Infraestructura Visible

Entrega 2

Autores:
Miguel Ángel Acosta (201914976)
Andrés Felipe Rincón (201914118)
Ángela Liliana Jiménez (201912941)

Tabla de Contenidos

Identificar Necesidades Analíticas

La documentación de este inciso se encuentra en el documento xlsx EntregaAnalisisRequeridos.xlsx, igualmente a continuación se hacen aclaraciones o se responde alguna pregunta explícita.
Temas analíticos:Mortalidad en Colombia por Departamento, Mortalidad en Colombia por Municipio, Mortalidad en Colombia por Evento y Pronóstico de muertes por evento.
Procesos identificados: Registro de muertes y registro de población.

Priorización de los procesos de negocio

La documentación de este inciso se encuentra en el documento xlsx EntregaAnalisisRequeridos.xlsx, igualmente a continuación se hacen aclaraciones o se responde alguna pregunta explícita.
Fuentes de datos:

- Terridata: Se sacó la informacion municipio por municipio del numero de personas en un rango de edad especifico en un año especifico y de un sexo especifico.
- Data 2018-2020: Información de las muertes suministrada por los profesores.
- Col_deptos: Un archivo csv un poco modificado que contiene latitud y longitud de todos los departamentos.
- Col_muni: Un archivo csv un poco modificado que contiene latitud y longitud de todos los municipios.

Forma de medir la factibilidad: Los análisis de visualización dependen de saber la población con la que se puede contrastar el conteo de muertes, si se cuenta con esta información el análisis es viable, si no, no. Por ejemplo en la entrega pasada no era viable un análisis municipal porque no había registro de la población a nivel municipal. Por otro lado, a nivel de la herramienta de ml, la factibilidad son las métricas que nos entregan, nos basamos en ellas para decir que en qué medida se supone que podemos pronosticar muertes para un grupo determinado.
Plan estratégico de la organización: Realmente lo único que dice en la página es que ellos quieren publicar datos, así que realmente todo lo que hacemos apoya a los procesos de negocio. Si algo pudimos clasificar la información para indicar a qué divulgación de información estamos apoyando.

Proceso de negocio priorizado: Se prioriza el proceso de Mortalidad en Colombia por Departamento, es el más completo, tenemos dos tableros en los cuales basarnos y además es el que más gente podría interesarse en consultar ya que nivel departamental cubre a más personas que nivel municipal.

Modelado de Data Marts

Se plantean dos **procesos** de negocios, por un lado el registro de muertes y por el otro el registro de población.

Las **granularidades** escogidas son:

- Registro de muertes: Conjunto de muertes registradas en un lugar dado, con una demografía específica por una causa dada.
- Registro de población: Registro de población registrada en un lugar dado, con una demografía específica.

Las **dimensiones** que explicarán los hechos antes mencionados:

- Fecha:
  - fecha_key: Llave subrogada de la dimensión.
  - anio: Año en el que se registra la muerte o la población.
- Lugar:
  - lugar_key: Es la llave subrogada de la dimensión.
  - codigo_depto: Es la llave natural del atributo departamento, es útil para comparar con otras bases de datos que usen este número.
  - nombre_depto: Nombre común del departamento, permite analizar a nivel geográfico las muertes.
  - lati_depto: Latitud del departamento, útil para ubicarlo geográficamente.
  - long_depto: Longitud del departamento, útil para ubicarlo geográficamente.
  - codigo_muni: Es la llave natural del atributo municipio, es útil para comparar con otras bases de datos que usen este número.
  - nombre_muni: Nombre común del municipio, permite analizar a nivel geográfico las muertes.
  - lati_muni: Latitud del municipio, útil para ubicarlo geográficamente.
  - long_muni: Longitud del municipio, útil para ubicarlo geográficamente.
- Perfil:
  - perfil_key: Llave subrogada de la dimensión.
  - Edad: Es el grupo de edad en el que se encuentra un conjunto de muertes o una población, permite analizar a nivel de estructura de población las muertes.
  - Sexo: Es el sexo relacionado a un grupo de muertos o a una población, permite análizar a nivel de sexo las muertes.
- Evento:
  - Evento key: Llave subrogada de la dimensión.
  - Codigo_evento: Es la llave natural del atributo evento, es útil para comparar con otras bases de datos que usen este número.
  - Evento: Es el nombre común del evento causante de muerte, permite analizar a nivel de causas las muertes.

La **tabla de hechos** contienen

- Registro de muertes: Llaves foráneas a fecha, lugar, evento y perfil además de un contador de casos de muerte, el cual es una medida aditiva
- Registro de población: Llaves foráneas a fecha, lugar y perfil además de un contador de población, el cual es una medida aditiva

El **modelo dimensional propuesto** que se diseñó, evolucionó al siguiente:

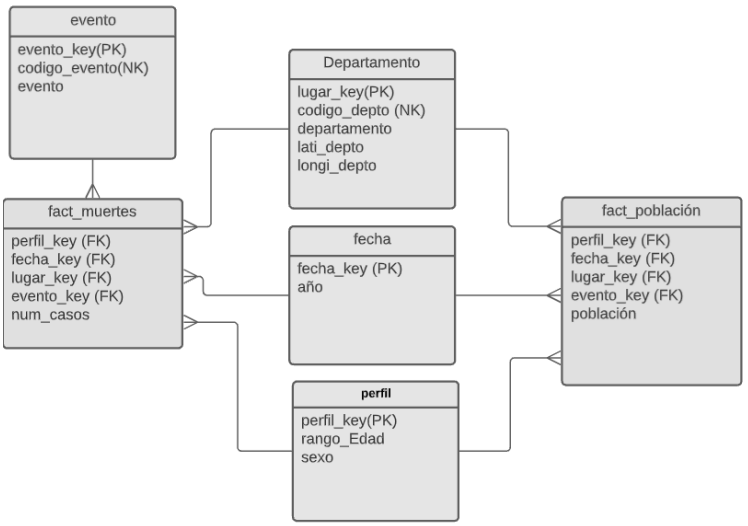


Figura 1: Modelo dimensional

Perfilamiento de Datos

Descripción y perfilamiento:

En la entrega anterior, se nos entregó un consolidado de muertes entre 2010 y 2017, en esta entrega, se nos entregó uno entre 2018 y 2020. Lo más importante a señalar del perfilamiento es lo siguiente:

- Conteo de casos por departamento 2010-2017 vs 2018-2020

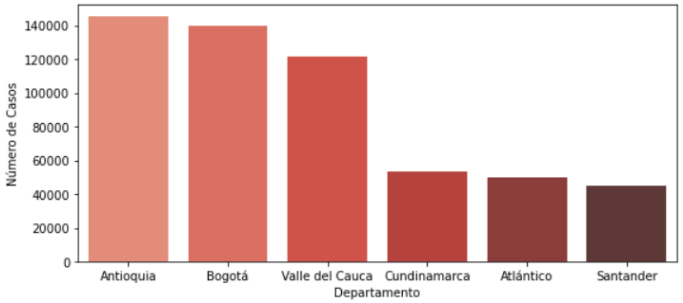


Figura 2: Conteo de casos por departamento 2010-2017

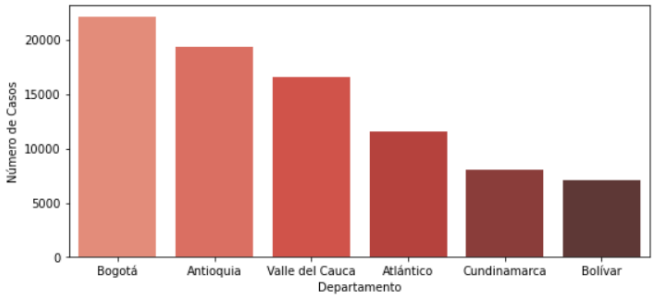


Figura 3: Conteo de casos por departamento 2018-2020

- Conteo de casos por evento 2010-2017 vs 2018-2020

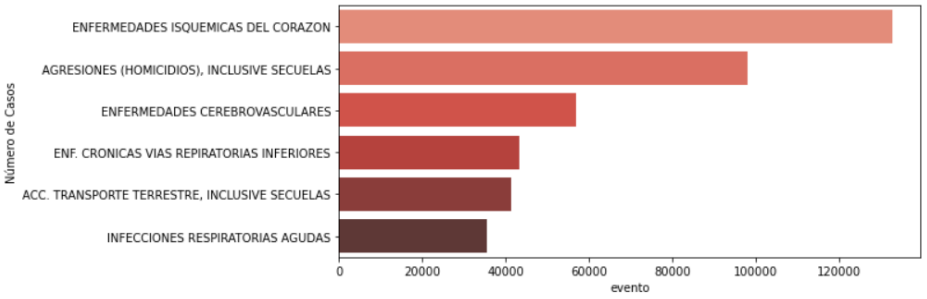


Figura 4: Conteo de casos por evento 2010-2017

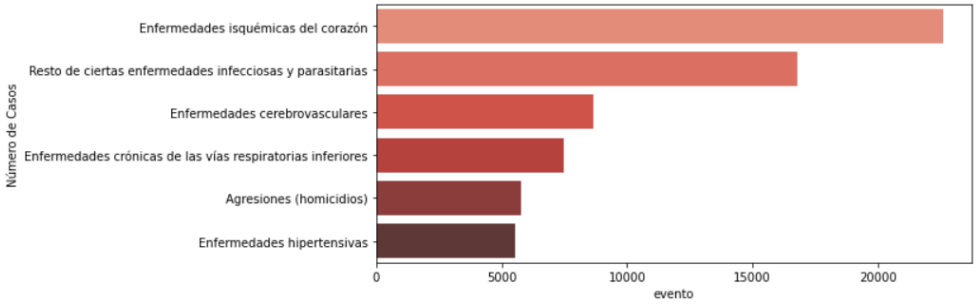


Figura 5: Conteo de casos por evento 2018-2020

- Por la forma en la que se juntaron los datos (es decir los joins entre fuentes) al final no se conservó información de 2018 ni de 2019

Ahora sobra nuestras otras fuentes:

1. Fuente de datos geográficos: colombia_depto.csv
   Esta fuente simplemente nos permitirá unir los departamentos con sus respectivas referencias geográficas, a continuación podemos ver los puntos geográficos.

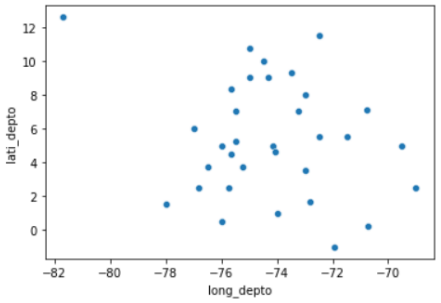
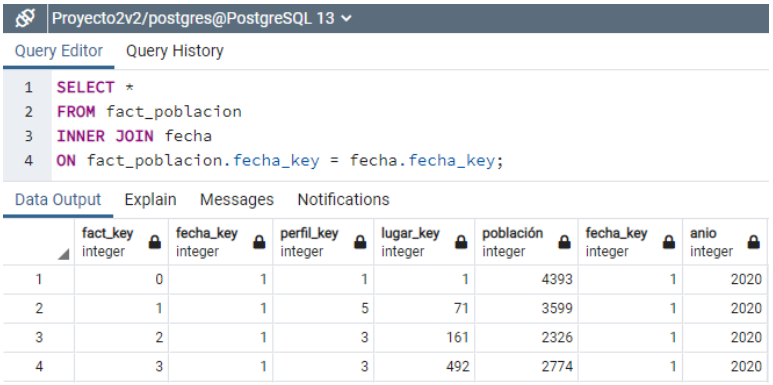


Figura 6: Ubicaciones geográficas

   Podemos ver que los puntos son coherentes al representar Colombia (el punto a la izquierda es San Andrés)
2. Fuente de datos geográficos: colombia_muni.csv
   Esta fuente simplemente nos permitirá unir los municipios con sus respectivas referencias geográficas, no hay mucho más que decir que lo que se dijo del anterior.

Integrar Fuentes de Datos Suministradas: Este punto no fue posible dado que como se ha mencionado anteriormente, para poder entender los casos de muerte es necesario saber la población a la que están relacionados, en el proceso de la entrega pasada usamos información a nivel departamental porque no existe información a nivel municipal, en esta entrega si existe esa información pero no es compatible con la entrega anterior, así que conservamos la entrega anterior como un modelo de visualización 2010-2017 y para esta entrega es un modelo de visualización 2020, por lo explicado de que solo hay información del 2020. compatible con terridata.



	fact_key integer	fecha_key integer	perfil_key integer	lugar_key integer	población integer	fecha_key integer	anio integer
1	0	1	1	1	4393	1	2020
2	1	1	5	71	3599	1	2020
3	2	1	3	161	2326	1	2020
4	3	1	3	492	2774	1	2020

Figura 7: Ejemplo datos cargados en postgre

Podemos ver como hay datos en esta base de datos con información de 2020

El **proceso de ETL**, en sí, el proceso se conservó, con algunos cambios para adaptarse al nuevo modelo. Tenemos tres etapas, consolidación de datos, transformación a modelo dimensional y spoon.
La primera fue la de consolidación de información, en esta etapa recolectamos información de terridata con una herramienta de web scraping para juntarla con la información de data 2018-2020, esto se hizo con carga_de_datos.ipynb.

	codigo_depto	nombre_depto	codigo_muni	nombre_muni	edad	anio	sexo	población
0	41	Huila	41885	Yaguará	0-4 años	2020	Hombres	310
1	41	Huila	41885	Yaguará	0-4 años	2020	Mujeres	323
2	41	Huila	41885	Yaguará	15-44 años	2020	Hombres	1744

Figura 8: Ejemplo datos terridata

	anio	codigo_depto	nombre_depto	nombre_muni	codigo_muni	codigo_evento	evento	edad	sexo	num_casos	población
0	2020	05	Antioquia	Abejorral	05002	102	Tuberculosis	15-44 años	Hombres	1	4393
1	2020	05	Antioquia	Abejorral	05002	501	Accidentes de transporte terrestre	15-44 años	Hombres	1	4393
2	2020	05	Antioquia	Abejorral	05002	511	Lesiones autoinfligidas intencionalmente (suic...	15-44 años	Hombres	1	4393

Figura 9: Ejemplo datos consolidados

Entonces, la idea era juntar la información de los 1103 municipios de colombia de archivos terridata (como los de la figura 8) con data 2018-2020 lo que dio como resultado datos como los de la figura 9.
El paso siguiente fue pasar del archivo consolidado a dimensiones y hechos, como se puede ver a continuación un ejemplo de dimensión y de tabla de hechos

	evento_key	codigo_evento	evento
0	1	102	Tuberculosis
1	2	501	Accidentes de transporte terrestre
2	3	511	Lesiones autoinfligidas intencionalmente (suic...

Figura 10: Muestra Dimensión evento

	fact_key	num_casos	fecha_key	perfil_key	lugar_key	evento_key
0	0	1	1	1	1	1
1	1	1	1	5	71	1

Figura 11: Muestra Fact Table casos_muerte

Luego con esos CSV's se empieza un proceso de carga a la base de datos en spoon, tenemos un job principal que se encarga de cargar los datos a una base de datos postgresql, a continuación el job principal:

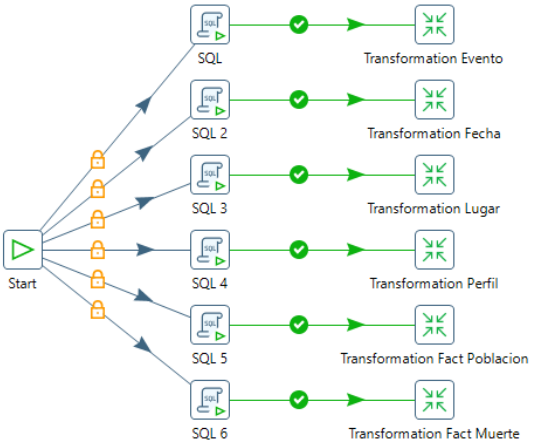


Figura 12: Job principal

Implementación tableros de control

Se implementaron dos tableros de control pensando en los requerimientos analíticos de visualización por departamento y visualización por municipio.

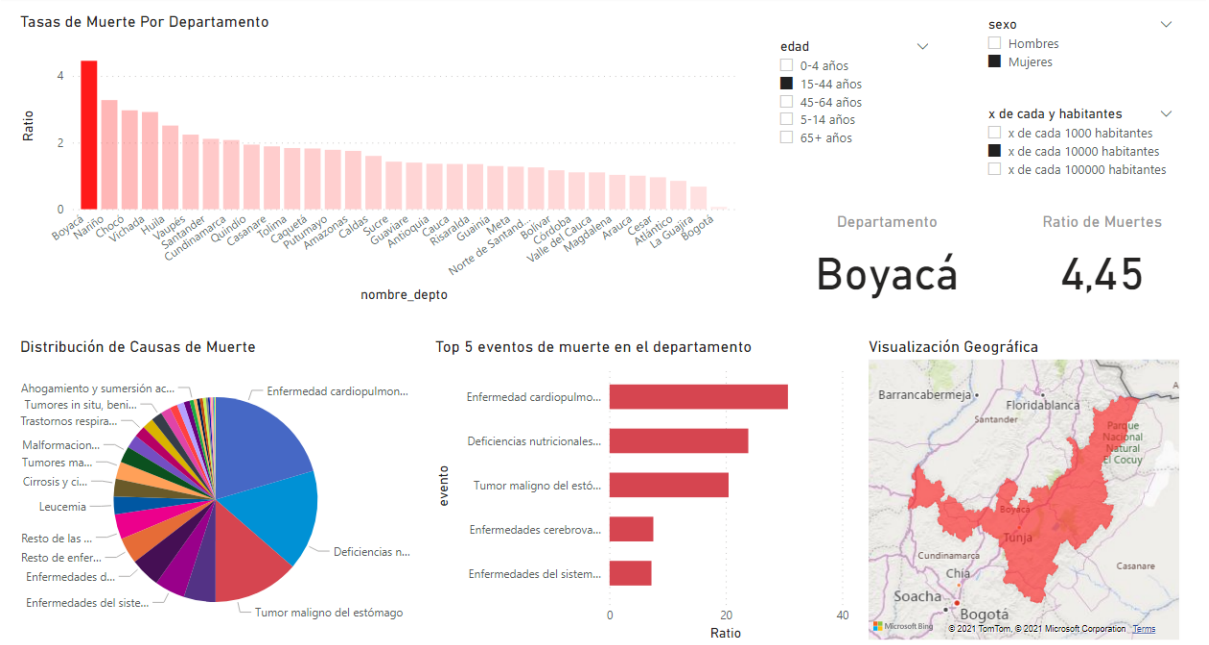


Figura 13: Tablero visualización por departamento

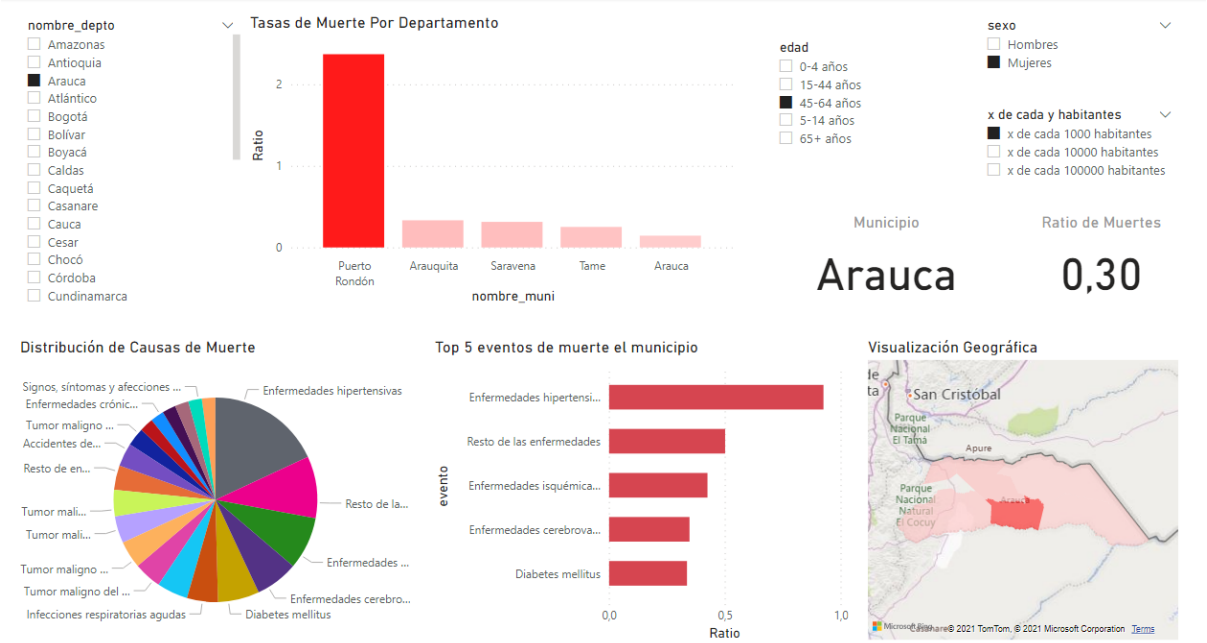


Figura 14: Tablero visualización por municipio

Tarea de Aprendizaje de Máquina

Diseño de una tarea de aprendizaje de máquina:

Tras un detallado análisis al negocio pudimos concluir que serían de su interés 2 cosas:

1. Encontrar grupos por los cuales se pueda clasificar
2. Proyectar las causas de muerte que pueden generarse en diferentes grupos poblacionales.

Es por esto que optamos por hacer clustering que permitiera analizar diferentes formas de categorización y Regresión logística para la proyección.

Tanto los modelos como el perfilamiento y limpieza se encuentran en el archivo Perfilamiento y Limpieza.ipynb.

Para el perfilamiento se pasó a través de diferentes etapas: en la primera se eliminaron las columnas que dificultarían el análisis, tales como año, codigo_depto, nombre_muni codigo_muni, codigo_evento. Las cuales o bien tenían un solo valor (año) o tenían muchas categorías (nombre_muni, codigo_muni) o eran reiterativas (codiigo_evento,codigo_depto)

Adicionalmente se creó una columna “c/10000” que nos permite obtener los casos que ocurrieron cada 10000 habitantes en un lugar en específico, el cual se calculó de la siguiente forma (num_casos/población) 100000. El paréntesis nos dá el % de habitantes del grupo poblacional que fue afectado y al multiplicarlo por 10 mil obtenemos cuantos de dicha población tendrían la causa de muerte asociada.

Por último se volvieron columnas por categoría los siguientes atributos:

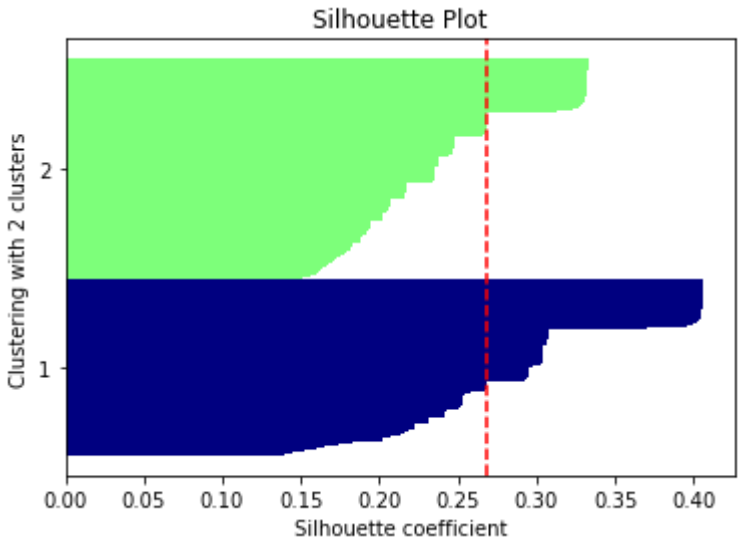
Evento, el cual se agrupó por tipo de causa, siendo los siguientes posibles tipos: CAUSA_NATURAL, MUERTE_ACCIDENTAL, SUICIDIO, HOMICIDIO, INDETERMINADA

Sexo, creandó una para sexo_Hombre y otra para Sexo_Mujeres

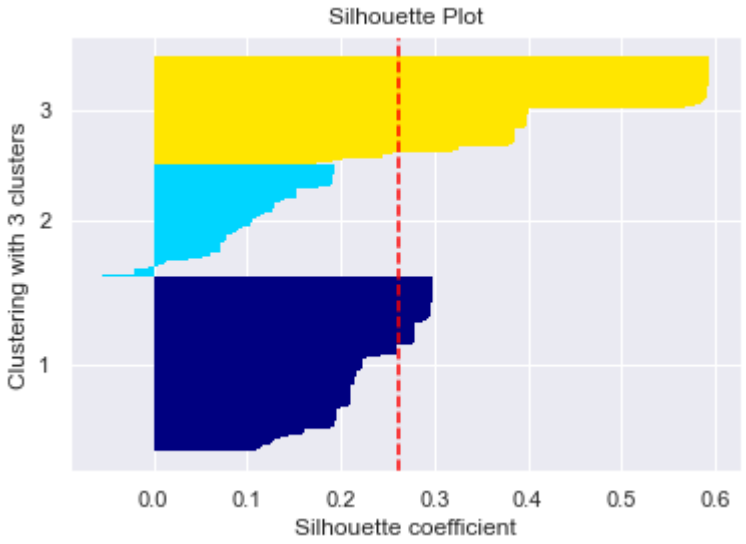
Departamento, agrupando en región para poder volverlos en columna, esto porque al tener muchas categorías en una variable se dificulta su estudio, se crearon las siguientes regiones: Andina, Caribe, Amazonía, Pacífico y Orinoquía.
La tabla final es la siguiente:

	evento	edad	sexo_Hombres	sexo_Mujeres	c/10000	region
0	CAUSA_NATURAL	15-44 años	1	0	2.27635	Andina
1	MUERTE_ACCIDENTAL	15-44 años	1	0	2.27635	Andina
2	SUICIDIO	15-44 años	1	0	2.27635	Andina
3	HOMICIDIO	15-44 años	1	0	2.27635	Andina
4	CAUSA_NATURAL	65+ años	1	0	7.36377	Andina

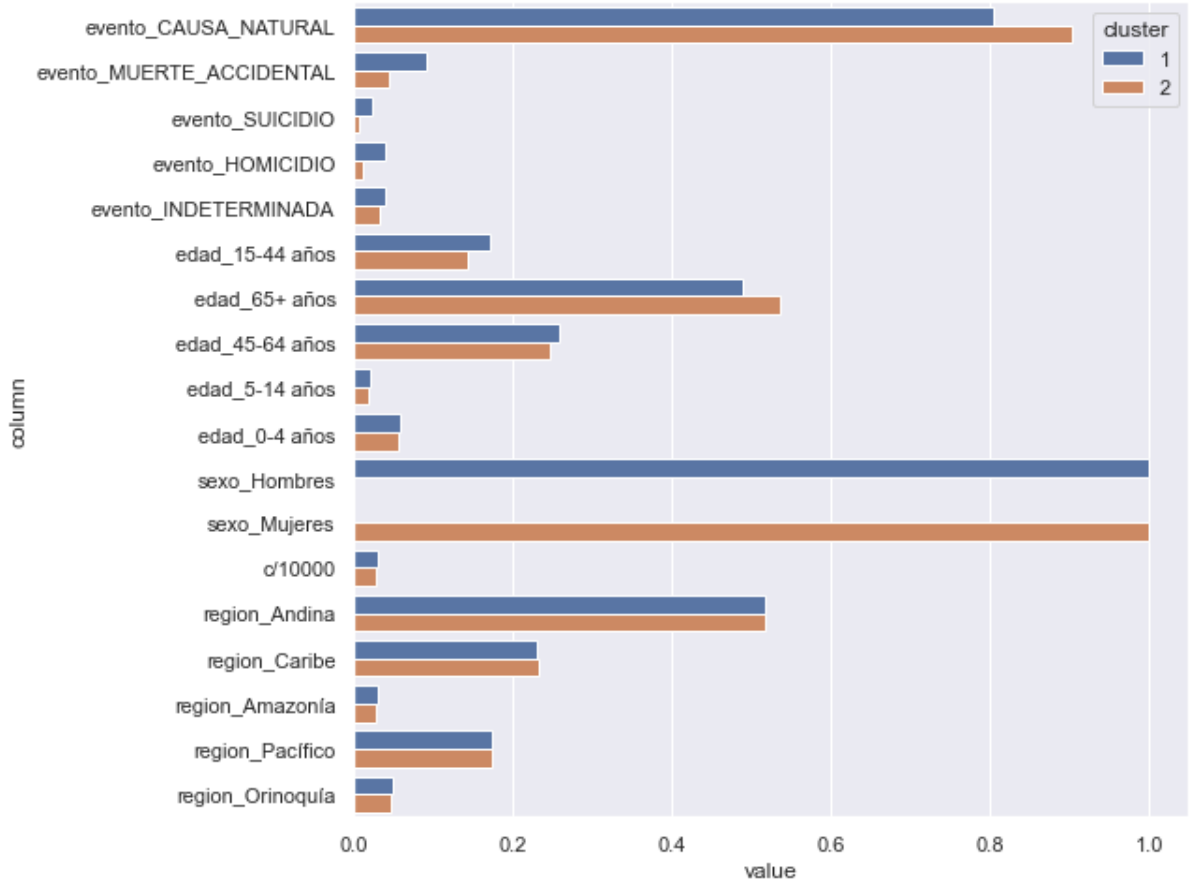
Para el clustering utilizamos K-means ya que según nuestro estudio es el que mejor permite proyectar grupos en esta situación. Es por esto que para poder usarlo tuvimos que agregar 2 pasos al perfilamiento: El primero es volver los tipos de eventos cada uno como una columna, lo que agregaría 5 columnas más (CAUSA_NATURAL, MUERTE_ACCIDENTAL, SUICIDIO, HOMICIDIO, INDETERMINADA). El segundo fue normalizar los valores (de 0 a 1), para lo que usamos MinMaxScaler().
En la primera versión, con 2 clusters, obtuvimos lo siguiente:



Por lo que se realizó una segunda versión con 3 clusters:



Dado que este ya da negativos, se consideran 2 clusters como la mejor. Sin embargo al generar el modelo y visualizar las variables obtenidas, se obtuvo lo siguiente:



Lo que significa que este modelo pudo agrupar por 2 clusters muy parecidos, en donde se encuentran divididos los hombres de las mujeres, pero el resto de columnas son muy parecidas, lo que significa que la principal agrupación es por sexo del grupo poblacional.

Dado que en un principio no se tenía claro cómo hacer un análisis de los datos del proyecto porque en su mayoría eran de tipo categórico, se tomó la decisión de seleccionar como

variable objetivo el evento de causa de muerte, puesto que era una variable que se podía relacionar con las demás, permitiendo segmentar los datos dentro de distintos grupos demográficos. Como se quería estudiar la relación entre distintas variables, se podría pensar en la implementación de un modelo de Regresión Lineal. Sin embargo, este modelo contempla la variable dependiente como un dato numérico, lo cual no resultaba conveniente segun el analisis que se hizo previamente y por la razón de que las variable independientes también tendrían que ser de tipo numérico, lo cual ocasionaria distintos problemas dada la cantidad de variables categóricas que se tenían que manejar. Por esta razón se optó por una regresión logística, el cual es un modelo estadístico que se utiliza para modelar una variable categórica. Así, teniendo la variable categórica como la variable dependiente, se quería conocer cómo otras variables numéricas incidían en su comportamiento. Para el caso de estudio se quiso determinar cómo las columnas de sexo, edad, numero de muertes y departamento guardaban relación con la causa de muerte registrada.

```
score = logisticRegr.score(x_test, y_test)
print(score)
```

```
0.84967815221507
```

Esto significa que la regresión Logística se proyectó correctamente con un 85% de precisión, lo que es un buen porcentaje. Sin embargo esto puede estar sesgado por el gran número de causas naturales.

Video

https://www.youtube.com/watch?v=lkLBEEX5NO8

Trabajo en Grupo

Miguel Angel Acosta (87.5 pts): Tablero 1, Tablero 2, Carga de datos, Proceso ETL

Andres Felipe Rincón (6.25 pts): Herramienta de aprendizaje de máquina

Angela Liliana Jiménez (6.25 pts): Herramienta de aprendizaje de máquina