Acadia DOE OBD Data Exploratory Analysis

Dingchao Zhang 11/12/2016

Acadia DOE Data Scope

- 7 System Errors, all Event Driven
- 8 Parameters
- 124 Tests from 3/25/2015 to 10/2/2015

SEID	Name	CriticalParam	Units	LSL	USL	ThresholdValue
5365	NOX_IN_SENSOR_IR_HI_MOTOR_ERR	EONOx_IRH_Mot_Cusum_	PPM	NULL	C_EONOx_IRH	2125
5366	NOX_IN_SENSOR_IR_LO_MOTOR_ERR	EONOx_IRL_Mot_Cusum_	PPM	NULL	C_EONOx_IRL	1125
5976	NOX_OUT_SENSOR_IR_HI_MOTOR_ERR	P_SCD_ppm_NOxOff_Filt1	ppm	NULL	C_SCD_ppm_	1500
5976	NOX_OUT_SENSOR_IR_HI_MOTOR_ERR_decision	V_SCD_ppm_AvgNOxOff_	ppm	NULL	NULL	
5978	NOX_OUT_SENSOR_IR_LO_MOTOR_ERR	P_SCD_ppm_NOxOff_Filt1	ppm	C_SCD_pp	NULL	-30
5978	NOX_OUT_SENSOR_IR_LO_MOTOR_ERR_decision	V_SCD_ppm_AvgNOxOff_	ppm	NULL	NULL	
10102	SCR_CAT_SUBSTRATE_MISSING_ERR	P_SCDE_CM_NormEff_EW	None	C_SCDE_C	NULL	-5
12493	SCR_EFFICIENCY4_DEGRADED_ERR	P_SCDE_CE4_EWMA_FiltV	None	NULL	C_SCDE_CE4_	1000
5369	NOX_IN_SENSOR_DITHER_ERR	EONOx_SIR_Delta	ppm	C_EONOx_	NULL	0

Plots Types

- Individual System Error Plots:

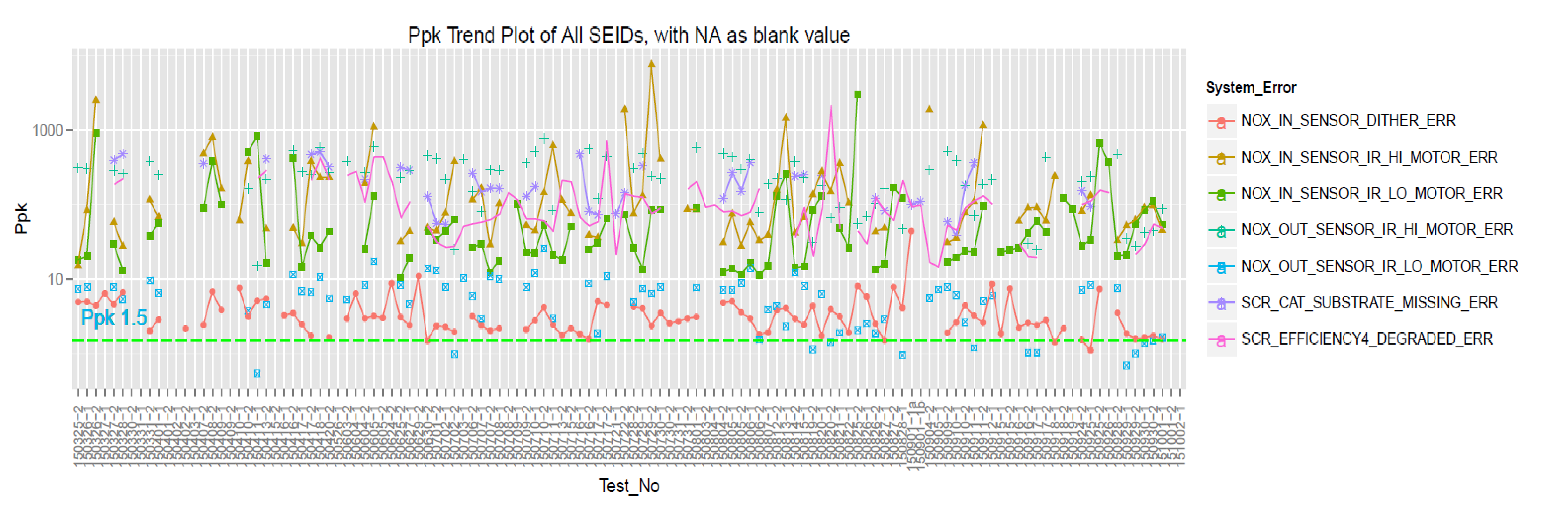
1. Mean, Max, Min range Plot(Mean,Max,Min on the same plot)

- All System Errors Plot:

1. Ppk Trend Plot
2. Mean+-1Std Plot
3. Mean, Max, Min range Plot(Mean,Max,Min on the same plot)

R output graphs can be found in
\\CIDCSDFS01\EBU_Data01$\NACTGx\Common\DL_Diag\Data Analysis\Storage\Knowledge base\Acadia DOE\graphs

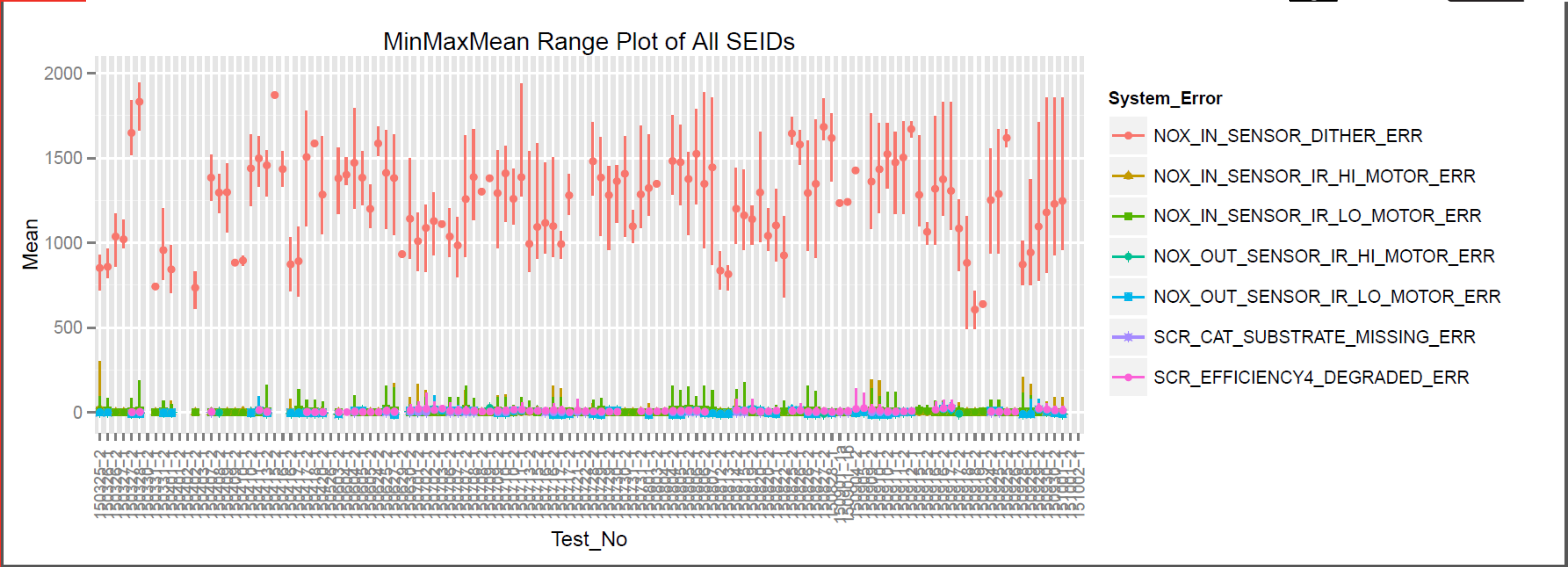
Ppk Trend All System Error Plot



Observations:

- NOX_OUT_SENSOR_IR_LO_MOTOR_ERR has a few tests whose Ppk are below 1.5
- NOX_IN_SENSOR_DITHER_ERR has one test whose Ppk is below 1.5
- All other System Errors have Ppk above 1.5 in all tests

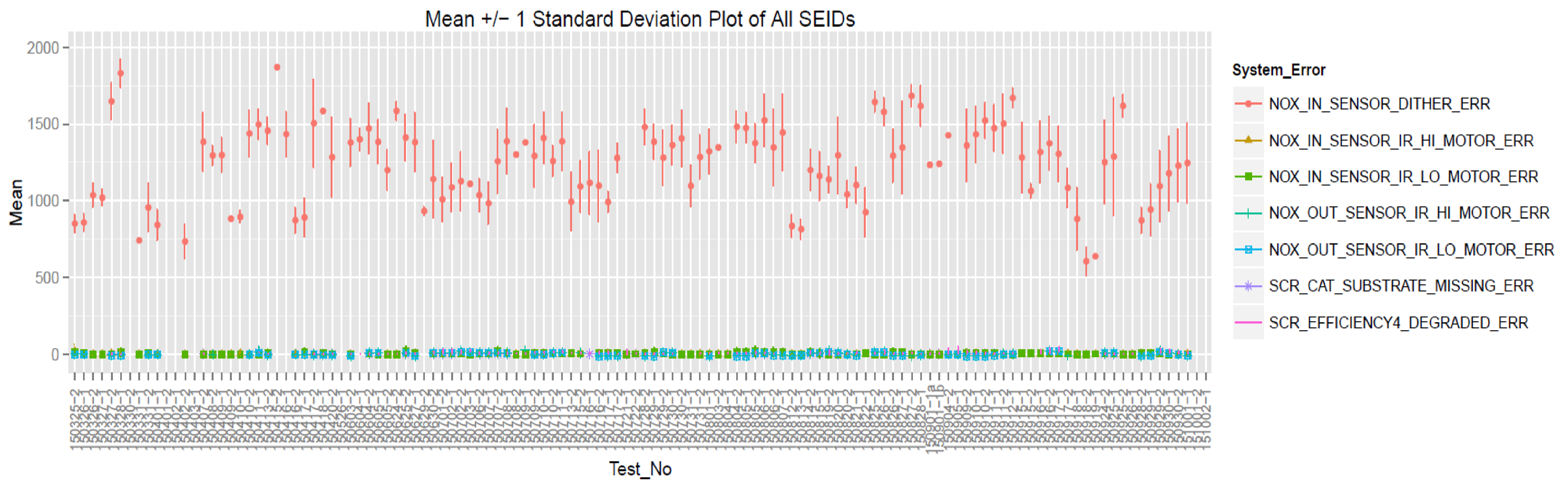
MinMaxMean Range All System Error Plot



Observations:

- NOX_IN_SENSOR_DITHER_ERR has larger variance than other System Errors
- NOX_IN_SENSOR_DITHER_ERR appears to have a mean below 1000 or above 1500 in the first few tests, while in the more recent tests, the mean appears between 1000 and 1500

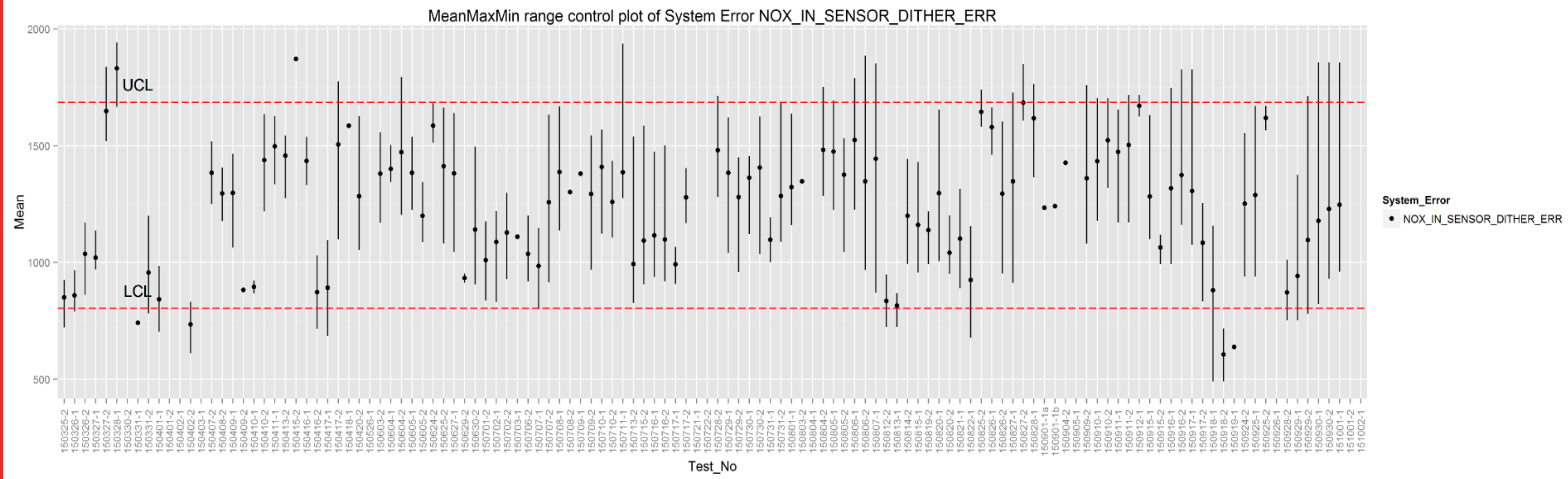
Mean+-1Std All System Error Plot



Observations:

- NOX_IN_SENSOR_DITHER_ERR has the largest variance, and the variance appears gets larger as tests carry on

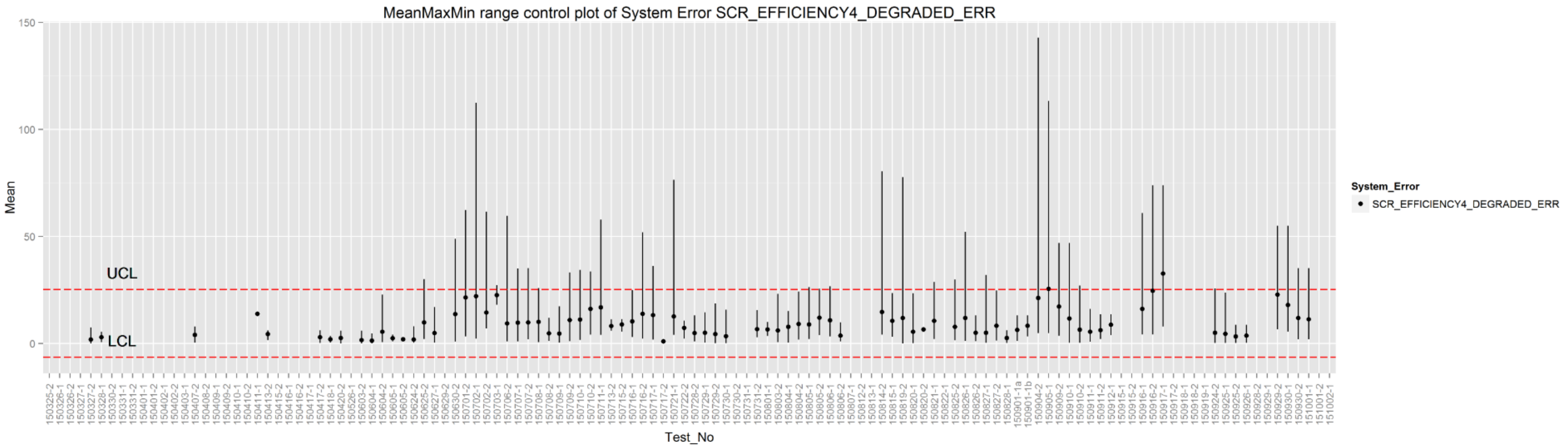
NOX_IN_SENSOR_DITHER_ERR



Observations:

- NOX_IN_SENSOR_DITHER_ERR appears to have a mean below 1000 or above 1500 in the first few tests, while in the more recent tests, the mean appears between 1000 and 1500
- NOX_IN_SENSOR_DITHER_ERR’s variance appears gets larger as tests carry on

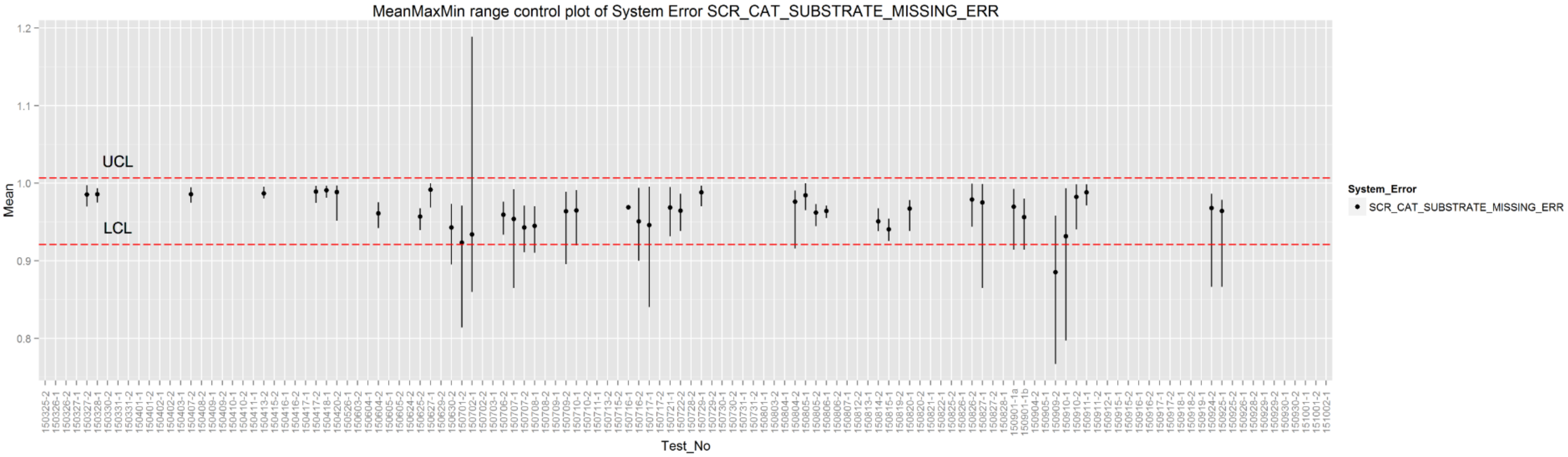
SCR_EFFICIENCY4_DEGRADED_ERR



Observations:

- SCR_EFFICIENCY4_DEGRADED_ERR appears not making many decisions in the first few tests
- Majority of tests show the maximum values are about 2 times larger than mean values, with a few tests showing maximum values as large as 3 times of mean values.

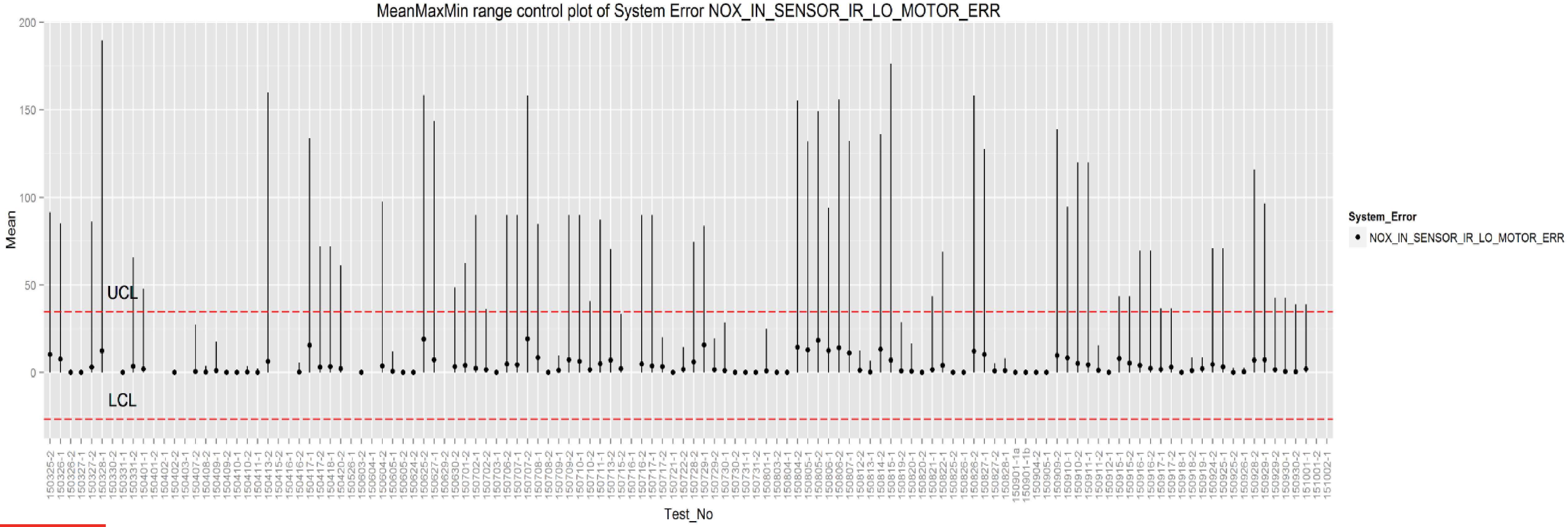
SCR_CAT_SUBSTRATE_MISSING_ERR



Observations:

- SCR_CAT_SUBSTRATE_MISSING_ERR appears not making many decisions in the first few tests
- SCR_CAT_SUBSTRATE_MISSING_ERR has a few tests records showing minimum or maximum value at a larger difference than mean values

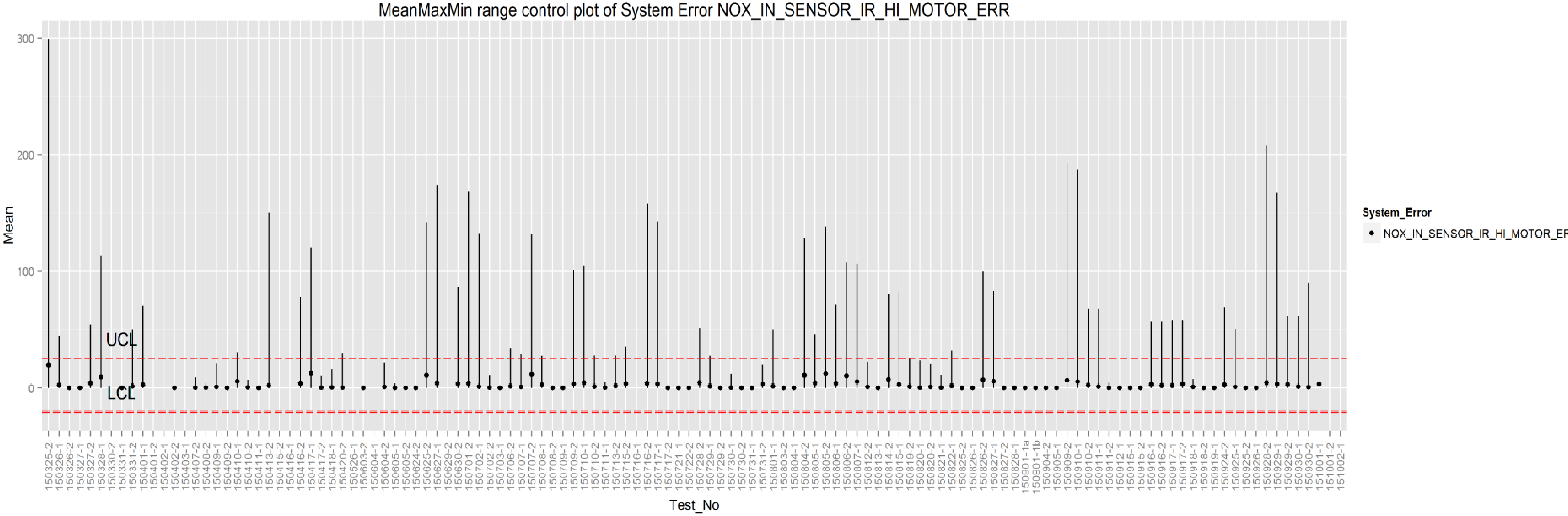
NOX_IN_SENSOR_IR_LO_MOTOR_ERR



Observations:

- NOX_IN_SENSOR_IR_HI_MOTOR_ERR's performances are correlated with NOX_IN_SENSOR_IR_LO_MOTOR_ERR's performances

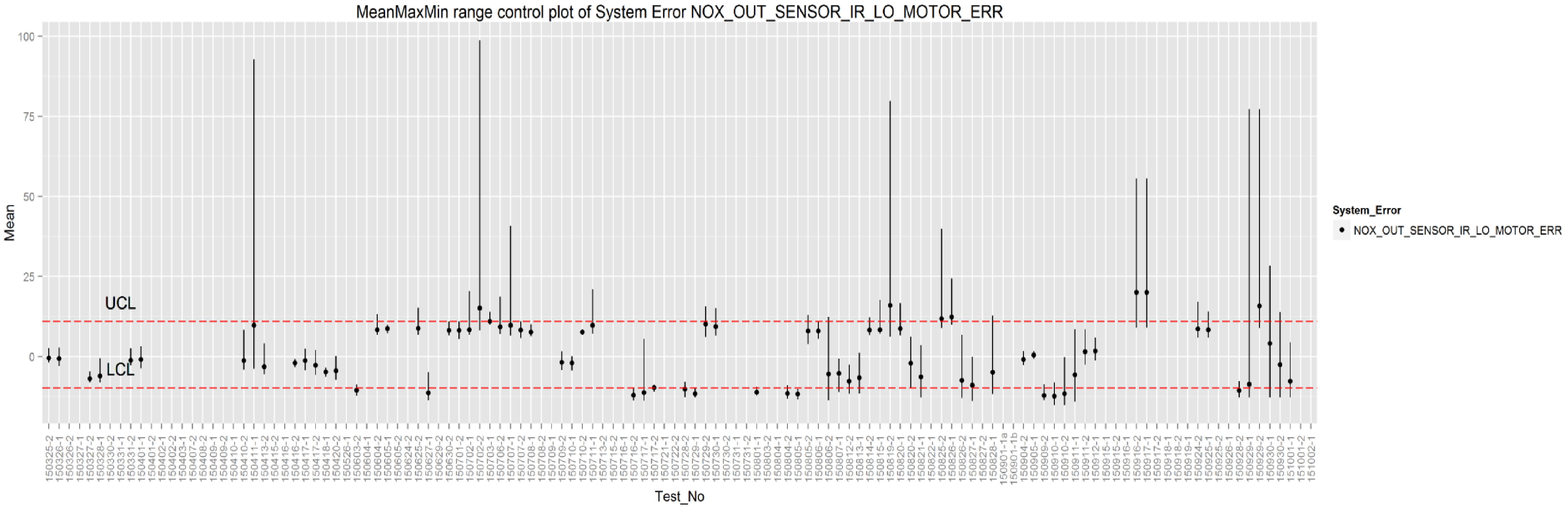
NOX_IN_SENSOR_IR_HI_MOTOR_ERR



Observations:

- NOX_IN_SENSOR_IR_HI_MOTOR_ERR’s performances are correlated with NOX_IN_SENSOR_IR_LO_MOTOR_ERR’s performances
- A few tests showing very large maximum values

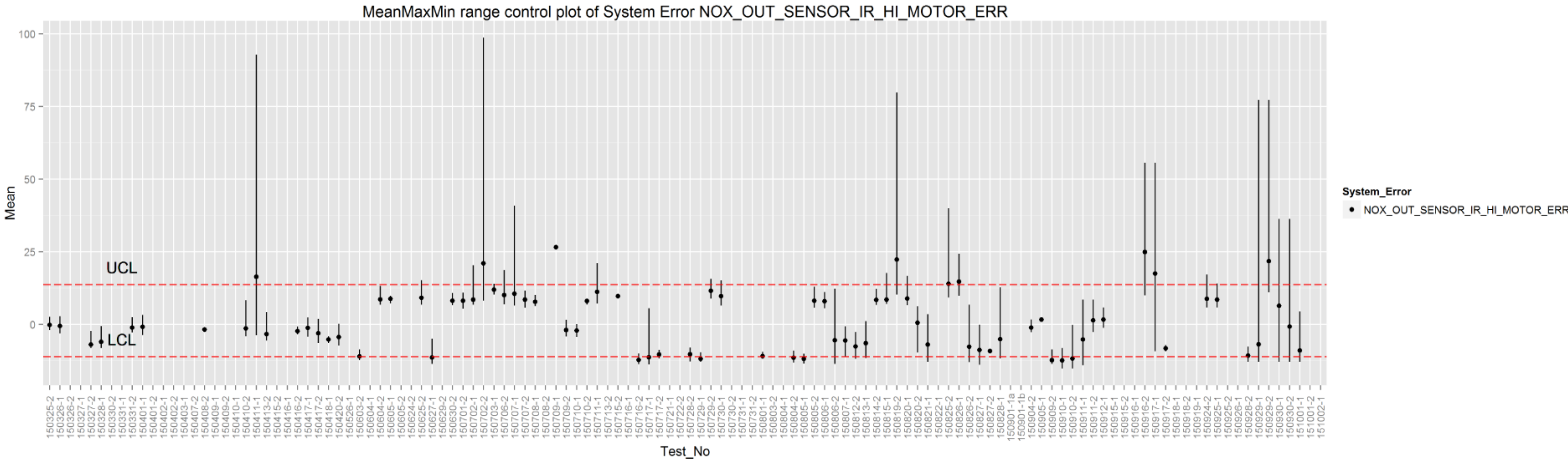
NOX_OUT_SENSOR_IR_LO_MOTOR_ERR



Observations:

- NOX_OUT_SENSOR_IR_LO_MOTOR_ERR's performances are highly correlated with NOX_OUT_SENSOR_IR_HI_MOTOR_ERR's performances
- A few tests showing very large maximum values

NOX_OUT_SENSOR_IR_HI_MOTOR_ERR



Observations:

- NOX_OUT_SENSOR_IR_LO_MOTOR_ERR's performances are highly correlated with NOX_OUT_SENSOR_IR_HI_MOTOR_ERR's performances
- A few tests showing very large maximum values

Unsupervised Machine Learning Practices

To provide insight into the 1st goal of the Acadia DOE tests listed by Paul

Overall, There are several goals of the project:

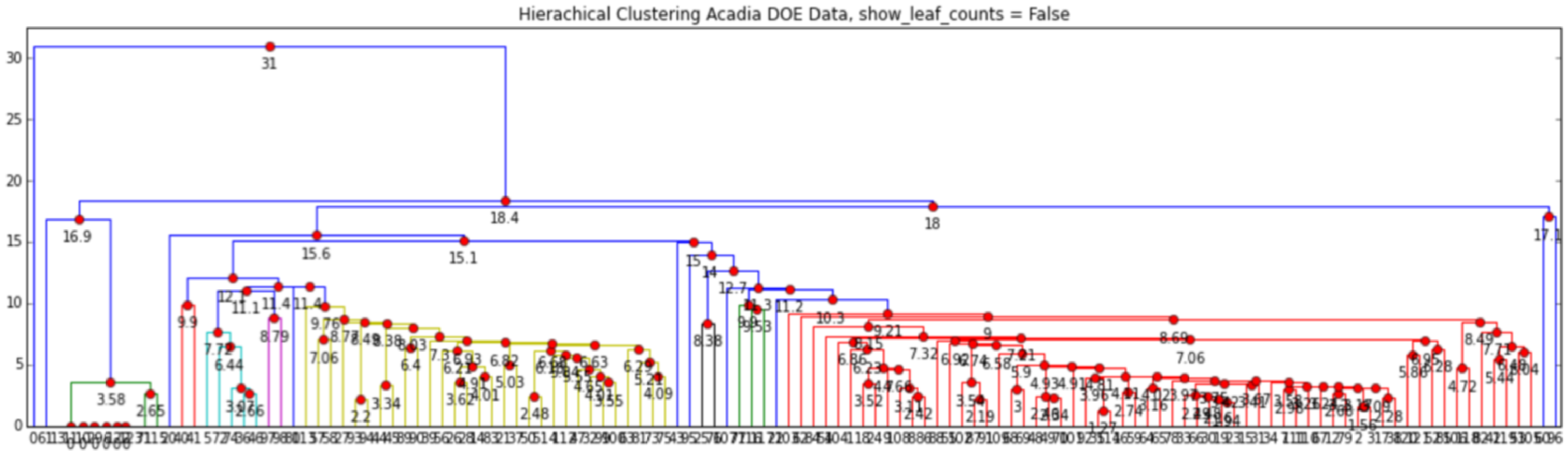
1. On a macro-level, create a field test data set that is richer in variance than a normal field test. I am certain that our DOE data set is richer (more variance) than a single truck. But, I need to quantify how much richer. I hope to compare various inter/intra group comparisons of the DOE data set. I also hope to compare variance and COV of DOE to other Acadia test trucks.

Tools:
Python scikit, pandas, numpy, matplot packages are used to complete the following analysis.
Scripts are not attached.

Hierarchical Clustering

Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together. We begin with a distance matrix which contains the distances between every pair of objects in our database.

Dendrogram of Acadia DOE Tests using OBD Data



Hierarchical Clustering Groups

Refers to the color in the dendrogram in slide 15

r	['40', '41', '18', '24', '8', '86', '108', '9', '1', '104', '87', '91', '102', '68', '69', '49', '70', '48', '35', '114', '92', '16', '59', '64', '65', '19', '23', '30', '66', '33', '15', '31', '7', '111', '12', '79', '2', '3', '17', '38', '67', '2', '110', '34', '78', '101', '109', '55', '88', '120', '121', '52', '85', '54', '84', '106', '118', '42', '119', '53', '105', '82', '62', '103']
	['25', '76']
c	['36', '46', '74', '72', '5']
m	['97', '98']
y	['57', '58', '93', '94', '44', '45', '89', '90', '26', '28', '14', '83', '56', '21', '37', '50', '51', '99', '100', '32', '47', '112', '4', '73', '75', '81', '63', '39', '27', '113']
b	['61', '80', '22', '107', '95', '43', '20', '60', '96', '0']
g	['122', '123', '6', '29', '10', '11', '13', '71', '115', '116', '117', '77']

Principle Component Analysis

- **Principal component analysis** (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components**.
- The major goal of principal components analysis is to reveal hidden structure in a data set. In so doing, we may be able to

• identify how different variables work together to create the dynamics of the system

• reduce the dimensionality of the data

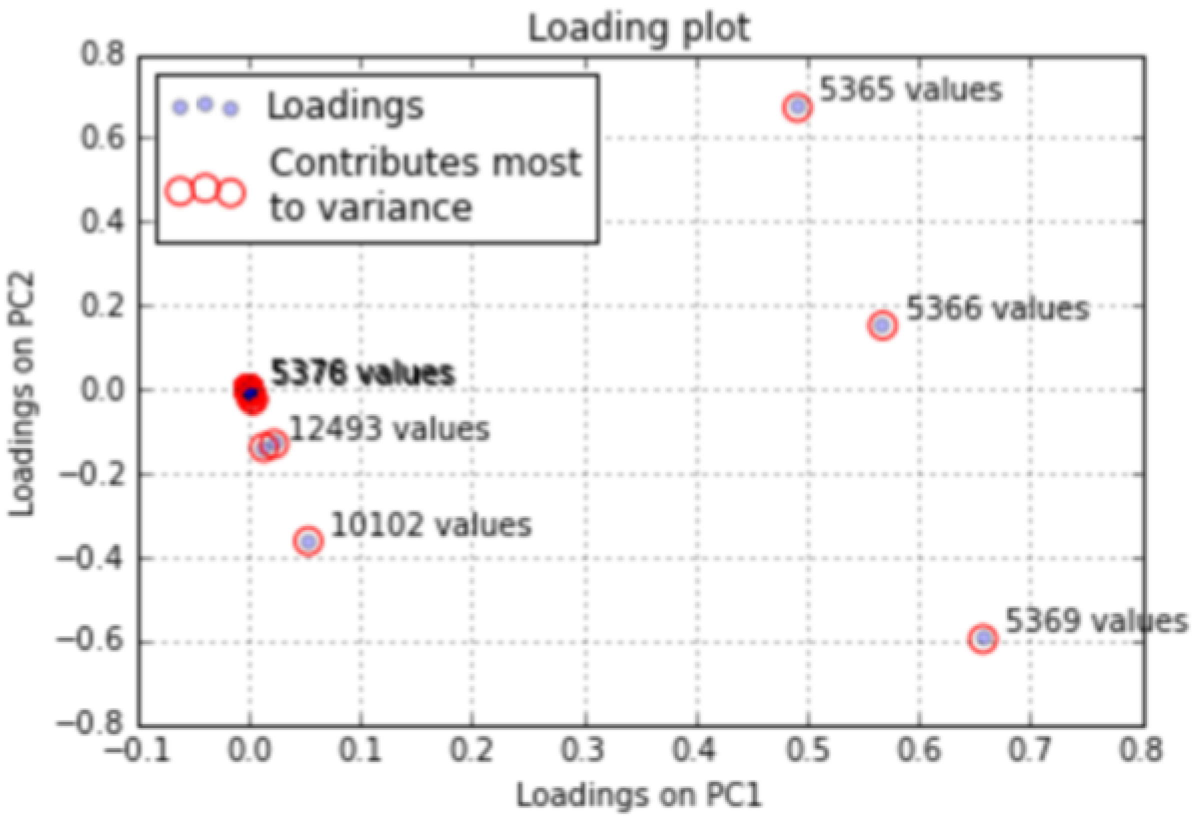
• decrease redundancy in the data

• filter some of the noise in the data

The principal components are orthogonal because they are the eigenvectors of the covariance matrix.

PCA Factors Loading Analysis and Scatter Plot of 1st, 2nd Components of PCA

Factors Loading Score Plot



PCA Scatter Plot

X axis : 1st component of Eigenvector

Y axis: 2nd component of Eigenvector

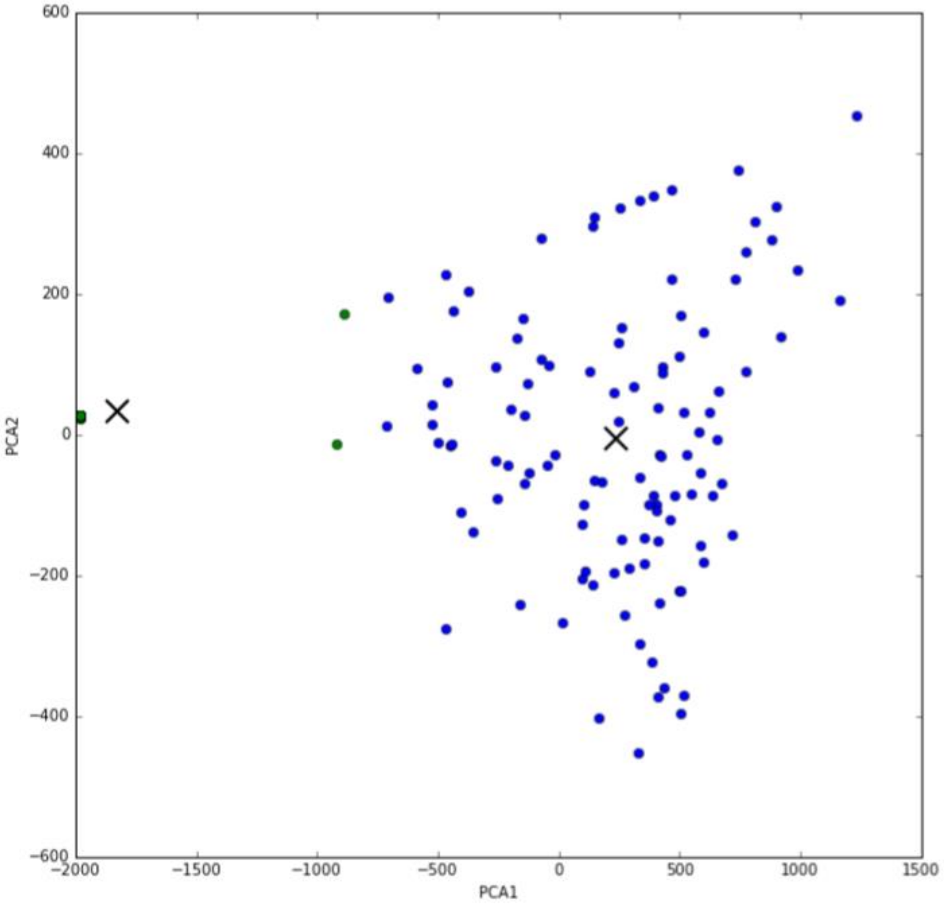
K Means Clustering

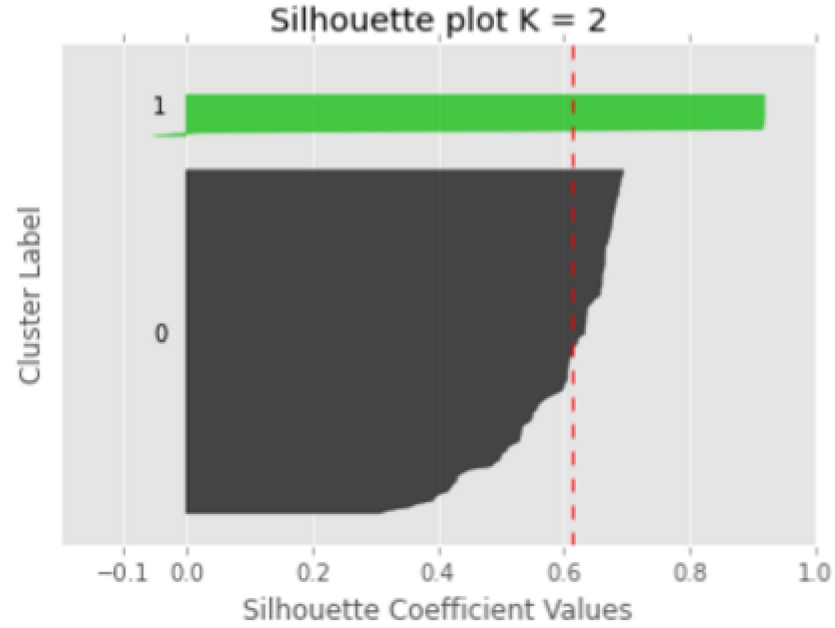
Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS). In other words, its objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

K Means Clustering, K = 2



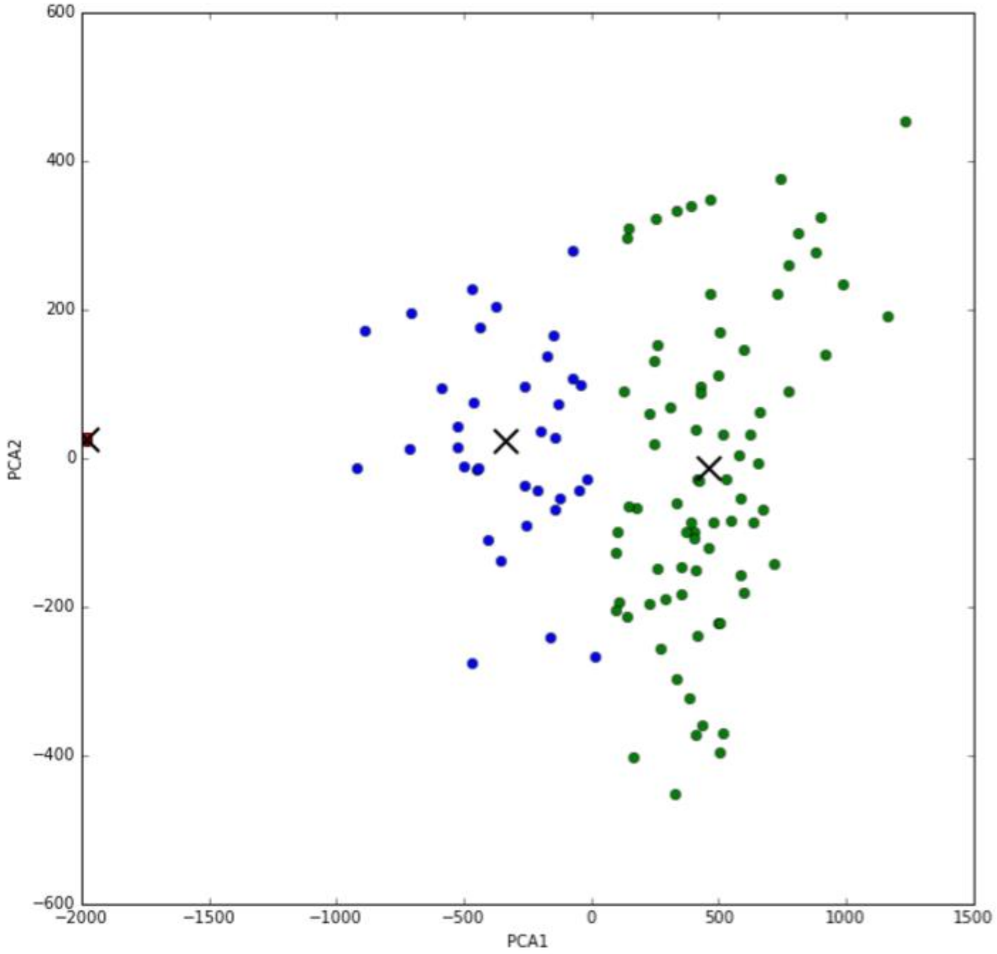


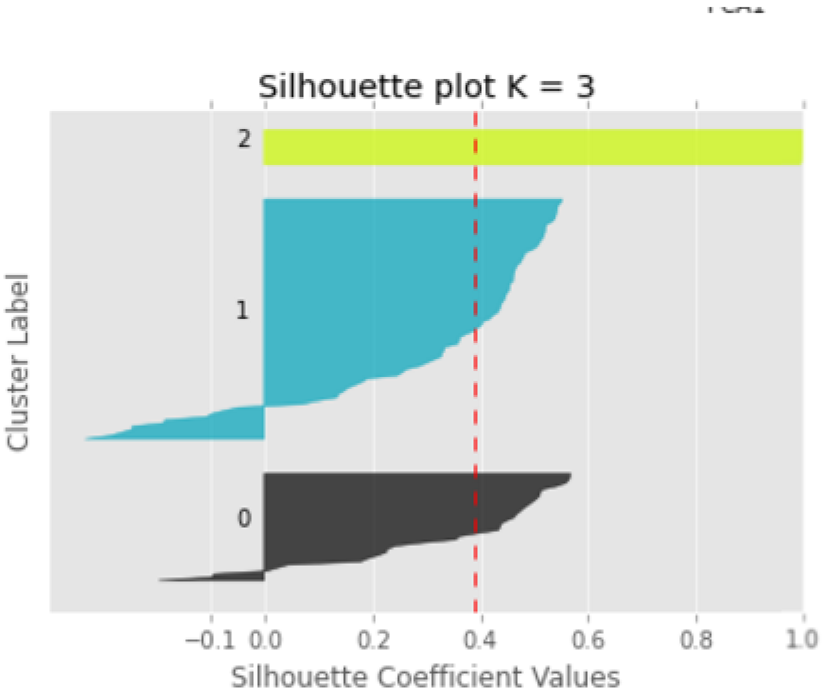
At K = 2, Average Silhouette Score = 0.612675925343

K Means Clustering Subgroups, K = 2

```
['150325-2', '150326-1', '150326-2', '150327-1', '150327-2', '150328-1', '150330-2', '150331-1', '150331-2', '150401-1', '15040
1-2', '150402-1', '150402-2', '150403-1', '150407-2', '150408-2', '150409-1', '150409-2', '150410-1', '150410-2', '150411-1',
'150413-2', '150415-2', '150416-1', '150416-2', '150417-1', '150417-2', '150420-2', '150526-1', '150603-2', '150604-1', '15060
4-2', '150605-1', '150605-2', '150624-2', '150625-2', '150627-1', '150629-2', '150630-2', '150701-2', '150702-1', '150702-2',
'150703-1', '150706-2', '150707-1', '150707-2', '150708-1', '150708-2', '150709-1', '150709-2', '150710-1', '150710-2', '15071
1-1', '150713-2', '150715-2', '150716-1', '150716-2', '150717-1', '150717-2', '150721-1', '150722-2', '150728-2', '150729-1',
'150729-2', '150730-1', '150730-2', '150731-1', '150731-2', '150801-1', '150803-2', '150804-1', '150804-2', '150805-1', '15080
5-2', '150806-1', '150806-2', '150807-1', '150812-2', '150813-1', '150814-2', '150815-1', '150819-2', '150820-1', '150820-2',
'150821-1', '150822-1', '150825-2', '150826-1', '150826-2', '150827-1', '150827-2', '150828-1', '150904-2', '150905-1', '15090
9-2', '150910-1', '150910-2', '150911-1', '150911-2', '150912-1', '150915-1', '150915-2', '150916-1', '150916-2', '150917-1',
'150917-2', '150918-1', '150918-2', '150919-1', '150924-2', '150925-1', '150925-2', '150926-1', '150928-2', '150929-1', '15092
9-2', '150930-1', '150930-2', '151001-1', '151001-2', '151002-1']

['150418-1', '150901-1a', '150901-1b']
```

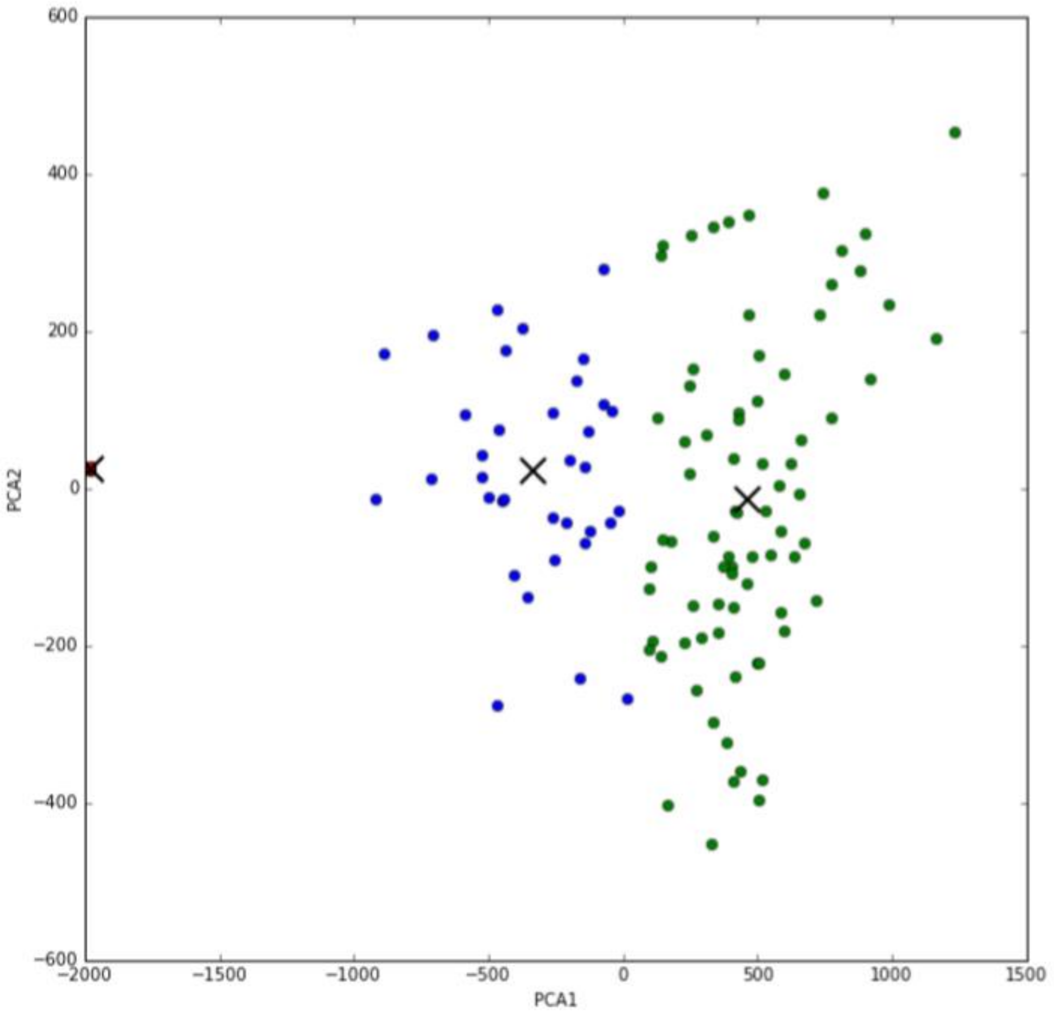
K Means Clustering Subgroups, K = 3

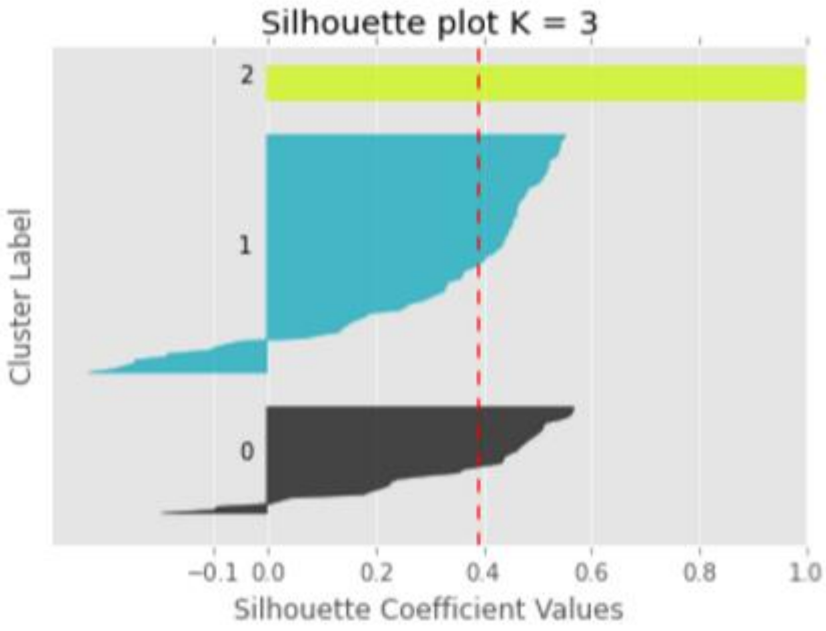




At K = 3, Average Silhouette Score = 0.387455915216

K Means Clustering, K = 3



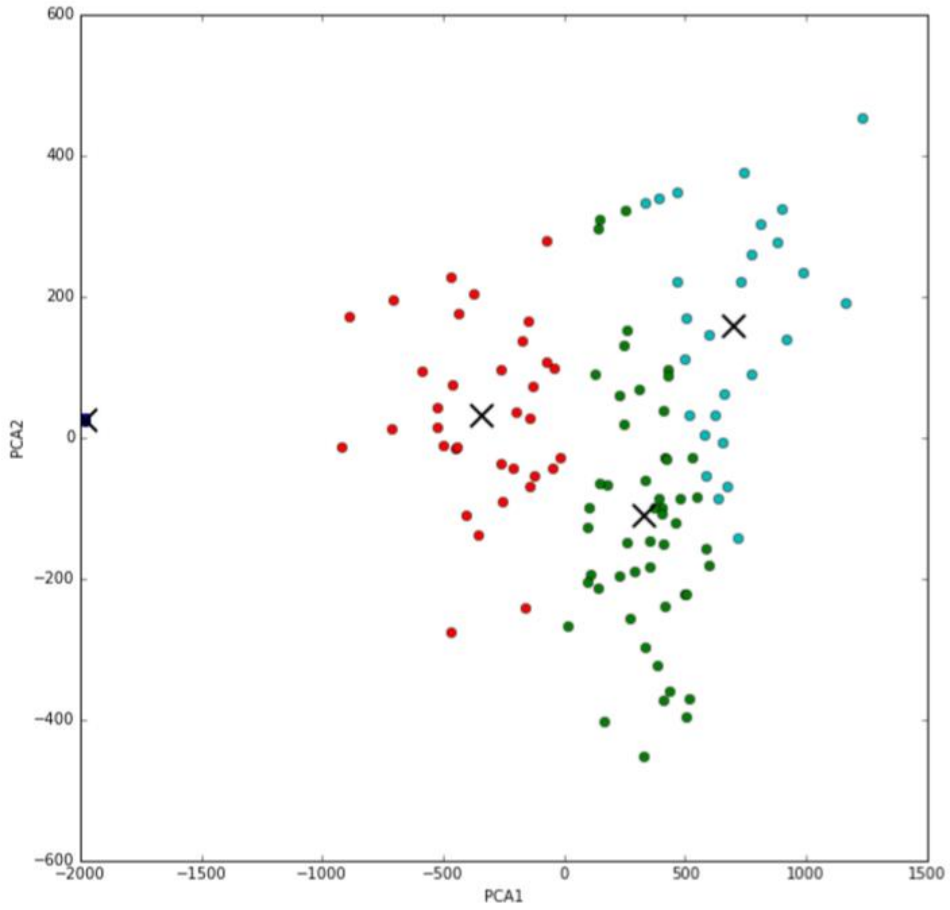


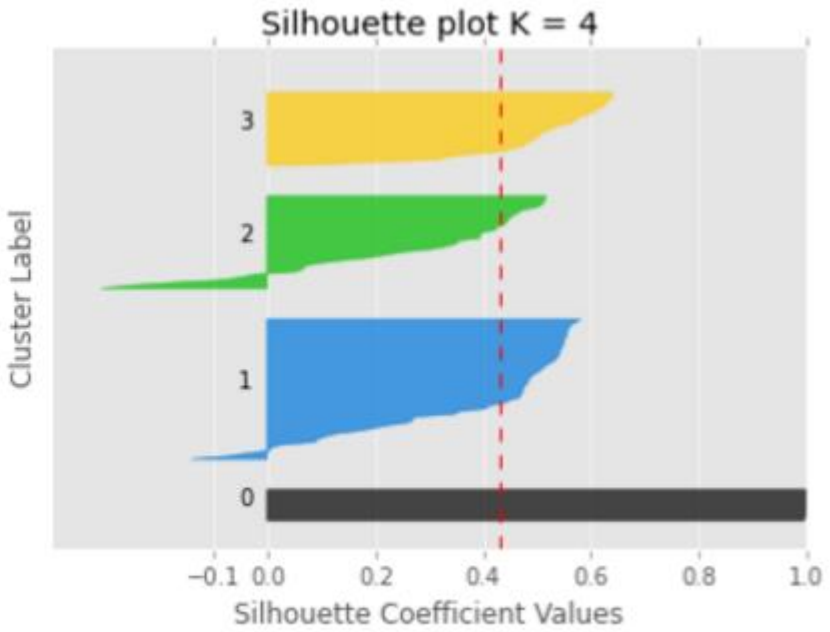
At K = 3, Average Silhouette Score = 0.387455915216

K Means Clustering Subgroups, K = 3

```
['150327-2', '150328-1', '150407-2', '150411-1', '150413-2', '150417-2', '150418-1', '150420-2', '150603-2', '150604-1', '15060
4-2', '150605-1', '150605-2', '150624-2', '150625-2', '150627-1', '150630-2', '150701-2', '150702-2', '150703-1', '150707-1',
'150707-2', '150708-1', '150708-2', '150709-1', '150710-2', '150711-1', '150713-2', '150715-2', '150716-1', '150717-2', '15072
1-1', '150722-2', '150728-2', '150729-1', '150729-2', '150730-1', '150731-2', '150801-1', '150803-2', '150804-1', '150804-2',
'150805-1', '150805-2', '150806-1', '150806-2', '150814-2', '150815-1', '150819-2', '150820-1', '150820-2', '150821-1', '15082
5-2', '150826-1', '150826-2', '150827-1', '150827-2', '150828-1', '150901-1a', '150901-1b', '150904-2', '150905-1', '150909-2',
'150910-1', '150910-2', '150911-1', '150911-2', '150912-1', '150916-1', '150916-2', '150917-1', '150924-2', '150925-1', '15092
5-2', '150926-1', '150929-2', '150930-1', '150930-2', '151001-1']

['150325-2', '150326-1', '150326-2', '150327-1', '150330-2', '150331-1', '150331-2', '150401-1', '150401-2', '150402-1', '15040
2-2', '150403-1', '150408-2', '150409-1', '150409-2', '150410-1', '150410-2', '150415-2', '150416-1', '150416-2', '150417-1',
'150526-1', '150629-2', '150730-2', '150731-1', '150807-1', '150812-2', '150813-1', '150822-1', '150915-1', '150915-2', '15091
7-2', '150918-1', '150918-2', '150919-1', '150928-2', '150929-1', '151001-2', '151002-1']

['150702-1', '150706-2', '150709-2', '150710-1', '150716-2', '150717-1']
```

K Means Clustering, K = 4



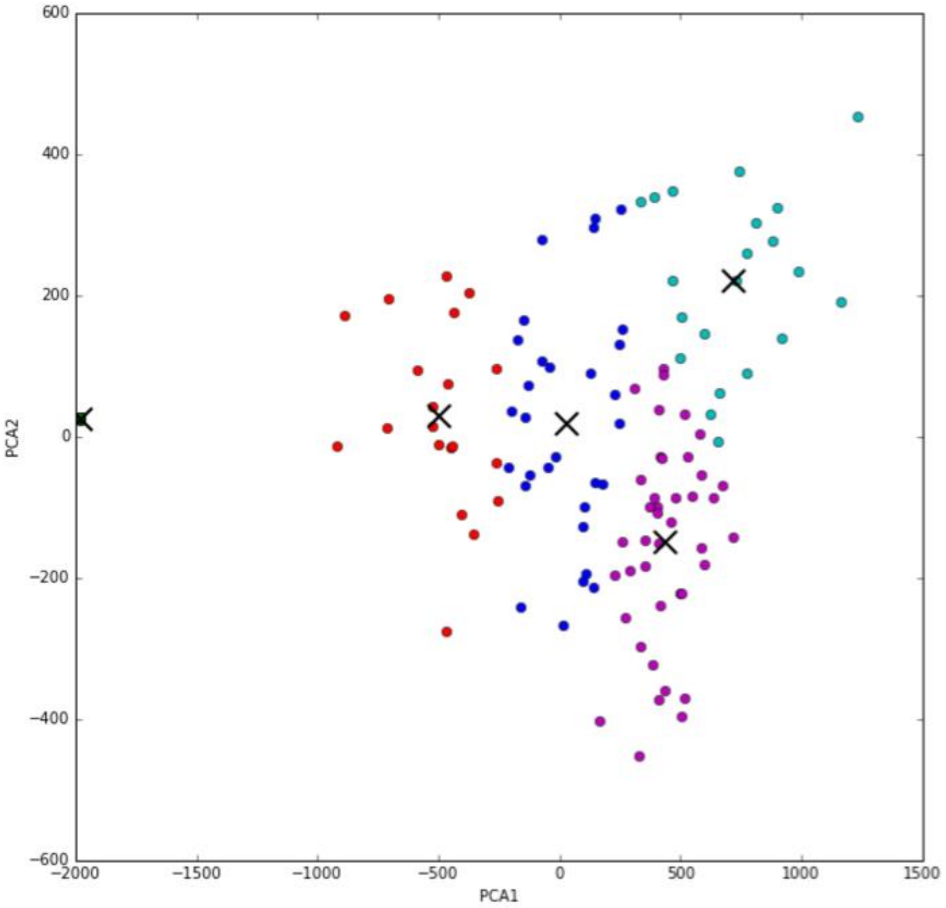


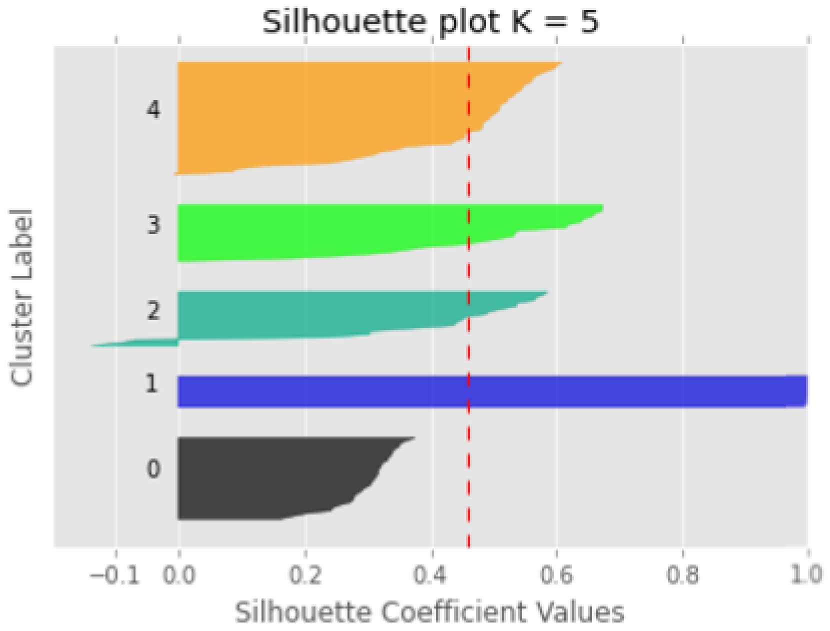
At K = 4, Average Silhouette Score = 0.432004661447

K Means Clustering Subgroups, K = 4

```
['150325-2', '150326-1', '150326-2', '150327-1', '150327-2', '150328-1', '150330-2', '150331-1', '150331-2', '150401-1', '15040
1-2', '150402-1', '150402-2', '150403-1', '150407-2', '150408-2', '150409-1', '150409-2', '150410-1', '150410-2', '150411-1',
'150413-2', '150415-2', '150416-1', '150416-2', '150417-1', '150417-2', '150418-1', '150420-2', '150526-1', '150603-2', '15060
4-1', '150604-2', '150605-1', '150624-2', '150625-2', '150627-1', '150629-2', '150630-2', '150702-1', '150702-2', '150703-1',
'150706-2', '150707-1', '150707-2', '150708-1', '150708-2', '150709-1', '150709-2', '150710-1', '150710-2', '150711-1', '15071
3-2', '150715-2', '150716-1', '150716-2', '150717-1', '150717-2', '150721-1', '150722-2', '150728-2', '150729-1', '150729-2',
'150730-1', '150730-2', '150731-1', '150731-2', '150801-1', '150803-2', '150804-1', '150804-2', '150805-1', '150805-2', '15080
6-1', '150806-2', '150807-1', '150812-2', '150813-1', '150814-2', '150815-1', '150819-2', '150820-1', '150820-2', '150821-1',
'150822-1', '150825-2', '150826-1', '150826-2', '150827-1', '150827-2', '150828-1', '150901-1a', '150901-1b', '150904-2', '1509
05-1', '150909-2', '150911-2', '150912-1', '150915-1', '150915-2', '150916-1', '150916-2', '150917-1', '150917-2', '150918-1',
'150918-2', '150919-1', '150924-2', '150925-1', '150925-2', '150926-1', '150928-2', '150929-1', '150929-2', '150930-1', '15093
0-2', '151001-1', '151001-2', '151002-1']

['150605-2']

['150910-1', '150910-2', '150911-1']

['150701-2']
```

K Means Clustering, K = 5





At K = 5, Average Silhouette Score = 0.459481236495

K Means Clustering Subgroups, K = 5

```
['150325-2', '150326-1', '150326-2', '150327-1', '150328-1', '150330-2', '150331-1', '150331-2', '150401-1', '150401-2', '150402-1', '150402-2', '150403-1', '150407-2', '150408-2', '150409-1', '150409-2', '150410-1', '150410-2', '150411-1', '150413-2', '150415-2', '150416-1', '150416-2', '150417-1', '150417-2', '150418-1', '150526-1', '150603-2', '150604-1', '150604-2', '150605-1', '150605-2', '150624-2', '150625-2', '150627-1', '150629-2', '150630-2', '150701-2', '150702-1', '150702-2', '150703-1', '150706-2', '150707-1', '150707-2', '150708-1', '150708-2', '150709-1', '150709-2', '150710-1', '150710-2', '150711-1', '150713-2', '150715-2', '150716-1', '150716-2', '150717-1', '150717-2', '150721-1', '150722-2', '150728-2', '150729-1', '150729-2', '150730-1', '150730-2', '150731-1', '150731-2', '150801-1', '150803-2', '150804-1', '150804-2', '150805-1', '150805-2', '150806-1', '150806-2', '150807-1', '150812-2', '150813-1', '150814-2', '150815-1', '150819-2', '150820-1', '150820-2', '150821-1', '150825-2', '150826-1', '150826-2', '150827-1', '150827-2', '150828-1', '150901-1a', '150901-1b', '150904-2', '150905-1', '150909-2', '150910-1', '150910-2', '150911-1', '150911-2', '150912-1', '150915-1', '150915-2', '150916-1', '150916-2', '150917-1', '150917-2', '150918-1', '150918-2', '150919-1', '150924-2', '150925-1', '150928-2', '150929-1', '150929-2', '150930-1', '150930-2', '151001-1', '151001-2', '151002-1']

['150822-1']

['150327-2']

['150420-2']

['150925-2', '150926-1']
```