



Universidad de Valladolid

**FACULTAD DE CIENCIAS
TRABAJO DE FIN DE GRADO**

Grado en Estadística

**Investigación estadística
en el ámbito de la criminología**

Alumno: Iván López de Munain Quintana

Tutora: Lourdes Barba Escribá

Índice

Índice de figuras	3
Índice de cuadros	4
Resumen	5
I Memoria del proyecto	6
1. Introducción	6
1.1. Estado del arte	7
2. Objetivos perseguidos en la investigación	7
3. Metodología de trabajo	8
3.1. Planificación del proyecto	8
3.2. Herramientas utilizadas	9
4. Plan de investigación	10
4.1. Ámbito del estudio	10
4.2. Variables analizadas	10
4.3. Preprocesamiento de los datos	11
5. Análisis descriptivo	12
6. Técnicas descriptivas avanzadas	24
6.1. Análisis de correspondencias	24
6.2. Análisis de componentes principales	29
6.2.1. Clasificación mediante K-Medias	32
7. Fase de modelado	34
7.1. Análisis de la varianza	36
7.2. Modelo de Poisson	37
7.2.1. Predicciones y residuales	42
7.2.2. Capacidad predictiva	43
7.3. Redes neuronales recursivas	44
7.3.1. Red neuronal recursiva simple	44
7.3.2. Red neuronal LSTM (Long Short-Term Memory)	45
8. Futuras investigaciones	46
9. Conclusiones	47
Referencias	48

II Anexos	50
A. Aclaraciones	50
B. Configuración Rmarkdown	50
C. Librerías	50
D. Análisis descriptivo	51
D.1. Mapa interactivo de Europa	51
D.2. Mapas España: valores absolutos y tasas	55
D.3. Tabla valores medios infracciones 2016-2018	61
D.4. Barplots: tipología, sexo y edad	62
D.5. Diagrama de sankey: territorio y tipología	66
D.6. Boxplot MENAS	69
D.7. Barplot: comparativa tasa de paro y criminalidad	73
D.8. Gráfico animado	78
D.9. Gráfico evolución infracciones por edad y sexo	81
E. Técnicas descriptivas avanzadas	83
E.1. Análisis de correspondencias	83
E.2. Análisis de componentes principales	93
E.2.1. Clasificación K-Medias	98
F. Fase de modelado	100
F.1. Normalidad y ANOVA	100
F.2. Modelo de Poisson	106
F.3. Redes neuronales recursivas	115
F.3.1. Procesamiento	115

Índice de figuras

1.	Tasa de infracciones cometidas por países europeos (Interactivo)	13
2.	Infracciones cometidas por comunidad en 2018.	15
3.	Infracciones cometidas por provincia en el año 2018.	16
4.	Número de delitos por cada 1000 habitantes en 2018.	17
5.	Distribución de los delitos según su tipología (Interactivo).	18
6.	Distribución de los delitos según su tipología sin clases mayoritarias (Interactivo) . . .	19
7.	Distribución de los delitos según territorio y tipo por 100.000 habitantes (Interactivo). .	19
8.	Distribución de los delitos según el sexo y edad del infractor (Interactivo).	21
9.	Comparación tasas de paro y criminalidad por sexo y comunidad (Interactivo)	21
10.	Relación entre el PIB y la población reclusa por comunidades (Interactivo)	22
11.	Número de delitos cometidos por cada 1000 menores según su nacionalidad (Inter- activo)	23
12.	Distribución de los delitos según año, edad y sexo (Interactivo)	24
13.	Porcentaje de varianza explicada por dimensión (Interactivo)	26
14.	Contribución de los perfiles fila en las dimensiones escogidas (Interactivo)	26
15.	Contribución de los perfiles columna en las dimensiones escogidas (Interactivo) . .	27
16.	Distribución de los niveles de las covariables.	28
17.	Porcentaje de la varianza explicada por componente (Interactivo)	29
18.	Nube de variables.	30
19.	Nube de variables e individuos.	31
20.	Representación 3D de la nube de variables e individuos (Interactivo)	32
21.	Relación entre número de clusters y varianza explicada (Interactivo)	33
22.	Clasificación comunidades (Interactivo)	33
23.	Gráfico Q-Q para la variable respuesta.	34
24.	Histograma para estudiar la normalidad de la variable respuesta.	35
25.	Gráfico Q-Q para la variable respuesta por grupos.	35
26.	Residuos frente a predichos del modelo de análisis de la varianza.	36
27.	Estimaciones de los parámetros del modelo escogido (Interactivo)	41
28.	Predicciones para el logaritmo del número de infracciones por año.	42
29.	Residuos frente a predichos (Interactivo)	42
30.	Qqnorm para los residuos.	43
31.	Porcentaje de acierto/error del modelo ajustado.	43
32.	Grafo de interconexiones de una red neuronal recursiva simple	44
33.	Tasa de éxitos en una red neuronal recursiva simple.	45
34.	Grafo de interconexiones de una red neuronal LSTM	45
35.	Tasa de éxitos en una red neuronal recursiva LSTM.	46

Índice de cuadros

1.	Calendarización del proyecto.	8
2.	Evolución del número de infracciones entre los años 2016-2018	14
3.	Tasa de agresiones sexuales por cada 100.000 habitantes	20
4.	Tasa de agresiones sexuales con penetración por cada 100.000 habitantes	20
5.	Comparativa para menores según su nacionalidad para el año 2018.	23
6.	Número de detenidos e investigados por edad y comunidad en el año 2018.	25
7.	Primeras observaciones datos Europa	51
8.	Primeras observaciones: número de infracciones	62
9.	Primeras observaciones: infracciones según tipología	63
10.	Primeras observaciones: infracciones según sexo y edad	65
11.	Primeras observaciones: infracciones MENAS	70
12.	Primeras observaciones: comparativa tasa paro y criminalidad	75
13.	Primeras observaciones: infracciones por edad y territorio	85
14.	Primeras observaciones: tasa criminalidad por tipología y territorio	94
15.	Primeras observaciones: datos empleados modelo Poisson	100
16.	Primeras observaciones: serie temporal	118

Resumen

Este proyecto pretende aplicar diversas técnicas estadísticas en el entorno criminológico con el fin de extraer conclusiones que aporten claridad a aspectos actuales y/o polémicos. Se trata de un análisis, mayoritariamente a nivel nacional, en el que se aspira a configurar el perfil base de un delincuente, además de estudiar la evolución y distribución de la criminalidad sobre el territorio español. Asimismo, mediante factores socioeconómicos, trata de establecer posibles causas de estos comportamientos delictivos consiguiendo centrar la atención en la raíz del problema y así hallar medidas para su prevención.

Abstract

This project aims to apply different statistical techniques in the criminological field in order to draw conclusions that bring clarity to current and/or controversial aspects. It is based on an analysis, mainly at national level, in which the aim is to configure the basic profile of an offender, besides studying the evolution and distribution of crime over Spanish territory. Additionally, through socioeconomic factors, it attempts to establish the causes of these criminal behaviours by focusing attention on the root of the problem, thus finding measures to prevent it.

Keywords

Tasa de criminalidad, número de infracciones, población reclusa, producto interior bruto, modelo lineal generalizado, distribución de Poisson, red neuronal recursiva, análisis de correspondencias, análisis de componentes principales, análisis de la varianza, Rstudio, Python, plotly, ggplot2.

Parte I

Memoria del proyecto

1. Introducción

El comportamiento humano ha sido, es y seguirá siendo uno de los principales focos en los que se centra la ciencia. Para poder comprender el por qué de nuestras acciones, tratamos de buscar patrones regulares que definan nuestro comportamiento. Ahora bien, aquí se encuentra el factor imprevisible, aquello que es guiado por la voluntad del individuo, su libertad como ser racional. He aquí donde se encuentra el verdadero reto, tratar de modelar lo puramente aleatorio.

Prácticamente en cualquier ámbito de la sociedad se ha tratado de predecir el comportamiento del ser humano para obtener beneficios, destacando sobre todo los entornos económicos, sociales, políticos y comerciales. El interés de este proyecto reside en el ámbito social, donde cabe destacar la aplicación de la *Estadística a la Criminología y a la Justicia Criminal*.

La criminología es una ciencia multidisciplinar que abarca muchas ramas del conocimiento como la psicología, educación social, derecho o medicina; aunque los métodos estadísticos representan la columna vertebral sobre la que se sostiene el avance del conocimiento en las investigaciones criminológicas. Esto se debe a la necesidad de recopilar información y obtener estadísticas nuevas, o bien extraer datos actualizados y compararlos con anteriores para poder evaluar su evolución en el tiempo o incluso realizar predicciones futuras.

Esto supone el estudio de conjuntos de datos numéricos sobre los crímenes y los delincuentes, siendo estos extraídos de los registros de organismos oficiales. Las estadísticas criminales pueden iniciarse a través de distintas vías [1]:

- *Vía policial*: datos recogidos y depurados por la policía. Principalmente cubren delitos e infracciones que son registrados de las distintas dependencias administrativas.
- *Vía judicial*: proporcionada por los jueces y magistrados penales del país de acuerdo con los procesos iniciados y las providencias dictadas en su desarrollo.
- *Vía penitenciaria*: almacena información relacionada con la población reclusa del país. Proporciona clasificaciones en función de distintos factores como la categoría del delito cometido, procedencia, sexo, edad y más aspectos personales de los presos.

Las investigaciones en criminología son muy diversas tanto atendiendo a su finalidad como a su naturaleza propia. La gran mayoría se basa en evaluar distintas teorías, compuestas por hipótesis, que tratan de aportar explicaciones verosímiles a unos determinados eventos. Los principales tipos de investigación son *exploratorios* (cuando el objetivo no ha sido abordado previamente), *descriptivos* (el objetivo principal es describir situaciones y eventos) y *experimentales* (estudio prospectivo en donde el investigador manipula el factor de estudio).

En resumen, es inmediato darse cuenta de la importancia de que los criminólogos desarrollen su función de forma efectiva y así puedan asesorar a los gobiernos y a las sociedades acerca de la problemática de la delincuencia. Para ello resulta vital disponer de estadísticas de todo tipo que ayuden a los investigadores, y a la sociedad en general, a examinar el problema en su total amplitud.

1.1. Estado del arte

A lo largo de la historia, pese a la dificultad que conlleva, se han realizado una gran cantidad de informes y artículos versando sobre este tema. Esta complejidad es debida a la escasez y dispersión de datos, además de la problemática incluida en la adaptación de los métodos estadísticos a las investigaciones criminológicas. Hasta ahora, la metodología general que se ha utilizado para la investigación de la delincuencia puede dividirse en dos vertientes: métodos *cuantitativos* y *cualitativos* [2].

Los métodos *cuantitativos* nos permiten conocer la frecuencia con la que suceden distintos acontecimientos y describir sus características, además de definir variables que pueden estar influyendo en la ocurrencia y magnitud del evento. Se utilizan para conocer el desarrollo y evolución del fenómeno, así como inferir resultados que nos permitan hacer predicciones futuras acerca del mismo. En esta vertiente se puede encontrar la aplicación de modelos aditivos lineales, regresiones múltiples, modelos de clustering, etcétera [3]. Por contra, los métodos *cualitativos* van más allá de los datos ofrecidos por los organismos oficiales. Esto se consigue mediante la realización de encuestas, ya sean de victimización o de autoinforme¹, que permiten explicar la producción del fenómeno.

Debido a la naturaleza de ambas vertientes metodológicas este proyecto se centrará en la descrita en primer lugar, focalizándose en *modelos lineales generalizados (GLM)*, *análisis de correspondencias (AC)*, *análisis de componentes principales (ACP)* y *redes neuronales recursivas*.

2. Objetivos perseguidos en la investigación

Como bien se ha comentado anteriormente, el grueso de la investigación pretende aplicar distintas técnicas estadísticas al entorno judicial y criminológico para tratar de comprender mejor las causas y circunstancias de los delitos, buscando aportar claridad sobre temas actuales y controvertidos. Para lograr estas metas, se ha fijado una serie de objetivos a cumplir por el proyecto:

1. Estudio superficial sobre la tasa de criminalidad europea en los últimos años, teniendo como fin poner en contexto al lector sobre la situación española respecto el resto de países europeos en términos delictivos.
2. Evolución del número de delitos e investigados a lo largo de los años, desde el nivel nacional hasta el provincial.
3. Análisis de la distribución de la tasa de delincuencia sobre el territorio español, desde el nivel nacional hasta el provincial. Así es posible realizar clasificaciones territoriales en función de dicha tasa.
4. Crear posibles perfiles de tipología de una persona más propensa a delinquir, atendiendo a su nacionalidad, sexo, edad, etcétera. De esta manera, se puede extraer un modelo de personalidad de los delincuentes y emplearlo para su represión.
5. Estudio de la delincuencia desde el punto de vista socioeconómico, buscando posibles causas que lleven a una persona a traspasar la ley. Se tendrán en cuenta indicadores como el paro o el PIB.

¹Las encuestas de victimización recogen si los individuos han sido víctimas de algún tipo de delito, mientras que las encuestas de autoinforme registran si han cometido alguna infracción en un periodo de tiempo. Ambas realizadas sobre poblaciones representativas.

6. Examinar la tipología de los delitos entendiendo su distribución y composición, de tal forma que se pueda focalizar la atención en una categoría en concreto que se considere más relevante.
7. Ahondar en problemas actuales como es la violencia de género, estudiando cuáles son las comunidades con mayor tasa de agresiones sexuales para así tomar medidas para su prevención.
8. Esclarecer la cuestión sobre la supuesta relación entre los menores extranjeros no acompañados y el alto grado de delincuencia.
9. Aplicación de distintos tipos de modelos para estudiar y entender mejor el fenómeno criminalógico.

3. Metodología de trabajo

3.1. Planificación del proyecto

El progreso del *Trabajo de Fin de Grado* se ha monitorizado mediante reuniones presenciales entre tutora y alumno durante el curso académico del año 2019-2020, además de las correspondientes dudas resueltas a través de la plataforma de mensajería de la universidad. El desarrollo del proyecto queda definido en cuatro etapas principales, mostradas a continuación en orden cronológico de ejecución:

1. Ingeniería de datos → búsqueda y contrastación de fuentes fiables de datos; realización de un posterior procesamiento de dichos datos para poder desempeñar las siguientes tareas.
2. Fase exploratoria → definición del problema y del marco de estudio, planteamiento de hipótesis, enfoque y análisis descriptivo.
3. Fase de modelado → ajuste de modelos y estudio tanto de los residuales como de la capacidad predictiva.
4. Interpretación de resultados → extracción de conclusiones.

Con cierto grado de abstracción, se emplea un diagrama de *Gantt* para mostrar la organización, el orden de ejecución y el tiempo de dedicación a cada una de las tareas realizadas.

	Enero	Febrero	Marzo	Abril	Mayo	Junio
Búsqueda de datos						
Procesado de datos						
Análisis exploratorio						
Elaboración de hipótesis						
Fase de modelado						
Interpretación resultados						
Extracción de conclusiones						

Cuadro 1: Calendarización del proyecto.

3.2. Herramientas utilizadas

A la hora de desarrollar el proyecto han sido necesarios distintos programas software, los cuales son descritos a continuación:

- **R (Rstudio):** se trata de un entorno y lenguaje de programación enfocado al análisis estadístico. Debido a que es un software libre, se ha convertido en uno de los lenguajes más populares en *machine learning* y *minería de datos*. Además de que puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como *Python*. Otro de sus puntos fuertes es su capacidad para la representación gráfica mediante librerías como *plotly* o *ggplot2**. Representa el software más potente y utilizado durante el desarrollo de este proyecto.
- **Overleaf (LATEX):** se trata de la herramienta online usada para realizar la documentación pues es un sistema de composición de textos. Es habitual emplearlo en artículos académicos donde se requiera una alta calidad tipográfica.
- **Python (Anaconda):** se trata de un lenguaje de programación multiparadigma ya que soporta orientación a objetos, programación imperativa y funcional. Es interpretado, dinámico y multiplataforma. Se emplea durante la parte de ingeniería de datos, sobre todo en aquellos casos que *R* no sea eficiente, además de en el ajuste de las redes neuronales recursivas.
- **SAS:** se utiliza únicamente como apoyo para los resultados obtenidos a través de los lenguajes de *Python* y *R*. Esto es debido a que no es un software libre y su manejo queda restringido a los ordenadores de la universidad. Se caracteriza por tener pasos *data* y *procedimientos* que permiten realizar operaciones sobre los datos, además de disponer de un intérprete *SQL* y basarse en *macros*.
- **Gitlab:** se trata de un servicio web de control de versiones y desarrollo de software colaborativo basado en Git. En este proyecto se ha utilizado el dominio proporcionado por la Universidad de Valladolid. Esta herramienta se ha empleado básicamente para generar las páginas *html* de los gráficos interactivos.

*Comentarios sobre ggplot2 y plotly:

Respecto a las librerías *ggplot2* y *plotly* cabe remarcar que la primera proporciona gráficos ‘estáticos’ mientras que la segunda los ofrece ‘interactivos’. Por consiguiente, en la mayoría de gráficos mostrados en el posterior desarrollo existe la posibilidad de clickar en el pie de foto de tal forma que se redirija a una página *html* donde se mostrará dicha representación interactiva. De esta forma se consigue explotar al máximo su potencial alcanzando un mayor grado de representación e ilustración del problema abordado. Dichas figuras se identifican fácilmente puesto que en sus pies de foto aparece la etiqueta ‘Interactivo’. Huelga decir que para acceder a dichos *html* es necesario disponer de conexión a internet. Para lograr estos resultados, se ha subido el archivo local *html* generado por las funciones de *plotly* a *Gitlab*, donde se creará una página *html* pública.

A la hora de acceder a dichas páginas existe la posibilidad de que dé la sensación de que el gráfico está desconfigurado, bastará con que el usuario ajuste el zoom (*Ctrl + / Ctrl -*) hasta que se vislumbre adecuadamente. Esto ocurre debido a que se trata de páginas *html* configuradas en un determinado ordenador con sus características específicas, mientras que podrán ser accedidas por otras computadoras con propiedades diferentes (e.g. resolución y tamaño de pantalla).

Por último, se va a describir de forma genérica las posibilidades de interacción en dichos gráficos. Para empezar existe la posibilidad de ‘recortar’ directamente en el gráfico para poder focalizar la atención en una determinada parte de la representación. Además de esto, existe un menú en la parte superior derecha en la que se proporcionan funcionalidades como el zoom, selección de una parte del gráfico, resetear y autoescalar a la situación inicial, disponer de un modo de ‘arrastre’ para desplazarse por la ilustración en caso de haber empleado el zoom, selección de variables (clickando sobre su nombre), etcétera. Existen más opciones de interacción que las aquí citadas pero todas son intuitivas y de fácil exploración por parte del usuario, destacar que dependiendo del tipo de gráfico habrá diferentes posibilidades.

4. Plan de investigación

4.1. Ámbito del estudio

Este proyecto se desarrolla con datos procedentes de todo el territorio español desde el año 2010 hasta la actualidad, remarcando que la información del 2019 no se tiene en cuenta debido a que se encuentran de forma incompleta. Además, en aquellos casos donde se empleen subconjuntos de datos quedará especificado su composición de forma explícita. Las principales fuentes de extracción de información son, a nivel nacional, los datos abiertos del Ministerio del Interior [4], del Instituto Nacional de Estadística [5] y del Poder Judicial de España [6], mientras que a nivel europeo son los datos abiertos de la Unión Europea [7] y del Eurostat [8].

4.2. Variables analizadas

El ámbito criminológico es un entorno de gran complejidad que requiere de un amplio número de covariables para su explicación. Debido al formato pobre y disperso de los datos abiertos no se puede estudiar todos los factores que se consideran relevantes. Las variables principales analizadas son:

1. **Número de infracciones:** se trata de la variable respuesta por excelencia de este proyecto, en definitiva, se puede medir el grado de delincuencia de un territorio a través de la cantidad de delitos cometidos por su población. Se trata de una variable numérica cuya distribución variará en función de las covariables que serán explicadas a continuación. Es decir, se pretende estudiar la distribución de la delincuencia según el sexo, edad y nacionalidad de las personas; la comunidad y/o provincia donde residen; el año en el que ha ocurrido y el tipo de delito cometido.
2. **Tasa de infracciones:** se trata de una variación de la variable anterior en la que se representa el número de infracciones cometidas por cada mil (o cien mil, dependiendo el caso) habitantes de un territorio, queriendo estudiar su distribución según las mismas características explicadas en el punto anterior.
3. **Sexo:** variable dicotómica (femenino o masculino). Es interesante conocer el sexo predominante de los infractores, esta distinción puede ayudar a explicar la naturaleza de la variable objetivo.
4. **Edad:** variable continua que se transforma a variable categórica ordinal. Los distintos niveles del factor edad son: 14-17, 18-30, 31-40, 41-64 y +64.

5. **Comunidad/Provincia:** variable categórica que indica el territorio (comunidad o provincia) de donde provienen los datos. Cabe remarcar que en los desarrollos posteriores se especifica el nivel de profundidad territorial empleado.
6. **Año:** variable numérica discreta que representa los años cuando son cometidos los delitos.
7. **Tipología:** variable categórica que representa una clasificación de las infracciones según su naturaleza y ejecución.
8. **Nacionalidad:** variable dicotómica (español o extranjero). Gracias a esta covariable se puede estudiar el peso de la procedencia del individuo en la variable respuesta.

4.3. Preprocesamiento de los datos

Como bien se ha resaltado previamente, el escaso contenido de las fuentes junto con su descentralización en distintos organismos oficiales, hacen de esta fase un hito crucial en el proyecto. De esta manera, se disponen los datos de la forma requerida por los modelos desarrollados, pudiendo variar estas configuraciones en función del objetivo a lograr. Remarcar que en este apartado se realiza una descripción genérica de todos los procedimientos llevados a cabo, describiéndose de forma más detallada posteriormente en aquellos casos que así se requiera.

Al extraer los datos de diferentes fuentes es importante integrarlos de forma única para poder trabajar con ellos de manera uniforme y homogénea. Para lograrlo, es preciso crear una tabla de autoridad para las comunidades y provincias puesto que cada fuente lo almacena de una forma distinta (e.g. Comunidad Foral de Navarra - Navarra). Además de haber provincias dispuestas en otras lenguas como el catalán o euskera; o en algunas ocasiones las tildes no se reconocen y son sustituidas por caracteres alfanuméricos. Mediante estas tablas de autoridad, se puede realizar *merges* de columnas y filas procedentes de distintos conjuntos de datos.

Para la realización de la mayoría de gráficos, se tienen que llevar a cabo las operaciones correspondientes de agregación y desagregación para hacer agrupaciones según el sexo, edad, nacionalidad o año, entre otros. Específicamente para un ejemplo, cabe resaltar la carga de procesamiento requerida para realizar la Fig. 10 en la que se integra información de tres fuentes de datos distintas para mostrar la evolución temporal de la población reclusa y el PIB por comunidades.

Los datos extraídos a nivel europeo suponen un esfuerzo de procesamiento mayor debido a que los dataset mezclan distintos caracteres de separación, llegando a utilizar en un mismo *csv* el punto y coma, la coma y la tabulación de forma simultánea. Para resolver esto y facilitar así su lectura en *R*, se emplean expresiones regulares de tal forma que en el fichero solo se utilice una vía de separación de datos. En suma, también se tiene que utilizar la codificación ISO3 de los países debido a que los nombres de los países pueden variar; aunque otra posible acción es modificar dichas denominaciones manualmente al tratarse de pocos casos (e.g. Reino Unido vs Inglaterra-Escocia-Gales-Irlanda del Norte).

Otro aspecto recurrente en la mayoría de conjuntos descargados es la necesidad de conversión de tipos, por ejemplo, una variable numérica como lo puede ser el número de infracciones que se encuentre como categórica en el fichero.

Por último, se quiere destacar la dificultad para crear el conjunto de datos empleado para la red neuronal recursiva. A la hora de construir la secuencia temporal (explicada su composición en [7.3-**Redes neuronales recursivas**]) los años más recientes se obtienen del Ministerio del Interior mientras que los registros más antiguos de los históricos del INE. Esto supone la necesidad de

adecuarlos puesto que el INE los dispone desagregados según la tipología del delito, en suma, no se puede estudiar un marco temporal más amplio debido a que los históricos del INE a partir del 2015 se encuentran almacenados como tablas en PDF y no como conjuntos de datos descargables.

Todos estos procedimientos son desarrollados mayoritariamente mediante el software *Rstudio*.

5. Análisis descriptivo

Antes de comenzar el estudio ya expuesto, es conveniente examinar de forma minuciosa la situación actual de España en el ámbito criminológico. Por lo que surgen una serie de dudas, ¿somos conscientes de cuántas infracciones cometan los ciudadanos españoles durante un año? ¿Esta tasa de ilegalidades está por encima del resto de países europeos? ¿En qué comunidades y/o provincias españolas es mayor el grado de delincuencia? ¿A qué pueden deberse estas diferencias entre distintos territorios? Estas cuestiones serán contestadas empleando un enfoque descendente, a saber, desde un nivel europeo hasta llegar a una profundidad provincial.

Para estudiar el grado de delincuencia a nivel europeo se realiza un mapa (Fig. 1) en el que se representa el número de delitos cometidos por cada 100.000 habitantes, pudiendo seleccionar entre las estadísticas de los años 2015, 2016 y 2017 para observar la evolución en el tiempo de la criminalidad (no se dispone de registros más actuales). Además de esto, al desplazar el ratón por encima de los países aparece una ventana en la que se dispone la información delictiva relativa a cada territorio, mostrando la tasa total de infracciones y su distribución entre las distintas categorías especificadas por Eurostat (todo esto se puede realizar en la representación interactiva). En el caso de datos desconocidos se muestran como ‘NA’. El formato de la ventana es el siguiente:

- **Country name:**
- **Total rate:**

- **Categorization of crime in 2017**
 - Intentional homicide:
 - Attempted intentional homicide:
 - Assault:
 - Kidnapping:
 - Sexual violence:
 - Rape:
 - Sexual assault:
 - Robbery:
 - Burglary:
 - Burglary of private residential premises:
 - Theft:
 - Theft of a motorized land vehicle:
 - Unlawful acts involving controlled drugs or precursors:

La Fig. 1 representa la tasa delictiva de las regiones europeas para el año 2017, donde claramente se aprecia que los países con mayor índice de criminalidad son Reino Unido, Dinamarca y Suecia. Por el contrario, en el otro extremo se encuentran países como Albania, Montenegro o Macedonia,

teniendo en cuenta que esto puede ocurrir debido a que hay muchos datos desconocidos para dichos países. Respecto al resto de años, se observa el mismo patrón aunque destaca cómo Francia, Grecia y Bélgica se suman a los países más delictivos durante los años 2015 y 2016. Puede afirmarse que España se encuentra en el ecuador del espectro delictivo europeo durante estos últimos años.

Pese a las comparaciones aquí realizadas, es importante ser consciente de que los datos pueden ser engañosos. Esto se debe a que las diferencias entre países, en términos de códigos penales y procedimientos legales, es sustancial. Además de las divergencias culturales entre regiones (e.g. en materia de violencia de género). Solo se podrían hacer comparaciones totalmente rigurosas y fiables en el caso de que todos los países dispusiesen del mismo sistema judicial y legislativo, además de compartir mismas raíces culturales.

Para realizar el gráfico interactivo a nivel europeo se utiliza la función *plot_geo*, requiriendo ésta el nombre de los países o su codificación ISO3 para identificarlas. También es posible emplear las coordenadas longitudinales y latitudinales para especificar las regiones que se quieren representar, siendo esta última opción recomendable en aquellos casos en los que se pretende focalizar la atención en niveles más concretos como por ejemplo el nacional.

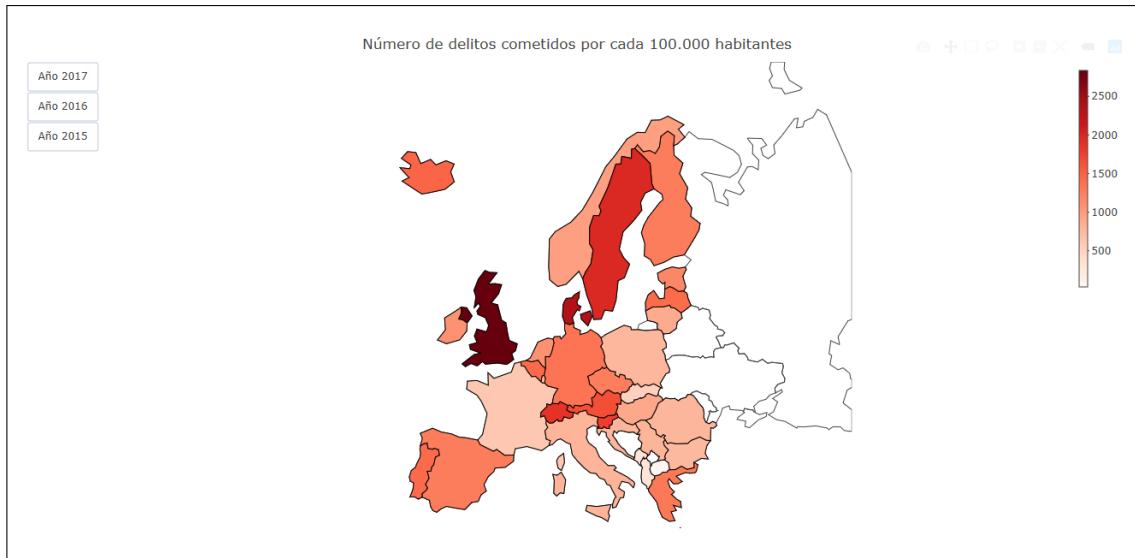


Figura 1: Tasa de infracciones cometidas por países europeos (Interactivo).

A continuación, descendiendo ya a nivel nacional, se muestra el número de infracciones penales documentadas por comunidad autónoma entre los años 2016 y 2018 (incluidos), además de enseñar los totales y el número medio.

Comunidades autónomas	Año 2016	Año 2017	Año 2018	Total	Nº medio
Andalucía	331836	334331	333199	999366	<u>333122</u>
Aragón	38146	37878	40255	116279	38759
Asturias	26093	25652	26453	78198	26066
Islas Baleares	68160	72157	72944	213261	71087
Canarias	88782	91359	90566	270707	<u>90235</u>
Cantabria	17002	17440	17655	52097	17365
Castilla-La Mancha	63579	64006	68115	195700	65233
Castilla y León	72037	71874	74744	218655	72885
Cataluña	403928	422286	471389	1297603	<u>432534</u>
Ceuta	4633	4503	4673	13809	4603
Melilla	4832	4744	5465	15041	5013
Comunidad Valenciana	230209	227102	231581	688892	<u>229630</u>
Extremadura	26286	26411	26535	79232	26410
Galicia	74405	74174	77969	226548	<u>75516</u>
Madrid	373402	381242	389955	1144599	<u>381533</u>
Murcia	51309	52690	54893	158892	52964
Navarra	25704	24828	27287	77819	25939
País Vasco	82809	83687	90214	256710	85570
La Rioja	7997	7877	8010	23884	7961
Total nacional	2009690	2045785	2131424	6186899	2062299

Cuadro 2: Evolución del número de infracciones entre los años 2016-2018

Se puede apreciar que el número medio anual de infracciones cometidas es superior a los dos millones, siendo las comunidades de Andalucía, Cataluña y Madrid en las que hay mayor grado de delincuencia (todas corresponden con los territorios más poblados). Además, también se puede destacar cómo aquellas zonas que reciben turismo ‘líquido’, como es el caso de las Islas Baleares o las Islas Canarias, tienen una mayor tasa de infracciones respecto al tamaño de su población. Mismamente, Canarias, pese a ser casi 60.000 habitantes menos que Galicia, cometan de media cerca de 15.000 infracciones más. Es clara la existencia de una correlación entre el turismo ‘etílico’ y la cantidad de delitos cometidos.

Estos resultados nos permiten observar que Cataluña es la comunidad autónoma con mayor número de infracciones, pero antes de precipitarse, cabe destacar que es el segundo territorio con mayor población por detrás de Andalucía. También es importante darse cuenta de la situación actual de la sociedad catalana debido al movimiento independentista. Una fecha muy importante en el

desarrollo de esta corriente fue la imposibilitación del referéndum del 1 de Octubre de 2017. Esta paralización supuso posteriores altercados, quedando éstos reflejados en un aumento de casi 50.000 infracciones más en el año 2018 con respecto 2017. En ninguna comunidad se aprecia una variación tan grande como ocurre en este caso.

Para poder reforzar estas ideas de una forma más ilustrativa, se presenta un mapa de las comunidades españolas en función del número total de infracciones durante el año 2018. Este gráfico es realizado mediante el lenguaje de programación *R*, usando las librerías *sp* (para leer los archivos *.RDS*²) y *RColorBrewer* (para definir la paleta de colores). Remarcarse que la información territorial y espacial se ha obtenido a través de *GADM* [9] que proporciona los datos sobre las subdivisiones de países de todo el mundo. Finalmente, tras la adecuación de los datos, se emplea la función *spplot* para obtener el siguiente gráfico:

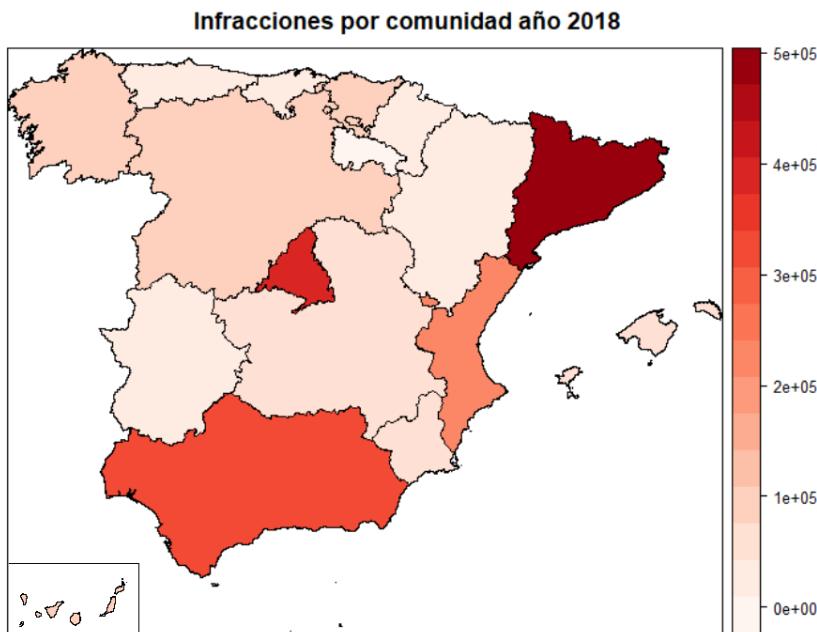


Figura 2: Infracciones cometidas por comunidad en 2018.

²En nuestro caso, se tratan de archivos que almacenan un *SpatialPolygonsDataFrame* definido en el paquete *sp* donde quedan descritas las distintas subdivisiones de un país (pudiendo escoger el nivel: nacional, provincial, etcétera).

Podemos observar cómo mediante la figura anterior (Fig. 2) se acentúan las ideas ya expuestas previamente. Es por eso, que siguiendo el mismo método, se muestra a continuación un mapa con un nivel más de profundidad, el provincial:

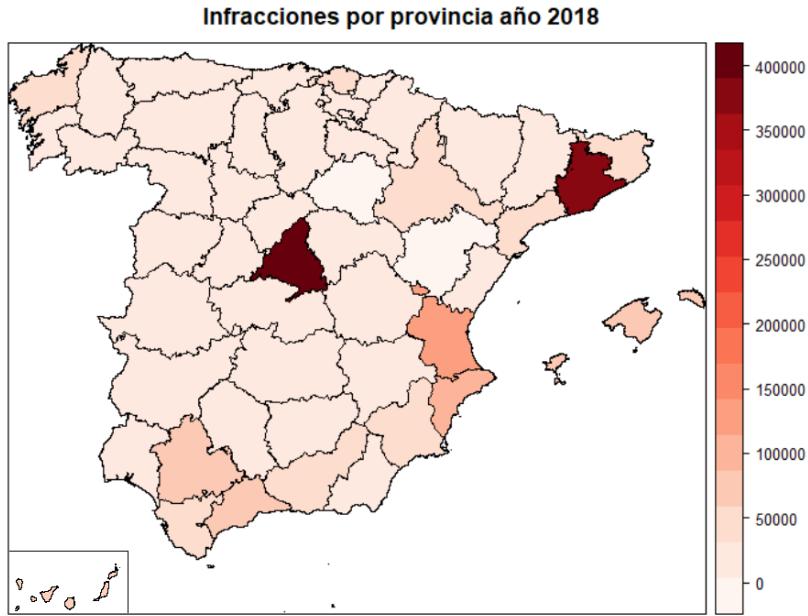


Figura 3: Infracciones cometidas por provincia en el año 2018.

Gracias a Fig. 3 comprobamos que la provincia de Barcelona es la que mayor número de delitos aporta a la comunidad de Cataluña, mientras que Madrid se convierte ahora en el mayor exponente de delincuencia. Por contra, en Andalucía, no destaca ninguna provincia en especial. Cabe remarcar aquellas provincias de la España vaciada, como son Soria o Teruel, que obviamente corresponden con los territorios con menor número de delitos.

Realizar estos mapas, según el número absoluto de infracciones, nos sirve para saber dónde focalizar las fuerzas policiales y ser conscientes de las necesidades de cada territorio. Ahora bien, para llevar a cabo comparativas consistentes es preciso tener en cuenta la población de cada región. De esta forma, se obtienen los siguientes mapas (Fig. 4) en donde se manifiesta la tasa de delitos cometidos por cada mil habitantes durante el año 2018 según la comunidad y/o provincia.

Con esta nueva información se refuerza y fundamenta la idea ya expuesta sobre las causas negativas del turismo ‘líquido’, pues las Islas Baleares se encuentran entre las comunidades con más delincuencia compitiendo contra Madrid o Cataluña (siendo su población la sexta y séptima parte respectivamente). Además, pese a que no se aprecie en el gráfico, Ceuta con 54 delitos/1.000 habitantes y Melilla con 63 delitos/1.000 habitantes, también conforman un núcleo de delincuencia de gran nivel, llegando a ser casi mayores que las citadas anteriormente.

A nivel provincial, destaca la alta tasa de Barcelona llegando a casi 70 delitos/1.000 habitantes. Otras provincias destacables, que no habían sido detectadas hasta ahora, son: Alicante (48 delitos/1.000 habitantes), Málaga (47 delitos/1.000 habitantes), Vizcaya (46 delitos/1.000 habitantes) y Valencia (46 delitos/1.000 habitantes). Mientras que en el otro extremo, representando la región más segura en el territorio nacional, se encuentra Teruel con tan solo 20 delitos/1.000 habitantes.

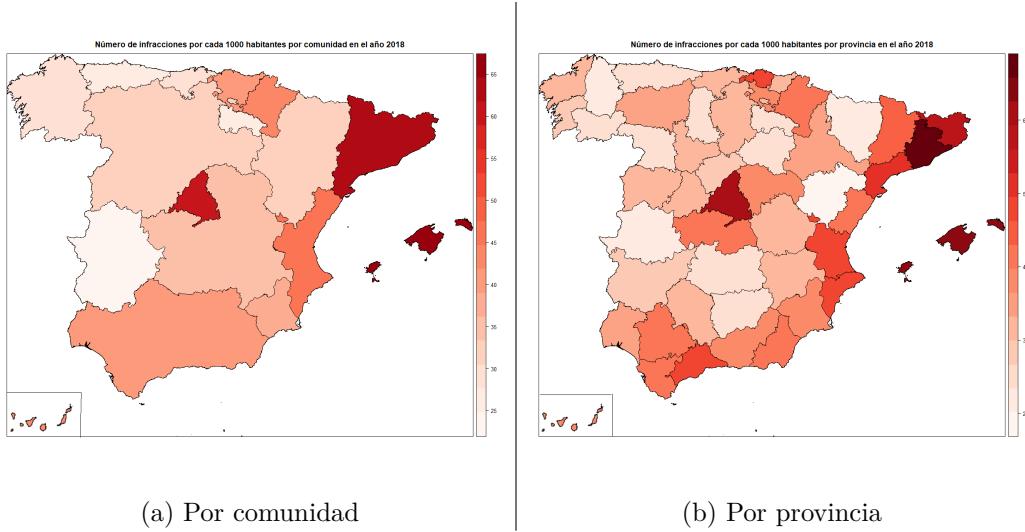


Figura 4: Número de delitos por cada 1000 habitantes en 2018.

Hasta ahora se ha analizado la distribución por territorios del total de hechos conocidos, entendiendo por esto el conjunto de infracciones penales y administrativas que han sido conocidas por las distintas Fuerzas y Cuerpos de Seguridad (bien por medio de denuncia interpuesta o por actuación policial realizada motu proprio). Ahora bien, respecto a la naturaleza de los sujetos responsables de la comisión de las infracciones se computan las siguientes categorías:

- **Investigado:** se trata de una persona física o jurídica a la que se atribuya la participación en una ilegalidad, aunque no se adoptan medidas restrictivas de libertad para esa persona imputada.
- **Detenido:** supone la lectura de derechos de la persona física, privándole de libertad y poniéndolo a disposición judicial.
- **Víctima:** se engloba mediante el término de victimización, representando hechos denunciados por personas en los cuales manifiestan ser víctimas de alguna infracción penal.

En posteriores desarrollos se focaliza la atención en los datos que aporten claridad sobre los rasgos y características de los detenidos e investigados. Los datos abiertos del Ministerio del Interior [4] categorizan los delitos de la siguiente forma:

- **A)** Asesinatos consumados.
- **B)** Asesinatos en grado tentativo.
- **C)** Delitos de lesiones.
- **D)** Secuestro.
- **E)** Delitos contra la libertad e indemnidad sexual.
- **F)** Agresión sexual con penetración.
- **G)** Resto de delitos contra la libertad.

- **H)** Robos con violencia e intimidación.
- **I)** Robo con fuerza en establecimientos.
- **J)** Robo con fuerza en domicilios.
- **K)** Hurtos.
- **L)** Sustracción de vehículos.
- **M)** Tráfico de drogas.
- **N)** Resto de delitos.

Mediante el siguiente *barplot* (Fig. 5), realizado con los datos del 2018, se puede observar cómo los hurtos representan la infracción más mayoritaria tras la categoría *N* que aglutina a los delitos cuya clasificación no es tan obvia. Cabe destacar que el delito con mayor frecuencia después del hurto es el robo con fuerza, ambas prácticas se caracterizan por apoderarse de cosas muebles ajenas diferenciándose en la necesidad del empleo de la fuerza.

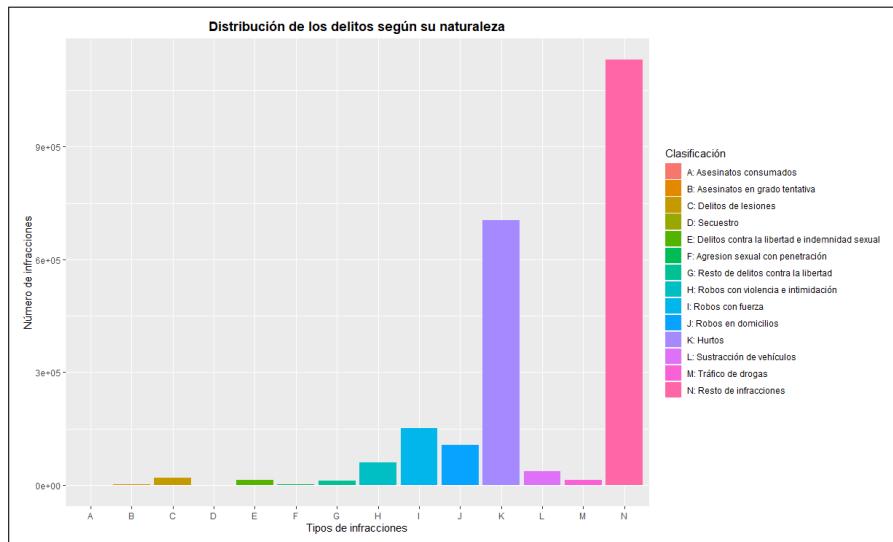


Figura 5: Distribución de los delitos según su tipología (Interactivo).

Para poder apreciar mejor la distribución del resto de tipos de delitos, a continuación, se muestra un *barplot* idéntico al anterior pero sin los grupos mayoritarios ya nombrados; esto se puede conseguir o bien haciendo zoom en el gráfico interactivo o seleccionando los tipos que se quieran ocultar (clickando sobre dichas categorías). Se concluye que las infracciones más frecuentes corresponden con variaciones del robo, ya sea por la ubicación (e.g en domicilios llegando a superar las cien mil ocurrencias), por la metodología llevada a cabo (mediante la fuerza o la intimidación, con 150.579 y 60.677 eventualidades respectivamente) o por el bien sustraído (e.g vehículos con 36.152 delitos). Cabe resaltar, aunque sea menor que los ya citados, el número considerable de infracciones relacionadas con delitos de lesiones y delitos contra la libertad e indemnidad sexual.

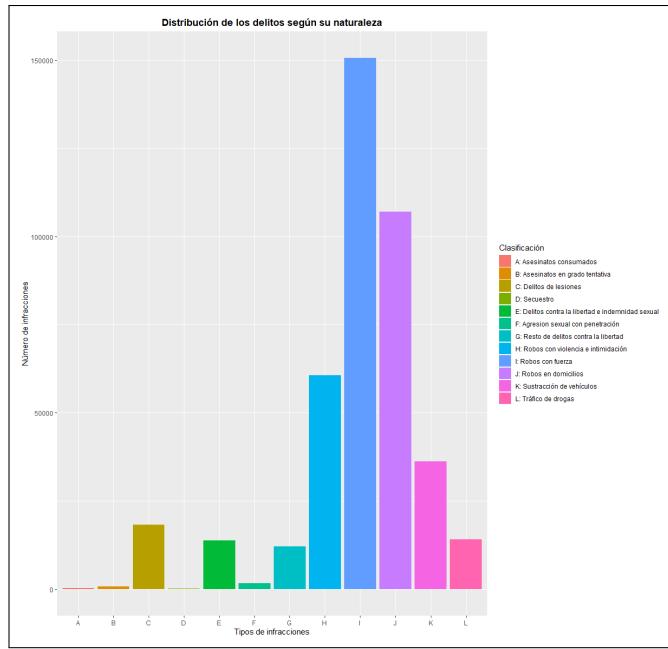


Figura 6: Distribución de los delitos según su tipología sin clases mayoritarias (Interactivo).

Una vez estudiada la distribución de la ilegalidad según territorio y tipología por separado, resulta interesante observar la combinación de ambos factores. Esto se consigue mediante un diagrama *Sankey* (Fig. 7) en donde se representa cómo la tasa de los distintos tipos de delitos se reparten entre las diferentes comunidades. Aunque normalmente la tasa de delincuencia se realiza por cada 1.000 habitantes, aquí se obtiene por cada 100.000 porque al categorizar por tipología en muchas ocasiones la tasa es prácticamente cero. En suma, es elegida una gama de colores violetas que conforme aumenta el tono corresponde con un delito cuya tasa de criminalidad es mayor. Esta representación se realiza mediante las librerías *networkD3* y *dplyr* de *R*, usando la función *sankeyNetwork*.

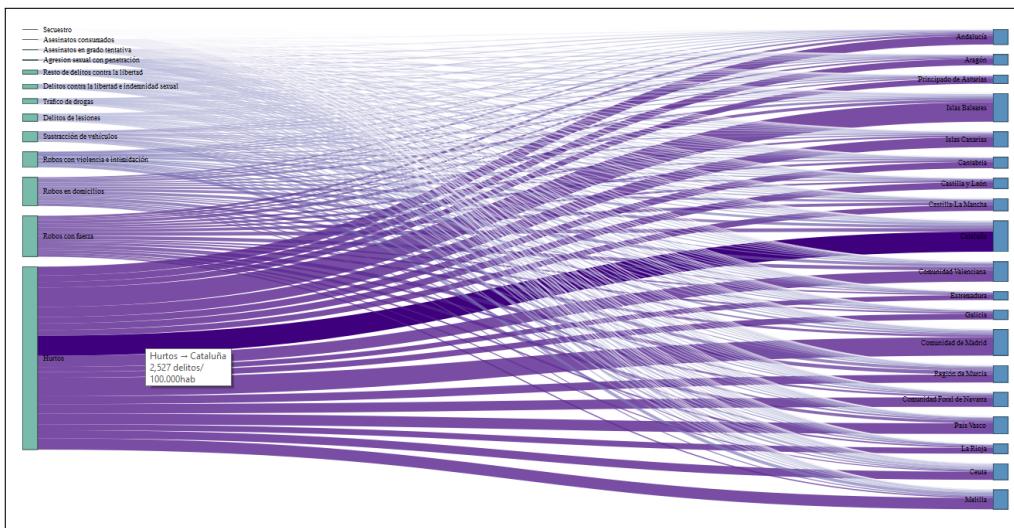


Figura 7: Distribución de los delitos según territorio y tipo por 100.000 habitantes (Interactivo).

Las conclusiones que pueden obtenerse son similares a las ya expuestas en el estudio de las variables por separado, aunque ahora es posible comparar la categorización de la delincuencia por comunidades. En todos los territorios el mayor acto delictivo es el hurto, seguido del robo y sus variaciones. Además de esto, se puede observar una gran diferencia entre comunidades en materia de delitos contra la libertad e indemnidad sexual (corresponden éstos en su mayoría con abusos sexuales, violaciones, acoso, etcétera). La siguiente tabla muestra las cinco regiones con mayor tasa de este tipo de delitos:

Comunidad Autónoma	Nº Agresiones/100.000 habitantes
Ceuta	58.72
Islas Baleares	51.73
Navarra	41.85
Islas Canarias	35.30
Cataluña	34.18

Cuadro 3: Tasa de agresiones sexuales por cada 100.000 habitantes

Profundizando más en el tema, resulta triste comprobar cómo en las Islas Baleares es donde ocurren mayor número de agresiones sexuales con penetración, posiblemente ligado al tipo de turismo que reciben. A continuación se expone una tabla con las cinco comunidades autónomas con mayor tasa de este tipo de delito ³; esta información debería usarse para saber dónde focalizar el esfuerzo en educación y concienciación de la población para evitar estas atrocidades.

Comunidad Autónoma	Violaciones/100.000 habitantes
Islas Baleares	6.47
Cataluña	6.16
Ceuta	5.87
Navarra	5.56
País Vasco	5.05

Cuadro 4: Tasa de agresiones sexuales con penetración por cada 100.000 habitantes

Otro punto muy importante es tratar de configurar el perfil de una persona más propensa a realizar delitos. Para ello, gracias al siguiente *barplot*, se observa que los hombres cometan hasta más del triple de ilegalidades que las mujeres. Mientras que las edades más propensas a cometer delitos, para ambos sexos, son la post-adolescencia (18-30 años) y la adulterez (41-54 años).

³Cabe recordar que en nuestro país se distingue entre abuso sexual (se refiere a cualquier contacto sexual no deseado) y violación (agresión sexual con penetración). En este caso, se están mostrando estadísticas sobre la peor de las situaciones, la violación.

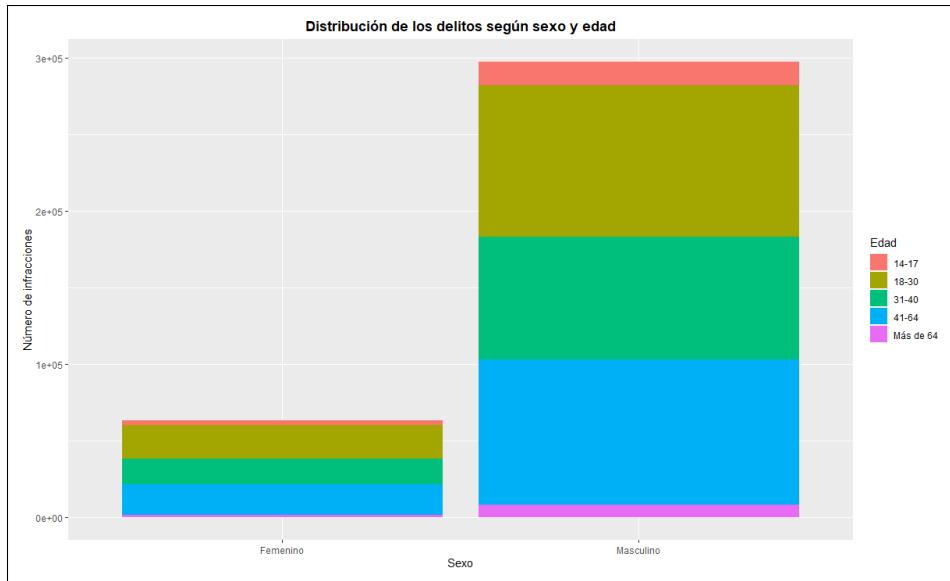


Figura 8: Distribución de los delitos según el sexo y edad del infractor (Interactivo).

Una vez analizadas las principales características de los detenidos e investigados, resulta imperativo tratar de encontrar causas que puedan provocar dichos comportamientos. Un factor económico, como es el paro, puede convertirse en un detonante que guíe a los ciudadanos a cometer delitos en situaciones de desesperación. En la siguiente representación (Fig. 9) se ha contrapuesto las tasas de paro y criminalidad, calculadas por cada 1.000 habitantes, distribuidas por sexo y comunidad durante el año 2018. A primera vista, no se puede apreciar una clara relación entre ambas tasas aunque sí se observa que en la mayoría de regiones con más paro hay un mayor grado de delincuencia. Lo que sí destaca sustancialmente, aunque no de forma sorprendente, es la gran diferencia entre hombres y mujeres en términos de empleo, llegando en algunos casos casi a doblar la tasa de paro en el caso del sexo femenino. Esto tan solo recuerda la dura e injusta realidad para dicho sexo y la reminiscencia de fenómenos como el ‘techo de cristal’.

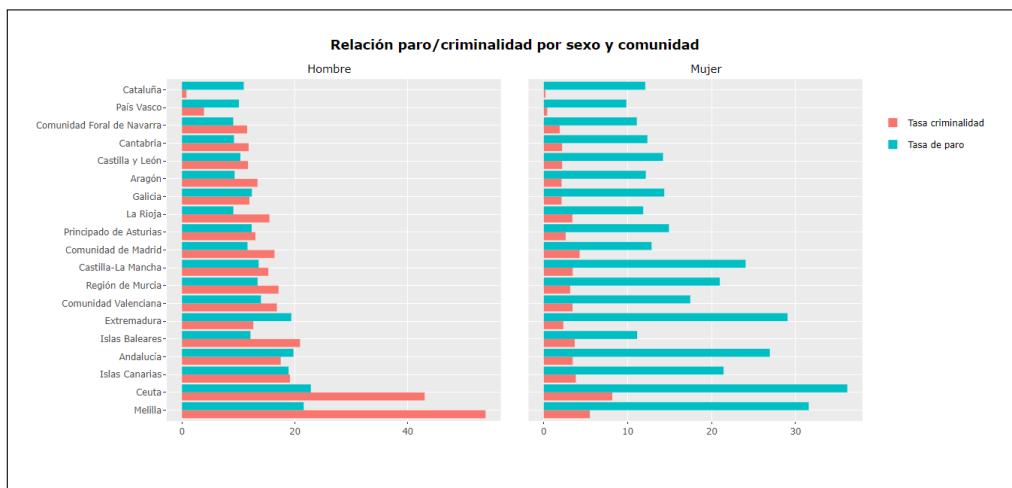


Figura 9: Comparación tasas de paro y criminalidad por sexo y comunidad (Interactivo).

La economía parece estar correlacionada con el grado de delincuencia, en el sentido de que a mayor nivel económico, traduciéndose esto en una mejor calidad de vida, supone un descenso en el índice de criminalidad de una región. Para comprobar esta hipótesis, se ha realizado la Fig. 10 en donde se muestra la evolución del tamaño de la población reclusa y el PIB por comunidad entre los años 2006 y 2018. En suma, el tamaño de cada burbuja depende de la población total de dicha comunidad. Se trata de un gráfico animado en el que se puede observar cómo se interrelacionan el aspecto económico y el delictivo, viendo en términos generales cómo conforme aumenta el PIB de un territorio, por lo que las condiciones sociales mejoran, disminuye la población de presos, es decir, se reduce el nivel de ilegalidades. Esta tendencia se aprecia de forma clara en los últimos años para todas las comunidades, mientras que durante los años 2008-2014 se muestra una recesión en el PIB debido a la crisis económica que sufría España en aquel entonces. En dicho periodo, aunque empeora la economía española, hay algunas comunidades en las que disminuye el número de reclusos. Destaca el cambio brusco del año 2008 al 2009 (coincidencia con el inicio de la crisis), donde casi la totalidad de regiones sufren un empeoramiento económico y un aumento de la población presa. Por último, es aconsejable interactuar con el slider en el que se puede controlar manualmente el cambio de año (en vez del uso automático clickando en el botón *Play*).

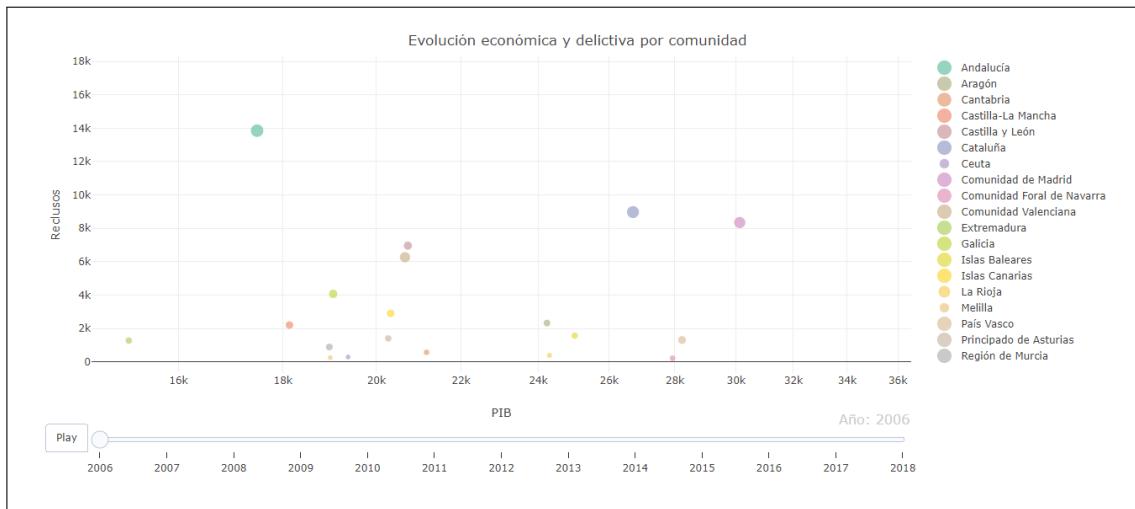


Figura 10: Relación entre el PIB y la población reclusa por comunidades (Interactivo).

Actualmente, existe una gran controversia sobre la situación de los menores extranjeros no acompañados (*MENAS*). Esto es debido a que algunos sectores de la sociedad afirman que el índice de criminalidad aumenta a causa de éstos, pudiendo difundir así discursos de odio. En el siguiente *boxplot* (Fig. 11) se representa la distribución del número de infracciones cometidas por cada 1000 menores en las distintas comunidades del territorio español, distinguiendo los clusters de extranjero y español.

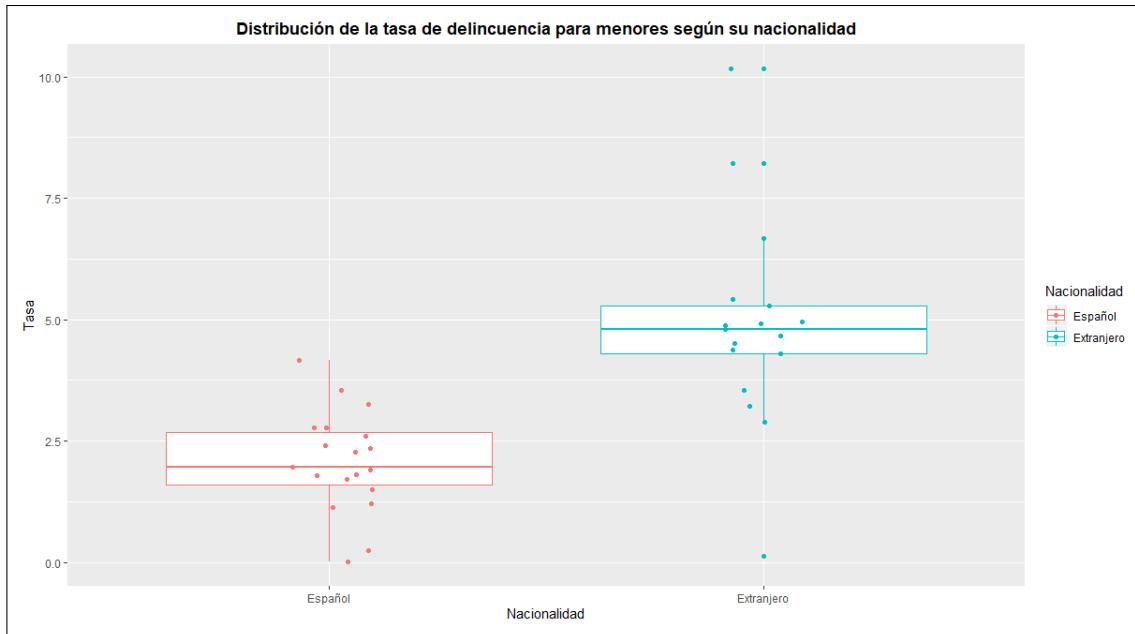


Figura 11: Número de delitos cometidos por cada 1000 menores según su nacionalidad (Interactivo).

Antes de analizar el gráfico, cabe destacar que no se han representado las comunidades de Ceuta y Melilla ya que se tratan de *outliers*, llegando a obtener tasas tan altas como 70 y 63 delitos respectivamente (ambas para menores extranjeros). También es vital entender que se está tratando de aportar claridad a un ámbito en el que el desconocimiento es inherente a la cuestión, es decir, la población de menores extranjeros empleada en este análisis representa una cota inferior del número que verdaderamente residen aquí; esto se debe a que su situación actual en España es ilegal. En suma, es habitual que el motivo por el que los menores extranjeros son censados es que han sido detenidos al realizar una infracción. Por lo que en definitiva, aunque se muestre una mayor tasa de delincuencia en menores extranjeros que en nacionales, esta conclusión no es fiable pues la tasa en el caso de extranjeros disminuiría si se conociese el tamaño real de su población. A continuación se muestra a nivel nacional la población, el número de delitos y la tasa para cada cluster.

Nacionalidad	Población	Nº Infracciones	Infracción/Persona
Españoles	8274338	14374	1.73
Extranjeros	913873	4296	4.70

Cuadro 5: Comparativa para menores según su nacionalidad para el año 2018.

Por último, para finalizar el análisis descriptivo, se va a extender el estudio hasta el año 2010 para poder observar la tendencia del número de detenciones en función del año en el que se ha cometido, el sexo y la edad del infractor.

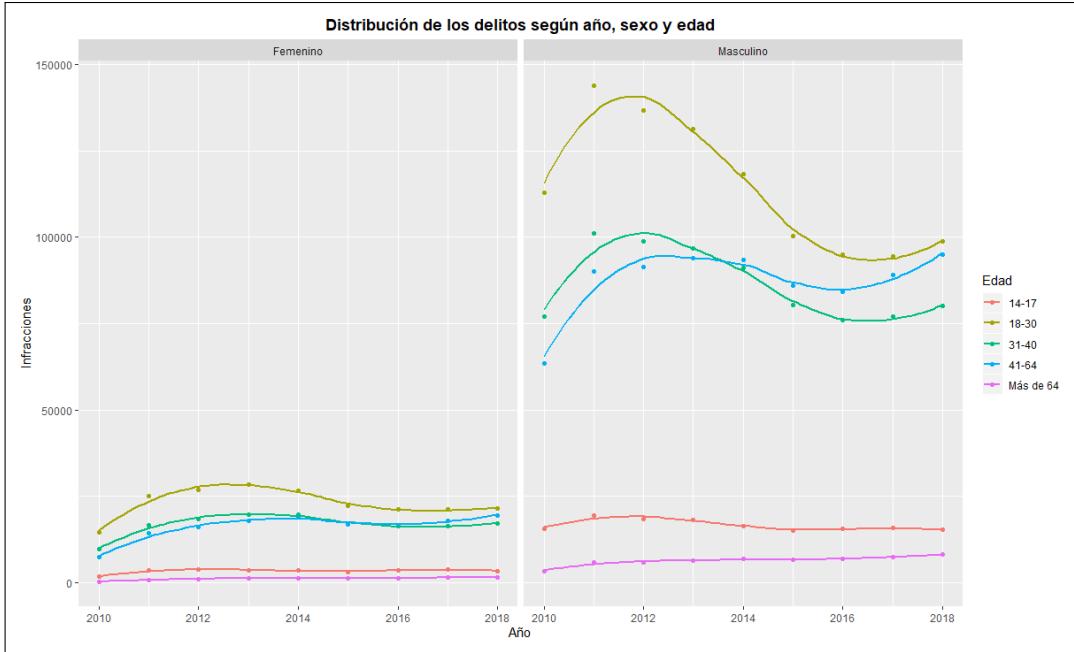


Figura 12: Distribución de los delitos según año, edad y sexo (Interactivo).

En Fig. 12 se concluye que durante la última década el número de ilegalidades cometidas por mujeres se mantiene constante y de una magnitud bastante menor que las realizadas por los hombres. Además en el sexo masculino, sobre todo para edades jóvenes, se observa una clara parábola en donde el máximo número de delitos se cometió en 2011-2012 disminuyendo hasta el 2017, dando la sensación de una posible tendencia en la que en años cercanos podría haber otro aumento. El pico del año 2011-2012 puede deberse a la precaria situación social y económica que predominaba ese periodo, caracterizado por una alta tasa de paro, amplios recortes y un gran número de desahucios que se tradujeron en huelgas generales o movimientos como el 25-S. [12][13]

6. Técnicas descriptivas avanzadas

6.1. Análisis de correspondencias

El análisis de correspondencias se asemeja con una técnica descriptiva de interdependencia para análisis multivariados de datos categóricos, considerándose una extensión del análisis de componentes principales (*ACP*) para variables no numéricas. Para el caso bivariante, el cual se emplea aquí, se representa la información mediante una tabla de contingencia en la que se puede mostrar tanto los valores absolutos como las frecuencias relativas de los factores que se están analizando. El objetivo final es representar dichas variables en un espacio de dimensión menor donde se puedan observar posibles agrupamientos y relaciones entre ellas. En definitiva, este tipo de estudio tiene dos objetivos básicos:

- Buscar similitudes entre categorías de una misma variable para ver si sus niveles pueden ser combinados.
- Buscar similitudes entre categorías de variables distintas, estudiando si existe relación entre los niveles de las filas y de las columnas.

En primer lugar, se va a aplicar un análisis de correspondencia para estudiar la posible asociación entre las distintas comunidades y la edad de los detenidos e investigados durante el año 2018. Siendo la tabla de contingencia la siguiente:

Comunidades autónomas	14-17	18-30	31-40	41-64	+64	Total
Andalucía	3870	28652	24308	27861	2332	87023
Aragón	1000	3276	2624	2916	227	10043
Asturias	292	2173	2006	2934	385	7790
Islas Baleares	616	4820	4001	4134	277	13848
Canarias	796	7433	6659	8724	665	24277
Cantabria	175	1289	1046	1357	114	3981
Castilla-La Mancha	1035	6532	4951	5986	520	19024
Castilla y León	788	5223	3963	5911	675	16560
Cataluña	48	1308	1072	1213	56	3697
Ceuta	154	1084	518	429	15	2200
Melilla	322	1335	543	373	16	2589
Comunidad Valenciana	3000	15892	13507	16066	1315	49780
Extremadura	380	2480	1885	2970	268	7983
Galicia	539	5224	4990	6954	814	18521
Madrid	4090	24340	17865	18481	1525	66301
Murcia	973	4930	4121	4645	330	14999
Navarra	188	1553	1171	1322	94	4328
País Vasco	228	1939	1189	1251	52	4659
La Rioja	176	893	816	989	87	2961
Total	18670	120376	97235	114516	9767	360564

Cuadro 6: Número de detenidos e investigados por edad y comunidad en el año 2018.

A primera vista, se puede ver cómo en las columnas destacan las categorías de 18-30 y de 41-64 años, mientras que en las filas sobresalen regiones como Andalucía o Madrid. A continuación, se elabora un *scree-plot* en donde se expone el porcentaje de la varianza explicada por cada dimensión. Viendo el gráfico se concluye que con extraer dos dimensiones es suficiente puesto que explican el 90 % de la variabilidad total del modelo.

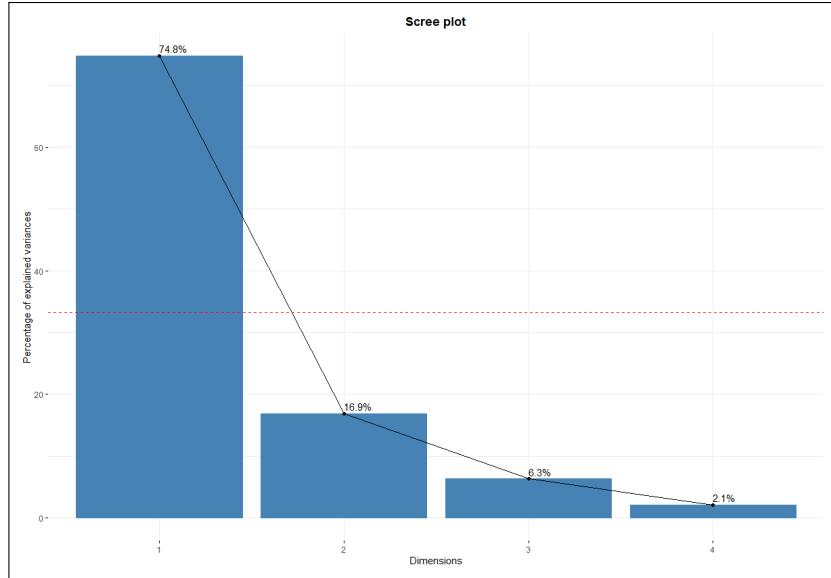


Figura 13: Porcentaje de varianza explicada por dimensión (Interactivo).

Una vez seleccionado el número de dimensiones a extraer, se analiza la contribución a cada una de ellas por parte de los perfiles fila y perfiles columna. Comenzando por las comunidades, se puede observar que la mayor contribución en la primera dimensión corresponde con Melilla y en la segunda con Aragón. Seguidos a estos, también con un importante peso, se encuentran las regiones de Galicia y Madrid para la primera componente y Cataluña y Valencia para la segunda. Toda esta información es muy relevante a la hora de entender y extraer conclusiones de Fig. 16.

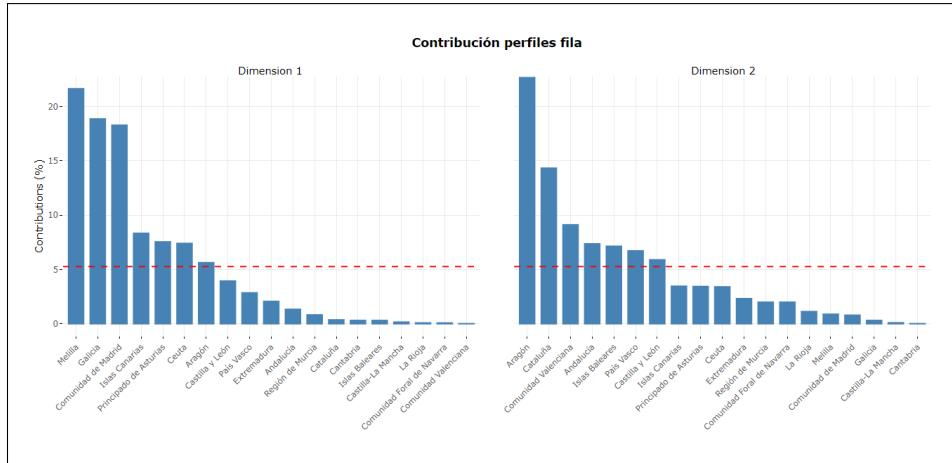


Figura 14: Contribución de los perfiles fila en las dimensiones escogidas (Interactivo).

Respecto a las edades, correspondientes con los perfiles columna, para la primera dimensión todos los niveles contribuyen de forma bastante significativa a excepción de las edades de 31-40 años. Mientras que para la segunda componente destaca la amplia contribución de la categoría de 14-17 años.

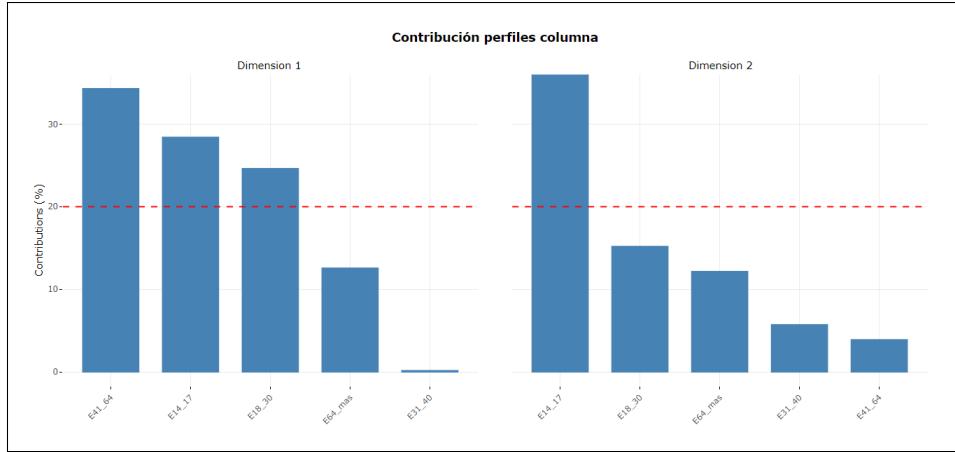


Figura 15: Contribución de los perfiles columna en las dimensiones escogidas (Interactivo).

Llegados a este punto, en el que se ha escogido el número de dimensiones a extraer y estudiado su construcción mediante las contribuciones de los perfiles, se procede a analizar el *biplot* (Fig. 16) correspondiente de las covariables. Aquellos niveles que se encuentren próximos al centro de coordenadas significa que son los que menos aportan a la definición de ambas dimensiones (e.g. Castilla-La Mancha). Ocurre lo mismo para la construcción de cada una de las dimensiones por separado, es decir, próximos a la recta $x = 0$ (no aportan a la definición de la segunda componente como Madrid o Cantabria) o $y = 0$ (no aportan a la definición de la primera componente como Valencia). Por contra, las categorías más alejadas se caracterizan por su gran contribución en la elaboración de las dimensiones. Es remarcable que se rechaza la hipótesis de independencia entre las variables edad y comunidad ya que el valor del estadístico chi-cuadrado es de 4943.211, suponiendo esto un p-valor igual a cero.

Además se observa cómo la primera dimensión discrimina por edades jóvenes y ancianas, correspondiendo el sector negativo con las personas más avanzadas en edad. Mientras que para la segunda componente se aprecia cómo separa entre edades extremas, situándose las edades menores y mayores en la región positiva. Yendo más allá, se encuentra una posible asociación entre la región de Aragón y las edades más jóvenes (14-17 años), pudiendo concluir que en dicha comunidad la delincuencia debido a menores de edad es mayor que en el resto de regiones. Ambas categorías están bien representadas ya que, en el caso de Aragón, dispone de una buena representación en ambas dimensiones con valores del coseno al cuadrado superiores a 0.5, mientras que las edades de 14-17 años tan solo disponen de buena representación en la primera componente (0.67)⁴.

Buscando asociaciones entre distintos territorios, se observa la clara cercanía entre Navarra y las Islas Baleares (ambas con altas contribuciones relativas) significando que ambas poseen perfiles de edad criminológicos similares. A priori, esto puede resultar chocante debido a que son regiones distintas (peninsular-insular) aunque cabe recordar que ambas comarcas se encuentran entre las comunidades con mayor número de delitos sexuales, además de ser de las provincias con mayor tasa de delitos (Fig. 4).

⁴Cabe recordar que tanto la contribución relativa (valor entre 1-100) de una categoría como su coseno al cuadrado (valor entre 0-1) son indicadores de la calidad de representación de la categoría, siendo mejor ésta cuanto mayor sean los valores de dichos indicadores.

También puede distinguirse, aunque sus contribuciones relativas sean menores, una cierta proximidad entre territorios pertenecientes al interior de la península como son Castilla y León, Extremadura y La Rioja, traduciéndose esto en que dichas regiones comparten cierta semejanza en la distribución delictiva sobre las distintas edades.

Profundizando más al respecto, se advierte que en territorios como Cataluña, Andalucía o Madrid destaca la delincuencia en edades intermedias con respecto a otras comarcas. Mientras que en los extremos de la edad, se encuentra el territorio asturiano liderando en tasa criminológica para ancianos y las ciudades de Melilla y Ceuta en menores.

Finalmente, citar que dichos procedimientos se ejecutan mediante las funciones *CA*, *fviz_screeplot*, *fviz_contrib* y *fviz_ca_biplot* de las librerías *FactoMineR* y *factoextra*.

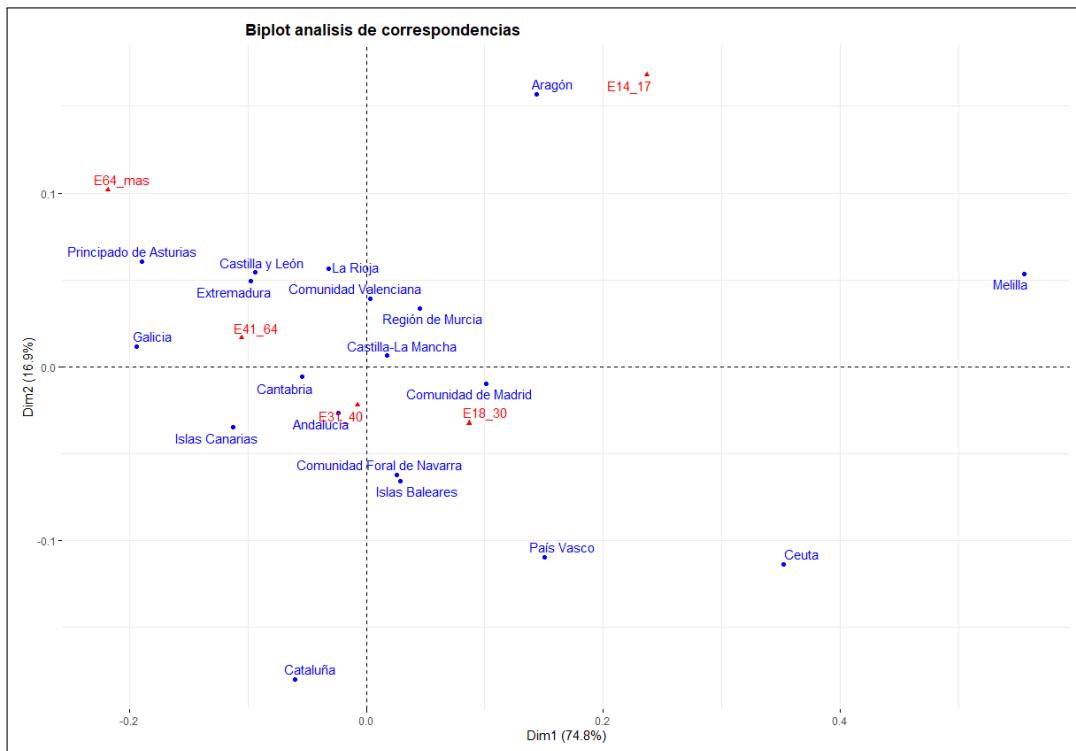


Figura 16: Distribución de los niveles de las covariables.

6.2. Análisis de componentes principales

El análisis de componentes principales es un método de análisis no supervisado cuya principal aplicación es la reducción de la dimensionalidad de conjuntos de datos de gran tamaño, tratando de perder la menor cantidad de información posible. Cuando se dispone de un gran número de variables cuantitativas, estando posiblemente correlacionadas entre ellas, *ACP* permite reducirlas a un número menor de variables transformadas e incorreladas entre sí (componentes principales) que expliquen la mayoría de la variabilidad en los datos. Dichas componentes principales corresponden con una combinación lineal de las variables originales. Además, se trata de una herramienta muy útil para la visualización y representación de los datos. Aportando una visión geométrica, esta técnica puede considerarse una rotación de los ejes del sistema de coordenadas de las variables originales a unos nuevos ejes ortogonales, de manera que éstos coincidan con la dirección de máxima varianza de los datos.

A continuación, se va a aplicar un *ACP* al conjunto de datos en el que cada comunidad corresponde con un individuo y las variables asociadas son las tasas de los distintos tipos de delitos por cada 100.000 habitantes durante el año 2018. A razón de que el cálculo de las componentes principales depende de las unidades de medida y la magnitud de las variables, resulta aconsejable estandarizarlas para que tengan media 0 y desviación estándar 1, puesto que en caso contrario, las de mayor varianza dominarían al resto.

Una vez ajustado el modelo, es necesario decidir el número de componentes que se van a extraer. Observando el *scree-plot* (Fig. 17) se ve cómo las tres primeras componentes explican más del 80 % de la variabilidad de los datos, además de ser las únicas componentes cuyos autovalores son superiores a la unidad (5.49, 3.01 y 2.06 respectivamente).

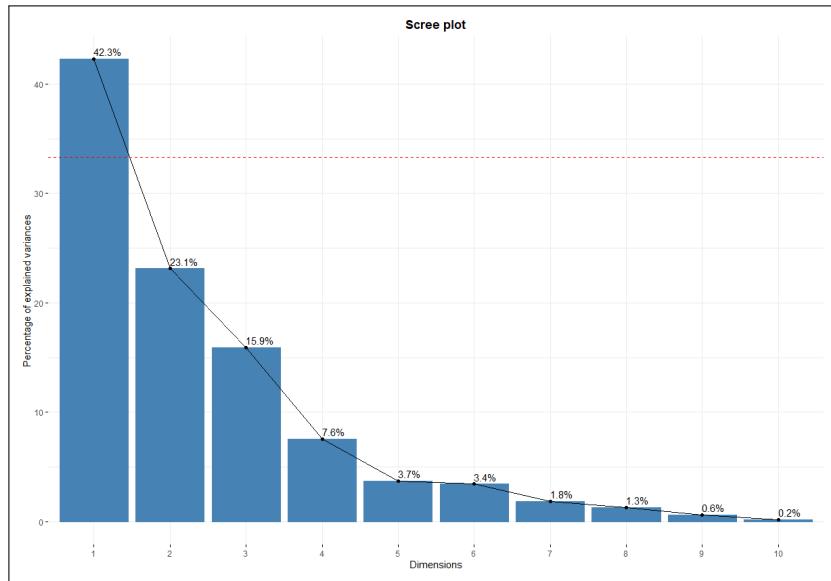


Figura 17: Porcentaje de la varianza explicada por componente (Interactivo).

Aunque el número más adecuado de dimensiones sea tres, los siguientes gráficos tan solo muestran las dos primeras para poder extraer conclusiones gráficamente de forma más sencilla. En la Fig. 18 se expone la nube de variables en la que se observa cómo todas tienen el mismo signo para la primera componente, por lo que tienen la misma tendencia, interpretándose esto en que dicha componente mide la seguridad de una región. Destaca que los delitos con mayor peso y mejor representados son los que atentan contra la libertad e indemnidad sexual, seguidos del tráfico de drogas. Respecto a la segunda componente no se aprecian conclusiones claras aunque sí es posible afirmar que separa delitos en función de su gravedad y naturaleza.

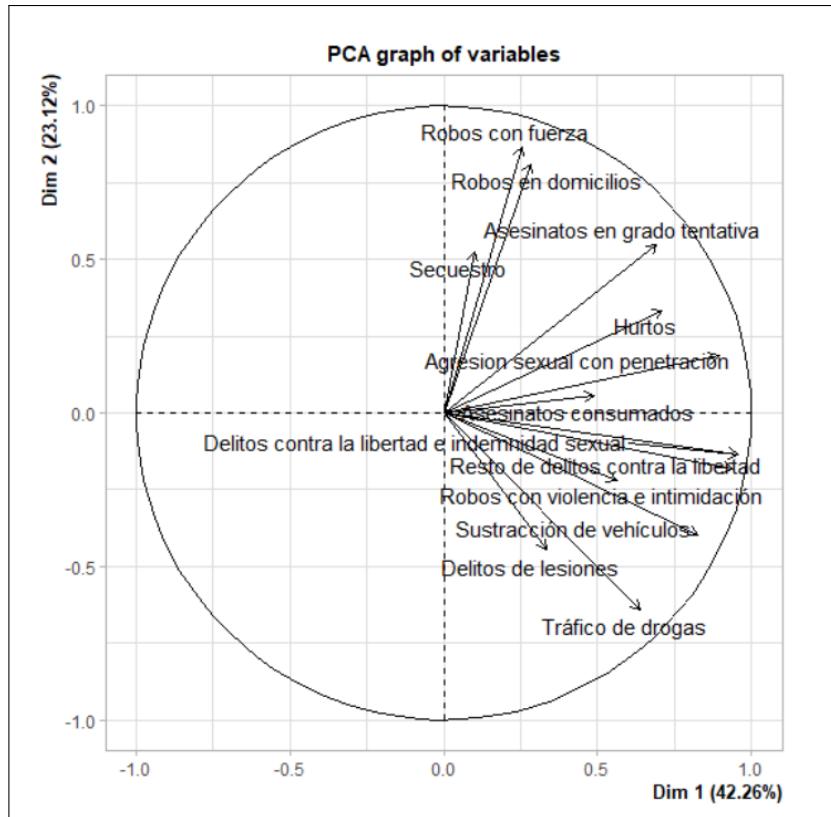


Figura 18: Nube de variables.

En el siguiente *biplot* (Fig. 19) se representan conjuntamente la nube de individuos y la de variables. Se quiere remarcar que el eje inferior e izquierdo representan la escala de valores de las puntuaciones de las observaciones, mientras que el eje superior y derecho representan la escala de los *loadings*, comprendida en un rango [-1, 1].

Pese a la mezcolanza de etiquetas, se puede observar cómo las comunidades con menor grado de delincuencia son Castilla y León, Asturias, Extremadura, La Rioja, Galicia, Aragón y Cantabria, estando altamente asociadas entre ellas (significando que dichas regiones poseen similitudes en la distribución tipológica de los delitos cometidos); Castilla-La Mancha también se puede añadir a este grupo de baja criminalidad aunque en menor medida. Mientras que en el otro extremo del espectro, se encuentran territorios como las Islas Baleares, Ceuta, Navarra y Cataluña.

También se distingue una asociación entre Cataluña y los hurtos, uno de los principales motivos puede deberse a la gran cantidad de turistas que recibe (convirtiéndose en fáciles objetivos de posibles carteristas). En suma, se aprecia como las Islas Baleares y Navarra están bastante asociadas con las agresiones sexuales, cosa que ya se había comentado previamente (Tabla 4). Además, de forma bastante precisa, podría establecerse una relación entre el delito de tráfico de drogas con las ciudades autónomas de Melilla y Ceuta.

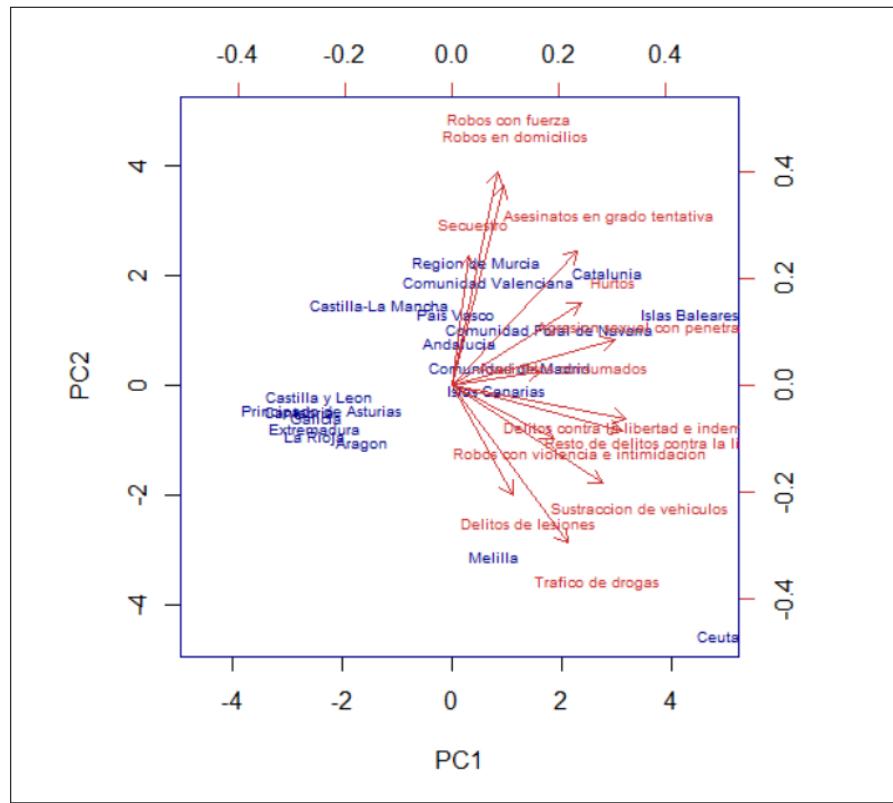


Figura 19: Nube de variables e individuos.

Como bien se ha comentado previamente, las tres primeras componentes principales explican más del 80 % de la variabilidad de los datos. Por este motivo, a continuación, se muestra un gráfico 3D en el que se representa la nube de variables e individuos. Remarcar que las etiquetas de las comunidades se han incluido pese a que dificultaban la visión de la ilustración. La elipse está construida a un nivel de confianza del 75 %. Se pueden apreciar conclusiones a mayores de las ya extraídas con los gráficos 2D de las dos primeras componentes, por ejemplo, se corrobora que las componentes sirven para discriminar entre los territorios más seguros y los más delictivos. En suma, enfrentando la primera contra la tercera componente, se concluye cómo esta última discrimina por tipología de delito entre hurtos y robos (en cualquiera de sus modalidades) frente al resto (a excepción de delitos de lesiones). Esta tercera componente podría interpretarse como un diferenciador según la gravedad y naturaleza de la infracción. Al igual que antes, se reafirma la clara asociación entre Cataluña y los hurtos o las Islas Baleares y las agresiones sexuales con penetración, aunque no ocurre lo mismo entre Navarra y este último delito, sino más bien se aprecia una posible relación entre este territorio y los asesinatos consumados. Por último, se remarca que podría existir una asociación entre Ceuta y el tráfico de drogas y entre Melilla y los delitos de lesiones.

Al tratarse de una representación en tres dimensiones, es recomendable acceder al modo interactivo para poder rotar y hacer zoom con el fin de poder observar cómo se distribuyen las variables e individuos respecto a las componentes principales. Este gráfico se realiza mediante la función *pca3D* de la librería *pca3D* y se exporta a un *html* local mediante la función *writeWebGL* de la librería *rgl*.

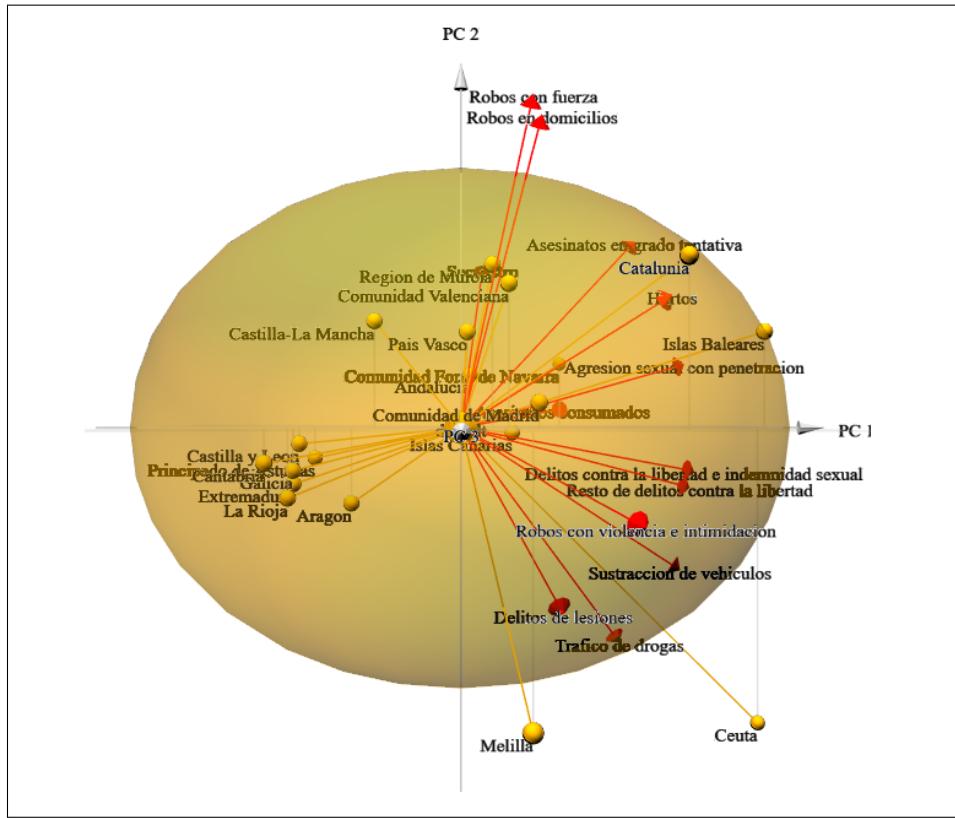


Figura 20: Representación 3D de la nube de variables e individuos (Interactivo).

6.2.1. Clasificación mediante K-Medias

Mediante el análisis de componentes principales se han extraído una serie de conclusiones sobre la similitud entre algunos territorios. Esto se traduce en que para dichas regiones existe una similaridad en el número y tipo de delitos cometidos por sus habitantes. Buscando confirmar dichas deducciones se aplica un método de clustering no supervisado como es el *K-Medias*.

De forma muy resumida, esta técnica se basa en la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo centroide es más cercano, se trata de un proceso iterativo. Al ser un agrupamiento no supervisado el primer paso es decidir el número de clusters. Para este cometido se utiliza el ‘método del codo’ aplicado al porcentaje de varianza explicada en función del número de clusters empleados. Así, usando este criterio, se observa si añadir un cluster mejora significativamente o no el ajuste del modelo. Gracias a Fig. 21 se aprecia cómo el número óptimo de clusters es igual a tres.

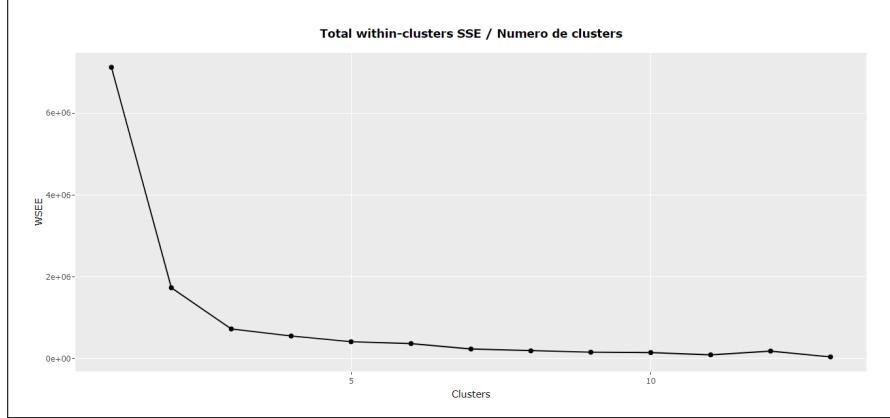


Figura 21: Relación entre número de clusters y varianza explicada (Interactivo).

A partir de aquí, mediante la función *kmeans* de *R*, se aplica un *K-Medias* de 3 clusters. En el siguiente gráfico (Fig. 22), se representan las comunidades en las tres primeras componentes principales y están coloreadas en función del grupo al que se han asignado. Además, el tamaño de cada burbuja depende del número medio de delitos de cada una de las regiones, es decir, a mayor volumen mayor es la delincuencia en ese territorio.

Las conclusiones obtenidas de este procedimiento apenas distan de las ya extraídas previamente. El primer cluster está compuesto por comunidades como Cataluña, Islas Baleares o Madrid; tratándose del grupo con mayor tasa de criminalidad. Mientras que por el contrario, el segundo cluster está formado por las regiones más seguras (en su mayoría interiores) como son Castilla y León, Aragón o Asturias; cabe remarcar que estas agrupaciones ya habían sido descubiertas en el apartado anterior. Finalmente, las ciudades autónomas de Ceuta y Melilla, que pese a tener un alto índice de delincuencia similar al primer cluster, conforman un grupo independiente. Esto principalmente se debe al número y tipo de los delitos allí cometidos, distintos del resto de territorios.

Se aconseja acceder al gráfico interactivo para poder rotarlo y así observar mejor los cluster.

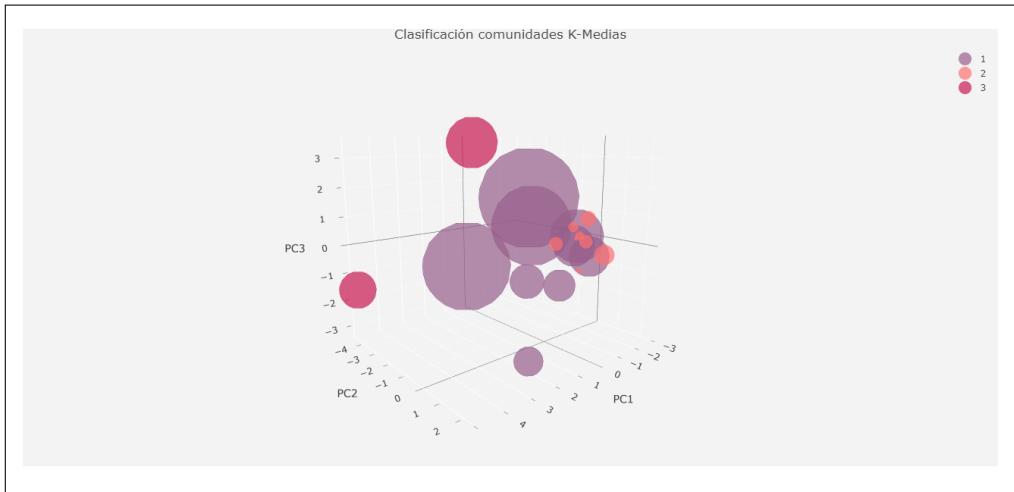


Figura 22: Clasificación comunidades (Interactivo).

7. Fase de modelado

Pese a que los datos estudiados ya han sido descritos previamente, a continuación se presenta la estructura final del conjunto de datos, obtenido tras las técnicas de procesado, que se aplica en los modelos (siendo el número de infracciones la variable objetivo y el resto las variables explicativas):

Infracciones	Sexo	Edad	Año
15285	Masculino	14-17	2018
3385	Femenino	14-17	2018
98793	Masculino	18-30	2018
21583	Femenino	18-30	2018
...

Antes de aplicar ningún modelo, se estudia si la distribución del número de infracciones es normal. Con el fin de comprobarlo, se emplea tanto el test de *Kolmogorov-Smirnov* como el test de *Lilliefors*. En ambos casos se rechaza la hipótesis nula de normalidad. Además, se refuerza este resultado mediante el gráfico de cuantiles teóricos (Fig. 23) donde se compara los cuantiles de la distribución observada con los cuantiles teóricos de una distribución normal (con misma media y desviación estándar que los datos). Es claramente visible que no se adecuan a la recta $y = x$ por lo que los datos no se aproximan a una distribución normal.

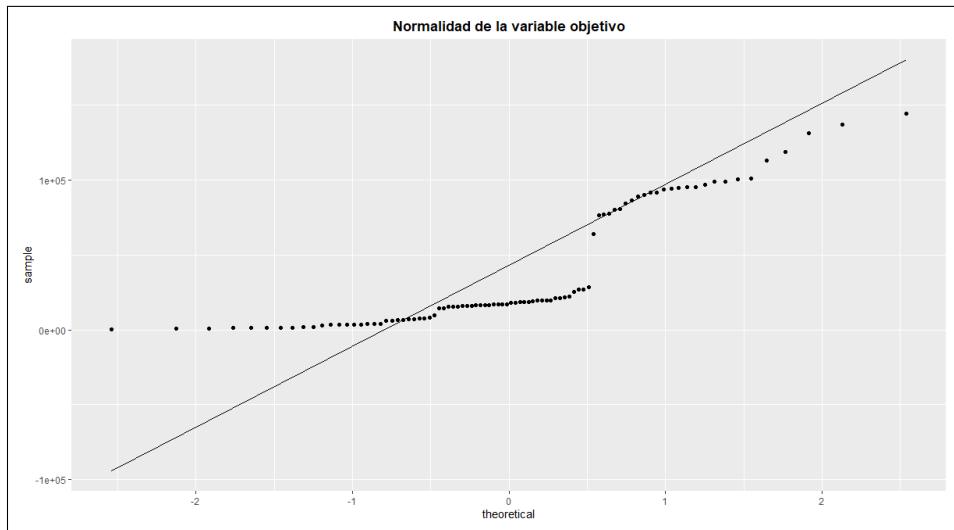


Figura 23: Gráfico Q-Q para la variable respuesta.

Para reforzar el rechazo de normalidad se representa un histograma de los datos, Fig. 24, en donde no se aprecia una distribución suave que se asemeje a la campana de *Gauss*. Además se observa una clara división en dos grupos, una con menor número de infracciones que otro. Esto se debe principalmente a la diferencia de delincuencia atendiendo al sexo y a la población de las regiones.

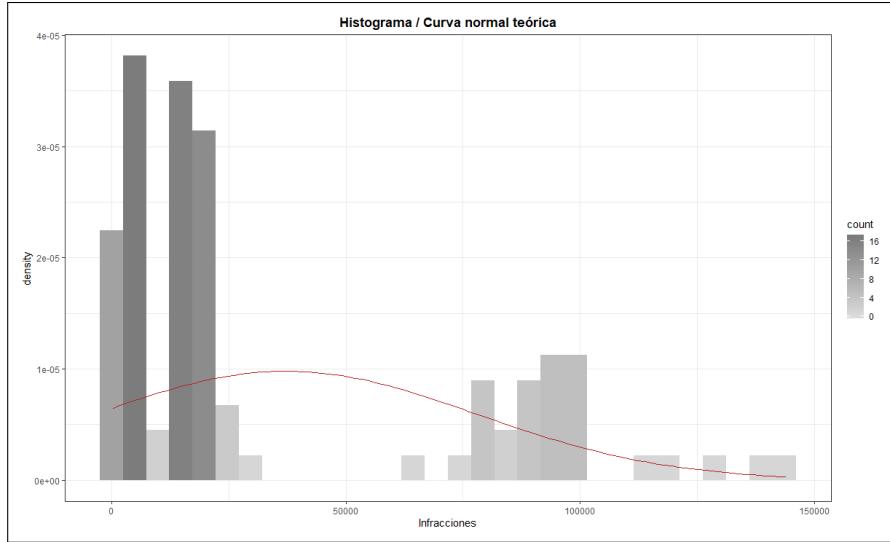


Figura 24: Histograma para estudiar la normalidad de la variable respuesta.

Mediante estos resultados se descarta que los datos se ajusten a una única distribución normal, pero no la existencia de una mezcla de distribuciones normales cuyos parámetros dependan de los valores de algunas variables explicativas. Como consecuencia, se enseña seguidamente los gráficos de cuantiles teóricos asociados a los distintos grupos que conforman el sexo y la edad:

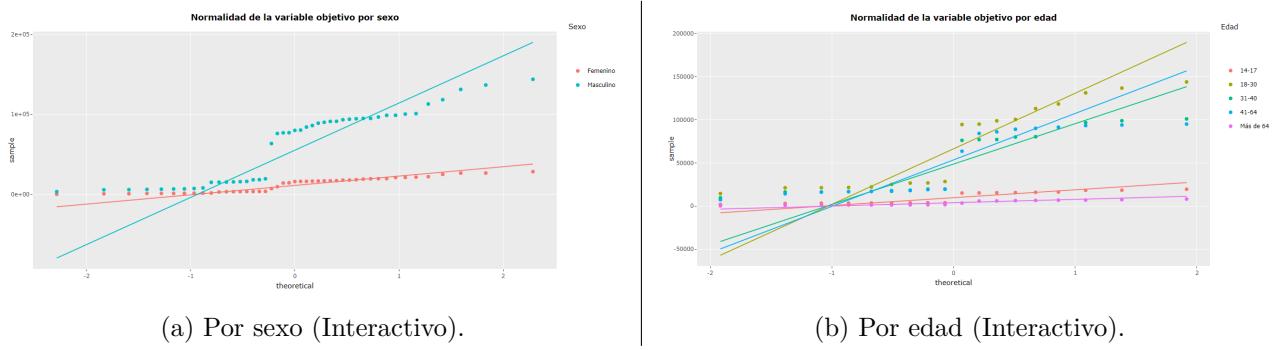


Figura 25: Gráfico Q-Q para la variable respuesta por grupos.

Para los grupos formados por el sexo, es observable la falta de normalidad para los hombres aunque para las mujeres no sería tan obvia dicha deducción. Lo mismo ocurre para los factores que conforman la edad, en todos los casos se rechaza la normalidad de forma clara salvo para los grupos de 14-17 años y de más de 64 años. Para apreciar bien estas conclusiones, es recomendable emplear el gráfico interactivo clickando en cada uno de los niveles para observarlos de forma independiente. Debido a estas posibles dudas, se va a ajustar un modelo de análisis de la varianza conformado por dichos factores para estudiar la normalidad de los residuos.

7.1. Análisis de la varianza

El análisis de la varianza sirve para comparar las medias de varios grupos (en este caso los formados por los factores edad y sexo) en una variable cuantitativa (el número de infracciones). En el *anova* ajustado no se tiene en cuenta las interacciones entre ambos factores, quedando así definido por la siguiente fórmula:

$$y_{ij} = \mu + Sexo_i + Edad_i + \epsilon_{ij}$$

y_{ij} : valor observado j-ésimo del nivel i-ésimo

μ : media general de la variable respuesta

$Sexo_i$: factor que determina los niveles según sexo

$Edad_i$: factor que determina los niveles según edad

ϵ_{ij} : error aleatorio

Se aplica este modelo para esclarecer el estudio de la existencia de normalidad, puesto que en caso afirmativo los residuos resultantes deberían ser normales. Finalmente, se puede corroborar mediante Fig. 26 la falta de normalidad en los residuos, se aprecia una cierta tendencia parabólica en su distribución.

Debido a estos resultados, donde ni la variable objetivo ni los distintos factores siguen una distribución normal, no es adecuado pensar en aplicar un modelo lineal general (e.g. regresión simple, regresión múltiple, análisis de la varianza, etcétera). Por consiguiente, en la siguiente sección [7.2-Modelo de Poisson], se ajusta un modelo lineal generalizado (*GLM*) donde la variable respuesta sigue una distribución perteneciente a la familia exponencial de dispersión.

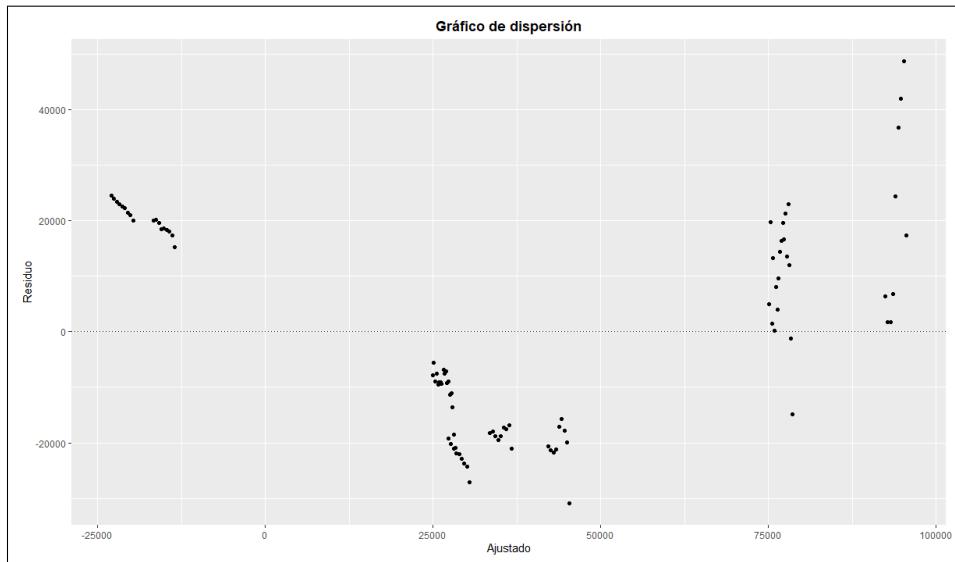


Figura 26: Residuos frente a predichos del modelo de análisis de la varianza.

7.2. Modelo de Poisson

La distribución de Poisson es frecuente en aquellos modelos donde se cuenta la ocurrencia de un evento, en este caso, el registro de una infracción. Antes de avanzar en el estudio cabe recordar ciertas características de dicha distribución [10]:

- **Función de probabilidad:** proporciona la probabilidad de que el evento ocurra y veces. El parámetro μ es positivo y representa el número de ocurrencias esperadas del evento, en nuestro caso está descrito en valores absolutos del número de infracciones cometidas.

$$P(Y = y) = \frac{e^{-\mu} \cdot \mu^y}{y!} ; y = 0, 1, 2, \dots$$

- **Esperanza y varianza:** tanto el valor esperado como la varianza de dicha distribución coinciden y son iguales a μ .

$$E(Y) = Var(Y) = \mu$$

- **Función de enlace:** como la distribución de la variable respuesta no es normal, es necesario buscar una forma de enlazar las distintas covariables de las observaciones (X) con el valor esperado de la variable de interés (μ). Es decir, proporciona la relación entre el predictor lineal y la media de la función de distribución.

$$\begin{aligned} E(Y) &= \mu / \mu \neq X^t \cdot B \\ &\downarrow \\ f(\mu) &= X^t \cdot B / \mu = f^{-1}(X^t \cdot B) \end{aligned}$$

- **Función de enlace canónica:** existen muchas funciones de enlace de uso común, y su elección se basa en distintos criterios. La función de enlace canónica es aquella que expresa θ (el parámetro canónico) en términos de μ , es decir, $\theta = f(\mu)$. En el caso de la distribución de Poisson dicha función es el logaritmo, la cual será usada en este estudio.

$$\log(\mu) = X^t \cdot B \rightarrow \mu = e^{X^t \cdot B}$$

También es importante especificar las covariables que participan en el modelo para tratar de describir la naturaleza del concepto. Las elegidas son la edad y el sexo del infractor, además del año en el que se ha cometido el delito. Primero se estudiará el efecto de cada variable explicativa por separado, observando su significancia. Posteriormente, se irán obteniendo nuevos modelos anidados, ya sea mediante adición de covariables o nuevas interacciones entre ellas, sobre los que se realizará un test *Chi-Cuadrado* para ver cuál es más adecuado. Estos procedimientos se obtienen mediante las funciones *glm* y *anova* en *Rstudio*.

- Solo variable explicativa sexo:

	Estimación	Std Error	P-valor
Intercept	9.402	0.001	<2e-16
SexoMasculino	1.637	0.002	<2e-16

Para ajustar este modelo se toma como clase base el sexo femenino. Se distingue cómo todos los parámetros son significativamente distintos de cero, por lo que su influencia sobre la respuesta es relevante. También se puede concluir, mediante el coeficiente del sexo masculino (siendo éste positivo), que la relación entre el número de delitos y los hombres es positiva. Este hecho significa que el número de delitos esperado aumenta si el sexo es masculino y, además, que es mayor que en el caso del sexo femenino.

- Solo variable explicativa año:

	Estimación	Std Error	P-valor
Intercept	32.262	0.426	<2e-16
Año	-0.011	0.0002	<2e-16

En este caso se ajusta el modelo con una variable numérica discreta, obteniendo que los parámetros son significativos a un nivel 0.05 de confianza, aunque la influencia del año es reducida debido a que su coeficiente es próximo a cero.

- Solo variable explicativa edad:

	Estimación	Std Error	P-valor
Intercept	9.211	0.002	<2e-16
Edad 18-30	1.928	0.002	<2e-16
Edad 31-40	1.640	0.002	<2e-16
Edad 41-64	1.643	0.003	<2e-16
Edad +64	-0.975	0.004	<2e-16

En esta ocasión se modeliza tomando a los más jóvenes como clase base, a saber, los comprendidos entre 14 y 17 años. Se concluye que el factor edad es significativo para explicar la variable objetivo. En suma, puede observarse que para las edades más jóvenes la relación con el número de infracciones es positiva, es decir, cuanto más joven sea una persona más probabilidad tiene de cometer un delito. Mientras que los ancianos (+64 años), tienen una relación negativa, traduciéndose esto en una disminución delictiva conforme se adentra en la vejez. Siendo esto totalmente coherente con la realidad.

Mediante estos resultados se comprueba que todas las variables estudiadas son significativas por separado aunque se ha detectado la existencia de sobredispersión, por lo que no son conclusiones totalmente fiables. Esto se puede apreciar en que la varianza de la distribución es mayor que la media, traduciéndose en un exceso de deviance residual que no puede ser explicada por la distribución. Este fenómeno ocurre debido a que los grupos de la distribución de Poisson se corresponden con distintas combinaciones de variables categóricas. Un método para solucionarlo es utilizar la distribución Binomial Negativa usando el comando *glm.nb* de la librería *MASS*, con el que se va a repetir el mismo procedimiento para estudiar si hay cambios en las conclusiones obtenidas.

- **Solo variable explicativa sexo:**

	Estimación	Std Error	P-valor
Intercept	9.402	0.138	<2e-16
SexoMasculino	1.637	0.196	<2e-16

En esta ocasión prácticamente se elimina la sobredispersión y los parámetros siguen siendo significativos, extrayéndose las mismas conclusiones que en el caso anterior. Solo varía ligeramente las estimaciones de los errores estándar.

- **Solo variable explicativa año:**

	Estimación	Std Error	P-valor
Intercept	33.254	94.754	0.726
Año	-0.011	0.047	0.810

Ahora que se ha paliado la sobredispersión, para el parámetro año no puede rechazarse la hipótesis nula, por lo que su influencia en la respuesta no sería significativa. Podría convertirse relevante en presencia de otros factores.

- **Solo variable explicativa edad:**

	Estimación	Std Error	P-valor
Intercept	9.211	0.182	<2e-16
Edad 18-30	1.928	0.258	8.38e-14
Edad 31-40	1.640	0.258	2.15e-10
Edad 41-64	1.643	0.258	1.99e-10
Edad +64	-0.975	0.258	0.00016

Al eliminar la sobredispersión se observa cómo la estimación de los parámetros varía poco aunque sí su error estándar. Aun así, siguen siendo todos significativos a un nivel 0.05 y se extraen las mismas conclusiones que antes.

A continuación, se estudian modelos más complejos teniendo en cuenta no sobreajustar los datos. Aplicando el test *Chi-cuadrado* se puede elegir cuál es el más adecuado. En *R*, se puede calcular de forma automática gracias a la función *anova* o manualmente mediante:

$$M0 \subset M1$$

$$df = df.residual(M0) - df.residual(M1)$$

$$1 - pchisq(2 \cdot (logLik(M1) - logLik(M0)), df)$$

Los resultados obtenidos se muestran en la siguiente tabla:

Modelo	Df	Test	Estadístico	P-valor
[1] Sexo	88			
[2] Sexo+Edad	84	[1] vs [2]	286.842	0
[3] Sexo+Edad+Año	83	[2] vs [3]	7.212	0.007

Se ve claramente que el mejor modelo es el que contiene todas las variables. Una vez deducido esto, se va a estudiar modelos anidados en los que se consideren las posibles interacciones entre factores, teniendo en cuenta evitar el problema del sobreajuste. Mismamente, teniendo presente tres covariables sin interacciones entre ellas ya hay siete parámetros, estando el conjunto de datos compuesto por tan solo 90 observaciones.

Modelo	Df	Test	Estadístico	P-valor
[1] Sexo+Edad+Año	83			
[2] Sexo+Edad*Año	79	[1] vs [2]	27.907	1.302e-05
[3] Sexo*Edad+Año	79	[1] vs [3]	0.785	0.940
[4] Sexo*Año+Edad	82	[1] vs [4]	8.672	0.003
[5] Sexo*Edad*Año	70	[1] vs [5]	43.605	3.566e-05

Entre los modelos estudiados, se aprecia que todas las interacciones son significativas, a excepción del sexo y la edad (3). Además, pese a la significancia del test *Anova*, se descartan tanto el caso (5) como el (2) porque se tratan de modelos no aceptables debido a que sobreajustan los datos y la mayoría de sus interacciones no son significativas a un nivel 0.05. Por consiguiente, se concluye que **el modelo más adecuado** es el formado por las covariables sexo, edad y año, considerando la interacción entre sexo y año (4). A continuación se muestran las estimaciones del modelo:

	Estimación	Std Error	Z Value	P-valor
Intercept	-92.99520	22.88716	-4.063	4.84e-05
Edad 18-30	1.93444	0.06547	29.545	<2e-16
Edad 31-40	1.63011	0.06548	24.896	<2e-16
Edad 41-64	1.61151	0.06548	24.611	<2e-16
Edad +64	-1.01871	0.06568	-15.511	<2e-16
Sexo Masculino	102.85239	32.32237	3.182	0.00146
Año	0.05019	0.01136	4.417	1.00e-05
SexoMasculino*Año	-0.05025	0.01605	-3.131	0.00174

Una vez encontrado el modelo más adecuado, en el que todas las covariables son significativas, es interesante estudiar cómo se relacionan con la variable de interés. Se puede ver que la clase base del sexo es la femenina, siendo su tendencia a cometer delitos mucho menor que la de los hombres; además hay una clara correlación positiva entre el sexo masculino y la criminalidad. En el caso de la edad, la clase base corresponde con los más jóvenes (14 a 17 años). Existe una correlación positiva entre todas las edades y el número de infracciones cometidas, salvo para las personas de más de 64 años ya que la estimación de su coeficiente es negativa, traduciéndose esto en una disminución de delincuencia conforme se adentra en la vejez. En el siguiente gráfico se puede apreciar de forma más ilustrativa las correlaciones positivas y negativas entre el número de infracciones y las distintas covariables, observando la enorme influencia del sexo masculino.

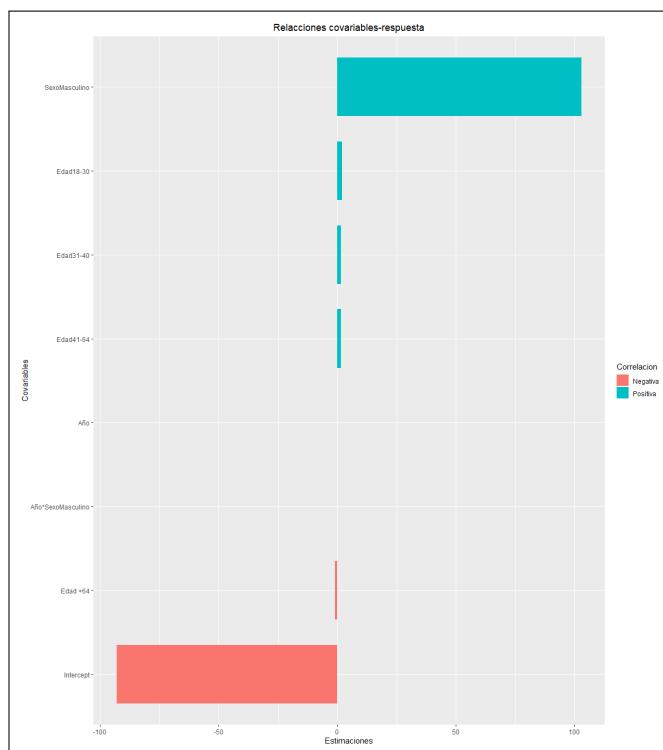


Figura 27: Estimaciones de los parámetros del modelo escogido (Interactivo).

7.2.1. Predicciones y residuales

Ahora que se ha escogido el modelo con la composición más adecuada, se pueden realizar predicciones para analizar la calidad de ajuste que se ha aplicado. Estas predicciones pueden ser para el predictor lineal (que se trataría del logaritmo del número de infracciones) o bien para los valores ajustados del número medio de infracciones. La Fig. 28 muestra el logaritmo del número real y el logaritmo del número predicho de infracciones frente a los años en los que se han cometido.

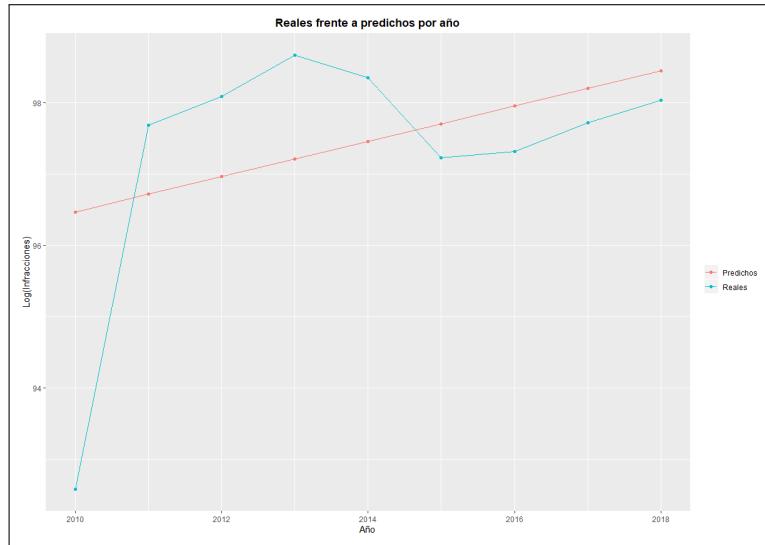


Figura 28: Predicciones para el logaritmo del número de infracciones por año.

Llegados a este punto, es importante estudiar la distribución de los residuos de los valores ajustados. Seguidamente se muestra un diagrama de dispersión (Fig. 29) en el que no se aprecia ningún patrón específico, distribuyéndose los residuos de forma aleatoria indicando cierta normalidad. Se comprueba la existencia de unos *outliers* que corresponden todos con casos femeninos.

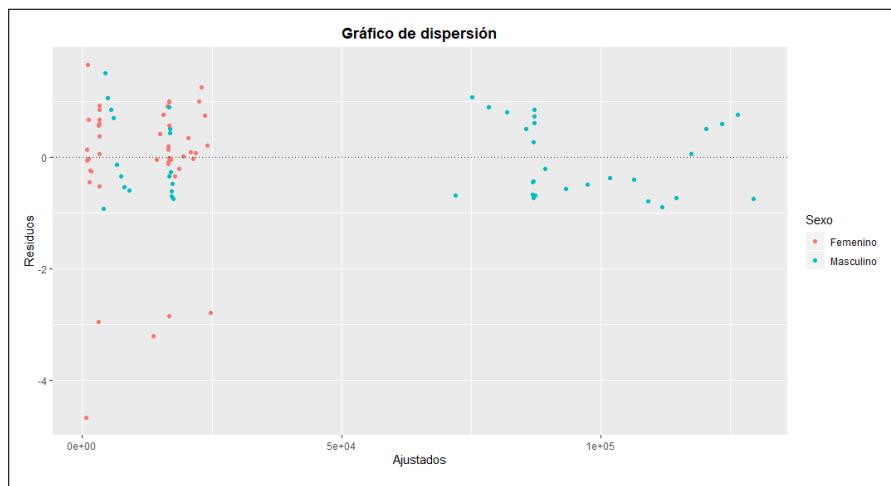


Figura 29: Residuos frente a predichos (Interactivo).

Además, para confirmar la normalidad de los residuos, se va a representar un gráfico Q-Q (Fig 30) en el que se aprecia cómo la mayoría de los residuos se ajustan a la recta $y = x$. Localizando los mismos *outliers* ya comentados en el gráfico anterior.

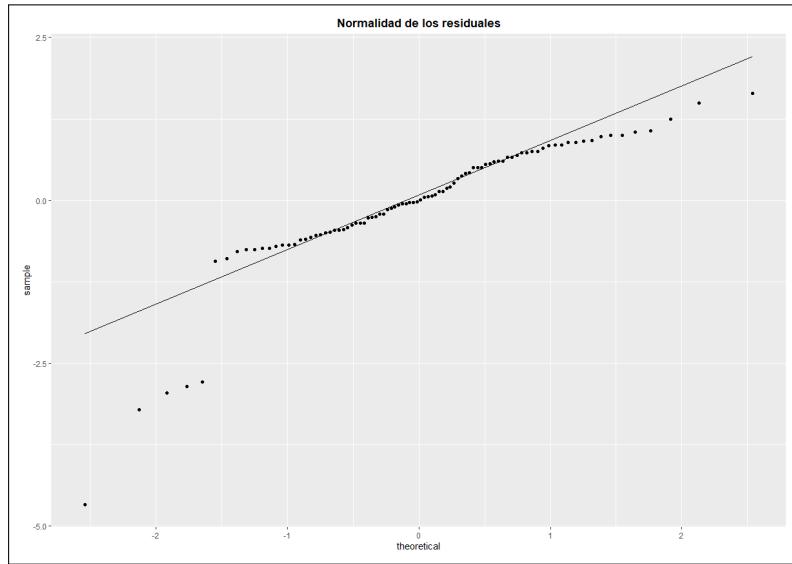


Figura 30: Qqnorm para los residuos.

7.2.2. Capacidad predictiva

Los resultados obtenidos en el subapartado anterior [7.2.1-Predicciones y residuales] son fruto de ajustar un modelo en donde se entrena el clasificador con todas las observaciones del conjunto de datos. En cambio, para poder analizar la capacidad predictiva es necesario dividir los datos en conjuntos de entrenamiento, prueba y validación. En este caso, al no disponer de muchos datos, se prescinde del conjunto de validación. El procedimiento llevado a cabo es un *cross validation 10-Fold* para obtener una estimación de la tasa de aciertos menos sesgada y más precisa que mediante *Hold-out*. Se remarca que se considera como éxito aquellas predicciones que no distan más del 15 % del total del valor real.

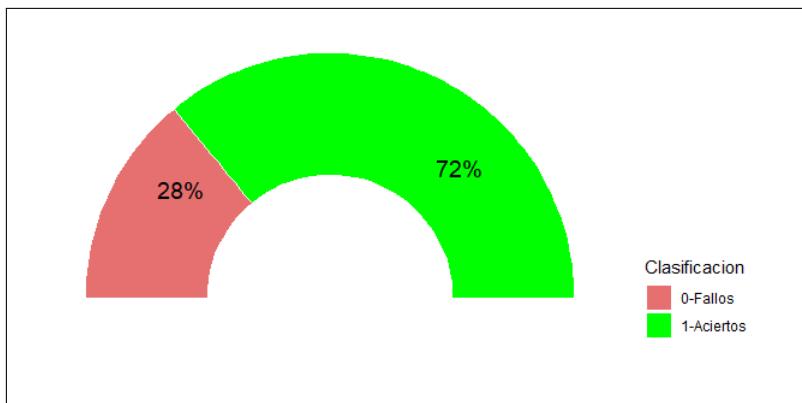


Figura 31: Porcentaje de acierto/error del modelo ajustado.

7.3. Redes neuronales recursivas

Las redes neuronales representan uno de los modelos computacionales con mayor aceptación en la actualidad, siendo una de sus variantes las redes neuronales recursivas. Esta recursividad supone que las neuronas se comuniquen consigo mismas, ya sea de forma directa o indirecta, creando ciclos en el grafo que define sus interconexiones. Este hecho hace que sean adecuadas para detectar patrones en secuencias temporales [11]. Aprovechando las capacidades de este tipo de redes, se pretende estudiar la posible existencia de regularidades en la ejecución de delitos a lo largo de los años.

Para lograr este objetivo, se obtienen las infracciones cometidas por comunidad durante cada trimestre desde el año 2015 hasta el tercer trimestre del año 2019. De esta forma, se crea una secuencia temporal de 361 valores. Pese a que el preprocesamiento se ha llevado a cabo con *R*, el ajuste de las siguientes redes neuronales recursivas se ha realizado con *Anaconda (Python)*, usando las funciones *SimpleRNN* y *LSTM* del módulo de *Keras*. El procedimiento llevado a cabo consiste en dividir la secuencia temporal en sucesiones más cortas, de igual tamaño al número de comunidades, y predecir el último valor de dichas sucesiones, repitiendo aleatoriamente el proceso durante un número determinado de épocas. La capacidad predictiva se analizará mediante *Hold-Out*.

7.3.1. Red neuronal recursiva simple

Este tipo de redes se caracteriza por concatenar la salida de una neurona a su propia entrada, consiguiendo así que su nueva salida dependa del estado anterior. El grafo asociado a una neurona de dicho sistema es representado mediante Fig. 32.

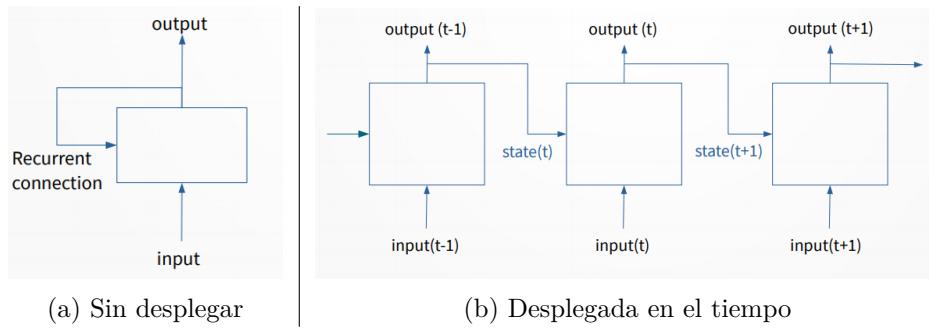


Figura 32: Grafo de interconexiones de una red neuronal recursiva simple

La red implementada consta de tres capas, estando compuesta cada una de ellas de 30, 10 y 5 neuronas respectivamente. Para analizar la validez de dicha red, se usa un tercio de los datos, seleccionados de forma aleatoria, como conjunto de prueba. De éste, se consideran instancias mal clasificadas aquellas cuyo valor predicho difiere más de un 15 % del valor real. Remarcando que para llevarlo a cabo se ha utilizado el optimizador *Adams* y la función de pérdida *MSE*.

Los resultados obtenidos durante 100 épocas, ilustrados en Fig. 33, muestran cómo se llega a conseguir un 70 % de éxito tanto en el conjunto de entrenamiento como en el de validación. Se trata de un porcentaje bastante bueno debido a que no se está analizando una secuencia temporal excesivamente amplia, en caso contrario, sería aconsejable ajustar otro tipos de redes como *LSTM* o *GRU*.

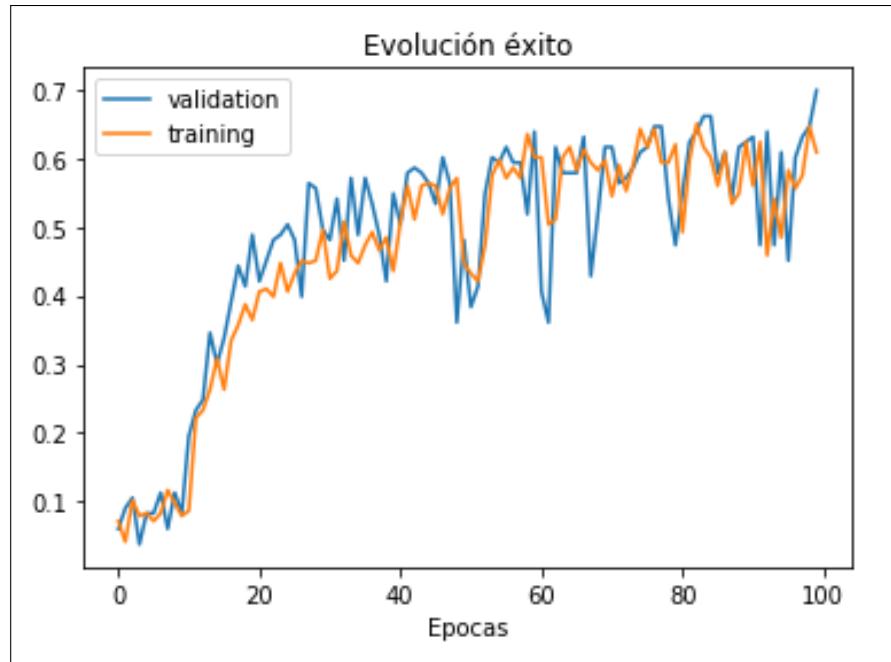


Figura 33: Tasa de éxitos en una red neuronal recursiva simple.

7.3.2. Red neuronal LSTM (Long Short-Term Memory)

A diferencia de la red explicada en el subapartado anterior [7.3.1-Red neuronal recursiva simple], el modelo *LSTM* se trata de una red neuronal recursiva de memoria a largo plazo. Este hecho la hace más adecuada para proyectos en los que se estudien secuencias temporales que abarquen grandes periodos de tiempo. Dicha ventaja sobre las redes recursivas simples es debida a que poseen una “cinta transportadora” que les permite recordar entradas anteriores incluso alejadas en el tiempo, y no limitarse tan solo a la última salida. Además, mediante *LSTM*, también se soluciona el problema de la evanescencia del gradiente (propia de secuencias temporales largas en las que la magnitud del gradiente va disminuyendo de una capa a otra hasta llegar a ser casi imperceptible en las más profundas). El grafo asociado a dicho modelo computacional es representado mediante Fig. 34.

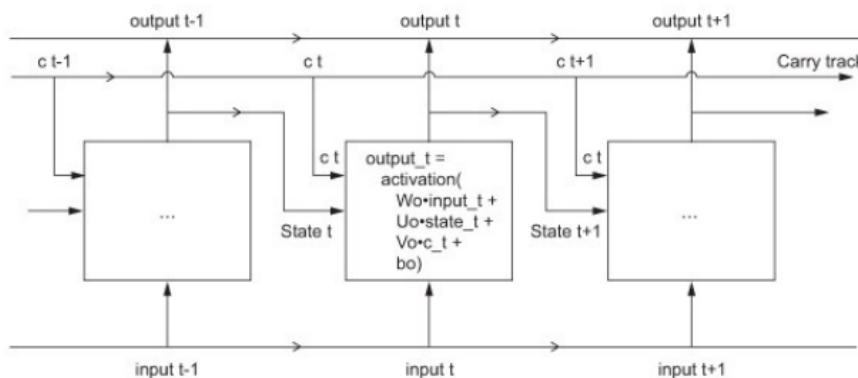


Figura 34: Grafo de interconexiones de una red neuronal LSTM

Al igual que antes, la red implementada consta de tres capas compuestas cada una de ellas por 30, 10 y 5 neuronas respectivamente. También se ha aplicado un *Hold-Out* en el que 1/3 de los datos se ha utilizado como conjunto de prueba y el criterio de clasificación es el mismo que el explicado previamente, habiendo usado idéntico optimizador y función de pérdida.

Los resultados obtenidos son peores que los conseguidos con la red neuronal recursiva simple, esto no es sorprendente puesto que la secuencia temporal con la que se ajusta el modelo no abarca una gran ventana temporal. Tras 100 épocas tan solo se alcanza un 45% de éxito tanto en el conjunto de entrenamiento como en el de validación, bastante inferior que en el modelo de memoria a corto plazo.

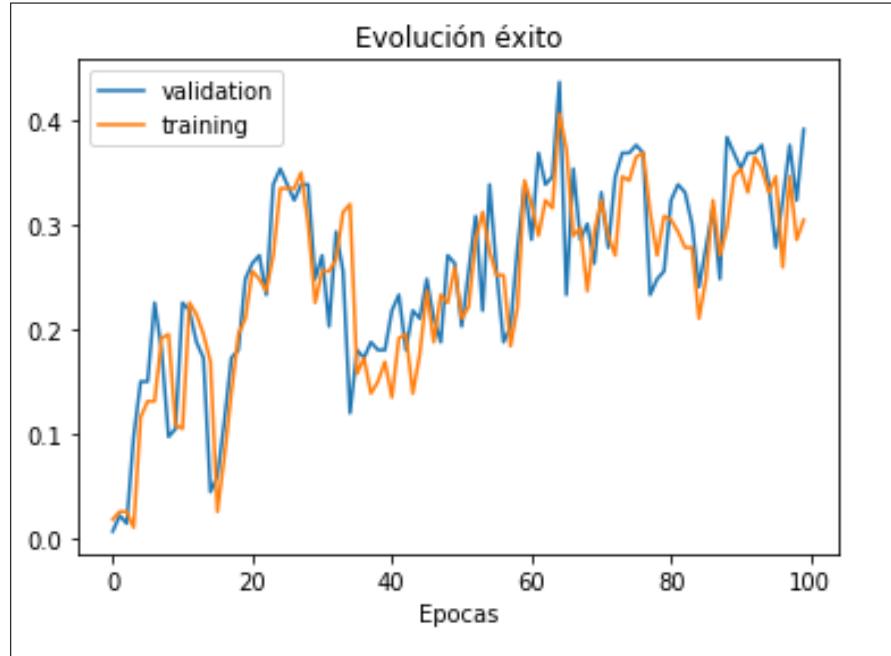


Figura 35: Tasa de éxitos en una red neuronal recursiva LSTM.

8. Futuras investigaciones

En primera instancia, todos los estudios y procedimientos desarrollados a lo largo del proyecto podrían ser extrapolables a niveles europeos o mundiales, de tal forma que permitiesen realizar comparaciones entre países de distintos continentes (teniendo en cuenta las limitaciones de dichas conclusiones debido a las diferencias culturales, jurídicas o sociales, entre otras muchas más, que distinguen a los distintas regiones del globo terráqueo). Esta nueva línea de investigación requerirá de un gran esfuerzo de procesamiento e integración de la información debido al grado de globalización que se quiere analizar, además de lo ineficientes y dispersos que suelen ser los datos abiertos en estos niveles.

Además, todos los aspectos aquí versados pueden ser estudiados con un mayor grado de profundidad y detalle, llegando incluso a un nivel comarcal en el que las conclusiones extraídas puedan ayudar a ciudades concretas en la toma de decisiones sobre materia policial y criminológica.

9. Conclusiones

El estudio realizado con este proyecto pretende abarcar el máximo número de aspectos, que se han considerado relevantes y actuales, con el mayor detalle posible. Resulta categórico remarcar cómo el género masculino tiende a infringir la ley con mayor habitualidad que las mujeres, llegando a ser el triple de detenidos e investigados que en el sexo femenino. Respecto a la edad, es coherente que los individuos jóvenes y de mediana edad sean los estratos con mayor tasa de criminalidad.

A nivel territorial, se ha obtenido tanto las regiones con más infracciones en valores absolutos (Madrid y Barcelona) traduciéndose esto en una mayor necesidad de fuerzas policiales, como aquellas con mayor tasa de criminalidad (observando cómo sufren un alto índice de delincuencia aquellos territorios de costa que reciben turismo ‘etílico’).

Utilizando indicadores como la tasa de paro y el PIB, se ha comprobado la influencia del factor económico en el ámbito delictivo, concluyendo que cuando las condiciones de vida de la población son óptimas la delincuencia disminuye considerablemente.

Respecto a la polémica situación de los MENAS, es complejo extraer deducciones rigurosas y fiables que aporten claridad debido al desconocimiento adherido de la propia cuestión. Por este motivo, a título personal, considero que no es responsable incendiar la opinión de la población mediante discursos de odio.

Otro aspecto, que personalmente considero muy relevante y tristemente actual, es el machismo intrínseco de la sociedad que se traduce en hechos como la violencia de género o el ‘techo de cristal’. En este proyecto se ha puesto de manifiesto tanto la diferencia entre hombres y mujeres en términos laborales, como la gran cantidad de mujeres que sufren violencia física y psíquica por parte del sexo contrario. Aunque los datos mostrados en la Tabla 3 y la Tabla 4 correspondan con agresiones sexuales sufridas por ambos géneros, es necesario recordar que en más del 85 % de los casos las víctimas son mujeres [14]. Con la información aquí obtenida, se pueden focalizar los esfuerzos de educación y concienciación en aquellos territorios que tienen mayores carencias al respecto, ya sea fomentando talleres o charlas orientadas sobre todo a gente joven; en definitiva, ellos son el futuro.

Mediante la fase de modelado se ha descubierto que el número de infracciones y detenidos se ajusta de forma bastante adecuada a un modelo de *Poisson*, mientras que se rechaza que siga una distribución *Normal*. En suma, gracias al análisis de correspondencias se han podido descubrir fenómenos como la asociación entre la delincuencia juvenil y la comunidad de Aragón. Otra idea destacable, aunque esta vez extraída a través del *ACP*, es que los territorios interiores (e.g. Castilla y León) son más seguros que los costeros (e.g. Islas Baleares). En la última fase del modelado, mediante las redes neuronales recursivas, se infiere que el número de infracciones a lo largo de los años puede ser modelada como una secuencia temporal para tratar de predecir el grado de delincuencia que habrá en tiempos futuros.

Finalmente, remarcar que las deducciones obtenidas en este estudio buscan poder ayudar a la sociedad a avanzar y prosperar; ya que, durante todo el desarrollo del proyecto, el principal objetivo ha sido poder aportar este pequeño y humilde grano de arena en pos de una mejora social.

Referencias

- [1] Roberto Alonso Ramos Erosa. 2015. Estadística: el preludio de la Criminología Científica.
Acceso: 15 de Febrero, 2020.
- [2] José Aureliano Martín Segura. 3º época, nº 1, 2009. La ciencia estadística y la criminología.
Revista de derecho penal y criminología de la UNED.
Acceso: 26 de Febrero, 2020.
- [3] David P. Farrington, Roger Tarling. 1985. Prediction in Criminology.
- [4] Datos abiertos del Ministerio de Interior del gobierno de España.
Acceso: Febrero y Marzo, 2020.
- [5] Instituto Nacional de Estadística.
Acceso: Febrero y Marzo, 2020.
- [6] Poder Judicial de España.
Acceso: Febrero y Marzo, 2020.
- [7] Portal de datos abiertos de la Unión Europea.
Acceso: Febrero y Marzo, 2020.
- [8] Portal de Eurostat.
Acceso: Febrero y Marzo, 2020.
- [9] GADM maps and data.
Acceso: 20 de Febrero, 2020.
- [10] Annette J. Dobson, Adrian G. Barnett. Mayo, 2008. An introduction to generalized linear models.
- [11] Juan Antonio Pérez Ortiz. Julio, 2002. Modelos predictivos basados en redes neuronales recurrentes de tiempo discreto. Tesis universidad de Alicante.
- [12] Las 10 noticias más relevantes de 2012 en España. - EcoDiario.
Acceso: 13 de Marzo, 2020.
- [13] Las noticias más destacadas de España en 2011, de la A a la Z. - Diario de Navarra.
Acceso: 17 de Marzo, 2020.
- [14] El 85,8% de las víctimas mortales a manos de su pareja o expareja son mujeres asesinadas por hombres. - El País.
Acceso: 17 de Marzo, 2020.
- [15] M. Grünhut. Part II, 1951. Journal of the Royal Statistical Society (Statistics in Criminology).
- [16] Santiago de la Fuente Fernández. 2011. Análisis correspondencias simples y múltiples. Universidad Autónoma de Madrid.

Parte II

Anexos

A. Aclaraciones

Los siguientes apartados se encuentran nombrados igual que en la documentación del proyecto, así es posible localizar los fragmentos de código de cada uno de los procedimientos llevados a cabo. Al ser el *pdf* un formato estático, los gráficos interactivos no pueden mostrarse. Debido a esto, a continuación se ofrece un link que redirige a un *html* que posee más información que este anexo ya que se trata de un entorno que permite ilustrar las representaciones interactivas y dinámicas. En consecuencia, en caso de disponer de internet, se recomienda utilizar el *notebook html* en vez del *pdf* para ver los anexos. El link es el siguiente:

- **Anexos-HTML**

Ambas documentaciones se han realizado mediante **Rmarkdown**.

B. Configuración Rmarkdown

```
---
```

```
title: "Anexos"
author: "Iván López de Munain"
#date: "`r format(Sys.time(), '%d%B, %Y')`"
always_allow_html: true
documentclass: article
lang: es
output:
  pdf_document:
    toc: true
    toc_depth: 3
    number_sections: true
    fig_caption: true
    df_print: kable
    keep_tex: true

---

knitr::opts_chunk$set(echo = TRUE, warning = F, message = F)
```

C. Librerías

```
library(plotly)
library(RColorBrewer)
library(kableExtra)
library(sp)
```

```

library(xlsx)
library(ggplot2)
library(networkD3)
library(dplyr)
library(tidyverse)
library(MASS)
library(scales)
library(caret)
library(FactoMineR)
library(factoextra)
library(pca3d)

```

D. Análisis descriptivo

D.1. Mapa interactivo de Europa

```

rm(list=ls())

# ===== Mapa interactivo Europa =====

datosEuropa<-read.csv("europaTasaCrimenes.csv", header=T, sep=",") 

kable(head(datosEuropa[,c("TIME","GEO","ICCS","Value")]),
      "latex", caption = "Primeras observaciones datos Europa",
      booktabs = T) %>%
kable_styling(latex_options = c("striped","hold_position"))

```

Cuadro 7: Primeras observaciones datos Europa

TIME	GEO	ICCS	Value
2015	Belgium	Intentional homicide	2.05
2015	Belgium	Attempted intentional homicide	7.83
2015	Belgium	Assault	614.32
2015	Belgium	Kidnapping	10.42
2015	Belgium	Sexual violence	60.00
2015	Belgium	Rape	28.59

```

#tasa por cada 100000 habitantes
datosEuropa <- datosEuropa %>%
  mutate(Value=as.numeric(as.character(datosEuropa$Value)))

levels(datosEuropa$GEO)<-
  c(levels(datosEuropa$GEO), "United Kingdom")

df17<- datosEuropa %>% filter(TIME=="2017")

```

```

df17$GEO[which(df17$GEO=="England and Wales")]<-"United Kingdom"
df17<- df17[order(df17$GEO),]

df16<- datosEuropa %>% filter(TIME=="2016")
df16$GEO[which(df16$GEO=="England and Wales")]<-"United Kingdom"
df16<- df16[order(df16$GEO),]

df15<- datosEuropa %>% filter(TIME=="2015")
df15$GEO[which(df15$GEO=="England and Wales")]<-"United Kingdom"
df15<- df15[order(df15$GEO),]

aa<-levels(df17$GEO)[-10]
aa[13]<-"Germany"
aa[19]<-"Kosovo"

a17<-tapply( df17$Value[!is.na(df17$Value)] , 
             df17$GEO[!is.na(df17$Value)],sum)
a16<-tapply( df16$Value[!is.na(df16$Value)] , 
             df16$GEO[!is.na(df16$Value)],sum)
a15<-tapply( df15$Value[!is.na(df15$Value)] , 
             df15$GEO[!is.na(df15$Value)],sum)

datos17<-data.frame(value=a17[-10], pais=as.factor(aa))
datos17<- datos17[order(datos17$pais),]

datos16<-data.frame(value=a16[-10], pais=as.factor(aa))
datos16<- datos16[order(datos16$pais),]

datos15<-data.frame(value=a15[-10], pais=as.factor(aa))
datos15<- datos15[order(datos15$pais),]

=====concatenacion de los textos=====
aux <- with(df17, paste('<br>->', ICCS, ":", Value))

i<-0
conca<-NULL
while(i < length(aux)){

  conca[(1+i/13)]<-paste(aux[i+1], aux[i+2],
                           aux[i+3], aux[i+4],
                           aux[i+5], aux[i+6],
                           aux[i+7], aux[i+8],
                           aux[i+9], aux[i+10],
                           aux[i+11], aux[i+12],
                           aux[i+13])
}

```

```

    i<-i+13
}

datos17$texto<-conca
datos17$texto<-paste("<br>Categorization of crimes in
2017:",datos17$texto)

#2016

aux <- with(df16, paste('<br>->', ICCS, ":", Value))

i<-0
conca<-NULL
while(i < length(aux)){

  conca[(1+i/13)]<-paste(aux[i+1], aux[i+2],
                           aux[i+3], aux[i+4],
                           aux[i+5], aux[i+6],
                           aux[i+7], aux[i+8],
                           aux[i+9], aux[i+10],
                           aux[i+11], aux[i+12],
                           aux[i+13])

  i<-i+13
}

datos16$texto<-conca
datos16$texto<-paste("<br>Categorization of crimes in
2016:",datos16$texto)

#2015

aux <- with(df15, paste('<br>->', ICCS, ":", Value))

i<-0
conca<-NULL
while(i < length(aux)){

  conca[(1+i/13)]<-paste(aux[i+1], aux[i+2],
                           aux[i+3], aux[i+4],
                           aux[i+5], aux[i+6],
                           aux[i+7], aux[i+8],
                           aux[i+9], aux[i+10],
                           aux[i+11], aux[i+12],
                           aux[i+13])

  i<-i+13
}

```

```

datos15$texto<-conca
datos15$texto<-paste("<br>Categorization of crimes in
2015:",datos15$texto)

updatemenus <- list(
  list(
    active = 0,
    type= 'buttons',
    showactive = F,
    buttons = list(
      list(
        label = "Año 2017",
        method = "update",
        args = list(list(visible = c(T, T, F,F))),
      list(
        label = "Año 2016",
        method = "update",
        args = list(list(visible = c(T,F, T,F))))
      ,list(
        label = "Año 2015",
        method = "update",
        args = list(list(visible = c(T,F, F,T))))
    )))
  )

cols=brewer.pal(9, "Reds")

plot_geo() %>%
  add_trace(z = ~value, data = datos17, locations = c("x"),
            locationmode= "country names", color = ~value,
            colors = cols,visible=T,showscale=T) %>%
  add_trace(z = ~value, data = datos17, text= ~texto,
            locations = ~pais, locationmode= "country names",
            color = ~value, colors =cols,
            visible=T,showscale=FALSE) %>%
  add_trace(z = ~value, data = datos16, text= ~texto,
            locations = ~pais, locationmode= "country names",
            color = ~value, colors =cols,
            visible=F,showscale=FALSE) %>%
  add_trace(z = ~value, data = datos15, text= ~texto,
            locations = ~pais, locationmode= "country names",
            color = ~value, colors = cols,
            visible=F,showscale=F) %>%
  layout(geo = list(scope="europe", projection=list(scale=1.6)),
         title="Número de delitos cometidos por
         cada 100.000 habitantes",

```

```

    margin=list(l=0,r=0,b=0,t=30),
    updatemenus=updatemenus) %>%
colorbar(title = "", thickness=10, which=1,
      yanchor="bottom", x=1, y=0.5)

```

D.2. Mapas España: valores absolutos y tasas

```

-----Mapa por provincias: valores absolutos-----

mapaProv<-readRDS('gadm36_ESP_2_sp.rds')

#comentado porque se usa para realizar mapa sin las Canarias
#y asi disponerlo más bonito

#mapaProv <- mapaProv[mapaProv$NAME_2!="Santa Cruz de Tenerife",]
#mapaProv <- mapaProv[mapaProv$NAME_2!="Las Palmas",]

datosProv <- read.xlsx('total-infrac-prov.xlsx', sheetIndex = 1)

infracProv <- which(datosProv[,2]!="<NA>")

#quitamos el total y el extranjero
infracProv <- fracProv[3:(length(infracProv)-1)]

#tabla de autoridad
provDefinit<-c("Álava", "Albacete", "Alicante" , "Almería",
               "Ávila", "Badajoz" , "Baleares",
               "Barcelona","Burgos","Cáceres", "Cádiz",
               "Castellón","Ciudad Real" , "Córdoba" ,
               "A Coruña", "Cuenca","Girona",
               "Granada", "Guadalajara" , "Guipúzcoa",
               "Huelva", "Huesca","Jaén" , "León",
               "Lleida","La Rioja" , "Lugo","Madrid",
               "Málaga","Murcia", "Navarra" , "Ourense",
               "Asturias", "Palencia", "Las Palmas",
               "Pontevedra", "Salamanca" ,
               "Santa Cruz de Tenerife",
               "Cantabria", "Segovia", "Sevilla",
               "Soria" , "Tarragona", "Teruel",
               "Toledo" , "Valencia", "Valladolid" ,
               "Vizcaya" , "Zamora", "Zaragoza" ,
               "Ceuta", "Melilla" )

#which(provDefinit=="Santa Cruz de Tenerife")
#which(provDefinit=="Las Palmas")

```

```

datosProv <- as.numeric(as.vector(
  datosProv[infracProv[c(1:length(infracProv)),2]]))

datosAbsProv<-data.frame(NAME_2=provDefinit,
                           Infracciones=datosProv)

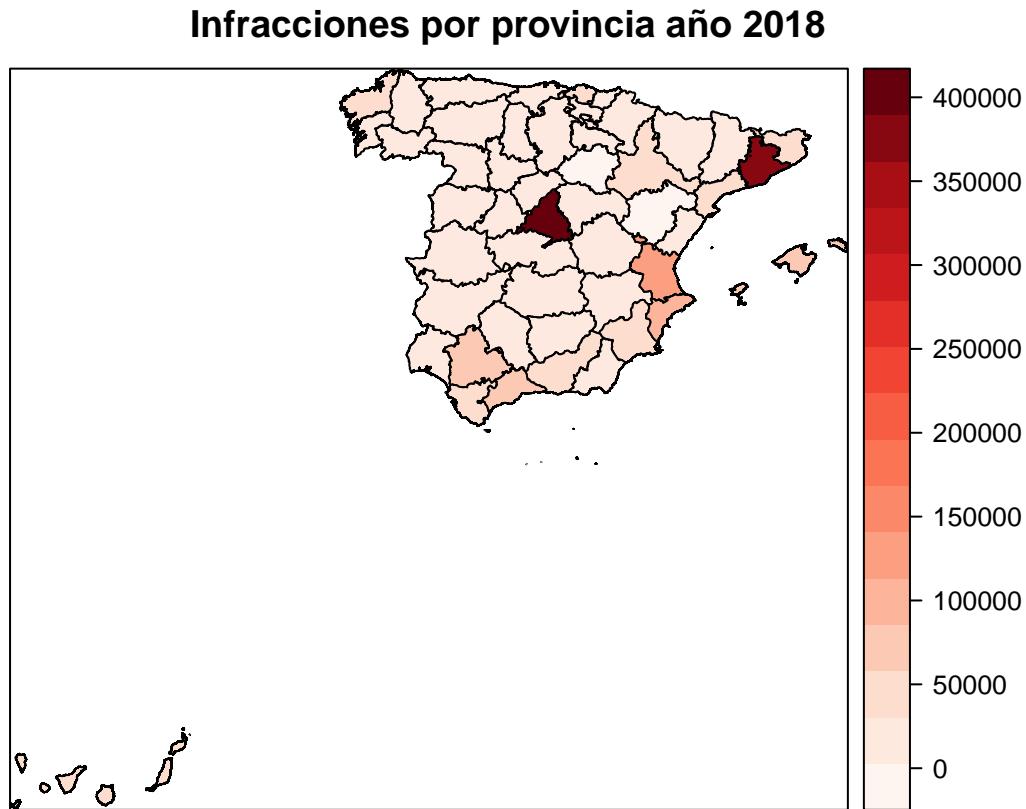
merged<- merge(mapaProv@data, datosAbsProv)

correct.ordering <- match(mapaProv$NAME_2, merged$NAME_2)
mapaProv@data <- merged[correct.ordering,]

my.palette <- colorRampPalette(
  brewer.pal(9, "Reds"))(max(mapaProv@data$Infracciones))

spplot(mapaProv, "Infracciones", col.regions=my.palette,
       main="Infracciones por provincia año 2018")

```



#-----Mapa por provincias: tasa de criminalidad -----

```

poblacionProv<-read.xlsx('poblacionProvincias18.xlsx',
                           sheetIndex = 1)

```

```

#tabla de autoridad
provPobl<-c("Albacete", "Alicante", "Almería", "Álava", "Asturias",
           "Ávila", "Badajoz", "Baleares", "Barcelona", "Vizcaya",
           "Burgos", "Cáceres", "Cádiz", "Cantabria", "Castellón",
           "Ciudad Real", "Córdoba", "A Coruña",
           "Cuenca", "Guipúzcoa", "Girona",
           "Granada", "Guadalajara", "Huelva", "Huesca", "Jaén",
           "León", "Lleida", "Lugo", "Madrid", "Málaga", "Murcia",
           "Navarra", "Ourense",
           "Palencia", "Las Palmas",
           "Pontevedra", "La Rioja", "Salamanca",
           "Santa Cruz de Tenerife",
           "Segovia", "Sevilla", "Soria", "Tarragona",
           "Teruel", "Toledo", "Valencia", "Valladolid",
           "Zamora", "Zaragoza", "Ceuta", "Melilla")

#which(provPobl=="Santa Cruz de Tenerife")
#which(provPobl=="Las Palmas")
aux <- which(poblacionProv[,2] != "<NA>")
aux <- aux[3:(length(aux))]
pobProv18 <- as.numeric(as.vector(poblacionProv[aux,2]))

dataPoblProv <- data.frame(NAME_2=provPobl, poblacion=pobProv18)

datosUnion<-merge(datosAbsProv,dataPoblProv)

#variable tasa por mil
tasaProv<-1000*datosUnion$Infracciones/datosUnion$poblacion

tasaDatosProv<-data.frame(NAME_2=datosUnion$NAME_2, Tasa=tasaProv)

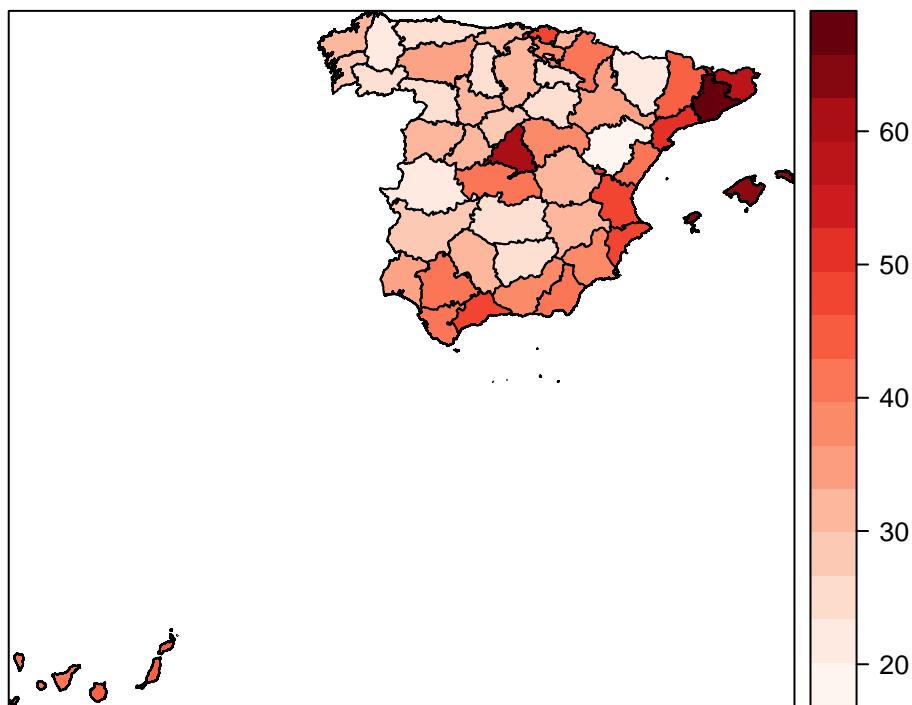
merged<- merge(mapaProv@data, tasaDatosProv)

correct.ordering <- match(mapaProv$NAME_2, merged$NAME_2)
mapaProv@data <- merged[correct.ordering,]

my.palette <- colorRampPalette(
  brewer.pal(9, "Reds"))(max(mapaProv$Tasa))
spplot(mapaProv, "Tasa", col.regions=my.palette,
       main="Número de infracciones por cada 1000
habitantes por provincia en el año 2018")

```

Número de infracciones por cada 1000 habitantes por provincia en el año 2018



```
#===== POR COMUNIDADES ======

#tabla de autoridad por comunidades

comunidades<-c("Andalucía" , "Aragón", "Principado de Asturias",
               "Islas Baleares", "Islas Canarias",
               "Cantabria" , "Castilla y León" ,
               "Castilla-La Mancha" ,
               "Cataluña" , "Comunidad Valenciana",
               "Extremadura", "Galicia",
               "Comunidad de Madrid", "Región de Murcia",
               "Comunidad Foral de Navarra",
               "País Vasco" , "La Rioja" ,
               "Ceuta y Melilla")


#-----poblacion por comunidades-----

poblacion<-read.xlsx('poblacionComunidades.xlsx', sheetIndex = 1)


aux <- which(poblacion[,2] != "<NA>")
aux <- aux[3:(length(aux))]
```

```

pob18 <- as.numeric(as.vector(poblacion[aux,2]))
pob17 <- as.numeric(as.vector(poblacion[aux,3]))
pob16 <- as.numeric(as.vector(poblacion[aux,4]))
pob <- data.frame(com=poblacion[aux,1],
                   pob16=pob16, pob17=pob17, pob18=pob18)
pobFinal<-data.frame(
  com=pob$com,PobMedia=apply(pob[,c(2,3,4)],1,mean))

#obtener de menor a mayor las poblaciones
#pobFinal$com[order(pobFinal$PobMedia)]
```

#-----

```

#----- Mapa por comunidades: valores absolutos -----
```

```

mapa<-readRDS('gadm36_ESP_1_sp.rds')
#mapa <- mapa[mapa$NAME_1!="Islas Canarias",]

data <- read.xlsx('total-infrac-com.xlsx', sheetIndex = 1)

infrac <- which(data[,2]!="<NA>")
infrac <- frac[3:(length(infrac)-1)]

datos <- as.numeric(as.vector(
  data[infrac[c(1:5,6:(length(infrac)-1))],2]))
datos[length(datos)]<-
  datos[length(datos)]+
  as.numeric(as.vector(data[infrac[length(infrac)],2]))

datosFinales<-data.frame(NAME_1=comunidades, Infracciones=datos)

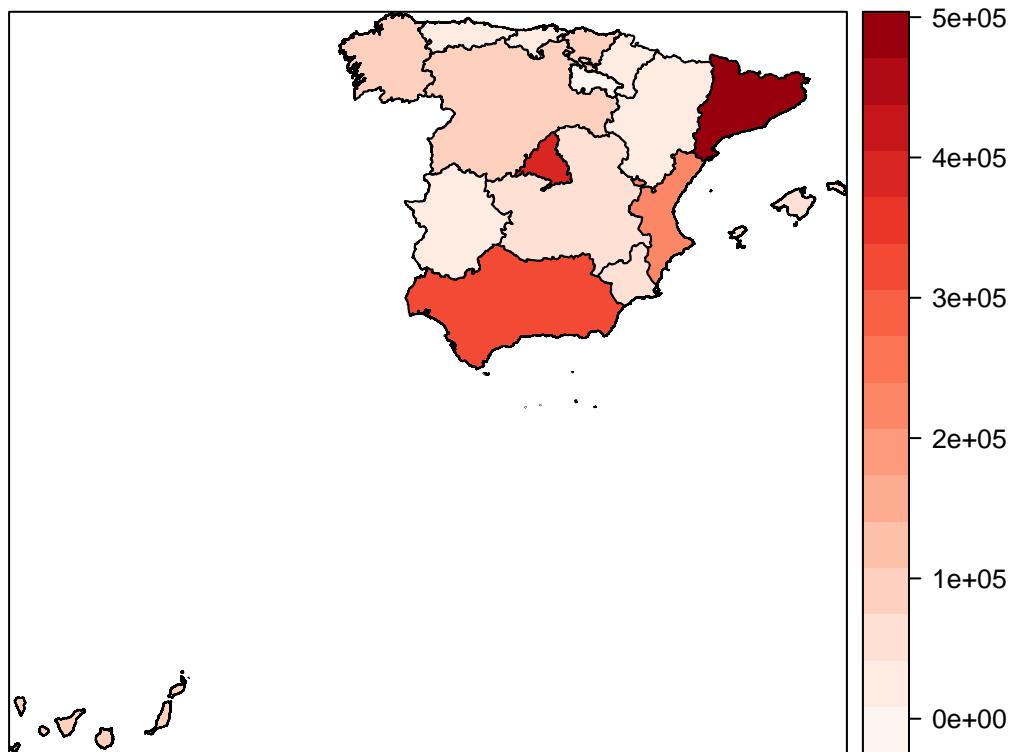
merged<- merge(mapa@data, datosFinales)

correct.ordering <- match(mapa$NAME_1, merged$NAME_1)
mapa@data <- merged[correct.ordering,]

my.palette <- colorRampPalette(
  brewer.pal(8, "Reds"))(max(mapa$Infracciones))
spplot(mapa, "Infracciones", col.regions=my.palette,
       main="Infracciones por comunidad año 2018")

```

Infracciones por comunidad año 2018



```
#----- Mapa por comunidades: tasa de criminalidad -----
```

```
#para juntar la poblacion de ceuta y melilla  
#pues en el mapa se consideran juntas  
  
pob18[length(pob18)-1]<-  
  pob18[length(pob18)-1]+pob18[length(pob18)]  
  
#ademas no coger el total de poblacion en España  
pob18<-pob18[2:(length(pob18)-1)]
```

```
#variable tasa por mil  
tasa<-1000*datosFinales$Infracciones/pob18
```

```
tasaDatos<-data.frame(NAME_1=comunidades, Tasa=tasa)
```

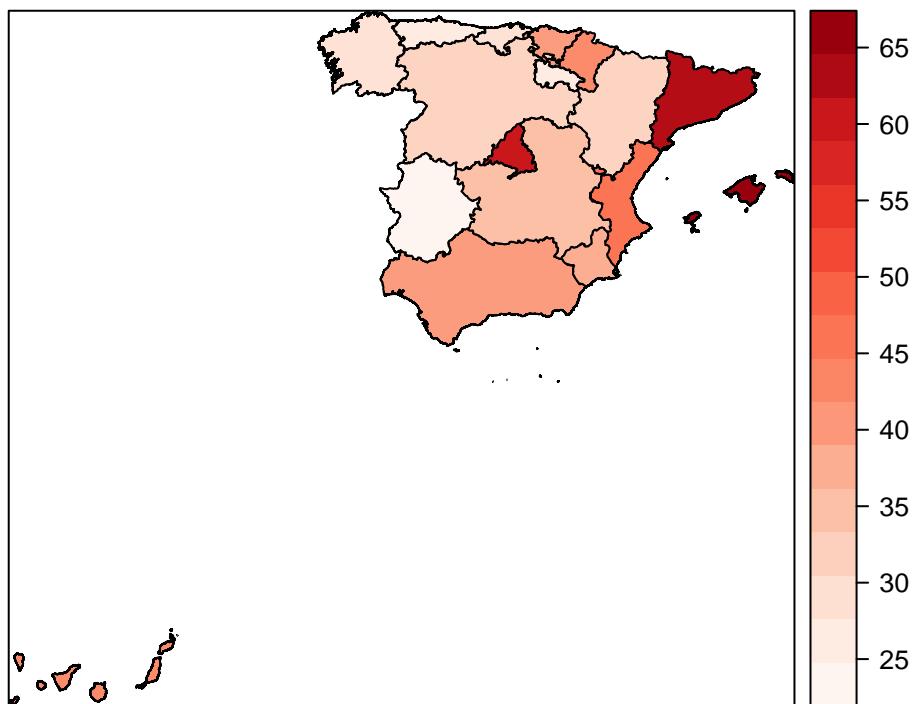
```
merged<- merge(mapa@data, tasaDatos)
```

```
correct.ordering <- match(mapa$NAME_1, merged$NAME_1)  
mapa@data <- merged[correct.ordering,]
```

```
my.palette <- colorRampPalette(  
  brewer.pal(8, "Reds"))(max(mapa$Tasa))
```

```
spplot(mapa, "Tasa", col.regions=my.palette,
       main="Número de infracciones por cada 1000
             habitantes por comunidad en el año 2018")
```

Número de infracciones por cada 1000 habitantes por comunidad en el año 2018



D.3. Tabla valores medios infracciones 2016-2018

```
=====Tabla valores medios infracciones 2016-2018=====

d18 <- read.xlsx('total-infrac-com.xlsx', sheetIndex = 1)
d1617 <- read.xlsx('total-infrac-com-1617.xlsx', sheetIndex = 1)

infrac18 <- which(d18[,2] != "<NA>")
infrac18 <- frac18[2:(length(infrac18)-1)]

aux18 <- as.numeric(as.vector(d18[infrac18,2]))

dfin18<-data.frame(com=d18[infrac18-1,1], Infracciones18=aux18)

infrac1617 <- which(d1617[,2] != "<NA>")
infrac1617 <- frac1617[2:(length(infrac1617)-1)]

aux16 <- as.numeric(as.vector(d1617[infrac1617,2]))
```

```

aux17 <- as.numeric(as.vector(d1617[infrac1617,3]))

dfin1617<-data.frame(com=d1617[infrac1617-1,1],
                      Infracciones16=aux16, Infracciones17=aux17)

unidas<- merge(dfin1617,dfin18)

#medias por comunidad

unidas<-cbind(unidas,apply(unidas[,c(2,3,4)],1,sum))
colnames(unidas)<-c("com", "Infracciones16", "Infracciones17",
                     "Infracciones18", "Total")
unidas<-cbind(unidas, unidas$Total/3)
colnames(unidas)<-c("com", "Infracciones16", "Infracciones17",
                     "Infracciones18", "Total", "Media")

kable(head(unidas),
      "latex", caption = "Primeras observaciones:
número de infracciones", booktabs = T) %>%
kable_styling(latex_options =
              c("striped","hold_position","scale_down"))

```

Cuadro 8: Primeras observaciones: número de infracciones

com	Infracciones16	Infracciones17	Infracciones18	Total	Media
ANDALUCÃA	331836	334331	333199	999366	333122.00
ARAGÃ“N	38146	37878	40255	116279	38759.67
ASTURIAS (PRINCIPADO DE)	26093	25652	26453	78198	26066.00
BALEARS (ILLES)	68160	72157	72944	213261	71087.00
CANARIAS	88782	91359	90566	270707	90235.67
CANTABRIA	17002	17440	17655	52097	17365.67

D.4. Barplots: tipología, sexo y edad

```

#===== Bar Plots =====

#-----Barplot: Por tipología -----

data <- read.xlsx('total-tiposInfrac-18.xlsx', sheetIndex = 1)

tipos <- which(data[,2] != "<NA>")
tipos <- tipos[2:(length(tipos))]

#-->comentados para realizarlo sin hurtos y sin resto de delitos

auxTipos <- as.numeric(as.vector(data[tipos,2]))
#auxTipos <- as.numeric(as.vector(data[tipos[c(1:10,12,13)],2]))

```

```

names <- c("A: Asesinatos consumados",
          "B: Asesinatos en grado tentativa",
          "C: Delitos de lesiones", "D: Secuestro",
          "E: Delitos contra la libertad e indemnidad sexual",
          "F: Agresion sexual con penetración",
          "G: Resto de delitos contra la libertad",
          "H: Robos con violencia e intimidación",
          "I: Robos con fuerza", "J: Robos en domicilios",
          "K: Hurtos", "L: Sustracción de vehículos",
          "M: Tráfico de drogas",
          "N: Resto de infracciones")

#names <- c("A: Asesinatos consumados",
#"B: Asesinatos en grado tentativa",
#"C: Delitos de lesiones", "D: Secuestro",
#"E: Delitos contra la libertad e indemnidad sexual",
#"F: Agresion sexual con penetración",
#"G: Resto de delitos contra la libertad",
#"H: Robos con violencia e intimidación",
#"I: Robos con fuerza", "J: Robos en domicilios",
#"K: Sustracción de vehículos", "L: Tráfico de drogas")

ind <- c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J", "K", "L", "M", "N")
#ind <- c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J", "K", "L")

finalTipos<-data.frame(Clasificación=names,
                         NumInfr=auxTipos, Identificador=ind)

kable(head(finalTipos),
      "latex", caption = "Primeras observaciones:
      infracciones según tipología", booktabs = T) %>%
kable_styling(latex_options =
              c("striped", "hold_position", "scale_down"))

```

Cuadro 9: Primeras observaciones: infracciones según tipología

Clasificación	NumInfr	Identificador
A: Asesinatos consumados	289	A
B: Asesinatos en grado tentativa	798	B
C: Delitos de lesiones	18252	C
D: Secuestro	81	D
E: Delitos contra la libertad e indemnidad sexual	13811	E
F: Agresion sexual con penetración	1702	F

```

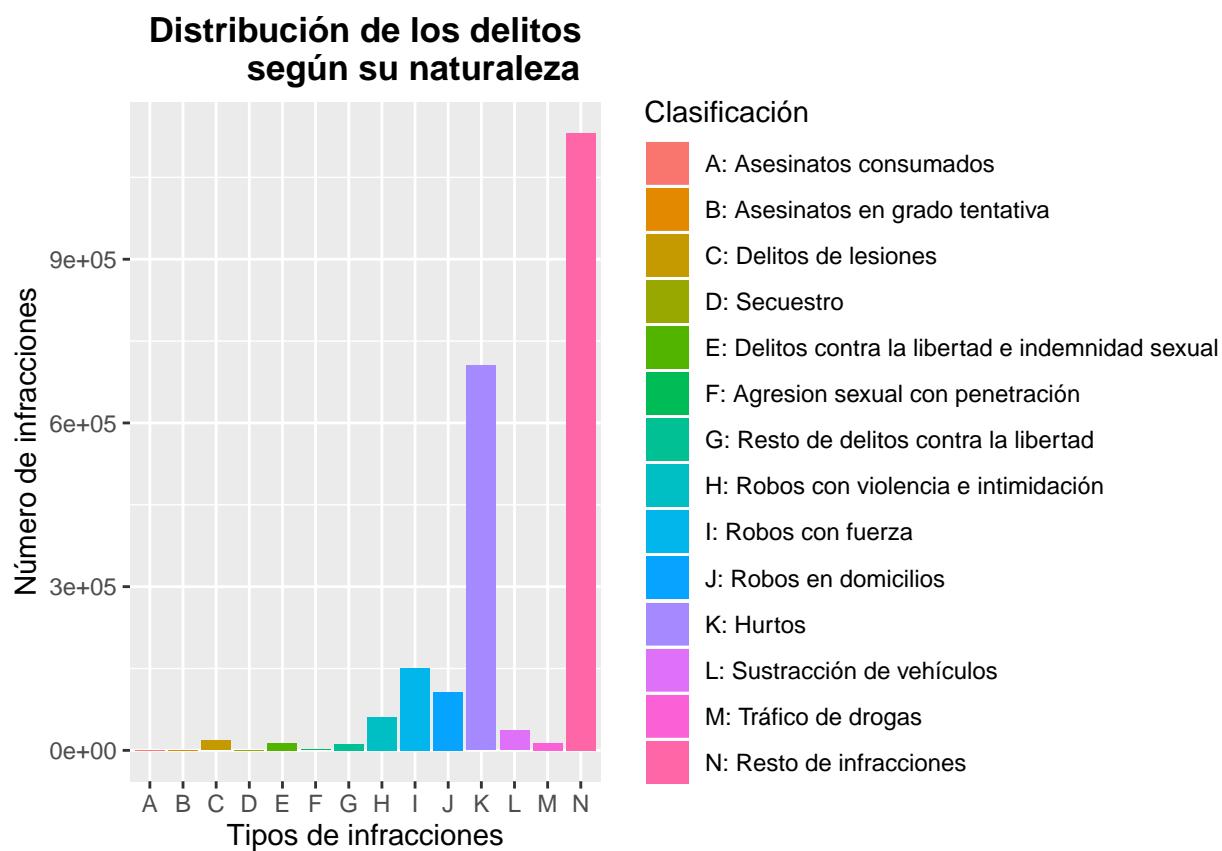
graf<-ggplot(data=finalTipos,
             aes(x=Identificador,
                  y=NumInfr, fill=Clasificación)) +
  geom_bar(stat="identity", position="dodge")

graf<-graf +
  ggtitle("Distribución de los delitos
           según su naturaleza") +
  xlab("Tipos de infracciones") +
  ylab("Número de infracciones") +
  theme(plot.title = element_text(hjust = 0.5, face="bold"))

#con ggplotly se puede observar qué tipos quieres ver, zoom, etc
ggplotly(graf)

graf

```



```

#----- barplot: distinción sexo y edad -----

data <- read.xlsx('total-infracSexoEdad-18.xlsx', sheetIndex = 1)

tipos <- which(data[,2] != "<NA>")
tipos <- tipos[4:(length(tipos))]

```

```

auxMasc <- as.numeric(as.vector(data[tipos,2]))
auxFem <- as.numeric(as.vector(data[tipos,3]))

edad<-c("14-17","18-30","31-40","41-64","Más de 64")
finalSex<-data.frame(Edad=edad, Masculino=auxMasc,
                      Femenino=auxFem)

sex<-c(rep("Masculino",5),rep("Femenino",5))
infr<-c(finalSex$Masculino,finalSex$Femenino)

fin<-data.frame(Sexo=sex, NumInfr=infr,
                  Edad=rep(finalSex$Edad,2))

kable(head(fin),
      "latex", caption = "Primeras observaciones:
      infracciones según sexo y edad", booktabs = T) %>%
kable_styling(latex_options =
              c("striped","hold_position"))

```

Cuadro 10: Primeras observaciones: infracciones según sexo y edad

Sexo	NumInfr	Edad
Masculino	15285	14-17
Masculino	98793	18-30
Masculino	80054	31-40
Masculino	95023	41-64
Masculino	8117	Más de 64
Femenino	3385	14-17

```

grafSex<-ggplot(data=fin, aes(x=Sexo, y=NumInfr, fill=Edad)) +
  geom_bar(stat="identity", position="stack")

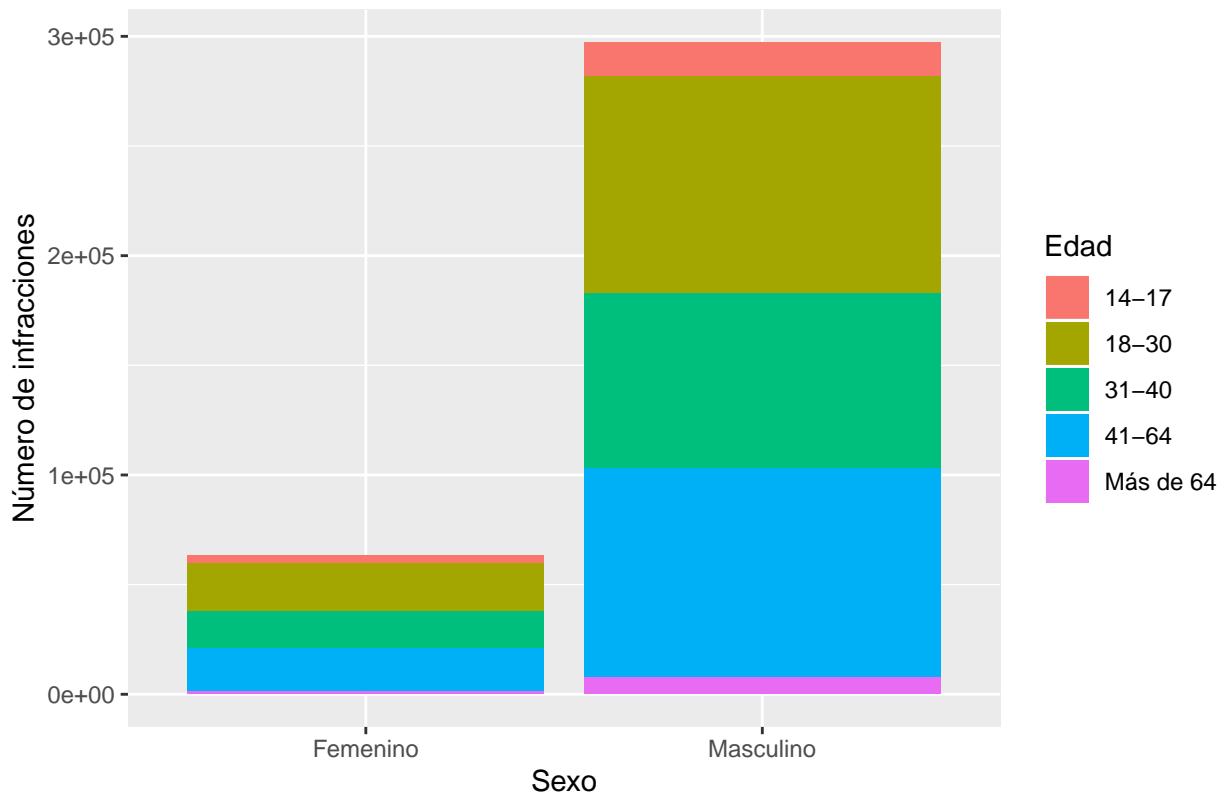
grafSex<-grafSex +
  ggtitle("Distribución de los delitos según sexo y edad") +
  xlab("Sexo") + ylab("Número de infracciones") +
  theme(plot.title = element_text(hjust = 0.5,face="bold"))

ggplotly(grafSex)

grafSex

```

Distribución de los delitos según sexo y edad



D.5. Diagrama de sankey: territorio y tipología

```
=====Shankey: Territorio y tipologia =====
```

```
data <- read.xlsx('total-tiposInfracComSinResto-18.xlsx',
                  sheetIndex = 1)

comunidades<-c("Andalucía" , "Aragón", "Principado de Asturias",
               "Islas Baleares", "Islas Canarias",
               "Cantabria" , "Castilla y León" ,
               "Castilla-La Mancha" ,
               "Cataluña" , "Comunidad Valenciana",
               "Extremadura", "Galicia", "Comunidad de Madrid",
               "Región de Murcia","Comunidad Foral de Navarra",
               "País Vasco" , "La Rioja" , "Ceuta" , "Melilla")

nombresDelitos <- c("Asesinatos consumados",
                     "Asesinatos en grado tentativa",
                     "Delitos de lesiones", "Secuestro",
                     "Delitos contra la libertad e indemnidad
                     sexual","Agresion sexual con penetración",
                     "Resto de delitos contra la libertad",
```

```

    "Robos con violencia e intimidación",
    "Robos con fuerza",
    "Robos en domicilios","Hurtos",
    "Sustracción de vehículos","Tráfico de drogas")

tipos <- which(data[,2] != "<NA>")
tipos <- tipos[2:(length(tipos))]

valoresCom_Tipo <- as.numeric(as.vector(data[tipos,2]))
tasaPorCienMil<-
  100000*valoresCom_Tipo/pob$pob18[2:length(pob$pob18)]

tiposDelitos<-rep(nombresDelitos,each=length(comunidades))
repComunidades<- rep(comunidades,length(nombresDelitos))

links <- data.frame(
  source=tiposDelitos,
  target=repComunidades,
  value=round(tasaPorCienMil,2),
  ind=0
)

write.csv(links, file="tasaTiposCom-18.csv")

#ordenar los grupos en funcion de la tasa
#para representarlos en orden creciente
ordenado<- aggregate(links$value,list(links$source),sum)
ordenado$Group.1 <- ordenado$Group.1[order(ordenado$x)]
ordenado <- dplyr::select(ordenado, -x) %>% mutate(ind=c(1:13))

for(i in 1:length(ordenado$Group.1)){
  links$ind[which(links$source==ordenado$Group.1[i])]<-
    ordenado$ind[i]
}

links$source<- links$source[order(links$ind)]
links$target<- links$target[order(links$ind)]
links$value<- links$value[order(links$ind)]
links$ind<-links$ind[order(links$ind)]

# ---el codigo comentado para las tablas de agresiones sexuales--

#orden de comunidades por agresion sexual
#links$value[57+order(
# links$value[which(

```

```

#      links$source=="Agresion sexual con penetración")])]

#links$target[57+order(
#  links$value[which(
#    links$source=="Agresion sexual con penetración")])]

#links$value[95+order(
#  links$value[which(
#    links$source==
#      "Delitos contra la libertad e indemnidad sexual")])]

#links$target[95+order(
#  links$value[which(
#    links$source==
#      "Delitos contra la libertad e indemnidad sexual")])]

nodes <- data.frame(
  name=c(as.character(links$source),
         as.character(links$target)) %>% unique()
)

nodes$group <- as.factor(
  c(rep("a",length(nombresDelitos)),rep("b",length(comunidades)))) 

links$group <- as.factor(
  c(rep(c("1","2","3","4","5","6","7","8","9","10","11","12","13"),
        each=length(comunidades)))) 

#para obtener los colores
#brewer.pal(13, "Purples")

my_color <- 'd3.scaleOrdinal() .domain(["a", "b","1","2","3","4",
  "5","6","7","8","9","10","11","12","13"]) .range([
  "#69b3a2",
  "steelblue", "#FCFBFD", "#EFEDF5", "#DADAEB", "#BCBDDC", "#BCBDDC",
  "#BCBDDC", "#BCBDDC", "#BCBDDC", "#9E9AC8", "#807DBA",
  "#6A51A3", "#54278F", "#3F007D"])' 

links$IDsource <- match(links$source, nodes$name)-1
links$IDtarget <- match(links$target, nodes$name)-1

p <- sankeyNetwork(Links = links, Nodes = nodes,
                    Source = "IDsource", Target = "IDtarget",
                    Value = "value", units="delitos/100.000hab",
                    NodeID = "name",
                    colourScale=my_color, LinkGroup = "group",
                    NodeGroup="group",

```

```
    iterations=0)
```

```
p
```

D.6. Boxplot MENAS

```
#===== Boxplot MENAS =====

comunidades<-c("Andalucía" , "Aragón", "Principado de Asturias",
              "Islas Baleares", "Islas Canarias",
              "Cantabria" , "Castilla y León" ,
              "Castilla-La Mancha" ,
              "Cataluña" , "Comunidad Valenciana",
              "Extremadura", "Galicia", "Comunidad de Madrid",
              "Región de Murcia","Comunidad Foral de Navarra",
              "País Vasco" , "La Rioja" , "Ceuta" , "Melilla")

pobMenores<-read.xlsx('poblacionMenoresNacioCom-18.xlsx',
                       sheetIndex = 1)

indice <- which(pobMenores[,2] != "<NA>")
indice <- indice[4:(length(indice))]

auxPobEsp<- as.numeric(as.vector(pobMenores[indice,2]))
auxPobExtr <- as.numeric(as.vector(pobMenores[indice,3]))

aux_pob<-data.frame(Españoles=auxPobEsp,
                      Extranjeros=auxPobExtr,
                      ident=rep(1:19, each=4))

pob_menores18<- data.frame(
  Españoles=tapply(aux_pob$Españoles,aux_pob$ident,sum),
  Extranjeros=tapply(aux_pob$Extranjeros,aux_pob$ident,sum))

data <- read.xlsx('total-infrComunExtr-18.xlsx', sheetIndex = 1)

ind <- which(data[,2] != "<NA>")
ind <- ind[4:(length(ind))]

auxEsp<- as.numeric(as.vector(data[ind,2]))
auxExtr <- as.numeric(as.vector(data[ind,3]))

delitos<-c(auxEsp,auxExtr)
nacionalidad <- c(rep("Español",19),rep("Extranjero",19))

dataExtr<- data.frame(Comunidades=comunidades,
```

```

Nacionalidad=nacionalidad,
Infracciones=delitos)

dataExtr<-dataExtr %>%
  mutate(Poblacion=
    c(pob_menores18$Españoles,pob_menores18$Extranjeros))%>%
  mutate(Tasa=Infracciones*1000/Poblacion)

kable(head(dataExtr),
  "latex", caption = "Primeras observaciones:
  infracciones MENAS", booktabs = T) %>%
kable_styling(latex_options =
  c("striped","hold_position","scale_down"))

```

Cuadro 11: Primeras observaciones: infracciones MENAS

Comunidades	Nacionalidad	Infracciones	Poblacion	Tasa
Andalucía	Español	3277	1666217	1.966731
Aragón	Español	695	214113	3.245950
Principado de Asturias	Español	244	143484	1.700538
Islas Baleares	Español	457	194016	2.355476
Islas Canarias	Español	687	360899	1.903580
Cantabria	Español	146	96996	1.505217

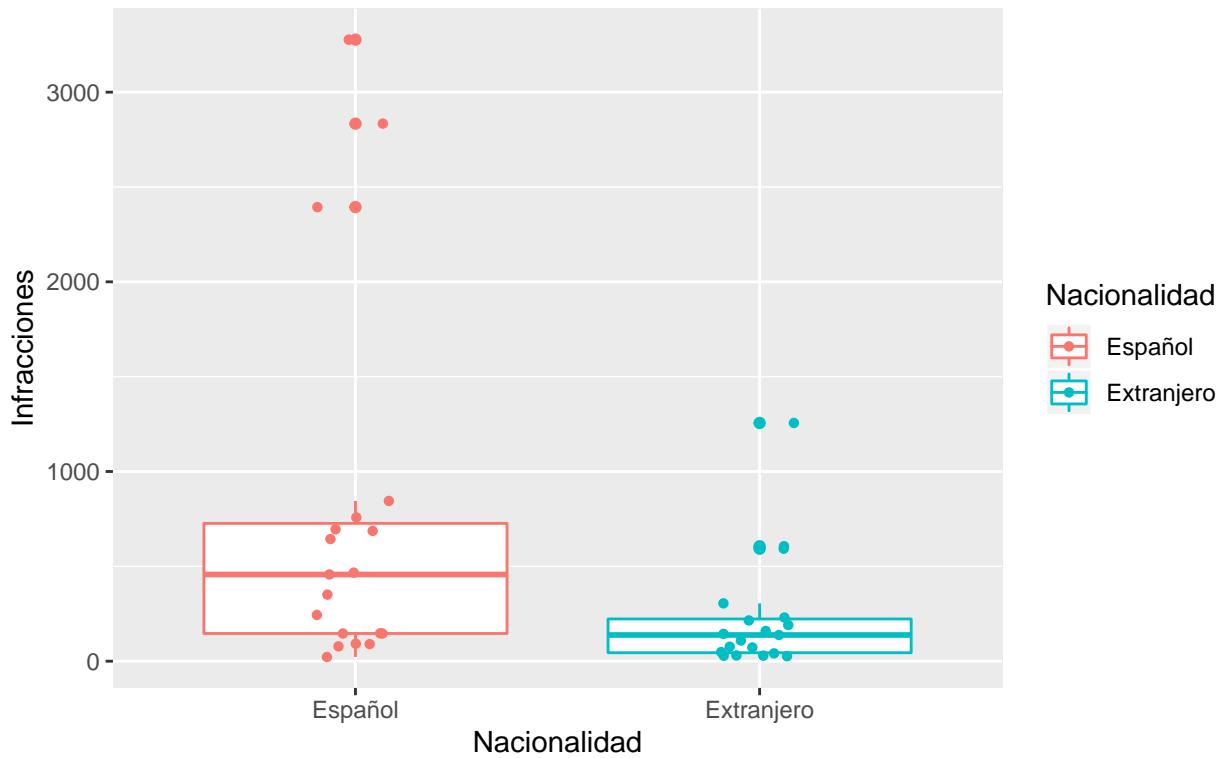
```

#Infracciones totales -> No es muy util
grafExtr<-
  ggplot(dataExtr, aes(x=Nacionalidad,
                        y=Infracciones,
                        color=Nacionalidad)) +
  geom_boxplot()

grafExtr + geom_jitter(shape=16, position=position_jitter(0.1)) +
  ggtitle("Distribución de los delitos
  para menores según su nacionalidad") +
  theme(plot.title = element_text(hjust = 0.5,face="bold"))

```

Distribución de los delitos para menores según su nacionalidad



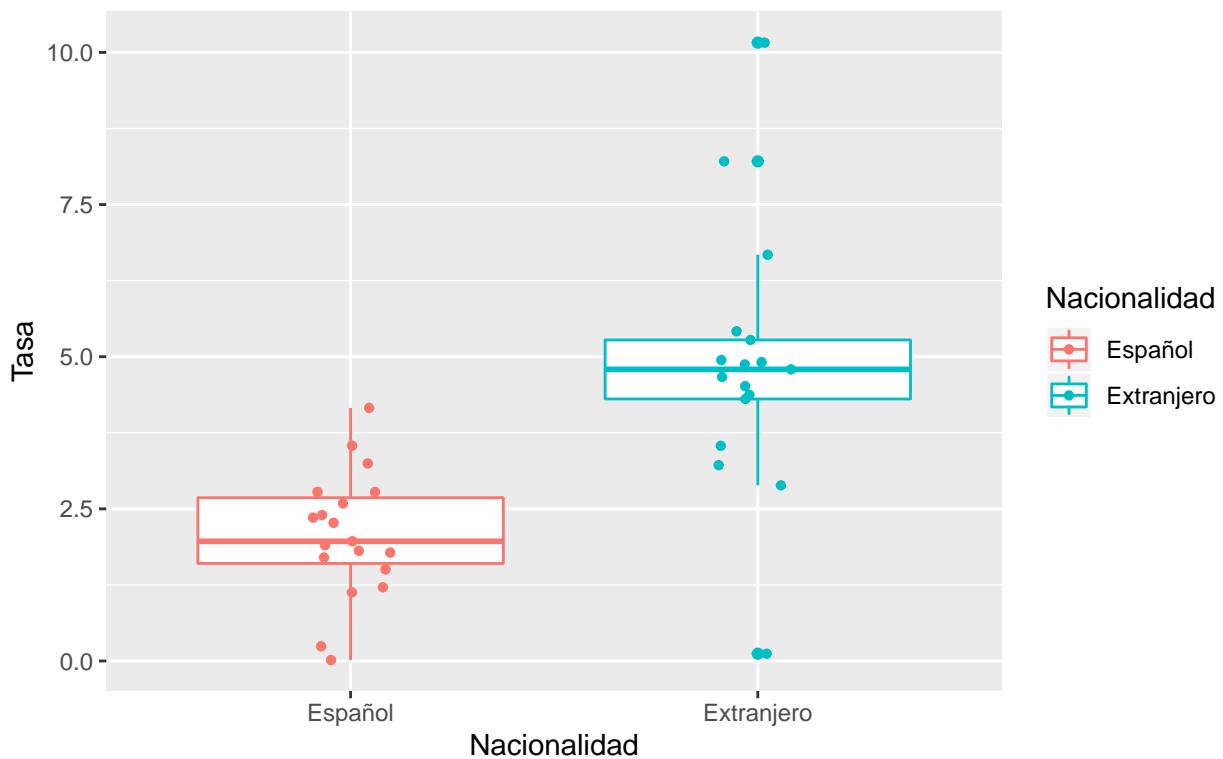
```
#Tasa infracciones por mil menores
```

```
grafExtr<-
  ggplot(dataExtr[1:(dim(dataExtr)[1]-2),],
         aes(x=Nacionalidad, y=Tasa, color=Nacionalidad)) +
  geom_boxplot()

grafExtr<-
  grafExtr +
  geom_jitter(shape=16, position=position_jitter(0.1)) +
  ggtitle("Distribución de la tasa de delincuencia
           para menores según su nacionalidad") +
  theme(plot.title = element_text(hjust = 0.5, face="bold"))

grafExtr
```

Distribución de la tasa de delincuencia para menores según su nacionalidad



```

grafExtr<-
  ggplot(dataExtr[1:(dim(dataExtr)[1]-2),],
         aes(x=Nacionalidad, y=Tasa, color=Nacionalidad,
              text=paste(
                "Tasax1000:", round(Tasa,2),
                "<br>Nacionalidad:", Nacionalidad,
                "<br>Comunidad:", Comunidades))) +
  geom_boxplot()

grafExtr<-
  grafExtr +
  geom_jitter(shape=16, position=position_jitter(0.1)) +
  ggtitle("Distribución de la tasa de delincuencia
           para menores según su nacionalidad") +
  theme(plot.title = element_text(hjust = 0.5, face="bold"))

ggplotly(grafExtr, tooltip = "text")

# TABLA: obtencion de las respectivas poblaciones y tasas
pobMenoresEsp<-sum(pob_menores18$Españoles)
pobMENAS<-sum(pob_menores18$Extranjeros)
infEsp<- sum(dataExtr$Infracciones[1:19])
infExt<-sum(dataExtr$Infracciones[20:38])

```

```
tasaEsp<-1000*infEsp/pobMenoresEsp
tasaExt<-1000*infExt/pobMENAS
```

D.7. Barplot: comparativa tasa de paro y criminalidad

```
#----poblacion----

comunidades<-c("Andalucía" , "Aragón", "Principado de Asturias",
              "Islas Baleares", "Islas Canarias",
              "Cantabria" , "Castilla y León" ,
              "Castilla-La Mancha" ,
              "Cataluña" , "Comunidad Valenciana",
              "Extremadura", "Galicia", "Comunidad de Madrid",
              "Región de Murcia","Comunidad Foral de Navarra",
              "País Vasco" , "La Rioja" , "Ceuta" , "Melilla")

poblacion<-read.xlsx('poblacionComSexo-18.xlsx', sheetIndex = 1)

aux <- which(poblacion[,2] != "<NA>")
aux <- aux[3:(length(aux))]

pobMasc <- as.numeric(as.vector(poblacion[aux,2]))
pobFem <- as.numeric(as.vector(poblacion[aux,3]))


#----datos infracciones---

data <- read.xlsx('total-infracSexoEdadCom-18.xlsx',
                   sheetIndex = 1)

tipos <- which(data[,2] != "<NA>")
tipos <- tipos[3:(length(tipos))]

auxMasc <- as.numeric(as.vector(data[tipos,2]))
auxFem <- as.numeric(as.vector(data[tipos,3]))

edad<-rep(c("14-17","18-30","31-40","41-64","Más de 64"),19)
sex<-c(rep("Masculino",5*19),rep("Femenino",5*19))

dfComSexoEdad<-data.frame(Edad=edad, Sexo=sex,
                           Infracciones=c(auxMasc,auxFem),
                           Indicador=as.factor(rep(c(1:19),each=5)))
)

infMasc<- dfComSexoEdad %>% filter(Sexo=="Masculino")
tasaMasc<-
  1000*tapply(infMasc$Infracciones, infMasc$Indicador,sum)/pobMasc
```

```

infFem<- dfComSexoEdad %>% filter(Sexo=="Femenino")
tasaFem<-
  1000*tapply(infFem$Infracciones, infFem$Indicador,sum)/pobFem

#===== datos paro=====

datos<-read.xlsx('tasaParoSexo-18.xlsx', sheetIndex = 1)

tipos <- which(datos[,2]!="<NA>")
tipos <- tipos[4:(length(tipos))]

auxMasc <- (as.numeric(as.vector(datos[tipos,2])) +
             as.numeric(as.vector(datos[tipos,3])) +
             as.numeric(as.vector(datos[tipos,4])) +
             as.numeric(as.vector(datos[tipos,5]))) / 4
auxFem <- (as.numeric(as.vector(datos[tipos,6])) +
             as.numeric(as.vector(datos[tipos,7])) +
             as.numeric(as.vector(datos[tipos,8])) +
             as.numeric(as.vector(datos[tipos,9]))) / 4

#=====Barplot compuesto=====

Sexo<-rep(rep(c("Masculino","Femenino"),each=19),2)
valores<-c(tasaMasc,tasaFem, auxMasc, auxFem )

data=data.frame(
  Comunidad=rep(comunidades,4),
  Sexo=Sexo ,
  Tasa=round(valores,2),
  Clasificacion=
    rep(c("Tasa criminalidad", "Tasa de paro"), each=19*2))

kable(head(data),
      "latex", caption = "Primeras observaciones:
comparativa tasa paro y criminalidad", booktabs = T) %>%
kable_styling(latex_options =
  c("striped","hold_position","scale_down"))

#---no interactivo---

p1<-
  ggplot(data %>% filter(Sexo=="Masculino"),
         aes(fill=Clasificacion, y=Tasa,

```

Cuadro 12: Primeras observaciones: comparativa tasa paro y criminalidad

Comunidad	Sexo	Tasa	Clasificacion
Andalucía	Masculino	17.50	Tasa criminalidad
Aragón	Masculino	13.38	Tasa criminalidad
Principado de Asturias	Masculino	13.02	Tasa criminalidad
Islas Baleares	Masculino	20.93	Tasa criminalidad
Islas Canarias	Masculino	19.14	Tasa criminalidad
Cantabria	Masculino	11.82	Tasa criminalidad

```

x=reorder(Comunidad,-Tasa))) +
geom_bar(position="dodge", stat="identity") +
coord_flip() +
ggtitle("Relación paro/criminalidad por
        sexo MASCULINO y comunidad")+
theme(plot.title = element_text(hjust = 0.5, face="bold"))+
labs(fill="")

aa<-c("Cataluña", "País Vasco", "Comunidad Foral de Navarra",
      "Cantabria", "Castilla y León", "Aragón", "Galicia",
      "La Rioja", "Principado de Asturias",
      "Comunidad de Madrid",
      "Castilla-La Mancha", "Región de Murcia",
      "Comunidad Valenciana", "Extremadura", "Islas Baleares",
      "Andalucía", "Islas Canarias", "Ceuta", "Melilla")

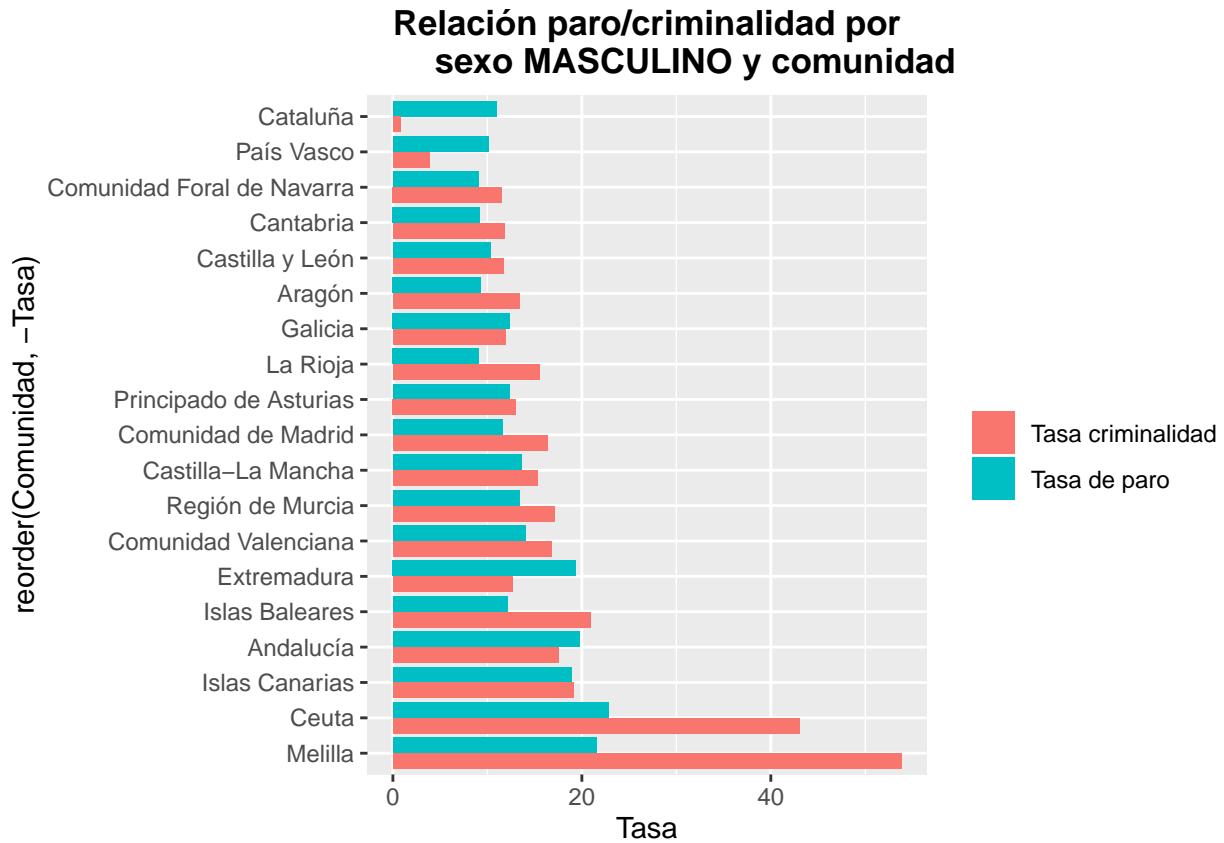
aa<-aa[length(aa):1]
auxiliar<-data.frame(Comunidad=aa)
auxFem<-data %>% filter(Sexo=="Femenino")

auxFinal<-inner_join(auxiliar, auxFem)
auxFinal$Comunidad <-
  factor(auxFinal$Comunidad, levels = auxiliar$Comunidad)

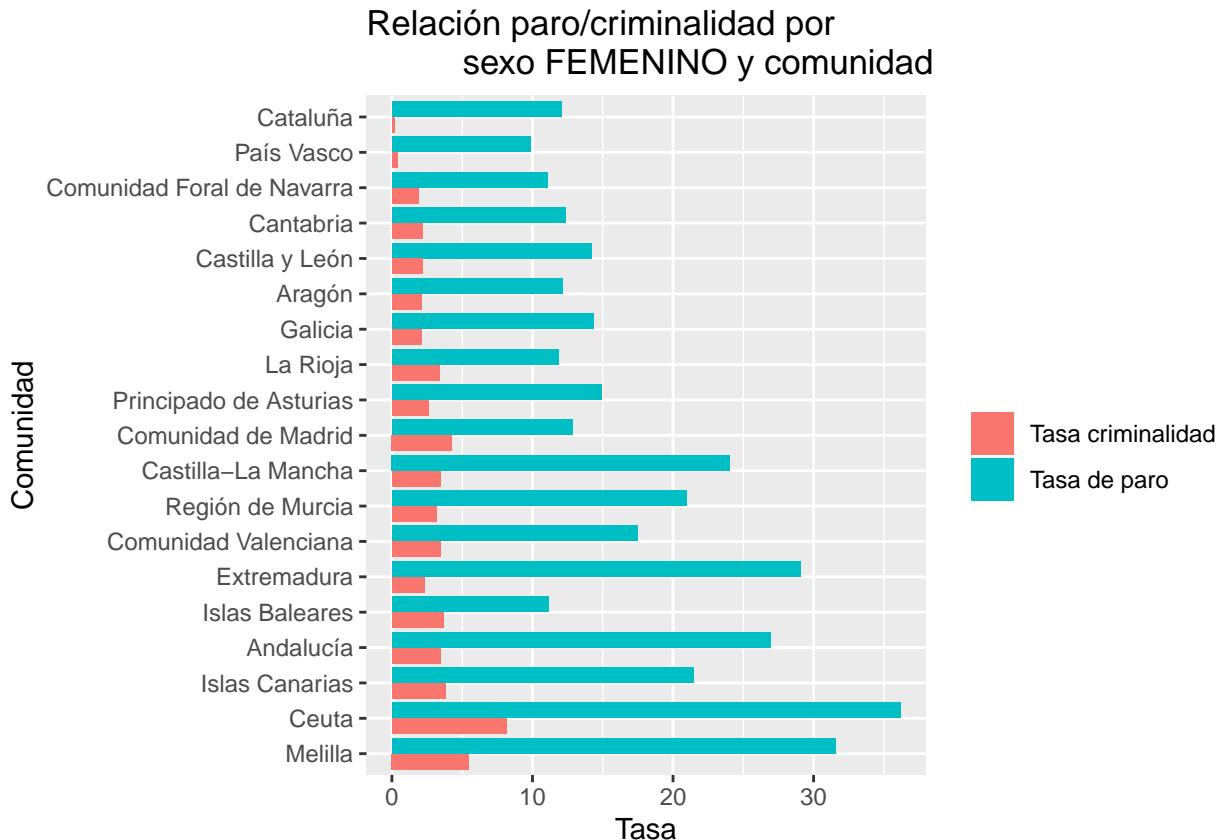
p2<-ggplot(auxFinal,
            aes(fill=Clasificacion, y=Tasa, x=Comunidad))+
  geom_bar(position="dodge", stat="identity") +
  coord_flip()+
  ggtitle("Relación paro/criminalidad por
          sexo FEMENINO y comunidad")+
  labs(fill="")

```

p1



p2



-----interactivo-----

```
p1<-
ggplot(data %>% filter(Sexo=="Masculino"),
       aes(fill=Clasificacion, y=Tasa,
           x=reorder(Comunidad,-Tasa),
           text=paste("Tipo:",Clasificacion,
                      "<br>Comunidad:",Comunidad,
                      "<br>Tasa:",Tasa," delitos/1.000hab")))+
  geom_bar(position="dodge", stat="identity") +
  coord_flip() +
  ggtitle("Relación paro/criminalidad por sexo y comunidad")+
  theme(plot.title = element_text(hjust = 0.5, face="bold"))+
  labs(fill="")

aa<-c("Cataluña","País Vasco","Comunidad Foral de Navarra",
      "Cantabria", "Castilla y León", "Aragón", "Galicia",
      "La Rioja", "Principado de Asturias",
      "Comunidad de Madrid",
      "Castilla-La Mancha", "Región de Murcia",
      "Comunidad Valenciana", "Extremadura", "Islas Baleares",
      "Andalucía", "Islas Canarias", "Ceuta","Melilla")
```

```

aa<-aa[length(aa):1]
auxiliar<-data.frame(Comunidad=aa)
auxFem<-data %>% filter(Sexo=="Femenino")

auxFinal<-inner_join(auxiliar, auxFem)
auxFinal$Comunidad <-
  factor(auxFinal$Comunidad, levels = auxiliar$Comunidad)

p2<-ggplot(auxFinal,
            aes(fill=Clasificacion, y=Tasa, x=Comunidad,
                text=paste("Tipo:",Clasificacion,
                           "<br>Comunidad:",Comunidad,
                           "<br>Tasa:",Tasa, " delitos/1.000hab")))+
  geom_bar(position="dodge", stat="identity") +
  coord_flip()+
  labs(fill="")

pFinal<-subplot(ggplotly(p1,tooltip="text"),
                 style(ggplotly(p2,tooltip="text"),
                        showlegend=FALSE ),
                 shareY = TRUE, titleY = FALSE, heights = 0.9)

pFinal %>%
  layout(annotations = list(
    list(x = 0.2 , y = 1, text = "Hombre",
         showarrow = F, xref='paper', yref='paper'),
    list(x = 0.8 , y =1, text = "Mujer",
         showarrow = F, xref='paper', yref='paper')))
```

D.8. Gráfico animado

```

=====poblacion reclusa=====

datos<-read.xlsx('Estadistica diciembre 2018.xlsx',
                  sheetIndex = 3)
a18<-as.numeric(as.vector(datos[3:21,4]))

datos<-read.xlsx('Estadistica diciembre 2017.xlsx',
                  sheetIndex = 3)
a17<-as.numeric(as.vector(datos[3:21,4]))

datos<-read.xlsx('Estadistica diciembre 2016.xlsx',
                  sheetIndex = 3)
a16<-as.numeric(as.vector(datos[3:21,4]))
```

```

datos<-read.xlsx('Estadistica diciembre 2015.xlsx',
                 sheetIndex = 3)
a15<-as.numeric(as.vector(datos[3:21,4]))


datos<-read.xlsx('Estadistica diciembre 2014.xlsx',
                 sheetIndex = 3)
a14<-as.numeric(as.vector(datos[3:21,4]))


datos<-read.xlsx('Estadistica diciembre 2013.xlsx',
                 sheetIndex = 3)
a13<-as.numeric(as.vector(datos[3:21,4]))


datos<-read.xlsx('Estadistica diciembre 2012.xlsx',
                 sheetIndex = 3)
a12<-as.numeric(as.vector(datos[3:21,4]))


datos<-read.xlsx('Estadistica diciembre 2011.xlsx',
                 sheetIndex = 3)
a11<-as.numeric(as.vector(datos[3:21,4]))


datos<-read.xlsx('Estadistica diciembre 2010.xlsx',
                 sheetIndex = 3)
a10<-as.numeric(as.vector(datos[3:21,4]))


datos<-read.xlsx('Estadistica diciembre 2009.xlsx',
                 sheetIndex = 3)
a09<-as.numeric(as.vector(datos[3:21,4]))


datos<-read.xlsx('Estadistica diciembre 2008.xlsx',
                 sheetIndex = 3)
a08<-as.numeric(as.vector(datos[3:21,4]))


datos<-read.xlsx('Estadistica diciembre 2007.xlsx',
                 sheetIndex = 3)
a07<-as.numeric(as.vector(datos[3:21,4]))


datos<-read.xlsx('Estadistica diciembre 2006.xlsx',
                 sheetIndex = 3)
a06<-as.numeric(as.vector(datos[3:21,4]))


poblacionReclusa<- data.frame(Año18= a18, Año17=a17, Año16=a16,
                                 Año15=a15, Año14=a14, Año13=a13,
                                 Año12=a12, Año11=a11, Año10=a10,
                                 Año09=a09, Año08=a08,Año07=a07,
                                 Año06=a06)

```

```

pRec<-as.vector(t(poblacionReclusa))

#=====poblacion comunidad

datos<-read.xlsx('2915.xlsx', sheetIndex = 1)

poblacionTotal<-as.numeric(as.vector(t(datos[8:26,2:14])))

#===== PIB per capita =====

datos<-read.xlsx('pr_cre.xlsx', sheetIndex = 3)

#which(datos[4,]==2006)
#which(datos[4,]=="2018 (A)")
secuencias<- seq(from=25, to=73, by=4)
pib<-as.numeric(
  as.vector(
    t(
      datos[
        c(6,15,19,20,21,24,25,35,41,46,
          50,53,58,59,60,61,65,66,67),secuencias])))
}

#=====datos finales

comunidad<-c("Andalucía" , "Aragón", "Principado de Asturias",
             "Islas Baleares", "Islas Canarias",
             "Cantabria", "Castilla y León" ,
             "Castilla-La Mancha" , "Cataluña" ,
             "Comunidad Valenciana", "Extremadura", "Galicia",
             "Comunidad de Madrid", "Región de Murcia",
             "Comunidad Foral de Navarra", "País Vasco" ,
             "La Rioja" , "Ceuta" , "Melilla")

datosFin<-data.frame(Reclusos=pRec,
                      Poblacion=poblacionTotal,
                      Año=rep(c(2018:2006),19),
                      Comunidad=rep(comunidad, each=13)
)

datosFin2<-data.frame(PIB=pib,
                      Año=rep(c(2006:2018),19),
                      Comunidad=rep(comunidad, each=13))

```

```

datosFinales<-merge(datosFin, datosFin2)

n <- 19
qual_col_pals =
  brewer.pal.info[brewer.pal.info$category == 'qual',]

col_vector =
  unlist(mapply(brewer.pal, qual_col_pals$maxcolors,
    rownames(qual_col_pals)))

colores<-sample(col_vector, n)

datosFinales <- datosFinales %>% mutate(Color=rep(colores,13))

p <- datosFinales %>%
  plot_ly(
    x = ~PIB,
    y = ~Reclusos,
    size = ~Poblacion,
    color=~Comunidad,
    frame = ~Año,
    text = paste("Comunidad:",datosFinales$Comunidad,
      "<br>Población:", datosFinales$Poblacion,
      "<br>Población reclusa:",datosFinales$Reclusos,
      "<br>PIB:",datosFinales$PIB),
    hoverinfo = "text",
    type = 'scatter',
    mode = 'markers'
  ) %>%
  layout(
    xaxis = list(
      type = "log"
    ),
    title="Evolución económica y delictiva por comunidad"
  )

p %>% animation_opts(
  1000, easing = "elastic", redraw = FALSE
)

```

D.9. Gráfico evolución infracciones por edad y sexo

```

datos<-read.xlsx('datosSexEdadYear-Modelos.xlsx', sheetIndex = 1)

ind<- which(datos[,2] != "<NA>")

```

```

infr <- as.numeric(as.vector(datos[ind[3:length(ind)],2]))
sexo<- as.factor(rep(c("Masculino", "Femenino"),9*5))
edad<-as.factor(
  rep(rep(
    c("14-17","18-30","31-40","41-64","Más de 64"),each=2), 9))
anio<-(rep
  (c(2018,2017,2016,2015,2014,2013,2012,2011,2010), each=10))

dataProc <- data.frame(Infracciones=infr,
                        Sexo=sexo,
                        Edad=edad,
                        Año=anio)

#=====Grafico descriptivo =====

#----no interactivo-----

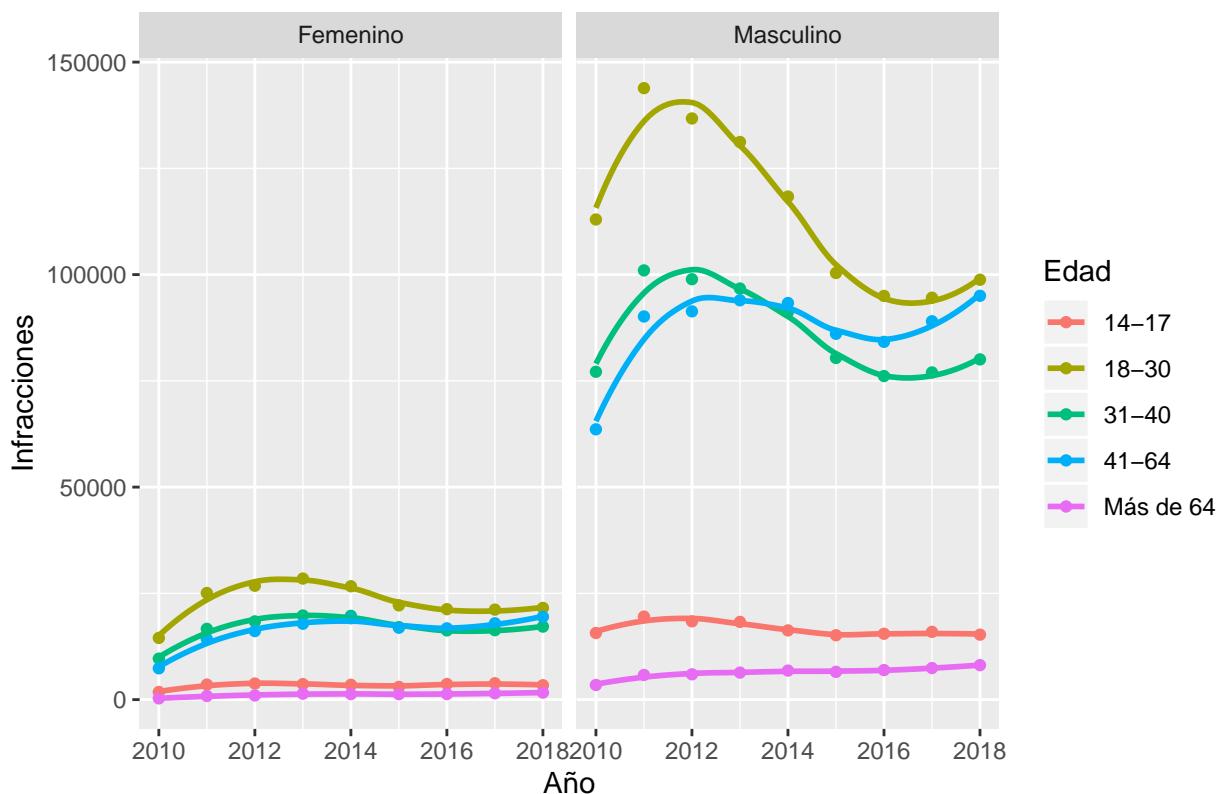
graf <- ggplot(dataProc, aes(Año, Infracciones, color=Edad)) +
  geom_point()

graf<-graf + stat_smooth(se =FALSE) + facet_wrap(~ Sexo) +
  ggtitle("Distribución de los delitos según año, sexo y edad") +
  theme(plot.title = element_text(hjust = 0.5,face="bold"))

graf

```

Distribución de los delitos según año, sexo y edad



```
#----interactivo----
```

```
graf <- ggplot(dataProc, aes(Año, Infracciones, color=Edad)) +
  geom_point(aes(text=paste("Año:", Año,
                        "<br>Sexo:", Sexo,
                        "<br>Edad:", Edad,
                        "<br>Número de infracciones:",
                        Infracciones)))

graf<-graf + stat_smooth(se =FALSE) + facet_wrap(~ Sexo) +
  ggtitle("Distribución de los delitos según año, sexo y edad") +
  theme(plot.title = element_text(hjust = 0.5,face="bold"))

ggplotly(graf,tooltip="text")
```

E. Técnicas descriptivas avanzadas

E.1. Análisis de correspondencias

```
=====Análisis de correspondencias=====
```

```

comunidad<-c("Andalucia" , "Aragon" , "Principado de Asturias",
           "Islas Baleares" , "Islas Canarias",
           "Cantabria" , "Castilla y Leon" ,
           "Castilla-La Mancha" ,
           "Catalunia" , "Comunidad Valenciana" , "Extremadura",
           "Galicia" , "Comunidad de Madrid",
           "Region de Murcia","Comunidad Foral de Navarra",
           "Pais Vasco" , "La Rioja" , "Ceuta" , "Melilla")

-----por edad-----

datos<-read.xlsx('datosEdadCom-AC.xlsx', sheetIndex = 1)

ind<- which(datos[,2] !="<NA>")

prim <- as.numeric(as.vector(datos[ind[4:length(ind)],2]))
seg <- as.numeric(as.vector(datos[ind[4:length(ind)],3]))
ter <- as.numeric(as.vector(datos[ind[4:length(ind)],4]))
cuart <- as.numeric(as.vector(datos[ind[4:length(ind)],5]))
quin <- as.numeric(as.vector(datos[ind[4:length(ind)],6]))

dataACEedad<- data.frame(E14_17=prim,
                           E18_30=seg,
                           E31_40=ter,
                           E41_64=cuart,
                           E64_mas=quin)

#tabla para ver los totales
#dataACEedad$Total <- apply(dataACEedad, 1, sum)
#dataACEedad<-
#as.data.frame(rbind(dataACEedad, apply(dataACEedad, 2, sum)))
#rownames(dataACEedad)<-c(comunidad, "Total")

rownames(dataACEedad)<-comunidad

kable(head(dataACEedad),
      "latex", caption = "Primeras observaciones:
      infracciones por edad y territorio", booktabs = T) %>%
kable_styling(latex_options =
              c("striped","hold_position","scale_down"))

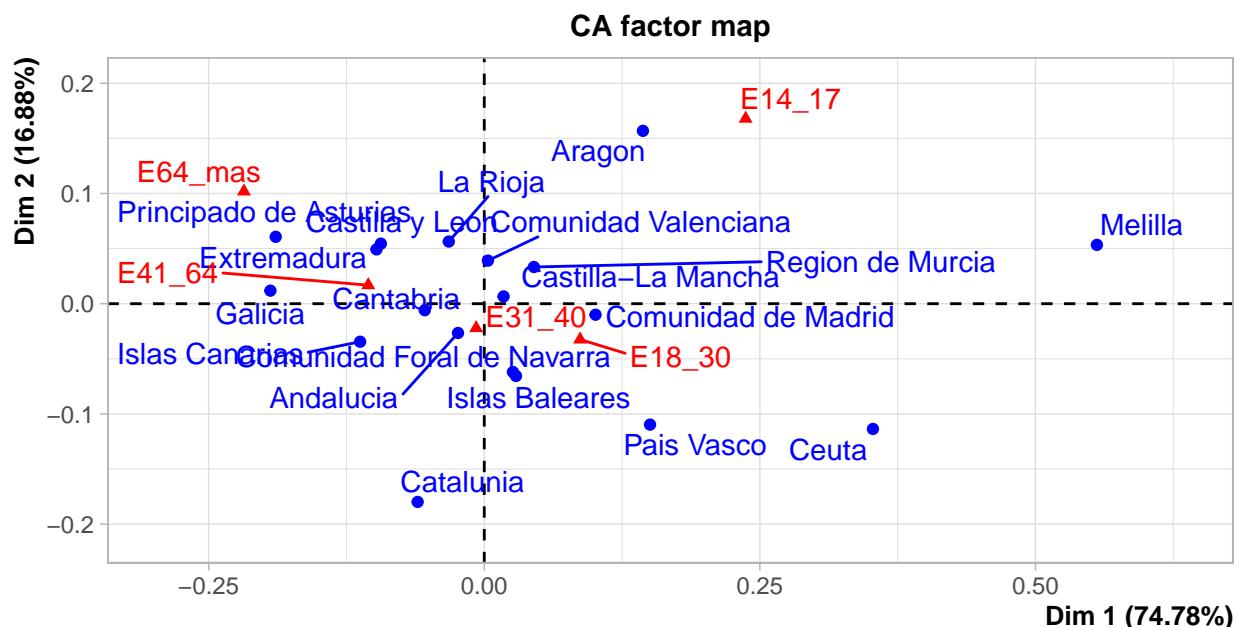
propor_contig_edad<-
  as.data.frame(prop.table(as.matrix(dataACEedad[1:19,1:5])))

ac<-CA(dataACEedad)

```

Cuadro 13: Primeras observaciones: infracciones por edad y territorio

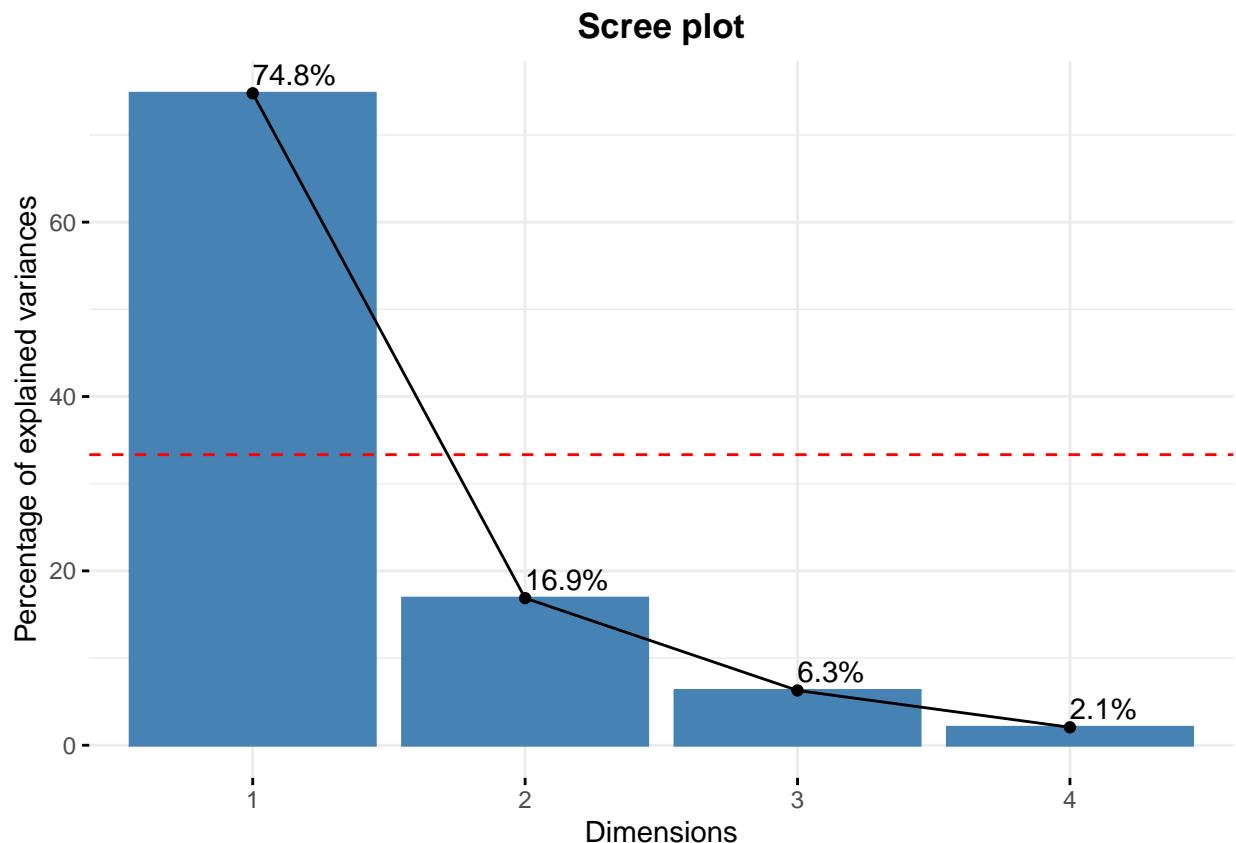
	E14_17	E18_30	E31_40	E41_64	E64_mas
Andalucía	3870	28652	24308	27861	2332
Aragón	1000	3276	2624	2916	227
Principado de Asturias	292	2173	2006	2934	385
Islas Baleares	616	4820	4001	4134	277
Islas Canarias	796	7433	6659	8724	665
Cantabria	175	1289	1046	1357	114



```
#summary(ac)
#get_eigenvalue(ac)

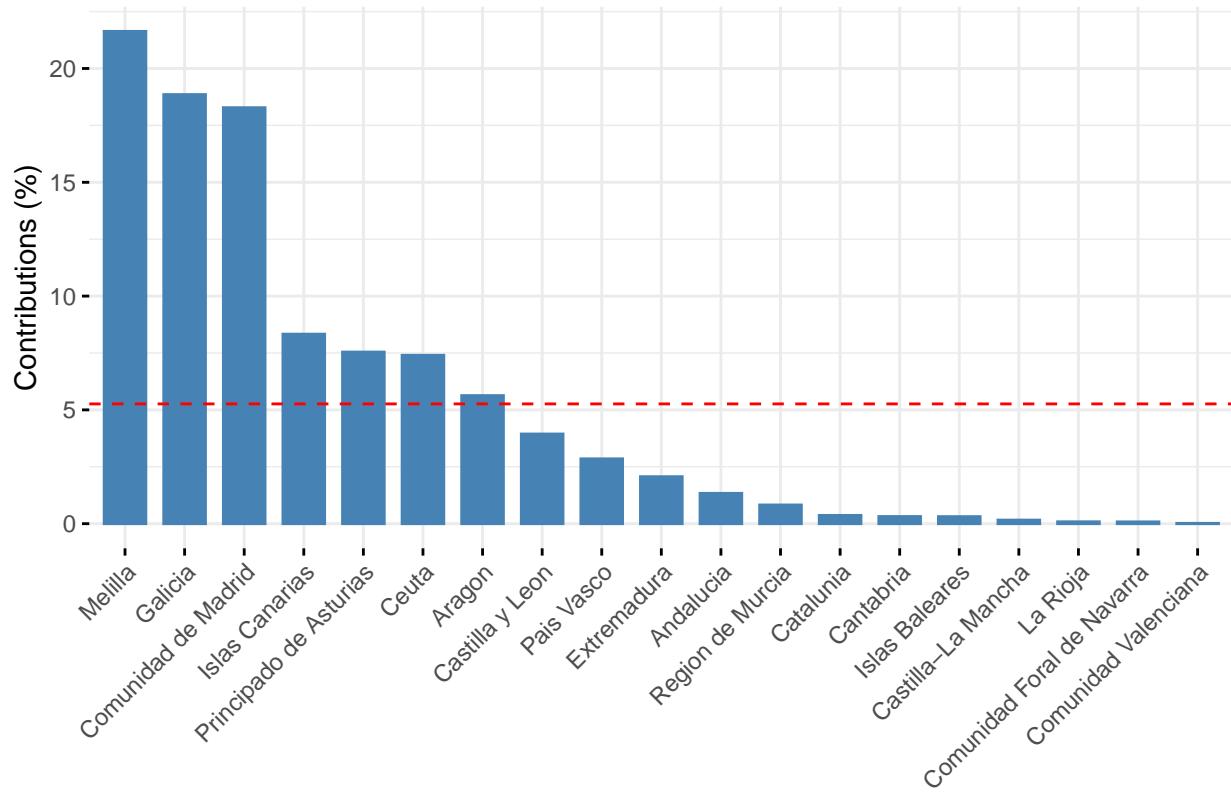
#grafico de contribucion de cada dimension
#interactivo
ggplotly(fviz_screenplot(ac, addlabels=TRUE) +
  geom_hline(yintercept=33.33, linetype=2, color="red") +
  theme(plot.title = element_text(hjust =
  0.5,face="bold")))
```

```
#no interactiva
fviz_screeplot(ac, addlabels=TRUE) +
  geom_hline(yintercept=33.33, linetype=2, color="red") +
  theme(plot.title = element_text(hjust =
  0.5, face="bold"))
```



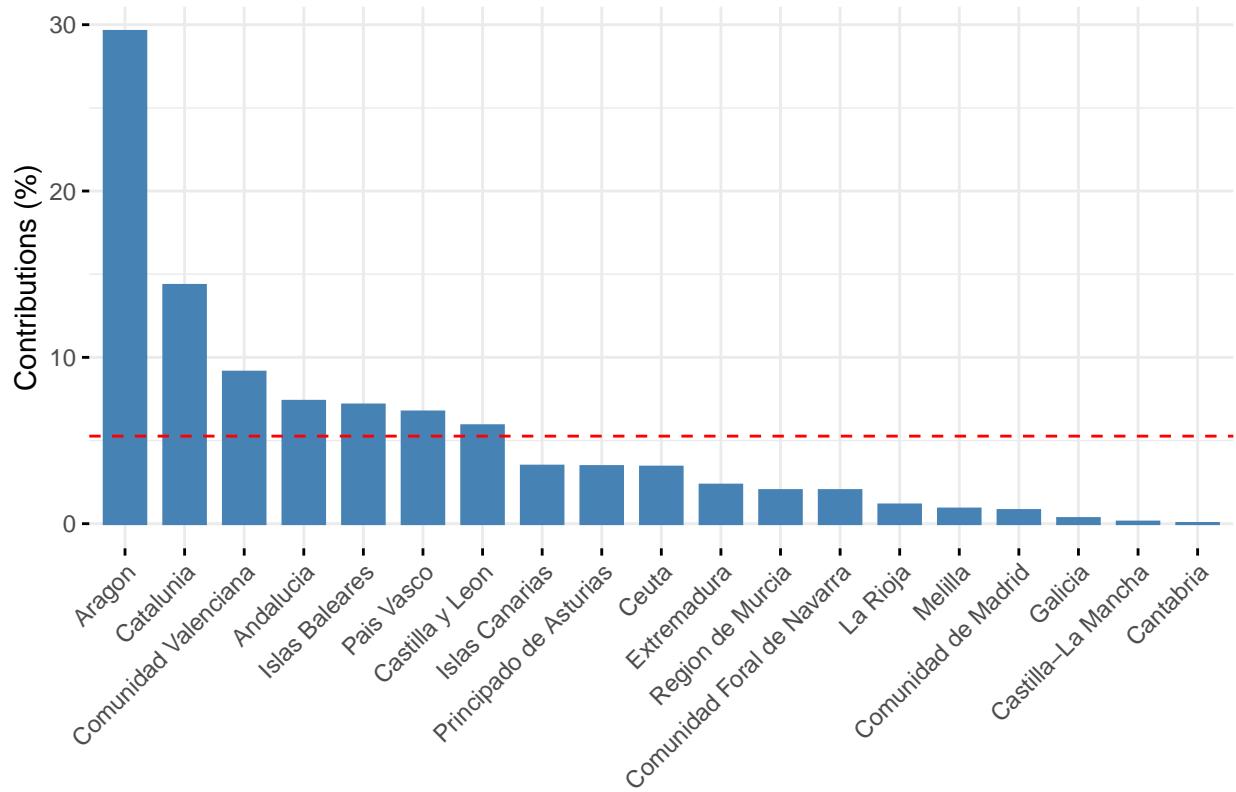
```
#contribuciones filas y columnas a las dimensiones 1 y 2
#interactivo
rdim1<-ggplotly(fviz_contrib(ac, choice = "row", axes = 1) +
  ggtitle("Contribución perfiles fila")+
  theme(plot.title = element_text(hjust = 0.5,
  face="bold")))
#no interactiva
fviz_contrib(ac, choice = "row", axes = 1) +
  ggtitle("Contribución perfiles fila")+
  theme(plot.title = element_text(hjust = 0.5,
  face="bold"))
```

Contribución perfiles fila



```
#interactivo
rdim2<-ggplotly(fviz_contrib(ac, choice = "row", axes = 2)+
  ggttitle("Contribución perfiles fila")+
  theme(plot.title = element_text(hjust = 0.5,
                                    face="bold")))
#no interactivo
fviz_contrib(ac, choice = "row", axes = 2)+
  ggttitle("Contribución perfiles fila")+
  theme(plot.title = element_text(hjust = 0.5,
                                    face="bold"))
```

Contribución perfiles fila



```

filas<-subplot(rdim1, rdim2,
               shareY = TRUE, heights = 0.9)

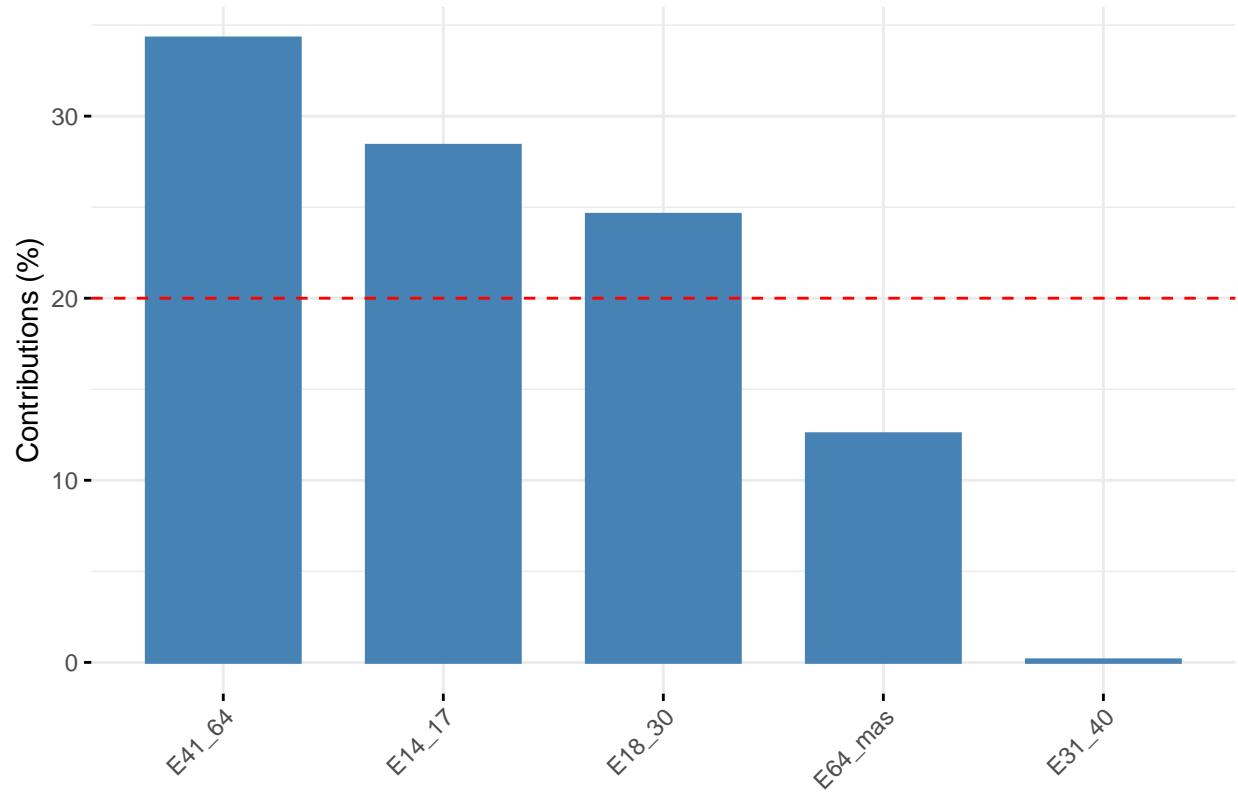
filas %>% layout(annotations = list(
  list(x = 0.2 , y = 1, text = "Dimension 1", showarrow = F,
       xref='paper', yref='paper'),
  list(x = 0.8 , y = 1, text = "Dimension 2", showarrow = F,
       xref='paper', yref='paper'))
)

#interactivo
cdim1<-ggplotly(fviz_contrib(ac, choice = "col", axes = 1) +
  ggtitle("Contribución perfiles columna")+
  theme(plot.title = element_text(hjust = 0.5,
                                  face="bold")))

#no interactivo
fviz_contrib(ac, choice = "col", axes = 1) +
  ggtitle("Contribución perfiles columna")+
  theme(plot.title = element_text(hjust = 0.5,
                                  face="bold"))

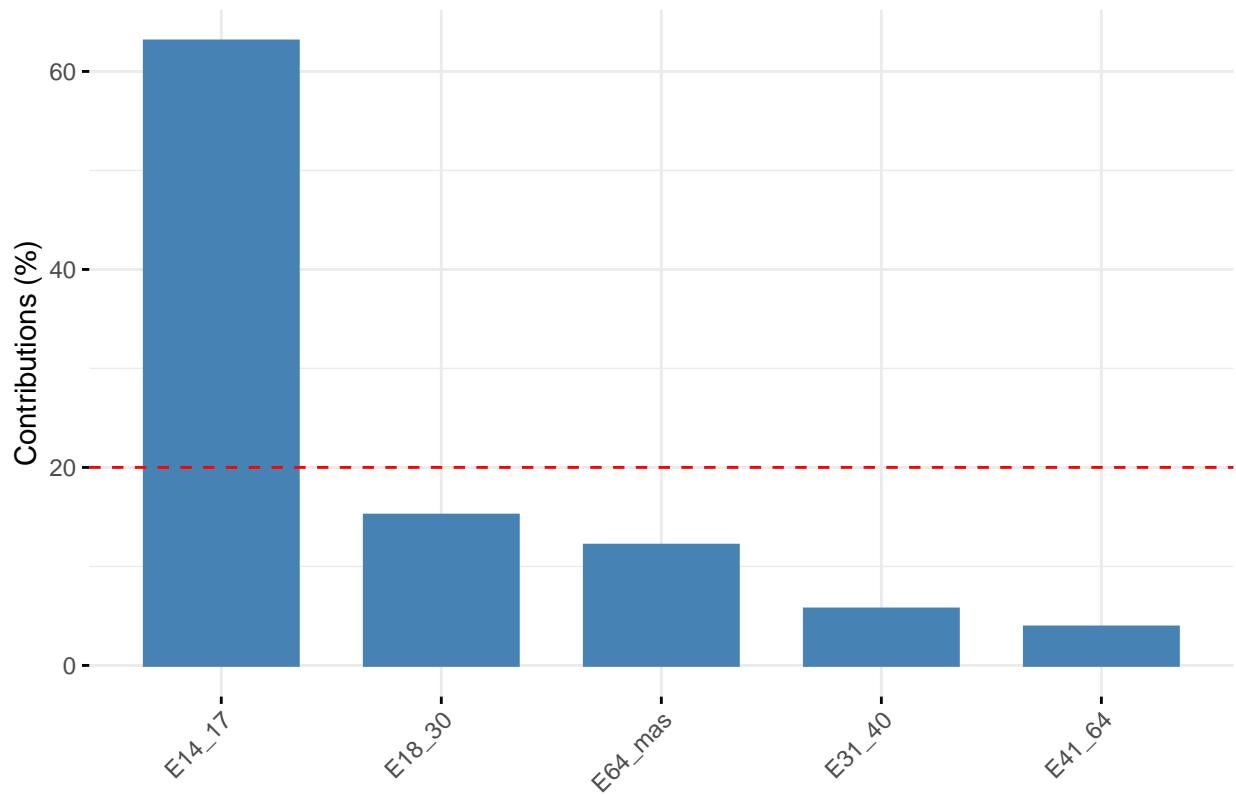
```

Contribución perfiles columna



```
#interactivo
cdim2<-ggplotly(fviz_contrib(ac, choice = "col", axes = 2) +
  ggttitle("Contribución perfiles columna")+
  theme(plot.title = element_text(hjust = 0.5,
                                    face="bold")))
#no interactivo
fviz_contrib(ac, choice = "col", axes = 2) +
  ggttitle("Contribución perfiles columna")+
  theme(plot.title = element_text(hjust = 0.5,
                                    face="bold"))
```

Contribución profiles columna

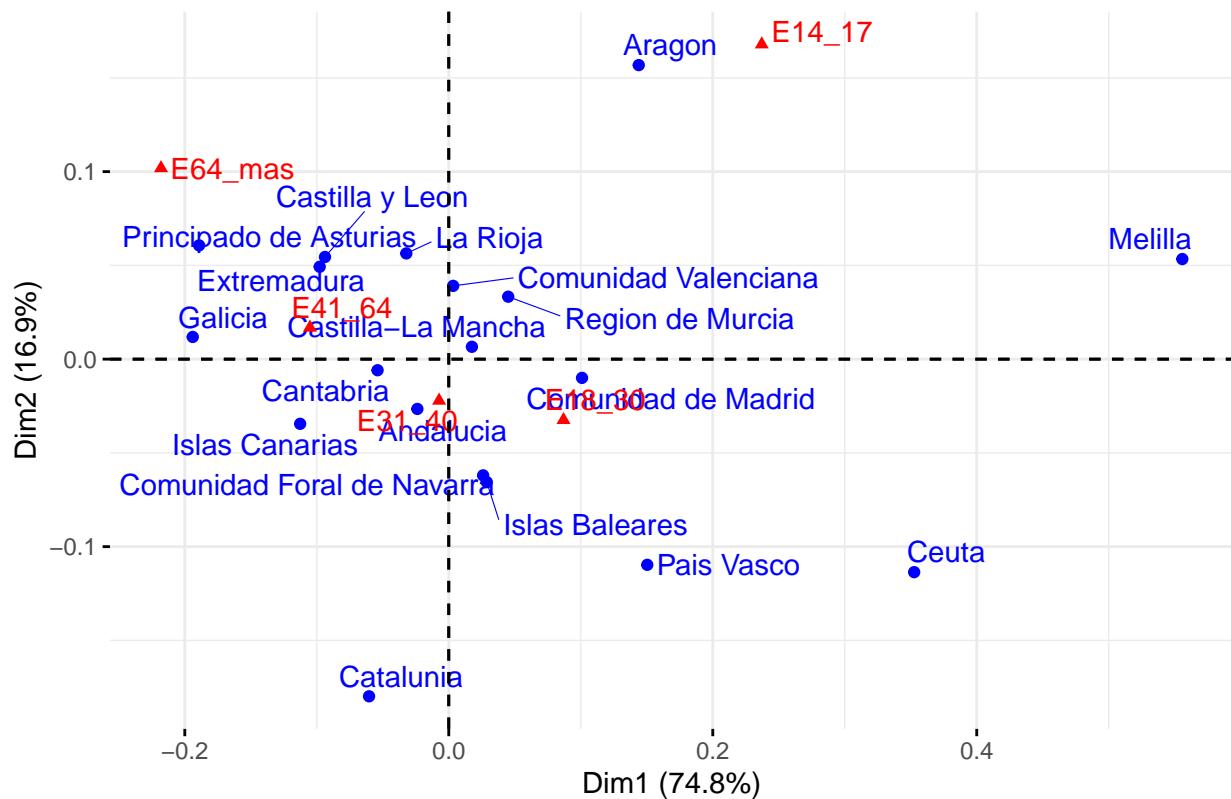


```
columnas<-subplot(cdim1, cdim2,
                     shareY = TRUE, heights = 0.9)

columnas %>% layout(annotations = list(
  list(x = 0.2 , y = 1, text = "Dimension 1", showarrow = F,
       xref='paper', yref='paper'),
  list(x = 0.8 , y = 1, text = "Dimension 2", showarrow = F,
       xref='paper', yref='paper'))
)

#representacion categorias
#fviz_ca_col(ac)
#fviz_ca_row(ac, repel = TRUE)
fviz_ca_biplot(ac, repel = TRUE)+
  ggtitle("Biplot analisis de correspondencias")+
  theme(plot.title = element_text(hjust = 0.25, face="bold"))
```

Biplot análisis de correspondencias



-----por tipología (NO SE HA INCLUIDO EN LA MEMORIA)-----

```
data <- read.xlsx('total-tiposInfracComSinResto-18.xlsx',
                  sheetIndex = 1)

nombresDelitos <- c("Asesinatos consumados",
                     "Asesinatos en grado tentativa",
                     "Delitos de lesiones", "Secuestro",
                     "Delitos contra la libertad e indemnidad
                     sexual", "Agresion sexual con penetracion",
                     "Resto de delitos contra la libertad",
                     "Robos con violencia e intimidacion",
                     "Robos con fuerza", "Robos en domicilios",
                     "Hurtos", "Sustraccion de vehiculos",
                     "Trafico de drogas")

tipos <- which(data[,2] != "<NA>")
tipos <- tipos[2:(length(tipos))]

valoresCom_Tipo <- as.numeric(as.vector(data[tipos,2]))
```

```

final<-matrix(nrow = 19, ncol = 13 )
i<-1
z<-1
while(i < length(valoresCom_Tipo)){

  final[,z]<-valoresCom_Tipo[i:(18+i)]

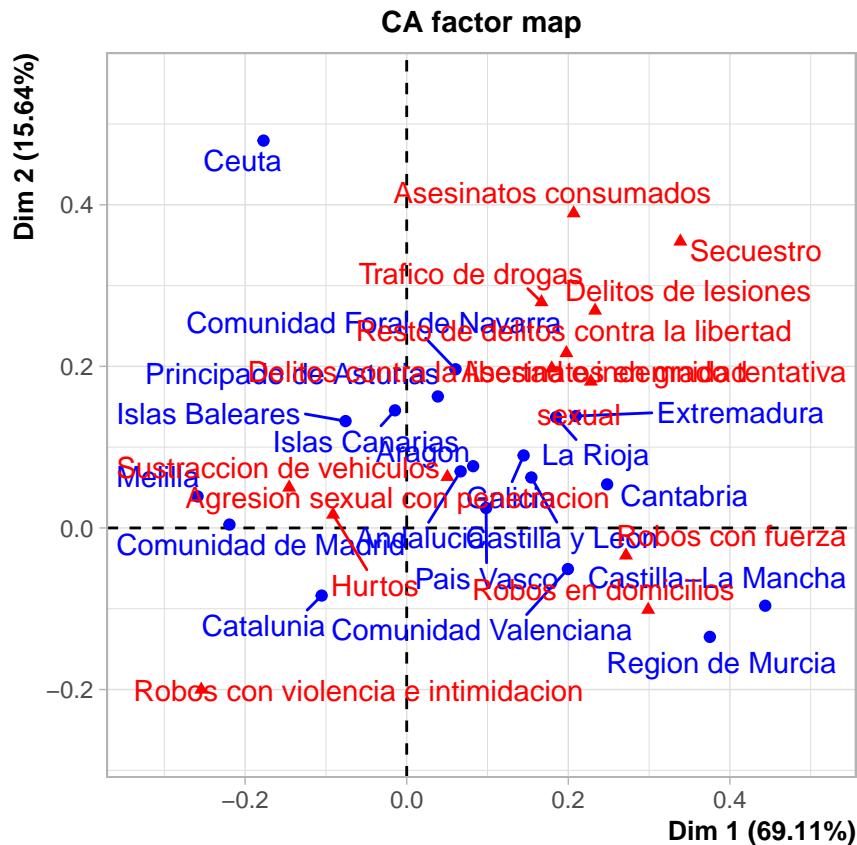
  #final<-cbind(final,aux)
  i<-i+19
  z<-z+1

}

final<-as.data.frame(final)
colnames(final)<-nombresDelitos
rownames(final)<-comunidad

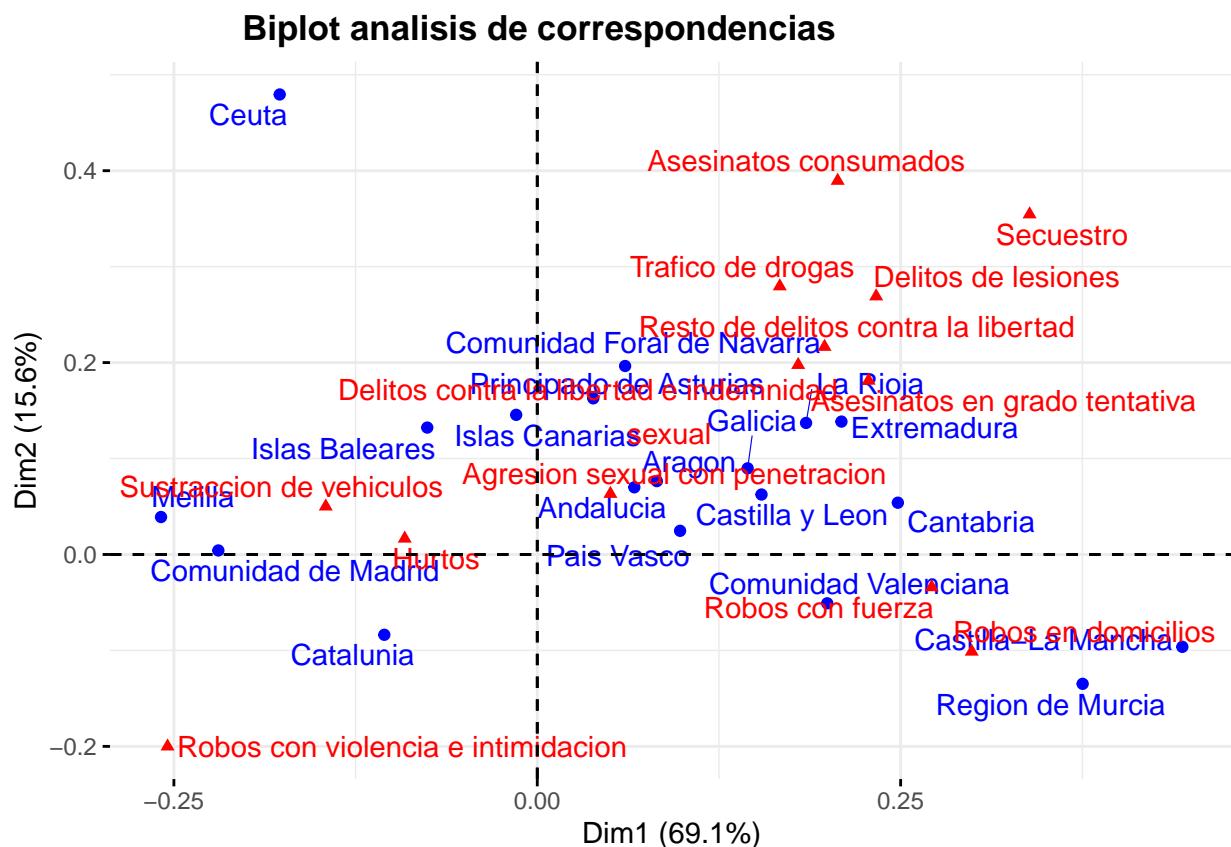
acTipo<-CA(final)

```



```
#summary(acTipo)

fviz_ca_biplot(acTipo, repel = TRUE) +
  ggtitle("Biplot analisis de correspondencias") +
  theme(plot.title = element_text(hjust = 0.25, face="bold"))
```



E.2. Análisis de componentes principales

```
=====Analisis de componentes principales =====

datos<-read.csv("tasaTiposCom-18.csv", header=T)

finalACP<-matrix(nrow = 19, ncol = 13 )
i<-1
z<-1
while(i < length(datos$value)){

  finalACP[,z]<-datos$value[i:(18+i)]

  #final<-cbind(final,aux)
```

```

i<-i+19
z<-z+1

}

finalACP<-as.data.frame(finalACP)
colnames(finalACP)<-nombresDelitos
rownames(finalACP)<-comunidad

kable(head(finalACP),
  "latex", caption = "Primeras observaciones:
  tasa criminalidad por tipología y territorio",
  booktabs = T) %>%
kable_styling(latex_options =
  c("striped","hold_position","scale_down"))

```

Cuadro 14: Primeras observaciones: tasa criminalidad por tipología y territorio

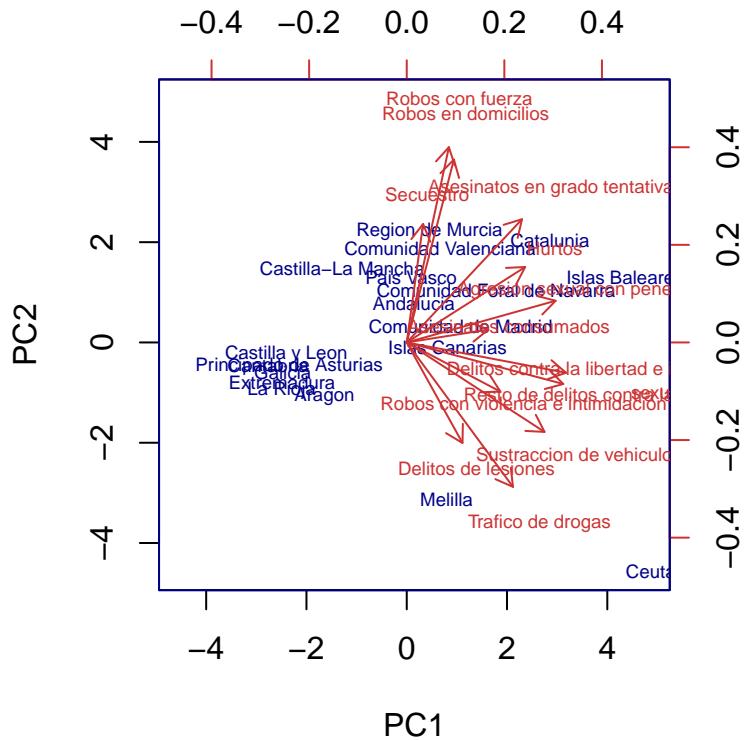
	Asesinatos consumados	Asesinatos en grado tentativo	Delitos de lesiones	Secuestro	Delitos contra la libertad e indemnidad sexual	Agresión sexual con penetración	Rusto de delitos contra la libertad	Robos con violencia e intimidación	Robos con fuerza	Robos en domicilio	Homic.	Sustacción de vehículos	Trafico de drogas
Andalucía	0.96	2.37	34.79	0.31	29.00	2.49	25.61	85.31	274.79	187.37	1112.69	82.70	43.09
Aragón	0.69	0.61	37.44	0.00	26.29	2.52	23.76	71.06	190.87	133.41	812.16	21.93	25.67
Principado de Asturias	0.68	1.07	36.68	0.29	20.13	1.65	34.62	144.71	84.42	701.10	21.20	15.62	
Isla. Baleares	0.71	1.9	70.59	0.00	51.73	6.47	45.26	109.49	389.11	270.91	1403.03	124.77	45.51
Isla. Canarias	1.13	1.93	49.91	0.09	35.30	4.14	31.16	70.26	235.56	151.90	1282.00	64.58	29.28
Cantabria	0.34	0.52	47.40	0.00	17.58	1.38	16.20	31.02	271.27	167.86	802.10	21.37	14.99

```

pca <- prcomp(finalACP, scale = TRUE)
#summary(pca)
#pca$rotation

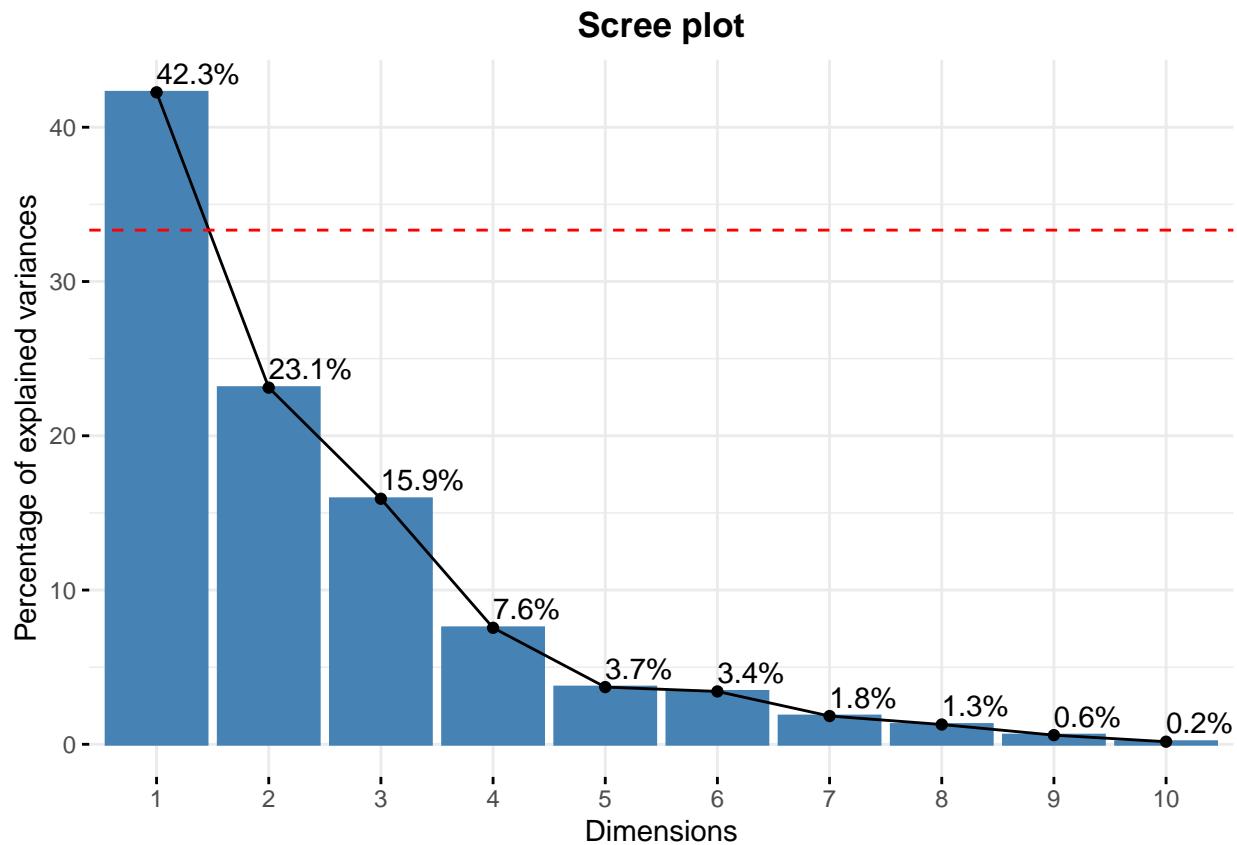
pca$rotation <- -pca$rotation
pca$x           <- -pca$x
biplot(x = pca, scale = 0, cex = 0.6, col = c("blue4", "brown3"))

```

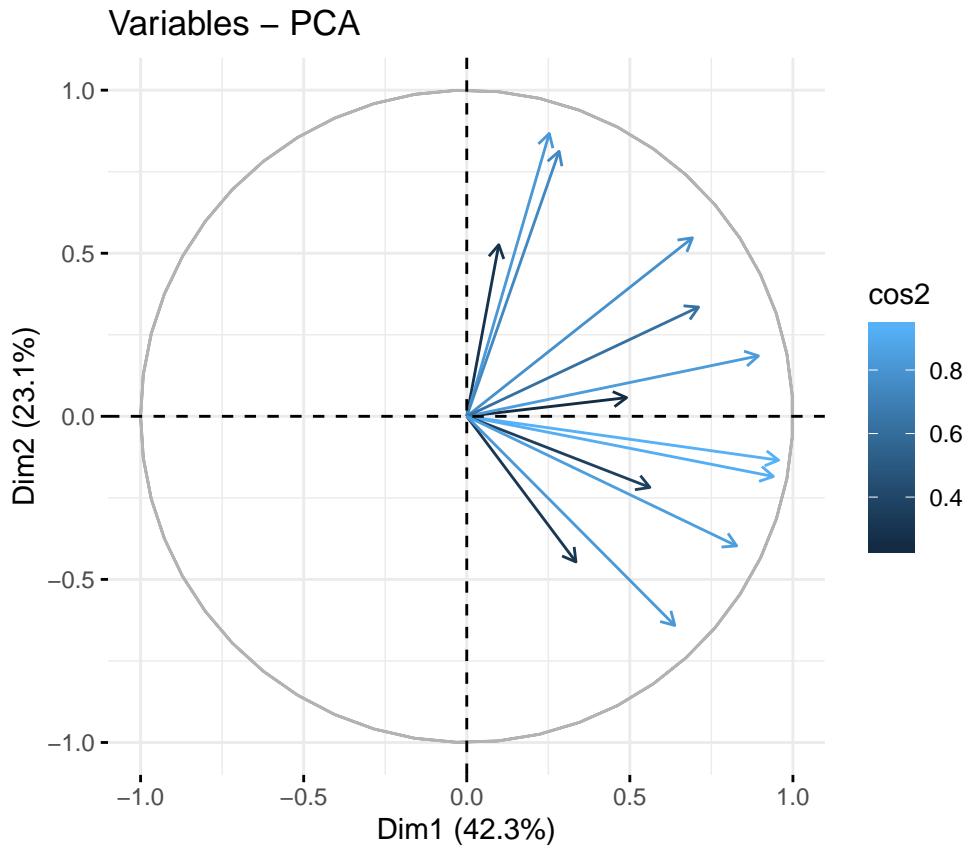


```
#screeplot
#interactivo
ggplotly(fviz_screeplot(pca, addlabels=TRUE) +
  geom_hline(yintercept=33.33, linetype=2, color="red") +
  theme(plot.title = element_text(hjust =
  0.5,face="bold")))
```

```
#no interactivo
fviz_screeplot(pca, addlabels=TRUE) +
  geom_hline(yintercept=33.33, linetype=2, color="red") +
  theme(plot.title = element_text(hjust =
  0.5,face="bold"))
```



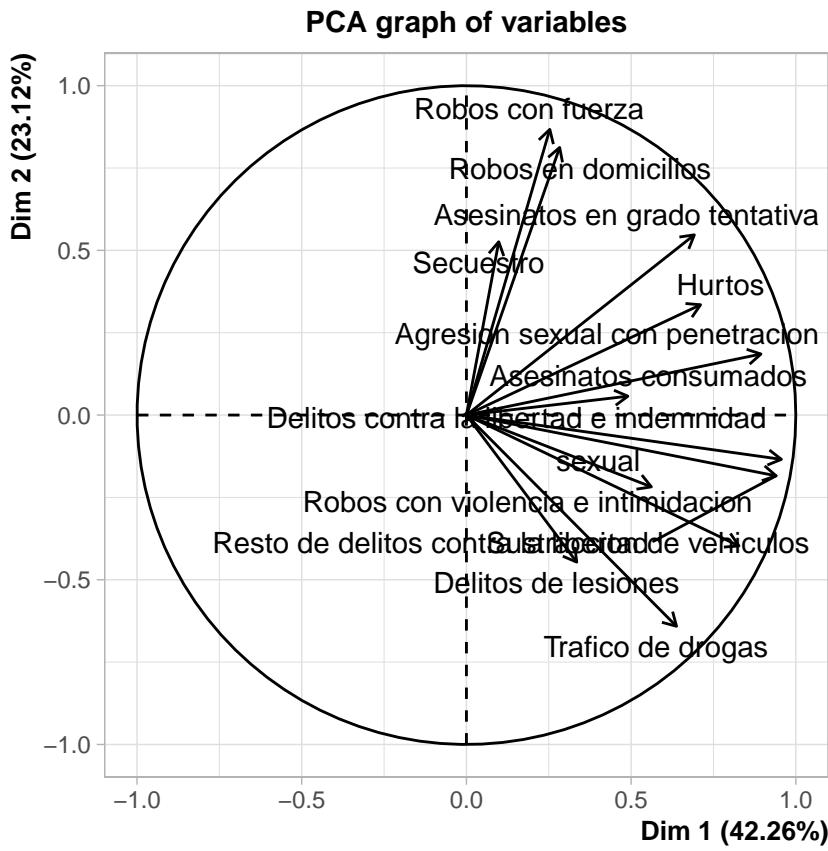
```
#nube de variables bonito
fviz_pca_var(pca, col.var = "cos2",
             geom.var = "arrow",
             labelszie = 2,
             repel = FALSE)
```



```
#biplot 3d interactivo
#pca3d(pca,show.ellipses=TRUE,
#       ellipse.ci=0.75, show.plane=TRUE, biplot=TRUE,fancy=TRUE,
#       title="Representación 3D de las componentes principales",
#       radius = 1.25)

#obtener el html del viewer
#rgl::writeWebGL("pca3D.html")

#para obtener nube de variables de forma alternativa
a<-PCA(finalACP)
```



```
#summary(a)
#plot(a, choix="var")
```

E.2.1. Clasificación K-Medias

```
===== Clasificacion: KMEANS =====

#mirar suma de cuadrado para ver el número de cluster

wss <- (nrow(finalACP)-1)*sum(apply(finalACP,2,var))

for (i in 2:13){
  wss[i] <- sum(kmeans(finalACP, centers=i)$tot.withinss)
}

aux<-data.frame(Clusters=c(1:13),
                 WSEE=wss)

#con el plot se ve que se cogen 3 cluster, cuando se ve el codo
ggplotly(
  ggplot(aux, aes(x=Clusters, y=WSEE)) +
    geom_point() +
    geom_line() +
```

```

ggttitle("Total within-clusters SSE / Numero de clusters")+
  theme(plot.title = element_text(hjust = 0.5, face="bold"))

comp<- pca$x[,1:3]

k <- kmeans(comp, 3, nstart=25, iter.max=1000)

comp<-as.data.frame(comp)
comp<- comp %>%
  mutate(size=round(apply(finalACP,1, mean),2)) %>%
  mutate(Comunidad= as.factor(rownames(finalACP))) %>%
  mutate(Cluster=as.factor(k$cluster))

colors <- c('#965F8A', '#FF7070', '#C61951')
fig <-
  plot_ly(comp, x = ~PC1, y = ~PC2, z = ~PC3,
          color = ~Cluster, size = ~size, colors = colors,
          marker = list(symbol = 'circle',
                        sizemode = 'diameter'),
          sizes = c(5, 150),
          text = ~paste('Comunidad:', Comunidad,
                        '<br>Tasa media criminalidad:', size,
                        '<br>Cluster:', Cluster))

fig <- fig %>%
  layout(title = 'Clasificación comunidades K-Medias',
         scene = list(
           xaxis = list(title = 'PC1',
                         gridcolor = 'rgb(255, 255, 255)',
                         zerolineWidth = 1,ticklen = 5,
                         gridwidth = 2),
           yaxis = list(title = 'PC2',
                         gridcolor = 'rgb(255, 255, 255)',
                         zerolineWidth = 1,ticklen = 5,
                         gridwidth = 2),
           zaxis = list(title = 'PC3',
                         gridcolor = 'rgb(255, 255, 255)',
                         zerolineWidth = 1,ticklen = 5,
                         gridwidth = 2)),
         paper_bgcolor = 'rgb(243, 243, 243)',
         plot_bgcolor = 'rgb(243, 243, 243)')

fig

```

F. Fase de modelado

F.1. Normalidad y ANOVA

```
datos<-read.xlsx('datosSexEdadYear-Modelos.xlsx', sheetIndex = 1)

ind<- which(datos[,2] != "<NA>")

infr <- as.numeric(as.vector(datos[ind[3:length(ind)],2]))
sexo<- as.factor(rep(c("Masculino", "Femenino"),9*5))
edad<-as.factor(
  rep(rep(
    c("14-17", "18-30", "31-40", "41-64", "Más de 64"),each=2), 9))
anio<-rep(
  c(2018,2017,2016,2015,2014,2013,2012,2011,2010), each=10))

dataProc <- data.frame(Infracciones=infr,
                        Sexo=sexo, Edad=edad,
                        Año=anio)

kable(head(dataProc),
      "latex", caption = "Primeras observaciones:
      datos empleados modelo Poisson", booktabs = T) %>%
kable_styling(latex_options =
              c("striped","hold_position","scale_down"))
```

Cuadro 15: Primeras observaciones: datos empleados modelo Poisson

Infracciones	Sexo	Edad	Año
15285	Masculino	14-17	2018
3385	Femenino	14-17	2018
98793	Masculino	18-30	2018
21583	Femenino	18-30	2018
80054	Masculino	31-40	2018
17181	Femenino	31-40	2018

```
attach(dataProc)
```

```

#=====normalidad =====

library(nortest)

lillie.test(Infracciones)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: Infracciones
## D = 0.29982, p-value < 2.2e-16

ks.test(Infracciones, pnorm ,
        mean(Infracciones), sd(Infracciones))

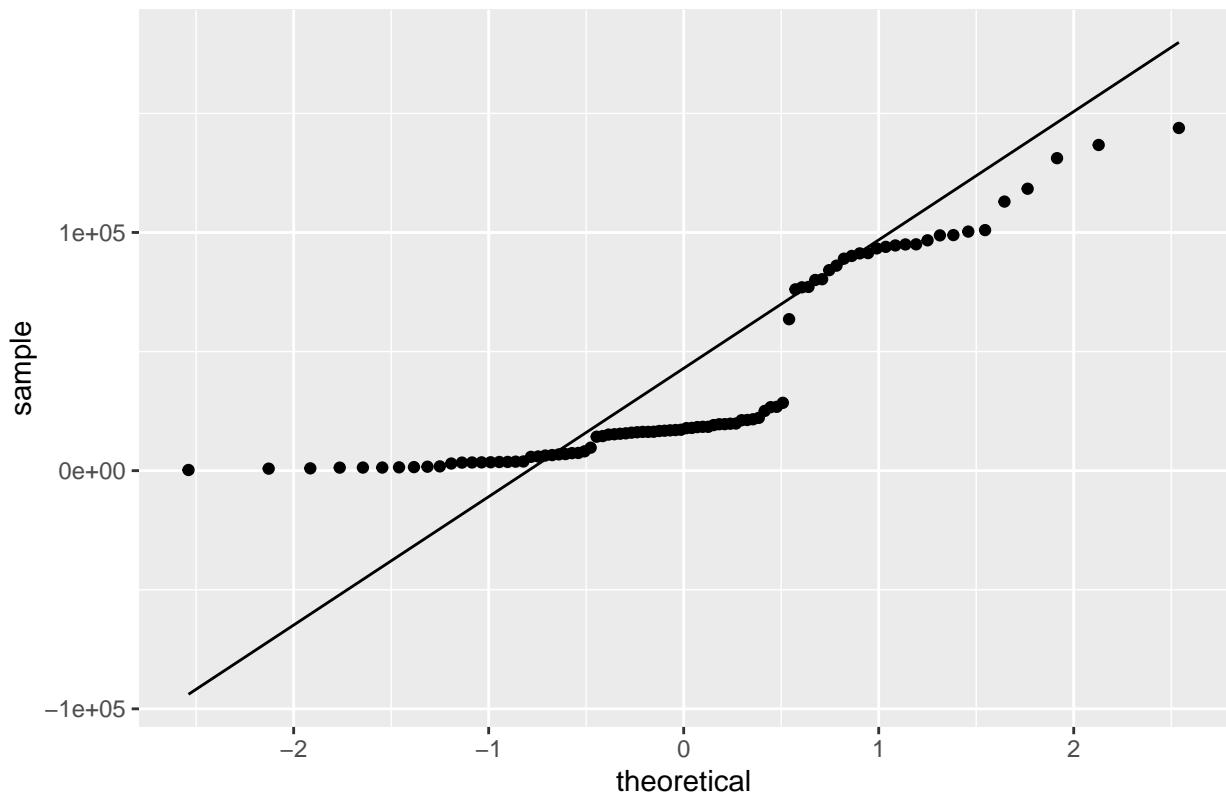
##
##  One-sample Kolmogorov-Smirnov test
##
## data: Infracciones
## D = 0.29982, p-value = 1.878e-07
## alternative hypothesis: two-sided

#graficos normalidad respuesta
grafQQ<-ggplot(dataProc, aes(sample = Infracciones)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Normalidad de la variable objetivo") +
  theme(plot.title = element_text(hjust = 0.5,face="bold"))

grafQQ

```

Normalidad de la variable objetivo



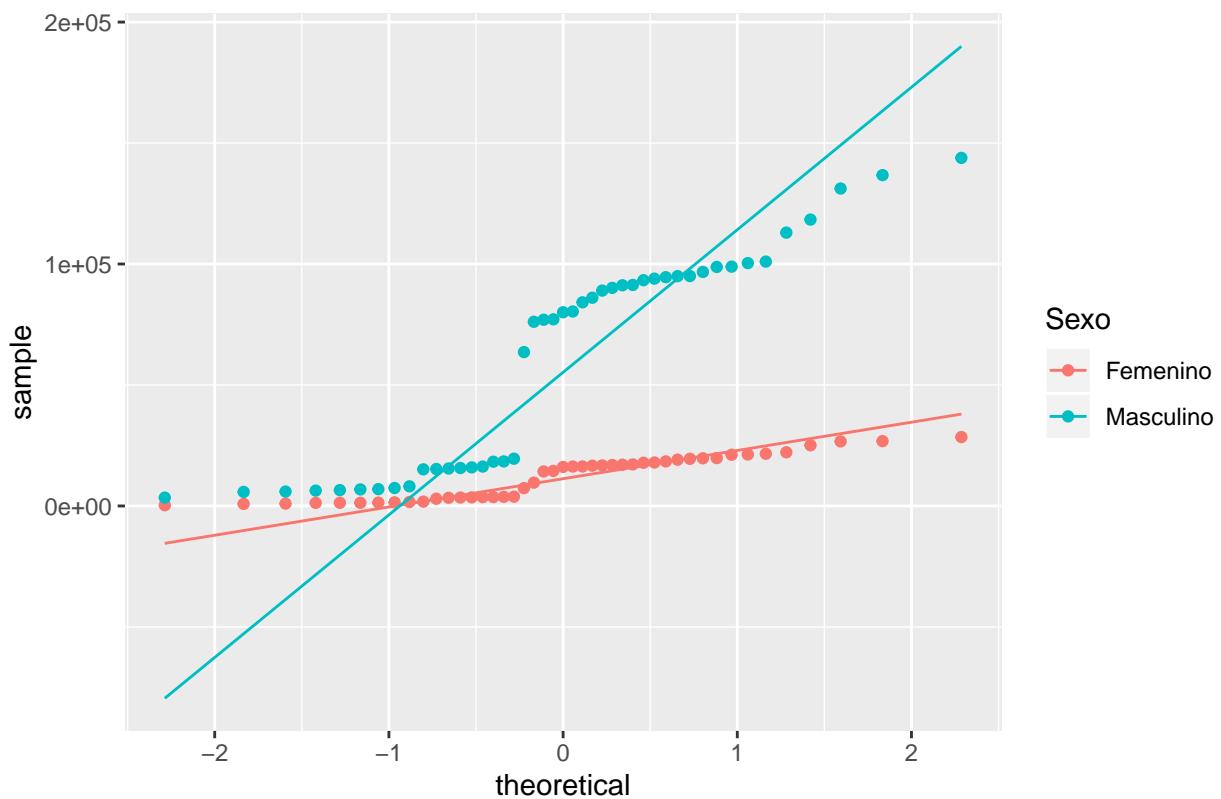
```
#ggplotly(grafQQ)

#graficos normalidad respuesta por grupos sexo
grafQQsexo<-ggplot(dataProc,
                     aes(sample = Infracciones,
                         colour=Sexo, text=paste("Sexo:", Sexo))) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Normalidad de la variable objetivo por sexo") +
  theme(plot.title = element_text(hjust = 0.5, face="bold"))

ggplotly(grafQQsexo, tooltip = "text")

grafQQsexo
```

Normalidad de la variable objetivo por sexo

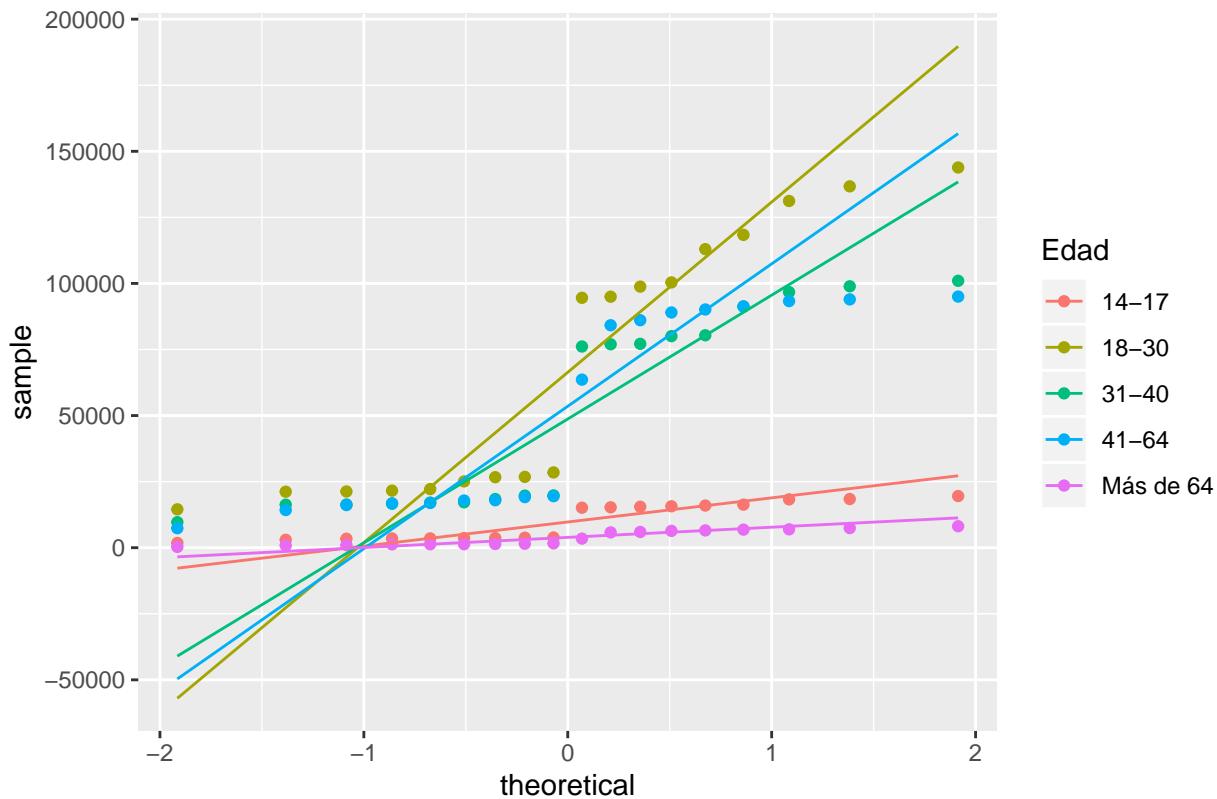


```
#graficos normalidad respuesta por grupos edad
grafQQedad<-ggplot(dataProc,
                     aes(sample = Infracciones, colour=Edad,
                         text=paste("Edad:", Edad))) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Normalidad de la variable objetivo por edad") +
  theme(plot.title = element_text(hjust = 0.5, face="bold"))

ggplotly(grafQQedad, tooltip = "text" )

grafQQedad
```

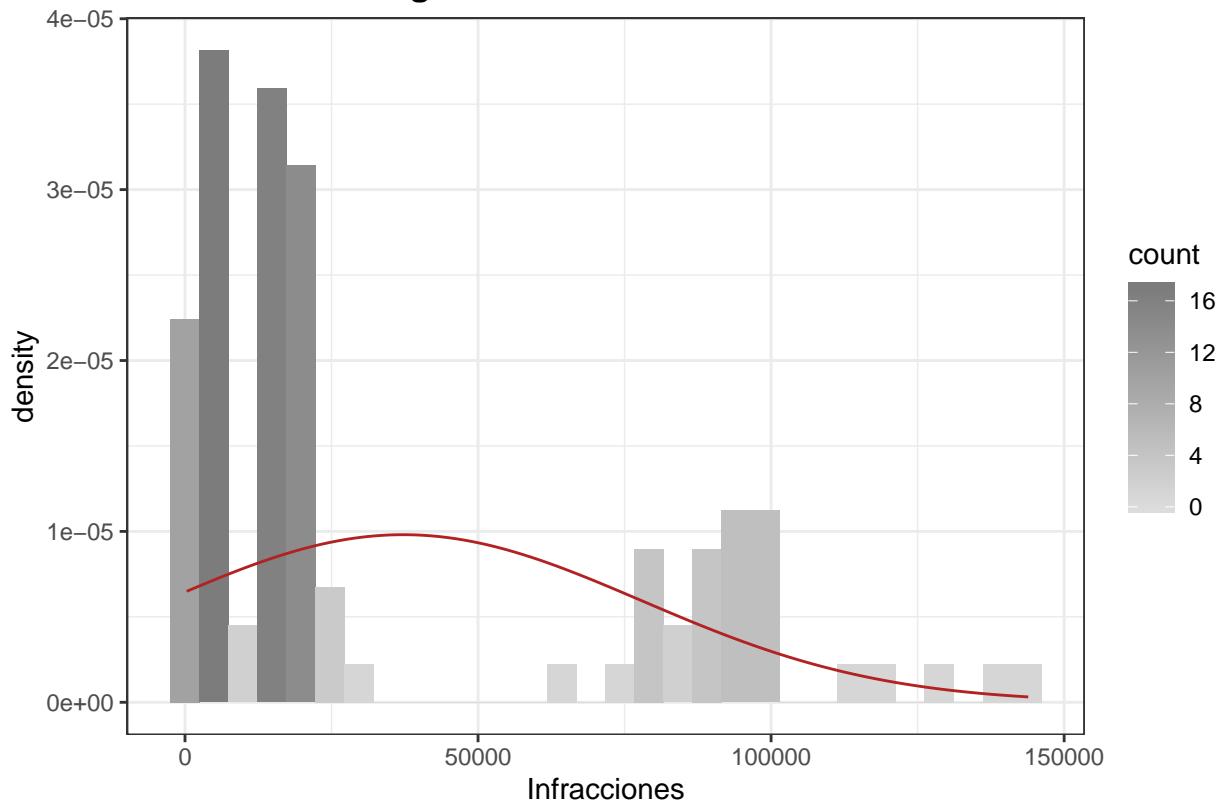
Normalidad de la variable objetivo por edad



```
grafHist<-ggplot(data = dataProc, aes(x = Infracciones)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
                args = list(mean = mean(dataProc$Infracciones),
                            sd = sd(dataProc$Infracciones))) +
  ggtitle("Histograma / Curva normal teórica") +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5,face="bold"))

#ggplotly(grafHist)
grafHist
```

Histograma / Curva normal teórica



```
#=====Modelo anova ======
```

```
anovaInfr <- aov(Infacciones~Sexo+Edad+Año, data=dataProc)
```

```
#summary(anovaInfr)
```

```
residuales<-data.frame(Residuo=anovaInfr$residuals,
```

```
                      Ajustado=anovaInfr$fitted.values,
```

```
                      Sexo=dataProc$Sexo,
```

```
                      Edad=dataProc$Edad)
```

```
grafResi<-ggplot(residuales, aes(x=Ajustado, y = Residuo)) +
```

```
  geom_point(aes(color=Sexo)) +
```

```
  geom_hline(yintercept = 0, lty=3) +
```

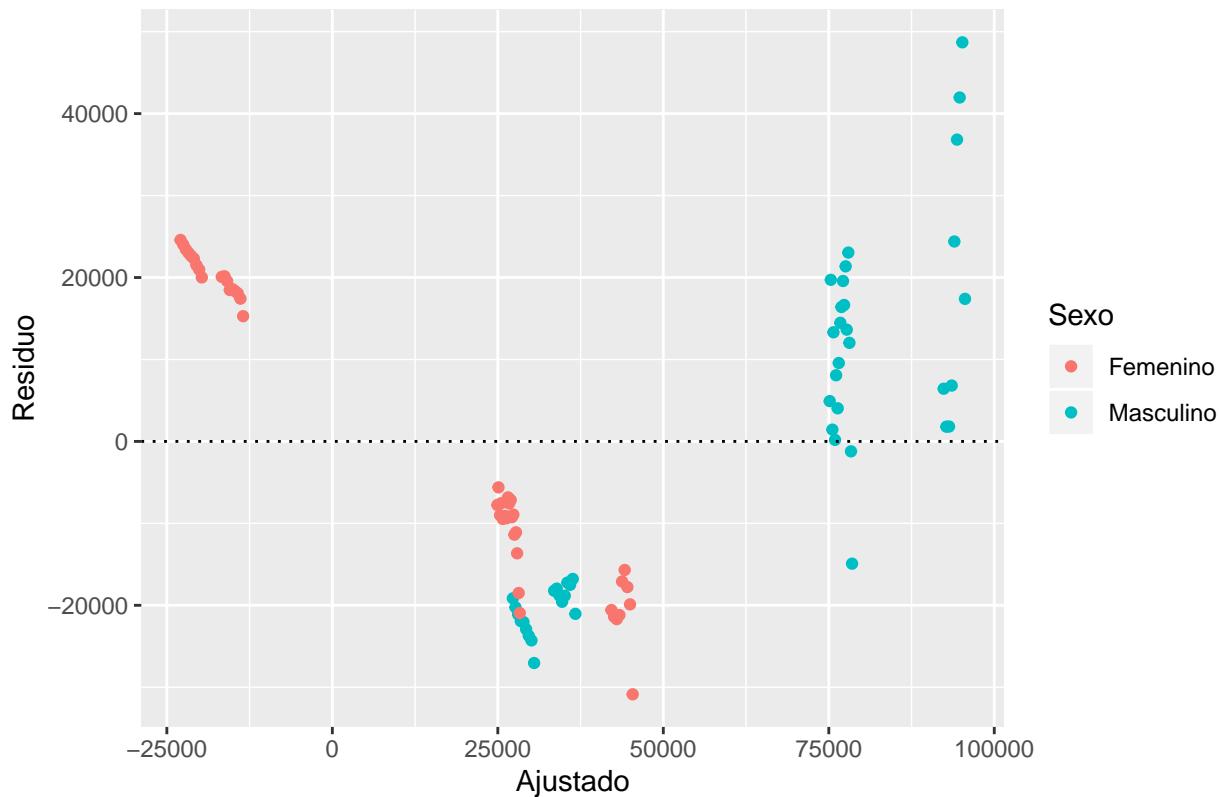
```
  ggtitle("Gráfico de dispersión") +
```

```
  theme(plot.title = element_text(hjust = 0.5,face="bold"))
```

```
ggplotly(grafResi)
```

```
grafResi
```

Gráfico de dispersión



F.2. Modelo de Poisson

```

datos<-read.xlsx('datosSexEdadYear-Modelos.xlsx', sheetIndex = 1)

ind<- which(datos[,2] != "<NA>")

infr <- as.numeric(as.vector(datos[ind[3]:length(ind)],2)))
sexo<- as.factor(rep(c("Masculino", "Femenino"),9*5))
edad<-as.factor(
  rep(rep(
    c("14-17", "18-30", "31-40", "41-64", "Más de 64"),each=2), 9))
anio<-rep(
  c(2018,2017,2016,2015,2014,2013,2012,2011,2010), each=10))

dataProc <- data.frame(Infracciones=infr,
                        Sexo=sexo,
                        Edad=edad,
                        Año=anio)

#===== GLM poisson =====

```

```

#modelos usando covariables individuales -> HAY SOBREDISPERSION

modSexo<- glm (Infracciones ~ Sexo,
                 family=poisson(link="log") ,data = dataProc)
#kable(modSexo$coefficients,
#      "latex", caption = "Modelo solo sexo",
#      booktabs = T) %>%
#kable_styling(latex_options =
#              c("striped","hold_position"))

modAño<- glm (Infracciones ~ Año,
                 family=poisson(link="log") ,data = dataProc)
#kable(modAño$coefficients,
#      "latex", caption = "Modelo solo año",
#      booktabs = T) %>%
#kable_styling(latex_options =
#              c("striped","hold_position"))

modEdad<- glm (Infracciones ~ Edad,
                 family=poisson(link="log") ,data = dataProc)
#kable(modEdad$coefficients,
#      "latex", caption = "Modelo solo edad",
#      booktabs = T) %>%
#kable_styling(latex_options =
#              c("striped","hold_position","scale_down"))

#modelos covariables individuales -> PALIANDO LA SOBREDISPERSION

modSexoNB<-glm.nb(Infracciones ~ Sexo, link="log",
                    data= dataProc)
#kable(modSexoNB$coefficients,
#      "latex", caption = "Modelo solo sexo sin sobredispersión",
#      booktabs = T) %>%
#kable_styling(latex_options =
#              c("striped","hold_position"))

modEdadNB<-glm.nb(Infracciones ~ Edad, link="log",
                    data= dataProc)
#kable(modEdadNB$coefficients,
#      "latex", caption = "Modelo solo edad sin sobredispersión",
#      booktabs = T) %>%
#kable_styling(latex_options =
#              c("striped","hold_position","scale_down"))

modAñoNB<-glm.nb(Infracciones ~ Año, link="log",
                   data= dataProc)

```

```

#kable(modAñoNB$coefficients,
#       "latex", caption = "Modelo solo año sin sobredispersión",
#       booktabs = T) %>%
#kable_styling(latex_options =
#              c("striped","hold_position"))

#vemos que por separado son todas significativas salvo edad
#añadimos Edad
modSexoEdadNB<- glm.nb(Infracciones ~ Sexo + Edad,link="log",
                         data= dataProc)
#kable(modSexoEdadNB$coefficients,"latex",
#caption = "Modelo sexo y edad sin sobredispersión",
#booktabs = T) %>%
#kable_styling(latex_options =
#              c("striped","hold_position","scale_down"))

#añadimos año
modSexoEdadAñoNB<-
  glm.nb (Infracciones ~ Sexo + Edad + Año,link="log",
          data= dataProc)
#kable(modSexoEdadAñoNB$coefficients,
#      "latex", caption = "Modelo sexo, edad y año sin
#sobredispersión",
#      booktabs = T) %>%
#kable_styling(latex_options =
#              c("striped","hold_position","scale_down"))

#comparacion los 3 modelos
#anova(modSexoNB, modSexoEdadNB, modSexoEdadAñoNB)

#manual (ejemplo)
#chis<-2*(logLik(modSexoEdadAñoNB)-logLik(modSexoEdadNB))
#1-pchisq(chis[1],1)

#===== Interacciones =====

mod1 <- glm.nb(Infracciones ~ Año * Edad + Sexo , link="log",
                data= dataProc)
#kable(mod1$coefficients,
#      "latex", caption = "Modelo interacciones (1)",
#      booktabs = T) %>%
#kable_styling(latex_options =
#              c("striped","hold_position","scale_down"))

```

```

#descartado
mod2 <- glm.nb(Infacciones ~ Año + Edad * Sexo , link="log",
                data= dataProc)
#kable(mod2$coefficients,
#      "latex", caption = "Modelo interacciones (2)",
#      booktabs = T) %>%
#kable_styling(latex_options =
#               c("striped","hold_position","scale_down"))

mod3 <- glm.nb(Infacciones ~ Edad + Año * Sexo , link="log",
                data= dataProc)
#kable(mod3$coefficients,
#      "latex", caption = "Modelo interacciones (3)",
#      booktabs = T) %>%
#kable_styling(latex_options =
#               c("striped","hold_position","scale_down"))

mod4 <- glm.nb(Infacciones ~ Año * Edad * Sexo , link="log",
                data= dataProc)
#kable(mod4$coefficients,
#      "latex", caption = "Modelo interacciones (4)",
#      booktabs = T) %>%
#kable_styling(latex_options =
#               c("striped","hold_position","scale_down"))

#descartamos mod2
#del resto nos quedamos con mod3 que es el único válido

anova(modSexoEdadAñoNB, mod1)
anova(modSexoEdadAñoNB, mod2)
anova(modSexoEdadAñoNB, mod3)
anova(modSexoEdadAñoNB, mod4)

#===== MODELO ESCOGIDO: mod3 =====

===== grafico coeficientes: Barplot horizontal=====

auxDf<-data.frame(Estimación=coef(mod3)[order(coef(mod3))],
                    Covariante=c("Intercept","Edad +64",
                                "Año*SexoMasculino",
                                "Año","Edad41-64",
                                "Edad31-40","Edad18-30",
                                "SexoMasculino")) %>%
  mutate(
    Correlación=ifelse(Estimación>0,"Positiva","Negativa"))%>%

```

```

arrange(Estimación) %>%
mutate(Covariable=factor(Covariable, levels = Covariable))

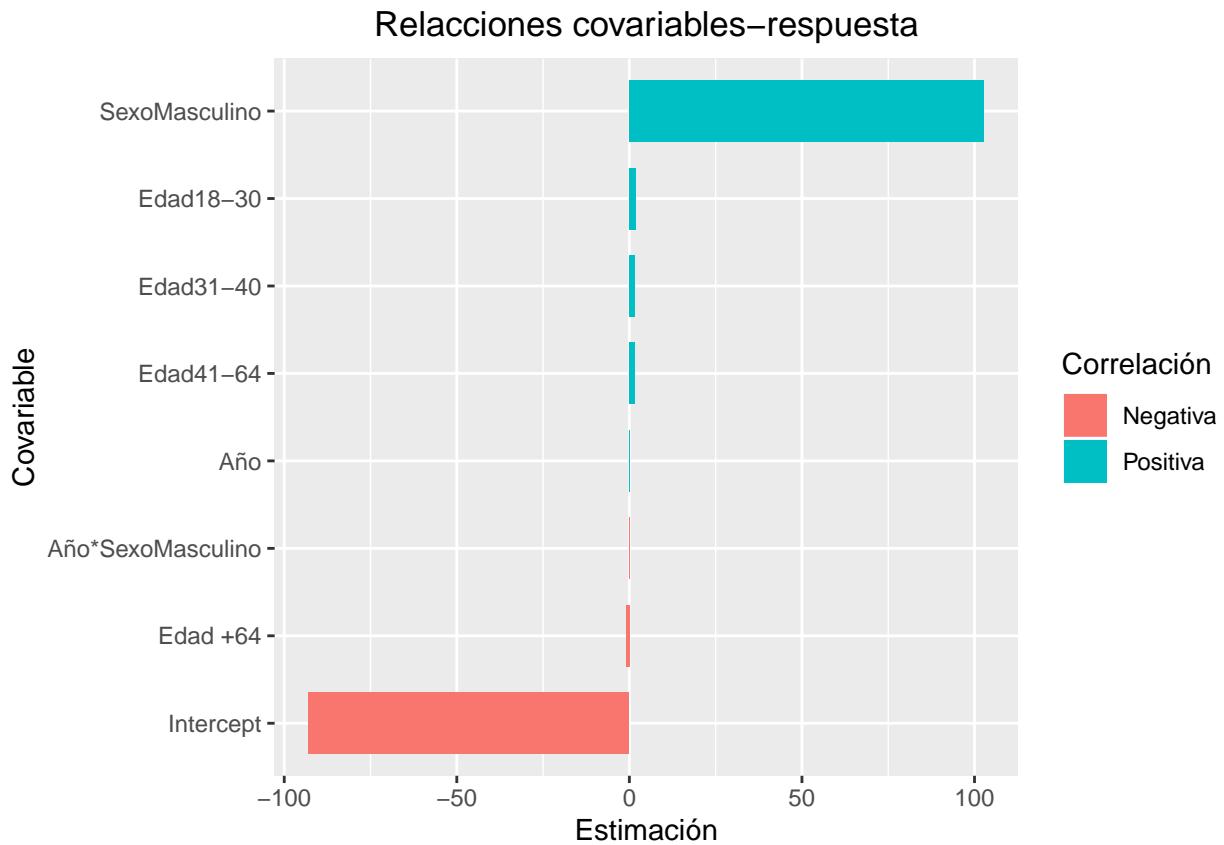
p <- ggplot(auxDf, aes(x = Covariable, y = Estimación)) +
  geom_col(aes(fill = Correlación), width = 0.7)

p<-
  p+ coord_flip() + ggtitle("Relaciones covariables-respuesta")+
  theme(plot.title = element_text(hjust=0.5))

ggplotly(p)

```

p



```

=====predicciones y graficos correspondientes=====

predicciones <- predict(mod3,type="link" ,se.fit =T)

#grafico por años
auxAño<-
  data.frame(pre=predicciones$fit, year=rep(c(18:10),each=10))

auxAño<-

```

```

data.frame(
  inf=tapply(log(dataProc$Infracciones), dataProc$Año, sum),
  predichos =
    tapply(auxAño$pre, auxAño$year, sum), año=c(2010:2018))

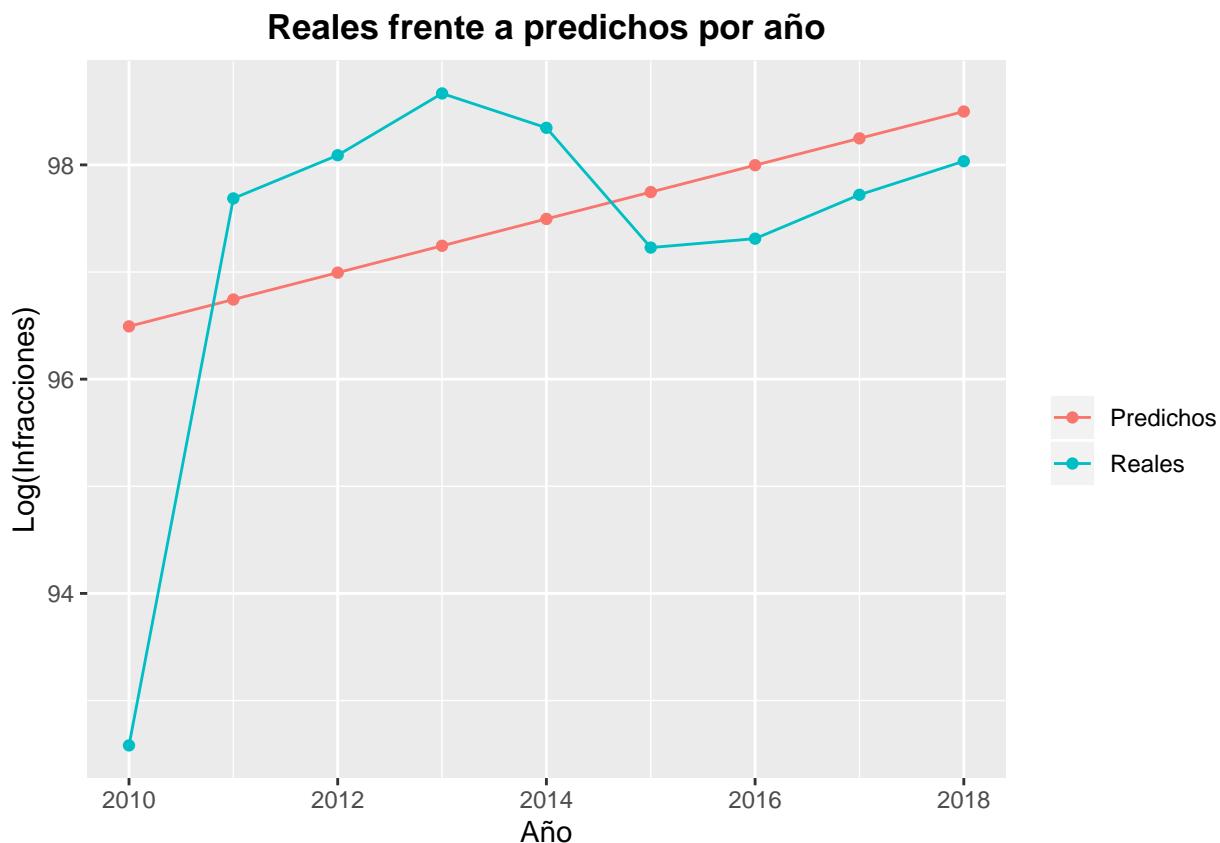
auxAño<-
  data.frame(Infracciones=c(auxAño$inf,auxAño$predichos),
             Año=rep(c(2010:2018),2),
             ind=rep(c("Reales","Predichos"),each=9))

graf <- ggplot(auxAño, aes(Año, Infracciones)) +
  geom_line(aes(col=ind)) + geom_point(aes(col=ind))
#+ geom_smooth(aes(col=ind))

graf<-graf + ggtitle("Reales frente a predichos por año") +
  labs(colour="") + ylab("Log(Infracciones)") +
  theme(plot.title = element_text(hjust = 0.5,face="bold"))

#ggplotly(graf)
graf

```



```
#===== residuales =====
```

```

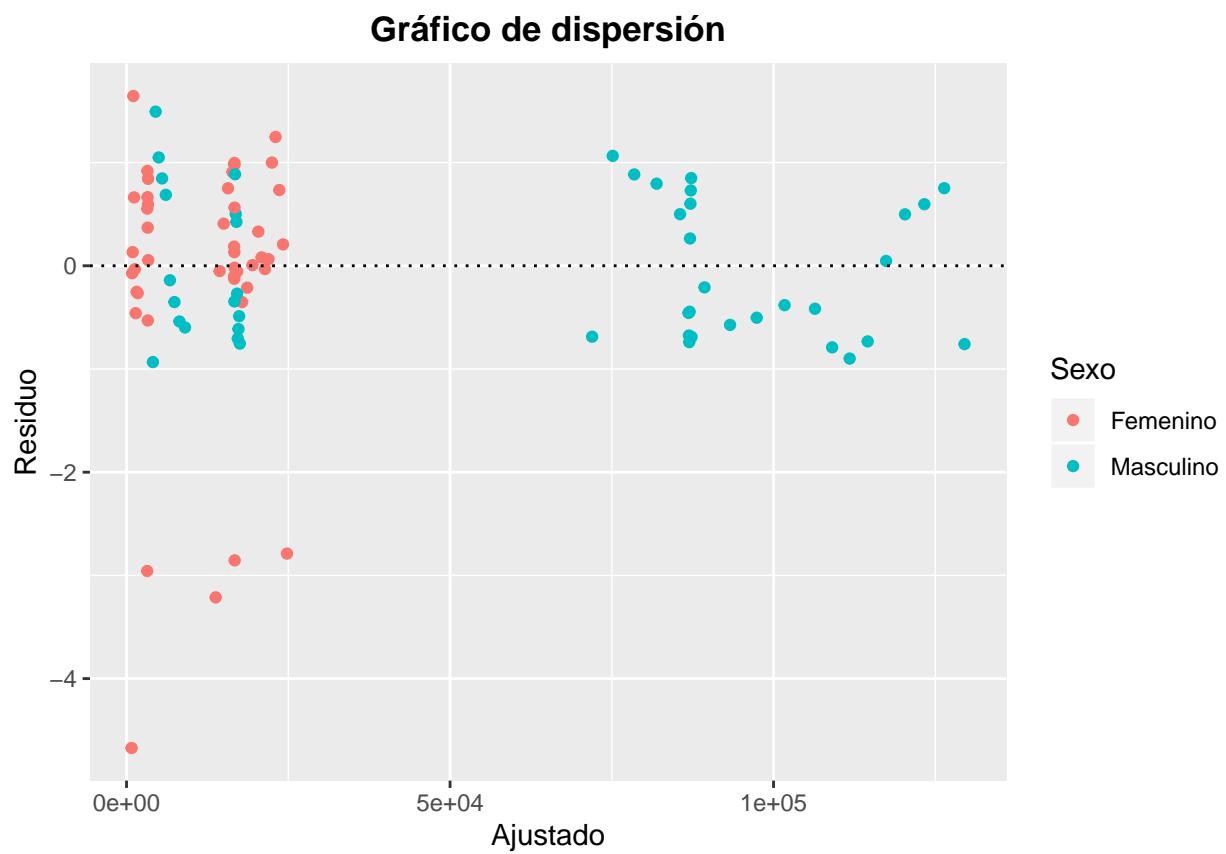
resi<-data.frame(Residuo=resid(mod1),
                  Ajustado=fitted(mod1),
                  Edad=dataProc$Edad, Sexo=dataProc$Sexo)

#distinguiendo por sexo
grafResi<-
  ggplot(resi,
         aes(x=Ajustado, y = Residuo, text=paste("Sexo:",Sexo,
                                                     "<br>Edad:",Edad,
                                                     "<br>Valor ajustado:",
                                                     round(Ajustado,2),
                                                     "<br>Residuo:",round(Residuo,2)))) +
  geom_point(aes(color=Sexo)) +
  geom_hline(yintercept = 0, lty=3) +
  ggtitle("Gráfico de dispersión") +
  theme(plot.title = element_text(hjust = 0.5,face="bold"))

ggplotly(grafResi,tooltip = "text")

```

grafResi



```

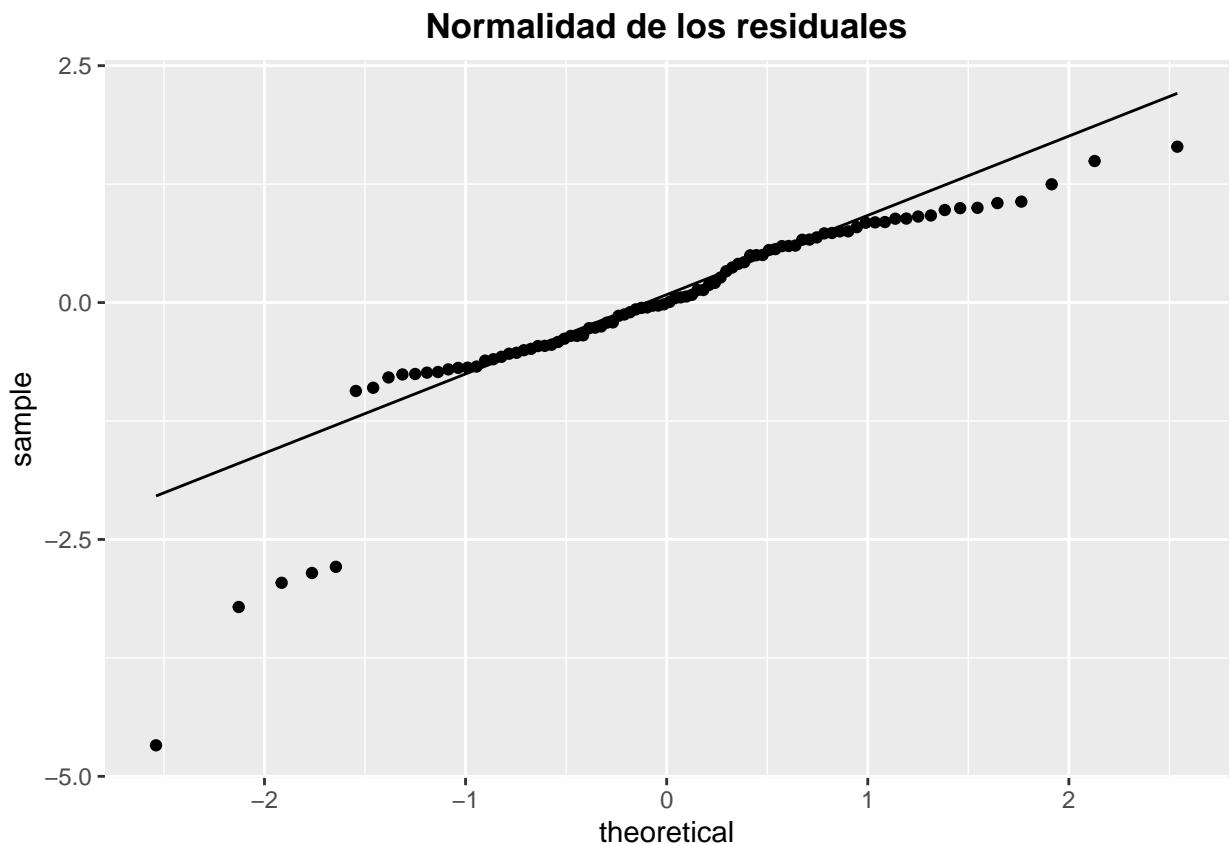
#colour=
ggplot(resi, aes(sample = Residuo)) +

```

```

stat_qq() +
stat_qq_line() +
ggtitle("Normalidad de los residuales") +
theme(plot.title = element_text(hjust = 0.5, face="bold"))

```



```

=====Capacidad predictiva de mod3 ---> Validacion cruzada 10Fold

folds<- createFolds(dataProc$Infracciones, k=10)

tasaExito<-lapply(folds, function(x){

  entre<-dataProc[-x,]
  test<-dataProc[x,]
  mod<-glm.nb(Infracciones~Edad+Año*Sexo, link="log", data=entre)
  predicc<-predict(mod, newdata=test, type="response")
  aciertos<-0
  for(i in 1:length(predicc)){
    if(predicc[i]<test$Infracciones[i]*1.15 &&
      predicc[i]>test$Infracciones[i]*0.85){
      aciertos<-aciertos+1
    }
  }
  return(aciertos/length(test$Infracciones))
})

```

```

})

tasaMedia<-mean(as.numeric(tasaExito))

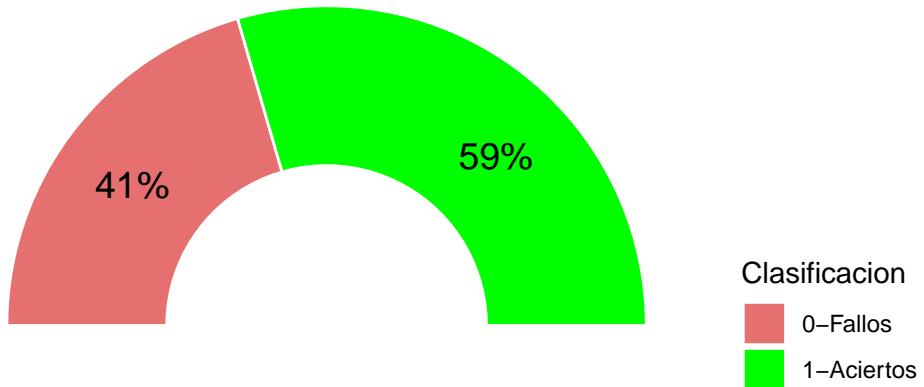
df<-data.frame(Clasificacion=c("0-Fallos","1-Aciertos"),
                 Tasa=c(1-tasaMedia,tasaMedia))

graf<-ggplot(df, aes(fill = Clasificacion, ymax = cumsum(Tasa),
                      ymin = c(0, head(cumsum(Tasa), n=-1)),
                      xmax = 2, xmin = 1)) +
  geom_rect(color="white") +
  coord_polar(theta = "y",start=-pi/2) + xlim(c(0, 2)) +
  ylim(c(0,2))

graf + theme_void() +
  scale_fill_manual(values=c("#e67070", "#00ff00")) +
  geom_text(aes(x=c(1.5,1.5), y=c(0.2,0.75),
                label = percent(Tasa)), size=5) +
  ggtitle("Porcentaje de acierto/error") +
  theme(plot.title = element_text(hjust=0.55, size=20))

```

Porcentaje de acierto/error



F.3. Redes neuronales recursivas

F.3.1. Procesamiento

```
#años 2018-2017
c1.18.17<-read.xlsx('1cuatri-1817.xlsx', sheetIndex = 1)
c2.18.17<-read.xlsx('2cuatri-1817.xlsx', sheetIndex = 1)
c3.18.17<-read.xlsx('3cuatri-1817.xlsx', sheetIndex = 1)
c4.18.17<-read.xlsx('4cuatri-1817.xlsx', sheetIndex = 1)

ind<- which(c1.18.17[,2]!="<NA>")

#cuatrimestres
primerCuatri17<-
  as.numeric(as.vector(c1.18.17[ind[2:length(ind)],2]))
primerCuatri18<-
  as.numeric(as.vector(c1.18.17[ind[2:length(ind)],3]))
segundoCuatriAux17<-
  as.numeric(as.vector(c2.18.17[ind[2:length(ind)],2]))
segundoCuatriAux18<-
  as.numeric(as.vector(c2.18.17[ind[2:length(ind)],3]))

segundoCuatri17<-segundoCuatriAux17-primerCuatri17
segundoCuatri18<-segundoCuatriAux18-primerCuatri18

tercerCuatriAux17<-
  as.numeric(as.vector(c3.18.17[ind[2:length(ind)],2]))
tercerCuatriAux18<-
  as.numeric(as.vector(c3.18.17[ind[2:length(ind)],3]))

tercerCuatri17<-tercerCuatriAux17-segundoCuatriAux17
tercerCuatri18<-tercerCuatriAux18-segundoCuatriAux18

cuartoCuatriAux17<-
  as.numeric(as.vector(c4.18.17[ind[2:length(ind)],2]))
cuartoCuatriAux18<-
  as.numeric(as.vector(c4.18.17[ind[2:length(ind)],3]))

cuartoCuatri17<-cuartoCuatriAux17-tercerCuatriAux17
cuartoCuatri18<-cuartoCuatriAux18-tercerCuatriAux18

años1718<-
  c(primerCuatri17,segundoCuatri17,tercerCuatri17,
    cuartoCuatri17, primerCuatri18,
    segundoCuatri18,tercerCuatri18,cuartoCuatri18)
```

```

#años 2016-2015
c1.16.15<-read.xlsx('1cuatri-1615.xlsx', sheetIndex = 1)
c2.16.15<-read.xlsx('2cuatri-1615.xlsx', sheetIndex = 1)
c3.16.15<-read.xlsx('3cuatri-1615.xlsx', sheetIndex = 1)
c4.16.15<-read.xlsx('4cuatri-1615.xlsx', sheetIndex = 1)

ind<- which(c1.16.15[,2]!="<NA>")

#cuatrimestres
ident<-c(rep(c(1:19),each=8))
primerCuatri15<-
  data.frame(
    inf=as.numeric(
      as.vector(c1.16.15[ind[2:length(ind)],2])), ident=ident)

primerCuatri15<-
  tapply(primerCuatri15$inf,primerCuatri15$ident,sum)

primerCuatri16<-
  data.frame(
    inf=as.numeric(
      as.vector(c1.16.15[ind[2:length(ind)],3])), ident=ident)

primerCuatri16<-
  tapply(primerCuatri16$inf,primerCuatri16$ident,sum)

segundoCuatriAux15<-
  data.frame(
    inf=as.numeric(
      as.vector(c2.16.15[ind[2:length(ind)],2])), ident=ident)
segundoCuatriAux15<-
  tapply(segundoCuatriAux15$inf,segundoCuatriAux15$ident,sum)

segundoCuatriAux16<-
  data.frame(
    inf=as.numeric(
      as.vector(c2.16.15[ind[2:length(ind)],3])), ident=ident)
segundoCuatriAux16<-
  tapply(segundoCuatriAux16$inf,segundoCuatriAux16$ident,sum)

segundoCuatri15<-segundoCuatriAux15-primerCuatri15
segundoCuatri16<-segundoCuatriAux16-primerCuatri16

#tercero
tercerCuatriAux15<-
  data.frame(
    inf=as.numeric(

```

```

    as.vector(c3.16.15[ind[2:length(ind)],2])), ident=ident)
tercerCuatriAux15<-
  tapply(tercerCuatriAux15$inf,tercerCuatriAux15$ident,sum)

tercerCuatriAux16<-
  data.frame(
    inf=as.numeric(
      as.vector(c3.16.15[ind[2:length(ind)],3])), ident=ident)
tercerCuatriAux16<-
  tapply(tercerCuatriAux16$inf,tercerCuatriAux16$ident,sum)

tercerCuatri15<-tercerCuatriAux15-segundoCuatriAux15
tercerCuatri16<-tercerCuatriAux16-segundoCuatriAux16

#cuarto
cuartoCuatriAux15<-
  data.frame(
    inf=as.numeric(
      as.vector(c4.16.15[ind[2:length(ind)],2])), ident=ident)
cuartoCuatriAux15<-
  tapply(cuartoCuatriAux15$inf,cuartoCuatriAux15$ident,sum)

cuartoCuatriAux16<-
  data.frame(
    inf=as.numeric(
      as.vector(c4.16.15[ind[2:length(ind)],3])), ident=ident)
cuartoCuatriAux16<-
  tapply(cuartoCuatriAux16$inf,cuartoCuatriAux16$ident,sum)

cuartoCuatri15<-cuartoCuatriAux15-tercerCuatriAux15
cuartoCuatri16<-cuartoCuatriAux16-tercerCuatriAux16

años1516<-
  c(primerCuatri15,segundoCuatri15,tercerCuatri15,
    cuartoCuatri15, primerCuatri16,
    segundoCuatri16,tercerCuatri16,cuartoCuatri16)

años15161718<-c(años1516,años1718)

datosFinales<-
  data.frame(Infracciones=años15161718,
             Cuatrimestre= rep(rep(c(1:4),each=19),4),
             Año=rep(c(2015:2018),each=4*19))

dim(datosFinales)

## [1] 304   3

```

```

===== Datos de prueba: Año 2019 =====

c1.19<-read.xlsx('1cuatri-19.xlsx', sheetIndex = 1)
c2.19<-read.xlsx('2cuatri-19.xlsx', sheetIndex = 1)
c3.19<-read.xlsx('3cuatri-19.xlsx', sheetIndex = 1)

ind<- which(c1.19[,2] != "<NA>")

#cuatrimestres
primerCuatri19<-
  as.numeric(as.vector(c1.19[ind[2:length(ind)],2]))
segundoCuatriAux19<-
  as.numeric(as.vector(c2.19[ind[2:length(ind)],2]))

segundoCuatri19<-segundoCuatriAux19-primerCuatri19

tercerCuatriAux19<-
  as.numeric(as.vector(c3.19[ind[2:length(ind)],2]))

tercerCuatri19<-tercerCuatriAux19-segundoCuatriAux19

test19<-data.frame(
  Infracciones=c(primerCuatri19, segundoCuatri19, tercerCuatri19),
  Cuatrimestre=rep(c(1:3),each=19), Año=rep(2019,57))

conjuntoEnter0<-as.data.frame(rbind(datosFinales,test19))
dim(conjuntoEnter0)

## [1] 361   3

kable(head(datosFinales),
      "latex", caption = "Primeras observaciones: serie temporal",
      booktabs = T) %>%
  kable_styling(latex_options = c("striped","hold_position"))

```

Cuadro 16: Primeras observaciones: serie temporal

Infracciones	Cuatrimestre	Año
131219	1	2015
14641	1	2015
9778	1	2015
17011	1	2015
36013	1	2015
5933	1	2015

```
#===== Exportar datos para PYTHON =====

#file<-paste(getwd(), "/test2019.xlsx",sep="")
#write.xlsx(test19, file)

#file<-paste(getwd(), "/entrenamiento15-18.xlsx",sep="")
#write.xlsx(datosFinales, file)

#file<-paste(getwd(), "/datos15-19.xlsx",sep="")
#write.xlsx(conjuntoEntero, file)
```

F.3.2. Recursiva Simple

```
===== PYTHON =====

import numpy as np
from keras.models import Sequential
from keras.layers import SimpleRNN
from keras.layers import LSTM
from keras.layers import GRU
from keras.layers import Dense
import pandas as pd
import keras.backend as K
import matplotlib.pyplot as plt
from sklearn.preprocessing import minmax_scale

def porcentaje_margen(y_true, y_pred):
    margen = 0.15
    yy = K.sum(K.cast(K.less(
        K.abs((y_pred/y_true)-1.0), margen), dtype=float))
    return yy/K.cast(K.shape(y_pred)[0], dtype=float)

def split_sequence(sequence,n_steps):
    X,y = list(),list()
    for i in range(len(sequence)):
        end_ix = i+n_steps
        if end_ix > len(sequence)-1:
            break
        seq_x,seq_y = sequence[i:end_ix],sequence[end_ix]
        X.append(seq_x)
        y.append(seq_y)
    return np.array(X),np.array(y)

tabla_historico=pd.read_excel('datos15-19.xlsx')
#print(tabla_historico)

#Tamaño de las secuencias
```

```

#en nuestro caso por trimestre cada comunidad
T=19

# numero de caracteristicas
n_features=1

#conjunto X (19 observaciones por serie)
# observacion y (ultimo valor de cada serie)
x,y=split_sequence(tabla_historico['Infracciones'],T)

x=x.reshape(x.shape[0], x.shape[1],1)

#Train y test
#desordeno aleatoriamente los indices
#cojo 2/3 para entrenamiento y 1/3 para test

indices=np.arange(y.shape[0],dtype=int)
np.random.shuffle(indices)
train=indices[0:int(y.shape[0]*2/3)]
test=indices[int(y.shape[0]*2/3):]

#Entrenamiento y validación
SRNN = Sequential()
SRNN.add(SimpleRNN(
    input_shape=(T,n_features),
    units=30,activation='relu',
    return_sequences=True))

SRNN.add(SimpleRNN(10,return_sequences=True,activation="relu"))
SRNN.add(SimpleRNN(5,return_sequences=False,activation="relu"))

#Capa de salida de un perceptrón multicapa
SRNN.add(Dense(1))

SRNN.compile(
    optimizer="adam",loss="mse",metrics=[porcentaje_margen])

history=SRNN.fit(
    x[train], y[train], epochs=100, verbose=1,
    validation_data=(x[test],y[test]))

plt.plot(history.epoch,
         history.history['val_porcentaje_margen'],
         label='validation')

plt.plot(history.epoch,

```

```

    history.history['porcentaje_margen'],
    label="training")

plt.legend()
plt.title('Evolución éxito')
plt.xlabel('Epocas')
plt.show()

```

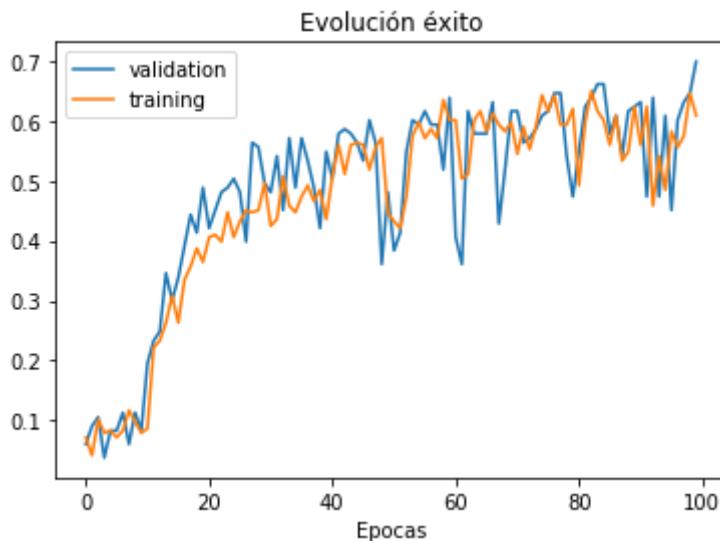


Figura 36: Red neuronal recursiva simple

F.3.3. LSTM

```

#Entrenamiento y validación
LSTM_model = Sequential()

LSTM_model.add(LSTM(
    input_shape=(T,n_features),
    units=30,activation="relu",
    return_sequences=True))

LSTM_model.add(LSTM(10,return_sequences=True,activation="relu"))
LSTM_model.add(LSTM(5,return_sequences=False,activation="relu"))

#Capa de salida de un perceptrón multicapa
LSTM_model.add(Dense(1))

LSTM_model.compile(
    optimizer="adam",loss="mse",metrics=[porcentaje_margen])

historyLSTM=LSTM_model.fit(

```

```

x[train], y[train], epochs=100, verbose=1,
validation_data=(x[test],y[test]))
```

```

plt.plot(historyLSTM.epoch,
         historyLSTM.history['val_porcentaje_margen'],
         label='validation')

plt.plot(historyLSTM.epoch,
         historyLSTM.history['porcentaje_margen'],
         label="training")
```

```

plt.legend()
plt.title('Evolución éxito')
plt.xlabel('Epocas')
plt.show()
```

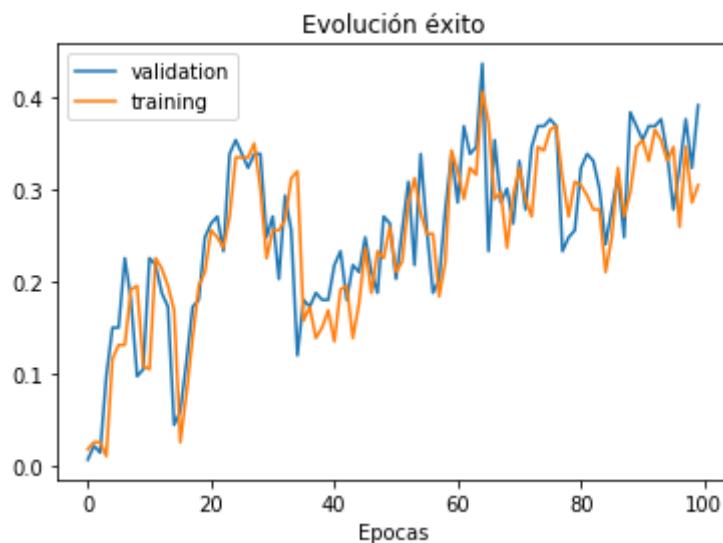


Figura 37: Red neuronal recursiva LSTM

G. Gráficos no incluidos en la memoria

```

#----poblacion----
```

```

comunidades<-c("Andalucía" , "Aragón", "Principado de Asturias",
               "Islas Baleares", "Islas Canarias",
               "Cantabria" , "Castilla y León" ,
               "Castilla-La Mancha" ,
               "Cataluña" , "Comunidad Valenciana",
               "Extremadura", "Galicia", "Comunidad de Madrid",
               "Región de Murcia", "Comunidad Foral de Navarra",
               "País Vasco" , "La Rioja" , "Ceuta" , "Melilla")
```

```

poblacion<-read.xlsx('poblacionComSexo-18.xlsx', sheetIndex = 1)

aux <- which(poblacion[,2] != "<NA>")
aux <- aux[3:(length(aux))]

pobMasc <- as.numeric(as.vector(poblacion[aux,2]))
pobFem <- as.numeric(as.vector(poblacion[aux,3]))


#----datos infracciones---

data <- read.xlsx('total-infracSexoEdadCom-18.xlsx',
                    sheetIndex = 1)

tipos <- which(data[,2] != "<NA>")
tipos <- tipos[3:(length(tipos))]

auxMasc <- as.numeric(as.vector(data[tipos,2]))
auxFem <- as.numeric(as.vector(data[tipos,3]))

edad<-rep(c("14-17","18-30","31-40","41-64","Más de 64"),19)
sex<-c(rep("Masculino",5*19),rep("Femenino",5*19))

dfComSexoEdad<-data.frame(Edad=edad, Sexo=sex,
                           Infracciones=c(auxMasc,auxFem),
                           Indicador=as.factor(rep(c(1:19),each=5)))
)

infMasc<- dfComSexoEdad %>% filter(Sexo=="Masculino")
tasaMasc<-
  1000*tapply(infMasc$Infracciones, infMasc$Indicador,sum)/pobMasc

infFem<- dfComSexoEdad %>% filter(Sexo=="Femenino")
tasaFem<-
  1000*tapply(infFem$Infracciones, infFem$Indicador,sum)/pobFem

#===== datos paro=====

datos<-read.xlsx('tasaParoSexo-18.xlsx', sheetIndex = 1)

tipos <- which(datos[,2] != "<NA>")
tipos <- tipos[4:(length(tipos))]

auxMasc <- (as.numeric(as.vector(datos[tipos,2])) +
             as.numeric(as.vector(datos[tipos,3])) +
             as.numeric(as.vector(datos[tipos,4])) +
             as.numeric(as.vector(datos[tipos,5]))) / 4

```

```

auxFem <- (as.numeric(as.vector(datos[tipos,6])) +
            as.numeric(as.vector(datos[tipos,7]))+
            as.numeric(as.vector(datos[tipos,8])) +
            as.numeric(as.vector(datos[tipos,9]))) /4

#===== Barplot circular =====

Sexo<-rep(rep(c("Masculino","Femenino"),each=19),2)
valores<-c(tasaMasc,tasaFem, auxMasc, auxFem )

data=data.frame(
  individual=rep(comunidades,4),
  group=
    c( rep('A', 19), rep('B', 19), rep('D', 19), rep('C', 19)),
  value=valores
)

data = data %>% arrange(group, value)

empty_bar=3
to_add =
  data.frame(
    matrix(NA, empty_bar*nlevels(data$group), ncol(data)))

colnames(to_add) = colnames(data)
to_add$group=rep(levels(data$group), each=empty_bar)
data=rbind(data, to_add)
data=data %>% arrange(group)
data$id=seq(1, nrow(data))

label_data=data
number_of_bar=nrow(label_data)
angle= 90 - 360 * (label_data$id-0.5) /number_of_bar
label_data$hjust<-ifelse( angle < -90, 1, 0)
label_data$angle<-ifelse(angle < -90, angle+180, angle)

base_data=data %>%
  group_by(group) %>%
  summarize(start=min(id), end=max(id) - empty_bar) %>%
  rowwise() %>%
  mutate(title=mean(c(start, end)))

grid_data = base_data
grid_data$end =
  grid_data$end[ c( nrow(grid_data), 1:nrow(grid_data)-1)] + 1

```

```

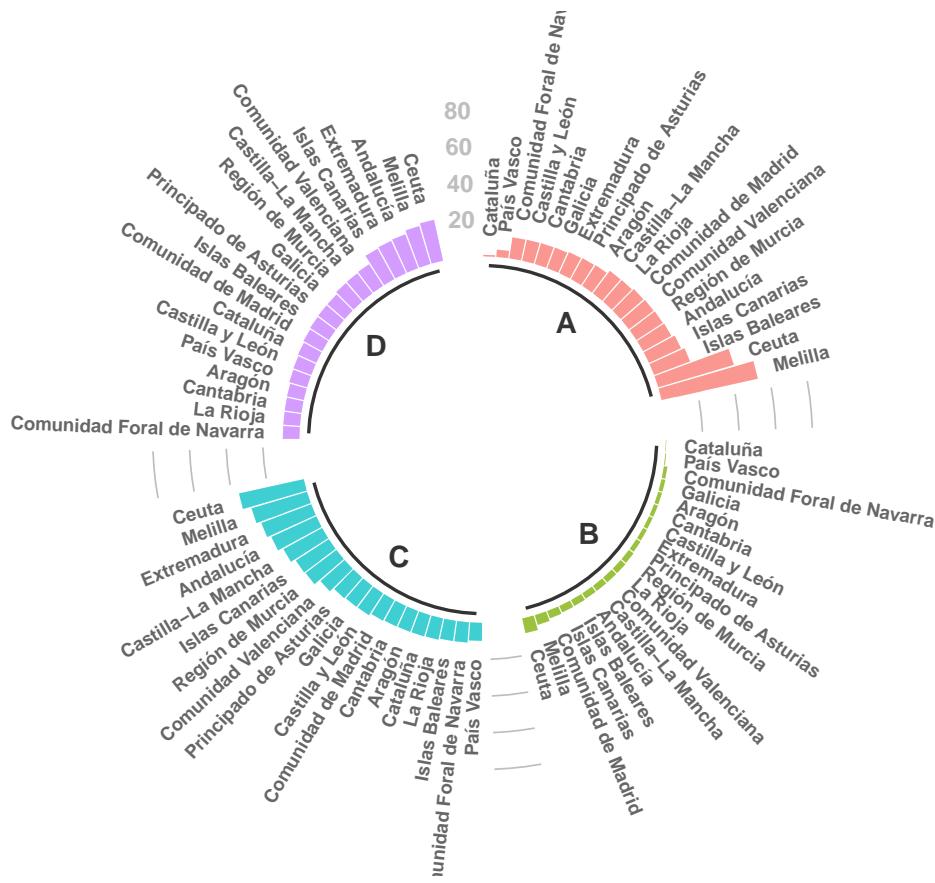
grid_data$start = grid_data$start - 1
grid_data=grid_data[-1,]

p =
  ggplot(data, aes(x=as.factor(id), y=value, fill=group)) +
  geom_bar(aes(x=as.factor(id), y=value, fill=group),
            stat="identity", alpha=0.5) +
  geom_segment(data=grid_data, aes(x = end, y = 80,
                                    xend = start, yend = 80),
               colour = "grey", alpha=1, size=0.3 ,
               inherit.aes = FALSE ) +
  geom_segment(data=grid_data, aes(x = end, y = 60,
                                    xend = start, yend = 60),
               colour = "grey", alpha=1, size=0.3 ,
               inherit.aes = FALSE ) +
  geom_segment(data=grid_data, aes(x = end, y = 40,
                                    xend = start, yend = 40),
               colour = "grey", alpha=1, size=0.3 ,
               inherit.aes = FALSE ) +
  geom_segment(data=grid_data, aes(x = end, y = 20,
                                    xend = start, yend = 20),
               colour = "grey", alpha=1, size=0.3 ,
               inherit.aes = FALSE ) +
  annotate("text", x = rep(max(data$id),4), y = c(20, 40, 60, 80),
           label = c("20", "40", "60", "80") , color="grey",
           size=3 , angle=0, fontface="bold", hjust=1) +
  geom_bar(aes(x=as.factor(id), y=value, fill=group),
            stat="identity", alpha=0.5) +
  ylim(-100,120) +
  theme_minimal() +
  theme(
    legend.position = "none",
    axis.text = element_blank(),
    axis.title = element_blank(),
    panel.grid = element_blank(),
    plot.margin = unit(rep(-1,4), "cm")) +
  coord_polar() +
  geom_text(data=label_data, aes(x=id, y=value+10,
                                 label=individual, hjust=hjust),
            color="black", fontface="bold",alpha=0.6, size=2.5,
            angle= label_data$angle, inherit.aes = FALSE ) +
  geom_segment(data=base_data, aes(x = start, y = -5,
                                    xend = end, yend = -5),
               colour = "black", alpha=0.8, size=0.6 ,
               inherit.aes = FALSE ) +
  geom_text(data=base_data, aes(x = title, y = -18,
                                label=group), hjust=c(1,1,0,0),
            color="black", fontface="bold",alpha=0.6, size=2.5,
            angle= base_data$title, inherit.aes = FALSE )

```

```
colour = "black", alpha=0.8, size=4,
fontface="bold", inherit.aes = FALSE)
```

p



```
===== Boxplot violin -> sexo y edad =====
```

```
data <- read.xlsx('total-infracSexoEdadCom-18.xlsx',
sheetIndex = 1)

tipos <- which(data[,2] != "<NA>")
tipos <- tipos[3:(length(tipos))]

auxMasc <- as.numeric(as.vector(data[tipos,2]))
auxFem <- as.numeric(as.vector(data[tipos,3]))

comunidades<-c("Andalucía" , "Aragón", "Principado de Asturias",
"Islas Baleares", "Islas Canarias",
"Cantabria" , "Castilla y León" ,
"Castilla-La Mancha" ,
```

```

    "Cataluña" , "Comunidad Valenciana",
    "Extremadura", "Galicia", "Comunidad de Madrid",
    "Región de Murcia","Comunidad Foral de Navarra",
    "País Vasco" , "La Rioja" , "Ceuta" , "Melilla")

edad<-rep(c("14-17","18-30","31-40","41-64","Más de 64"),19)
sex<-c(rep("Masculino",5*19),rep("Femenino",5*19))

dfComSexoEdad<-data.frame(Edad=edad, Sexo=sex,
                           Infracciones=c(auxMasc,auxFem),
                           Comunidad=rep(comunidades,each=5),
                           Poblacion=pob$pob18[2:dim(pob)[1]]
                           )
dfComSexoEdad <- dfComSexoEdad %>%
  mutate(Tasa=1000*Infracciones/Poblacion)

kable(head(dfComSexoEdad),"latex",
      caption = "Primeras observaciones:
tasa infracciones por edad, comunidad y sexo",
      booktabs = T) %>%
kable_styling(latex_options =
              c("striped","hold_position","scale_down"))

```

Cuadro 17: Primeras observaciones: tasa infracciones por edad, comunidad y sexo

Edad	Sexo	Infracciones	Comunidad	Poblacion	Tasa
14-17	Masculino	3141	Andalucía	8384408	0.3746239
18-30	Masculino	23758	Andalucía	1308728	18.1535048
31-40	Masculino	20183	Andalucía	1028244	19.6286096
41-64	Masculino	23286	Andalucía	1128908	20.6270130
Más de 64	Masculino	1993	Andalucía	2127685	0.9366988
14-17	Masculino	890	Aragón	580229	1.5338771

```

p <- dfComSexoEdad %>%
  plot_ly(type = 'violin') %>%
  add_trace(
    x = ~Edad[dfComSexoEdad$Sexo == 'Femenino'],
    y = ~Infracciones[dfComSexoEdad$Sexo == 'Femenino'],
    legendgroup = 'Femenino',
    scalegroup = 'Femenino',
    name = 'Femenino',
    side = 'negative',
    box = list(
      visible = T
    ),
    meanline = list(

```

```

    visible = T
),
color = I("#8dd3c7")
) %>%
add_trace(
  x = ~Edad[dfComSexoEdad$Sexo == 'Masculino'],
  y = ~Infracciones[dfComSexoEdad$Sexo == 'Masculino'],
  legendgroup = 'Masculino',
  scalegroup = 'Masculino',
  name = 'Masculino',
  side = 'positive',
  box = list(
    visible = T
),
meanline = list(
  visible = T
),
color = I("#bebada")
) %>%
layout(
  title = "Densidad y cuartiles",
  xaxis = list(
    title = "Edad"
),
yaxis = list(
  title = "Número de infracciones",
  zeroline = F
),
violingroupgap = 0,
violingroupgap = 0,
violinmode = 'overlay'
)
)

```

p