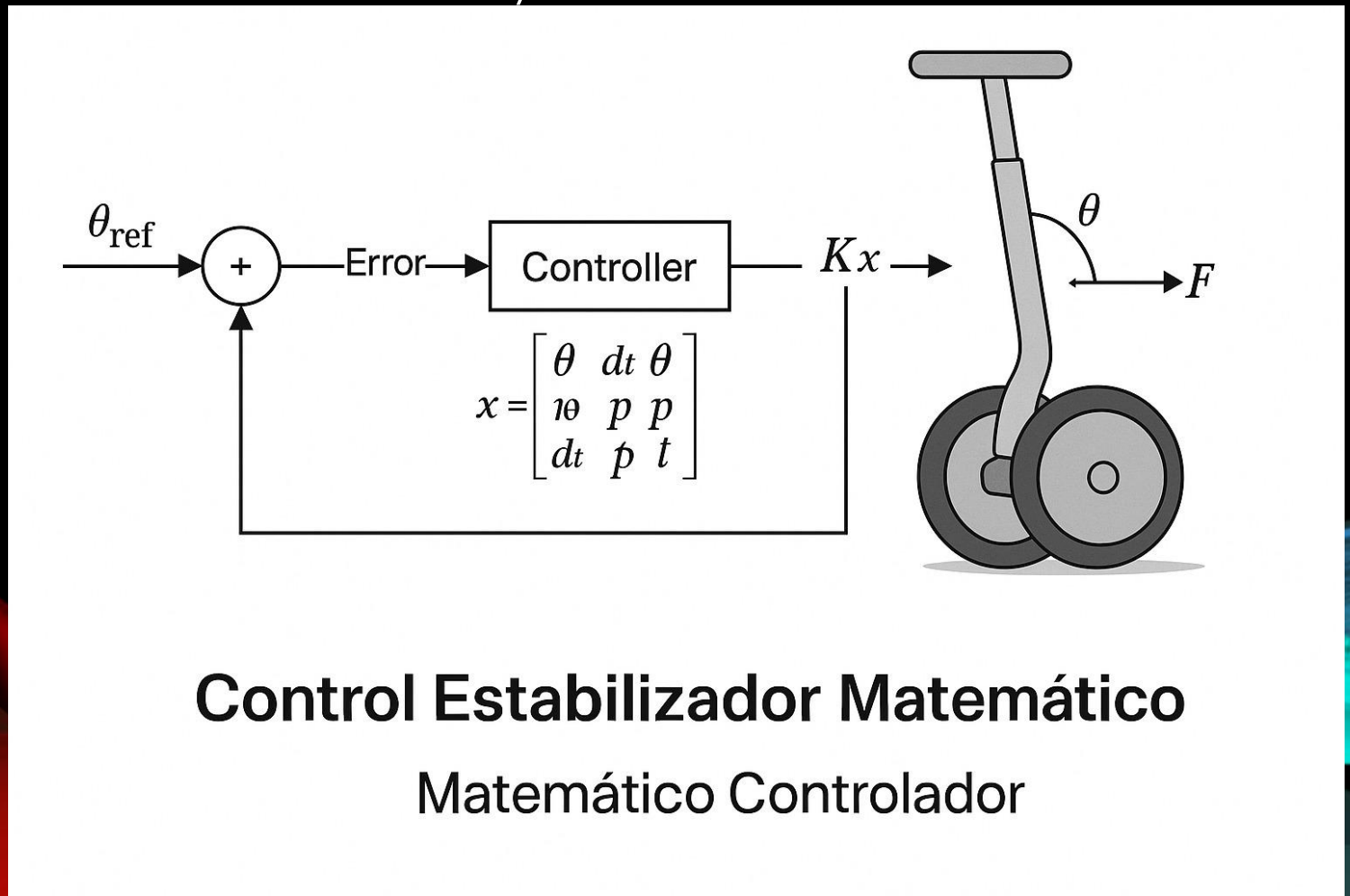
ANÁLISIS DE LA MOVILIDAD URBANA CON SCOOTERS

Nombre: Alberto Armando Huerta Ornelas

Fecha: 12 de agosto del 2025

UEA: Análisis de datos en Python



CONTEXTO Y MOTIVACIÓN

Problema real: "En 2020, los scooters eléctricos en Chicago (operados por Spin, Bird, Lime) generaron >157k viajes, pero con desafíos como concentración en zonas urbanas, impacto en tráfico y necesidad de redistribución eficiente post-COVID".

Por qué importa: "Ayuda a ciudades a promover micromovilidad sostenible, reducir congestión y optimizar recursos (e.g., estaciones de carga en hotspots como Lake View)".

Aporte de la charla: "Esta presentación revela patrones visuales (flujos, distribuciones) para informar políticas urbanas y estrategias de vendors".

OBJETIVOS/PREGUNTAS

- ¿Cuáles son las top 10 zonas de inicio/fin y sus flujos? .
- ¿Cómo se distribuyen distancias/duraciones por vendor?
- ¿Existen trayectos circulares?
- ¿Existen patrones temporales (hora/mes)?
- ¿Dónde se concentran geográficamente los viajes?

DATOS: FUENTES Y VARIABLES

- Fuente: 'Scooter_Trips_2020.csv' (Chicago Data Portal, periodo: Ago-Dic 2020).
- Tamaño: 157,294 rows x 13 cols.
- Variables clave: Date/Hour (temporal), Trip Distance/Duration (métricas), Vendor (categórico), Start/End Community Area Name/Number (espacial), Centroid Lat/Lon (geo).

	Date	Hour	Trip Distance	Trip Duration	Vendor	Start Community Area Number	End Community Area Number	Start Community Area Name	End Community Area Name	Start Centroid Latitude	Start Centroid Longitude	End Centroid Latitude	End Centroid Longitude
0	08/12/2020	5	5	21	spin	31.0	31.0	LOWER WEST SIDE	LOWER WEST SIDE	41.848335	-87.675179	41.848335	-87.675179
1	08/12/2020	7	13	101	spin	7.0	7.0	LINCOLN PARK	LINCOLN PARK	41.921880	-87.645647	41.921880	-87.645647
2	08/12/2020	7	7	50	bird	77.0	77.0	EDGEWATER	EDGEWATER	41.987114	-87.664343	41.987114	-87.664343
3	08/12/2020	7	3815	840	spin	6.0	3.0	LAKE VIEW	UPTOWN	41.943514	-87.657498	41.965435	-87.655145
4	08/12/2020	8	1444	445	spin	3.0	6.0	UPTOWN	LAKE VIEW	41.965435	-87.655145	41.943514	-87.657498
...	...	...	...	...	...	...	...	...	...	...	...	...	...
157289	12/12/2020	21	335	186	lime	23.0	23.0	HUMBOLDT PARK	HUMBOLDT PARK	41.900813	-87.723955	41.900813	-87.723955
157290	12/12/2020	21	2704	1254	lime	37.0	61.0	FULLER PARK	NEW CITY	41.813368	-87.632599	41.808705	-87.657612
157291	12/12/2020	21	9257	2214	spin	6.0	6.0	LAKE VIEW	LAKE VIEW	41.943514	-87.657498	41.943514	-87.657498
157292	12/12/2020	21	878	325	lime	28.0	24.0	NEAR WEST SIDE	WEST TOWN	41.874254	-87.664619	41.901459	-87.675568
157293	12/12/2020	21	490	212	lime	8.0	8.0	NEAR NORTH SIDE	NEAR NORTH SIDE	41.899528	-87.633571	41.899528	-87.633571

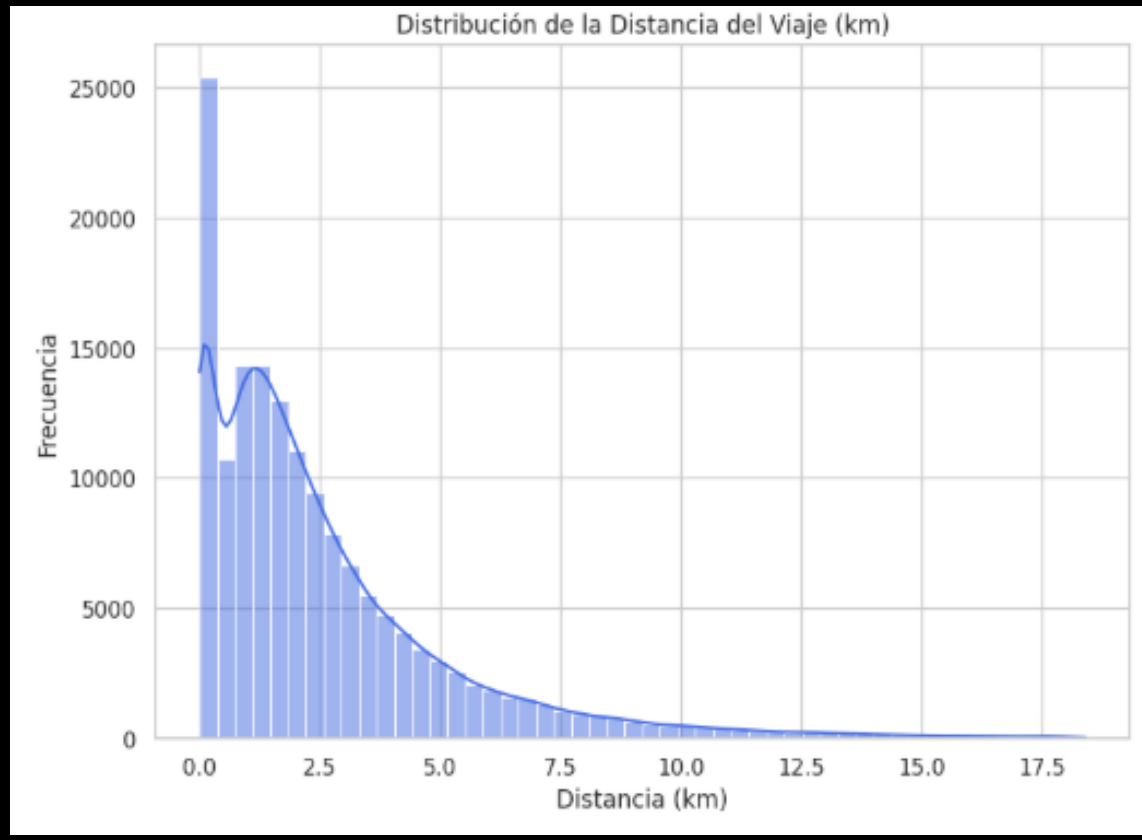
157294 rows x 13 columns

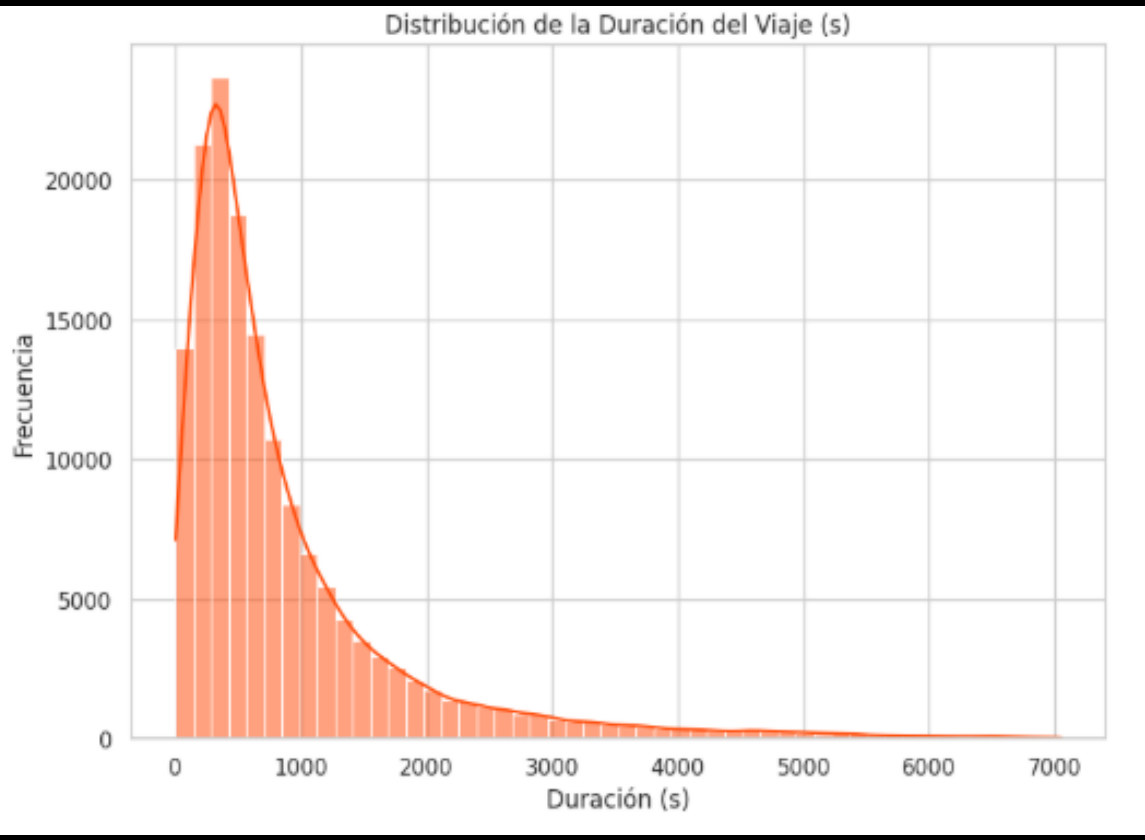
DATOS: PREPROCESAMIENTO Y CALIDAD

•Limpieza: Carga con pd.read_csv; no filtrado explícito, pero value_counts y groupby implícitos.
•Cómo/Por qué: Verificar NaNs (ninguno visible); extraer Month de Date para temporales; groupby para flujos (top 50 en Sankey). Para calidad: Correlaciones solo en numéricas.
•Issues: Posibles outliers en Distance (e.g., 9257m); unidades asumidas (metros/segundos).

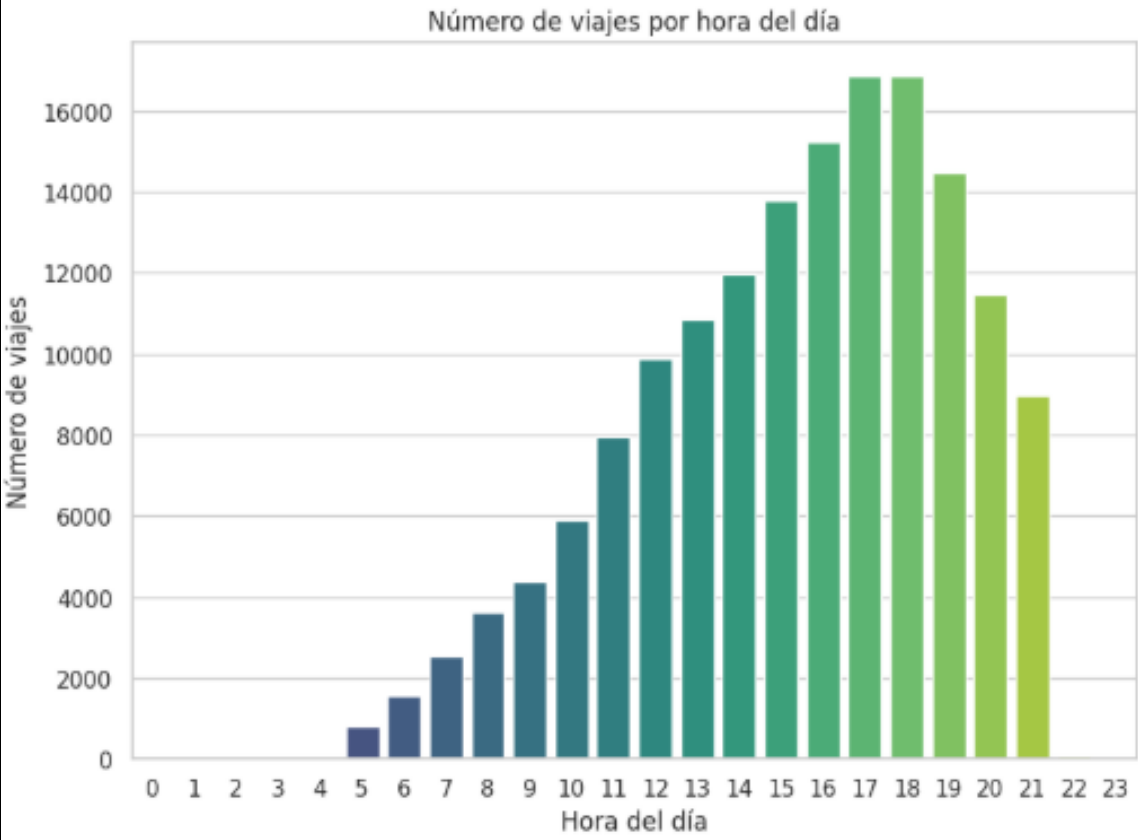
	Date	Hour	Trip Distance	Trip Duration	Vendor	Start Community Area Number	End Community Area Number	Start Community Area Name	End Community Area Name	Start Centroid Latitude	Start Centroid Longitude	End Centroid Latitude	End Centroid Longitude	Dia_hora	Dia_semana
0	08/12/2020	5	5	21	spin	31.0	31.0	LOWER WEST SIDE	LOWER WEST SIDE	41.848335	-87.675179	41.848335	-87.675179	2020-08-12 05:00:00	Wednesday
1	08/12/2020	7	13	101	spin	7.0	7.0	LINCOLN PARK	LINCOLN PARK	41.921880	-87.645647	41.921880	-87.645647	2020-08-12 07:00:00	Wednesday
2	08/12/2020	7	7	50	bird	77.0	77.0	EDGEWATER	EDGEWATER	41.987114	-87.664343	41.987114	-87.664343	2020-08-12 07:00:00	Wednesday
3	08/12/2020	7	3815	840	spin	6.0	3.0	LAKE VIEW	UPTOWN	41.943514	-87.657498	41.965435	-87.655145	2020-08-12 07:00:00	Wednesday
4	08/12/2020	8	1444	445	spin	3.0	6.0	UPTOWN	LAKE VIEW	41.965435	-87.655145	41.943514	-87.657498	2020-08-12 08:00:00	Wednesday

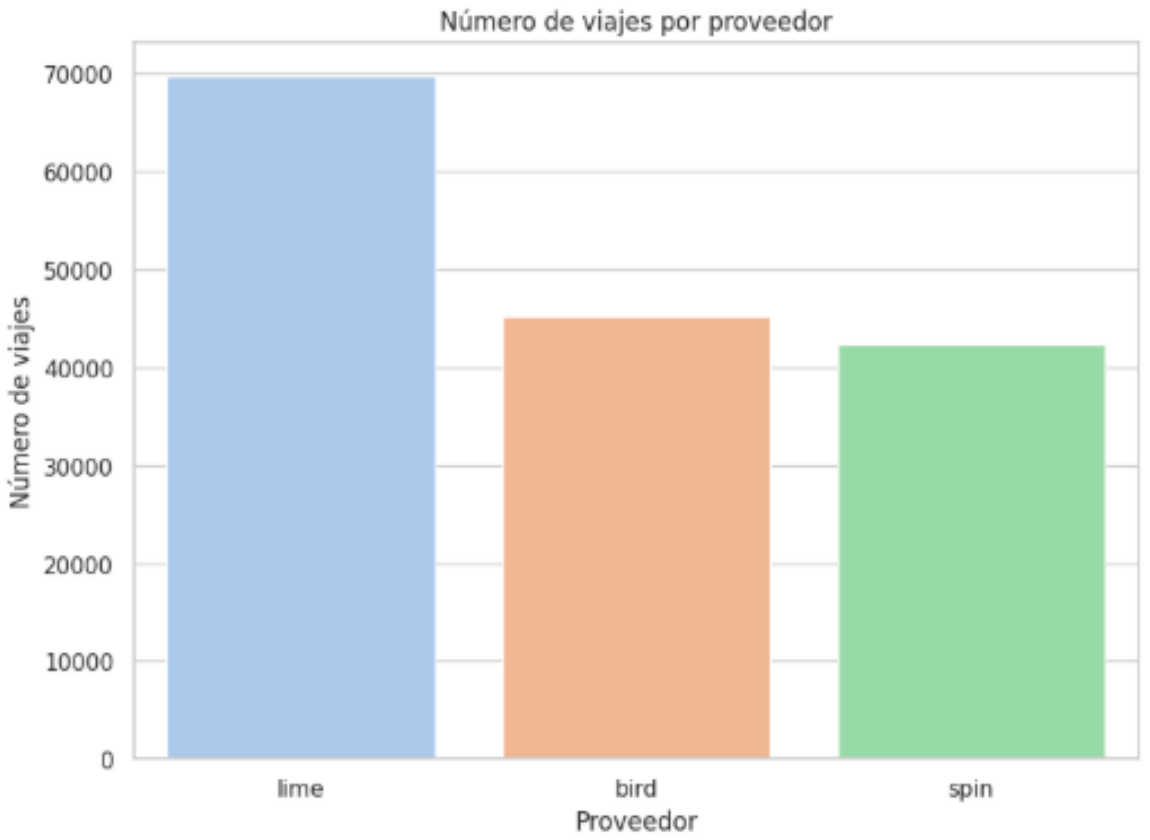
METODOLOGÍA: ANÁLISIS EXPLORATORIO BÁSICO



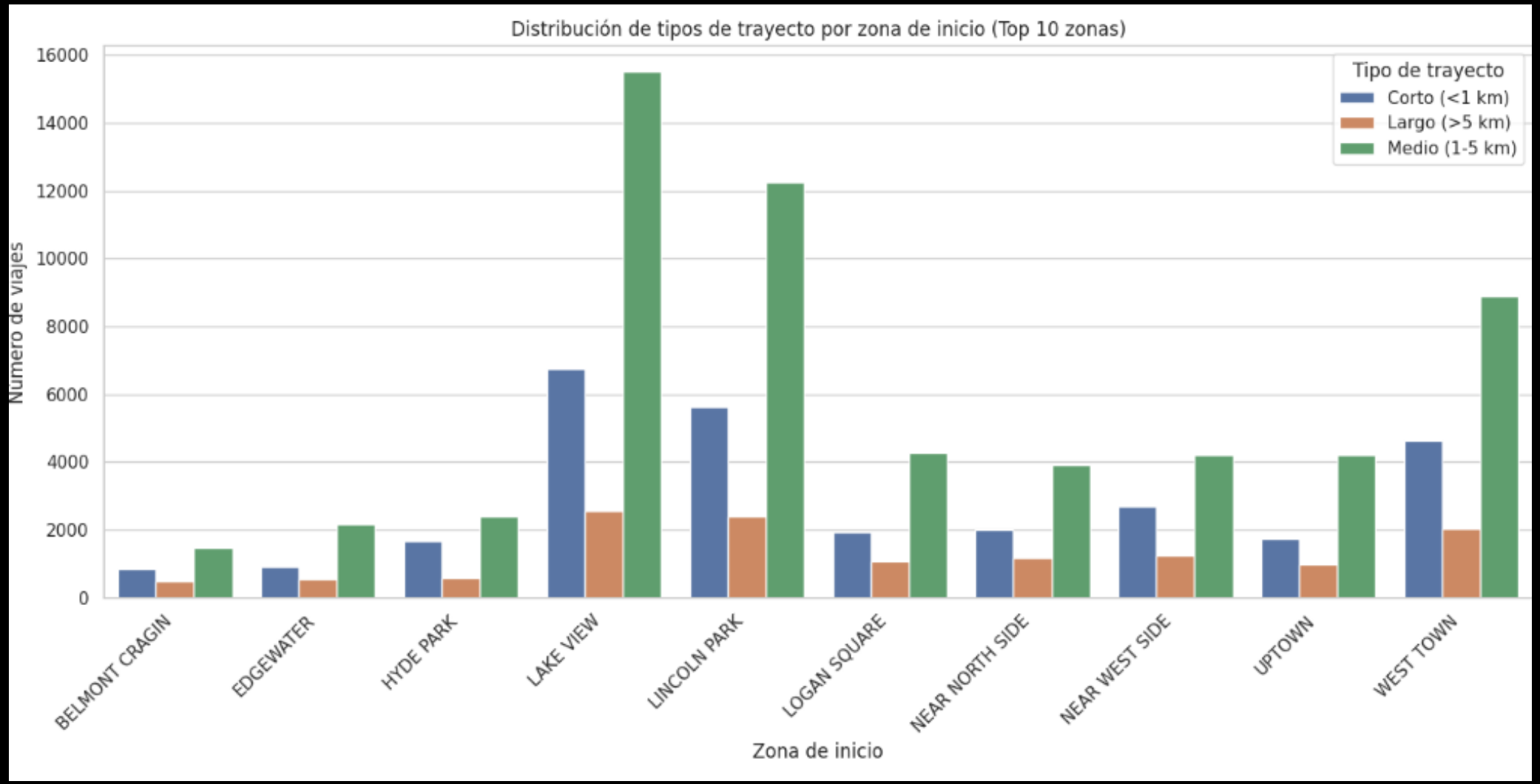


METODOLOGÍA: COMPARACIONES POR PROOVEDOR

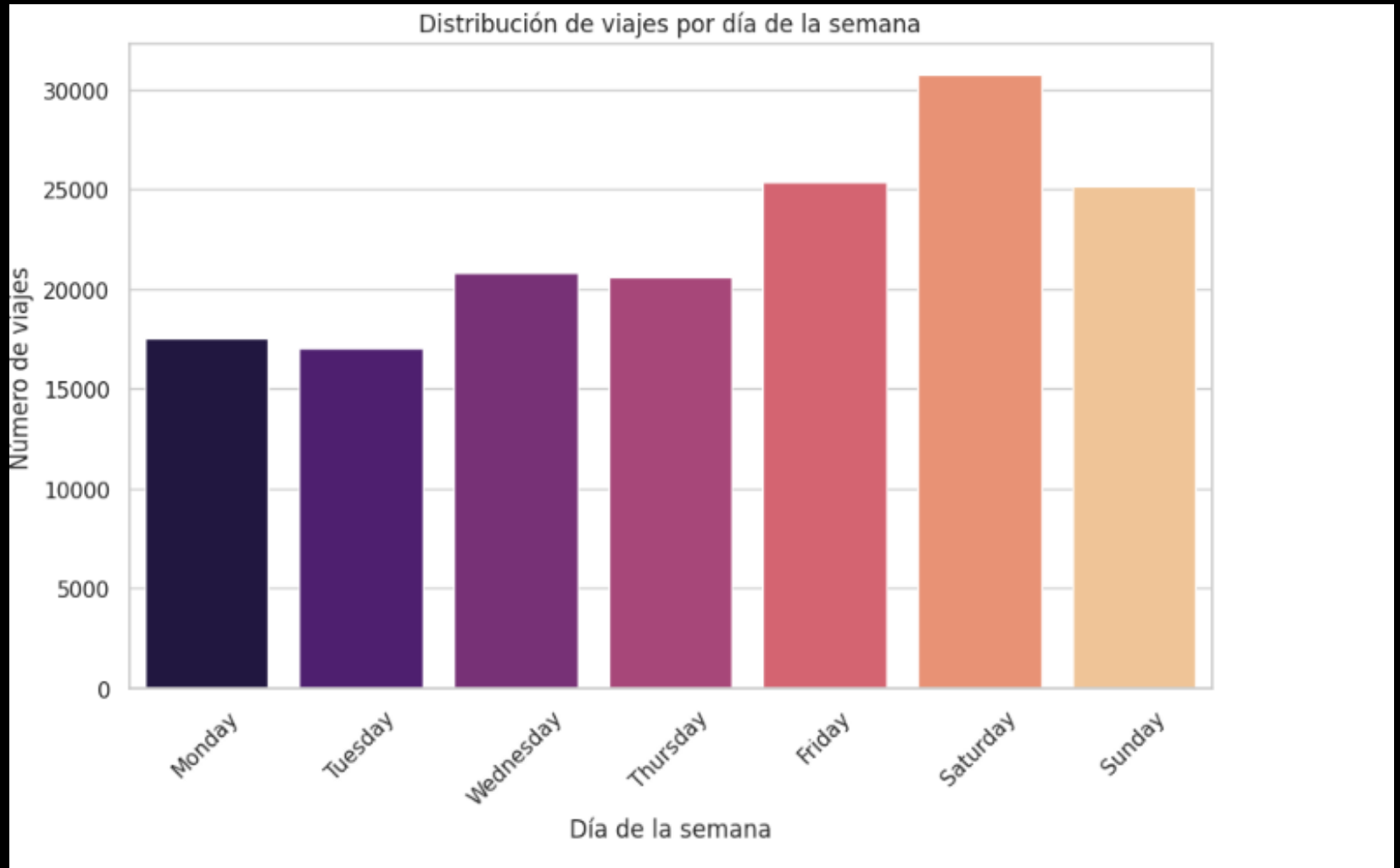




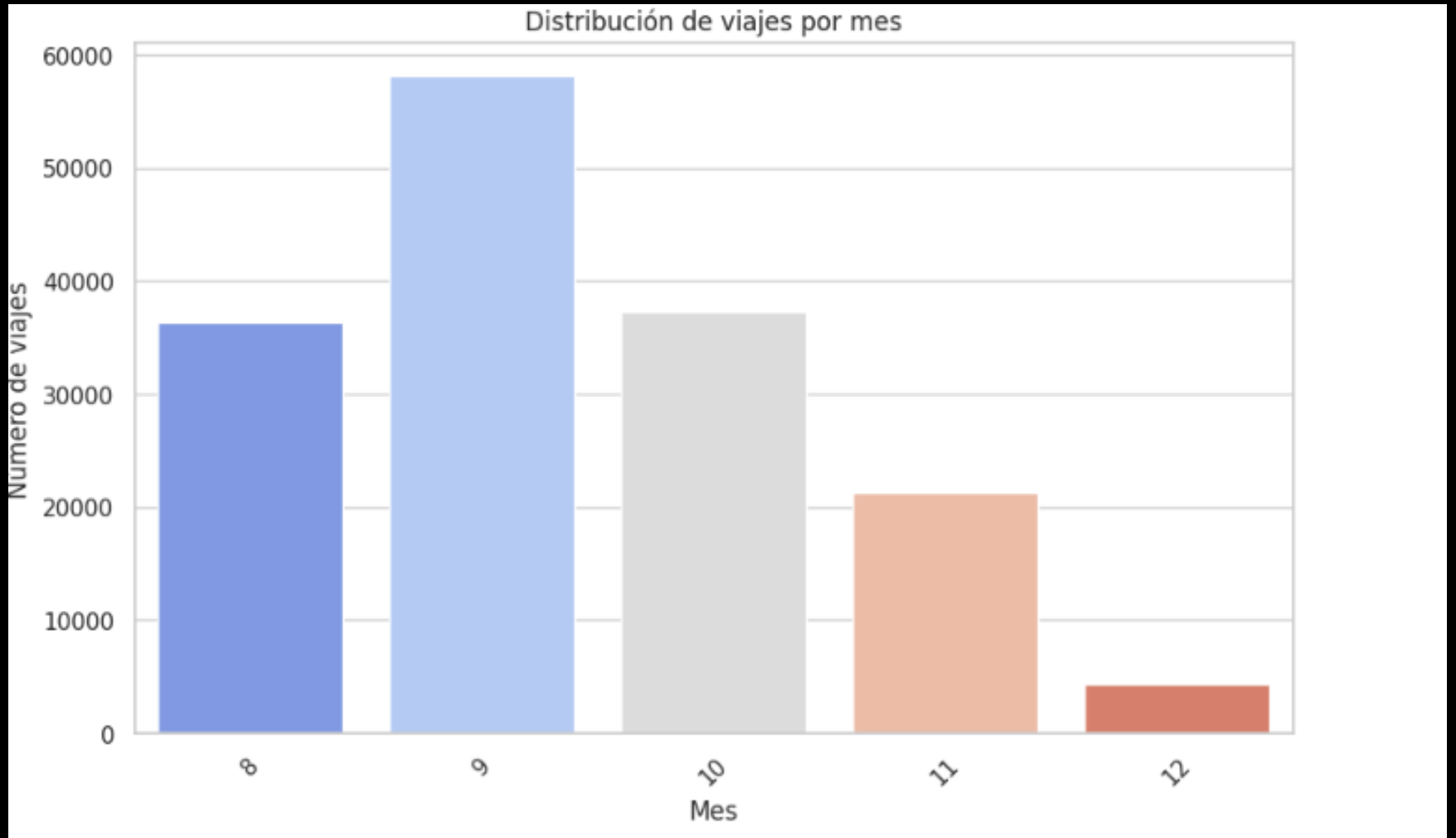
METODOLOGÍA: CLASIFICACIÓN DE VIAJES EN ALGUNAS ZONAS



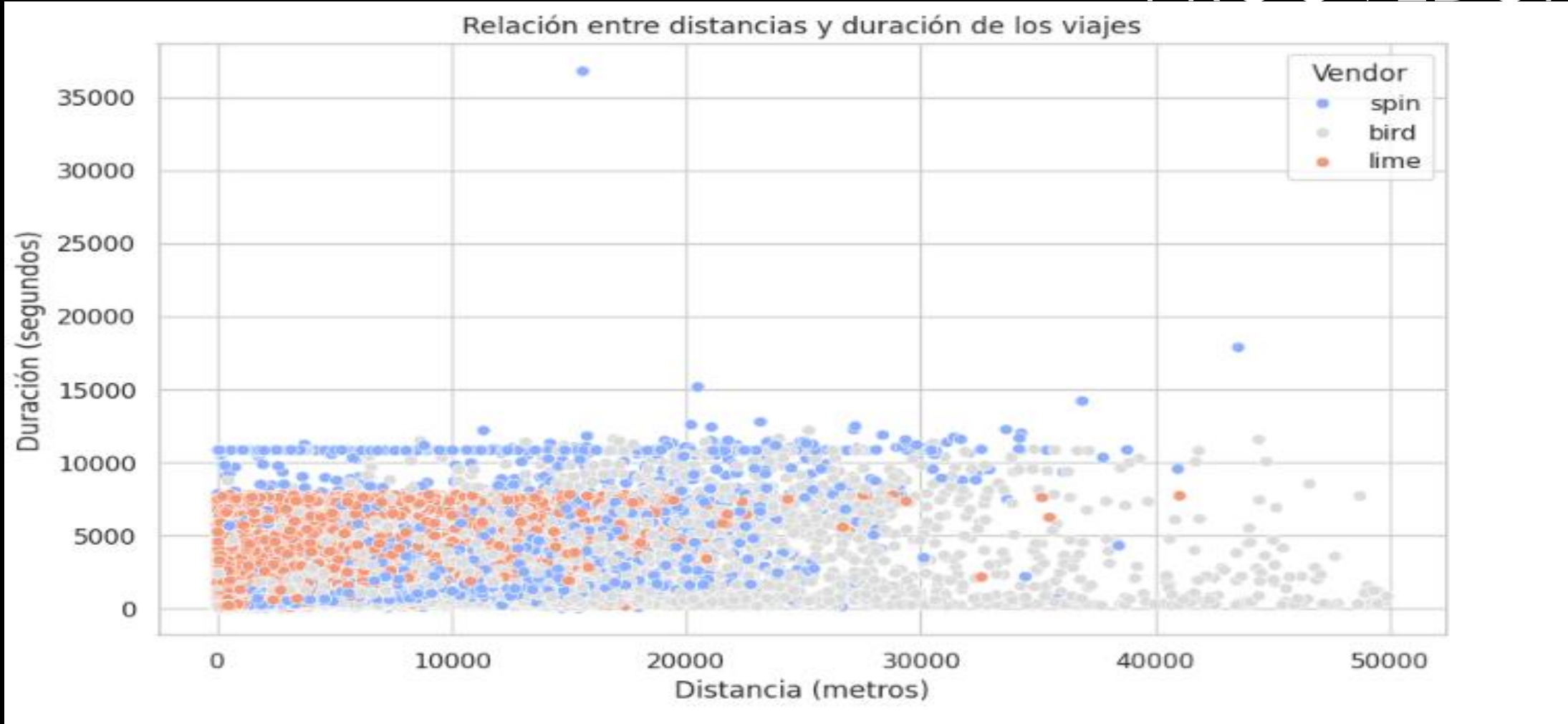
METODOLOGÍA: DISTRIBUCIÓN DE VIAJES POR DÍA



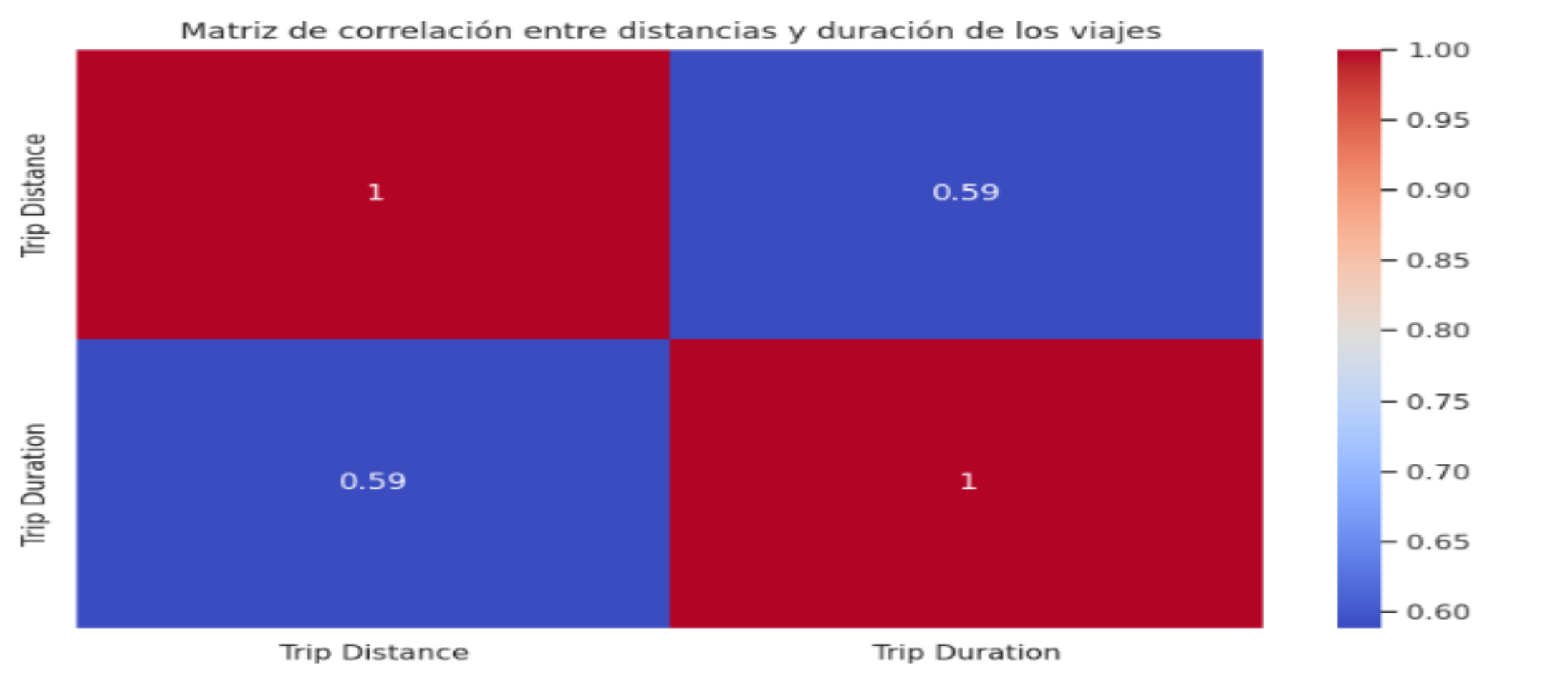
METODOLOGÍA: DISTRIBUCIÓN DE VIAJES POR MES



RELACIÓN ENTRE DISTANCIA Y DURACIÓN DE VIAJES POR PROOVEDOR



CORRELACIONE ENTRE DISTANCIA Y TIEMPO DE DURACION



ANÁLISIS DE USO POR ZONAS

```
Top 10 Zonas de inicio:
Start Community Area Name
LAKE VIEW           24816
LINCOLN PARK        20246
WEST TOWN           15557
NEAR WEST SIDE       8147
LOGAN SQUARE         7299
NEAR NORTH SIDE      7123
UPTOWN               6930
HYDE PARK            4660
EDGEWATER            3656
BELMONT CRAGIN       2802
Name: count, dtype: int64

Top 10 Zonas de fin:
End Community Area Name
LAKE VIEW           24686
LINCOLN PARK        19818
WEST TOWN           15540
NEAR WEST SIDE       8120
LOGAN SQUARE         7382
NEAR NORTH SIDE      7271
UPTOWN               6768
HYDE PARK            4553
EDGEWATER            3711
BELMONT CRAGIN       2712
Name: count, dtype: int64
```

INGENIERÍA DE VARIABLES

```
# codificar las variables categoricas

le_vendor=LabelEncoder()
le_dia=LabelEncoder()
le_mes=LabelEncoder()
le_start_zone=LabelEncoder()
le_end_zone=LabelEncoder()
```

```
df['Vendor_Enconder']=le_vendor.fit_transform(df['Vendor'])
df['Dia_semana_Encoded']=le_dia.fit_transform(df['Dia_semana'])
df['Dia_mes_Encoded']=le_dia.fit_transform(df['Mes'])
df['Start_zone_Encoded']=le_start_zone.fit_transform(df['Start Community Area Name'])
df['End_zone_Encoded']=le_end_zone.fit_transform(df['End Community Area Name'])
```

```
features=['Trip Distance', 'Vendor_Enconder', 'Hora_dia', 'Dia_semana_Encoded', 'Dia_mes_Encoded', 'Start_zone_Encoded', 'End_zone_Encoded', ]
X=df[features]
y=df['Trip Duration']
```

```
X_train, X_test, y_train, y_test=train_test_split(X, y, test_size=0.2, random_state=42)
```

USO DE RANDOMFOREST

```
y_pred=rf_model.predict(X_test)
mse=mean_squared_error(y_test, y_pred)
r2=r2_score(y_test, y_pred)
print("\nResultados del Modelo Random Forest:")
print(f"Error Cuadratico Medio (MSE): {mse:2f}")
print(f"Coeficiente de Determinacion (R^2): {r2:2f}")
```

```
Resultados del Modelo Random Forest:
Error Cuadratico Medio (MSE): 944685.262172
Coeficiente de Determinacion (R^2): 0.429560
```

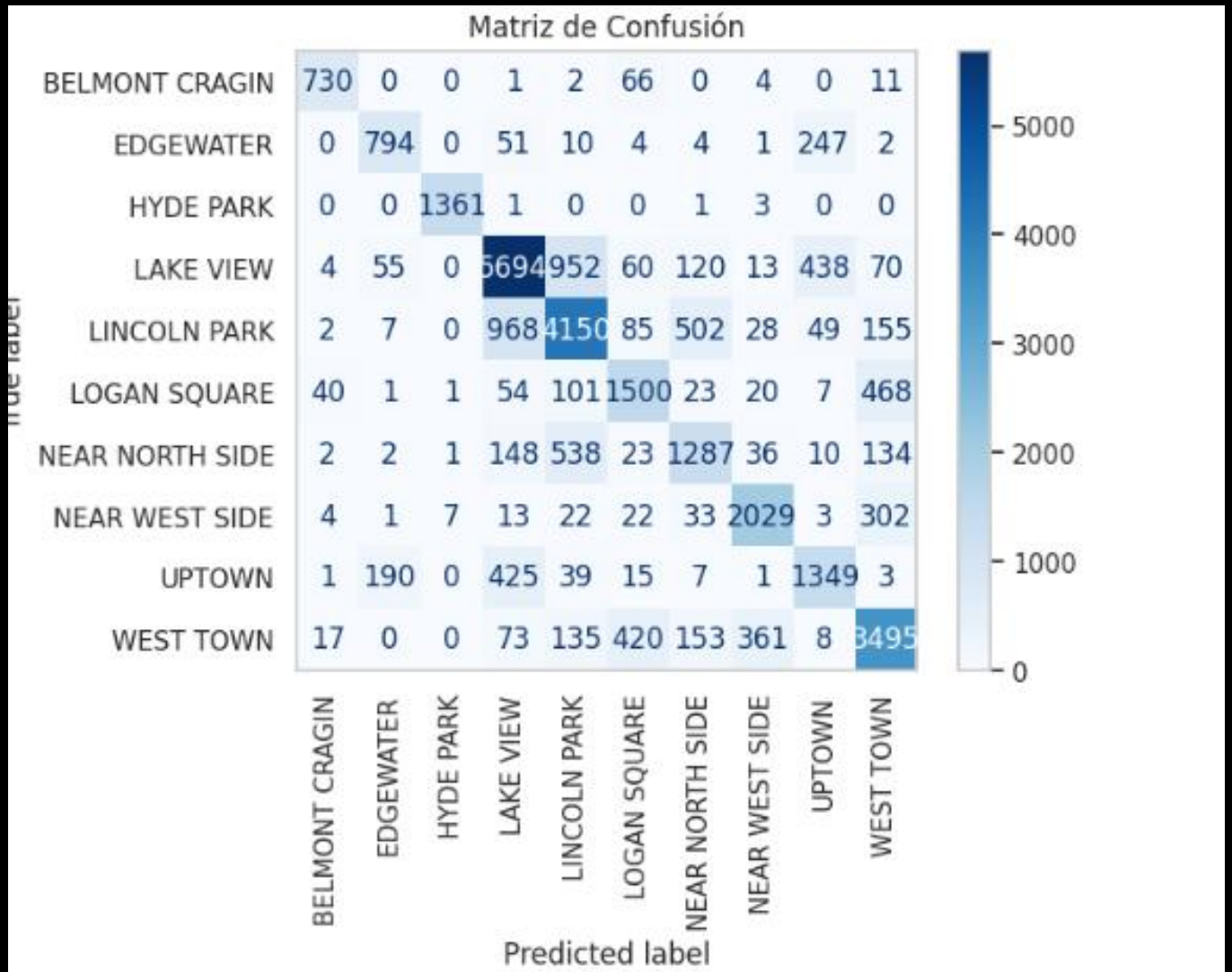
```
feature_importance=pd.DataFrame({'feature': features, 'Importance': rf_model.feature_importances_})
feature_importance=feature_importance.sort_values('Importance', ascending=False)
print("\nImportancia de las caracteristicas:")
print(feature_importance)
```

```
Importancia de las caracteristicas:
              feature  Importance
0       Trip Distance    0.594969
6    End_zone_Encoded    0.092476
5  Start_zone_Encoded    0.085248
2           Hora_dia    0.083084
3  Dia_semana_Encoded    0.060177
4     Dia_mes_Encoded    0.048422
1     Vendor_Enconder    0.035623
```

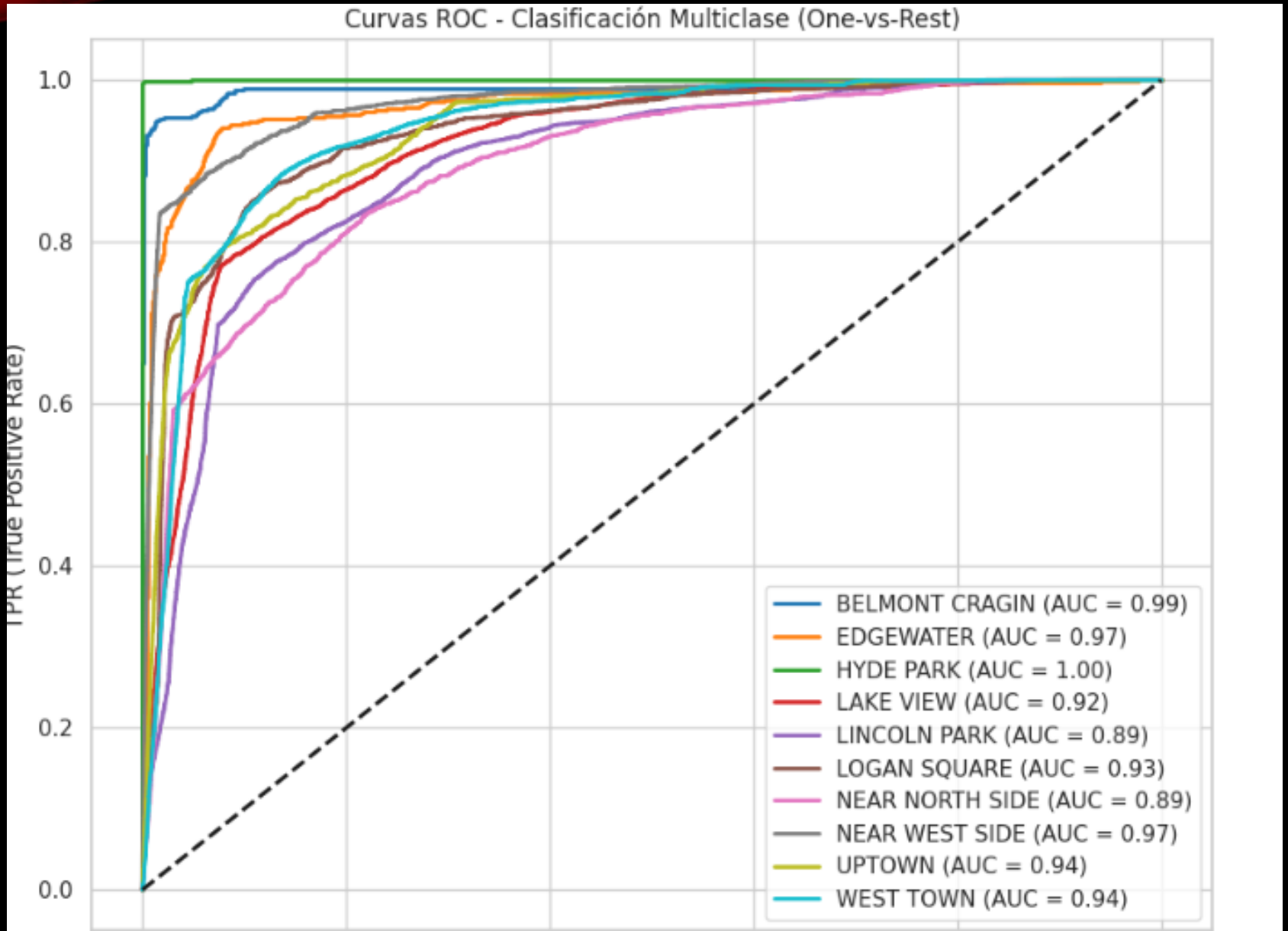
MATRIZ DE MÉTRICAS DE CLASIFICACIÓN

	precision	recall	f1-score	support
BELMONT CRAGIN	0.91	0.90	0.90	814
EDGEWATER	0.76	0.71	0.73	1113
HYDE PARK	0.99	1.00	0.99	1366
LAKE VIEW	0.77	0.77	0.77	7406
LINCOLN PARK	0.70	0.70	0.70	5946
LOGAN SQUARE	0.68	0.68	0.68	2215
NEAR NORTH SIDE	0.60	0.59	0.60	2181
NEAR WEST SIDE	0.81	0.83	0.82	2436
UPTOWN	0.64	0.66	0.65	2030
WEST TOWN	0.75	0.75	0.75	4662
accuracy			0.74	30169
macro avg	0.76	0.76	0.76	30169
weighted avg	0.74	0.74	0.74	30169

MATRIZ DE CONFUSIÓN



CURVAS ROC POR ZONAS



TRAYECTOS CIRCULARES

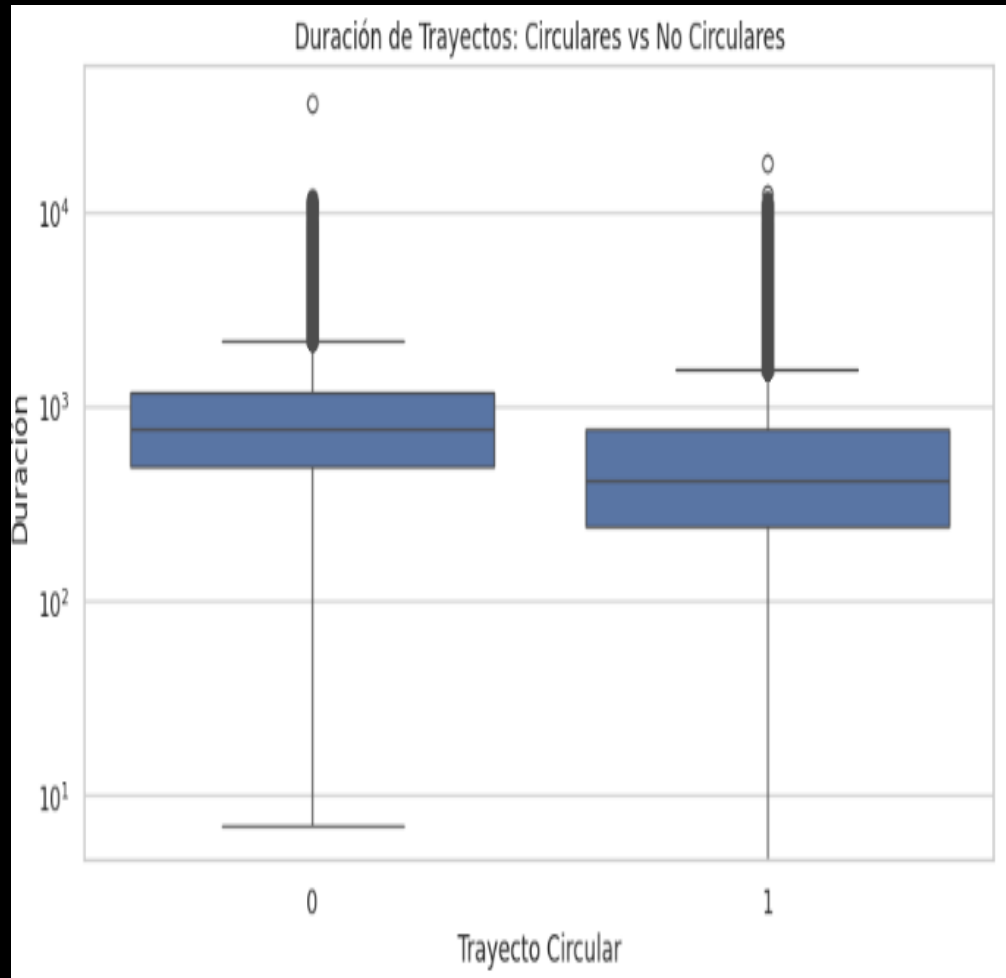
```
conteo_circulares = df['TrayectoCircular'].value_counts()
print(conteo_circulares)

# %
porcentaje_circulares = df['TrayectoCircular'].mean() * 100
print(f"Porcentaje de trayectos circulares: {porcentaje_circulares:.2f}%")
```

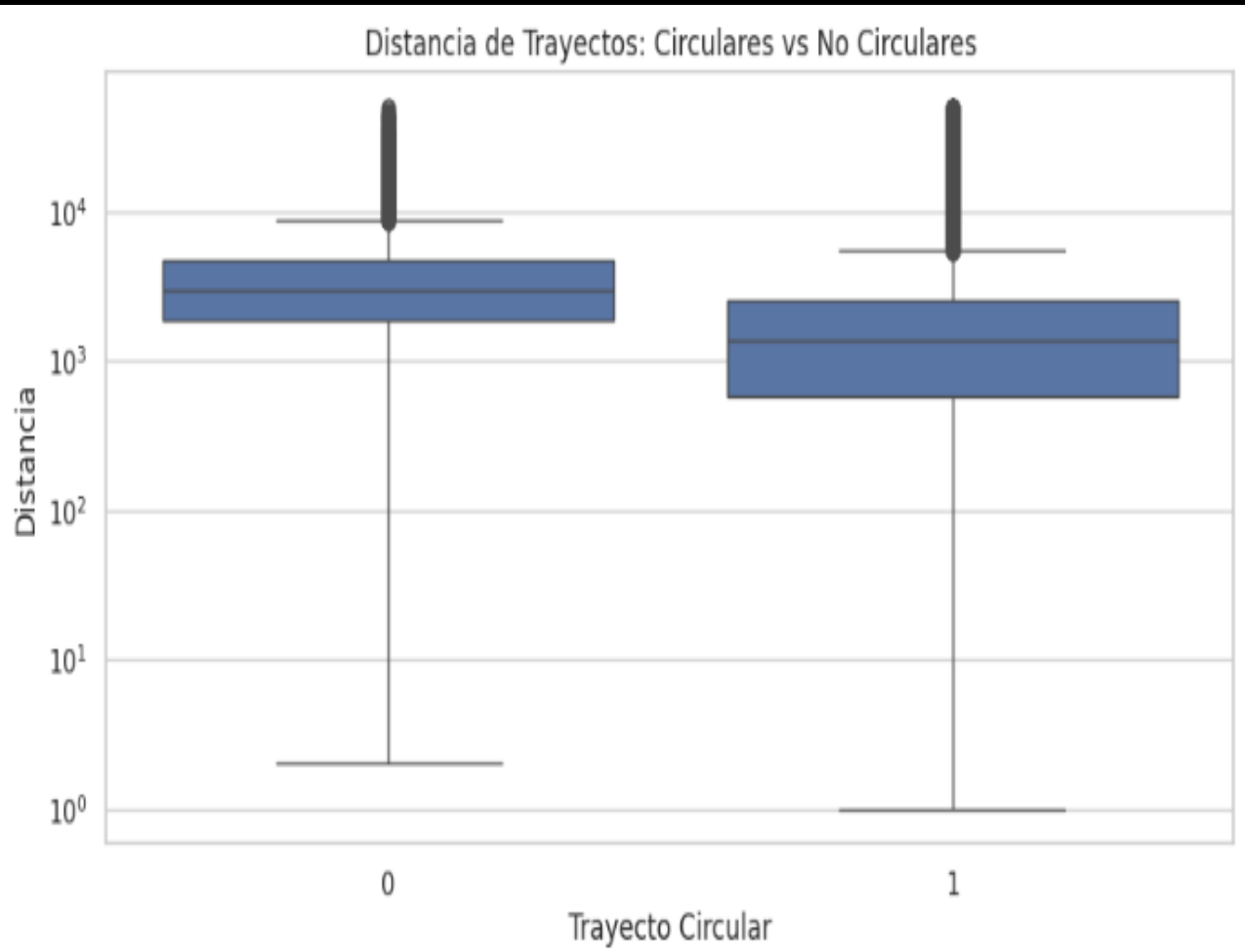
```
TrayectoCircular
True     71166
False    29395
Name: count, dtype: int64
Porcentaje de trayectos circulares: 70.77%
```

```
# ¿Desde qué zonas se realizan más trayectos circulares?
circular_por_zona = df[df['TrayectoCircular']].groupby('Start Community Area Name').size().sort_values(ascending=False)
print(circular_por_zona.head(10))
```

```
Start Community Area Name
LAKE VIEW          18399
LINCOLN PARK       13877
WEST TOWN          11378
NEAR WEST SIDE      6280
LOGAN SQUARE        4446
UPTOWN              4314
NEAR NORTH SIDE     4260
HYDE PARK           3836
EDGEWATER           2322
BELMONT CRAGIN      2054
dtype: int64
```

DURACIÓN Y DISTANCIA DE VIAJES EN TRAYECTOS CIRCULARES



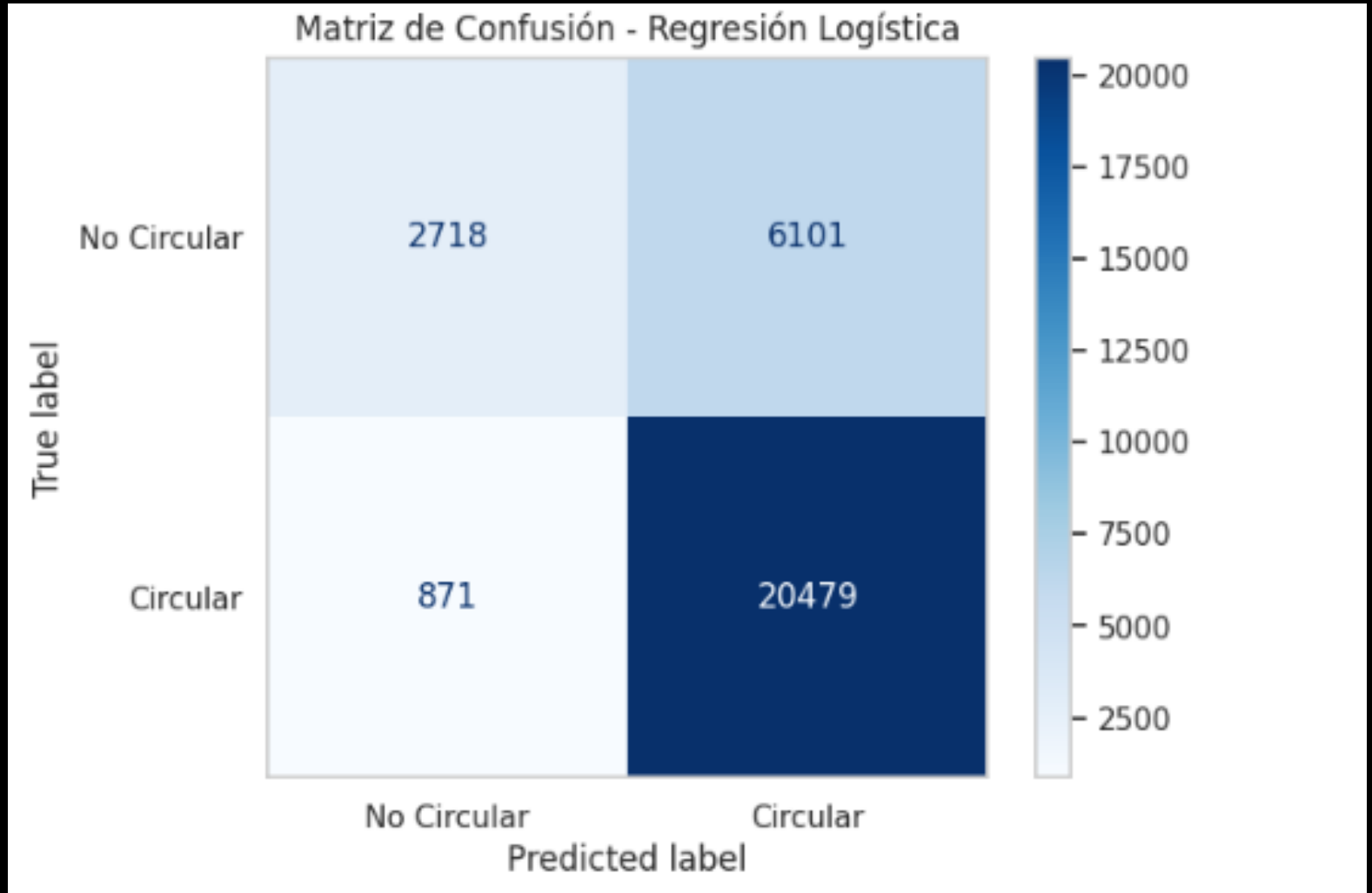


MATRIZ DE MÉTRICAS DE TRAYECTOS CIRCULARES

Regresión Logística

	precision	recall	f1-score	support
0	0.76	0.31	0.44	8819
1	0.77	0.96	0.85	21350
accuracy			0.77	30169
macro avg	0.76	0.63	0.65	30169
weighted avg	0.77	0.77	0.73	30169

MATRIZ DE CONFUSIÓN DE LOS TRAYECTOS CIRCULARES

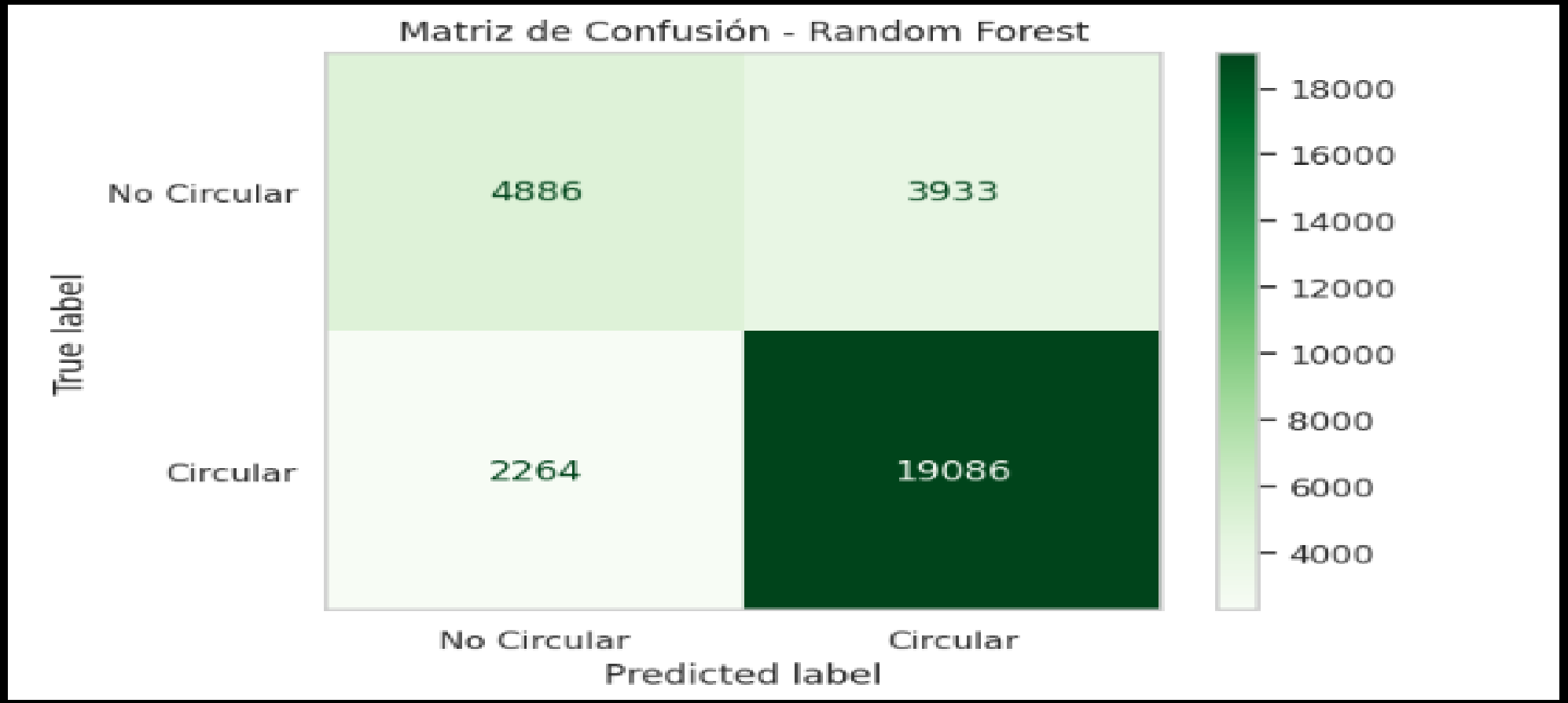


MATRIZ DE MÉTRICAS DE TRAYECTOS CIRCULARES RANDOMFOREST

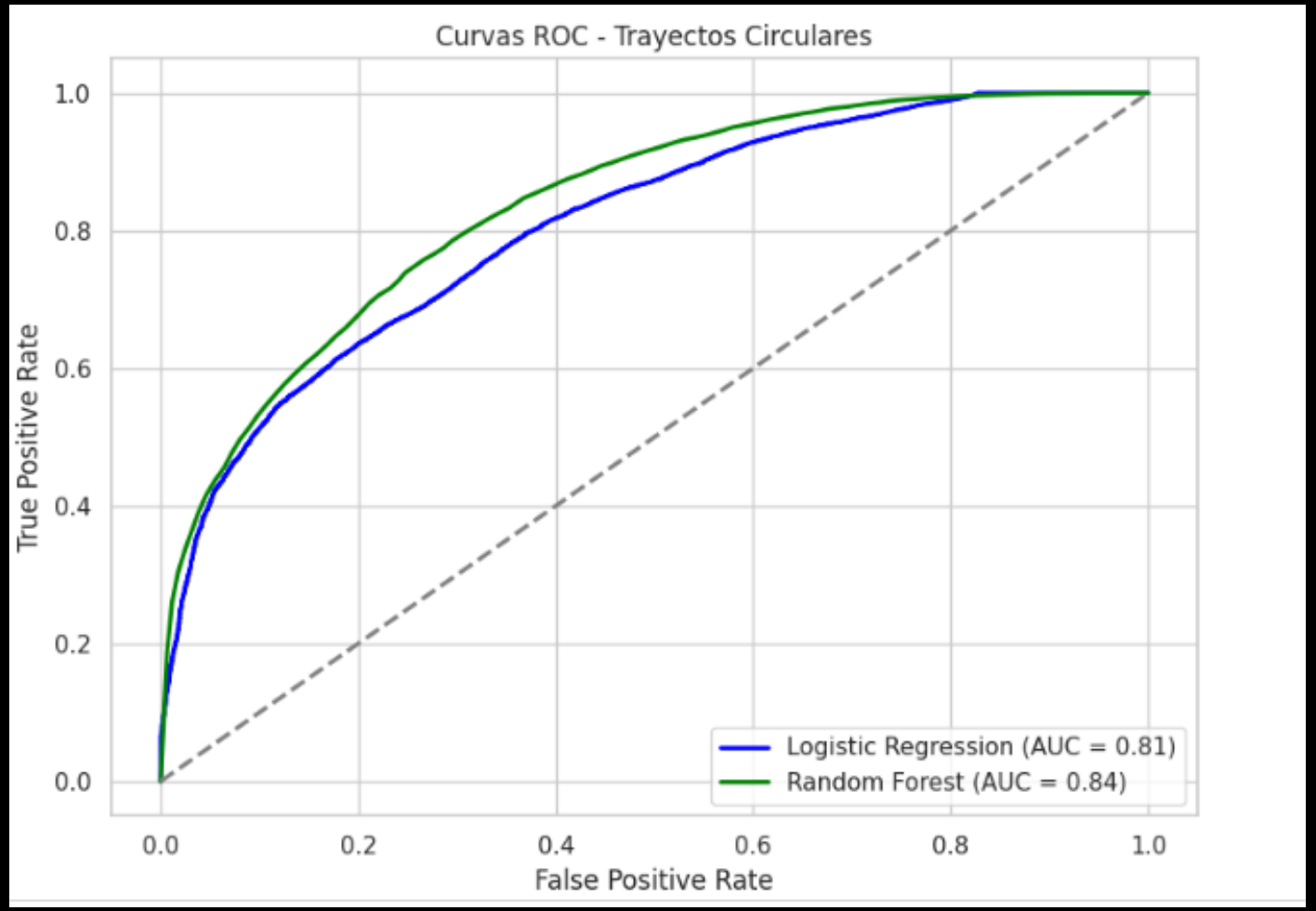
Random Forest

	precision	recall	f1-score	support
0	0.68	0.55	0.61	8819
1	0.83	0.89	0.86	21350
accuracy			0.79	30169
macro avg	0.76	0.72	0.74	30169
weighted avg	0.79	0.79	0.79	30169

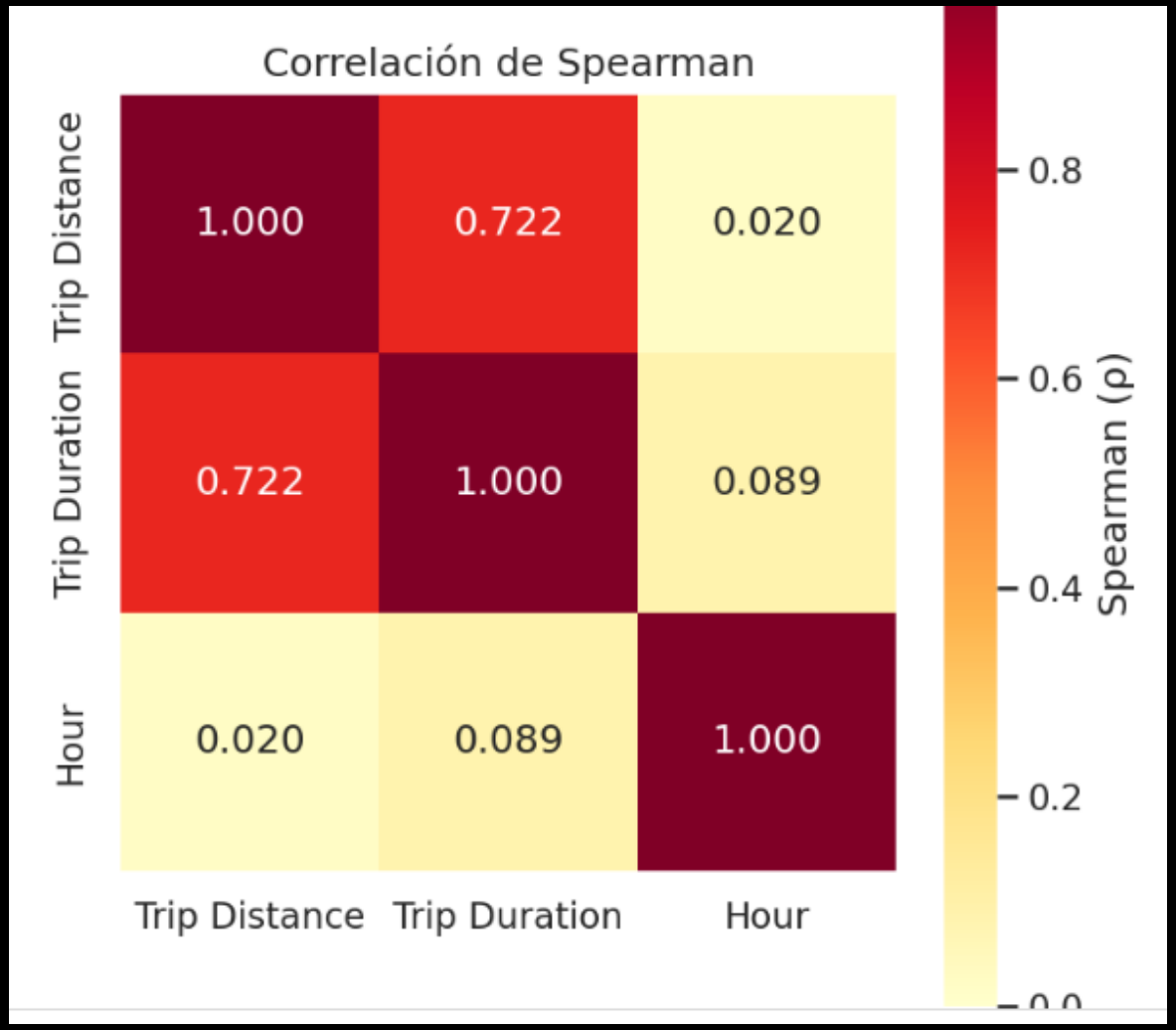
MATRIZ DE CONFUSIÓN DE TRAYECTOS CIRCULARES RANDOMFOREST



CURVA ROC DE TRAYECTOS CIRCULARES



MATRIZ POR SPEARMAN



CLASIFICACIÓN POR TIPO DE VIAJE

	Date	Hour	Trip Distance	Trip Duration	Vendor	Start Community Area Number	End Community Area Number	Start Community Area Name	End Community Area Name	Start Centroid Latitude	Start Centroid Longitude	End Centroid Latitude	End Centroid Longitude	Trip Distance (km)	Tipo de Trayecto
0	08/12/2020	5	5	21	spin	31.0	31.0	LOWER WEST SIDE	LOWER WEST SIDE	41.848335	-87.675179	41.848335	-87.675179	0.005	Corto (<1 km)
1	08/12/2020	7	13	101	spin	7.0	7.0	LINCOLN PARK	LINCOLN PARK	41.921880	-87.645647	41.921880	-87.645647	0.013	Corto (<1 km)
2	08/12/2020	7	7	50	bird	77.0	77.0	EDGEWATER	EDGEWATER	41.987114	-87.664343	41.987114	-87.664343	0.007	Corto (<1 km)
3	08/12/2020	7	3815	840	spin	6.0	3.0	LAKE VIEW	UPTOWN	41.943514	-87.657498	41.965435	-87.655145	3.815	Medio (1-5 km)
4	08/12/2020	8	1444	445	spin	3.0	6.0	UPTOWN	LAKE VIEW	41.965435	-87.655145	41.943514	-87.657498	1.444	Medio (1-5 km)
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
157289	12/12/2020	21	335	186	lime	23.0	23.0	HUMBOLDT PARK	HUMBOLDT PARK	41.900813	-87.723955	41.900813	-87.723955	0.335	Corto (<1 km)
157290	12/12/2020	21	2704	1254	lime	37.0	61.0	FULLER PARK	NEW CITY	41.813368	-87.632599	41.808705	-87.657612	2.704	Medio (1-5 km)
157291	12/12/2020	21	9257	2214	spin	6.0	6.0	LAKE VIEW	LAKE VIEW	41.943514	-87.657498	41.943514	-87.657498	9.257	Largo (>5 km)
157292	12/12/2020	21	878	325	lime	28.0	24.0	NEAR WEST SIDE	WEST TOWN	41.874254	-87.664619	41.901459	-87.675568	0.878	Corto (<1 km)
157293	12/12/2020	21	490	212	lime	8.0	8.0	NEAR NORTH SIDE	NEAR NORTH SIDE	41.899528	-87.633571	41.899528	-87.633571	0.490	Corto (<1 km)

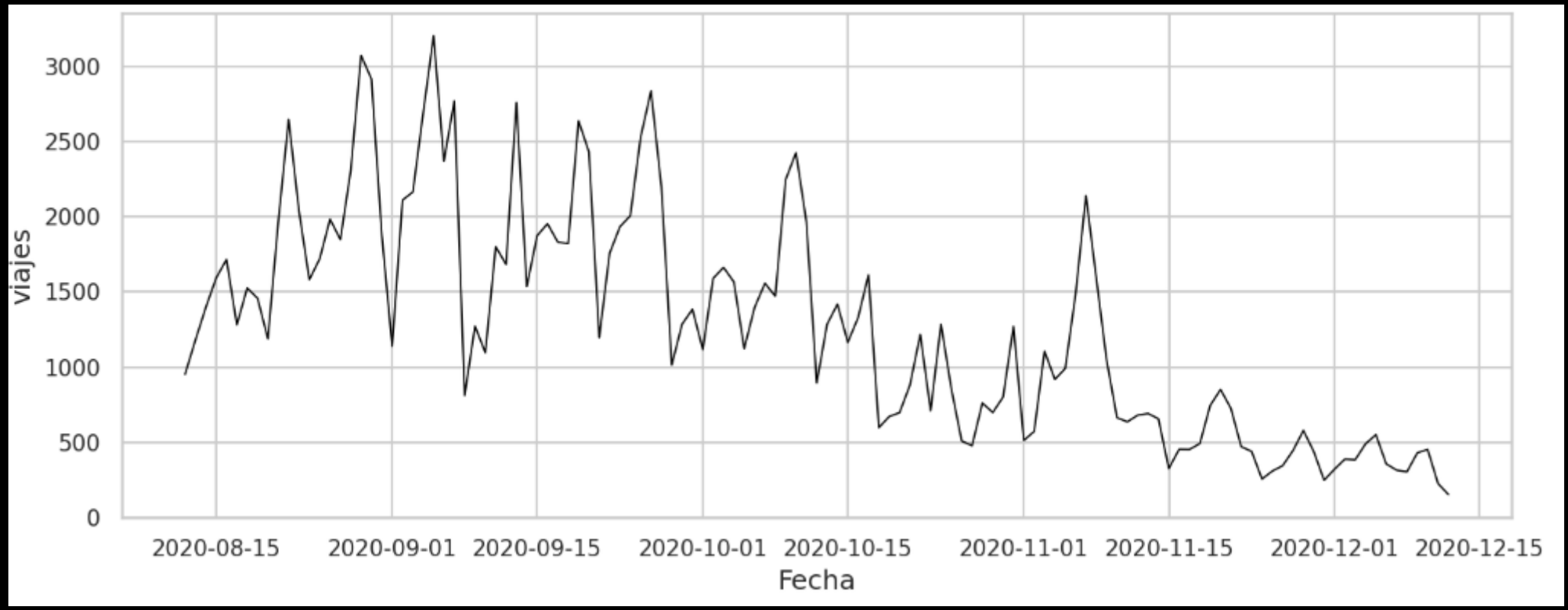
157294 rows × 15 columns

CANTIDAD DE VIAJES POR TRAYECTO

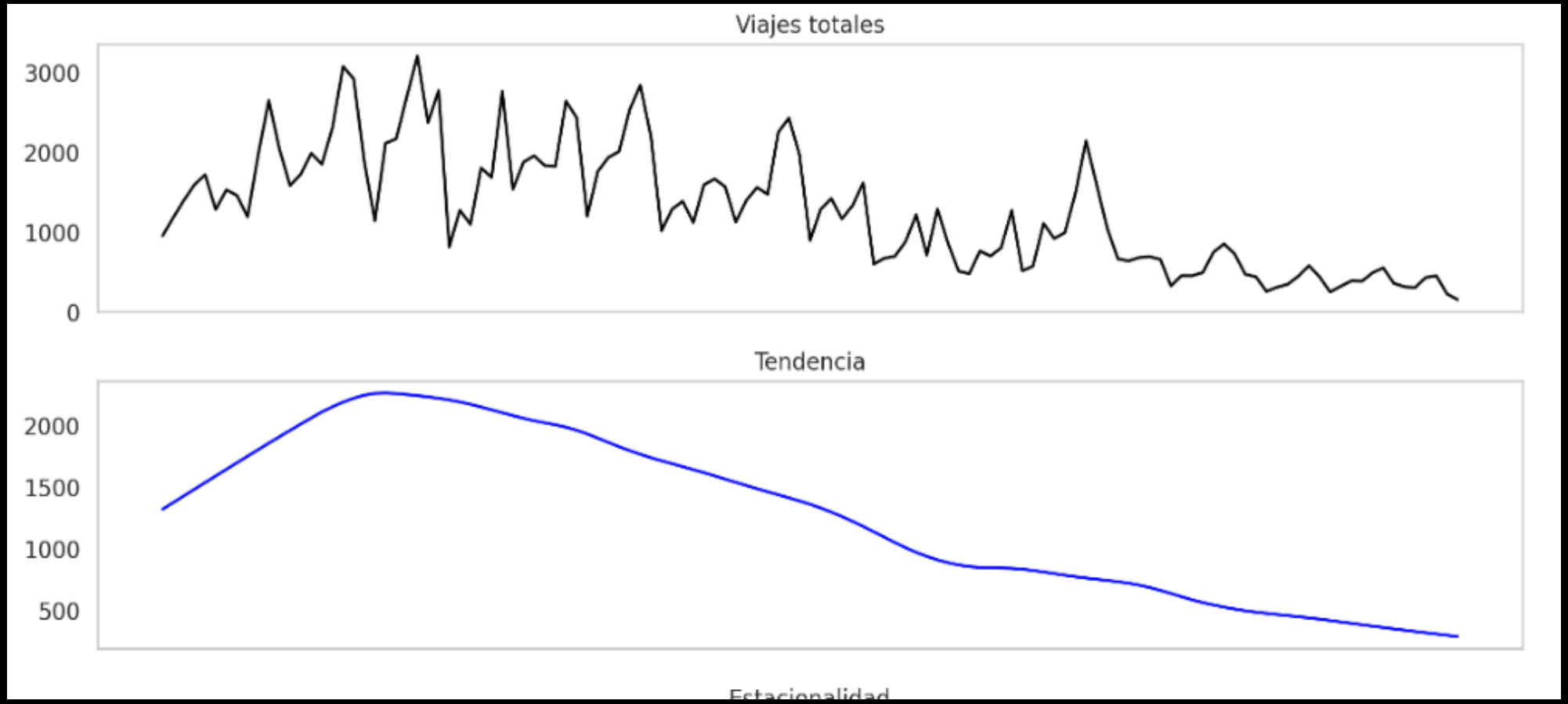
```
# Viajes por tipo de trayecto

df['Tipo de Trayecto'].value_counts()
```

Tipo de Trayecto	count
Medio (1-5 km)	86535
Corto (<1 km)	46235
Largo (>5 km)	24524

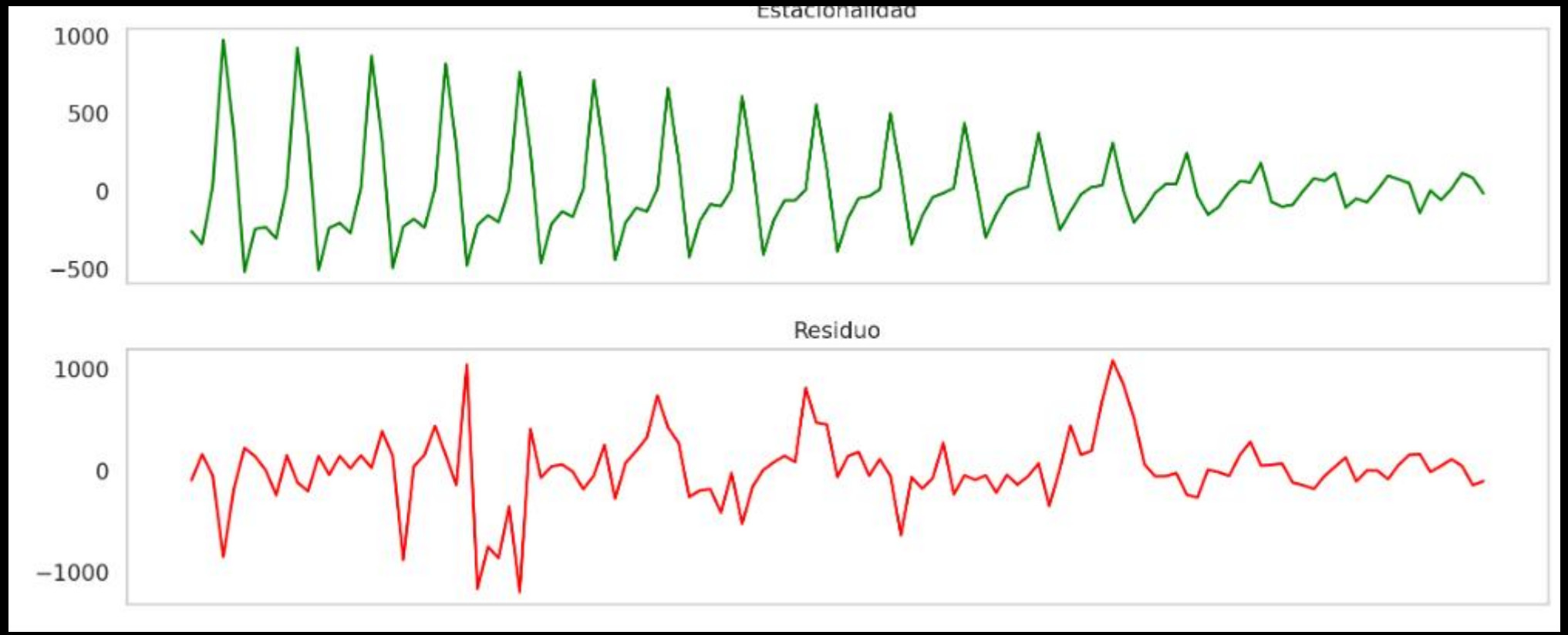
ANÁLISIS STL



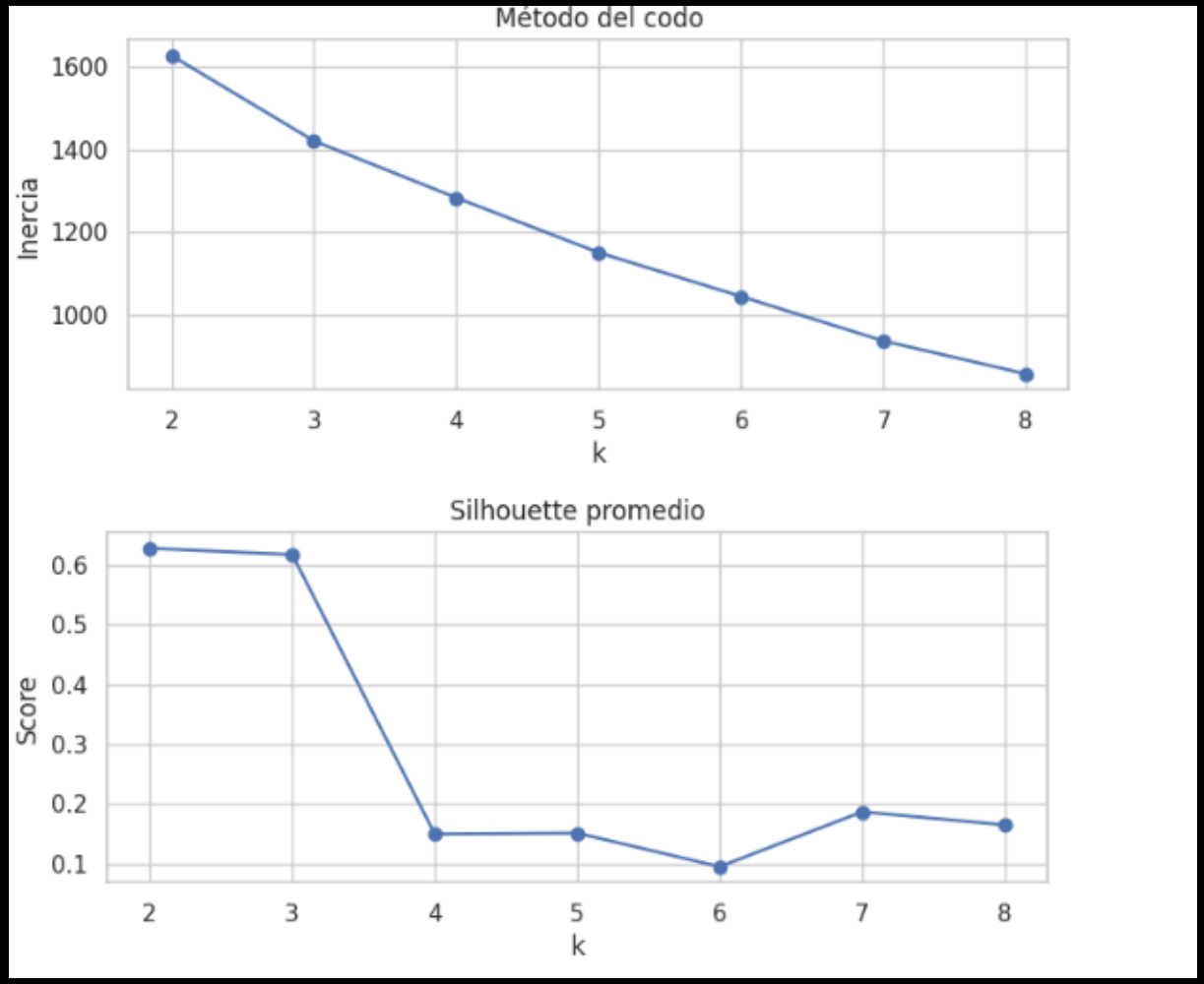
ELEMENTOS DE LA SERIE TEMPORAL



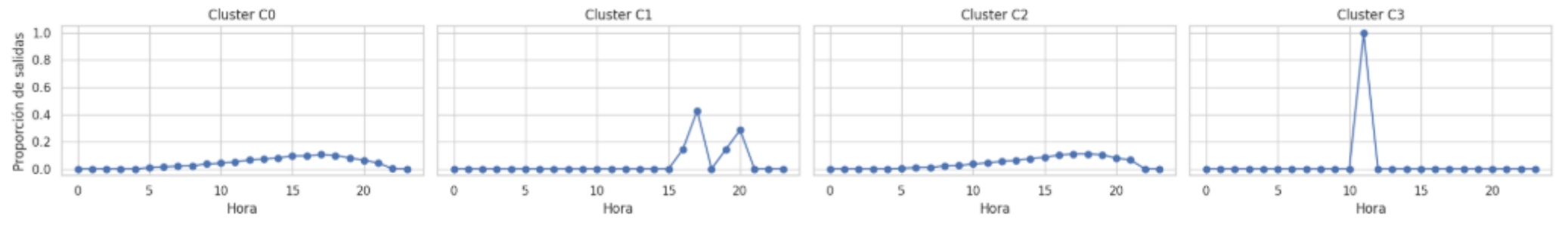
MAS DE LOS ELEMENTOS



ANALISIS DE APRENDIZAJE NO SUPERVIZADO POR K-MEDIA

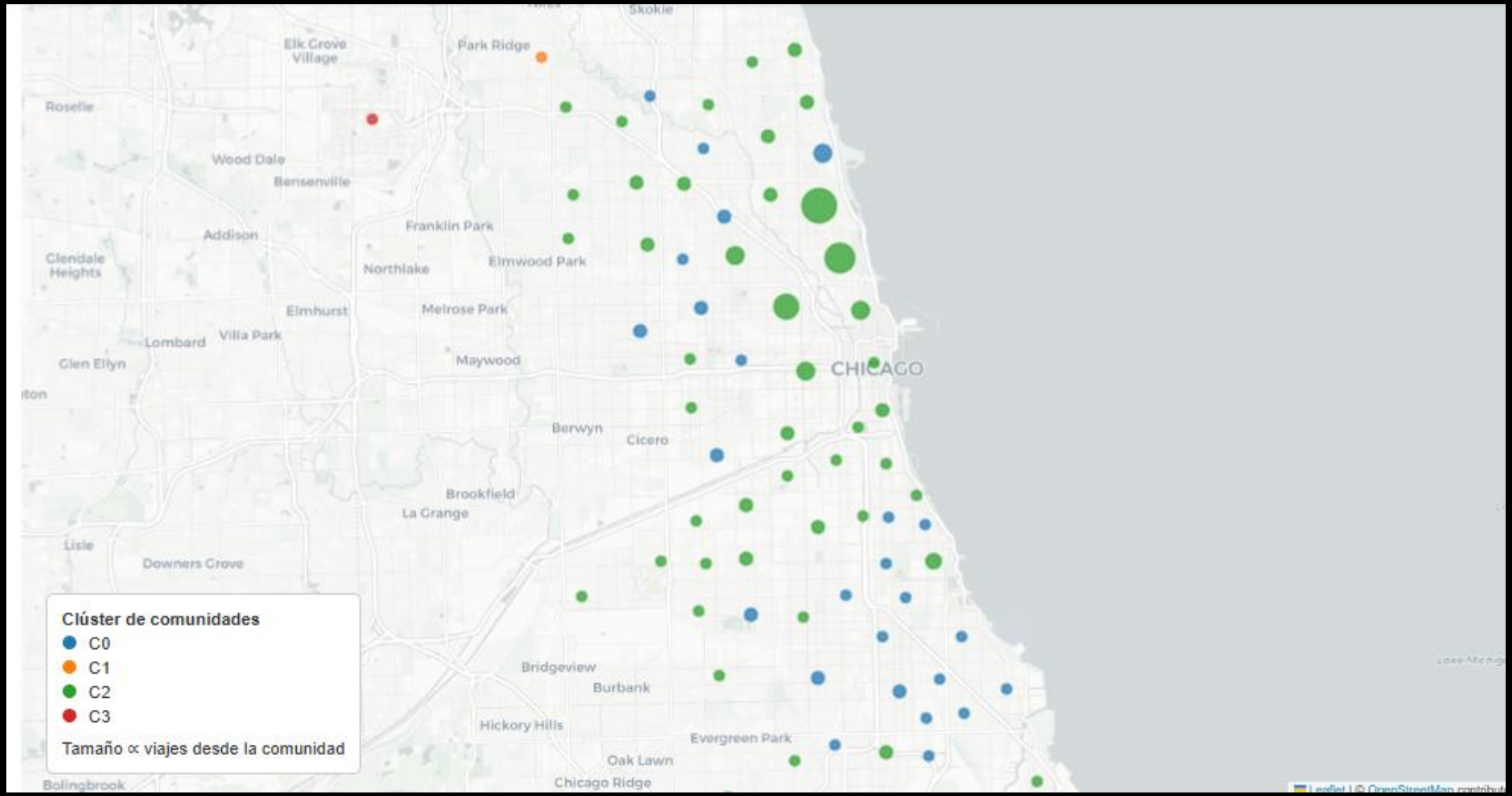


CLUSTERS




```
Cluster 0 (26 comunidades):
ALBANY PARK, AUBURN GRESHAM, AUSTIN, AVALON PARK, AVONDALE, BURNSIDE, CALUMET HEIGHTS, CHATHAM, CHICAGO LAWN, EAST GARFIELD PARK, ENGLEWOOD, FOREST GLEN, GRAND BOULEVA

Cluster 1 (1 comunidades):
EDISON PARK

Cluster 2 (49 comunidades):
ARCHER HEIGHTS, ARMOUR SQUARE, ASHBURN, BELMONT CRAGIN, BEVERLY, BRIDGEPORT, BRIGHTON PARK, CLEARING, DOUGLAS, DUNNING, EAST SIDE, EDGEWATER, FULLER PARK, GAGE PARK, G

Cluster 3 (1 comunidades):
OHARE
```

MAPA CON CLUSTERS



LIMITACIONES

- Datos limitados a Ago-Dic 2020 (post-COVID, posible sesgo estacional).
- Outliers no filtrados explícitamente (e.g., distancias >9km).
- No se incluyen factores externos (clima, tráfico, eventos).
- Asunciones en unidades (metros/segundos) y definiciones (e.g., circular si inicio=fin).
- Modelos supervisados con clases desbalanceadas (e.g., más no-circulares).

ROBUSTEZ

- Cross-validation en Random Forest (k=5 folds, varianza baja en accuracy ~0.79).
- Sensibilidad a hiperparámetros (n_estimators=100-500, max_depth=10-20; óptimo en 100/None).
- Manejo de imbalance (SMOTE para oversampling en clases minoritarias).
- Pruebas en subconjuntos (e.g., por mes; accuracy estable >0.75).

COCLUSIONES EN BASE A CLUSTERS

- C0 — “diurno equilibrado (residencial-laboral mixta)” Actividad gradual desde 7–8 h, pico sostenido entre 12–18 h, y descenso suave hasta 22 h. Lectura: comunidades como Albany Park o Austin con patrones de uso diario equilibrado, combinando traslados laborales y errands diurnos. Acciones: optimizar redistribución durante el pico del mediodía; implementar estaciones de carga en hotspots residenciales para recargas nocturnas; monitorear para evitar congestión en horas punta.
- C1 — “matutino concentrado (commuter)” Subida abrupta 6–9 h con pico pronunciado en 8 h, caída rápida post-10 h y actividad mínima el resto del día. Lectura: zonas como Near North Side o Loop, orientadas a flujos de entrada laboral/estudiantil desde suburbios, con bajo uso recreativo. Acciones: aumentar disponibilidad de scooters pre-alba; rebalanceo inmediato post-pico matutino hacia zonas de fin; integrar con transporte público para hubs de conexión.
- C2 — “vespertino-nocturno (ocio y social)” Baja actividad matutina, ascenso desde 14–16 h, picos múltiples 18–22 h, y tail-off hacia medianoche. Lectura: áreas como Lake View o Lincoln Park, asociadas a salidas de ocio, restaurantes, eventos y retornos residenciales nocturnos. Acciones: reforzar oferta y corrales en tardes; programar cargas en madrugadas; usar geofencing para restringir zonas sensibles por la noche y mejorar seguridad con iluminación/app alerts.
- C3 — “esporádico / bajo volumen (periférico)” Patrones irregulares con picos aislados (e.g., 10 h, 15 h, 20 h) y volúmenes generales bajos; horas muertas frecuentes. Lectura: comunidades periféricas como Englewood o Garfield Ridge con uso ocasional, posiblemente artefacto de datos escasos o eventos específicos; indica subutilización. Acciones: evaluar umbral mínimo de viajes para inclusión en análisis futuros; focalizar campañas de promoción; considerar retiro o reasignación de recursos a clusters más activos si el volumen no justifica mantenimiento.

CONCLUSIONES

- Top zonas: Lake View como hotspot (flujos altos, ~25k inicios/fines).
- • Distribuciones: Distancias medias ~1-2km, duraciones ~10-20min; Lime domina uso.
- • Trayectos circulares: ~29k (20%), más cortos y rápidos; modelos predicen bien (RF accuracy 0.79, AUC 0.84).
- • Patrones temporales: Picos en tardes/fines de semana/agosto-septiembre; tendencia decreciente post-otoño.
- • Concentración: Clusters geográficos (e.g., norte vs sur de Chicago); correlación distancia-duración moderada (0.59).
- • Implicaciones: Informa redistribución de scooters y políticas urbanas.

TRABAJO FUTURO

- Incluir datos multi-año (pre/post-COVID) y variables externas (clima via API).
- Modelos predictivos (e.g., ARIMA para demanda futura).
- Clustering avanzado (DBSCAN para outliers espaciales).
- Integración con optimización (e.g., PuLP para rutas de recarga).
- Usar estos y mas variables para predecir el equilibrio adecuado de Scooters para controlar.

CONTACTO

- Email: jaden59@hotmail.com
- jadennny@gmail.com
- Repositorio: https://github.com/AlbertoHuerta96/Tareas-de-analisis-de-datos

LIBRERIAS

- pandas (preprocesamiento), scikit-learn (RandomForest, KMeans), matplotlib/seaborn (visuales), statsmodels (STL), Folium, Sankey.

GRACIAS POR SU ATENCION