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:

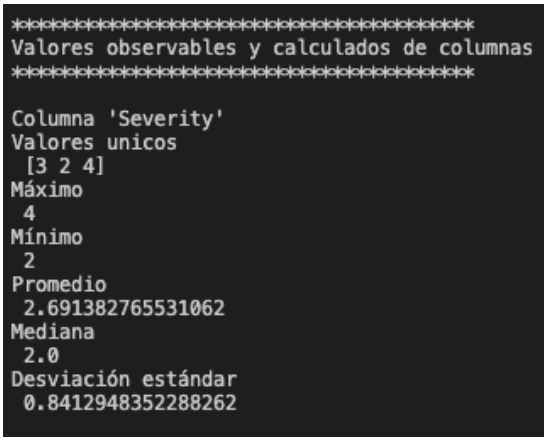
```
ID                      object
Severity                 int64
Start_Time              object
End_Time                object
Start_Lat              float64
Start_Lng              float64
End_Lat                float64
End_Lng                float64
Distance(mi)           float64
Description             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
```

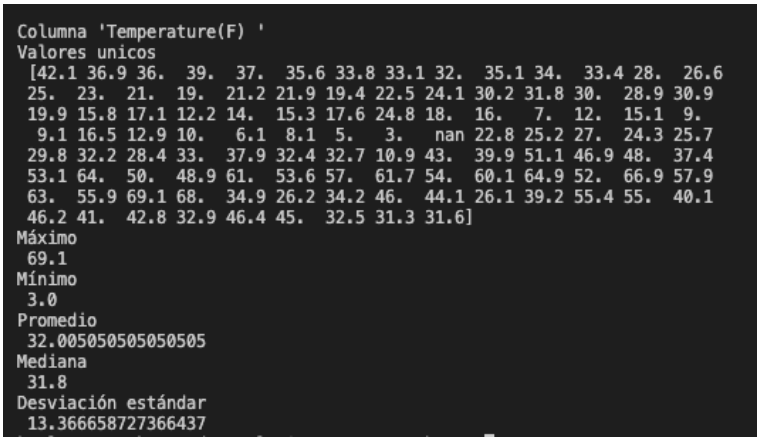
```
Wind_Chill(F)          float64
Humidity(%)            float64
Pressure(in)           float64
Visibility(mi)         float64
Wind_Direction          object
Wind_Speed(mph)        float64
Precipitation(in)      float64
Weather_Condition       object
Amenity                   bool
Bump                      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.

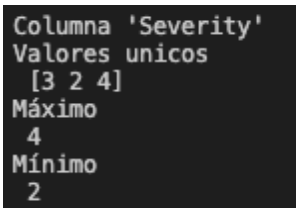
Acerca de las columnas seleccionadas

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.

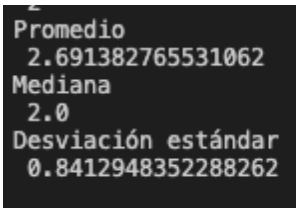




Primero haremos un análisis de la severidad, la cual indica la severidad que tuvo el accidente con valores que en teoría deberían ir del 1 al 4, siendo 1 siendo el menos severo o menor gravedad y 4 el de mayor gravedad o más severo. A pesar de está teoría, vemos que sus valores únicos son el 2, 3 y 4,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