Noms : MARIE, DELAUNAY, LORRET-DESPRET

Prénoms : Corentin, Robin, Noé

Classe : A3

Titre du document : Projet - Big Data

Date et heure d'envoi : Lundi 16 juin 11h

Nombre de mots : 4172 mots

☑ En adressant ce document à l'enseignant, je certifie que ce travail est le mien et que j'ai pris connaissance des règles relatives au référencement et au plagiat.

Sommaire

Gantt de la semaine

Diagramme de Gantt - Projet électif Big Data

Liste des tâches à faire - Dates clés du projet - Répartition du travail

Membres de l'équipe (avec couleur attribué) :	Corentin	Noé (Scrum master)	Robin

Début du projet : 10/06/2025 — Fin du projet : 16/06/2025

TÂCHE	ATTRIBUÉE À	DÉBUT	FIN	Demi-journée 1	2	3	4	5	6	7	8 (WK)
Phase 0 : Fondamentaux du projet	Équipe entière	0	1								
Phase 1 : Fonctionnalité 1	Robin	2	3								
Phase 2 : Fonctionnalité 2	Corentin	2	3								
Phase 3 : Fonctionnalité 3	Noé	2	3								
Phase 4 : Fonctionnalité 4	Corentin	4	6								
Phase 5 : Fonctionnalité 5	Robin	4	6								
Phase 6 : Fonctionnalités subsidiaires	Noé	5	7								
Phase 7 : Écriture du rapport	Robin et Corentin	7	7								
Phase 8 : Préparation de la présentation + Soutenance	Équipe entière	8	8								

Introduction et mise en place du projet

Pour ce premier “sprint” du projet d’année 3, nous allons traiter de la partie Big Data qui consistera à analyser et traiter les données brutes qui sont l’entrée du sujet. Ce sont des données AIS, elles proviennent du système d’identification automatique des bateaux dans le golfe du Mexique. Nous allons devoir extraire les données de la database, visualiser ce grand volume de données sous différentes formes (histogrammes, map, etc.) et appliquer plusieurs modèles statistiques pour analyser/prédire.

Après avoir téléchargé le grand volume de données et créer un repository Git partagé, nous avons commencé à détailler les besoins à travers les différentes fonctionnalités qui nous sont demandées. Nous avons également creuser nos recherches sur les variables composant chacune de nos colonnes de la database (Documentation AIS : ref-navstat et coast.noaa.gov, pour les types codes notamment).

Dès le début, après avoir identifié clairement les attentes du projet et les différentes tâches/fonctionnalités sur lesquelles nous nous attelées, nous nous sommes réunis pour nous attribuer les tâches et sous-tâches découlant de notre analyse préliminaire et du cahier des charges. Pour formaliser cela, nous avons réalisé un diagramme de Gantt (Redirection vers la partie 8) qui témoigne de la structure que prendra notre organisation du projet, nous nous réservons le droit que celui-ci évolue en cours de projet car nous sommes en mode projet “Agile”, ce qui implique une certaine liberté dans la gestion du projet.
Pour nous conformer à la méthode SCRUM, que nous avons étudiée, nous avons déterminé un Scrum Master qui s’assurera de la bonne tenue des délais, du Gantt et de la bonne direction générale du projet pour atteindre les objectifs au cours des trois semaines à venir. Chaque début de session (demi-journée), nous allons faire 10 minutes de “scrum” (réunion rapide ayant pour but de “confronter” les participants au projet sur les tâches/problèmes dans le passé, en cours ou futurs), toujours conformément à la méthode SCRUM.

Description et exploration des données

En suivant rigoureusement le cahier des charges, il nous est demandé premièrement, pour la fonctionnalité 1 de décrire les données et de donner des statistiques descriptives univariées, pour cela nous nous servons simplement de la fonction summary(), intégré dans r-base. Cette fonction nous donne des informations descriptives sur l’ensemble du dataframe, le nombre de lignes, pour les variables numériques, on obtient le maximum/minimum, la moyenne, la médiane, les valeurs à différents quartiles.

La partie sur laquelle nous avons le plus travaillée sur la fonctionnalité 1, c’est le nettoyage de données. Nous nous sommes premièrement assuré que les valeurs numériques étaient considérées comme numérique, sinon nous les transformons, nous avons également transformé les “\N” en NA pour des questions de facilitation de gestion). Nous nous sommes profondément renseignés sur la nature de chaque variable pour comprendre comment réaliser nos choix au moment de nous séparer de données qui n’étaient pas représentatives (gestion des valeurs aberrantes). Voici les limites que nous avons choisis

pour chaque donnée (on notera que nous remplaçons tout ce qui est en dehors des limites par NA) :

- Latitude entre -90 et 90, et longitude entre -110 et 110. Nous avons fait le choix de ces bornes de manière arbitraire, cela correspond grossièrement à la zone du golfe du Mexique (dans les parties suivantes, il est possible que nous ayons réajusté ces bornes).
- SOG (Speed over ground) correspondant à la vitesse du navire en noeud. Nous avons constaté plusieurs navires qui avaient excessivement rapide nous avons donc ajusté la borne à 27 nœuds, une vitesse supérieure à cela relativement exagéré si l'on suit les recherches que nous avons réalisées. Nous n'avons pas fixé de borne inférieure (ou 0) car les arrêts d'un navire nous intéresserons pour faire travailler sur les trajets et les temps d'arrêts d'un bateau.
- COG (Course over ground) correspondant au cap du navire. Pour cela nos bornes correspondent à l'étendue d'un angle, donc entre 0 et 360. Nous avons remarqué la sur-présence de valeurs "511", après recherche cela correspond à une incapacité de savoir le COG du navire en question pour l'AIS, ces valeurs sont en dehors des bornes et seront donc traitées comme des NA (cela tombe bien car nous devons les traiter comme tel).
- Pour la longueur, la largeur et le tirant d'eau (Draft), nous nous sommes simplement assuré que les valeurs soient "raisonnables" (correspondance à la taille et la largeur min/max d'un bateau existant à la période d'enregistrement des données).
- Pour VesselType et Status, nous avons été consulté la documentation sur le site de l'AIS et nous avons appris que les typecodes (car ces variables sont sous formes numériques mais sont des typescodes) "0" pour VesselType et "15" pour Status corresponde à un NA, nous les remplaçons donc par NA.
- Nous avons creusé la documentation et réalisé de nombreuses observations dans les données sur le sujet de la variables Cargo. Il semblerait que ce soit également un typecode, un typecode qui correspond très très souvent à celui de VesselType, par exemple, on trouvera très régulièrement un code 80 VesselType accompagné de 83 en code Cargo. Si on suit les documentations, cela nous dit simplement que ce sont des Tankers. Nous avons donc décidé de procéder simplement : si l'une des deux valeurs est NA, nous attribuons à la variable jumelle le même typecode, si l'un des deux typecodes et plus "précis" (car un chiffre différent de 0 à l'unité donnent plus d'informations sur la nature du navire), nous attribuons à la variable moins "précise", le code de l'autre.

Nous avons également traité les valeurs "doublons" grâce à la fonction unique(), mais après constat qu'il n'y pas (ou très peu) de lignes qui se dédoublent (car la variable BaseDateTime combiné à MMSI nous permet de voir qu'un même bateau n'envoie pas deux fois les données à un même timing) nous avons décidé de retirer unique() car cela allonge fortement le temps d'exécution du programme de la partie 1.

Nous avons réfléchi à plusieurs possibilités pour traiter des NA que nous avions bien remplacé pour chaque variable. Nous pouvions simplement retirer les lignes avec des NA (mais cela retiré vraiment beaucoup de lignes à la base de donnée), nous pouvions remplacer par les valeurs médiane/moyenne pour les variables numériques, mais nous

avons choisi d'imputer des valeurs via Random Forest qui en corrélation avec les autres données de la ligne et en entraînant sur l'ensemble de la base de donnée, nous a permis de "remplir" nos valeurs NA par des valeurs "cohérentes" (au moins avec la base de donnée en l'état), et cela malgré que nous soyons conscient du "biais" que cela peut imputer à la pertinence de l'ensemble des données. Nous avons utilisé la fonction randomForest() de la librairie qui nous permet de réaliser le nombre d'arbres de décision voulus avec le nombre de feuilles voulues. Cela agit comme un système de vote qui met en relation un ensemble d'arbres qui acquiert une "expertise" sur un ou plusieurs choix en fonction des variables qu'on lui donne (le peu qu'il faut savoir c'est que les données prédites sont imputés selon les autres variables un peu à la façon d'une corrélation). Nous notons de plus que si une ligne contient plusieurs NA, nous substituons les NA par la médiane de la variable (pour les variables numériques) et la valeur mode de la variable (pour les valeurs catégorielles qui peuvent avoir des NA problématiques (Status)), cela temporairement, juste pendant l'entraînement du modèle avec Random Forest.

Dernièrement, nous notons que nous ne nous occupons pas des NA pour les valeurs suivantes : IMO, CallSign et VesselName. La raison est que cela aura peu d'importance pour la suite du projet (prédictions, etc…) car la variable MMSI est toujours présente et suffit à elle seule pour "traquer" un navire sur l'ensemble des données que l'on reçoit de celui-ci.

Avec les choix que nous avons réalisés, nous ne supprimons donc que très peu de ligne (moins de 2%). Malgré un léger biais qui s'installe, cela nous permet de préserver l'intégrité des données.

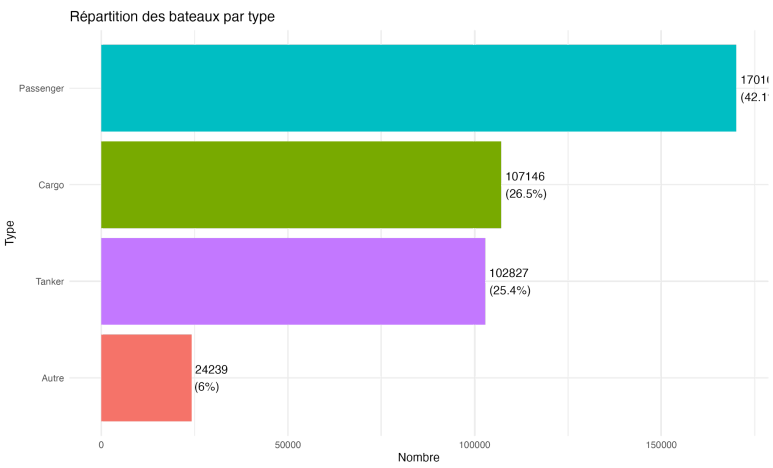
Visualisation des données sur des graphiques

La fonctionnalité 2 a pour objectif de faire des représentations graphiques des données pour nous permettre de visualiser et comprendre la distribution des données AIS dans le golfe du Mexique. Cette étape va nous permettre de mieux comprendre et de trouver des éléments incohérents ou cohérents sur lesquels se pencher pour l'analyse des données. Grâce à cette visualisation nous allons comprendre les caractéristiques des différents types de navires (Cargo, Tanker, Passager).

Avant de commencer à afficher des données suivant les types de navires, il fallait les classer en fonction de leur type ("VesselType" dans le .csv). Avec l'aide d'une datasheet sur l'indication du numéro associé à un type de navire (source: Marine Traffic), on sait que les navires "Passenger" ont un VesselType entre 60 et 69, les "Cargo" entre 70 et 79, et les "Tanker" entre 80 et 89.

Comme le fichier .csv est composé principalement de ces types de navires, nos visualisations se porteront sur "Passager", "Cargo", "Tanker". De plus, comme second filtre pour ces données et la limitation au golfe du mexique, on fait un filtre suivant les données en longitude/latitude du golfe pour être sûr de ne pas analyser des navires d'autres régions.

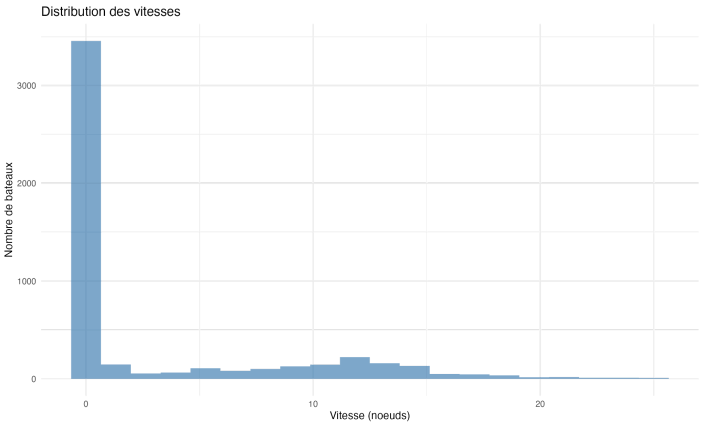
Pour la première visualisation on va d'abord regarder le nombre de navires par type, ça va déjà nous donner un ordre d'idée des principaux types de navires dans le golfe du Mexique.



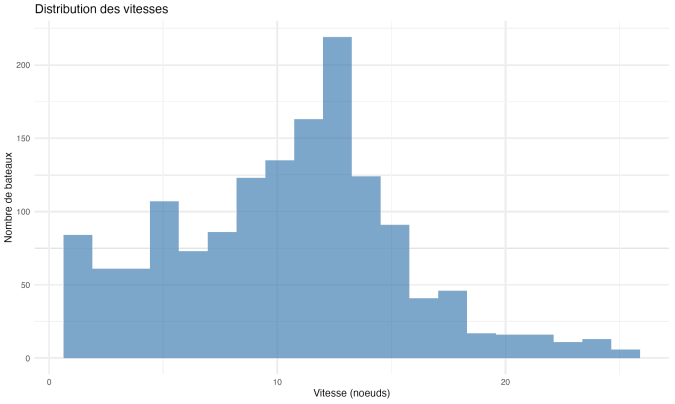
Comme on peut le voir sur le graphique, on constate que le nombre de navires Passager est très élevé ce qui montre beaucoup de tourisme dans la région. Il serait par exemple intéressant de comparer ce graphique à un autre moment (1 an après), ce qui pourrait nous montrer si les pays au bord du golfe du Mexique sont en déclin touristique ou non.On voit aussi qu'il y a environ une égalité de nombre entre les Cargo et Tanker - des données intéressantes si on les compare à un autre moment (été-hiver) ce qui nous montrerait si le marché fonctionne tout autant dans cet endroit du monde pendant ces périodes différentes. Il faut noter que ce graphique montre le nombre de lignes dans le jeu de données étant passager/cargo/tanker/autre. À savoir qu'un seul navire apparaît environ 2500-3000 fois dans le fichier à cause du système AIS qui enregistre les positions régulièrement.

La vitesse

Ce graphique montre une forte concentration de navires à une vitesse nulle ou très faible (≤1 nœud), ce qui est normal si on prend en compte les arrêts des croisières, les attentes au port, etc. Il faut savoir qu'il y a peu de navires en mer à 0 nœud exact car il y a toujours une dérive à cause de la houle. Ensuite on observe une distribution plus étalée pour les vitesses de navigation normale (5-20 nœuds).

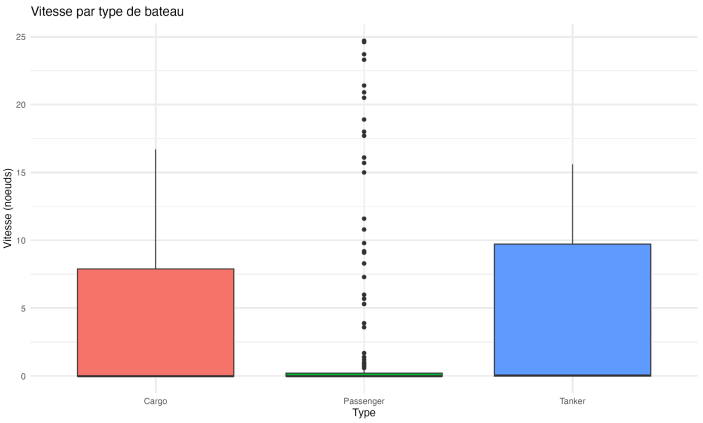


Pour mieux pouvoir voir la distribution des navires en mouvement, on a fait un deuxième graphique excluant les navires à <1 nœud.



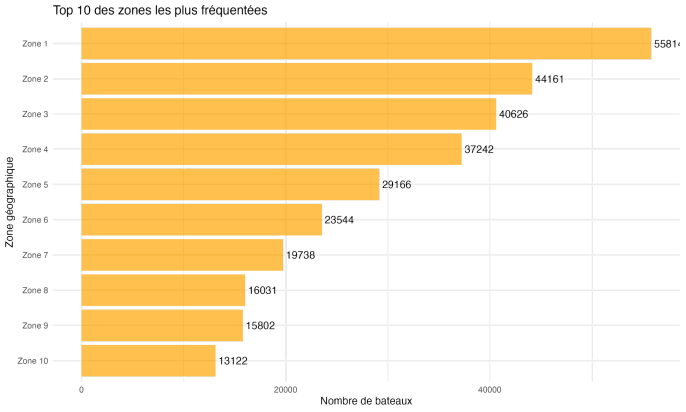
Vitesse par type de navire

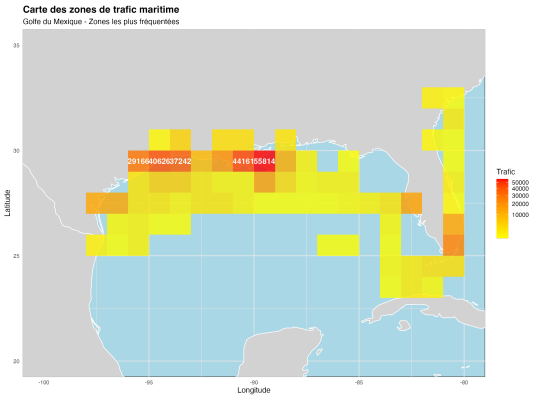
Le graphique nous montre des résultats intéressants : les navires les plus souvent à l'arrêt sont les "Passenger" mais c'est aussi ceux qui vont le plus vite. C'est dû au temps mort en port ou proche de la côte pour observer et profiter du pont extérieur sans avoir la nuisance du bruit et sans la pollution des cheminées. On voit que les Cargo ont environ 6 nœuds qui représente le compromis entre efficacité et économie de carburant. Les Tankers font des longs trajets mais font des pauses pour les opérations de chargement/déchargement, ce qui explique leur profil de vitesse particulier.



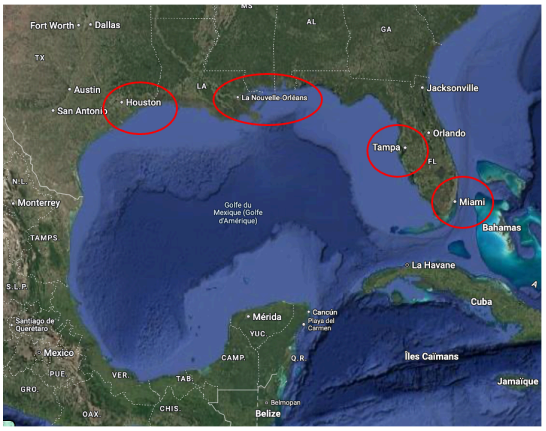
Analyse géographique du trafic

Pour identifier les zones de forte activité maritime, on a divisé le golfe du Mexique en grilles de 1 degré x 1 degré, et compté le nombre de navires dans chaque zone. Pour connaître où se trouvent les ports, nous cherchons les 10 zones les plus fréquentées.





Le graphique nous donne ces résultats, et pour associer ces zones à une carte, on a la "Carte des zones de trafic maritime" qui l'accompagne.
Nous voyons bien que les zones les plus fréquentées sont possiblement des endroits où il y a des ports, et on confirme ces données grâce à une vraie carte des villes portuaires du golfe. Grâce à ça on peut faire un classement des villes portuaires du golfe du Mexique les plus fréquentées:

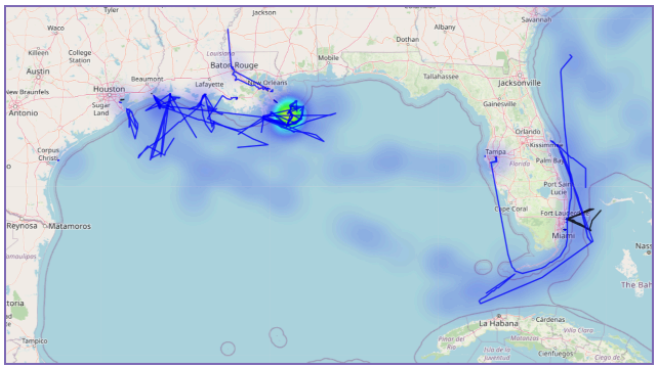
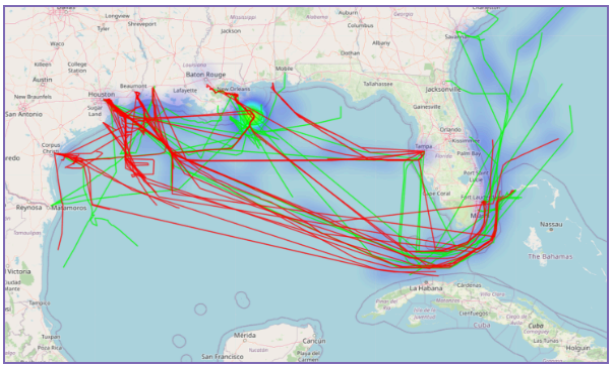


- La Nouvelle-Orléans
- Houston
- Miami
- Tampa

Les graphiques nous montrent aussi, en plus des ports les plus fréquentés, les zones offshore les plus fréquentées (sûrement les plateformes pétrolières). Les visualisations de cette partie sont les prémices des graphiques plus complexes des prochaines fonctionnalités du projet.

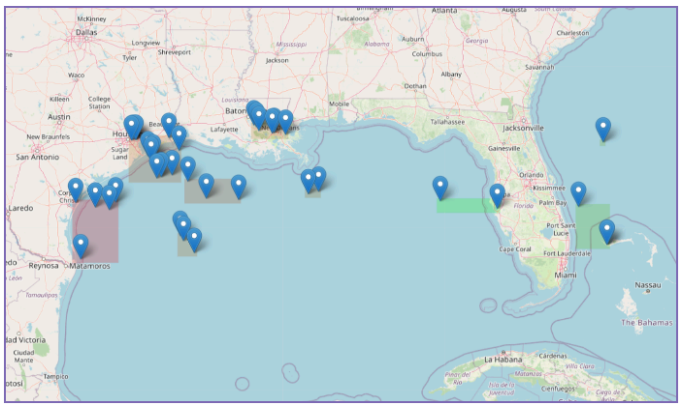
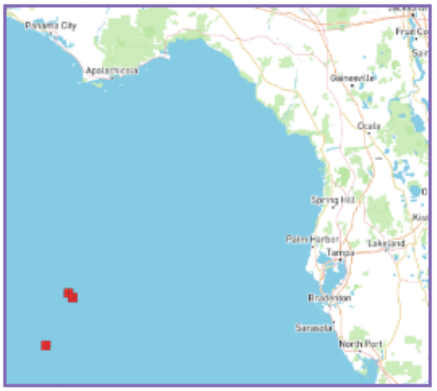
Visualisation des données sur une carte

Nous allons maintenant pousser plus loin l'étude des données en observant si des tendances se remarquent sur une carte.
Tout d'abord, on remarque une importante différence entre les trajets des navires de passagers / de plaisance (gauche) et les navires de cargaisons / pétrole (droite)

Les navires de passagers ont tendance à faire des petits trajets proches de côtes, non loin des ports principaux. À l'inverse, les navires de cargaisons et pétroliers réalisent des trajets réguliers et plus longs d'un port à l'autre, suivant des routes fixes. On voit facilement se dessiner ici des routes commerciales.

Une autre observation que l'on peut tirer de l'observation des cartes, plus spécifiquement pour les navires de cargaison/pétroliers, est leur chargement déchargement. On ne possède malheureusement pas directement cette information, mais il est possible d'en faire une estimation en comparaison le "draft" (niveau du bateau sous la mer) d'un bateau.

On voit que des changements de draft ont fréquemment lieu au niveau des ports, ce qui nous permet de déduire si les ports sont plutôt des ports de chargement (en vert) ou de déchargement (en rouge). On peut même noter que certains changements de draft en pleine mer, de la part des pétroliers, correspondent à des gisements de pétrole !

Étude des corrélations entre variables

Cette fonctionnalité va permettre de quantifier les relations entre les variables du dataset AIS. Ça va nous permettre de comprendre les interactions entre les caractéristiques des navires et leur comportement.

Dans cette partie il y aura:

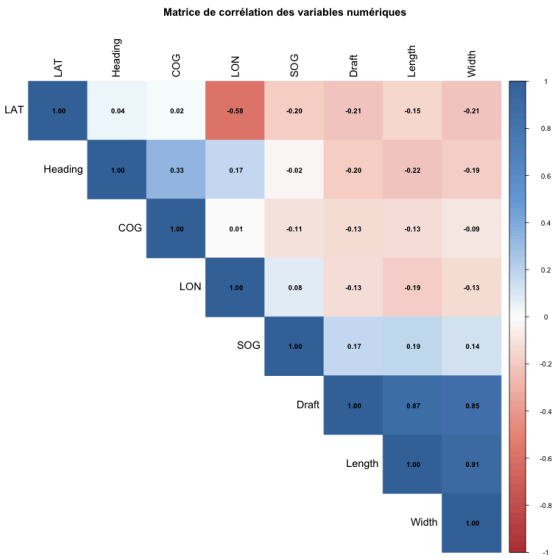
- Matrice de corrélation pour les variables quantitatives
- Analyse bivariées avec visualisations
- Tests d'indépendance du Chi² pour les variables quantitatives
- ANOVA pour les relations quantitatif/qualitatif

Corrélation

Pour faire les corrélations entre les variables quantitatives on utilise la formule de Pearson, qui mesure la force et la direction de la relation linéaire:

$$pearson = \frac{Cov(X,Y)}{\sqrt{Var(X)} \times \sqrt{Var(Y)}}$$

Matrice de corrélation



On a fait la matrice de corrélation pour les valeurs quantitatives pour 8 variables:

- SOG: vitesse du navire
- Length: Longueur du navire
- Width: Largeur du navire
- Lat/Lon: coordonnées
- Draft: tirant d'eau
- Heading/COG: Cap et route

Comme on peut le voir sur la Matrice de corrélation des variables numériques, celle-ci révèle plusieurs corrélations importantes:

Entre Length et Width: r=0.91 : c'est une corrélation très forte ce qui montre une proportionnalité dans les dimensions des navires. C'est logique du point de vue de la construction navale.

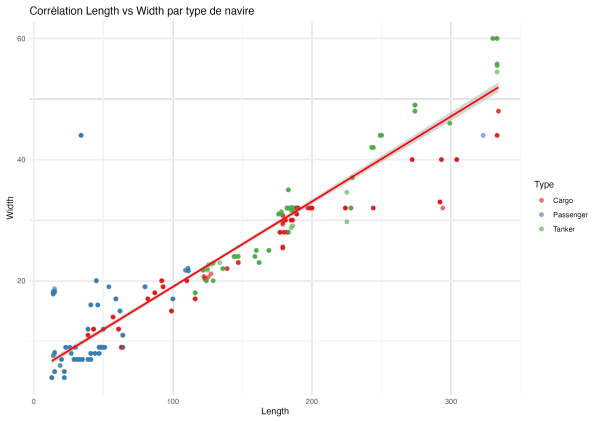
Length et Draft: r=0.85 : ça montre que les navires les plus longs ont souvent un tirant d'eau plus important (inverse pour les navires plus petits). C'est cohérent avec les contraintes de stabilité.

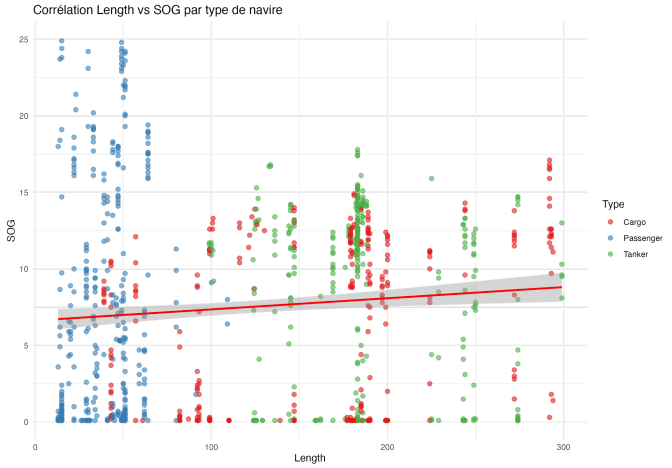
Width et Draft: r=0.87 : qui rejoint la corrélation précédente.

Par contre, un exemple de corrélation faible est le SOG et Length, ce qui montre que la vitesse ne dépend pas de la taille du navire mais plutôt d'autres facteurs opérationnels.

Analyse bivariées

Les graphiques nous donnent plus de détails:





Length vs Width: pour les Tankers et les Cargo c'est très corrélé ce qui montre qu'ils ont plus des tailles standards, alors que les Passagers sont beaucoup plus variables ce qui montre que les navires de plaisance n'ont pas de normes strictes sur la longueur/largeur à suivre.

Pour la vitesse il n'y a pas de corrélation (on le savait déjà), mais le graphique d'analyse bivariée montre par exemple la variété de différences de vitesse pour les navires Passenger et qu'il y a comme des lignes qui se forment, ce qui montre que beaucoup de navires d'un même type ont le même Length. Pour les passagers le length est environ entre 20 et 70 mètres, pour les tankers et cargo c'est uniquement au-dessus de 70 mètres et jusqu'à 300 mètres pour les plus gros navires.

Les résultats des corrélations par type sont:

Cargo :

SOG_Length : 0.266
SOG_Width : 0.213
Length_Width : 0.956

Tanker :

SOG_Length : –0.016
SOG_Width : –0.021
Length_Width : 0.974

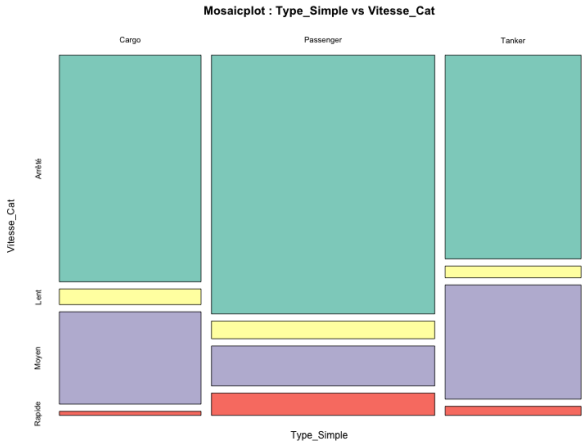
Passenger :

SOG_Length : 0.148
SOG_Width : 0.007
Length_Width : 0.457

Test du Chi² pour faire les corrélation entre 2 variables qualitatives.

Par exemple nous allons prendre la vitesse où on a créé des catégories:
0<SOG<1 : arrêté
1<SOG<5: lent
5<SOG<15: moyen
15<SOG<25: rapide

Si nous faisons la corrélation entre le type de navire et le Vitesse_Categorie que nous venons de créer, on obtient avec le plot Mosaicplot



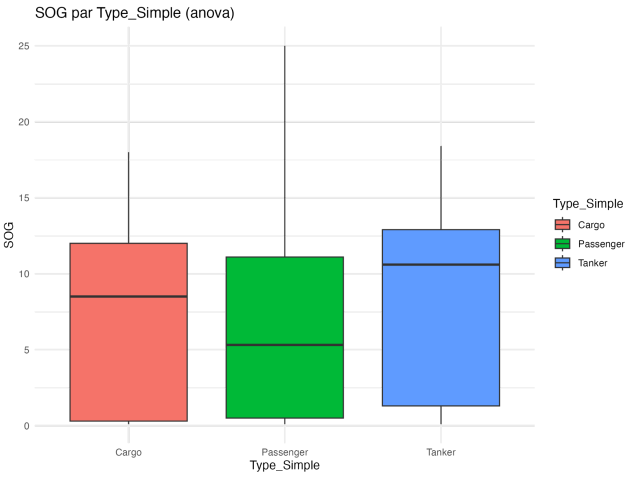
Le mosaicplot confirme nos observations sur le fait qu'il y a beaucoup de navires Passenger arrêtés, que très peu de Tankers et Cargo sont rapides, etc.

On obtient un Chi² = 24044.17 pour tout le dataset et = 250 pour 5000 lignes aléatoires (pour faire un d'échantillon) avec une p-value proche de 0, ce qui veut dire qu'il y a une forte dépendance entre le type du navire et la vitesse.

ANOVA

L'ANOVA permet de tester l'égalité des moyennes entre groupes:

$$F = \frac{MSB}{MSW}$$



Le graphique nous montre les résultats: p-value proche de 0. et F = 30 pour un échantillon de 5000 navires. Vitesses moyennes: Passenger = 8.2 nœuds, Cargo = 6.1 nœuds, Tanker = 4.3 nœuds. A savoir que plus F est élevé, plus les groupes sont différents.

Ces différences de vitesses moyennes s'expliquent par les différents modes opérationnels de chaque type de navire. Les navires à passagers alternent entre des phases d'arrêt (observation, escales) et des transits rapides, d'où une moyenne élevée. Les cargos optimisent le rapport vitesse/consommation, tandis que les tankers privilégient la sécurité avec des vitesses plus modérées.

Prédictions de VesselType

De nouveau pour cette partie du cahier des charges, nous avons réfléchi à plusieurs solutions pour réussir au mieux à “prédire” la nature du VesselType.

Cela à tout d’abord passé par plusieurs réflexions :

- Certaines variables sont d’offices à exclure car elles n’ont pas de corrélation logique (fondamentalement ou en l’état). C’est le cas pour les suivantes que nous avons décrété non pertinentes : id, MMSI, IMO, CallSign, VesselName.
- Nous avons après plusieurs constats au cours d’essais, de supprimer COG, Heading, LAT et LON (nous reviendrons juste après sur le cas de ces variables, mais, en l’état, ligne par ligne, elles ne sont, selon nous, pas pertinentes à mettre en corrélation)
- Il n’y que 4 bateaux (4 MMSI donc) sur l’ensemble des données de la base qui n’ont pas de VesselType entre 60 et 89 (Passengers, Cargos, Tankers), il est donc inutile de chercher à les prédire car cela va juste perturber le modèle de prédiction (on

appelle cela le déséquilibrage de classe pour l'entraînement du modèle dans ce cas présent).

- Nous notons, comme nous l'avons dit dans la partie, que étant donné que Cargo suit très fortement le typecode de VesselType, c'est une variable légèrement "trichée" car elles sont fortement similaires (en tout cas on constate qu' avec Cargo on a une très bonne prédiction et sans Cargo on a une très mauvaise prédiction (on verra que ce n'est pas une fatalité)).

Désormais que nous avons établis ces quelques observations, nous avons un ensemble de variables qui peut-être, avec une certaine combinaison, pourrons nous permettre de prédire au mieux VesselType. Ainsi, nous avons programmé le choix de notre modèle de prédiction de la manière suivante (version simplifiée de notre programme):

- Nous testons toutes les combinaisons possibles avec nos variables choisies pour prédire VesselType par une régression multinomiale.
- Ces différents entraînements de modèle se font sur 80% des données pour l'entraînement et les 20% restants pour le test du modèle.
- On calcule ensuite la précision du modèle sur les données test (souvent relativement bon dû à l'overfitting, donc le fait que le modèle soit très entraîné sur les données testées).
- On réalise un classement des modèles avec les meilleures précisions de prédiction. Si on garde Cargo, la meilleure combinaison est SOG + Cargo avec ~87% de précision sur les données tests. Si on retire Cargo, la meilleure combinaison est Status, Longueur, Largeur et Tirant d'eau avec ~73% de précision.
- Après avoir choisi le modèle final (celui qui est premier au classement, donc la meilleure précision), nous échantillonnons notre base de donnée avec la fonction sample_n(), cela nous permet de "prélever" un échantillon aléatoire (quasiment, selon la seed fixée préalablement) qui nous permet de re-tester notre modèle mais sur des données "quelconques". Cela démystifie l'overfitting du modèle est nous donne les résultats finaux suivants pour notre fonctionnalité 5 :
    - Pour la combinaison Status + Longueur + Largeur + Tirant d'eau, on obtient 31.8% de précision "optimisé" sur un test "réel".
    - Pour la combinaison Cargo + SOG (donc avec Cargo), on obtient ~30.5% de précision "optimisé" sur un test "réel". Il semblerait que malgré la présence de la variable Cargo, la précision soit basse mais on peut imputer la responsabilité au fait qu'il n'y a que 2 valeurs pour prédire.

Nous ajoutons également qu'étant donné le caractère aléatoire du "sampling" au moment de réaliser l'échantillon de test "réel", d'une seed à l'autre choisie, les résultats seront différents.

Nous aurions aimé pousser la prédiction beaucoup plus loin. Vous retrouverez dans le rendu du code, des programmes R dans le dossier "advancedFeatures" plusieurs fichiers R qui ont pour but de calculer des super paramètres à partir des données brutes pour gagner grandement en pertinence/information et notamment pouvoir intégrer les 4 données que nous n'avons pas intégré à la corrélation (LAT, LON, COG, Heading), que nous avons utilisé dans notre programme "trajectorySegmentation.r", dans lequel nous avons formé un csv qui "segmente" les trajets des navires de manières ordonnées, temporellement et géographiquement, cela, nous permettant de connaître les temps d'arrêts, les moments

d’arrêts, la distance des trajets, la vitesse de croisière moyenne, le temps d'accélération ou même les routes empruntées et cela en fonction du VesselType.

Nous avons également composé dans le fichier “computingEnhancedParameters.r”, un coefficient entre la largeur et la longueur d’un navire, ce qui est plus représentatif que simplement les valeurs séparées mises en corrélation, car plus généralement les Tankers auront des coefficients entre 8-12, les Cargos environ 5-6 et les Passengers encore moins.

Dans le fichier “offShoreDraftVariation.r”, nous avons abouti à pouvoir savoir si un navire se stoppe en plein milieu du golfe du Mexique et se charge ou se décharge, un paramètre qui nous permettrait de détecter avec une grande précision si les données transmises d’un navire, nous permettrait de prédire un Tanker, car ces navires s’arrêtent souvent à des plateformes pétrolières (nous avons utiliser des librairies (geosphere, rnaturalearth) qui nous permettent, combinées ensemble, de savoir si une position géographique LAT-LON correspond à un point de terre ferme).

Nous avons aussi combiné deux de nos programmes dans notre programme “coastDistanceTrajectory.r” pour avoir la distance moyenne d’un navire à la terre ferme sur un “segment” (trajet d’un navire). Cela pourrait nous permettre de détecter avec une bonne précision si un navire est un Passenger, car après constat grâce à notre fonctionnalité 3 et 3+, on peut constater que les navires de croisières font des trajets proches des côtes (ce fait est réputé).

Malheureusement, le temps qui nous est donné pour la réalisation du projet ne nous permettra pas de finir de mettre en application notre modèle “plus poussé” de prédiction (certains membres de l’équipe ont pour désir de continuer de travailler le projet en dehors de la période accordée). Nous aurions aimé pour l’implémentation du modèle poussée, d’améliorer grandement la recherche du modèle optimisé via les techniques suivantes :

- Test de plusieurs modèles de prédictions et comparaisons entre eux, voire même des combinaisons et pondérations de ces modèles : RandomForest, régression logistique, XGBoost, SVM.
- Test de plusieurs configurations d’hyperparamètres pour chacun de ces modèles.
- Gestion des déséquilibrages des classes (Cas quand notre modèle pour s’entraîner en masse sur des cas de Cargo et moins sur Tankers ou Passengers, ce qu'il encouragerait à prédire par Cargo).
- Test de la solution SMOTE avec la validation croisée à plusieurs plis.
- Plusieurs occurrences de train/test pour écarter des résultats de précisions chanceuses ou malchanceuses.
- Validation de robustesse améliorée et visualisation automatique.