Stat-415/615 Project - Tidying Data & EDA

Naomi Carrigg, Conor Gillingham, Emily Randolph, Connor Rempe

10/31/24

```
## cleaning/tidying the data
AirQuality_data_raw <- read_csv("Air_Quality_History.csv")
```

```
## Rows: 45505 Columns: 30
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (15): STATE_NAME, COUNTY_NAME, PARAMETER_NAME, DATETIME_LOCAL, DATUM, UN...
## dbl (15): AQSID, SITE_NUM, STATE_CODE, PARAMETER_CODE, POC, LATITUDE, LONGIT...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
AirQuality_data <- AirQuality_data_raw %>%
  mutate(across(starts_with("PARAMETER_NAME"),
                ~ str_replace_all(., " ", "_") %>%
                  str_to_lower() %>%
                  str_remove_all("[^a-z_]"))) %>%
  mutate(SITE_NAME = case_match(SITE_NUM,
                                41 ~ "River_Terrace_NE",
                                43 ~ "McMillan_NW",
                                50 ~ "Takoma_Recreation_NW",
                                51 ~ "Anacostia_Freeway_NE",
                                53 ~ "Greenleaf_Recreation_SW",
                                42 ~ "Hains_Point_SW")) %>%
  select(-c(LONGITUDE, LATITUDE, STATE_CODE, STATE_NAME, COUNTY_NAME, POC, DATUM, OBJECTID, UNITS_OF_MEA
  mutate(DATETIME_LOCAL = as.POSIXct(DATETIME_LOCAL, tz = "UTC"),
         Year = year(DATETIME_LOCAL),
         Month = month(DATETIME_LOCAL, label = TRUE, abbr = TRUE) %>% as.character()) %>%
  mutate(Season = case_when(
    Month %in% c("Dec", "Jan", "Feb") ~ 'Winter',
    Month %in% c("Mar", "Apr", "May") ~ 'Spring',
    Month %in% c("Jun", "Jul", "Aug") ~ 'Summer',
    TRUE ~ 'Fall'
  ))

# Grouping and summarizing
AirQuality_data_means <- AirQuality_data %>%
  group_by(Season, SITE_NAME, Month, Year, PARAMETER_NAME) %>%
  summarize(ARITHMETIC_MEAN = mean(ARITHMETIC_MEAN, na.rm = TRUE), .groups = 'drop')
```

```r
AirQuality_data_aqi <- AirQuality_data %>%
  group_by(Season, SITE_NAME, Month, Year, PARAMETER_NAME) %>%
  summarize(AQI = mean(AQI, na.rm = TRUE), .groups = 'drop')

# Pivot the data,
# this is where the issue is with the NAs  because there are some cases which don't have any data for t
AirQuality_data_means <- AirQuality_data_means %>%
  pivot_wider(names_from = PARAMETER_NAME, values_from = ARITHMETIC_MEAN)

AirQuality_data_aqi <- AirQuality_data_aqi %>%
  pivot_wider(names_from = PARAMETER_NAME, values_from = AQI)
```

```r
head(AirQuality_data_means)
```

```
## # A tibble: 6 x 92
##   Season SITE_NAME            Month  Year barometric_pressure carbon_monoxide
##   <chr>  <chr>                <chr> <dbl>               <dbl>           <dbl>
## 1 Fall   Anacostia_Freeway_NE Nov    2021               1018.           0.433
## 2 Fall   Anacostia_Freeway_NE Nov    2022               1019.           0.367
## 3 Fall   Anacostia_Freeway_NE Oct    2021               1012.           0.233
## 4 Fall   Anacostia_Freeway_NE Oct    2022               1015.           0.323
## 5 Fall   Anacostia_Freeway_NE Sep    2021               1013.           0.2
## 6 Fall   Anacostia_Freeway_NE Sep    2022               1014.           0.1
## # i 86 more variables: nitrogen_dioxide_no <dbl>, outdoor_temperature <dbl>,
## #   pm__local_conditions <dbl>, relative_humidity <dbl>,
## #   wind_direction__resultant <dbl>, wind_speed__resultant <dbl>,
## #   aluminum_pm_lc <dbl>, ammonium_ion_pm_lc <dbl>, antimony_pm_lc <dbl>,
## #   arsenic_pm_lc <dbl>, arsenic_pm_stp <dbl>, average_ambient_pressure <dbl>,
## #   average_ambient_pressure_for_urgn <dbl>, average_ambient_temperature <dbl>,
## #   average_ambient_temperature_for_urgn <dbl>, barium_pm_lc <dbl>, ...
```

```r
tail(AirQuality_data_means)
```

```
## # A tibble: 6 x 92
##   Season SITE_NAME            Month  Year barometric_pressure carbon_monoxide
##   <chr>  <chr>                <chr> <dbl>               <dbl>           <dbl>
## 1 Winter Takoma_Recreation_NW Feb    2021                  NA              NA
## 2 Winter Takoma_Recreation_NW Feb    2022                  NA              NA
## 3 Winter Takoma_Recreation_NW Feb    2023                  NA              NA
## 4 Winter Takoma_Recreation_NW Jan    2021                  NA              NA
## 5 Winter Takoma_Recreation_NW Jan    2022                  NA              NA
## 6 Winter Takoma_Recreation_NW Jan    2023                  NA              NA
## # i 86 more variables: nitrogen_dioxide_no <dbl>, outdoor_temperature <dbl>,
## #   pm__local_conditions <dbl>, relative_humidity <dbl>,
## #   wind_direction__resultant <dbl>, wind_speed__resultant <dbl>,
## #   aluminum_pm_lc <dbl>, ammonium_ion_pm_lc <dbl>, antimony_pm_lc <dbl>,
## #   arsenic_pm_lc <dbl>, arsenic_pm_stp <dbl>, average_ambient_pressure <dbl>,
## #   average_ambient_pressure_for_urgn <dbl>, average_ambient_temperature <dbl>,
## #   average_ambient_temperature_for_urgn <dbl>, barium_pm_lc <dbl>, ...
```

```r
head(AirQuality_data_aqi)
```

```
## # A tibble: 6 x 92
##   Season SITE_NAME            Month  Year barometric_pressure carbon_monoxide
##   <chr>  <chr>                <chr> <dbl>               <dbl>           <dbl>
## 1 Fall   Anacostia_Freeway_NE Nov    2021                 NaN            9.47
## 2 Fall   Anacostia_Freeway_NE Nov    2022                 NaN            8.85
## 3 Fall   Anacostia_Freeway_NE Oct    2021                 NaN            6.3
## 4 Fall   Anacostia_Freeway_NE Oct    2022                 NaN            8
## 5 Fall   Anacostia_Freeway_NE Sep    2021                 NaN            7.27
## 6 Fall   Anacostia_Freeway_NE Sep    2022                 NaN            6.56
## # i 86 more variables: nitrogen_dioxide_no <dbl>, outdoor_temperature <dbl>,
## #   pm__local_conditions <dbl>, relative_humidity <dbl>,
## #   wind_direction__resultant <dbl>, wind_speed__resultant <dbl>,
## #   aluminum_pm_lc <dbl>, ammonium_ion_pm_lc <dbl>, antimony_pm_lc <dbl>,
## #   arsenic_pm_lc <dbl>, arsenic_pm_stp <dbl>, average_ambient_pressure <dbl>,
## #   average_ambient_pressure_for_urgn <dbl>, average_ambient_temperature <dbl>,
## #   average_ambient_temperature_for_urgn <dbl>, barium_pm_lc <dbl>, ...
```

```r
tail(AirQuality_data_aqi)
```

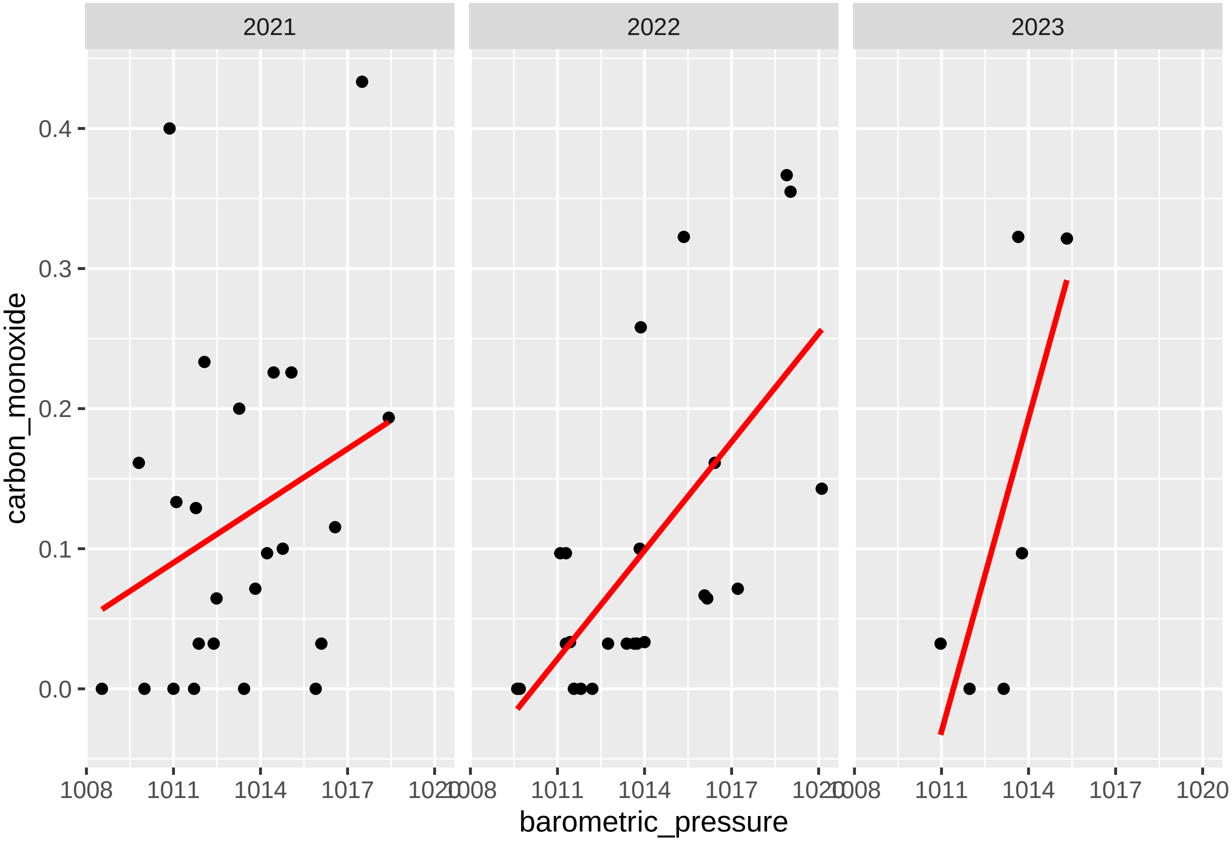
```
## # A tibble: 6 x 92
##   Season SITE_NAME            Month  Year barometric_pressure carbon_monoxide
##   <chr>  <chr>                <chr> <dbl>               <dbl>           <dbl>
## 1 Winter Takoma_Recreation_NW Feb    2021                  NA              NA
## 2 Winter Takoma_Recreation_NW Feb    2022                  NA              NA
## 3 Winter Takoma_Recreation_NW Feb    2023                  NA              NA
## 4 Winter Takoma_Recreation_NW Jan    2021                  NA              NA
## 5 Winter Takoma_Recreation_NW Jan    2022                  NA              NA
## 6 Winter Takoma_Recreation_NW Jan    2023                  NA              NA
## # i 86 more variables: nitrogen_dioxide_no <dbl>, outdoor_temperature <dbl>,
## #   pm__local_conditions <dbl>, relative_humidity <dbl>,
## #   wind_direction__resultant <dbl>, wind_speed__resultant <dbl>,
## #   aluminum_pm_lc <dbl>, ammonium_ion_pm_lc <dbl>, antimony_pm_lc <dbl>,
## #   arsenic_pm_lc <dbl>, arsenic_pm_stp <dbl>, average_ambient_pressure <dbl>,
## #   average_ambient_pressure_for_urgn <dbl>, average_ambient_temperature <dbl>,
## #   average_ambient_temperature_for_urgn <dbl>, barium_pm_lc <dbl>, ...
```

```r
#Columns with at least 1 non-NA AQI
AirQuality_data_aqi_reduced <- subset(AirQuality_data_aqi, select = c("Season", "SITE_NAME","Month", "Ye
```

```r
# means df with columns of interest
reduced_data <- AirQuality_data_means %>%
  subset(select = c("Season", "SITE_NAME","Month", "Year", "barometric_pressure", "carbon_monoxide", "ni
```

```r
reduced_data %>%
  ggplot(aes(x = barometric_pressure, y = carbon_monoxide)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE, color = 'red') +
  facet_wrap(~Year)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 97 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 97 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



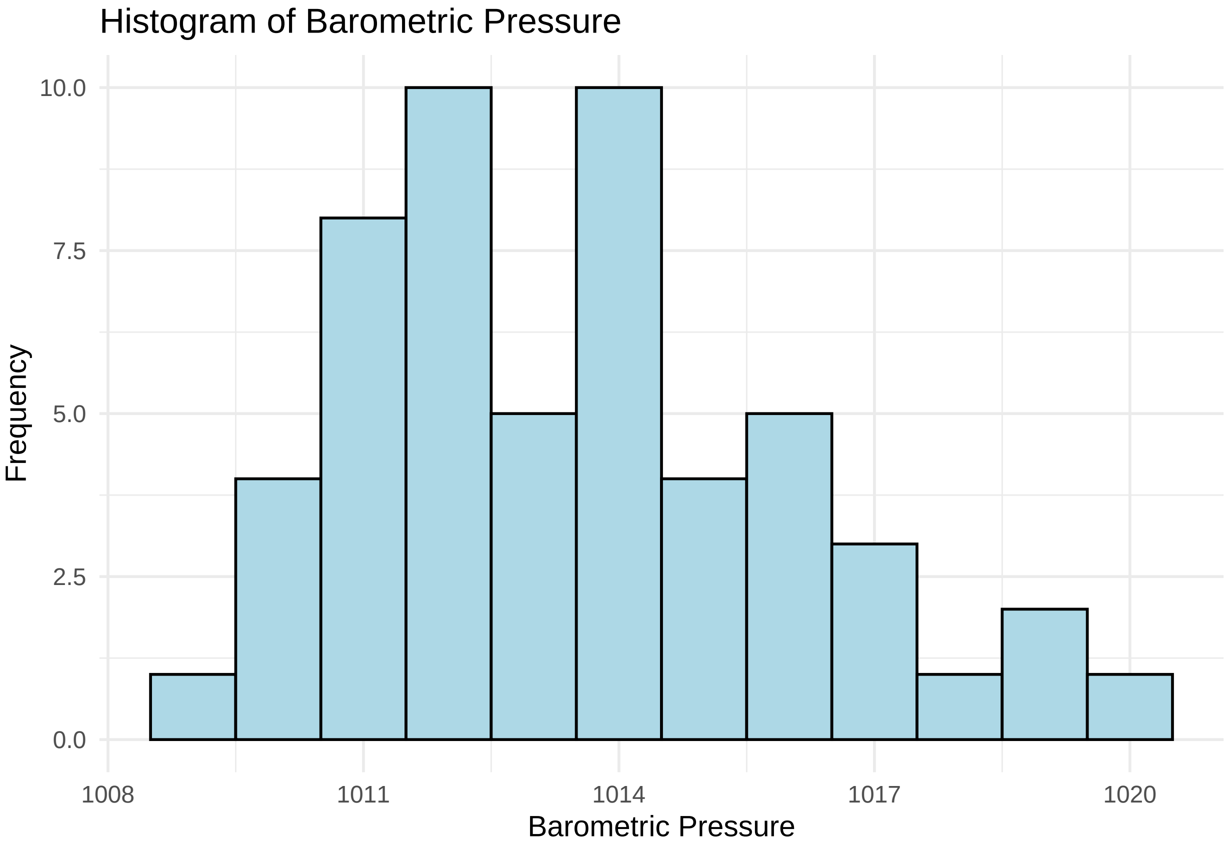
```
# Get Descriptive Statistics for numeric colmns in means
descriptive_stats_means <- data.frame()

for (col in names(AirQuality_data_means)) {
  if (is.numeric(AirQuality_data_means[[col]])) {
    working_means <- AirQuality_data_means %>%
      summarise(
        variable = col,
        mean = mean(.data[[col]], na.rm = TRUE),
        median = median(.data[[col]], na.rm = TRUE),
        sd = sd(.data[[col]], na.rm = TRUE),
        min = min(.data[[col]], na.rm = TRUE),
        max = max(.data[[col]], na.rm = TRUE),
        n = sum(!is.na(.data[[col]]))
      )
        descriptive_stats_means <- bind_rows(descriptive_stats_means, working_means)
  }
}

print(descriptive_stats_means)
```

```
##                                  variable         mean       median          sd
## 1                                    Year 2.021841e+03 2.022000e+03  0.78394486
## 2                     barometric_pressure 1.013526e+03 1.013410e+03  2.62212574
## 3                         carbon_monoxide 9.768574e-02 6.451613e-02  0.11486733
## 4                     nitrogen_dioxide_no 1.125621e+01 1.183871e+01  3.78513921
## 5                     outdoor_temperature 5.799322e+01 5.600000e+01 15.03198489
## 6                    pm__local_conditions 8.527484e+00 7.787097e+00  3.20137538
## 7                       relative_humidity 5.923924e+01 5.909301e+01  6.50243919
## 8               wind_direction__resultant 2.026929e+02 2.047419e+02 16.89839318
## 9                   wind_speed__resultant 3.819026e+00 3.942972e+00  2.05045478
## 10                         aluminum_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 11                     ammonium_ion_pm_lc 2.052940e-01 1.500000e-01  0.21492046
## 12                         antimony_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 13                          arsenic_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 14                         arsenic_pm_stp 7.450000e-01 6.000000e-01  0.49342710
## 15               average_ambient_pressure 7.579630e+02 7.578889e+02  2.08900995
## 16      average_ambient_pressure_for_urgn 7.552293e+02 7.542000e+02  3.12886875
## 17            average_ambient_temperature 1.501486e+01 1.400000e+01  8.26178769
## 18 average_ambient_temperature_for_urgn 1.547348e+01 1.415000e+01  8.29684644
## 19                           barium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 20                       beryllium_pm_stp 0.000000e+00 0.000000e+00  0.00000000
## 21                 black_carbon_pm_at__nm 2.997265e-01 2.909946e-01  0.18804425
## 22                          bromine_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 23                          cadmium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 24                         cadmium_pm_stp 2.000000e-02 0.000000e+00  0.06102572
## 25                          calcium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 26                           cerium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 27                           cesium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 28                         chloride_pm_lc 1.753247e-02 0.000000e+00  0.05431650
## 29                         chlorine_pm_lc 7.142857e-03 0.000000e+00  0.03779645
## 30                         chromium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 31                        chromium_pm_stp 1.813889e+00 1.366667e+00  1.19789153
## 32                           cobalt_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 33                           copper_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 34            ec_csn_rev_unadjusted_pm_lc 2.570192e-01 2.333333e-01  0.07793014
## 35        ec_csn_rev_unadjusted_pm_lc_tor 5.340651e-01 5.000000e-01  0.18558227
## 36        ec_csn_rev_unadjusted_pm_lc_tot 2.627860e-01 2.500000e-01  0.16494462
## 37                               ec_pm_lc 2.535560e-01 2.333333e-01  0.07806460
## 38                           ec_pm_lc_tor 5.340651e-01 5.000000e-01  0.18558227
## 39                           ec_pm_lc_tot 2.592146e-01 2.500000e-01  0.16293395
## 40                           indium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 41                             iron_pm_lc 3.968254e-03 0.000000e+00  0.02099803
## 42                             lead_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 43                            lead_pm_stp 0.000000e+00 0.000000e+00  0.00000000
## 44                        magnesium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 45                        manganese_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 46                       manganese_pm_stp 4.985556e+00 3.816667e+00  3.34405956
## 47                           nickel_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 48                          nickel_pm_stp 5.950000e-01 4.000000e-01  0.56193341
## 49            oc_csn_rev_unadjusted_pm_lc 3.799784e-01 3.623737e-01  0.14770238
## 50        oc_csn_rev_unadjusted_pm_lc_tor 2.182602e+00 2.050000e+00  0.60353634
## 51        oc_csn_rev_unadjusted_pm_lc_tot 2.348078e+00 2.190909e+00  0.63327244
## 52                               oc_pm_lc 3.363739e-01 3.250000e-01  0.13276410
## 53                           oc_pm_lc_tor 2.053257e+00 1.904545e+00  0.56631469
## 54                         oc_pm_lc_tot 2.217739e+00 2.100000e+00  0.55169705
## 55      op_csn_rev_unadjusted_pm_lc_tor 2.136724e-01 2.000000e-01  0.21213003
## 56      op_csn_rev_unadjusted_pm_lc_tot 4.929963e-01 4.000000e-01  0.26578964
## 57                         op_pm_lc_tor 1.846320e-01 1.055556e-01  0.20648771
## 58                         op_pm_lc_tot 4.718924e-01 4.000000e-01  0.24344100
## 59                                ozone 0.000000e+00 0.000000e+00  0.00000000
## 60                     phosphorus_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 61                      pm_total_um_stp 1.479723e+01 1.467742e+01  2.53761960
## 62                  potassium_ion_pm_lc 1.785714e-02 0.000000e+00  0.09449112
## 63                      potassium_pm_lc 1.785714e-02 0.000000e+00  0.09449112
## 64             reconstructed_mass_pm_lc 6.527404e+00 6.533333e+00  1.21055448
## 65                       rubidium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 66    sample_flow_rate_cv__nylon_filter 1.000000e+00 1.000000e+00  0.00000000
## 67   sample_flow_rate_cv__quartz_filter 0.000000e+00 0.000000e+00  0.00000000
## 68   sample_flow_rate_cv__teflon_filter 1.000000e+00 1.000000e+00  0.00000000
## 69          sample_volume__nylon_filter 9.992857e+00 1.000000e+01  0.04002906
## 70         sample_volume__quartz_filter 3.197850e+01 3.200000e+01  0.05006154
## 71         sample_volume__teflon_filter 9.988889e+00 1.000000e+01  0.03271023
## 72                       selenium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 73                        silicon_pm_lc 7.142857e-03 0.000000e+00  0.03779645
## 74                         silver_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 75                     sodium_ion_pm_lc 3.571429e-03 0.000000e+00  0.01889822
## 76                         sodium_pm_lc 3.246753e-03 0.000000e+00  0.01718020
## 77                           soil_pm_lc 3.318182e-01 2.474747e-01  0.29932938
## 78                      strontium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 79                        sulfate_pm_lc 9.672799e-01 1.000000e+00  0.20017042
## 80                       sulfur_dioxide 5.627035e-01 5.672970e-01  0.39219032
## 81                         sulfur_pm_lc 9.689755e-02 4.545455e-02  0.12226872
## 82                            tin_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 83                       titanium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 84                  total_nitrate_pm_lc 9.156836e-01 8.000000e-01  0.84026608
## 85                  uv_carbon_pm_at__nm 4.325175e-01 3.935484e-01  0.17855664
## 86                       vanadium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 87                           zinc_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 88                      zirconium_pm_lc 0.000000e+00 0.000000e+00  0.00000000
## 89        light_absorption_coeffiecient 0.000000e+00 0.000000e+00  0.00000000
##             min          max   n
## 1  2.021000e+03 2.023000e+03 151
## 2  1.008533e+03 1.020107e+03  54
## 3  0.000000e+00 4.333333e-01  66
## 4  4.935484e+00 1.925806e+01 119
## 5  3.338710e+01 8.090323e+01  54
## 6  5.233333e+00 2.366667e+01 120
## 7  4.745161e+01 7.277419e+01  54
## 8  1.569677e+02 2.383226e+02  54
## 9  1.032258e+00 8.064516e+00  54
## 10 0.000000e+00 0.000000e+00  28
## 11 0.000000e+00 8.888889e-01  28
## 12 0.000000e+00 0.000000e+00  28
## 13 0.000000e+00 0.000000e+00  28
## 14 2.000000e-01 2.200000e+00  30
## 15 7.538333e+02 7.624000e+02  28
## 16 7.513000e+02 7.628000e+02  28
## 17 1.600000e+00 2.740000e+01  28
## 18 2.700000e+00 2.770000e+01  28
## 19 0.000000e+00 0.000000e+00  28
## 20 0.000000e+00 0.000000e+00  30
## 21 6.451613e-02 7.419355e-01  24
## 22 0.000000e+00 0.000000e+00  28
## 23 0.000000e+00 0.000000e+00  28
## 24 0.000000e+00 2.000000e-01  30
## 25 0.000000e+00 0.000000e+00  28
## 26 0.000000e+00 0.000000e+00  28
## 27 0.000000e+00 0.000000e+00  28
## 28 0.000000e+00 2.000000e-01  28
## 29 0.000000e+00 2.000000e-01  28
## 30 0.000000e+00 0.000000e+00  28
## 31 1.000000e+00 5.400000e+00  30
## 32 0.000000e+00 0.000000e+00  28
## 33 0.000000e+00 0.000000e+00  28
## 34 1.481481e-01 5.333333e-01  28
## 35 1.818182e-01 8.888889e-01  28
## 36 0.000000e+00 6.000000e-01  28
## 37 1.481481e-01 5.333333e-01  28
## 38 1.818182e-01 8.888889e-01  28
## 39 0.000000e+00 6.000000e-01  28
## 40 0.000000e+00 0.000000e+00  28
## 41 0.000000e+00 1.111111e-01  28
## 42 0.000000e+00 0.000000e+00  28
## 43 0.000000e+00 0.000000e+00  30
## 44 0.000000e+00 0.000000e+00  28
## 45 0.000000e+00 0.000000e+00  28
## 46 1.600000e+00 1.800000e+01  30
## 47 0.000000e+00 0.000000e+00  28
## 48 0.000000e+00 2.600000e+00  30
## 49 1.500000e-01 8.250000e-01  28
## 50 1.500000e+00 4.600000e+00  28
## 51 1.500000e+00 4.800000e+00  28
## 52 1.136364e-01 7.500000e-01  28
## 53 1.400000e+00 4.200000e+00  28
## 54 1.500000e+00 4.300000e+00  28
## 55 0.000000e+00 1.100000e+00  28
## 56 1.818182e-01 1.500000e+00  28
## 57 0.000000e+00 1.000000e+00  28
## 58 1.818182e-01 1.400000e+00  28
## 59 0.000000e+00 0.000000e+00  87
## 60 0.000000e+00 0.000000e+00  28
## 61 1.096667e+01 2.096774e+01  27
## 62 0.000000e+00 5.000000e-01  28
## 63 0.000000e+00 5.000000e-01  28
## 64 4.714286e+00 9.111111e+00  28
## 65 0.000000e+00 0.000000e+00  28
## 66 1.000000e+00 1.000000e+00  28
## 67 0.000000e+00 0.000000e+00  28
## 68 1.000000e+00 1.000000e+00  28
## 69 9.888889e+00 1.011111e+01  28
## 70 3.180000e+01 3.200000e+01  28
## 71 9.888889e+00 1.000000e+01  28
## 72 0.000000e+00 0.000000e+00  28
## 73 0.000000e+00 2.000000e-01  28
## 74 0.000000e+00 0.000000e+00  28
## 75 0.000000e+00 1.000000e-01  28
## 76 0.000000e+00 9.090909e-02  28
## 77 0.000000e+00 1.400000e+00  28
## 78 0.000000e+00 0.000000e+00  28
## 79 4.000000e-01 1.444444e+00  28
## 80 0.000000e+00 1.129032e+00  32
## 81 0.000000e+00 4.000000e-01  28
## 82 0.000000e+00 0.000000e+00  28
## 83 0.000000e+00 0.000000e+00  28
## 84 0.000000e+00 3.333333e+00  28
## 85 1.935484e-01 7.333333e-01  24
## 86 0.000000e+00 0.000000e+00  28
## 87 0.000000e+00 0.000000e+00  28
## 88 0.000000e+00 0.000000e+00  28
## 89 0.000000e+00 0.000000e+00  12
```

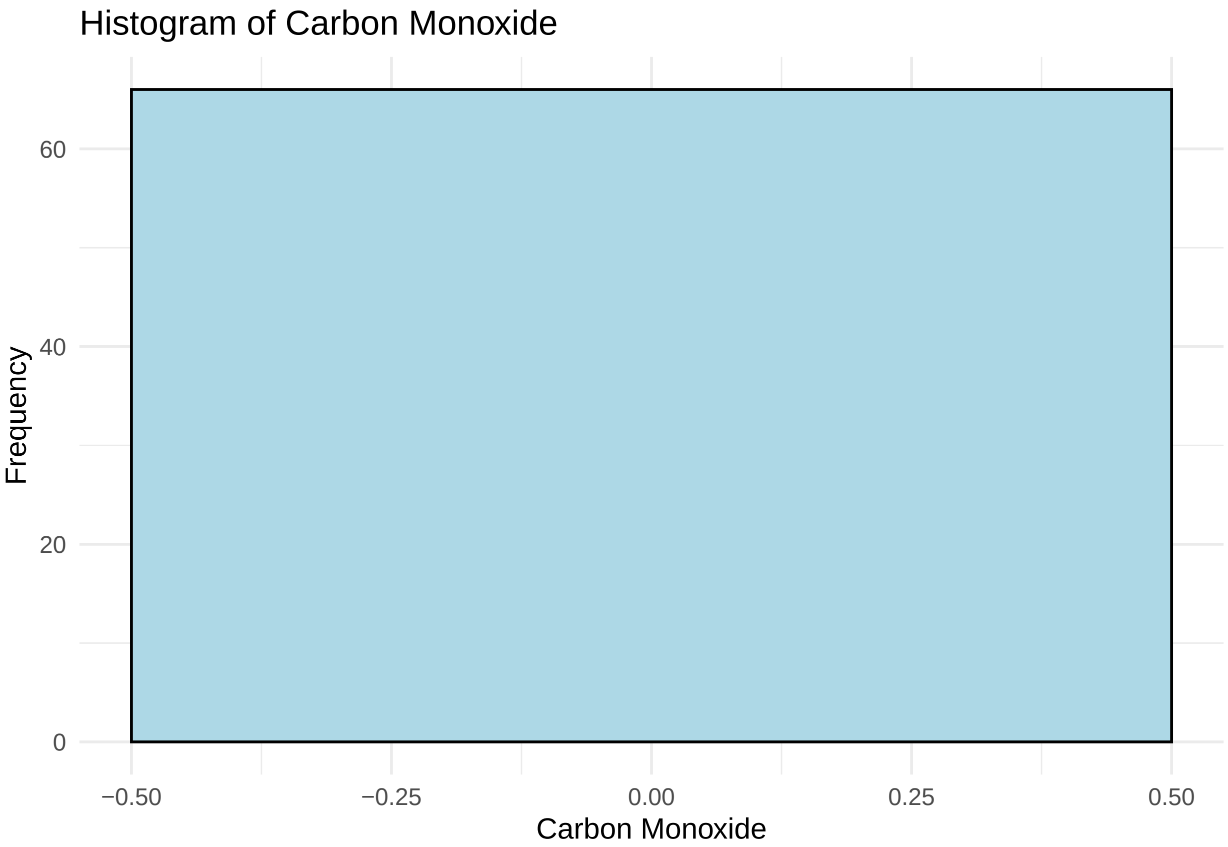
```
# Descriptive stats for AQI
descriptive_stats_aqi <- data.frame()

for (col in names(AirQuality_data_aqi_reduced)) {
  if (is.numeric(AirQuality_data_aqi_reduced[[col]])) {
    working_means_aqi <- AirQuality_data_aqi_reduced %>%
      summarise(
        variable = col,
        mean = mean(.data[[col]], na.rm = TRUE),
        median = median(.data[[col]], na.rm = TRUE),
        sd = sd(.data[[col]], na.rm = TRUE),
        min = min(.data[[col]], na.rm = TRUE),
        max = max(.data[[col]], na.rm = TRUE),
        n = sum(!is.na(.data[[col]]))
      )
        descriptive_stats_aqi <- bind_rows(descriptive_stats_aqi, working_means_aqi)
  }
}

print(descriptive_stats_aqi)
```

```
##                 variable         mean       median         sd         min
## 1                   Year 2021.8410596 2022.0000000  0.7839449 2021.000000
## 2       carbon_monoxide    5.6218548    5.8585608  1.7061617    2.000000
## 3  nitrogen_dioxide_no   20.7754686   21.8064516  5.7207652    9.290323
## 4 pm__local_conditions   33.6653438   31.4589744  9.4344155   18.454545
## 5                  ozone   35.6366324   36.0322581 10.8655649    7.741935
## 6       pm_total_um_stp   13.7440122   13.7741935  2.5238856    9.176471
## 7        sulfur_dioxide    0.5682122    0.5104167  0.3597947    0.000000
##           max   n
## 1 2023.000000 151
## 2    9.888889  66
## 3   32.161290 119
## 4   72.750000 120
## 5   60.933333  87
## 6   19.516129  27
## 7    1.166667  20
```

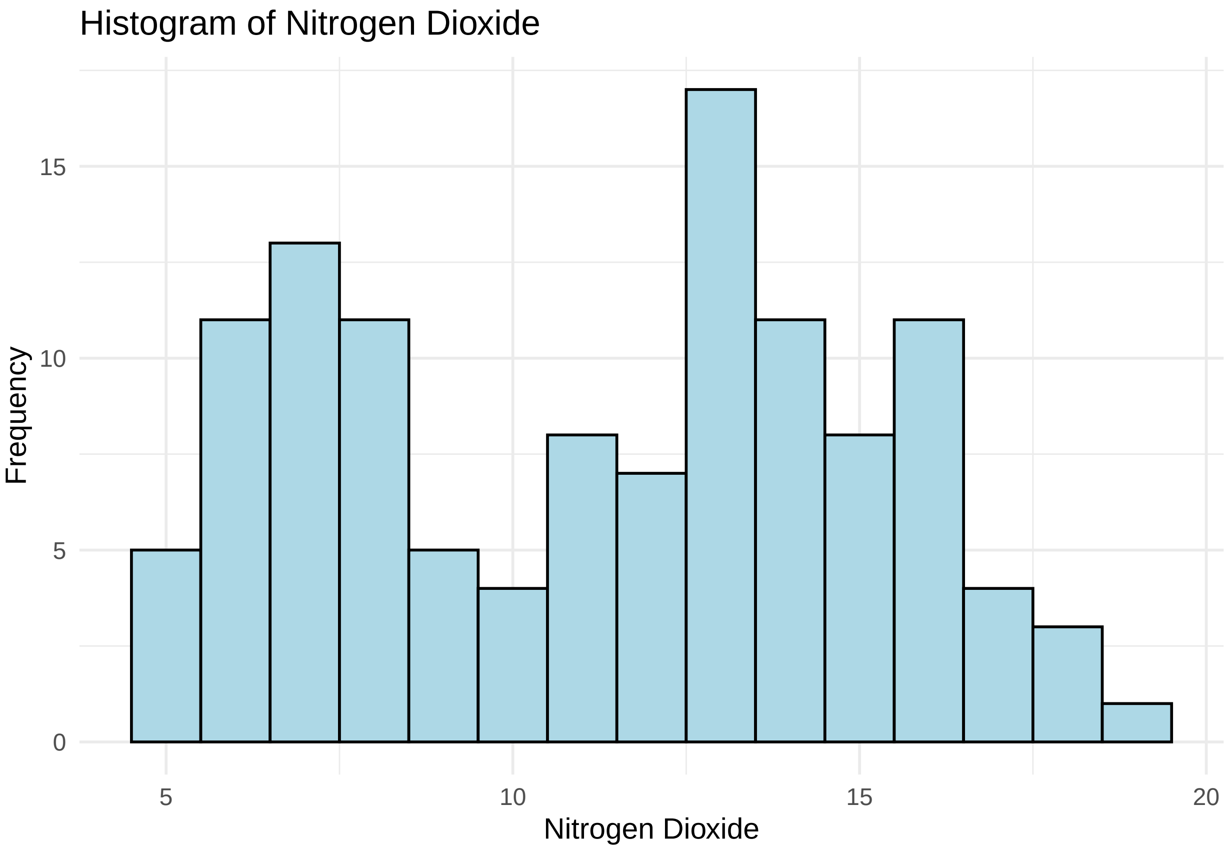
```
# Create a list of histograms for the specified columns
graph <- function(pollutant, name) {
  ggplot(reduced_data, aes(x = pollutant)) +
    geom_histogram(binwidth = 1, fill = 'lightblue', color = 'black', na.rm = TRUE) +
    labs(title = paste0('Histogram of ', name), x = name, y = 'Frequency') +
    theme_minimal()
}

graph(reduced_data$barometric_pressure, 'Barometric Pressure')
```



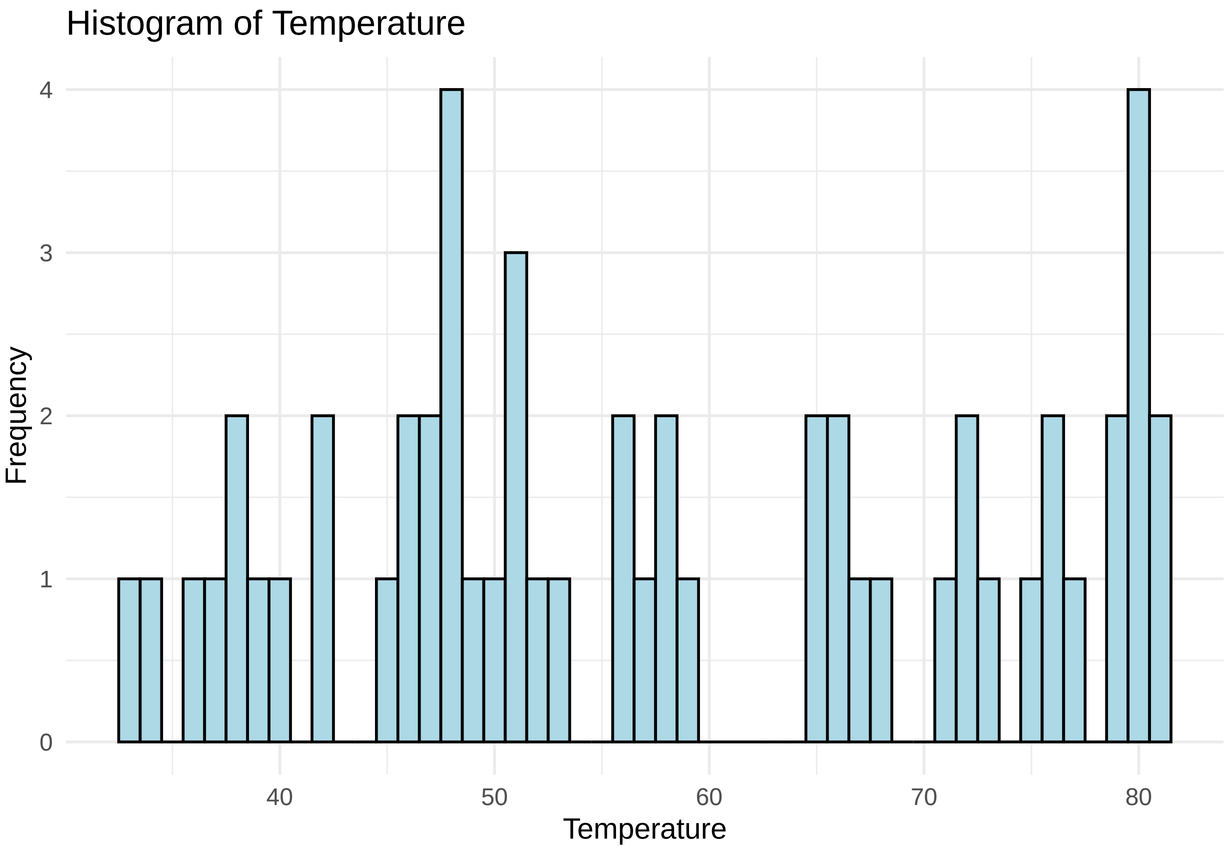
```
graph(reduced_data$carbon_monoxide, 'Carbon Monoxide')
```



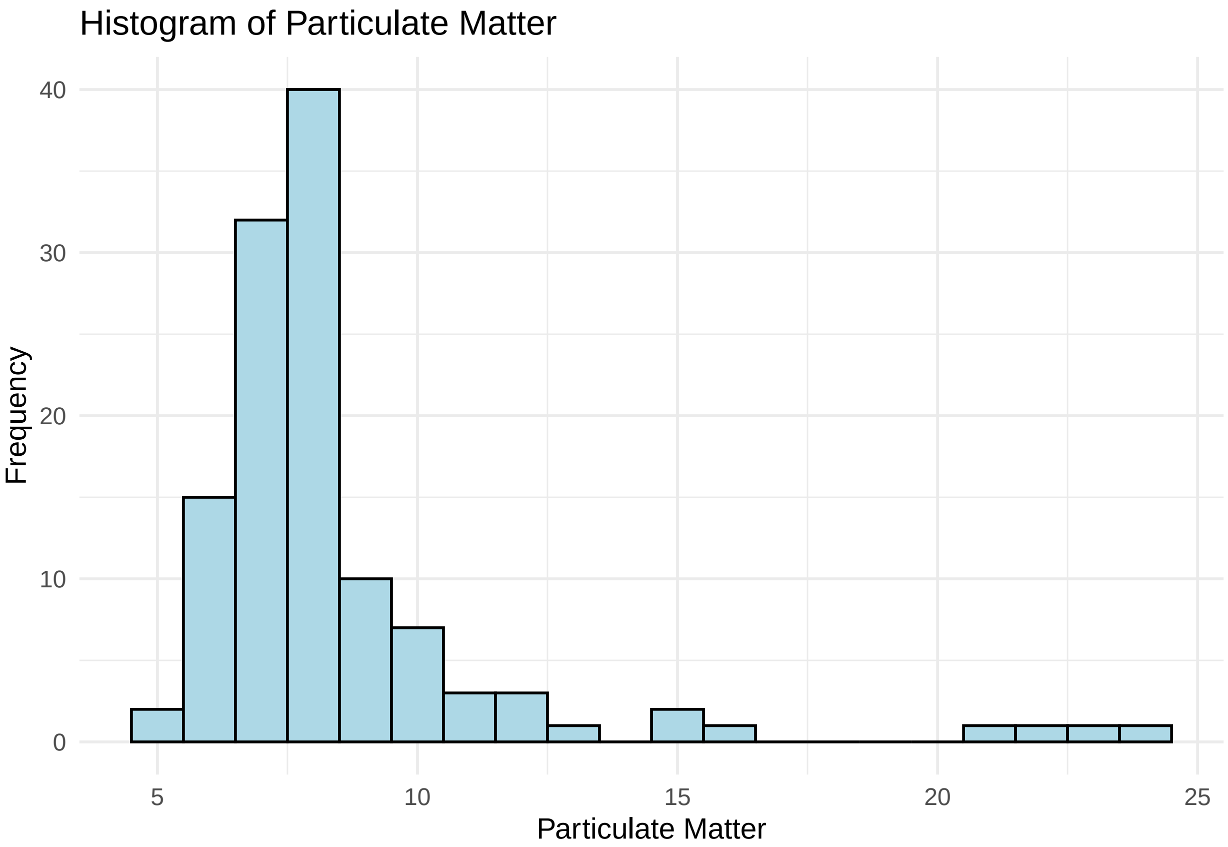
```
graph(reduced_data$nitrogen_dioxide_no, 'Nitrogen Dioxide')
```



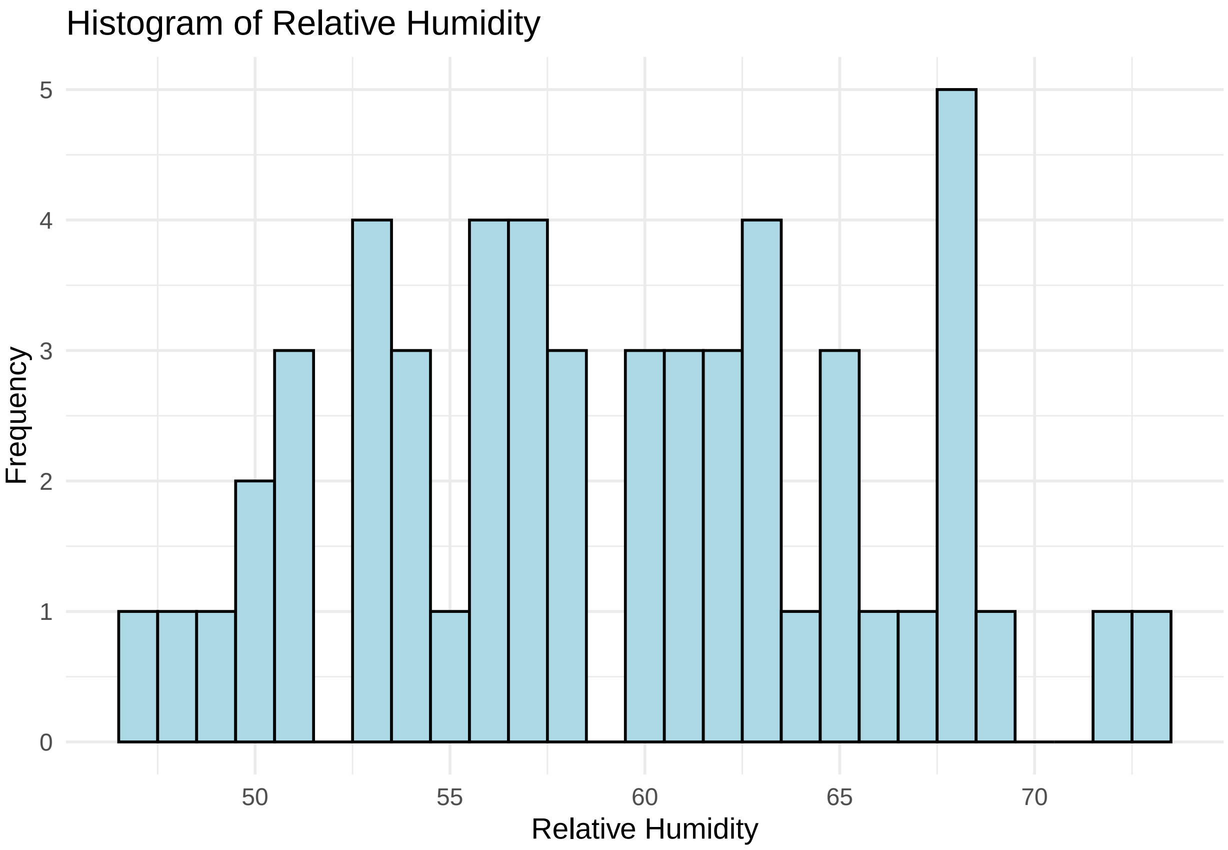
```
graph(reduced_data$outdoor_temperature, 'Temperature')
```



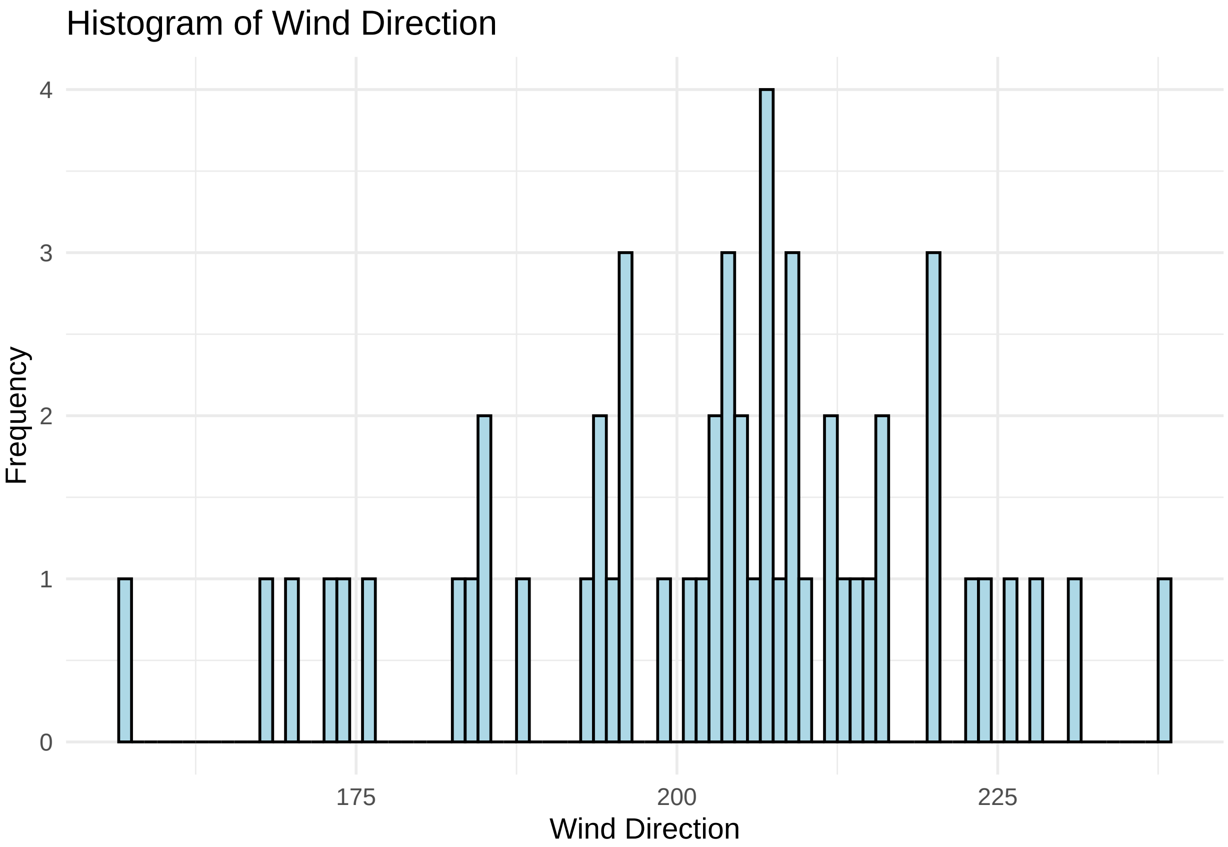
```
graph(reduced_data$pm__local_conditions, 'Particulate Matter')
```



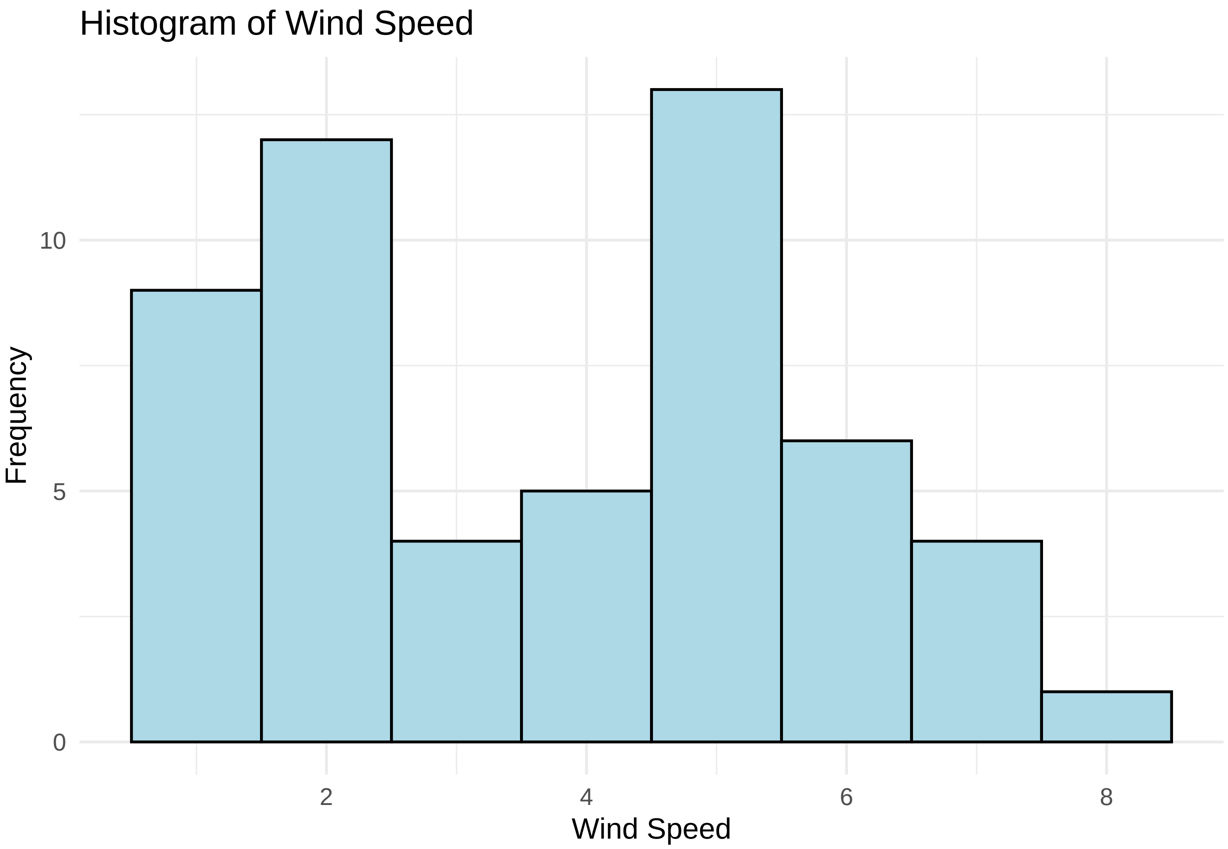
```r
graph(reduced_data$relative_humidity, 'Relative Humidity')
```



```
graph(reduced_data$wind_direction__resultant, 'Wind Direction')
```



```
graph(reduced_data$wind_speed__resultant, 'Wind Speed')
```



Model Fitting

Model 1: Predicting Carbon Monoxide with Nitrogen Dioxide and season using AQI

```
model_1_data <- na.omit(subset(AirQuality_data_aqi_reduced, select = c("Season", "Month", "Year", "SITE
model_1 <- lm(carbon_monoxide ~ Season + nitrogen_dioxide_no, data = model_1_data)
summary(model_1)
```

```
##
## Call:
## lm(formula = carbon_monoxide ~ Season + nitrogen_dioxide_no,
##     data = model_1_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9513 -0.7807 -0.1537  0.4701  3.1542
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.37003    0.62718   3.779 0.000369 ***
## SeasonSpring           -1.91869    0.39292  -4.883 8.32e-06 ***
## SeasonSummer        -0.80663    0.40476  -1.993 0.050910 .
## SeasonWinter        -1.35124    0.41845  -3.229 0.002030 **
## nitrogen_dioxide_no  0.20580    0.02577   7.985 5.81e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.052 on 59 degrees of freedom
## Multiple R-squared:  0.6131, Adjusted R-squared:  0.5869
## F-statistic: 23.37 on 4 and 59 DF,  p-value: 1.303e-11
```

Hypothesis test (Can Nitrogen be dropped from the model)

```
reduced <- lm(carbon_monoxide ~ Season, data = model_1_data)
anova(reduced, model_1)
```

```
## Analysis of Variance Table
##
## Model 1: carbon_monoxide ~ Season
## Model 2: carbon_monoxide ~ Season + nitrogen_dioxide_no
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     60 135.732
## 2     59  65.237  1    70.495 63.755 5.808e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null is that the coefficient is zero, at significance level of 0.05 we reject the null and say that carbon monoxide cannot be dropped from the model. Suggests that two pollutants do occur in tandem

Hypothesis test for Overall Significance

```
null <- lm(carbon_monoxide ~ 1 , data = model_1_data)
anova(null,model_1)
```

```
## Analysis of Variance Table
##
## Model 1: carbon_monoxide ~ 1
## Model 2: carbon_monoxide ~ Season + nitrogen_dioxide_no
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     63 168.613
## 2     59  65.237  4    103.38 23.373 1.303e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model is significant overall.

Model 2: Predicting Nitrogen Dioxide AQI using all other predictors. Trying to see if relationship exists in the opposite direction and whether location has an effect.

```
model_2_data <- na.omit(subset(AirQuality_data_aqi_reduced, select = c("Season", "Month", "Year", "SITE_

model_2_red <- lm(nitrogen_dioxide_no ~ SITE_NAME + Season + Month + Year, data = model_1_data)

model_2_full <- lm(nitrogen_dioxide_no ~ SITE_NAME + Season + Month + Year + carbon_monoxide, data = mo

anova(model_2_red, model_2_full)
```

```
## Analysis of Variance Table
##
## Model 1: nitrogen_dioxide_no ~ SITE_NAME + Season + Month + Year
## Model 2: nitrogen_dioxide_no ~ SITE_NAME + Season + Month + Year + carbon_monoxide
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     50 372.29
## 2     49 369.27  1    3.0202 0.4008 0.5296
```

Carbon monoxide can be dropped from the model when we include season, month, year, and Site_Name. Now we perform best subset selection on this model

```r
step(model_2_red, direction = "both")
```

```
## Start:  AIC=140.69
## nitrogen_dioxide_no ~ SITE_NAME + Season + Month + Year
##
##
## Step:  AIC=140.69
## nitrogen_dioxide_no ~ SITE_NAME + Month + Year
##
##             Df Sum of Sq     RSS    AIC
## - Year       1      4.17  376.46 139.40
## <none>                    372.29 140.69
## - Month     11    867.63 1239.92 195.69
## - SITE_NAME  1   1041.42 1413.71 224.09
##
## Step:  AIC=139.4
## nitrogen_dioxide_no ~ SITE_NAME + Month
##
##             Df Sum of Sq     RSS    AIC
## <none>                    376.46 139.40
## + Year       1      4.17  372.29 140.69
## - Month     11    872.48 1248.95 194.16
## - SITE_NAME  1   1041.42 1417.88 222.27
```

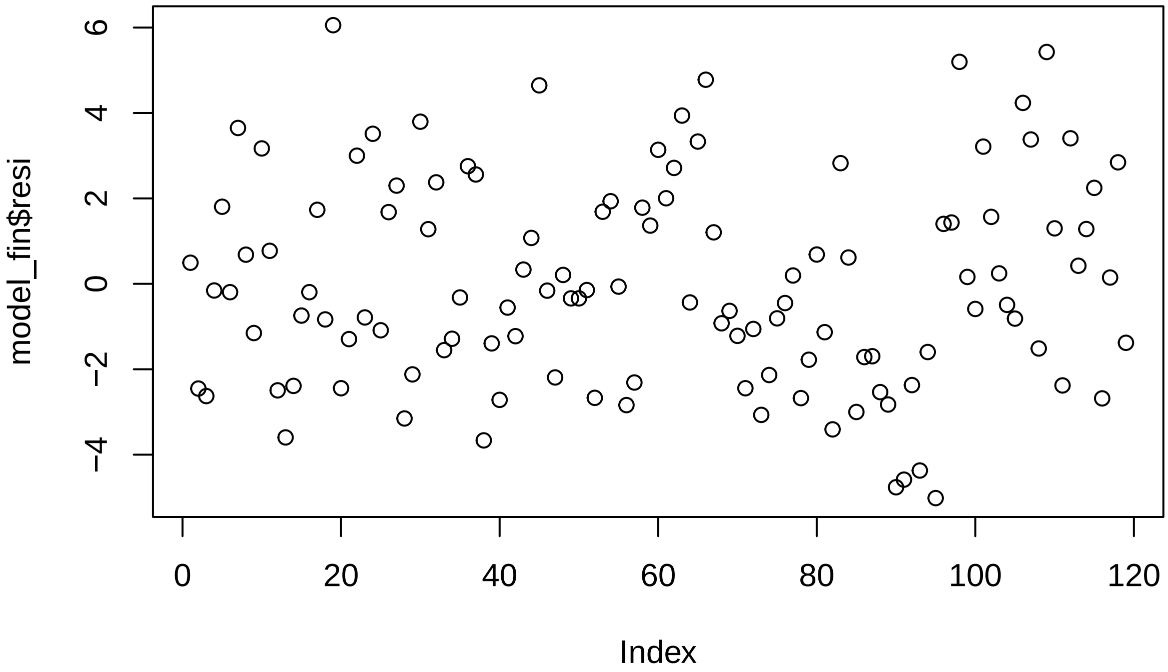
```
##
## Call:
## lm(formula = nitrogen_dioxide_no ~ SITE_NAME + Month, data = model_1_data)
##
## Coefficients:
##          (Intercept)  SITE_NAMEMcMillan_NW              MonthAug
##               25.445                -8.068                -3.658
##             MonthDec              MonthFeb              MonthJan
##                4.524                 5.267                 3.578
##             MonthJul              MonthJun              MonthMar
##               -4.986                -3.499                 4.196
##             MonthMay              MonthNov              MonthOct
##               -1.228                 4.500                -1.642
##             MonthSep
##               -3.211
```

Only site name and month are retained, using Anacostia NE and April as baselines

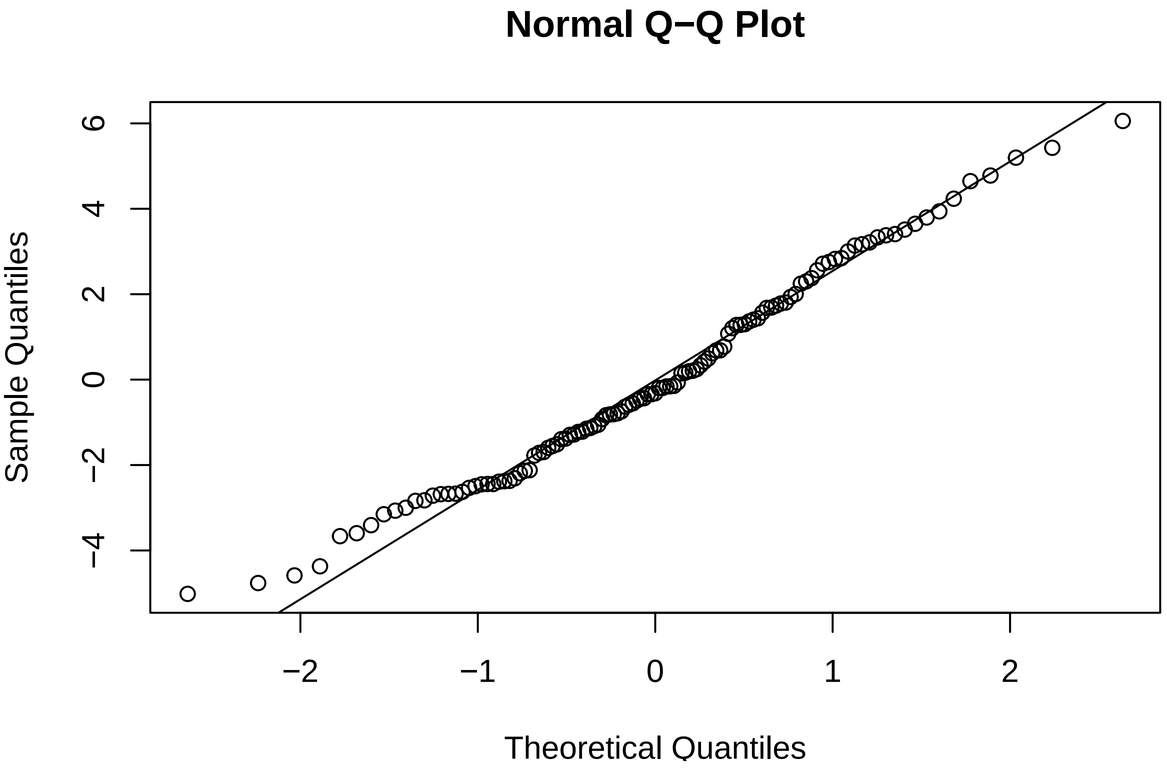
```
model_fin <- lm(nitrogen_dioxide_no ~ SITE_NAME + Month, data = model_2_data)
summary(model_fin)
```

```
##
## Call:
## lm(formula = nitrogen_dioxide_no ~ SITE_NAME + Month, data = model_2_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0165 -1.7460 -0.3195  1.7102  6.0560
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    25.2872     0.8450  29.925  < 2e-16 ***
## SITE_NAMEMcMillan_NW           -8.0678     0.6439 -12.530  < 2e-16 ***
## SITE_NAMERiver_Terrace_NE      -5.9283     0.6439  -9.207 3.97e-15 ***
## SITE_NAMETakoma_Recreation_NW  -7.4745     0.7125 -10.491  < 2e-16 ***
## MonthAug                       -4.4894     1.1066  -4.057 9.63e-05 ***
## MonthDec                        5.3128     1.1756   4.519 1.65e-05 ***
## MonthFeb                        6.5834     1.0515   6.261 8.71e-09 ***
## MonthJan                        4.7293     1.0515   4.498 1.79e-05 ***
## MonthJul                       -5.4850     1.1066  -4.957 2.79e-06 ***
## MonthJun                       -3.6178     1.1066  -3.269 0.001462 **
## MonthMar                        4.5750     1.0515   4.351 3.17e-05 ***
## MonthMay                       -1.5345     1.1066  -1.387 0.168484
## MonthNov                        4.0646     1.1756   3.458 0.000791 ***
## MonthOct                       -1.2926     1.1756  -1.100 0.274058
## MonthSep                       -3.0253     1.2267  -2.466 0.015285 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.576 on 104 degrees of freedom
## Multiple R-squared:  0.8214, Adjusted R-squared:  0.7973
## F-statistic: 34.15 on 14 and 104 DF,  p-value: < 2.2e-16
```

```
plot(model_fin$resi)
```



```
qqnorm(model_fin$resi)
qqline(model_fin$resi)
```



Residual plots indicate that linear model is appropriate.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
dwtest(model_fin)
```

```
##
##  Durbin-Watson test
##
## data:  model_fin
## DW = 1.554, p-value = 0.0004329
## alternative hypothesis: true autocorrelation is greater than 0
```

Caveat, DW test reveals autocorrelation in this mod

Model 3: Does weather have an affect?

```
model_3_data <- na.omit(subset(reduced_data, select = c("Season", "Month", "Year", "SITE_NAME", "nitrog

dim(model_3_data)
```

```
## [1] 54 10
```

```r
model_3 <- lm(nitrogen_dioxide_no ~ ., data = model_3_data)
summary(model_3)
```

```
##
## Call:
## lm(formula = nitrogen_dioxide_no ~ ., data = model_3_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.49364 -0.85579  0.07087  0.72594  1.91021
##
## Coefficients: (3 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1349.41567  737.71655   1.829  0.07590 .
## SeasonSpring                -0.25424    1.45586  -0.175  0.86238
## SeasonSummer                -1.00657    1.03302  -0.974  0.33655
## SeasonWinter                 1.81272    2.32565   0.779  0.44095
## MonthAug                     1.00280    0.97751   1.026  0.31199
## MonthDec                     3.08157    0.97862   3.149  0.00335 **
## MonthFeb                     1.25646    0.83696   1.501  0.14226
## MonthJan                          NA         NA      NA       NA
## MonthJul                    -0.52508    0.94720  -0.554  0.58287
## MonthJun                          NA         NA      NA       NA
## MonthMar                     2.16001    0.98412   2.195  0.03490 *
## MonthMay                     0.52864    1.19336   0.443  0.66051
## MonthNov                     3.99520    1.75348   2.278  0.02891 *
## MonthOct                     0.24209    1.16241   0.208  0.83623
## MonthSep                          NA         NA      NA       NA
## Year                        -0.66704    0.36504  -1.827  0.07619 .
## SITE_NAMEMcMillan_NW        -8.58799    1.34564  -6.382 2.43e-07 ***
## carbon_monoxide             -5.31438    2.25566  -2.356  0.02420 *
## outdoor_temperature         -0.10897    0.06489  -1.679  0.10197
## relative_humidity            0.17517    0.06608   2.651  0.01198 *
## wind_direction__resultant    0.04458    0.01595   2.796  0.00836 **
## wind_speed__resultant        0.44480    0.33829   1.315  0.19711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.249 on 35 degrees of freedom
## Multiple R-squared:  0.9304, Adjusted R-squared:  0.8946
## F-statistic: 25.98 on 18 and 35 DF,  p-value: 4.235e-15
```

```r
step(model_3, direction = "both")
```

```
## Start:  AIC=38.57
## nitrogen_dioxide_no ~ Season + Month + Year + SITE_NAME + carbon_monoxide +
##     outdoor_temperature + relative_humidity + wind_direction__resultant +
##     wind_speed__resultant
##
##
## Step:  AIC=38.57
## nitrogen_dioxide_no ~ Month + Year + SITE_NAME + carbon_monoxide +
##     outdoor_temperature + relative_humidity + wind_direction__resultant +
##     wind_speed__resultant
##
##                             Df Sum of Sq     RSS    AIC
## <none>                                    54.574 38.571
## - wind_speed__resultant      1     2.696  57.270 39.175
## - outdoor_temperature        1     4.398  58.972 40.756
## - Year                       1     5.206  59.781 41.492
## - carbon_monoxide            1     8.655  63.229 44.520
## - relative_humidity          1    10.957  65.531 46.451
## - wind_direction__resultant  1    12.186  66.760 47.454
## - Month                     11    56.168 110.742 54.784
## - SITE_NAME                  1    63.511 118.085 78.251
```

```
##
## Call:
## lm(formula = nitrogen_dioxide_no ~ Month + Year + SITE_NAME +
##     carbon_monoxide + outdoor_temperature + relative_humidity +
##     wind_direction__resultant + wind_speed__resultant, data = model_3_data)
##
## Coefficients:
##               (Intercept)                   MonthAug
##                1349.16143                    0.25047
##                  MonthDec                   MonthFeb
##                   5.14853                    3.32342
##                  MonthJan                   MonthJul
##                   2.06696                   -1.27741
##                  MonthJun                   MonthMar
##                  -0.75233                    2.16001
##                  MonthMay                   MonthNov
##                   0.52864                    4.24944
##                  MonthOct                   MonthSep
##                   0.49633                    0.25424
##                      Year       SITE_NAMEMcMillan_NW
##                  -0.66704                   -8.58799
##           carbon_monoxide        outdoor_temperature
##                  -5.31438                   -0.10897
##         relative_humidity  wind_direction__resultant
##                   0.17517                    0.04458
##     wind_speed__resultant
##                   0.44480
```

```r
model_3_step <- lm(formula = nitrogen_dioxide_no ~ Month + Year + SITE_NAME +
    carbon_monoxide + outdoor_temperature + relative_humidity +
    wind_direction__resultant + wind_speed__resultant, data = model_3_data)
summary(model_3_step)
```

```
##
## Call:
## lm(formula = nitrogen_dioxide_no ~ Month + Year + SITE_NAME +
##     carbon_monoxide + outdoor_temperature + relative_humidity +
##     wind_direction__resultant + wind_speed__resultant, data = model_3_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.49364 -0.85579  0.07087  0.72594  1.91021
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1349.16143  737.74679   1.829 0.075968 .
## MonthAug                     0.25047    1.86651   0.134 0.894018
## MonthDec                     5.14853    1.35250   3.807 0.000544 ***
## MonthFeb                     3.32342    1.36326   2.438 0.019994 *
## MonthJan                     2.06696    1.60646   1.287 0.206659
## MonthJul                    -1.27741    1.86334  -0.686 0.497515
## MonthJun                    -0.75233    1.59829  -0.471 0.640769
## MonthMar                     2.16001    0.98412   2.195 0.034897 *
## MonthMay                     0.52864    1.19336   0.443 0.660506
## MonthNov                     4.24944    1.17841   3.606 0.000959 ***
## MonthOct                     0.49633    1.34171   0.370 0.713670
## MonthSep                     0.25424    1.45586   0.175 0.862377
## Year                        -0.66704    0.36504  -1.827 0.076191 .
## SITE_NAMEMcMillan_NW        -8.58799    1.34564  -6.382 2.43e-07 ***
## carbon_monoxide             -5.31438    2.25566  -2.356 0.024201 *
## outdoor_temperature         -0.10897    0.06489  -1.679 0.101975
## relative_humidity            0.17517    0.06608   2.651 0.011976 *
## wind_direction__resultant    0.04458    0.01595   2.796 0.008356 **
## wind_speed__resultant        0.44480    0.33829   1.315 0.197109
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.249 on 35 degrees of freedom
## Multiple R-squared:  0.9304, Adjusted R-squared:  0.8946
## F-statistic: 25.98 on 18 and 35 DF,  p-value: 4.235e-15
```

Anova test for significance of weather predictors

```
model_3_red <- lm(formula = nitrogen_dioxide_no ~ Month + Year + SITE_NAME +
    carbon_monoxide, data = model_3_data)

anova(model_3_red, model_3_step)
```

```
## Analysis of Variance Table
##
## Model 1: nitrogen_dioxide_no ~ Month + Year + SITE_NAME + carbon_monoxide
## Model 2: nitrogen_dioxide_no ~ Month + Year + SITE_NAME + carbon_monoxide +
##     outdoor_temperature + relative_humidity + wind_direction__resultant +
##     wind_speed__resultant
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     39 78.592
## 2     35 54.574  4    24.018 3.8508 0.01074 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All weather variables cannot be dropped from the model, weather has some impact on concentration.

Overall findings

1. Concentration of other pollutants doesn't have a significant effect when Location and Month are included in the model
2. In general, winter and fall months tend to have higher AQI
3. Season and Site location are strong predictors, suggests disparities accross DC. NE worse off than NW.
4. It appears that weather has an affect