STAT 448 Final Project

Lisa Silverstein, Jessica Fornek, Jace Rice

Introduction

Air pollution poses a significant global environmental issue, adversely affecting the health of ecosystems and the overall well-being of humans. Various locations across the world measure key pollutants, which include PM2.5 (particulate matter with a diameter of 2.5 μm or less), PM10 (particulate matter with a diameter of 10 μm or less), SO2 (sulfur dioxide), NO2 (nitrogen dioxide), CO (carbon monoxide), and O3 (ozone). These measurements serve as the basis for calculating the air quality index for various cities globally. Additionally, meteorological variables such as precipitation, temperature, dew point, air pressure, wind speed, and wind direction play an influential role in the dispersion or concentration of these pollutants.

The data set we used for our project was the Beijing Multi-Site Air-Quality Data which was published in the UCI Machine Learning Repository. The UCI Machine Learning Repository site notes that "This data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017." The data set includes 420768 instances, though it's important to note that there are some missing values which are denoted by "NA."

For our project , there were a handful of analyses to complete, all centering around the idea of observing how air pollutants act within our environment. Our report will cover all of the analyses we completed and the findings we derived from them.

Methods

Descriptive Overview of the Data

The first part of our project assigned us with providing a descriptive overview of the meteorological and air pollutant variables. The initial overview of the data was done to be a general overview of the variables so that we could see what the dataset contained. We did this by using *proc univariate* and *proc means* to see what the variables would look like when tests of

normality were supported against the variables and when they were not. However, we decided that the general overview was not enough and we wanted to look at specific relationships between the continuous and categorical variables. To do this, we ran each categorical variable against continuous variables that we believed would have the greatest impact and identifiable relationship between air pollutants and meteorological variables. We chose month, year, and wind direction and used *proc reg* and *proc anova* procedures to look for the relationships between these variables.

Check For Correlation

The second part of the project our group was tasked with was finding a potential correlation between the air pollutants and meteorological variables. Our initial thought process was to find any significant interactions between the variables and then proceed to check for multicollinearity to see if there could be any relationship that would not have shown up in the initial correlation test. These two procedures were performed for each air pollutant and meteorological variable.

Principal Component Analysis

The third part of our project tasked us with finding the most representative air pollutant by performing a Principle Component Analysis (PCA) using the air pollutant variables and by retaining the principal components that explain most of the variation. First, we standardized the air pollutant variables with a proc standard statement. Standardizing the data provides us with more meaningful and interpretable results, and helps us uncover results which otherwise would have been obscured by differences in the scales of the variables. Next, we went in with a proc princomp statement to perform the PCA. We listed all the air pollutant variables to be included in the principle component analysis and requested a scree plot to be produced upon running the code. From our output, we could decide how many principal components to keep and discover which principle component explains the most variation in the data.

Comparing Air Pollutant Concentrations by Month

The next part of our project focused on determining whether there are statistically significant differences of the mean air pollutant concentrations for the different months of the year. To

normalize the data for the eventual ANOVA, we log-transformed each air pollutant variable. Then, for each individual variable, we set up a *proc glm* procedure that would return class levels, number of observations, overall ANOVA, fit statistics, Type I and III model ANOVA, Levene's HoV test, and a distribution of the specific air pollutant by month.

Modeling Air Pollutant Concentrations Against Meteorological Variables

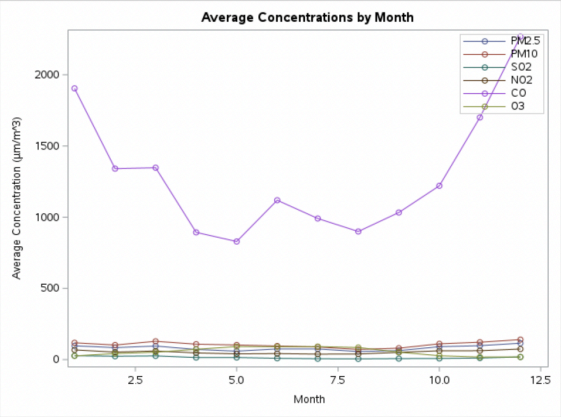
The next step of our analysis entailed fitting a regression model to the first principal component of all pollutants as a function of the meteorological variables. Similarly to before, we standardized the air pollutant variables with a *proc standard* statement. Then, we performed PCA analysis using the standardized pollutant variables. After getting a result for the first principal component, we modeled it against all the meteorological variables. This regression model will help us understand how the meteorological variables contribute to the variation in the first principal component, providing insights into the relationship between air pollutants and meteorological conditions. However, we noticed some diagnostics issues with this model. Once those issues were remedied (see Appendix B for details), we proceeded with stepwise selection to get our final model.

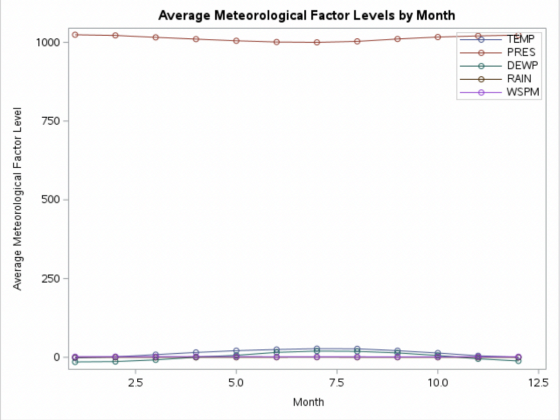
Further Analysis: Generalized Linear Model

We were also curious about how Poisson and Gamma models might fit the data. To find out, we used the generalized linear model procedure and compared how the Poisson model fit the data vs the Gamma model. We also analyzed some of the residual diagnostics for each.

Conclusions

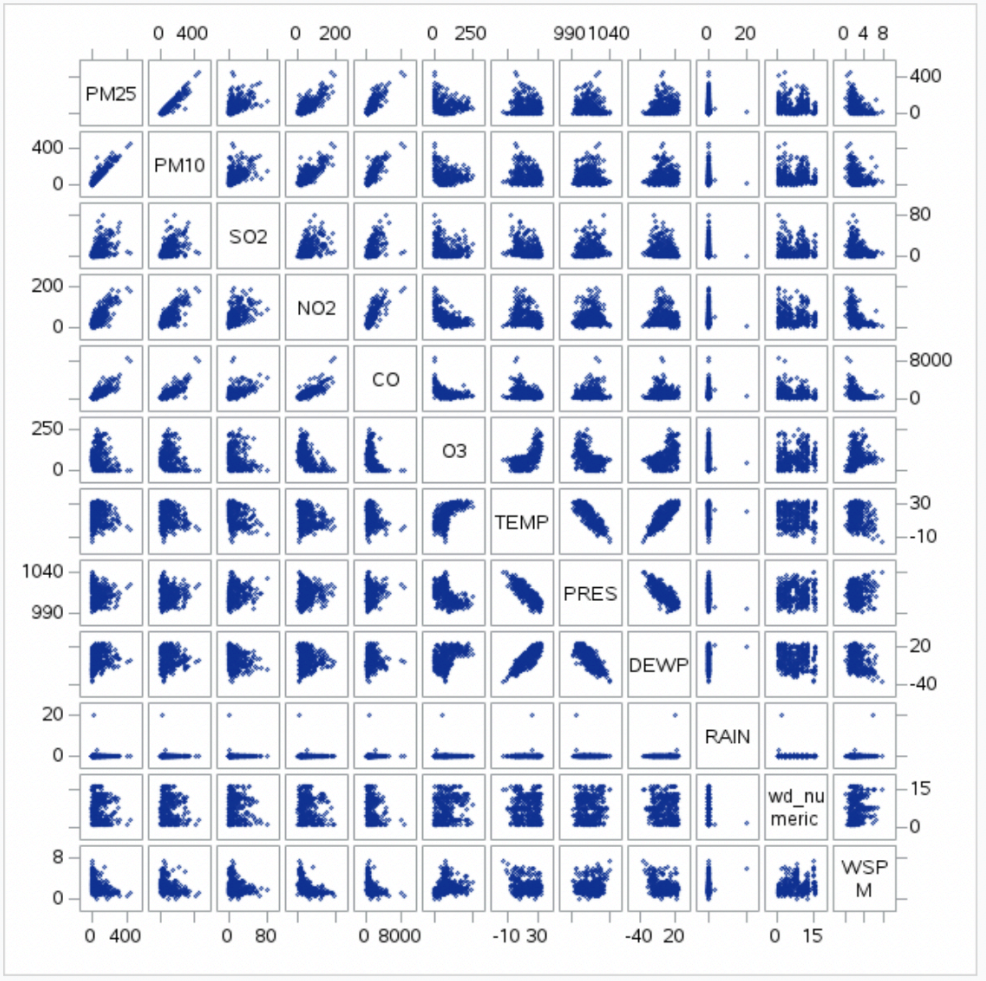
Descriptive Overview of the Data





For the first part of our group's analysis, we concluded that, of the categorical variables, month seemed to be the most significant in terms of interactions with the other meteorological variables. The data does not appear to be normally distributed so we took a deeper look into the data set. After performing the tests for each we found that the F-value returned a highly significant p-value for almost every test. However, each test for the meteorological variables against the month returned a relatively small R-squared value which meant that no one variable was explaining the pollution concentration of that month. This could be because of the large volume of observations within the data set. The largest R-squared value returned from our tests was 0.0816 or 8.16% of the data is explained by the month. As stated earlier, this is small but considering the large size of the data set, over 32 thousand observations, this is actually quite a significant number. These steps gave us a good understanding of not only what is inside the dataset but how it reacts when put up against tests of normality and when checked against each other which ultimately ended up being useful for analysis later in the project.

Correlation Between Variables



The next task we were assigned was to look for possible correlations between the air pollutants and meteorological variables. When looking at the Pearson Correlation Coefficients (see Appendix C), we can see that there is either a negative or positive correlation and that each is highly significant under a hypothesis test between the two variable sets. In fact a majority of the relationships were negatively correlated but still had a correlated relationship nevertheless. We also looked at significant interactions between the two variables set and checked for multicollinearity and found that for each model, there was a highly significant relationship between each air pollutant variable and meteorological variable.The strongest positive relationship was shown to be between O3 and WSMP with a Pearson Correlation Coefficient of 0.33317. The strongest negative relationship was shown to be between WSPM and NO2 with a Pearson Correlation Coefficient of -0.41396. We defined both of these as a medium level of

correlation strength while the other relationships were smaller and weaker. This led us to believe that, in these two cases, when one variable is present, so is other one.

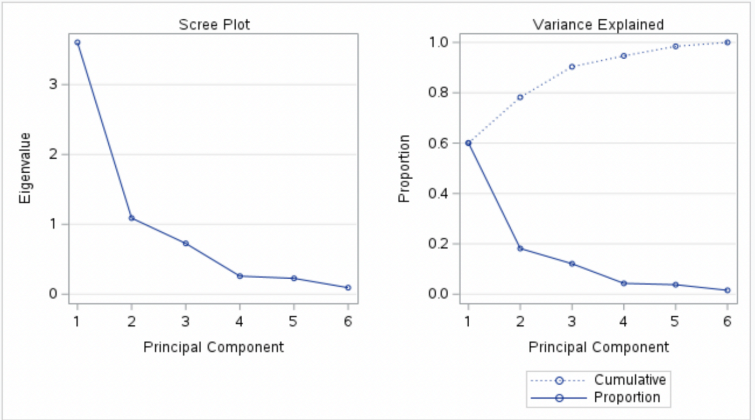
Principal Component 1 Explains Most of the Variation in the Data

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.60375880	2.51437144	0.6006	0.6006
2	1.08938736	0.36232297	0.1816	0.7822
3	0.72706440	0.46813113	0.1212	0.9034
4	0.25893327	0.03178784	0.0432	0.9465
5	0.22714543	0.13343470	0.0379	0.9844
6	0.09371073		0.0156	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
PM25	0.478451	0.210072	-.283454	-.156608	-.171097	-.769940
PM10	0.459463	0.258028	-.288943	-.514390	-.134424	0.596792
SO2	0.295323	0.307512	0.890836	-.063561	-.142614	-.015921
NO2	0.453949	-.292170	0.082911	-.018615	0.837399	-.010451
CO	0.471331	0.001067	-.130016	0.813259	-.221376	0.224823
O3	-.211047	0.842237	-.137110	0.212370	0.426700	0.011109



From the third part of our analysis for this project, we concluded that the first principal component explains the most variation in the data. The first principal component explained 60.06% of the total variance in the data and by adding the second principal component, 78.22% of the variance in the data is explained. Additionally, from our scree plot, we see the “elbow” of the graph at around 2 principal components, meaning that we likely want to select the first two principal components in this data to explain the variance in air pollutant concentrations.

Specifically looking at Prin1 in the eigenvector table created, we see that O3 is the only pollutant that has a negative contribution. The other five air pollutants display positive contributions to Prin1. We see that, compared to the other eigenvectors, Prin1 best captures a common pattern or underlying source that influences multiple air pollutants simultaneously. The positive contributions of PM2.5, PM10, SO2, NO2, and CO to Prin1 suggest that higher concentrations of

these pollutants tend to occur together. When Prin1 increases, it indicates a higher overall level of all these pollutants.

We also see in the correlation matrix (see Appendix C) that O3 is the only air pollutant that displays negative correlation to any other air pollutants. All other air pollutants have positive correlation values, meaning that as one air pollutant increases, the other does too. O3 exhibits an inverse relationship with all the other variables, so when any other air pollutant increases, O3 will decrease instead. It is possible that because O3 is the only non-man-made air pollutant from this data set, it may act differently than the other man-made pollutants.

Difference in Mean Air Pollutant Concentrations for Different Months

Level of month	N	PM10		SO2		NO2		CO		O3	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
1	2877	117.997567	111.030998	27.4258603	27.2740514	68.0238095	37.4122228	1904.51860	1645.52553	25.8593326	24.0575715
2	2486	102.360418	105.356253	23.5321802	29.4392183	52.6275141	33.0776118	1340.82864	1182.86320	42.3793242	30.1630132
3	2814	128.520469	102.280346	26.0394456	28.4247444	60.4687278	34.4864964	1347.40156	1060.77618	52.1958067	39.4814914
4	2735	108.388044	70.513522	14.4727605	16.1275517	47.1047861	26.1079916	893.32431	583.22407	72.3903568	56.2761891
5	2422	102.566061	68.684569	14.7968621	19.3686604	40.1159785	24.7010430	829.37077	549.94001	91.2476186	67.4718386
6	2730	95.492747	68.195928	9.1278388	10.9134371	42.1219780	22.8906331	1119.19048	977.16454	90.6835165	71.6251453
7	2841	88.510384	55.257738	5.9992960	6.3726970	38.3632524	21.0284610	990.87786	469.36067	91.3717001	73.0068670
8	2872	71.436943	47.549933	5.0702646	5.0081898	39.1718663	20.7935061	899.58217	460.21798	85.0164345	70.3282334
9	2748	80.224527	56.494703	6.9647016	8.6839834	50.9498180	23.6199422	1032.89338	604.69422	52.0701237	54.7548535
10	2867	110.892222	86.130736	7.9487269	11.5170837	61.8440879	30.7490649	1220.50924	809.51125	27.0958297	36.1145966
11	2820	121.907482	101.000055	10.9929078	13.4243094	62.3293262	33.7051382	1700.31915	1323.19468	17.8919728	20.6793556
12	2631	140.323793	125.921592	18.5748385	20.5728010	74.2755226	40.8604194	2268.68111	2011.63718	18.2256545	20.4338226

The fourth task in our analysis was to determine whether we have statistically significant differences of the mean air pollutant concentrations for the different months of the year. When running the ANOVA models for all the air pollutants, we found that all of them display a significant difference in air pollutant levels across various months (see Appendix A for F-values, p-values, and boxplots). For all pollutants, the overall model p-values were <0.001, which is significantly below the typical threshold of 0.05, so we can reject the null hypothesis that various air pollutant concentrations do not vary by month. The majority of the air pollutants had their highest concentrations in December. PM10, PM2.5, NO2, and CO were all highest in December with concentrations of 140.32, 115.06, 74.28, 2268.68, respectively. Deviating from the December norm, SO2 and O3 reported highest concentrations in January and July, respectively.

In January, SO2 levels were found to be 27.43 and in July, O3 levels were at 91.37. Distribution graphs for all of the air pollutants were also generated and they displayed very clear patterns in air pollutant concentration over the course of a year.

Final Regression Model For Predicting Air Pollutant Concentration Variability

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	40240	6706.67809	2785.72	<.0001
Error	32824	79024	2.40752		
Corrected Total	32830	119265			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	39.18417	1.61887	1410.48442	585.87	<.0001
TEMP	-0.14470	0.00181	15383	6389.58	<.0001
PRES	-0.03635	0.00159	1264.61363	525.28	<.0001
DEWP	0.07814	0.00139	7642.27011	3174.33	<.0001
RAIN	-0.23974	0.01530	591.21792	245.57	<.0001
wd_numeric	0.01016	0.00195	65.25409	27.10	<.0001
WSPM	-0.35552	0.00818	4548.31822	1889.21	<.0001

We saw in the Principal Component Analysis step that the first principal component explains most of the variation in the data. Therefore, the first principal component can be used as a sort of smaller subset of data that is easier to interpret. When modeling the first principal component against all the meteorological variables, our final model included all meteorological variables as predictors. The F-value for the overall ANOVA was 2785.7 and the p-value was <0.00l. This shows that this model, with all meteorological variables, is better than the intercept-only model in explaining the variation in the first principal component. The p-values of the individual parameter estimates are all <0.001. Therefore, every meteorological variable is significant in this model, and should be kept in the model to obtain the most accurate results for predicting the first principal component. Using the first principal component as a response variable serves as a way to predict all air pollutant concentrations from all meteorological variables.

The parameter estimate for TEMP tells us that for every one-unit change in temperature, the expected Prin1 will decrease by 0.145. The parameter estimate for PRES tells us that for every one-unit change in pressure, the expected Prin1 will decrease by 0.036. The parameter estimate for DEWP tells us that for every one-unit change in dew point temperature, the expected Prin1

will increase by 0.078. The parameter estimate for RAIN tells us that for every one-unit change in precipitation, the expected Prin1 will decrease by 0.240. The parameter estimate for WD tells us that for every one-unit change in wind direction, the expected Prin1 will increase by 0.010. The parameter estimate for WSPM tells us that for every one-unit change in wind speed, the expected Prin1 will decrease by 0.356. Higher principal components are characterized by having values in the original variables that contribute more to the pattern represented by that principal component. Based on these individual parameter estimates, higher temperatures, pressure, precipitation, and wind speeds tend to decrease the first principal component and are likely associated with higher variation in air pollutant concentrations. On the other hand, higher dew point temperature and higher wind direction (i.e. wind going in the south and west directions) tend to increase the first principal component and are likely associated with less variation in air pollutant concentrations.

Further Analysis: Generalized Linear Model

When we used *proc genmod* to see how Poisson and Gamma distributions fit the data, these were the results:

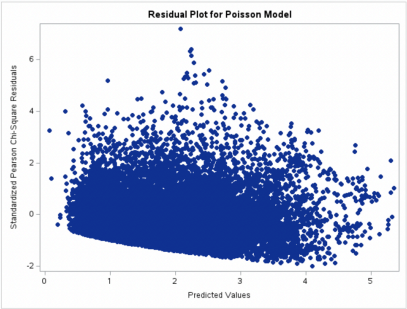
Analysis Of Maximum Likelihood Parameter Estimates

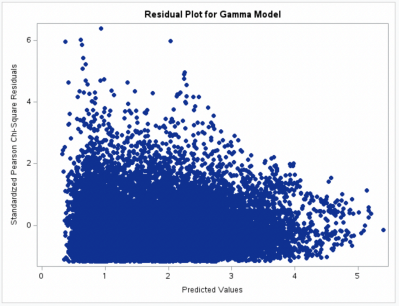
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	20.5708	1.2973	18.0281	23.1135	251.42	<.0001
TEMP		1	-0.0147	0.0020	-0.0187	-0.0108	53.27	<.0001
PRES		1	-0.0189	0.0013	-0.0214	-0.0164	221.91	<.0001
DEWP		1	0.0420	0.0016	0.0389	0.0451	723.07	<.0001
RAIN		1	-0.0858	0.0305	-0.1455	-0.0261	7.93	0.0049
wd	E	1	0.3110	0.0415	0.2298	0.3923	56.26	<.0001
wd	ENE	1	0.3079	0.0412	0.2271	0.3888	55.76	<.0001
wd	ESE	1	0.3172	0.0428	0.2334	0.4010	55.04	<.0001
wd	N	1	0.3070	0.0496	0.2098	0.4042	38.31	<.0001
wd	NE	1	0.2486	0.0421	0.1661	0.3311	34.92	<.0001
wd	NNE	1	0.2209	0.0476	0.1275	0.3143	21.49	<.0001
wd	NNW	1	0.1736	0.0615	0.0530	0.2942	7.96	0.0048
wd	NW	1	0.2273	0.0571	0.1155	0.3392	15.86	<.0001
wd	S	1	0.1804	0.0513	0.0798	0.2810	12.36	0.0004
wd	SE	1	0.1831	0.0463	0.0924	0.2737	15.66	<.0001
wd	SSE	1	0.2421	0.0501	0.1439	0.3403	23.35	<.0001
wd	SSW	1	0.0864	0.0528	-0.0171	0.1899	2.68	0.1017
wd	SW	1	0.0313	0.0497	-0.0662	0.1288	0.40	0.5289
wd	W	1	0.1056	0.0575	-0.0071	0.2183	3.38	0.0662
wd	WNW	1	0.2152	0.0608	0.0960	0.3344	12.52	0.0004
wd	WSW	0	0.0000	0.0000	0.0000	0.0000	.	.
WSPM		1	-0.0115	0.0093	-0.0296	0.0067	1.53	0.2164
month	1	1	-0.0704	0.0218	-0.1132	-0.0275	10.38	0.0013
month	2	1	-0.2632	0.0252	-0.3126	-0.2138	109.15	<.0001
month	3	1	-0.6532	0.0280	-0.7081	-0.5983	543.62	<.0001
month	4	1	-1.6222	0.0438	-1.7081	-1.5364	1371.67	<.0001
month	5	1	-1.9413	0.0573	-2.0536	-1.8291	1148.63	<.0001
month	6	1	-2.1257	0.0578	-2.2391	-2.0124	1350.84	<.0001
month	7	1	-2.7569	0.0719	-2.8977	-2.6160	1470.89	<.0001
month	8	1	-2.8783	0.0817	-3.0385	-2.7181	1240.39	<.0001
month	9	1	-2.1692	0.0556	-2.2781	-2.0602	1522.77	<.0001
month	10	1	-1.2734	0.0386	-1.3492	-1.1977	1085.95	<.0001
month	11	1	-0.6832	0.0254	-0.7330	-0.6334	722.90	<.0001
month	12	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	19.7936	1.5971	16.6634	22.9239	153.60	<.0001
TEMP		1	-0.0103	0.0023	-0.0149	-0.0058	19.95	<.0001
PRES		1	-0.0182	0.0016	-0.0213	-0.0151	135.50	<.0001
DEWP		1	0.0421	0.0019	0.0384	0.0458	498.39	<.0001
RAIN		1	-0.0268	0.0129	-0.0520	-0.0016	4.36	0.0369
wd	E	1	0.3064	0.0438	0.2206	0.3923	48.90	<.0001
wd	ENE	1	0.3044	0.0437	0.2188	0.3899	48.61	<.0001
wd	ESE	1	0.3096	0.0454	0.2207	0.3985	46.61	<.0001
wd	N	1	0.3697	0.0543	0.2633	0.4761	46.39	<.0001
wd	NE	1	0.2458	0.0448	0.1580	0.3336	30.07	<.0001
wd	NNE	1	0.2384	0.0510	0.1384	0.3385	21.82	<.0001
wd	NNW	1	0.2357	0.0646	0.1091	0.3623	13.31	0.0003
wd	NW	1	0.2317	0.0630	0.1083	0.3552	13.53	0.0002
wd	S	1	0.1686	0.0542	0.0624	0.2747	9.68	0.0019
wd	SE	1	0.1863	0.0486	0.0910	0.2816	14.68	0.0001
wd	SSE	1	0.2286	0.0528	0.1252	0.3320	18.77	<.0001
wd	SSW	1	0.1010	0.0539	-0.0046	0.2066	3.51	0.0609
wd	SW	1	0.0079	0.0514	-0.0930	0.1087	0.02	0.8785
wd	W	1	0.0987	0.0607	-0.0203	0.2178	2.64	0.1040
wd	WNW	1	0.2010	0.0669	0.0699	0.3321	9.03	0.0027
wd	WSW	0	0.0000	0.0000	0.0000	0.0000	.	.
WSPM		1	-0.0078	0.0101	-0.0276	0.0119	0.60	0.4370
month	1	1	-0.0547	0.0314	-0.1162	0.0069	3.03	0.0817
month	2	1	-0.2199	0.0349	-0.2884	-0.1515	39.68	<.0001
month	3	1	-0.6473	0.0362	-0.7182	-0.5764	320.08	<.0001
month	4	1	-1.6325	0.0477	-1.7259	-1.5391	1172.89	<.0001
month	5	1	-1.9514	0.0607	-2.0703	-1.8325	1034.68	<.0001
month	6	1	-2.1936	0.0663	-2.3236	-2.0636	1093.90	<.0001
month	7	1	-2.8401	0.0739	-2.9849	-2.6952	1475.96	<.0001
month	8	1	-2.9550	0.0770	-3.1059	-2.8042	1473.77	<.0001
month	9	1	-2.2083	0.0596	-2.3251	-2.0914	1371.90	<.0001
month	10	1	-1.3140	0.0479	-1.4078	-1.2201	752.89	<.0001
month	11	1	-0.6481	0.0330	-0.7128	-0.5833	384.63	<.0001
month	12	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	1.3610	0.0155	1.3310	1.3918		

In both models, wind speed is the only non-significant variable with a p-value of 0.2164 for the Poisson model, and 0.4370 for the Gamma model. We can remove this variable from each model. Every other variable has a parameter estimate with a p-value less than 0.05, so they are all significant in predicting the first principal component. Month and wind direction are significant categorical variables. We then obtained plots of the residuals of each model to see which is a better fit for the data:

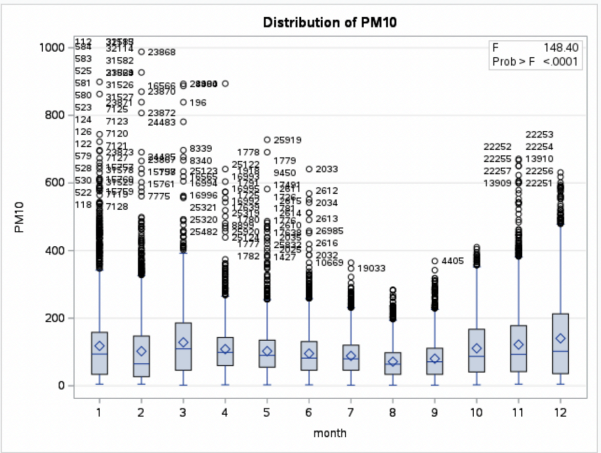


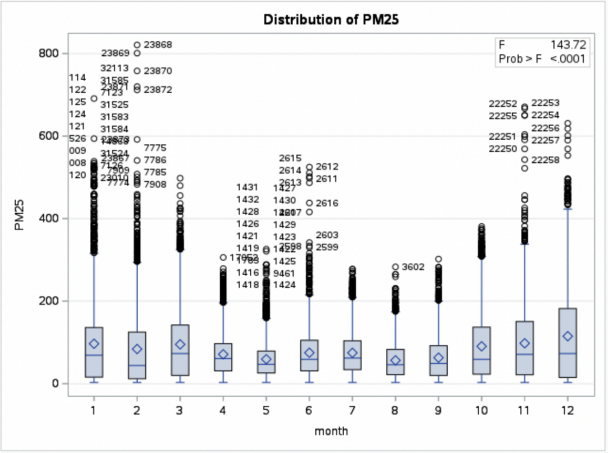


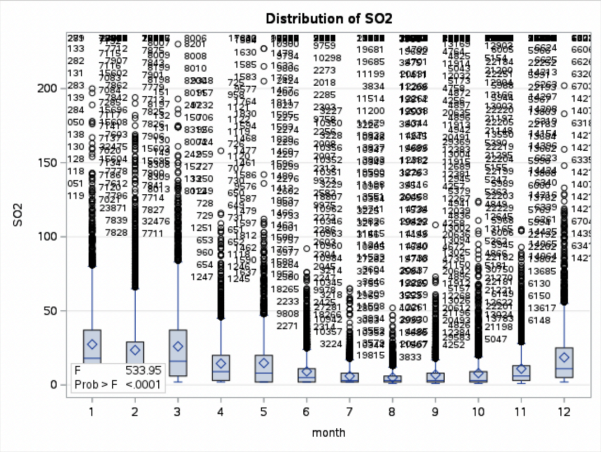
The residuals for the Poisson model are slightly more evenly distributed, so the Poisson model is a better fit for the data. We may use this model for further study about how meteorological variables and month affect air pollutant concentration.

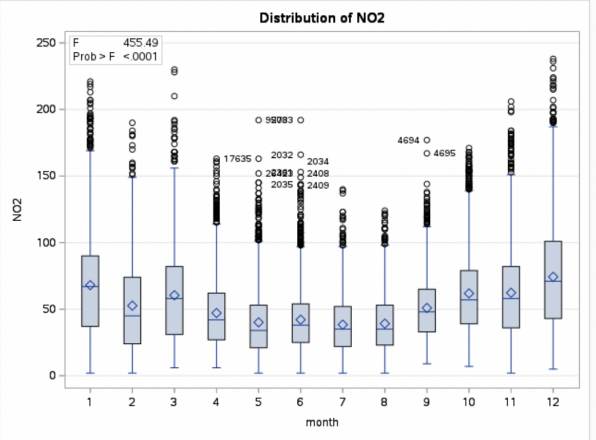
Appendix

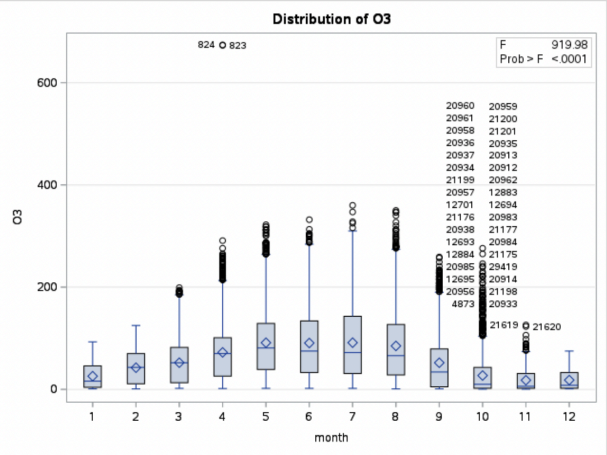
A)

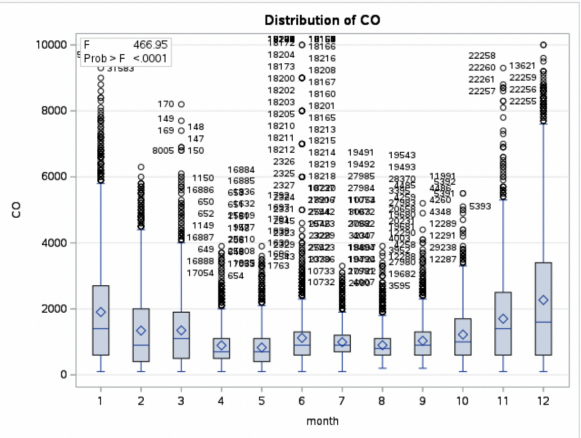








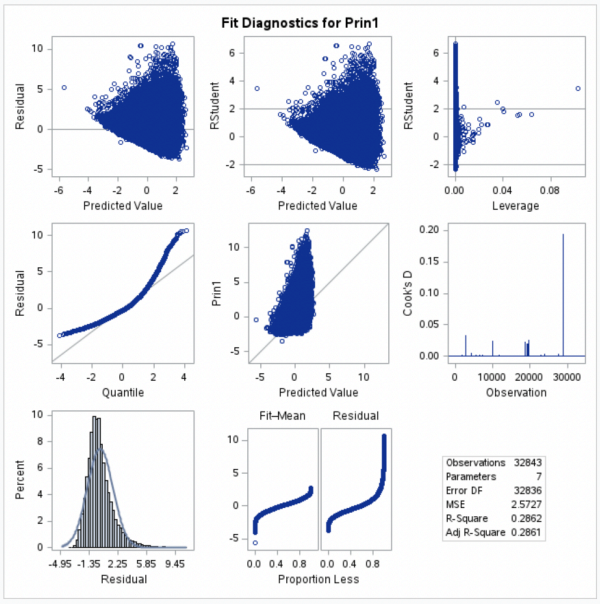




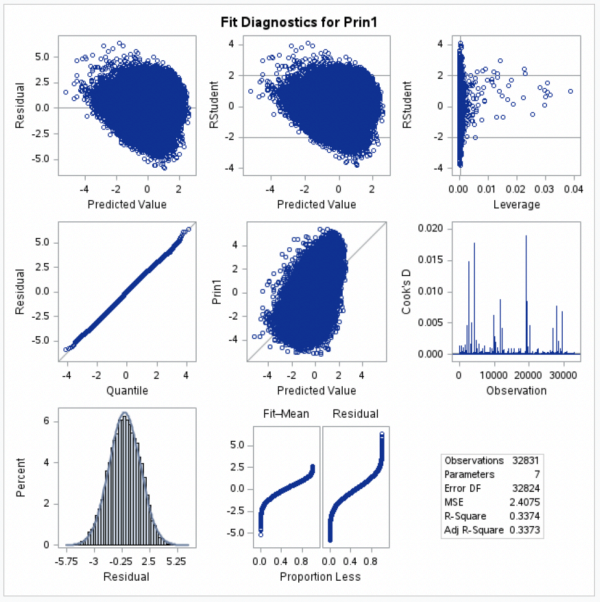
These plots show the distributions of each pollutant by month, as well as the F-values and p-values of each ANOVA test performed.

B)

Model before remedial measures:



Final model after remedial measures:



Before taking remedial measures, the residual vs. predicted values plot shows that the variance is not constant. It increases significantly as the predicted values increase. The QQ-plot and histogram show that there are deviations from the normal distribution because the points do not fall in a straight line and the histogram is right-skewed. There are also several high leverage points, highly influential points, and possible outliers. To remedy the homoscedasticity and normality issues, we performed a log transformation on each of the pollutant variables and repeated the steps above, redoing the standardization and PCA with the log-transformed data. Then, we removed highly influential points one by one based on which Cook's Distance was the highest. Once there were no more observations with a Cook's Distance much higher relative to other observations, we stopped removing observations and proceeded with stepwise selection to get our final model.

C)

Pearson Correlation Coefficients, N = 32843 Prob > |r| under H0: Rho=0												
	PM25	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	wd_numeric	WSPM
PM25	1.00000	0.89384 <.0001	0.40525 <.0001	0.66772 <.0001	0.79912 <.0001	-0.16889 <.0001	-0.14610 <.0001	-0.00050 0.9282	0.11932 <.0001	-0.01529 0.0056	-0.13902 <.0001	-0.29744 <.0001
PM10	0.89384 <.0001	1.00000	0.40022 <.0001	0.62843 <.0001	0.71905 <.0001	-0.12459 <.0001	-0.12339 <.0001	-0.02217 <.0001	0.05930 <.0001	-0.02889 <.0001	-0.09034 <.0001	-0.18476 <.0001
SO2	0.40525 <.0001	0.40022 <.0001	1.00000	0.41214 <.0001	0.41122 <.0001	-0.04860 <.0001	-0.22366 <.0001	0.15062 <.0001	-0.22102 <.0001	-0.03854 <.0001	-0.00538 0.3298	-0.03926 <.0001
NO2	0.66772 <.0001	0.62843 <.0001	0.41214 <.0001	1.00000	0.71664 <.0001	-0.54147 <.0001	-0.32121 <.0001	0.18152 <.0001	-0.07824 <.0001	-0.04844 <.0001	-0.20784 <.0001	-0.41396 <.0001
CO	0.79912 <.0001	0.71905 <.0001	0.41122 <.0001	0.71664 <.0001	1.00000	-0.32104 <.0001	-0.31667 <.0001	0.14640 <.0001	-0.03255 <.0001	-0.01461 0.0081	-0.20957 <.0001	-0.32930 <.0001
O3	-0.16889 <.0001	-0.12459 <.0001	-0.04860 <.0001	-0.54147 <.0001	-0.32104 <.0001	1.00000	0.57478 <.0001	-0.42394 <.0001	0.27153 <.0001	0.01875 0.0007	0.29311 <.0001	0.33317 <.0001
TEMP	-0.14610 <.0001	-0.12339 <.0001	-0.22366 <.0001	-0.32121 <.0001	-0.31667 <.0001	0.57478 <.0001	1.00000	-0.83370 <.0001	0.82149 <.0001	0.03886 <.0001	0.14858 <.0001	0.03839 <.0001
PRES	-0.00050 0.9282	-0.02217 <.0001	0.15062 <.0001	0.18152 <.0001	0.14640 <.0001	-0.42394 <.0001	-0.83370 <.0001	1.00000	-0.77151 <.0001	-0.06728 <.0001	-0.07986 <.0001	0.05082 <.0001
DEWP	0.11932 <.0001	0.05930 <.0001	-0.22102 <.0001	-0.07824 <.0001	-0.03255 <.0001	0.27153 <.0001	0.82149 <.0001	-0.77151 <.0001	1.00000	0.08890 <.0001	-0.04580 <.0001	-0.28631 <.0001
RAIN	-0.01529 0.0056	-0.02889 <.0001	-0.03854 <.0001	-0.04844 <.0001	-0.01461 0.0081	0.01875 0.0007	0.03886 <.0001	-0.06728 <.0001	0.08890 <.0001	1.00000	-0.01653 0.0027	0.02587 <.0001
wd_numeric	-0.13902 <.0001	-0.09034 <.0001	-0.00538 0.3298	-0.20784 <.0001	-0.20957 <.0001	0.29311 <.0001	0.14858 <.0001	-0.07986 <.0001	-0.04580 <.0001	-0.01653 0.0027	1.00000	0.22903 <.0001
WSPM	-0.29744 <.0001	-0.18476 <.0001	-0.03926 <.0001	-0.41396 <.0001	-0.32930 <.0001	0.33317 <.0001	0.03839 <.0001	0.05082 <.0001	-0.28631 <.0001	0.02587 <.0001	0.22903 <.0001	1.00000

This is a correlation coefficient matrix of the air pollutant and meteorological variables. We can use it to find variables that are highly correlated. Highly correlated variables may cause collinearity in a model and, therefore, inflated variance.