Task 1: Vehicle Segmentation

Objective: Cluster vehicles into distinct groups based on their specifications and identify meaningful patterns.

Task steps:

1. **Data Preparation:** Select relevant features for clustering. Consider techniques for handling categorical variables and missing data. Justify your feature selection and data preprocessing choices.

Data preprocessing approach:

I. Handling Missing Values: Some columns have a very high percentage of missing values (e.g., over 50%), and imputing these values could potentially introduce bias or inaccuracies. The following features will be removed: 'First_registration_date', 'CO_2 emission', and 'First Owner'. Group variables with less than 1% missing values are imputed as follows:

- Fill missing values in Mileage_km using the condition-specific medians.
- Fill missing values in Power and Displacement with the median of each column.
- Fill missing values in the Transmission and Door number using the mode.

II. The car prices are represented in both PLN and EUR. To ensure consistency, all prices are converted to EUR.

III. Detecting and removing noise from the data (e.g., values like '55' for the number of doors, which is clearly incorrect).

IV. Handling categorical variables: After examining the categorical features, it was found that some features have a large number of unique levels (e.g., Vehicle Brand with 108 levels). Hence, categories with a frequency of less than 5% were combined and defined as the

'Other' group. For categorical features with a high number of levels and low frequency (e.g., Vehicle Model with 1203 levels and a maximum frequency of 0.8% for any level), were dropped. Subsequently, one-hot encoding was applied to the remaining categorical features.

V. There is a column representing car features; however, around 40% of the ads do not include this information. Hence, a new column was created to indicate whether an ad includes car features or not.

Feature selection approach:

A heatmap was used to visualize the correlation between the features, shown in Figure 1. It is clear that fuel types like gasoline and diesel are mutually exclusive for a car, a single vehicle cannot run on both simultaneously. Since one variable can predict the other, keeping both is redundant. Therefore, the 'diesel' feature was removed for simplicity.

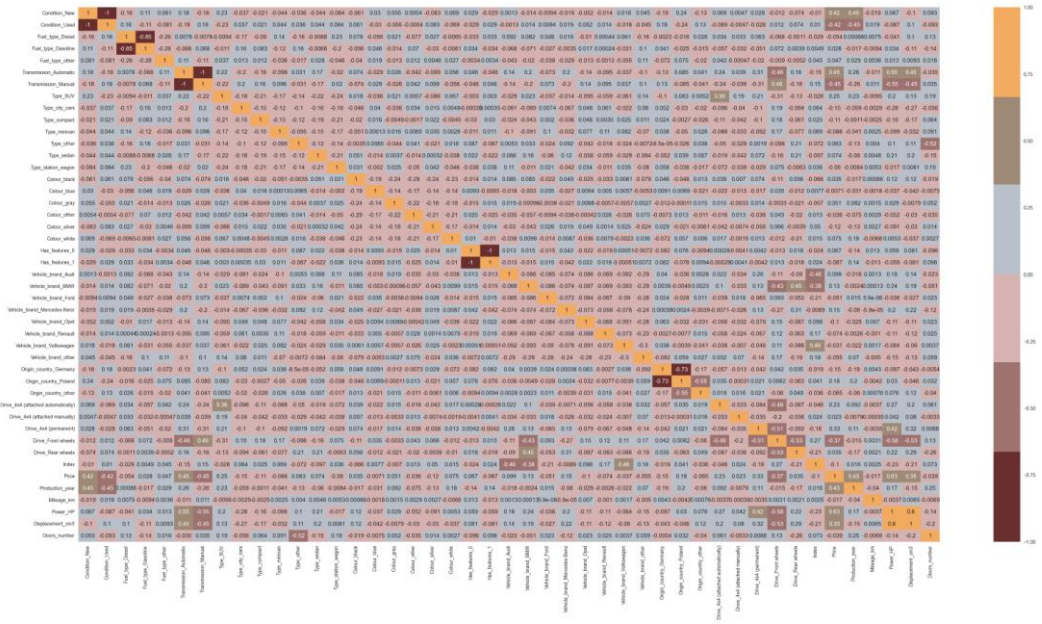

Figure 1. Correlation analysis for feature selection.

Columns for clustering: Based on the reasons mentioned in the privous steps, the following feature were considered: Price, Condition, Vehicle_brand, Production_year, Mileage_km, Power_HP, Displacement_cm^3, Fuel_type, Drive, Transmission, Type, Doors_number, Colour, Origin_country, and Has_features.

2. **Dimensionality Reduction (Optional)**: Apply dimensionality reduction techniques to visualize the data and aid in cluster interpretation. If you use dimensionality reduction, discuss its impact on the clustering results.

PCA with 3 components is applied to address high dimensionality, as shown in Figure 2. If clustering results significantly change after dimensionality reduction, it suggests that the high-dimensional space contains valuable features that were lost during the reduction.

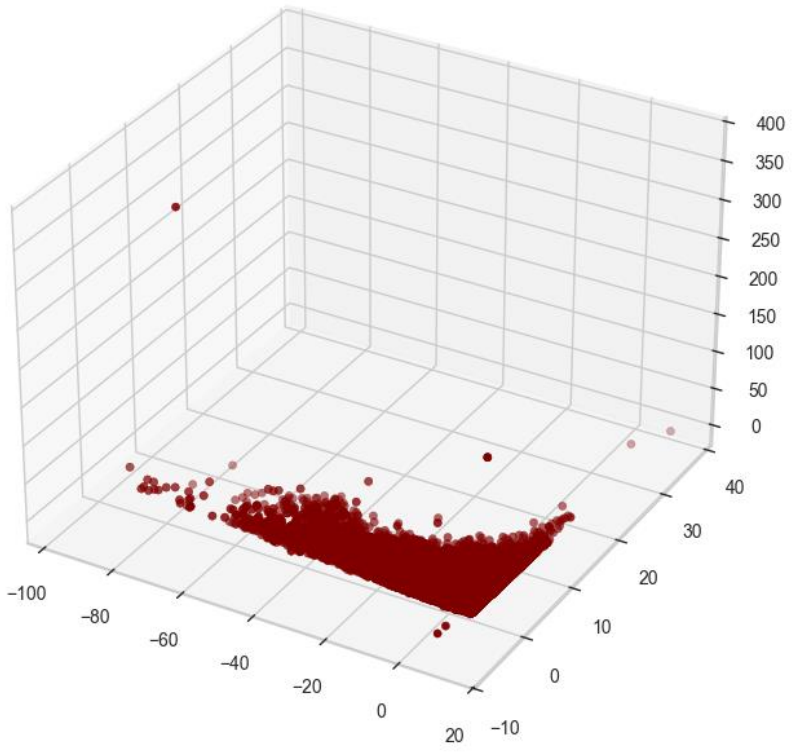

Figure 2. 3D projection of data in the reduction dimension.

3. **Clustering**: Implement a suitable clustering algorithm. Explore different algorithms and parameter settings to find the best solution. Clearly document your approach, including algorithm selection, parameter tuning, and results.

Algorithm selection approach:

As the dataset contains outliers and is not in spherical form, the preferred model is the DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise). Additionally, this algorithm does not force every data point into a cluster, further improving its robustness in the presence of outliers.

However, by applying this method, DBSCAN generated more than 20 clusters. After attempting to fine-tune its parameters for better results, the process became computationally expensive, eventually leading to a system error, as follows:

✗ *"The Kernel crashed while executing code in the current cell or a previous cell. Please review the code in the cell(s) to identify a possible cause of the failure. Click here for more info. View the Jupyter log for further details."*

After failing the DBSCAN algorithm, the approach shifted to the K-Means algorithm.

Parameter Tuning: To determine the optimal number of clusters, the Elbow Method was applied and the Silhouette Score across a range of K values was also calculated, as shown in Figure 3. The most optimal number of clusters was found to be K=4.

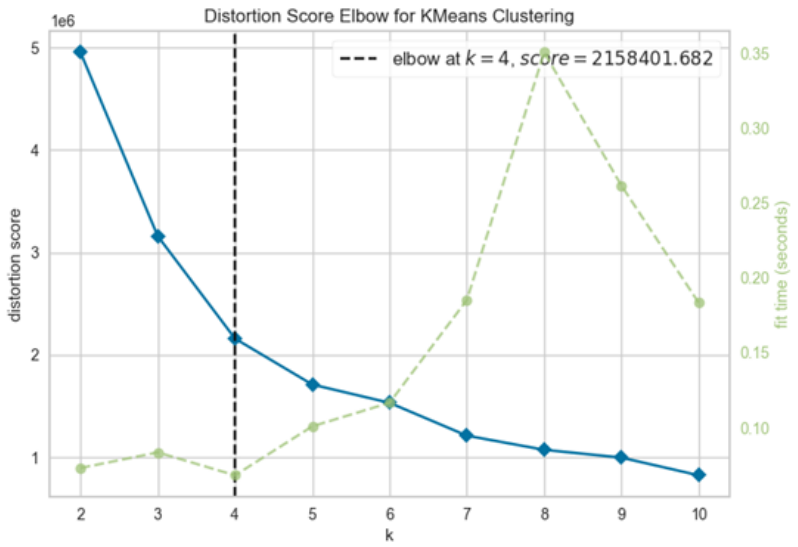

Figure 3. Distribution score Elbow for K-means clustering.

4. **Cluster Evaluation**: Evaluate the quality of your clustering results using appropriate metrics. Justify your choice of evaluation metrics.

To create groups that are internally homogeneous (similar within the group) and externally heterogeneous (distinct from other groups), the following metrics used are as follow:

Silhouette Score: 0.50

The Silhouette Score measures how similar points within a cluster are to points in other clusters. A score of 0.50 suggests that developed clustering is moderately well-defined, as values closer to 1 indicate better-defined clusters, while values close to 0 or negative may indicate poor clustering.

Davies-Bouldin Score: 0.61

The Davies-Bouldin Score evaluates the average similarity of each cluster to its most similar cluster. Lower values indicate better clustering, with values closer to 0 being ideal. A score of 0.61 is relatively low, suggesting that the clustering is relatively well-separated with minimal overlap.

Overall, these metrics suggest that the developed clustering algorithm is performing reasonably well, with well-defined clusters, though there is still some room for improvement. Algorithms like DBSCAN and Agglomerative Clustering may work better.

5. **Interpretation and Insights**: Analyze the characteristics of each cluster and provide a meaningful interpretation of the segments. Consider the business implications of your findings and propose potential applications (e.g., targeted marketing, product development).

The characteristics of each cluster were analyzed, and the important findings are summarized below:

Cluster 0: Most of the car types in this cluster are compact and station wagon. The cars in this cluster were produced in 1992-2008, and recognized as moderately older cars.

Cluster 1: This category has the lowest price average with 3,352 €. The majority of cars have rare brands with 42% like Talbot and Vanderhall. The advertised cars in this cluster were produced in 2009-2015, and recognized as moderately new cars. Most of the car types in this cluster are sedan and station wagon.

Cluster 2: This category comprises both used and new cars (condition) compared to other categories that only have used cars. This category has the highest price average with 26,822 €. The advertised cars in this cluster are produced in 2014-2021, and recognized as newer or recent cars. The majority of the cars are SUV type and over 50% of the cars have automatic transmission.

Cluster 3: Most of the car advertisements feature 2 or 4 doors. 30% of the cars are Mercedes-Benzes in this category with the highest frequency. The majority of cars (87%) are fueled by

gasoline. 65% of the advertisements do not include any description of the cars. This category includes used cars and mostly old ones from 1985-1989. Since the colors and types of cars in this category are not regular and common (e.g., violet convertible), they can be recognized as old luxury or rare cars.

Marketing strategies:

Cluster 1: Highlight affordability to attract first-time car buyers, students, or individuals with limited budgets.

Cluster 2: Emphasize the appeal of recent models (2014-2021) and automatic transmissions in marketing campaigns, targeting buyers seeking newer technology and convenience.

Cluster 3: Since the cars in this category are primarily old luxury or rare models (1985-1989), marketing efforts could target collectors, vintage car enthusiasts, and those seeking unique vehicles.

Task 2: Anomaly Detection

Objective: Identifying unusual or suspicious vehicle offers within the dataset.

Task Steps:

1. **Exploratory Data Analysis (EDA)**: Perform EDA to understand the characteristics of the data and identify potential areas for anomaly detection. Visualize data distributions and look for unusual patterns or outliers.

In the first step, the distribution of the numerical features of the dataset, including price, power, mileage and displacement, was investigated. Moreover, in order to check outliers in relation to categorical features, patterns of the numerical variables (price, power, and displacement) were evaluated across categories.

- **Price (€):**

Table 1 shows the statistical distribution of price variables. The histogram and box plot for the price distribution were plotted in Figure 4.

Table 1. Statistical distribution of price variable

Count, number of total data	208,304.00
Mean	13,871.93
Std	19,479.03
Min	128.17
25%	2,899.98
50%	7,843.78
75%	16651.60
Max	1,533,480.90

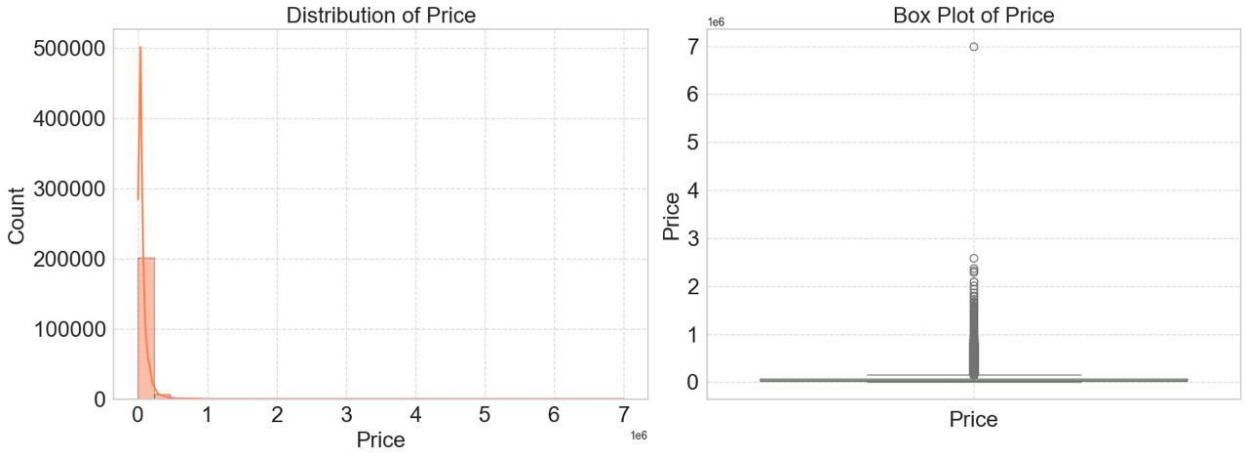

Figure 4. Histogram, left plot, and box plot, right plot, of the price distribution.

According to Figure 4, prices are skewed to the right, indicating significant positive skewness. This skewness reflects the presence of a few high-priced vehicles. Most vehicles fall into the low-to-moderate price range, but the inclusion of listings with extraordinarily high prices stretches the distribution, resulting in a long right tail.

The extreme prices, as highlighted in the box plot, suggest potential anomalies or unusual vehicle listings, which may be attributed to factors such as:

- Luxury or rare vehicles: High-end models or limited editions commanding premium prices.

- Pricing errors or ad fraud: Incorrect listings or intentional misrepresentation.

These observations warrant further investigation to determine whether the extreme values are legitimate or indicative of data issues.

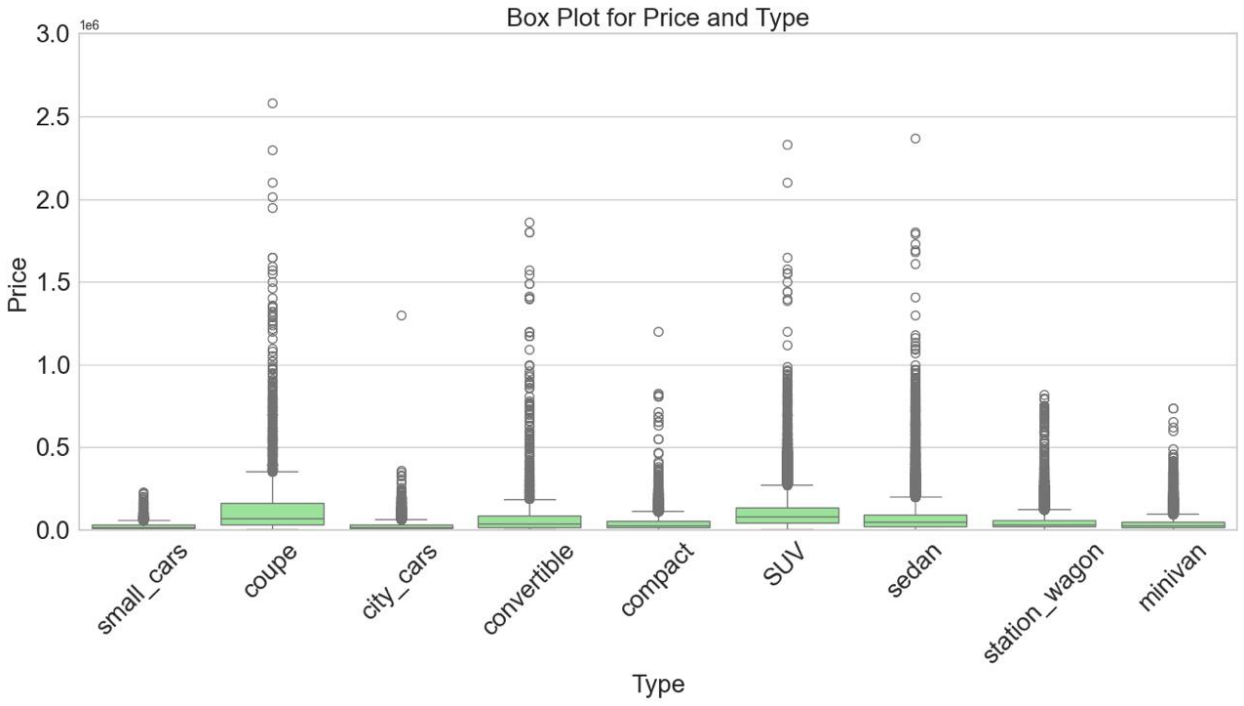

Figure 5. Box plot for price and type of the car.

According to Figure 5, a particularly notable outlier exists in the SUVs, sedans, coupe, and convertibles categories, representing an extreme deviation from the norm. These findings suggest that specific vehicle types, especially coupes and SUVs, may include high-end models, luxury editions, or rare vehicles that significantly influence their price distributions.

- **Power (HP)**

Table 2 shows the statistical distribution of the power variable. The histogram and box plot for power distribution were plotted in Figure 6. According to Figure 6, the majority of vehicles have power values ranging from 100 to 200 HP, which is typical for most passenger cars. A few

extremely high values (above 700 HP) are rare but noticeable, indicating potential outliers. Additionally, vehicles with power values exceeding 400 HP are considered outliers, likely representing high-powered sports cars, trucks, or specialty vehicles.

Table 2. Statistical distribution of power variable.

Count, number of total data	207,661.00
Mean	151.84
Std	77.68
Min	1.00
25%	105.00
50%	136.00
75%	172.00
Max	1,398.00

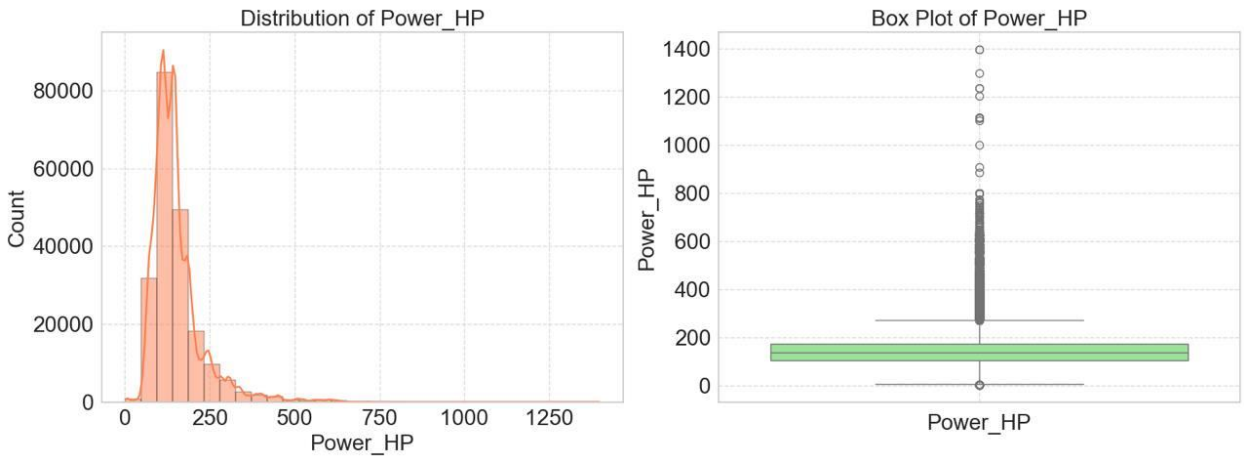

Figure 6. Histogram and box plot for power distribution.

The histogram plot shows the distribution of vehicle type versus power (HP), shown in Figure 7. Categories such as coupes, convertibles, and SUVs exhibit several extreme values beyond 1,200 HP, suggesting the presence of powerful, and niche vehicles in these groups.

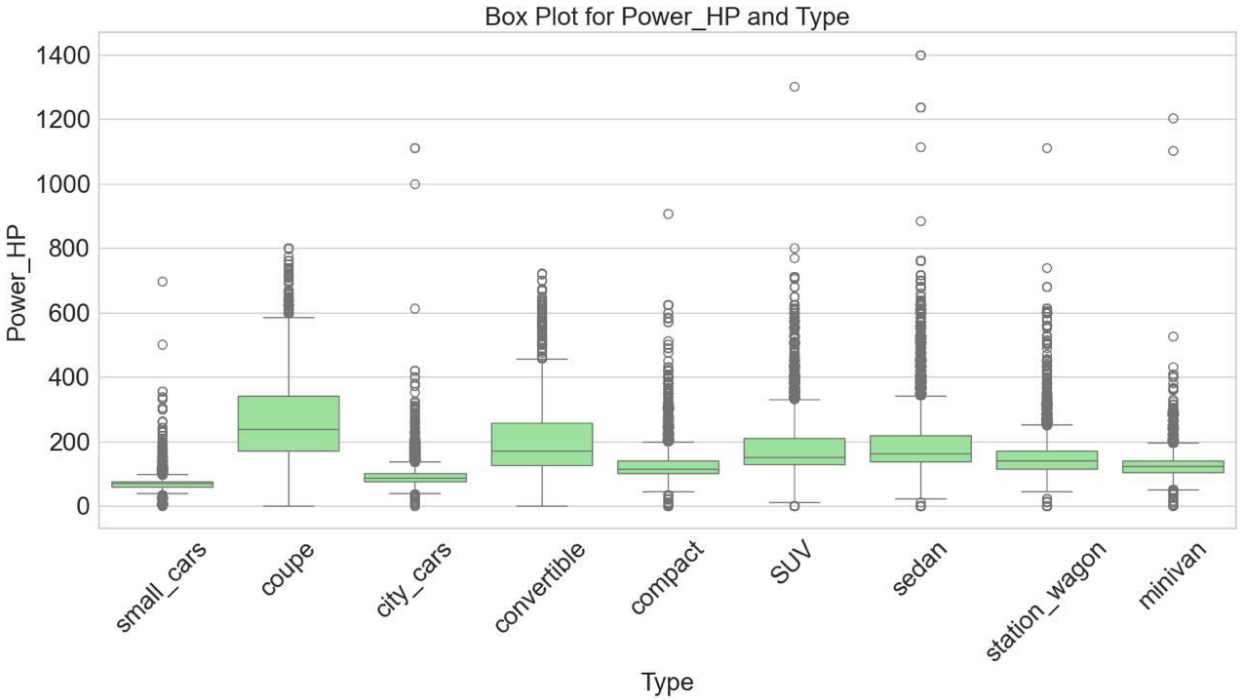

Figure 7. Histogram plot for the distribution of vehicle type versus power (HP).

- **Displacement (cm^3)**

Table 3 shows the statistical distribution of displacement variables. The histogram and box plot for displacement distribution were plotted in Figure 8. Based on Table 4 and Figure 8, 75% of vehicles have engine displacements below 2,000 cm^3, with the majority (50-75%) falling within the range of 1,798 to 1,997 cm^3. This is typical for standard regular cars.

Table 3. Statistical distribution of displacement variable.

Count, number of total data	206,338.00
Mean	1,882.57
Std	729.61
Min	400.00
25%	1,461.00
50%	1,798.00
75%	1,997.00
Max	8,400.00

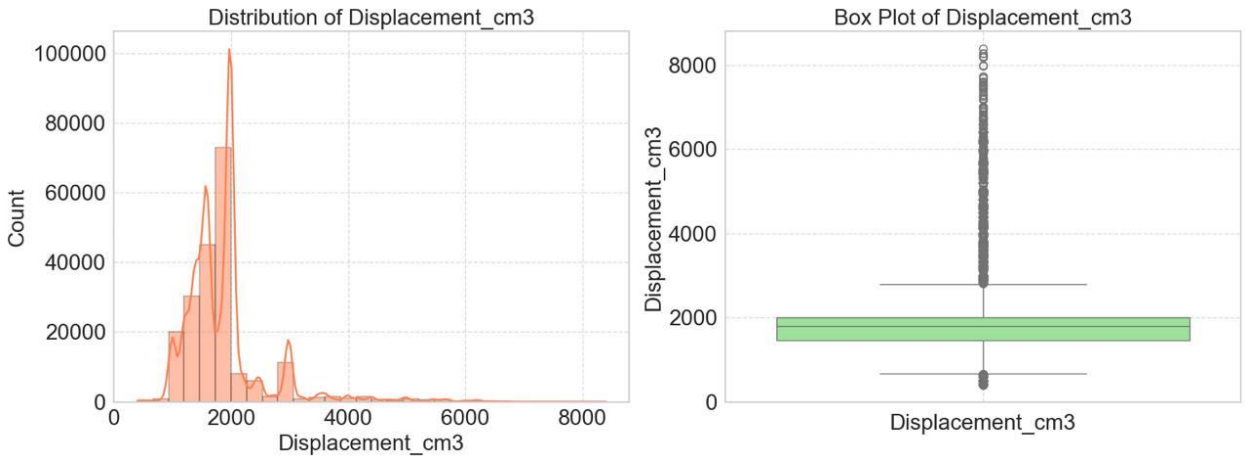

Figure 8. Histogram and box plot for displacement distribution.

According to Figure 9, there are numerous outliers for most car types, especially for SUVs, sedans, where engine displacement goes above 7000 cm^3. These are likely corresponding to specialty high-performance vehicles or outliers.

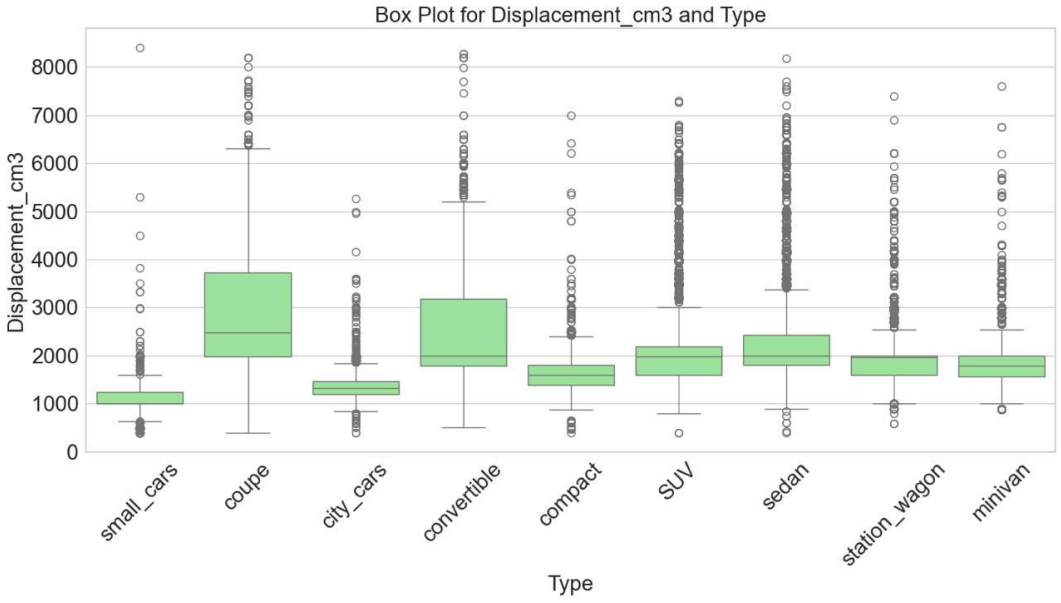

Figure 9. Box plot for displacement and type of the car.

2. **Anomaly Detection Method Selection**: Choose an appropriate anomaly detection method or methods. Consider the nature of the data and the types of anomalies you are trying to detect. Justify your choice of methods.

To select an appropriate anomaly detection method, the nature of the data and the types of anomalies were considered, as below:

1. Data characteristics:

Numerical and categorical*:* Primary variables like price, mileage, power, and displacement are considered numerical, while categorical features include vehicle brand, fuel type, body type, etc. that can help segment anomalies for deeper insights.

Large dataset and high-dimensional*:* When combining multiple features, the data becomes high-dimensional. With over 200,000 records, computational efficiency is critical.

Skewed distributions*:* Features like price and power are highly skewed with extreme values.

2. Types of anomalies to detect:

Global outliers*:* Data points that are extreme across all features (e.g., an unusually high-priced car with standard power and mileage).

Contextual outliers*:* Anomalies within a specific group (e.g., a city car with extremely high power) are ideal for identifying rare vehicles, like SUVs with extreme power.

Cluster-based anomalies*:* Data points that do not fit into any of the natural clusters formed in the data.

3. **Anomaly Detection Implementation:** Implement your chosen method(s) and detect anomalies in the data. Document your approach, including parameter settings and results.

In the first step, rare or unusual ad listings were detected using an Isolation Forest, and then the identified anomalies based on their scores were filtered. After isolating the anomalies, clustering

algorithms were applied to group similar anomalies into natural clusters, while flagging any points that deviated significantly from these groups. The following is the detailed approach used:

Workflow approach:

Global anomaly detection*:* The Isolation Forest model was applied to identify overall anomalies in key numerical features (price, mileage, power).

Cluster-based detection*:* The DBSCAN model was applied to capture vehicles that do not belong to any cluster.

Parameter setting:

Silhouette Score across a range of K values was calculated. The best parameters include eps=0.2, min samples=15 with Silhouette Score=0.9, resulting in detecting 1961 anomalies.

4. **Anomaly Analysis and Interpretation:** Analyze the detected anomalies and interpret your findings. Consider potential explanations for the anomalies and discuss any limitations of your approach.

Based on the obtained results, 1,961 ads were identified as anomalies, with 80% of them representing new cars (2018-2021) featuring low mileage and high prices, as shown in Figure 10. Among these, BMW and Mercedes-Benz had the highest frequencies, accounting for 25% and 16% of the ads, respectively. Additionally, most of the new cars with anomalies were equipped with 4×4 drive modes (either permanent or automatic), predominantly found in SUVs.

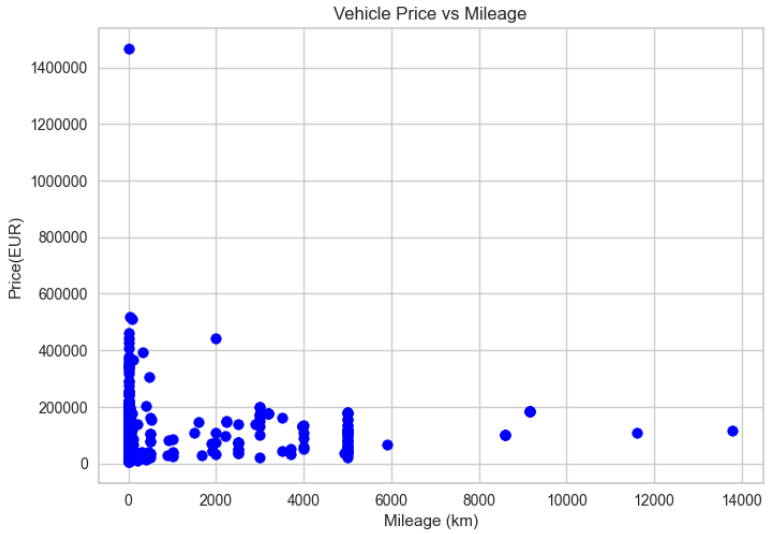

Figure 10. Distribution of vehicle price and mileage in new cars in the anomalies category.

Limitation of the approach used: There can be an overlap between true anomalies and noise, such as typos or erroneous values, which may result in the detection of outliers that are not meaningful or actionable. Incorporating domain knowledge is essential to focus on anomalies that are relevant within the dataset's context. Standard techniques like the Z-score often perform well for anomaly detection when the data is normally distributed. However, addressing skewness in features remains a challenge. For instance, when applying a cubic root transformation to reduce skewness, significant improvements were not achieved. Exploring alternative transformations, such as log or Box-Cox, or using other metrics tailored to non-normal distributions, could enhance anomaly detection accuracy.

Task 3: Time-Based Analysis

Objective: Explore temporal trends in the data.

1. Data Preparation: Identify relevant time-based features in the dataset. Prepare the data for time-series analysis. This may involve data transformation, handling missing values, or feature engineering.

To effectively prepare the data for time-series analysis, the key temporal features were investigated:

- Offer_publication_date: This timestamp can help analyze temporal trends.
- First_registration_date: Over half of its values are missing (122,083 out of 208,304 entries). Due to the high proportion of missing data, the First_registration_date column was dropped to avoid introducing inaccuracies through imputation.

Steps for Data Preparation:

I. ***Convert Dates to Datetime Format:*** Transform the Offer_publication_date column into a standardized datetime format to ensure compatibility for time-series analysis.

II. ***Handle Missing Values:*** Confirm that Offer_publication_date has no missing values. Address missing values in other features using appropriate strategies (e.g., dropping columns, imputation, or marking missing data).

III. ***Feature Engineering:*** Extract time-based features from Offer_publication_date to enhance analysis, such as Year, Month, Week, Day of the Week, and Day of the Month. These features will allow us to capture seasonal patterns, weekly trends, and other time-based insights.

2. **Exploratory Data Analysis (EDA):** Perform EDA to explore temporal trends in the data. Visualize the data to identify potential patterns or anomalies over time.

In this part, analyzing the number of car advertisements across different months and days was focused to identify trends and anomalies. The key steps are as follows:

I. ***Monthly and Daily Comparisons:*** The number of car advertisements was analyzed to observe variations across months and days.

II. ***Anomaly Detection Using IQR Method:*** The Interquartile Range (IQR) method was applied with a moderate threshold (k=1.5) to detect anomalies.

Figure 11 shows the number of car offers by month and monthly trend with IQR-based anomalies highlighted. It is clear that the activity for car offers peaks significantly during April and May, suggesting a seasonal trend, while the rest of the months have much lower levels of activity. Any anomalies have not been observed. Overall, the data follows a consistent seasonal pattern without any unusual deviations. This rapid growth in ads during spring aligns with seasonal behavior, warmer weather motivates people to buy or sell cars, while the rest of the months remain relatively quiet.

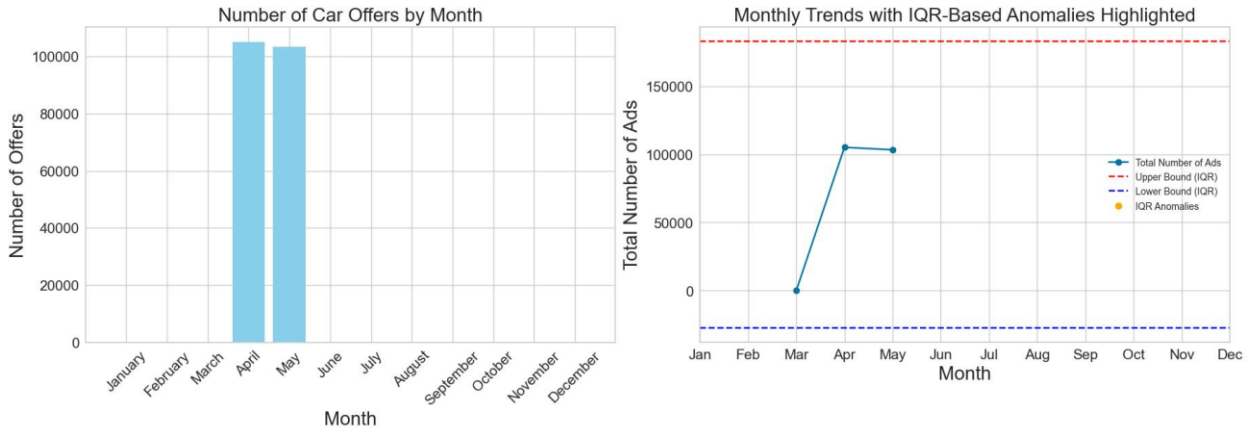

Figure 11. Number of cars offers by month, left side, and monthly trend with IQR-based anomalies highlighted, right side. The red dashed line at the top and the blue dashed line at the bottom represent the normal range for the number of ads.

According to Figure 12, most people post their car ads at the beginning of the week, especially on Monday and Tuesday, while fewer ads are posted on Wednesday.

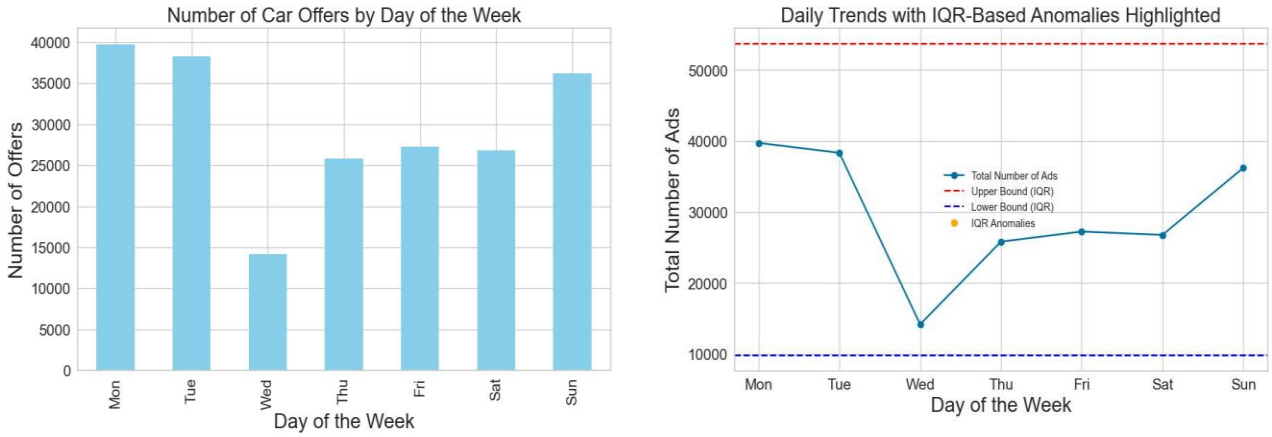

Figure 12. The number of cars offers by day of the week and daily trends with IQR-based anomalies highlighted.

3. **Time-Series Analysis**: Choose appropriate time-series analysis techniques to investigate the data. This could involve techniques such as decomposition, autocorrelation analysis, or modeling.

In order to investigate the data, decomposition characteristics were investigated as follows:

Seasonality: does the data have a clear cyclical/periodic pattern?

Trend: does the data represent a general upward or downward slope?

Noise: what are the outliers or missing values that are not consistent with the rest of the data?

4. **Interpretation and Insights:** Analyze your findings and interpret the results. Discuss any trends or patterns you have discovered and provide meaningful insights based on your analysis. Consider the business implications of your findings.

According to Figure 13, the number of car advertisements (blue and orange lines) remains flat at the beginning, starts increasing sharply in late April, and then drops abruptly in early May, possibly caused by an external or sudden event (e.g., promotional campaign, external shock, or unplanned demand).

The green line oscillates consistently over time, indicating a recurring weekly pattern. The number of ads at the beginning of the week, especially on Monday and Tuesday, is higher, while fewer ads are posted on Wednesday.

In the earlier period, residuals (red line) are relatively stable with ±1000 number/day due to randomness, but as the trend (orange line) increases sharply, the residuals become larger, and some days see unexpected spikes and dips in the number of ads. This suggests that unusual fluctuations occurred may be due to changes in user behaviour or website traffic.

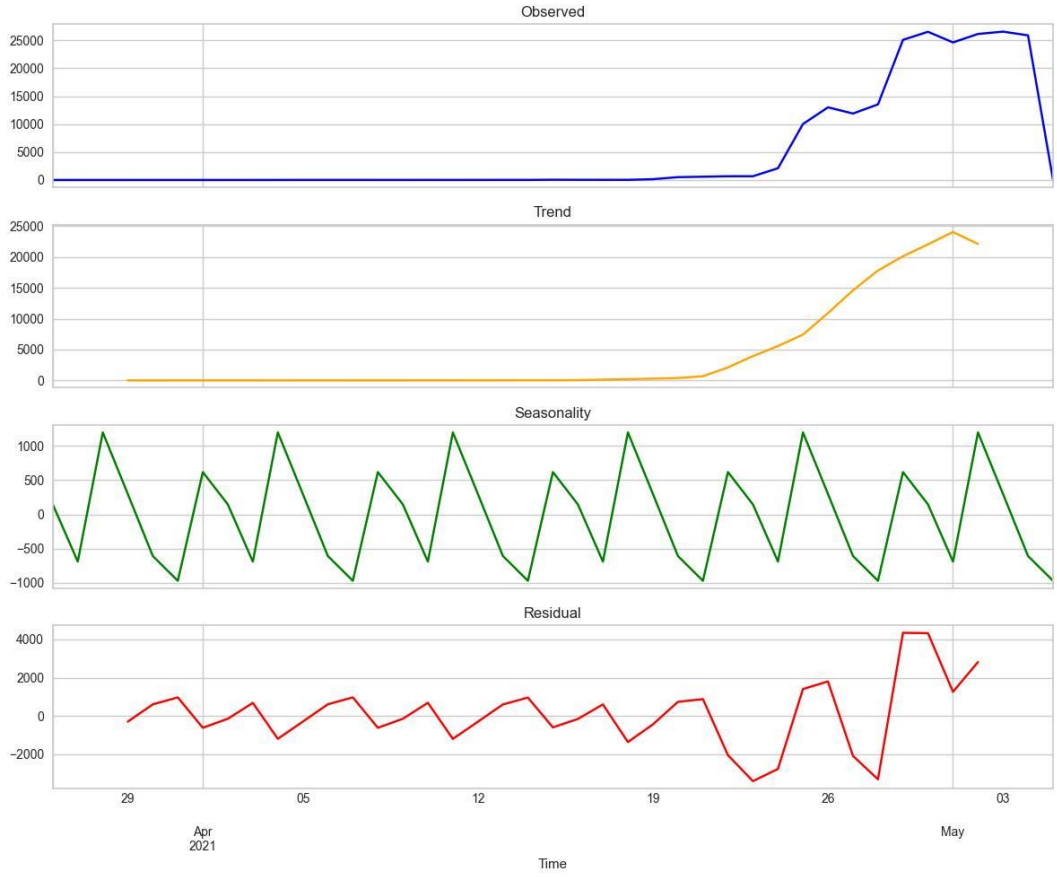

Figure 13. Time series the composition analysis.