Data Quality Report – Assignment 1 – COMP47350 – Stephen McMahon

1. Overview

In this report we will be analyzing the data collected by CAL FIRE with the aspiration of developing a data analytics solution to predict risk for structural damage due to wild-land fires. This report will summarize the data present in cal-wildfires-24241048.csv, describe the features present in the dataset, highlight the data quality issues present and finally outline how these data quality issues will be addressed. Analysis of the data quality of the issues will be conducted using a number of visualizations including histograms, box plots and bar plots, and the logical integrity of the data will be checked with a number of tests.

The data set itself has 33 features and 10,000 entries and describes structures in the state of California that have been affected by wildfires in the area. The earliest incident date in the dataset is 07/08/2013 and the most recent incident date is 07/01/2025. Upon initial inspection the data appears to be relatively clean with no duplicate features and only one duplicate row but the data does have several constant columns and several features that have over 80% Nan values which require further evaluation.

2. Summary

The vast majority of the features in the dataset are categorical with only 6 of the features being continuous. Two of the main overall issues relating to the categorical features is that of nan values with several features having large proportion of missing values including 'Structure Defense Action Taken' with 72.94% alongside other important features such as 'Zip Code' that has 39.97%. However, it is important to note that a large number of these missing zip code entries have a zip code included in the 'Site Address (parcel)' feature. For a large number of these categorical features that contain a high percentage of missing values the imputation of a value entitled 'missing' will better ensure data consistency and will be implemented to the data set. However, it is possible that some features have too many nan values and may need to be dropped as a result.

The entire dataset only contained one duplicate row and tests have been carried to confirm that each 'Object Id' entry – which can be considered the primary key – is unique. However, there are two constant features in the dataset 'State' and 'Hazard Type', which each have a cardinality of 1. In the case of the 'State' feature the only value is 'CA' and for 'Hazard Type' the only value is 'Fire'. Given that the data we are analyzing is considering structures in California that have either been at risk or damaged by wildfires these two features are clearly redundant. As a result, both features will be dropped.

The continuous features are also plagued by a large amount of missing values or '0' values which we can only assume indicates nan. In the case of '# Units in Structure (if multi unit)' 70.10% of values are missing, and for the feature 'Year Built (parcel)' 24.18% are missing. Alongside this, a large number of entries for the feature 'Year Built (parcel)' are '0' which given that we are looking at years does not make sense. We can assume that '0' is equivalent to missing but this should be verified with the domain expert first. As a result, the imputation of a new value for any missing entries is required. However, we must be careful when imputing new values for any '0' entries as we do not fully understand what this '0' value means as of yet.

Alongside this, there were a significant number of outliers for a number of the continuous features including 'Year Built (parcel) – which can be accounted for by the large number of '0' values – 'Assessed improved value (parcel)' and 'Units in Structure'. However, with the exception of 'Year Built (parcel) these outliers do make sense. For instance in the case of 'Units in Structure' the largest outlier is 60 but when examining the structure type and category we can see that this entry is a 'multi family residence multi story' and as result it follows that there would be 60 units. However, in the case of year built the large number of '0' values does not make sense and needs to be examined further and addressed.

To conclude, the main issue with this data set is that of missing values and the use of unclear values such as '0' for features such as 'Year Built (parcel)' which does not make any logical sense and reduces consistency. As a

result, imputing the value ‘missing’ for any NaN values in certain features and the removal of both features that are constant (‘State’, ‘Hazard Type’) will greatly improve the quality and consistency of this data set.

3. Review Logical Integrity

A number of tests were carried out to verify the logical integrity of various features.

- Test 1 - Ensure that each ‘Object ID’ entry is unique. The ‘Object ID’ can be considered the primary key of the dataset and as a result it is imperative that each entry is unique.
  - All Object ID values are unique.

- Test 2 - Check to ensure that the year built of each structure makes sense and took place within the last 200 years.
  - 834 invalid cases found (all ‘0’ values).

- Test 3 - Ensure that the incident start dates are logical and none take place in the future.
  - All 'Incident Start Date' values are valid (no future dates).

- Test 4 - Verify that all entries in the Assessed Improved Value (parcel) feature are positive numbers.
  - All 'Assessed Improved Value' entries are positive (greater than or equal to 0).

- Test 5 - Verify that all Latitude Values are unique.
  - 8 entries have longitudes that are present elsewhere in the data but 6 of these 8 entries have different longitudes which confirms that they are not duplicates.

- Test 6 – Verify that all Longitude values are unique.
  - 6 entries have longitudes that are present elsewhere in the data but 4 of these 6 entries have different latitudes which confirms that they are not duplicates.

- Test 7 – Verify that there are no entries that have both the same latitude and longitude.
  - 2 entries have latitudes and longitudes that are present elsewhere in the data. From examining both entries – ObjectID’s 74416 and 74451 - we can see that that they are almost identical but both have different Structure Types (‘Single Family Residence Single Story’ and ‘Utility Misc Structure’). This indicates that there are two different structures on a single property that were damaged. This explains why both entries have the exact same longitude and latitude and confirms that they are not duplicate entries or errors.

4. Categorical Features Review

4.1 Descriptive Statistics

There are a total of 6 continuous features in the data set. Descriptive details for each feature can be found below.

- *IncidentStartDate*
  - Treating this feature as data type ‘datetime’ and as a continuous feature as the data can be ordered and measured numerically. This feature has a count of 9999 values and a cardinality of 146. The feature has a low number of missing values.
  - Overall, there does not appear to be any issues with this feature and it does not have anywhere near as many missing values as other features in the dataset.

- *#UnitsinStructure(ifmultiunit)*
  - This continuous feature outlines the number of units in each structure. There is a total of 2990 entries in the feature but – similar to much of the data set – this feature has a large amount of missing values (7010 null entries in total). The feature has outliers with the max entry being the value '60'. However, all of these outliers appear plausible with the entry that has a value of 60 being associated with the type 'multi family residence multi story'.
  - Due to the fact that this feature has 70.10% missing values and that the data this feature contains does not provide much utility in relation to predicting structural risk this feature will be dropped. Imputing another value for 70% of this features entries is not advisable.

- *AssessedImprovedValue(parcel)*
  - This feature outlines the assessed value that the property has improved by. This feature has a low number of missing values 5.37% and a cardinality of 7621. There are some outliers with 4 structures having an assessed improved value of > $100,000,000. However, from examining the Structure type and category of these entries we can see that these structures are multi story commercial buildings, mixed residential buildings and utility structures which explains why the assessed improved value is so high.
  - Overall, the distribution of this feature appears to be normal and unlike many of the features of this data set it has a low level of missing values.

- *YearBuilt(parcel)*
  - This feature details the year that the structure was built but does not include the date and time and as a result the decision was made to treat this feature as a continuous feature and not as type date time. This feature has a count of 7582 entries and 24.18% of the values are missing. However, of these 7582 entries 833 have the outlier value of '0'. Given that it does not make any logical sense for any of these structures to have been built in the year 0 and as we are unable to consult a domain expert we will assume that the value 0 equates to missing. Alongside this, there is a value of '89' which also does not make any logical sense.
  - This feature has a large number of outliers that have illogical values given the nature of the dataset. Alongside this, the feature has a high number of null values. Imputing the value 'missing' for both entries that contain the value '0' and for nan values will ensure better data consistency.

- *Latitude*
  - This feature outlines the latitude of the structure in question. The feature has a count of 10000 values and from examining the histogram it also has a normal multi-modal distribution and a 0% missing value rate. The only concern with this feature is that the cardinality is 9995 indicating that there are a number of duplicate latitudes in the data set, which given that each structure is supposed to be unique does not follow. However, while there may be multiple instances of duplicate latitudes as long as there is a unique longitude value this confirms that each structure is in a different Geo-location.

- *Longitude*
  - This feature details the longitude of the structure in question. Akin to the latitude feature this feature has a count of 10000, a multi-modal distribution and a 0% missing value rate. However, the features cardinality is also 9996 which indicates duplicate values. From test 6 we can see that there are a number of entries that have duplicate longitudes but most have differing latitude values which confirms that they are unique locations.

4.2 Histograms

A histogram has been created for each feature and can be seen in section 8 of this report. The features 'Latitude' and 'Longitude' have a multi-modal distribution that appears normal. From reviewing the histograms for the features '#UnitsinStructure(ifmultiunit)', 'YearBuilt(parcel)' and 'AssessedImprovedValue(parcel)' we can see that there are outliers however all of these outliers appear plausible with the exception of 'YearBuilt(parcel)' which has a number of 0 values which is illogical

given that the feature details the year a structure was built. When plotting the histogram for the feature 'IncidentStartDate' due to the large number of unique values the decision was made is grouped by year. The 'IncidentStartDate' histogram appears to have a normal multi-modal distribution with a spike in 2024/2025 which aligns with the recent wildfires in California.

4.3 Box Plots

A box plot has also been completed for each continuous feature. The box plots more clearly illustrate the existence of outliers in the '#UnitsinStructure(ifmultiunit)', 'YearBuilt(parcel)' and 'AssessedImprovedValue(parcel)' features. As discussed previously, all of these outliers are plausible with the exception of 'YearBuilt(Parcel) where the '0' values will be treated as 'missing' going forward to ensure data consistency. Once again the box plot of for 'IncidentStartDate' has been grouped by year and there are no outliers present.

5. Review of Categorical Features

5.1 Descriptive Statistics

There are a total of 29 categorical features in the data set. Please note that while the feature could 'OBJECTID' could be considered a categorical feature as it is a unique identifier for the dataset it has not been included in the descriptive statistics.

- Damage (Count 1)
  - This feature has a cardinality of 2 and the value is either 'No Damage' or 'Destroyed (>50%). The feature's mode is 'Destroyed (>50%)' and it has a very low number of missing values of 0.01%.
  - Overall, this feature appears normal with a low level of missing values.

- Location Data (Count 9)
  - The following features fall under this grouping - '*StreetNumber', '*StreetName', '*StreetType(e.g.road,drive,lane,etc.)', '*City', 'ZipCode', '*CALFIREUnit', 'County', 'Community' and 'State'.
  - The main issue with several of these features including 'Community' and 'ZipCode' is that they have a large number of missing values (58.29% and 39.97% respectively). 'ZipCode' also has a large number of '0' values. The feature 'community' could be considered redundant given that we already have a feature 'SiteAddress(parcel)' which only has a 5.90% missing rate and includes the data that is present in community. As a result, the feature 'Community' will be dropped. However, the feature 'ZipCode' will remain despite its large amount of missing values because having a numeric identifier for geographic locations could be useful when it comes to predicting structural risk due to wildfires.
  - The feature 'State' has a cardinality of 1 and as a result is constant. This feature was dropped before carrying out descriptive statistics.
  - In order to ensure better data consistency NaN and '0' values should be imputed with a consistent value such as missing.

- Structural Data (Count 16)
  - The remaining categorical features outline structural details. The following features fall under this grouping 'StructureDefenseActionsTaken', '*StructureType', 'StructureCategory', '*RoofConstruction', '*Eaves', '*VentScreen', '*ExteriorSiding', '*WindowPane', '*Deck/PorchOnGrade', '*Deck/PorchElevated', '*PatioCover/CarportAttachedtoStructure', '*FenceAttachedtoStructure', 'SiteAddress(parcel)', 'Distance-PropaneTanktoStructure' and 'Distance-ResidencetoUtility/MiscStructure>120SQFT'.
  - Once again the main issue plaguing a large number of these features is that of missing values with features such as StructureDefenseActionsTaken having 72.94% missing values and other features such as Distance-PropaneTanktoStructure and Distance-ResidencetoUtility/MiscStructure&gt120SQFT having over 80% missing values. Given, the large number of missing values for these three features and the fact that they would not provide

much utility in helping us predict our target value (predicting risk for structural damage) these three features will be dropped. Given the high number of missing values imputation is not recommended.

- Several features including *RoofConstruction, *Eaves, *VentScreen, *ExteriorSiding, *WindowPane, *Deck/PorchOnGrade, *Deck/PorchElevated, *PatioCover/CarportAttachedtoStructure, *FenceAttachedtoStructure, SiteAddress(parcel), Distance-PropaneTanktoStructure have a large number of 'unknown' values but also having a large number of NaN values.
- Feature 'HazardType' has a cardinality of 1 and as a result is a constant. This feature was dropped prior to completing descriptive statistics.
- In order to ensure better data consistency throughout this dataset the imputation of a common value to represent NaN is necessary for features with less than 30% of missing values.

5.2 Bar-plots

A bar-plot has been completed for each categorical feature and can be found in section 8 of this report. Due to the large cardinality of several features the decision has been made to only display the Top 10 most frequent entries for the sake of readability. The bar plots further illustrate the prevalence of nan and 'unknown' values throughout the categorical features with features such as Distance-PropaneTanktoStructurehave having other 80% NaN values. The Bar-plots further emphasize the need for data consistency in this data set when it comes to NaN, '0' and 'unknown' values.

6. Actions to Take

The following 6 actions will be taken.

- Drop constant columns 'State' and 'HazardType'.
- Drop the only duplicate row in the dataset (OBJECTID 74364).
- Drop categorical features 'Community', 'StructureDefenseActionsTaken', 'Distance-PropaneTanktoStructure' and 'Distance-ResidencetoUtility/MiscStructure&gt120SQFT' as at least 58% of each of these features are missing and the data present in each feature is either redundant or does not provide much utility in terms of predicting structural risk prevention.
- Drop continuous feature #UnitsinStructure(ifmultiunit) as it has over 70% missing value and is not of much benefit toward are objective.
- For all categorical features the imputation of the value 'Missing' for all 'Unknown', '0' or Nan values.
- For continuous features AssessedImprovedValue(parcel) and 'YearBuilt(parcel)' the imputation of the mean value for any entries with NaN or '0' values. The imputation of the mean value will also be done for the entry that has a value of '89'.

8. Appendix

8.1 Continuous Features

Descriptive Statistics

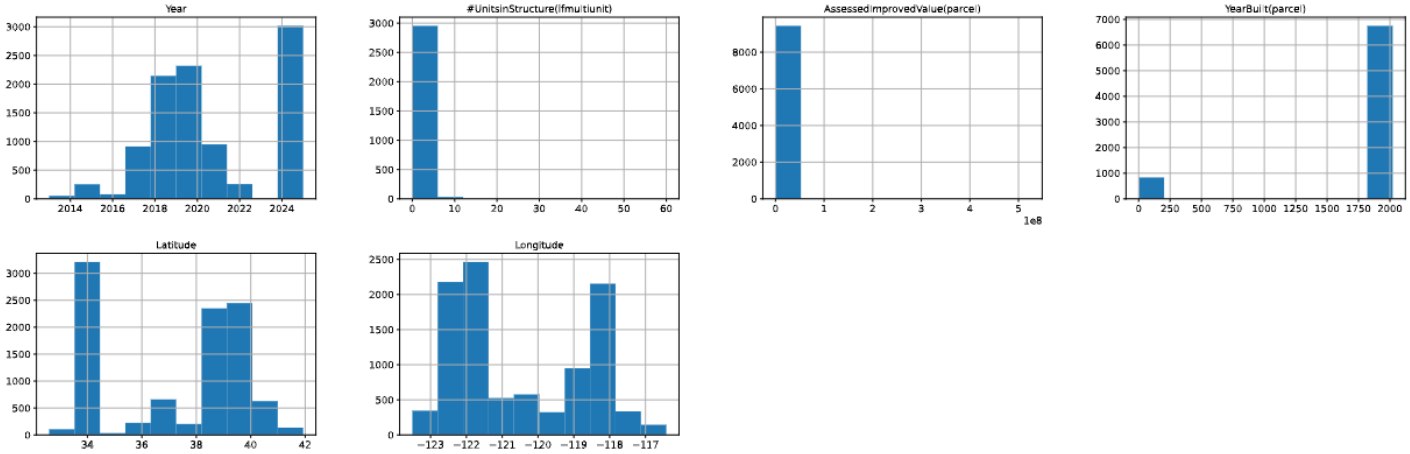
Feature	count	mean	min	25%	50%	75%	max	std	%missing	cardinality
IncidentStartDate	9999	2021-03-28 06:17:38	2013-08-07 00:00:00	2018-11-08 00:00:00	2020-09-09 00:00:00	2024-11-06 00:00:00	2025-01-07 21:48:00		0.01	146
#UnitsinStructure(ifmultiunit)	2990	0.264214046822742	0	0	0	0	60	1.59910384980223	70.1	14
AssessedImprovedValue(parcel)	9462	690735.933417882	0	68695.5	162328.5	335173.5	522652568	8790280.39074692	5.38	7621
YearBuilt(parcel)	7582	1748.74149300976	0	1939	1961	1981	2022	615.294591465377	24.18	133
Latitude	9999	37.3302559952914	32.59425401	34.19436555	38.4706091305689	39.7404104686597	41.9237421167365	2.5032077452995	0.01	9995
Longitude	9999	-120.487988295121	-123.508980408764	-122.111404867937	-121.376983649627	-118.53792425	-116.4181628	1.82508984784828	0.01	9996

8.2 Categorical Features

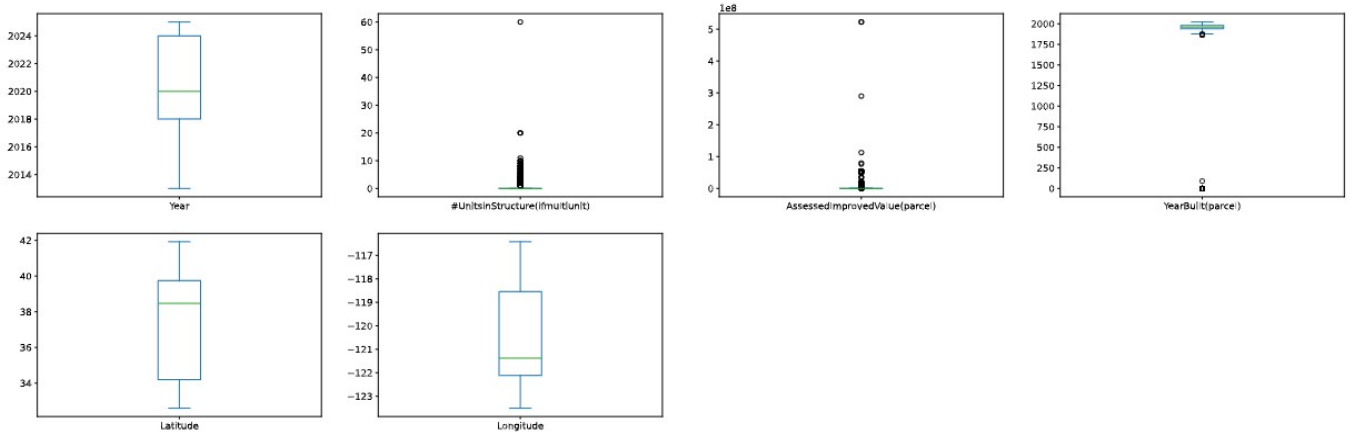
Descriptive Statistics

Feature	mode	freq_mode	%mode	2ndmode	freq_2ndmode	%2ndmode	%missing	cardinality
*Damage	Destroyed (>50%)	5692	0.5692569	No Damage	4307	0.4307430	7.01	2
*StreetNumber	0	545	0.0563774	580	19	0.0019654	5.33	5621
*StreetName	Pacific Coast	76	0.0079357	Pentz	52	0.0054296	8.4.23	3976
*StreetType(e.g.road,drive,lane,etc.)	Road	3427	0.3863585	Drive	1568	0.1767756	5 11.3	18
*City	Unincorporated	1169	0.154405	Altadena	1141	0.1507066	4 24.29	257
ZipCode	0	1853	0.308679	91001	1123	0.1870731	3 39.97	148
*CALFIREUnit	LAC	2642	0.2642264	BTU	2210	0.2210221	0.01	27
County	Los Angeles	2642	0.2643057	Butte	2206	0.2206882	8 0.04	45
Community	Paradise	579	0.1388156	Paradise	578	0.1385758	8 58.29	413
StructureDefenseActionsTaken	Unknown	2127	0.786031	Engine Company Actions	247	0.0912786	4 72.94	10
*StructureType	Single Family Residence Single Story	3521	0.3521352	Utility Misc Structure	2766	0.2766276	6 0.01	17
StructureCategory	Single Residence	6616	0.6616662	Other Minor Structure	2771	0.2771277	1 0.01	6
*RoofConstruction	Asphalt	4519	0.4685815	Metal	1568	0.1625881	4 3.56	9
*Eaves	Unknown	4051	0.4217156	Unenclosed	3304	0.3439517	3.94	5
*VentScreen	Unknown	3091	0.3217446	Mesh Screen > 1/8""	2660	0.2768814	4 3.93	6
*ExteriorSiding	Stucco Brick Cement	2701	0.2809152	Wood	2255	0.2345293	8 3.85	10
*WindowPane	Multi Pane	3264	0.3391874	Single Pane	2912	0.3026083	3 3.77	4
*Deck/PorchOnGrade	Masonry/Concrete	3067	0.3655542	No Deck/Porch	2993	0.3567342	1 6.1	5
*Deck/PorchElevated	No Deck/Porch	4762	0.5675805	Wood	1587	0.1891537	5 16.1	5
*PatioCover/CarportAttachedtoStructure	No Patio Cover/Carport	4383	0.5224076	Combustible	2023	0.2411203	8 16.1	4
*FenceAttachedtoStructure	No Fence	4231	0.5042908	Non Combustible	1779	0.2120381	4 16.1	4
SiteAddress(parcel)	No Address Available	45	0.004772	LAKESHORE CA 93634	45	0.004772	5.7	8527
Distance-PropaneTanktoStructure	Unknown	531	0.2917582	>30	467	0.2565934	1 81.8	5
Distance-ResidencetoUtility/MiscStructure&gt120SQFT	<30'	811	0.5395875	>50'	365	0.2428476	4 84.97	5

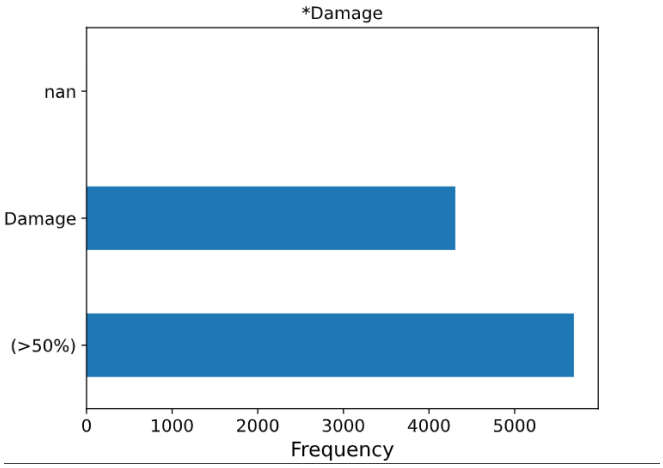
8.3 Histograms of Continuous Features (larger versions can be found in the accompanying notebook).

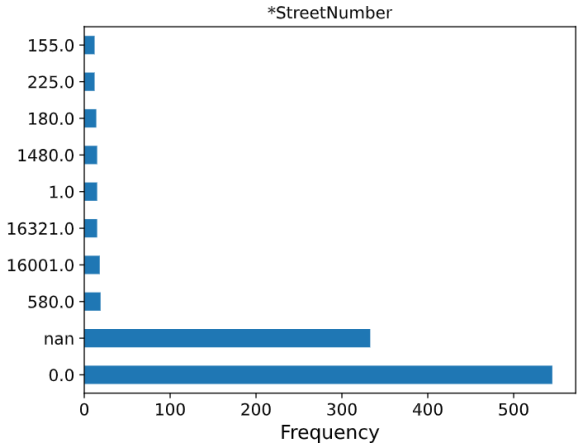


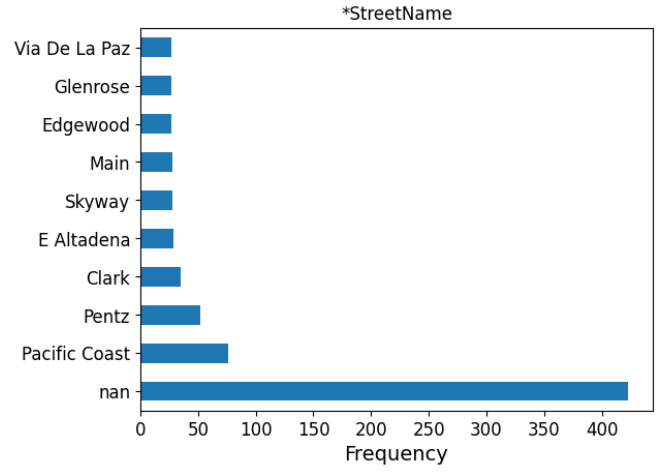
8.4 Box-plots of Continuous Features (larger versions can be found in the accompanying notebook).

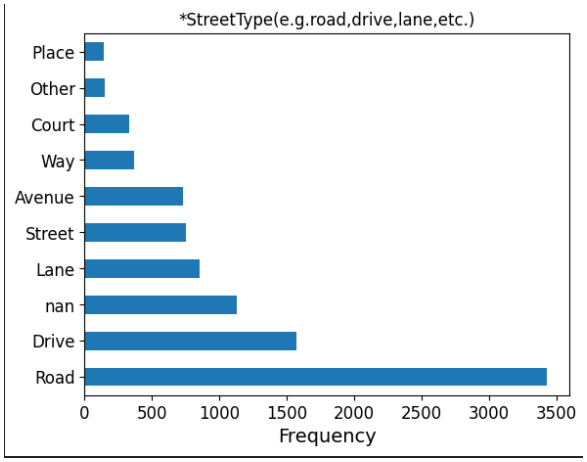


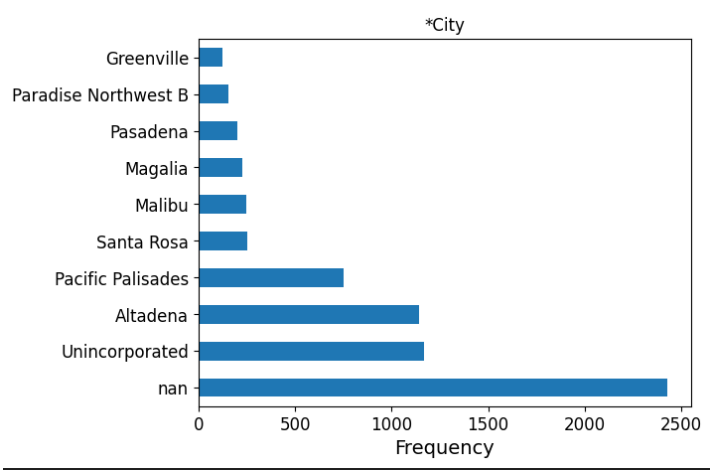
8.5 Bar-plots of Categorical Features (larger versions can be found in the accompanying notebook).

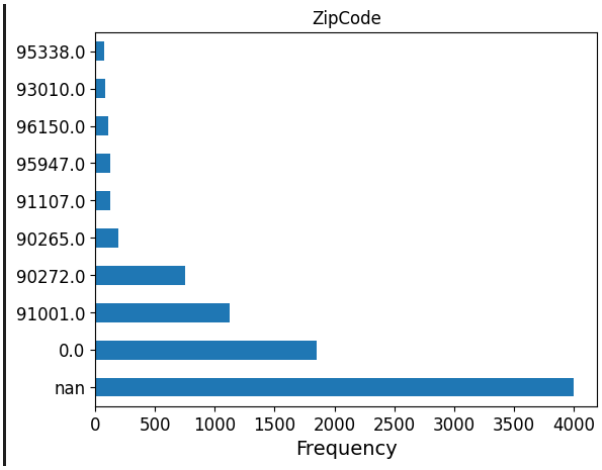


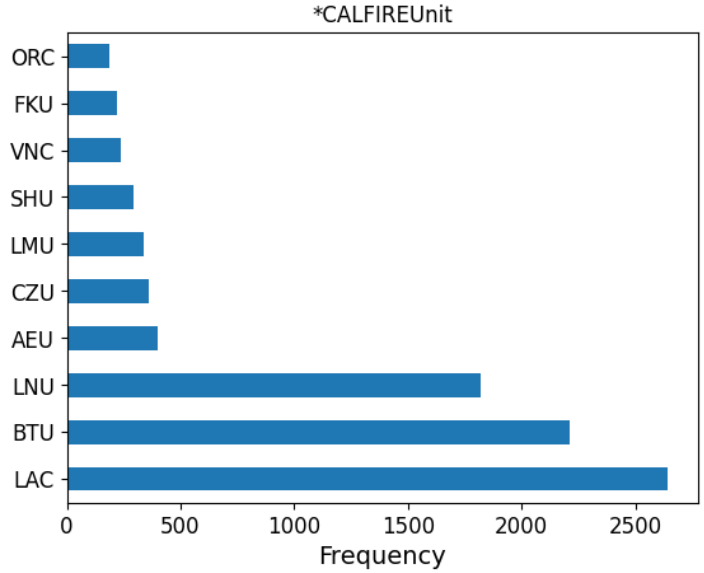


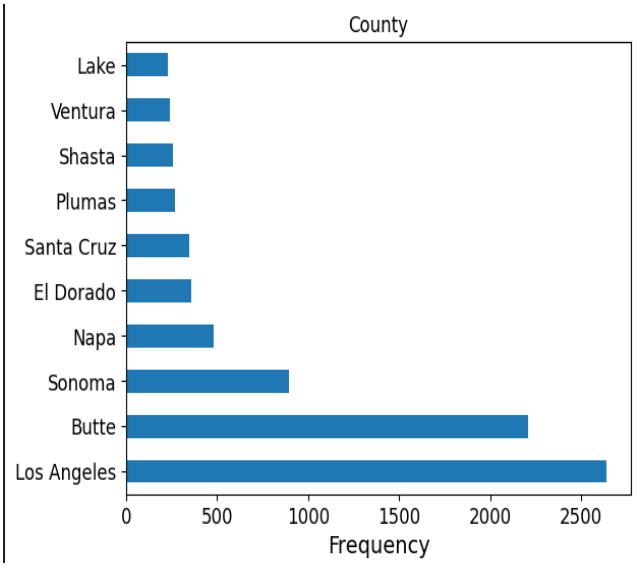


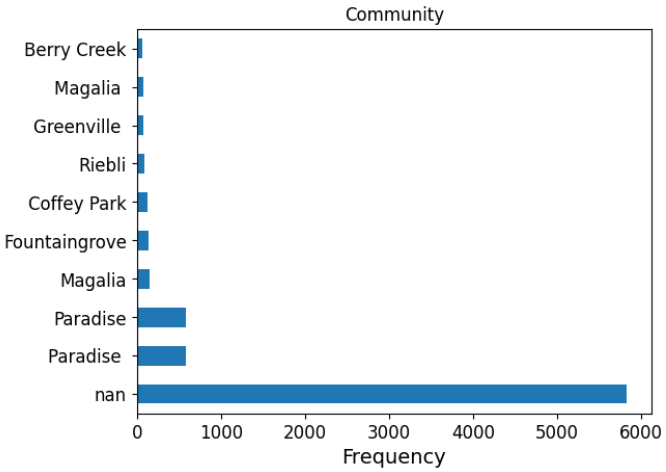


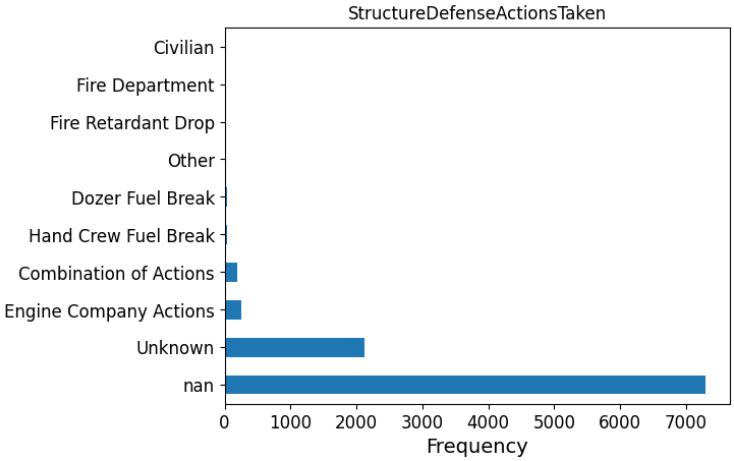

*City
Greenville
Paradise Northwest B
Pasadena
Magalia
Malibu
Santa Rosa
Pacific Palisades
Altadena
Unincorporated
nan
0
500
1000
1500
2000
2500
Frequency

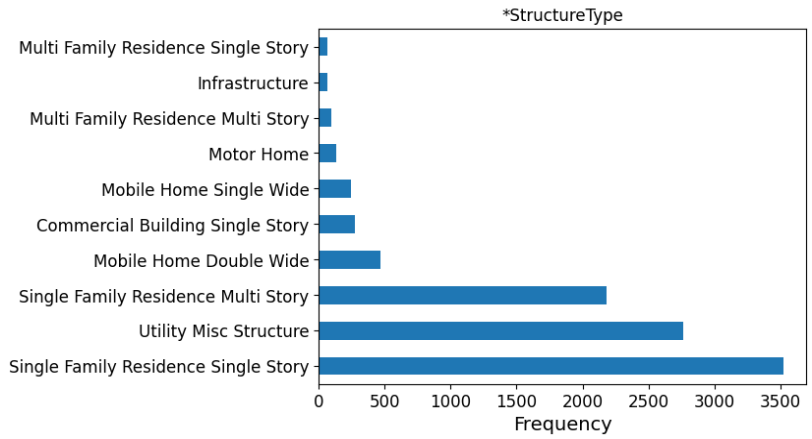

ZipCode
95338.0
93010.0
96150.0
95947.0
91107.0
90265.0
90272.0
91001.0
0.0
nan
0
500
1000
1500
2000
2500
3000
3500
4000
Frequency

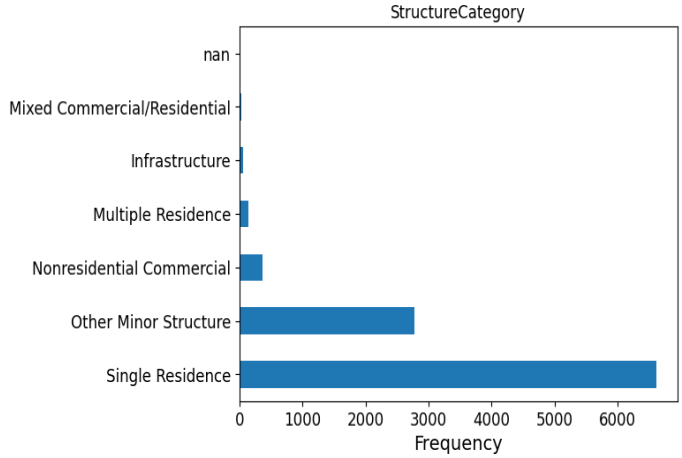

*CALFIREUnit
ORC
FKU
VNC
SHU
LMU
CZU
AEU
LNU
BTU
LAC
0
500
1000
1500
2000
2500
Frequency

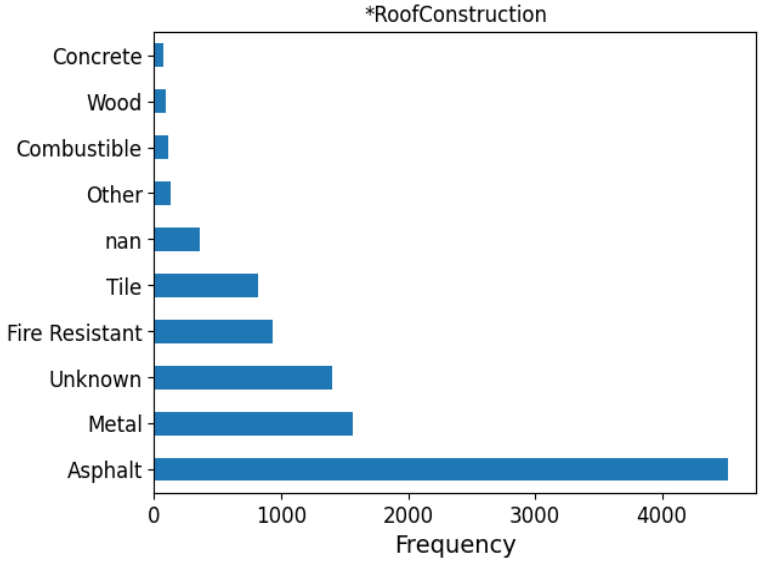

County
Lake
Ventura
Shasta
Plumas
Santa Cruz
El Dorado
Napa
Sonoma
Butte
Los Angeles
0
500
1000
1500
2000
2500
Frequency

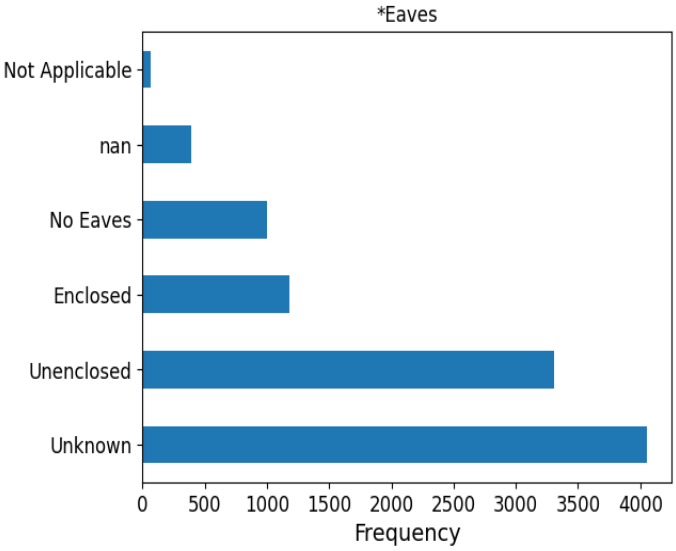

Community
Berry Creek
Magalia
Greenville
Riebli
Coffey Park
Fountaingrove
Magalia
Paradise
Paradise
nan
0
1000
2000
3000
4000
5000
6000
Frequency

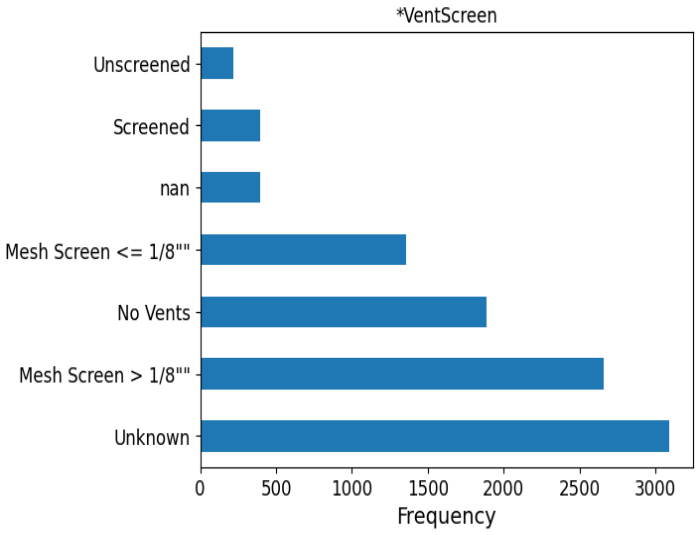

StructureDefenseActionsTaken
Civilian
Fire Department
Fire Retardant Drop
Other
Dozer Fuel Break
Hand Crew Fuel Break
Combination of Actions
Engine Company Actions
Unknown
nan
0
1000
2000
3000
4000
5000
6000
7000
Frequency

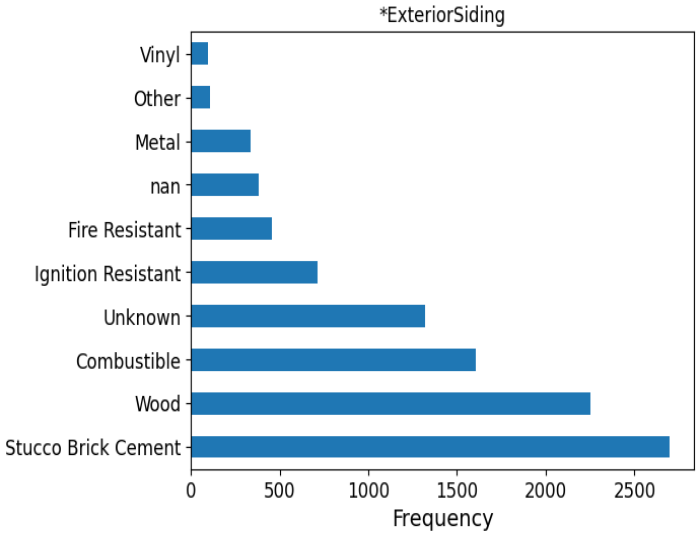

*StructureType
Multi Family Residence Single Story
Infrastructure
Multi Family Residence Multi Story
Motor Home
Mobile Home Single Wide
Commercial Building Single Story
Mobile Home Double Wide
Single Family Residence Multi Story
Utility Misc Structure
Single Family Residence Single Story
0
500
1000
1500
2000
2500
3000
3500
Frequency

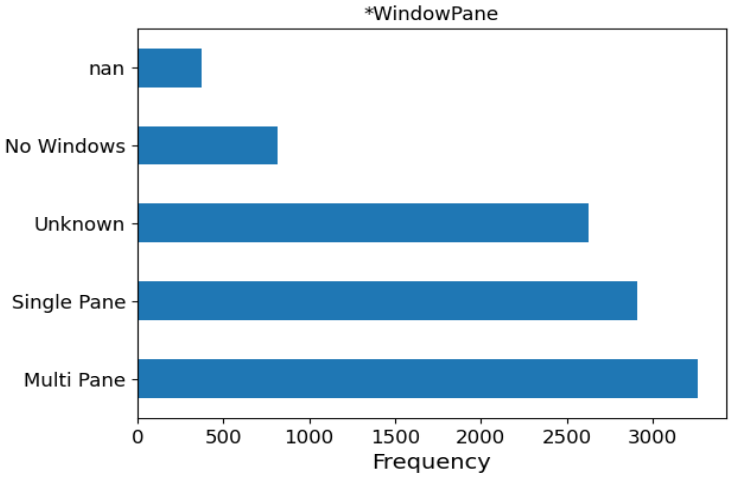

StructureCategory
nan
Mixed Commercial/Residential
Infrastructure
Multiple Residence
Nonresidential Commercial
Other Minor Structure
Single Residence
0
1000
2000
3000
4000
5000
6000
Frequency

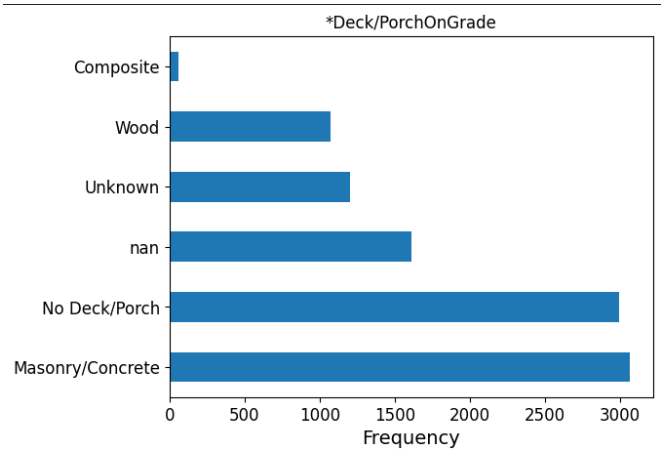

*RoofConstruction
Concrete
Wood
Combustible
Other
nan
Tile
Fire Resistant
Unknown
Metal
Asphalt
0
1000
2000
3000
4000
Frequency

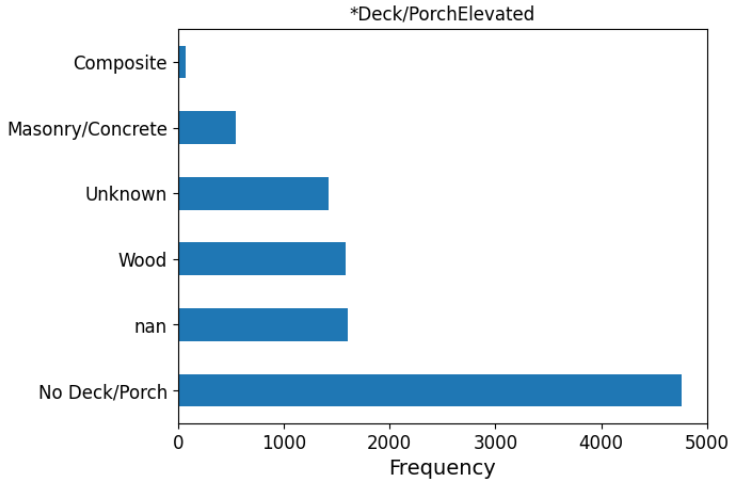

*Eaves
Not Applicable
nan
No Eaves
Enclosed
Unenclosed
Unknown
0
500
1000
1500
2000
2500
3000
3500
4000
Frequency

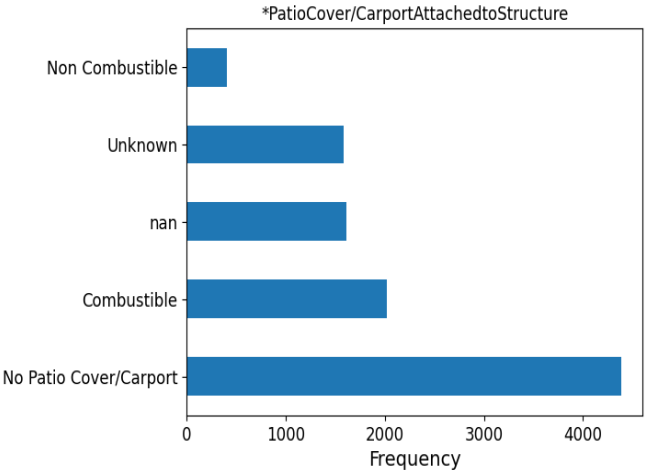

*VentScreen
Unscreened
Screened
nan
Mesh Screen <= 1/8""
No Vents
Mesh Screen > 1/8""
Unknown
0
500
1000
1500
2000
2500
3000
Frequency

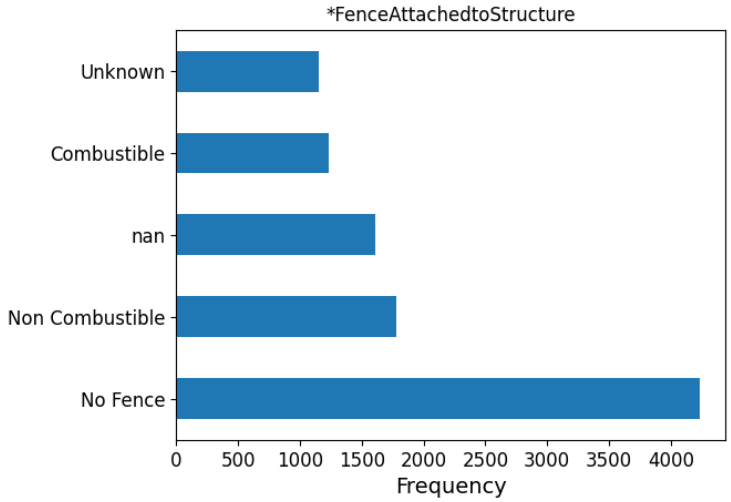

*ExteriorSiding
Vinyl
Other
Metal
nan
Fire Resistant
Ignition Resistant
Unknown
Combustible
Wood
Stucco Brick Cement
0
500
1000
1500
2000
2500
Frequency

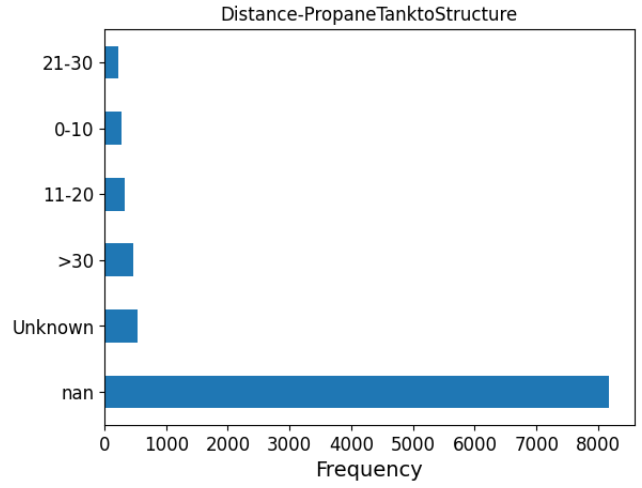

*WindowPane
nan
No Windows
Unknown
Single Pane
Multi Pane
0
500
1000
1500
2000
2500
3000
Frequency

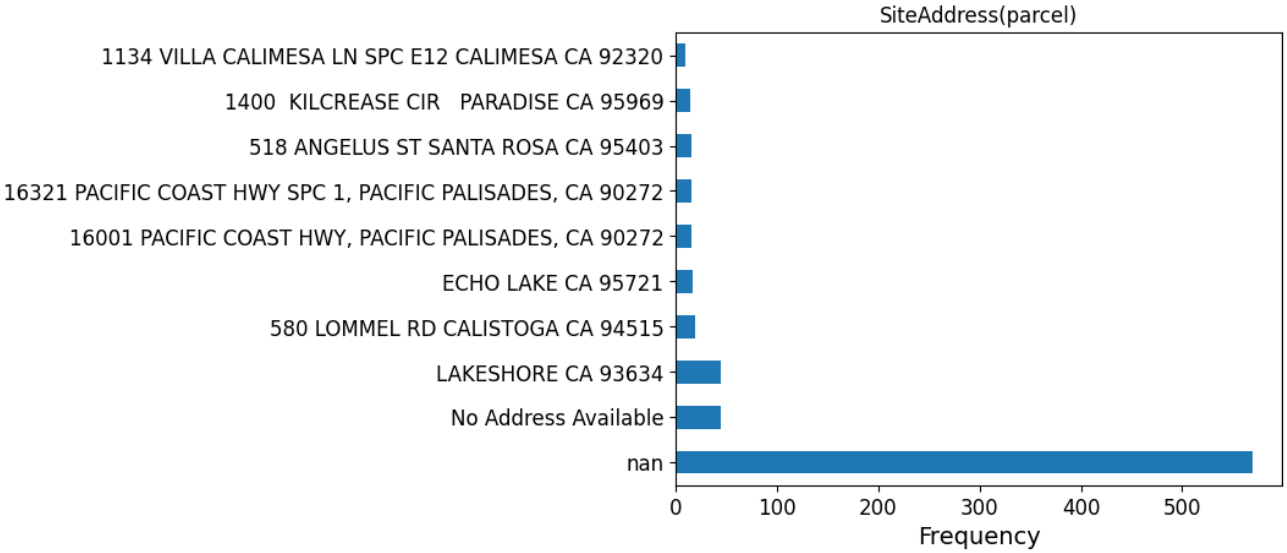

*Deck/PorchOnGrade
Composite
Wood
Unknown
nan
No Deck/Porch
Masonry/Concrete
0
500
1000
1500
2000
2500
3000
Frequency

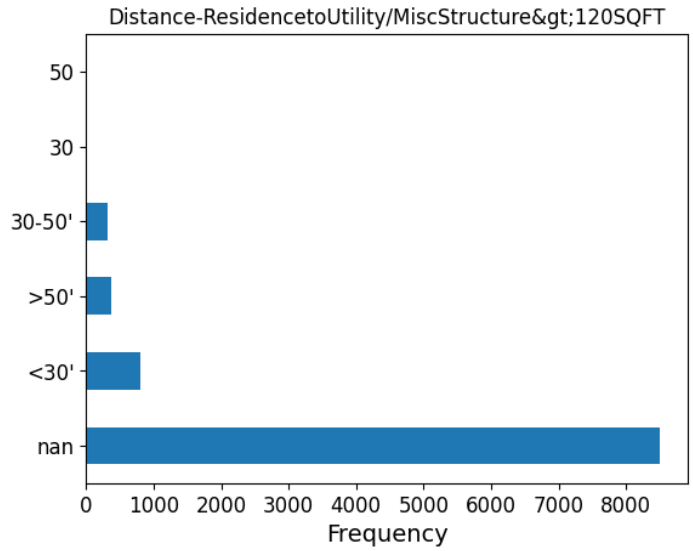

*Deck/PorchElevated
Composite
Masonry/Concrete
Unknown
Wood
nan
No Deck/Porch
0
1000
2000
3000
4000
5000
Frequency


*PatioCover/CarportAttachedtoStructure
Non Combustible
Unknown
nan
Combustible
No Patio Cover/Carport
0
1000
2000
3000
4000
Frequency


*FenceAttachedtoStructure
Unknown
Combustible
nan
Non Combustible
No Fence
0
500
1000
1500
2000
2500
3000
3500
4000
Frequency


Distance-PropaneTanktoStructure
21-30
0-10
11-20
>30
Unknown
nan
0
1000
2000
3000
4000
5000
6000
7000
8000
Frequency


SiteAddress(parcel)
1134 VILLA CALIMESA LN SPC E12 CALIMESA CA 92320
1400 KILCREASE CIR PARADISE CA 95969
518 ANGELUS ST SANTA ROSA CA 95403
16321 PACIFIC COAST HWY SPC 1, PACIFIC PALISADES, CA 90272
16001 PACIFIC COAST HWY, PACIFIC PALISADES, CA 90272
ECHO LAKE CA 95721
580 LOMMEL RD CALISTOGA CA 94515
LAKESHORE CA 93634
No Address Available
nan
0
100
200
300
400
500
Frequency


Distance-ResidencetoUtility/MiscStructure>120SQFT
50
30
30-50'
>50'
<30'
nan
0
1000
2000
3000
4000
5000
6000
7000
8000
Frequency