QUE ES LA CALIDAD DE DATOS?

MG. CECILIA ANA RUZ

1

AGENDA

- Que es la calidad de datos
- Medidas Intrínsecas de la calidad
- Una primera aproximación
- Tipos de Variables
- Gráficos Univariados
- Análisis bivariado

2

AGENDA

- **Que es la calidad de datos**
- Medidas Intrínsecas de la calidad
- Una primera aproximación
- Tipos de Variables
- Gráficos Univariados
- Análisis bivariado

3

QUÉ ES CALIDAD DE DATOS?

- La calidad de los datos es la "idoneidad" de los mismos para el uso que se le quiere dar
- Ejemplos:
  - Los datos tienen alta calidad cuando son apropiados para el objetivo que se quiere alcanzar.
  - Los datos tienen baja calidad cuando presentan uno o más problemas que los hacen no aptos para el objetivo
- Es obvio que calidad de datos tiene que ver con limpiar datos sucios, faltantes, incorrectos o inválidos.
- Calidad de datos es un concepto multidimensional que permite evaluar por qué la calidad es buena o mala.

4

CATEGORÍAS DE CALIDAD DE DATOS

- Lógica: grado de adherencia a reglas lógicas de estructuras de datos y atributos y relaciones
  - Un campo con el monto de capital de un préstamo tiene todos valores positivos.
- Semántica: las entidades y atributos representan de manera verídica las entidades y situaciones del mundo real.
  - Fechas de nacimiento de clientes: muchas fechas con valor 01/01/1980
  - Pólizas de accidentes personales por 'fractura de brazo' porque era el valor por defecto en la carga de datos.
- Metadatos: cada entidad y atributos está nomenclado y tiene una definición clara.
  - Ejemplos de definición de cliente:
  - Un cliente es una persona que tiene un registro en la tabla de clientes
  - Un cliente es una persona física que utiliza o utilizó en el pasado por lo menos un producto. Los productos contemplados son préstamos, tarjetas de crédito, caja de ahorro o plazo fijo.

5

AGENDA

- Que es la calidad de datos
- **Medidas Intrínsecas de la calidad**
- Una primera aproximación
- Tipos de Variables
- Gráficos Univariados
- Análisis bivariado

6

DIMENSIONES INTRÍNSECAS DE CALIDAD DE DATOS

- Precisión (accuracy): es la medida en que los atributos reflejan la realidad de los objetos.
- Imprecisión: cuando los datos reflejan valores incorrectos o desactualizados
- Ejemplos:
  - La captura de datos desde un termómetro no refleja la temperatura real por mal funcionamiento
  - Los ingresos del cliente están desactualizados por inflación.
  - El nombre de la calle de una dirección está mal escrita.
  - Al número de teléfono le faltan dígitos

7

7

DIMENSIONES INTRÍNSECAS DE CALIDAD DE DATOS

- Completitud (completenesss): es la presencia o ausencia de características, atributos y relaciones.
  - A nivel negocio
    - ¿Que proporción de las entidades requeridas se encuentra en la base de datos? Por ejemplo, cuantos de nuestros clientes están en la base de clientes.
  - A nivel lógico (modelo de datos):
    - La tabla de cliente los identifica por CUIL/CUIT, pero otras tablas lo hacen por tipo + nro de documento
  - A nivel físico:
    - Tenemos datos de los préstamos, pero están faltando datos de las cuotas de algunos préstamos en la tabla de cuotas.
    - En los archivos de seguros hay seguros dados de baja pero la fecha de baja es nula (el campo existe).

8

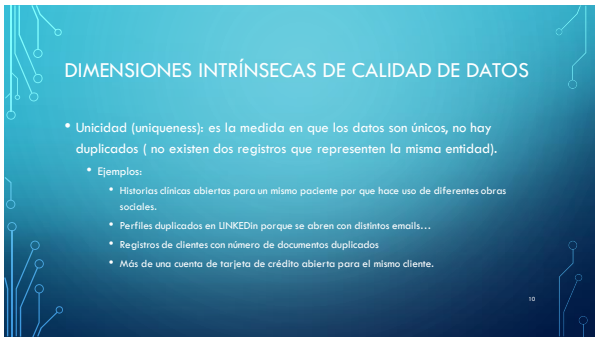
8

DIMENSIONES INTRÍNSECAS DE CALIDAD DE DATOS

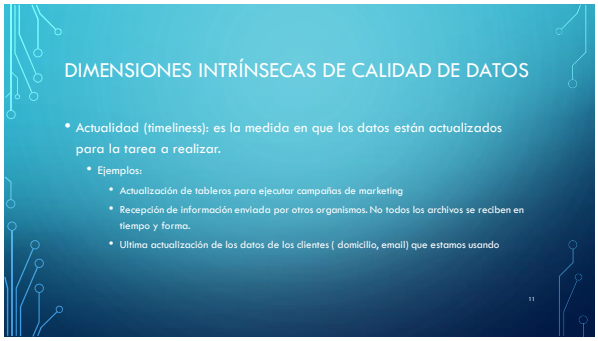
- Consistencia (consistency):
  - es la coherencia de los datos representados cuando los mismos se encuentran en múltiples copias
    - Ejemplos:
      - Cantidad de hijos declarados en el formulario de ganancias versus cantidad de hijos informados por el ANSES
      - Stock en deposito versus sumas de compras – suma de ventas
  - También, es la coherencia cuando se verifican reglas de negocio:
    - Ejemplo: un atributo indica que el cliente no tiene mora, pero en otra tabla la última cuota de un préstamo no tiene fecha de pago y esta vencida.
  - La redundancia suele afectar la consistencia, la única forma de que un valor de siempre igual es que este "almacenado" en un único lugar.
    - Un ejemplo clásico es cuando se decide guardar el saldo de la factura en la tabla factura. Normalmente al comparar la suma de los saldos con la suma de los montos facturados menos los pagos no coinciden

9

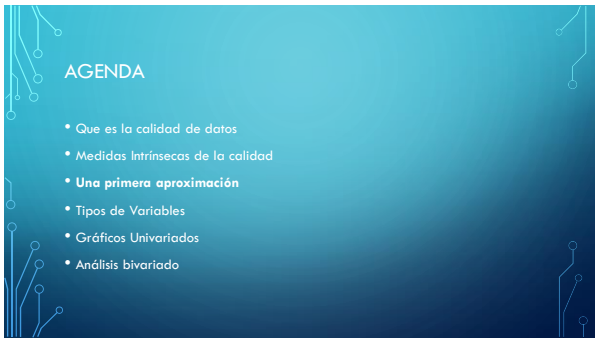
9


DIMENSIONES INTRÍNSECAS DE CALIDAD DE DATOS
• Unicidad (uniqueness): es la medida en que los datos son únicos, no hay duplicados (no existen dos registros que representen la misma entidad).
• Ejemplos:
• Historias clínicas abiertas para un mismo paciente por que hace uso de diferentes obras sociales.
• Perfiles duplicados en LINKEDin porque se abren con distintos emails…
• Registros de clientes con número de documentos duplicados
• Más de una cuenta de tarjeta de crédito abierta para el mismo cliente.
10

10


DIMENSIONES INTRÍNSECAS DE CALIDAD DE DATOS
• Actualidad (timeliness): es la medida en que los datos están actualizados para la tarea a realizar.
• Ejemplos:
• Actualización de tableros para ejecutar campañas de marketing
• Recepción de información enviada por otros organismos. No todos los archivos se reciben en tiempo y forma.
• Ultima actualización de los datos de los clientes (domicilio, email) que estamos usando
11

11


AGENDA
• Que es la calidad de datos
• Medidas Intrínsecas de la calidad
• Una primera aproximación
• Tipos de Variables
• Gráficos Univariados
• Análisis bivariado

12

EJEMPLO

- Supongamos que tenemos la una tabla de Empleado con los siguientes atributos
  - Nombre y Apellido
  - Fecha de Nacimiento
  - Fecha de Ingreso
  - Nacionalidad
  - Sueldo mensual
  - Sexo (1 : Femenino, 2: Masculino)
  - Días de vacaciones por año
  - Tipo y numero de documento
  - Domicilio

13

PRIMEROS PASOS

- Para cada atributo
  - Definir que representa
  - Establecer el dominio
  - Efectuar un análisis descriptivo
- Para el conjunto de atributos
  - Evaluar completitud, precisión, duplicidad.
  - Definir reglas de negocio aplicables a atributos "cruzadas" (por ejemplo que la fecha de ingreso debe ser posterior a la fecha de nacimiento + 18). Esto apunta a determinar consistencia
  - Efectuar un análisis bivariado
- Como se podría determinar la "actualidad"?

14

AGENDA

- Que es la calidad de datos
- Medidas Intrínsecas de la calidad
- Una primera aproximación
- **Tipos de Variables**
- Gráficos Univariados
- Análisis bivariado

15

TIPOS DE VARIABLES

- Categóricas
  - Binarias (SI / NO)
  - Nominales, por ejemplo nacionalidad
  - Ordinales
    - Son las que , a pesar de ser categóricas tienen un orden
    - Por ejemplo Nivel Educativo
- Numéricas

16

ANÁLISIS UNIVARIADO

- Tabla de frecuencia, para cada valor de una variable indica cuantas veces aparece en un conjunto de datos
- Mediana: es el valor medio cuando los valores se ordenan de menor a mayor
- Moda: es el valor mas frecuente
- Media: es el valor que se obtiene al sumar todos los valores y dividirlo por la cantidad de valores.
- Medidas de **dispersión**, parámetros estadísticos que indican como se alejan los datos respecto de la media aritmética. Indican la variabilidad de los datos. Las medidas de **dispersión** más utilizadas son el rango, la desviación estándar y la varianza

17

TIPOS DE OPERACIONES

- Sobre las variables categóricas nominales y binarias se puede "contar", lo que permite hacer una tabla de frecuencia y calcular la moda
- Sobre las variables categóricas ordinales , además de contar, se puede ordenar y calcular los cuartiles y la mediana
- Sobre las variables numéricas, además de las operaciones anteriores se puede calcular el promedio y la dispersión
- El tipo de operación que se puede aplicar **NO DEPENDE DE COMO ESTA ALMACENADA UNA VARIABLE, SINO DE LO QUE REPRESENTA**

18

VEÁMOSLO EN EL EJEMPLO

Variable	Frecuencia	Mediana	Media
Nombre y Apellido	X	-	-
Fecha de Nacimiento	X	X(*)	X(*)
Fecha de Ingreso	X	X(*)	X(*)
Nacionalidad	X	-	-
Sueldo mensual	X	X	X
Sexo	X	-	-
Días de vacaciones por año	X	X	X
Tipo y Numero de Documento	X	-	-
Domicilio	-	-	-

X(*) Suele convertirse a "años" y hacer las operaciones sobre esa variable
El domicilio es un capítulo aparte, vamos a hablarlo mas adelante

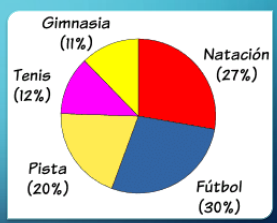
19

AGENDA

- Que es la calidad de datos
- Medidas Intrínsecas de la calidad
- Una primera aproximación
- Tipos de Variables
- **Gráficos Univariados**
- Análisis bivariado

20

GRAFICO DE TORTAS



Fuente: http://www.disfrutalasmatematicas.com/definiciones/grafico-de-pastel-sectores.html

21

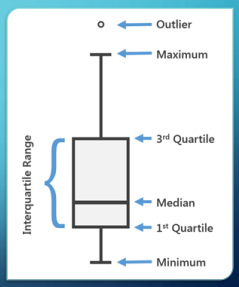
HISTOGRAMA

Fuente:http://jugandoenmiclase.blogspot.com/2011/06/graficos-de-barras.html

Fuente : https://ar.pinterest.com/pin/664140276268372322/

22

BOXPLOT



- Fuente: https://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/charts/box-plot.htm

23

AGENDA

- Que es la calidad de datos
- Medidas Intrínsecas de la calidad
- Una primera aproximación
- Tipos de Variables
- Gráficos Univariados
- Análisis bivariado

24

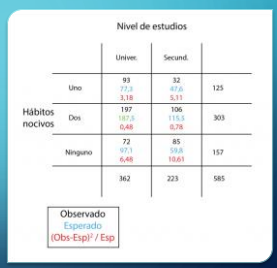
ANÁLISIS BIVARIADO

- Dependiendo del tipo de variables que quiero analizar son las herramientas disponibles

	Categórica	Numérica
Categórica	Tabla de contingencia Grafica de barras apiladas Prueba de Chi Cuadrado	Grafico de barras Boxplot comparativos Histogramas
Numérica	Diferencia de medias	Coeficiente de correlación Diagrama de dispersión

25

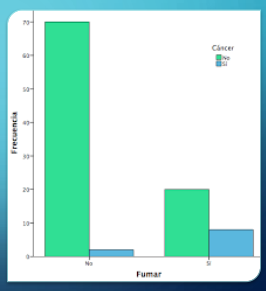
TABLA DE CONTINGENCIA – PRUEBA DE CHI CUADRADO



- Fuente : https://www.cienciasinseso.com/tag/tabla-de-contingencia/

26

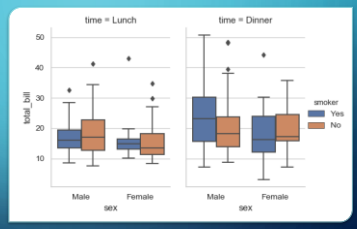
BARRAS



- Fuente : https://estadisticalidia.com/tema-4-analisis-conjunto-de-dos-variables/

27

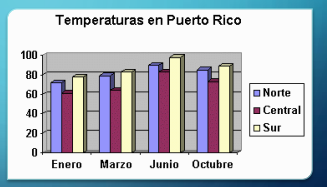
BOX PLOT BIVARIADOS



- Fuente: https://seaborn.pydata.org/generated/seaborn.boxplot.html

28

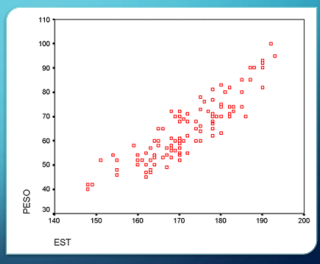
GRAFICO DE BARRAS



Fuente-http://myfaculty.metro.inter.edu/jahumada/amp/histogramas/histogramas_p2.htm

29

DIAGRAMA DE DISPERSIÓN



Fuente: http://www.ub.edu/aplica_infor/spss/cap3-2.htm

30

PREGUNTAS

31