Universidad de los Andes

Facultad de Ingeniería
Departamento de Ingeniería de Sistemas
Inteligencia de Negocios
Periodo 2023-10

Laboratorio 1:
Clustering

Esteban Gonzales Ruales - 202021225
Mariana Díaz Arenas - 202020993
Juan Diego Yepes Parra - 202022391

10 de febrero de 2022
Bogotá D.C.

1. Introducción

BiciAlpes es una organización que ofrece un servicio solidario de alquiler de bicicletas con múltiples beneficios para sus usuarios. Su principal objetivo es fomentar planes de movilidad sostenible en áreas urbanas para reducir las emisiones de gases de efecto invernadero que contribuyen al calentamiento global. Sin embargo, la organización ha notado una falta de información sobre la seguridad en las carreteras, lo que está afectando el uso de bicicletas como medio de transporte. Por lo tanto, BiciAlpes decidió llevar a cabo un estudio para determinar los factores que influyen en los accidentes viales que involucran a ciclistas. Este conocimiento podría ser útil para las autoridades y los planificadores urbanos para implementar medidas que reduzcan la frecuencia de los accidentes y promuevan la movilidad sostenible. Para llevar a cabo el estudio, BiciAlpes ha recopilado datos de accidentes que involucran a ciclistas de fuentes abiertas de la Alcaldía.

Nota: En este documento presentamos unos datos en anexos. No es necesario verlos pero pueden complementar nuestro desarrollo.

2. Entendimiento de los datos

En primer lugar, para entender bien los datos que tenemos a nuestra disposición hicimos un análisis de lo que quiere el negocio y lo que podemos encontrar con los datos que tenemos[1]. Para ello, cargamos los datos en un dataFrame de Pandas. Primero vemos que tenemos 5338 filas y 14 columnas. Se ve de la siguiente forma:

	Time	Number_of_Casualties	Day_of_Week	Road_Type	Speed_limit	Light_Conditions	Weather_Conditions	Road_Surface_Conditions	Urban_or_Rural_Area	Did_Police_Officer_Attend_Scene_of_Accident
0	1	1	1	6	30.0	1	1	1	1	1
1	1	1	1	6	30.0	1	1	1	1	1
2	2	1	2	6	30.0	1	1	1	1	1
3	2	2	1	6	30.0	1	1	1	1	1
4	1	2	1	6	30.0	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...
5333	2	1	1	6	20.0	1	1	2	1	1
5334	3	2	1	6	30.0	1	1	1	2	2
5335	2	1	1	6	30.0	4	5	2	1	2
5336	3	1	1	6	30.0	1	1	1	1	1
5337	1	1	1	3	70.0	1	1	1	1	2

5317 rows × 14 columns

En este dataframe tenemos tipos de variables numéricas, como `Speed_Limit`, `Time` y `Number_of_Casualties`, para nombrar algunas, y el resto son categóricas. Para las categóricas algunas tenían ya su un valor numérico asociado a la misma, pero algunas otras no. Por ejemplo, `Day_of_Week` tiene los posibles valores día laboral y fin de semana, pero `Accident_Severity` si tiene su respectiva enumeración: (Fatal = 1, Serio = 2, Leve = 3).

De igual forma, aquí se muestran algunas de las medidas de tendencia central de cada columna, en concordancia con la estadística descriptiva que se pide:

[1] Esta sección hace referencia a lo desarrollado en la sección 1 del notebook

	$\bar{x}$	$\tilde{x}$	σ_x^2	σ_x
Number_of_Casualties	1.11	1.00	0.10	0.33
Day_of_Week	1.27	1.00	0.20	0.44
Time	1.96	2.00	0.59	0.77
Road type	5.59	6.00	1.99	1.41
Speed_limit	33.51	30.00	104.11	1.45
Light_conditions	1.71	1.00	2.11	1.45

Tabla 1: Estadísticas sobre los datos[2]

Siguiendo con el análisis, pudimos ver que algunas de las columnas o filas no eran relevantes para el estudio y las podíamos quitar de nuestro data set. Esto se debe a que su valor era constante en toda la tabla, como el valor de `Vehicle_Type`, o porque tenían valores nulos como `Day_of_Week`, y demás.

2.1. Completitud

```
(df_roads.isnull().sum() / df_roads.shape[0]).sort_values(ascending = False)
```

Para ver la completitud de los datos ejecutamos el siguiente código que nos determina cuáles columnas tienen valores nulos y el porcentaje de los mismos. Esto nos arrojó que la única columna con valores nulos era `Day_of_Week`

2.2. Unicidad

En el caso de nuestro estudio, hacemos la suposición de que pueden haber valores duplicados puesto que cada una de las filas representa un evento; y puede pasar que los eventos se repitan con las mismas condiciones.

2.3. Consistencia, validez

En cuanto a consistencia y validez (los agrupamos porque son muy similares) lo que hicimos fue revisar si los valores tenían concordancia con lo que se espera que sean; es decir, que no hayan diferentes formas de escribirse (como 'Día laboral' o 'DIA LABORAL') y que estén dentro de los rangos indicados. Sobre ello obtuvimos que `Did_Police_Officer_ Attend_Scene_of_Accident` tenía valores negativos, lo que es inadecuado, y que eran 2 valores erróneos, los cuales eliminamos para nuestro estudio.

3. Preparación de los datos

Para limpiar los datos que no nos sirven en nuestro estudio, hicimos lo que se presenta en el notebook en la sección 2. En este orden:

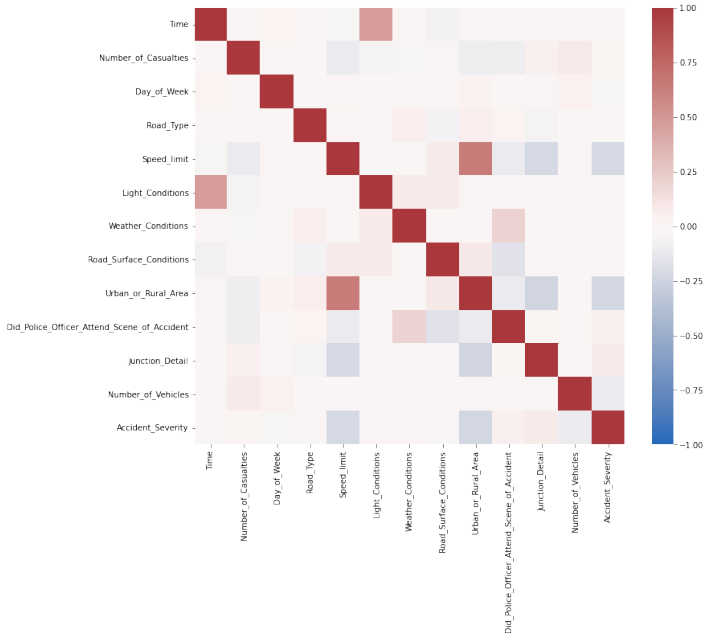
1. Hacemos `.drop()` de las columnas `Vehicle_Type` y de las filas en donde `Did_Police...` tiene un valor negativo, lo cual no es coherente.

[2] Algunos de los features aquí fueron transformados para ser numéricas. Ver sección 3 de este documento

2. Hacemos `.dropna()` de la columna `Day_of_Week` para eliminar los valores nulos
3. Reemplazamos los valores categóricos no numéricos por numéricos.

4. Modelado

Para saber qué queremos comparar y cómo lo primero que hicimos fue hallar la correlación entre los datos, ya limpios, que tenemos. Esto lo visualizamos mediante dos gráficos distintos[3].

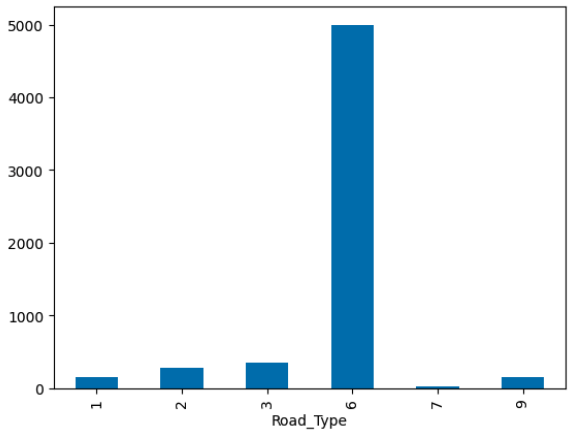


4.1. Análisis Exploratorio

Antes de empezar es importante hacer la aclaración de que en este estudio se están asumiendo muchas cosas de los datos. Es posible que hayan inclinaciones o discordancias porque no sabemos nada de la fuente, ni del lugar, ni de cómo fueron recolectados los datos, ni por quién, ni cuál era su objetivo al tomarlos. De igual forma, la mayoría de las columnas por las que estamos haciendo investigación son categóricas.

Nuestro primer acercamiento a resolver lo que plantea el ejercicio es hacer un análisis exploratorio dadas las columnas que tenemos. Es decir, con base en los datos que tenemos hacemos visualizaciones de algunos subconjuntos de datos para saber si existe alguna tendencia que pueda ser importante en el estudio. En los anexos se pueden ver todas las gráficas, sin embargo, aquí mostramos algunas que nos parecen interesantes.

[3] El otro de los gráficos y demás visualizaciones grandes se encuentran en anexos. 8



Número de accidentes contra tipo de vía

		casualties			
vehicles		1	2	3	4
	1	4700	601	8	2
	2	0	6	0	0

Número de accidentes por número de vehículos[4]

De esta primera aproximación ya podríamos empezar a sacar conclusiones útiles para el negocio, como por ejemplo, que la mayoría de accidentes suceden en vías tipo 6 (es decir, en calzadas), o que normalmente se ve involucrado solo un vehículo, por lo que no debe ser un factor muy influyente en la accidentalidad. Sin embargo, para profundizar nuestro análisis vamos a hacer uso de algoritmos no supervisados de machine learning que nos ayuden a categorizar y encontrar otros patrones.

5. Algoritmo K-means[5]

El funcionamiento de este algoritmo se explica en el notebook, sin embargo, lo que vale la pena resaltar es que se iteran sobre la ubicación de los centroides en un espacio donde estén todos los features a analizar.

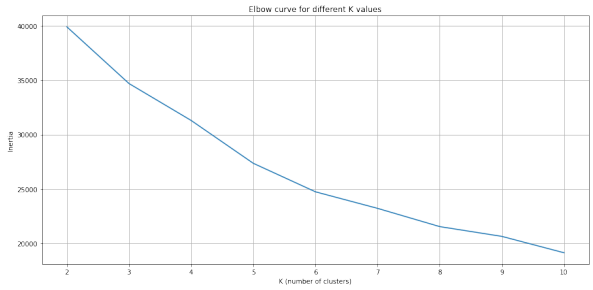
5.1. Validación Cuantitativa

Para la validación cuantitativa de este algoritmo tuvimos varios procesos. En principio, pensamos que definir un número alto de clusters nos iba a dar el punto de codo; sin embargo, caímos en cuenta que entre más clusters hacíamos, efectivamente ibamos a llegar al número de datos del estudio, luego aunque si iban a ser clusters definidos y con datos separados, dada la naturaleza de los datos esto no es posible; por lo que decidimos ir por una cifra mucho menor. Si desea ver las pruebas de los 100 clusters diríjase al Anexo8.

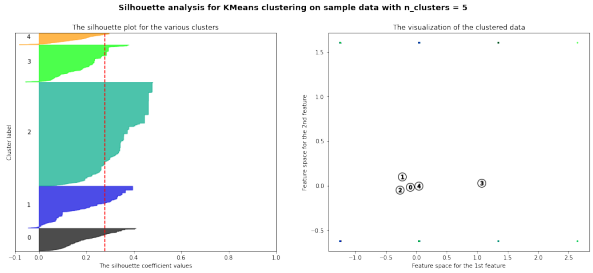
[4] Esta matriz y las demás que hicimos (ver anexos: 8) surgen a partir de la correlación positiva cercana a 1

[5] Realizado por Juan Diego Yepes

En ese orden de ideas, esta es la elbow curve que encontramos después de correr el algoritmo de clustering.



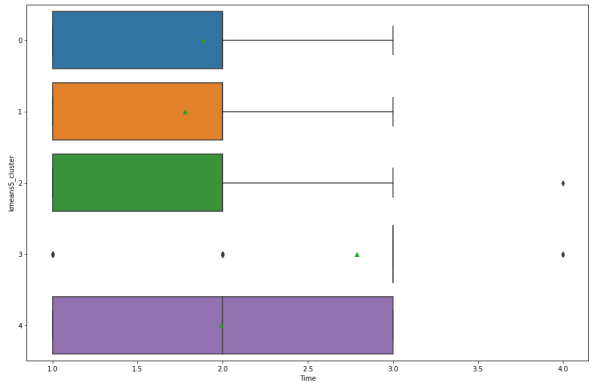
De igual forma, esta es la gráfica de silueta para corroborar lo anterior.

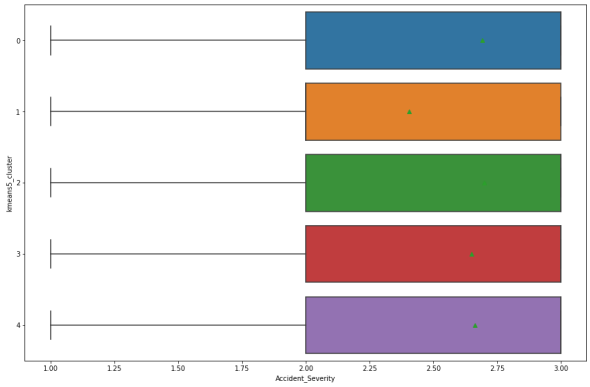


Podemos ver que nuestro codo tiene sentido por la forma de la gráfica.

5.2. Validación Cualitativa

La información más interesante que encontramos está en las columnas de `Time` y `Accident_Severity`. Pudimos encontrar que los clusters mejor distribuidos estaban en dichas features, como lo mostramos a continuación:





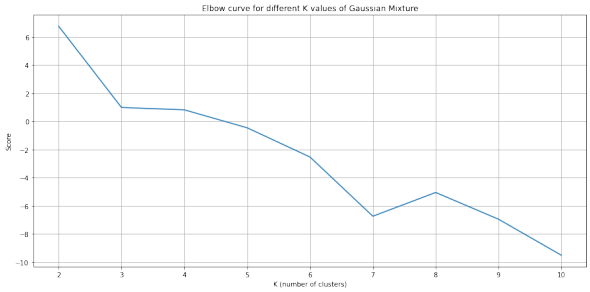
Esto nos puede estar indicando que la cantidad de accidentes, en la mayoría de los clusters suele estar en las horas de la mañana (ya que la mayoría de los clusters se dirigen allí), y que la severidad del accidente es independiente de las condiciones en las que se desarrolle; es por ello que la mayoría de nuestros datos tienden a la derecha.

6. Algoritmo Gaussian Mixture[6]

Decidimos utilizar Gaussian Clustering porque es capaz de producir las probabilidades de pertinencia por instancia a cada uno de los clusters. Esto es importante porque queremos saber los factores más importantes a la hora de causar accidentes.

6.1. Validación Cuantitativa

Para poder saber que numero de clusters usar con el algoritmo, se corre el algoritmo con un numero incremental de clusters y a estos clusters se les saca el valor del "score". Este valor es parecido al valor de la inercia del algoritmo de KMeans pero para Gaussian Mixture. Con esto se puede hacer una grafica de codo para intentar obtener una idea de cual seria un buen numero de nodos a usar en el algoritmo.

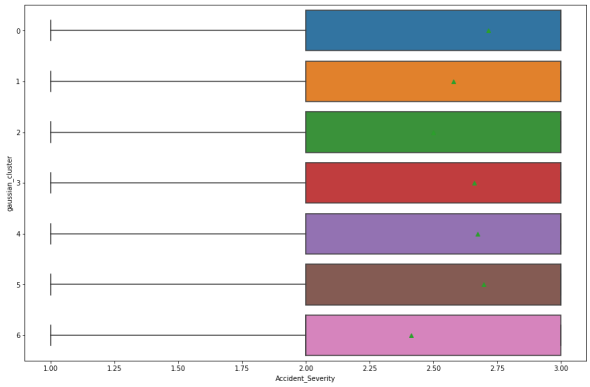


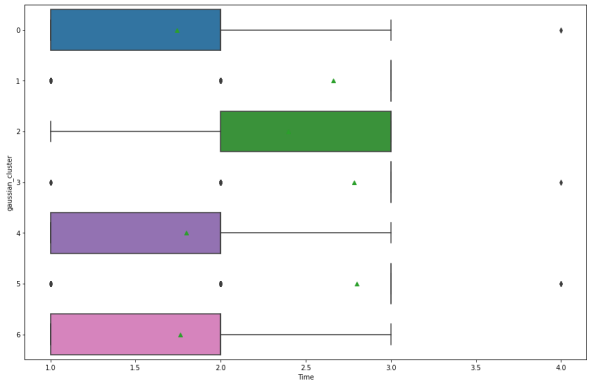
Decidimos que en este caso el codo estaba en 7.

[6]Realizado por Esteban Gonzalez

6.2. Validación Cualitativa

De nuevo, la información más interesante que encontramos está en las columnas de `Time` y `Accident_Severity`.





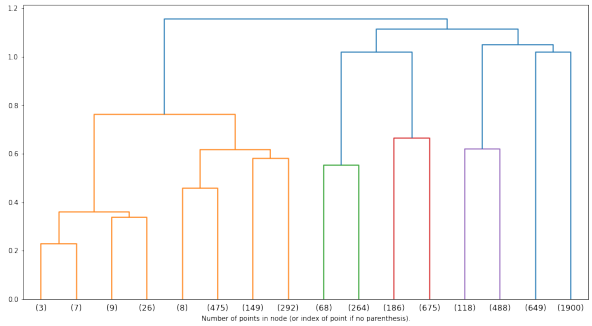
Esto corrobora lo que vimos en el anterior algoritmo.

7. Algoritmo Agglomerative Clustering[7]

7.1. Validación Cuantitativa

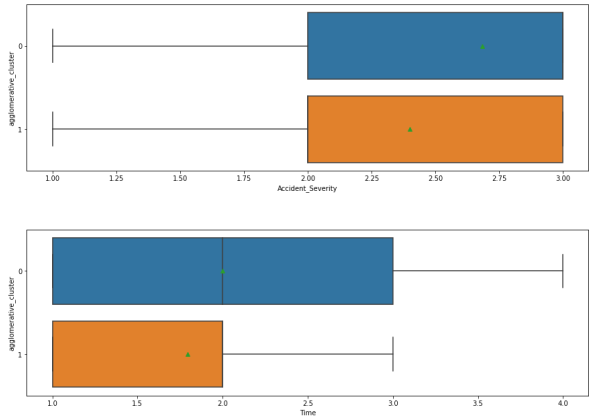
Para este algoritmo no se hace validación cuantitativa puesto que determina el número de clusters. Sin embargo, en el siguiente dendograma se muestra cómo el algoritmo fue escogiendo los clusters:

[7] Realizado por Mariana Díaz



7.2. Validación Cualitativa

Igualmente, la información más interesante que encontramos está en las columnas de `Time` y `Accident_Severity`.



Esto corrobora lo que vimos en el anterior algoritmo.

7.3. Conclusiones

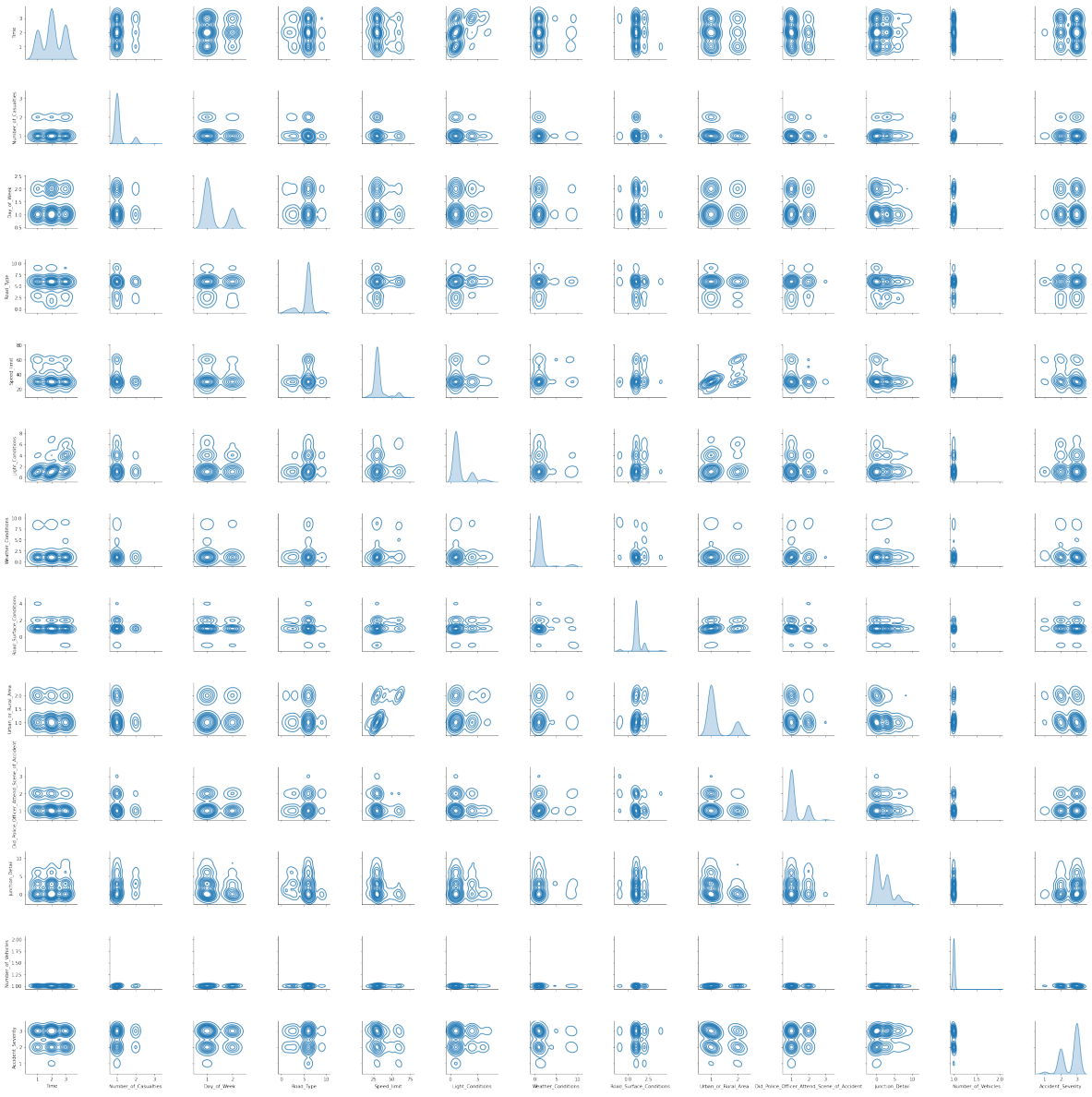
En forma de conclusión general, en este set de datos encontramos que nos estaba haciendo una categorización general más que una específica por clusters. Esto quiere decir que todos los datos son muy parecidos, y el hacer clustering lo único que nos estaba indicando era la medida de tendencia central. Por ejemplo, la mayoría de los accidentes suceden en la mañana, tienen una severidad mayor a 2, suceden entre semana, las condiciones de luz son de día, etc.

Por lo tanto, a la empresa le sirve que le informemos las medidas de tendencia central y estadística (media, mediana, desviación estándar) de sus datos, algo que podemos corroborar con los clústers porque todos los datos apuntan a lo mismo.

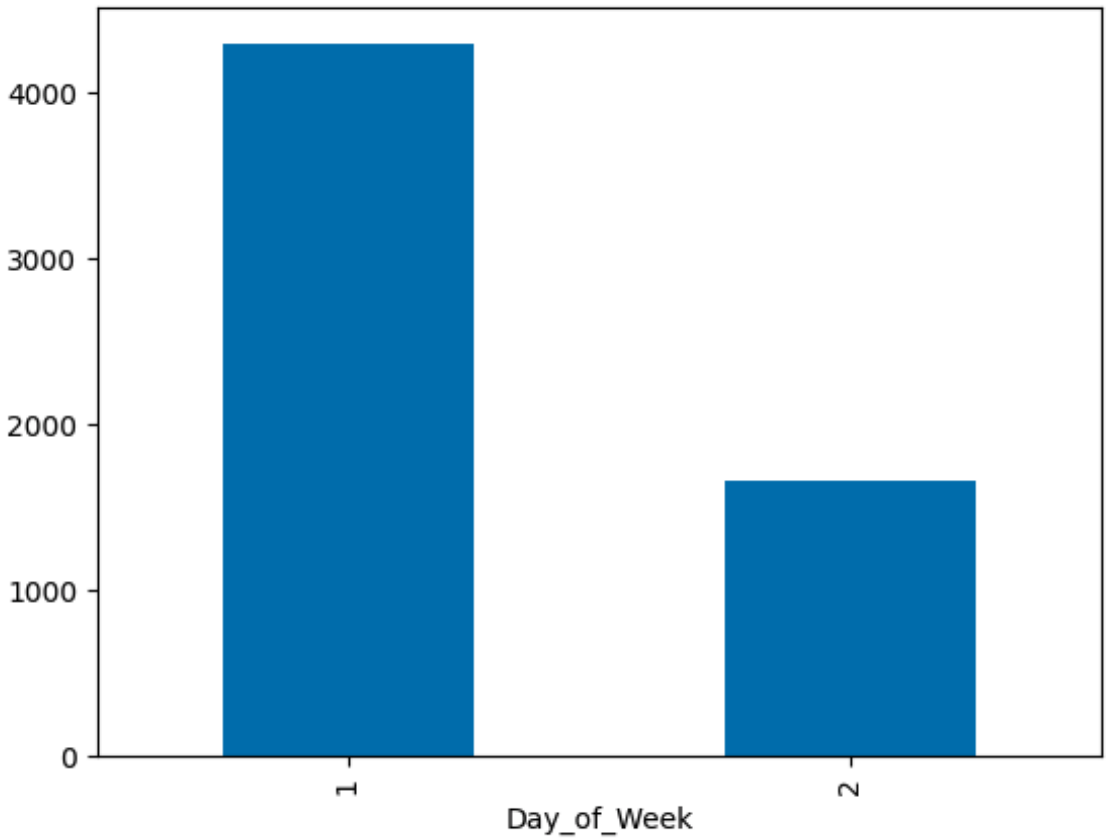
8. Anexos

Esta sección no cuenta dentro de las 8 páginas obligatorias del documento, son gráficas de apoyo que de otra forma había que incluir (Porque no aparecen en el notebook)

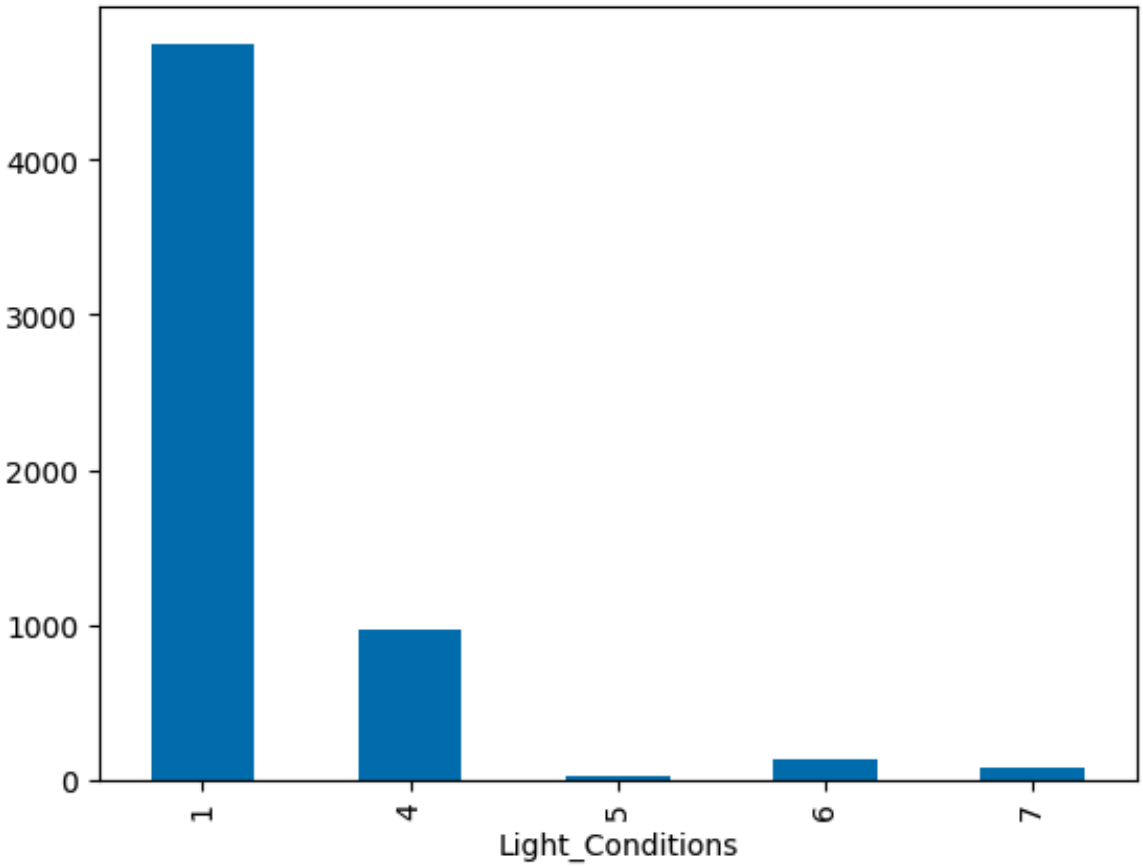
8.1. Correlaciones y análisis exploratorio



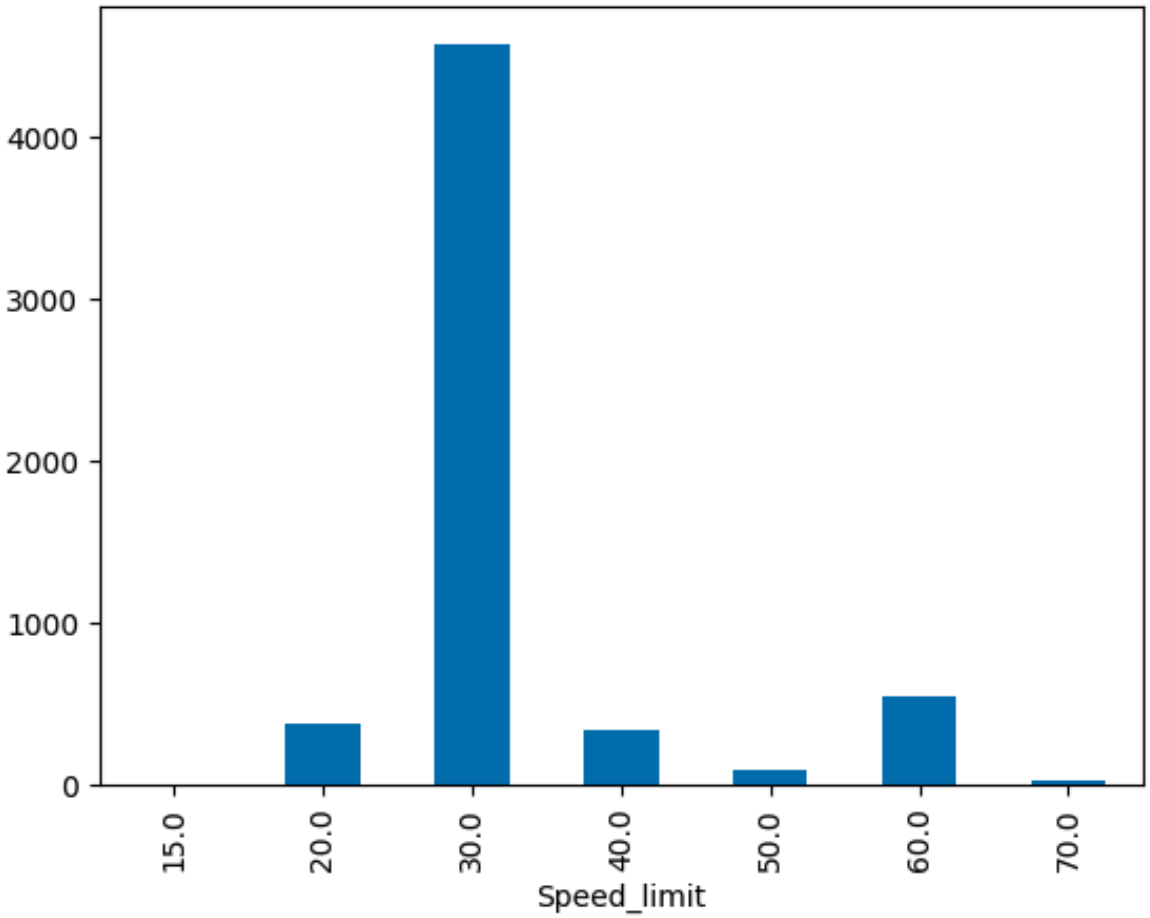
Correlación entre todas las columnas visualizada en curvas de nivel.



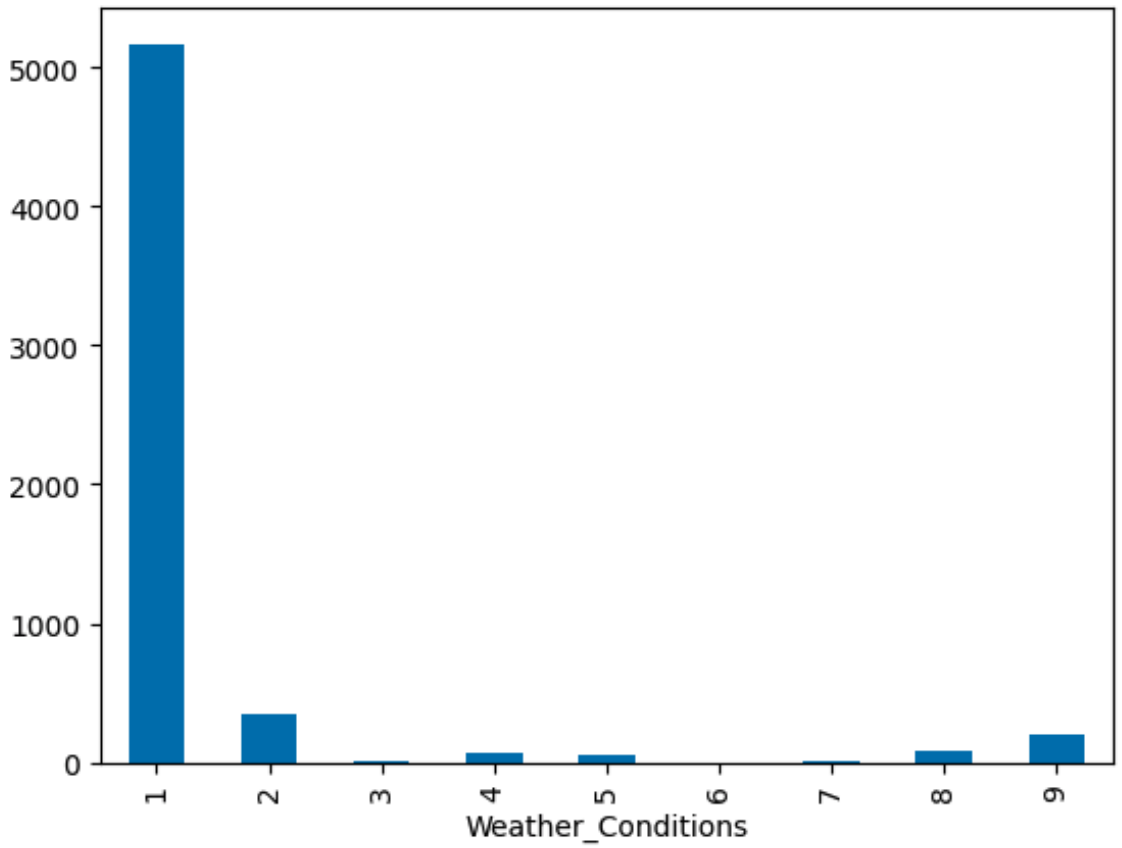
Número de accidentes v.s. día de la semana



Número de accidentes v.s. condiciones de luz



Número de accidentes v.s. límite de velocidad



Número de accidentes v.s. condiciones climáticas

		casualties			
vehicles		1	2	3	4
	1	4700	601	8	2
	2	0	6	0	0

Número de vehículos v.s. accidentes

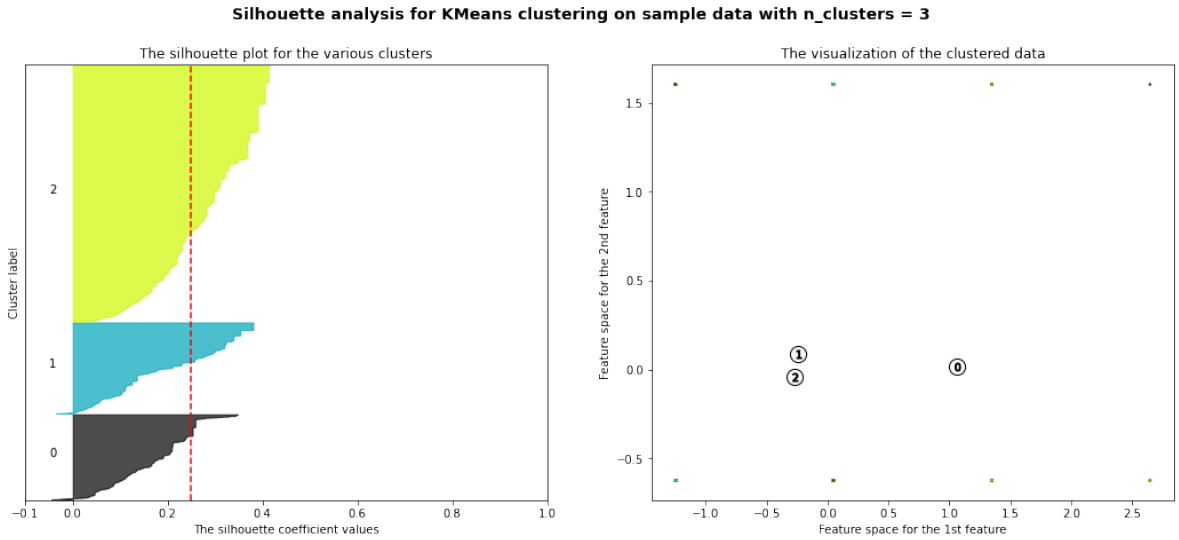
	Day of the week		
Urban/ Rural		1	2
	1	2942	1082
	2	888	405

Lugar urbano o rural v.s. día de la semana

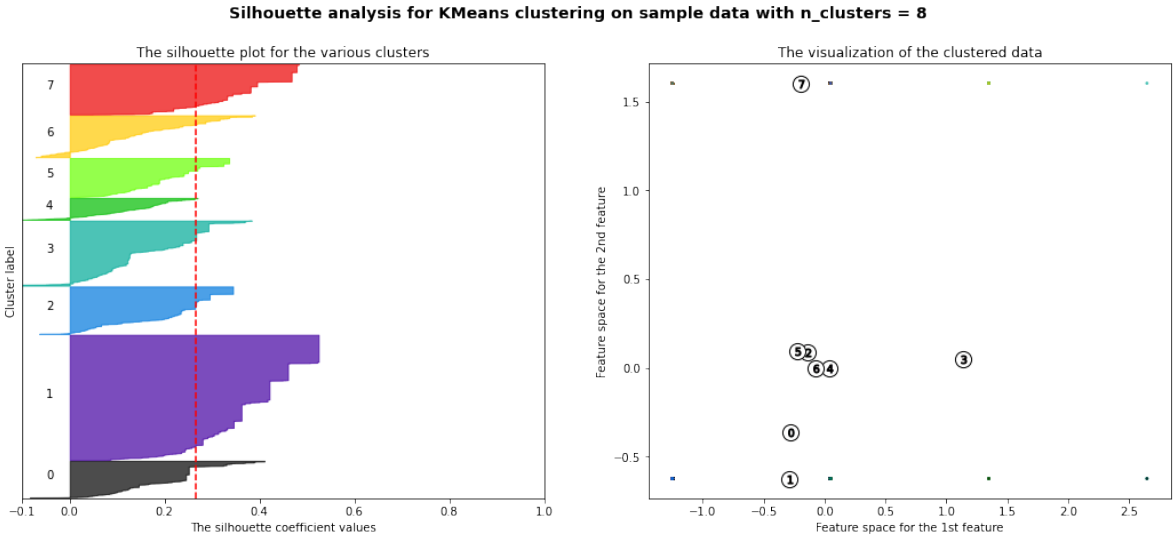
	Did police officer attend scene			
Weather Condition		1	2	3
	1	3592	981	27
	2	246	74	1
	3	9	3	0
	4	46	13	0
	5	39	7	0
	6	1	0	0
	7	6	1	0
	8	56	23	1
	9	49	126	16

Condiciones climáticas v.s. Número de policías que atendieron la emergencia

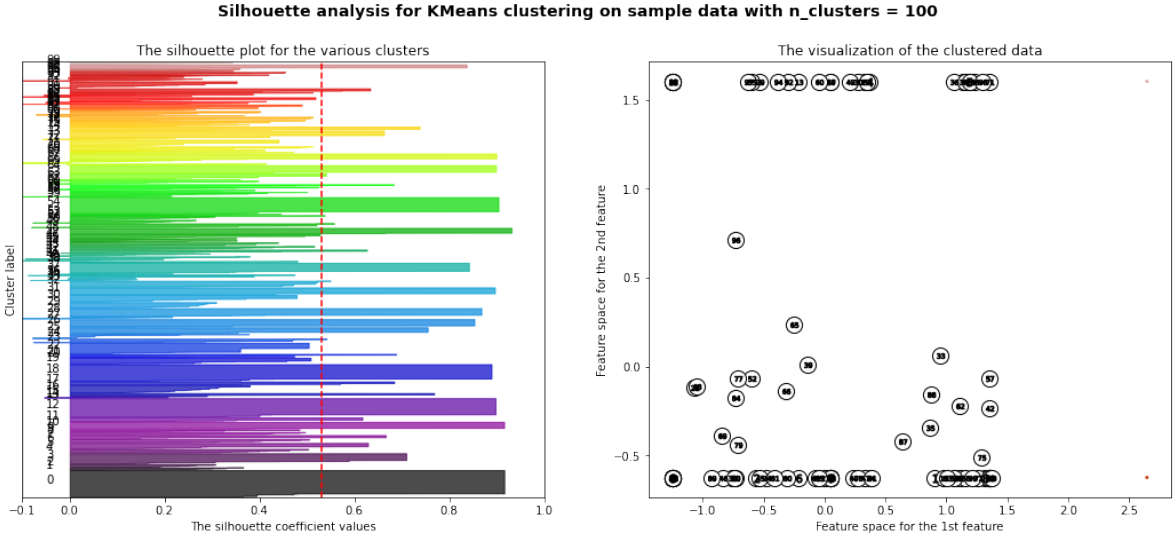
8.2. KMeans



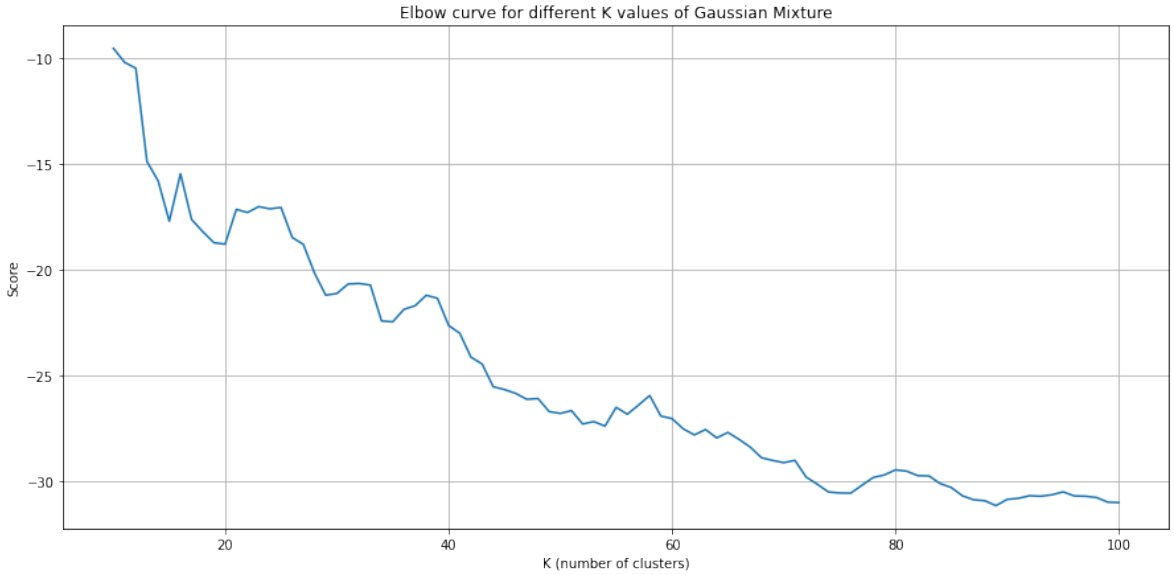
Silueta para 3 clusters de Kmeans



Silueta para 8 clusters de Kmeans

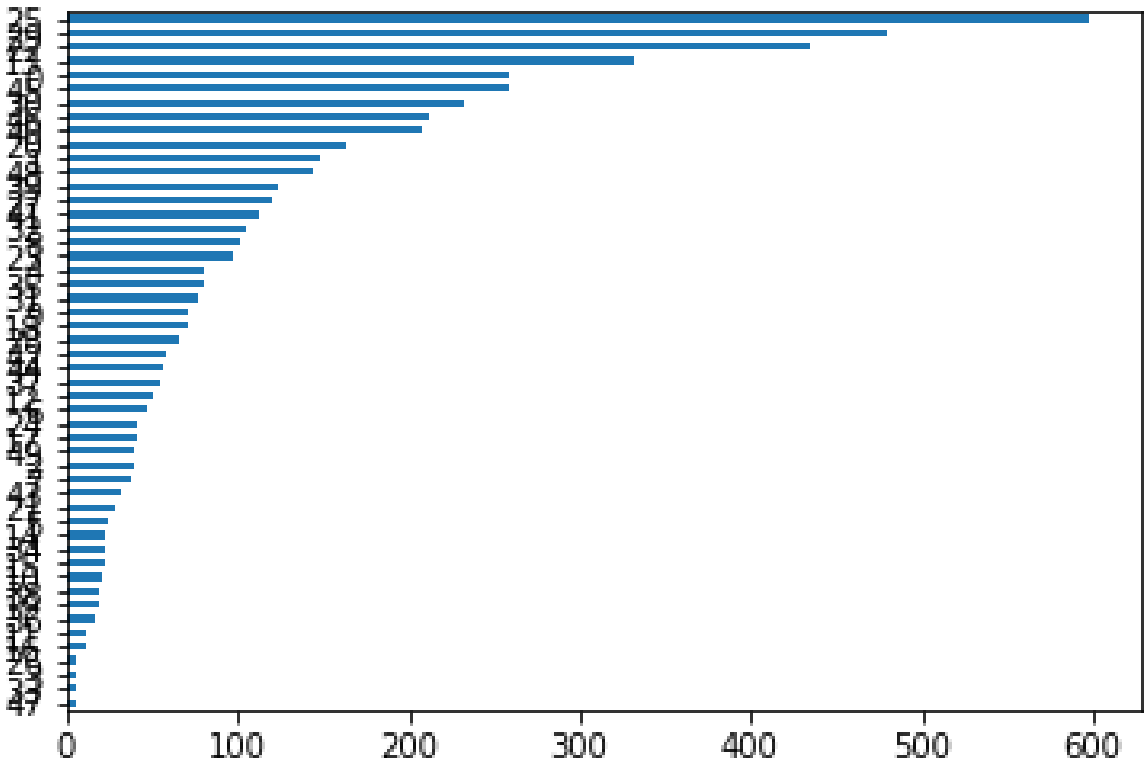


Silueta para 100 clusters de Kmeans

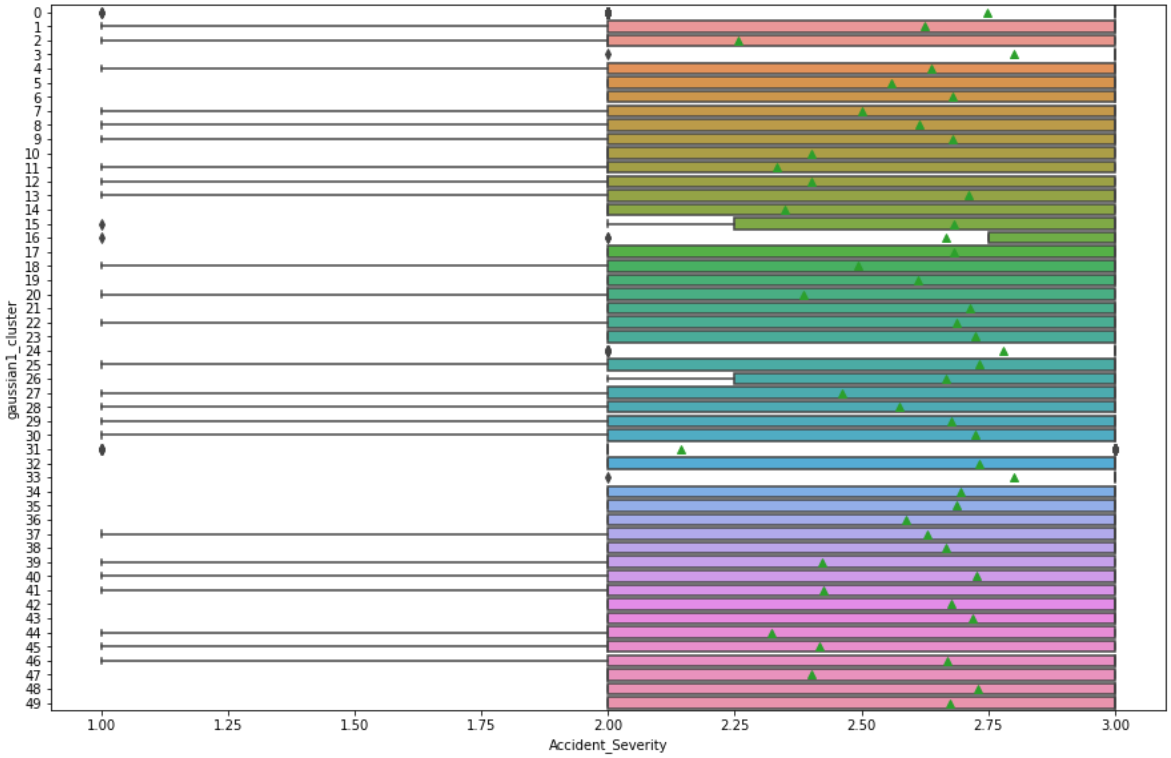


Elbow para 100 clusters de Kmeans

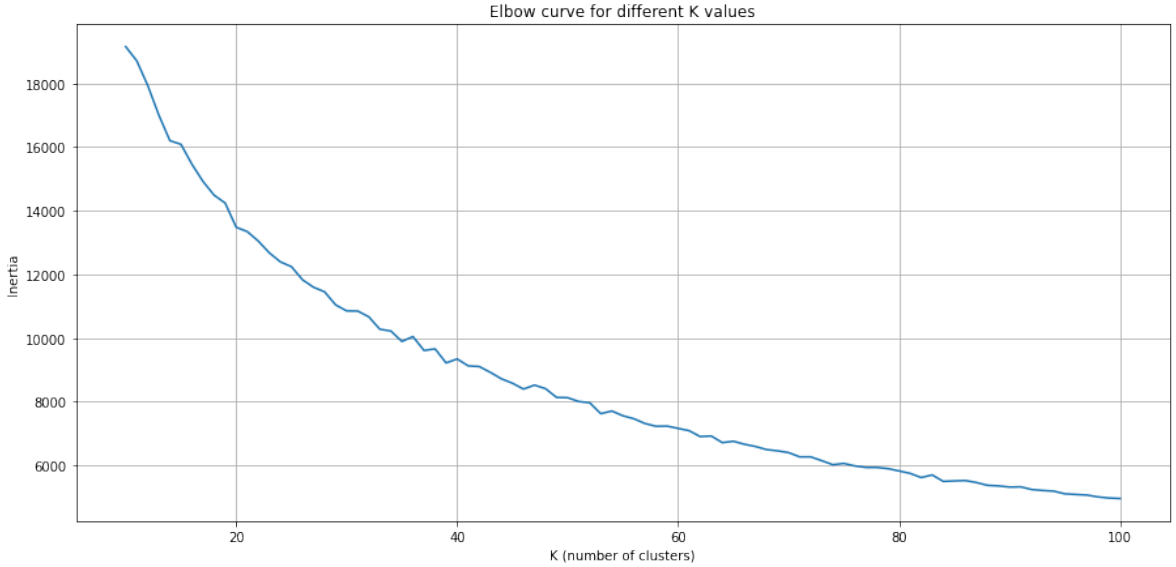
8.3. Gaussian



Distribución de los 100 clusters en Gaussian



Accident Severity de los 100 clusters en Gaussian



Elbow de los 100 clusters en Gaussian