Remark that the minimum of 'Vertical_Distance_To_Hydrology' is negative. I suppose that it is normal because some area can be lower than the hydrology source. But since it is negative value, we have to notice that we w ill not be able to use filter method with chi-sqaured to perform the dimensional reduction.

Regroup features

This will serve for future study and implementation since our features are composed of binary and continuous features. For example, if it is neccesary we will have to create new features using linear combination of con tinuous features or we will have to do reverse one-hot encoding for binary features.

In [7]:
```
continuous_features=train_data.loc[:,'Elevation':'Horizontal_Distance_To_Fire_Points']
binary_features=train_data.loc[:,'Wilderness_Area1':'Soil_Type40']
wilderness_features=train_data.loc[:,'Wilderness_Area1': 'Wilderness_Area4']
soil_features=train_data.loc[:,'Soil_Type1':'Soil_Type40']
```

Check for Anomalies & Outliers

Extreme Outliers

Extreme outliers are data points that are significantly different from the rest of the data in a dataset. They can have a significant impact on the statistical properties of the data and can affect the performance of machine learning models. The extreme outliers is the points where

- x < Q1 - 3 * IQR
- x > Q3 + 3 * IQR

where IQR = third_quartile - first_quartile

In [8]:
```
def outlier_function(df, col_name):
    first_quartile = np.percentile(np.array(df[col_name].tolist()), 25)
    third_quartile = np.percentile(np.array(df[col_name].tolist()), 75)
    IQR = third_quartile - first_quartile

    upper_limit = third_quartile+(3*IQR)
    lower_limit = first_quartile-(3*IQR)
    outlier_count = 0

    for value in df[col_name].tolist():
```