

Tipología y Ciclo de Vida de los Datos: PRA2 - Limpieza y validación de los datos

Autores: Joel Bustos - Iván Ruiz

Junio 2020

Tabla de Contenidos

Introducción.....	2
Presentación.....	2
1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende resolver?	3
2. Integración y selección de los datos de interés a analizar.....	5
2.1 Resumen de tratamientos previos.....	10
2.2 Carga del nuevo archivo tras el procesado de datos.	17
3. Limpieza de los datos.	21
3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?	21
3.2 Identificación y tratamiento de valores extremos.....	23
4. Análisis de los datos.....	36
4.1 Planificación de los análisis a aplicar.....	36
4.2 Análisis particulares, y comprobaciones de normalidad y distribución de la varianza en las variables bajo estudio.....	38
4.2.1 Análisis temporal de los ciberataques.....	38
4.2.2 Análisis territorial de los ciberataques.	45
4.2.3 Análisis de la tipología de ataques.....	49
5. Representación de los resultados a partir de tablas y gráficas.....	55
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?	64
Contribuciones.....	68
Bibliografía	68

Introducción

Presentación.

A lo largo de esta segunda práctica, vamos a tratar de profundizar en el análisis planteado durante la primera parte de la asignatura. Así, hacemos referencia tanto al repositorio que creamos, como a la documentación generada a través del siguiente enlace: <https://github.com/iruiper/Cyberattacks-History>

A pesar de las dificultades que plantea el set de datos sobre el que trabajaremos, nos gustaría tratar de cerrar las inquietudes y motivaciones que nos llevaron, en primera instancia, a trabajar sobre la problemática de los ataques de ciberseguridad. En este sentido, nos gustaría iniciar esta segunda exposición rescatando la motivación planteada en el proyecto de obtención de datos:

“Los equipos de seguridad han necesitado incorporar, cada vez más, perfiles técnicos en el área de la ciberseguridad. Estos equipos técnicos, normalmente con un conocimiento muy específico, en ocasiones no disponen de demasiadas herramientas que les permitan ser proactivos y anticiparse a las nuevas tendencias y técnicas de ciberataque. De esta forma, acaban adaptando un comportamiento reactivo, realizando tareas de mantenimiento y de respuesta ante incidentes.”

“Nos planteábamos, como contexto para la presente práctica, recopilar datos históricos de ciberataques con el objetivo de crear un modelo predictivo que sirviese de soporte al equipo de seguridad de una empresa. Idealmente, estudiando lo que está ocurriendo en relación a delitos cibernéticos, los equipos internos de las distintas entidades, podrían tratar de prepararse mejor contra aquellos riesgos a los que los modelos estadísticos les pudieran sugerir que se encuentran más expuestos.”

Bajo esta situación, vamos a tratar de plantear un problema concreto que podría analizarse utilizando los datos del sitio web <https://www.hackmageddon.com/>.

Cabe remarcar que, para abordar este segundo proyecto, será necesario aplicar tareas de procesamiento de datos, ya que la calidad de la información extraída durante la primera práctica, no es la adecuada. De esta forma, se pretenderán resolver problemas de calidad, tales como la falta de integración, la existencia de datos incompletos e inconsistentes, o incluso la carencia de variables relevantes para el análisis.

A pesar de que la propuesta de práctica se centra en tareas de limpieza y acondicionado mediante R, también haremos uso del lenguaje de programación Python, con el objetivo de ampliar las herramientas destinadas a mejorar la calidad de datos, y así, potenciar su posterior análisis.

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende resolver?

Tal y como se expone en la introducción, disponemos de un set de datos con información de ataques cibernéticos producidos desde el año 2017 hasta la actualidad.

Con esta información, cabe preguntarse hasta qué punto podemos anticiparnos a los problemas que sobrevendrán a una entidad, o cómo estos datos pueden ayudar en asuntos financieros y presupuestarios, para responder a preguntas como:

- ¿Hasta qué punto mi compañía se encuentra más expuesta a ciberataques, en función del sector en el que se encuentra?
- ¿Existe algún periodo del año en que debería ampliar el presupuesto y los medios necesarios, al estar más expuesto a recibir ciberataques?
- ¿A qué tipología de ciberataque mi compañía está más expuesta? ¿Debería por ello, analizar la documentación cualitativa, así como los detalles técnicos referentes a este tipo de ataque?
- ¿Hasta qué punto el encontrarme en un mundo globalizado, estoy expuesto a recibir ataques internacionales?
- ¿Existen atacantes bien conocidos, con pautas concretas, a los que me encuentre particularmente expuesto?

Todas las preguntas anteriores, podrían resolverse a partir del análisis de la información contenida en el set de datos obtenido durante la primera práctica, ya que:

- Disponemos de datos sobre las entidades o sectores afectados por los ciberataques que aparecen documentados en el dataset.
- Disponemos de la fecha en la que se han producido los distintos ataques.
- Sabemos el alcance del impacto de los ataques, dado que existe información sobre si el ataque afectó a un país en concreto, o a varios de ellos.
- Tenemos información acerca de la autoría de los ataques.
- Conocemos la tipología de ataque de los distintos incidentes reportados.

En concreto, para que los resultados y los contrastes que llevemos a cabo sean lo más concretos y realistas posibles, vamos a centrarnos en el caso de que **formemos parte del equipo de seguridad de un organismo público**, por lo que nuestros análisis cuantitativos y cualitativos, tratarán de poner en relieve las diferencias entre nuestro sector y los demás. Esta distinción, también puede tener como derivada interesante, averiguar el nivel de gasto e inversión acometido por entidades de otros sectores en

materia de ciberseguridad, y a partir de nuestra evaluación del riesgo específico, estudiar si puede ser necesaria la aplicación de nuevas partidas presupuestarias para la defensa contra estas amenazas.

Como hemos visto en los materiales didácticos de Subirats, Pérez y Calvo [1], existen muchos retos diferentes a la hora de integrar y asegurar la calidad de los datos, con el objetivo de dar respuesta a las inquietudes de cualquier analítica de datos.

Adicionalmente, según se desprende de dicho material, así como se muestra en los ejemplos basados en sets de datos estructurados, gran parte de los problemas de calidad se originan en variables cuantitativas. Estas, posteriormente, serán objeto de análisis con el fin de obtener algún tipo de conocimiento.

En nuestro caso, la principal dificultad presente en el set de datos seleccionado, es precisamente la escasez de variables numéricas. En consecuencia, gran parte del reto al que nos enfrentamos, y gran parte de los problemas de procesamiento de datos que vamos a desarrollar a lo largo de la práctica, irán dirigidos a “crear” datos numéricos, así como estadísticos que permitan resolver las cuestiones planteadas.

2. Integración y selección de los datos de interés a analizar.

El primer reto que deberemos resolver en esta etapa de procesamiento de datos, consistirá en integrar el conjunto de datos recopilados durante la primera práctica. En concreto, es de vital importancia destacar la presencia de heterogeneidad entre las distintas fuentes de información ya que, aunque todas procedan del mismo sitio web, presentan estructuras de datos distintas.

A modo de recordatorio, a continuación, se exponen brevemente los distintos conjuntos de datos recopilados del sitio web <https://www.hackmageddon.com/>.

- **Timeline:** Contiene ciberataques producidos desde el 2019 hasta la actualidad. Esta base de datos, está formada por informes quincenales, que presentan una estructura de datos específica.
- **Master Table 2018:** Ciberataques producidos durante el año 2018 contenidos en una única tabla.
- **Master Table 2017:** Ciberataques producidos durante el año 2017. Al igual que en el caso anterior, se encuentran agregados en forma tabular, pero con una estructura de campos distinta.

En consecuencia, el primer problema con el que nos encontramos, es la existencia de varios archivos csv con una estructura de campos específica y con información de distintos periodos temporales.

El primer paso pues, consistirá en unir todos los archivos en bruto, con el objetivo de crear un único conjunto de datos, que servirá para identificar las tareas de procesamiento necesarias para cada campo, así como la tipología de datos presente en estos.

Cabe remarcar que este conjunto de datos, no pretende ser el archivo definitivo con la información procesada, a partir del cual, se realizarán tareas de minería de datos, sino que es un primer borrador, sobre el que se analizarán las distintas estrategias de procesamiento de datos que se deberían de aplicar, para obtener el conocimiento necesario, que permita resolver los planteamientos propuestos. Este archivo, “*DatosAtaques_2017_2020_RAW.csv*”, que corresponde más bien a un archivo en bruto ya que lo utilizaremos para un análisis exploratorio muy inicial, lo hemos ubicado en la carpeta *data/00_raw* del repositorio de Github.

```
# Almacenamos el set de datos bruto en el frame "attacks_Raw" para un análisis preliminar de algunos de los campos que utilizaremos
attacks_Raw <- read.csv2(file='DatosAtaques_2017_2020_RAW.csv', stringsAsFactors = TRUE)
```

Mostramos la estructura del archivo recién cargado

```
str(attacks_Raw)
```

```
## 'data.frame':    4468 obs. of  11 variables:
## $ ID             : int  1 2 3 4 5 6 7 8 9 10 ...
## $ Date           : Factor w/ 1054 levels "01/01/2017","01/01/2018",...: 1
1 1 1 71 104 104 104 104 138 ...
## $ Author         : Factor w/ 564 levels "", "LulzSecITA",...: 17 7 166 7
38 7 289 7 7 186 ...
## $ Target         : Factor w/ 3518 levels "", "City of Del Rio",...: 2547
2954 1070 2743 1385 1417 1231 969 2080 2528 ...
## $ Description    : Factor w/ 4453 levels "", "Malaysia's Computer Emerge
ncy Response Team (MyCERT) reveal the details of a campaign carried out b
y APT40, ta"|__truncated__,...: 3591 3443 1072 4046 92 1412 1522 1040 389
5 1885 ...
## $ Attack         : Factor w/ 327 levels "\"view as\" vulnerability",...:
3 280 220 177 91 177 97 177 9 263 ...
## $ Target.Class   : Factor w/ 25 levels "C Manufacturing",...: 13 2 13 16
13 13 7 16 15 23 ...
## $ Attack.Class   : Factor w/ 13 levels "CC", ">1", "CC",...: 4 6 3 3 10 4
3 3 3 4 ...
## $ Country        : Factor w/ 158 levels "", ">1", "AE", "AF",...: 47 129 139
139 47 59 16 139 139 2 ...
## $ Link           : Factor w/ 1144 levels "", "http://abcnews.go.com/Polit
ics/fbi-probing-attempted-hack-trump-organization-officials/story?id=4765
2150",...: 268 174 68 714 1042 44 1032 724 815 69 ...
## $ Tags           : Factor w/ 927 levels "", "#OpIsrael, #OpUSA, Anonymous
",...: 352 831 213 754 381 556 456 261 582 245 ...
```

A partir de esta carga inicial, observamos que existen campos con mucha información cualitativa, que resultarán irrelevantes para resolver los problemas planteados. En consecuencia, una primera estrategia consistirá en realizar tareas de **reducción de dimensionalidad**, descartando aquellos atributos que no vamos a necesitar:

- **ID:** Identificador único dentro de cada informe.
- **Target:** Nombre de la entidad atacada.
- **Description:** Explicación detallada de cada incidente.
- **Attack:** De manera análoga a *Description*, es un campo con información detallada sobre de ataque realizado.
- **Link:** Enlace URL a la noticia del incidente reportado.
- **Tags:** Contiene hashtags o etiquetas con palabras clave contenidas en la descripción del ciberataque. Únicamente presente en *Master Table 2017*.

Por otra parte, el campo fecha (*Date*) tiene un nivel de granularidad excesivo, ya que resultará muy difícil encontrar varios ataques producidos en un mismo día. Por ello, el nivel de granularidad que definiremos será el número de ataques reportados en un mes concreto. Para realizar este cambio de granularidad, será necesario agregar la información contenida en un mismo mes, creando así, dos variables que contendrán los campos *Año – mes*.

En relación al tipo de ataque, *Attack Class*, al ser una variable categórica, se analizarán los valores contenidos en ella.

Analizaremos la calidad de los datos de la variable Tipo de Ataque

```
tipoAtaque <- attacks_Raw$Attack.Class
table(tipoAtaque)
```

```
## tipoAtaque
##          CC          >1          CC          CE
##          1          1         1469         232
##          CW          CW?    Cyber Crime  Cyber Espionage
##          50          1         1094         172
## Cyber Warfare          H    Hacktivism          N/A
##          33          58          38          5
## Not Found
##        1314
```

Sobre la tabla anterior, observamos que:

- Será necesario homogenizar valores. Existen distintas representaciones/etiquetas para un mismo valor, como, por ejemplo: “*Cyber Crime*” - “*CC*” o “*Cyber Warfare*” - “*CW*”.
- Se deberán de tratar los valores ausentes representados a través de las etiquetas “*N/A*” y “*Not Found*”.
- Existen etiquetas que contienen el carácter “?”.

Para el tipo de entidad, *Target Class*, se realizará el mismo análisis que en *Attack Class*. De esta forma, el rango de valores contenidos en esta variable es:

Analizaremos la calidad de los datos de la variable Entidad Atacada

```
tipoEntidad <- attacks_Raw$Target.Class
table(tipoEntidad)
```

```
##
##
##
##          D Electricity gas steam and air conditioning supply
##
## E Water supply, sewerage waste management, and remediation activities
##
```

##	G Wholesale and retail trade	
##		86
##	H Transportation and storage	
##		38
##	I Accommodation and food service activities	
##		71
##	J Information and communication	
##		201
##	K Financial and insurance activities	
##		193
##	L Real estate activities	
##		4
##	M Professional scientific and technical activities	
##		68
##	N Administrative and support service activities	
##		33
##		Not Found
##		1314
##	O Public administration and defence, compulsory social security	
##		263
##	O Public administration, defence, compulsory social security	
##		167
##		P Education
##		201
##	Q Human health and social work activities	
##		281
##	R Arts entertainment and recreation	
##		119
##	S Other service activities	
##		57
##	U Activities of extraterritorial organizations and bodies	
##		25
##		V Fintech
##		88
##		X Individual
##		672
##	Y Multiple Industries	
##		352
##	Y Multiple targets	
##		78
##	Y Multiple Targets	
##		9
##		Z Unknown
##		13

Sobre la tabla anterior, observamos que:

- Cada valor está formado por un Identificador único (Primera letra), seguido de una breve descripción. En este caso, utilizaremos el identificador único para encontrar inconsistencias en los datos y realizar tareas de homogenización.

- Existen valores perdidos, representados mediante las etiquetas “*Not Found*” y “*Z-Unknown*”.

En relación al autor, campo *Author*, dado el alto número de posibles valores (564 niveles), dicotomizaremos esta variable con el objetivo de distinguir, si los ataques son producidos por delincuentes u organizaciones bien identificadas, o por autores anónimos o desconocidos. De esta forma, se creará una clase con los valores ‘*Desconocido*’ o ‘*Conocido*’ que podría ayudarnos a identificar si estamos más expuestos a ataques de redes conocidas y, por lo tanto, dedicar recursos a analizar sus sistemas y tipos de ataques; o si necesitamos mecanismos de defensa mucho más heterogéneos debido al alto número de atacantes anónimos.

Por último, en relación al país afectado, *Country*, observamos que el número de casos documentados es muy alto (158 niveles). En particular, con el objetivo de realizar un proceso de reducción de datos basado en la cantidad, se discretizarán los valores de estas variables en función del *Continente* en el que se encuentre cada país.

Por otra parte, existen tipologías de ataques de rango internacional, cuya etiqueta es ‘>1’. Del mismo modo, el set de datos contiene observaciones con distintos países separados por saltos de línea. A estos casos, también se les asignará el valor ‘*International*’.

```
>>> attacks_Raw.Country.unique()
array(['GB', 'TR', 'US', 'IN', 'BR', '>1', 'IL', 'CA', 'JP', 'VE', 'TH',
      'IT', 'CZ', 'RU', 'KR', 'FI', 'NL', nan, 'SA', 'CN', 'SE', 'HK',
      'INT', 'AU', 'AT', 'RU\n AT', 'PL', 'US\n JP', 'RU\n BY', 'NO',
```

Finalmente, este campo también contiene la presencia de valores perdidos que deberán ser tratados.

```
>>> attacks_Raw[attacks_Raw.Country.isnull()].__len__()
116
```

2.1 Resumen de tratamientos previos

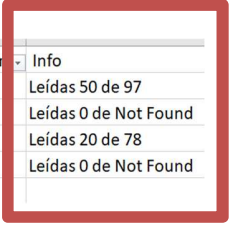
Como indicábamos en la sección introductoria, en el proyecto de limpieza de datos que hemos planteado, puede que uno de los principales retos con nuestro set de datos, sea precisamente todo lo relativo a tareas de **integración, selección, reducción y conversión de datos**, porque partimos de una situación en la que ni siquiera tenemos información tabulada, con variables cuantitativas necesarias para el estudio.

Adicionalmente, en los procesos de obtención de datos, es posible que se produzcan errores por distintos motivos, o que incluso en el propio informe de la web no se haya registrado el dato. Sea por el motivo que sea, hay que tomar una decisión sobre qué hacer con dichos valores, como se nos indica en [1].

Una de las posibilidades pasaría por completar los datos manualmente, como indica Osborne [3], o corregir los procesos de obtención de datos, contemplando nuevas particularidades en la fase de *scraping*.

Durante la primera práctica, se implantó un proceso de control de calidad, en el cual se generaba un archivo *csv* con información referente a cómo había ido el proceso de *scraping*. De esta forma, a través de este archivo, se podrían detectar errores durante el proceso de extracción de datos, y relanzar dichos procesos a partir de funcionalidades concretas del scraper, con el objetivo de completar el set de datos.

A continuación, se muestra una captura del archivo *scraping 2020-05-17 08.46.29.csv*, contenido en la carpeta *logging* del repositorio de Github. En esta imagen, se resaltan procesos en los cuales, la extracción de datos no ha sido completa o ha fallado.



Title	URL	Status	Register	Scraping Time	Request Time	Info
16-29 February 2020 Cyber Attacks Timeline	https://www.hackmagedc	OK	50	6.79	0.24	Leídas 50 de 97
1-15 October 2019 Cyber Attacks Timeline	https://www.hackmagedc	OK	0	2.37	2.30	Leídas 0 de Not Found
1-15 August 2019 Cyber Attacks Timeline	https://www.hackmagedc	OK	20	10.51	1.71	Leídas 20 de 78
1-15 February 2019 Cyber Attacks Timeline	https://www.hackmagedc	OK	0	3.30	2.36	Leídas 0 de Not Found

Con el objetivo pues, de obtener toda la información de las fuentes de origen, se ha relanzado el *scraper* para las casuísticas concretas reflejadas en el archivo de *logging*, creando un nuevo fichero que será integrado con el resto de archivos *csv*. Esta nueva fuente de información, *error-files.csv*, está disponible en la carpeta *data/00_raw* del repositorio de Github.

Antes de realizar la integración de las fuentes de información, será necesario aplicar un conjunto de controles, que asegurarán la calidad de datos. Estos controles, consistirán en filtrar los posibles errores contenidos en la fecha de reporte, eliminando aquellos registros que no pertenezcan al periodo temporal contenido en cada archivo. Adicionalmente, en esta fase, se seleccionarán y adecuarán los nombres de los campos presentes en todas las estructuras de datos.

```
def filter_master(data_raw, year_master, subset):
    # Selección de columnas de interes
    data_filter_cols = data_raw.copy()

    # Se convierte la columna date a formato fecha
    data_filter_cols.Date = pd.to_datetime(data_filter_cols.Date, errors='coerce', format='%Y-%m-%d', utc=True)

    # Se eliminan aquellos registros que estén fuera del 2017
    data_filtered = data_filter_cols.loc[data_filter_cols.Date.dt.year == year_master, subset]

    return data_filtered
```

En el caso del *Timeline*, se añadirá un nivel más de control, eliminando aquellos registros cuya fecha de ataque sea posterior a la fecha de Reporte.

```
def filter_timeline(data_raw, subset):
    rename_col = {
        'id': 'ID',
        'date': 'Date',
        'author': 'Author',
        'target': 'Target',
        'description': 'Description',
        'attack': 'Attack',
        'target_class': 'Target Class',
        'attack_class': 'Attack Class',
        'country': 'Country',
        'link': 'Link',
        'author_report': 'author_report',
        'date_report': 'date_report',
        'views': 'views',
    }

    data_raw.date = pd.to_datetime(data_raw.date, errors='coerce', format='%d/%m/%Y', utc=True)
    data_raw.date_report = pd.to_datetime(data_raw.date_report, errors='coerce', utc=True)

    data_raw.drop(index=data_raw[data_raw.date > data_raw.date_report].index, inplace=True)
    data_raw.drop(index=data_raw[data_raw.date.dt.year < 2019].index, inplace=True)

    data_raw.rename(columns=rename_col, inplace=True)
    data_filtered = data_raw.loc[:, subset]

    return data_filtered
```

Una vez filtrados los archivos originales, y seleccionado el conjunto de datos comunes en ellos, el siguiente paso consistirá en integrarlos verticalmente, con el objetivo de crear una única estructura de datos, sobre la cual aplicar las de tareas de limpieza restantes.

```
# Integración de los datos. Se realiza una fusión vertical para integrar la información 2017 - Actualidad
data = pd.concat(objs=[df_2017_filtered, df_2018_filtered, df_timeline_filtered, df_errors_filtered],
                  ignore_index=True)
```

El siguiente paso, consistirá en garantizar la unicidad de las observaciones, eliminando duplicidades de datos, generadas al reportar el mismo ataque en distintos reportes.

```
# Se eliminan duplicados. De esta forma se p
data.drop_duplicates(inplace=True)
data.reset_index(drop=True, inplace=True)
```

A continuación, se procesará el campo *Attack Class* con el objetivo de solucionar las casuísticas detectadas durante el análisis realizado al inicio de esta sección. Para ello, se homogenizarán los datos a través de la siguiente tabla.

CE	Cyber Espionage
CW	Cyber Warfare
CC	Cyber Crime
H	Hacktivism
>1	Multiple
UK	Unknown

Posteriormente, en el apartado 3.1, se explicará cómo se han tratado los valores perdidos, categorizados a través de la clase *'Unknown'*.

Adicionalmente, a partir de este campo, crearemos una nueva variable dicotómica *ProblemasQC*, que informará de aquellas observaciones que han sufrido algún tipo de procesamiento especial, que pueda afectar a los resultados obtenidos en análisis posteriores.

Una vez realizadas las tareas de homogenización del tipo de ataque, el siguiente paso consistirá en procesar el campo *Author*. En este caso, se realizará una segmentación en dos niveles, en autores conocidos o desconocidos.

A través del análisis del rango de posibles valores que toma esta variable, para determinar si un autor es desconocido, deberá contener el carácter '?' o tener las etiquetas: 'unknown', 'anonymous', '>1'.

```
data.loc[:, 'Author_processed'] = np.where((data.Author.str.contains('unknown', case=False)) |
                                           (data.Author.str.contains('\?', case=False)) |
                                           (data.Author.str.contains('anonymous', case=False)) |
                                           (data.Author.str.contains('>1', case=False)),
                                           'Desconocido', 'Conocido')
```

El siguiente paso, consistirá en procesar la información presente en el campo *Country*. Para ello, primero se asignará la etiqueta '>1' a aquellos valores que presenten diversos países separados por saltos de línea.

```
# Procesado Country
df_country = \
    pd.merge(left=data,
             right=data.Country.fillna('Unknown').str.replace('\n', ' ').str.split(' ').explode(),
             right_index=True,
             left_index=True)[['ID', 'Country_y']]

df_country.reset_index(drop=True, inplace=True)
df_country.drop(df_country[df_country.Country_y == ''].index, inplace=True)
df_country.set_index('ID', drop=False, inplace=True)
df_country.loc[:, 'recuento'] = df_country.groupby(level='ID').nunique()['Country_y']

df_country.loc[:, 'Country_processed'] = df_country.Country_y.where(df_country.recuento == 1, '>1')
```

Una vez codificado este valor, se realizará una discretización asignando el continente al cual pertenece cada país, a través del código ISO, contenido en el campo *Country*.

```
data.loc[:, 'Country_processed'] = df_country[['ID', 'Country_processed']].drop_duplicates().set_index('ID')
data.loc[:, 'Continent'] = data.Country_processed.map(df_continent['Continent'].to_dict())
data.loc[:, 'Pais'] = data.Country_processed.map(df_continent['Nombre Pais'].to_dict())
```

Para realizar esta discretización, se hará uso del documento externo *continente_country.xlsx*, contenido en la carpeta *data/99_aditonal* del repositorio de Github.

Este *Excel* está compuesto por los siguientes campos:

A	B	C	D
ISO ▼	Nombre País ▼	Continente ▼	
AE	Emiratos Árabes Unidos	Asia	
AF	Afganistán	Asia	
AM	Armenia	Asia	
AR	Argentina	América	
AT	Austria	Europa	
AU	Australia	Australia y Oceanía	

El siguiente campo a procesar será *Target*. En este caso, tal como hemos comentado al inicio de esta sección, nos encontramos ante un problema de inconsistencia de datos.

```
array(['O Public administration and defence, compulsory social security',
      'Y Multiple targets', 'Y Multiple Targets',
      'O Public administration, defence, compulsory social security',
      'Y Multiple Industries'], dtype=object)
```

En este caso, existen dos descripciones distintas para las tipologías 'O' e 'Y'. De esta forma, se deberá de realizar una homogenización de los datos.


```
target_class_df = pd.DataFrame(data=data['Target Class'].str.split(' ', n=1, expand=True).values,
                               columns=['Code_target_class', 'Desc_target_class'])

# Homogenización de Descripciones de target class
target_class_dict = {
    'Y': 'Multiple Industries',
    'O': 'Public administration and defence, compulsory social security'
}

target_class_df.loc[:, 'Desc_target_class'] = \
    np.where(target_class_df['Code_target_class'].map(target_class_dict).isnull(),
             target_class_df['Desc_target_class'],
             target_class_df['Code_target_class'].map(target_class_dict))
```

Finalmente, a partir de la variable *Date*, crearemos dos nuevas variables *Año* y *mes*, que servirán para aumentar la granularidad de los datos y tener observaciones más significativas.

```
# Procesado Mes y Año
data.loc[:, 'Year'] = pd.to_datetime(data.Date, utc=True).dt.year
data.loc[:, 'Mes'] = pd.to_datetime(data.Date, utc=True).dt.month
```

Una vez realizadas todas las tareas de limpieza, homogenización y conversión de datos, se seleccionarán aquellos campos de interés para nuestros análisis.

```
# Columnas seleccionadas para sacar la información relevante del dataset
analysis_cols = ['Year', 'Mes', 'Continent', 'Country_name', 'Code_target_class', 'Desc_target_class',
                 'ProblemasQC', 'Author_processed', 'Code_attack_class']
data_complete = data.loc[:, analysis_cols]
```

Con la información seleccionada, el siguiente paso, consistirá en crear un conjunto de variables numéricas que recojan el número de ataques por tipología, en función del periodo de tiempo en el que se originan; el lugar dónde se producen; y el objetivo que afectan, es decir, el tipo de entidad atacada.

De esta forma, se creará una estructura del tipo *crosstab* [2], que contendrá la distribución de los ataques según su tipología. Esta información, nos será de utilidad a la hora de tratar de analizar si hay algún tipo de ataque que afecte en mayor medida a nuestra entidad. Además, podremos estudiar el tipo de relación existente entre distintas tipologías de ataques, en cuestión de tendencias, correlaciones, etc.

Adicionalmente, también será de interés, conocer el número de atacantes conocidos o desconocidos, con el objetivo de poder realizar análisis de tendencias, y averiguar a qué tipología de atacante estamos más expuestos. Por este motivo, se añadirá esta variable cuantitativa, a las ya mencionadas anteriormente.

```
# Creación de las variables dummy de ataques para hacer joins

data_complete_attacks = pd.get_dummies(data_complete[analysis_cols],
                                         columns=['Author_processed', 'Code_attack_class'])

aggregated_data = data_complete_attacks.groupby(analysis_cols[:-2]).sum().reset_index()
```

Como último paso previo a la realización de tareas de análisis, volcaremos la información transformada en el fichero ***EstadisticasAtaques2017_2020_Input.csv***, contenido en la carpeta *Data/01_Clean* del repositorio de Github.

```
aggregated_data.to_csv(path.join(CLEAN_DATA_PATH, 'EstadisticasAtaques2017_2020_Input.csv'),
                          index=False, sep=';', decimal=',')
```

Tal y como comentamos previamente, se complementa parte del código R que incluye el presente informe, con el lenguaje de programación *Python* a través del código *data_processing.py*, con el cual, se resuelven todos los problemas que hemos expuesto y analizado anteriormente. Este *script* se encuentra en la carpeta *src* del repositorio de Github.

A modo de resumen, a continuación, se expondrán las variables finales que componen el fichero *EstadisticasAtaques2017_2020_Input.csv*.

- **Year:** Año en el que se producen los ataques recogidos para cada observación.
- **Mes.** Mes en el que se producen los ataques recogidos para cada observación.
- **Code_target_class.** Código del tipo de entidad afectada por el número de ataques recogidos para cada observación. Este campo contiene las letras iniciales de la variable original *Target Class*.
- **Desc_target_class.** Contiene el descriptivo del tipo de entidad afectada por el número de ataques recogidos para cada observación.
- **Country_name:** nombre del país afectado por el ataque. Este, ha sido obtenido a partir del código ISO de la variable *Country*.
- **Continent:** Continente al que pertenece el país afectado por el ataque.
- **ProblemasQC:** Identificador que señala si la observación concreta ha tenido problemas de calidad identificados en la etapa de generación de la información. Nos indica que los valores cuantificados pueden ser imprecisos, estando clasificados en categorías genéricas, o en entidades no identificadas.
- **Author_processed_conocido:** Número de ataques con autor conocido e identificable que se han producido para ese tipo de entidad, en el país, mes y año correspondiente. Variable creada a partir de la variable original *Author*.
- **Author_processed_conocido.** Número de ataques con autor desconocido que se han producido para ese tipo de entidad, en el país, mes y año correspondiente. Variable creada a partir de la variable original *Author*.

- **Code_attack_class_CC:** Número de ataques del tipo “Cyber Crime” que se han producido para ese tipo de entidad, en el país, mes y año correspondiente. Variable creada a partir del recuento de casos, a través de la variable original *Attack Class*.
- **Code_attack_class_CE:** Número de ataques del tipo “Cyber Espionage” que se han producido para ese tipo de entidad, en el país, mes y año correspondiente. Variable creada a partir del recuento de casos, a través de la variable original *Attack Class*.
- **Code_attack_class_CW:** Número de ataques del tipo “Cyber Warfare” que se han producido para ese tipo de entidad, en el país, mes y año correspondiente. Variable creada a partir del recuento de casos, a través de la variable original *Attack Class*.
- **Code_attack_class_H:** Número de ataques del tipo “Hacktivism” que se han producido para ese tipo de entidad, en el país, mes y año correspondiente. Variable creada a partir del recuento de casos, a través de la variable original *Attack Class*.
- **Code_attack_class_UK:** Número de ataques de tipo ‘Unknown’ que se han producido para ese tipo de entidad, en el país, mes y año correspondiente. Variable creada a partir del recuento de casos, a través de la variable original *Attack Class*.
- **Code_attack_class_>1.** Número de ataques con múltiples tipologías, codificadas a través de valor “>1”, y que se han producido para ese tipo de entidad, en el país, mes y año correspondiente. Variable creada a partir de conteo de casos, a través de la variable original *Attack Class*
- **NumeroAtaques:** Número de ataques total producido para un tipo de entidad, país, mes y año concreto. Variable creada a partir del conteo de casos de la variable original *Attack Class*.

2.2 Carga del nuevo archivo tras el procesamiento de datos.

A continuación, cargaremos el nuevo fichero de *input*, y explicaremos brevemente cómo vamos a avanzar en las siguientes secciones con dichos datos.

```
# Almacenamos el nuevo set de datos en el frame "attacks_Input" para una
# validación adicional, y para explicar algunos de los tratamientos que lle
# varemos a cabo a lo largo del análisis
attacks_Input <- read.csv2(file='EstadisticasAtaques2017_2020_Input.csv',
stringsAsFactors = TRUE)
attacks_Input$Year <- as.factor(attacks_Input$Year)
attacks_Input$Mes <- as.factor(attacks_Input$Mes)

# Creamos una nueva variable con el número total de ataques por observaci
# ón
attacks_Input$NumeroAtaques <- attacks_Input$Code_attack_class_1+attacks
_Input$Code_attack_class_CC+attacks_Input$Code_attack_class_CE+attacks_In
put$Code_attack_class_CW+attacks_Input$Code_attack_class_H+attacks_Input$
Code_attack_class_UK

# Mostramos la estructura del archivo recién cargado
str(attacks_Input)

## 'data.frame':    1748 obs. of  16 variables:
## $ Year                : Factor w/  4 levels "2017","2018",...:
1 1 1 1 1 1 1 1 1 1 ...
## $ Mes                 : Factor w/ 12 levels "1","2","3","4",.
.: 1 1 1 1 1 1 1 1 1 1 ...
## $ Continent           : Factor w/  7 levels "África","América"
, ...: 2 2 2 2 2 2 2 2 2 2 ...
## $ Country_name        : Factor w/ 112 levels "Afganistán","Al
emania", ...: 15 15 18 33 33 33 33 33 33 ...
## $ Code_target_class   : Factor w/ 21 levels "C","D","E","G",.
.: 7 21 13 1 6 7 10 11 12 13 ...
## $ Desc_target_class   : Factor w/ 21 levels "Accommodation an
d food service activities", ...: 11 19 5 12 1 11 15 3 16 5 ...
## $ ProblemasQC         : logi  FALSE FALSE FALSE FALSE FALSE F
ALSE ...
## $ Author_processed_Conocido : int   1 0 0 0 0 1 0 0 1 0 ...
## $ Author_processed_Desconocido: int   0 1 1 3 3 5 1 1 4 10 ...
## $ Code_attack_class_1    : int   0 0 0 0 0 0 0 0 0 0 ...
## $ Code_attack_class_CC   : int   1 1 1 3 3 6 1 1 5 10 ...
## $ Code_attack_class_CE   : int   0 0 0 0 0 0 0 0 0 0 ...
## $ Code_attack_class_CW   : int   0 0 0 0 0 0 0 0 0 0 ...
## $ Code_attack_class_H    : int   0 0 0 0 0 0 0 0 0 0 ...
## $ Code_attack_class_UK   : int   0 0 0 0 0 0 0 0 0 0 ...
## $ NumeroAtaques         : int   1 1 1 3 3 6 1 1 5 10 ...
```

Mostramos un resumen de Los datos por cada campo

summary(attacks_Input)

```
##      Year      Mes      Continent
## 2017:478    1      :186  África      : 32
## 2018:587    2      :183  América     :531
## 2019:434    3      :183  Asia       :363
## 2020:249   10      :161  Australia y Oceanía: 66
##           11      :158  Desconocido      : 84
##           4       :157  Europa          :476
##           (Other):720  International     :196
##           Country_name Code_target_class
## Estados Unidos de América:386 Z      :276
## International              :196 O      :227
## United Kingdom             :114 X      :169
## Desconocido                : 84 K      :138
## Canadá                    : 69 J      :118
## Italia                     : 63 Y      :104
## (Other)                    :836 (Other):716
##
##                               Desc_target_class
## Unknown                      :276
## Public administration and defence, compulsory social security:227
## Individual                   :169
## Financial and insurance activities :138
## Information and communication :118
## Multiple Industries          :104
## (Other)                      :716
## ProblemasQC      Author_processed_Conocido Author_processed_Desconocido
## Mode :logical    Min.   : 0.0000      Min.   : 0.000
## FALSE:1520       1st Qu.: 0.0000      1st Qu.: 1.000
## TRUE :228        Median : 0.0000      Median : 1.000
##                  Mean    : 0.4491      Mean    : 2.231
##                  3rd Qu.: 1.0000      3rd Qu.: 2.000
##                  Max.    :13.0000      Max.    :57.000
##
## Code_attack_class_.1 Code_attack_class_CC Code_attack_class_CE
## Min.   :0.000000      Min.   : 0.000      Min.   :0.0000
## 1st Qu.:0.000000      1st Qu.: 0.000      1st Qu.:0.0000
## Median :0.000000      Median : 1.000      Median :0.0000
## Mean   :0.001144      Mean   : 1.594      Mean   :0.2677
## 3rd Qu.:0.000000      3rd Qu.: 1.000      3rd Qu.:0.0000
## Max.   :1.000000      Max.   :31.000      Max.   :7.0000
##
## Code_attack_class_CW Code_attack_class_H Code_attack_class_UK
## Min.   :0.0000      Min.   : 0.0000      Min.   : 0.00000
## 1st Qu.:0.0000      1st Qu.: 0.0000      1st Qu.: 0.00000
## Median :0.0000      Median : 0.0000      Median : 0.00000
## Mean   :0.0595      Mean   : 0.6899      Mean   : 0.06751
## 3rd Qu.:0.0000      3rd Qu.: 0.0000      3rd Qu.: 0.00000
```

```
## Max.      :4.0000      Max.      :60.0000      Max.      :17.00000
##
## NumeroAtaques
## Min.      : 1.00
## 1st Qu.: 1.00
## Median : 1.00
## Mean     : 2.68
## 3rd Qu.: 2.00
## Max.     :69.00
##
```

Con una función exploratoria tan sencilla como “summary”, somos capaces de empezar a entender e interpretar buena parte de los datos del dataset, con algunos aspectos especialmente interesantes como los siguientes:

- Se aprecia un número de casos relativamente estable en cada año (hay que tener en cuenta que para 2020 sólo tenemos datos de los primeros meses).
- Parece que el número de casos por mes también se mantiene relativamente estable, aunque los 3 primeros meses tienen un número de casos ligeramente superior.
- El país con mayor impacto de ataques es Estados Unidos, y por continentes los principales objetivos de ataques son America y Europa.
- Un gran número de casos tienen como objetivo de ataque el sector desconocido (*unknown*), seguido del sector público, el cual, corresponde al tipo de entidad que estamos representando.
- En cada una de las observaciones, dado el nivel de granularidad escogido (Año/Mes/País/Entidad), el número absoluto de ataques de cada tipo (campos *Code_attack_class_XX*) son relativamente bajos, con medidas de tendencia central en torno al valor 1. Esto sugiere y recomienda que, para responder a las distintas cuestiones que hemos planteado en la sección inicial, se utilicen distintos niveles de agregación que permitan obtener datos globales con mayor significatividad.

Todo lo anterior resulta de gran utilidad para tener una mejor composición de lugar, y apunta de manera temprana algunas cautelas que deberemos tener en consideración a lo largo del ejercicio.

Adicionalmente, haremos una comprobación sobre una de las variables más relevantes de nuestros análisis numéricos: el número de ataques totales. Para ello, los segmentaremos a partir de las distintas tipologías del atributo *Attack Class*.

Análisis de integridad en la carga: comprobación de la completitud del dataset y la consistencia entre variables numéricas.

```
colSums(attacks_Input[7:15])

##              ProblemasQC      Author_processed_Conocido
##              228              785
## Author_processed_Desconocido      Code_attack_class_.1
##              3899              2
##      Code_attack_class_CC      Code_attack_class_CE
##              2786              468
##      Code_attack_class_CW      Code_attack_class_H
##              104              1206
##      Code_attack_class_UK
##              118

cat("Suma del número de ataques por categoría: ",sum(colSums(attacks_Input[8:15]))[3:8]),"\n")

## Suma del número de ataques por categoría:  4684

cat("Suma del número de ataques de acuerdo a si se conoce el atacante: ",sum(colSums(attacks_Input[8:15]))[1:2]),"\n")

## Suma del número de ataques de acuerdo a si se conoce el atacante:  4684

cat("Valor acumulado en la variable NumeroAtaques: ",sum(attacks_Input$NumeroAtaques))

## Valor acumulado en la variable NumeroAtaques:  4684
```

Con lo anterior, ya podemos considerar que tenemos un primer set de datos, con información consistente, exacta, única y válida; garantizando de esta forma, un nivel de calidad óptimo, que nos permitirá realizar tareas de análisis.

Aun así, cabe remarcar que quedarán pendientes tareas de procesamiento de datos, como la normalización/estandarización de variables; el análisis de valores extremos, o la verificación de suposiciones de normalidad y homocedasticidad con el objetivo de poder aplicar distintas pruebas estadísticas.

3. Limpieza de los datos.

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

En el dataset utilizado para la realización de esta práctica, nos hemos encontrado ante un problema de pérdida total de la información (casos que sí están documentados en la fuente de datos, pero que no se han extraído). En este caso, a través del fichero de *logging*, se han detectado un conjunto de registros que, por errores en la etapa de extracción de datos, no se han leído adecuadamente, reduciendo de esta forma el número total de observaciones posibles. En este caso, para solucionar esta casuística, se ha relanzando el proceso de *scraping*.

Por otra parte, a lo largo de la etapa de acondicionamiento de datos, se han detectado observaciones que contienen valores perdidos, generando así, un problema de pérdida parcial de la información. En este sentido, para solventar dicha casuística, se han empleado metodologías distintas, en función de la variable afectada.

La primera variable que contiene *missing values*, es *Attack Class*. Para este campo, se han considerado como valores perdidos, las etiquetas que contenían el carácter '?' o los valores 'unknown', 'not found' y 'n/a'. En este caso, debemos diferenciar dos metodologías distintas de tratamiento de valores perdidos.

La primera, intentará solventar los valores representados por la etiqueta 'Not found'. Esta, hace referencia a la falta de información producida por un error durante la etapa de obtención de datos. Ante tal situación, una de las opciones propuestas por Osborne [3], sería la de recoger los datos de forma manual, siempre que suponga una inversión de tiempo aceptable. En concreto, existen un total de 1314 casos *Not Found*, por lo que esta opción ha sido descartada. La alternativa, será realizar la imputación de estos valores perdidos, a través de una medida de tendencia central. De esta forma, se calculará cual es la clase de ataque más representativa, es decir, aquella con una frecuencia de aparición mayor, en función de la variable *Attack*.

```
# Ahora se intentarán informar los datos perdidos a partir del campo 'Ataque'. Para ello, primero
# se obtiene la información de aquellos campos informados correctamente. Aquellos que no son conocidos.
data_known = data[data.Desc_attack_class != 'Unknown']

# El metodo de llenar los missing values, será a partir del elemento que presente máxima frecuencia entre
# la variable categorica Attack
map_atk = data_known.groupby(['Attack', 'Desc_attack_class']).count().max(level=[0, 1])\
    .reset_index(1)['Desc_attack_class'].to_dict()

# Se informa el campo mediante el mapping obtenido
data.loc[data.Desc_attack_class == 'Unknown', 'Ffill_Desc_attack_class'] = \
    data.loc[data.Desc_attack_class == 'Unknown', 'Attack'].map(map_atk)

data.loc[:, 'Desc_attack_class'] = data['Desc_attack_class'].where(data['Ffill_Desc_attack_class'].isnull(),
    data['Ffill_Desc_attack_class'])
```

Cabe destacar, que a parte de la imputación de valores perdidos a través de medidas de tendencia central, existen otras técnicas más precisas que utilizan métodos probabilísticos entrenados a través del resto de información presente en el set de datos. Entre estas técnicas de imputación de datos más avanzadas destacan el *K-Nearest-Neighbors* o el *MissForest*. [1].

La segunda metodología empleada, será la de asignar una constante, tanto a aquellos valores perdidos representados por las etiquetas '*unknown*', '?' y '*n/a*', como a los valores '*not found*', que no se han imputado correctamente.

La siguiente variable que contiene registros vacíos es *Country*. En este caso, se les asignará la constante '*desconocido*', y en función del análisis que realicemos, descartaremos o no la información referente a esta variable. De esta forma, podremos aprovechar el resto de campos informados en los análisis que no requieran información del país. Esta técnica de análisis mencionada, es conocida como *pairwise* [4].

```
df_country = \
    pd.merge(left=data,
            right=data.Country.fillna('Unknown').str.replace('\n', ' ').str.split(' ').explode(),
            right_index=True,
            left_index=True)[['ID', 'Country_y']]
```

Otra variable en la que encontramos valores perdidos es *target_class*. En este caso, los valores perdidos se representan a través de las etiquetas '*Not Found*' y '*Z Unknown*'. Del mismo modo que con el atributo *Attack Class*, nos encontramos ante una pérdida parcial de la información producida por errores en la etapa de extracción de los datos. En concreto, la volumetría de valores perdidos para este campo ('*Not Found*') es de 1382 registros, por lo que se ha decidido asignarles la clase '*Z Unknown*', y realizar técnicas de *pairwise* en las tareas de análisis y extracción de conocimiento. Adicionalmente, cabe destacar que se ha descartado la posibilidad de eliminar todas observaciones que contienen valores perdidos, debido la gran cantidad de información que se perdería.

```
data.loc[:, 'Target Class'] = data['Target Class'].where(data['Target Class'] != 'Not Found', 'Z Unknown')
```

Finalmente, la última variable que contiene valores perdidos es *Author*. Tal como se expone en la introducción, uno de los objetivos será analizar si estamos estadísticamente más expuestos, a ataques realizados por autores conocidos o desconocidos. La finalidad, será saber si debemos centrar nuestros recursos en la creación sistemas de defensa versátiles, o focalizados en técnicas de ataque características de cada autor.

En este sentido, para el campo *Autor*, asignaremos a todos los valores perdidos la variable '*Desconocido*'. En concreto, se consideran las etiquetas '*unknown*', '*anonymous*', '>1', y aquellas clases que contienen el símbolo '?', como valores perdidos.

```
data.loc[:, 'Author_processed'] = np.where((data.Author.str.contains('unknown', case=False)) |
                                           (data.Author.str.contains('\?', case=False)) |
                                           (data.Author.str.contains('anonymous', case=False)) |
                                           (data.Author.str.contains('>1', case=False)),
                                           'Desconocido', 'Conocido')
```

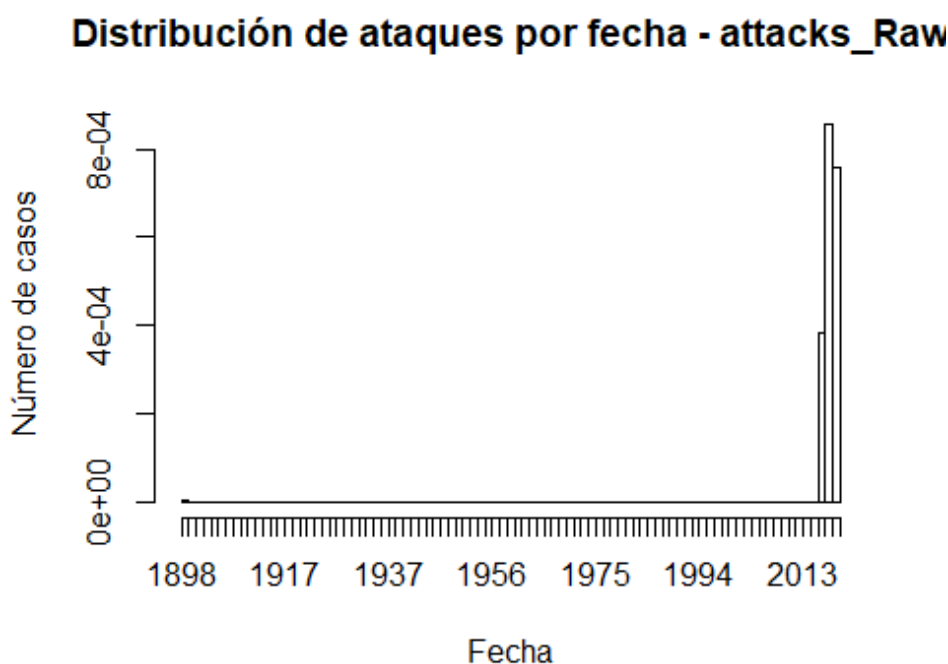
3.2 Identificación y tratamiento de valores extremos

Como hemos visto en secciones anteriores, gran parte de las variables que utilizamos en nuestro estudio, son cualitativas. Sin embargo, cuando hablamos de valores extremos o *outliers*, las variables afectadas, son las cuantitativas. De hecho, según la definición de Subirats, Pérez y Calvo [1], “*son observaciones que se desvían tanto del resto, que levantan sospechas sobre si fueron generadas mediante el mismo mecanismo*”.

A partir de la definición anterior, durante la etapa de filtrado de los datos de origen, se detectó que la variable *Date*, contenía un conjunto de valores anómalos.

```
# Extracción y conversión del campo que registra la fecha del incidente
fechas <- as.Date(attacks_Raw$Date, format="%d/%m/%Y")

# Análisis gráfico de la distribución de incidentes por fecha
hist(fechas, breaks=100, main="Distribución de ataques por fecha - attacks_Raw", xlab = "Fecha", ylab="Número de casos")
```



Tal y como se puede observar, existen unas cuantas observaciones que presentan el año 1900, suponiendo más de 100 años de diferencia, con el resto de las observaciones del set de datos.

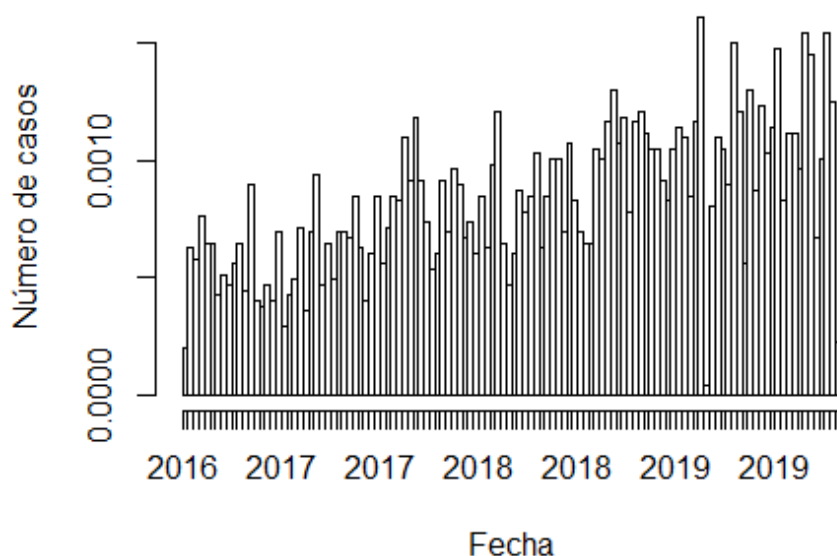
En este caso, para evitar incrementar el error en la varianza de la dimensión temporal y evitar sesgos en la realización de cálculos y estimaciones, se eliminarán estas muestras del set de datos.

```
# Se observa que necesariamente hay valores erróneos (el típico registro
de valor nulo y/o conversión a 01/01/1900). Veámoslo en modo tabla.
table(format(fechas,"%Y"))

##
## 1900 2017 2018 2019 2020
##    4  951 1338 1693  482

# Analicemos de nuevo la distribución de casos por fecha si eliminamos lo
s casos de 1900
hist(fechas[format(fechas,"%Y")!="1900"], breaks=100, main="Distribución
de ataques por fecha - attacks_Raw", xlab = "Fecha", ylab="Número de caso
s")
```

Distribución de ataques por fecha - attacks_Raw



Por otra parte, tal como se ha comentado en secciones anteriores, las variables cuantitativas presentes, han sido construidas por nosotros a partir del set de datos original. Este método de construcción, se basa en el recuento de tipologías de ataque, teniendo en cuenta el país, el mes, el año y la entidad de cada observación.

Al realizar esta cuantificación de datos, nos aseguramos evitar heterogeneidades producidas por la presencia de distintas unidades de medida en las fuentes de datos. Sin embargo, los recuentos obtenidos, podrían contener algún problema intrínseco que afectase a nuestros análisis, por lo que será necesario realizar una inspección de valores atípicos, a lo largo de las distintas variables cuantitativas. De esta forma, se analizará la presencia de *outliers*, a partir del nivel de granularidad más bajo con el que se ha realizado la cuantificación de valores, es decir, a partir de la combinación de *Año/Mes/Entidad/País*.

Para cada una de estas observaciones, disponemos de las siguientes variables numéricas:

- **Author_processed_Conocido:** Número de casos observados con autor reconocido y documentado. En el análisis renombraremos la variable como “*casosConAutor*”.
- **Author_processed_Desconocido:** Número de casos observados con autor no identificado, y para los cuales, resultaría mucho más difícil analizar patrones o detectar reincidencia, por parte del autor, del uso de metodologías concretas. En el análisis renombraremos la variable como “*casosAnonimos*”
- **NumeroAtaques:** Número total de incidentes de seguridad o ataques que se han documentado para la selección *Año/Mes/Entidad/País*.
- **Code_attack_class_>1:** Número de incidentes con más de un tipo de ataque empleado. En el análisis renombraremos la variable como “*casosMultiataque*”
- **Code_attack_class_CC:** Número de incidentes de tipo *Cyber Crime*. En el análisis renombraremos la variable como “*casosCyberCrime*”
- **Code_attack_class_CE:** Número de incidentes de tipo *Cyber Espionage*. En el análisis renombraremos la variable como “*casosCyberEspionage*”
- **Code_attack_class_CW:** Número de incidentes de tipo *Cyber Warfare*. En el análisis renombraremos la variable como “*casosCyberWarfare*”
- **Code_attack_class_H:** Número de incidentes de tipo *Hacktivism*. En el análisis renombraremos la variable como “*casosHacktivism*”
- **Code_attack_class_UK:** Número de incidentes de tipo *Unknown*. En el análisis renombraremos la variable como “*casosTipoAtaqueDesconocido*”

Creación de Las variables anteriores

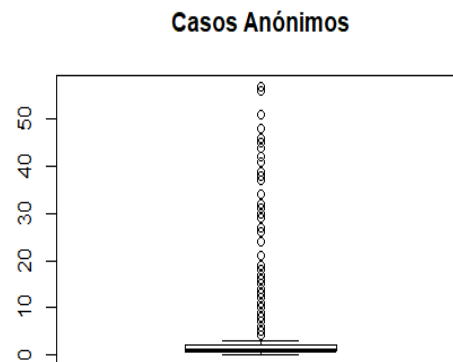
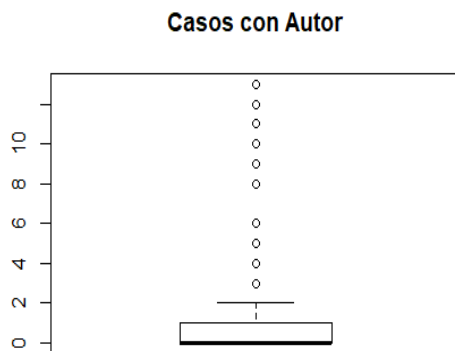
```
casosConAutor <- attacks_Input$Author_processed_Conocido
casosAnonimos <- attacks_Input$Author_processed_Desconocido
NumeroAtaques <- attacks_Input$NumeroAtaques
casosMultiataque <- attacks_Input$Code_attack_class_.1
casosCyberCrime <- attacks_Input$Code_attack_class_CC
casosCyberEspionage <- attacks_Input$Code_attack_class_CE
casosCyberWarfae <- attacks_Input$Code_attack_class_CW
casosHacktivism <- attacks_Input$Code_attack_class_H
casosTipoAtaqueDesconocido <- attacks_Input$Code_attack_class_UK
```

En primer lugar, vamos a ver, mediante gráficos de cajas, si encontramos valores extremos.

Representación de Los boxplot para Las variables numéricas

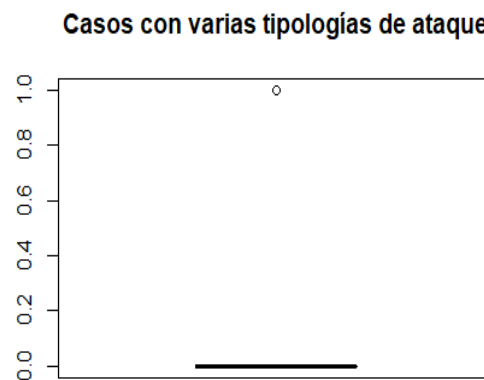
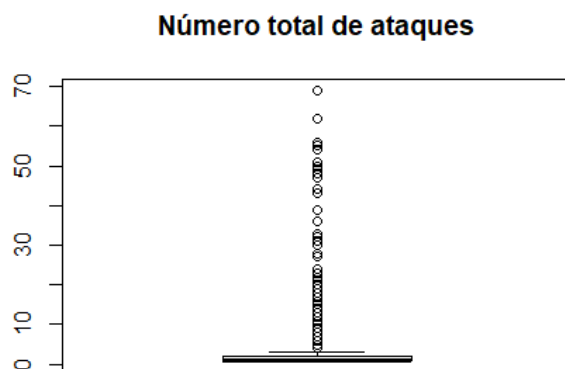
```
boxplot(casosConAutor, main="Casos con Autor")
```

```
boxplot(casosAnonimos, main="Casos Anónimos")
```



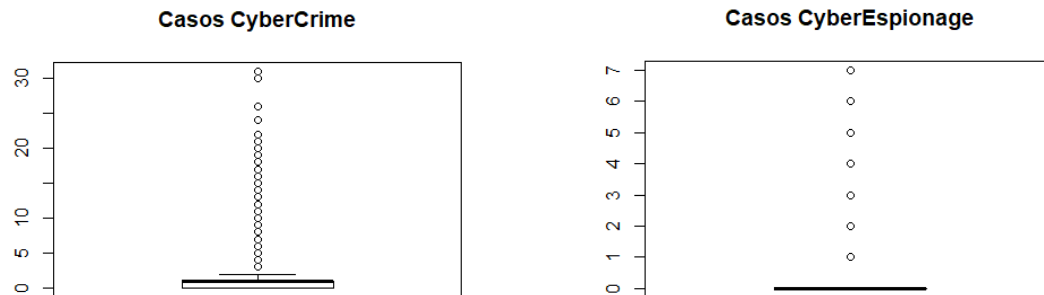
```
boxplot(NumeroAtaques, main="Número total de ataques")
```

```
boxplot(casosMultiataque, main="Casos con varias tipologías de ataque")
```



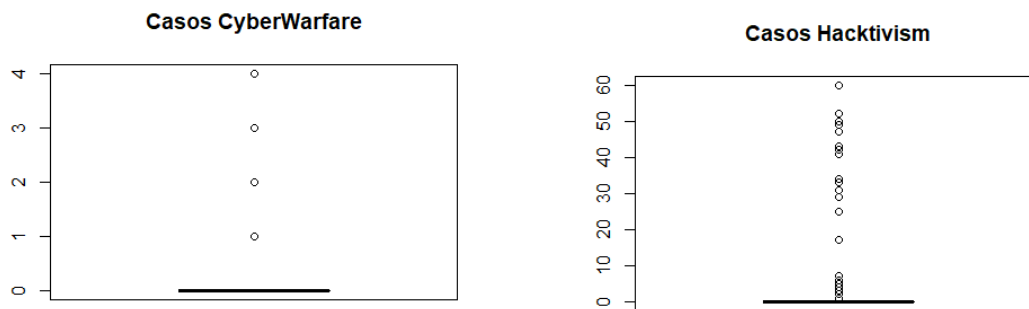
```
boxplot(casosCyberCrime, main="Casos CyberCrime")
```

```
boxplot(casosCyberEspionage, main="Casos CyberEspionage")
```

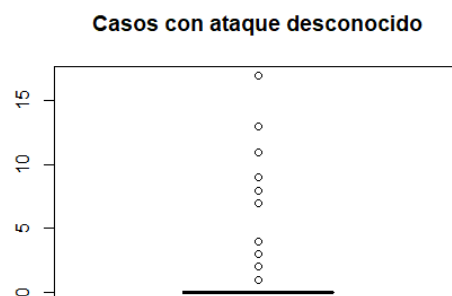


```
boxplot(casosCyberWarfare, main="Casos CyberWarfare")
```

```
boxplot(casosHacktivism, main="Casos Hacktivism")
```



```
boxplot(casosTipoAtaqueDesconocido, main="Casos con ataque desconocido")
```



Tal como se puede observar a través del análisis gráfico de *outliers*, las distribuciones del número de ataques, presentan medidas de tendencia central con valores muy pequeños, de prácticamente 0. Este hecho es producido por el grado de granularidad elegido, basado en cuatro variables cualitativas (Año, Mes, Entidad y País). De esta forma, se justifica que en los análisis que vayamos a hacer, se tengan que agregar hasta 3 de las dimensiones anteriores, produciendo modificaciones en los valores de las

variables cuantitativas. Estos nuevos valores, exigirán de nuevos procesos de análisis y tratamientos particulares, que serán explicados y justificados a lo largo de la sección 4.

Aunque, como decimos, estas agregaciones podrán depender del análisis concreto que vayamos a hacer, vamos a valorar la utilidad de los mecanismos de agregación a través de la comparación de los gráficos anteriores, con los que obtendríamos utilizando el nuevo granulo de la dimensión geográfica.

Para profundizar en la utilidad de los mecanismos de agregación, hemos seguido el sistema de trabajo propuesto en [5]. Adicionalmente, para no sobrecargar el informe de gráficos, se reducirá el número de variables a analizar, considerando el número total de ataques, sin tener en cuenta su tipología.

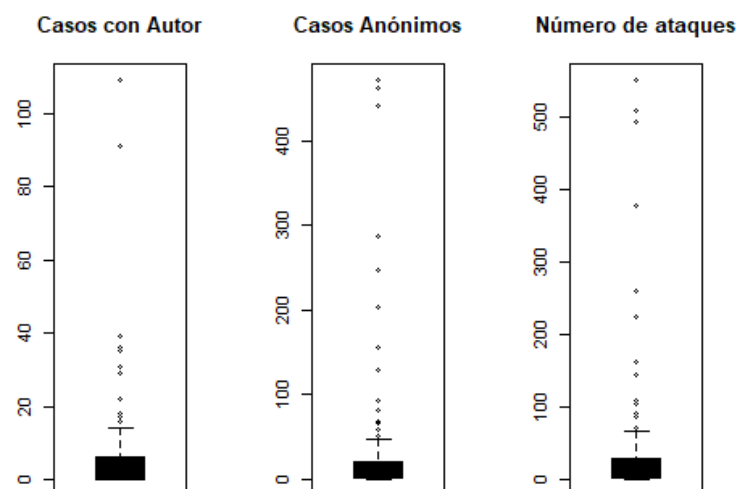
En primer lugar necesitaremos reconstruir la agregación de datos considerando las cuatro variables cualitativas seleccionadas. El cambio principal con respecto al set de datos inicial es el cambio en la dimensión geográfica de país a continente

```
library("dplyr") # Librería que utilizaremos para las tareas de agregación [3]
```

```
attacks_agg1 <- attacks_Input %>% group_by(Code_target_class, Desc_target_class, Continent) %>% summarize(NumeroAtaques=sum(NumeroAtaques), casosConAutor=sum(Author_processed_Conocido), casosAnonimos=sum(Author_processed_Desconocido))
```

Representación de los boxplot para las variables numéricas
`par(mfrow=c(1,3))`

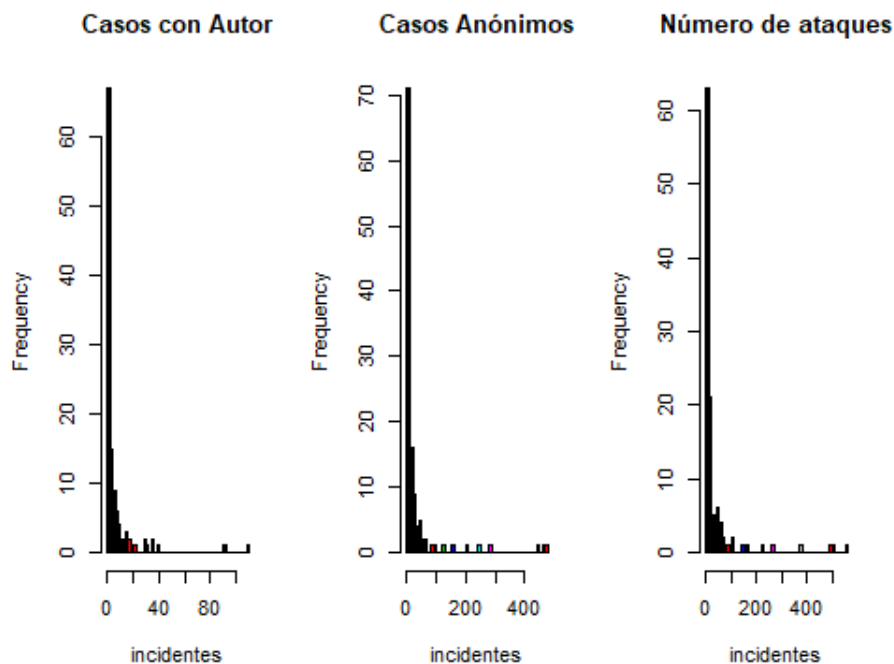
```
boxplot(attacks_agg1$casosConAutor, col=attacks_agg1$Code_target_class, main="Casos con Autor")
boxplot(attacks_agg1$casosAnonimos, col=attacks_agg1$Code_target_class, main="Casos Anónimos")
boxplot(attacks_agg1$NumeroAtaques, col=attacks_agg1$Code_target_class, main="Número de ataques")
```



Representación de los histogramas para las variables numéricas

```
par(mfrow=c(1,3))

hist(attacks_agg1$casosConAutor, col=attacks_agg1$Code_target_class, breaks = 50, main="Casos con Autor", xlab="incidentes")
hist(attacks_agg1$casosAnonimos, col=attacks_agg1$Code_target_class, breaks = 50, main="Casos Anónimos", xlab="incidentes")
hist(attacks_agg1$NumeroAtaques, col=attacks_agg1$Code_target_class, breaks = 50, main="Número de ataques", xlab="incidentes")
```

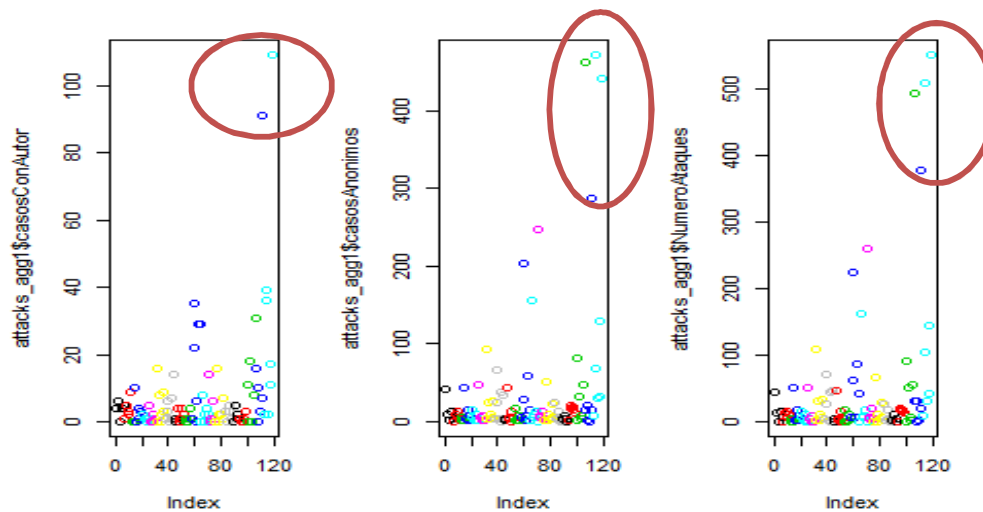


A partir de los gráficos obtenidos, seguimos observando valores muy alejados, por lo que será interesante analizar de manera visual, si están relacionados con sectores de empresas concretos.

Representación de los gráficos de dispersión de valores

```
par(mfrow=c(1,3))

plot(attacks_agg1$casosConAutor, col=attacks_agg1$Code_target_class)
plot(attacks_agg1$casosAnonimos, col=attacks_agg1$Code_target_class)
plot(attacks_agg1$NumeroAtaques, col=attacks_agg1$Code_target_class)
```



A través de los diagramas de dispersión, se puede observar que los valores alejados pertenecen a 3 entidades distintas, representadas por los colores verde, azul claro y azul marino. Vamos a ver en modo tabla a qué entidades son.

Análisis de entidades que ocasionan valores extremos

```
temp <- attacks_agg1 %>% group_by(Code_target_class, Desc_target_class) %
>% summarize(sum(NúmeroAtaques))
temp2 <- as.data.frame(temp)
head(temp2[order(-temp2$sum(NúmeroAtaques)),])
```

##	Code_target_class	Desc_target_class	sum(NúmeroAtaques)
## 21	Z	Unknown	1397
## 19	X	Individual	704
## 20	Y	Multiple Industries	468
## 12	O	Public administration and defence, compulsory social security	443
## 14	Q	Human health and social work activities	300
## 7	J	Information and communication	204

A partir de los valores obtenidos en la tabla, se aprecia que, los casos que hemos clasificado como *desconocidos*, así como los valores *individual* o *multiple industries*, están introduciendo un ruido que hace incomparables los valores entre sectores.

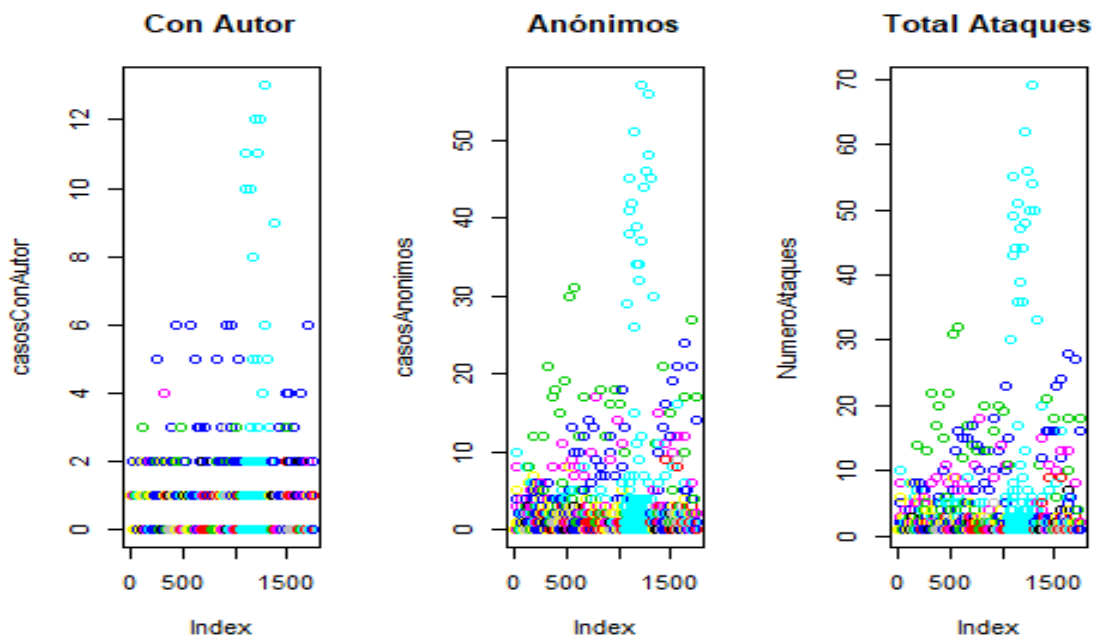
Seguramente la mejor decisión que deberemos tomar para estos valores es excluirllos de los análisis sectoriales, o incluso tratar de reasignar los valores desconocidos a alguno de los otros tipos de entidades. Comentaremos más sobre este aspecto, en la sección 4 de análisis.

Por el momento, como último análisis visual, vamos a representar los gráficos sin considerar los valores que hemos identificado como *outliers*.

Representación de Los gráficos de dispersión de valores

```
par(mfrow=c(1,3))

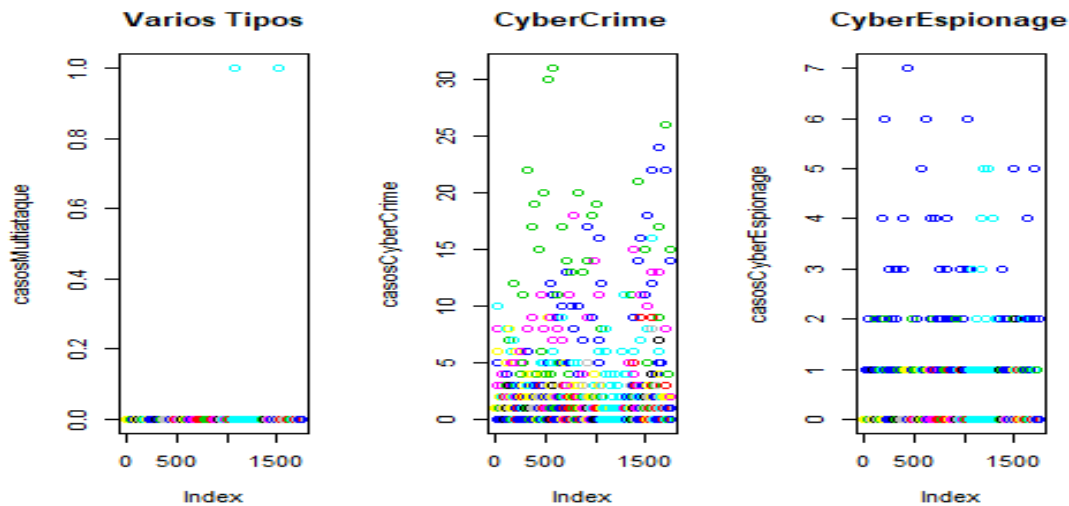
plot(casosConAutor,col=attacks_Input$Code_target_class, main="Con Autor")
plot(casosAnonimos, col=attacks_Input$Code_target_class, main="Anónimos")
plot(NumeroAtaques, col=attacks_Input$Code_target_class, main = "Total Ataques")
```



```

plot(casosMultiataque, col=attacks_Input$Code_target_class, main="Varios
Tipos")
plot(casosCyberCrime, col=attacks_Input$Code_target_class, main="CyberCri
me")
plot(casosCyberEspionage, col=attacks_Input$Code_target_class, main="Cybe
rEspionage")

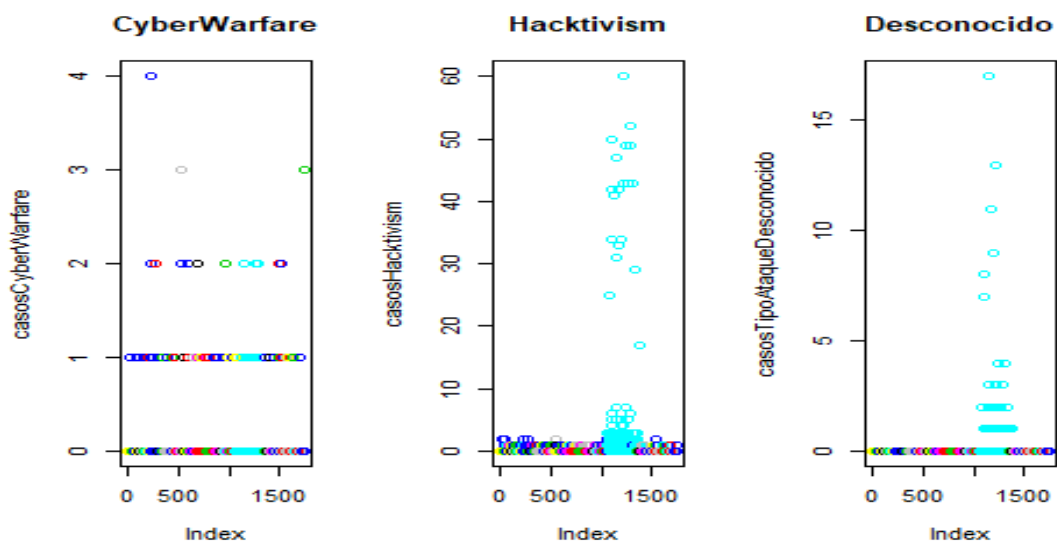
```



```

plot(casosCyberWarfare, col=attacks_Input$Code_target_class, main="CyberW
arfare")
plot(casosHacktivism, col=attacks_Input$Code_target_class, main="Hacktivi
sm")
plot(casosTipoAtaqueDesconocido, col=attacks_Input$Code_target_class, mai
n="Desconocido")

```



Representación de los gráficos de dispersión de valores, excluyendo las entidades que generan valores extremos

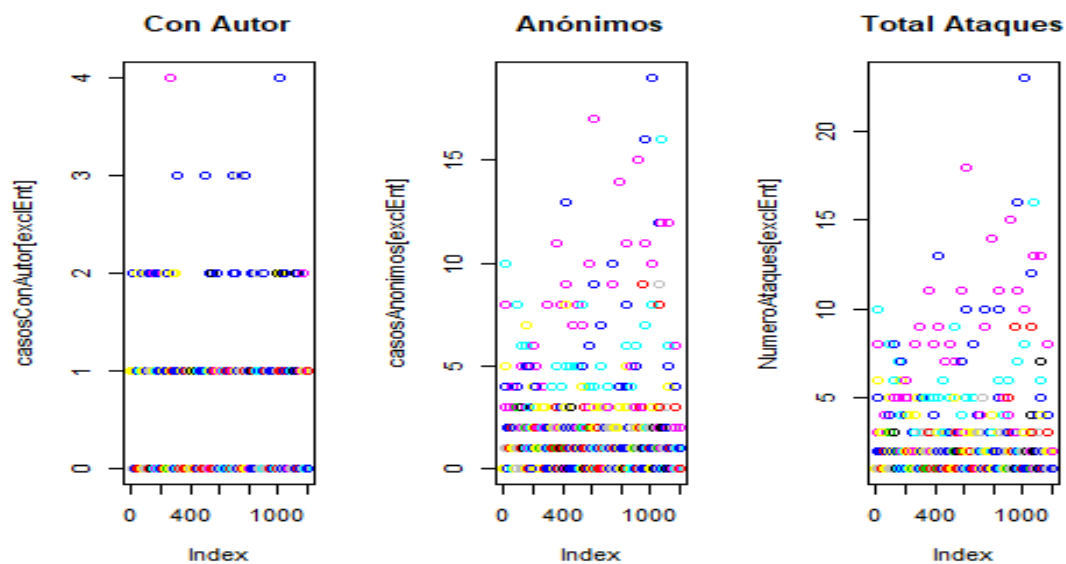
```
par(mfrow=c(1,3))
```

```
exclEnt <- attacks_Input$Code_target_class!="X" & attacks_Input$Code_target_class!="Y" & attacks_Input$Code_target_class!="Z"
```

```
plot(casosConAutor[exclEnt], col=attacks_Input$Code_target_class[exclEnt], main = "Con Autor")
```

```
plot(casosAnonimos[exclEnt], col=attacks_Input$Code_target_class[exclEnt], main = "Anónimos")
```

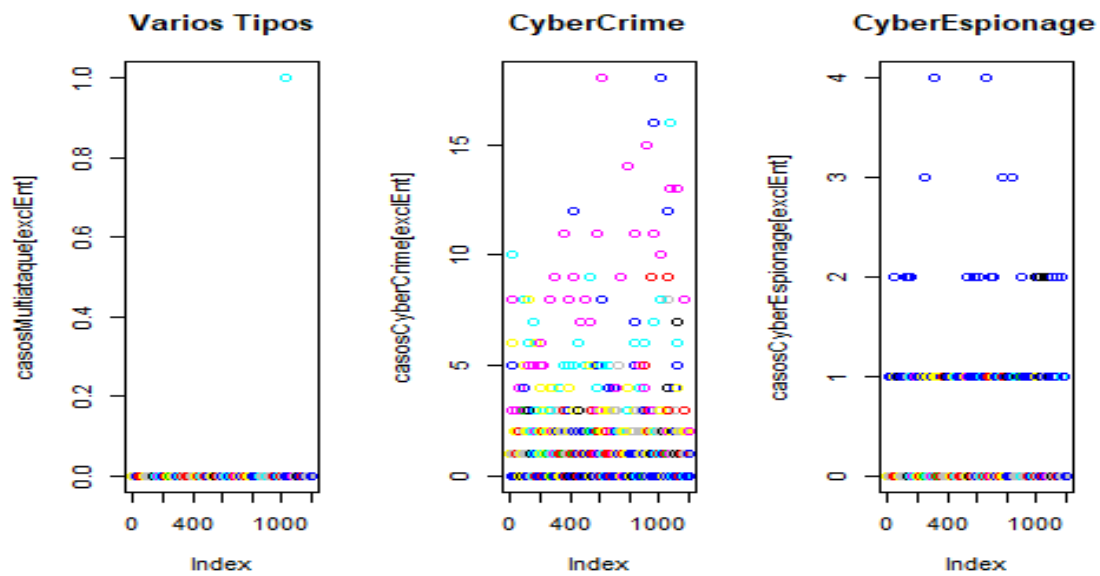
```
plot(NumeroAtaques[exclEnt], col=attacks_Input$Code_target_class[exclEnt], main = "Total Ataques")
```



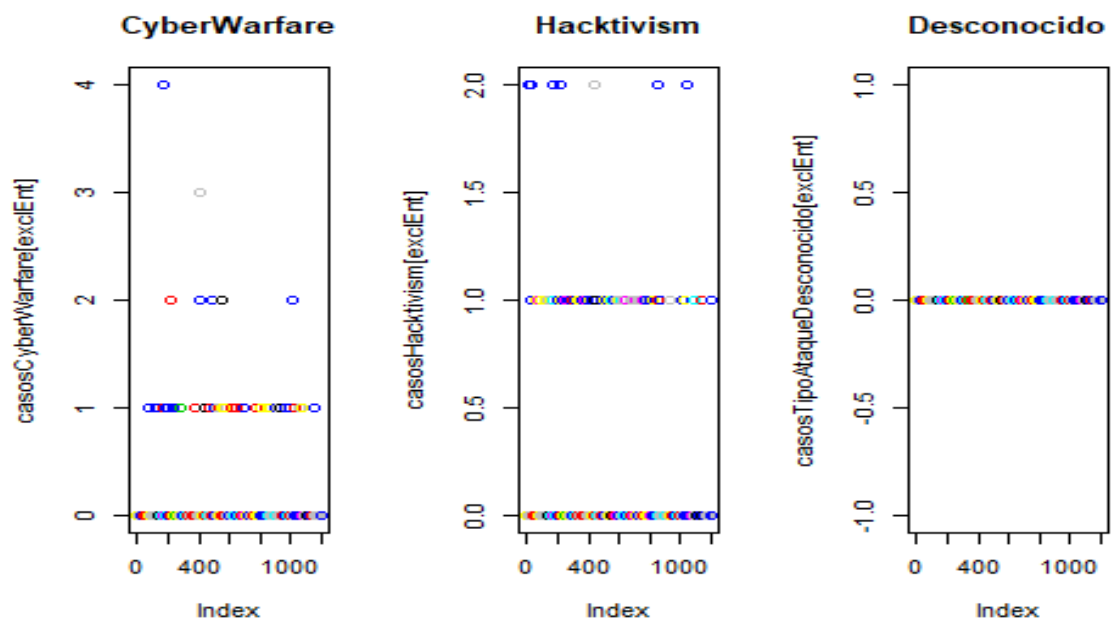
```
plot(casosMultiataque[exclEnt], col=attacks_Input$Code_target_class[exclEnt], main="Varios Tipos")
```

```
plot(casosCyberCrime[exclEnt], col=attacks_Input$Code_target_class[exclEnt], main="CyberCrime")
```

```
plot(casosCyberEspionage[exclEnt], col=attacks_Input$Code_target_class[exclEnt], main="CyberEspionage")
```



```
plot(casosCyberWarfare[exclEnt], col=attacks_Input$Code_target_class[exclEnt], main="CyberWarfare")
plot(casosHacktivism[exclEnt], col=attacks_Input$Code_target_class[exclEnt], main="Hacktivism")
plot(casosTipoAtaqueDesconocido[exclEnt], col=attacks_Input$Code_target_class[exclEnt], main="Desconocido")
```



De todo lo anterior, podemos extraer varias conclusiones muy interesantes sobre la presencia de valores extremos:

- Comprobamos que, las categorías X, Y, Z de la variable *Target Class*, constituyen casos particulares a considerar separadamente. En concreto, el valor *Unknown* no debería contemplarse en aquellos análisis que consideren como atributo relevante, el tipo de entidad, ya que pueden incluir valores agregados de otros tipos de entidades, y en consecuencia generan valores extremos.
- Observamos que la tipología de ataques '*Multiataque*' tiene un único valor espurio, siendo siempre 0, por lo que constituye una variable que puede omitirse sin pérdida de generalidad.
- Vemos que una vez excluidas las categorías X, Y, Z, no existen observaciones con tipo de ataque desconocido, por lo que ambas variables vienen a recoger los valores perdidos.
- A través de las representaciones realizadas en esta sección, se pueden observar distribuciones con colas a derechas, creando preponderancia a la existencia de valores bajos. Será necesario pues, corregir este hecho en todos aquellos análisis que supongan distribuciones normales.

4. Análisis de los datos.

4.1 Planificación de los análisis a aplicar.

Tal y como se expone en la introducción, una de las principales motivaciones para la realización de este proyecto, es la de ser capaces de adoptar un comportamiento proactivo ante las amenazas cibernéticas presentes en el mundo actual. Para ello, será necesario obtener, a través del análisis de los datos procesados, el conocimiento necesario para responder las preguntas planteadas en la introducción.

Para ello, como primera aproximación, se han planteado un conjunto de análisis que podrían aplicarse sobre los datos, para responder cada una de las cuestiones iniciales. A continuación, se expondrán cada uno de estos planteamientos.

¿Hasta qué punto mi compañía se encuentra más expuesta a ciberataques, en función del sector en el que se encuentra?

En esta prueba, a partir del set de datos, crearíamos tantas muestras independientes como sectores existan, descartando aquellos considerados *outliers* (sectores X, Y, Z) en la sección 3.2. Una vez dividido el set de datos, utilizaríamos la muestra de datos con observaciones pertenecientes al sector público (sector O), para compararla con el resto de sectores mediante un contraste de hipótesis sobre la media de ataques recibidos.

Por otra parte, sería interesante construir para cada tipología de ciberataque, el intervalo de confianza de la media de ataques documentados. De esta forma, seríamos capaces de analizar hasta qué punto una empresa del sector público está expuesta a recibir un tipo de ataque en concreto.

¿Existe algún periodo del año en que debería ampliar el presupuesto y los medios necesarios, al estar más expuesto a recibir ciberataques?

Para responder esta pregunta, sería interesante crear 12 muestras distintas a partir de las observaciones pertenecientes a cada mes del año. A continuación, un posible análisis consistiría en realizar alguna prueba estadística que sirviese para comparar más de dos grupos de datos, y que nos diese información sobre si los meses son estadísticamente similares. Por otra parte, también sería interesante realizar pruebas estadísticas en las que se compararían dos grupos de datos, como contrastes de hipótesis sobre la media de ataques recibidos. En este caso, se cogería como referencia aquel mes con mayor cantidad de amenazas documentadas.

Adicionalmente, al encontrarnos ante una variable temporal, será interesante analizar la evolución de los ataques y ser capaces de utilizar la información recopilada, para intentar predecir el volumen de eventos futuros. En este sentido, se utilizaría un modelo de regresión lineal a partir del número de ataques totales producidos a empresas del sector público.

Un tipo de análisis más sofisticado y, seguramente, mejor adaptado al tipo de cuestión que nos estamos planteando, pasaría por el análisis de series temporales [6]. Este tipo de análisis, sin embargo, excede en tiempo y recursos las posibilidades de esta práctica.

¿A qué tipología de ciberataque mi compañía está más expuesta? ¿Debería por ello, analizar la documentación cualitativa, así como los detalles técnicos referentes a este tipo de ataque?

Para responder esta cuestión, se podrían realizar distintos planteamientos. El primero de estos consistiría en analizar la correlación presente entre los distintos tipos de ataques en función del sector al que afectan. De esta forma, las empresas del sector público, estarán más expuestas a aquellas tipologías de ataque cuya correlación sea más elevada.

Por otra parte, se podría plantear un modelo de regresión logística para analizar como varían los *ODDS* de que una empresa del sector público sea atacada, a través de los coeficientes estimados para cada tipología de ataque. En este sentido, se crearía una variable dicotómica que adoptaría el valor 1 en aquellas observaciones pertenecientes al sector público; y el valor 0, para el resto de casos. Adicionalmente, se debería de realizar un procesamiento de datos para transformar las variables que contienen información de los tipos de ataque, en una única variable categórica. Por consiguiente, el modelo de regresión logística utilizaría una variable explicativa categórica.

¿Hasta qué punto el encontrarme en un mundo globalizado, estoy expuesto a recibir ataques internacionales?

En esta prueba, primero de todo, sería interesante analizar si existen diferencias significativas entre la media de ataques producidos a nivel internacional, y los ataques que han tenido como objetivo un único continente. Para ello, el método de análisis utilizado sería un contraste de hipótesis sobre la media de ataques.

Como siguiente paso, a partir de los resultados obtenidos, sería interesante obtener información acerca de si existen diferencias significativas entre continentes, para las distribuciones de ataques producidos a empresas del sector público. Para ello, aplicaríamos un método estadístico que permitiese determinar si existen diferencias significativas entre los distintos continentes, o si por el contrario, las distribuciones de ataques son similares.

¿Existen atacantes bien conocidos, con pautas concretas, a los que me encuentre particularmente expuesto?

Para responder esta cuestión, sería interesante realizar, sobre una muestra de datos con observaciones del sector público, un contraste de hipótesis sobre la media de atacantes desconocidos y conocidos. En caso de obtener resultados que nos diesen diferencias significativas, el siguiente paso sería analizar cuál de los grupos es el más peligroso para adoptar una estrategia de defensa concreta. Por otra parte, en caso de no obtener diferencias significativas, sería interesante considerar otras variables como la tipología de ataque, y ser capaces de identificar correlaciones entre el tipo de atacante y la metodología empleada.

4.2 Análisis particulares, y comprobaciones de normalidad y distribución de la varianza en las variables bajo estudio

A continuación, vamos a presentar distintos análisis destinados a responder algunas de las cuestiones iniciales que motivaron a realizar el proyecto. Para ello, primero de todo, realizaremos el planteamiento de la cuestión a resolver y, acto seguido, se expondrán los datos seleccionados para realizar dicha tarea. A continuación, se comprobará si se cumplen las suposiciones de normalidad y homocedasticidad para los datos seleccionados y, finalmente, se aplicarán pruebas estadísticas que nos permitirán resolver la cuestión inicial.

4.2.1 Análisis temporal de los ciberataques.

El objeto de análisis de este subapartado, estará destinado a resolver el siguiente planteamiento: *‘¿Existe algún periodo del año en que debería ampliar el presupuesto y los medios necesarios, al estar más expuesto a recibir ciberataques?’*

Para responder a esta pregunta, podrían plantearse múltiples técnicas estadísticas. Por ejemplo, al entrar en juego la dimensión temporal a través de las variables *Year* y *Mes*, podrían plantearse técnicas basadas en el análisis de series temporales, tal como se muestra en [6], aun así, debido al alto grado de complejidad presente, trataremos de simplificar la tarea y realizar análisis más sencillos.

En concreto, consideraremos que la variable *Mes* constituye una variable categórica, que segmentará el set de datos en 12 muestras distintas, una para cada uno de los meses del año. Adicionalmente, nos interesará obtener conocimiento acerca del sector al que pertenecemos. Para ello, seleccionaremos exclusivamente aquellas observaciones del set de datos pertenecientes a la administración pública en tareas de defensa y administración. Finalmente, la variable de análisis que nos permitirá medir el riesgo al cual estamos expuestos, será el campo *TotalAtaques*, el cual representa el número total de incidentes de seguridad reportados

De esta forma, a partir de las observaciones pertenecientes al Sector Público, fragmentaremos el set de datos en 12 submuestras, sobre las que analizaremos si existen diferencias estadísticamente significativas realizando contrastes sobre la media de ataques recibidos.

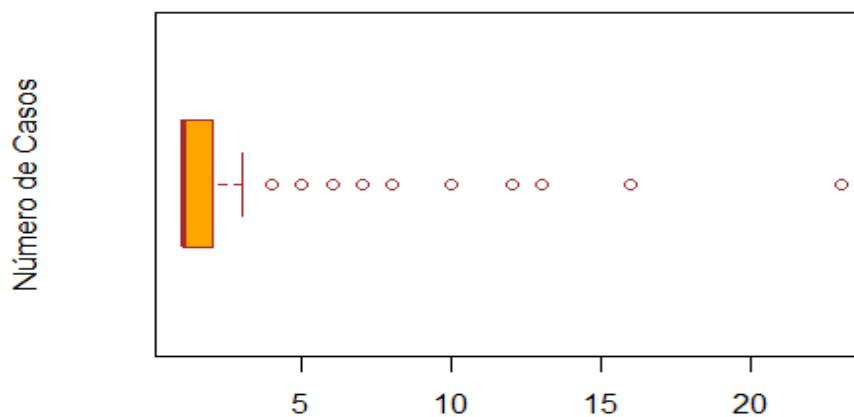
Para realizar los contrastes, escogeremos como mes de referencia aquel con mayor media de ataques, y lo contrastaremos con cada uno de los meses restantes para detectar diferencias significativas.

A continuación, se muestra cómo seleccionaremos el set de datos con el que realizaremos este análisis.

```
# Set de datos necesario para el análisis de impacto temporal
attacks_temp <- attacks_Input[attacks_Input$Code_target_class=="0",c('Year', 'Mes', 'NumeroAtaques')]
monthlyAttacks <- attacks_temp$NumeroAtaques

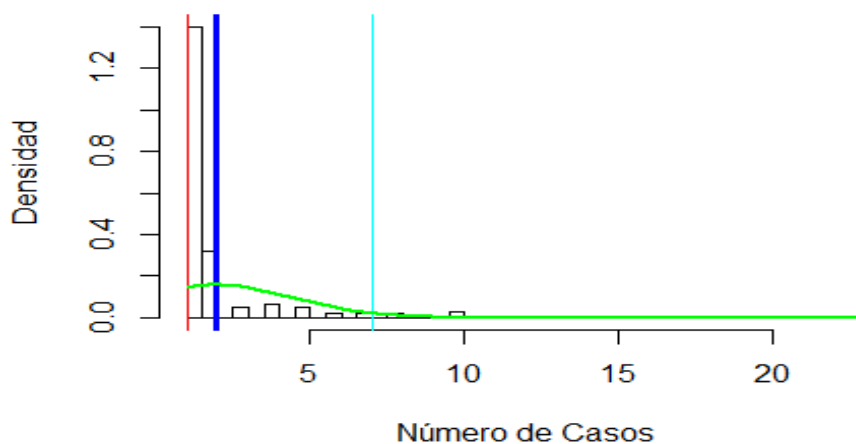
# Graficamos para la variable seleccionada un diagrama de boxplot, y la distribución de frecuencias de los distintos valores.
boxplot(monthlyAttacks,
        main="Ataques Mensuales para Sector Público",
        ylab="Número de Casos",
        col="orange",
        border="brown",
        horizontal=TRUE)
```

Ataques Mensuales para Sector Público



```
hist(monthlyAttacks, freq=FALSE, breaks=50, main="Ataques Mensuales para Sector Público", xlab="Número de Casos", ylab="Densidad")
abline(v=mean(monthlyAttacks), col="blue", lwd=3)
abline(v=median(monthlyAttacks), col="red")
abline(v=mean(monthlyAttacks)-2*sd(monthlyAttacks), col="cyan")
abline(v=mean(monthlyAttacks)+2*sd(monthlyAttacks), col="cyan")
curve(dnorm(x, mean=mean(monthlyAttacks), sd=sd(monthlyAttacks)), add=TRUE, col="green", lwd=2)
```

Ataques Mensuales para Sector Público

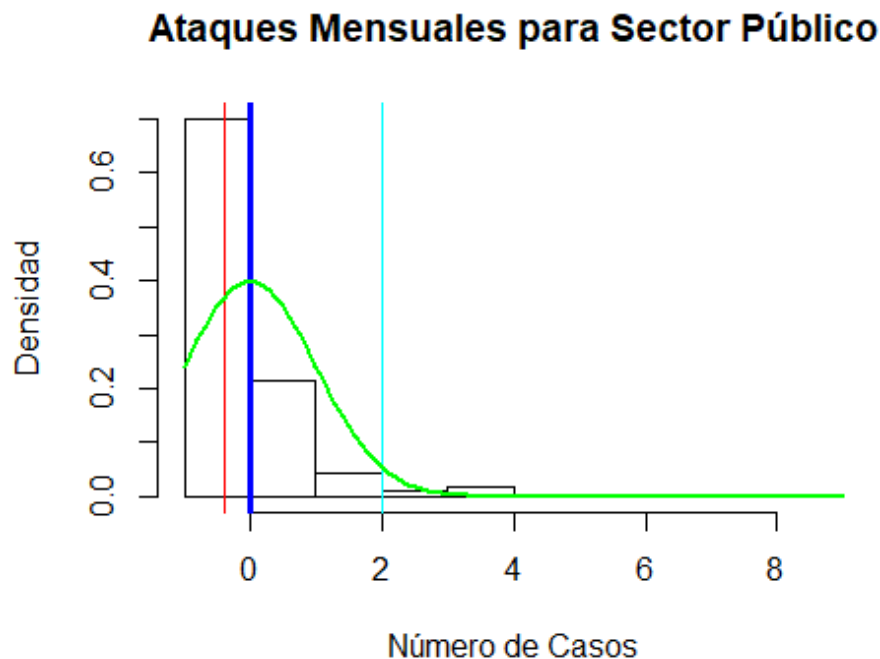


A partir de los gráficos anteriores, se puede observar como la distribución de ataques muestra una cola a derechas. Este resultado, induce a intuir que no se cumple la asunción de normalidad para la variable *NumeroAtaques*. Ante este resultado, vamos a intentar tipificar [7] esta variable.

Tipificación de La variable Número de Ataques

```
monthlyAttacks_N <- scale(monthlyAttacks, center = T, scale = T)
```

```
hist(monthlyAttacks_N, freq=FALSE, main="Ataques Mensuales para Sector Público", xlab="Número de Casos", ylab="Densidad")
abline(v=mean(monthlyAttacks_N), col="blue", lwd=3)
abline(v=median(monthlyAttacks_N), col="red")
abline(v=mean(monthlyAttacks_N)-2*sd(monthlyAttacks_N), col="cyan")
abline(v=mean(monthlyAttacks_N)+2*sd(monthlyAttacks_N), col="cyan")
curve(dnorm(x, mean=mean(monthlyAttacks_N), sd=sd(monthlyAttacks_N)), add=TRUE, col="green", lwd=2)
```



Después de realizar la tipificación de la variable, observamos una ligera mejoría en el aspecto visual, en cuanto a la asunción de normalidad, sin embargo, la calidad del ajuste sigue sin ser del todo buena. Vamos a analizar qué nos dice un test de normalidad sobre dicha variable.

Test de normalidad para La variable tipificada que hemos creado, para el análisis del número de casos mensuales.

Carga de librerías para tests de normalidad

```
library(normtest)
library(nortest)
```



```

library(moments)

# Varios tests de normalidad, todos ellos rechazando la hipótesis nula de
normalidad

sf.test(monthlyAttacks_N) # Shapiro-Francia

##
##  Shapiro-Francia normality test
##
## data:  monthlyAttacks_N
## W = 0.41688, p-value < 2.2e-16

jb.norm.test(monthlyAttacks_N) # Jarque-Bera

##
##  Jarque-Bera test for normality
##
## data:  monthlyAttacks_N
## JB = 7713.7, p-value < 2.2e-16

agostino.test(monthlyAttacks_N) # Agostino

##
##  D'Agostino skewness test
##
## data:  monthlyAttacks_N
## skew = 4.6251, z = 12.9676, p-value < 2.2e-16
## alternative hypothesis: data have a skewness

```

A partir de los distintos test, observamos que la variable que tratamos de analizar no cumple con la hipótesis de normalidad. Sin embargo, teniendo en cuenta que “*si una muestra es lo bastante grande ($n > 30$), sea cual sea la distribución de la variable de interés, la distribución de la media muestral será aproximadamente una normal. Además, la media será la misma que la de la variable de interés, y la desviación típica de la media muestral será aproximadamente el error estándar.*” (Rovira, 2009) [8], podremos asumir que **la media de variable NumeroAtaques sigue una distribución normal**, dado que tenemos un tamaño de muestra grande (superior a 30) y el contraste de hipótesis se realizará sobre la media.

En relación a la homocedasticidad, según se indica en [1], podemos aplicar el test de *Fligner-Killeen* cuando la distribución de los datos no cumple con la condición de normalidad.

```

# test de homocedasticidad para muestras que no cumplen la condición de n
ormalidad: Fligner-Killeen

fligner.test(NumeroAtaques ~ Mes, data = attacks_temp)

##
##  Fligner-Killeen test of homogeneity of variances

```

```
##
## data: NumeroAtaques by Mes
## Fligner-Killeen:med chi-squared = 10.518, df = 11, p-value =
## 0.4845
```

En este caso, a través del test, podemos afirmar que no existe una diferencia significativa entre las varianzas de los ataques a lo largo de los distintos grupos de meses.

En consecuencia, podremos considerar la asunción de normalidad de la media a través del Teorema del Límite Central; y la homocedasticidad a través del test *Fligner-Killeen*. De esta forma, podremos utilizar técnicas paramétricas para realizar contrastes de hipótesis sobre la media. En concreto, para esta sección, utilizaremos el *t-test* ya que *"el t-test sigue siendo suficientemente robusto, aunque un test no paramétrico basado en la mediana (Mann-Whitney-Wilcoxon) o un test de bootstrapping podrían ser más adecuados."* (Amat, 2007) [9].

A continuación, realizaremos el contraste de hipótesis a partir del *t-test*, tomando como mes de referencia aquel con mayor media de ataques.

```
# Mes con máximo valor medio de ataques

attacks_meanByMonth <- attacks_temp %>% group_by(Mes) %>% summarize(mean(
NumeroAtaques))

attacks_meanByMonth[order(-attacks_meanByMonth$`mean(NumeroAtaques)`),]

## # A tibble: 12 x 2
##   Mes   `mean(NumeroAtaques)`
##   <fct>          <dbl>
## 1 1             2.59
## 2 12            2.52
## 3 6             2.5
## 4 7             2.2
## 5 10            2.11
## 6 2             1.94
## 7 8             1.88
## 8 4             1.73
## 9 5             1.73
## 10 11           1.58
## 11 3            1.54
## 12 9            1.38
```

A partir de la tabla anterior, observamos que enero es el mes con mayor media de ataques, por lo que vamos a contrastar si su media es estadísticamente diferente a la del resto de meses. En este caso, utilizaremos una hipótesis alternativa unilateral, en la cual plantearemos que la media del resto de meses es inferior al de referencia.

$$\begin{cases} H_0: \mu_{\text{Enero}} - \mu_{\text{meses}} = 0 \\ H_1: \mu_{\text{Enero}} - \mu_{\text{meses}} > 0 \end{cases} \text{ con meses} \in (\text{Febrero} \dots \text{Diciembre})$$

Contraste de hipótesis sobre igualdad en la media del número de ataques entre enero y el resto de meses, de manera iterativa.

```
for(i in seq(2,12)){
  a <- t.test(attacks_temp[attacks_temp$Mes==1,3], attacks_temp[attacks_t
emp$Mes==i,3], alternative = "less")
  cat("p-valor para el contraste entre el mes 1 y el mes",i,":", a$p.valu
e, "\n")
}

## p-valor para el contraste entre el mes 1 y el mes 2 : 0.7193594
## p-valor para el contraste entre el mes 1 y el mes 3 : 0.8444329
## p-valor para el contraste entre el mes 1 y el mes 4 : 0.7898819
## p-valor para el contraste entre el mes 1 y el mes 5 : 0.7785468
## p-valor para el contraste entre el mes 1 y el mes 6 : 0.5286045
## p-valor para el contraste entre el mes 1 y el mes 7 : 0.6124608
## p-valor para el contraste entre el mes 1 y el mes 8 : 0.7374304
## p-valor para el contraste entre el mes 1 y el mes 9 : 0.877013
## p-valor para el contraste entre el mes 1 y el mes 10 : 0.6631326
## p-valor para el contraste entre el mes 1 y el mes 11 : 0.8297838
## p-valor para el contraste entre el mes 1 y el mes 12 : 0.5206724
```

Como se puede apreciar a través de los *p-values* obtenidos, no se puede rechazar la hipótesis nula para ninguna pareja de meses, suponiendo de esta forma, que no existen diferencias significativas y concluyendo que el número de ataques se mantiene relativamente estable a lo largo del año.

Por otra parte, la tendencia de los ataques, es otro asunto relacionado con la variable temporal que resultaría interesante de analizar. A través de este análisis, se podría concluir si se prevé que el número de ataques aumente, disminuya o se mantenga constante.

En este caso, aunque la metodología más ortodoxa sea a través del análisis de series temporales, vamos a aplicar un modelo de regresión lineal que nos permitirá analizar la tendencia de los ataques.

Primero de todo, vamos a “ordenar” de alguna manera el número de ataques, creando una variable que recoja la secuencialidad de las observaciones a partir del transcurso del tiempo, agregando los casos por mes y año, y asignando a esta variable “*t*” un valor secuencial.

*# Agregación por Mes y Año, y creación de una variable temporal secuencia
L “t”*

```
attacks_sumByMonth <- attacks_temp %>% group_by(Year, Mes) %>% summarize(
sumAtaques = sum(NúmeroAtaques))
attacks_sumByMonth <- as.data.frame(attacks_sumByMonth)
attacks_sumByMonth$t <- seq(1,dim(attacks_sumByMonth)[1])
```

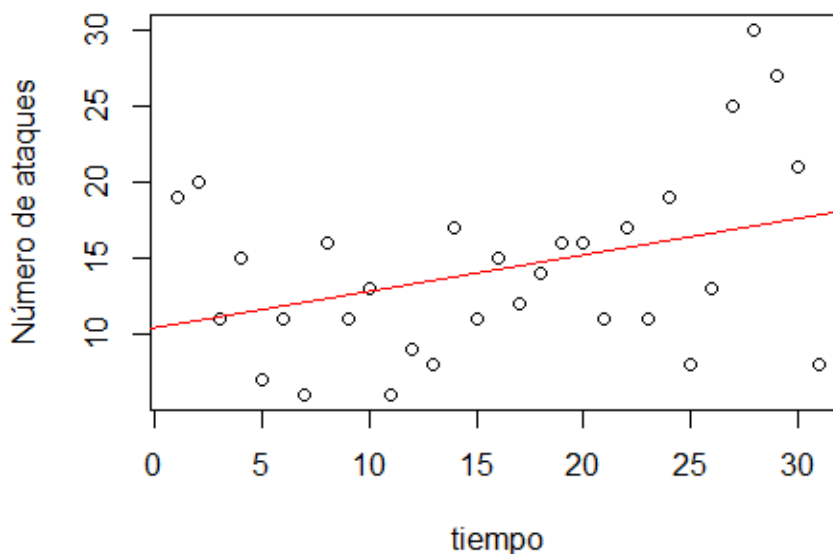
```
# Regresión del número de ataques sobre el paso del tiempo (variable secuencial "t")
reg_attacks_temp <- lm(sumAtaques ~ t, data = attacks_sumByMonth)

# Mostramos el resumen de la regresión, y la representación gráfica de la recta de regresión sobre el diagrama de dispersión de ambas variables
summary(reg_attacks_temp)

##
## Call:
## lm(formula = sumAtaques ~ t, data = attacks_sumByMonth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.8770 -4.4099 -0.1819  3.2750 12.8403
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.4645     2.0940   4.997 2.56e-05 ***
## t              0.2391     0.1142   2.093  0.0452 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.689 on 29 degrees of freedom
## Multiple R-squared:  0.1312, Adjusted R-squared:  0.1013
## F-statistic: 4.381 on 1 and 29 DF,  p-value: 0.0452

plot(x=attacks_sumByMonth$t, y=attacks_sumByMonth$sumAtaques, main="Regresion lineal NumeroAtaques ~ t. Sector Público", xlab='tiempo', ylab='Número de ataques')
abline(reg_attacks_temp, col=2)
```

Regresion lineal NumeroAtaques ~ t. Sector Público



A partir del modelo de regresión lineal, parece que la tendencia en el número de ataques que esperamos recibir, será creciente. De hecho, si consideramos que el factor tiempo tiene una granularidad mensual, el parámetro 0.23 de la pendiente, indica que en cada mes se espera un crecimiento del número medio de casos superior al 20%. Adicionalmente, esta pendiente es estadísticamente significativa (*p-value* de 0.045), para un nivel de significancia del 0.05.

Es cierto que el número medio de incidentes mensuales sigue siendo muy bajo, por lo que el crecimiento en valor absoluto no es muy alto. Sin embargo, este hecho, debe darnos una idea sobre la previsión de inversión futura, en la que cada vez, será necesaria una mayor dedicación y esfuerzo para atajar estos problemas.

Por otra parte, el valor del coeficiente de autodeterminación R^2 es bastante bajo, mientras que el valor de significatividad global del modelo (observado a través del estadístico F), no rechaza la hipótesis nula, suponiendo un buen ajuste global del modelo a los datos. Esta información aparentemente contradictoria, puede significar que el modelo utilizado es excesivamente simple considerando únicamente la variable tiempo. De esta forma, gran parte de la varianza de los ataques, seguramente estará determinada por otras variables que no están presentes en el análisis. Además, puede que parte del buen ajuste del modelo, se produzca por el hecho de que se han descartado diferencias estadísticamente significativas entre los ataques producidos entre meses.

Finalmente, teniendo en cuenta que lo que queríamos era una primera aproximación sobre la evolución del número de ataques en nuestro sector, parece razonable afirmar que se producirá un incremento.

4.2.2 Análisis territorial de los ciberataques.

El objeto de análisis de este subapartado, estará destinado a resolver el siguiente planteamiento: *¿Hasta qué punto el encontrarme en un mundo globalizado, estoy expuesto a recibir ataques internacionales?*

Con este planteamiento, pretendemos conocer hasta qué punto nuestro sector está expuesto a recibir ataques internacionales, y determinar si existen diferencias significativas entre los ataques internacionales y los realizados exclusivamente a un continente. Por consiguiente, realizaremos un contraste de hipótesis sobre la diferencia en la medida de tendencia central entre los ataques internacionales y continentales.

Para realizar este análisis, agregaremos los datos a nivel de continente, utilizando de esta forma, la granularidad Año – Mes – Continente. Adicionalmente, descartaremos aquellas observaciones que presenten valores perdidos, identificadas mediante la etiqueta '*Desconocido*'. Por último, seleccionaremos exclusivamente aquellas observaciones pertenecientes a nuestro sector, es decir, aquellas que presenten el valor 0 para el atributo *Code_target_class*.

```
# Selección de las variables de interés y agrupación.
# Se seleccionan las empresas de nuestro sector
# Las columnas de análisis serán año, mes y continente agregando el número de ataques totales
df_analysis <- df[(df$Code_target_class == 'O') & (df$Continent != 'Desconocido'),] %>%
  select(Year, Mes, Continent, NumeroAtaques) %>%
  group_by(Year, Mes, Continent) %>%
  summarize(NumeroAtaques=sum(NumeroAtaques)) %>%
  ungroup()

# Creamos una variable que servirá para segregar entre ataques internacionales o continentales
df_analysis$Attack_Kind <- ifelse(df_analysis$Continent == 'International', yes: 'International', no: 'SingleContinent')
```

Una vez adecuados los datos, comprobaremos si la distribución de valores para la variable *NumeroAtaques* sigue una distribución normal. Para ello, se realizará el test *Shapiro-Wilk*. Este test, supone como hipótesis nula que la distribución analizada se distribuye de forma normal. En caso de obtener un *p-value* superior al nivel de significación marcado, daremos dicha suposición como válida.

```
```{r}
shapiro.test(df_analysis$NumeroAtaques)
```
```

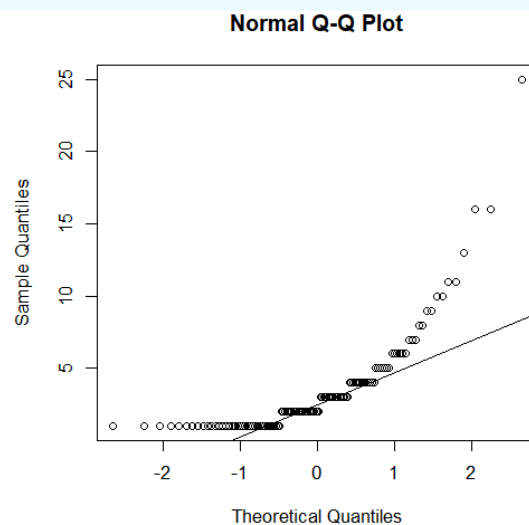
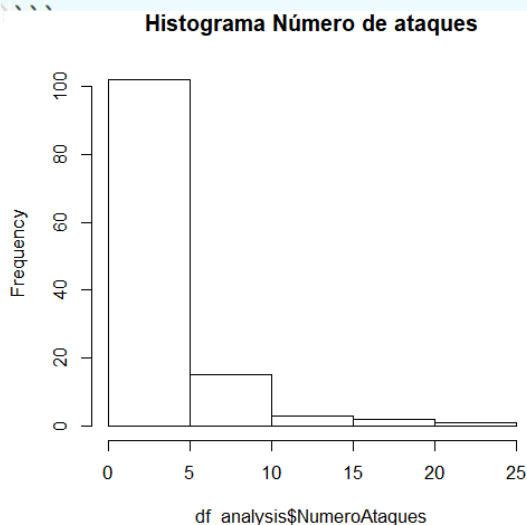
```
Shapiro-Wilk normality test

data:  df_analysis$NumeroAtaques
W = 0.68887, p-value = 8.438e-15
```

Como se puede observar, el *p-value* es inferior a un nivel de significación de 0.05, rechazando la hipótesis nula en favor de la hipótesis alternativa. En este caso, no podemos aceptar la suposición de normalidad.

Adicionalmente, si realizamos el histograma y el gráfico *qq-plot* para la variable *NumeroAtaques*, podemos observar cómo su distribución no se ajusta a una normal.

```
```{r}
par(mfrow=c(1,2))
hist(df_analysis$NumeroAtaques, main='Histograma Número de ataques')
qqnorm(df_analysis$NumeroAtaques)
qqline(df_analysis$NumeroAtaques)
```
```



Ante tal situación, con el objetivo de mejorar la normalidad de la variable y poder aplicar pruebas estadísticas paramétricas, aplicaremos una transformación *Box-Cox*.

```
```{r}
library(package= 'DescTools')

Realizamos la normalización del número de ataques
df_analysis$NumeroAtaques_Norm <- BoxCox(df_analysis$NumeroAtaques, lambda=BoxCoxLambda(df_analysis$NumeroAtaques))
```
```

Una vez realizada la transformación, comprobaremos de nuevo a través del test *Shapiro-wilk*, si la variable sigue una distribución normal o no.

```
# Volvemos a aplicar el test
shapiro.test(df_analysis$NumeroAtaques_Norm)
```
```

```
Shapiro-Wilk normality test

data: df_analysis$NumeroAtaques_Norm
W = 0.89671, p-value = 1.003e-07
```

Después de la transformación, observamos que el *p-value*, ha incrementado considerablemente respecto la primera prueba, aun así, sigue siendo inferior a un nivel de significación de 0.05, por lo que deberemos rechazar la suposición de normalidad.

Adicionalmente, aunque el número de observaciones es suficiente para aplicar el teorema del límite central y suponer que la media de la muestra sigue una distribución normal, en este caso, optaremos por adoptar una posición más conservadora y realizar pruebas no paramétricas para realizar los siguientes análisis.

Por consiguiente, para comprobar que las varianzas son constantes e iguales, se aplicará el test de *Fligner-Killeen* sobre los datos originales. Este test, considera como hipótesis nula la homocedasticidad entre muestras, de forma que, en caso de obtener un *p-value* superior al nivel de significación fijado, fallaremos a favor de dicha hipótesis. Por el contrario, las muestras presentarán heterocedasticidad.

```
```{r}
fligner.test(NumeroAtaques ~ Attack_Kind, data=df_analysis)
```

Fligner-Killeen test of homogeneity of variances

data: NumeroAtaques by Attack_Kind
Fligner-Killeen:med chi-squared = 14.119, df = 1, p-value = 0.0001716
```

Tal como puede observarse, el *p-value* obtenido es inferior a un nivel de significación de 0.05, por lo que nuestros datos no presentan varianzas constantes.

A través de los análisis realizados, podemos determinar que la variable *NumeroAtaques*, no sigue una distribución normal, ni presenta homocedasticidad. En consecuencia, para comprobar el contraste de medidas de tendencia central utilizaremos la prueba *Man-*



*Whitney* para muestras independientes. En esta prueba se realiza una comparación de medianas, siendo esta, una medida de tendencia central más robusta que la media ante la presencia de *outliers* y la no asunción de normalidad de la muestra.

En este caso, realizaremos un contraste bilateral, con el cual definiremos como hipótesis alternativa que los ataques internacionales  $\mu_I$  son distintos de los que se producen exclusivamente en un continente en concreto  $\mu_C$ .

$$\begin{cases} H_0: \mu_I - \mu_C = 0 \\ H_1: \mu_I - \mu_C \neq 0 \end{cases}$$

```
```{r}
wilcox.test(NúmeroAtaques ~ Attack_Kind, data=df_analysis)
```

Wilcoxon rank sum test with continuity correction

data: NúmeroAtaques by Attack_Kind
W = 721, p-value = 0.002325
alternative hypothesis: true location shift is not equal to 0
```

A través del test, se ha obtenido un *p-value* inferior a un nivel de significación de 0.05, en consecuencia, existen diferencias significativas entre las medianas de ambas distribuciones.

Ante este resultado, observaremos cuál de las dos medias es superior, con el objetivo de saber si es necesario analizar los ataques producidos en exclusividad en nuestro continente, o realizar un análisis a nivel mundial.

```
```{r}
paste('Media de ataques internacionales', round(mean(df_analysis[df_analysis$Attack_Kind == 'International',]$NúmeroAtaques), digits= 0))
paste('Media de ataques a continentes', round(mean(df_analysis[df_analysis$Attack_Kind == 'SingleContinent',]$NúmeroAtaques), digits= 0))
```

[1] "Media de ataques internacionales 2"
[1] "Media de ataques a continentes 4"
```

A partir de los resultados obtenidos, se puede observar que el número de ataques exclusivos a continentes es superior al número de ataques internacionales.

Ante esta situación, será interesante analizar si los ataques se distribuyen de la misma forma en todos los continentes, es decir, si presentan las mismas distribuciones a lo largo de la variable *Continent*. Esta información, será de utilidad a la hora de mejorar nuestros sistemas de seguridad, ya que, si los ataques son similares, deberíamos de recoger datos acerca de los ciberataques y de los sistemas de defensa de empresas del sector público situadas en otros continentes. En caso contrario, deberíamos centrarnos en exclusividad a estudiar las empresas y los ataques, de nuestro continente.

Teniendo en cuenta los resultados anteriores, en los que se mostraba como la variable *NúmeroAtaques* no presentaba una distribución normal ni homocedasticidad, realizaremos el análisis propuesto utilizando un método no paramétrico que nos permita realizar comparaciones entre más de dos grupos. De esta forma, aplicaremos



el test de *Kruskal – Wallis*, el cual supone como hipótesis nula, la no existencia de diferencias significativas entre los ataques de continentes.

```
```{r}
# Selección de las variables de interés y agrupación.
# Se seleccionan las empresas de nuestro sector
# Las columnas de análisis serán año, mes y continente agregando el número de ataques totales
df_analysis_continent <- df[(df$Code_target_class == 'O') & (df$Continent != 'Desconocido') & (df$Continent != 'International'),] %>%
  select(Year, Mes, Continent, NumeroAtaques) %>%
  group_by(Year, Mes, Continent) %>%
  summarize(NumeroAtaques=sum(NumeroAtaques)) %>%
  ungroup()

kruskal.test(NumeroAtaques ~ Continent, data=df_analysis_continent)
```
```

```
Kruskal-Wallis rank sum test

data: NumeroAtaques by Continent
Kruskal-Wallis chi-squared = 53.64, df = 4, p-value = 6.259e-11
```

El *p-value* obtenido es inferior a un nivel de significación de 0.05, por lo que los ataques se distribuyen de forma diferente a través de los distintos continentes. En consecuencia, deberemos analizar aquellos ataques producidos exclusivamente en Europa, con el objetivo de mejorar nuestros sistemas de seguridad.

#### 4.2.3 Análisis de la tipología de ataques.

El objeto de análisis de este subapartado, estará destinado a resolver el siguiente planteamiento: *¿A qué tipología de ciberataque mi compañía está más expuesta? ¿Debería por ello, analizar la documentación cualitativa, así como los detalles técnicos referentes a este tipo de ataque?*

En este contexto, para ser capaces de identificar a qué tipología de ataques están más expuestas las empresas del sector público, realizaremos un modelo de regresión logística.

Para realizar este modelo, crearemos una variable dicotómica que nos indicará si el ataque afecta a una empresa del sector público o no. Esta variable, recibirá el nombre de *Attacked* y constituirá el termino dependiente del modelo.

Por otra parte, para crear la variable dependiente, utilizaremos los campos *Code\_attack\_class\_CW*, *Code\_attack\_class\_CE*, *Code\_attack\_class\_CC* y *Code\_attack\_class\_H*. En este sentido, el objetivo será constituir una única variable categórica con la información referente a la tipología de ataques. Esta variable, recibirá el nombre de *CodigoAtaque* y constituirá el termino independiente del modelo.

Una vez construido el modelo, analizaremos los coeficientes estimados para cada una de las posibles categorías presentes en la variable explicativa, para determinar cómo aumentan o disminuyen los *ODDS* y, en consecuencia, las posibilidades de que sea atacada una empresa del sector público, en función de la tipología de ataque.

Adicionalmente, antes de generar el modelo de regresión logística, únicamente nos quedaremos con aquellas observaciones de ataques producidas en Europa. Esta decisión, ha sido tomada a través del análisis anterior, en el cual se determinaba que la distribución de ataques era distinta en función del continente. De esta forma, al ser una empresa europea, nos interesará analizar cómo afectan las distintas tipologías de ataque exclusivamente en nuestro continente. Por otra parte, se descartarán aquellas muestras clasificadas como *outliers* en el apartado 3.2. En este contexto, no se considerarán aquellas observaciones cuya tipología de empresa sea *X*, *Y* o *Z*.

A continuación, se muestra el código de R utilizado para realizar el procesado de datos.

```
Procesado de datos. Se descartan las tipologías de empresa X, Y, Z y nos quedamos con la información de europa.
Adicionalmente, seleccionamos los datos de las distintas tipologías de ataque juntamente con el tipo de empresa.
Por último, a través de melt concatenamos los valores de las tipologías de ataque en una única columna.
df_analysis <- df[!(df$Code_target_class %in% c('Z','Y','X')) & (df$Continent == 'Europa'),] %>%
 select(Code_target_class, Code_attack_class_CW, Code_attack_class_CE, Code_attack_class_CC, Code_attack_class_H) %>%
 melt(value.name = 'NumeroAtaques', variable.name = 'CodigoAtaque')

Eliminamos aquellos registros con 0 números de ataque, ya que simbolizará que no se ha producido dicha categoría de ataque
df_analysis_filtered <- df_analysis[df_analysis$NumeroAtaques != 0,]

En función del número de ataques, repetiremos los registros n veces.
df_analysis_final <- as.data.frame(lapply(df_analysis_filtered, rep, df_analysis_filtered$NumeroAtaques))

Codificación de la variable Attacked.
df_analysis_final$Attacked <- ifelse(df_analysis_final$Code_target_class == 'O', yes= 1, no= 0)
```

Una vez realizado el procesado de datos, ya podremos crear el modelo de regresión logística a partir del predictor *CodigoAtaque*. En este caso como la variable dependiente es categórica y presenta más de un nivel, deberemos asignar una de las tipologías de ataque como referencia. De esta forma, los coeficientes estimados estarán basados en función de este nivel de referencia. En nuestro caso, este nivel será la categoría *CC*, obteniendo el siguiente modelo logístico:

$$\text{logit}(ODDS) = \widehat{\beta}_0 + \widehat{\beta}_1 x_{CW} + \widehat{\beta}_2 x_{CE} + \widehat{\beta}_3 x_H$$

Con:

$$ODDS = \frac{P(Y = \text{Empresa Sector Público Europea Atacada} \mid X = x)}{1 - P(Y = \text{Empresa Sector Público Europea Atacada} \mid X = x)}$$

$$x_{CE} = \text{ataque ciber espionage}$$

$$x_{CW} = \text{ataque ciber warfare}$$

$$x_H = \text{ataque hacktivism}$$

A continuación, se muestra el código R utilizado para realizar el modelo logístico y asignar como nivel de referencia la categoría CC.

```
Asignamos CC como nivel de referencia
df_analysis_final$CodigoAtaque <- relevel(df_analysis_final$CodigoAtaque, ref='Code_attack_class_CC')

Creamos el modelo logico
log_model <- glm(formula=Attacked ~ CodigoAtaque, data=df_analysis_final, family=binomial)
```

Una vez se ha obtenido el modelo de regresión logística, lo analizaremos a través de *summary*.

```
summary(log_model)
'''

Using Code_target_class as id variables

Call:
glm(formula = Attacked ~ CodigoAtaque, family = binomial, data = df_analysis_final)

Deviance Residuals:
 Min 1Q Median 3Q Max
-1.2968 -0.4914 -0.4914 -0.4914 2.0852

Coefficients:
 Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.0532 0.1972 -10.410 < 2e-16 ***
CodigoAtaqueCode_attack_class_CW 1.8121 0.4486 4.039 5.36e-05 ***
CodigoAtaqueCode_attack_class_CE 2.3295 0.3447 6.757 1.41e-11 ***
CodigoAtaqueCode_attack_class_H 1.7247 0.3667 4.704 2.56e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

 Null deviance: 405.73 on 373 degrees of freedom
Residual deviance: 343.16 on 370 degrees of freedom
AIC: 351.16

Number of Fisher Scoring iterations: 4
```

A partir de los *p-values* asignados a los coeficientes estimados del modelo, observamos que todas las tipologías de ataque son significativas, mostrando valores inferiores a una significancia del 0.05.

Cabe recordar, que todos los coeficientes estimados del modelo, están referenciados en función de la tipología de ataque CC, de esta forma, al ser todos positivos, nos indican que el los *Ciber Crimes*, son la tipología de ataque a la que una empresa del sector público está menos expuesta. Adicionalmente, observando los valores *beta*, parece que la tipología de ataque CE (correspondiente a casos de ciber espionaje), es la que presenta una relación significativa mayor, con el hecho de que una empresa del sector público, sea atacada. En concreto, se produce un incremento del logaritmo de los *ODDS* del 2.33, suponiendo un incremento promedio de  $e^{2.33} \approx 10.28$  unidades con respecto

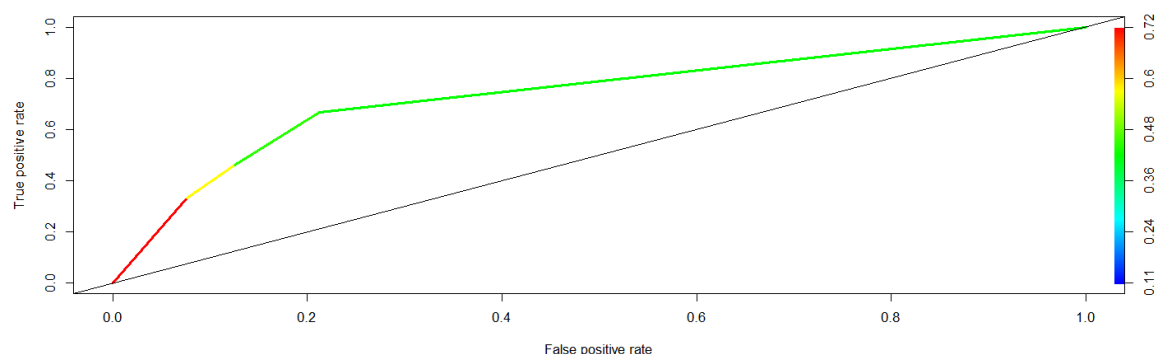
a un ataque CC. Por otra parte, los ataques del tipo *CW* (guerra cibernética) y *H* (actividades de hacktivismo), incrementan el logaritmo de los *ODDS* de forma similar.

Atendiendo a la realidad social y a las funciones del sector público, resulta razonable que los ataques de tipo espionaje (entre países), ataques desestabilizadores (guerra cibernética) y protestas de grupos sociales (hacktivismo) sean más significativos que aquellos más asociados con el fraude o los delitos económicos, objetivos aparentemente más atractivos para un *hacker* en el sector privado.

Finalmente, con el objetivo de evaluar la bondad de ajuste del modelo de regresión logístico obtenido, representaremos la curva ROC y mediremos el área contenida debajo de esta curva (AUROC).

Primero, realizaremos esta evaluación sobre el set de datos utilizado para obtener el modelo.

```
```{r}
library( package= 'ROCR')
predictions <- prediction(log_model$fitted, df_analysis_final$Attacked)
perf <- performance(predictions, measure= 'tpr', x.measure= 'fpr')
plot(perf, colorize=T, lwd=3)
abline(a=0, b=1)
```
```



```
```{r}
paste('AUROC de:', performance(predictions, measure="auc")@y.values[1])
```
```

```
[1] "AUROC de: 0.737394369017582"
```

A través de los resultados obtenidos, el valor del área bajo la curva es de 0.734, indicando la necesidad de encontrar otras variables que, juntamente con el tipo de ataque producido, puedan determinar si una empresa del sector público será el objetivo de ataque o no.

Aunque la bondad de ajuste para el set de datos no ha sido del todo buena, a continuación, realizaremos la evaluación del modelo a través de un set de datos de test.

Para obtener este set de datos de test se ha relanzado el *scraper* recopilando de esta forma, un reporte de datos nuevo publicado durante la realización de la práctica en la página web <https://www.hackmageddon.com/>. Esta nueva fuente de datos, puede encontrarse en la carpeta *data/00\_raw* con el nombre *scraping 2020-06-05 17.25.27.csv*.

Una vez obtenido el set de datos en crudo, se le ha aplicado el mismo procesado de datos, para adecuar la información a la contenida en la base de datos de entrenamiento. Este nuevo archivo, *datos\_test.cs*, puede encontrarse en la carpeta *data/01\_clean* del repositorio de Github.

Finalmente, una vez adecuados los datos del set de test, se obtendrán las predicciones con el objetivo de representar la curva ROC y medir el área contenida debajo de esta curva (AUROC).

```
```{r}
path_test <- 'https://raw.githubusercontent.com/iruiper/Cyberattacks-
History/master/data/01_clean/datos_test.csv'

# Obtención del dataset
df_test <- read.csv(file=path_test, encoding='latin-1')

# Procesado de datos. Se descartan las tipologías de empresa X, Y, Z y
nos quedamos con la información de europa.
# Adicionalmente, seleccionamos los datos de las distintas tipologías de
ataque juntamente con el tipo de empresa.
# Por último, a través de melt concatenamos los valores de las tipologías
de ataque en una única columna.
df_test_model <- df_test[!(df_test$Code_target_class %in% c('Z','Y','X'))
& (df_test$Continent == 'Europa'),] %>%
  select(Code_target_class, Code_attack_class_CW, Code_attack_class_CE,
Code_attack_class_CC, Code_attack_class_H) %>%
  melt(value.name = 'NumeroAtaques', variable.name = 'CodigoAtaque')

# Eliminamos aquellos registros con 0 números de ataque, ya que
simbolizará que no se ha producido dicha categoria de ataque
df_test_model_filtered <- df_test_model[df_test_model$NumeroAtaques !=
0,]

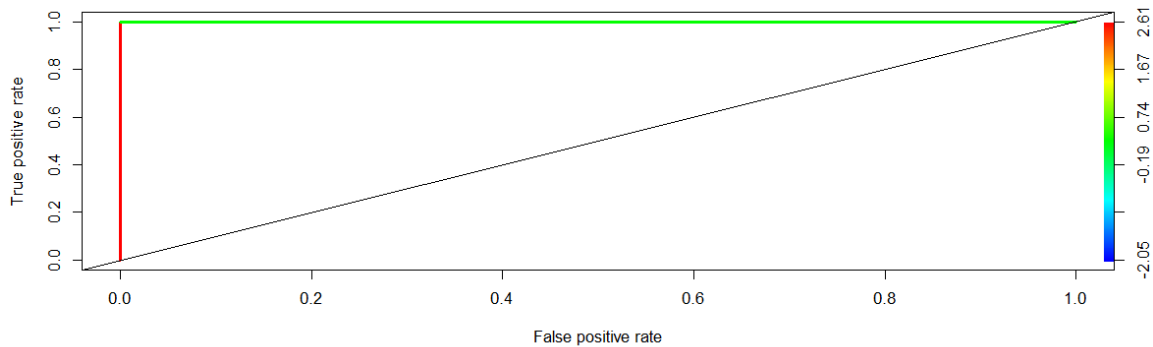
# En función del número de ataques, repetiremos los registros n veces.
df_test_final <- as.data.frame(lapply(df_test_model_filtered, rep,
df_test_model_filtered$NumeroAtaques))

# Codificación de la variable Attacked.
df_test_final$Attacked <- ifelse(df_test_final$Code_target_class == 'O',
1, 0)

# Asignamos CC como nivel de referencia
df_test_final$CodigoAtaque <- relevel(df_test_final$CodigoAtaque,
ref='Code_attack_class_CC')

# Creación del gráfico ROC
predictions_test <- prediction(predict(log_model, df_test_final),
df_test_final$Attacked)
test <- performance(predictions, 'tpr', 'fpr')
```

```
plot(predictions_test, colorize=T, lwd=3)
abline(a=0, b=1)
```



```
```{r}
paste('AUROC de:', performance(predictions_test, measure="auc")@y.values[1])
```
```

```
[1] "AUROC de: 1"
```

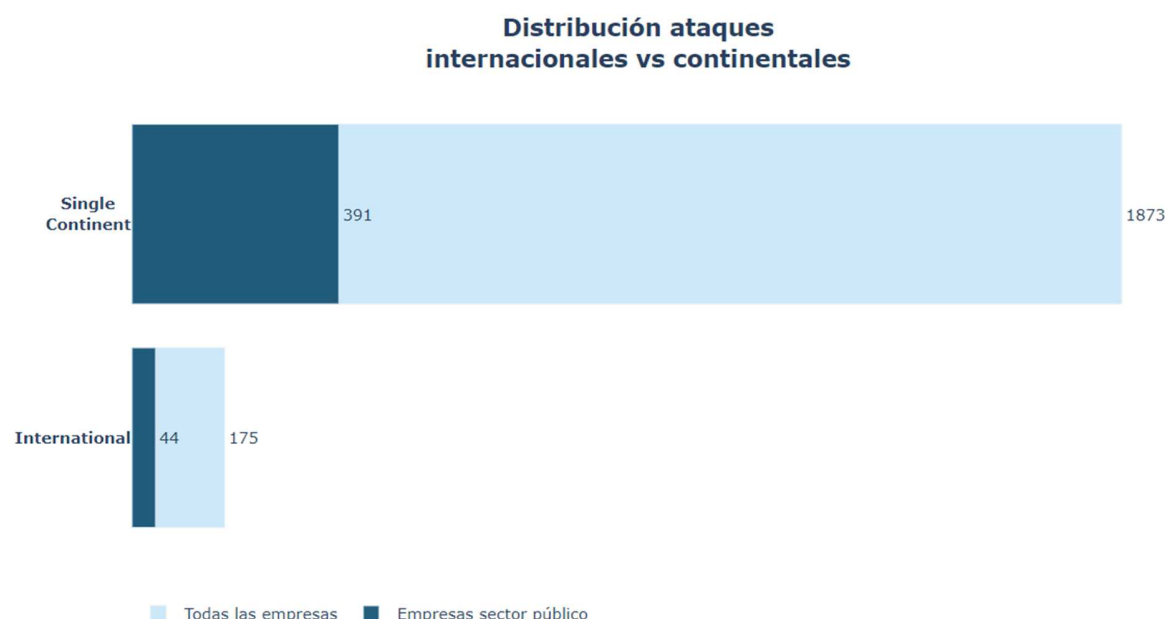
A través de los resultados obtenidos, el valor del área bajo la curva es de 1 consiguiendo predecir todas las muestras del set de test.

Este resultado confuso, en el que la evaluación del set de datos de entrenamiento es inferior a la evaluación del set de datos de test, es producido por dos causas. La primera de ellas proviene del bajo número de registros presente en la base de datos de test; por otra parte, la segunda causa puede originarse por un sub-entrenamiento del modelo, necesitando un mayor número de muestras y de iteraciones para lograr adecuar correctamente los parámetros beta estimados.

5. Representación de los resultados a partir de tablas y gráficas

Una vez finalizados los análisis de la sección 4, en el presente apartado, vamos a complementar los resultados obtenidos, a través del uso de gráficas y tablas que permitirán explicar con mayor detalle algunas de las conclusiones que posteriormente desarrollaremos en la sección 6. De esta forma, utilizaremos la potencia presente en las representaciones gráficas, con el objetivo de comprender mejor los resultados, y obtener así, más conocimiento del set de datos.

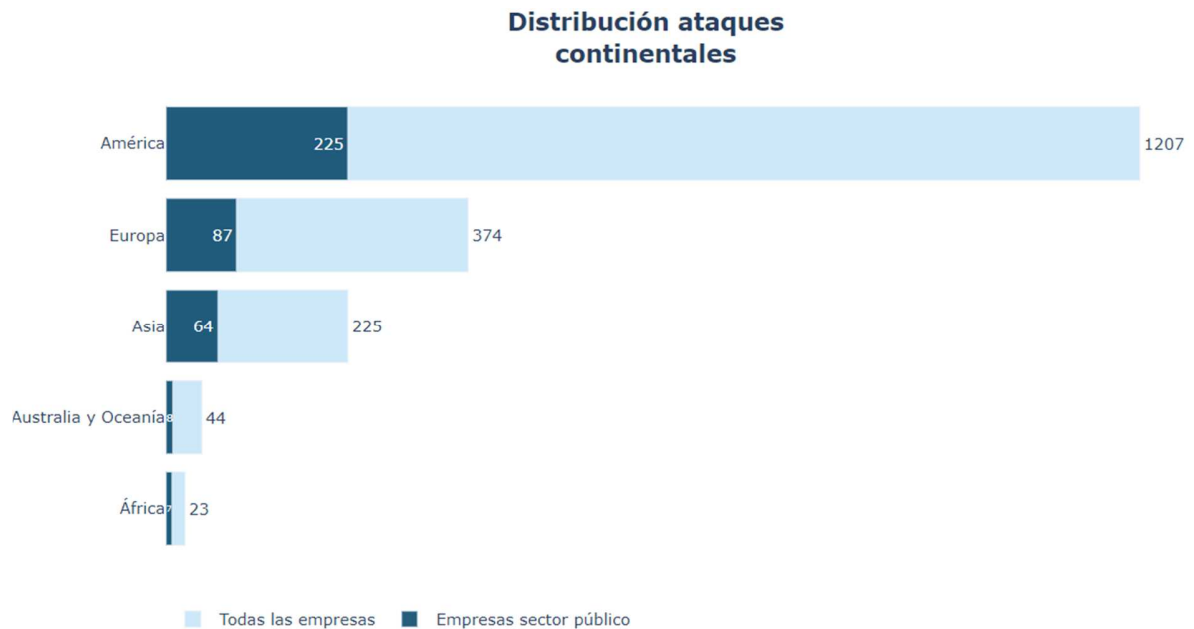
En primera instancia, a través de los análisis de la sección 4, se observó que existían diferencias significativas entre los ataques producidos a nivel internacional, y los que se producían exclusivamente en continentes concretos. Sin embargo, a través del análisis, no se realizó una cuantificación de los datos que nos permitiese obtener una idea acerca de la volumetría de eventos producidos para cada categoría. De esta forma, a continuación, se utilizará un gráfico de barras que nos permitirá comparar el rango de ataques producidos en cada una de las categorías anteriores.



Tal y como se puede observar en el gráfico, esta casuística no es exclusiva para las empresas del sector público, sino que parece reproducirse para el resto de sectores contenidos en el dataset.

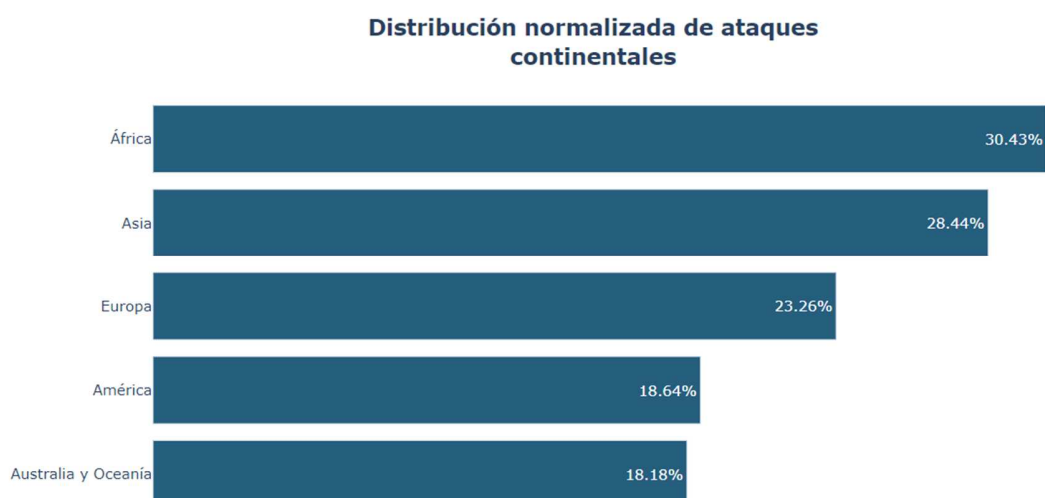
Además, parece ser que los ataques continentales son mucho más frecuentes que los internacionales, produciéndose aproximadamente entre 10 y 11 ataques continentales, por cada ataque internacional.

Ante el resultado anterior, se avanzó con el análisis estadístico para determinar si los ataques entre continentes se distribuían de forma similar a lo largo de todos los continentes. Por consiguiente, a continuación, se representará la volumetría de ataques producidos en cada continente.



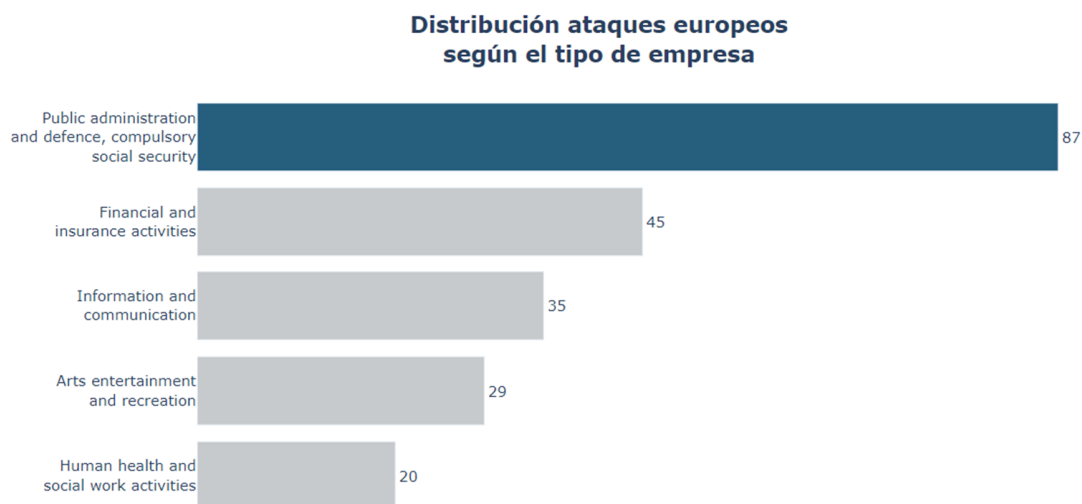
Tal como se puede observar, la volumetría de ataques es muy dispar, siendo América el continente que más ataques cibernéticos recibe. Por otra parte, África parece ser el continente con menos amenazas documentadas. En cuanto a Europa, continente al cual pertenece nuestra empresa, se sitúa en la posición número 2 con 87 ataques, teniendo muy cerca Asia con 64 ataques.

Por otra parte, si contextualizamos los ataques producidos a empresas del sector público, teniendo en cuenta el número total de ataques producidos en cada continente, los resultados cambian.



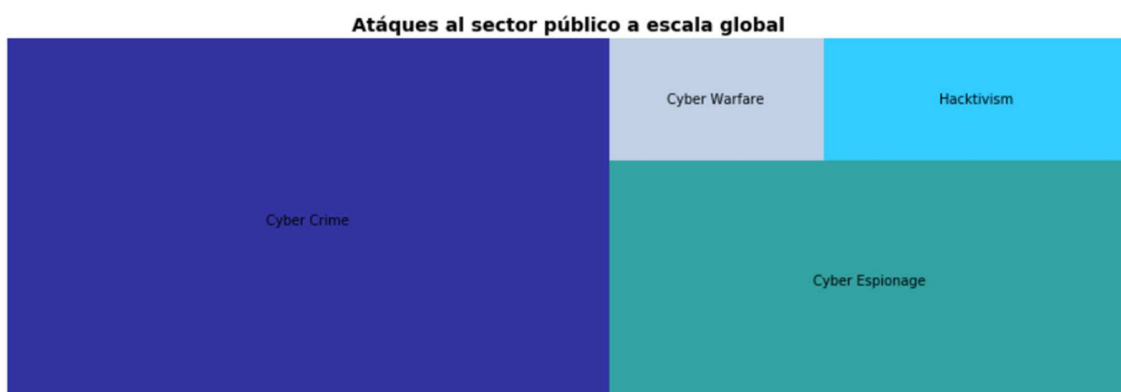
A través del gráfico normalizado, se han cambiado drásticamente las posiciones de los continentes. En este caso, África pasa a ser el continente con mayor ratio de ataques al sector público. En cambio, América, siendo el continente con mayor número de amenazas, presenta un ratio de ataques a empresas del sector público menor al 20%.

En cuanto a Europa, aproximadamente 1 de cada 4 ataques producidos van dirigidos a empresas del sector público. Ante este resultado, deberemos destinar partidas presupuestarias más grandes que las que deberían de destinar empresas pertenecientes a otros sectores. En este sentido, resultará interesante observar cuales son los sectores más atacados en Europa.



A través del gráfico de barras obtenido, en Europa, las empresas del sector público son las más atacadas, duplicando a las empresas del sector financiero.

Por otra parte, será interesante observar qué tipo de ciberataque es la más común, con el objetivo de adecuar correctamente nuestros sistemas de seguridad. Recordemos que en la sección 4, se detectó que los ciber espionajes eran las tipologías de ataque que producían un mayor aumento de los *ODDS* de recibir un ataque, sin embargo, vamos a visualizar como es la volumetría de ataques a través de *TreeMaps* [10]. Primero, se realizará una visualización a nivel Global que nos permita contextualizar los datos a escala mundial.



A través del *TreeMap* se puede observar que los *Ciber Crimes* son la tipología de ataque más frecuente a nivel global, mientras que los ataques de *Cyber Warfare*, son la tipología de ataque menos frecuente. En la siguiente tabla se cuantifican dichos ataques, segregándolos por continentes.

| | Cyber Crime | Cyber Espionage | Cyber Warfare | Hacktivism |
|----------------------------|-------------|-----------------|---------------|------------|
| Continent | | | | |
| América | 178.0 | 24.0 | 7.0 | 16.0 |
| Asia | 12.0 | 44.0 | 6.0 | 2.0 |
| Australia y Oceanía | 5.0 | 3.0 | 0.0 | 0.0 |
| Europa | 29.0 | 29.0 | 11.0 | 18.0 |
| International | 3.0 | 34.0 | 4.0 | 3.0 |
| África | 6.0 | 0.0 | 0.0 | 1.0 |
| Total | 233.0 | 134.0 | 28.0 | 40.0 |

En la tabla, se puede observar como los *Cyber Crimes* es la tipología de ataque más frecuente en empresas del sector público americanas, mientras que, en Europa esta tipología de ataque se encuentra en la misma medida que los *Cyber Espionages*.

Este resultado, puede llegar a ser confuso según el modelo de regresión logística obtenido en la sección 4 ya que, según los resultados de la tabla, los *Cyber Crimes* deberían de ser la tipología de ataque a la que una empresa del sector público esté más expuesta.

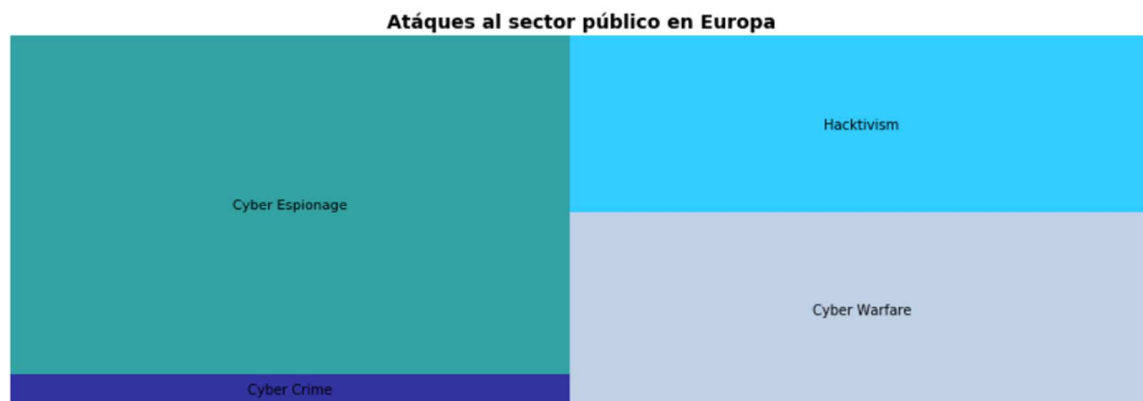
Para realizar una correcta interpretación, también se deberían tener en cuenta aquellas tipologías de ataque que afectan a otros sectores, ya que podría ser que los *Cyber Crimes* se produjesen con más frecuencia en otros sectores, y que la volumetría de ataques, se origine por otros aspectos como la facilidad de realizar el ataque.

| | Cyber Crime | Cyber Espionage | Cyber Warfare | Hacktivism |
|----------------------------|-------------|-----------------|---------------|------------|
| Continent | | | | |
| América | 956.0 | 13.0 | 4.0 | 9.0 |
| Asia | 124.0 | 16.0 | 10.0 | 10.0 |
| Australia y Oceanía | 33.0 | 3.0 | 0.0 | 0.0 |
| Europa | 226.0 | 22.0 | 14.0 | 25.0 |
| International | 105.0 | 21.0 | 3.0 | 2.0 |
| África | 15.0 | 0.0 | 1.0 | 0.0 |
| Total | 1459.0 | 75.0 | 32.0 | 46.0 |

Tal como se puede observar a través de la tabla anterior, los *Cyber Crimes* son la tipología de ataque más común entre empresas que no se encuentran dentro del sector público. De esta forma, a continuación, vamos a medir la proporción de ataques que afectan a empresas de nuestro sector.

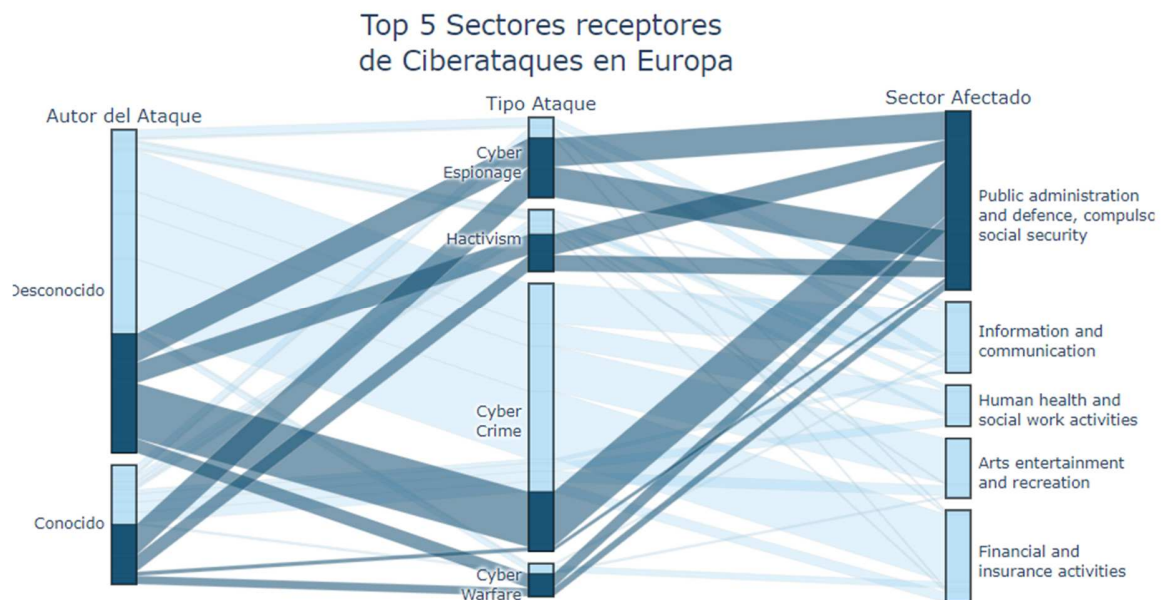
| | Cyber Crime | Cyber Espionage | Cyber Warfare | Hacktivism |
|---------------------|-------------|-----------------|---------------|------------|
| Continent | | | | |
| América | 18.62 | 184.62 | 175.00 | 177.78 |
| Asia | 9.68 | 275.00 | 60.00 | 20.00 |
| Australia y Oceanía | 15.15 | 100.00 | 0.00 | 0.00 |
| Europa | 12.83 | 131.82 | 78.57 | 72.00 |
| International | 2.86 | 161.90 | 133.33 | 150.00 |
| África | 40.00 | 0.00 | 0.00 | inf |
| Total | 15.97 | 178.67 | 87.50 | 86.96 |

Finalmente, observando el total de ataques, se puede apreciar como los resultados se adecuan a los *ODDS* obtenidos a través del modelo de regresión logística, ya que los *cyber espionajes* son la tipología de ataque que más afecta al sector público. De esta forma, gran parte de los recursos destinados a asegurar los datos de nuestra empresa, deberán ir destinados a la investigación de los ciber crímenes. Adicionalmente, a través de un *TreeMap*, observaremos como es la volumetría de estos ataques en Europa.

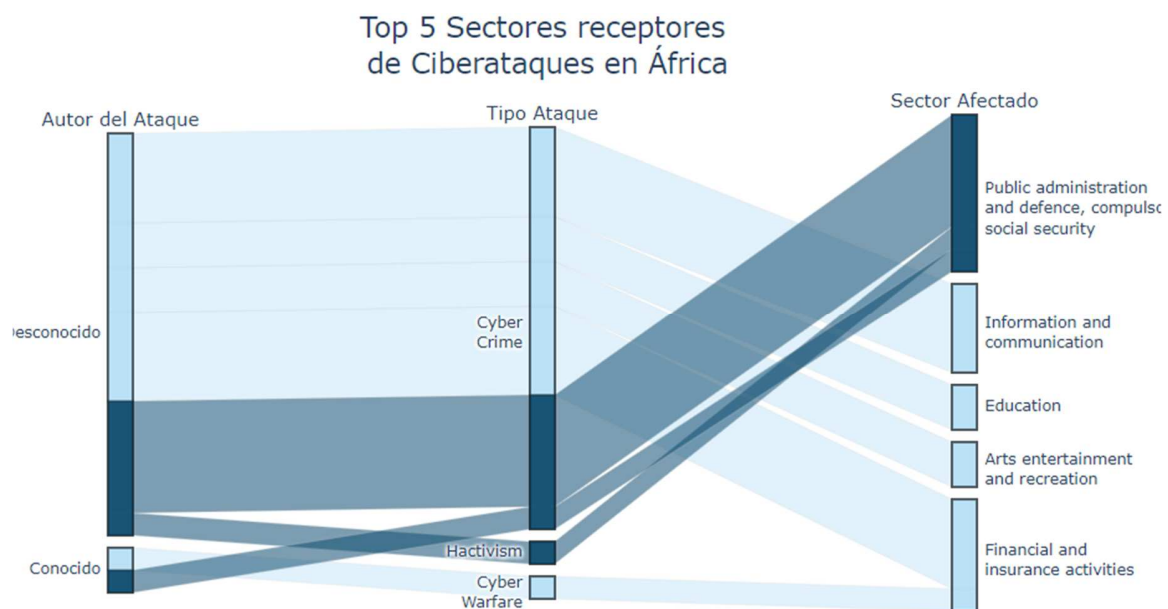


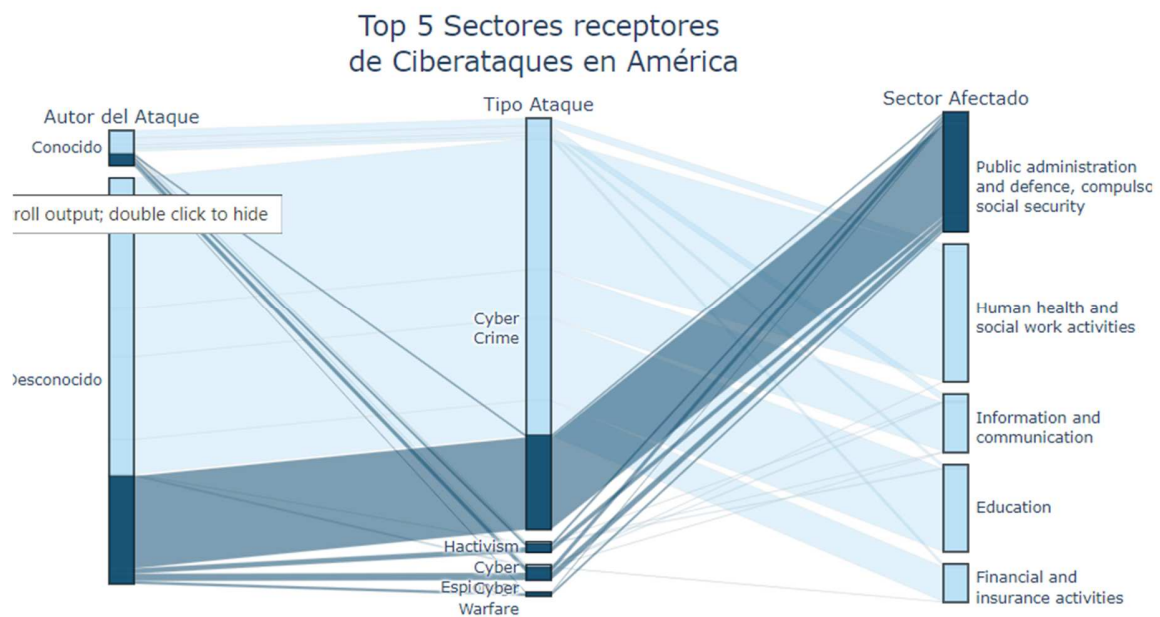
Como última aproximación, realizaremos un conjunto de *Parallel Sets* [11] que nos permitirán visualizar las relaciones presentes entre los autores de los ciberataques y las tipologías de ataque que utilizan, a lo largo de los distintos continentes. De esta forma, a través de estas visualizaciones, podremos contestar a una de las cuestiones planteadas inicialmente con el objetivo de saber si estamos más expuestos a amenazas producidas por autores conocidos o desconocidos. Adicionalmente, podremos observar cuales son las técnicas de ataque predominantes para cada tipo de atacante.

Para realizar dichas visualizaciones, utilizaremos datos de los 5 sectores empresariales más afectados en cada continente.

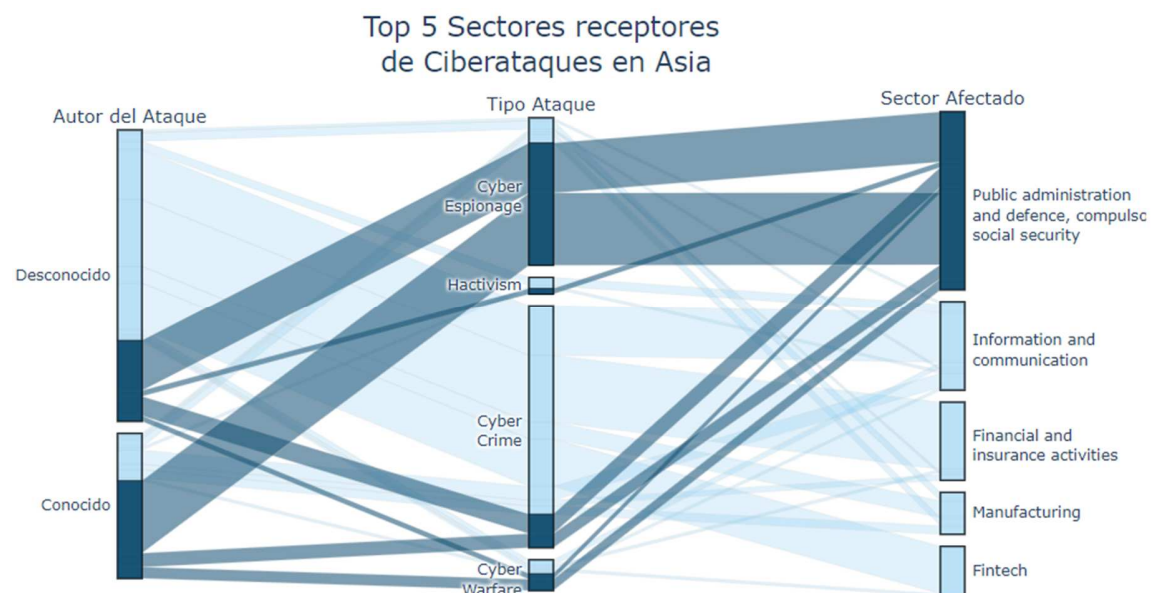


En Europa, parece que las empresas del sector público son mayoritariamente atacadas por autores desconocidos, los cuales utilizan principalmente técnicas de *Cyber Espionage* y de *Cyber Crime*. Por otra parte, aproximadamente un 50% de los autores conocidos, dirigen sus ataques a empresas de nuestro sector.

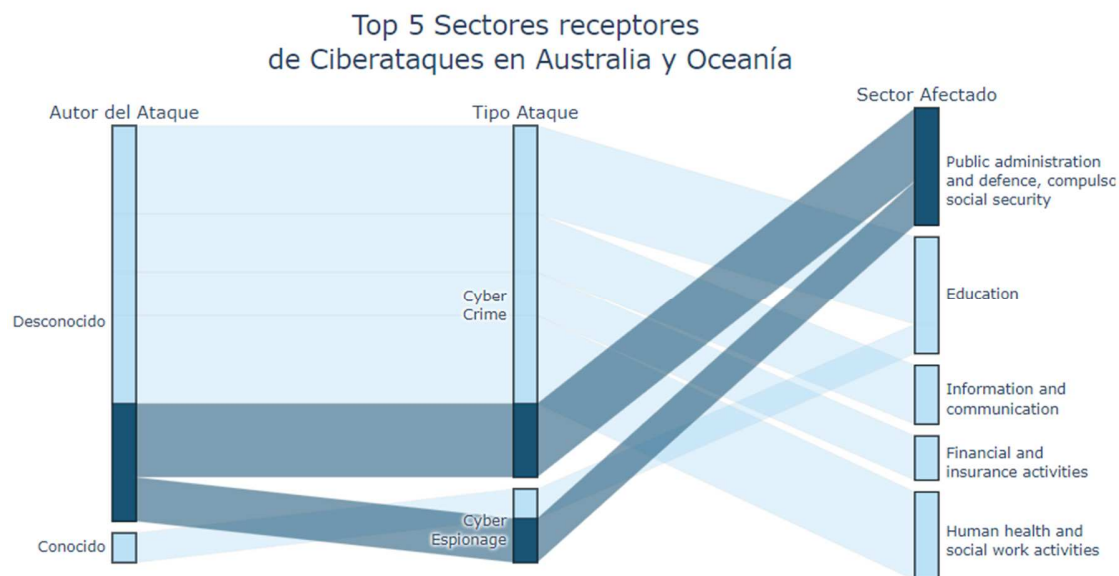




Tanto África como América, se parecen a Europa, dónde la mayor proporción de ciber atacantes son desconocidos y utilizan técnicas de *Cyber Crime*. Adicionalmente, en África, los *Hactivism*, son tipologías de ataque exclusivas en autores desconocidos. Por otra parte, parece que en África no se han registrado ataques de *Cyber Espionage*.



En Asia la situación cambia, siendo los atacantes conocidos los mayores causantes de ciberataques a empresas del sector público. Adicionalmente, estos ataques son *Cyber Espionages*.



En Australia y Oceanía, exclusivamente los autores desconocidos son los causantes de ciberataques a empresas del sector público. Estos, además, utilizan técnicas de *Cyber Crime* y *Cyber Espionages*. Finalmente, parece que aún no se han registrado ataques del tipo *Hactivism* o *Cyber Warfare*.

Tal como se ha observado a través de los *Parallel Sets* de cada continente, existe similitud entre los patrones de ataques a empresas del sector público a lo largo de todo el mundo, dónde, la mayoría de atacantes son desconocidos. Adicionalmente, parece curioso que, el sector público se encuentra en el top 5 de sectores receptores de ciberataques en todos los continentes. Este hecho nos indica que gran parte de la inversión, deberá ir destinada a tareas de seguridad.

Finalmente, incluiremos unas visualizaciones que nos ayudarán a interpretar la perspectiva temporal de los ataques. En este sentido, rescataremos la recta de regresión de la sección 4.2.1, con la que se podía explicar el crecimiento del número de ataques producidos en el sector público a lo largo del tiempo:

$$\text{NumeroAtaques} = 10.4645 + t \cdot 0.2391$$

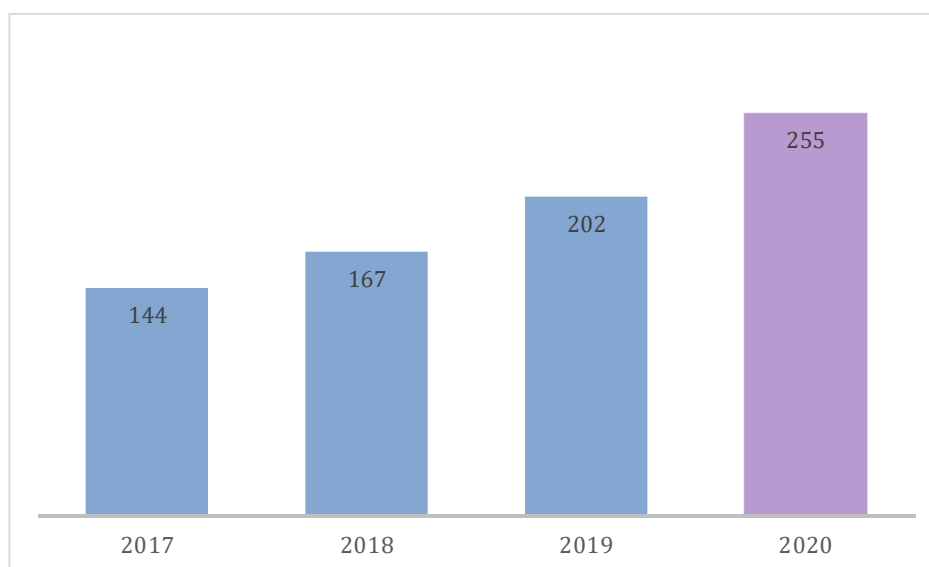
A partir de este análisis, puede presentarse para el año 2020, una tabla con el número de ataques pronosticados para los próximos meses (en rojo) y el número de ataques reales cuantificados hasta la fecha.

| Año | Mes | Número Ataques |
|------|-----|----------------|
| 2020 | 1 | 30 |
| 2020 | 2 | 27 |
| 2020 | 3 | 21 |
| 2020 | 4 | 8 |
| 2020 | 5 | 20 |
| 2020 | 6 | 21 |

| <i>Año</i> | <i>Mes</i> | <i>Número Ataques</i> |
|-------------------|-------------------|------------------------------|
| 2020 | 7 | 21 |
| 2020 | 8 | 21 |
| 2020 | 9 | 21 |
| 2020 | 10 | 21 |
| 2020 | 11 | 22 |
| 2020 | 12 | 22 |

Para la creación del cuadro anterior, hemos hecho uso de la ecuación de regresión temporal, en la que se estudió que el parámetro de la pendiente resultaba estadísticamente significativo, para obtener los valores predichos de ataques en los meses que están por llegar (hay que tener en cuenta que en el momento de la presentación del presente trabajo aún no se disponía del dato de mayo).

De manera gráfica, podemos ver en forma de diagrama de barras, la evolución temporal (por años) del número de ataques de nuestro sector.



Este tipo de representación, a pesar de resultar muy básica, también ayuda a la hora de comparar el número de casos históricos (representado en azul), frente al número de casos estimado para el presente año (representado en morado). Claramente la tendencia creciente se deberá traducir en una dedicación de recursos acorde al volumen de incidentes que se espera que puedan producirse.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

En esta sección se resumirán las conclusiones obtenidas a través de los análisis realizados en la sección 4 utilizando adicionalmente, las gráficas y tablas generadas en la sección 5. De este modo, tal como se planteó en la introducción, las conclusiones irán dirigidas a responder cada una de las siguientes cuestiones:

¿Hasta qué punto mi compañía se encuentra más expuesta a ciberataques, en función del sector en el que se encuentra?

Tal como se muestra a lo largo de la sección 5, las empresas del sector público, aparecen en el top 5 de empresas receptoras de ciberataques en todos los continentes. Este resultado demuestra que somos la compañía con mayor probabilidad de recibir un ataque, y por lo tanto, a diferencia de otras empresas, deberemos destinar una gran parte de nuestros recursos a garantizar la protección cibernética.

¿Existe algún periodo del año en que debería ampliar el presupuesto y los medios necesarios, al estar más expuesto a recibir ciberataques?

Según nos ha demostrado el análisis que hemos llevado a cabo, a pesar de que la tendencia en el número de ataques que recibiremos como entidad pública será creciente, las pruebas de contraste efectuadas no han arrojado una incidencia estadísticamente significativa en algún mes en concreto en comparación con el resto.

Este hecho, presupuestariamente hablando, debería animarnos a dotarnos de unos medios y unos recursos uniformes a lo largo de todo el año, a pesar de que en valores absolutos hemos observado que típicamente se producen más ataques en el primer trimestre del año.

En conclusión, si dispusiéramos de un equipo de vigilancia y de respuesta ante incidentes relativamente constante a lo largo del tiempo, resultará interesante vigilar con especial énfasis los periodos de final de año y de inicio del nuevo año, pudiéndose incrementar ligeramente la realización de tareas de protección ante ciberataques.

Este año, posiblemente, podríamos protegernos en aquellos momentos que puedan producirse picos de ataques no esperados, con personal eventual o contratando servicios externos de seguridad.

¿A qué tipología de ciberataque mi compañía está más expuesta? ¿Debería por ello, analizar la documentación cualitativa, así como los detalles técnicos referentes a este tipo de ataque?

Tal como se muestra a través del modelo de regresión logística creado en la sección 4.2.3, las empresas del sector público están más expuestas a recibir ataques de ciber espionaje, mientras que los ciber crímenes, son la tipología de ataque con menor amenaza. Este resultado cobra sentido ya que, al realizar tareas dentro del ámbito social del sector público, la mayor parte de los ataques que recibiremos irán destinados a la obtención de información confidencial. En consecuencia, deberemos adaptar nuestros sistemas de seguridad con el objetivo de protegernos ante esta técnica de ataque concreta, sin descuidar las otras tipologías que, aunque presenten un nivel de amenaza inferior, siguen estando presentes.

Aunque excede del ámbito del presente estudio, y quizás toca áreas más relacionadas con los sistemas de protección de datos, estos resultados animan a explorar sistemas de niveles de seguridad y de confidencialidad de la información. Es decir, si sabemos que uno de los principales riesgos a los que nos exponemos es al intento de robo de información pública sensible, o de acceso a datos con fines de divulgación, será relevante reforzar los mecanismos de segregación de funciones, la creación de niveles de acceso a información sensible, y la dotación de medios de seguridad física y lógica para aquella información especialmente confidencial que manejen nuestros sistemas.

¿Hasta qué punto el encontrarme en un mundo globalizado, estoy expuesto a recibir ataques internacionales?

A través de las pruebas estadísticas realizadas en la sección 4.2.2, los ataques internacionales se producen en menor medida que los ataques continentales, por consiguiente, estaremos más expuestos a recibir ataques originados en nuestro continente, Europa.

Adicionalmente, a través de la prueba de *kruskal-wallis*, se observa que existen diferencias significativas entre los ataques producidos entre continentes, por lo que deberemos destinar gran parte de nuestros recursos a la investigación de los ataques producidos a empresas europeas pertenecientes al sector público.

Por otra parte, a través del análisis visual realizado en la sección 5, Europa se sitúa en la parte media de la tabla de continentes receptores de ataque, por lo que en comparación a Asia y América, deberemos de destinar inversiones menor en tareas de seguridad.

¿Existen atacantes bien conocidos, con pautas concretas, a los que me encuentre particularmente expuesto?

Tal como se muestra a lo largo de la sección 5, los autores desconocidos son los que más amenaza suponen para las empresas de nuestro sector, por lo que nuestros sistemas de seguridad deberán de ser suficientemente heterogéneos para soportar distintas metodologías de ataque procedentes de personas no identificadas. De esta forma, se

destinarán, en menor medida, recursos a investigar atacantes conocidos para descubrir sus tendencias y hábitos de ataque más utilizados.

A partir de las conclusiones obtenidas, podemos observar que el uso de metodologías de recopilación, limpieza y tratamiento de datos, pueden ayudar en gran medida a la mitigación de riesgos cibernéticos.

Pese a las limitaciones de tiempo, recursos y tipología de datos (con la carencia de variables numéricas) a los que nos hemos enfrentado para la realización de esta práctica, hemos conseguido obtener un conjunto de pautas generales y recomendaciones específicas para tratar distintos ataques cibernéticos ocurridos en una empresa concreta. En este sentido, se pone en relieve la importancia presente en procesos como la extracción, la limpieza y el procesamiento de datos, así como las técnicas de análisis y *data mining*, para obtener el conocimiento necesario que nos permita realizar tareas de prevención y mitigación de riesgos informáticos.

De hecho, con el entorno de trabajo desarrollado, se podrían plantear diversos casos de estudio con los que se podrían obtener resultados más precisos y valiosos. Algunos de estos casos de estudio serían:

- El estudio de fuentes de información en las que se documenten técnicas de resolución de incidentes. Es decir, de igual modo que hemos enfocado nuestro estudio desde la perspectiva del atacante, un estudio similar podría desplegarse para la recopilación y el análisis de datos sobre mecanismos de seguridad, haciendo énfasis en su grado de éxito en la mitigación de ciberataques.
- La utilización de modelos de predicción basados en la obtención temprana de datos. A partir del estudio que hemos diseñado, podría programarse una recopilación temprana de incidentes que están ocurriendo en nuestro continente (ya que hemos visto que son el tipo de incidentes que pueden afectarnos con mayor probabilidad), de manera que cuando se notifique un incidente en el sector público, de una determinada tipología en otro país europeo, se activen mecanismos de refuerzo en previsión de que podamos acabar siendo afectados.
- Aunar esfuerzos internacionales en atrapar a delincuentes cibernéticos. A pesar de que nuestro estudio no ha profundizado en la correlación entre determinadas tipologías de incidentes y autores concretos, un estudio con mayor detalle en esta línea podría tratar de poner en contacto a entidades públicas y privadas, que estuvieran siendo atacadas de manera recurrente por los mismos atacantes, para aunar esfuerzos en la resistencia ante estos ataques, pudiendo incluso dar respuesta y armar un contraataque.
- El resultado y el valor de nuestra respuesta estará directamente relacionado con la calidad de los datos que seamos capaces de recopilar. A través de todas las dificultades y retos que hemos necesitado resolver a lo largo de la práctica (errores de raspado de datos, errores inherentes en las propias fuentes de datos, valores perdidos o atípicos, etc.), hemos experimentado la dificultad que entraña el asegurarnos que podemos dar por concluido de forma satisfactoria, el proceso de

acondicionamiento de datos. Según hemos podido comprobar, aquellos datos cuya calidad no es lo suficientemente buena, pueden estropear totalmente un experimento estadístico, o incluso llevar a conclusiones erróneas. En concreto, en ámbitos como el que hemos tratado de estudiar, el hecho de concluir incorrectamente sobre los orígenes de los ataques o las tendencias de los mismos, pueden derivar en una mayor exposición a la recepción de ciberataques. Por este motivo, creemos que un proceso de gestión del ciclo de vida de los datos que considere y pondere de forma adecuada cada etapa, puede contribuir a resolver grandes problemas de análisis.

Contribuciones

| | |
|-----------------------------|-----------|
| Investigación Previa | JBP – IRP |
| Redacción de las respuestas | JBP – IRP |
| Desarrollo código | JBP – IRP |

Bibliografía

- [1] Subirats, Laia - Pérez, Diego O. - Calvo, Mireia (2019). “Introducción a la limpieza y análisis de los datos”, Universidad Oberta de Catalunya
- [2] Bock, Tim (2019). “What is a Crosstab”, Display R Blog. [en línea] [Última consulta: 15/May/2020] <https://www.displayr.com/what-is-a-crosstab/>
- [3] Osborne, Jason W. (2013). “Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data.” Thousand Oaks, CA. Sage Publications.
- [4] How to Handle Missing Data. (2020). Retrieved 21 May 2020, from <https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>
- [5] Datacarpentry (2018). “Aggregating and analyzing data with dplyr” [en línea] [Última consulta: 20/May/2020] <https://datacarpentry.org/R-genomics/04-dplyr.html>
- [6] Peña, D. (2010), “Análisis de Series Temporales”, Alianza Editorial.
- [7] Parada, Luis F. (2019). “Tipificación de variables”. RPubS. [en línea] [Última consulta: 24/May/2020] <https://www.rpubs.com/F3rnando/521190>
- [8] Rovira, Carles (2009), “Teorema del límite central”, Universidad Oberta de Catalunya.
- [9] Amat, Joaquín (2016), “T-test: Comparación de medias poblacionales independientes”, RpubS. [en línea] [Última consulta: 25/May/2020] https://rpubs.com/joaquin_AR/218467
- [10] Tree Map (2020), Learn about this chart and tools to create it. [en línea] [Última consulta: 26/May/2020] <https://datavizcatalogue.com/methods/treemap.html>
- [11] Parallel Sets (2020), Data Viz Project. [en línea] [Última consulta: 26/May/2020] <https://datavizproject.com/data-type/parallel-sets/>