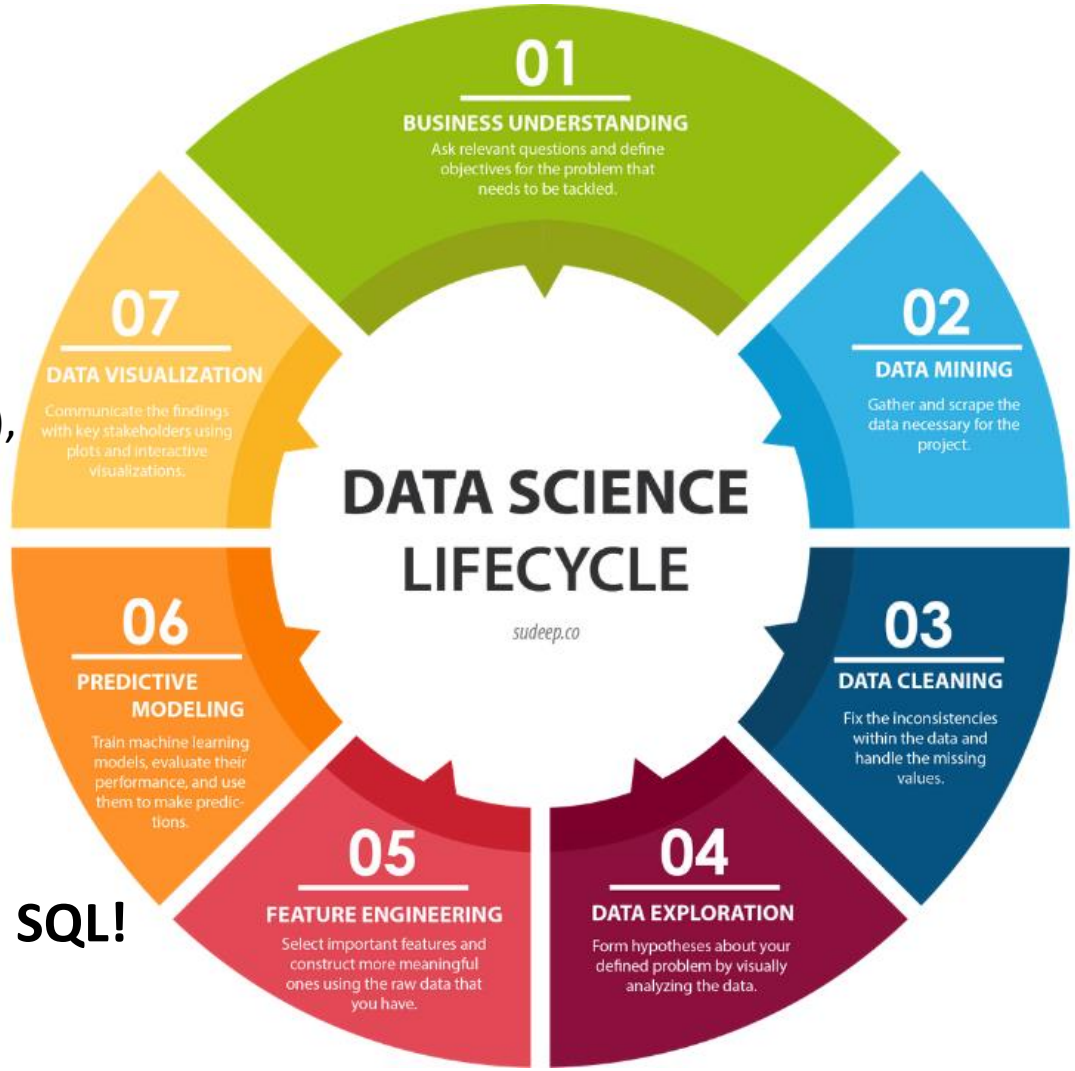
EDA (Exploratory Data Analysis) con SQL

DSML - Facultad de Telemática. Universidad de Colima – rherrera@ucol.mx

El Proceso del Data Science

Etapas del Data Science Process

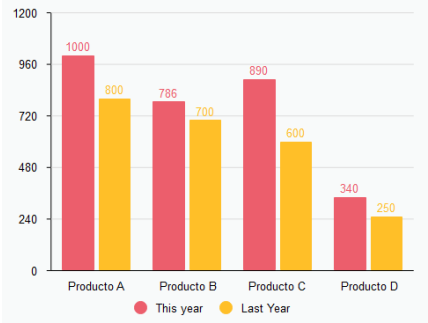
- Conocimiento del negocio/problema,
- Adquisición de datos,
- Exploración de datos (EDA),
- Preparación de datos, Limpieza de datos,
- Aplicación de ML
  (Analisis predictivo, Extracción de Conocimiento),
- Visualización de Datos (dataviz) y StoryTelling
- Toma de decisiones

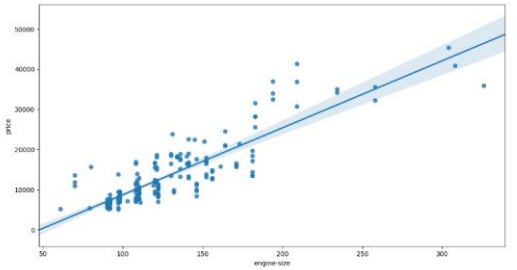


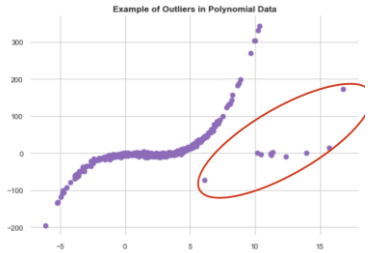
… en esta ocasión nos enfocaremos
a la etapa del EDA, pero … abordándolo con **SQL!**

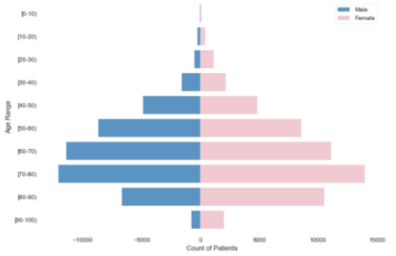
EDA con SQL para datos en formato de BD relacional

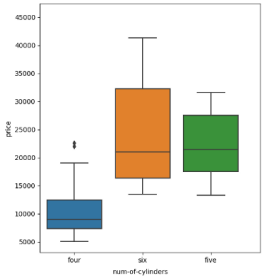
- ***Exploración del dataset***:
  estructura, tablas, campos, relaciones, y
  tipos de datos (numéricos, fechas, textos, categorías),
- ***Dimensiones y datos de variables***:
  valores mínimos, máximos, promedios, desv-standard,
  frecuencia, agrupamiento con respecto a otras variables,
  valores faltantes (nulos), valores fuera de rango (outliers),
- ***Representación visual con gráficos***:
  distribución, correlación, histogramas,
  comportamiento a lo largo del tiempo
  (patrones, tendencias, periodicidades)

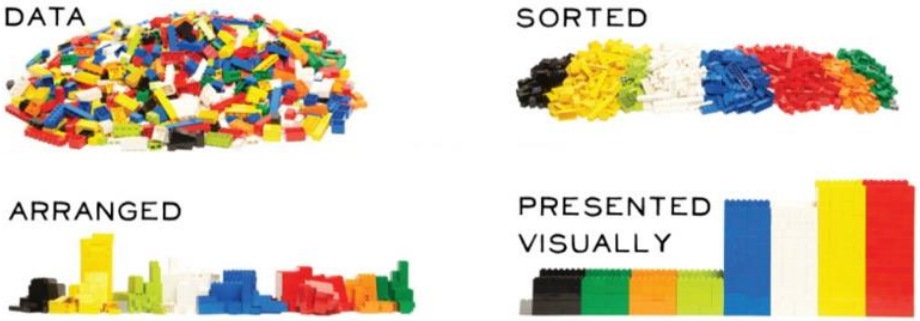








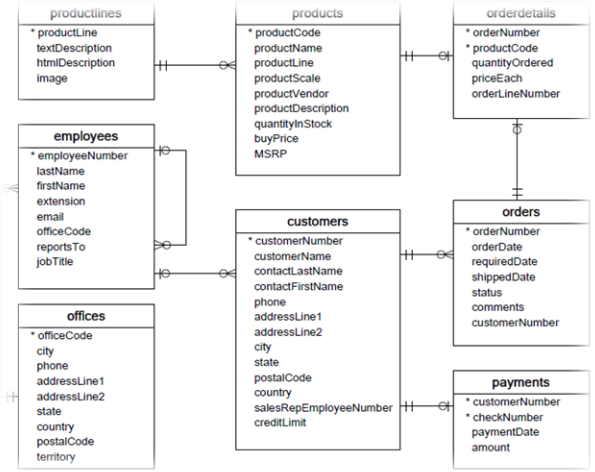




EDA con SQL para datos en formato de BD relacional

Uso de BD relacional y MySQL para EDA

- **Para hacer consultas con joins entre varias tablas** (para filtrados y agrupamientos)
- **Para generar Transformaciones de datos** (hacer categorías a partir de otros campos)
- **Para generar Vistas Integradas de datos** (simplificar consultas con el detalle requerido)
- **Para generar análisis de estadística descriptiva básica** (valores mínimo, máximo, promedio, desviación estándar, moda, mediana, Q1, Q2, Q3 y Q4, etc)



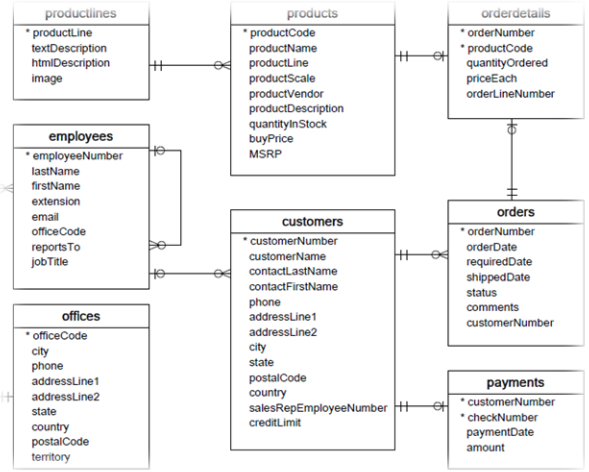
Caso práctico - BD Classicmodels

```
-- ***************************************************************************
-- Para hacer la TABLA de Estadistica Basica Descriptiva,
--   Tomamos el CreditLimit que es un campo numerico de relevancia:
-- ---------------------------------------------------------------------------
SELECT max(creditlimit) as Maximo,          -- $ 227,600.00
       min(creditlimit) as Minimo,          -- $       0.00
       avg(creditlimit) as Promedio,        -- $  67,659.02
       std(creditlimit) as DesvEstandar     -- $  44,858.39
       -- median(creditlimit) as Mediana       $       ?
       -- Q1 (25%)                            $       ?
       -- Q2 (50%)                            $       ?
       -- Q3 (75%)                            $       ?
       -- Q4 (100%) -> Max()                  $ 227,600.00
       -- Q0 --> Min()                        $       0
FROM customers;
```

EDA con SQL para datos en formato de BD relacional

Caso práctico - BD Classicmodels

```
-- SOLUCION:
-- Se Define una funcion UDF para obtener los Cuuartiles del CreditLimit
  SELECT max(creditlimit) as Maximo,         -- $ 227,600.00
         min(creditlimit) as Minimo,         -- $       0.00
         avg(creditlimit) as Promedio,       -- $  67,659.02
         std(creditlimit) as DesvEstandar,   -- $  44,858.39
         quartilCreditLimit(0.50) as Mediana,
         quartilCreditLimit(0.25) as Q1, -- (25%)
         quartilCreditLimit(0.50) as Q2, -- (50%)
         quartilCreditLimit(0.75) as Q3, -- (75%)
         quartilCreditLimit(1.00) as Q4  -- (100%) -> Max()
  FROM customers;
```

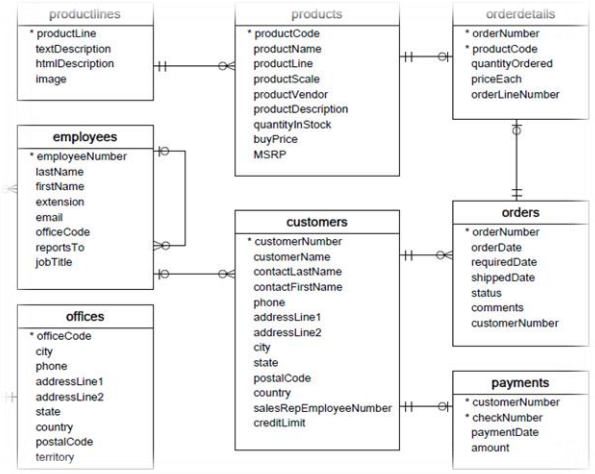


… habrá que implementar una función propia que contenga la lógica necesaria para obtener los diferentes valores de los cuartiles (incluso se puede generalizar para obtener cualquier decil) ..

```
-- Maqueta de la funcion UDF
DELIMITER $$
CREATE FUNCTION quartilCreditLimit(param_quartil double) RETURNS INTEGER
    -- param_quartil:  q1=(1/4)=0.25 / q2=(1/2)=0.5 / q3=(3/4)=0.75
  BEGIN
      DECLARE resultado_quartil, total_filas integer;
  -- ...  cuerpo de la funcion
  -- ...
      RETURN resultado_quartil;
  END$$
  DELIMITER ;
```

EDA con SQL para datos en formato de BD relacional

Caso práctico - BD Classicmodels



… y esto es solo el comienzo..
Manos a la obra...!!

```
-- ********************************************************************************
-- EDA básico.  Queries SQL genéricas para conocer la estructura, relaciones
--              integrar datos, agrupar y filtrar información,
--              transformar datos, encapsular consultas elaboradas en vistas
-- ********************************************************************************

-- Sobre los clientes:
    SELECT * FROM customers;
    SELECT max(customerNumber), min(customerNumber) FROM customers ORDER BY 1 DESC;

-- Cuántos clientes hay de cada país/continente?
    SELECT country, count(*) as cant FROM customers
    GROUP BY country ORDER BY 2 DESC ;

-- para considerar la región geográfica habria que sacar ese datos de Oficinas
    SELECT country, territory, count(*) as cant FROM offices
    GROUP BY country, territory ORDER BY 2 DESC ;

-- Cuál es el Límite de Credito máximo, mínimo y promedio, devStandard?
    SELECT max(creditLimit), min(creditLimit) FROM customers;

 -- Hay clientes sin límite de crédito?
    SELECT creditLimit, count(creditLimit) as cant FROM customers
    GROUP BY creditLimit ORDER BY 1 DESC ;    -- 93 filas
```

EDA con SQL para datos en formato de BD relacional

Referencias de información

Libros

- Herrera-Morales, J. R., Peralta-Domínguez, G., & Herrera-Espinoza, A. R. (2020). *El internet de las cosas y la ciencia de datos*. En A. R. Gallardo, J. R. Herrera-Morales, S. Sandoval Carrillo, & M. E. Cabello-Espinosa (Coords.), El internet de las cosas y su impacto en la educación (pp. 117-143). Editorial Universidad de Colima. ISBN 9786078549887.
- Cielen, D., Meysman, A. D. B., & Ali, M. (2016*). Introducing Data Science: Big data, machine learning, and more, using Python tools*. Manning Publications.
- McKinney, W. (2022). *Python for data analysis: Data wrangling with Pandas, NumPy, and Jupyter* (3ª ed.). O'Reilly Media.

Base de datos Classic Models

- MySQL Tutorial. (s.f.). *MySQL sample database* [BD]. https://www.mysqltutorial.org/getting-started-with-mysql/mysql-sample-database/