

Daniel Antares

Data Analyst · Data Scientist

Semarang, Indonesia (UTC+7) | daniel.antares@yahoo.com | [GitHub](#) | [LinkedIn](#)

Tiga proyek end-to-end, semuanya sudah di-deploy dan dapat diproduksi ulang. Setiap grafik dihasilkan langsung oleh kode di repository, dan tiap proyek menyertakan tautan ke dashboard atau dokumentasi yang bisa dibuka tanpa instalasi. Semua repository bersifat publik, lengkap dengan README berisi metodologi, sumber data, dan batasan analisis.

1 Data Job Market Intelligence

Skill data apa yang benar-benar dicari, berapa bayarannya, dan seberapa jauh perbedaannya antara pasar Indonesia dan pasar remote internasional?

▶ [Baca ulasan lengkap](#)

[Dashboard interaktif \(bangun ~30 detik\)](#)

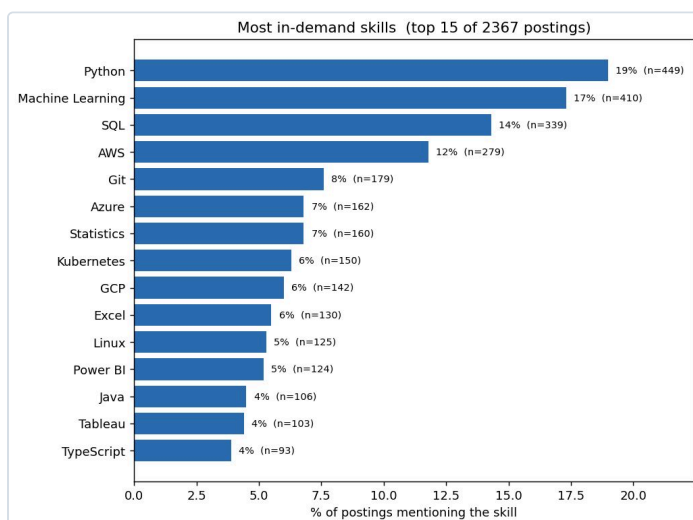
[GitHub](#)

YANG SAYA BANGUN

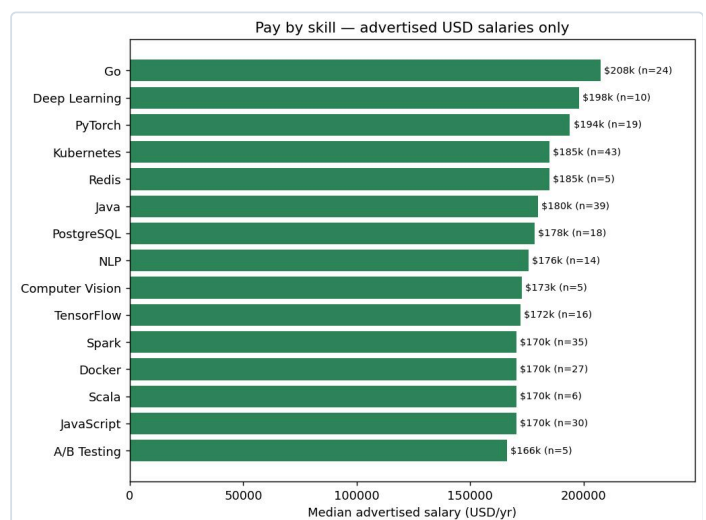
- Pipeline otomatis yang mengumpulkan lowongan data dari **6 API publik** (Remotive, Jobicy, RemoteOK, Adzuna, Jooble, JSearch) di **5 pasar**, dinormalisasi ke satu skema tunggal dan diupdate secara mingguan.
- **Parser gaji** yang mengubah teks bebas ("80k - 100k", "18/hr") menjadi nilai terstruktur dalam USD/tahun memakai tabel kurs bertanggal, serta **ekstraktor skill** dengan taksonomi ~70 skill yang mampu membedakan nama ambigu (R vs "R&D", Java vs JavaScript).
- Warehouse Parquet terdeduplikasi, dashboard Streamlit interaktif dengan estimator gaji, serta **pembaruan data terjadwal mingguan** dan pengujian CI melalui GitHub Actions.

TEMUAN UTAMA (2.367 lowongan)

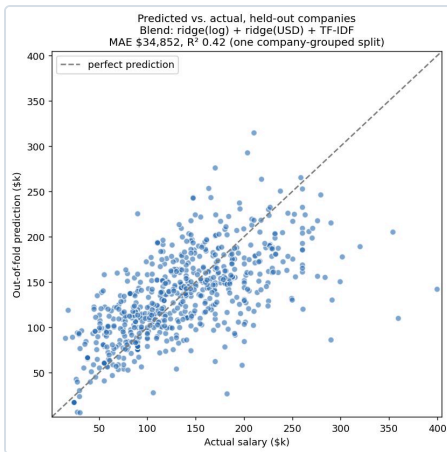
- **Pasar lokal dan remote menuntut hal yang berbeda.** Lowongan di Indonesia cenderung analyst/BI: SQL **61%**, Python **58%**, Tableau **30%**. Malaysia lebih ML-first (**43%**) dengan Power BI. Lowongan remote internasional didominasi Python dan AWS, dengan porsi cloud serta MLOps yang jauh lebih besar.
- **Skill spesialis dan infrastruktur memberi premi yang jelas:** Deep Learning **\$198k**, PyTorch **\$194k**, Kubernetes **\$185k**, dibanding **\$120k hingga \$135k** untuk tooling analitik umum. Per peran, ML Engineer **\$162k** memimpin atas Data Analyst **\$112k**.
- **Model gaji naik dari R^2 0,12 menjadi 0,46 tanpa mengganti algoritmanya.** Seluruh peningkatan berasal dari pengerjaan data: normalisasi kurs memulihkan 143 baris yang sebelumnya terbuang, kolom yang sudah ada di warehouse (location, source, market) memberi lompatan terbesar, dan penggalan teks deskripsi menambah sisanya. Gradient boosting justru kalah dari model linear yang diregularisasi.



Skill yang paling banyak diminta di 2.367 lowongan.



Median gaji yang diiklankan per skill, dinormalisasi ke USD/tahun.



Prediksi vs aktual pada data perusahaan di luar training. Setiap perusahaan hanya masuk train atau test.

Teknologi: Python · pandas · PyArrow/Parquet · scikit-learn (ridge, TF-IDF, GroupKFold) · Streamlit · matplotlib · pytest · GitHub Actions (ETL terjadwal + CI)

2 Brazilian E-Commerce Analytics: SQL ke Power BI

Apa yang mendorong pendapatan, siapa pelanggan terbaik, dan di mana operasional bermasalah pada marketplace dengan 100 ribu pesanan?

[▶ Lihat dashboard Power BI](#) [GitHub](#)

YANG SAYA BANGUN

- Sembilan tabel operasional mentah dimuat ke **PostgreSQL** dan dimodelkan menjadi **star schema** terdokumentasi: dua tabel fakta yang memakai dimensi yang sama untuk customer, produk, seller, dan tanggal, sehingga angka dari keduanya bisa dibandingkan secara konsisten.
- Pustaka **SQL analitik**: cohort retention, segmentasi RFM, Pareto kategori, dan tren berbasis window function. Satu pertanyaan bisnis per file, disajikan ke BI (Business Intelligence) melalui tujuh reporting view.
- Dashboard **Power BI** empat halaman (Executive, Customers, Operations, Products & Payments) dengan ~15 DAX measure di atas star schema.

TEMUAN UTAMA (~97.9k pesanan, R\$15.7M, 2017-01 s.d. 2018-08)

- Bisnis ini bergantung pada pembeli satu kali transaksi. 97% pelanggan tidak pernah kembali**, menghasilkan 94% pendapatan, dengan tingkat pembelian ulang di bawah 1% setelah bulan pertama. Padahal 3% yang kembali bernilai sekitar **2× lebih besar** (R\$308 vs R\$161), dan di situlah peluang retensinya.
- Pendapatan sangat terkonsentrasi**. Analisis RFM menunjukkan **16% pelanggan menguasai 31% pendapatan**; São Paulo sendiri 37% dari pendapatan, dan SP+RJ+MG jika digabung mencapai 62%.
- Keterlambatan pengiriman paling merusak kepuasan pelanggan**. Rata-rata rating pesanan yang terlambat **2,57★**, **4,29★** untuk pesanan tepat waktu. Secara keseluruhan 8% pesanan terlambat, naik menjadi **21%** di bulan promo.



Halaman Executive overview.



Segmentasi pelanggan dan cohort retention.

Teknologi: PostgreSQL · SQL tingkat lanjut (window function, CTE, NTILE) · dimensional modelling · Power BI + DAX · Docker · DuckDB

3 NYC Mobility: Analytics Engineering dengan dbt

Tiga sumber data ride-hailing dengan format berbeda, disatukan menjadi satu tabel yang bisa diandalkan, dengan pengujian otomatis sebagai buktinya.

[▶ Lihat dokumentasi warehouse](#)[GitHub](#)

TEMUAN KUNCI

Dua penyedia sistem argo mencatat perjalanan yang sama dengan cara berbeda, sehingga menjumlahkan semua kolom tarif membuat \$2,50 terhitung dua kali pada 27% perjalanan. Satu vendor menggabungkan biaya kemacetan dan bandara ke dalam kolom extra; vendor lain mencatatnya terpisah. Pada Q1 2022 itu berarti **2,34 juta perjalanan kelebihan hitung sekitar \$5,9 juta**. Tidak ada tanda apa pun di data sumbernya, dan tidak ada grafik yang akan terlihat aneh.

YANG SAYA BANGUN

- Warehouse dbt bertingkat untuk **9,28 juta perjalanan**: 9 model dari data mentah sampai siap pakai, **70 pengujian data otomatis**, plus snapshot, seed, dan macro. Waktu build hanya ~34 detik di laptop.
- Cara pencatatan tarif diperiksa **baris per baris, bukan sekadar percaya pada VendorID**, karena ada 120 baris yang menyimpang dari aturan vendornya sendiri.
- **Data bulanan diperbarui dengan cara ditimpa, bukan ditambah**, karena TLC sering merevisi bulan yang sudah terbit. Kalau kuncinya per perjalanan, revisi itu akan masuk sebagai duplikat. CI menjalankan build dua kali lalu membandingkan jumlah barisnya, memastikan hasilnya selalu sama.
- Tingkat peringatan diatur per pengujian: **error** untuk kesalahan sendiri, **warn** untuk cacat data dari sumber yang tidak bisa diperbaiki. Batasnya diambil dari sebaran data nyata, bukan angka bulat kira-kira.

YANG DITEMUKAN OLEH PENGUJIAN

MASALAH YANG DITEMUKAN OLEH TES YANG GAGAL	BARIS
Jarak perjalanan tidak masuk akal	105.198
Perjalanan dibatalkan atau disengketakan	49.201
Durasi tidak masuk akal	10.038
Waktu turun lebih awal dari waktu naik	8.739
Tip negatif (3 dengan kode pembayaran tidak terdaftar)	894

Tidak ada yang dibuang diam-diam. Semuanya ditandai, dihitung, dan ditampilkan sebagai kolom pada tabel harian, sehingga sumber data yang rusak tetap bisa dibedakan dari penurunan permintaan yang sebenarnya.

SUMBER	STAGING	INTERMEDIATE	MARTS
yellow_tripdata_*	-> stg_tlc__yellow_trips		fct_trips (incremental)
green_tripdata_*	-> stg_tlc__green_trips	-> int_trips__	agg_trips_daily
fhvvh_tripdata_*	-> stg_tlc__fhvvh_trips	conformed	agg_zone_performance
taxi_zone_lookup	-> stg_tlc__zones		dim_zone, dim_date

Teknologi: dbt-core 1.9 · dbt-duckdb · DuckDB · dbt_utils · Python 3.12 · GitHub Actions (build, test, cek hasil konsisten, deploy dokumentasi)