Milestone 1

Anh Le, Eve Wicksteed

```
## Warning: package 'ggplot2' was built under R version 3.5.2
## Warning: package 'tibble' was built under R version 3.5.2
## Warning: package 'tidyr' was built under R version 3.5.2
## Warning: package 'purrr' was built under R version 3.5.2
## Warning: package 'dplyr' was built under R version 3.5.2
## Warning: package 'stringr' was built under R version 3.5.2
## Warning: package 'DT' was built under R version 3.5.2
## Warning: package 'knitr' was built under R version 3.5.2
## Warning: package 'tidyquant' was built under R version 3.5.2
## Warning: package 'PerformanceAnalytics' was built under R version 3.5.2
## Warning: package 'zoo' was built under R version 3.5.2
## Warning: package 'quantmod' was built under R version 3.5.2
## Warning: package 'TTR' was built under R version 3.5.2
## Warning: package 'cowplot' was built under R version 3.5.2
```

Air Quality Data

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.

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)
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

Exploring the dataset

The dataset is shown below:

```r
# first we read the data in
airq <- readr::read_csv(here::here("Data","airquality.csv"))
```

```
## Parsed with column specification:
## cols(
##   Date = col_date(format = ""),
##   Time = col_time(format = ""),
##   `CO(GT)` = col_double(),
##   `PT08.S1(CO)` = col_integer(),
##   `NMHC(GT)` = col_integer(),
##   `C6H6(GT)` = col_double(),
##   `PT08.S2(NMHC)` = col_integer(),
##   `NOx(GT)` = col_integer(),
##   `PT08.S3(NOx)` = col_integer(),
##   `NO2(GT)` = col_integer(),
##   `PT08.S4(NO2)` = col_integer(),
##   `PT08.S5(O3)` = col_integer(),
##   T = col_double(),
##   RH = col_double(),
##   AH = col_double()
## )
```

```r
DT::datatable(airq)
```

Show 10 entries Search:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
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
2	2004-03-10	19:00:00	2	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
3	2004-03-10	20:00:00	2.2	1402	88	9	939	131	1140	114	1555	1074	11.9	54	0.7502
4	2004-03-10	21:00:00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11	60	0.7867
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
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
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
8	2004-03-11	01:00:00	1	1136	31	3.3	672	62	1453	76	1333	730	10.7	60	0.7702
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
10	2004-03-11	03:00:00	0.6	1010	19	1.7	561	-200	1705	-200	1235	501	10.3	60.2	0.7517

Showing 1 to 10 of 9,357 entries Previous 1 2 3 4 5 … 936 Next

Summary Statistics

The following shows the five-number stats summary for each variable:

```r
# Five-number summary for each variable
summary(airq)
```

```
##       Date                Time               CO(GT)          PT08.S1(CO)
##  Min.   :2004-03-10   Length:9357        Min.   :-200.00   Min.   :-200
##  1st Qu.:2004-06-16   Class1:hms         1st Qu.:   0.60   1st Qu.: 921
##  Median :2004-09-21   Class2:difftime    Median :   1.50   Median :1053
##  Mean   :2004-09-21   Mode  :numeric     Mean   : -34.21   Mean   :1049
##  3rd Qu.:2004-12-28                      3rd Qu.:   2.60   3rd Qu.:1221
##  Max.   :2005-04-04                      Max.   :  11.90   Max.   :2040
##     NMHC(GT)          C6H6(GT)         PT08.S2(NMHC)       NOx(GT)
##  Min.   :-200.0   Min.   :-200.000   Min.   :-200.0   Min.   :-200.0
##  1st Qu.:-200.0   1st Qu.:   4.000   1st Qu.: 711.0   1st Qu.:  50.0
##  Median :-200.0   Median :   7.900   Median : 895.0   Median : 141.0
##  Mean   :-159.1   Mean   :   1.866   Mean   : 894.6   Mean   : 168.6
##  3rd Qu.:-200.0   3rd Qu.:  13.600   3rd Qu.:1105.0   3rd Qu.: 284.0
##  Max.   :1189.0   Max.   :  63.700   Max.   :2214.0   Max.   :1479.0
##   PT08.S3(NOx)      NO2(GT)          PT08.S4(NO2)    PT08.S5(O3)
##  Min.   :-200   Min.   :-200.00   Min.   :-200   Min.   :-200.0
##  1st Qu.: 637   1st Qu.:  53.00   1st Qu.:1185   1st Qu.: 700.0
##  Median : 794   Median :  96.00   Median :1446   Median : 942.0
##  Mean   : 795   Mean   :  58.15   Mean   :1391   Mean   : 975.1
##  3rd Qu.: 960   3rd Qu.: 133.00   3rd Qu.:1662   3rd Qu.:1255.0
##  Max.   :2683   Max.   : 340.00   Max.   :2775   Max.   :2523.0
##        T                  RH                AH
##  Min.   :-200.000   Min.   :-200.00   Min.   :-200.0000
##  1st Qu.:  10.900   1st Qu.:  34.10   1st Qu.:   0.6923
##  Median :  17.200   Median :  48.60   Median :   0.9768
##  Mean   :   9.778   Mean   :  39.49   Mean   :  -6.8376
##  3rd Qu.:  24.100   3rd Qu.:  61.90   3rd Qu.:   1.2962
##  Max.   :  44.600   Max.   :  88.70   Max.   :   2.2310
```

The following shows some preliminary info on the air quality dataset that we are using. We record the number of total observations, number of missing observations, percentage of missing values and the number of usable observations.

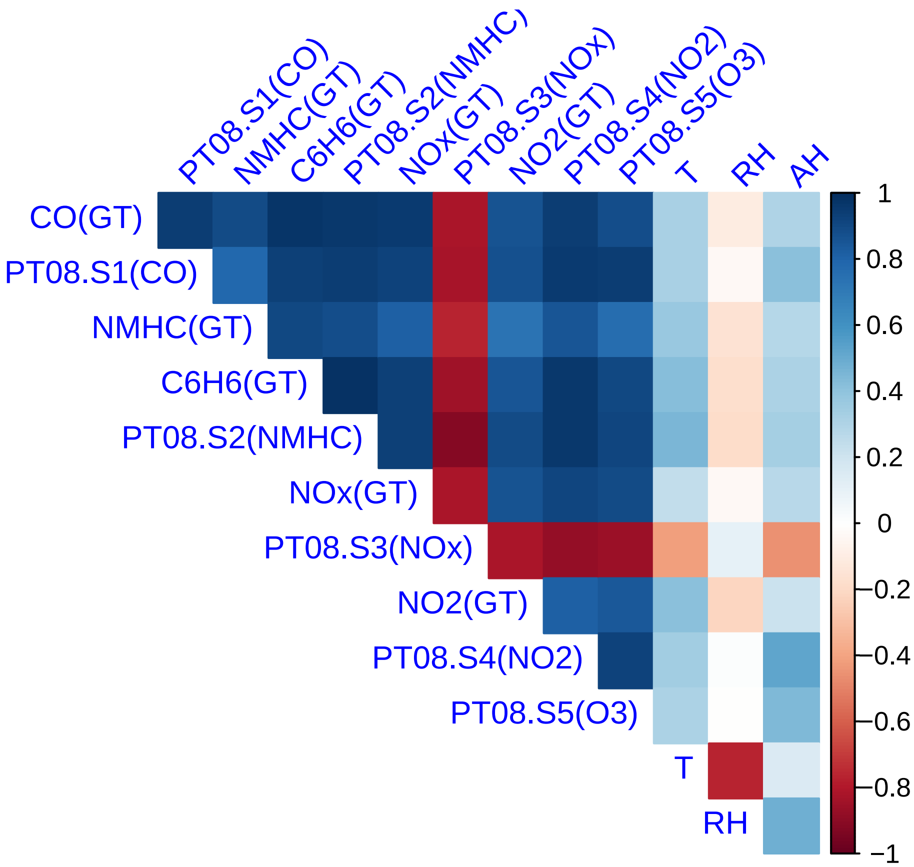
```r
# Look at missing values for each variable
missing = list()
for(i in 1:15) {
  l = length(which(airq[i] == -200))
  missing[[i]] = l
}
obs = list()
for(i in 1:15) {
  o = length(airq[[i]])
  obs[[i]] = o
}
dfmissing = data.frame(Variables,
                       matrix(unlist(missing), nrow=length(missing), byrow=T),
                       matrix(unlist(obs), nrow=length(missing), byrow=T))
names(dfmissing)[names(dfmissing) == "matrix.unlist.missing...nrow...length.missing...byrow...T."] = "Co
names(dfmissing)[names(dfmissing) == "matrix.unlist.obs...nrow...length.missing...byrow...T."] = "Total
dfmissing %>%
  mutate(`% Missing Values` = `Count of Missing Values`/`Total Observations`*100) %>%
  mutate(`Usable Observations` = `Total Observations` - `Count of Missing Values`)
```

```
##        Variables Count of Missing Values Total Observations
## 1           Date                       0               9357
## 2           Time                       0               9357
## 3         CO(GT)                    1683               9357
## 4    PT08.S1(CO)                     366               9357
## 5       NMHC(GT)                    8443               9357
## 6       C6H6(GT)                     366               9357
## 7  PT08.S2(NMHC)                     366               9357
## 8        NOx(GT)                    1639               9357
## 9   PT08.S3(NOx)                     366               9357
## 10       NO2(GT)                    1642               9357
## 11  PT08.S4(NO2)                     366               9357
## 12   PT08.S5(O3)                     366               9357
## 13             T                     366               9357
## 14            RH                     366               9357
## 15            AH                     366               9357
##    % Missing Values Usable Observations
## 1           0.00000                9357
## 2           0.00000                9357
## 3          17.98653                7674
## 4           3.91151                8991
## 5          90.23191                 914
## 6           3.91151                8991
## 7           3.91151                8991
## 8          17.51630                7718
## 9           3.91151                8991
## 10         17.54836                7715
## 11          3.91151                8991
## 12          3.91151                8991
## 13          3.91151                8991
## 14          3.91151                8991
## 15          3.91151                8991
```

From this we see that for many of the observations less than 4% of the data is missing. This is adequate for the research we are conducting.

Graph 1: Correlogram of pollutants

```
#convert Missing Values (tagged with -200 value) to NA
airq[airq == -200] = NA

# Convert numeric columns to 'double' type
airq[3:15] <- sapply(airq[3:15], as.double)
airq_corr <- cor(airq[3:15], use = "complete.obs")

# Round the values to 2 decimal places
airq_corr <- round(airq_corr,2)
corrplot(airq_corr,
         type="upper",
         method="color",
         tl.srt=45,
         tl.col = "blue",
         diag = FALSE)
```

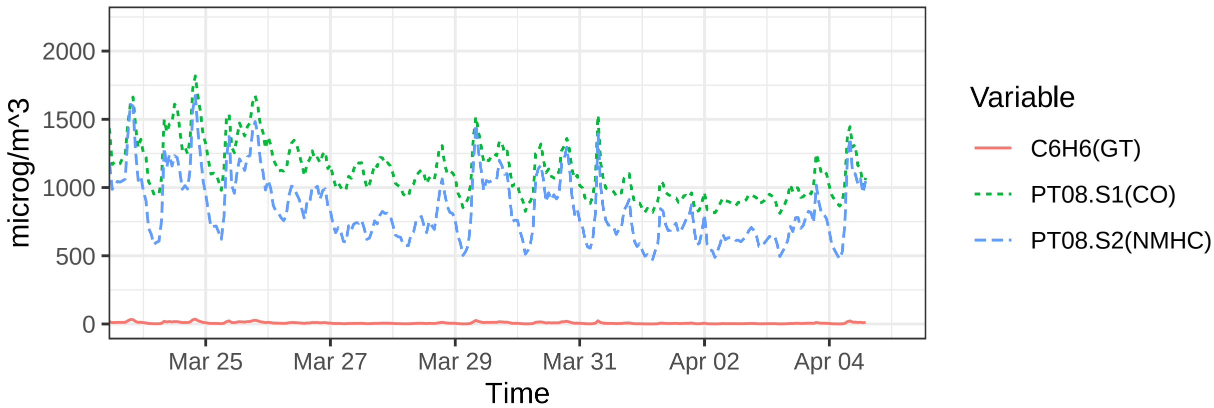


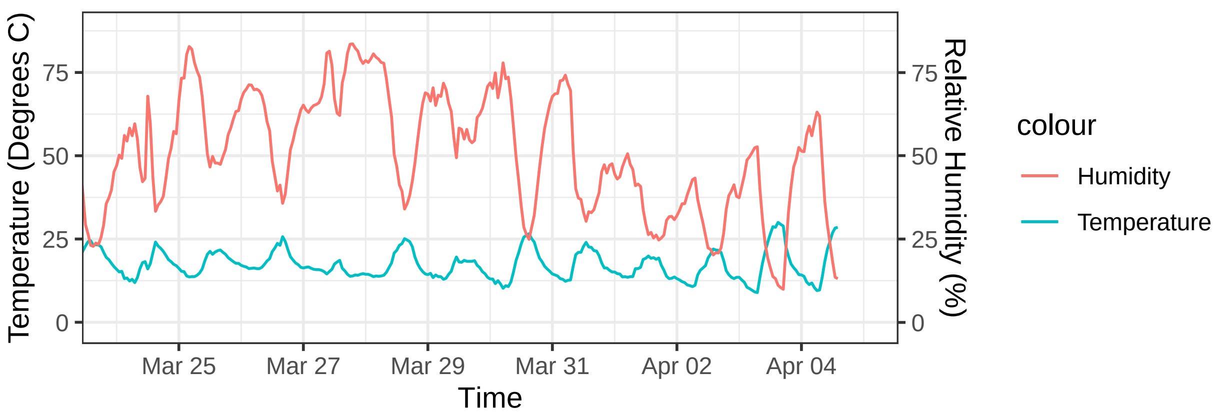
Looking at the correlations of the pollutants with weather, we can see that for all pollutants except NOx, temperature (T) is positively correlated, although weakly so. This means that higher temperatures correspond to higher concentrations of the gases. Relative humidity (RH) is negatively and correlated 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

The following plot shows the **hourly** concentrations of some of the pollutants (tin oxide, benzene, and Titania) for a month. Also plotted are temperature and relative humidity.

```
#create a time stamp with both date and time (might want to move this to earlier section?)
airq = airq %>%
  mutate(Date_Time = ymd_hms(paste(airq$Date, airq$Time)))

#data preparation: short to long, select pollutants to be included
airq.long = airq %>%
  #select(Date_Time, `PT08.S1(CO)`, `C6H6(GT)`, `PT08.S2(NMHC)`,`PT08.S3(NOx)`, `PT08.S4(NO2)`, `PT08.S
  select(Date_Time, `PT08.S1(CO)`,`C6H6(GT)`, `PT08.S2(NMHC)`) %>%
  gather(key = "Variable", value = "Value", -Date_Time)
#head(airq.long)

weather.long = airq %>%
  select(Date_Time,`T`, `RH`) %>%
  gather(key = "Variable", value = "Value", -Date_Time)

#Graph of pollutants' concentration by hours
plot_1 <- airq.long %>%
  drop_na(Value) %>%
  ggplot(aes(x = Date_Time, y = Value)) +
  geom_line(aes(color = Variable, linetype = Variable)) +
  theme_bw() +
  coord_x_datetime(xlim = c("2005-03-24 01:00:00", "2005-04-04 23:00:00")) +
  xlab("Time") +
  ylab("microg/m^3")


plot_2 <- airq %>%
  ggplot(aes(x = Date_Time)) +
  geom_line(aes(y=T, colour = "Temperature")) +
  theme_bw() +
  coord_x_datetime(xlim = c("2005-03-24 01:00:00", "2005-04-04 23:00:00")) +
  xlab("Time") +
  ylab("Temperature (Degrees C)") +
  geom_line(aes(y=RH, colour = "Humidity")) +
  scale_y_continuous(sec.axis = sec_axis(~., name = "Relative Humidity (%)"))

plot_grid(plot_1, plot_2, ncol=1)
```





Graph 3: Concentration of some Air Pollutants, Temperature, Humidity over Time, daily average

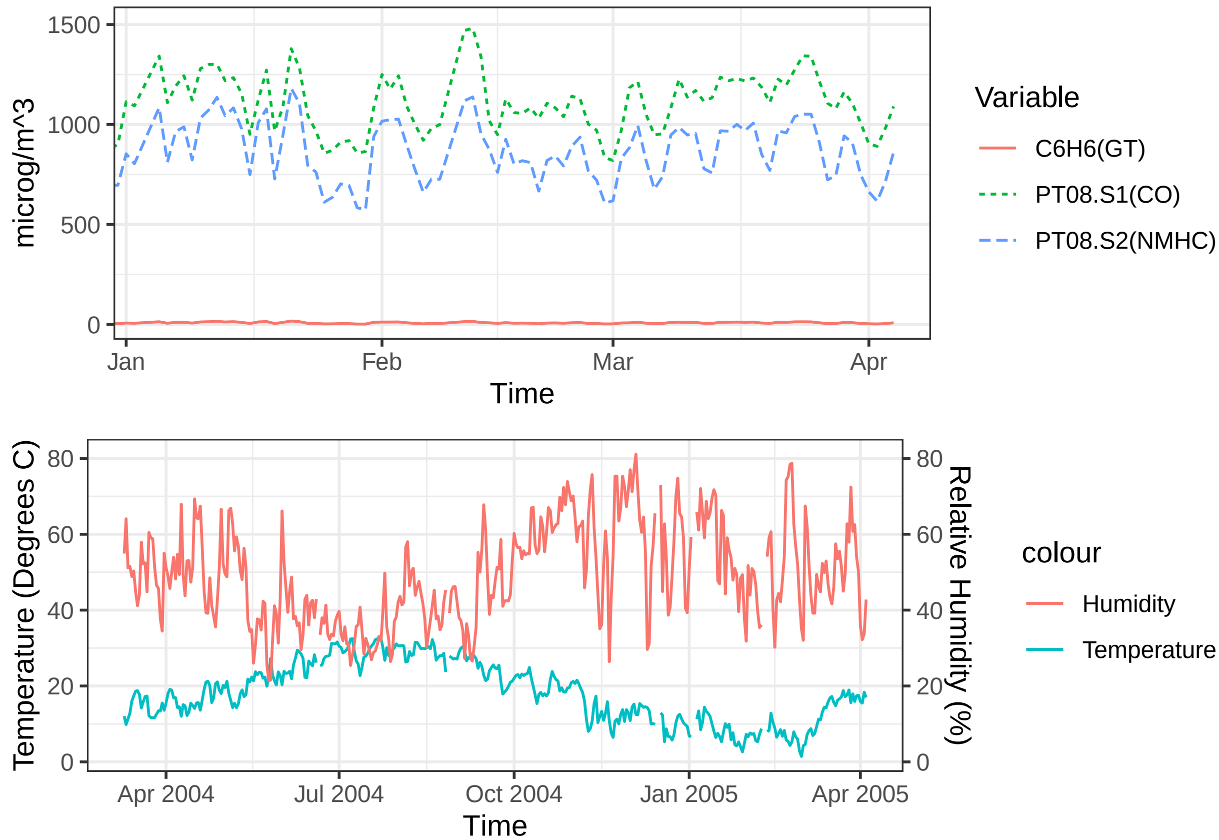
The plots below now show the **daily** averaged concentrations of some of the pollutants (tin oxide, benzene, and Titania) for a year. Also plotted are daily temperature and relative humidity.

```
#Aggregate Daily Average
airq_daily = airq %>%
  group_by(Date) %>%
  summarise_all(funs(mean), na.rm = TRUE)
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

```
#data preparation: short to long, select pollutants to be included
airq.lg.d = airq_daily %>%
  #select(Date, `PT08.S1(CO)`, `C6H6(GT)`, `PT08.S2(NMHC)`,`PT08.S3(NOx)`, `PT08.S4(NO2)`, `PT08.S5(O3)
```

```r
  select(Date, `PT08.S1(CO)`,`C6H6(GT)`, `PT08.S2(NMHC)`) %>% #, `T`, AH) %>%
  gather(key = "Variable", value = "Value", -Date)
#head(airq.lg.d)

weather.dly.long = airq_daily %>%
  select(Date,`T`, `RH`) %>%
  gather(key = "Variable", value = "Value", -Date)


#Graph of pollutants' concentration by day
plot_3 <- airq.lg.d %>%
  drop_na(Value) %>%
  ggplot(aes(x = Date, y = Value)) +
  geom_line(aes(color = Variable, linetype = Variable)) +
  theme_bw() +
  coord_x_date(xlim = c("2005-01-04", "2005-04-04")) +
  xlab("Time") +
  ylab("microg/m^3")

plot_4 <- airq_daily %>%
  ggplot(aes(x = Date)) +
  geom_line(aes(y=T, colour = "Temperature")) +
  theme_bw() +
  #coord_x_datetime(xlim = c("2005-01-04", "2005-04-04")) +
  xlab("Time") +
  ylab("Temperature (Degrees C)") +
  geom_line(aes(y=RH, colour = "Humidity")) +
  scale_y_continuous(sec.axis = sec_axis(~., name = "Relative Humidity (%)"))

plot_grid(plot_3, plot_4, ncol=1)
```

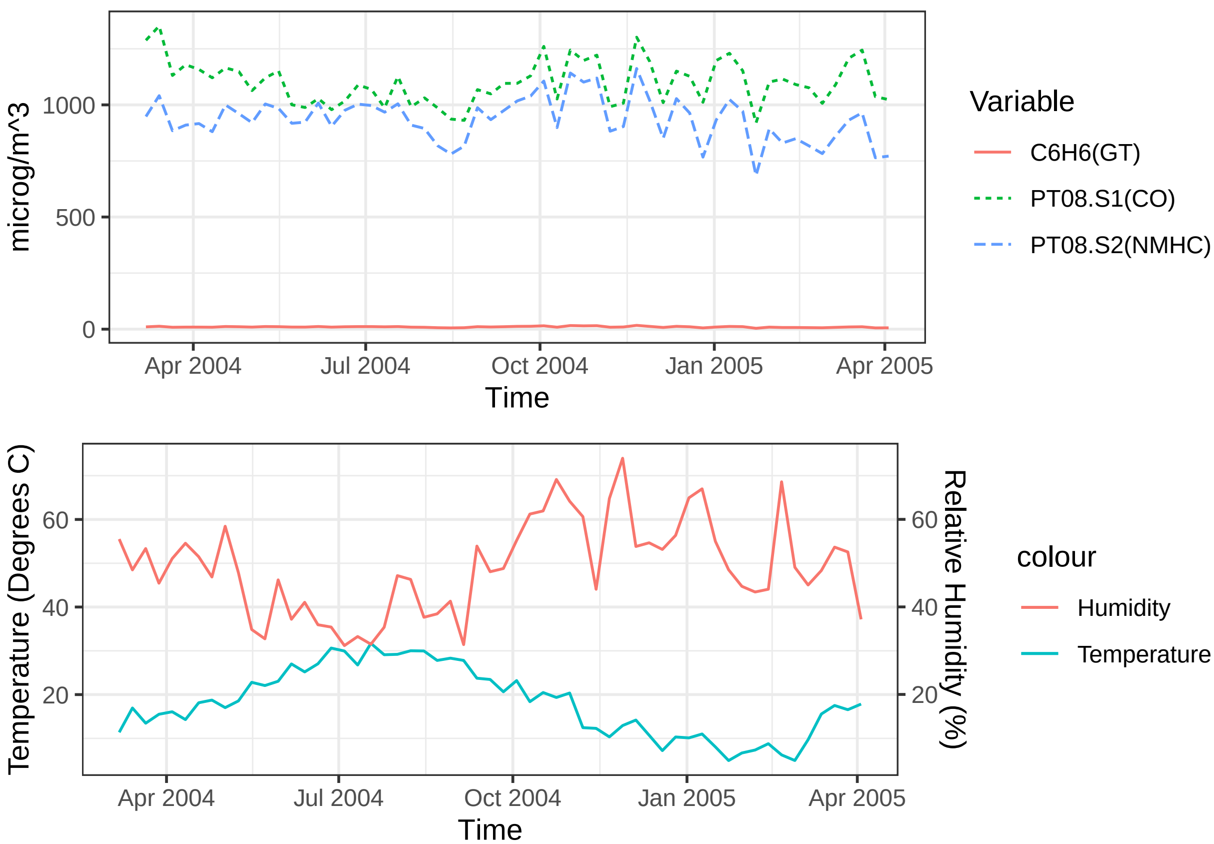


Graph 4: Concentration of some Air Pollutants, Temperature, Humidity over Time, weekly average

We now show the **weekly** averaged concentrations of some of the pollutants (tin oxide, benzene, and Titania) for a year. Also plotted are daily temperature and relative humidity.

```
#Aggregate Weekly Average
airq_weekly = airq %>%
  mutate(Week = floor_date(Date_Time, unit = "week")) %>%
  group_by(Week) %>%
  summarise_all(funs(mean), na.rm = TRUE)

#data preparation: short to long, select pollutants to be included
airq.lg.w = airq_weekly %>%
  #select(Week, `PT08.S1(CO)`, `C6H6(GT)`, `PT08.S2(NMHC)`,`PT08.S3(NOx)`, `PT08.S4(NO2)`, `PT08.S5(O3)
  select(Week, `PT08.S1(CO)`,`C6H6(GT)`, `PT08.S2(NMHC)`) %>% #, `T`, AH) %>%
  gather(key = "Variable", value = "Value", -Week)
#head(airq.lg.w)



#Graph of pollutants' concentration by week
plot_5 <- airq.lg.w %>%
  drop_na(Value) %>%
  ggplot(aes(x = Week, y = Value)) +
  geom_line(aes(color = Variable, linetype = Variable)) +
```

```
  theme_bw() +
  #coord_x_date(xlim = c("2005-01-04", "2005-04-04")) +
  xlab("Time") +
  ylab("microg/m^3")


plot_6 <- airq_weekly %>%
  ggplot(aes(x = Week)) +
  geom_line(aes(y=T, colour = "Temperature")) +
  theme_bw() +
  #coord_x_datetime(xlim = c("2005-01-04", "2005-04-04")) +
  xlab("Time") +
  ylab("Temperature (Degrees C)") +
  geom_line(aes(y=RH, colour = "Humidity")) +
  scale_y_continuous(sec.axis = sec_axis(~., name = "Relative Humidity (%)"))

plot_grid(plot_5, plot_6, ncol=1)
```

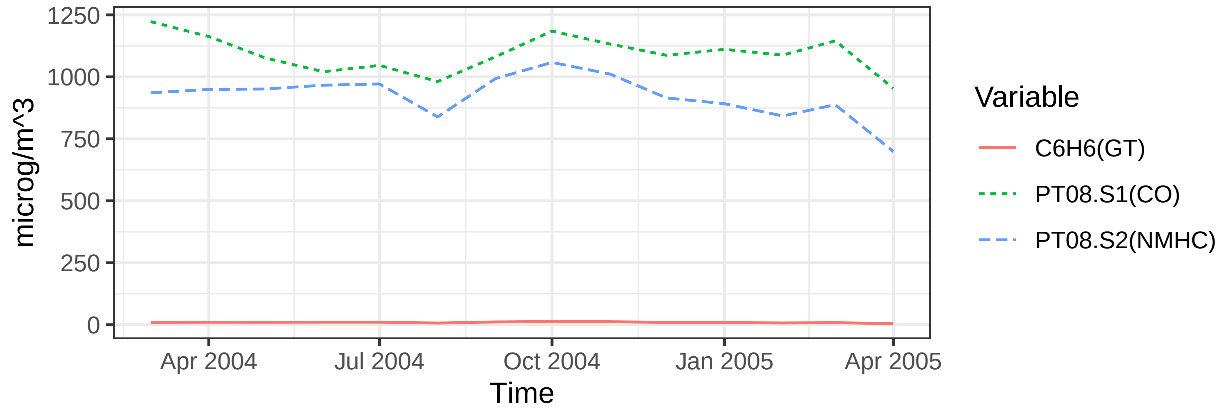


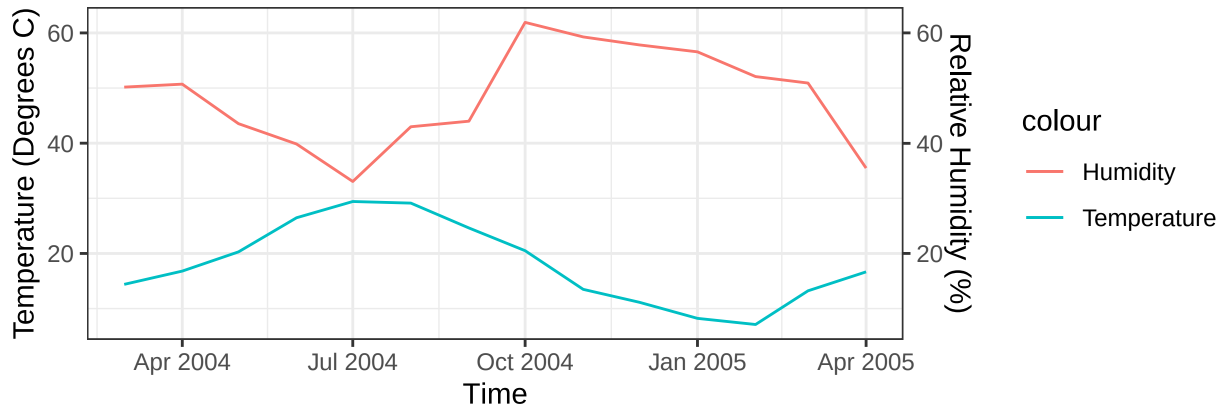
Graph 5: Concentration of some Air Pollutants, Temperature, Humidity over Time, monthly average

We now show the **monthly** averaged concentrations of some of the pollutants (tin oxide, benzene, and Titania) for a year. Also plotted are daily temperature and relative humidity.

```
#Aggregate Daily Average
airq_monthly = airq %>%
```

```r
  mutate(Month = floor_date(Date_Time, unit = "month")) %>%
  group_by(Month) %>%
  summarise_all(funs(mean), na.rm = TRUE)

#data preparation: short to long, select pollutants to be included
#Graph of pollutants' concentration by month
plot_7 <- airq_monthly %>%
  #select(Month, `PT08.S1(CO)`, `C6H6(GT)`, `PT08.S2(NMHC)`,`PT08.S3(NOx)`, `PT08.S4(NO2)`, `PT08.S5(O3
  select(Month, `PT08.S1(CO)`,`C6H6(GT)`, `PT08.S2(NMHC)`) %>%  #, `T`, AH) %>%
  gather(key = "Variable", value = "Value", -Month) %>%
  drop_na(Value) %>%
  ggplot(aes(x = Month, y = Value)) +
  geom_line(aes(color = Variable, linetype = Variable)) +
  theme_bw() +
  #coord_x_date(xlim = c("2005-01-04", "2005-04-04")) +
  xlab("Time") +
  ylab("microg/m^3")

#Graph of temperature and humidity by month

plot_8 <- airq_monthly %>%
  ggplot(aes(x = Month)) +
  geom_line(aes(y=T, colour = "Temperature")) +
  theme_bw() +
  #coord_x_datetime(xlim = c("2005-01-04", "2005-04-04")) +
  xlab("Time") +
  ylab("Temperature (Degrees C)") +
  geom_line(aes(y=RH, colour = "Humidity")) +
  scale_y_continuous(sec.axis = sec_axis(~., name = "Relative Humidity (%)"))

plot_grid(plot_7, plot_8, ncol=1)
```






```
#airq


newnames <- c("Date", "Time", "CO", "Tin_oxide", "Hydro_carbons", "Benzene", "Titania", "NOx", "Tungste

for (i in 1:ncol(airq)){
  print(i)
  names(airq)[i]=newnames[i]
}
```

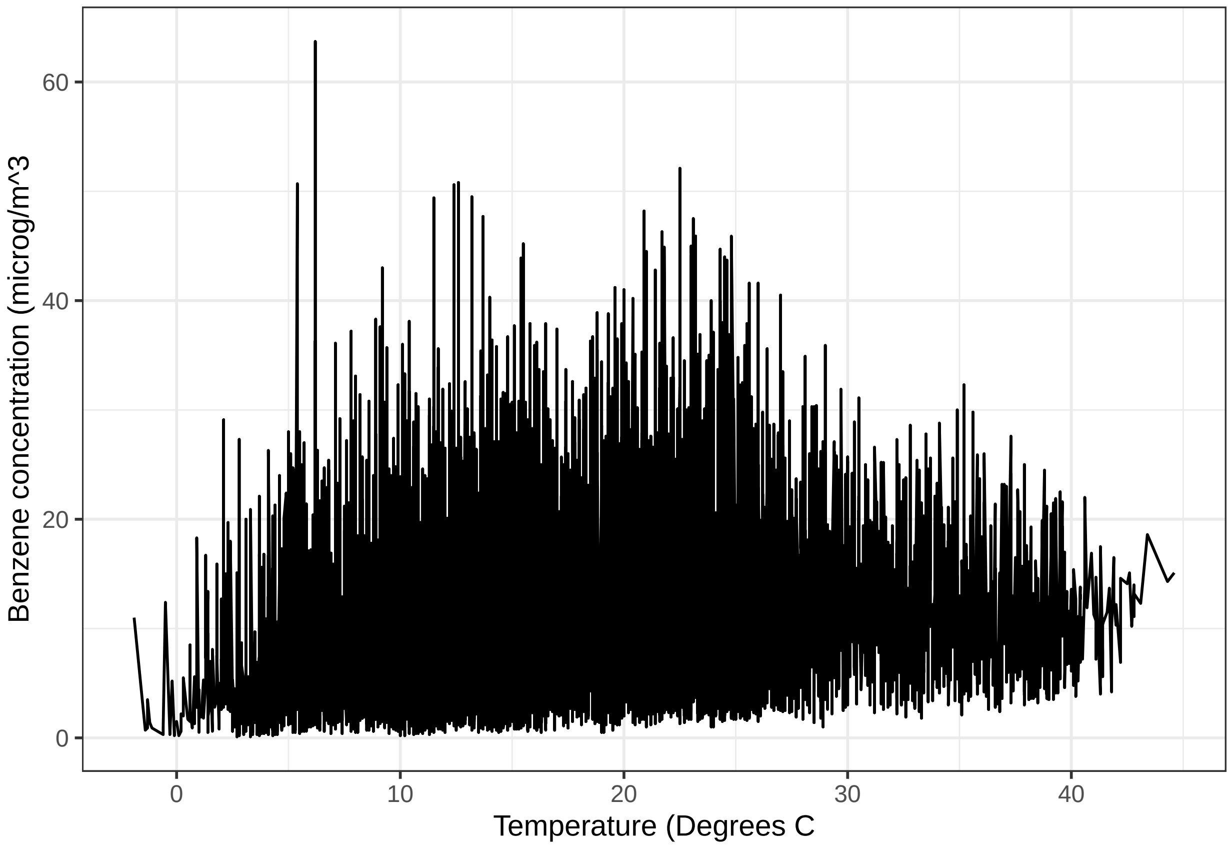
```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
```

```r
airq
```

```
## # A tibble: 9,357 x 16
##    Date       Time     CO Tin_oxide Hydro_carbons Benzene Titania   NOx
##    <date>     <drt> <dbl>     <dbl>         <dbl>   <dbl>   <dbl> <dbl>
##  1 2004-03-10 18:00   2.6      1360           150    11.9    1046   166
##  2 2004-03-10 19:00   2        1292           112     9.4     955   103
##  3 2004-03-10 20:00   2.2      1402            88     9       939   131
##  4 2004-03-10 21:00   2.2      1376            80     9.2     948   172
##  5 2004-03-10 22:00   1.6      1272            51     6.5     836   131
##  6 2004-03-10 23:00   1.2      1197            38     4.7     750    89
##  7 2004-03-11 00:00   1.2      1185            31     3.6     690    62
##  8 2004-03-11 01:00   1        1136            31     3.3     672    62
##  9 2004-03-11 02:00   0.9      1094            24     2.3     609    45
## 10 2004-03-11 03:00   0.6      1010            19     1.7     561    NA
## # ... with 9,347 more rows, and 8 more variables:
## #   Tungsten_oxide_NOx <dbl>, NO2 <dbl>, Tungsten_oxide_NO2 <dbl>,
## #   Indium_oxide <dbl>, Temp <dbl>, RH <dbl>, AH <dbl>, NA <dttm>
```

```r
# rename_f = function(col, newname){
#   names(airq)[col]=newname
# }

#new <- map(newnames, ~rename_f(cols,.x))


#new
airq %>%
  ggplot() +
  geom_line(aes(y=Benzene, x=Temp)) +
  theme_bw() +
  xlab("Temperature (Degrees C") +
  ylab("Benzene concentration (microg/m^3")
```

```
## Warning: Removed 366 row(s) containing missing values (geom_path).
```



Research question

In this analysis, we will attempt to determine the effects of temperature and humidity on the concentration of air pollutants.

Plan of action

With our research question, we are interested in the hourly averaged concentrations of air pollutants, temperature and humidity. We will ignore variables which have too many missing data to increase the precision of this analysis. After dealing with the missing data, we will perform a linear regression analysis using OLS (ordinary least square) method. Coefficients of relevant variables will be plotted with confidence intervals.

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.