Universidad Autónoma de Baja California

Maestría y Doctorado en Ciencias e Ingeniería

2. Conociendo tus datos

Minería de Datos

Conociendo tus datos…

- Conocimiento sobre tus datos es útil para la etapa de preprocesamiento de datos, el cúal es la primer etapa principal durante el proceso de minería.
- Querrás saber lo siguiente:
  - ¿Qué tipos de atributos o campos componen a tus datos?
  - ¿Qué tipo de valores contiene cada atributo de tus datos?
  - ¿Cuáles atributos son discretos, y cuáles son valores continuos?
  - ¿Cómo son tus datos?
  - ¿Cómo están distribuidos tus datos?
  - ¿Hay forma de visualizar los datos para entender mejor el problema?
  - ¿Podemos indentificar datos anómalos?
  - ¿Podemos identificar la similitud de algunos objetos de datos con respecto otros?
  - ¿Entendiendo más los datos ayudará con el análisis que se realiza?

Objetos de datos y tipos de atributos

- Los datasets, o bases de datos, se componen de objetos de datos.
- Un objeto de datos representa una entidad.
- Los objetos de datos típicamente son descritos por atributos.
- A los objetos de datos también se les pueden llamar:
  - Muestras
  - Ejemplos
  - Objetos
  - Instancias
  - Etc.

¿Qué es un atributo?

- Un atributo es un campo que representa alguna característica o rasgo de un objeto de datos.
- También se le pude decir:
  - Atributo. Término usado por profesionales de minería de datos y bases de datos.
  - Dimensión. Término usado en almacenes de datos.
  - Característica (feature). Término usado en aprendizaje máquina.
  - Variable. Término usado por estadístas
- El tipo de atributo es determinado por el conjunto de posibles valores
  - Nominales
  - Binarios
  - Ordinales
  - Numéricos

Ejemplo de atributos e instancias

no. instancia	estatura (m)	edad (años)	ingreso_anual (pesos)	sexo
1	1.67	34	800000	mujer
2	1.8	23	80000	hombre
3	1.55	67	200000	mujer
4	1.76	34	150000	hombre
5	1.6	45	900000	mujer

Atributos nominales

- Por nominal se entiende ”perteneciente o relativo al nombre”.
- Los valores en atributos nominales son *símbolos* o *nombre de cosas*.
- Cada valor representa algun tipo de categoría, código, estado, etc.
- También son referidos como atributos de categoría.

no. instancia	estatura (m)	edad (años)	Ciudad de residencia	sexo
1	1.67	34	Tijuana	mujer
2	1.8	23	Tecate	hombre
3	1.55	67	Ensenada	mujer
4	1.76	34	Ensenada	hombre
5	1.6	45	Mexicali	mujer

Atributos binarios

- Un atributo binario es un atributo nomonal con sólamente dos categorías o estados: 0 o 1, donde 0 típicamente significa que el atributo no está presente, y 1 significa que está presente.
- A los atributos binarios también se les refere como booleanos, donde cada uno de sus estados corresponde a verdadero y falso.

no. instancia	estatura (m)	edad (años)	ingreso_anual (pesos)	Estado credicitico
1	1.67	34	800000	bien
2	1.8	23	80000	mal
3	1.55	67	200000	bien
4	1.76	34	150000	bien
5	1.6	45	900000	mal

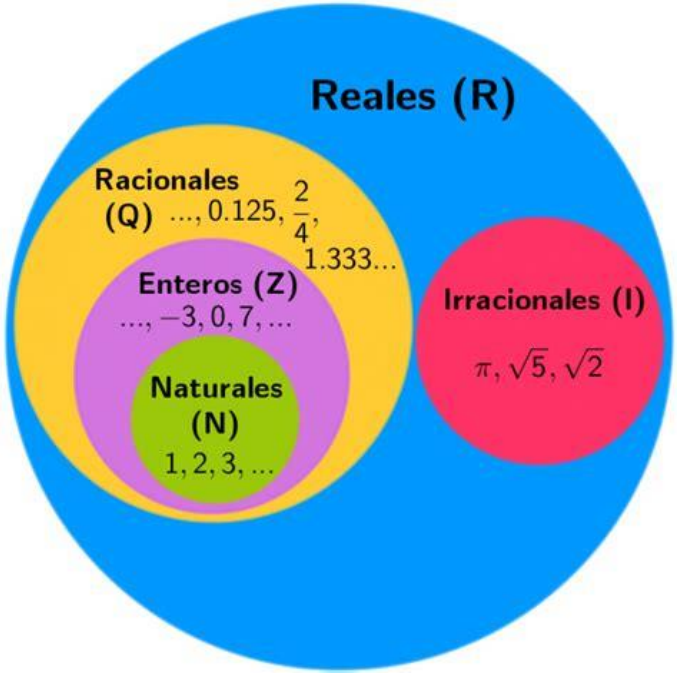
Atributos ordinales

- Un atributo ordinal es un atributo con posibles valores que tienen un orden significativo o una jerarquía entre ellos, pero la magnitud entre los valores sucesivos es desconocido.
- Los atributos ordinales también pudieran ser obtenidos mediante la discretización de cantidades numéricas dividiendo el rango de valores en una cantidad finita de categorías numéricas ordenadas.

no. instancia	estatura (m)	edad (años)	Carrera que estudia	sexo
1	1.67	34	7001	mujer
2	1.8	23	7005	hombre
3	1.55	67	4501	mujer
4	1.76	34	1002	hombre
5	1.6	45	8211	mujer

Atributos numéricos

- Un atributo numérico es quantitativo; esto es, es una cantidad medible representada por valores enteros o reales.



Atributos discretos vs continuos

- Un atributo discreto tiene una cantidad finita de valores posibles, el cual puede ser o no representado por enteros.
- Si un atributo no es discreto, por lo tanto es continuo.
- En práctica, los atributos discretos son representados por *enteros*, mientras que los continuos son representados por números *flotantes*.

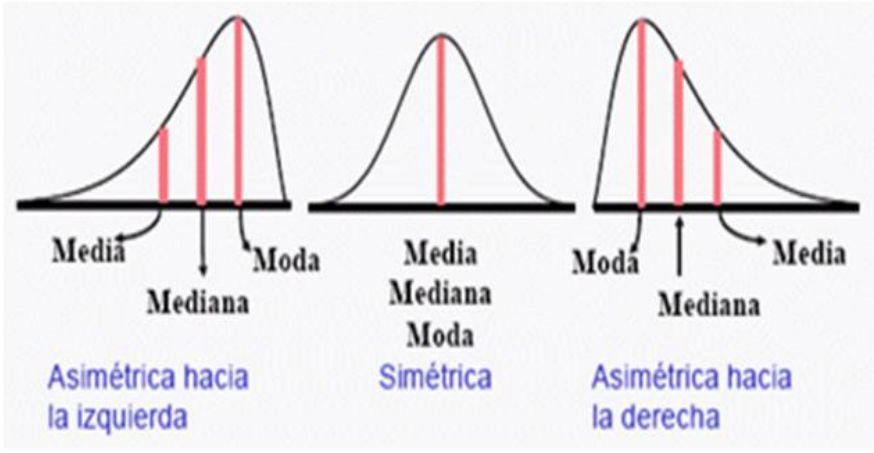
no. instancia	estatura (m)	edad (años)	ingreso_anual (pesos)	sexo
1	1.67	34	801,235.15	mujer
2	1.8	23	80,998.76	hombre
3	1.55	67	207,789.01	mujer
4	1.76	34	157,124.64	hombre
5	1.6	45	900,000.01	mujer

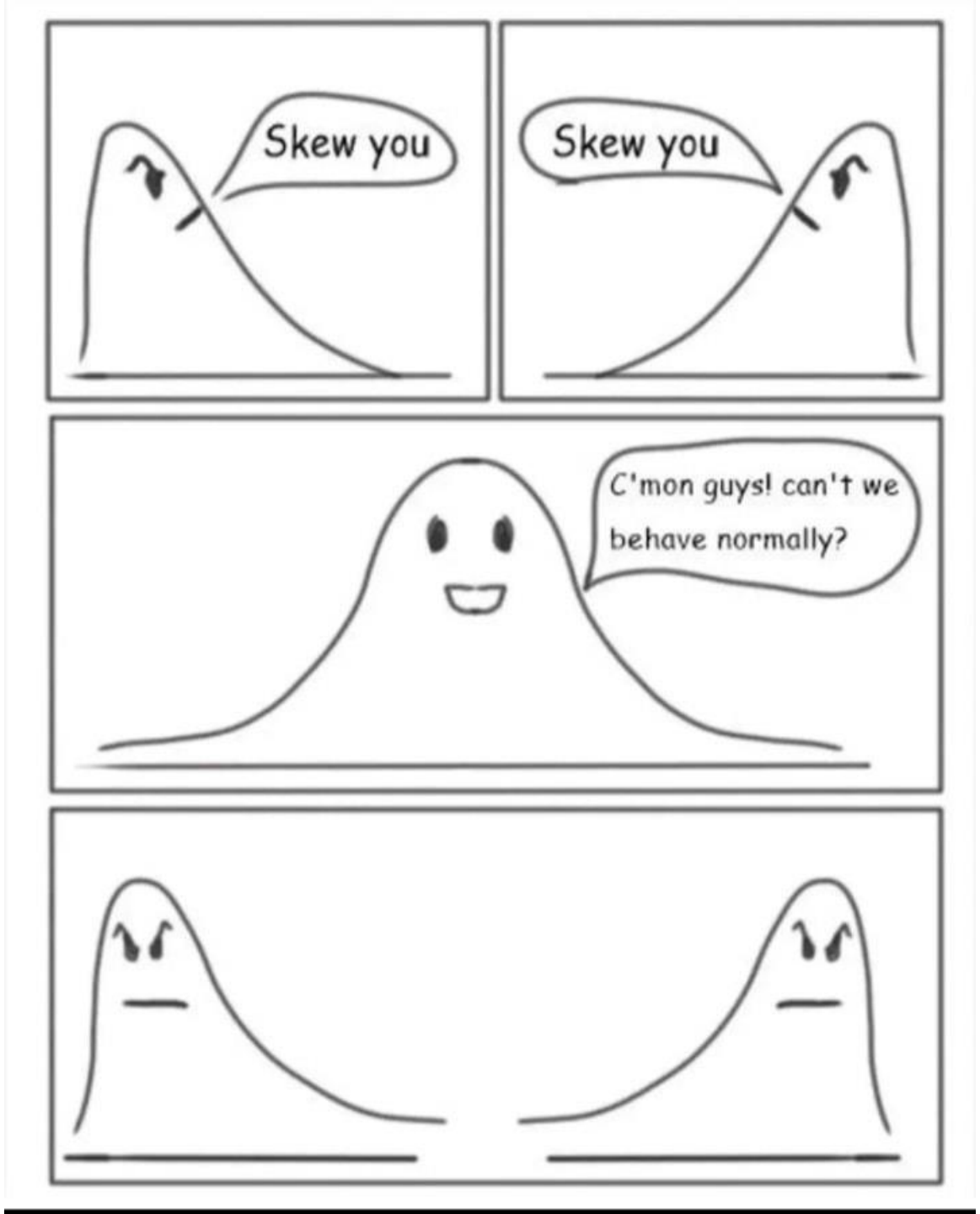
Descripciones estadísticas básicas de datos

- Descripciones estadísticas básicas pueden ser usados para identificar propiedades en los datos y resaltar cuales valores en los datos pudieran ser tratados como datos anómalos.

Midiendo la tendencia central: promedio (media), mediana y moda

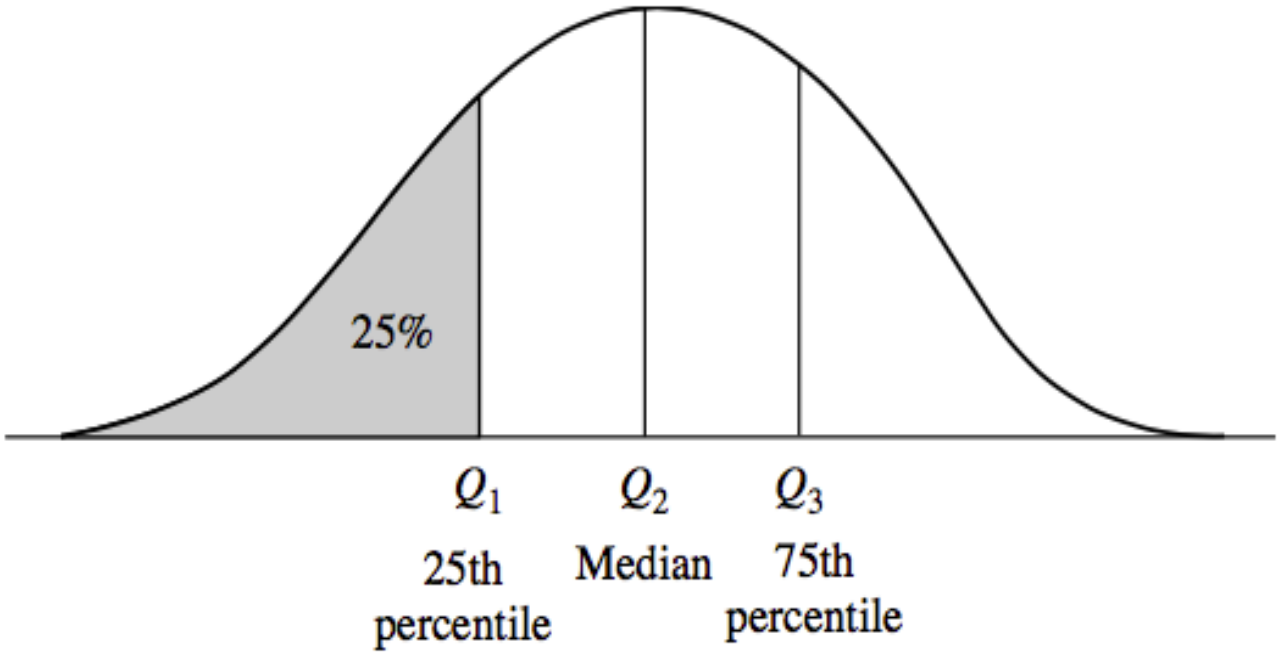
- La medida numérica más común y efectiva es la medida del centro e un conjunto de datos, esto es, el promedio.
- Un problema con promedios es que son sensibles a datos anómalos, una pequeña cantidad de éstos datos pueden corromper el promedio.
- Para datos sesgados, una mejor medida del centro de los datos es la mediana, que es el valor central en un conjunto de datos ordenados.
- La moda en un conjunto de datos es el valor que ocurre con mayor frecuencia.




Skew you
Skew you
C'mon guys! can't we behave normally?

Midiendo la dispersión: rango, cuantiles, y rango intercuantil

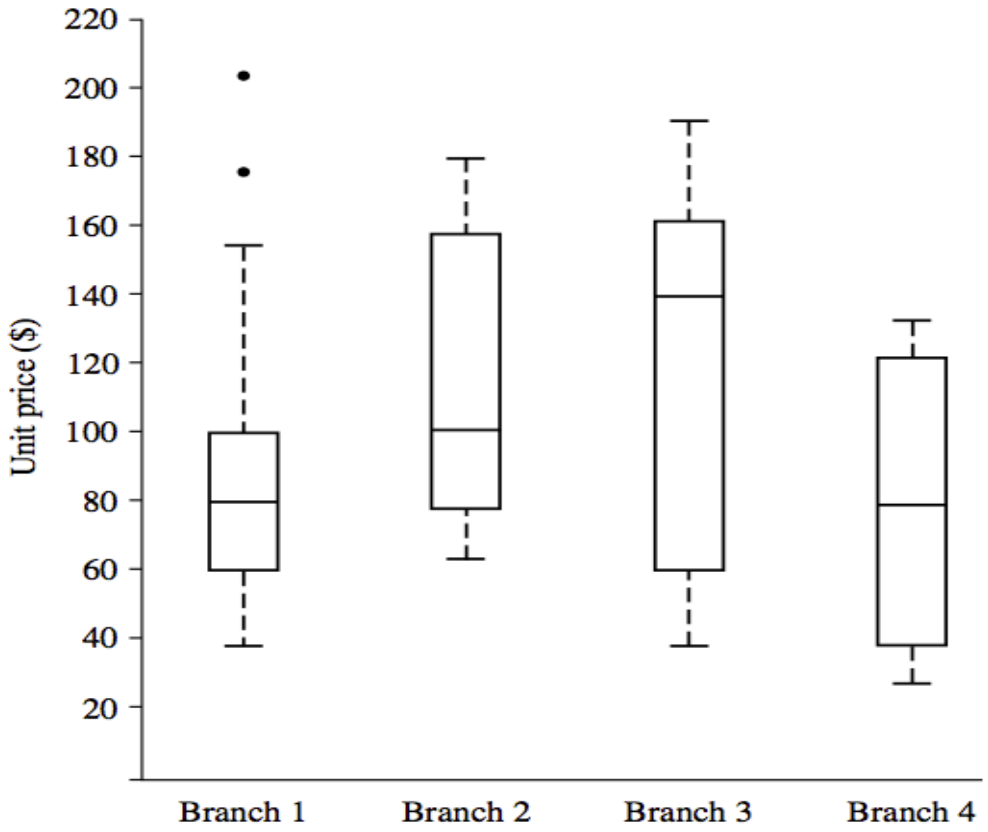
- El rango de un set es la diferencia entre el valor máximo y mínimo de un conjunto.
- Cuantiles son puntos tomando en intervalos regulares dentro de una distribución de datos, dividiendo el conjunto en intervalos iguales. Por ejemplo, 2-cuantiles es el punto que divide en dos desde la mediana hacia el lado izquierdo y otro hacia el lado derecho. 4-cunatiles son 3 puntos que dividen en 4 secciones desde la mediana hacia el lado izquierdo y otros hacia el lado derecho.
- El rango intercuantil es una medida simple de extensión que da el rango cubierto por la mitad media de los datos. Se define como $IQR=Q_3-Q_1$
- Una regla genérica para identificar posibles datos anómalos, es descartar valores dentro de un mínimo de 1.5xIQR por encima del tercer cuantil o debajo del primer cuantil.


25%
Q_1
25th
percentile
Q_2
Median
Q_3
75th
percentile

Midiendo la dispersión: resumen de cinco-números, diagrama de caja, y datos anómalos

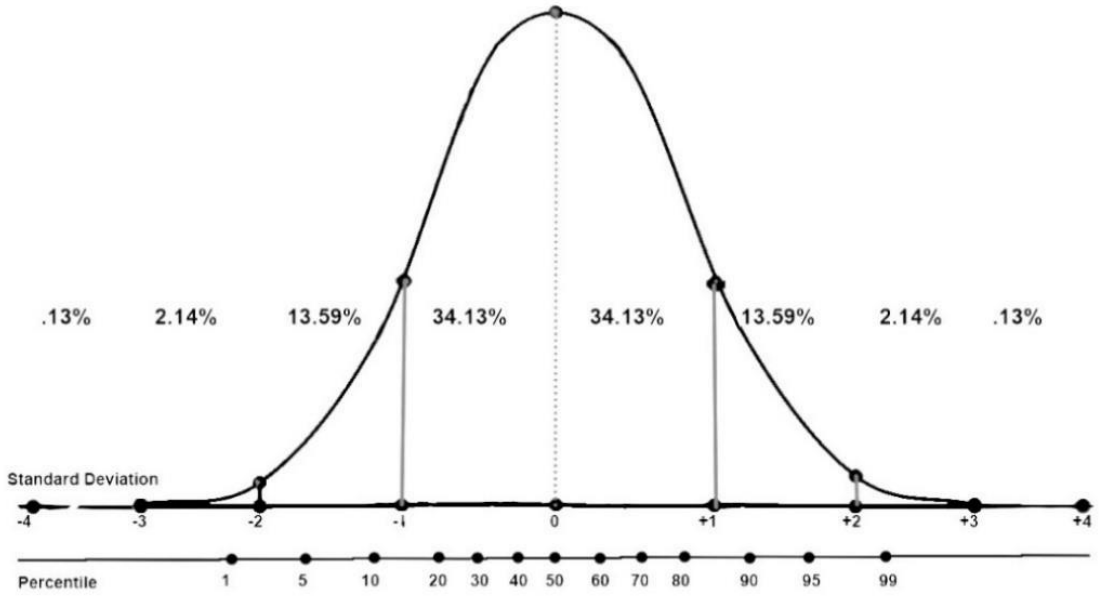
- El resumen de cinco-números de una distribución consiste en la mediana (Q_2), los cuantiles Q_1 y Q_3, y las observaciones más pequeñas y más grandes, ordenando en el orden de Mínimo, Q_1, Mediana, Q_3, y Máximo.
- Los diagramas de caja son una manera popular de visualizar una distribución.

- Un diagrama de caja incorpora un resumen de cinco-número:
  - Típicamente, los extremos de la caja están en los cuantiles tal que la longitud de la caja es el rango intercuantil.
  - La mediana se marca por una línea dentro de la caja.
  - Dos líneas (llamados bigotes) fuera de la caja se extienden hacia las observaciones más pequeñas (Mínimo) y más grandes (Máximo).



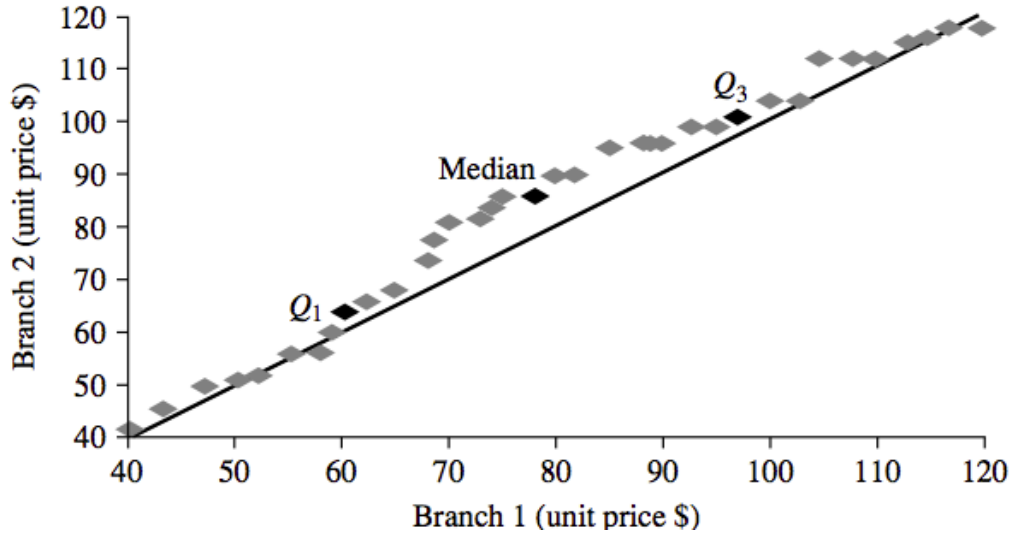
Midiendo la tendencia central: desviación estándar

- La desviación estándar son medidas de dispersión. Indican que tan esparcido está una distribución de datos.
- Una desviación estándar baja indica que las observaciones en los datos tienden a a estar muy cercanos al promedio, mientras que una desviación estándar alta indica que los datos están esparcidos a través del rango de valores.



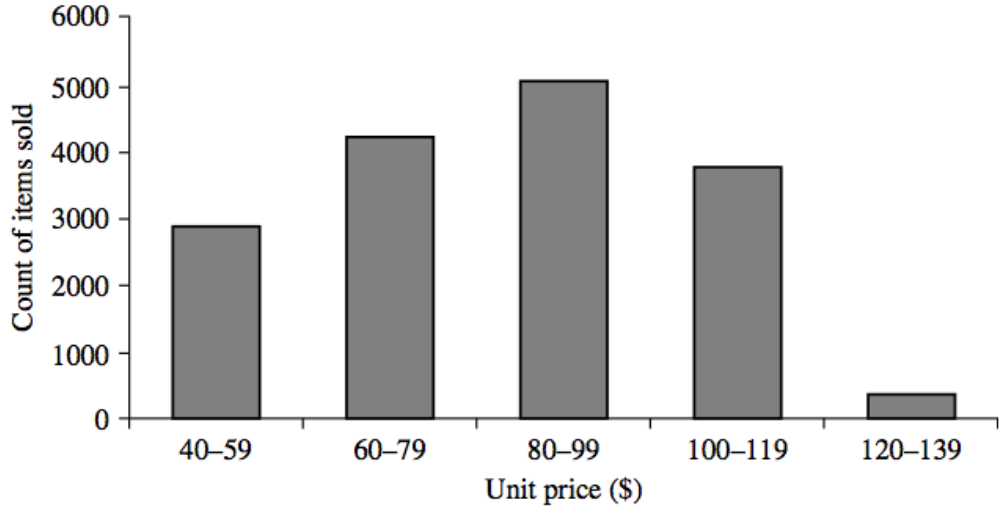
Métodos gráficos de descripciones estadísticas básicas de datos

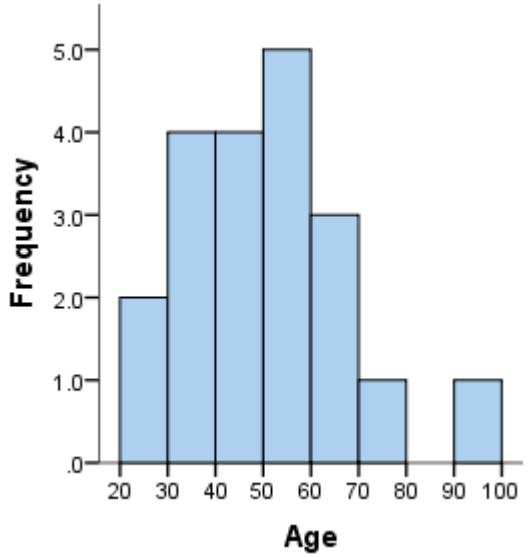
- Un gráfico cuantil (q plot) es una manera simple, y efectiva de visualizar una distribución univariable.
- Un gráfico cuantil-cuantil (q-q plot) grafica los cuantiles de una distribución univariable contra los cuantiles correspondientes de otra distribución univariable correspondiente.



Histogramas

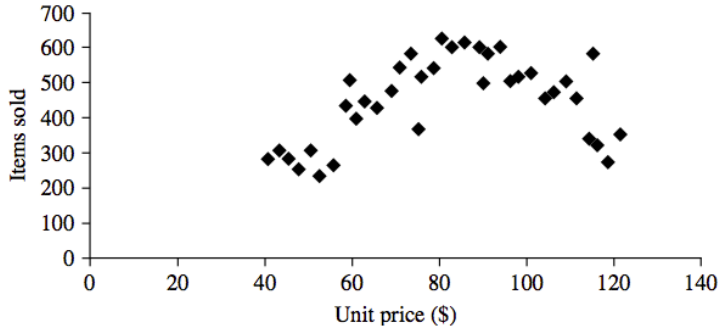
- Los histogramas son métodos gráficos para resumir la distribución de un atributo dado.

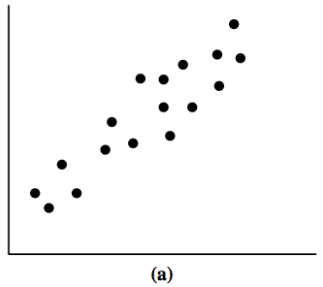


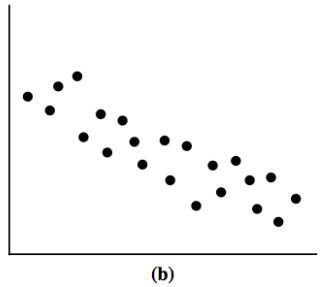


Diagramas de dispersión y correlación en datos

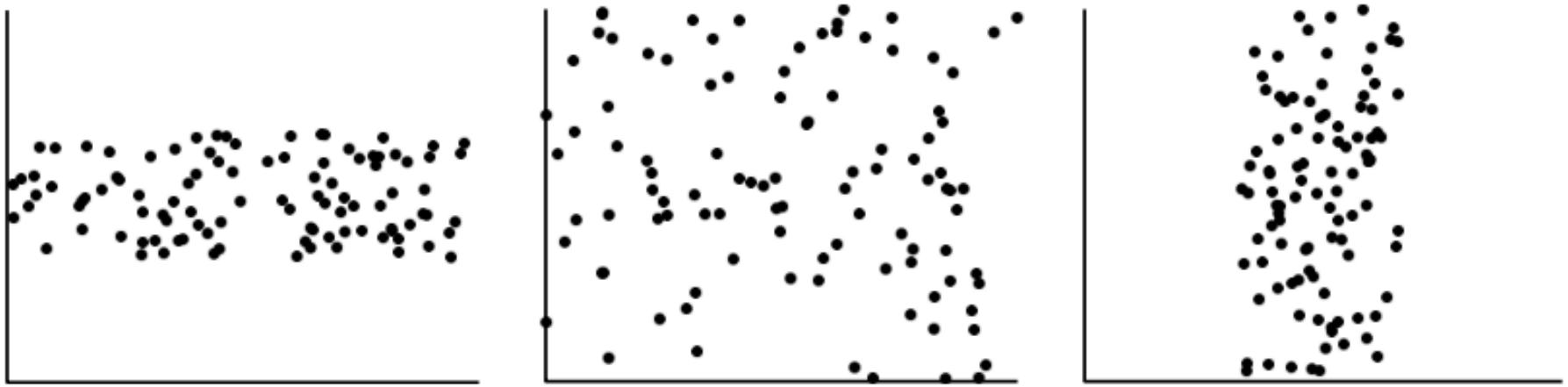
- Un diagrama de dispersión (scatter plot) es uno de los métodos gráficos más efectivos para determinar si hay alguna relación patrón, o tendencia entre dos atributos numéricos.
- Para construir un diagrama de dispersión, cada par de valores entre dos atributos es manejado como una coordenada.
- Sirve para identificar agrupaciones, datos anómalos, así como explorar la posibilidad de correlaciones.



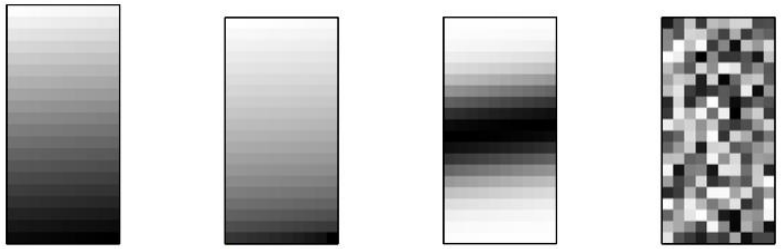

(a)


(b)

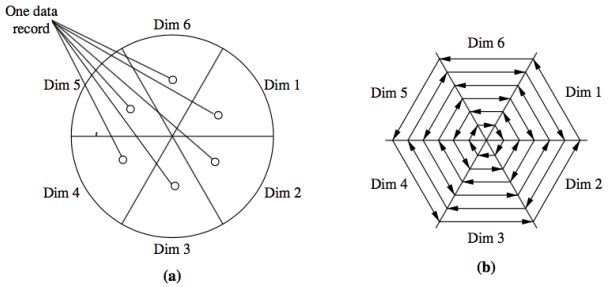
Casos donde no existe correlación entre dos atributos usando diagramas de dispersión

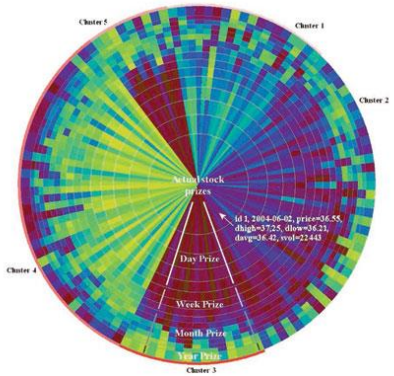


Visualización de datos basado en la técnica de visualización orientada a pixeles

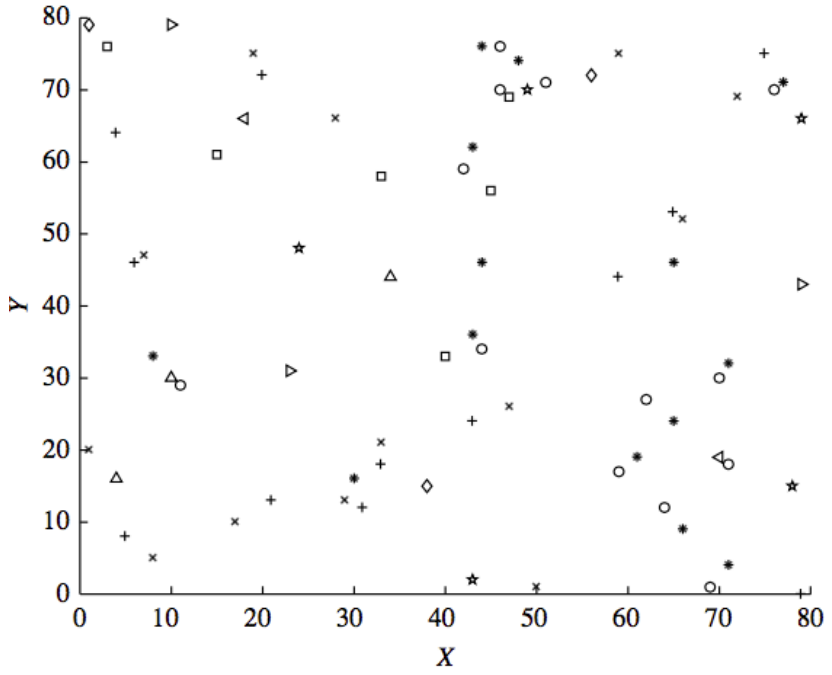


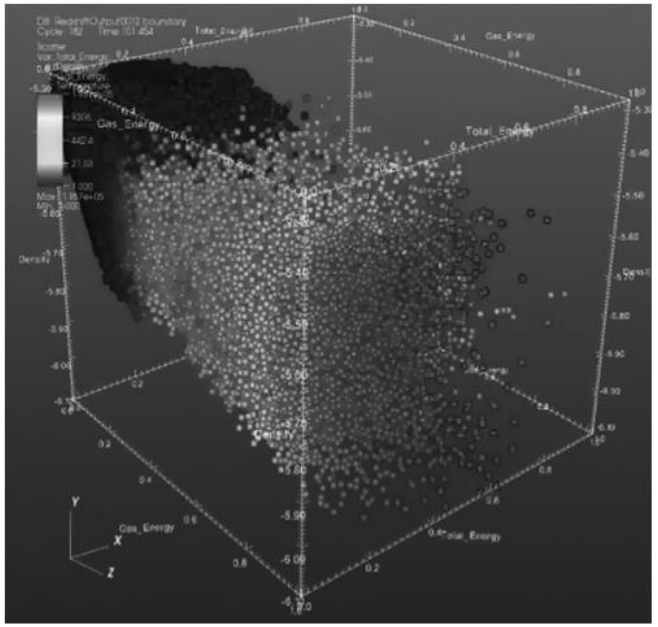
(a) *income* (b) *credit_limit* (c) *transaction_volume* (d) *age*



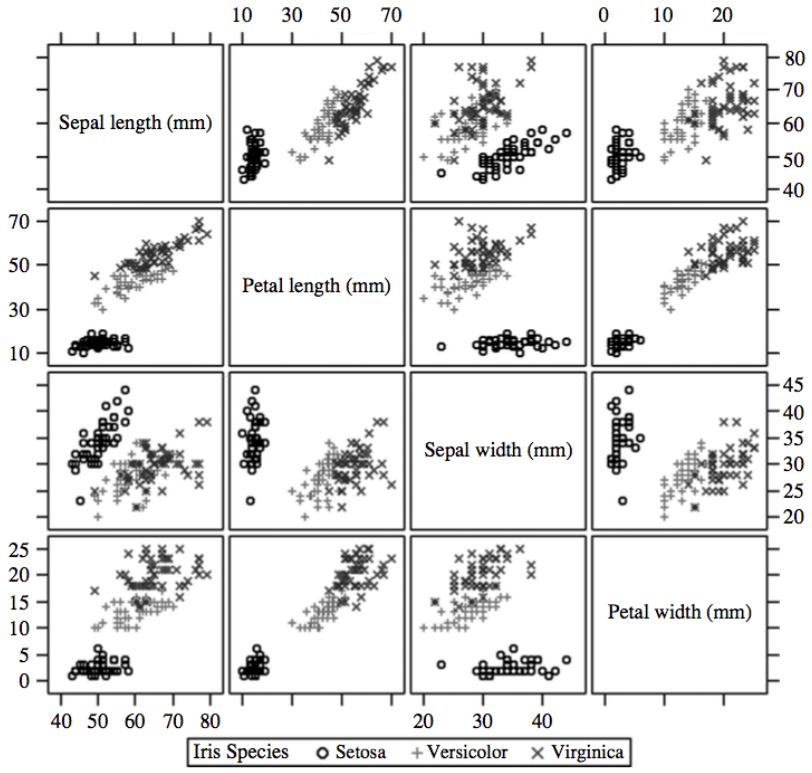


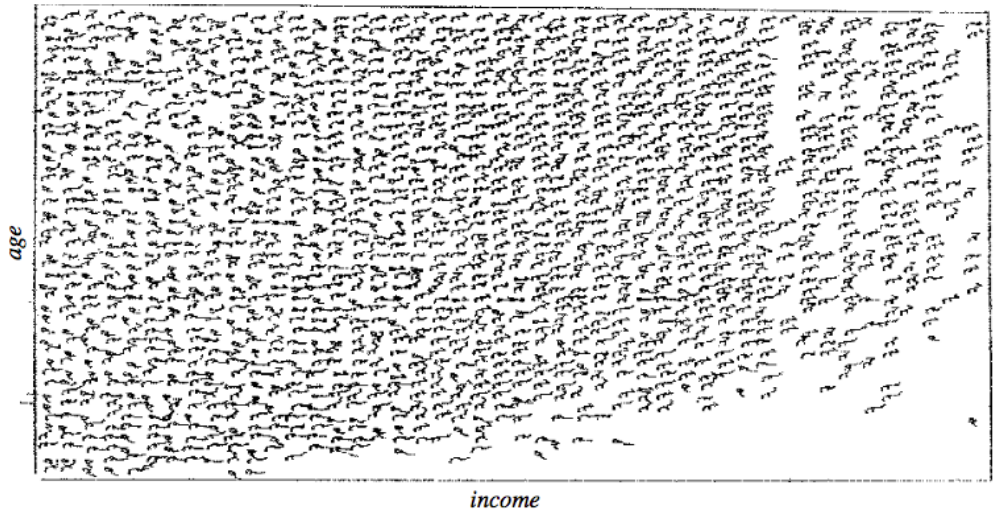
Visualización de datos basado en la técnica de proyección geométrica

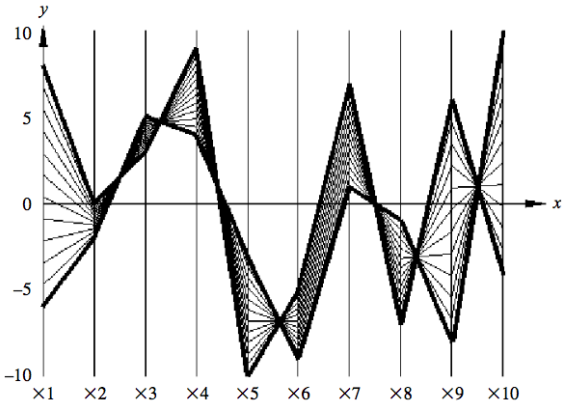




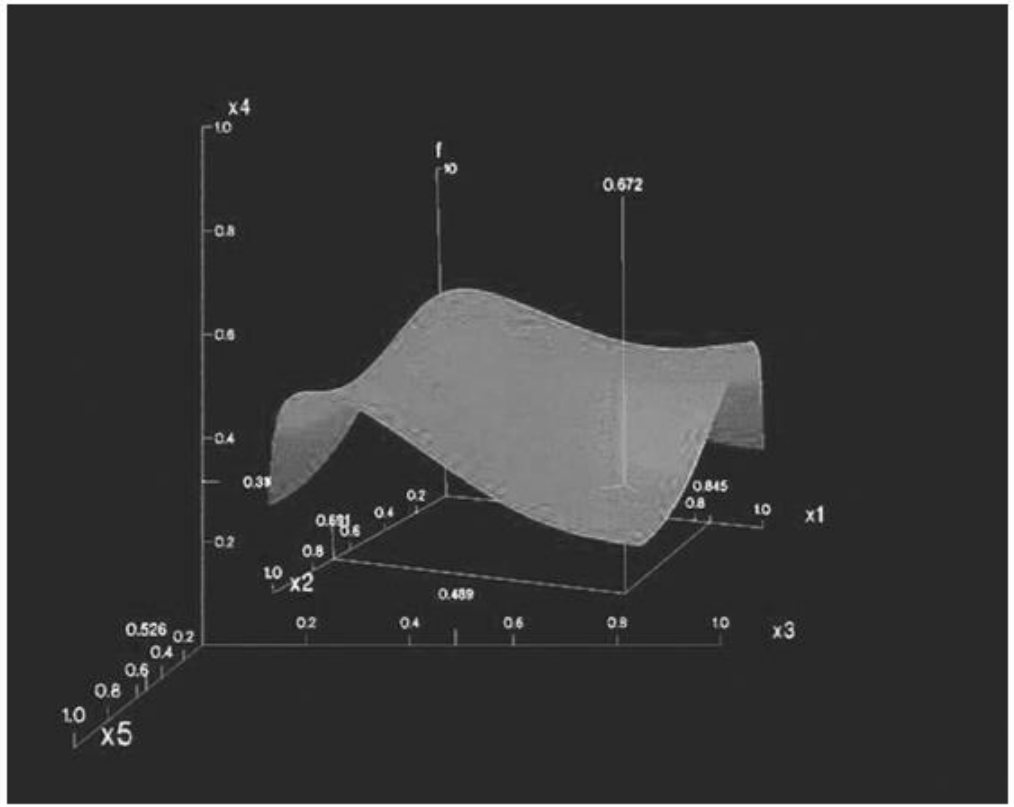
Visualización de datos basado en la técnica a base de iconos







Visualización de datos basado en la técnica de jerarquías



Visualizando datos y relaciones complejas

Actividad

- A partir de los siguientes datasets
  - https://www.kaggle.com/datasets/himanshunakrani/iris-dataset
  - https://www.kaggle.com/datasets/mdwaquarazam/shoe-dataset
  - https://www.kaggle.com/datasets/unsdsn/world-happiness
- Aplicar todo lo visto en esta presentación para describir y entender tus datos (usar *Pandas*)
  - Cómo estan distribuidos?
  - Faltan datos?
  - Hay tendencias?
  - Hay outliers?
  - Visualizarlos
  - Cualquier otra cosa que encuentren que se vea interesante
- Preparar una presentación (formato libre) invidivual para la siguiente clase