

**Aim:** Introduction to Data science and Data preparation using Pandas steps.

### Theory:

Data preparation is a crucial step in data science, involving cleaning and transforming raw data into an analyzable format. Using Pandas, we can perform operations such as handling missing values, encoding categorical data, and scaling numerical features. Proper preprocessing ensures the dataset is reliable for analysis and modeling by addressing inconsistencies, missing data, and outliers.

### Problem Statement:

The Vehicle Safety Recall dataset, provided by NHTSA, contains 15 columns detailing various aspects of recall events, such as manufacturers, affected components, and corrective actions. This analysis focuses on:

- **Manufacturer Trends:** Identifying manufacturers prone to frequent recalls or specific defects.
- **Impact Analysis:** Understanding recall types affecting the largest populations and assessing average completion rates.
- **Temporal Patterns:** Detecting trends in recalls over time and seasonal spikes.
- **Safety Implications:** Investigating critical safety advisories like "Do Not Drive" or "Park Outside" and their resolution rates.

By cleaning the dataset and applying data preprocessing steps, the goal is to enhance its quality and draw actionable insights for stakeholders.

### Dataset Overview:

The dataset provides detailed information about vehicle safety recalls managed by the National Highway Traffic Safety Administration (NHTSA). It contains 15 columns, each capturing specific aspects of recall events. Below is a breakdown of the columns and their relevance:

1. **Report Received Date:** Date the recall was officially reported.
2. **NHTSA ID:** A unique identifier for each recall event.
3. **Recall Link:** A hyperlink to the recall details on the NHTSA website.
4. **Manufacturer:** Name of the vehicle or product manufacturer responsible for the recall.
5. **Subject:** Brief description of the recall issue.

6. **Component:** The affected part of the vehicle/product (e.g., "POWER TRAIN").
7. **Mfr Campaign Number:** Manufacturer's internal reference for the recall.
8. **Recall Type:** Type of product involved (e.g., vehicle, tire, or car seat).
9. **Potentially Affected:** Number of units potentially impacted by the recall.
10. **Recall Description:** Detailed explanation of the defect or issue.
11. **Consequence Summary:** Description of the risks or consequences associated with the defect.
12. **Corrective Action:** Steps taken to address the defect.
13. **Park Outside Advisory:** Indicates whether there's an advisory to park outside for safety.
14. **Do Not Drive Advisory:** Indicates whether there's an advisory not to drive the affected vehicle.
15. **Completion Rate %:** Percentage of affected vehicles repaired or addressed.

## Steps:

### 1. Loading The Dataset

```
[1] import pandas as pd
```

```
[2] df = pd.read_csv('recalls.csv')
```

### 2. Description of the dataset

#### a. Information about dataset

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28671 entries, 0 to 28670
Data columns (total 15 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Report Received Date                     28671 non-null  object
1   NHTSA ID                                28671 non-null  object
2   Recall Link                              28671 non-null  object
3   Manufacturer                             28671 non-null  object
4   Subject                                  28671 non-null  object
5   Component                                28671 non-null  object
6   Mfr Campaign Number                      28624 non-null  object
7   Recall Type                              28671 non-null  object
8   Potentially Affected                     28630 non-null  float64
9   Recall Description                       26270 non-null  object
10  Consequence Summary                     23783 non-null  object
11  Corrective Action                       26283 non-null  object
12  Park Outside Advisory                    28671 non-null  object
13  Do Not Drive Advisory                    28671 non-null  object
14  Completion Rate % (Blank - Not Reported) 10007 non-null  float64
dtypes: float64(2), object(13)
memory usage: 3.3+ MB
```

## b. Description of Dataset

```
# Get the dataset's shape and basic statistics
print(f"Dataset Shape: {df.shape}")
print(df.describe(include='all'))
```

|        | Report Received Date | NHTSA ID \ |
|--------|----------------------|------------|
| count  | 28671                | 28671      |
| unique | 10023                | 28671      |
| top    | 10/17/2013           | 25E002000  |
| freq   | 42                   | 1          |
| mean   | NaN                  | NaN        |
| std    | NaN                  | NaN        |
| min    | NaN                  | NaN        |
| 25%    | NaN                  | NaN        |
| 50%    | NaN                  | NaN        |
| 75%    | NaN                  | NaN        |
| max    | NaN                  | NaN        |

|        | Recall Link \  |
|--------|--|
| count  | 28671  |
| unique | 28671  |
| top    | Go to Recall ( <a href="https://www.nhtsa.gov/recalls?nh...">https://www.nhtsa.gov/recalls?nh...</a> ) |
| freq   | 1  |
| mean   | NaN  |
| std    | NaN  |
| min    | NaN  |
| 25%    | NaN  |
| 50%    | NaN  |
| 75%    | NaN  |
| max    | NaN  |

|        | Mfr Campaign Number | Recall Type | Potentially Affected \ |
|--------|---------------------|-------------|------------------------|
| count  | 28624               | 28671       | 2.863000e+04           |
| unique | 11341               | 4           | NaN                    |
| top    | NR (Not Reported)   | Vehicle     | NaN                    |
| freq   | 16602               | 24940       | NaN                    |
| mean   | NaN                 | NaN         | 4.572011e+04           |
| std    | NaN                 | NaN         | 3.730381e+05           |
| min    | NaN                 | NaN         | 0.000000e+00           |
| 25%    | NaN                 | NaN         | 9.900000e+01           |
| 50%    | NaN                 | NaN         | 6.860000e+02           |
| 75%    | NaN                 | NaN         | 6.385500e+03           |
| max    | NaN                 | NaN         | 3.200000e+07           |

|        | Recall Description \                              |
|--------|---|
| count  | 26270   |
| unique | 25523   |
| top    | ON CERTAIN TRAILERS EQUIPPED WITH SEALCO SPRIN... |
| freq   | 28  |
| mean   | NaN   |
| std    | NaN   |
| min    | NaN   |
| 25%    | NaN   |
| 50%    | NaN   |
| 75%    | NaN   |
| max    | NaN   |

|        | Consequence Summary \                             |
|--------|---|
| count  | 23783   |
| unique | 17015   |
| top    | RELEASE OF COOLANT UNDER CERTAIN CONDITIONS CO... |
| freq   | 128   |
| mean   | NaN   |
| std    | NaN   |
| min    | NaN   |
| 25%    | NaN   |
| 50%    | NaN   |
| 75%    | NaN   |
| max    | NaN   |

|        | Corrective Action \                               |
|--------|---|
| count  | 26283   |
| unique | 25579   |
| top    | DEALERS WILL EQUIP AIR SYSTEMS WITH A PRESSURE... |
| freq   | 18  |
| mean   | NaN   |
| std    | NaN   |
| min    | NaN   |
| 25%    | NaN   |
| 50%    | NaN   |
| 75%    | NaN   |
| max    | NaN   |

|        | Park Outside Advisory | Do Not Drive Advisory \ |
|--------|-----------------------|-------------------------|
| count  | 28671                 | 28671                   |
| unique | 2                     | 2                       |
| top    | No                    | No                      |
| freq   | 28601                 | 28510                   |
| mean   | NaN                   | NaN                     |
| std    | NaN                   | NaN                     |
| min    | NaN                   | NaN                     |
| 25%    | NaN                   | NaN                     |
| 50%    | NaN                   | NaN                     |
| 75%    | NaN                   | NaN                     |
| max    | NaN                   | NaN                     |

|        | Completion Rate % (Blank - Not Reported) |
|--------|--|
| count  | 10007.000000                             |
| unique | NaN                                      |
| top    | NaN                                      |
| freq   | NaN                                      |
| mean   | 67.874214                                |
| std    | 29.937993                                |
| min    | 0.000000                                 |
| 25%    | 48.350000                                |
| 50%    | 76.390000                                |
| 75%    | 93.765000                                |
| max    | 100.000000                               |

## 3. Drop columns that aren't useful.

Columns that might not be necessary for analysis include Recall Link, Mfr Campaign Number, Park Outside Advisory, Completion rate(%). These columns do not provide much insight in the context of data analysis for recall trends or consequences. Therefore, you can drop them to simplify the dataset.

```
# Remove leading/trailing spaces from column names
df.columns = df.columns.str.strip()

# List of columns to drop
cols = ["Recall Link", "Mfr Campaign Number", "Park Outside Advisory", "Do Not Drive Advisory", "Completion Rate % (Blank - Not Reported)"]

# Drop the columns that are present in the DataFrame
df = df.drop(cols, axis=1)

# Display the updated DataFrame
print(df.head())
```

|   | Report Received Date | NHTSA ID             | Manufacturer               | Subject  | Component         | Recall Description                                | Consequence Summary                               | Corrective Action                                 |
|---|----------------------|----------------------|----------------------------|--|-------------------|---|---|---|
| 0 | 01/14/2025           | 25E002000            | GKN Automotive             | Driveshaft Can Break                             | POWER TRAIN       | GKN Automotive (GKN) is recalling certain repl... | A cracked or broken driveshaft can cause a los... | GKN will reimburse the cost of a replacement d... |
| 1 | 01/13/2025           | 25E001000            | N&B Mobility Solutions LLC | Charger Adapter May Cause Arcing or Shock Risk   | ELECTRICAL SYSTEM | N&B Mobility Solutions LLC (Nivion) is recalli... | Inadequate clearance between DC busbars may ca... | Nivion will replace the defective adapters, fr... |
| 2 | 01/13/2025           | 25V005000            | Forest River, Inc.         | Cooktop Burner Tube May Crack and Cause Gas Leak | EQUIPMENT         | Forest River, Inc. (Forest River) is recalling... | A gas leak in the presence of an ignition sour... | Owners are advised not to use the cooktop unti... |
| 3 | 01/13/2025           | 25V006000            | Kia America, Inc.          | Loss of Headlights and Taillights/FMVSS 108      | ELECTRICAL SYSTEM | Kia America, Inc. (Kia) is recalling certain 2... | A loss of headlights and taillights can reduce... | Dealers will update the BDC software, free of ... |
| 4 | 01/13/2025           | 25V007000            | Winnebago Industries, Inc. | Spare Tire Carrier May Detach                    | EQUIPMENT         | Winnebago Industries, Inc. (Winnebago) is reca... | A detached spare tire carrier can become a roa... | Dealers will inspect, replace, and correctly t... |
|   | Recall Type          | Potentially Affected |                            |  |                   |   |   |   |
| 0 | Equipment            | 18.0                 |                            |  |                   |   |   |   |
| 1 | Equipment            | 130.0                |                            |  |                   |   |   |   |
| 2 | Vehicle              | 396.0                |                            |  |                   |   |   |   |
| 3 | Vehicle              | 74469.0              |                            |  |                   |   |   |   |
| 4 | Vehicle              | 107.0                |                            |  |                   |   |   |   |

Thus the columns now present in dataset are:

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28671 entries, 0 to 28670
Data columns (total 10 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Report Received Date                 28671 non-null  object
1   NHTSA ID                             28671 non-null  object
2   Manufacturer                         28671 non-null  object
3   Subject                             28671 non-null  object
4   Component                           28671 non-null  object
5   Recall Type                         28671 non-null  object
6   Potentially Affected                 28630 non-null  float64
7   Recall Description                   26270 non-null  object
8   Consequence Summary                 23783 non-null  object
9   Corrective Action                   26283 non-null  object
dtypes: float64(1), object(9)
memory usage: 2.2+ MB
```

#### 4. Take care of missing data.

##### a. Drop rows with maximum missing values.

Here we drop the rows which have more than 50% missing values. these can be done by `dropna()` function with threshold parameter=0.5.

```

▶ print(f"Dataset Shape before Dropping Rows: {df.shape}")
# Drop rows with the highest number of missing values
threshold = len(df.columns) * 0.5 # Drop rows where over 50% of columns are missing
df = df.dropna(thresh=threshold)

print(f"Dataset Shape After Dropping Rows: {df.shape}")

```

```

➡ Dataset Shape before Dropping Rows: (28671, 10)
Dataset Shape After Dropping Rows: (28671, 10)

```

```

▶ print(df.isnull().sum())

```

```

➡ Report Received Date      0
   NHTSA ID                  0
   Manufacturer              0
   Subject                   0
   Component                 0
   Recall Type               0
   Potentially Affected      41
   Recall Description        2401
   Consequence Summary       4888
   Corrective Action         2388
   dtype: int64

```

## b. Handle Missing Data

Here the above information says Potential Affected ,Recall Description ,Consequence Summary and corrective action contain some null values thus we need to handle missing data. For these columns, either fill in with a placeholder (e.g., "Unknown") or drop the rows if the missing data is significant.

```

[12] # Fill missing numerical values with the median
df['Potentially Affected'] = df['Potentially Affected'].fillna(df['Potentially Affected'].median())
# Fill missing categorical values with a placeholder
df['Recall Description'] = df['Recall Description'].fillna('Not Known')
df['Consequence Summary'] = df['Consequence Summary'].fillna('Unknown')
df['Corrective Action'] = df['Corrective Action'].fillna('Unknown')

print(df.isnull().sum()) # Verify no missing values remain

```

```

➡ Report Received Date      0
   NHTSA ID                  0
   Manufacturer              0
   Subject                   0
   Component                 0
   Recall Type               0
   Potentially Affected      0
   Recall Description         0
   Consequence Summary       0
   Corrective Action         0
   dtype: int64

```

## 5. Create dummy variables

For columns containing categorical data (e.g., Recall Type), we can create dummy variables. This is helpful for machine learning models.

```
# Convert categorical columns into dummy variables
df = pd.get_dummies(df, columns=['Recall Type'], drop_first=True)

print(df.head())
```

|   | Report Received Date | NHTSA ID  | Manufacturer \             |
|---|----------------------|-----------|----------------------------|
| 0 | 01/14/2025           | 25E002000 | GKN Automotive             |
| 1 | 01/13/2025           | 25E001000 | N&B Mobility Solutions LLC |
| 2 | 01/13/2025           | 25V005000 | Forest River, Inc.         |
| 3 | 01/13/2025           | 25V006000 | Kia America, Inc.          |
| 4 | 01/13/2025           | 25V007000 | Winnebago Industries, Inc. |

  

|   | Subject  | Component \       |
|---|--|-------------------|
| 0 | Driveshaft Can Break                             | POWER TRAIN       |
| 1 | Charger Adapter May Cause Arcing or Shock Risk   | ELECTRICAL SYSTEM |
| 2 | Cooktop Burner Tube May Crack and Cause Gas Leak | EQUIPMENT         |
| 3 | Loss of Headlights and Taillights/FMVSS 108      | ELECTRICAL SYSTEM |
| 4 | Spare Tire Carrier May Detach                    | EQUIPMENT         |

  

|   | Potentially Affected | Recall Description \                              |
|---|----------------------|---|
| 0 | 18.0                 | GKN Automotive (GKN) is recalling certain repl... |
| 1 | 130.0                | N&B Mobility Solutions LLC (Nivion) is recalli... |
| 2 | 396.0                | Forest River, Inc. (Forest River) is recalling... |
| 3 | 74469.0              | Kia America, Inc. (Kia) is recalling certain 2... |
| 4 | 107.0                | Winnebago Industries, Inc. (Winnebago) is reca... |

|   | Consequence Summary \                             |
|---|---|
| 0 | A cracked or broken driveshaft can cause a los... |
| 1 | Inadequate clearance between DC busbars may ca... |
| 2 | A gas leak in the presence of an ignition sour... |
| 3 | A loss of headlights and taillights can reduce... |
| 4 | A detached spare tire carrier can become a roa... |

  

|   | Corrective Action                                 | Recall Type_Equipment \ |
|---|---|-------------------------|
| 0 | GKN will reimburse the cost of a replacement d... | True                    |
| 1 | Nivion will replace the defective adapters, fr... | True                    |
| 2 | Owners are advised not to use the cooktop unti... | False                   |
| 3 | Dealers will update the BDC software, free of ... | False                   |
| 4 | Dealers will inspect, replace, and correctly t... | False                   |

  

|   | Recall Type_Tire | Recall Type_Vehicle |
|---|------------------|---------------------|
| 0 | False            | False               |
| 1 | False            | False               |
| 2 | False            | True                |

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28671 entries, 0 to 28670
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Report Received Date                 28671 non-null  object
1   NHTSA ID                             28671 non-null  object
2   Manufacturer                         28671 non-null  object
3   Subject                             28671 non-null  object
4   Component                           28671 non-null  object
5   Potentially Affected                 28671 non-null  float64
6   Recall Description                   28671 non-null  object
7   Consequence Summary                 28671 non-null  object
8   Corrective Action                   28671 non-null  object
9   Recall Type_Equipment                28671 non-null  bool
10  Recall Type_Tire                     28671 non-null  bool
11  Recall Type_Vehicle                 28671 non-null  bool
dtypes: bool(3), float64(1), object(8)
memory usage: 2.1+ MB
```

## 6. Find out outliers (manually):

Outliers can be detected by looking at numerical columns like Potentially Affected. One method for identifying outliers is by visualizing the data using box plots or using statistical methods like the Z-score.

**First Quartile (Q1):**=QUARTILE(H2:H28719, 1)

Q1 = 99

**Third Quartile (Q3):**=QUARTILE(H2:H28719, 3)

Q3 = 6386

**Interquartile Range (IQR):**=Q3 - Q1

$IQR = 6386 - 99 = 6287$

### Outlier Boundaries:

**Lower Bound:**  $= Q1 - 1.5 * IQR$

Lower Bound  $= 99 - (1.5 * 6287) = -9331.5$

**Upper Bound:**  $= Q3 + 1.5 * IQR$

Upper Bound  $= 6386 + (1.5 * 6287) = 15816.5$

**Identifying Outliers:** Any value less than -9331.5 or greater than 15816.5 is considered an outlier.

|            |           |                             |                      |                   |           |        |
|------------|-----------|-----------------------------|----------------------|-------------------|-----------|--------|
| 01/31/2025 | 25V048000 | Ford Motor Company          | BACK OVER PREVENTIO  | 25S05             | Vehicle   | 72624  |
| 01/30/2025 | 25V043000 | Jayco, Inc.                 | EQUIPMENT            | 9901617           | Vehicle   | 412    |
| 01/30/2025 | 25V045000 | Autocar, LLC                | ELECTRICAL SYSTEM    | ACTT-2501         | Vehicle   | 130    |
| 01/28/2025 | 25V037000 | Mitsubishi Fuso Truck of Am | ELECTRICAL SYSTEM    | C10129            | Vehicle   | 233    |
| 01/24/2025 | 25V034000 | Forest River, Inc.          | EQUIPMENT            | 203-1889          | Vehicle   | 64     |
| 01/23/2025 | 25V029000 | Winnebago Towable           | EQUIPMENT            | CAM0000041        | Vehicle   | 144    |
| 01/23/2025 | 25V031000 | Honda (American Honda Mo    | ELECTRICAL SYSTEM    | EL1, AL0          | Vehicle   | 294612 |
| 01/23/2025 | 25V033000 | Subaru of America, Inc.     | WHEELS               | WRB-25            | Vehicle   | 20366  |
| 01/23/2025 | 25V030000 | Mack Trucks, Inc.           | SERVICE BRAKES, AIR  | SC0474            | Vehicle   | 142    |
| 01/23/2025 | 25V032000 | Honda (American Honda Mo    | BACK OVER PREVENTIO  | RKZ               | Vehicle   | 9221   |
| 01/22/2025 | 25V027000 | Forest River Bus, LLC       | STRUCTURE            | 05-1890           | Vehicle   | 37     |
| 01/22/2025 | 25V028000 | Toyota Motor Engineering &  | FUEL SYSTEM, GASOLIN | 25TA01 / 25LA01   | Vehicle   | 858    |
| 01/21/2025 | 25V026000 | Mack Trucks, Inc.           | SERVICE BRAKES, AIR  | SC0473            | Vehicle   | 21     |
| 01/21/2025 | 25E006000 | Oshkosh Corporation         | VEHICLE SPEED CONTR  | NR (Not Reported) | Equipment | 500    |
| 01/17/2025 | 25V021000 | Forest River, Inc.          | EQUIPMENT            | 503-1887          | Vehicle   | 18     |
| 01/17/2025 | 25E004000 | Cummins, Inc.               | FUEL SYSTEM, DIESEL  | C7111             | Equipment | 715    |
| 01/17/2025 | 25V024000 | Kia America, Inc.           | ELECTRICAL SYSTEM    | SC332             | Vehicle   | 80255  |
| 01/17/2025 | 25T001000 | Pirelli Tire, LLC           | TIRES                | NR (Not Reported) | Tire      | 2023   |
| 01/17/2025 | 25V023000 | Mercedes-Benz USA, LLC      | TIRES                | NR (Not Reported) | Vehicle   | 165    |
| 01/17/2025 | 25V020000 | Ford Motor Company          | POWER TRAIN          | 25S03             | Vehicle   | 259    |
| 01/17/2025 | 25V019000 | Ford Motor Company          | ELECTRICAL SYSTEM    | 25S02             | Vehicle   | 272817 |
| 01/17/2025 | 25V025000 | Ford Motor Company          | SUSPENSION           | 25S01             | Vehicle   | 149449 |

## 7. standardization and normalization of column

Standardization and normalization are crucial when dealing with numerical data that varies in scale, especially for machine learning algorithms.

```

from sklearn.preprocessing import StandardScaler, MinMaxScaler
# Standardization: Transform data to have a mean of 0 and a standard deviation of 1
standard_scaler = StandardScaler()
df['Potentially Affected (Standardized)'] = standard_scaler.fit_transform(df[['Potentially Affected']])

# Normalization: Scale data between 0 and 1
min_max_scaler = MinMaxScaler()
df['Potentially Affected (Normalized)'] = min_max_scaler.fit_transform(df[['Potentially Affected']])

# Display the updated DataFrame
print(df[['Potentially Affected', 'Potentially Affected (Standardized)', 'Potentially Affected (Normalized)']].head())

```

|   | Potentially Affected | Potentially Affected (Standardized) \ |
|---|----------------------|---------------------------------------|
| 0 | 18.0                 | -0.122429                             |
| 1 | 130.0                | -0.122129                             |
| 2 | 396.0                | -0.121415                             |
| 3 | 74469.0              | 0.077295                              |
| 4 | 107.0                | -0.122190                             |

  

|   | Potentially Affected (Normalized) |
|---|-----------------------------------|
| 0 | 5.625000e-07                      |
| 1 | 4.062500e-06                      |
| 2 | 1.237500e-05                      |
| 3 | 2.327156e-03                      |
| 4 | 3.343750e-06                      |

**Conclusion:**

This experiment demonstrated effective data cleaning and preparation techniques. Issues such as missing values, irrelevant data, and outliers were addressed, and the dataset was scaled for uniformity. These steps are essential for ensuring high-quality data and reliable model outcomes.