# **URE Review - I**

# Title: Air quality forecasting using statistical and machine learning methods

**Methods for which experiment is completed:**

* Moving Average (MA)
* Weighted Moving Average (WMA)
* Exponential Smoothing (double/triple)

## **Experiment 1: Missing values imputation**

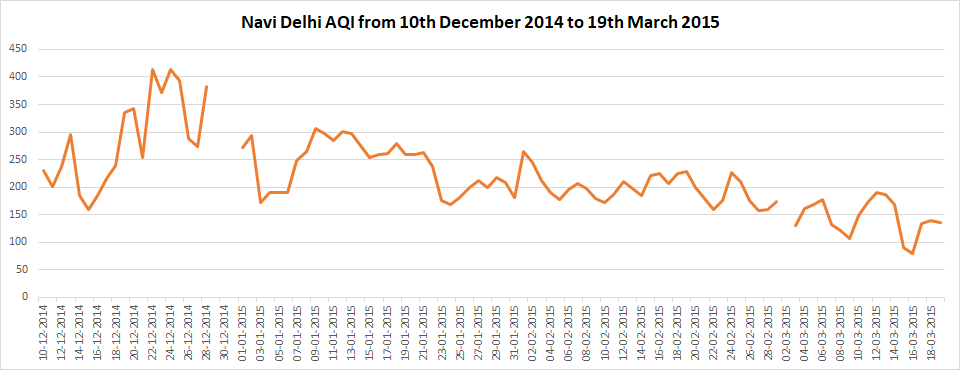


Figure 1. Trendline with missing values

|  |  |
| --- | --- |
| Total observations | 2476 |
| Missing observations out of total | 142 |
| **Percentage of missing observations out of total** | **5.7** |

There were several missing values found during investigation. The table above shows the basic stats about the dataset. Missing values are imputed with results of Exponential Smoothing method using Excel software.

## **Experiment 2: Moving Average**

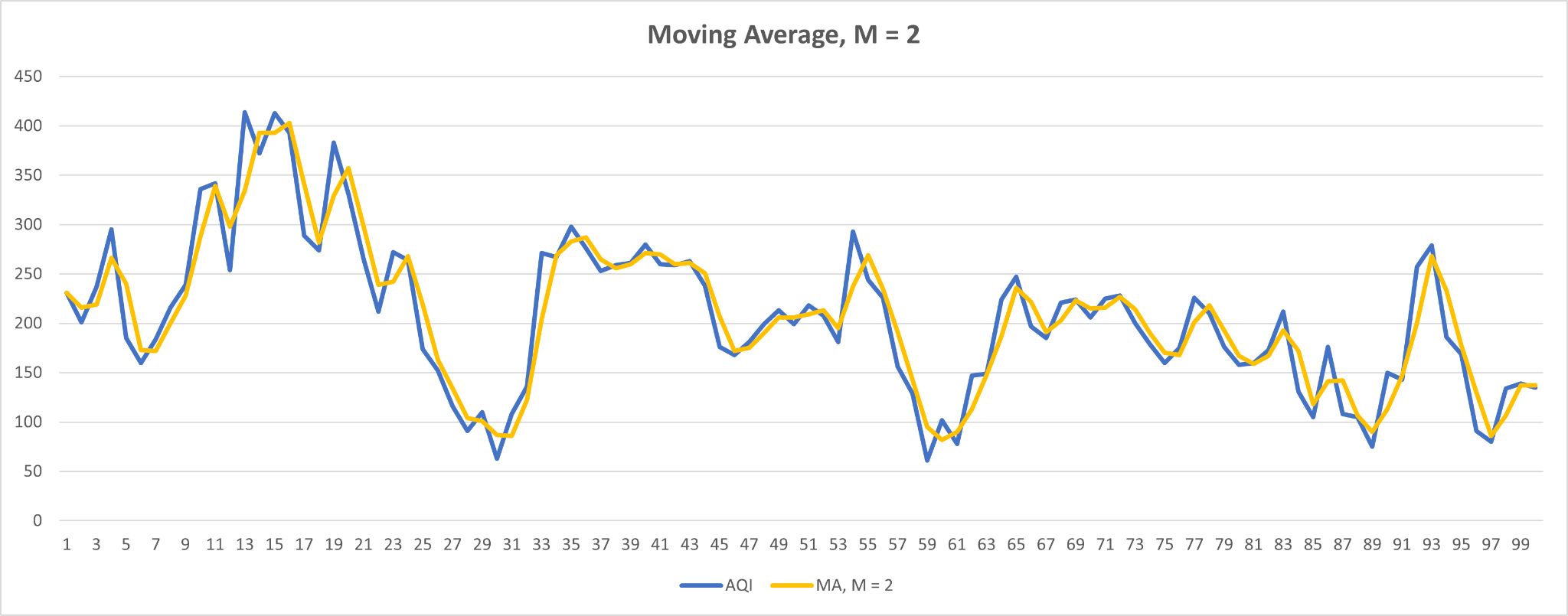


Figure 2. Moving average, M = 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Mean estimator** | **MA, M = 2** | **MA, M = 3** | **MA, M = 5** |
| **Mean Squared Error** | 5896 | **631** | 1274 | 2349 |

## **Experiment 3: Exponential Smoothing**

Figure 3. Exponential smoothing with alpha = 0.8

Following table is the comparisons of various moving average estimator configurations. The value for ɑ = 0.8 gives the minimum Mean Squared Error (MSE) value. Still, the MSE achieved in the MA method is far better.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ɑ value** | **0.2** | **0.3** | **0.5** | **0.8** | **0.9** | 1 |
| **MSE** | 3955 | 3411 | 2834 | 2522 | **2519** | 2558 |

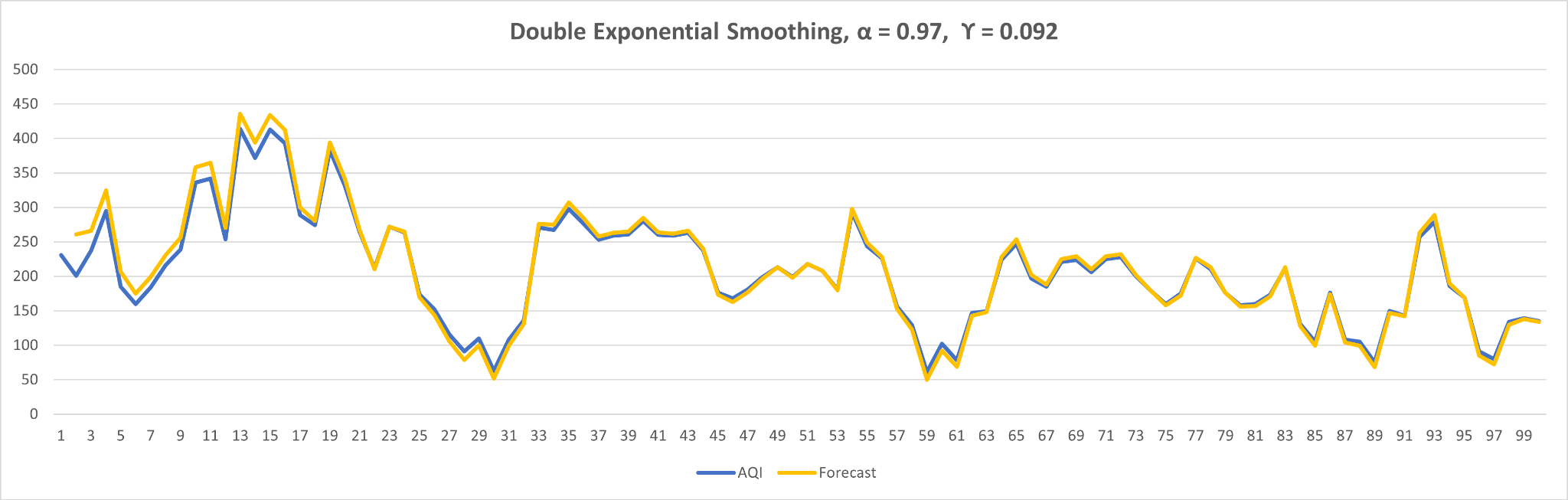
## **Experiment 4: Double Exponential Smoothing**

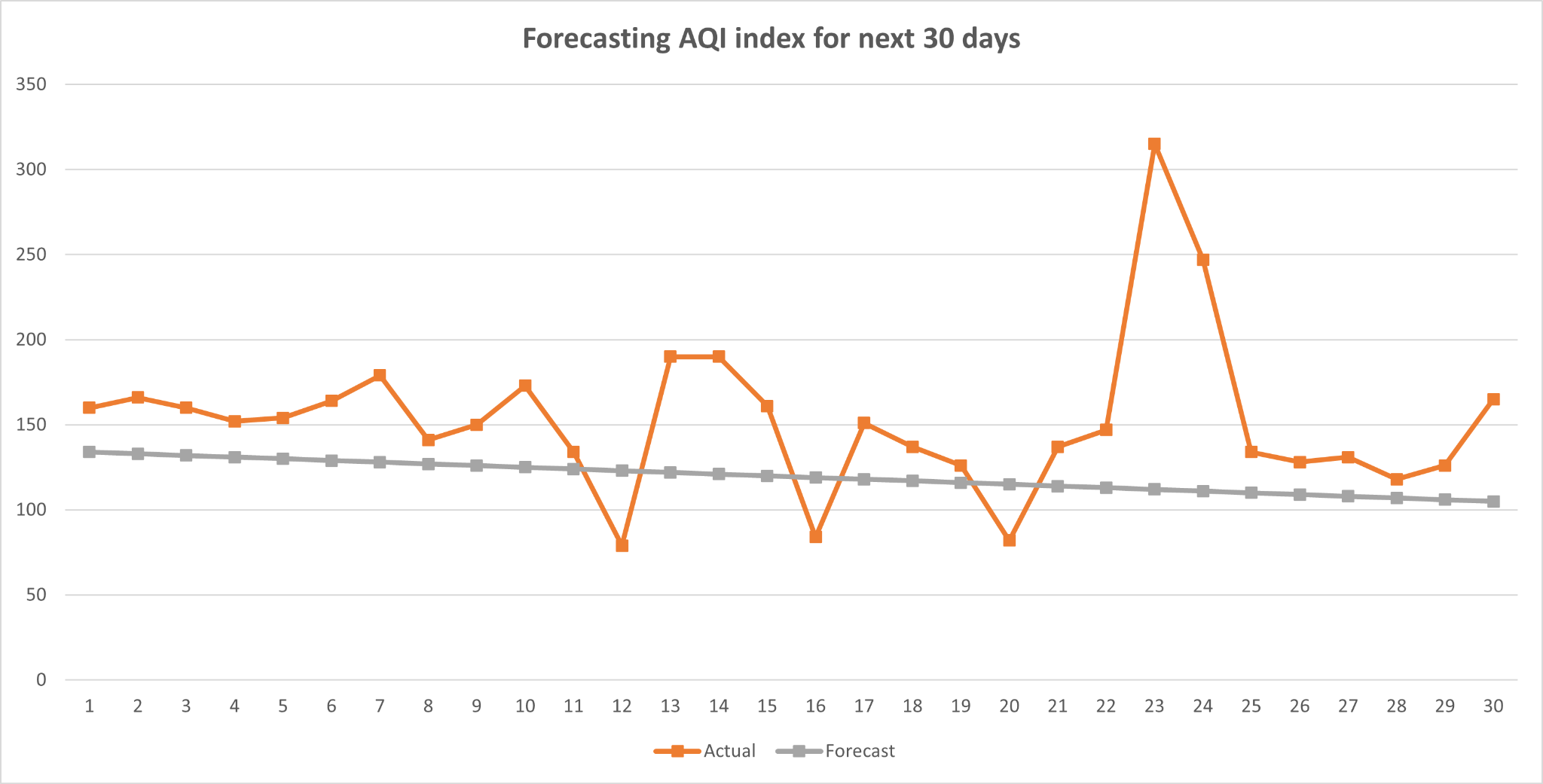
Following table is the experimentation for finding Mean Squared Error using all possible values of ɑ and Ƴ. The value pair for (ɑ, Ƴ) = (1, 0.1)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ɑ \ Ƴ** | **0.0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
| **0.0** | 3131318 | \* | \* | \* | \* | \* | \* | \* | \* | \* | \* |
| **0.1** | 86205 | 9433 | 7160 | 7340 | 8036 | 9306 | 8238 | 8862 | 8753 | 10265 | 10581 |
| **0.2** | 24739 | 4103 | 3788 | 3965 | 4098 | 4569 | 4420 | 4430 | 4455 | 4553 | 4471 |
| **0.3** | 11824 | 2443 | 2338 | 2426 | 2416 | 2585 | 2534 | 2611 | 2687 | 2758 | 2766 |
| **0.4** | 6836 | 1516 | 1443 | 1492 | 1546 | 1624 | 1639 | 1695 | 1739 | 1785 | 1838 |
| **0.5** | 4461 | 962 | 913 | 947 | 981 | 1079 | 1096 | 1172 | 1239 | 1318 | 1382 |
| **0.6** | 2995 | 604 | 569 | 609 | 665 | 769 | 813 | 887 | 969 | 1084 | 1172 |
| **0.7** | 2125 | 370 | 358 | 412 | 478 | 586 | 667 | 773 | 886 | 1017 | 1157 |
| **0.8** | 1542 | 216 | 233 | 310 | 402 | 530 | 644 | 790 | 949 | 1137 | 1327 |
| **0.9** | 1170 | 145 | 192 | 293 | 421 | 585 | 747 | 951 | 1171 | 1435 | 1732 |
| **1** | 927 | **132** | 217 | 358 | 535 | 755 | 994 | 1294 | 1640 | 2062 | 2573 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Ɑ \ Ƴ** | **0.090** | **0.091** | **0.092** | **0.093** | **0.094** | **0.095** | **0.096** | **0.097** | **0.098** | **0.099** | **0.1** |
| **0.95** | 129 | 131 | \* | 132 | 133 | 132 | 128 | \* | 130 | 131 | 134 |
| **0.96** | 130 | \* | 128 | 130 | \* | 131 | 126 | \* | 129 | 130 | 136 |
| **0.97** | 130 | 131 | 125 | \* | 128 | \* | 129 | \* | 128 | 129 | 130 |
| **0.98** | 126 | \* | \* | 127 | 128 | 127 | 129 | 128 | 132 | 128 | 129 |
| **0.99** | 125 | 127 | \* | 128 | 131 | 132 | 129 | \* | 128 | 128 | 131 |
| **1** | 128 | \* | 131 | 132 | 131 | \* | 127 | 129 | 132 | \* | 2573 |

Refining the values of ɑ and Ƴ to find minimum MSE

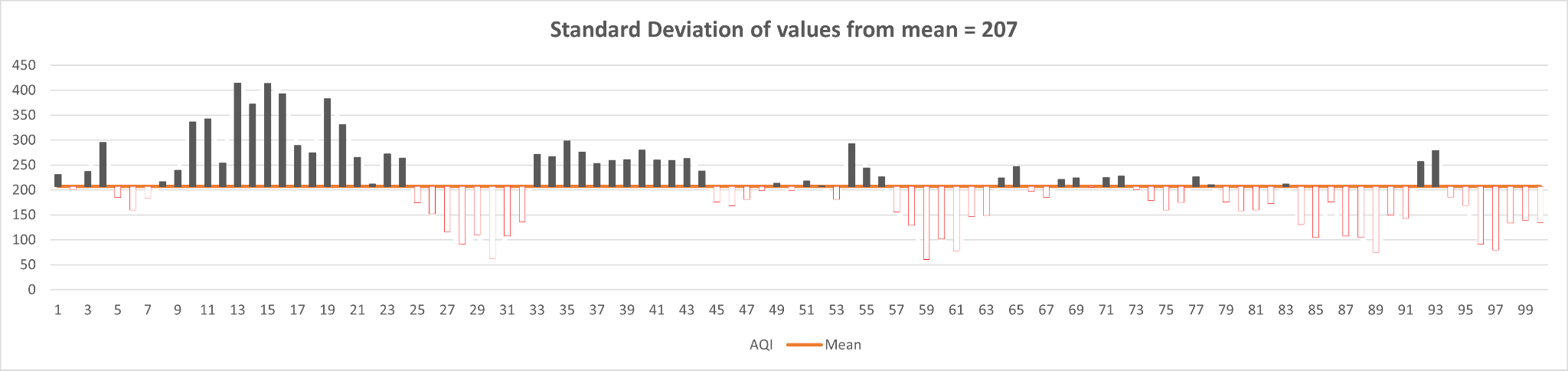




Even though DES trendline matched exactly to the originally observed trendline, it didn’t perform well during forecasting. The MSE = 3148 which was 125 during training.

## **Experiment 5: Standard Deviation**

We calculate the standard deviation of the given input data.



Standard deviation is widely used in weather forecasting to understand how much variation exists in daily and monthly temperatures in different cities. For example: *A weatherman who works in a city with a small standard deviation in temperatures year-round can confidently predict what the weather will be on a given day since temperatures don’t vary much from one day to the next.*

## **Findings and future work**

1. The given trendline is almost non-linear in its nature because of varying effect of hidden parameters on the AQI.
2. Compared statistical methods are alone not sufficient to get a good result i.e., accurate prediction or almost full accuracy.
3. The compared or yet to compare methods can be used to know the trend function or the nature of trend in terms of mathematical model.
4. The RNN or more sophisticated Machine Learning algorithms are necessary in order to find out non-linear relationship among the data points.
5. Standard deviation is useful to minimize the errors in prediction. Higher deviation value will eventually affect the prediction and trendline.