Exploring the Air Pollution Dataset

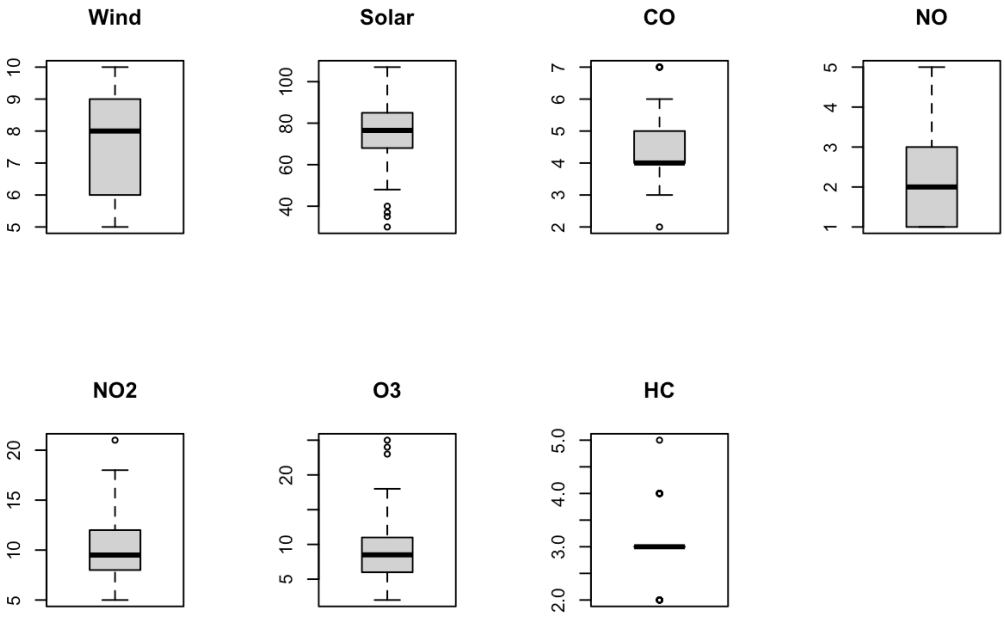
Jeremy Maslanko

Introduction

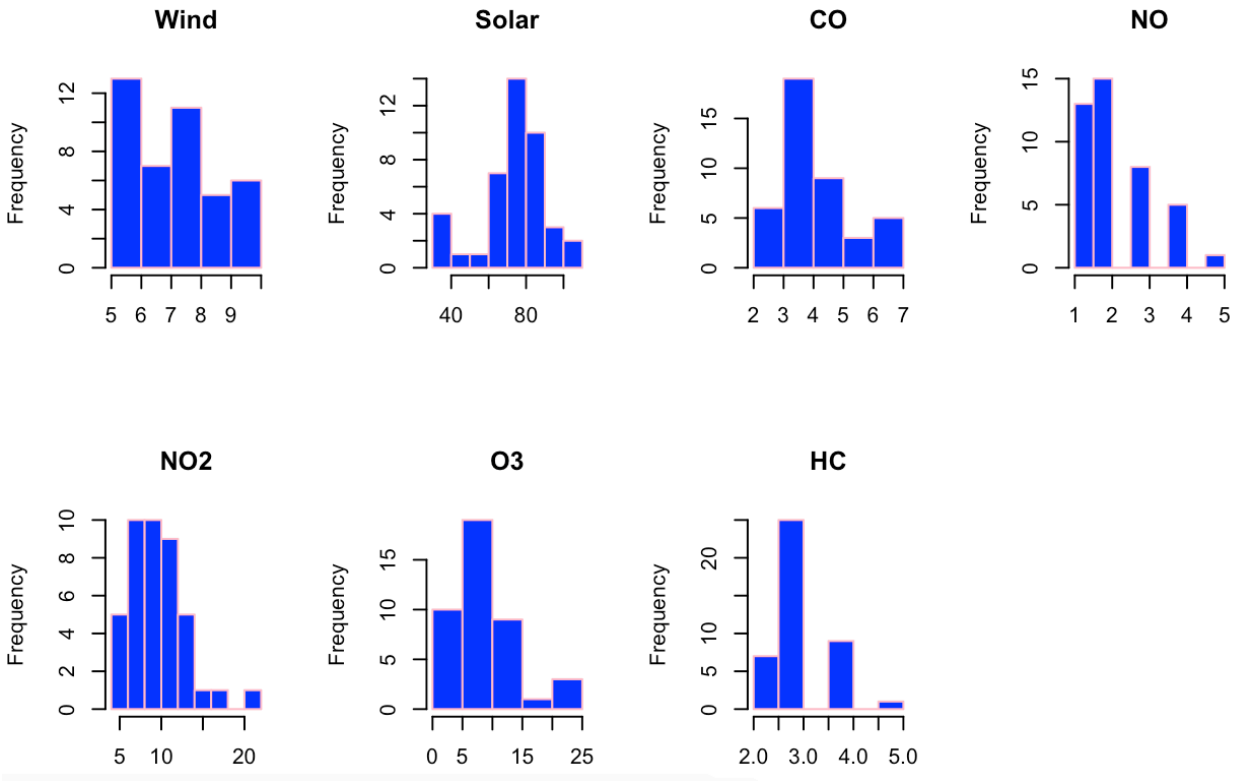
In this paper, we will be exploring the Air Pollution dataset that is provided in JW's book. The data is found in table 1.5 and contains 42 records of data taken at noon in Los Angeles on different days. The datapoints are Wind, Solar Radiation, CO, NO, NO_2, O_3, and HC. Air pollution levels can greatly impact the health of individuals, so having an understanding of this data is crucial. Our goal will be to perform exploratory data analysis that will set us up for further modeling.

Analysis

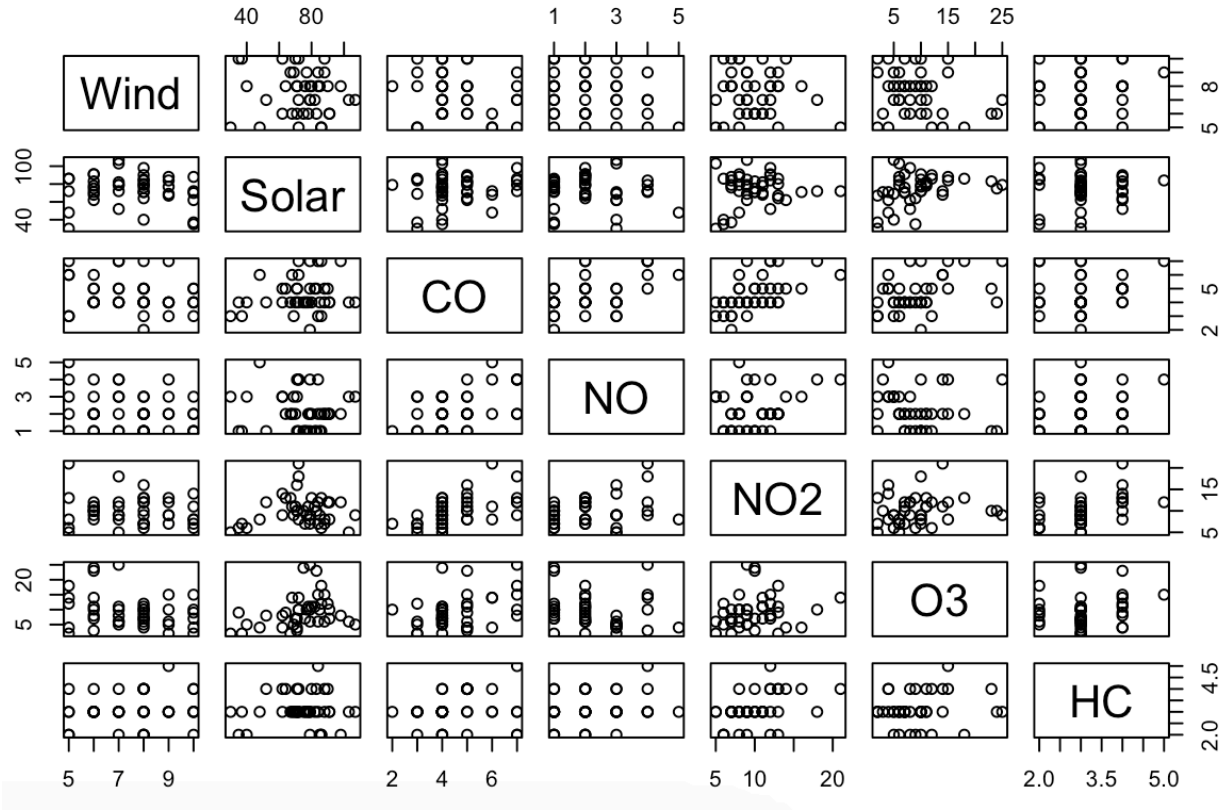
Let us begin the analysis by looking at box plots of the data.



There are a few observations that pop out right away. First, the variable HC appears to have a very tight range of values with just a couple outliers away from the rest of the data. Additionally, Wind has a wide range of values and we can see that NO may be skewed as the median appears to be further away from the rest of the range of data points. Looking at the histograms of the variables below, we are able to confirm what we noticed in the box plots.



Most of the variables appear to be showing some skew, although Solar looks the most normal. We will check these variables for normality, however it would be beneficial to first look at how correlated they are. We will start by looking the scatterplots for all of the variables, and then turn our attention to a table with the correlation coefficients.

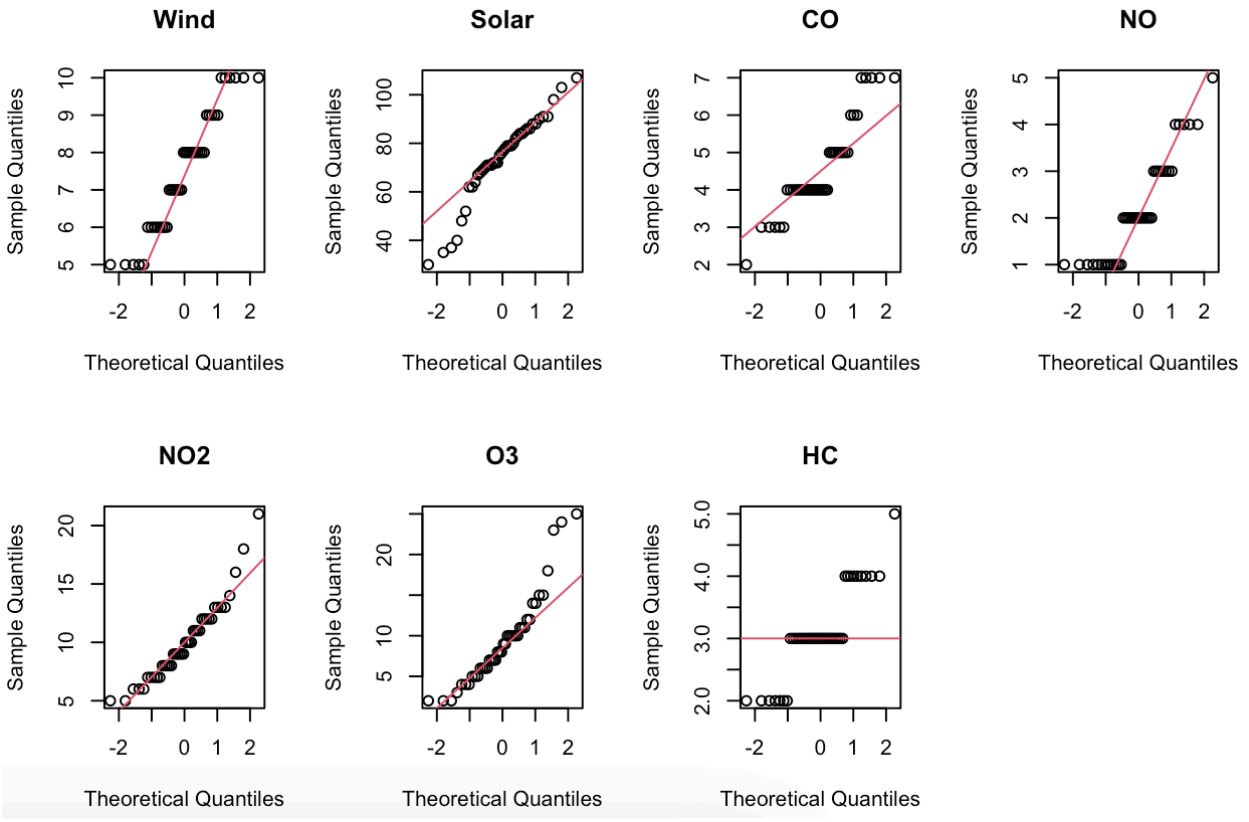


Just looking at the scatterplots, there are no strong relationships that can be seen. However, we do see that some of the variables appear to be discrete data. Lets now look at the correlation coefficients.

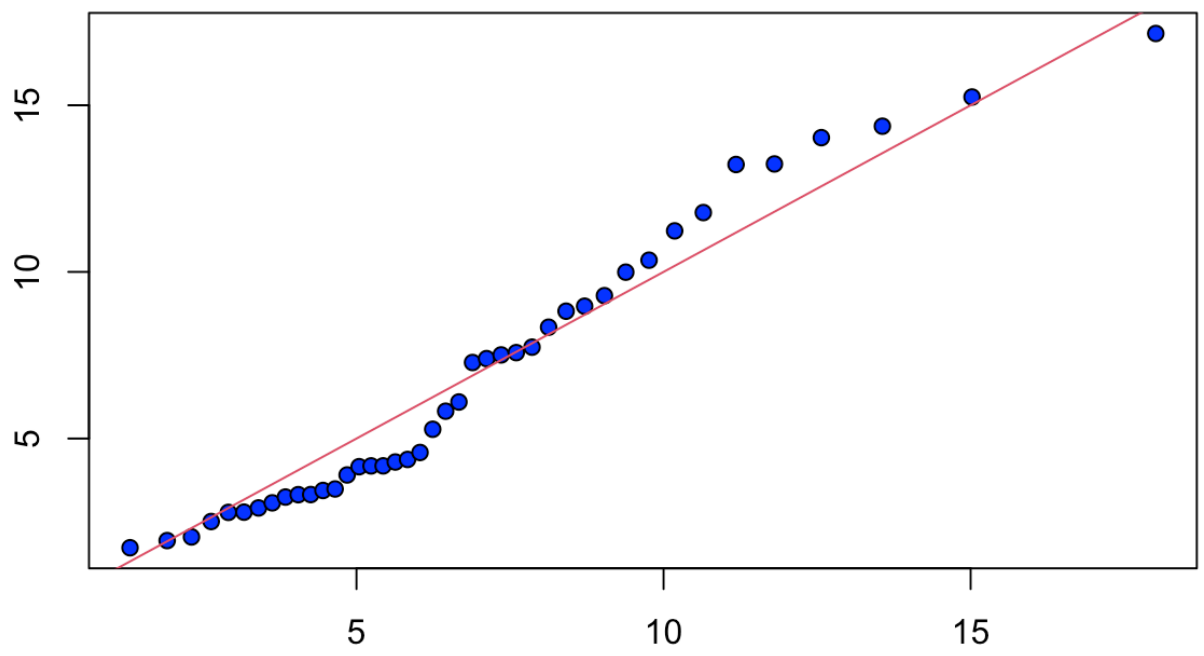
	Wind	Solar	CO	NO	NO2	O3	HC
Wind	1.0000000	-0.10144191	-0.1938032	-0.26954261	-0.1098249	-0.2535928	0.15609793
Solar	-0.1014419	1.00000000	0.1827934	-0.07356907	0.1157320	0.3191237	0.05201044
CO	-0.1938032	0.18279338	1.0000000	0.50215246	0.5565838	0.4109288	0.16603235
NO	-0.2695426	-0.07356907	0.5021525	1.00000000	0.2968981	-0.1339521	0.23470432
NO2	-0.1098249	0.11573199	0.5565838	0.29689814	1.0000000	0.1666422	0.44776780
O3	-0.2535928	0.31912373	0.4109288	-0.13395214	0.1666422	1.0000000	0.15445056
HC	0.1560979	0.05201044	0.1660323	0.23470432	0.4477678	0.1544506	1.00000000

A few of the fields show closer to a medium relationship, such as NO_2 and CO as well as NO and CO. However, like we saw with the scatterplots, most of the variables show a weak relationship.

Having looked at the relationship among variables, let's now return to assessing the normality of the data.



We again see that some of the variables are discrete, but we are able to learn more about the continuous variables. All three of the continuous variables show some outliers that are impacting their normality. From the above plots, NO_2 is the most normal, with Solar and O_3 showing heavier skew at the tails. With our data being multivariate and not univariate, it is important to check normality using a ChiSquare test. The plot for this can be seen below.



Just looking at this, it is sort of difficult to confidently say whether or not our data is normally distributed. It would be best to test this with the Shapiro-Wilk test for normality. When we do this, we get a p-value of 0.000272. This number is less than an alpha value of 0.05, leading us to reject the null hypothesis that our data is normally distributed.

Conclusion

To summarize, we analyze the air pollution data set that contained 42 observations recorded on different days at noon in Los Angeles. We saw that HC had a tight range of values, which was confirmed from the scatterplots showing that the data was discrete. We then checked the relationships of the variables, and found that most of them had weak relationships, with the exception of NO_2 and CO as well as NO and CO. Lastly, we checked the normality of the data. While some of the variables showed

signs of being normal at the univariate level, we concluded form the Shapiro-Wilk test that the data as a whole was not normal under the multivariate case.