



Actividad Evaluable: Análisis final

Herramientas computacionales: el arte de la analítica (Gpo 570)

- Alumna: Karla Yazmín Trejo Pichardo
- Matricula: A01422942
- Profesor: Gilberto Huesca Juárez

14 de Enero del 2022

Instrucciones de la actividad

1. Con las retroalimentaciones obtenidas en las tareas, haz los cambios pertinentes a tu análisis
2. Prepara una presentación para exponer tus datos, los análisis ejecutados y los resultados obtenidos.
3. Con base en tu presentación, graba un video de máximo 4 minutos, donde describas los resultados y cambios hechos.

Acerca del conjunto

El conjunto elegido habla sobre los distintos registros de accidentes automovilísticos registrados entre 2016 y 2020 en Estados Unidos en 49 estados distintos. Podremos observar distintas variables en torno a estos registros como el clima, la temperatura, donde ocurrieron, entre otros aspectos.

Dicho dataset contiene 499 registros y 47 variables.

El dataset está dentro de las opciones que el profesor dispuso en la plataforma de Canvas, pero podemos encontrarlo formalmente en la siguiente liga: <https://www.kaggle.com/sobhanmoosavi/us-accidents>.

```
*****
Datos generales acerca de la tabla
*****

Cantidad de registros y variables sin depurar "(Con renglones con NaN)"
(499, 47)

Cantidad de registros y variables depurada "(Sin renglones con NaN en cualquiera de las columnas, pero no en todos)"
(19, 47)

Index(['ID', 'Severity', 'Start_Time', 'End_Time', 'Start_Lat', 'Start_Lng',
      'End_Lat', 'End_Lng', 'Distance(mi)', 'Description', 'Number', 'Street',
      'Side', 'City', 'County', 'State', 'Zipcode', 'Country', 'Timezone',
      'Airport_Code', 'Weather_Timestamp', 'Temperature(F)', 'Wind_Chill(F)',
      'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Direction',
      'Wind_Speed(mph)', 'Precipitation(in)', 'Weather_Condition', 'Amenity',
      'Bump', 'Crossing', 'Give_Way', 'Junction', 'No_Exit', 'Railway',
      'Roundabout', 'Station', 'Stop', 'Traffic_Calming', 'Traffic_Signal',
      'Turning_Loop', 'Sunrise_Sunset', 'Civil_Twilight', 'Nautical_Twilight',
      'Astronomical_Twilight'],
      dtype='object')
```

Como podremos observar en la imagen que se encuentra arriba, las 47 variables y el tipo de dato que contienen son los siguientes:

ID	object	Wind_Chill(F)	float64
Severity	int64	Humidity(%)	float64
Start_Time	object	Pressure(in)	float64
End_Time	object	Visibility(mi)	float64
Start_Lat	float64	Wind_Direction	object
Start_Lng	float64	Wind_Speed(mph)	float64
End_Lat	float64	Precipitation(in)	float64
End_Lng	float64	Weather_Condition	object
Distance(mi)	float64	Amenity	bool
Description	object	Bump	bool
Number	float64	Crossing	bool
Street	object	Give_Way	bool
Side	object	Junction	bool
City	object	No_Exit	bool
County	object	Railway	bool
State	object	Roundabout	bool
Zipcode	object	Station	bool
Country	object	Stop	bool
Timezone	object	Traffic_Calming	bool
Airport_Code	object	Traffic_Signal	bool
Weather_Timestamp	object	Turning_Loop	bool
Temperature(F)	float64	Sunrise_Sunset	object
		Civil_Twilight	object
		Nautical_Twilight	object
		Astronomical_Twilight	object
		dtype:	object

Como observamos en el listado, la mayoría de las variables tienen que ver con las condiciones ambientales del lugar donde ocurrió el accidente automovilístico y la mayoría no son valores matemáticos, sino objetos (strings) o valores booleanos, pocas de las variables son numéricas, sin embargo, podemos observar datos muy interesantes de que tenemos disponibles y que en este caso examinaremos dos de ellas más a fondo de lo que posiblemente podremos observar a simple vista.

Acerca de las columnas seleccionadas

Como podremos observar de las dos imágenes que encontraremos a continuación, las columnas seleccionadas fueron las relacionadas a la severidad y la temperatura, de las cuales por el momento solo se obtuvieron los valores únicos, su máximo, mínimo, promedio, mediana y la desviación estándar, tal y como se señala en las imágenes.

```
*****
Valores observables y calculados de columnas
*****

Columna 'Severity'
Valores unicos
[3 2 4]
Máximo
4
Mínimo
2
Promedio
2.691382765531062
Mediana
2.0
Desviación estándar
0.8412948352288262
```

```
Columna 'Temperature(F)'
Valores unicos
[42.1 36.9 36. 39. 37. 35.6 33.8 33.1 32. 35.1 34. 33.4 28. 26.6
 25. 23. 21. 19. 21.2 21.9 19.4 22.5 24.1 30.2 31.8 30. 28.9 30.9
 19.9 15.8 17.1 12.2 14. 15.3 17.6 24.8 18. 16. 7. 12. 15.1 9.
 9.1 16.5 12.9 10. 6.1 8.1 5. 3. nan 22.8 25.2 27. 24.3 25.7
 29.8 32.2 28.4 33. 37.9 32.4 32.7 10.9 43. 39.9 51.1 46.9 48. 37.4
 53.1 64. 50. 48.9 61. 53.6 57. 61.7 54. 60.1 64.9 52. 66.9 57.9
 63. 55.9 69.1 68. 34.9 26.2 34.2 46. 44.1 26.1 39.2 55.4 55. 40.1
 46.2 41. 42.8 32.9 46.4 45. 32.5 31.3 31.6]
Máximo
69.1
Mínimo
3.0
Promedio
32.005050505050505
Mediana
31.8
Desviación estándar
13.366658727366437
```

Primero haremos un análisis de la severidad, la cual indica la severidad que tuvo el accidente con valores que en teoría deberían ir del 1 al 4, siendo 1 siendo el menos severo o menor gravedad y 4 el de mayor gravedad o más severo. A pesar de está teoría, vemos que sus valores únicos son el 2, 3 y 4, siendo el mínimo 2 y el máximo 4, indicándonos que dentro de los registros, los accidentes tienen siempre un grado de gravedad considerable, siendo que el vehículo o las personas probablemente sufren un daño relevante siempre que ocurre un accidente.

```
Columna 'Severity'
Valores unicos
[3 2 4]
Máximo
4
Mínimo
2
```

Por otro lado, cuando observamos el promedio y la mediana, podemos ver que estos valores tienen una importante diferencia considerando los rangos de esta variable, siendo que el promedio es de 2.69 aproximadamente y la mediana de 2, lo que nos diría que encontraremos solamente valores de 2 en al menos la mitad de los registros, pero el resto se distribuyen en posiblemente en muchos accidentes de grado de severidad 4 y algunos en 3 para obtener un promedio que es bastante cercano al 3 y una mediana de 2, lo que nos confirma la desviación estándar, que muestra que es de .84 aproximadamente, hablándonos que la distribución de datos deben ser más cercanos al 2 y al 4 que al 3 (valor cercano al promedio). Por lo tanto podemos decir que los accidentes podemos decir que por lo general no serán tan severos, pero si no son tan severos, lo más posible es que hayan sido bastante graves.

```
Promedio
2.691382765531062
Mediana
2.0
Desviación estándar
0.8412948352288262
```

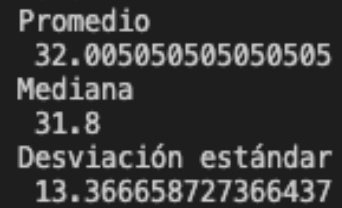
Ahora hablando de la temperatura, podemos observar una gran cantidad de valores únicos, lo que tiene sentido, ya que los registros ocurren en muchos estados y en distintas épocas del año, por lo que la variación de la temperatura es algo común.

```
Valores unicos
[42.1 36.9 36. 39. 37. 35.6 33.8 33.1 32. 35.1 34. 33.4 28. 26.6
25. 23. 21. 19. 21.2 21.9 19.4 22.5 24.1 30.2 31.8 30. 28.9 30.9
19.9 15.8 17.1 12.2 14. 15.3 17.6 24.8 18. 16. 7. 12. 15.1 9.
9.1 16.5 12.9 10. 6.1 8.1 5. 3. nan 22.8 25.2 27. 24.3 25.7
29.8 32.2 28.4 33. 37.9 32.4 32.7 10.9 43. 39.9 51.1 46.9 48. 37.4
53.1 64. 50. 48.9 61. 53.6 57. 61.7 54. 60.1 64.9 52. 66.9 57.9
63. 55.9 69.1 68. 34.9 26.2 34.2 46. 44.1 26.1 39.2 55.4 55. 40.1
46.2 41. 42.8 32.9 46.4 45. 32.5 31.3 31.6]
```

Por otro lado, vemos que los registros muestran un máximo de 69.1 y un mínimo 3.0 en grados fahrenheit, los cuales equivalen a 20.61 y un -16.11 respectivamente, lo cual confirma que la variabilidad de la temperatura es grande probablemente por ser registros en distintas épocas y estados.

```
Máximo
69.1
Mínimo
3.0
```

Ahora viendo otra parte muy interesante de la temperatura, podremos observar el promedio, la mediana y la desviación estándar, las cuales se calcularon en 32 aproximadamente, en 31.8 y en 13.37 respectivamente, como podemos observar, a diferencia de la severidad, el promedio y la



Promedio	32.0050505050505
Mediana	31.8
Desviación estándar	13.36658727366437

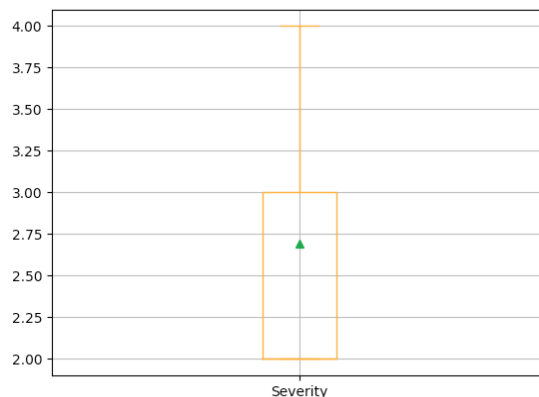
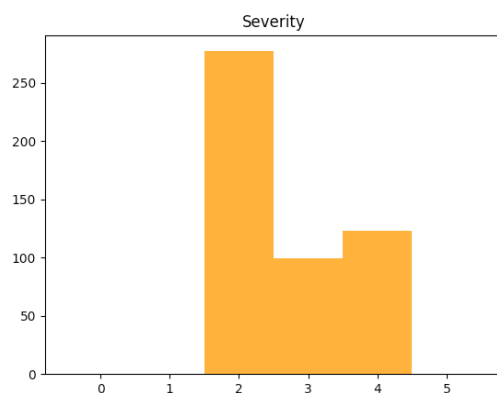
mediana son bastante cercanos, siendo una diferencia muy pequeña, y la desviación estándar es entendible (considerando que la diferencia entre el máximo y mínimo es de 85 aproximadamente) por las mismas razones que el máximo y mínimo tenían sentido, que es debido a que las temperaturas varían bastante por época y lugar, pero lo que es interesante es que el promedio y mediana son temperaturas muy frías (0 en celsius), lo cual nos diría que la mayoría de accidentes suceden en épocas o lugares fríos, y por la desviación estándar podríamos ver que es así, pues muestra que al menos la mayoría de accidentes son en épocas frías (lo cual tal vez en México no lo vemos tan claro, pero en Estados Unidos debido al hielo y la nieve, esto tiene sentido).

Analizando los datos de las columnas seleccionadas

En esta actividad tuvimos la oportunidad de revisar con mayor profundidad el comportamiento de los datos de las columnas que nos interesaron, en mi caso me gusto mucho lo que aprendí de las columnas anteriores y me siento intrigada de confirmar las suposiciones que había hecho con respecto al comportamiento de las columnas de severidad y temperatura (las trabajadas anteriormente).

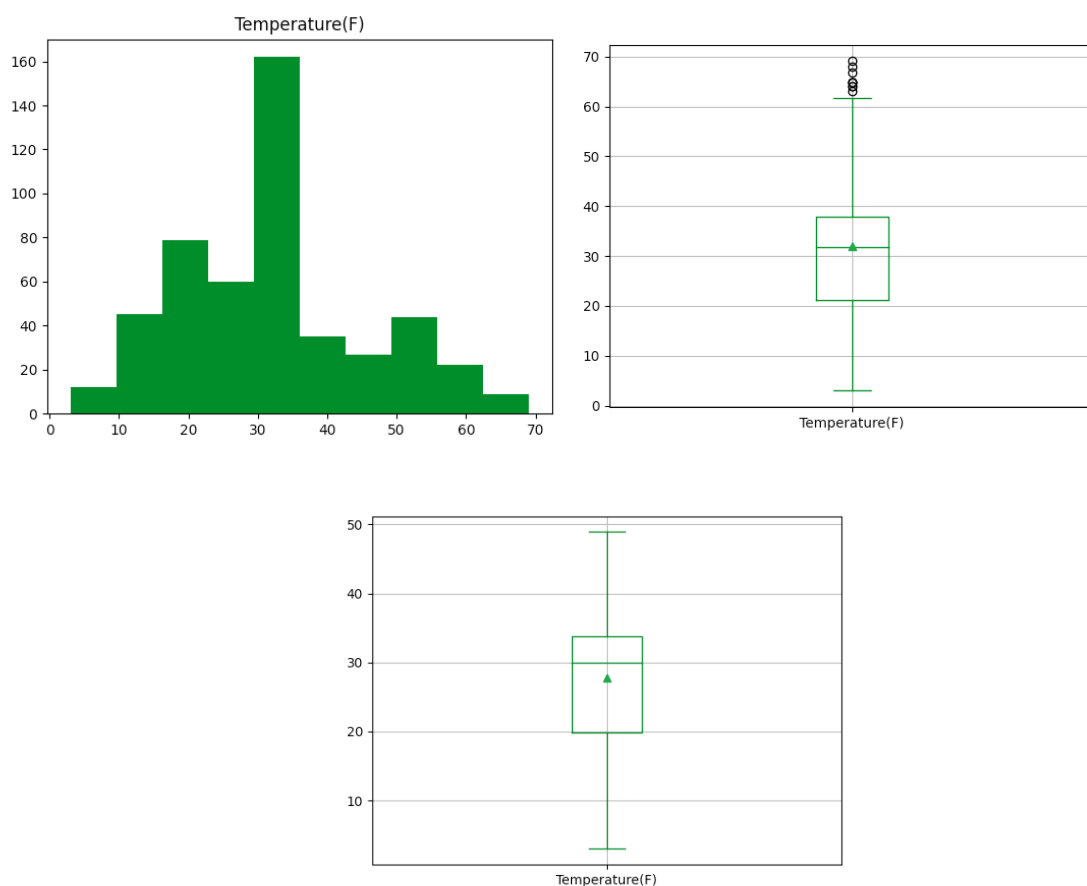
Además de revisar la correlación existente entre todas las variables del dataset, algo muy interesante, pues es normal que cuando ocurre un accidente de este tipo asociamos algunos elementos y me gustaría saber si los datos pueden confirmar este tipo de cosas son solo suposiciones sin fundamento o ciertamente están relacionadas.

Estas imágenes que observemos a continuación son de la columna de severidad, como habríamos predicho anteriormente y se observa claramente en el histograma, la mayoría de accidentes son de grado 2, después de grado 4, el más severo, y por último de grado 3, aunque había pensado que serían muchos más de tres, esto sigue explicando por que el promedio era mucho más elevado que la mediana.

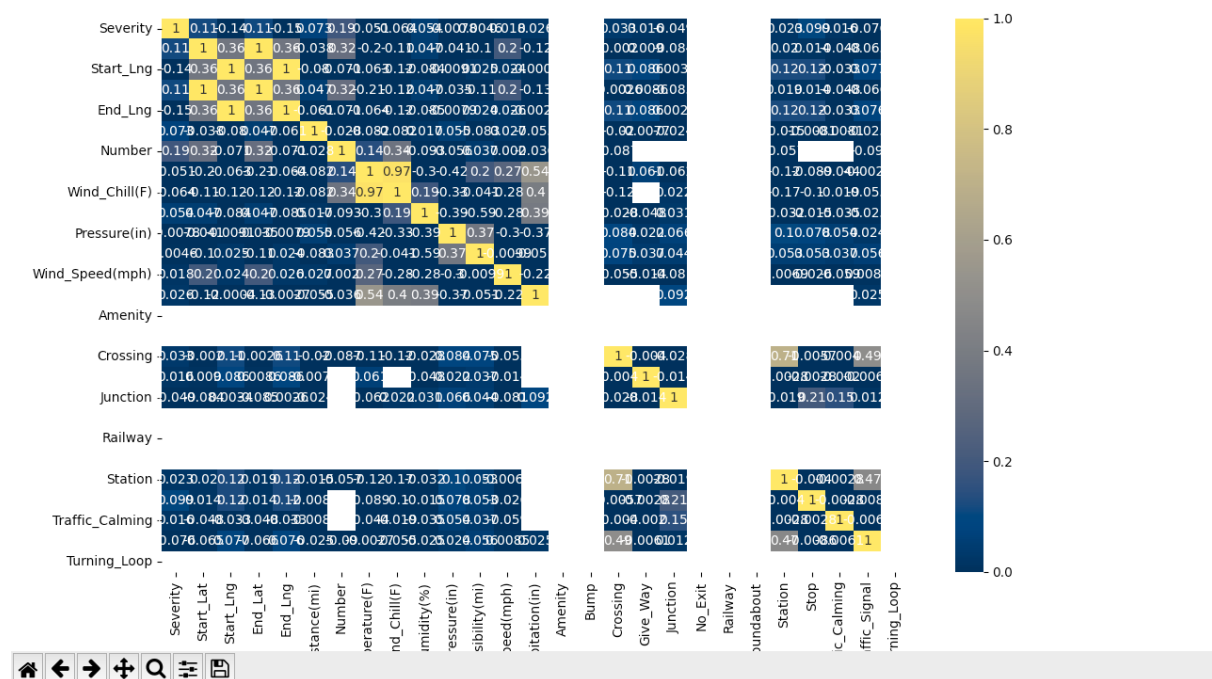


Por otro lado, observamos que en el diagrama de caja y bigotes que la distribución de los datos es bastante cercana y sin outliers, probablemente debido a que el rango de datos es bastante pequeño y bien definida, por lo que la variación no es tan grande, permitiendo que a pesar de que puede ser algo extremo, siendo la caja muy cargada en el valor 2 (y con justa razón), siga siendo aceptable, y podemos apreciar fácilmente que el promedio es cercano al 3 (siendo 2.69 aproximadamente).

Ahora observando los diagramas y gráficas de la columna de temperatura a continuación podemos apreciar que el histograma es bastante variado y extenso, lo cual tiene sentido, pues la temperatura a diferencia de la severidad es muy cambiante, con muchas variaciones debido a las distintas estaciones del año y lugares donde se de el accidente, sin embargo vemos que un valor muy elevado y que sobresale con más registros que todos es en la temperatura de aproximadamente 32 grados fahrenheit (0 grados celsius aproximadamente), además de que valores cercanos a este, especialmente los que son temperaturas más frías, demuestran tener más registros que las temperaturas más calientes, lo cual confirma la sospecha que los accidente ocurren más en lugares o épocas donde hace mucho frío, probablemente debido a los problemas asociados con la nieve, hielo o ventiscas.



Ahora observando el diagrama de caja y bigotes de arriba, en la primera ocasión que lo hice con todos los datos podemos apreciar muchos outliers, y después de delimitar la temperatura a menores a 50 fahrenheit (la máxima temperatura es 69.1) vemos que no hay más outliers y de hecho el diagrama se ve bastante bien y sin muchos cambios en el promedio, quedando este de 30, lo cual confirma nuevamente que hay muchos más registros en épocas frías y muy pocos en épocas calurosa.

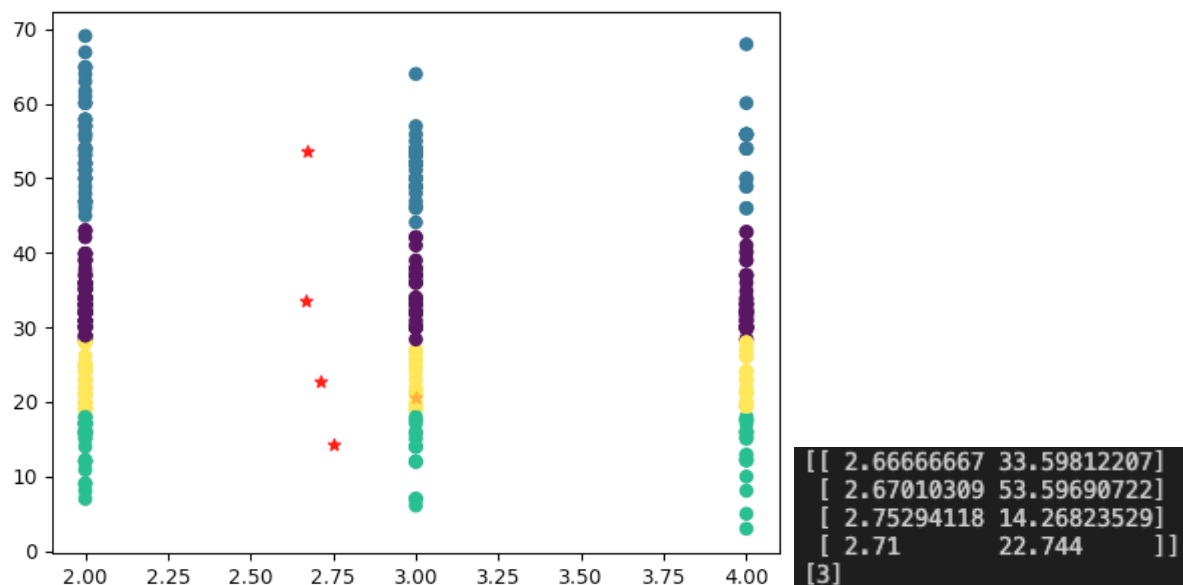


Finalmente con el mapa de calor que se muestra justo arriba, vemos que son muy pocas las observaciones que podemos hacer de variables altamente relacionadas con otra, algunas como las de latitud o longitud están muy relacionadas, pero creo que no son algo tan interesante de observar. En el caso de las variables que hemos analizado, vemos que su correlación es negativa y baja, siendo de -0.051 aproximadamente, así que se puede decir muy poco que cuando la temperatura es elevada, la severidad del accidente será más baja por ejemplo, pero están muy poco relacionadas para poder decir que esto es siempre cierto o la gran mayoría de veces, así pasa con prácticamente todas las variables, realmente no note ninguna que quisiera destacar en esta ocasión.

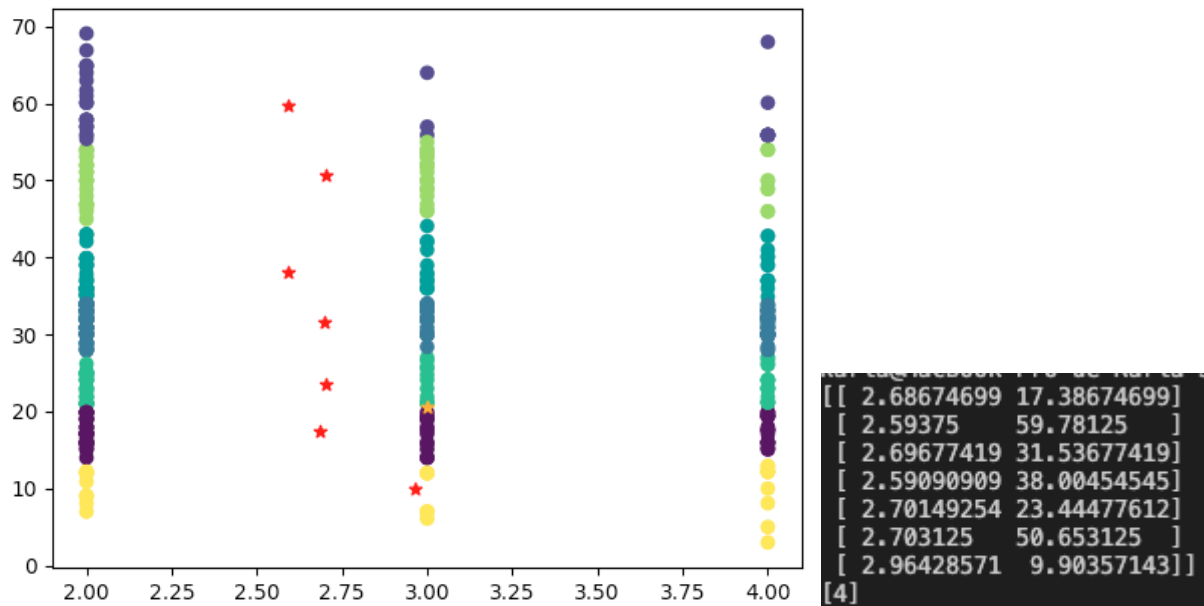
Analizando con Algoritmo K-medias.

Como lo vimos durante la clase y de lo que nos centraremos es en el uso del algoritmo k-means o k-medias para analizar las variables que encontramos en nuestro dataset y algunos comportamientos interesantes que podemos llegar a encontrar en ellos cuando relacionamos las dos variables de severidad y temperatura, a pesar de que en el reporte de la actividad 3 la correlación nos mostró que estas dos variables, al contrario de mi hipótesis inicial de que ambas tendrían que estar fuertemente relacionadas, no estaban muy correlacionadas, y en general prácticamente ninguna variable estaba muy relacionada con otra, pero creo que si se usa el algoritmo de k-medias, pueda llegar a ver algo que la correlación no me mostró.

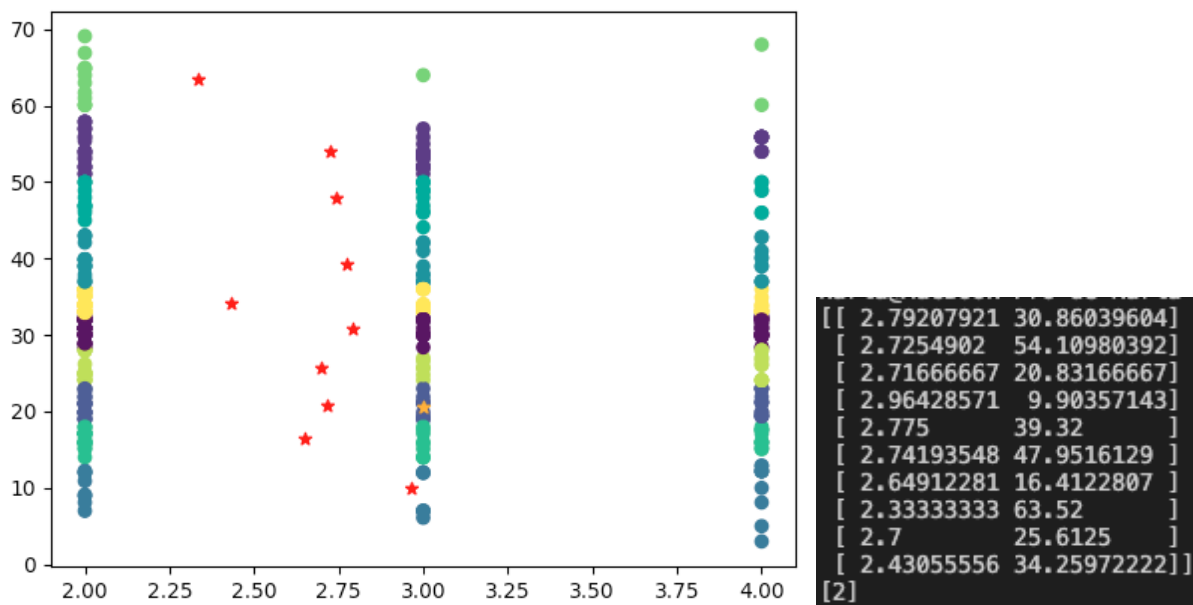
A continuación mostraré el resultado del algoritmo de k-medias haciendo uso de tres valores de k para analizar ambas columnas en diferentes perspectivas.



Estos son los resultados usando una k de 4



Estos son los resultados usando una k de 7.



Finalmente, estos últimos son los resultados usando una k de 10.

Cabe mencionar que el dato del cual predecimos su grupo (la estrella amarilla en todas las gráficas), tiene una severidad de 3 y una temperatura de 20.5 en todas las gráficas, además de que me gustaría mencionar que realmente el valor de k no fue algo aleatorio, elegí esos valores basándose en el tamaño del dataset y especialmente en el rango máximo y mínimo de la temperatura, quería probar, especialmente con el último resultado, si había algo que pudiera observar cuando se hacían separaciones más claras entre severidad y la temperatura.

En este caso y después de los comentarios realizados por el profesor en la entrega de la actividad 4, a pesar de que había pensado que era posible que haciendo uso de este algoritmo cambiaría la idea de que severidad y temperatura no estaban relacionadas, pues realmente creía posible mi hipótesis que tal vez era por que las la mezcla de todas las temperaturas o tal vez analizando de manera más particular detalles muy particulares podría ayudar con esto, pero no fue así, pues tanto la correlación como la dispersión de los datos en las gráficas que se muestran relacionadas al algoritmo k-medias muestran contundentemente que no hay manera que estas dos variables tengan una relación o comportamiento del cual podamos podamos decir que una de las variables determina o nos puede ayudar a predecir el valor de la otra, por lo tanto la hipótesis de que estaban relacionadas queda totalmente descartada en este caso.

Conclusiones.

En un comienzo estaba bastante emocionada con este dataset porque en general me gusta el area automotriz de una forma muy propia de ingeniería (como funcionan ciertas partes del coche o la explicación misma de las mecánicas que se realizan al momento de manejar y que por lo general no se le da tanta importancia a comprender con detalle) y quería probar si algo que había pensado como algo muy “lógico” a llegar podría ser probado o desmentido con el comportamiento de los datos, en mi caso quería probar que la temperatura y severidad eran cosas relacionadas y que una podría ser muy determinante de la otra.

Conforme iba analizando fui descubriendo cosas interesantes de ambas variables, pero cuando observe el mapa de calor fue sorprendente observar que estas variables estaban prácticamente nada correlacionadas y fue la inexperiencia y la falta de conocimiento por mi parte el creer que en este último algoritmo podría encontrar algo que me ayudará a decir que mi hipótesis era válida, pero ya fue después que después de hacermelo notar y yo misma comprender el porqué de tal afirmación con la gráfica, sin embargo no puedo decir que no obtuve resultados de todo esto, a pesar de ser inesperados.

A pesar de que mi hipótesis fue rechazada, me di cuenta que los datos y este tipo de análisis nos habrían los ojos y nos hacían darnos de cosas que perdemos a simple vista y que pueden ayudarnos a comprender el por que de muchas cosas o que pensamos que son cosas sin importancia, como yo asumiendo que la variable de start_lng y end_lng a pesar de estar fuertemente relacionadas no les tome importancia, pero descubrí que pueden ayudarnos a comprender cosas relacionadas a las ubicaciones de los accidentes y descubrir cosas que van más allá de un pensamiento muy simple. Tal vez no encontré una respuesta que yo esperaré, pero ciertamente aprendí que los datos y su análisis pueden ayudar a no solo fundamentar el por que de un comportamiento, si no también ayudar a descubrir y aprender cosas que uno ni siquiera imagina que podrían ser determinantes para aquello que estamos viendo o analizando.