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

Anexo 4.1. Exploración de datos: VesselBalancedSample

Oscar Antón

diciembre de 2023

Carga de librerías y datos

```
# Librería                        # Propósito
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

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

```
# Cargar el dataframe VesselBalancedSample (50% barcos con incidentes, 50% barcos sin incidentes)
VesselBalancedSample <- as.data.table(readRDS("../1.DataPreprocess/DataMergedActivity/VesselBalancedSample.r
ds"))
```

Descripción estadística

```
# Descripción de datos balanceados
skim(VesselBalancedSample)
```

Data summary

Name	VesselBalancedSample
Number of rows	109836
Number of columns	12
Key	NULL
________________________	
Column type frequency:	
character	9
numeric	3
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
imo_number	0	1	0	7	90591	9728	0
vessel_name	0	1	0	50	5	62340	0
vessel_class	0	1	5	23	0	16	0
build_year	0	1	4	4	0	131	0
flag_abbr	0	1	0	2	50	152	0
classification_society	0	1	6	58	0	41	0
solas_desc	0	1	9	16	0	3	0
event_type	0	1	4	30	0	27	0
damage_status	0	1	7	35	0	5	0

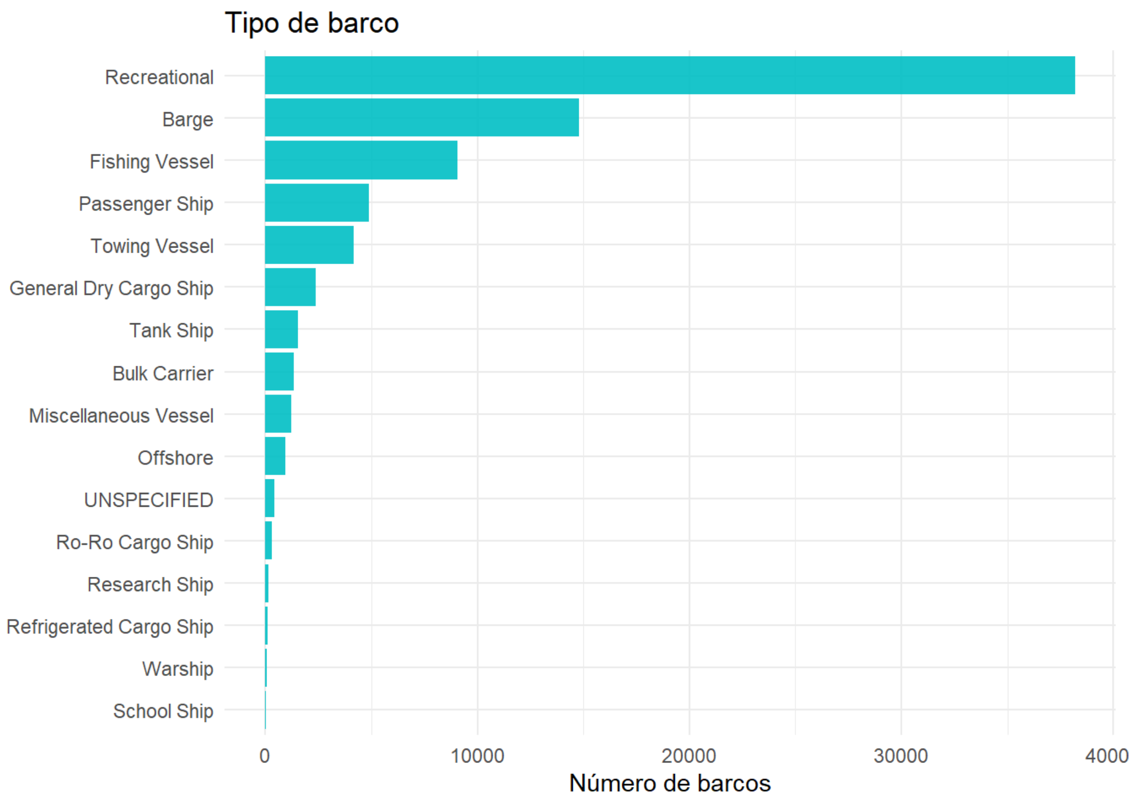
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
vessel_id	0	1	340600.76	299164.28	18.0	119634.2	241684	459737	1348019.0	▇▅▁▁▁
gross_ton	0	1	2946.08	11476.77	1.0	16.0	67	734	234627.0	▇▁▁▁▁
length	0	1	136.59	174.86	6.8	36.3	60	195	1203.8	▇▁▁▁▁

1. Características de los barcos

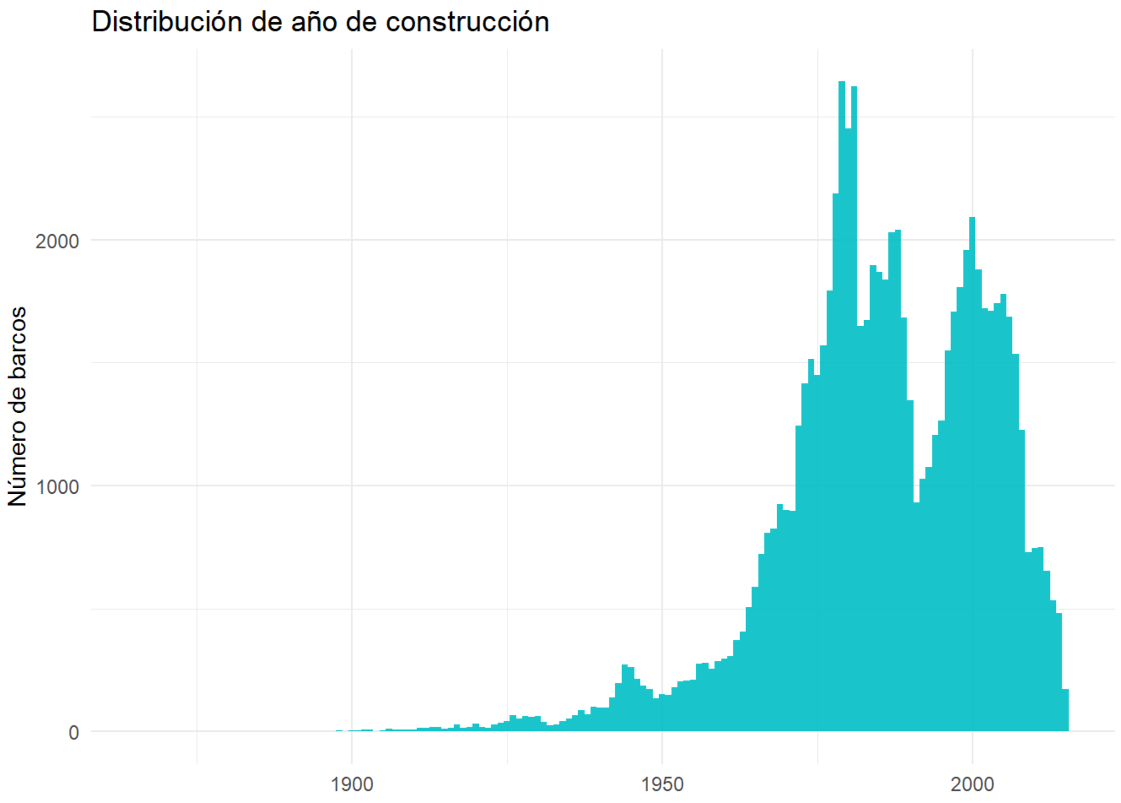
1.1. Tipo de barco (vessel_class)

```r
# Frecuencia por tipo de barco
VesselBalancedSample %>%
  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") +
  theme_minimal() +
  coord_flip()
```



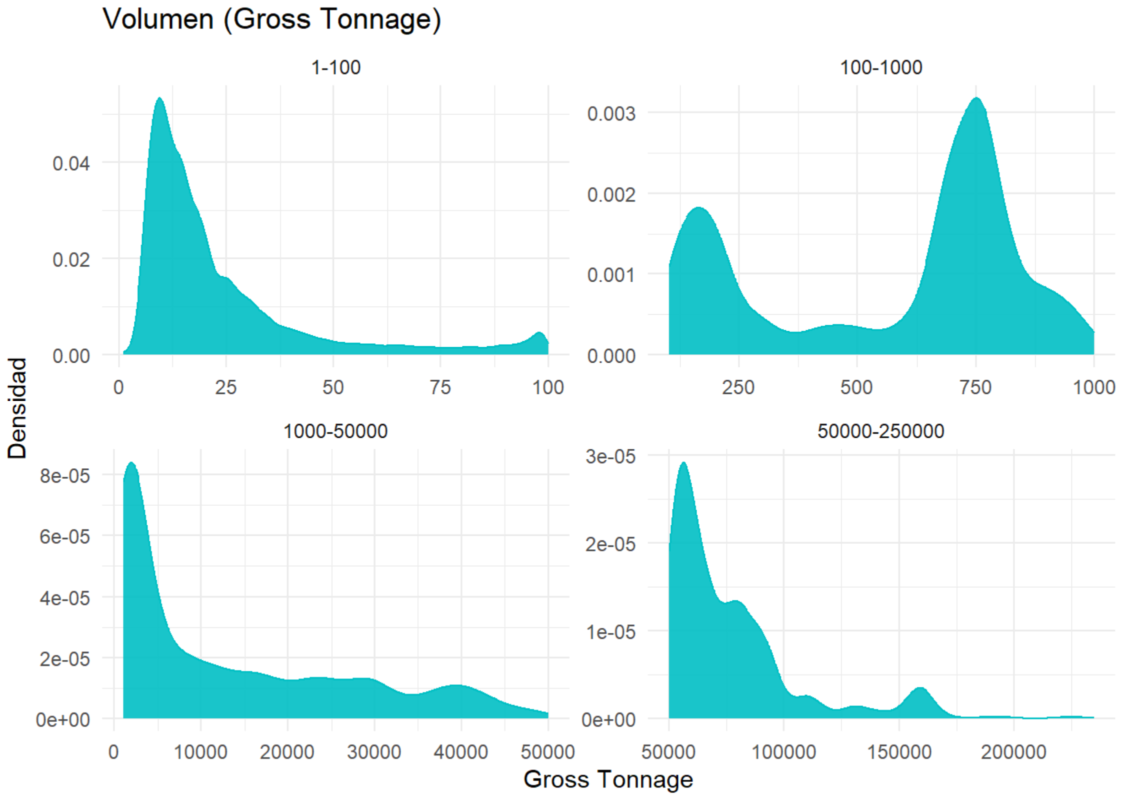
1.2. Año de construcción (build_year)

```
# Frecuencia por año de construcción
VesselBalancedSample %>%
  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") +
  theme_minimal()
```



1.3. Volumen (gross_ton)

```
# Gráficos de densidad por tramos para gross_ton
VesselBalancedSample %>%
  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, 100, 1000, 50000, 250000), labels = c("1-100", "100-1000", "1000-
50000", "50000-250000")), nrow = 2, scales = "free") +
  labs(title = "Volumen (Gross Tonnage)", x = "Gross Tonnage", y = "Densidad") +
  theme_minimal()
```



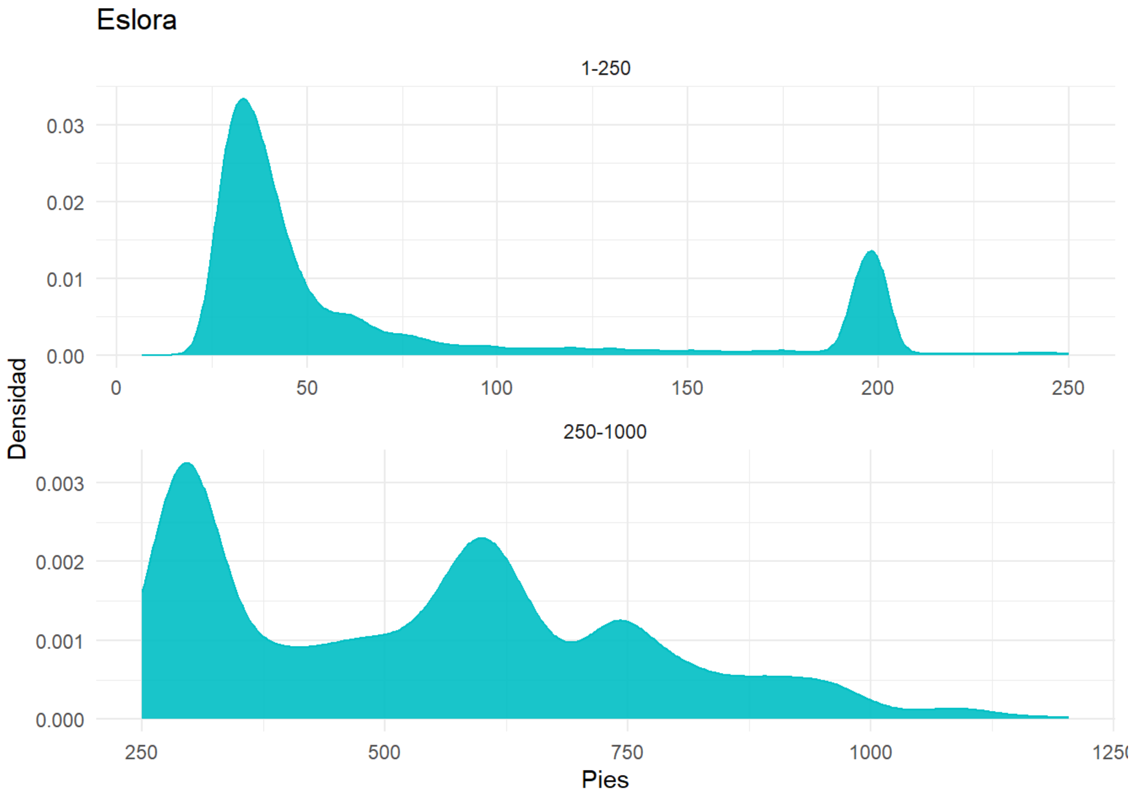
```
# Barcos con mayor Gross Tonnage
VesselBalancedSample %>%
  select(vessel_id, imo_number, vessel_name, build_year, gross_ton, length) %>%
  arrange(desc(gross_ton)) %>%
  unique() %>%
  head(10) %>%
  knitr::kable("html")%>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"), full_width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

vessel_id	imo_number	vessel_name	build_year	gross_ton	length
299897	7376525	KAPETAN GIANNIS	1977	234627	406.6
933483	9383936	OASIS OF THE SEAS	2009	225282	1187.0
933484	9383948	ALLURE OF THE SEAS	2010	225282	1181.0
324613	7370301	KAPETAN PANAGIOTIS	1977	218447	362.3
224539	7376989	CHEVRON SOUTH AMERICA	1976	198951	1200.4
260729	7373298	AURIGA	1976	194992	378.0
228357	7708302	FOLK MOON	1981	188728	1117.0

vessel_id	imo_number	vessel_name	build_year	gross_ton	length
228358	7708314	BERGE PIONEER	1980	188728	1071.7
881546	9266102	YM SKY	2003	179037	172.0
275226	7389534	BERGE INGERID	1977	169752	362.6

1.4. Eslora (length)

```
# Gráficos de densidad por tramos para length
VesselBalancedSample %>%
  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 = "fr
ee") +
  labs(title = "Eslora", x = "Pies", y = "Densidad") +
  theme_minimal()
```



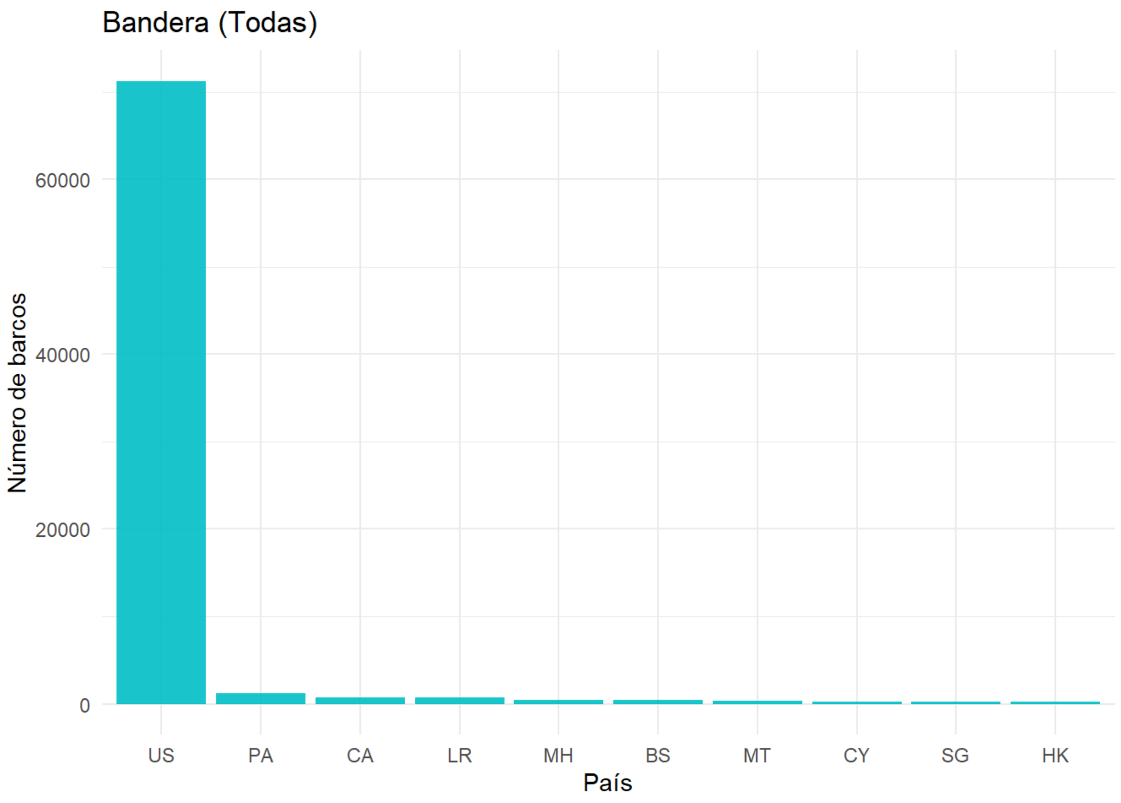
```
# Barcos con mayor eslora
VesselBalancedSample %>%
  select(vessel_id, imo_number, vessel_name, build_year, gross_ton, length) %>%
  arrange(desc(length)) %>%
  unique() %>%
  head(10) %>%
  knitr::kable("html")%>%
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"), full_width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

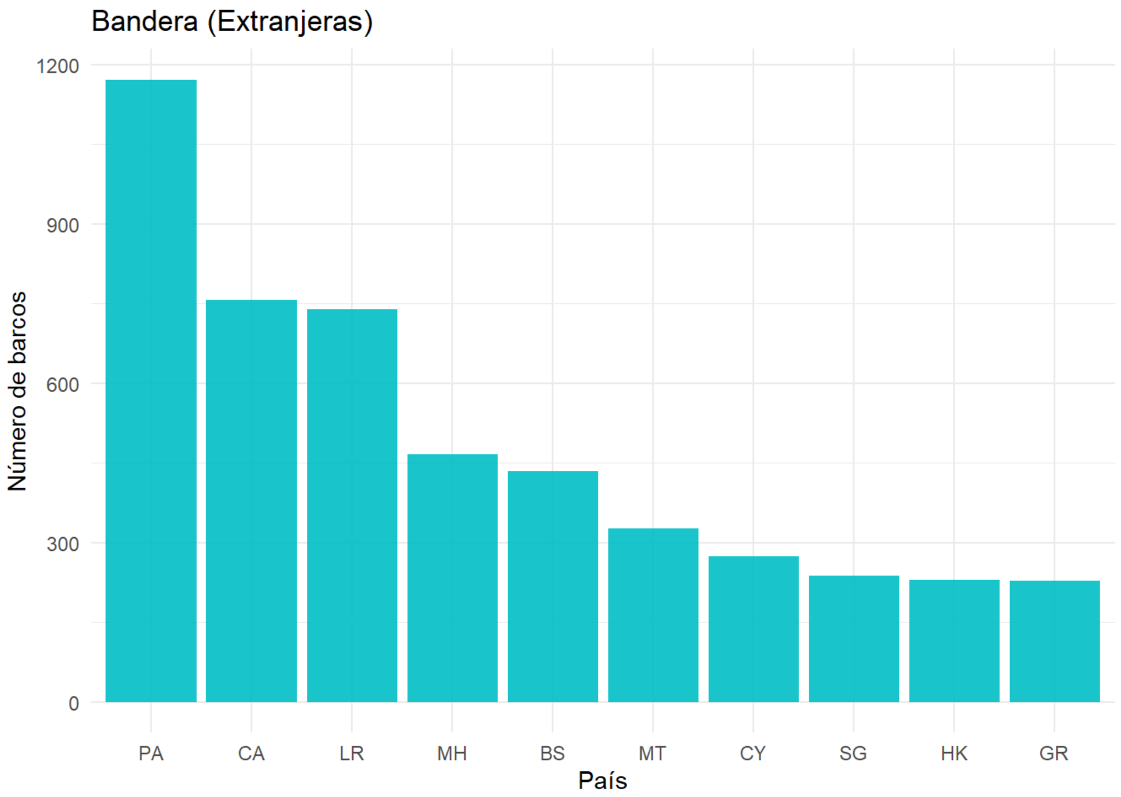
vessel_id	imo_number	vessel_name	build_year	gross_ton	length
1008200	9359040	MARIT MAERSK	2009	98268	1203.7
1028411	9359052	MATHILDE MAERSK	2009	98268	1203.7
999387	9359014	MARCHEN MAERSK	2007	98268	1203.7
1277844	9472127	COSCO FORTUNE	2012	141823	1202.2
1325557	9447902	MSC FILLIPPA	2011	140259	1201.0
224539	7376989	CHEVRON SOUTH AMERICA	1976	198951	1200.4
1171505	9398371	MSC IVANA	2008	131771	1192.9
274926	7359058	KAROLINE	1976	158475	1192.0

1.5. Bandera (flag_abbr)

```
# Gráfico de barras con top10 banderas
VesselBalancedSample %>%
  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 = "Número de barcos") +
  theme_minimal()
```

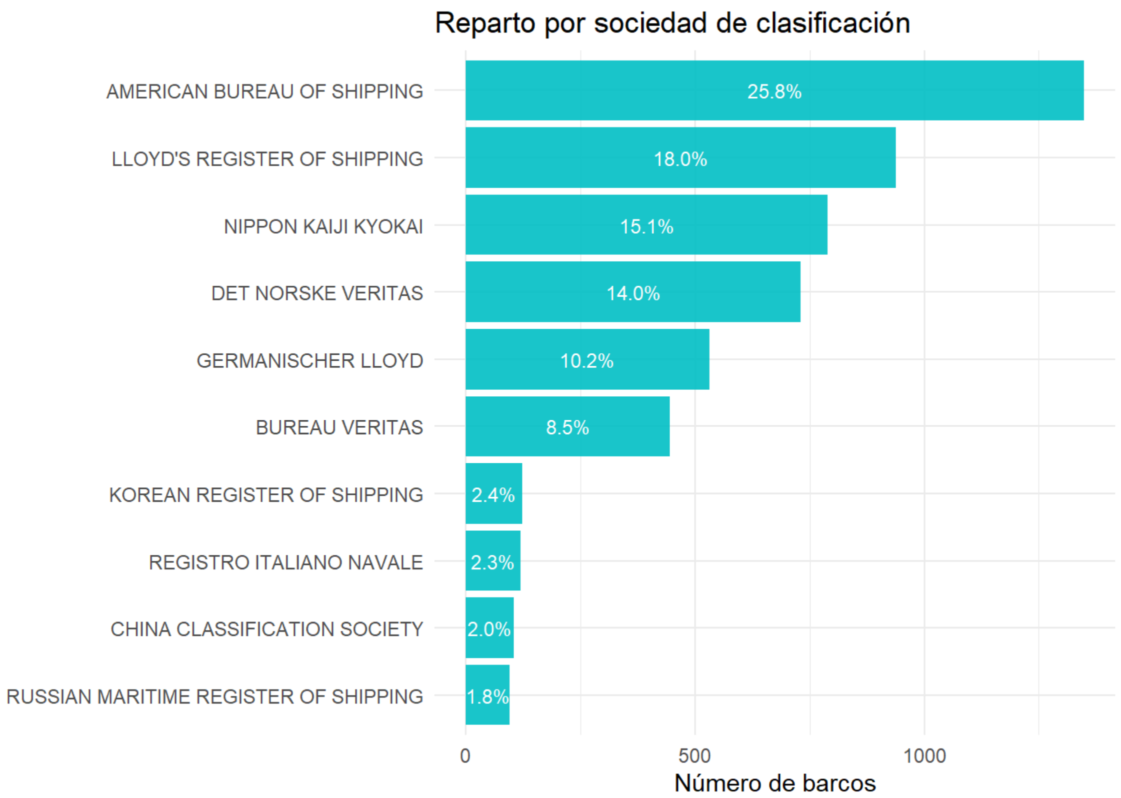


```
# Top 10 sin bandera local (EEUU)
VesselBalancedSample %>%
  filter(flag_abbr != "US") %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  group_by(flag_abbr) %>%
  summarise(frecuencia = n()) %>%
  arrange(desc(frecuencia)) %>%
  head(10) %>%
  ggplot(aes(x = fct_reorder(flag_abbr, frecuencia, desc),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



1.6. Sociedad de clasificación (classification_society)

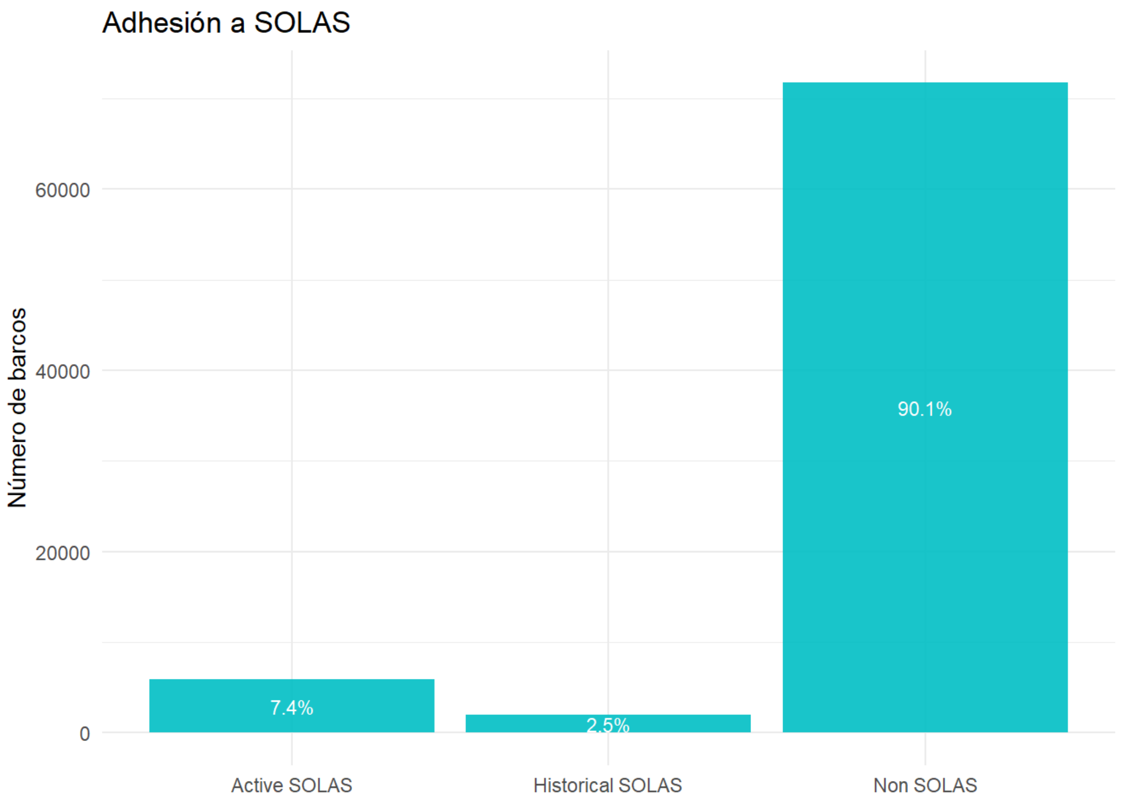
```
# Gráfico de barras horizontales para top10 sociedad de clasificación
VesselBalancedSample %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(classification_society != "UNSPECIFIED") %>%
  group_by(classification_society) %>%
  summarise(frecuencia = n()) %>%
  arrange(desc(frecuencia)) %>%
  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 = "whi
te", size = 3) +
  labs(title = "Reparto por sociedad de clasificación", x = NULL, y = "Número de barcos") +
  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

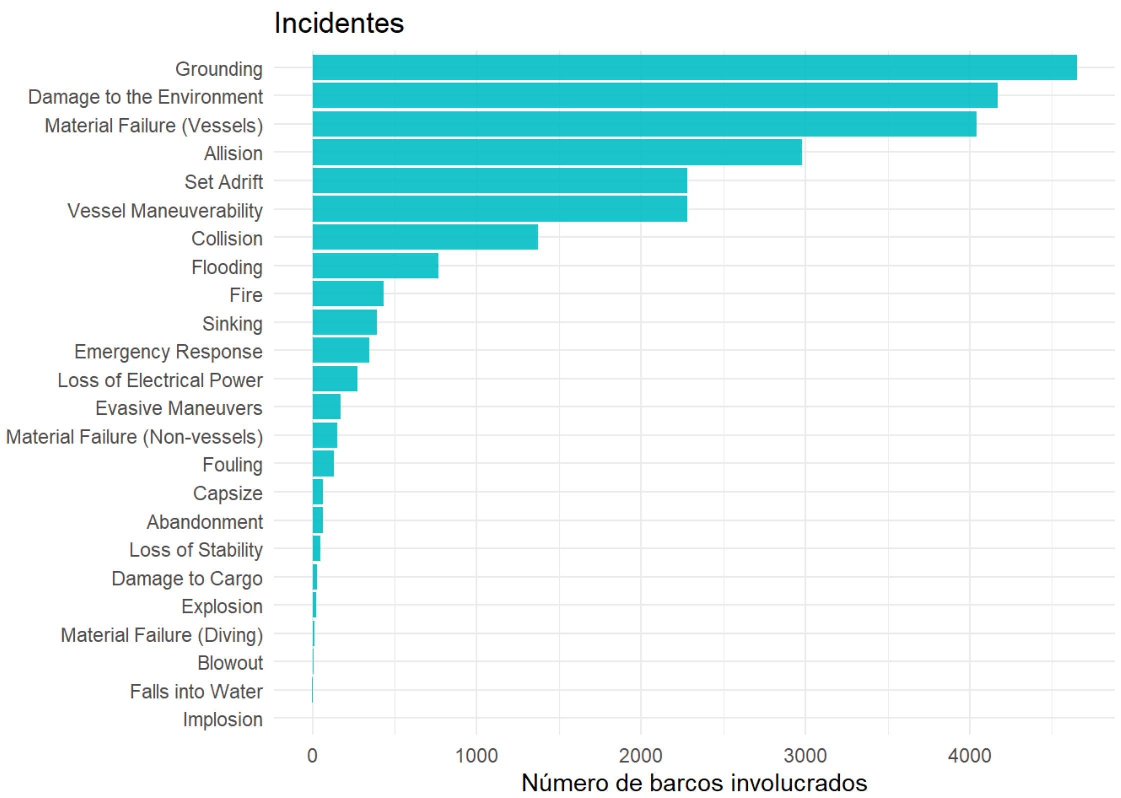
```
# Gráfico de barras para SOLAS
VesselBalancedSample %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  group_by(solas_desc) %>%
  summarise(frecuencia = n()) %>%
  mutate(porcentaje = frecuencia / sum(frecuencia) * 100) %>%
  ggplot(aes(x = solas_desc, y = frecuencia)) +
  geom_bar(stat = "identity",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



2. Incidentes

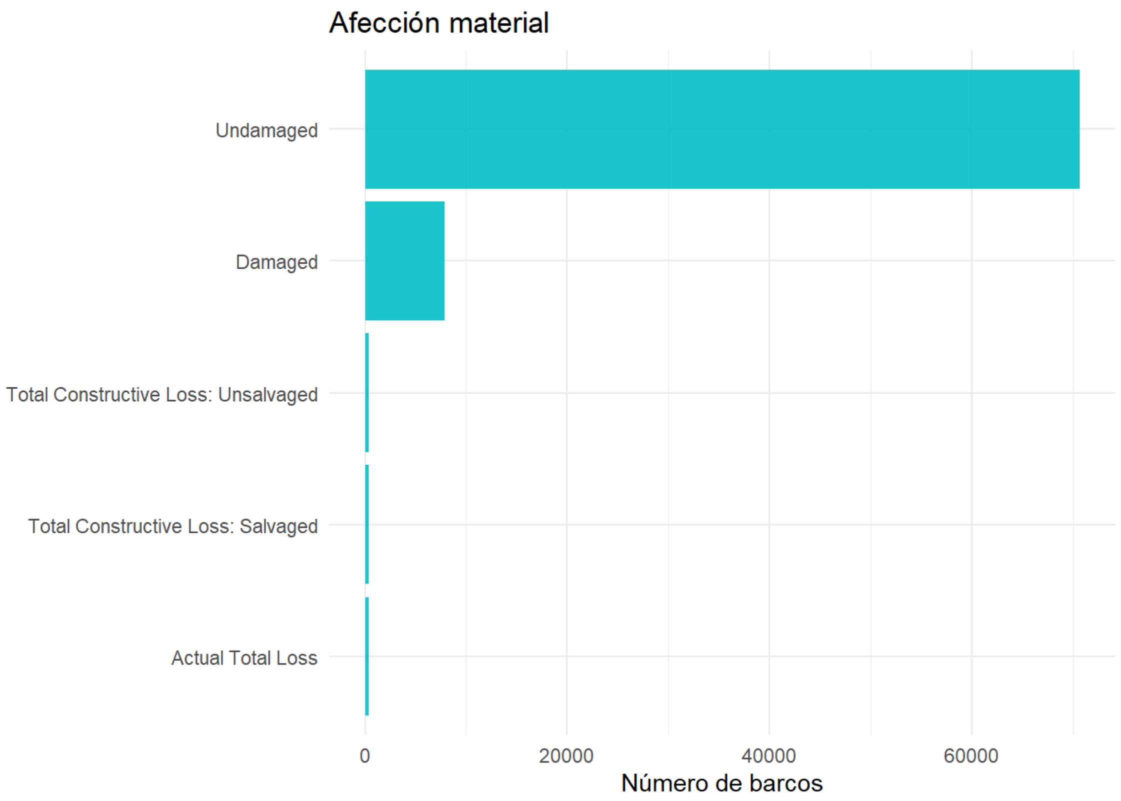
2.1 Tipo de incidente (event_type)

```
# Gráfico de barras horizontales para event_type
VesselBalancedSample %>%
  filter(event_type != "No event") %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  group_by(event_type) %>%
  summarise(frecuencia = n()) %>%
  arrange(desc(frecuencia)) %>%
  ggplot(aes(x = fct_reorder(event_type, frecuencia), y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  labs(title = "Incidentes", x = NULL, y = "Número de barcos involucrados") +
  theme_minimal() +
  coord_flip()
```



2.2. Daños (damage_status)

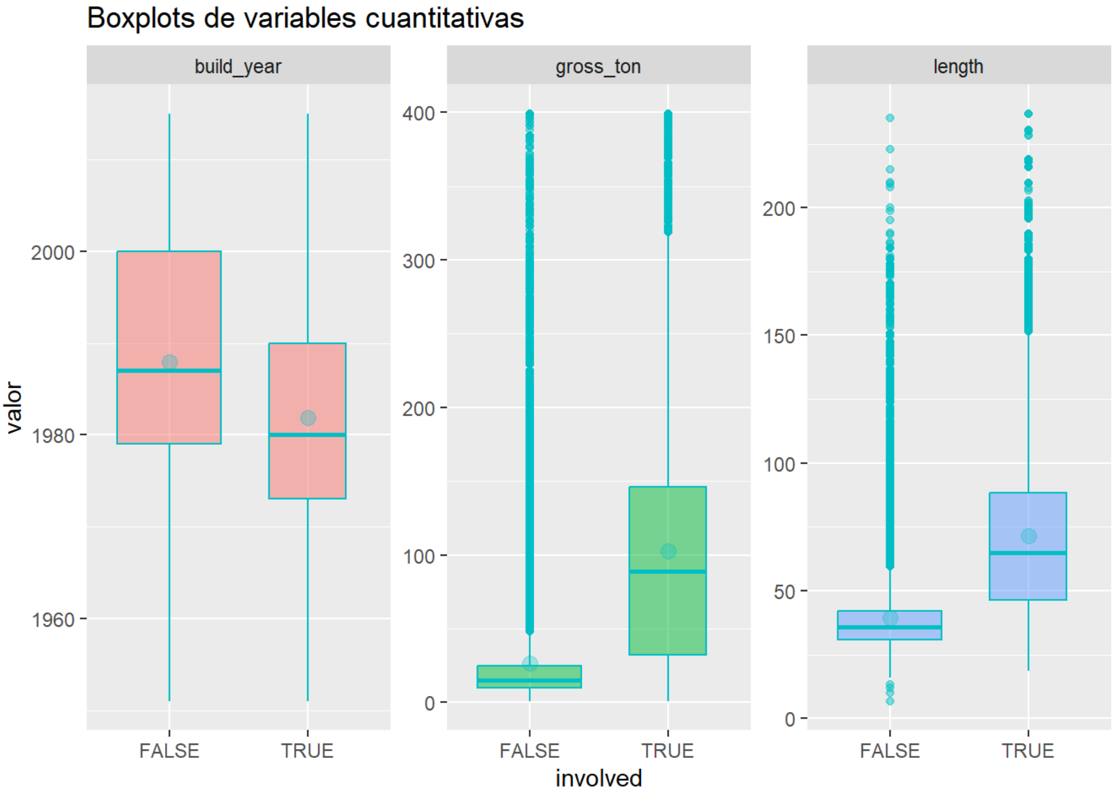
```
# Gráfico de barras horizontales para damage_status
VesselBalancedSample %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  group_by(damage_status) %>%
  summarise(frecuencia = n()) %>%
  ggplot(aes(x = fct_reorder(damage_status, frecuencia), y = frecuencia)) +
  geom_bar(stat = "identity", fill = "#00bfc4", alpha = 0.9) +
  labs(title = "Afección material", x = NULL, y = "Número de barcos") +
  theme_minimal() +
  coord_flip()
```



3. Involucración en accidente / Variables explicativas

3.1. Involucración en accidente / Características del barco

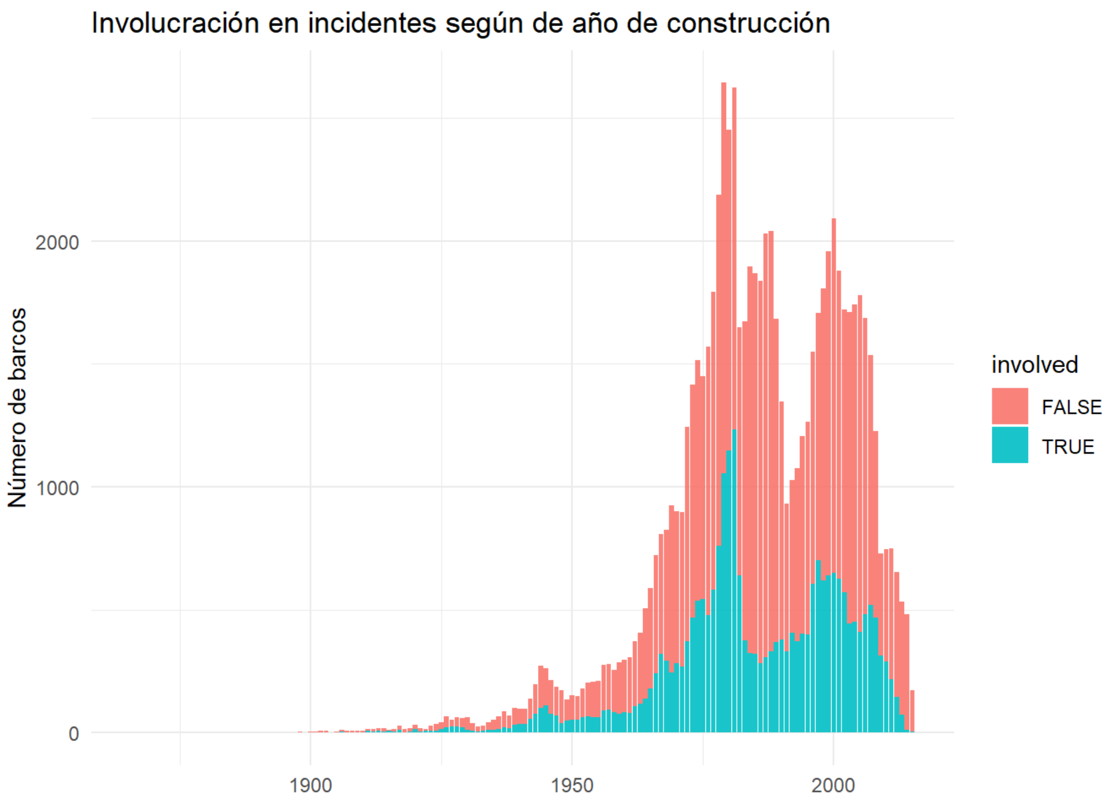
```
# Boxplots de variables no factoriales
VesselBalancedSample %>%
  mutate(involved = as.factor(as.character(event_type != "No event"))) %>%
  mutate(build_year = as.numeric(build_year)) %>%
  filter(gross_ton < 400, length < 250, build_year > 1950) %>%
  select(involved, gross_ton, length, build_year) %>%
  pivot_longer(cols = -involved, names_to = "variable", values_to = "valor") %>%
  ggplot(aes(y = valor, x = involved, 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 cuantitativas")
```



```
# Frecuencia por año de construcción
VesselBalancedSample %>%
  distinct(vessel_id, .keep_all = TRUE) %>%
  filter(build_year >= 1800 & build_year <= 2015) %>%
  mutate(involved = as.factor(as.character(event_type != "No event"))) %>%
  group_by(build_year, involved) %>%
  summarise(frecuencia = n()) %>%
  ggplot(aes(x = as.numeric(build_year), y = frecuencia, fill = involved)) +
  geom_histogram(stat = "identity", alpha = 0.9) +
  labs(title = "Involucración en incidentes según de año de construcción", x = NULL, y = "Número de barcos")
+
  theme_minimal()
```

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

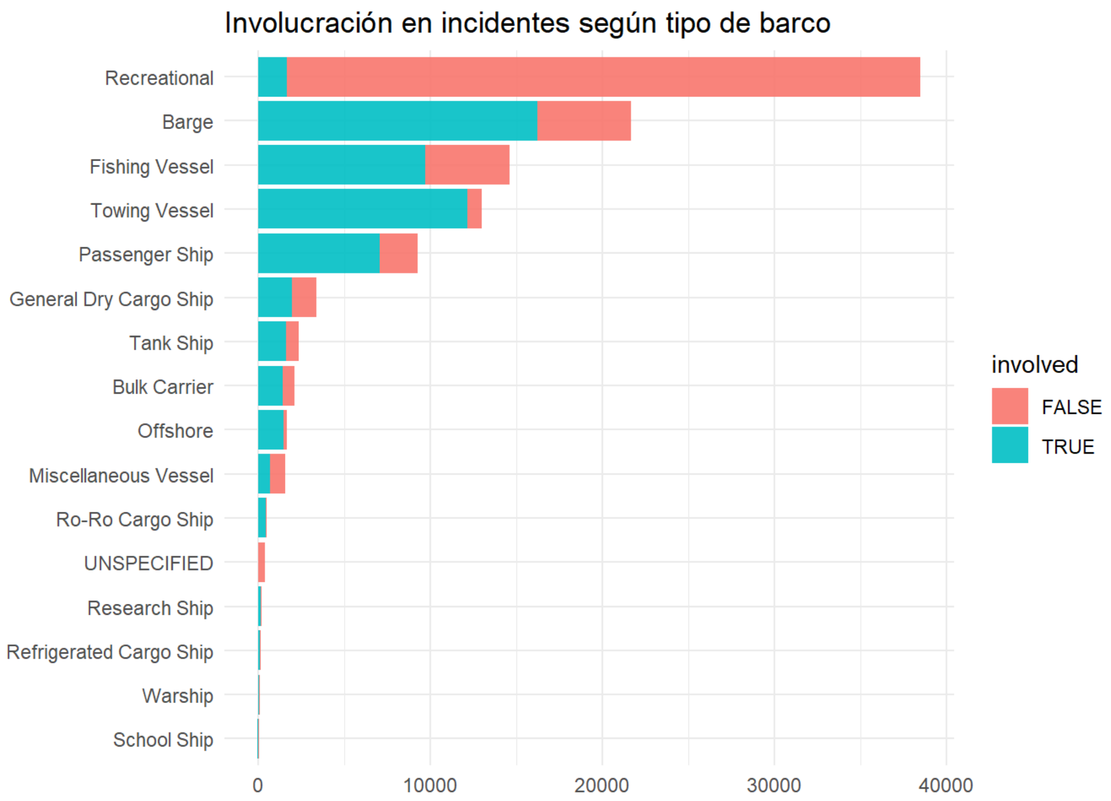
```
## Warning in geom_histogram(stat = "identity", alpha = 0.9): Ignoring unknown
## parameters: `binwidth`, `bins`, and `pad`
```



3.2. Involucración en accidente / Tipo de barco

```
# Gráfico de barras apiladas horizontales
VesselBalancedSample %>%
  mutate(involved = as.factor(as.character(event_type != "No event"))) %>%
  group_by(vessel_class, involved) %>%
  summarise(frecuencia = n()) %>%
  ggplot(aes(x = fct_reorder(vessel_class, frecuencia), y = frecuencia, fill = involved)) +
  geom_bar(stat = "identity", alpha = 0.9) +
  labs(title = "Involucración en incidentes según tipo de barco", x = NULL, y = NULL) +
  theme_minimal() +
  coord_flip()
```

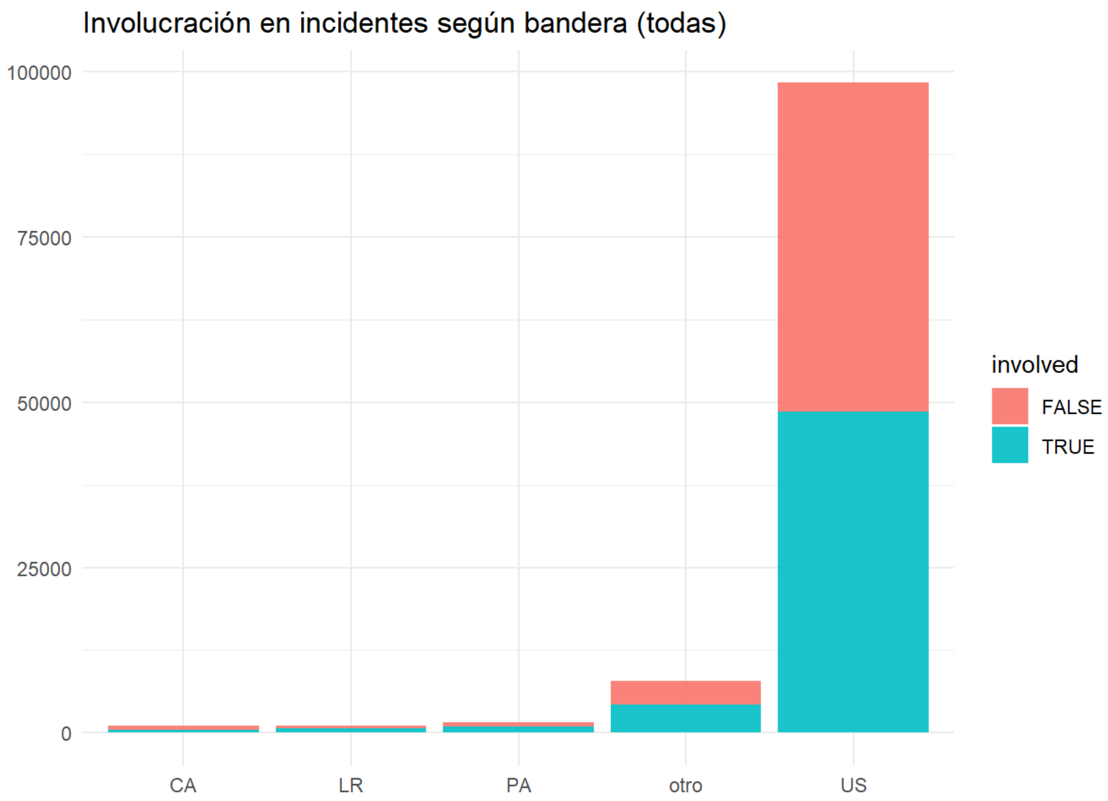
```
## `summarise()` has grouped output by 'vessel_class'. You can override using the
## `.groups` argument.
```



3.3. Involucración en accidente / Bandera

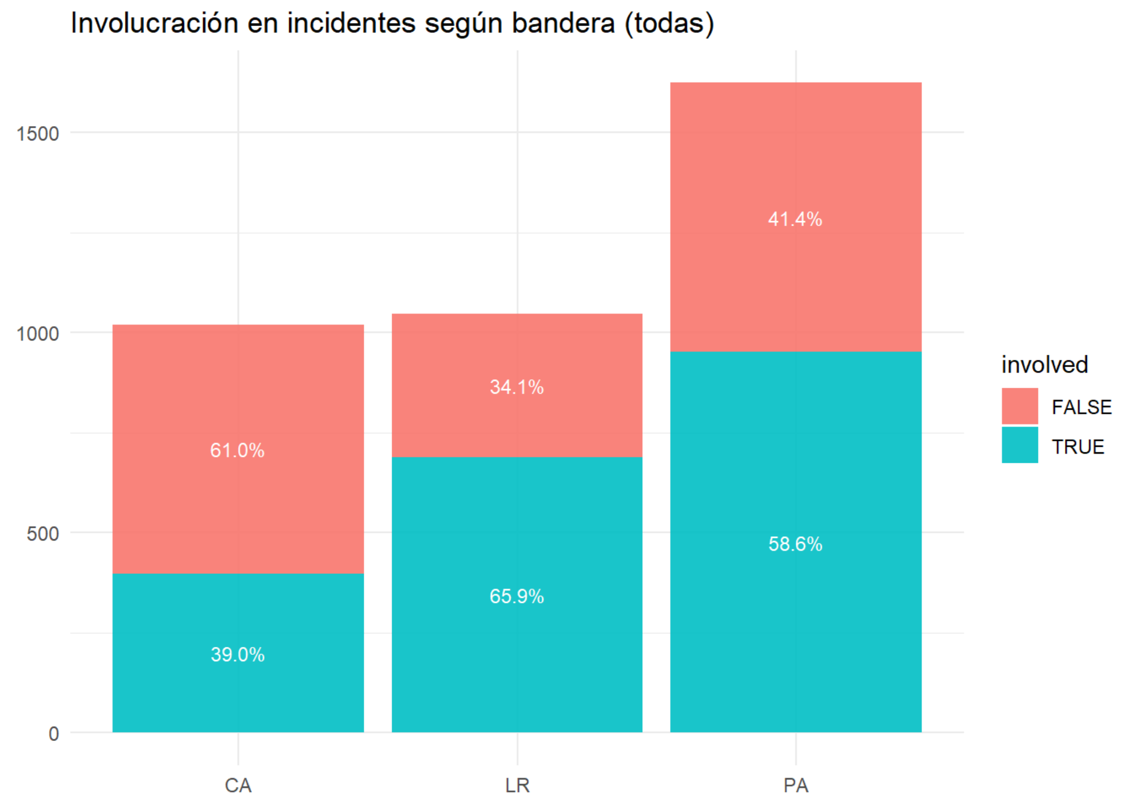
```
lump_factorials <- function(factor_var) {
  fct_lump(factor_var, prop = 0.008, other_level = "otro")
}

# Gráfico de barras apiladas para todas las banderas
VesselBalancedSample %>%
  mutate(involved = as.factor(as.character(event_type != "No event"))) %>%
  mutate(flag_abbr = lump_factorials(flag_abbr)) %>%    # Reducción de la variabilidad
  group_by(flag_abbr, involved) %>%
  summarise(frecuencia = n()) %>%
  ggplot(aes(x = fct_reorder(flag_abbr, frecuencia), y = frecuencia, fill = involved)) +
  geom_bar(stat = "identity", alpha = 0.9) +
  #geom_text(aes(label = frecuencia), color = "white", position = position_stack(vjust = 0.5)) +
  labs(title = "Involucración en incidentes según bandera (todas)", x = NULL, y = NULL) +
  theme_minimal()
```

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



```
# Gráfico de barras apiladas para banderas extranjeras
VesselBalancedSample %>%
  mutate(involved = as.factor(as.character(event_type != "No event"))) %>%
  mutate(flag_abbr = lump_factorials(flag_abbr)) %>%    # Reducción de la variabilidad
  filter(flag_abbr != "US", flag_abbr != "otro" ) %>%
  group_by(flag_abbr, involved) %>%
  summarise(frecuencia = n()) %>%
  mutate(porcentaje = frecuencia / sum(frecuencia) * 100) %>%
  ggplot(aes(x = fct_reorder(flag_abbr, frecuencia), y = frecuencia, fill = involved)) +
  geom_bar(stat = "identity", alpha = 0.9) +
  geom_text(aes(label = sprintf("%.1f%%", porcentaje)), position = position_stack(vjust = 0.5), color = "whi
te", size = 3) +
  labs(title = "Involucración en incidentes según bandera (todas)", x = NULL, y = NULL) +
  theme_minimal()
```

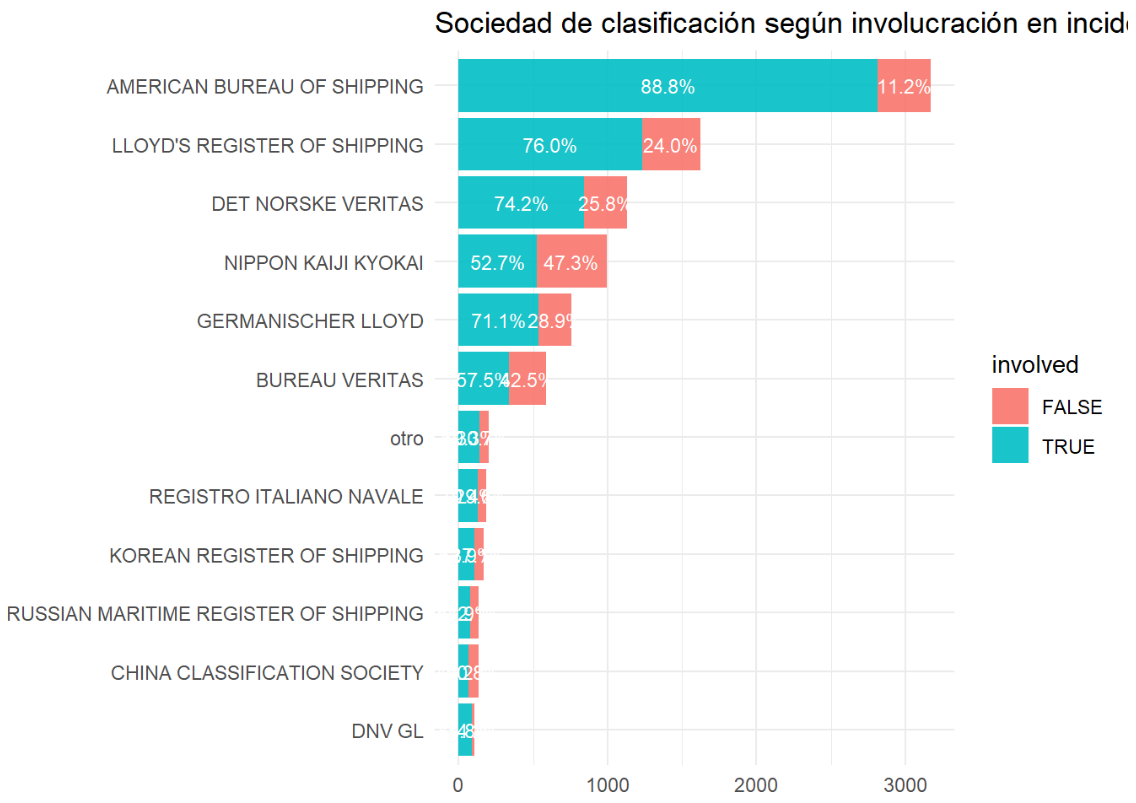
```
## `summarise()` has grouped output by 'flag_abbr'. You can override using the
## `.groups` argument.
```



3.4. Involucración en accidente / Sociedad de clasificación

```
# Gráfico de barras apiladas horizontales
VesselBalancedSample %>%
  filter(classification_society != "UNSPECIFIED") %>%
  mutate(involved = as.factor(as.character(event_type != "No event"))) %>%
  mutate(classification_society = lump_factorials(classification_society)) %>%    # Reducción de la variabil
idad
  group_by(classification_society, involved) %>%
  summarise(frecuencia = n()) %>%
  mutate(porcentaje = frecuencia / sum(frecuencia) * 100) %>%
  ggplot(aes(x = fct_reorder(classification_society, frecuencia), y = frecuencia, fill = involved)) +
  geom_bar(stat = "identity", alpha = 0.9) +
  geom_text(aes(label = sprintf("%.1f%%", porcentaje)), position = position_stack(vjust = 0.5), color = "whi
te", size = 3) +
  labs(title = "Sociedad de clasificación según involucración en incidentes", x = NULL, y = NULL) +
  theme_minimal() +
  coord_flip()
```

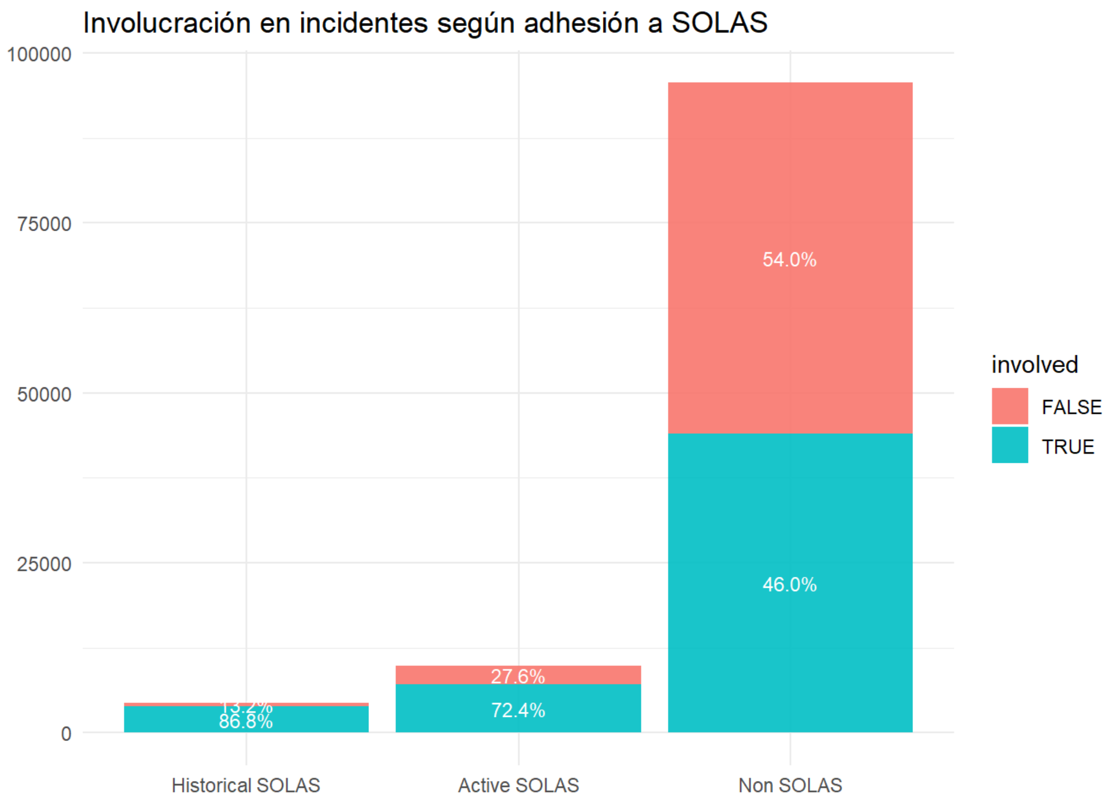
```
## `summarise()` has grouped output by 'classification_society'. You can override
## using the `.groups` argument.
```



3.5. Involucración en accidente / SOLAS

```
# Gráfico de barras apiladas para adhesión a SOLAS
VesselBalancedSample %>%
  mutate(involved = as.factor(as.character(event_type != "No event"))) %>%
  group_by(solas_desc, involved) %>%
  summarise(frecuencia = n()) %>%
  mutate(porcentaje = frecuencia / sum(frecuencia) * 100) %>%
  ggplot(aes(x = fct_reorder(solas_desc, frecuencia), y = frecuencia, fill = involved)) +
  geom_bar(stat = "identity", alpha = 0.9) +
  geom_text(aes(label = sprintf("%.1f%%", porcentaje)), position = position_stack(vjust = 0.5), color = "whi
te", size = 3) +
  labs(title = "Involucración en incidentes según adhesión a SOLAS", x = NULL, y = NULL) +
  theme_minimal()
```

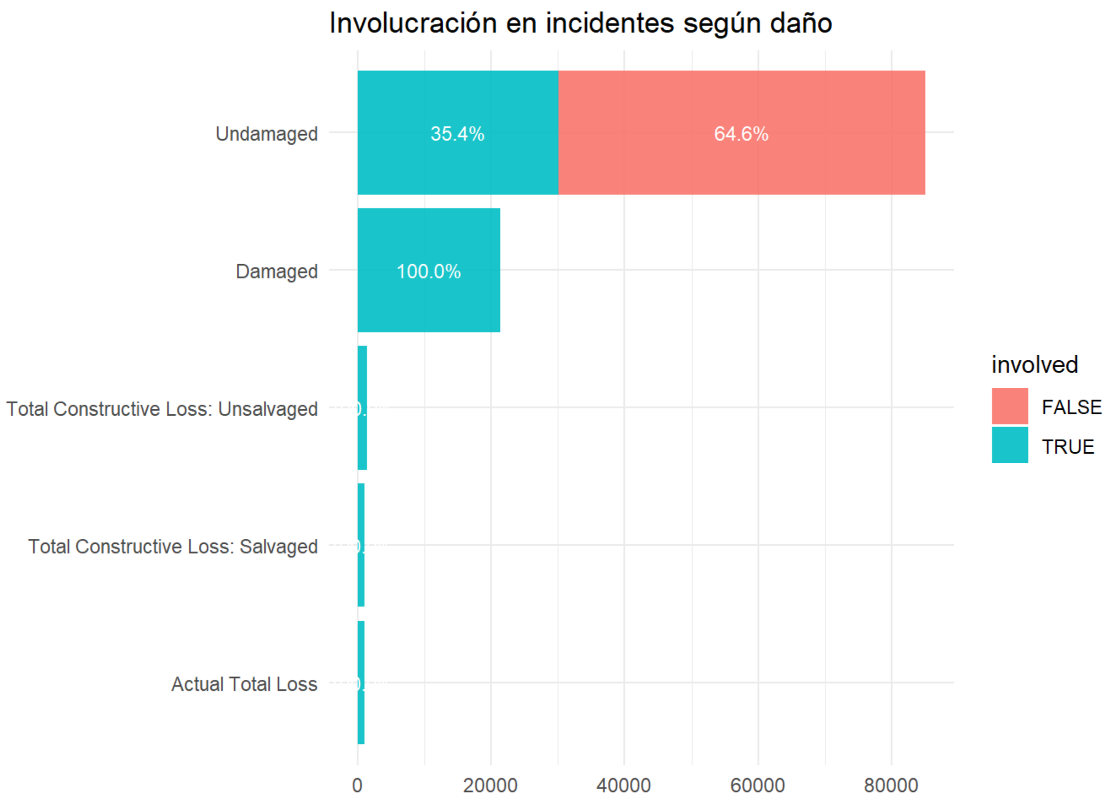
```
## `summarise()` has grouped output by 'solas_desc'. You can override using the
## `.groups` argument.
```



3.6. Involucración en accidente / Daño

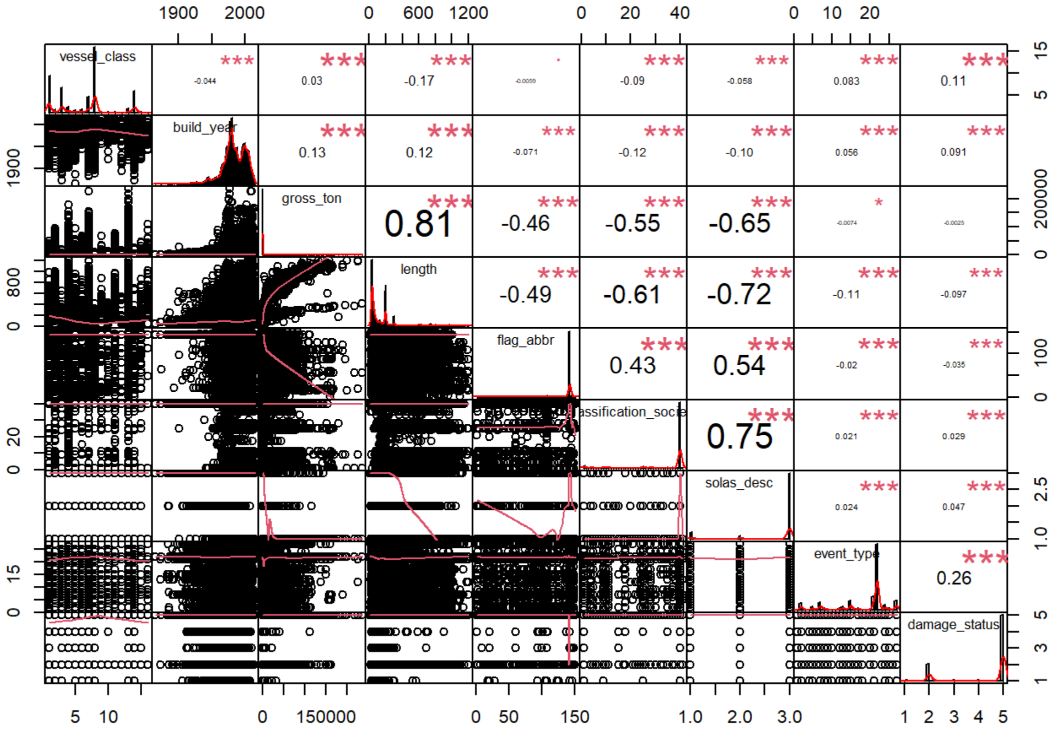
```
# Gráfico de barras apiladas para adhesión a SOLAS
VesselBalancedSample %>%
  mutate(involved = as.factor(as.character(event_type != "No event"))) %>%
  group_by(damage_status, involved) %>%
  summarise(frecuencia = n()) %>%
  mutate(porcentaje = frecuencia / sum(frecuencia) * 100) %>%
  ggplot(aes(x = fct_reorder(damage_status, frecuencia), y = frecuencia, fill = involved)) +
  geom_bar(stat = "identity", alpha = 0.9) +
  geom_text(aes(label = sprintf("%.1f%%", porcentaje)), position = position_stack(vjust = 0.5), color = "whi
te", size = 3) +
  labs(title = "Involucración en incidentes según daño", x = NULL, y = NULL) +
  theme_minimal() +
  coord_flip()
```

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



4. Correlaciones

```
VesselBalancedSample %>%
  select(-vessel_id, -imo_number, -vessel_name) %>%
  mutate_at(vars(vessel_class, flag_abbr, classification_society, solas_desc, event_type, damage_status), factor) %>%
  mutate_all(~as.integer(.)) %>%
  chart.Correlation(histogram = T, pch = 19)
```



Hay correlaciones destacadas entre:

· *length* y *gross_ton* (mayor eslora, implica mayor volumen)

· *classification_society* y *solas_desc* Normalmente, lo barcos con mayor volumen están obligados a atenerse a ambas cuestiones