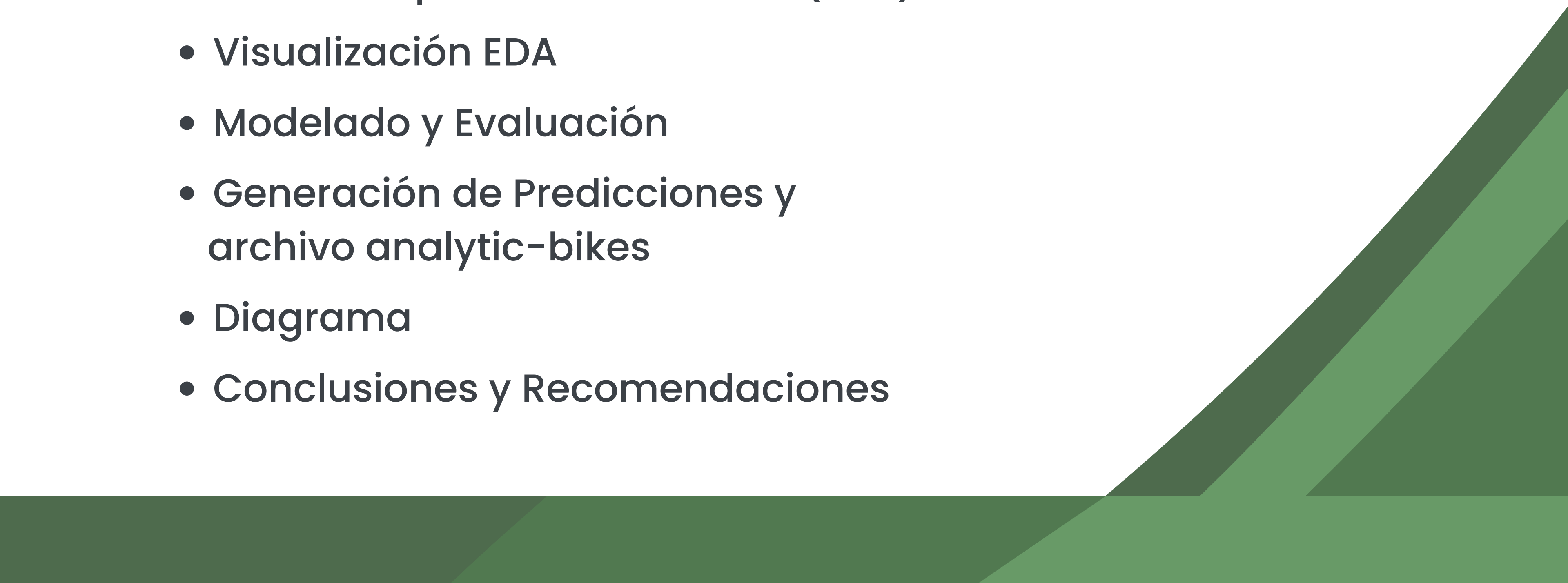


# PROYECTO BICICLETAS

De La Ciudad de los Angeles

# INDICE

- Introducción, Objetivo y Alcance
  - Análisis Exploratorio de Datos (EDA)
  - Visualización EDA
  - Modelado y Evaluación
  - Generación de Predicciones y  
archivo analytic-bikes
  - Diagrama
  - Conclusiones y Recomendaciones
- 

# INTRODUCCION, OBJETIVO & ALCANCE

En Los Ángeles existe un sistema compartido de bicicletas que brinda datos anónimos acerca del uso del servicio que ofrecen. La tabla que se contiene proporciona el histórico de viajes que se han realizado durante cerca de 9 meses, y contiene una columna que es de particular interés y que se buscará analizar a más profundidad.

Para el problema analítico, se desea saber si es posible inferir si el tipo de pase es “Monthly Pass” tomando en cuenta las demás variables de viaje.

## **Objetivo:**

Es predecir el tipo de pase (passholder\_type) que utiliza cada usuario en un sistema compartido de bicicletas, utilizando datos históricos que incluyen información de viajes (horarios, ubicaciones, duración, etc.). Además, se busca generar un archivo de predicciones en el formato especificado (analytic-bikes.csv) y diseñar un flujo de trabajo completo para la puesta en producción del modelo.

## **Alcance:**

- Realizar un análisis exploratorio de datos (EDA) para entender la demanda del servicio.
- Preprocesar y transformar los datos para extraer características relevantes.
- Construir y evaluar modelos predictivos (ej. Random Forest, Logistic Regression, LGBMClassifier).
- Generar predicciones en el formato requerido.
- Documentar y diseñar el flujo de producción del modelo, incluyendo un diagrama del pipeline.



# ANÁLISIS EXPLORATORIO DE DATOS (EDA)

*Comprender la distribución y tendencias de los datos para identificar patrones relevantes que ayuden en la selección de variables y en la estrategia del modelado. Uno de los patrones es entender el uso de la saturación del servicio y la distribución de los planes.*

## 1. Carga y Visualización de DataSet

- Se cargan los conjuntos de datos train.csv y test.csv.
- Se visualiza la información general (número de filas, columnas, tipos de datos) y se identifican valores nulos.

## 2. Análisis de Valores Faltantes

- Se revisa el porcentaje de datos faltantes en columnas críticas (ej. start\_lat, end\_lat, plan\_duration, passholder\_type).
- Se decide la estrategia de manejo: eliminar filas con passholder\_type nulo (variable objetivo) y aplicar imputación (mediana o valor más frecuente) para otros campos.

# EDA – CARGA Y VISUALIZACION DATASET

La importancia de la variable "passholder\_type" (Plan del passholder) está intrínsecamente ligada a los objetivos y metas del proyecto bicicletas, y su valor se reflejará en la capacidad del modelo para proporcionar información valiosa para la toma de decisiones estratégicas.

```
# Leyendo el dataset de train
train_data = pd.read_csv(train_data_path)
RangeIndex: 700000 entries, 0 to 699999
Data columns (total 14 columns):
#   Column          Non-Null Count  Dtype
---  -
0   trip_id         700000 non-null  int64
1   duration        700000 non-null  int64
2   start_time      700000 non-null  object
3   end_time        700000 non-null  object
4   start_lat       694437 non-null  float64
5   start_lon       694437 non-null  float64
6   end_lat         681426 non-null  float64
7   end_lon         681426 non-null  float64
8   bike_id         700000 non-null  object
9   plan_duration   699792 non-null  float64
10  trip_route_category 700000 non-null  object
11  passholder_type  697424 non-null  object
12  start_station   700000 non-null  int64
13  end_station     700000 non-null  int64
dtypes: float64(5), int64(4), object(5)
memory usage: 74.8+ MB
```

A primera instancia se visualizan los valores faltantes en los datos *star\_lat, star\_lon, end\_lat, end\_lot, plan\_duration, passholder\_type*. También existen valores nulos.

```
#Leyendo el dataset de test.
test_data = pd.read_csv(test_data_path)
RangeIndex: 569886 entries, 0 to 569885
Data columns (total 12 columns):
#   Column          Non-Null Count  Dtype
---  -
0   trip_id         569886 non-null  int64
1   duration        569886 non-null  int64
2   start_time      569886 non-null  object
3   end_time        569886 non-null  object
4   start_lat       565264 non-null  float64
5   start_lon       565264 non-null  float64
6   end_lat         554995 non-null  float64
7   end_lon         554995 non-null  float64
8   bike_id         569886 non-null  object
9   trip_route_category 569886 non-null  object
10  start_station   569886 non-null  int64
11  end_station     569886 non-null  int64
dtypes: float64(4), int64(4), object(4)
memory usage: 52.2+ MB
None
```

A primera instancia se visualizan los valores faltantes en los datos *star\_lat, star\_lon, end\_lat, end\_lot*. También existen valores nulos.

Conclusión: Las variables numéricas (star\_lat, star\_lon, end\_lat, end\_lot) se Imputaron con la mediana. y la variables categóricas (passholder\_type, plan\_duration) con "desconocido"



# EDA – ANALISIS DE VALORES FALTANTES

*En realidad se visualiza un porcentaje pequeño de valores faltantes. Pero son importantes para el análisis de los datos en este proyecto.*

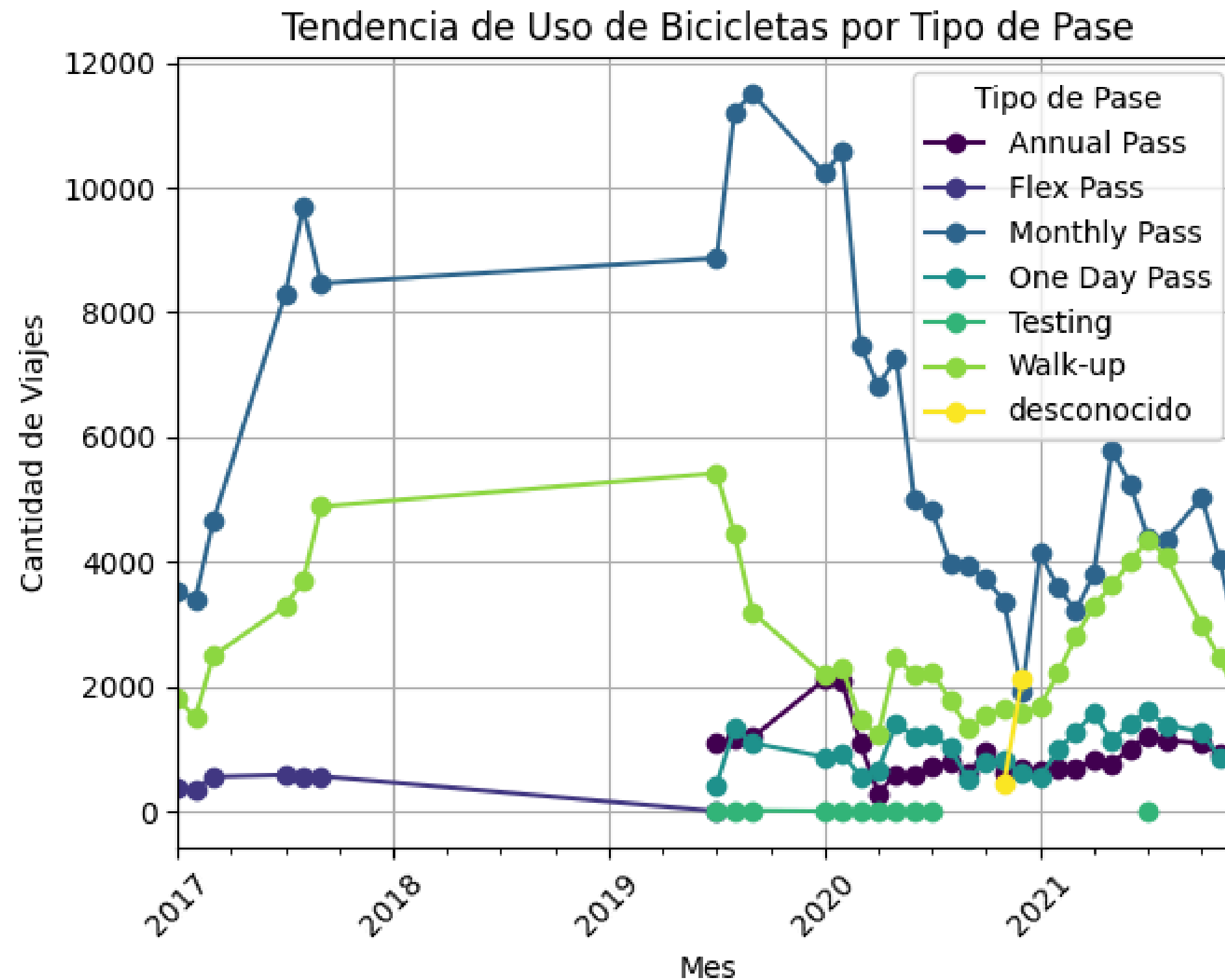
- Las Columnas de start\_lat, start\_lon, end\_lat, end\_lon. Podemos reemplazarla con la mediana para evitar los valores atipicos.
- Las Columnas de passholder\_type y plan\_duration. Como son variables categóricas entonces podemos reemplazarla con el valor "desconocido".

Train DataSet	% of Missing Values
end_lat	2.65
end_lon	2.65
start_lat	0.79
start_lon	0.79
passholder_type	0.37
plan_duration	0.03
trip_id	0.00
duration	0.00
start_time	0.00
end_time	0.00
bike_id	0.00
trip_route_category	0.00
start_station	0.00
end_station	0.00

Test DataSet	% of Missing Values
end_lat	2.61
end_lon	2.61
start_lat	0.81
start_lon	0.81
trip_id	0.00
duration	0.00
start_time	0.00
end_time	0.00
bike_id	0.00
trip_route_category	0.00
start_station	0.00
end_station	0.00

# VISUALIZACIÓN EDA

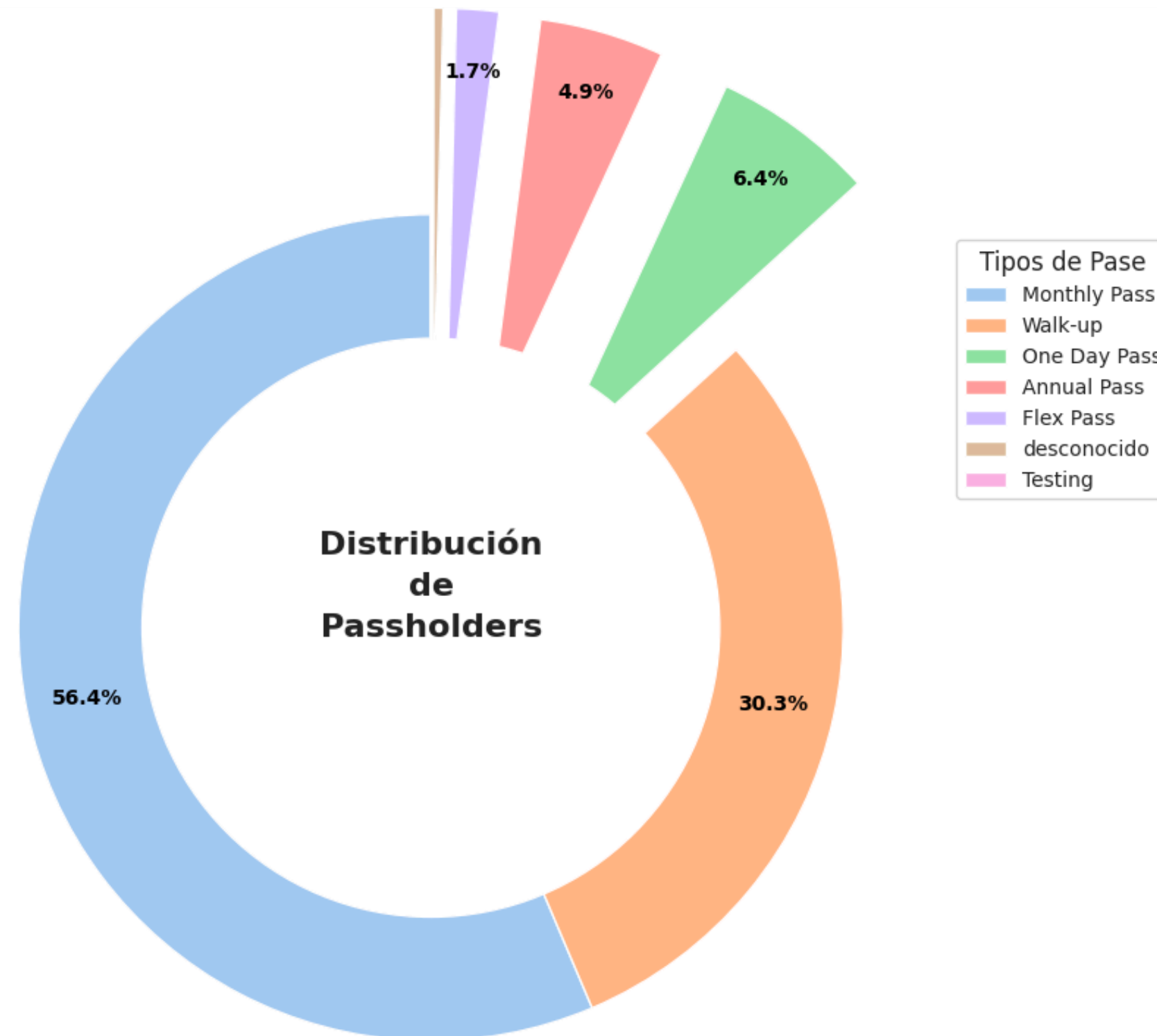
## TENDENCIA USO DE BICICLETAS POR PASSHOLDER / AÑO



- En la gráfica de tendencia se visualiza cual es el tipo de pase y el crecimiento. Identificando que el pase por "MES" es el mas usado. Aunque se visualiza que va a la baja.
- El segundo mas usado es Walk-up. Se visualiza un pequeño aumento en el pase Anual.
- Podemos afirmar que los usuarios prefieren el pase MONTHLY.
- Se puede decir que los usuarios de WALK-UP, son usuarios ocasionales pero es un ingreso considerable especialmente en las horas pico.
- Se puede decir que la caída en 2020 podría relacionarse con la pandemia, afectando a todos los planes.

# VISUALIZACIÓN EDA

## DISTRIBUCION DE LOS PASES PASSHOLDERS

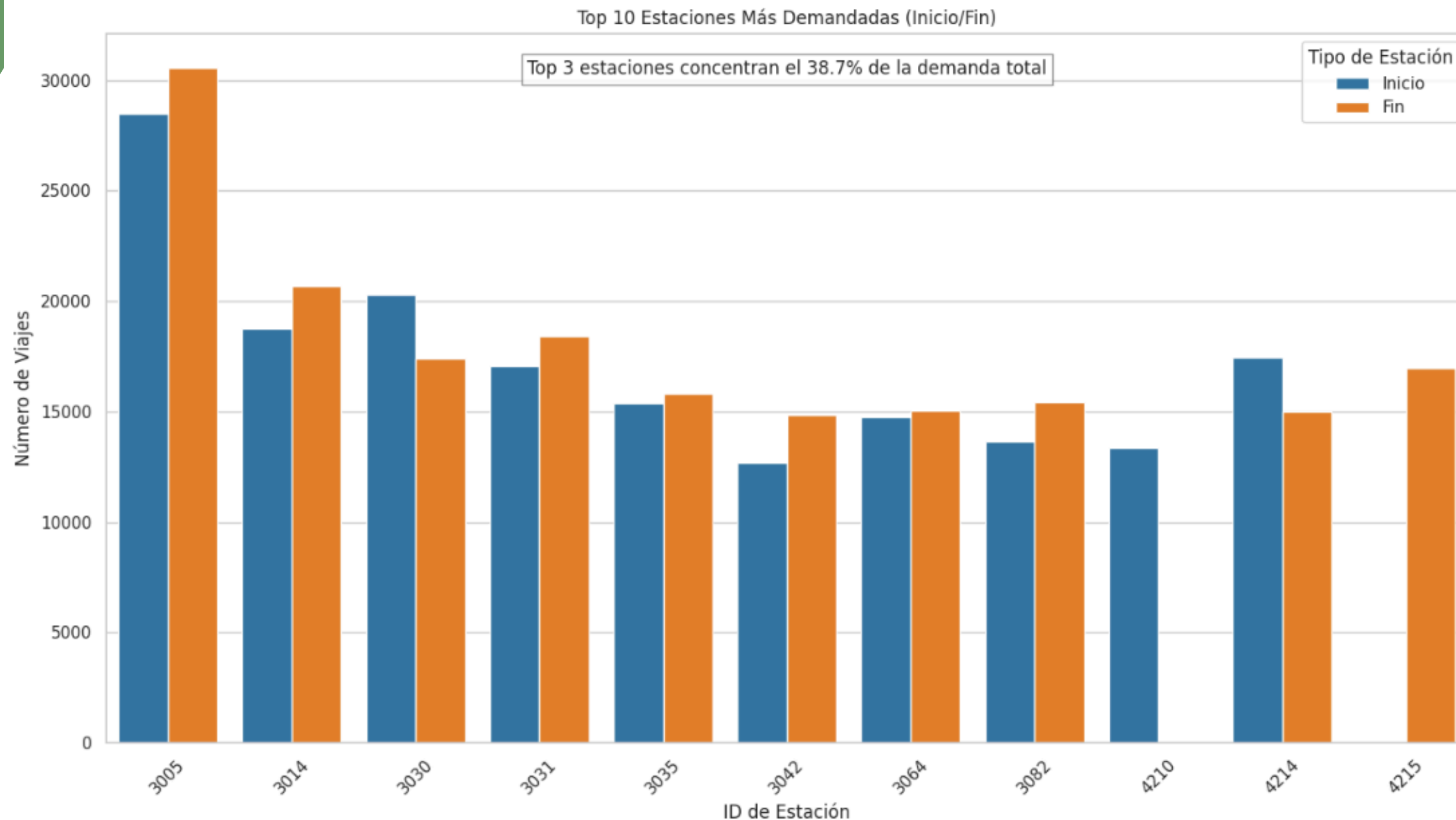


- Dominancia del Monthly Pass (56.4%): Indica que el servicio es utilizado principalmente por usuarios recurrentes (ej: viajes laborales/estudio).
- Walk-up (30.3%): Segmento importante de usuarios ocasionales, clave para ingresos adicionales.
- One Day Pass (6.4%): Potencial para captar turistas con promociones estacionales.
- Annual Pass (4.9%): Baja adopción sugiere necesidad de mejores incentivos (ej: descuentos por renovación).
- Alerta de datos: Categoría "desconocido" (0.4%) requiere investigación (posibles errores en registro).



# VISUALIZACIÓN EDA

## TOP 10 DE LAS ESTACIONES MAS USADAS (INCIO/FIN)

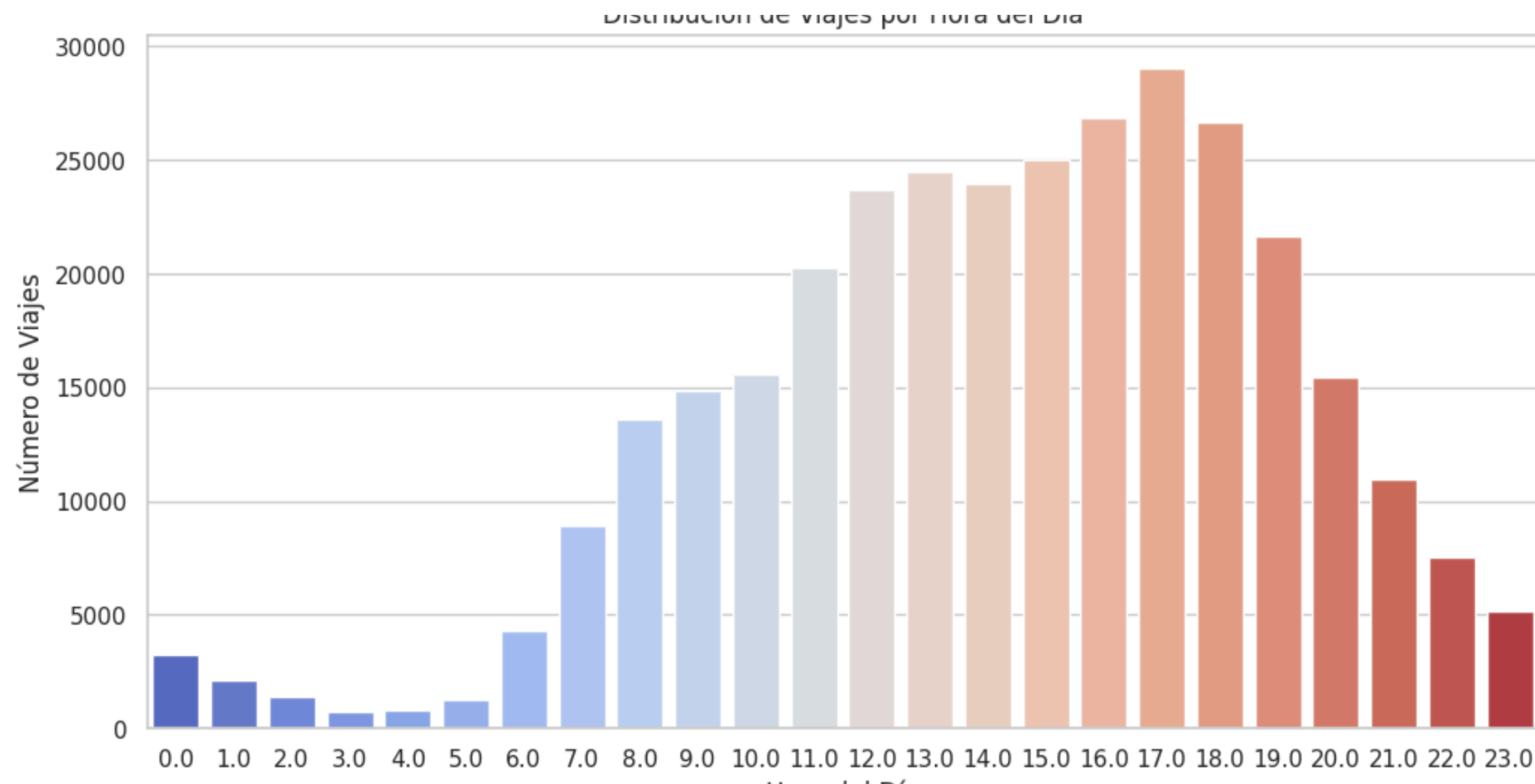


- Estación 3005 es el principal hub del sistema, con 28,490 viajes de inicio y 30,576 viajes de fin, siendo la más crítica para la operación.
- Concentración geográfica: Las 3 estaciones principales (3005, 3014, 3030) acumulan el 38.7% de la demanda total, lo que sugiere:
  - Necesidad de reforzar mantenimiento y capacidad en estas estaciones.
  - Posible saturación en horarios pico.
- Patrón de viajes: Las estaciones más demandadas son consistentes en inicio y fin, indicando flujos predecibles (ej: viajes casa-trabajo).

# VISUALIZACIÓN EDA

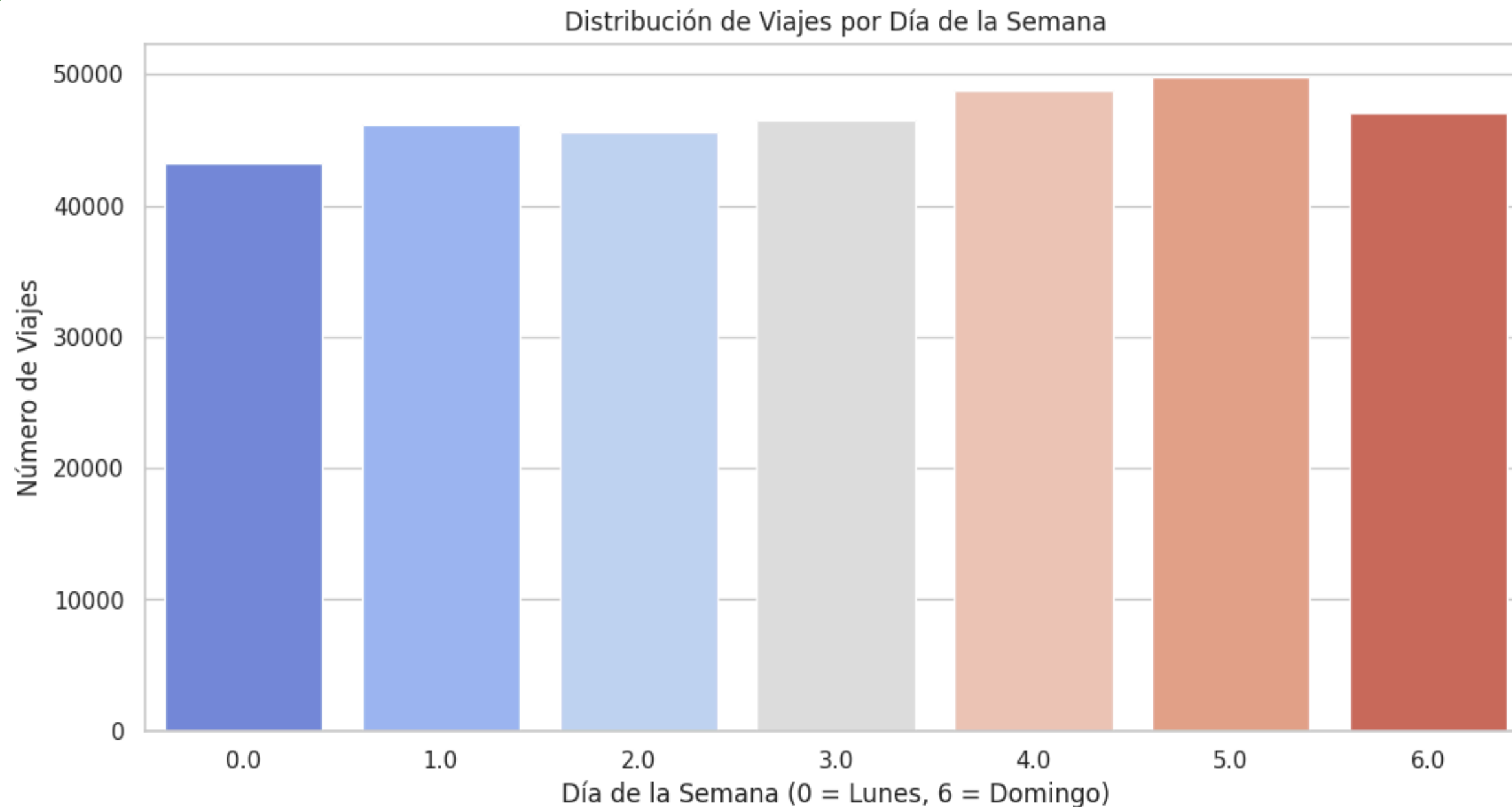
## DISTRIBUCIÓN DE VIAJES POR HORA DEL DÍA

- Horarios pico claros:
- Mañana (8-9 AM): 13,603 viajes.
- Tarde (5-6 PM): 29,022 viajes (máximo absoluto).
- Demanda nocturna mínima: Solo 5.2% de los viajes ocurren entre 12 AM y 5 AM.
- Se puede priorizar rebalanceo de bicicletas antes de horas pico.
- Promover planes "Walk-up" o "One Day Pass" para uso nocturno con tarifas reducidas.



# VISUALIZACIÓN EDA

## DISTRIBUCIÓN DE VIAJES POR DÍA DE LA SEMANA



- **Uso progresivo:** La demanda aumenta durante la semana, alcanzando su pico los viernes (49,847 viajes).
- **Patrón semanal:**
  - Lunes-Viernes: Promedio de 46,000 viajes/día (uso laboral).
  - Sábado (47,139 viajes): Demanda estable, posiblemente para ocio.
- **Recomendación:** Ofrecer planes flexibles los fines de semana para atraer usuarios recreativos.

# CONCLUSIONES DEL EDA

- El pase Monthly Pass es el más utilizado, mientras que otros tipos (como One Day Pass o Flex Pass) presentan variaciones en la demanda.
- Existen patrones temporales claros: picos de uso en determinadas horas y días.
- Algunas estaciones concentran un alto número de viajes, lo que puede reflejar ubicaciones estratégicas.
- La presencia de valores nulos en algunas columnas sugiere la necesidad de un tratamiento cuidadoso en la fase de preprocesamiento.

# HALLAZGO CLAVE

- El sistema funciona principalmente como transporte laboral, con alta dependencia de usuarios recurrentes (Monthly Pass).
- Para escalar, es crucial:
  - Atraer nuevos segmentos (turistas, usuarios ocasionales).
  - Garantizar experiencia perfecta en estaciones estratégicas.
  - Usar datos de horarios/demandas para predecir necesidades de mantenimiento.
- Patrones Temporales: 70% de los viajes ocurren en días laborables.

# RECOMENDACIONES ESTRATEGICAS

- Enfoque en estaciones críticas:
  - Instalar más estacionamientos en estaciones 3005, 3030 y 3014.
  - Monitoreo en tiempo real de disponibilidad durante horas pico.
- Optimización de flota:
  - Rebalancear bicicletas hacia el centro urbano antes de 7 AM.
  - Incentivar devoluciones en zonas periféricas post 6 PM con descuentos.
- Estrategia comercial:
  - Paquetes "Fin de Semana" para usuarios Walk-up.
  - Programa de fidelización para convertir usuarios Monthly Pass a Annual Pass.
- Mejora de datos:
  - Investigar y corregir registros en "desconocido".
  - Implementar encuestas para entender necesidades de usuarios Flex Pass.

# MODELADO Y EVALUACION

Construir y evaluar modelos que permitan predecir de forma precisa el tipo de pase.

Métricas de RandomForest:

	precision	recall	f1-score	support
Annual Pass	0.58	0.57	0.57	1951
Flex Pass	0.52	0.64	0.57	1727
Monthly Pass	0.99	0.99	0.99	41934
One Day Pass	0.25	0.49	0.33	3621
Testing	1.00	1.00	1.00	8
Walk-up	0.90	0.75	0.82	25274
accuracy			0.87	74515
macro avg	0.71	0.74	0.71	74515
weighted avg	0.90	0.87	0.88	74515

## MODELO RANDOM FOREST/ BOSQUE ALEATORIO

- Métricas
  - Precision: 87%
  - Macro: 71%,
  - Ponderado: 88%

### OBSERVACIONES DEL DESEMPEÑO DEL MODELO

- **Monthly Pass:** tiene un desempeño casi perfecto (0.99 en precisión, recall y F1-score) y, al mismo tiempo, es la clase con mayor cantidad de ejemplos. Esto influye de forma significativa en el weighted average, haciendo que este promedio sea muy alto (por ejemplo, precisión de 0.90).
- **One Day Pass:** Presenta métricas bajas (precisión 0.25, recall 0.49, F1-score 0.33). Esto sugiere que el modelo tiene dificultades para identificar correctamente esta clase, lo que podría ser consecuencia de un menor número de ejemplos o de características menos discriminativas para esta categoría.
- **Flex Pass y Annual Pass:** Tienen métricas moderadas, en torno a 0.57 en F1-score, lo que indica un desempeño aceptable pero con margen de mejora.
- **Testing** muestra métricas perfectas (1.00 en todas), pero solo tiene 8 muestras. Esto podría no ser representativo y es importante tenerlo en cuenta al interpretar el desempeño general.
  - **La diferencia entre el macro average (0.71 de precisión y 0.71 de F1) y el weighted average (0.90 de precisión y 0.88 de F1) es significativa.**
  - **Esto indica que a pesar de que el modelo funciona muy bien en la clase dominante, su desempeño en clases minoritarias (como One Day Pass) es inferior.**
  - **En contexto se puede afirmar que las clases estan desbalanceadas. y MACRO AVG es un indicador del rendimiento de las clases.**

# MODELADO Y EVALUACION

Construir y evaluar modelos que permitan predecir de forma precisa el tipo de pase.

Mejores hiperparámetros encontrados:  
`{'max_depth': None, 'min_samples_split': 5, 'n_estimators': 200}`

F1 Macro score: 0.6990

## MODELO VALIDACIÓN CRUZADA - RANDOM FOREST

- Métricas
  - Accuracy: 4.82%
  - Macro: 0.0270
  - Recall: 0.0396
  - F1-Score: 0.0315

### OBSERVACIONES DEL DESEMPEÑO DEL MODELO

- La Métrica de Precision indica que solo el 4.82% de las predicciones totales fueron correctas.
- Estas métricas son extremadamente bajas, lo que sugiere que el modelo está fallando en la clasificación de la mayoría de las muestras cuando se evalúa en datos que no formaron parte del entrenamiento.
- **Monthly Pass:** A pesar de tener el mayor número de ejemplos (41934), sus métricas son muy bajas (por ejemplo, F1 de 0.01).
- **One Day Pass y Testing:** No se predijeron correctamente, con F1-score de 0.00.
- **Walk-up:** Aunque presenta resultados algo mejores (F1 de 0.09), siguen siendo muy bajos.
- **Existe una discrepancia muy marcada entre lo que muestra la validación cruzada (F1 Macro  $\approx$  0.69) y la evaluación en el conjunto de validación (F1 Macro  $\approx$  0.03). Esto puede indicar:**
  - **Sobreajuste (overfitting):** El modelo se ajustó demasiado a las particiones de la validación cruzada y no generaliza a datos nuevos.
  - **Cambio en la distribución:** El conjunto de validación podría tener una distribución diferente a la de los datos usados en el entrenamiento y la validación cruzada.
  - **Problemas con el preprocesamiento o con el muestreo de datos:** Quizás existan diferencias en cómo se tratan o distribuyen las clases en el conjunto de validación.

Evaluación en el conjunto de validacion

Accuracy: 0.0482  
Precision (macro): 0.0270  
Recall (macro): 0.0396  
F1-score (macro): 0.0315

Reporte de clasificación:

	precision	recall	f1-score	support
Annual Pass	0.04	0.07	0.05	1951
Flex Pass	0.03	0.04	0.04	1727
Monthly Pass	0.02	0.01	0.01	41934
One Day Pass	0.00	0.00	0.00	3621
Testing	0.00	0.00	0.00	8
Walk-up	0.07	0.12	0.09	25274
accuracy			0.05	74515
macro avg	0.03	0.04	0.03	74515
weighted avg	0.03	0.05	0.04	74515



# MODELADO Y EVALUACION

*Construir y evaluar modelos que permitan predecir de forma precisa el tipo de pase.*

## MODELO VALIDACIÓN CRUZADA - RANDOM FOREST

Importancia de las variables  
duration: 0.1560  
day\_of\_week: 0.0155  
hour\_of\_day: 0.0236  
round\_trip: 0.0171  
plan\_duration: 0.7879

- Importancia de las variables
  - **plan\_duration** (78.79%):
    - La mayor parte de la “información” para la toma de decisiones del modelo proviene de la variable plan\_duration. Esto quiere decir que, para el modelo, esta característica es la más relevante para clasificar las muestras.
  - **duration** (15.60%):
    - Es la segunda más importante, pero su aporte es mucho menor comparado con plan\_duration.
  - **day\_of\_week, hour\_of\_day y round\_trip** (1.5% - 2.4% cada una):
    - Estas variables aportan muy poco en la toma de decisiones según el modelo.
- **Se observa que la variable plan\_duration domina la importancia (alrededor del 78% del total). Esto puede indicar que, en el conjunto de datos, este predictor es extremadamente informativo para distinguir entre los tipos de pase.**
- **Sin embargo, depender excesivamente de una sola variable puede generar vulnerabilidades, especialmente si esa variable presenta sesgos o no está disponible en nuevos datos.**

# MODELADO Y EVALUACION

*Construir y evaluar modelos que permitan predecir de forma precisa el tipo de pase.*

## MODELO REGRESION LOGISTICA

- Métricas
  - Puntuaciones macro de validación cruzada F1:  
0,645,0,644,0,632,0.640,0,6440,645, 0,644, 0,632,  
0,640, 0,6440,645 ,0.644 ,0,632 ,0.640 ,0,644
    - **Media: ~0.641**

```
Logistic Regression Cross-Validation F1 Macro scores:  
[0.64540049 0.6437831 0.63241592 0.63980369 0.6442584 ]  
Mean F1 Macro score (Logistic Regression): 0.6411
```

- ***La regresión logística presenta un desempeño ligeramente inferior al de Random Forest en CV (macro F1 de ~0.64 vs. ~0.69). Esto sugiere que la capacidad de capturar relaciones no lineales o interacciones entre variables es menor en la regresión logística en este caso.***

# MODELADO Y EVALUACION

*Construir y evaluar modelos que permitan predecir de forma precisa el tipo de pase.*

Evaluación en el conjunto de validación  
Accuracy: 0.9194  
Precision (macro): 0.7561  
Recall (macro): 0.7177  
F1-score (macro): 0.6995

## MODELO LGBMCLASSIFIER (LIGHTBGM)

- Métricas
  - **Accuracy:** 0.9194
  - **Precisión:** 0.7561
  - **Recall :** 0.7177
  - **F1-score :** 0.6995
  - **Clase 0:** Precisión 0.65, Recall 0.61, F1 0.63
  - **Clase 1:** Precisión 0.56, Recall 0.72, F1 0.63
  - **Clase 2:** Precisión 0.99, Recall 1.00, F1 0.99
  - **Clase 3:** Precisión 0.48, Recall 0.02, F1 0.04
  - **Clase 4:** Precisión 1.00, Recall 1.00, F1 1.00
  - **Clase 5:** Precisión 0.87, Recall 0.96, F1 0.91
- Informe de clasificación:
  - **"Pase Mensual"** se predice con F1 muy alto (0,99).
  - **"One Day Pass"** presenta un desempeño muy bajo (F1 ~0.04), lo que sugiere dificultad en capturar esta clase, posiblemente debido a su baja representación o alta variabilidad.
- El modelo LightGBM ofrece un desempeño muy bueno en el conjunto de validación (accuracy  $\approx$  92% y macro F1  $\approx$  0.70) y sus resultados en CV (macro F1 promedio  $\approx$  0.7054) son consistentes con la evaluación en validación. Sin embargo, se observa que algunas clases (por ejemplo, la clase 3 con F1 de 0.04) tienen resultados muy bajos, lo que puede deberse a:
  - Un problema de desbalanceo entre clases.
  - Que la clase en cuestión tenga características difíciles de distinguir.

Reporte de clasificación:

	precision	recall	f1-score	support
0	0.65	0.61	0.63	1951
1	0.56	0.72	0.63	1727
2	0.99	1.00	0.99	41934
3	0.48	0.02	0.04	3621
4	1.00	1.00	1.00	8
5	0.87	0.96	0.91	25274
accuracy			0.92	74515
macro avg	0.76	0.72	0.70	74515
weighted avg	0.90	0.92	0.90	74515

LightGBM Cross-Validation F1 Macro scores:  
[0.7038935 0.70174278 0.70100687 0.70886052 0.70804885]  
Mean F1 Macro score: 0.7047

# MODELADO Y EVALUACION

*Construir y evaluar modelos que permitan predecir de forma precisa el tipo de pase.*

## MODELO LGBMCLASSIFIER (LIGHTGBM)

- Importancia de las variables

Importancia de las variables

duration: 22321.0000

day\_of\_week: 10197.0000

hour\_of\_day: 14939.0000

round\_trip: 2547.0000

plan\_duration: 2272.0000

- **valores absolutos, no normalizados**
- **En contraste con Random Forest, aquí la variable “duration” es la más importante, seguida de cerca por “hour\_of\_day” y “day\_of\_week”. Las variables “round\_trip” y “plan\_duration” aportan menos información al modelo. Esto indica que, para LightGBM, el comportamiento del modelo es distinto y se está basando en otros aspectos de los datos.**

# CONCLUSIONES DEL MODELADO Y EVALUCIÓN

- **RandomForest:**

- Aunque en validación cruzada presenta métricas aceptables, su desempeño en el conjunto de validación final es muy bajo, lo que indica problemas de generalización o posibles inconsistencias en el preprocesamiento.
- La dependencia excesiva de "plan\_duration" podría estar provocando un sesgo en el modelo.

- **Regresión Logística:**

- Desempeño algo inferior en CV (macro F1 ~0.64), lo que indica que quizás las relaciones en los datos no son puramente lineales.

- **LGBMClassifier:**

- Se destaca por su robustez y mejores métricas en el conjunto de validación (precisión ~0.92 y F1 Macro ~0.70). Aunque algunas clases (como "One Day Pass") siguen siendo un reto, LightGBM ofrece un equilibrio más adecuado en la importancia de las variables y un mejor desempeño global y validación.

## RECOMENDACIONES ESTRATEGICAS

### 1. Balance de Clases:

- Considerar técnicas de re-muestreo (over-sampling o under-sampling) o ajustar pesos de clase para mejorar la predicción de las clases menos representadas, especialmente "One Day Pass".

### 2. Ingeniería de Características Adicionales:

- Explorar la inclusión de variables geográficas (por ejemplo, calcular distancias reales, agrupaciones de estaciones) o interacciones entre variables que puedan aportar información adicional.

### 3. Ajuste de Hiperparámetros:

- Continuar refinando los hiperparámetros de LightGBM y considerar la incorporación de técnicas de optimización bayesiana para explorar el espacio de parámetros de manera más eficiente.

### 4. Validación y Monitoreo:

- Asegurarse de que el pipeline de preprocesamiento y la división de datos sean consistentes, ya que la discrepancia en el desempeño de RandomForest sugiere posibles problemas en la forma en que se generan los conjuntos de validación.

# Generación de Predicciones y archivo analytic-bikes

***DURANTE EL PROCESO DE GENERACIÓN DEL ARCHIVO ANALYTICS-BIKES.CSV, SE IDENTIFICARON Y RESOLVIERON VARIOS DESAFÍOS RELACIONADOS CON LA LIMPIEZA DE DATOS, EL PREPROCESAMIENTO Y LA PREDICCIÓN DEL PASSHOLDER\_TYPE.***

**OBJETIVO: GENERAR UN ARCHIVO ANALYTIC-BIKES.CSV CON EL FORMATO REQUERIDO PARA SUBIR LAS PREDICCIONES.**

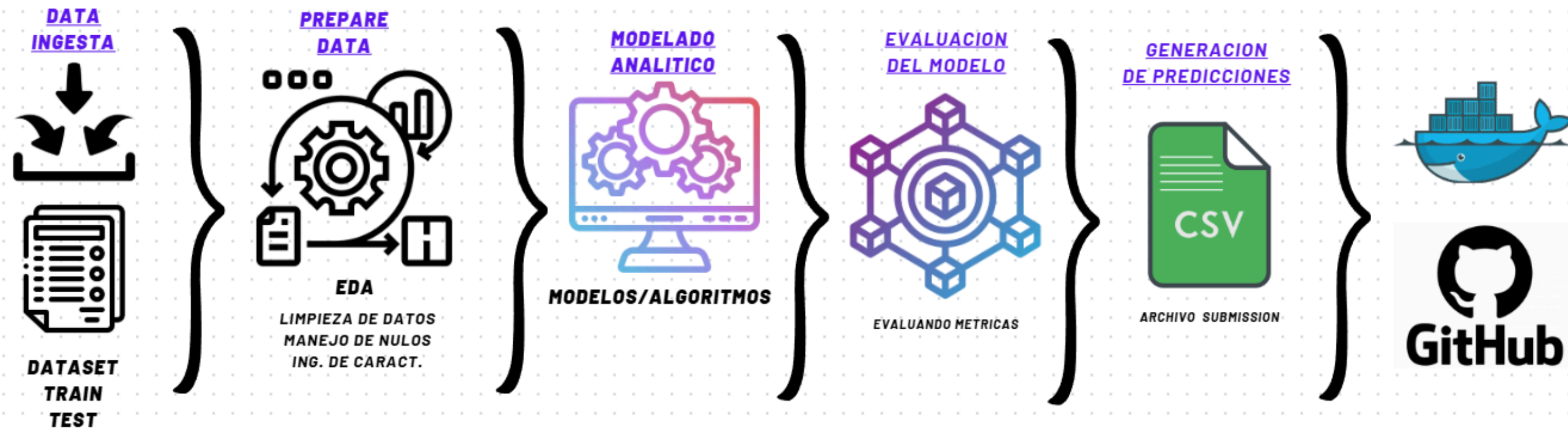
Pasos de Creación:

- Se corrigió la filtración excesiva de datos en la fase de limpieza para asegurar que todas las instancias relevantes estuvieran presentes en la predicción.
- Errores en la Etiquetación de Predicciones:
  - El modelo estaba devolviendo valores numéricos en lugar de las etiquetas originales ("Monthly Pass", "Walk-up", etc.).
  - Se implementó una decodificación del labelencoder, asegurando que las predicciones finales coincidan con los nombres originales de los tipos de pase.
- Uso Correcto del Conjunto de Prueba:
  - Se estandarizaron las características de entrada con el mismo scaler utilizado durante el entrenamiento del modelo.
- Formato del Archivo de Salida:
  - Se ajustó la estructura del archivo analytics-bikes.csv para garantizar que incluya las columnas trip\_idy & passholder\_type con los valores correctos.
- Se validó que el formato final cumpliera con las especificaciones requeridas para su posterior análisis.



# Diagrama

## **PROYECTO PIPELINE DE DATOS** **PARA EL SISTEMA DE BICICLETA DE LOS ANGELES**



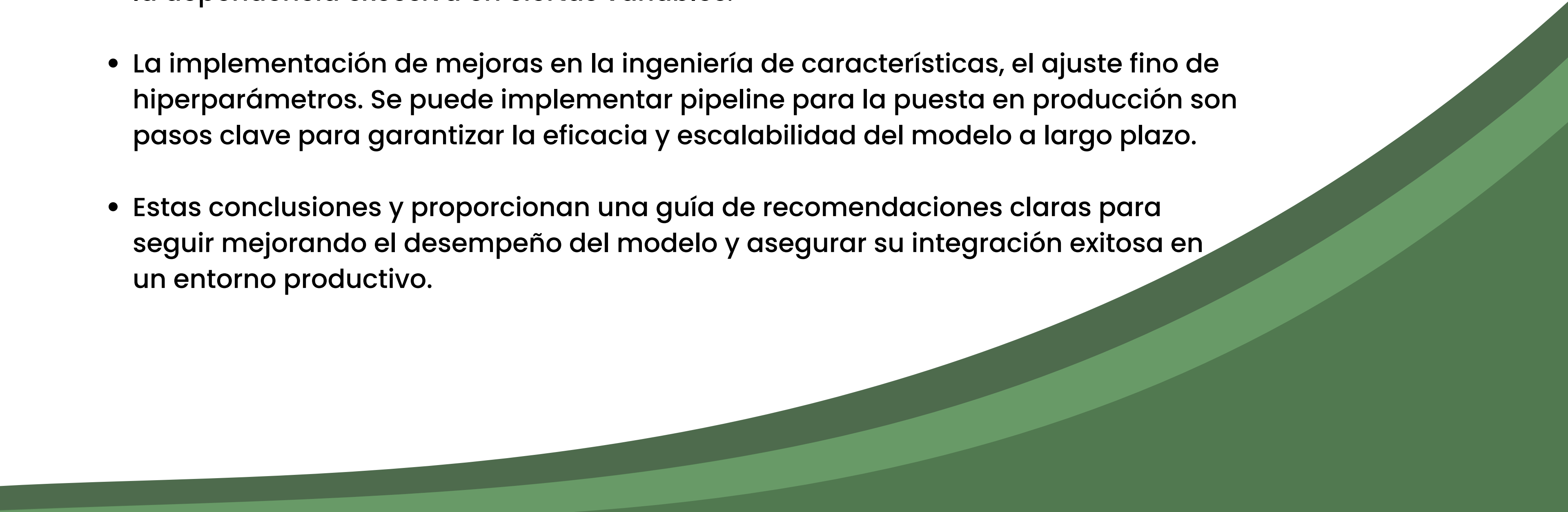
# CONCLUSIONES

AREA	ACCIÓN
Comprensión del comportamiento del usuario	<ul style="list-style-type: none"><li>• El análisis exploratorio mostró que existen patrones temporales y geográficos muy relevantes en el uso del sistema de bicicletas. Se identifican picos de demanda en horarios y días específicos, y ciertas estaciones (tanto de inicio como de fin) concentran un alto volumen de viajes.</li><li>• El pase "Monthly Pass" es, sin duda, el más utilizado, lo que indica una fuerte preferencia por la suscripción mensual. Sin embargo, otros tipos de pase, como el "One Day Pass", presentan un comportamiento más irregular, lo que sugiere oportunidades de mejora en la segmentación y oferta de estos productos.</li></ul>
Desempeño del Modelo Predictivo	<ul style="list-style-type: none"><li>• Entre los modelos evaluados, RandomForest y Logistic Regression presentaron un desempeño aceptable en validación cruzada, pero mostraron problemas al generalizar en el conjunto de validación final.</li><li>• LightGBM se destacó por su robustez, logrando mejores métricas generales (precisión alrededor del 92% y F1 Macro aproximadamente de 0.70) en la evaluación en validación. Esto sugiere que el modelo de LightGBM es el más adecuado para este problema, aunque todavía existen desafíos, especialmente en la predicción de clases minoritarias (por ejemplo, "One Day Pass").</li></ul>
Importancia de las variables	<ul style="list-style-type: none"><li>• Se observará que la variable plan_duration es la más determinante para la predicción del tipo de pase, seguida de variables relacionadas con la duración del viaje y características temporales. Esto indica que el modelo se apoya en factores clave que reflejan tanto la duración del uso del servicio como la planificación del usuario.</li><li>• La dependencia excesiva de una sola variable (en este caso, "plan_duration") puede generar sesgos, por lo que es importante explorar la incorporación de más características, especialmente aquellas relacionadas con la dimensión geográfica (como distancias y agrupaciones de estaciones).</li></ul>

# RECOMENDACIONES

AREA	ACCIÓN
Mejora en el Manejo del Desequilibrio de Clases	<ul style="list-style-type: none"><li>• Implementar técnicas de balanceo, como re-muestreo (oversampling de clases minoritarias o undersampling de clases mayoritarias) o ajustar los pesos de las clases durante el entrenamiento, para mejorar la predicción de los pases menos representados (por ejemplo, "One Day Pass").</li></ul>
Ampliar y Refinar la Ingeniería de Características	<ul style="list-style-type: none"><li>• Incluir nuevas variables geográficas (por ejemplo, distancias reales entre estaciones, clusters de ubicación) que permitan captar mejor la variabilidad espacial del servicio.</li><li>• Explorar interacciones entre variables, como la relación entre la hora del día y la estación de inicio, o la interacción entre duración del viaje y plan_duration.</li></ul>
Optimización y Monitoreo del Modelo	<ul style="list-style-type: none"><li>• Continuar afinando los hiperparámetros del modelo, utilizando técnicas avanzadas como la optimización bayesiana, para explorar de manera más exhaustiva el espacio de parámetros.</li><li>• Implementar un sistema de monitoreo en producción que permita evaluar el desempeño del modelo en tiempo real y detectar posibles degradaciones, lo que facilitará el reentrenamiento o ajuste del modelo cuando sea necesario.</li></ul>

# RESUMEN FINAL

- El proyecto ha permitido identificar patrones críticos en el uso de bicicletas compartidas y construir un modelo predictivo capaz de anticipar el tipo de paso utilizado por los usuarios.
  - Aunque LightGBM se presenta como el mejor modelo en términos de métricas generales, se deben abordar desafíos relacionados con el desequilibrio de clases y la dependencia excesiva en ciertas variables.
  - La implementación de mejoras en la ingeniería de características, el ajuste fino de hiperparámetros. Se puede implementar pipeline para la puesta en producción son pasos clave para garantizar la eficacia y escalabilidad del modelo a largo plazo.
  - Estas conclusiones y proporcionan una guía de recomendaciones claras para seguir mejorando el desempeño del modelo y asegurar su integración exitosa en un entorno productivo.
- 

# **¡GRACIAS!**

*Proyecto de Bicicletas*

*Ing. Ma. Luisa Ramos Peñaloza*

*Feb 2025*