Data Quality Report

1. Overview:

This report will outline the initial findings based on the cleaned dataset (cal-wildfires-24220492.csv). It will summarise the data, describe various data quality issue observed and how we will fix them. Please see appendix for some background on the data. Appendix includes descriptive statistics and plots of changes made to the original dataset. This also includes histograms, box plots and bar plots.

On first look, the dataset does not look very clean. There are lots of missing values and few features where wrongly marked as continuous features. The dataset contains a high amount of categorical feature and very low number of continuous features. Few features have outliers that must be dealt with. One more thing that must be done is we have to clean the feature names.

2. Summary:

There is not much logical integrity test that we can do on the dataset since there are very little number of continuous features available. The major problems are with the categorical features.

Few of the categorical features have very high number of missing values. We must investigate if we can derive them from other features or delete those features altogether. There are also values in the categorical columns called “Unknown” which I have decided to keep it the same way.

3. Logical integrity test:

There are very few continuous features (5) and we have to check the logical integrity for few of them

Test 1: Year built cannot be a very low value or <= 0 ---- 459 values found
Test 2: Assessed improved value cannot be <= 0 ---- 448 values found

4. Review continuous features:

4.1. Descriptive Statistics:

There are not big issues with the continuous features. We found there are few logically wrong values in Assessed improved values and Year built features. These values cannot be derived from other columns.

Looking at the box plots, we notice the values are centred around the mean with not much deviation but there are few rows with very high values. On investigation, it seems possible for the feature to have such high values for few rows since it’s the value of houses and few luxury houses may have those kinds of high values.

The rest of the features namely Incident start date, latitude and longitude don't have any problems.

4.2. Histograms:

All histograms can be found in the appendix section. These plots can be found in the notebook as well.

4.3. Box plots:

All box plots can be found in the appendix section. These plots can be found in the notebook as well.

5. Review categorical features:

5.1. Descriptive statistics

Firstly, features with very high missing values. These must be investigated further to see if these can be derived from other features or dropped from the dataset.

Feature	Miss %
Community	4266 (57%)
Structure Défense Actions Taken	2645 (74%)
Units in structure	3048 (70%)
Distance - Propane Tank to Structure	1861 (81%)
Distance - Residence to Utility/Misc Structure	1403 (86%)

There is another list of features that are all location-based features. They are:

1. Street number
2. Street name
3. City
4. State
5. Zip code
6. County
7. Community

Looking at the data of these features, many of them have very high cardinality and some of them have 1 cardinality. But the only column with low cardinality and provide some information about location is county. With high cardinality columns, there comes a point where there is need to group up categories together. So, its best that we drop most of these columns and keep only county.

Next comes features that define the house. These features have little missing values and should be addressed. The best way to fill these is by replacing the null values with the mode values which makes sense.

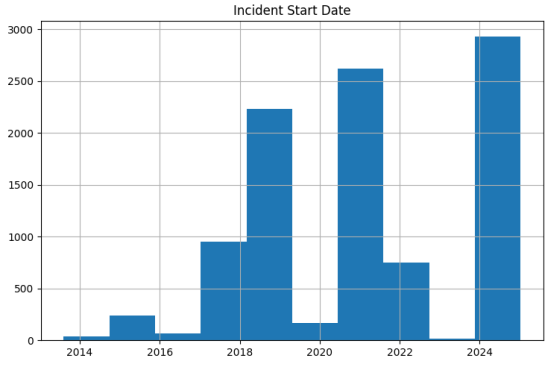
Next comes the high cardinality and low cardinality values. Both cannot be used properly. So, its best that we drop the high cardinality features.

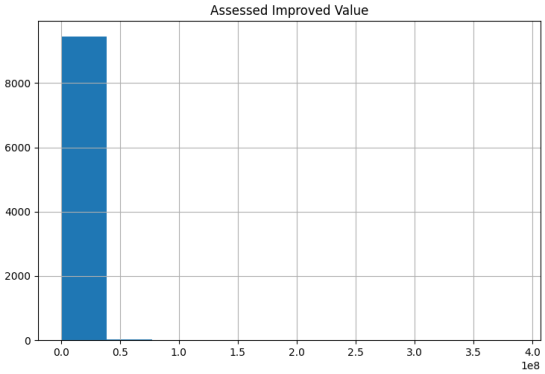
5.2. Bar plots:

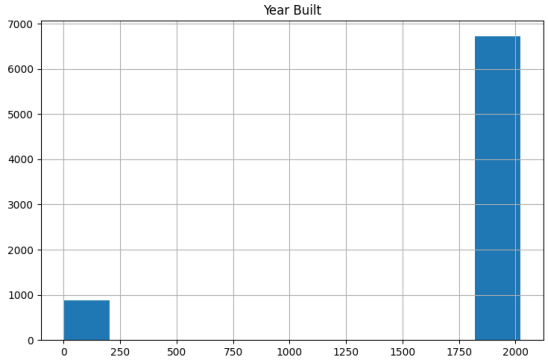
Bar plots are provided in the appendix. There are also given as part of the notebook.

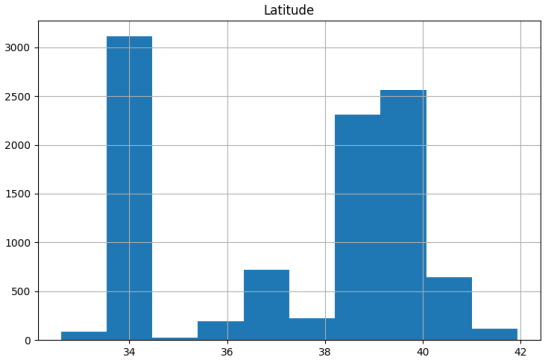
6. Appendix:

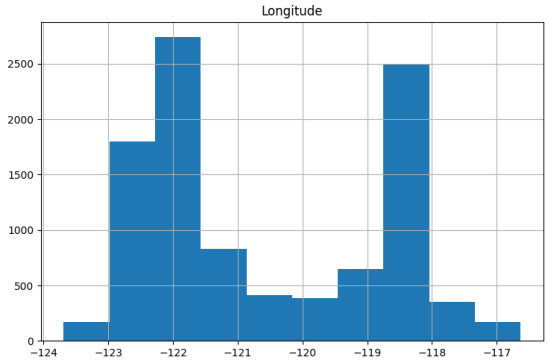
6.1. Histograms of continuous features:



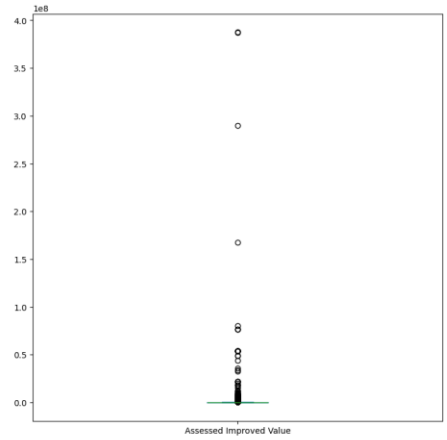


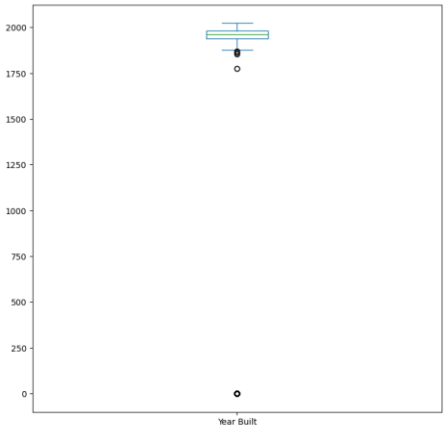


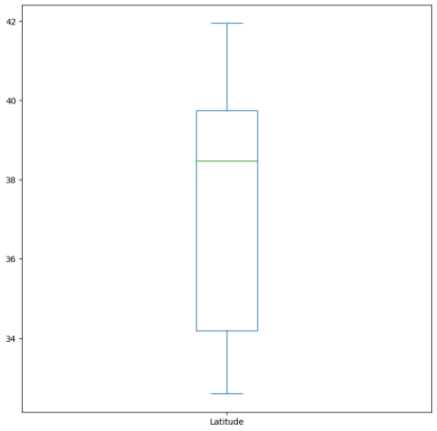


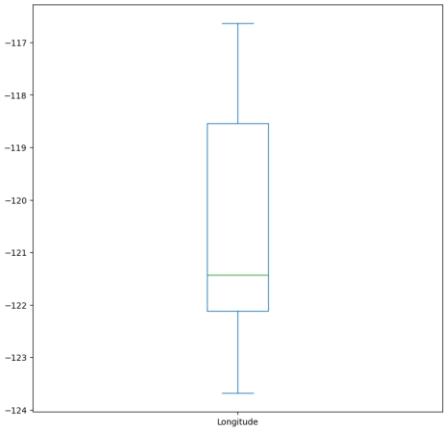


6.2. Box plots of continuous features:









6.3. Bar plots of categorical features:

There are so many categories to plot and this is given individually in the notebook.