fond (SOG) et le type de navire

Le graphique met en évidence la relation entre la vitesse sur le fond (SOG) et le type de navire. Les navires de passagers présentent les vitesses les plus élevées, suivis des cargos, tandis que les pétroliers sont globalement plus lents.

Cette tendance est cohérente avec les caractéristiques opérationnelles de chaque type de navire. La vitesse sur le fond constitue donc une variable explicative pertinente pour la classification des navires.

Analyse de la variance (ANOVA)

Objectif:

L'objectif de cette analyse est d'évaluer si certaines variables numériques (vitesse, dimensions, tirant d'eau, aire) permettent de distinguer significativement les types de navires (passagers, cargos, pétroliers). Pour cela, une analyse de variance (ANOVA) a été appliquée.

Méthodologie

Pour chaque variable (SOG, Draft, Length, Width, Area), un modèle de type variable ~ VesselGroupLabel a été ajusté. L'ANOVA permet de comparer les moyennes de chaque variable entre les trois groupes de navires. Les résultats sont interprétés à partir de la valeur

de F (qui mesure la variabilité intergroupes) et de la p-value (qui indique la significativité statistique).

Résultats

Les résultats montrent des valeurs de F très élevées et des p-values proches de zéro pour toutes les variables testées :

```
--- ANOVA pour SOG ---
                     Df  Sum Sq Mean Sq F value Pr(>F)
VesselGroupLabel      2  477848  238924    6325 <2e-16 ***
Residuals        169563 6405656      38
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

--- ANOVA pour Draft ---
                     Df  Sum Sq Mean Sq F value Pr(>F)
VesselGroupLabel      2  598154  299077   35603 <2e-16 ***
Residuals        169563 1424400       8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

--- ANOVA pour Length ---
                     Df    Sum Sq  Mean Sq F value Pr(>F)
VesselGroupLabel      2 111156943 55578471   12279 <2e-16 ***
Residuals        169563 767524422     4526
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

--- ANOVA pour Width ---
                     Df   Sum Sq Mean Sq F value Pr(>F)
VesselGroupLabel      2  4184358 2092179   20798 <2e-16 ***
Residuals        169563 17057339     101
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

--- ANOVA pour Area ---
                     Df    Sum Sq   Mean Sq F value Pr(>F)
VesselGroupLabel      2 3.016e+11 1.508e+11    8200 <2e-16 ***
Residuals        169563 3.118e+12 1.839e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6 résultats anova

Ces résultats indiquent que les différences observées entre les groupes sont hautement significatives. Cela signifie que ces variables sont pertinentes pour différencier les types de navires. Par exemple, les pétroliers ont des tirants d’eau et des dimensions nettement supérieurs, tandis que les navires de passagers présentent des vitesses plus élevées.

Matrice de corrélation

Objectif:

Cette seconde analyse vise à explorer les relations linéaires entre les variables numériques afin d'identifier les redondances ou les dépendances structurelles.

Méthodologie

Une matrice de corrélation de Pearson a été calculée entre les variables : SOG, Draft, Length, Width, Area. Le coefficient de corrélation varie entre -1 (corrélation négative parfaite) et 1 (corrélation positive parfaite).

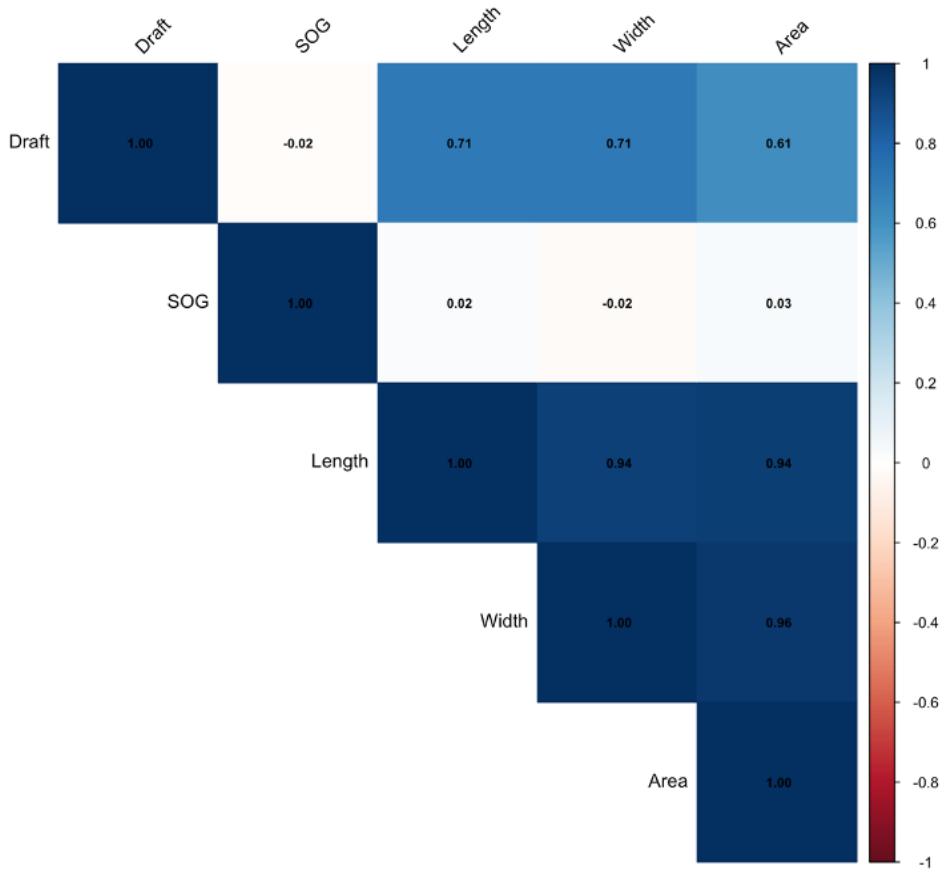


Figure 7 matrice de corrélation

Interprétation

- Les dimensions du navire sont fortement corrélées entre elles, ce qui est attendu.
- Le tirant d'eau est modérément corrélé aux dimensions, ce qui reflète une cohérence structurelle.

- La vitesse (SOG) est faiblement corrélée aux autres variables, ce qui suggère qu'elle dépend davantage du comportement du navire que de sa structure.

Conclusion de l'analyse exploratoire

L'analyse statistique menée sur les données AIS a permis de mettre en évidence plusieurs éléments clés pour la suite du projet. D'une part, l'analyse de variance (ANOVA) a montré que les variables telles que le tirant d'eau, la longueur, la largeur, l'aire et la vitesse présentent des différences significatives selon le type de navire. Ces résultats confirment que ces variables sont pertinentes pour la classification des navires en catégories (passagers, cargos, pétroliers).

D'autre part, la matrice de corrélation a révélé une forte redondance entre les dimensions physiques du navire (longueur, largeur, aire), ce qui suggère qu'il pourrait être judicieux de ne conserver qu'un sous-ensemble de ces variables pour éviter la multicolinéarité dans les modèles prédictifs. En revanche, la vitesse apparaît comme une variable faiblement corrélée aux autres, ce qui en fait un complément intéressant dans une approche multivariée.

Ces analyses constituent une étape essentielle de préparation à la modélisation, en permettant de sélectionner les variables les plus informatives et de mieux comprendre la structure des données. Elles serviront de base pour la construction des modèles de prédiction du type de navire dans la suite du projet.

2. Analyse de l'association entre variables qualitatives

a. Statut de navigation vs Catégorie de vitesse (SOG)

Une variable SOG_cat a été créée pour regrouper les vitesses selon cinq classes :

- Immobilisée (< 1 nœud),
- Lente (1–5),
- Modérée (5–15),
- Rapide (15–30),
- Très rapide (>30).

Un test du Chi² a été réalisé sur la table croisant le Status (code AIS de navigation) et SOG_cat. La p-value est bien inférieure à 0.05, indiquant une dépendance significative.

	Immobilisé	Lente	Modérée	Rapide	Très rapide
En route (moteur) [0]	11344	8417	60110	9201	444
Au mouillage [1]	41176	35	16	0	0
Manœuvre restreinte [3]	3320	497	331	32	0
Amarré [5]	34614	0	0	0	0

Figure 1 : *Tableau croisé du statut de navigation vs catégorie de vitesse*

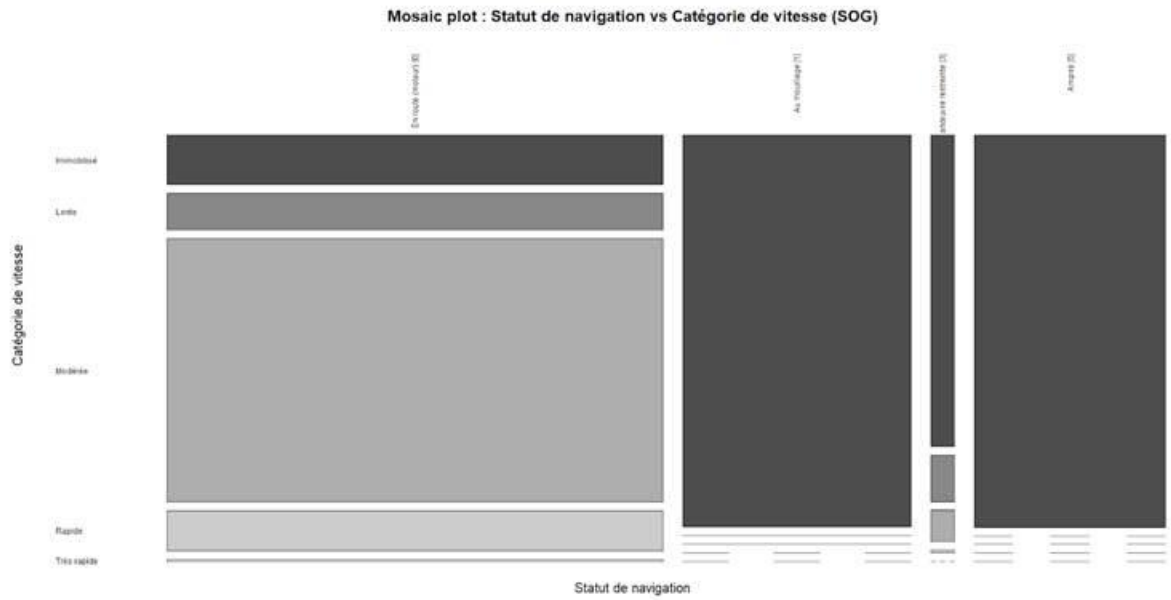
```
        Pearson's Chi-squared test

data:  table_filtered
X-squared = 127646, df = 12, p-value < 2.2e-16
```

Figure 2 : *Résultat test Chi2 du statut de navigation vs catégorie de vitesse*

Le test de Chi² (p-value < 2.2e-16) indique une dépendance significative entre le statut de navigation et la vitesse.

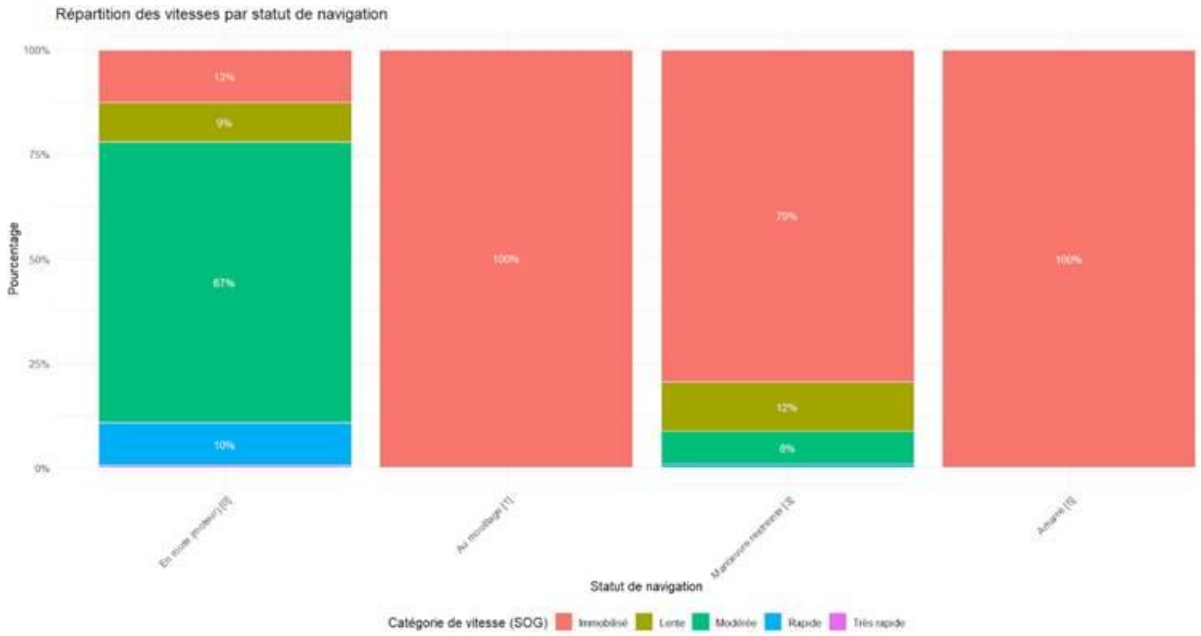
Conclusion:

- Les navires en route (Status 0) sont majoritairement dans les catégories modérée à rapide.
- Les navires amarrés ou au mouillage (Status 1 ou 5) sont très majoritairement immobiles.



Graphique 1 : *Mosaicplot : Status vs SOG_cat.*

Ce graphique permet de visualiser l'association entre le statut du navire et la catégorie de vitesse. Les statuts fixes comme 'Amarré' ou 'Mouillage' apparaissent en majorité dans les vitesses faibles, alors que 'En route (moteur)' couvre une gamme plus large.



Graphique 2 : *Barplot empilé (ggplot2) des pourcentages de SOG_cat par Status.*

Ce graphique donne une vue comparative du profil de vitesse moyen selon chaque statut. Il met en évidence les différences de comportement selon le statut de navigation.

b. Statut de navigation vs Type de navire

Nous avons testé l'association entre le StatusLabel et la variable regroupée VesselGroupLabel (Navires de passagers, Cargo, Pétroliers).

Certaines valeurs du Status (ex. : [2] Non manœuvrable) ont été supprimés pour éviter les faibles effectifs (< 5 par case), condition nécessaire à la validité du test du Chi².

	Navires de passagers	Cargo	Pétroliers
En route (moteur) [0]	10887	41515	37114
Au mouillage [1]	0	16357	24870
Manœuvre restreinte [3]	0	0	4180
Amarré [5]	1331	16605	16678

Figure 3 : *Tableau croisé du statut de navigation vs type de navire*

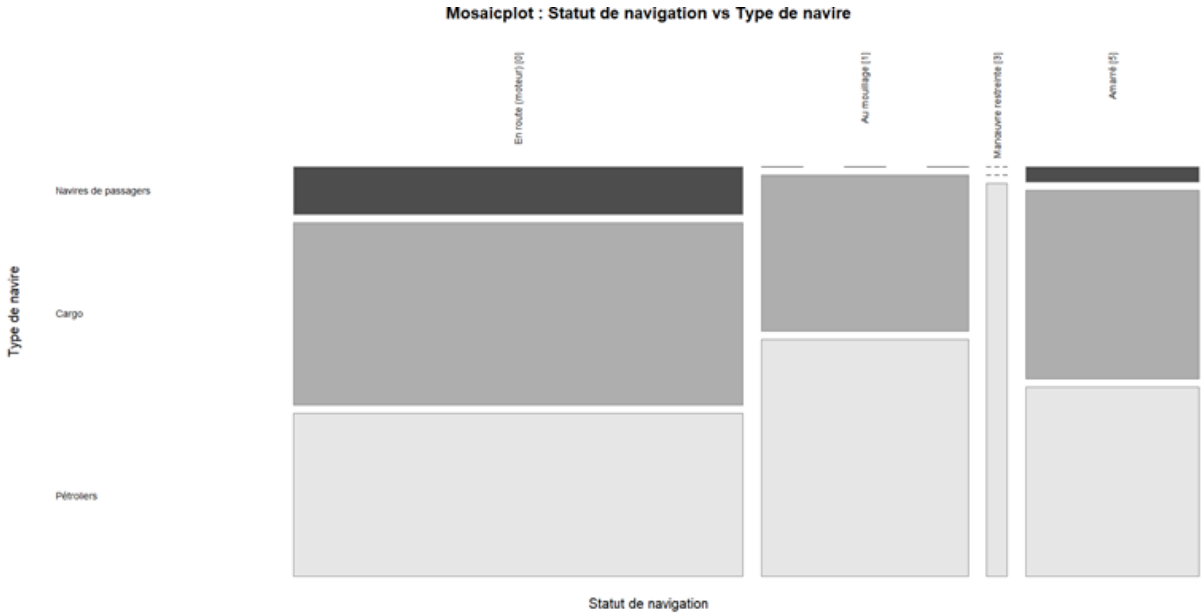
```
        Pearson's Chi-squared test

data:  table_filtered
X-squared = 13475, df = 6, p-value < 2.2e-16
```

Figure 4 : *Résultat du test Chi2 du statut de navigation vs type de navire*

Le test a montré une p-value extrêmement faible, confirmant une dépendance entre le type de navire et son statut de navigation.

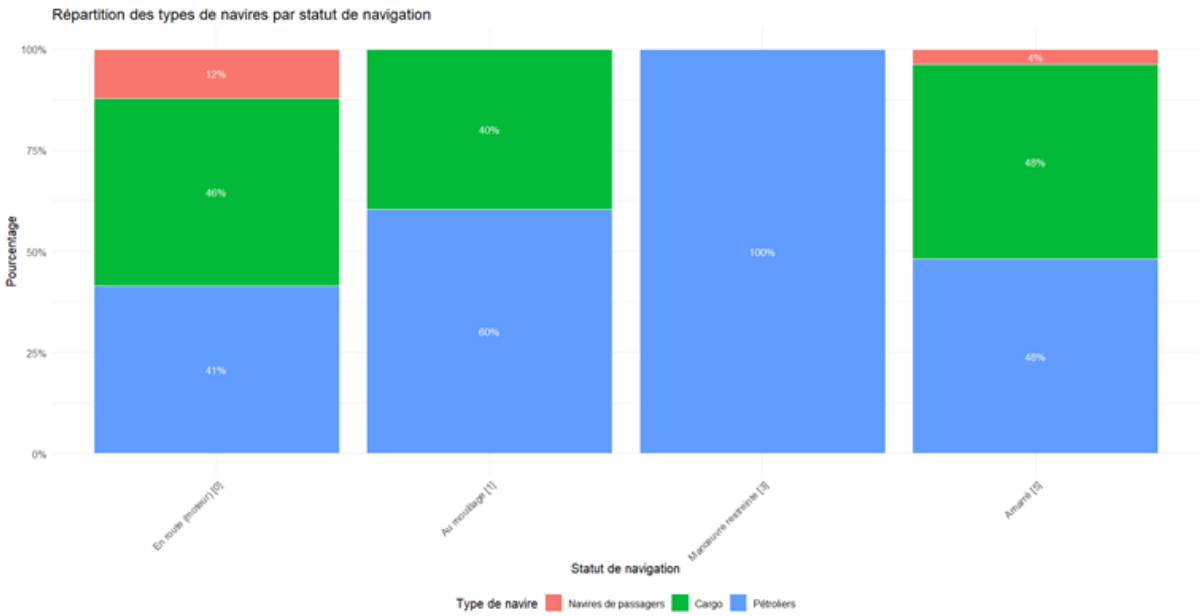
Exemples observés :

- Les cargos sont fréquemment en route ou au mouillage.
- Les navires de passagers sont le plus souvent en mouvement.
- Les pétroliers sont observés dans divers statuts.



Graphique 3 : *Mosaicplot : StatusLabel vs VesselGroupLabel.*

Ce mosaicplot illustre la répartition des types de navires en fonction des statuts de navigation. On observe que les pétroliers ont une plus grande diversité de statuts que les passagers.



Graphique 4 : *Barplot empilé en pourcentages des types de navires par statut.*

Ce graphe permet de comparer les proportions de chaque type de navire dans les différents statuts. Il confirme que les cargos sont majoritaires dans les statuts au mouillage ou en route.

3. Synthèse : variables explicatives du type de navire

L'ensemble des analyses met en évidence que les caractéristiques physiques telles que Length, Width, Draft, Area ainsi que SOG permettent de distinguer efficacement les types de navires. En particulier:

- Draft est très discriminant pour les pétroliers,
- Length et Area sont utiles pour différencier cargos et passagers,
- SOG est moins discriminant seul, mais pertinent croisé avec le statut.

Conclusion de la fonctionnalité 4

Les tests statistiques de corrélation et d'indépendance ont permis de valider l'utilisation de plusieurs variables dans le modèle de prédiction du type de navire. L'association significative entre Status, SOG_cat et VesselGroupLabel confirme que le comportement dynamique du navire est un bon indicateur de sa catégorie. Ces résultats sont exploités dans la fonctionnalité suivante (prédiction par régression logistique).

Fonctionnalité 5 – Prédiction du type de navire et estimation indirecte de la vitesse

Objectif

L'objectif de cette fonctionnalité est d'évaluer si la régression logistique multinomiale, construite pour prédire le type de navire (passagers, cargos, pétroliers), peut être utilisée comme base pour estimer la vitesse individuelle des navires (à travers la variable SOG). L'idée est de mesurer la qualité de cette estimation et de déterminer dans quelle mesure le type de navire prédit peut servir de proxy pour estimer sa vitesse réelle.

Méthodologie

1. Construction du modèle

- Utilisation d'une régression logistique multinomiale (multinom() du package **nnet**) pour prédire VesselGroupLabel.
- Variables explicatives : Length, Width, Area, Draft & Cargo.

2. Sélection des données

- Données nettoyées (élimination des valeurs aberrantes et incohérences).
- Suppression des navires avec SOG <1 noeud car cela fausse les MAE.
- Suppression des statuts peu exploitables (8 : voile, 15 : non défini).
- Filtrage final : seulement les types 60, 70, 80 (passagers, cargos, pétroliers).

3. Entraînement et test

- Séparation en train (70%) et test (30%) avec set.seed(123).

4. Estimation de la vitesse

- Pour chaque classe prédite, la **vitesse moyenne réelle** a été calculée sur les données d'entraînement.
- Chaque navire du jeu de test a reçu cette vitesse moyenne selon son type prédit.

5. Évaluation de l'erreur

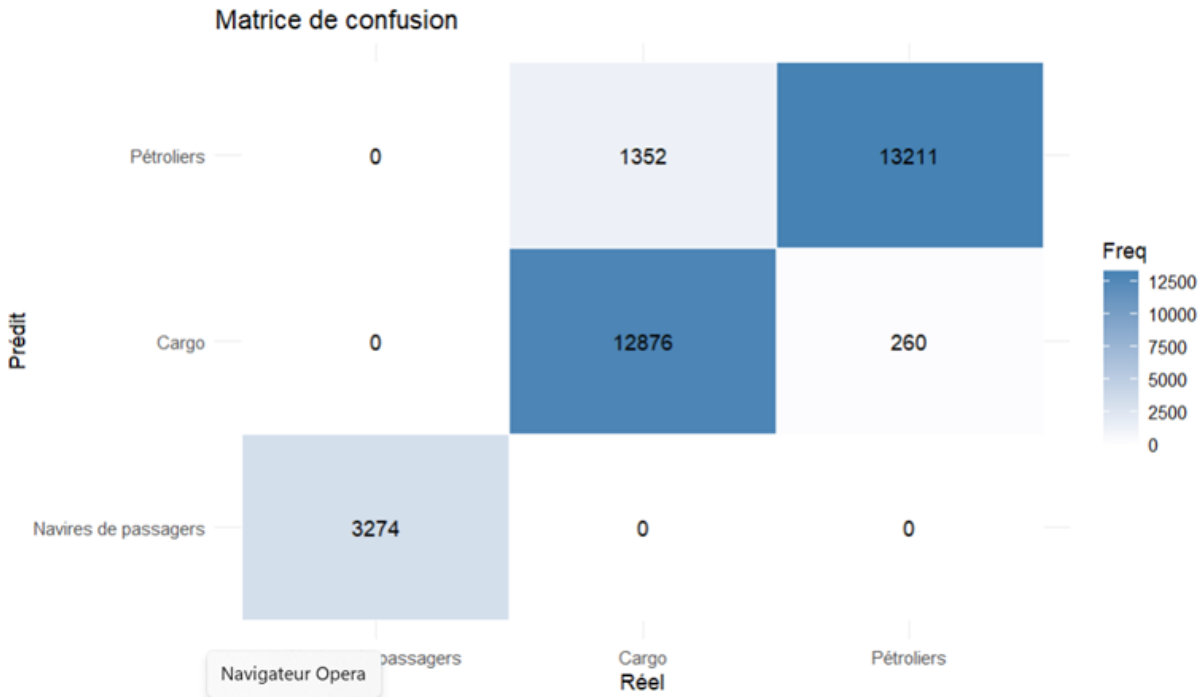
- **MAE** (Mean Absolute Error)
- **RMSE** (Root Mean Square Error)
- **MAE%** (Erreur relative moyenne), uniquement pour les navires avec SOG >= 1.

Résultats (en supprimant les vitesses nulles)

- **Précision du modèle** : 94.8%
- **MAE** : 5.115 nœuds

- **RMSE** : 6.037 nœuds
- **Erreur relative moyenne MAE%** : 68.42 % (après filtrage des navires très lents)

L'erreur relative moyenne de 68.42% est un résultat trop élevé pour que l'on puisse dire que l'on peut donné la vitesse d'un bateau uniquement à partir de son type. Cela peut s'expliquer par la variabilité intra-classe (deux cargos n'ont pas la même vitesse selon la zone, l'opération…) mais également l'absence de facteurs dynamiques (conditions météo, chargement, etc.).



Graphique 1: *Matrice de confusion*

Cette représentation de la matrice permet de visualiser la précision du modèle de classification en comparant les types réels et prédits.

	Classe	Prevalence	Sensibilite	Specificite	Precision	Exactitude
1	Navires de passagers	0.106	1.000	1.000	1.000	1.000
2	Cargo	0.459	0.905	0.984	0.980	0.948
3	Pétroliers	0.435	0.981	0.923	0.907	0.948

Figure 1: *Métriques calculées à partir de la matrice de confusion*

Un tableau récapitulatif des métriques pour chaque classe (sensibilité, spécificité, précision, etc.) permet d'identifier les types bien classés (passagers) et ceux sujets à confusion (cargo vs pétroliers).

Détail des performances de classification

<u>Navires de passagers</u>

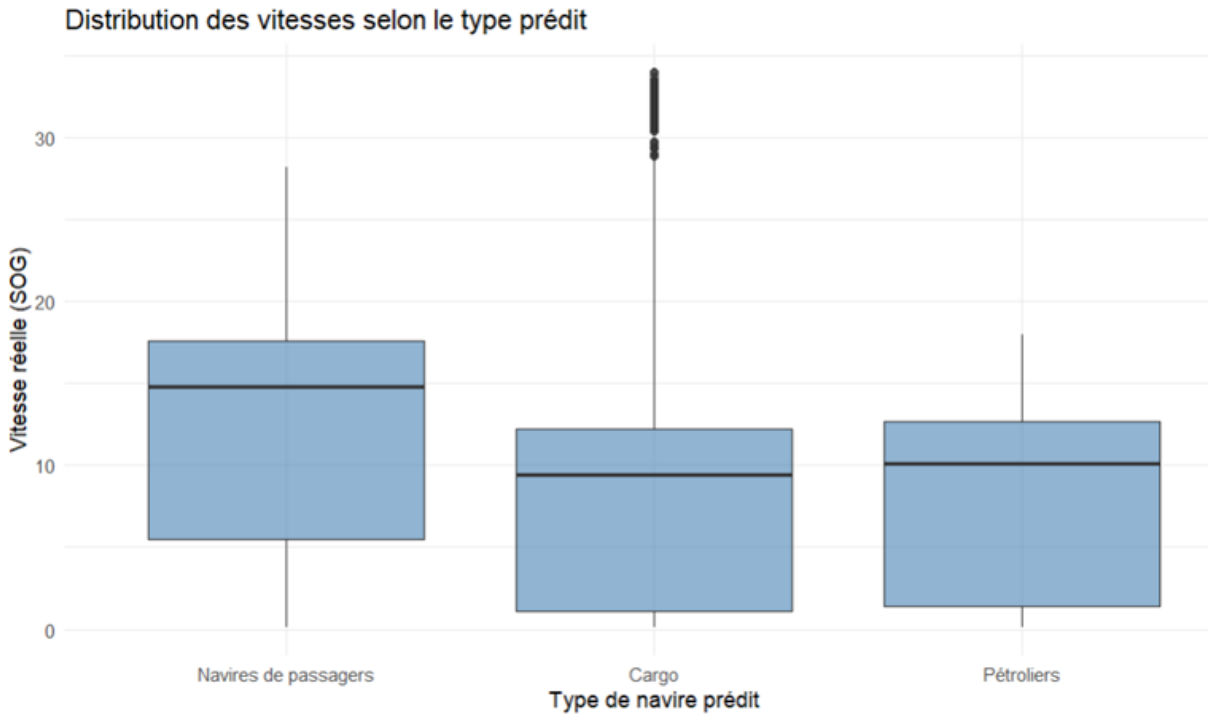
- Prévalence : 10.6 %
- Sensibilité : 1.000
- Spécificité : 1.000
- Précision : 1.000
- Exactitude : 1.000
- Interprétation : Modèle parfait pour cette classe, mais faible fréquence.

<u>Cargo</u>

- Prévalence : 45.9 %
- Sensibilité : 0.905
- Spécificité : 0.984
- Précision : 0.980
- Exactitude : 0.948
- Interprétation : Bon résultat, peu de confusions.

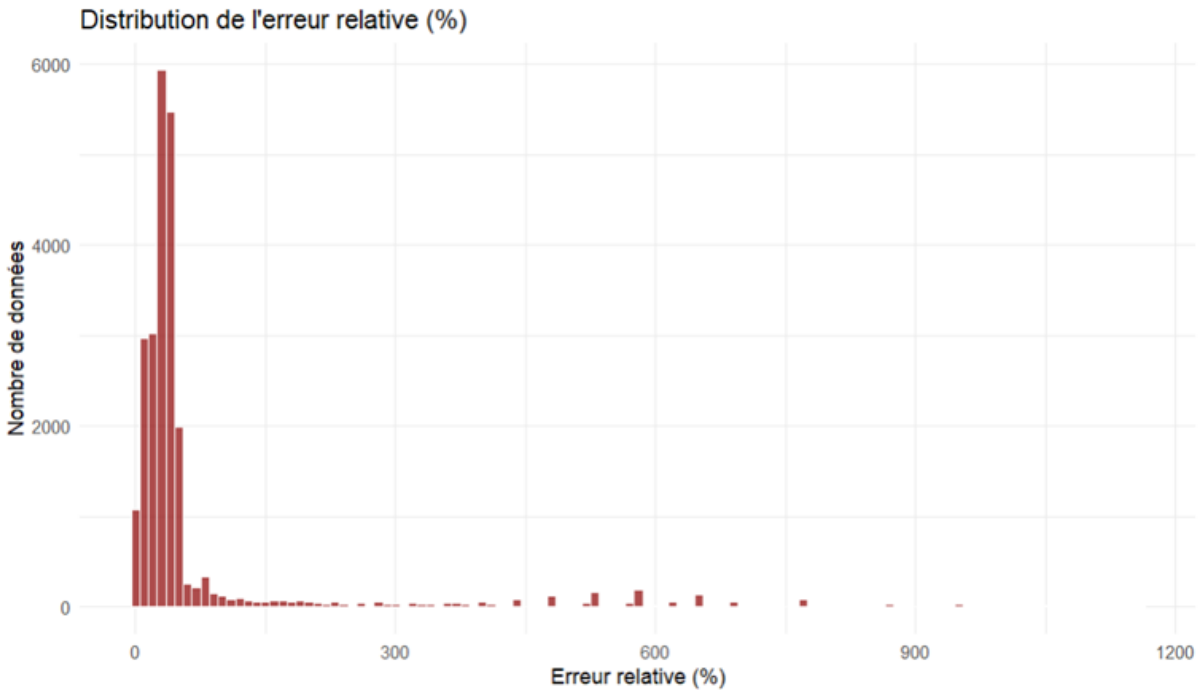
<u>Pétroliers</u>

- Prévalence : 43.5 %
- Sensibilité : 0.981
- Spécificité : 0.923
- Précision : 0.907
- Exactitude : 0.948
- Interprétation : Bonne performance, quelques confusions avec cargos.



Graphique 2: *vitesses réelles en fonction des types De navires prédit*

Un boxplot a été utilisé pour illustrer la variabilité de la vitesse réelle au sein de chaque classe prédite. On remarque une forte dispersion chez les cargos.



Graphique 3: *Distribution de l'erreur relative (en %)*

Un histogramme de l'erreur relative permet de montrer que la majorité des erreurs sont comprises entre 30 % et 100 %, indiquant une estimation très approximative.

Conclusion

Prédire la vitesse des navires à partir du type prédit est une approche simple mais peu précise. La performance du modèle est correcte pour la classification, mais l'utilisation de cette sortie pour estimer une variable continue comme la vitesse introduit une erreur importante. L'analyse montre que **le type de navire n'est pas suffisamment discriminant** pour prédire la vitesse avec fiabilité. Une modélisation directe (régression sur SOG) serait plus adaptée pour cette tâche. Cette approche indirecte par classification a l'avantage d'être simple, mais pour une prédiction plus fine de la vitesse, une régression dédiée intégrant des variables dynamiques serait nécessaire.

Conclusion générale

Ce projet nous a permis de plonger concrètement dans l'univers du Big Data appliqué au domaine maritime, en exploitant les données AIS pour mieux comprendre le comportement des navires. À travers les différentes étapes — nettoyage, visualisation, analyse statistique et modélisation — nous

avons pu mettre en œuvre une démarche complète, allant de la préparation des données brutes jusqu'à la prédiction du type de navire.

Le travail de filtrage s'est révélé essentiel : sans lui, les données étaient trop bruitées pour permettre une analyse fiable. Une fois nettoyées, elles ont révélé des tendances claires et cohérentes, notamment sur les dimensions, la vitesse ou encore les statuts de navigation.

L'analyse statistique nous a permis d'identifier les variables les plus pertinentes pour différencier les types de navires, et les premiers modèles prédictifs ont montré qu'il était possible de classifier efficacement les navires à partir de leurs caractéristiques. Même si certaines approches, comme l'estimation de la vitesse à partir du type, restent approximatives, elles ouvrent des pistes intéressantes pour aller plus loin.

En résumé, ce projet nous a permis de mobiliser des compétences en traitement de données, en visualisation, en statistiques et en modélisation, tout en travaillant sur un cas concret et riche de sens. Il constitue une base solide pour des applications futures en intelligence artificielle, en logistique maritime ou en surveillance des zones portuaires