DATA SCIENCE

Federico Baiocco
baioccofede@gmail.com
3512075440

Clase 9 - Agenda

ANÁLISIS EXPLORATORIO DE DATOS

- Para qué sirve
- Outliers
- Analicemos un dataset

¿ Dudas de la clase pasada ?

Análisis exploratiorio de los datos (EDA)

Un paso esencial en proyectos de data science para "entender" el dataset.

ICARO 2021

¿A qué nos referimos con "entender el dataset" ?

- Extraer variables "importantes" y descartar variables que no nos son útiles
- Identificar outliers
- Conocer cuantos datos faltantes tenemos (si es que tenemos)
- Identificar errores humanos
- Entender la relación (o ausencia de la misma) entre variables

Para todo esto es muy importante entender el problema que queremos resolver con datos

¿Qué preguntas me gustaría responder?
¿Qué preguntas podrán responder con ese dataset? (A veces no puedo responder todo lo que quiero con los datos que tengo)

“garbage in, garbage out”

¿A qué nos referimos con "entender el dataset" ?

- Extraer variables "importantes" y descartar variables que no nos son útiles
- Identificar outliers
- Conocer cuantos datos faltantes tenemos (si es que tenemos)
- Identificar errores humanos
- Entender la relación (o ausencia de la misma) entre variables

Para todo esto es muy importante entender el problema que queremos resolver con datos

Primero: Entender mis variables

- ¿ Qué tipos de datos tengo ?
    - cuantitativos (altura de una persona)
    - categóricos (género)
    - discretos (cantidad de personas)
- ¿ Cómo se distribuyen mis datos ?
- Medidas de resumen

Outliers

Los outliers son valores atípicos que difieren significativamente del resto de las observaciones.
¿ Por qué difieren ?

- Error de medición del instrumento.
- Error al introducir un dato.
- Estamos trabajando con muestras/poblaciones que no son tan homogéneas como creíamos.

Muchas veces los outliers son datos erroneos (por ejemplo un número negativo en la edad de una persona), pero esto no es siempre así.

A veces lo que queremos (nuestro problema a solucionar) es encontrar outliers!! -> Detección de fraude

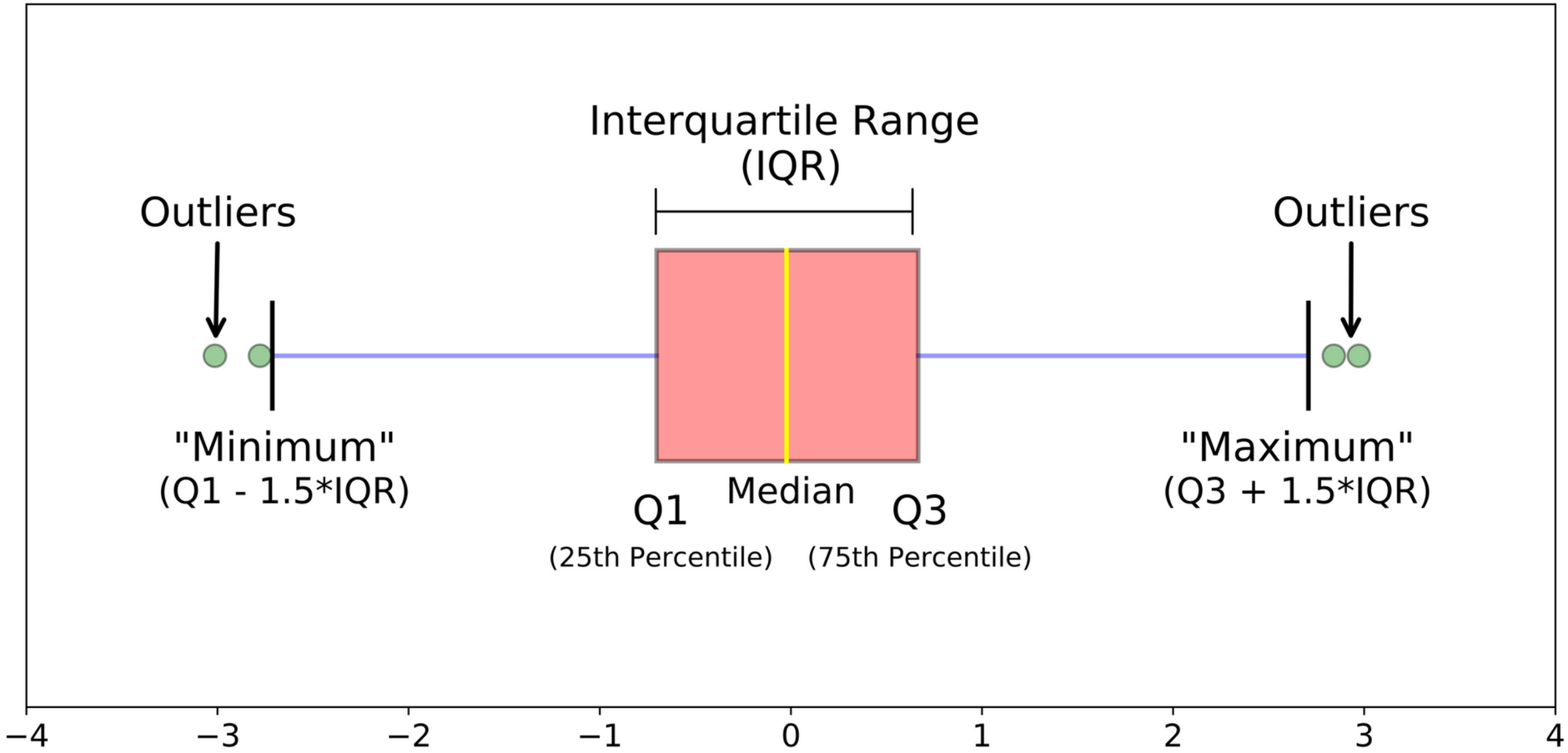
Outliers - ¿ Cómo los identifico?

Como muchas cosas en la ciencia de datos ... DEPENDE DEL PROBLEMA!!

Algunas técnicas que podemos utilizar para identificarlos son:

- Mediante visualización con un tipo de gráfico llamado boxplot
- Rango intercuartílico
- Regla de las 3 sigmas

Outliers - Boxplot



Outliers - IQR

Una forma de decidir que es un outlier y que no lo es, es definiendo valores mínimos y máximos. Pero.. ¿ Cómo definimos este mínimo y máximo ?

A veces, la variable misma nos lo indica. Por ejemplo, una edad no puede ser un número negativo.

En otros casos, una forma de definirlos es utilizando el rango intercuartílico:

mínimo = Q1 - 1.5 x IQR
máximo = Q3 + 1.5 x IQR

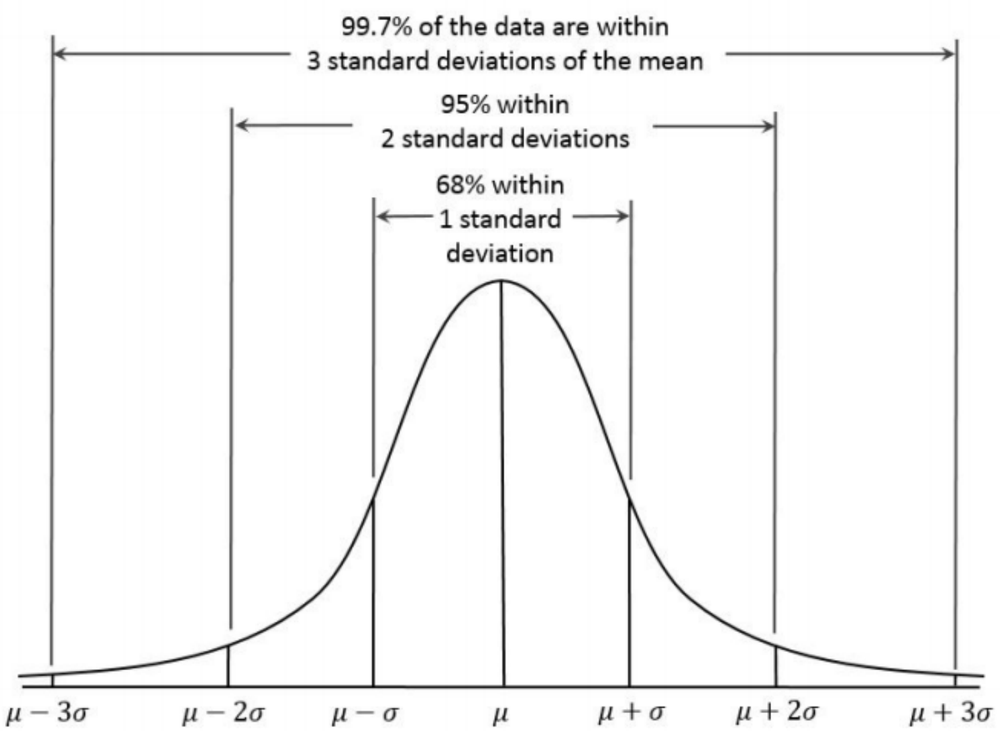
Donde IQR es = Q3 - Q1

Outliers - Regla de las 3 sigmas

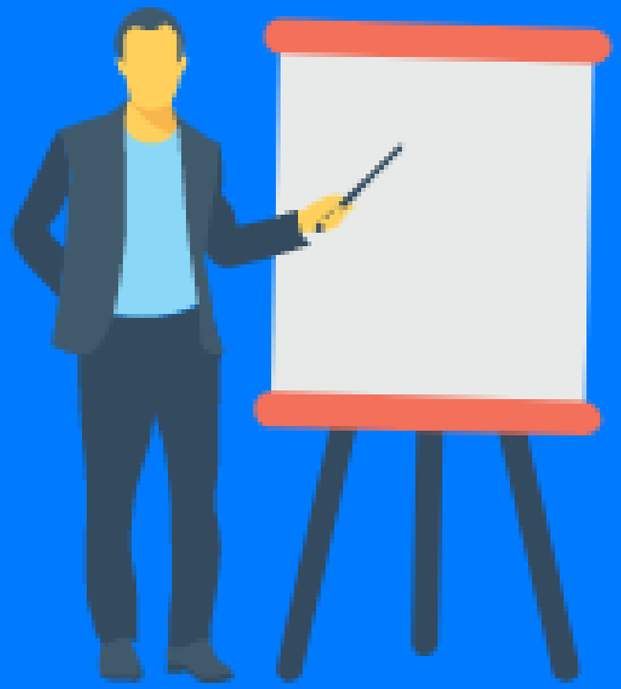
La regla de las 3 sigmas es otra forma de definir un mínimo y un máximo.

En este caso, en lugar de usar los cuartiles, usamos la desviación estandar.

- Mínimo = media - 3 SD
- Máximo = media + 3 SD



Abrimos notebook clase 9



Próxima clase

- Grupos de 2 o 3 personas
- 1 dataset distinto por grupo
- Análisis exploratorio y presentación:
    - Presentación de los datos
    - Explorar y buscar relaciones entre variables
    - Presentación para personas que no son data scientists
- Presentación: máximo 15 min

Preguntas?

Para la próxima clase:

- Terminar ejercicios.
- Seleccionar dataset con el que van a trabajar.

Vamos a comenzar la clase revisando dudas y después presentan los primeros grupos.