Nombre: Anahí Andrade **Fecha:**13-11-2025

Código: 00323313

Data Mining – Deber 4

1. Evidencia – Ingesta de años 2015–2025 (Yellow/Green):

Imprimir conteos por año/mes, duración:

```
================================================================================
INICIANDO INGESTA MASIVA NYC TLC -> POSTGRES
================================================================================
Inicio: 2025-11-11 13:19:00.976454
Servicios: ['yellow', 'green']
Rango: 2015-2025
Run ID: run_001
================================================================================

Checkpoint cargado: 258 archivos registrados.


================================================================================
Servicio: YELLOW
================================================================================

  yellow 2015-01: Ya procesado (SKIP) [12741035 filas, 494.63848s]
  yellow 2015-02: Ya procesado (SKIP) [12442394 filas, 339.948858s]
  yellow 2015-03: Ya procesado (SKIP) [13342951 filas, 364.81764s]
  yellow 2015-04: Ya procesado (SKIP) [13063758 filas, 350.321475s]
  yellow 2015-05: Ya procesado (SKIP) [13157677 filas, 359.700977s]
  yellow 2015-06: Ya procesado (SKIP) [12324936 filas, 334.870962s]
  yellow 2015-07: Ya procesado (SKIP) [11559666 filas, 312.419438s]
  yellow 2015-08: Ya procesado (SKIP) [11123123 filas, 308.906263s]
  yellow 2015-09: Ya procesado (SKIP) [11218122 filas, 303.17454s]
  yellow 2015-10: Ya procesado (SKIP) [12307333 filas, 336.394519s]
  yellow 2015-11: Ya procesado (SKIP) [11305240 filas, 307.264311s]
```

```
================================================================================
Servicio: GREEN
================================================================================

  green 2015-01: Ya procesado (SKIP) [1508493 filas, 45.402486s]
  green 2015-02: Ya procesado (SKIP) [1574830 filas, 47.060252s]
  green 2015-03: Ya procesado (SKIP) [1722574 filas, 51.600149s]
  green 2015-04: Ya procesado (SKIP) [1664394 filas, 50.882582s]
  green 2015-05: Ya procesado (SKIP) [1786848 filas, 53.026007s]
  green 2015-06: Ya procesado (SKIP) [1638868 filas, 54.647252s]
  green 2015-07: Ya procesado (SKIP) [1541671 filas, 48.479503s]
  green 2015-08: Ya procesado (SKIP) [1532343 filas, 49.18342s]
  green 2015-09: Ya procesado (SKIP) [1494927 filas, 46.274572s]
  green 2015-10: Ya procesado (SKIP) [1630536 filas, 52.876946s]
  green 2015-11: Ya procesado (SKIP) [1529984 filas, 48.926557s]
  green 2015-12: Ya procesado (SKIP) [1608297 filas, 54.280331s]
  green 2016-01: Ya procesado (SKIP) [1445292 filas, 48.692978s]
```

Summary final:

```
================================================================================
RESUMEN FINAL DE INGESTA
================================================================================
Exitosos:       0
Omitidos:       258
No encontrados: 0
Fallidos:       0
Rango:          2015-2025
Servicios:      yellow, green
Run ID:         run_001
Checkpoint:     /home/jovyan/work/checkpoint_ingesta.json
Duracion total: 0.00s (0.00 min)
Fin: 2025-11-11 13:19:00.984427
================================================================================
```

2. Evidencia - Imprimir conteos por año/mes, duración y summary final (OBT):

Imprimir conteos por año/mes, duración:

```
[2025-11-11 19:14:38] Iniciando script build_obt.py (OPTIMIZADO CON COPY)
[2025-11-11 19:14:38] Argumentos: Namespace(mode='full', year_start=2020, year_end=2022, overwrite=True)
[2025-11-11 19:14:38] Intentando conectar a PostgreSQL...
[2025-11-11 19:14:38] Conexion exitosa a PostgreSQL con optimizaciones
[2025-11-11 19:14:38] ============================================================
[2025-11-11 19:14:38] INICIANDO CONSTRUCCION OBT (MODO OPTIMIZADO)
[2025-11-11 19:14:38] ============================================================
[2025-11-11 19:14:38] Rango: 2020-2022
[2025-11-11 19:14:38] Overwrite: True
[2025-11-11 19:14:38] RUN_ID: run_001
[2025-11-11 19:14:38] ============================================================

[2025-11-11 19:14:38] Creando tabla analytics.obt_trips...
[2025-11-11 19:14:38] Tabla analytics.obt_trips creada/verificada

[2025-11-11 19:14:38] ============================================================
[2025-11-11 19:14:38] Servicio: YELLOW
[2025-11-11 19:14:38] ============================================================

[2025-11-11 19:14:38] --- Anio 2020 ---
[2025-11-11 19:14:38] Procesando: yellow 2020-01
[2025-11-11 19:14:38]   -  Extrayendo datos con COPY...
[2025-11-11 19:14:39]   - Insertando datos con COPY...
[2025-11-11 19:14:43]   - COMPLETADO: 6,405,008 filas en 92.6s (1525240 filas/seg)
[2025-11-11 19:14:43] Procesando: yellow 2020-02
[2025-11-11 19:14:43]   -  Extrayendo datos con COPY...
[2025-11-11 19:14:44]   - Insertando datos con COPY...
[2025-11-11 19:14:48]   - COMPLETADO: 6,299,367 filas en 88.4s (1536431 filas/seg)
[2025-11-11 19:14:48] Procesando: yellow 2020-03
[2025-11-11 19:14:48]   -  Extrayendo datos con COPY...
[2025-11-11 19:14:49]   - Insertando datos con COPY...
[2025-11-11 19:14:51]   - COMPLETADO: 3,007,687 filas en 83.1s (1253203 filas/seg)
```

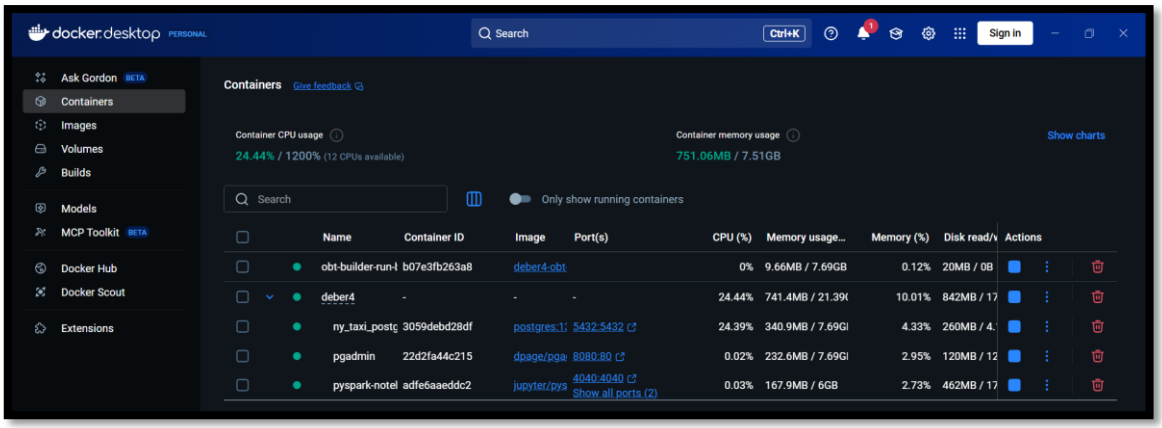
```
[2025-11-11 19:18:30] ============================================================
[2025-11-11 19:18:30] Servicio: GREEN
[2025-11-11 19:18:30] ============================================================

[2025-11-11 19:18:30] --- Anio 2020 ---
[2025-11-11 19:18:30] Procesando: green 2020-01
[2025-11-11 19:18:30]  -  Extrayendo datos con COPY...
[2025-11-11 19:18:30]  - Insertando datos con COPY...
[2025-11-11 19:18:31]  - COMPLETADO: 447,770 filas en 95.9s (497522 filas/seg)
[2025-11-11 19:18:31] Procesando: green 2020-02
[2025-11-11 19:18:31]  -  Extrayendo datos con COPY...
[2025-11-11 19:18:31]  - Insertando datos con COPY...
[2025-11-11 19:18:32]  - COMPLETADO: 398,632 filas en 98.4s (498290 filas/seg)
[2025-11-11 19:18:32] Procesando: green 2020-03
[2025-11-11 19:18:32]  -  Extrayendo datos con COPY...
[2025-11-11 19:18:32]  - Insertando datos con COPY...
[2025-11-11 19:18:33]  - COMPLETADO: 223,496 filas en 80.7s (446992 filas/seg)
```

Summary final:

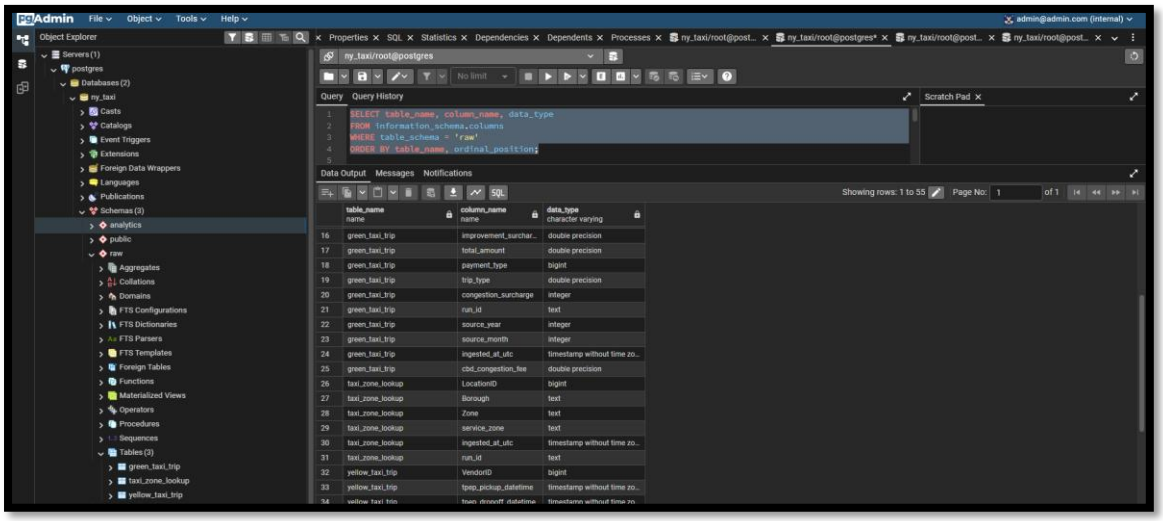
```
[2025-11-11 19:22:47] ============================================================
[2025-11-11 19:22:47] RESUMEN FINAL
[2025-11-11 19:22:47] ============================================================
[2025-11-11 19:22:47] Total filas insertadas: 96,478,663
[2025-11-11 19:22:47] Finalizado: 2025-11-11 19:22:47
[2025-11-11 19:22:47] ============================================================
```

3. Evidencia – Docker Compose ejecutándose:

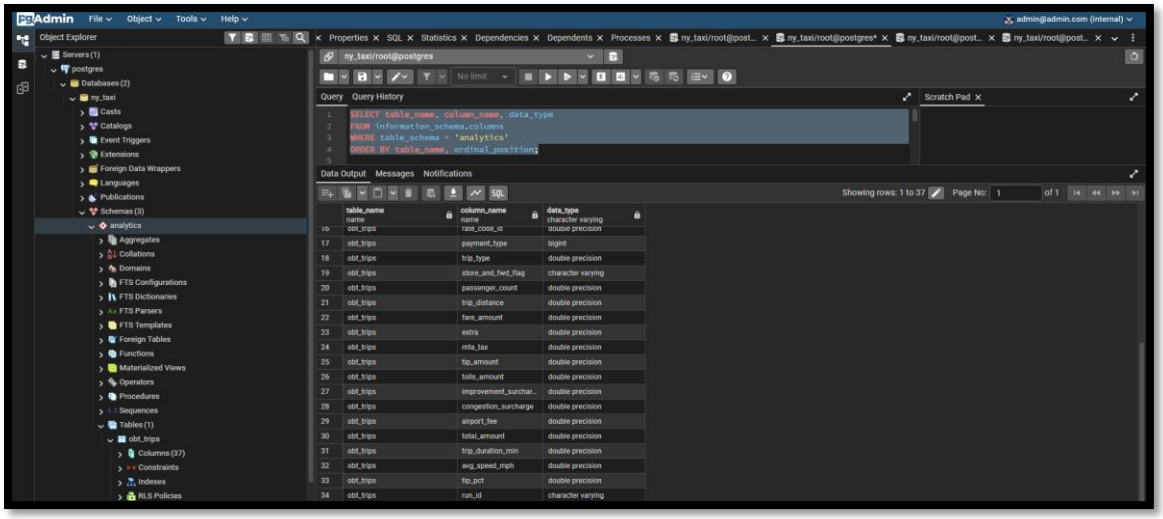


4. Evidencia – Postgress:

Esquema RAW:



Esquema ANALYTICS:



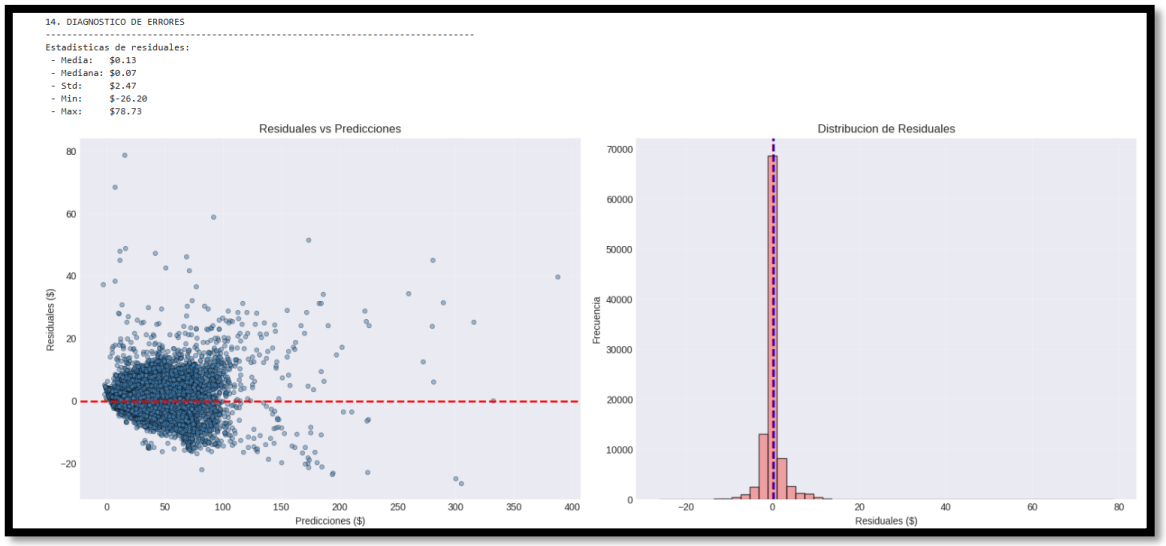
5. **Evidencia – Tabla comparativa de todos los modelos (propios y sklearn) con métricas:**

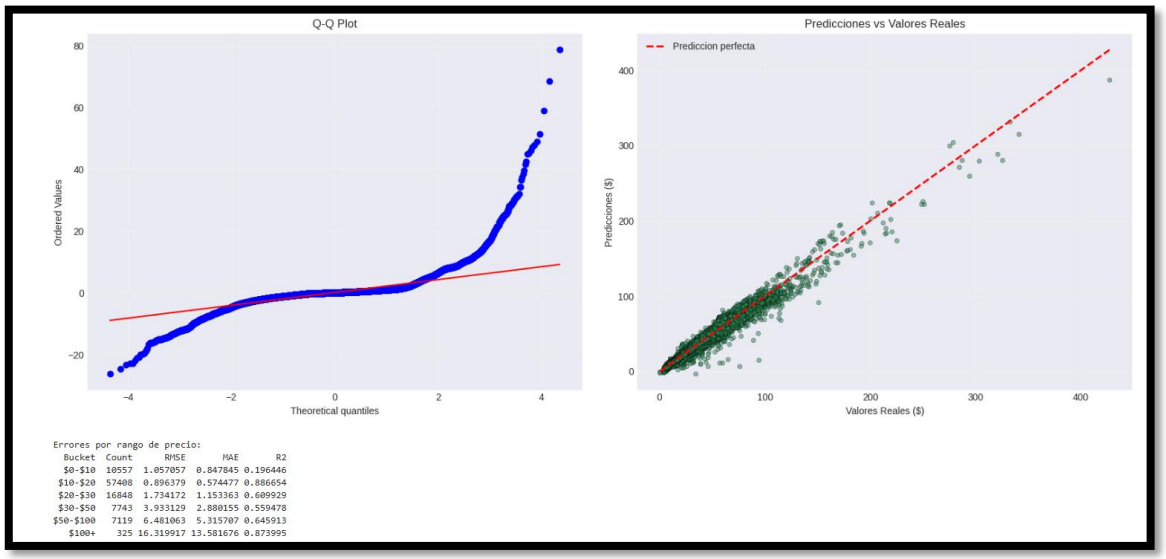
Tabla comparativa completa:

Modelo	Tipo	RMSE_Val	MAE_Val	R2_Val	Tiempo (s)
Ridge_Scratch	From-Scratch	2.203395	1.121511	0.975824	0.198644
Ridge_Sklearn	Scikit-Learn	2.203395	1.121511	0.975824	0.230908
Lasso_Sklearn	Scikit-Learn	2.209805	1.119803	0.975684	2.135232
Lasso_Scratch	From-Scratch	2.209948	1.119719	0.975680	53.260162
SGD_Scratch	From-Scratch	2.230374	1.133703	0.975229	499.708629
ElasticNet_Scratch	From-Scratch	2.281987	1.149258	0.974069	35.541106
ElasticNet_Sklearn	Scikit-Learn	2.282110	1.149357	0.974066	2.188510
SGD_Sklearn	Scikit-Learn	8.621060	1.776164	0.629904	78.973437

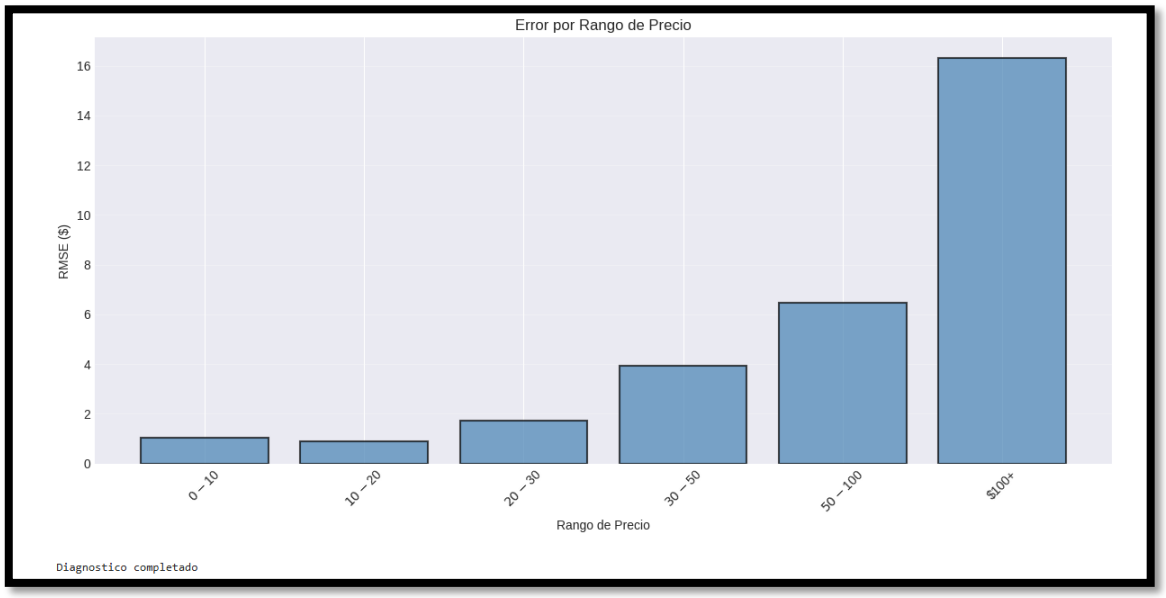
6. **Evidencia – Gráficos de residuales/errores por bucket:**

```
14. DIAGNOSTICO DE ERRORES
--------------------------------------------------------------------------------
Estadisticas de residuales:
 - Media:   $0.13
 - Mediana: $0.07
 - Std:     $2.47
 - Min:     $-26.20
 - Max:     $78.73
```






```
Errores por rango de precio:
   Bucket  Count      RMSE       MAE       R2
    $0-$10  10557  1.057057  0.847845 0.196446
   $10-$20  57408  0.896379  0.574477 0.886654
   $20-$30  16848  1.734172  1.153363 0.609929
   $30-$50   7743  3.933129  2.880155 0.559478
  $50-$100   7119  6.481063  5.315707 0.645913
    $100+    325 16.319917 13.581676 0.873995
```




```
Diagnostico completado
```