

## **PEC 4: Proyecto de Visualización de Datos**

Carlos Giner Baixauli

# Índice

## I. Introducción y comunicación.

Abstract .....	3
Introducción .....	3

## II. Datos.

Descripción del dataset .....	4
Análisis exploratorio de los datos .....	5
Preprocesado .....	6
Relaciones .....	6
Gráficas .....	7
Gobernanza de datos .....	7

## III. Diseño.

Colores .....	8
Tipografía .....	8
Estructura y bloques .....	8
Formato .....	8
Mockup .....	9

## IV. Conclusiones.

Visualización .....	9
Evolución del proyecto .....	10
Descripción técnica .....	10
Licencia .....	11
Códigos .....	11
Referencias consultadas .....	12

## I. Introducción y comunicación.

### Abstract

This project introduces an analysis of the dataset PA-X Peace Agreements, a peace agreements database created and managed by the University of Edinburgh. The dataset is available at [www.peaceagreements.org](http://www.peaceagreements.org) and the current version was released on February 19, 2018.

We will present the interaction between some of the most relevant variables in the dataset in order to visualize their relationship and acquire new knowledge. The visualization will be a dashboard, which will include 5 visualizations: Stage by country, Agreement status by country, Agreement conflict level by conflict nature, Agreement status by conflict nature and Agreement status by region.

### Introducción

Hemos seleccionado este proyecto por el interés que presenta extraer conocimiento acerca de los conflictos en el mundo y de los acuerdos de paz, tanto de aquellos en los que ya se han alcanzado como aquellos en los que los tratados están en progreso.

Sería particularmente interesante conocer relaciones de causa-efecto como poder relacionar el nivel de acuerdo o su estado según el tipo de conflicto, así como cruzar esta información con datos geospaciales para conocer cómo se distribuyen los acuerdos y en qué etapa se encuentran en las distintas zonas del mundo.

Nuestro objetivo es presentar la información descrita de manera visual sin renunciar al rigor estadístico de los datos, por lo que la visualización incluirá tanto elementos gráficos como la posibilidad de obtener información numérica de forma interactiva.

## II. Datos.

### Descripción del dataset

El dataset contiene 1789 observaciones y 12 variables. Estas variables son las siguientes:

- *X.U.FEFF.AgreementId* es una variable numérica que identifica el acuerdo de paz.
- *Name* es una variable cualitativa que consiste en el nombre del acuerdo de paz.
- *Region* es una variable cualitativa que identifica la región del acuerdo de paz. Contiene 6 niveles (Europe and Eurasia, Cross-regional, Middle East and North Africa, Africa (excl MENA), Asia and Pacific y Americas).
- *Country* es una variable cualitativa que identifica el país del acuerdo de paz. Contiene 169 niveles.
- *Peace.Process* es una variable numérica que codifica los procesos de paz. Contiene 152 niveles.
- *Peace.Process.Name* es una variable cualitativa que identifica el nombre de los procesos de paz codificados en la variable *Peace.Process*. Contiene 152 niveles.
- *Stage* es una variable cualitativa que identifica la etapa del proceso de paz. Contiene 7 niveles: SubPar (Marco-sustantivo parcial), Imp (Implementación / Renegociación), Pre (Pre-negociación en proceso), SubComp (Marco-sustantivo comprensivo), Cea (Alto el fuego), Ren (Renovación) y Oth (Otros).
- *Signed.Date* es una variable numérica de tipo fecha que identifica la fecha en la que fue firmado el acuerdo.
- *Agreement.Conflict.Level* es una variable cualitativa que indica el nivel del acuerdo. Contiene 4 niveles: Inter (Entre estados), InterIntra (Entre estados y dentro de un estado), Intra (Dentro de un estado) e IntraLocal (Dentro de un estado, a nivel local).
- *Conflict.Nature* es una variable cualitativa que identifica el tipo de conflicto. Contiene 5 niveles (Government, Government/territory, Inter-group, Territory y Other).
- *Agreement.Status* es una variable cualitativa que indica el estado final del acuerdo. Contiene 4 niveles (Agreement with subsequent status, Multiparty signed/agreed, Status nuclear y Unilateral document).
- *Agreement.Text* es una variable cualitativa que contiene el texto del acuerdo.

Para obtener información visual del dataset, las variables más interesantes son las cualitativas estructuradas en niveles. Estas variables agrupan los datos en categorías, de manera que contando las observaciones por grupo podemos conocer su distribución. Además, si cruzamos estas variables podemos obtener nueva información. En concreto utilizaremos las variables *Stage*, *Agreement status*, *Agreement conflict level*, *Conflict nature*, *Country* y *Region*.

## Análisis exploratorio de los datos

De las variables discretas citadas en el apartado anterior, obtenemos los siguientes conteos por nivel:

### Region

Africa (excl MENA)	529
Americas	195
Asia and Pacific	376
Cross-regional	34
Europe and Eurasia	407
Middle East and North Africa	248

### Agreement.Conflict.Level

Inter	72
InterIntra	234
Intra	1293
IntraLocal	190

### Stage

Cea	369
Imp	325
Oth	9
Pre	505
Ren	40
SubComp	106
SubPar	435

### Conflict.Nature

Government	662
Government/territory	813
Inter-group	206
Other	1
Territory	107

### Agreement.Status

Agreement with subsequent status	7
Multiparty signed/agreed	1652
Status unclear	10
Unilateral document	120

De los resultados anteriores y de las visualizaciones obtenemos algunas conclusiones, entre las que destacan las siguientes:

- El estado del acuerdo “Multiparty signed/agreed” es, con diferencia, el más frecuente tanto a nivel global como a nivel de región.
- La naturaleza del conflicto implica al gobierno en la mayoría de los casos.
- Los acuerdos alcanzados se corresponden con conflictos a nivel interno de los estados en la mayoría de los casos.

Con las variables descritas puede ser interesante conocer la distribución geográfica de los acuerdos con comparaciones como la etapa del acuerdo por país, así como el estado del acuerdo por país y por región. También podemos obtener información valiosa de tipo causa-efecto si comparamos la naturaleza del conflicto con el nivel del conflicto el acuerdo y con el estado del acuerdo. Las variables que seleccionamos para el proyecto son las siguientes:

- Stage
- Agreement status
- Agreement conflict level
- Conflict nature
- Region
- Country

Finalmente resaltamos las tendencias que hemos indicado en este apartado, así como la presencia de un outlier, el valor “Other” en Conflict.Nature y en el cruce de esta variable con otras variables cualitativas discretas.

## Preprocesado

Según lo expuesto en los apartados anteriores, las variables numéricas y categóricas están bien definidas, no obstante, tenemos que hacer un preprocesado en dos de ellas:

- En la variable Conflict nature eliminaremos el valor “Other”, que solo tiene una observación.
- En la variable Country eliminaremos los nombres de países que no existen en la actualidad y dejaremos sólo el nombre de un país para que pueda ser representado en un mapa.

## Relaciones

En la visualización queremos mostrar las siguientes relaciones:

- Stage by country: En esta visualización queremos obtener la etapa del proceso de paz más frecuente por país para conocer su distribución en el mundo y poder hacer comparaciones entre países y regiones.
- Agreement status by country: En esta visualización queremos obtener el estado final del acuerdo de paz más frecuente por país para conocer su distribución en el mundo y poder hacer comparaciones entre países y regiones.
- Agreement conflict level by conflict nature: En este caso tratamos de buscar una relación de causa-efecto entre la naturaleza del conflicto y el nivel de acuerdo, que obtendremos al observar las proporciones de ambas variables cruzadas.
- Agreement status by conflict nature: Aquí tratamos de buscar una relación de causa-efecto entre la naturaleza del conflicto y el estado del acuerdo, que obtendremos al observar las proporciones de ambas variables cruzadas.
- Agreement status by region: En esta visualización queremos comparar los estados del acuerdo de paz por regiones para conocer su distribución en el mundo y poder hacer comparaciones entre regiones.

## Gráficas

Las relaciones del apartado anterior se representarán con las siguientes gráficas:

- Stage by country: Esta visualización consistirá en un mapa del mundo en el que el color de cada país indicará la etapa del proceso de paz más frecuente. Este tipo de gráfico permite visualizar fácilmente las variables geoespaciales.
- Agreement status by country: Esta visualización consistirá en un mapa del mundo en el que el color de cada país indicará el estado final del acuerdo de paz más frecuente. Este gráfico permite ubicar la información en el espacio.
- Agreement conflict level by conflict nature: Aquí mostraremos las proporciones de las variables cruzadas con áreas de cuadrados. Puesto que ambas variables tienen pocas categorías, este método permite hacer la comparación de manera rápida y visual.
- Agreement status by conflict nature: En este caso, como también hay pocas categorías, mostraremos las proporciones de las variables cruzadas con áreas de cuadrados para obtener la información de manera más intuitiva.
- Agreement status by region: En esta visualización utilizaremos gráficos de barras verticales agrupados por región, de esta manera se podrán comparar los estados de los acuerdos entre las distintas regiones.

## Gobernanza de datos

Las visualizaciones descritas en los apartados anteriores se integrarán en un dashboard, de manera que el usuario podrá obtener información visual acerca de las relaciones que hemos establecido.

Puesto que el dashboard es interactivo, el usuario podrá, si lo desea, obtener información más detallada en los gráficos seleccionando los elementos de la visualización con el cursor. Al seleccionar un país, una caja de área o una barra se desplegará una ventana con el valor de las variables en el cruce.

La infografía consistirá en 5 visualizaciones integradas en una historia global en la que todas tendrán la consideración de gráficas primarias. Tratamos de mostrar información relevante de distinta clase de la manera más visual posible.

## III. Diseño.

### Colores

Como Podemos leer en “What to Consider When Choosing Colors for Data Visualization”, en los mapas utilizaremos distintas tonalidades con gradiente del mismo color puesto que, al tratarse de superficies y no de puntos aislados, se pueden distinguir fácilmente. Seguiremos este mismo criterio en los diagramas de cajas para visualizar las distintas combinaciones de pares de variables. Finalmente, en el diagrama de barras utilizaremos un color para cada región, pero el mismo color dentro de cada categoría para evitar la saturación de información. En el conjunto de la visualización utilizaremos tonalidades de azul y verde, ya que son colores que se asocian con la confianza y la paz, cualidades muy adecuadas para mostrar información relativa a tratados de paz.

### Tipografía

Siguiendo la referencia “Data Visualization”, utilizaremos un tamaño de fuente ligeramente mayor para el título general que para los títulos de las 5 gráficas. En cuanto a la fuente, Tableau Book es una buena opción puesto que no sobrecarga el dashboard, permite su lectura rápida y no desvía la atención del usuario, que debe dirigirse al contenido de las gráficas.

### Estructura y bloques

El proyecto de visualización se estructurará en 5 bloques que representarán las 5 gráficas mencionadas en los apartados anteriores. En la parte superior se encontrará el título de la visualización. A continuación, en los dos cuadrantes superiores se situarán los mapas, en el cuadrante inferior izquierdo encontraremos las dos gráficas de áreas y en el cuadrante inferior derecho el gráfico de barras.

### Formato

Puesto que queremos ofrecer al usuario la posibilidad de ampliar información de manera selectiva sobre los datos de la visualización de forma interactiva, el formato de la visualización será un dashboard digital.

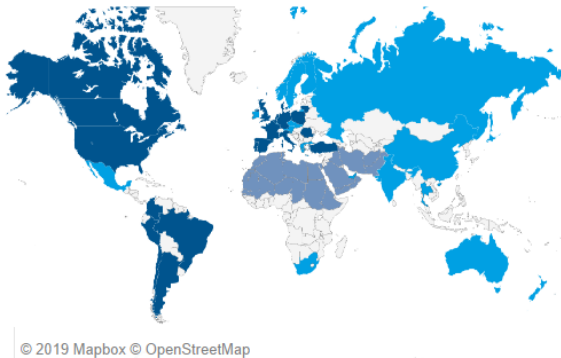


## Mockup

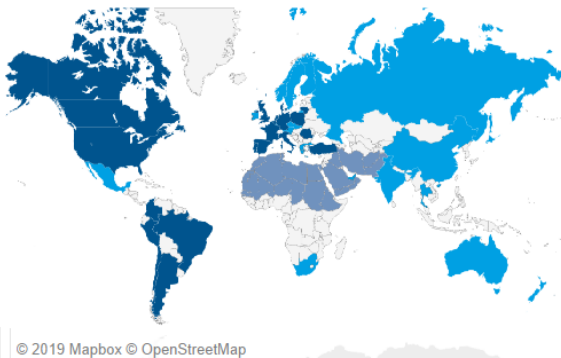
El aspecto final de la visualización, según hemos descrito en los apartados anteriores será el siguiente:

### Conflicts and Peace Agreements around the World

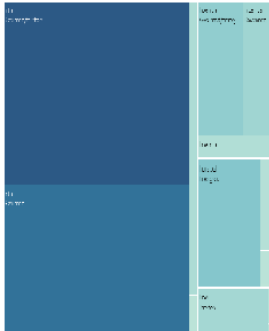
Stage by country



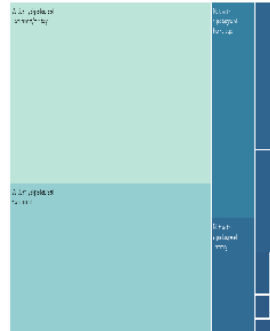
Agreement status by country



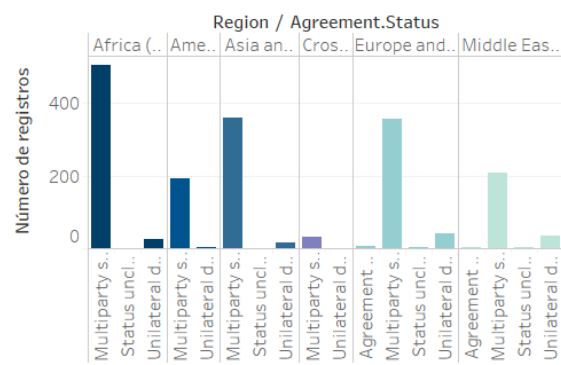
Agreement conflict level by conflict



Agreement status by conflict nature



Agreement status by region



## IV. Conclusiones.

### Visualización

El título seleccionado es “Conflictos y acuerdos de paz en el mundo”. No obstante, el dashboard contiene 5 visualizaciones que muestran la etapa del acuerdo por país, el estado del acuerdo por país, el nivel del acuerdo por tipo de conflicto, el estado del acuerdo por tipo de conflicto y el estado del acuerdo por región.

Palabras clave: peace, conflict, agreement, world

El dashboard está disponible en el siguiente enlace:

[https://public.tableau.com/profile/carlos.giner#!/vizhome/dashboard\\_15776646316290/ConflictsandPeaceAgreementsaroundtheWorld](https://public.tableau.com/profile/carlos.giner#!/vizhome/dashboard_15776646316290/ConflictsandPeaceAgreementsaroundtheWorld)

El código está disponible en el siguiente enlace:

<https://github.com/ginerbaixauli/visualization>

## Evolución del proyecto

Como es habitual en los proyectos de visualización, entre la idea inicial y los resultados finales hay elementos que se han mantenido y elementos que han cambiado. El esquema general del dashboard se ha mantenido, puesto que consideramos que con el dataset de acuerdos de paz, los cruces de variables que más información aportan son las seleccionadas.

No obstante, un estudio pormenorizado de la naturaleza de los datos nos ha obligado a aplicar a estos un preprocesado con el fin de limpiarlos y obtener unas visualizaciones más claras. En el ámbito de la visualización también hemos realizado algunos cambios con respecto a la idea original para hacer la obtención visual de información más accesible al público mediante el uso de gráficas más adecuadas a la naturaleza de los datos, así como la selección de formas geométricas y colores que facilitan la comprensión y la asociación de las ideas que pretendemos transmitir.

Por los motivos descritos, consideramos que el esfuerzo realizado ha dado lugar a un dashboard más intuitivo y completo que el presentado con anterioridad.

## Descripción técnica

Para la realización del apartado anterior se ha procedido a contar los casos en cada categoría de las variables seleccionadas para observar su distribución, su tendencia y la presencia de outliers. También se han cruzado algunas de estas variables para obtener grupos de casos en función de combinaciones de variables.

Para las operaciones con una variable se ha utilizado el lenguaje de programación R, con el que se pueden realizar los cálculos mencionados de forma muy sencilla. No obstante, para la combinación de variables ha sido preferible el uso de Tableau, ya que permite la visualización de patrones y características de una manera mucho más rápida y sencilla.

En concreto, para realizar el análisis en R hemos necesitado las librerías `xlsx` (para manipular ficheros con esta extensión) y `stringr` (para realizar operaciones de sustitución con expresiones regulares). Una vez realizadas las comprobaciones y el preprocesado descritos anteriormente, hemos guardado el dataset preparado. En el siguiente paso hemos cargado este dataset en Tableau y hemos construido la visualización con las especificaciones establecidas y haciendo uso de la interfaz. Como observación hay que remarcar que Tableau presenta una gran cantidad de diseños predefinidos y que su uso es muy intuitivo, por lo que es una excelente aplicación para la creación de dashboards y visualizaciones.

## Licencia

Released Under CC0: Public Domain License

Hemos seleccionado esta licencia para facilitar el uso libre puesto que ofrecemos la visualización al dominio público, liberándolo de todos los derechos de propiedad intelectual, incluyendo todos los derechos conexos. Pueden copiar, modificar, distribuir la obra y hacer comunicación pública.

## Códigos

A continuación, se presentan los códigos de R utilizados en el análisis y tratamiento de los datos:

```
data <- read.csv("C:/Users/Carlos/Desktop/Visualización de
datos/A5 - PEC2/pax_corpus_1789_agreements_17-11-19.csv", header
= T, encoding = "UTF-8")
```

```
# Comprobaciones
```

```
length(unique(data$X.U.FEFF.AgreementId))
```

```
length(unique(data$Name))
```

```
table(data$Region)
```

```
unique(data$Region)
```

```
table(data$Country)
```

```
unique(data$Country)
```

```
length(unique(data$Peace.Process))
```

```
table(data$Peace.Process.Name)
```

```
table(data$Stage)
```

```
unique(data$Stage)
```

```
length(unique(data$Signed.Date))
```

```
table(data$Agreement.Conflict.Level)
```

```
unique(data$Agreement.Conflict.Level)
```

```

table(data$Conflict.Nature)

unique(data$Conflict.Nature)

table(data$Agreement.Status)

unique(data$Agreement.Status)

length(unique(data$Agreement.Text))

# Preprocesado

data <- data[!data$Conflict.Nature=="Other",]

library(stringr)

data$Country <- str_extract(data$Country, "[^/]+")

data$Country <- gsub("[:punct:]", "", data$Country)

# Guardado

library(xlsx)

data <- as.data.frame(data)

write.xlsx(data, "C:/Users/Carlos/Desktop/Visualización de
datos/A9 - PEC4/data.xlsx", col.names = T, row.names = F, showNA
= F)

```

## Referencias consultadas

- What to Consider When Choosing Colors for Data Visualization: <https://www.dataquest.io/blog/what-to-consider-when-choosing-colors-for-data-visualization/>
- Data visualization: <https://material.io/design/communication/data-visualization.html#>
- Data Visualization, Best Practices and Foundations: <https://www.toptal.com/designers/data-visualization/data-visualization-best-practices>
- A Reader on Data Visualization: [https://mschermann.github.io/data\\_viz\\_reader/](https://mschermann.github.io/data_viz_reader/)
- What Makes A Good Data Visualization?. Information is beautiful. <https://informationisbeautiful.net/visualizations/what-makes-a-good-data-visualization/>
- 5 Data Visualization Best Practices. GoodData. <https://www.gooddata.com/blog/5-data-visualization-best-practices-0>