

Data Predictive Insights ETL Pipeline

A pipeline that turns raw powerlifting results into a clean database, a dashboard and a baseline model.

BUILT WITH

Python

PostgreSQL

ETL

Power BI

Data Engineering

SSIS

8

normalised tables, plus a failed-rows table

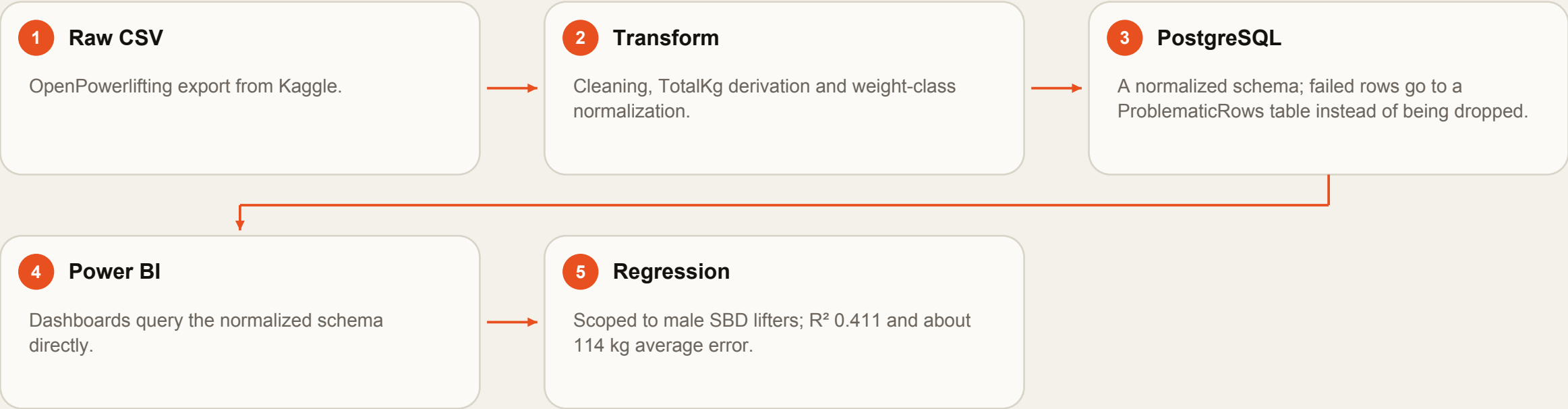
R2 0.411

regression baseline

~114 kg

average prediction error (RMSE)

How it works



WHAT EACH PART DOES

Extract	OpenPowerlifting CSV from Kaggle.
Transform	Cleaning, TotalKg derivation and weight-class normalisation.
Load	Normalised PostgreSQL schema for athletes, meets, lifts and performance.
Report and model	Power BI dashboards plus a scikit-learn regression with an auto-generated report.

Key decisions

01

A normalised schema, not one flat table

Splitting athletes, meets, equipment and lifts makes the data answer questions the way an analyst asks them.

02

A table for rows that fail

Failed inserts go to ProblematicRows with the error and raw data, so one bad row never aborts the load or vanishes.

03

Model a coherent group

The regression covers male competitors in full SBD meets with outliers removed, because mixed groups blur the signal.

04

Report the numbers as they are

41.1% of variance explained and a roughly 114 kg error are stated as a baseline for spotting trends, not a precise predictor.

Stack, status and what is next

BUILT WITH



KNOWN LIMITS

- The model is a baseline for trends, not a precision predictor.

SOURCE

<https://github.com/wailethewave/Data-Predictive-Insights-ETL>

