



PROYECTO FINAL

Línea 1: Desarrollo de soluciones – Ágil

Plataforma de Análisis Interactivo y Modelo de Predicción de Emisiones de Carbono en Taxis de NYC: Información Accesible para Decisiones Sostenibles y Basadas en Datos

Integrantes:

Hospital Roman, Oscar David (0000-0002-5298-6268)
Jofré, Leandro Gastón (0009-0002-3075-7883)
Oria García, Pedro Santos (0009-0004-8777-6991)
Alarcón Bothia, Ilbert Ferney (0009-0009-3556-3421)
Cáceres, Juliana (0009-0004-0482-109X)

ASESOR

Lila Alves
Jonathan Deiloff

2023-02

TABLA DE CONTENIDO

TABLA DE CONTENIDO.....	2
INDICE DE FIGURAS	4
INDICE DE TABLAS.....	9
CAPITULO 1: DEFINICIÓN DEL PROYECTO	11
1.1 DEFINICIÓN DEL PROBLEMA	11
1.1.1 OBJETO DE ESTUDIO	11
1.1.2 DOMINIO DEL PROBLEMA	11
1.1.3 LIENZO LEAN CANVAS.....	12
1.2 DEFINICIÓN DE LA SOLUCIÓN.....	13
1.2.1 OBJETIVOS DEL PROYECTO.....	13
1.2.2 INDICADORES DE ÉXITO.....	14
1.2.3 ALCANCE	16
1.2.4 RESTRICCIÓN FUNCIONAL DEL ALCANCE	16
CAPITULO 3: DESARROLLO DE LA SOLUCIÓN	17
3.1 INCEPCIÓN ÁGIL.....	17
3.1.1. VISIÓN DEL PRODUCTO	17
3.1.2 DEFINICIÓN DE ROLES	17
3.1.3 HISTORIAS DE USUARIO.....	19
3.1.4 TABLERO SCRUM BASE.....	23
3.1.5 DIAGRAMA DE GANTT.	24
3.1.6 CRONOLOGÍA DEL PROYECTO.....	25
3.2 ANÁLISIS PRELIMINAR DE CALIDAD DE DATOS.....	25
3.3 ARQUITECTURA DE SOFTWARE	26

3.3.1 ARQUITECTURA TÉCNICA BASE.....	26
3.4 DESARROLLO DE SPRINT 1.....	28
3.4.1 SPRINT PLANNING 1.....	28
3.4.2 CONSTRUCCIÓN DEL SPRINT 1.....	30
3.4.3 DAILY MEETING	33
3.4.4 SPRINT RETROSPECTIVE.....	35
3.5 DESARROLLO DE SPRINT 2.....	35
3.5.1 SPRINT PLANNING 2.....	35
3.5.2 CONSTRUCCIÓN DEL SPRINT 2.....	38
3.5.3 DAILY MEETING	67
3.5.4 SPRINT RETROSPECTIVE.....	70
 CAPITULO 7: CONCLUSIONES DEL PROYECTO	94
 7.1 CONCLUSIONES.....	94
7.2 RECOMENDACIONES	94
 GLOSARIO	96
 SIGLARIO.....	97
 REFERENCIAS BIBLIOGRÁFICAS	99
 ANEXOS	99
 I. ACTAS	ERROR! BOOKMARK NOT DEFINED.
II. GESTIÓN DEL PROYECTO	100

INDICE DE FIGURAS

Figura 1. Lienzo Canvas	12
Figura 2. Visión del producto.....	17
Figura 3. Historias de usuario	19
Figura 4. Desglose del product Backlog	21
Figura 5. Historia de usuario por épicas	22
Figura 6. Tablero scrum base.....	23
Figura 7. Diagrama de gannt.....	24
Figura 8. Cronología del proyecto.....	25
Figura 9. Arquitectura técnica Base	27
Figura 10. Historia de usuario HU001-Epica01-Sprint01	29
Figura 11. Mockup de la HU001	31
.....	31
Figura 12. Product Backlog Refinado - Sprint01	32
Figura 13. Burdown Chart -Sprint01	33
Figura 14. Sprint Retrospective 01	35
Figura 15. Historia de usuario HU002 -Epica01-Sprint02.....	36
.....	36

Figura 16. Historia de usuario HU003 -Epica02-Sprint02.....	36
.....	37
Figura 17. Mockup de la HU002.....	38
Figura 18. Mockup de la HU003.....	38
Figura 19. Product backlog -Sprint02.....	40
Figura 20. Incremento del Sprint-HU002	41
.....	42
Figura 21. Interfaz Mongo Compas.....	46
Figura 22. Servicio de Atlas en MongoDB	47
Figura 23. Estructura del Data Warehouse	48
Figura 24. Creación de Usuarios	48
Figura 25. Dataset ETL_AirQ.csv	53
Figura 26. Dataset ETL_C_Combust.csv	53
Figura 27. Dataset Taxis Rutas	54
Figura 28. Dataset Taxis Zonas.....	54
Figura 29. Dataset Sonora.....	55
Figura 30. Workflow Detallado	56
Figura 31. Gráfico de viajes por año.	57

Figura 32. Gráfico heatmap de las correlaciones de los borough de New York.....	57
.....	58
Figura 33. Gráfico de dispersión de la tabla de Combustión y CO2.....	58
Figura 34. Gráfico de barras sobre los borough de New York.....	59
.....	60
Figura 35. Gráfico de barras de las tarifas diarias por distrito.....	60
.....	61
Figura 36. Gráfico de Líneas de la contaminación en el aire.....	61
Figura 37. Gráfica de barras sobre los tipos de taxi en New York.....	62
Figura 38. Grafica de barras sobre las emisiones de CO2	62
Figura 39. Grafica de distribución por día de taxi por distrito.....	63
Figura 40. Grafica de barras.....	64
Figura 41. Gráfico de los viajes por Borough.....	64
Figura 42. Gráfico de barras sobre la cantidad de viajes por año.....	65
Figura 43. Gráfico de línea sobre el promedio de emisiones de CO2 por año.....	66
Figura 44. Gráfico de línea sobre el promedio de tarifa por distrito al año.....	66
Figura 45. Sprint Burdonwn Chart -Sprint 02.....	69
Figura 46. Sprint Retrospective 02	70

Figura 47.	Historia de usuario HU004 -Épica03-Sprint03.....	71
Figura 48.		71
Figura 49.	Historia de usuario HU005 -Épica04-Sprint03.....	72
Figura 50.	Historia de usuario HU006-Épica04-Sprint03.....	72
Figura 51.	Historia de usuario HU007-Épica05-Sprint03.....	72
Figura 52.	Mockup de la HU004.....	73
Figura 53.	Mockup de la HU005.....	74
Figura 54.	Mockup de la HU006.....	75
Figura 55.	Mockup de la HU007.....	75
Figura 56.	Product backlog -Sprint03.....	76
Figura 57.	Model elección	78
Figura 58.	ML producción	81
Figura 59.	Estructura del dashboard	81
Figura 60.	xxxxx	82
Figura 61.	Xxx	82
Figura 62.	Despliegue en la nube AWS.....	83
Figura 63.	Creación de un dominio en aws	84
Figura 64.	Dataset seleccionado para el despliegue en bucket S3.....	85

Figura 65.	Bucket S3 AIMPEC_2023	85
Figura 66.	Creación del modelo Sagemaker	86
Figura 67.	Generando el target en el modelo	86
Figura 68.	Configuración de las columnas en forecasting.....	87
Figura 69.	Valores de la predicción.....	89
Figura 70.	Intervalo de predicción del conjunto de datos actual: 2023-12-04 al 2024-01-03.....	89
Figura 71.	Visualización en aws Quicksight.....	90
Figura 72.	Sprint Burdonwn Chart -Sprint 03.....	92
Figura 73.	Sprint Retrospective 03	93

INDICE DE TABLAS

Tabla 1. Indicadores de éxito	14
Tabla 2. Análisis de calidad de datos.....	26
Tabla 3. Resumen sprint planning 1	30
Tabla 4. Sprint 01 – Revisión del progreso.	33
Tabla 5. Resumen sprint planning 2	37
Tabla 6. Diseño del modelo conceptual de datos	42
Tabla 7. Detalle de los Pipelines	44
Tabla 8. Tabla detalle de las Reglas de Negocio	49
Tabla 9. Tabla detalle de la Validación de scripts de la regla del Negocio.....	51
Tabla 10. Sprint 02 – Revisión del progreso.	67
Tabla 11. Glosario de términos usados en el documento.....	96
Tabla 12. Siglario de las siglas usadas en el documento.....	97

RESUMEN

El proyecto "NYC Taxis & Carbon Emission" es una iniciativa de análisis de datos con enfoque en el sector de transporte de taxis de la ciudad de Nueva York y su relación con la contaminación del aire y la sonora. La idea central es entender los patrones de uso de los taxis, las emisiones que generan y cómo estos factores influyen en la calidad del aire y la contaminación sonora de la ciudad.

La empresa de servicios de transporte de pasajeros contratante está interesada en expandir su negocio hacia el sector de taxis y busca entender el panorama actual para tomar decisiones informadas. Su visión de futuro contempla la posibilidad de introducir vehículos eléctricos en su flota, lo que podría disminuir la contaminación y mejorar la calidad del aire.

Para lograr los objetivos del proyecto, se utilizarán diversas fuentes de datos, incluyendo datos de viajes de taxis, condiciones climáticas, contaminación sonora, contaminación del aire, emisiones de CO₂ y otros factores relevantes. La estrategia de análisis implica la recopilación y limpieza de estos datos, el análisis estadístico para identificar tendencias y patrones, y la implementación de modelos de aprendizaje automático para realizar predicciones y proporcionar recomendaciones basadas en los hallazgos.

Finalmente, las métricas y los indicadores clave de rendimiento (KPI) se utilizarán para evaluar la eficacia del proyecto y guiar futuras decisiones. El objetivo final es proporcionar un conjunto de datos y análisis robustos que la empresa de transporte pueda utilizar para mejorar sus servicios y contribuir a un futuro menos contaminado.

CAPITULO 1: DEFINICIÓN DEL PROYECTO

1.1 Definición del problema

El problema central es la falta de información sobre las emisiones de los taxis y su impacto ambiental, lo que impide a la empresa de transporte tomar decisiones estratégicas para reducir la contaminación y mejorar la calidad del aire en la ciudad de Nueva York. Esta falta de información dificulta la implementación de políticas y medidas efectivas para abordar el problema de la contaminación del aire causada por los taxis.

Además, sin información detallada sobre las emisiones de los taxis, es difícil establecer metas realistas y medibles para reducir la contaminación y monitorear el progreso en el tiempo. La empresa podría implementar medidas generales, como reemplazar gradualmente su flota por vehículos eléctricos, pero sin una comprensión precisa de las emisiones actuales y las áreas donde se concentran los niveles más altos de contaminación, es difícil optimizar los esfuerzos y asignar recursos de manera eficiente.

1.1.1 *Objeto de estudio*

El objetivo principal es analizar los datos relacionados con los viajes de taxis, las condiciones climáticas, la contaminación sonora, la contaminación del aire, las emisiones de CO₂ y otros factores relevantes. Se busca recopilar y limpiar estos datos para realizar un análisis estadístico y aplicar modelos de aprendizaje automático con el fin de identificar tendencias, patrones y realizar predicciones.

1.1.2 *Dominio del Problema*

Si bien la contaminación del aire en general es un problema importante en las áreas urbanas, el enfoque principal de este proyecto se centra en las emisiones de CO₂ debido a su papel significativo en el calentamiento global y el cambio climático. El CO₂ es uno de los principales gases de efecto invernadero que contribuyen al atrapamiento del calor en la atmósfera, lo que conduce al calentamiento global y sus consecuencias asociadas.

En este proyecto en particular, nos enfocaremos específicamente en la contaminación de dióxido de carbono (CO₂) en la ciudad de Nueva York. Si bien el dominio del problema abarca la contaminación del aire y la contaminación sonora en el transporte de taxis.

1.1.3 Lienzo Lean Canvas

Se trabajó con un lienzo Lean Canvas en el proyecto porque nos proporciona una estructura clara y concisa para definir la propuesta de valor, identificar oportunidades y desafíos, y mantener un enfoque ágil y centrado en el cliente. Ayuda a comunicar de manera efectiva la visión del proyecto y a tomar decisiones informadas para lograr el éxito.

Figura 1.

Lienzo Canvas



Nota. El presente diagrama representa la visión final del proyecto

1.2 Definición de la solución

El proyecto consiste en desarrollar una solución integral que incluye un dashboard interactivo y un modelo de predicción. El objetivo principal es brindar al cliente la capacidad de obtener ratios y métricas relevantes para la toma de decisiones relacionadas con el sector de taxis en la ciudad de Nueva York.

La solución permitirá visualizar de manera intuitiva y en tiempo real diversos indicadores clave, como la duración de viajes, porcentaje de tarifas, viajes inter e intra boroughs, días y semanas con mayor cantidad de viajes, distancia promedio por pasajero, propinas promedio por pasajero, tipos de pago más utilizados, entre otros.

Además, se desarrollará un modelo de predicción basado en técnicas de machine learning, que utilizará datos históricos y variables relevantes para estimar las emisiones de carbono generadas por los taxis. Esto permitirá al cliente tener una visión clara del impacto ambiental de su flota de vehículos y evaluar la viabilidad de implementar vehículos eléctricos o tomar medidas para reducir las emisiones.

La solución se basará en una plataforma en la nube para garantizar el acceso remoto y la escalabilidad, y se utilizarán tecnologías como Hugging Face para el desarrollo del modelo de predicción y Power BI para la creación del dashboard interactivo.

Con esta solución, el cliente podrá obtener información en tiempo real, identificar patrones y tendencias, y tomar decisiones informadas y sostenibles para mejorar la eficiencia y calidad del servicio de taxis, al tiempo que se consideran aspectos ambientales y de sostenibilidad.

1.2.1 Objetivos del proyecto

Objetivo General

Es de optimizar los servicios de taxis en NYC, identificando patrones de demanda y oportunidades de mejora. Se desarrollará un modelo de pronóstico con una precisión del 95% para estimar las emisiones de carbono, impulsando decisiones más sostenibles y eficientes.

Objetivos específicos

Por cada sprint se cumplirá lo siguiente:

- **OE-01:** Definir el alcance del proyecto y establecer la infraestructura necesaria, utilizando metodologías ágiles. Colaborar en la organización y distribución de tareas, asignando roles específicos y creando el repositorio de código, fuentes de datos y documentación.
- **OE-02:** Implementar pipelines de extracción, transformación y carga (ETL) de datos hacia estructuras de almacenamiento, como Data Warehouse, Data Lake o Data Lakehouse. Considerar la carga incremental de datos y utilizar herramientas de big data y/o servicios en la nube para optimizar el proceso.
- **OE-03:** Desarrollar un dashboard interactivo y un modelo de Machine Learning utilizando los datos procesados. Incluir los KPIs relevantes para el análisis realizado y preparar un storytelling efectivo para presentar los resultados. Elaborar reportes y dashboards, definir y calcular los KPIs, y poner en producción el modelo de Machine Learning.

1.2.2 Indicadores de éxito

El cumplimiento de los objetivos del proyecto se mide a través de los indicadores de logro detallados en la siguiente tabla:

Tabla 1.

Indicadores de éxito.

	Indicador de éxito	Objetivo
I-01	<p>Reducción de la emisión de contaminantes.</p> <p>Objetivo: Reducir la emisión de contaminantes en un 5% en los próximos 3 años tomados semestralmente.</p> <p>Formula:</p> $\%PR = ((Pis - Pfs) / Pis) * 100$ <p><i>Donde:</i></p> <p><i>PR: Reducción de la emisión de contaminantes.</i></p> <p><i>Pis: Polución al inicio del primer semestre(2022-I,2021-I,2020-I).</i></p> <p><i>Pfs: Polución al final del segundo semestre(2022-II,2021-II-I,2020-II).</i></p>	OE-01

Medir la eficiencia del consumo de combustibles.

Objetivo: Aumentar la eficiencia del consumo de combustibles en un 10% durante los próximos 3 años tomados semestralmente.

Entonces: el indicador de eficiencia de consumo de combustible podría definirse de la siguiente manera:

Ecc=Eficiencia de consumo de combustible (MIL/L) = 1 / "Fuel Consumption(City (L/100 MIL))"

Ecc: Cantidad de millas que un vehículo puede recorrer por cada litro de combustible que consume en ciudad.

I-02

OE-02

Formula:

$$\%Ef = ((Ef_f - Ef_i) / Ef_i) * 100$$

Donde:

Ef: Eficiencia de consumo de combustible.

Ef_f: Eficiencia al final del segundo semestre (2022-II, 2021-II, 2020-II).

Ef_i: Eficiencia al inicio del primer semestre(2022-I, 2021-I, 2020-I).

Mejorar las ventas promedio de los viajes.

Objetivo: Aumentar las ventas promedio de los viajes en un 20% durante los próximos 1 año, medido semestralmente.

Definición del indicador: Ventas Promedio de los Viajes (VPV)

VPV = "Total recaudado por día" / "Viajes por día"

VPV: representa la cantidad de dinero promedio que se gana por cada viaje realizado.

Formula:

I-03

OE-03

$$\%Ip = ((Ip_f - Ip_i) / Ip_i) * 100$$

Donde:

Ip= Ventas Promedio de los Viajes.

I_f= Ventas Promedio de los Viajes al final del semestre (2022-II, 2021-II).

I_i = Ventas Promedio de los Viajes al inicio del semestre (por ejemplo, 2022-I, 2021-I).

Aumento de los ingresos

Objetivo: Aumentar el farebox_per_day en un 10% en 3 meses.

Formula.

$$\%RI = ((If - Ii) / Ii) * 100$$

I-04

Donde:

OE-03

RI: Aumento de los ingresos.

Ii: Ingresos al inicio (farebox_per_day al inicio del período).

If: Ingresos al final (farebox_per_day al final del período).

1.2.3 Alcance

El proyecto de taxis incluye la recolección y procesamiento de datos, el diseño e implementación de un Data Warehouse, el análisis exploratorio de datos, el desarrollo de modelos de Machine Learning, la creación de un dashboard interactivo y la implementación del modelo de Machine Learning en producción. Estas tareas se llevarán a cabo siguiendo una metodología ágil, con roles y responsabilidades asignados, reuniones periódicas de seguimiento (Daily) y herramientas colaborativas para facilitar la comunicación y el intercambio de información. El proyecto se desarrollará dentro de un marco de tiempo definido de 3 semanas, con hitos y entregables establecidos en la documentación de HENRY, y se gestionarán los riesgos de manera proactiva para maximizar el valor entregado a los Stakeholders.

1.2.4 Restricción Funcional del alcance

Las restricciones de este proyecto incluyen un plazo de tiempo específico de 3 semanas para su finalización, el desarrollo de este proyecto se limita a la ciudad de New York no se puede ampliar su aplicación a otras ciudades, este proyecto no contempla factores topográficos o de tránsito. Los recursos de cómputo son limitados conforme al stack tecnológico, así como las consideraciones de los stakeholders identificados inicialmente, adicional el proyecto cumple con las leyes y regulaciones, y la posibilidad de ajustar el alcance ante cambios significativos en el negocio. Estas restricciones se establecen para garantizar una gestión efectiva y exitosa del proyecto, optimizando los

recursos disponibles y asegurando el cumplimiento de los objetivos dentro de los límites establecidos.

CAPITULO 3: DESARROLLO DE LA SOLUCIÓN

3.1 Incepción Ágil

3.1.1. Visión del Producto

En este punto, se obtiene la estrategia general del negocio con el fin de establecer las necesidades, los usuarios y los beneficios del producto al negocio.

Figura 2.

Visión del producto



Nota. El presente diagrama es la estrategia general del negocio

3.1.2 Definición de Roles

1. Data Engineers:

- Leandro Gastón Jofré: Responsable de la recopilación, limpieza y transformación de los datos necesarios para el proyecto "NYC Taxis & Carbon Emission". Trabaja en la implementación y mantenimiento de los sistemas de almacenamiento y procesamiento de datos.
- Oria García Pedro Santos: Encargado de desarrollar y mantener las infraestructuras de datos necesarias para el proyecto. Colabora en la implementación de flujos de datos eficientes y en la integración de diversas fuentes de datos.

2. Data Scientist:

- Ilbert Ferney Alarcón Bothia: Especialista en análisis de datos y modelos de aprendizaje automático. Se encarga de aplicar técnicas estadísticas y algoritmos de machine learning para descubrir patrones, identificar tendencias y realizar predicciones relacionadas con los patrones de uso de los taxis y su impacto en la contaminación del aire y la contaminación sonora.

3. Data Analyst:

- Juliana Caceres: Responsable de analizar los datos recopilados y realizar análisis exploratorios para extraer información relevante y generar insights que ayuden en la toma de decisiones. Trabaja en la identificación de métricas clave y en la generación de informes y visualizaciones para comunicar los hallazgos de manera efectiva.

4. Project Manager:

- Hospital Roman Oscar David: Encargado de liderar y coordinar el proyecto "NYC Taxis & Carbon Emission". Gestiona el alcance, los recursos y los plazos del proyecto, además de facilitar la comunicación entre los diferentes roles. Supervisa la ejecución del proyecto y asegura que se cumplan los objetivos establecidos.

Cada uno de estos roles desempeña funciones específicas y contribuye de manera integral al éxito del proyecto, desde la recopilación y procesamiento de datos, pasando por el análisis y modelado, hasta la gestión general del proyecto. Trabajando en equipo, estos profesionales colaboran para generar un conjunto de datos y análisis robustos que permitan mejorar los servicios de transporte y contribuir a un futuro menos contaminado.

3.1.3 Historias de Usuario

Como descripción de las historias de usuario, en donde se especifica el rol, la necesidad funcional y el valor o beneficio para el negocio.

Figura 3. Historias de usuario

HISTORIAS DE USUARIO				
Identificador de la Historia	Nombre de la Historia	Como (rol)	Necesito (característica)	Descripción de la Historia de Usuario
HU001	Limpieza de datos	Data Engineer	Deseo limpiar y preprocesar los datos para garantizar la calidad y fiabilidad de los mismos.	Mejora la calidad de los datos y la confiabilidad de los resultados para la ingestión en mi Base de datos.
HU002	Integración de datos	Data Engineer	Deseo integrar datos de diferentes fuentes para tener una visión completa del problema.	Proporciona una visión integral y mejora la precisión de los análisis y mi relación final en mi Base de Datos.
HU003	Análisis exploratorio de datos	Data Analyst	Deseo analizar los datos para identificar tendencias y patrones.	Revela insights valiosos y guía la toma de decisiones.
HU004	Modelo de predicción	Data Scientist	Deseo desarrollar un modelo de predicción para estimar las emisiones de carbono.	Ayuda a la empresa a planificar estrategias sostenibles y cumplir con las regulaciones medioambientales.
HU005	Desarrollo de Dashboard	Data Analyst	Deseo desarrollar un dashboard interactivo para visualizar los indicadores.	Facilita la interpretación de los datos y la comunicación de los resultados.
HU006	Validación de Dashboard	Data Analyst	Deseo validar y ajustar el dashboard para asegurarme de que cumple con las necesidades del usuario.	Asegura que la solución es útil y relevante para el cliente.
HU007	Despliegue en la nube	Arquitecto Cloud	Necesito desplegar la solución en la nube para que esté accesible.	Permite el acceso remoto y garantiza la escalabilidad de la solución.

Product Backlog.

El producto backlog del proyecto, enumera todas las características, funcionalidades, requisitos, mejoras y correcciones que constituyen cambios a realizarse sobre el producto para entregas futuras. Los elementos de la “Lista de Producto” tienen como atributos la descripción, el orden, la estimación y el valor. Es así como, según las técnicas y metodologías explicadas anteriormente, es que se ha llegado a refinar nuestra lista de productos hasta conseguir una proporcionalidad según la prioridad requerida. A continuación, se detalla el refinamiento del Product Backlog asociado.

Figura 4. Desglose del product Backlog



Figura 5.

Historia de usuario por épicas

HISTORIAS DE USUARIO										
Nº	Épica	Identificador de la Historia	Nombre de la Historia	D descripción de la Historia de Usu Como (rol)	Estado	Dimensión /Esfuerzo	Iteración (Sprint)	Prioridad	Criterio de Aceptación	Comentarios
1	ETL	HU001	Limpieza de datos	Data Engineer	Planificada	88	Sprint 1	89	CA1	Leandro J.
2		HU002	Integración de datos	Data Engineer	Planificada	70	Sprint 2	61.5	CA2	Pedro O.
4	EDA	HU003	Análisis exploratorio de datos	Data Analyst	Planificada	42	Sprint 2	44.5	CA3	Juliana C.
5	ML	HU004	Modelo de predicción	Data Scientist	Planificada	67	Sprint 3	40	CA4	Ilbert A.
6	Dashboard	HU005	Desarrollo de Dashboard	Data Analyst	Planificada	51	Sprint 3	38	CA5	Juliana C.
7		HU006	Validación de Dashboard	Data Analyst	Planificada	50	Sprint 3	34	CA6	Juliana C.
9	Cloud	HU007	Despliegue en la nube	Arquitecto Cloud	Planificada	65	Sprint 3	33	CA7	David H.

3.1.4 Tablero Scrum Base.

Para la realización del tablero scrum base, se está utilizando la herramienta “Click Up”, en la que se consideran los estados: por hacer (to do), haciendo (doing) y terminado (done) como se detalla. Se debe tener en cuenta que, para realizar la definición del tablero, se hizo uso de las historias de usuario ya establecidas en el producto backlog.

Figura 6.

Tablero scrum base

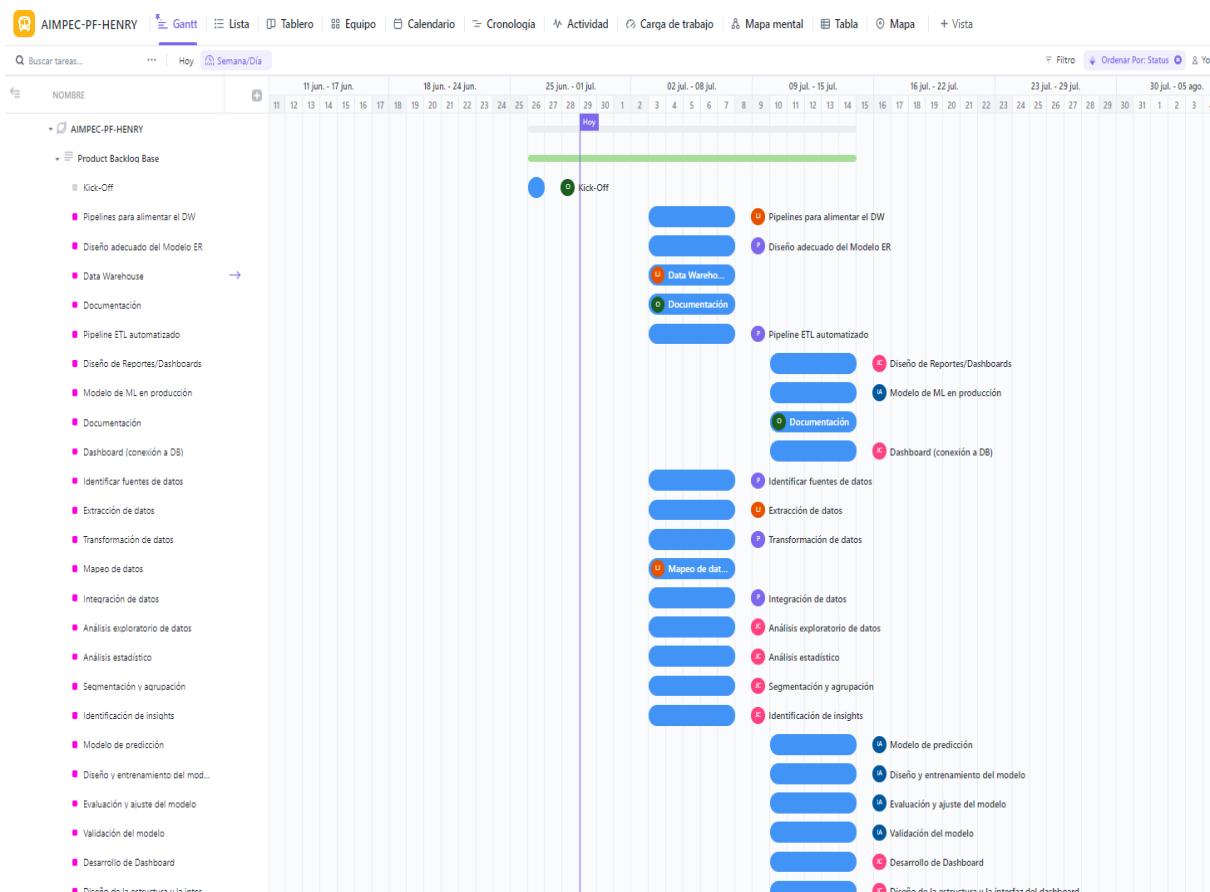
AIMPEC-PF-HENRY						Gantt	Lista	Tablero	Equipo	Calendario	Cronología	Actividad	Carga de trabajo
#	NOMBRE DE LA TAREA	PERSONA ASIGNADA	ESTADO	FECHA LÍMITE	PRIORIDAD	Filtro		Agrupar por: N					
1	Crear Lienzo Lean Canvas	OscarHospital	DONE	Ayer	■								
2	Completar los Issues y Uniqu...	Leandro Gastón Jofré	OPEN		■								
3	Completar Special Advanta...	Juliana Caceres	OPEN		■								
4	Completar KPIs	Ilbert Alarcon	OPEN		■								
5	Completar Solution,Cost St...	OscarHospital	OPEN		■								
6	Completar Lienzo Sources ...	pedrooria23@gmail.com	OPEN		■								
7	Kick-Off	OscarHospital	OPEN	Hace 3 días	■								
8	Organización y División de tar...	OscarHospital	DONE	Hace 2 días	■								
9	Entendimiento de la situación ...	Juliana Caceres	DOING	Mañana	■								
10	Objetivos	Ilbert Alarcon	DOING	Mañana	■								
11	Alcance	Ilbert Alarcon	DOING	Hoy	■								
12	Objetivos y KPIs asociados (pla...	Ilbert Alarcon	DONE	Hoy	■								
13	Creación del Repositorio Github	Ilbert Alarcon	DONE	Ayer	■								
14	Solución propuesta	OscarHospital	DONE	Ayer	■								
15	Incluir stack tecnológico	Leandro Gastón Jofré	DONE	Hoy	■								
16	Creación del Equipo de trabaj...	Juliana Caceres	DONE	Ayer	■								
17	Creación del Cronograma gen...	OscarHospital	DOING	Mañana	■								
18	Análisis preliminar de calidad ...	pedrooria23@gmail.com	DOING	Mañana	■								
19	Realizar un análisis inicial de lo...	Leandro Gastón Jofré	DOING	Mañana	■								
20	Eliminación de datos irrelevant...	Leandro Gastón Jofré	DOING	Mañana	■								
21	Tratamiento de valores faltantes	Leandro Gastón Jofré	DOING	Mañana	■								
22	Limpieza de errores y valores a...	pedrooria23@gmail.com	DOING	Mañana	■								
23	Verificación de integridad y co...	pedrooria23@gmail.com	DOING	Mañana	■								
24	Pipelines para alimentar el DW	Leandro Gastón Jofré	TO DO	7/7/23	■								
25	Diseño adecuado del Modelo ...	pedrooria23@gmail.com	TO DO	7/7/23	■								
26	Data Warehouse	Leandro Gastón Jofré	TO DO	7/7/23	■								

3.1.5 Diagrama de Gantt.

La razón de utilizar un diagrama de Gantt en el proyecto es para visualizar y planificar de manera efectiva las tareas, el tiempo y la secuencia de actividades a lo largo de todo el proyecto. El diagrama de Gantt permite mostrar las fechas de inicio y finalización de cada tarea, así como las dependencias entre ellas, lo que facilita la coordinación y el seguimiento del progreso del proyecto. Además, el diagrama de Gantt proporciona una visión clara del cronograma del proyecto, permitiendo identificar posibles superposiciones de tareas, retrasos o adelantos, y ajustar el plan en consecuencia. En resumen, el diagrama de Gantt es una herramienta visual esencial para la planificación y gestión eficiente del proyecto, asegurando que se cumplan los plazos y se logren los objetivos establecidos.

Figura 7.

Diagrama de gannt

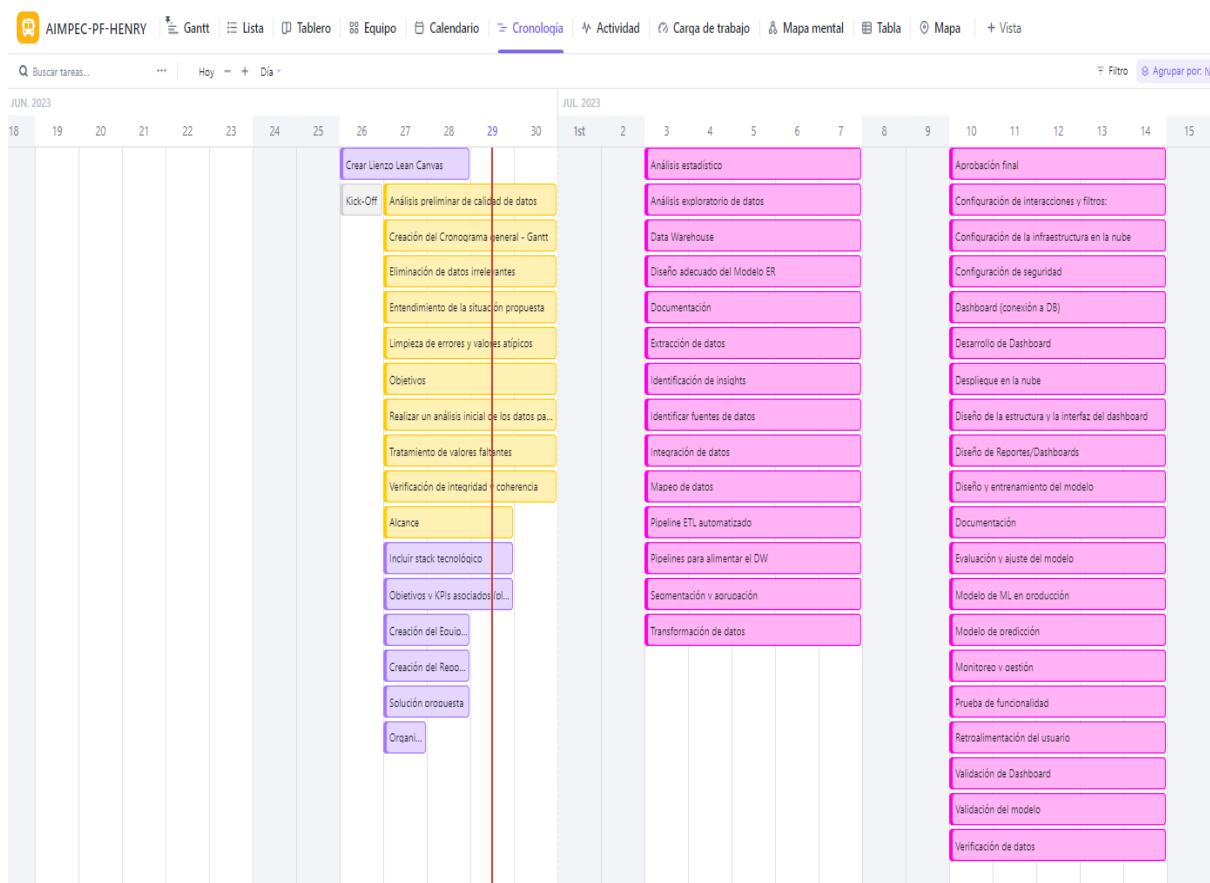


3.1.6 Cronología del Proyecto.

La cronología del proyecto se desarrollará en distintas etapas, siguiendo un enfoque ágil y flexible. A continuación, se presenta un resumen de la cronología:

Figura 8.

Cronología del proyecto



3.2 Análisis preliminar de calidad de datos

El análisis de calidad de datos en el proyecto "Plataforma de Análisis Interactivo y Modelo de Predicción de Emisiones de Carbono en Taxis de NYC" es un proceso fundamental que garantiza la precisión, consistencia e integridad de los datos antes de que sean utilizados para el modelado y la toma de decisiones.

Tabla 2.

Análisis de calidad de datos

PROCESOS/ DATA	OBJETIVOS	EXTRACCION	LIMPIEZA	INSPECCION FINAL	MUESTREO	OBSERVACIONES
TAXIS	-Normalizar la data. -Identificar posibles variables importantes.	-csv -parquet -páginas web	-Desorden en las columnas. -Presencia de valores nulos.	-Separar columnas importantes. -Ordenar datos.	-Separar data para análisis.	Datos faltantes en los registros 264 y 265, debido a que la columna de borough contaba con distritos Unkonwn.
CLIMA	-Normalizar la data. -Identificar posibles variables importantes.	-csv	-Desorden en el formato. -Presencia de valores nulos.	-Ordenar y setear datos.	-Separar data para análisis.	Datos faltantes para los registros después del 22 de junio, no se encontró registros sobre niebla.
CONTAMINACION SONORA	-Normalizar la data. -Identificar posibles variables importantes.	-csv	-Presencia de valores nulos. -Presencia de vacíos. -Datos duplicados.	-Separar las columnas importantes. -Ordenar y Normalizar los datos.	-Separar data para análisis.	No se logró encontrar datos actualizados, contamos con datos hasta 2019.
CONTAMINACION DEL AIRE	-Normalizar la data. -Identificar posibles variables importantes.	CSV	Transformación de columnas. Quitar nulos y columnas adicionales	Escoger los parámetros para el proyecto.	-Separar data para análisis.	Falta información relacionada al CO2

3.3 Arquitectura de software

El resultado de los esfuerzos relacionados con el diseño de arquitectura de software de alto nivel se desarrollará en este capítulo. Se detallarán los objetivos y las limitaciones, los mecanismos arquitectónicos, las perspectivas lógicas, la implementación y la entrega, así como la prueba de concepto para esta arquitectura de software.

3.3.1 Arquitectura Técnica Base

3.2.1.2 Arquitectura Técnica Base TO-BE

El punto "Arquitectura Técnica Base TO-BE" se refiere a la descripción de la arquitectura técnica que se utilizará en el proyecto. A continuación, se detallan las tecnologías que se utilizarán en cada etapa:

1.ETL (Extracción, Transformación y Carga de datos):

. Databricks: Plataforma que proporciona un entorno de trabajo colaborativo para realizar tareas de ETL y análisis de datos utilizando herramientas como PySpark.

. Jupyter Notebook: Entorno interactivo de programación que permite escribir y ejecutar código en Python para realizar tareas de extracción y transformación de datos.

. Beautiful Soup: Biblioteca de Python utilizada para el scraping web, permitiendo extraer datos de páginas HTML o XML.

Visual Studio Code: Entorno de desarrollo integrado (IDE) utilizado para escribir y editar el código en Python.

2.Deploy y despliegue:

.Base de datos en formato .parquet: Se utilizará el formato de archivo Parquet para almacenar los datos procesados, ya que ofrece una alta compresión y un acceso eficiente a los datos.

. Conexión a un S3 de AWS: Se establecerá una conexión con el servicio de almacenamiento en la nube Amazon S3 para almacenar los archivos Parquet y facilitar el acceso y la escalabilidad de los datos.

. Modelo en Machine Learning: Se utilizará un modelo de Machine Learning desarrollado previamente para estimar las emisiones de carbono en los taxis de NYC.

. Plataforma de Hugging Face: Se aprovechará la plataforma de Hugging Face para implementar y gestionar el modelo de Machine Learning, brindando una interfaz para su uso y despliegue eficiente.

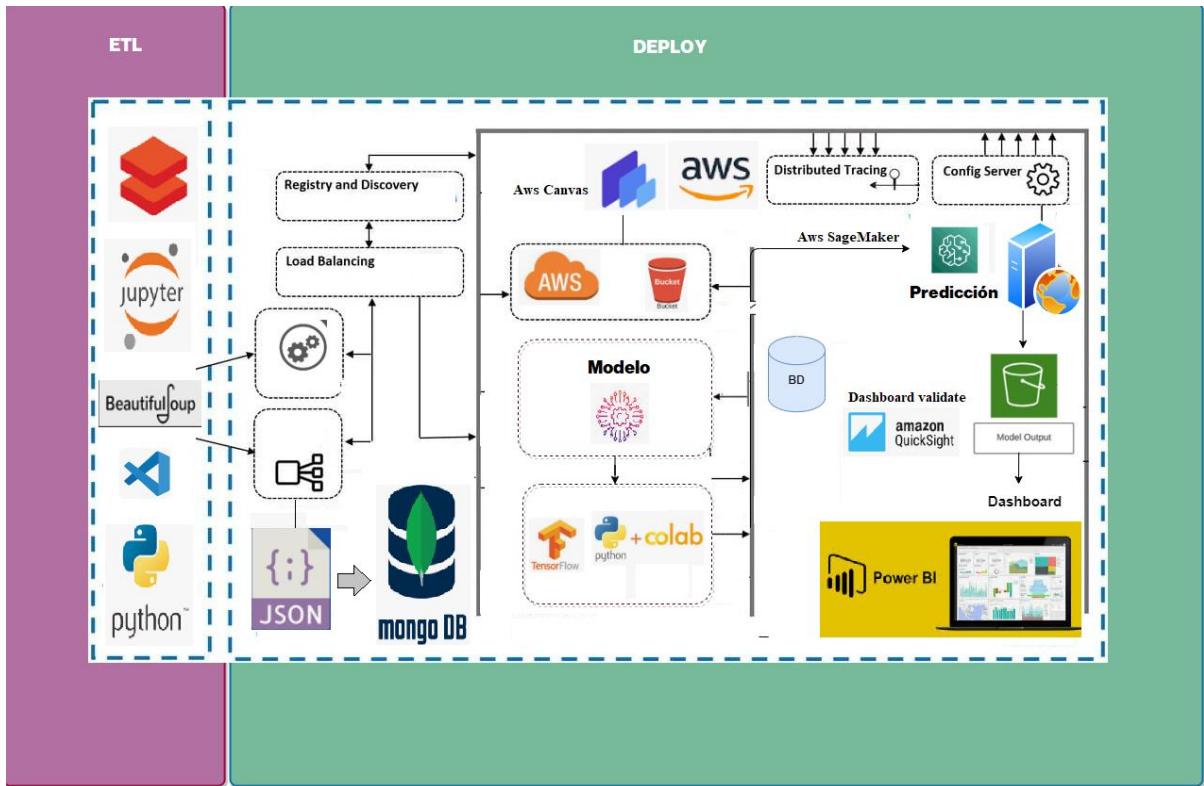
3.Output: Dashboard en Power BI

Power BI: Herramienta de visualización de datos que permitirá crear un dashboard interactivo y dinámico para presentar los resultados del análisis de datos y las predicciones del modelo de Machine Learning.

En resumen, la arquitectura técnica propuesta utiliza tecnologías como Databricks, Jupyter Notebook, Beautiful Soup, Visual Studio Code, Parquet, AWS S3, Hugging Face y Power BI para realizar el ETL de los datos, almacenarlos en un formato eficiente, desplegar el modelo de Machine Learning y finalmente presentar los resultados a través de un dashboard interactivo en Power BI.

Figura 9.

Arquitectura técnica Base



3.4 Desarrollo de Sprint 1

3.4.1 Sprint Planning 1

Definition of Done (DoD)

A fin de garantizar la calidad del incremento y de cumplir con las condiciones requeridas por el producto, se debe cumplir lo siguiente:

- Todas las pruebas unitarias y funcionales deben ser correctas y validadas por los desarrolladores.
- El código debe estar completo de acuerdo con los estándares de desarrollo del equipo.
- El código de desarrollo del aplicativo debe estar versionado y en GitHub.
- El despliegue del proyecto debe estar en un entorno Dev.
- Todos los criterios de aceptación deben cumplirse.
- Todos los bugs deben estar corregidos.

- Los módulos dentro del alcance del sprint 1 deben ser aceptados por el Product Owner.

Historias de usuario para el Sprint 1

Las historias de usuario pertenecientes al Sprint 01, están compuesta por HU001 (Limpieza de datos), dentro de las cuales, se detallará por cada una los puntos de historia, criterios de aceptación y tareas técnicas correspondientes.

Figura 10.

Historia de usuario HU001-Epica01-Sprint01

DOING	10 TAREAS	PERSONA ASIGNADA	FECHA LÍMITE	PRIORIDAD
●	Análisis preliminar de calidad de datos entregable 01	P	Mañana	■
●	Realizar un análisis inicial de los datos para comprender su estructura, calidad y características epica01 hu001 sprint01	U	Mañana	■
●	Eliminación de datos irrelevantes epica01 hu001 sprint01	U	Mañana	■
●	Tratamiento de valores faltantes epica01 hu001 sprint01	U	Mañana	■
●	Limpieza de errores y valores atípicos epica01 hu001 sprint01	P	Mañana	■
●	Verificación de integridad y coherencia epica01 hu001 sprint01	P	Mañana	■
+ Nueva tarea				

Resumen del Sprint Planning 1

La duración del Sprint 1 tiene una duración total de 1 semana.

Tabla 3.

Resumen sprint planning 1

<i>Fecha Inicio del Sprint</i>	<i>Fecha Fin del Sprint</i>	<i>Días</i>
26/06/2023	30/06/2023	05

Sprint Backlog 1

Dentro del Sprint Backlog se detallan las tareas técnicas específicas para cada historia de usuario.

Tablero Scrum

El avance de la historia de usuario especificadas para el Sprint 01, se encuentra concluidas en fecha y presenta el detalle.

3.4.2 Construcción del Sprint 1

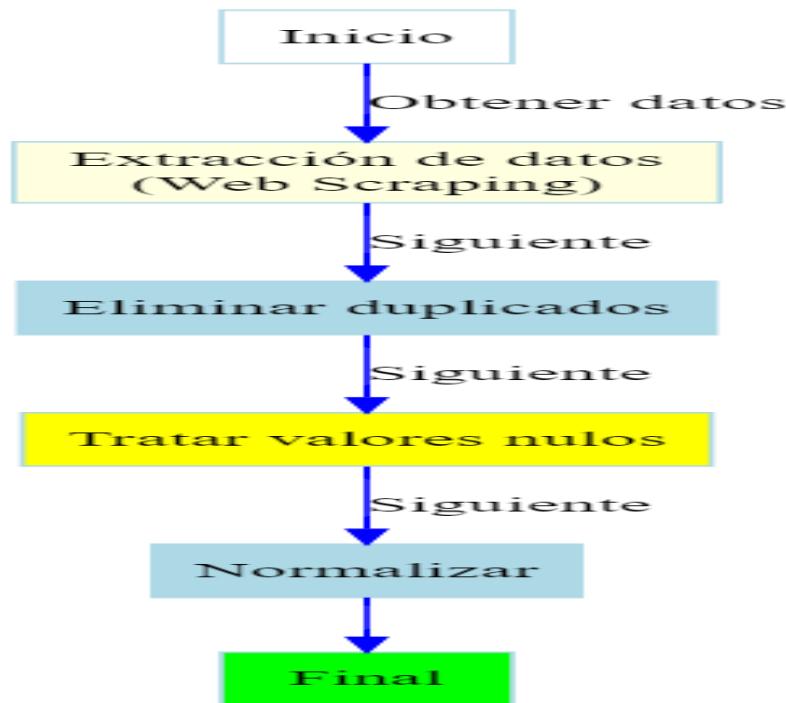
Diseño de prototipos

Los prototipos pertenecientes a cada historia de usuario, es decir, HU001, se detallan como prototipos realizados y con componentes actualizados.

Aquí se muestra un diagrama de flujo o diagrama de proceso que detalla las etapas de limpieza de datos y las decisiones que se toman en cada etapa. Por ejemplo, eliminación de duplicados, tratamiento de valores nulos, normalización, etc.

Figura 11.

Mockup de la HU001



Product Backlog Refinado

El Product Backlog Refinado del Sprint 01 proporciona una visión clara de las tareas a realizar y establece una base sólida para el inicio del desarrollo durante el primer sprint.

Figura 12.

Product Backlog Refinado - Sprint01

The screenshot shows a product backlog refinement board for Sprint 01. The board is organized into four columns: OPEN (1), TO DO (34), DOING (10), and DONE (7). The DOING column is highlighted with a red box. Each task card includes a title, description, due date, duration, and epics/hotfixes/sprints assigned. The tasks in the DOING column are:

- Realizar un análisis inicial de los datos para comprender su estructura, calidad y características (Hace 2 días - Mañana 15h)
- Eliminación de datos irrelevantes (Hace 2 días - Mañana 4h)
- Tratamiento de valores faltantes (Hace 2 días - Mañana 4h)
- Limpieza de errores y valores atípicos (Hace 2 días - Mañana 4h)

Other tasks visible in the TO DO and DONE columns include:

- Pipelines para alimentar el DW (entregable02)
- Diseño adecuado del Modelo ER (entregable02)
- Data Warehouse (entregable02)
- Documentación (entregable02)
- Roles y responsabilidades (entregable01)
- Incluir stack tecnológico (entregable01)
- Organización y División de tareas (entregable01)
- Creación del Repositorio Github (entregable01)

3.4.3 Daily Meeting

Gestión del tablero en el Sprint

- La frecuencia de las reuniones será de cinco veces por semana, de lunes a viernes.
- Cada reunión tendrá una duración máxima de 45 minutos.
- Durante la reunión se realizará la revisión de los avances presentes y futuros por cada historia de usuario.

Revisión del progreso. Se han realizado todas las historias de usuario con un progreso concluido del 100% tal como se especifica:

Tabla 4.

Sprint 01 – Revisión del progreso.

Historia de Usuario	Progreso	Estado
HU001	100%	Terminado

Identificación de impedimentos, bloqueos, dependencias y riesgos. Dentro de estas, se encuentran los siguientes impedimentos:

- Zonas horarias diferentes entre los desarrolladores del proyecto.

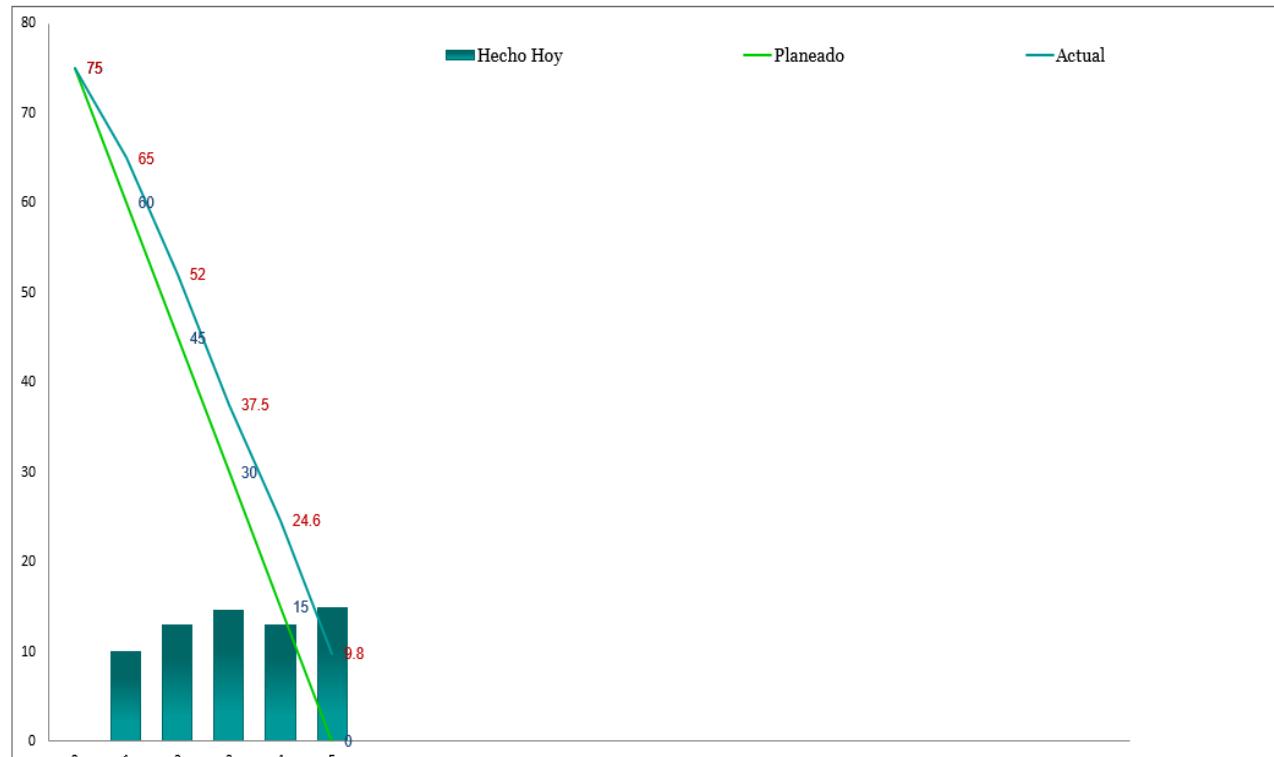
Sprint Burndown Chart

El gráfico de Burndown Chart se ingresa diariamente, después de cada dayli meeting y está realizado por el Scrum Master del proyecto con la finalidad de representar el esfuerzo esperado versus el real para el Sprint 1. Para poder obtener el porcentaje de avance diario, se toman los puntos de historia definidos en el Product Backlog, y se establece el esfuerzo diario obtenido.

Figura 13.

Burndown Chart -Sprint01

Burn Down Chart

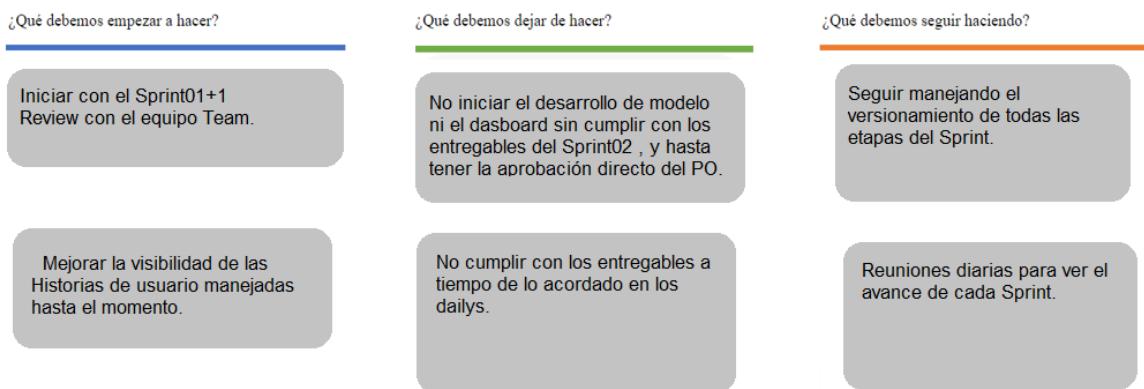


HU001	75
HU002	
HU003	
HU004	
HU005	
HU006	
HU007	
Total	75

3.4.4 Sprint Retrospective

Dentro de los planes de mejora contemplados en la reunión realizada el día 30/06/2023, se obtuvieron los puntos detallados.

Figura 14. Sprint Retrospective 01



3.5 Desarrollo de Sprint 2

3.5.1 Sprint Planning 2

Definition of Done (DoD)

A fin de garantizar la calidad del incremento y de cumplir con las condiciones requeridas por el producto, se debe cumplir lo siguiente:

- Todas las pruebas unitarias y funcionales deben ser correctas y validadas por los desarrolladores.
- El código debe estar completo de acuerdo con los estándares de desarrollo del equipo.
- El código de desarrollo del aplicativo debe estar versionado y en GitHub.

- El despliegue del proyecto debe estar en un entorno Dev.
- Todos los criterios de aceptación deben cumplirse.
- Todos los bugs deben estar corregidos.
- Los módulos dentro del alcance del sprint 2 deben ser aceptados por el Product Owner.

Historias de usuario para el Sprint 2

Las historias de usuario pertenecientes al Sprint 02, están compuesta por HU002 (Integración de datos) y HU003(Análisis exploratorio de datos), dentro de las cuales, se detallará por cada una los puntos de historia, criterios de aceptación y tareas técnicas correspondientes.

Figura 15.

Historia de usuario HU002 -Epica01-Sprint02

DOING	14 TAREAS	PERSONA ASIGNADA	FECHA LÍMITE	PRIORIDAD
Product Backlog Base				
Integración de datos	= epica01 hu002 sprint02	p	vie.	!
Product Backlog Base				
Mapeo de datos	= epica01 hu002 sprint02	U	vie.	!
Product Backlog Base				
Transformación de datos	= epica01 hu002 sprint02	p	vie.	!
Product Backlog Base				
Extracción de datos	= epica01 hu002 sprint02	U	vie.	!
Product Backlog Base				
Identificar fuentes de datos	= epica01 hu002 sprint02	p	vie.	!

Figura 16.

Historia de usuario HU003 -Epica02-Sprint02

DOING	14 TAREAS	PERSONA ASIGNADA	FECHA LIMITE	PRIORIDAD
Product backlog base		JC	vie.	!
Identificación de insights	= epica02 hu003 sprint02			
Product Backlog Base		JC	vie.	!
Segmentación y agrupación	= epica02 hu003 sprint02			
Product Backlog Base		JC	vie.	!
Ánálisis estadístico	= epica02 hu003 sprint02			
Product Backlog Base		JC	vie.	!
Ánálisis exploratorio de datos	= epica02 hu003 sprint02			

Resumen del Sprint Planning 2

La duración del Sprint 2 tiene una duración total de 1 semana.

Tabla 5.

Resumen sprint planning 2

Fecha Inicio del Sprint	Fecha Fin del Sprint	Días
03/07/2023	07/07/2023	05

Sprint Backlog 2

Dentro del Sprint Backlog se detallan las tareas técnicas específicas para cada historia de usuario.

Tablero Scrum

El avance de la historia de usuario especificadas para el Sprint 02, se encuentra concluidas en fecha y presenta el detalle.

3.5.2 Construcción del Sprint 2

Diseño de prototipos

Los prototipos pertenecientes a cada historia de usuario, es decir, HU002 y HU003, se detallan como prototipos realizados y con componentes actualizados. Aquí se muestra un diagrama de flujo o diagrama de proceso que detalla las etapas de integración y análisis exploratorio de datos y las decisiones que se toman en cada etapa.

Figura 17.

Mockup de la HU002

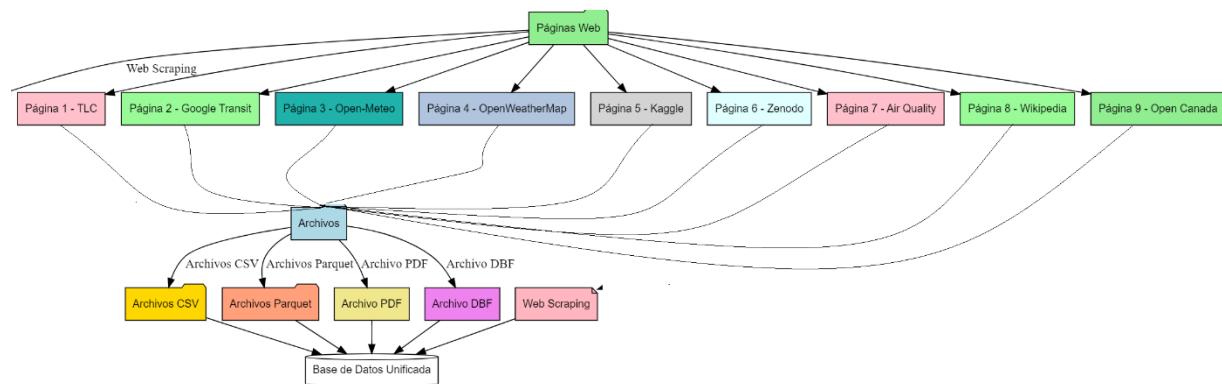
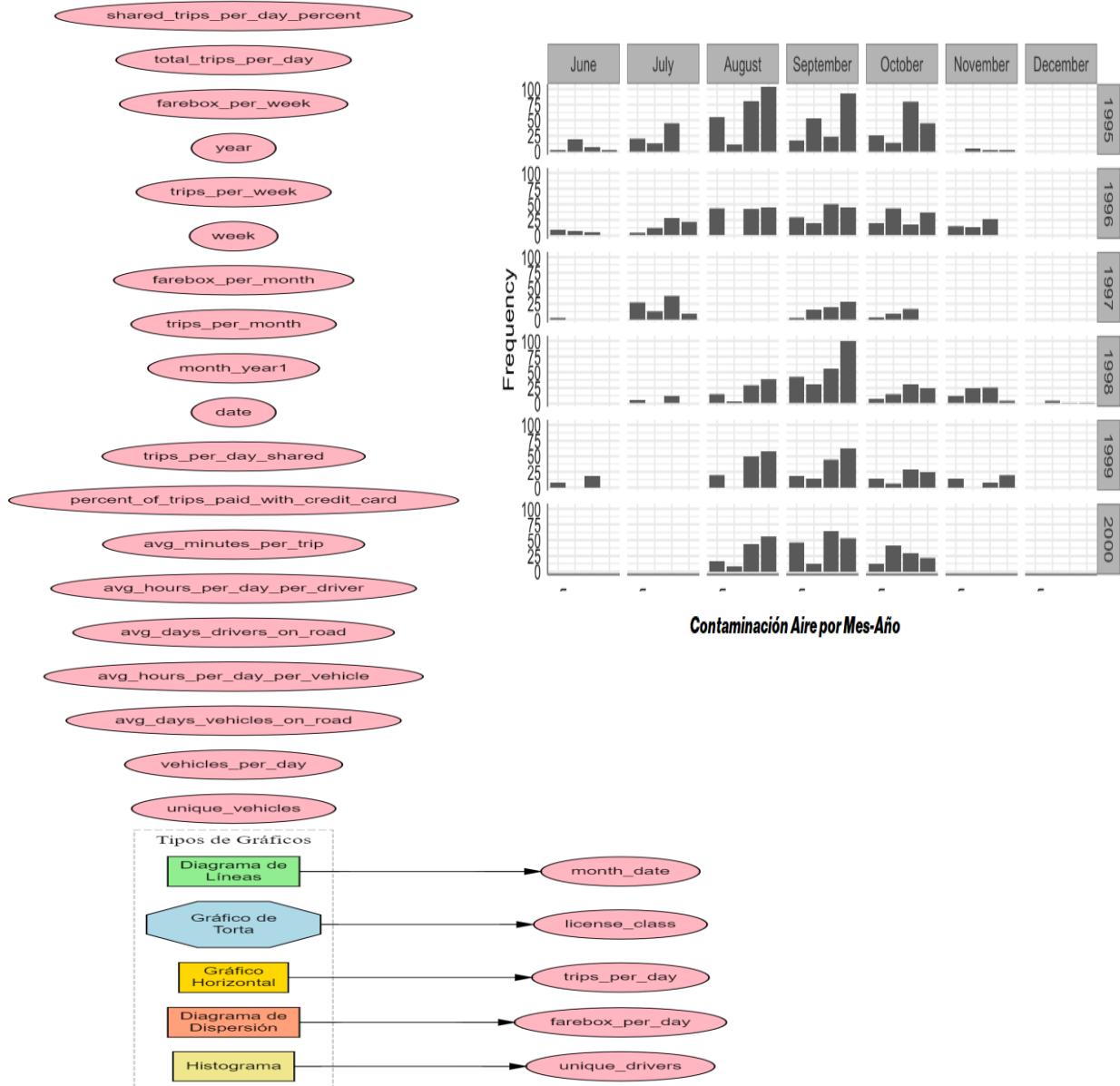


Figura 18.

Mockup de la HU003

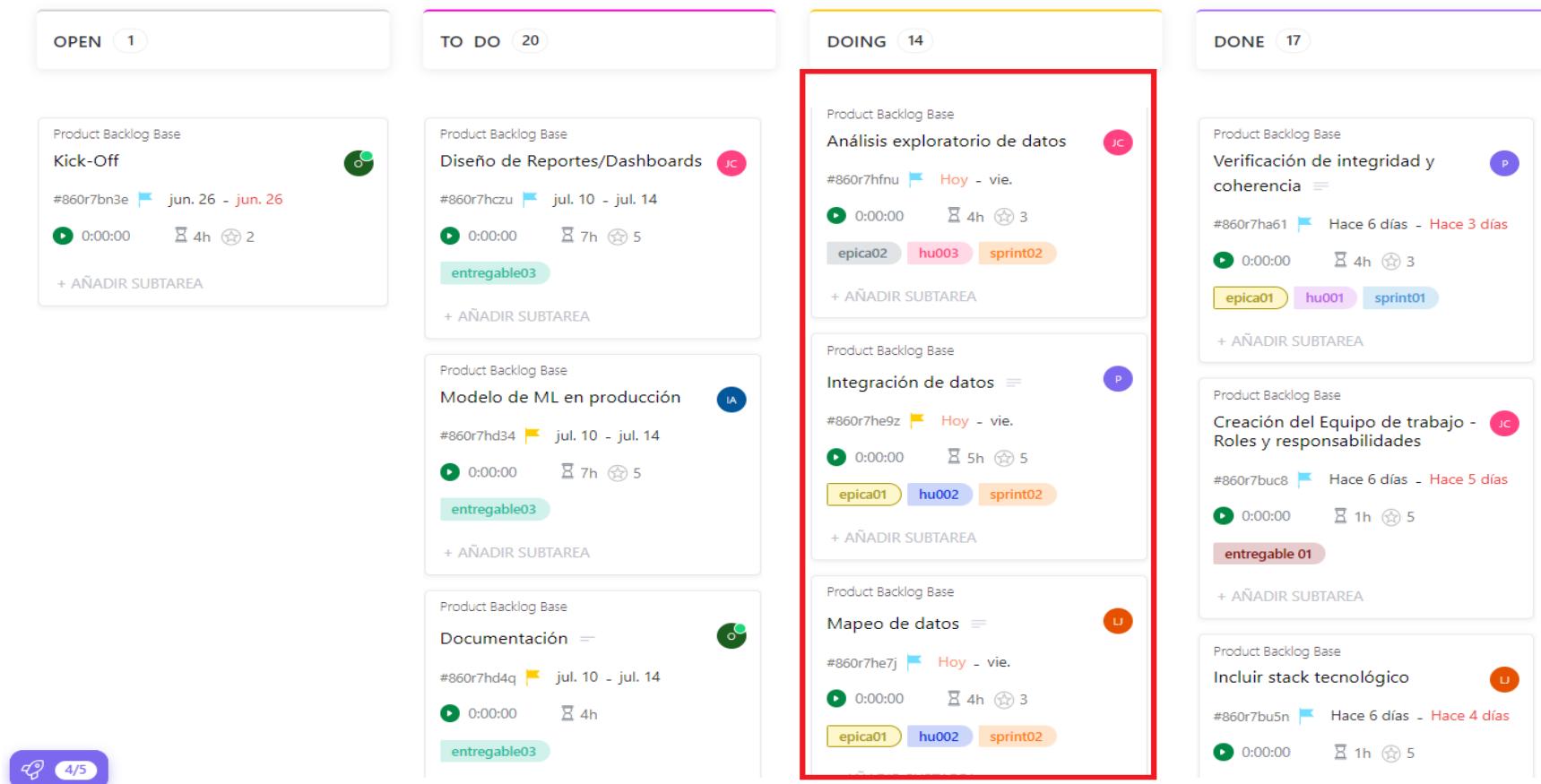


Product Backlog Refinado

El Product Backlog Refinado del Sprint 02 proporciona una visión clara de las tareas a realizar y establece una base sólida para el inicio del desarrollo durante el segundo sprint.

Figura 19.

Product backlog -Sprint02



Incremento del Sprint

El incremento del Sprint actual se centra en la HU002, que implica la integración de datos y la creación de un manual de usuario para MongoDB. A continuación, se detalla el proceso de instalación de MongoDB y MongoDB Compass, junto con los pasos para configurar y activar el servidor en Windows.

Una vez completada la instalación, se mencionan los comandos básicos de MongoDB, como la visualización de colecciones y la búsqueda de información utilizando filtros. Además, se incluyen ejemplos de consultas SQL y su equivalente en MQL (MongoDB Query Language).

También se proporciona información sobre la actualización de datos, incluyendo la modificación de un único ID y la modificación de varios IDs simultáneamente.

Además, se explica el proceso de asignación de privilegios a los usuarios y la creación de colecciones, así como la carga de información utilizando MongoDB Compass.

Finalmente, se menciona la posibilidad de invocar la información desde Visual Studio Code (VSC) para trabajar con los datos de MongoDB.

En resumen, el incremento del Sprint se enfoca en la integración de datos y la creación de un manual de usuario para MongoDB. Se detallan los pasos de instalación, los comandos básicos, la actualización de datos, la asignación de privilegios, la creación de colecciones y la carga de información. También se menciona la opción de invocar la información desde VSC.

Figura 20.

Incremento del Sprint-HU002



MongoDB

Autor: Jofré, Leandro Gastón

[Manual](#) —> para ver el manual de comandos de trabajo de mongo, en su web oficial.

- Instalacion MongoDB y Compass
- Instalacion de NoSQLBooster
- Commandos Basicos
- Formato de visualizacion
- SQL en NoSQLBooster
- Actualizar Datos
- Usuarios y asignacion de privilegios
- Creacion de Collections y cargar informacion (MongoDB Compass)
- Invocar la informacion desde VSC
- Glosario

3.5.2.1 Construcción de la HU002

La Historia de Usuario HU002 se centra en la construcción de la integración de datos. A continuación, se hará un resumen de los pasos clave involucrados en esta tarea:

Diseño adecuado del modelo

En esta etapa, se realiza un diseño cuidadoso de nuestro modelo que representa la estructura de la base de datos. Se identifican las entidades, atributos y relaciones entre ellas, y se definen las restricciones de integridad necesarias para garantizar la consistencia de los datos.

Tabla 6.

Diseño del modelo conceptual de datos

Nº	Colección	Descripción
C001	Calidad del Aire	<p>La colección "Calidad del Aire" contiene información sobre la calidad del aire, centrándose en las moléculas detectadas por los sensores. Está estructurada por nueve columnas (llaves) que describen diferentes aspectos de la medición:</p> <ul style="list-style-type: none"> -Name: Nombre de la molécula medida. Por ejemplo, "Ozone (O3)". -Measure: Tipo de medida realizada. Por ejemplo, "Mean" (media). -Measure Info: Unidad de medida utilizada. Por ejemplo, "ppb" (partes por billón).

		<ul style="list-style-type: none"> -Geo Type Name: Tipo de ubicación geográfica. Por ejemplo, "CD" (distrito censal). -Geo Join ID: Identificador único de la ubicación geográfica. Por ejemplo, 313. -Geo Place Name: Nombre del lugar geográfico. Por ejemplo, "Coney Island (CD13)". -Time Period: Período de tiempo en el que se realizó la medición. Por ejemplo, "Summer 2013" (verano 2013). -Start_Date: Fecha y hora de inicio del período de medición. Por ejemplo, 2013-06-01T00:00:00.000+00:00. -Data Value: Valor numérico de la medida. Por ejemplo, 34.64.
C002	Combustion y CO2	<p>La colección "Combustion y CO2" detalla las emisiones de vehículos contaminantes de CO2, categorizados por modelo y marca. Está estructurada por las siguientes columnas:</p> <ul style="list-style-type: none"> - Model(Year): Año y modelo del vehículo (ejemplo: 2023). - Make: Marca del vehículo (ejemplo: "Acura"). - Model: Modelo específico del vehículo. - Vehicle Class: Clase del vehículo (ejemplo: "Full-size"). - Engine Size(L): Tamaño del motor en litros (ejemplo: 1.5). - Cylinders: Número de cilindros del motor. - Transmission: Tipo de transmisión del vehículo. - Fuel (Type): Tipo de combustible utilizado (ejemplo: 4). - Fuel Consumption(City (L/100 km)): Consumo de combustible en ciudad en litros por cada 100 km (ejemplo: 7.9). - CO2 Emissions(g/km): Emisiones de CO2 en gramos por kilómetro (ejemplo: 167). - CO2(Rating): Calificación de emisiones de CO2 (ejemplo: 6). <p>En esta colección se detalla los diferentes sonidos producidos por un taxi que contaminan New York. Está estructurada por 6 columnas (llaves):</p> <ul style="list-style-type: none"> -Fecha: 2016-05-01T00:00:00.000+00:00 (formato de fecha y hora UTC). -id_borough: Identificador de la zona geográfica correspondiente al borough (distrito) donde se registraron los sonidos (ejemplo: 1) -engine_sounds: Calificación de los sonidos del motor (valor numérico). (ejemplo: 5) -alert_signal_sounds: Calificación de los sonidos de las señales de alerta (valor numérico). (ejemplo: 3) -total_sounds: Calificación total de sonidos (valor numérico). (ejemplo: 8) -borough_name: Nombre del borough (distrito) donde se registraron los sonidos. (ejemplo: "manhattan")
C003	Contaminación sonora	<p>En esta colección se detallan las áreas y longitudes de los distritos y zonas de New York. Está estructurada por 6 columnas(llaves):</p> <ul style="list-style-type: none"> -Shape_Leng: Longitud de la forma (ejemplo: 0.11635). -Shape_Area: Área de la forma (ejemplo: 0.00078). -LocationID: Identificador de ubicación (ejemplo: 1). - -Borough: Nombre del borough (distrito) correspondiente a la ubicación (ejemplo:). -Zone: Zona geográfica de la ubicación. (ejemplo:"Newark Airport"). -service_zone: Zona de servicio de la ubicación. (ejemplo:"EWR")
C004	Taxis zonas	

C005	Taxis rutas	<p>En esta colección se detallan las rutas, tarifas, pasajeros y tipos de pago que fueron evaluados por un controlador y un taxímetro en la ciudad de New York. Está estructurado por 8 columnas (llaves):</p> <ul style="list-style-type: none"> -Fecha: Fecha del registro de datos (formato: "2022-01-01"). -Pasajeros por día: Número de pasajeros transportados durante el día. -Viajes por día: Número total de viajes realizados durante el día. -Tarifario por día: Monto total del tarifario registrado durante el día. -Total recaudado por día: Monto total recaudado durante el día. -Pago con tarjeta: Número de viajes pagados con tarjeta durante el día. -Pago con efectivo: Número de viajes pagados en efectivo durante el día. -Tipo de Taxi: Tipo de taxi utilizado para los viajes (ejemplo: "green").
------	-------------	---

Pipelines para alimentar el DW

Se crean los pipelines de extracción, transformación y carga (ETL) para obtener datos de diversas fuentes y alimentar el Data Warehouse (DW). Estos pipelines se encargan de recolectar, limpiar y preparar los datos antes de cargarlos en el DW.

Tabla 7.

Detalle de los Pipelines

Nº	Pipelines	Descripción
01	Extracción	<p>Fuentes internet: Total Bruto: 50, Total Neto: 8</p> <p>Archivos csv (17) -----</p> <p>Air_Quality.csv Light Duty Vehicles.csv Alternative Fuel Vehicles US.csv Electric and Alternative Fuel Charging Stations.csv ElectricCarData_Clean.csv ElectricCarData_Norm.csv MY2012-2023 Plug-in Hybrid Electric Vehicles.csv MY2012-2023 Battery Electric Vehicles.csv annotations.csv Clima2020-2023.csv energy.csv MY2023 Fuel Consumption Ratings.csv Vehicle Fuel Economy Data.csv data_reports_monthly-1.csv df_taxis.csv taxi+_zone_lookup.csv test_taxis_kaggle.csv</p>

Archivos Parquet (32) -----
green_tripdata_2022-01.parquet
green_tripdata_2022-02.parquet
green_tripdata_2022-03.parquet
green_tripdata_2022-04.parquet
green_tripdata_2022-05.parquet
green_tripdata_2022-06.parquet
green_tripdata_2022-07.parquet
green_tripdata_2022-08.parquet
green_tripdata_2022-09.parquet
green_tripdata_2022-10.parquet
green_tripdata_2022-11.parquet
green_tripdata_2022-12.parquet
green_tripdata_2023-01.parquet
green_tripdata_2023-02.parquet
green_tripdata_2023-03.parquet
green_tripdata_2023-04.parquet
yellow_tripdata_2022-01.parquet
yellow_tripdata_2022-02.parquet
yellow_tripdata_2022-03.parquet
yellow_tripdata_2022-04.parquet
yellow_tripdata_2022-05.parquet
yellow_tripdata_2022-06.parquet
yellow_tripdata_2022-07.parquet
yellow_tripdata_2022-08.parquet
yellow_tripdata_2022-09.parquet
yellow_tripdata_2022-10.parquet
yellow_tripdata_2022-11.parquet
yellow_tripdata_2022-12.parquet
yellow_tripdata_2023-01.parquet
yellow_tripdata_2023-02.parquet
yellow_tripdata_2023-03.parquet
yellow_tripdata_2023-04.parquet
Archivos dbf (1)
taxi_zones.dbf
Data final (3)
df_taxis.csv
taxis_zonas.csv
dfConta_sonora.csv

MY2023 Fuel Consumption Ratings.csv

- Eliminación de nulos en las columnas y filas
- Eliminación de columnas innecesarias
- Renombramiento de los campos

	Reestablecer los tipos de campos
	Exportar JASON
	Air_Quality.csv
	Eliminación de nulos en las columnas y filas
	Eliminación de columnas innecesarias
	Cambio de formato necesario Date
	Reestablecer los tipos de campos
	df_taxis.csv
	Eliminación de nulos en columnas y filas
	Conversión de las fechas a formato yyyy-mm-dd
	Eliminación de columnas innecesarias
	Eliminación de valores atípicos
	Renombrar los nombres de las columnas
	taxis_zonas.csv
	Eliminación de los valores nulos en filas y columnas
	Merge de las zonas con distritos
	Eliminación de columnas repetidas
	Eliminación de valores atípicos
	dfContaSonora.csv
	Eliminación de valores nulos en filas y columnas
	Eliminación de valores atípicos
	Conversión a formato yyyy-mm-dd
03	Carga
	Ingesta en colecciones en Mongo Bd.json
	Archivos JSON (2) -----
	ETL_Air_Quality.JSON
	ETL_combust_MY2023.JSON
	Archivos csv (5) -----
	taxis_zonas.csv
	df_taxis.csv
	dfContaSonora.csv

Data Warehouse

Aquí se almacenan los datos integrados y estructurados de manera que sean fácilmente accesibles y permitan realizar consultas y análisis eficientes.

Figura 21.

Interfaz Mongo Compas

MongoDB Compass - Mongo Atlas/PROYECTO

Connect Edit View Help

Mongo Atlas

My Queries Databases PROYECTO

Collections + Create collection Refresh

View Sort by Collection Name

Calidad del Aire

Storage size: 905.22 kB Documents: 16 K Avg. document size: 268.00 B Indexes: 1 Total index size: 962.56 kB

Combustion y CO2

Storage size: 73.73 kB Documents: 823 Avg. document size: 309.00 B Indexes: 1 Total index size: 53.26 kB

Contaminación sonora

Storage size: 32.77 kB Documents: 434 Avg. document size: 142.00 B Indexes: 1 Total index size: 28.67 kB

Taxis rutas

Storage size: 90.11 kB Documents: 11 K Avg. document size: 216.00 B Indexes: 1 Total index size: 49.16 kB

Taxis zonas

Storage size: 32.77 kB Documents: 263 Avg. document size: 154.00 B Indexes: 1 Total index size: 24.58 kB

Nota. Aquí se muestra la interfaz de Mongo Compas en la cual se detalla las colecciones ingestadas en nuestro datawarehouse.

Figura 22.

Servicio de Atlas en MongoDB

Atlas pedro's Org ... Access Manager Billing

All Clusters Get Help pedro

Project O Data Services App Services Charts

DEPLOYMENT Database

DATA SERVICES Triggers Data API Data Federation Search Stream Processing

SECURITY Backup Database Access Network Access Advanced Goto

ClusterO PEDRO'S ORG - 2023-07-02 > PROJECT O > DATABASES VERSION 6.0.6 REGION AWS Sao Paulo (sa-east-1)

CREATE COLLECTION

DATABASES: 1 COLLECTIONS: 6

PROYECTO

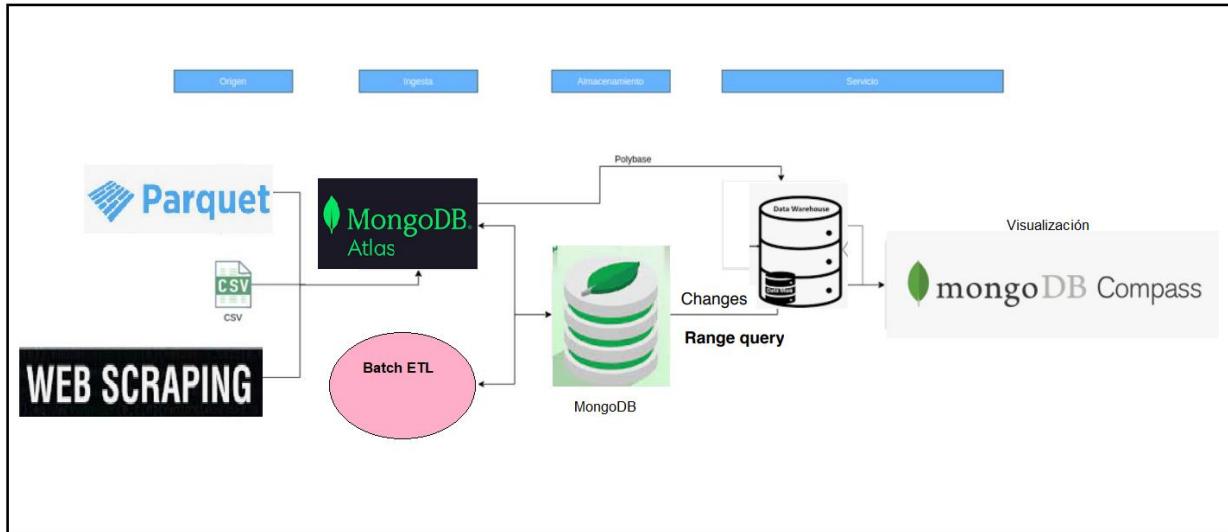
LOGICAL DATA SIZE: 4.71MB STORAGE SIZE: 1.9MB INDEX SIZE: 1.07MB TOTAL COLLECTIONS: 6

Collection Name	Documents	Logical Data Size	Avg Document Size	Storage Size	Indexes	Index Size	Avg Index Size
Calidad del Aire	16122	4.13MB	269B	1.66MB	1	940KB	940KB
Combustion y CO2	823	248.9KB	310B	84KB	1	52KB	52KB
Contaminación sonora	434	60.39KB	143B	32KB	1	28KB	28KB
Taxis rutas	1134	239.74KB	217B	100KB	1	48KB	48KB
Taxis zonas	263	39.70KB	165B	32KB	1	24KB	24KB

Nota. Proporciona una plataforma completamente administrada y escalable para almacenar y gestionar bases de datos MongoDB en la nube.

Figura 23.

Estructura del Data Warehouse



Automatización

Monitoreo y notificaciones: Implementamos un sistema de monitoreo para verificar el estado y el rendimiento de los pipelines automatizados. Si ocurre algún error o se detecta un problema, se enviarían notificaciones automáticas a los responsables para que puedan tomar medidas correctivas de inmediato.

Figura 24.

Creación de Usuarios

The screenshot shows the MongoDB Atlas Database Access page. On the left, there's a sidebar with sections for Deployment, Services (Triggers, Data API, Data Federation, Search, Stream Processing), Security (Backup), and Database Access (Network Access, Advanced). The Database Access section is currently selected. In the main area, there are tabs for 'Database Users' and 'Custom Roles'. Under 'Database Users', two users are listed: 'pedroorio23' (SCRAM, atlasAdmin@admin) and 'proyecto' (SCRAM, readWriteAnyDatabase@admin). Each user has 'All Resources' assigned and edit/delete buttons. A green button at the top right says '+ADD NEW DATABASE USER'. At the bottom of the page, there's a note about system status ('All Good'), copyright information ('©2023 MongoDB, Inc.'), and links to Status, Terms, Privacy, Atlas Blog, Contact, and Sales.

Validación de datos

Definir criterios de validación: Identifica los criterios que se utilizarán para verificar la calidad de los datos. Esto puede incluir reglas de negocio, restricciones de integridad, formatos esperados, rangos válidos, entre otros. Desarrolla scripts, consultas SQL u otras herramientas para automatizar las validaciones de datos. Esto te permitirá detectar errores y anomalías en los datos de manera eficiente y escalable.

Tabla 8.

Tabla detalle de las Reglas de Negocio

Regla Negocio	Descripción	Collection	Valor para el Negocio
RN01	En la colección Contaminación sonora, total_sounds debe ser igual a la suma de engine_sounds y alert_signal_sounds.	Contaminación sonora	Estas reglas verifican la coherencia de las mediciones del sonido. Esto es importante para evitar análisis erróneos basados en mediciones de sonido incorrectas.
RN02	En la colección Combustión y CO2,	Combustión y CO2	Estas reglas aseguran que los datos de los vehículos y sus emisiones sean coherentes y lógicos. Esto es

	Engine Size(L) no debe ser menor a 0.		crucial para realizar análisis precisos y confiables sobre las emisiones de los vehículos.
RN03	En la colección combustión y CO2, CO2 Emissions(g/km) no puede ser menor a 0.	Combustión y CO2	Garantizar que los datos sean valores únicos
RN04	En la colección Taxis zonas, LocationID debe ser un valor único.	Taxis zonas	Estas reglas aseguran que los datos sobre las zonas de taxis sean lógicos y únicos. Estos son fundamentales para evitar confusiones al analizar las zonas de taxis.
RN05	Validar que las fechas se registren en un formato específico y que sean válidas (YYYY/MM/DD).	Taxis rutas	Esta regla es fundamental para el correcto seguimiento y análisis de los datos a lo largo del tiempo. Asegura que todas las fechas estén en un formato consistente, lo que facilita la realización de operaciones de comparación de fechas, análisis de series temporales y generación de informes periódicos. Además, la validación de la fecha también asegura que no se introduzcan datos erróneos en el sistema, lo que podría llevar a interpretaciones incorrectas y a decisiones de negocio erróneas.
RN06	En la colección Taxis rutas, Pasajeros por día, Viajes por día, Tarifario por día, Total recaudado por día, Pago con tarjeta, Pago con efectivo no deben ser menor a 0.	Taxis rutas	Estas reglas verifican la coherencia de los datos de las rutas de taxis. Esta coherencia es crucial para realizar análisis correctos sobre los patrones de viaje y los ingresos de los taxis.
RN07	Todos los formatos de datos ingestados deben estar en .json.	Calidad del Aire, combustión y CO2, Contaminación sonora, Taxis zonas, Taxis rutas	El valor principal de esta regla de negocio es la coherencia y la facilidad de uso. Los archivos JSON son un formato común para la transmisión de datos, especialmente en aplicaciones web, y son fáciles de leer y escribir tanto para las máquinas como para los humanos. Tener todos los datos en un solo formato facilita el procesamiento y el análisis, ya que no es necesario escribir y mantener diferentes rutinas de carga y transformación para cada formato de datos. También minimiza el riesgo de errores en la

			interpretación de los datos debido a diferencias en los formatos de archivo.
--	--	--	--

Tabla 9.

Tabla detalle de la Validación de scripts de la regla del Negocio

Regla Negocio	Ejecución del Script para su Validación
RN01	assert (df_contaminacion_sonora['engine_sounds'] + df_contaminacion_sonora['alert_signal_sounds'] == df_contaminacion_sonora['total_sounds']).all(), "Total sound score doesn't match the sum of engine and alert sounds"
RN02	assert (df_combustion_co2['Engine Size(L)'] >= 0).all(), "Negative engine size found"
RN03	assert (df_combustion_co2['CO2 Emissions(g/km)'] >= 0).all(), "Negative CO2 emissions found"
RN04	assert (df_contaminacion_sonora[['engine_sounds', 'alert_signal_sounds', 'total_sounds']] >= 0).all().all(), "Negative sound score found"
RN05	<pre> df = pd.read_json('taxis_rutas.json') # Definir una función para comprobar si una fecha es válida y está en el formato correcto def check_date_format(date_str): try: pd.to_datetime(date_str, format='%Y/%m/%d') return True except ValueError: return False # Aplicar la función de comprobación a la columna de fecha df['is_date_valid'] = df['Fecha'].apply(check_date_format) # Verificar si hay alguna fecha no válida invalid_dates = df[~df['is_date_valid']] print(invalid_dates) </pre>
RN06	assert (df_taxis_rutas[['Pasajeros por día', 'Viajes por día', 'Tarifario por día', 'Total recaudado por día', 'Pago con tarjeta', 'Pago con efectivo']] >= 0).all().all(), "Negative values found in Taxi Routes data"

RN07	<pre> import pandas as pd # Lista de colecciones collections = ['calidad_del_aire.json', 'combustion_y_co2.json', 'contaminacion_sonora.json', 'taxis_zonas.json', 'taxis_rutas.json'] # Definir una función para comprobar si un archivo es un JSON válido def is_valid_json(file): try: pd.read_json(file) return True except ValueError: return False # Comprobar cada archivo for collection in collections: if is_valid_json(collection): print(f'{collection} es un archivo JSON válido.') else: print(f'{collection} no es un archivo JSON válido.') </pre>
------	---

Diccionario de Datos

Figura 25.

Dataset ETL_AirQ.csv

Nombre	Registros	Campos							
ETL_AirQ.csv	16112	9							
Name	Nombre del indicador								
Measure	Como el indicador es medido								
Measure Info	Informacion (como unidades) acerca de la medicion								
Geo Type Name	Tipo de Geografia								
Geo Join ID	Identificador del area geografica del vecindario								
Geo Place Name	Nombre del vecindario								
Time Period	Descripcion de el tiempo que el dato aplica								
Start_Date	Valor del dato para el inicio del periodo de tiempo								
Data Value	El valor del dato actual para este indicador, medida, valor, etc.								
Apariencia									
Name	Measure	Measure Info	Geo Type Name	Geo Join ID	Geo Place Name	Time Period	Start Date	Data Value	
0 Ozone (O3)	Mean	ppb	CD	313	Coney Island (CD13)	Summer 2013	2013-06-01	34.64	
1 Ozone (O3)	Mean	ppb	CD	313	Coney Island (CD13)	Summer 2014	2014-06-01	33.22	

Figura 26.

Dataset ETL_C_Combust.csv

Nombre	Registros	Campos										
ETL_C_Combust.csv	823	12										
Model(Year)	Año del modelo del vehiculo											
Make	Fabricante											
Model.1	Modelo de la marca del vehiculo											
Vehicle Class	Clase del vehiculo											
Engine Size(L)	Tipo del Motor en litros											
Cylinders	Cilindros											
Transmission	Transmission											
Fuel(Type)	Tipo del combustible											
Fuel Consumption(City (L/100km))	Consumision de combustible en litros por cada 100 km											
CO2 Emissions(g/km)	Cantidad de emisiones de CO2 en gramos por km											
CO2(Rating)	Indicador de contaminacion de CO2 del vehiculo											
Smog(Rating)	Indicador de contaminacion de Smog del vehiculo											
Apariencia												
Model(Year)	Make	Model.1	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel(Type)	Fuel Consumption(City (L/100 km))	CO2 Emissions(g/km)	CO2(Rating)	Smog(Rating)	
0	2023	Acura	Integra	Full-size	1.5	4.0	AV7	Z	7.9	167	6	7
1	2023	Acura	Integra A-SPEC	Full-size	1.5	4.0	AV7	Z	8.1	172	6	7

Figura 27.

Dataset Taxis Rutas

Nombre	Registros	Campos							
df_taxis.csv	1136	8							
Fecha	registro del viaje								
Pasajeros por dia	clase de la licencia								
Viajes por dia	viajes por dia								
Tarifario por dia	cantidad de dinero que se recauda por dia								
Total recaudado por dia	Cantidad de contuctores unicos								
Pago con tarjeta	Cantidad de vehiculos que transitan								
Pago con efectivo	promedio de dias de los vehiculos en calle								
Tipo de taxi	promedio de dias de los conductores en calle								
Apariencia									
Fecha	Pasajeros por dia	Viajes por dia	Tarifario por dia	Total recaudado por dia	Pago con tarjeta	Pago con efectivo	Tipo de Taxi		
0	2022-01-01	1320	1273	20361.61	25128.72	568	508	green	
1	2022-01-02	1739	1500	21448.61	26851.77	830	543	green	
2	2022-01-03	2716	2332	30798.87	38778.99	1216	917	green	
3	2022-01-04	2438	2165	30469.21	38390.10	1190	793	green	
4	2022-01-05	2650	2259	30853.43	39222.21	1266	806	green	

Figura 28.

Dataset Taxis Zonas

Nombre	Registros	Campos								
taxis_zonas.csv	263	6								
Shape_Leng			area de la forma de los borough							
Shape_Area			longitud de la forma de los borough							
LocationID			identificador de ubicación							
Borough			Nombre de los distritos o zonas							
Zone			Nombre de la zona geográfica							
service_zone			nombre del servicio de la zona							
Apariencia										
Shape_Leng	Shape_Area	LocationID	Borough		Zone	service_zone				
0	0.116357	0.000782	1	EWR	Newark Airport	EWR				
1	0.433470	0.004866	2	Queens	Jamaica Bay	Boro Zone				
2	0.084341	0.000314	3	Bronx	Allerton/Pelham Gardens	Boro Zone				
3	0.043567	0.000112	4	Manhattan	Alphabet City	Yellow Zone				
4	0.092146	0.000498	5	Staten Island	Arden Heights	Boro Zone				

Figura 29.

Dataset Sonora

Nombre	Registros	Campos								
dfContaSonora.csv	434	6								
fecha			Fecha de los registros							
id_borough			Id del distrito							
engine_sounds			Cantidad de sonidos de motor registrados							
alert_signal_sounds			Cantidad de sonidos de señal de alarmas registradas							
total_sounds			Cantidad total de sonidos registrados							
borough_name			Nombre del distrito							
Apariencia										
fecha	id_borough	engine_sounds	alert_signal_sounds	total_sounds	borough_name					
2	2016-05-01	1	5	3	8	manhattan				
3	2016-05-02	1	9	14	23	manhattan				
4	2016-05-03	1	4	3	7	manhattan				

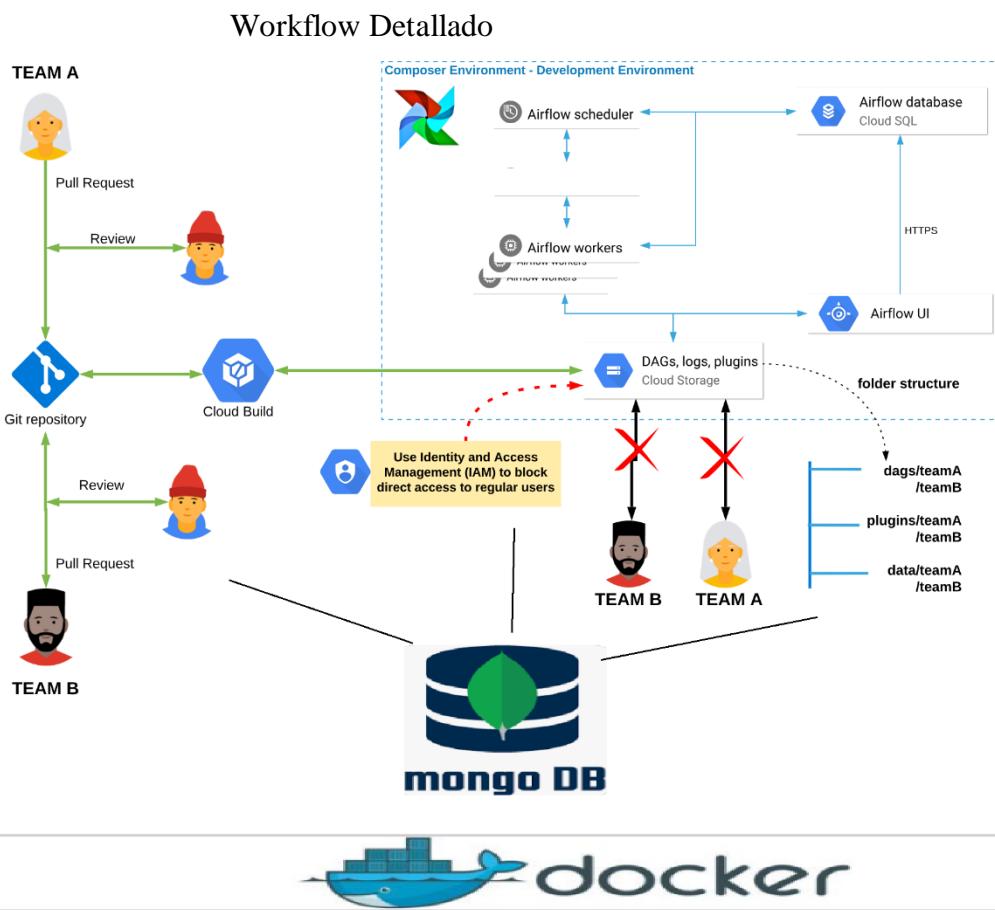
Workflow detallado en tecnologías

Docker: Utilizas Docker para crear y administrar contenedores virtuales que contienen todas las dependencias necesarias para ejecutar tu aplicación, incluyendo MongoDB y Airflow. Docker facilita la creación de entornos consistentes y reproducibles, lo que garantiza que tu aplicación se ejecute de la misma manera en diferentes entornos.

MongoDB: Utilizas MongoDB como tu base de datos para almacenar y gestionar tus datos. Puedes utilizar las capacidades de MongoDB para realizar operaciones CRUD (Create, Read, Update, Delete) en tus datos, así como para realizar consultas y agregaciones avanzadas. MongoDB te permite almacenar datos estructurados y no estructurados de forma flexible y escalable.

Airflow: Utilizas Apache Airflow como tu orquestador de tareas. Airflow te permite definir y programar flujos de trabajo complejos mediante la creación de DAGs (Directed Acyclic Graphs). Cada DAG consta de tareas individuales que se ejecutan en un orden determinado y pueden tener dependencias entre sí. Airflow te proporciona una interfaz fácil de usar para programar y monitorear tus flujos de trabajo, así como para gestionar el flujo de datos entre las tareas.

Figura 30.



3.5.2.2 Construcción de la HU003

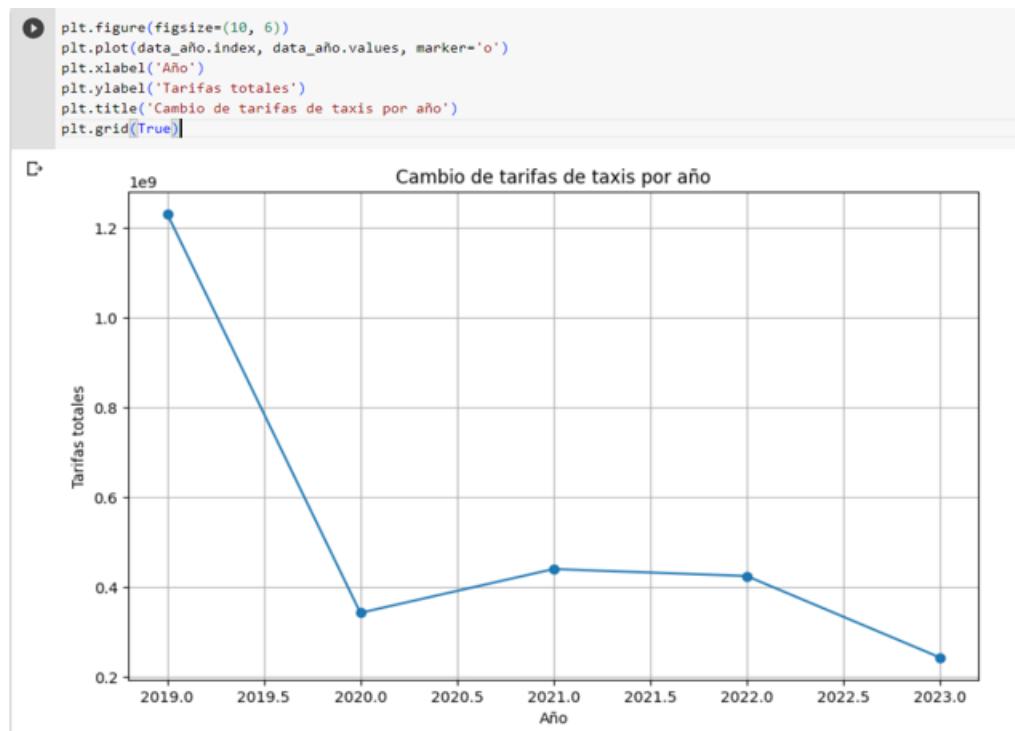
La Historia de Usuario HU003 se centra en el Análisis exploratorio de datos. A continuación, se hará un resumen de los pasos clave involucrados en esta tarea:

Análisis exploratorio de datos

- Taxis rutas

Figura 31.

Gráfico de viajes por año.

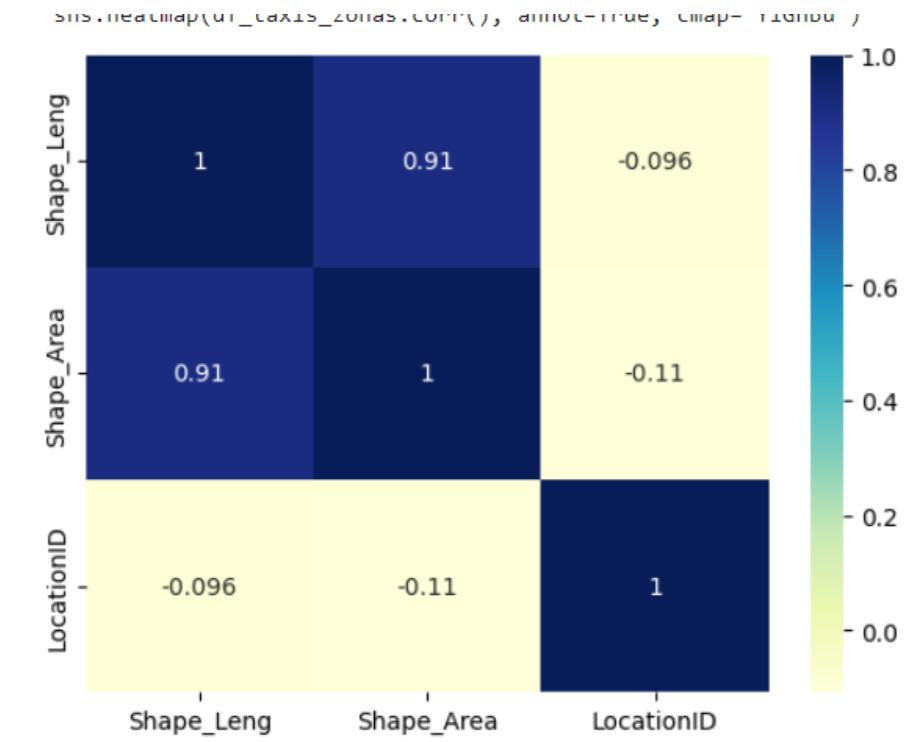


Nota: Realizamos este grafico con el fin de identificar las demandas que existen por años en la ciudad de Nueva York.

- Taxis zonas

Figura 32.

Gráfico heatmap de las correlaciones de los borough de New york.



Nota: Podemos observar el mapa de correlación con las variables: Shape_area: Es la forma del área de los distritos y zonas de new york city en millas cuadradas. Shape_leng: es la forma de la longitud de los distritos y zonas de new york en millas. location_id: es el identificador de cada zona de new york.

- **Combustion y CO2**

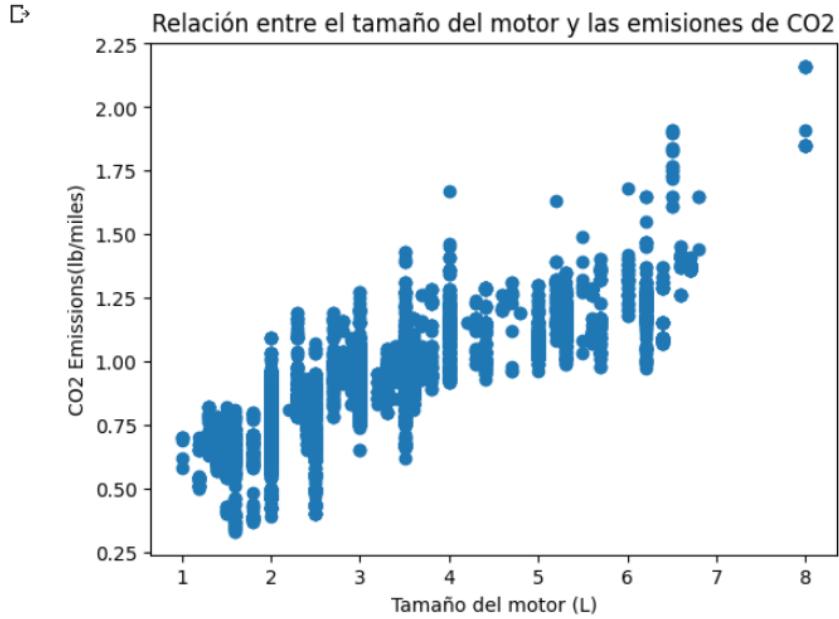
Figura 33.

Gráfico de dispersión de la tabla de Combustión y CO2.

```

plt.scatter(df_comb_CO2['Engine Size(L)'], df_comb_CO2['CO2 Emissions(lb/miles)'])
plt.xlabel('Tamaño del motor (L)')
plt.ylabel('CO2 Emissions(lb/miles)')
plt.title('Relación entre el tamaño del motor y las emisiones de CO2')
plt.show()

```

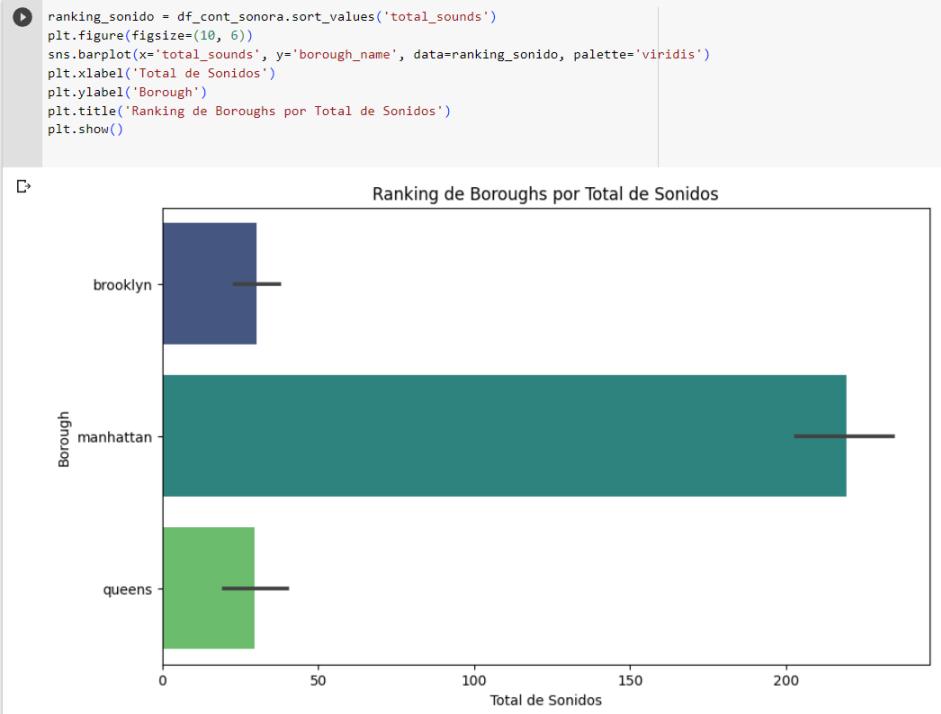


Nota: Gráfico de dispersión de la tabla de Combustión y CO2. Este gráfico de dispersión nos permite comprender la relación entre el tamaño del motor y las emisiones de CO2. Nos muestra claramente que los motores más grandes tienden a generar mayores emisiones de CO2. Sin embargo, también nos revelan que la concentración de emisiones se encuentra en los motores de menor tamaño debido a una mayor cantidad de vehículos equipados con ellos.

- **Contaminación sonora**

Figura 34.

Gráfico de barras sobre los borough de New York.

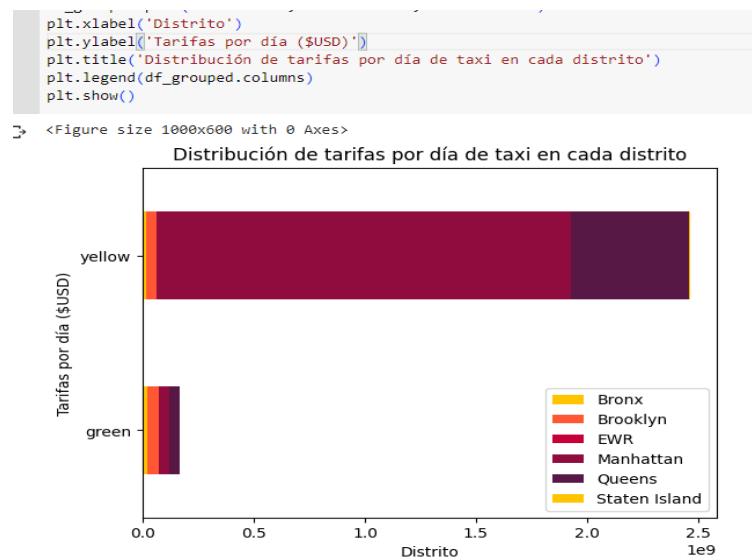


Nota: Graficó de barras horizontales que muestra el nivel de contaminación sonora por cada distrito de New York.

- **New Taxis**

Figura 35.

Gráfico de barras de las tarifas diarias por distrito.

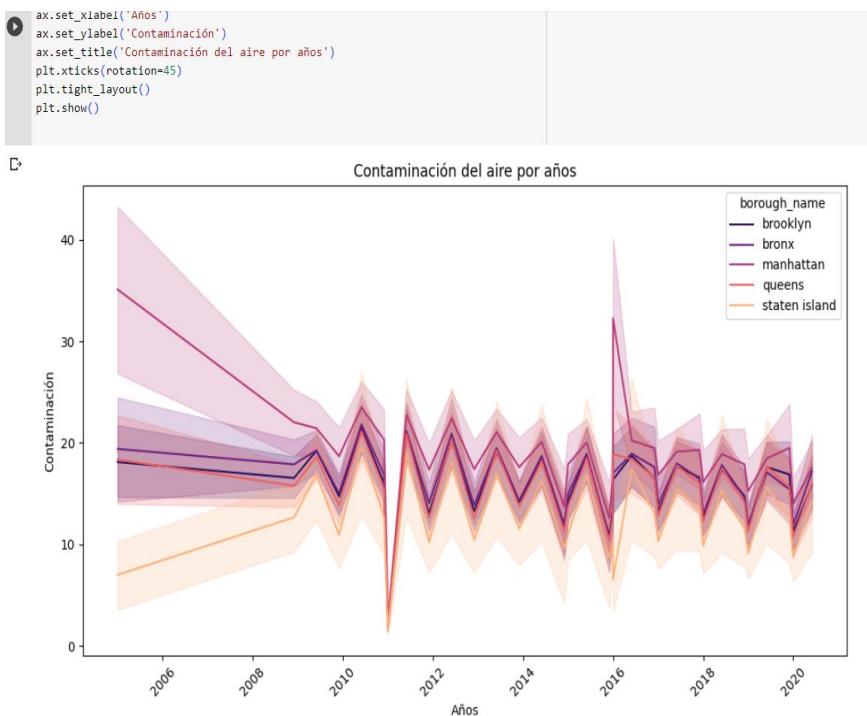


Nota: En este grafico de barras horizontales se puede diferenciar la tarifa diaria entre los taxis amarillos y verdes categorizados por distrito.

- **Calidad del Aire**

Figura 36.

Gráfico de Líneas de la contaminación en el aire.



Nota: En este grafico podemos ver la contaminación que hay en el aire por distrito de New york.

Segmentación y agrupación

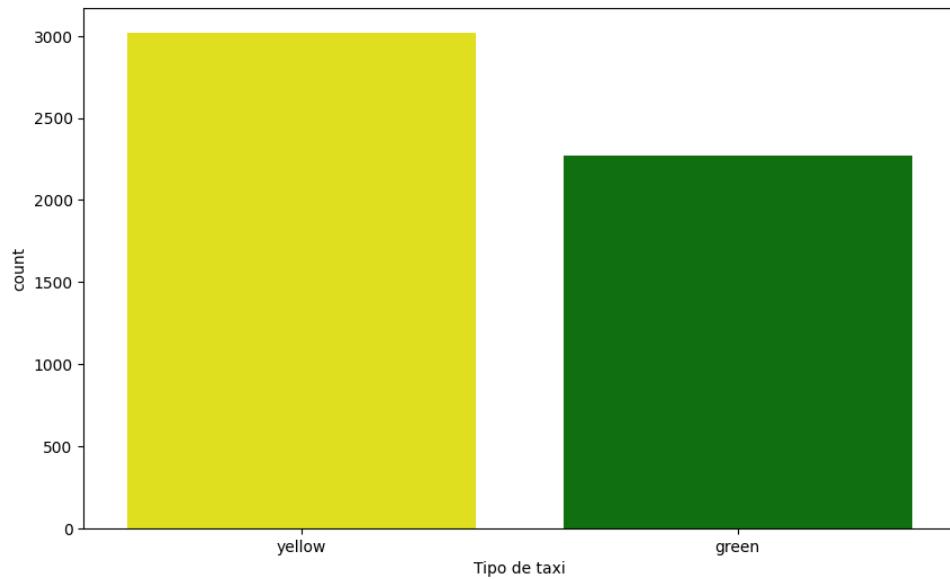
La segmentación consta de importar los datos y examinar su estructura, utilizando técnicas como clustering o segmentación basada en reglas para agrupar los datos en diferentes segmentos.

Analizar los segmentos: Examinar las características y comportamientos de cada segmento para comprender mejor las diferencias y similitudes entre ellos.

Figura 37.

Gráfica de barras sobre los tipos de taxi en New York.

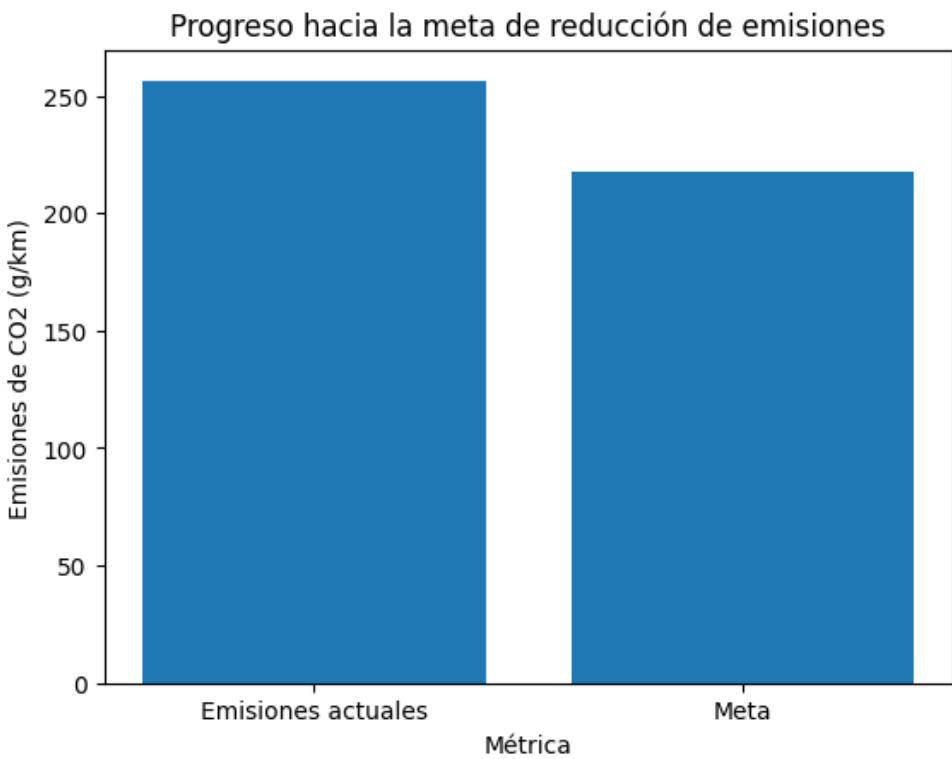
```
##Exploramos la relación entre variables categóricas utilizando gráficos de barras:  
plt.figure(figsize=(10, 6))  
sns.countplot(x='Tipo de taxi', data=df_NEW_TAXIS, palette=['yellow', 'green'])  
plt.show()
```



Nota: En este gráfico sirve para ver la diferencia en registros de taxis amarillos con respecto a taxis verdes.

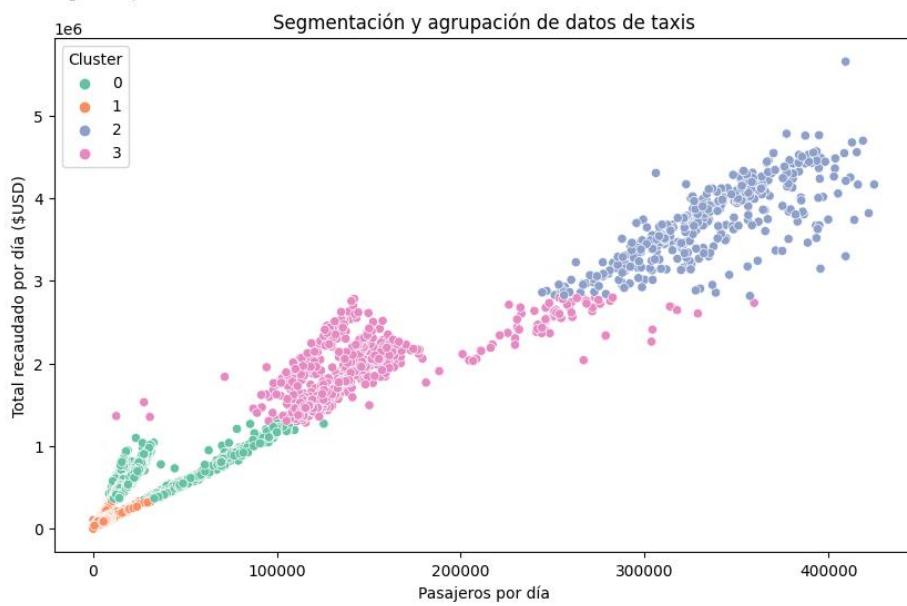
Figura 38.

Grafica de barras sobre las emisiones de CO2



Nota: Con esta grafica podemos proyectarnos a como serían las emisiones después de nuestra implementación.

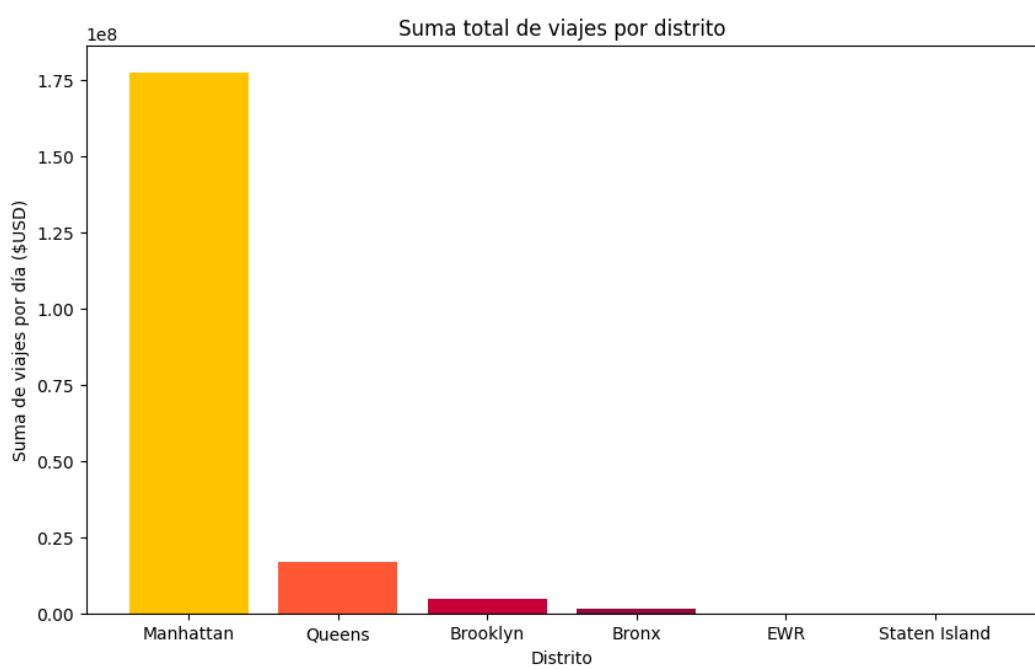
Figura 39.
Grafica de distribución por día de taxi por distrito.



Nota: En esta gráfica se presentará un análisis visual con el objetivo de identificar los tipos de vehículos que emiten mayores cantidades de CO2. Este análisis permitirá entender mejor la problemática ambiental relacionada al uso de automóviles.

Figura 40.

Grafica de barras

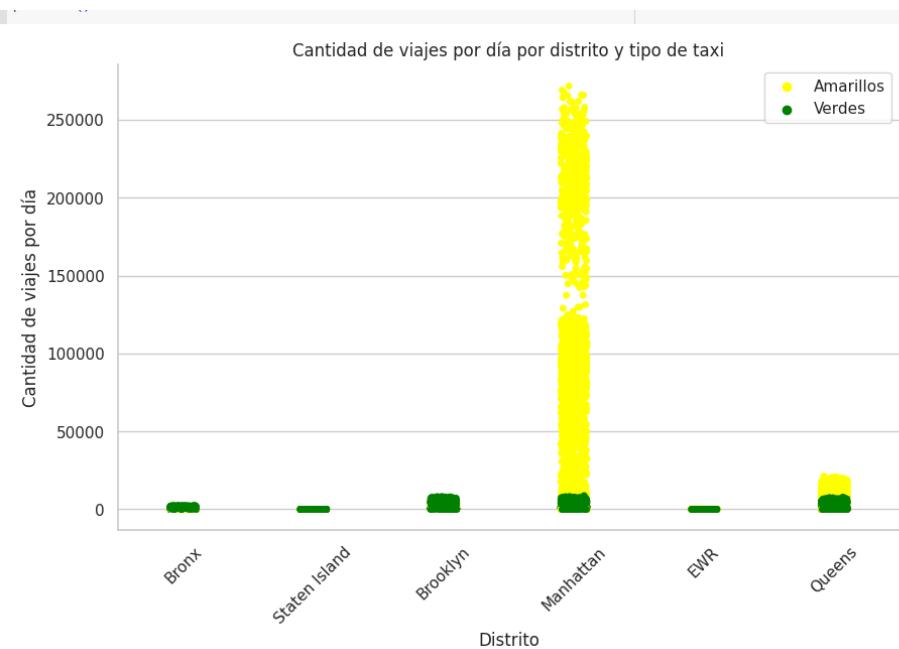


Identificación de insights

Este proceso consta de analizar relaciones y patrones realizando un análisis exploratorio más profundo para identificar relaciones y patrones interesantes entre variables, Identificando valores atípicos para buscar excepciones que puedan proporcionar información relevante o insights únicos, realizando análisis comparativo en diferentes grupos o segmentos para identificar valores significativos y obtener insights adicionales, como por ejemplo la investigación de features como CO2, contaminación sonora, combustión y CO2.

Figura 41.

Gráfico de los viajes por Borough.



Nota: En este grafico usamos stripplot para ver si nuestros datos se sobrelapan o si hay algún cambio, la información más importante de aquí es que la cantidad de taxis amarillos es significativamente superior en Manhattan.

Figura 42.

Gráfico de barras sobre la cantidad de viajes por año.



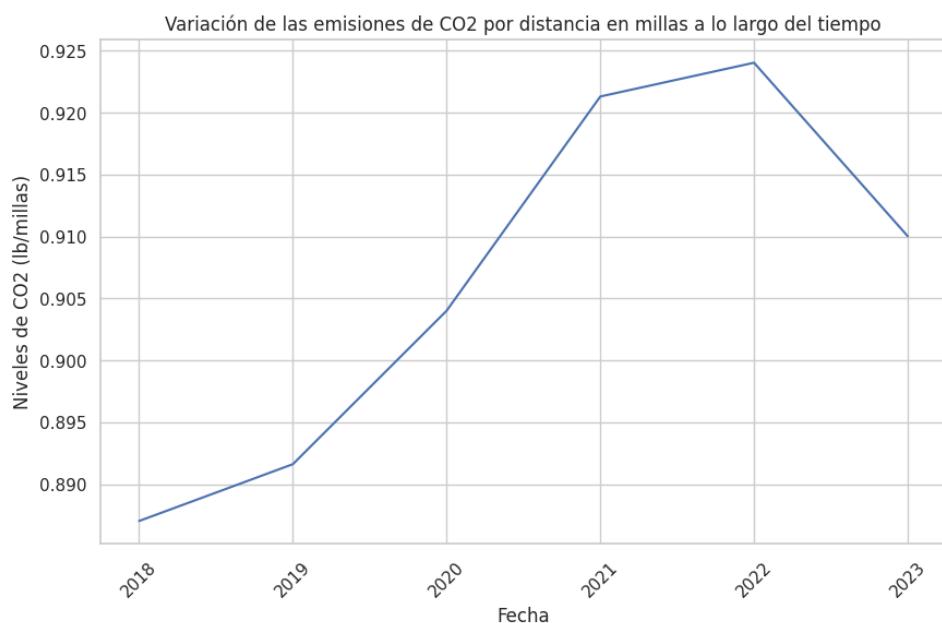
Nota: Este gráfico nos da la información sobre que taxi tuvo más viajes y en qué año, por ejemplo, antes de pandemia se nota una gran demanda de taxis.

Análisis estadístico

Los valores estadísticos se identificaron para las variables numéricas validando la media, mediana y moda.

Figura 43.

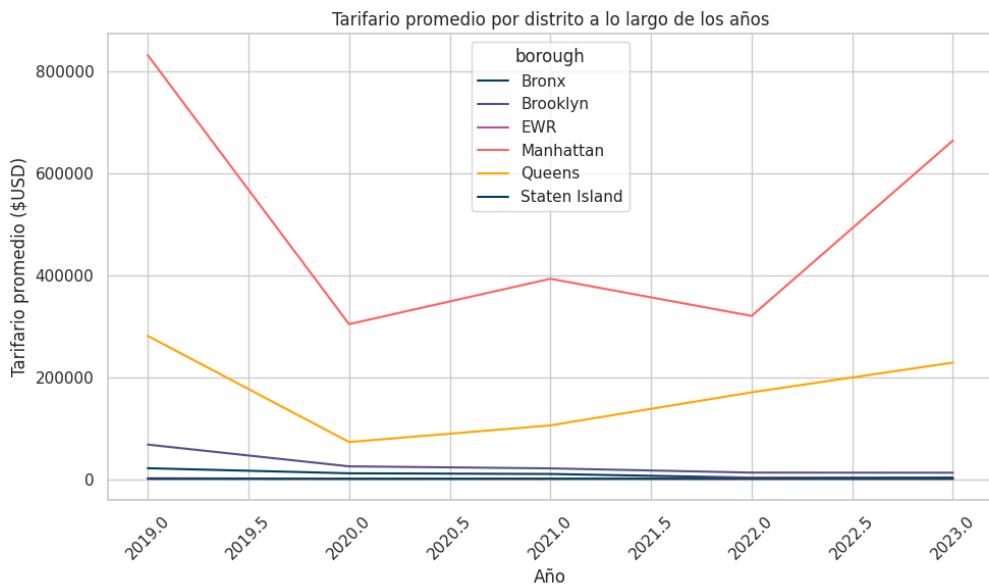
Gráfico de línea sobre el promedio de emisiones de CO2 por año.



Nota: Esta gráfica nos entrega el año con el mayor promedio de emisión de CO2, que sería el 2022 seguido del 2021.

Figura 44.

Gráfico de línea sobre el promedio de tarifa por distrito al año.



Nota: Este gráfico nos brinda una valiosa información que respalda la idea de invertir o investigar más en el distrito de Manhattan. Podemos observar claramente que Manhattan tiene la mayor tarifa promedio de todos los distritos, y lo más interesante es que esta tendencia se ha mantenido incluso después de la pandemia. Esto sugiere que el mercado de taxis en Manhattan es sólido y presenta oportunidades potenciales para aquellos interesados en invertir en este sector o llevar a cabo investigaciones más detalladas.

3.5.3 Daily Meeting

Gestión del tablero en el Sprint

- La frecuencia de las reuniones será de cinco veces por semana, de lunes a viernes.
- Cada reunión tendrá una duración máxima de 45 minutos.
- Durante la reunión se realizará la revisión de los avances presentes y futuros por cada historia de usuario.

Revisión del progreso. Se han realizado todas las historias de usuario con un progreso concluido del 100% tal como se especifica:

Tabla 10.

Sprint 02 – Revisión del progreso.

Historia de Usuario	Progreso	Estado
HU002	100%	Terminado

HU003	100%	Terminado
-------	------	-----------

Identificación de impedimentos, bloqueos, dependencias y riesgos. Dentro de estas, se encuentran los siguientes impedimentos:

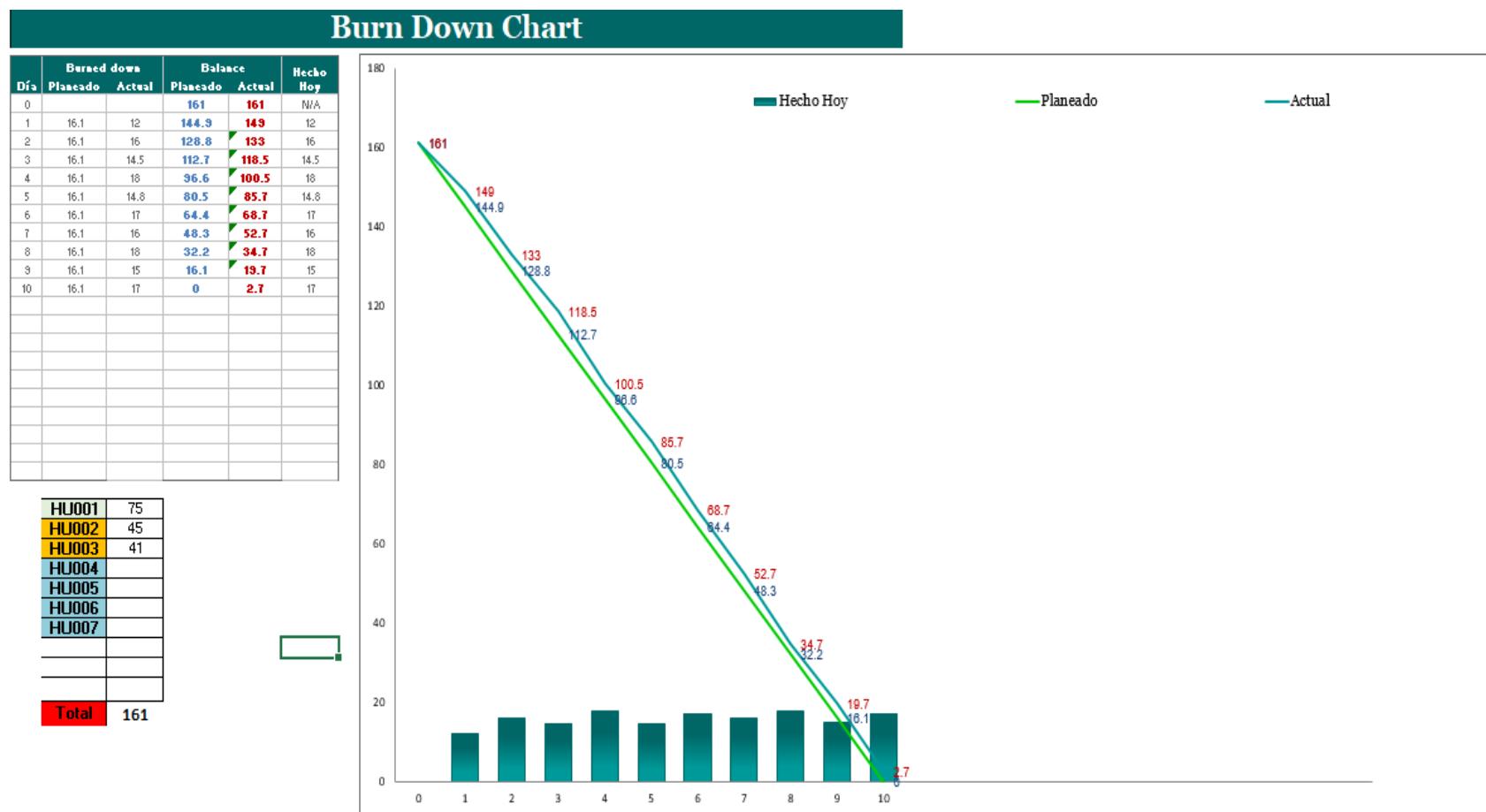
- Zonas horarias diferentes entre los desarrolladores del proyecto.

Sprint Burndown Chart

El gráfico de Burndown Chart se ingresa diariamente, después de cada dayli meeting y está realizado por el Scrum Master del proyecto con la finalidad de representar el esfuerzo esperado versus el real para el Sprint 2. Para poder obtener el porcentaje de avance diario, se toman los puntos de historia definidos en el Product Backlog, y se establece el esfuerzo diario obtenido.

Figura 45.

Sprint Burdonwn Chart -Sprint 02

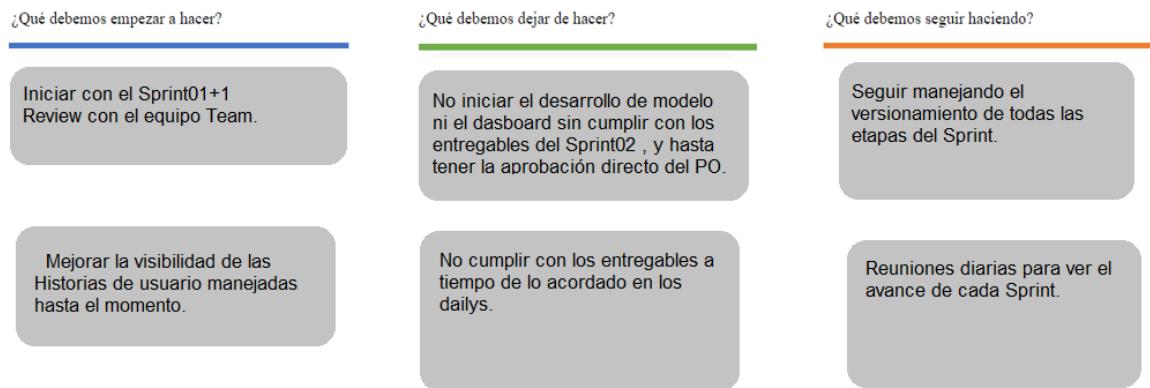


3.5.4 Sprint Retrospective

Dentro de los planes de mejora contemplados en la reunión realizada el día 07/07/2023, se obtuvieron los puntos detallados.

Figura 46.

Sprint Retrospective 02



3.6 Desarrollo de Sprint 3

3.6.1 Sprint Planning 3

Definition of Done (DoD)

A fin de garantizar la calidad del incremento y de cumplir con las condiciones requeridas por el producto, se debe cumplir lo siguiente:

- Todas las pruebas unitarias y funcionales deben ser correctas y validadas por los desarrolladores.

- El código debe estar completo de acuerdo con los estándares de desarrollo del equipo.
- El código de desarrollo del aplicativo debe estar versionado y en GitHub.
- El despliegue del proyecto debe estar en un entorno Dev.
- Todos los criterios de aceptación deben cumplirse.
- Todos los bugs deben estar corregidos.
- Los módulos dentro del alcance del sprint 2 deben ser aceptados por el Product Owner.

Historias de usuario para el Sprint 3

Las historias de usuario pertenecientes al Sprint 03, están compuesta por HU004 (Modelo de predicción), HU005(Desarrollo de dashboard), HU006(Validación de dashboard) y HU007(Despliegue en la nube), dentro de las cuales, se detallará por cada uno de los puntos de historia, criterios de aceptación y tareas técnicas correspondientes.

Figura 47.

Historia de usuario HU004 -Épica03-Sprint03

DOING	21 TAREAS	PERSONA ASIGNADA	FECHA LÍMITE	PRIORIDAD
Product Backlog Base	Diseño y entrenamiento del modelo = epica03 hu004 sprint03	IA	vie.	!
Product Backlog Base	Validación del modelo = epica03 hu004 sprint03	IA	vie.	!
Product Backlog Base	Modelo de predicción = epica03 hu004 sprint03	IA	vie.	!
Product Backlog Base	Evaluación y ajuste del modelo = epica03 hu004 sprint03	IA	vie.	!

Figura 48.

Figura 49.

Historia de usuario HU005 -Épica04-Sprint03

DOING	21 TAREAS	PERSONA ASIGNADA	FECHA LÍMITE	PRIORIDAD
Product Backlog Base	Diseño de la estructura y la interfaz del dashboard = epica04 hu005 sprint03	JC	vie.	!
Product Backlog Base	Desarrollo de Dashboard epica04 hu005 sprint03	JC	vie.	!
Product Backlog Base	Configuración de interacciones y filtros: epica04 hu005 sprint03	JC	vie.	!

Figura 50.

Historia de usuario HU006-Épica04-Sprint03

DOING	21 TAREAS	PERSONA ASIGNADA	FECHA LÍMITE	PRIORIDAD
Product backlog base	Verificación de datos epica04 hu006 sprint03	JC	vie.	!
Product Backlog Base	Validación de Dashboard epica04 hu006 sprint03	JC	vie.	!
Product Backlog Base	Prueba de funcionalidad epica04 hu006 sprint03	JC	vie.	!
Product Backlog Base	Retroalimentación del usuario epica04 hu006 sprint03	O	vie.	!
Product Backlog Base	Aprobación final epica04 hu006 sprint03	O	vie.	!

Figura 51.

Historia de usuario HU007-Épica05-Sprint03

DOING	21 TAREAS	PERSONA ASIGNADA	FECHA LÍMITE	PRIORIDAD
Product Backlog Base	Aprobación final epica04 hu006 sprint03	O	vie.	!
Product Backlog Base	Monitoreo y gestión epica05 hu007 sprint03	O	vie.	!
Product Backlog Base	Configuración de seguridad epica05 hu007 sprint03	O	vie.	!
Product Backlog Base	Configuración de la infraestructura en la nube epica05 hu007 sprint03	O	vie.	!
Product Backlog Base	Despliegue en la nube epica05 hu007 sprint03	O	vie.	!

Resumen del Sprint Planning 3

La duración del Sprint 3 tiene una duración total de 1 semana.

Tabla 11.

Resumen sprint planning 3

<i>Fecha Inicio del Sprint</i>	<i>Fecha Fin del Sprint</i>	<i>Días</i>
10/07/2023	14/07/2023	05

Sprint Backlog 3

Dentro del Sprint Backlog se detallan las tareas técnicas específicas para cada historia de usuario.

Tablero Scrum

El avance de la historia de usuario especificadas para el Sprint 03, se encuentra concluidas en fecha y presenta el detalle.

3.5.2 Construcción del Sprint 3

Diseño de prototipos

Los prototipos pertenecientes a cada historia de usuario, es decir, HU004, HU005, HU006 y HU007 se detallan como prototipos realizados y con componentes actualizados. Aquí se muestra un diagrama de flujo o diagrama de proceso que detalla las etapas de integración y análisis exploratorio de datos y las decisiones que se toman en cada etapa.

Figura 52.

Mockup de la HU004

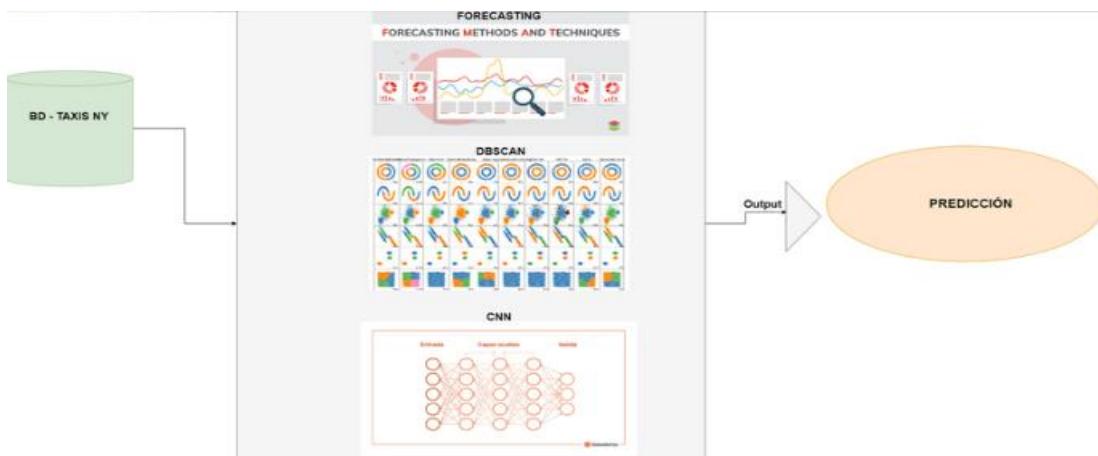


Figura 53.

Mockup de la HU005

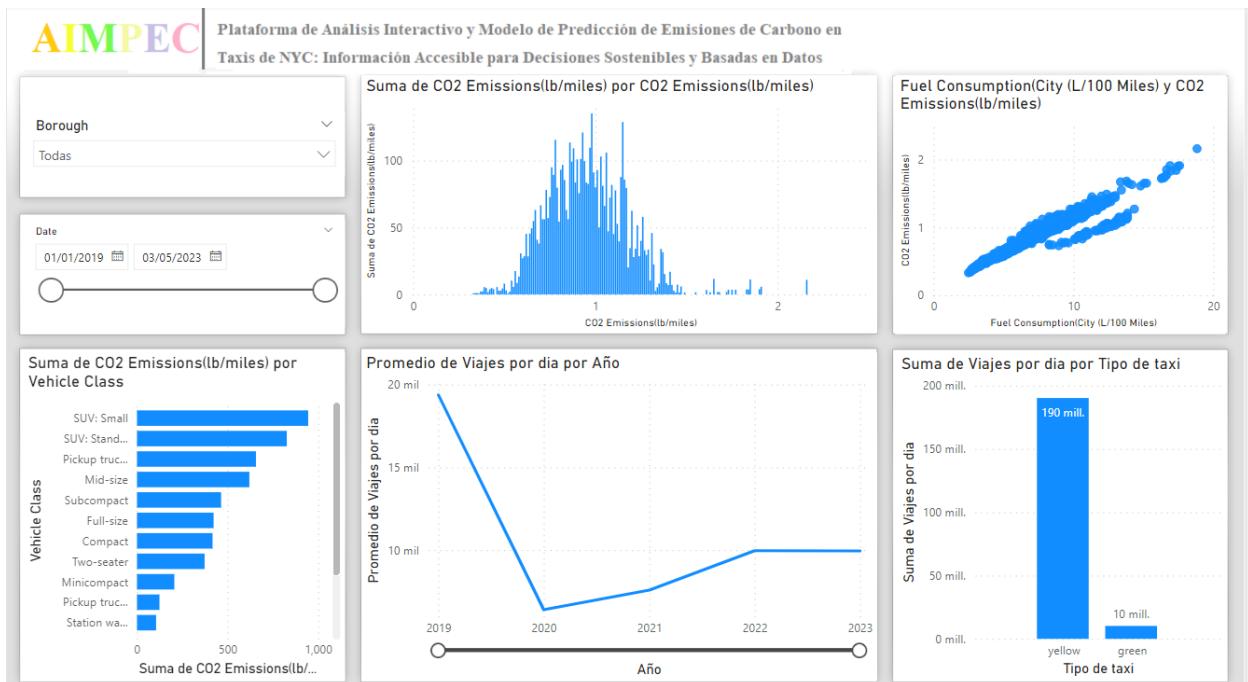


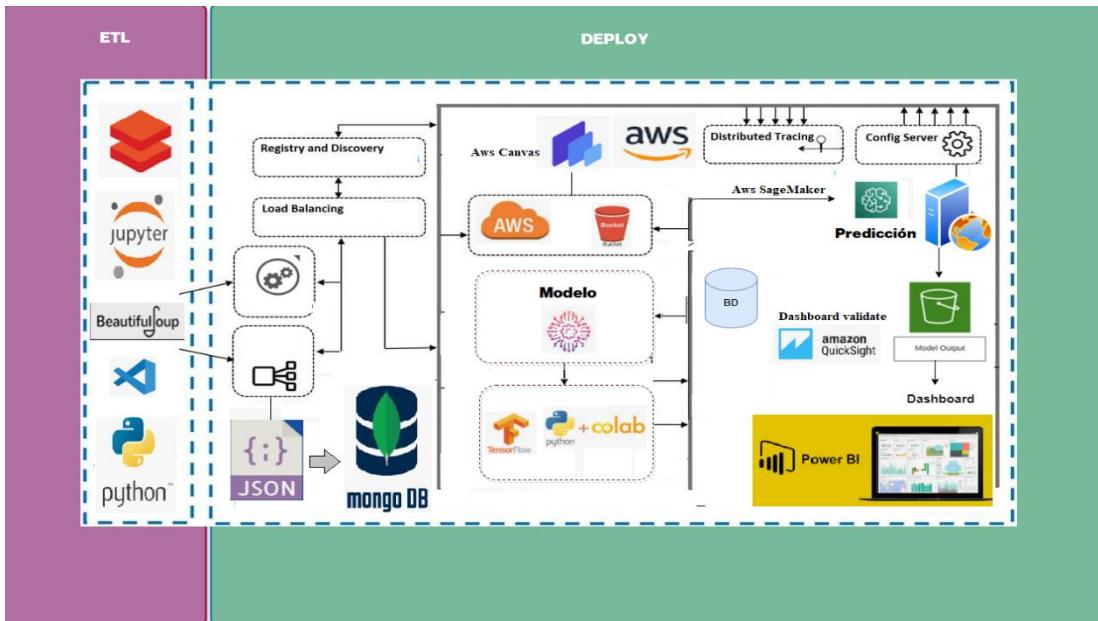
Figura 54.

Mockup de la HU006



Figura 55.

Mockup de la HU007

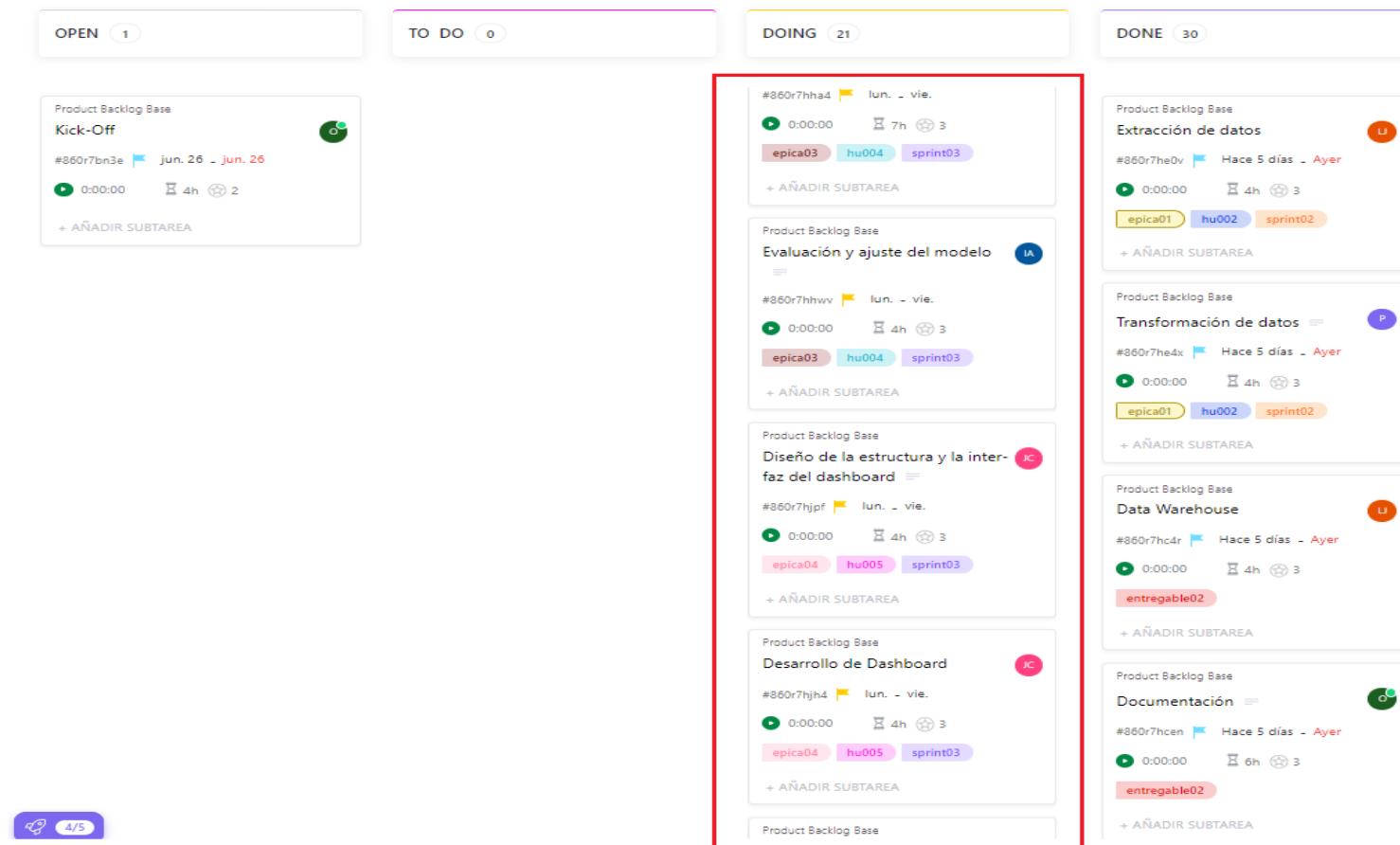


Product Backlog Refinado

El Product Backlog Refinado del Sprint 03 proporciona una visión clara de las tareas a realizar y establece una base sólida para el inicio del desarrollo durante el segundo sprint.

Figura 56.

Product backlog -Sprint03



3.5.2.1 Construcción de la HU004

La Historia de Usuario HU004 se centra en la construcción del modelo de predicción. A continuación, se hará un resumen de los pasos clave involucrados en esta tarea:

Evaluación y ajustes del modelo

Random Forest Regression: El modelo Random Forest es un algoritmo de aprendizaje automático que combina múltiples árboles de decisión para realizar predicciones. Cada árbol en el bosque se entrena con una muestra aleatoria del conjunto de datos y se utiliza para realizar una predicción. Luego, se toma la predicción promedio de todos los árboles para obtener el resultado final. Esto proporciona una mayor estabilidad y precisión en las predicciones, ya que cada árbol puede compensar las debilidades de los demás.

CNN (Convolutional Neural Network): Las redes neuronales convolucionales (CNN) son un tipo de modelo de aprendizaje profundo especialmente diseñado para procesar datos en forma de matrices, como imágenes. Utilizan capas convolucionales que aplican filtros a través de la imagen para extraer características relevantes y capas de agrupación para reducir la dimensionalidad. Las CNN son muy efectivas en tareas de reconocimiento de patrones y clasificación de imágenes debido a su capacidad para capturar características locales y aprender jerarquías de características.

Regresión lineal: La regresión lineal es un modelo estadístico utilizado para predecir un valor numérico continuo en función de una o más variables predictoras. Se basa en la relación lineal entre las variables y busca encontrar la mejor línea recta que se ajuste a los datos. El modelo calcula los coeficientes de la línea recta mediante el método de mínimos cuadrados y utiliza estos coeficientes para realizar predicciones. La regresión lineal es ampliamente utilizada debido a su simplicidad y facilidad de interpretación, pero asume una relación lineal entre las variables, lo que puede ser limitante en algunos casos.

Mean Squared Error (MSE): Es una métrica utilizada para evaluar la calidad de un modelo de regresión. Calcula la media de los errores al cuadrado entre las predicciones del modelo y los valores reales. Cuanto menor sea el valor del MSE, mejor será el ajuste del modelo a los datos. El MSE penaliza los errores grandes de manera más significativa que los errores pequeños, lo que lo hace útil para identificar discrepancias significativas entre las predicciones y los valores reales.

Coefficient of Determination (R^2): También conocido como coeficiente de determinación, es una medida que indica la proporción de la varianza de la variable dependiente que puede ser explicada por el modelo. R^2 varía entre 0 y 1, donde 0 indica que el modelo no explica nada de la variabilidad de los datos y 1 indica que el modelo explica perfectamente la variabilidad. Un valor de R^2 cercano a 1 indica un buen ajuste del modelo, mientras que un valor cercano a 0 indica un ajuste deficiente. R^2 se interpreta como el porcentaje de la varianza total explicada por el modelo.

Elección del modelo

Figura 57.

Model elección

MODELO	R^2
Random Forest	99.995%
Regresion Lineal	89.428%
CNN	63.043%

Conforme a lo descrito en el anterior capítulo, luego de realizar la evaluación de cada uno de los modelos, el modelo seleccionado que mejor desempeño tiene es el modelo de **Random Forest**, el cual será el que se aplicará para el modelo de predicción.

Diseño y Entrenamiento del modelo

Para el diseño y entrenamiento del modelo de Random Forest, se siguieron los siguientes pasos:

Preprocesamiento de datos: Se realizó una limpieza y transformación de los datos. Se eliminaron las columnas no relevantes y se aplicaron técnicas de codificación para variables categóricas.

División de datos: El conjunto de datos se dividió en conjuntos de entrenamiento y prueba utilizando la función `train_test_split` de la biblioteca scikit-learn. Se utilizó un tamaño de prueba del 20% y se fijó una semilla aleatoria para garantizar la reproducibilidad de los resultados.

Creación del modelo: Se configuraron los parámetros del modelo de Random Forest, como el número de árboles, la profundidad máxima de los árboles y el número mínimo de muestras requeridas para dividir un nodo. Estos parámetros pueden ajustarse para obtener un equilibrio entre el rendimiento del modelo y la capacidad de generalización.

Entrenamiento del modelo: El modelo se entrenó utilizando el conjunto de entrenamiento mediante el método `fit`. Durante el entrenamiento, los árboles de decisión individuales se ajustaron a diferentes subconjuntos aleatorios del conjunto de entrenamiento.

Evaluación del modelo: Una vez entrenado el modelo, se realizaron predicciones en el conjunto de prueba utilizando el método `predict`. Se calcularon dos métricas para evaluar el rendimiento del modelo: el Mean Squared Error (MSE) y el Coefficient of Determination (R^2).

Validación del modelo

Carga del modelo entrenado: Se cargó el modelo entrenado previamente, que se encuentra almacenado en el archivo 'Modelo_12072023.pkl'. Esto permite utilizar el modelo para realizar predicciones en los nuevos datos.

Realización de predicciones: Utilizando el modelo cargado, se realizaron predicciones en los nuevos datos. Esto se hizo utilizando el método `predict` del modelo de Random Forest. Se obtuvieron las predicciones correspondientes a la variable objetivo del problema, en este caso, las emisiones de CO₂.

Evaluación de las predicciones: Se evaluaron las predicciones realizadas por el modelo en los nuevos datos. Esto se hizo comparando las predicciones con los valores reales disponibles en los nuevos datos. Se utilizaron métricas adecuadas para evaluar el rendimiento del modelo en esta

validación específica. Estas métricas pueden incluir el Mean Squared Error (MSE) y el Coefficient of Determination (R^2), entre otros.

Resultados y análisis: Se analizaron los resultados obtenidos en la validación del modelo. Se evaluó la precisión y la capacidad de generalización del modelo en los nuevos datos. Se realizaron comparaciones con los resultados obtenidos en la evaluación anterior del modelo para obtener una perspectiva completa del rendimiento del modelo en diferentes conjuntos de datos.

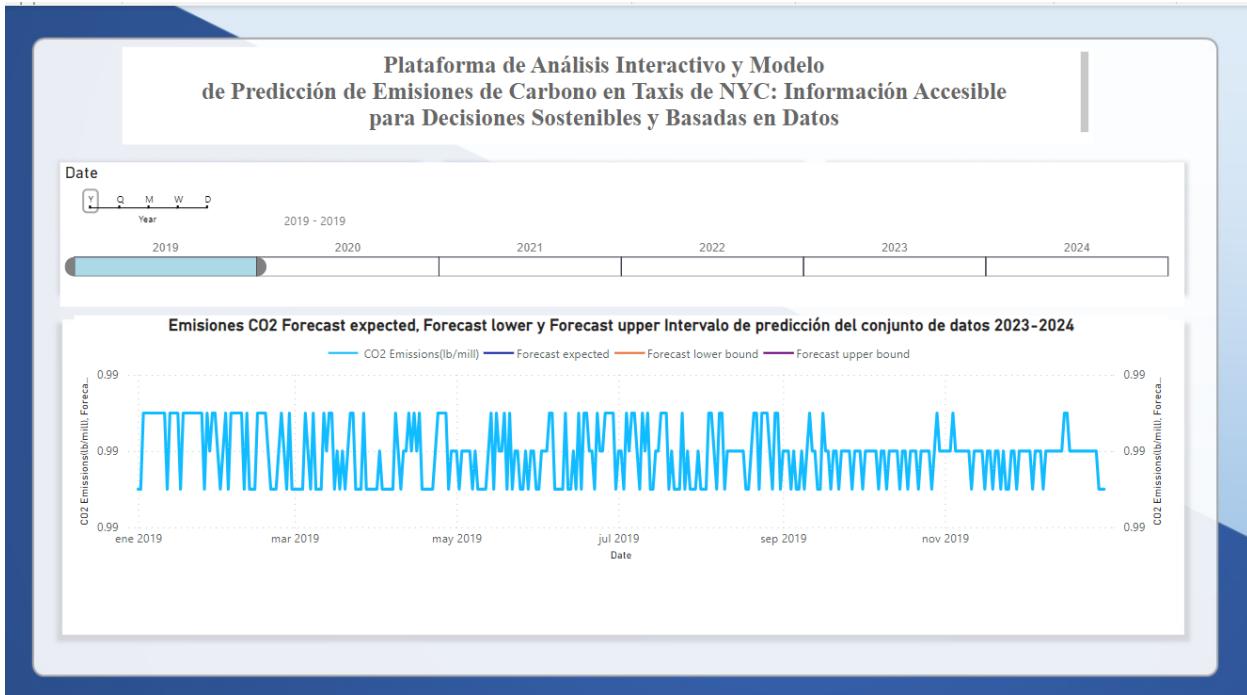
Modelo ML en producción

El Modelo de Aprendizaje Automático (ML) en producción en el proyecto “AIMPEC” ha pasado por varios pasos y herramientas. Aquí está una descripción detallada:

- **SageMaker:** El modelo ha pasado por dos filtros en SageMaker, lo que implica que se ha utilizado la plataforma de aprendizaje automático de Amazon para realizar tareas como entrenamiento, ajuste de parámetros y validación del modelo. Entrenamiento con Forescasting en SageMaker Canvas: Se extrajo SageMaker Canvas, una interfaz visual en SageMaker, para entrenar el modelo y realizar análisis exploratorio de datos. Este proceso permitió obtener métricas de rendimiento, como la precisión de predicción promedio y el impacto de diferentes columnas en las predicciones.
- **Almacenamiento en Bucket S3:** El conjunto de datos se guardó en un bucket de Amazon S3. S3 es un servicio de almacenamiento en la nube altamente escalable y duradero de Amazon Web Services (AWS).
- **Validación en QuickSight:** Para validar el modelo y realizar análisis adicional, se obtuvo QuickSight, una herramienta de visualización de datos en la nube de AWS. QuickSight permite crear paneles interactivos y realizar análisis exploratorio de datos.
- **Consumo en Power BI:** Finalmente, el modelo se consumió en Power BI, una plataforma de análisis y visualización de datos de Microsoft. Se mantendrá un enlace público que permite a los usuarios interactuar con los filtros y explorar los datos.

Puede acceder al enlace público() proporcionado para interactuar con los filtros y explorar los datos en el entorno de Power BI.

Figura 58.
ML producción



Nota.<https://app.powerbi.com/view?r=eyJrIjoiYWNmNzFiNTItMmI3NC00YTk2LWIwN2QtMWZhMjM3YTE5N2I3IiwidCI6ImRmODY3OWNkLWE4MGUtNDVkcOC05OWFjLWM4M2VkN2ZmOTVhMCJ9>

3.5.2.2 Construcción de la HU005

La Historia de Usuario HU005 se centra en el desarrollo de dashboard. A continuación, se hará un resumen de los pasos clave involucrados en esta tarea:

Diseño de la estructura y la interfaz del dashboard

Completar aquí Juliana

Figura 59.
Estructura del dashboard

Desarrollo de Dashboard

Completar aquí Juliana

Figura 60.

XXXXX

Configuración de interacciones y filtros

Completar aquí Juliana

Figura 61. XXX

3.5.2.3 Construcción de la HU006

La Historia de Usuario HU006 se centra en la validación del dashboard. A continuación, se hará un resumen de los pasos clave involucrados en esta tarea:

Verificación de Kpis

Completar aquí Julianasdsddsdsdssdsdsdsdsdsdsdasasaasassassaa

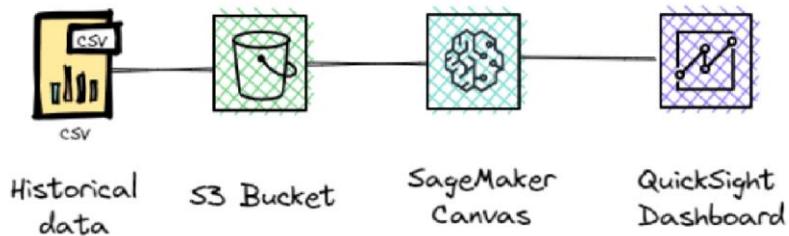


3.5.2.3 Construcción de la HU007

La Historia de Usuario HU007 se centra en el despliegue en la nube. A continuación, se hará un resumen de los pasos clave involucrados en esta tarea:

Figura 62.

Despliegue en la nube AWS



Configuración de la infraestructura en la nube

La Historia de Usuario HU007 se enfoca en el uso en la nube para el proyecto de taxis. Implica la configuración de la infraestructura en la nube, utilizando servicios como un bucket S3 para almacenar los datos, SageMaker Canvas para el entrenamiento y desarrollo del modelo, y QuickSight para la validación de los datos. Sin embargo, debido a la consideración de costos, el uso final del modelo se llevará a cabo en Power BI utilizando el modelo perturbado en SageMaker. Este enfoque permite aprovechar la escalabilidad y flexibilidad de la nube, optimizando los recursos y asegurando un uso eficiente y rentable del modelo de predicción de taxis.

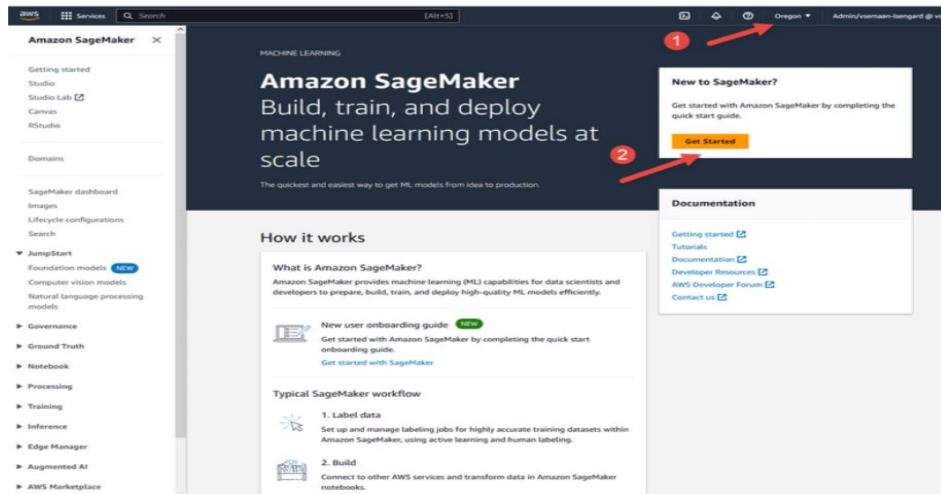
Parte 1: configuración de requisitos previos

Para comenzar con Amazon SageMaker Canvas, primero debemos crear un dominio. Se puede crear un dominio como un almacén central donde la configuración, los cuadernos y otros artefactos se almacenarán y compartirán entre los usuarios.

Para crear un dominio, abra la Consola de AWS y luego busque SageMaker. Seleccione una región que le gustaría usar. Haga clic en el botón Comenzar.

Figura 63.

Creación de un dominio en aws



Parte 2: Creación del almacenamiento S3.

Un depósito de Amazon S3 es un contenedor de almacenamiento proporcionado por Amazon Web Services (AWS) que permite a los usuarios almacenar y recuperar datos en la nube. Funciona como una carpeta o directorio donde puede cargar y organizar archivos, como documentos, imágenes o videos. Ofrece una solución de almacenamiento escalable y confiable, accesible desde cualquier lugar en Internet, y puede integrarse con varias aplicaciones y servicios para fines de almacenamiento y respaldo de datos. Se puede crear un depósito nuevo o utilizar cualquier depósito existente. Para crear un nuevo depósito de S3, en la Consolade aws.

Figura 64.

Dataset seleccionado para el despliegue en bucket S3

A	B	C	D	E	F	G	H	I	J	K	L	
1	Ticker	Date	Viajes por dia	Pasajeros por dia	Distancia(millas)	borough	Tipo de taxi	Model	CO2 Emissions(lb/millil)	Fuel Consumption(City (l/millil)	CO2_Trip_Lb	Fuel_Ltr_Lt
2	AIMPEC	01/01/2022	93	97	272.33 Bronx	yellow	2021	1.05	8.99	285.946	2448.247	
3	AIMPEC	01/01/2022	55593	84497	265662.88 Manhattan	yellow	2021	1.05	8.99	278946.024	2388309.29	
4	AIMPEC	01/01/2022	43	88	15.04 EWR	yellow	2021	1.05	8.99	15.792	135.21	
5	AIMPEC	02/01/2022	37	60	51.55 EWR	yellow	2021	1.05	8.99	54.128	463.434	
6	AIMPEC	03/01/2022	7	5	103.92 Staten Island	yellow	2021	1.05	8.99	109.116	934.241	
7	AIMPEC	03/01/2022	15	29	26.29 EWR	yellow	2021	1.05	8.99	27.604	236.347	
8	AIMPEC	05/01/2022	327	370	1460.56 Brooklyn	yellow	2021	1.05	8.99	1533.588	13130.434	
9	AIMPEC	06/01/2022	72592	97618	214861.78 Manhattan	yellow	2021	1.05	8.99	225604.869	1931607.4	
10	AIMPEC	06/01/2022	139	142	108906.9 Bronx	yellow	2021	1.05	8.99	114352.245	979073.031	
11	AIMPEC	08/01/2022	75744	109547	177263.6 Manhattan	yellow	2021	1.05	8.99	186126.78	1593599.76	
12	AIMPEC	08/01/2022	377	397	1653.54 Brooklyn	yellow	2021	1.05	8.99	1736.217	14865.325	
13	AIMPEC	09/01/2022	4	4	29.58 Staten Island	yellow	2021	1.05	8.99	31.059	265.924	
14	AIMPEC	09/01/2022	7083	9982	89763.31 Queens	yellow	2021	1.05	8.99	94251.476	806972.157	
15	AIMPEC	10/01/2022	6123	8418	73522.3 Queens	yellow	2021	1.05	8.99	77198.415	660965.477	
16	AIMPEC	11/01/2022	364	369	105281.82 Brooklyn	yellow	2021	1.05	8.99	110545.911	946483.562	
17	AIMPEC	11/01/2022	5087	7110	61570.16 Queens	yellow	2021	1.05	8.99	64648.668	553515.738	
18	AIMPEC	13/01/2022	5205	7361	63767.9 Queens	yellow	2021	1.05	8.99	66956.295	573273.421	
19	AIMPEC	14/01/2022	513	531	2201.61 Brooklyn	yellow	2021	1.05	8.99	2311.69	19792.474	
20	AIMPEC	15/01/2022	425	451	1836.21 Brooklyn	yellow	2021	1.05	8.99	1928.02	16507.528	
21	AIMPEC	17/01/2022	8099	11405	102013.66 Queens	yellow	2021	1.05	8.99	107114.343	917102.803	
22	AIMPEC	19/01/2022	5	5	96 Staten Island	yellow	2021	1.05	8.99	100.8	863.04	
23	AIMPEC	20/01/2022	156	139	768.59 Bronx	yellow	2021	1.05	8.99	807.02	6909.624	
24	AIMPEC	20/01/2022	8	13	0.41 EWR	yellow	2021	1.05	8.99	0.431	3.686	
25	AIMPEC	22/01/2022	111	101	577.55 Bronx	yellow	2021	1.05	8.99	606.428	5192.174	
26	AIMPEC	23/01/2022	AA7	AA9	1886.78 Brooklyn	yellow	AA1	AA5	AA9	1086.561	14047.657	

Figura 65.

Bucket S3 AIMPEC_2023

Amazon S3 > Buckets > myprueba23

myprueba23 [Información](#)

Objetos | Propiedades | Permisos | Métricas | Administración | Puntos de acceso

Objetos (1)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a s que concederes permisos de forma explícita. [Más información](#)

[Cargar](#)

Buscar objetos por prefijo

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
AIMPEC_2023.csv	CSV	12 Jul 2023 11:07:42 PM -04	1.4 MB	Estándar

Parte 3: Creación de predicciones con SageMaker Canvas.

En el menú de la izquierda, se crea un modelo y luego en el botón Nuevo. Proporcione un nombre, por ejemplo, AIMPEC Predictions, y haga clic en el botón crear. En la siguiente pantalla, seleccione su conjunto de datos y haga clic en el botón Seleccionar conjunto de datos en la parte inferior.

Figura 66.

Creación del modelo Sagemaker

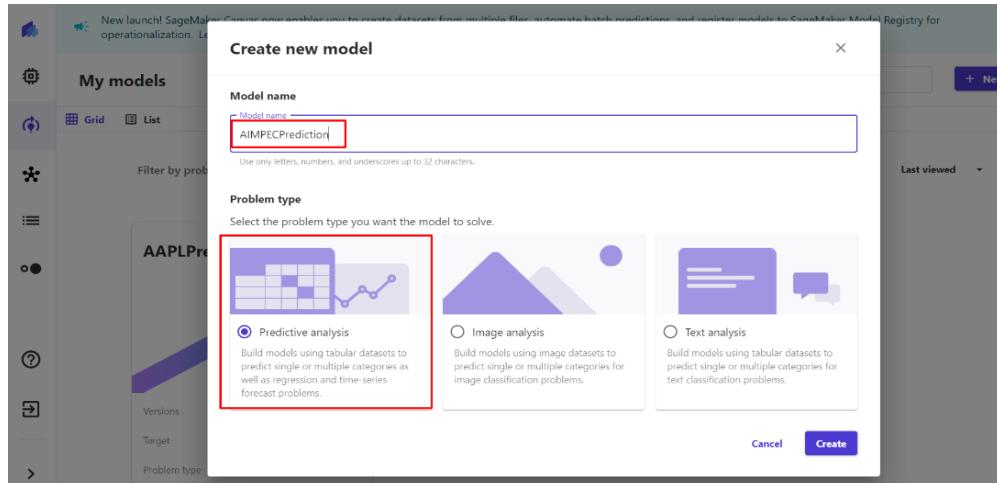
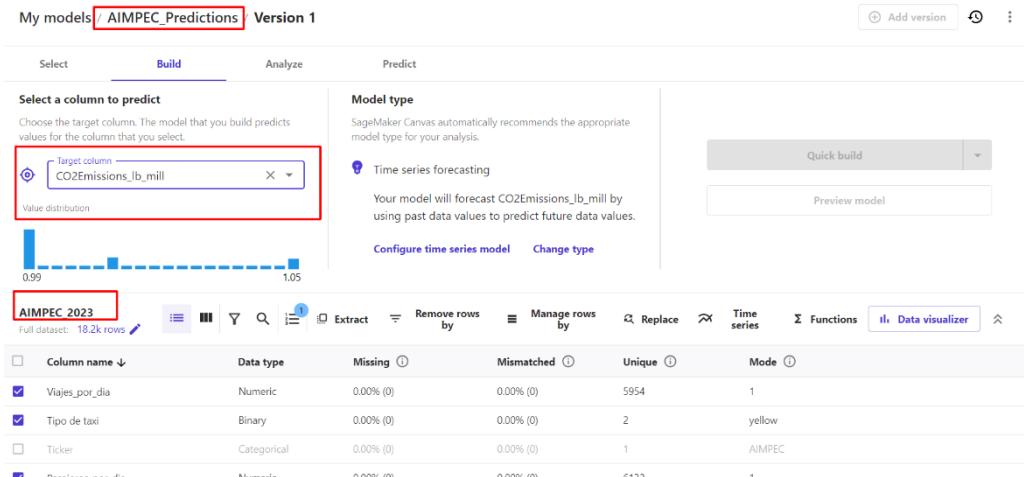


Figura 67.

Generando el target en el modelo



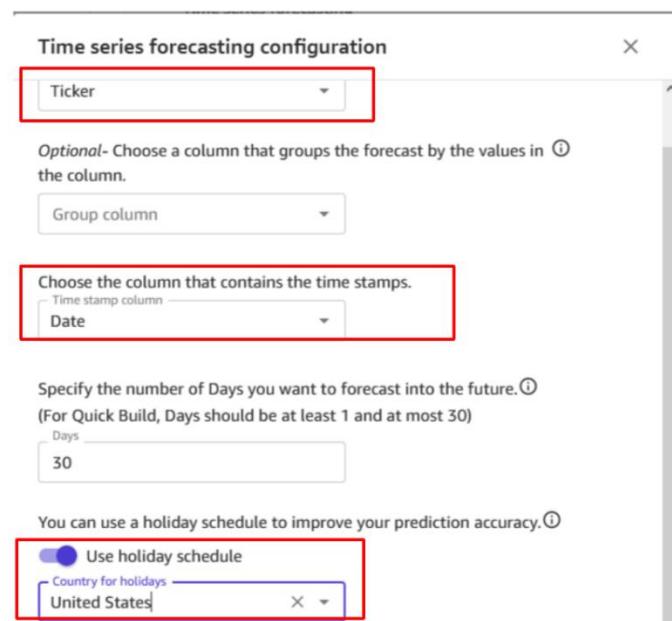
Haga clic en Configurar modelo de serie temporal y complete la configuración en la ventana emergente de la siguiente manera:

- . La columna que identifica de forma única los elementos en el conjunto de datos: Ticker.
- . La columna que contiene las marcas de tiempo: Date.
- . Especifique el número de días para el pronóstico: 30.

Usar programación de días festivos: habilite y elija Estados Unidos. El mercado de valores NYC, está cerrado durante las vacaciones de EE. UU.

Figura 68. Configuración de las columnas en forecasting

Haga clic en el botón **Guardar** en la parte inferior.



Parte 4: uso del modelo para generar predicciones.

Después de finalizar el entrenamiento del modelo de predicción de emisiones de CO₂ en el proyecto de taxis, se accede a la pestaña "Analizar". Allí se puede verificar la precisión promedio de las predicciones y examinar cómo diferentes columnas se derivan de los resultados. Es importante tener en cuenta que los valores reales pueden diferir ligeramente de los mostrados en la captura de pantalla debido a la naturaleza estocástica del proceso de aprendizaje automático, que implica cierta incertidumbre y aleatoriedad en los algoritmos de ML.

En SageMaker Canvas, el conjunto de datos se divide en conjuntos de entrenamiento y prueba. El conjunto de entrenamiento se utiliza para construir el modelo, mientras que el conjunto de prueba se utiliza para evaluar el rendimiento del modelo con nuevos datos. En la pestaña de evaluación del rendimiento, se puede observar cómo se desempeñó el modelo en el conjunto de prueba. Para

obtener más detalles y métricas sobre el rendimiento del modelo, se puede consultar la documentación de evaluación del rendimiento de los modelos en Amazon SageMaker Canvas.

Model status: Indica el estado del modelo alterado en SageMaker. Puede tener diferentes valores como "Completed", "Failed", "InProgress", entre otros, que indican el estado actual del modelo.

- **Error porcentual absoluto medio:** Es una métrica que calcula el error promedio en términos de porcentaje entre los valores reales y las predicciones del modelo. Un valor de 0,042 indica que, en promedio, el modelo tiene un error absoluto del 4,2% en las predicciones.
- **Error porcentual absoluto ponderado:** Es similar al error porcentual absoluto medio, pero tiene en cuenta el peso de cada muestra en el conjunto de datos. Un valor de 0.993 indica que, al considerar el peso de cada muestra, el error absoluto ponderado promedio es del 99.3%.
- **Root Mean Square Error:** Es una métrica que calcula la raíz cuadrada del error cuadrático medio entre los valores reales y las predicciones. Un valor de 2,54 indica que, en promedio, el modelo tiene un error de 2,54 en las predicciones.
- **Mean Absolute Scaled Error:** Es una métrica que calcula el error absoluto promedio ajustado en función de la escala de los datos. Proporcionar una medida de la precisión del modelo en relación con la obtención de los datos. Un valor de 1.322 indica que el modelo tiene un error absoluto promedio escalado de 1.322.
- **Promedio Weighted Quantile Loss:** Es una métrica que calcula la pérdida promedio ponderada por cuantil. Proporciona una medida de la precisión del modelo en la estimación de cuantiles específicos. Un valor de 0,87 indica que, en promedio, el modelo tiene una pérdida ponderada promedio de 0,87 en la estimación de los cuantiles.

Estas métricas son útiles para evaluar el rendimiento del modelo y determinar su precisión en las predicciones de las emisiones de CO2.

Figura 69.

Valores de la predicción

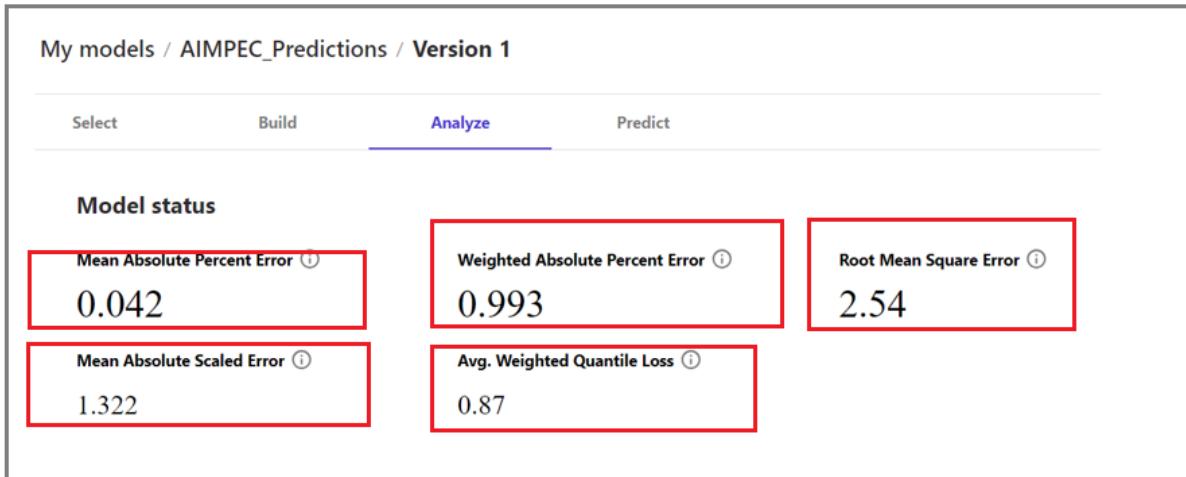
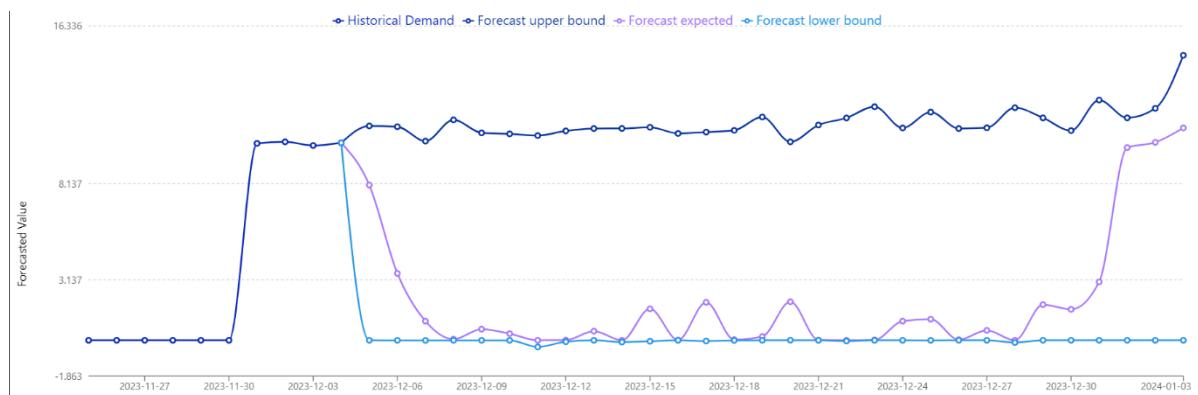


Figura 70. Intervalo de predicción del conjunto de datos actual: 2023-12-04 al 2024-01-03

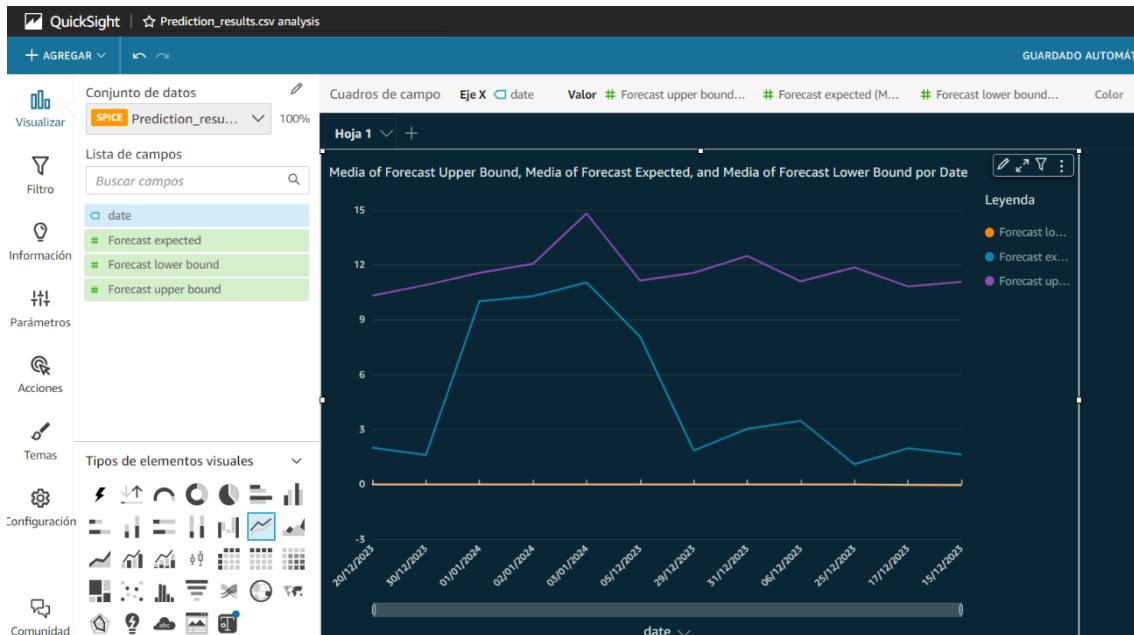


Parte 5: creación de un panel de visualización con QuickSight para su validación

Amazon Quicksight le permite crear paneles de visualización y admite diferentes tipos de visualización. Puede agregar campos calculados, aplicar filtros y cambiar campos y tipos de datos.

Figura 71.

Visualización en aws Quicksight



3.5.3 Daily Meeting

Gestión del tablero en el Sprint

- La frecuencia de las reuniones será de cinco veces por semana, de lunes a viernes.
- Cada reunión tendrá una duración máxima de 45 minutos.
- Durante la reunión se realizará la revisión de los avances presentes y futuros por cada historia de usuario.

Revisión del progreso. Se han realizado todas las historias de usuario con un progreso concluido del 100% tal como se especifica:

Tabla 12.

Sprint 02 – Revisión del progreso.

Historia de Usuario	Progreso	Estado
HU004	100%	Terminado
HU005	100%	Terminado
HU006	100%	Terminado
HU007	100%	Terminado

Identificación de impedimentos, bloqueos, dependencias y riesgos. Dentro de estas, se encuentran los siguientes impedimentos:

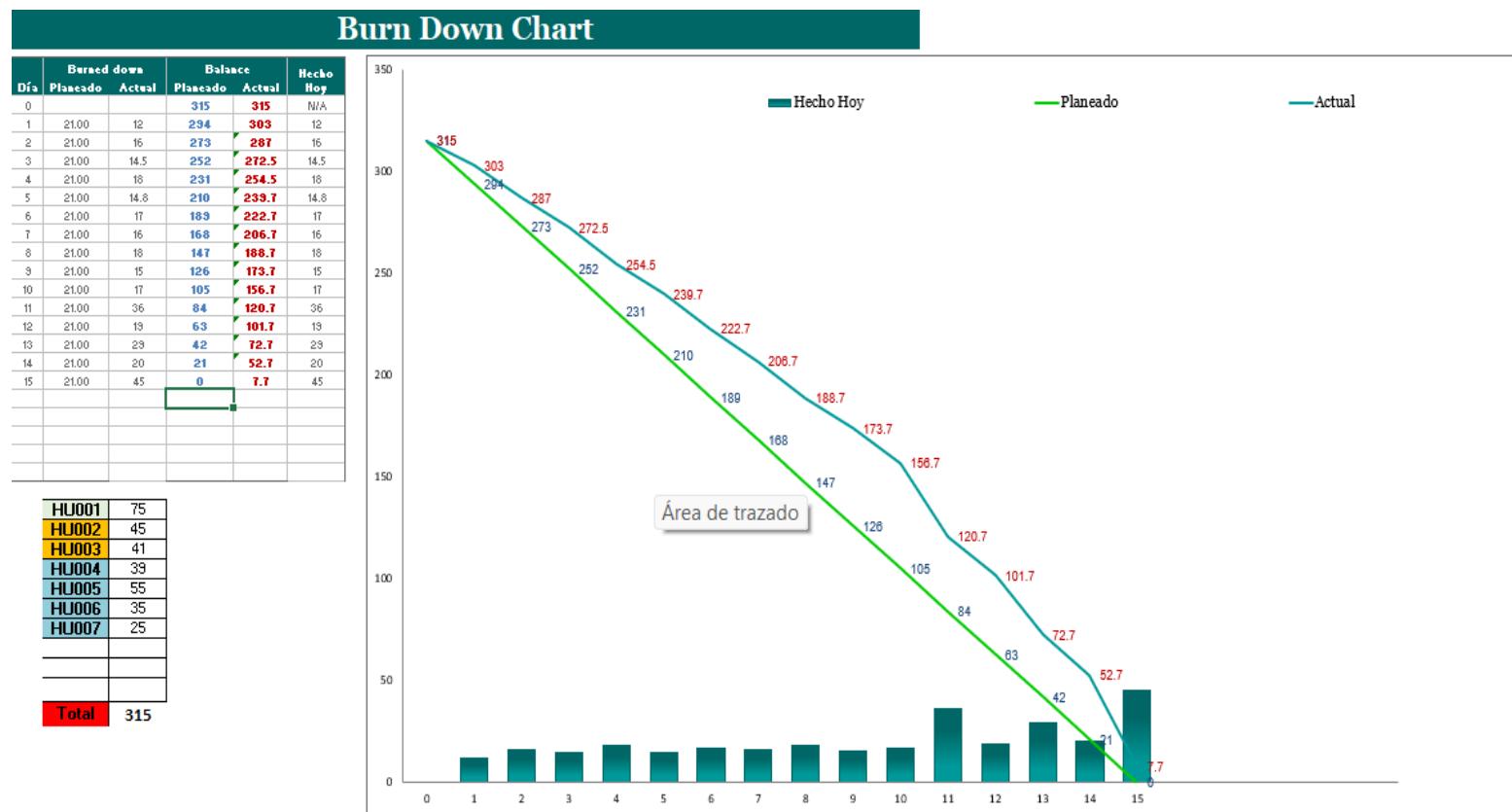
- Zonas horarias diferentes entre los desarrolladores del proyecto.

Sprint Burndown Chart

El gráfico de Burndown Chart se ingresa diariamente, después de cada dayli meeting y está realizado por el Scrum Master del proyecto con la finalidad de representar el esfuerzo esperado versus el real para el Sprint 3. Para poder obtener el porcentaje de avance diario, se toman los puntos de historia definidos en el Product Backlog, y se establece el esfuerzo diario obtenido.

Figura 72.

Sprint Burdonwn Chart -Sprint 03

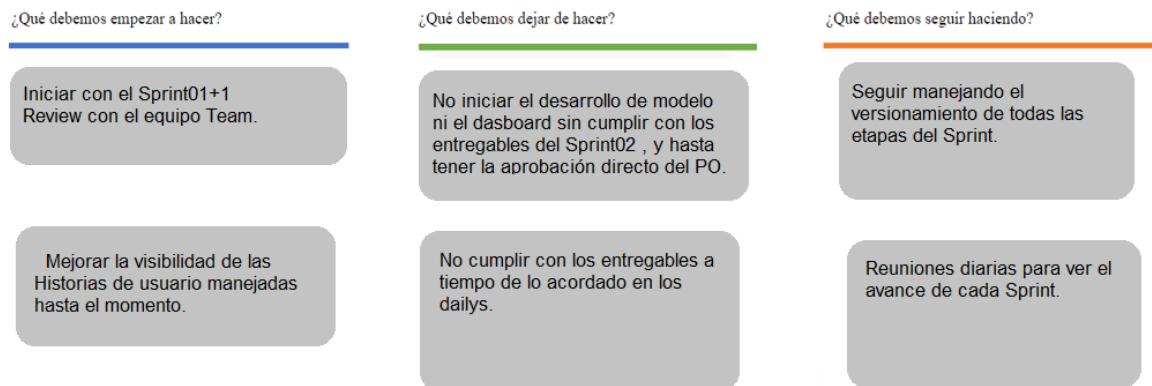


3.5.4 Sprint Retrospective

Dentro de los planes de mejora contemplados en la reunión realizada el día 07/07/2023, se obtuvieron los puntos detallados.

Figura 73.

Sprint Retrospective 03



CAPITULO 7: CONCLUSIONES DEL PROYECTO

7.1 Conclusiones

El proyecto "NYC Taxis & Carbon Emission" proporcionó con éxito una comprensión integral de los patrones de uso de taxis, el impacto de las emisiones y las oportunidades de mejora. Los hallazgos destacaron el potencial para reducir la contaminación, aumentar los viajes compartidos, mejorar la eficiencia del conductor y aumentar los ingresos. Con estos conocimientos, la empresa de servicios de transporte de pasajeros puede tomar decisiones informadas, incorporando potencialmente vehículos eléctricos a su flota y contribuyendo a un ecosistema de transporte más ecológico y sostenible en la ciudad de Nueva York.

7.2 Recomendaciones

Basado en el análisis realizado en el proyecto "NYC Taxis & Carbon Emission", se presentan las siguientes recomendaciones para la empresa de servicios de transporte de pasajeros interesada en expandirse al sector de los taxis y lograr los objetivos establecidos:

1. Incorporación de vehículos eléctricos: Considere la transición hacia una flota de taxis eléctricos para reducir significativamente las emisiones de contaminantes. Evaluar la viabilidad económica y logística de la adquisición e implementación de vehículos eléctricos, así como la infraestructura de carga necesaria.
2. Promoción de viajes compartidos: Implementar estrategias para fomentar los viajes compartidos entre los pasajeros, como incentivos económicos, campañas de concientización y mejoras en la plataforma de reservas. Establecer acuerdos y alianzas con empresas y organizaciones para promover la cultura del viaje compartido y maximizar la utilización de los vehículos.

3. Mejora de la experiencia del cliente: Continuar mejorando la calidad y comodidad del servicio ofrecido a los pasajeros. Esto puede incluir opciones de pago más convenientes, como pagos electrónicos y tarjetas de crédito.
4. Análisis continuo de datos: Establecer un sistema de monitoreo y análisis de datos en tiempo real para evaluar constantemente el desempeño de la flota de taxis, la demanda de los pasajeros y los resultados de las estrategias implementadas. Usar datos para tomar decisiones informadas y ajustar estas estrategias en consecuencia.

GLOSARIO

A continuación, se detalla la recopilación de definiciones o explicaciones de palabras que se mencionan en el presente documento:

Tabla 13.

Glosario de términos usados en el documento.

Término	Definición
API	Interfaz de programación de aplicaciones constituido por conjuntos de funciones, procedimientos y subrutinas, que permite realizar la comunicación entre dos aplicaciones.
AWS	Conjunto de servicios de computación sobre la nube pública ofrecida por la empresa Amazon a través de internet.
Burndown chart	“Diagrama de Quemado” refiera a una representación gráfica del trabajo por hacer en un proyecto en el tiempo.
Deep learning	Campo de estudio del aprendizaje de máquinas (“Machine Learning”) enfocado en el uso de algoritmos inspirados en la estructura y función del cerebro denominado Redes Neuronales Artificiales.
EDT	La Estructura de Desglose de Trabajo (EDT), representa una descomposición jerárquica orientada al producto entregable del proyecto por ejecutar por el equipo de trabajo.
Machine learning	Campo de estudio que brinda a las computadoras la habilidad de aprender sin ser explícitamente programadas. Contempla la programación de computadoras para que puedan aprender de una data de entrada.
Redes neuronales convolucionales.	Es un tipo de red neuronal artificial donde las neuronas artificiales, corresponden a campos receptivos de una manera muy similar a las neuronas en la corteza visual primaria de un cerebro biológico.
SCRUM	Marco de referencia empleada para el desarrollo ágil de un software.
Sprint	Periodo de tiempo en el cual se realiza y evalúa un trabajo específico.
PySpark	Es un lenguaje de programación compatible con Apache Spark.
Hugging Face	Es una plataforma de procesamiento de lenguaje natural o NLP basadas en inteligencia artificial.

MongoDB	Es un sistema de base de datos NoSQL, orientado a documentos y de código abierto.
SQL	Es un lenguaje específico de dominio, diseñado para administrar, y recuperar información de sistemas de gestión de bases de datos relacionales

SIGLARIO

A continuación, se detalla la recopilación de los significados de las abreviaturas o que se mencionan en el presente documento:

Tabla 14.

Siglario de las siglas usadas en el documento.

Abreviatura	Significado
AWS	Servicios de Amazon Web
BD	Base de Datos
CNN	Redes neuronales convolucionales.
EDT	Estructura de desglose de trabajo
HU	Historia de usuario
IDE	Entorno de desarrollo integrado
KPI	Indicador clave de desempeño
PMI	Asociación profesional de gestores de proyecto
TI	Tecnología de la información
UX	Experiencia de usuario
EDA	Análisis Exploratorio de Datos
ETL	Extracción, Transformación y Carga de datos

VSC

Visual Studio Code.

CRUD

Create (Crear), Read (Leer), Update (Actualizar) y Delete (Borrar).

REFERENCIAS BIBLIOGRÁFICAS

- MongoDB. (s. f.). MongoDB Atlas: Cloud Document Database.
https://www.mongodb.com/cloud/atlas/lp/try4?utm_source=google&utm_campaign=search_gs_pl Evergreen_atlas_core_prosp-brand_gic-null_amers-cl_ps-all_desktop_eng_lead&utm_term=mongodb&utm_medium=cpc_paid_search&utm_ad=&utm_ad_campaign_id=12212624314&adgroup=115749715423&cq_cmp=12212624314&gad=1&gclid=CjwKCAjwzJmlBhBBEiwAEJyLu5s7yvVRzR-Y23YVgsFU1ktojXsvzxEvZ3NthwVYnbEpTZ2Y1_9eBoCvg8QAvD_BwE
- AWS | Cloud Computing - Servicios de informática en la nube. (s. f.). Amazon Web Services, Inc. <https://aws.amazon.com/es/>
- Fundamentos de la metodología Agile. (s. f.). <https://www.wrike.com/es/project-management-guide/fundamentos-de-la-metodologia-agile/>
- Madedios. (2022b). Scrum: qué es y cómo funciona este marco de trabajo. www.wearemarketing.com. <https://www.wearemarketing.com/es/blog/metodologia-scrum-que-es-y-como-funciona.html>

ANEXOS

I. Gestión del Proyecto

Anexo I. Diccionario EDT, según sus entregables.

Nombre	Descripción	Responsables	Supuestos	Riesgos	Dependencias
1.2.11 Product Backlog	Permite hallar los principales requerimientos y requisitos para la construcción del aplicativo.	Responsable Principal: David Hospital	Tener el Project Chárter terminado y aprobado	No se logre terminar el Project Chártero se tendrá claro que roles son necesarios para el proyecto.	Project Charter, y EDT.

1.3.1.1 Acta de Sprint Planning	Sirve para inspeccionar el Backlog del Producto (Product Backlog) y que el equipo de desarrollo seleccione los Product Backlog Ítems en los que va a trabajar durante el siguiente Sprint.	Responsable Principal: David Hospital	Tener el Product Backlog aprobado y priorizado según su esfuerzo y dimensiones.	No se logre terminar el Product Backlog aprobado y priorizado según su esfuerzo y dimensiones. Que son necesarios para el proyecto.	Project Charter y Product Backlog.
1.3.1.2 Acta de Sprint Backlog	Es la suma del Objetivo del Sprint, los elementos del Product Backlog elegidos para el Sprint, más un plan de acción	Responsable Principal: David Hospital	Tener el Product Backlog aprobado y priorizado según su esfuerzo y dimensiones y el Sprint Planning.	No se logre terminar el Product Backlog y Sprint Planing.	Project Charter y Product Backlog y Sprint Planing.

1.3.1.3 Acta de Daily Scrum	Reunión para volver a planificar y tomar decisiones.	Responsable Principal: David Hospital	Tener el Product Backlog aprobado y priorizado según su esfuerzo y dimensiones y el Sprint Planning y el Sprint Backlog.	No se logre terminar el Product Backlog y Sprint Planing y el Sprint Backlog.	Product Backlog y Sprint Planing y Sprint Backlog.
1.3.1.4 Acta de Sprint Retrospective	Último evento del Sprint y es la oportunidad para que el Equipo Scrum se analice a sí mismo y haga una propuesta de mejoras para que el desarrollo del siguiente Sprint sea más eficiente.	Responsable Principal: David Hospital	Tener el Sprint Planning acabado.	No se logre terminar Sprint Planing .	Sprint Planing.

1.3.1.5 Acta de Sprint Review.	reunión informal a la que asiste el equipo Scrum con el objetivo de ofrecer una demostración del prototipo del producto y determinar qué pendientes fueron terminados.	Responsable Principal: David Hospital	Tener el Sprint Retrospective acabado y prototipos con su criterio de aceptación.	No se logre hacer el Sprint Retrospective los prototipos.	Sprint Retrospective, prototipos con su criterio de aceptación.
1.4.1 Tablero Kanban	Herramienta ágil de gestión de proyectos diseñada para ayudar a visualizar el trabajo, limitar el trabajo en curso y maximizar la eficiencia (o el flujo). Puede ayudar tanto a los equipos ágiles como a los de DevOps a definir el orden de su trabajo diario.	Responsable Principal: David Hospital	Tener la gestión de cronograma y EDT acabado.	No se logre hacer el cronograma.	Cronograma.

1.4.2 Burndown Chart	Representación gráfica del trabajo por hacer en un proyecto en el tiempo. Usualmente el trabajo remanente se muestra en el eje vertical y el tiempo en el eje horizontal. Es decir, el diagrama representa una serie temporal del trabajo pendiente.	Responsable Principal: David Hospital	Tener los Sprint y product backlog y cronograma del proyecto acabado.	No se logre hacer el cronograma y tablero Kanban.	Cronograma y tablero Kanban.
1.5.3.2 Entrega del producto final	Es la entrega Final del producto	Responsable Principal: David Hospital	Tener el modelo terminado y validado.	No se logre testear el modelo	Validaciones y test

