Auto & Commodity Data Collection & Analysis

Analyzing the Relationship Between Used Car & Commodity Prices

Group 206 — Dongyuan Gao, Ramiro Diez-Liebana, Cyriel Van Helleputte

November 2025

Contents

1 Project Overview

1.1 Business Problem

AutoHelvetia AG, a national used-car dealer, struggles to align pricing and procurement with volatile commodity markets. We were tasked with quantifying how raw material costs spill over into Swiss used car prices.

1.2 Our Solution

We built a web + API data pipeline, harmonized listings with commodity curves, and analyzed correlations so the client can time purchases and price vehicles with commodity context.

1.3 Key Findings Summary

- **Moderate commodity influence**: The composite index spiked post-2020, yet used-car medians barely moved—commodity effects exist but are muted.
- **Minimal powertrain contrast**: Correlations stay below 0.20 across petrol, diesel, and electric segments; vehicle attributes dominate.
- **Brand sensitivities**: Volvo/Porsche align with battery metals, Toyota/Cupra skew negative; copper and oil drive most divergences.
- **Recommendations**: Track copper/cobalt before sourcing premium EV stock, price luxury models with commodity signals, and lean on resilient brands when metals rally.

2 Feasibility

- **Ethical**: Scraping targets only public listing pages allowed by AutoScout24.ch's robots.txt and stays within academic-use terms while avoiding gated or user data.
- **Technical**: A Selenium + BeautifulSoup pipeline with modest throttling reliably captures listings and Yahoo Finance prices without hitting rate limits.
- **Analytical**: The combined dataset supports correlation, regression, and time-series exploration, yielding actionable pricing diagnostics for AutoHelvetia AG.

3 Data Sources & Collection

3.1 Web Scraping Implementation

3.1.1 Target Website

- **Primary Source**: AutoScout24.ch (https://www.autoscout24.ch).
- **Target Path**: `/de/autos/alle-marken` (All car listings).
- **Scope**: Used car listings across all makes and models available on the platform.

3.1.2 Technical Implementation

3.1.2.1 Core Toolkits

- **Selenium WebDriver**: For browser automation and dynamic content loading.
- **BeautifulSoup4**: For HTML parsing and data extraction.
- **Custom Parser**: Combines multiple extraction methods(json, html, css, regex) for robustness.

3.1.2.2 Scraping Methodology

1. **Pagination Handling**:
   - Iterates through listing pages systematically and click on next page.

- Implements smart navigation with randomized delays (5-15s between pages).

2. **Data Extraction Strategy**:
   - **Primary Method**: Structure-aware parsing using SVG icon titles and sibling elements.
   - **Combination of Methods**:
     - JSON structured data extraction.
     - CSS class-based element targeting.
     - Regular expression fallbacks for critical fields.

3.1.3 Data Points Collected

Data Field	Description	Example
`car_model`	Full vehicle make and model	"Volkswagen Golf 2.0 TDI"
`price_chf`	Listing price in CHF	25,900
`mileage`	Vehicle mileage in km	85,200
`engine_power_hp`	Engine power in HP	150
`power_mode`	Fuel/power type	Diesel, Petrol, Electric, Hybrid
`transmission`	Transmission type	Automat, Manuell, Halbautomatik
`production_date`	Production date	2018
`listing_url`	Direct URL to the listing	[Link]

3.2 Yahoo FinanceAPI Integration

We fed our commodity pipeline using the **yfinance Python library**. This library (over 20k stars in Github, (https://github.com/ranaroussi/yfinance)) gives access Yahoo Finance's public endpoints **without requiring API authentication**.It is not affiliated, to Yahoo, Inc. It's an **open-source tool that uses Yahoo's publicly available APIs**.

3.2.1 Technical Implementation

fetched historical **daily closing prices** for all tickers in the list from Yahoo Finance. If it doesn't find closing price, it falls back to an adj close column.

The core yfinance function used is `yf.download()`. It fetches historical daily closing prices for all tickers in the list from Yahoo Finance. If it doesn't find closing price, it falls back to an adj close column.

3.2.2 Data Points Collected

Data Field	Description	Example
`Date`	Trading day	`2024-07-31`
`Month`	Month and year in `MM-YYYY` format	`07-2024`
`WTI_Spot`	Closing price of crude oil (`CL=F`)	`81.32`
`Copper_Spot`	Closing price of COMEX copper futures (`HG=F`)	`4.32`
`Lithium_Spot`	Proxy for lithium prices (`LIT`)	`57.89`
`Aluminium_Spot`	Closing price of LME aluminum futures (`ALI=F`)	`2235.00`
`Steel_Spot`	Closing price of U.S. steel futures (`HRC=F`)	`1015.00`
`Nickel_Spot`	Global nickel prices – (`NIC.AX`)	`17345.00`
`Cobalt_Spot`	Proxy for cobalt prices – (`603799.SS`)	`92.40`

- **Rate limits and handling**: While it does not have official request limits to call the tool, it still accesses Yahoo, and if the website implements changes or rate limits per IP or token that could be a problem with a more frequent use of the tool.

4 Data Cleaning and Transformation

Across the project we apply a standard data-science cleaning cadence: validate dataframe, coerce types, handle missing values with data quality strategies, and normalize key features before exporting analysis-ready datasets.

4.0.1 Autoscout Listing Standardization

Script: `Data/clean_data/Autoscout_Cleaner_Standardizer.py`

- **Brand & Model Extraction**: Uses regex-based patterns to parse the `car_model` field into `brand` and base `model` tokens, removing unwanted information (e.g., VW TIGUAN TSI 2.0 S VERSION BERN TOP ZUSTAND vs VW TIGUAN TSI 2.0 S).

4.0.2 Commodity Price Cleaning

Script: `Data/clean_data/load_data_cleaning.py`

- **Type Coercion**: Converts `Date` to datetime and commodity columns to numeric by replacing European decimal separators.
- **Missing Value Strategy**:
  - Reports gaps before/after processing for clarity.
  - Fills missing commodity prices with a 7-day rolling mean, then rounds to two decimals.

4.0.3 Scraper Output Post-processing

Script: `Data/Scraping/Scraper.py` - **Schema Consistency**: The scraper normalizes fuel types, transmission labels, and numeric fields ("N/A" fallbacks) to reduce later data cleaning steps.

- **Hybrid Extraction**: Combines JSON-LD fields with HTML parsing, ensuring critical attributes (`price_chf`, `mileage`, `production_date`, `listing_url`, etc.) are captured.

4.0.4 Research Dataset Post-Processing

Script: `Analysis/RQ3/RQ3_Analysis.py` - **Autoscout Cleaning & Missing Value Imputation**:

- Converts `production_date` strings (including "Neues Fahrzeug") to October 2025 and makes a `Month` period column.
- Imputes continuous fields (`price_chf`, `mileage`, `engine_power_hp`) with rounded means.
- Customized fill for `consumption_l_per_100km` (EV=0, then model mean, brand mean, global mean).

5 Analysis & Visualization

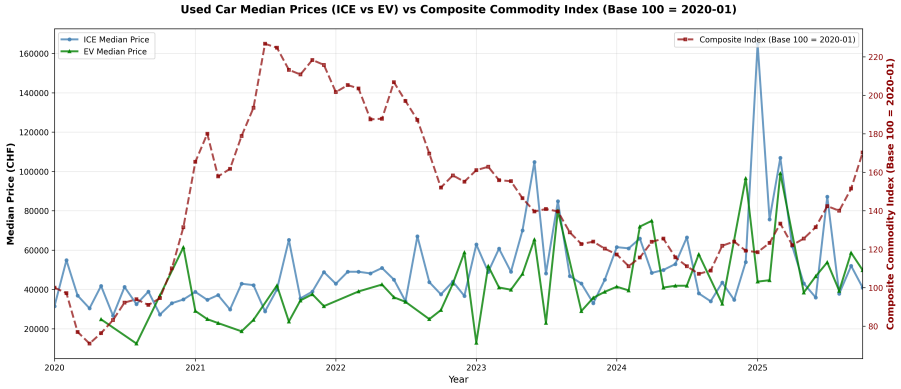
5.1 Research Questions

We tested three hypotheses: (1) whether a composite commodity index co-moves with Swiss used car prices, (2) whether powertrains (petrol, diesel, electric, hybrid) react differently to commodity swings, and (3) whether brand segments exhibit divergent exposure to raw material costs.

6 Results and Findings

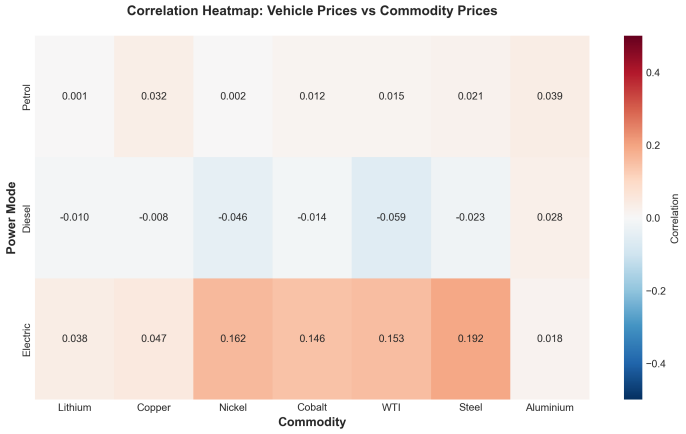
6.1 RQ1 — Commodity Index vs. Used Car Prices

Commodity prices crashed in early 2020, then spiked through 2022 while used car medians stayed comparatively flat—yielding only a modest co-movement signal.



6.2 RQ2 — Do different power modes exhibit different commodity price sensitivities?

We expected powertrains to react differently to commodity swings, yet all correlations stayed below 0.20.

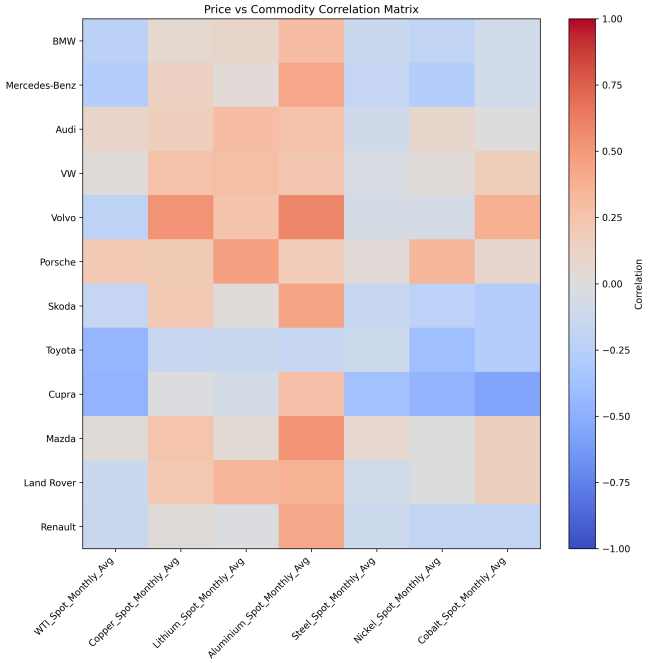


Vehicle characteristics (mileage, engine power, age) overshadow commodity effects in the used market.

6.3 RQ3 — Brand-Level Differences

Key takeaways:

- **Brand dispersion**: Volvo and Porsche track battery metals positively, while Toyota and Cupra tilt negative—showing uneven commodity pass-through.
- **Luxury vs. volume**: Premium marques retain pricing power despite metal swings; mass-market European brands mirror industrial metal moves more closely.
- **Commodity highlights**: Copper is the most consistent positive correlate; higher oil prices often coincide with softer ICE valuations.



7 Conclusion and Limitations

7.1 Conclusion

Used car prices exhibit only moderate co-movement with commodity markets overall, but brand and powertrain nuances matter. The 2021–2022 commodity surge highlighted where AutoHelvetia can react fastest: premium electrified brands shadow battery metals, while mass-market ICE models respond more to oil and industrial metals.

7.2 Recommendations for our Client

1. **Time purchasing** around copper/cobalt dips before buying premium EV-heavy inventory.
2. **Embed commodity indices** as leading signals in pricing dashboards for luxury segments.
3. **Adjust stock mix** in bull markets—favor resilient Toyota/Cupra models, apply surcharges to exposed volume brands, and keep premium pricing steady.
4. **Explore new-vehicle corners** if tighter commodity transmission is needed; production costs reflect market moves faster than used valuations.

7.3 Limitations

- **Temporal**: 2020–2025 includes pandemic-era shocks that may not generalize.
- **Causality**: Correlations capture association only; policy and consumer shifts remain confounders.
- **Geography**: Swiss market dynamics limit direct exportability.
- **Data quality**: Web listings can be inconsistent; premium data sources (CME, Bloomberg) would tighten accuracy.