Evaluation Report

# Descriptive Statistics (Overall)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | count | mean | std | min | 25% | 50% | 75% | max |
| IBS | 320.000 | 53.154 | 15.837 | 0.000 | 48.148 | 56.351 | 62.869 | 100.000 |
| CFS | 320.000 | 38.032 | 27.452 | 0.000 | 16.582 | 28.571 | 60.000 | 100.000 |
| CFS without Floor and Walls | 320.000 | 72.932 | 23.594 | 0.000 | 57.143 | 79.239 | 91.837 | 100.000 |
| VCIS | 320.000 | 93.861 | 9.075 | 0.000 | 93.177 | 95.925 | 97.970 | 100.000 |
| NCIS | 320.000 | 96.618 | 5.496 | 62.382 | 97.016 | 98.670 | 99.333 | 100.000 |

# Descriptive Statistics (Duration & Elements)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | count | mean | std | min | 25% | 50% | 75% | max |
| PipelineDurationSeconds | 320.000 | 1086.657 | 369.403 | 365.675 | 801.506 | 1089.910 | 1322.384 | 2210.346 |
| TotalElementsCount | 320.000 | 42.031 | 16.224 | 5.000 | 28.000 | 42.500 | 55.000 | 97.000 |

# Descriptive Statistics by Validator

## with Validator

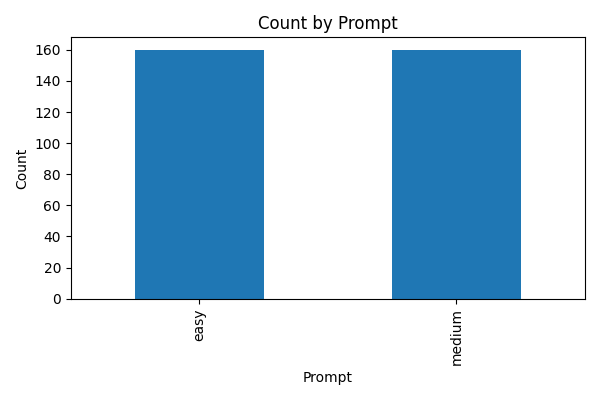
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | count | mean | std | min | 25% | 50% | 75% | max |
| IBS | 160.000 | 54.205 | 15.218 | 0.000 | 48.433 | 56.639 | 63.409 | 100.000 |
| CFS | 160.000 | 49.077 | 28.784 | 0.000 | 23.986 | 47.013 | 73.021 | 100.000 |
| CFS without Floor and Walls | 160.000 | 74.767 | 24.715 | 0.000 | 60.831 | 82.609 | 96.282 | 100.000 |
| VCIS | 160.000 | 94.306 | 11.079 | 0.000 | 94.216 | 97.111 | 98.969 | 100.000 |
| NCIS | 160.000 | 96.357 | 6.371 | 62.382 | 97.137 | 98.683 | 99.379 | 100.000 |

## without Validator

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | count | mean | std | min | 25% | 50% | 75% | max |
| IBS | 160.000 | 52.103 | 16.414 | 3.704 | 46.360 | 55.338 | 62.069 | 86.207 |
| CFS | 160.000 | 26.987 | 20.927 | 0.000 | 12.914 | 19.379 | 33.750 | 91.837 |
| CFS without Floor and Walls | 160.000 | 71.097 | 22.343 | 0.000 | 55.525 | 73.861 | 89.474 | 100.000 |
| VCIS | 160.000 | 93.416 | 6.488 | 54.057 | 92.828 | 94.962 | 96.851 | 99.733 |
| NCIS | 160.000 | 96.879 | 4.458 | 78.190 | 97.007 | 98.656 | 99.283 | 99.862 |

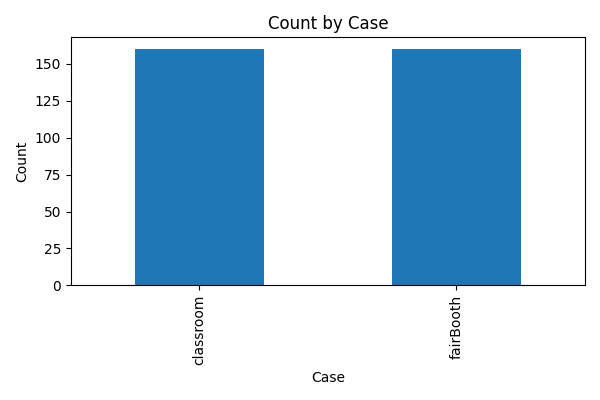
# Counts by Prompt

|  |  |
| --- | --- |
| Prompt | Count |
| easy | 160 |
| medium | 160 |



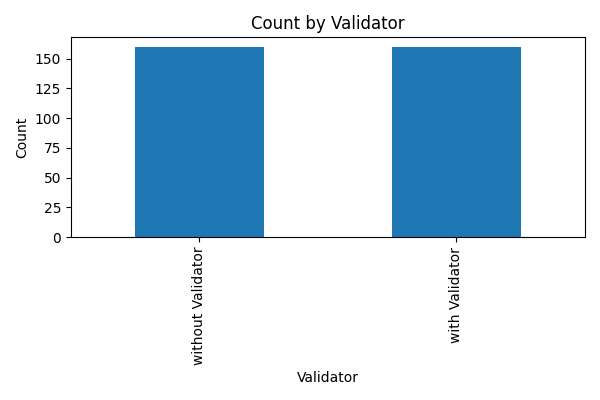
# Counts by Case

|  |  |
| --- | --- |
| Case | Count |
| classroom | 160 |
| fairBooth | 160 |



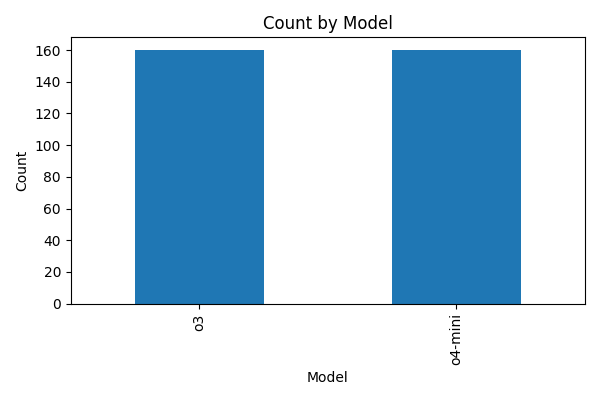
# Counts by Validator

|  |  |
| --- | --- |
| Validator | Count |
| without Validator | 160 |
| with Validator | 160 |



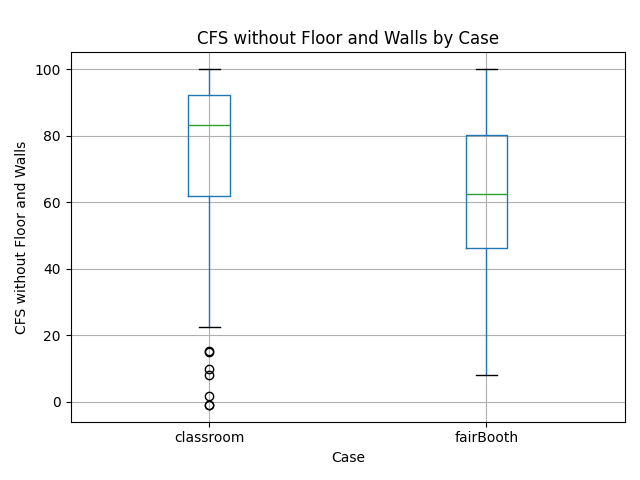
# Counts by Model

|  |  |
| --- | --- |
| Model | Count |
| o3 | 160 |
| o4-mini | 160 |

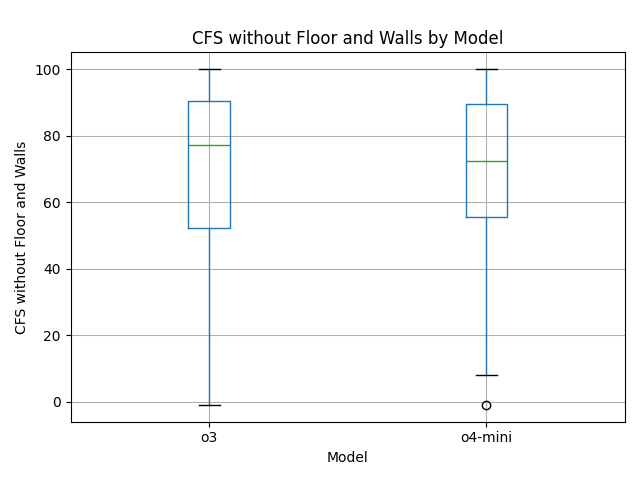


# All Other Plots

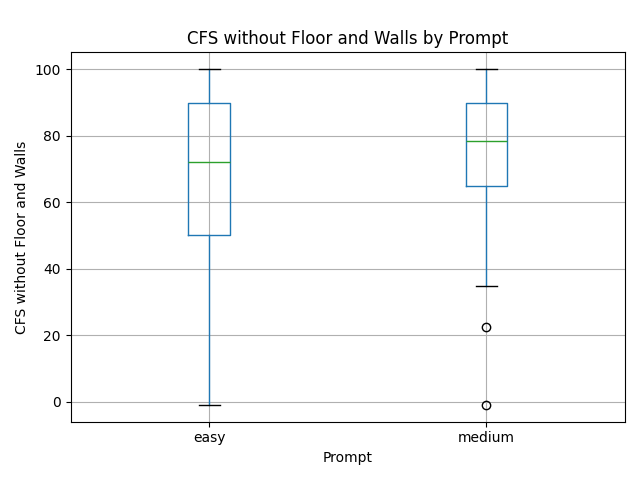
## Box Cfs Without Floor And Walls Case



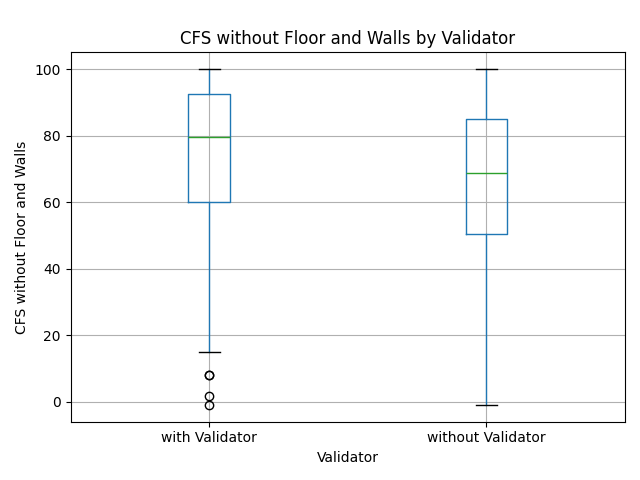
## Box Cfs Without Floor And Walls Model



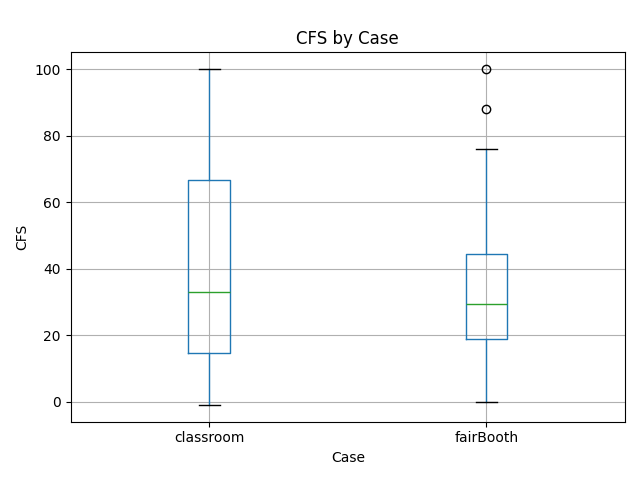
## Box Cfs Without Floor And Walls Prompt



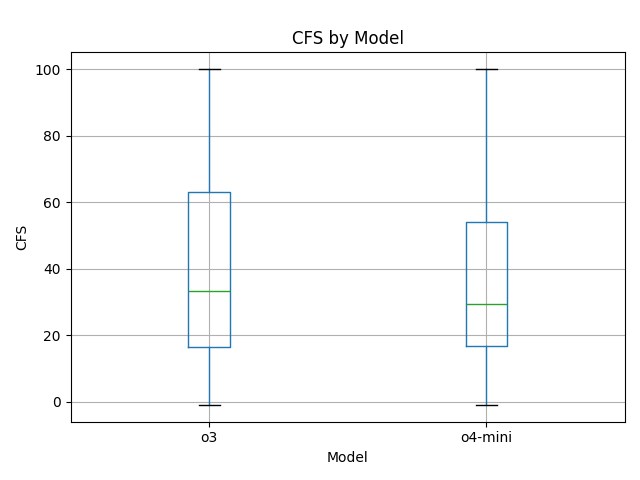
## Box Cfs Without Floor And Walls Validator



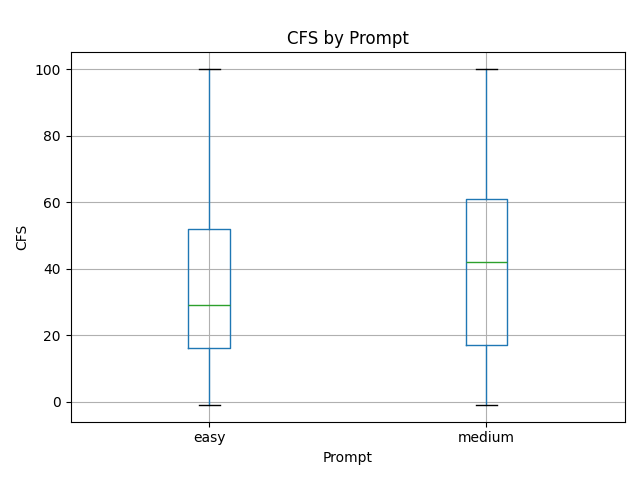
## Box Cfs Case



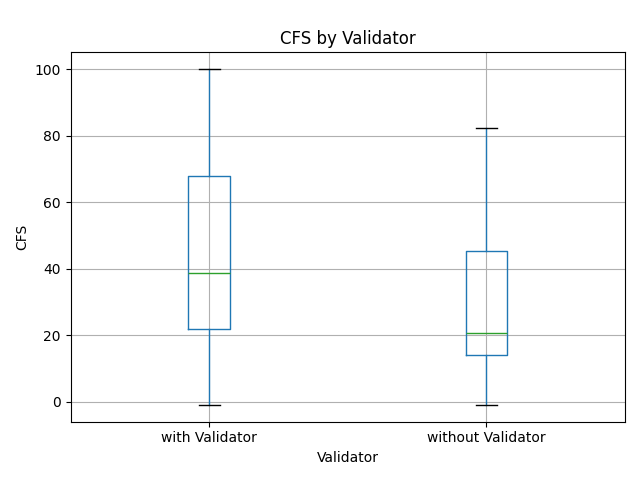
## Box Cfs Model



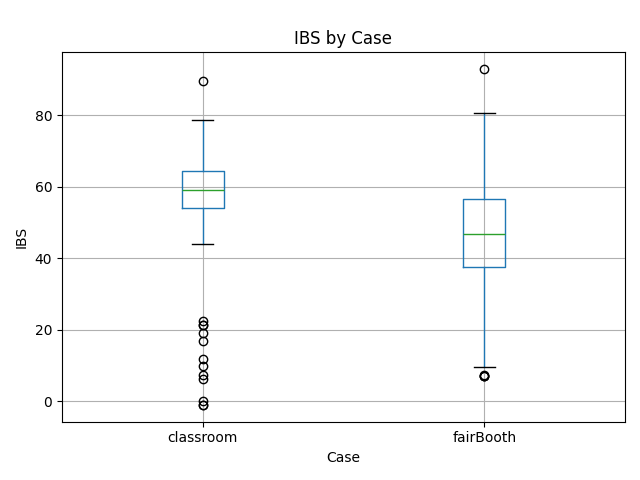
## Box Cfs Prompt



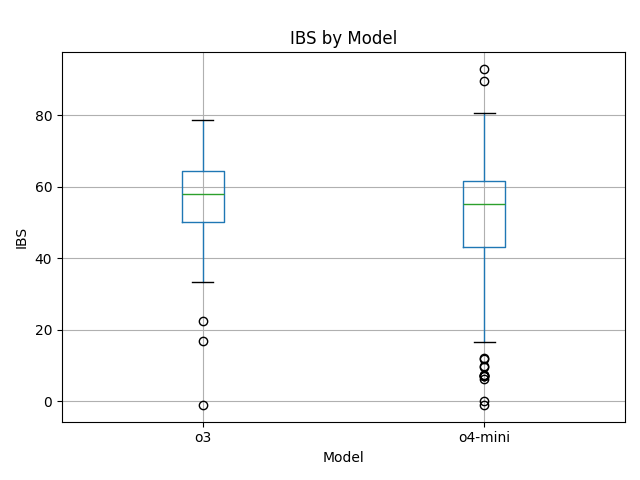
## Box Cfs Validator



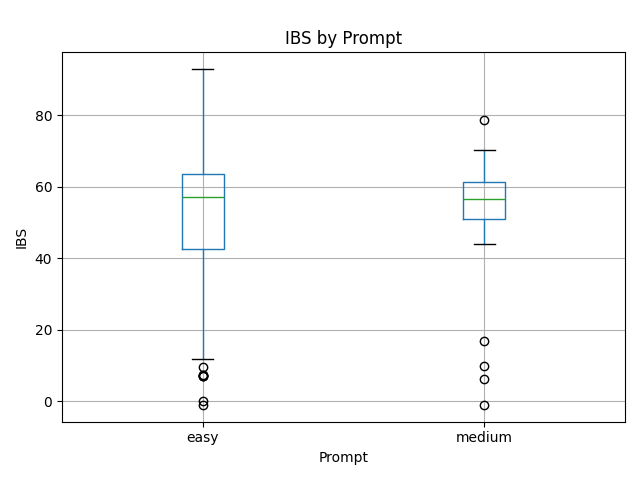
## Box Ibs Case



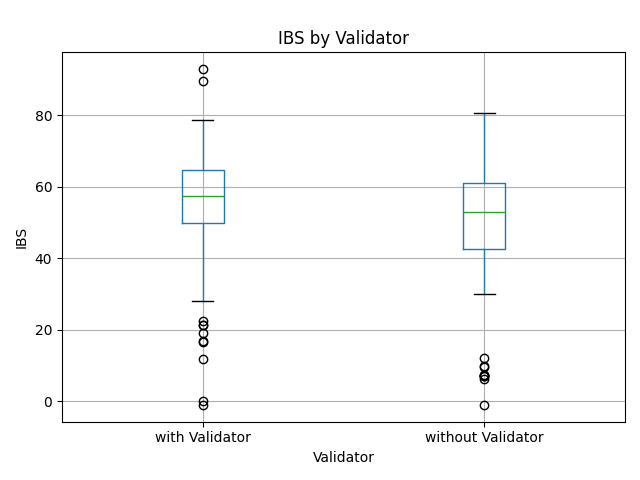
## Box Ibs Model



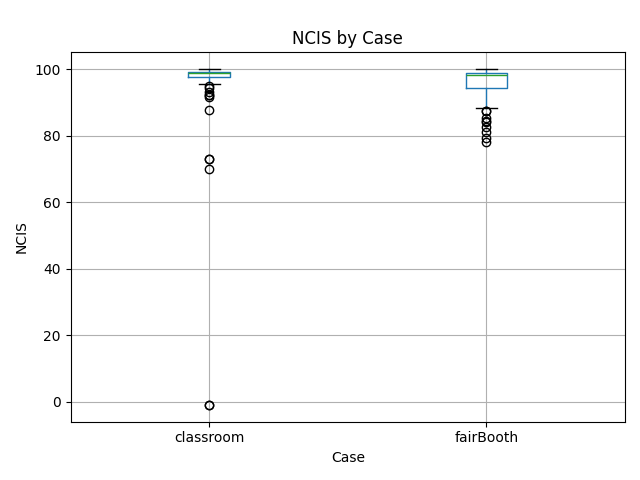
## Box Ibs Prompt



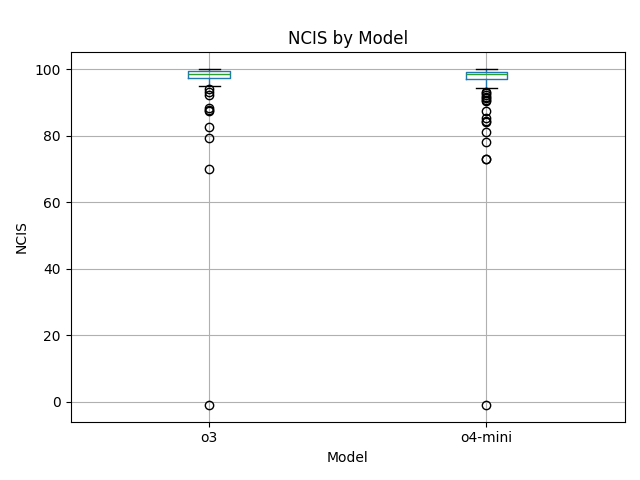
## Box Ibs Validator



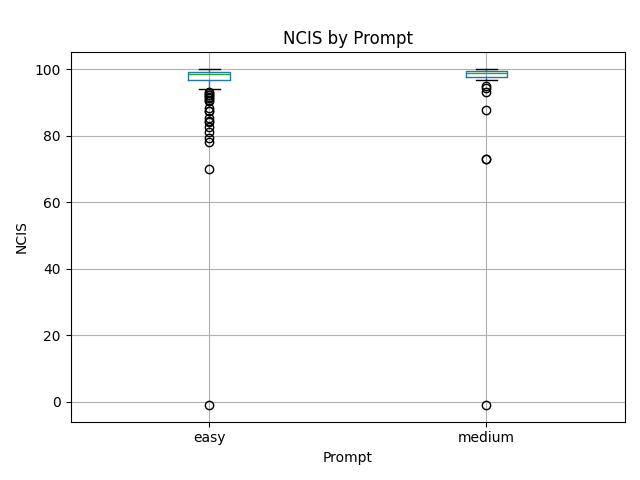
## Box Ncis Case



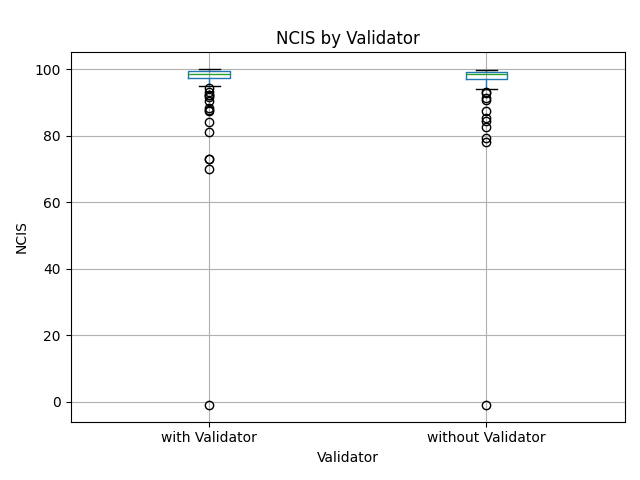
## Box Ncis Model



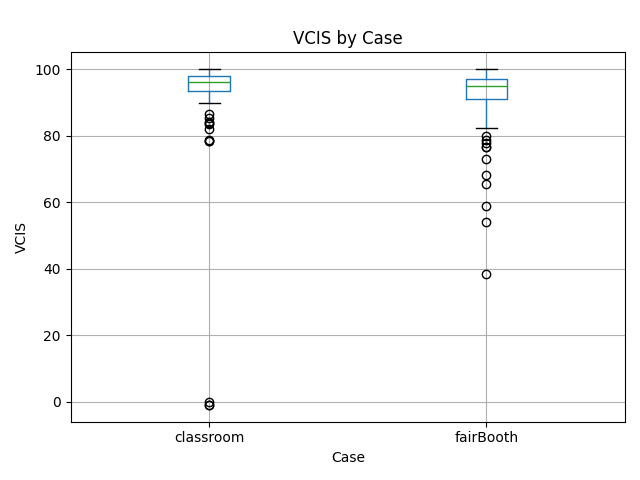
## Box Ncis Prompt



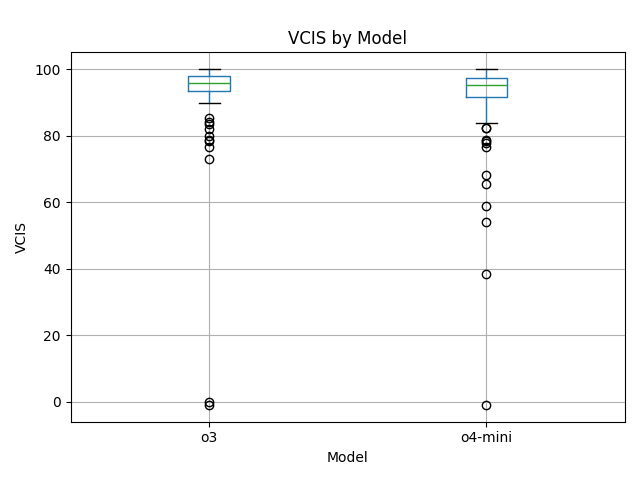
## Box Ncis Validator



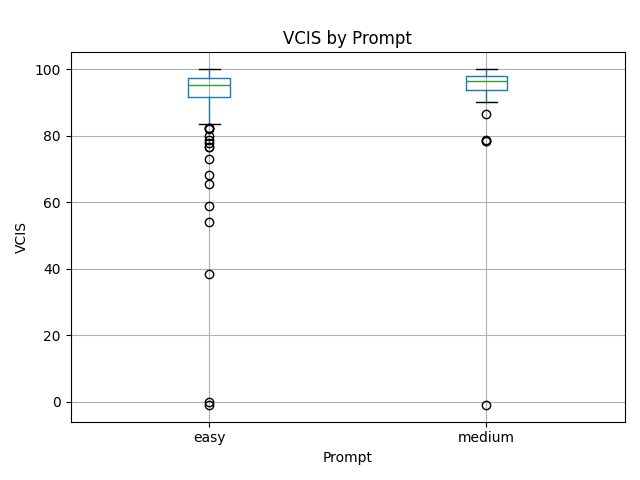
## Box Vcis Case



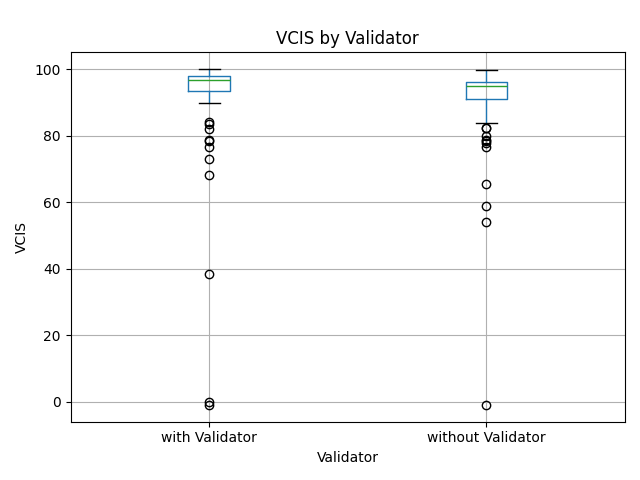
## Box Vcis Model



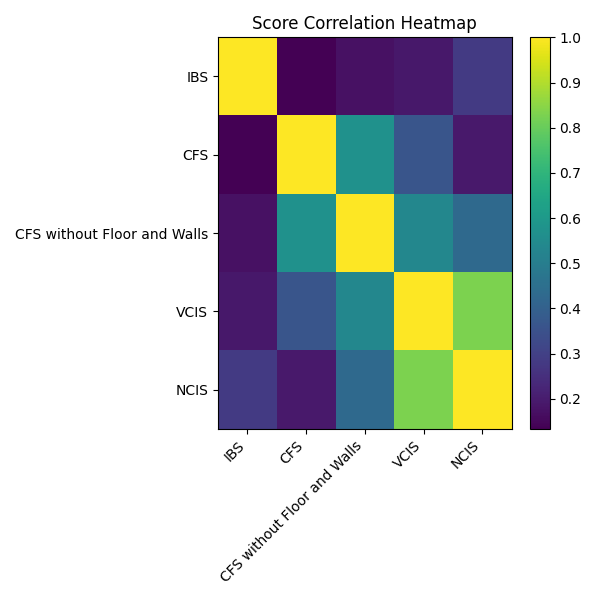
## Box Vcis Prompt



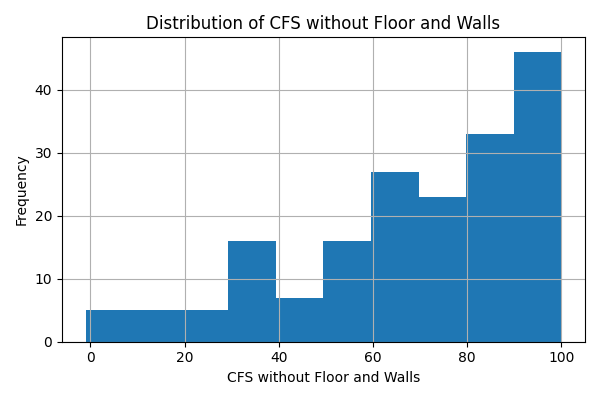
## Box Vcis Validator



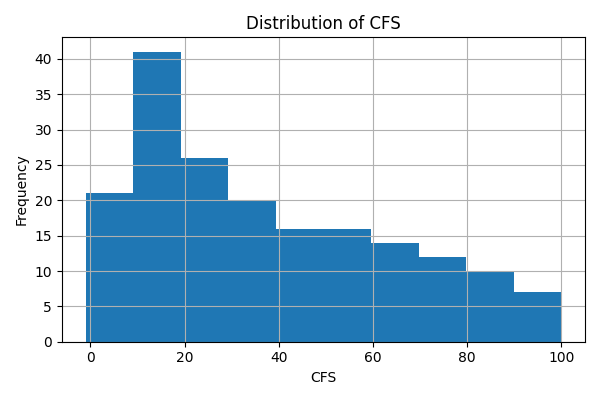
## Heatmap Correlation



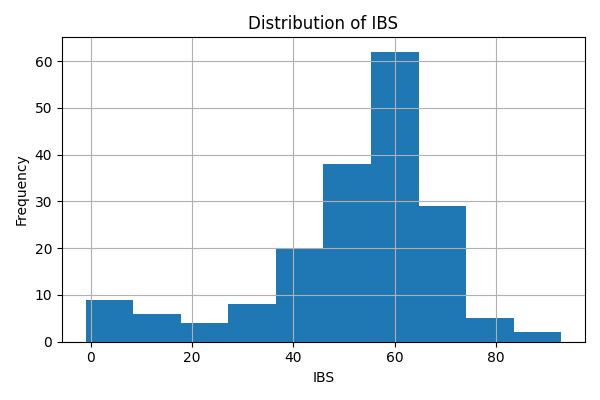
## Hist Cfs Without Floor And Walls



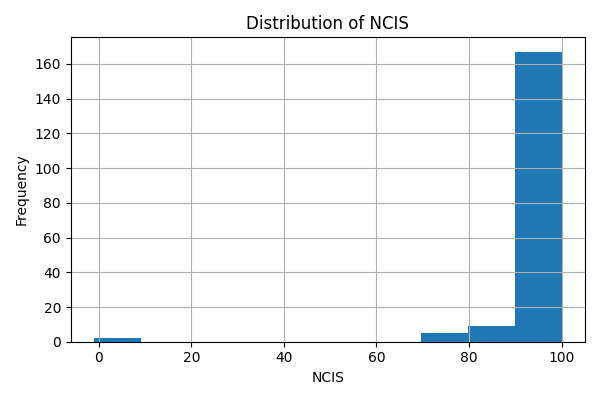
## Hist Cfs



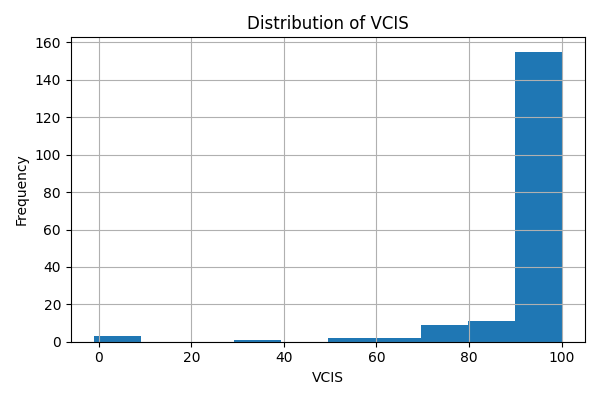
## Hist Ibs



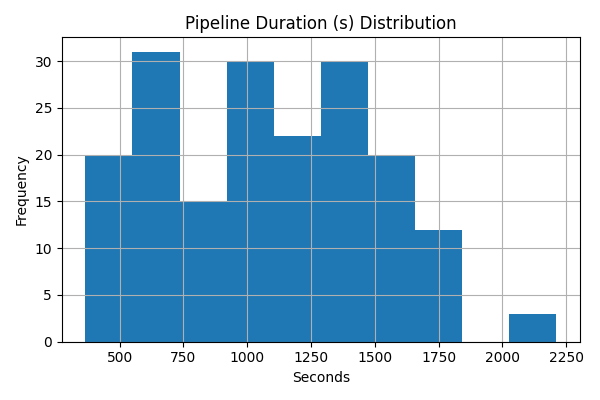
## Hist Ncis



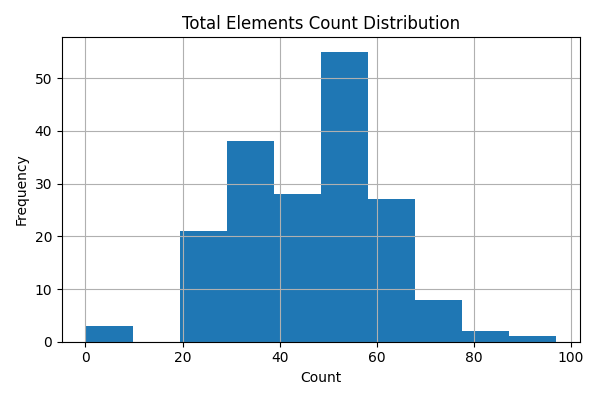
## Hist Vcis



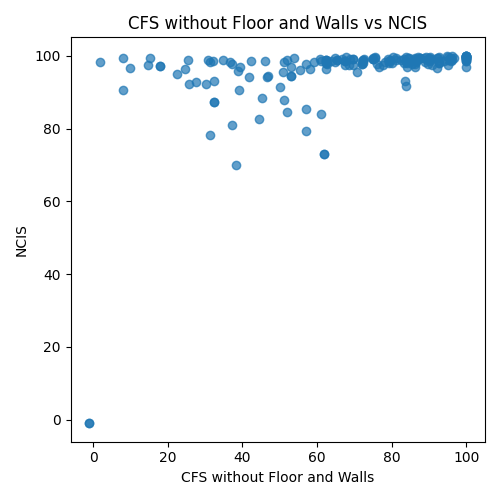
## Hist Pipeline Duration



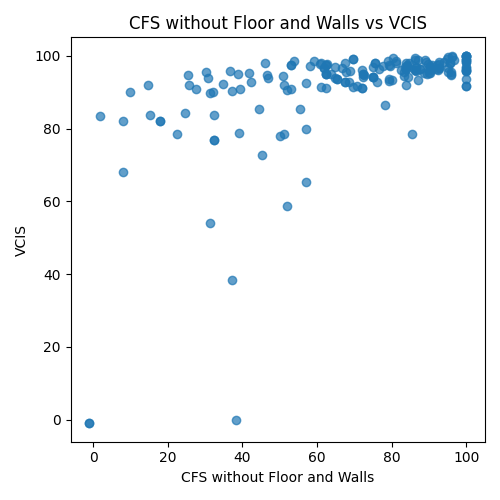
## Hist Total Elements



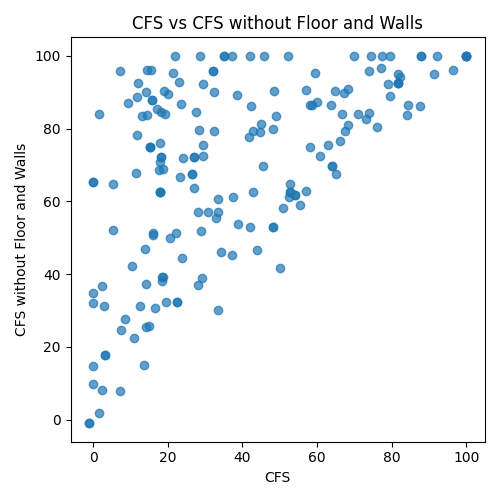
## Scatter Cfs Without Floor And Walls Ncis



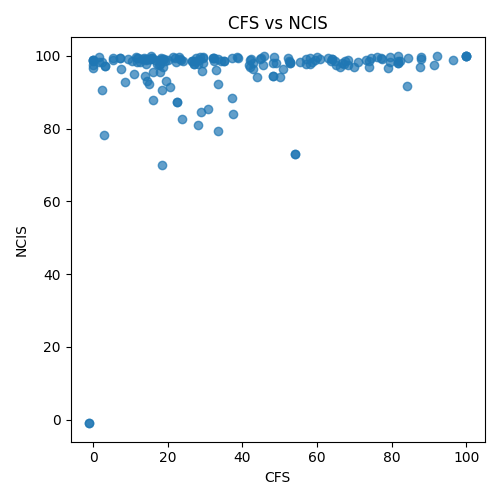
## Scatter Cfs Without Floor And Walls Vcis



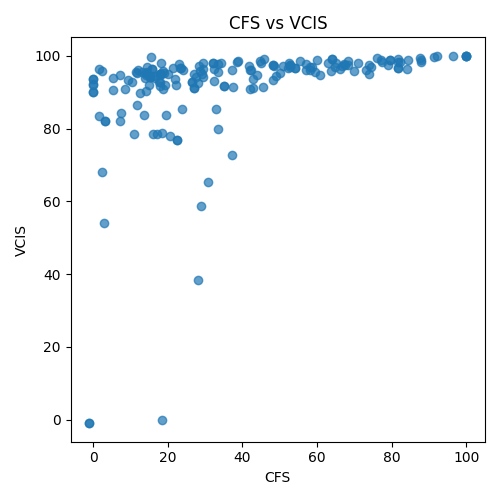
## Scatter Cfs Cfs Without Floor And Walls



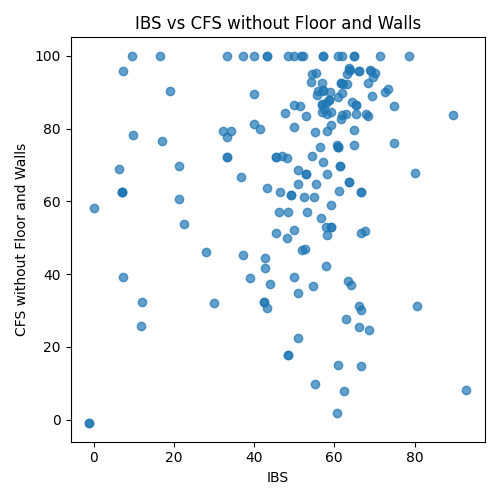
## Scatter Cfs Ncis



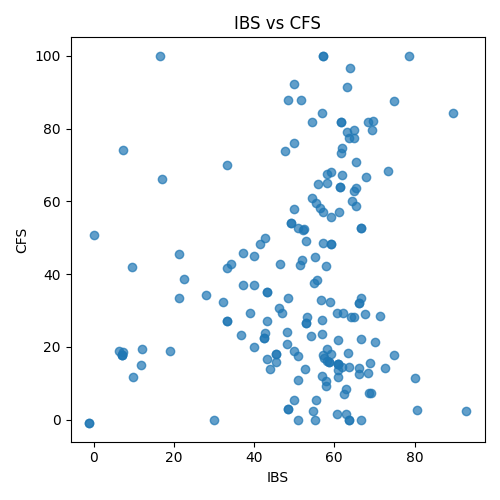
## Scatter Cfs Vcis



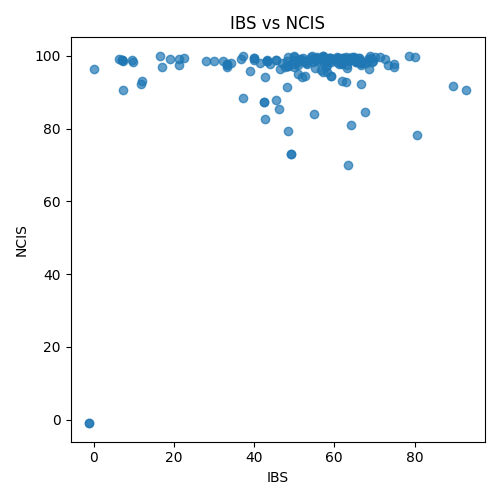
## Scatter Ibs Cfs Without Floor And Walls



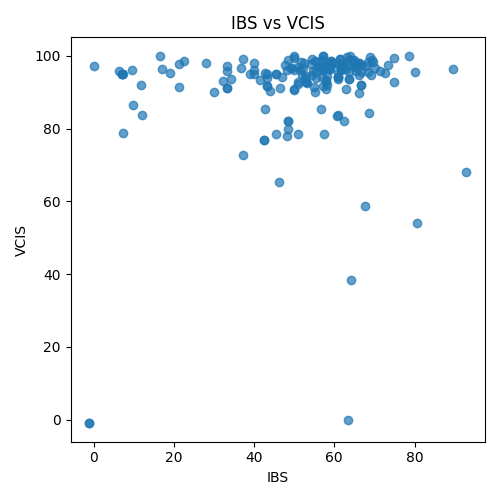
## Scatter Ibs Cfs



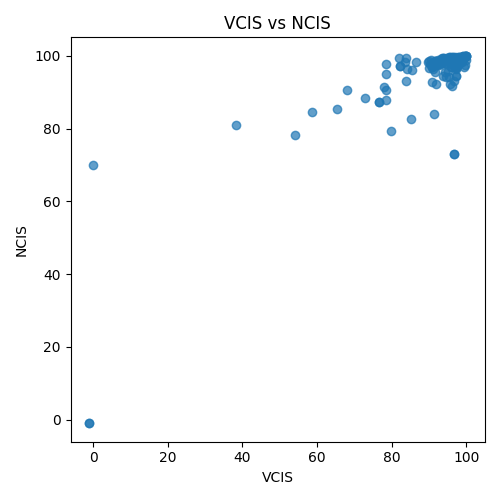
## Scatter Ibs Ncis



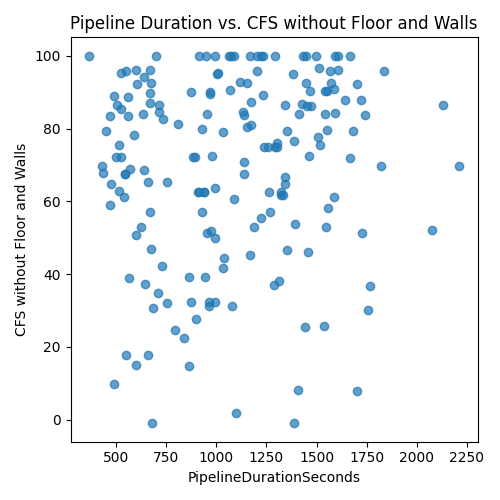
## Scatter Ibs Vcis



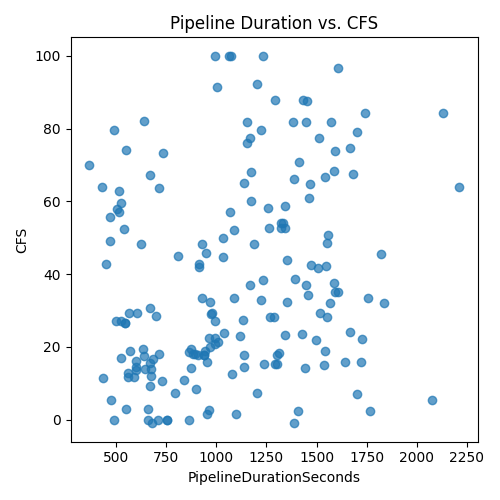
## Scatter Vcis Ncis



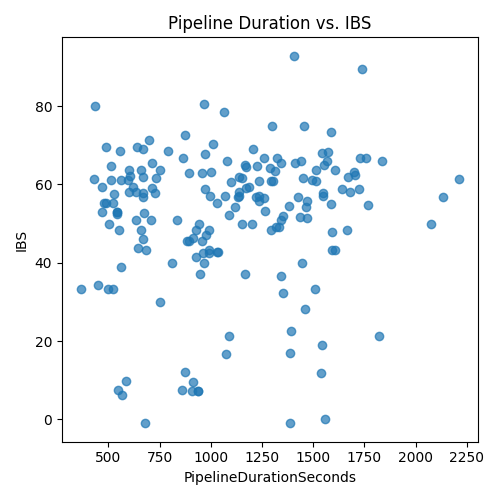
## Scatter Duration Cfs Without Floor And Walls



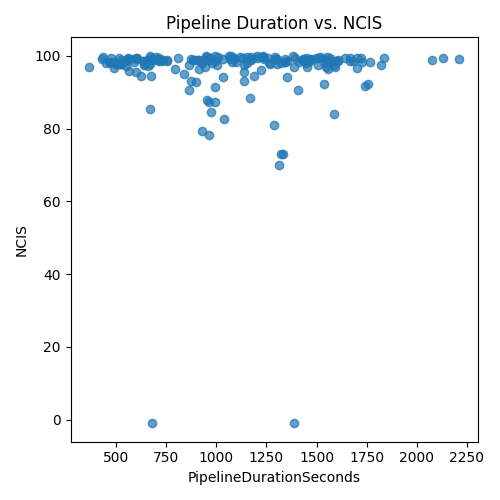
## Scatter Duration Cfs



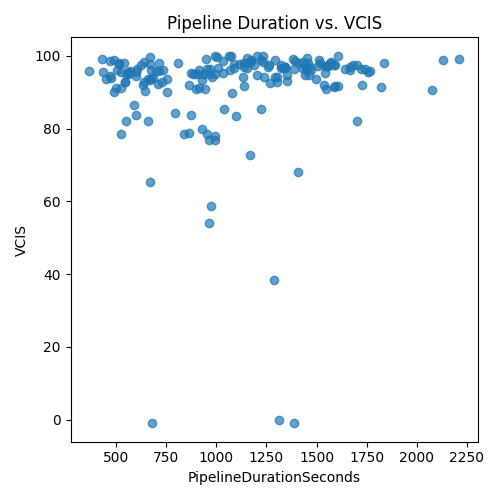
## Scatter Duration Ibs



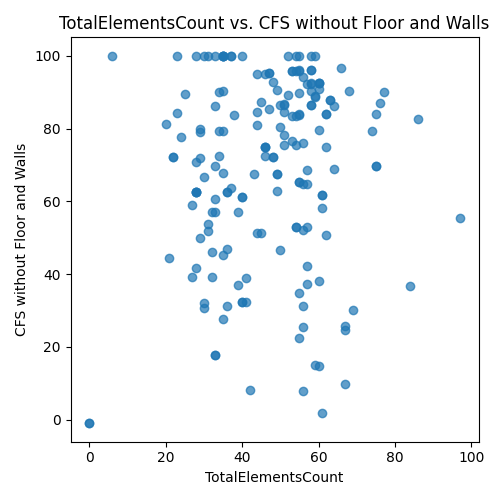
## Scatter Duration Ncis



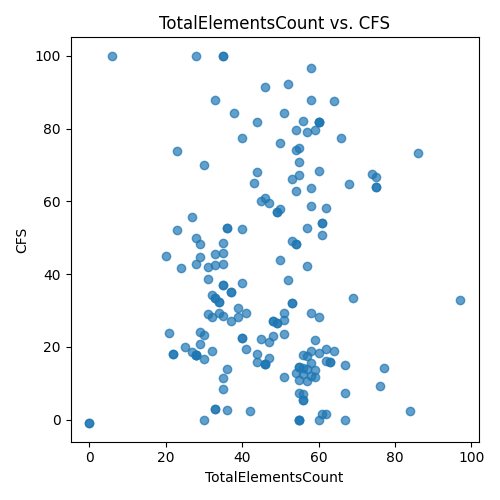
## Scatter Duration Vcis



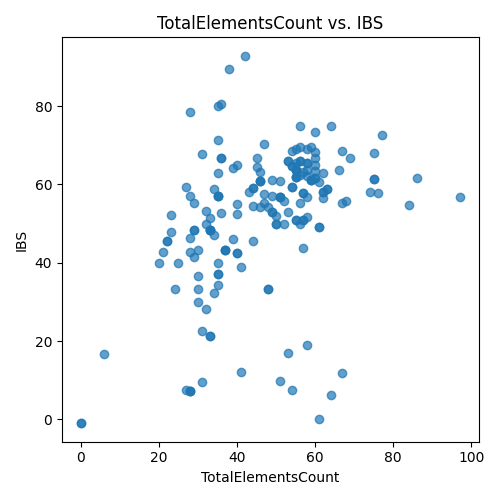
## Scatter Elements Cfs Without Floor And Walls



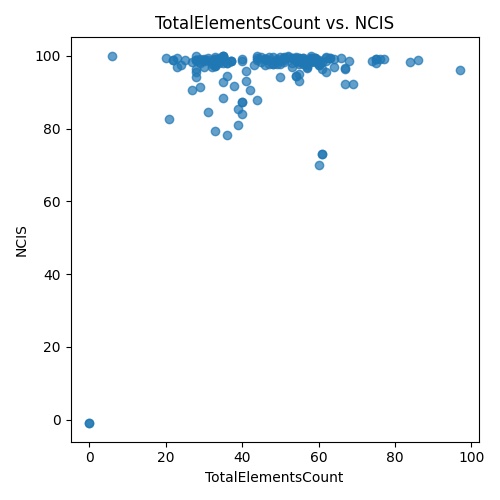
## Scatter Elements Cfs



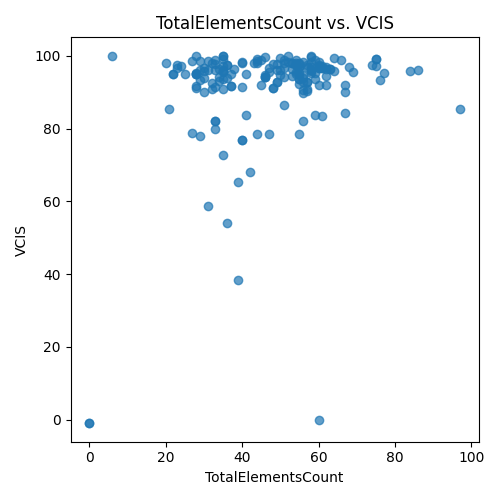
## Scatter Elements Ibs



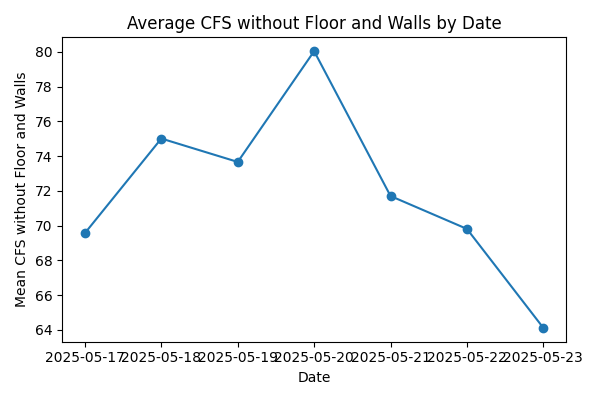
## Scatter Elements Ncis



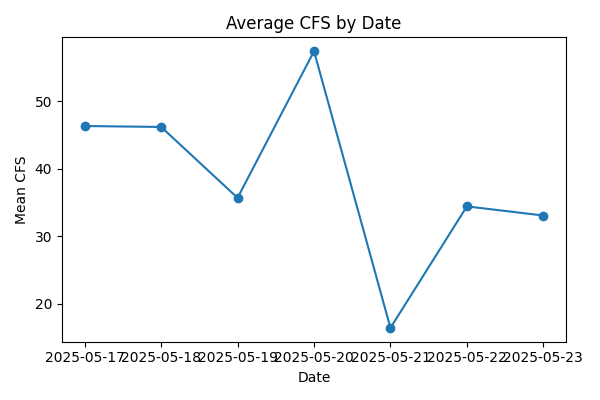
## Scatter Elements Vcis



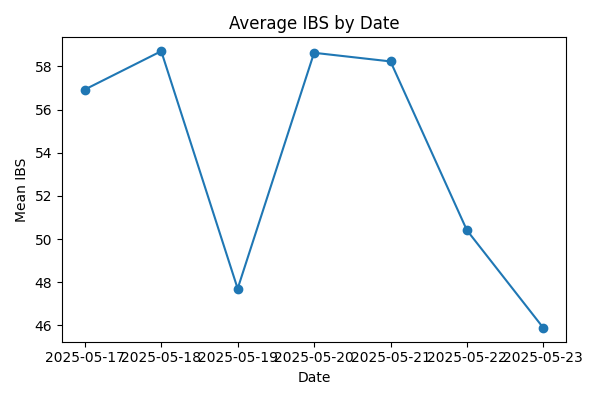
## Timeseries Cfs Without Floor And Walls



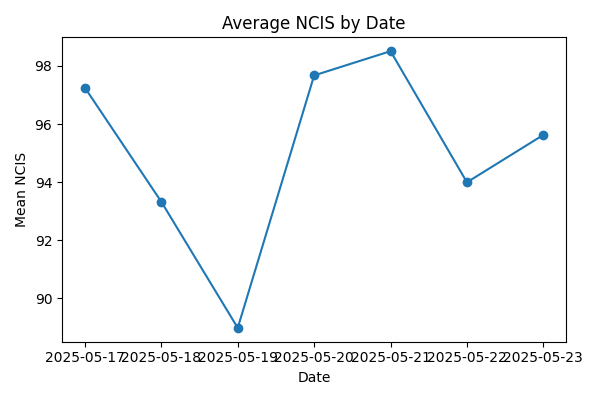
## Timeseries Cfs



## Timeseries Ibs



## Timeseries Ncis



## Timeseries Vcis

