CSE Capstone - HSPH EDA

Keyan Halperin

February 11, 2018

```
library(tidyverse)
library(lubridate)
library(psych)
```

```
data = read.csv('random_subset.csv', stringsAsFactors = F)

#67,000 observations, 116 variables
dim(data)
```

```
## [1] 66943   116
```

```
data$date = as.Date(data$date)
data$month = month(data$date)
```

```
#Proportion of missing values by variable
missing.table = sapply(data, function(x) round( mean(is.na(x)), 4))
sort(missing.table, decreasing = T)[1:40] #78% of monitor data is missing
```

```
##                  REANALYSIS_windspeed_10m_1Day
##                                         1.0000
##        MAIACUS_Optical_Depth_047_Aqua_Nearest4
##                                         0.7881
##        MAIACUS_Optical_Depth_055_Aqua_Nearest4
##                                         0.7881
##                                    MonitorData
##                                         0.7813
##       MAIACUS_Optical_Depth_047_Terra_Nearest4
##                                         0.7312
##       MAIACUS_Optical_Depth_055_Terra_Nearest4
##                                         0.7312
##                OMSO2e_ColumnAmountSO2_PBL_Mean
##                                         0.6659
## OMNO2d_ColumnAmountNO2StratoCloudScreened_Mean
##                                         0.6179
##                                    MOD04L2_550
##                                         0.5596
##                   MOD11A1_LST_Day_1km_Nearest4
##                                         0.5401
##                 MOD11A1_Clear_day_cov_Nearest4
##                                         0.5401
##                 MOD11A1_LST_Night_1km_Nearest4
##                                         0.5363
##               MOD11A1_Clear_night_cov_Nearest4
##                                         0.5363
##                   OMAERUVe_VISAerosolIndex_Mean
##                                         0.4572
##                    OMAERUVe_UVAerosolIndex_Mean
##                                         0.4564
##                   OMAERUVd_UVAerosolIndex_Mean
```

```
##                                     0.4481
##                         OMTO3e_ColumnAmountO3
##                                     0.4425
##                           OMUVBd_UVindex_Mean
##                                     0.4184
##                                     OMO3PR
##                                     0.3135
##                      MAIACUS_cosVZA_Aqua_Nearest
##                                     0.1805
##                       REANALYSIS_gflux_DailyMean
##                                     0.1468
##                       REANALYSIS_soilm_DailyMean
##                                     0.1468
##                     MAIACUS_cosVZA_Terra_Nearest
##                                     0.0672
##                      Nearby_Peak2_MaxTemperature
##                                     0.0604
##                     Nearby_Peak2_MeanTemperature
##                                     0.0604
##                      Nearby_Peak2_MinTemperature
##                                     0.0604
##                  Nearby_Peak2Lag1_MaxTemperature
##                                     0.0604
##                 Nearby_Peak2Lag1_MeanTemperature
##                                     0.0604
##                  Nearby_Peak2Lag1_MinTemperature
##                                     0.0604
##                  Nearby_Peak2Lag3_MaxTemperature
##                                     0.0604
##                 Nearby_Peak2Lag3_MeanTemperature
##                                     0.0604
##                  Nearby_Peak2Lag3_MinTemperature
##                                     0.0604
##                                MOD13A2_Nearest4
##                                     0.0338
##                                    MOD09A1
##                                     0.0273
##                      RoadDensity_prisecroads1000
##                                     0.0136
##                     RoadDensity_prisecroads10000
##                                     0.0136
##                           RoadDensity_roads1000
##                                     0.0124
##                               Nearby_Peak2_NO2
##                                     0.0048
##                             Nearby_Peak2Lag1_NO2
##                                     0.0046
##                             Nearby_Peak2Lag3_NO2
##                                     0.0043
```

```
range(data$site)
```

```
## [1]    1 2156
```

```r
#All 2156 sites are accounted for in this subset
length(unique(data$site))
```

```
## [1] 2156
```

```r
#Each site has at least 15 observations, at most 48
data %>% count(site) %>% summarize(min_count = min(n), avg_count = mean(n), max_count = max(n))
```

```
## # A tibble: 1 × 3
##   min_count avg_count max_count
##       <int>     <dbl>     <int>
## 1        15  31.04963        48
```

```r
#Sites with lowest avg pollution
data %>% group_by(site) %>% summarize(avg_pollution = mean(MonitorData, na.rm = T)) %>%
  arrange(avg_pollution) %>% slice(1:10)
```

```
## # A tibble: 10 × 2
##     site avg_pollution
##    <int>         <dbl>
## 1    964      0.563500
## 2   2117      0.732700
## 3   1174      0.781950
## 4    278      0.816670
## 5   1092      0.890620
## 6   1013      1.033652
## 7    943      1.100000
## 8   1364      1.300000
## 9   2003      1.500000
## 10   884      1.504170
```

```r
#Sites with highest avg pollution
data %>% group_by(site) %>% summarize(avg_pollution = mean(MonitorData, na.rm = T)) %>%
  arrange(desc(avg_pollution)) %>% slice(1:10)
```

```
## # A tibble: 10 × 2
##     site avg_pollution
##    <int>         <dbl>
## 1   1630          46.2
## 2   1876          39.2
## 3    223          35.4
## 4   1504          32.8
## 5   1401          32.4
## 6   1676          32.2
## 7   1935          31.5
## 8   2026          31.2
## 9   1700          27.5
## 10  1884          26.5
```

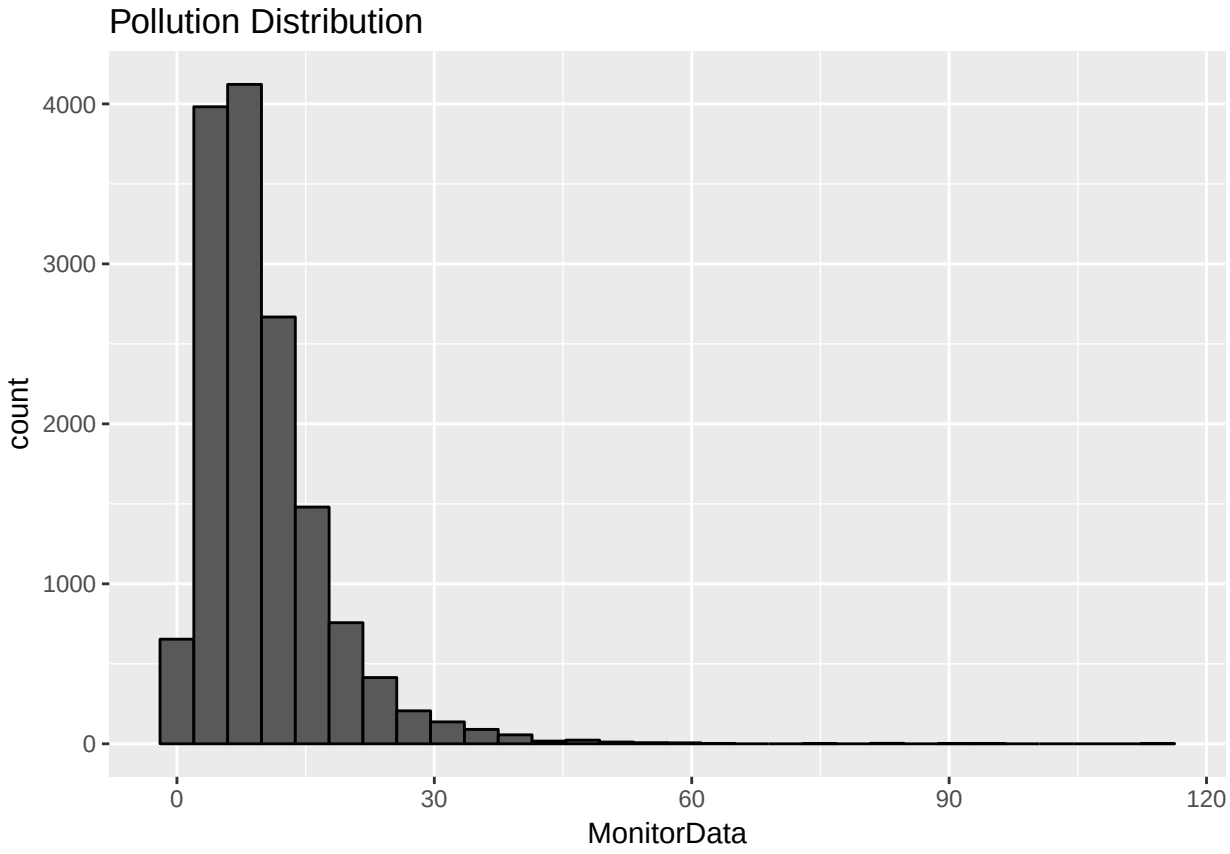
```r
#Monitor Data
describe(data$MonitorData, skew = F)
```

```
##    vars     n  mean   sd min    max  range   se
## X1    1 14639 10.04 7.39   0 114.32 114.32 0.06
```

```r
ggplot(data, aes(MonitorData)) +
  geom_histogram(col = 'black') +
```

```
    ggtitle('Pollution Distribution')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 52304 rows containing non-finite values (stat_bin).
```



Data appears to be consistent over time

```
#Date range
range(data$date)
```

```
## [1] "2000-01-01" "2016-12-31"
```

```
#Number of observations by year
data %>% count(year)
```

```
## # A tibble: 17 × 2
##     year     n
##    <int> <int>
## 1   2000  3969
## 2   2001  3915
## 3   2002  3842
## 4   2003  3882
## 5   2004  3965
## 6   2005  3909
## 7   2006  3872
## 8   2007  3901
## 9   2008  3948
```

```
## 10  2009  4026
## 11  2010  3832
## 12  2011  4041
## 13  2012  3825
## 14  2013  4011
## 15  2014  3933
## 16  2015  4029
## 17  2016  4043
```

```
#Proportion of missing monitor data by year
data %>% group_by(year) %>% summarize(missing.monitor = mean(is.na(MonitorData)))
```

```
## # A tibble: 17 × 2
##     year missing.monitor
##    <int>           <dbl>
## 1   2000       0.7986898
## 2   2001       0.7816092
## 3   2002       0.7727746
## 4   2003       0.8019062
## 5   2004       0.7977301
## 6   2005       0.8012279
## 7   2006       0.8070764
## 8   2007       0.8162010
## 9   2008       0.7983789
## 10  2009       0.7873820
## 11  2010       0.7802714
## 12  2011       0.7720861
## 13  2012       0.7728105
## 14  2013       0.7486911
## 15  2014       0.7332825
## 16  2015       0.7230082
## 17  2016       0.7914915
```

```
#Number of observations by month
data %>% count(month)
```

```
## # A tibble: 12 × 2
##    month     n
##    <dbl> <int>
## 1      1  5643
## 2      2  5208
## 3      3  5712
## 4      4  5594
## 5      5  5634
## 6      6  5412
## 7      7  5728
## 8      8  5666
## 9      9  5441
## 10    10  5604
## 11    11  5579
## 12    12  5722
```

```
#Proportion of missing monitor data by month
data %>% group_by(month) %>% summarize(missing.monitor = mean(is.na(MonitorData)))
```

```
## # A tibble: 12 × 2
```

```
##    month missing.monitor
##    <dbl>           <dbl>
## 1      1       0.7830941
## 2      2       0.7743856
## 3      3       0.7738095
## 4      4       0.7699321
## 5      5       0.7870075
## 6      6       0.7810421
## 7      7       0.7950419
## 8      8       0.7749735
## 9      9       0.7689763
## 10    10       0.7817630
## 11    11       0.7892095
## 12    12       0.7953513
```

```
#PM2.5 appears to be going doum over time!
data %>% group_by(year) %>% summarize(avg_pollution = mean(MonitorData, na.rm = T))
```

```
## # A tibble: 17 × 2
##     year avg_pollution
##    <int>         <dbl>
## 1   2000     12.794712
## 2   2001     11.895841
## 3   2002     11.966771
## 4   2003     11.449758
## 5   2004     10.852584
## 6   2005     12.163334
## 7   2006     10.843709
## 8   2007     11.748573
## 9   2008      9.767332
## 10  2009      9.584481
## 11  2010      9.562141
## 12  2011      9.574281
## 13  2012      8.826494
## 14  2013      8.495890
## 15  2014      8.006517
## 16  2015      8.143072
## 17  2016      7.259983
```

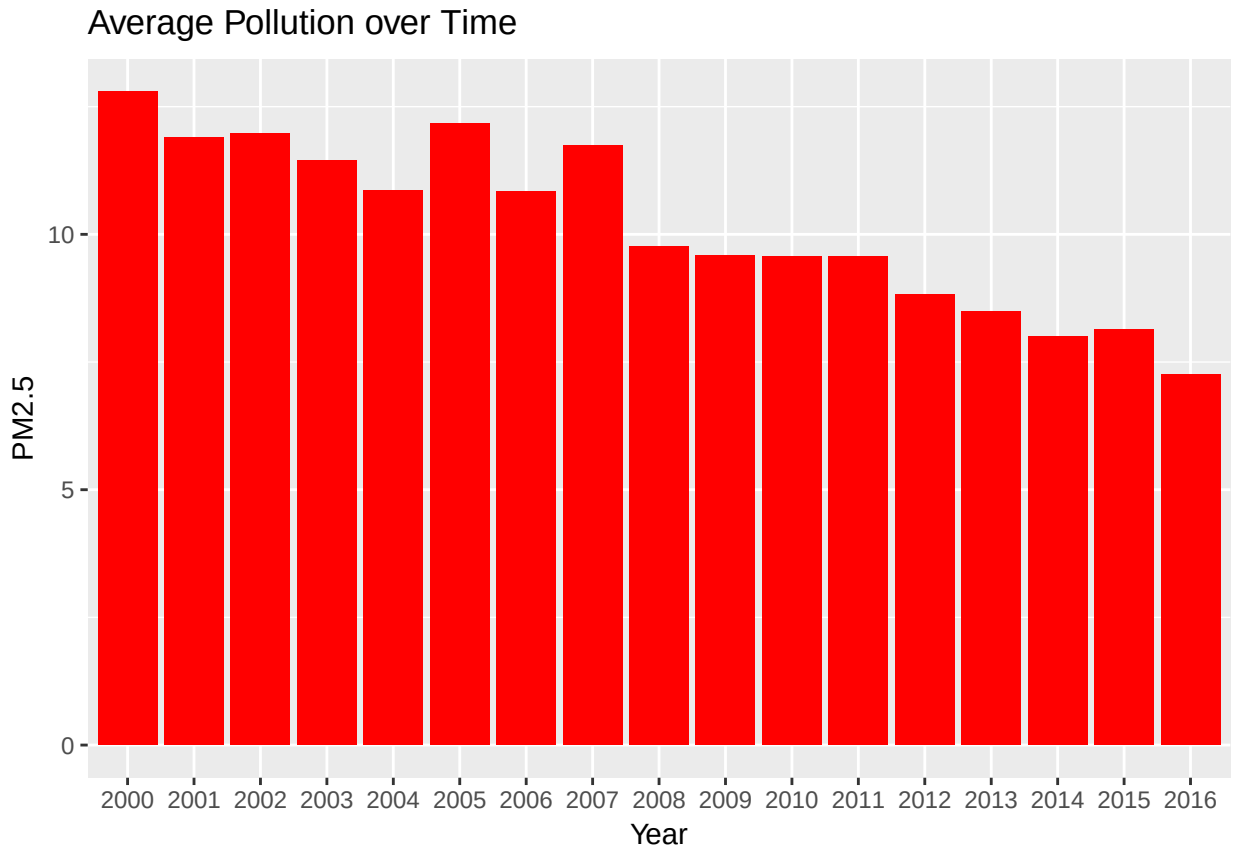
```
#PM2.5 levels also seem to vary by month
data %>% group_by(month) %>% summarize(avg_pollution = mean(MonitorData, na.rm = T))
```

```
## # A tibble: 12 × 2
##    month avg_pollution
##    <dbl>         <dbl>
## 1      1     11.178927
## 2      2     10.383440
## 3      3      9.308232
## 4      4      8.201018
## 5      5      9.205504
## 6      6     10.560507
## 7      7     12.129970
## 8      8     11.005990
## 9      9      9.407819
## 10    10      8.700584
## 11    11     10.139112
```

```
## 12    12     10.525167
```

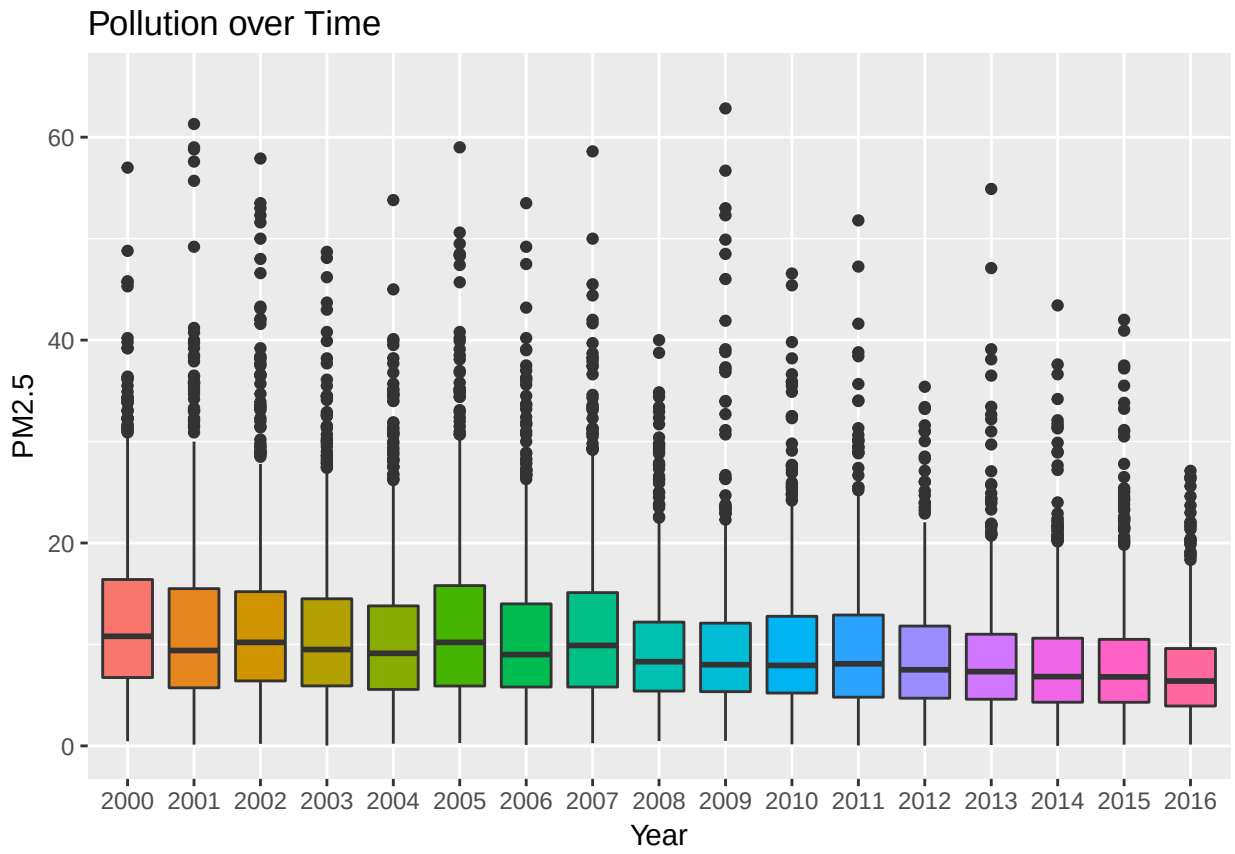
```
ggplot(data, aes(x = factor(year), y = MonitorData)) +
  stat_summary(fun.y = 'mean', geom = 'bar', fill = 'red') +
    ggtitle('Average Pollution over Time') + xlab('Year') + ylab('PM2.5')
```

```
## Warning: Removed 52304 rows containing non-finite values (stat_summary).
```



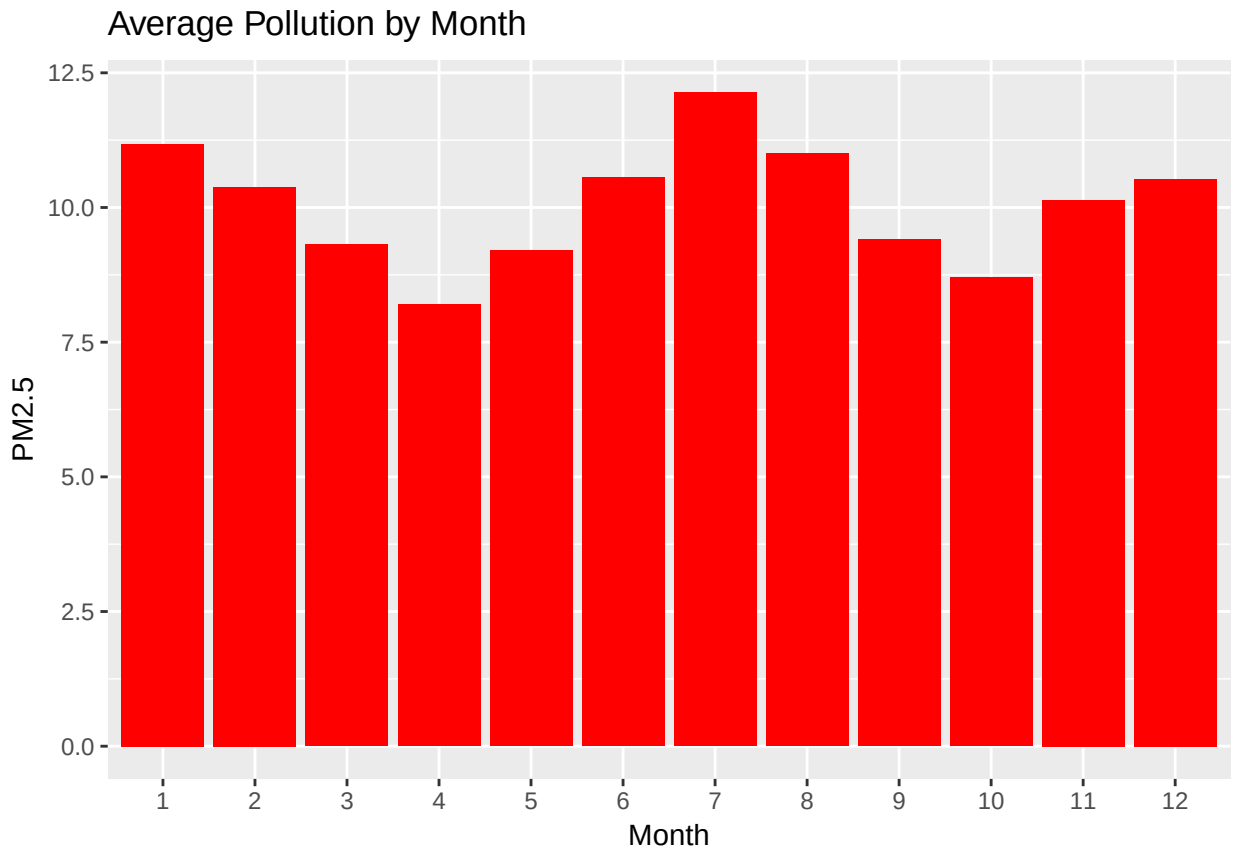

```
ggplot(data, aes(x = factor(year), y = MonitorData, fill = factor(year))) +
  geom_boxplot() + theme(legend.position = 'none') + ylim(0, 65) +
    ggtitle('Pollution over Time') + xlab('Year') + ylab('PM2.5')
```

```
## Warning: Removed 52311 rows containing non-finite values (stat_boxplot).
```



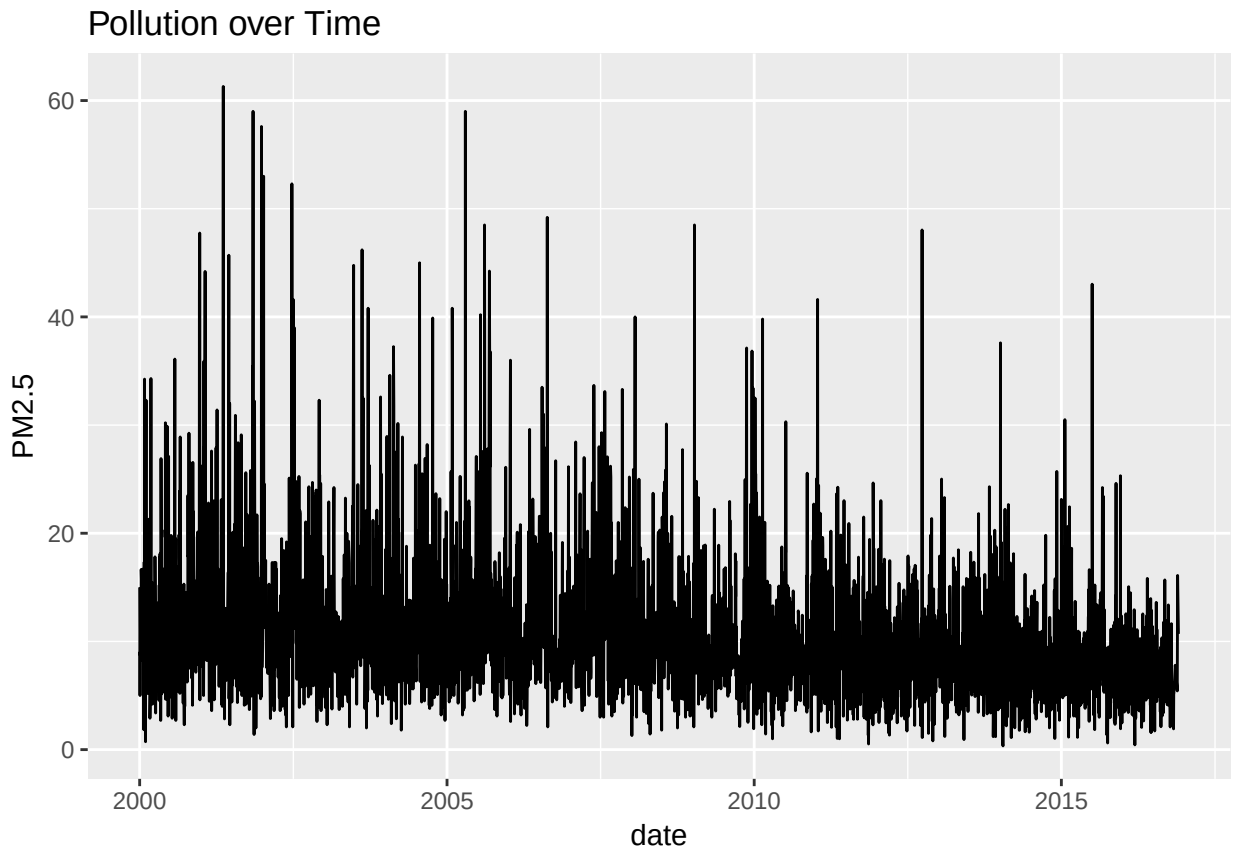

```
ggplot(data, aes(x = factor(month), y = MonitorData)) +
  stat_summary(fun.y = 'mean', geom = 'bar', fill = 'red') +
    ggtitle('Average Pollution by Month') + xlab('Month') + ylab('PM2.5')
```

```
## Warning: Removed 52304 rows containing non-finite values (stat_summary).
```




```
#Time Series of average PM2.5 levels by day over time
ggplot(data, aes(x = date, y = MonitorData)) +
  stat_summary(fun.y = 'mean', geom = 'line', size = .5) +
    ylab('PM2.5') + ggtitle('Pollution over Time')
```

```
## Warning: Removed 52304 rows containing non-finite values (stat_summary).
```



We will now explore the relationship between the various covariates and the PM2.5 monitor data

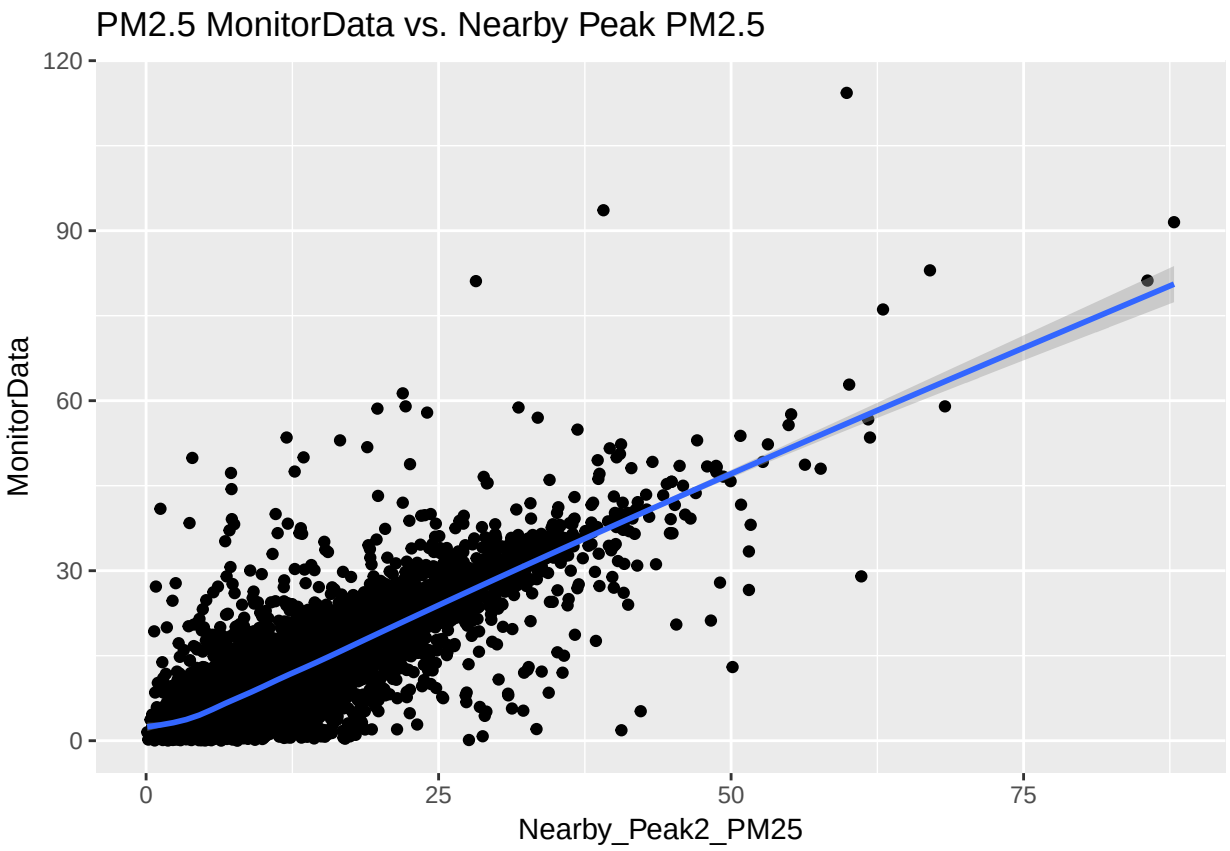
```
#Remove rows with missing monitor data and remove non-predictor variables
monitor.na.omit = data %>% filter(!is.na(MonitorData)) %>% select(-c(site:date, month))

#Look at correlation between each predictor and PM2.5 monitor data
num.cols = ncol(monitor.na.omit)
cors = matrix(rep(NA, 3*num.cols), nrow = num.cols)
colnames(cors) = c('variable', 'correlation', 'num.complete.cases')

for(i in 2:num.cols){

  monitor.na.omit2 = na.omit(monitor.na.omit[, c(1,i)])
  cors[i,2] = cor(monitor.na.omit2$MonitorData, monitor.na.omit2[,2])

  n = nrow(monitor.na.omit2)
  cors[i,3] = n
  #cors[i,4] = round(n/nrow(data), 3)

}

cors = as.data.frame(cors)
cors[,1] = names(monitor.na.omit)
```

```
#Variables with that are most correlated with PM2.5
cors %>% arrange(desc(abs(correlation))) %>% slice(1:20)
```

```
##                                  variable correlation num.complete.cases
## 1                        Nearby_Peak2_PM25   0.8593813              14639
## 2                     Nearby_Peak2Lag1_PM25   0.5967797              14639
## 3  MAIACUS_Optical_Depth_047_Terra_Nearest4   0.4217809               4047
## 4  MAIACUS_Optical_Depth_055_Terra_Nearest4   0.4145451               4047
## 5                     Nearby_Peak2Lag3_PM25   0.3848342              14639
## 6                          Nearby_Peak2_NO2   0.3374003              14639
## 7   MAIACUS_Optical_Depth_047_Aqua_Nearest4   0.3273920               3191
## 8                                MOD04L2_550   0.3265802               6659
## 9                      Nearby_Peak2Lag1_NO2   0.3263739              14639
## 10  MAIACUS_Optical_Depth_055_Aqua_Nearest4   0.3230416               3191
## 11                REANALYSIS_hpbl_DailyMean  -0.2949095              14638
## 12                 REANALYSIS_hpbl_DailyMin  -0.2795266              14638
## 13                     REANALYSIS_hpbl_1Day  -0.2525072              14638
## 14                       USElevation_max100  -0.2234779              14638
## 15                       USElevation_mea100  -0.2227626              14603
## 16                       USElevation_med100  -0.2225470              14630
## 17                       USElevation_bln100  -0.2216269              14623
## 18                     Nearby_Peak2Lag3_NO2   0.2212935              14639
## 19                       USElevation_min100  -0.2212438              14592
## 20                     USElevation_max10000  -0.2184453              14638
```

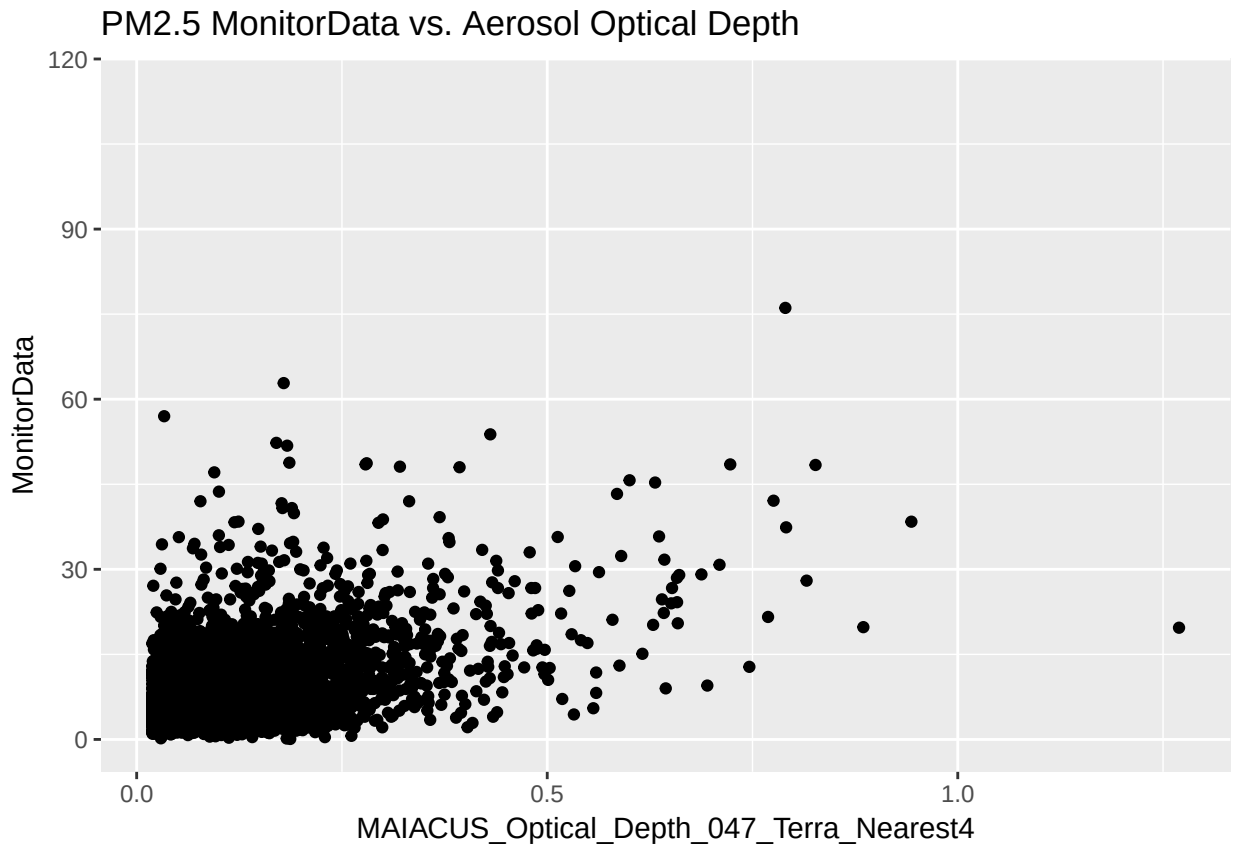
```
ggplot(monitor.na.omit, aes(x = Nearby_Peak2_PM25, y = MonitorData)) + geom_point() +
  ggtitle('PM2.5 MonitorData vs. Nearby Peak PM2.5') + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```

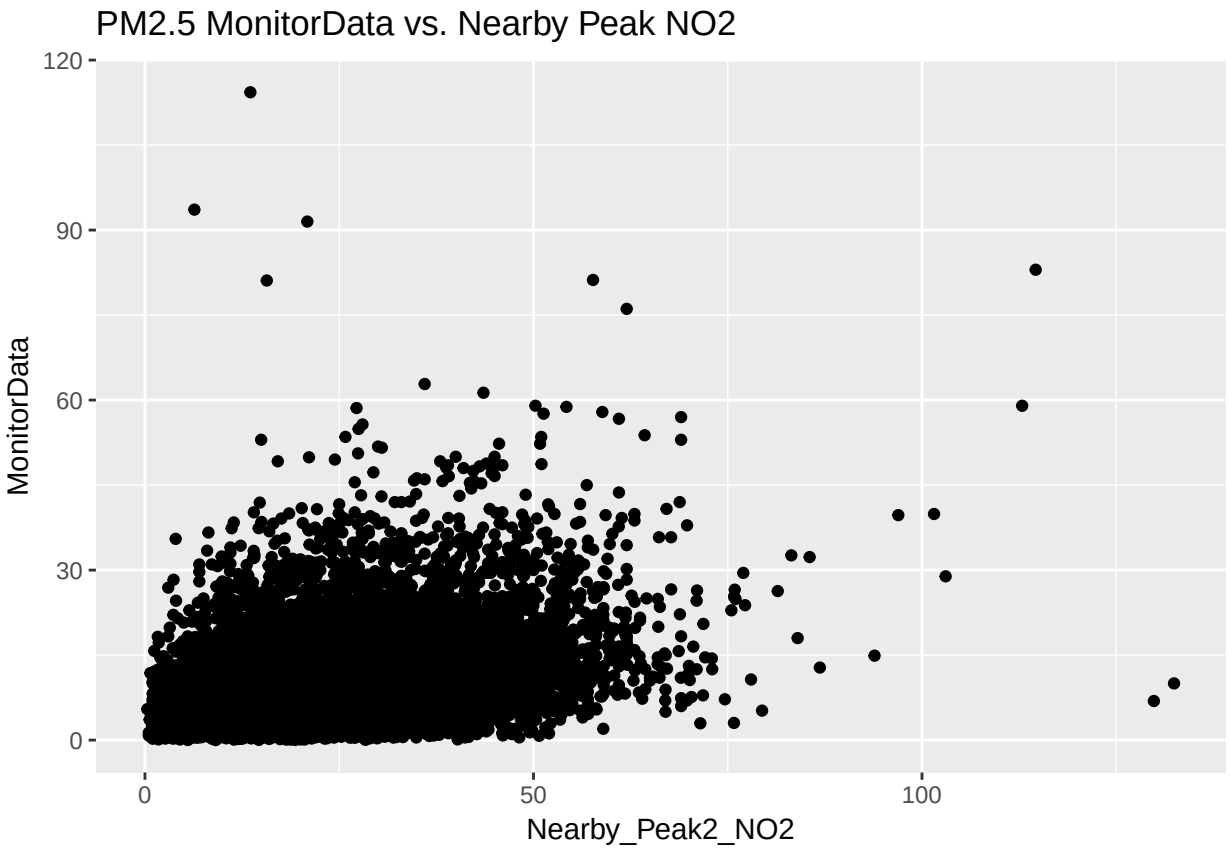


```
ggplot(monitor.na.omit, aes(x = MAIACUS_Optical_Depth_047_Terra_Nearest4, y = MonitorData)) +
  geom_point() + ggtitle('PM2.5 MonitorData vs. Aerosol Optical Depth')
```

```
## Warning: Removed 10592 rows containing missing values (geom_point).
```




```
ggplot(monitor.na.omit, aes(x = Nearby_Peak2_NO2, y = MonitorData)) +
  geom_point() + ggtitle('PM2.5 MonitorData vs. Nearby Peak NO2')
```


PM2.5 MonitorData vs. Nearby Peak NO2
MonitorData
120
90
60
30
0
0
50
100
Nearby_Peak2_NO2