# Outline:

Results:

Showing results and interpretation of clustering.

Data gathering:

Process that was taken to get data and initial cleaning.

Data Processing

Process that was done to clean, aggregate, and transform data for modeling.

Modeling:

Process that was done to cluster data.

# Results:

Cluster 0:

On average models take up the most time of the analyst, require the most model maintenance, and tend to be APR models.

* Models alert more often.
* Touched more often.
* Have higher save and build frequencies.
* Lest deactivate models

Cluster 1:

On average models take up the least time of the analyst, almost never alert, to be Fixed limit and Rolling average.

* have the lowest build and save frequency.
* Have the most deactivate models.
* Have the lowest model scores.
* almost never touched.
* Equipment with the fewest number of cases
* On average models are on equipment with the least case frequency

Cluster 2:

Over 50% of all models. Models tend to be in between cluster 0 and 1.

* Tend to have almost as many APRs as cluster zero.
* Some fixed limit and rolling average.
* On average models are on equipment with the most case frequency

Chart, bar chart

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Decription | Cluster 0 | Cluster 1 | Cluster 2 |
| Upper Vs Actaul Perc diff | Lowest | Highest | Middle |
| Lower Vs Actual | Lowest | Highest | Middle |
| Currently False | Lowest | Highest | Middle |
| Model Age | Highest | Lowest | Middle |
| Model Build Freq | Highest | Lowest | Middle |
| Model Save Freq | Highest | Lowest | Middle |
| No Alert Freq | Lowest | Highest (see box plot) | Middle |
| Touch Freq | Highest | Lowest (zero) | middle |
| Model Score | Highest | Lowest | Middle |
| Model Type Rolling Avg | Lowest | Highest | Middle |
| Model Type Fixed Limit | Lowest | Highest | Middle |
| Model Type APR | Highest | Lowest | Middle |
| OP-P | Middle | Highest | Lowest |
| MP-Pump | Lowest | Highest | Middle |
| MP-Fan\_Turbn Grbx\_other | Highest | Middle | Lowest |
| MP-Comp | Middle | Highest | Lowest |
| IP-Valve | Highest | Lowest | Middle |
| IP-Loops\_xmtr\_other | Highest | Middle | Lowest |
| FP-Tank-Vessel | Lowest | Highest | Middle |
| FP-Fired | Highest By far | Middle | Lowest |
| FP-Exchanger | About same as 1 | About same as 0 | Lowest |
| FP-Comp\_valve\_pipe\_other | Highest By far | Lowest | Middle |
| EP | Lowest | Highest | Lowest |

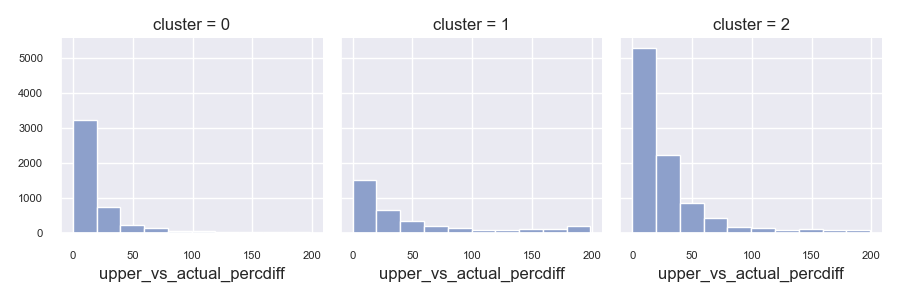
UPPER VS ACTUAL

Chart, box and whisker chart

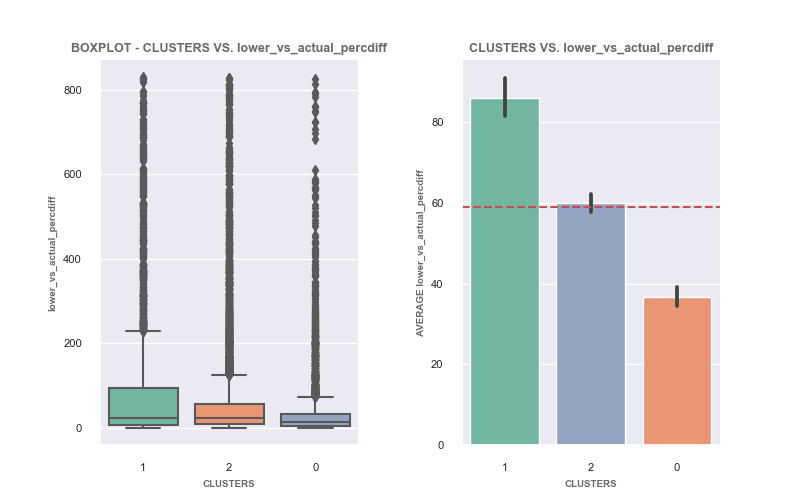
Description automatically generated

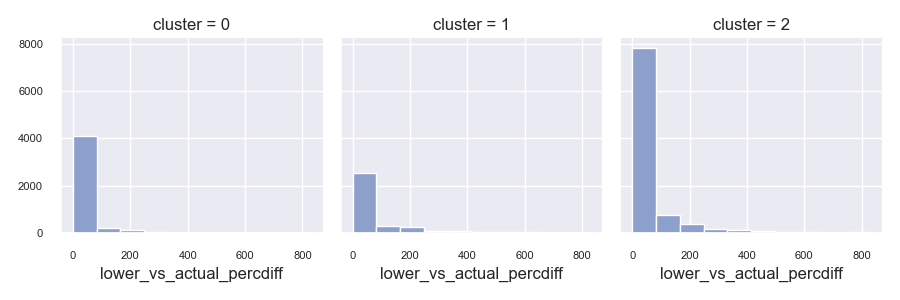
Chart

Description automatically generated



LOWER VS ACTUAL:





Chart

Description automatically generated

CURRENTLY NOT ACTIVE

Chart, bar chart

Description automatically generated

MODEL AGE:

Chart, box and whisker chart

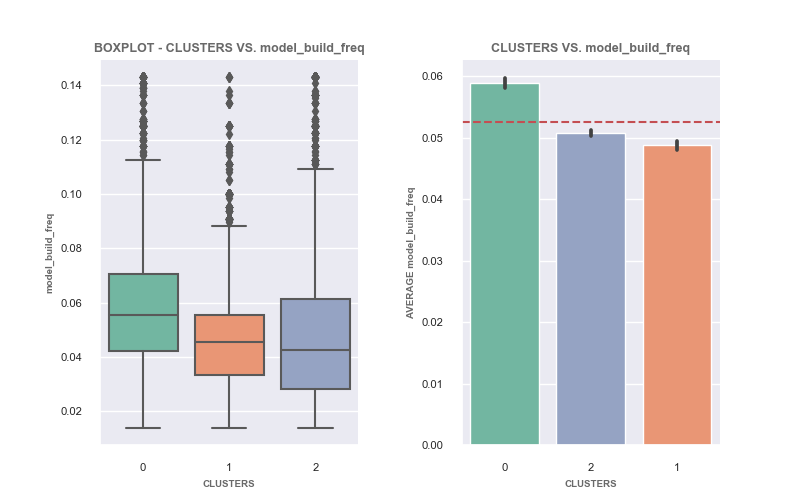
Description automatically generated



Chart

Description automatically generated

MODEL BUILD FREQ



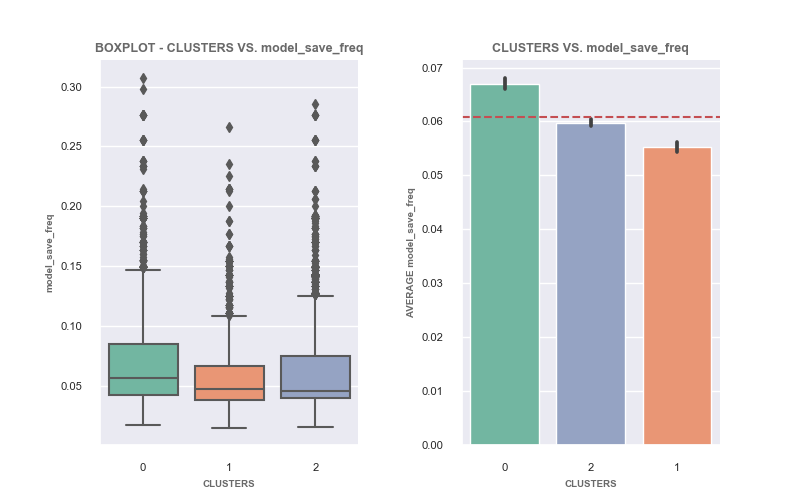
Chart, histogram

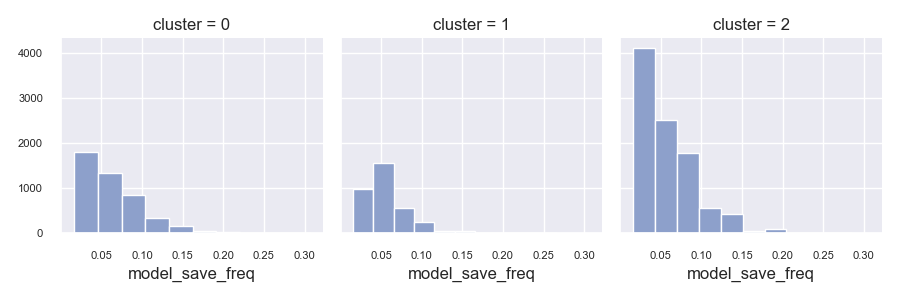
Description automatically generated

Chart

Description automatically generated

MODEL SAVE FREQ





Chart, bar chart

Description automatically generated

NO ALERT FREQ

Chart, box and whisker chart

Description automatically generated

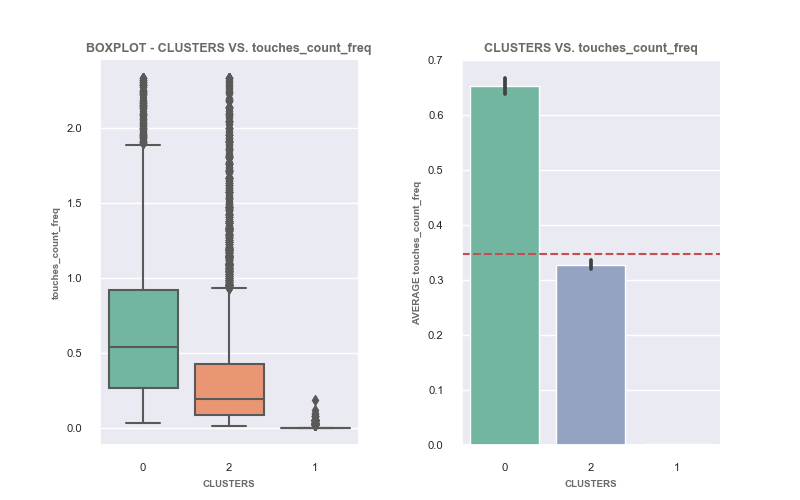
Chart

Description automatically generated

Chart, histogram

Description automatically generated

MODEL TOUCHES



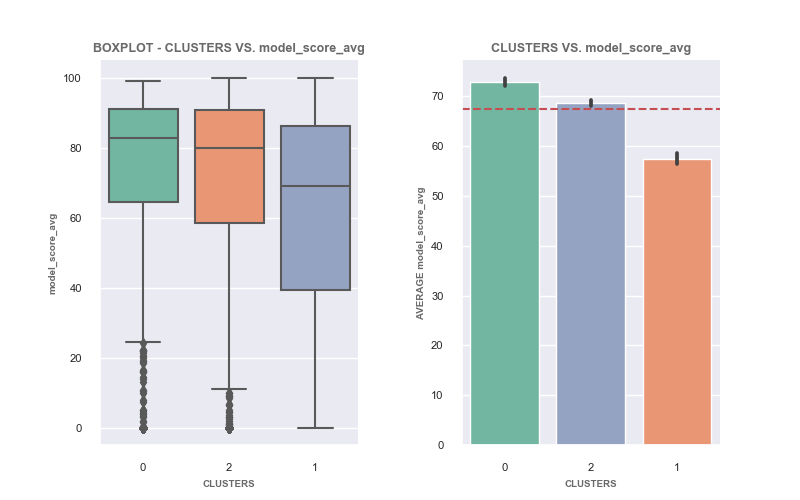
Chart

Description automatically generated

Table

Description automatically generated with medium confidence

Model Score:

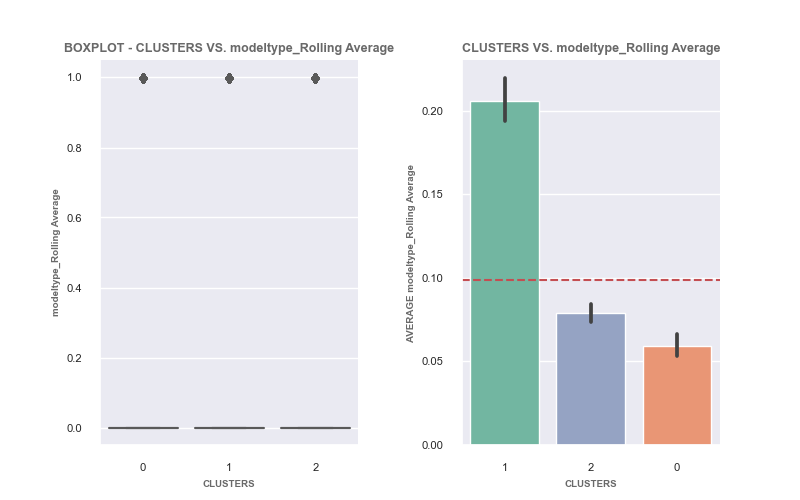


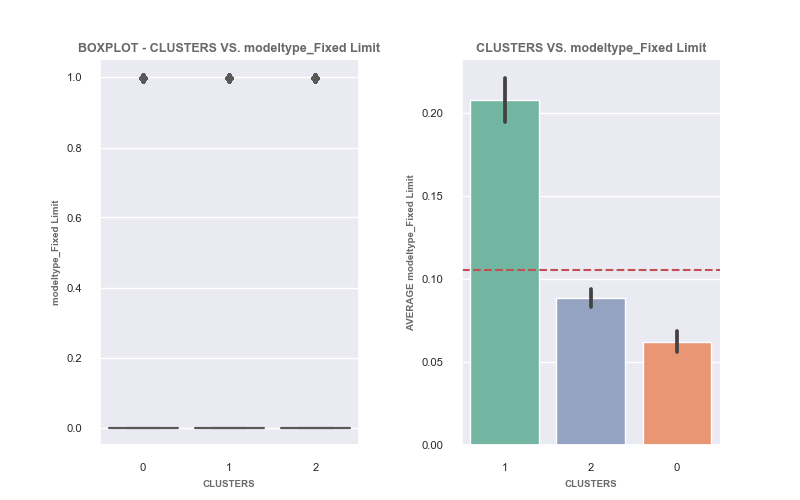


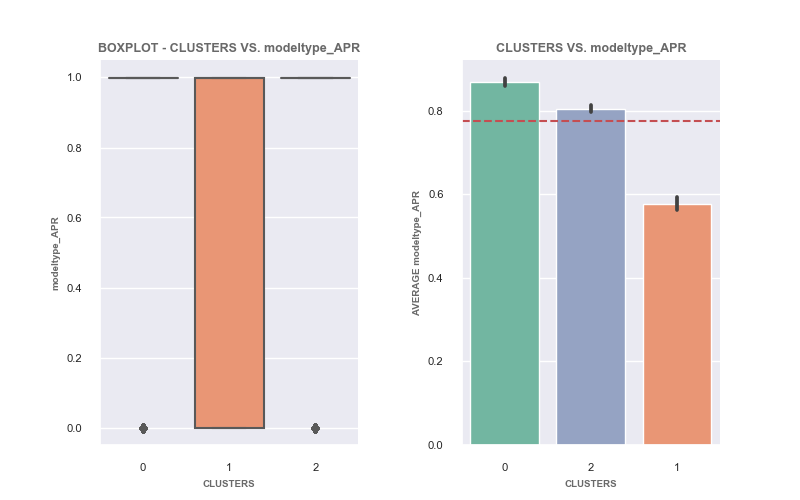
Chart, bar chart

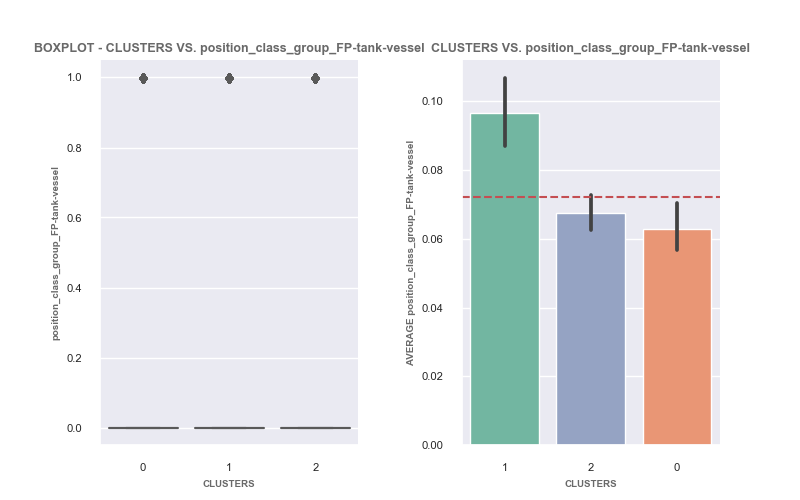
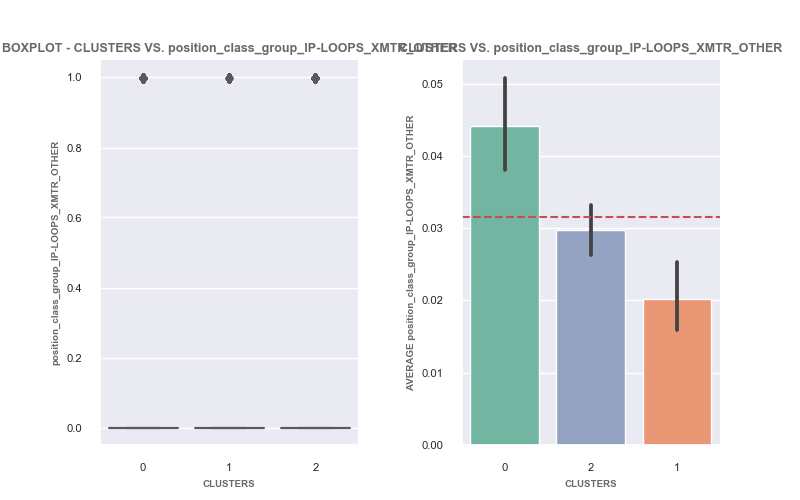
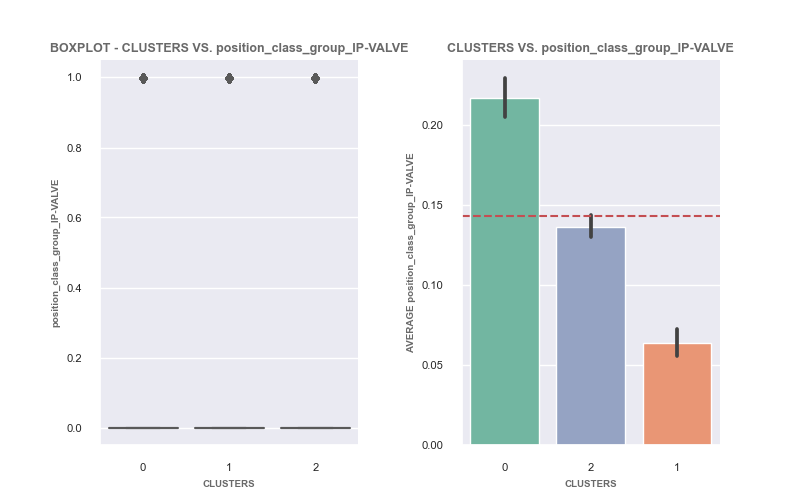
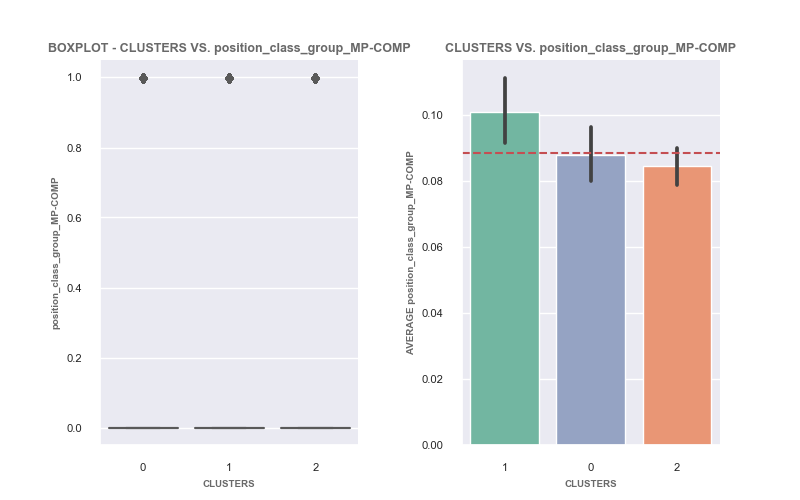
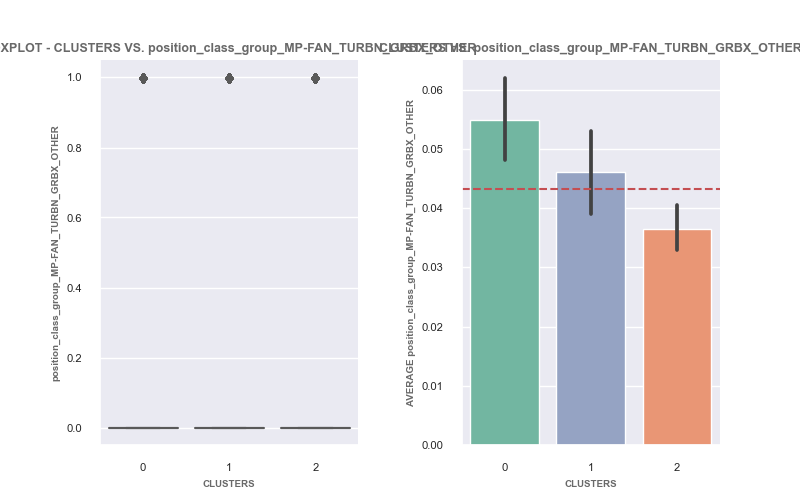
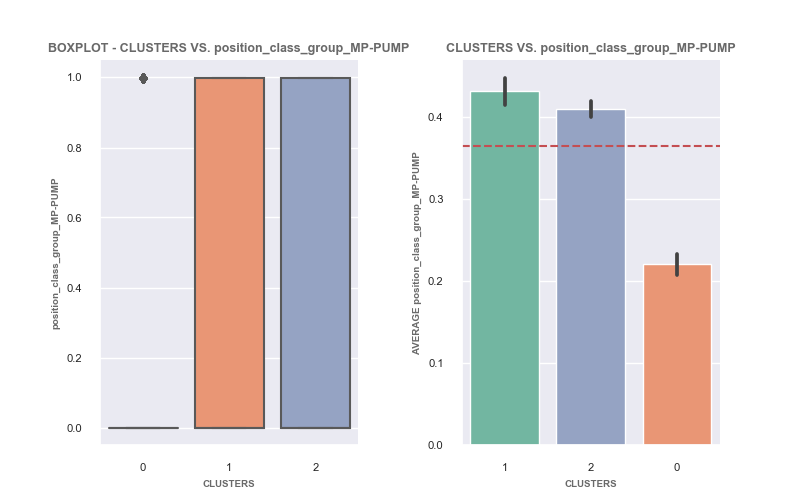
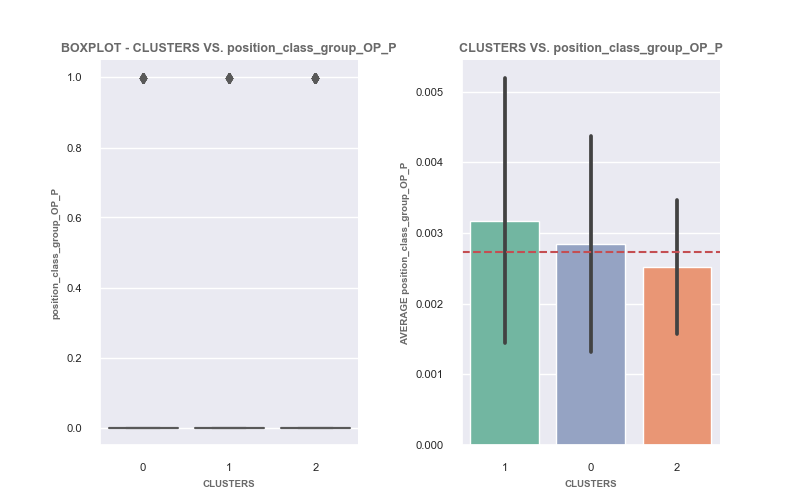
Description automatically generated

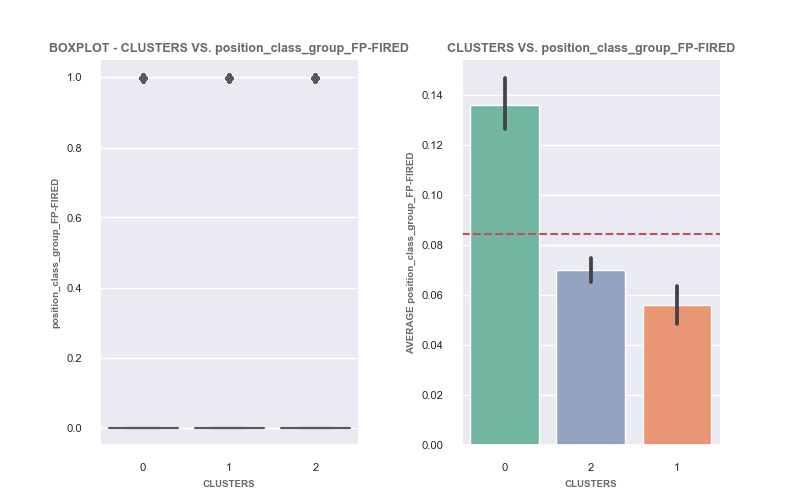
OTHER GENERAL:

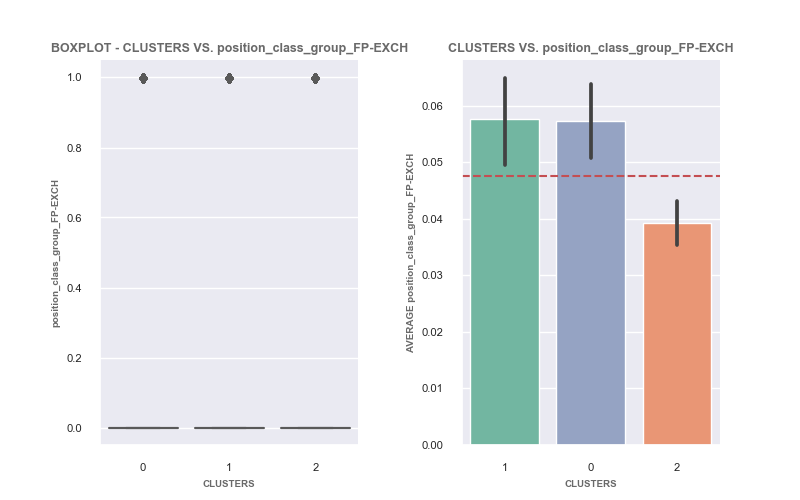


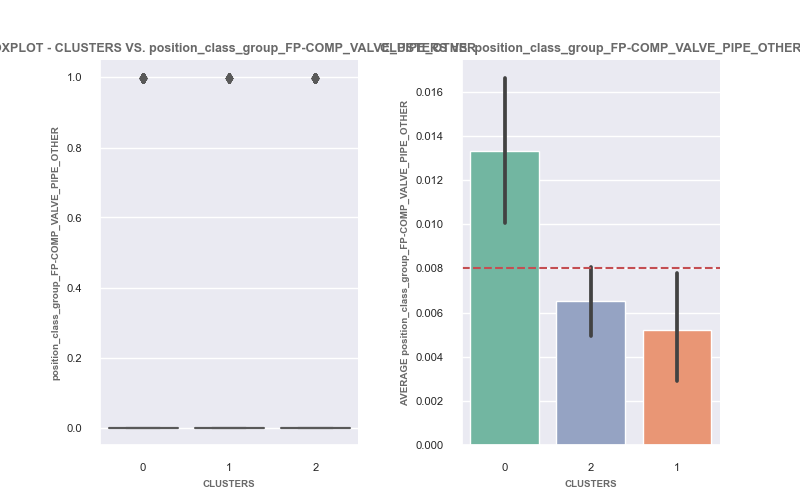


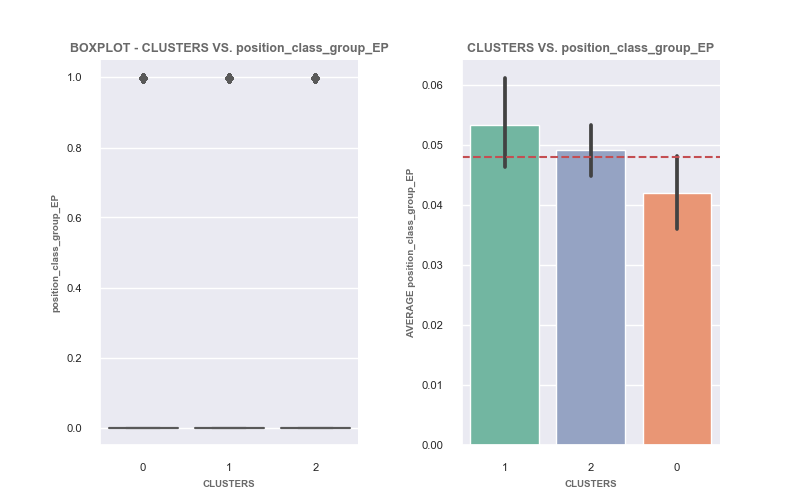


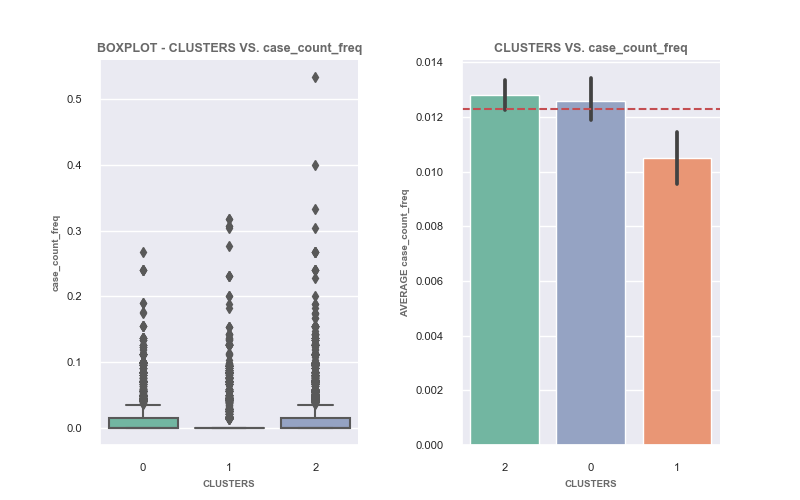


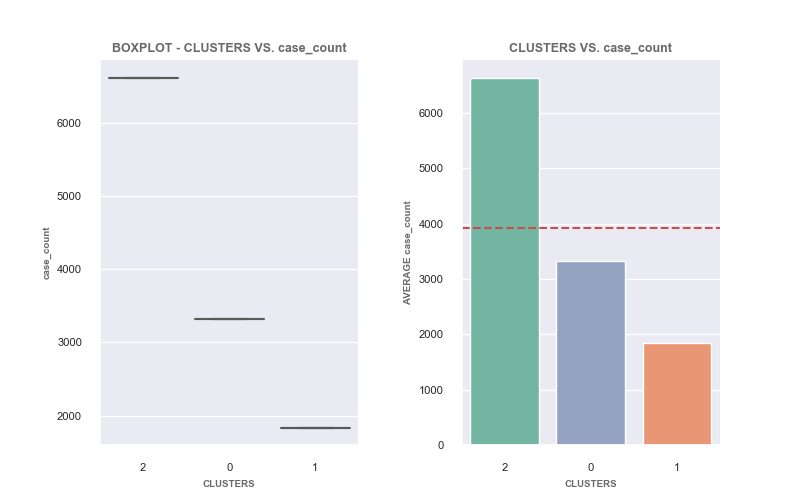












# Data gathering:

Data Pull from Denodo Data Lake

fv\_asset360\_models\_compress

Weekly data of model info data



fv\_asset360\_modelstate\_compress

Weekly atta of model values (expected, actual, etc), also alerts



fv\_asset360\_modelactions\_compress

Each ingestion is all actions for a model (ie. Touches)



eam\_casemanagement\_header

Cases written



eam\_position\_record\_view

Equipment types

fv\_asset360\_assets\_compress\_current

Needed to link model ID to asset ID to get site (ie, CC, PB, or pipelines)

 Issue with model active:

sets 'active' field in table. this value will be used for all other tables to know when a model was active or not.

active status will == false for all past records if model is not currently active.

case 1: current ingestion\_ts active == true

action make all entries active == true

case 2: current ingestion\_ts active == false

2.1: there is an active == true before any null values

action: make all entries after the active == true ==> active == true

make all entries before the active == true ==> active == false

2.2: there are no active == true before any null values

we know that the model is currently deactivate or deleted, there must be a save in order to make a deactivate or deleted,

therefore last save date is point where model was deactivate or deleted

action find last save point and drop any rows after that ingestion point

to be able to tell which models are not active anymore (ie. they were removed for some reason) set all active == to "current\_active\_value"

this may be a helpful metric to identify patterns of low quality models

REMOVED MODELS WITH NO BUILD STATUS:

Looking at the rows with no lastbuildtime and comparing them to model state table, it looks these are not good models, most the values for actual / expected ..etc are 0. the belief is that they are models that were added but had no data to backfill and were never fully trained

action: remove model IDs that have no build status

ADDED Position ID, Equipment Type, site, columns

A lot of work searching and filtering to get these values

Required to get cases linked to models

Required to get equipment type

Remove rows from model state when model was not active

This was needed to get accurate model age

Fix Model types that changed

#it was found that the model type may change over time (ie. starts as a fixed limit, then at some point changes to APR).

#the consequence of this is a model score or other frequencies will not be accurate represented due to age being incorrect.

# 1333 models have changed types, for now remove these rows. but in future create function to find change date and add as new model

# Data Processing

Data is currently in the form of many samples of the same models at different periods of time Aggregating in datal lake could of been done but with lack of user rights to data lake it is easier to do it here.

Model info Table:

Sum: Model save count and Model build count

Averaged respective model scores into one column 'model\_score\_avg':

Fixedlimitscore, gffscore, linregscore, logregscore, rollingaveragescore

Position Class :

There were 32 different positions

'MP-PUMP', 'MP-COMP', 'FP-TANK', 'FP-VESSL', 'EP-MTR', 'FP-FIRED',

'IP-VALVE', 'IP-XMTR', 'FP-COMP', 'FP-EXCH', nan, 'MP-FAN',

'IP-CNTRL', 'MP-TRBN', 'IP-INDCT', 'IP-LOOPS', 'EP-HTR', 'FP-PIPE',

'MP-GRBX', 'MP-OTHER', 'FP-COLL', 'FP-OTHER', 'EP-GEN', 'IP-ANALZ',

'FP-VALVE', 'P-CC', 'OP-BLDME', 'MP-ENGN', 'IP-SGNL', 'P-UNIT',

'OP-OTHER', 'IP-ELMNT'

There were 7 different parent positions with wide range of models in each class

Text

Description automatically generated with medium confidence

To try and even out number of models in each class but still have specific insights

position class was Split Position class into 12 categories

Text

Description automatically generated

Final Columns:

'age\_inweeks', 'modeltype', 'site', 'currently\_active\_False',

'currently\_active\_True', 'model\_save\_count', 'model\_build\_count',

'model\_score\_avg', 'position\_class\_group

Model Actions:

Average model values for 'actual', 'expected', 'upper', and 'lower':

Did not end up using these values because not every week these values are recorded like they are with Model stats, they are the recording at the time the action was done. Using these values could lead to sample bias for models that are regular touched over models not touched.

LENGTH OF NOTE: get average length of notes for each action on each model ID

ACTION TYPE COUNT: get count of each type of actiontype for each Model ID.

Final Columns:

'actual', 'expected', 'upper', 'lower', 'action\_note\_len\_avg',

'actiontype\_Clear Alert Status', 'actiontype\_Diagnose Cleared',

'actiontype\_Diagnose Set', 'actiontype\_Ignore Expiration',

'actiontype\_Ignore Set', 'actiontype\_Model Maintenance Cleared',

'actiontype\_Model Maintenance Set', 'actiontype\_Note Added',

'actiontype\_Quick Watch Set', 'actiontype\_Stop Ignoring',

'actiontype\_Watch Cleared', 'actiontype\_Watch Expiration',

'actiontype\_Watch Override', 'actiontype\_Watch Set', 'touches\_count'

Model States:

Sum ALERT TYPES:

alerts are doubled up, ie on cell may have 'AAS,HHA' and another may have 'HHA'

needed to split each row and add columns

sum up all each alert type for each model ID

Calculate model age. Each row is a week (samples taken at 1 week interval)

Average VALUES: convert values to percent difference more important is how model is predicting relative to self(ie. actual vs expected, upper vs actual, and lower vs actual)

Create in\_alert column: sum of all alerts for each model

Final columns:

'AAF', 'HHA', 'FRQ', 'noAlert', 'WOR', 'OSC', 'LLA', 'AAS', 'model\_age',

'in\_alert', 'actual\_vs\_expected\_percdiff', 'upper\_vs\_actual\_percdiff',

'lower\_vs\_actual\_percdiff'

Model Cases:

group by case ID to get count of each equipment case

map equipment ID back to final\_info DF to get case count

Final columns:

Case\_count

Then merged Model info, Model States , Model Actions, and Model Cases and calculated frequency using model states “model Age”

Final Columns:

'modeltype', 'site', 'currently\_active\_False', 'currently\_active\_True',

'model\_save\_freq', 'model\_build\_freq', 'model\_score\_avg',

'position\_class\_group', 'LLA\_freq', 'AAS\_freq', 'AAF\_freq',

'noAlert\_freq', 'WOR\_freq', 'FRQ\_freq', 'HHA\_freq', 'OSC\_freq',

'model\_age', 'in\_alert\_freq', 'actual\_vs\_expected\_percdiff',

'upper\_vs\_actual\_percdiff', 'lower\_vs\_actual\_percdiff', 'case\_count',

'action\_note\_len\_avg', 'actiontype\_Clear Alert Status\_freq',

'actiontype\_Diagnose Cleared\_freq', 'actiontype\_Diagnose Set\_freq',

'actiontype\_Ignore Expiration\_freq', 'actiontype\_Ignore Set\_freq',

'actiontype\_Model Maintenance Cleared\_freq',

'actiontype\_Model Maintenance Set\_freq', 'actiontype\_Note Added\_freq',

'actiontype\_Quick Watch Set\_freq', 'actiontype\_Stop Ignoring\_freq',

'actiontype\_Watch Cleared\_freq', 'actiontype\_Watch Expiration\_freq',

'actiontype\_Watch Override\_freq', 'actiontype\_Watch Set\_freq',

'touches\_count\_freq', 'case\_count\_freq'

# Modeling

Tried Kmeans, DbScan, and Hierarchical Clustering.

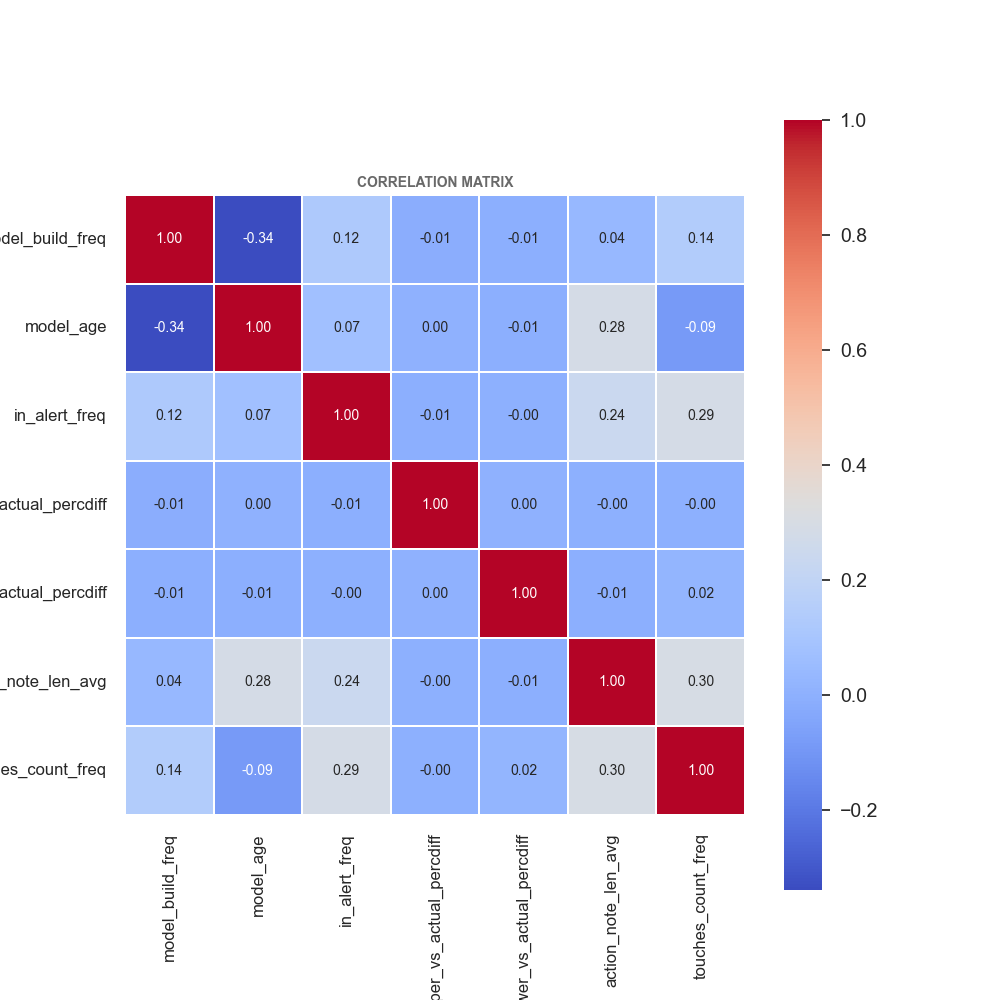
Kmeans had the best scores and best clustering and was used for final report

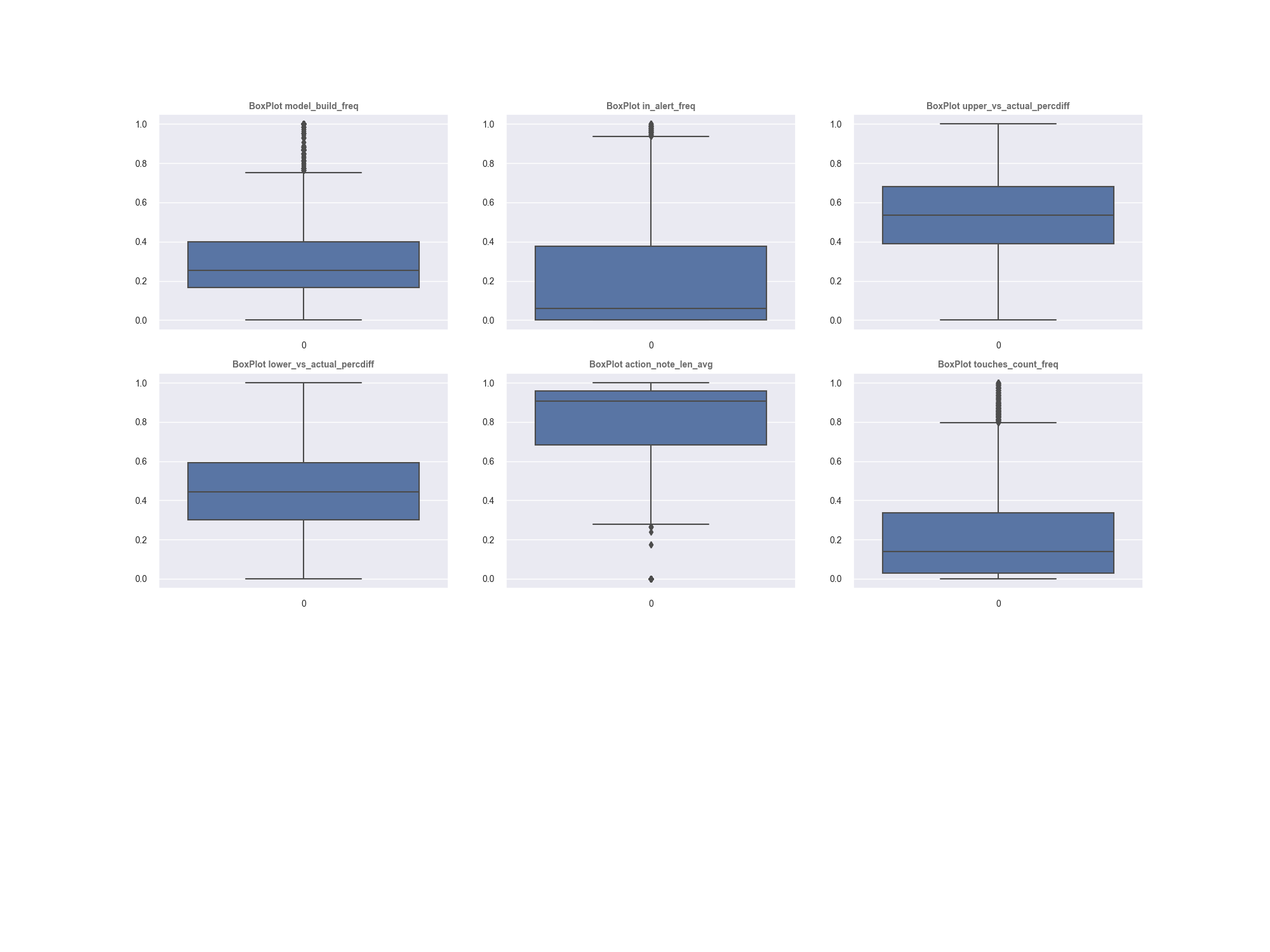
Features Used in model:

'model\_build\_freq', 'in\_alert\_freq',

'upper\_vs\_actual\_percdiff', 'lower\_vs\_actual\_percdiff',

'action\_note\_len\_avg', 'touches\_count\_freq'



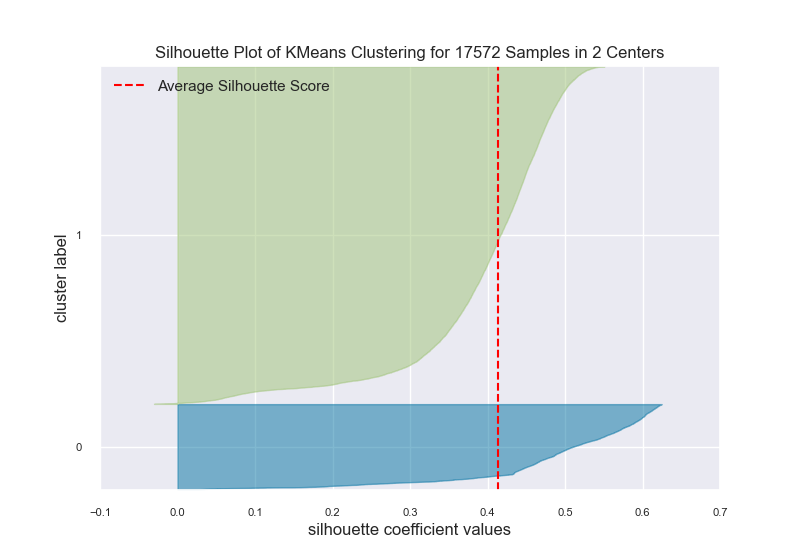


Chart, histogram

Description automatically generated

Graphical user interface, chart, line chart

Description automatically generated



Chart, funnel chart

Description automatically generated

Chart

Description automatically generated

