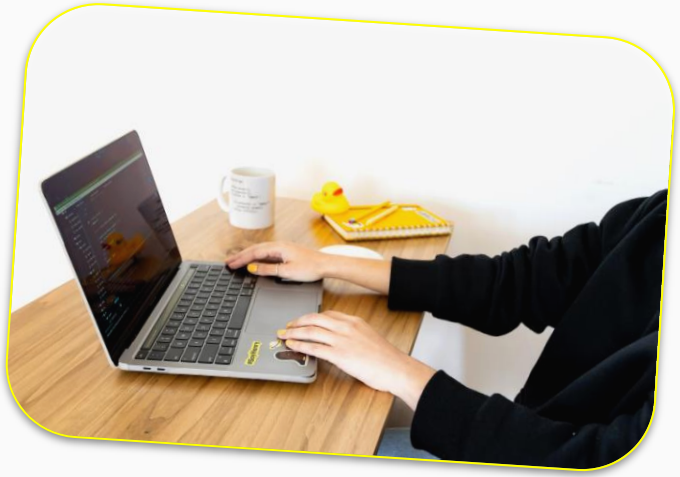
HENRY

ETL

Data Science

Agenda



- ETL
- Extracción
- Transformación
- Outliers
- Diagrama de caja
- Regla de las tres Sigmas

OBJETIVOS DE LA CLASE

Al finalizar esta lecture estarás en la capacidad de...

→ **conocer** y **entender** el concepto de ETL

Gracias a los procesos ETL es posible que cualquier organización:

- Mueva datos desde una o múltiples fuentes.
- Reformate esos datos y los limpie, cuando sea necesario.
- Cargue en otro lugar los datos como una base de datos unificada.
- Una vez alojados en un destino, esos datos se analizan.
- cuando ya están cargados en su ubicación definitiva se empleen en otro sistema operacional, para apoyar un proceso de negocio.

Fases de extracción

Fases de extracción

01 Extraer los datos desde los sistemas de origen.

02 Analizar los datos extraídos obteniendo un resultado

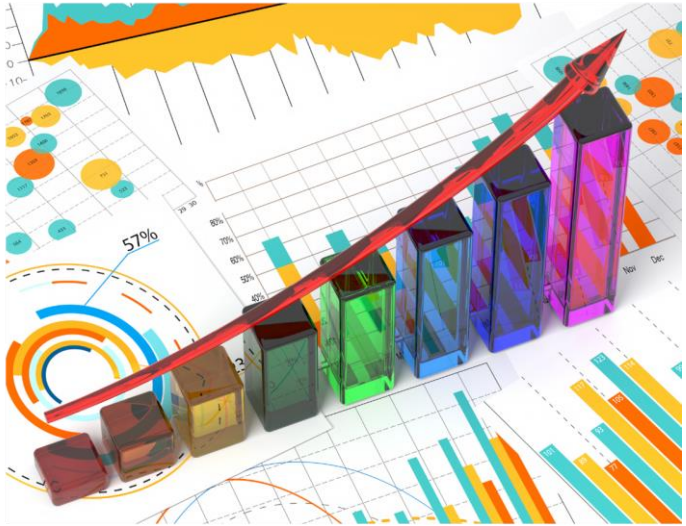
03 Interpretar el resultado para verificar que los datos extraídos cumplen con la pauta. Si no fuese así, los datos deberían ser rechazados.

04 Convertir los datos a un formato preparado para iniciar el proceso de transformación.

Fases de transformación

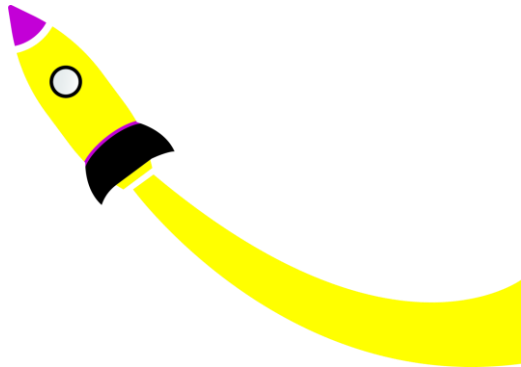
¿Qué es?

Un proceso ETL aplica una serie de reglas de negocio o funciones, sobre los datos extraídos para **convertirlos en datos** que serán cargados.



Las reglas deben de ser

- Independientes
- Claras
- Utilidad
- Declarativas
- Inteligibles



Outliers

¿Qué es?

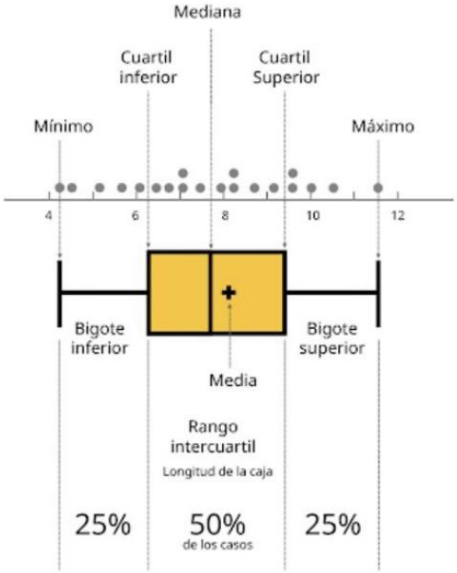
Un elemento fundamental a descubrir dentro de las tareas de identificación del ruido son los **valores atípicos o outliers**.

Los **outliers** son elementos que por su comportamiento se apartan notoriamente del comportamiento general. Esto se puede deber a un error en los datos o a un dato correcto que representa anomalías en la realidad.

Diagrama de caja

¿Para qué sirve?

Permite observar la distribución completa de los datos al mismo tiempo que su mediana y sus cuartiles. También, muestra los elementos que se escapan del universo, los outliers.



Rango intercuartílico o IQR:

* mínimo = Q1 – 1.5 x IQR

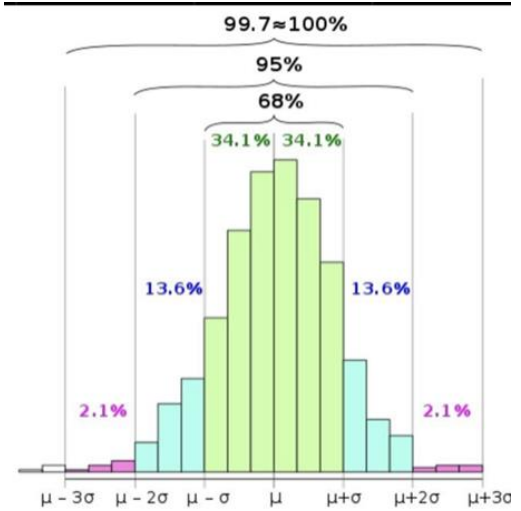
* máximo = Q3 + 1.5 x IQR

Reglas de las tres sigmas

¿Qué es?

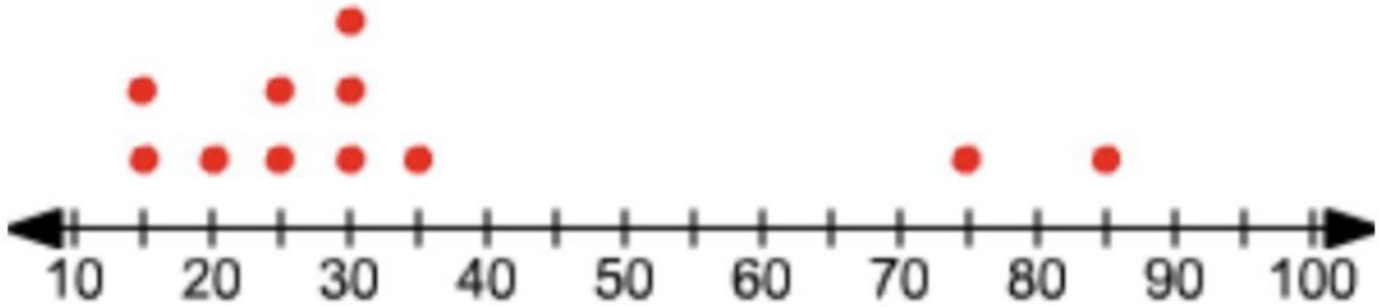
La Regla de las Tres Sigmas se basa en el valor promedio y la desviación estándar para obtener el rango, fuera del cual, podemos asumir que un valor es atípico.

* mínimo = Promedio – 3 * Desviación Estándar
* máximo = Promedio + 3 * Desviación Estándar



consideraciones

A veces, es la variable la que nos lo indica. Por ejemplo, la asistencia a un curso no puede ser menor que cero o mayor al número de alumnos que tiene el curso.



¿PREGUNTAS?

Resumen

¿Alguien dijo Homework?

HENRY

Próxima lecture

Query optimization


HENRY