Octavio Andrick Sánchez Perusquia
A01378649

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

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

Actividad

Cuarta entrega

Octavio Andrick Sánchez Perusquia A01378649

jueves 13 de enero de 2022

Profesor
Gilberto Huesca Juárez

Octavio Andrick Sánchez Perusquia
A01378649

Actividad 2

Se trata de una base de datos acerca de accidentes automovilísticos que sucedieron dentro del territorio estadounidense donde se recopilan distintas características del suceso como el lugar, momento del día y condiciones climáticas. La base de datos se proporcionó por el profesor en la plataforma de Canvas.

La base de datos contiene 498 registros de accidentes donde se describen 47 variables por registro. La descripción de cada variable se presenta a continuación, cabe destacar que un “object” se trata de un string:

```
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
```

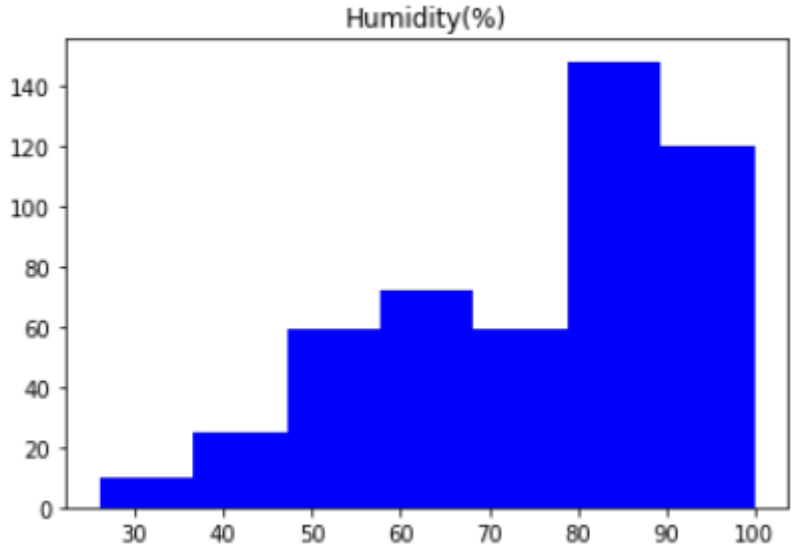
Octavio Andrick Sánchez Perusquia
A01378649

Escogí la severidad del accidente como un rubro a describir, el cual tiene un rango de 2 a 4, sin valores de punto flotante. De igual manera, escogí el valor del porcentaje de humedad, cuyo intervalo en este dataset es 26 a 100 por ciento. Juzgando con base en las estadísticas descriptivas, la media y la mediana de la severidad se acercan al 2, por lo que se asume que la mayoría de los accidentes forman parte del umbral de severidad considerado en estas mediciones. Al mirar a su desviación estándar, se observa que un accidente de severidad 3 puede tener una probabilidad plausible, pero un accidente de severidad 4 sería un evento improbable, ya que implica casi dos desviaciones estándar respecto a la media.
Respecto a la humedad, la media se encuentra arriba del 75%, por lo que se considera un ambiente arriba de lo normal (50%). La desviación estándar indica que es normal encontrarse con humedades significativamente más altas. Aunque la humedad no necesariamente implica que la calle está mojada, puede influir en fenómenos como la niebla que reduce visibilidad o presión de las llantas que a su vez reduce la capacidad de tracción o frenado. Por lo tanto, es plausible hacer una hipótesis de que la humedad es factor con correlación positiva a la probabilidad de un accidente, aunque un mayor análisis sería necesario para entender a profundidad.

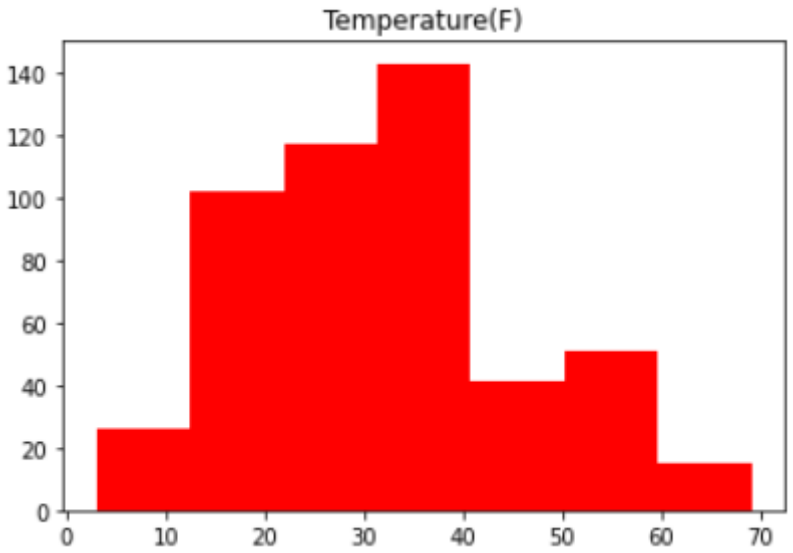
Octavio Andrick Sánchez Perusquia
A01378649

Actividad 3

Escogí la humedad relativa (HR) nuevamente como la actividad pasada, no obstante, cambié la severidad por la temperatura porque dicha variable tiene un mayor rango de variables para demostrar dentro las herramientas usadas para describir los datos. Primeramente, se mostrará el histograma de la HR, cuyo rango es de 26 a 100 por ciento, utilizando 7 intervalos que dividen dicho intervalo:

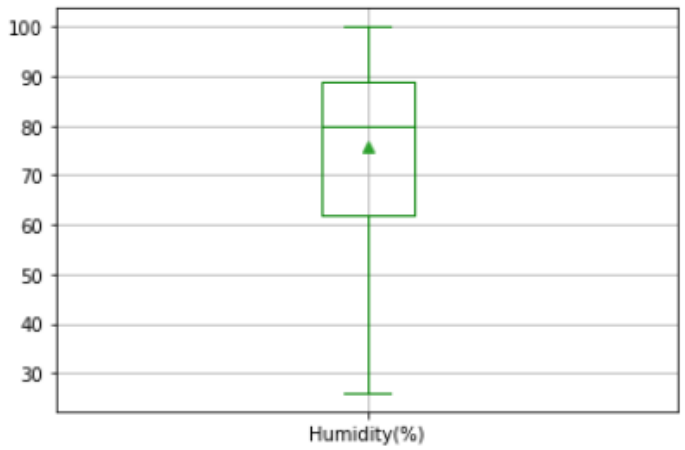


Posteriormente, la temperatura tiene un rango de 3 a 69.1. También se dividió dicho rango en 7 intervalos para la visualización:

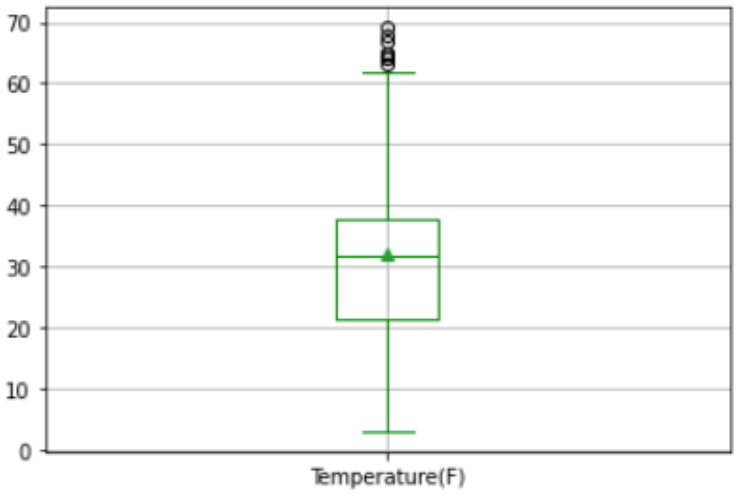


Octavio Andrick Sánchez Perusquia
A01378649

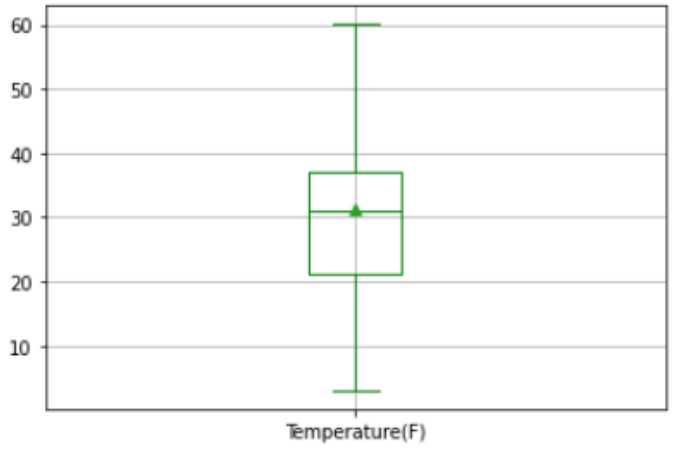
El diagrama de cajas y bigote de la HR muestra que el rango intercuartil se acerca al rango de 60 a 90, pero no se encuentran outliers:



A primera vista, el diagrama de la temperatura tiene un rango intercuartil que se acerca al 20 y 40, pero los valores mayores a 60 se consideran como outliers:

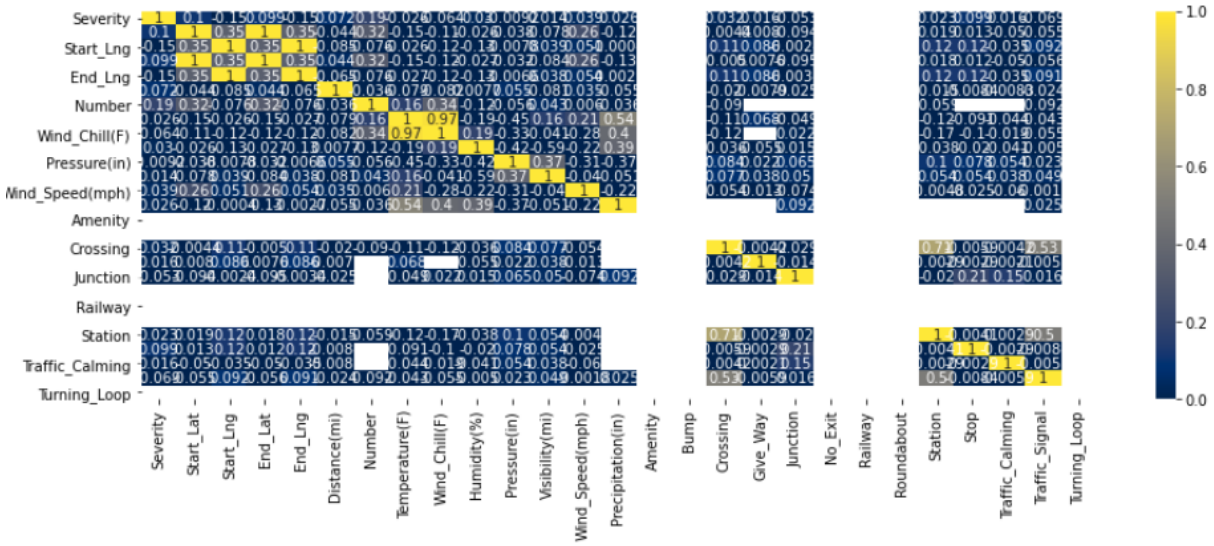


Entonces, se realiza el ajuste para no incluir valores mayores que 61° Fahrenheit:



Octavio Andrick Sánchez Perusquia
A01378649

También, se muestran las correlaciones entre todas las variables numéricas dentro de un mapa de calor:

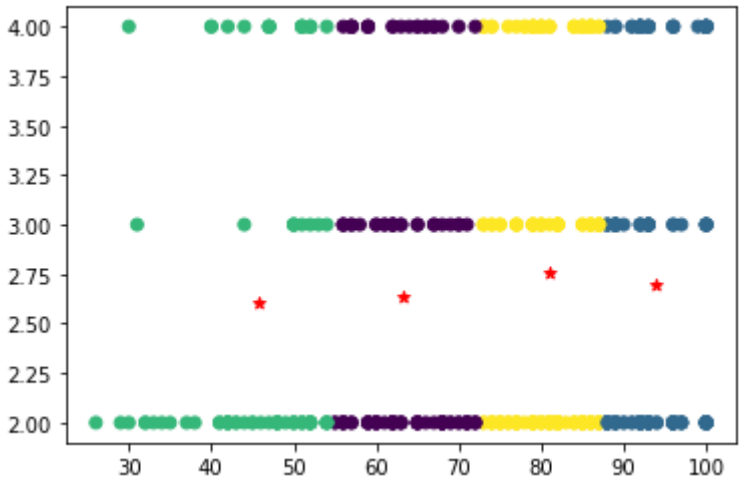


Referente a las dos variables seleccionadas, se demuestra una correlación débil (menor que 0.3). Al considerar su correlación con la severidad, es significativamente débil, ya que los valores son 0.02 y 0.06 respectivamente.

Octavio Andrick Sánchez Perusquia
A01378649

Actividad 4

Utilizando las variables de entregas pasadas, decidí relacionar la severidad y la HR para formar categorías, a pesar de la carencia de correlación entre ellas vista en la actividad pasada. Dicho esto, se muestra el output del algoritmo de K-means para dichas variables, considerando la HR como el eje de las abscisas y la Severidad como las ordenadas:



A primera vista, al analizar dimensionalmente los centros, los cuatro centros están esparcidos casi uniformemente en el dominio de la gráfica, con excepción de una ligera separación adicional del primer centroide con respecto al inicio de la gráfica. Esto significa que no se presentaron datos de accidentes con la misma densidad en eventos de HRs más bajas, pero dicha densidad de eventos comenzó a normalizarse a lo largo del resto del dominio. Lo anterior podría sugerir una hipótesis de que los accidentes son más probables a partir de un umbral de humedad, probablemente 40 en este caso. Respecto al eje y, los centroides se mantienen muy cercanos entre sí, con excepción de un ligero incremento en los últimos dos. Similarmente al primer planteamiento, esto podría sugerir una hipótesis que indique que en cierto rango de humedad se maximiza la severidad probabilísticamente. En este caso, esta cúspide se observa después de 70 de HR y antes de 90 de HR. Sin embargo, también podría significar que la infraestructura de las localidades donde es común esa clase de HR tengan otro problema como un diseño deficiente de vialidades.

En caso de incluir outliers, los centroides se descentralizarían en una magnitud proporcional a la lejanía de dichos datos. Por ejemplo, si se incluyen humedades menores a 23, entonces el primer centroide experimentaría un desplazamiento a la izquierda, pero sería improbable que dicho comportamiento se propague a los otros centroides si no es una cantidad abundante de datos. No obstante, no se encontró ningún outlier en la HR inicialmente al ser analizada con la gráfica de cajas y bigotes.

Se escogieron 4 centroides ya que de esta forma se representa de mejor manera los rangos de HR. Es decir, en caso de usar menos centroides, existe menos información respecto a las categorías generadas al ser formadas por rangos extensos, ya que el objetivo de esta observación es enfocarse en el comportamiento de la severidad de un accidente dependiendo de los rangos específicos de humedad. Entonces, con 4 centroides se obtiene la granularidad requerida sin reducir los rangos de HR significativamente.

Octavio Andrick Sánchez Perusquia
A01378649

Por ejemplo, una humedad saludable se considera en un rango de 30 a 50 para hogares, y utilizando este número de K, se aprecia dicho rango en una categoría.