Universidad del Valle de Guatemala
Facultad de ingeniería

Proyecto 1
Análisis Exploratorio y Clustering

Maria Isabel Solano 20504
Christopher García 20541
José Pablo Monzón 20309
Priscilla Gonzalez 20689

Guatemala 13 de abril del 2023

Situación problemática

En Guatemala, las defunciones -fallecimientos o muertes- son un factor de mucha importancia para medir la mortalidad de una población. Durante los últimos 10 años, ha existido un incremento en las defunciones alrededor del país, causadas por diversas razones que abarcan temas de salud pública y atención médica, violencia, enfermedades hereditarias o contagiosas, calidad de vida baja que conduce a padecer enfermedades que no reciben ningún tipo de atención y con el tiempo empeoran, entre otras; por lo que se busca identificar cuál es el grupo de población más vulnerable a fallecer, con el fin de poder desarrollar estrategias de prevención y tratamientos enfocados en salud y políticas que apoyen al grupo en mayor riesgo para reducir las tasas de mortalidad del mismo.

Problema científico

¿Cuáles son los factores sociodemográficos, ambientales y de salud que influyen en la mortalidad de grupos poblaciones en Guatemala?

Esto con el fin de comprender de una mejor manera las causas de altas tasas de mortalidad en algunos grupos poblacionales, y así de esa manera diseñar estrategias para prevenirlas o reducirlas a través de programas de apoyo, enfoque de recursos hacía dicho grupo y otras actividades de la misma índole.

Objetivos

- General
    - Identificar y comprender los factores que influyen en la tasa de mortalidad en diferentes grupos de la población, con el fin de desarrollar estrategias óptimas para reducir o prevenir las tasas de mortalidad en los grupos más vulnerables de Guatemala.

- Específicos
    - Analizar los datos demográficos y de salud de diferentes grupos de población en Guatemala para identificar las características que los hacen más vulnerables a las altas tasas de mortalidad.
    - Evaluar la efectividad de las estrategias y políticas de salud implementadas para realizar los ajustes necesarios en base a los resultados obtenidos.
    - Desarrollar un análisis exploratorio para determinar las relaciones entre los diferentes factores (como lo es la edad, el departamento, el sexo, el lugar de ocurrencia de las defunciones, el servicio de salud utilizado y el tipo de hospital que pudo haber visitado) y las tasas de mortalidad en los diferentes grupos de la población.

Análisis exploratorio y descripción de los datos

Para realizar el análisis exploratorio de los datos se utilizaron bases de datos obtenidas a partir de la página oficial del Instituto Nacional de Estadística Guatemala. Se consultaron distintos documentos de excel y se combinó información de ellos para obtener una base de datos robusta que abarcara información desde el 2009, hasta el 2021, la cual es la información más actualizada hasta el momento.

A continuación se explicarán los datos y los hallazgos encontrados por conjunto de datos trabajado.

Defunciones en función de la edad:

Se realizó un análisis inicial utilizando pandas profiler para obtener datos iniciales del dataset:

Overview | Alerts 4 | Reproduction

Dataset statistics

Number of variables	6
Number of observations	4399
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	206.3 KiB
Average record size in memory	48.0 B

Variable types

Numeric	4
Categorical	2

Seguido se utilizaron las gráficas que quickDA para analizar las variables individualmente.

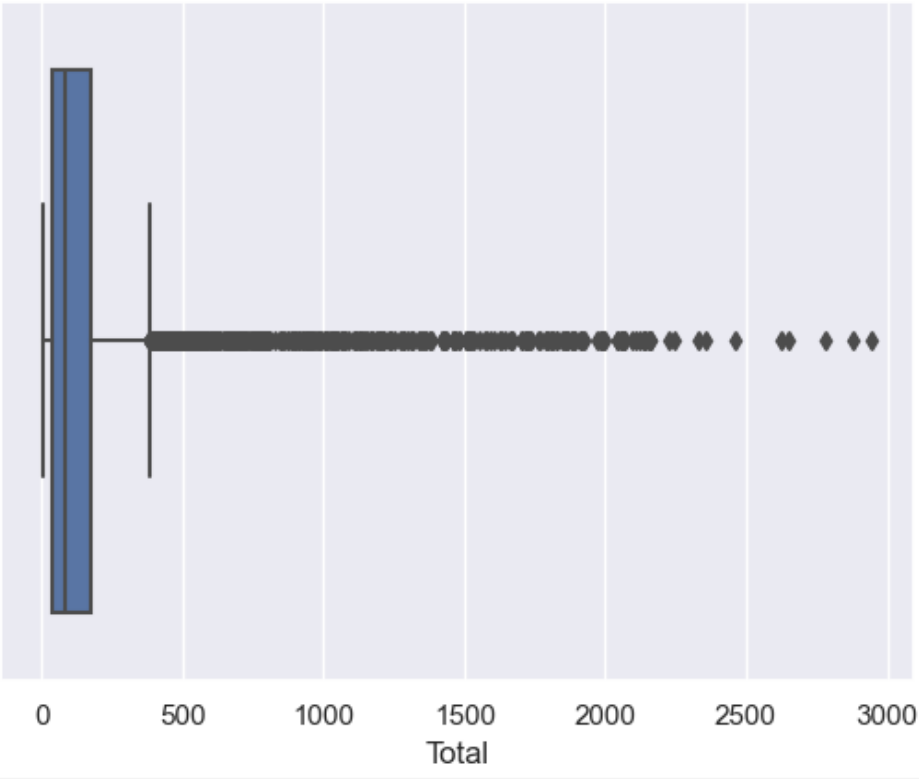
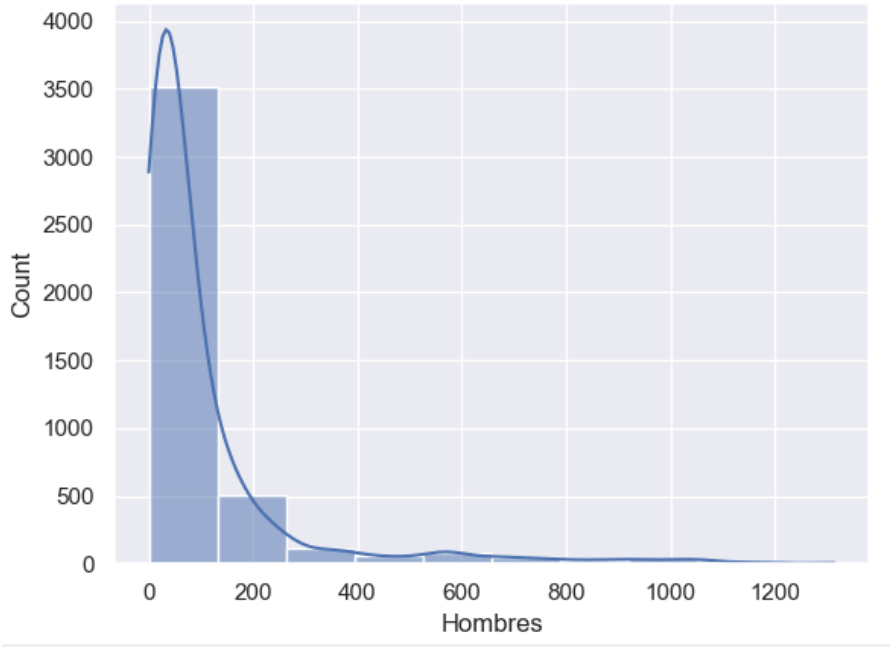
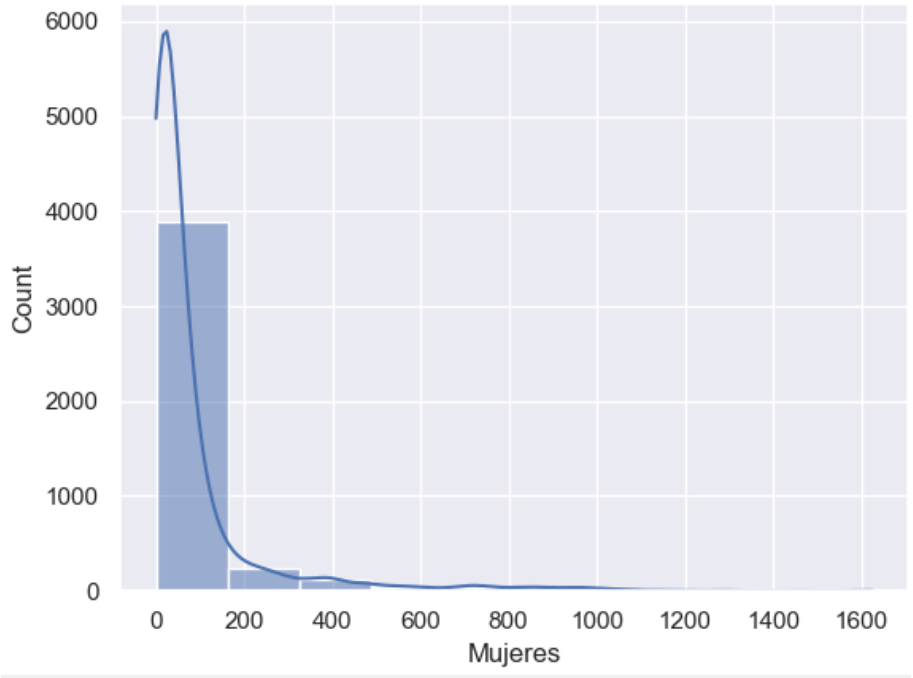


Diagrama de caja y bigotes de total de defunciones

Podemos observar un gran número de outliers y que lo más común es que se muera un aproximado de 50 personas por causa.

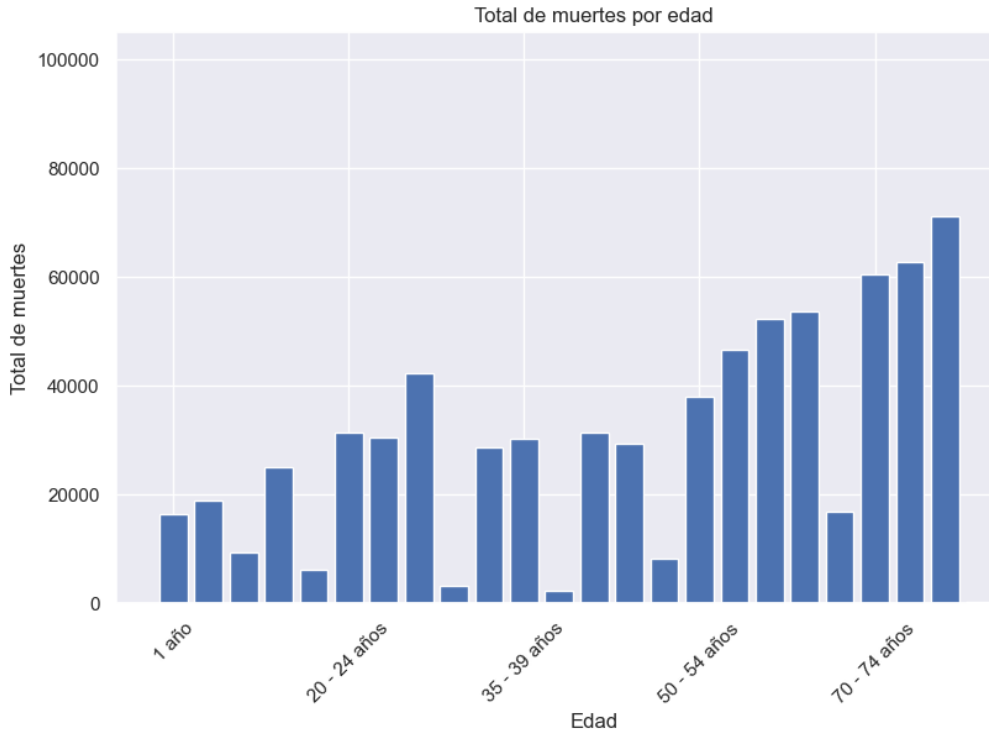


Histograma de total de defunciones masculinas



Histograma de total de defunciones femeninas

Luego utilizando matplotlib se creó una visualización de las muertes en función de la edad de la persona:



Muertes en función de la edad de la persona a lo largo del tiempo.

Defunciones en función del departamento, género y causas de muerte:

Este conjunto de datos era uno de los más simples pero proporcionó un vistazo general de cómo se encontraban los datos relacionados a las defunciones a nivel departamental y si existía alguna relación con el género y la causa de las mismas.

Lo primero que se realizó con los datos (correspondientes a los años 2009-2021) fue una limpieza simple que buscaba eliminar caracteres especiales, sustituir valores vacíos por valores numéricos que representaran lo mismo, como por ejemplo sustituir un vacío por un valor 0 en la cantidad de personas; Convertir variables al tipo que les correspondía como tipo numérico u objeto para así diferenciar variables categóricas y numéricas; También se eliminar filas que no aportaban mayor información como aquellas causas de muerte que se refiere a "otros" en general.

Para lograr esto fue necesario realizar lo siguiente:

- Con ayuda de la función unicode se transformaron todos los caracteres especiales en caracteres normal (i.e. letras con tilde: á se transformaron a letras sin tilde: a)
- Se eliminaron las comas de las variables numéricas para que no se confundieran con decimales y se realizó la transformación a tipo Integer.
- Se sustituyeron los guiones y espacios vacíos que significaban que no había datos por 0's para manejar únicamente números
- Se redujo el data frame utilizado a un data frame sin filas correspondientes a las causas de muertes denominadas: "Otras causas" o "Síntomas no clasificados" puesto que no ofrecían mucho información sobre lo que se desea investigar que tiene relación con causas que se puedan detallar y permitan concluir acerca de ello.

	Año	Total	Hombres	Mujeres
count	4863.000000	4863.000000	4863.000000	4863.000000
mean	2014.998972	199.616081	113.998972	89.013161
std	3.742344	483.774747	265.371592	225.451335
min	2009.000000	11.000000	0.000000	0.000000
25%	2012.000000	43.000000	24.000000	15.000000
50%	2015.000000	77.000000	45.000000	30.000000
75%	2018.000000	157.500000	98.000000	71.000000
max	2021.000000	8285.000000	4395.000000	3877.000000

Descripción de las variables numéricas
(Las que más resaltan para la investigación sería la información de Hombres y Mujeres)

Dataset statistics

Number of variables	6
Number of observations	4863
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	228.1 KiB
Average record size in memory	48.0 B

Variable types

Numeric	4
Categorical	2

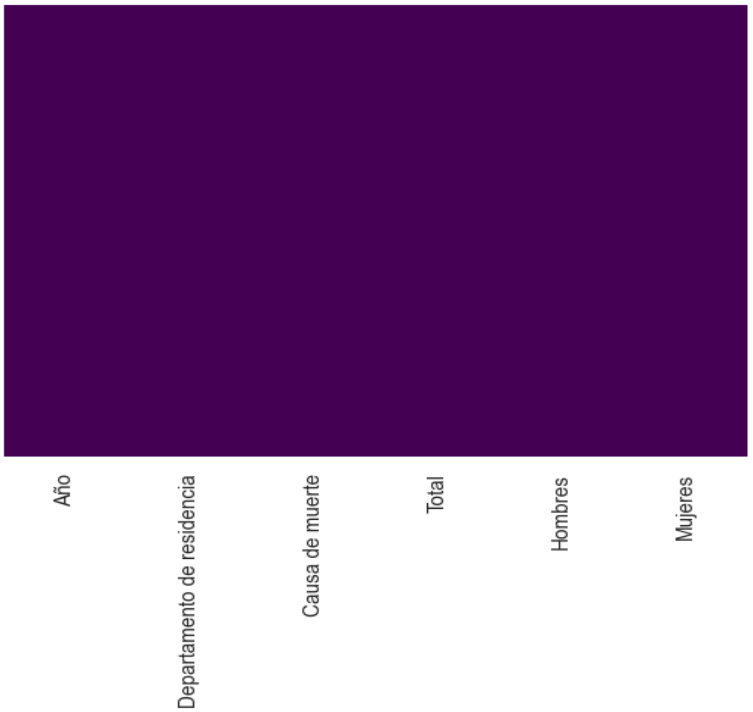
Utilizando un Profiler se obtuvo información más detallada sobre la forma que tenían los datos

Las variables numéricas fueron Año, Total de defunciones, Cantidad de Hombres y Mujeres que fallecieron. Las variables categóricas fueron el Departamento de residencia y la Causa de muerte

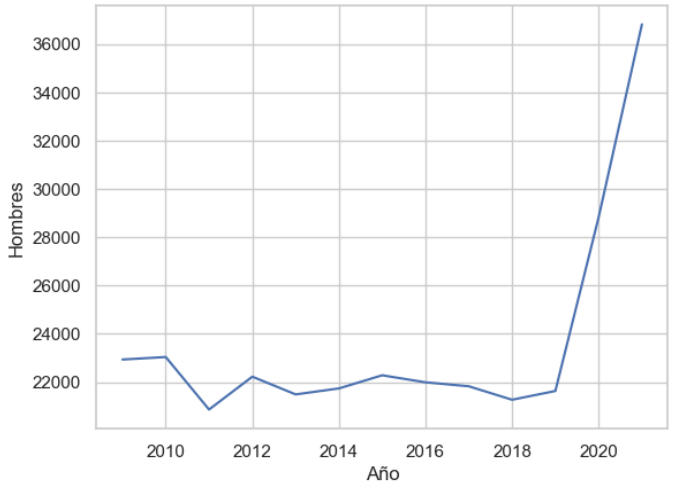
```
RangeIndex: 4863 entries, 0 to 4862
Data columns (total 6 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Año                       4863 non-null   int64
 1   Departamento de residencia  4863 non-null   object
 2   Causa de muerte           4863 non-null   object
 3   Total                     4863 non-null   object
 4   Hombres                   4863 non-null   object
 5   Mujeres                   4863 non-null   object
dtypes: int64(1), object(5)
```

```
Int64Index: 4291 entries, 0 to 4860
Data columns (total 6 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Año                       4291 non-null   int64
 1   Departamento de residencia  4291 non-null   object
 2   Causa de muerte           4291 non-null   object
 3   Total                     4291 non-null   int64
 4   Hombres                   4291 non-null   int64
 5   Mujeres                   4291 non-null   int64
dtypes: int64(4), object(2)
```

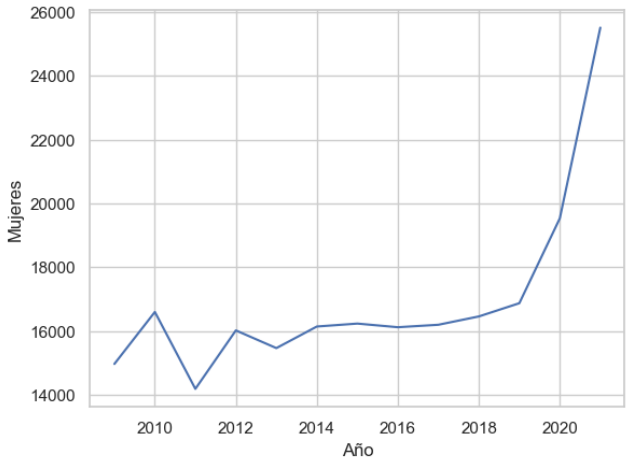
Antes y después de la limpieza y conversión de variables



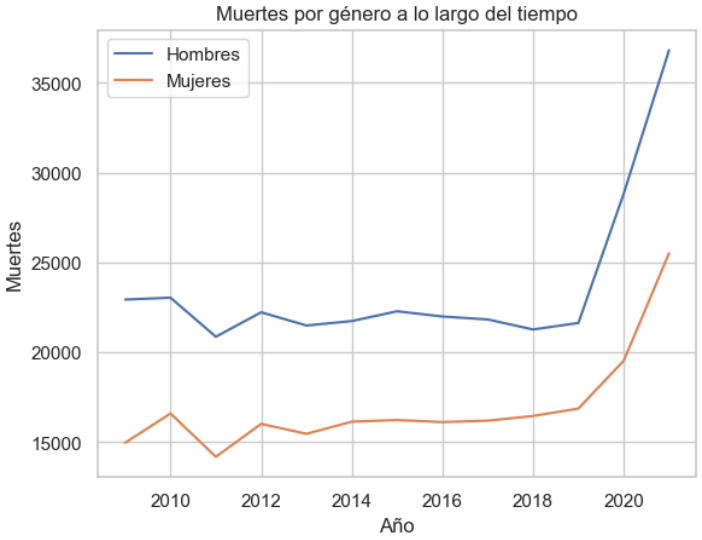
Mapa de calor para verificar que no haya datos nulos o faltantes



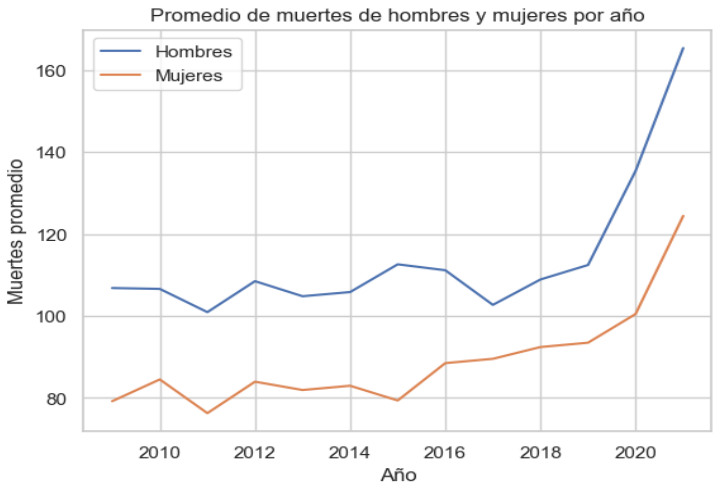
Gráfica lineal que relaciona la cantidad de hombres que fallecieron a lo largo de los 10 años



Gráfica lineal que relaciona la cantidad de mujeres que fallecieron a lo largo de los 10 años



Gráfica lineal que compara los resultados de los dos gráficos anteriores



Gráfica lineal que muestre el promedio de muertes por año de hombres y mujeres

El acercamiento que se tomó para realizar las gráficas lineales consistió en dos partes:

- Para las gráficas individuales de hombres y mujeres únicamente se graficaron los datos correspondientes
- Para la gráfica de promedio se agruparon los datos por año y se graficaron los promedios de hombres y mujeres que fallecieron.

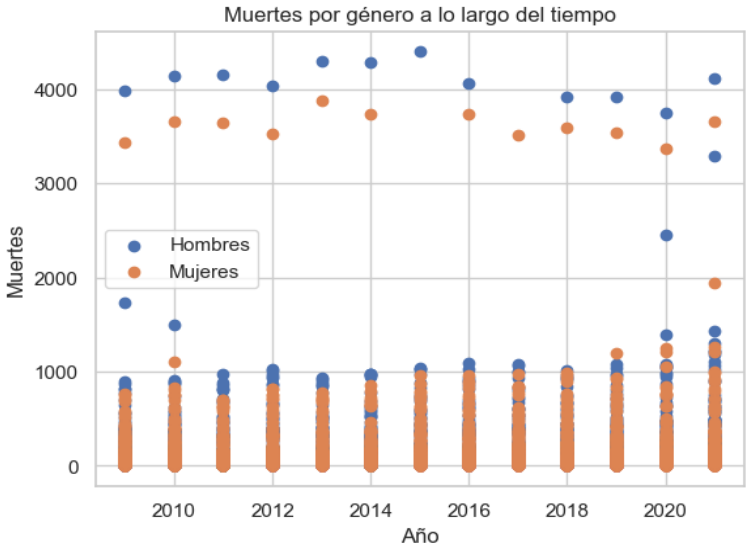
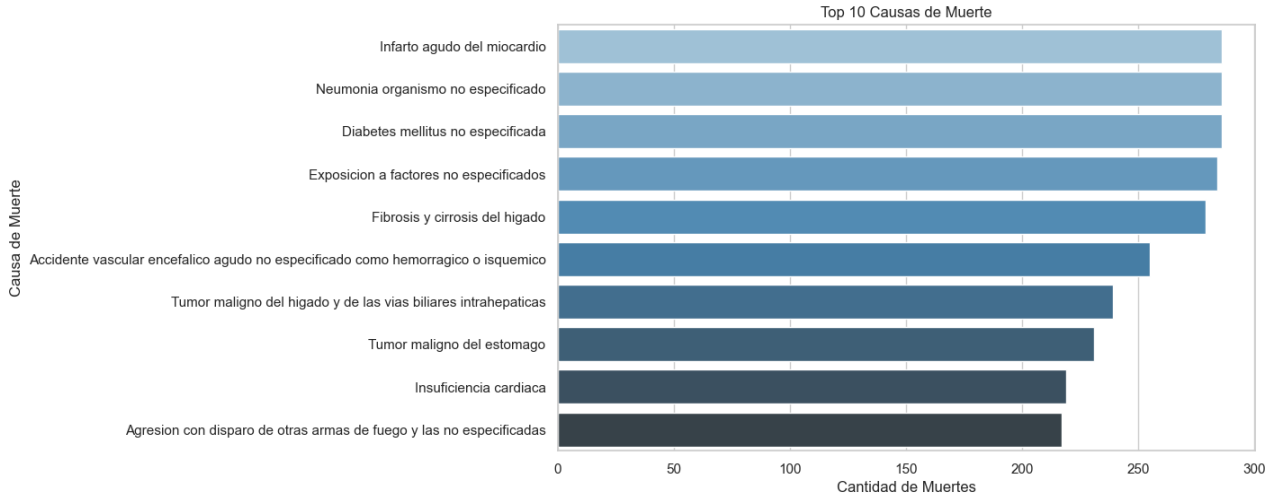
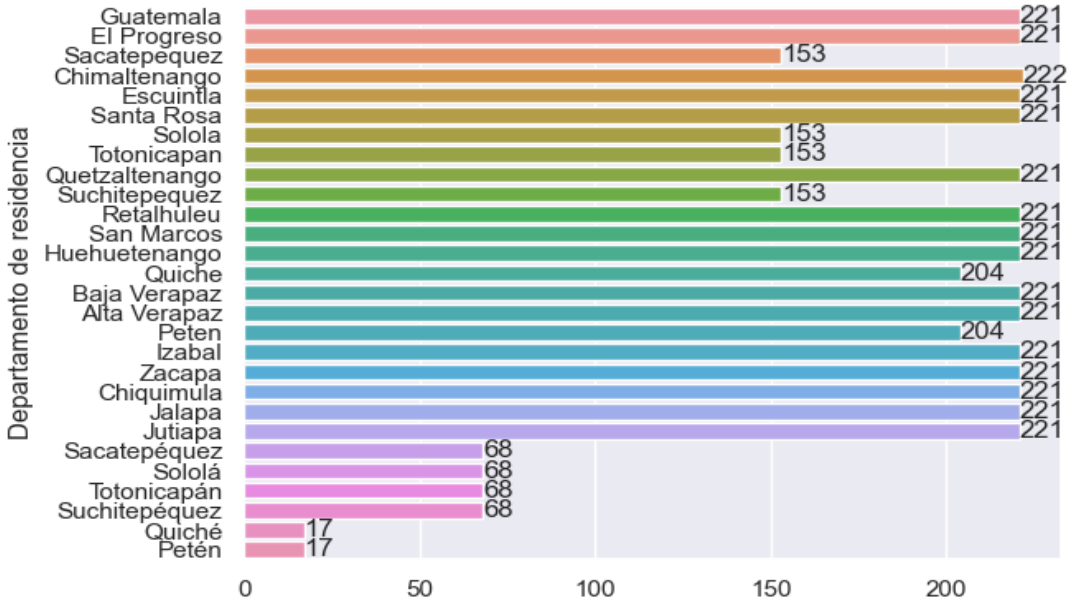


Diagrama de dispersión sobre las muertes por género a lo largo de los 10 años

En el gráfico de dispersión se tomaron los datos de cantidad de hombres y mujeres que fallecieron a lo largo de los diez años y se observa que la mayoría de datos se encuentran cercanos a las 1000 muertes y algunos datos atípicos superan las 2500 muertes mostrando que existen cierto tipo de causas que provocan más muertes que otras.

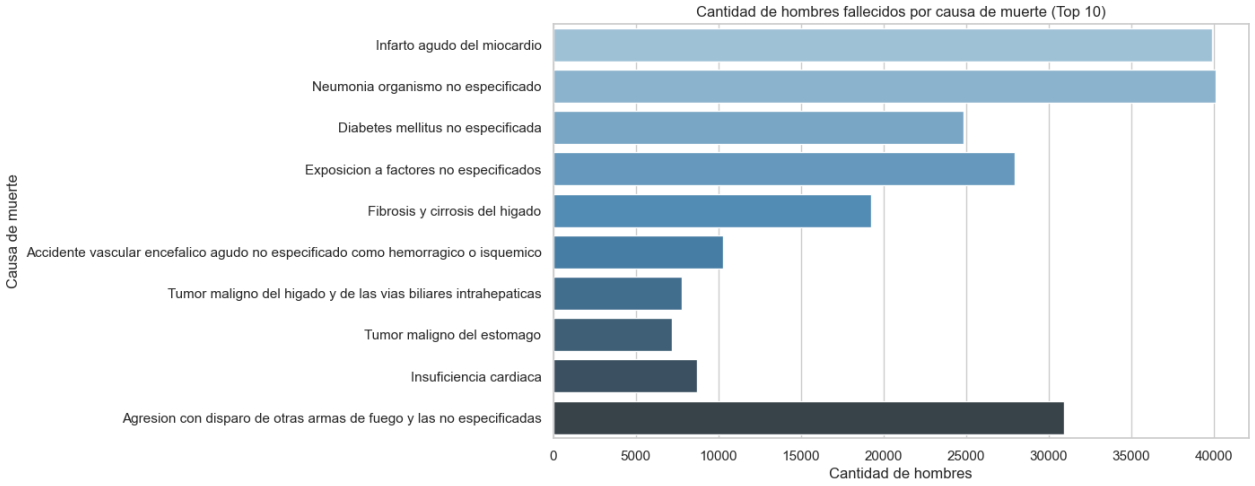


Top 10 Causas de muerte en el país a lo largo de los 10 años

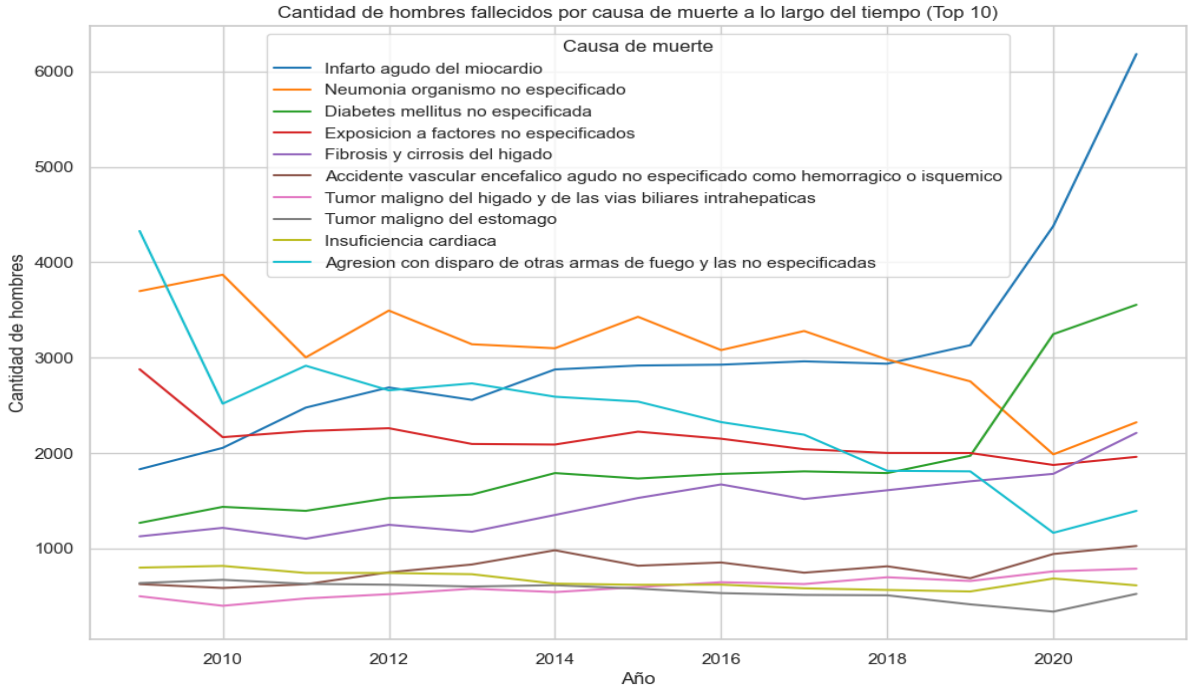


Cantidad de registros de defunciones por departamentos

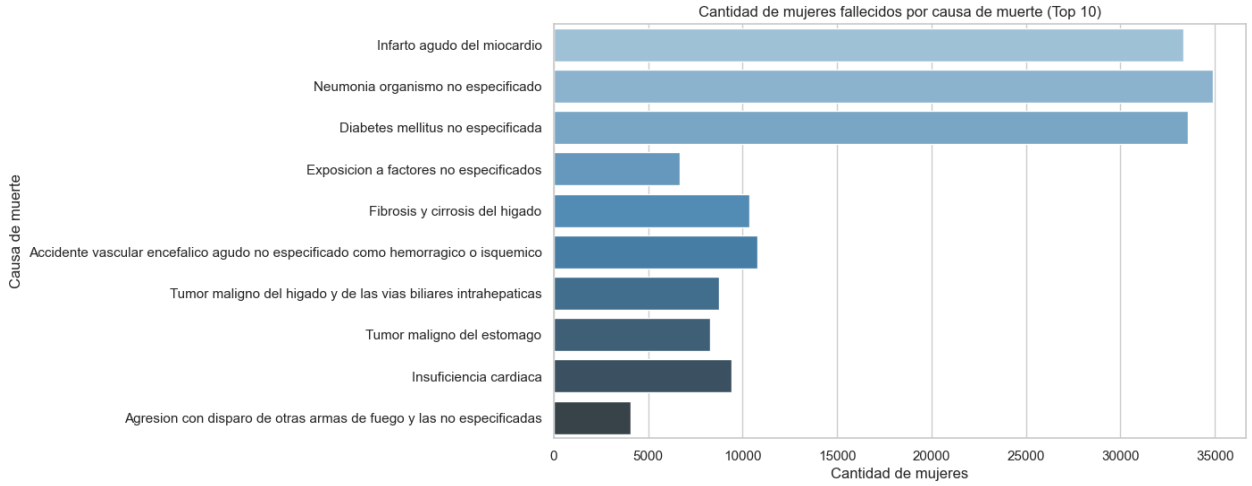
En estas gráficas de barras se realizaron conteos tanto de hombres como mujeres y también causas de muerte para conocer la cantidad de registros y las enfermedades que más muertes provocaban para reducir el campo de investigación a aquellos departamentos donde la tendencia a fallecer es más alta e investigar aquellas enfermedades que provocan un mayor número de muertes en la población.



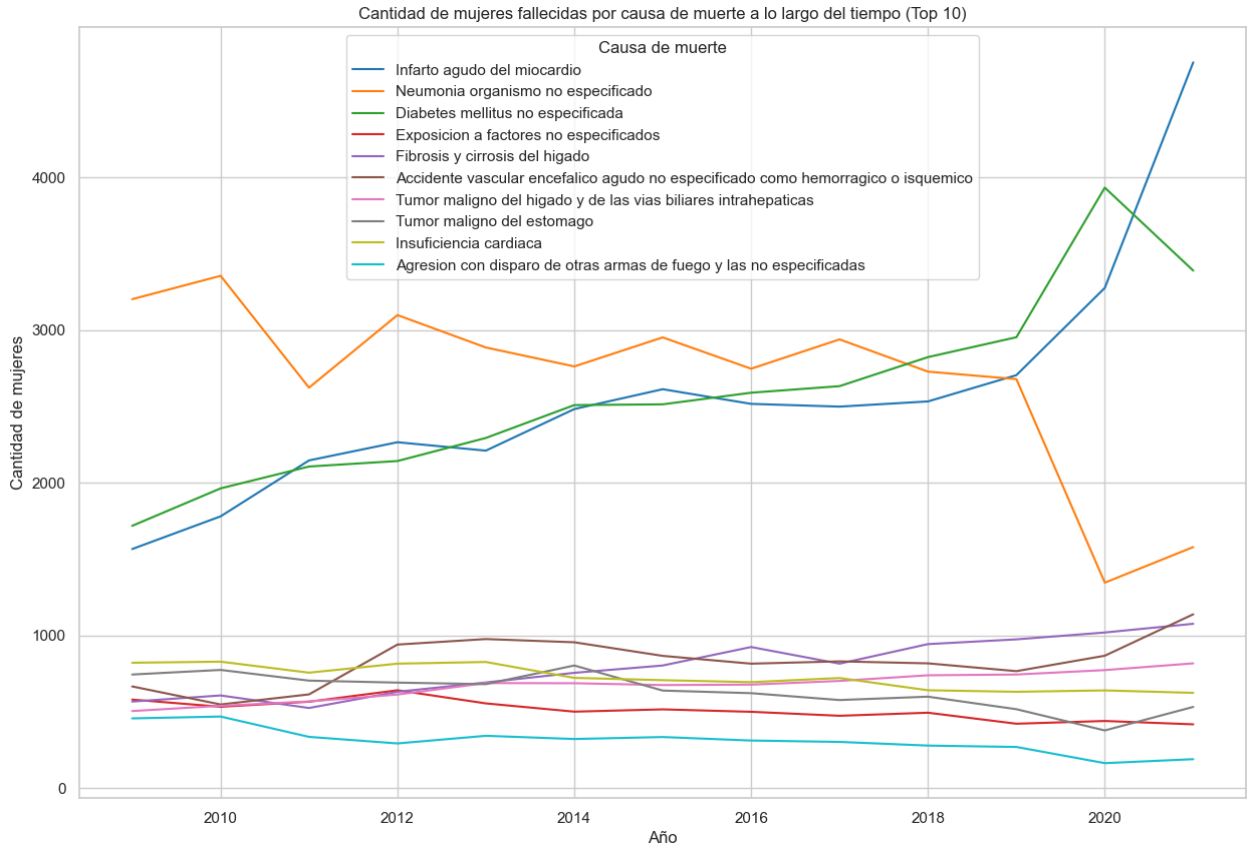
Gráfico



Top 10 Causas de muerte de hombres en el país a lo largo de los 10 años



Gráfico



Top 10 Causas de muerte de mujeres en el país a lo largo de los 10 años

Para estas últimas gráficas se buscó reducir los grupos por género y se agruparon por causa de muerte y año permitiendo conocer qué causas provocaban más muertes en hombres y en mujeres (independientes en cada caso) durante estos diez años y se observan datos interesantes como picos en causas de violencia para el género masculino pero no tanto para el género femenino.

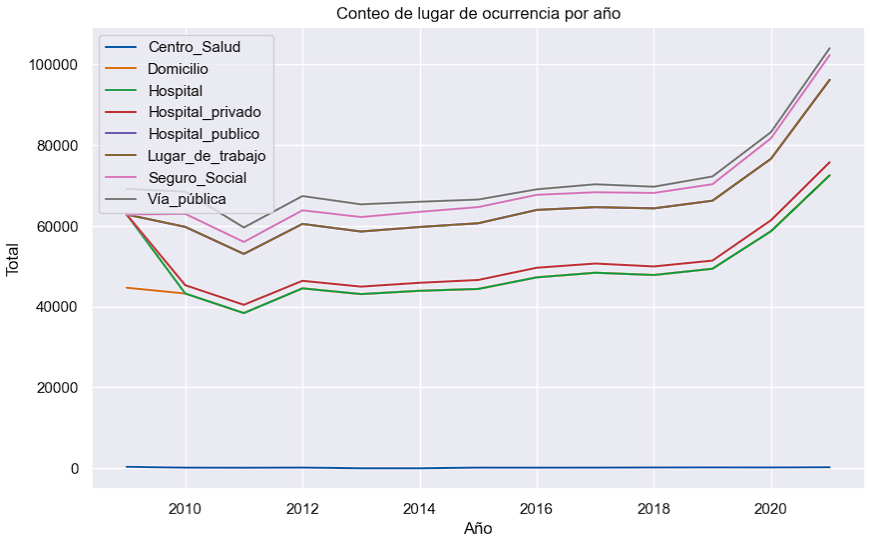
Defunciones en función del departamento, lugar de ocurrencia por departamento:

El siguiente set de datos tiene el objetivo de relacionar la información de: en qué departamento y en qué tipo de lugar ocurrió la defunción.

Para ello, los tipos de dato que se guardan son:

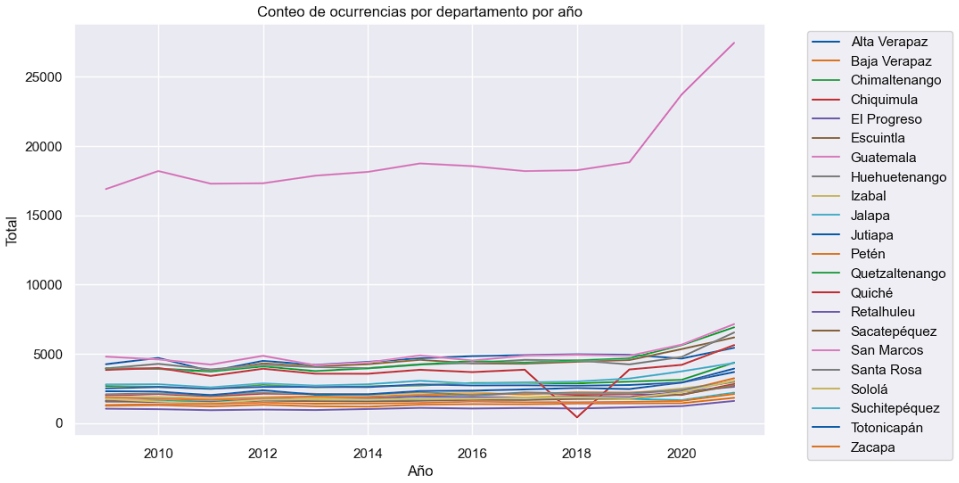
- Año: Año del conteo
- Departamento: Departamento en el que ocurrió el fallecimiento
- Total: Total de Fallecimientos en cierto año y cierto departamento
- Hospital_público: Total de fallecimientos por año y departamento ocurridos en hospitales públicos
- Hospital_privado: Total de fallecimientos por año y departamento ocurridos en hospitales privados
- Hospital: Total de fallecimientos por año y departamento ocurridos en hospitales sin reconocer privados de públicos
- Centro_Salud: Total de fallecimientos por año y departamento ocurridos en centros de salud.
- Seguro_Social: Total de fallecimientos por año y departamento ocurridos en el IGSS (Instituto Guatemalteco de Seguridad Social)
- Vía_pública: Total de fallecimientos por año y departamento ocurridos en calles y carreteras públicas
- Domicilio: Total de fallecimientos por año y departamento ocurridos en hogares o viviendas
- Lugar_de_trabajo: Total de fallecimientos por año y departamento ocurridos en áreas relacionadas al trabajo de las personas.
- Otro: Total de fallecimientos por año y departamento ocurrido en cualquier otro lugar que no sea de los establecidos anteriormente.

A partir de las variables descritas anteriormente se pudo obtener las siguientes gráficas y realizar las siguientes observaciones.



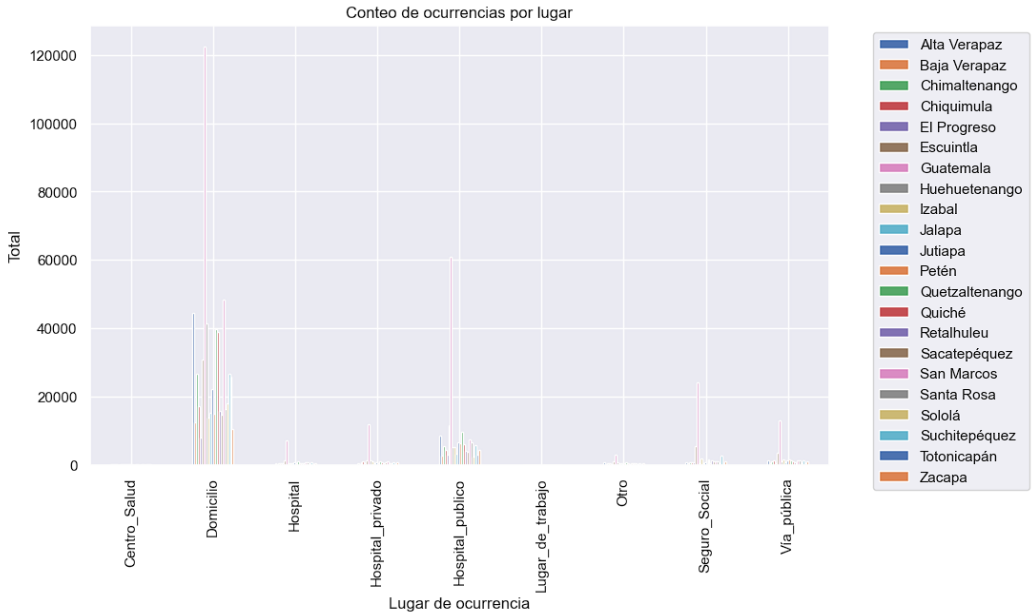
Conteo de lugares de ocurrencia por año

El gráfico anterior nos muestra la cantidad de defunciones por lugar de ocurrencia a través del tiempo. Información relevante que esta ofrece es que la mayoría de defunciones ocurrieron en vías públicas. Por el contrario, a comparación de los otros lugares, casi no ocurrió ninguna defunción en centros de salud.



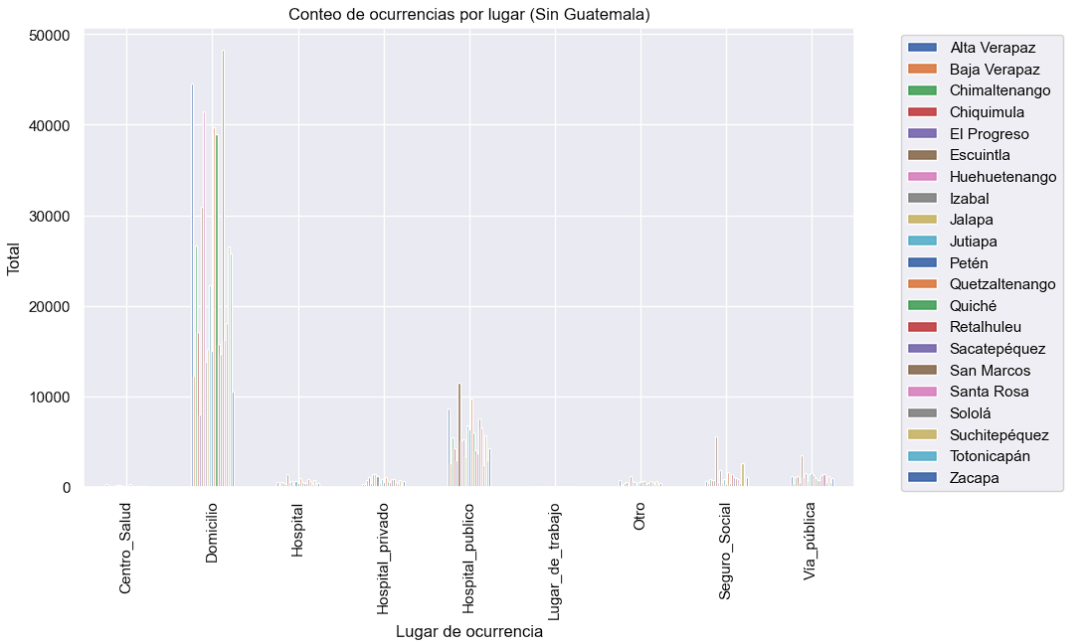
Conteo de defunciones por departamento por año

Este gráfico nos presenta la cantidad de defunciones ocurridas por departamento y por año. Partes relevantes del mismo son que en el departamento de Guatemala es el que mayor cantidad de defunciones por más de diez mil muertes. Otra curiosidad es el pico que hubo en 2018 en el departamento de Chiquimula.



Conteo de defunciones por lugar de ocurrencia por departamento

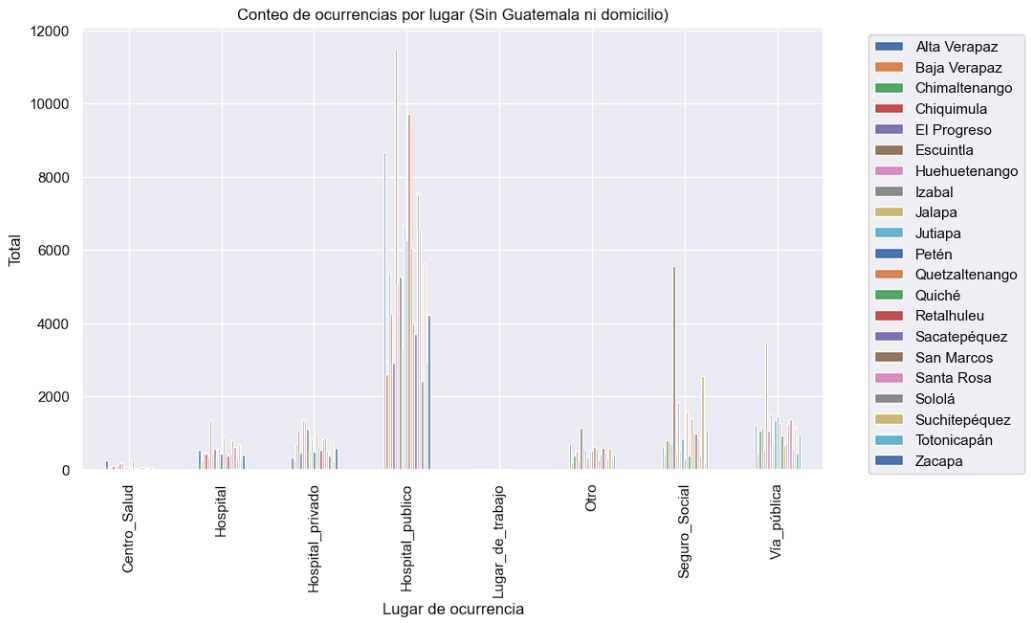
Debido a la gran cantidad de departamentos y de lugares, es un poco difícil de observar la información. A pesar de ello, este gráfico nos indica que la mayoría de defunciones a través de los años han ocurrido en los domicilios de las personas, especialmente en el departamento de Guatemala.



Conteo de defunciones por lugar de ocurrencia por departamento sin Guatemala

Para poder observar los datos de mejor manera y ver más información sobre los otros departamentos, se decidió hacer unos pequeños cambios a los datos y omitir datos que sesguen al set de datos, para que de esa manera se puedan hacer mejores comparaciones entre los datos no sesgados. Por ello primero se decidió esconder el departamento de Guatemala,

para poder observar las defunciones de todos los otros departamentos. Gracias a ello pudimos observar que el segundo departamento con más defunciones es Quetzaltenango.



Conteo de defunciones por lugar de ocurrencia sin Domicilios por departamento sin Guatemala

Siguiendo con el objetivo de poder hacer mejores observaciones sin datos que sesguen al set de datos, en esta gráfica se decidió no solo omitir el departamento, sino que también al lugar de ocurrencia, domicilios. Gracias a ello, podemos observar que los lugares en los que ocurren más defunciones luego de los domicilios son los hospitales públicos y el Seguro Social, principalmente en Escuintla.

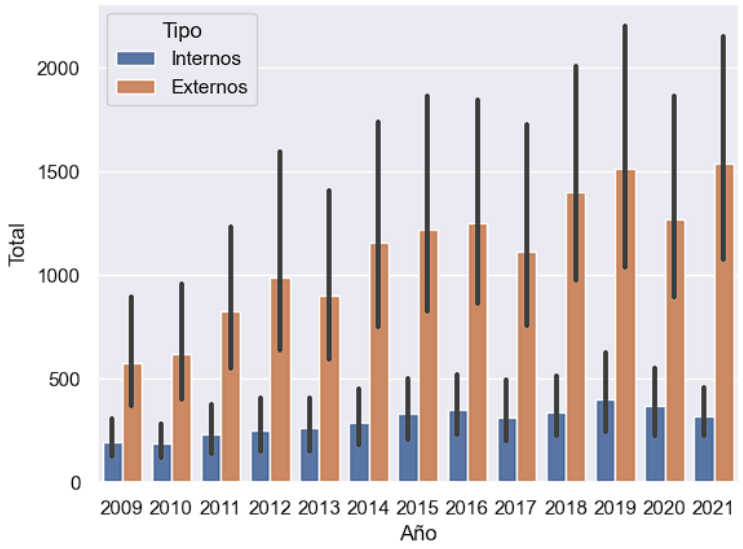
Servicios médicos en función de departamento, tipo y sexo:

Para cumplir el objetivo de identificar factores que influyen en la mortalidad de la población Guatemalteca, se decidió obtener datos generales sobre servicios médicos. Los datos que se obtuvieron son los siguientes.

- Año: Año en el que los servicios médicos fueron brindados
- Tipo: Si el servicio fue interno o externo, esto significando si fueron internos, los pacientes fueron ingresados a un establecimiento que brinda servicios médicos tal como hospitalización, operaciones, y consultas. Mientras que los externos se brindan fuera del entorno clínico tal como son las consultas a domicilio o clínicas ambulatorias.
- Todas las causas: Las causas por las cuales las personas solicitaron y se les ofreció servicios médicos.
- Total: Total de casos de personas atendidas por año, tipos, departamento y causa.
- Hombres: Total de hombres atendidos por año, tipo, departamento y causa.
- Mujeres: Total de mujeres atendidas por año, tipo, departamento y causa.

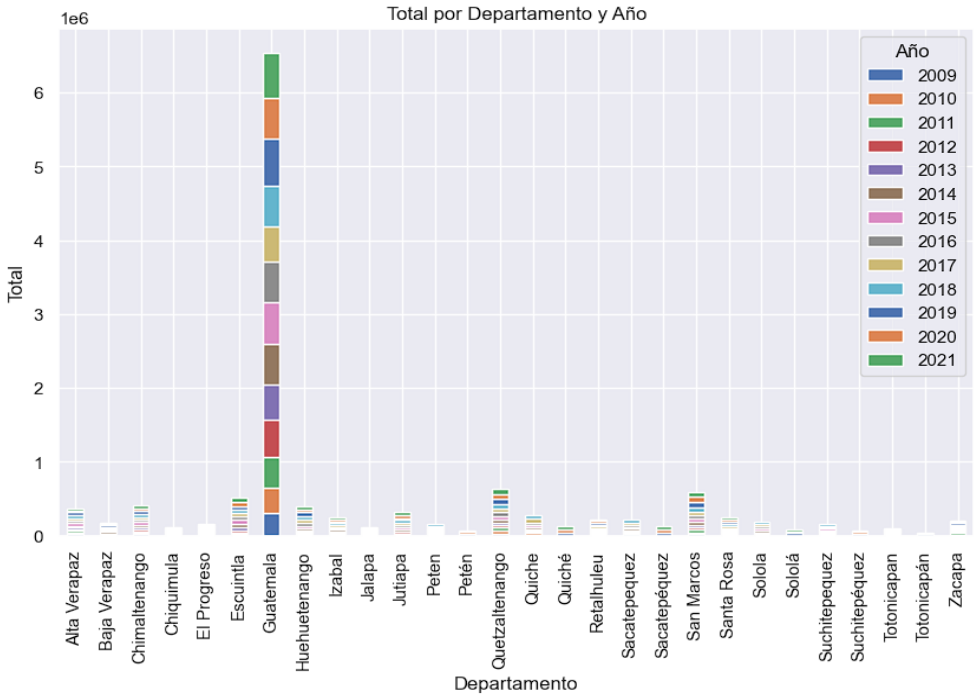
- Ignorado: Total de personas sin haber almacenado el sexo de la misma atendidos por año, tipo, departamento y causa.

A partir de las variables descritas anteriormente se pudo obtener las siguientes gráficas y realizar las siguientes observaciones.



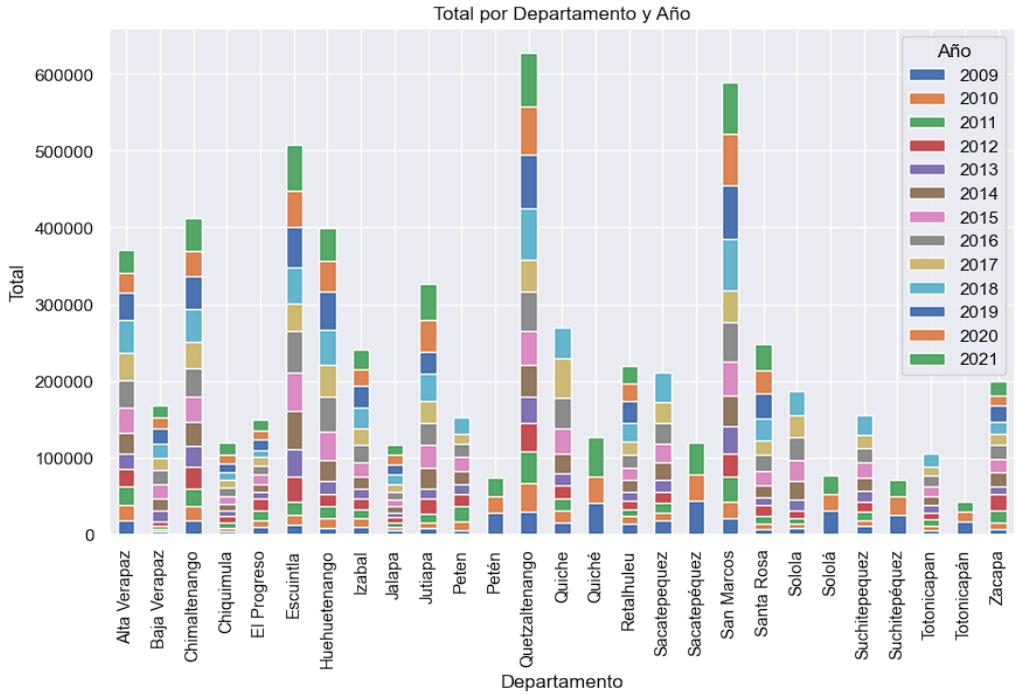
Tipo de servicio médico por año

La primera gráfica nos presenta la cantidad de servicios por tipo de servicio. De ella podemos observar que en todos los años, la mayoría de servicios son ofrecidos de manera externa, y no de manera interna.



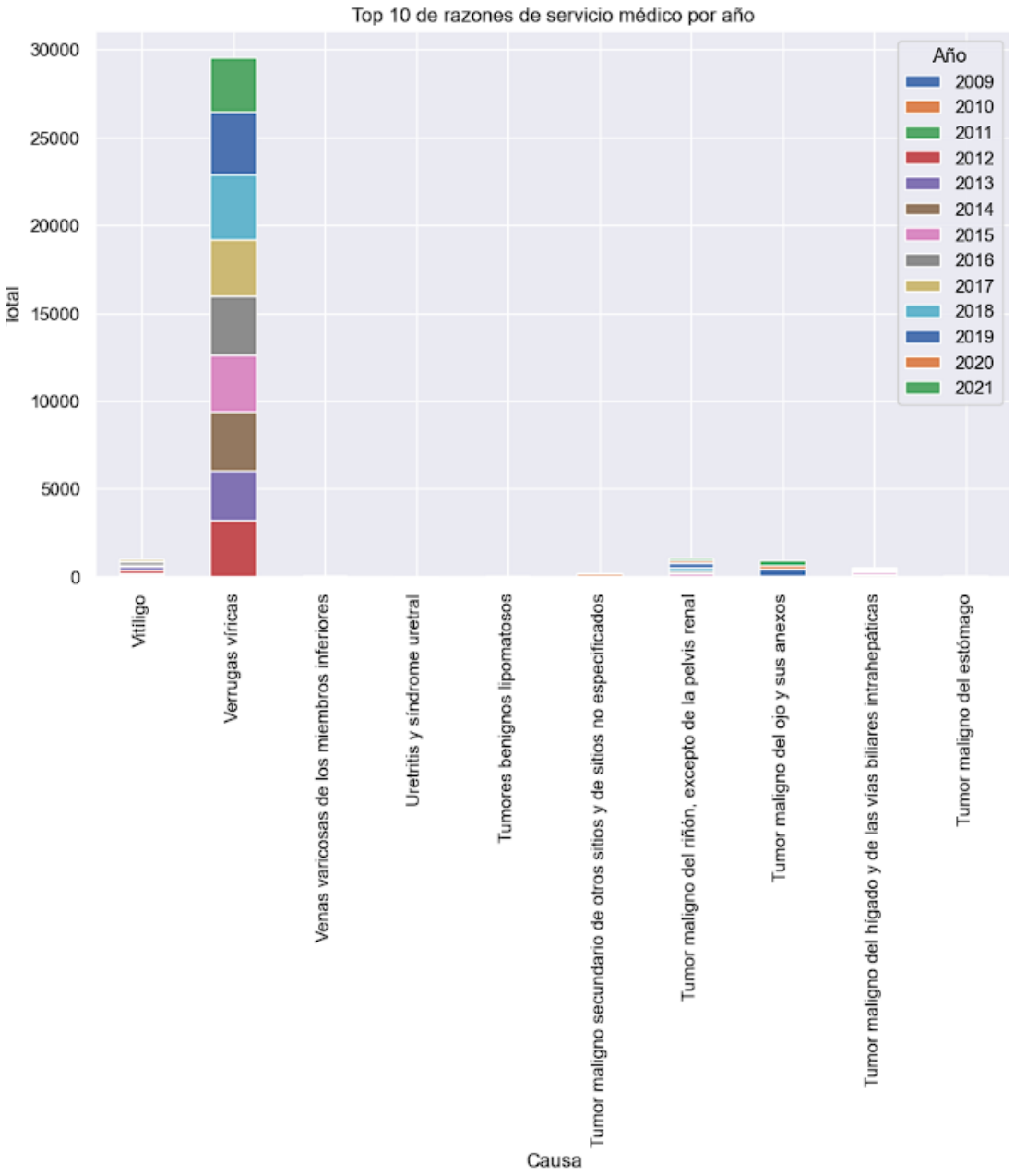
Total de servicios médicos por departamento y por año

En esta tabla nos muestra que la mayoría de servicios médicos son ofrecidos en el departamento de Guatemala, especialmente en el año 2021,



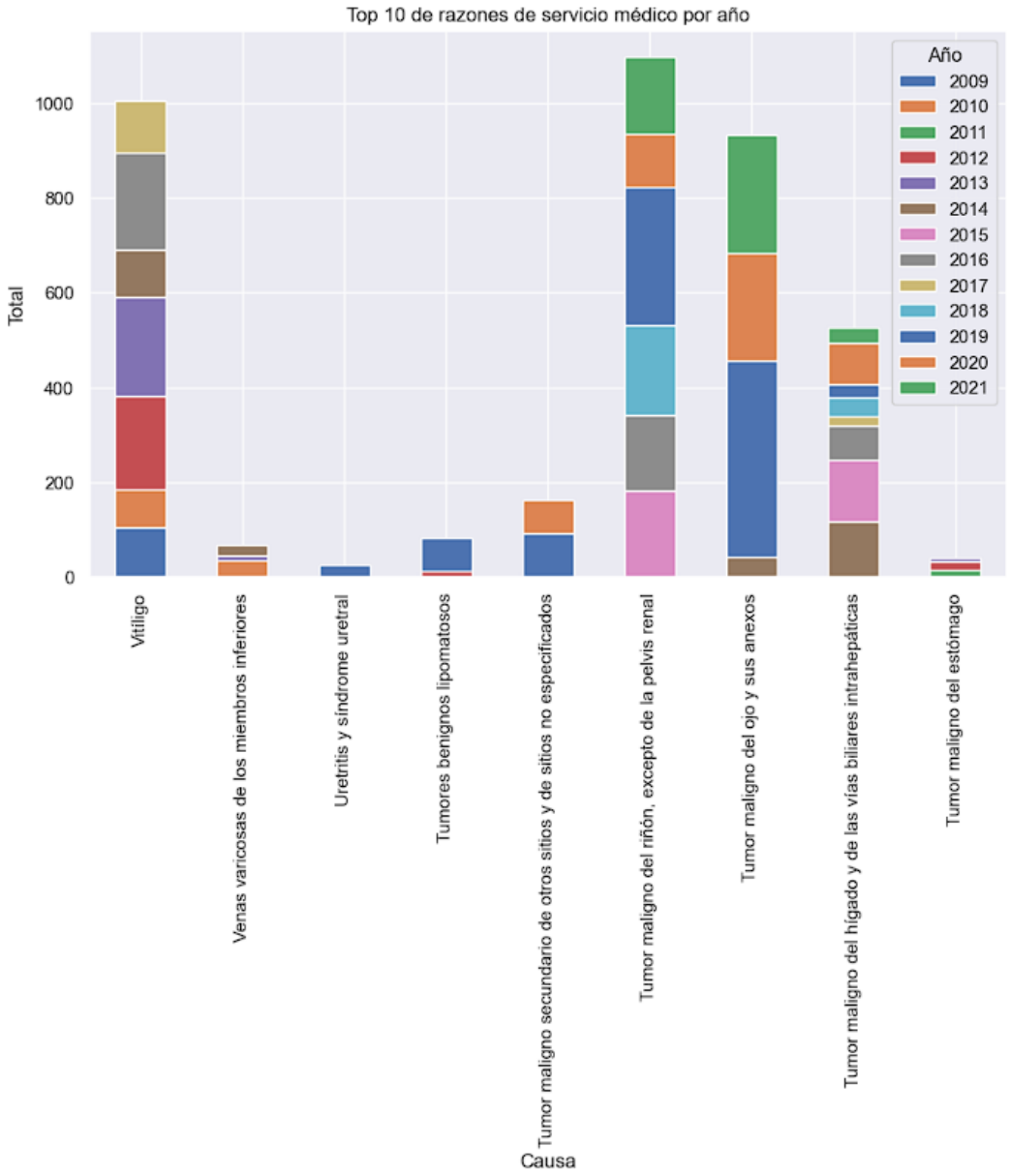
Total de servicios médicos por departamento sin Guatemala y por año

Para poder hacer mejores observaciones, en esta tabla también se decidió ignorar las variables que sesgan al set de datos, en este caso nuevamente se decidió omitir al departamento de Guatemala para poder observar la información de todos los otros departamentos. Gracias a esto podemos saber que luego de Guatemala, los otros 2 departamentos que más brindaron servicios de salud fueron Quetzaltenango y San Marcos.



Top 10 de causas de servicio médico

Este gráfico nos ayuda a observar cuál, a través de todos los años, fue la razón de mayores consultas médicas, y por una gran ventaja esta es por verrugas víricas, las cuales son consecuencias de infección vírica de la piel por algunos tipos de virus del papiloma humano (VPH).

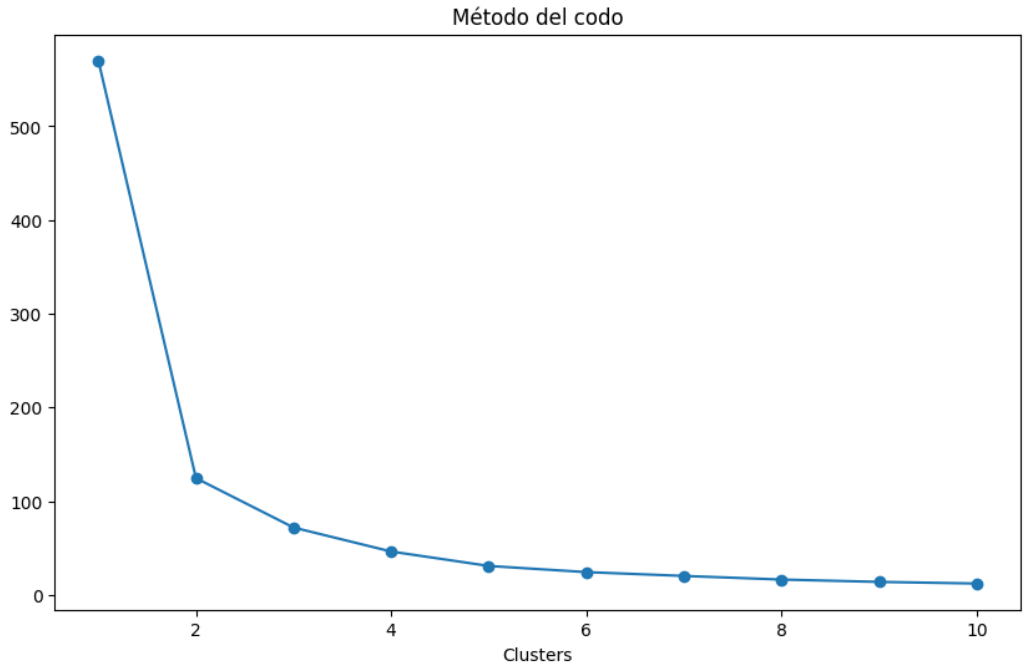


Top 10 de causas de servicio médico sin incluir VPH

Debido a que el VPH es el síntoma con la mayor cantidad de ocurrencias, nuevamente se decidió omitir el dato para poder observar de mejor manera las otras razones de servicios médicos. Por ello, gracias a este gráfico podemos observar que la segunda y tercera mayor razón del ofrecimiento de servicios médicos fueron tumores malignos en el riñón y en el ojo.

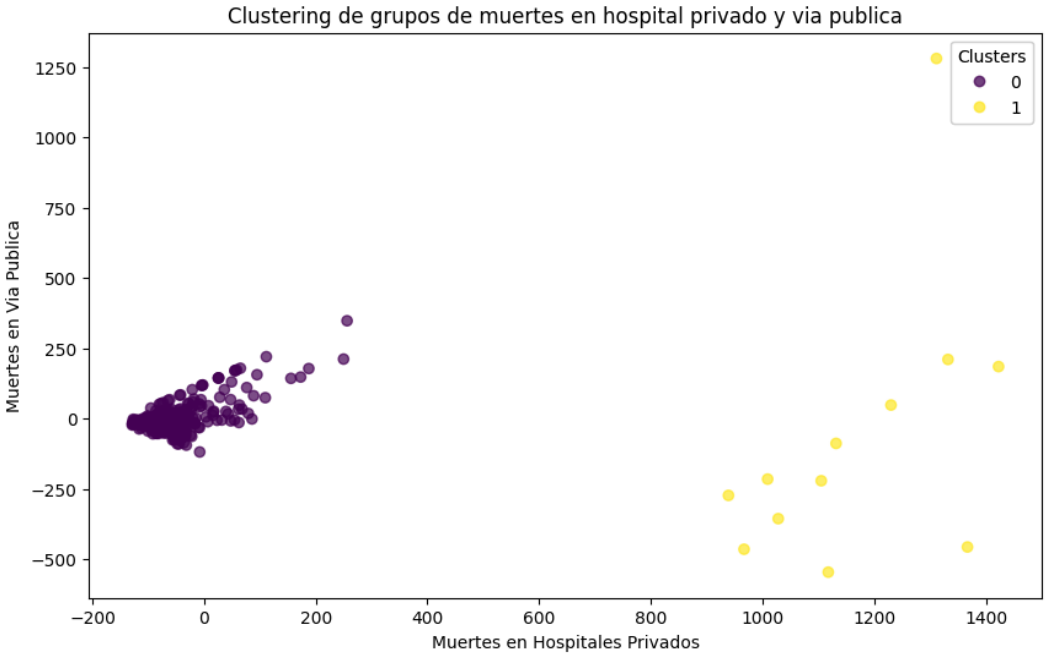
Clustering

Para realizar el Clustering primero se designaron los datos a utilizar, en este caso se tomó la tabla de defunciones por lugar de ocurrencia. Luego se realizo el método del codo para encontrar el número óptimo de clusters:



Método del codo para encontrar el número óptimo de clusters para hospitales privados vs via pública

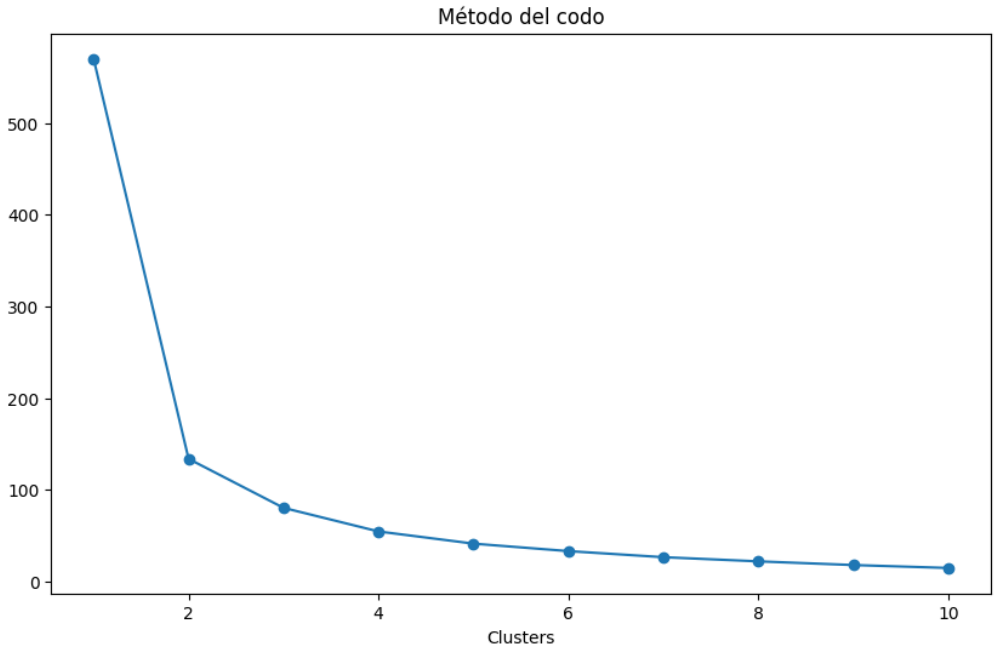
Con los resultados de la tabla observamos que el codo se encuentra en el punto 2, de modo que íbamos a encontrar dos regiones:



Clustering de grupos de muertes en hospital privado y vía pública

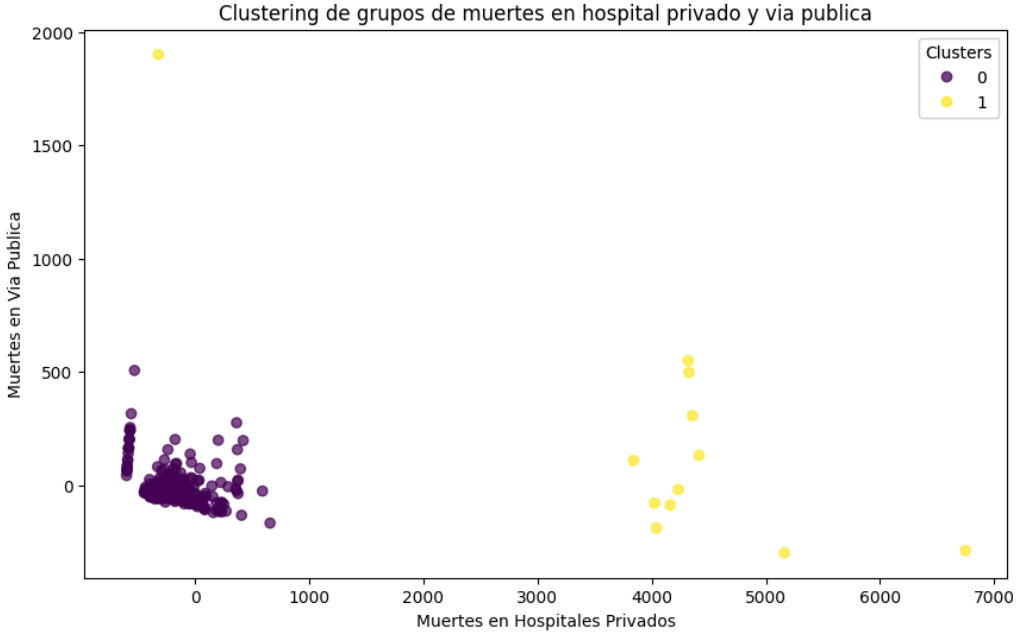
De esa manera se encuentra un grupo de departamentos que tienen una mayor captación de pacientes a hospitales privados ya que disminuye las muertes en la vía pública.

Asimismo se hizo el mismo procedimiento con los hospitales públicos para hallar alguna diferencia.



Método del codo para encontrar el número óptimo de clusters para hospitales privados vs via pública

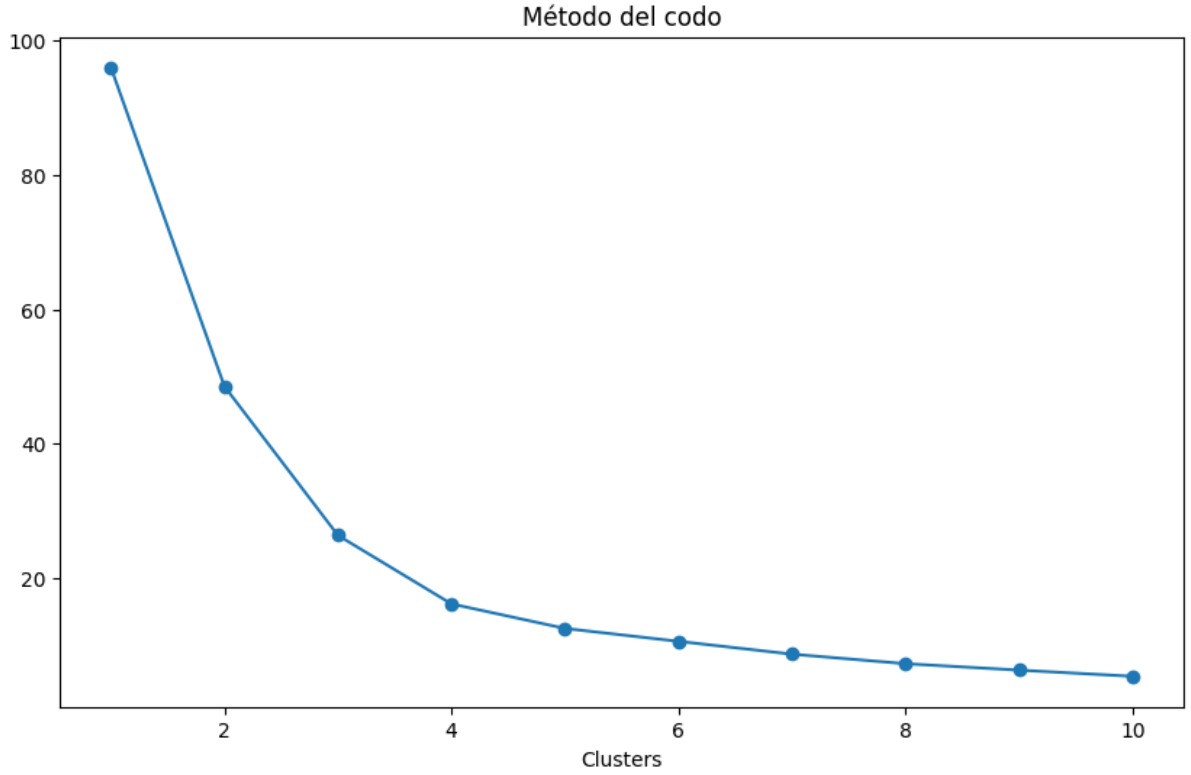
En el que igualmente se encuentran dos grupos



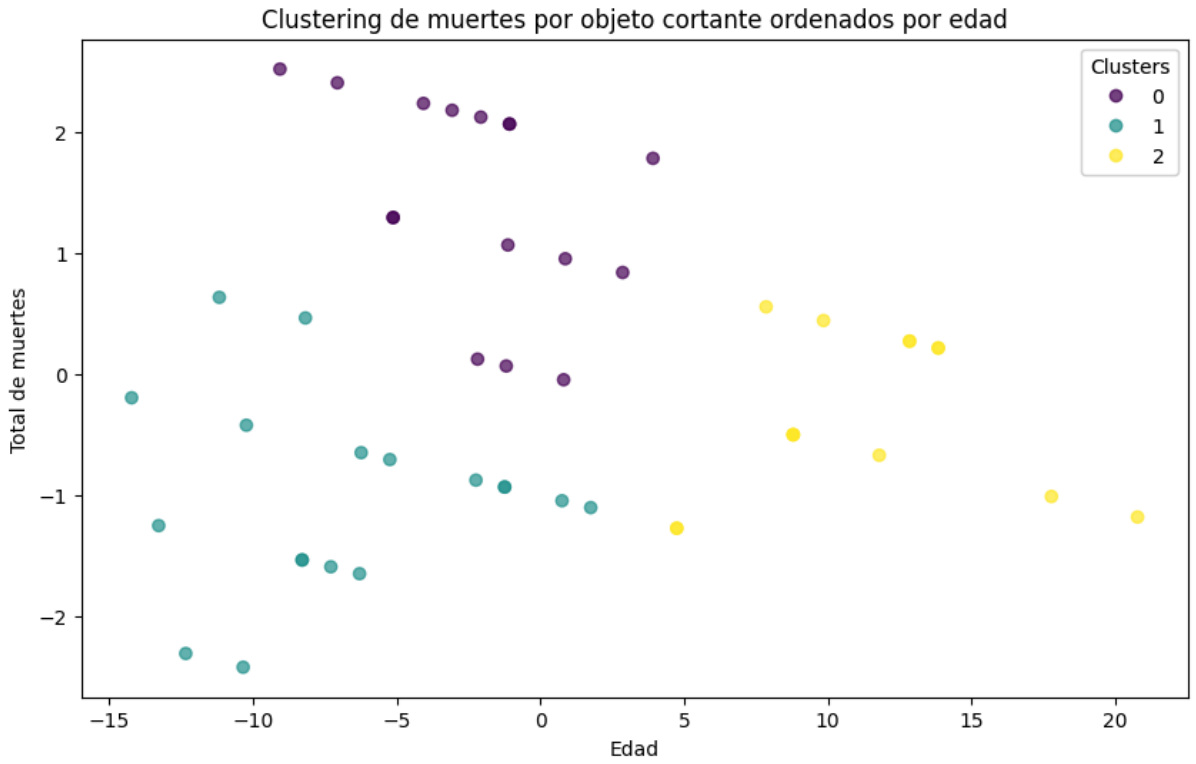
Clustering de grupos de muertes en hospital privado y vía pública

Y se nota que en este caso el impacto de los hospitales públicos es menor, ya que al momento de aumentar las muertes captadas a hospital público, el impacto en muertes en vía público es menor que el de hospitales privados.

Luego se decidió formar un cluster con la causa de muerte “Agresión con objeto cortante” para ello se normalizaron las edades a números y se aislaron las demás variables:



En este método del codo se observa que el número óptimo de clusters es 3:



Se pueden observar tres grupos, a los que definiremos como dos grupos menos activos: el 1 y el 2 que se podrían catalogar como accidentes. Y un grupo activo: el 0, que se puede considerar más “buscapleitos”

Hallazgos y conclusiones

1. Luego de haber realizado el proceso de limpieza en los conjuntos de datos utilizados la cantidad de datos restante era de una proporción lo suficientemente amplia para llevar a cabo nuestro análisis exploratorio donde encontramos la relación entre las diferentes variables que existían y esto nos dejó una visión general acerca de las defunciones ocurridas en los últimos 10 años.

    a. Por cada una de las causas que se presentan en los datos existen al menos 50 defunciones. Aparte de esto, dado el conjunto que describe todas las edades se encontró que los grupos que más defunciones presentan son los mayores de edad y las personas cercanas a los 25 años.
    b. En cuanto al conjunto de datos que describen las defunciones por departamento de residencia y otras variables muestra varias cosas:
        i. La cantidad de defunciones masculinas durante los últimos 10 años supera por miles la cantidad de defunciones femeninas.
        ii. En el año 2011 la cantidad de defunciones para ambos géneros tuvo un descenso bastante grande en comparación a los demás años.
        iii. En el año 2015 hubo un caso especial en el promedio de muertes por año ya que mientras la cantidad de defunciones femeninas disminuyó

respecto al año anterior, la cantidad de defunciones masculinas aumentó.

iv. En el año 2020 y 2021 se observa un pico enorme en la cantidad de defunciones que hacen sentido y coinciden con todo el tema del SARS-CoV-2.

v. Durante los últimos 10 años el total de defunciones por género y por año se mantuvo desde las 1000 hasta las 4500 defunciones aproximadamente.

vi. Las causas de muerte que más se presentaron fueron: Infarto agudo del miocardio, Neumonía, Diabetes mellitus, Tumores, agresiones con disparo de arma de fuego, entre otras.

vii. Los departamentos que más defunciones registraron fueron Guatemala, El Progreso, Quetzaltenango, Chimaltenango, Escuintla, entre otros. Y los que menos defunciones presentaron están Quiché y Petén.

viii. Las causas de muerte que más defunciones masculinas generan son: Infarto agudo del miocardio, Neumonía y Agresión con Arma de fuego. Y las causas de muerte que más defunciones femeninas genera son: Infarto agudo del miocardio, Neumonía y Diabetes mellitus

c. El conjunto de defunciones por departamento y lugar de ocurrencia de las mismas mostró los siguientes hallazgos:

i. La mayoría de las defunciones ocurre en vías públicas, casi no se muestran registros en centros de salud.

ii. La mayor parte de las defunciones se han registrado en la capital, los demás departamentos presentan cantidades similares, exceptuando Chiquimula que tuvo un descenso en la cantidad de defunciones en el año 2018.

iii. En cuanto a lugares de ocurrencia, el domicilio de las personas es el lugar más común donde fallecen. Y dentro de estos datos, Quetzaltenango es el segundo departamento con más defunciones registradas en domicilios, el primero es la capital. El segundo y tercer lugar son los hospitales públicos y el seguro social respectivamente.

d. Finalmente en el conjunto de datos, de servicios médicos, se encontraron los siguientes resultados:

i. La mayoría de servicios se ofrecen de manera externa y no tanto de manera interna. Con esto se hace referencia a lo siguiente:

1. Servicios externos: Consultas médicas, visitas a enfermerías, visitas a domicilios, etc.
2. Servicios internos: Hospitales, internados, presencia en los mismos por tiempo indefinido.

ii. La mayor parte de servicios médicos se ofrecen en la capital. Siendo los departamentos de Escuintla, Chimaltenango, San Marcos, y Quetzaltenango los que ofrecen más servicios de este tipo luego de Guatemala.

iii. Una de las razones por las que se realizaban más consultas médicas era por verrugas víricas, las cuales son consecuencias de infección vírica de la piel por algunos tipos de virus del papiloma humano (VPH).
    1. Fuera de esta razón las que más se presentaban eran tumores malignos en el riñón y los ojos.

2. Como siguientes pasos en la investigación basándonos en los resultados obtenidos tenemos que:
    a. Investigar más a fondo sobre lo que representan las causas catalogadas como las más recurrentes para determinar más relaciones entre el ambiente en el que se llegan a desarrollar y qué grupos se ven más afectados por las mismas.
    b. Determinar qué influencia existió en los años donde se mostró un declive en la cantidad de defunciones registradas y también por qué en dichos establecimientos o departamentos se presentó este comportamiento.
    c. Establecer otros clusters para encontrar más grupos basándonos en diferentes variables y que nos permitan desarrollar un mejor trayecto en la búsqueda de los grupos más vulnerables ante las causas encontradas y otros factores.

Link del repositorio: https://github.com/MaIsabelSolano/UVG_Mineria_Proyecto.git
Link de documento de google drive:
https://docs.google.com/document/d/1SVzWuh12hG6KiqTlrELyGFX9DXAoTAApcSAPweKU5J0/edit?usp=sharing