AutoCommodity 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 **3**

1.1 The Storyline (Potential Business Problem) . . . 3

1.2 Our Solution . . . 3

1.3 Project Structure . . . 3

1.4 Initiall Findings . . . 3

2 Feasibility Research **3**

2.1 Ethical Feasibility of Web Scraping AutoScout24.ch . . . 3

2.1.1 Technical Feasibility . . . 4

2.1.2 Analytical Feasibility . . . 4

3 Data Sources & Collection **4**

3.1 Web Scraping Implementation . . . 4

3.1.1 Target Website . . . 4

3.1.2 Technical Implementation . . . 4

3.1.3 Data Points Collected . . . 4

3.2 Yahoo FinanceAPI Integration . . . 5

3.2.1 Technical Implementation . . . 5

3.2.2 Data Points Collected . . . 5

4 Data Transformation and Cleaning **5**

4.0.1 Autoscout Listing Standardization . . . 6

4.0.2 Commodity Price Cleaning . . . 6

4.0.3 Scraper Output Post-processing . . . 6

4.0.4 Research Dataset Post-Processing . . . 6

5 Analysis & Visualization **7**

5.1 Research Questions 7

6 Results and Findings **7**

6.1 RQ1 — Commodity Index vs. Used Car Prices 7

6.2 RQ2 — Powertrain Sensitivity 8

6.3 RQ3 — Brand-Level Differences 8

7 Conclusion and Limitations **9**

7.1 Conclusion 9

7.2 Recommendations for our Client 9

7.3 Limitations 9

Group 206

Dongyuan Gao · Ramiro Diez-Liebana · Cyriel Van Helleputte

---

1 Project Overview

1.1 The Storyline (Potential Business Problem)

The Swiss used car market is highly competitive. Our **fictional client** AutoHelvetia AG, a leading national **used car dealer**, faces the challenge of optimizing their pricing & purchasing strategy. In recent years, commodity prices are volatile and affects pricing of cars. So AutoHelvetia AG delegated the task to us: to understand the relationship between used car prices and commodity prices.

1.2 Our Solution

This project delivers an **advanced data collection** and **anylsis** framework. Our goal is to collect valuable **market data** and uncover relationships between used car prices and key commodity markets. We develop a tool box using **web scraping of AutoScout24.ch** and integrating with **Yahoo Finance commodity data**, to provide AutoHelvetia AG data-driven insights for:

- **Optimize Pricing Strategies**
- **Gain Competitive Advantage**

1.3 Project Structure

```
project_scraping_CIP_analysis_car_commodity_price/
   Analysis/                # Analysis notebooks and scripts
       RQ1/                 # Research Question 1 script & analysis
       RQ2/                 # Research Question 2 script & analysis
       RQ3/                 # Research Question 3 script & analysis
   Data/                    # Data storage
       API_data_pull/       # API-fetched commodity data & script
       clean_data/          # Processed and cleaned datasets & script
       Scraping/            # Web scraped data and scripts & scraper script
   Documentation.md         # This documentation file
   README.md                # Project overview
   AI_Disclosure.md         # Gen AI usage disclosure and guidelines
   requirements.txt         # Project dependencies
   .gitignore
```

1.4 Initiall Findings

2 Feasibility Research

2.1 Ethical Feasibility of Web Scraping AutoScout24.ch

This web scraping project was evaluated for both technical and legal feasibility. We focused on the academic research context and our analysis of AutoScout24.ch's robots.txt file and terms of service indicates that the project operates within acceptable boundaries for academic research purposes. - **robots.txt Analysis**: - Allowed: General listing pages without filters - Restricted: User account pages (`/de/account/`, `/de/member/`) - Restricted: Filtered search results with specific URL parameters (e.g., `sort=`, `pricefrom=`) - Restricted: Administrative functions

2.1.1 Technical Feasibility

- **Data Extraction**: Ethically extracts vehicle specifications, pricing, and listing details with Scraper and Yahoo Finance API, involving selenium and beautifulsoup.
- **Data Availability**: We found consistent and abundant data, which is appropriate for analysis for both used car listings and Commodity Data.

2.1.2 Analytical Feasibility

- **Statistical Methods**: Appropriate statistical methods can be appllied for analysis, including correlation analysis, regression analysis, and time series analysis, etc.
- **Potential Conclusions**: The project can provide potential valuable insights into the relationship between used car prices and commodity prices, helping stakeholders make informed decisions.

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 Transformation and Cleaning

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).
- **Model Normalization**: Applies a replacement dictionary to outlier variants (e.g., "TESLA Model Y" → "Model Y"). Keeping consistent models during analysis.
- **Field Selection & Export**: Outputs a curated schema (`brand`, `model`, `car_model`, pricing, powertrain, and URL fields) to `Autoscout_Cleaned_Standardized.csv`, preserving only analytics-ready columns.

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.
- **Temporal Date Standardization**: Generates a formatted `Date` string (`%d-%m-%Y`) for later joins and saves the cleaned series as `Data/Final Data/yahoo_spot_cleaned.csv`.

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).
- Categorical fills (`power_mode`, `transmission`) based on model majority vote, defaulting to "Unknown".

- **Commodity Cleaning & Missing Value Imputation**:
  - Builds monthly periods, fills gaps from daily `Date` entries, and aggregates to one row per month.
  - Keeps `_Monthly_Avg` features and drops duplicates for consistent joins.
- **Merge Output**: Produces `Final_Merged_Data_RQ3.csv`, the **master clean dataset**.

5 Analysis & Visualization

5.1 Research Questions

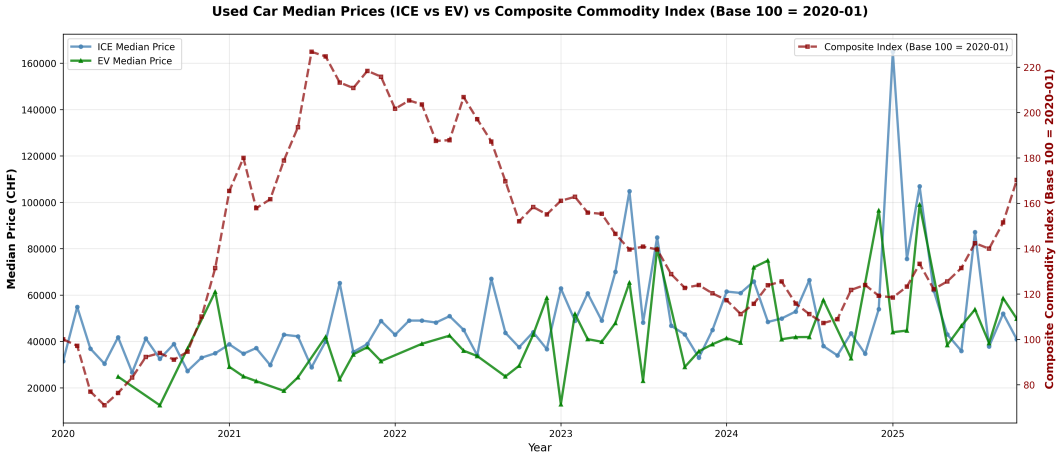
In our feasibility study we set out three potential questions for the client:

1. **RQ1 — Commodity Index vs. Used Car Prices**: How do used car prices in Switzerland correlate with historical commodity price indices for key automotive materials (steel, aluminum, copper, crude oil)? We can construct a weighted composite commodity index and examine Pearson/Spearman correlations with median used car prices while controlling for vehicle age and mileage.
2. **RQ2 — Powertrain Sensitivity**: Do different vehicle power modes (petrol, diesel, electric, hybrid) exhibit distinct sensitivity to commodity price movements? We expect electric vehicles to react more to battery metals, while ICE vehicles respond to energy and structural metals.
3. **RQ3 — Brand-Level Differences**: How does the relationship between commodity prices and used car values vary across popular Swiss car brands? We segment by brand to detect whether luxury vs. volume manufacturers show different exposure to raw material cost pressures.

6 Results and Findings

6.1 RQ1 — Commodity Index vs. Used Car Prices

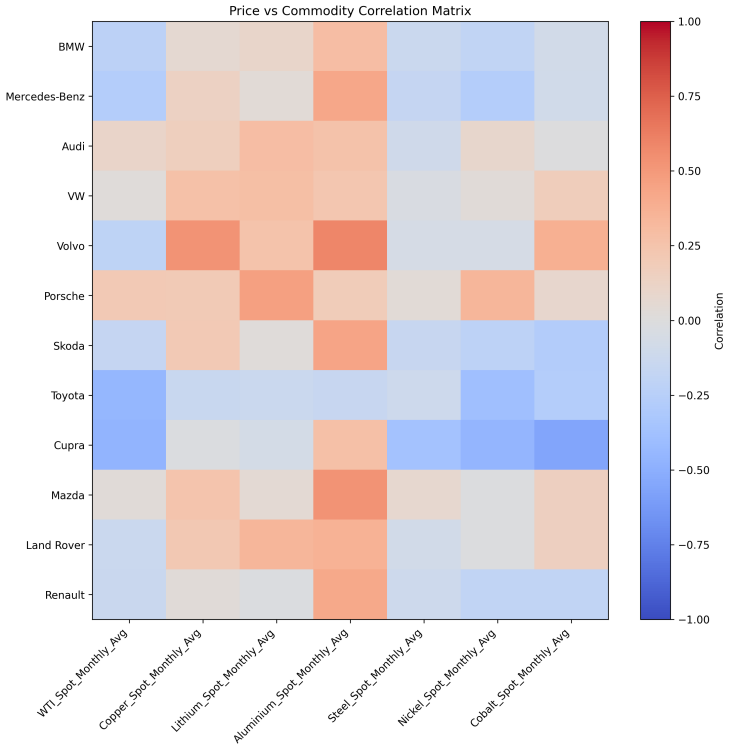
The index is rebased at **January 2020**, shortly before most commodities tanked 40 or 50% and resources so important like oil were traded at negative values (suppliers had to incentivize buyers with negative prices to clear out their reserves). Eventually things would rebound, making the index climb sharply, while both EV and combustion-engined median prices remained relatively flat during the length of the analysis.



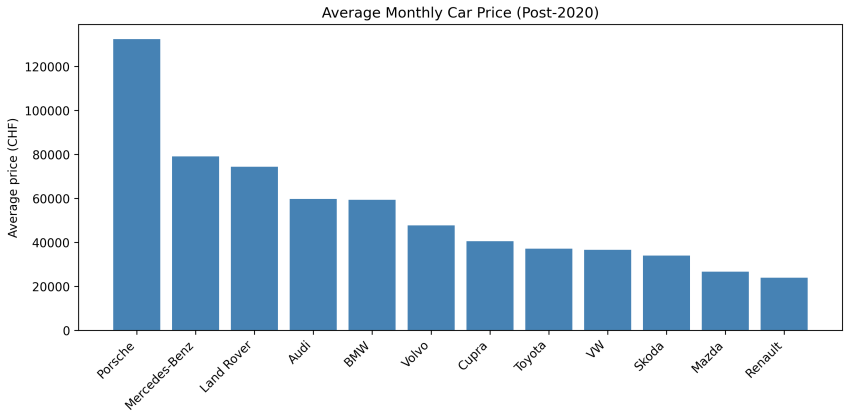
6.2 RQ2 — Powertrain Sensitivity

6.3 RQ3 — Brand-Level Differences

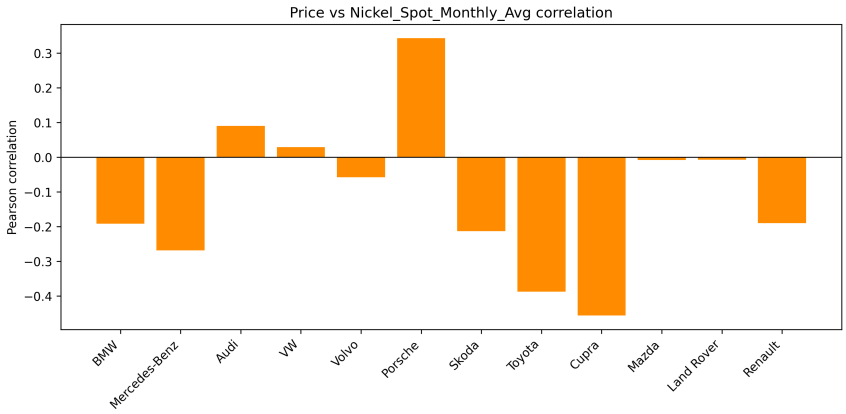
1. **Volvo & Porsche Track Battery Metals** — Volvo prices co-move with Copper (~0.52) and Cobalt (~0.38), while Porsche aligns with Nickel (~0.34), underscoring the exposure of electrified and luxury lineups to battery-heavy alloys.



2. **European Mass-Market Brands Show Mild Sensitivity** — VW, Audi, and Skoda maintain small positive links to copper and steel, suggesting moderate effect of industrial metal costs.



3. **Asian Brands Move Differently** — Toyota and potentially other Asian brands exhibit negative correlations with multiple commodities (WTI -0.45, Nickel -0.39), hinting at alternative supply strategies or value positioning that buffer raw material shocks.



Overall, brand equity and product mix modulate how commodity swings influence Swiss used car prices.

7 Conclusion and Limitations

7.1 Conclusion

This project aimed to analyze the relationship between used car prices and commodity markets in Switzerland. Through web scraping AutoScout24.ch and integrating Yahoo Finance commodity data, we uncovered interesting relationships into how raw material costs influence the Swiss used car market.

The results suggest **moderate correlation between commodity prices and the median price of used vehicles**, but the magnitude and direction vary substantially by brand and powertrain type. The composite commodity index peaked in 2021–2022 before normalizing, while used car prices remained relatively stable—suggesting that consumer behavior, policy incentives, and supply chain factors play equally important roles alongside raw material costs.

7.2 Recommendations for our Client

For our fictional client, we have the folliwng findings:

- **Timing Inventory Purchases**: Monitor battery metal prices (especially copper and cobalt) when acquiring electrified premium brands.
- **Pricing Models**: Incorporate commodity indices as leading indicators, particularly for luxury and EV segments.

7.3 Limitations

- **Temporal Scope**: Analysis covers 2020–2025, a period marked by extraordinary events (pandemic, supply chain disruptions) that may not represent typical market conditions.
- **Causality**: Correlations do not establish causal relationships; confounding factors like consumer preferences and policy changes were not fully controlled.
- **Geographic Scope**: Limited to the Swiss market, which may exhibit unique characteristics not generalizable to other regions.
- **Data Quality**: Web-scraped data subject to listing inconsistencies and potential sampling bias toward certain brands or price ranges. API Connection to market-ready sources such as CME's API, Shanghai Metals, or Bloomberg Terminal would be ideal.