Milestone 3

Anh Le, Eve Wicksteed

The impact of weather on air quality

Introduction

The adverse affects of air pollution on health are well documented and air pollution can lead to a large range of diseases and increased morbidity and mortality (Younger et al., 2008). Adverse health impacts include, but are not limited to, lung cancer risk, respiritory infections, allergic disease and asthma (Younger et al., 2008; Shea et al., 2008). These health risks can affect a large proportion of the population as many different groups are vulnerable to the effects of air pollution including infants, children, the elderly, people with impaired immune systems, and people who work or are physically active outdoors (Matooane et al., 2004).

Because of the many, and severe, impacts of air quality, it is important to understand patterns in the data. We have a dataset of air quality observations as well as temperature and humidity data which we will use to gain understanding of the patterns and impacts of weather on air quality.

Research question

As stated above we would like to understand the impact of weather on air quality. For this reason our research question is: - What is the affect of temperature and humidity on the concentration of air pollutants, such as benzene, titania, and tin oxide?

Data and methods

Data Description

The air quality dataset used in this analysis was obtained from the University of California Irvine Machine learning Repository. It was contributed by Saverio De Vito from the National Agency for New Technologies, Energy and Sustainable Economic Development.

The dataset contains 15 variables and 9358 observations of hourly averaged responses from an Air Quality Chemical Multisensor Device. Data were recorded from March 2004 to February 2005, in a significantly polluted area, at road level, within a city in Italy. Variables include the date and time each response was recorded, and the corresponding concentrations of 13 air pollutants analyzed by the sensor device. Missing values are tagged with -200 value. Below is the entire variable set:

Variables	Type	Description
Date	character	Date (DD/MM/YYYY)
Time	time	Time (HH.MM.SS)
CO(GT)	double	True hourly averaged concentration CO in mg/m^3 (reference analyzer)
PT08.S1(CO)	integer	PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
NMHC(GT)	integer	True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (ref
C6H6(GT)	double	True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
PT08.S2(NMHC)	integer	PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
NOx(GT)	integer	True hourly averaged NOx concentration in ppb (reference analyzer)
PT08.S3(NOx)	integer	PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
NO2(GT)	integer	True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)

Variables	Type	Description
PT08.S4(NO2)	integer	PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
PT08.S5(O3)	integer	PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
T	double	Temperature in °C
RH	double	Relative Humidity (%)
AH	double	AH Absolute Humidity

Here you can see the data that we use:

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 7 parsing failures.
## row # A tibble: 5 x 5 col     row col          expected       actual file
## ... ................. ... ..........................................................................
## See problems(...) for more details.
```

Show 10 entries Search:

	Date	Time	CO	Tin_oxide	Hydro_carbons	Benzene	Titania	NOx	Tungsten_oxide_NOx	NO2	Tungsten_oxide_NO2	Indium_oxide	Temp	RH	AH	Date_time
1	2004-03-10	18:00:00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578	2004-03-10T18:00:00Z
2	2004-03-10	19:00:00	2	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255	2004-03-10T19:00:00Z
3	2004-03-10	20:00:00	2.2	1402	88	9	939	131	1140	114	1555	1074	11.9	54	0.7502	2004-03-10T20:00:00Z
4	2004-03-10	21:00:00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11	60	0.7867	2004-03-10T21:00:00Z
5	2004-03-10	22:00:00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888	2004-03-10T22:00:00Z
6	2004-03-10	23:00:00	1.2	1197	38	4.7	750	89	1337	96	1393	949	11.2	59.2	0.7848	2004-03-10T23:00:00Z
7	2004-03-11	00:00:00	1.2	1185	31	3.6	690	62	1462	77	1333	733	11.3	56.8	0.7603	2004-03-11T00:00:00Z
8	2004-03-11	01:00:00	1	1136	31	3.3	672	62	1453	76	1333	730	10.7	60	0.7702	2004-03-11T01:00:00Z
9	2004-03-11	02:00:00	0.9	1094	24	2.3	609	45	1579	60	1276	620	10.7	59.7	0.7648	2004-03-11T02:00:00Z
10	2004-03-11	03:00:00	0.6	1010	19	1.7	561		1705		1235	501	10.3	60.2	0.7517	2004-03-11T03:00:00Z

Showing 1 to 10 of 50 entries Previous 1 2 3 4 5 Next

Methods

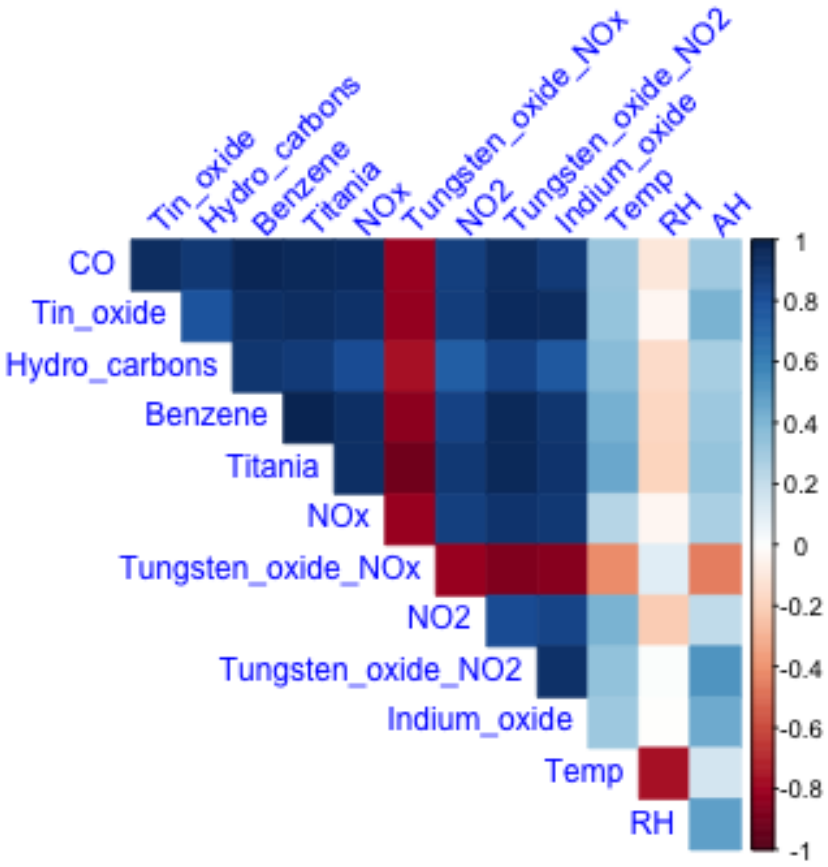
We are interested in the hourly averaged concentrations of air pollutants, temperature and humidity. We ignore variables which have too many missing data to increase the precision of this analysis. The air pollutants that we will focus on are benzene, titania and tin oxide. After dealing with the missing data, we will perform a linear regression analysis using the OLS (ordinary least squares) method. Coefficients of relevant variables will be plotted with confidence intervals.

Results

We first performed some exploratory data analysis of the air quality data.

Exploratory data analysis

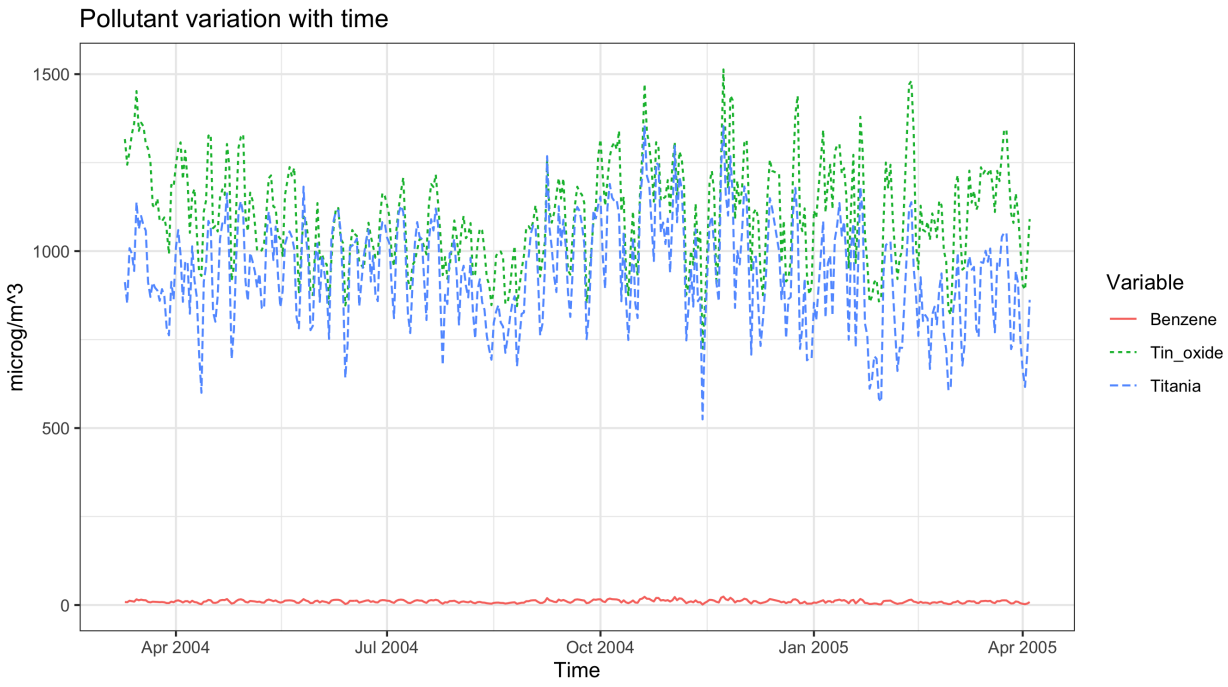
Graph 1: Correlogram of pollutants



Looking at the correlations of the pollutants with weather, we can see that for all pollutants except NOx, temperature (T) is positively correlated, although weakly so. This means that higher temperatures correspond to higher concentrations of the gases. Relative humidity (RH) is negatively and correlated to temperature and has a weak negative correlation to the concentrations of pollutants, except NOx. Absolute humidity (AH) has stronger correlations, mostly positive, although, like temperature, it has a negative correlation with NOx.

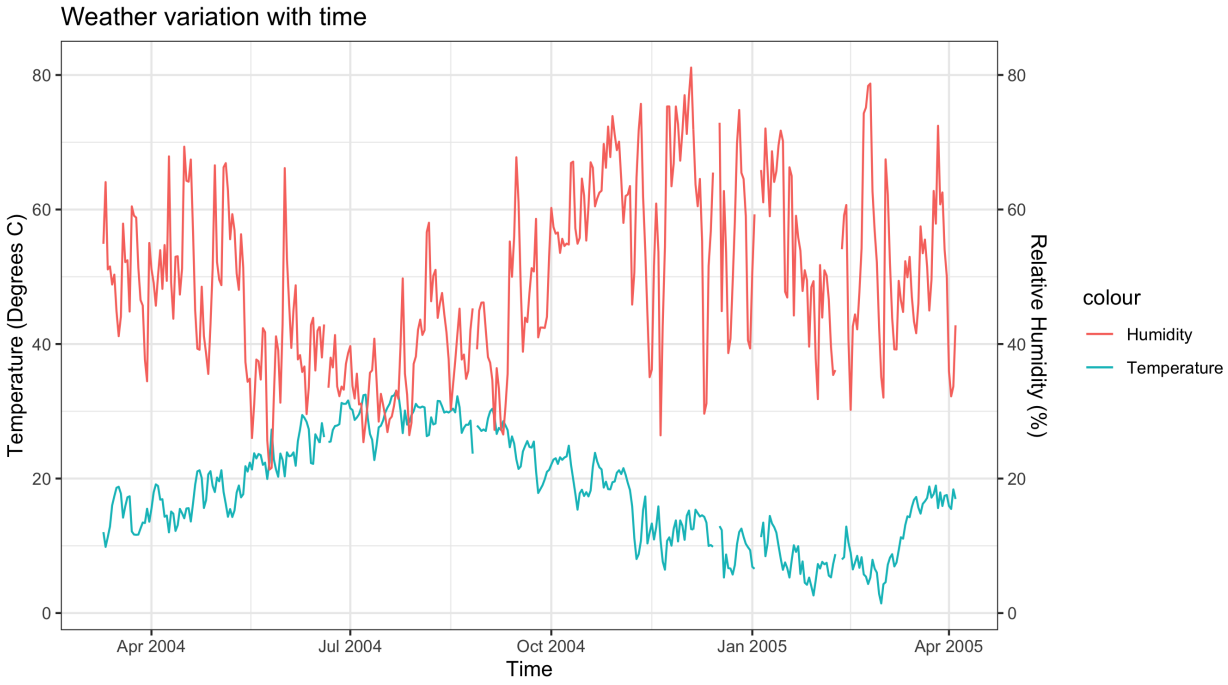
Graph 2: Concentration of some Air Pollutants, Temperature, Humidity over Time, daily average

The plot below shows the **daily** averaged concentrations of some of the pollutants (tin oxide, benzene, and Titania) for a year.



Graph 3: Concentration Temperature and Humidity over Time

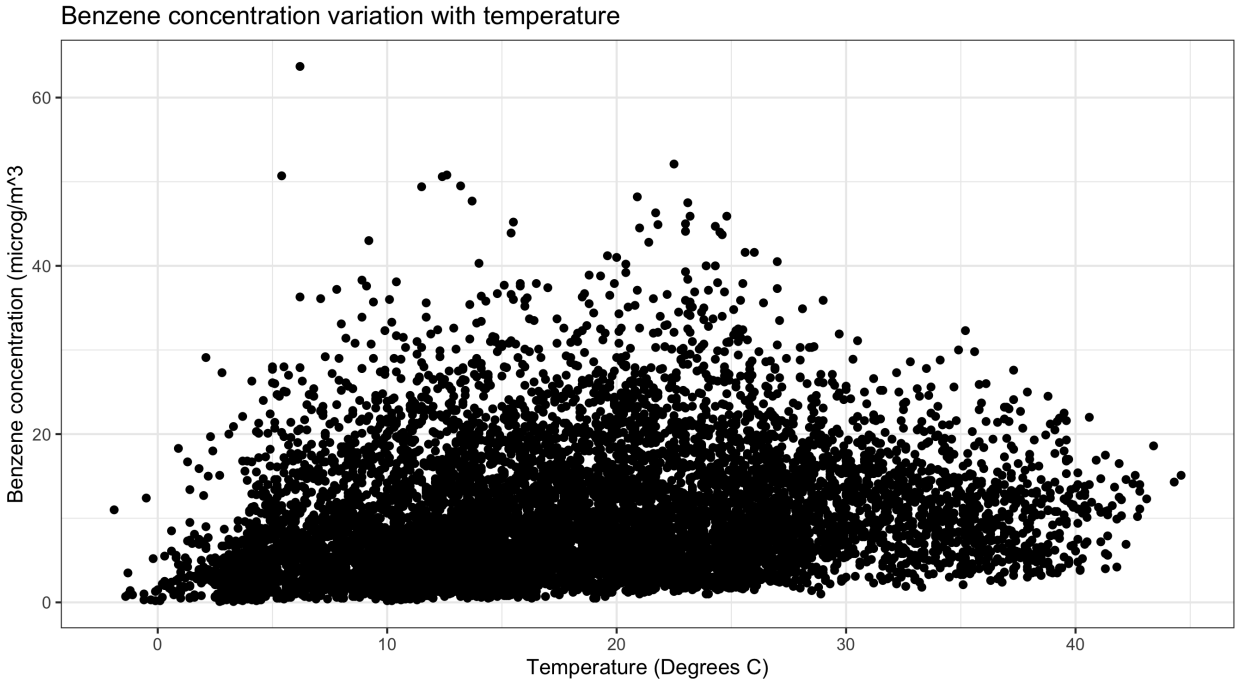
The plot below show the **daily** averaged values of temperature and humidity for a year.



Graph 4: Temperature vs. Benzene concentration

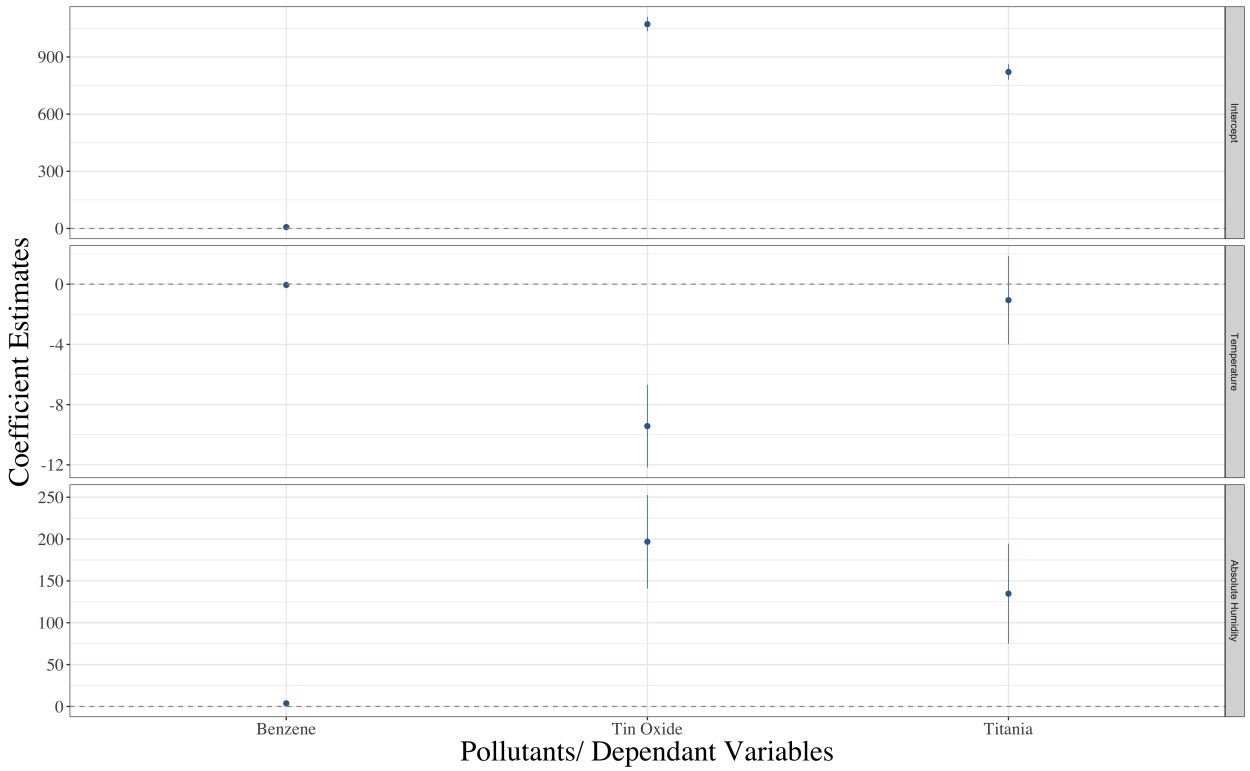
The following graph shows the relationship of benzene to temperature over the year in which data was recorded. The plot suggests there is perhaps a slight relationship. Linear regression in future work will help

to clarify the relationships between weather and pollutant concentrations.



Linear regression

We then perform linear regression of all the separate pollutants with temperature and absolute humidity. The graph below shows the coeffecients for linear regression for temperature and humidity with all the various pollutants.



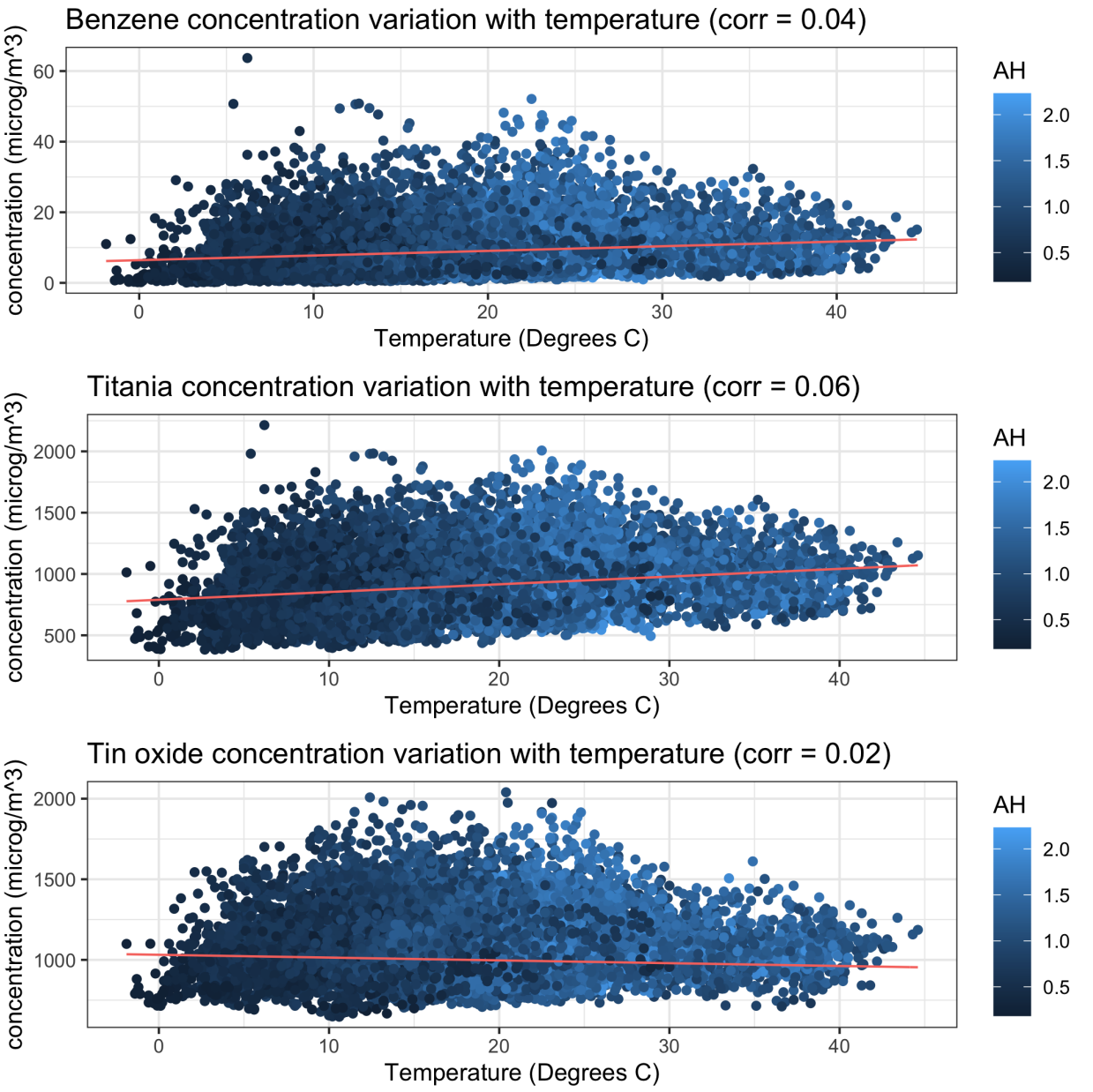
<<<<<<< HEAD

The table below shows an example of the output of the linear regression model. Dependent variable is Benzene, independent variables are temperature (Temp), and absolute humidity (AH).

The table below shows an example of the output of the linear regression model. >>>>>>> upstream/master

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   7.27      0.576      12.6  9.40e-31
## 2 Temp         -0.0523    0.0418     -1.25 2.12e- 1
## 3 AH            3.69      0.850       4.34 1.83e- 5
```

The next plot shows three of the pollutants (Benzene, Titania and Tin Oxide) plotted with temperature with the linear regression line also plotted. From looking at the plots we can tell the the linear regression line does not represent a lot of the data well. This is shown in the low correlations values (in the plot titles).



Discussion and Conclusion

It is clear that just looking at the values of temperature and humidity on their own do not provide sufficient information to explain or predict the given concentrations of pollutants. This can be seen in the figures produced above where the linear regression line does not capture well the variation of pollutants. Although weather will affect some pollutants, a more important determinant of pollutant concentrations is emissions. These often vary with time of day and human activity and thus a model incorporating weather as well as these other important variables would likely be more accurate.

References

S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical, Volume 129, Issue 2, 22 February 2008, Pages 750-757, ISSN 0925-4005.

Matooane, M., John, J., Oosthuizen, R., and Binedell, M. 2004. Vulnerability of South African communities to air pollution. In: 8th World Congress on Environmental Health. Durban, South Africa: Document Transformation Technologies.

Shea, K., Truckner, R., Weber, R., and Peden, D. 2008. Climate change and allergic disease. Journal of Allergy and Clinical Immunology, 122(3): 443-453.

Younger, M., Morrow-Almeida, H., Vindigni, S., and Dannenberg, A. 2008. The Built Environment, Climate Change, and Health Opportunities for Co-Benefits. American Journal of Preventative Medicine, 35 (5): 517-526.