



ETL SOLUTION FOR AIR QUALITY ANALYSIS IN INDIA

- Aguilar Ramírez Carlos Francisco
- Arista Romero Juan Ismael
- Vazquez Martin Marlene Gabriela

Introducción

La contaminación del aire es uno de los mayores problemas ambientales en India, impactando la salud pública y el medio ambiente.

El proyecto tiene como propósito principal analizar de manera integral la calidad del aire en India mediante un proceso ETL que permita transformar datos crudos en información útil y procesable.

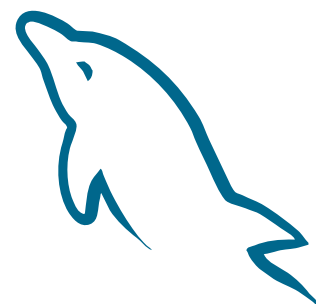
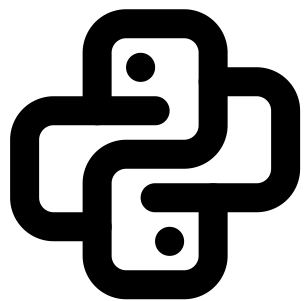


Entre los objetivos específicos del proyecto se incluyen la identificación de patrones de contaminación, la evaluación de regiones críticas y la creación de un dashboard interactivo que facilite la toma de decisiones basada en datos.

Proceso ETL

El proceso ETL desarrollado consta de tres etapas principales: **extracción, transformación y carga**. Estas etapas están diseñadas para garantizar que los datos sean procesados y presentados de manera efectiva, asegurando su integridad y utilidad analítica.

Herramientas utilizadas:



Proceso ETL

Extracción de Datos

Los datos fueron recopilados desde un conjunto de archivos CSV que contienen métricas sobre la calidad del aire en diversas ciudades y estados de la India.

Transformación de Datos

Se diseñaron reglas específicas para limpiar y estandarizar los datos:

- Los valores numéricos nulos, como los de PM2.5 y PM10, fueron rellenos utilizando la mediana o interpolación.
- Los valores categóricos nulos, como Status, se reemplazaron con "Desconocido".
- Se eliminó cualquier dato duplicado.

Carga de Datos


Los datos procesados fueron migrados a Databricks Community Edition, donde se registraron tablas temporales para realizar análisis SQL. Los reportes generados se exportaron como archivos CSV para su posterior visualización en Power BI.

Análisis de Datos

El análisis de datos realizado permitió extraer insights significativos sobre la calidad del aire en la India, utilizando herramientas SQL para procesar grandes volúmenes de información y generar reportes detallados.

Insights Clave Identificados:

- Temporadas más contaminadas.
 - El invierno mostró los niveles más altos de contaminación, seguido por el otoño.



Season	Avg_AQI
Winter	204.10201612903225
Autumn	173.431972276631
Spring	143.355415860735
Summer	116.06168198273461

Insights Clave Identificados

Ciudades con peor calidad del aire

Ahmedabad y Delhi presentaron los promedios de AQI más altos, destacándose como las ciudades más críticas.

Estados mas afectados

Delhi y Haryana registraron los niveles más altos de PM2.5 y PM10.

Los siguientes reportes fueron clave para este análisis:

- **Calidad del aire por ciudades y estados.**
- **Comparaciones por temporada y hora del día.**
- **Niveles de contaminantes específicos (PM2.5 y PM10) en regiones críticas.**

Dashboard en Power BI

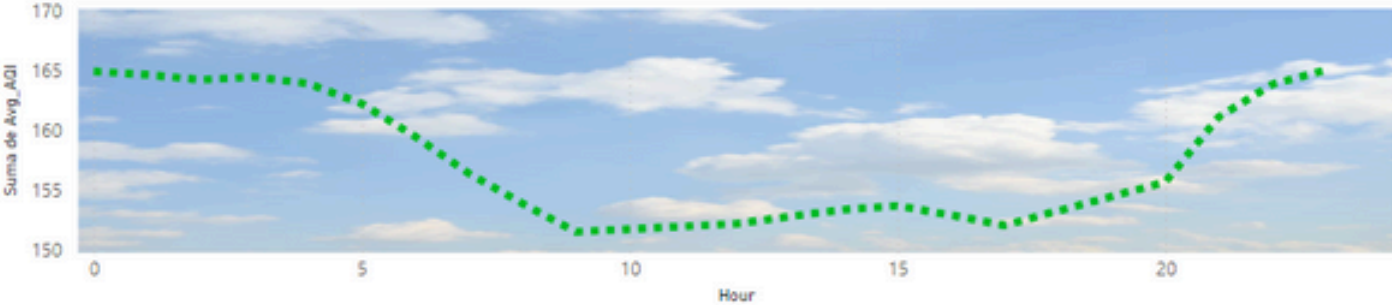
El dashboard creado en Power BI proporciona una visualización clara y detallada de los datos analizados, permitiendo explorar tendencias clave y realizar comparaciones interactivas.

Este dashboard integra múltiples visualizaciones diseñadas para facilitar la toma de decisiones y resaltar patrones significativos.

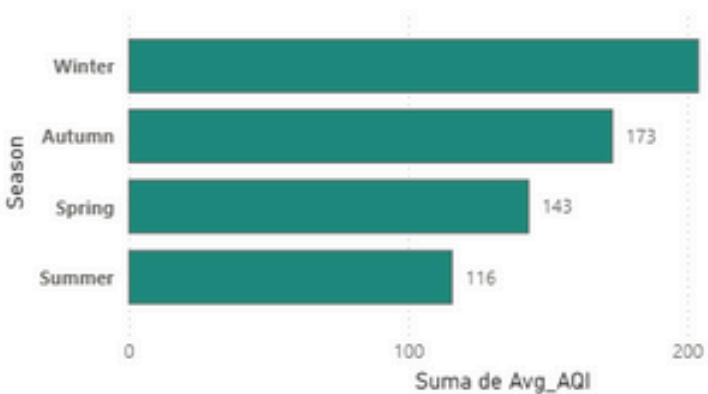
Calidad del Aire en India: Análisis Visual

State	Avg_PM25	Avg_PM10
Mizoram	17.69	23.99
Meghalaya	34.49	54.10
Kerala	28.77	54.85
Karnataka	36.75	85.47
Chandigarh	41.50	85.66
Tamil Nadu	49.37	88.37
Telangana	47.12	92.59
Andhra Pradesh	43.76	93.73
Uttar Pradesh	106.61	95.68
Maharashtra	43.35	96.08
Bihar	110.55	98.87
Gujarat	61.83	99.51
Punjab	54.75	114.20
West Bengal	63.30	114.28
Assam	63.66	116.60
Madhya Pradesh	50.01	118.59
Rajasthan	54.44	123.13
Odisha	59.94	135.47
Jharkhand	53.65	136.96
Haryana	110.82	137.91
Delhi	117.13	227.55

Suma de Avg_AQI por Hour



Suma de Avg_AQI por Season



City y Avg_AQI



Elementos de Dashboard

- **Tabla dinámica:**

Resumen interactivo que permite filtrar por estado, ciudad y nivel de contaminación.

- **Mapa interactivo:**

Muestra la distribución geográfica del promedio de AQI por ciudad, destacando las regiones más afectadas.

State	Avg_PM25	Avg_PM10
Mizoram	17.69	23.99
Meghalaya	34.49	54.10
Kerala	28.77	54.85
Karnataka	36.75	85.47
Chandigarh	41.50	85.66
Tamil Nadu	49.37	88.37
Telangana	47.12	92.59
Andhra Pradesh	43.76	93.73
Uttar Pradesh	106.61	95.68
Maharashtra	43.35	96.08
Bihar	110.55	98.87
Gujarat	61.83	99.51
Punjab	54.75	114.20
West Bengal	63.30	114.28
Assam	63.66	116.60
Madhya Pradesh	50.01	118.59
Rajasthan	54.44	123.13
Odisha	59.94	135.47
Jharkhand	53.65	136.96
Haryana	110.82	137.91
Delhi	117.13	227.55

Elementos de Dashboard

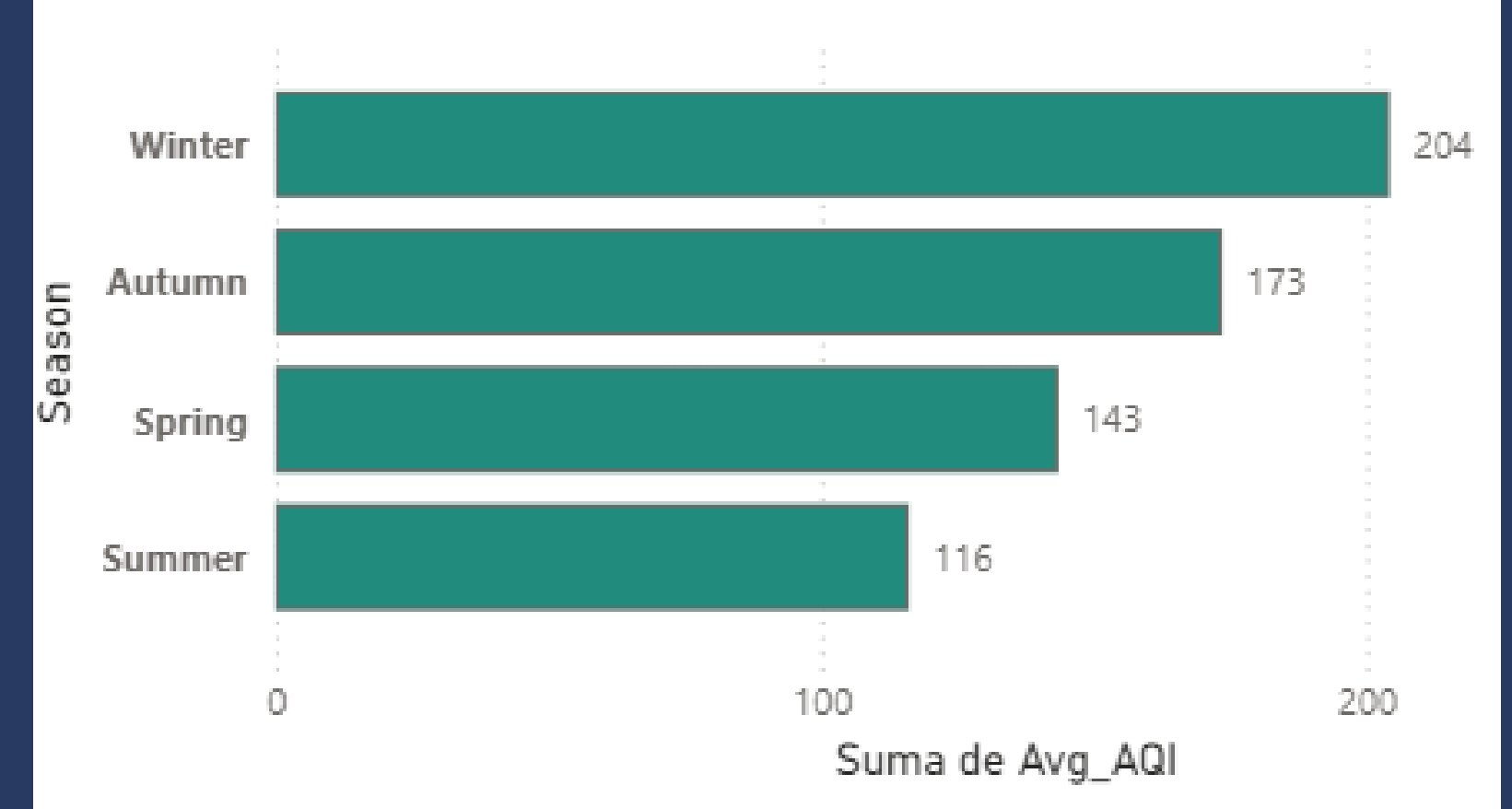
- **Gráfico de barras:**

Compara los promedios de AQI por temporada (invierno, primavera, verano, otoño) para identificar las épocas del año más críticas.

- **Gráfico de líneas:**

Representa la evolución del AQI promedio por hora del día, señalando las horas pico de contaminación.

Suma de Avg_AQI por Season



Insights Extraídos

- Las regiones más afectadas están claramente identificadas en el mapa.
- Las horas nocturnas presentan consistentemente niveles altos de contaminación, lo cual es crítico para la salud pública.
- El invierno es la temporada más problemática, confirmando la necesidad de medidas específicas en este periodo.



Conclusión



El proyecto ETL para analizar la calidad del aire en India logró cumplir con éxito los objetivos planteados. Mediante un flujo de datos estructurado y eficiente, se logró transformar información cruda en insights clave que facilitan la toma de decisiones informadas.

El proceso ETL integró herramientas modernas (Python, Databricks, Power BI) que garantizaron la limpieza, transformación y visualización eficiente de los datos.

El dashboard interactivo proporciona una herramienta poderosa para explorar los datos y comprender los patrones de contaminación de manera intuitiva.

Gracias por su atencion

