# ENTERPRISE DATA SOLUTION FOR CAR LISTING PROJECT

Presented by: Khaled Alruwita

Supervised by: Bandar Alqurashi

# OVERVIEW

01 Introduction

02 Business Context

03 Gathering Requirements

04 Data Modeling

05 Data Engineering

06 Data Quality

07 Business Intelligence

08 Predictive Analytics

09 Next Projects

# INTRODUCTION

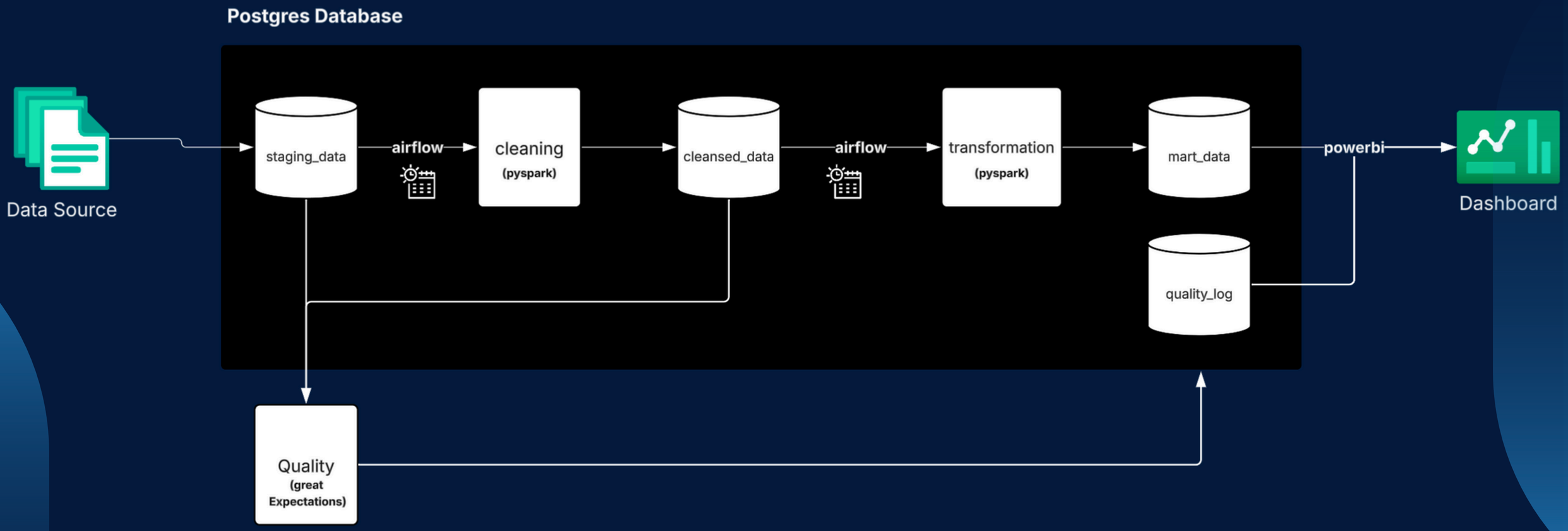
## BUSNIESS CONTEXT

A Car Listing Company wanted to implement a data infrastructure to gain deeper insights from their operational data

# GATHERING REQUIREMENTS

- Centralized Data Storage
- Automated Data Pipelines
- Data Quality Validations
- Business Intelligence Dasboard & Reporting
- Predictive Model for Car Prices

# SOLUTION DESIGN



01

# DATA MODELING

# GATHERING REQUIREMENTS

MOST SELLING MANUFACTURERS

MOST SELLING SIZE PER STATE

AVERAGE CAR AGE IN MARKET

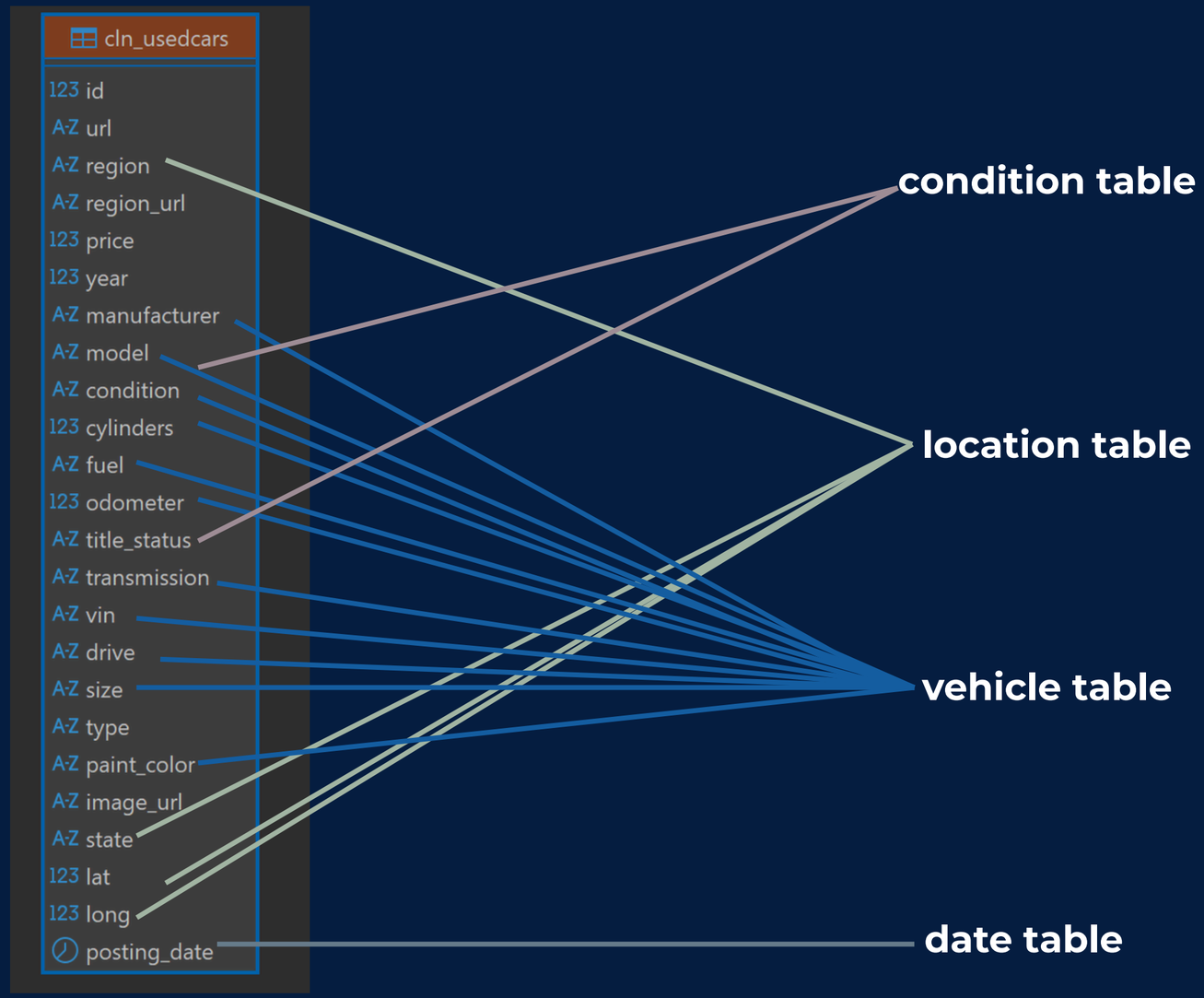
SUPPORT PREDICTIVE MODELING FOR PRICING

# DATA GRAIN

| Listing ID | Manufacturer | Model | Year | State | Price | Odometer |
|---|---|---|---|---|---|---|
| 1 | Toyota | Corolla | 2,018 | NM | $15,500 | 45,000 |
| 2 | Hyundai | i10 | 2,018 | NY | $1,200 | 43,000 |
| 3 | Kia | K5 | 2,022 | IL | $22,212 | 13,000 |

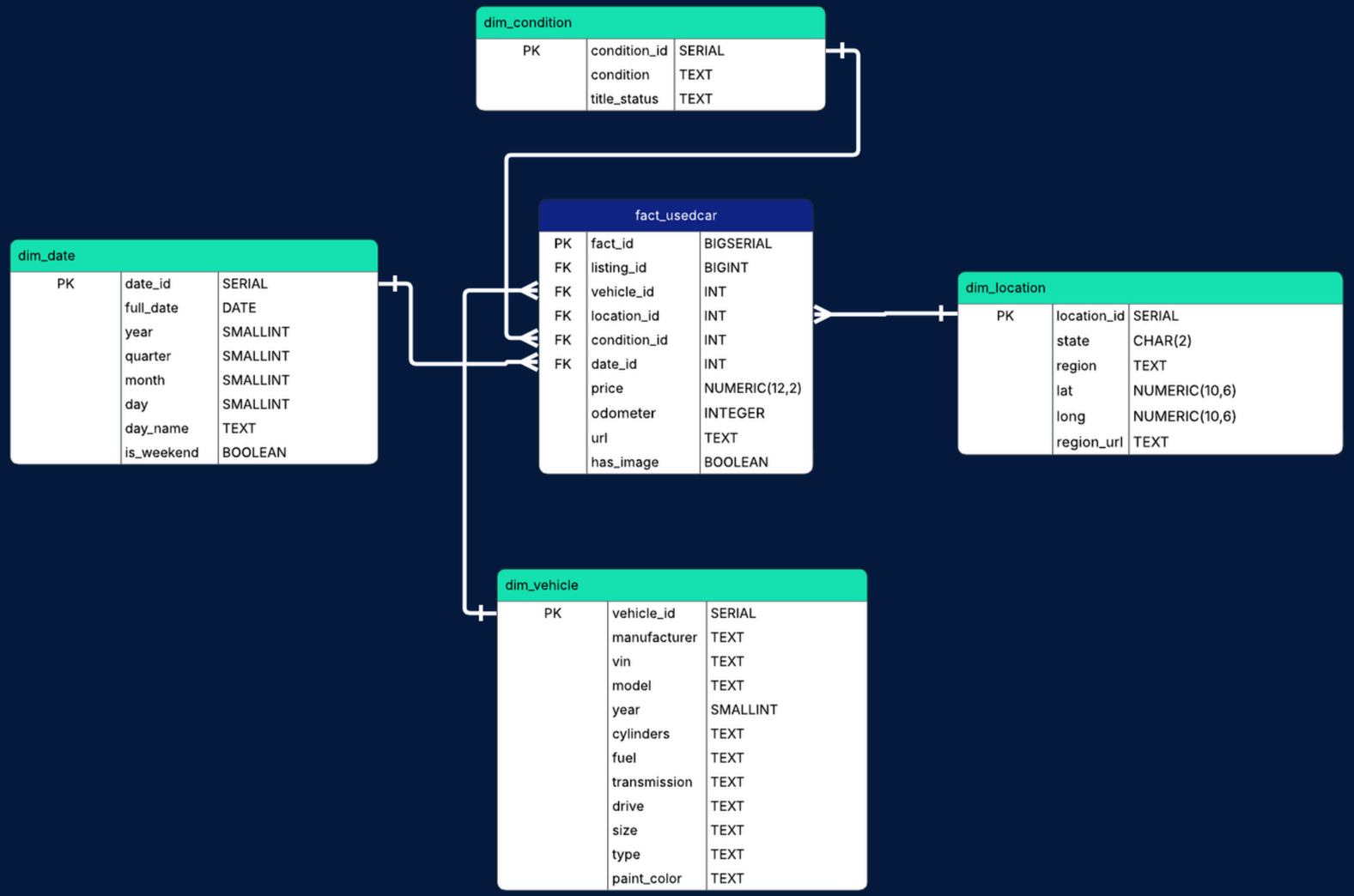
**We had to define the grain — the lowest level of detail. row represents one car listing.**

# DESIGNING STAR SCHEMA




dividing the data into a dims

# DESIGNING STAR SCHEMA




**model the data into a star-schema structure to align with the business requirements for analysis and reporting**

# DATA DICTIONARY

| Table Name | Description | Key Fields | Example Attributes |
|---|---|---|---|
| fact_usedcars | Main fact table | id | price, year, odometer |
| dim_vehicle | Vehicle details | vehicle_id | manufacturer, model |
| dim_location | Geographic info | location_id | region, state, lat, long |
| dim_condition | Car condition | condition_id | condition, title_status |
| dim_date | Posting date info | date_id | posting_date, year, month |

**defines each table in the data model, describing its purpose, key fields, and important attributes.**

02

# DATA ENGENERING

# TOOL USED

## PYSPARK

Cleaning and transforming large volumes of raw data efficiently.

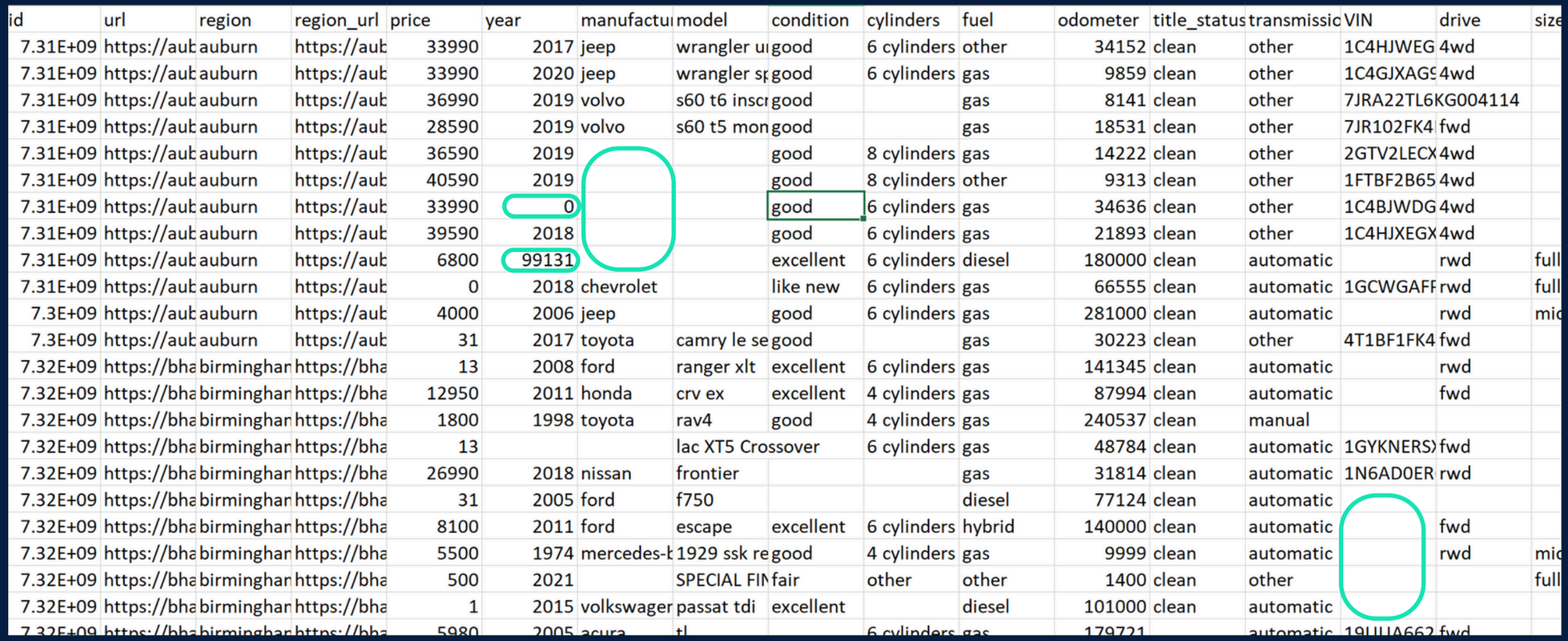
## POSTGRESQL

Centralized data warehouse and schema management

## APACHE AIRFLOW

Workflow orchestration, scheduling, and automation

# CLEANING THE DATA



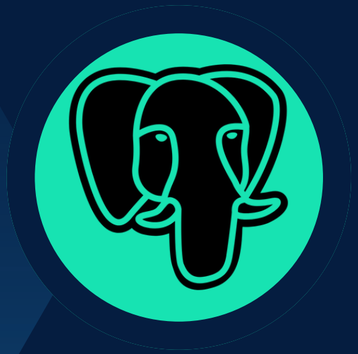
| id | url | region | region_url | price | year | manufactu | model | condition | cylinders | fuel | odometer | title_status | transmissio | VIN | drive | size |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7.31E+09 | https://aub | auburn | https://aub | 33990 | 2017 | jeep | wrangler u | good | 6 cylinders | other | 34152 | clean | other | 1C4HJWEG | 4wd | |
| 7.31E+09 | https://aub | auburn | https://aub | 33990 | 2020 | jeep | wrangler sp | good | 6 cylinders | gas | 9859 | clean | other | 1C4GJXAG5 | 4wd | |
| 7.31E+09 | https://aub | auburn | https://aub | 36990 | 2019 | volvo | s60 t6 inscr | good | | gas | 8141 | clean | other | 7JRA22TL6 | KG004114 | |
| 7.31E+09 | https://aub | auburn | https://aub | 28590 | 2019 | volvo | s60 t5 mon | good | | gas | 18531 | clean | other | 7JR102FK4 | fwd | |
| 7.31E+09 | https://aub | auburn | https://aub | 36590 | 2019 | | | good | 8 cylinders | gas | 14222 | clean | other | 2GTV2LEC | 4wd | |
| 7.31E+09 | https://aub | auburn | https://aub | 40590 | 2019 | | | good | 8 cylinders | other | 9313 | clean | other | 1FTBF2B65 | 4wd | |
| 7.31E+09 | https://aub | auburn | https://aub | 33990 | 0 | | | good | 6 cylinders | gas | 34636 | clean | other | 1C4BJWDG | 4wd | |
| 7.31E+09 | https://aub | auburn | https://aub | 39590 | 2018 | | | good | 6 cylinders | gas | 21893 | clean | other | 1C4HJXEG | 4wd | |
| 7.31E+09 | https://aub | auburn | https://aub | 6800 | 99131 | | | excellent | 6 cylinders | diesel | 180000 | clean | automatic | | rwd | full |
| 7.31E+09 | https://aub | auburn | https://aub | 0 | 2018 | chevrolet | | like new | 6 cylinders | gas | 66555 | clean | automatic | 1GCWGAFF | rwd | full |
| 7.3E+09 | https://aub | auburn | https://aub | 4000 | 2006 | jeep | | good | 6 cylinders | gas | 281000 | clean | automatic | | rwd | mic |
| 7.3E+09 | https://aub | auburn | https://aub | 31 | 2017 | toyota | camry le se | good | | gas | 30223 | clean | other | 4T1BF1FK4 | fwd | |
| 7.32E+09 | https://bha | birminghan | https://bha | 13 | 2008 | ford | ranger xlt | excellent | 6 cylinders | gas | 141345 | clean | automatic | | rwd | |
| 7.32E+09 | https://bha | birminghan | https://bha | 12950 | 2011 | honda | crv ex | excellent | 4 cylinders | gas | 87994 | clean | automatic | | fwd | |
| 7.32E+09 | https://bha | birminghan | https://bha | 1800 | 1998 | toyota | rav4 | good | 4 cylinders | gas | 240537 | clean | manual | | | |
| 7.32E+09 | https://bha | birminghan | https://bha | 13 | | | lac XT5 Cro | ssover | 6 cylinders | gas | 48784 | clean | automatic | 1GYKNERS | fwd | |
| 7.32E+09 | https://bha | birminghan | https://bha | 26990 | 2018 | nissan | frontier | | | gas | 31814 | clean | automatic | 1N6AD0ER | rwd | |
| 7.32E+09 | https://bha | birminghan | https://bha | 31 | 2005 | ford | f750 | | | diesel | 77124 | clean | automatic | | | |
| 7.32E+09 | https://bha | birminghan | https://bha | 8100 | 2011 | ford | escape | excellent | 6 cylinders | hybrid | 140000 | clean | automatic | | fwd | |
| 7.32E+09 | https://bha | birminghan | https://bha | 5500 | 1974 | mercedes-b | 1929 ssk re | good | 4 cylinders | gas | 9999 | clean | automatic | | rwd | mic |
| 7.32E+09 | https://bha | birminghan | https://bha | 500 | 2021 | | SPECIAL FIN | fair | other | other | 1400 | clean | other | | | full |
| 7.32E+09 | https://bha | birminghan | https://bha | 1 | 2015 | volkswager | passat tdi | excellent | | diesel | 101000 | clean | automatic | | | |
| 7.32E+09 | https://bha | birminghan | https://bha | 5980 | 2005 | acura | tl | | 6 cylinders | gas | 179721 | | automatic | 19UUA662 | fwd | |

- **null critical field**
- **invalid year**
- **Remove extreme outliers in price and odometer.**
- **Fix formatting (**

# PYSPARK

- **REMOVING DUPLICATE** — Eliminate repeated records to ensure unique and consistent data entries.
- **HANDLE MISSING DATA** — Fill, remove, or flag incomplete values to maintain data integrity.
- **STANDARDIZE THE DATA** — Ensure consistent formats (e.g., date, casing, units) across all records.
- **NORMLIZE DATA TYPES** — Convert columns to proper data types for accurate processing and analysis.
- **CHECK FOR OUTLIERS** — Detect and handle abnormal values that may distort insights.

# POSTGRESQL



## POSTGRESQL SETUP

installed and configured a PostgreSQL server as the target database for storing and modeling the cleaned cars data.

## DIMENSION TABLES

created dimension tables first (dim_vehicle, dim_location, dim_date, dim_condition)

## FACT TABLE

built the central fact_usedcar table to hold listing-level with foreign keys referencing all dimension tables.

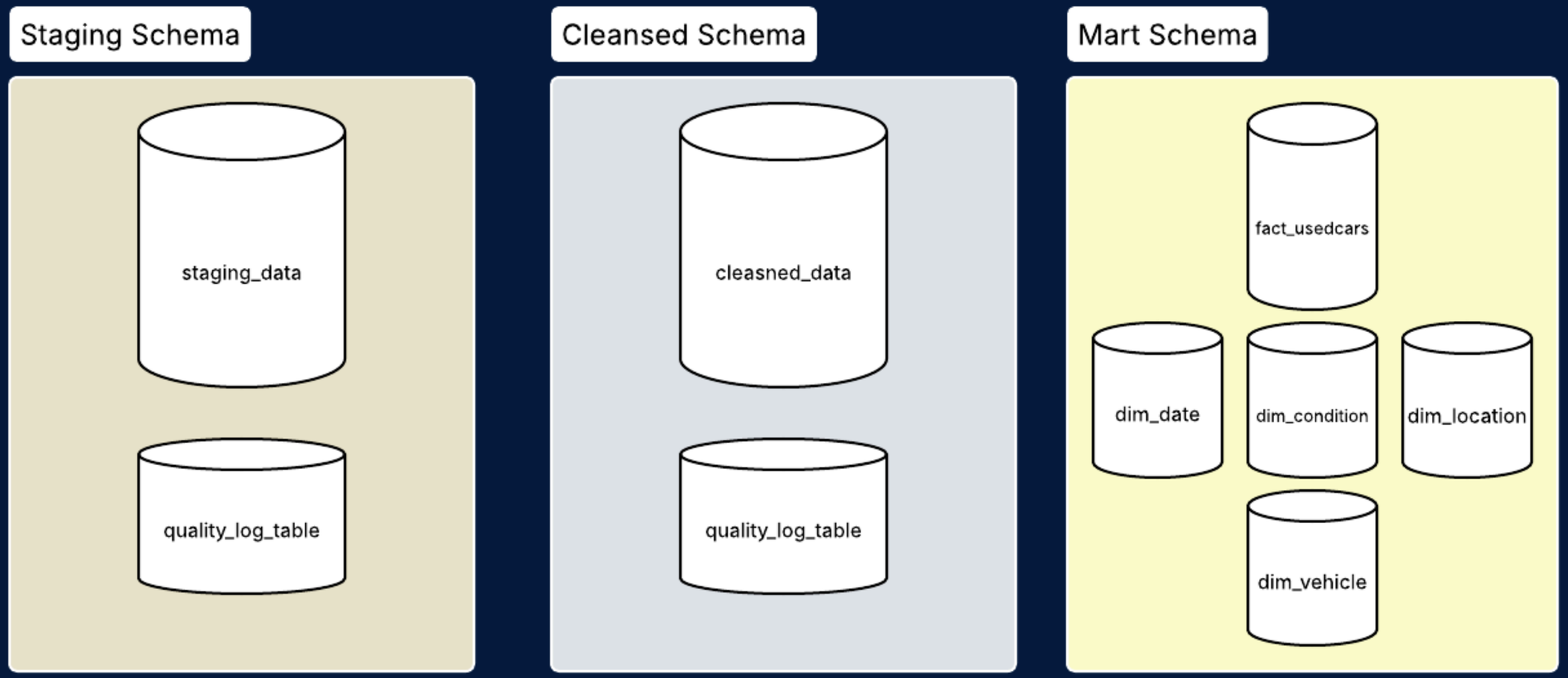
## DATA LOADING & VALIDATION

inserted fact records by joining to the dims.& verified counts and checked for rows to ensure integrity.

## PERFORMANCE OPTIMIZATION

Created indexes, views, and optimized query execution plans for faster Power BI reporting and analytical queries.

# DATA WAREHOUSE SCHEMA ARCHITECTURE




**architecture illustrates how data is organized in the PostgreSQL data warehouse.**

# APACHE AIRFLOW

## AIRFLOW SETUP

Installed and configured Apache Airflow to orchestrate the ETL workflow, automating data extraction, cleaning, transformation, and loading tasks.

## DAG CREATION

Developed custom DAGs (Directed Acyclic Graphs) to define task dependencies for staging, cleaning, and transformation steps using PySpark and SQL scripts.

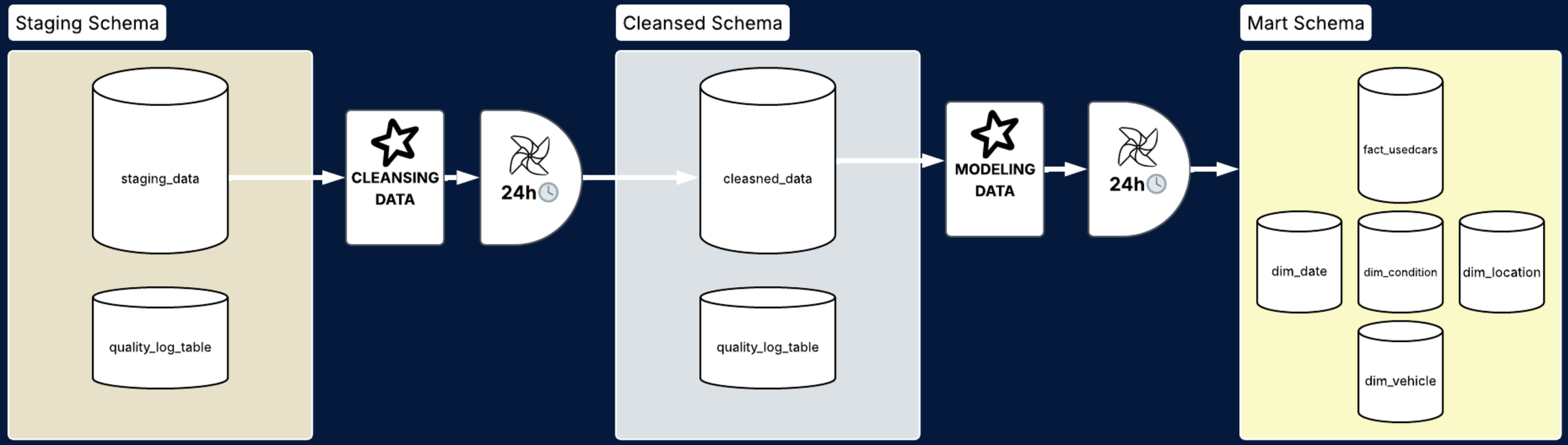
## SCHEDULING & AUTOMATION

Scheduled DAGs to run daily for continuous data updates, ensuring that new listings and quality checks are processed automatically.

## MONITORING & ALERTS

Enabled task monitoring through Airflow's UI and configured email/Slack alerts for task failures or performance delays.

# DATA PIPELINE ORCHESTRATING




orchestrating the ETL workflow: staging → cleansed → mart

03

# DATA QUALITY

# TOOL USED

## GREAT EXPECTATION

Used for running automated data quality checks and validating datasets to ensure accuracy and completeness.

## POSTGRESQL

Serves as the central database for storing validated data and logging all quality check results (**quality_log**).

## POWER BI

Used for visualizing insights and monitoring data quality metrics through interactive dashboards.

# DATA QUALITY DIMENSIONS

### COMPLETENESS

All required data fields are filled with no missing or null values.

### ACCURACY

Data correctly represents real-world values without errors or distortion.

### CONSISTENCY

Data remains uniform and aligned across different tables and sources.

### TIMELINESS

Data is updated and available when needed for decision-making.

### VALIDITY

Data follows defined rules, formats, and business constraints.

### UNIQUENESS

Each record is distinct with no duplicates across the dataset.

# GREAT EXPECTATIONS

## DATA SOURCE CONNECTION

Connected (schema: cleansed) as a data source and loaded the cln_usedcars table into the validation context.

## SUITE CREATION

Built a suite named containing 24 automated validation rules to assess data quality across all key columns.

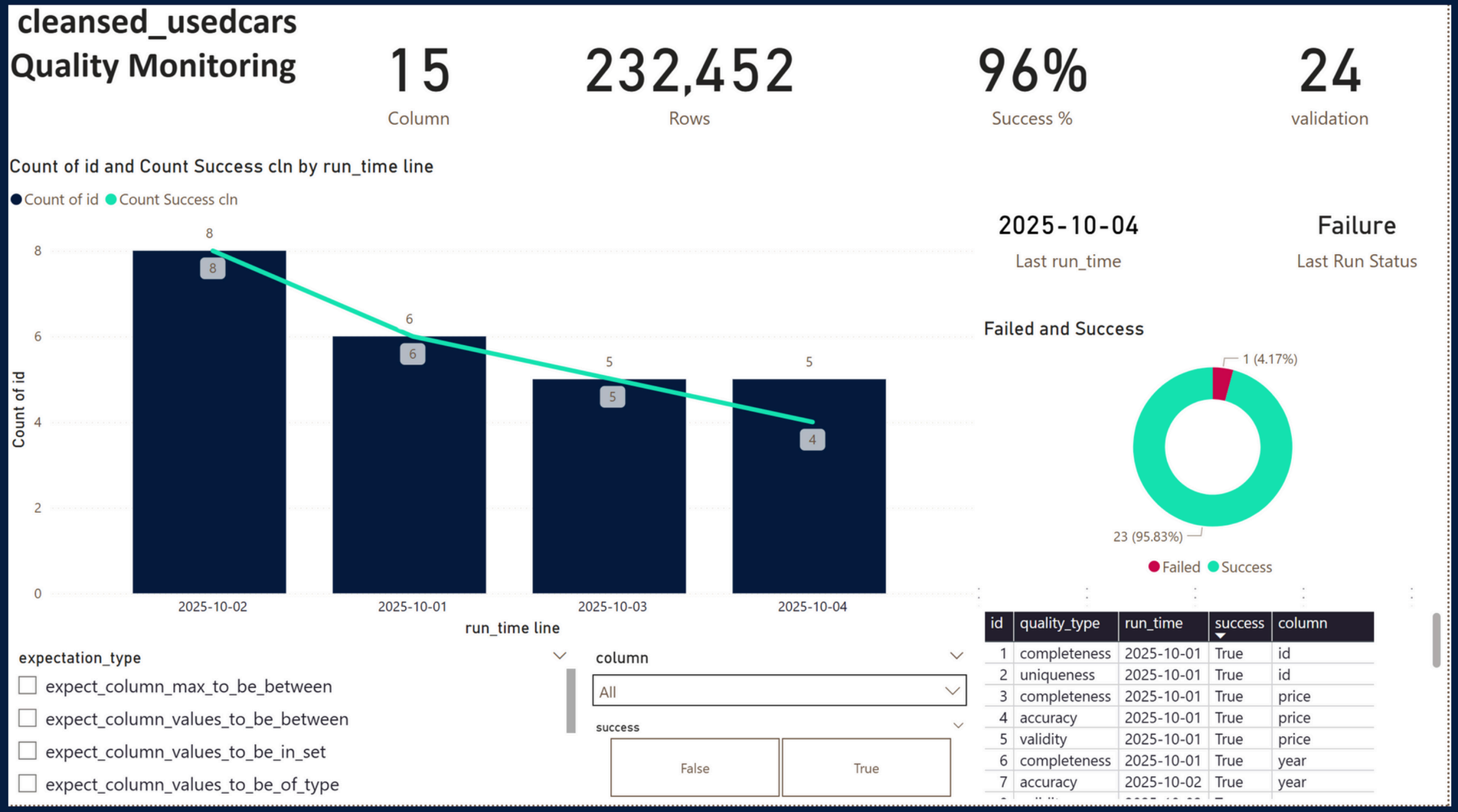
## VALIDATION TESTS

for Completeness, Accuracy, Consistency, Timeliness, Validity, and Uniqueness

## RESULT LOGGING

Stored validation outcomes (24 successful checks, 100% pass rate) in log DataFrame

# DATA QUALITY MONITOR

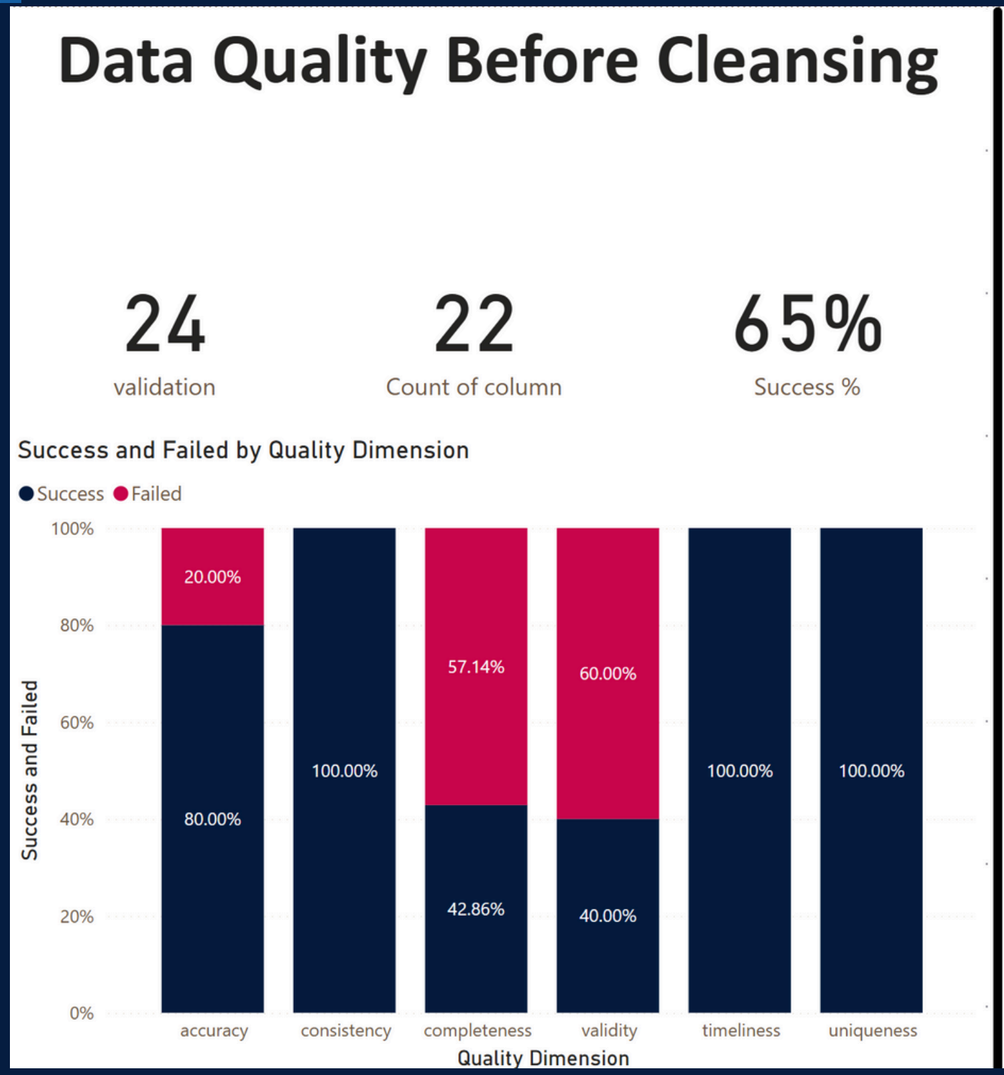


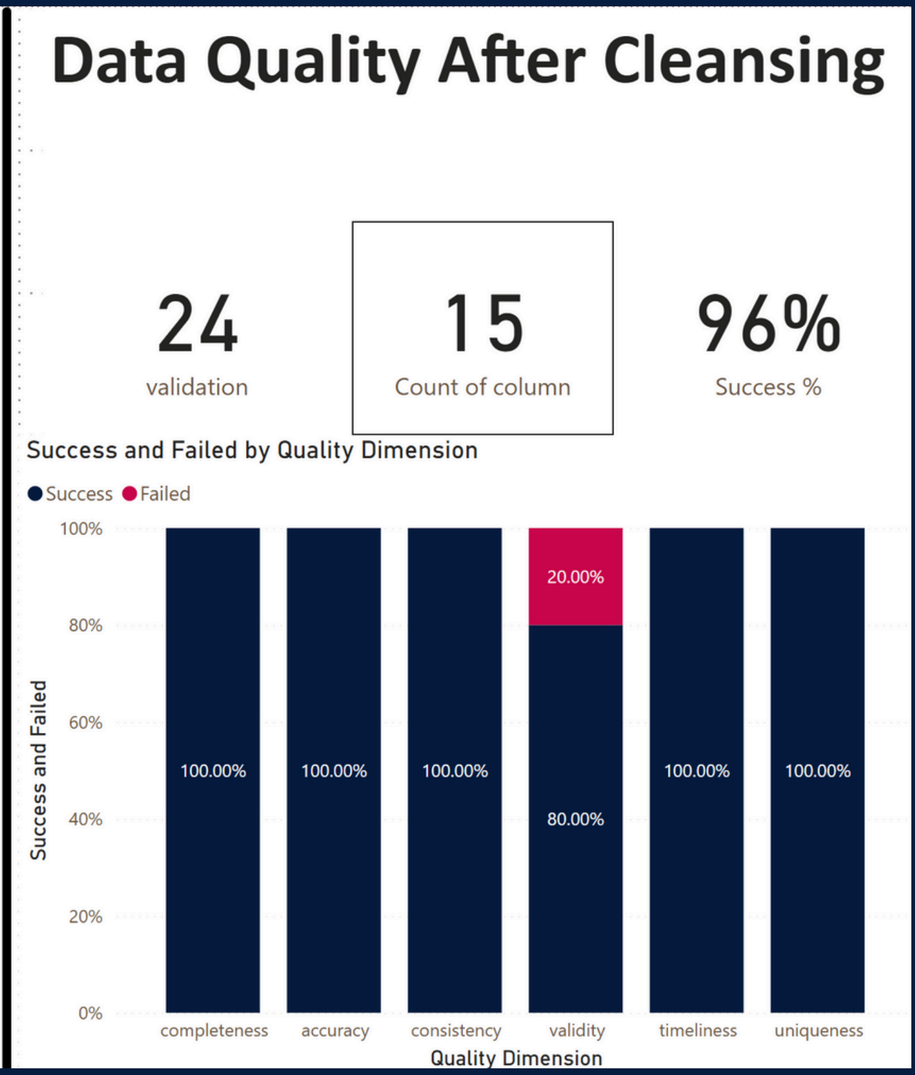

**Real-time Interactive Power BI dashboard showing daily data quality results from Great Expectations, tracking validation success rates**

# DATA QUALITY MONITOR


success rates across all data quality dimensions

# DATA QUALITY MONITOR




## Comparison dashboard BRFORE VS AFTER Cleansing

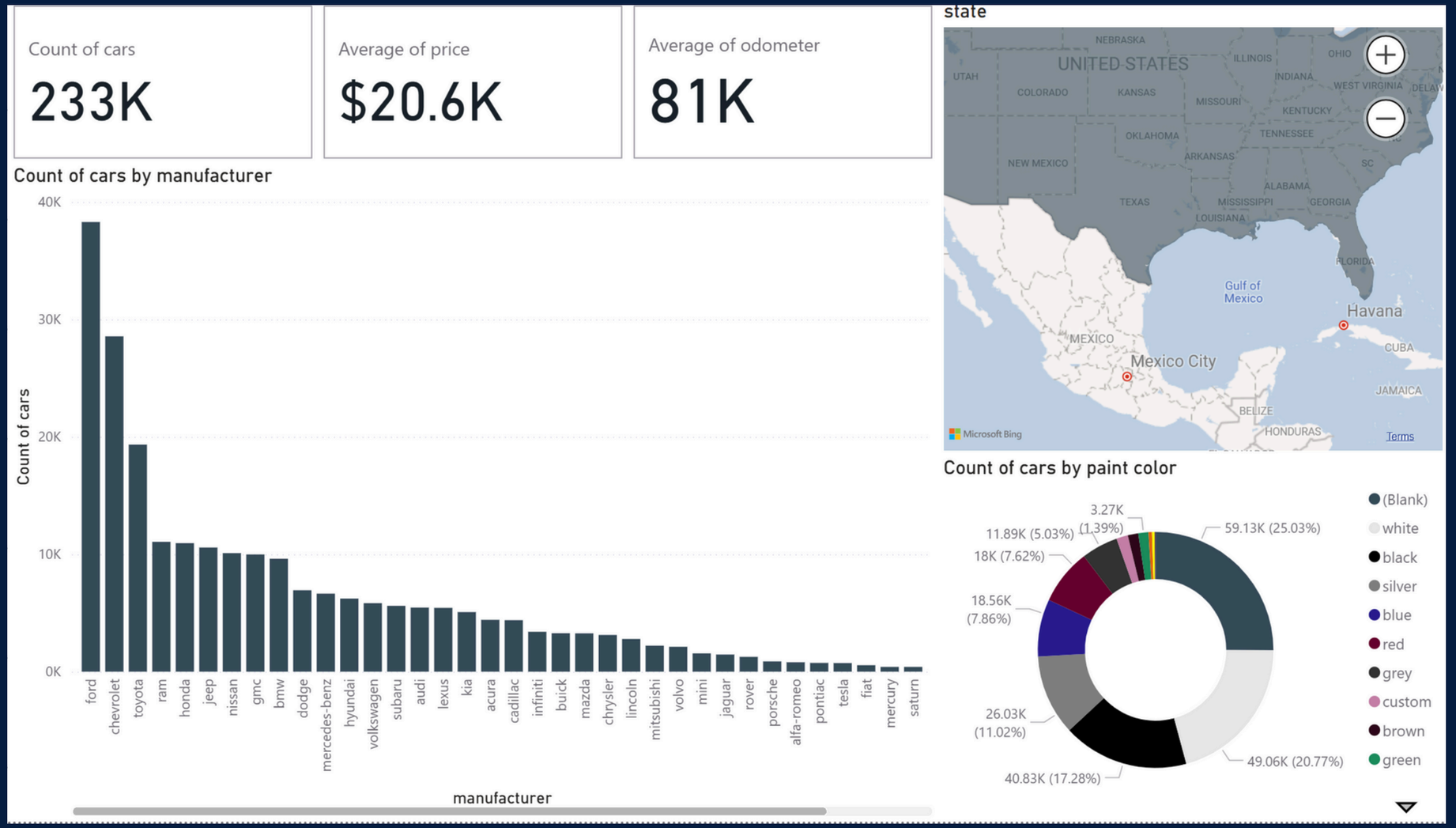
04

# DATA DASHBOARDS

# BUSINESS DASHBOARD


**Interactive dashboard visualizing car listings — showing total listings, average prices, mileage, posting trends over time**

# BUSINESS DASHBOARD




**Interactive dashboard visualizing car listings — showing total listings, per manufacturer**

05

# DATA PREDECTION

# TOOL USED

## SCIKIT-LEARN

for data preprocessing and model evaluation with metrics like $R^2$, MAE, RMSE, and MAPE.

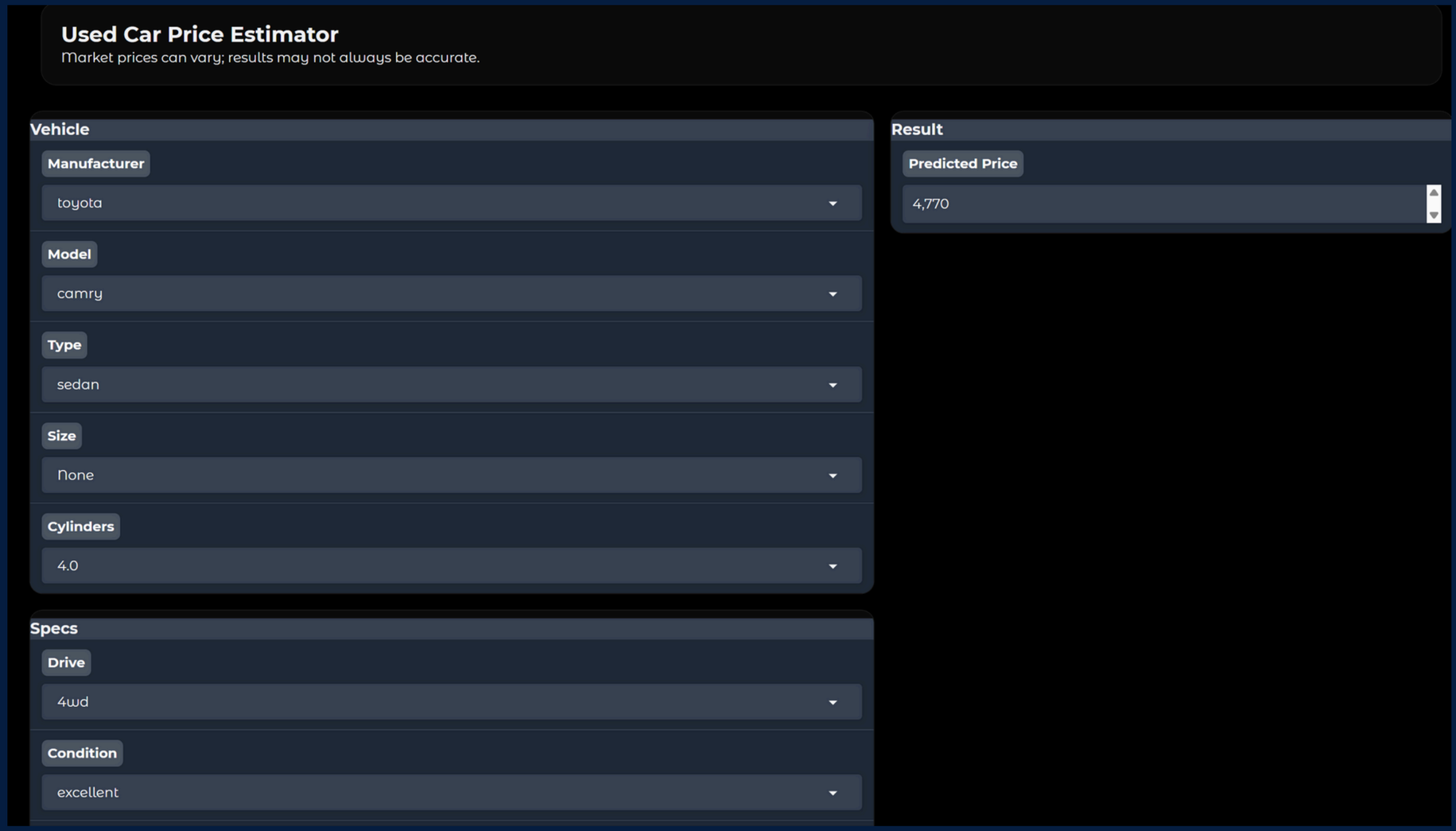
## LIGHTGBM

Implemented the regression model for predicting car prices

## GRADIO

Built an interactive web interface allowing users to input car details and get a predicted price

# PRICE SUGGESTION SYSTEM




## Price Suggestion System using the prediction model

# PREDECTION RESULTS

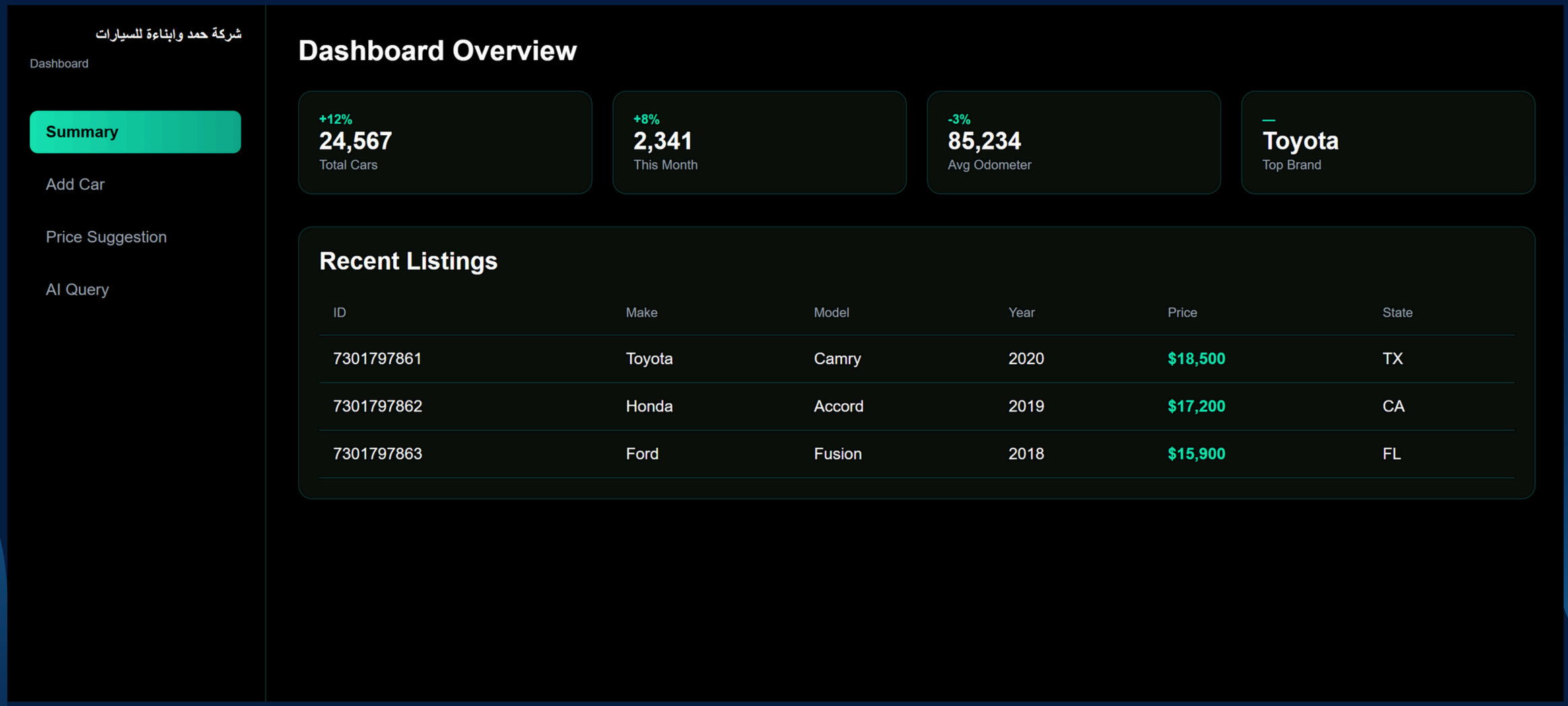
**85%** | $R^2$ SCORE

**15%** | MAPE

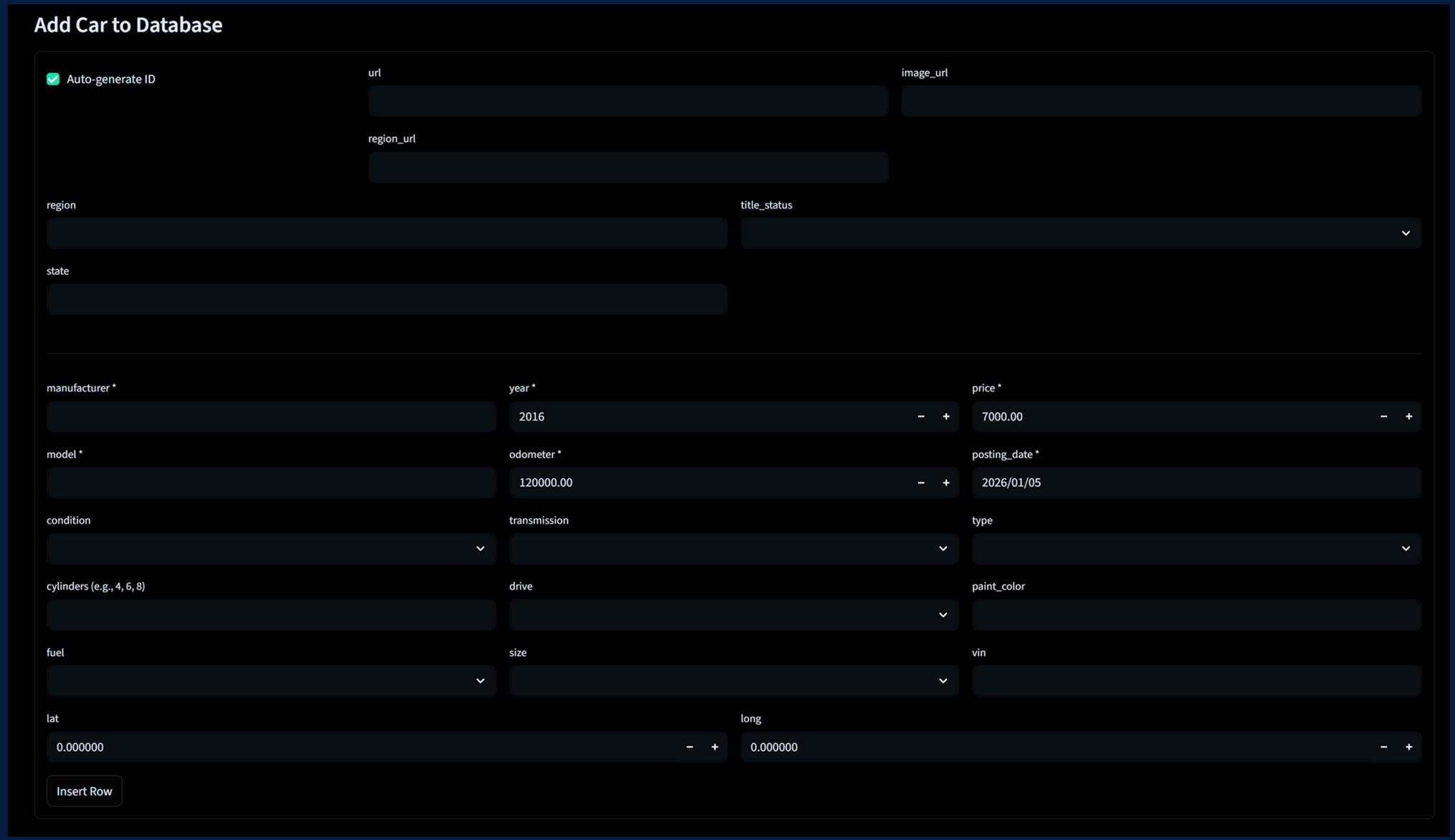
06

# FRONT END SYSTEM

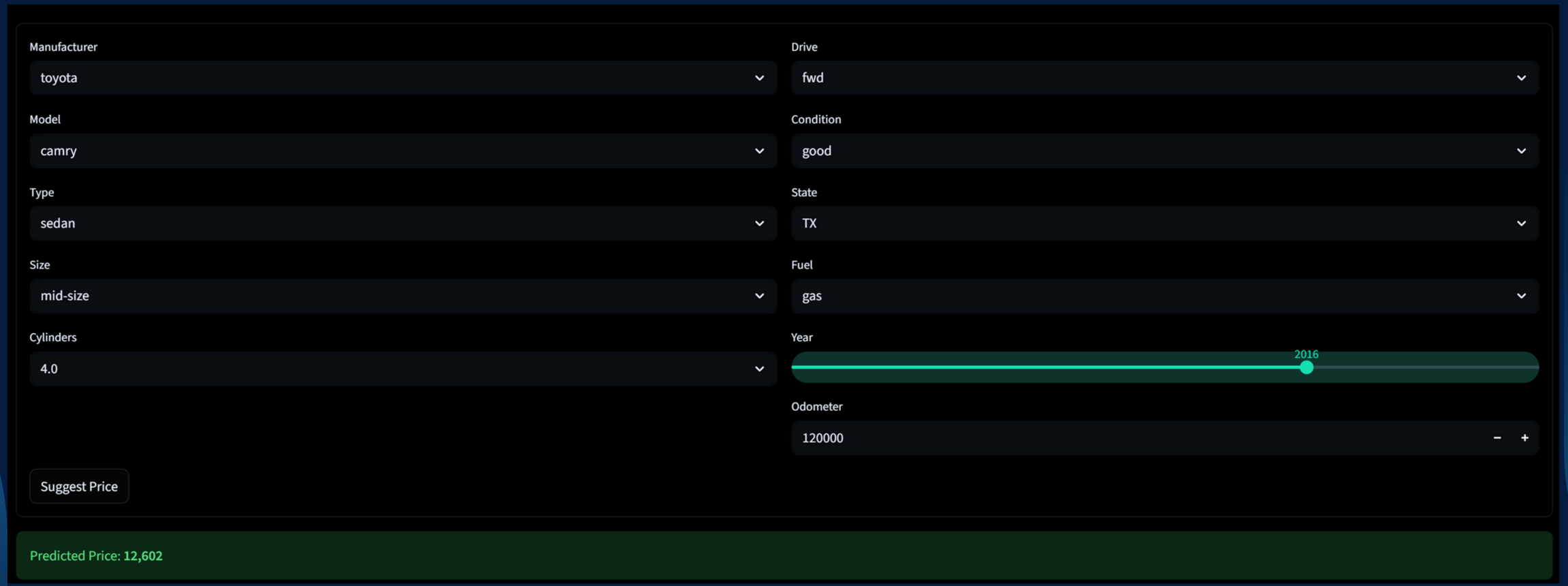
# HOME PAGE


شركة حمد وابناءة للسيارات
Dashboard
Summary
Add Car
Price Suggestion
AI Query
Dashboard Overview
+12%
24,567
Total Cars
+8%
2,341
This Month
-3%
85,234
Avg Odometer
—
Toyota
Top Brand
Recent Listings
ID
Make
Model
Year
Price
State
7301797861
Toyota
Camry
2020
$18,500
TX
7301797862
Honda
Accord
2019
$17,200
CA
7301797863
Ford
Fusion
2018
$15,900
FL

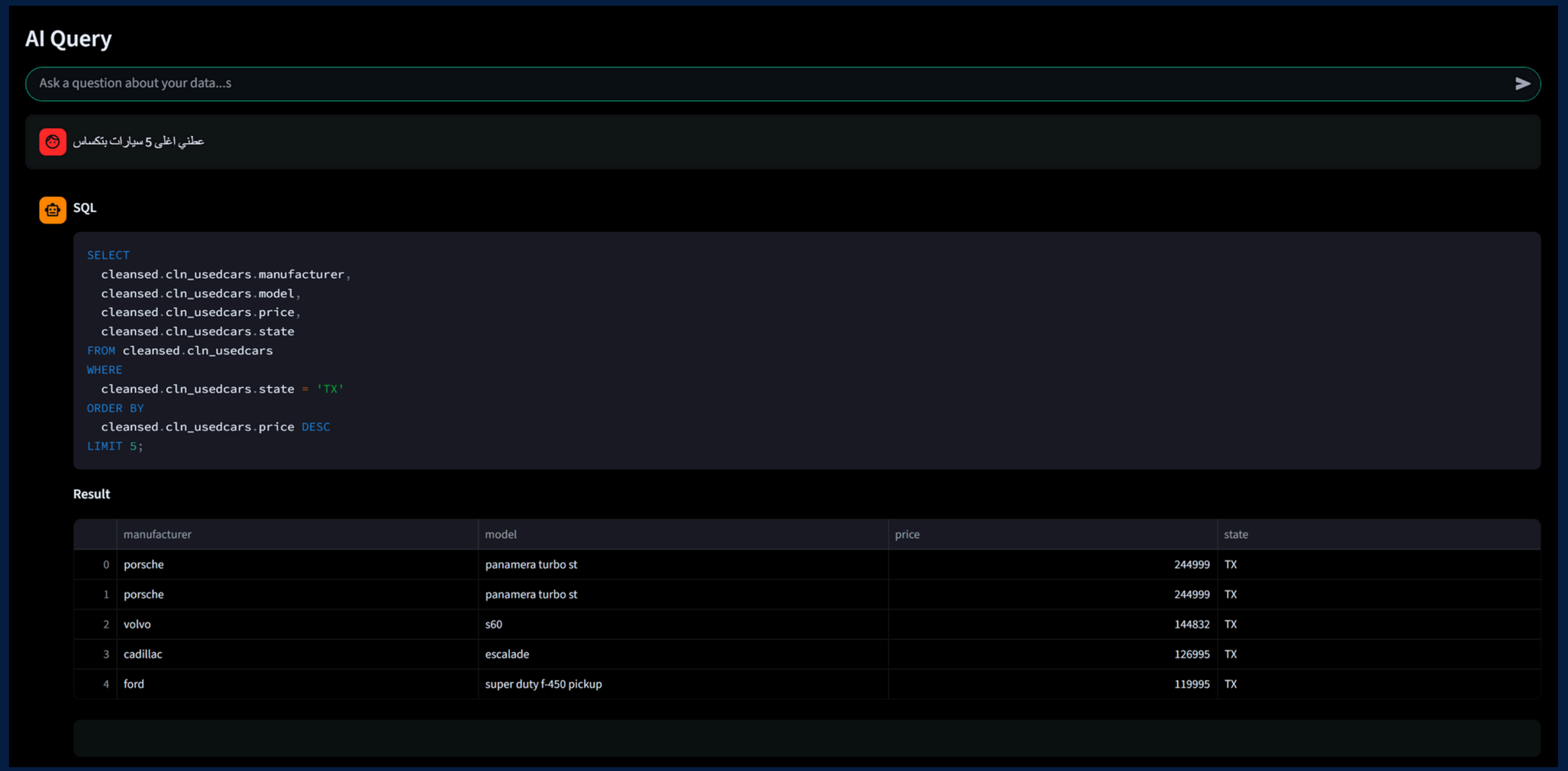
# INSERT PAGE


Add Car to Database
Auto-generate ID
url
image_url
region_url
region
title_status
state
manufacturer *
year *
2016
price *
7000.00
model *
odometer *
120000.00
posting_date *
2026/01/05
condition
transmission
type
cylinders (e.g., 4, 6, 8)
drive
paint_color
fuel
size
vin
lat
0.000000
long
0.000000
Insert Row

# CAR PRICE SUGGESTION



# AI QUERY


AI Query
Ask a question about your data...s
عطني اغلى 5 سيارات بتكساس
SQL
SELECT
cleansed.cln_usedcars.manufacturer,
cleansed.cln_usedcars.model,
cleansed.cln_usedcars.price,
cleansed.cln_usedcars.state
FROM cleansed.cln_usedcars
WHERE
cleansed.cln_usedcars.state = 'TX'
ORDER BY
cleansed.cln_usedcars.price DESC
LIMIT 5;
Result
manufacturer model price state
0 porsche panamera turbo st 244999 TX
1 porsche panamera turbo st 244999 TX
2 volvo s60 144832 TX
3 cadillac escalade 126995 TX
4 ford super duty f-450 pickup 119995 TX

# THANK YOU