Project Report

On

Smartphone Dataset for Anomaly Detection in Crowds

By

V Venkata Sai Kumar

(214g1a33c0@srit.ac.in)

Mentor: Mrs. Chitra Pandey

INDEX

Introduction

This project explores applying anomaly detection techniques to smartphone datasets in densely populated environments. By combining statistical and machine learning methods, we aim to create a model capable of effectively discerning between typical and unusual behavior patterns in smartphone usage. This versatile model can serve multiple purposes, including detecting suspicious activities, identifying potential risks, and enhancing our understanding of crowd dynamics.

Anomalous behavior vs non-anomalous behavior

According to any dataset, outliers represent the meaning of anomalous data which are irrelevant to the dataset. Likewise, many actions in real-time timeline series of dataset "outliers represent anomalous behavior". For example, if a person is walking freely representing the normal behavior and same person walks slowly in crowded environment can be represented as anomalous behavior.

Crowded Media

Generally, crowds represent an important activity which may be either safe or unsafe. For example, People cross the road according to traffic rules, scenario of market and stores during festivals etc. which represent normal behavior in crowd but in same situations suddenly people moving in one direction due to threats like fire, short circuits etc. represents anormal behavior in crowd. However, it is not easier to categorize the normal and abnormal behavior in crowds because some scenarios like people suddenly moved to shelters due to rain represents threat as per our assumptions but in reality, it is not.

Why is Smartphone Dataset Required?

In the present generation, every family has at least one or more smartphones and mostly carried with in routine life from morning workout activities to night sleep. Which shows that smartphone dataset can be utilized for anomaly detection in crowd. In addition to this smart phone contains many sensors like from intensity light detection i.e. proximity sensor to movement, orientation, acceleration sensors. Hence smartphone dataset is prior to analyze the behavior of crowd.

Module-1: Data Collection

Objectives

- To collect smartphone dataset which contain maximum features supporting to the crowded behavior.
- To identify and select the dataset which possesses the data of both normal and abnormal situations.
- To ensure features with appropriate labels based on sensory data of smartphone.

Problem Statement:

This project delves into the application of anomaly detection to smartphone datasets in crowded environments. By utilizing a combination of statistical and machine learning techniques, we will build a model that can effectively distinguish between normal and anomalous behaviour patterns in smartphone usage. This model can be used for various purposes, such as detecting suspicious activities, identifying potential hazards, or even understanding crowd dynamics better.

Problem statement defines that smartphone dataset is crucial for analysis and it must contain features to detect activities. Here is the dataset collected from Kaggle.com

timestamp	X	Y	Speed	Heading	AgentCour	Density	Acc	LevelOfCrc	label	label2	Severity_level
00:05:36	0.4225	19.1176	1.1432	89.1222	81	0.81	-0.0027	1	0	normal	0
00:05:37	0.3704	19.513	1.1476	89.5976	83	0.83	-0.0027	1	0	normal	0
00:05:38	0.3999	19.8316	1.1466	89.4905	86	0.86	-0.0051	1	0	normal	0
00:05:39	0.3787	20.0386	1.1521	89.2123	88	0.88	-0.0009	1	0	normal	0
00:05:40	0.4031	20.4625	1.1499	89.2521	88	0.88	-0.0012	1	0	normal	0
00:05:41	0.4104	20.6724	1.1406	89.5428	90	0.9	-0.0066	1	0	normal	0
00:05:42	0.4054	20.7604	1.1492	89.9155	91	0.91	0.0183	1	0	normal	0
00:05:43	0.3843	20.8616	1.1419	89.7962	90	0.9	-0.0017	1	0	normal	0
00:05:44	0.3603	21.0586	1.1503	88.7938	92	0.92	0.009	1	0	normal	0
00:05:45	0.3448	20.7365	1.1566	88.6091	91	0.91	0.0063	1	0	normal	0
00:05:46	0.3712	20.8257	1.1435	88.3917	90	0.9	-0.0036	1	0	normal	0
00:05:47	0.4543	20.893	1.1202	88.0303	89	0.89	-0.0066	1	0	normal	0
00:05:48	0.4425	20.9571	1.1267	88.3365	86	0.86	0.0106	1	0	normal	0
00:05:49	0.4587	20.8961	1.1278	88.9066	86	0.86	0.0062	1	0	normal	0
00:05:50	0.4576	21.2929	1.1303	88.9831	86	0.86	0.005	1	0	normal	0
00:05:51	0.4247	21.791	1.1329	89.16	87	0.87	0.0063	1	0	normal	0
00:05:52	0.3583	21.9679	1.1274	89.1392	85	0.85	0.0045	1	0	normal	0
00:05:53	0.3654	21.5336	1.1456	89.01	84	0.84	0.0216	1	0	normal	0
00:05:54	0.3674	20.9545	1.1483	89.3064	82	0.82	0.0104	1	0	normal	0
00:05:55	0.3688	20.9422	1.1535	88.8171	81	0.81	0.0084	1	0	normal	0
00:05:56	0.3892	20.568	1.1523	88.8102	81	0.81	0.0048	1	0	normal	0
00:05:57	0.3401	20.4463	1.1595	88.9378	81	0.81	0.0098	1	0	normal	0
00:05:58	0.3544	20.3336	1.1632	89.2561	83	0.83	0.0094	1	0	normal	0
00:05:59	0.3857	20.2172	1.1629	89.3949	81	0.81	0.001	1	0	normal	0
00:06:00	0.3755	20.457	1.1562	88.7305	80	0.8	-0.0019	1	0	normal	0
00:06:01	0.3404	20.214	1.1583	88.9206	81	0.81	0.0072	1	0	normal	0

Overview of Dataset:

Number of Features: 12

Number of Rows: 24123

Data Dictionary

1. Timestamp: Time of the observation.
2. X: X-coordinate.
3. Y: Y-coordinate.
4. Speed: Speed of the agent.
5. Heading: Direction of the agent.
6. AgentCount: Number of agents.
7. Density: Density of the crowd.
8. Acc: Acceleration.
9. LevelOfCrowdness: Level of crowdness (discrete levels).
10. Label: Binary label indicating normal (0) or anomalous (1) behavior.
11. Label2: Text label indicating the type of behavior (e.g., normal).
12. Severity_level: Severity of the situation (discrete levels).

Module-2: Exploratory Data Analysis (EDA) and Data Preprocessing

Missing Values

Data preprocessing is mandatory to format the dataset into useful dataset to the analysis. Detecting the missing values is one of the important preprocessing techniques. Here, in the dataset there are 104 values are missing for the Acceleration feature. These values should be treated with other values.

```
timestamp            0
X                    0
Y                    0
Speed                0
Heading              0
AgentCount           0
Density              0
Acc                104
LevelOfCrowdness     0
label                0
label2               0
Severity_level       0
dtype: int64
```

Treatment of missing values

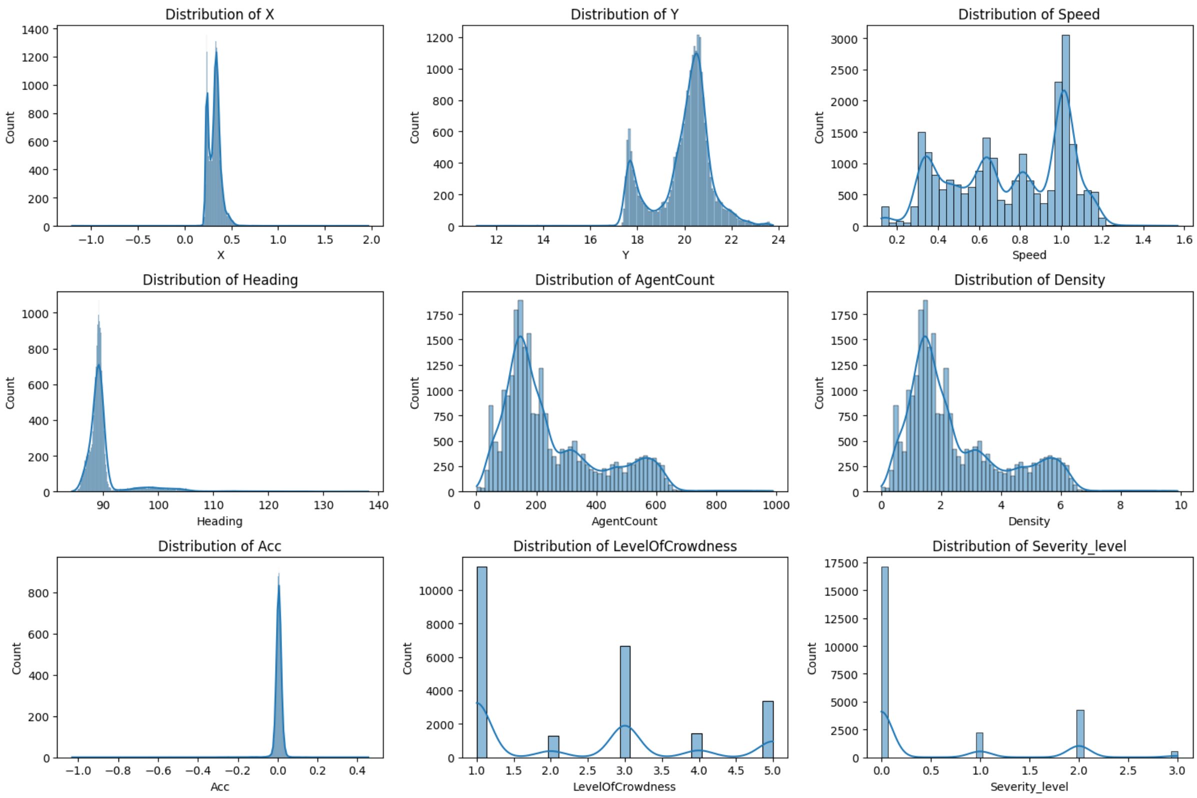
Missing Values are bottleneck to the analysis because which effect on results hence normalizing them is important task in preprocessing of dataset. As we observed acceleration feature contains 104 null values (missing values), we normalize the values with mean distribution. Here initially we calculate mean value for distribution of acceleration feature and then all null values (missing values) replaced with mean value.

```
timestamp           0
X                   0
Y                   0
Speed               0
Heading             0
AgentCount          0
Density             0
Acc                 0
LevelOfCrowdness    0
label               0
label2              0
Severity_level      0
dtype: int64
```

Above output shows after treating the missing values that it concludes there are no missing values in dataset.

Analyzing the Distribution of Features

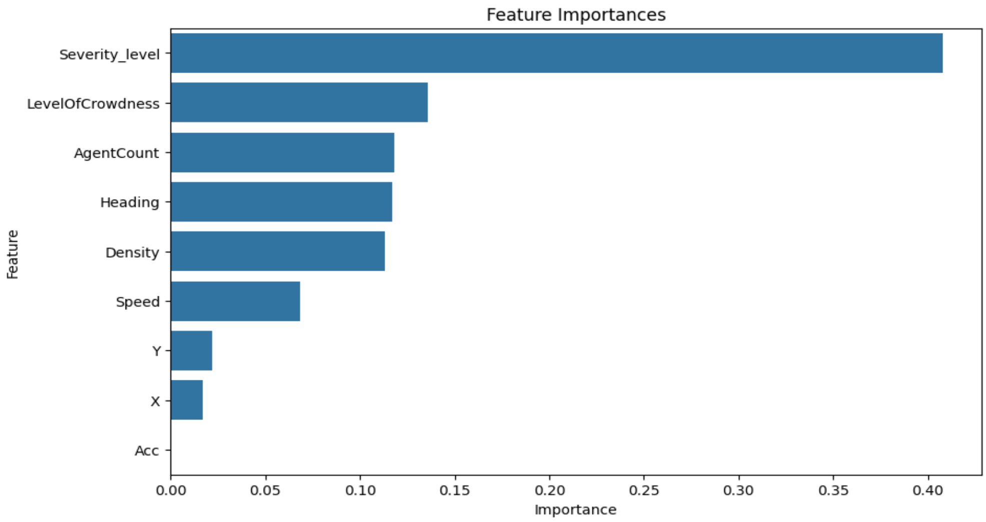
In dataset, there are total 24,233 rows which means the inputs are broadly greater in number. Hence it is important to find out distribution of features in dataset.



Above plots describes about features of dataset. For feature ‘X’ the values are between 0.1 to 0.6, for feature ‘Y’ values are 17 to 23.5, for feature ‘Speed’ values are between 0.1 to 1.3, for feature ‘Heading’ values are between 80 to 105, for feature ‘Agent Count’ values are between 0 to 650, for feature ‘Density’ values are between 0 to 7, for feature ‘Acc(Acceleration) values are between -0.1 to 0.1, for feature ‘Level of Crowdness’ values are between 1.0 to 5.0.

From this analysis, we observe that distribution of features such as ‘Y’, ‘Speed’, ‘Agent Count’, ‘Density’ are volatile in nature which represents that these features could play key role in anomaly detection.

Importance of features



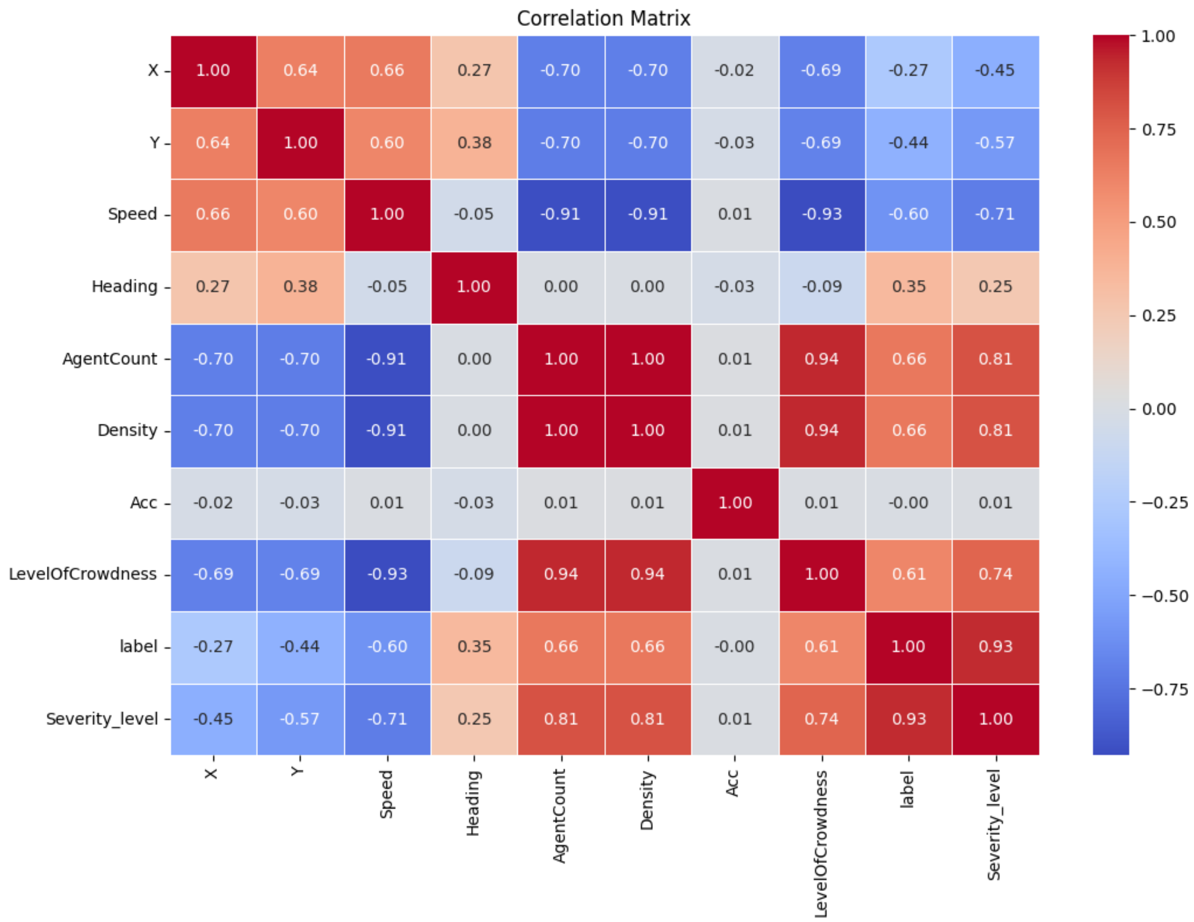
From the feature importance chart, we observe the following key points:

- **Severity_level** is the most significant feature, with an importance score of over 0.40. This indicates that the severity of detected events or conditions plays a critical role in identifying anomalies. It suggests that severe deviations from normal behavior are highly indicative of anomalous activity.
- **LevelOfCrowdness** follows with an importance score slightly above 0.20. The level of crowd density is a crucial factor in understanding crowd dynamics and detecting unusual patterns in dense environments.
- **AgentCount** and **Heading** both show moderate importance, with scores around 0.15. The number of agents and the directional movement are essential for understanding how individuals navigate and interact within a crowd.
- **Density** also has a moderate importance score, similar to AgentCount and Heading. This feature provides additional context on the spatial distribution of the crowd, complementing the LevelOfCrowdness.
- **Speed** is another important feature, with a score around 0.10. Unusual speed patterns can indicate anomalies such as sudden movements or unexpected stops.
- The spatial coordinates **Y** and **X** have lower importance scores, around 0.05 each. While they contribute to tracking the smartphone's position, their individual impact on anomaly detection is less significant compared to other features.
- **Acc** (accelerometer data) has the least importance, with a score close to 0. This suggests that in the context of this dataset and model, acceleration data alone is not a strong indicator of anomalous behavior.

The feature importance analysis highlights the critical factors that contribute to anomaly detection in crowded environments. Understanding the significance of each feature helps in refining the model and improving its performance.

By focusing on the most important features, such as Severity_level and LevelOfCrowdness, we can enhance the model's ability to accurately detect and respond to anomalies in real-time.

Correlation Matrix (excluding non-numerical features)



Insights:

The correlation matrix displayed in the image provides insights into the relationships between various numerical features used in our anomaly detection model. Understanding these correlations helps in feature selection and model optimization by identifying which features are closely related and how they contribute to the target variable.

X and Y Coordinates:

- The X and Y coordinates have a moderate positive correlation with each other (0.64), suggesting some spatial relationship in movement.
- Both X and Y coordinates show a negative correlation with features like AgentCount, Density, and LevelOfCrowdness, indicating that higher values of these features are associated with lower spatial coordinates.

Speed:

- Speed has a strong negative correlation with AgentCount (-0.91) and Density (-0.91), indicating that higher crowd densities are associated with lower movement speeds.
- Speed also shows a moderate negative correlation with LevelOfCrowdness (-0.93), aligning with the idea that more crowded environments restrict movement speed.

Heading:

- Heading shows weak correlations with most features, except for a moderate positive correlation with the label (0.35) and a mild positive correlation with Severity_level (0.25). This suggests that directional movement has some relevance to detecting anomalies.

AgentCount, Density, and LevelOfCrowdness:

- These features are highly correlated with each other (AgentCount and Density: 1.00, AgentCount and LevelOfCrowdness: 0.94, Density and LevelOfCrowdness: 0.94). This high correlation indicates that they are measuring similar aspects of the crowd environment.
- They all show strong positive correlations with the label and Severity_level, suggesting that higher crowd density and agent counts are associated with higher anomaly severity.

Accelerometer Data (Acc):

- The accelerometer data (Acc) shows weak correlations with most other features, indicating that it might not be as impactful in distinguishing anomalies within this context.

LevelOfCrowdness and Severity_level:

- LevelOfCrowdness has a strong positive correlation with Severity_level (0.74) and the label (0.61), indicating that crowdedness is a significant factor in determining the severity of anomalies.
- Severity_level is highly correlated with the label (0.93), emphasizing its importance in the model's predictions.

Implications for Model Development

1. Due to the high correlations between AgentCount, Density, and LevelOfCrowdness, including all three features in the model might introduce redundancy. It could be beneficial to select one or two of these features to avoid multicollinearity.
2. Severity_level, LevelOfCrowdness, and AgentCount are crucial features due to their strong correlations with the label and each other. These features should be prioritized in model development and tuning.
3. The X and Y coordinates and Acc show weaker correlations with the label and other features. While they provide spatial and movement context, their impact on anomaly detection might be limited.

Module-3: Preliminary Statistical Models (IQR)

The Interquartile Range (IQR) is a measure of statistical dispersion, or how spread out the values in a dataset are. The IQR is used to describe the middle 50% of values, providing a robust measure of variability that is less influenced by outliers or extreme values.

$$IQR = Q3 - Q1$$

Where,

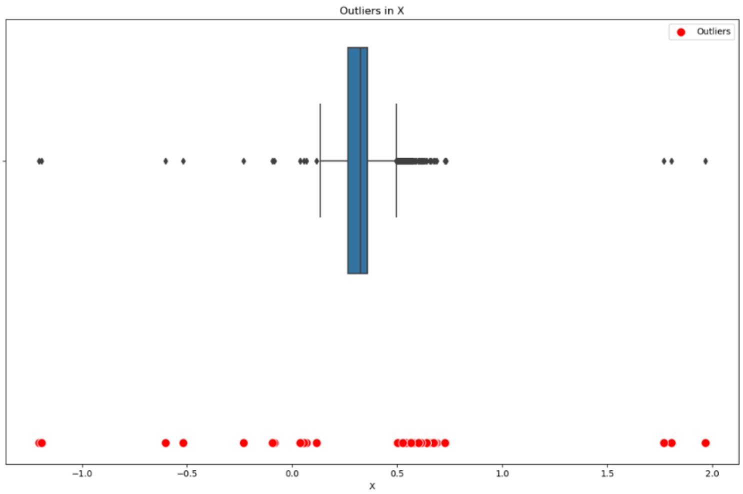
Q1 (First Quartile): The median of the first half of the dataset (25th percentile).

Q3 (Third Quartile): The median of the second half of the dataset (75th percentile).

Any data point falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR is considered a potential outlier.

IQR for Feature 'X'

```
IQR Values:
{'X': 0.0923}

Outliers:
280
      timestamp       X        Y   Speed   Heading  AgentCount  Density
2884    0:53:40  0.5134  21.2985  0.7035  102.0650         283     2.83
2885    0:53:41  0.5183  21.4306  0.7333  103.5605         281     2.81
2886    0:53:42  0.5083  21.4887  0.7637  105.3477         278     2.78
2887    0:53:43  0.5070  21.6273  0.8043  106.4613         274     2.74
4150    0:59:07  0.5301  20.2118  1.0353   91.0069          49     0.49
...         ...     ...      ...     ...       ...         ...      ...
16710   0:00:22  0.6000  16.2390  1.4521   90.6531           9     0.09
16711   0:00:23  0.7257  16.4256  1.4208   89.4937          11     0.11
16713   0:00:25  0.5652  16.8410  1.3910   90.4042          16     0.16
16821   0:02:13  0.5021  21.3055  1.0841   89.8225          45     0.45
16900   0:03:32  0.5273  21.1877  1.1483   90.4889          53     0.53

          Acc  LevelOfCrowdness  label   label2  Severity_level
2884   0.0320                 3      1  anomaly               1
2885   0.0483                 3      1  anomaly               1
2886   0.0386                 3      1  anomaly               1
2887   0.0482                 3      1  anomaly               1
4150  -0.0080                 1      0   normal               0
...       ...               ...    ...      ...             ...
16710  0.0033                 1      0   normal               0
16711 -0.0034                 1      0   normal               0
16713 -0.0005                 1      0   normal               0
16821 -0.0134                 1      0   normal               0
16900  0.0159                 1      0   normal               0

[280 rows x 12 columns]
```

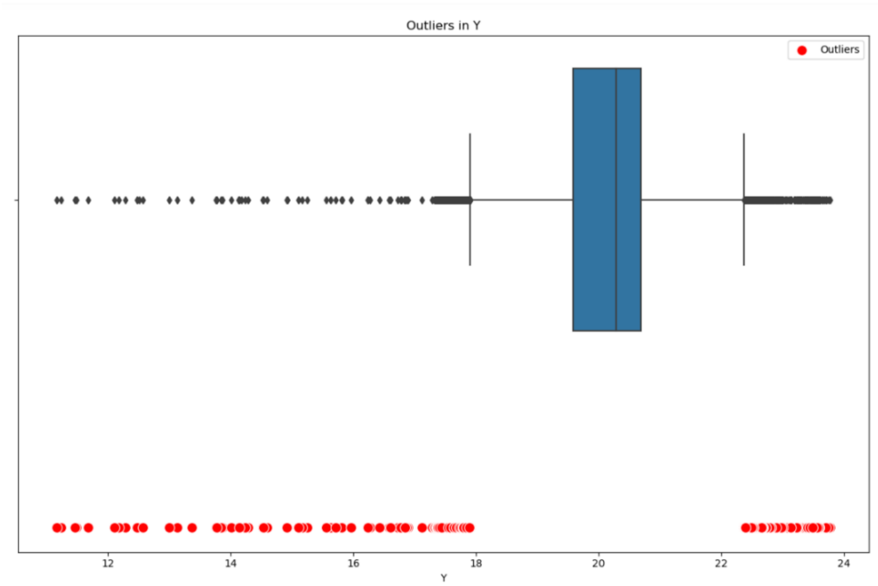


- The IQR value for the variable X is 0.0923. This indicates the spread of the middle 50% of the X values.
- A total of 280 outliers were detected based on the X variable.

IQR for Feature 'Y'

```
IQR Values:
{'Y': 1.1172000000000004}

Outliers:
3004
      timestamp       X        Y   Speed  Heading  AgentCount  Density  \
1719    0:34:15  0.2314  17.8719  0.3453  87.4521         509     5.09
1720    0:34:16  0.2295  17.8558  0.3355  86.9938         510     5.10
1729    0:34:25  0.2429  17.8711  0.3345  87.9196         519     5.19
1731    0:34:27  0.2342  17.8566  0.3380  87.5961         524     5.24
1733    0:34:29  0.2443  17.8939  0.3311  86.9780         525     5.25
...         ...     ...      ...     ...      ...         ...      ...
23084   0:41:56  0.2490  17.8727  0.3875  87.1785         483     4.83
23085   0:41:57  0.2441  17.8590  0.3915  87.9671         480     4.80
23086   0:41:58  0.2398  17.8562  0.3955  87.8023         483     4.83
23107   0:42:19  0.2485  17.8408  0.4224  87.3489         459     4.59
23108   0:42:20  0.2472  17.8950  0.4364  87.8707         459     4.59

          Acc  LevelOfCrowdness  label   label2  Severity_level
1719   0.0222                 5      1  anomaly               2
1720   0.0053                 5      1  anomaly               2
1729   0.0153                 5      1  anomaly               2
1731   0.0159                 5      1  anomaly               2
1733   0.0048                 5      1  anomaly               2
...       ...               ...    ...      ...             ...
23084 -0.0029                 4      1  anomaly               2
23085  0.0202                 4      1  anomaly               2
23086  0.0132                 4      1  anomaly               2
23107 -0.0055                 4      1  anomaly               2
23108  0.0127                 4      1  anomaly               2

[3004 rows x 12 columns]
```

- The IQR value for the Y variable is 1.1172. This indicates the spread of the middle 50% of the Y values.
- A total of 3004 outliers were detected based on the Y variable.

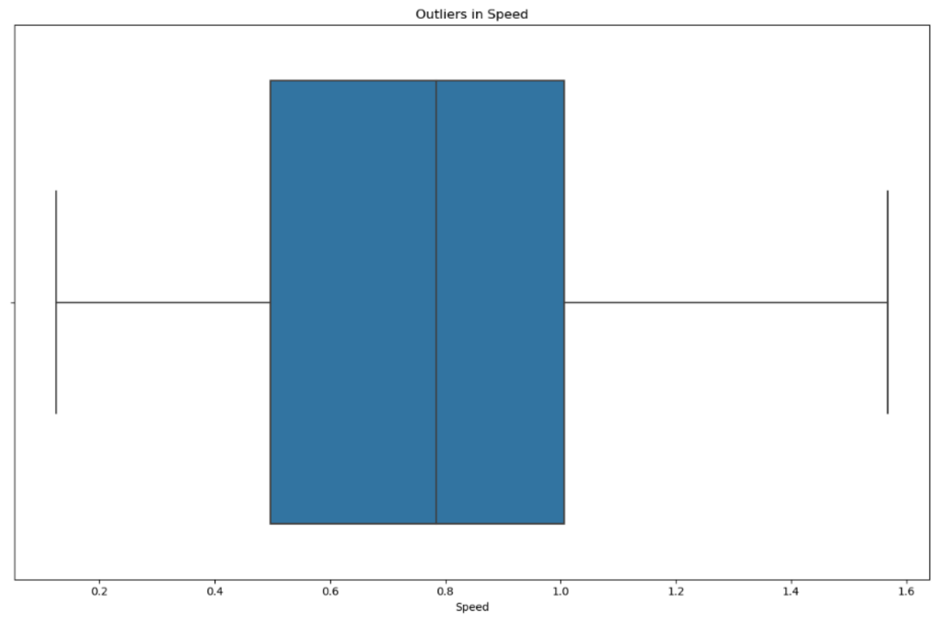


Examples of outliers:

- At timestamp `0:34:15`, the `Y` value is 17.8719, and the anomaly labels are marked as 1.
- At timestamp `0:41:56`, the `Y` value is 17.8727, and the anomaly labels are marked as 1 (indicating an anomaly).

IQR for Feature ‘Speed’

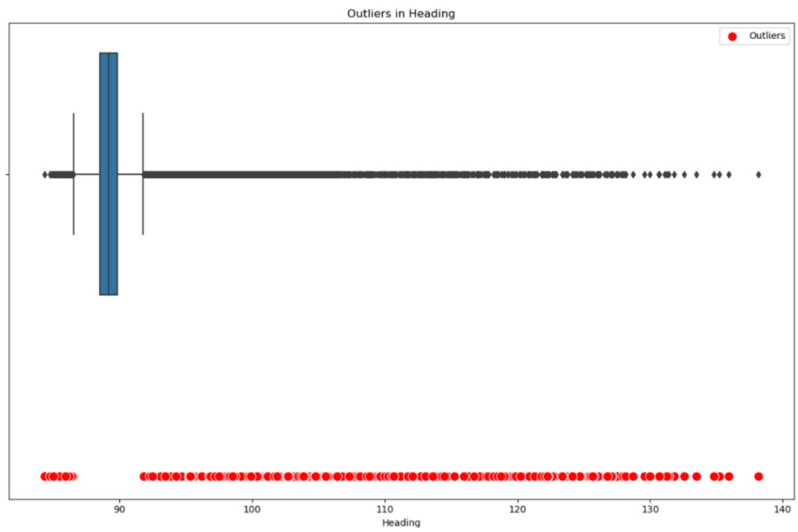
```
IQR Values:
{'Speed': 0.5092000000000001}

Outliers:
0
Empty DataFrame
Columns: [timestamp, X, Y, Speed, Heading, AgentCount, Density, Acc, LevelOfCrowdness, label, label2, Severity_level]
Index: []
```



- The IQR value for the Speed variable is 0.5092. This indicates the spread of the middle 50% of the Speed values.
- No outliers were detected in the Speed variable.
- The box plot visualization effectively illustrates the distribution of Speed values, showing that all data points are within the normal range.

IQR for Feature ‘Heading’

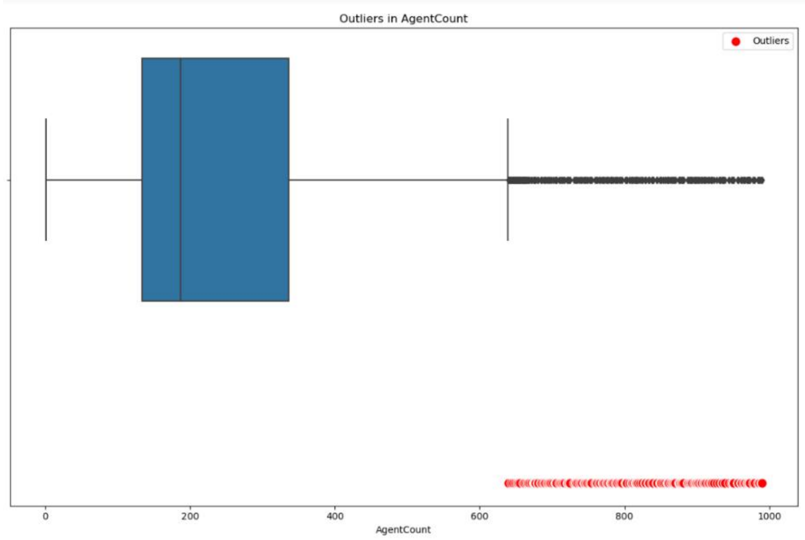
```
IQR Values:
{'Heading': 1.3221499999999935}

Outliers:
3461
       timestamp       X        Y   Speed  Heading  AgentCount  Density
1279     0:26:55  0.3217  19.3262  0.5203  92.3444         233     2.33
1421     0:29:17  0.2529  19.7438  0.4777  86.3473         334     3.34
1422     0:29:18  0.2655  19.7065  0.4736  86.4754         332     3.32
1503     0:30:39  0.2485  19.3132  0.4333  86.4520         372     3.72
1505     0:30:41  0.2525  19.4052  0.4179  86.3024         373     3.73
...          ...     ...      ...     ...      ...         ...      ...
23032    0:41:04  0.2282  17.8689  0.3740  86.3509         510     5.10
23033    0:41:05  0.2205  17.8110  0.3673  86.1579         511     5.11
23034    0:41:06  0.2225  17.8313  0.3700  86.1503         509     5.09
23083    0:41:55  0.2418  17.8679  0.4046  86.1618         481     4.81
23119    0:42:31  0.2502  18.2235  0.4310  85.9236         459     4.59

          Acc  LevelOfCrowdness  label    label2  Severity_level
1279  -0.0283                 3      0    normal               0
1421   0.0007                 3      0    normal               0
1422   0.0145                 3      0    normal               0
1503   0.0197                 3      0    normal               0
1505   0.0045                 3      0    normal               0
...       ...               ...    ...       ...             ...
23032  0.0220                 5      1   anomaly               2
23033  0.0102                 5      1   anomaly               2
23034  0.0207                 5      1   anomaly               2
23083  0.0120                 4      1   anomaly               2
23119 -0.0022                 4      1   anomaly               2

[3461 rows x 12 columns]
```



- The IQR value for the Speed variable is 1.3221. This indicates the spread of the middle 50% of the Speed values.
- A total of 3461 outliers were detected based on the Heading variable.

IQR for Feature ‘AgentCount’

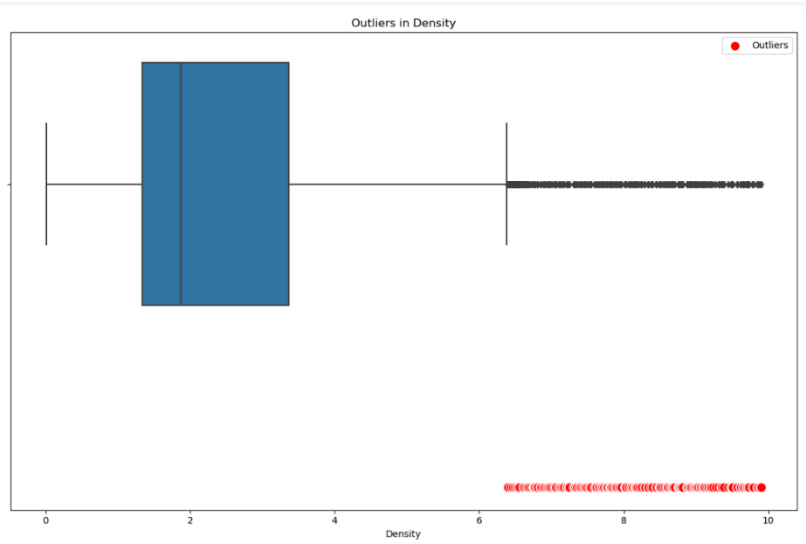
```
IQR Values:
{'AgentCount': 202.0}

Outliers:
284
       timestamp       X        Y   Speed   Heading  AgentCount  Density  \
14782    0:37:21  0.2257  17.6054  0.2927   86.8057         641     6.41
14785    0:37:24  0.2317  17.5431  0.2944   86.1967         642     6.42
14793    0:37:32  0.2445  17.4699  0.2877   86.2804         640     6.40
14794    0:37:33  0.2321  17.4633  0.2803   86.9579         641     6.41
14795    0:37:34  0.2343  17.4788  0.2789   87.3341         643     6.43
...          ...     ...      ...     ...       ...         ...      ...
22423    0:59:55  0.2476  20.7841  0.1411  128.1307         986     9.86
22424    0:59:56  0.2464  20.8038  0.1388  128.6967         986     9.86
22425    0:59:57  0.2477  20.7936  0.1405  126.7505         989     9.89
22426    0:59:58  0.2465  20.7839  0.1420  125.6554         990     9.90
22427    0:59:59  0.2458  20.7864  0.1422  124.1256         990     9.90

          Acc  LevelOfCrowdness  label    label2  Severity_level
14782  0.0102                 5      1   anomaly               2
14785  0.0085                 5      1   anomaly               2
14793  0.0065                 5      1   anomaly               2
14794  0.0039                 5      1   anomaly               2
14795  0.0089                 5      1   anomaly               2
...       ...               ...    ...       ...             ...
22423  0.0021                 5      1   anomaly               3
22424 -0.0023                 5      1   anomaly               3
22425  0.0014                 5      1   anomaly               3
22426  0.0016                 5      1   anomaly               3
22427  0.0003                 5      1   anomaly               3

[284 rows x 12 columns]
```



- The IQR value for the Speed variable is 202.0. This indicates the spread of the middle 50% of the Speed values.
- A total of 284 outliers were detected based on the Heading variable.

IQR for Feature ‘Density’

```
IQR Values:
{'Density': 2.0199999999999996}

Outliers:
286
       timestamp       X        Y   Speed   Heading  AgentCount  Density
14782    0:37:21  0.2257  17.6054  0.2927   86.8057         641     6.41
14783    0:37:22  0.2252  17.5493  0.2961   86.9792         639     6.39
14785    0:37:24  0.2317  17.5431  0.2944   86.1967         642     6.42
14793    0:37:32  0.2445  17.4699  0.2877   86.2804         640     6.40
14794    0:37:33  0.2321  17.4633  0.2803   86.9579         641     6.41
...          ...     ...      ...     ...       ...         ...      ...
22423    0:59:55  0.2476  20.7841  0.1411  128.1307         986     9.86
22424    0:59:56  0.2464  20.8038  0.1388  128.6967         986     9.86
22425    0:59:57  0.2477  20.7936  0.1405  126.7505         989     9.89
22426    0:59:58  0.2465  20.7839  0.1420  125.6554         990     9.90
22427    0:59:59  0.2458  20.7864  0.1422  124.1256         990     9.90

          Acc  LevelOfCrowdness  label    label2  Severity_level
14782  0.0102                 5      1   anomaly               2
14783  0.0149                 5      1   anomaly               2
14785  0.0085                 5      1   anomaly               2
14793  0.0065                 5      1   anomaly               2
14794  0.0039                 5      1   anomaly               2
...       ...               ...    ...       ...             ...
22423  0.0021                 5      1   anomaly               3
22424 -0.0023                 5      1   anomaly               3
22425  0.0014                 5      1   anomaly               3
22426  0.0016                 5      1   anomaly               3
22427  0.0003                 5      1   anomaly               3

[286 rows x 12 columns]
```

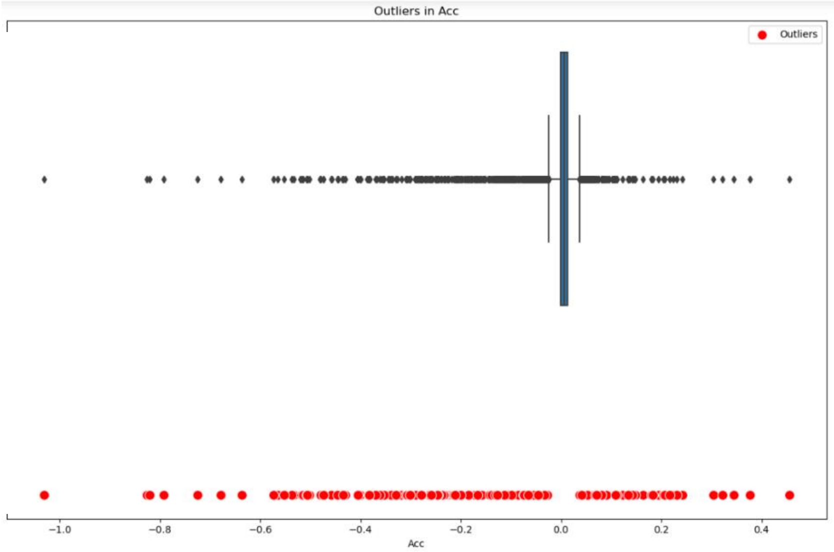


- The IQR value for the Speed variable is 2.0199. This indicates the spread of the middle 50% of the Speed values.

- A total of 286 outliers were detected based on the Heading variable.

IQR for Feature 'Acc'

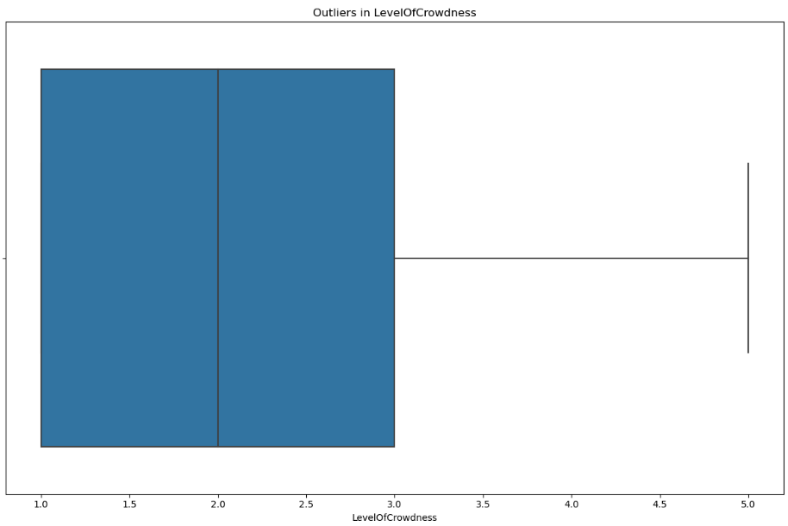
```
IQR Values:
{'Acc': 0.0153}

Outliers:
857
       timestamp       X        Y   Speed  Heading  AgentCount  Density  \
624      0:16:00  0.3495  20.4847  0.9018  88.3933         157     1.57
699      0:17:15  0.3506  20.5183  0.7720  89.6749         183     1.83
708      0:17:24  0.3844  20.5947  0.8444  89.9760         187     1.87
748      0:18:04  0.3045  21.0239  0.8950  88.2255         163     1.63
749      0:18:05  0.3153  20.7040  0.9082  88.2997         160     1.60
...          ...     ...      ...     ...      ...         ...      ...
23511    0:49:03  0.3345  20.8622  1.1506  89.2948         142     1.42
23533    0:49:25  0.2995  20.9276  1.1693  89.0261         136     1.36
23557    0:49:49  0.3156  20.6709  1.1337  88.4769         134     1.34
23561    0:49:53  0.3335  20.9106  1.1244  89.4679         137     1.37
23566    0:49:58  0.3387  20.5836  1.1565  88.8676         133     1.33

          Acc  LevelOfCrowdness  label  label2  Severity_level
624    0.0407                 1      0  normal               0
699   -0.0394                 2      0  normal               0
708    0.0382                 2      0  normal               0
748    0.0379                 1      0  normal               0
749    0.0384                 1      0  normal               0
...       ...               ...    ...     ...             ...
23511 -0.0742                 1      0  normal               0
23533 -0.1271                 1      0  normal               0
23557 -0.0653                 1      0  normal               0
23561 -0.0461                 1      0  normal               0
23566  0.0704                 1      0  normal               0

[857 rows x 12 columns]
```



- The IQR value for the Speed variable is 0.0153. This indicates the spread of the middle 50% of the Speed values.
- A total of 857 outliers were detected based on the Heading variable.

IQR for Feature 'LevelofCrowdness'

```
IQR Values:
{'LevelOfCrowdness': 2.0}

Outliers:
0
Empty DataFrame
Columns: [timestamp, X, Y, Speed, Heading, AgentCount, Density, Acc, LevelOfCrowdness, label, label2, Severity_level]
Index: []
```

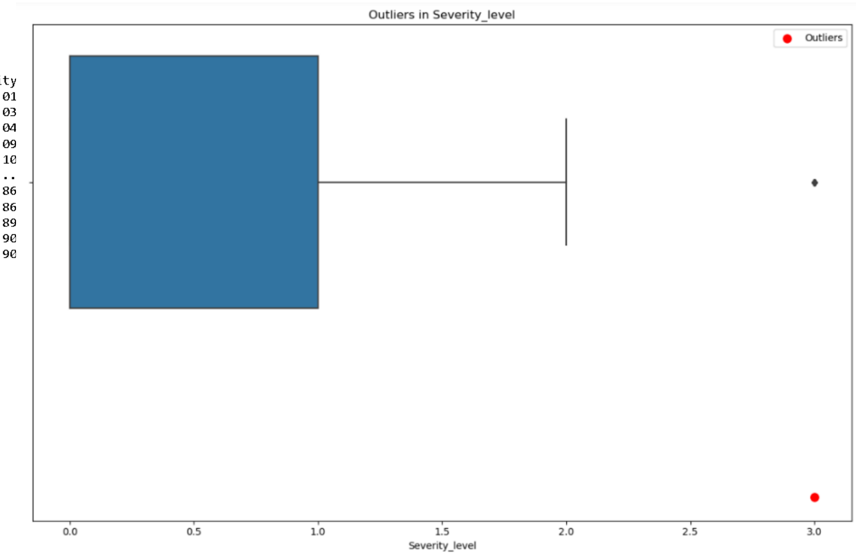


- The IQR value for the Speed variable is 2.0. This indicates the spread of the middle 50% of the Speed values.

- No outliers were detected in the Speed variable.

IQR for Feature ‘Severity_level’

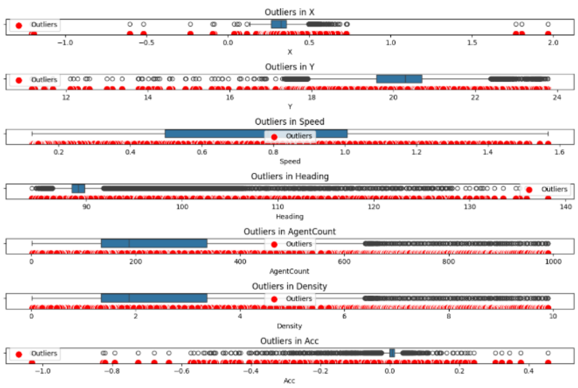
```
IQR Values:
{'Severity_level': 1.0}

Outliers:
534
      timestamp       X        Y   Speed   Heading  AgentCount  Density
3067    0:56:43  0.2718  21.4688  0.2834  101.3003         401     4.01
3068    0:56:44  0.2762  21.4441  0.2894   99.8226         403     4.03
3069    0:56:45  0.2862  21.4480  0.2841   98.9516         404     4.04
3070    0:56:46  0.2860  21.3002  0.2917   98.0279         409     4.09
3071    0:56:47  0.2801  21.2719  0.2962   98.2284         410     4.10
...         ...     ...      ...     ...       ...         ...      ...
22423   0:59:55  0.2476  20.7841  0.1411  128.1307         986     9.86
22424   0:59:56  0.2464  20.8038  0.1388  128.6967         986     9.86
22425   0:59:57  0.2477  20.7936  0.1405  126.7505         989     9.89
22426   0:59:58  0.2465  20.7839  0.1420  125.6554         990     9.90
22427   0:59:59  0.2458  20.7864  0.1422  124.1256         990     9.90

          Acc  LevelOfCrowdness  label   label2  Severity_level
3067  -0.5656                 4      1  anomaly               3
3068   0.0051                 4      1  anomaly               3
3069  -0.0049                 4      1  anomaly               3
3070  -0.0011                 4      1  anomaly               3
3071   0.0029                 4      1  anomaly               3
...       ...               ...    ...      ...             ...
22423  0.0021                 5      1  anomaly               3
22424 -0.0023                 5      1  anomaly               3
22425  0.0014                 5      1  anomaly               3
22426  0.0016                 5      1  anomaly               3
22427  0.0003                 5      1  anomaly               3

[534 rows x 12 columns]
```



- The IQR value for the Speed variable is 1.0. This indicates the spread of the middle 50% of the Speed values.
- A total of 534 outliers were detected based on the Heading variable.

Total outliers

```
IQR Values:
{'X': 0.0923, 'Y': 1.1172000000000004, 'Speed': 0.5092000000000001, 'Heading': 1.3221499999999935, 'AgentCount': 202.0, 'Density': 2.0199999999999996, 'Acc': 0.0153}

Outliers:
6202
      timestamp       X        Y   Speed   Heading  AgentCount  Density  \
2884    0:53:40  0.5134  21.2985  0.7035  102.0650         283     2.83
2885    0:53:41  0.5183  21.4306  0.7334  104.5605         281     2.81
2886    0:53:42  0.5083  21.4887  0.7637  105.3477         278     2.78
2887    0:53:43  0.5070  21.6273  0.8043  106.4613         274     2.74
4150    0:59:07  0.5301  20.2118  1.0353   91.0069          49     0.49
...         ...     ...      ...     ...       ...         ...      ...
23511   0:49:03  0.3345  20.8622  1.1506   89.2948         142     1.42
23533   0:49:25  0.2995  20.9276  1.1693   89.0261         136     1.36
23557   0:49:49  0.3156  20.6709  1.1337   88.4769         134     1.34
23561   0:49:53  0.3335  20.9106  1.1244   89.4679         137     1.37
23566   0:49:58  0.3387  20.5836  1.1565   88.8676         133     1.33

          Acc  LevelOfCrowdness  label   label2  Severity_level  outlier_lof
2884   0.0320                 3      1  anomaly               1            1
2885   0.0433                 3      1  anomaly               1            1
2886   0.0486                 3      1  anomaly               1            1
2887   0.0482                 3      1  anomaly               1            1
4150  -0.0080                 1      0   normal               0            1
...       ...               ...    ...      ...             ...          ...
23511 -0.0742                 1      0   normal               0            1
23533 -0.1271                 1      0   normal               0            1
23557 -0.0653                 1      0   normal               0            1
23561 -0.0461                 1      0   normal               0            1
23566  0.0704                 1      0   normal               0            1
```



By using IQR statistical method, we identified that there are total 6202 potential outliers in dataset which are considered as anomalies (abnormal behaviour).

Module-4: Preliminary Statistical Models (Z-Score)

The Z-score, also known as the standard score, is a statistical measurement that describes a value's relationship to the mean of a group of values. It can be calculated by

$$z = (x - \text{mean}) / \text{standard deviation}$$

1) The Z-score standardizes data, transforming it into a common scale with a mean of 0 and a standard deviation of 1. This makes it easier to compare different data points and detect outliers.

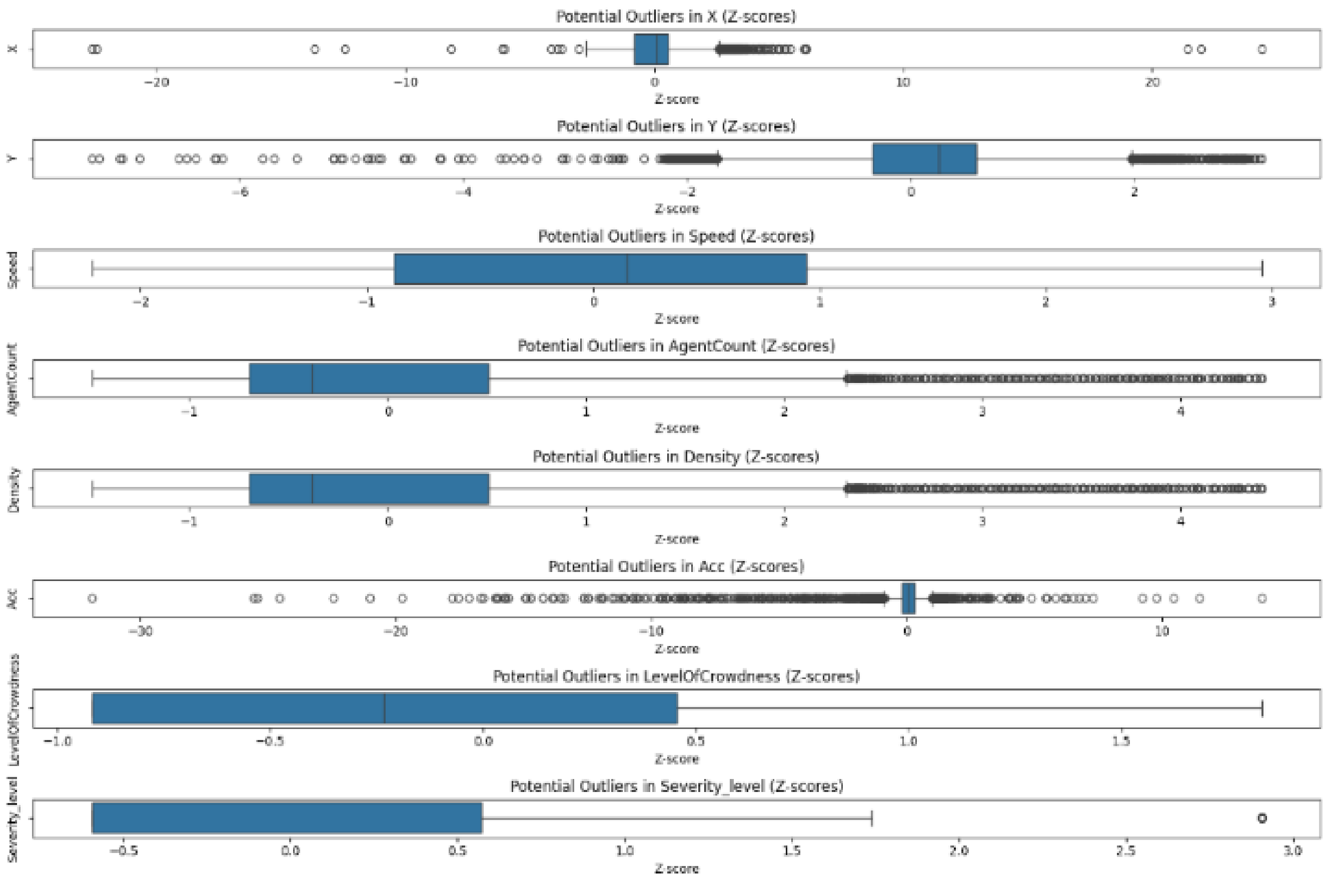
2) Outliers are data points that are significantly different from the rest of the data. Z-scores help in identifying these anomalies by quantifying how many standard deviations a data point is from the mean. Points with high absolute Z-scores are considered outliers.

3) A common threshold for identifying outliers using Z-scores is 3 or -3. This means any data point with a Z-score greater than 3 or less than -3 is considered an outlier. This threshold is based on the properties of the normal distribution, where about 99.7% of data points lie within three standard deviations of the mean.

In our dataset, we can apply on all numerical features i.e. X, Y, Speed, Acc etc. Z-Score value varies based on data for features in dataset. These are Z-Score values for first 5rows of dataset.

	X	Y	Speed	AgentCount	Density	Acc	LevelOfCrowdness	Severity_level
0	1.520675	-0.723185	1.432952	-1.014795	-1.014795	-0.212691	-0.918124	-0.594115
1	0.749318	-0.395174	1.448705	-1.002863	-1.002863	-0.212691	-0.918124	-0.594115
2	1.186075	-0.130874	1.445125	-0.984965	-0.984965	-0.286634	-0.918124	-0.594115
3	0.872202	0.040846	1.464815	-0.973032	-0.973032	-0.157233	-0.918124	-0.594115
4	1.233452	0.392500	1.456939	-0.973032	-0.973032	-0.166476	-0.918124	-0.594115

We need to find out potential outliers for all sensory features as well as other numerical features based on threshold value. If z-score value greater than 3 or less than -3 considered as potential outlier.



Z-score technique identified 543 outliers in dataset based on threshold condition, where 300 outliers are having z-scores greater than 3 and 243 outliers are having z-scores less than -3.

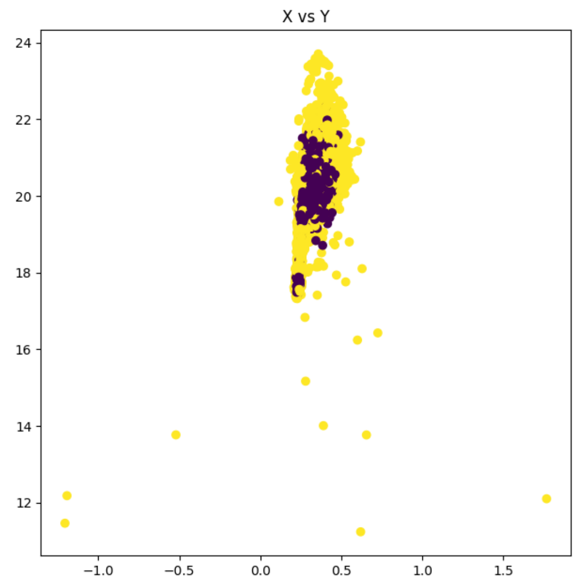
```
Number of outliers: 543
              X         Y     Speed  AgentCount   Density        Acc  \
845    0.696019  0.597652  0.366453   -0.459944 -0.459944  -5.610555
990   -0.147885 -0.374186 -0.386433   -0.316756 -0.316756  -3.931425
993   -0.073858 -0.262360 -0.371039   -0.263061 -0.263061  -6.642681
1113   0.069754  0.390343 -0.438344   -0.251129 -0.251129  -4.572267
1114   0.001649  0.330863 -0.363879   -0.251129 -0.251129  -3.888291
...         ...       ...       ...         ...       ...        ...
23224 -0.902956 -0.208190 -0.274020    0.524470  0.524470 -15.719226
23281 -0.580200  0.309128 -0.024848    0.250028  0.250028  -8.103062
23291 -0.131599  0.289634  0.040310    0.130705  0.130705   3.207189
23335 -0.233755  0.654062  0.544382   -0.227264 -0.227264  -3.614085
23533 -0.300379  0.778331  1.526392   -0.686657 -0.686657  -4.045421

       LevelOfCrowdness  Severity_level
845           -0.918124       -0.594115
990           -0.231452       -0.594115
993            0.455219       -0.594115
1113           0.455219       -0.594115
1114           0.455219       -0.594115
...                 ...             ...
23224          0.455219       -0.594115
23281          0.455219       -0.594115
23291          0.455219       -0.594115
23335          0.455219       -0.594115
23533         -0.918124       -0.594115

[543 rows x 8 columns]
```

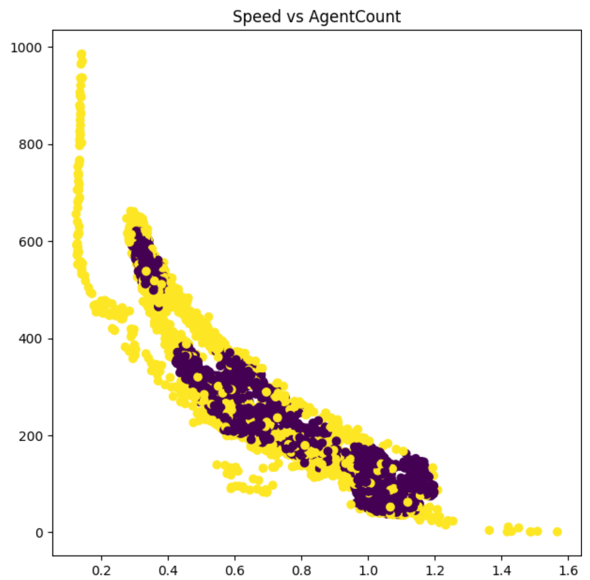
Module-5: Machine Learning Technique for Anomaly Detection (Isolation Forest)

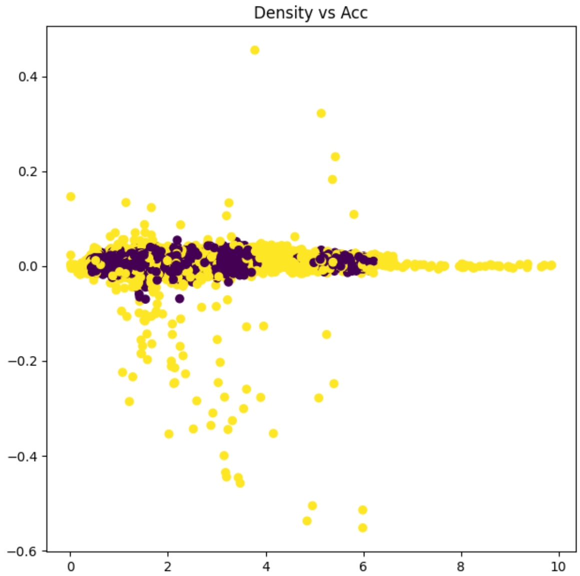
Isolation Forest Approach:

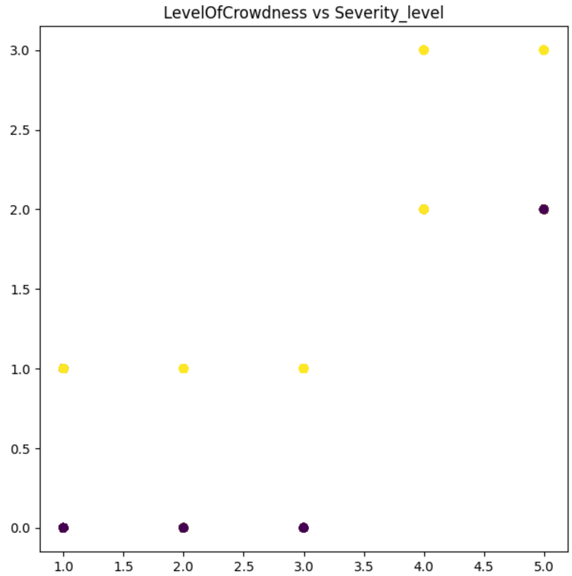
- The algorithm randomly selects features and partitions the data, isolating individual points.
- Anomalies are easier to isolate (shorter paths to isolate) than normal points (longer paths to isolate).
- An anomaly score is assigned based on how easily a point is isolated.

As dataset is already cleaned and formatted, to implement isolation forest we need to split dataset into training and test datasets. We can split train and test datasets in any ratio, but it is better to have more data for training. Hence, we splitted dataset for 75% training and 25% testing then trained model. By using 25% test dataset model tested and checked accuracy to determine whether it is working in efficient or not. Initially model got 73% accuracy score but after hypertuning the model observed that 82% accuracy score is max using isolation forest.





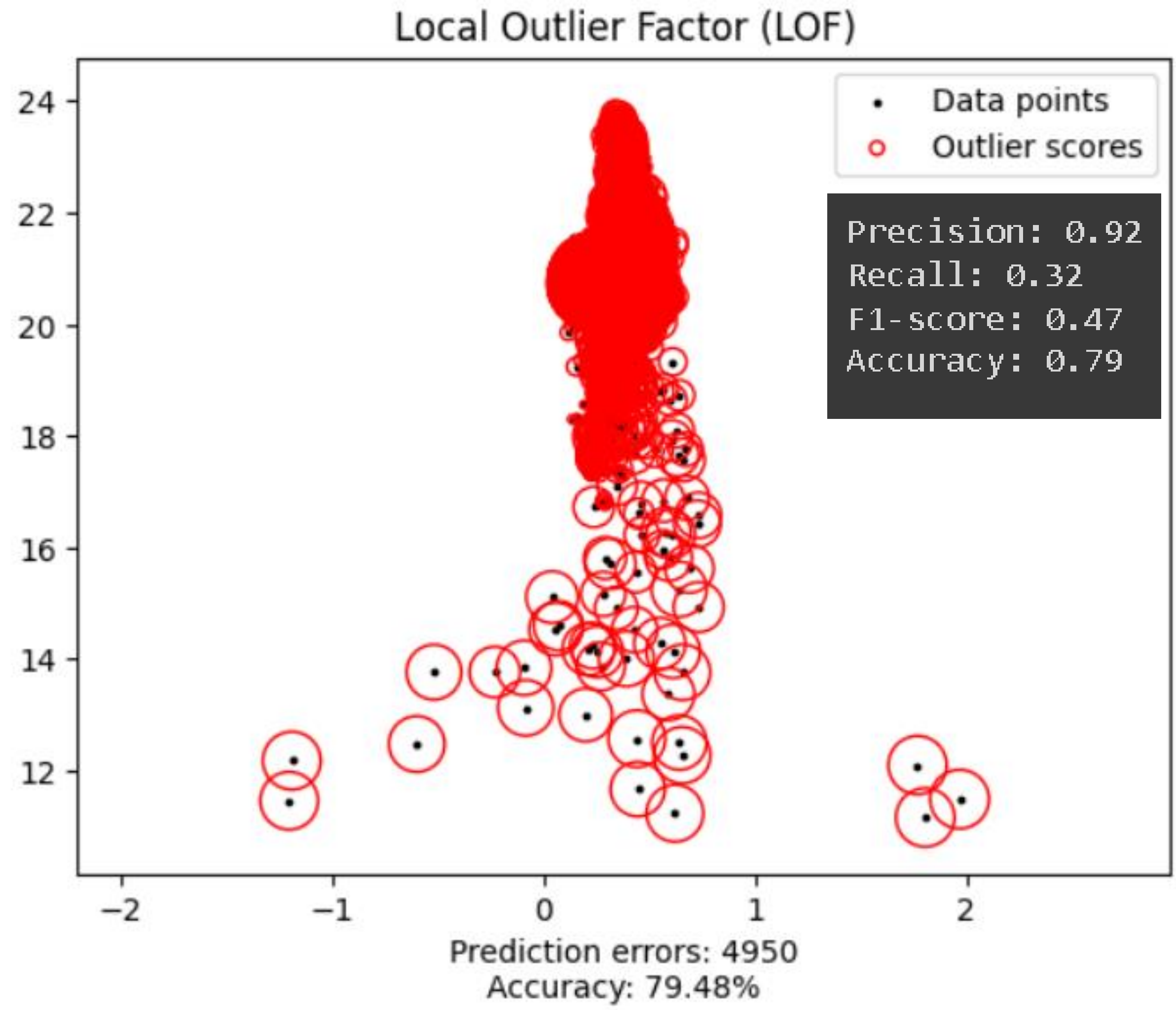




These visualizations are generated based on anomaly score of each numerical feature in dataset.

Module-6: Machine Learning Technique for Anomaly Detection (Local Outlier Factor)

The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbors. It considers as outliers the samples that have a substantially lower density than their neighbors.



Initially accuracy of model using local outlier factor is 63% but after hypertuning parameters in model it reaches to 79.48% accuracy.

Conclusion:

From the observations of accuracy in different models it cleared that they are not very high accurate for the prediction and due to illusion of identifying anomalies in regular conditions can leads to false predictions. However, these models can be used for understanding behaviour of ML models in anomalies detection of crowd which can further helps deep learning model analysis.

References:

https://archive.ics.uci.edu/dataset/613/smartphone+dataset+for+anomaly+detection+in+crowds

https://www.sciencedirect.com/science/article/pii/S0957417422008065