FDZ Data Description: Data on the Used Car Market in Europe (RWI-GEO-CARMKT V1) - Used Car Listings on AutoScout24 01/2024 - 06/2024

Patrick Thiel

May 2025

Impressum

FDZ Data Description: Data on the Used Car Market in Europe (RWI-GEO-CARMKT V1) - Used Car Listings on AutoScout24 01/2024 - 06/2024

May 2025

Patrick Thiel

Contents

List of Figures

List of Tables

Abstract

The FDZ Ruhr provides a dataset on car advertisements in Europe, drawing from information obtained from the online marketplace AutoScout24. The dataset includes a diverse range of vehicle listings, covering various manufacturers, models, fuel types, and other key characteristics. The dataset offers detailed regional information and a comprehensive set of attributes, making it suitable for analyzing car market trends, pricing, and consumer preferences. It is important to note that the data are available for scientific research purposes only.

This data report provides an overview of the dataset, its structure, limitations, and key details. It is intended to support (potential) users in preparing and utilizing the data for research purposes.

This data report refers to RWI-GEO-CARMKT V1, which covers data from 01/2024 to 06/2024 and includes currently the country of Germany.

1.   Introduction and Short Description

1.1.   Introduction

The research data center (FDZ) Ruhr at the RWI provides a unique dataset on the European car market, obtained from AutoScout24. The dataset contains information on car listings, including advertised prices and various vehicle characteristics that influence a car's market value. The data is collected on a regular basis, with the present dataset (V1) covering 01/2024 to 06/2024.[1]

AutoScout24 is one of the largest online marketplaces for vehicles in Europe, enabling both private individuals and professional dealerships to advertise their cars. The platform covers a wide range of vehicles, from compact cars to luxury models, across multiple countries.[2] Listings on AutoScout24 include extensive details about each vehicle, including price, mileage, model specifications, and optional features. RWI-GEO-CARMKT only includes listings for used cars.

To list a vehicle, sellers complete a standardized form that captures key attributes such as brand, model, fuel type, and power output. The advertised price represents the seller's asking price and does not necessarily reflect final transaction prices. While price data is available for nearly all listings, additional details depend on seller input, leading to some variation in data completeness.

The dataset is distinguished by its broad coverage of the online car market, setting it apart from other automotive datasets. Several key features highlight its value:

Firstly, the dataset is extensive, comprising approximately 2,700,000 observations in the latest version (V1). Regular updates ensure a continuously growing database, facilitating in-depth analyses of market trends and consumer preferences.

Secondly, RWI-GEO-CARMKT provides detailed listing information. With 38 variables available, researchers can analyze key factors such as pricing, fuel types, transmission types, and regional market differences. This allows for flexibility in tailoring the data to specific research needs.

Thirdly, the dataset offers high-resolution geographical and temporal data. Each listing includes location information on the zip-code level, enabling spatial analyses of car market trends. In addition, each observation includes the listing date, allowing for daily analysis.

---

[1]It should be noted that the temporal coverage indicates active listings during the specified period. This does not necessarily imply that the listed vehicles were registered or that the advertisements themselves were created during that particular interval.

[2]The current version of RWI-GEO-CARMKT (V1) only includes listings from Germany, but other countries like France and Italy will be added in future versions.

1.2. Short Description

Table 1
Data Overview

Data unit	Used car sale listings offered on AutoScout24
Included countries	Germany
Period covered	01/2024 until 06/2024
Time reference	Daily
Regional structure	Zip-code information of seller
Sample	Full population of all used car listings offered on the platform
Update frequency	Continuously
Data access	Available for scientific research only
Current version	01/2024–06/2024 V1.0

1.3. List of Variables

The dataset contains the following variables.

Table 2
List of Variables

Variable	Variable name	Variable label	Variable type
Identifiers	ofid	Identifier for the listing	character
	uniqueID_gen	Unique identifier (RWI)	numeric
Features related to exterior	body_color	Body color of the car	character
	num_doors	Number of doors of the car	numeric
	co2_emissions	CO_2 emissions of the car (in g/km)	numeric
	efficiency_class	Efficiency class of the car	character
	emission_class	Emission class of the car	character
	emission_sticker	Emission sticker of the car	character
	fuel_consumption_city	Fuel consumption within city (in l/km)	numeric
	fuel_consumption_mixed	Mixed fuel consumption (in l/km)	numeric
	fuel_consumption_rural	Fuel consumption outside city (in l/km)	numeric
	fuel_type	Fuel type of the car	character
Features related to interior	equipment_first	First listed equipment of the car	character
	interior_color	Interior color of the car	character
	seats	Number of seats of the car	numeric
	cylinders	Number of cylinders of the car	numeric
	displacement	Displacement of the car (in cc)	numeric
	gears	Number of gears of the car	numeric

	power	Power of the car (in kW)	numeric
	transmission	Transmission type of the car	character
Features related to price	asking_price	Indicator for price being asking price	categorical
	price	Price of the car	numeric
	vat_deductible	Indicator for VAT being deductible	categorical
Generated features	carmkt_delivery	Classification of data delivery period	character
	carmkt_version	Classification of data version	character
Regional information	city	City of the seller	character
	country_code	Label for country of the seller	character
	country_zip_code	Country zip-code combination of the seller	character
	zipcode	Zip-code of the seller	character
Temporal information	created_date	Creation date of the listing	character
	first_registration	First registration date of the car	character
Vehicle characteristics	brand	Brand of the car	character
	mileage	Mileage of the car (in km)	numeric
	model	Model of the car	character
	vehicle_type	Type of the vehicle	character
	weight	Weight of the car (in kg)	numeric
Other features	num_previous_owners	Number of previous owners of the car	numeric
	provider_type	Type of the provider	character

1.4. Data Access

The data are available to researchers for non-commercial use. The data can be obtained in both .csv and .parquet file formats. It should be noted that data access to both versions requires a signed data use agreement. Both versions are restricted to non-commercial research; only researchers from scientific institutions are eligible to apply for data access.

Data access is provided by the Research Data Centre Ruhr at the RWI – Leibniz-Institute for Economic Research (FDZ Ruhr). The data can be accessed at https://www.rwi-essen.de/en/research-advice/further/research-data-center-ruhr-fdz/data-access. The application form includes a brief description and title of the project, potential cooperation, information on the applying department, expected duration of data usage as well as further participants in the project.

Data users shall cite the datasets properly with the respective DOIs. The DOIs of the current version (V1) of the dateset is:

Scientific Use File (SUF)

10.7807/as24:carmkt:suf:v1

2.   Data Preparation

2.1.   Data Source

The FDZ Ruhr at the RWI acquires the raw data directly from AutoScout24 in the form of spell data, wherein a spell denotes an advertisement. The FDZ Ruhr at the RWI procures all active advertisements of used car listings between two deliveries.

Users submit information regarding their car offers. The platform is available to both private and commercial users (i.e., car dealers), who can utilize it for the purpose of advertising their offers. These users are guided through an online questionnaire to create their advertisements. The majority of the questionnaire's fields are mandatory, including but not limited to brand, model, price, and mileage. Optional information encompasses the number of previous owners and equipment. Furthermore, it is possible to a brief description as well as picture to the listing. It is important to note that we do not obtain the name and contact information of the user who has placed the advertisement, nor do we receive the description and pictures of the car.

Listing a vehicle for sale on AutoScout24 is a free service. However, AutoScout24 also offers premium features that increase the visibility of the advertisement for a limited time. It should be noted that the obtained raw data does not provide insight into whether an advertisement has been designated as a premium feature.

2.2.   Anonymization

We do not implement additional anonymization procedures. The rationale and origins of this decision are rooted in the inherent characteristics of the data structure. Firstly, the raw data set does not contain any personally identifiable information but rather encompasses information relevant to the listed vehicle. Secondly, the smallest regional unit is the zip code information of the seller. It is important to acknowledge that the provided zip code does not necessarily correspond to the seller's actual residence; rather, it signifies the location and point of sale of the vehicle in question. This discrepancy in geographical location decreases the probability that a specific individual could be matched with the sold car and, consequently, could be re-identified. Thirdly, the dataset consists of a considerable amount of data points, which would scale any efforts to re-identify person-related information.

2.3.   Missings

Missing values are consistently encoded in the following manner:

Table 3
Missing Values Classification

Value	Description
-5	Implausible value (set by FDZ Ruhr RWI)
-6	Not available for specific delivery
-7	Not specified
-9	Other missing

We employ a data-driven approach to censor implausible values for numerical variables, which

we encode as -5, as they appear to be the result of typing errors (see Table 4 for details). However, it is plausible that there are still typing errors and implausible values remaining. Unlike in surveys, there is no interviewer to assist in the measurement process, which suggests that the prevalence of mismeasured values may be higher than in standard surveys. However, it is not possible to identify these values with certainty.

Table 4
Censoring Thresholds for Implausible Values

Variable	Threshold	Threshold type	Value
price	0.001	percentile	498.0
price	0.999	percentile	378,990.7
mileage	0.001	percentile	2.0
mileage	0.999	percentile	444,097.1
power	0.001	percentile	19.0
power	0.999	percentile	544.0
co2_emissions	0.001	percentile	6.0
co2_emissions	0.999	percentile	389.0
cylinders	0.001	percentile	2.0
cylinders	0.999	percentile	12.0
displacement	0.001	percentile	498.0
displacement	0.999	percentile	6,400.0
equipment_first	0.001	percentile	5.0
equipment_first	0.999	percentile	217.0
fuel_consumption_city	0.001	percentile	0.1
fuel_consumption_city	0.999	percentile	24.3
fuel_consumption_country	0.001	percentile	0.1
fuel_consumption_country	0.999	percentile	13.3
fuel_consumption_mixed	0.001	percentile	1.0
fuel_consumption_mixed	0.999	percentile	16.7
gears	0.001	percentile	1.0
gears	0.999	percentile	10.0
num_of_doors	0.001	percentile	2.0
num_of_doors	0.999	percentile	5.0
num_previous_owners	0.001	percentile	1.0
num_previous_owners	0.999	percentile	5.0
seats	0.001	percentile	2.0
seats	0.999	percentile	9.0
weight	0.001	percentile	720.0
weight	0.999	percentile	3,021.0

Notes: Values outside these thresholds have been recoded as implausible (-5).

Missing data encoded as -6 indicates variables that are not part of the delivery. -7 represents entries where the user chose not to provide additional information regarding the listing. -9 summarizes all other types of missings.

2.4. Data Problems

This section discusses issues the user should be aware of when working with the data.

Lack of observations

Since most information is not mandatory, some variables are only filled for few observations.

Partial cleaning and other notes

City. In addition to the zip code indicating the vehicle's location, RWI-GEO-CARMKT also provides the city of the seller as the next largest regional unit. The city variable has undergone a process of refinement, resulting in the elimination of implausible values that are readily apparent, such as those comprising solely numerals or the characters like "–" and "...". However, it is possible that less obvious implausible values remain part of the dataset. The responsibility for determining the appropriate course of action in such instances rests with the researcher.

3. Variable Description

Each variable is described below. The variables are arranged by category.

3.1. Identifier

Table 5
Detailed Variable Descriptions for Identifiers

Feature	Description
Label	Identifier for the listing
Name	ofid
Data type	Character
Description	Each listing is distinguished by a unique artificial ID number. These IDs are specific to each listing and remain constant over time. However, it is possible to pause an offer so that the offer ID (ofid) reappears in subsequent deliveries.
Label	Unique identifier (RWI)
Name	uniqueID_gen
Data type	Numeric
Description	*uniqueID_gen* is a unique identifier for each advertisement. It has been artificially generated and is defined as a running ID.

3.2. Features Related to Exterior

Table 6
Detailed Variable Descriptions for Features Related to Exterior

Feature	Description
Label	Body color of the car
Name	body_color
Data type	Character
Description	*body_color* is a character variable, indicating the main exterior color of the car.
Label	Number of doors of the car
Name	num_doors
Data type	Numeric
Description	*num_doors* is a numeric variable which refers to the number of doors of the car.
Label	CO_2 emissions of the car (in g/km)
Name	co2_emissions
Data type	Numeric
Description	*co2_emissions* is a numeric variable, denoting the emitted CO_2 emissions of the car in g/km.
Label	Efficiency class of the car
Name	efficiency_class
Data type	Character

Description	*efficiency_class* is a character variable that categorizes the car's fuel efficiency based on regulatory standards.
Label	Emission class of the car
Name	emission_class
Data type	Character
Description	*emission_class* is a character variable that represents the car's emission standard, indicating compliance with environmental regulations (e.g., Euro 6).
Label	Emission sticker of the car
Name	emission_sticker
Data type	Character
Description	*emission_sticker* is a character variable that specifies the environmental sticker classification assigned to the car, typically used in low-emission zones.
Label	Fuel consumption within city (l/km)
Name	fuel_consumption_city
Data type	Numeric
Description	*fuel_consumption_city* is a numeric variable that represents the car's fuel consumption in urban driving conditions, measured in liters per kilometer (l/km).
Label	Mixed fuel consumption (l/km)
Name	fuel_consumption_mixed
Data type	Numeric
Description	*fuel_consumption_mixed* is a numeric variable that represents the car's average fuel consumption under a combination of city and highway driving conditions, measured in liters per kilometer (l/km).
Label	Fuel consumption outside city (l/km)
Name	fuel_consumption_rural
Data type	Numeric
Description	*fuel_consumption_rural* is a numeric variable that represents the car's fuel consumption in outside-city driving conditions, measured in liters per kilometer (l/km).
Label	Fuel type of the car
Name	fuel_type
Data type	Character
Description	*fuel_type* is a character variable that specifies the type of fuel the car uses, such as petrol, diesel, electric, or hybrid.

3.3.   Features Related to Interior

Table 7
Detailed Variable Descriptions for Features related to Interior

Feature	Description
Label	First listed equipment of the car
Name	equipment_first
Data type	Character
Description	*equipment_first* is a character variable listing the first mentioned equipment of the car.
Label	Interior color of the car
Name	interior_color
Data type	Character
Description	*interior_color* is a character variable that specifies the color of the vehicle's interior.
Label	Number of seats of the car
Name	seats
Data type	Numeric
Description	*seats* is a numeric variable representing the number of seats in the vehicle.
Label	Number of cylinders of the car
Name	cylinders
Data type	Numeric
Description	*cylinders* is a numeric variable representing the number of cylinders in the vehicle's engine.
Label	Displacement of the car (in cc)
Name	displacement
Data type	Numeric
Description	*displacement* is a numeric variable that indicates the total volume of all the cylinders in the engine, measured in cubic centimeters (cc).
Label	Number of gears of the car
Name	gears
Data type	Numeric
Description	*gears* is a numeric variable representing the number of gears in the vehicle's transmission system.
Label	Power of the car (in kW)
Name	power
Data type	Numeric
Description	*power* is a numeric variable indicating the engine's power output, measured in kilowatt (kW).
Label	Transmission type of the car
Name	transmission
Data type	Character

Description	*transmission* is a character variable that specifies the type of transmission in the vehicle, such as manual, automatic, or semi-automatic.

3.4. Features Related to Price

Table 8
Detailed Variable Descriptions for Features Related to Price

Feature	Description
Label	Indicator for price being asking price
Name	asking_price
Data type	Categorical
Description	*asking_price* is a categorical variable indicating whether the listed price is an asking price.
Label	Price of the car
Name	price
Data type	Numeric
Description	*price* is a numeric variable representing the listed price of the vehicle in the advertisement.
Label	Indicator for VAT being deductible
Name	vat_deductible
Data type	Categorical
Description	*vat_deductible* is a categorical variable indicating whether the price of the vehicle is VAT-deductible.

3.5. Generated Features

Table 9
Detailed Variable Descriptions for Generated Features

Feature	Description
Label	Classification of data delivery period
Name	carmkt_delivery
Data type	Character
Description	*carmkt_delivery* is a character variable indicating the delivery the observations belongs to. This variable has more internal than external use.
Label	Classification of data version
Name	carmkt_version
Data type	Character
Description	*carmkt_version* is a character variable representing the specific version of the dataset. This variable has more internal than external use.

3.6. Regional Information

Table 10
Detailed Variable Descriptions for Regional Information

Feature	Description
Label	City of the seller
Name	city
Data type	Character
Description	*city* is a character variable representing the name of the city where the vehicle is listed.
Label	Label for country of the seller
Name	country_code
Data type	Character
Description	*country_code* is a character variable containing the country code of the listing.
Label	Country zip-code combination of the seller
Name	country_zip_code
Data type	Character
Description	*country_zip_code* is a character variable that provides the country-specific postal code of the vehicle listing.
Label	Zip-code of the seller
Name	zipcode
Data type	Character
Description	*zipcode* is a character variable representing the postal code of the listing location.

3.7. Temporal Information

Table 11
Detailed Variable Descriptions for Temporal Information

Feature	Description
Label	Creation date of the listing
Name	created_date
Data type	Character
Description	*created_date* is a character variable representing the date when the vehicle listing was created on the platform.
Label	First registration date of the car
Name	first_registration
Data type	Character
Description	*first_registration* is a character variable indicating the date of the vehicle's first registration.

3.8.   Vehicle Characteristics

Table 12
Detailed Variable Descriptions for Vehicle Characteristics

Feature	Description
Label	Brand of the car
Name	brand
Data type	Character
Description	*brand* is a character variable representing the manufacturer or brand of the vehicle.
Label	Mileage of the car (in km)
Name	mileage
Data type	Numeric
Description	*mileage* is a numeric variable indicating the total distance the vehicle has traveled, measured in kilometers (km).
Label	Model of the car
Name	model
Data type	Character
Description	*model* is a character variable specifying the model name of the vehicle.
Label	Type of the vehicle
Name	vehicle_type
Data type	Character
Description	*vehicle_type* is a character variable categorizing the type of vehicle, such as sedan, SUV, or convertible.
Label	Weight of the car (in kg)
Name	weight
Data type	Numeric
Description	*weight* is a numeric variable indicating the vehicle's weight in kilograms (kg).

3.9. Other Features

Table 13
Detailed Variable Descriptions for Other Features

Feature	Description
Label	Number of previous owners of the car
Name	num_previous_owners
Data type	Numeric
Description	*num_previous_owners* is a numeric variable representing the number of previous owners the vehicle has had.
Label	Type of the provider
Name	provider_type
Data type	Character
Description	*provider_type* is a character variable indicating the type of provider, such as dealership or private seller

4.   Further Material

Users might find the coding for the data preparation useful. Please refer to our GitHub repository (https://github.com/PThie/RWI-GEO-AS24), which should be cited as: Patrick Thiel (2025). PThie/RWI-GEO-AS24: RWI-GEO-CARMKT V1 (V1.0). Zenodo. https://doi.org/10.5281/zenodo.15423311.

The variable and value labels are also delivered to the user as separate files, in conjunction with the dataset files.