Why Does My City Smell bad ?

GROUP-1
by
Syed Hani Haider, Ramzi Adil, Karan Paresh

Introduction

- The city of Portland has been collecting complaints from citizens regarding strong odors and weather data from multiple devices scattered along the Fore River where the odors are thought to originate from.

- There is a general public concern that the foul odors may be indicative of harmful particles floating in the air. These particles may cause the air quality to worsen and impact the health of the citizens.

- In response, the local government of Portland has requested that we investigate odor complaints along with the collected weather data.

- We want to investigate the relationship between the odor complaints and weather

Goals:

- Coordinate the data into 1 data frame
- Visualize weather data
- Visualize merged data
- Create a map to show complaints locations
- Perform an analysis of the merged data with models

Files

- Multiple csv files with only weather information
- Multiple csv files with only complaint information
- The files are uploaded to a github repo where our functions can directly access them via a github link

Steps of Tidying

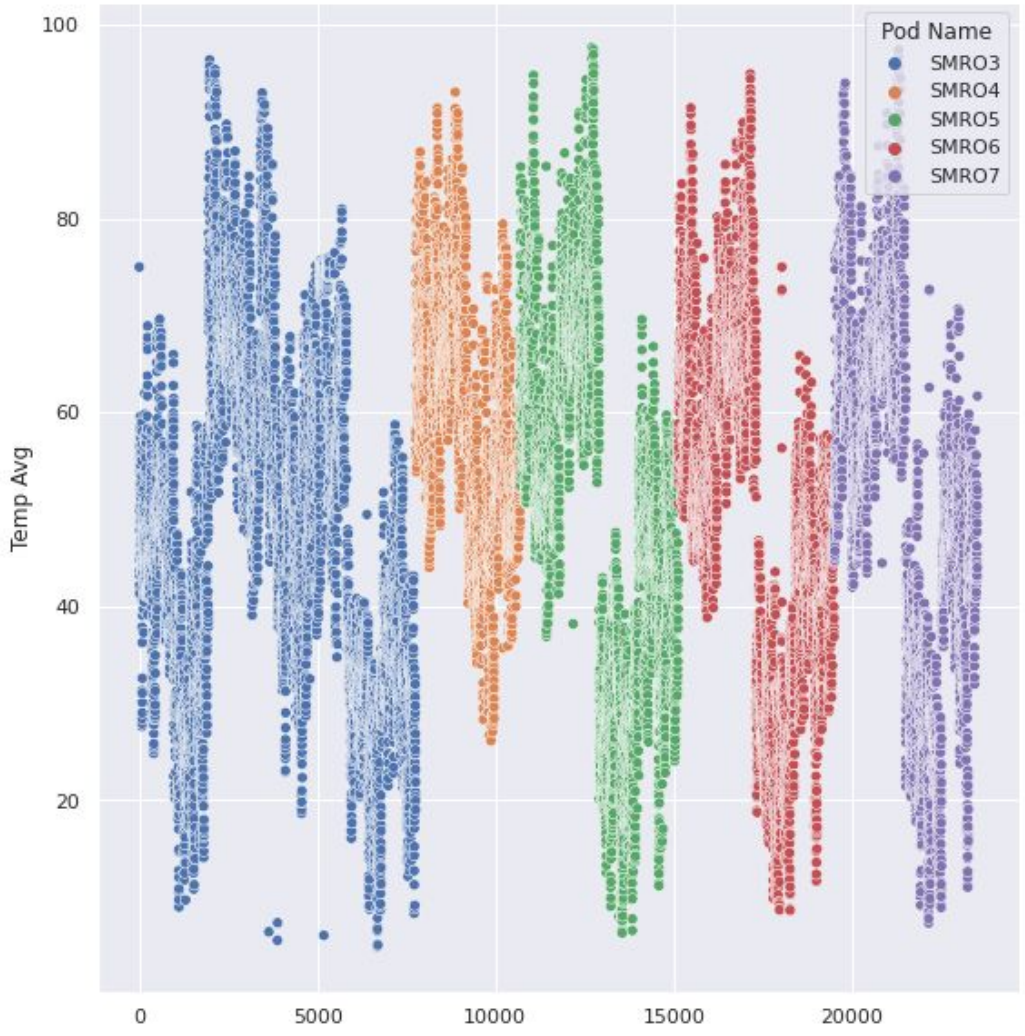
- Appending all weather files together into a singular data frame containing all weather information for all available time ranges (weather data is aggregated by hour)
- The smell data was also merged into a singular dataframe (This step was performed by another group).
- We then aggregated the complaints by hour.
- Then we merged that weather and smell data together resulting in a dataframe with all the weather events and complaint totals by hour

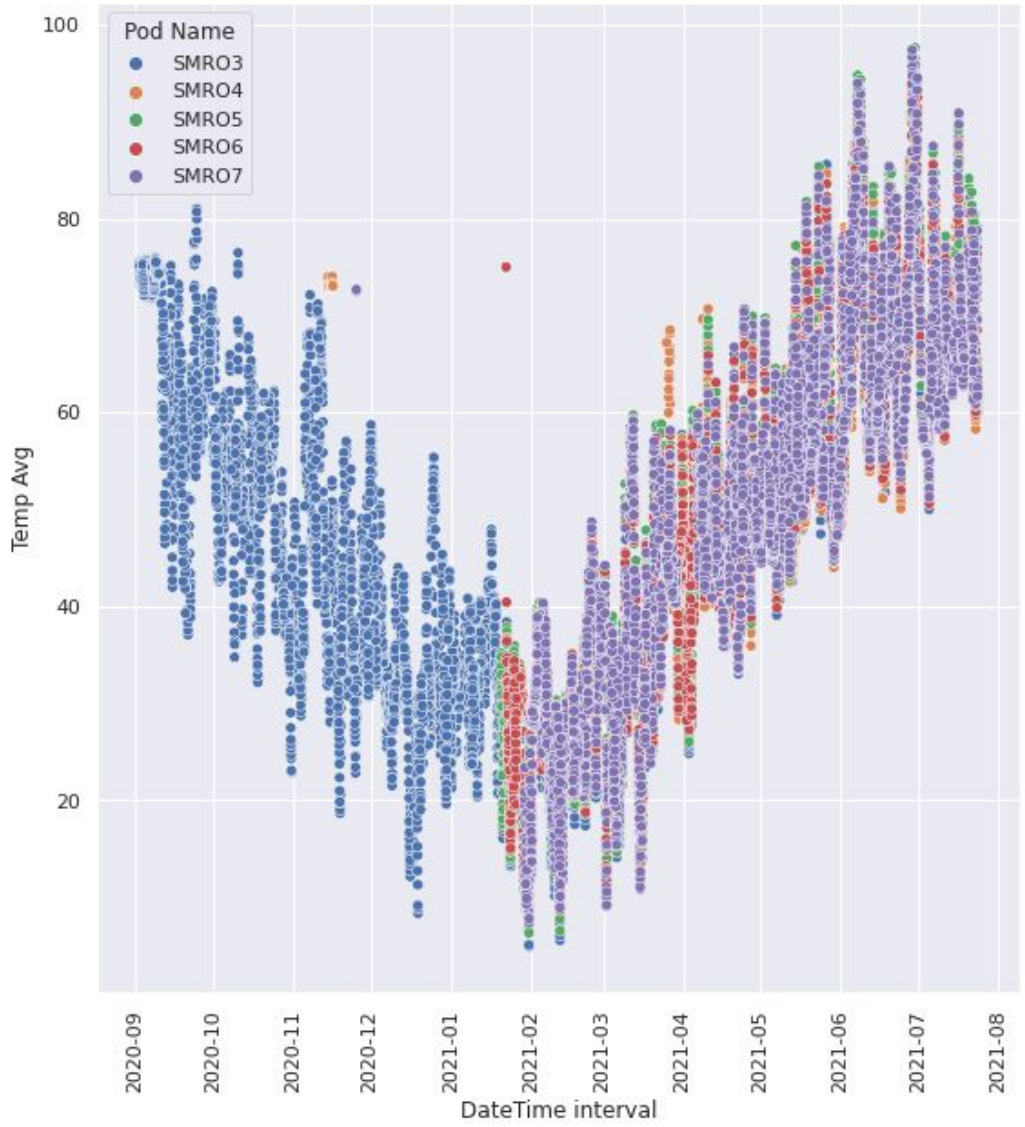
Weather EDA

- We are interested in investigating the following weather features: temperature, air pressure, humidity, windspeed, wind direction, dew point, heat index, wind chill

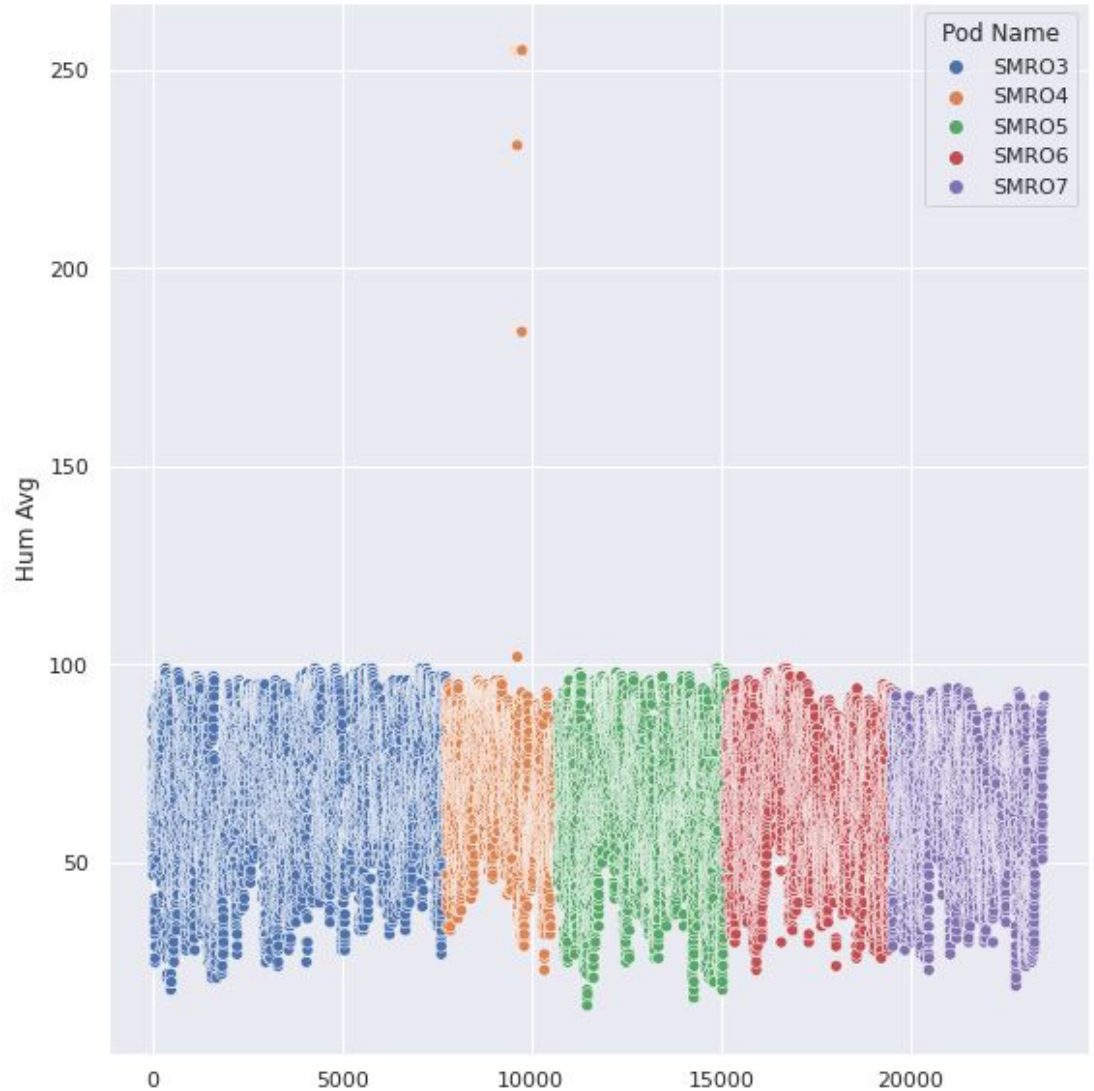
DateTime interval	Temp Avg	Baro Avg	Windspeed	Gust	Wind Direction	Cardinal Direction	Heat Index	Wind Chill	Dew Point
2020-09-02 19:00:00	75.5900	29.8200	0.000000	0.000	135.00	South-East	75.5900	75.600	61.200
2020-09-02 20:00:00	75.1300	29.8000	0.000000	0.000	135.00	South-East	75.1300	75.100	60.700
2020-09-02 21:00:00	74.7200	29.7900	0.000000	0.000	135.00	South-East	74.7200	74.700	60.500
2020-09-02 22:00:00	74.3300	29.7800	0.000000	0.000	135.00	South-East	74.3300	74.300	60.400
2020-09-02 23:00:00	74.1100	29.7600	0.000000	0.000	135.00	South-East	74.1100	74.100	60.200
...	...	...	...	...	...	...	...	...	...
2021-07-23 04:00:00	61.1050	30.0825	1.388335	4.250	253.00	South-West	61.1050	61.100	57.025
2021-07-23 05:00:00	60.4850	30.0850	1.231668	3.625	264.50	West	60.4850	60.475	57.000
2021-07-23 06:00:00	60.5300	30.0950	1.045000	3.775	252.75	None	60.5300	60.525	57.025
2021-07-23 07:00:00	61.1050	30.1075	0.692085	3.350	263.75	None	61.1050	61.125	57.050
2021-07-23 08:00:00	65.8975	30.1125	1.368748	5.900	263.75	None	65.8975	65.875	57.475

Average Temp

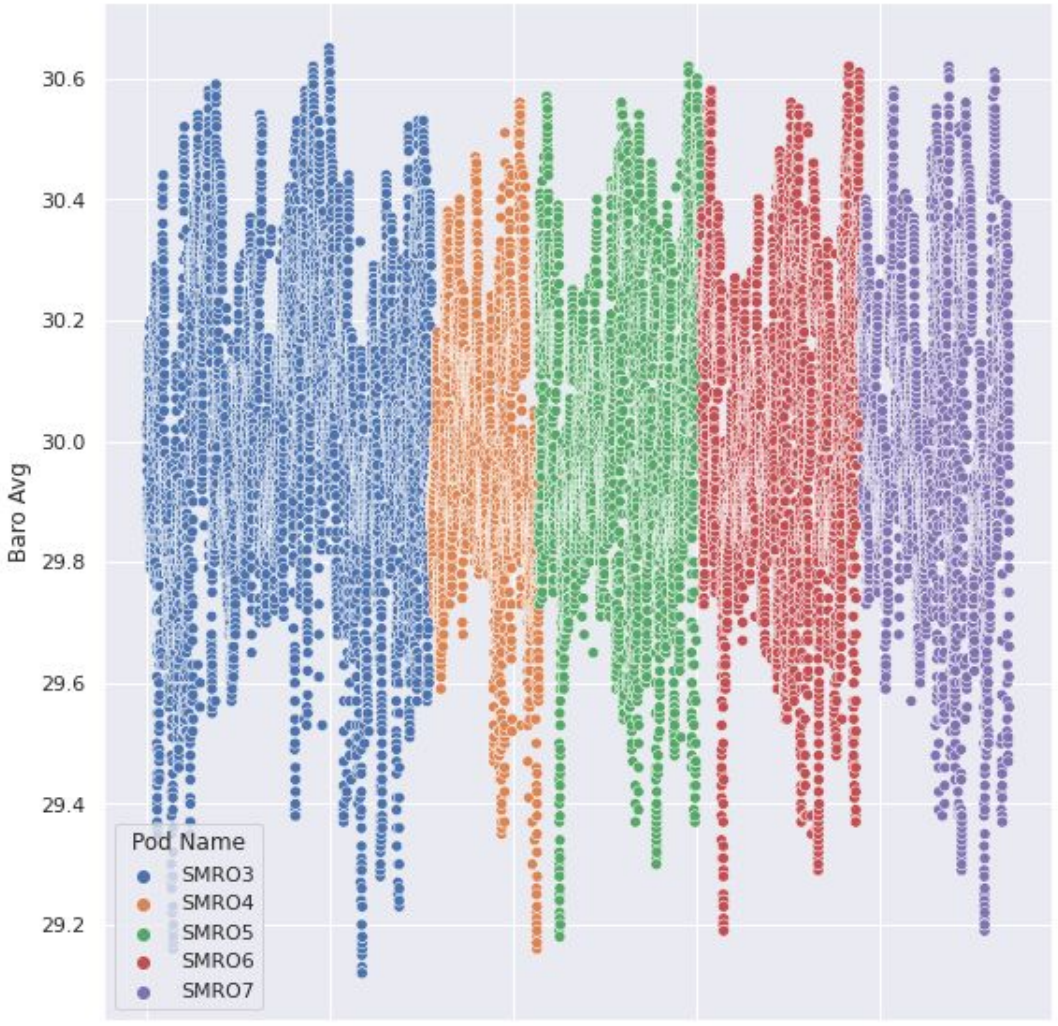


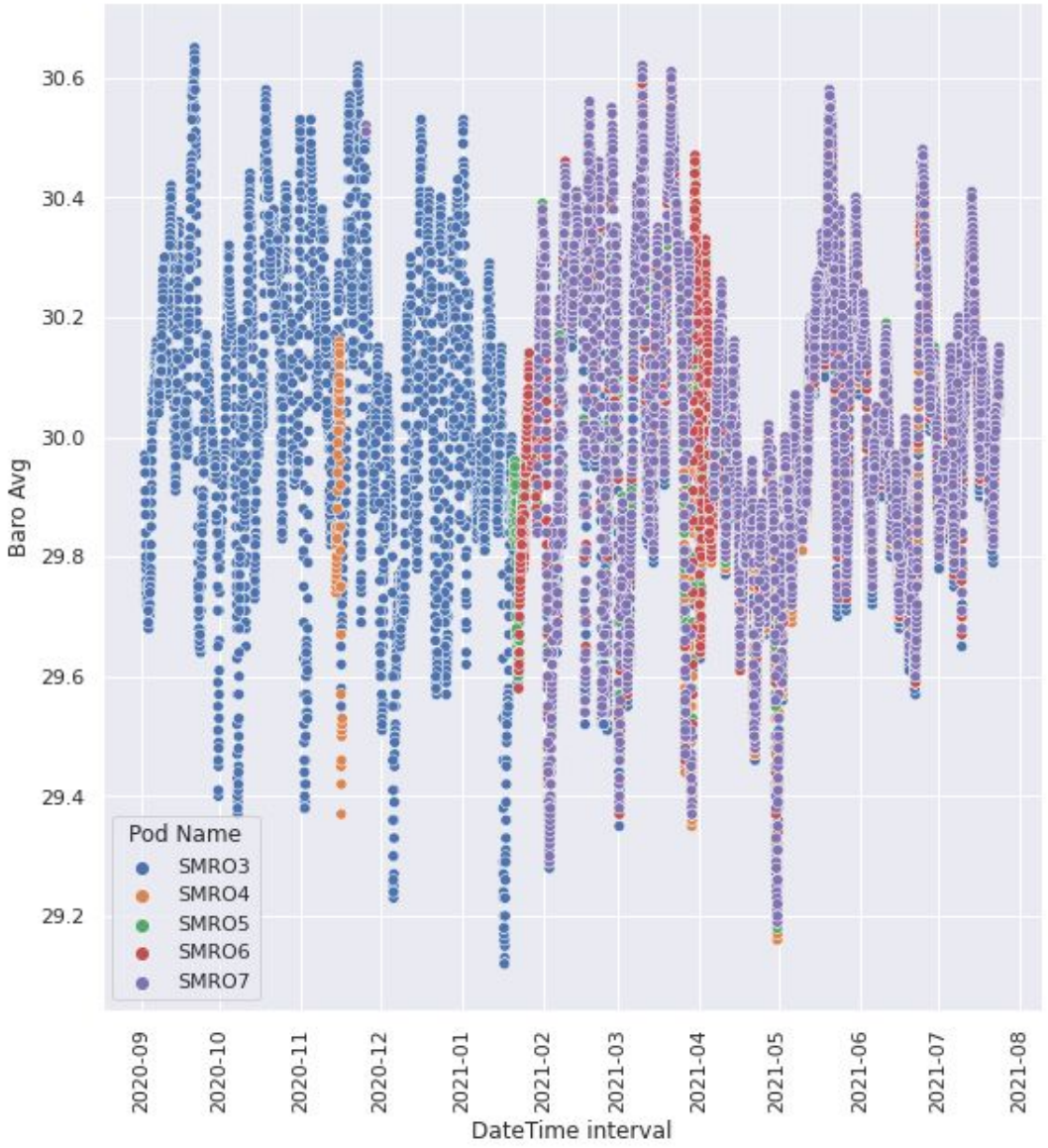


Average Humidity

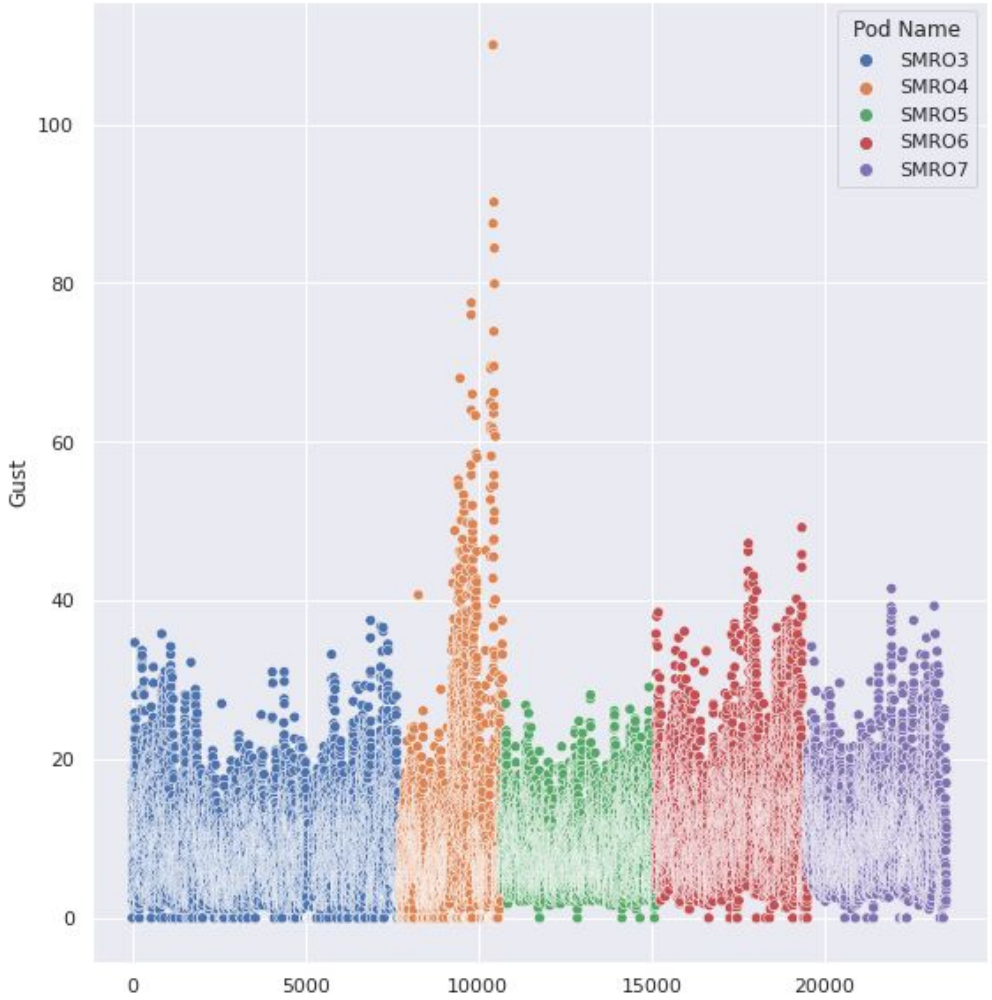


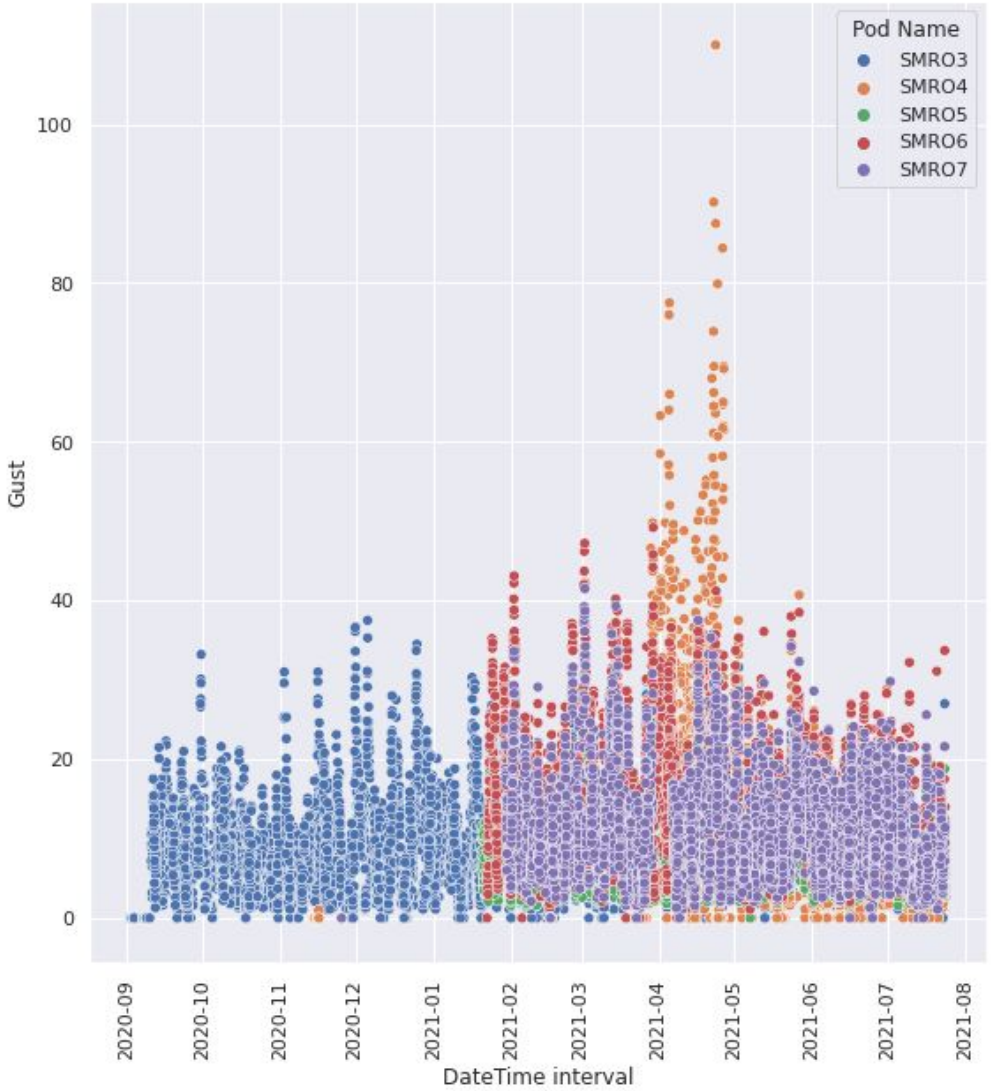
Air Pressure



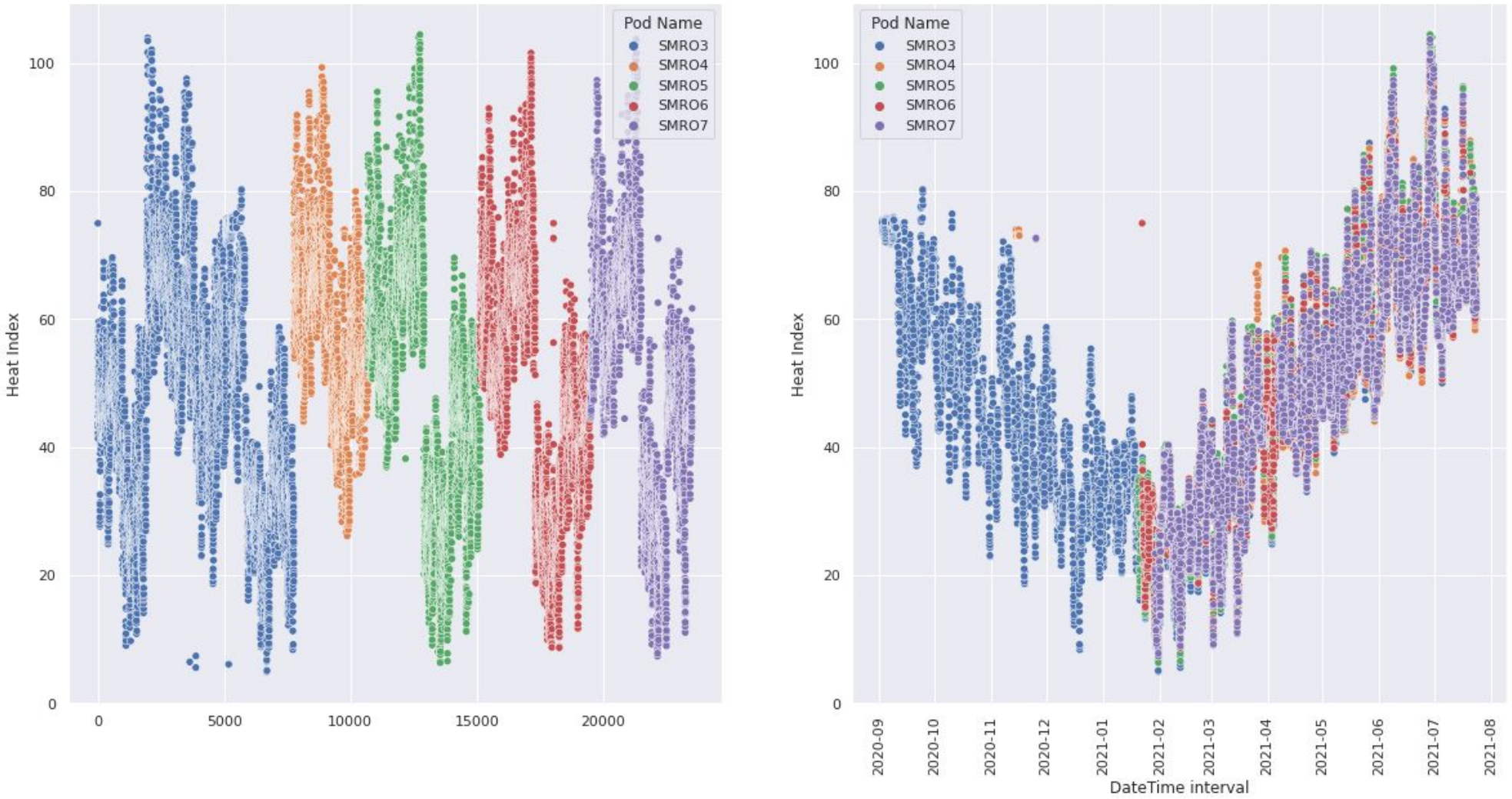


Gust

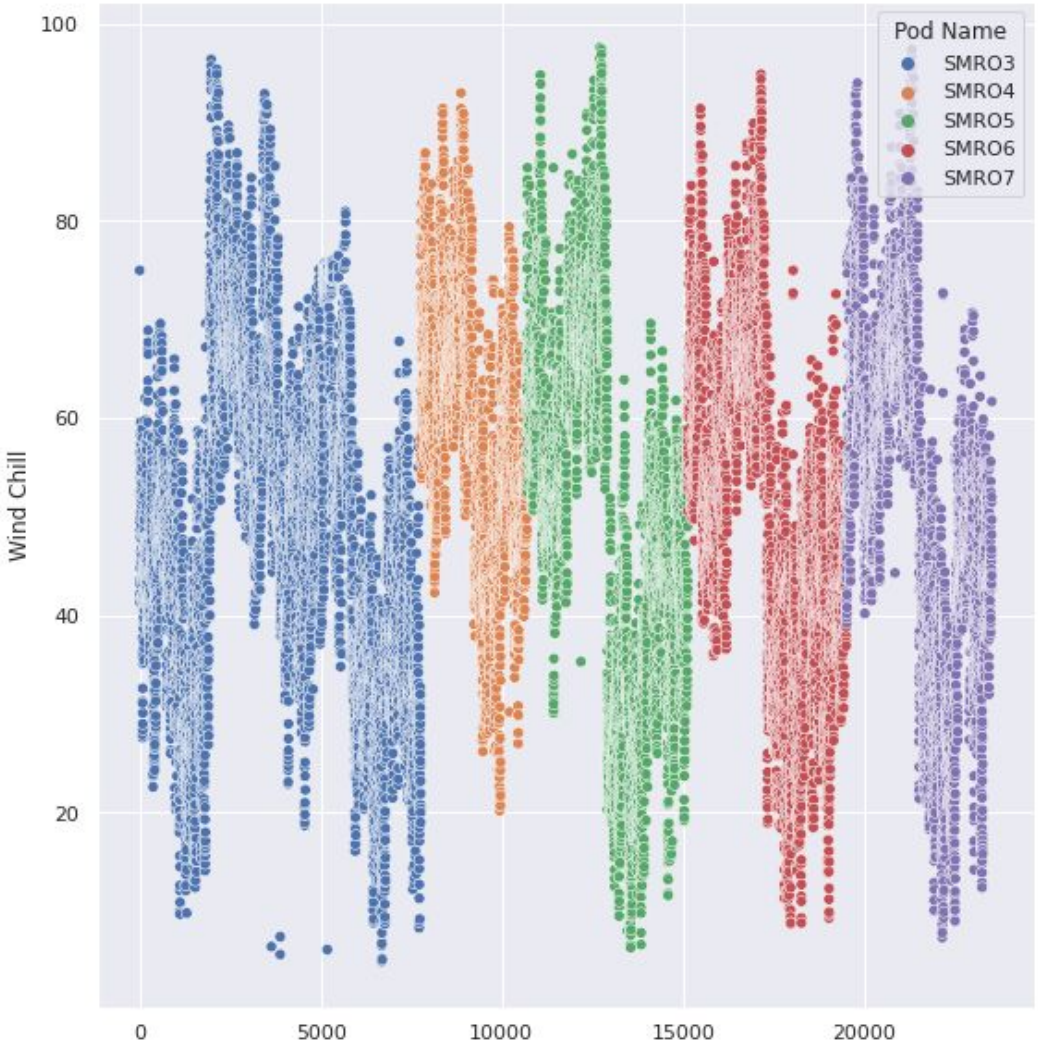

Pod Name
SMRO3
SMRO4
SMRO5
SMRO6
SMRO7
Gust
0
20
40
60
80
100
0
5000
10000
15000
20000

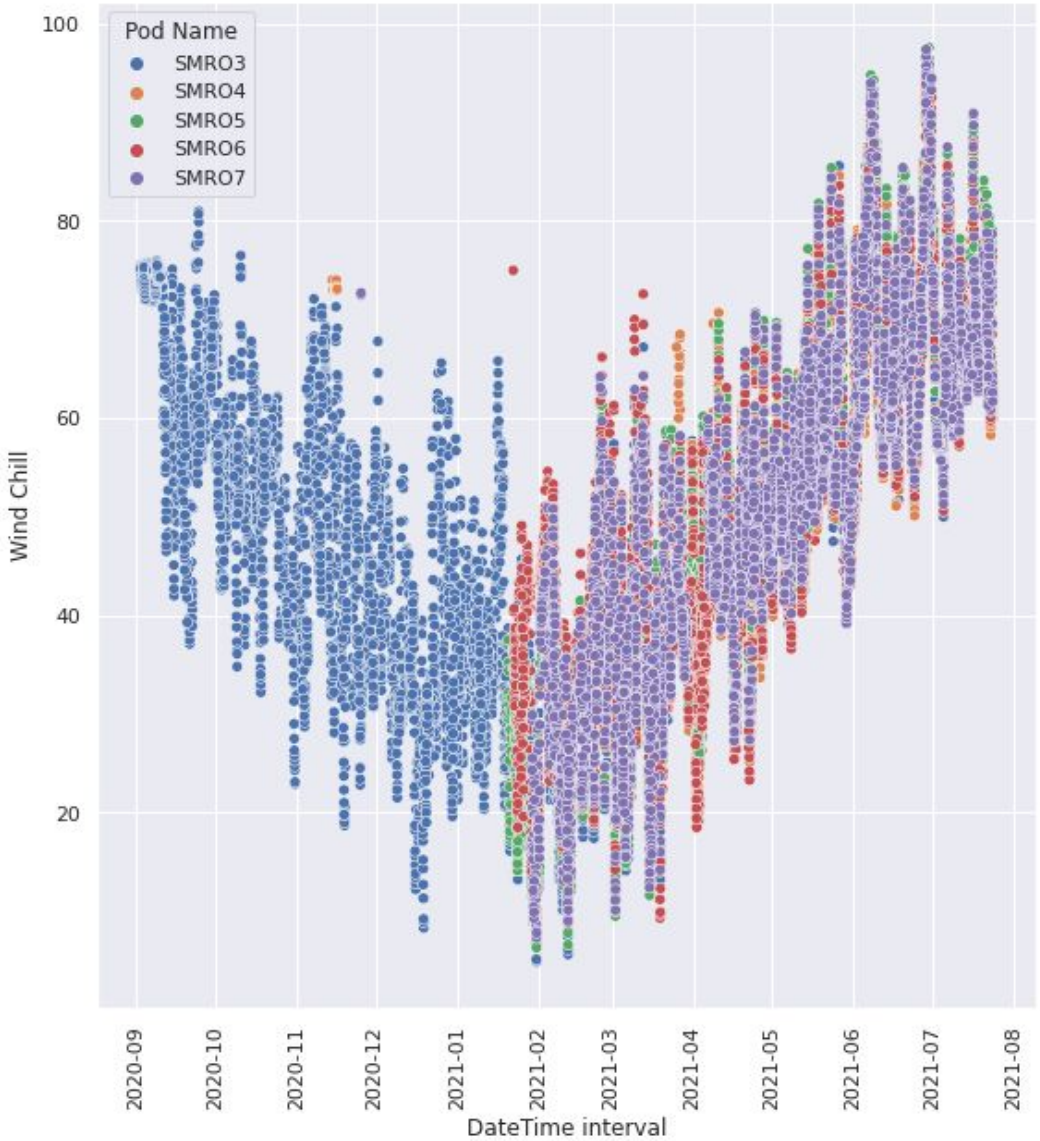

Pod Name
SMRO3
SMRO4
SMRO5
SMRO6
SMRO7
Gust
0
20
40
60
80
100
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
DateTime interval

Heat Index

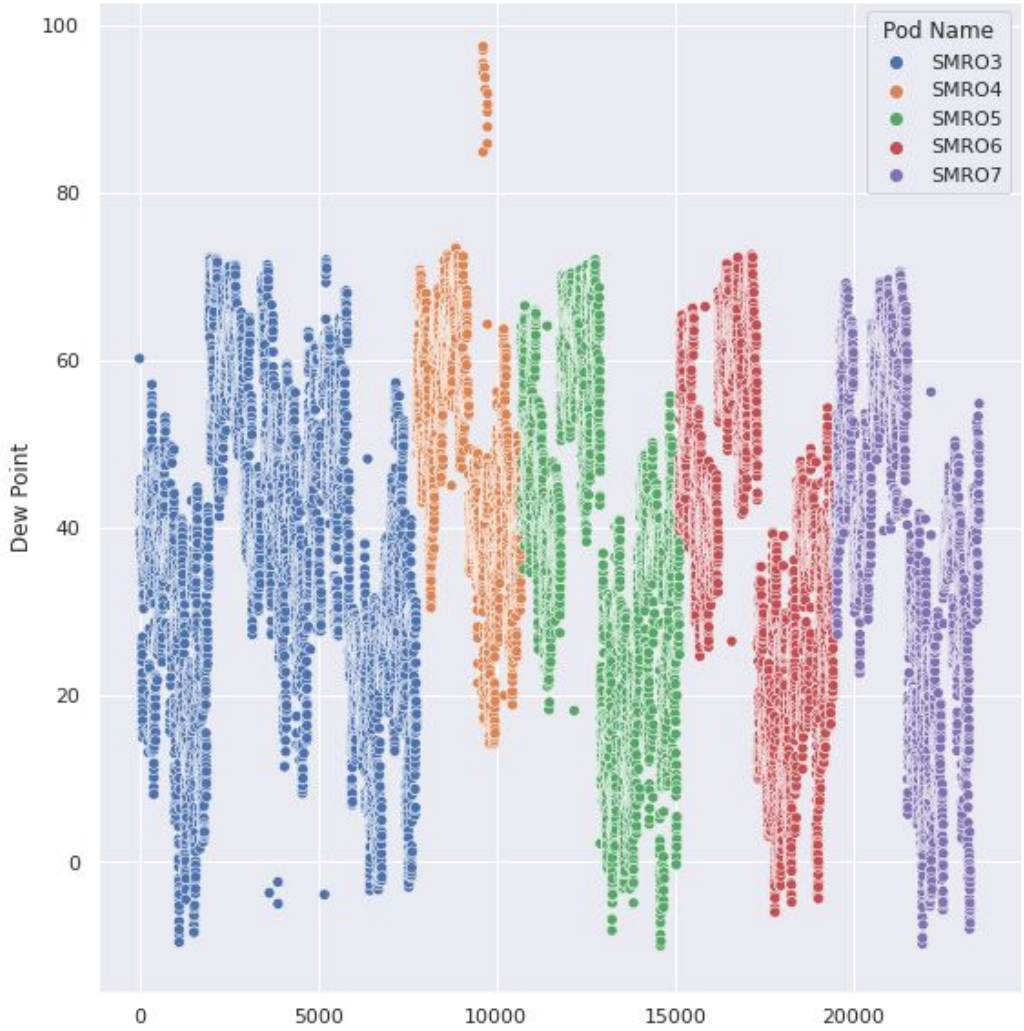

Pod Name
SMRO3
SMRO4
SMRO5
SMRO6
SMRO7
Heat Index
0
20
40
60
80
100
0
5000
10000
15000
20000
Pod Name
SMRO3
SMRO4
SMRO5
SMRO6
SMRO7
Heat Index
0
20
40
60
80
100
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
DateTime interval

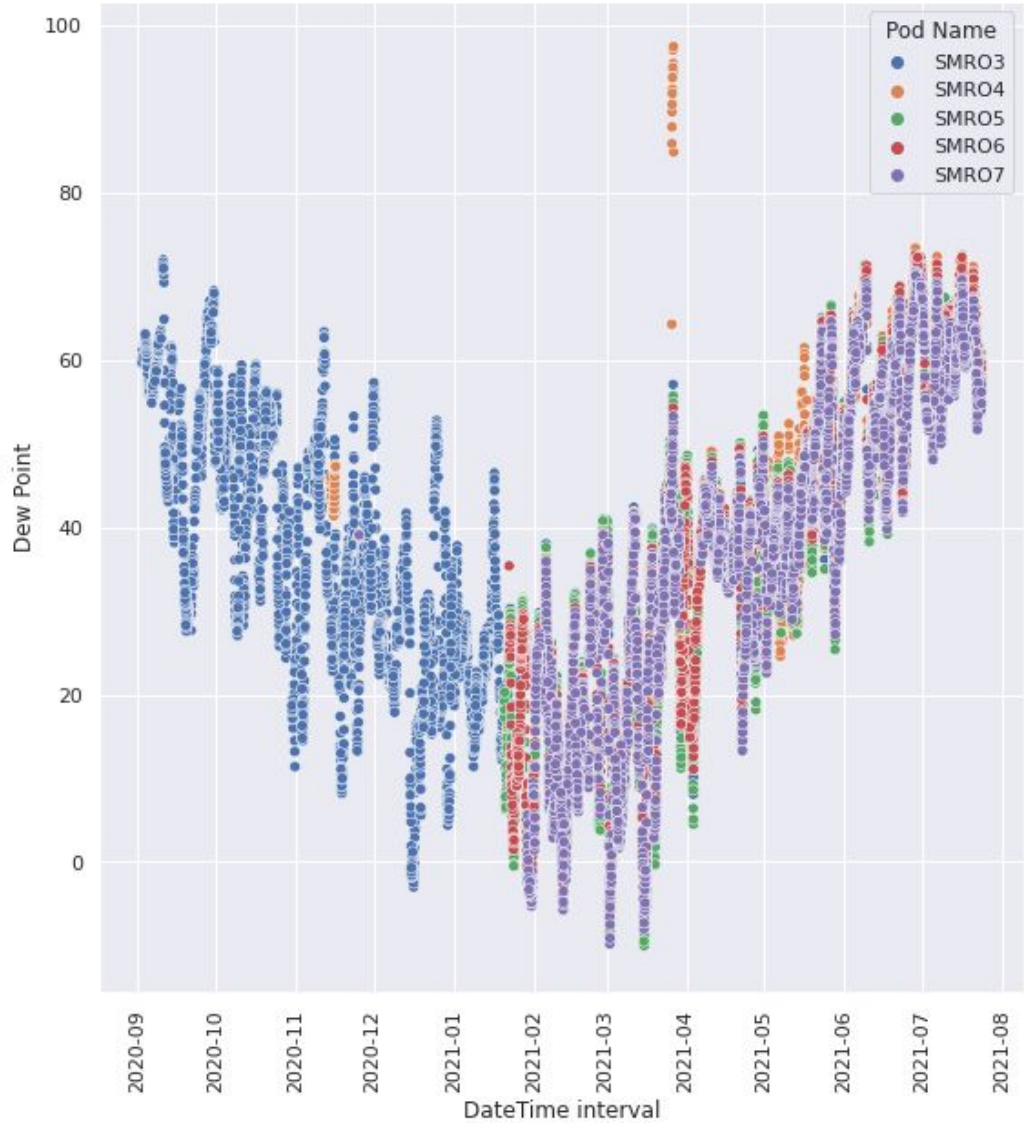
Wind Chill


Pod Name
SMRO3
SMRO4
SMRO5
SMRO6
SMRO7
Wind Chill
100
80
60
40
20
0
5000
10000
15000
20000


Pod Name
SMRO3
SMRO4
SMRO5
SMRO6
SMRO7
Wind Chill
100
80
60
40
20
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
DateTime interval

Dew Point

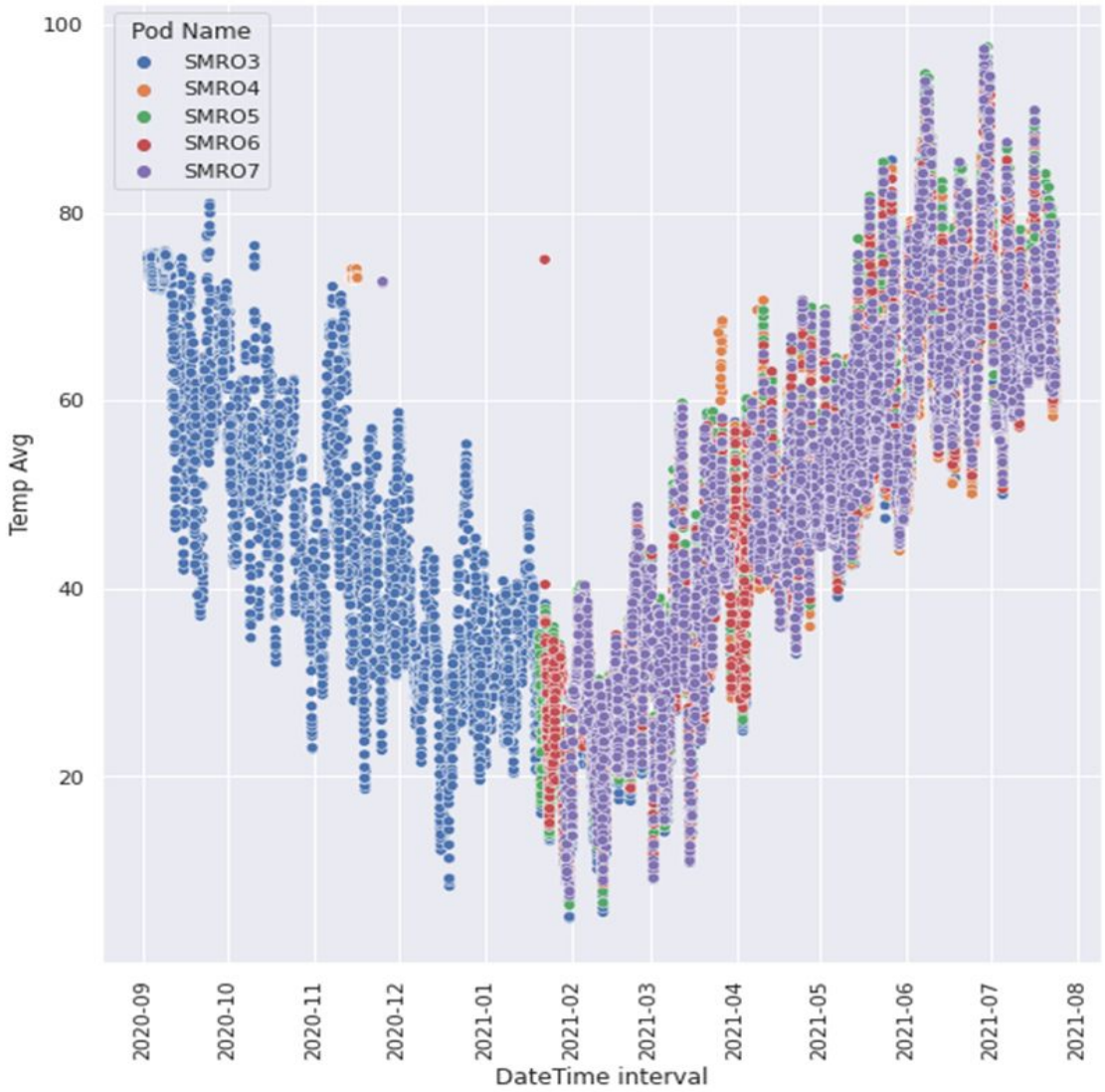

Pod Name
SMRO3
SMRO4
SMRO5
SMRO6
SMRO7
Dew Point
100
80
60
40
20
0
0
5000
10000
15000
20000

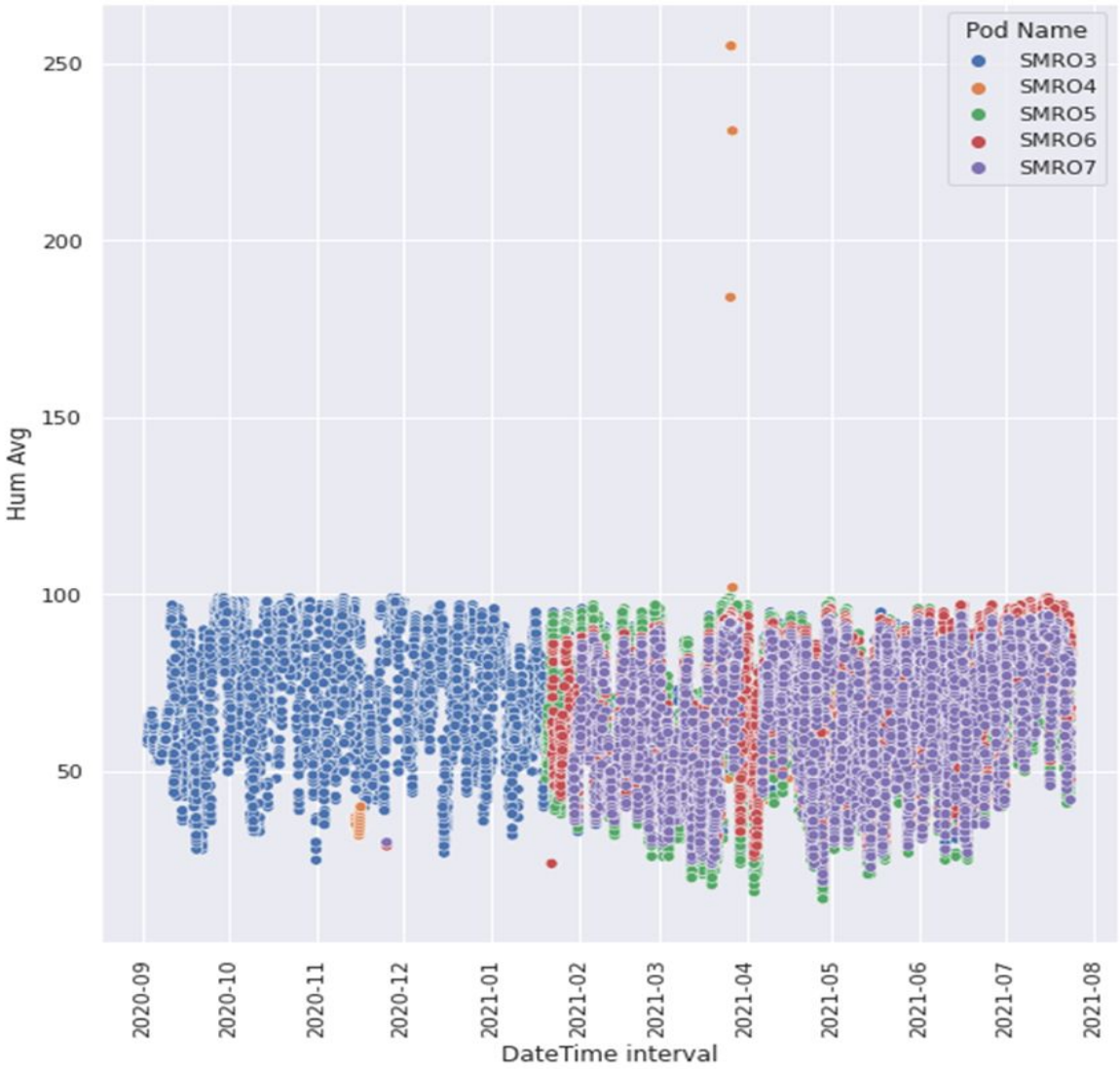

Pod Name
SMRO3
SMRO4
SMRO5
SMRO6
SMRO7
Dew Point
100
80
60
40
20
0
2020-09
2020-10
2020-11
2020-12
2021-01
2021-02
2021-03
2021-04
2021-05
2021-06
2021-07
2021-08
DateTime interval

EDA Results

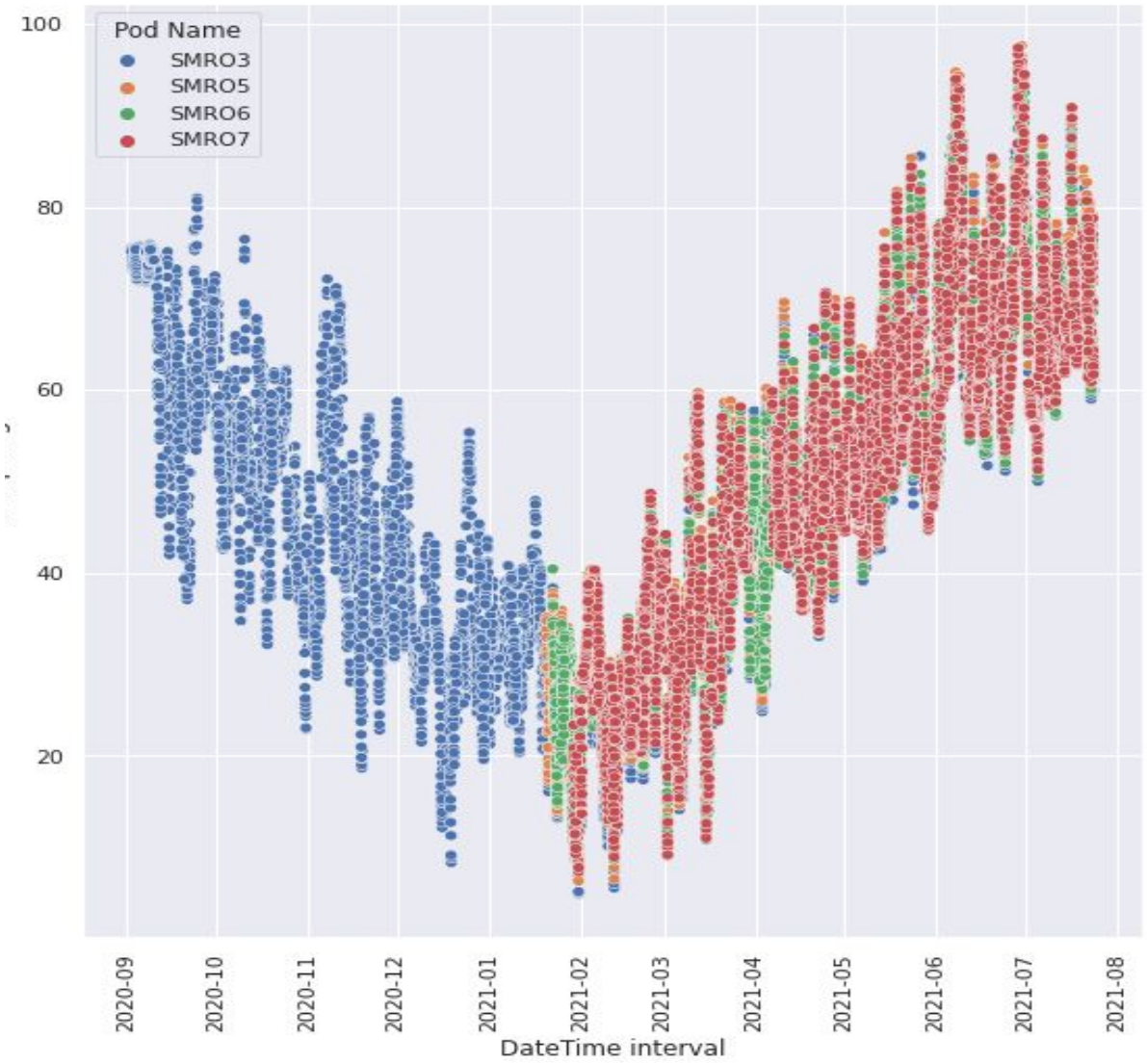
- There are 5 pods (devices that measure weather values)
- SMRO4 looks like it is producing a lot of abnormally high values/outliers, so we decided to exclude SMRO4
- The next slides will show how the removal of outliers affects the data

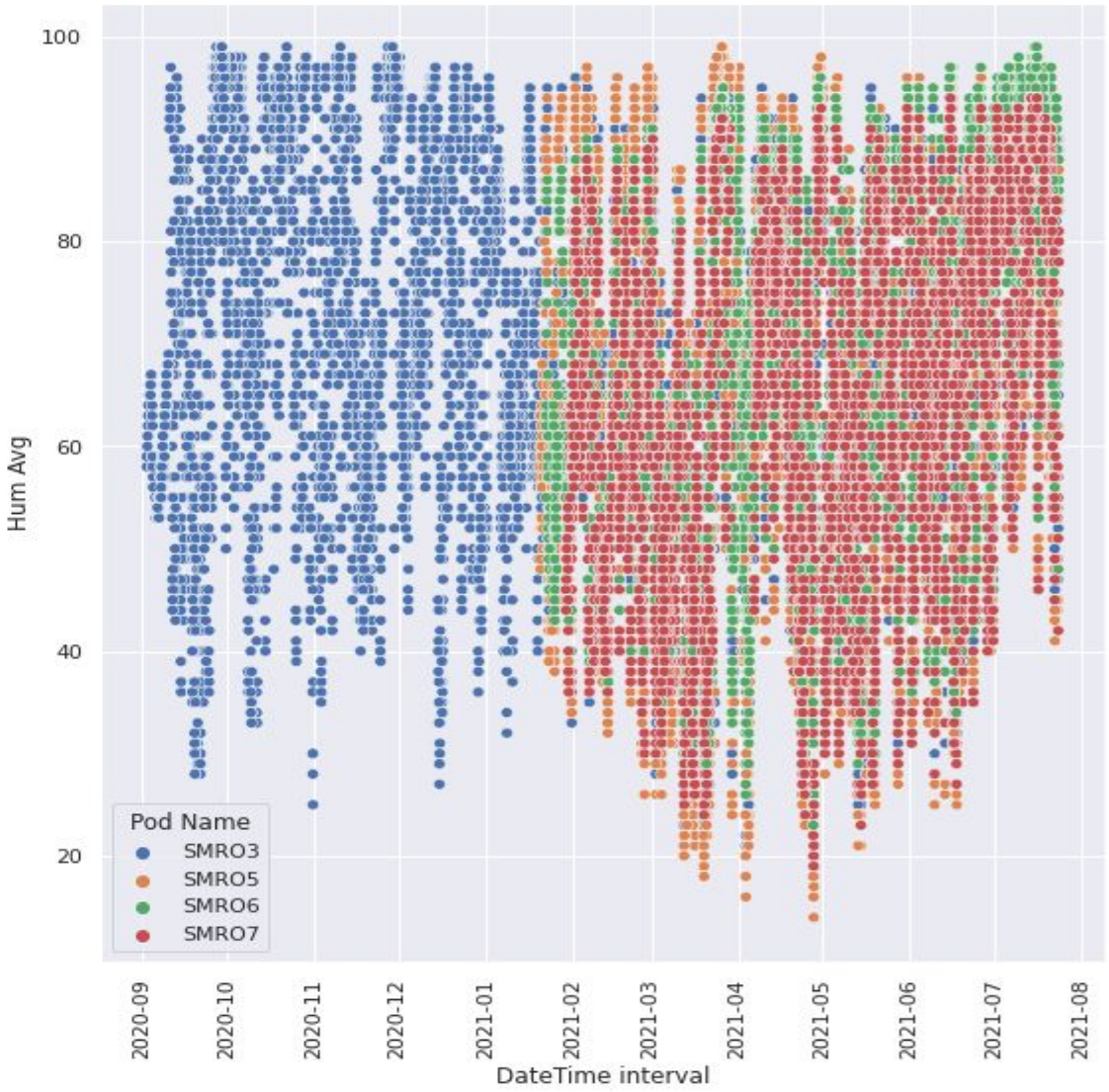
Before Outliers Removed



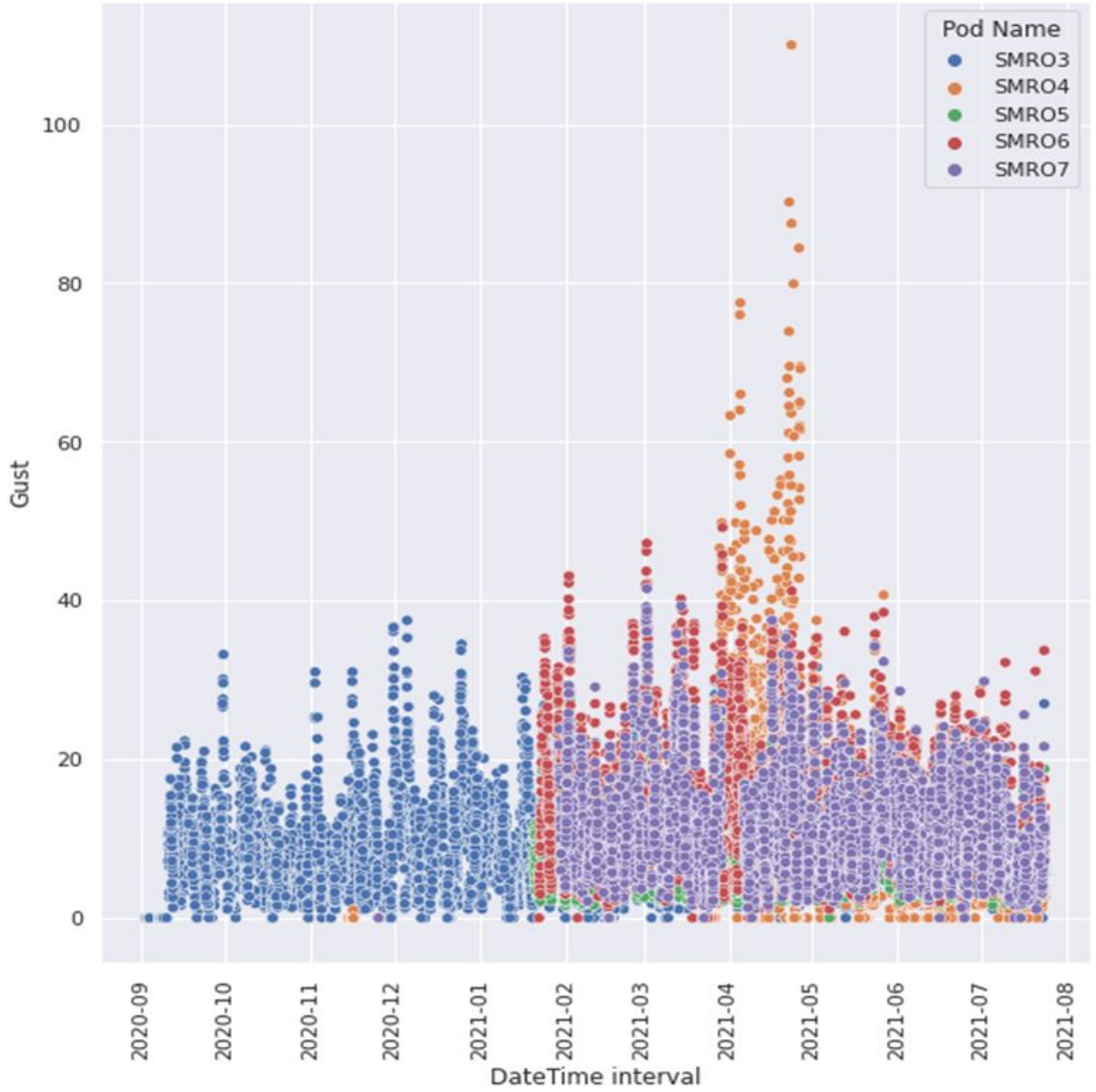


After Outliers Removed





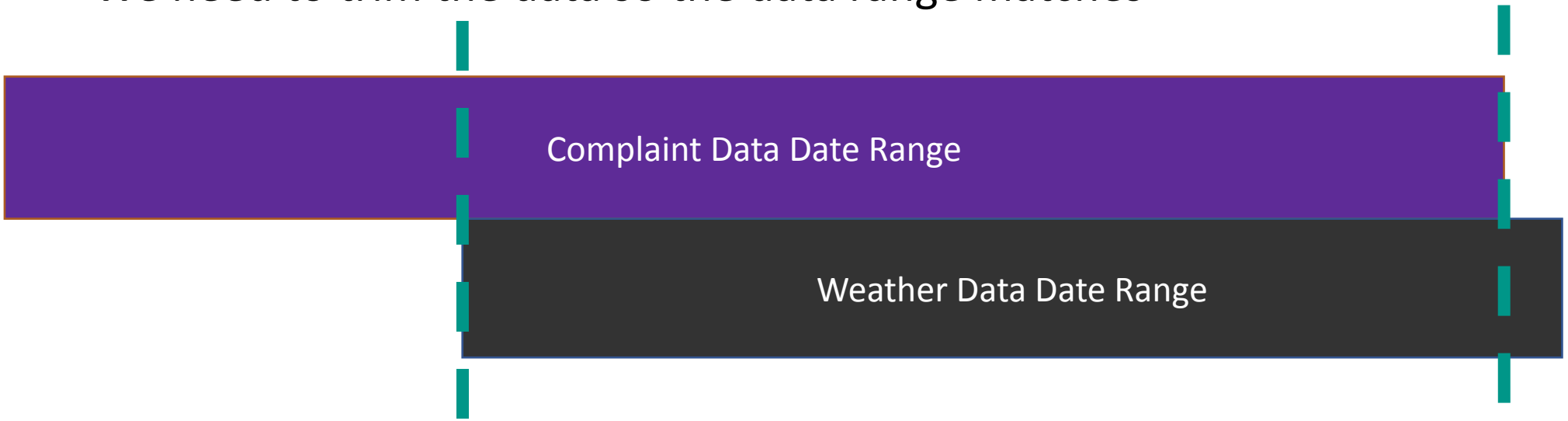
More comparisons post outlier removal



Merging Data

- We merge the complaints from the smell data, creating a df with weather events and complaints by the hour

- We need to trim the data so the data range matches



Pre-trimming results:
for weather data the minimum datetime is = 2020-09-02 12:00:00
for weather data the maximum datetime is = 2021-07-24 00:00:00
for smell data the minimum datetime is = 2019-05-22 09:00:00
for smell data the maximum datetime is = 2021-07-23 08:00:00

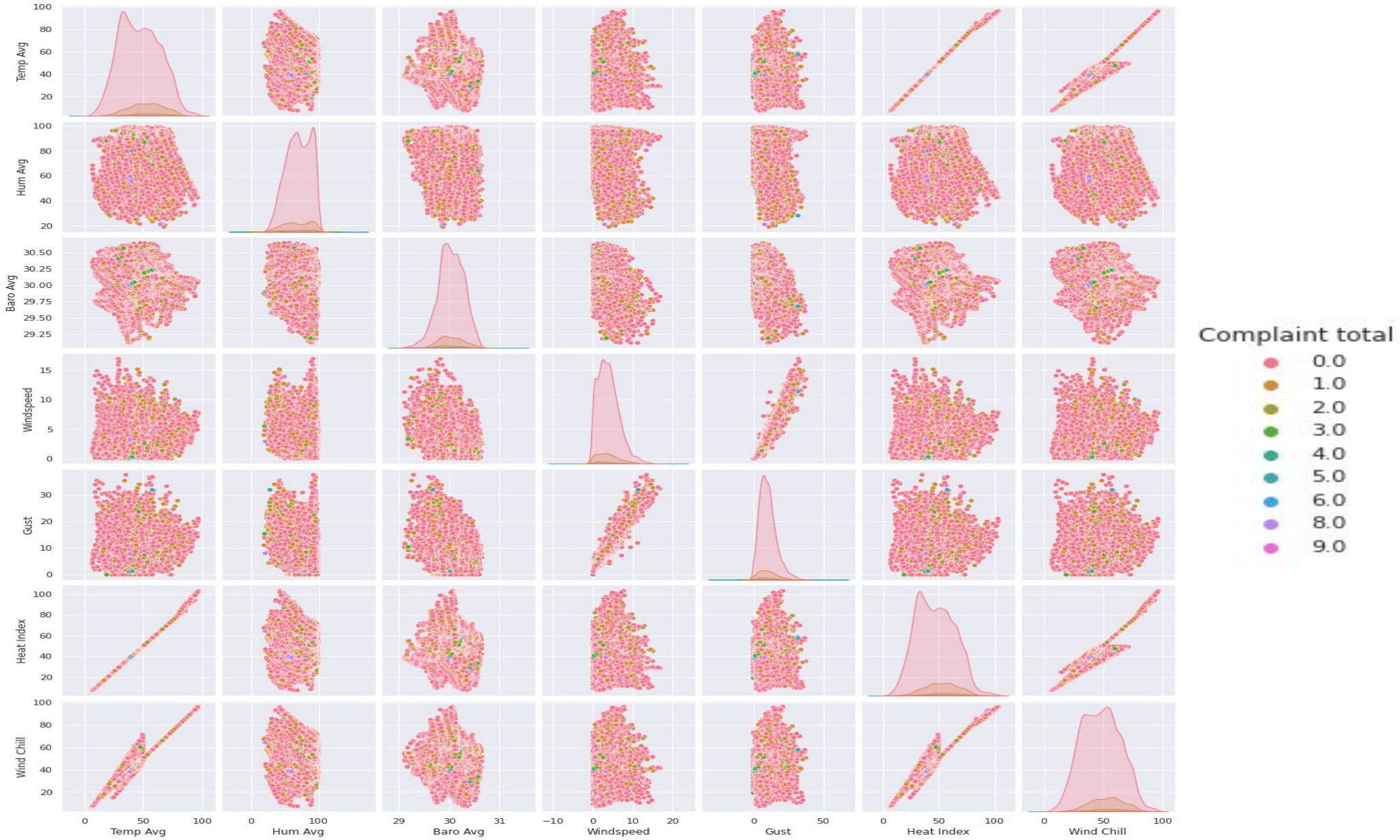
Post trimming results:
for weather data the minimum datetime is = 2020-09-02 19:00:00
for weather data the maximum datetime is = 2021-07-23 08:00:00
for smell data the minimum datetime is = 2020-09-02 19:00:00
for smell data the maximum datetime is = 2021-07-23 08:00:00

Merged Table

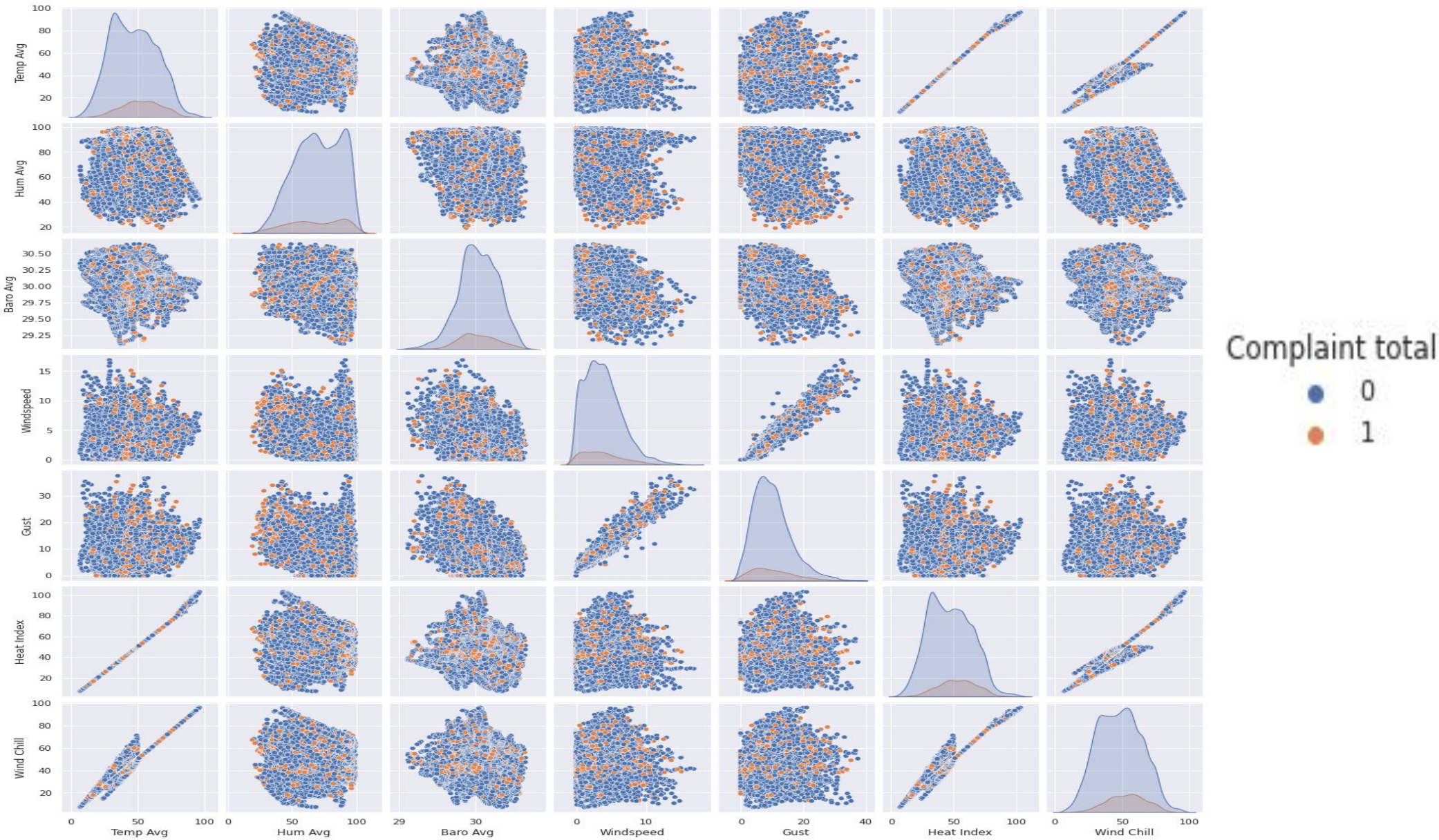
DateTime interval	Complaint total	Hour Group	Day_of_week	Month_of_year	Season	Temp Avg	Temp Low	Temp High	Hum Avg	Hum Low	Hum
2020-09-02 19:00:00	3.0	19:00:00	Wednesday	September	Fall	75.5900	75.200	75.700	61.00	61.00	
2020-09-02 20:00:00	1.0	20:00:00	Wednesday	September	Fall	75.1300	75.000	75.400	61.00	61.00	
2020-09-02 21:00:00	0.0	21:00:00	Wednesday	September	Fall	74.7200	74.500	75.000	61.00	61.00	
2020-09-02 22:00:00	1.0	22:00:00	Wednesday	September	Fall	74.3300	74.200	74.500	62.00	62.00	
2020-09-02 23:00:00	0.0	23:00:00	Wednesday	September	Fall	74.1100	74.000	74.200	62.00	62.00	
...	...	...	...	...	...	...	...	...	...	...	
2021-07-23 04:00:00	0.0	04:00:00	Friday	July	Summer	61.1050	60.600	61.500	86.50	86.00	
2021-07-23 05:00:00	0.0	05:00:00	Friday	July	Summer	60.4850	60.175	60.725	88.25	87.00	
2021-07-23 06:00:00	0.0	06:00:00	Friday	July	Summer	60.5300	60.175	60.900	88.00	87.25	
2021-07-23 07:00:00	0.0	07:00:00	Friday	July	Summer	61.1050	60.350	62.125	86.50	84.25	
2021-07-23 08:00:00	2.0	08:00:00	Friday	July	Summer	65.8975	62.125	69.900	75.00	65.25	

ws × 28 columns

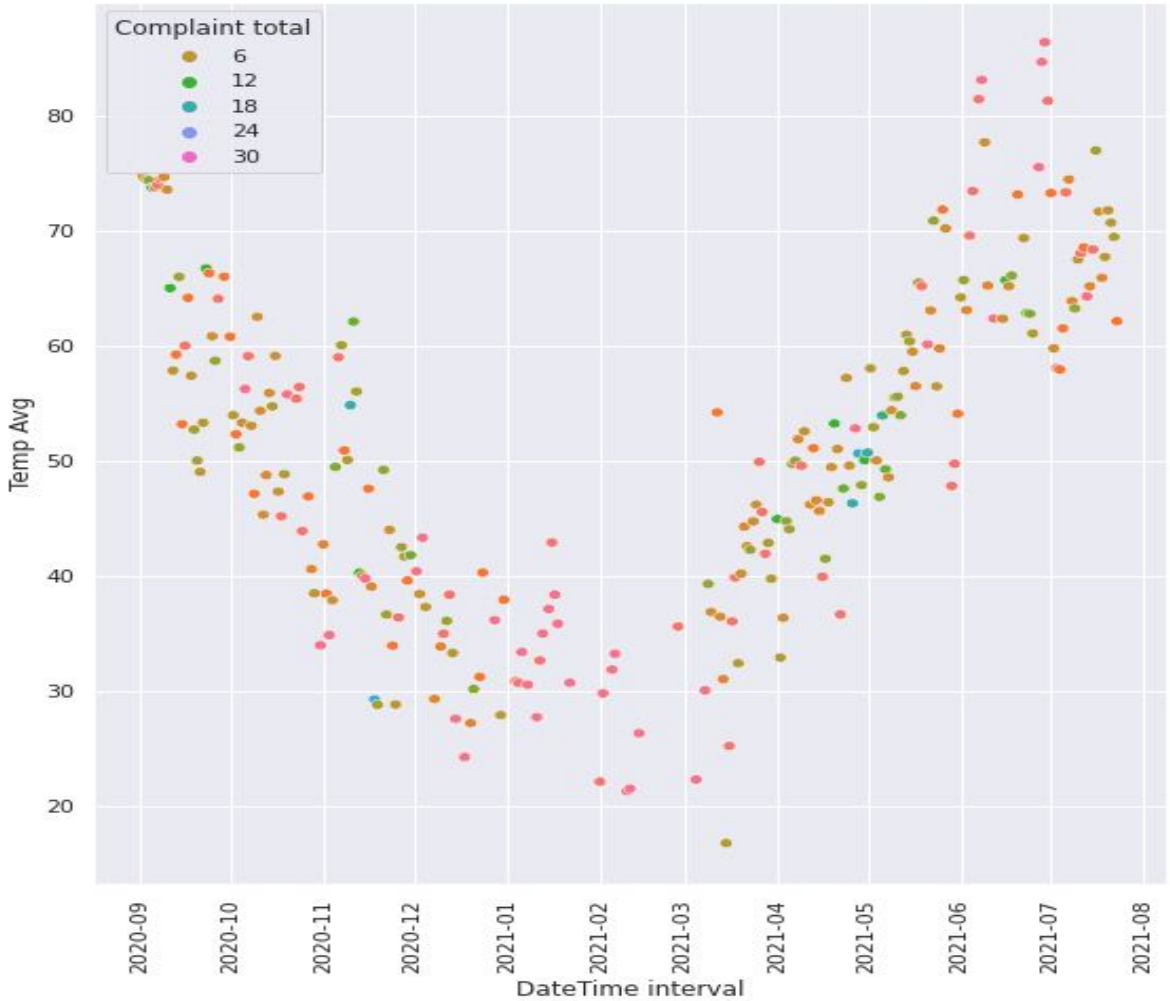
EDA Complaint Groups

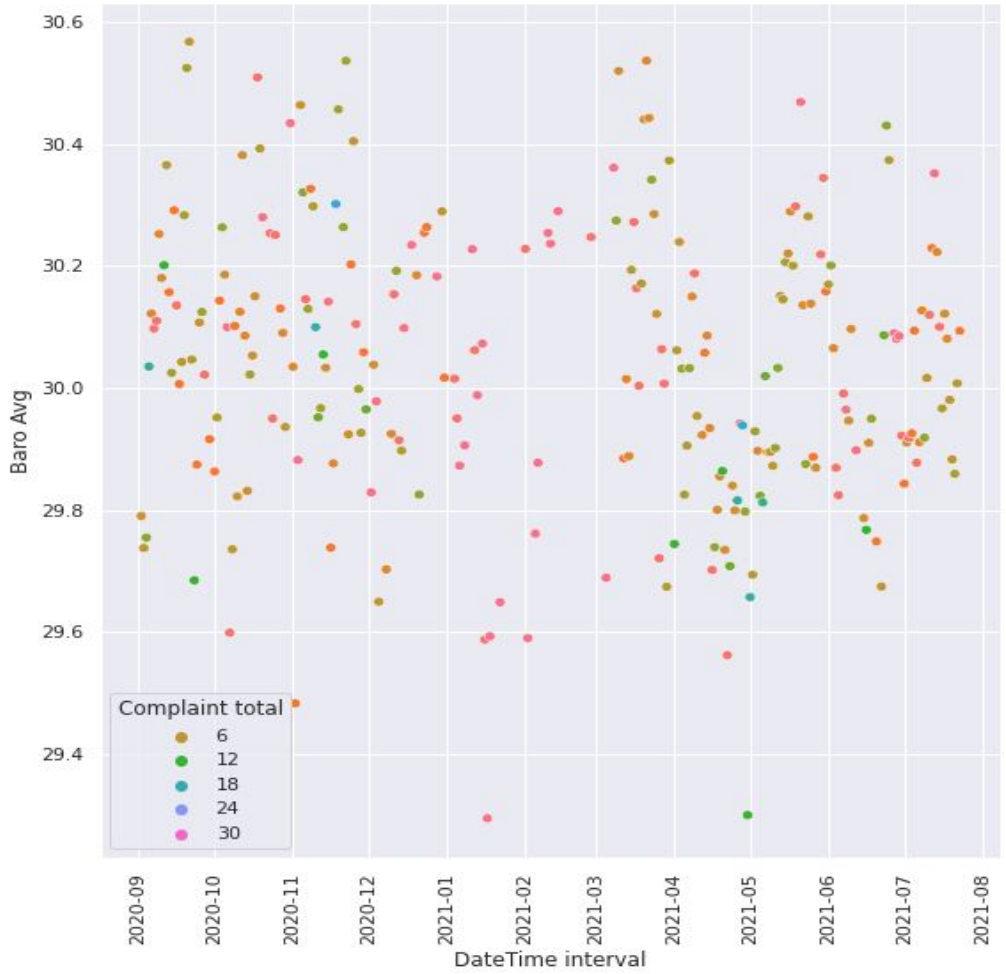


EDA Complaint Groups Continued

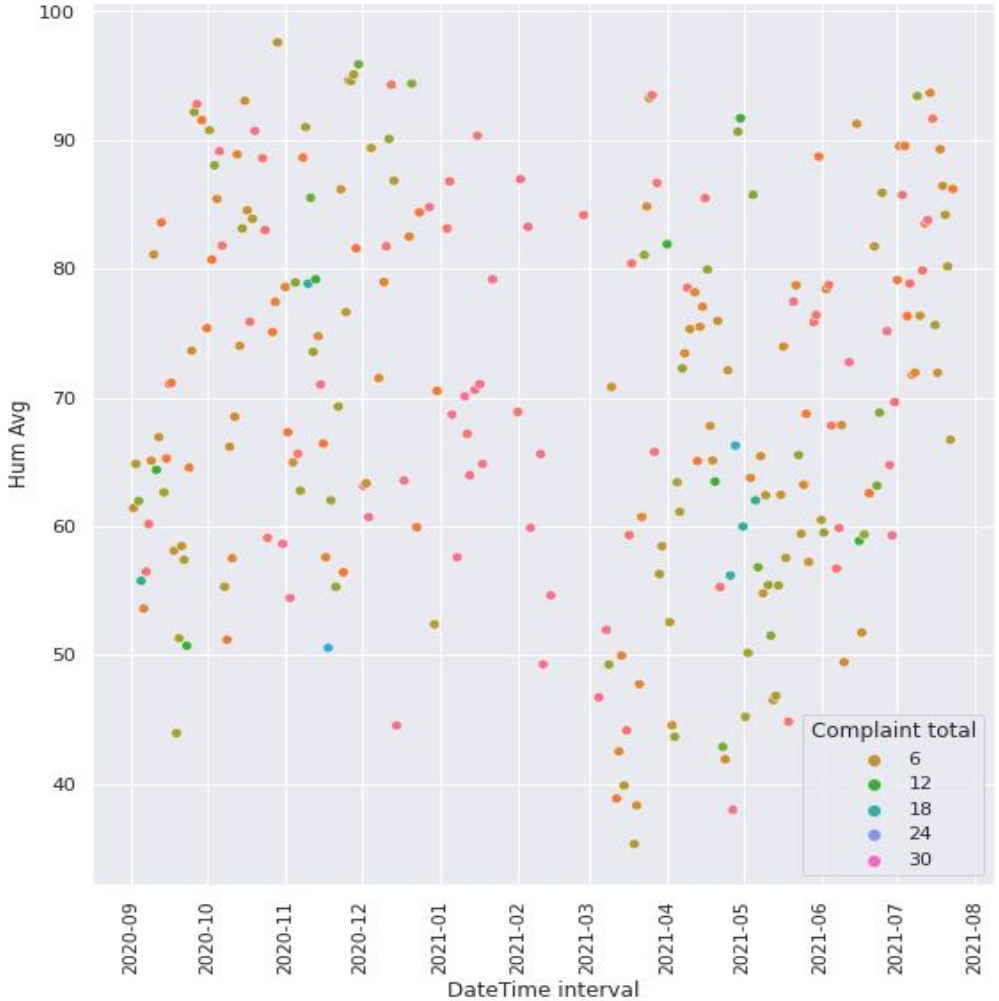


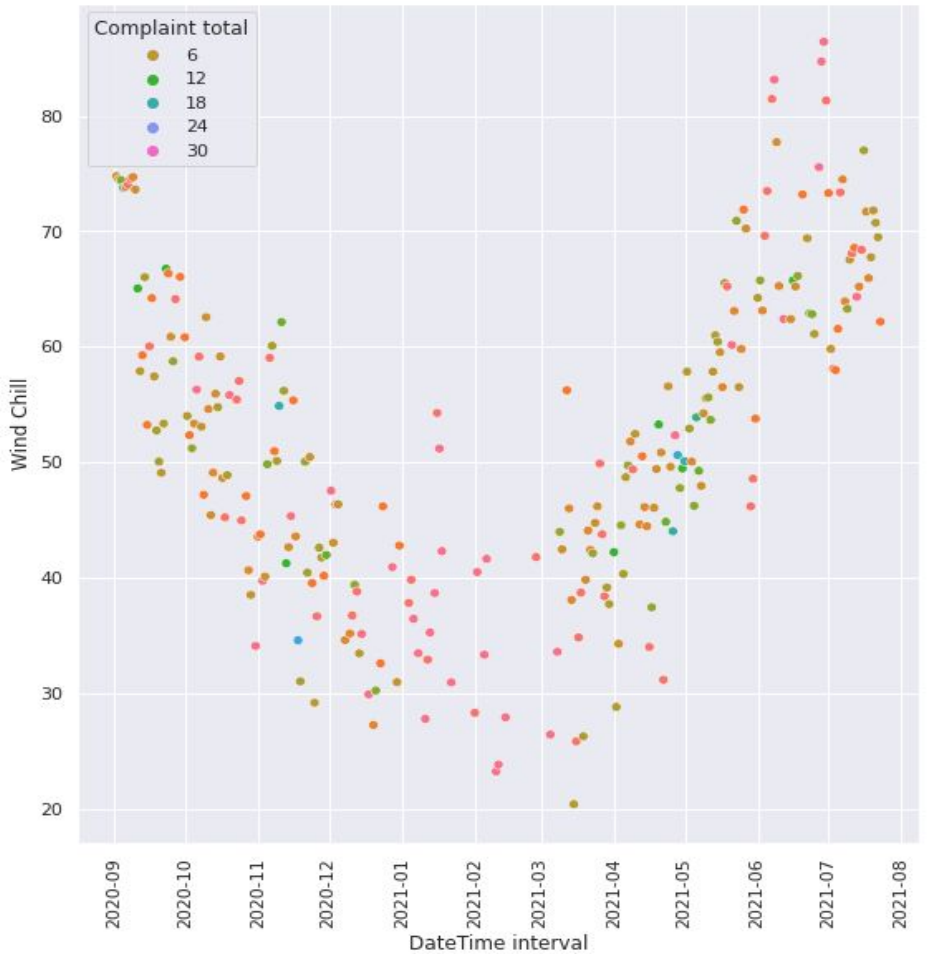
Weather events over time with complaints



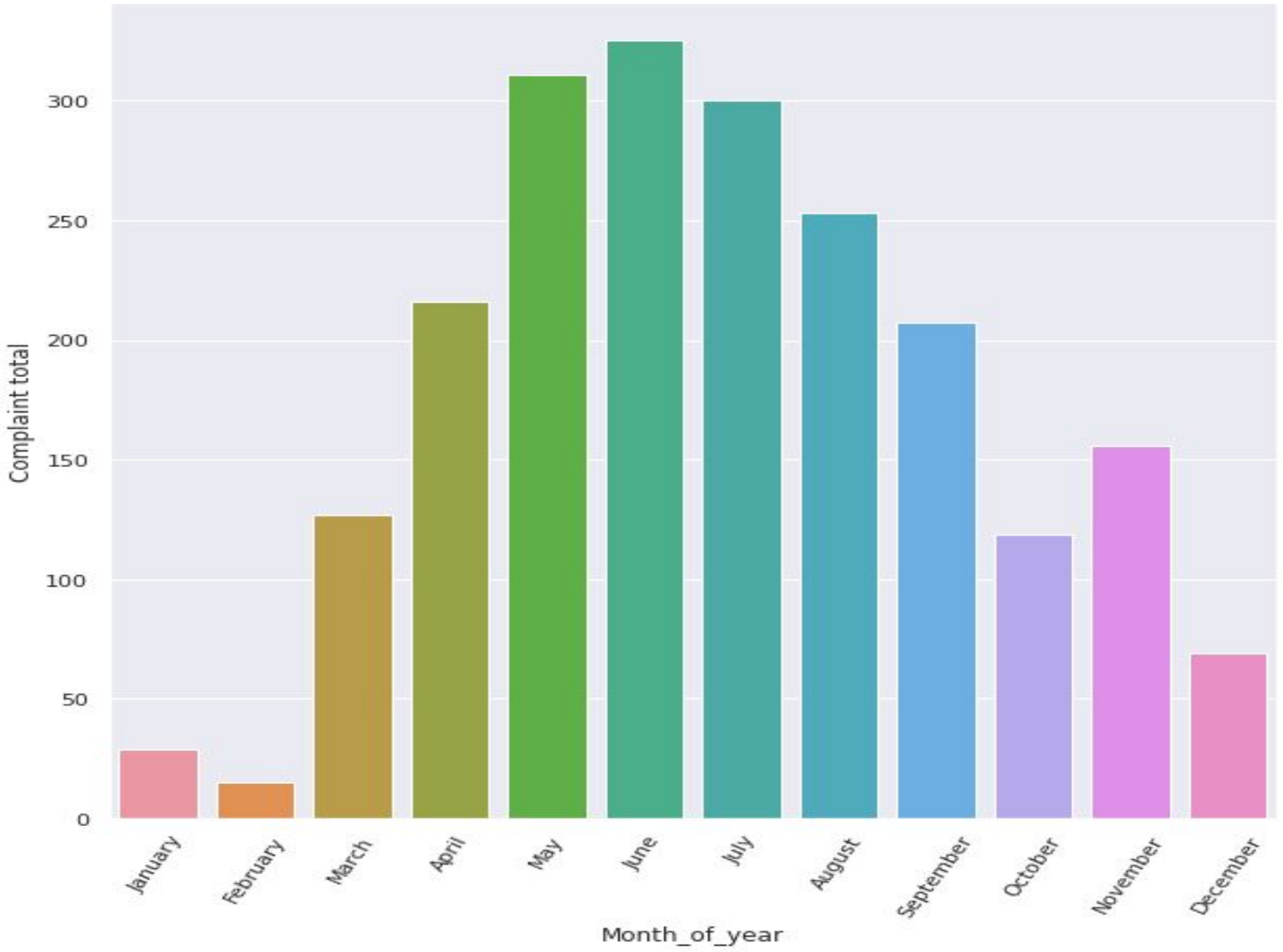


Weather events over time with complaints



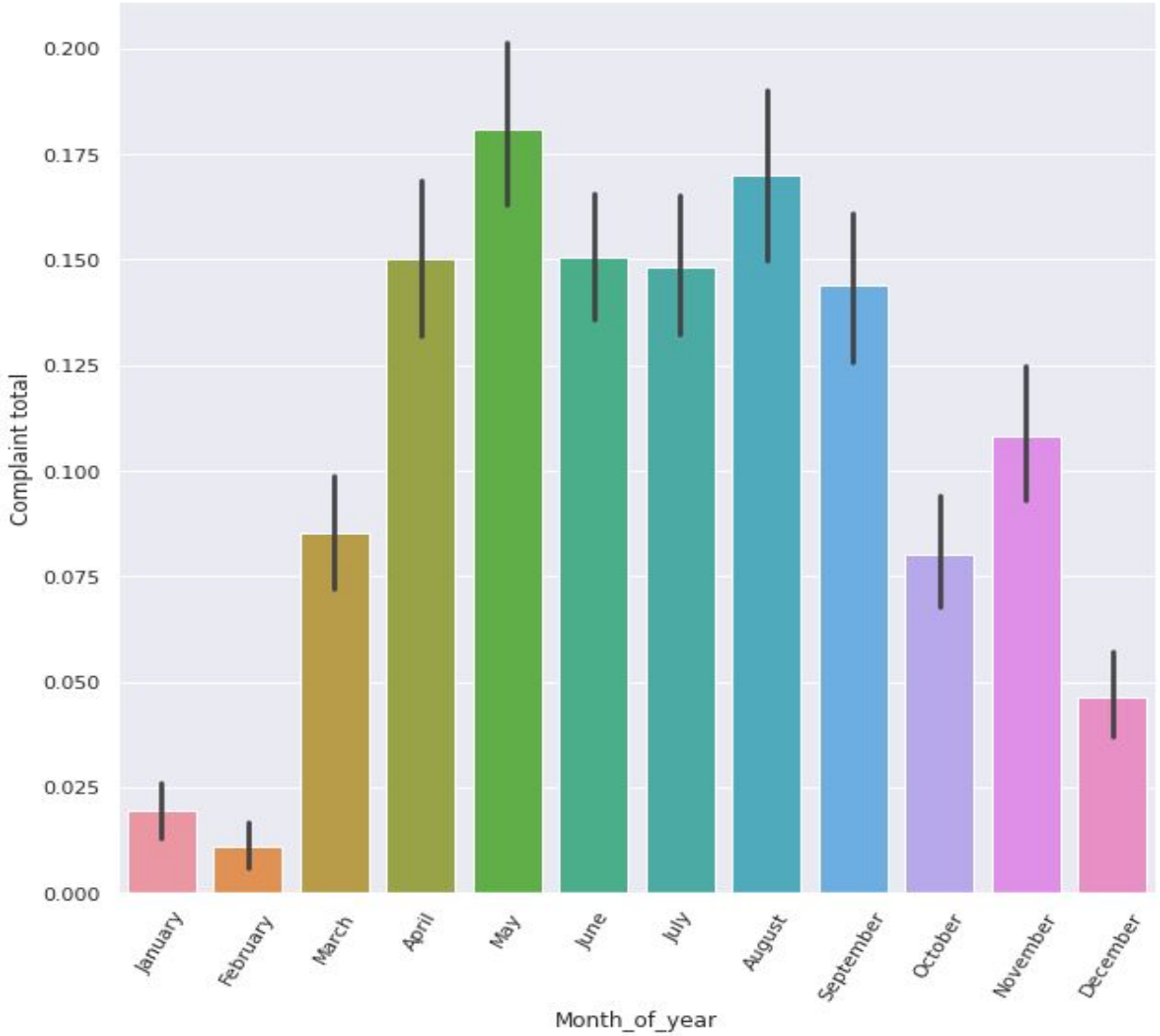


Complaints by Month



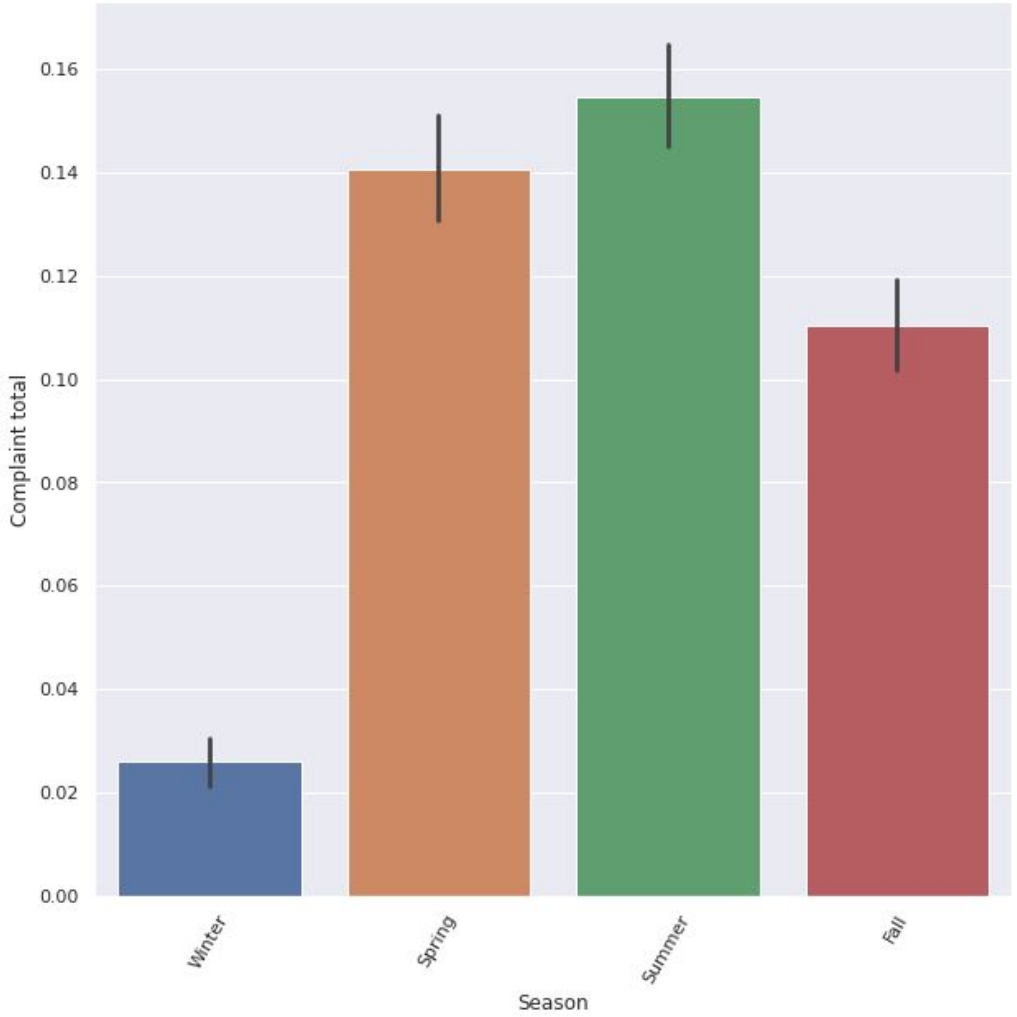
- This is a barplot of the absolute the complaints grouped by month
- May, June, and July have the most complaints
- December, January, and February have the least complaints

Complaints by Month Continued

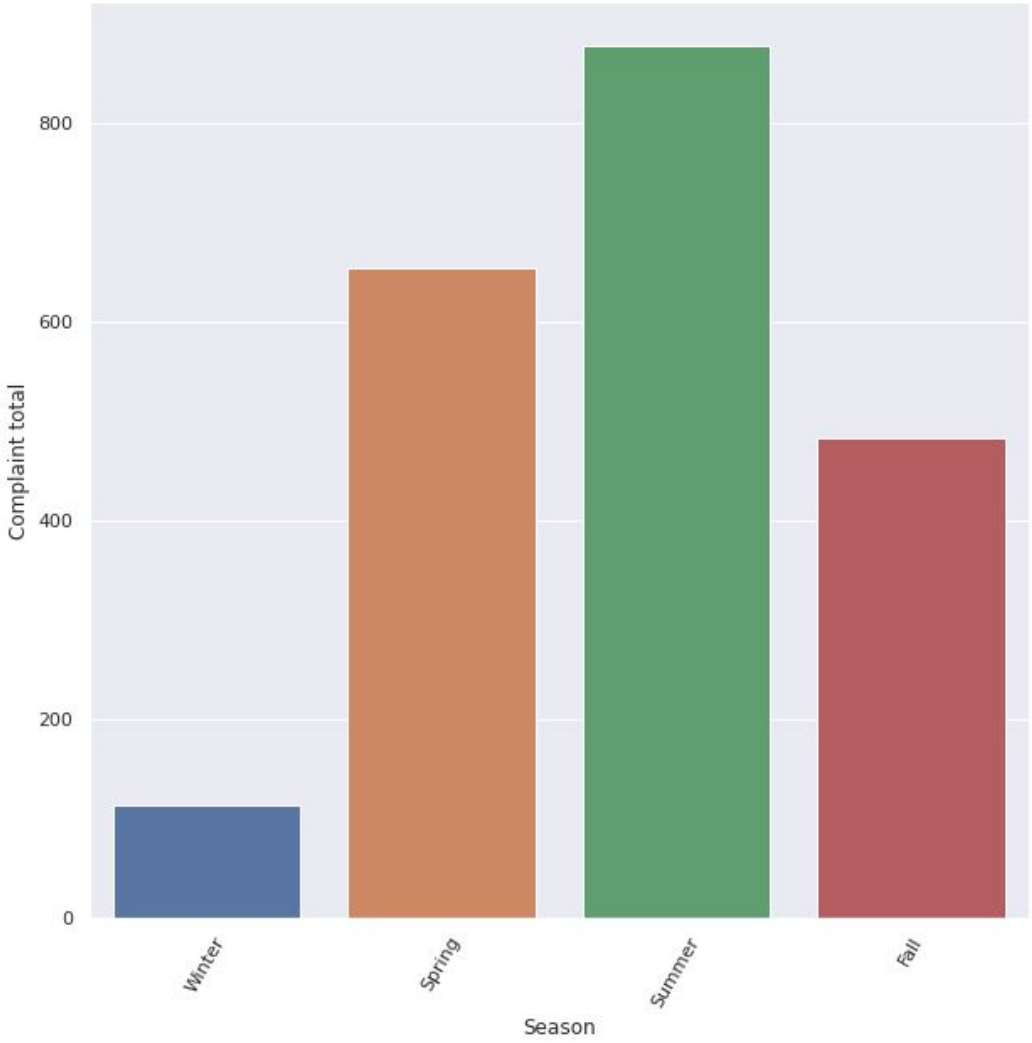


- This is a barplot of all average complaints grouped by month on average
- May, June, and July and August have the most complaints
- December, January, and February continue to have the least have the least complaints

Complaints by Season

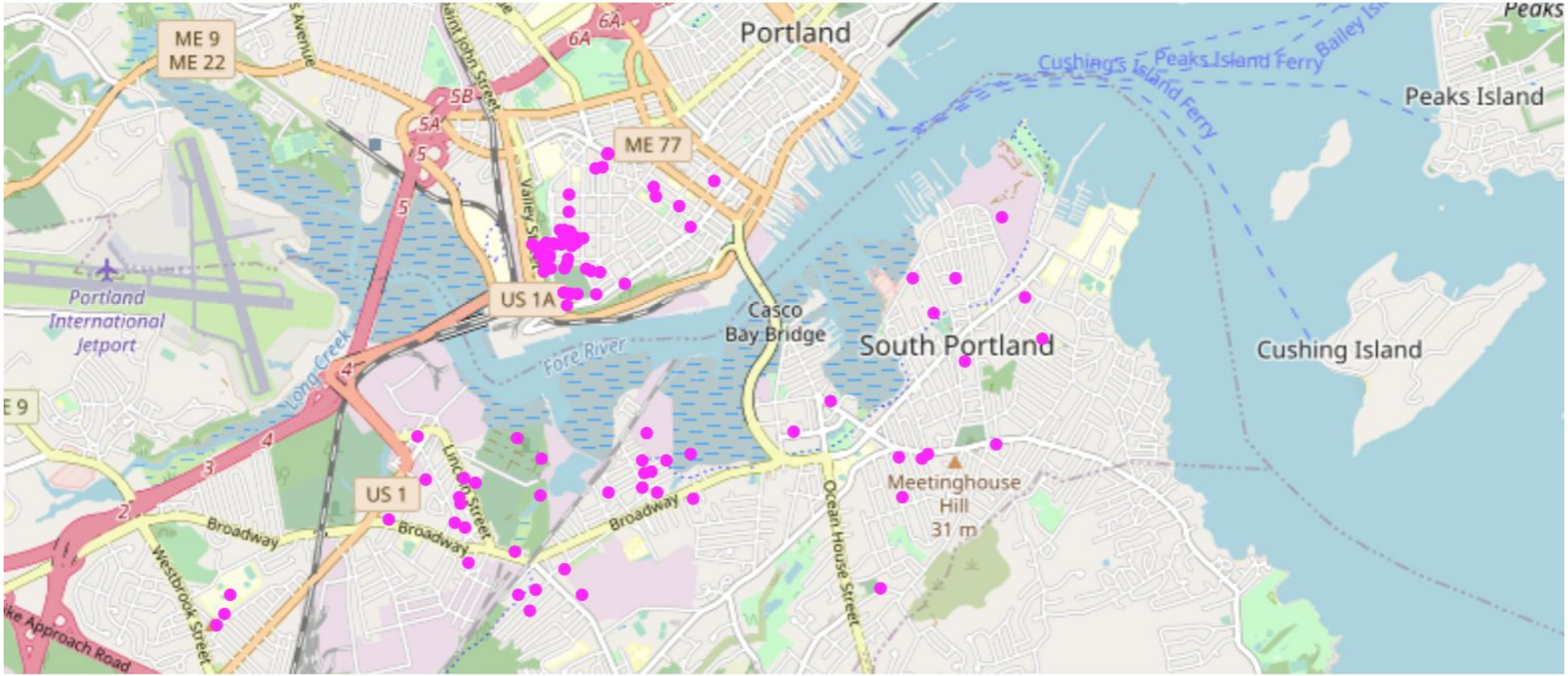


This is the average complaints per season

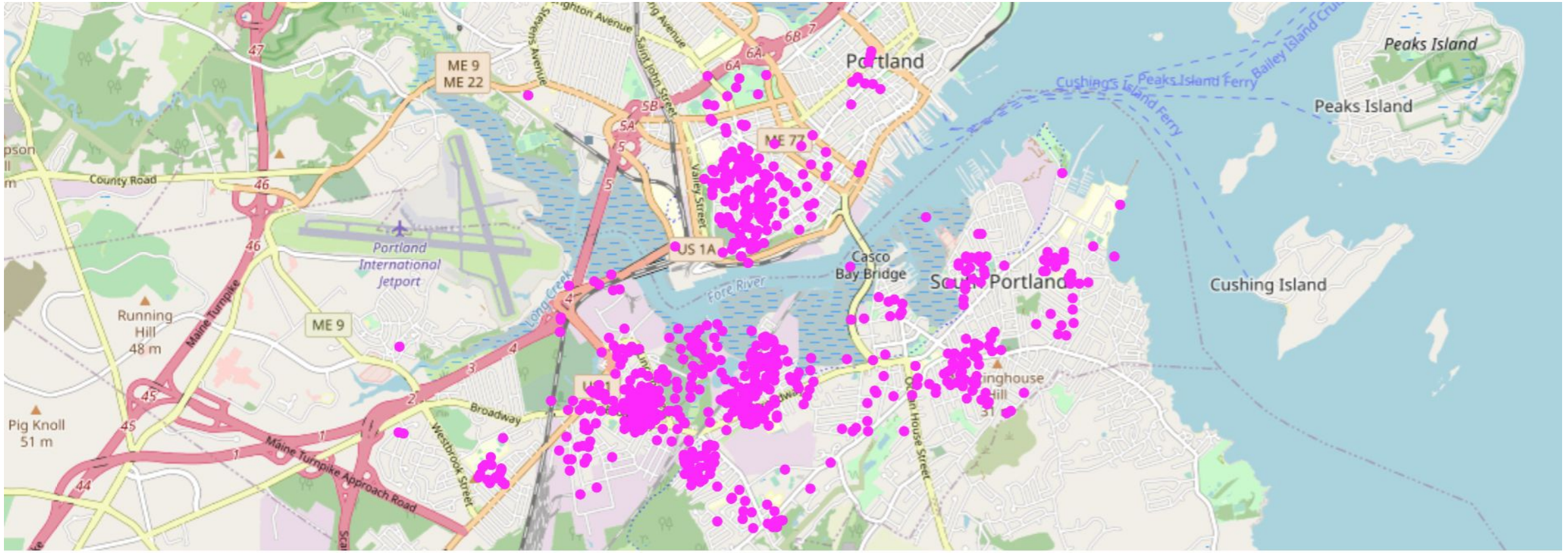


This is the absolute complaints per season

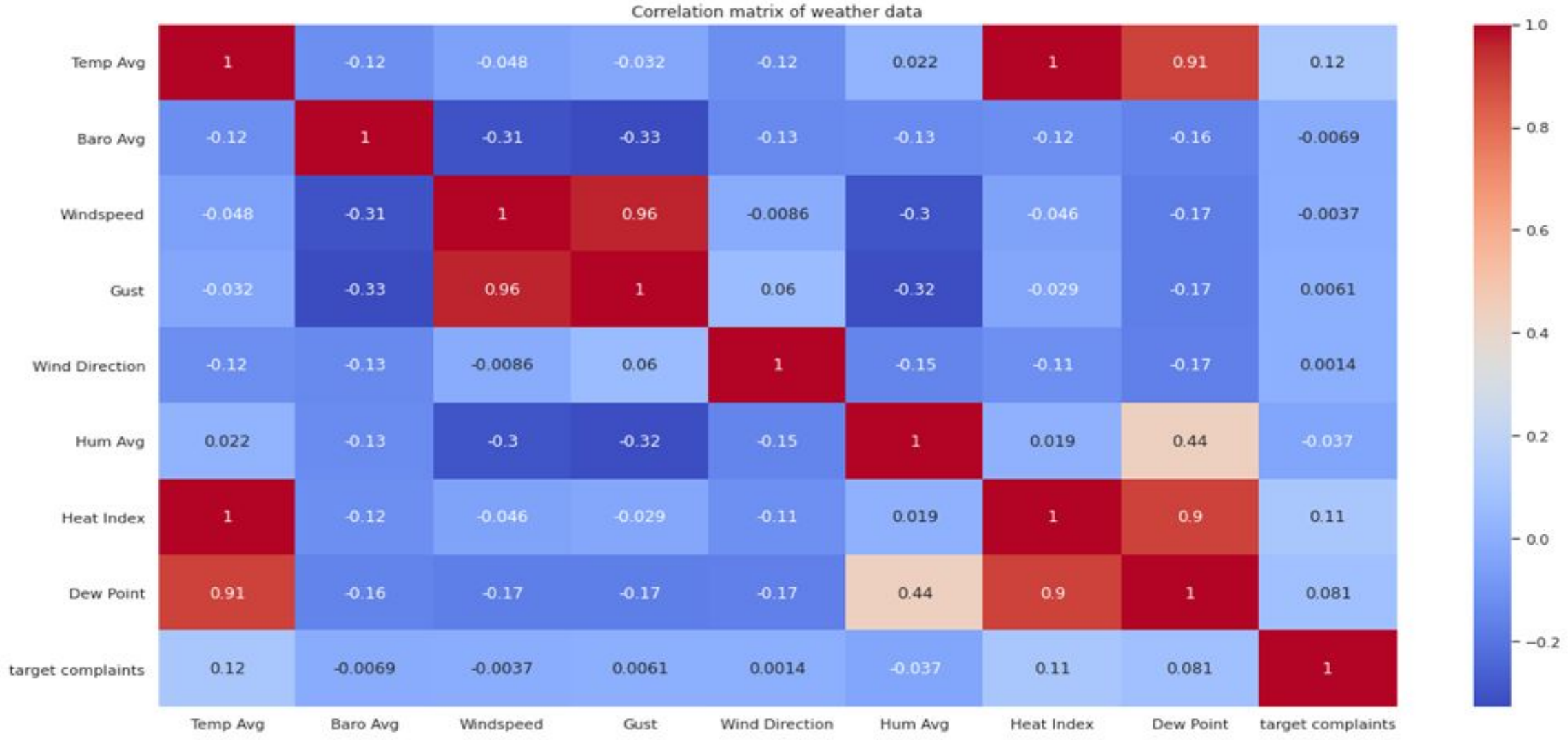
Winter Map


Portland
ME 9
ME 22
ME 77
US 1A
US 1
Portland International Jetport
Long Creek
Fore River
Casco Bay Bridge
South Portland
Cushing Island
Peaks Island
Cushings Island Ferry
Peaks Island Ferry
Meetinghouse Hill 31 m
Broadway
Westbrook Street
Lincoln Street
Ocean House Street
Valley Street
Saint John Street

Summer Map


Portland
Peaks Island
Peaks Island
Cushing Island
Casco Bay Bridge
Portland International Jetport
Running Hill 48 m
Pig Knoll 51 m
ME 9
ME 22
ME 77
US 1A
Maine Turnpike
Maine Turnpike Approach Road
Broadway
Valley Street
Saint John Street
Long Creek
Fore River
Westbrook Street
Ocean House Street
County Road
Cushing's Island Ferry
Peaks Island Ferry
Bailey Island Cruise

Correlation Matrix:-



- The correlation matrix shows the relation between the dependent (target complaints) and the independent variables.
- There is no strong correlation with the target complaint and any of the weather features on an individual level.

OLS Model:-

OLS Regression Results

Dep. Variable:	target complaints	R-squared:	0.029
Model:	OLS	Adj. R-squared:	0.028
Method:	Least Squares	F-statistic:	56.82
Date:	Tue, 17 Aug 2021	Prob (F-statistic):	2.52e-47
Time:	00:34:21	Log-Likelihood:	-2304.7
No. Observations:	7739	AIC:	4619.
Df Residuals:	7734	BIC:	4654.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.9178	0.107	-8.576	0.000	-1.128	-0.708
Temp Avg	0.0615	0.008	7.691	0.000	0.046	0.077
Hum Avg	0.0103	0.001	8.456	0.000	0.008	0.013
Heat Index	-0.0331	0.007	-4.415	0.000	-0.048	-0.018
Dew Point	-0.0275	0.003	-9.259	0.000	-0.033	-0.022

Omnibus:	3074.589	Durbin-Watson:	1.642
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9099.271
Skew:	2.180	Prob(JB):	0.00
Kurtosis:	6.033	Cond. No.	3.08e+03

- The backward p-value selection is done using the ordinary least square technique.
- These are the features which maintained significant p-value during backward feature selection.
- These features collectively affect the target variable (target complaints).
- Features affecting the target variables are Temp avg, Hum avg, Heat Index and Dew point.

Logistic Model:

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

- To verify the OLS model the logistic model is used and the output verifies that there is no change in the feature selection.
- The features remain the same as it showed in the OLS model.

Accuracy Score:-

```
[106] grid.best_params_
```

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

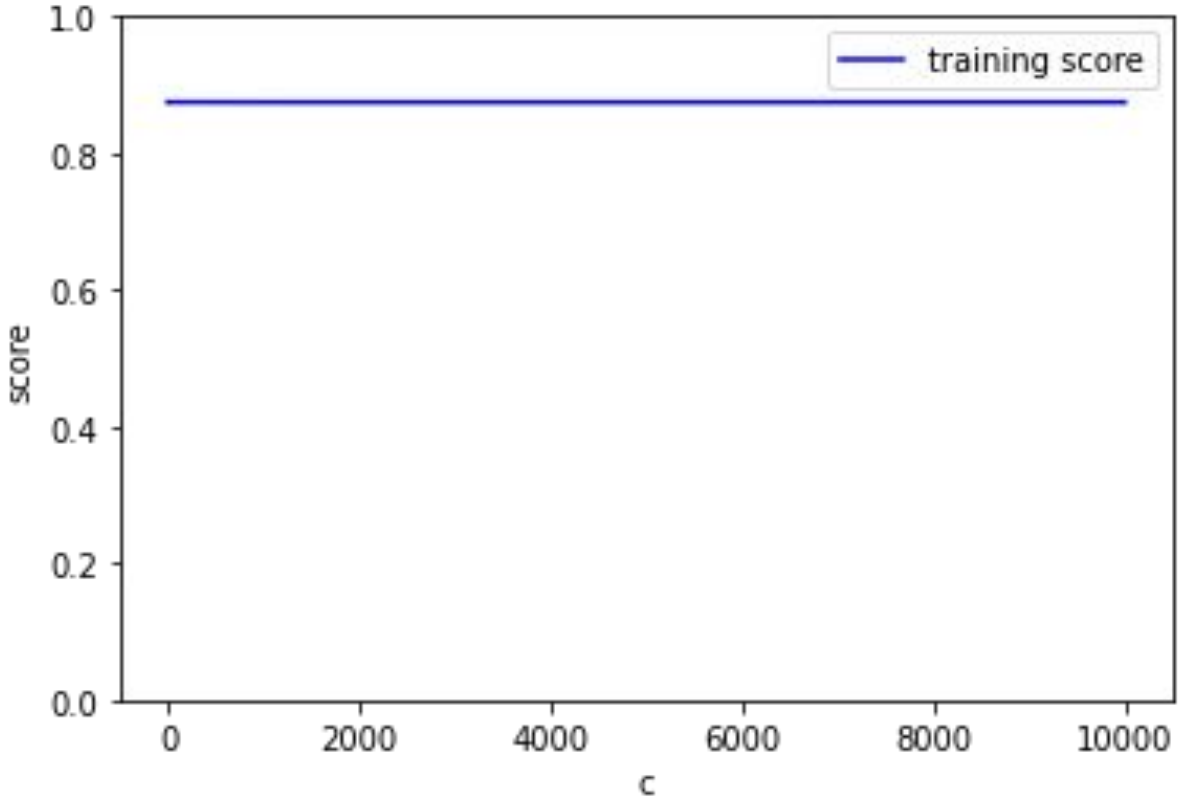
```
[107] lr = LogisticRegression(C=1) #c is regularization parameter
      lr.fit(X, y)
      y_pred = lr.predict(X)
```

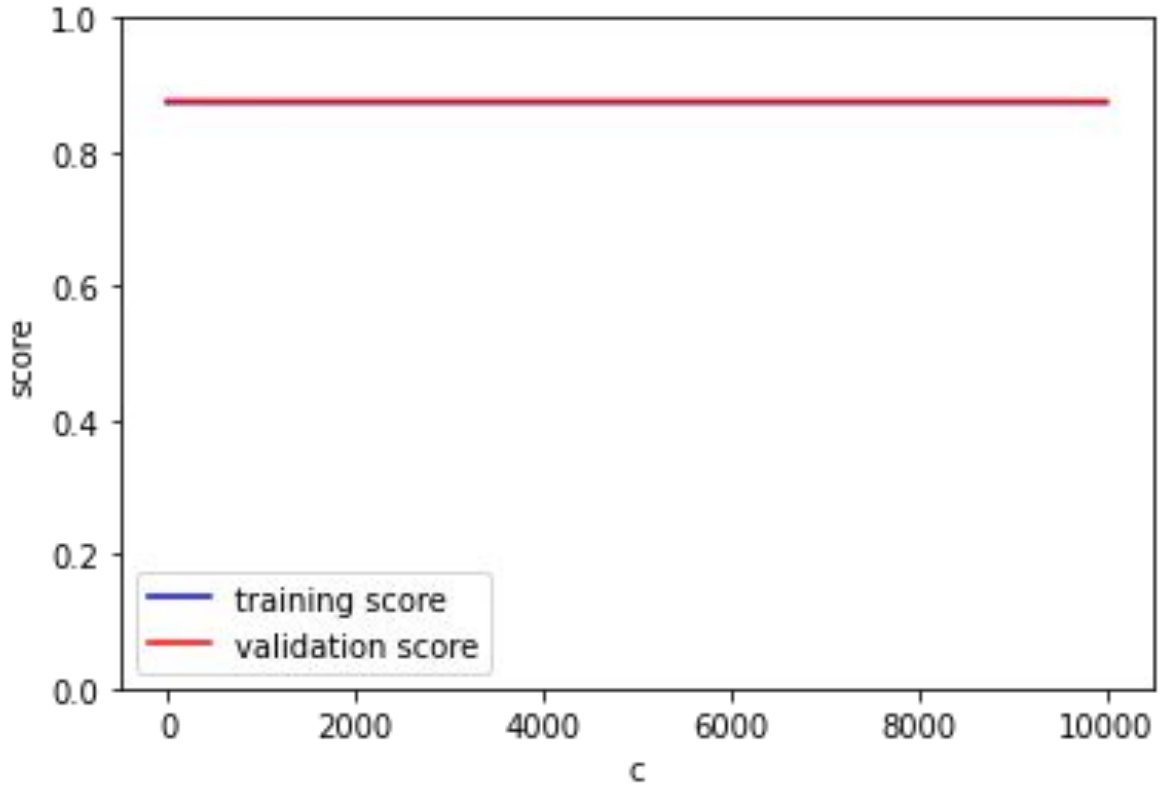
```
[108] from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score


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

```
Accuracy: 0.875
```

- The accuracy score to predict whether we get a complaint or not is 87.5% .
- The optimum regularization parameter for the model is c=1(10**0).
- The prediction depends on the independent variables which affect the possibility of getting complaints or not.

Validation Curve:



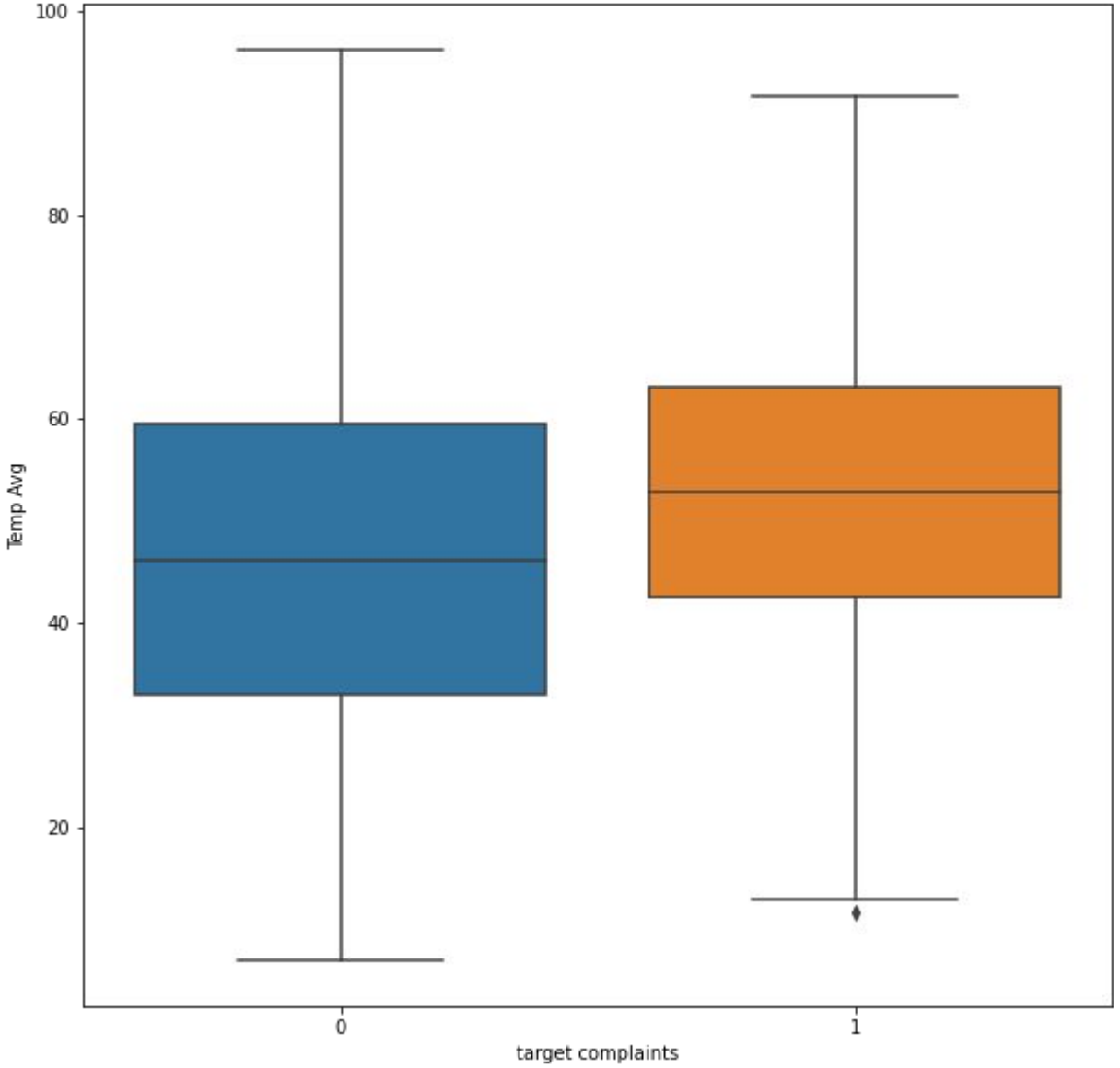


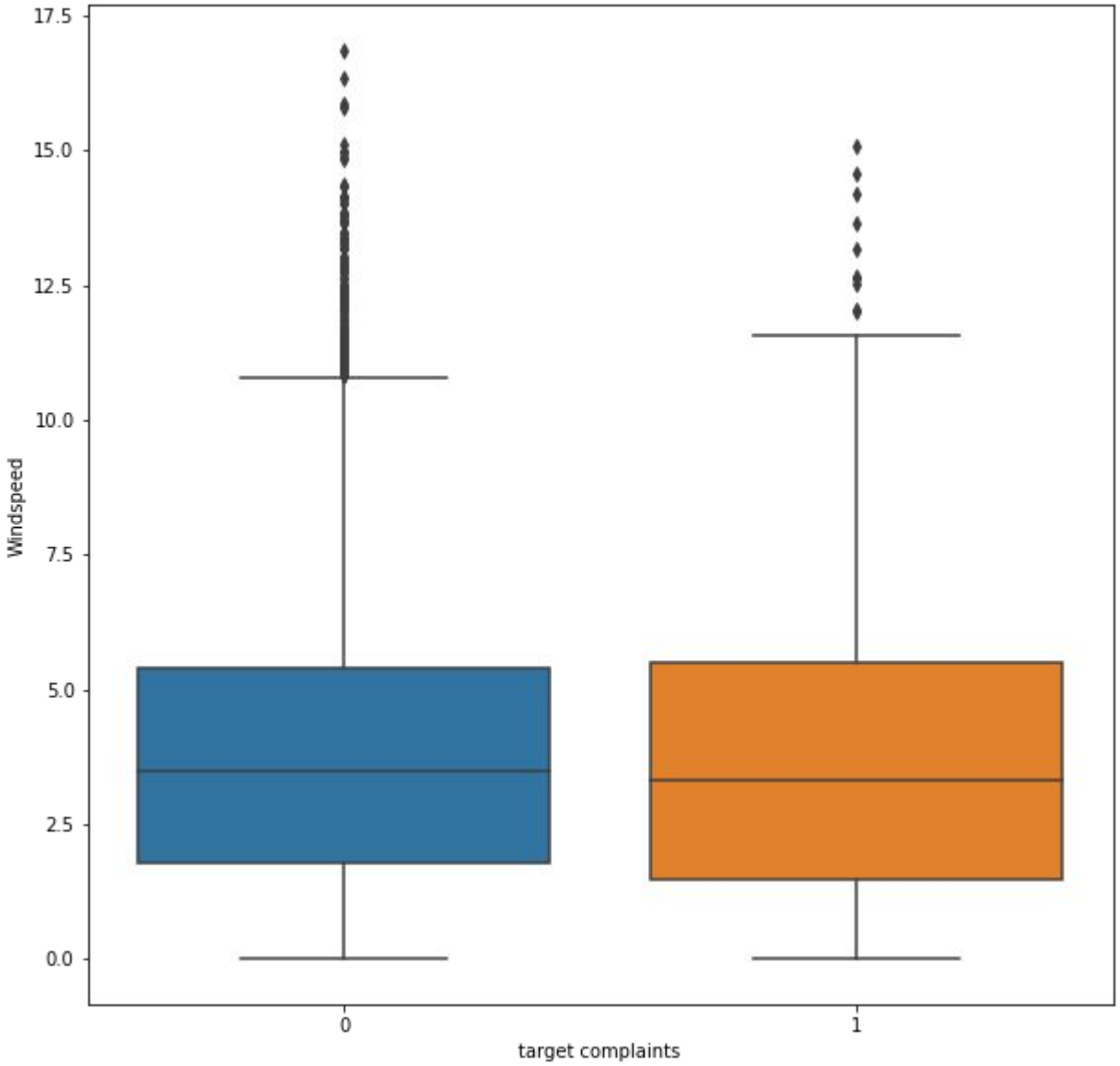
- The Geo dataset keeps increasing its training accuracy if you let the model be as complex as it likes.

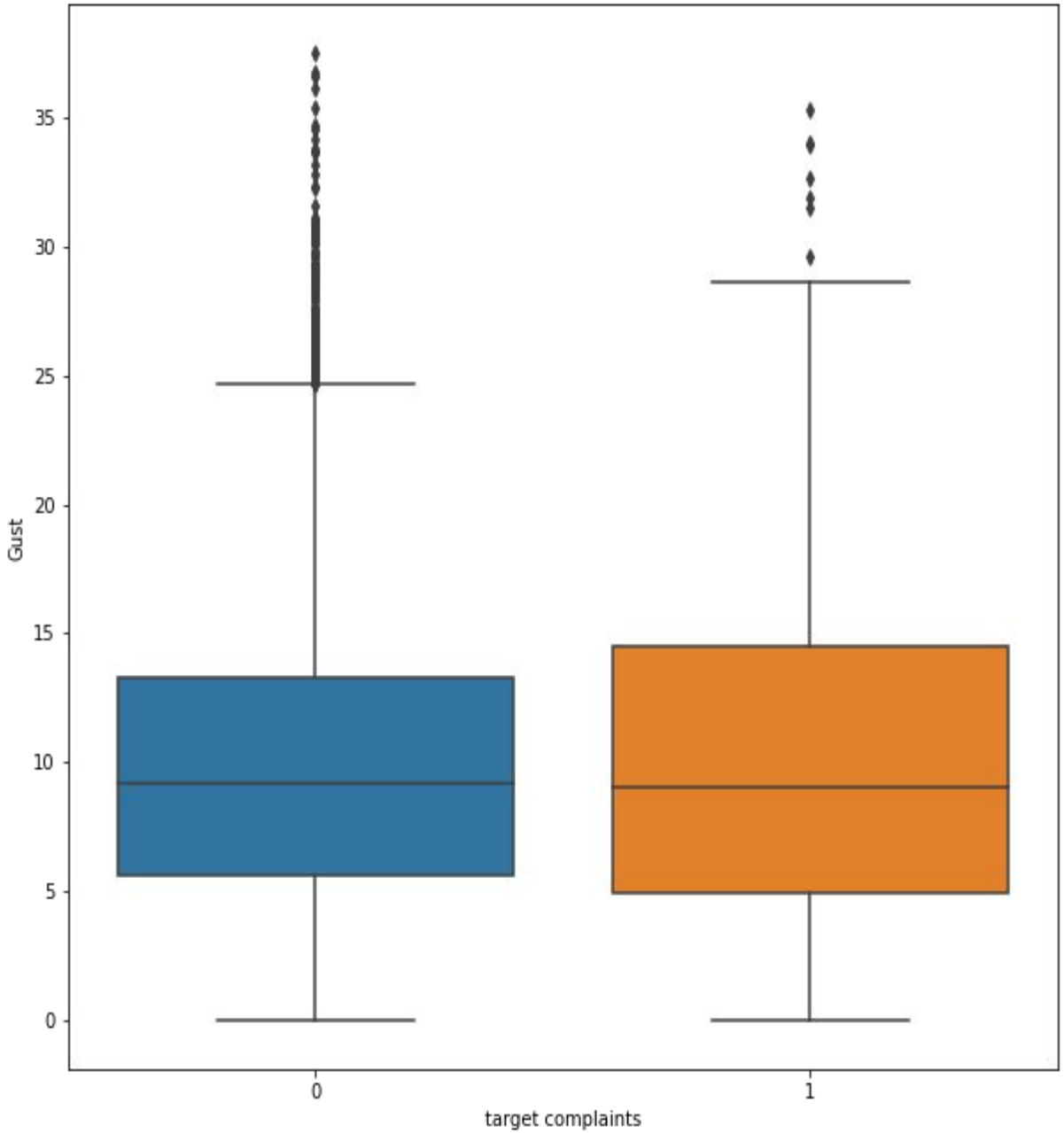
So, which C do we use?

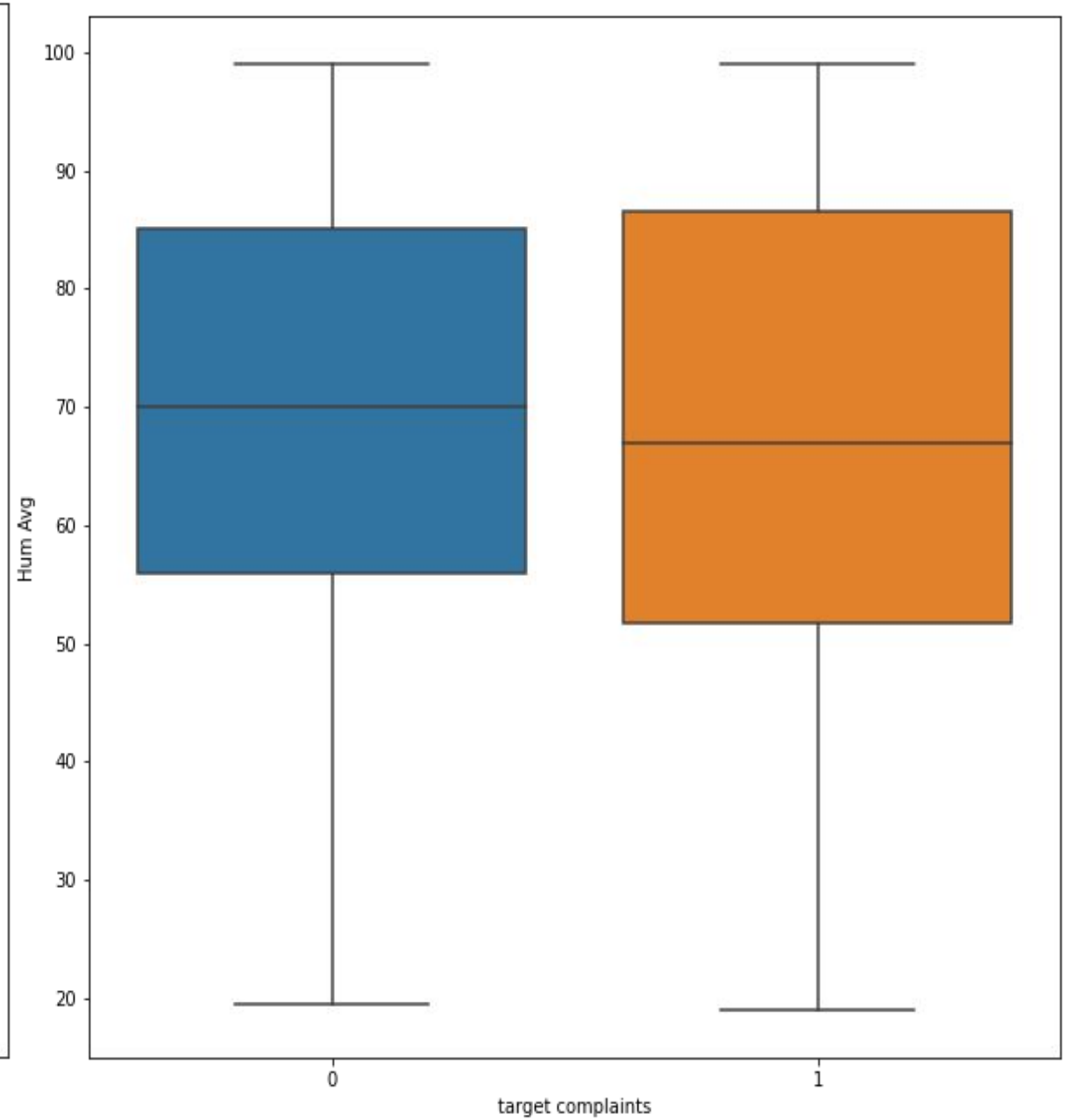
You might have guessed it, but the idea is to choose that C which offers the smallest difference between the training and testing accuracy (remember we want to generalise our model to unseen data)

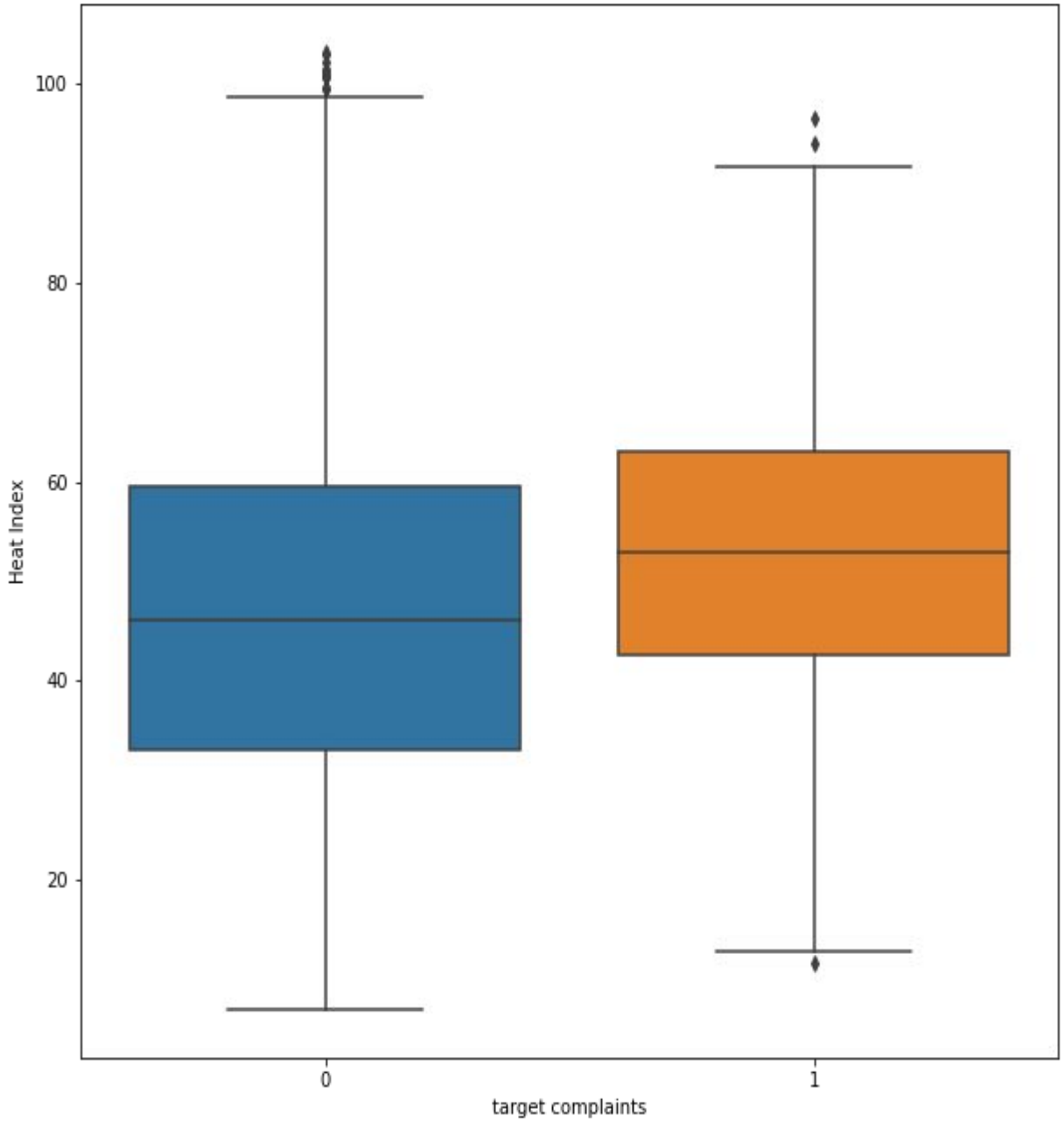
Box plots :- (Distribution of the all the features with the target variable)

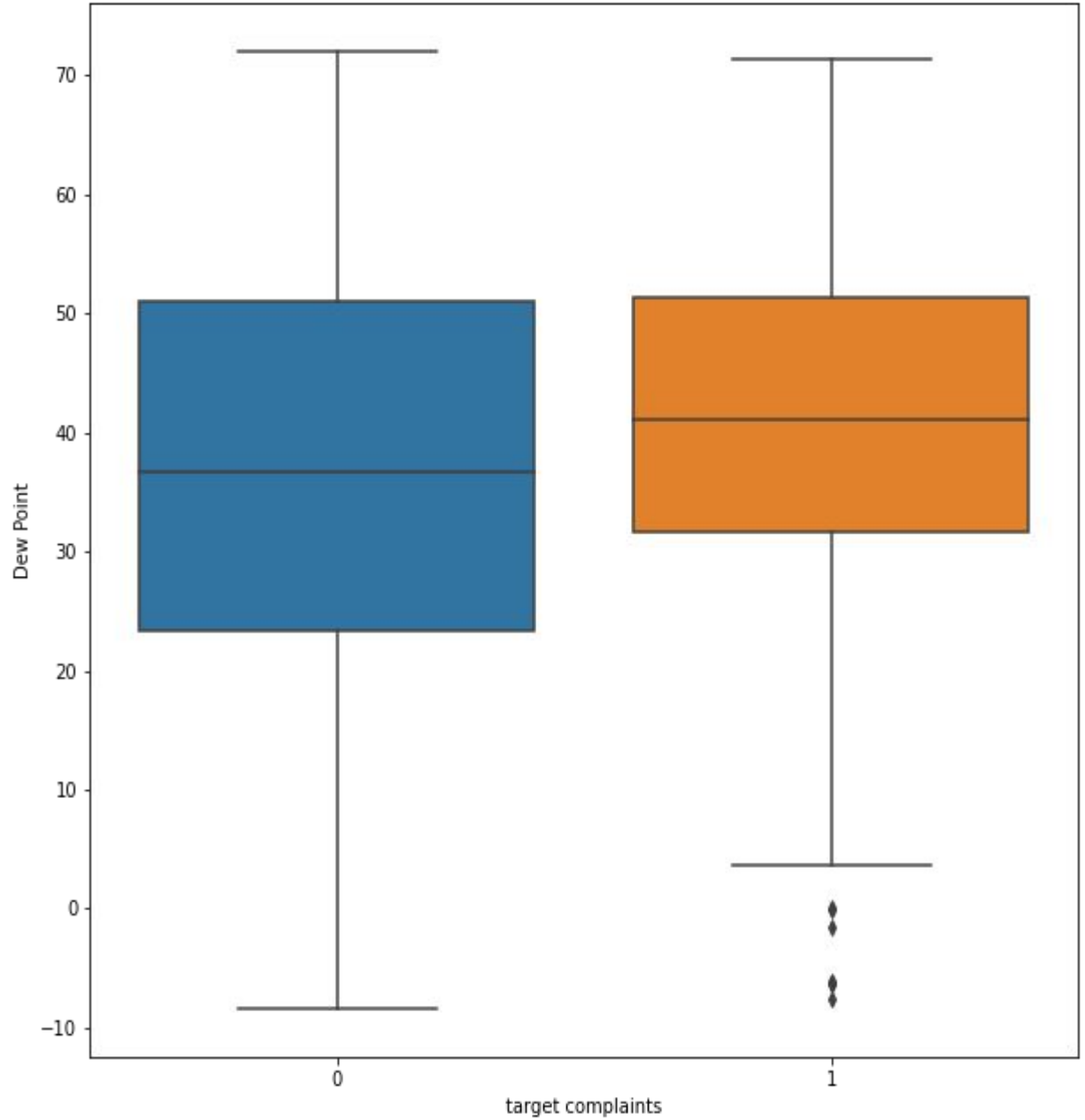




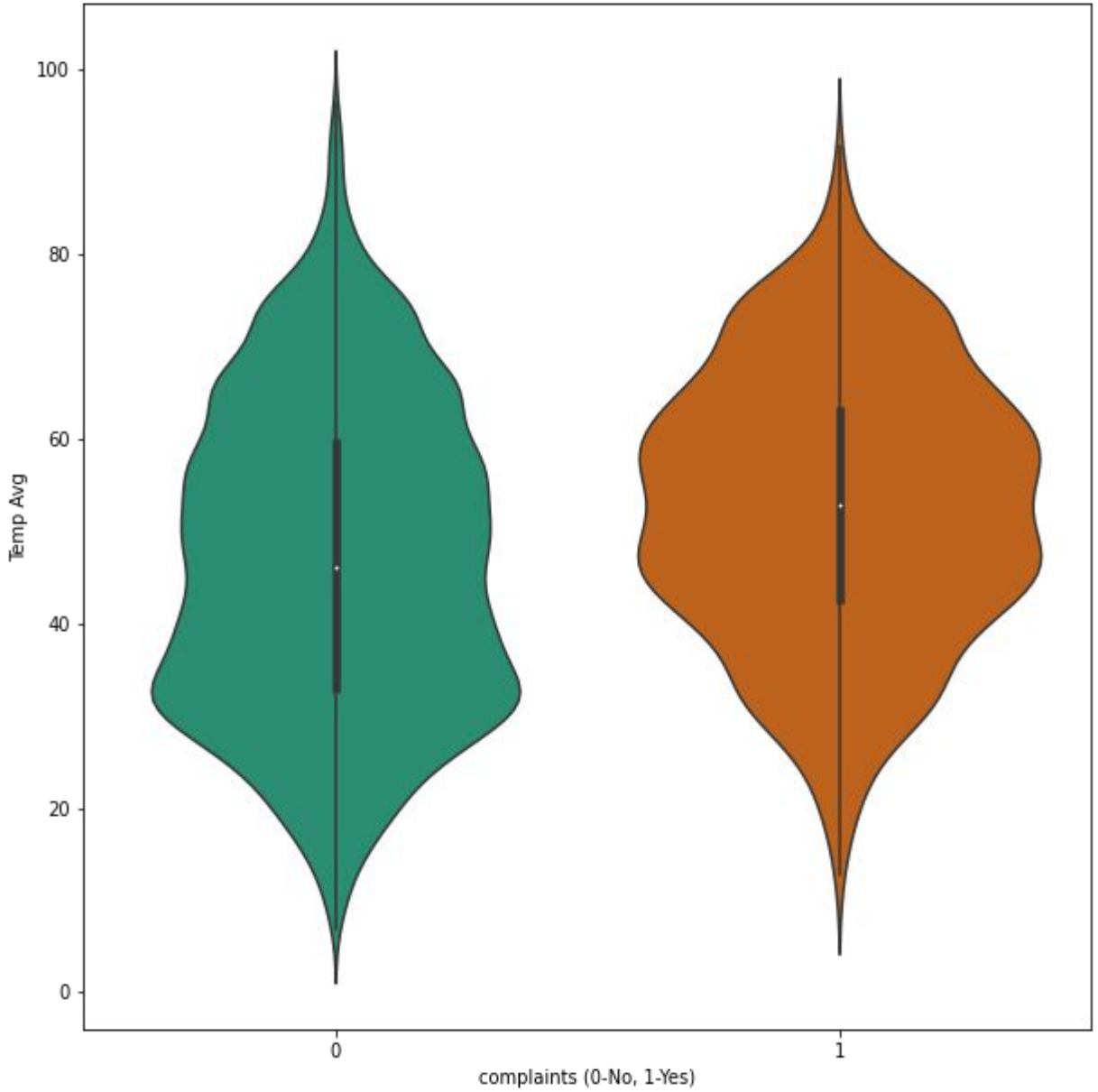

Gust
35
30
25
20
15
10
5
0
0
1
target complaints

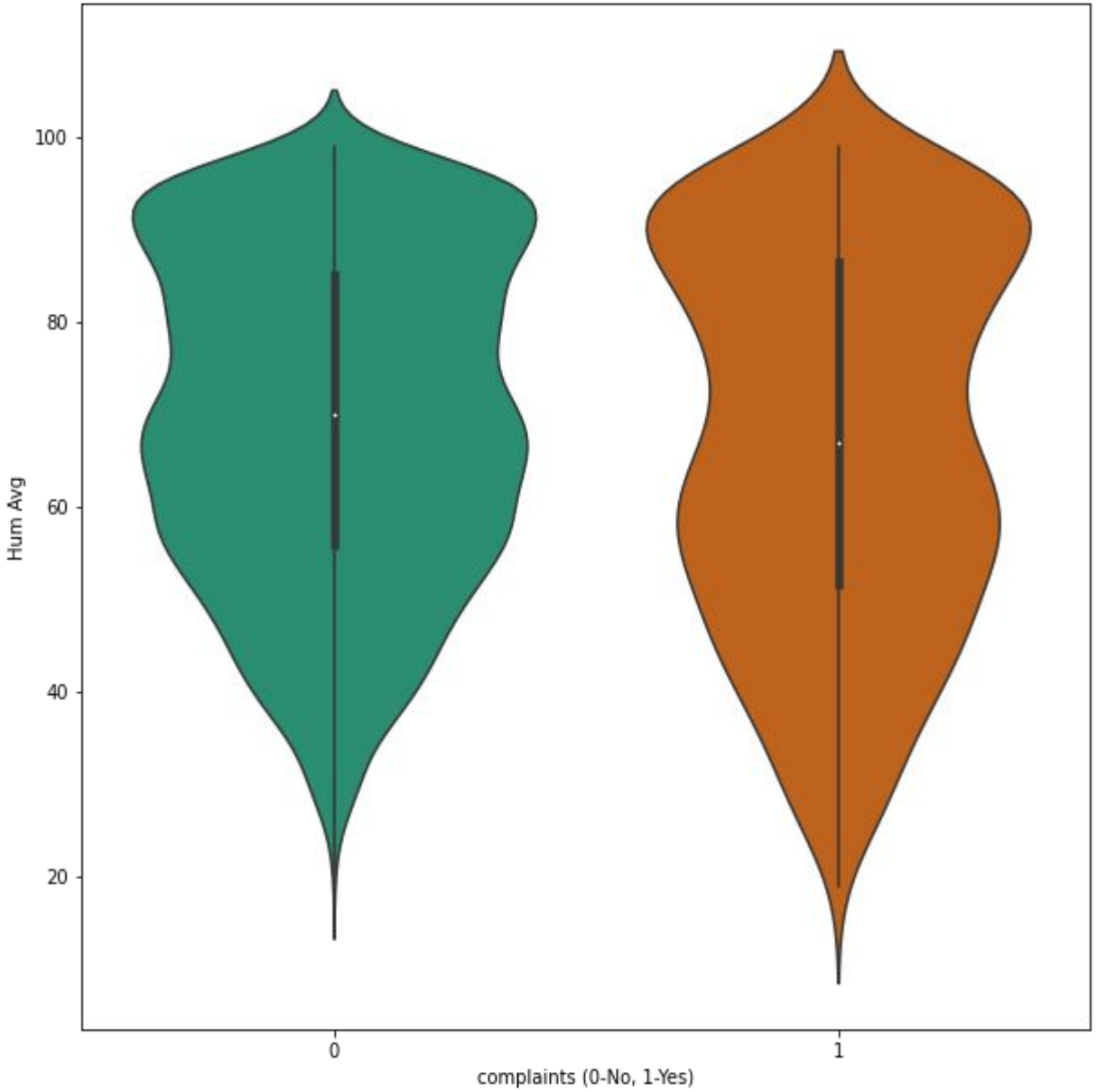

Hum Avg
100
90
80
70
60
50
40
30
20
0
1
target complaints

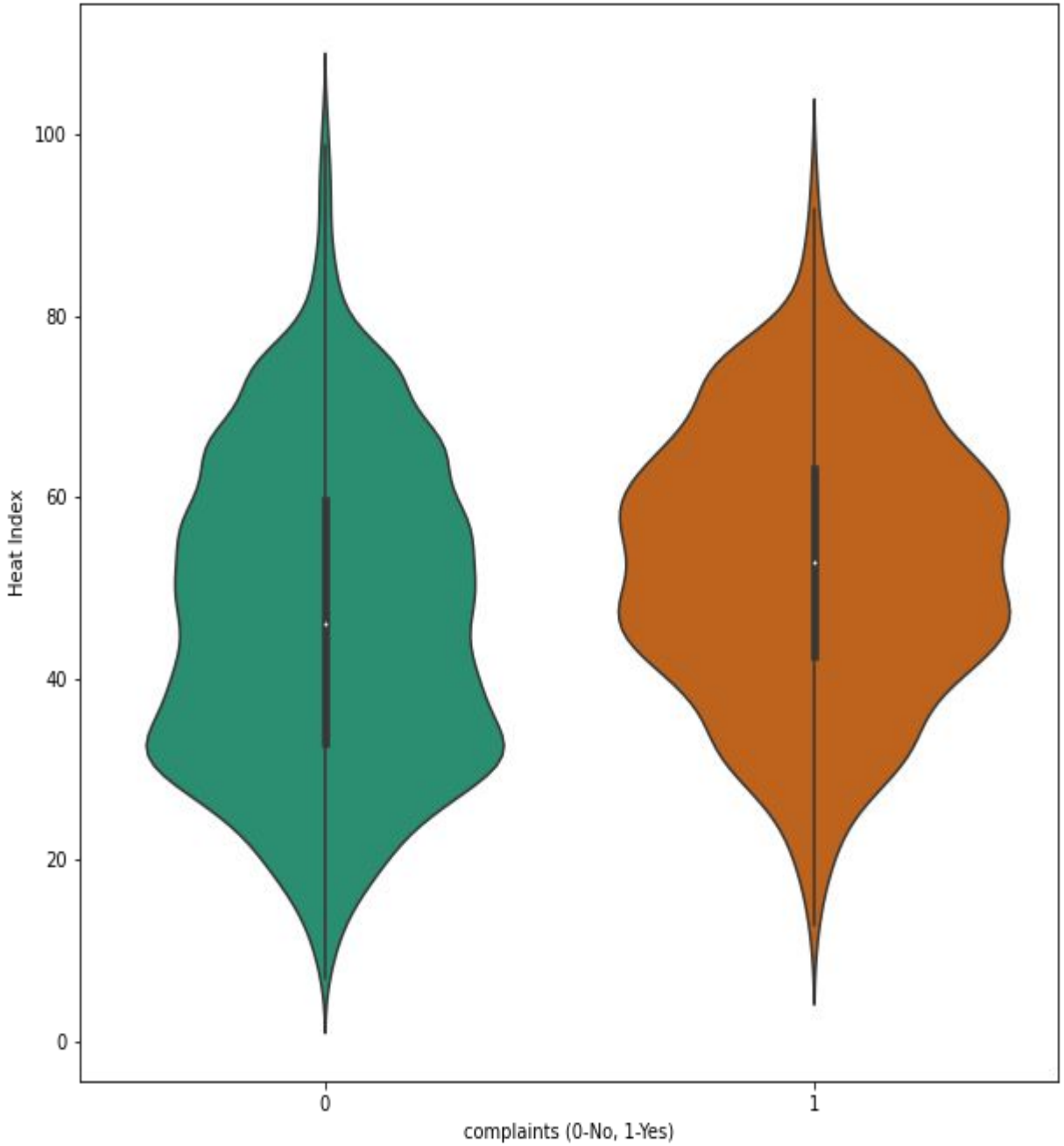

Heat Index
100
80
60
40
20
0
1
target complaints

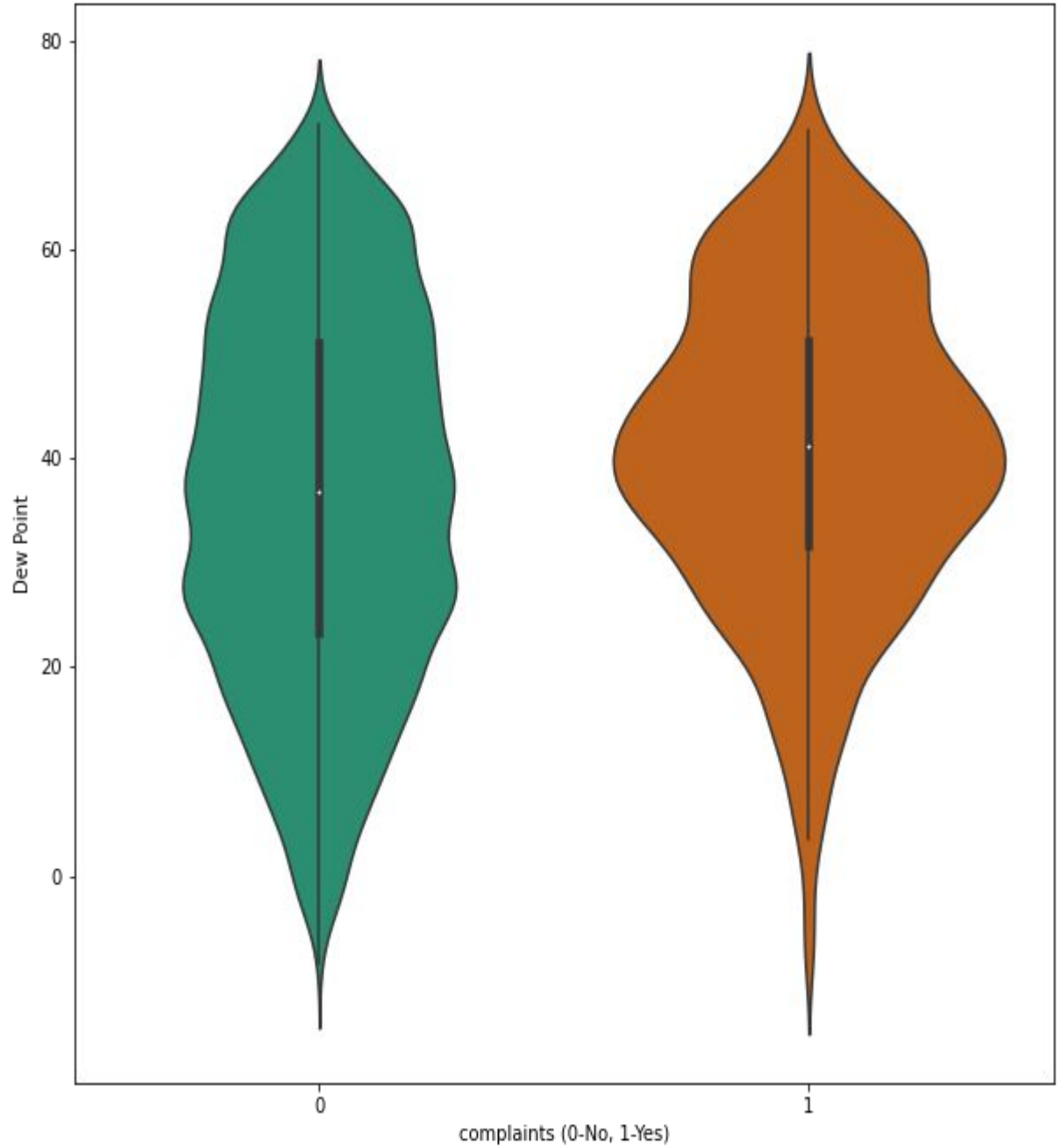

Dew Point
70
60
50
40
30
20
10
0
−10
0
1
target complaints

Violin plots:- (Distribution and KDE of the independent features)

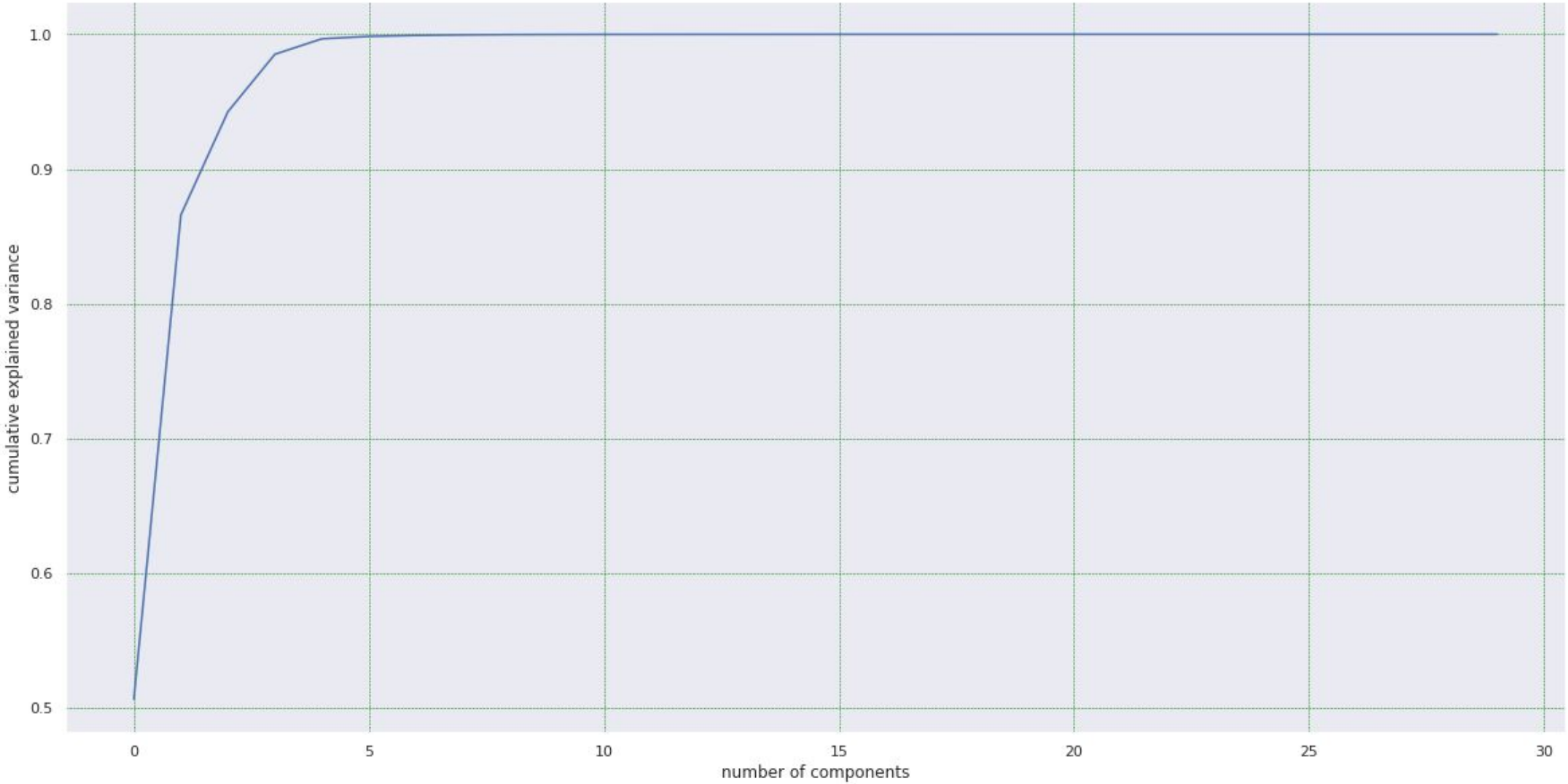





Heat Index
100
80
60
40
20
0
0
1
complaints (0-No, 1-Yes)


Dew Point
80
60
40
20
0
0
1
complaints (0-No, 1-Yes)

Principal Component analysis (Wind Field):-



- The number of components is 4 after which the explained variance becomes constant.

Conclusion And Future Process

Most of the complaints are made in the summer months - June, July, August

The least complaints are made in the winter months

According to our model, the features that correlate heavily with the the occurrence of complaints are temperature, humidity, heat index, and dew point

Conclusion: Complaints are more likely to occur in warmer and more humid weather.

Future Process: Add more weather and complaint data overtime, utilize airport weather data

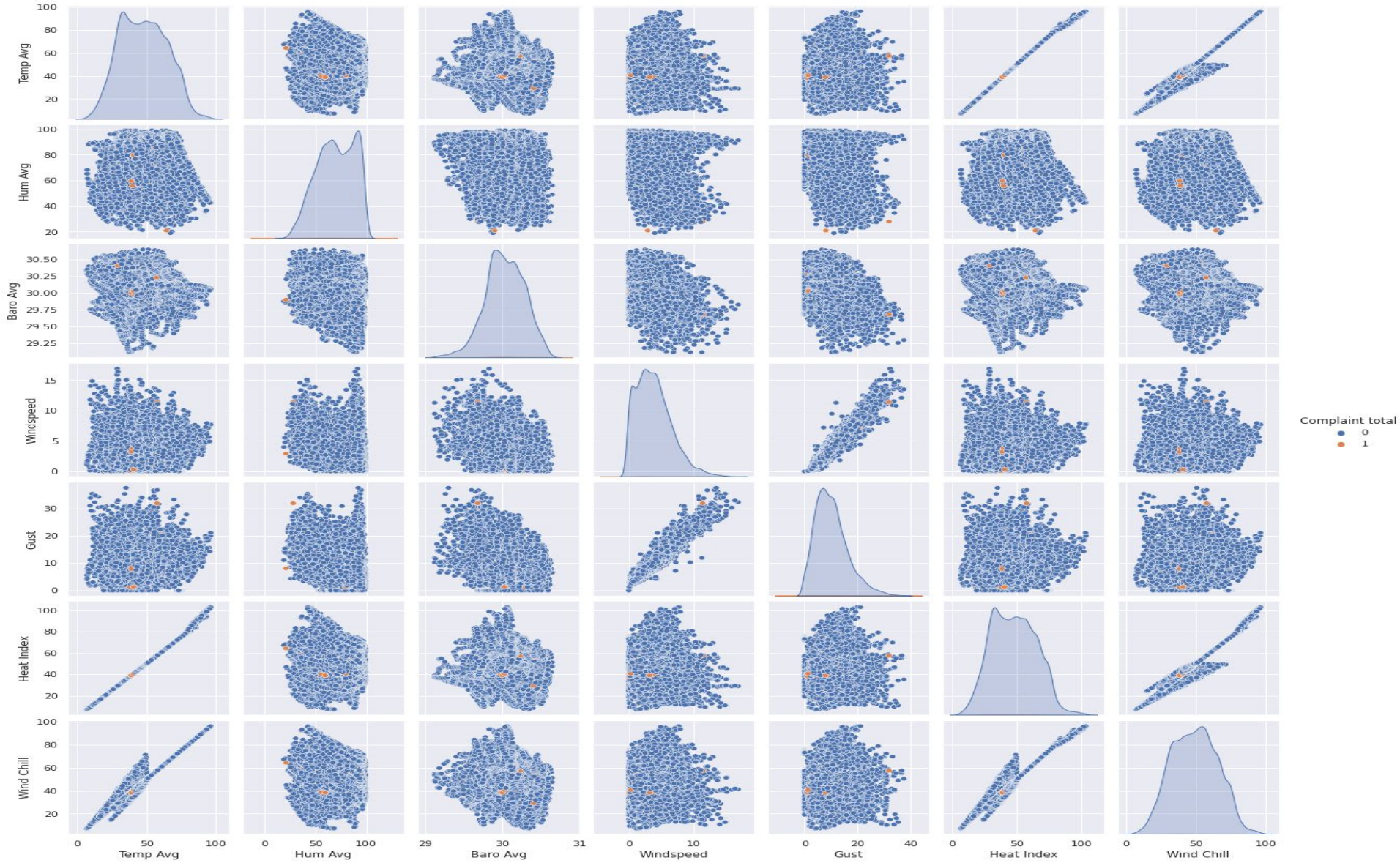
THE END

Supplementary Slides

Pairplot of merged data (Complaints = 0 removed)

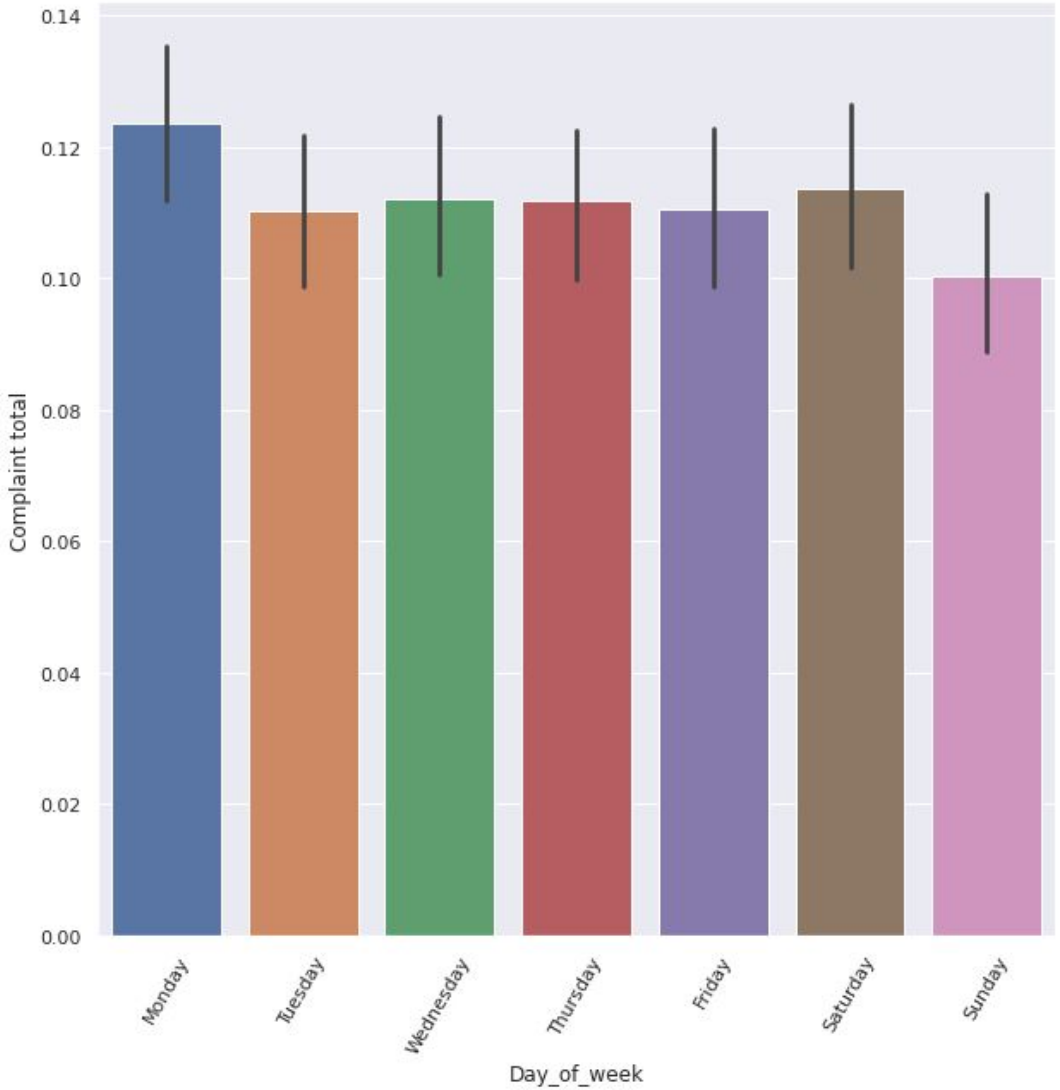
Supplementary Slide:

Pairplot of complaints >3 = 1 and complaints < 4 = 0

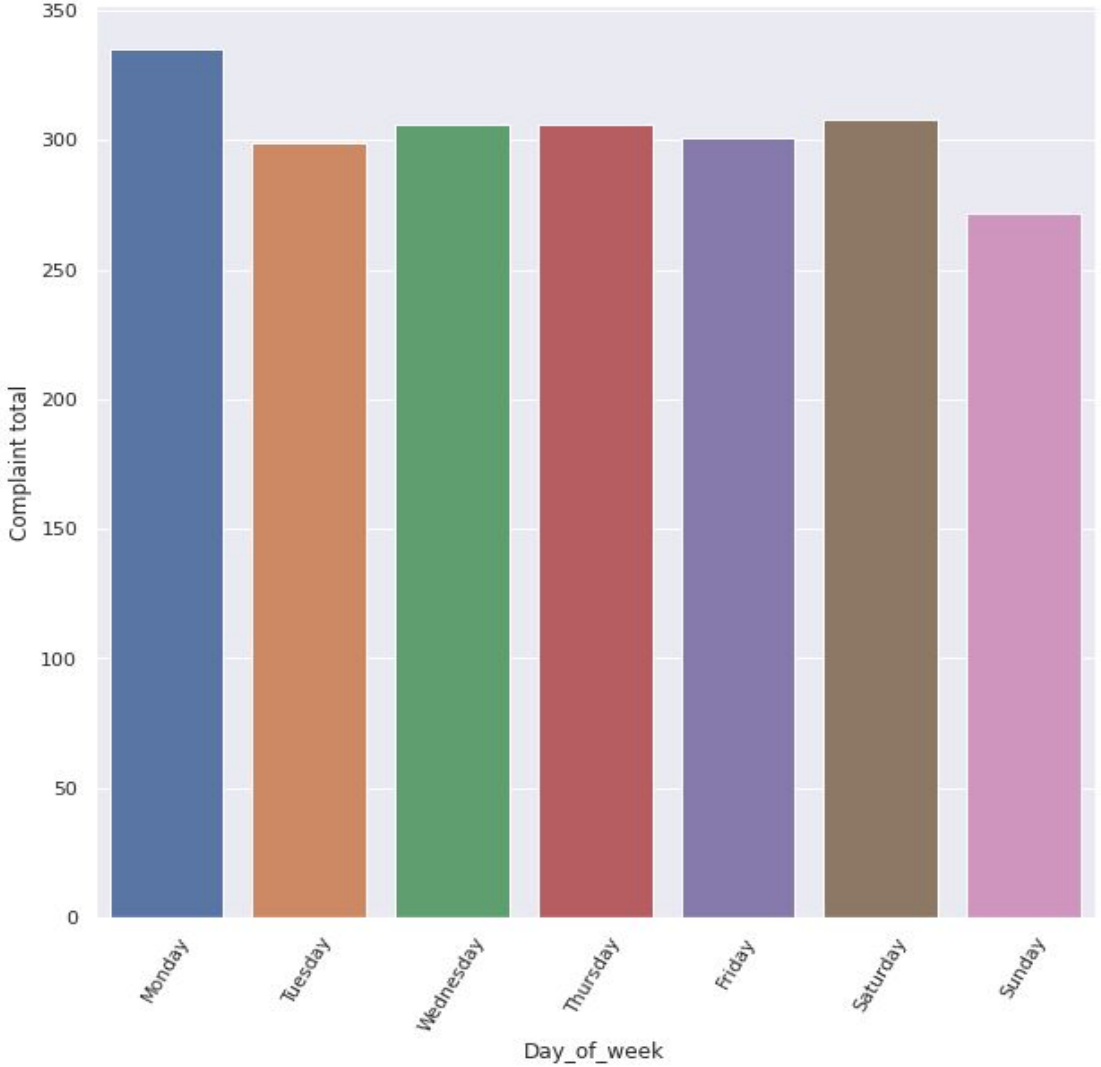


Supplementary Slide:

Complaints by weekday



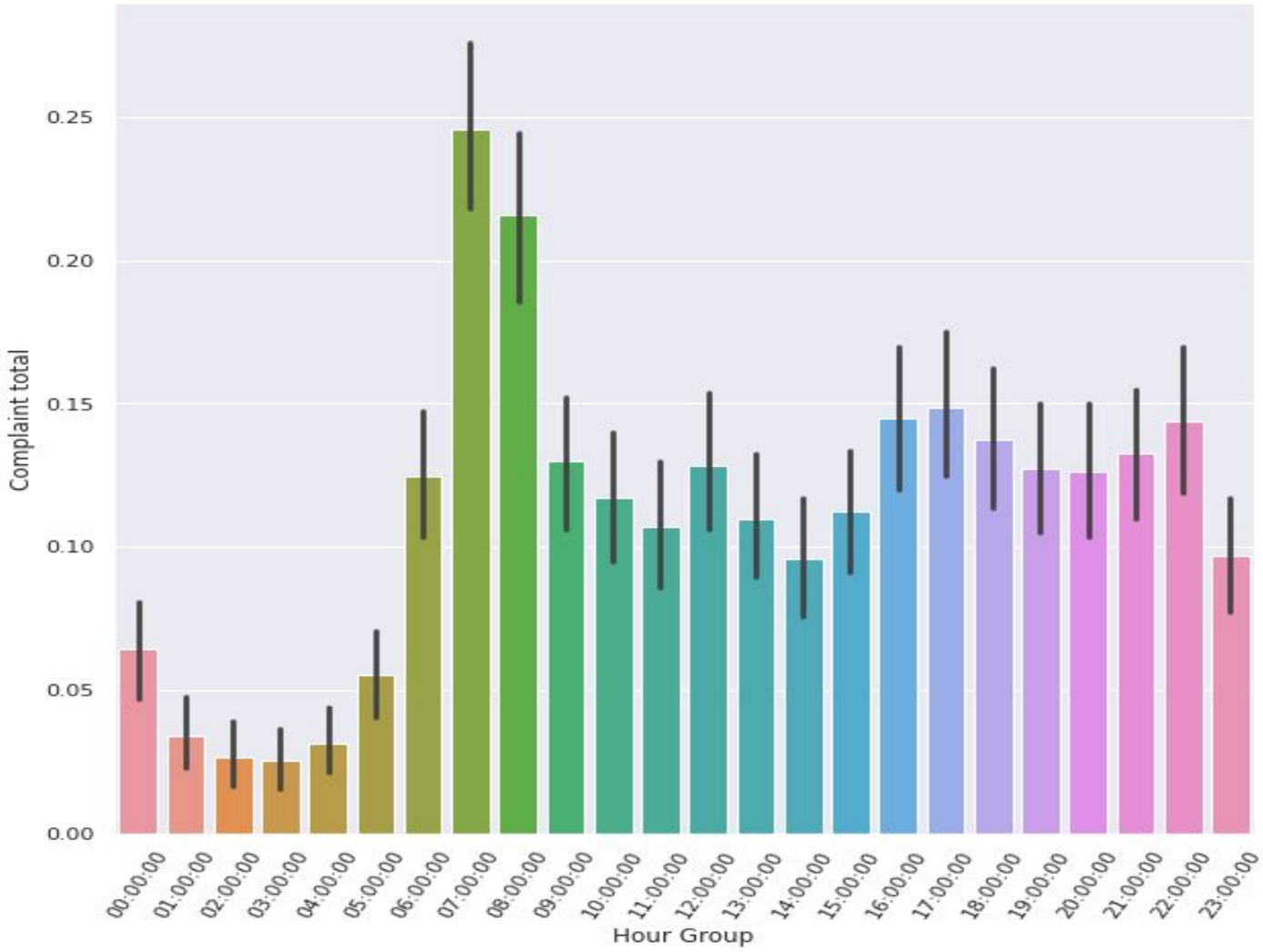
This is the average complaints per day of the week



This is the absolute complaints per day of the week

Supplementary Slide:

Complaints by hour

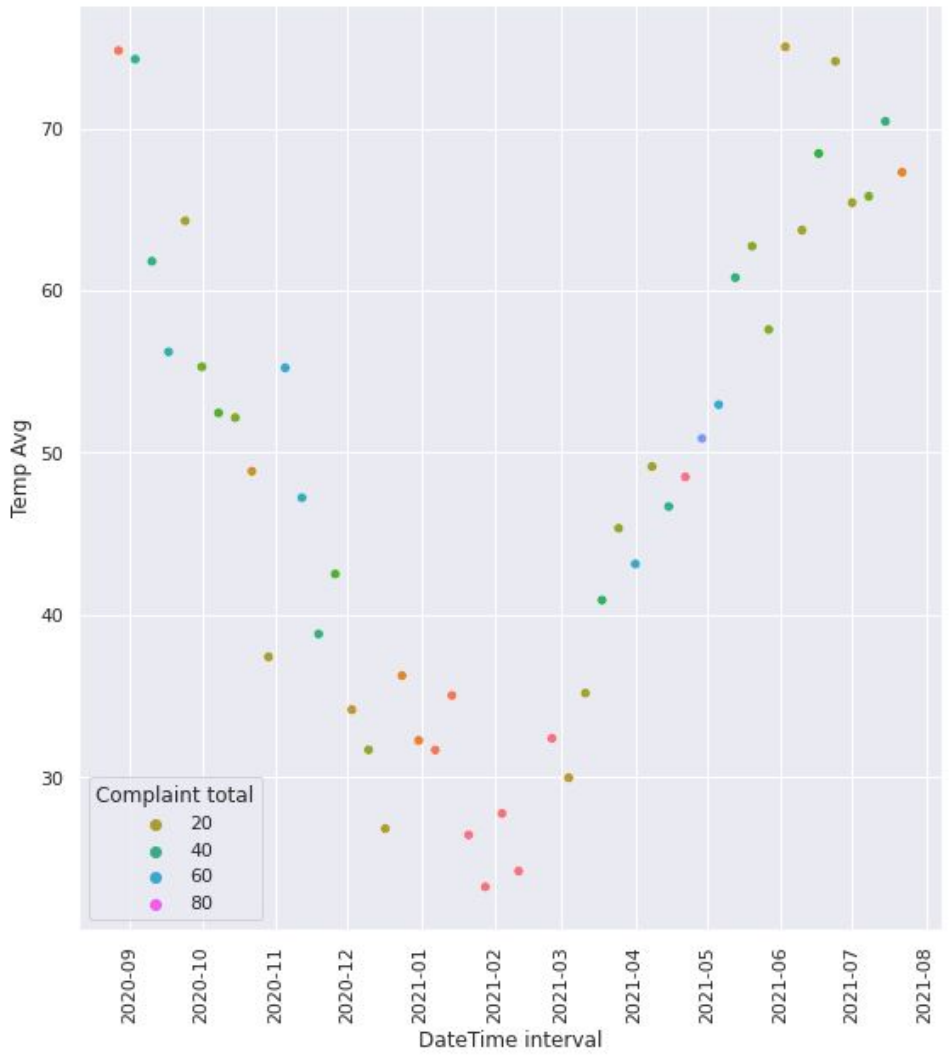


Average amount of complaints by the hour

Supplementary Slide:

Average temperature in 12 hours intervals
(complaints = 0 removed)

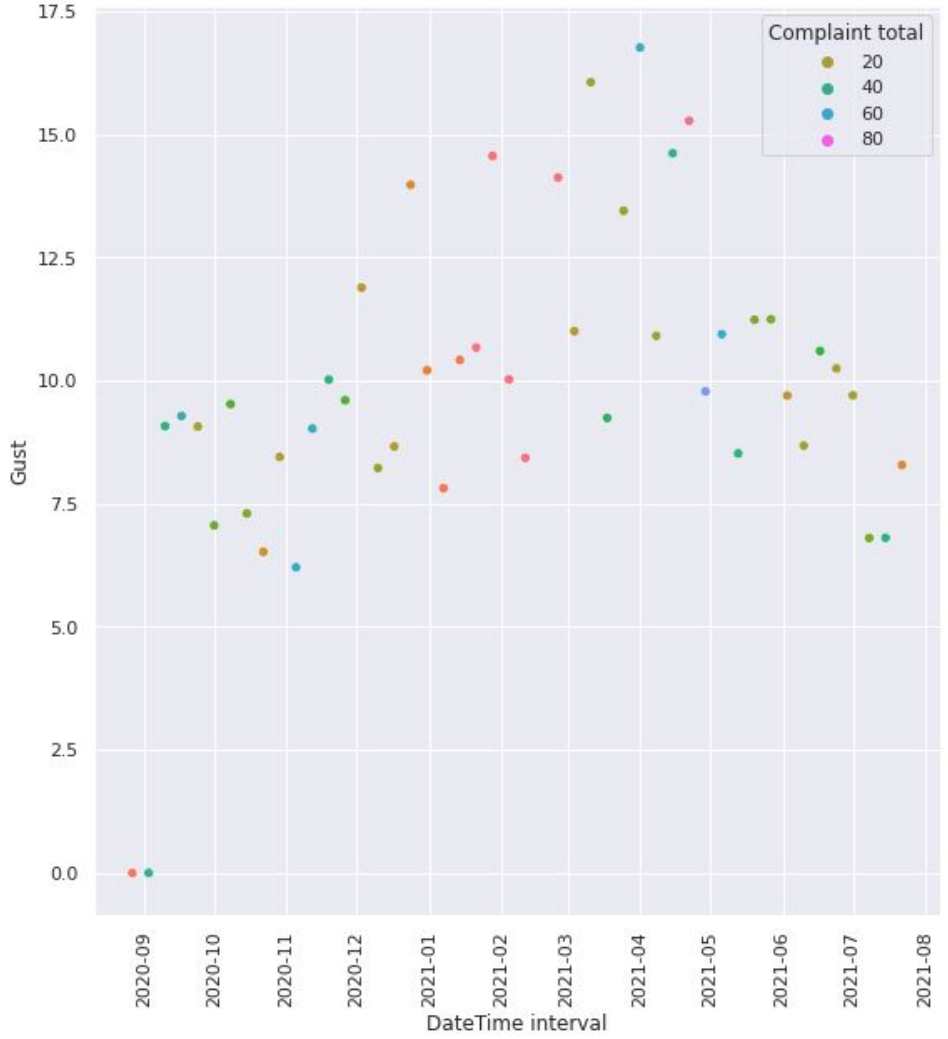
Average temperature in weekly intervals
(complaints = 0 removed)



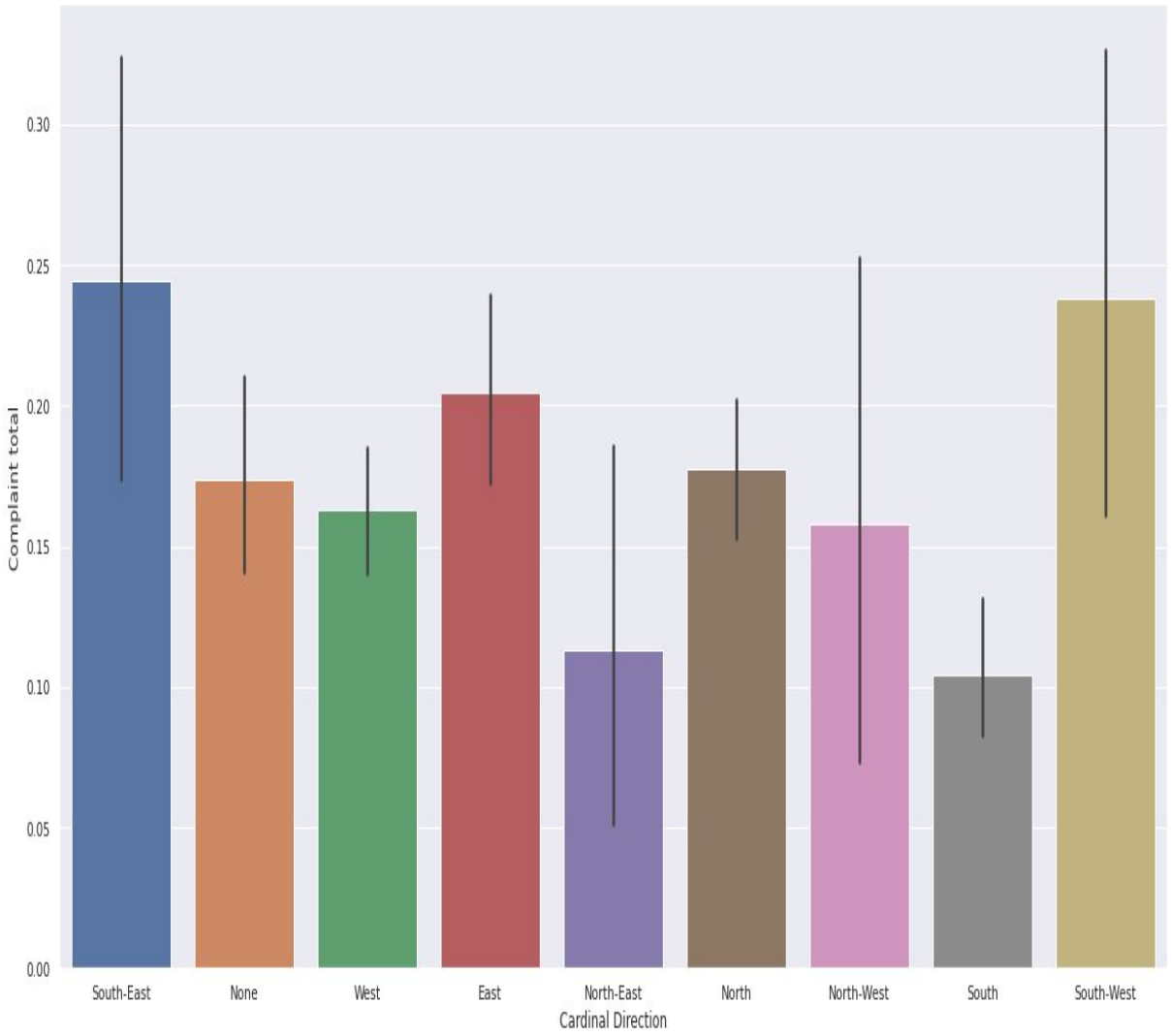
Supplementary Slide:

Average Gust in 12 hour intervals
(complaints = 0 removed)

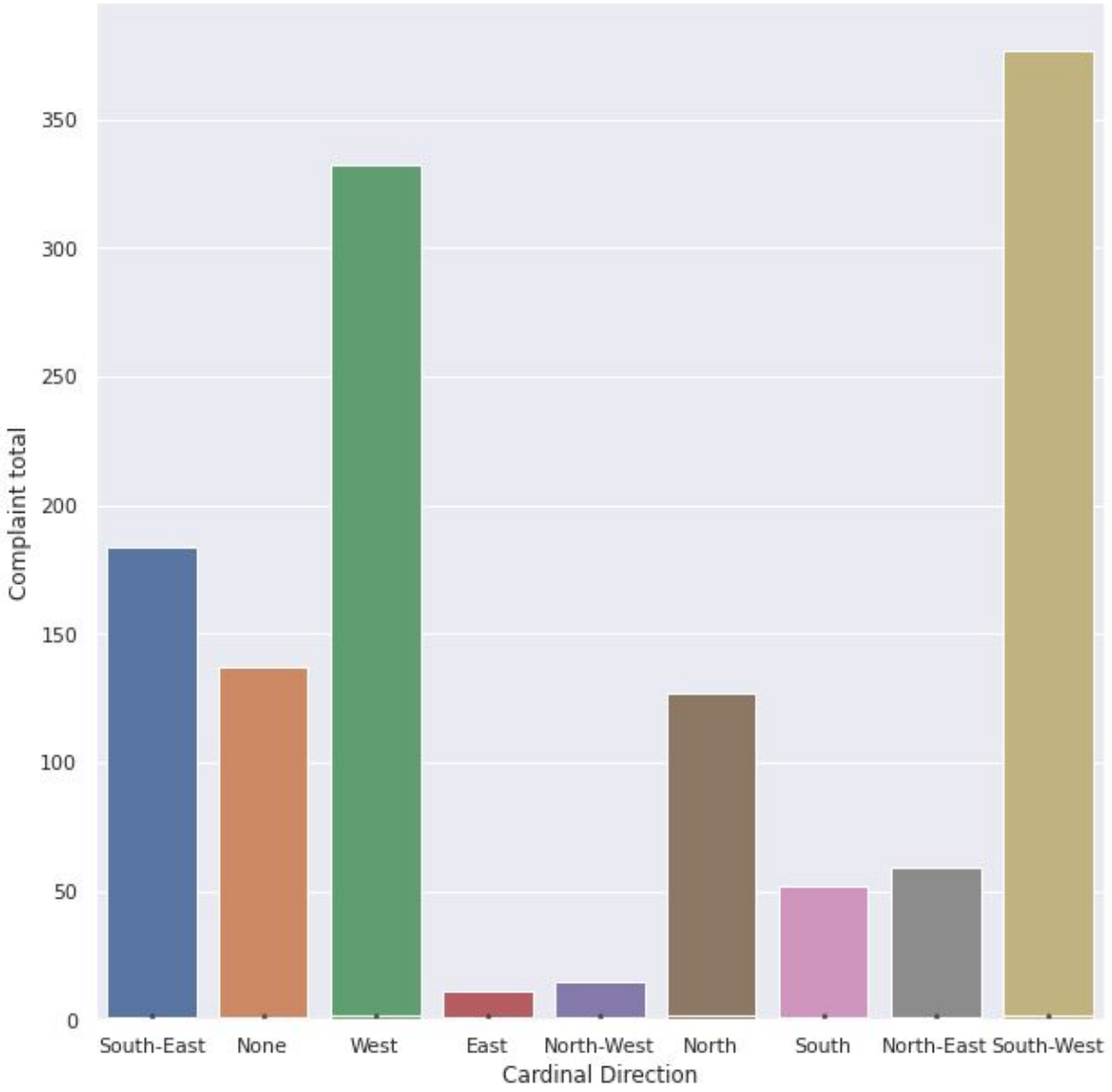
Average Gust in weekly intervals
(complaints = 0 removed)



Supplementary Slide:



Average complaints by cardinal wind direction



Absolute complaints by cardinal wind direction