**DIAPOSITIVA 4(ARQUITECTURA DEL PROYECTO):**

El proyecto está basado en la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining)

Que es un modelo de implantación utilizado en proyectos de minería de datos y se divide en las fases de:

**DIAPOSITIVA 4(ETL):**

En la fase de Preparación de los datos, hemos trabajado con ETL (Extracción, Transformación y Limpieza)

Los datos presentados en esta diapositiva han sido obtenidos en el Portal de datos abiertos del Ayuntamiento de Madrid

Hemos importado al proyecto, un total de 107 ficheros CSV, siendo 53 ficheros de mediciones de contaminantes atmosféricos, 53 ficheros de meteorología y 1 fichero de Estaciones.

Cómo podéis ver, en el CSV original no viene las columnas del ICA. Para obtener el ICA hemos creado una función en el proyecto que pudiese calcular este índice.

Después de finalizada la transformación de los datos hemos generado un nuevo CSV para ser usado en la próxima fase del proyecto y este CSV contiene los mismos datos y estructura de columnas presentados en esta diapositiva.

**DIAPOSITIVA 4(TRATAMIENTO DE OUTLIERS):**

Antes de entrar en el tema de los outliers, comentar que, en este punto, ya teníamos tratado los valores nulos, él cuál iremos comentar más adelante con más detalles en la presentación.

Para verificar si había valores atípicos en nuestro conjunto de datos, hemos pintado esta gráfica de Caja y Bigotes, el famoso “BoxPlot”, donde hemos identificado valores muy fuera del rango y lo hemos eliminado usando la técnica de los “Cuartiles”.

Luego después de hacer el tratamiento de valores atípicos, hemos pintado nueva gráfica para confirmar el resultado. Y se quedó como presentado en esta imagen, finalizando de esta manera, el tratamiento de valores atípicos.

**DIAPOSITIVA 4(POWER BI - Portada):**

Para desarrollar las visualizaciones de las gráficas en Power BI, vuelvo comentaros que ha sido cargado el mismo fichero CSV generado en la fase final del proyecto en Python.

**DIAPOSITIVA 4(POWER BI - ESTACIONES):**

Aquí están presentadas todas las estaciones que hacen medición de los contaminantes atmosféricos, son las mismas que están contenidas en el CSV de Estaciones.

Como podéis ver, están todas, pero es importante enfatizar, que no todas hacen las mediciones de todos los contaminantes, pero, independientemente están incluidas dentro de los CSVs de mediciones.

Al verificar el conjunto de datos, luego después de la fase de carga y transformación de los datos, nos dimos cuenta de la gran cantidad de valores nulos que venía en el CSV y analizando la configuración de las estaciones, nos dimos cuenta, que eran falsos valores nulos, porque, de hecho, estas estaciones no están configuradas para hacer mediciones de ciertos contaminantes.

De este modo, hemos verificado que solamente la estación 8, hace la medición de todos los contaminantes y la hemos elegido para el proyecto.

Luego de elegir y seleccionar solamente los datos de medición de esta estación, hemos comprobado que todavía había valores nulos, los cuales, hemos rellenado usando el algoritmo de **IterativeImputer** donde Cada característica se imputa secuencialmente, una tras otra, lo que permite que los valores imputados anteriores se utilicen como parte de un modelo para predecir características posteriores.