Data Quality Report – Initial Findings

1. Overview

The dataset under analysis pertains to structural damages caused by fire incidents in California. Throughout this report, I will outline the initial findings based on the cleaned dataset 'cal-wildfires-24110699_Updated_part1.csv'. I will summarise the data, describe any data quality issues found and discuss potential solutions to address them. Please refer to the appendix for tables, histograms, boxplots and boxplots of continuous and categorical features respectively

The dataset contains no duplicate rows or columns. Two constant columns (State and Hazard) were found. These have been removed from the dataset as they have no variability and are therefore irrelevant for predictive analysis purposes. OBJECTID, Zip Code and Street Number have all been converted into categorical features because they all can act as regional identifiers. The dataset contains one datetime feature (Incident Start Date) and five continuous features (# Units in Structure (if multi unit), Assessed Improved Value (parcel), Year Built (parcel), Latitude and Longitude). The remaining features are all categorical in nature.

Furthermore, a considerable number of features were shown to contain missing values. Other issues include several features containing very high cardinalities. Finally, some logical integrity tests were performed, which have highlighted some additional consistencies within the data.

2. Summary

2.1 Missing values

The number of missing values for the dataset ordered from highest to lowest is the following:

- **Structure Defense Actions Taken** – 7,258 instances (72.6%)
- **Units in Structure (if multi unit)** – 6.956 instances (69.6%)
- **Community** – 5,819 instances (58.2%)
- **Zip Code** – 4,024 instances (40.2%)
- **City –** 2,517 instances (25.2%)
- **Year Built (parcel)** - 2,401 instances (24.0%)
- **Street Type** – 1,114 instances (11.1%)
- **Street Name** – 419 instances (4.2%)
- **Site Address (parcel)** - 597 instances (6.0%)
- **Street Number** – 333 instances (3.3%)
- **Vent Screen** – 399 instances (4.0%)
- **Eaves** – 389 instances (3.9%)
- **Window Pane** – 379 instances (3.8%)
- **Exterior Sliding** – 370 instances (3.7%)
- **Roof Construction** – 348 instances (3.5%)
- **County** - 1 instance (0.01%)

When a column contains 60% or more missing values, unreliable conclusions can be drawn. This is because confidence intervals become wider and less precise. Furthermore, important patterns in the data might become invisible. Therefore, selection bias is created. Imputation with the median is not appropriate when this proportion of values are missing as greater noise than signal is produced. In short, variance will be greatly reduced, and it will be difficult to make reasonable estimates. Therefore, these columns will be dropped. This strategy is implemented in part 2 of the homework.

2.2 Primary Key

Each entry in the dataset has a unique OBJECTID, which acts a primary key for each record.

2.3 Managing Logical Errors

- **Year Built (parcel)** – All values should be greater than zero. However, there are 836 instances where the value is zero, which is illogical for a building construction year.

3. Review Logical Integrity

Eight respective tests were carried out to assess the logical integrity of the data. These highlighted 14,934 cumulative issues, which impact data quality and the accuracy of analysis. Below are the findings:

- **Test 1: Check if any entries have an Incident Date that occurs before the building was constructed.**
  This is theoretically impossible. The dataset shows that zero rows failed the test, confirming the consistency of the data.
- **Test 2: Check that single-family residences do not have multiple residences.**
  Single-family residences are designed for and occupied by only one family or household at a time. If a building has multiple units, it would be classified under an alternative name such as a multi-family residence or duplex. 16 rows were shown to fail the test. Albeit this is a small number in the dataset of 10,000 rows, it highlights a data integrity issue.
- **Test 3: Check that no address components have zero or missing values. A valid fire incident should have a complete address.**
  3913 rows were shown to fail the test. This means that almost 40% of the dataset contains missing address information, which is critical for fire incident analysis.
- **Test 4: Check for unknown construction materials in structural elements. This is significant for assessing fire risk. Missing construction material data affects modelling potential.**
  2785 rows were shown to fail the test meaning fire risk modelling capabilities are limited.
- **Test 5: Check that the properties were destroyed despite having defense actions in place. This helps to assess the effectiveness of defense measures and to deliberate on which defense strategies are impactful.**
  1651 rows were shown to fail the test.
- **Test 6: Check that there is consistency between cities and communities. Communities should typically contain or relate to the city name. Consistency here is vital as concordant response planning to fire incidents is affected.**
  2494 rows were shown to fail the test.
- **Test 7: Check for any mismatches between city and zip codes.**
  4075 rows were shown to fail the test meaning there is substantial (approximately 41%) potential discrepancies.
- **Test 8: Check that all coordinates fall within California's geographic boundaries (32.5°-42.0°N, 124.5°-114.0°W). If any coordinates are out-of-bounds, this is indicative of data entry errors or misclassified incidents.**
  0 rows were shown to fail the test confirming valid coordinates and that all fire incidents occurred within the state of California.

4. Review Continuous Features

4.1 Descriptive Statistics

- **Units in Structure (if multi-unit)** - Median and upper quartile are both zero, which indicates that most properties are single-unit structures. Some outliers are present that have multiple units (maximum of 60). Frequency distribution data recorded in the notebook reveals the following: Many instances (2,760 instances) are recorded as zero. There are just 141 instances of single-unit structures. The distribution is very heavily skewed towards the right as very few properties have more than 10 units. It can be assumed that values greater than or equal to one for this feature are reserved for multi-unit structures.

- **Assessed Improved Value (parcel)** - Extreme variability present with a standard deviation of 14.6 million and a maximum value of 1.22 billion. This confirms the presence of significant outliers. The distribution is heavily skewed towards the right as there is a large gap between the median value of $160,240 and the mean of $807,870. 465 entries have a $0 value, which represents missing data**.** The heavy skewing of the distribution to the right is further confirmed as of the presences of extremely high values such as $54.14 million.
- **Year Built (parcel)** - 836 entries have a value of zero, which are likely missing values causing the mean (1748.7) to be skewed and much lower than the median (1962). There are logical errors present. The year should always be greater than zero and within the range of 1800 – 2025, as this is reasonable. The year of 1990 appears frequently at 106 instances denoting that it may be the default value for older buildings with unknown construction dates.
- **Latitude & Longitude** – All results are within the state of California, with small standard deviations. All numeric features besides latitude and longitude contain zeros, which may represent missing data rather than actual zero values. Therefore, both are reliable for data visualisation purposes.

4.2 Histograms

- **Units in Structure (if multi-unit)** – The distribution is extremely right skewed. This confirms that values greater than or equal to one for this feature are reserved for multi-unit structures.
- **Assessed Improved Value (parcel)** – This distribution confirms the detection of extreme outliers as discussed above.
- **Year Built (parcel)** - This distribution is bimodal.

4.3 Boxplots

- **Units in Structure (if multi-unit)** – Extreme outliers are shown at 40 and 60 respectively.
- **Assessed Improved Value (parcel)** – Extreme outliers are shown at $400 million, $500 million, and $1.2 billion respectively.
- **Year Built (parcel)** - There are outliers towards zero, both represent missing data. The upper whisker extends to around 2010 – 2020, which is representative of newer constructions.

5. Review Categorical Features

5.1 Descriptive Statistics

- **External structural elements -** Deck/Porch on Grade, Deck/Porch Elevated, Patio Cover/Carport Attached to Structure and Fences Attached to structure have 3,050, 4,786, 4,341 and 4,376 no values respectively. These features warrant further investigation as they often serve as pathways for fire to reach the main structure. Examining a combination of these features and damage outcomes could be crucial in revealing key insights surrounding structural susceptibility to fire.
- **Construction Material Features** - These display a high percentage of unknown values, for the following materials: Eaves (4,079 unknown values, 41.1%), Vent Screen (3,117 unknown values, 31.5%). For Roof Construction, the most frequent material used is "Asphalt" (4,419 instances, 44.4%). For Exterior Sliding, the most frequent material used is "Stucco Brick Cement" (2,591 instances, 26.1%). For Window Pane, the most frequent material used is "Multi Pane" (3,257 instances or 32.8%). The latter three features have significantly better data quality than the former two features, where "Unknown" is the dominant category and so they are more useful for data visualisation purposes.
- **Damage Status Distribution** - The "Damage" feature is of great importance within the dataset and classifies structures into two distinct categories of "Destroyed (>50%)" and "No Damage". A total of 5,662 structures or 56.6% are classified as "Destroyed(>50%)". It can be

assumed that the remaining, 4,338 structures or 43.4% are under the "No Damage" category. A slight imbalance is created between these classes and must be considered during predictive modelling. Since this imbalance does not exceed the 60/40 split (as specified earlier), a standard modelling approach can still be applied.

- **Structure Type and Categories -** Single-family residences dominate. This is proven by the fact that "Single Family Residence Single Story" accounts for 3,589 instances or 35.9% of cases. The "Structure Category" feature is indicative of "Single Residence" representing 6,598 cases or 66.0% of the dataset. However, the dataset also contains unusual categories of Hospital in 2 cases and Agriculture in one case only. These categories are clearly outliers and have limited contribution to meaningful statistical analysis and pose modelling challenges. These two categories should be verified to ensure that they accurately represent the structures under examination. This can be done via cross-referencing other data sources.

5.2 Bar Charts

Refer to the accompanying pdfs to see all the bar charts. The high cardinality bar charts are not particularly useful, but they do highlight the issue of high cardinality and give an idea of the distribution of the most frequent values within these features.

6. Review Datetime Feature

6.1 Descriptive Statistics

The “Incident Start Date” Column was converted to datetime format. There are no missing values within this column. All records within this column are valid. No logical errors or outliers were observed.

6.2 Histogram

A histogram visualising the datetime distribution was generated and saved in pdf format.

7. Appendix

7.1 Table of descriptive statistics for all Categorical features

	count	unique	top	freq
OBJECTID	10000	10000	131535	1
* Damage	10000	2	Destroyed (>50%)	5662
* Street Number	9667.0	5609.0	0.0	544.0
* Street Name	9581	3974	Pacific Coast	78
* Street Type (e.g. road, drive, lane, etc.)	8886	22	Road	3434
* City	7486	258	Unincorporated	1177
Zip Code	5976.0	145.0	0.0	1878.0
* CAL FIRE Unit	10000	27	LAC	2570
County	9999	47	Los Angeles	2570
Community	4198	424	Paradise	585
Structure Defense Actions Taken	2742	10	Unknown	2150
* Structure Type	10000	18	Single Family Residence Single Story	3589
Structure Category	10000	7	Single Residence	6598
* Roof Construction	9957	10	Asphalt	4419
* Eaves	9916	6	Unknown	4079
* Vent Screen	9907	7	Unknown	3117
* Exterior Siding	9936	11	Stucco Brick Cement	2591
* Window Pane	9927	5	Multi Pane	3257
* Deck/Porch On Grade	10000	6	No Deck/Porch	3050
* Deck/Porch Elevated	10000	6	No Deck/Porch	4786
* Patio Cover/Carport Attached to Structure	10000	5	No Patio Cover/Carport	4341
* Fence Attached to Structure	8412	4	No Fence	4376
Distance - Propane Tank to Structure	1883	5	Unknown	547
Distance - Residence to Utility/Misc Structure > 120 SQFT	1413	5	<30'	748
Site Address (parcel)	9611	8467		151

7.2 Table of descriptive statistics for Continuous features

	count	mean	std	min	25%	50%	75%	max
# Units in Structure (if multi unit)	3044.0	0.314	1.936e+00	0.000	0.000	0.000	0.000	6.000e+01
Assessed Improved Value (parcel)	9499.0	807870.078	1.460e+07	0.000	70000.000	160240.000	335016.000	1.220e+09
Year Built (parcel)	7599.0	1748.705	6.158e+02	0.000	1939.000	1962.000	1983.000	2.020e+03
Latitude	10000.0	37.381	2.487e+00	32.598	34.196	38.474	39.744	4.192e+01
Longitude	10000.0	-120.521	1.810e+00	-123.673	-122.125	-121.406	-118.542	-1.164e+02

7.3 Histograms for all Continuous features

See the following accompanying pdfs:

- continuous_histograms.pdf
- continuous_histograms_1-1.pdf

7.4 Boxplots for all Continuous features

See the following accompanying pdfs:

- continuous_boxplots_1-1.pdf
- continuous_boxplots_summary1-1.pdf

7.5 Bar plots for all Categorical features

See the following accompanying pdfs:

- categorical_high_cardinality_summary.pdf
- categorical_low_cardinality_summary.pdf
- categorical_lowcardinality_barcharts.pdf
- categorical_plots_high_cardinality.pdf

7.6 Histogram for Datetime Feature

See the following accompanying pdf:

- datetime_histograms_1-1.pdf