2024 Vehicle Data Set

A Data Cleaning Project

07/12/24

Contents

[The Vehicle Dataset 2024 2](#_Toc171665871)

[Dataset Variables 2](#_Toc171665872)

[Data Cleaning 3](#_Toc171665873)

[Cleaning Findings 3](#_Toc171665874)

[Mitigation Methods 4](#_Toc171665875)

[Summary of Outcomes 4](#_Toc171665876)

[Export Clean Dataset 4](#_Toc171665877)

[Limitations 4](#_Toc171665878)

[Impact of Limitations 4](#_Toc171665879)

[Sources 4](#_Toc171665880)

# The Vehicle Dataset 2024

The "Vehicle Dataset 2024" dataset was found through Kaggle.com and is available under the ODC Attribution License. <https://www.kaggle.com/datasets/kanchana1990/vehicle-dataset-2024>

## Dataset Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** | **Variable Example** |
| **name** | Text | The vehicle's full name, including make, model, and trim. | "2024 Jeep Wagoneer Series II" |
| **description** | Text | A brief vehicle description, often including key features and selling points. | "White Knuckle Clearcoat 2023 Dodge Durango Pur…" |
| **make** | Categorical | The vehicle manufacturer (Ford, Toyota, BMW). | "Jeep" |
| **model** | Categorical | The model name of the vehicle. | "Wagoneer" |
| **type** | Categorical | The type of vehicle. (New, Used). | "New" |
| **year** | Continuous | The year the vehicle was manufactured | "2024" |
| **price** | Discrete | The price of the vehicles is in USD. | "74600.0" |
| **engine** | Text | Details about the engine, including type and specifications | "24V GDI DOHC Twin Turbo" |
| **cylinders** | Discrete | The number of cylinders in a vehicle's engine. | "6.0" |
| **fuel** | Categorical | The type of fuel the vehicle uses. (Gasoline, Diesel, Electric) | "Gasoline" |
| **mileage** | Continuous | The vehicle's mileage. | "32.0" |
| **transmission** | Categorical | The transmission type. (Automatic, Manual) | "8-Speed Automatic" |
| **trim** | Categorical | The trim level of the vehicle indicating different feature sets or packages | "Series II" |
| **body** | Categorical | The vehicle's body style. (SUV, Sedan, Pickup Truck) | "SUV" |
| **doors** | Discrete | The number of doors on the vehicle | "4.0" |
| **exterior\_color** | Categorical | The exterior color of the vehicle | "White" |
| **interior\_color** | Categorical | The interior color of the vehicle | "Global Black" |
| **drivetrain** | Categorical | The vehicle's drivetrain. (All-wheel Drive, Front-wheel Drive | "Four-Wheel Drive" |

# Data Cleaning

## Cleaning Findings

* **Duplicate row findings**
  + After removing the description column, 31 duplicate rows were found in the dataset.
* **Missing value findings**
  + **223** missing values were found in the price, engine, cylinders, fuel, mileage, transmission, trim, body, doors, exterior\_color, and interior\_color columns.
  + Count of missing values per column:
    - price: 23 missing,
    - engine: 2 missing,
    - cylinders: 103 missing,
    - fuel: 7 missing,
    - mileage: 32 missing,
    - transmission: 2 missing,
    - trim: 1 missing,
    - body: 3 missing,
    - doors: 7 missing,
    - exterior\_color: 5 missing, and
    - interior\_color: 38 missing.
  + Data Sparsity of < 2%.

A black and white image of a chart

Description automatically generated with medium confidence

Figure 1 – Visualization of missing values in the dataset. Solid bars indicate no missing values.

***Source Code****: Bilogur, (2018). Missingno: a missing data visualization suite. Journal of Open Source Software, 3(22), 547,* [*https://doi.org/10.21105/joss.00547*](https://doi.org/10.21105/joss.00547)

* **Outlier findings**
  + The quantitative variables inspected for outliers were the price, cylinders, mileage, and doors columns.
    - price: 28 outliers found, right-skew distribution.

A graph with numbers and a bar

Description automatically generated

* + - cylinders: 104 outliers found, left-skew distribution.

A graph with green bars and numbers

Description automatically generated

* + - mileage: 112 outliers found, right-skew distribution.

A graph with numbers and a number

Description automatically generated

* + - doors: 47 outliers found, left-skew distribution.

A graph with green and red lines

Description automatically generated with medium confidence

## Mitigation Methods

* **Dropping the description column**
  + A new DataFrame named vehicle\_only\_df was created by dropping the description column, which is best suited for analysis using NLP techniques and will be excluded from the DataFrame.
* **Dropping duplicate observations**
  + The Pandas .drop\_duplicates() function removed all identified duplicates.

## Summary of Outcomes

## Export Clean Dataset

# Limitations

## Impact of Limitations

# Sources

Bilogur, (2018). Missingno: a missing data visualization suite. *Journal of Open Source Software, 3(22), 547*, <https://doi.org/10.21105/joss.00547>

Kanchana1990. (2024, May 29). Vehicle Dataset 2024. *Kaggle.com*. Retrieved May 31, 2024, from <https://www.kaggle.com/datasets/kanchana1990/vehicle-dataset-2024>