Data Quality Report

# Unnamed Columns Count (Dataset Level):

WARNING: There are 9 columns found which had no name.  
SOLUTION: Assign column names manually so as to remove confusion during data file handling.

# Duplicate Rows Percentage (Dataset Level):

INFO: There are 0.0% duplicate rows in the uploaded dataset.

# Outliers Detection (Column Level):

WARNING: There are 3 columns contain outlier values. They are as follows:

• Column Consumer Total contains 42 outlier values ranging from 1264 to 195580.

• Column Corporate Total contains 19 outlier values ranging from 1417 to 121885.

• Column Home Office Total contains 11 outlier values ranging from 1311 to 74255.  
  
SOLUTION: Outliers in some columns could be an issue as they might impact statistical measures such as mean and standard deviation in a wrong way so that they show influenced results. Moreover, if one plans to use the outlier affected columns in regression models, then outliers can distort the relationship between variables in the regression model and impact the model's coefficients and overall fit disproportionately. If the values in the range above are same, then it might be due to a singular value being the sole outlier but having high frequency in the column.

# Null Values Proportion (Column Level):

WARNING: The Null Values Proportion of all columns are as follows:

• Column Segment>> contains 0.0% of Na Values

• Column Consumer contains 91.12% of Na Values

• Column Unnamed: 2 contains 96.47% of Na Values

• Column Unnamed: 3 contains 88.81% of Na Values

• Column Unnamed: 4 contains 69.1% of Na Values

• Column Consumer Total contains 45.86% of Na Values

• Column Corporate contains 95.01% of Na Values

• Column Unnamed: 7 contains 99.27% of Na Values

• Column Unnamed: 8 contains 93.8% of Na Values

• Column Unnamed: 9 contains 81.39% of Na Values

• Column Corporate Total contains 69.83% of Na Values

• Column Home Office contains 98.42% of Na Values

• Column Unnamed: 12 contains 99.15% of Na Values

• Column Unnamed: 13 contains 97.08% of Na Values

• Column Unnamed: 14 contains 89.05% of Na Values

• Column Home Office Total contains 84.06% of Na Values  
  
SOLUTION: A high count of Na values inside a column might lead to ineffective analysis results and therefore na's have to be replaced either by a mean of the non-na values of the column or a manual value which makes business or practical sense

# Inconsistent Data Types Check (Column Level):

WARNING: There are 12 columns with inconsistent data type values. They are as follows:

• Column Consumer contains multiple data type values, namely float and str.

• Column Unnamed: 2 contains multiple data type values, namely float and str.

• Column Unnamed: 3 contains multiple data type values, namely str and float.

• Column Unnamed: 4 contains multiple data type values, namely float and str.

• Column Corporate contains multiple data type values, namely float and str.

• Column Unnamed: 7 contains multiple data type values, namely float and str.

• Column Unnamed: 8 contains multiple data type values, namely float and str.

• Column Unnamed: 9 contains multiple data type values, namely float and str.

• Column Home Office contains multiple data type values, namely float and str.

• Column Unnamed: 12 contains multiple data type values, namely float and str.

• Column Unnamed: 13 contains multiple data type values, namely float and str.

• Column Unnamed: 14 contains multiple data type values, namely float and str.  
  
SOLUTION: A single column cannot contain values of multiple data types. Use pandas functions such as astype() or pd.to\_datetime() or pd.to\_numeric() to coerce each relevant column values into a single data type. The next section of this report calculates summary statistics and if one of the data inconsistency affected column is supposed to be integer or float type, then it won't show up in the summary statistics report. This is one such example of the disadvantages of a column having values of multiple data types.

# Summary Statistics (Column Level):

INFO: Summary statistics for the number-related columns are as follows:

• Column Consumer Total has a non-null row count of 445, mean of 879, median 173, mode 15, minimum value 2, maximum value 195580, standard deviation 9285, variance 86217144, range 195578, first quartile 36, third quartile 517, skewness 20, kurtosis 438, and has 377 na values.

• Column Corporate Total has a non-null row count of 248, mean of 982, median 158, mode 12, minimum value 1, maximum value 121885, standard deviation 7761, variance 60237256, range 121884, first quartile 31, third quartile 583, skewness 15, kurtosis 241, and has 574 na values.

REMARKS: If you find insufficient number-related columns above then please check if some columns which should be number-related have multiple data types values inside them (as mentioned the above section). Please make sure no such column is being considered as a multiple data type column.

# Unnamed Columns Count (Dataset Level):

WARNING: There are 9 columns found which had no name.  
SOLUTION: Assign column names manually so as to remove confusion during data file handling.

# Duplicate Rows Percentage (Dataset Level):

INFO: There are 0.0% duplicate rows in the uploaded dataset.

# Outliers Detection (Column Level):

WARNING: There are 3 columns contain outlier values. They are as follows:

• Column Consumer Total contains 42 outlier values ranging from 1264 to 195580.

• Column Corporate Total contains 19 outlier values ranging from 1417 to 121885.

• Column Home Office Total contains 11 outlier values ranging from 1311 to 74255.  
  
SOLUTION: Outliers in some columns could be an issue as they might impact statistical measures such as mean and standard deviation in a wrong way so that they show influenced results. Moreover, if one plans to use the outlier affected columns in regression models, then outliers can distort the relationship between variables in the regression model and impact the model's coefficients and overall fit disproportionately. If the values in the range above are same, then it might be due to a singular value being the sole outlier but having high frequency in the column.

# Null Values Proportion (Column Level):

WARNING: The Null Values Proportion of all columns are as follows:

• Column Segment>> contains 0.0% of Na Values

• Column Consumer contains 91.12% of Na Values

• Column Unnamed: 2 contains 96.47% of Na Values

• Column Unnamed: 3 contains 88.81% of Na Values

• Column Unnamed: 4 contains 69.1% of Na Values

• Column Consumer Total contains 45.86% of Na Values

• Column Corporate contains 95.01% of Na Values

• Column Unnamed: 7 contains 99.27% of Na Values

• Column Unnamed: 8 contains 93.8% of Na Values

• Column Unnamed: 9 contains 81.39% of Na Values

• Column Corporate Total contains 69.83% of Na Values

• Column Home Office contains 98.42% of Na Values

• Column Unnamed: 12 contains 99.15% of Na Values

• Column Unnamed: 13 contains 97.08% of Na Values

• Column Unnamed: 14 contains 89.05% of Na Values

• Column Home Office Total contains 84.06% of Na Values  
  
SOLUTION: A high count of Na values inside a column might lead to ineffective analysis results and therefore na's have to be replaced either by a mean of the non-na values of the column or a manual value which makes business or practical sense

# Inconsistent Data Types Check (Column Level):

WARNING: There are 12 columns with inconsistent data type values. They are as follows:

• Column Consumer contains multiple data type values, namely float and str.

• Column Unnamed: 2 contains multiple data type values, namely float and str.

• Column Unnamed: 3 contains multiple data type values, namely str and float.

• Column Unnamed: 4 contains multiple data type values, namely float and str.

• Column Corporate contains multiple data type values, namely float and str.

• Column Unnamed: 7 contains multiple data type values, namely float and str.

• Column Unnamed: 8 contains multiple data type values, namely float and str.

• Column Unnamed: 9 contains multiple data type values, namely float and str.

• Column Home Office contains multiple data type values, namely float and str.

• Column Unnamed: 12 contains multiple data type values, namely float and str.

• Column Unnamed: 13 contains multiple data type values, namely float and str.

• Column Unnamed: 14 contains multiple data type values, namely float and str.  
  
SOLUTION: A single column cannot contain values of multiple data types. Use pandas functions such as astype() or pd.to\_datetime() or pd.to\_numeric() to coerce each relevant column values into a single data type. The next section of this report calculates summary statistics and if one of the data inconsistency affected column is supposed to be integer or float type, then it won't show up in the summary statistics report. This is one such example of the disadvantages of a column having values of multiple data types.

# Summary Statistics (Column Level):

INFO: Summary statistics for the number-related columns are as follows:

• Column Consumer Total has a non-null row count of 445, mean of 879, median 173, mode 15, minimum value 2, maximum value 195580, standard deviation 9285, variance 86217144, range 195578, first quartile 36, third quartile 517, skewness 20, kurtosis 438, and has 377 na values.

• Column Corporate Total has a non-null row count of 248, mean of 982, median 158, mode 12, minimum value 1, maximum value 121885, standard deviation 7761, variance 60237256, range 121884, first quartile 31, third quartile 583, skewness 15, kurtosis 241, and has 574 na values.

REMARKS: If you find insufficient number-related columns above then please check if some columns which should be number-related have multiple data types values inside them (as mentioned the above section). Please make sure no such column is being considered as a multiple data type column.