ENGINEERING
TEXAS A&M UNIVERSITY

Multivariate Analysis

ISEN-614 Project: Phase I Analysis

Submitted by: Team #13

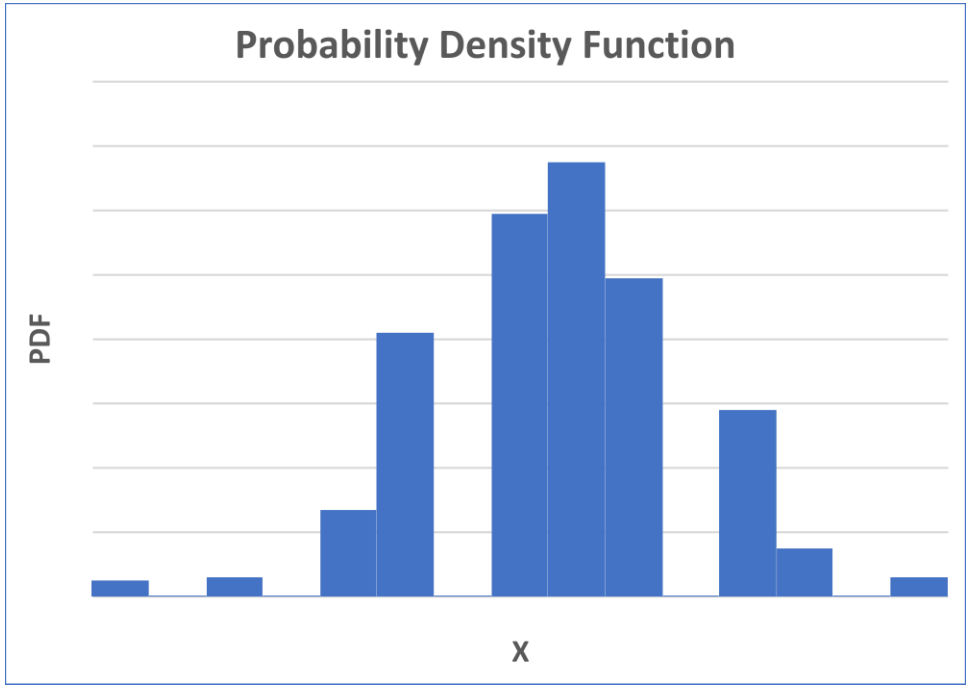
Vraj Thakkar (329006917)
Deep Patel (130004781)
Vraj Patel (829009617)

Approach:

Training Data	The training dataset is for a manufacturing process having 552 data records with 209 features. Thus, values for the 'dimension' and 'sample size' will be 209 and 1 respectively
T^2 & EWMA	Used T^2 and EWMA charts at initial stage to identify the Out of Control points mainly due to Large spikes' and 'small mean shifts' respectively
Selection of FEW	Applied the principle of sparsity to select the variable having the most contribution towards variances following its statement: There are 'Vital Few' instead of 'Trivial Many'
Dimension Reduction	Reduced dimensions using the three popular methods namely Pareto Chart, Scree plot and Minimum Description Length
Again T^2 & EWMA	Used T^2 to identify the Out of Control points due to 'Large spikes after dimension reduction' and EWMA for points due to 'Small mean shifts after removal of large spikes data points'
Compare & Establish	Compared the results obtained by observing the training data before vs. after dimension reduction and established a selection criterion for the procedure
Parameters Finalization	Following the selection criteria & performing the procedure until no out of control points left, the Mean Vector & the Variance matrix were finalized which can be used for phase II analysis

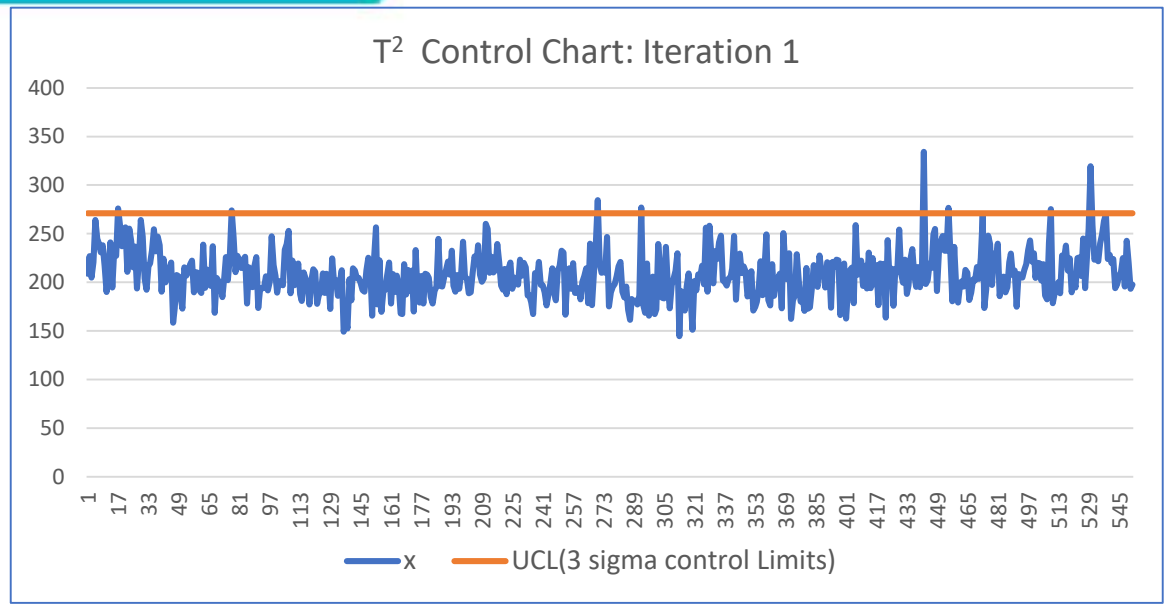
Training Data:

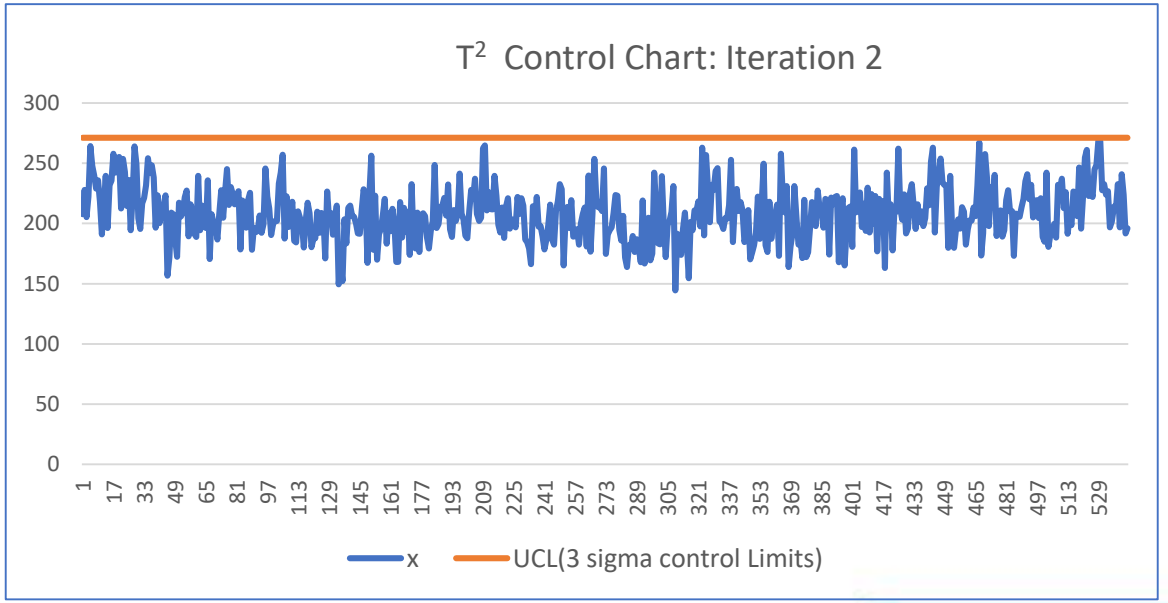
- The training dataset is for a manufacturing process having 552 data records with 209 values per each
- Values for the ‘dimensions’ and ‘sample size’ will be 209 and 1 respectively
- The probability density function (pdf) of the data is close to Gaussian Distribution
- We plotted the pdf for 20 variables and found that it approximately follows the normal distribution and the plot is shown
- This means that the data set is a continuous data set and the control chart methods that can be used are limited to T^2, m-CUSUM and m-EWMA.



T^2 Chart:

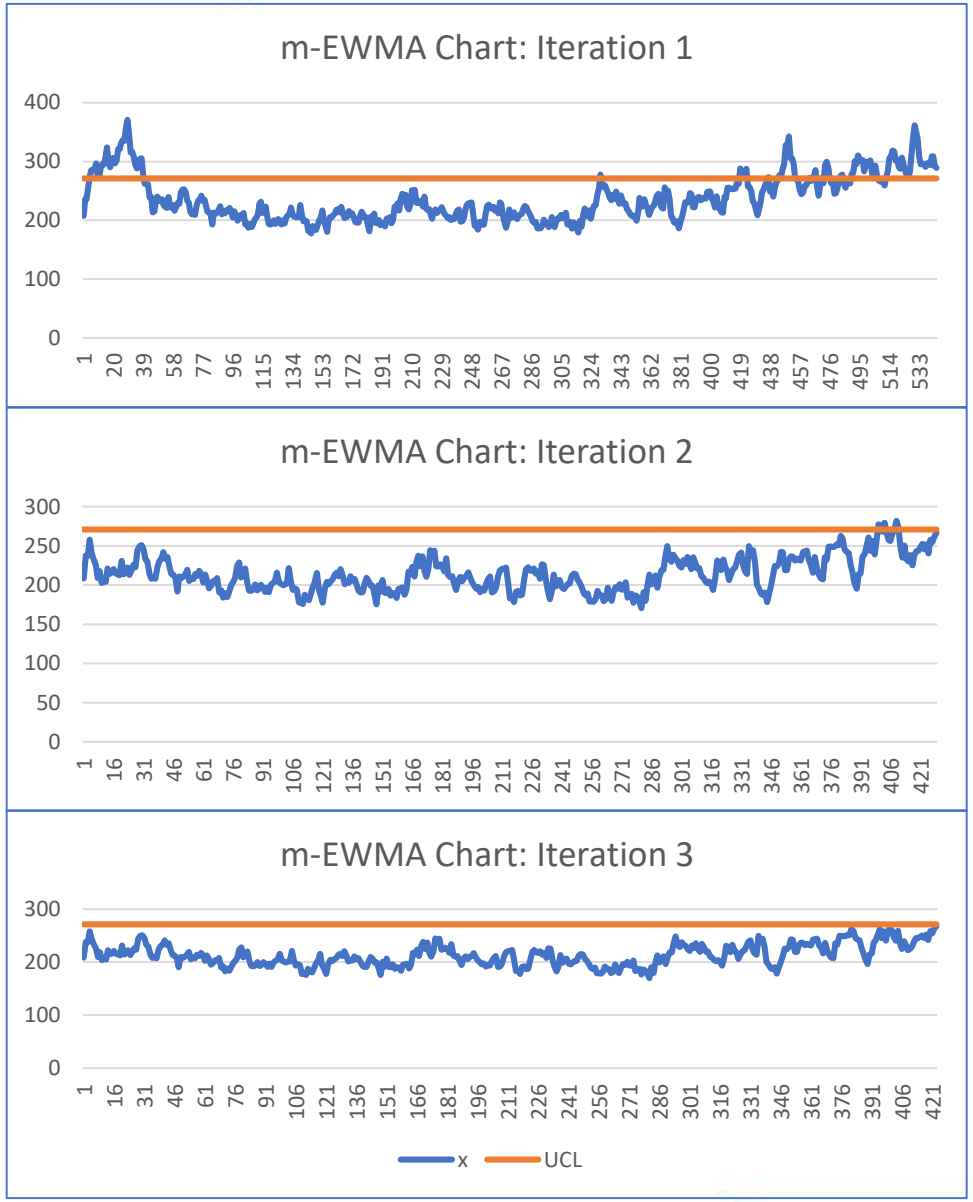
- The reasons for using the T^2 chart are: To evaluate the behavior of the simulated dataset without losing any information and easy identification of Large spikes due to weighted effect of each Principle Components
- The Dataset initially consists of 552 data records with 209 values in each record
- The parameters used for evaluating T^2 Chart are:
  $\alpha = 0.0027(3\sigma$ control limits) and $p = 209$
- The UCL for T^2 Chart: $\chi^2_{1-0.0027}(209) = 271.012$
- Performing 1st iteration, total 8 datapoints were found to be out of control in T^2 Chart. Removing the detected 8 out of control points, performing 2nd iteration produced an in-control dataset of 544 data records and 209 values in each record.
- Although T2 is a good control chart for successfully eliminating large spikes and mean shifts, small mean shifts cannot be detected.





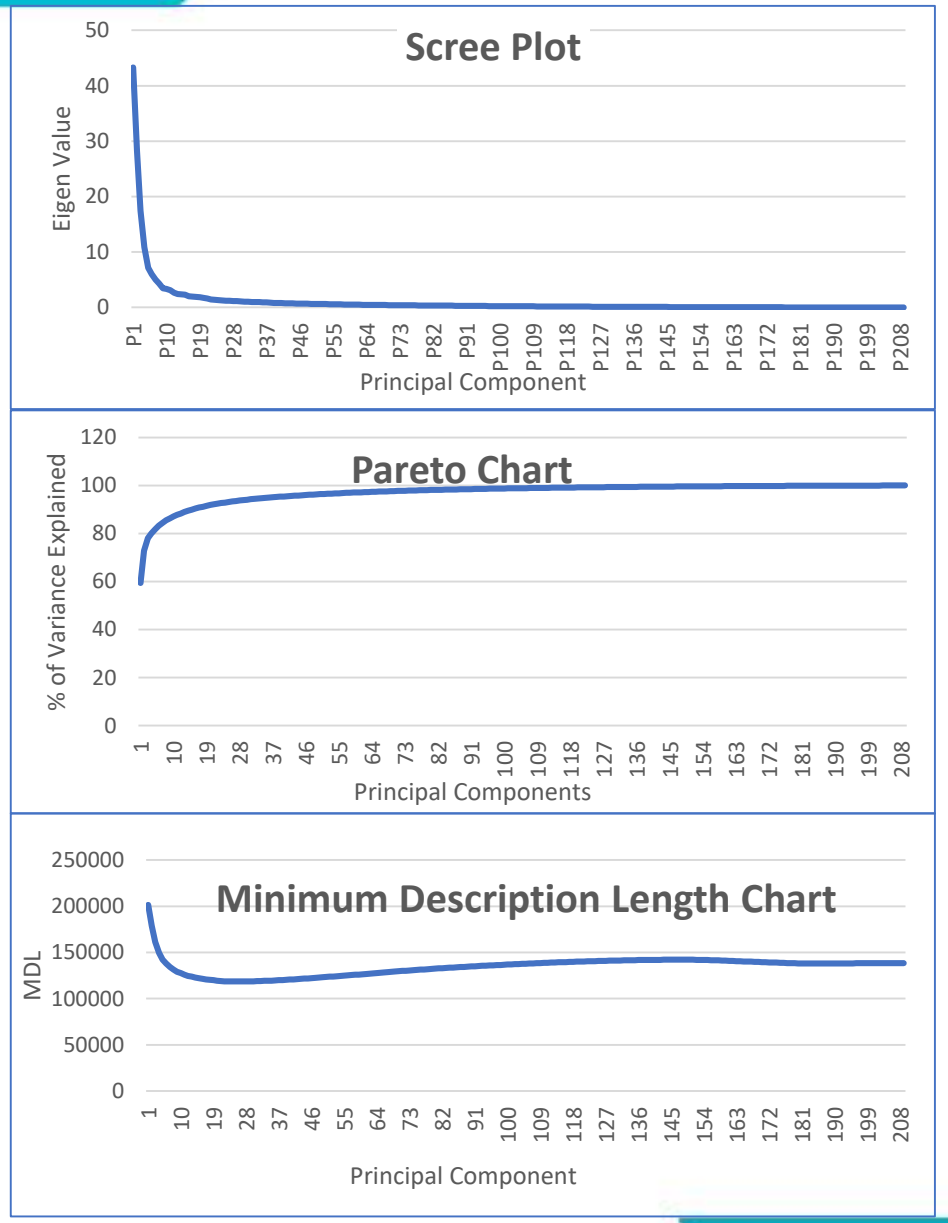
m-EWMA Chart:

- The reasons for using the m-EWMA chart are: To detect the Out of control points due to small mean shifts which T^2 chart is unable to detect
- The Dataset initially consists of 544 data records with 209 values in each record
- The parameters used for evaluating m-EWMA Chart are: $\alpha = 0.0027(3\sigma$ control limits) and $p = 209$
- Since under H_0, the EWMA statistic follows Chi-square distribution, we can calculate an approximated UCL for m-EWMA chart as: $\chi^2_{1-0.0027}(209) = 271.012$
- Performing 1st iteration, total 115 datapoints were found to be out of control and were removed. Performing 2nd iteration, total 6 datapoints were found to be out of control and were removed.
- Finally, 3rd iteration produced an in-control dataset of 423 data records and 209 values in each record.



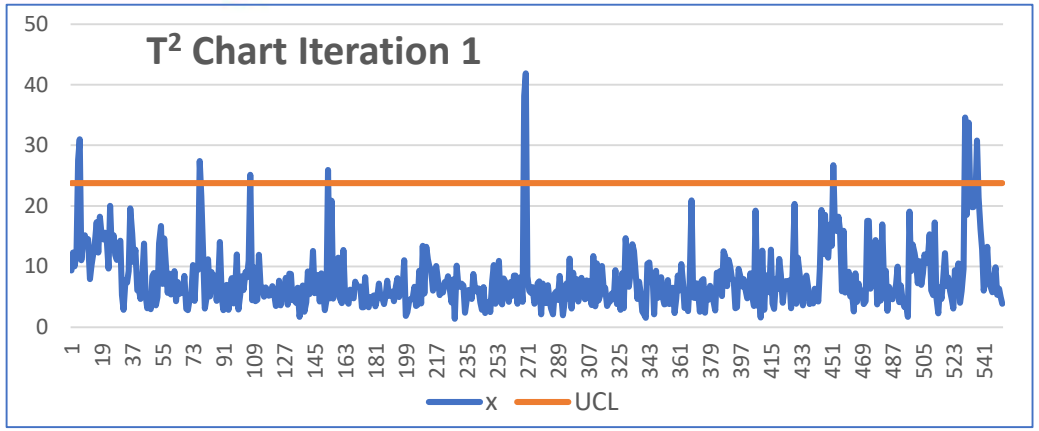
PCA for Dimension Reduction:

- We considered the given training dataset as a standardized dataset, assuming that the respective physical unit of the variables have no relative consideration or are same.
- Than applied PCA on the standardized dataset, to reduce the dimension and nullify the aggregated noise effect which can overwhelm the signal effects i.e. the effect of 'Curse of Dimensionality'.
- Minimum description length indicates that 26 Pcs can be considered as the 'vital few' but this is still significantly large number of PCs
- Calculating principal components with the help of variance-covariance matrix obtained earlier, [E λ] pairs were formed. Than we arranged these eigen values in a descending order and generated a scree plot.
- It is clearly visible that there is an elbow bend formation at 6th PC, but it constitutes to approx. 60% of the total variance as per Pareto Chart.
- After referring the generated Scree plot, Pareto chart and MDL charts, we decided to reduce the data dimension to 552 data records and 8 variables which accounts to 85% of the total variance.

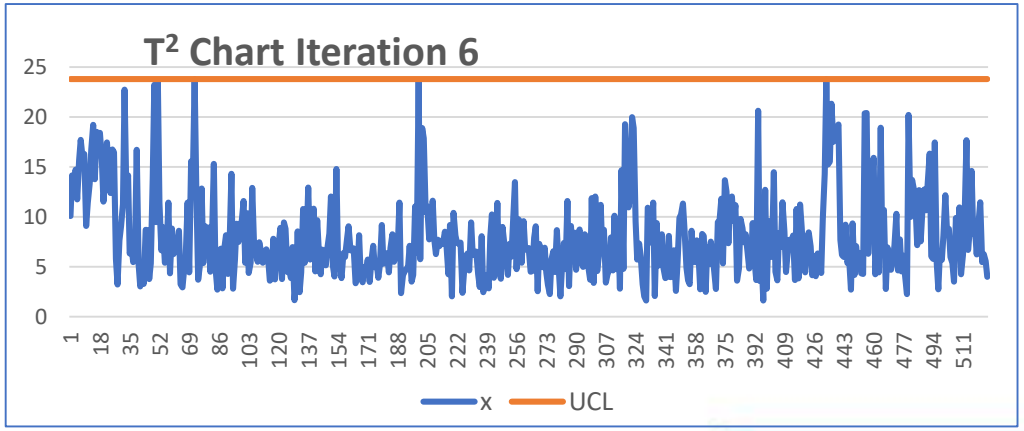


T^2 Chart post Dimension Reduction:

- Performing dimension reduction using PCA, later we used T^2 Chart again for identifying any large spikes (change in mean or variance) leading to out of control points
- The Dataset initially consists of 552 data records and 8 values per each record now
- The parameters used for evaluating T^2 Chart are:
  $\alpha = 0.0027(3\sigma$ control limits), $p = 8$
- The UCL for T^2 Chart is: $\chi^2_{1-0.0027} = 23.7927$, found by Interpolating
- After performing 6 iterations, total **27** datapoints were found to be out of control limits in total and were removed assuring the elimination of large spikes present in the dataset
- Use of Multiple Univariate Charts was rejected because of two reasons: It will be time consuming to generate 8 different univariate plots for the analysis and still will not be able to detect changes due to correlation. Also, there is a possibility that α and β may get inflated.

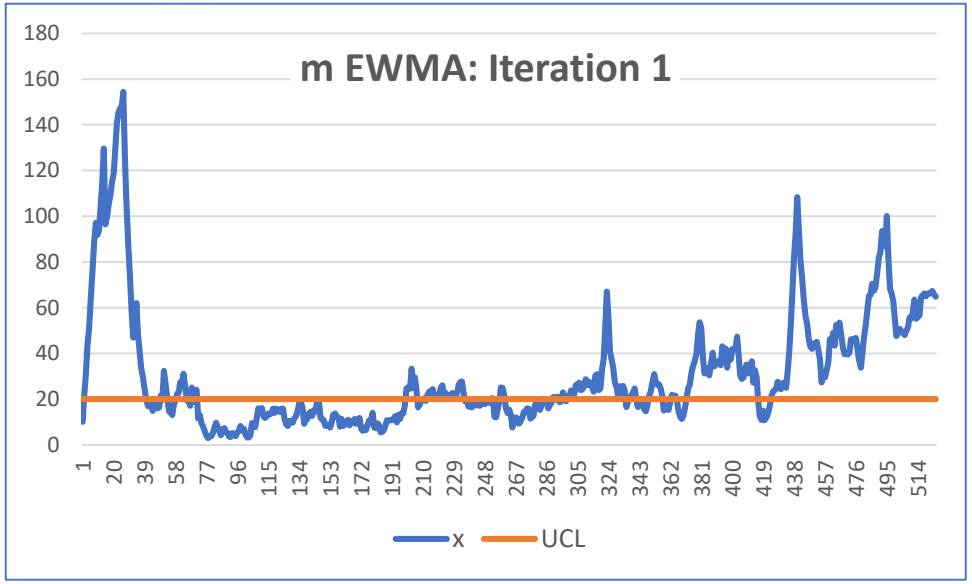


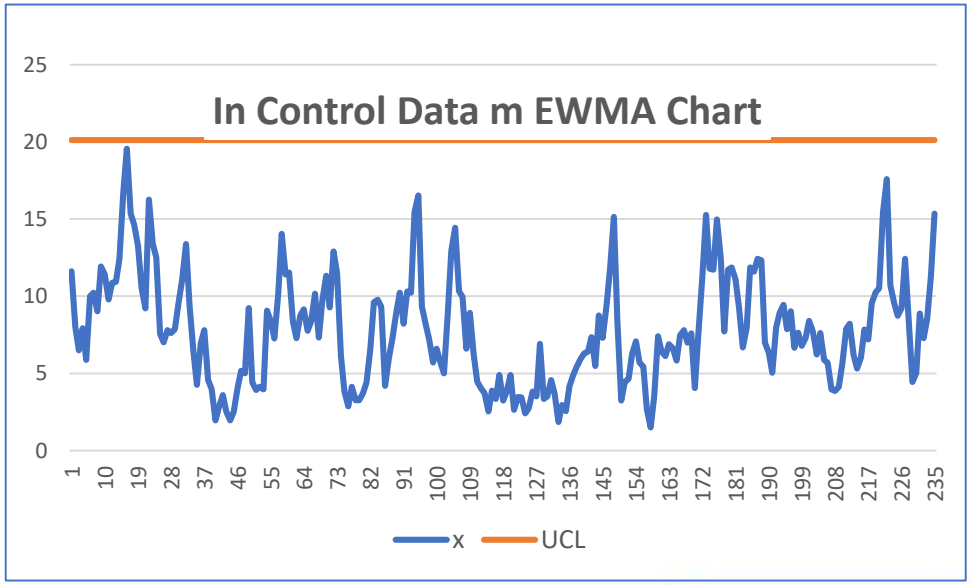
Iteration Number	Number of Out of Control Points	Number of Data Records Considered
Iteration 1	11	552
Iteration 2	8	541
Iteration 3	6	533
Iteration 4	1	527
Iteration 5	1	526
Iteration 6	0	525



m-EWMA Chart post Dimension Reduction:

- After eradicating the large spikes and mean shifts with the help of T^2 chart post dimension reduction, we now use m-EWMA seeking elimination of small mean shifts in the dataset.
- The Dataset initially consists of 525 data records and 8 variables
  The parameters for m-EWMA are : ARL_0= 370 and α = 0.0027
- With the aim of detecting the small mean shifts, which can be detected by a chart having longer memory, we chose value of 'r' to be 0.1.
- The value of UCL: 20.11067 is selected for small mean shift detection as per the interpolation and shown in table
- After Plotting m-EWMA chart for the dataset, in first iteration we found 290 data records out of control and were removed
- In second iteration, it produced an in-control dataset of total 235 datapoints and 8 values in each record.





Comparison, establishment & Finalization:

- Comparison for the “Control charts without Dimension reductions” and “Control charts with Dimensions reductions” is shown in table
- We found very less out of control points in former method than latter and hence can say dimension reduction enhances detection reducing the noise i.e. less contributing variables
- Established selection says that monitoring the given manufacturing process do not require to keep a watch on all 209 values but instead can establish a reasonable check using just 8 values providing 85% of the explanation for the process variability
- This makes easy interpretability for the reason causing out of detection points easier
- Finalized values for the ‘Mean vector’ and ‘Variance-Covariance Matrix’ is shown in table

Methods	Total out of control points	Final size of the dataset	Dimensions
T^2	8	544	544 x 209
m-EWMA	121	423	423 x 209
Final Output:	**129**	**423**	**423 x 209**
T2 post PCA	27	525	525 x 8
M-EWMA post PCA	290	235	235 x 8
Final Output post PCA:	**317**	**235**	**235 x 8**

Incontrol Parameters

Mean

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
-13.6792	-14.5815	10.36355	-1.15733	1.262921	-1.30874	3.521522	-0.0785

Covariance Matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
PC1	4813.82	507.7968	577.9889	93.74237	-48.3542	20.50776	43.42157	207.9479
PC2	507.7968	1116.787	381.3418	-18.3061	82.29268	-37.0603	80.25456	18.53225
PC3	577.9889	381.3418	1038.464	98.59908	-29.9471	-62.193	-15.135	64.86246
PC4	93.74237	-18.3061	98.59908	377.1398	32.81779	83.1876	23.7011	5.18534
PC5	-48.3542	82.29268	-29.9471	32.81779	429.2751	-15.5063	-41.3048	-24.1288
PC6	20.50776	-37.0603	-62.193	83.1876	-15.5063	263.8562	40.96823	9.529688
PC7	43.42157	80.25456	-15.135	23.7011	-41.3048	40.96823	274.1759	-7.90781
PC8	207.9479	18.53225	64.86246	5.18534	-24.1288	9.529688	-7.90781	151.0356

Conclusions:

- Identifying and after removing all the out of control data points using the approach described in slide 2, if any more data points turn out to be out of control, one can infer the process to be out of control

- Comparing the results obtained by T^2 and m-EWMA charts before and data reduction one can conclude that use of PCA can help us eliminate the aggregate noise and eliminate more out of control data records

- These control charts can be used to determine if the manufacturing process in question meets predefined quality standards as per the established Vital few variables by performing Phase I analysis

- By this project, we gained more pragmatic knowledge about the different methods and techniques to use them to perform Multivariate Phase-I Analysis and finalizing parameters for the Phase-II

- Finally, it can be concluded that, the selection of control charts and other supplemental techniques like PCA dimension reduction plays an important role in industrial processes monitoring and quality control