PROYECTO DE VISUALIZACIÓN DE DATOS

PEC 4: Creación de la visualización

Juan Herranz Martin

8 de enero de 2020

Contents

Título

Situación geográfica y evolución anual de los acuerdos de paz internacionales con referencias sobre corrupción, drogas y/o crimen organizado.

URL de la visualización

- Dashboard sobre los acuerdos con el total de referencias:

https://public.tableau.com/profile/juan.herranz.mart.n#!/vizhome/Libro2_15783128716970/Dashboard1

- Dashboard sobre los acuerdos con referencias de corrupción:

https://public.tableau.com/profile/juan.herranz.mart.n#!/vizhome/Libro2_15783128716970/Dashboard2

- Dashboard sobre los acuerdos con referencias de crimen organizado:

https://public.tableau.com/profile/juan.herranz.mart.n#!/vizhome/Libro2_15783128716970/Dashboard3

- Dashboard sobre los acuerdos con referencias de drogas:

https://public.tableau.com/profile/juan.herranz.mart.n#!/vizhome/Libro2_15783128716970/Dashboard4

- Repositorio Github público con el informe, el dataset y el archivo .twbx de a visualización:

https://github.com/jherranzma/ProyectoVisualizacion

Descripción

Este documento trata de implementar un proyecto de visualización de datos que dé solución a ciertas cuestiones que nos planteamos desde el inicio del proyecto. Algunas de estas cuestiones son:

- ¿De qué datos disponemos?
- ¿De qué forma presentamos los datos?
- ¿Qué información queremos transmitir?

En concreto:

- ¿Qué países están involucrados en acuerdos de paz internacionales?
- ¿En cuántos acuerdos de paz figura cada país?
- ¿Cuántas referencias se hacen sobre corrupción, crimen organizado y/o drogas en cada país en estos acuerdos?
- ¿En qué años se hicieron esas referencias y cuál fue su cantidad?
- ¿Qué países presentan una mayor cantidad de referencias sobre corrupción, crimen organizado y/o drogas?
- ¿En qué años se produjo mayor cantidad de estas referencias?
- ¿Permite comparar las cantidades de diferentes países y años?

Veremos si el proyecto es capaz de responder a estas cuestiones y en qué grado, y qué limitaciones nos encontramos. Se valorará el esfuerzo invertido y los resultados obtenidos,y se describirá la elaboración técnica del proyecto.

Cumplimientos, limitaciones y valoraciones

Considero que con la visualización se ha conseguido cumplir la mayoría de los principios de una buena visualización de datos, como pueden ser: selección adecuada del tipo de gráficos y de los colores, orientada a un público general, intuitiva y estéticamente atractiva, resalta la información más importante, el título contextualiza los gráficos, permite interactividad sin sobrecarga de información, es ordenada, equilibrada y evita distracciones.

Concretamente:

El proyecto cumple con la mayoría de las cuestiones planteadas anteriormente. En efecto, podemos ver en el mapa qué países forman parte de algún acuerdo de paz y cuáles no. Se puede comprobar el recuento de acuerdos en los que figura cada país y el número de referencias (sumando los tipos 1, 2 y 3) sobre corrupción, crimen organizado y/o drogas. En el diagrama de barras (o en las descripciones emergentes) puede comprobarse en qué años se hicieron esas referencias y su cantidad. También, se puede analizar los países que presentan mayor cantidad de referencias y los años con mayor cantidad de éstas, fijándonos en el tamaño de los circulos en el mapa y en el tamaño de las barras del diagrama de barras, respectivamente; algo que también permite comparar países y años.

En esta parte de la práctica he aprendido a crear diferentes tipos de campos calculados, a mostrar diferentes cantidades y recuentos, y a ser capaz de mostrar gráficos como descripciones emergentes dentro de otros gráficos.

Sim embargo, he encontrado ciertas limitaciones técnicas que han dejado en el tintero algunas de mis espectativas, ya sea por mi falta de experiencia con el uso de Tableau (lo más probable) o por limitaciones del propio software (menos probable). Algunas de éstas han sido: no ser capaz de obtener una columa con todos los países de los acuerdos que contienen más de un país. He sido capaz de obtener varias columnas para los acuerdos con varios países pero sólo he podido hacer uso de la primera de ellas. Tampoco he podido obtener un sólo dashboard que fuera capaz de cambiar el tipo de referencia (corrupción, crimen organizado, drogas o totales) en el propio dashboard, para que no fuera necesario obtener cuatro dashboards.

El esfuerzo invertido en la exploración y compresión de los datos, su limpieza, integración y transformación, y en la elaboración de toda la visualización, creo que ha sido el adecuado como para obtener unos resultados que considero que han sido buenos. Quizás se podría haber invertido más esfuerzo para solventar las limitaciones anteriores, aunque considero que este reto no conllevaría una notable mejora de los resultados.

Cabe añadir que a medida que avanzaba en el proyecto me he dado cuenta que no merecía la pena llevar a cabo algunos de los retos planteados en la PEC3 (como la elaboración de diagramas de tartas, histogramas de frecuencias o gráficos lineales), ya sea porque consideraba que éstos sobrecargarían la visualización y dificultarían su comprensión, porque no aportarían información relevante, porque no era el mejor tipo de gráfico, o porque su información ya podía extraerse de la visualización.

Descripción técnica del proyecto

Para realizar las labores de limpieza, integración, transformación y exportación de los datos se hizo uso del lenguaje R en la PEC2. Además, para hacer ciertos análisis exploratorios visuales de los datos se utilizó la librería ggplot2, que ofrece mayores posibilidades de visualización en R. Para la elaboración de los informes, se hizo uso de R Markdown en la PEC2 y en esta PEC4; y para el informe de la PEC3 se utilizó el lenguaje LaTeX en TeXworks.

Para la elaboración de la visualización de datos utilizamos el software de análisis de datos: Tableau. En mi caso, pese a tener licencia de estudiante de la UOC para hacer uso de la versión Tableau Desktop, he optado por hacer uso de Tableau Public, ya que permite publicar la visualización en una URL pública accesible. Desde allí puede descargarse la visualización en formato .twbx.

En cuanto a la descripción del proceso de creación de la visualización y los retos encontrados:

- Lo primero que hago es conectarme desde Tableau al archivo de datos (p2.xlsx) y obtener todos los campos de la fuente de datos. Aquí, se me plantea el reto de dividir en columnas el campo de los países para separar los acuerdos con varios países en varias columnas. Esto lo hago con la opción “División personalizada” seleccionando el slash “/” como separador. Trabajaré con la primera de estas columnas.
- El segundo paso es añadir la primera Hoja de trabajo. Añado la dimensión de país al panel central y Tableau crea un primer mapa de puntos. Añado una de los tres tipos de referencias que tengo en el panel de medidas (referencias sobre corrupción, crimen organizado y drogas) o creo un campo calculado con la suma de esas tres y lo añado como medida. Arrastro la medida a la marca de color como medida

de SUMA y de nuevo la arrastro a la marca de tamaño. De esta forma se crean puntos de diferente color y tamaño sobre el mapa. Personalizo el color según cada caso. Arrastro también la dimensión de acuerdos de paz al panel de marcas y lo establezco como medida de "Recuento (Definido)".

- El tercer paso es añadir la segunda Hoja de trabajo. En esta Hoja agrego a las columnas la dimensión Year y a las filas la SUMA de la medida de referencias. Se crea un diagrama de barras. Añado al panel de las marcas la dimensión país y al color la medida de SUMA de referencias, y selecciono el color según el caso. Se obtiene de esta manera una diagrama de barras apiladas de un color con diferentes intensidades.
- En el cuarto paso se plantea el reto de añadir diagramas de barras como descripción emergente en el mapa. Para el vuelvo a la hoja del mapa y pulso en la marca de Descripción emergente. Aquí, voy a la pestaña insertar y selecciono la hoja de diagrama de barras anterior, le doy un tamaño adeacuado y pulsamos "Aceptar".
- El quinto y último paso será crear el dashboard con las dos hojas anteriores. Para ello pulsamos en Nuevo dashboard y arrastramos al panel las dos hojas: encima el mapa y debajo las barras. Quitamos del dashboard las leyendas que puedan aparecer y le damos un nombre adecuado a cada gráfico (a cada hoja).

Repetimos los pasos segundo, tercero, cuarto y quinto otras tres veces para cada caso, de forma que obtenemos cuatro dashboards, uno por cada tipo de referencias (sobre corrupción, crimen organizado, drogas y totales). Finalmente, se ocultan las hojas y se guardan los dashboards en Tableau Public.

Visualización realizada

En cuanto a las cuestiones sobre la visualización realizada:

- 1. La temática que presenta es, en general, la presentada desde el inicio en la PEC2. La temática trata sobre qué países figuran en acuerdos de paz internacionales relativos a temas de corrupción, drogas y crimen organizado; cuál es la cantidad de referencias a estos temas en cada país involucrado y en qué año tuvieron lugar. A diferencia con el inicio del proyecto, en esta parte han emergido nuevos datos como resultado del campo calculado correspondiente al total de referencias (corrupción + drogas + crimen organizado) y que han pasado a formar parte de la visualización.
- 2. Se ha añadido capacidad de interacción para mostrar datos de forma innovadora. En particular, se ha conseguido mostrar sobre el mapa visualizaciones emergentes de los diagramas de barras propios de cada país.
- 3. Considero que se muestran los datos de forma efectiva ya que se responden a la cuestiones planteadas en la descripción. A modo de ejemplo, las visualizaciones emergentes, con sólo pasar el ratón por encima de un país, ofrecen una rápida información acerca de los años en los que ese país figura en algún acuerdo de paz, y qué cantidad de referencias se hacen sobre corrupción, drogas y/o crimen organizado en cada uno de esos años.
- 4. El diseño final apenas varía con respecto al diseño presentado en el informe del proyecto (PEC3). Los tipos de gráficos seleccionados (mapas del mundo y diagramas de barras), su disposición y estructura, la tipografía y el color (Morado: referencias totales, Marrón: referencias de corrupción, Rojo: referencias de crimen organizado, Verde: referencias de drogas) y sus diferentes intensidades; se consideran los adecuados para este proyecto. Las URL públicas facilitadas, tanto las de Tableau Public como la de Github, hacen que este proyecto sea fácilmente accesible; dando accesibilidad al dataset utilizado, al archivo .twbx de la visualización, al informe del proyecto y a la propia visualización en línea en Tableau Public.

Anexo 1. PEC 2: Familiarízate con los datos

Título:

Referencias sobre corrupción, drogas y crimen organizado en acuerdos de paz según la región o país involucrado; y su evolución anual.

Descripción:

En este documento se presenta información acerca de la relevancia que toman las medidas sobre corrupción, drogas y crimen organizado en los diferentes acuerdos de paz internacionales. Este hecho se presentará en función de las regiones o países involucrados en los acuerdos, y de la evolución anual de los acuerdos.

De esta forma se intentará comprobar cuáles son las regiones o países más afectados y una posible correlación entre esas variables.

Los datos se obtendrán a través de la base de datos contenida en la web: https://www.peaceagreements.org/ , propiedad de: The University Of Edinburgh. Desde esta web se descargará el dataset con los 1789 acuerdos y sus 266 variables, actualizado el día 13-11-2019.

Los datos: presentación, exploración, procedimientos y herramientas.

El proceso de presentación y exploración de los datos se llevará a cabo mediante procedimientos y funciones propias en la herramienta RStudio, haciendo uso del lenguaje de programación R. También se hará uso de la librería ggplot2 para explorar los datos con visualizaciones más avanzadas.

En primer lugar, vamos a cargar el conjunto de datos descargado y comprobar sus dimensiones.

```
pax<-read.csv("pax_data_1789_agreements_13-11-19.csv")
dim(pax)
```

```
## [1] 1789  266
```

```
#tambien podemos ver todos los nombres de las variables con el comando colnames(pax)
```

A continuación, seleccionamos el subconjunto con las variables que vamos a estudiar, hacemos las primeras transformaciones y exportamos ese subconjunto en formato xlsx.

```
#seleccionamos las variables siguientes:
p<-pax[,c("ï..Con","Reg","PPName","Cor","SsrDrugs","SsrCrOcr")]

#añadimos la variable "year" tranformando la variable "Dat" para quedarnos sólo con el año
p$year<-format(as.Date(pax$Dat),"%Y")

#renombramos la variable país con un nombre más "amable"
colnames(p)[1]<-"Con"

#exportamos el subconjunto, que se utilizará para la elaboración de los dashboards.
library(xlsx)
write.xlsx(p,"p.xlsx")
```

Ahora vamos a comprobar si hay alguna columna con valores NA (perdidos o desconocidos).

```
colSums(is.na(p))
```

```
##      Con      Reg   PPName      Cor SsrDrugs SsrCrOcr     year
##        0        0        0        0        0        0        0
```

No se observan valores NA en ninguna de las variables. Veamos ahora la estructura de los datos con la función str():

```r
str(p)
```

```
## 'data.frame':    1789 obs. of  7 variables:
##  $ Con     : Factor w/ 169 levels "(Bougainville)/(United Nations)",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Reg     : Factor w/ 6 levels "Africa (excl MENA)",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ PPName  : Factor w/ 152 levels "","Abkhazia peace process",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Cor     : int  0 1 2 1 2 2 1 2 0 0 ...
##  $ SsrDrugs: int  0 0 2 1 2 2 0 2 1 1 ...
##  $ SsrCrOcr: int  0 0 2 0 0 2 0 2 1 2 ...
##  $ year    : chr  "2016" "2014" "2012" "2011" ...
```

Vemos que:

- La variable "Con" se presenta como una variable categórica (Factor) con 169 niveles, cada uno correspondiente a un país o conjunto de países implicados en los acuerdos.
- La variable "year" se presenta como una variable de caracteres, en la que se especifica el año en que tuvo lugar el acuerdo. Por tanto, convendría factorizar esta variable para transformarla en categórica.
- La variable "Reg" se presenta como categórica, en la que cada una de las 6 categorias (factores) hace referencia a una región del mundo.
- La variable "PPname" también se presenta como categórica donde cada categoría se corresponde al nombre formal del proceso de paz
- Las variables "Cor", "SsrDrugs" y "SsrCrOcr" se presentan como variables numéricas de enteros. Sabemos que cada número de estas variables (del 0 al 3) indica el tipo de mención que se hace en el acuerdo en lo que a medidas contra la corrupción, drogas o crimen organizado (respectivamente) se refiere. Consultando la documentación vemos que el 0 corresponde cuando no se hace mención, el 1 cuando se hace una referencia general, el 2 cuando se referencia un proceso concreto de forma general y el 3 cuando se referencia un proceso concreto especificado en detalle. Convendría, por tanto, transformar estas variables en categóricas ya que cada registro hace referencia a un suceso.

Realizamos las trasformaciones anteriormente descritas, con la función as.factor():

```r
p$year<-as.factor(p$year)

#Creamos 3 nuevas variables categóricas a partir de las numéricas:
p$Cor_fac<-as.factor(p$Cor)
p$Drugs_fac<-as.factor(p$SsrDrugs)
p$CrOr_fac<-as.factor(p$SsrCrOcr)

#Comprobamos la nueva estructura:
str(p)
```

```
## 'data.frame':    1789 obs. of  10 variables:
##  $ Con      : Factor w/ 169 levels "(Bougainville)/(United Nations)",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Reg      : Factor w/ 6 levels "Africa (excl MENA)",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ PPName   : Factor w/ 152 levels "","Abkhazia peace process",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Cor      : int  0 1 2 1 2 2 1 2 0 0 ...
##  $ SsrDrugs : int  0 0 2 1 2 2 0 2 1 1 ...
##  $ SsrCrOcr : int  0 0 2 0 0 2 0 2 1 2 ...
##  $ year     : Factor w/ 30 levels "1990","1991",..: 27 25 23 22 22 21 21 21 20 20 ...
##  $ Cor_fac  : Factor w/ 4 levels "0","1","2","3": 1 2 3 2 3 3 2 3 1 1 ...
##  $ Drugs_fac: Factor w/ 4 levels "0","1","2","3": 1 1 3 2 3 3 1 3 2 2 ...
##  $ CrOr_fac : Factor w/ 4 levels "0","1","2","3": 1 1 3 1 1 3 1 3 2 3 ...
```

Vemos que la varible "year" ahora es una variable categórica con 30 categorías, una por cada año; y que se

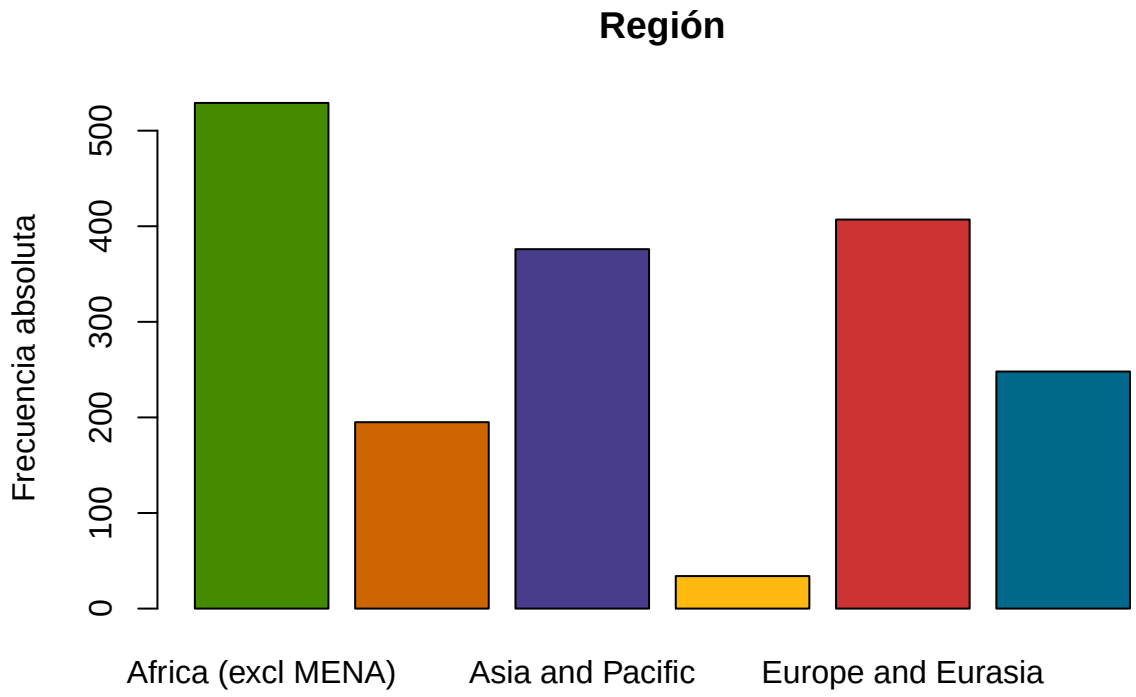
han creado las nuevas variables categóricas con las 4 categorías indicadas anteriormente.

Realizamos a continuación, un análisis exploratorio visual de los datos; con el objetivo de encontrar evidencias o tendencias que ayuden a conocer mejor los datos de los que se dispone.

Hacemos uso de diagramas de barras para analizar las variables categóricas, con el objetivo de comprobar la frecuencia de cada categoría y cuál es la predominante.

En primer lugar, veamos la frecuencia de las regiones en los acuerdos:

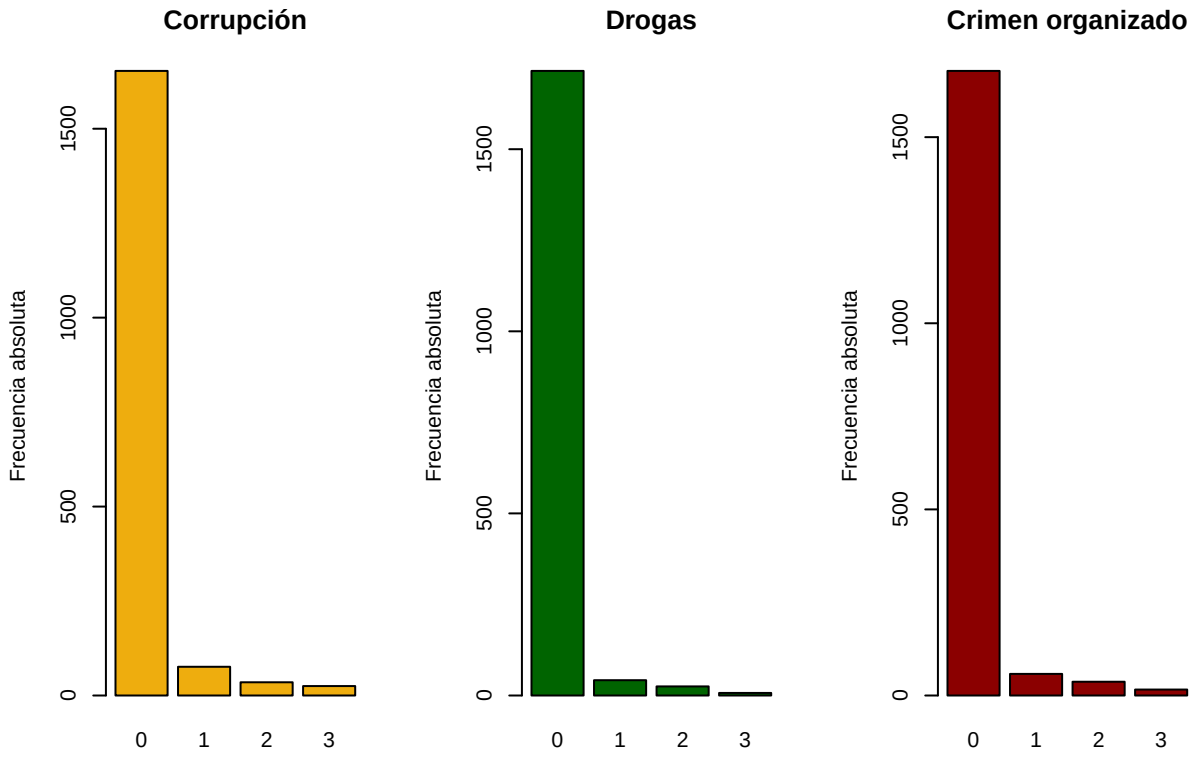
```
plot(x=p$Reg, main="Región", ylab="Frecuencia absoluta", col = c("chartreuse4", "darkorange3",
                                    "darkslateblue","darkgoldenrod1","brown3","deepskyblue4"))
```



Predominan los acuerdos que involucran a África (excluyendo la parte del medio-este y el norte), seguidos de los de Europa y Eurasia.

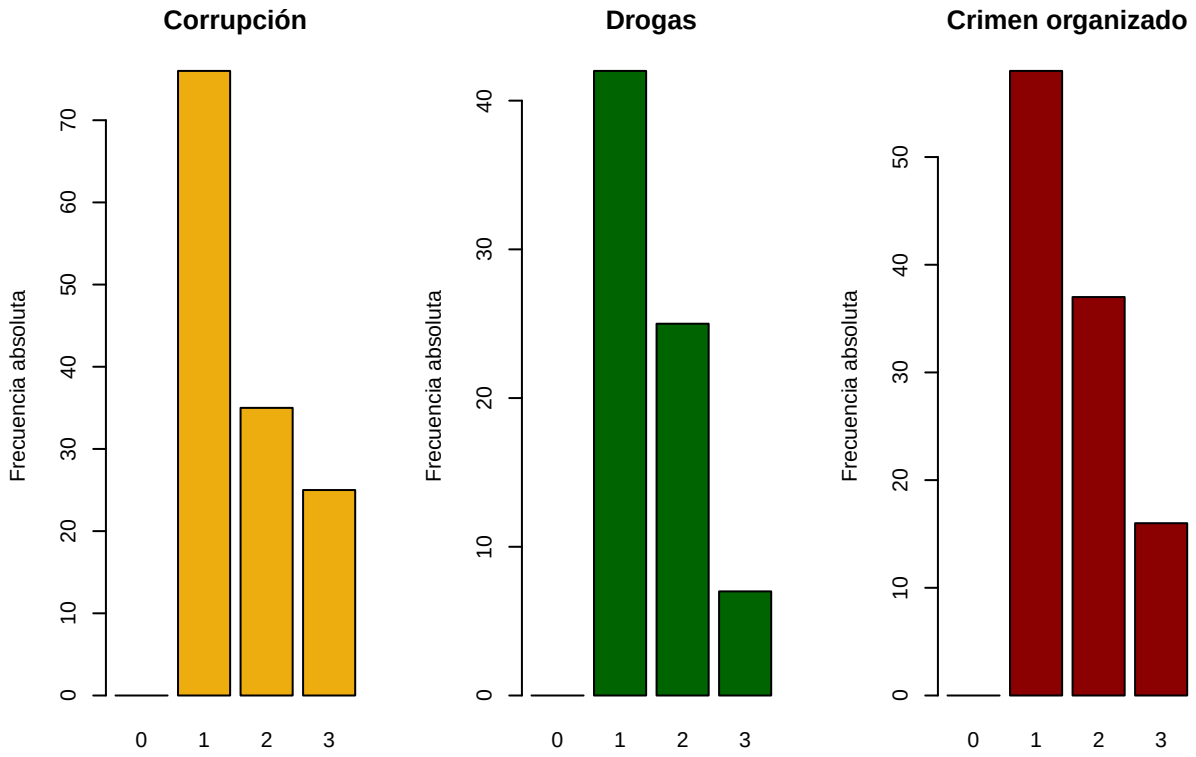
En segundo lugar, veamos la frecuencia de los diferentes tipos de referencia (0, 1, 2 o 3) en los acuerdos en relación a corrupción, drogas y crimen organizado:

```
par(mfrow=c(1,3))
plot(x=p$Cor_fac, main="Corrupción", ylab="Frecuencia absoluta", col="darkgoldenrod2")
plot(x=p$Drugs_fac, main="Drogas", ylab="Frecuencia absoluta", col="darkgreen")
plot(x=p$CrOr_fac, main="Crimen organizado", ylab="Frecuencia absoluta", col="darkred")
```



Vemos que claramente predomina el número de acuerdos en los que no se hace mención a estas tres variables. Además, se observa una tendencia decreciente en el número de casos según crece el nivel de detalle de las medidas de los acuerdos. Veamos con más detalle el número de acuerdos en los que si se hace mención a estas variables, prescindiendo de los casos que no hacen mención (categoría 0).

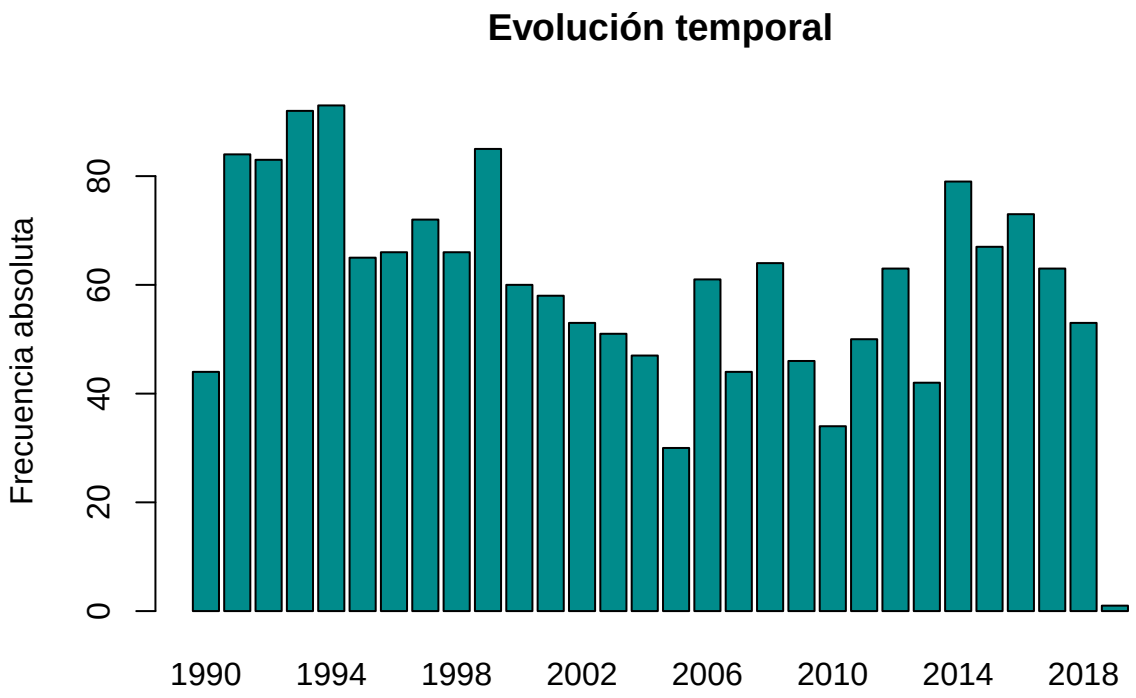
```
par(mfrow=c(1,3))
plot(x=p$Cor_fac[p$Cor_fac!=0], main="Corrupción", ylab="Frecuencia absoluta",
     col="darkgoldenrod2")
plot(x=p$Drugs_fac[p$Drugs_fac!=0], main="Drogas", ylab="Frecuencia absoluta",
     col="darkgreen")
plot(x=p$CrOr_fac[p$CrOr_fac!=0], main="Crimen organizado", ylab="Frecuencia absoluta",
     col="darkred")
```



Predominan los acuerdos con referencias generales (1) a la corrupción, los acuerdos con referencias a procesos concretos (2) de crimen organizado, y los acuerdos con referencias detalladas a procesos (3) de corrupción. Se observa una distribución similar entre las 3 variables, sobre todo entre las variables Drogas y Crimen organizado.

En tercer lugar, veamos la frecuencia de acuerdos por año, para tener una idea la evolución temporal en cantidad de acuerdos que han tenido lugar desde 1990:

```
plot(x=p$year, main="Evolución temporal", ylab="Frecuencia absoluta",col="cyan4")
```

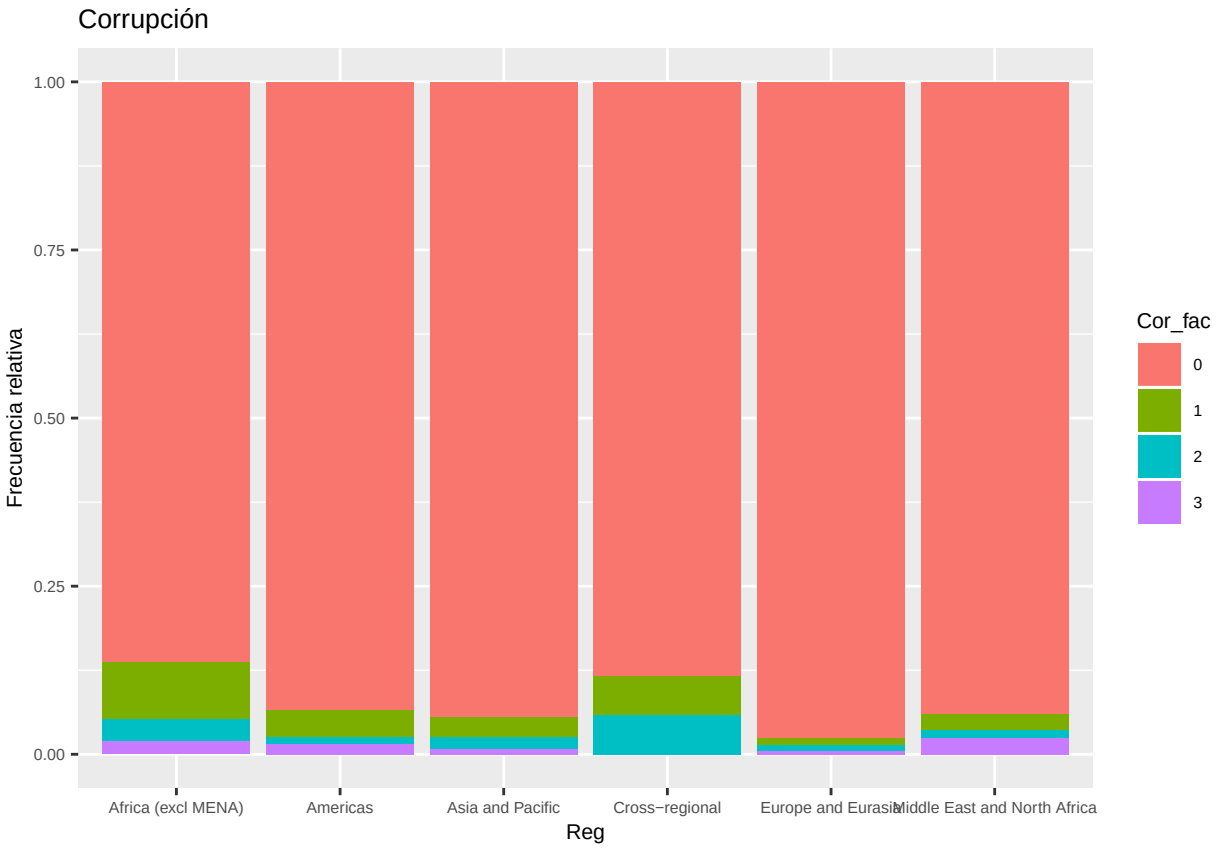


Se observa una evolución con dos crestas y un valle. La primera cresta tuvo lugar hacia el año 1994, el valle hacia el año 2005 y la segunda cresta hacia el año 2014. Se observa una tendencia de disminución en los últimos años.

Hagamos una exploración más detallada con ayuda del paquete ggplot2.

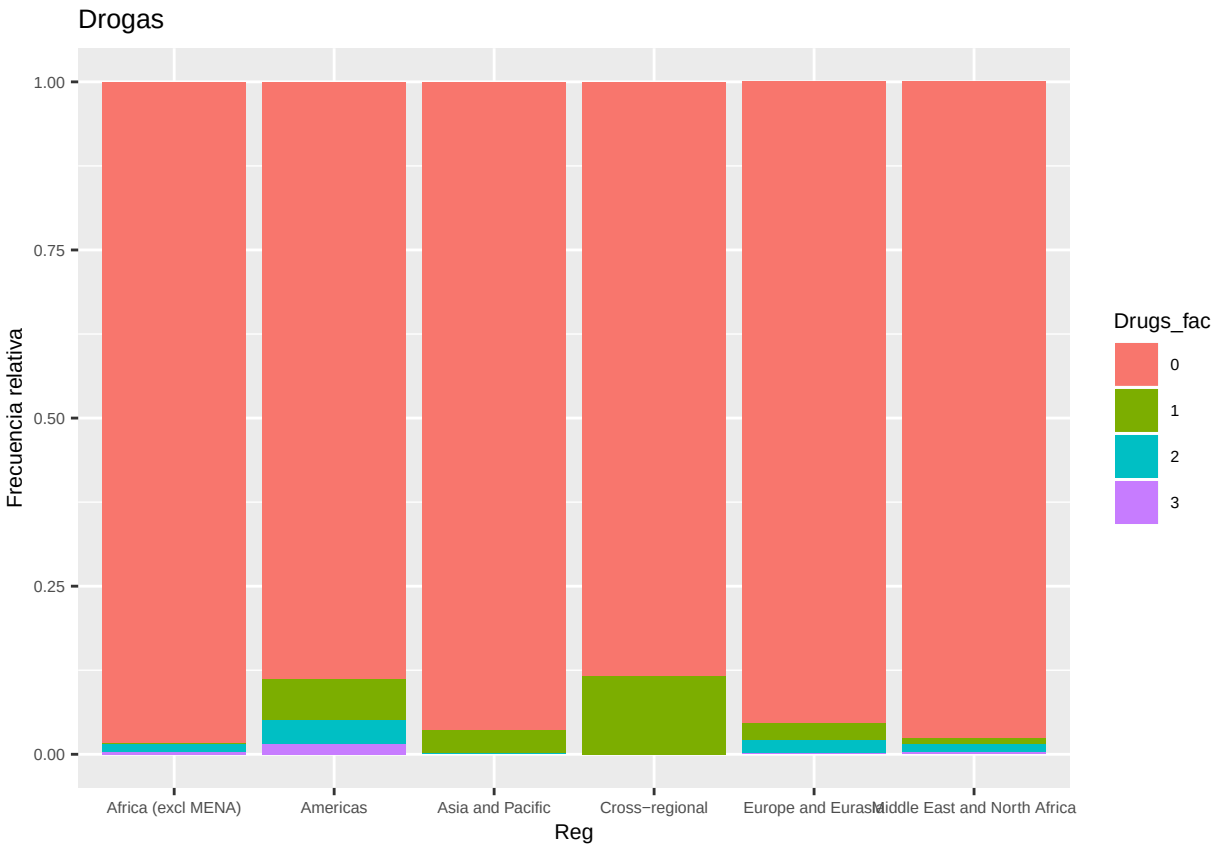
A continuación, exploramos las menciones de corrupción, drogas y crimen organizado por región, en proporción, haciendo uso de la función geom_bar() y la frecuencia relativa:

```
library(ggplot2)
ggplot(p, aes(x = Reg, fill = Cor_fac)) +geom_bar(position="fill")+
  ylab("Frecuencia relativa")+ggtitle("Corrupción")+theme(text = element_text(size=8))
```



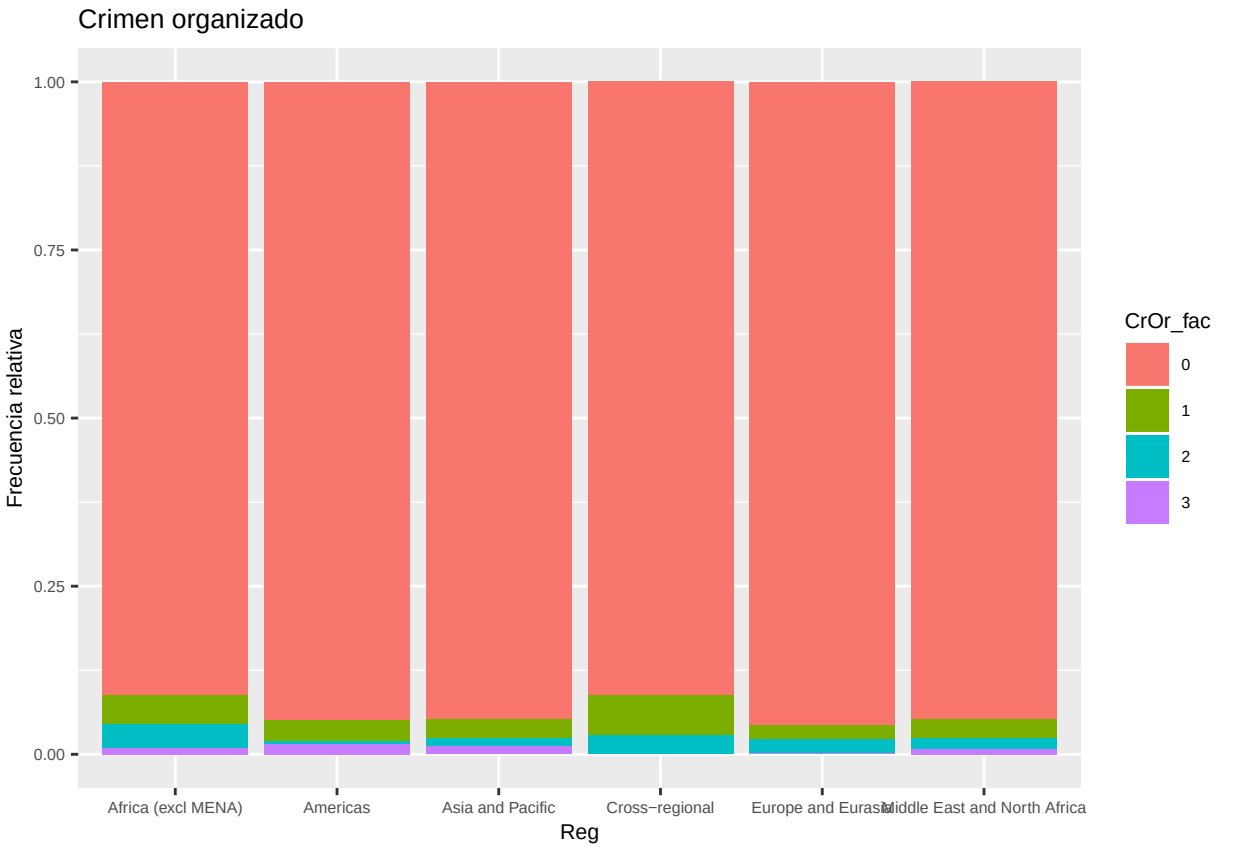
En este caso, la región que más menciones tiene en sus acuerdos (en proporción) es África, seguido del grupo de las regiones cruzadas. La mención tipo 1 es más frecuente en los acuerdos de África, la tipo 2 en regiones cruzadas y la tipo 3 en el medio-este y norte de África.

```
ggplot(p, aes(x = Reg, fill = Drugs_fac)) +geom_bar(position="fill")+
  ylab("Frecuencia relativa")+ggtitle("Drogas")+theme(text = element_text(size=8))
```



En el caso de menciones relacionadas con drogas, las regiones con mayor frecuencia son América y el grupo de regiones cruzadas. La mención tipo 1 es más frecuente en los acuerdos de la regiones cruzadas, la tipo 2 y la tipo 3 en América.

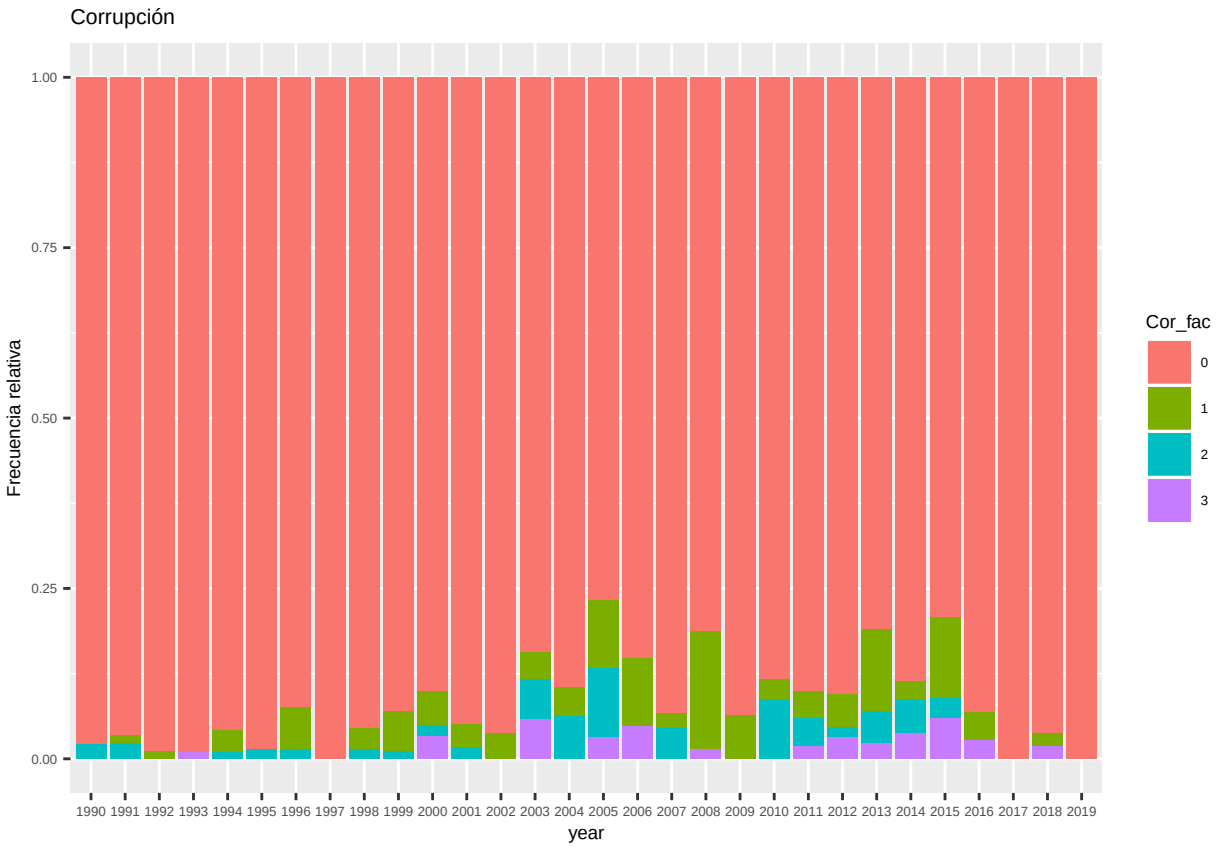
```
ggplot(p, aes(x = Reg, fill = CrOr_fac)) +geom_bar(position="fill")+
  ylab("Frecuencia relativa")+ggtitle("Crimen organizado")+theme(text = element_text(size=8))
```



En cuanto a menciones del crimen organizado, las regiones con mayor frecuencia son África y el grupo de regiones cruzadas. La mención tipo 1 es más frecuente en los acuerdos de la regiones cruzadas, la tipo 2 en África y la tipo 3 en América.

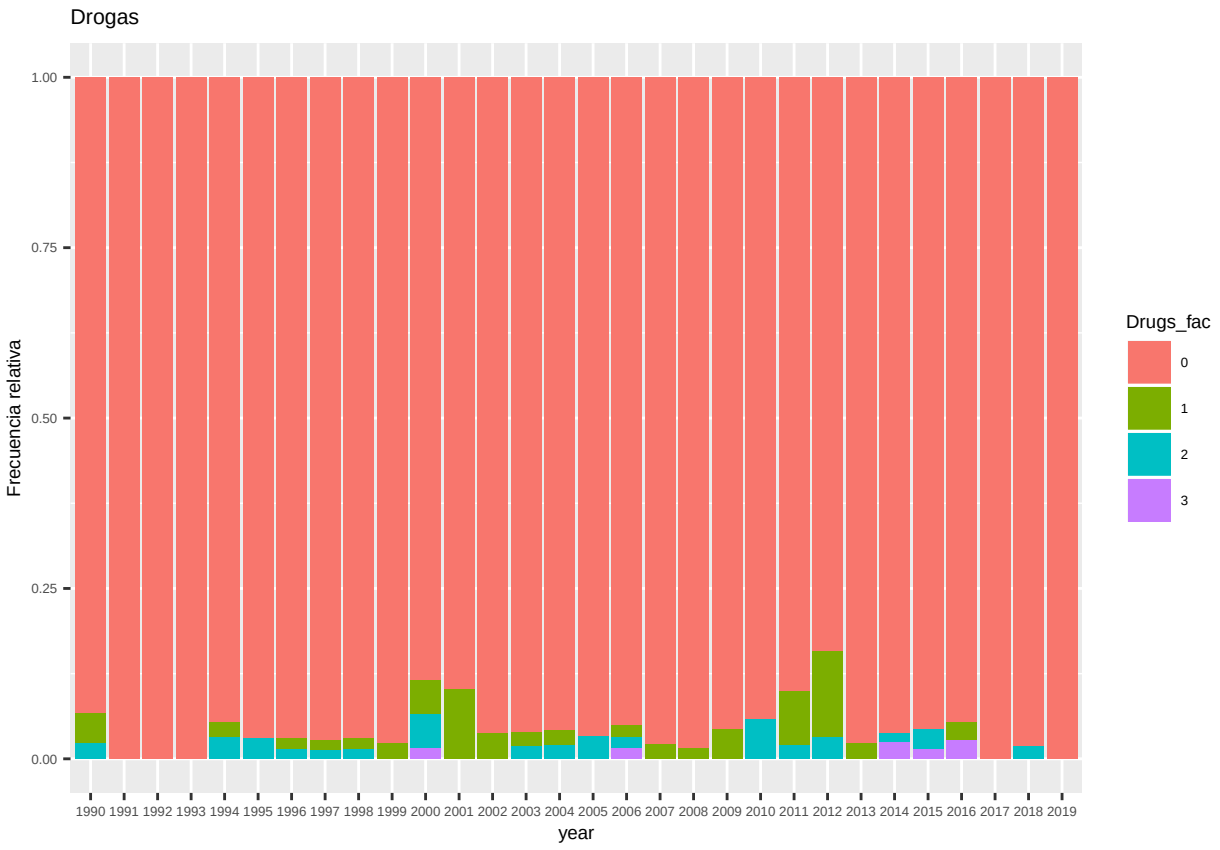
Por último, exploramos las menciones de corrupción, drogas y crimen organizado a lo largo del tiempo, en proporción, con la frecuencia relativa al igual que antes.

```
ggplot(p, aes(x = year, fill = Cor_fac)) +geom_bar(position="fill")+
  ylab("Frecuencia relativa")+ggtitle("Corrupción")+theme(text = element_text(size=6.5))
```



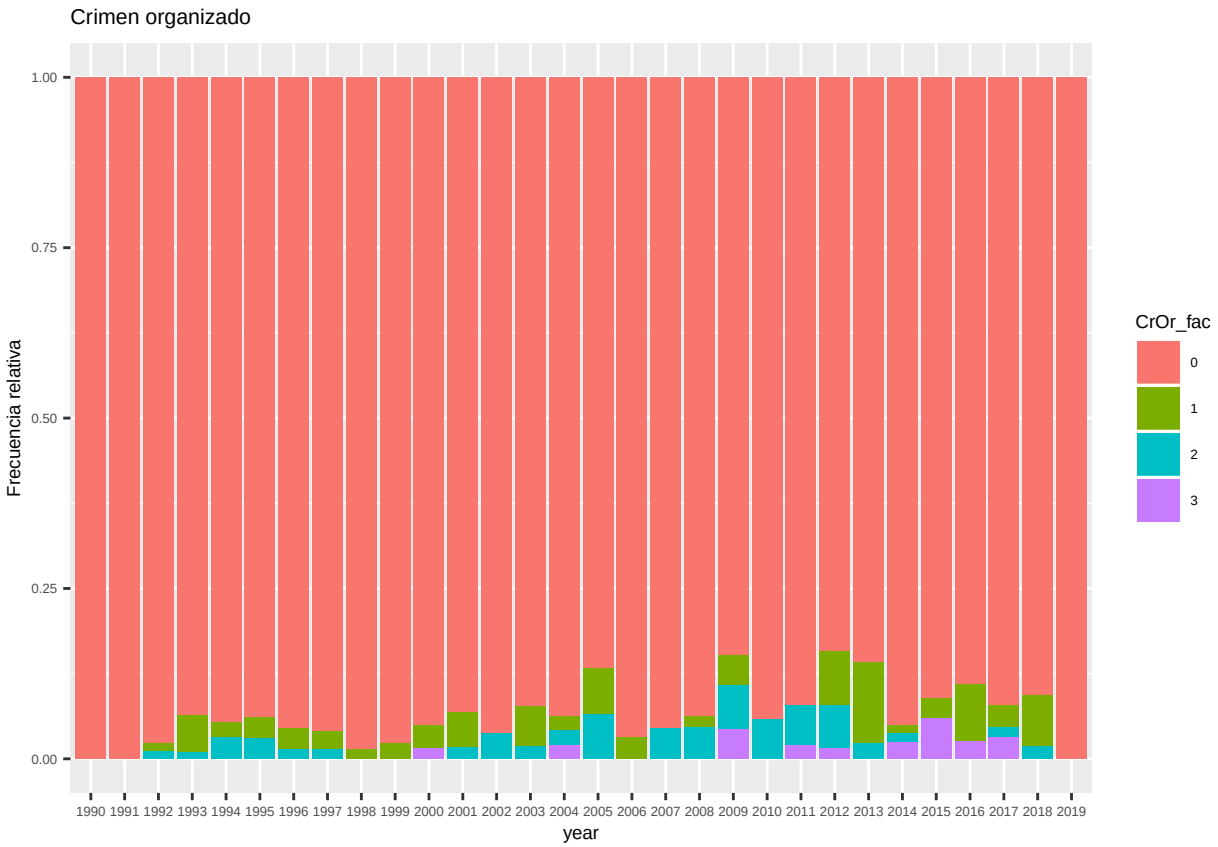
El año con más referencias de corrupción en los acuerdos (en proporción) fue 2005. La mención tipo 1 fue más frecuente en 2008, la tipo 2 en 2005 y 2010, y la tipo 3 en 2003 y 2015.

```
ggplot(p, aes(x = year, fill = Drugs_fac)) +geom_bar(position="fill")+
  ylab("Frecuencia relativa")+ggtitle("Drogas")+theme(text = element_text(size=6.5))
```



En cuanto a referencia sobre drogas, el año con mayor frecuencia fue el 2012. La mención tipo 1 fue más frecuente en 2012 y 2001, la tipo 2 en 2010, y la tipo 3 en 2016.

```
ggplot(p, aes(x = year, fill = CrOr_fac)) +geom_bar(position="fill")+
  ylab("Frecuencia relativa")+ggtitle("Crimen organizado")+theme(text = element_text(size=6.5))
```



Y en cuanto a referencias sobre crimen organizado, el año con mayor frecuencia también fue el 2012. La mención tipo 1 fue más frecuente en 2013, la tipo 2 en 2005 y entre 2009 y 2012, y la tipo 3 en 2015.

Dashboards sobre los datos.

Finalmente, se elaboran unas visualizaciones con el total (la suma) de referencias en los acuerdos sobre corrupción, drogas y crimen organizado, por país y por año. Para ello se hace uso de la herramienta Tableau, donde creamos tres dashboards con dos hojas por dashboard. Cada dashboard hace referencia a una de las tres variables de referencias (corrupción, drogas y crimen organizado) con dos visualizaciones interactivas: un mapa donde se muestran las referencias, en proporción, en cada país; y un diagrama de barras con la evolución anual de las referencias y la proporción de cada país. Las URL de estos tres dashboard se facilitan a continuación:

- Dashboard sobre los acuerdos con referencias de corrupción:

https://public.tableau.com/views/LibroPEC2/Dashboard1?:display_count=y&:origin=viz_share_link

- Dashboard sobre los acuerdos con referencias de drogas:

https://public.tableau.com/views/LibroPEC2/Dashboard2?:display_count=y&:origin=viz_share_link

- Dashboard sobre los acuerdos con referencias de crimen organizado:

https://public.tableau.com/views/LibroPEC2/Dashboard3?:display_count=y&:origin=viz_share_link

Anexo 2. PEC 3: El informe del proyecto

Abstract

Este proyecto consta de tres partes:

La primera parte consistía en familiarizarse con los datos, llevando a cabo exploraciones que permitían descubrir algunas tendencias o evidencias subyacentes en los datos. Para ello se hacía una exploración previa del tipo o estructura de los datos, un breve proceso de limpieza y transformaciones, y una serie de calculos y operaciones llevadas a cabo por diferentes herramientas de visualización que permitían obtener una perspectiva de los datos. Finalmente se elaboraba un dashboard para visualizar los datos con cierta capacidad de exploración e interacción.

En esta segunda parte elaboramos el informe del proyecto con el registro y el rigor académicos adecuados. Haremos una introducción del proyecto, especificando motivos, propósitos y expectativas de la visualización. Iniciaremos un bloque sobre los datos, que constará de secciones como la descripción del conjunto de datos y del proceso de trabajo que se llevará a cabo, el uso de evidencias obtenidas en la primera parte, el tipo de relaciones que se establecerán sobre los datos, la descripción justificada de los gráficos que se utilizarán y otra descripción sobre el gobierno de los datos en la visualización. A continuación, se abrirá el bloque del diseño donde se concretarán y justificarán los aspectos que seleccionemos, tales como: los colores seleccionados, la tipografía de la fuente, la estructura que seguiremos y su división en varios bloques y el tipo de formato. Se añadirá, además, un esquema gráfico con las ideas principales del proyecto de visualización. Finalmente añadiremos las referencias bibliográficas que se consulten para la elaboración del proyecto.

La tercera y última parte consistirá en la propia creación de la visualización de los datos, describiendo y explicando todo lo aprendido, la limitaciones encontradas, descripciones técnicas... todo ello junto con la entrega final de la URL accesible con la visualización final de los datos.

Introducción

El motivo principal que nos lleva a la selección del proyecto (y del conjunto de datos en concreto) es conseguir comprobar con datos una idea acerca de algún tema más o menos conocido y poder obtener infromación más haya de lo que cabía esperar. En particular, se ha selecconado el subconjunto presentado en la primera parte (obtenido de la web https://www.peaceagreements.org/) con el propósito de intentar conseguir información, mediante la visualización de los datos de acuerdos de paz internacionales, sobre cuáles son los países o las regiones afectadas por casos de corrupción, drogas y/o crimen organizado y en qué grado.

Además, añadiendo al estudio la variable temporal, se espera poder conocer también la evolución temporal de cada país con respecto a esas tres variables (corrupción, drogas y crimen organizado).

Con un análisis más exhaustivo en la visualización se podría responder incluso a la cuestión de si existe alguna correlación entre esas variables o si pueden considerarse independientes.

Descripción del dataset

El conjunto de datos seleccionado consta de 1789 observaciones y 7 variables. Presentamos a continuación esas variables y su tipo:

- La variable “Con” se presenta como una variable cualitativa, nominal y categórica (Factor), con 169 niveles, cada uno correspondiente a un país o conjunto de países implicados en los acuerdos de paz.
- La variable “year” se presenta como una variable de caracteres (en formato fecha), ordinal y discreta, en la que se especifica el año en que tuvo lugar el acuerdo. Por tanto, convendría factorizar esta variable para transformarla en categórica, donde cada factor corresponda a un año.
- La variable “Reg” se presenta como cualitativa, nominal y categórica, en la que cada una de las 6 categorias (factores) hace referencia a una región del mundo implicada en el correspondiente acuerdo.

- La variable “PPname” también se presenta como cualitativa, nominal y categórica, donde cada una de las 152 categorias que contiene se corresponde al nombre formal del proceso de paz.
- Las variables “Cor”, “SsrDrugs” y “SsrCrOcr” se presentan como variables numéricas de enteros, ordinales y discretas.Sabemos que cada número de estas variables (del 0 al 3) indica el tipo de mención que se hace en el acuerdo en lo que a medidas contra la corrupción, drogas o crimen organizado (respectivamente) se refiere. Consultando la documentación vemos que el 0 corresponde cuando no se hace mención, el 1 cuando se hace una referencia general, el 2 cuando se referencia un proceso concreto de forma general y el 3 cuando se referencia un proceso concreto especificado en detalle. Convendría, por tanto,transformar estas variables en tres nuevas variables categóricas, ya que cada registro hace referencia a un suceso. Sin embargo, también conviene conservar estas variables como numéricas para poder realizar cáculos con ellas.

Proceso de trabajo

El proceso de trabajo consistirá, en primer lugar, en llevar a cabo un proceso de limpieza para comprobar si existen valores perdidos, desconocidos, outliers..., y realizar las acciones pertinentes. A continuación, se transformarán las variables como se indicaba en la descripción del dataset. Después, se podrá hacer un breve análisis exploratorio visual de los datos para tener una perspectiva previa de los datos, con ayuda de las evidencias y conocimientos obtenidos en la primera parte del proyecto. Una vez hecho esto, ya estaremos en disposición de utilizar los datos en el diseño de la visualización, con unos objetivos de análisis claros.

Tipo de relaciones con los datos

En este proyecto se pretenden establecer diferentes relaciones entre los datos, como pueden ser:

- Comparaciones: entre los diferentes países o regiones en lo referente a menciones sobre corrupción drogas o crimen organizado en los diferentes acuerdos. También podrán hacerse comparaciones sobre esas menciones entre los diferentes años involucrados.
- Relaciones: entre las referencias asociadas a un país con respecto al total o entre las referencias asociadas a un periodo con respecto a la evolución temporal total, por ejemplo. También podrán buscarse correlaciones entre las diferentes variables de referencias.
- Distribuciones: se podrá ver la distribución que sigue el factor temporal con respecto a cada una de las variables de referencias (corrupción, drogas y crimen organizado)
- Adición: para poder llevar a cabo algunas de las relaciones anteriores, puede ser necesario realizar la operación de “suma” de las diferentes categorías de las variables de referencias, es decir, puede ser necesario sumar los valores 0, 1, 2 y 3 para obetener un cómputo global de la variable, que de una idea del “peso” total de las referencias.

Justificación de los gráficos seleccionados

- Se seleccionará un mapa geográfico del mundo dividido por países donde se mostrará el cómputo de las diferentes referencias por cada páis. La cantidad de referencias de cada país se representará sobre el lugar geográfico del país con un punto de diferente tamaño y color, según esa cantidad.
- En caso de no sumar todas las clases de referencias de las variables “Cor”, “SsrDrugs” y “SsrCrOcr” (0+1+2+3), podrían usarse gráficos de tartas para ver la proporción de cada una.
- Para el caso de la visualización temporal (anual) se seleccionará un gráfico de barras, donde cada barra corresponda a un año (ya que son pocas categorías) y cada barra se podrá dividir en la proporciones correspondientes a las distintas referencia de cada país. Un gráfico lineal también puede ser útil para analizar la evolución temporal y las posibles correlaciones lineales entre variables.
- Se podrán utilizar también histogramas para visualizar la frecuencias absolutas de cada región o de las categorias de las variables “Cor”, “SsrDrugs” y “SsrCrOcr”.

Descripción de la gobernanza de datos

- En el mapa con la situación geográfica de los países, el usuario podrá navegar y hacer zoom en las regiones que más le interese. También podrá seleccionar cada país para obtener información más concreta en forma de etiquetas, como el nombre los acuerdos de paz que involucran a ese pais o el número concreto de referencias. Podrá filtrar las variables que desee visualizar ("Cor", "SsrDrugs" y "SsrCrOcr").
- En la evolución anual el usuario podrá seleccionar el año que le interese inspeccionar y la porción de la barra correspondiente a cierto país, donde obtendrá información numérica más detallada en una etiqueta auxiliar.
- Se podrá disponer de gráficos secundarios como resultado de un filtro de variables por parte del usuario, como podría ser la evolución anual de cierto país en materia de drogas o el ranking de países con mayor corrupción en un año concreto.

Diseño

En cuanto al color en el diseño escogeremos el color marrón claro para las referencias de corrupción, el verde para referencias de drogas y el rojo para el caso de crimen organizado. Variará la intensidad de estos colores según la frecuencia de aparición de referencias en los acuerdos. Para la evolución temporal total utilizamos un color "suave" como puede ser el azul claro o morado. La clase 0 de las variables "Cor", "SsrDrugs" y "SsrCrOcr" la representaremos con un color neutro como el gris, ya que hace referencia a una ausencia de referencias. Para las otras tres clases (1, 2 y 3) podríamos utilizar un mismo color (como el rosa o morado) con diferentes intensidades, según el nivel de detalle de las referencias (menos intenso para 1 y más para 3).

No se hará demasiado hincapié en la tipografía, siempre y cuando ésta corresponda a un registro académico. La tipografía de serie de Tableau puede ser suficientemente rigurosa.

Seguiremos una estructura donde se presentarán, para una misma variable ("Cor", "SsrDrugs" o "SsrCrOcr"), sus correspondiente gráficas en un dashborad. Primero aparecerá el mapa, con un mayor tamaño que el resto de gráficas, y más abajo los gráficos de barras y/o lineales de evolución. En paralelo, se dispodrá de los dashboard correspondientes a las otras variables y un dashboard principal con el total de las variables que esté conectado a cada uno de los otros tres; por lo que podemos pensar que el proyecto se puede dividir en unos cuatro bloques.

Este proyecto será sólo en formato digital.

Bibliografía

Se presentan a continuación algunos de los ejemplos y recursos consultados para coger ideas para la elaboración del dashboard:

- https://public.tableau.com/es-es/s/resources
- https://www.tableau.com/solutions/gallery/austin-teacher-turnover-storypoint
- https://www.tableau.com/solutions/gallery/super-sample-superstore
- https://www.tableau.com/solutions/gallery/over-hill
- https://www.tableau.com/sites/default/files/whitepapers/10_best_practices_for_building_effective_dashboardswp_es-es.pdf?reg-delay=dae159a528f8cb7b0b8b8adc5bef118f
- https://towardsdatascience.com/a-step-by-step-guide-to-making-sales-dashboards-34c999cfc28b

Mockup del proyecto

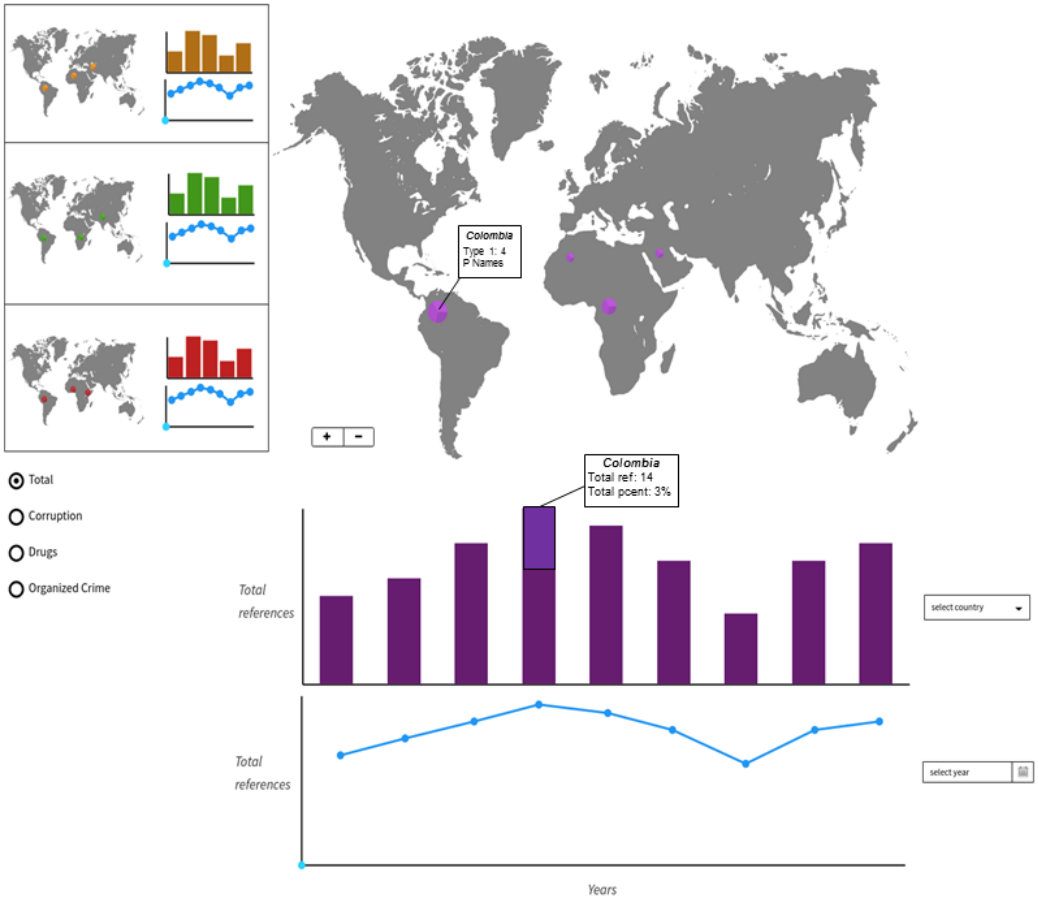


Figure 1: Se presenta a continuación un esquema (mockup) del proyecto de visualización de datos