

Big Data A3

Analyse des données
maritimes dans le golfe du
Mexique



Exploration de la Data Base



Description des données

On utilise les fonctions `glimpse()` et `head()`

st	nmz	longitude	lat	lon	gdp
41	France	12.141	45.76	12.141	64.4
42	France	12.141	45.76	12.141	64.4
43	France	12.141	45.76	12.141	64.4



Statistiques descriptives

On parvient à analyser les données grâce à `summary()`

LAT	lon	gdp
Min. :23.39	Min. : -118.19	Min. : 0.000
1st Qu.:27.81	1st Qu.: -84.43	1st Qu.: 0.000
Median :29.28	Median : -90.21	Median : 0.000
Mean :29.72	Mean : -90.09	Mean : 4.534
3rd Qu.:29.73	3rd Qu.: -89.07	3rd Qu.: 8.800
Max. :36.69	Max. : -75.00	Max. :200.000

Nettoyage des données

```
variable <dbl>
#> [1] 1.0000000000000000e+000
#> [2] 1.0000000000000000e+000
#> [3] 1.0000000000000000e+000
#> [4] 1.0000000000000000e+000
#> [5] 1.0000000000000000e+000
#> [6] 1.0000000000000000e+000
#> [7] 1.0000000000000000e+000
#> [8] 1.0000000000000000e+000
#> [9] 1.0000000000000000e+000
#> [10] 1.0000000000000000e+000
#> [11] 1.0000000000000000e+000
#> [12] 1.0000000000000000e+000
#> [13] 1.0000000000000000e+000
#> [14] 1.0000000000000000e+000
#> [15] 1.0000000000000000e+000
#> [16] 1.0000000000000000e+000
#> [17] 1.0000000000000000e+000
#> [18] 1.0000000000000000e+000
#> [19] 1.0000000000000000e+000
#> [20] 1.0000000000000000e+000
```



Description des données

On utilise les fonctions `glimpse()` et `head()`

	id	MMSI	BaseDateTime	LAT	LON	SOG	COG
	<dbl>	<dbl>	<dtm>	<dbl>	<dbl>	<dbl>	<dbl>
1	82	636017833	2023-05-29 00:00:06	29.1	-89.3	13.4	228.
2	84	367387350	2023-05-29 00:00:01	25.8	-80.1	5.6	176



Statistiques descriptives

On parvient à analyser les données grâce à `summary()`

LAT		LON		SOG	
Min.	:23.39	Min.	:-118.19	Min.	: 0.000
1st Qu.:	27.91	1st Qu.:	-94.43	1st Qu.:	0.000
Median	:29.28	Median	: -90.21	Median	: 0.000
Mean	:28.72	Mean	: -90.09	Mean	: 4.534
3rd Qu.:	29.73	3rd Qu.:	-89.07	3rd Qu.:	6.800
Max.	:36.69	Max.	: -75.00	Max.	:102.300

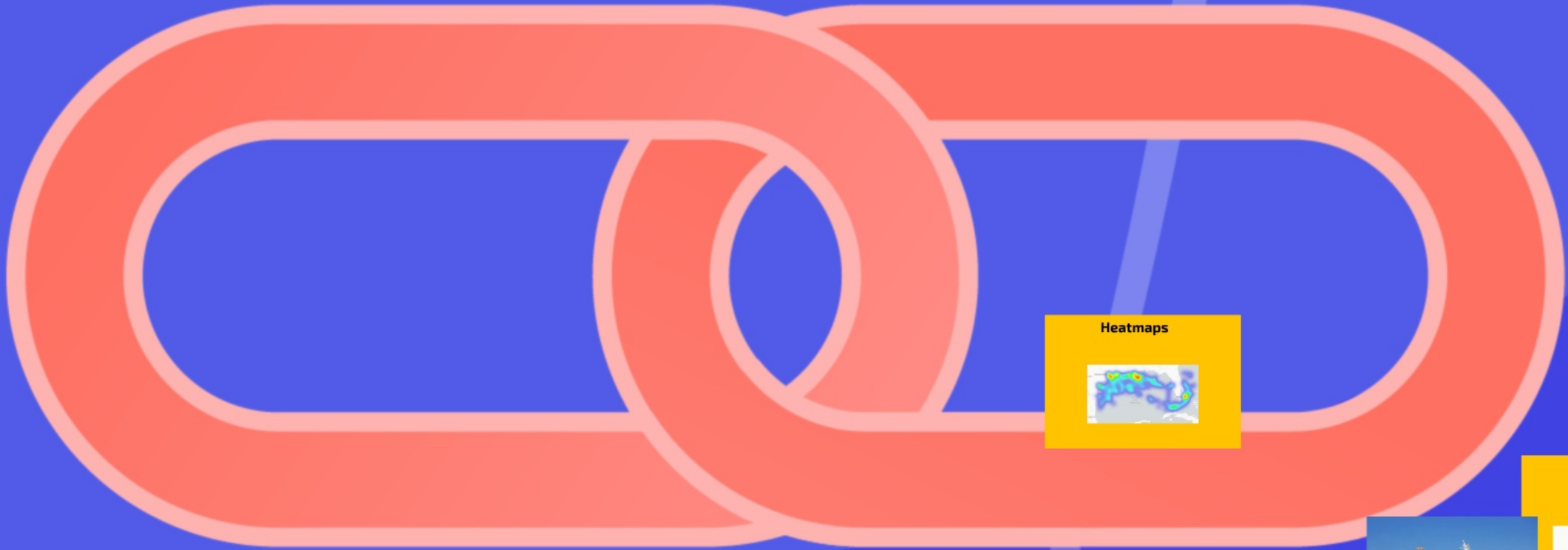
Nettoyage des données

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

Big Data A3

Analyse des données
maritimes dans le golfe du
Mexique





Heatmaps

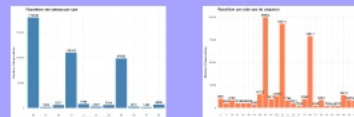


Trajectoire des bateaux



Représentation graphique

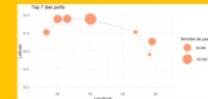
Histogramme 1.



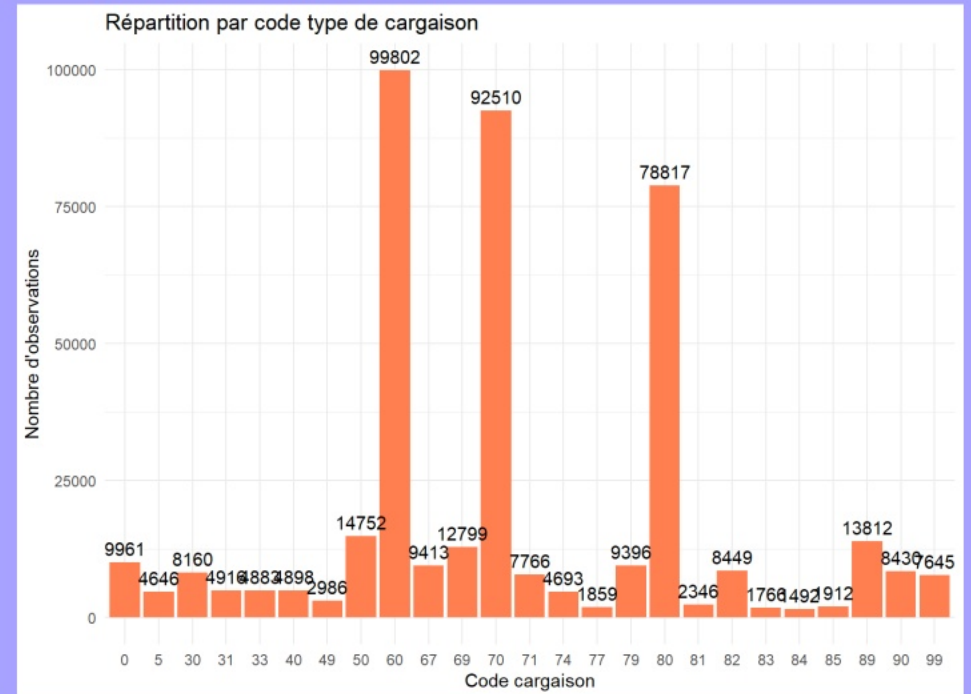
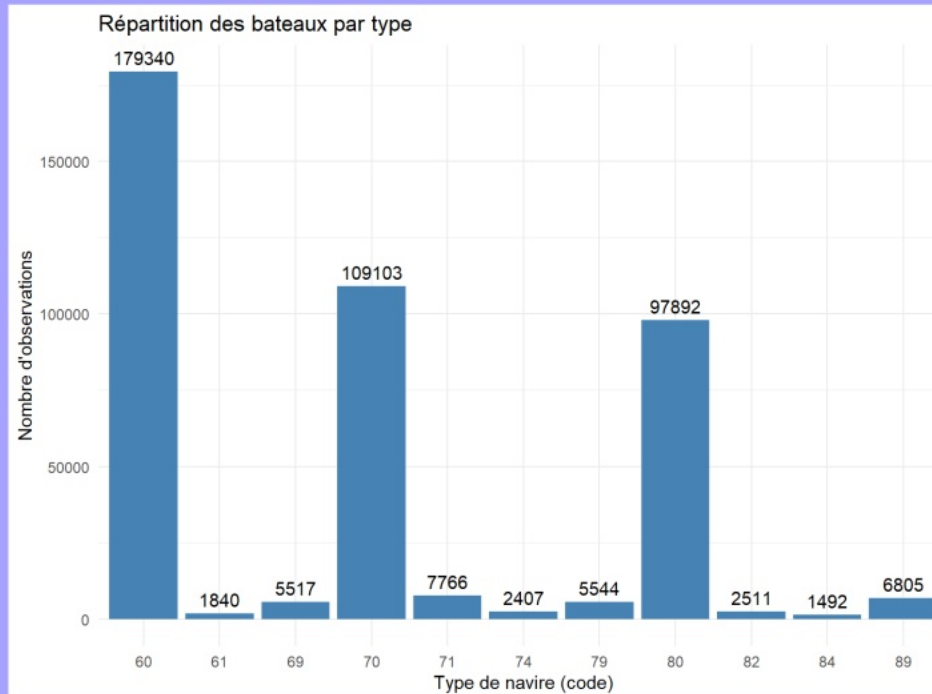
Histogramme 2.



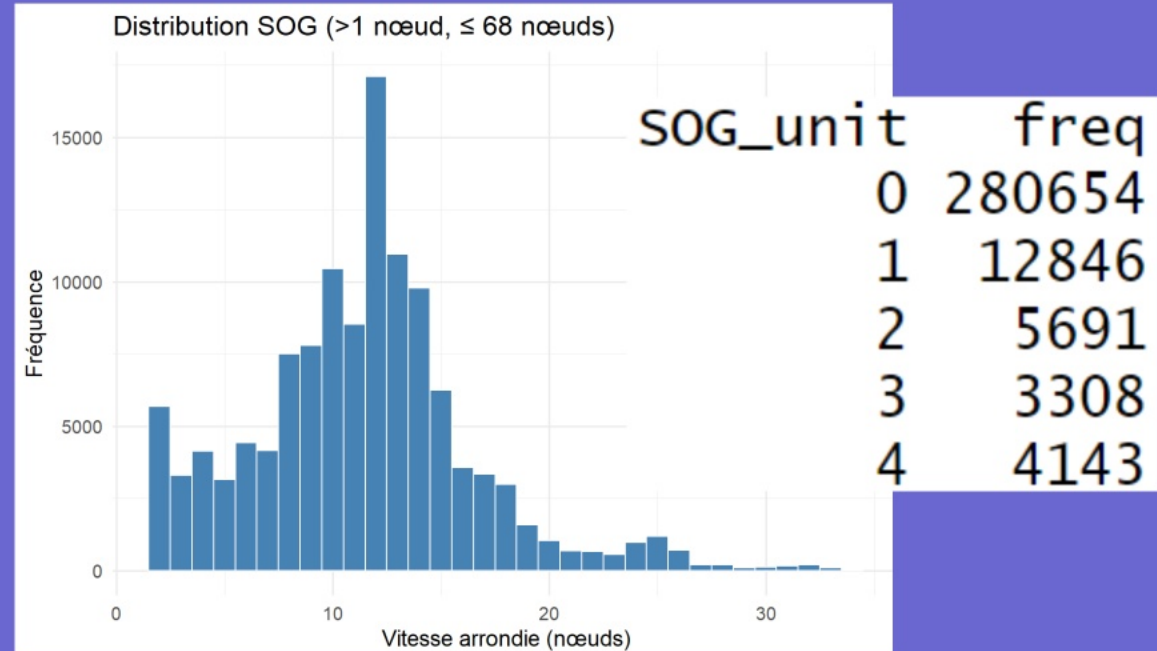
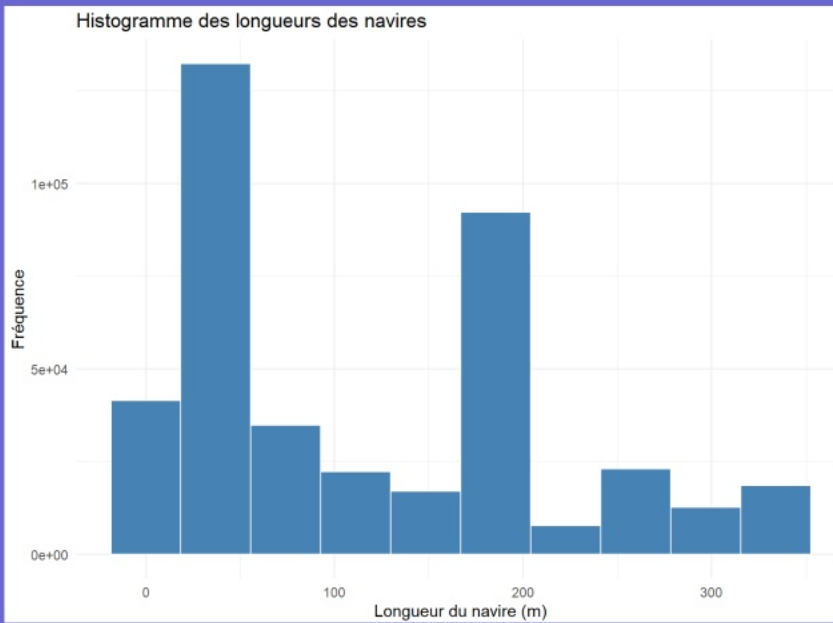
Ports principaux



Histogramme 1.

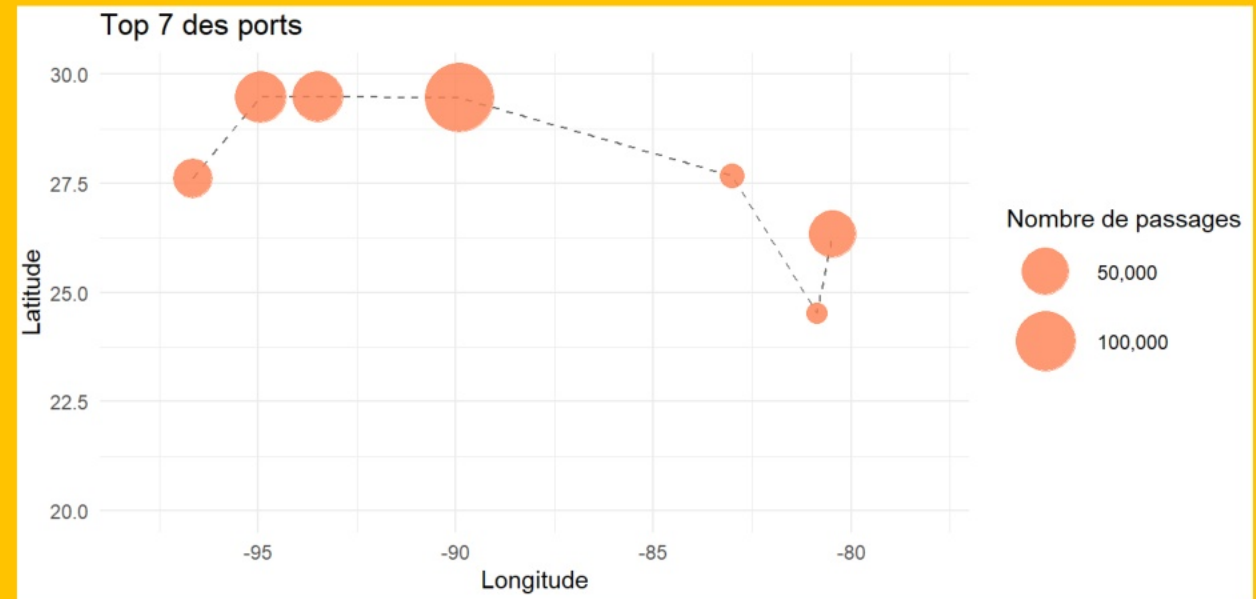


Histogramme 2.



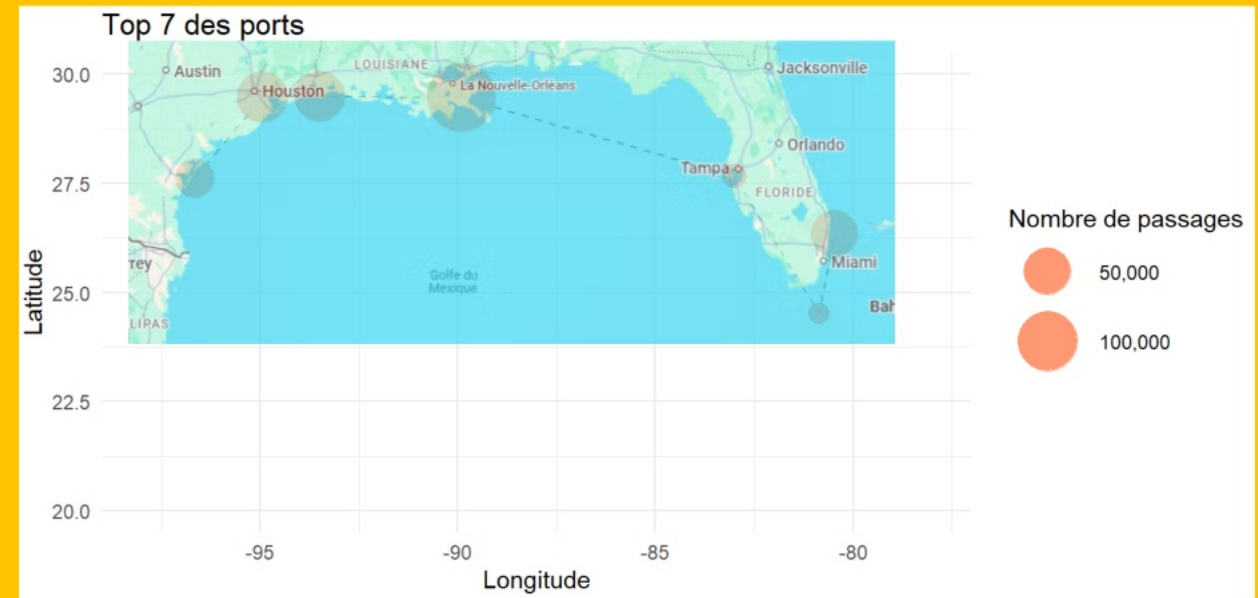


Ports principaux

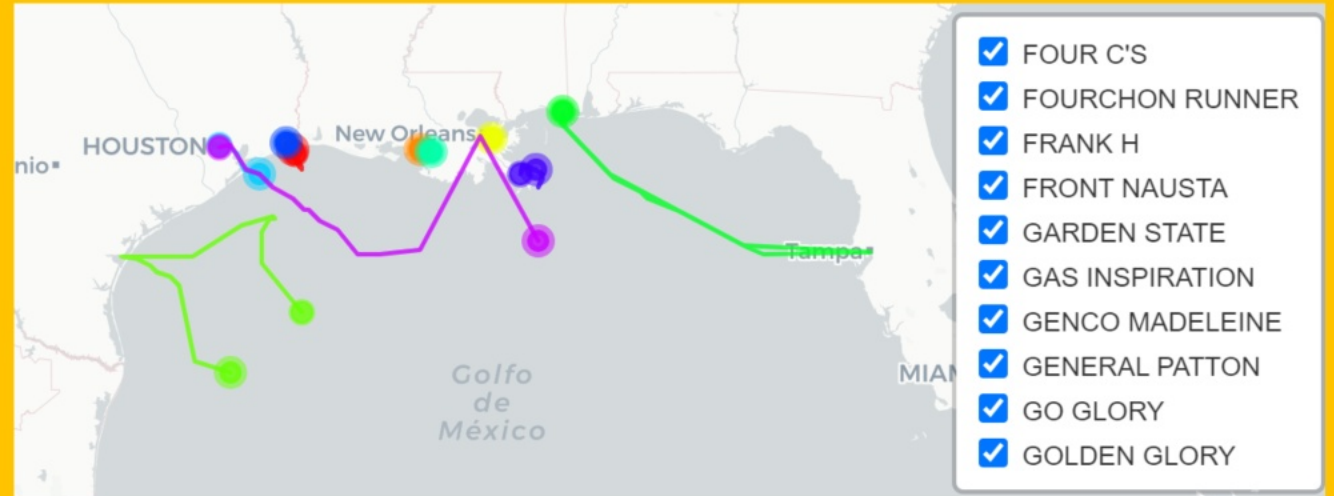




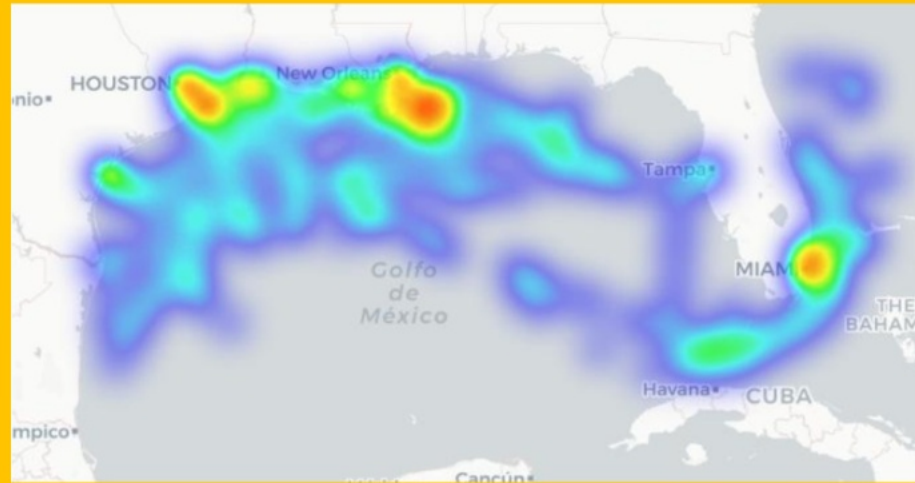
Ports principaux



Trajectoire des bateaux

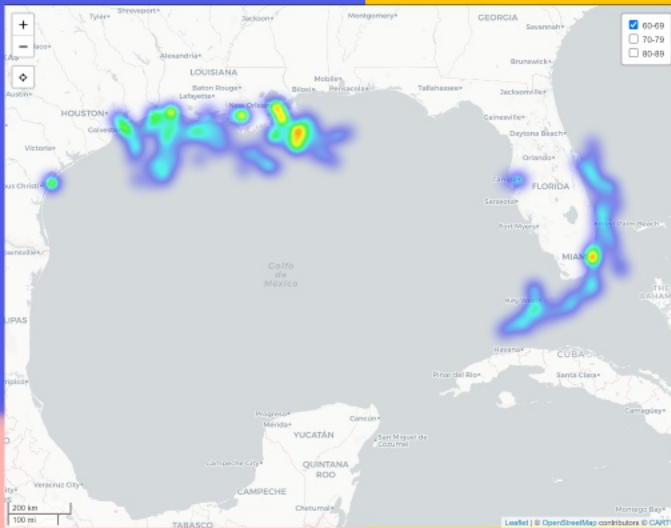


Heatmaps

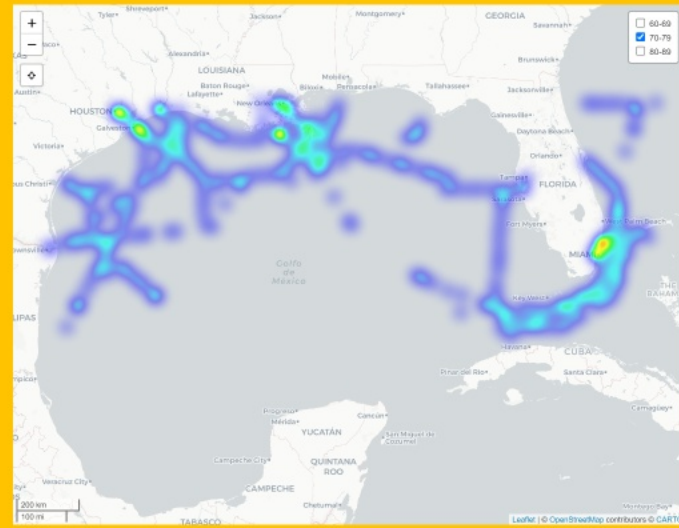
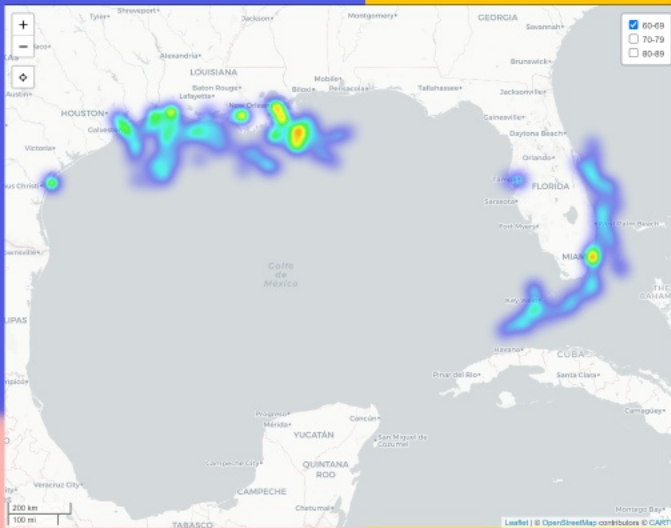


Heatmaps

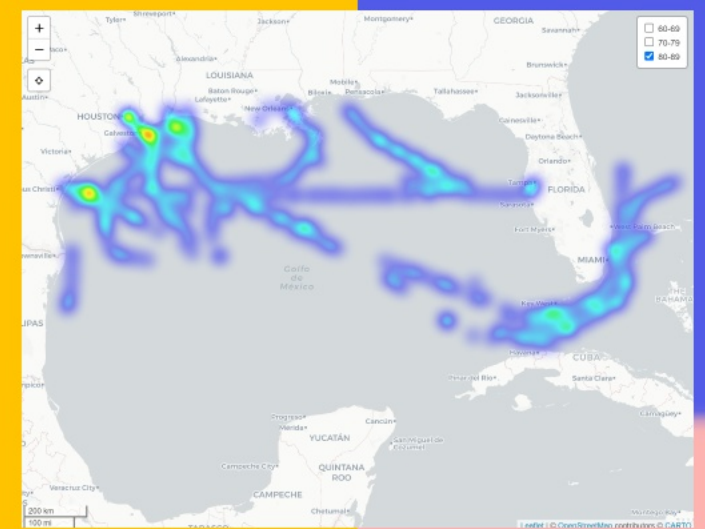
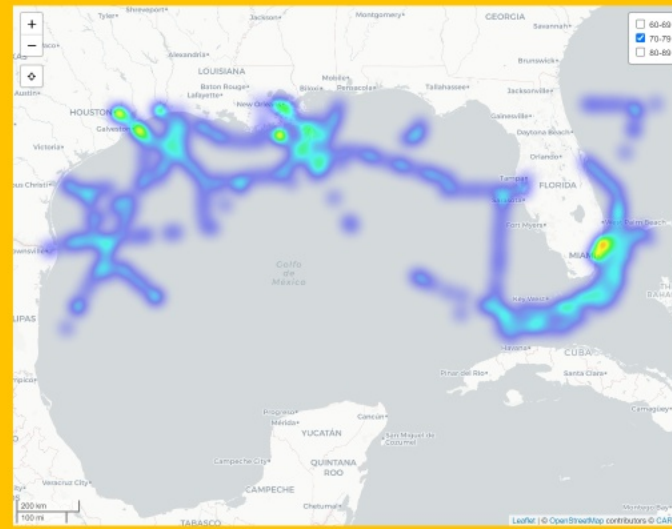
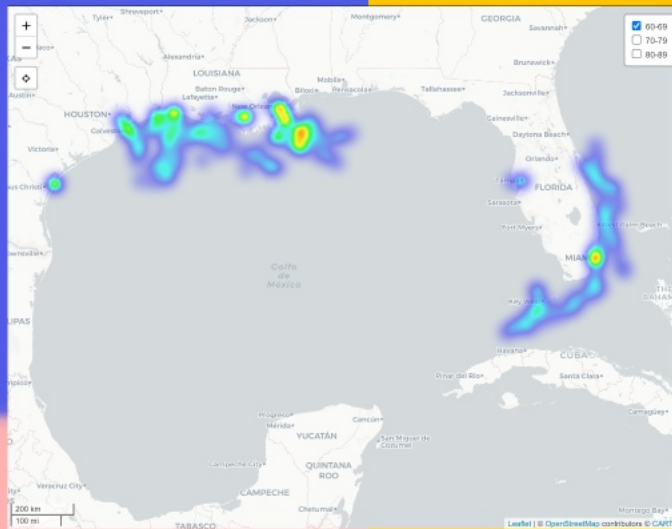
Heatmaps



Heatmaps



Heatmaps



Big Data A3

Analyse des données
maritimes dans le golfe du
Mexique

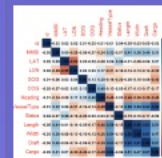


Etudes des corrélations entre variables

Analyse des données quantitatives

Attention aux types de données !

Matrice de corrélation :



Analyse des données qualitatives

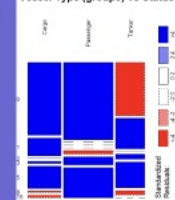
Mosaicplot :

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).

→ On rejette l'hypothèse d'indépendance

Vessel Type (groupe) vs Status



Analyse des données qualitatives + quantitatives

Boxplots

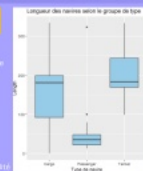
- Les cargoes présentent une grande dispersion autour de 150 m.
- Les passagers sont concentrés entre 70 et 80 m.
- Les tankers varient majoritairement de 210 à 290 m.

Analyse de variance (ANOVA)

Les données montrent une distribution de test F de 29.785 et une p-value extrêmement faible (1.2e-16).

Cela signifie que l'on peut rejeter l'hypothèse nulle d'égalité des moyennes avec une confiance quasi absolue.

→ Le type de navire influence donc bien sa longueur.

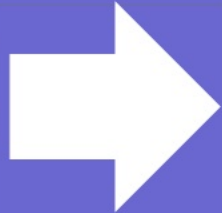
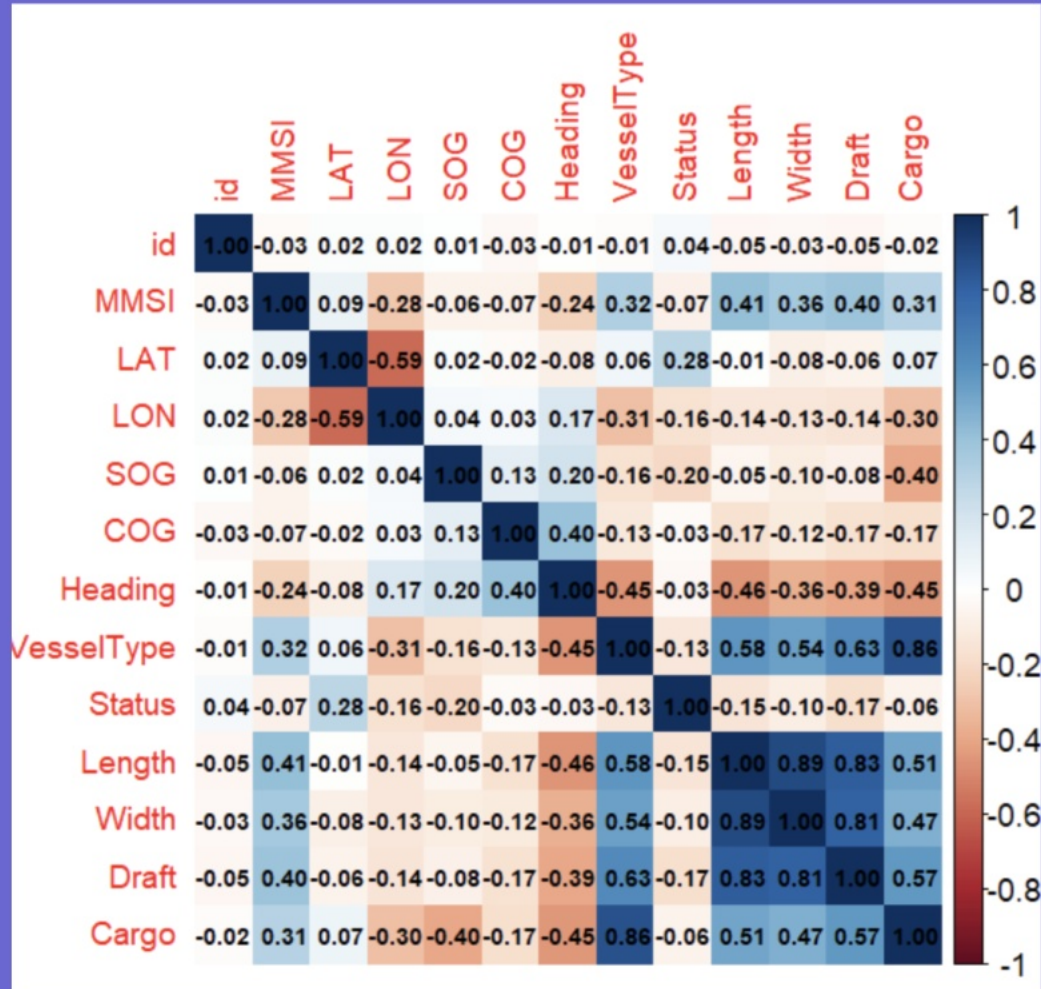


	df	Sum Sq	Mean Sq	F value	Pr(>F)
VesselType	2	2.131e+09	1.066e+09	29.785	<2e-16 ***
Residuals	401917	1.487e+09	3.699e+03		

Analyse des données quantitatives

Attention aux types de données !

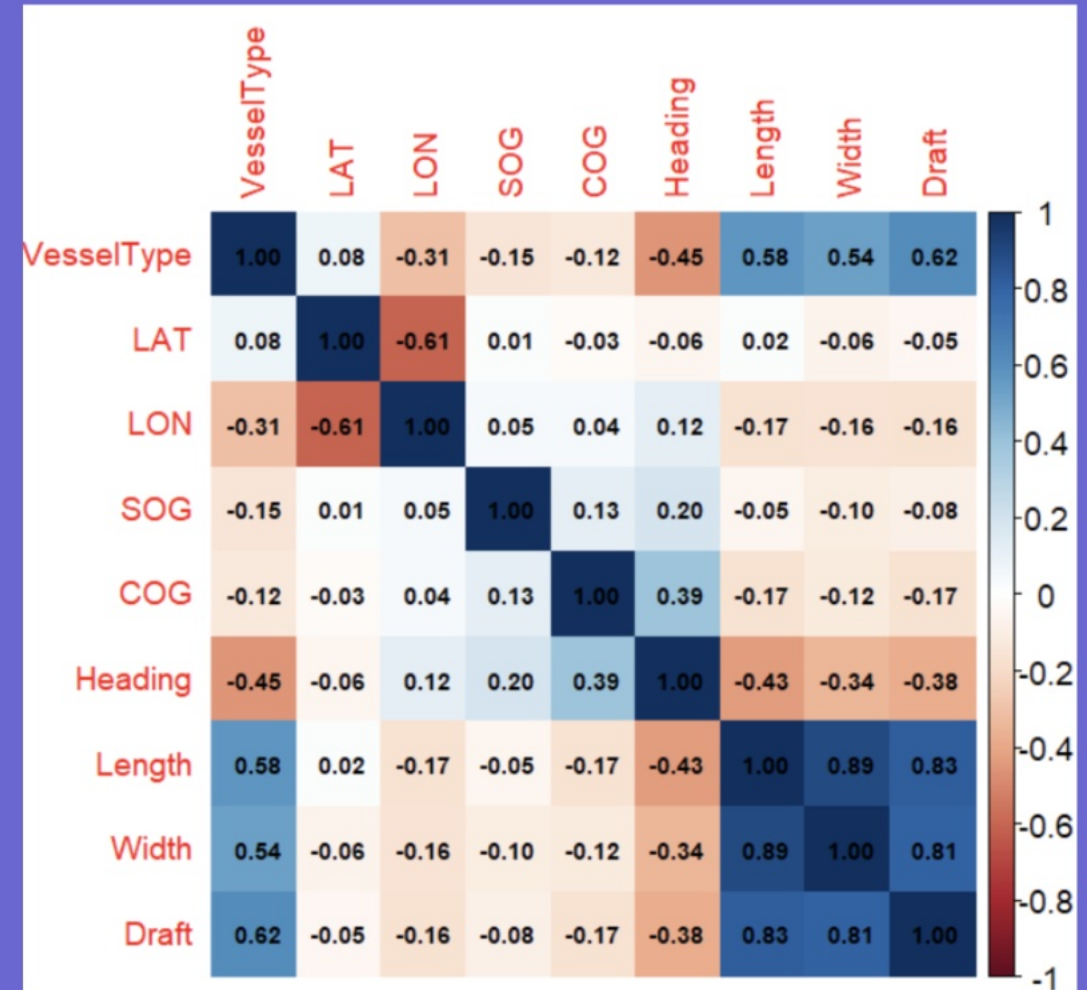
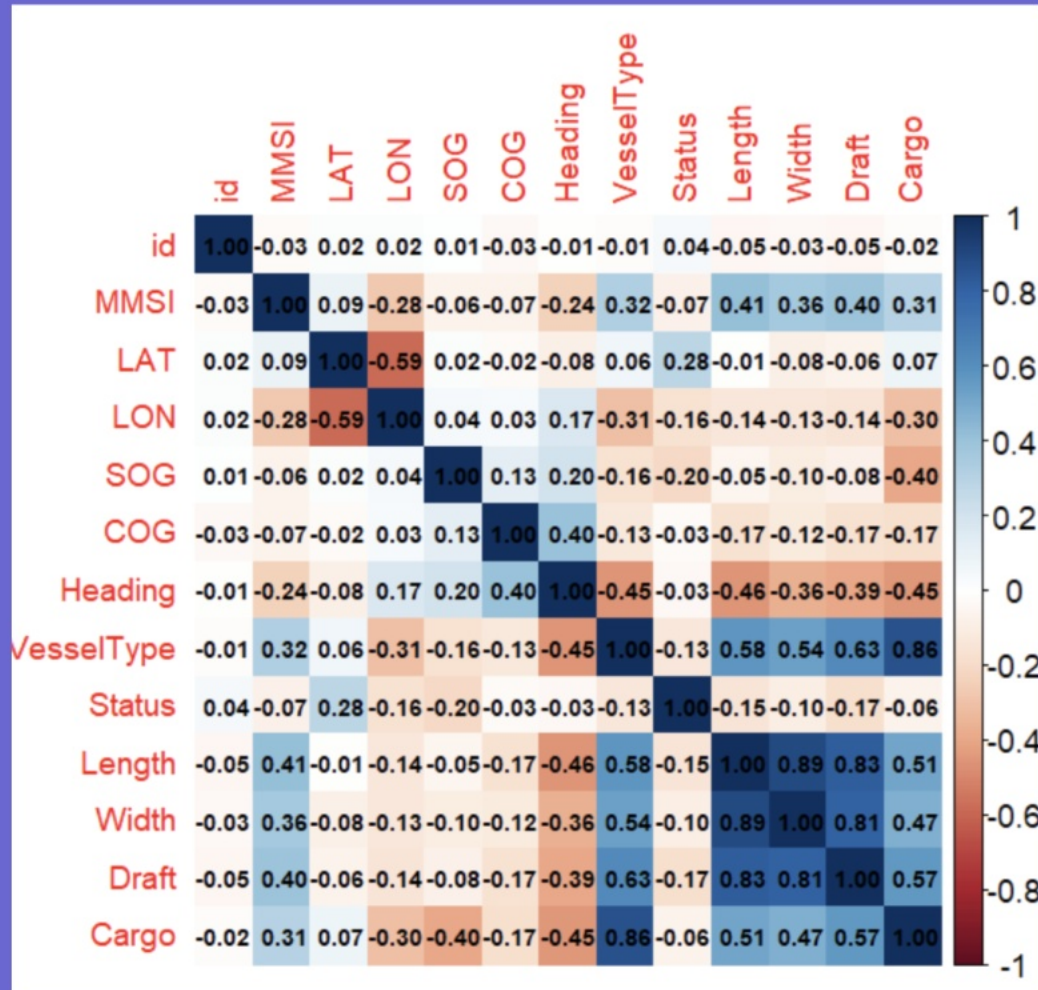
Matrice de corrélation :



Analyse des données quantitatives

Attention aux types de données !

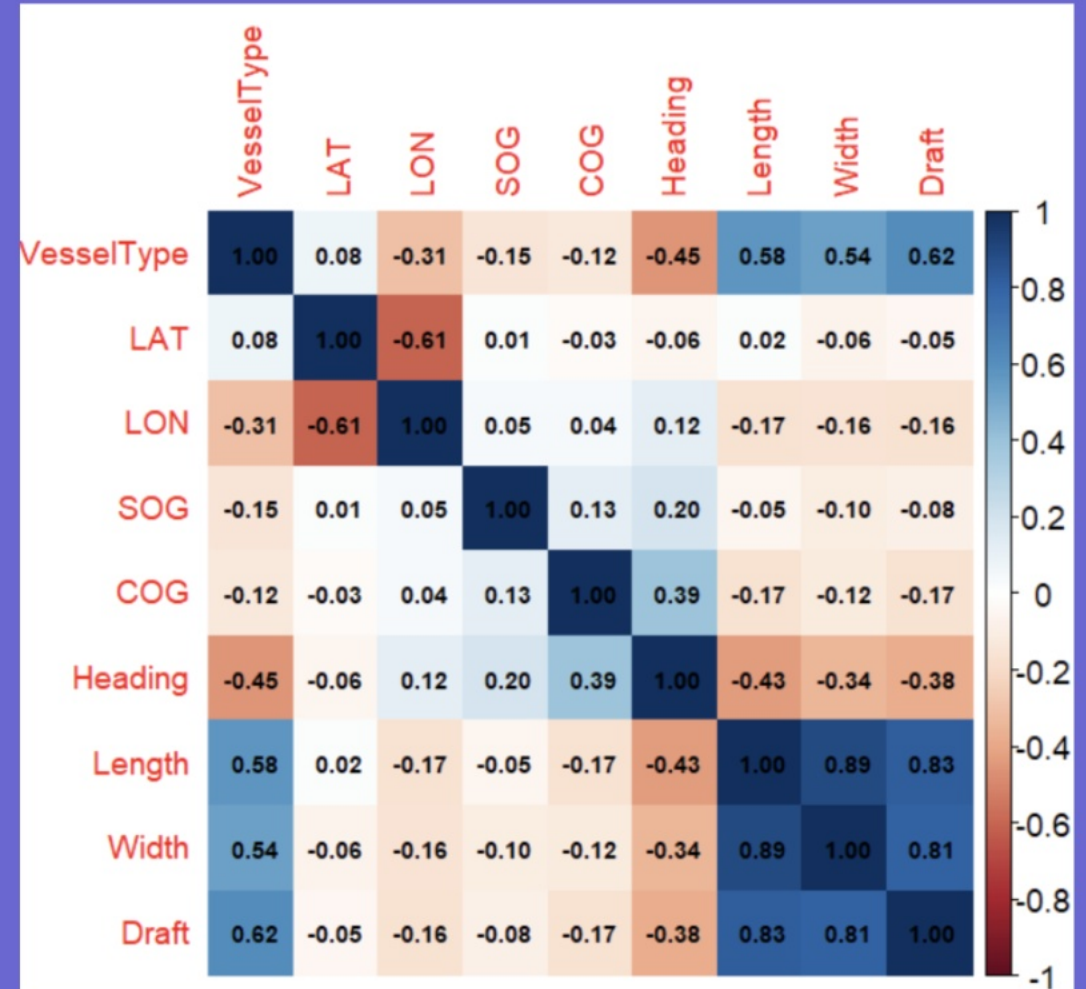
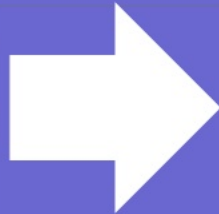
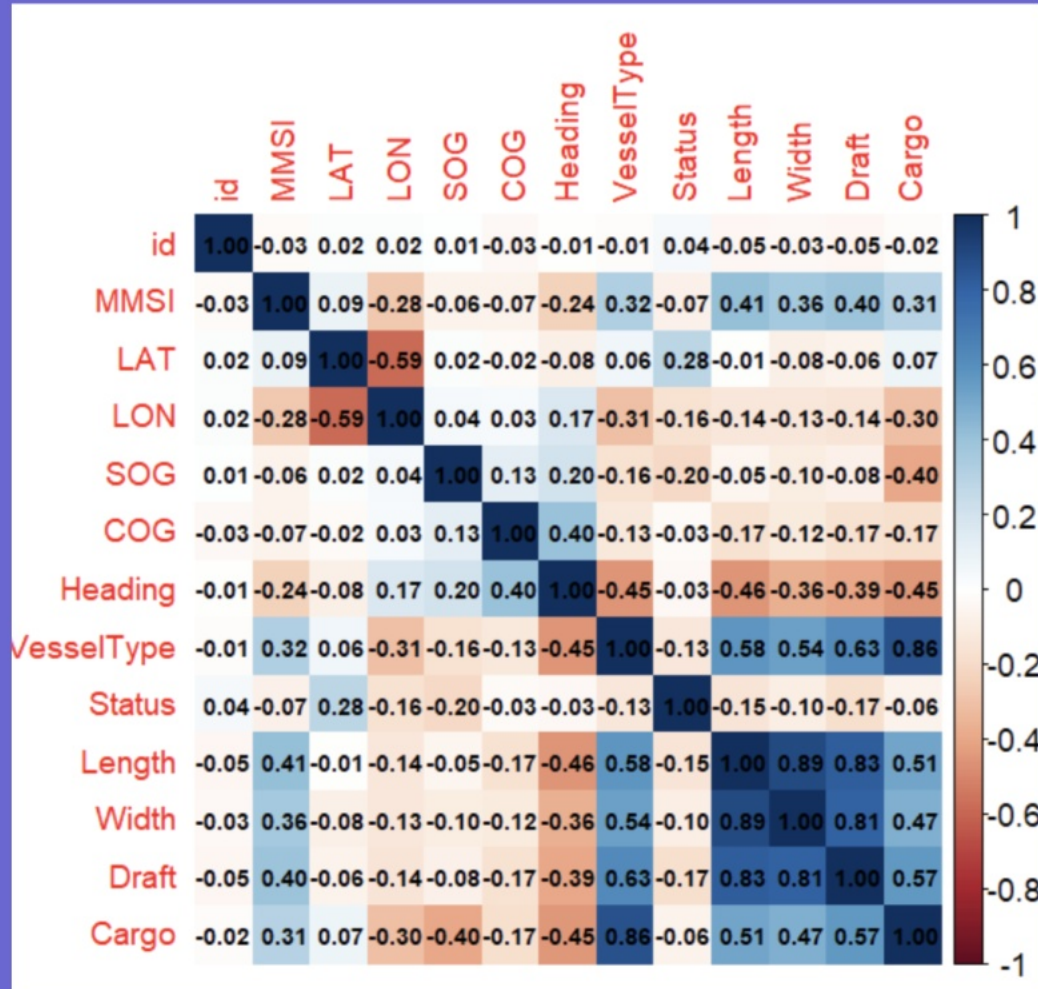
Matrice de corrélation :



Analyse des données quantitatives

Attention aux types de données !

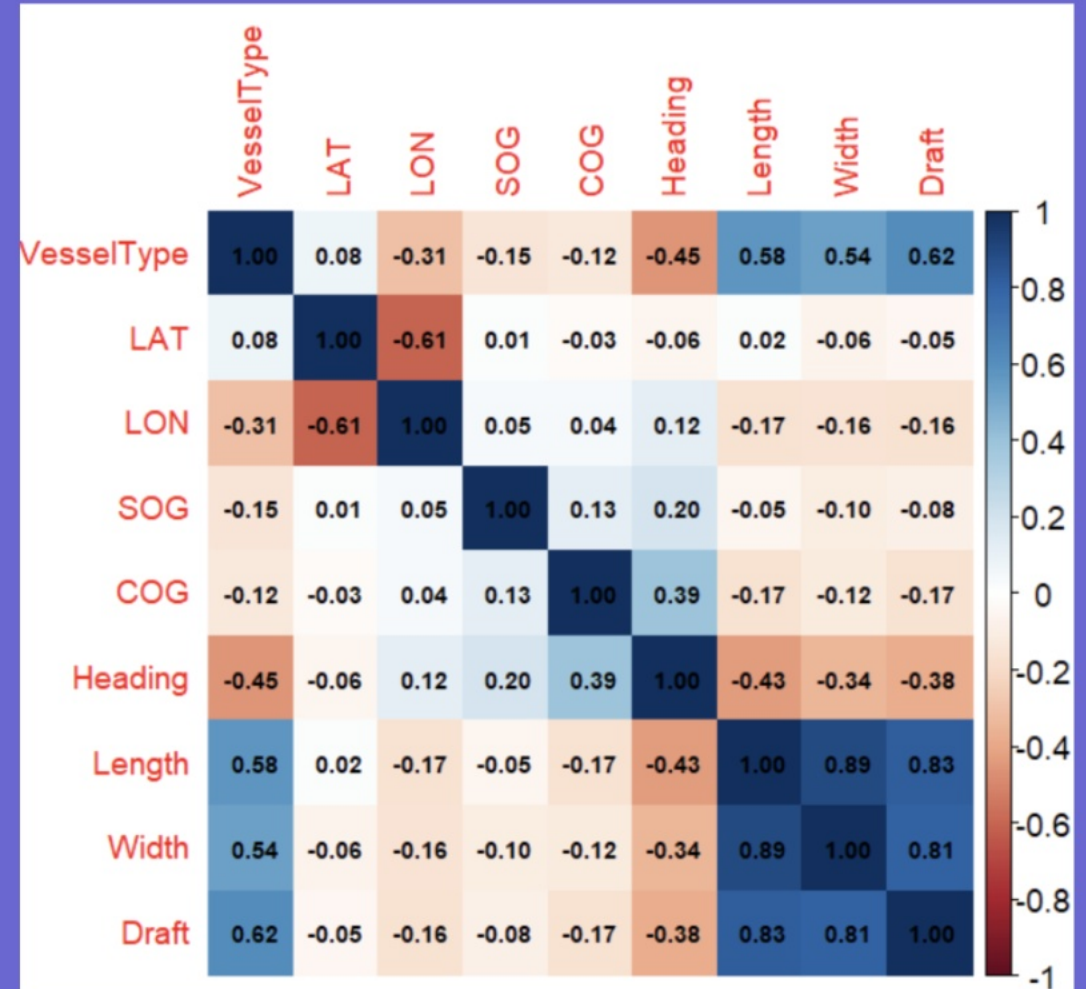
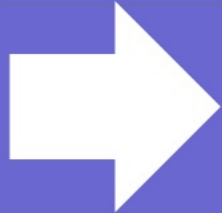
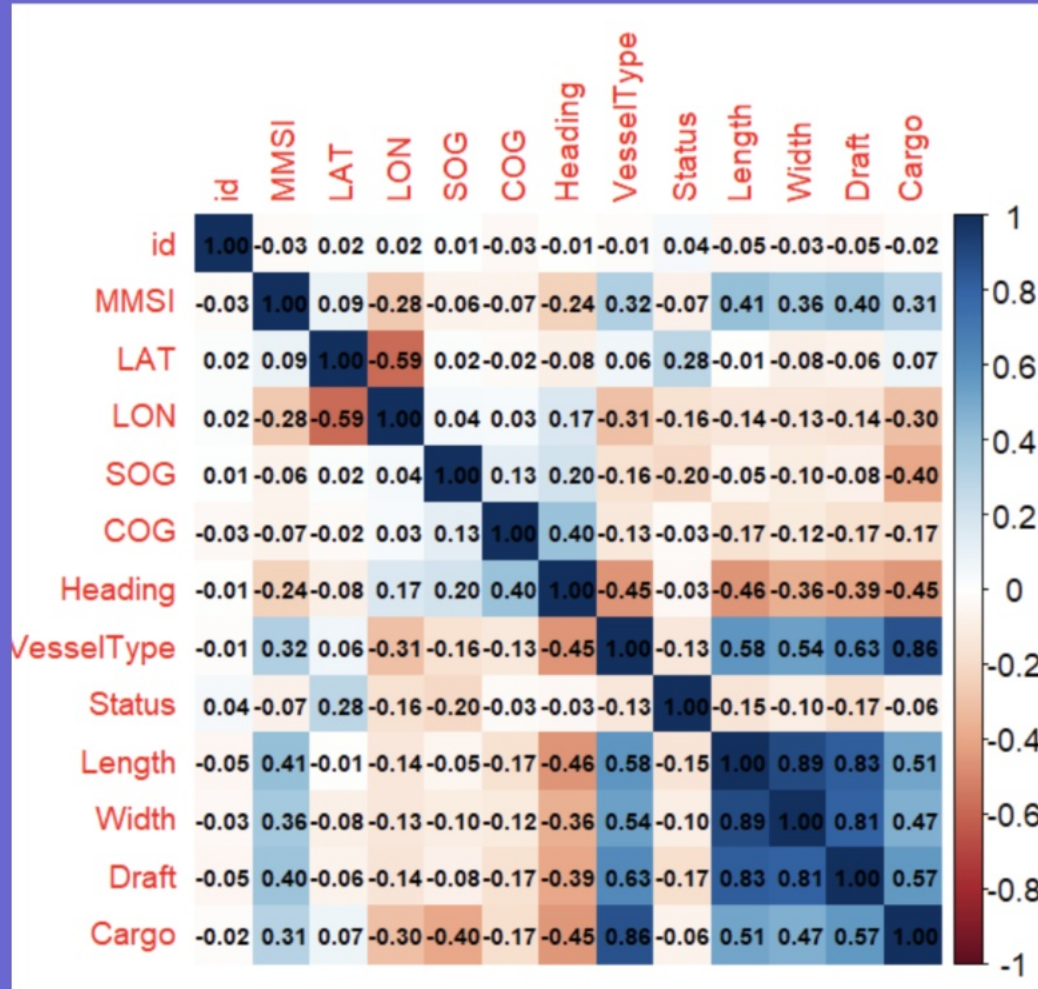
Matrice de corrélation :



Analyse des données quantitatives

Attention aux types de données !

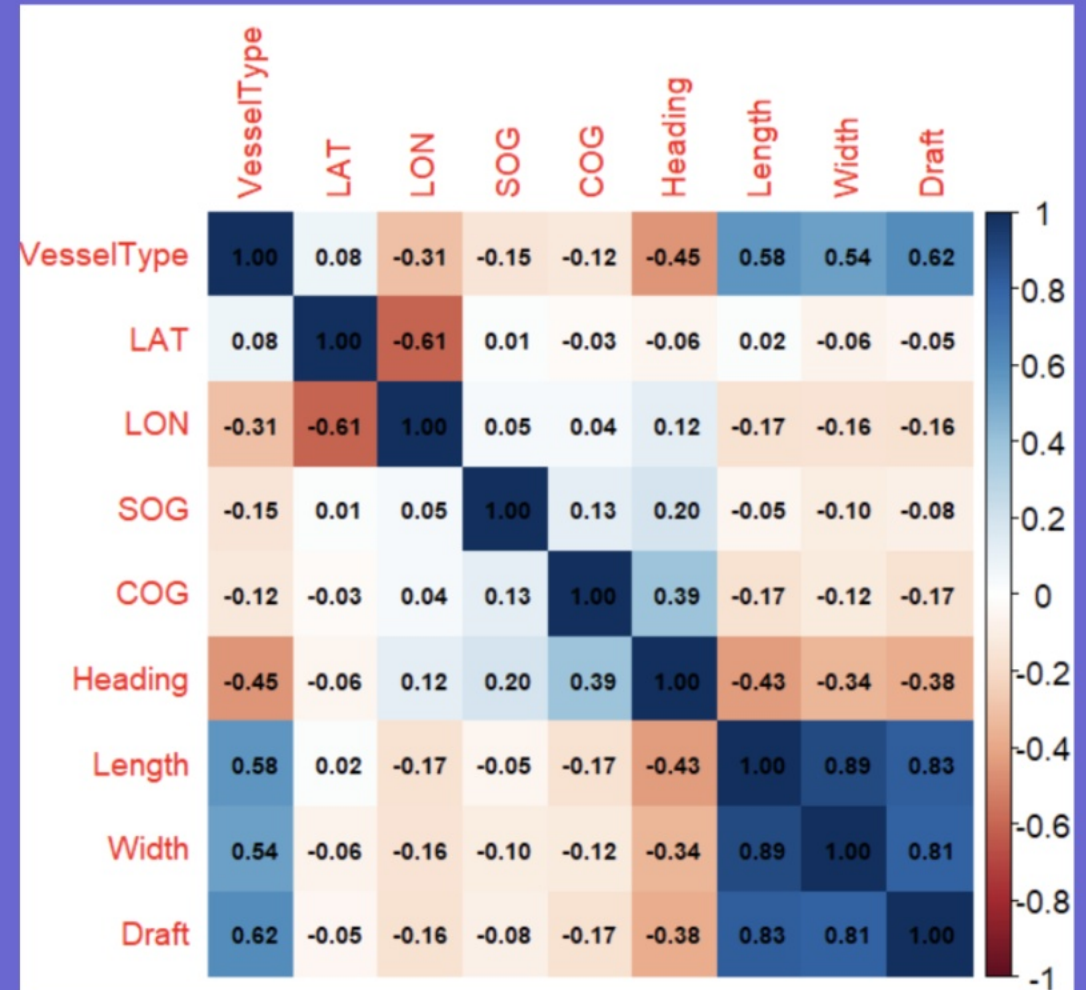
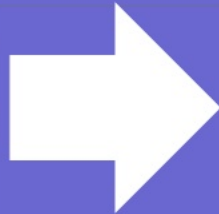
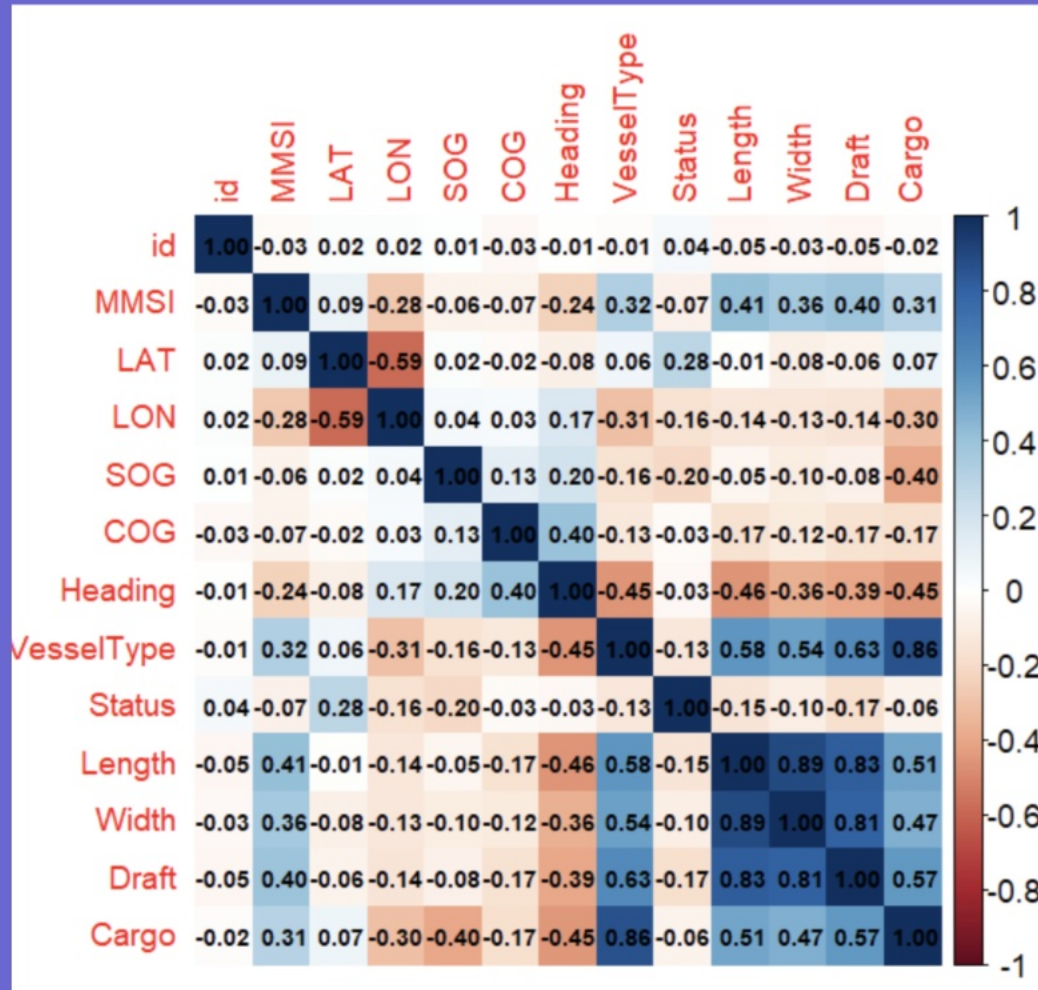
Matrice de corrélation :



Analyse des données quantitatives

Attention aux types de données !

Matrice de corrélation :



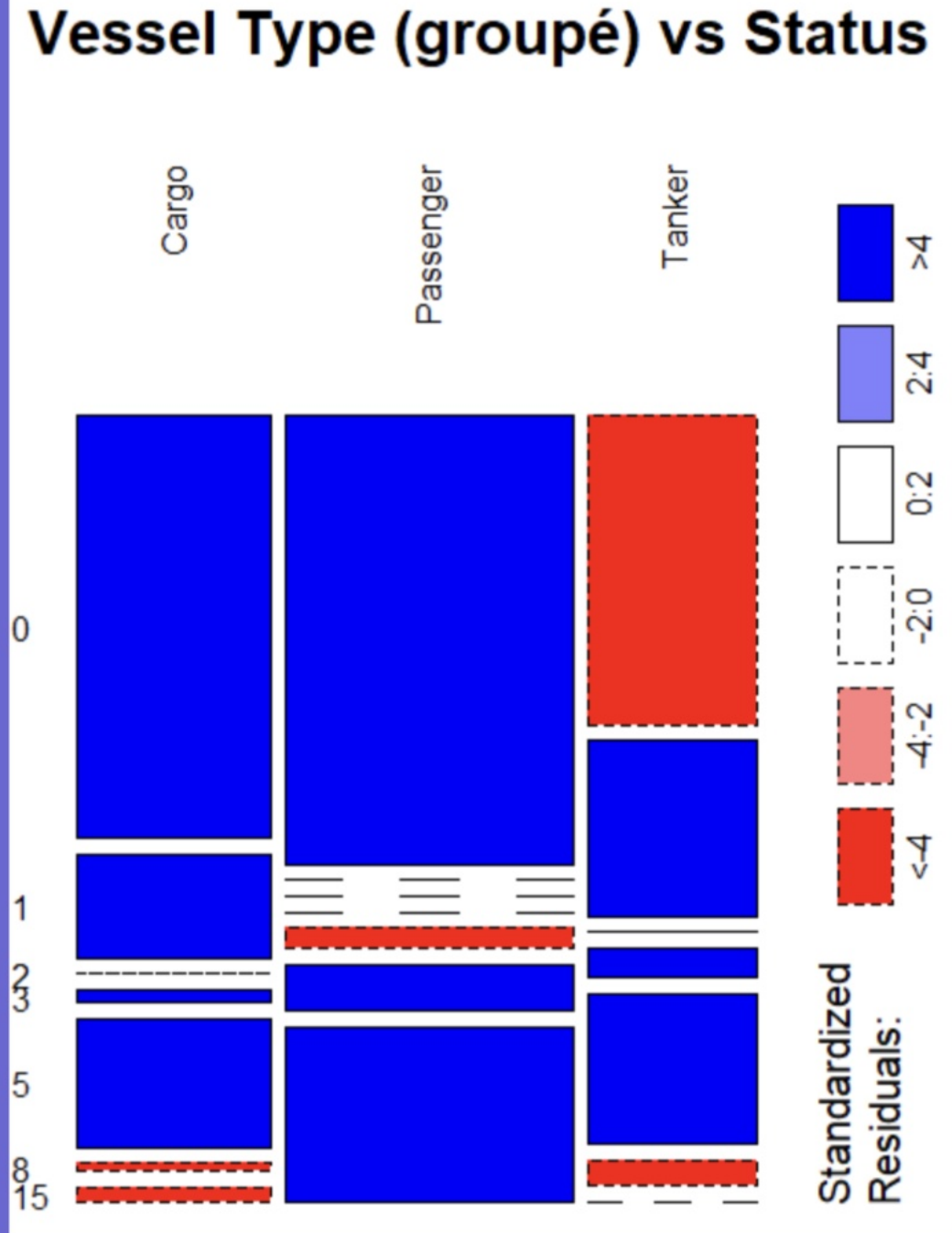
Analyse des données qualitatives

Mosaicplot :

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).

-> On rejète l'hypothèse d'indépendance



Analyse des données qualitatives × quantitatives

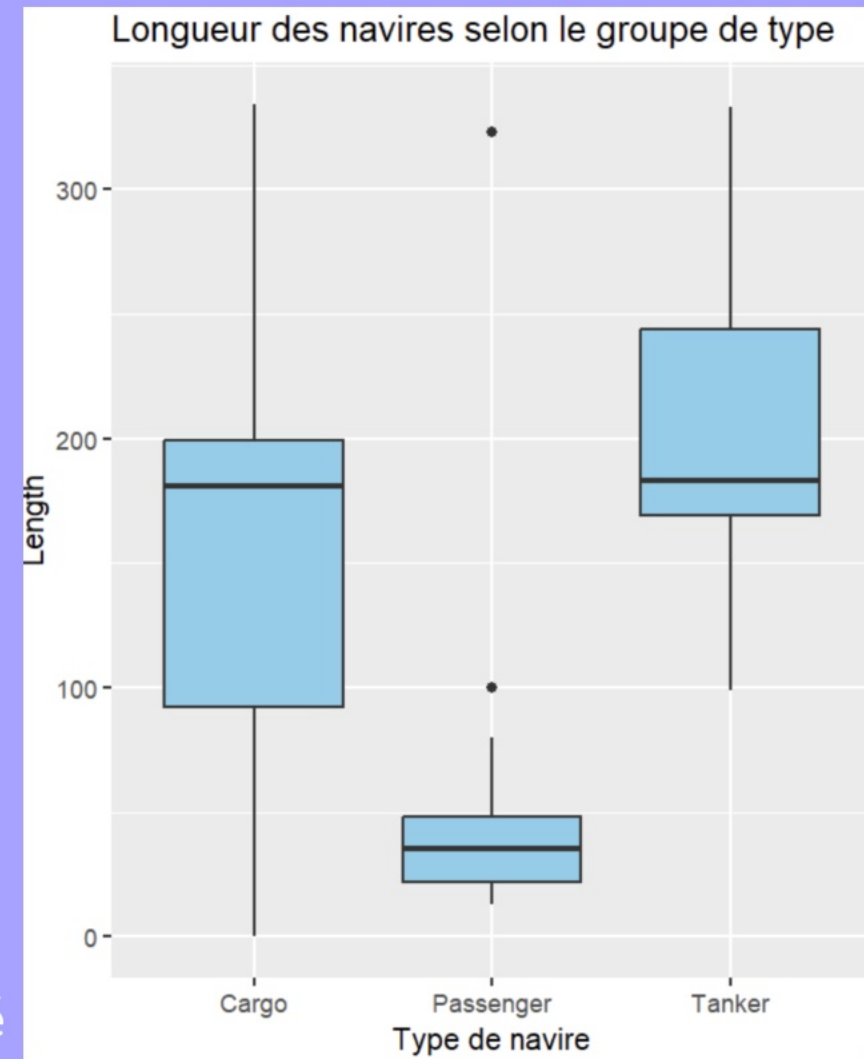
Boxplots

- Les cargos présentent une grande dispersion autour de 180 m
- Les passagers sont concentrés entre 20 et 60 m
- Les tankers varient majoritairement de 150 à 260 m.

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.



	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			

-> Le type de navire influence donc bien sa longueur.

Big Data A3

Analyse des données
maritimes dans le golfe du
Mexique



Prédiction/régression

Prédiction de VesselType

-On se base sur les variables avec forte corrélation, on s'entraîne sur 30 000 données (10/30)
-Validation croisée (5 folds)



Prédiction de la vitesse



On regarde les variables
RMSE : 5.803
R² : 0.078



Merci !



Prédiction de VesselType

- On se base sur les variables avec forte corrélation, on s'entraine sur 30 000 données (70/30)
- Validation croisée (5 folds)



	VesselType	LAT	LON	SOG	COG	Heading	Length	Width	Draft
VesselType	1.00	0.08	-0.31	-0.15	-0.12	-0.45	0.58	0.54	0.62

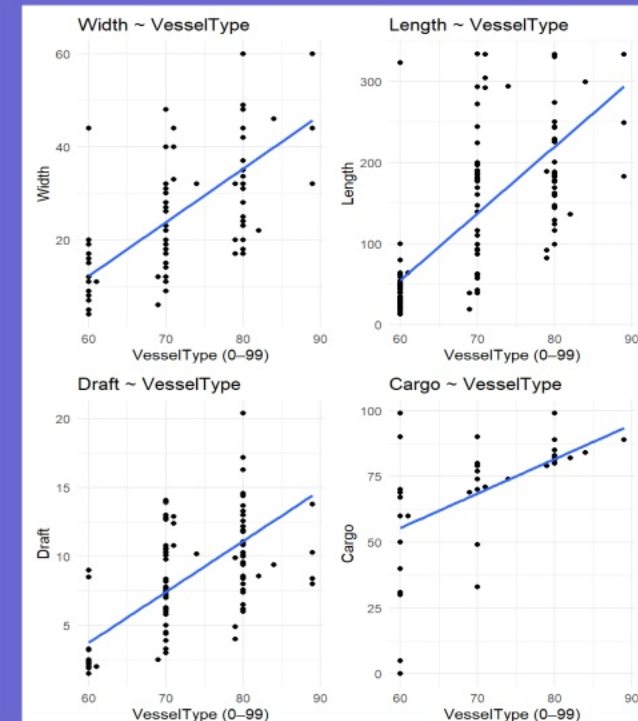
Prédiction de VesselType

- On se base sur les variables avec forte corrélation, on s'entraîne sur 30 000 données (70/30)
- Validation croisée (5 folds)

	VesselType	LAT	LON	SOG	COG	Heading	Length	Width	Draft
VesselType	1.00	0.08	-0.31	-0.15	-0.12	-0.45	0.58	0.54	0.62

Prédiction de VesselType

- On se base sur les variables avec forte corrélation, on s'entraîne sur 30 000 données (70/30)
- Validation croisée (5 folds)



	VesselType	LAT	LON	SOG	COG	Heading	Length	Width	Draft
VesselType	1.00	0.08	-0.31	-0.15	-0.12	-0.45	0.58	0.54	0.62

Prédiction de la vitesse

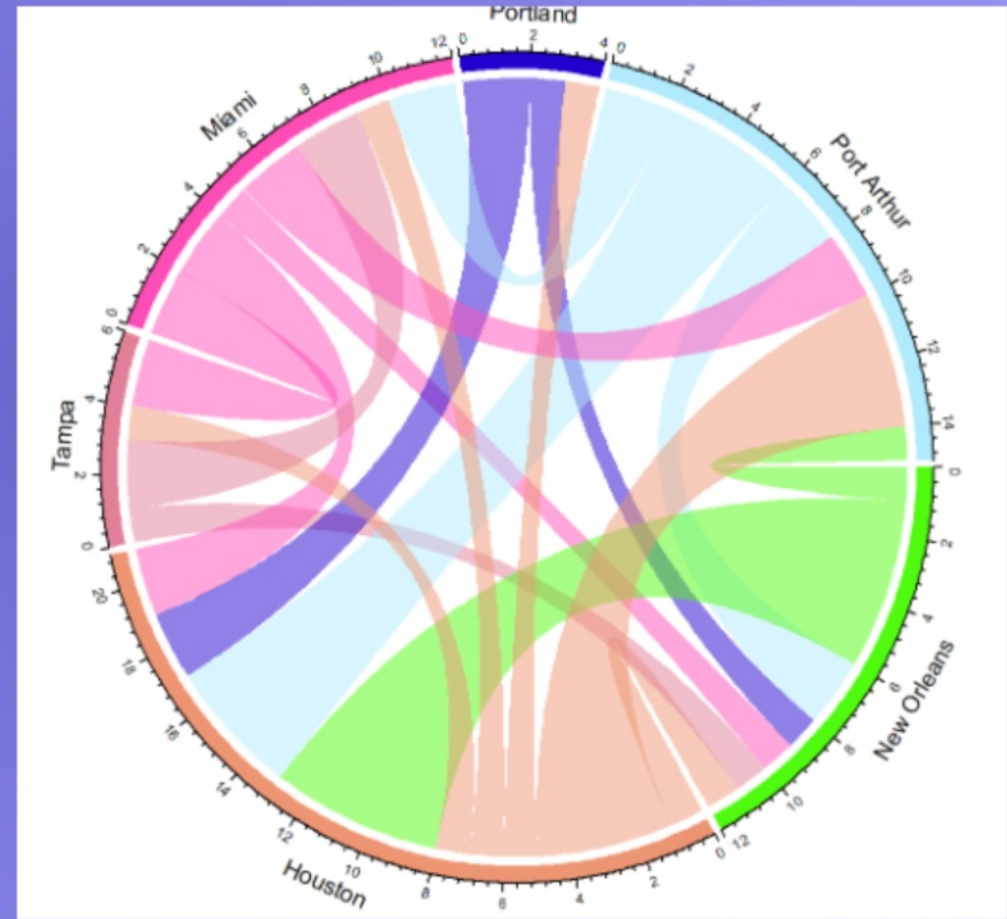
	VesselType	LAT	LON	SOG	COG	Heading	Length	Width	Draft
SOG	-0.15	0.01	0.05	1.00	0.13	0.20	-0.05	-0.10	-0.08

On regarde les variables

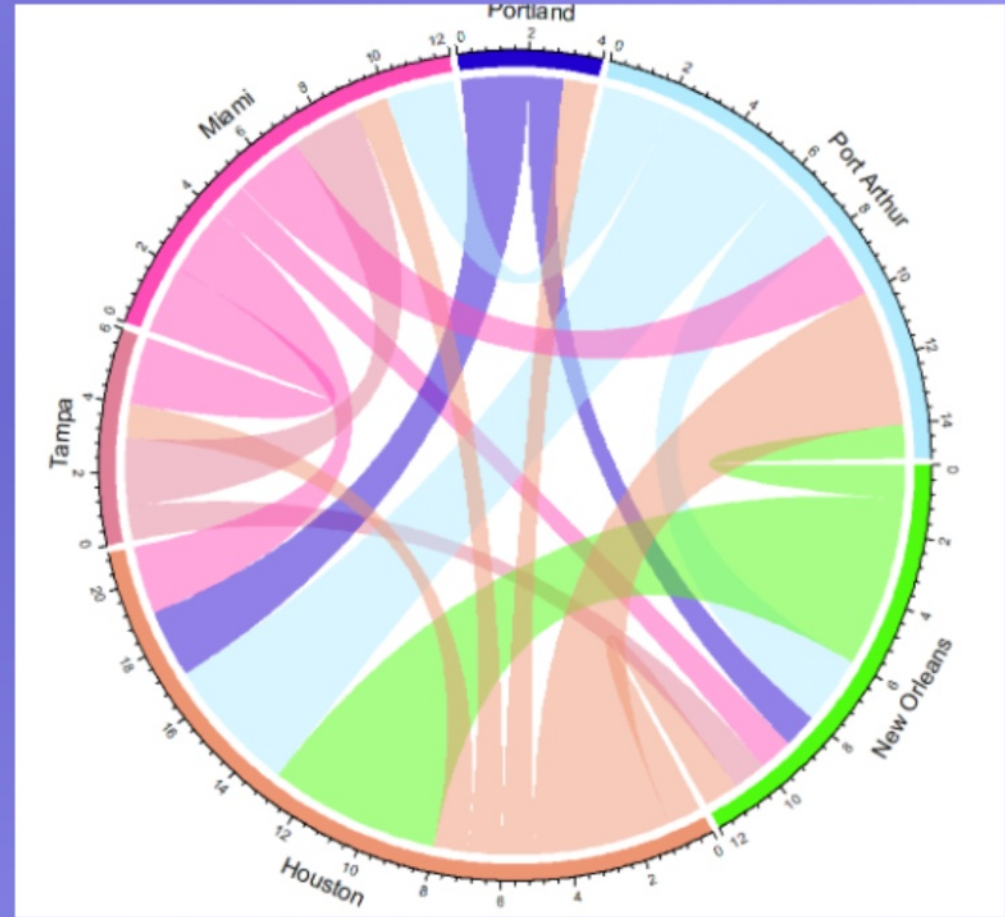
RMSE : 5.803
R² : 0.078



Bonus



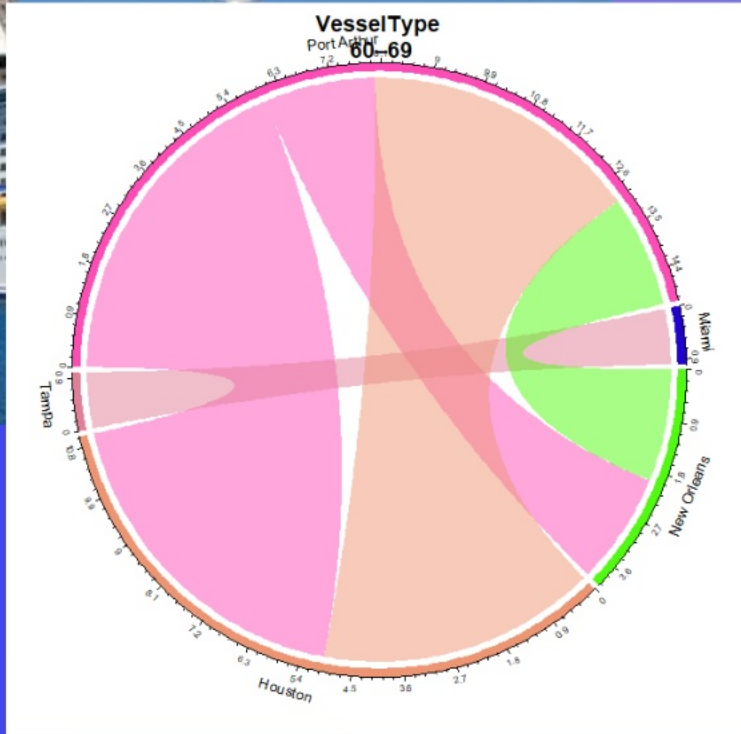
Bonus



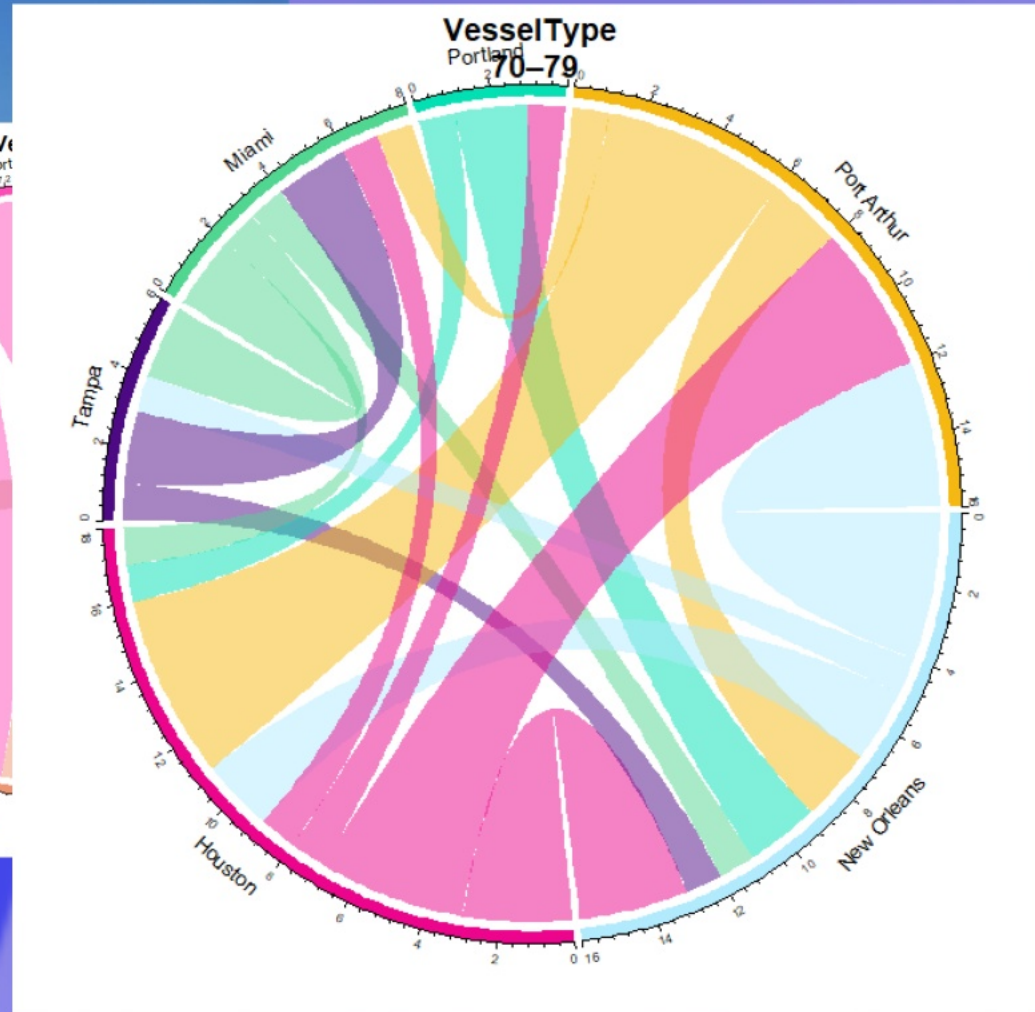
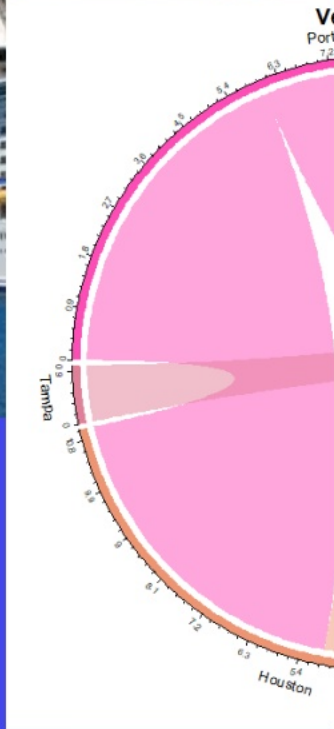
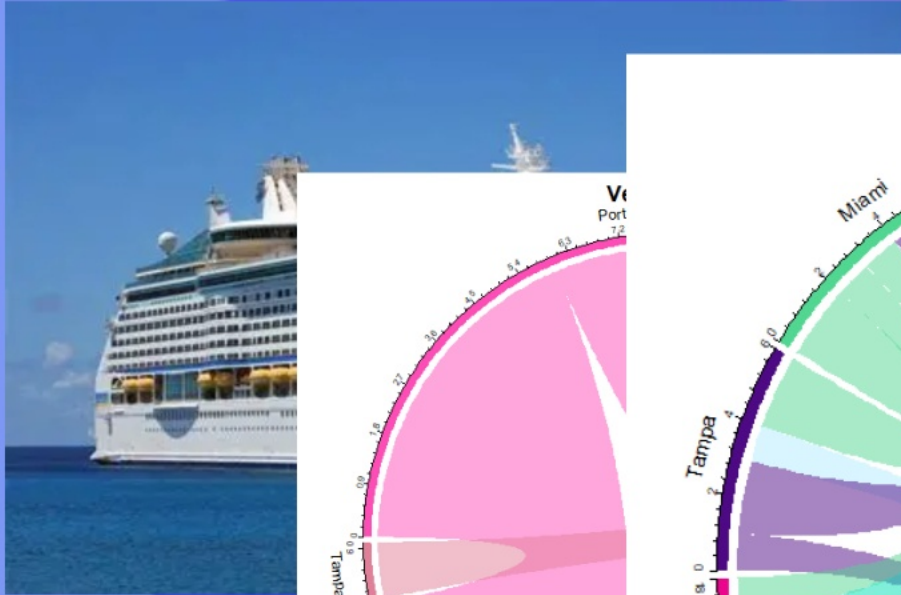
Bonus



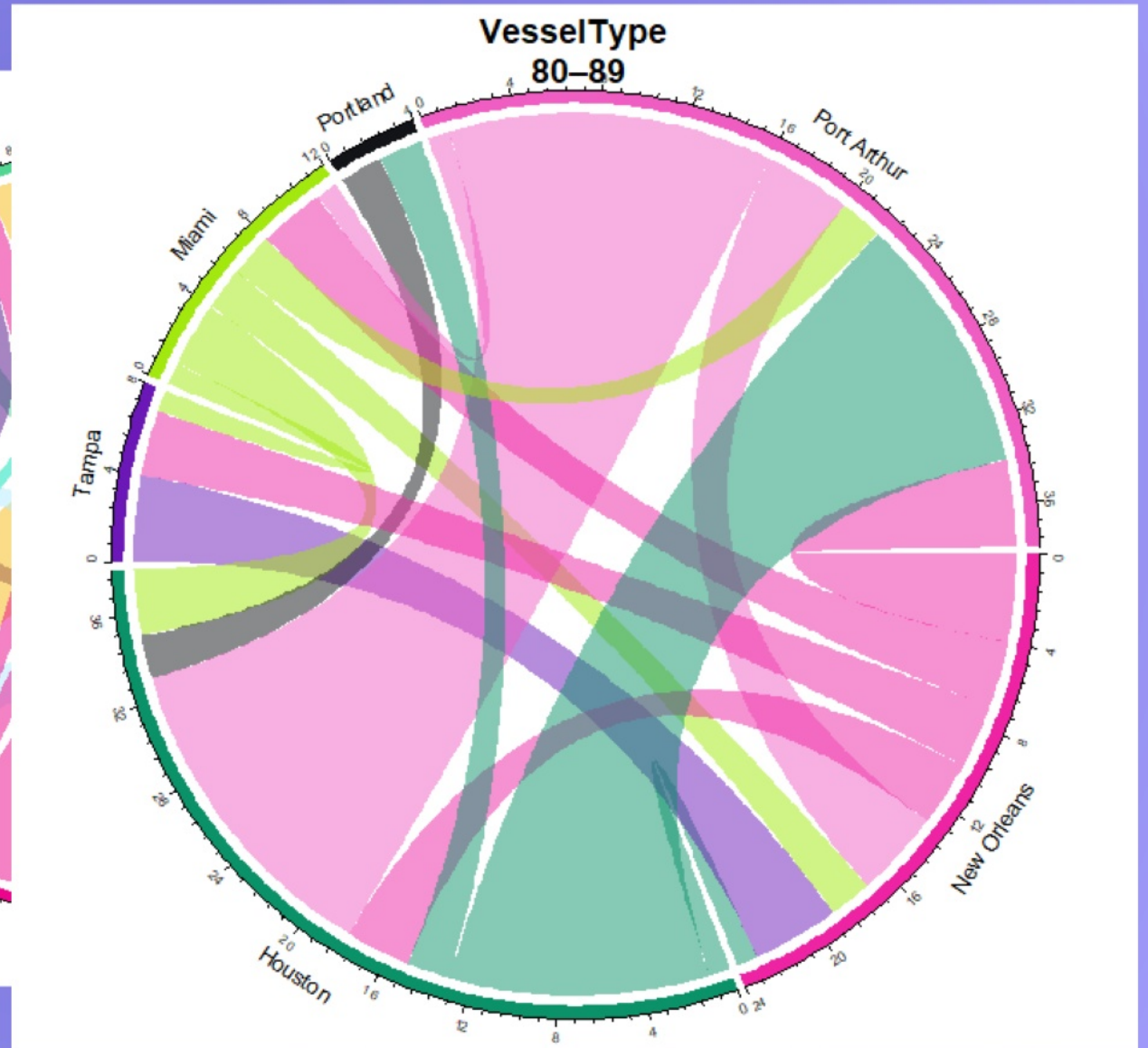
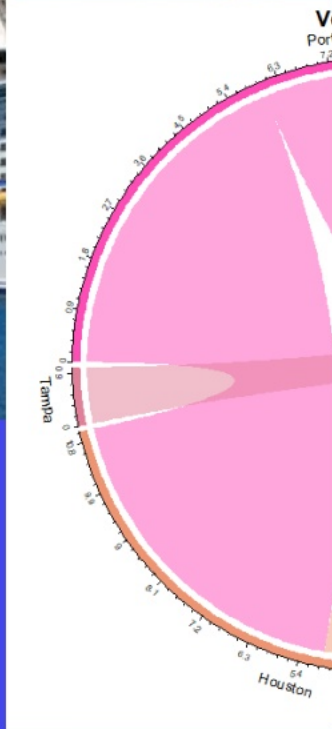
Bonus



Bonus



Bonus



Merci !

Merci !

Merci !

Merci !

Merci !