Unidad 4

Limpieza y transformación

Modelizado de Minería de Datos - Q22025

El Arte de Preparar los Datos: Limpieza, Selección y Transformación 🧹✨

La calidad de los datos es fundamental para el éxito de la minería de datos.

La transformación de los datos es a menudo necesaria para adaptarlos a las técnicas de minería de datos.

Aunque la limpieza e integración pueden realizarse durante la construcción de un almacén de datos, se tratan por separado debido a que la minería de datos no siempre requiere un almacén de datos.

El orden de las técnicas no es fijo, y depende del problema y las características de los datos.

Es fundamental conocer el dominio de los datos para realizar los procesos de **integración, limpieza, selección y transformación**.

Reconocimiento de Datos: La Primera Inspección 🕵️

Antes de limpiar o transformar, el reconocimiento es la fase exploratoria para comprender la calidad y estructura de los datos. Es el primer paso para detectar problemas como valores faltantes o anomalías y es la base de todo análisis posterior.

Métodos Estadísticos: Usar describe() para ver el resumen numérico (media, cuartiles) y info() para identificar tipos de datos y valores nulos.

Visualizaciones: Aplicar gráficos como histogramas para ver distribuciones, box plots para detectar valores atípicos, y scatter plots para visualizar relaciones.

Reconocimiento

`info()`: proporciona un resumen conciso de un DataFrame.

- Esencial para la exploración inicial de los datos y la detección de posibles problemas como valores faltantes o tipos de datos incorrectos.

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 88883 entries, 0 to 88882
Data columns (total 85 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Respondent            88883 non-null  int64
 1   MainBranch            88331 non-null  object
 2   Hobbyist              88883 non-null  object
 3   OpenSourcer           88883 non-null  object
 4   OpenSource            86842 non-null  object
 5   Employment            87181 non-null  object
 6   Country               88751 non-null  object
```

Reconocimiento

`describe()`: genera estadísticas descriptivas de las columnas numéricas del DataFrame.

- Permite identificar rápidamente la tendencia central, la dispersión y la forma de la distribución de los datos numéricos.
- Útil para detectar valores atípicos y comprender la variabilidad de los datos.

```
In [5]: frame = pd.read_csv('ratings.csv')
        frame.describe()
```

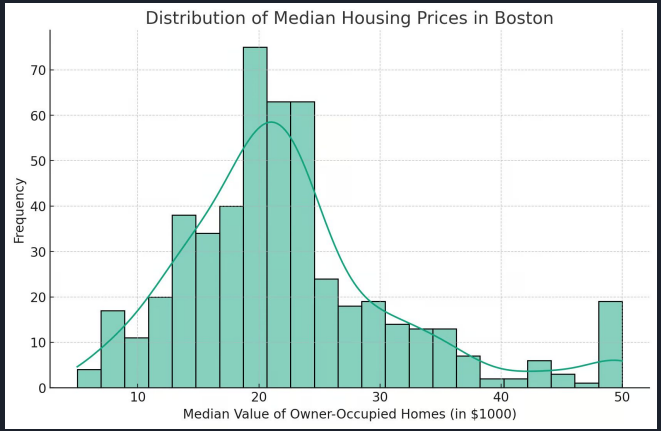
Out[5]:

	placeID	rating	food_rating	service_rating
count	1161.000000	1161.000000	1161.000000	1161.000000
mean	134192.041344	1.199828	1.215332	1.090439
std	1100.916275	0.773282	0.792294	0.790844
min	132560.000000	0.000000	0.000000	0.000000
25%	132856.000000	1.000000	1.000000	0.000000
50%	135030.000000	1.000000	1.000000	1.000000
75%	135059.000000	2.000000	2.000000	2.000000
max	135109.000000	2.000000	2.000000	2.000000

Reconocimiento

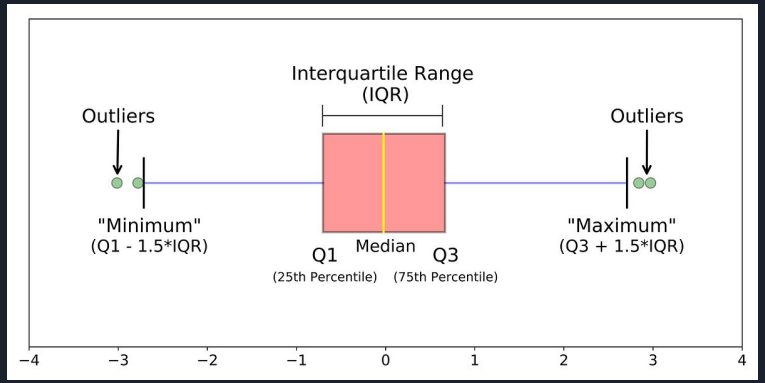
Los histogramas proporcionan una visión general de la distribución de una variable.

- Permiten visualizar la frecuencia de los valores en diferentes intervalos.
- Ayudan a identificar la forma de la distribución (simétrica, sesgada, etc.) y la presencia de valores atípicos.



Reconocimiento

Los box plots resumen los histogramas y son eficaces para la detección de outliers (valores atípicos)



Reconocimiento

Las gráficas de dispersión (scatterplots) son útiles para visualizar la relación entre dos atributos numéricos.

El Arte de Preparar los Datos: Limpieza, Selección y Transformación 🧹✨

- **Limpieza e Integración**: Unir y corregir datos de distintas fuentes para garantizar su precisión y consistencia. 🧹
- **Transformaciones**: Adaptar el formato de los datos (ej., de texto a números) para que los modelos puedan procesarlos.
- **Selección**: Elegir las variables y registros más relevantes para optimizar el análisis y la minería.

Limpieza e integración

El Gran Problema: Los datos suelen estar incompletos, sucios y dispersos en múltiples sistemas.

Integración: El proceso de unificar los datos de diferentes fuentes para crear un repositorio coherente.

Limpieza: La etapa de corregir errores y tratar los valores faltantes y atípicos.

Técnicas Clave: El uso de visualizaciones (histogramas, box plots) y el manejo estratégico de los valores nulos (ignorarlos, reemplazarlos o crear una señal).

Transformación de Atributos: El Arte de Adaptar los Datos 🛠️

Modificar la forma de los datos no solo permite una comprensión más sencilla sino que facilita el aprendizaje de los modelos de ML

- **Numerización:** (“variables dummies”). La numerización "1 a n" crea variables indicadoras (dummy) para cada valor nominal
- **Normalización:** (“scalers”) necesaria cuando se integran datos de diferentes fuentes con diferentes escalas, o para algoritmos basados en distancias.
- **Discretización**: convierte valores numéricos en nominales ordenados (intervalos). Ejemplo: calificaciones numéricas a {suspenso, aprobado, notable, sobresaliente}.
- El **Análisis de Componentes Principales (PCA)** es una técnica para reducir la dimensionalidad de conjuntos de datos con muchas variables.