UNIVERSIDAD NACIONAL DE INGENIERÍA

OTI UNI

CIENCIA DE DATOS I:

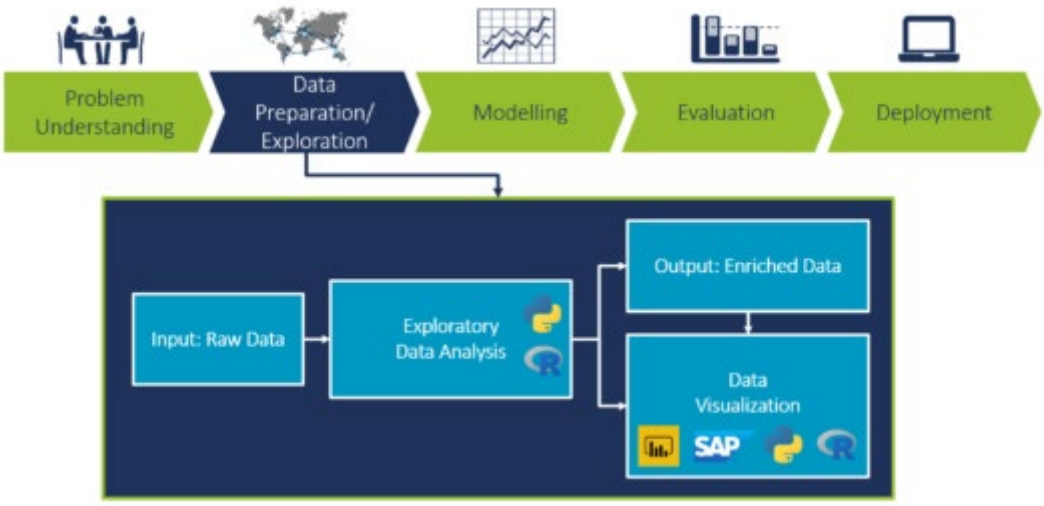
SESION 02

EXPLORATORY DATA ANALYSIS

Todas las imágenes pertenecen a sus autores, se usan solo con fines educativos

ÍNDICE

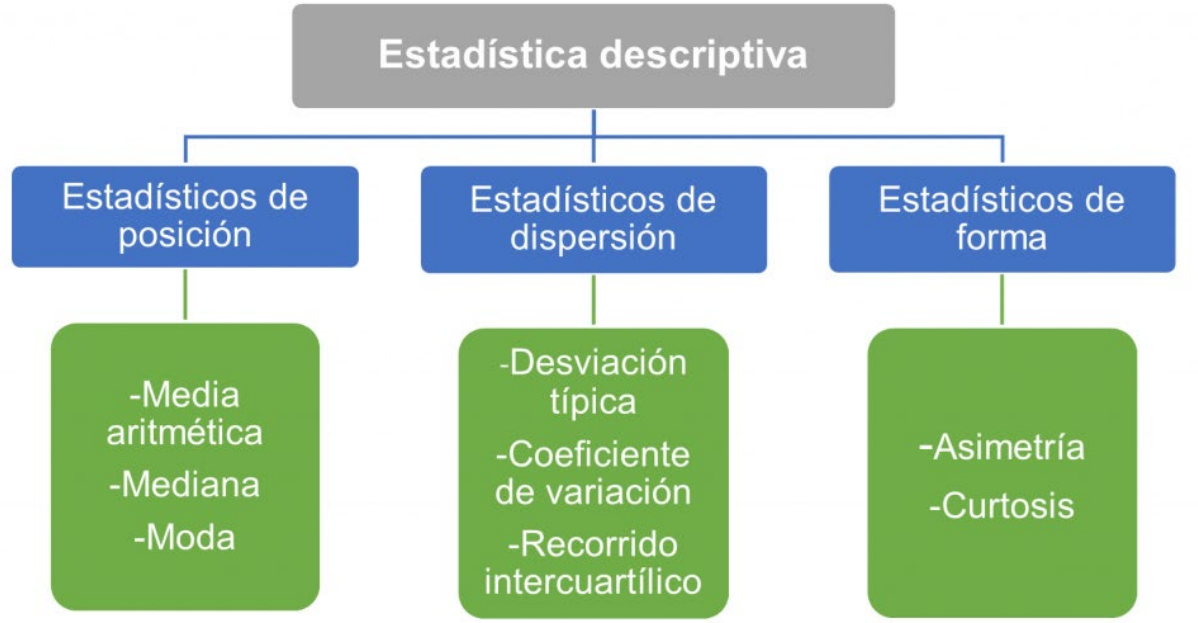
OBJETIVO



El objetivo de esta semana es realizar una revisión de la estadística descriptiva y aplicarla al Análisis Exploratorio de Datos (EDA)

En esta segunda sesión la práctica de laboratorio consistirá en llevar a cabo un análisis exploratorio de datos de la tragedia del Titanic para ello utilizaremos las principales librerías de Python, como son: numpy, pandas, scikit learn, matplotlib.

INTRODUCCION



La estadística descriptiva es el término que se le da al análisis de datos que ayuda a describir, mostrar o resumir datos de una manera significativa de modo que puedan surgir patrones a partir de dichos datos, eso se realiza con la ayuda de gráficos o valores de resumen. Sin embargo, la estadística descriptiva no nos permitirá sacar conclusiones más allá de los datos que hemos analizado ni llegar a conclusiones con respecto a las hipótesis que podríamos haber hecho. Es simplemente una manera de describir nuestros datos.

La estadística descriptiva es muy importante porque si presentamos nuestros datos sin procesar, sería difícil visualizar lo que muestran los datos, especialmente si son muchos. Por lo tanto, la estadística descriptiva nos permite presentar los datos de una manera más significativa, lo que permitirá una interpretación más sencilla de los datos. Por ejemplo, si tenemos los resultados de 100 exámenes de un curso, podríamos estar interesados en el desempeño general de los estudiantes. También estaríamos interesados en la distribución o difusión de las notas. La estadística descriptiva nos permite hacer esto.

ANÁLISIS EXPLORATORIO DE DATOS - EDA

Antes de realizar análisis de datos, con fines estadísticos o predictivos, usando por ejemplo técnicas de aprendizaje automático, es necesario entender la materia prima (raw data) con la que vamos a trabajar. Es necesario comprender y evaluar la calidad de los datos para, entre otros aspectos, detectar y tratar los datos atípicos (outliers) o incorrectos, evitando posibles errores que puedan repercutir en los resultados del análisis.

EDA consiste en aplicar un conjunto de técnicas estadísticas destinadas a explorar, describir y resumir la naturaleza de los datos, de forma que podamos entender claramente como están relacionadas nuestras variables de interés.

Todo esto nos permite identificar posibles errores, revelar la presencia de outliers, comprobar la relación entre variables (correlaciones) y su posible redundancia, y realizar un análisis descriptivo de los datos mediante representaciones gráficas y resúmenes de los aspectos más significativos.

MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central son formas de describir la posición central de una distribución de frecuencia para un grupo de datos. En nuestro ejemplo anterior, la distribución de frecuencias es simplemente la distribución y el patrón de calificaciones obtenidas por los 100 estudiantes desde la más baja hasta la más alta. Podemos describir esta posición central utilizando una serie de estadísticas, incluida la moda, la mediana y la media.

Una medida de tendencia central es un valor único que intenta describir un conjunto de datos mediante la identificación de la posición central dentro de ese conjunto de datos. Como tales, las medidas de tendencia central a veces se denominan medidas de ubicación central. También se clasifican como estadísticas de resumen. La media (a menudo llamada promedio) es probablemente la medida de tendencia central que más conocemos, pero hay otras, como la mediana y la moda.

La media, la mediana y la moda son todas medidas válidas de tendencia central, pero en diferentes condiciones, algunas medidas de tendencia central se vuelven más apropiadas que otras. En las siguientes secciones, veremos la media, la moda y la mediana, y aprenderemos cómo calcularlos y en qué condiciones son más apropiados para su uso.

Media

La media (o promedio) es la medida de tendencia central más popular y conocida. Se puede usar tanto con datos discretos como continuos, aunque su uso es más frecuente con datos continuos. La media es igual a la suma de todos los valores del conjunto de datos dividida por el número de valores del conjunto de datos.

Media de la muestra:

$$\overline{x} = \frac{\sum x}{n}$$

Media de la población:

$$\mu = \frac{\sum x}{n}$$

Una propiedad importante de la media es que incluye todos los valores de su conjunto de datos como parte del cálculo. Además, la media es la única medida de tendencia central donde la suma de las desviaciones de cada valor de la media es siempre cero.

La media tiene una desventaja principal: es particularmente susceptible a la influencia de valores atípicos. Estos son valores que son inusuales en comparación con el resto del conjunto de datos por ser especialmente pequeños o grandes en valor numérico.

Mediana

La mediana es la puntuación media de un conjunto de datos que se ha organizado en orden de magnitud. La mediana se ve menos afectada por los valores atípicos y los datos sesgados.

Para calcular la mediana, supongamos que tenemos los siguientes datos (número impar de datos):

65	55	89	56	35	14	56	55	87	45	92

Primero ordenamos los datos en orden ascendente:

14	35	45	55	55	**56**	56	65	87	89	92

En el siguiente ejemplo tenemos un numero par de datos:

65	55	89	56	35	14	56	55	87	45

Primero ordenamos los datos en orden ascendente:

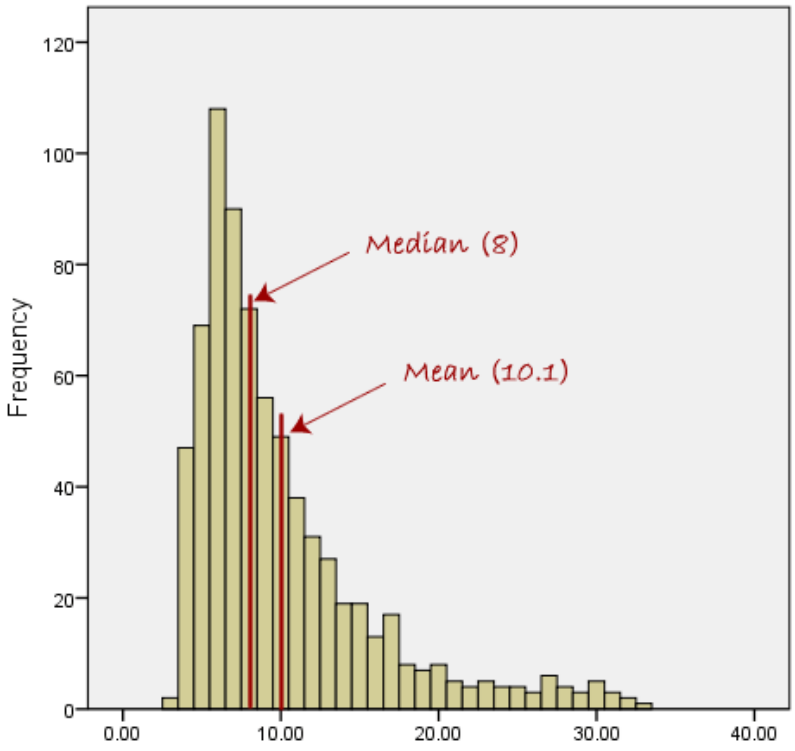
14	35	45	55	**55**	**56**	56	65	87	89

La mediana será el resultado del promedio de las notas 5 y 6 es decir: 55.5

Preferiremos usar la mediana sobre la media (o la moda) cuando nuestros datos están sesgados (es decir, la distribución de frecuencia de nuestros datos está sesgada). Si consideramos la distribución normal, ya que es la más evaluada en estadística, cuando los datos son perfectamente normales, la media, la mediana y la moda son idénticas. Además, todos representan el valor más típico en el conjunto de datos. Sin embargo, a medida que los datos se vuelven sesgados, la media pierde su capacidad de proporcionar la mejor ubicación

central para los datos porque los datos sesgados los alejan del valor típico. Sin embargo, la mediana conserva mejor esta posición y no está tan fuertemente influenciada por los valores sesgados.

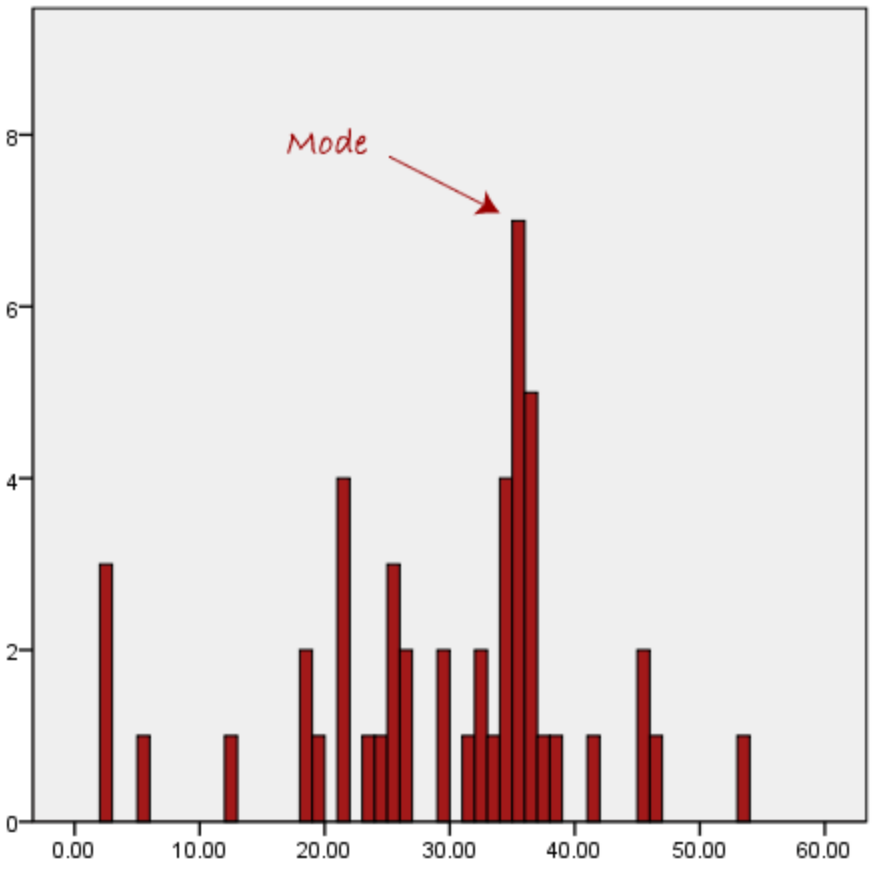
En el caso de datos sesgados, en la siguiente imagen encontramos que la media está siendo arrastrada en el sentido directo del sesgo. En estas situaciones, generalmente se considera que la mediana es el mejor representante de la ubicación central de los datos. Cuanto más sesgada sea la distribución, mayor será la diferencia entre la mediana y la media, y se debe poner mayor énfasis en usar la mediana en lugar de la media. Un ejemplo clásico de la distribución sesgada hacia la derecha salario, donde los que ganan más brindan una representación falsa del ingreso típico si se expresan como una media y no como una mediana.



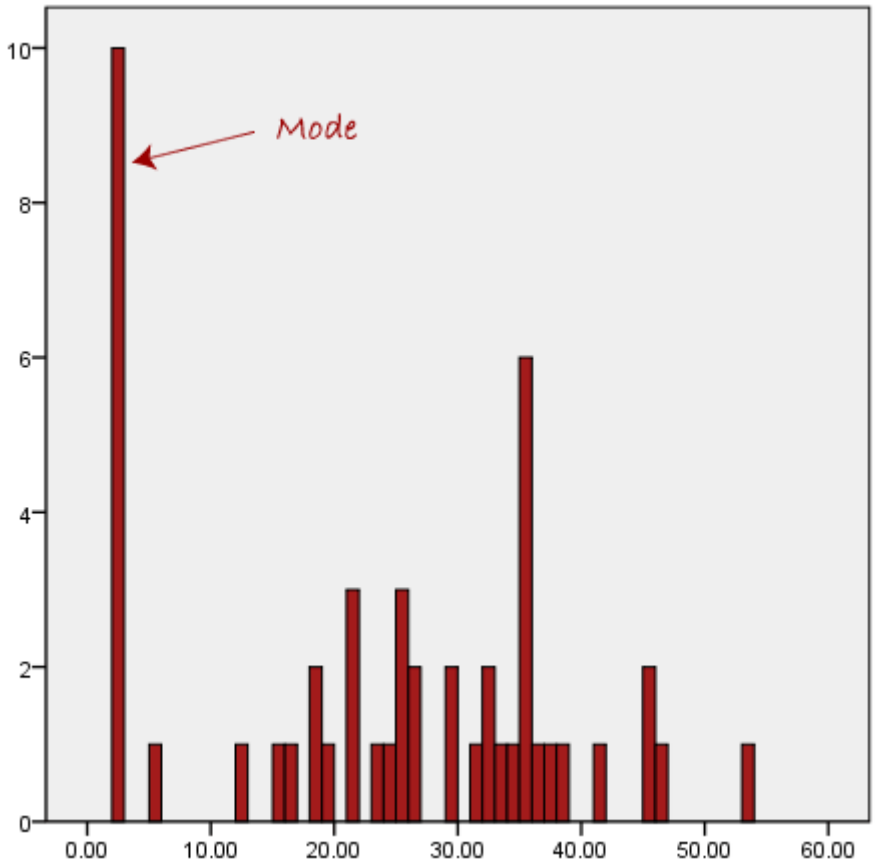
Moda

La moda es la puntuación más frecuente en nuestro conjunto de datos. En un histograma representa la barra más alta en un gráfico de barras o histograma. Por lo tanto, a veces puede considerar la moda como la opción más popular.

A continuación, se presenta un ejemplo de una moda:



Un problema con la moda es que no nos proporcionará una muy buena medida de tendencia central cuando la marca más común está lejos del resto de los datos en el conjunto de datos, como se muestra en el siguiente diagrama:



MEDIDAS DE DISPERSIÓN

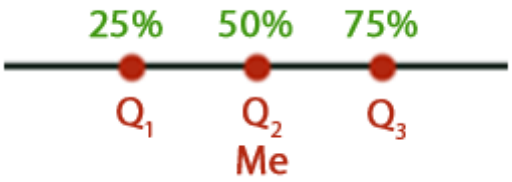
Introducción

Las medidas de dispersión, se utiliza para describir la variabilidad en una muestra o población. Por lo general, se usa junto con una medida de tendencia central, como la media o la mediana, para proporcionar una descripción general de un conjunto de datos.

Por ejemplo, la puntuación media de los 100 alumnos puede ser de 65 sobre 100. Sin embargo, no todos los alumnos habrán obtenido 65 puntos. Más bien, sus puntajes se distribuirán. Unos serán más bajos y otros más altos. Las medidas de dispersión nos ayudan a resumir cuán dispersas están estas puntuaciones. Para describir este diferencial, tenemos a nuestra disposición una serie de estadísticas, algunas de ellas son los cuartiles, la varianza, la desviación standard y la correlación.

Cuartiles

Los cuartiles nos informan sobre la dispersión de un conjunto de datos dividiéndolo en cuartos, al igual que la mediana lo divide por la mitad. Por ejemplo, considere las calificaciones de los 100 estudiantes, a continuación, que se han ordenado de la calificación más baja a la más alta. En este caso datos no agrupados.



$$Q_1 = x_i + d \cdot (x_{i+1} - x_i)$$

Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

N = 100;
Primer Cuartil => (N + 1)/4 = 25.25;
Segundo Cuartil => 2 (N + 1)/4 = 50.50;
Tercer Cuartil => 3(N + 1)/4 = 75.75
Q1 = 55 + 0.25(45 – 45) = 25
Q2 = 58 + 0.50(59 – 58) = 58.5
Q3 = 71 + 0.75(71 – 71) = 71

Varianza

Los cuartiles son útiles, pero también son algo limitados porque no tienen en cuenta todas las notas de nuestro grupo de datos. Para tener una idea más representativa de la dispersión, debemos tener en cuenta los valores reales de cada puntaje en un conjunto de datos. La varianza y la desviación estándar son tales medidas.

La varianza alcanza valores positivos elevando al cuadrado cada una de las desviaciones. La suma de estas desviaciones al cuadrado nos da la suma de los cuadrados, que luego podemos dividir por el número total de notas en nuestro grupo de datos (en otras palabras, 100 porque hay 100 estudiantes) para encontrar la varianza. Por lo tanto, para nuestros 100 estudiantes, la varianza es 211,89, como se muestra a continuación:

$$variance = \frac{\sum(X-\mu)^2}{N}$$

$$= \frac{21188.75}{100}$$

$$= 211.89$$

Where μ = mean, X = score, ∑= the sum of, N = number of scores, ∑X = "add up all the scores"

Como medida de variabilidad, la varianza es útil. Si las notas en nuestro grupo de datos están muy dispersas, la varianza será un número grande. Por el contrario, si las notas se distribuyen muy cerca de la media, la varianza será un número menor. Sin embargo, hay dos problemas potenciales con la varianza. En primer lugar, debido a que las desviaciones de las notas con respecto a la media se elevan al cuadrado, esto da más peso a las puntuaciones extremas. Si nuestros datos contienen valores atípicos (en otras palabras, uno o un pequeño número de puntajes que están particularmente lejos de la media y quizás no representan bien nuestros datos en su conjunto), esto puede deshacer el peso de estas notas. En segundo lugar, la varianza no está en las mismas unidades que las puntuaciones en nuestro conjunto de datos: la varianza se mide en unidades al cuadrado. Esto significa que no podemos ubicarlo en nuestra distribución de frecuencia y no podemos relacionar directamente su valor con los valores de nuestro conjunto de datos. Por lo tanto, la cifra de 211,89, nuestra varianza, parece algo arbitraria. Calcular la desviación estándar en lugar de la varianza corrige este problema. No obstante, el análisis de la varianza es extremadamente importante en algunos análisis estadísticos.

Desviación Standard

La desviación estándar es una medida de la dispersión de puntajes dentro de un conjunto de datos. Por lo general, estamos interesados en la desviación estándar de una población. Sin embargo, como a menudo se nos presentan datos de una muestra solamente, podemos estimar la desviación estándar de la población a partir de una desviación estándar de la muestra. Estas dos desviaciones estándar (desviaciones estándar de la muestra y de la población) se calculan de manera diferente. En estadística, generalmente se nos presenta el tener que calcular las desviaciones estándar de la muestra, aunque también se mostrará la fórmula para una desviación estándar de la población.

La desviación estándar se usa junto con la media para resumir datos continuos, no datos categóricos. Además, la desviación estándar, como la media, normalmente solo es adecuada cuando los datos continuos no están significativamente sesgados o tienen valores atípicos.

Fórmula de la desviación standard para una muestra:

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

Fórmula de la desviación estándar para una población:

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{n}}$$

Covarianza

La desviación típica es un indicador de dispersión de una variable. ¿Qué pasa cuando tienes más de una variable? ¿Existe alguna forma de saber cómo se relaciona una con la otra?

La Covarianza es la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas.

$$\sigma_{XY} = \sum_{i=1}^{N} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{N}$$

La covarianza positiva: Cuando uno variable crece la otra variable también. Tienen una relación directa.

La covarianza negativa: Cuando una variable crece la otra variable decrece. Tienen una relación Inversa.

Coeficiente de Correlación

La correlación es un indicador para saber si hay relación (LINEAL) entre dos variables numéricas para esto se utiliza el coeficiente de correlación o correlación de Pearson.

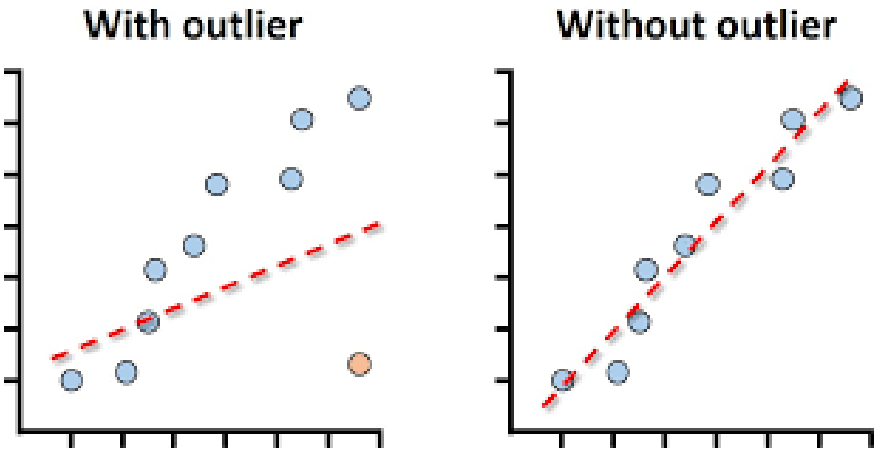
La correlación contesta preguntas como las siguientes:

- ¿La práctica de algún deporte está relacionada con una vida más longeva?
- ¿Existe una relación entre la cantidad de carne ingerida diariamente y el cáncer?
- ¿Mayor estudio implica mejores notas en un examen?

La correlación es un ratio entre la dispersión entre las dos variables conjuntamente (covarianza) y la desviación standard de cada variable.

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

OUTLIERS



Ya sabemos que el análisis exploratorio de datos (EDA) es crucial cuando se trabaja en proyectos de ciencia de datos. Conocer sus datos por dentro y por fuera puede simplificar la toma de decisiones con respecto a la selección de características, algoritmos e hiperparámetros.

Uno de los pasos más importantes como parte del preprocesamiento de datos es detectar y tratar los *valores atípicos*, ya que pueden afectar negativamente el análisis estadístico y el proceso de entrenamiento de un algoritmo de aprendizaje automático, lo que resulta en una menor precisión.

La detección de valores atípicos o anomalías es uno de los problemas centrales en la minería de datos. La expansión emergente y el crecimiento continuo de los datos y la difusión de los dispositivos IoT nos hacen repensar la forma en que abordamos las anomalías y los casos de uso que se pueden construir al observar esas anomalías.

Ahora tenemos relojes y pulseras inteligentes que pueden detectar los latidos de nuestro corazón cada poco minuto. La detección de anomalías en los datos de los *latidos del corazón* puede ayudar a predecir enfermedades del corazón. Las anomalías en los patrones de tráfico pueden ayudar a predecir accidentes. También se puede utilizar para identificar *cuellos de botella* en la infraestructura de red y el tráfico entre servidores. Por lo tanto, los casos de uso y la solución construida sobre la detección de anomalías son ilimitados.

Otra razón por la que necesitamos detectar anomalías es que, al preparar conjuntos de datos para modelos de aprendizaje automático, es realmente importante detectar todos los valores atípicos y deshacerse de ellos o analizarlos para saber por qué los tenía allí en primer lugar.

¿QUE ES UN OUTLIER?

En estadística, los valores atípicos son puntos de datos que no pertenecen a una determinada población. Es una observación anormal que se encuentra muy lejos de otros valores. Un valor atípico es una observación que diverge de los datos bien estructurados.

Por ejemplo, puede ver claramente el valor atípico en esta lista:

[20, 24, 22, 19, 29, 18, 4300, 30, 18]

Es fácil identificarlo cuando las observaciones son solo unos pocos números y son unidimensionales, pero cuando tiene miles de observaciones o multidimensionales, necesitará formas más inteligentes de detectar esos valores.

Esto es lo que cubriremos en esta sesión.

¿Por qué ocurren?

Un outlier puede ocurrir debido a la variabilidad en los datos o debido a un error experimental/error humano.

Pueden indicar un error experimental o una gran asimetría en los datos (distribución de cola pesada).

¿Qué afectan?

En estadística, tenemos tres medidas de tendencia central, a saber, Media, Mediana y Moda. Ellas nos ayudan a describir los datos.

Observemos el siguiente grupo de datos, ¿qué podemos deducir de él?:

[15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10 , 9]

with outlier	without outlier
Mean: 20.08	Mean: 12.72
Median: 14.0	Median: 13.0
Mode: 15	Mode: 15
Variance: 614.74	Variance: 21.28
Std dev: 24.79	Std dev: 4.61

¿QUE ES UN OUTLIER?

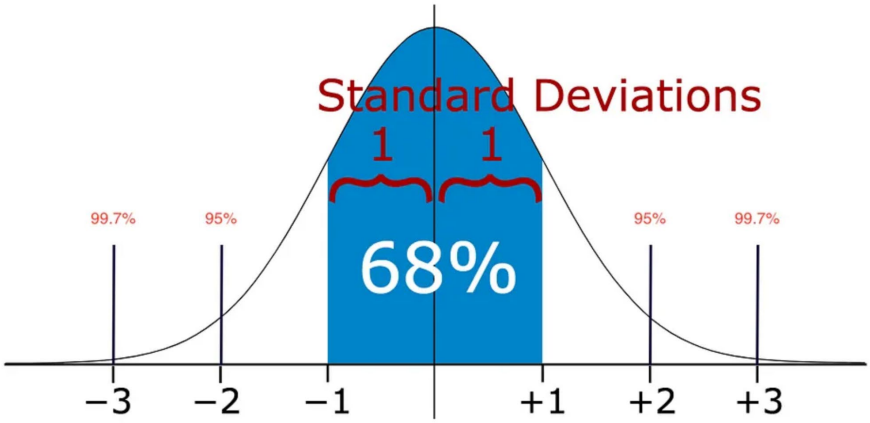
La media es la medida adecuada para describir los datos cuando no tenemos valores atípicos presentes. La mediana se usa si hay un valor atípico en el conjunto de datos. La moda se usa si hay un valor atípico y aproximadamente la mitad o más de los datos son iguales.

La "media" es la única medida de tendencia central que se ve afectada por los valores atípicos, lo que a su vez afecta la desviación estándar.

DETECTANDO OULIERS

Método 1 — Desviación estándar (Z-scores):

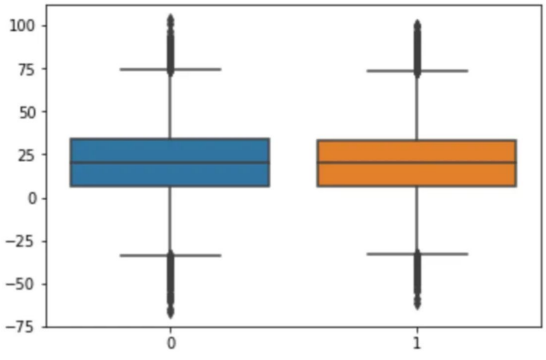
En estadística, si una distribución de datos es aproximadamente normal, aproximadamente el 68 % de los valores de los datos se encuentran dentro de una desviación estándar de la media, aproximadamente el 95 % se encuentran dentro de dos desviaciones estándar y aproximadamente el 99,7 % se encuentran dentro de tres desviaciones estándar.



Por lo tanto, si tenemos algún punto de datos que es más de 3 veces la desviación estándar, es muy probable que este punto sea anómalo o atípico.

Método 2 — Boxplots:

Los boxplots son una representación gráfica de datos numéricos a través de sus cuantiles. Es una forma muy simple pero efectiva de visualizar valores atípicos. Piense en los bigotes superior e inferior como los límites de la distribución de datos. Cualquier punto de datos que se muestre por encima o por debajo de los bigotes puede considerarse atípico o anómalo.

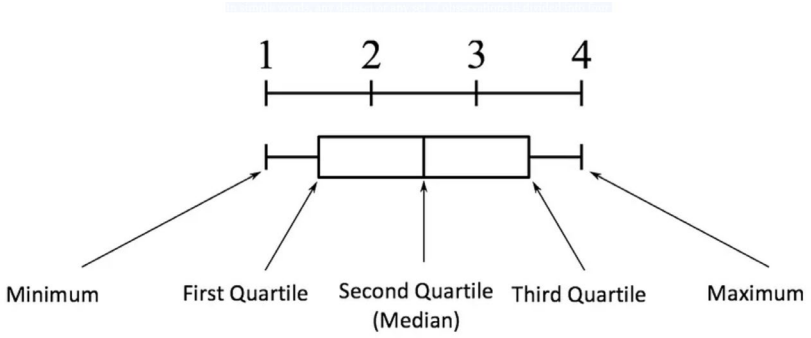


DETECTANDO OUTLIERS

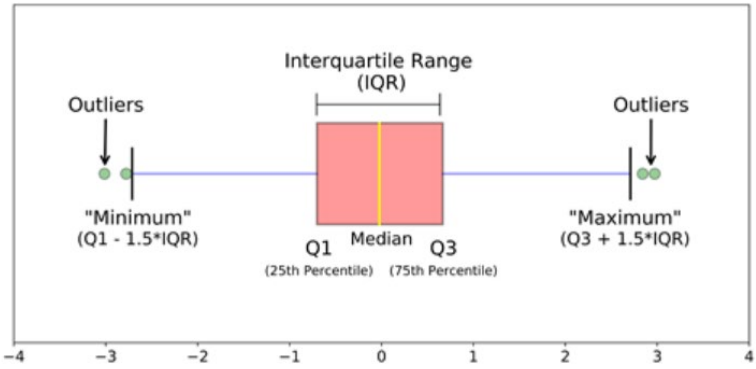
Anatomía del boxplot:

El concepto de rango intercuartílico (IQR) se utiliza para construir los gráficos de diagramas de caja. IQR es un concepto en estadística que se utiliza para medir la dispersión estadística y la variabilidad de los datos al dividir el conjunto de datos en cuartiles.

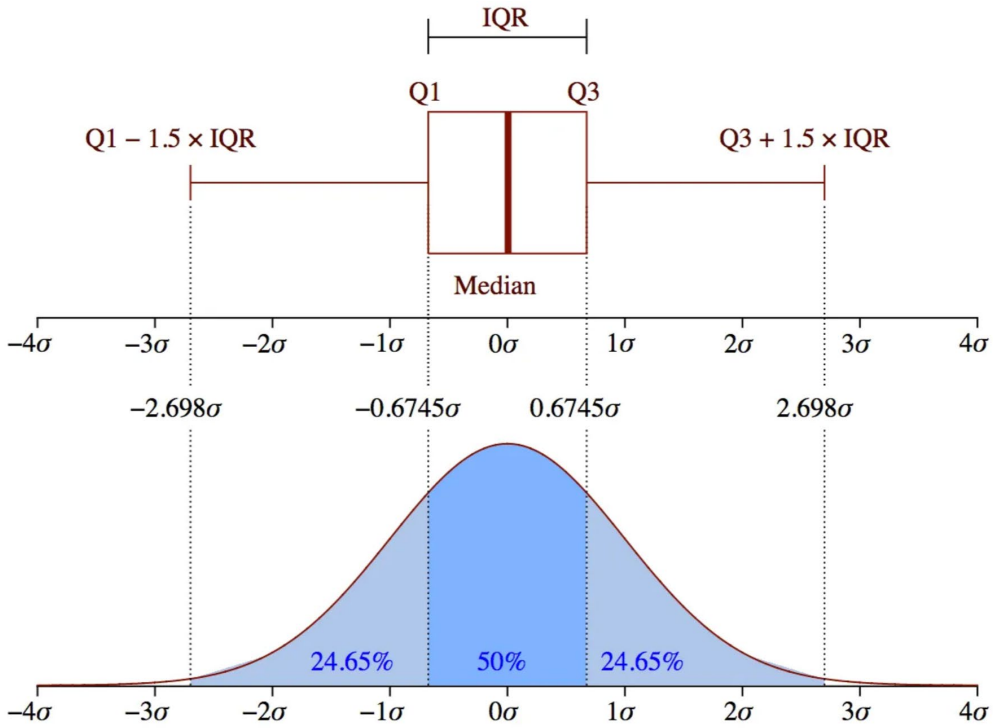
En palabras simples, cualquier conjunto de datos o cualquier conjunto de observaciones se divide en cuatro intervalos definidos según los valores de los datos y cómo se comparan con el conjunto de datos completo. Un cuartil es lo que divide los datos en tres puntos y cuatro intervalos.



El rango intercuartílico (IQR) es importante porque se utiliza para definir los valores atípicos. Es la diferencia entre el tercer cuartil y el primer cuartil (RIC = Q3 -Q1). Los valores atípicos en este caso se definen como las observaciones que están debajo (Q1 − 1.5x IQR) o el bigote inferior del diagrama de caja o arriba (Q3 + 1.5x IQR) o el bigote superior del diagrama de caja.



DETECTANDO OUTLIERS



Método 3— Agrupación de DBScan:

DBScan (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de clustering que se utiliza para agrupar puntos en un conjunto de datos según su proximidad y densidad. A diferencia de otros algoritmos de clustering, como K-means y agrupamiento jerarquico, DBSCAN no requiere que se especifique el número de clústeres de antemano.

La idea principal detrás de DBSCAN es que los puntos que están cerca unos de otros y tienen una alta densidad son parte del mismo clúster, mientras que los puntos que están aislados o tienen baja densidad se consideran ruido o puntos atípicos.

El algoritmo DBSCAN comienza seleccionando un punto aleatorio y determina si hay suficientes puntos cercanos a su alrededor para formar un clúster. Si sí, se expande a esos puntos y así sucesivamente, formando un clúster. Luego, se selecciona otro punto no visitado y se repite el proceso hasta que todos los puntos hayan sido visitados.

El algoritmo tiene dos parámetros importantes:

Epsilon (ε): especifica la distancia máxima entre dos puntos para que sean considerados vecinos.

DETECTANDO OUTLIERS

Mínimo de puntos (MinPts): especifica el número mínimo de puntos dentro de la vecindad de un punto para que sea considerado núcleo (core).

En resumen, DBSCAN encuentra áreas densas en el espacio de características y agrupa los puntos que están cerca unos de otros en clústeres, mientras que los puntos aislados se consideran ruido. Es un algoritmo muy utilizado en la minería de datos y análisis de patrones para descubrir estructuras interesantes en conjuntos de datos no etiquetados.

DBScan tiene tres conceptos importantes:

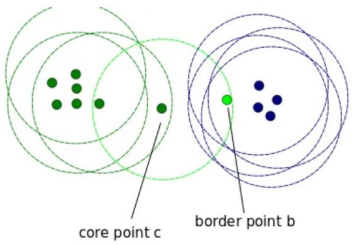
En DBSCAN, durante el proceso de clustering, los puntos se clasifican en tres tipos diferentes: puntos núcleo (core points), puntos frontera (border points) y puntos de ruido (noise points). Estos tipos se determinan en función de la densidad de puntos en su vecindad.

Puntos núcleo (core points): Son puntos que tienen al menos un número mínimo de puntos (definido por el parámetro MinPts) dentro de su vecindad de radio ε. Los puntos núcleo son el núcleo de un clúster y ayudan a formar y expandir el clúster. Todos los puntos alcanzables desde un punto núcleo, incluido el propio punto núcleo, pertenecen al mismo clúster.

Puntos frontera (border points): Son puntos que no cumplen la condición de tener el número mínimo de puntos dentro de su vecindad, pero están a una distancia ε de un punto núcleo. Los puntos frontera son considerados parte del clúster, pero no son puntos núcleo. Estos puntos ayudan a unir diferentes clústeres y se encuentran en los límites del clúster.

Puntos de ruido (noise points): Son puntos que no cumplen las condiciones anteriores. Estos puntos no pertenecen a ningún clúster y se consideran puntos aislados o ruido en el conjunto de datos.

En resumen, los puntos núcleo son el centro de un clúster, los puntos frontera están cerca de los puntos núcleo pero no tienen suficientes vecinos para ser considerados núcleo, y los puntos de ruido no pertenecen a ningún clúster. Pueden ser anómalos o no anómalos y necesitan más investigación.



DETECTANDO OUTLIERS

Método 4— Isolation Forest:

Isolation Forest es un algoritmo de aprendizaje no supervisado que pertenece a la familia de árboles de decisión *ensemble*. Este enfoque es diferente de todos los métodos anteriores. Todos los anteriores estaban tratando de encontrar la región normal de los datos y luego identifica cualquier cosa fuera de esta región definida como un valor atípico o anómalo.

Este método funciona de manera diferente. Aísla explícitamente las anomalías en lugar de perfilar y construir puntos y regiones normales mediante la asignación de una puntuación a cada punto de datos. Aprovecha el hecho de que las anomalías son los puntos de datos minoritarios y que tienen valores de atributo que son muy diferentes a los de las instancias normales. Este algoritmo funciona muy bien con conjuntos de datos de dimensiones muy altas y demostró ser una forma muy efectiva de detectar anomalías. Dado que esta sesión se centra en la implementación más que en analizarlo a fondo, no profundizaremos más en cómo funciona el algoritmo. Sin embargo, los detalles completos sobre cómo funciona están cubiertos en este documento:

https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf

Ejemplo: Supongamos que tenemos un conjunto de datos que representa transacciones bancarias. Queremos identificar si hay alguna transacción sospechosa que pueda indicar fraude.

Construcción del bosque: Creamos un bosque con varios árboles de decisión. Cada árbol se construye seleccionando aleatoriamente una característica (por ejemplo, el monto de la transacción) y un valor de división dentro del rango de valores en nuestro conjunto de datos.

Aislamiento de instancias: Comenzamos a dividir el conjunto de datos en subconjuntos utilizando las características y los valores de división seleccionados. Por ejemplo, si el valor de división es 1000 dólares, dividimos el conjunto de datos en dos grupos: uno con transacciones de monto superior a 1000 dólares y otro con transacciones de monto inferior.

Repetición del proceso: Continuamos dividiendo los subconjuntos resultantes en subconjuntos más pequeños utilizando diferentes características y valores de división. Repetimos este proceso hasta que cada transacción esté aislada individualmente o alcancemos un límite predefinido.

Cálculo de puntajes de anomalía: Calculamos un puntaje de anomalía para cada transacción en función del número promedio de divisiones necesarias para aislarla en todos los árboles

del bosque. Cuanto menor sea el puntaje de anomalía, más probable es que la transacción sea sospechosa.

Identificación de anomalías: Podemos identificar las transacciones con puntajes de anomalía más altos como posibles transacciones fraudulentas. Por ejemplo, si encontramos una transacción con un puntaje de anomalía muy alto, esto podría indicar que es una transacción inusual o atípica en comparación con el resto.

En resumen, el Isolation Forest es un algoritmo que utiliza la rapidez con la que se puede aislar una instancia en un bosque de árboles para determinar su nivel de anomalía. En el caso de las transacciones bancarias, las transacciones que se aíslan rápidamente se consideran más sospechosas y podrían indicar posibles fraudes.

Método 5— Robust Random Cut Forest:

El algoritmo Random Cut Forest (RCF) es el algoritmo no supervisado de Amazon para detectar anomalías. Funciona asociando también una puntuación de anomalía. Los valores de puntuación bajos indican que el punto de datos se considera "normal". Los valores altos indican la presencia de una anomalía en los datos. Las definiciones de "bajo" y "alto" dependen de la aplicación, pero la práctica común sugiere que las puntuaciones superiores a tres desviaciones estándar de la puntuación media se consideran anómalas. Los detalles del algoritmo se pueden encontrar en este artículo:

http://proceedings.mlr.press/v48/guha16.pdf

Lo mejor de este algoritmo es que funciona con datos dimensionales muy altos. También puede funcionar con datos de transmisión en tiempo real (incorporados en AWS Kinesis Analytics), así como con datos sin conexión.

El paper muestra algunos puntos de referencia de rendimiento en comparación con Isolation Forest. Estos son los resultados del paper que muestra que RCF es mucho más preciso y rápido que Isolation Forest.

Ejemplo: Supongamos que tenemos un conjunto de datos que representa la temperatura diaria en una ciudad durante un año. Queremos identificar días en los que la temperatura fue inusualmente alta o baja.

Construcción del bosque aleatorio: Creamos un bosque aleatorio de cortes. Cada árbol del bosque se construye seleccionando aleatoriamente una característica (por ejemplo, la temperatura) y un valor de corte dentro del rango de valores en nuestro conjunto de datos.

DETECTANDO OUTLIERS

División de instancias: Comenzamos a dividir el conjunto de datos en subconjuntos utilizando las características y los valores de corte seleccionados. Por ejemplo, si el valor de corte es 25 grados Celsius, dividimos el conjunto de datos en dos grupos: uno con días de temperatura superior a 25 grados y otro con días de temperatura inferior.

Repetición del proceso: Continuamos dividiendo los subconjuntos resultantes en subconjuntos más pequeños utilizando diferentes características y valores de corte. Repetimos este proceso hasta que cada día esté aislado individualmente o alcancemos un límite predefinido.

Cálculo de la profundidad de los días: Calculamos la profundidad de cada día en el bosque, que representa la cantidad de divisiones necesarias para aislar el día en todos los árboles del bosque.

Identificación de anomalías: Podemos identificar los días con una profundidad más baja como posibles días anómalos. Días que requirieron menos divisiones para ser aislados se consideran menos representativos del resto de los días y podrían indicar temperaturas inusuales.

En resumen, el Robust Random Cut Forest es un algoritmo que utiliza un bosque aleatorio de cortes para detectar anomalías en un conjunto de datos. Cuanto menor sea la profundidad de una instancia en el bosque, más probable es que sea una anomalía. En el ejemplo de la temperatura diaria, los días con una menor profundidad en el bosque podrían indicar temperaturas inusuales y ser identificados como posibles anomalías.

Table 1. Comparison of Baseline Isolation Forest to proposed Robust Random Cut Forest

Method	Sample Size	Positive Precision	Positive Recall	Negative Precision	Negative Recall	Accuracy	AUC
IF	256	0.42 (0.05)	0.37 (0.02)	0.96 (0.00)	0.97 (0.01)	0.93 (0.01)	0.83 (0.01)
RRCF	256	0.87 (0.02)	0.44 (0.04)	0.97 (0.00)	1.00 (0.00)	0.96 (0.00)	0.86 (0.00)
IF	512	0.48 (0.05)	0.37 (0.01)	0.97 (0.01)	0.96 (0.00)	0.94 (0.00)	0.86 (0.00)
RRCF	512	0.84 (0.04)	0.50 (0.03)	0.99 (0.00)	0.97 (0.00)	0.96 (0.00)	0.89 (0.00)
IF	1024	0.51 (0.03)	0.37 (0.01)	0.96 (0.00)	0.98 (0.00)	0.94 (0.00)	0.87 (0.00)
RRCF	1024	0.77 (0.03)	0.57 (0.02)	0.97 (0.00)	0.99 (0.00)	0.96 (0.00)	0.90 (0.00)

Method	Segment Precision	Segment Recall	Time to Detect Onset	Time to Detect End	Prec@5	Prec@10	Prec@15	Prec@20
IF	0.40 (0.09)	0.80 (0.09)	22.68 (3.05)	23.30 (1.54)	0.52 (0.10)	0.50 (0.00)	0.34 (0.02)	0.28 (0.03)
RRCF	0.65 (0.14)	0.80 (0.00)	13.53 (2.05)	10.85 (3.89)	0.58 (0.06)	0.49 (0.03)	0.39 (0.02)	0.30 (0.00)

problems in the business or building a proactive solution to potentially

Notemos que En lugar de construir un bosque aleatorio de cortes como en RRCF, Isolation Forest utiliza árboles de aislamiento. Estos árboles se construyen seleccionando aleatoriamente una característica y un valor de corte, al igual que en RRCF. Sin embargo, en lugar de dividir los datos en subconjuntos, los árboles de aislamiento dividen los datos en función de las comparaciones con los valores de corte.

TRATANDO LOS OUTLIERS

¿Qué métodos se utilizan para tratar los valores atípicos?

Existen varios métodos comúnmente utilizados para tratar los valores atípicos, según la naturaleza de los datos y el análisis específico que se realice. Aquí hay algunos enfoques comunes:

1. *Eliminación*: este método consiste simplemente en eliminar los valores atípicos del conjunto de datos. Puede ser apropiado cuando se cree que los valores atípicos se deben a errores de entrada de datos o errores de medición. Sin embargo, se debe tener precaución ya que eliminar valores atípicos sin una buena razón puede sesgar el análisis y distorsionar los resultados.

2. *Transformación*: la transformación de los datos mediante funciones matemáticas a veces puede reducir el impacto de los valores atípicos. Las transformaciones comunes incluyen tomar el logaritmo, la raíz cuadrada o el recíproco de los datos. Estas transformaciones pueden ayudar a que los datos se distribuyan de manera más normal y a estabilizar la varianza.

3. *Winsorización*: Winsorización reemplaza valores de datos extremos con valores menos extremos. El proceso implica limitar o truncar los valores extremos en un determinado percentil (por ejemplo, reemplazar los valores por encima del percentil 95 con el valor en el percentil 95). Este enfoque reduce la influencia de los valores atípicos al mismo tiempo que conserva parte de la información de los valores extremos.

4. *Topeo:* es una técnica similar a la anterior que también se utiliza para tratar valores extremos, pero en lugar de reemplazar los valores extremos, los "topea" o limita a un valor específico. Por ejemplo, si se aplica el topeo a 100 en un conjunto de datos, todos los valores por encima de 100 se ajustarán a 100. Esto se hace para evitar que los valores extremos influyan demasiado en los análisis o modelos.

5. *Imputación*: en lugar de eliminar los valores atípicos, se pueden reemplazar con valores estimados. Las técnicas de imputación incluyen reemplazar los valores atípicos con la media, la mediana u otro valor adecuado según las características de los datos. La imputación debe hacerse con cuidado, ya que puede introducir sesgos si no se maneja adecuadamente.

6. *Métodos robustos*: Los métodos estadísticos robustos están diseñados para ser menos sensibles a los valores atípicos. Estos métodos estiman parámetros usando estimadores robustos que no están fuertemente influenciados por valores extremos. Por ejemplo, la

mediana es una medida robusta de tendencia central que se ve menos afectada por los valores atípicos en comparación con la media.

7. Enfoques basados en modelos: en algunos casos, los valores atípicos pueden detectarse y tratarse mediante modelos específicos. Por ejemplo, en el análisis de regresión, los valores atípicos influyentes se pueden identificar utilizando medidas de diagnóstico como la distancia de Cook o los residuos *studentizados*. Una vez identificados, los valores atípicos se pueden reducir o excluir del análisis.

La elección del método depende del contexto específico, la naturaleza de los datos y los objetivos del análisis. Es importante considerar las razones subyacentes de los valores atípicos y el impacto potencial de su tratamiento en los resultados.

¿Cuándo debo eliminar los valores atípicos?

Decidir cuándo eliminar los valores atípicos es un juicio que depende del contexto y los objetivos de su análisis. Aquí hay algunas situaciones en las que se puede considerar la eliminación de valores atípicos:

1. Errores de ingreso de datos o errores de medición: si tiene pruebas sólidas o sospechas de que los valores atípicos se deben a errores en el ingreso de datos o la medición, puede ser apropiado eliminarlos. Por ejemplo, si tiene un conjunto de datos de alturas humanas y observa una entrada que es claramente un error tipográfico (por ejemplo, una altura de 8 pies), tendría sentido eliminar ese valor atípico.

2. Violaciones de suposiciones: algunos análisis estadísticos asumen ciertas distribuciones o relaciones entre variables. Si los valores atípicos están causando violaciones significativas de estos supuestos y es poco probable que formen parte de la población subyacente o del proceso que está estudiando, su eliminación puede estar justificada. *Por ejemplo, si está realizando un análisis de regresión lineal y los valores atípicos provocan una desviación sustancial de la linealidad,* eliminarlos puede ayudar a garantizar la validez del modelo de regresión.

3. Análisis sensible: en ciertos casos, los valores atípicos pueden tener un impacto desproporcionado en los resultados, lo que lleva a estimaciones sesgadas o errores estándar inflados. En tales situaciones, se puede considerar eliminar los valores atípicos para obtener resultados más precisos y confiables. Sin embargo, es crucial documentar y justificar la eliminación de valores atípicos, ya que su eliminación puede afectar la interpretación del análisis.

TRATANDO LOS OUTLIERS

4. Mejora del rendimiento del modelo: los valores atípicos a veces pueden afectar negativamente al rendimiento de los modelos predictivos. Si los valores atípicos influyen significativamente en las predicciones del modelo o conducen a un rendimiento deficiente del modelo, eliminarlos podría mejorar la precisión o la generalización del modelo.

5. Conocimiento del dominio específico o criterio experto: el conocimiento del dominio o la experiencia en la materia pueden proporcionar información sobre si los valores atípicos son observaciones significativas o anómalas. Si tiene una buena comprensión del proceso de generación de datos y sabe que ciertos valores extremos son inverosímiles o no están relacionados con el fenómeno que se está estudiando, podría ser razonable eliminarlos.

Sin embargo, es importante tener cuidado al eliminar los valores atípicos. *La eliminación ciega de valores atípicos sin una justificación sólida o una comprensión clara de su origen puede conducir a resultados sesgados o engañosos.* Es recomendable evaluar cuidadosamente el impacto de los valores atípicos en su análisis, explorar métodos alternativos para manejar los valores atípicos y considerar técnicas estadísticas sólidas que sean menos sensibles a los valores extremos antes de decidir eliminarlos. Además, *documentar la justificación y los pasos tomados para la eliminación de valores atípicos es crucial para la transparencia y la reproducibilidad de su análisis.*

TIPOS DE OUTLIERS

Se debe distinguir entre valores atípicos univariados y multivariados. Los valores atípicos univariados son valores extremos en la distribución de una variable específica, mientras que los valores atípicos multivariados son una combinación de valores en una observación que es poco probable. Por ejemplo, un valor atípico univariante podría ser una medición de la edad humana de 120 años o una medición de la temperatura en la Antártida de 50 grados centígrados.

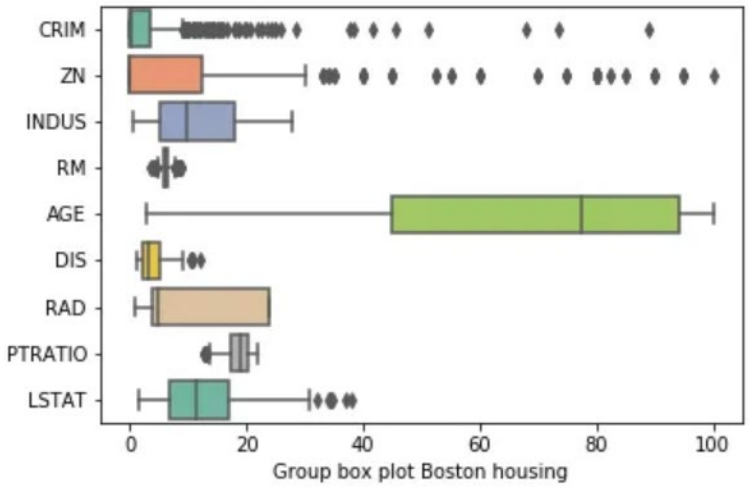
Un valor atípico multivariado podría ser una observación de un ser humano con una altura de 2 metros (en el percentil 95) y un peso de 50 kg (en el percentil 5). Ambos tipos de valores atípicos pueden afectar el resultado de un análisis, pero se detectan y tratan de manera diferente.

Los valores atípicos se pueden descubrir de varias maneras, incluidos métodos estadísticos, métodos basados en la proximidad o detección de valores atípicos supervisada. En esta introducción nos centraremos únicamente en los métodos estadísticos de uso común.

Visualización de valores atípicos:

Un primer y útil paso para detectar valores atípicos univariados es la visualización de la distribución de las variables. Por lo general, cuando se realiza una EDA, esto debe hacerse individualmente para todas las variables interesantes de un conjunto de datos. Una manera fácil de resumir visualmente la distribución de una variable es el diagrama de caja.

En un diagrama de caja, introducido por John Tukey en 1970, los datos se dividen en cuartiles. Por lo general, muestra una caja rectangular que representa el 25%-75% de las observaciones de una muestra, extendida por los llamados bigotes que alcanzan la entrada de datos mínima y máxima. Las observaciones que se muestran fuera de los bigotes son valores atípicos.



OUTLIERS UNIVARIADOS

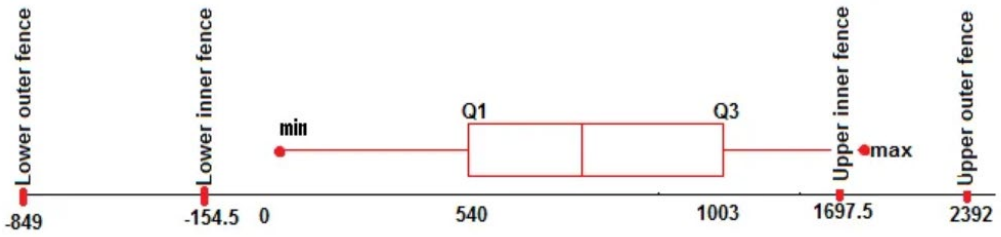
Parece que hay tres variables, precisamente AGE, INDUS y RAD, SIN observaciones atípicas univariadas. Todas las variables restantes tienen puntos de datos más allá de sus bigotes.

Los diagramas de caja son excelentes para resumir y visualizar la distribución de variables de manera fácil y rápida. Sin embargo, no identifican los índices reales de las observaciones periféricas. A continuación, analizaré tres métodos cuantitativos comúnmente utilizados en estadística para la detección de valores atípicos univariados:

- Método de diagrama de caja de Tukey
- Residuales studentizados internamente (método AKA z-score)
- Método de desviación absoluta mediana

Método de diagrama de caja de Tukey:

Además de sus beneficios visuales, el diagrama de caja proporciona estadísticas útiles para identificar observaciones individuales como valores atípicos. Tukey distingue entre valores atípicos posibles y probables. Un *posible* valor atípico se encuentra entre la valla interior y la exterior, mientras que un *probable* valor atípico se encuentra fuera de la valla exterior.



Example of a box plot including the inner and outer fences and minimum and maximum observations (known as whiskers). Image by Stephanie Glen on statisticsHowTo.com

Si bien la *inner* (a menudo confundida con los *whiskers*) y la *outer fence* generalmente no se muestran en el diagrama de caja real, se pueden calcular usando el rango intercuartílico (IQR) de esta manera:

IQR =Q3 - Q1, mientras que q3: = cuartil 75 y q1: = cuartil 25

Inner fence = [Q1-1.5*IQR, Q3+1.5*IQR]

Outer fence = [Q1–3*IQR, Q3+3*IQR]

La valla interior de la distribución se define como 1,5 x IQR por debajo de Q1 y 1,5 x IQR por encima de Q3. La valla exterior se define como 3 x IQR por debajo de Q1 y 3 x IQR por encima de Q3. Siguiendo a Tukey, solo se tratan los valores atípicos probables, que se encuentran fuera de la valla exterior.

OUTLIERS UNIVARIADOS

La gran ventaja del método de diagrama de caja de Tukey es que las estadísticas (por ejemplo, IQR, cerca interna y externa) son sólidas para los valores atípicos, lo que significa que encontrar un valor atípico es independiente de todos los demás valores atípicos. Además, las estadísticas son fáciles de calcular.

Además, este método no requiere una distribución normal de los datos, que a menudo no se garantiza en entornos de la vida real. Si una distribución está muy sesgada (generalmente se encuentra en datos de la vida real), el método Tukey se puede extender al método log-IQ. Aquí, cada valor se transforma en su logaritmo antes de calcular las vallas interior y exterior.

Finalmente, es importante destacar que el mínimo y el máximo de los whiskers no siempre coinciden con los límites establecidos por Q1-1.5IQR y Q3+1.5IQR. Esto se debe a que los límites de los whiskers se ajustan según los valores reales presentes en los datos. Si no hay valores fuera de estos límites, los whiskers se extienden hasta el valor mínimo y máximo de los datos. En algunos casos, esto puede resultar en una extensión de los whiskers más allá de los límites establecidos por Q1-1.5IQR y Q3+1.5IQR.

Por lo tanto, es posible que en ciertos diagramas de caja el mínimo de la caja no coincida exactamente con Q1-1.5IQR, especialmente si hay valores atípicos presentes en los datos. Es importante considerar el contexto y la interpretación de los diagramas de caja al analizar los resultados.

Residuales studentizados internamente, también conocido como método de puntuación z:

Otro método comúnmente utilizado para detectar valores atípicos univariados son los residuos estandarizados internamente, también conocido como el método de puntuación z. Para cada observación (Xn), se mide cuántas desviaciones estándar está el punto de datos de su media ($\bar{X}$).

$$z_n = \frac{X_n - \bar{X}}{SD_X}$$

Siguiendo una regla general común, si z > C, donde C generalmente se establece en 3, la observación se marca como un valor atípico. Esta regla se deriva del hecho de que, si una variable se distribuye normalmente, el 99,7 % de todos los puntos de datos se ubican 3 desviaciones estándar alrededor de la media.

Sin embargo, este método es muy limitado ya que la media y la desviación estándar de las distribuciones son sensibles a los valores atípicos. Esto significa que encontrar un valor atípico depende de otros valores atípicos, ya que cada observación afecta directamente a la media.

OUTLIERS UNIVARIADOS

Además, el método de puntuación z supone que la variable de interés se distribuye normalmente. Un método más robusto que se puede utilizar en su lugar son los residuos studentizados externamente. Aquí, la influencia del punto de datos examinado se elimina del cálculo de la media y la desviación estándar, así:

$$\overline{X}_{(n)} = \frac{1}{N-1} \sum_{i \in \mathbb{N}_{(n)}} X_i$$

$$SD_{X(n)} = \sqrt{\frac{1}{N-1} \sum_{i \in \mathbb{N}_{(n)}} (X_i - \overline{X}_{(n)})^2}$$

Sin embargo, los residuales studentizados externamente tienen limitaciones ya que la media y las desviaciones estándar aún son sensibles a otros valores atípicos y aún esperan que la variable de interés X tenga una distribución normal.

Método de desviación absoluta mediana:

El método de la desviación absoluta de la mediana (MAD) reemplaza la media y la desviación estándar con estadísticas más sólidas, como la desviación absoluta de la mediana y la mediana. La desviación absoluta mediana se define como:

$$\text{MAD} = \text{median}(\left|X_i - \overline{X}\right|)$$

La estadística de prueba se calcula como el puntaje z utilizando estadísticas sólidas. Además, para identificar observaciones periféricas, se utiliza el mismo punto de corte de 3. Si la estadística de prueba se encuentra por encima de 3, se marca como un valor atípico. En comparación con los residuos studentizados internamente (puntuación z) y externamente, este método es más resistente a los valores atípicos y supone que X se distribuye paramétricamente (Ejemplos de distribuciones paramétricas discretas y continuas).

Hay diferentes formas de detectar valores atípicos univariados, cada uno con sus ventajas y desventajas. El puntaje z debe aplicarse de manera crítica debido a su sensibilidad a la media y la desviación estándar y su suposición de una variable distribuida normalmente. El método MAD se usa a menudo en su lugar y sirve como una alternativa más sólida. El método de diagrama de caja de Tukey ofrece resultados sólidos y se puede ampliar fácilmente cuando los datos están muy sesgados.

Para decidir el enfoque correcto para su propio conjunto de datos, examine de cerca la distribución de sus variables y utilice su conocimiento del dominio.

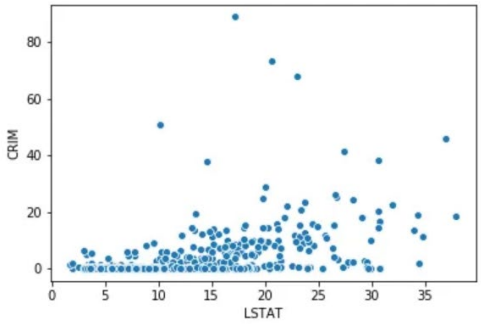
OUTLIERS MULTIVARIADOS

Un valor atípico multivariante es una combinación inusual de valores en una observación a través de varias variables. Por ejemplo, podría ser un ser humano con una altura de 2 metros (en el percentil 95) y un peso de 50 kg (en el percentil 5).

Visualización:

Una forma común de trazar valores atípicos multivariados es el diagrama de dispersión. Tenga en cuenta que la visualización de valores atípicos multivariados en más de dos variables no es factible en un espacio 2D. Por lo tanto, nos ceñiremos a los valores atípicos encontrados en dos variables para la visualización, los llamados valores atípicos bivariados.

El diagrama de dispersión visualiza la relación entre dos variables (numéricas). En un diagrama de dispersión, cada observación se representa como un punto con dos coordenadas (X, Y) que representan dos variables. Aquí, por ejemplo, X representa el valor de la variable 1 e Y el valor de la variable 2.



Scatterplot: crime rate per capita by town (CRIM) against the percentage of lower status in the population

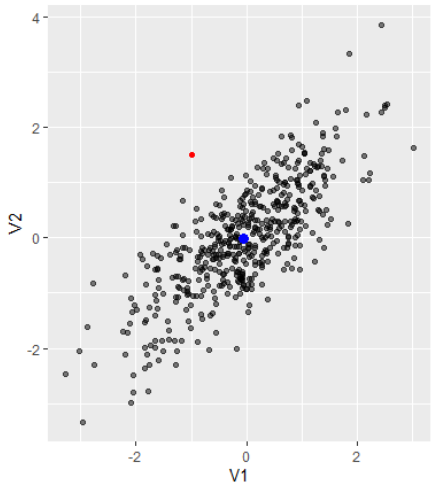
Al igual que los diagramas de caja, los diagramas de dispersión visualizan muy bien las observaciones periféricas, pero no las identifican ni las marcan para facilitar el tratamiento. Cuando se trata de valores atípicos multivariados, las métricas de distancia pueden ser útiles para la detección. Con métricas de distancia, se determina la distancia entre dos vectores. Estos dos vectores pueden ser dos observaciones diferentes (filas) o una observación (fila) comparada con el vector medio (fila de medias de todas las columnas). Las métricas de distancia se pueden calcular independientemente del número de variables en el conjunto de datos (columnas).

DISTANCIA DE MAHALANOBIS

Uno de los objetivos principales -y difíciles- de la estadística es determinar cuándo dos elementos son parecidos y cuándo no. Resolver adecuadamente esta cuestión es la base de multitud de procedimientos y algoritmos, desde contrastes de hipótesis hasta métodos de clasificación. Usamos la ambigua palabra elementos, porque a veces interesa comparar magnitudes escalares (números), pero otras veces queremos comparar observaciones multivariantes (vectores). Incluso podemos necesitar saber si dos funciones se parecen, o elementos más complicados, como fotografías, novelas, tuits o fragmentos de ADN.

P. C. Mahalanobis (1893-1972) se graduó en Física en Calculta en 1912 y completó sus estudios en Cambridge, donde fue compañero de habitación del famoso matemático Srinavasa Ramanujan. En Cambrige, Mahalanobis acabó centrando su investigación en el área de estadística. Es especialmente conocido por definir la distancia que lleva su nombre, aunque realizó también contribuciones relevantes en la planificación de muestreos a gran escala.

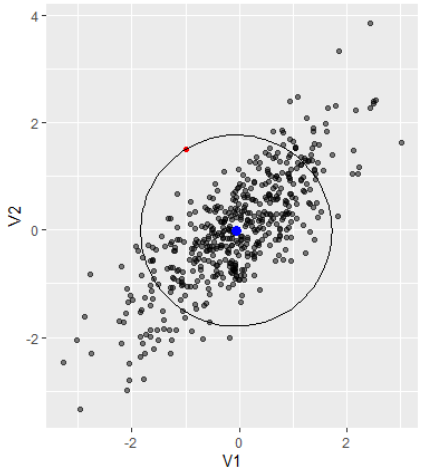
La distancia de Mahalanobis para vectores:



La relación entre las variables en el punto rojo es ligeramente distinta que en el resto, por eso destaca.

Si usamos la distancia euclídea habitual para medir la distancia a la media no tendremos en cuenta la relación entre las variables. Muchos puntos que estadísticamente aparecen como más "normales" (y están fuera del círculo en la figura siguiente) distarán de la media más que el punto rojo, cuando estadísticamente destacan menos.

DISTANCIA DE MAHALANOBIS



Para definir una distancia que tenga sentido estadístico tenemos que tener en cuenta las correlaciones entre las variables. Esto motiva el uso de la distancia de Mahalanobis.

La distancia de Mahalanobis es una medida que se utiliza para calcular la distancia entre un punto y un conjunto de puntos en un espacio multidimensional. A diferencia de otras medidas de distancia, como la distancia euclidiana, la distancia de Mahalanobis tiene en cuenta la covarianza entre las variables.

La matriz de covarianza captura la relación entre las diferentes variables y también proporciona información sobre la dispersión o variabilidad de los datos.

Al considerar la correlación, la distancia de Mahalanobis ajusta la medida de distancia según la relación lineal entre las variables. *Si hay una alta correlación entre dos variables, significa que varían juntas en la misma dirección.* En este caso, la distancia de Mahalanobis tomará en cuenta esta relación y ajustará la medida de distancia en consecuencia.

Además, la distancia de Mahalanobis también tiene en cuenta la varianza de los datos. La varianza mide la dispersión o variabilidad de una variable. Si una variable tiene una alta varianza, significa que los datos están más dispersos alrededor de su media. *La distancia de Mahalanobis considera la varianza de cada variable en el cálculo de la distancia, lo que implica que los puntos que están lejos de la media en términos de variabilidad tendrán una mayor distancia de Mahalanobis.*

En términos sencillos, la distancia de Mahalanobis mide cuánto se aleja un punto de un conjunto de puntos teniendo en cuenta la correlación y la variabilidad de las variables en ese

OUTLIERS MULTIVARIADOS

conjunto. Esto significa que tiene en cuenta la forma y orientación de los datos en el espacio multidimensional.

El cálculo de la distancia de Mahalanobis equivale a:

1. Trasladar los datos para que su nueva media sea el origen de coordenadas: Esto implica restar la media de cada variable a todos los puntos de datos. Al hacerlo, los datos se desplazan de manera que la nueva media sea el punto de origen de coordenadas (0, 0, ..., 0). Este paso asegura que la media de los datos trasladados sea igual a cero en todas las variables.
2. Rotar los datos para que las correlaciones entre las variables sean lo mas cercano a cero: El objetivo de este paso es reducir las correlaciones entre las variables. Esto se logra mediante una transformación lineal que rota el sistema de coordenadas original.

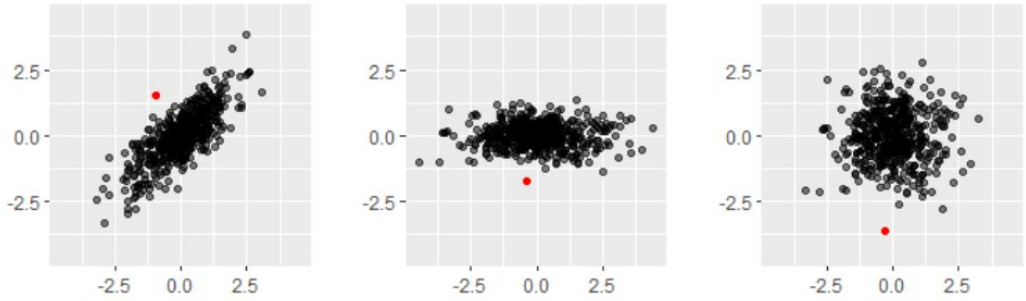
   Esta matriz de rotación se elige de tal manera que las nuevas variables resultantes tengan una correlación mínima o igual a cero entre sí. Dado que la matriz de rotación se elige para minimizar las correlaciones, es posible que las nuevas variables no sean completamente ortogonales entre sí. Sin embargo, se espera que las correlaciones entre las nuevas variables sean muy bajas o cercanas a cero, lo que significa que las variables están casi des correlacionadas. La idea detrás de esta rotación es desacoplar las variables originales y reducir la redundancia de información, lo que ayuda a evitar la duplicación o el sesgo en el cálculo de la distancia de Mahalanobis.

3. Aplicar una estandarización para que todas las varianzas sean igual a uno: Después de trasladar y rotar los datos, se realiza una estandarización para asegurar que todas las varianzas sean iguales a uno. Esto se logra dividiendo cada variable por su desviación estándar. La estandarización es importante para asegurar que las variables tengan el mismo rango y no dominen la distancia de Mahalanobis debido a su escala.

4. Calcular la distancia euclidiana al origen de los puntos resultantes: Finalmente, se calcula la distancia euclidiana desde el origen de coordenadas (0, 0, ..., 0) hasta cada uno de los puntos de datos transformados. La distancia euclidiana es una medida de la distancia entre dos puntos en un espacio multidimensional y se calcula utilizando la fórmula matemática de la distancia euclidiana.

OUTLIERS MULTIVARIADOS

El siguiente gráfico representa el resultado de los tres primeros pasos con los datos del ejemplo. He marcado en rojo el punto que habíamos identificado como atípico en un principio. Se puede ahora comprobar que, en el sentido de la distancia de Mahalanobis, es uno de los puntos más lejanos a la media, lo que coincide con la intuición que teníamos al mirar los datos.



Distancia de Mahalanobis:

Una métrica de distancia ampliamente utilizada para la detección de valores atípicos multivariados es la distancia de Mahalanobis (MD). El MD es una medida que determina la distancia entre un punto de datos x y una distribución D. Es una generalización multivariante de los residuos estudentizados internamente (puntuación z) presentados en mi último artículo. Esto significa que el MD define a cuántas desviaciones estándar x se aleja de la media de D.

Se define como:

$$d_n = \sqrt{(x_n - \hat{\mu}_X)\, C^{-1}\, (x_n - \hat{\mu}_X)^T}$$

Aquí, x representa un vector de observación y μ el vector medio aritmético de variables independientes (columnas) en la muestra. C (-1) es la matriz de covarianza inversa de las variables independientes en la muestra. Para comprender mejor la intuición matemática detrás del MD, recomiendo leer el punto cuatro de esta publicación de blog.

Al igual que la puntuación z, la MD de cada observación se compara con un punto de corte. Suponiendo una distribución normal multivariante de los datos con K variables, la distancia de

OUTLIERS MULTIVARIADOS

Mahalanobis sigue una distribución chi-cuadrado con K grados de libertad. Utilizando un nivel de significación razonable (por ejemplo, 2,5 %, 1 %, 0,01 %), el punto de corte se define como:

$$C = \sqrt{\chi^2_{K,q}}$$

Un inconveniente del MD es que utiliza la media aritmética y la matriz de covarianza y, con eso, es muy sensible a los valores atípicos en los datos. Existen varios métodos que usan estimaciones robustas para μ y C. En el siguiente pasaje, explicaré el método del Determinante de Covarianza Mínima, presentado por Rousseeuw, como ejemplo.

Distancia robusta de Mahalanobis:

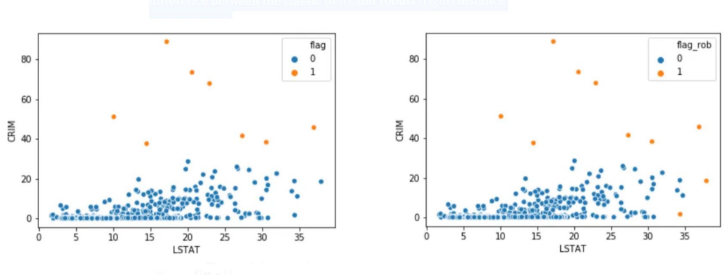
El método del determinante de covarianza mínima (MCD) proporciona estimaciones sólidas para μ y C utilizando solo un subconjunto de la muestra. Solo utiliza las observaciones donde el determinante de la matriz de covarianza es lo más pequeño posible. Se define como la MD clásica, pero con estimaciones robustas para la media y la covarianza:

Distancia robusta de Mahalanobis:

El método del determinante de covarianza mínima (MCD) proporciona estimaciones sólidas para μ y C utilizando solo un subconjunto de la muestra. Solo utiliza las observaciones donde el determinante de la matriz de covarianza es lo más pequeño posible. Se define como la MD clásica, pero con estimaciones robustas para la media y la covarianza:

$$rd_n = \sqrt{(x_n - \hat{\mu}_{R,X})\, C_R^{-1}\, (x_n - \hat{\mu}_{X,R})^T}$$

Los dos gráficos a continuación (basados en el ejemplo del conjunto de datos 2D) ayudan a visualizar la diferencia entre la medición de distancia clásica (izquierda) y robusta (derecha):



OUTLIERS MULTIVARIADOS

Además de los valores atípicos univariados, también es importante examinar un conjunto de datos subyacente para los valores atípicos multivariados. Ambos tipos de valores atípicos pueden afectar significativamente los resultados de un análisis de datos o proyectos de aprendizaje automático. En esta publicación, aprendimos que los valores atípicos multivariados son una combinación única de valores en una observación y se pueden detectar a través de métricas de distancia.

Una métrica de distancia comúnmente utilizada es la distancia de Mahalanobis. Su definición clásica se basa en la media y la covarianza entre todas las variables de un conjunto de datos. Por lo tanto, es sensible a los valores atípicos. Una forma de recibir estimaciones más sólidas para la media y la covarianza es el método de Determinantes de Covarianza Mínima (MCD).

Finalmente, es importante tener en cuenta que existen otras formas de detectar valores atípicos univariados y multivariados. Otros métodos populares son k-vecinos más cercanos, DBSCAN o bosques de aislamiento, solo por nombrar algunos. No existe un método correcto o incorrecto, pero uno podría ser más apropiado que otro para su conjunto de datos. Al decidir el método de detección de valores atípicos que le gustaría usar, le recomiendo basar su decisión en la distribución de los datos, el tamaño de la muestra y la cantidad de dimensiones.