Data Mining – Deber 5

Integrantes:

- Anahí Andrade (00323313)
- Erick Suárez (00325769)
- Jesús Alarcón (00324826)
- Giselle Cevallos (00325549)

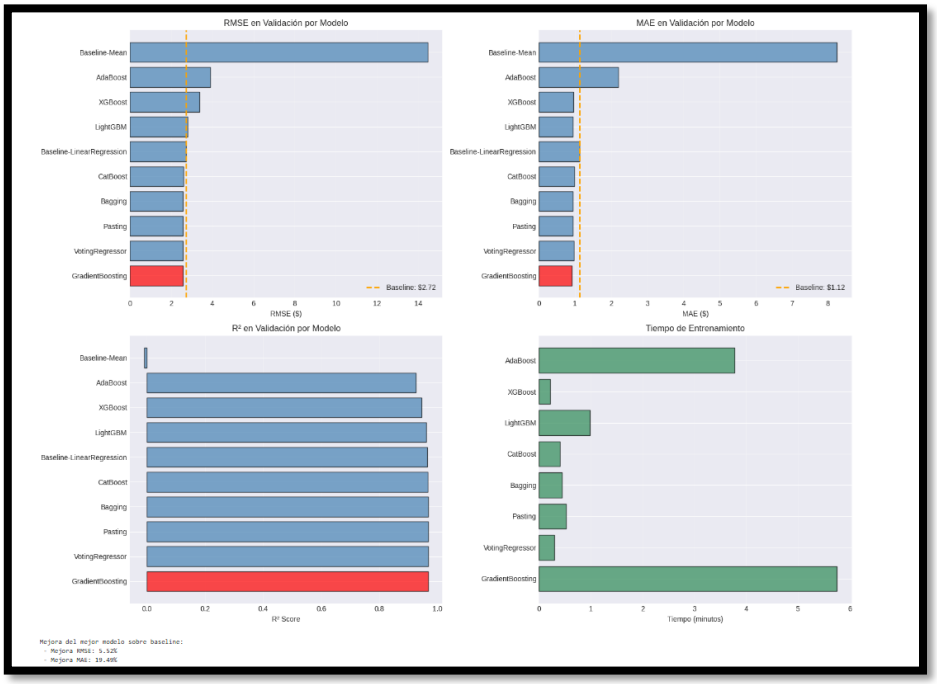
1. Evidencia – Tablas comparativas (RMSE/MAE/R²):

pset5_ensemble_regression.ipynb:

```
16. COMPARATIVA COMPLETA
--------------------------------------------------------------------------------

Tabla comparativa (ordenada por RMSE en Validación):
                      Modelo             Tipo  RMSE_Val  MAE_Val    R2_Val  Tiempo_Train
          GradientBoosting         Boosting  2.568338 0.904003  0.968048    345.067705
           VotingRegressor  Ensemble-Voting  2.581056 0.964625  0.967731     17.636931
                   Pasting Ensemble-Pasting  2.585935 0.944586  0.967609     31.681893
                   Bagging Ensemble-Bagging  2.587734 0.943347  0.967564     26.717474
                  CatBoost         Boosting  2.598098 0.980608  0.967303     24.099514
Baseline-LinearRegression         Baseline  2.718496 1.122803  0.964203      0.000000
                  LightGBM         Boosting  2.808612 0.931513  0.961790     59.046815
                   XGBoost         Boosting  3.389028 0.951536  0.944366     12.764911
                  AdaBoost         Boosting  3.913961 2.191863  0.925796    226.354443
             Baseline-Mean         Baseline 14.440150 8.248850 -0.010034      0.000000

================================================================================
 MEJOR MODELO: GradientBoosting
   RMSE Validación: $2.57
   MAE Validación: $0.90
   R² Validación: 0.9680
================================================================================
```

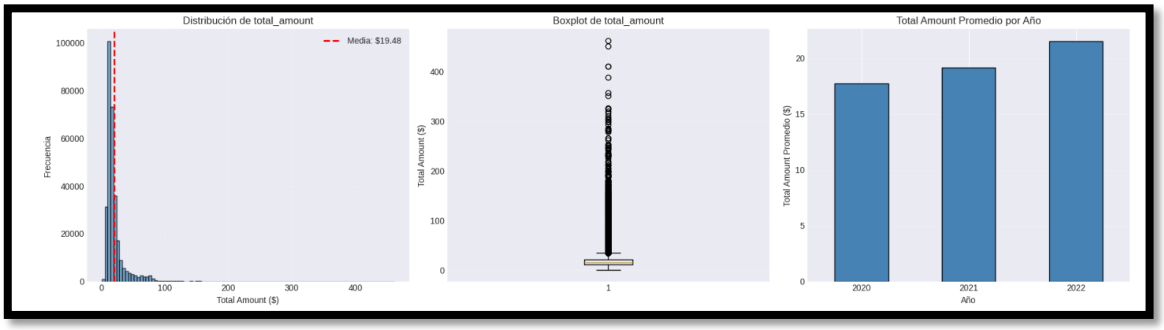


evidencias/results_comparison_pset5.csv:

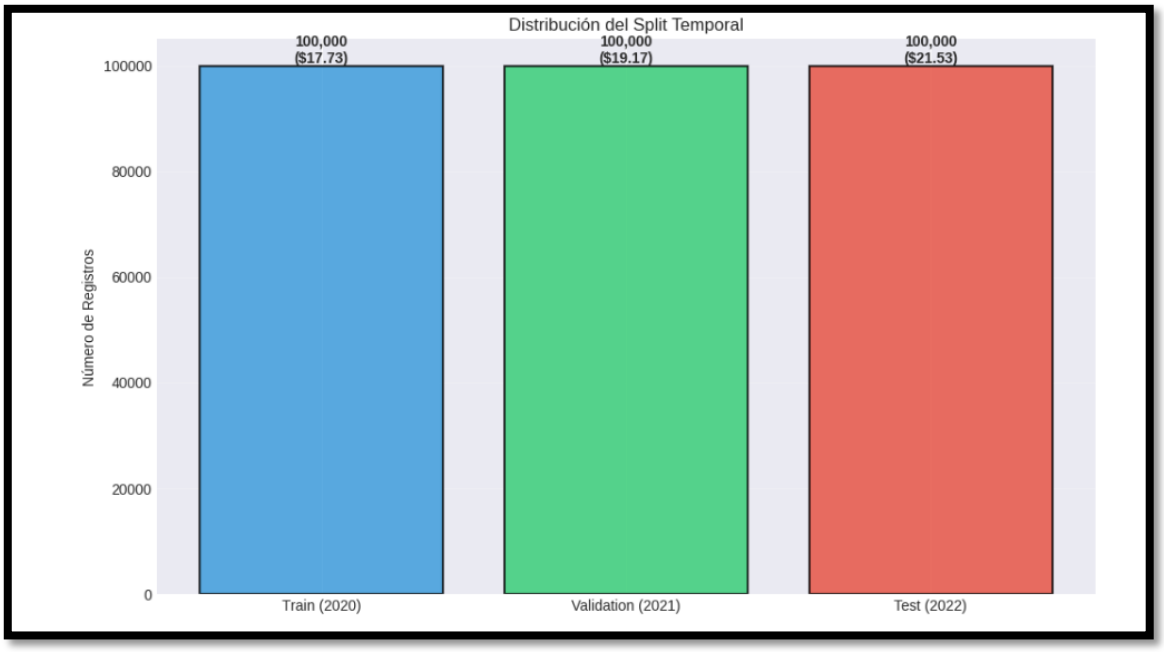
	Modelo	Tipo	RMSE_Val	MAE_Val	R2_Val	Tiempo_Train	Best_Params
1	GradientBoosting	Boosting	2.5683378649346205	0.9040031920109356	0.9680480868401601	345.067705	rs': 200, 'subsample': 0.8}
2	VotingRegressor	Ensemble-Voting	2.581056117577182	0.9646253851819243	0.9677308554700327	17.636931	
3	Pasting	Ensemble-Pasting	2.5859346074746346	0.9445858830124689	0.9676087554739776	31.681893	
4	Bagging	Ensemble-Bagging	2.5877342623408066	0.9433468790930065	0.9675636550720609	26.717474	
5	CatBoost	Boosting	2.5980984766317996	0.9806077963165462	0.9673033111476893	24.099514	agging_temperature': 0.5}
6	aseline-LinearRegression	Baseline	2.718496133285426	1.1228025379394533	0.9642027225987193	0.0	
7	LightGBM	Boosting	2.8086119495339212	0.9315126338744573	0.9617900880861926	59.046815	5, 'bagging_fraction': 0.8}
8	XGBoost	Boosting	3.3890284747004054	0.9515357934947102	0.9443656531751787	12.764911	5, 'colsample_bytree': 1.0}
9	AdaBoost	Boosting	3.9139613288465758	2.1918629664282627	0.9257962901255783	226.354443	': 0.05, 'n_estimators': 50}
10	Baseline-Mean	Baseline	14.440149507384461	8.248849732640002	-0.01003387009735212 7	0.0	

2. Evidencia – Gráficos:

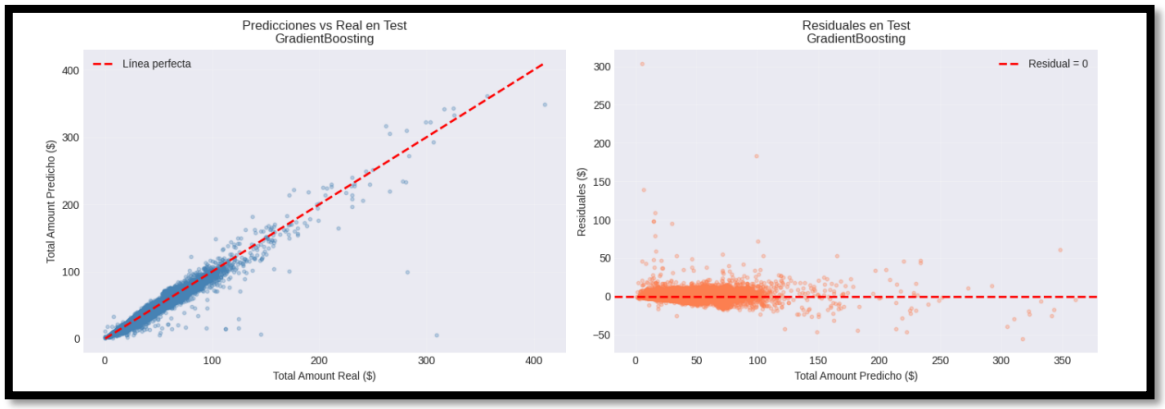
Análisis Exploratorio:



Split Temporal:



Evaluación en Test:



3. Evidencia – Logs de búsqueda:

A lo largo del notebook se detalla el registro completo de esta actividad. Por ejemplo:

```
11. ADABOOST (BOOSTING PRINCIPAL)
================================================================================
Configurando Grid Search para AdaBoost...
 - Parámetros a buscar: {'n_estimators': [50, 100, 200], 'learning_rate': [0.05, 0.1, 0.5]}
 - CV: TimeSeriesSplit
 - Base learner: DecisionTree con max_depth=3
 - Combinaciones totales: 18 fits

Entrenando AdaBoost con Grid Search...
(Estimado: 5-10 minutos...)
Fitting 2 folds for each of 9 candidates, totalling 18 fits

================================================================================
✓ Mejores parámetros AdaBoost: {'learning_rate': 0.05, 'n_estimators': 50}
✓ Mejor score CV: $3.36
================================================================================

Resultados AdaBoost en Validación:
 - RMSE: $3.91
 - MAE: $2.19
 - R²: 0.9258
 - Tiempo: 226.35s (3.8 min)

Análisis de sensibilidad (AdaBoost):

Todas las combinaciones probadas:
param_n_estimators param_learning_rate  mean_rmse
                50                0.05   3.360250
                50                 0.1   3.567953
               100                 0.1   3.585563
               200                0.05   3.592858
               100                0.05   3.601231
               200                 0.1   4.388471
                50                 0.5   4.653048
               100                 0.5   5.862130
               200                 0.5   7.907964
```

4. Evidencia – Tiempos:

A lo largo del notebook se detalla el registro completo de esta actividad. Algunos ejemplos:

```
9. VOTING REGRESSOR
--------------------------------------------------------------------------------

✓ Modelos base seleccionados (3):
  1. Ridge Regression (lineal, regularizado)
  2. DecisionTree (no lineal, flexible)
  3. GradientBoosting (ensemble, robusto)

✓ Justificación:
  - Diversidad: Combina modelos lineales, árboles y boosting
  - Complementariedad: Ridge captura tendencias lineales,
    Tree captura no linealidades, GB reduce bias

✓ Strategy: Promedio simple (voting='uniform')
✓ Pesos: [1, 1, 1] (iguales para todos los modelos)
--------------------------------------------------------------------------------
Entrenando Voting Regressor...
 - Modelos base: Ridge, DecisionTree, GradientBoosting

Resultados Voting Regressor:
 - RMSE: $2.58
 - MAE: $0.96
 - R²: 0.9677
 - Tiempo entrenamiento: 17.64s
```

```
12. GRADIENT BOOSTING
--------------------------------------------------------------------------------
Configurando Grid Search para Gradient Boosting...
 - Parámetros a buscar: {'n_estimators': [100, 200], 'learning_rate': [0.05, 0.1], 'max_depth': [3, 5], 'subsample': [0.8]}
 - Combinaciones totales: 16 fits

Entrenando Gradient Boosting con Grid Search...
(Estimado: 4-8 minutos...)
Fitting 2 folds for each of 8 candidates, totalling 16 fits

✓ Mejores parámetros GBDT: {'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 200, 'subsample': 0.8}
✓ Mejor score CV: $1.99

Resultados Gradient Boosting en Validación:
 - RMSE: $2.57
 - MAE: $0.90
 - R²: 0.9680
 - Tiempo: 345.07s (5.8 min)
```

5. Evidencia – Gráficos de residuales/errores por bucket:

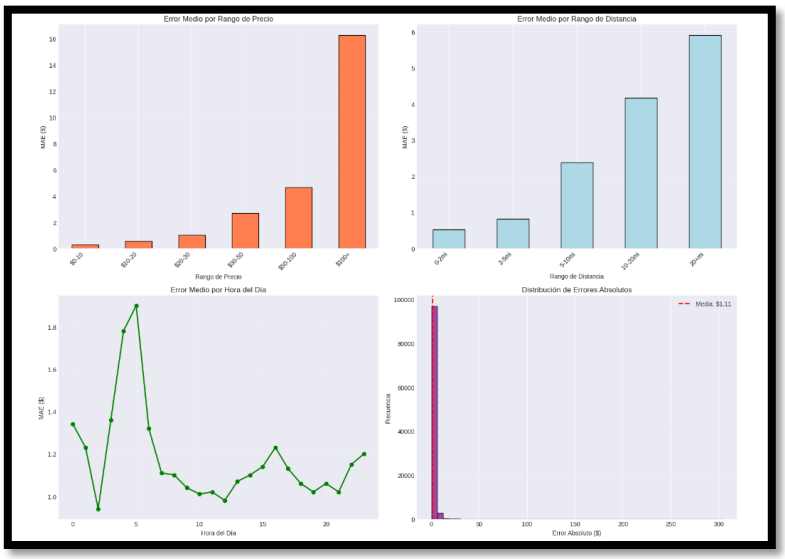
```
18. ANÁLISIS DE ERROR POR BUCKETS
--------------------------------------------------------------------------------

1. Error por rango de precio:
                MAE  Median_Error  Std_Error  Count
price_bucket
$0-10          0.28          0.09       0.53  10677
$10-20         0.53          0.29       0.66  57305
$20-30         1.02          0.53       1.25  16486
$30-50         2.68          1.83       2.59   7991
$50-100        4.66          4.11       3.66   7222
$100+         16.24         11.19      25.01    319

2. Error por rango de distancia:
                  MAE  Median_Error  Count
distance_bucket
0-2mi            0.52          0.24  52903
2-5mi            0.82          0.41  29487
5-10mi           2.38          1.38   9081
10-20mi          4.17          3.54   7397
20+mi            5.90          4.11   1132

3. Error por hora del día:
                MAE  Median_Error  Count
pickup_hour
0              1.34          0.53   2727
1              1.23          0.50   1794
2              0.94          0.49   1178
3              1.36          0.46    735
4              1.78          0.75    503
5              1.90          0.73    549
6              1.32          0.41   1474
7              1.11          0.36   2803
8              1.10          0.37   3845
9              1.04          0.35   4300
10             1.01          0.34   4769
11             1.02          0.33   5262

4. Error por borough:
                   MAE  Median_Error  Count
pu_borough
Bronx             2.74          1.72    125
Brooklyn          1.67          0.58    711
EWR               8.70          8.41      4
Manhattan         0.81          0.33  88896
Queens            3.65          2.98   9080
Staten Island    13.80         14.23      3
Unknown           2.58          0.89   1154
```



Evidencias extras

6. 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
================================================================================
```

7. 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)
```

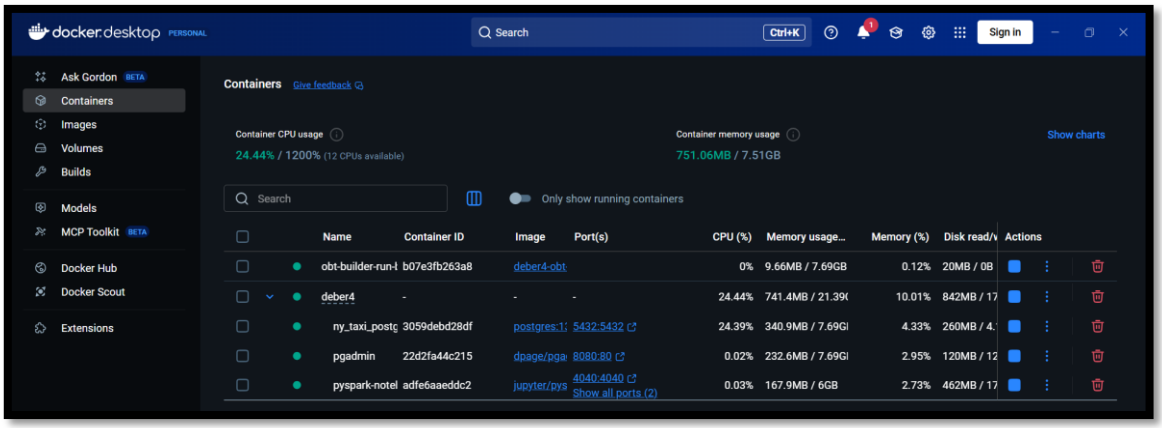
```
[2025-11-11 19:18:30] ============================================================
[2025-11-11 19:18:30] Servicio: GREEN
[2025-11-11 19:18:30] ============================================================

[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:

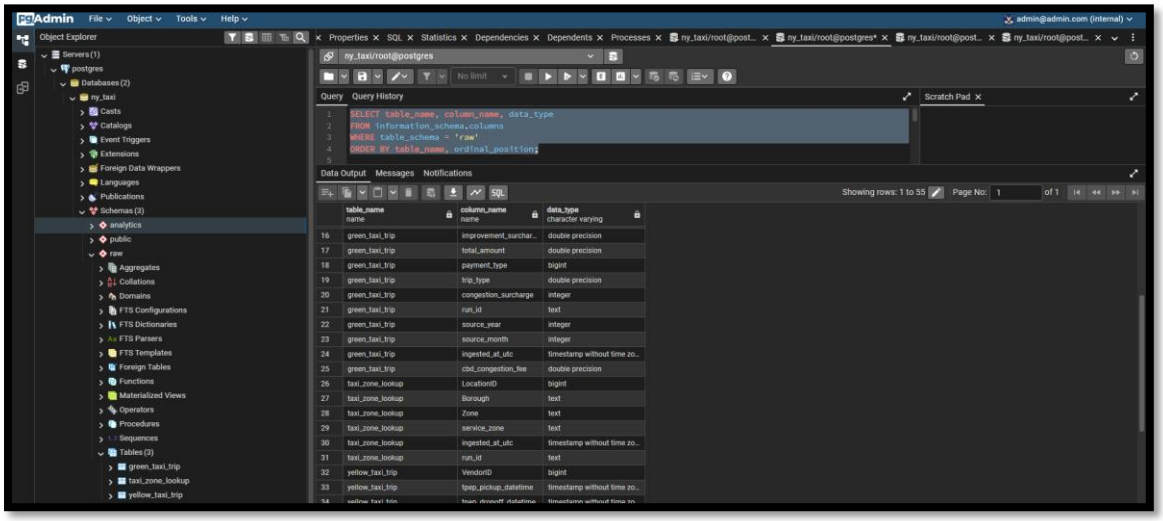
```
[2025-11-11 19:22:47] ============================================================
[2025-11-11 19:22:47] RESUMEN FINAL
[2025-11-11 19:22:47] ============================================================
[2025-11-11 19:22:47] Total filas insertadas: 96,478,663
[2025-11-11 19:22:47] Finalizado: 2025-11-11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

8. Evidencia – Docker Compose ejecutándose:



9. Evidencia – Postgress:

Esquema RAW:



Esquema ANALYTICS: