M2.851 · Tipología y ciclo de vida de los datos
Máster universitario en Ciencia de datos (*Data science*)
Estudios de Informática, Multimedia y Telecomunicación

Práctica 2: ¿Cómo realizar la limpieza y análisis de datos?

Agustin Rovira Quezada & Adrian Vega Morales

- 1 Descripción del dataset
- 2 Integración y selección
- 3 Limpieza de los datos
  - 3.1 Gestiona elementos nulos
  - 3.1 Gestiona outliers
- 4 Análisis de los datos
  - 4.1 Seleccion de Grupos
  - 4.2 Comprobar Normalidad
  - 4.3 Pruebas Estadisticas
- 5 Representación de los resultados
- 6 Resolución del problema

1. Descripción del dataset

¿Por qué es importante y qué pregunta/problema pretende responder?

	ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Descripti
0	A-1	3	2016-02-08 00:37:08	2016-02-08 06:37:08	40.108910	-83.092860	40.112060	-83.031870	3.230	Betwe Sawmill Rd/E 20 and O 315/Olentan
1	A-2	2	2016-02-08 05:56:20	2016-02-08 11:56:20	39.865420	-84.062800	39.865010	-84.048730	0.747	At OH-4/O 235/Exit 4 Accide
2	A-3	2	2016-02-08 06:15:39	2016-02-08 12:15:39	39.102660	-84.524680	39.102090	-84.523960	0.055	At I-71/U 50/Exit Accide
3	A-4	2	2016-02-08 06:51:45	2016-02-08 12:51:45	41.062130	-81.537840	41.062170	-81.535470	0.123	At D Ave/Exit 2 Accide
4	A-5	3	2016-02-08 07:53:43	2016-02-08 13:53:43	39.172393	-84.492792	39.170476	-84.501798	0.500	At Mitch Ave/Exit Accide

5 rows × 47 columns

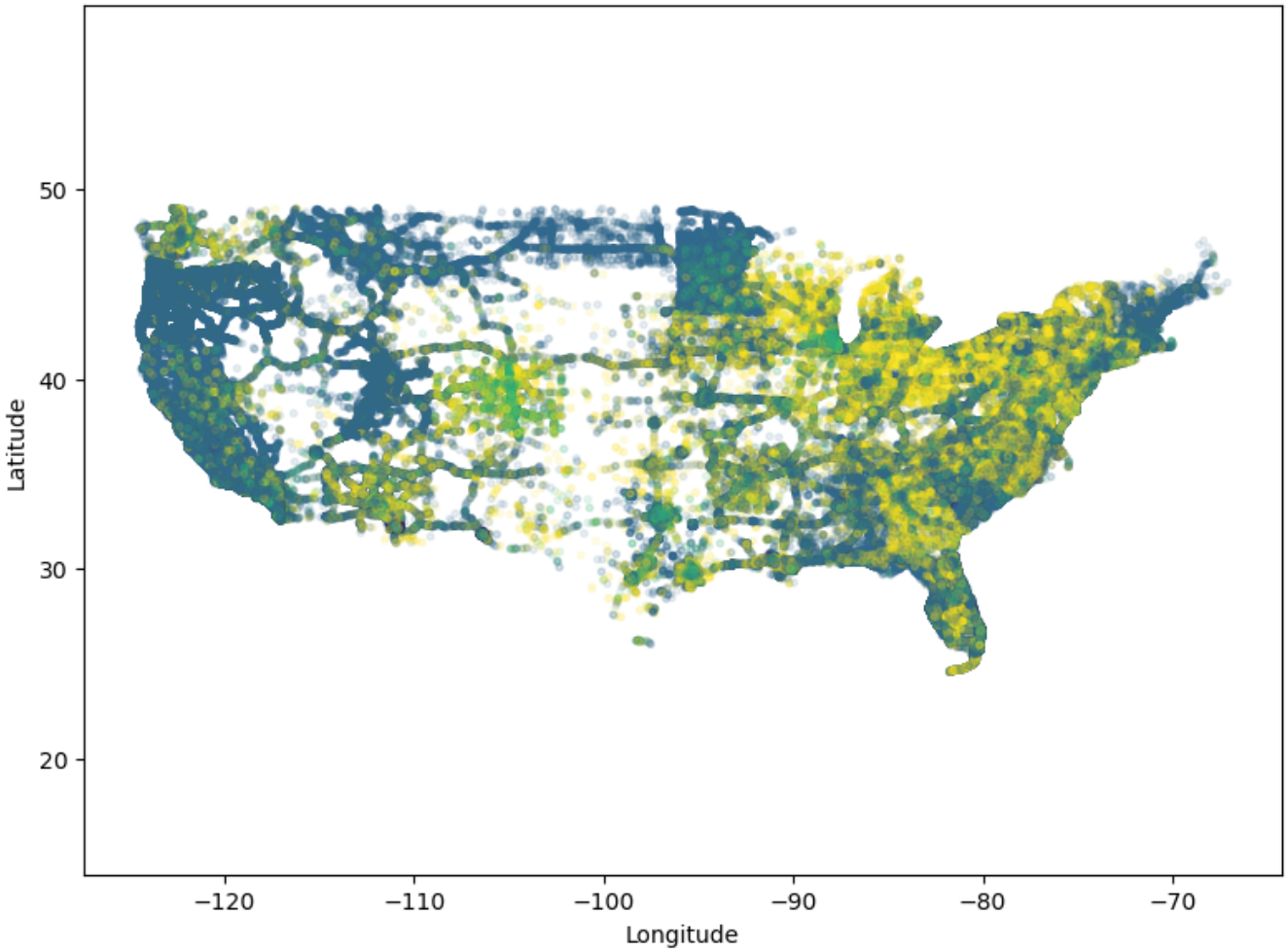
Reducir los accidentes de trafico es un gran desafío para la seguridad pública en todo el mundo. En concreto en Estados Unidos donde se ha visto una enorme cantidad de accidentes diarios en los ultimos 5 años. Asi lo demuestra el total de 2 millones de accidentes recogidos en el dataset de Kaggle US Accidents (2016 - 2021)[1].

Este es un conjunto de datos de accidentes automovilísticos de todo el país, que cubre 50 estados de los EE. UU. Los datos de accidentes se recopilan desde febrero de 2016 hasta diciembre de 2021, utilizando múltiples API que proporcionan transmisión de datos de incidentes (o eventos) de tráfico. Estas API transmiten datos de tráfico capturados por una variedad de entidades, como los departamentos de transporte estatales y de EE. UU., agencias de aplicación de la ley, cámaras de tráfico y sensores de tráfico dentro de las redes de carreteras.

Con la informacion de este dataset buscamos realizar un analisis exploratorio con la finalidad de identificar los factores que tengan influencia sobre la severidad de los accidentes. Por otro, es tambien se quiere encontrar los focos de concentracion de accidentes con la finalidad de posibilitar a las autoridades competentes la toma de decisiones necesarias en los puntos geograficos estratégicos para reducir el numero de accidentes o al menos la severidad de estos.

[1]https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

```
<module 'matplotlib.pyplot' from 'C:\\Users\\agust\\.conda\\envs\\Proyects\\lib\\site-package
s\\matplotlib\\pyplot.py'>
```



2. Integración y selección

Tras una primera exploracion descriptiva y una lectura comprensiva del significado de cada variable, podemos ver que es posible seleccionar un subconjunto de variables con el que trabajar.

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

	Variable	Tipo
35	Railway	bool
36	Roundabout	bool
37	Station	bool
38	Stop	bool
39	Traffic_Calming	bool
40	Traffic_Signal	bool
41	Turning_Loop	bool
42	Sunrise_Sunset	object
43	Civil_Twilight	object
44	Nautical_Twilight	object
45	Astronomical_Twilight	object

```
(2845342, 47)
```

Columns tipo Bool

Empezando con las variables tipo bolean podemos ver que las 13 variables no estan balanceadas, todas tienen mas de un 98% de frecuencia para el valor "False". Por lo tanto estas variables pueden ser eliminadas ya que no aportan informacion relevante.

Como se aplicamos una subseleccion con .iloc

	count	unique	top	freq
Amenity	2845342	2	False	99
Bump	2845342	2	False	100
Crossing	2845342	2	False	93
Give_Way	2845342	2	False	100
Junction	2845342	2	False	90
No_Exit	2845342	2	False	100
Railway	2845342	2	False	99
Roundabout	2845342	2	False	100
Station	2845342	2	False	98
Stop	2845342	2	False	98
Traffic_Calming	2845342	2	False	100
Traffic_Signal	2845342	2	False	91
Turning_Loop	2845342	1	False	100

Columns tipo Object

Se observa que muchas de las columnas de este tipo estan desbalanceadas, tiene un 70% de frecuencia para una de los factores. Por lo tanto los eliminaremos las columnas ['Country', 'Astronomical_Twilight', 'Civil_Twilight', 'Sunrise_Sunset', 'Nautical_Twilight', 'Side']

Por otro lado, hay otras columnas que carecen de significado para el analisis que se va a realizar por lo cual tambien se eliminaran. Estas son las siguientes [Timezone,ID,Description,County]

	count	unique	top	freq
Country	2845342	1	US	2845342
Astronomical_Twilight	2842475	2	Day	2176983
Civil_Twilight	2842475	2	Day	1929103
Sunrise_Sunset	2842475	2	Day	1811935
Nautical_Twilight	2842475	2	Day	2063472
Side	2845342	3	R	2353309
Timezone	2841683	4	US/Eastern	1221927
Wind_Direction	2771567	24	CALM	433622
State	2845342	49	CA	795868
Weather_Condition	2774706	127	Fair	1107194
County	2845342	1707	Los Angeles	234122
Airport_Code	2835793	2004	KCQT	52790
City	2845205	11681	Miami	106966
Street	2845340	159651	I-95 N	39853
Zipcode	2844023	363085	91761	6162
Weather_Timestamp	2794606	474214	2021-12-17 14:53:00	640
Description	2845342	1174563	A crash has occurred causing no to minimum del...	7978
Start_Time	2845342	1959333	2021-01-26 16:16:13	214
End_Time	2845342	2351505	2021-11-22 08:00:00	88
ID	2845342	2845342	A-1	1

Coulmns tipo numeric

De las columnas tipo numericas solo tenemos que eliminar Number ya que pose mas de un 90% de NANs

	count	mean	std	min	25%	50%	75%	
Severity	100	2.137572	0.478722	1.000000	2.000000	2.000000	2.000000	4.00
Start_Lat	100	36.245201	5.363797	24.566027	33.445174	36.098609	40.160243	4.90
Start_Lng	100	-97.114633	18.317819	-124.548074	-118.033113	-92.418076	-80.372431	-6.71
End_Lat	100	36.245321	5.363873	24.566013	33.446278	36.097987	40.161049	4.90
End_Lng	100	-97.114387	18.317632	-124.545748	-118.033331	-92.417718	-80.373383	-6.71
Distance(mi)	100	0.702678	1.560361	0.000000	0.052000	0.244000	0.764000	1.55
Number	39	8089.408114	18360.093995	0.000000	1270.000000	4007.000000	9567.000000	9.99
Temperature(F)	98	61.793556	18.622629	-89.000000	50.000000	64.000000	76.000000	1.96
Wind_Chill(F)	83	59.658231	21.160967	-89.000000	46.000000	63.000000	76.000000	1.96
Humidity(%)	97	64.365452	22.874568	1.000000	48.000000	67.000000	83.000000	1.00
Pressure(in)	98	29.472344	1.045286	0.000000	29.310000	29.820000	30.010000	5.89
Visibility(mi)	98	9.099391	2.717546	0.000000	10.000000	10.000000	10.000000	1.40
Wind_Speed(mph)	94	7.395044	5.527454	0.000000	3.500000	7.000000	10.000000	1.08
Precipitation(in)	81	0.007017	0.093488	0.000000	0.000000	0.000000	0.000000	2.40

Transformacion de columnas

Como resultado nos quedaremos con 21 columnas de 47.

['Severity','Start_Lat', 'Start_Lng', 'End_Lat', 'End_Lng', 'Distance(mi)', 'Street', 'State', 'Zipcode', 'Weather_Timestamp', 'Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Direction', 'Wind_Speed(mph)', 'Precipitation(in)', 'Weather_Condition',"Duration","Year","Time"]

3. Limpieza de los datos

3.1 ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos

	NANs	Porcentage
Precipitation(in)	549458	19.0
Wind_Chill(F)	469643	17.0
Wind_Speed(mph)	157944	6.0
Wind_Direction	150309	5.0
Humidity(%)	73092	3.0
Weather_Condition	70636	2.0
Visibility(mi)	70546	2.0
Temperature(F)	69274	2.0
Pressure(in)	59200	2.0
Airport_Code	9549	0.0
Zipcode	1319	0.0
City	137	0.0
Street	2	0.0

Se observo en el bloque anterior que habia missing values en el aiport_code y aunque es una de las columnas a eliminar porque no son necesarias para el analisis posterior si que la vamos a usar para eliminar registros de condiciones climatologicas como Precipitation. Esto se decidio asi ya que si no hay registro de condiciones meteorologicas para un accidente no podemos usarlos para el analisis de agrupacion o regresion.

Por otro lado las variables Precipitation(in) y Wind_Chill(F) tiene un porcentage de NaNs de casi un 20%, un valor muy alto para una inputacion. Por lo cual optaremos por eliminar esas filas.

```
                   NANs  Porcentage
Visibility(mi)     6739    0.302797
Weather_Condition  6531    0.293451
Humidity(%)        3201    0.143827
Pressure(in)       1445    0.064927
Wind_Direction       14    0.000629
Street                1    0.000045
```

Vamos a imputar los valores con una media una ventana movil en un rango de una hora. Para dicha imputacion se tendra que cargar el dataset otra vez y clasificar los valores por estado y horas. Esto se hara para Visibility(mi), Weather_Condition, Humidity(%), Pressure(in)

```
                   NANs  Porcentage
Weather_Condition  6531    0.293451
Visibility(mi)     1097    0.049290
Humidity(%)         259    0.011637
Pressure(in)        130    0.005841
Wind_Direction       14    0.000629
Street                1    0.000045
```

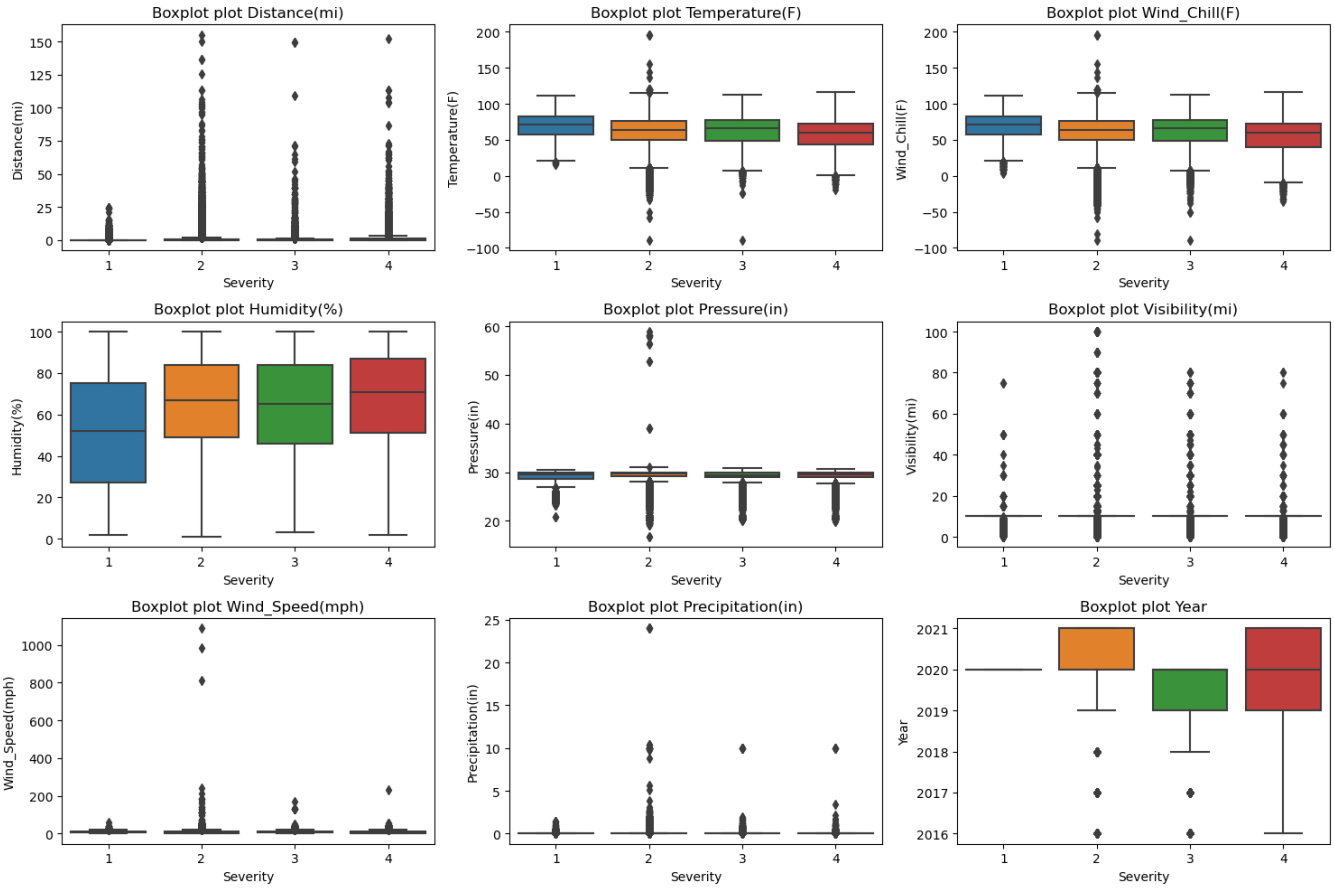
Tras la imputacion vemos que no hay demasiados valores con NaNs y que representan un valor insingnificante de porcentage. Por lo cual procedemos simplemente a eliminar esas filas

3.2 Identifica y gestiona los valores extremos

Visualizamos como se distribuyen los factores de severity a lo largo del dataset. Como se vera en el plot la clase no esta balanceada.

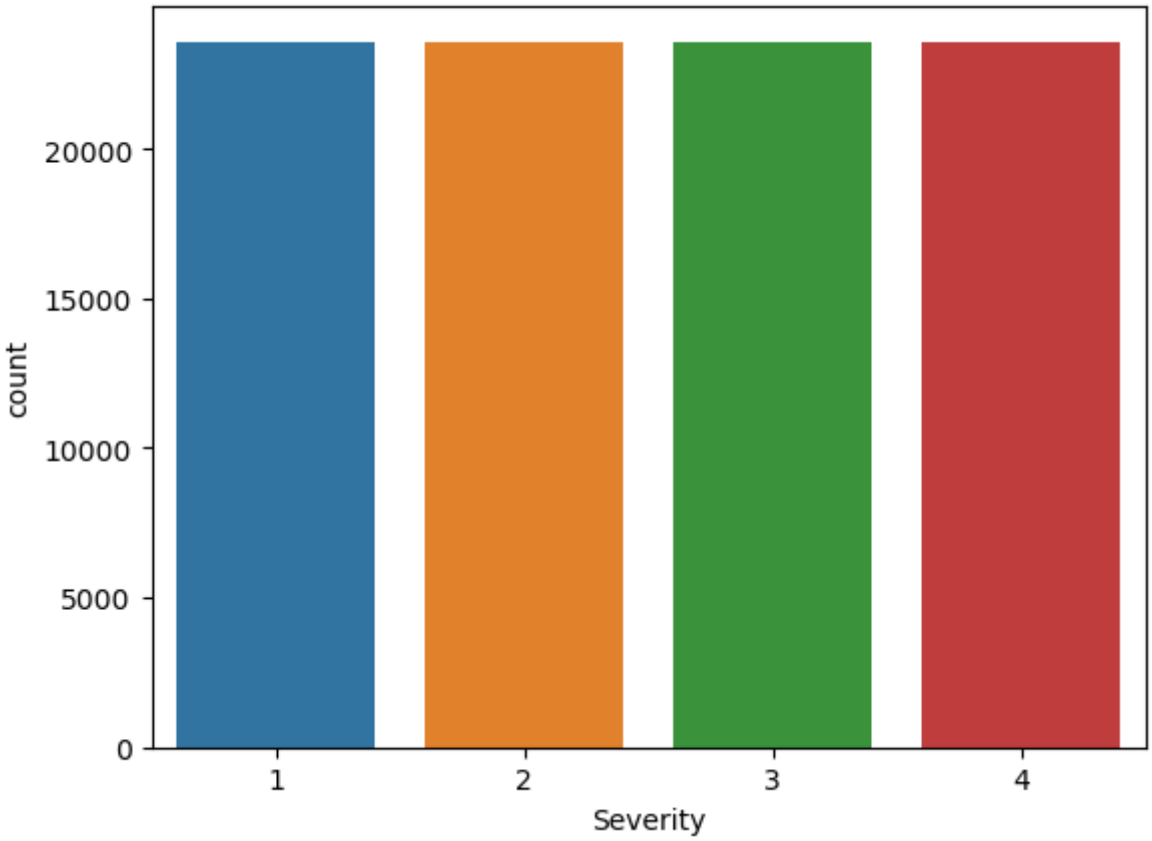
Los valores atípicos en los datos categóricos también pueden atribuirse al problema del desequilibrio de clases. Esto significa que los datos para cada clase no están en una proporción similar. En tal situación, utilizaremos tecnicas de muestreo, concretamente Muestra aleatoria simple con sustitución

```
2    2066877
3      64824
4      62733
1      23581
Name: Severity, dtype: int64
```



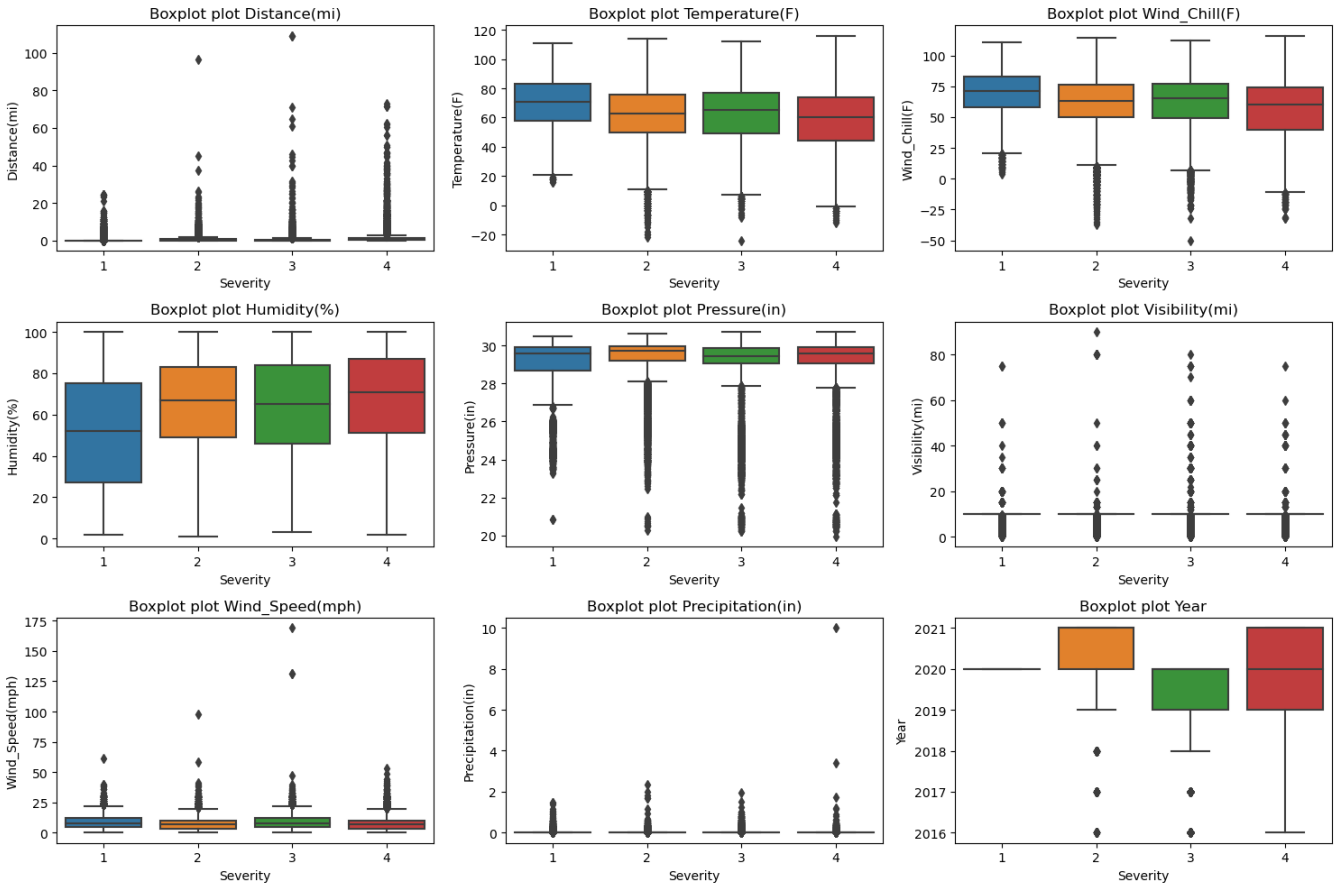

```
1    23581
2    23581
3    23581
4    23581
Name: Severity, dtype: int64
```

```
<AxesSubplot:xlabel='Severity', ylabel='count'>
```

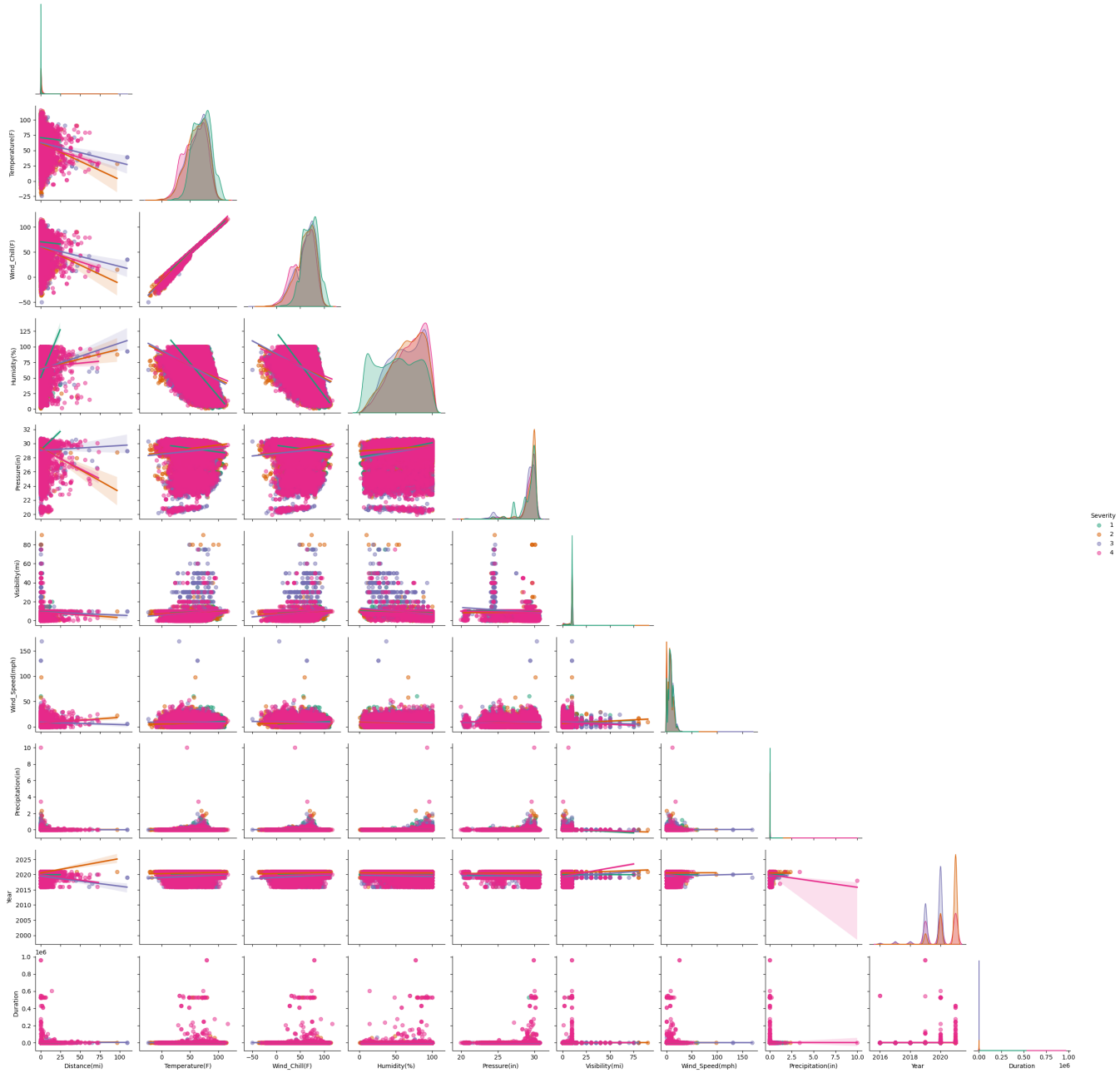


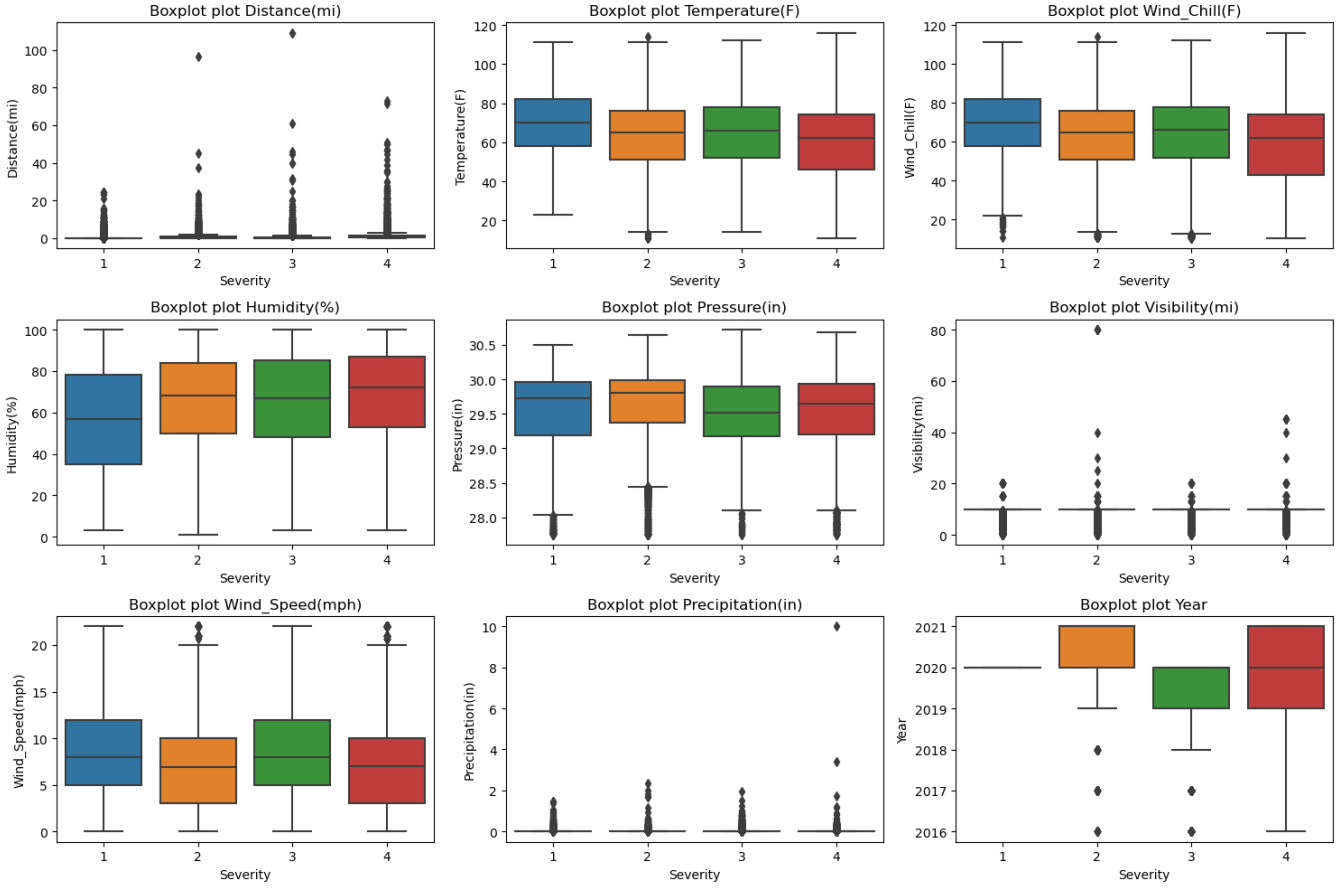
A continuacion visualizamos las variables independientes numericas contra la severidad para ver por una parte si las clases estan balanceadas respecto a las variables independientes, y por otro lado usaremos un boxplot para tomar la decision sobre que se puede considerar un valor extremo.

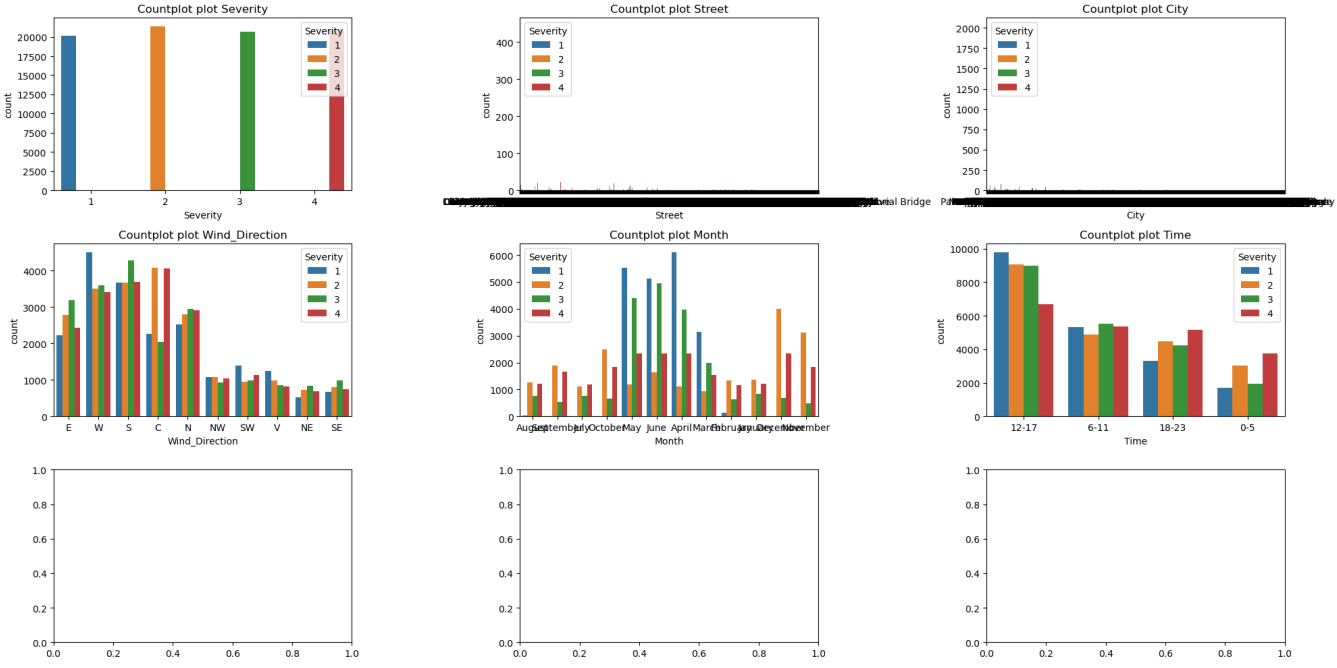
Para ciertas variables consideraremos outliers todo lo que se encuentra alejado 3 desviaciones estándar con respecto a la media del conjunto es un outlier.



<Figure size 2000x600 with 0 Axes>


Severity
1
2
3
4
Distance(mi)
Temperature(F)
Wind_Chill(F)
Humidity(%)
Pressure(in)
Visibility(mi)
Wind_Speed(mph)
Precipitation(in)
Year
Duration
1e6





Vemos que hay una gran cantidad de outliers sobre todo para los accidentes con severidad 2. Se va a limpiar los outliers usando la desviacion estandart y luego se volvera a comparar.

4. Análisis de los datos

4.1 Seleccion de Grupos

Selección de los grupos de datos que se quieren analizar/comparar (p.ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?

Dado que en el apartado 4.3 se va realizar un analisis de agrupacion y un analisis de clasificacion, separaremos el dataset en 2 datasets distintos con las variables necesarias para el analisis correspondiente.

1. Analisis de la regresion usaremos las siguientes variables atmosfericas:

    - Distance(mi)
    - Temperature(F)
    - Wind_Chill(F)
    - Humidity(%)
    - Pressure(in)
    - Visibility(mi)
    - Wind_Direction
    - Wind_Speed(mph)
    - Precipitation(in)
    - Weather_Condition

    Se aplicara un modelo de regresion logistica multinomial, que incluira las variables citadas como predictoras del nivel de severidad de los accidentes.

1. En el analisis del clustering nos interesa saber los puntos geograficos, para observar donde se agrupan la mayoria de accidentes de trafico. Se usan las siguientes variables:

    - Start_Lat
    - Start_Lng

4.2 Comprobar Normalidad

Comprobación de la normalidad y homogeneidad de la varianza

Normalidad

Como se observa el sample size es mucho mayor que 100 por lo cual hay que acudir a un test distinto del Shapiro-Wilk Test para testear la normalidad. Por lo tanto recurrimos al test de Kolmogorov–Smirnov indicado para samples grandes.

Tal como se aplica aqui la prueba de una muestra realiza una prueba de la distribución F(x) de una variable aleatoria observada contra una distribución dada G(x) (es decir, una distribución normal).

Siendo las Hipotesis:

- Hipotesis Nula: Distribuciones identicas F(x)=G(x) para todo x
- Hipotesis Alternativa: Distribuciones no identicas

Para la realizacion de este test en particular hemos escogido un sample 100 para cada variable y asi porder testear la normalidad

	Stats	P-value	Normal	Variable
0	0.083707	4.604371e-01	False	Start_Lat
1	0.183916	1.965296e-03	False	Start_Lng
2	0.083589	4.622225e-01	False	End_Lat
3	0.183881	1.970525e-03	False	End_Lng
4	0.286710	9.037904e-08	False	Distance(mi)
5	0.082792	4.743975e-01	False	Temperature(F)
6	0.116084	1.246937e-01	False	Wind_Chill(F)
7	0.085095	4.396738e-01	False	Humidity(%)
8	0.087492	4.050550e-01	False	Pressure(in)
9	0.493671	5.083221e-23	False	Visibility(mi)
10	0.093427	3.267012e-01	False	Wind_Speed(mph)
11	0.475817	2.550479e-21	False	Precipitation(in)
12	0.285351	1.062930e-07	False	Year
13	0.453239	2.780255e-19	False	Duration

Observamos que ninguna de las variables del dataset es normal. Sin embargo, dado que tamaño del dataset es lo suficientemente grande (es decir,> 30), es posible asumir la normalidad de acuerdo con el Teorema del límite central.

homogeneidad de la varianza

Aunque asumimos que los datos siguen una distribucion normal, por el teorema del limite central. Ya que vamos a usar un sample pequeño que no sera normal usamos el test Levene para la comprobación de la homocedasticidad entre los grupos de la variable dependiente Severity.

Siendo las Hipotesis:

- Hipotesis Nula: Asume igualdad de varianzas en los diferentes grupos de datos, por lo que p-valores inferiores al nivel de significancia indicarán heterocedasticidad.
- Hipotesis Alternativa: No asume igualdad de varianzas en los diferentes grupos de datos

	W	P-value	equal_var	Variable
0	173.911815	2.093227e-112	False	Start_Lat
1	2158.699753	0.000000e+00	False	Start_Lng
2	173.744800	2.683690e-112	False	End_Lat
3	2158.585710	0.000000e+00	False	End_Lng
4	661.911845	0.000000e+00	False	Distance(mi)
5	180.077515	2.173739e-116	False	Temperature(F)
6	317.413338	6.144525e-205	False	Wind_Chill(F)
7	559.327543	0.000000e+00	False	Humidity(%)
8	259.190867	1.918641e-167	False	Pressure(in)
9	208.548892	8.922352e-135	False	Visibility(mi)
10	7.612379	4.372863e-05	False	Wind_Speed(mph)
11	38.841393	4.501355e-25	False	Precipitation(in)
12	5380.426586	0.000000e+00	False	Year
13	46.709993	3.755722e-30	False	Duration

Observamos que ninguna de las variables tiene igualdad de varianza entre los grupos de Severidad. Esto en parte se puede deber al sample size del dataset, ya que entre mas datos mayor sera la variabilidad entre los grupos.

Sin embargo, La regresión logística no hace muchas de las suposiciones clave de la regresión lineal y los modelos lineales generales que se basan en algoritmos de mínimos cuadrados ordinarios, en particular con respecto a la linealidad, la normalidad, la homocedasticidad y el nivel de medición.

Por lo tanto no aplicaremos ninguna transformacion como cuadrada,logistica,etc sobre los datos.

https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

4.3 Pruebas Estadisticas

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Analisis de correlacion

Análisis de los factores que influyen en la severidad. Correlaciones y regresión logística multinomial.

	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Direction	Wind
2	0.0	68.0	68.0	93.0	29.14	9.0	E	
3	0.0	98.0	98.0	18.0	28.28	10.0	W	
4	0.0	95.0	95.0	22.0	29.90	10.0	W	
5	0.0	81.0	81.0	79.0	29.20	10.0	S	
6	0.0	71.0	71.0	81.0	27.81	10.0	C	

	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_
Distance(mi)	1.000000	-0.053888	-0.057526	0.068994	0.001495	-0.036577	
Temperature(F)	-0.053888	1.000000	0.994832	-0.447813	-0.097641	0.279483	
Wind_Chill(F)	-0.057526	0.994832	1.000000	-0.432616	-0.088786	0.285641	
Humidity(%)	0.068994	-0.447813	-0.432616	1.000000	0.207449	-0.409717	
Pressure(in)	0.001495	-0.097641	-0.088786	0.207449	1.000000	-0.014960	
Visibility(mi)	-0.036577	0.279483	0.285641	-0.409717	-0.014960	1.000000	
Wind_Speed(mph)	-0.014219	0.114223	0.071549	-0.195245	-0.036661	0.032860	
Precipitation(in)	0.002098	-0.031880	-0.031853	0.139950	-0.002785	-0.228823	

Tras normalizar los datos y calcular los coeficientes de correlación observamos como *Temperature(F)* y *Wind_Chill(F)* están totalmente correlacionados.

A continuación comprobaré como se correlacionan las variables categóricas *weather_condition* y *wind_direction* con el resto de variables numéricas por medio del cálculo de ANOVAs.

```
ANOVAs for "Wind_Direction"
P-value for Wind_Direction and Distance(mi) ANOVA is 0.0009698818959663366
Rejected H0, both are correlated (95%)
P-value for Wind_Direction and Temperature(F) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Wind_Direction and Wind_Chill(F) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Wind_Direction and Humidity(%) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Wind_Direction and Pressure(in) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Wind_Direction and Visibility(mi) ANOVA is 1.3614657032668732e-275
Rejected H0, both are correlated (95%)
P-value for Wind_Direction and Wind_Speed(mph) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Wind_Direction and Precipitation(in) ANOVA is 2.3853159882842742e-29
Rejected H0, both are correlated (95%)
ANOVAs for "Weather_Condition"
P-value for Weather_Condition and Distance(mi) ANOVA is 3.008245014445999e-94
Rejected H0, both are correlated (95%)
P-value for Weather_Condition and Temperature(F) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Weather_Condition and Wind_Chill(F) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Weather_Condition and Humidity(%) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Weather_Condition and Pressure(in) ANOVA is 2.735808679477743e-131
Rejected H0, both are correlated (95%)
P-value for Weather_Condition and Visibility(mi) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Weather_Condition and Wind_Speed(mph) ANOVA is 0.0
Rejected H0, both are correlated (95%)
P-value for Weather_Condition and Precipitation(in) ANOVA is 0.0
Rejected H0, both are correlated (95%)
```

Correlación entre variables categóricas (V de Cramer)

```
0.09187909776290026
```

Este valor cercano a 0 nos indica la poca correlación entre ambas variables categóricas.

Analisis de Regresion

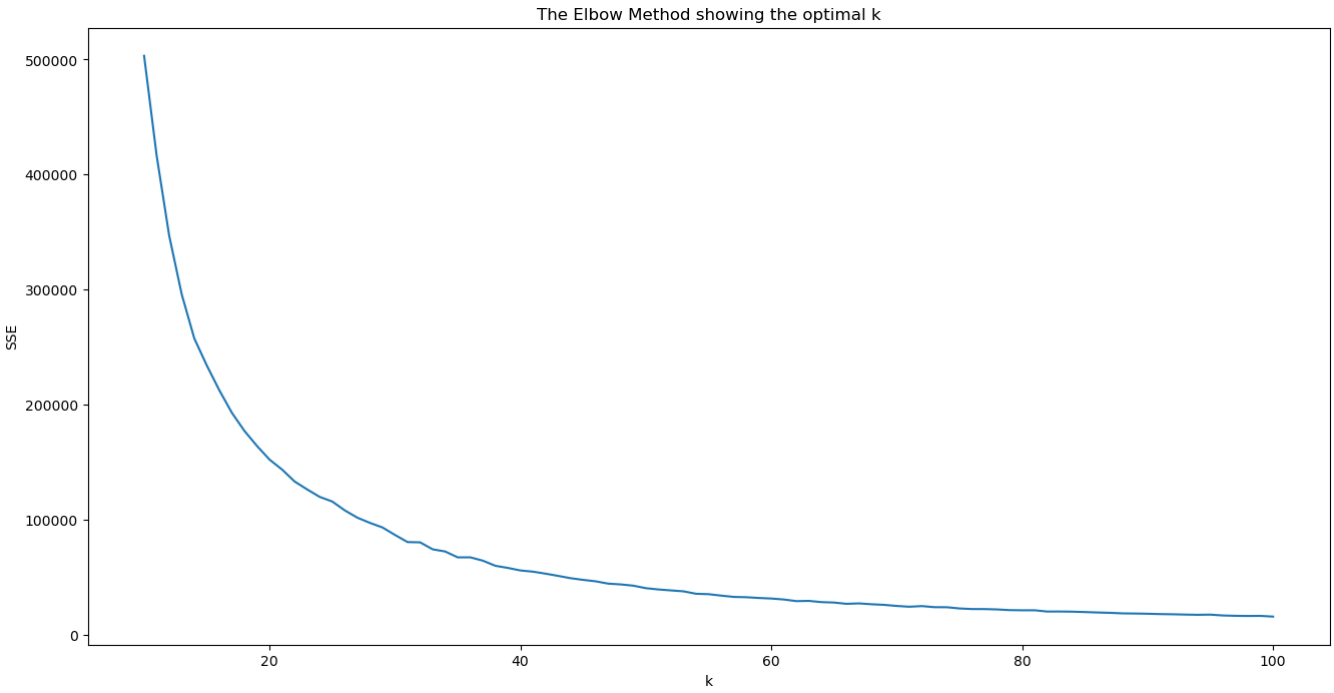
Aplicaré una regresión logística multinomial para comprobar como se relacionan estas variables con la severidad del accidente. Para las variables altamente correlacionadas puede omitirse una de ellas. Para aplicar correctamente regresiones logísticas e interpretarlas se ha seguido lo descrito en el siguiente enlace: https://www.datasklr.com/logistic-regression/multinomial-logistic-regression

```
Optimization terminated successfully.
         Current function value: 1.311815
         Iterations 7
```

Se coge como clase base la severidad nivel 1. A partir de esta se crean tres conjuntos de coeficientes para evaluar e interpretar las *odds* de que la severidad sea la clase base o la que se compara con esta. Cada coeficiente se corresponde con las variables introducidas en el modelo. En el primer caso (Severity = 0), aunque algo confuso, se está comparando la severidad de nivel 1 con la de nivel 2. La distancia en millas tiene un p-valor inferior a 0.05 por lo que es significativo a la hora de predecir la severidad del

accidente. Además, tiene un coeficiente de 1.18 en este caso y esto implica que por cada milla que aumente la distancia el accidente es 1.18 veces más posible que el accidente sea de severidad 2.

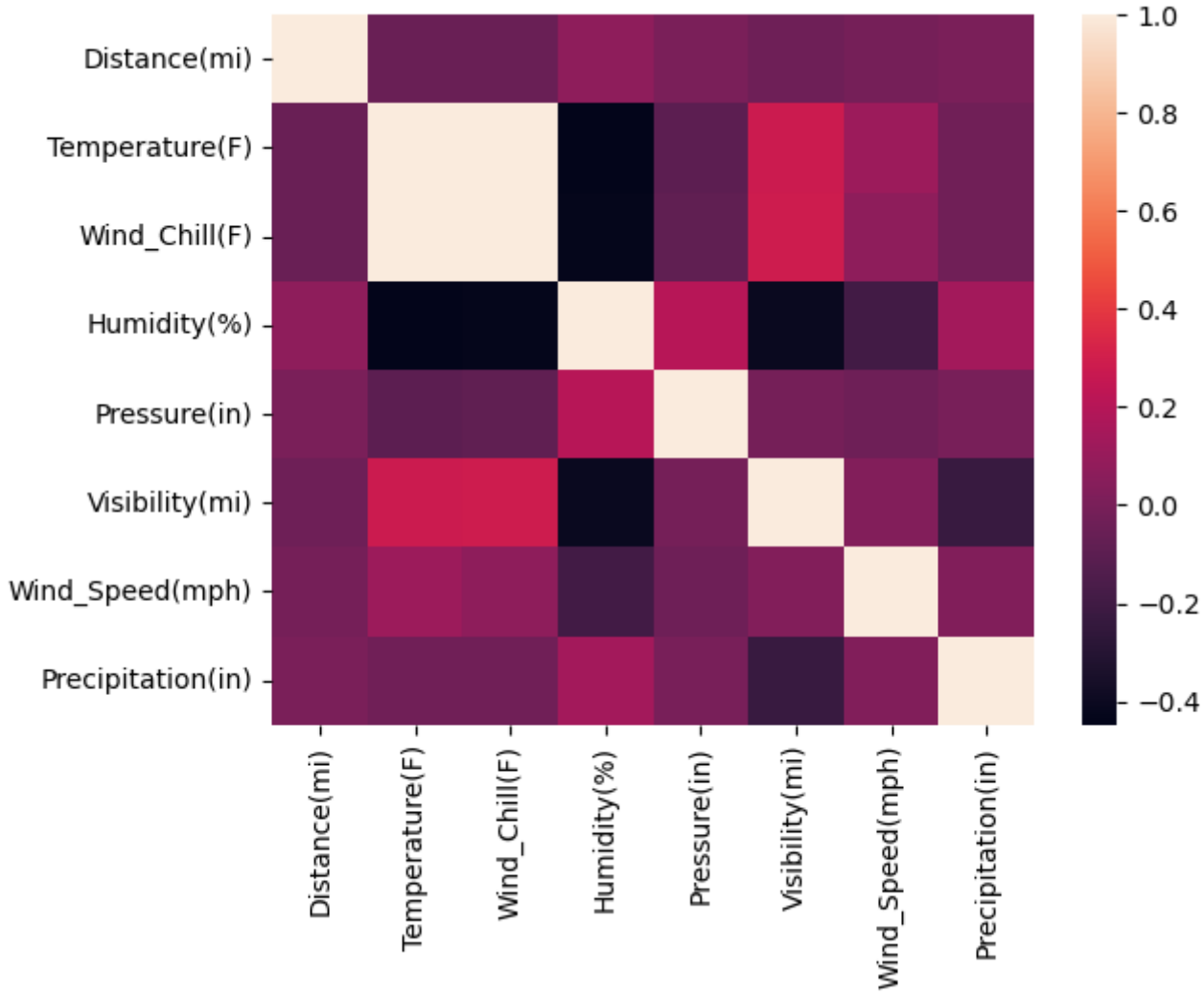
Analisis de agrupacion



5. Representación de los resultados

Correlacion Visualizacion

```
<AxesSubplot:>
```



multinomial Regression visualizacion

MNLogit Regression Results

Dep. Variable:	Severity	No. Observations:	66498
Model:	MNLogit	Df Residuals:	66474
Method:	MLE	Df Model:	21
Date:	Mon, 09 Jan 2023	Pseudo R-squ.:	0.05358
Time:	22:11:39	Log-Likelihood:	-87233.
converged:	True	LL-Null:	-92171.
Covariance Type:	nonrobust	LLR p-value:	0.000

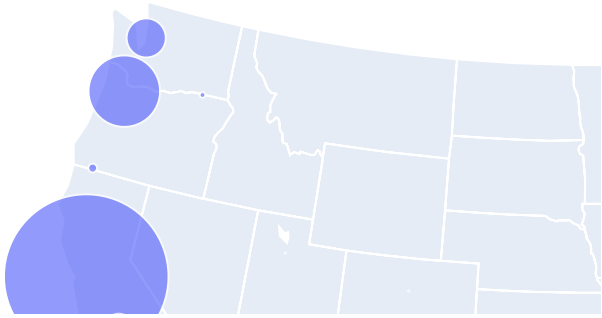
Severity=2	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2800	0.013	20.816	0.000	0.254	0.306
Distance(mi)	1.1800	0.033	35.233	0.000	1.114	1.246
Temperature(F)	-0.2891	0.013	-21.682	0.000	-0.315	-0.263
Humidity(%)	0.2026	0.014	14.531	0.000	0.175	0.230
Pressure(in)	0.1561	0.012	12.891	0.000	0.132	0.180
Visibility(mi)	-0.0158	0.015	-1.081	0.280	-0.044	0.013
Wind_Speed(mph)	-0.1819	0.012	-15.302	0.000	-0.205	-0.159
Precipitation(in)	-0.1116	0.019	-5.925	0.000	-0.149	-0.075

Severity=3	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2453	0.014	18.121	0.000	0.219	0.272
Distance(mi)	0.9728	0.034	28.375	0.000	0.906	1.040
Temperature(F)	-0.2316	0.013	-17.314	0.000	-0.258	-0.205
Humidity(%)	0.3191	0.014	22.771	0.000	0.292	0.347
Pressure(in)	-0.1258	0.012	-10.870	0.000	-0.148	-0.103
Visibility(mi)	-0.0639	0.014	-4.473	0.000	-0.092	-0.036
Wind_Speed(mph)	0.1850	0.012	16.045	0.000	0.162	0.208
Precipitation(in)	-0.0082	0.013	-0.653	0.514	-0.033	0.016

Severity=4	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2181	0.014	15.990	0.000	0.191	0.245
Distance(mi)	1.4091	0.033	42.300	0.000	1.344	1.474
Temperature(F)	-0.4600	0.014	-33.962	0.000	-0.487	-0.433
Humidity(%)	0.3615	0.014	25.005	0.000	0.333	0.390
Pressure(in)	-0.0898	0.012	-7.488	0.000	-0.113	-0.066
Visibility(mi)	0.0419	0.014	2.915	0.004	0.014	0.070
Wind_Speed(mph)	-0.1102	0.012	-9.115	0.000	-0.134	-0.087
Precipitation(in)	-0.1455	0.020	-7.204	0.000	-0.185	-0.106

Cluster visualizacion

US accident focus



6. Resolución del problema

A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

1. Resultados de las correlaciones

   - Tras el análisis observamos que estan altamente correlacionados la temperatura (Temperature(F)) y la sensacion térmica influida por el viento (Wind_Chill(F)). Por lo cual para la futura prediccion de la severidad de los accidentes podemos omitir una de ellas como variable independiente/predictor.

   - Por otro lado, las variables de direccion del viento ('Wind_Direction') y la percepcion general del tiempo ('Weather_Condition') estan correlacionadas con todas las variables meteorologicas ya que el conjunto de variables da una percepcion/clasificacion general del tiempo.

1. Resultados de las regresiones

   - Se observa, en la regresion, para los factores de severidad 1 & severidad 2 de la variable dependiente (Severidad) que el predictor visibilidad no es significativo.

- Ademas los factores distancia, humedad y presion tienen un efecto positivo en la probabilidad de ocurrencia del factor de severidad 2, mientras que las variables independientes como Temperatura, velocidad del viento y precipitacion tienen un efecto inverso en la ocurrencia de este factor.
- Para la comparativa entre el factor de severidad 1 & 3, no resulta significativa la variable precipitacion, y son las variables distancias y humedad las que presentan un efecto positivo, mientras que el resto presentan un efecto negativo.
- En el caso de los niveles de severidad de 1 y 4, todas las variables son significativas. Presentando un efecto positivo la distancia, la humedad y la visibilidad; siendo el resto de predictores negativos.

2. Resultados del clustering

- Se observa que el numero optimo de cluster se encuentra entorno a 40-60 grupos. Esto tiene cierta similitud con el numero de estados que conforman EEUU, por lo tanto el numero de agrupacion elegido sera 50.
- Se observa una mayor concentracion de accidentes en las carreteras correspondientes a estados que lindan con la costa. Por ejemplo Florida y California