Data Understanding: Data Quality Plan

Feature	Data Quality Issue	Handling Strategy
OBJECTID	This is a unique identifier	Drop feature
Damage	No issues	Keep as is
Street Number	The cardinality is very high considering we changed it to a categorical feature	Drop feature
Street Name	The cardinality is very high (4036)	Drop feature
Street Type	Few missing values (11%)	Consider imputation with mode value
City	Has higher cardinality (256)	Drop feature
State	Single valued feature	Drop feature
Zip Code	Very high missing values (41%)	Try getting it from site address column, If it has high cardinality after filling the missing rows, Drop feature
CAL FIRE unit	No issues	Keep as is
County	Has very low missing values (3 rows)	Consider imputation with mode value
Community	Very high missing values (57%)	Drop feature
Incident Start Date	No issues	Keep as is
Hazard type	Single valued feature	Drop feature
Structure Defense Actions Taken	It has very high missing values of 73.55% and over that has 2069 values marked as unknown. Which gives us just 576 values out of 10000 to work with.	Drop feature
Structure Type	No issues	Keep as is
Structure Category	No issues	Keep as is
Units in Structure	69.52% missing values. When plotted with the rest, all the quartiles are 0 and very few values for other values.	Drop feature
Roof Construction	Low missing values 4%	Consider imputation with mode value
Eaves	Low missing values 4%	Consider imputation with mode value
Vent Screen	Low missing values 4%	Consider imputation with mode value
Exterior Siding	Low missing values 4% and also has annotator inconsistencies for 1 feature	Consider imputation with mode value and fix the

Feature	Data Quality Issue	Handling Strategy
	value	annotator inconsistency
Window Pane	Low missing values 4%	Consider imputation with mode value
Deck/Porch On Grade	Low missing values 16%	Consider imputation with mode value
Deck/Porch Elevated	Low missing values 16%	Consider imputation with mode value
Patio Cover/Carport Attached to Structure	Low missing values 16%	Consider imputation with mode value
Fence Attached to Structure	Low missing values 16%	Consider imputation with mode value
Distance - Propane Tank to Structure	Very high missing values 81%	Drop feature
Distance - Residence to Utility/Misc Structure > 120 SQFT	Very high missing values 86%	Drop feature
Assessed Improved Value	Outliers	Consider Clampping
Year Built	24.02% missing values, on top of that, has 872 values marked as 0 which also be considered as missing. which comes out to overall 32.74% missing values. Also has outliers	Consider imputation with median value and do nothing with outliers
Site Address	Very high cardinality for a categorical feature	Drop feature
Latitude	No issues	Keep as is
Longitude	No issues	Keep as is