Análisis Exploratorio de Datos (AED

Maestría en Ciencia de Datos

Víctor Saquicela

Universidad de Cuenca

Junio 2023

UNIVERSIDAD
DE CUENCA

Contents

Introducción

Introducción

- La objetivo del AED es examinar los datos previamente a la aplicación de cualquier técnica estadística.
- El analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.
- El AED proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de datos, tratamiento y evaluación de datos ausentes, identificación de casos atípicos y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes.

Introducción

- ¿Existe algún tipo de estructura (normalidad, multimodalidad, asimetría, curtosis, linealidad, homogeneidad entre grupos, homocedasticidad, etc.) en los datos que voy a analizar?
- ¿Existe algún sesgo en los datos recogidos?
- ¿Hay errores en la codificación de los datos?
- ¿Cómo se sintetiza y presenta la información contenida en un conjunto de datos?
- ¿Existen datos atípicos (outliers)? ¿Cuáles son? ¿Cómo tratarlos?
- ¿Hay datos ausentes (missing)? ¿Tienen algún patrón sistemático? ¿Cómo tratarlos?

QUÉ ES EL ANÁLISIS EXPLORATORIO DE DATOS?

Que es el AED?

- Conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas.
- Proporciona métodos sistemáticos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de los mismos, tratamiento y evaluación de datos ausentes (missing), identificación de casos atípicos (outliers) y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes (normalidad, linealidad, homocedasticidad).
- El análisis previo de los datos es un paso necesario, que lleva tiempo, y que habitualmente se descuida por parte de los analistas de datos.
- Las tareas implícitas en dicho análisis pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico.

Etapas del Análisis Exploratorio de Datos

Etapas del Análisis Exploratorio de Datos

1. Preparar los datos para hacerlos accesibles a cualquier técnica estadística.
2. Realizar un examen gráfico de la naturaleza de las variables individuales a analizar y un análisis descriptivo numérico que permita cuantificar algunos aspectos gráficos de los datos.
3. Realizar un examen gráfico de las relaciones entre las variables analizadas y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre ellas.
4. Evaluar, si fuera necesario, algunos supuestos básicos subyacentes a muchas técnicas estadísticas como, por ejemplo, la normalidad, linealidad y homocedasticidad.
5. Identificar los posibles casos atípicos (outliers) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.
6. Evaluar, si fuera necesario, el impacto potencial que pueden tener los datos ausentes (missing) sobre la representatividad de los datos analizados.

Preparación de los Datos

Preparación de los Datos

- El primer paso en un AED es hacer accesible los datos a cualquier técnica estadística.
- Ello conlleva la selección del método de entrada (por teclado o importados de un archivo) y codificación de los datos así como la de un herramienta estadística adecuado para procesarlos.
- La codificación de los datos depende del tipo de variable. Las herramientas estadísticas existentes en el mercado proporcionan diversas posibilidades (datos tipo cadena, numéricos, nominales, ordinales, etc).
- La inmensa mayoría de los paquetes estadísticos permite realizar manipulaciones de los datos previas a un análisis de los mismos.

Preparación de los Datos

Algunas operaciones útiles son las siguientes:

- Combinar conjuntos de datos de dos archivos distintos
- Seleccionar subconjuntos de los datos
- Dividir el archivo de los datos en varias partes
- Transformar variables
- Ordenar casos
- Agregar nuevos datos y/o variables
- Eliminar datos y/o variables
- Guardar datos y/o resultados

Finalmente, y con el fin de aumentar la inteligibilidad de los datos almacenados, conviene asociar a la base de datos utilizada, un libro de códigos en el que se detallen los nombres de las variables utilizadas, su tipo y su rango de valores, su significado así como las fuentes de donde se han sacado los datos.

Análisis Estadístico Unidimensional

Análisis Estadístico Unidimensional

- Una vez organizados los datos, el segundo paso de un AED consiste en realizar una análisis estadístico gráfico y numérico de las variables del problema con el fin de tener una idea inicial de la información contenida en el conjunto de datos así como detectar la existencia de posibles errores en la codificación de los mismos.
- El tipo de análisis a realizar depende de la escala de medida de la variable analizada.

Análisis Estadístico Unidimensional

Escala de Medida	**Representaciones Gráficas**	**Medidas de Tendencia Central**	**Medidas de Dispersión**
Nominal	Diagrama de barras Diagrama de líneas Diagrama de sectores	Moda	
Ordinal	Boxplot	Mediana	Rango intercuartílico
Intervalo	Histograma Polígono de frecuencia	Media	Desviación típica
Razón		Media geométrica	Coeficiente de variaci

Table: Sugerencias en función de la escala de medida

Variables Cualitativas

- Las variables cualitativas son aquellas que no aparecen en forma numérica, sino como categorías o atributos como, por ejemplo, el sexo o la profesión de una persona.
- En dichas categorías puede haber un orden subyacente (variable ordinal) o no (variable nominal).
- Los datos correspondientes a variables cualitativas se agrupan de manera natural en diferentes categorías o clases y se cuenta el número de datos que aparecen en cada una de ellas.
- Se suelen representar mediante diagrama de barras, sectores o líneas.
- Utilizar tabla de frecuencias

Variables Cuantitativas

- Las variables cuantitativas son las que pueden expresarse numéricamente.
- Basado en el tipo de valores que puede tomar, permite distinguir entre variables cuantitativas discretas,que son, frecuentemente el resultado de contar y, por tanto, toman sólo valores enteros y variables cuantitativas continuas, que resultan de medir y pueden contener cifras decimales.
- Variables discretas son el número de lavadoras producidas por una empresa en un año.
- Variables continuas son aquellas cuyos valores pueden ser cualquier cantidad en un intervalo, como la temperatura, el peso o la altura de una persona o la superficie de las viviendas.
- Las variables cuantitativas discretas con un número pequeño de valores se tratarían de manera similar a las variables cualitativas antes descritas.
- Utilizar tabla de frecuencia cuando se tenga un número elevado de datos.
- Se representa gráficamente mediante histogramas, diagramas de tallos y hojas y boxplots con el fin de estudiar la forma de la distribución y analizar

Estudio de la Normalidad

Estudio de la Normalidad

- Muchos métodos estadísticos se basan en la hipótesis de normalidad de la variable objeto de estudio.
- Si la falta de normalidad de la variable es suficientemente fuerte, muchos de los contrastes utilizados en los análisis estadístico-inferenciales no son válidos.
- Incluso aunque las muestras grandes tiendan a disminuir los efectos perniciosos de la no normalidad, el investigador debería evaluar la normalidad de todas las variables incluidas en el análisis.
- Existen varios métodos para evaluar la normalidad de un conjunto de datos que pueden dividirse en dos grupos: los métodos gráficos y los contrastes de hipótesis.

Métodos Gráficos

- El método gráfico univariante más simple para diagnosticar la normalidad es una comprobación visual del histograma que compare los valores de los datos observados con una distribución normal.
- Aunque un histograma es atractivo por su simplicidad, este método es problemático para muestras pequeñas, donde la construcción del histograma puede distorsionar la representación visual de tal forma que el análisis sea poco fiable.
- Otras posibilidades, también basadas en información gráfica, consisten en realizar diagramas de cuantiles (Q-Q plots).

Contraste de Hipótesis

- No existe un contraste óptimo para probar la hipótesis de normalidad. La razón es que la potencia relativa depende del tamaño muestral y de la verdadera distribución que genera los datos.
- Desde un punto de vista poco riguroso, el contraste de Shapiro y Wilks es, en términos generales, el más conveniente en muestras pequeñas ($n < 30$), mientras que el contraste de Kolmogorov-Smirnov, en la versión modificada de Lilliefors es adecuado para muestras grandes.

Contraste de Hipótesis - test de Kolmogorov-Smirnov

- La hipótesis nula que se pone a prueba es que los datos proceden de una población con distribución normal frente a una alternativa de que no es así.
- Este contraste calcula la distancia máxima entre la función de distribución empírica de la muestra y la teórica.
- Si la distancia calculada es mayor que la encontrada en las tablas, fijado un nivel de significación, se rechaza el modelo normal.

Contraste de Hipótesis - contraste de Shapiro y Wilks

- Utilizado para muestras pequeñas (n<30)
- Trabajo

Contraste de Hipótesis - Otras

- Test de asimetría
- Test de curtosis

Transformaciones para alcanzar Normalidad

- En ocasiones la falta de normalidad de una variable puede arreglarse mediante una transformación de la misma.
- Trabajo

Forma de la Distribución	Transformación Aconsejada
Asimetría Positiva	Log(X+C)
Asimetría Negativa	Log(C-X)
Leptocurtosis	1/X
Platicurtosis	X^2

Table: Transformaciones para conseguir Normalidad

Análisis Estadístico Bidimensional y Multivariante

Análisis Estadístico Bidimensional

- Una vez realizado el estudio unidimensional de cada variable por separado, el siguiente paso consiste en analizar la existencia de posibles relaciones entre ellas.
- Este estudio puede realizarse desde una óptica bidimensional o multidimensional.
- Tres situaciones generales se pueden presentar: ambas variables son cualitativas, ambas variables son cuantitativas y una variable es cuantitativa y la otra cualitativa.

Análisis de dos Variables Cualitativas

- Se utiliza una tabla de contingencia que contiene en cada casilla la correspondiente frecuencia conjunta que representa el número de datos que pertenecen a la modalidad i-ésima de la primera variable y a la modalidad j-ésima de la segunda.
- A partir de dicha tabla podemos estudiar si las dos variables son o no independientes.
- Si son independientes no existe relación alguna entre ellas; en caso contrario analizaríamos el tipo y el grado de su dependencia tanto gráfica como numéricamente.
- Utilizar Chi-cuadrado

Análsis de dos Variables Cuantitativas

- La distribución conjunta de dos variables puede expresarse gráficamente mediante un diagrama de dispersión que proporciona una buena descripción de la relación entre las dos variables.
- La relación entre las variables también puede expresarse de forma numérica. Una medida de la relación entre dos variables que resuma la información del gráfico de dispersión y que no dependa de las unidades de medida es el coeficiente de correlación lineal.
- Cuando las variables están relacionadas linealmente de forma exacta, el coeficiente de correlación lineal será igual a uno en valor absoluto.
- Cuando las variables no están relacionadas linealmente entre sí, el coeficiente de correlación lineal es cero.
- Para interpretar este coeficiente conviene mirar siempre el diagrama de dispersión de los datos para comprobar que son homogéneos y que no existen datos atípicos.
- La existencia de correlación no implica una relación de causalidad entre las variables ni, en general, la no existencia de correlación permite deducir falta de causalidad.

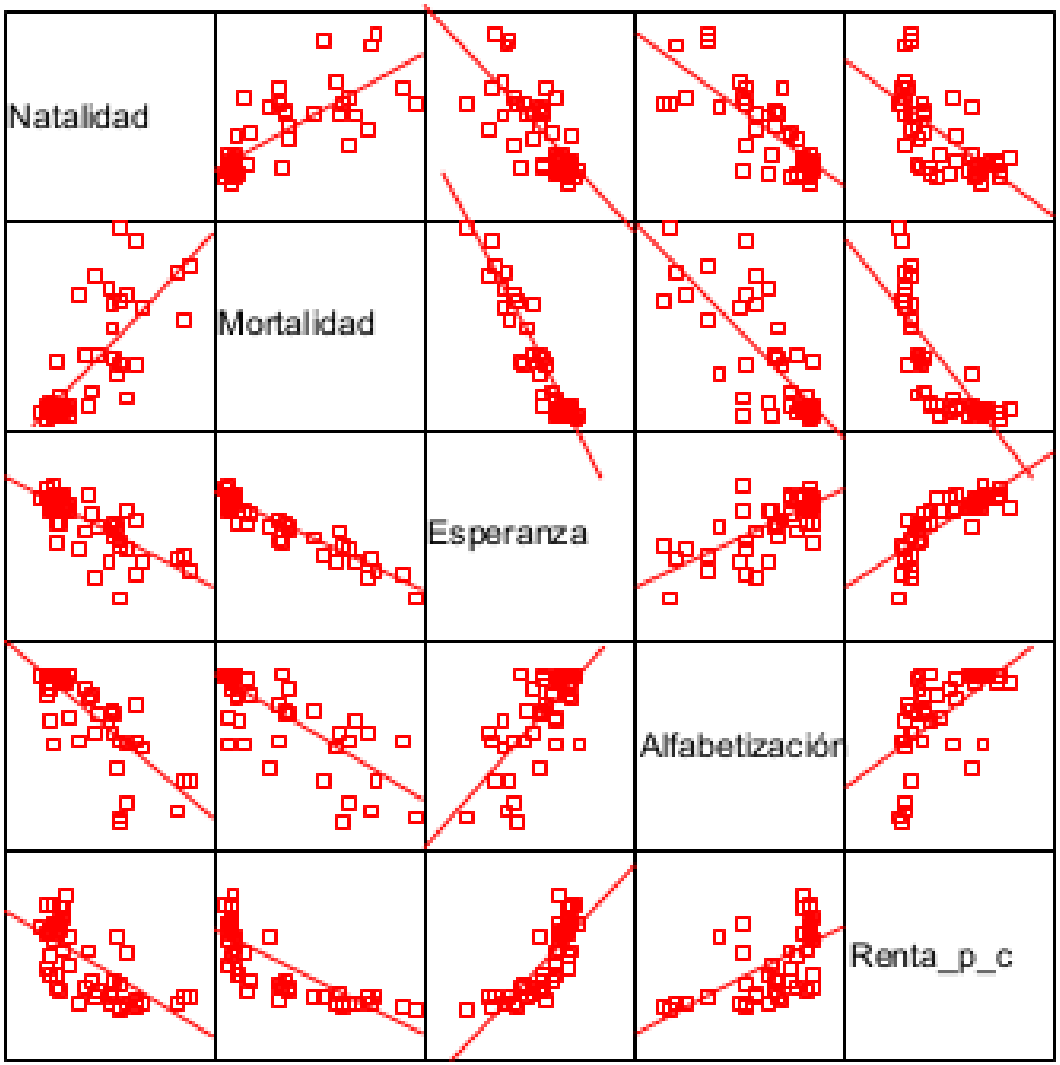
Análisis de dos Variables Cuantitativas - Linealidad

- La linealidad es un supuesto implícito de todas las técnicas multivariantes basadas en medidas de correlación, tales como la regresión múltiple, regresión logística, análisis factorial y los modelos de ecuaciones estructurales.
- Es, además, una forma indirecta de contrastar la normalidad conjunta de dos variables dado que si dicha hipótesis es cierta la relación existente entre ellas deberá ser lineal.
- Dado que las correlaciones representan sólo la asociación lineal entre variables, los efectos no lineales no estarán representados en el valor de la correlación.
- Como resultado, es siempre prudente examinar todas las relaciones para identificar cualquier desplazamiento de la linealidad que pueda impactar la correlación.
- La forma más común de evaluar la linealidad es examinar los gráficos de dispersión de las variables e identificar cualquier pauta no lineal en los datos.
- Una aproximación alternativa es ir a un análisis de regresión múltiple y examinar los residuos que reflejan la parte no explicada de la variable dependiente; por tanto, cualquier parte no lineal de la relación quedará reflejada en los residuos.

Análisis de dos Variables Cuantitativas - Diagramas de Dispersión Matriciales

- Existen muchos tipos de gráficos de dispersión, pero un formato que se ajusta particularmente cuando se aplican técnicas multivariantes son los llamados diagramas de dispersión matriciales que permiten analizar, de forma simultánea, las relaciones existentes entre un grupo de variables cuantitativas.
- Consisten en representar los diagramas de dispersión para todas las combinaciones de las variables analizadas.
- Con p variables existen, por lo tanto, p(p-1)/2 gráficos posibles, que pueden disponerse en forma de matriz para entender el tipo de relación existente entre los distintos pares de variables.
- En particular, estos gráficos son importantes para apreciar si existen relaciones no lineales, en cuyo caso la matriz de covarianzas puede no ser un buen resumen de la dependencia entre variables.

Análisis de dos Variables Cuantitativas - Diagramas de Dispersión Matriciales



Análisis de una Variable Cuantitativa y otra Variable Cualitativa

- Cuando se dispone de una variable cuantitativa y otra cualitativa, el estudio se enfoca como un problema de comparación del comportamiento de la variable numérica en las diferentes subpoblaciones que define la variable cualitativa.
- Ignorar la heterogeneidad debida a la presencia de subpoblaciones puede conducir a conclusiones equivocadas en el análisis.
- Una forma de realizar dicho análisis es mediante los diagramas de cajas y los test de diferencias de medias.

Análisis de una Variable Cuantitativa y otra Variable Cualitativa - Homocedasticidad

- La homocedasticidad es una hipótesis muy habitual en algunas técnicas estadísticas como el Análisis de la Varianza, el Análisis Discriminante y el Análisis de Regresión.
- Dicha hipótesis se refiere a suponer la igualdad de varianzas de las variables dependientes en diversos grupos formados por los distintos valores de las variables independientes.
- Si dicha hipótesis no se verifica puede alterar la potencia y el nivel de significación de los contrastes utilizados por dichas técnicas y de ahí el interés de analizar si se verifica o no y, en éste último caso, poner los remedios oportunos.
- Para ello se utilizan contrastes de hipótesis cuya finalidad es analizar la existencia de esta igualdad que, en muchas ocasiones, va ligada a una falta de normalidad de las variables analizadas.
- Aplicar ANOVA

Análisis Multivariante

- Es el conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente conjuntos de datos multivariantes en el sentido de que hay varias variables medidas para cada individuo ú objeto estudiado.
- Su razón de ser radica en un mejor entendimiento del fenómeno objeto de estudio obteniendo información que los métodos estadísticos univariantes y bivariantes son incapaces de conseguir.

Análisis Multivariante - Tipos de Técnicas

- Métodos de dependencia, suponen que las variables analizadas están divididas en dos grupos: las variables dependientes y las variables independientes. El objetivo de los métodos de dependencia consiste en determinar si el conjunto de variables independientes afecta al conjunto de variables dependientes y de qué forma.
- Métodos de interdependencia, estos métodos no distinguen entre variables dependientes e independientes y su objetivo consiste en identificar qué variables están relacionadas, cómo lo están y por qué.
- Métodos estructurales, suponen que las variables están divididas en dos grupos: el de las variables dependientes y el de las independientes. El objetivo de estos métodos es análizar, no sólo como las variables independientes afectan a las variables dependientes, sino también cómo están relacionadas las variables de los dos grupos entre sí.

Análisis Multivariante - Tipos de Técnicas

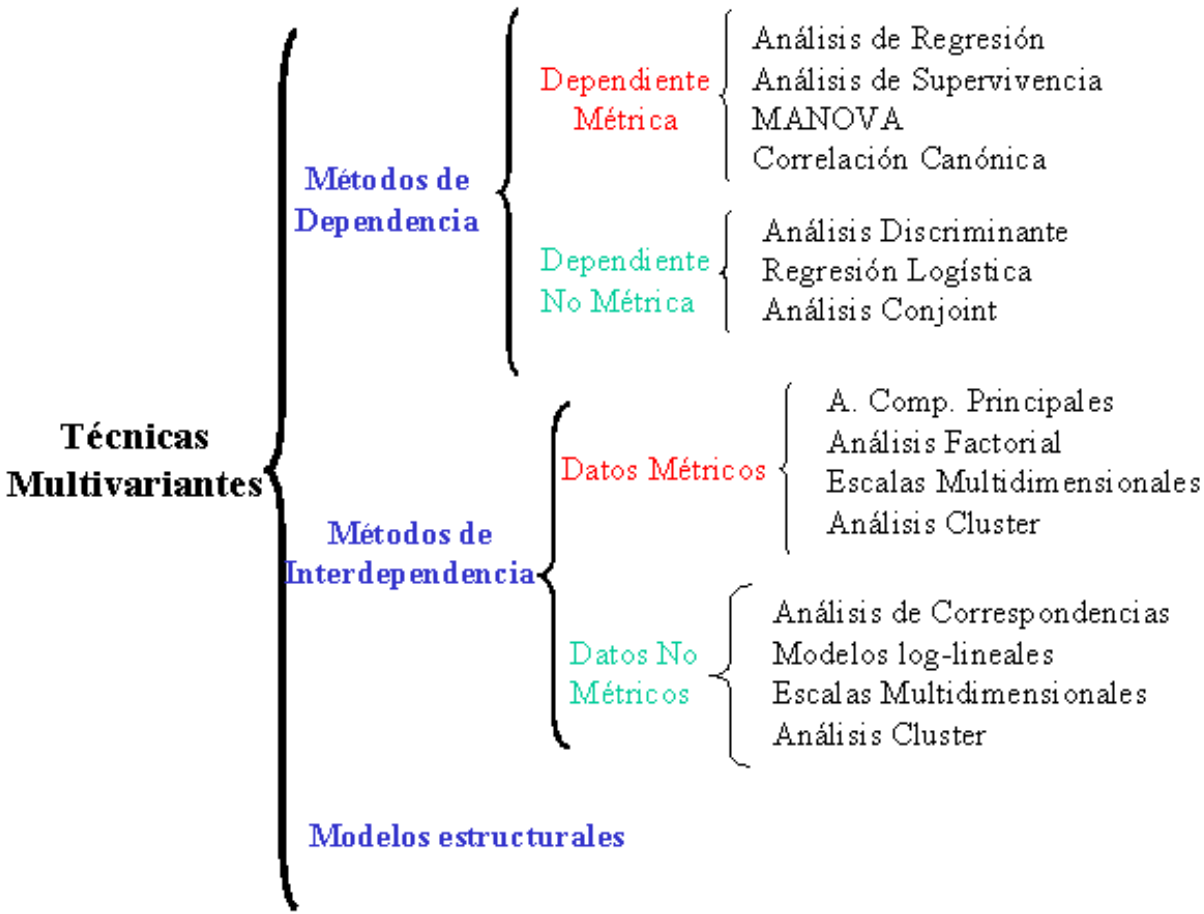


Figure: Técnicas Multivariantes

Datos Atípicos (Outliers)

Datos Atípicos (Outliers)

- Los casos atípicos son observaciones con características diferentes de las demás.
- Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar.
- Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos.
- Aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra.

Datos Atípicos (Outliers) - Tipos

Los casos atípicos pueden clasificarse en 4 categorías.

- La primera categoría contiene aquellos casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.
- La segunda clase es la observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento válido de la población y puede ser eliminado del análisis.

Datos Atípicos (Outliers) - Tipos

- La tercera clase contiene las observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables. Estas observaciones deberían ser retenidas en el análisis pero estudiando qué influencia ejercen en los procesos de estimación de los modelos considerados.
- La cuarta y última clase comprende las observaciones extraordinarias para las que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el por qué de dichas observaciones.

Identificación de Outliers

- Los casos atípicos pueden identificarse desde una perspectiva univariante o multivariante.
- La perspectiva univariante examina la distribución de observaciones para cada variable, seleccionando como casos atípicos aquellos casos cuyos valores caigan fuera de los rangos de la distribución. La cuestión principal consiste en el establecimiento de un umbral para la designación de caso atípico. Esto se puede hacer gráficamente mediante histogramas o diagramas de caja o bien numéricamente, mediante el cálculo de puntuaciones tipificadas.
- Además de la evaluación univariante, pueden analizarse conjuntamente pares de variables mediante un gráfico de dispersión. Casos que caigan manifiestamente fuera del rango del resto de las observaciones pueden identificarse como puntos aislados en el gráfico de dispersión.
- Existen procedimientos para detectar atípicos multivariantes. Dicha detección se puede hacer mediante un Análisis de Componentes Principales.

Datos Ausentes (Missing)

Datos Ausentes

Ver presentación ....

Conclusion

Conclusion

- El Análisis Exploratorio de Datos (AED) es un conjunto de técnicas estadísticas uni y multivariantes cuya finalidad es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.
- El AED proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de datos, el tratamiento y evaluación de datos ausentes, la identificación de casos atípicos y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes (linealidad, normalidad, homocedasticidad).
- Conviene hacer notar, finalmente, la importancia de estas técnicas y la necesidad de perder el tiempo en aplicarlas.
- La experiencia es que un AED hecho en profundidad muestra mucha información acerca de los datos objeto de análisis y que, en muchas ocasiones, la aplicación de técnicas estadísticas más sofisticadas del Análisis Multivariante no hace más que confirmar impresiones iniciales obtenidas a partir de un AED.

Gracias!
?????

Basado en

Salvador Figueras, M y Gargallo, P. (2003): "Análisis Exploratorio de Datos", [en línea] 5campus.com, Estadística <http://www.5campus.com/leccion/aed>