Technical University of Denmark

02450 Introduction to Machine Learning and Data Mining

Year 2019-2020

Group Project 1

Data: Feature Extraction and Visualization

Group Members

Name	Student Number
Altug Tosun	s181314
Mahsa Eskandarzadeh	s192933
Sukru Han Sahin	s192136

Contributions

Section	Author
1	Altug Tosun, Mahsa Eskandarzadeh, Sukru Han Sahin
2	Altug Tosun, Mahsa Eskandarzadeh, Sukru Han Sahin
3	Altug Tosun, Mahsa Eskandarzadeh, Sukru Han Sahin
4	Altug Tosun, Mahsa Eskandarzadeh, Sukru Han Sahin

1 Description of the Data Set

What is the problem of interest :

A data-set that will be used throughout this project is a data-set on ground-level Ozone pollution. Each point in the data-set consists of several attributes indicating the state of the weather on a given day. In addition to the state, there is an associated response. "The response, referred to as ozone, is actually the log of the daily maximum of the hourly-average ozone concentrations in Upland, California "[1]. The problem of interest posed by the data-set is predicting the ozone value given the other attributes.

Where we obtained the data :

The version of the data used here is that in Hastie and Tibshirani (1990) [3]. The data is obtained from the Stanford website with missing data removed, so that there are complete measurements for 330 days in 1976 [1].

What has previously been done to the data :

This data-set first appeared in Breiman and Friedman(1985) [2] and was analyzed extensively by Hastie and Tibshirani (1990) [3]. The goal is to use the meteorological covariates to predict ozone concentration, which is a pollutant at the level of human activity. For each of these days, the response variable of interest is the daily maximum one-hour-average ozone level in parts per million at Upland, California. A research paper by Leo Breiman reveals that the data-set was previously used to predict ozone levels of the Los Angeles Basin 12 hours prior to their actual occurrence, which gave government officials enough time to react and avoid danger to the public and themselves. Nevertheless, the data-set they used is a superset of the one at hand and was collected over a period of 7 years. Sadly, the efforts of the scientists were not as the best predictor they created, because it had a false alarm rate which was simply too high for use.

What is the primary machine learning modeling aim for the data :

The goal is to use the meteorological covariates to predict ozone concentration, which is a pollutant at the level of human activity. For each of these days, the response variable of interest is the daily maximum one-hour-average ozone level in parts per million at Upland, California. Regression is the task posed by the data-set. Particularly, the task is to predict the ozone value based on the other attributes. To accomplish this, the data will need to be normalized, which avoids errors due to relative scaling among attributes. For applying Principle Component Analysis (PCA) on the data, the mean for each attribute will be subtracted, and the attributes will be divided by their respective standard deviations. According to plan, PCA will identify the main and relevant air pollution sources. PCA is going to be used to transfer and reduce the number of predictive variables to new input variables, i.e. a set of Principle Components. In addition to that, all of the attributes are already numerical, so there is no need for transformation.

2 Explanation of Data-set Attributes

Description about the attributes :

This data set is a matrix containing the following columns :

Name	Description	Discrete /Continuous	Nominal/Ordinal /Interval/Ratio
Ozone	Daily maximum one-hour-average ozone reading (parts per million) at Upland, CA.	Discrete	Ratio
Pressure.Vand (vh)	500 millibar pressure height (m) measured at Vandenberg AFB.	Continuous	Ratio
Wind	Wind speed (mph) at Los Ange lesInternational Airport (LAX).	Continuous	Ratio
Humidity	Humidity in percentage at LAX.	Continuous	Ratio
Temp.Sand(temp)	Temperature (degrees F) measured at Sandburg, CA.	Continuous	Interval
Inv.Base.height(ibh)	Inversion base height (feet) at LAX.	Continuous	Ratio
Pressure.Grad(dpg)	Pressure gradient (mm Hg)from LAX Daggett, CA.	Continuous	Ratio
Inv.Base.Temp(ibt)	Inversion base temperature (degrees F) at LAX.	Continuous	Interval
Visilibity(vis)	Visibility (miles) measured at LAX.	Continuous	Ratio
doy	day of the year	Discrete	Ordinal

Table 1: Description of Attributes(descriptions taken from [3])

Description of data issues(i.e. missing values or corrupted data) :

The first thing to notice is that most variables have a strong non-linear association with the ozone level, making prediction feasible but requiring a flexible model to capture the nonlinear relationship. Most variables display a strong relationship with ozone, and all but the first are clearly nonlinear[5]. There are not any missing values in the samples which were taken, but there are days of the year in which measurements were not taken, as mentioned in Table 1. This is exemplified by the fact that the data-set contains only 330 samples instead of 365 and that there exist gaps between certain days. The gaps could complicate temporally based learning and graphs.

Since date is related with time series, we will set date object to the object. In that way, we won't let PCA or other statical analysis be disrupted by date object.

date	ozone	vh	wind	humidity	temp	ibh	dpg	ibt	vis
1976-01-04	3	5710	4	28	40	2693	-25	87	250
1976-01-05	5	5700	3	37	45	590	-24	128	100
1976-01-06	5	5760	3	51	54	1450	25	139	60
1976-01-07	6	5720	4	69	35	1568	15	121	60
1976-01-08	4	5790	6	19	45	2631	-33	123	100
1976-01-09	4	5790	3	25	55	554	-28	182	250
1976-01-10	6	5700	3	73	41	2083	23	114	120
1976-01-11	7	5700	3	59	44	2654	-2	91	120
1976-01-12	4	5770	8	27	54	5000	-19	92	120
1976-01-13	6	5720	3	44	51	111	9	173	150

Figure 1: doy attribute elimination

Summary statistics of the attributes :

Number of values are equal for each attribute which means there isn't any NaN values. Minimum and Maximum values of each attribute differs since their calculation range units are different. And standard deviation differences really differs as well. As it is seen from Figure 2, mean values of attributes are ranging from 5 to 5750. There is an important gap, which would misguide our analysis in terms of its value differences. For that reasons, we should standardize our data to compare features that have different units or scales and have further statistical analysis.[5]

Index	ozone	vh	wind	humidity	temp	ibh	dpg	ibt	vis
count	330	330	330	330	330	330	330	330	330
mean	11.7758	5750.48	4.89091	58.1303	61.7545	2572.88	17.3697	161.161	124.533
std	8.01128	105.708	2.29316	19.865	14.4587	1803.89	35.7172	76.6794	79.3624
min	1	5320	0	19	25	111	-69	-25	0
25%	5	5690	3	47	51	877.5	-9	107	70
50%	10	5760	5	64	62	2112.5	24	167.5	120
75%	17	5830	6	73	72	5000	44.75	214	150
max	38	5950	21	93	93	5000	107	332	350

Figure 2: Summary statistics of the attributes

3 Visualization

Issues with outliers in the data :

As it can be seen from boxplot of attributes(Figure 3), there exists outliers especially in vh, a little in wind, ozone and visibility. Therefore, only vh attribute has an issue with the outliers. The other attributes does not have values that differ greatly from the median.

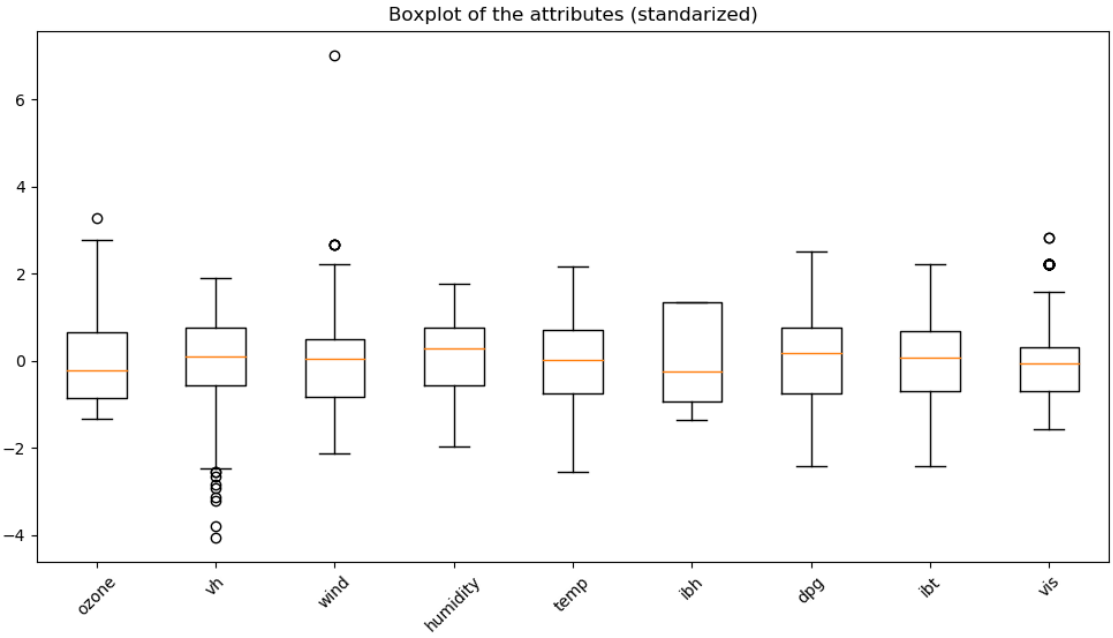


Figure 3: Boxplot of the attributes

Normal Distribution :

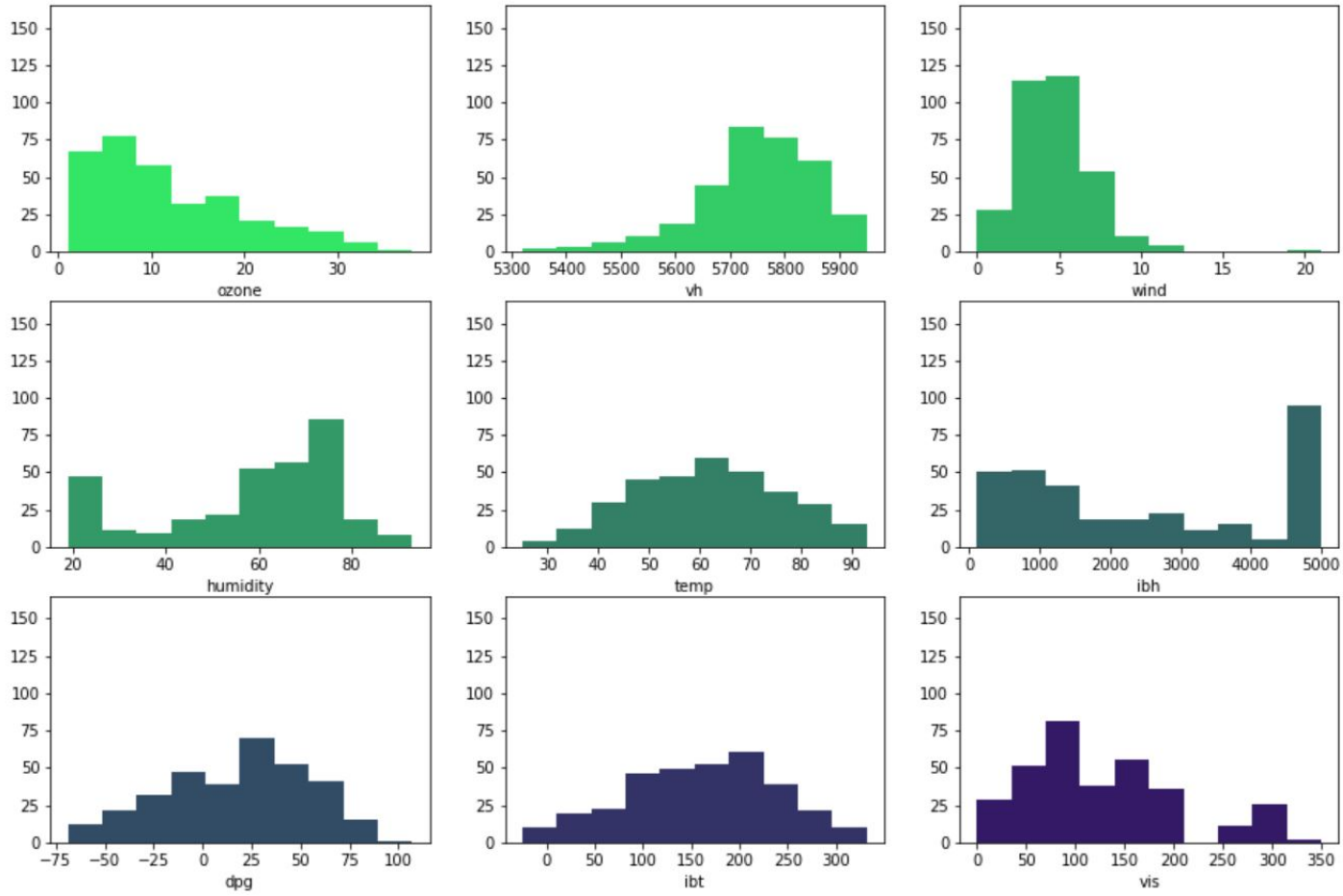


Figure 4: Histogram of the attributes

Normal Distribution depiction with histogram for each attribute are shown. At the "temp", "ibt" and maybe "dbg" and "dbg" attributes we can say that they seem as normally distributed. But for the other attributes it is not in that way. As it was obvious from the box-plot above wind attribute outlier has affected histogram as well.

	ozone	vh	wind	humidity	temp	ibh	dpg	ibt	vis
P-values	7.347073e-08	6.609357e-12	1.039288e-22	3.110651e-08	0.008593	0.0	0.000294	0.004159	1.997774e-07

Figure 5: P-Values Table

When the P-values are tested as shown, It supports our ideas of normality since greater values results in better normality. Another way to check it looking at Q-Q Plots.

When the Q-Q PLots are examined, we can verify our comments as aforementioned. We know that these plots will be later used to verify the prerequisites of Regression Analysis. "wind" attribute shows again its outlier. "temp", "ibt" and "dbg" are closer to normal distribution. "vis" and "ibh" are far away from normal distribution.

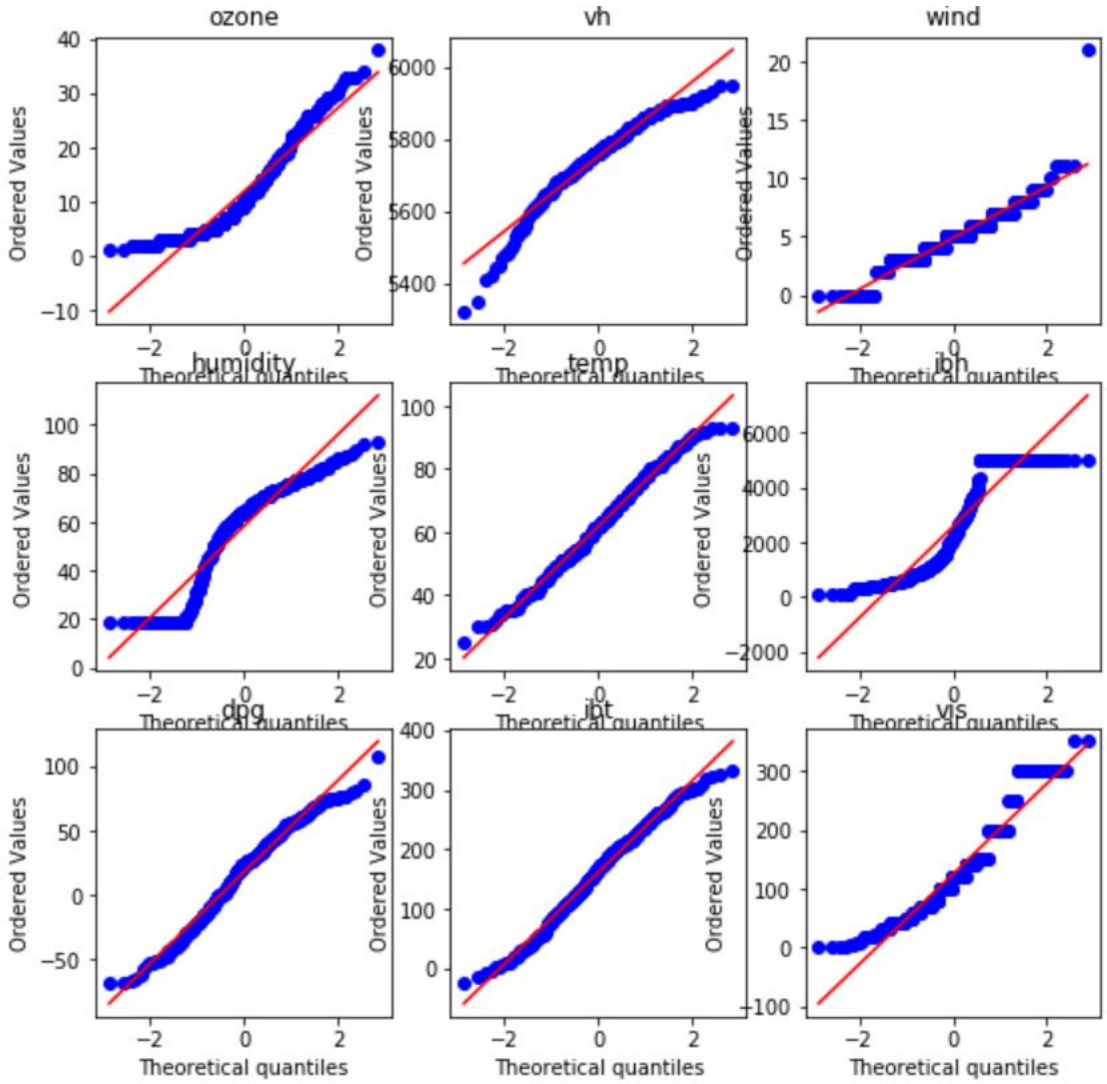


Figure 6: QQPlot of Attributes

Variable correlation :

Heat-map is created in order to determine the correlation level between the attributes(Figure 7).Lighter colors mean more correlation between attributes. Black means highly negatively correlation. As it can be seen from the heat-map, ibt attribute has high correlation with temp, vh and ozone. Also, temp attribute has high correlation with vh and ozone. Besides from these ones, the attributes have low level of correlation.

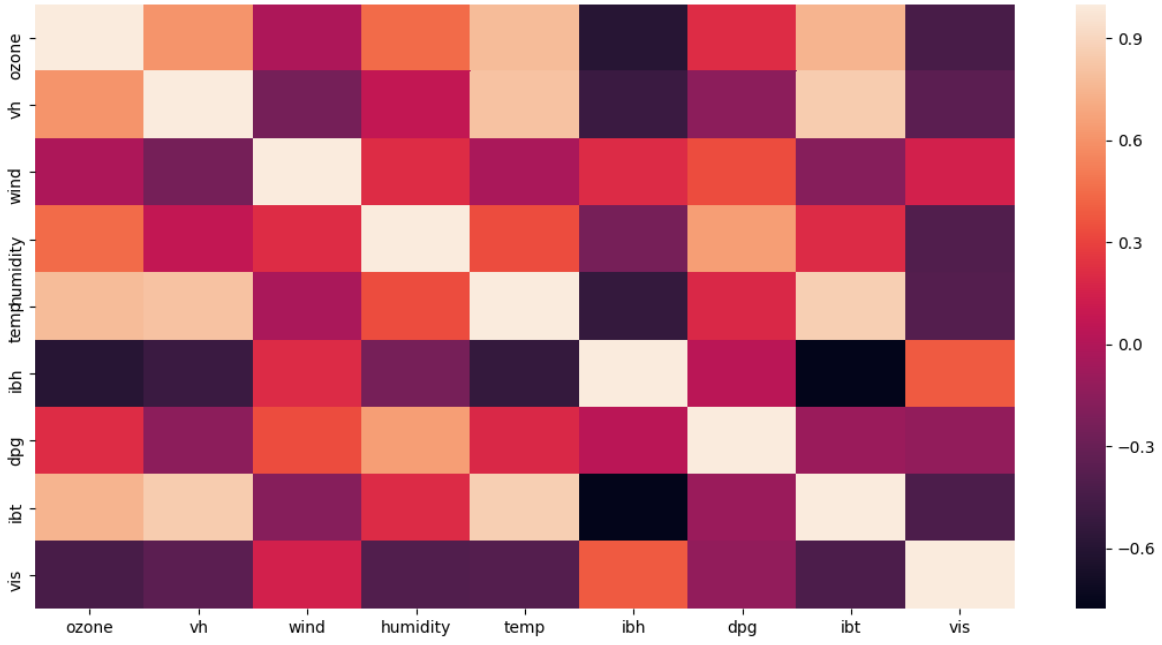


Figure 7: Correlation heat-map of the attributes

Feasibility of primary machine learning modelling :

Regression is the primary machine learning model for this data-set. The normality of residuals is required by the regression model. Unlike the normality of residuals, normality of the independent variables is not required by regression model. Therefore, the non-normality in the data-set is not a problem for our primary model. But, the independent variables should not be dependent in order to apply the regression. From heat-map(Figure 7) we concluded that ibt attribute has high correlation with temp, vh and ozone. Also, temp attribute has high correlation with vh and ozone. So, it might be good idea to exclude these attributes when creating the regression model.In addition to that, the effects of correlation can be reduced by omitting attributes using backward and forward selection.

The amount of variation using PCA :

Figure 8 represents the variants explained by the principal components with a chosen threshold of 0.9. According to figure, it is easy to see that the first 5 principal components explain more 90 percent of the variance. Therefore, only these 5 components will be included in further steps of analysis.

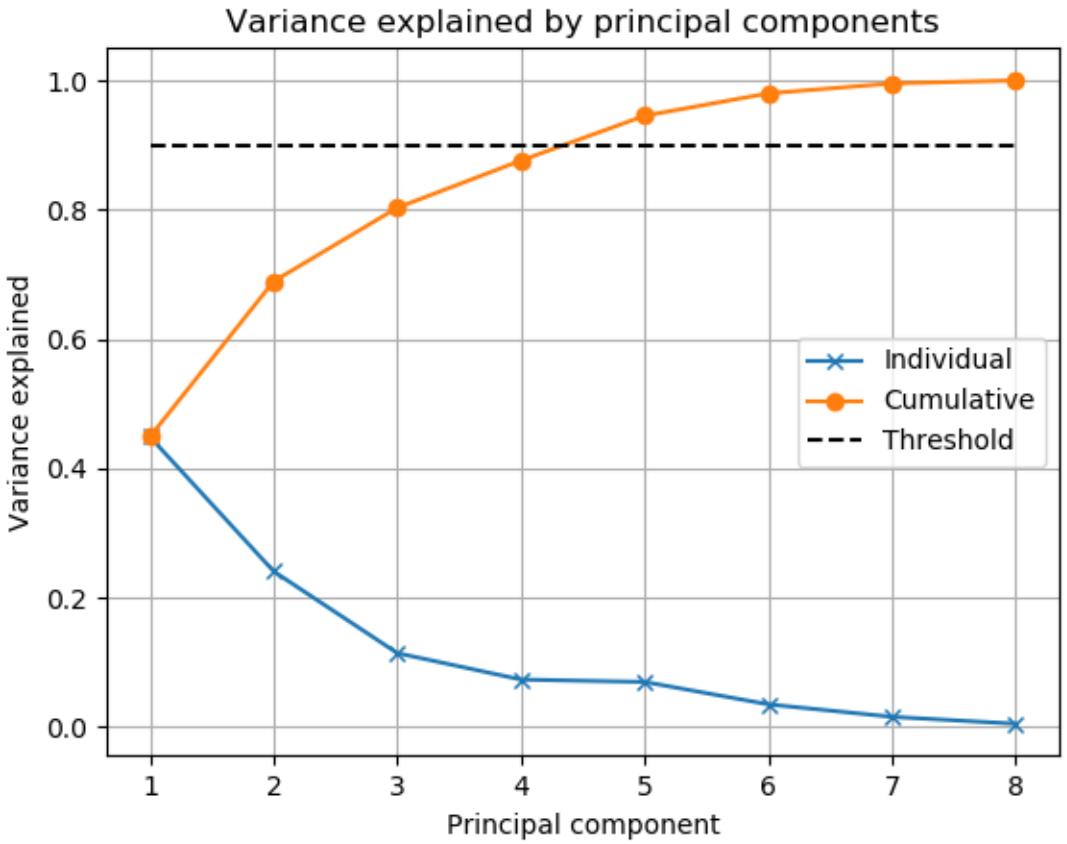


Figure 8: Variance explained by principle components

The principal directions of the considered PCA components :

Figure 9 and Figure 10 shows the direction of the principal components which will be taken into consideration. For instance, it can be seen that first principal component which explains more than 40 percent of the entire variance has a large magnitude in the dimensions of the attributes vh, temp and ibt.

Index	vh	wind	humidity	temp	ibh	dpg	ibt	vis
PCA1	-0.450714	0.112851	-0.201456	-0.470796	0.410164	-0.0354356	-0.502636	0.31584
PCA2	-0.189924	0.424827	0.575785	0.0749914	0.0673554	0.638077	-0.116508	-0.145071
PCA3	0.229855	0.654137	-0.233857	0.297646	0.0831759	-0.0343697	0.188289	0.576573
PCA4	-0.129269	0.607631	-0.110674	-0.198202	-0.143725	-0.435347	0.0230364	-0.59315
PCA5	-0.370659	0.0241199	0.122276	-0.247745	-0.795105	-0.0182645	0.109412	0.375773

Figure 9: The principal directions of the considered PCA components

At Figure 10, we can see clearly which attributes have been explained mostly by which PCAs. Therefore Figure 10 is a good visualization for Figure 9.

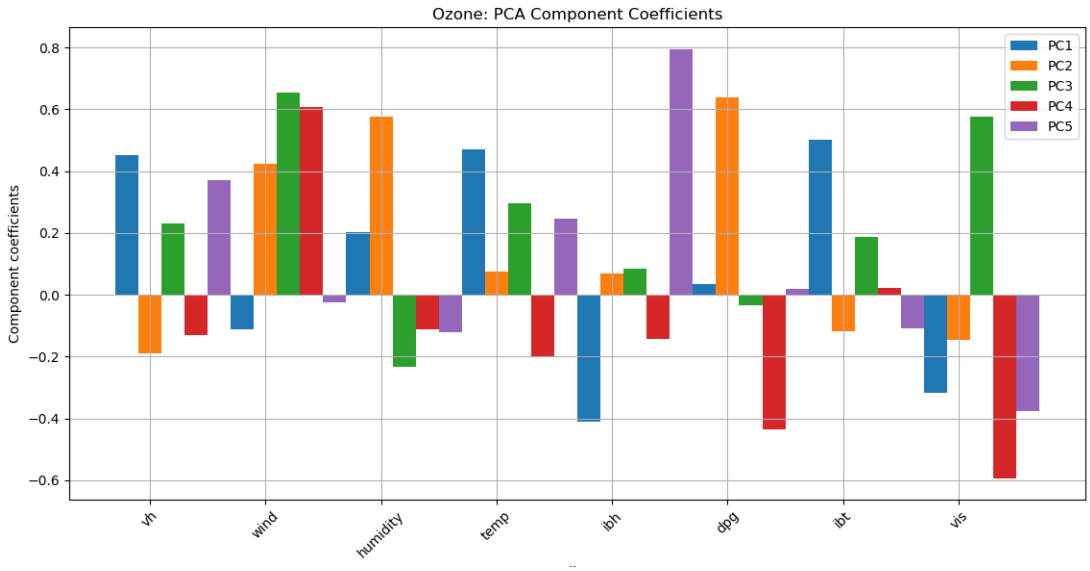


Figure 10: PCA component coefficients of selected principal components

The data projected onto the considered principal components :

The data was projected onto the first 5 principal components, and graphs are created accordingly. Points on the graphs colorized depending on their ozone levels. The projections onto the component 1 and component 5 are shown in Figure 11. All of the other projections can be seen in Appendix A.

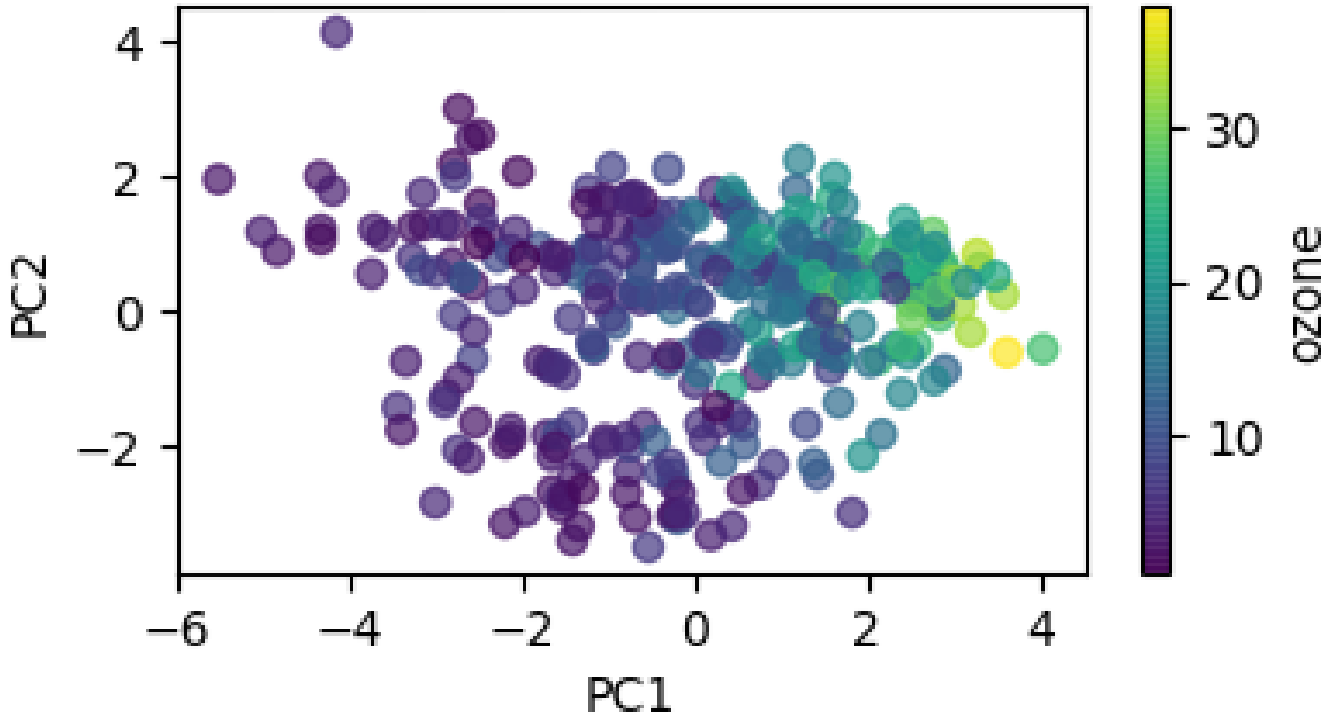


Figure 11: PCA projections for components 1 and 2

4 Conclusion

There are 10 attributes in LA Ozone data-set. Amongst these attributes ozone level is going to be considered as a response variable in the primary machine learning task. In the data-set, there are few missing measurements but data-set is mostly clean. In the data-set, a few number of outliers occurs. Therefore, these outliers should not affect the learning process. Most of the attributes seem to be not normally distributed, but it does not necessarily mean that regression will fail. Probably the biggest problem to predict ozone level according to these data-set is the level of correlation between attributes. But there are methods for dealing with problem of the lack of severeness in correlation levels. Still, it is easy to say that correlation levels is the most undesirable trait of the data. To sum up, by the light of all information we obtained, our primary machine learning goal, which is regression, is feasible.

We have chosen a threshold of 90 percent to explain the variants by the principal components. We concluded that only the first 5 principal component should suffice to explain variants more than 90 percent. The data projected onto the first and second component as it can be seen from figure 8. Figure reveals the high variation of ozone values along the x-axis. This variation implies that the first principal component could play an crucial role in establishing an accurate predictor for ozone levels.

5 References

[1] T. Hastie, R. Tibshirani, and J. Friedman, 'The Elements of Statistical Learning', 2009. [Online]. Available: https://web.stanford.edu/ hastie/ElemStatLearn. [Accessed : 29 - September 2019].

[2]Leo Breiman, Department of Statistics, UC Berkeley. Data used in Leo Breiman and Jerome H. Friedman (1985), Estimating optimal transformations for multiple regression and correlation, JASA, 80, pp. 580-598.

[3] Hastie, T., and R. Tibshirani. 1990. Generalized additive models. London: Chapman and Hall.

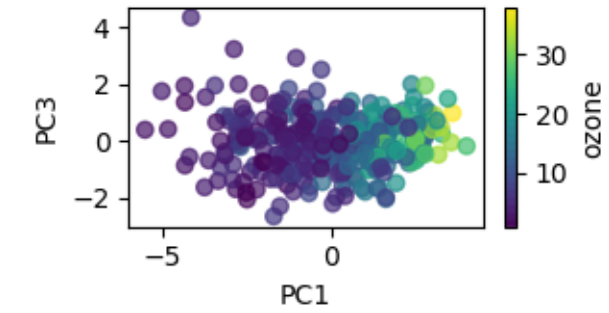
[4] Los Angeles ozone pollution data , ibr documentation built on May 2, 2019, 8:22 a.m available online at :.https://rdrr.io/cran/ibr/man/ozone.html

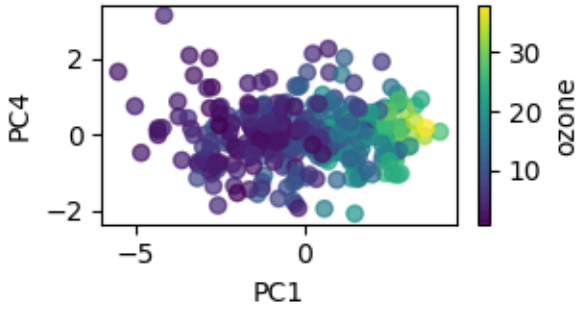
[5] Herbert K. H. Lee ,2004,Bayesian Nonparametrics via Neural Networks, University of California, Santa Cruz, California

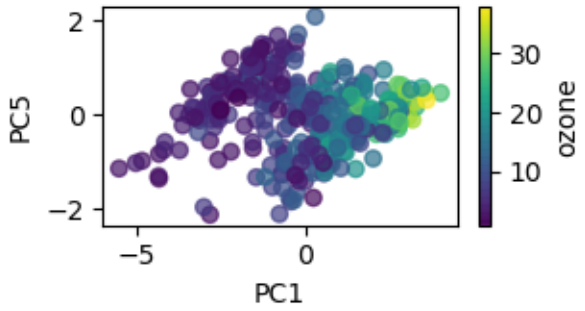
[5] https://medium.com/@rrfd/standardize-or-normalize-examples-in-python-e3f174b65dfc

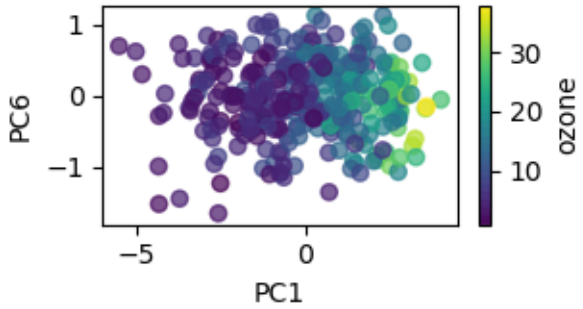
[6] https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

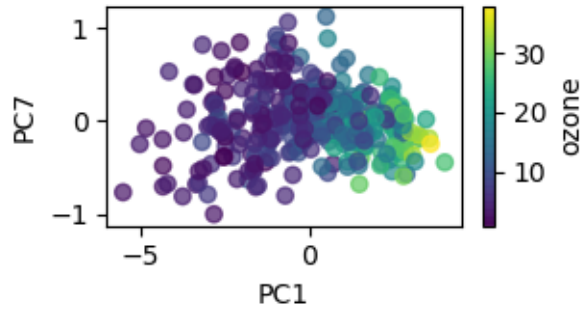
A Appendix A

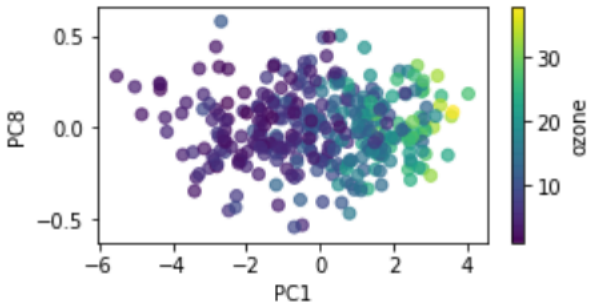


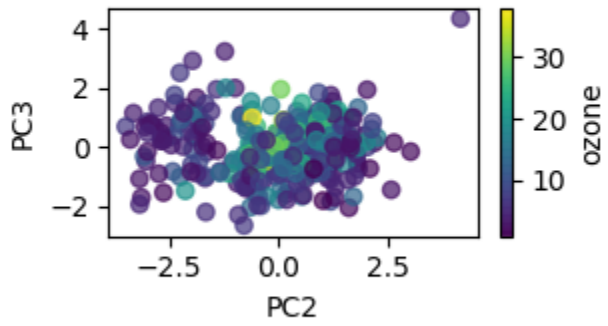


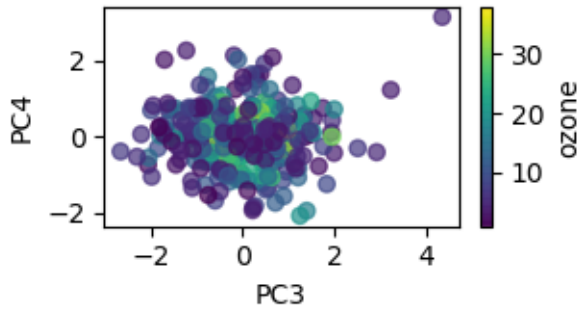

PC5
PC1
ozone
2
0
−2
−5
0
30
20
10

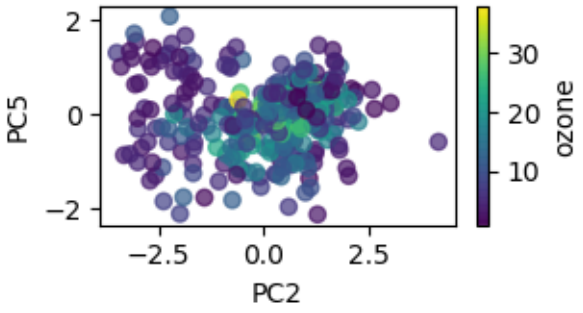

PC6
PC1
ozone
1
0
−1
−5
0
30
20
10

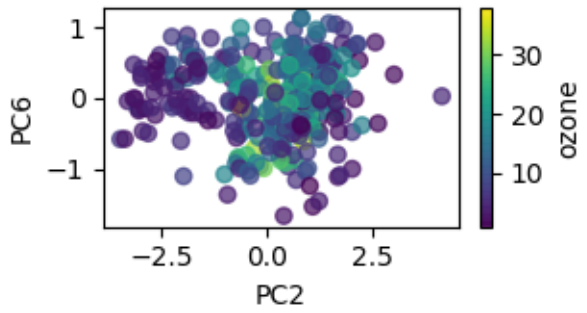

PC7
PC1
ozone
1
0
−1
−5
0
30
20
10

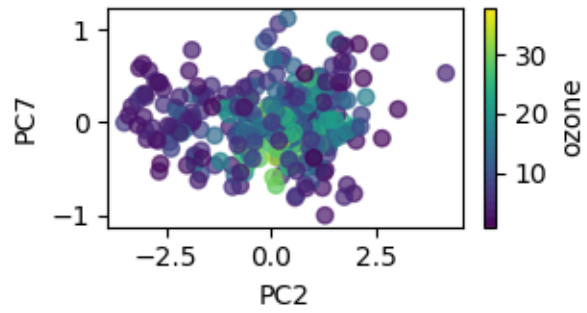

PC8
PC1
ozone
0.5
0.0
−0.5
−6
−4
−2
0
2
4
30
20
10

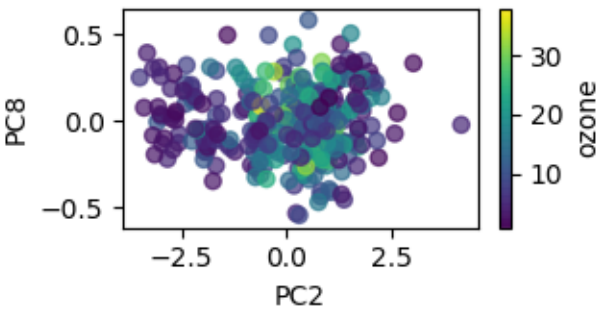

PC3
PC2
ozone
4
2
0
−2
−2.5
0.0
2.5
30
20
10

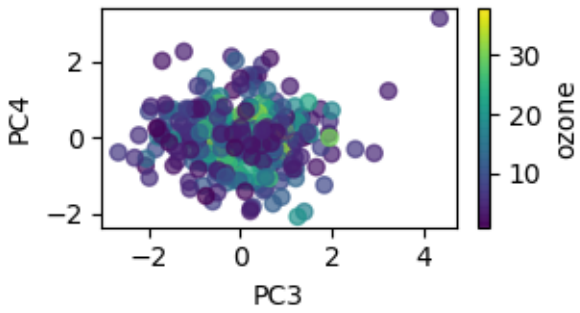

PC4
PC3
ozone
2
0
−2
−2
0
2
4
30
20
10

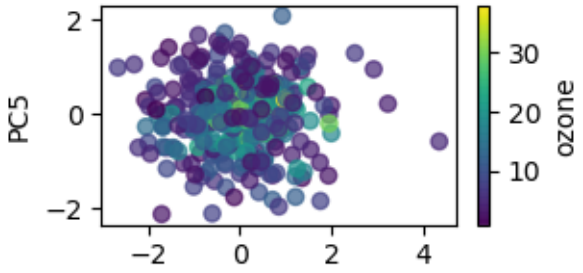

PC5
PC2
ozone
2
0
−2
−2.5
0.0
2.5
30
20
10

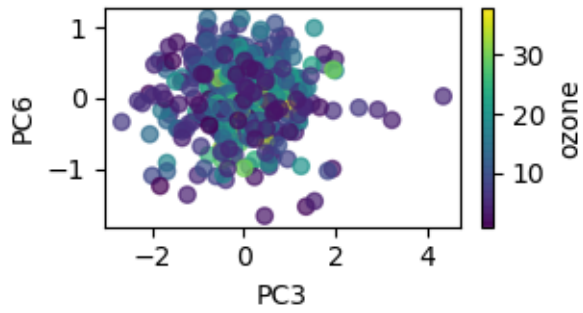

PC6
PC2
ozone
1
0
−1
−2.5
0.0
2.5
30
20
10

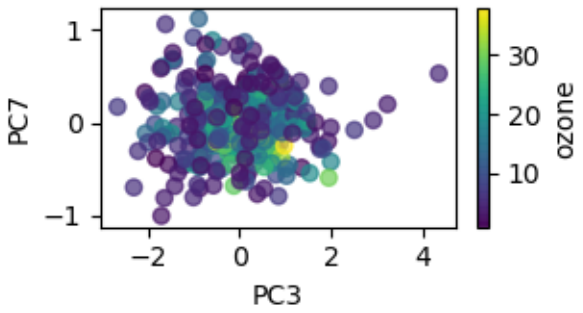

PC7
PC2
ozone
1
0
−1
−2.5
0.0
2.5
30
20
10

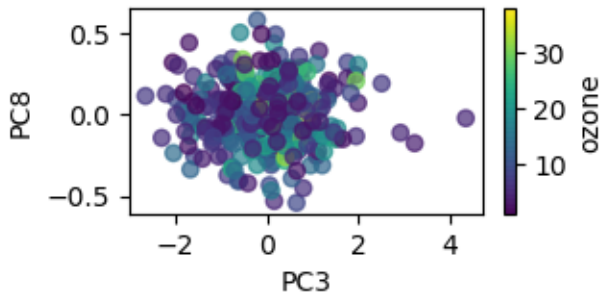

PC8
PC2
ozone
0.5
0.0
−0.5
−2.5
0.0
2.5
30
20
10

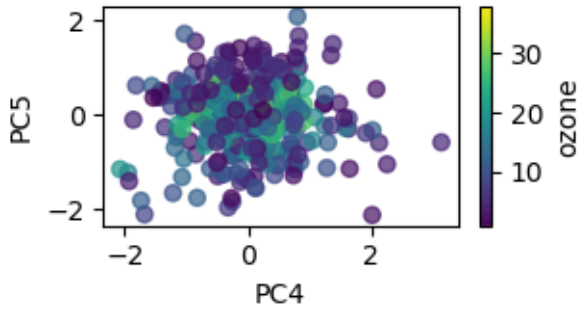

PC4
PC3
ozone
2
0
−2
−2
0
2
4
30
20
10

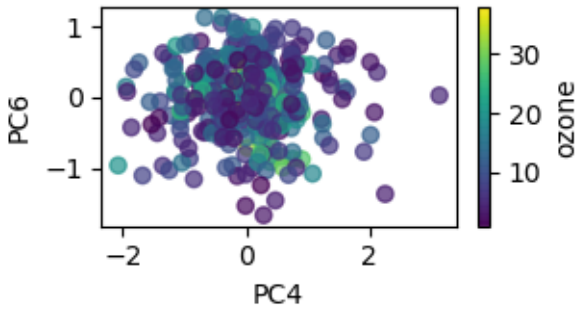

PC5
2
0
−2
−2
0
2
4
ozone
30
20
10

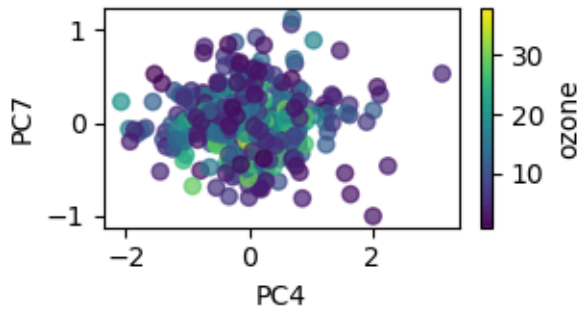

PC6
1
0
−1
−2
0
2
4
PC3
ozone
30
20
10

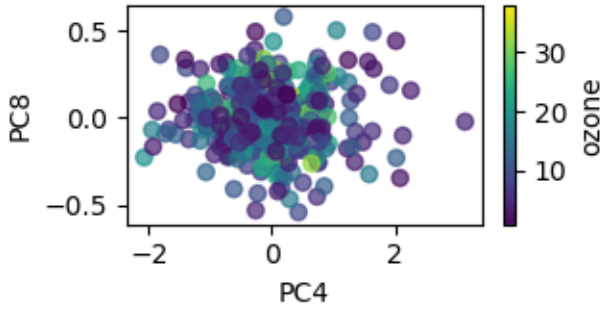

PC7
1
0
−1
−2
0
2
4
PC3
ozone
30
20
10

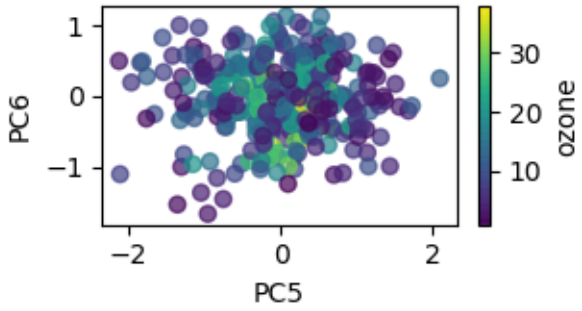

PC8
0.5
0.0
−0.5
−2
0
2
4
PC3
ozone
30
20
10

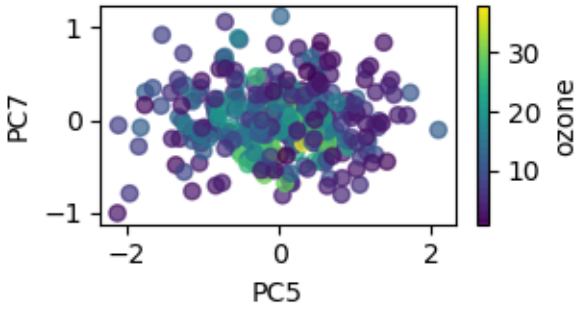

PC5
PC4
ozone
2
0
−2
−2
0
2
30
20
10

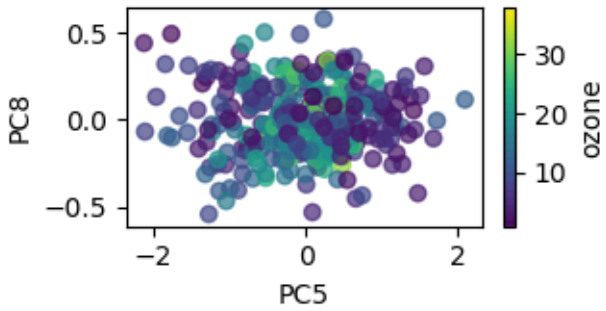

PC6
PC4
ozone
1
0
−1
−2
0
2
30
20
10

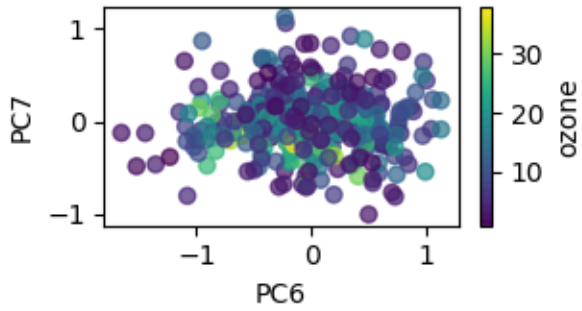

PC7
PC4
ozone
1
0
−1
−2
0
2
30
20
10

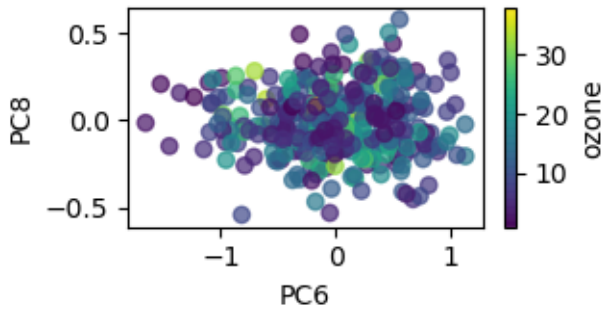

PC8
PC4
ozone
0.5
0.0
−0.5
−2
0
2
30
20
10


PC6
PC5
ozone
1
0
−1
−2
0
2
30
20
10


PC7
PC5
ozone
1
0
−1
−2
0
2
30
20
10


PC8
PC5
ozone
0.5
0.0
−0.5
−2
0
2
30
20
10


PC7
PC6
ozone
1
0
−1
−1
0
1
30
20
10


0.5
0.0
−0.5
PC8
−1
0
1
PC6
30
20
10
ozone