

Anomaly Detection (Web Portal)

**A Final Report for Software Engineering Project
Master of Engineering in Information Technology**

**Submitted By: Umair Junaid 1144240
 Waqas Ahmad 1144211**

**Fachbereich 2:
Informatik und Ingenieurwissenschaften
Frankfurt University of Applied Sciences**

Frankfurt, Germany

22 April 2016

Table of Contents

1.	Background	3
1.1.	An Introduction to Anomaly Detection	3
1.2.	Popular Methods	4
1.3.	Applications.....	4
1.4.	K-means Clustering	4
1.5.	K-means Algorithm	5
1.6.	Details of K-means	6
2.	Project Objective	6
3.	Project Requirement.....	6
4.	Web Portal Tabs	7
4.1.	Home	7
4.2.	Raw Data 2D.....	8
4.3.	K-means 2D.....	8
4.4.	Raw Data 3D.....	9
4.5.	K-means 3D.....	9
4.6.	Stats.....	10
5.	Conclusion.....	10

Document Version Control

Date	Version	Change Reference	Author
20.04.16	V1.0	Initial Draft	Umair & Waqas
21.04.16	V2.0	Final Draft	Umair & Waqas

1. Background

1.1. An Introduction to Anomaly Detection ¹

Anomaly detection denotes to the problem of finding patterns in data that do not follow expected behavior. These non-conforming configurations are often mentioned as anomalies, conflicting observations, exceptions, outliers, deviations, abnormalities, contaminants in different application domains. Of these, anomalies and outliers are two terms used most commonly in the context of anomaly detection; sometimes interchangeably. Anomaly detection finds broad use in a wide selection of applications such as fraud detection for credit cards, insurance or health care, invasion detection for cyber-security, fault recognition in safety critical systems, and military surveillance for enemy activities.

The prominence of anomaly detection is due to the fact that anomalies in data interpret to important (and often grave) actionable information in a wide variety of application domains. For example, an irregular traffic pattern in a computer network could mean that a hacked computer is sending out important data to an unsanctioned destination. An anomalous MRI image may indicate presence of harmful tumors. Anomalies in credit card transaction data could show credit card or identity theft or anomalous readings from a space craft sensor could signify a fault in some section of the space craft.

Precisely in the context of abuse and network intrusion detection, the interesting objects are often not rare objects, but unanticipated bursts in activity. This pattern does not follow the common statistical meaning of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been collected appropriately. As an alternative, a cluster analysis algorithm may be able to spot the micro clusters formed by these irregularities.

Three main types of anomaly detection techniques exist. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the supposition that most of the instances in the data set are normal by looking for instances that seem to be acceptable least to the remainder of the data set. Supervised anomaly detection techniques need a data set that has been categorized as "normal" and "abnormal" and includes teaching a classifier. Semi-supervised anomaly detection techniques build a model demonstrating normal behavior from a given normal

¹ https://en.wikipedia.org/wiki/Anomaly_detection

training data set, and then testing the probability of a test instance to be generated by the learnt model.

1.2. Popular Methods ²

Numerous anomaly detection methods have been proposed, some of the popular techniques are:

- Density-based techniques (k-nearest neighbor, local outlier factor, and many more variations of this concept).
- Subspace and correlation-based outlier detection for high-dimensional data.
- One class support vector machines.
- Replicator neural networks.
- Cluster analysis-based outlier detection.
- Deviations from association rules and frequent item sets.
- Fuzzy logic based outlier detection.
- Ensemble techniques, using feature bagging, score normalization and different sources of diversity.

1.3. Applications

Anomaly detection is applicable in a variety of domains, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting Eco-system disturbances. It is often used in preprocessing to remove anomalous data from the dataset. In supervised learning, removing the anomalous data from the dataset often results in a statistically significant increase in accuracy.

1.4. K-means Clustering ³

K-means clustering aims to divide n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a prior and must be computed from the data. The objective of K-means clustering is to minimize total intra-cluster variance, or, the squared error function.

² http://everything.explained.today/Anomaly_detection/

³ https://en.wikipedia.org/wiki/K-means_clustering

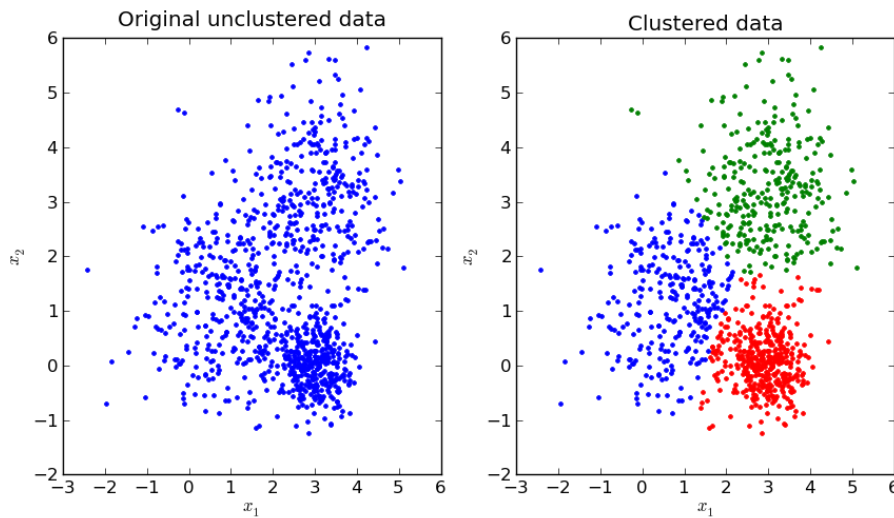


Figure 1 K-means Clustering⁴

1.5. K-means Algorithm⁵

1. Clusters the data into k groups where k is predefined.
2. Select k points at random as cluster centers.
3. Assign objects to their closest cluster center according to the Euclidean distance function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

K-means is comparatively an effective technique. Nevertheless, we need to lay down the number of clusters, in advance and the final outcomes are sensitive to initialization and often terminates at a local prime. Unluckily there is no global hypothetical method to find the ideal number of clusters. A real-world approach is to compare the results of multiple runs with different k and choose the best one based on a predefined condition. Overall, a large k perhaps decreases the error but increases the risk of over fitting.

⁴ http://pypr.sourceforge.net/images/kmeans_2d.png

⁵ http://www.saedsayad.com/clustering_kmeans.htm

1.6. Details of K-means ⁶

- Initial centroids are often chosen randomly.
 - Clusters produced vary from one run to another
- The centroid is (typically) the mean of the points in the cluster.
- ‘Closeness’ is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
 - Often the stopping condition is changed to ‘Until relatively few points change clusters’

2. Project Objective

In this project, we were supposed to build a web portal which could provide interesting cluster statistics and Graphical representation of the raw data presented to the API and display the K-means clustering results and the original (raw) data in 2D and 3D forms.

3. Project Requirement

Web Portal (Project 7 – Code: ML-KMEANWEB-00)

Implement a simple portal which provides interesting cluster information.

a. Cluster Statistics

Methods which return some usable cluster statistical data. For example

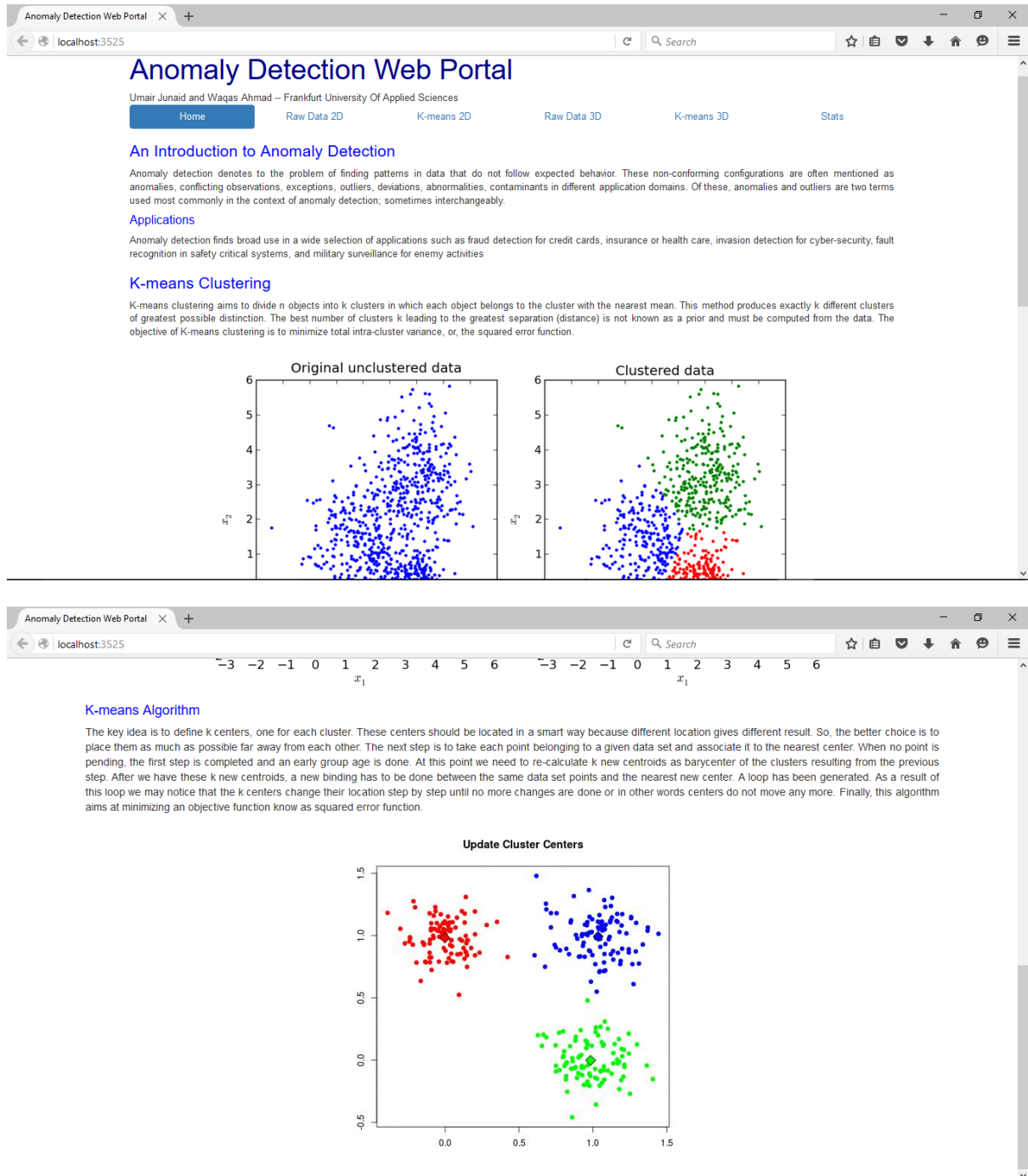
- i. Distance to farther sample from cluster centroid.
 - ii. Distance to next nearest cluster.
 - iii. Distance between nearest samples of nearest clusters.
- b. Graphical representation of Cluster-Centroids, Farthest Sample from Centroid Nearest Clusters, nearest samples from two clusters.
- c. How data in 2D and 3D by freezing remaining scalars.

⁶ <http://goo.gl/PDbaku>

4. Web Portal Tabs

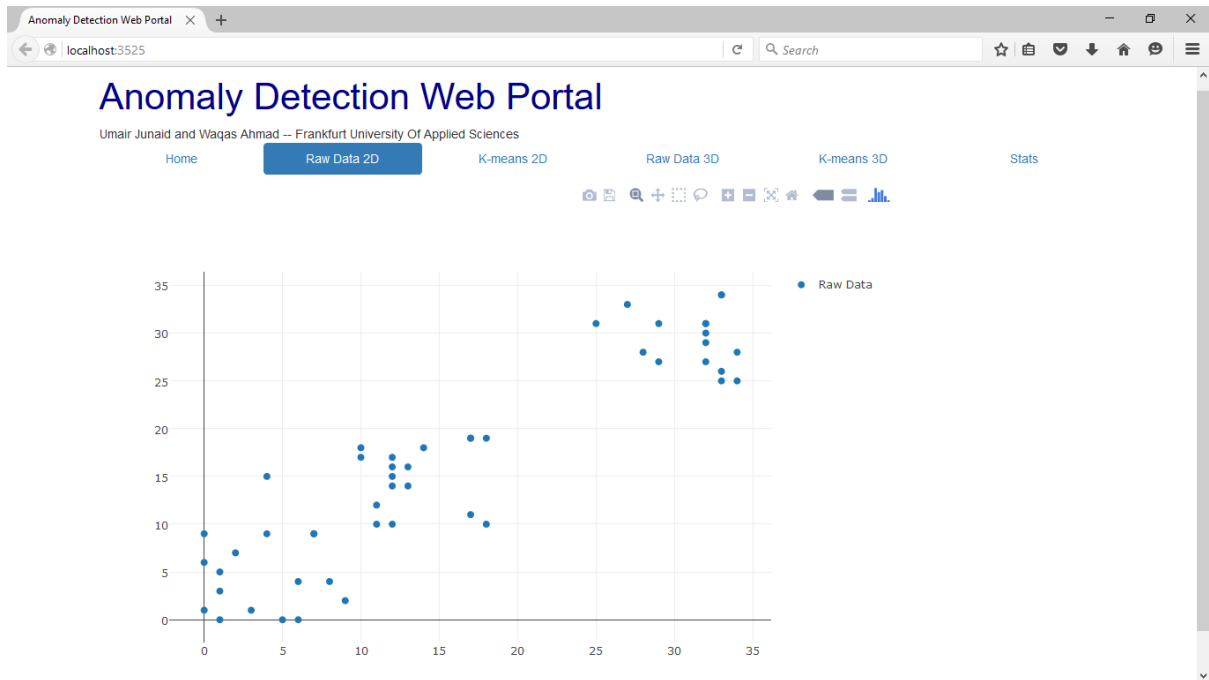
4.1. Home

The Home tab shows information about Anomaly Detection and K-means clustering.



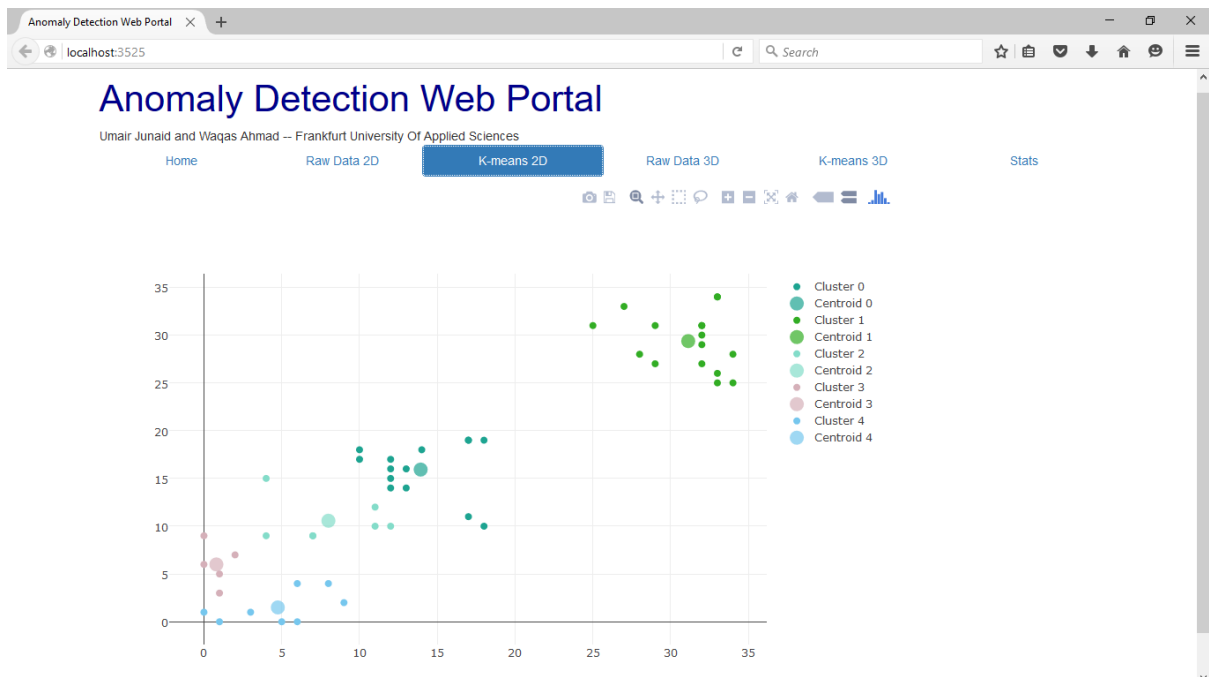
4.2. Raw Data 2D

The Raw Data 2D tab shows a 2-Dimensional scatter plot of the raw data.



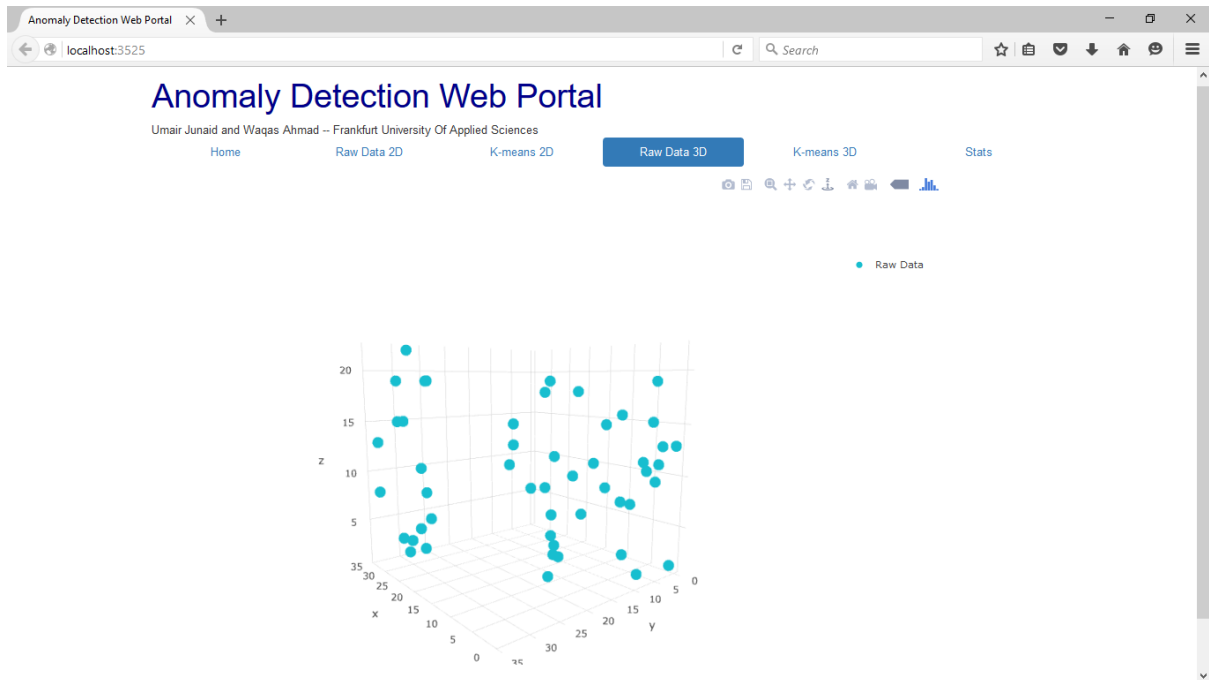
4.3. K-means 2D

The K-means 2D tab shows the 2-Dimensional plot of the result of K-means clustered data. The data is divided into different group clusters (based upon the number of k). The centroid of each cluster can also be seen.



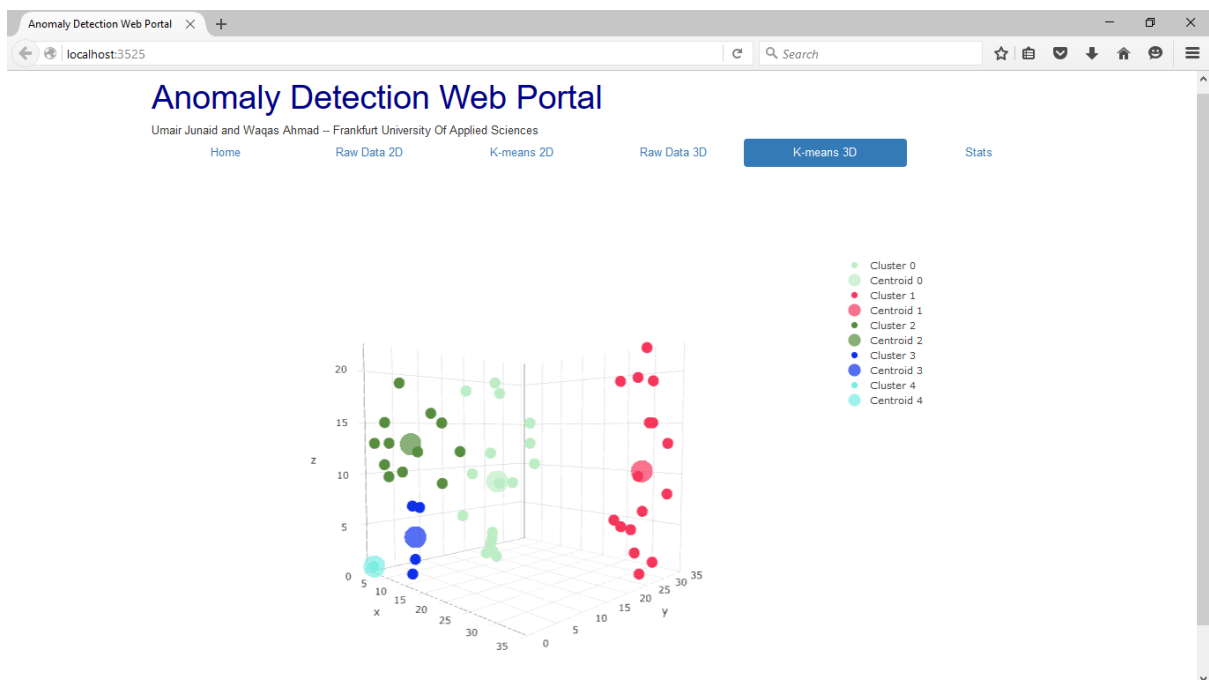
4.4. Raw Data 3D

The Raw Data 3D tab shows a 3-Dimensional scatter plot of the raw data.



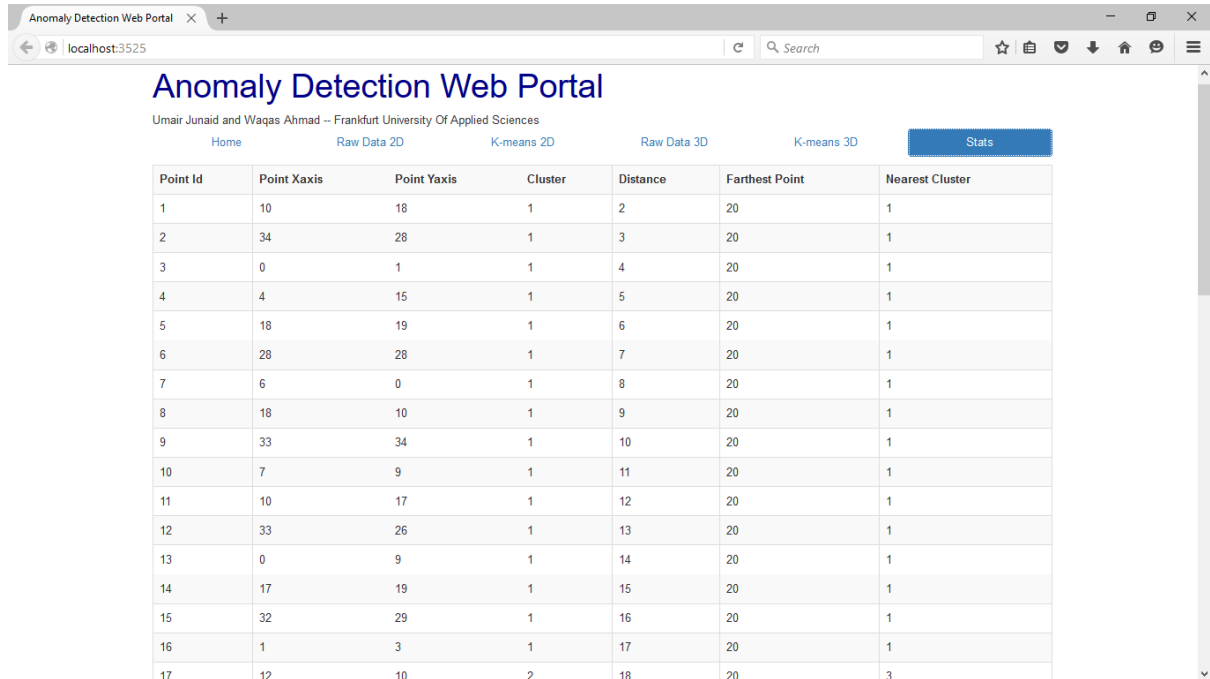
4.5. K-means 3D

The K-means 3D tab shows the 3-Dimensional plot of the result of K-means clustered data. Yet again, we can see different clusters, each with their centroid.



4.6. Stats

The last tab of the Web Portal is Stats. It shows the raw data points and the allocated cluster number after k-means clustering. It also shows the distance of a particular point to the farthest point and the nearest cluster to that point.



The screenshot shows a web browser window with the title "Anomaly Detection Web Portal". The address bar shows "localhost:3525". The page has a navigation bar with links: Home, Raw Data 2D, K-means 2D, Raw Data 3D, K-means 3D, and Stats (which is highlighted in blue). Below the navigation bar is a table with 7 columns: Point Id, Point Xaxis, Point Yaxis, Cluster, Distance, Farthest Point, and Nearest Cluster. The table contains 17 rows of data.

Point Id	Point Xaxis	Point Yaxis	Cluster	Distance	Farthest Point	Nearest Cluster
1	10	18	1	2	20	1
2	34	28	1	3	20	1
3	0	1	1	4	20	1
4	4	15	1	5	20	1
5	18	19	1	6	20	1
6	28	28	1	7	20	1
7	6	0	1	8	20	1
8	18	10	1	9	20	1
9	33	34	1	10	20	1
10	7	9	1	11	20	1
11	10	17	1	12	20	1
12	33	26	1	13	20	1
13	0	9	1	14	20	1
14	17	19	1	15	20	1
15	32	29	1	16	20	1
16	1	3	1	17	20	1
17	12	10	2	18	20	3

5. Conclusion

We believe that the small scale usage of K-means has helped us understand Anomaly detection. We were also able to appreciate how good, user friendly and powerful AngularJS is. Our code is able to display raw data and K-means clustered data in 2D and 3D graphical representations. Also, the code is able to show insightful statistics like farthest point, nearest cluster to the point etc.