

ISEN

ALL IS DIGITAL!

NANTES

**PROJET BIG DATA : CONCEVOIR ET DEVELOPPER UNE
APPLICATION D'ETUDE DES DONNEES AIS**

Trinôme 13 : Ewan Ménahéze, Clément Bolomey, Matthieu Lebeau

Nombre de mots : 4 710

Date rendu : 13/06/2025

Fonctionnalité 1 : Description et exploration des données.....	3
Description du jeu de données.....	3
Statistiques descriptives univariées	4
Nettoyage des données.....	6
Valeurs manquantes et valeurs aberrantes.....	6
Doublons.....	6
Fonctionnalité 2 : Visualisation des données sur des graphiques	7
Répartition des bateaux par type de bateau	7
Répartition des bateaux par type de cargaison	8
Répartition des bateaux par longueur de bateau	9
Répartition des bateaux en fonction de leur vitesse	10
Répartition de la longueur (largeur) des bateaux par type de bateau	11
Top 7 des ports les plus attractifs	12
Fonctionnalité 3 : Visualisation des données sur une carte.....	13
Fonctionnalité 4 : Etude des corrélations entre variables.....	16
Analyse des données quantitatives :.....	16
Analyse des variables qualitatives	18
Analyse variables qualitatives × quantitatives.....	21
Fonctionnalité 5 : Prédiction / Régression.....	23
Bonus diagramme de corde	25
Conclusion	27
Bibliographie.....	28

Fonctionnalité 1 : Description et exploration des données

Dans cette étude, nous nous intéresserons aux dynamiques du trafic maritime dans le Golfe du Mexique, une région d'importance stratégique pour le commerce international et les échanges énergétiques. Notre étude se base sur des données AIS (Automatic Identification System) qui permettent de suivre en temps réel les mouvements des navires.

Nous travaillons sur un jeu de données complet enregistré sur plusieurs jours, capturant les positions et caractéristiques d'une flotte variée de navires transitant par cette zone maritime. Il traite les positions de 150 bateaux les 25-27-29-31 mai 2023 toutes les x minutes. Les informations recueillies comprennent à la fois des paramètres techniques de navigation et des données descriptives sur les navires eux-mêmes.

L'objectif principal de cette étude est de comprendre les schémas de circulation dans cette région clé, d'identifier les axes maritimes principaux et d'analyser les comportements des différents types de navires. Cette analyse s'appuie sur des méthodes quantitatives pour traiter les données de positionnement, combinées à une approche qualitative pour interpréter les spécificités des routes maritimes.

Description du jeu de données

Dans notre jeu de données, chaque ligne représente la position et l'état d'un bateau à un instant précis, décrit par différentes caractéristiques. Voici la visualisation des 10 premières données :

	id	MMSI	BaseDateTime	LAT	LON	SOG	COG	Heading	VesselName	IMO	CallSign	VesselType	Status	Length	Width	Draft	Cargo
	<db1>	<db1>	<datetime>	<db1>	<db1>	<db1>	<db1>	<db1>	<chr>	<chr>	<chr>	<db1>	<chr>	<chr>	<chr>	<chr>	<chr>
1	82	636017833	2023-05-29 00:00:06	29.1	-89.3	13.4	228.	227	STOLT LOTUS	"IMO..."	D5MX5	80	"0"	183	"28"	"10"	"83"
2	84	367387350	2023-05-29 00:00:01	25.8	-80.1	5.6	176	511	ISLAND QUEEN	"\\N"	WDE6981	60	"\\N"	20	"7"	"\\N"	"\\N"
3	91	367477280	2023-05-29 00:00:06	29.3	-90.0	0	267.	511	HERO	"IMO..."	WDH2114	60	"0"	47	"9"	"2.3"	"60"
4	103	367037610	2023-05-29 00:00:06	29.9	-90.0	0.4	269	511	FRANK H	"\\N"	WDC5212	60	"0"	15	"5"	"\\N"	"0"
5	118	368112000	2023-05-29 00:00:10	29.8	-92.2	7.1	51.4	511	ADRI LAB	"IMO..."	WDF2836	60	"15"	49	"9"	"\\N"	"60"
6	119	366954420	2023-05-29 00:00:10	27.8	-82.6	0	265.	294	PROVINCETOWN ...	"IMO..."	WDB8185	60	"0"	30	"9"	"1.9"	"60"
7	139	367151450	2023-05-29 00:00:00	29.3	-94.8	11.3	333.	511	RAY STOKER; JR	"IMO..."	WDJ3395	60	"0"	80	"19"	"\\N"	"60"
8	159	367387060	2023-05-29 00:00:02	29.6	-89.9	0.1	277.	511	MISS CAITLIN	"\\N"	WCY8575	60	"15"	13	"\\N"	"\\N"	"31"
9	186	367458840	2023-05-29 00:00:02	25.8	-80.1	0	360	265	OSPREY	"\\N"	WCZ9995	60	"0"	50	"12"	"0"	"60"
10	195	367090270	2023-05-29 00:00:05	29.9	-90.1	0	360	511	TITUS	"\\N"	WCV9471	60	"0"	31	"7"	"\\N"	"99"

Figure 1 : Affichage des 10 premières données

L'ensemble des données comprend deux types principaux d'informations :

- Des mesures chiffrées (comme la position ou la vitesse)
- Des descriptions (comme le type de navire ou son identifiant)

On remarque qu'il y a des informations manquantes (« NA ») ou des incohérences (« \\N »). Ces particularités nécessiteront un travail de correction et d'uniformisation avant toute analyse approfondie. Les premières lignes montrées donnent un aperçu de la variété des informations recueillies, tandis qu'une documentation en figure 2 en précise la signification exacte.

	Name	Description	Example	Units	Resolution	Type	Size
1	MMSI	Maritime Mobile Service Identity value	477220100			Text	9
2	BaseDateTime	Full UTC date and time	2017-02-01T20:05:07		YYYY-MM-DD:HH-MM-SS	DateTime	
3	LAT	Latitude	42.35137	decimal degrees	XX.XXXXX	Double	8
4	LON	Longitude	-71.04182	decimal degrees	XXX.XXXXX	Double	8
5	SOG	Speed Over Ground	5.9	knots	XXX.X	Float	4
6	COG	Course Over Ground	47.5	degrees	XXX.X	Float	4
7	Heading	True heading angle	45.1	degrees	XXX.X	Float	4
8	VesselName	Name as shown on the station radio license	OOCL Malaysia			Text	32
9	IMO	International Maritime Organization Vessel number	IMO9627980			Text	7
10	CallSign	Call sign as assigned by FCC	VRME7			Text	8
11	VesselType	Vessel type as defined in NAIS specifications	70			Integer	short
12	Status	Navigation status as defined by the COLREGS	3			Integer	short
13	Length	Length of vessel (see NAIS specifications)	71.0	meters	XXX.X	Float	4
14	Width	Width of vessel (see NAIS specifications)	12.0	meters	XXX.X	Float	4
15	Draft	Draft depth of vessel (see NAIS specifications)	3.5	meters	XXX.X	Float	4
16	Cargo	Cargo type (see NAIS specification and codes)	70			Text	4
17	TransceiverClass	Class of AIS transceiver	A			Text	2

Figure 22 : Data dictionary

Statistiques descriptives univariées

Dans un premier temps, on souhaite lire les données pour ensuite s'intéresser aux statistiques descriptives univariées de nos données. On va donc chercher à afficher les statistiques descriptives pour l'ensemble de nos paramètres numériques :

id	MMSI	LAT	LON
Min. : 82	Min. : 205776000	Min. : 23.39	Min. : -118.19
1st Qu.: 7899115	1st Qu.: 367037610	1st Qu.: 27.91	1st Qu.: -94.43
Median : 16523442	Median : 367640910	Median : 29.28	Median : -90.21
Mean : 16829900	Mean : 404678555	Mean : 28.72	Mean : -90.09
3rd Qu.: 25793408	3rd Qu.: 374158000	3rd Qu.: 29.73	3rd Qu.: -89.07
Max. : 35186323	Max. : 671830000	Max. : 36.69	Max. : -75.00

SOG	COG	Heading	VesselType
Min. : 0.000	Min. : 0.0	Min. : 0.0	Min. : 60.0
1st Qu.: 0.000	1st Qu.: 127.6	1st Qu.: 127.0	1st Qu.: 60.0
Median : 0.000	Median : 228.2	Median : 301.0	Median : 70.0
Mean : 4.534	Mean : 211.6	Mean : 300.7	Mean : 68.6
3rd Qu.: 6.800	3rd Qu.: 302.3	3rd Qu.: 511.0	3rd Qu.: 80.0
Max. : 102.300	Max. : 360.0	Max. : 511.0	Max. : 89.0

Figure 3 3: Affichage des statistiques descriptives univariées

Le tableau affiche pour chaque variable numérique :

- **Min.** : Valeur minimale.
- **1st Qu.** : Premier quartile (25% des données $\leq$ cette valeur).
- **Median** : Médiane (50% des données).
- **Mean** : Moyenne.
- **3rd Qu.** : Troisième quartile (75% des données).
- **Max.** : Valeur maximale.

Interprétations pour la latitude LAT et la longitude LON :

La majorité des navires se situent entre 27.91°N à 29.73°N et -94.43°O à -89.07°O, ce qui correspond à la zone centrale du Golfe du Mexique. Ce sont les Q1 et Q3 de la latitude et la longitude.

En effet, dans notre cas, on limite la zone du Golfe du Mexique à une section qui correspond à une latitude comprise entre 20 et 30, et une longitude comprise entre -98 et -78.

De plus, les latitudes et longitudes appartiennent bien à l'intervalle [-180 ; 180] : les bateaux sont tous sur Terre (sur l'eau quand même). On respectera tout de même la section utilisée dans cette étude pour le Golfe du Mexique.

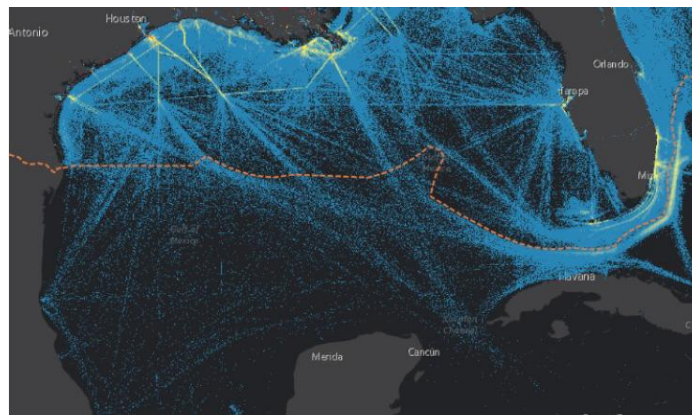


Figure 44 : Section utilisée pour le Golfe du Mexique

Interprétations pour la vitesse SOG :

Plus de la moitié des bateaux sont à l'arrêt puisque la valeur de la médiane est de 0. D'un autre côté, il semble y avoir un problème avec la valeur maximale car elle s'élève à hauteur de 102.3 nœuds. Cette valeur semble très élevée puisque par exemple, en moyenne, un cargo se déplace à une vitesse de 15 nœuds (<https://www.lonelyplanet.fr/article/voyage-en-cargo-mode-demploi#:~:text=Certaines%20destinations%20sont%20particulierement%20adapt%C3%A9es,27%20km%2Fh%20environ>).)

Attention :

Pour certains paramètres numériques, on mesure leurs statistiques descriptives parce qu'ils sont des variables quantitatives alors qu'ils représentent en réalité des variables qualitatives. C'est le cas pour les paramètres « id », « MMSI » et « VesselType ». Il n'y a donc aucun intérêt à interpréter les statistiques descriptives pour ces paramètres.

Il faut aussi faire attention aux valeurs extrêmes. On a vu que c'était le cas pour la vitesse SOG. C'est aussi le cas pour « Heading », qui représente la direction où l'on veut aller, qui

devrait se situer entre 0° et 360°. On remarque qu'à partir du 3^{ème} quartile Q3, toutes les valeurs sont égales à 511° (toutes les mêmes en plus...).

Nettoyage des données

Dans notre jeu de données, certaines données présentent des problèmes : valeurs manquantes, valeurs aberrantes, doublons.

Il faut donc gérer un à un ces problèmes pour nettoyer nos données.

Valeurs manquantes et valeurs aberrantes

Pour corriger les problèmes des valeurs manquantes ou aberrantes, on travaille par rapport à l'intervalle souhaité pour chaque paramètre. On établit donc une table pour définir les règles pour les paramètres :

Variable	Règle
:	-----
	Intervalle de validité
LON	[-98;-78], sinon NA
LAT	[20;30], sinon NA
Heading	[0;360], sinon si COG valide→copie; sinon 180
COG	[0;360], sinon si Heading valide→copie; sinon 180
SOG	[0;68], sinon moyenne SOG par type de navire
Length	[10;364], sinon moyenne Length par type
width	[3;364], sinon moyenne Width par type
Draft	[0,5;24], sinon moyenne Draft par type
Cargo	[0;99], sinon médiane par type
Status	[0;15], sinon médiane par type

Figure 55 : Table des règles

Vitesse du bateau le plus rapide : <https://www.annoncesbateau.com/conseils/bateaux-les-plus-rapides-du-monde/#:~:text=Le%20Vestas%20Sailrocket&text=Il%20a%20m%C3%A0me%20r%C3%A9ussi%20une,record%20est%20toujours%20%C3%A0%20battre.>

Remarque : ce moyen permet aussi de remplacer les « /n » par des « NA ».

Doublons

Pour corriger les doublons, on regarde si un ID est présent plusieurs fois dans notre jeu de données. Si c'est le cas on supprime entièrement les doublons pour ne garder que la première ligne associée à l'ID.

Fonctionnalité 2 : Visualisation des données sur des graphiques

À la suite de l'exploration et du nettoyage des données réalisés dans la Fonctionnalité 1, cette section vise à illustrer les caractéristiques clés du trafic maritime à travers des représentations graphiques. Les visualisations permettront de mieux appréhender la répartition des navires, leurs comportements, et les dynamiques spatiales et temporelles dans le Golfe du Mexique.

Répartition des bateaux par type de bateau

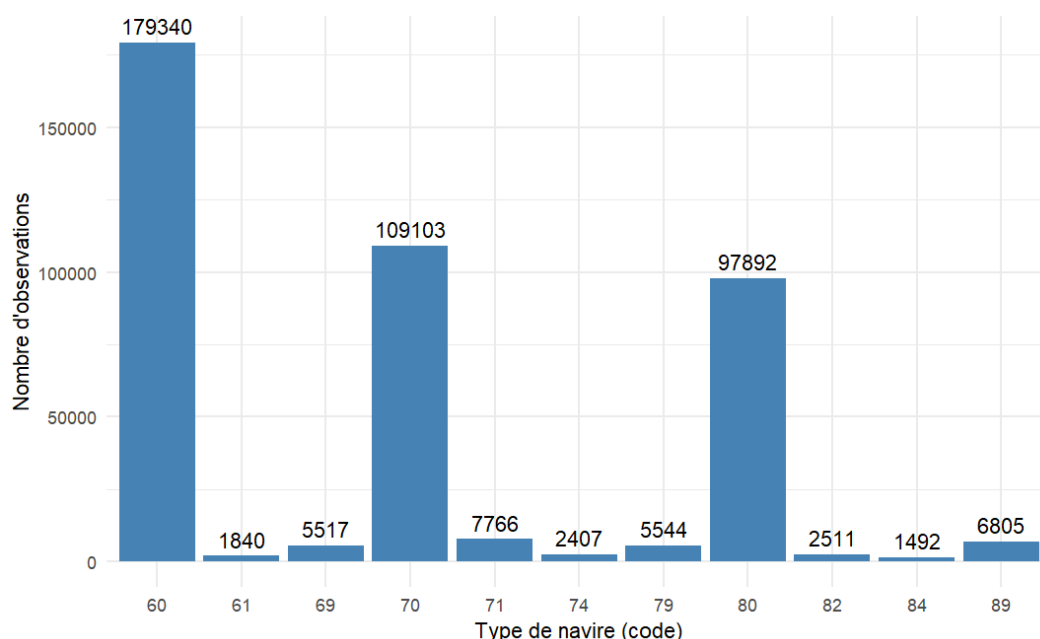


Figure 6 : Fréquence des bateaux en fonction du type de bateau

La répartition des bateaux dans le Golfe du Mexique révèle une prédominance marquée des navires de type « Passenger » (codes 60 à 69), avec environ 200 000 observations, ce qui en fait la catégorie la plus représentée. Ces navires incluent probablement des ferries, des bateaux de croisière et des embarcations destinées au transport de passagers, reflétant l'importance du tourisme et des déplacements locaux dans la région.

Les bateaux de type « Cargo » (codes 70 à 79), avec 125 000 observations, constituent la deuxième catégorie en importance. Cette présence significative souligne le rôle clé du Golfe du Mexique dans le commerce maritime international, notamment pour le transport de marchandises diverses.

Enfin, les navires « Tanker » (codes 80 à 89), au nombre d'environ 110 000, occupent une place centrale dans l'activité maritime régionale. Ces bateaux, spécialisés dans le transport d'hydrocarbures, témoignent de l'importance stratégique du Golfe pour l'industrie pétrolière et gazière.

Répartition des bateaux par type de cargaison

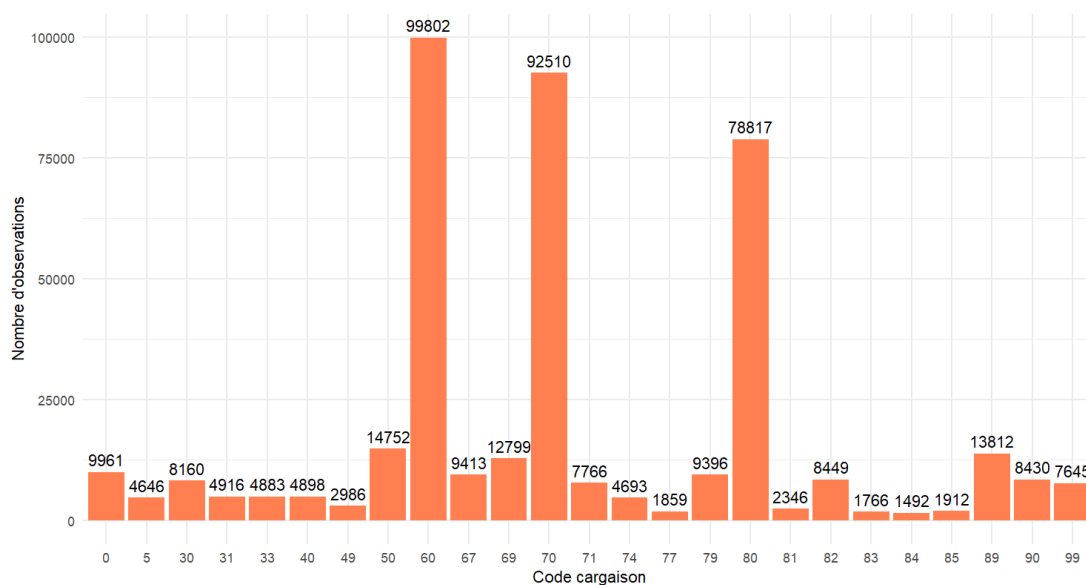


Figure 7 : Fréquence des bateaux en fonction du type de cargaison

La répartition des bateaux par type de cargaison dans le Golfe du Mexique montre que trois types de cargaison représentent à eux seuls l'essentiel du trafic maritime. Comme l'illustre la figure 6, les types 60, 70 et 80 sont largement prédominants, suggérant une spécialisation ou une forte demande pour ces catégories spécifiques. Cette concentration peut refléter la structure économique de la région, les flux commerciaux dominants ou encore la nature des infrastructures portuaires adaptées à ces cargaisons.

Répartition des bateaux par longueur de bateau

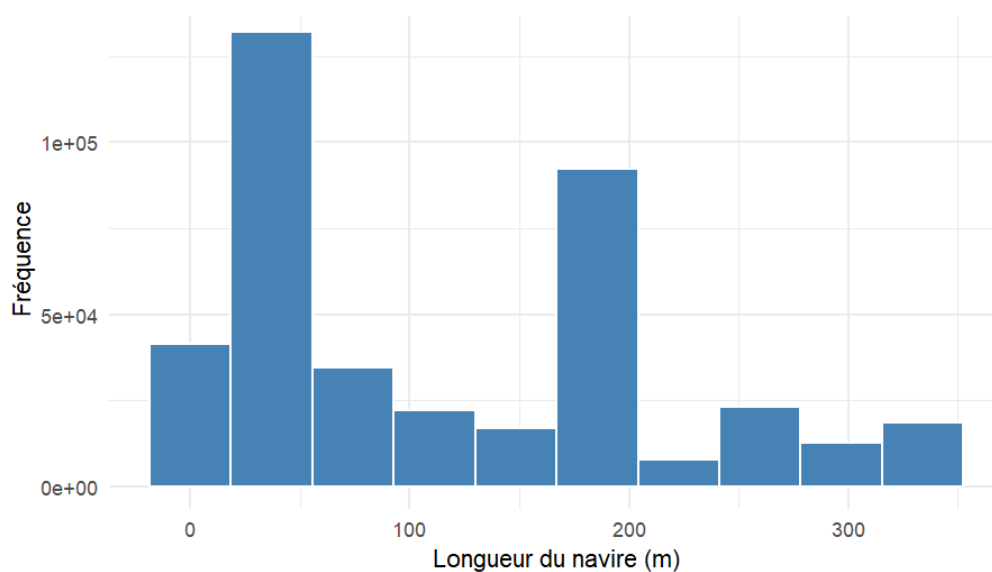


Figure 8 : Fréquence des bateaux en fonction de longueur du bateau

En ce qui concerne la répartition des bateaux selon leur longueur dans le Golfe du Mexique, deux catégories se distinguent nettement. Les navires mesurant entre 20 et 50 mètres sont les plus nombreux, atteignant environ 135 000 unités, ce qui suggère une forte présence de petits ou moyens navires, possiblement utilisés pour la pêche, les services portuaires ou le cabotage. En parallèle, on observe un second pic important pour les navires de grande taille, entre 180 et 200 mètres, avec près de 100 000 unités, ce qui pourrait refléter un trafic intense de cargos ou pétroliers de haute capacité. Cette double concentration traduit la diversité fonctionnelle de l'activité maritime dans la région, allant des opérations locales aux échanges internationaux.

Répartition des bateaux en fonction de leur vitesse

Dans notre cas, la majorité des bateaux sont soit à l'arrêt, soit se déplacent à une vitesse très faible, généralement autour d'un nœud. Cette forte concentration à basse vitesse a pour effet d'écraser la lisibilité de la distribution globale. Afin de mieux analyser la dynamique des navires en mouvement, nous avons donc choisi d'exclure ces deux vitesses du graphique. Par ailleurs, comme mentionné dans la section précédente, nous avons fixé une borne supérieure à 68 nœuds afin de filtrer les valeurs extrêmes peu représentatives. Cette approche permet de mieux visualiser la répartition des vitesses réelles de navigation, sans être biaisée par les situations d'arrêt ou les outliers.

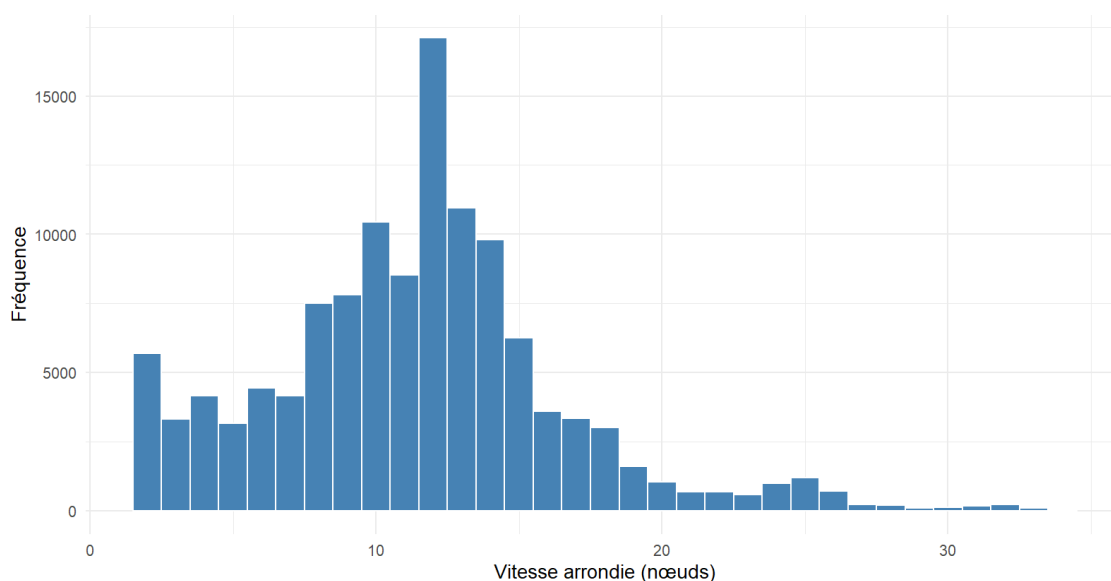


Figure 9 : Fréquence des bateaux en fonction de la vitesse des bateaux

On observe que la distribution des vitesses des bateaux dans le Golfe du Mexique est centrée autour de 12 nœuds, ce qui correspond approximativement à la vitesse moyenne des navires à l'échelle mondiale. Cette concentration suggère que la majorité des navires en déplacement adoptent une vitesse de croisière standard. Par ailleurs, on note une décroissance progressive du nombre de bateaux à mesure que la vitesse augmente, ce qui reflète une réalité opérationnelle : peu de navires naviguent à des vitesses élevées, probablement en raison de contraintes économiques, techniques ou réglementaires.

Répartition de la longueur (largeur) des bateaux par type de bateau

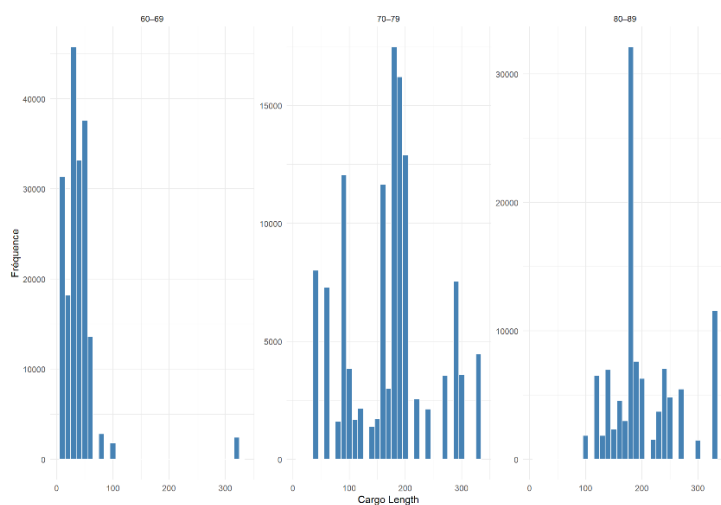


Figure 10 : Fréquence des bateaux en fonction de la longueur du bateau, par type de bateau

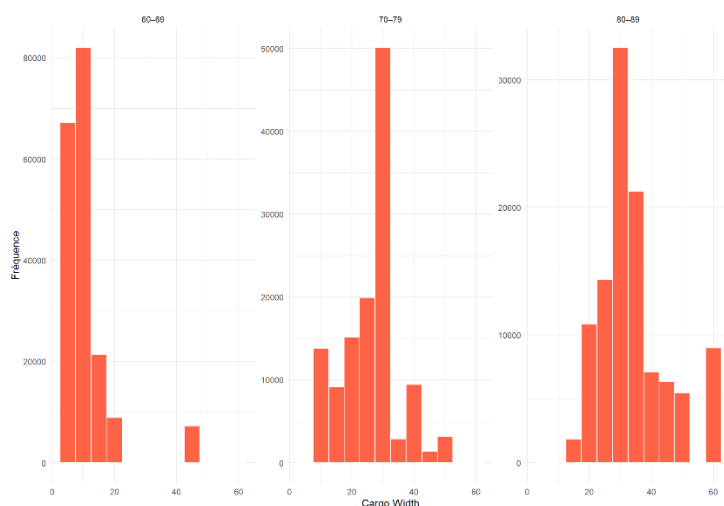


Figure 11 : Fréquence des bateaux en fonction de la largeur du bateau, par type de bateau

Les dimensions des navires varient de manière significative en fonction de leur type, chaque catégorie présentant des caractéristiques propres. Les bateaux de type "passenger" ont, dans la grande majorité des cas, une longueur inférieure à 100 mètres. Toutefois, une exception notable est observée pour quelques unités dépassant les 300 mètres, très probablement des paquebots de croisière.

Les navires de type "cargo" et "tanker" présentent quant à eux une distribution plus étendue des longueurs, allant de 40 à 330 mètres, avec une concentration marquée entre 180 et 200 mètres — ce qui reflète des standards fréquents dans le transport maritime international.

Les largeurs suivent une tendance similaire : plus un bateau est long, plus il est large. Cette corrélation entre longueur et largeur s'explique par des considérations d'équilibre, de capacité de charge et de stabilité, et reste cohérente à travers les différents types de navires.

Top 7 des ports les plus attractifs

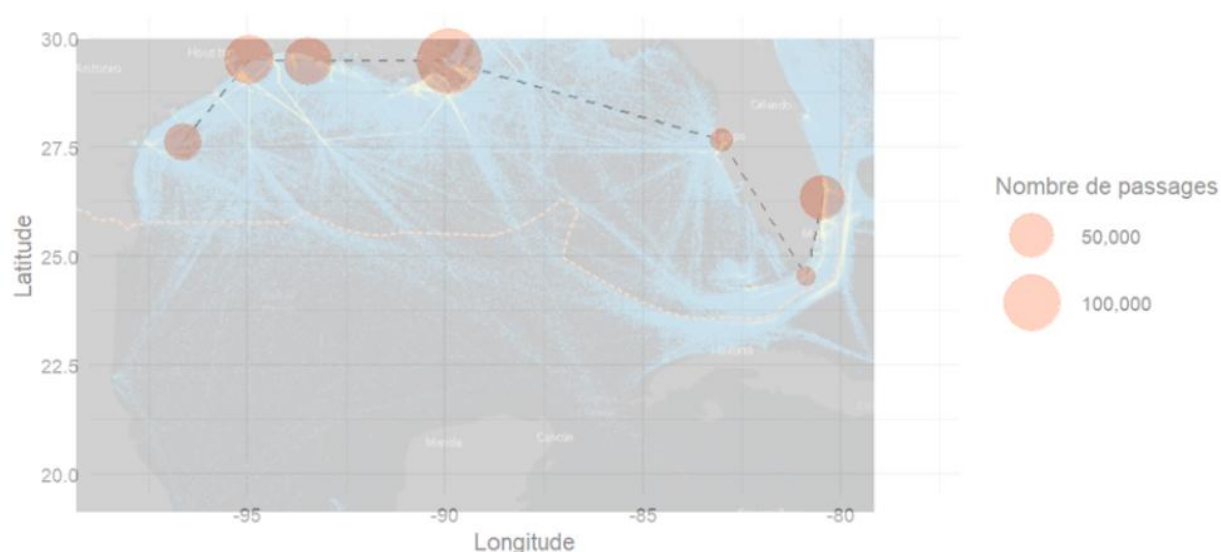


Figure 12 : Visualisation des ports les plus attractifs dans le Golfe du Mexique

Dans la zone du Golfe du Mexique, la densité portuaire favorise un flux continu de navires. D'après la figure 12, le port de La Nouvelle-Orléans se distingue nettement avec plus de 100 000 passages enregistrés, ce qui en fait le principal nœud maritime de la région. Il est suivi par les ports de Miami, Houston et Beaumont, chacun enregistrant plus de 50 000 passages, témoignant de leur importance dans les échanges commerciaux.

Le port de Portland, avec un trafic inférieur à 50 000 passages, et celui de Tampa, autour de 10 000 passages, affichent une activité plus modérée. Un autre point notable apparaît entre Miami et La Havane, avec environ 10 000 passages : ce couloir maritime semble constituer une voie de transit stratégique, probablement en lien avec des routes commerciales ou des corridors maritimes entre les États-Unis et les Caraïbes.

Fonctionnalité 3 : Visualisation des données sur une carte

Dans un premier temps, nous avons représenté les trajectoires des navires sur une carte, en utilisant les coordonnées de latitude et de longitude associées aux émissions successives de chaque bateau dans le temps. Cependant, avec un total de 150 navires, la superposition des trajectoires rendait la carte difficilement lisible. Pour améliorer la clarté, nous avons choisi de regrouper les navires par blocs de 10, triés par ordre alphabétique, et d'afficher leurs trajectoires par groupes de 10. La première carte ci-dessous illustre les déplacements de 10 bateaux.

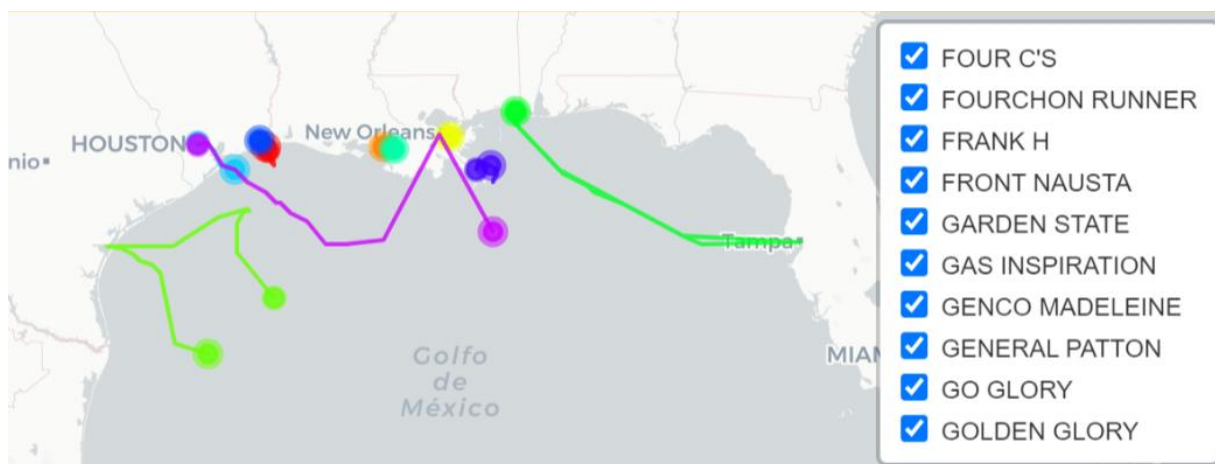


Figure 13 : Trajectoires des bateaux du groupe 6

Dans un second temps, nous avons voulu aller au-delà des trajectoires individuelles et visualiser la densité géographique des positions relevées. Pour cela, nous avons généré une carte de chaleur (heatmap), où les zones les plus fréquentées par les navires apparaissent en rouge, traduisant une plus forte densité. La première heatmap représente la densité globale sur l'ensemble du Golfe du Mexique, tous types de bateaux confondus.

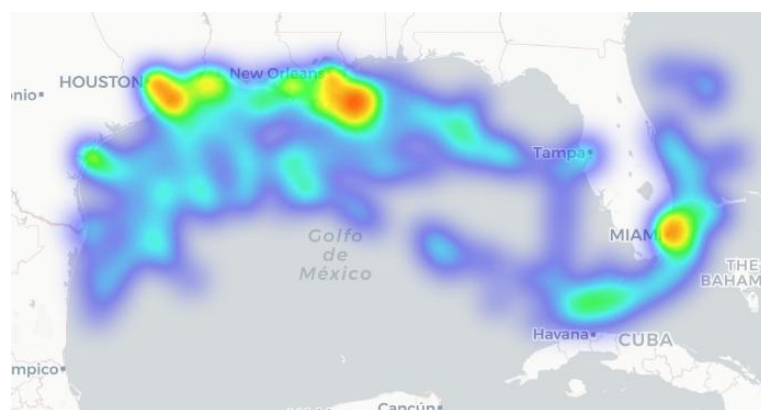


Figure 14 : Heatmap du Golfe du Mexique

Nous avons ensuite enrichi cette visualisation en permettant de filtrer la carte de chaleur selon le type de navire.

Cette fonctionnalité met en évidence des différences nettes de répartition. Par exemple, les bateaux de type "Passenger" sont très concentrés dans certaines zones fluviales ou côtières, correspondant à des lignes de navettes régulières :

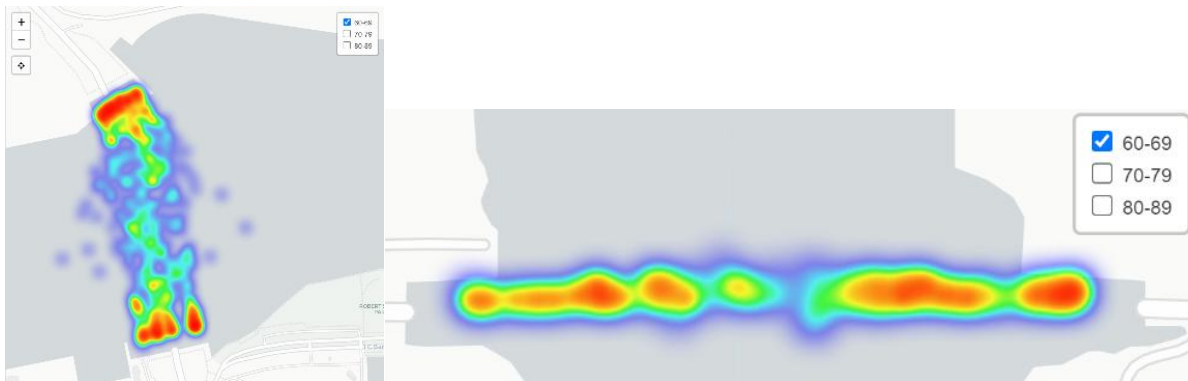


Figure 15 : Focus sur des trajets des bateaux "passenger"

En revanche, les cargos et tankers présentent des trajectoires plus larges et proches des routes commerciales maritimes, avec une légère flexibilité supplémentaire dans les parcours pour les cargos (car ils sont plus petits) :

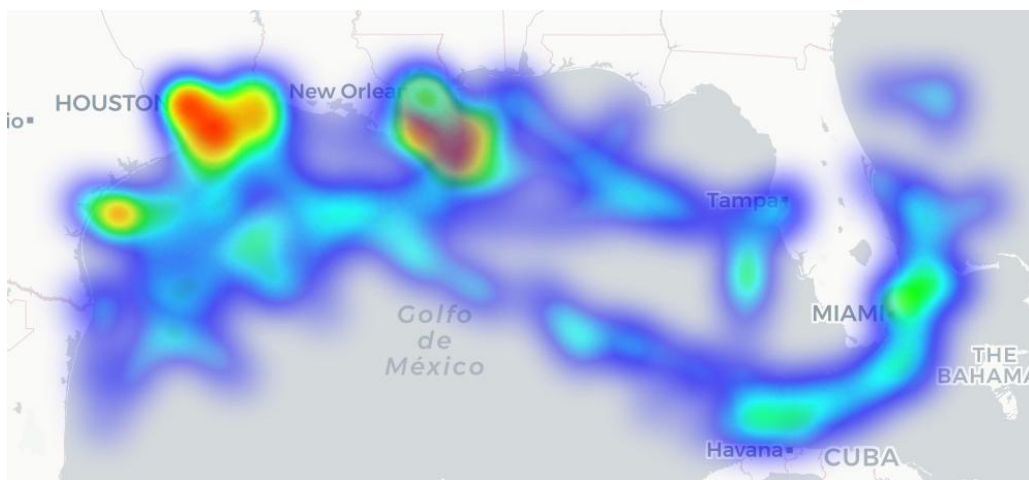


Figure 16 : Heatmap Tanker+Cargo

Pour affiner notre analyse, nous avons comparé deux heatmaps combinant à chaque fois les trajectoires des tankers avec un autre type de navire. Le choix de conserver les tankers dans les deux cas permet d'avoir une base visuelle constante :

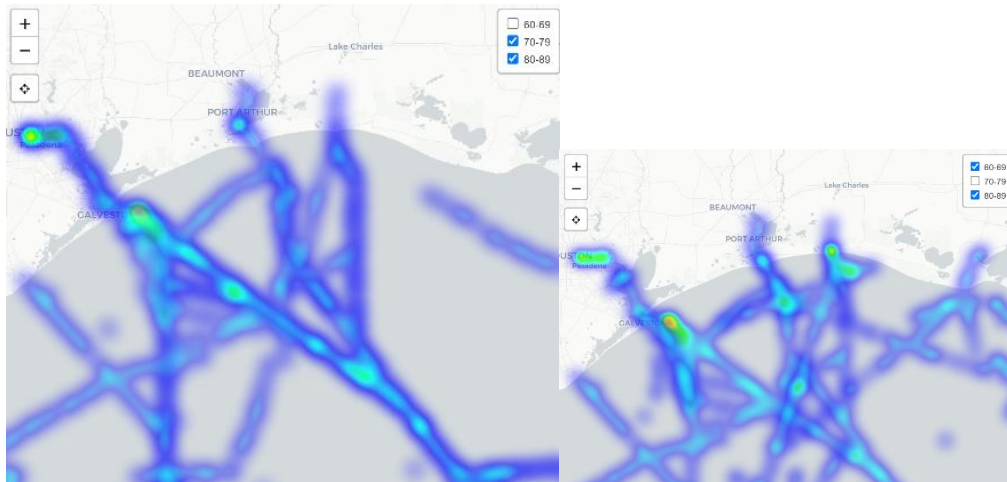


Figure 17: heatmap trajets tanker + cargo et tanker + passenger

Sur la carte Tanker + Cargo, on observe que les trajets des cargos se superposent presque parfaitement à ceux des tankers, ce qui révèle leur utilisation commune des routes commerciales maritimes. En revanche, sur la carte Tanker + Passenger, on distingue nettement d'autres zones de circulation, empruntées uniquement par les passagers : traversées fluviales, navettes locales... Cela confirme que les routes des navires passagers sont plus variées et moins sensibles aux itinéraires commerciaux classiques.

Fonctionnalité 4 : Etude des corrélations entre variables

Dans cette section, nous cherchons à identifier les relations linéaires existantes entre les variables numériques du jeu de données. L'objectif est double : repérer les variables qui évoluent de manière conjointe (corrélations fortes), et détecter d'éventuelles redondances ou dépendances pouvant orienter les analyses futures (réduction de dimensions, modélisation...).

Analyse des données quantitatives :

Pour amorcer l'analyse exploratoire, nous nous concentrons dans un premier temps sur les variables quantitatives. Afin de synthétiser les relations entre ces variables, nous utilisons une matrice de corrélation. Cet outil statistique permet de résumer l'ensemble des liens linéaires entre les variables numériques du jeu de données, facilitant ainsi la détection de dépendances, de redondances ou de structures sous-jacentes utiles pour les étapes d'analyse ultérieures.

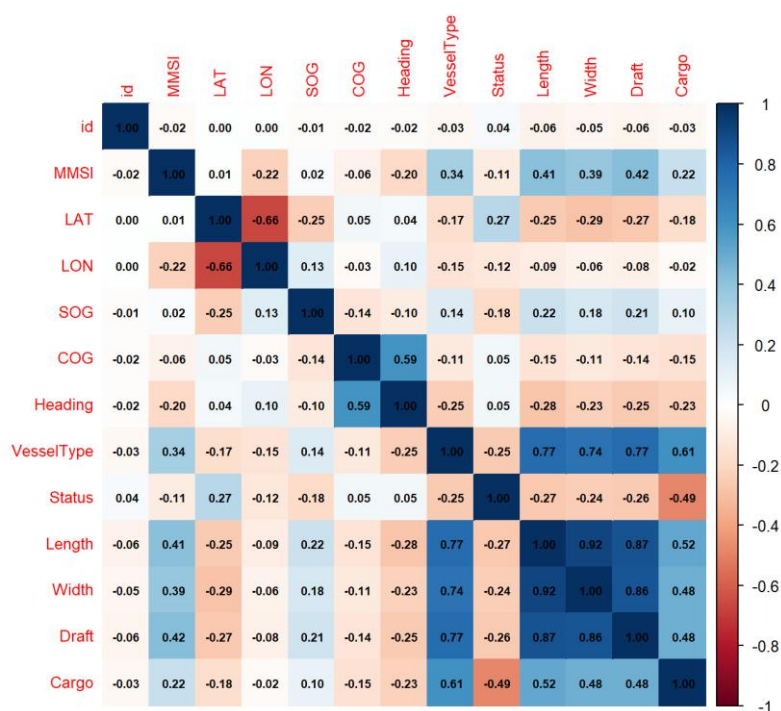


Figure 18 : Première matrice de corrélation

Cette matrice de corrélation est à critiquer, car elle inclut des variables qualitatives codées par des entiers, comme VesselType, Cargo ou Status. Or, ces codes n'ont a priori ni ordre logique, ni échelle de distance : Pearson les interprète alors à tort comme des valeurs continues, ce qui peut fausser les coefficients de corrélation.

Pour corriger cela, on pourrait envisager de vectoriser ces variables, mais cette transformation impliquerait d'abandonner la correspondance directe aux codes internationaux de l'AIS — et d'introduire une forme d'interprétation subjective.

Ainsi, la solution retenue est de reconstruire une matrice de corrélation nettoyée, en excluant les variables catégorielles, puis de traiter celles-ci à l'aide de méthodes plus adaptées (comme les tests d'indépendance du χ^2 ou les représentations de type mosaicplot).

Toutefois, corriger la matrice impose un travail d'analyse variable par variable. Et en s'intéressant plus précisément à la variable VesselType, on observe que ses indices semblent partiellement ordonnés : par exemple, les bateaux de passagers sont codés de 60 à 69, les cargos de 70 à 79, et les tankers de 80 à 89. Ce classement, bien qu'approximatif et non linéaire, reflète malgré tout un certain ordre croissant de taille des navires.

Dès lors, dans certains cas particuliers (comme une corrélation avec la longueur des navires), l'usage de ces codes dans la matrice de corrélation peut rester partiellement pertinent, à condition de l'interpréter avec prudence.

On propose ainsi une version corrigée de la matrice de corrélation, en y incluant tout de même la variable VesselType :

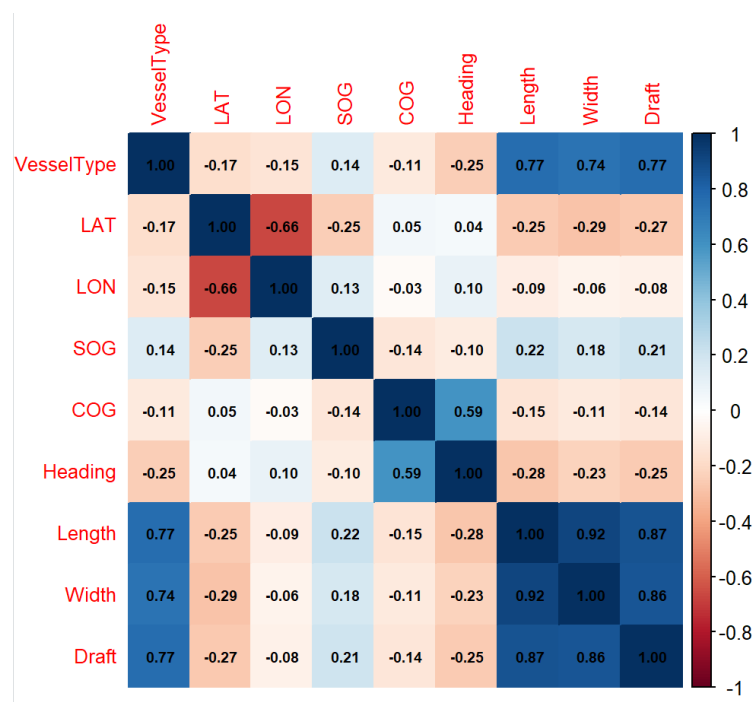


Figure 19 : Matrice de corrélation corrigée

Corrélation structurelle entre les dimensions des navires : plus un bateau est long, plus il a tendance à être large et profond (avec un plus grand tirant d'eau).

Corrélation géographique entre latitude et longitude : on remarque une corrélation négative assez forte entre la longitude et latitude. Ce qui signifie qu'en moyenne un bateau qui va vers l'ouest a tendance à aller vers le sud aussi. Lorsque l'on compare cette

donnée à la carte et aux routes maritimes, on remarque en effet que la route vers l'ouest impose de passer par le sud, entre Miami et La Havane.

Analyse des variables qualitatives

Nous passons maintenant à l'étude des variables qualitatives. Celles-ci n'ont pas pu être intégrées à la matrice de corrélation précédente, car le coefficient de Pearson nécessite des variables numériques continues, ce qui rend son application inadaptée à des données catégorielles.

Pour analyser les relations entre ces variables, nous utiliserons des tableaux croisés accompagnés de tests d'indépendance du χ^2 , qui permettent d'évaluer si deux variables qualitatives sont statistiquement liées. Les résultats seront ensuite représentés graphiquement à l'aide de *mosaicplots*, qui offrent une lecture visuelle claire des écarts à l'indépendance.

Vessel Type (groupé) vs Status

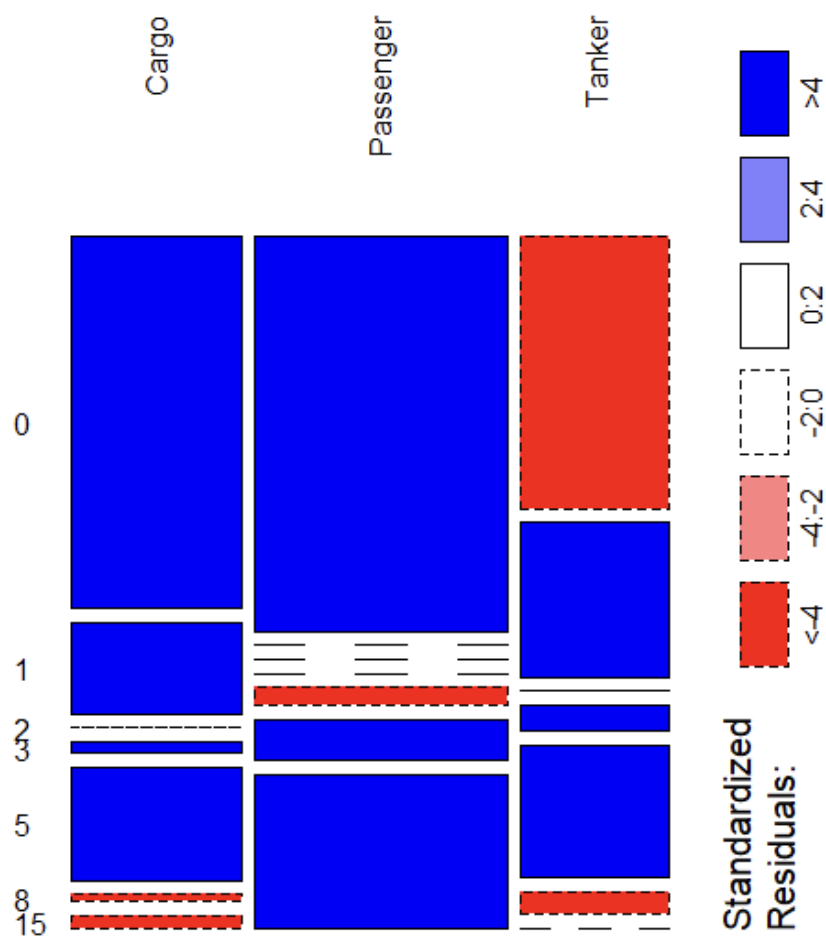


Figure 20 : MosaicPlot : VesselType et Status

Analyse de l'association entre VesselType et Status :

Afin d'examiner les relations entre variables qualitatives, nous avons utilisé un mosaicplot pour visualiser les interactions entre VesselType (type de navire) et Status (état de navigation). Ce type de graphique permet une lecture croisée des fréquences observées et attendues selon l'hypothèse d'indépendance, grâce à une représentation en rectangles colorés et proportionnels aux effectifs.

Résultats visuels et test du khi²

Le mosaicplot obtenu présente une distribution très contrastée :

- Des rectangles bleu foncé (résidus $> +4$) indiquent une surreprésentation marquée.
- Des rectangles rouge foncé (résidus < -4) signalent une sous-représentation très significative.

Ces écarts importants révèlent que certaines combinaisons VesselType $\times$ Status apparaissent bien plus (ou bien moins) souvent que ce que prévoirait un modèle d'indépendance. Par exemple :

- Les navires de type Passenger sont très fréquemment en navigation (Status = 0), bien au-delà de ce que le hasard anticiperait.
- Les navires de type Tanker, en revanche, sont nettement moins souvent en navigation et plus souvent à l'arrêt (Status 1 et 5).

Ces observations s'appuient sur des résidus standardisés, issus du test du khi², qui permettent de juger la significativité de l'écart entre fréquences observées et attendues. L'intervalle $[-4 ; 4]$ est une référence classique : au-delà de ± 2 , l'écart devient significatif ; au-delà de ± 4 , il est fortement significatif.

Conclusion

Le test du khi² et la visualisation par mosaicplot conduisent à rejeter l'hypothèse d'indépendance entre VesselType et Status. Ces deux variables sont donc fortement dépendantes, ce qui traduit des comportements spécifiques à chaque type de navire. L'échelle de représentation utilisée dans le graphique est appropriée, car les écarts observés sont effectivement très marqués, expliquant la faible présence de nuances intermédiaires.

Cette conclusion renforce l'idée qu'une analyse pertinente des données maritimes doit prendre en compte les interactions entre catégories, notamment dans les cas où les comportements opérationnels diffèrent selon le type d'embarcation.

Nous nous proposons d'examiner un second *mosaic plot* afin d'illustrer le cas d'indépendance entre deux variables qualitatives. Ici, le croisement concerne le *statut* du navire (*Status*) — c'est-à-dire s'il est en navigation, au mouillage, à quai, etc. — et le type d'émetteur-récepteur AIS qu'il utilise (*TransceiverClass*, classé en types A et B).

Ce croisement n'a a priori aucune raison de révéler une dépendance forte, car le statut opérationnel d'un navire à un instant donné n'est pas déterminé par son type de transpondeur AIS.

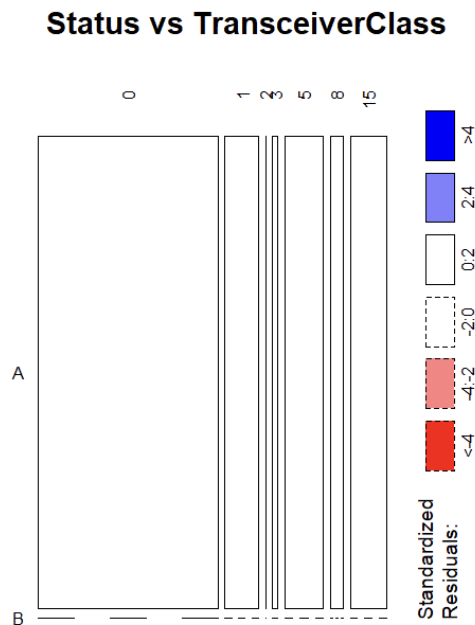


Figure 21 : *MosaicPlot : Status vs TransceiverClass*

Le *mosaic plot* résultant illustre bien cet état de fait. Tout d'abord, les couleurs des tuiles sont uniformes et pâles, indiquant que les résidus standardisés sont proches de zéro : cela signifie que les effectifs observés dans chaque croisement de modalités ne diffèrent pas significativement de ceux attendus si les deux variables étaient indépendantes.

En somme, cet exemple corrobore l'hypothèse d'indépendance entre *Status* et *TransceiverClass*, ce qui est attendu sur le plan logique et technique.

Analyse variables qualitatives × quantitatives

Après avoir exploré les relations entre variables numériques (via la matrice de corrélation) et entre variables qualitatives (à l'aide des tests du χ^2 et des mosaicplots), nous approfondissons ici les relations entre variables qualitatives et quantitatives. L'objectif est de vérifier si certaines caractéristiques numériques des navires (longueur, vitesse, tirant d'eau...) varient significativement selon des catégories comme leur type (VesselType), leur cargaison (Cargo) ou leur statut (Status).

Pour cela, deux approches complémentaires sont mobilisées :

- Des boxplots permettant une visualisation claire des distributions par groupe.
- Des tests d'ANOVA (analyse de variance) pour déterminer si les différences observées entre groupes sont statistiquement significatives.

Relation entre la longueur des navires et leur type (VesselGroup)

Dans un premier temps, nous avons analysé la variable *Length* (longueur des navires) en fonction du type de navire, regroupé en trois grandes catégories : Cargo, Passenger et Tanker. Le boxplot associé met en évidence des différences marquées entre les distributions de longueur pour ces trois groupes :

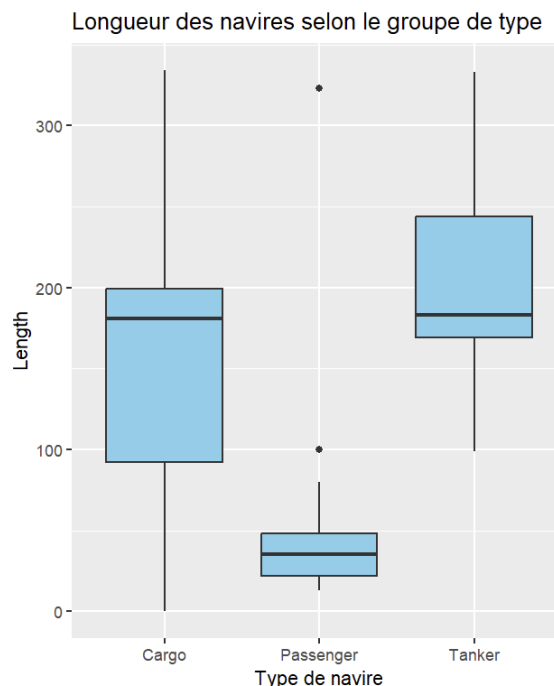


Figure 22 : Boxplot : Length vs VesselType

Les navires Passenger apparaissent comme étant les plus courts en moyenne, avec une médiane nettement inférieure à celle des autres groupes et une distribution concentrée sur les faibles valeurs. Les Cargos, quant à eux, présentent une plus grande variabilité, avec une médiane intermédiaire et des valeurs extrêmes allant des navires très courts à

d'autres très longs. Enfin, les Tankers (pétroliers) se distinguent par une médiane élevée et une dispersion importante, indiquant qu'ils sont globalement les plus longs parmi les trois catégories.

Pour confirmer visuellement cette différence, une analyse de variance (ANOVA) a été réalisée sur la variable *Length* en fonction du groupe de type de navire :

```

              Df    Sum Sq   Mean Sq  F value  Pr(>F)
VesselGroup    2 2.133e+09 1.066e+09   288285 <2e-16 ***
Residuals 401917 1.487e+09 3.699e+03
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
18297 observations effacées parce que manquantes

```

Figure 23 : Résultats analyse de variance ANOVA

Les résultats montrent une statistique de test très élevée ($F = 28\,285$) et une p-value extrêmement faible ($< 2e-16$). Cela signifie que l'on peut rejeter l'hypothèse nulle d'égalité des moyennes avec une confiance quasi absolue.

Ces résultats attestent statistiquement que la longueur moyenne des navires diffère significativement selon leur type : *le type de navire influence donc bien sa longueur.*

Fonctionnalité 5 : Prédiction / Régression

Dans cette partie, nous décrivons d'abord la mise en œuvre d'un modèle de régression logistique multinomiale pour prédire la variable qualitative VesselType à partir des attributs continus Length, Width, Draft et de la variable catégorielle Cargo, puis nous évoquons brièvement une tentative de prédiction de la vitesse (SOG), moins concluante, ainsi que l'analyse exploratoire par nuages de points.

Le jeu de données initial, issu du fichier `vessel-total-clean2.csv`, contient plusieurs centaines de milliers d'observations. Nous avons extrait un sous-ensemble de 50 000 lignes pour des raisons de temps de calcul, tout en conservant la distribution des niveaux de VesselType. Les variables VesselType et Cargo ont été converties en facteurs R avec un ordering naturel des niveaux.

Nous utilisons cette formule : `VesselType ~ Length + Width + Draft + Cargo`

70 % des observations ont été affectés au train set, les 30 % restants au test set, via la fonction `createDataPartition()`, on réalise également une validation croisée, en 5 folds chacun pour affiner la validité de notre régression en comparant les résultats.

```

      Reference
Prediction x60 x61 x69 x70 x71 x74 x79 x80 x82 x84 x89
x60      3346   71    0    0    0    0    0    0    0    0    0
x61       0    0    0    0    0    0    0    0    0    0    0
x69       0    0   96    0    0    0    0    0    0    0    0
x70       0    0    0 2585    0    0    0    0    0    0    0
x71       0    0    0    0 168    0    0    0    0    0    0
x74       0    0    0    0    0   57    0    0    0    0    0
x79       0    0    0    0    0    0 122    0    0    0    0
x80       0    0    0   29    0    0    0 2262    0    0   27
x82       0    0    0    0    0    0    0    0   40    0    0
x84       0    0    0    0    0    0    0    0    0   55    0
x89       0    0    0    0    0    0    0    0    0    0 137
> cat(sprintf("Test Accuracy : %.3f\n\n", cm_test$overall["Accuracy"]))
Test Accuracy : 0.986
```

Figure 24 : Matrice de confusion et précision de test

```

      Reference
Prediction x60 x61 x69 x70 x71 x74 x79 x80 x82 x84 x89
x60      7808  168    0    0    0    0    0    0    0    0    0
x61       0    0    0    0    0    0    0    0    0    0    0
x69       0    0   224    0    0    0    0    0    0    0    0
x70       0    0    0 6036    0    0    0    0    0    0    0
x71       0    0    0    0 395    0    0    0    0    0    0
x74       0    0    0    0    0 134    0    0    0    0    0
x79       0    0    0    0    0    0 286    0    0    0    0
x80       0    0    0   66    0    0    0 5278    0    0   72
x82       0    0    0    0    0    0    0    0   96    0    0
x84       0    0    0    0    0    0    0    0    0 129    0
x89       0    0    0    0    0    0    0    0    0    0 313
> cat(sprintf("CV Accuracy : %.3f\n", cm_cv$overall["Accuracy"]))
CV Accuracy : 0.985
```

Figure 25 : Matrice de confusion et précision des 5 folds

Nous observons une précision de 98,6%, et une matrice de confusion très concluante, on voit également que les valeurs sont très proches de celles trouvées avec les 5 folds, on en déduit que la régression est plutôt une réussite.

Dans un second temps, nous avons tenté d'expliquer la vitesse over ground (**SOG**) via une régression linéaire simple selon la formule :

$$\text{SOG} \sim \text{LAT} + \text{Length} + \text{Draft} + \text{Width}$$

RMSE	:	5.803
R ²	:	0.078

On déduit de ces résultats que la régression linéaire pour prédire le SOG n'est pas efficace ce qui fait sens car la matrice de corrélation montrait bien des corrélations très légères, et que de manière logique la vitesse est propre à elle-même et ne dépend pas de facteurs extérieurs.

Enfin nous réalisons, des nuages de points, croisant `VesselType` avec chacune des variables `Length`, `Width`, `Draft` et `Cargo`. Ces graphiques montrent des condensations de points mais ce n'est pas très concluant et ça limite l'information visuelle et confirme le besoin d'un modèle statistique plus structuré comme la multinomiale.

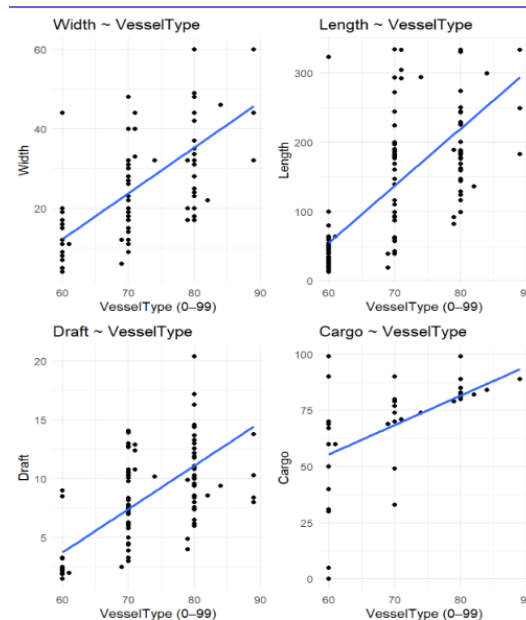


Figure 26 : nuages de points croisant `VesselType` et `Length`, `Width`, `Draft`, `Cargo`

Bonus diagramme de corde

L'utilisation d'un diagramme de corde permet de :

- D'intégrer en un coup d'œil l'ensemble des relations croisées entre ports, évitant la multiplication de graphiques individuels.
- De visualiser simultanément le volume global de trafic de chaque port (grandeur des arcs) et l'intensité spécifique de chaque lien (largeur des rubans).

Ce diagramme de corde représente, les flux de passages entre six ports clés du Golfe du Mexique. Chaque arc circulaire correspond à l'un des ports, et son étendue reflète le volume total de transitions (entrantes + sortantes) observées. Les rubans colorés relient un port « précédent » à un port « suivant », leur épaisseur étant proportionnelle au nombre de navires effectuant ce trajet précis.

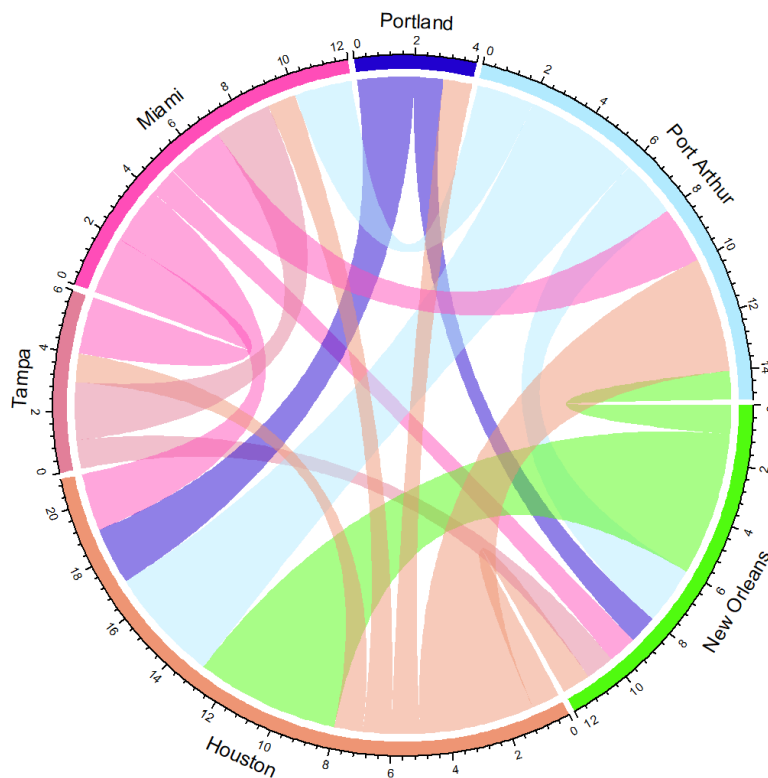
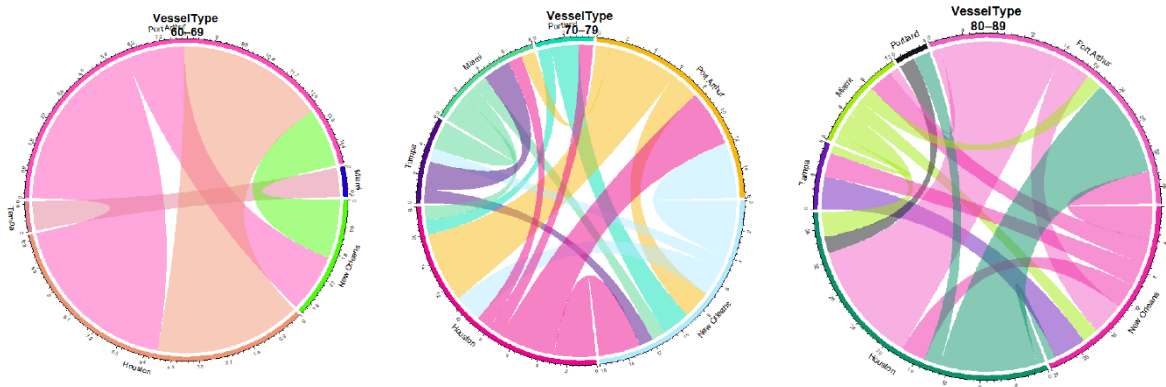


Figure 27 : Diagramme de corde des flux des passages de navires

Nous avons réalisé ce diagramme en récupérant les coordonnées de ces 6 ports sur internet, ensuite nous avons dit que les points proches à ces ports était passé par celui-ci, puis on regarde l'arrivée de ces mêmes bateaux, ce qui nous permet d'obtenir ce diagramme.

On affiche également des diagrammes de corde en fonction du type de bateau, on les range par intervalle (60,69), (70,79) et (80,89).



Voici nos 3 diagrammes de corde en fonction de VesselType, on remarque les ports principaux par type de bateau, par exemple les bateaux de passager utilise majoritairement port Arthur ou Houston, alors que les cargos ou tankers utilisent également beaucoup le port de Houston, on voit également que tous les imports de bateau passager de Tampa vont à Miami.

Conclusion

Ce projet d'analyse des données AIS dans le Golfe du Mexique a permis d'explorer les dynamiques complexes du trafic maritime à travers une approche méthodologique rigoureuse. L'étude a révélé une nette domination des navires passagers (60-69), cargos (70-79) et tankers (80-89), reflétant les trois piliers économiques de la région : tourisme, commerce et industrie pétrolière. Les visualisations graphiques et cartographiques ont mis en évidence des schémas de circulation distincts selon les types de navires, avec des routes privilégiées et des ports clés comme La Nouvelle-Orléans et Houston.

L'analyse des corrélations a montré des liens structurels entre les dimensions des navires, tandis que les tests statistiques (χ^2 , ANOVA) ont confirmé l'influence significative du type de navire sur ses caractéristiques opérationnelles. La modélisation par régression logistique multinomiale a atteint une précision remarquable (98,6%) pour prédire le type de navire, bien que la prédiction de vitesse se soit avérée moins concluante, soulignant la nature complexe et multifactorielle de cette variable.

Les limites rencontrées (données manquantes, valeurs aberrantes) ont été traitées systématiquement, garantissant la robustesse des résultats. Les diagrammes de corde ont offert une synthèse visuelle efficace des flux entre ports, complétant ainsi l'analyse quantitative.

Cette étude ouvre des perspectives pour des travaux futurs, comme l'intégration de données météorologiques ou économiques pour affiner l'analyse des routes commerciales. Elle démontre également l'utilité des données AIS pour optimiser la logistique maritime, améliorer la sécurité navigable et soutenir des décisions stratégiques dans cette région clé du transport maritime international.

Bibliographie

Pour exploiter les Cargos : <https://api.vtexplorer.com/docs/ref-aistypes.html>

Pour exploiter les Status : <https://documentation.spire.com/ais-fundamentals/how-to-interpret-navigational-status-codes/>