# Why Does My City Smell Bad
# By
# GROUP-1
# (Syed Hani Haider, Ramzi Adil, Karan Paresh)

## Summary:

The city of Portland has been collecting complaints from citizens regarding strong odors - in particular oil, that may be an underlying cause of symptoms such as headaches. As shown in the article [1], there is a general public concern that poor air quality may be causing health issues and that the poor air quality may be tied to the oil storage and processing facilities that are located along the Fore River. In response, the local government of Portland has reached out to our class. They requested that our team, along with another team, investigate odor complaints.

We want to determine if there are any correlations between weather events such as wind direction or temperature as well as smell complaints from citizens of Portland, and geographical locations. These data sets include relevant information such as odor description, location of complaint, temperature, etc. to achieve

## Hypotheses & Goals

We hypothesize that there is a correlation between weather and some features such as wind direction, temperature, windspeed, etc. Taking this into consideration, we also hypothesize that a model utilizing these features can predict the occurrence of a complaint.

We have 3 main goals:

1.) Clean up and process the weather files
2.) Consolidate the weather data and complaints together
3.) Analyze the consolidated dataset to determine if some weather features affect the occurrence of complaints.

## Data Set Description

Multiple raw csv files were obtained for this project. Each pod (weather measurement device) produced its own csv files. Each pod measured the same weather features. The measurements between pods can overlap in time. For example, Pod ‘SMRO3’ and pod ‘SMRO4’ have their individual measurements for a datetime of 09/15/2020 12:00:00 . The measurements taken by each pod were taken in hour intervals. The smell/complaint raw data files (see click fix and smell my city) obtained contains a description of the odor and the date and times the odor complaint was made.

## Data Processing

The multiple weather csv files were appended to each other to produce one weather csv. Visualization of the data showed that pod “SMRO4” produced many outliers so we excluded this dataset in further analyses. Since the pods overlapped each other in time, losing this pod was not a great loss in samples per hour.

The next step in processing was measurement aggregation. There are multiple measurements for the same time frames and our goal was to have a data frame that contained 1 set of weather measurements per hour. We did this by aggregating the pod measurements such that the measurements per hour represented the average measurement between all pods for that hour.

The processing for the complaints/smell data was performed by another team. This team appended the data from smell my city and see click fix together. We needed to merge this data to the weather data aggregation so we aggregated the complaints by hour. We then merged the data. This produced a new data frame with the average weather measurements per hour and the total number of complaints per hour. The columns composed of all NaN values were dropped, and we kept the columns we were interested in.

Merged Data Set

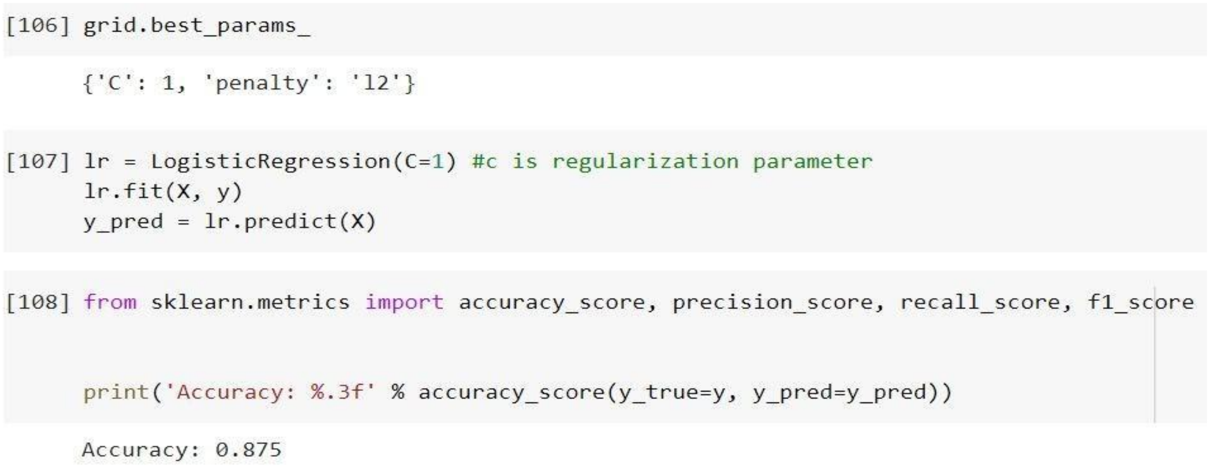
| Column | Datatype | Description |
|---|---|---|
| DateTime interval | datetime | The date and hour a set of measurements belong to |
| Complaint total | integer | The number of complaints made the particular Datetime interval |
| Hour Group | datetime.time | the hour that set of measurement were made, in 24 hour/military format |
| Day_of_week | string | The day of the week the set of measurement were made |
| Month_of_year | string | The month of the year the set of measurements belong to |
| Season | string | 'Winter', 'Spring', 'Summer' or 'Fall' |
| Temp Avg | float | The average temperature for a datetime interval |
| Baro Avg | float | The average barometer measurement for a datetime interval |
| Windspeed | float | The average wind speed measurement for a datetime interval |
| Gust | float | The average gust measurement for a datetime interval |
| Wind Direction | float | Wind direction represented in degrees (0-360) |
| Cardinal Direction | string | Wind direction represented as a string 'North','south'…etc |
| Heat Index | float | The average heat index for a datetime interval |
| Wind Chill | float | The average heat index for a datetime interval |
| Dew Point | float | The average heat index for a datetime interval |

## Modelling

**Logistic Regression:**
The model is used to predict whether there is a possibility of complaint occurrence for a particular hour, the data present has a time interval of an hour and using the logistic model the prospect of the complaints is calculated. As shown in fig 1

**Independent variables**: Temperature average, Humidity average, Heat Index and Dew point.

**Dependent variable:** Target complaints (0 =no complaints, 1= one or more complaints) in that hour.



```
[106] grid.best_params_

      {'C': 1, 'penalty': 'l2'}

[107] lr = LogisticRegression(C=1) #c is regularization parameter
      lr.fit(X, y)
      y_pred = lr.predict(X)

[108] from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score


      print('Accuracy: %.3f' % accuracy_score(y_true=y, y_pred=y_pred))

      Accuracy: 0.875
```

**Fig 1**

**Grid Search:**

The grid search technique is used to find the optimal value for the regularization parameter 'C' and best value for the parameter 'C' is (10**0=1). The penalty term used here is the ridge regularization (L2). The coefficients of the independent variables have a tendency to go towards zero before the optimal value reading.

**Model comparison:**

When we think of predicting the target variable at a certain hour/ time interval, it can be said that the independent variable will have a strong effect on how the model will perform. This is the main reason why Logistic Regression(lr) is preferred over the Ordinary Least Squares (OLS). Ordinary Least Square regression is not built for binary classification (0 and 1). Logistic regression performs a better job at classifying binary data points and has a better logarithmic loss function as opposed to least squares regression, theoretically speaking. Logistic regression is the best model to utilize in predicting the occurrence of complaints based on weather features.

## Data operation

The merged data had 7441 entries for each of the 28 columns present in the data which had been split in the ratio of 90:10 for training and testing, respectively. To optimize the data and in order to use it for the further analysis and modelling the columns with only null values were dropped and rows with any null values were dropped. Only two rows had null values so we still had 7439 entries to analyze.

**One-hot Encoding:**

The One-hot encoding is done for the cardinal direction column to convert the values like (North, South, East, West) to numeric which makes the principal component analysis easier as we reduce the dimensions of similar data type columns to the least possible dimension.

**Principal component analysis:**

The PCA modelling is done for the data with one hot encoding. The cumulative explained variance is calculated with respect to the number of components that needs to be assigned to the PCA class. This tells us how many components explain the variance in the dependent variable.

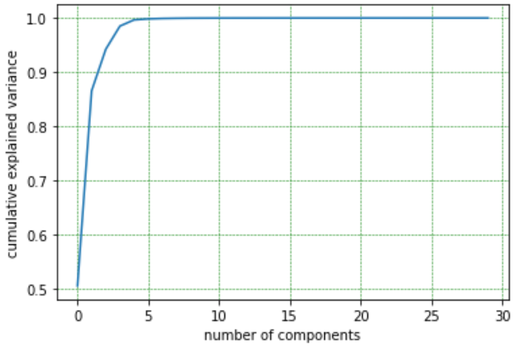
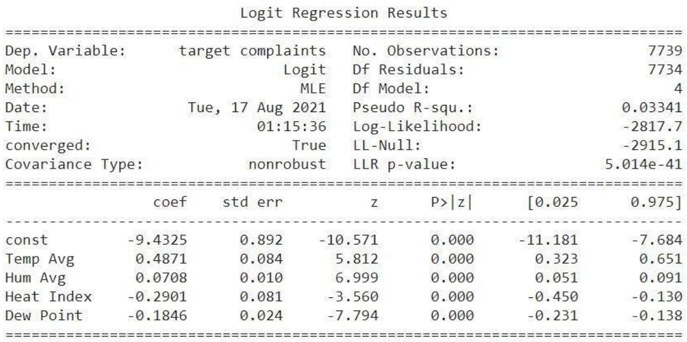


Figure **2:**The graph shows the number components required are "4".



```
                      Logit Regression Results
==============================================================================
Dep. Variable:      target complaints   No. Observations:                 7739
Model:                          Logit   Df Residuals:                     7734
Method:                           MLE   Df Model:                            4
Date:                Tue, 17 Aug 2021   Pseudo R-squ.:                 0.03341
Time:                        01:15:36   Log-Likelihood:                -2817.7
converged:                       True   LL-Null:                       -2915.1
Covariance Type:            nonrobust   LLR p-value:                 5.014e-41
==============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          -9.4325      0.892    -10.571      0.000     -11.181      -7.684
Temp Avg        0.4871      0.084      5.812      0.000       0.323       0.651
Hum Avg         0.0708      0.010      6.999      0.000       0.051       0.091
Heat Index     -0.2901      0.081     -3.560      0.000      -0.450      -0.130
Dew Point      -0.1846      0.024     -7.794      0.000      -0.231      -0.138
==============================================================================
```

Figure **3: stat model for feature selection**

**Feature selection:**

Figure 3 was produced by backwards selection. We removed every feature with a p-value over 0.05, one by one until only the most significant features remained. This method revealed the most important features which significantly affected the target variable. The most important features that affected the target variable were Temp avg, Hum avg, Heat index and dew point.

**Training:**

The model was trained using the k-fold cross validation technique with a value cv=18. The train-test split was done on the data with the test size of “0.10”. The model was evaluated using the confusion matrix.

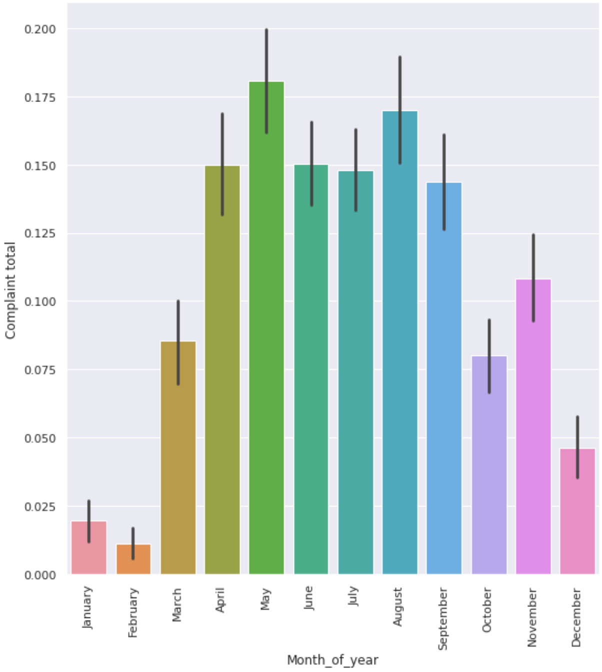
**Validation Curve:**

Post model training, the model was validated using the test data. The parameter range for the logistic model was set from ($10^{-10}$ to $10^{4}$). The mean scores for both train and test data sets were calculated. The training score and the testing score both ranged between 0.87 to 0.88 when the model was performing at its best.

## Results

### Complaints Grouped By Months

Plotting the weather features over time in one day intervals and observing the clustering of complaints show that between the months of January-March, there is a decrease in complaints. Plotting the average complaints by month show that the most complaints occur mostly in the summer months - June, July, August and partly in the spring month of April. The fewest complaints occur in the winter months - December, January, and February.



:

Figure 4: Complaints grouped by months. The summer months have the most complaints on average. The winter months have the fewest complaints.

### Comparing winter and summer complaint

The weather map provides more insight into this phenomenon. Each dot in figure 5 and 6 represents the location of a complaint made. There is a clear difference in the densities of the complaints between the summer and winter months. It is possible that since fewer people come out in the winter, less people make complaints about odors.

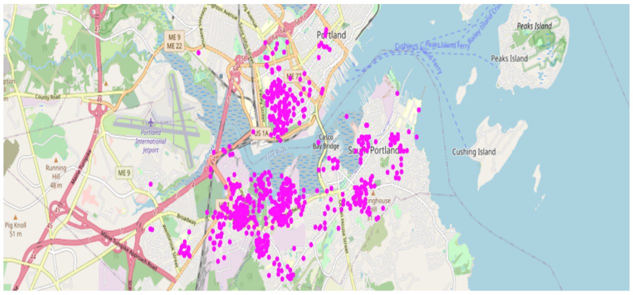


Figure 5: Summer complaints

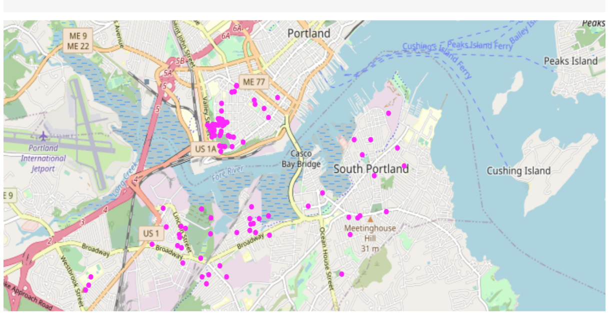
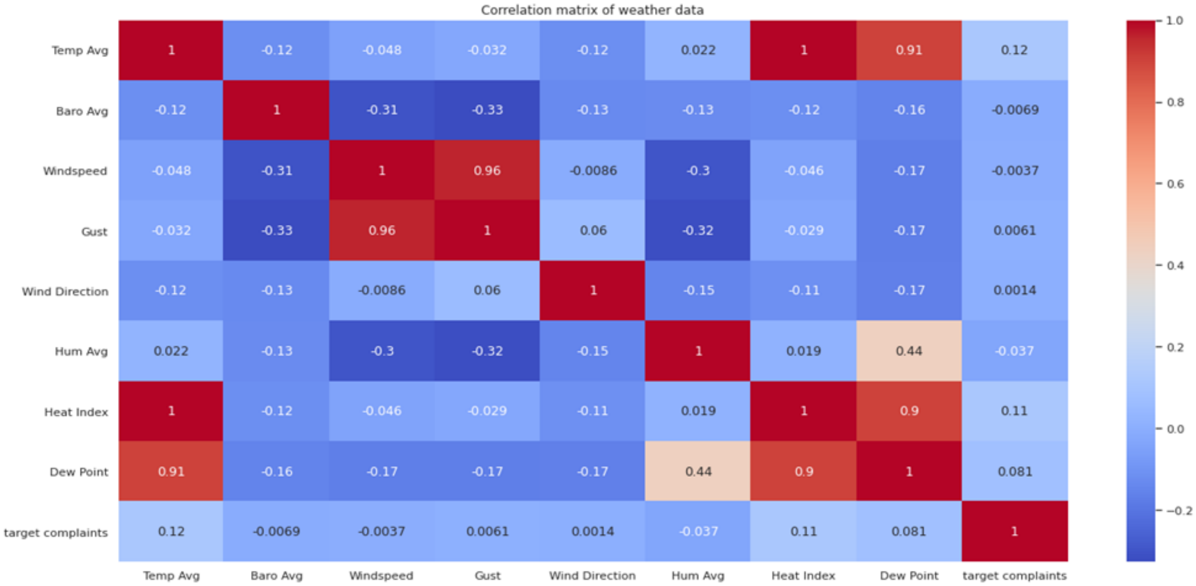


Figure 6: winter complaints

A correlation matrix is used to find any correlation between the independent and dependent variables.. This will help us evaluate which variable should be used to find a linear relationship with each other. The results of a correlation matrix analysis shows that the presence of complaints (target complaints) is not strongly correlated with any one weather feature as seen in figure 7

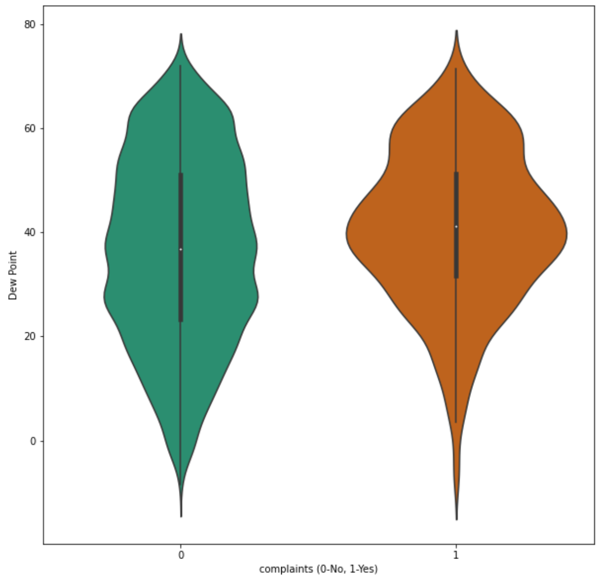

Correlation matrix of weather data

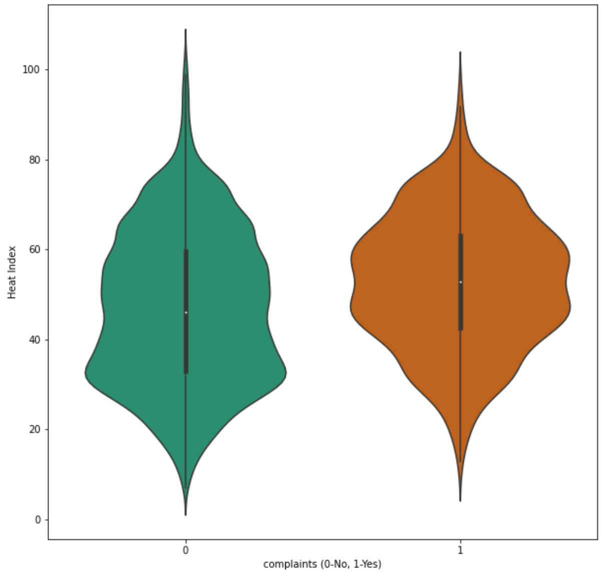
| | Temp Avg | Baro Avg | Windspeed | Gust | Wind Direction | Hum Avg | Heat Index | Dew Point | target complaints |
|---|---|---|---|---|---|---|---|---|---|
| Temp Avg | 1 | -0.12 | -0.048 | -0.032 | -0.12 | 0.022 | 1 | 0.91 | 0.12 |
| Baro Avg | -0.12 | 1 | -0.31 | -0.33 | -0.13 | -0.13 | -0.12 | -0.16 | -0.0069 |
| Windspeed | -0.048 | -0.31 | 1 | 0.96 | -0.0086 | -0.3 | -0.046 | -0.17 | -0.0037 |
| Gust | -0.032 | -0.33 | 0.96 | 1 | 0.06 | -0.32 | -0.029 | -0.17 | 0.0061 |
| Wind Direction | -0.12 | -0.13 | -0.0086 | 0.06 | 1 | -0.15 | -0.11 | -0.17 | 0.0014 |
| Hum Avg | 0.022 | -0.13 | -0.3 | -0.32 | -0.15 | 1 | 0.019 | 0.44 | -0.037 |
| Heat Index | 1 | -0.12 | -0.046 | -0.029 | -0.11 | 0.019 | 1 | 0.9 | 0.11 |
| Dew Point | 0.91 | -0.16 | -0.17 | -0.17 | -0.17 | 0.44 | 0.9 | 1 | 0.081 |
| target complaints | 0.12 | -0.0069 | -0.0037 | 0.0061 | 0.0014 | -0.037 | 0.11 | 0.081 | 1 |

Figure 7: Correlation matrix. This matrix shows that there is no significant correlation between the presence of complaints(target complaints) and the weather features.

### Violin plots

For certain features, the distribution of complaints seems to be affected by the variable values. For example, on the days when complaints were made, most of them occurred on days where the temperature was close to 40 degrees Fahrenheit. We see a similar value influence in the following weather variables: dew point, heat index.



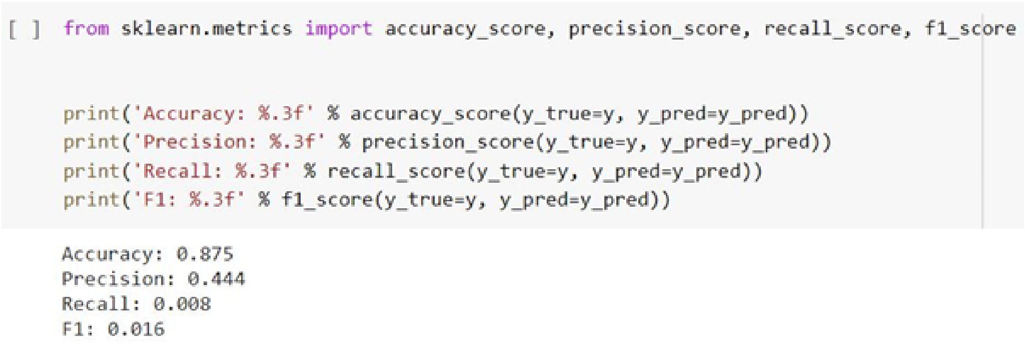


As shown in figure 3, The logistic model analysis showed a correlation between the following features: average temperature, humidity, heat index, and dew point. These features together strongly correlate with the occurrence of complaints even if they do not correlate with complaints individually.

## Model Evaluation

Model evaluation metrics are essential to quantify any predictive model's performance. The choice of evaluation metrics depends on a given machine learning task. For Model Evaluation, we use a confusion matrix that calculates the precision score, F1 score and the accuracy score of the model.

**Accuracy = (TP + TN) / (TP + TN + FP + FN) =# of correct predictions total # of samples in the dataset**



```
[ ]  from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score


     print('Accuracy: %.3f' % accuracy_score(y_true=y, y_pred=y_pred))
     print('Precision: %.3f' % precision_score(y_true=y, y_pred=y_pred))
     print('Recall: %.3f' % recall_score(y_true=y, y_pred=y_pred))
     print('F1: %.3f' % f1_score(y_true=y, y_pred=y_pred))
```

```
Accuracy: 0.875
Precision: 0.444
Recall: 0.008
F1: 0.016
```

The accuracy score is "87.5%" which concludes that our model predicts the occurrence of a complaint relatively well. This shows that we can predict the occurrence of the complaint based on the weather features of any new weather entries.

## Conclusion:

The results show that the winter months have the fewest complaints and that the summer months have the most complaints. The model analysis shows that the features that strongly correlate with the occurrence of complaints are temperature, humidity, heat index, and dew point. The results taken together, show that on hotter and more humid days, complaints are more likely to be made.

## Real World Use Case:

If the complaints are reflective of strong odors and if strong odors indicate poor air quality, then on hotter and more humid days, a warning can be issued. This warning would inform people that there will be a high likelihood of harmful air pollutants present in the air.. This would provide the people of Portland the opportunity to avoid exposing themselves to such air pollutants.

## Future Goals:

A future goal is to incorporate the oil vessel data sets into this project. It would be interesting to see if there is a relationship between weather features, amount of oil being brought/ type of oil being brought by the vessels, and complaints.

Another important future goal is to analyze the wind direction in terms of North/South and East/West relationships. Our data analyzes wind direction in degrees (0 - 360) which may not be the best for modeling. This type of analysis may indicate that wind direction may be an important feature in complaints as we initially believed.

## Statement of contribution:

| | |
|---|---|
| **Ramzi Adil** | **Data pre-processing, processing, and analysis** |
| **Karan Paresh** | **Data visualization and data modelling** |
| **Syed Hani Haider** | **data stream lining, file coordination, map visualization** |

**References :**

1. https://ecosorbindustrial.com/why-odors-smell-worse-warm-weather/

2.https://smellmycity.org/data

3.https://rainwise.net/weather/SMRO3

4.https://seeclickfix.com/portland_2

5.https://github.com/ds5110/stinky

6.www.wmtw.com/article/portland-south-portland-residents-ask-state-to-do-more-about-smell-from-large-oil-tanks/28777159

7. https://github.com/ds5110/stinky/blob/master/Geo_team.ipynb

## Appendix:

**Link to source code: [https://github.com/ds5110/stinky](https://github.com/ds5110/stinky)**

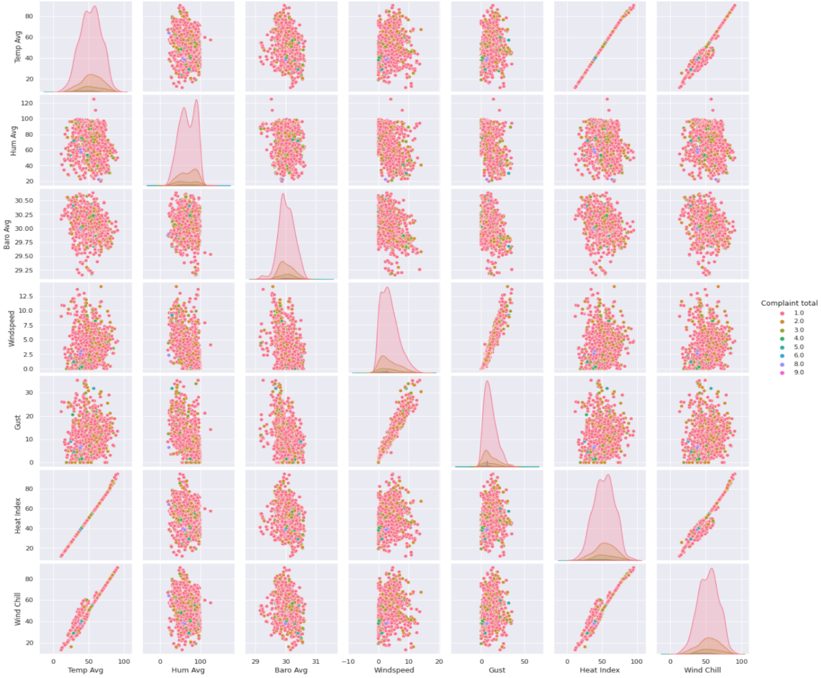


**Fig 1: The pairplot indicates that the classes overlap each other and there is no separable clustering**

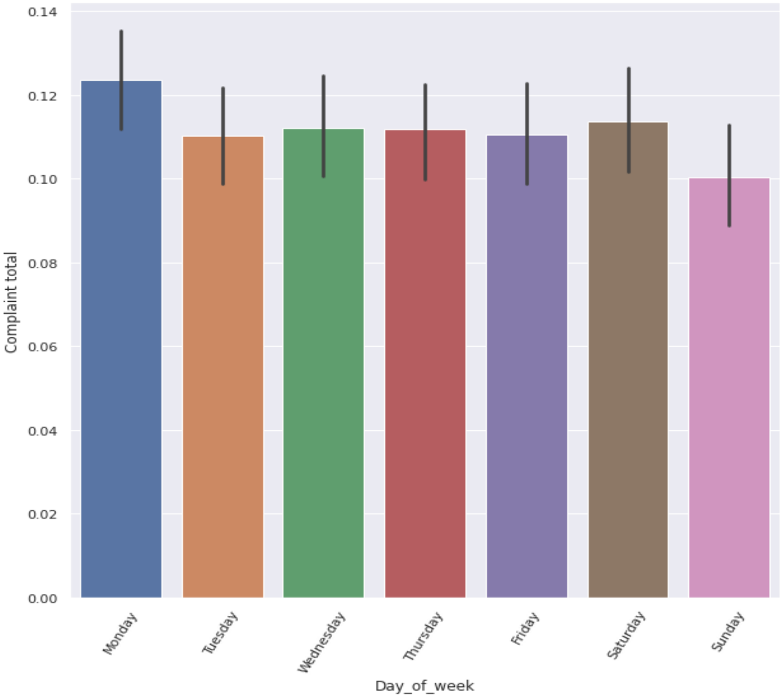


**Fig 2: The plot shows that the overall the number of complaints per weekday are similar.**

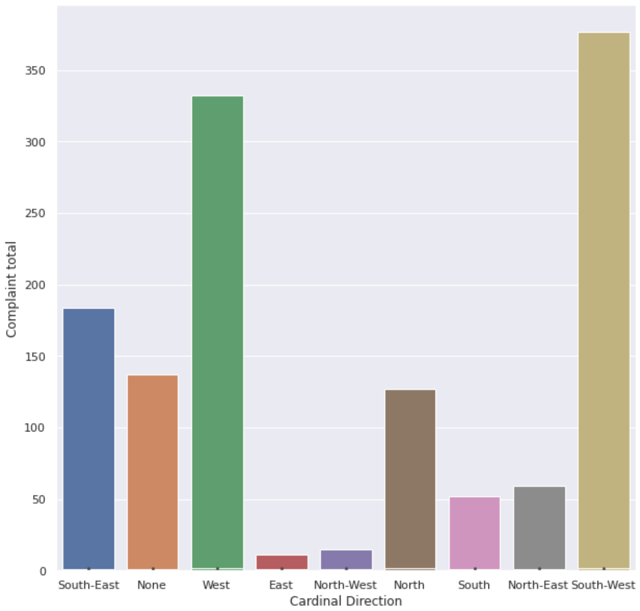


**Fig 3: The plot shows that the wind direction of south-west, west and south east impacted the number of complaints.**

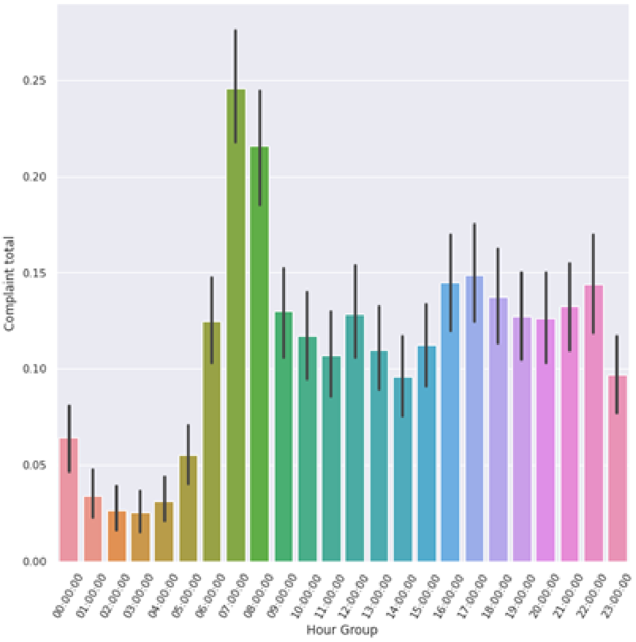


**Fig 4: The plot shows that the number of complaints occur more frequently in the morning hours (6- 8 am) and the evening hours (5 pm onwards).**

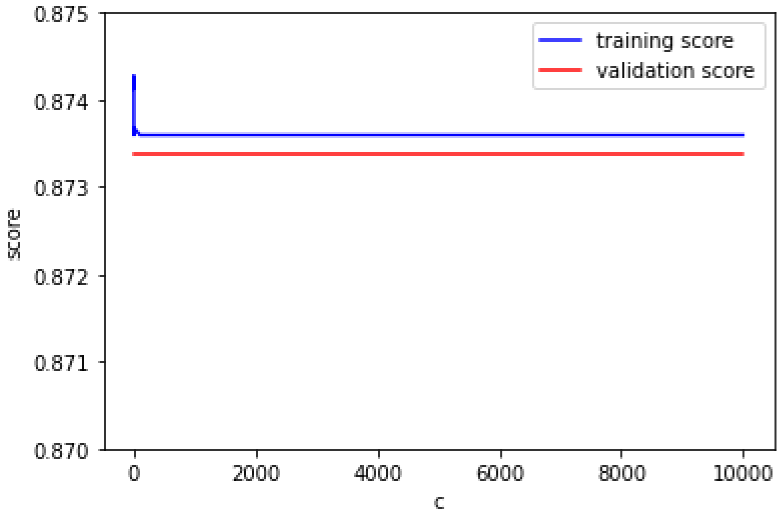


**Fig 5: The training score and the validation score has high bias before the optimum value c=1 and has high variance for the higher values.**