Data Mining - MIAV1E2 Proyecto Final

Docente:

MSc. Renzo Franck Claure Aracena

Integrantes Grupo 1:

- **Carolina Bello**
- **Nicolas Oporto**
- **Luis Martinez**
- **Oscar Loayza**
- **Joseph Thenier**

Comprensión del negocio

Contexto. Bluebikes es el sistema de bicicletas compartidas del área metropolitana de Boston. El objetivo de este proyecto es transformar datos de viajes en conocimiento accionable para distintas áreas (operaciones, planificación de estaciones, y analítica de demanda).

Preguntas de negocio guiadas:

1. ¿Cómo se distribuyen las duraciones de viaje y cuáles son los patrones por hora/día/estación del año?
2. ¿Qué calidad y cobertura geográfica tiene el histórico de viajes? ¿Hay outliers o registros inválidos?
3. ¿Podemos predecir la demanda por hora en estaciones para apoyar re-balanceo y staffing?

Entregables clave:

- Base “minable” integrada y limpia, en formato parquet.
- EDA con hallazgos sobre duraciones y geografía.
- Conjunto de features temporales y geoespaciales.
- Modelo de pronóstico por hora/estación con comparación contra un baseline ingenuo (Naive)

UAGRM SCHOOL OF ENGINEERING UNIDAD DE POSTGRADO

Comprensión de los datos

Cobertura temporal y esquemas. Se integró histórico de Bluebikes/Hubway 2015–2025. En la descarga y lectura se detectaron tres versiones de esquema:

- **Esquema 1** (legado): columnas tipo ride_id, rideable_type, started_at, ended_at, start_lat/lng, end_lat/lng, member_casual… — 27 archivos.
- **Esquema 2** (reciente): tripduration, starttime, stoptime, bikeid, usertype, postal code… — 35 archivos.
- **Esquema 3** (intermedio): como el 2 + birth year, gender — 64 archivos.
- (conteos impresos en el notebook de EDA).

Volumen. El consolidado supera 27 millones de viajes, lo que obligó a usar PySpark

Calidad inicial (resumen):

- Campos con mucha falta de datos en legados: birth_year, gender, postal_code y algunos rideable_type.
- Estaciones nulas en años recientes por la evolución “dockless” (viajes sin anclaje fijo).
- Coordenadas faltantes/cero y puntos fuera del área operativa.

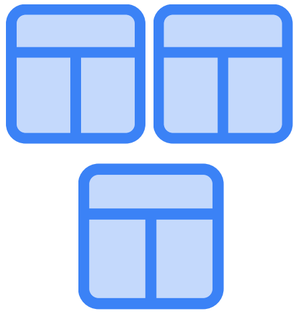
UAGRM SCHOOL OF ENGINEERING
UNIDAD DE POSTGRADO

Preparación de los datos

Flujo de integración

https://s3.amazonaws.com/hubway-data/

126 archivos .csv



Integración Multiesquema

Lectura de 27 M de registros

UAGRM SCHOOL OF ENGINEERING UNIDAD DE POSTGRADO

Preparación de los datos

Valores Nulos

columna	nulos	% nulos
birth_year	19377092	71.59
gender	18686522	69.04
postal_code	18253759	67.44
rideable_type	17220537	63.63
end_station_id	29036	0.11
end_station_name	28417	0.10
end_lat	21830	0.08
end_lng	21830	0.08

- Por negocio, **rideable_type** nulo ⇒ **docked_bike**.
- Estaciones nulas ⇒ etiquetas “Dockless start” / “Dockless end” para no perder viajes y poder agregarlos.
- Columnas con **nulos crónicos** y poco valor predictivo (birth_year, gender, postal_code) se descartaron en la base final.

UAGRM SCHOOL OF ENGINEERING
UNIDAD DE POSTGRADO

Preparación de los datos

Valores Nulos

columna	nulos	% nulos
birth_year	19377092	71.59
gender	18686522	69.04
postal_code	18253759	67.44
rideable_type	17220537	63.63
end_station_id	29036	0.11
end_station_name	28417	0.10
end_lat	21830	0.08
end_lng	21830	0.08

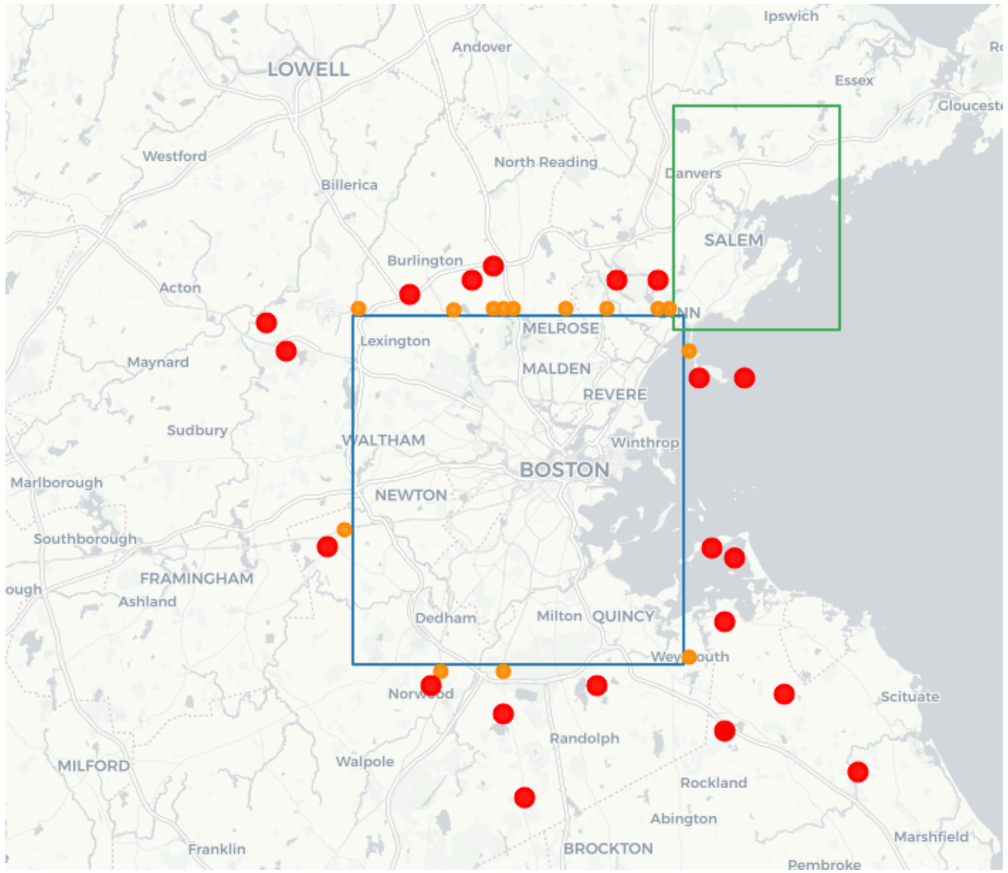
- Por negocio, **rideable_type** nulo ⇒ **docked_bike**.
- Estaciones nulas ⇒ etiquetas “Dockless start” / “Dockless end” para no perder viajes y poder agregarlos.
- Columnas con **nulos crónicos** y poco valor predictivo (birth_year, gender, postal_code) se descartaron en la base final.

UAGRM SCHOOL OF ENGINEERING
UNIDAD DE POSTGRADO

Preparación de los datos

Duraciones y coordenadas

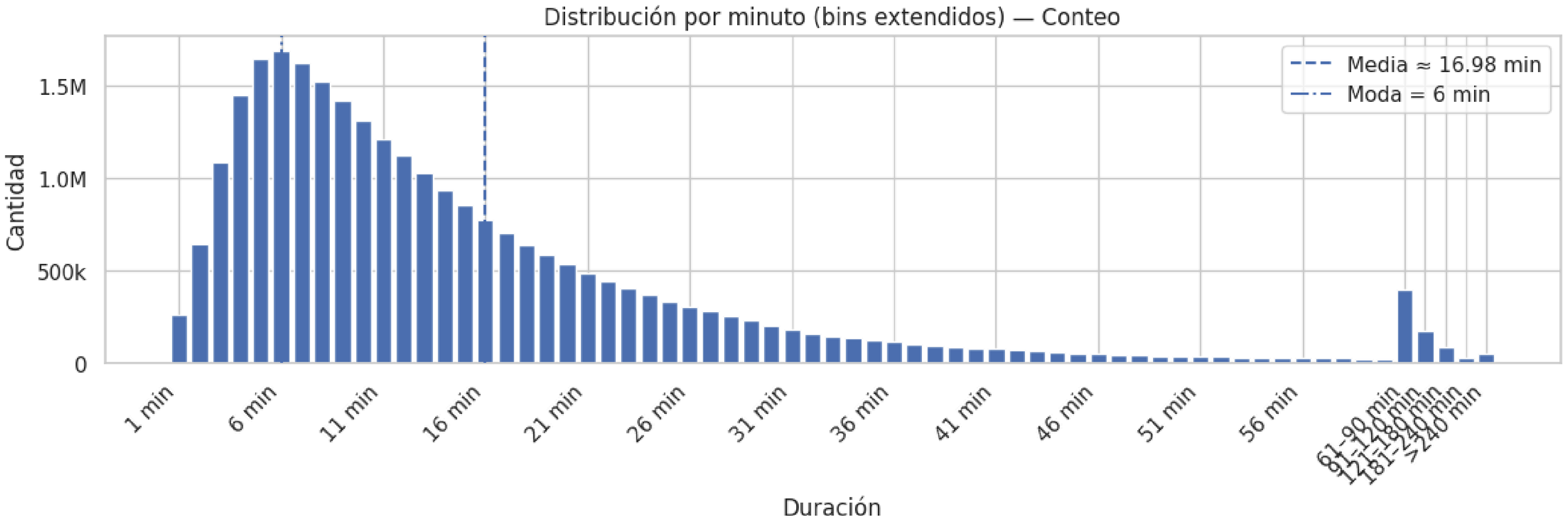
	bin_label	count	pct
0	0–30 min	24072068	89.14
1	30–45 min	1669216	6.18
2	45–60 min	490552	1.82
3	1–2 h	583348	2.16
4	2–6 h	135336	0.50
5	6–12 h	16728	0.06
6	12–24 h	16875	0.06
7	>24 h	19299	0.07



UAGRM SCHOOL OF ENGINEERING
UNIDAD DE POSTGRADO

Análisis Exploratorio

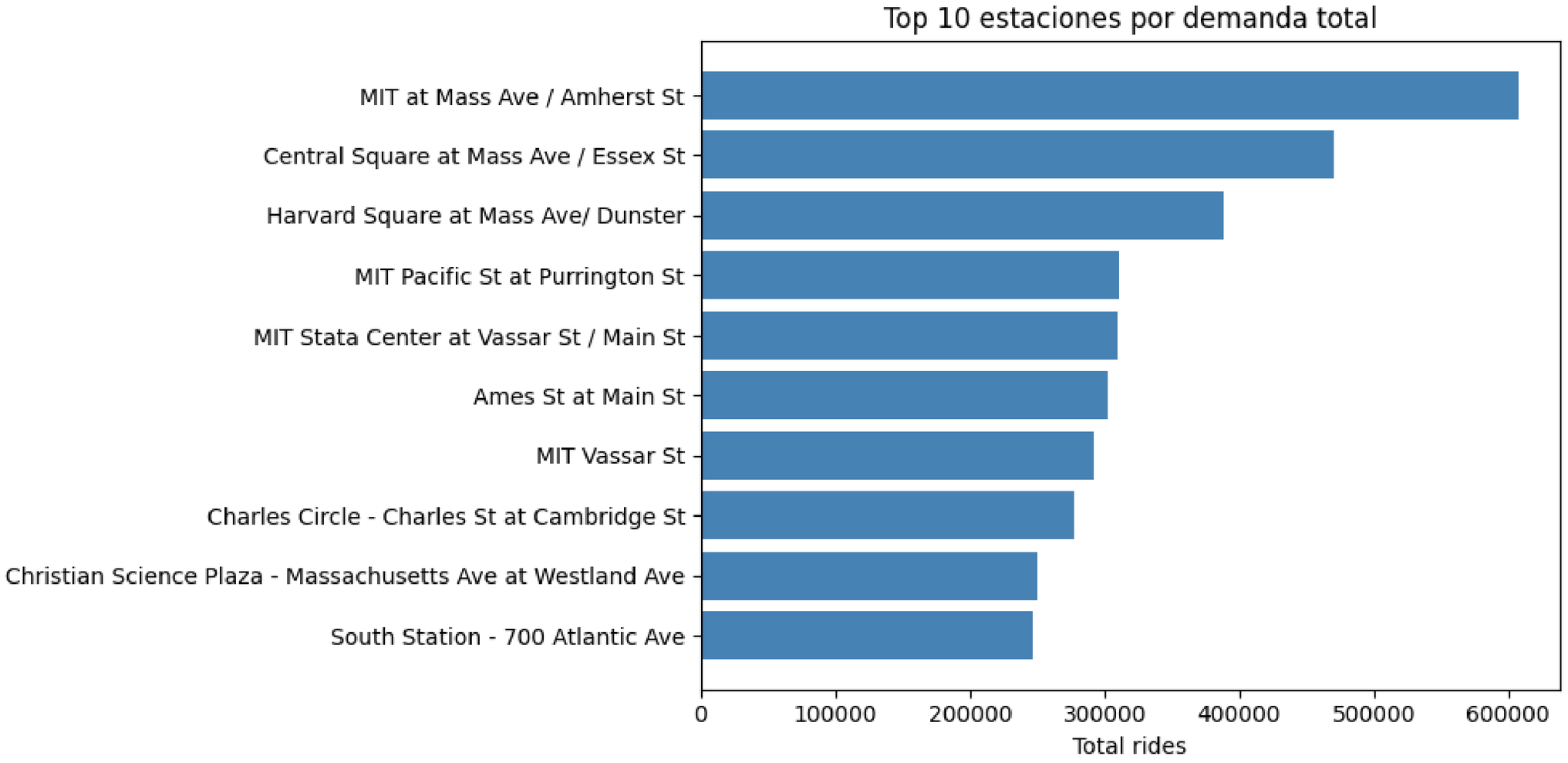
Duración de viajes



UAGRM SCHOOL OF ENGINEERING
UNIDAD DE POSTGRADO

Análisis Exploratorio

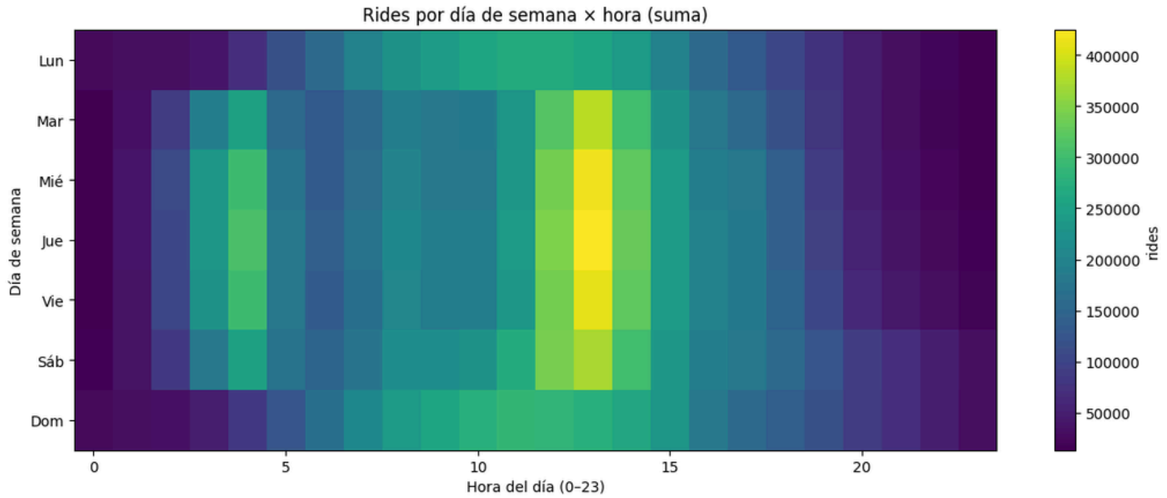
Demanda de viajes por estación

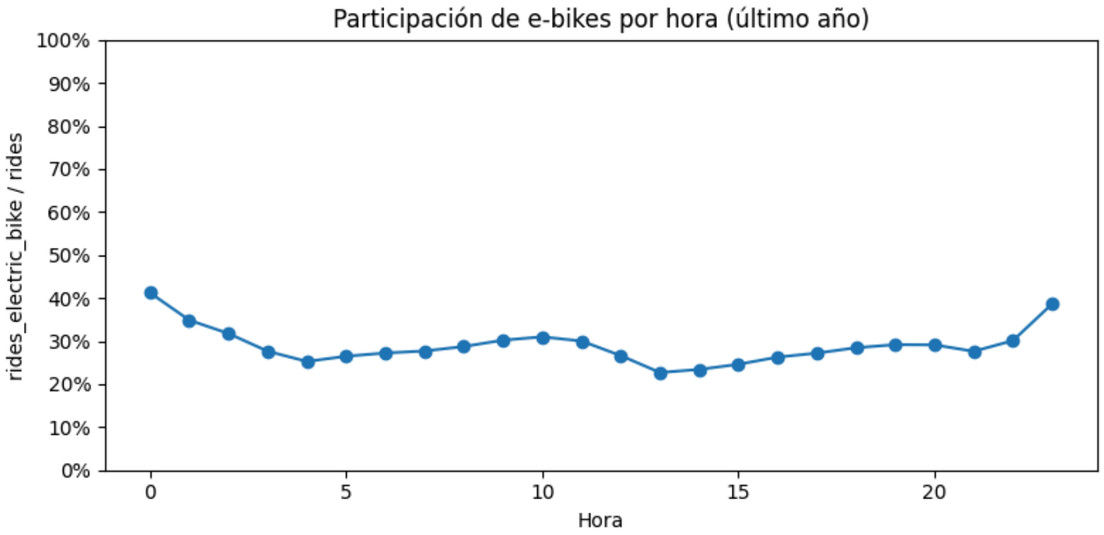


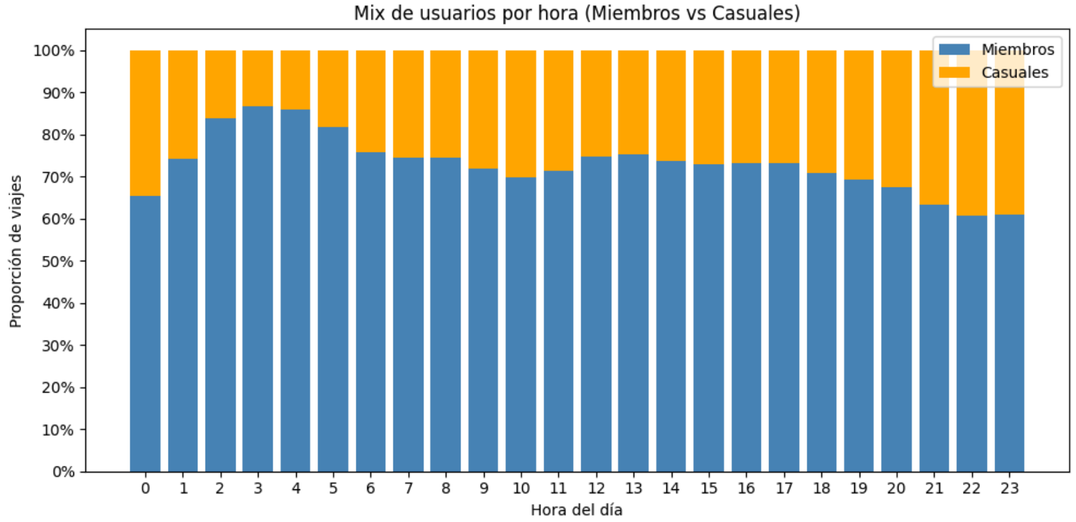
UAGRM SCHOOL OF ENGINEERING
UNIDAD DE POSTGRADO

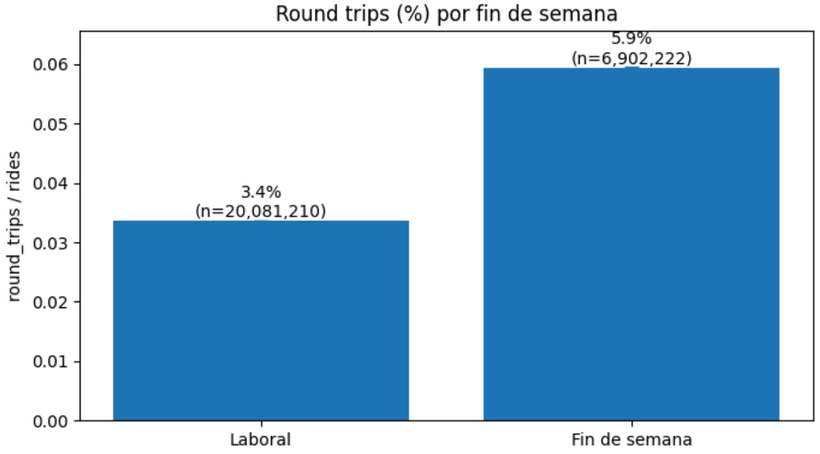
Análisis Exploratorio

Varios





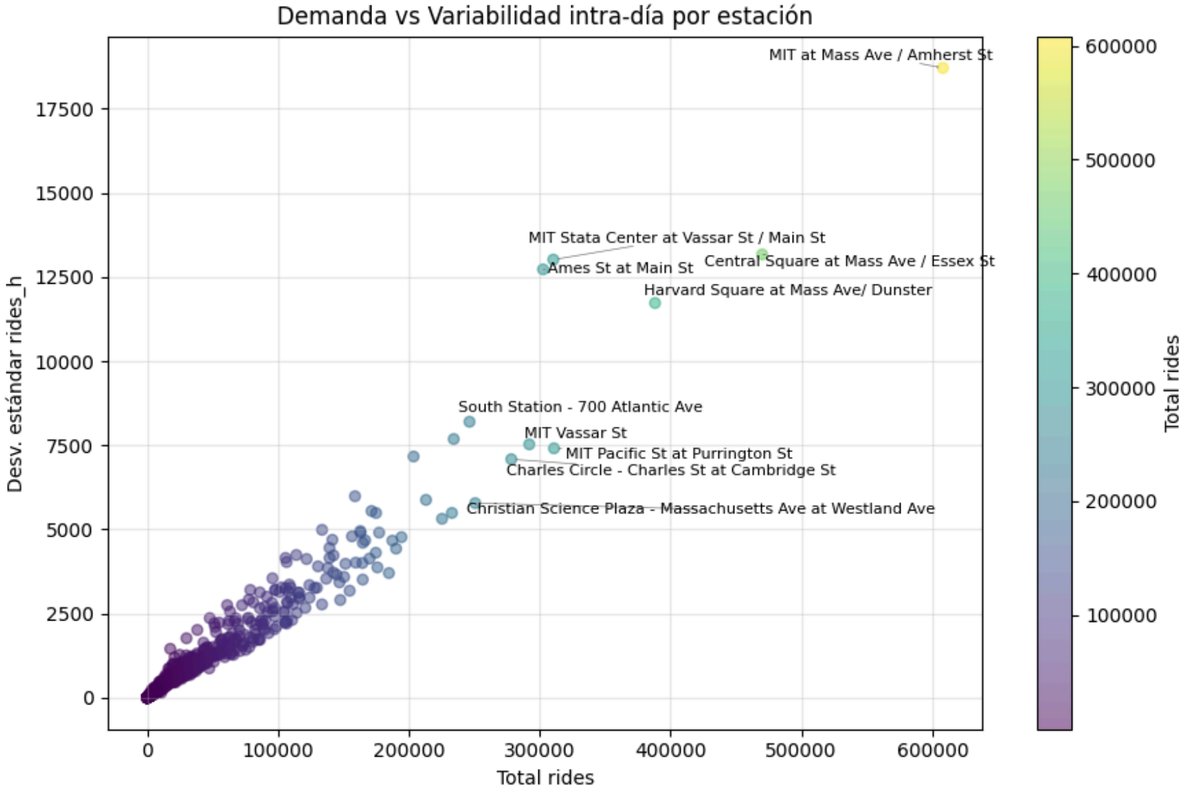


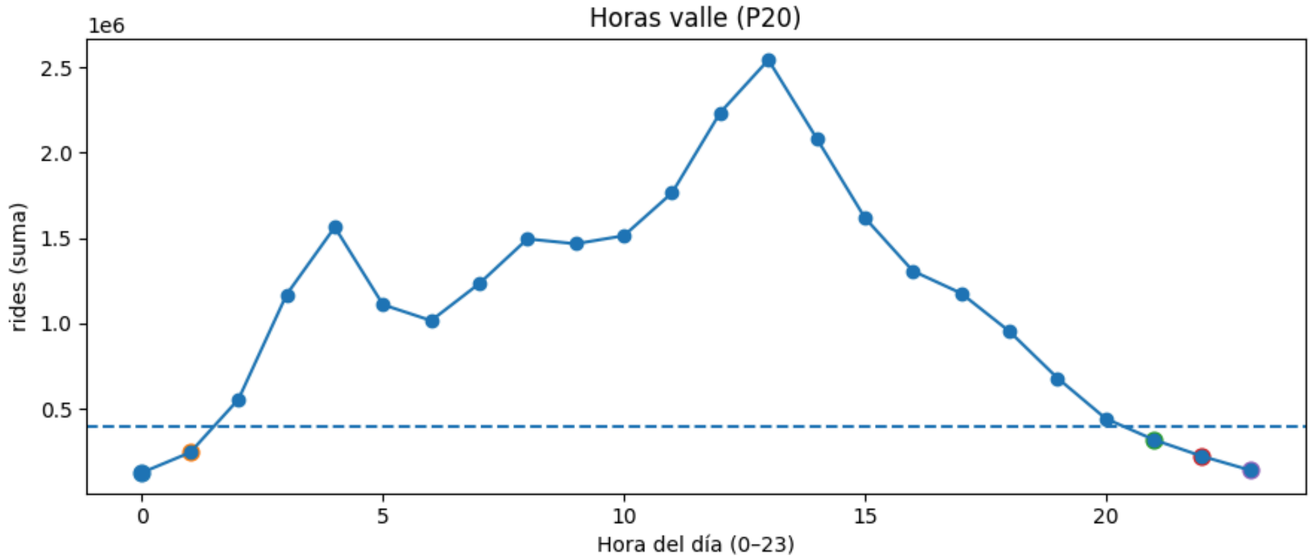


UAGRM
SCHOOL OF ENGINEERING
UNIDAD DE POSTGRADO

Análisis Exploratorio

Complejo vs simple

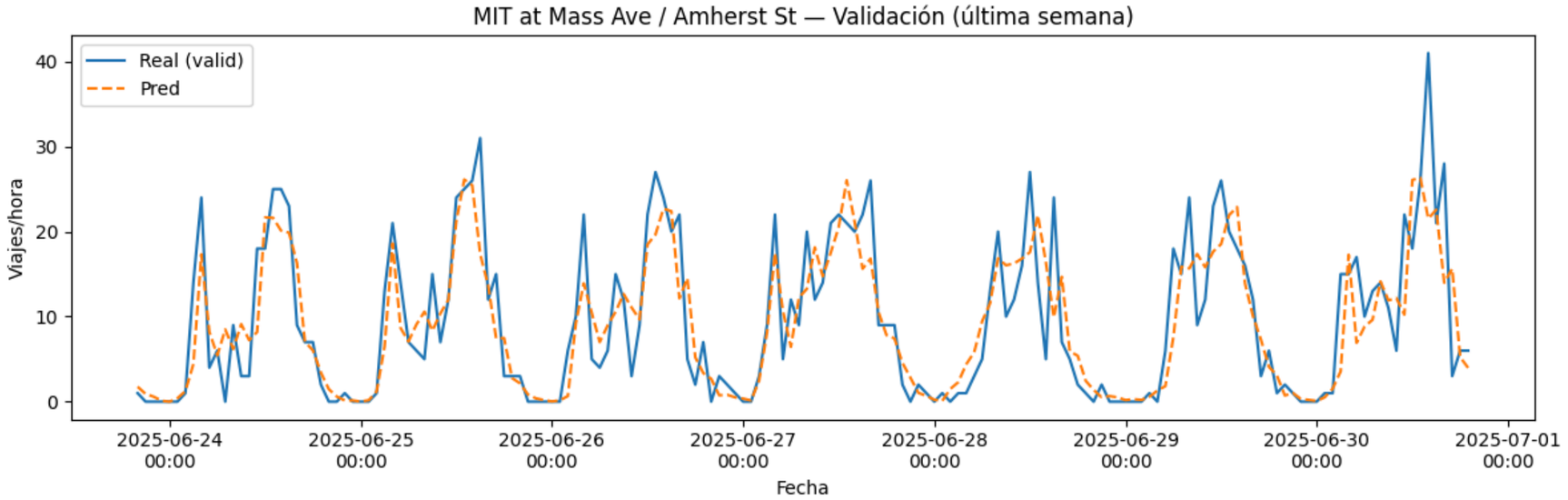


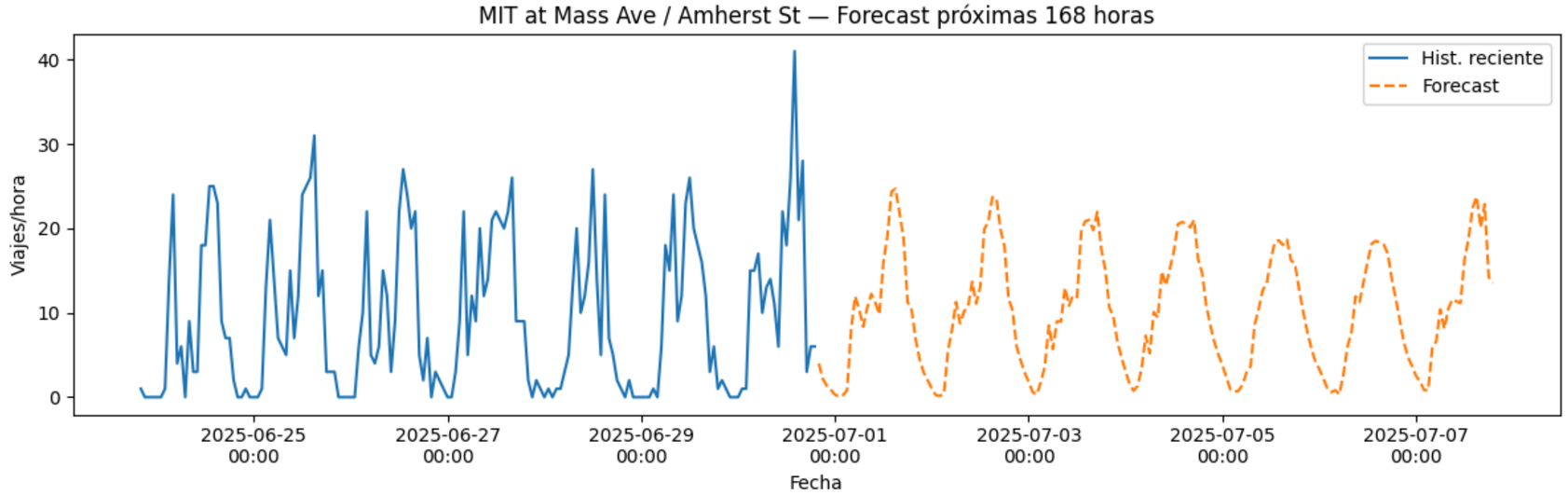


UAGRM SCHOOL OF ENGINEERING
UNIDAD DE POSTGRADO

Modelado

XGBoost Forecast para Series Temporales





MAE vs Naive −18,4% (de 4,083 a 3,332)

UAGRM SCHOOL OF ENGINEERING
UNIDAD DE POSTGRADO

Evaluación

Ajuste a objetivos de negocio.

- La base lograda es consistente, auditable y minable, adecuada para tableros y modelos.
- El pronóstico horario reduce el error significativamente frente a un enfoque ingenuo, habilitando acciones de re-balanceo y staffing más informadas.

Riesgos/limitaciones.

- Cambios de esquema en el tiempo implicaron supuestos (p.ej., imputar rideable_type nulo como docked_bike; etiquetar viajes “dockless”).
- No se integraron factores externos (clima, eventos), que podrían mejorar la precisión.
- Validaciones geográficas por BBOX+buffer son robustas, pero no sustituyen una máscara de polígonos exactos de cobertura.

UAGRM
SCHOOL OF
ENGINEERING
UNIDAD DE POSTGRADO