Data Curation Process Guide

This document outlines the key steps involved in the Data Curation process, providing a clear guide for structuring and cleaning data effectively. The purpose of this process is to ensure that the data is accurate, consistent, and ready for analysis.

Name: Al-Qasem AbuKashef

---

1. Data Sourcing

Purpose: Identifying and gathering datasets relevant to the project to ensure comprehensive analysis.

Deliverables:

- **Dataset Details**:

Aspect	Details
File Name	car details v4 -NEW.xlsx
Description	Dataset of car attributes including price, make, model, specifications, and ownership history.
Dataset Size	2,059 Rows & 20 Columns
File Size	385 KB
Source	Kaggle - Dataset Link

2. Data Profiling

Purpose: Assess the dataset's quality, identify inconsistencies, and prepare for cleaning.

Steps to Analyze Data Quality:

- **Structure Discovery**:
    - Verified checked each column formats in excel by navigating to the "format cells" option and confirming data types (e.g., text, numeric).
    - Identified missing values using "=COUNTBLANK" for each column such as engine (80 missing), max power (80 missing), and fuel tank capacity (113 missing) were counted calculate the total number of missing values in each column.
    - Outliers observed in price and kilometers columns by created histograms in excel to highlight extreme values.

- **Content Discovery**:
    - Found potential incorrect format in engine, max torque(Nm), make, model, and max power fields by scanned for inconsistent format columns like Engine and Max Power for patterns using text filters (e.g., contains "cc" or "bhp").
    - Verified categorical columns (Fuel Type, Transmission, and Drivetrain) for consistency through to the "Filter" option to ensure consistency.

- **Relationship Discovery**:
    - Cross-referenced related columns (e.g., Make, Model, and Engine) by grouping and filtering for unexpected combinations to ensure logical consistency.

Issue	Column(s) Affected	How It Was Discovered	Observations
Missing Values	Engine, Max Power, etc.	=COUNTBLANK in Excel for each column	80-113 missing values in key columns
Outliers	Price, Kilometer	Used histograms in Excel	Prices exceeded expected range; high mileage outliers
Inconsistent Formats	Engine, Max Power make, model	Text filter for units and patterns	Variations in format: "1198 cc" vs. "1,198 cc"
Cross-Column Issues	Make, Model, Engine	Grouped data by related columns in Excel	Logical relationships between columns appear consistent

    -

3. Data Wrangling

Purpose: Clean and transform the dataset for analysis.

Deliverables:

- **Data Cleaning Actions**:
    - Standardized numerical values (e.g., bhp, cc and Nm units in Max Power, Engine and Max Torque) using the function of "text to columns" in excel to extract values, units like "cc" or "bhp" were discarded after extraction.
    - The percentage that was removed from the data set is 5%.
    - Missing values in columns such as Engine (80 missing), Max Power (80 missing), and Fuel Tank Capacity (113 missing) was removed
    - Reformatting all columns fields by scanned columns for patterns using text filters to ensure consistency.
    - Renamed the column "Drivetrain" to Wheel Drive Type for clarify.
    - Removed outliers in Price and Kilometer using histograms and conditional formatting.

- **Data Table Schema**:

Column name	Type	Description	Action made & reason
Make	String	Manufacturer of the car.	standardized format for consistency
Model	String	Specific model name.	standardized format for consistency
Price	Integer	Price of the car in local currency.	removed outliers, standardized format for consistency
Year	Integer	Manufacturing year.	standardized format for consistency
Kilometer	Integer	Kilometers driven.	removed outliers, standardized format for consistency
Fuel type	String	Type of fuel (e.g., petrol, diesel).	standardized values to lowercase for consistency
Transmission	String	Gear type (e.g., manual, automatic).	standardized format for consistency
Location	String	Location of the seller.	standardized format for consistency
Color	String	Color of the car.	standardized format for consistency
Owner	String	Ownership history (e.g., first, second).	standardized format for consistency
Seller type	String	Type of seller (e.g., individual, dealer).	standardized format for consistency
Engine (cc)	Integer	Engine capacity in cc.	splitted to extract numerical values, removed null values
Max power (bhp)	Integer	Maximum power output.	splitted to extract numerical values, removed null values
Max Torque(Nm)	Integer	Maximum torque output.	splitted to extract numerical values, removed null values
Wheel drive type	String	Drivetrain type (e.g., fwd, rwd).	renamed for clarify, removed null values
Length (mm)	Integer	Length of the car in mm.	removed null values
Width (mm)	Integer	Width of the car in mm.	removed null values
Height (mm)	Integer	Height of the car in mm.	removed null values
Seating capacity	Integer	Number of seats.	removed null values
Fuel tank capacity (l)	Integer	Fuel tank capacity in liters.	removed null values