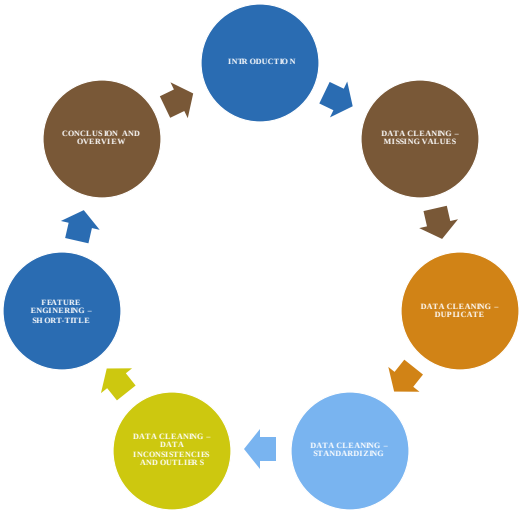
PRODUCT DATA CLEANING REPORT

HNG Data Analysis Stage ONE- HNG username(ACSP)

RELEVANT LINKS FOR THE TASK

- Jupyter Notebook
- CLEANED DATASET
- HNG Tech Internship
- HNG Tech Data Analysis

CONTENT LAYOUT



1. INTRODUCTION

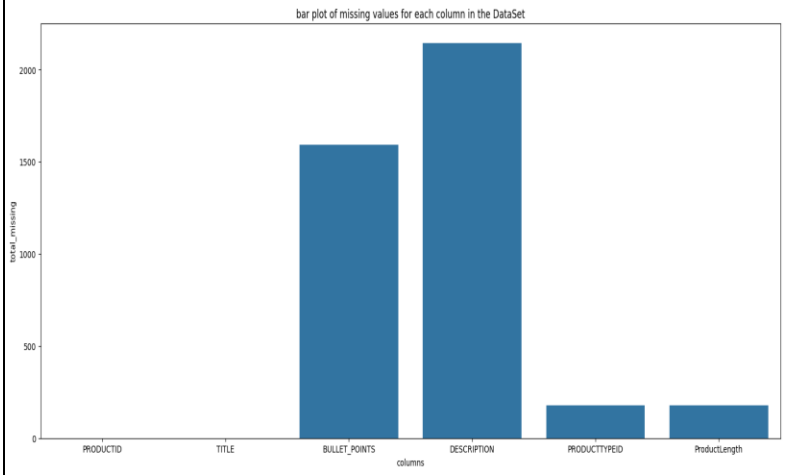
- Product Data
  - The above named was given dataset for this task.
  - This shape of the dataset is3847 rows and 6 columns
  - The data types comprises of two float64, one int64, and 3 object
  - The statistical view of the numerical columns (float and int) are given in the below table:

	PRODUCTID	PRODUCTTYPEID	ProductLength
count	3.847000e+03	3669.000000	3669.000000
mean	1.456557e+06	3932.736986	1150.529020
std	8.666684e+05	3970.908660	2665.897894
min	1.303000e+03	0.000000	1.000000
25%	6.922785e+05	154.000000	507.873000
50%	1.441218e+06	2879.000000	640.000000
75%	2.214798e+06	6337.000000	1023.622046
max	2.999397e+06	13330.000000	96000.000000

1. INTRODUCTION CONTD

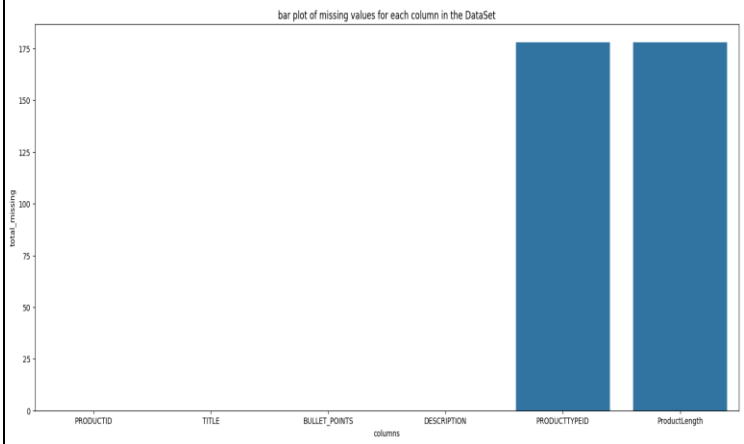
- The object columns all have high cardinality, that is, number of unique values, they are given below:
  - TITLE 3541
  - BULLET_POINTS 2116
  - DESCRIPTION 1609

2. DATA CLEANING(DC)- MISSING VALUES



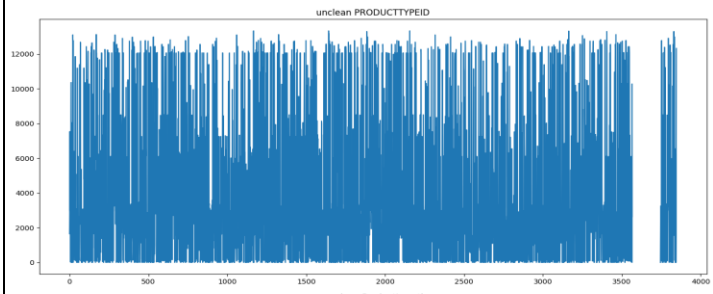
- The total sum of missing values as shown in his bar plot are:
  - Product id – 0
  - Title – 0
  - Bullet points - 1591
  - Description - 2144
  - Product type id - 178
  - Product length - 178

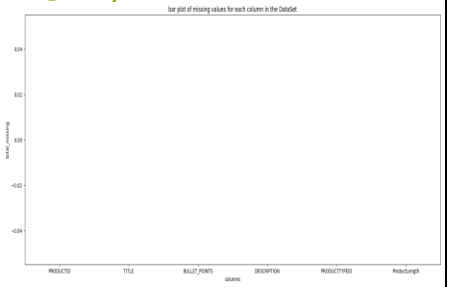
2. DC– MISSING VALUES -(CATEGORICAL)

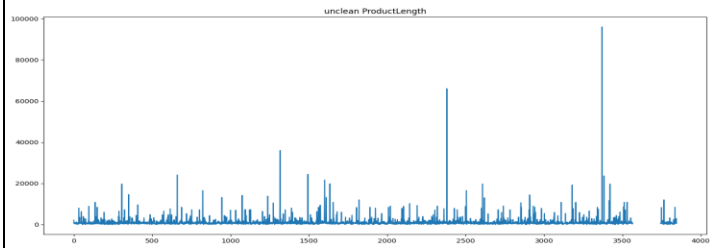


- Identified columns with missing values (categorical: object dtype).
- Applied fillna('Not Available') to fill missing values
- Bullet Points and Description columns were specifically imputed
- Result: Zero missing values in the respective columns
- All null values replaced with 'Not Available'

2. DC– MISSING VALUES -(NUMERICAL)

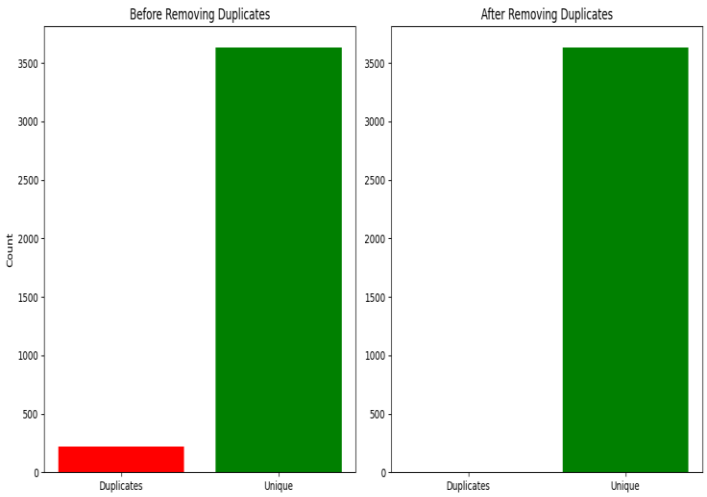






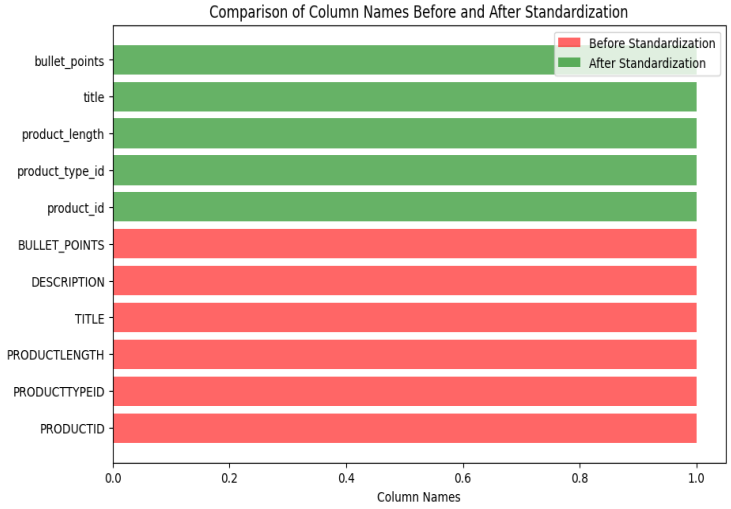
- Identified numerical columns with missing values (float64 dtype)
- Applied KNN Imputer with n_neighbors=5 for imputation
- Replaced missing values using the 5 nearest neighbors
- Result: Numerical columns have no missing values

2. DC- DUPLICATE



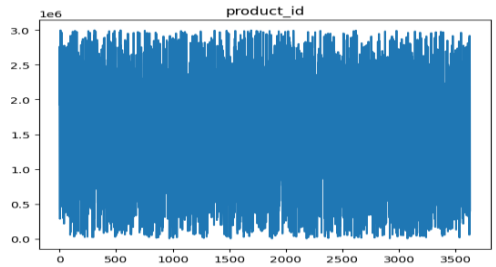
- There are 217 duplicate rows in the data set
- I checked for the duplicate rows using duplicated()
- I also remove duplicate using drop_duplicates()
- The index of the dataset after the operation was unsequential and so I reset index numbering using reset_index()

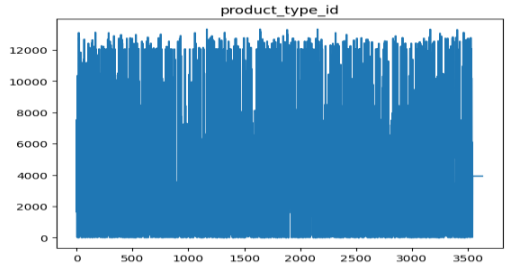
2. DC– STANDARDARDIZATION

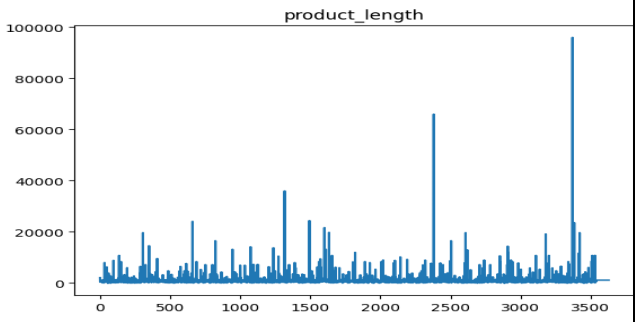


- The column names follows unstandard format as seen by the red bars.
- I thus, standardize the name by change them to the lowercase using 'df.columns.str.lower()'
- An underscore was use to join words. .str.replace().
- The standard column name are show with the green bar

2. DC – INCONSISTENTIES AND OUTLIERS



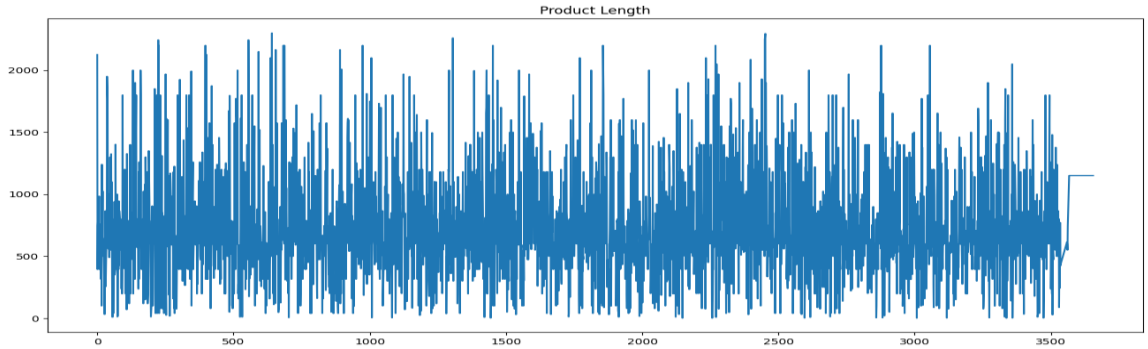




- Among the three numerical columns, outliers and inconsistencies are found in the product length
- These is shown in the spikes in the grapgh above.
- interquartile range(IQR) is a good measure of centered dispersion and it is not sensitive to outliers.
- Using the IQR, we will calculate for the lower limit and upper limit of outliers for the product length

2. DC – INCONSISTENTIES AND OUTLIERS CONTD

- It result in a two tail distribution
- the values above the lower limit (2311.982185312252)and the values below the upper limit (-1.3968146877475647) were dropped.
- Thus a total of 271 rows were further dropped from the dataset.



3. SHORT TITLE

- In order to extract key portions of text data for better readability
- I created a function to extract the first three and last two words joined with anhypen from each title row using the def and return parameters.
- the function was thus applied the function to the "title" column and assigned to a new "short_title" column
- The below is the final result of the first two rows in the data set.

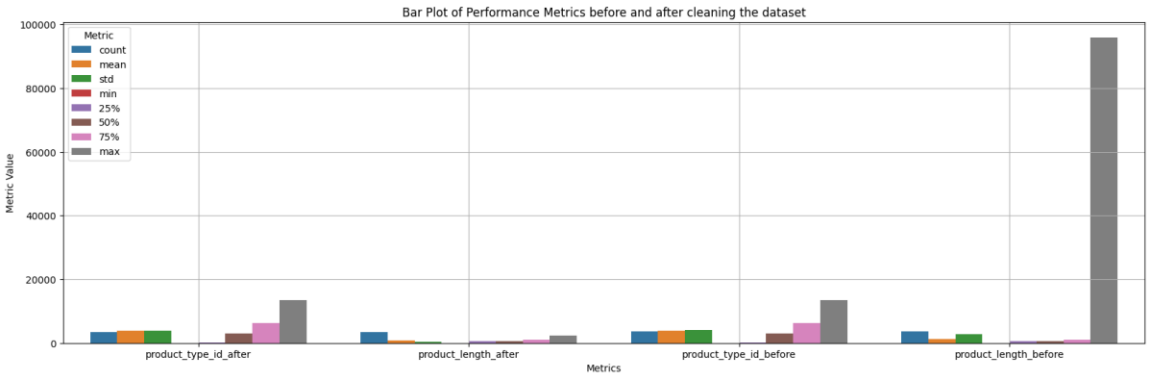
	product_id	title	bullet_points	description	product_type_id	product_length	short_title
0	1925202	ArtzFolio Tulip Flowers Blackout Curtain for D...	[LUXURIOUS & APPEALING: Beautiful custom-made ...	Not Available	1650.0	2125.98	ArtzFolio Tulip Flowers - 2 PCS
1	2673191	Marks & Spencer Girls' Pyjama Sets T86_2561C_N...	[Harry Potter Hedwig Pyjamas (6-16 Yrs),100% c...	Not Available	2755.0	393.70	Marks & Spencer - T86_2561 C_Navy Mix_9-10Y

4. CONCLUSION

- Reduction in Dataset Size
  - Before Cleaning: 3,847 records
  - After Cleaning: 3,359 records
  - Improvement: 488 rows were removed, likely containing missing values or outliers.
- Changes in Product Length Distribution
  - Mean Product Length: Reduced from 1,150.53 to 734.36, indicating the removal of extreme values.
  - Standard Deviation: Reduced from 2,665.90 to 407.85, showing a more consistent and less dispersed dataset.

	product_id	product_type_id	product_length
count	3.359000e+03	3359.000000	3359.000000
mean	1.425476e+06	3836.778682	734.362483
std	8.798367e+05	3931.671908	407.846727
min	1.303000e+03	0.000000	1.000000
25%	6.474105e+05	143.000000	500.000000
50%	1.385458e+06	2879.000000	614.000000
75%	2.206409e+06	6130.000000	942.440000
max	2.999397e+06	13330.000000	2300.000000

3. CONCLUSION – CONTD



4. CONCLUSION- CONTD

- Changes in Product Length Distribution Contd
  - Maximum Product Length: Dropped from 96,000 to 2,300, confirming the removal of extreme outliers.
- More Reliable Quartiles (Product Length):
  - quartiles after cleaning are now more stable and representative of the dataset
- Product Type ID Distribution.
  - Standard Deviation: Slight reduction from 3,970.91 to 3,931.67, suggesting a more controlled spread.
  - Mean Product Type ID: Slight decrease from 3,932.73 to 3,836.78, indicating adjustments in categorical data.
- Overall Improvements
  - Outliers removed, making the dataset more reliable.
  - Reduced standard deviation, leading to a more consistent distribution.
  - Lower maximum values, indicating extreme values were handled properly.
  - Refined quartiles, making statistical summaries more meaningful.

END OF REPORT