Task Brief: Lego inventory cleaning

Dear Team,

As you know the inventory was split among several groups to enter the data faster. Now, we want to merge all the data into a single database/inventory. It is very important for the business inventory that each piece has a unique identifier and that there are no missing values nor duplicates. At the end of the day, I want to have the database in an excel book. Therefore, from the file below, please create a clean inventory of all the pieces and create a small project report explaining what you did and showing an inspection of the dataset.

https://docs.google.com/spreadsheets/d/17o2TJJ3_pmrFsFNIdhxyPW3PFO0zyksoSVbrWzrDJoU/edit?usp=sharing

Project Report Suggested Structure

1. **Dataset Overview (*of the clean version*)**

Item	Description
Dataset name	
Authors	
Number of entries	
Number of features/variables	
Format file (.csv, .txt, etc)	

2. **Dataset Structure (*of the clean version*)**

Feature/variable	Data type	Description	Number of Unique values	Example values

3. **Descriptive statistics (*of the clean version*)**

Numeric columns

	Column 1	Column 2	Column 3
Count			
Mean			
Standard deviation			
Min			
25%			

50%			
75%			
Max			

Categorical/object columns

	Column 1	Column 2	Column 3
Count			
Number of unique values			
Most frequent value			
Most frequent value (frequency)			
Least frequent value			
Least frequent value (frequency)			

3. **Exploratory plots (optional)**

Feel free to create some basic plots if they are necessary to understand the dataset.

4. **Data cleaning procedure**

4.1 Major data inconsistencies:

Issue	Names of Columns affected	Description of the Issue	Action Taken
Inconsistent column labeling			
Wrong data types			
Missing values			
Duplicates			
Inconsistent categories			

*Feel free to change the format of this table or write it in plain text but well structured.

4.2 Minor data inconsistencies *(if some issues cannot be reported in the above table)*

5. **Recommendations for good practices regarding data collection**

6. **AI Disclaimer:**

If you had to use AI code in certain part of your code and where.

Suggested notebook structure

 # Title of the notebook

Libraries Loading

The libraries you are using

Data Loading

How you are loading the data

Data Initial Inspection

The initial status of the dataset in its raw form.

Data Cleaning

Very explicit step by step procedure of data cleaning. Imagine that other person from other group or department needs to understand what you did,

Data Final Inspection

The final status of the dataset after cleaning.

(Optional) Basic exploration plots

You can start visualizing here some basic distributions and by using histograms, boxplots and start creating some barplots. Nothing fancy.