Actividad Evaluable: Obtención de estadísticas descriptivas

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

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

11 de Enero del 2022

Instrucciones de la actividad

En esta actividad, trabajarás con el conjunto de datos que seleccionaste. Tienes que replicar los pasos vistos en clase.

Realiza un código que haga lo siguiente:

1. Cargar los datos.
2. Mostrar el número de variables y el número de registros.
3. Mostrar el nombre de las columnas.
4. Mostrar los tipos de datos.
5. Selecciones dos columnas que al momento parezcan interesantes. Para ellas:
    1. Mostrar los valores únicos
    2. Mostrar los valores máximos y mínimos
    3. Mostrar la media, la mediana y la desviación estándar.

Realiza un pequeño reporte que muestre lo siguiente:

1. Una introducción al conjunto de datos ¿Qué es? ¿De dónde se obtuvo? ¿Qué representa?
2. Cantidad de datos que tienes, las variables que contiene cada vector de datos y el tipo de variables.
3. Para las dos variables que escogiste:
    1. Los rangos de las variables que escogiste
    2. Basándote en la media, mediana y desviación estándar de cada variable, ¿qué conclusiones o asunciones puedes obtener de los datos? ¿Parecen muy dispersos? ¿Alcanzas a vislumbrar algún patrón? ¿Parecen relacionados?

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 predispuso 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 depurarada "(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:

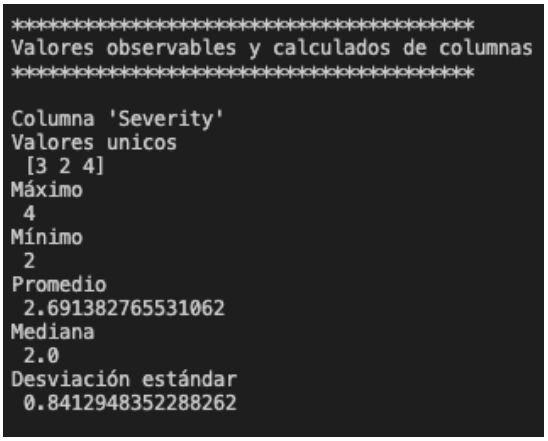
```
ID                      object
Severity                 int64
Start_Time              object
End_Time                object
Start_Lat              float64
Start_Lng              float64
End_Lat                float64
End_Lng                float64
Distance(mi)           float64
Description             object
Number                 float64
Street                  object
Side                    object
City                    object
County                  object
State                   object
Zipcode                 object
Country                 object
Timezone                object
Airport_Code            object
Weather_Timestamp       object
Temperature(F)         float64
```

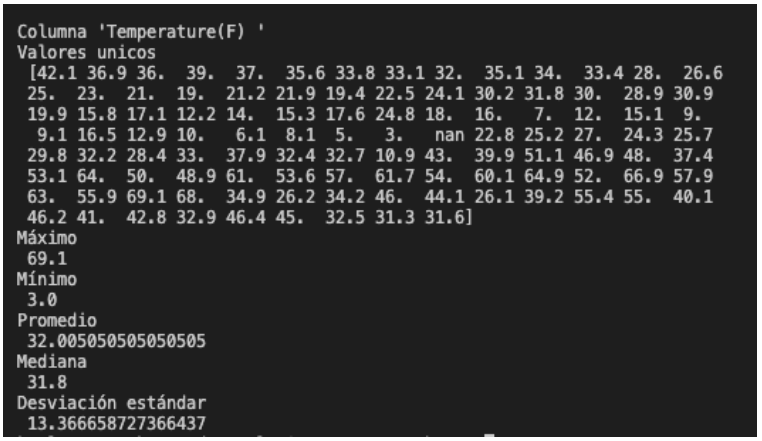
```
Wind_Chill(F)          float64
Humidity(%)            float64
Pressure(in)           float64
Visibility(mi)         float64
Wind_Direction          object
Wind_Speed(mph)        float64
Precipitation(in)      float64
Weather_Condition       object
Amenity                   bool
Bump                      bool
Crossing                  bool
Give_Way                  bool
Junction                  bool
No_Exit                   bool
Railway                   bool
Roundabout                bool
Station                   bool
Stop                      bool
Traffic_Calming           bool
Traffic_Signal            bool
Turning_Loop              bool
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.

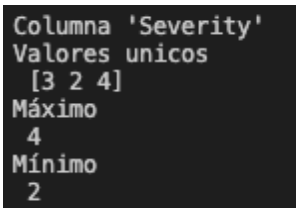
Acerca de las columnas seleccionadas

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.

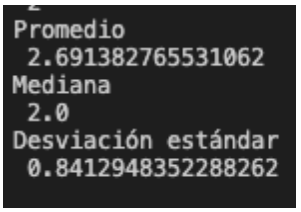




Primero haremos un análisis de la severidad, la cual indica la severidad que tuvo el accidente con valores que en teoría deberían ir del 1 al 4, siendo 1 siendo el menos severo o menor gravedad y 4 el de mayor gravedad o más severo. A pesar de está teoría, vemos que sus valores únicos son el 2, 3 y 4, 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.



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.



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 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).

```
Promedio
 32.005050505050505
Mediana
 31.8
Desviación estándar
 13.366658727366437
```

Actividad Evaluable: Mapas de calor y boxplots

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

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

12 de Enero del 2022

Instrucciones de la actividad

En esta actividad, trabajarás con el conjunto de datos que seleccionaste. Tienes que replicar los pasos vistos durante la clase.

Realiza un código que haga lo siguiente:

1. Cargar los datos.
2. Selecciones dos columnas que al momento parezcan interesantes. Para éstas:
    1. Desplegar su histograma.
    2. Desplegar el diagrama de cajas y bigotes.
    3. Si hay outliers, quítalos y vuelve a desplegar el diagrama con los nuevos datos.
3. Despliega el mapa de calor de las correlaciones entre todas las variables numéricas. Escoge la visualización más adecuada y que aporte la mayor información posible para el análisis.

Realiza un pequeño reporte que muestre lo siguiente:

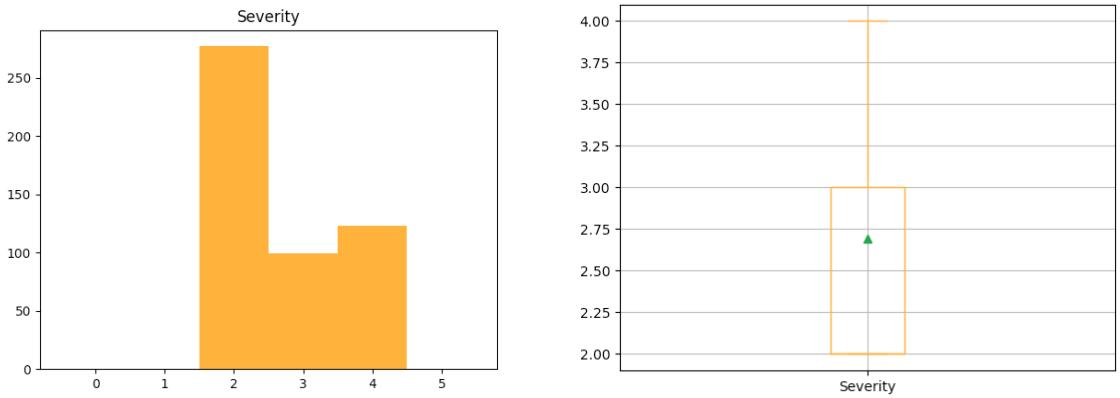
1. Describa el significado de las variables que seleccionaste.
2. Muestre su histograma y su diagrama de cajas y bigotes.
3. Muestre el mapa de calor de las correlaciones entre todas las variables numéricas.

Puedes agregar estos elementos al reporte anterior para ir mostrando el proceso completo.

Analizando los datos de las columnas seleccionadas

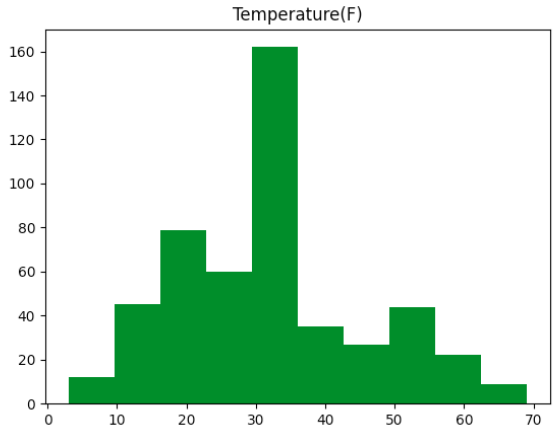
En esta actividad tuvimos la oportunidad de revisar con mayor profundidad el comportamiento de los datos de las columnas que nos interesaron, en mi caso me gusto mucho lo que aprendí de las columnas anteriores y me siento intrigada de confirmar las suposiciones que había hecho con respecto al comportamiento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

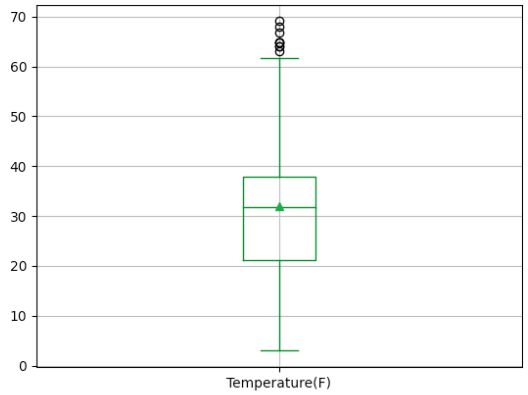
Además de revisar la correlación existente entre todas las variables del dataset, algo muy interesante, pues es normal que cuando ocurre un accidente de este tipo asociamos algunos elementos y me gustaría saber si los datos pueden confirmar este tipo de cosas son solo suposiciones sin fundamento o ciertamente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

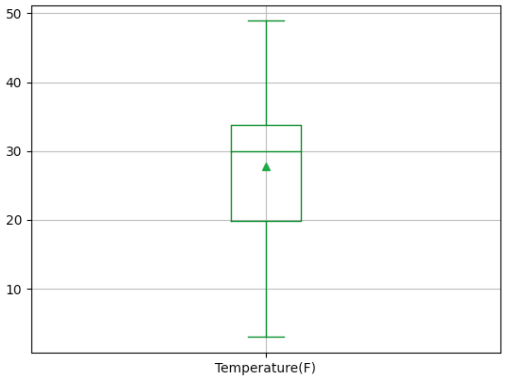


Estas imágenes que observamos arriba son de la columna de severidad, como habríamos predicho anteriormente y se observa claramente en el histograma, la mayoría de accidentes son de grado 2, después de grado 4, el más severo, y por último de grado 3, aunque había pensado que serían muchos más de tres, 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 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).

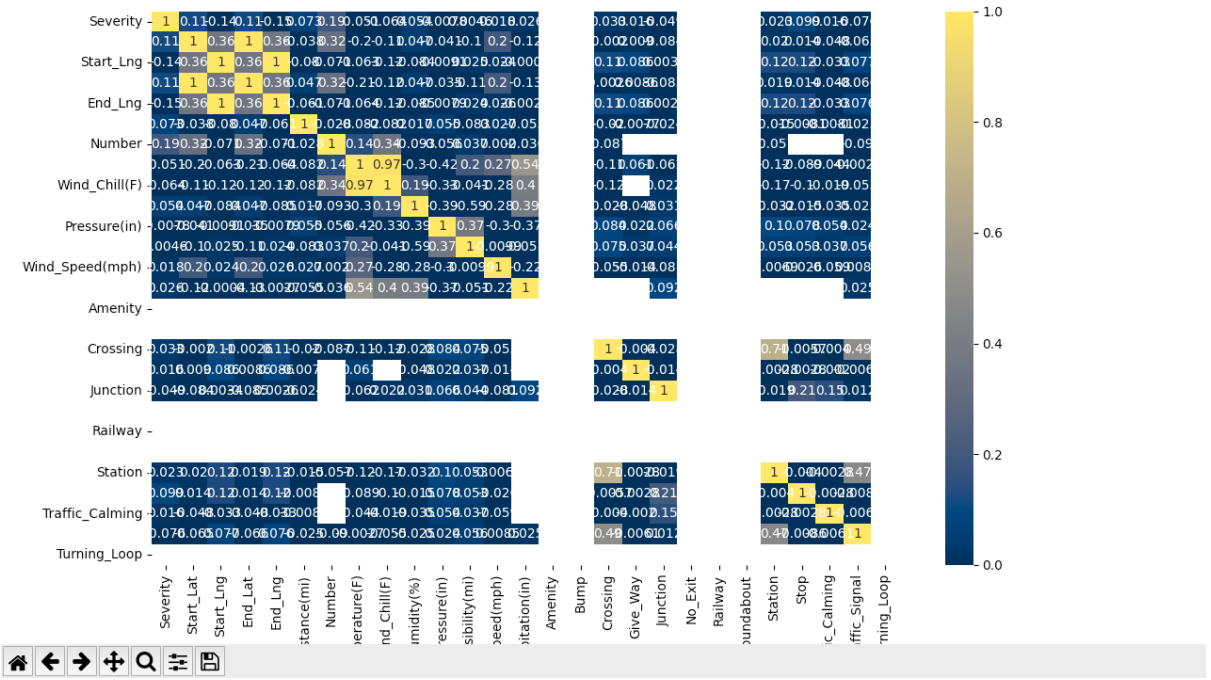






Ahora observando la columna de temperatura podemos apreciar que el histograma que vemos justo antes es bastante variado y extenso, lo cual tiene sentido, pues la temperatura a diferencia de la severidad es muy cambiante, con muchas variaciones debido a las distintas estaciones del año y lugares donde 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 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.

Actividad Evaluable: Patrones con K-means

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

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

13 de Enero del 2022

Instrucciones de la actividad

En esta actividad encontrarás patrones de tus datos utilizando la técnica de clustering k-means. Trabajarás con el conjunto de datos que seleccionaste. Tienes que replicar los pasos vistos durante la clase.

Realiza un código que haga lo siguiente:

1. Carga tus datos
2. Selecciona dos variables que consideres interesantes para este análisis.
3. Determina un valor de k de acuerdo a los datos que tienes, las variables y una pregunta que quieres contestar con este análisis.
4. Utilizando scikitlearn calcula los centros del algoritmo k-means

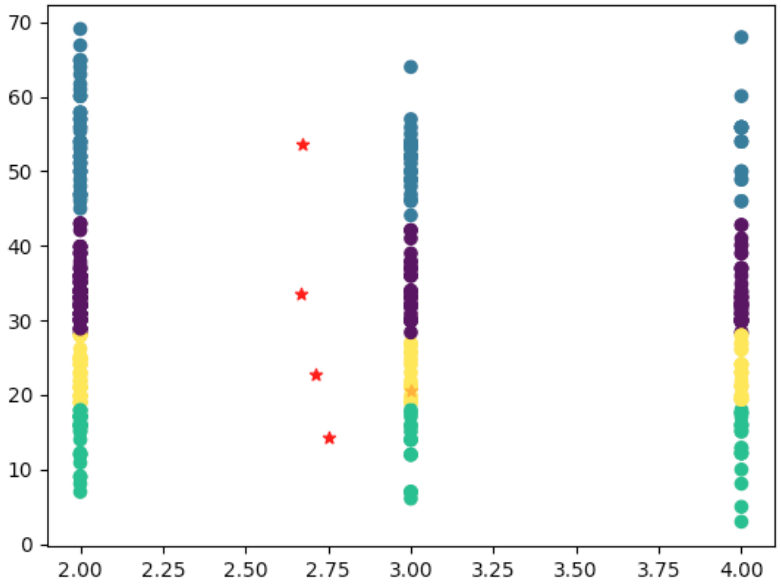
Realiza un pequeño reporte que muestre lo siguiente. Justifica tus respuestas con base en los centros encontrados.

1. Describa el significado de las variables que seleccionaste.
2. ¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?
3. ¿Por qué seleccionaste ese valor de k a usar?
4. ¿Los centros serían más representativos si usaras un valor k más alto? ¿Más bajo? ¿En qué cambiaría la interpretación?
5. ¿Qué distancia tienen los centros entre sí? ¿Hay alguno que este muy cercano a otros?
6. ¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?
7. ¿Qué puedes decir de los datos basándose en los centros?

Acerca de los centros encontrados y algo más.

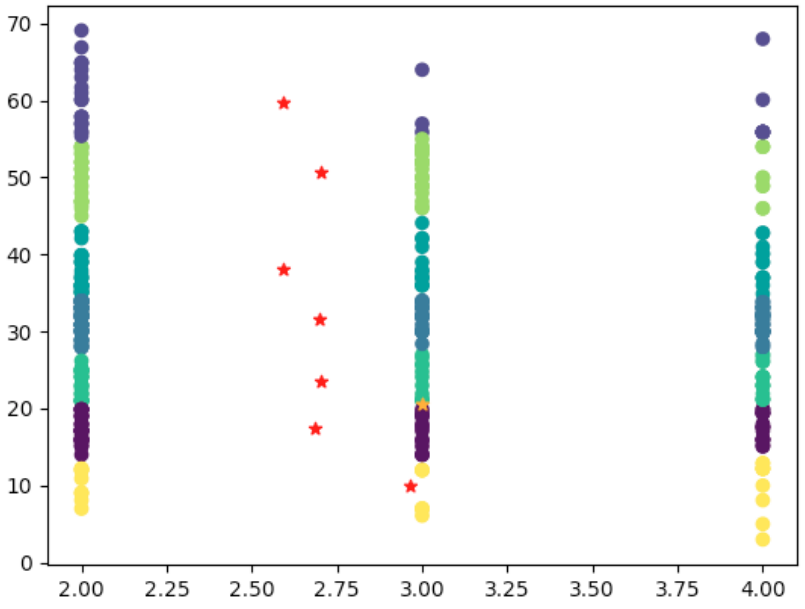
Como lo vimos durante la clase y de lo que nos centraremos es en el uso del algoritmo k-means o k-medias para analizar las variables que encontramos en nuestro dataset y algunos comportamientos interesantes que podemos llegar a encontrar en ellos cuando relacionamos las dos variables de severidad y temperatura, 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.



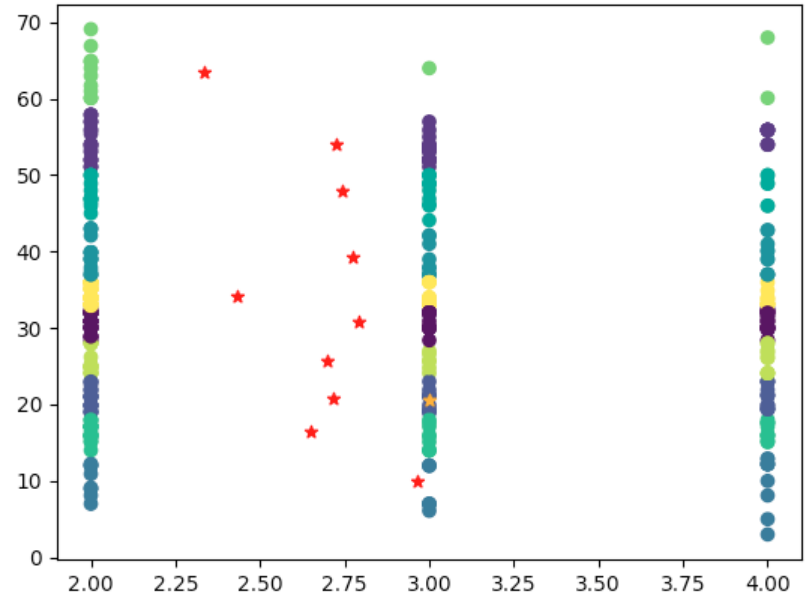

```
[[ 2.66666667 33.59812207]
 [ 2.67010309 53.59690722]
 [ 2.75294118 14.26823529]
 [ 2.71       22.744     ]]
[3]
```

Estos son los resultados usando una k de 4




```
[[ 2.68674699 17.38674699]
 [ 2.59375    59.78125   ]
 [ 2.69677419 31.53677419]
 [ 2.59090909 38.00454545]
 [ 2.70149254 23.44477612]
 [ 2.703125   50.653125  ]
 [ 2.96428571  9.90357143]]
[4]
```

Estos son los resultados usando una k de 7.




```
[[ 2.79207921 30.86039604]
 [ 2.7254902  54.10980392]
 [ 2.71666667 20.83166667]
 [ 2.96428571  9.90357143]
 [ 2.775      39.32       ]
 [ 2.74193548 47.9516129 ]
 [ 2.64912281 16.4122807 ]
 [ 2.33333333 63.52       ]
 [ 2.7        25.6125     ]
 [ 2.43055556 34.25972222]]
[2]
```

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.

Analizando un poco más lo que podemos observar de cada uno de los resultados que me parecieron bastante interesantes, primero vemos que en los primeros resultados con k=4, los centros están bastante alineados, siendo que tres de ellos tienen una distancia de aproximadamente 10 unidades en el eje de las y (la temperatura) y el último está bastante más arriba, siendo una distancia de aproximadamente 20 unidades más y abarca las temperaturas más caliente, por lo que podríamos decir que a pesar de los outliers que existen con la temperatura que observamos antes en el diagrama de caja y bigotes, los centros son estables y de hecho son muy cercanos al promedio de la severidad en todos los centros, por lo que con estos resultados podemos decir que la temperatura no afecta en ningún sentido a la severidad, probablemente por la cantidad de datos que encontramos en ellos el usar una k baja como esta no sea tan útil, por lo que los siguientes resultados a analizar decidí hacerlos con k más altas que puedan mostrarnos tal vez comportamientos más detallados y especificos.

Ahora analizando los siguientes resultados con k=7, podemos apreciar aquí que los centros presentan unas variaciones muy interesantes, siendo que a pesar de que 4 de ellos parecen estar bastante alineados, dos de ellos no, inclinándose hacia un grado menor de severidad (aproximadamente de 2.6 comparado a que los demás están en aproximadamente 2.71) con ciertos rangos de temperatura que son bastante elevados (alrededor de los 40's y arriba de 55 en la temperatura) y un último centro que está en aproximadamente de 3 de severidad y abarca las temperaturas de alrededor de 10, las que son las más bajas, además de que, cuando hablamos de distancia en el eje de la temperatura, los primeros 5 centros están aproximadamente a una distancia de 8-10 unidades entre cada uno y los últimos dos estan 20 unidades por encima de ellos aproximadamente y entre ellos tienen una distancia de 10 unidades. Todo esto podría decirnos dado esos centros un poco que la severidad de un accidente decrece cuando la temperatura del ambiente está aproximadamente cercana a 40 o arriba de 55 como lo vimos con estos resultados y que la severidad puede elevarse cuando la temperatura es cercana a los 10, pero hay que tomar en cuenta que los outliers pueden afectar los centros de estos resultados, dándonos centros poco realistas y bastante inclinados hacia algún extremo que no represente un comportamiento usual y descartando nuestra hipótesis, por lo que tendríamos que seguir analizando para hacer una afirmación certera.

Finalmente, analizaremos los siguientes resultados con k=10. Vemos que con esta k, los centros son bastante equidistante en el eje y (temperatura) siendo la distancia de alrededor de 8-10 unidades entre cada centro, pero lo más interesante es la variación de la severidad especialmente con los rangos aproximados que dimos en el análisis de los resultados cuando k es igual a 7, pues vemos que el centro que abarca las temperaturas más bajas, sigue diciéndonos que la severidad aumenta mucho (comparado a los demás centros) cuando la temperatura es baja, podemos tomar este dato como confirmado, pues los outliers no se encontraban en este rango de la temperatura. Por otro lado, los centros que abarcan entre los 32 al 38 aproximadamente y los que abarcan el arriba de 60, siguen marcando una menor severidad, tal vez es un poco más cuestionable que sea una diferencia tan marcada y más tomando en cuenta que se encontraron outliers anteriormente con las temperaturas elevadas, pero podemos armar una hipótesis mejor fundamentada en base a esto para futuros análisis para confirmar o desmentir con una mayor cantidad de registros.

Me gustaría destacar que estos últimos análisis fueron interesantes debido al aumento del valor k, más que nada debido a que hay bastantes registros y diseccionarlos un poco más podía darnos mejores resultados, como es el caso en está ocasión, siendo que si bien podríamos haber dejado de lado estas variables por los anteriores resultados en reportes anteriores, esta vez encontramos cosas realmente interesantes que tomar en cuenta.