TFM: Análisis predictivo de incidentes navales en EEUU, 2002 - 2015

Anexo 4.2. Exploración de datos: MergedActivity

Oscar Antón

diciembre de 2023

Carga de librerías y datos

```
# Librería                        # Propósito
library(skimr)                    # Exploración estadística. Resumen
library(PerformanceAnalytics)     # Exploración estadística. Análisis de correlaciones

library(tidyverse)                # Sintaxis para el manejo de datos. Incluye dplyr, ggplot2, etc.
library(data.table)               # Manejo eficiente de conjuntos de datos
library(leaflet)                  # Representación geográfica
```

```
# Cargar el dataframe MergedActivity (solo incidentes)
MergedActivity <- as.data.table(readRDS("../1.DataPreprocess/DataMergedActivity/MergedActivity.rds"))
```

Descripción estadística

```
# Descripción de datos de incidentes
skim(MergedActivity)
```

Data summary

Name	MergedActivity
Number of rows	68000
Number of columns	28
Key	NULL
________________________	
Column type frequency:	
character	15
Date	1
numeric	12
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
hour	0	1.00	5	5	0	1438	0
region	0	1.00	6	14	0	6	0
watertype	0	1.00	5	5	0	2	0
event_type	0	1.00	4	30	0	26	0
damage_status	0	1.00	7	35	0	5	0
imo_number	0	1.00	0	7	46583	6013	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
vessel_name	0	1.00	1	50	0	23854	0
vessel_class	0	1.00	5	23	0	16	0
build_year	0	1.00	4	4	0	122	0
flag_abbr	0	1.00	0	2	24	106	0
classification_society	0	1.00	6	58	0	36	0
solas_desc	0	1.00	9	16	0	3	0
casualty	65628	0.03	4	11	0	4	0
pollution	55049	0.19	0	3	73	131	0
event_class	0	1.00	15	19	0	5	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2002-01-01	2015-06-22	2008-07-05	4693

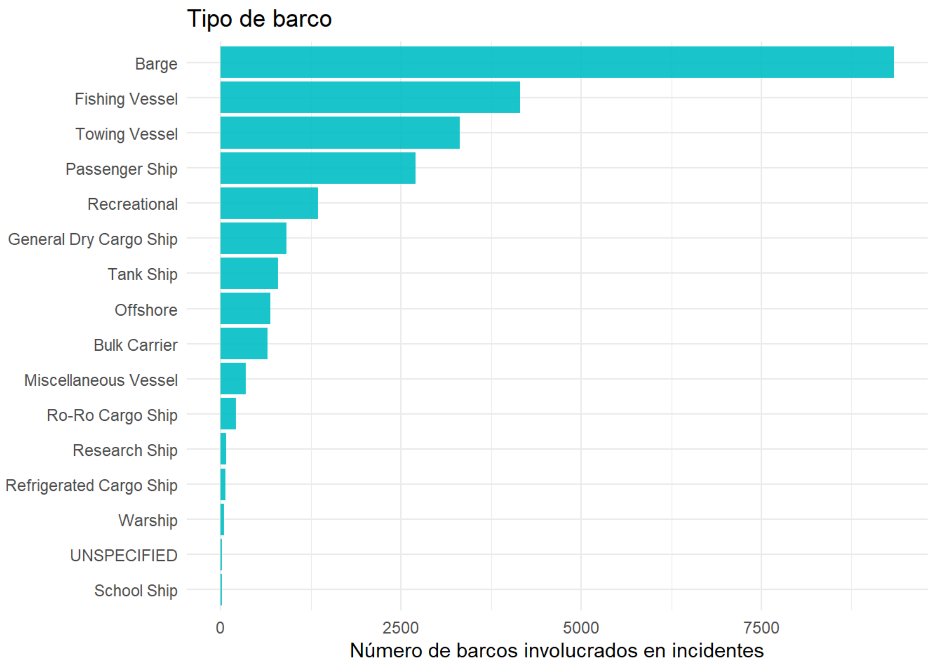
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
activity_id	0	1.00	3271440.45	942973.00	1838.0	2535937.00	3275581.00	4052092.75	5167891.00	▁▂▇▇
latitude	0	1.00	37.62	7.94	15.0	32.26	37.80	41.33	69.75	▁▇▇▁
longitude	0	1.00	-94.41	21.47	-179.8	-94.87	-89.55	-81.91	-46.26	▁▁▂▇
vessel_id	0	1.00	290399.44	268497.49	18.0	90388.00	191335.00	427193.00	1325666.00	▇▃▁▁
age	0	1.00	25.37	16.44	-7.0	12.00	26.00	34.00	138.00	▇▇▁▁
gross_ton	0	1.00	4391.51	13526.64	1.0	95.00	483.00	975.00	225282.00	▇▁▁▁
length	0	1.00	196.07	197.01	18.7	69.70	136.30	200.00	1203.80	▇▁▁▁
air_temp	5342	0.92	149.76	95.00	-230.5	82.00	153.50	230.71	350.00	▁▂▆▇
wind_speed	25494	0.63	50.58	30.65	0.0	29.00	44.14	65.17	350.00	▇▂▁▁
wave_hgt	55164	0.19	2.35	2.28	0.0	1.00	2.00	3.00	99.00	▇▁▁▁
visibility	63031	0.07	96.90	1.44	90.0	96.50	97.00	98.00	99.00	▁▁▂▇
damage_assessment	56	1.00	122959.80	3887042.37	0.0	0.00	0.00	10000.00	410000000.00	▇▁▁▁

1. Características de los barcos

1.1. Tipo de barco (vessel_class)

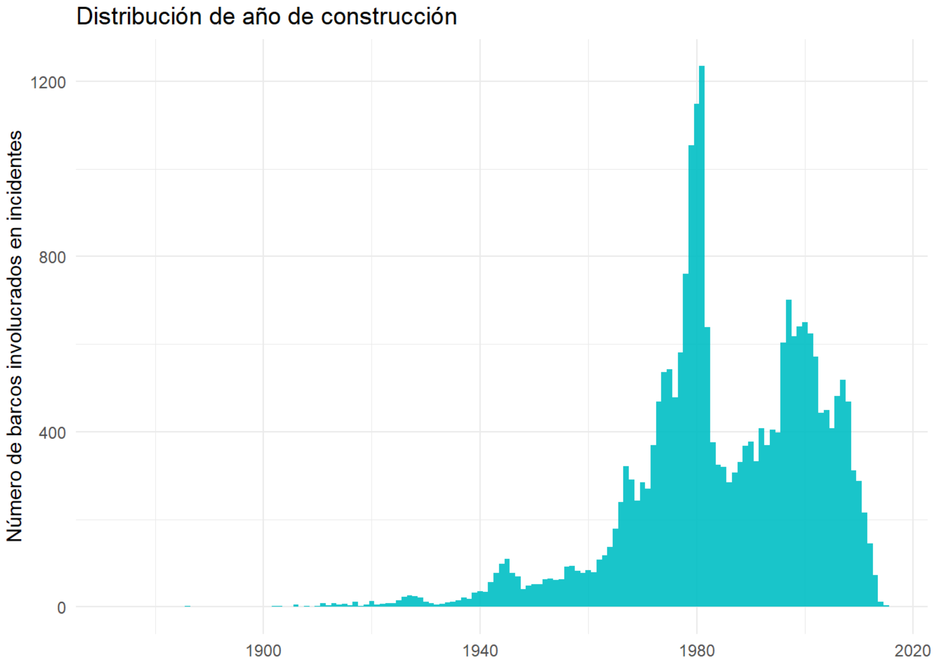
```
# Gráfico de barras para barcos con incidente
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  group_by(vessel_class) %>%
  summarise(frecuencia = n()) %>%
  ggplot(aes(x = fct_reorder(vessel_class, frecuencia), y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  labs(title = "Tipo de barco", x = NULL, y = "Número de barcos involucrados en incidentes") +
  theme_minimal() +
  coord_flip()
```



1.2. Año de construcción (build_year)

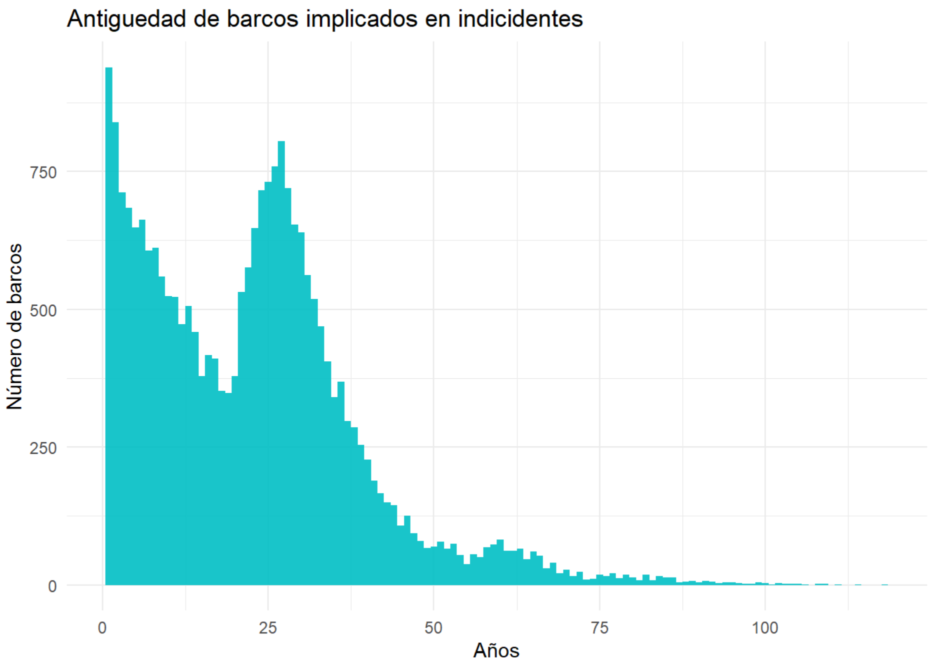
1.2.1. Construcción

```
# Gráfico de barras por año de construcción para barcos con incidentes
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(build_year >= 1800 & build_year <= 2015) %>%
  ggplot(aes(x = as.numeric(build_year))) +
  geom_histogram(binwidth = 1, fill = "#00bfc4", alpha = 0.9) +
  labs(title = "Distribución de año de construcción", x = NULL, y = "Número de barcos involucrados en incidentes") +
  theme_minimal()
```



1.2.2. Antiguedad en el accidente

```
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(age > 0, age < 120) %>%
  ggplot(aes(x = as.numeric(age))) +
  geom_histogram(binwidth = 1, fill = "#00bfc4", alpha = 0.9) +
  labs(title = "Antiguedad de barcos implicados en indicidentes", x = "Años", y = "Número de barcos") +
  theme_minimal()
```



1.2.3. Valores anómalos en las fechas

```
# Barcos con "antigüedad" negativa
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(build_year >= 1800 & build_year <= 2015) %>%
  mutate(antiguedad = year(as.Date(date)) - year(as.Date(paste0(build_year, "-01-01")))) %>%
  filter(antiguedad < 0) %>%
  select(vessel_id, vessel_name, imo_number, event_type, date, build_year, antiguedad) %>%
  knitr::kable("html")%>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size = 12)
```

vessel_id	vessel_name	imo_number	event_type	date	build_year	antiguedad
370425	BENNO C. SCHMIDT	9040546	Flooding	2002-02-20	2009	-7
813316	PEAPICKER		Damage to the Environment	2002-07-28	2006	-4
568186	MIDNIGHT SUN	9232278	Damage to the Environment	2002-09-08	2003	-1
722960	ALASKAN EXPLORER	9244661	Damage to the Environment	2004-10-04	2005	-1
567313	DOLPHIN SEAFARI		Material Failure (Vessels)	2008-03-25	2015	-7
1052777	OPTI-EX		Damage to the Environment	2010-01-15	2011	-1
1110524	FSV6	9664988	Grounding	2012-11-04	2013	-1
1229111	PACIFIC SPIRIT		Damage to the Environment	2013-02-12	2015	-2

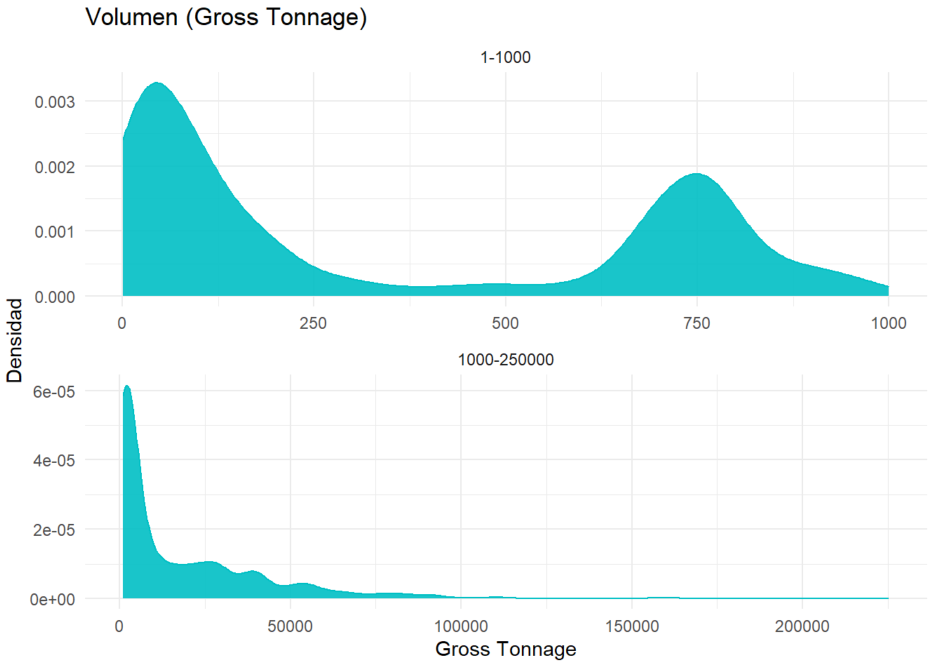
Los incidentes con antigüedad -1, pueden darse durante las pruebas de mar o en la fase de construcción. Sin embargo, -7 o -2 son valores anómalos. Tras revisar datos, se comprueba que se trata de errores en build_year, que se van a corregir:

```
MergedActivity$build_year[MergedActivity$vessel_id == "370425"] <- 1992

MergedActivity$build_year[MergedActivity$vessel_id == "813316"] <- 2001

MergedActivity$build_year[MergedActivity$vessel_id == "567313"] <- 2005

MergedActivity$build_year[MergedActivity$vessel_id == "1229111"] <- 2005

# Verificación
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(build_year >= 1800 & build_year <= 2015) %>%
  mutate(antiguedad = year(as.Date(date)) - year(as.Date(paste0(build_year, "-01-01")))) %>%
  filter(antiguedad < 0) %>%
  select(vessel_id, vessel_name, imo_number, event_type, date, build_year, antiguedad) %>%
  knitr::kable("html")%>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size = 12)
```

vessel_id	vessel_name	imo_number	event_type	date	build_year	antiguedad
568186	MIDNIGHT SUN	9232278	Damage to the Environment	2002-09-08	2003	-1
722960	ALASKAN EXPLORER	9244661	Damage to the Environment	2004-10-04	2005	-1
1052777	OPTI-EX		Damage to the Environment	2010-01-15	2011	-1
1110524	FSV6	9664988	Grounding	2012-11-04	2013	-1

1.3. Volumen (gross_ton)

```
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(gross_ton >= 1 & gross_ton <= 250000) %>%
  ggplot(aes(x = gross_ton)) +
  geom_density(fill = "#00bfc4", color = "#00bfc4", alpha = 0.9) +
  facet_wrap(~cut(gross_ton, breaks = c(0, 1000, 250000), labels = c("1-1000", "1000-250000")), nrow = 2, scales = "free") +
  labs(title = "Volumen (Gross Tonnage)", x = "Gross Tonnage", y = "Densidad") +
  theme_minimal()
```

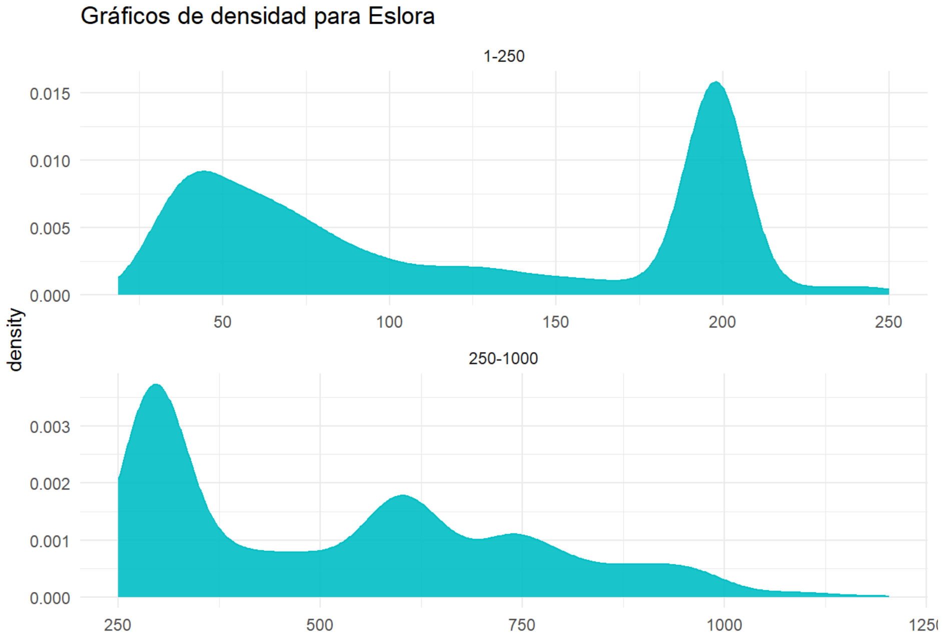


```
# Barcos con mayor Gross Tonnage
MergedActivity %>%
  select(vessel_id, imo_number, vessel_name, build_year, gross_ton, length) %>%
  arrange(desc(gross_ton)) %>%
  unique() %>%
  head(10) %>%
  knitr::kable("html")%>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size = 12)
```

vessel_id	imo_number	vessel_name	build_year	gross_ton	length
933483	9383936	OASIS OF THE SEAS	2009	225282	1187.0
933484	9383948	ALLURE OF THE SEAS	2010	225282	1181.0
228358	7708314	BERGE PIONEER	1980	188728	1071.7
437660	9102239	RAMLAH	1996	163882	1115.5
617142	9241114	ENERGY R	2003	161306	1092.4
586555	9230880	OVERSEAS MULAN	2002	161233	1092.0
938329	9315367	SPYROS	2007	161175	1092.4
1039985	9386964	DORRA	2009	160782	1092.6
606521	9247182	ABQAIQ	2002	159990	1093.4
881630	9312494	MAERSK NAUTILUS	2006	159911	1091.9

1.4. Eslora (length)

```
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(length >= 1 & length <= 1250) %>%
  ggplot(aes(x = length)) +
  geom_density(fill = "#00bfc4", color = "#00bfc4", alpha = 0.9) +
  facet_wrap(~cut(length, breaks = c(1, 250, 1250), labels = c("1-250", "250-1000")), nrow = 2, scales = "free") +
  labs(title = "Gráficos de densidad para Eslora", x = "") +
  theme_minimal()
```

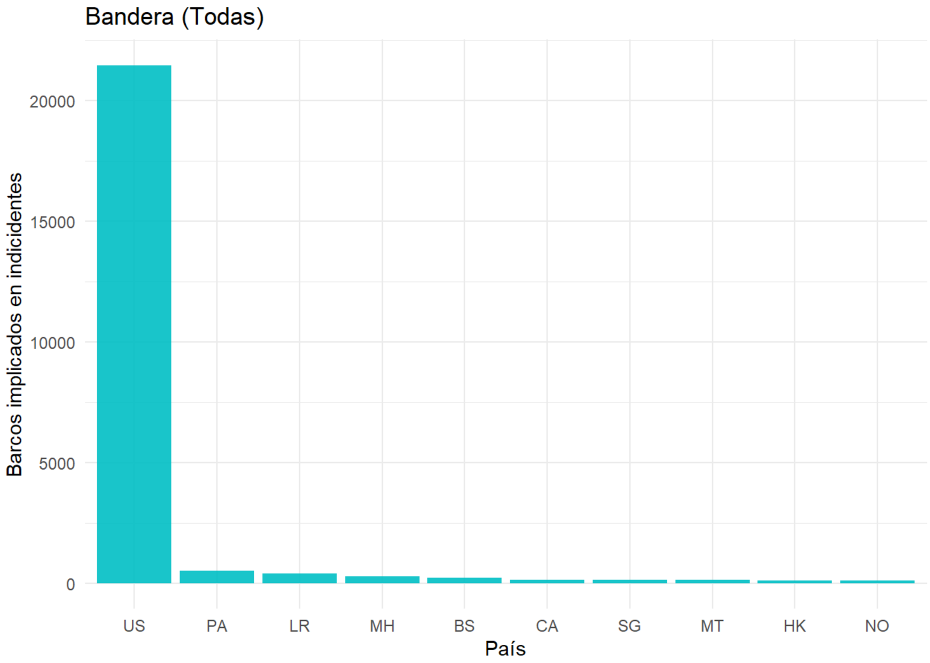


```
# Barcos con mayor eslora
MergedActivity %>%
  select(vessel_id, imo_number, vessel_name, build_year, gross_ton, length) %>%
  arrange(desc(length)) %>%
  unique() %>%
  head(10) %>%
  knitr::kable("html")%>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size = 12)
```

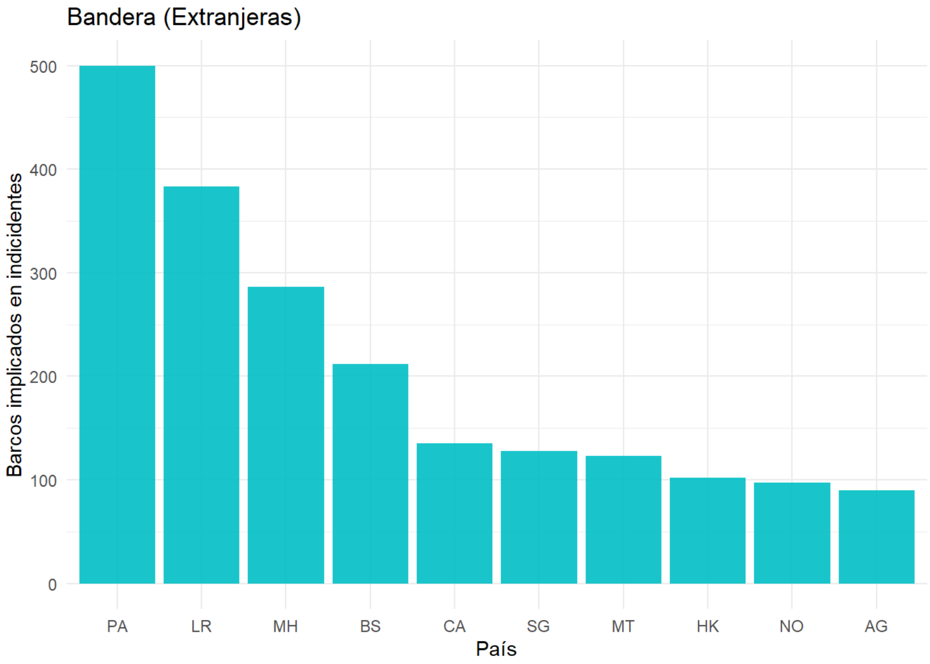
vessel_id	imo_number	vessel_name	build_year	gross_ton	length
1001188	9302889	GRETE MAERSK	2005	97933	1203.8
998455	9302877	GUDRUN MAERSK	2005	97933	1203.8
1008200	9359040	MARIT MAERSK	2009	98268	1203.7
1028411	9359052	MATHILDE MAERSK	2009	98268	1203.7
999387	9359014	MARCHEN MAERSK	2007	98268	1203.7
933483	9383936	OASIS OF THE SEAS	2009	225282	1187.0
933484	9383948	ALLURE OF THE SEAS	2010	225282	1181.0
1026695	9365805	CMA CGM IVANHOE	2008	111249	1148.0
489733	9166778	SVEND MAERSK	1999	91560	1138.3
500247	9166780	SOROE MAERSK	1999	91560	1138.3

1.5. Bandera (flag_abbr)

```
# Gráfico de barras con top10 banderas
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  group_by(flag_abbr) %>%
  summarise(frecuencia = n()) %>%
  arrange(desc(frecuencia)) %>%
  head(10) %>%
  ggplot(aes(x = fct_reorder(flag_abbr, frecuencia, desc), y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  labs(title = "Bandera (Todas)", x = "País", y = "Barcos implicados en indicidentes") +
  theme_minimal()
```

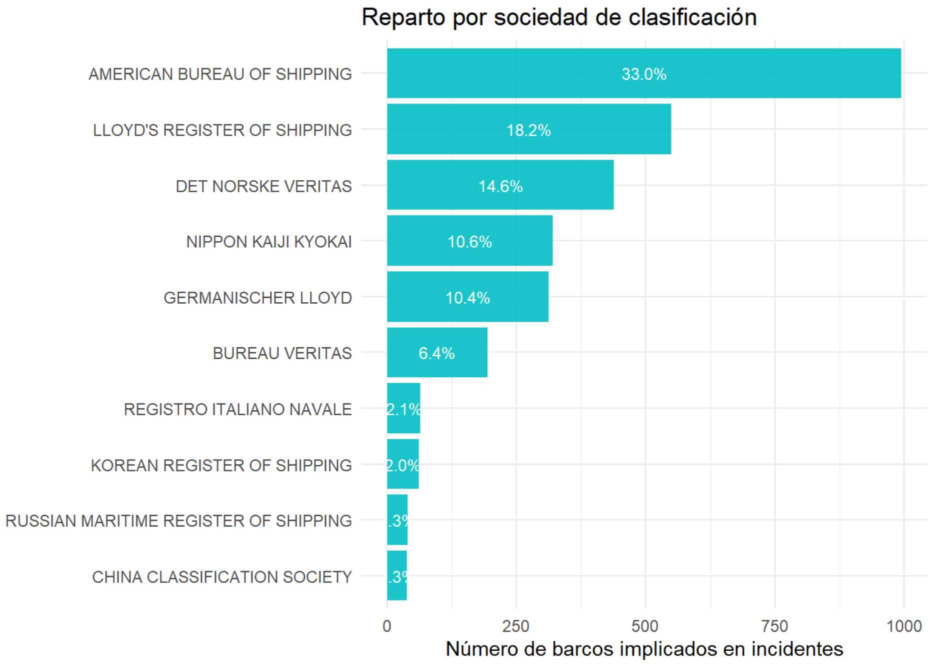


```
# Gráfico de barras top10 sin bandera local (EEUU)
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(flag_abbr != "US") %>%
  group_by(flag_abbr) %>%
  summarise(frecuencia = n()) %>%
  arrange(desc(frecuencia)) %>%
  head(10) %>%
  ggplot(aes(x = fct_reorder(flag_abbr, frecuencia, desc), y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  labs(title = "Bandera (Extranjeras)", x = "País", y = "Barcos implicados en indicidentes") +
  theme_minimal()
```



1.6. Sociedad de clasificación (classification_society)

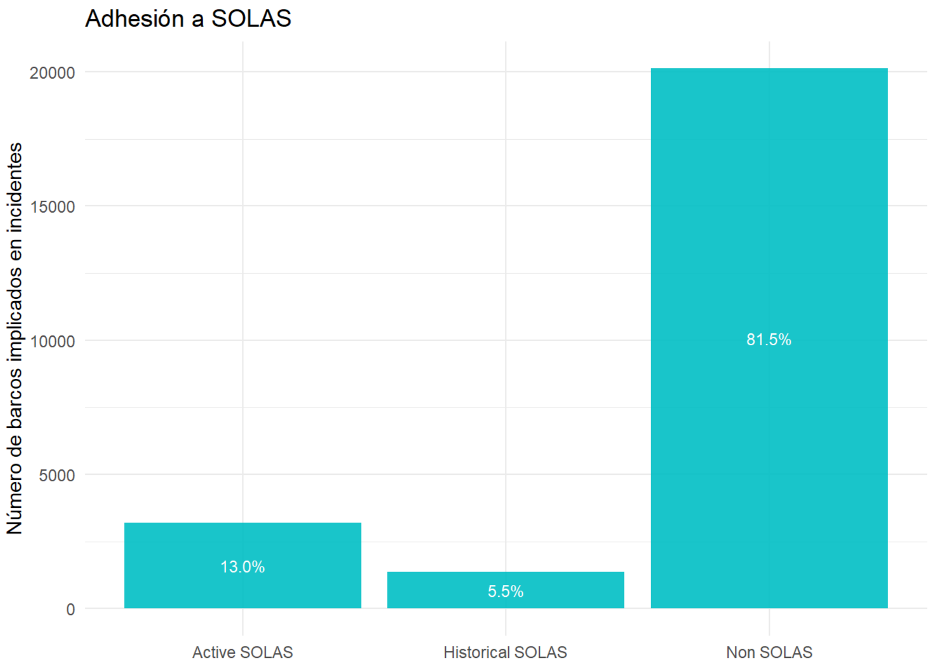
```
# Gráfico de barras horizontales para top10 sociedad de clasificación
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(classification_society != "UNSPECIFIED") %>%
  group_by(classification_society) %>%
  summarise(frecuencia = n()) %>%
  arrange(desc(frecuencia)) %>%
  head(10) %>%
  mutate(porcentaje = frecuencia / sum(frecuencia) * 100) %>%
  ggplot(aes(x = fct_reorder(classification_society, frecuencia), y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  geom_text(aes(label = sprintf("%.1f%%", porcentaje)), position = position_stack(vjust = 0.5), color = "white", size = 3) +
  labs(title = "Reparto por sociedad de clasificación", x = NULL, y = "Número de barcos implicados en incidentes") +
  theme_minimal() +
  coord_flip()
```



1.7. Safety of Life at Sea, SOLAS (solas_desc)

Adhesión al convenio Internacional para la Seguridad de la Vida Humana en el Mar

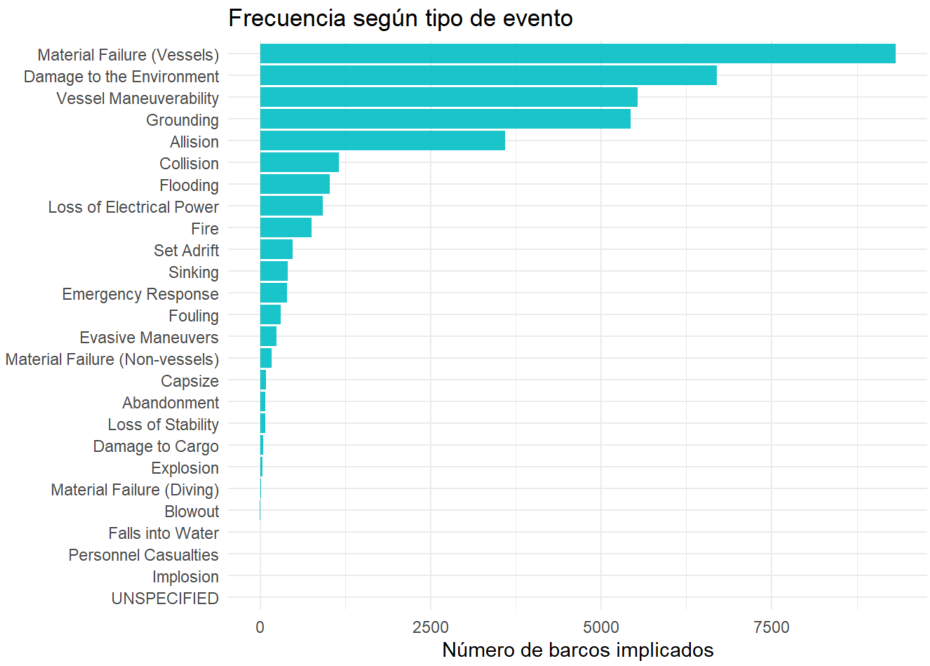
```
# Gráfico de barras para SOLAS
MergedActivity %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  group_by(solas_desc) %>%
  summarise(frecuencia = n()) %>%
  mutate(porcentaje = frecuencia / sum(frecuencia) * 100) %>%
  ggplot(aes(x = solas_desc, y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  geom_text(aes(label = sprintf("%.1f%%", porcentaje)), position = position_stack(vjust = 0.5), color = "white", size = 3) +
  labs(title = "Adhesión a SOLAS", x = NULL, y = "Número de barcos implicados en incidentes") +
  theme_minimal()
```



2. Incidentes

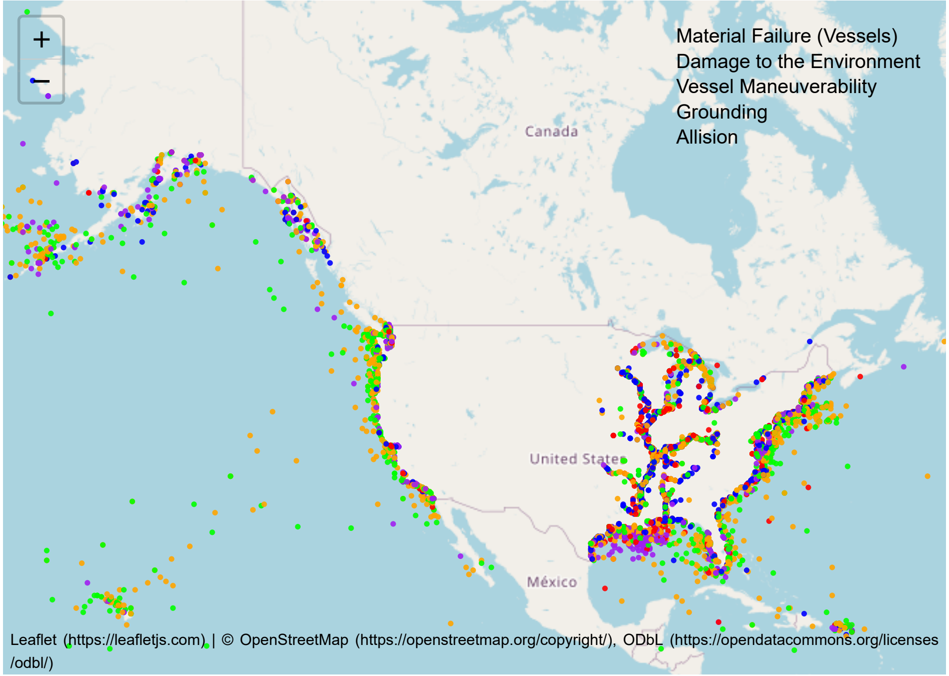
2.1 Tipo de incidente (envent_type)

```
# Gráfico de barras
MergedActivity %>%
  distinct(activity_id, .keep_all = TRUE) %>%
  group_by(event_type) %>%
  summarise(frecuencia = n()) %>%
  ggplot(aes(x = fct_reorder(event_type, frecuencia), y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  labs(title = "Frecuencia según tipo de evento", x = NULL, y = "Número de barcos implicados") +
  theme_minimal() +
  coord_flip()
```



2.2. Localización de incidentes (event_type)

```
# Eventos más frecuentes
top5MergedActivity <- MergedActivity %>%
  distinct(activity_id, .keep_all = TRUE) %>%
  group_by(event_type) %>%
  summarise(frecuencia = n()) %>%
  arrange(desc(frecuencia)) %>%
  head(5)

# Paleta de colores
pal <- colorFactor(
  palette = c('red', 'purple', 'blue', 'orange', 'green'),
  domain = top5MergedActivity$event_type
)

# Top5 MergedActivity
MergedActivity %>%
  filter(event_type %in% top5MergedActivity$event_type) %>%
  sample_frac(0.25) %>%
  # Representación sobre mapa
  leaflet() %>%
    setView(lng = -112, lat = 48, zoom = 3) %>%
    addTiles() %>%
    # Eventos
    addCircleMarkers(lat =~latitude, lng =~longitude,
      radius = 2,
      popup=~paste("activity id:", activity_id, "<br>",
                 "vessel_id:", vessel_id, "<br>",
                 "date:", date, "<br>",
                 "event_type:", event_type, "<br>",
                 "watertype:", watertype, "<br>",
                 "longitude:", longitude, "<br>",
                 "latitude:", latitude, "<br>"
                 ),
      fillOpacity = 0.9,
      color = ~pal(event_type),
      stroke = FALSE
    ) %>%
    # Legenda
    addLegend(position = "topright",
            colors = pal(top5MergedActivity$event_type),
            labels = top5MergedActivity$event_type,
            opacity = 0.5
  )
```


+
−
Material Failure (Vessels)
Damage to the Environment
Vessel Maneuverability
Grounding
Allision
Canada
United States
México
Leaflet (https://leafletjs.com) | © OpenStreetMap (https://openstreetmap.org/copyright/), ODbL (https://opendatacommons.org/licenses/odbl/)

2.3. Localización de incidentes (event_class)

```r
# Definir paleta de colores para cada zona
pal <- colorFactor(
  palette = c('red', 'purple', 'blue', 'orange', 'green', 'yellow'),
  domain = MergedActivity$event_class
)

# Representación sobre mapa (15% de observaciones para facilitar la visualización)
leaflet(data = MergedActivity %>% sample_frac(0.15)) %>%
  setView(lng = -112, lat = 48, zoom = 3) %>%
  addTiles() %>%
  # Color del área
  addRectangles(-122, 49, -180, 70, fillColor = pal("Alaska"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-45, 49, -122, 70, fillColor = pal("Canada"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-45, 15, -81.5, 49, fillColor = pal("East Coast"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-100, 15, -180, 49, fillColor = pal("West Coast"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-81.5, 15, -100, 31, fillColor = pal("Gulf of Mexico"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-81.5, 31, -100, 49, fillColor = pal("Mississippi"), fillOpacity = 0.1, stroke = FALSE) %>%
  # Eventos
  addCircleMarkers(lat =~latitude, lng =~longitude,
    radius = 2,
    popup=~paste("activity id:", activity_id, "<br>",
                 "vessel_id:", vessel_id, "<br>",
                 "date:", date, "<br>",
                 "event_type:", event_type, "<br>",
                 "watertype:", watertype, "<br>",
                 "longitude:", longitude, "<br>",
                 "latitude:", latitude, "<br>"
                 ),
    fillOpacity = 0.9,
    color = ~pal(event_class),
    stroke = FALSE
  ) %>%
  # Legenda
  addLegend(position = "topright",
          colors = pal(sort(unique(MergedActivity$event_class))),
          labels = sort(unique(MergedActivity$event_class)),
          title = "Clase de incidente"
  )
```

```
## Warning in pal("Mississippi"): Some values were outside the color scale and
## will be treated as NA
```

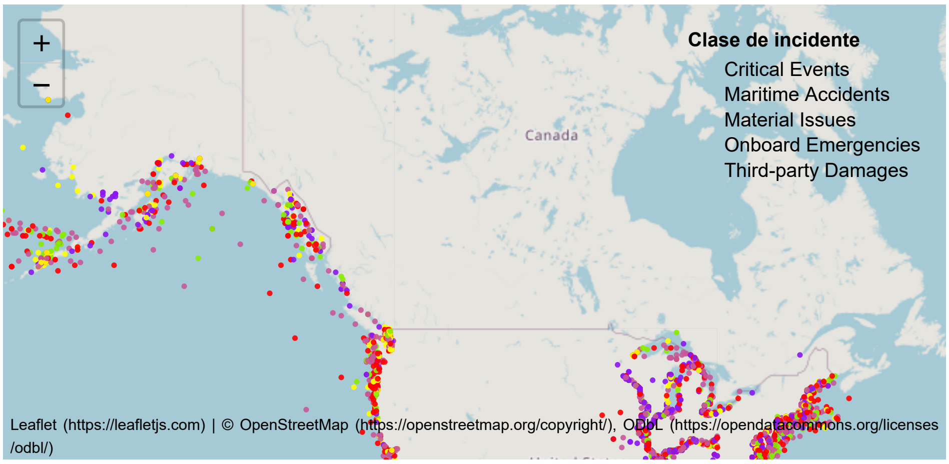
```
## Warning in pal("Gulf of Mexico"): Some values were outside the color scale and
## will be treated as NA
```

```
## Warning in pal("West Coast"): Some values were outside the color scale and will
## be treated as NA
```

```
## Warning in pal("East Coast"): Some values were outside the color scale and will
## be treated as NA
```

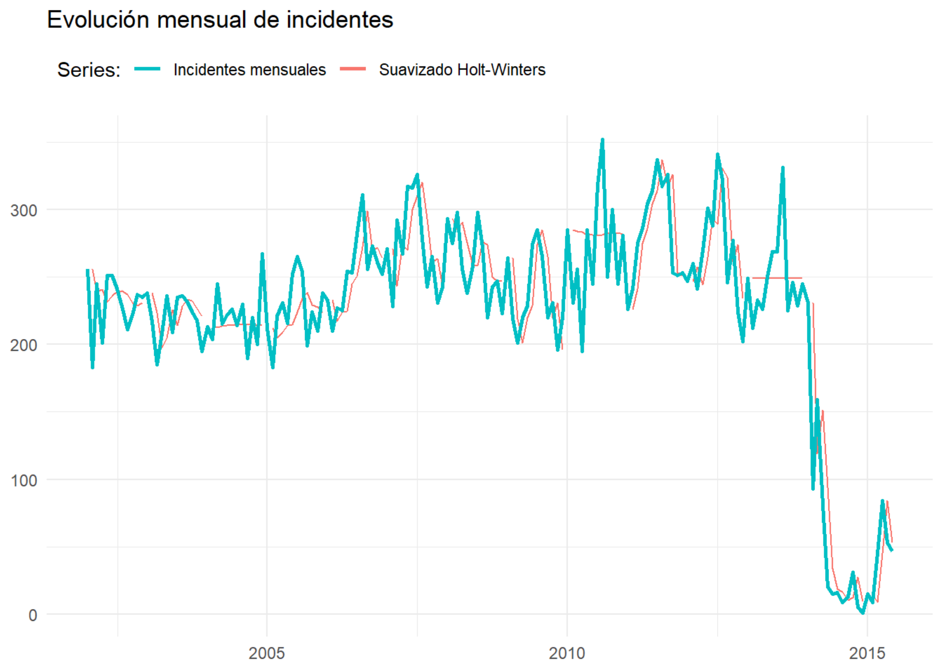
```
## Warning in pal("Canada"): Some values were outside the color scale and will be
## treated as NA
```

```
## Warning in pal("Alaska"): Some values were outside the color scale and will be
## treated as NA
```


+
−
Canada
Clase de incidente
Critical Events
Maritime Accidents
Material Issues
Onboard Emergencies
Third-party Damages
Leaflet (https://leafletjs.com) | © OpenStreetMap (https://openstreetmap.org/copyright/), ODbL (https://opendatacommons.org/licenses/odbl/)

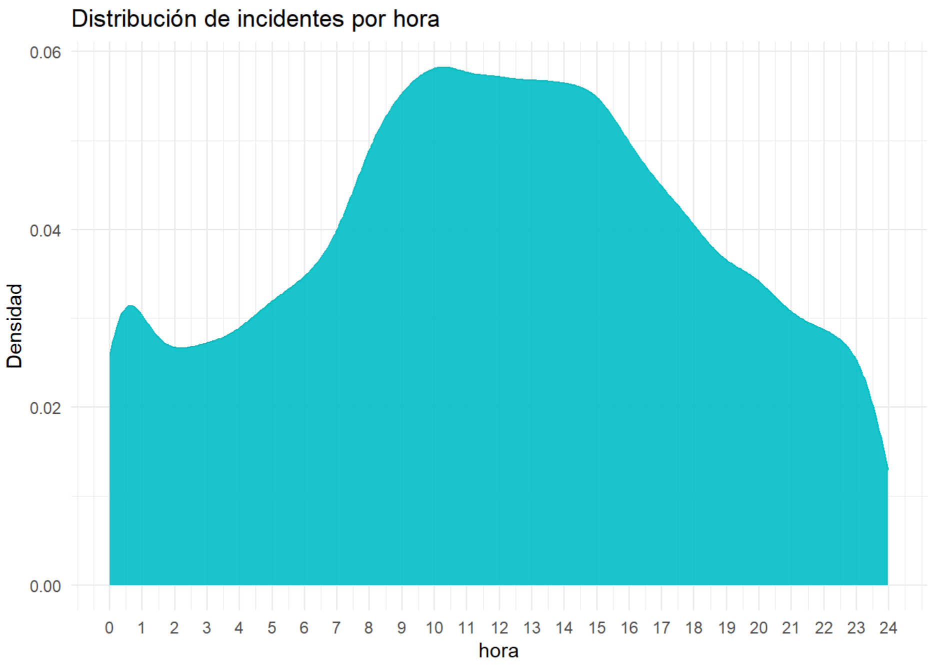
2.4. Evolución temporal (envent_type)

```
# Se añade un alisado tipo Holt-Winters
MergedActivity %>%
  distinct(activity_id, .keep_all = TRUE) %>%
  group_by(año = lubridate::year(date), mes = lubridate::month(date)) %>%
  summarise(incidentes_mes = n()) %>%
  mutate(Fecha = as.Date(paste(año, mes, "01", sep = "-"))) %>%
  mutate(incidentes_alisado = c(NA, HoltWinters(incidentes_mes, beta = FALSE, gamma = FALSE)$fitted[, "level"])) %>%
  arrange(año, mes) %>%
  ungroup() %>%
  ggplot() +
  geom_line(aes(x = Fecha, y = incidentes_alisado, color = "Suavizado Holt-Winters")) +
  geom_line(aes(x = Fecha, y = incidentes_mes, color = "Incidentes mensuales"), size = 1) +
  scale_color_manual(values = c("Incidentes mensuales" = "#00bfc4", "Suavizado Holt-Winters" = "#f8766d"), guide = guide_leg
end(title = "Series:")) +
  labs(title = "Evolución mensual de incidentes", x = NULL, y = NULL) +
  theme_minimal() +
  theme(legend.position = "top", legend.justification = "left")
```

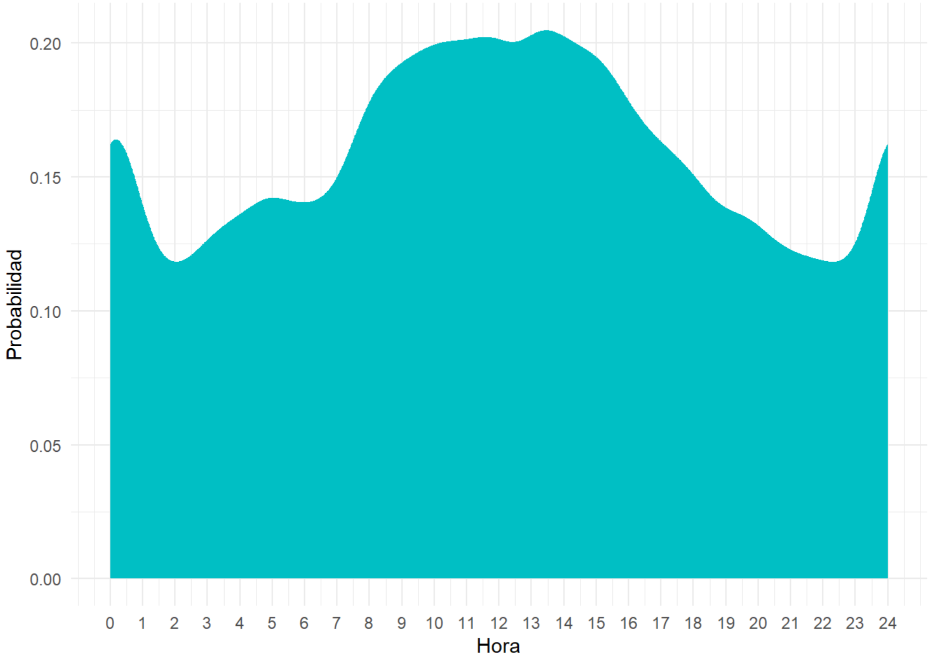


2.5. Hora de los incidentes

```
# Representación sencilla pero imprecisa
MergedActivity %>%
  distinct(activity_id, .keep_all = TRUE) %>%
  mutate(hora = round(as.numeric(sub(":.*", "", hour)) + (as.numeric(sub(".*:", "", hour)) / 60), 2)) %>%
  ggplot(aes(x = hora)) +
  geom_density(fill = "#00bfc4", color = "#00bfc4", alpha = 0.9) +
  scale_x_continuous(labels = 0:24, breaks = 0:24) +
  labs(title = "Distribución de incidentes por hora", y = "Densidad") +
  theme_minimal()
```



```r
# Teniendo en cuenta la "circularidad" de las horas
# Código basado en el post https://stackoverflow.com/questions/48407745/density-plot-based-on-time-of-the-day
datetimes = MergedActivity$hour %>%
  lubridate::parse_date_time("%h:%M")
times_in_decimal = lubridate::hour(datetimes) + lubridate::minute(datetimes) / 60
times_in_radians = 2 * pi * (times_in_decimal / 24)

# Estimación para ancho de banda
basic_dens = density(times_in_radians, from = 0, to = 2 * pi)

res = circular::density.circular(circular::circular(times_in_radians,
                                                    type = "angle",
                                                    units = "radians",
                                                    rotation = "clock"),
                                 kernel = "wrappednormal",
                                 bw = basic_dens$bw)

time_pdf = data.frame(time = as.numeric(24 * (2 * pi + res$x) / (2 * pi)), # Radianes a 24h
                      likelihood = res$y)

ggplot(time_pdf) +
  geom_area(aes(x = time, y = likelihood), fill = "#00bfc4") +
  scale_x_continuous("Hora", labels = 0:24, breaks = 0:24) +
  scale_y_continuous("Probabilidad") +
  theme_minimal()
```



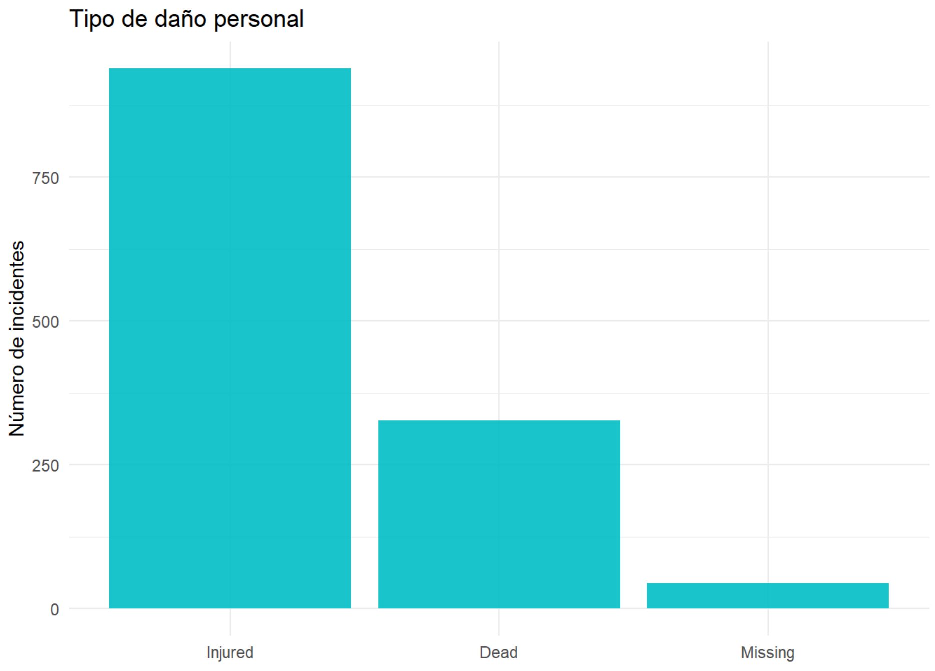
2.6. Valoración económica

```
# Top 10 incidentes con mayor perjuicio económico
MergedActivity %>%
  distinct(activity_id, .keep_all = TRUE) %>%
  select(activity_id, vessel_name, event_type, damage_assessment) %>%
  arrange(desc(damage_assessment)) %>%
  mutate(damage_assessment = format(damage_assessment, scientific = FALSE)) %>%
  head(10) %>%
  knitr::kable("html")%>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size = 12)
```

activity_id	vessel_name	event_type	damage_assessment
3960377	RIVER ELEGANCE	Evasive Maneuvers	350000000
3964637	C 533	Allision	350000000
3965662	SUNSET I	Grounding	350000000
1966333	RICHARD A BAKER	Grounding	98000000
1896064	JAY LUHR	Material Failure (Vessels)	85000000
1900683	RUBY RIVER	Allision	85000000
1902882	GILBERT TAYLOR	Material Failure (Vessels)	85000000
2865301	KIRBY 28037	Allision	60000000
2865600	KIRBY 30026B	Material Failure (Vessels)	60000000
2870902	HARRY J. BROCK	Grounding	60000000

2.7. Daños personales

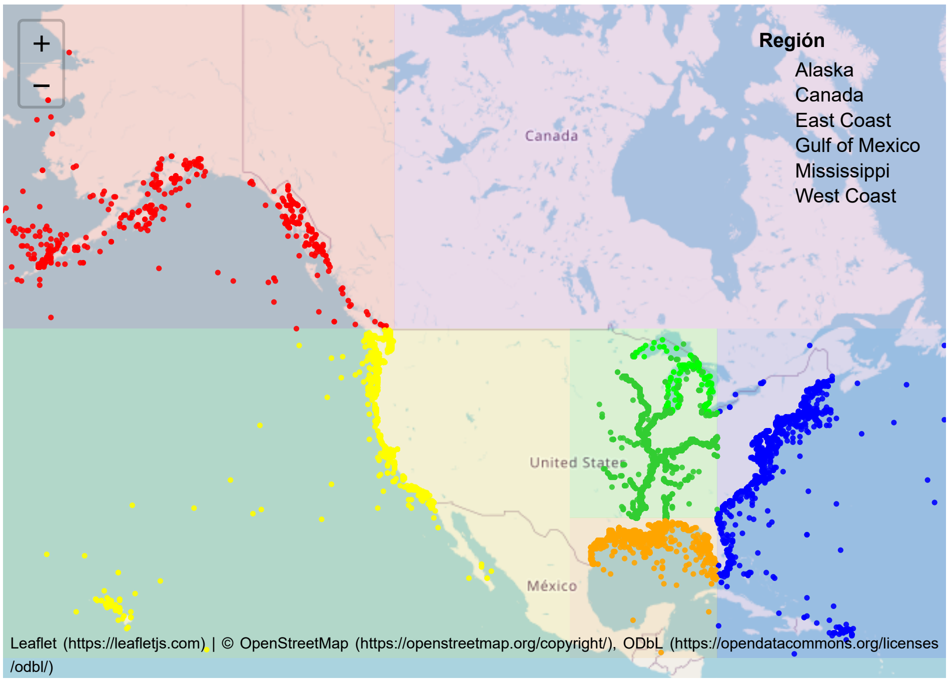
```
# Frecuencia de daños personales
# Gráfico de barras para barcos con incidente
MergedActivity %>%
  distinct(activity_id, .keep_all = TRUE) %>%
  filter(!is.na(casualty), casualty != "UNSPECIFIED") %>%
  group_by(casualty) %>%
  summarise(frecuencia = n()) %>%
  ggplot(aes(x = fct_reorder(casualty, frecuencia, desc), y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  labs(title = "Tipo de daño personal", x = NULL, y = "Número de incidentes") +
  theme_minimal()
```



3. Geografía

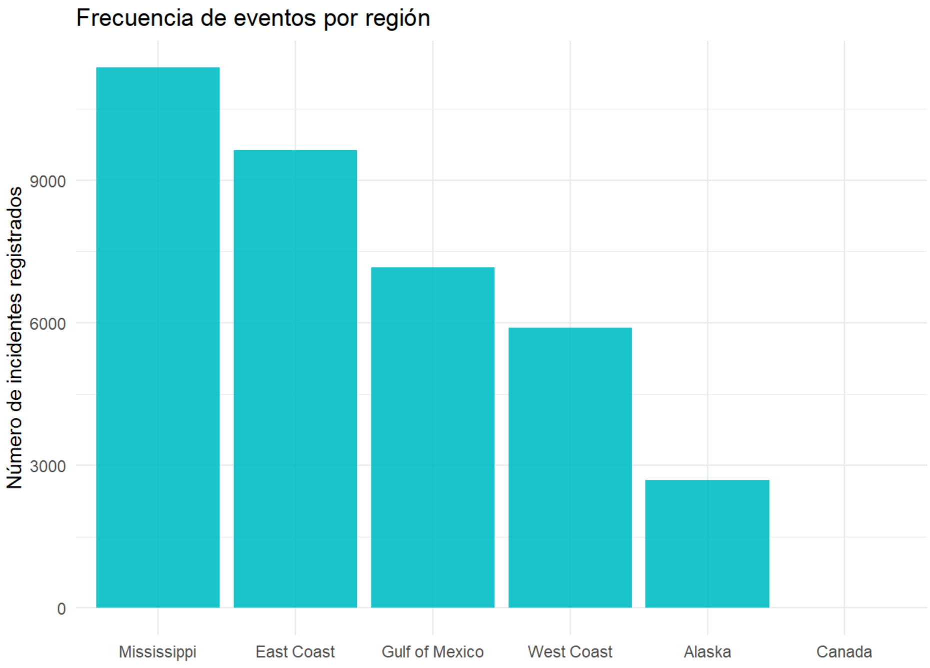
3.1. Localización de regiones

```
# Definir paleta de colores para cada zona
pal <- colorFactor(
  palette = c('red', 'purple', 'blue', 'orange', 'green', 'yellow'),
  domain = MergedActivity$region
)

# Representación sobre mapa (15% de observaciones para facilitar la visualización)
leaflet(data = MergedActivity %>% sample_frac(0.15)) %>%
  setView(lng = -112, lat = 48, zoom = 3) %>%
  addTiles() %>%
  # Color del área
  addRectangles(-122, 49, -180, 70, fillColor = pal("Alaska"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-45, 49, -122, 70, fillColor = pal("Canada"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-45, 15, -81.5, 49, fillColor = pal("East Coast"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-100, 15, -180, 49, fillColor = pal("West Coast"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-81.5, 15, -100, 31, fillColor = pal("Gulf of Mexico"), fillOpacity = 0.1, stroke = FALSE) %>%
  addRectangles(-81.5, 31, -100, 49, fillColor = pal("Mississippi"), fillOpacity = 0.1, stroke = FALSE) %>%
  # Eventos
  addCircleMarkers(lat =~latitude, lng =~longitude,
    radius = 2,
    popup=~paste("activity id:", activity_id, "<br>",
                 "vessel_id:", vessel_id, "<br>",
                 "date:", date, "<br>",
                 "event_type:", event_type, "<br>",
                 "watertype:", watertype, "<br>",
                 "longitude:", longitude, "<br>",
                 "latitude:", latitude, "<br>"
                 ),
    fillOpacity = 0.9,
    color = ~ifelse(watertype == "river", 'limegreen', pal(region)),
    stroke = FALSE
  ) %>%
  # Legenda
  addLegend(position = "topright",
            colors = pal(sort(unique(MergedActivity$region))),
            labels = sort(unique(MergedActivity$region)),
            title = "Región"
  )
```


+
−
Región
Alaska
Canada
East Coast
Gulf of Mexico
Mississippi
West Coast
Canada
United States
México
Leaflet (https://leafletjs.com) | © OpenStreetMap (https://openstreetmap.org/copyright/), ODbL (https://opendatacommons.org/licenses/odbl/)

3.2. Eventos por región

```
# Gráfico de barras
MergedActivity %>%
  distinct(activity_id, .keep_all = TRUE) %>%
  group_by(region) %>%
  summarise(frecuencia = n()) %>%
  ggplot(aes(x = fct_reorder(region, frecuencia, desc), y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Frecuencia de eventos por región", x = NULL, y = "Número de incidentes registrados") +
  theme_minimal()
```



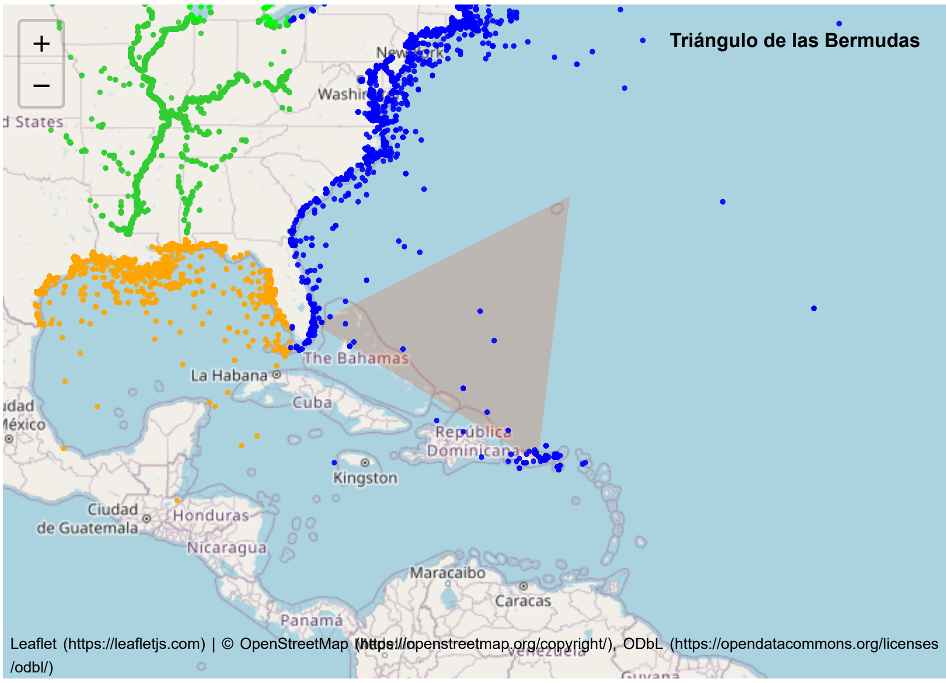
3.3. Evento más común en cada región

```
# Extracción del evento con mayor frecuencia en cada región
MergedActivity %>%
  distinct(activity_id, .keep_all = TRUE) %>%
  group_by(region) %>%
  mutate(num_sucesos_por_region = n()) %>%
  mutate(suceso_mas_frecuente = event_type[which.max(n())]) %>%
  select(region, suceso_mas_frecuente, num_sucesos_por_region) %>%
  unique() %>%
  arrange(desc(num_sucesos_por_region)) %>%
  knitr::kable("html")%>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size = 12)
```

region	suceso_mas_frecuente	num_sucesos_por_region
Mississippi	Grounding	11374
East Coast	Damage to the Environment	9630
Gulf of Mexico	Grounding	7163
West Coast	Vessel Maneuverability	5900
Alaska	Damage to the Environment	2692
Canada	Material Failure (Vessels)	1

3.4. Bonus: Triángulo de las bermudas

```
# Definir coordenadas de la zona
triang_bermuda <- data.frame(
  lng = c(-64, -80, -66),
  lat = c(33, 26, 18)
)

# Representación sobre mapa (15% de observaciones para facilitar la visualización)
leaflet(data = MergedActivity %>% sample_frac(0.15)) %>%
  setView(lng = -70, lat = 25, zoom = 4) %>%
  addTiles() %>%

  addPolygons(data = triang_bermuda, lat = ~lat, lng = ~lng, fillColor = "orangered", stroke = FALSE) %>%
  # Eventos
  addCircleMarkers(lat =~latitude, lng =~longitude,
    radius = 2,
    popup=~paste("activity id:", activity_id, "<br>",
                 "vessel_id:", vessel_id, "<br>",
                 "date:", date, "<br>",
                 "event_type:", event_type, "<br>",
                 "watertype:", watertype, "<br>",
                 "longitude:", longitude, "<br>",
                 "latitude:", latitude, "<br>"
                 ),
    fillOpacity = 0.9,
    color = ~ifelse(watertype == "river", 'limegreen', pal(region)),
    stroke = FALSE
  ) %>%
  # Legenda
  addLegend(position = "topright",
            colors = "orangered",
            labels = "",
            title = "Triángulo de las Bermudas"
  )
```


+
−
Triángulo de las Bermudas
New York
Washi
d States
The Bahamas
La Habana
Cuba
República Dominicana
Kingston
idad
éxico
Ciudad de Guatemala
Honduras
Nicaragua
Maracaibo
Caracas
Panamá
Venezuela
Guyana
Leaflet (https://leafletjs.com) | © OpenStreetMap (https://openstreetmap.org/copyright/), ODbL (https://opendatacommons.org/licenses/odbl/)

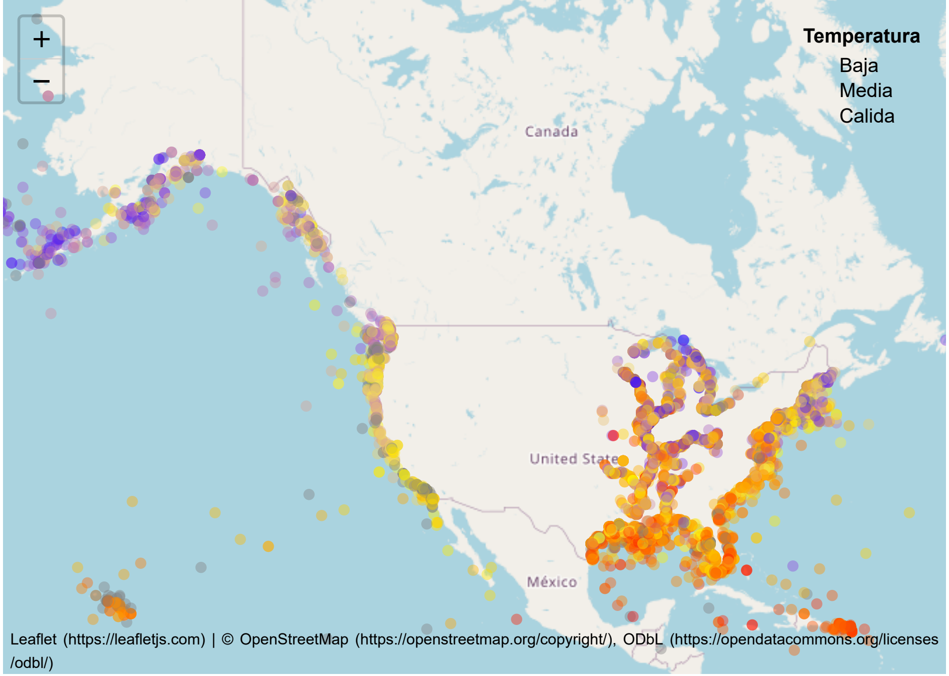
Nota: No se aprecia una mayor concentración de incidentes que otras zonas con distancias similares a la costa

4. Metereología

4.1. Temperatura

Mapa

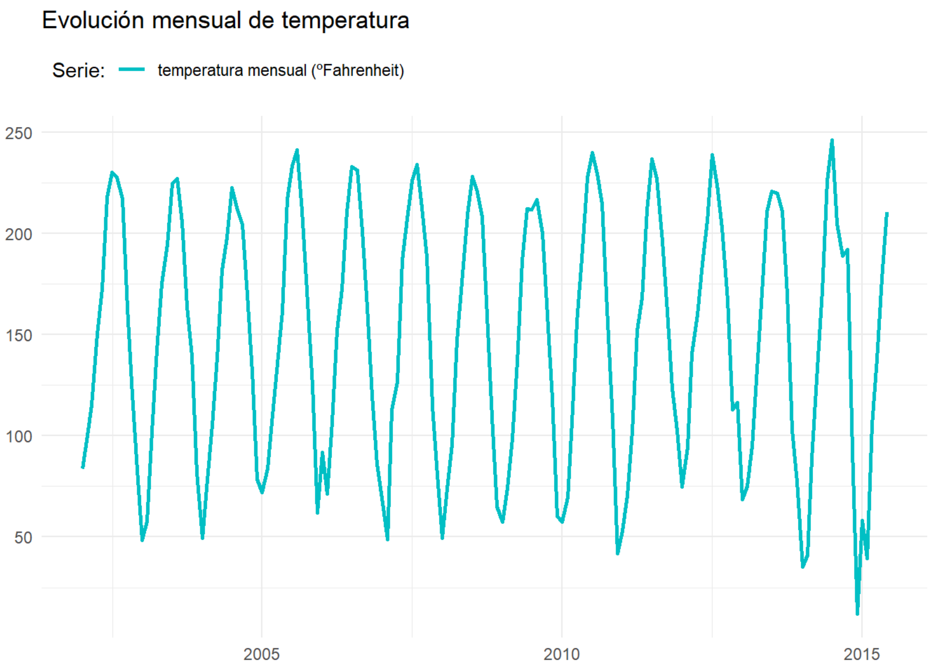
```
# Definir paleta de colores por intensidad
pal <- colorFactor(
  palette = c('blue', 'yellow','red'),
  domain = sort(MergedActivity$air_temp)
)

# Representación sobre mapa (15% de observaciones para facilitar la visualización)
leaflet(data = MergedActivity %>% sample_frac(0.15)) %>%
  setView(lng = -112, lat = 48, zoom = 3) %>%
  addTiles() %>%
  # Eventos
  addCircleMarkers(lat =~latitude, lng =~longitude,
    radius = 4,
    popup=~paste("activity id:", activity_id, "<br>",
                 "vessel_id:", vessel_id, "<br>",
                 "date:", date, "<br>",
                 "event_type:", event_type, "<br>",
                 "watertype:", watertype, "<br>",
                 "air_temp:", air_temp, "<br>",
                 "longitude:", longitude, "<br>",
                 "latitude:", latitude, "<br>"
                 ),
    fillOpacity = 0.4,
    color = ~pal(air_temp),
    stroke = FALSE
  ) %>%
  # Legenda
  addLegend(position = "topright",
          colors = c('blue', 'yellow','red'),
          labels = c("Baja", "Media", "Calida"),
          title = "Temperatura"
  )
```


+
−
Temperatura
Baja
Media
Calida
Canada
United States
México
Leaflet (https://leafletjs.com) | © OpenStreetMap (https://openstreetmap.org/copyright/), ODbL (https://opendatacommons.org/licenses/odbl/)

Temperatura durante el año

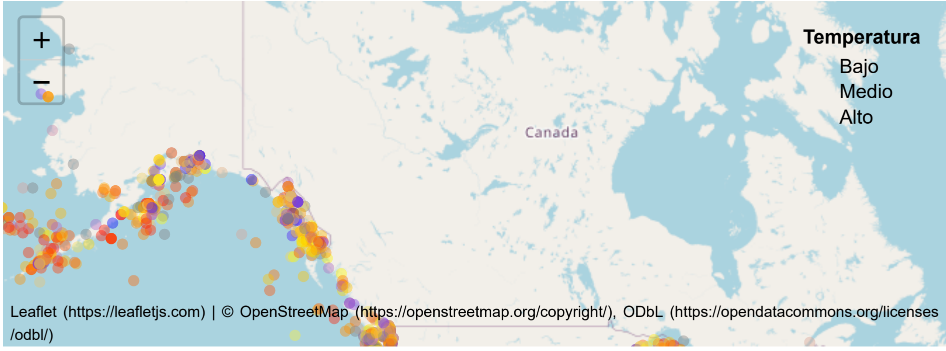
```
# Grafico de serie temporal de temperatura media mensual
MergedActivity %>%
  distinct(activity_id, .keep_all = TRUE) %>%
  group_by(año = lubridate::year(date), mes = lubridate::month(date)) %>%
  summarise(temperatura_mes = mean(air_temp, na.rm = TRUE)) %>%
  mutate(Fecha = as.Date(paste(año, mes, "01", sep = "-"))) %>%
  arrange(año, mes) %>%
  ungroup() %>%
  ggplot() +
  geom_line(aes(x = Fecha, y = temperatura_mes, color = "temperatura mensual (ºFahrenheit)"), size = 1) +
  scale_color_manual(values = c("temperatura mensual (ºFahrenheit)" = "#00bfc4"), guide = guide_legend(title = "Serie:")) +
  labs(title = "Evolución mensual de temperatura", x = NULL, y = NULL) +
  theme_minimal() +
  theme(legend.position = "top", legend.justification = "left")
```

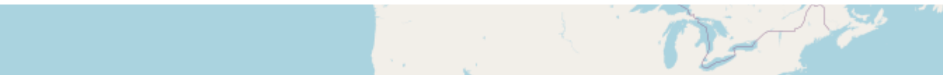
```
## `summarise()` has grouped output by 'año'. You can override using the `.groups`
## argument.
```



4.2. Mapa de viento

```
# Definir paleta de colores para cada zona
pal <- colorFactor(
  palette = c('blue', 'yellow','red'),
  domain = sort(MergedActivity$wind_speed)
)

# Representación sobre mapa (15% de observaciones para facilitar la visualización)
leaflet(data = MergedActivity %>% sample_frac(0.15)) %>%
  setView(lng = -112, lat = 48, zoom = 3) %>%
  addTiles() %>%
  # Eventos
  addCircleMarkers(lat =~latitude, lng =~longitude,
    radius = 4,
    popup=~paste("activity id:", activity_id, "<br>",
                 "vessel_id:", vessel_id, "<br>",
                 "date:", date, "<br>",
                 "event_type:", event_type, "<br>",
                 "watertype:", watertype, "<br>",
                 "longitude:", longitude, "<br>",
                 "latitude:", latitude, "<br>"
                 ),
    fillOpacity = 0.4,
    color = ~pal(wind_speed),
    stroke = FALSE
  ) %>%
  # Legenda
  addLegend(position = "topright",
            colors = c('blue', 'yellow','red'),
            labels = c("Bajo", "Medio", "Alto"),
            title = "Temperatura"
  )
```

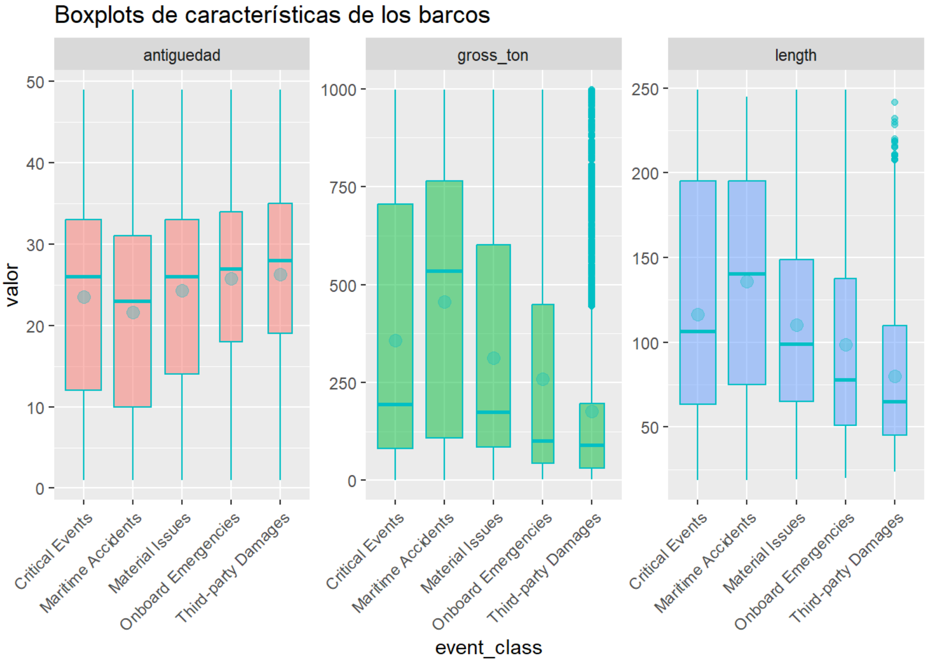

+
−
Canada
Temperatura
Bajo
Medio
Alto
Leaflet (https://leafletjs.com) | © OpenStreetMap (https://openstreetmap.org/copyright/), ODbL (https://opendatacommons.org/licenses/odbl/)



5. event_class / Variables explicativas

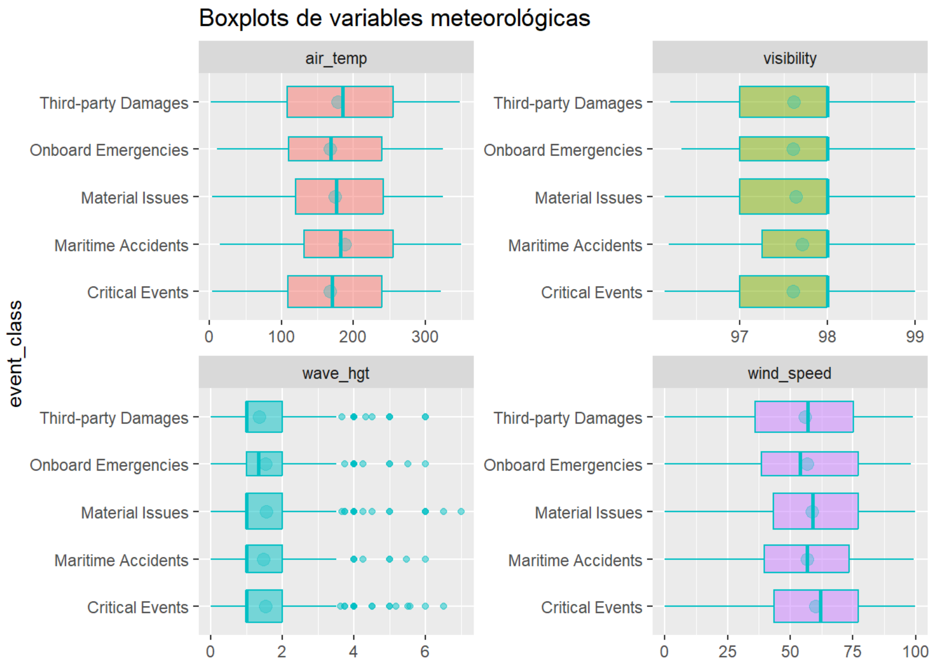
5.1. Variables cuantitativas de características de los barcos

```
# Boxplot de estas variables en un solo gráfico para una salida compacta
MergedActivity %>%
  mutate(build_year = as.numeric(build_year)) %>%
  mutate(antiguedad = year(as.Date(date)) - year(as.Date(paste0(build_year, "-01-01")))) %>%
  filter(gross_ton < 1000, length < 250, antiguedad > 0, antiguedad < 50) %>%
  select(event_class, gross_ton, length, antiguedad) %>%
  pivot_longer(cols = -event_class, names_to = "variable", values_to = "valor") %>%
  ggplot(aes(y = valor, x = event_class, fill = variable)) +
  geom_boxplot(varwidth = TRUE, color = "#00bfc4", alpha = 0.5) +
  stat_summary(fun = mean, geom = "point", color = "#00bfc4", size = 3, alpha = 0.3) +
  facet_wrap(~variable, scales = "free") +
  theme(legend.position="none") +
  labs(title = "Boxplots de características de los barcos") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



5.2. Variables meteorológicas

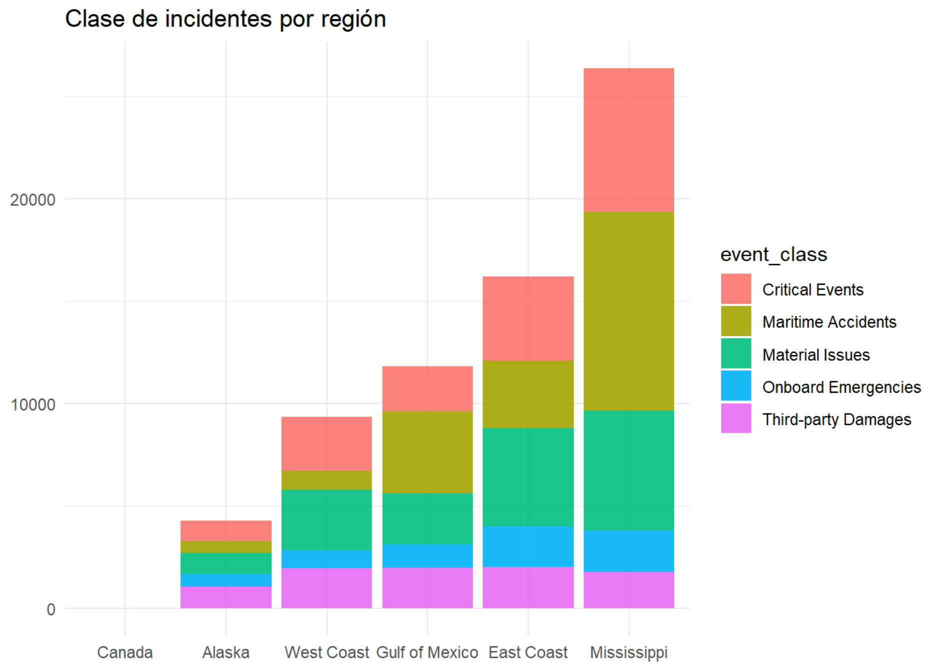
```
# Boxplot de estas variables en un solo gráfico para una salida compacta
MergedActivity %>%
  mutate(build_year = as.numeric(build_year)) %>%
  select(event_class, air_temp, wind_speed, wave_hgt, visibility) %>%
  filter(air_temp > 0, wind_speed < 100, wave_hgt < 10, visibility > 96) %>%
  pivot_longer(cols = -event_class, names_to = "variable", values_to = "valor") %>%
  ggplot(aes(y = valor, x = event_class, fill = variable)) +
  geom_boxplot(varwidth = TRUE, color = "#00bfc4", alpha = 0.5) +
  stat_summary(fun = mean, geom = "point", color = "#00bfc4", size = 3, alpha = 0.3) +
  facet_wrap(~variable, scales = "free") +
  theme(legend.position="none") +
  labs(title = "Boxplots de variables meteorológicas", y = NULL) +
  coord_flip()
```



5.3. Clase de incidentes por región

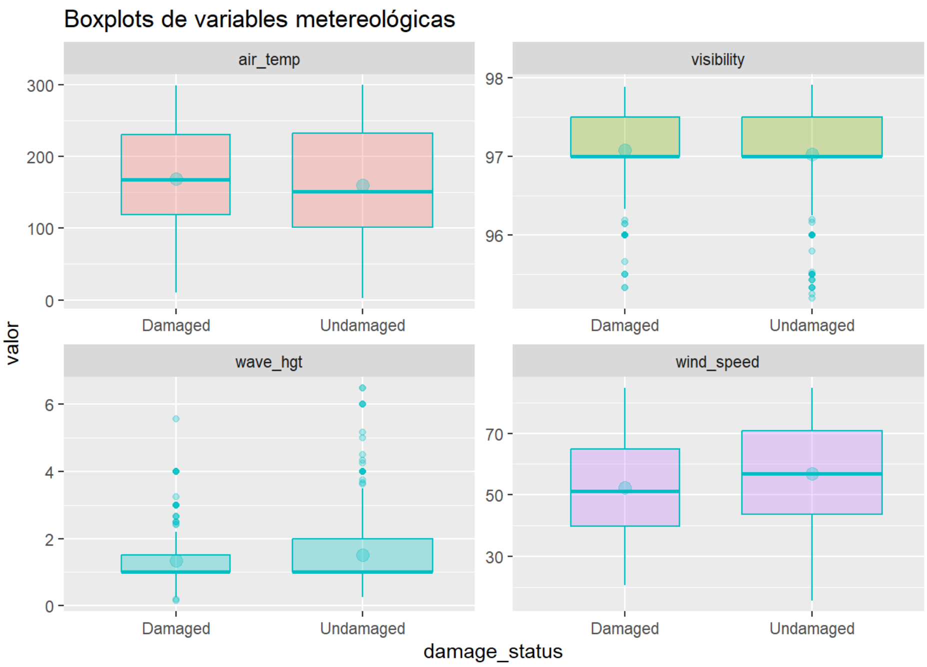
```
# Gráfico de barras apiladas
MergedActivity  %>%
  group_by(region, event_class) %>%
  summarise(frecuencia = n()) %>%
  ggplot(aes(x = fct_reorder(region, frecuencia), y = frecuencia, fill = event_class)) +
  geom_bar(stat = "identity", alpha = 0.9) +
  labs(title = "Clase de incidentes por región", x = NULL, y = NULL) +
  theme_minimal()
```

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```



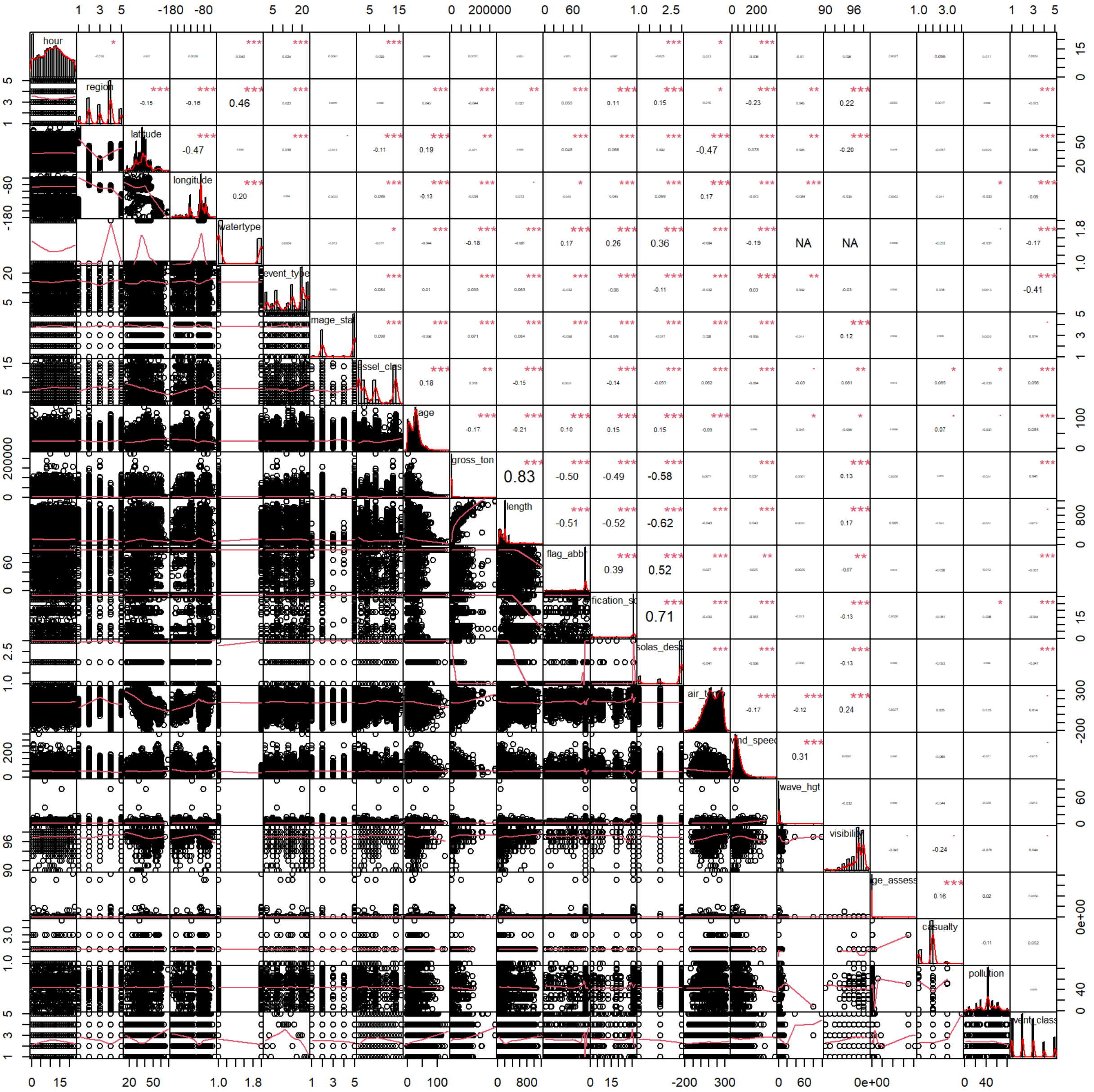
5.X. Daños / Variables meteorológicas

```
# Boxplots de variables meteorológicas
# Filtrados para encuadrar los valores centrales
MergedActivity %>%
  select(damage_status, air_temp, wind_speed, wave_hgt, visibility) %>%
  filter(damage_status == "Damaged" | damage_status == "Undamaged") %>%
  filter(air_temp > 0 & air_temp < 300) %>%
  filter(visibility > 95 & visibility < 98) %>%
  filter(wave_hgt > 0 & wave_hgt < 10) %>%
  filter(wind_speed > 15 & wind_speed < 85) %>%
  pivot_longer(cols = -damage_status, names_to = "variable", values_to = "valor") %>%
  ggplot(aes(y = valor, x = damage_status, fill = variable)) +
  geom_boxplot(varwidth = TRUE, color = "#00bfc4", alpha = 0.3) +
  stat_summary(fun = mean, geom = "point", color = "#00bfc4", size = 3, alpha = 0.3) +
  facet_wrap(~variable, scales = "free") +
  theme(legend.position="none") +
  labs(title = "Boxplots de variables metereológicas")
```



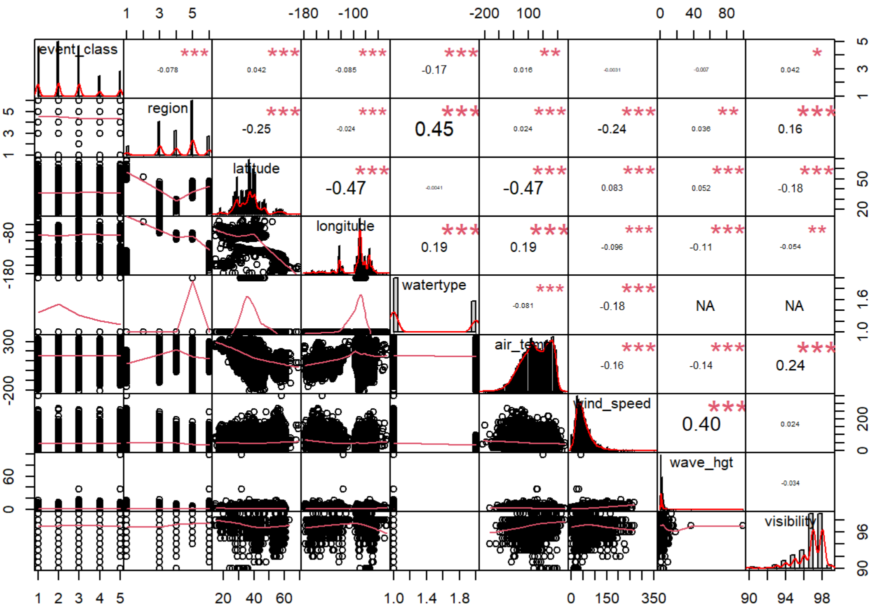
6. Correlaciones

```
# Cuadro conjunto
MergedActivity %>%
  sample_frac(0.3) %>%
  mutate(hour = round(as.numeric(sub(":.*", "", hour)) + (as.numeric(sub(".*:", "", hour)) / 60), 2)) %>%
  select(-activity_id, -date, -build_year, -vessel_id, -imo_number, -vessel_name) %>%
  mutate_at(vars(region, watertype, event_type, damage_status,
                 vessel_class, flag_abbr, classification_society, solas_desc,
                 casualty, pollution, event_class), factor ) %>%
  mutate_all(~as.integer(.)) %>%
  chart.Correlation(histogram = T, pch = 19)
```

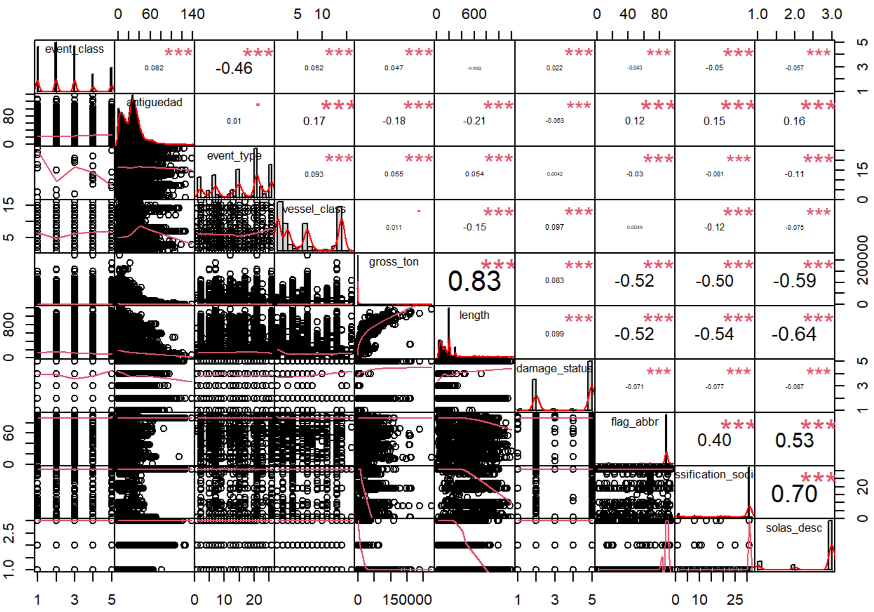


Más en detalle:

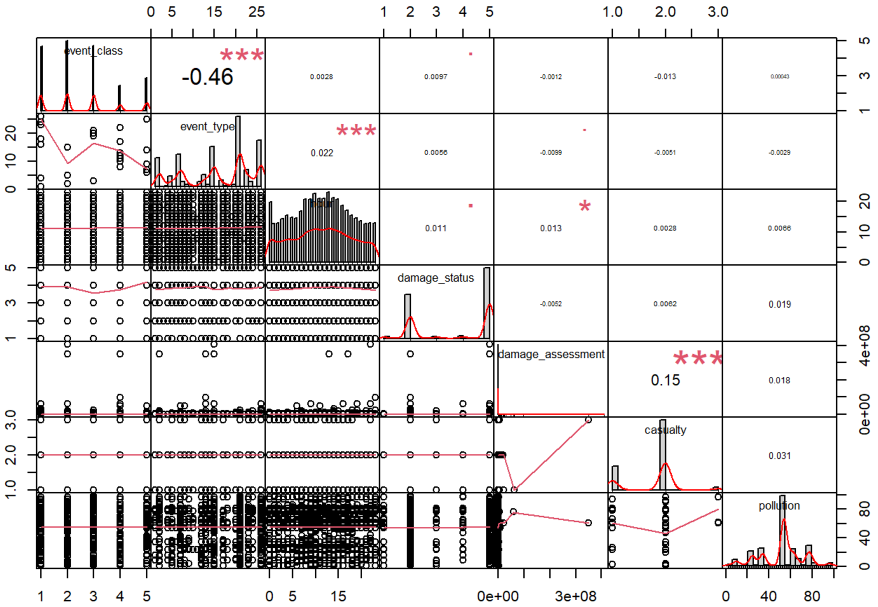
```
# Variables de localización y meteorología
MergedActivity %>%
  sample_frac(0.5) %>%
  select(event_class, region, latitude, longitude, watertype, air_temp, wind_speed, wave_hgt, visibility) %>%
  mutate_at(vars(region, watertype, event_class), factor ) %>%
  mutate_all(~as.integer(.)) %>%
  chart.Correlation(histogram = T, pch = 19)
```



```
# Características de barco
MergedActivity %>%
  sample_frac(0.5) %>%
  mutate(antiguedad = year(as.Date(date)) - year(as.Date(paste0(build_year, "-01-01")))) %>%
  select(event_class, antiguedad, event_type, vessel_class, gross_ton, length, damage_status, flag_abbr, classification_soci
ety, solas_desc) %>%
  mutate_at(vars(event_type, damage_status, vessel_class, flag_abbr, classification_society, solas_desc, event_class), facto
r ) %>%
  mutate_all(~as.integer(.)) %>%
  chart.Correlation(histogram = T, pch = 19)
```



```
# Incidentes
MergedActivity %>%
  sample_frac(0.5) %>%
  mutate(hour = round(as.numeric(sub(":.*", "", hour)) + (as.numeric(sub(".*:", "", hour)) / 60), 2)) %>%
  select(event_class, event_type, hour, damage_status, damage_assessment, casualty, pollution) %>%
  mutate_at(vars(event_type, damage_status, casualty, pollution, event_class), factor ) %>%
  mutate_all(~as.integer(.)) %>%
  chart.Correlation(histogram = T, pch = 19)
```


event_class
event_type
hour
damage_status
damage_assessment
casualty
pollution
-0.46 ***
0.0028
0.0097
-0.0012
-0.013
0.00043
0.022 ***
0.0056
-0.0099
-0.0051
-0.0029
0.011
0.013 *
0.0028
0.0066
-0.0052
0.0062
0.019
0.15 ***
0.018
0.031
0 5 15 25
1 2 3 4 5
1.0 2.0 3.0
5 3 1
0 10 20
0e+00 4e+08
0 40 80
1 2 3 4 5
0 5 15
0e+00 3e+08
0 40 80
20 10 0
5 3 1
3.0 2.0 1.0