Actividad Evaluable:

Patrones con K-means

Herramientas computacionales: el arte de la analítica

Javier Piña Camacho
A01701478

13 de enero de 2022

Introducción

Los datos que se utilizaron en esta actividad fueron obtenidos de choques de automóviles en Estados Unidos. Cuentan con una descripción del evento y datos relevantes sobre el ambiente, duración, fecha y lugar del suceso. Cada registro representa un accidente y recopila las variables que lo acompañaron.

Base de datos de: https://www.kaggle.com/sobhanmoosavi/us-accidents

Cantidad de datos, las variables que contiene cada vector de datos y el tipo de variables

La base de datos cuenta con 499 registros, los vectores contienen 47 variables de tipo entero, flotante, objeto y booleano.

1. ID object
2. Severity int64
3. Start_Time object
4. End_Time object
5. Start_Lat float64
6. Start_Lng float64
7. End_Lat float64
8. End_Lng float64
9. Distance(mi) float64
10. Description object
11. Number float64
12. Street object
13. Side object
14. City object
15. County object
16. State object
17. Zipcode object
18. Country object
19. Timezone object
20. Airport_Code object
21. Weather_Timestamp object
22. Temperature(F) float64
23. Wind_Chill(F) float64
24. Humidity(%) float64
25. Pressure(in) float64
26. Visibility(mi) float64
27. Wind_Direction object
28. Wind_Speed(mph) float64
29. Precipitation(in) float64
30. Weather_Condition object
31. Amenity bool
32. Bump bool
33. Crossing bool

34. Give_Way bool
35. Junction bool
36. No_Exit bool
37. Railway bool
38. Roundabout bool
39. Station bool
40. Stop bool
41. Traffic_Calming bool
42. Traffic_Signal bool
43. Turning_Loop bool
44. Sunrise_Sunset object
45. Civil_Twilight object
46. Nautical_Twilight object
47. Astronomical_Twilight object

Hípotesis

La mayoría de los accidentes de autos ocurren por la poca visibilidad causada por la humedad en el ambiente.

Variables escogidas para el análisis

Humidity(%), con un rango de valores de 26 a 100
Visibility(mi), con un rango de valores de 0.5 a 20.0

Basado en la media, mediana y desviación estándar de cada variable…

Los valores de la humedad varían mucho de acuerdo a su desviación estándar de 17.42. También, se puede notar un porcentaje de humedad alto en los accidentes de autos si vemos la media de 75.65%. La mediana de 80% apoya lo anterior, pues al menos la mitad de los accidentes contó con un alto porcentaje de humedad.

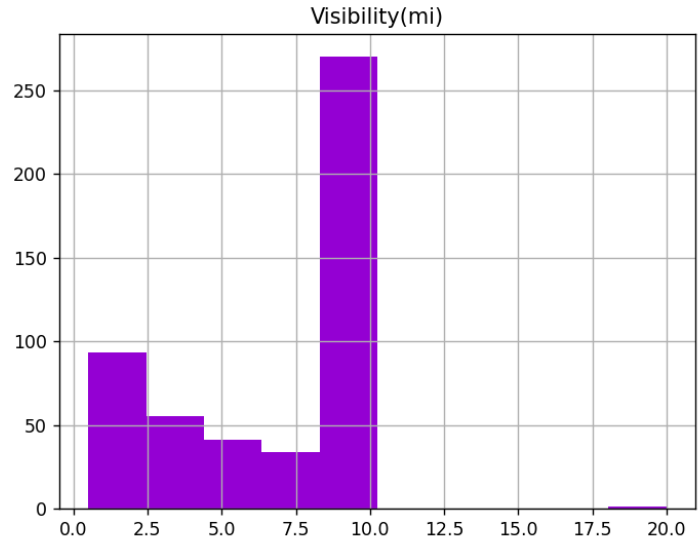

Figura 1. Histograma de la columna de visibilidad.

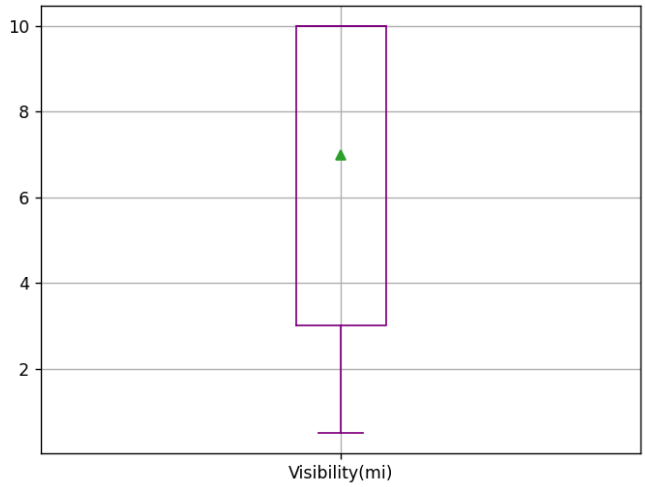

Figura 2. Diagrama de cajas y bigotes de la columna de visibilidad eliminando valores mayores a 15.

En cuanto a la visibilidad, la dispersión es menor con un valor de 3.597. Es por ello que se repiten más valores que en la columna de humedad. Tomando en cuenta que la media es de 7.029 y la mediana de 10, es decir se encuentran cercanos a la zona media media del rango (0.5 a 20), los accidentes no parecen ser muy influenciados por la visibilidad pero tienden ligeramente a suceder cuando la visibilidad disminuye.

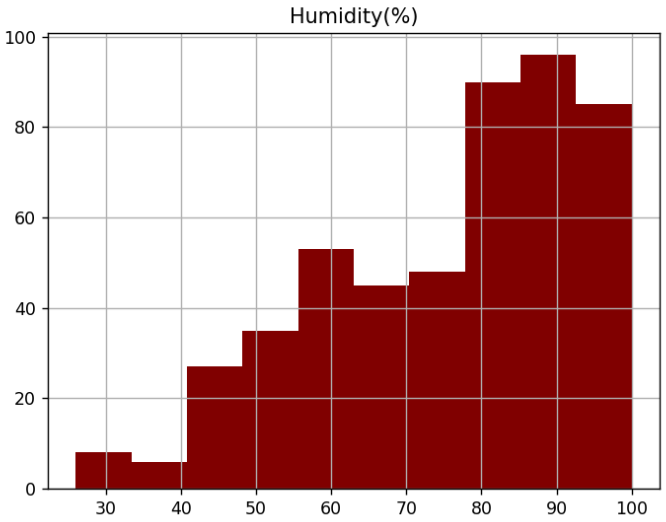

Figura 3. Histograma de la columna de humedad.

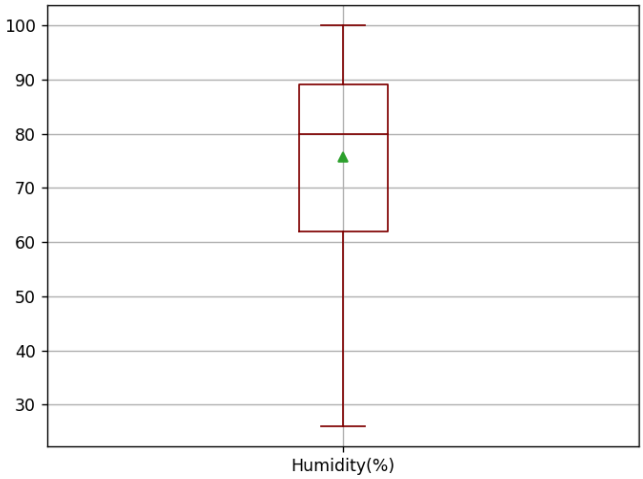

Figura 4. Diagrama de cajas y bigotes de la columna de humedad.

Existe una pequeña relación inversa entre estos dos valores. Puesto que la humedad tiende a ser alta y la visibilidad se inclina ligeramente a ser menor en los accidentes de coche.

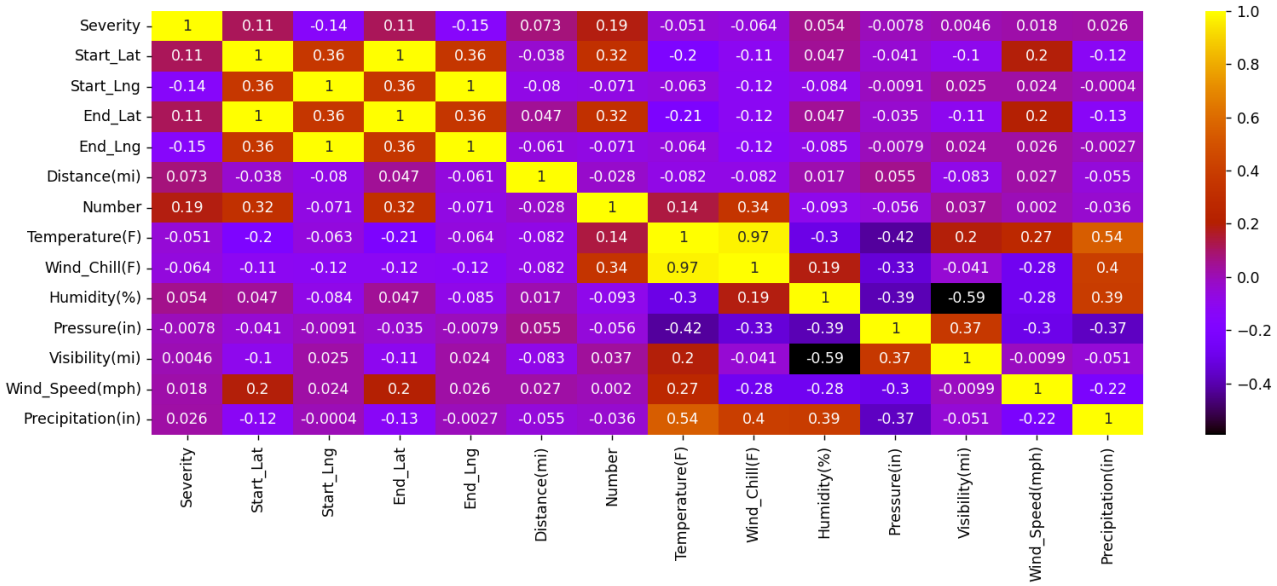

Figura 5. Mapa de calor de las correlaciones entre todas la variables númericas de la base de datos.

Se pueden hacer mejores observaciones de ambas variables con ayuda de las gráficas presentadas en cada figura. El mapa de calor de la figura 5 resulta ser lo más significativo, dandonos confirmación de una leve correlación inversa entre ambas variables. No obstante, podemos comprobar con los histogramas de las figuras 1 y 3, que aunque la visibilidad vaya disminuyendo gradualmente mientras que la humedad aumenta, la visibilidad tiene un incremento que representa la mayor parte de los datos cuando se llega a la visibilidad media. De acuerdo a los diagramas de cajas y bigotes (figura 2 y 4), podemos ver que los datos son significativos puesto

que no se encuentran valores atípicos. En el caso de la figura 2 se eliminaron datos mayores a 15 debido a la poca cantidad de ellos pero incluso si no se eliminaban en el diagrama no aparecen datos atípicos. Los diagramas de cajas también muestran visualmente la correlación inversa al tener la humedad su rango intercuartil en la parte baja y el rango de la visibilidad estar en la parte alta.

Por último se analizará una gráfica de dispersión para determinar si existen grupos de datos que demuestren la hipótesis.

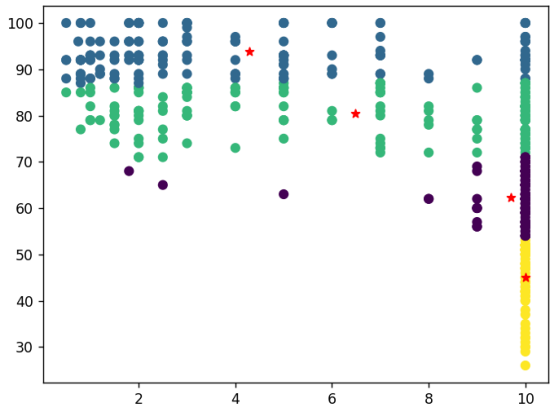

Figura 6. Gráfica de dispersión con el método de agrupación k-medias. La variable de visibilidad se encuentra en el eje y la variable de humedad en el eje y.

Se utilizó el método de k-medias para determinar los grupos que forma la base de datos tomando en cuenta las dos variables del análisis. Para el número de centros (k) se probaron los valores del 2 al 5. Se seleccionó un valor de k=4 por ser el valor que generaba divisiones más claras. Lo cual hace que los centros generados puedan ser representativos de los datos. Al disminuir el valor de k todos los grupos tienen valores que se separan de la visibilidad alta. Mientras que si se aumenta el valor, no se aprecian grupos nuevos relevantes. Por lo tanto, un menor valor dificultaría la interpretación y podría ser errada; un valor mayor no aporta nada nuevo a la interpretación.

Los centros obtenidos fueron: (9.69363636, 62.3), (6.51097561, 80.59146341), (10, 44.98529412) y (4.20167785, 93.87919463). Los más cercanos son el 2 y el 3. En caso de que hubiera muchos valores atípicos los centros se moverían. En caso que fuera de visibilidad se irían más a la derecha al no poder existir visibilidad menor a 0 y mayores a ese valor se mantendrían en el rango. Si fueran de humedad los centros subirían si fueran valores altos o bajarían si fueran valores bajos.

La Figura 6 nos muestra como va subiendo la cantidad de accidentes donde se deteriora la visibilidad cuando la humedad sube. Mientras la humedad se mantenga debajo o igual al 50% no habrá accidentes que presenten poca visibilidad en el ambiente. Conforme sube la humedad los grupos presentan accidentes donde la visibilidad se separa del valor de 10 mi. En los dos grupos más altos es donde los valores de visibilidad disminuyen más. El grupo más alto es el único que parece tener la mayor parte de sus valores separados de la visibilidad de 10 mi. Es claro entonces que los accidentes con poca visibilidad aumentan por la humedad pero el impacto se vuelve significativo a partir de un 70% de humedad.

Conclusión:

La hipótesis inicial era incorrecta, la mayoría de los accidentes no involucran poca visibilidad causada por la humedad. Curiosamente en la mayoría de los accidentes se registró una visibilidad alta. Sin embargo, sí existe una correlación inversa entre la visibilidad y la humedad del ambiente registrada en los accidentes. El número de accidentes en los que la visibilidad es baja aumenta cuando la humedad sube. Por lo cual, más probable que un accidente sea provocado por poca visibilidad cuando en el ambiente haya humedad mayor del 85%. En el resto de los casos es más probable que el accidente no se deba a la visibilidad del ambiente.