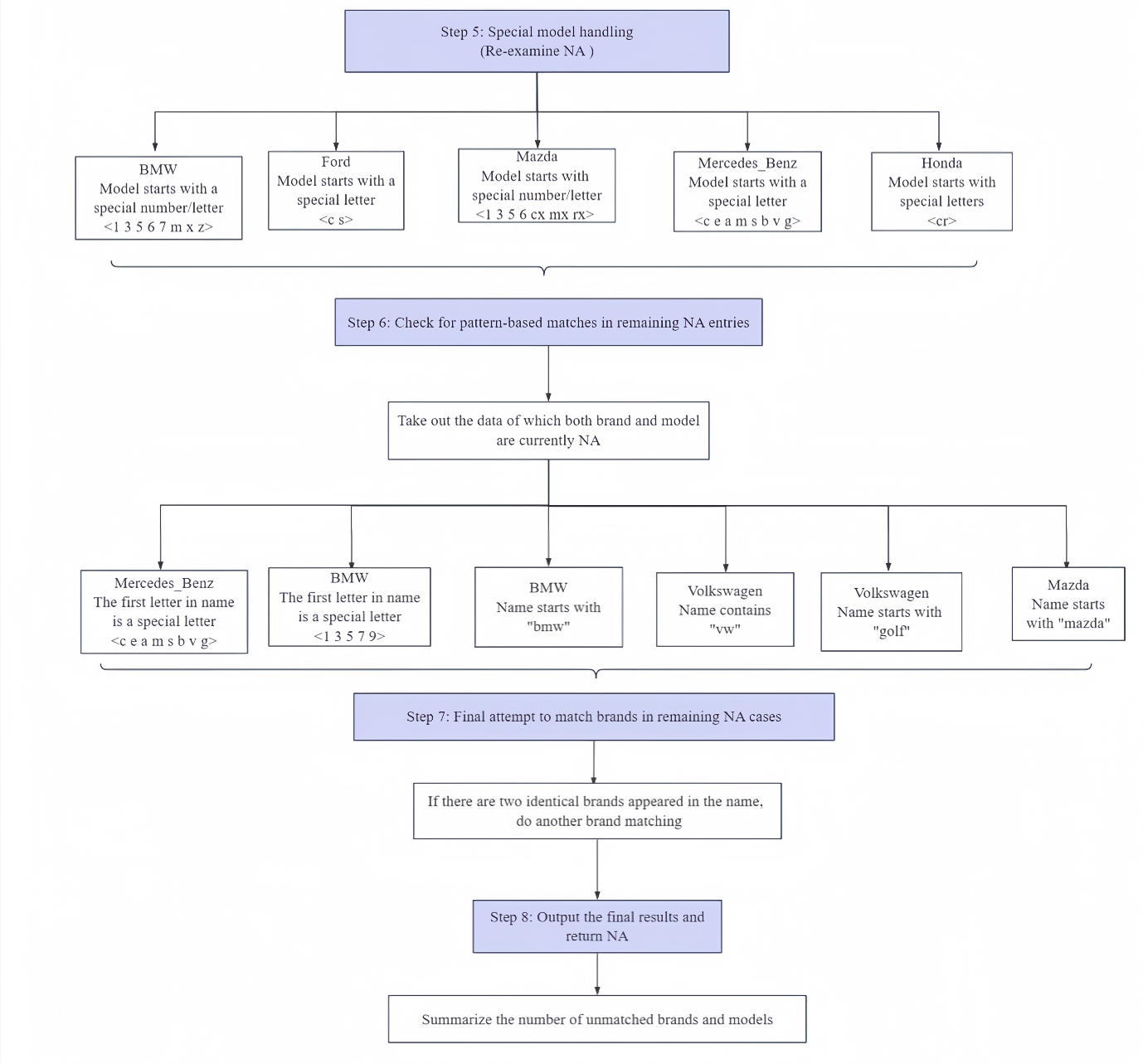


**Flowcharts**

(Continued on the following page)

(Continued)



# Car Data Matching Report

## Initial Data Analysis and Strategy Development

Our team began by analyzing two primary datasets: `autoswithout.csv` containing vehicle listings with unstructured name fields, and `model\_and\_brand.csv` containing standardized brand-model relationships. Early observations indicated that inconsistency in the "name" field, such as variations in brand or model names, irregular capitalization, and special characters, posed significant challenges to accurately matching vehicle information across the datasets. We also found the brand-model relationships highly inconsistent, ranging from names that included neither brand nor model, to those containing multiple brands or models. This complexity required a repeated matching process to identify and isolate unique brands accurately.

## Resolving Matching Challenges

To solve these problems, we developed a structured approach consisting of three main phases: Data Standardization, a Multi-Pass Matching Strategy, and Special Case Handling. This multi-phase workflow allowed us to address issues progressively, refining matches with each pass and ultimately increasing our accuracy and reducing ambiguities.

### Data Standardization

We standardized all entries to ensure consistency, starting by converting text to lowercase, splitting the name fields by underscores, and removing special characters. This process enabled easier component isolation, ensuring reliable matching criteria. Initial values for "brand" and "model" columns were set to "NA" to track unmatched entries across phases.

Besides, given that many entries had varied forms of brand names (e.g., "Mercedes" and "Benz" appearing separately), we standardized these variations. We expanded abbreviations and corrected common variations to reduce inconsistencies that could prevent accurate matches.

### Multi-Pass Matching Strategy

To address unique matching requirements, we developed a multi-pass approach:

#### First Pass: Direct Brand and Model Matching

We first focused on exact brand matches within each entry’s split components. Unique matches were assigned to their corresponding brands. Next, we want to infer the brand information from models. We then matched car models, assigning the model only when a unique match was found. If no match or multiple matches occurred, the entry was marked as "NA" to prevent incorrect assignments. This process ensured that only precise matches were retained, with unmatched entries flagged for further review.

#### Second Pass: Model-Based Brand Inference

Entries without direct brand matches were re-examined by inferring the brand from model names. When a model matched a brand, the corresponding brand was assigned. However, if multiple models or brands were found, or if the inference was ambiguous, the entry was set to "NA" to prevent misattribution. This ensures that only unique and confident matches are retained, while ambiguous cases are flagged for further review.

### Special Case Handling

A subset of entries presented additional challenges that required brand-specific processing. We created standardization rules to handle variations in abbreviations and series identifiers, ensuring consistent mapping of brands and models.

* + 1. Series Identification

We discovered that BMW's model names were too specific (e.g., "316i") while the given model list used broader categories (e.g., "3er"). To address this, we implemented pattern matching for BMW’s numeric series (e.g., "1er, " "3er, " "5er") and added special handling for the X, Z, and M series vehicles. This allowed us to correctly identify the model even when it included only numeric or letter indicators.

Similarly, we handled following brands:

**Mercedes-Benz:** Created specific handlers for class designations (e.g., "A-Class", "C-Class") and added logic to skip the brand name component due to complex name formats.

**Ford**: Added Ford models like "C-Max" and "S-Max" to prevent misclassification.

**Mazda**: Developed custom logic for Mazda’s series (e.g., "Mazda3", "Mazda6") to handle both numeric and letter-based model names.

**Honda**: Implemented recognition for Honda's "CR" series to ensure correct model classification.

**Volkswagen (VW)**: Standardized entries by implementing handlers for cases where "VW" appeared as an abbreviation.

#### Mercedes-Benz Class Processing Prefix Matching

We encountered an issue where some car names (e.g., "BMW316i") did not directly match the expected "BMW" brand due to the lack of a space between the brand and model. To address this, we implemented a pattern-based approach to detect entries starting with the prefix "bmw." Similarly, we handled the "golf" model, "vw" and "mazda" using the same logic.

#### Summary

All these specialized handlings for common patterns across brands helped us accurately resolve ambiguous entries, minimizing errors due to inconsistencies in naming.

## 横向组织结构图(1)Code Flow and Logic