Trabajemos con los datos

Agenda

- Que son los datos?
- Por que preprocesar los datos?
- Limpieza de datos
- Integración y transformación de los datos
- Reducción de datos
- Resumen

Agenda

- **Que son los datos?**
- Por que preprocesar los datos?
- Limpieza de datos
- Integración y transformación de los datos
- Reducción de datos
- Resumen

Que son los datos?

- Colecciones de objetos y sus atributos
- Un atributo es una propiedad o característica de un objeto
  - Ejemplo: posición arancelaria de una mercadería, precio FOB, país de origen.
  - Los atributos también son conocidos como variables, campos, características o aspectos
- Una colección de atributos describe un objeto
  - El objeto es también conocido como registro, punto, caso ejemplo o instancia.

Atributos

Objetos

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Valor de los atributos

- Diferencia entre los atributos y sus valores
- Algunos atributos pueden ser representados por medio de distintos valores
  - Ejemplo: la cantidad puede ser expresada en unidades estadísticas o en unidades comerciales

- Atributos distintos pueden tener el mismo «dominio» (conjunto de valores posibles)
  - Ejemplo: el peso de la mercadería y el precio son ambos números reales positivos.

Tipos de Atributos

- Hay
  - Nominal
    - Ejemplos: posición arancelaria, país de origen.
  - Ordinal
    - Ejemplos: rangos, nivel de consumo, nivel educativo
  - Continuo
    - Ejemplos: precio unitario, peso de la mercadería

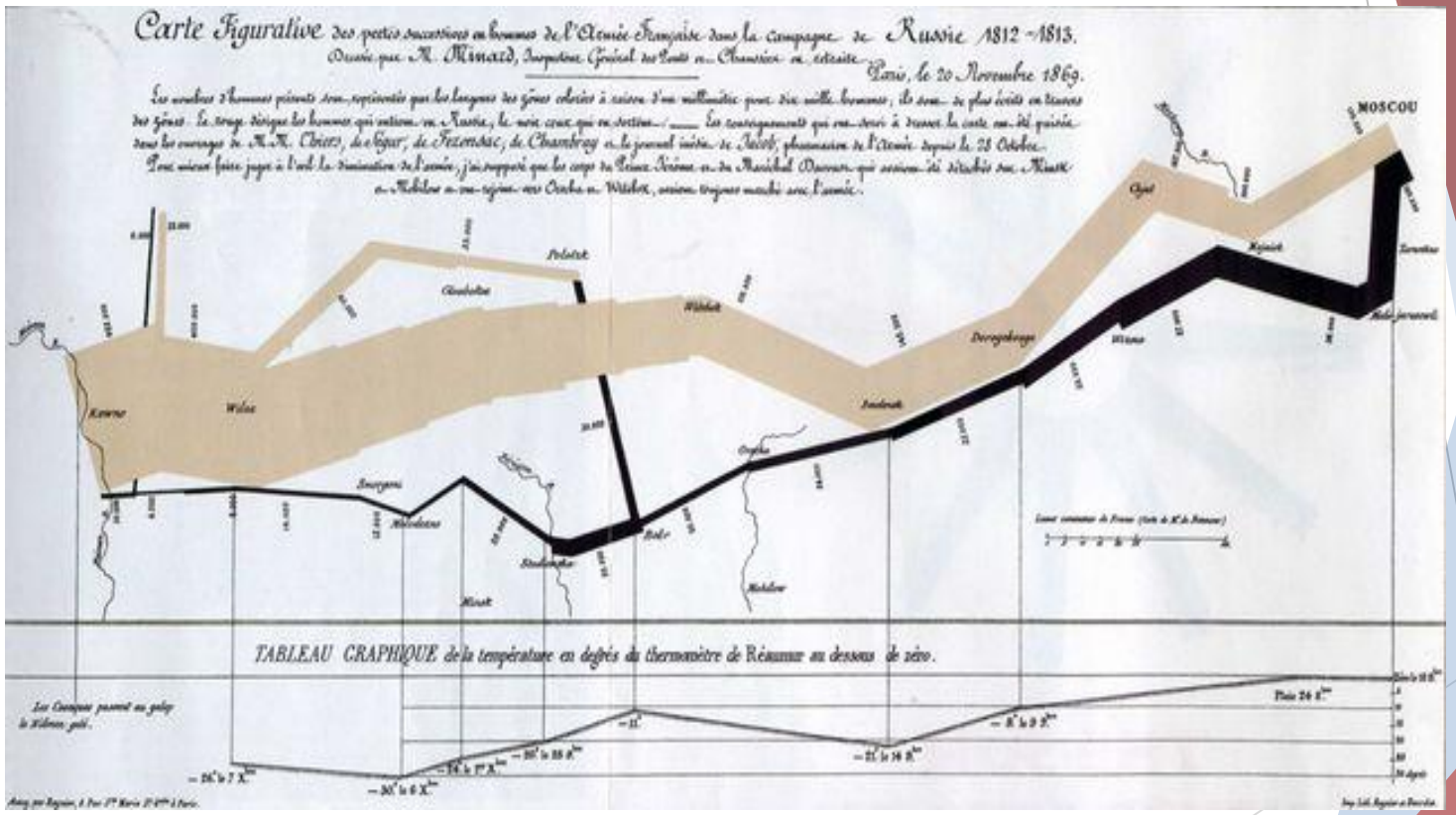
Atributos Discretos y Continuos

- Atributos Discretos
  - Los atributos ordinales y nominales son de este tipo.
  - Tienen un numero finito de valores o un conjunto numerable (que puede «emparejarse» con los números naturales)
  - Ejemplos: puertos, países de origen
  - Normalmente se representan como números enteros.
  - Los atributos binarios (o verdadero y falsos) son un tipo especial de atributos discretos.
- Atributos continuos
  - Su dominio son los números reales, normalmente de tipo punto flotante
  - Ejemplos : precios, peso

Como analizar los distintos tipos de datos?

- Medidas estadísticas que ayudan a la comprensión
- Gráficos, visualizaciones…

Una imagen vale mas que mil palabras, la Gioconda de la visualización



Variables Discretas

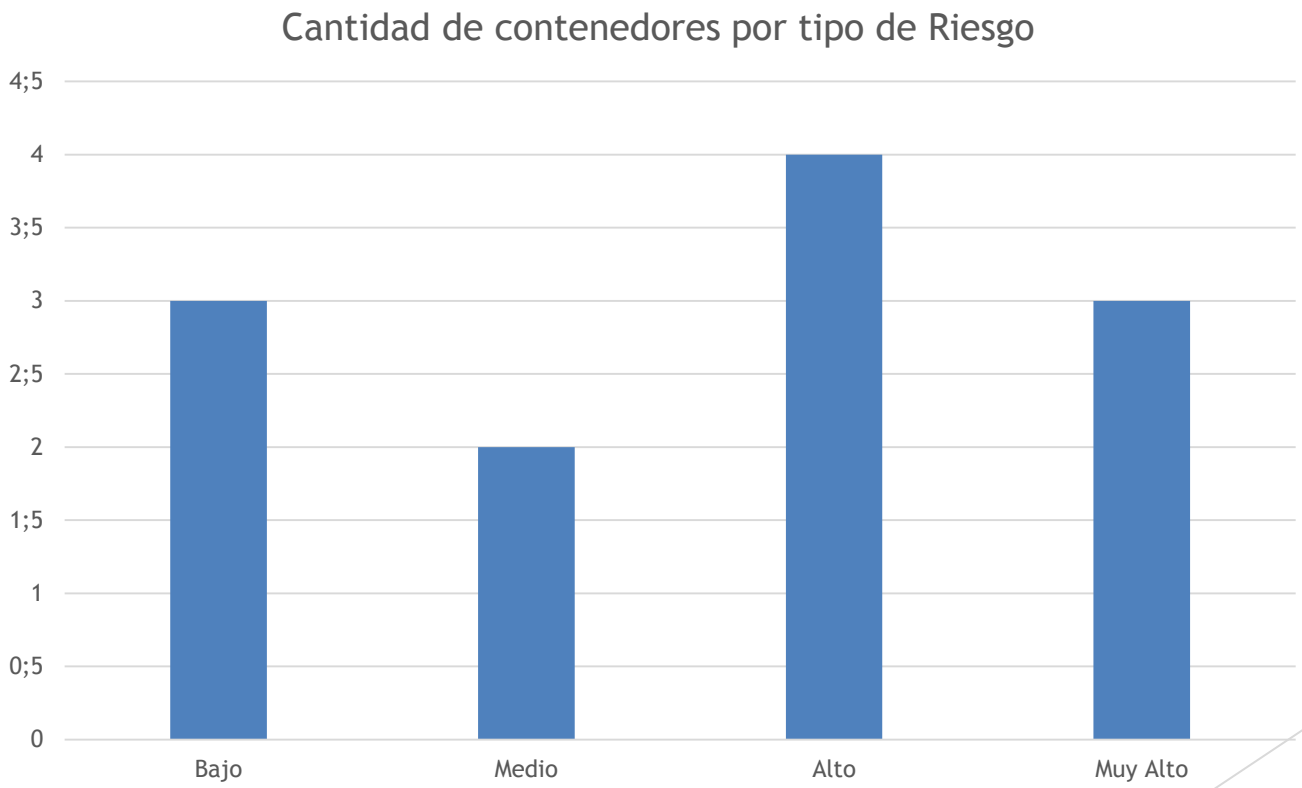
Medidas Estadísticas

- Moda: valor mas frecuente. Cual es el «principal» país de origen desde el que se importan los juguetes?
- Si son ordinales tiene sentido calcular la mediana y los cuartiles
  - La mediana es el «valor del medio»
  - El primer cuartil es el valor que se ubicar en «un cuarto de la tira de datos»
  - La mediana es el segundo cuartil
  - EL tercer cuartil es el valor que se ubicar en «la tercera parte de los datos»
- Veamos un ejemplo……

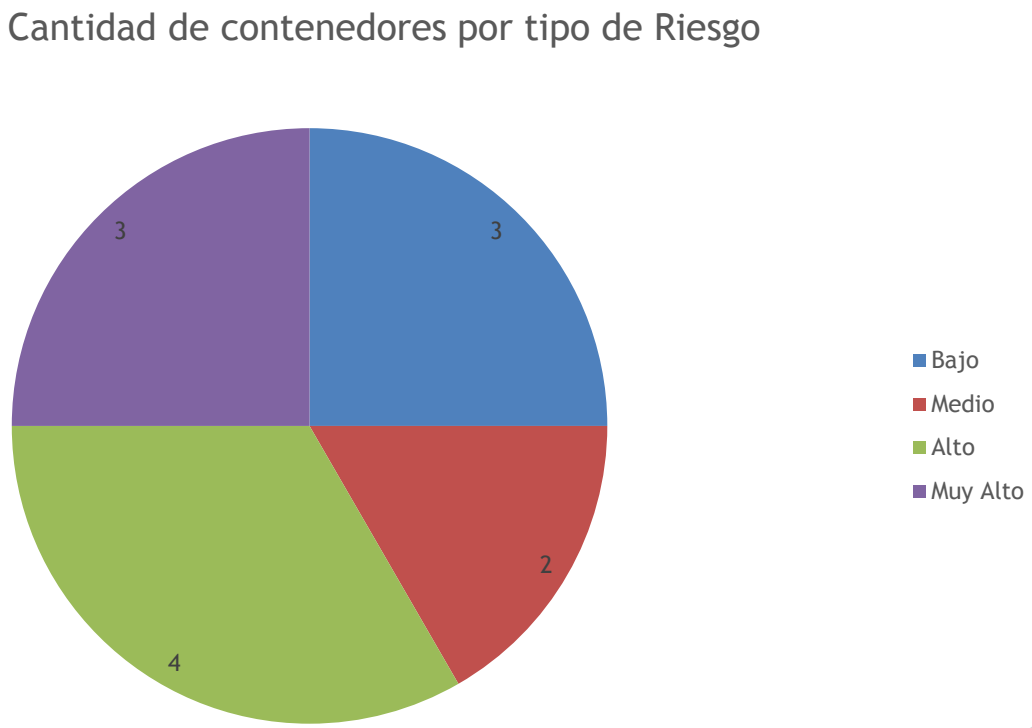
Ejemplo Medidas Estadísticas

- Supongamos que el riesgo asociado a una carga puede clasificarse en Bajo, Regular, Medio, Alto y Muy Alto
- Y que los 12 contenedores de un barco se clasificaron de la siguiente forma (ya ordenado por riesgo creciente y separados de a 3)
- Bajo, Bajo, Bajo,
- Medio, Medio, Alto,
- Alto, Alto, Alto,
- Muy Alto, Muy Alto, Muy Alto
- La moda es Alto
- El primer cuartil es Bajo, la mediana (segundo cuartil) es Alto y el tercer cuartil es Alto.
- Traducido en lenguaje común, en este ejemplo se puede decir que la mitad de los contenedores de ese barco son de riesgo «Alto» o «Muy Alto»

Como se puede graficar una variable discreta (1/ 2) ?



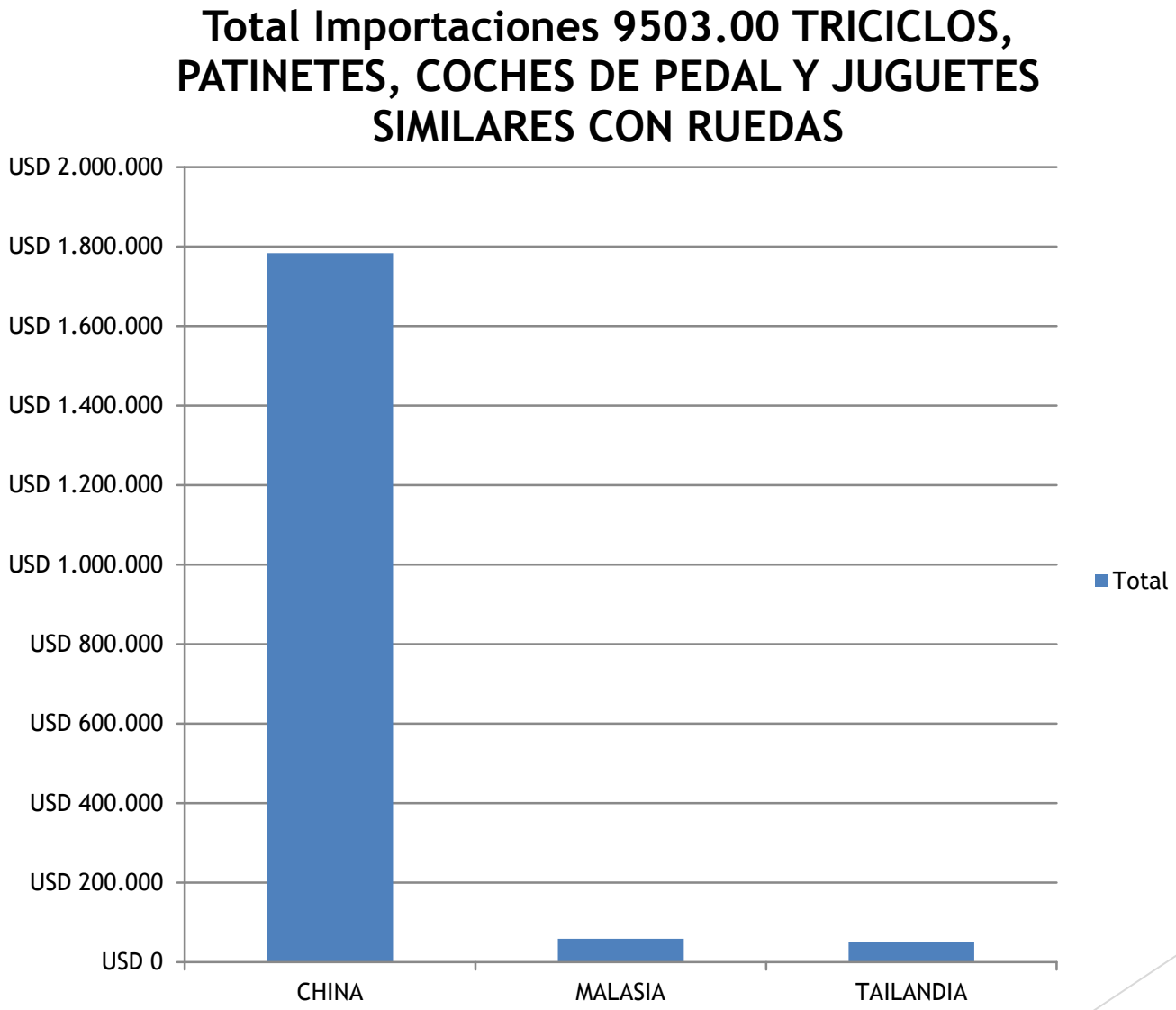
Como se puede graficar una variable discreta (2/ 2) ?

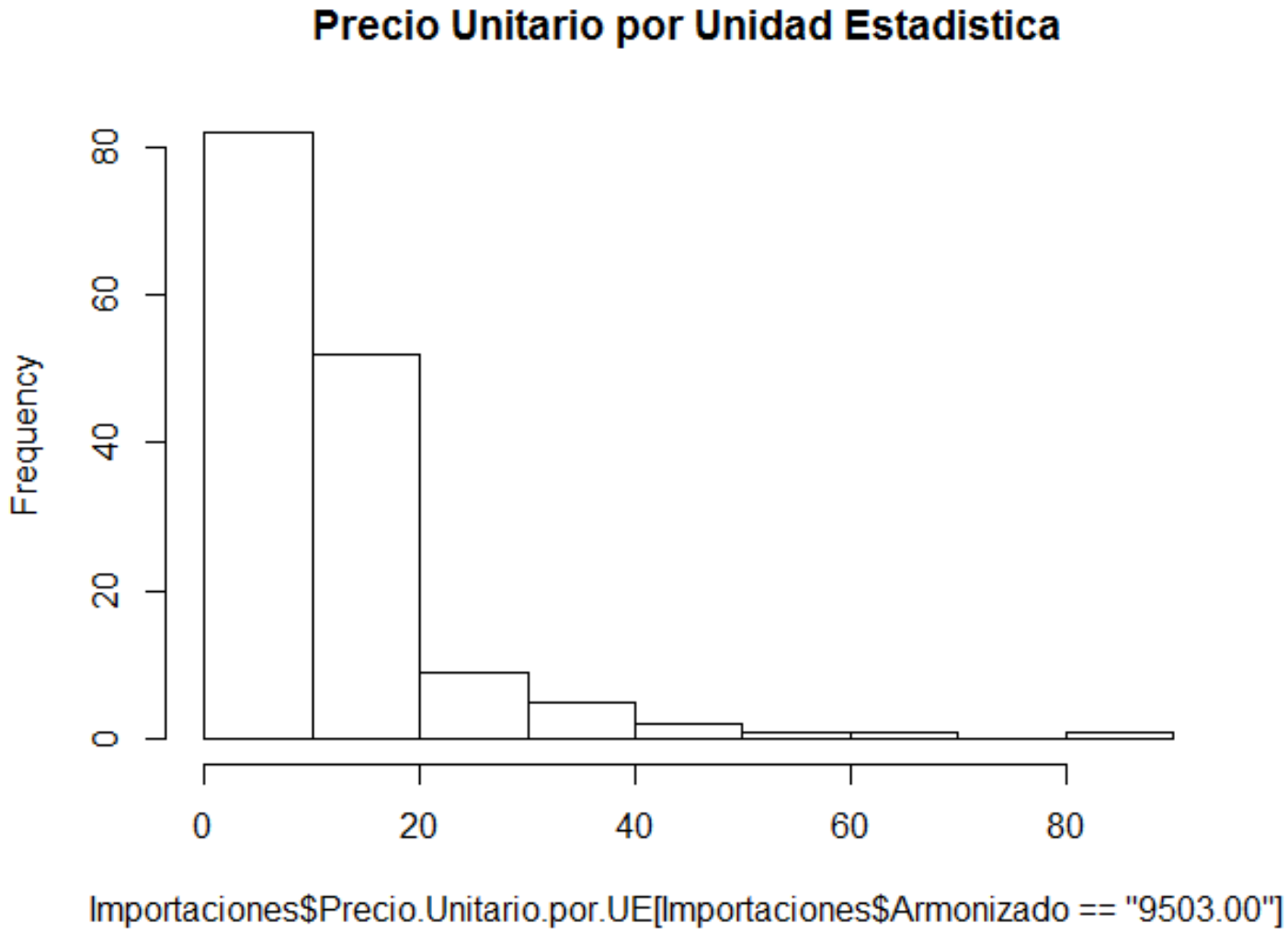


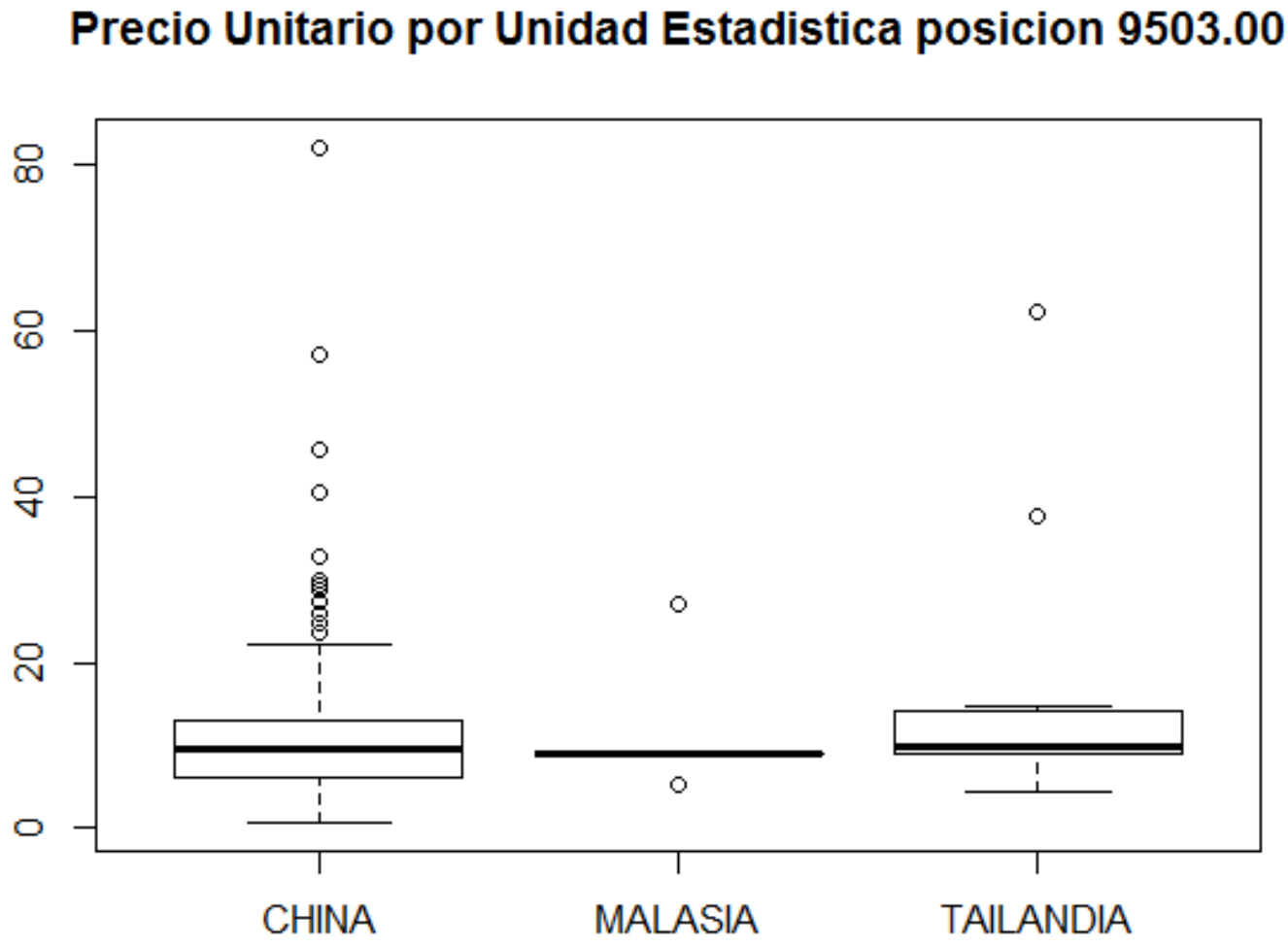
Variables Continuas

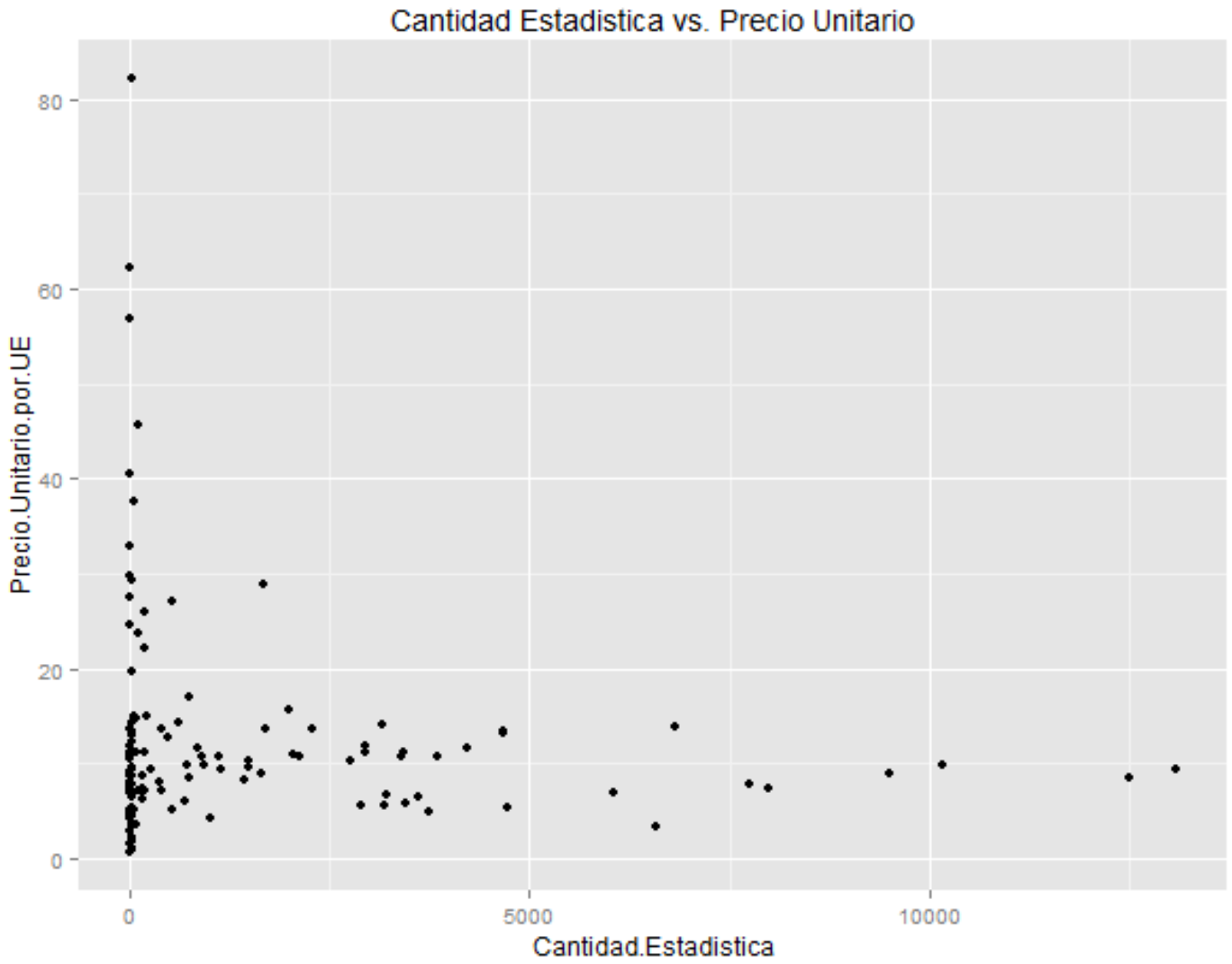
Medidas Estadísticas

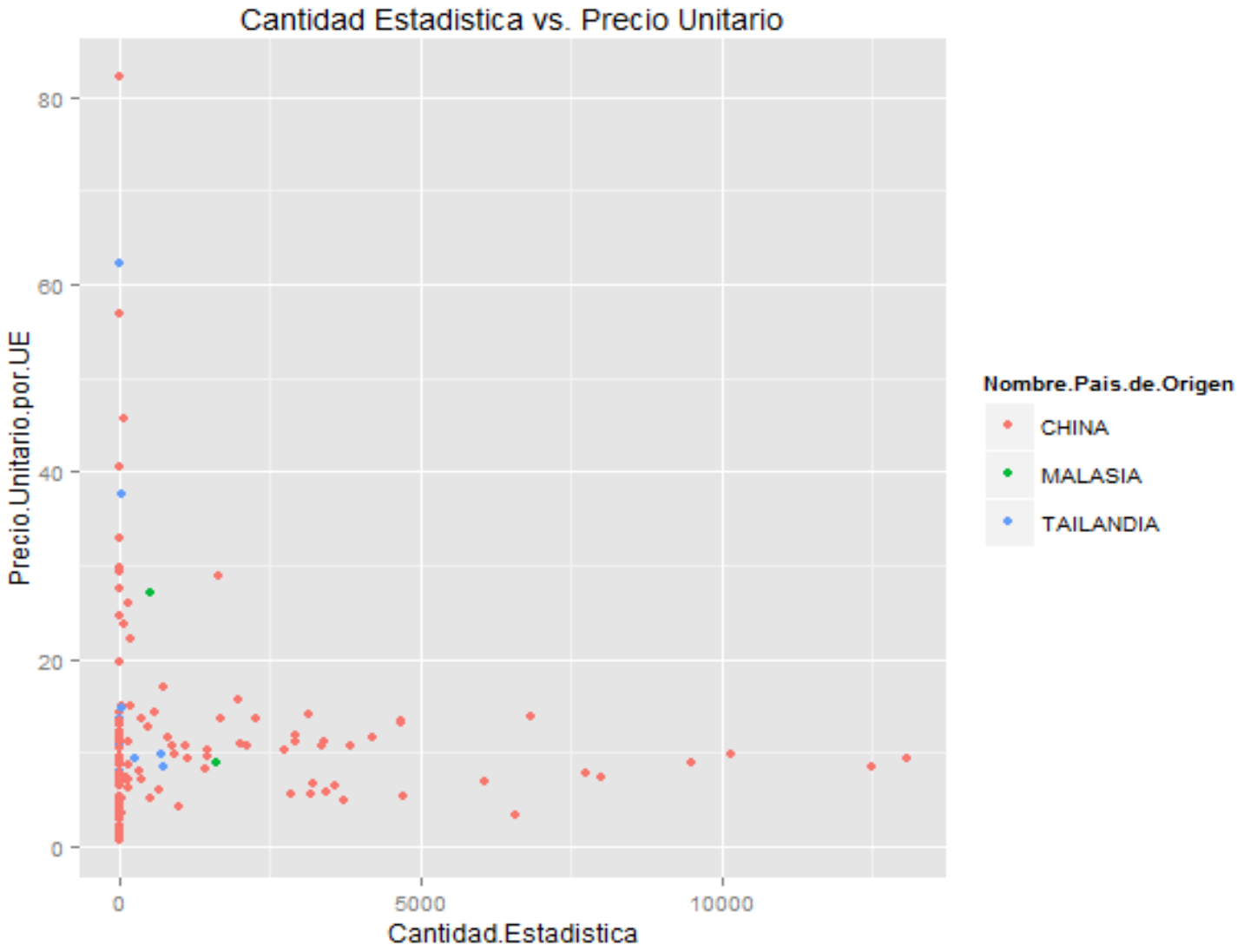
- Además de la moda, la mediana y los cuartiles
- Promedio
- Desviación Standard: cuan dispersos están los datos?


Total Importaciones 9503.00 TRICICLOS, PATINETES, COCHES DE PEDAL Y JUGUETES SIMILARES CON RUEDAS
USD 2.000.000
USD 1.800.000
USD 1.600.000
USD 1.400.000
USD 1.200.000
USD 1.000.000
USD 800.000
USD 600.000
USD 400.000
USD 200.000
USD 0
CHINA
MALASIA
TAILANDIA
Total


Precio Unitario por Unidad Estadistica
Frequency
0
20
40
60
80
0
20
40
60
80
Importaciones$Precio.Unitario.por.UE[Importaciones$Armonizado == "9503.00"]


Precio Unitario por Unidad Estadistica posicion 9503.00
0
20
40
60
80
CHINA
MALASIA
TAILANDIA


Cantidad Estadistica vs. Precio Unitario
Precio.Unitario.por.UE
80
60
40
20
0
0
5000
10000
Cantidad.Estadistica


Cantidad Estadistica vs. Precio Unitario
Precio.Unitario.por.UE
80
60
40
20
0
0
5000
10000
Cantidad.Estadistica
Nombre.Pais.de.Origen
CHINA
MALASIA
TAILANDIA

Tipos de data sets

- **Registro**
  - Matriz de datos
  - Documentos
  - Datos de transacciones
- **“Semi estructurado”**
  - XML, Jason
  - Non sql databases
- **Grafos**
  - Web
  - Estructuras moleculares
  - Redes sociales
- **Ordenados**
  - Datos Espaciales
  - Datos Temporales
  - Datos secuenciales
  - Stream Data

Registros

- Los datos consisten en un conjunto de registros, cada uno de los cuales consiste en un conjunto **fijo** de atributos

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Documentos

- Cada documento se representa como un vector de términos
  - Cada termino es un atributo del vector,
  - El valor de cada componente es la cantidad de veces que el termino aparece en el documento

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Datos de transacciones

- Un tipo especial de registro
  - Cada registro (transacción) involucra un conjunto de ítems
  - Por ejemplo las posiciones que se encuentran en una misma importacion

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

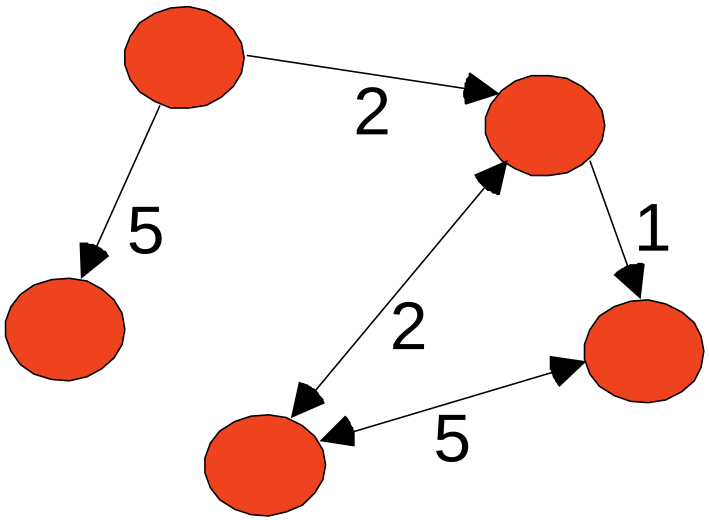
Semi estructurados, XML

- XML: es un lenguaje de marcación desarrollado por la WWW.
- Es de tipo jerárquico y se utiliza mucho para el intercambio de información.
- Existe una manera de «validar» el contenido mediante el uso de .xsd
- El intercambio de la OMA esta previsto en XML

- VersionSoloBorderTransportMean.xml

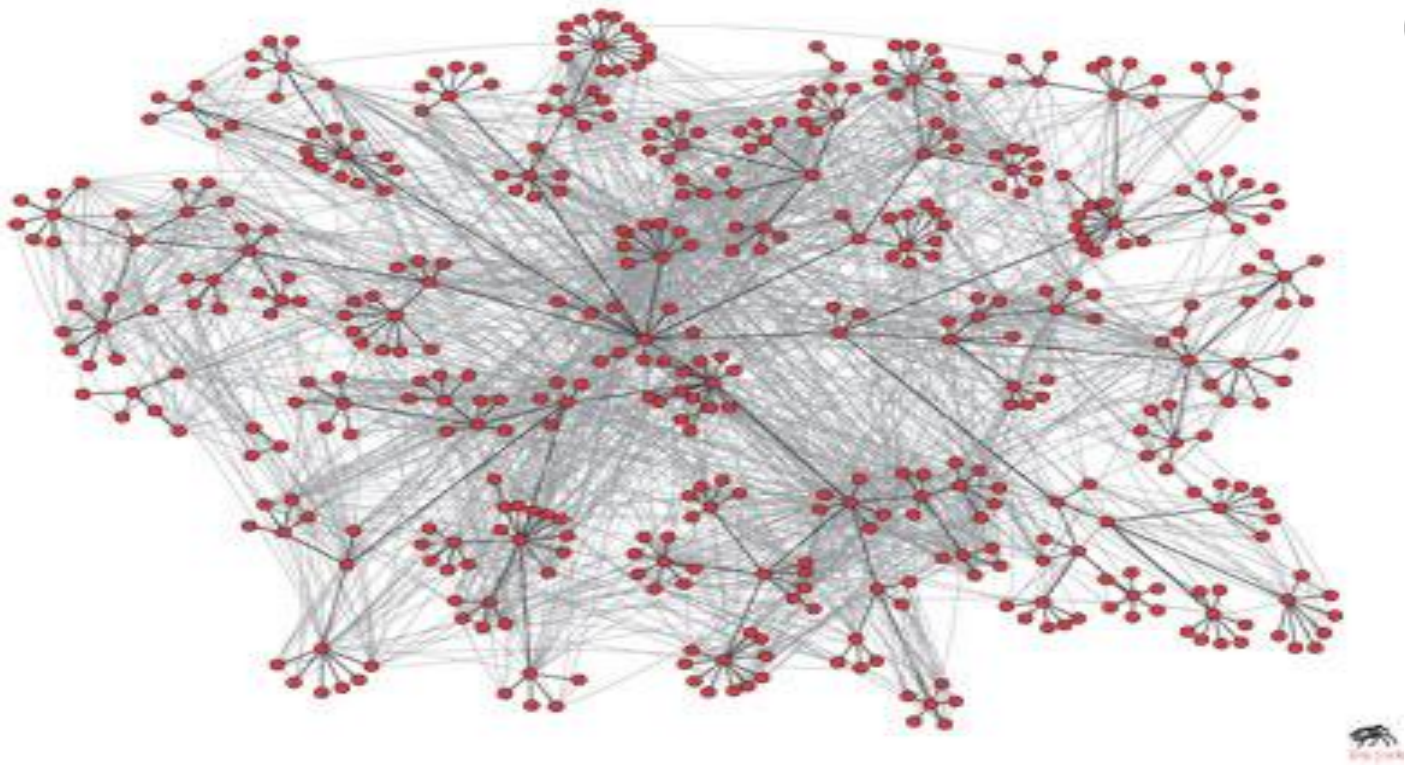
Grafos

- Ejemplos: Modo en que se vinculan las paginas




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

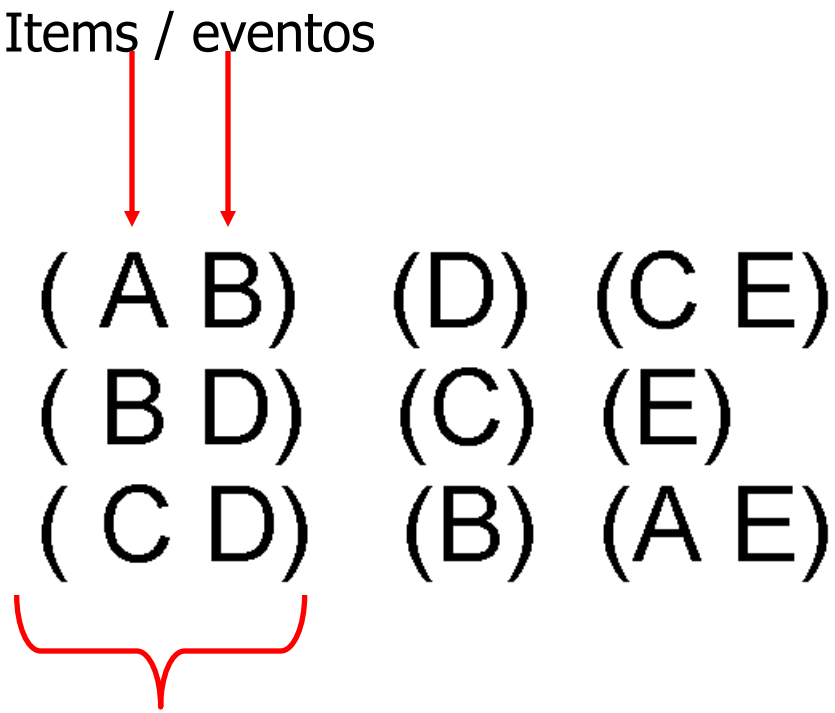
Redes Sociales



Patrón de intercambio del email en el laboratorio de investigación de Hewlett Packard superpuesto con la estructura de la organización . (Image from http://wwwpersonal.umich.edu/ ladamic/img/hplabsemailhierarchy.jpg)

Datos Ordenados

- Secuencia de transacciones



Datos Ordenados

Stream Data

- Los datos de tipo stream fluyen por un sistema de computadora en forma continua y con distintas velocidades.
- Están ordenados temporalmente, cambian rápidamente, son masivos y potencialmente infinitos
- IoT (internet of things) : para el año 2020 se calcula que va a haber 75 billones de dispositivos conectados.

Agenda

- Que son los datos?
- **Por que preprocesar los datos?**
- Limpieza de datos
- Integración y transformación de los datos
- Reducción de datos
- Resumen

Por que preprocesar los datos?

- Los datos del mundo real están “sucios”
  - incompletos: falta de valores en algunos atributos falta de atributos, datos que están solo agrupados.
  - “Ruidosos”: con errores u outliers
  - Inconsistentes : con discrepancias en los códigos o en los nombres
- Garbage in, garbage out!

La calidad de datos es multidimensional

- Algunos elementos:
  - Precisión
  - Completitud
  - Consistencia
  - En tiempo
  - Creíble
  - Agrega Valor
  - Entendible
  - Accesible

Principales Tareas en el Preprocesamiento

- Limpieza
  - Completar datos faltantes, suavizar el ruido, identificar outliers y resolver inconsistencias
- Integración
  - Integración de múltiples fuentes de datos
- Transformación
  - Normalización y Sumarización
- Reducción
  - Reduce el volumen, pero produce los mismos resultados
- Discretizacion
  - Transformación de variables numéricas en categorías

Agenda

- Que son los datos?
- Por que preprocesar los datos?
- **Limpieza de datos**
- Integración y transformación de los datos
- Reducción de datos
- Resumen

Limpieza de datos– Ejemplos de errores

- Fuera de Rango: Edad del Paciente= 185 ()
- No-Standard: Data Main Str, Main Street, Main ST, Main St.
- Datos inválidos: El dato puede ser “A” o “B” pero el valor es “C”
- Reglas culturales diferentes: Fecha= Enero1, 2002 o 1-1-2002 o 1 Ene 02
- Distintos Formatos: (919)674-2153 o [919]6742153 o 9196742153
- Cosméticos: jon j jones transformado en Jon J Jones
- Verificación: El código postal no corresponde a la ciudad o la dirección ingresada

Datos Faltantes

- Los datos faltantes pueden deberse a
  - Problemas en los equipos o en los programas
  - Inconsistencia con otras fuentes y por lo tanto se eliminaron
  - Los datos no se consideraron relevantes al momento de la carga (email en la cadena de electrodomésticos)
  - No se registra la historia de los cambios

Como manejar los datos faltantes

- Ignorar el registro : no puede hacerse si el porcentaje de atributos faltantes cambia mucho de un atributo a otro
- Completarlos
- Crear una clase nueva para los valores faltantes (“desconocido”). Esto es porque algunos algoritmos no pueden tratar los atributos con valores faltantes
- Completar los valores faltantes usando algún algoritmo de data Mining

Ruido

- Ruido: errores aleatorios en alguna variable
- Los valores incorrectos pueden deberse a
  - Errores en los instrumentos de recolección
  - Errores de data entry
  - Errores en la transmisión
  - Limitaciones tecnológicas
  - Inconsistencias en la forma de nombrar los objetos
- Otros problemas que requieren limpieza
  - Registros duplicados
  - Datos incompletos
  - Datos inconsistentes

Como manejar el ruido?

- Análisis univariado y bivariado
- Clustering
  - Detectar outliers
- Combinar técnicas automáticas y manuales
  - Detectar valores sospechosos y chequearlos manualmente

Agenda

- Que son los datos?
- Por que preprocesar los datos?
- Limpieza de datos
- **Integración y transformación de los datos**
- Reducción de datos
- Discretizacion y generación de jerarquía de conceptos
- Resumen

Integración de Datos

- Integración de datos :
  - Combina datos de múltiples fuentes en un único almacenamiento
- Integración de “Esquemas”
  - Integra los metadatos de diferentes fuentes
  - Problema de la identificación de entidades. Reconocer que A.cust-id ≡ B.cust-#
- Detección y resolución de conflictos de valores de datos
  - Para la misma entidad del mundo real los valores provenientes de distintas fuentes no coinciden.
  - Algunos motivos
    - Diferentes unidades de medida
    - Diferencias en la actualización de los datos, alguna fuente de datos se actualizo y otra no…

Transformacion de datos

- “Suavizar”: remover el ruido de los datos
- Agregación : Sumarización ,armado de cubos
- Generalización: subir en la jerarquía de conceptos, por ejemplo reemplazar un producto por su rubro
- Normalización
- Construcción de atributos
  - Atributos derivados de los existentes

Agenda

- Que son los datos?
- Por que preprocesar los datos?
- Limpieza de datos
- Integración y transformación de los datos
- **Reducción de datos**
- Resumen

Estrategias de Reducción de datos

- Los warehouse pueden tener terabytes de data, los análisis complejos pueden tardar mucho tiempo en correr en el dataset completo por eso se hacen necesarias técnicas de reducción de datos
- Reducción de datos
  - Obtener una representación reducida de los datos , que a pesar de tener mucho menos volumen produce el mismo resultado al aplicar técnicas de data mining
- Estrategias de reducción de datos
  - Cubos
  - Selección de atributos
  - Reducción de los casos (muestreo)
  - Discretizacion y generación de jerarquía de conceptos

Agenda

- Que son los datos?
- Por que preprocesar los datos?
- Limpieza de datos
- Integración y transformación de los datos
- Reducción de datos
- **Resumen**

Trabajemos con los datos

Muchas gracias!