Pollution and Mortality

Dvir Blander, Gabriel Felix-Martinez, Noah Kimball-Dembitzer, Naveen Thiruvazhi

3/24/2020

This dataset is used to look at air pollution and mortality, by using population, rainfall and mortality rates from cities across America. There are 60 cities from which the data was collected and all of the 17 varaibles other than the city names are continuous and numeric. Our question of interest is to see how air pollution affects mortality rate. This would make air pollution the independent variable and mortality rate the dependent variable. The data was collected from the U.S. Department of Labor Statistics. The variables are as follows: 1.city: City name 2.JanTemp: Mean January temperature (degrees Farenheit) 3.JulyTemp: Mean July temperature (degrees Farenheit) 4.RelHum: Relative Humidity 5.Rain: Annual rainfall (inches) 6.Mortality: Age adjusted mortality 7.Education: Median education 8.PopDensity: Population density 9.%NonWhite: Percentage of non whites 10.%WC: Percentage of white collar workers 11.pop: Population 12.pop/house: Population per household 13.income: Median income 14.HCPot: HC pollution potential 15.NOxPot: Nitrous Oxide pollution potential 16.SO2Pot: Sulfur Dioxide pollution potential 17.NOx: Nitrous Oxide

Let's now look at the question of interest, which are labeled below: 1. How much does pollution affects mortality? Pollution is the independent variable and mortality would be dependent.

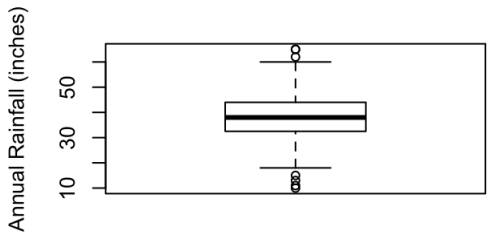
Note: We are only working with one response to focus our analysis as specified in the IE Step 3 comments.

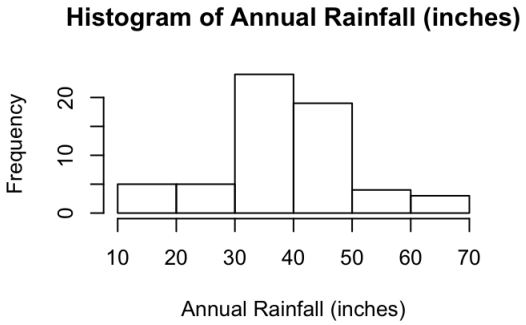
Descriptive Statistics Table for our variables of interest.

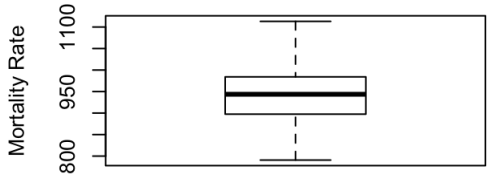
	Rain	Mort	Pop	HC PP	NO PP	SO PP	NO
Min	10	790.7	124833	1	1	1	1
Max	65	1113	8274961	648	319	278	319
Median	38	943.7	914427	14.5	9	30	9
Mean	38.38	940.3	1438037	37.85	22.6	53.77	22.6
SD	11.52	62.22	1541736	91.98	46.36	63.39	46.36

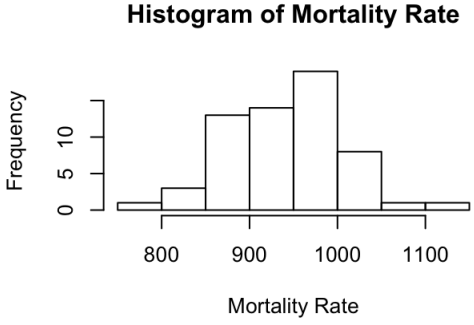
Note that PP is Pollution Potential.

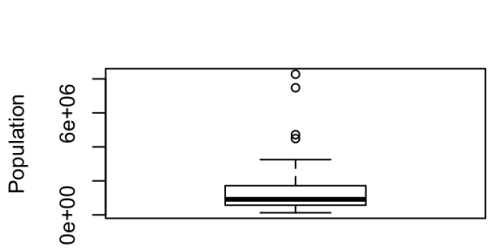
Plotting the Boxplots and Histograms for all variables.

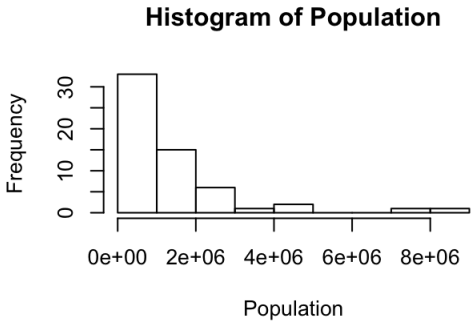


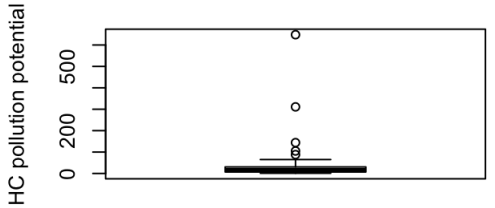


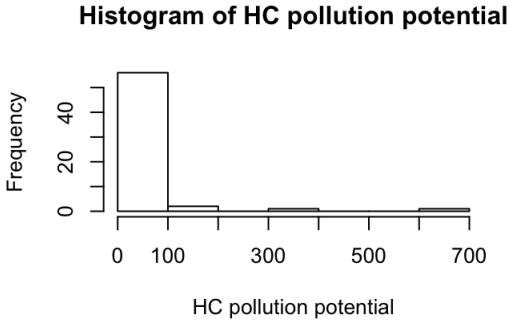


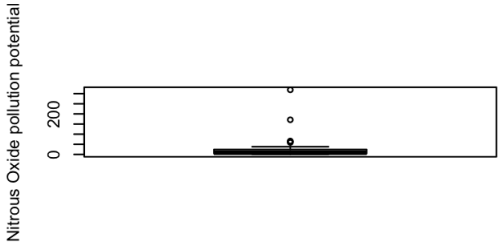


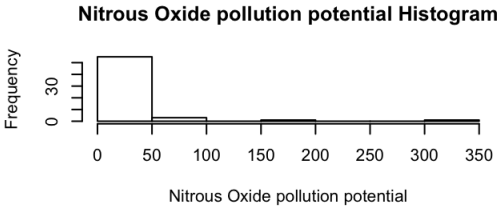


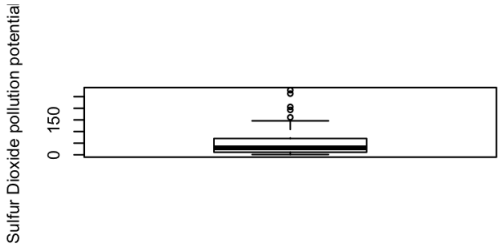


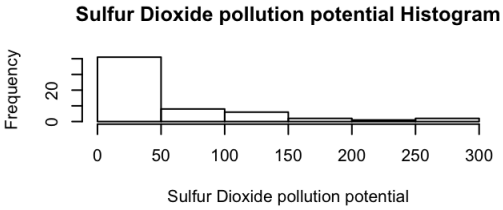


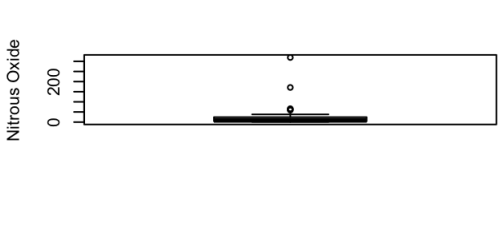


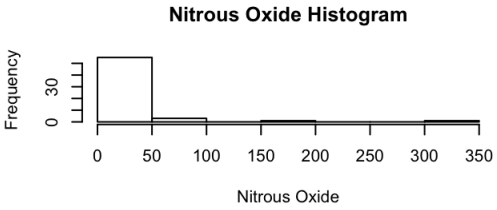




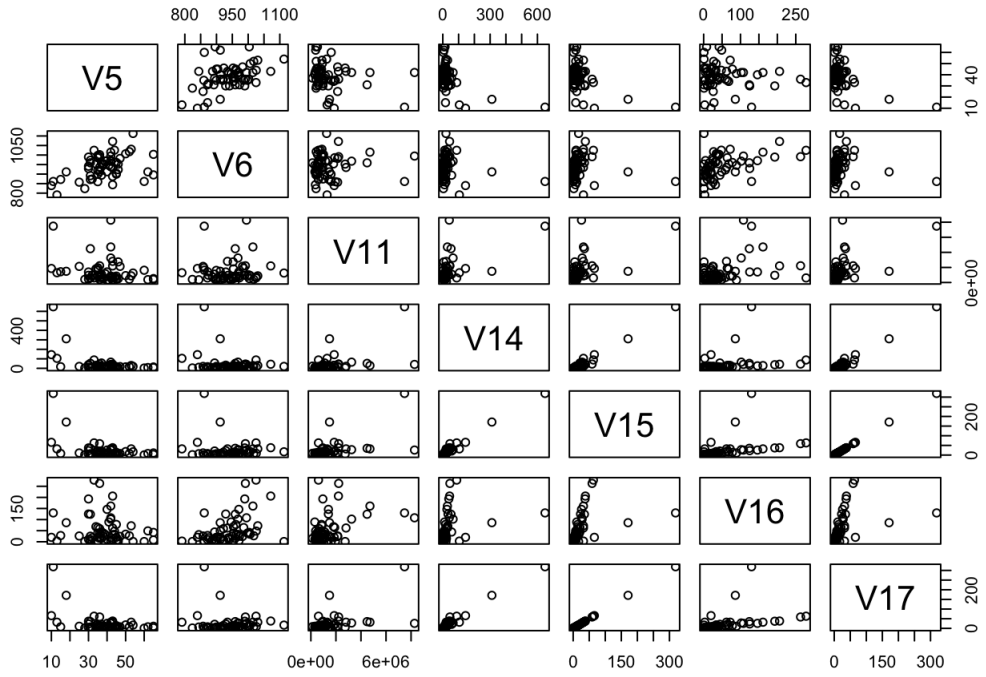








Constructing the scatter plots in addition to each correlation coefficient for each variable of interest.



	V5	V6	V11	V14	V15	V16	V17
V5	1	0.4379	NA	-0.4877	-0.452	-0.1205	-0.452
V6	0.4379	1	NA	-0.1782	-0.07769	0.4256	-0.07769
V11	NA	NA	1	NA	NA	NA	NA
V14	-0.4877	-0.1782	NA	1	0.9838	0.2823	0.9838
V15	-0.452	-0.07769	NA	0.9838	1	0.4098	1
V16	-0.1205	0.4256	NA	0.2823	0.4098	1	0.4098
V17	-0.452	-0.07769	NA	0.9838	1	0.4098	1

Note that the V11 is the population variable, which would not correlate with the other variables because it is not linear.

There seems to be a positive linear correlation between rainfall and mortality. There does not seem to be a relationship between rainfall and population. There seems to be a negative correlation between rainfall and pollution. However, this does not look linear and may simply be due to outliers. There does not seem to be a strong relationship between mortality and pollution. If we remove potential outiers, there may not be a relationship that is needed to be examined. There might be a slight increasing trend for Population and Pollution. This may need to be examined further.

Looking at outliers There may be an outlier with population and HC pollution potential, Nitrous Oxide pollution potential, and Nitrous Oxide. This can be seen by looking at the population scatter plots with these variables.

Let's now remove these outliers to make better models.

Questions of Interests

Let's first look at pollution as a whole and mortality specifically. Since there are many pollution variables, we will conduct a multiple linear-regression with multiple pollution predictors with the addition of rainfall and population.

The population predictor must be included because there is a small correlation between population and pollution as described above. There is also a correlation of the NOxPot with with HCPot so this also must be included in the model. As we can see, the SO2Pot predictor is not a significant predictor for this multiple linear regression due to the high p-value of 0.1154, so we will collectively decide not to remove this predictor from the data due to the adjusted R-squared is higher when this predictor is in the linear regression model than without.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	837	29.05	28.82	1.213e-33
NOxPot	2.044	1.236	1.654	0.1042
SO2Pot	0.271	0.1707	1.588	0.1184
HCpot	-1.07	0.6064	-1.764	0.08355

	Estimate	Std. Error	t value	Pr(>|t|)
rainfall	2.114	0.6588	3.209	0.002282
population	1.287e-06	6.991e-06	0.1841	0.8546

Fitting linear model: mortality ~ NOxPot + SO2Pot + NOx + HCpot + rainfall + population

Observations	Residual Std. Error	R^2	Adjusted R^2
58	48.61	0.4492	0.3962

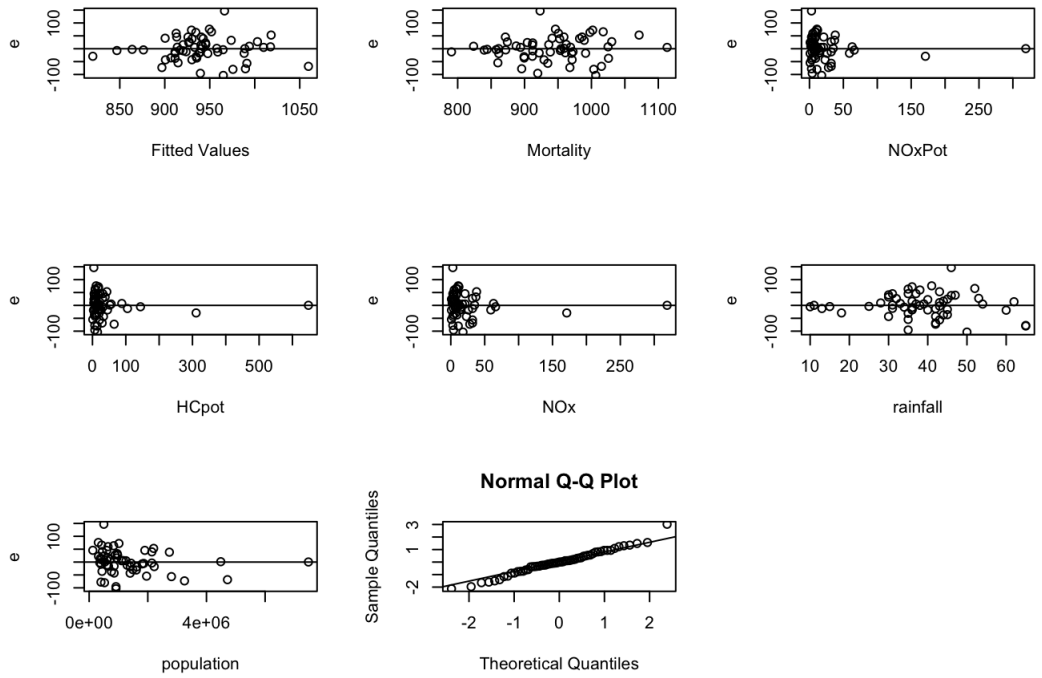
Now let's fit the data without the removed predictor.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	850.5	28.16	30.2	4.462e-35
NOxPot	3.552	0.8015	4.432	4.716e-05
HCpot	-1.795	0.4043	-4.441	4.589e-05
rainfall	1.902	0.6543	2.907	0.005316
population	3.164e-06	6.988e-06	0.4528	0.6525

Fitting linear model: mortality ~ NOxPot + NOx + HCpot + rainfall + population All the values above are statistically significant, except for population but this was discussed earlier, due to the high t-values. Note that R is conducting two-sided t-tests. The R squared is 0.4225 , which indicates that 42.25 % of the variation in the mortality rate is explained by the predictors.

Observations	Residual Std. Error	R^2	Adjusted R^2
58	49.3	0.4225	0.3789

Let's plot the residuals for the above linear regression to check for normality and constant variance.



The variability in the data seems consistent for all plots with the bulk of the residuals being in between -100 and 100 with near constant variance. The residuals are approximately normal due to most of the data being on or very near the 45 degree line. Also, the data seems to be equally above and below the residuals = 0 line.

Conclusion

In conclusion, the mortality rate can be predicted from 7 different predictors (NOxPot, NOx, HCpot,rainfall, population, SO2pot, and the constant). The R squared value of 0.4492 shows that 44.92% of the variability in the mortaility rate is the result of the predictors.

The covariate of Sulfur Dioxide pollution potential was not used due to that the predictor was not statistically significant in its use. Also, as a group, we made the conscious decision to include the predictors that we included for the model. 16 predictors would make the model too hard to interpret so we believe that we struck the right balance with the predctors that we chose.

Let's take a look at the estimated regression function based on the analysis.

$$\hat{\text{mortality}}_i = 837 + 2.044\text{NOxPot}_i + 0.271\text{SO2Pot}_i - 1.07\text{HCpot}_i + 2.114\text{rainfall}_i + 1.287e-06\text{population}_i \text{ for } i = 1, \ldots, n.$$

Thus, it seems that pollution heavily affects the mortaility rate because if we increase all the pollution indicators by 1 and hold all the other predictors constant, the moratility age adjusted mortality increases on average by 1.245.

What I learned

I have personally learned that the problem of itnerest was more complicated than expected. I thought that there would be only only pollution indicator in the final model. However, there were multiple pollution indicators and this complicated the model for the group to decide what indicators to include and what indicators to remove. Thankfully, there were many statistical tests in this class that taught the group how to do this and perform other anlysis that were necessary for the project.

In terms of working as a team, I learned a lot about myself. I learned how to delegate tasks and work with others in learning the software. I had the most familiarity with R in my group and I thus naturally became the leader because I could do everything the fastest in R. I learned mostly how to take constructive criticism and apply it to make the group's work better. This is a crucially important skill to develop because [illegible] and applying these twinks can drastically improve one's work.