Analysis of predictors for the concentration of fine particulate matter (PM2.5) in North Carolina during the year 2018.

https://github.com/fr55/DataAnalytics_FinalProject

Felipe Raby Amadori

Abstract

In this study, the main focus was in studying PM2.5 concentrations data in North Carolina during the year 2018, and other air pollutants, meteorological, and geographical parameters. Through the analysis, it was found that there is a significant difference between the mean daily PM 2.5 concentrations in the Piedmont and both the Mountains and the Coastal zone. It was also found that there is a significant linear relationship between the concentrations of PM2.5 and the concentrations of PM10, County Population, Daily Average Temperature, and Elevation. The distance to electricity generation combustion sites was also considered, but it was found that there is not a significant relationship with PM 2.5 concentrations measured in the North Carolina EPA stations.

Contents

List of Tables

List of Figures

1 Research Question and Rationale

Nowadays air pollution is one of the most relevant health issues in the world. It refers to the contamination of the air by chemicals, biological materials, and other types of pollutants that are harmful to human health. To solve the problem of air pollution, it's necessary to understand the problem, what are the causes, and search for solutions based on the findings.

Particulate matter with a diameter of less than 2.5 micrometers is called PM2.5, and it is a extremely harmful air pollutant because it consists of particles with diameters that are less than or equal to 2.5 microns in size, which can get deeply into the lung, and ultimately impair lung function.

This study focus on trying to understand how PM2.5 concentration in North Carolina vary with temperature, PM10 concentration, zoning (piedmont, coastal, mountain), population, elevation, and distance to combustion points for electricity generation. This last variable was included because according to the EPA combustion for electricity generation is the major point-source sector for PM2.5 in the USA (EPA, 2019).

The research question is: What are the effects of temperature, PM10 concentration, zoning (piedmont, coastal, mountain), population, elevation, distance to combustion points for electricity generation, in PM2.5 concentrations within North Carolina in the year 2018?

2 Dataset Information

For the analysis the following datasets were considered:

2.1 EPA PM2.5 Dataset

This dataset contains data from air quality monitoring of PM2.5 in North Carolina in 2018, and it was obtained using the Download Daily Data Tool in the United States Environmental Protection Agency (EPA) webpage https://www.epa.gov/outdoor-air-quality-data/download-daily-data where the options showed in Table 1 were selected:

Option	Selection
Pollutant	PM2.5
Year	2018
Geographic Area	North Carolina
Monitor Site	All Sites
Download	Download CSV (spreadsheet)

Table 1: Selections

The downloaded file was saved in the project folder path ./Data/Raw/ as EPAair_PM25_NC2018_raw.csv on 2019-03-31.

2.1.1 Data Content Information

The dataset contains daily mean PM2.5 concentration in ug/m3 in 2018. Data from 24 stations in 21 different counties of North Carolina with their location in NAD83 lat/long coordinates.

The dataset contains 19 columns, which are shown in Table 2. Column names without description are self-explanatory.

Column	Description
Date	mm/dd/YY
Source	AQS (Air Quality System)
Site ID	A unique number identifying the site.
POC	"Parameter Occurrence Code", distinguishes different instruments that measure the same parameter at the same site.
Daily Mean PM2.5 Concentration	
Units	Concentration Units
DAILY_AQI_VALUE	AQI = Air quality index
Site Name	
DAILY_OBS_COUNT	
PERCENT_COMPLETE	
AQS_PARAMETER_CODE	
AQS_PARAMETER_DESC	
CBSA_CODE	
CBSA_NAME	
STATE_CODE	A unique number identifying the State.
STATE	
COUNTY CODE	A unique number identifying the County.
COUNTY	
SITE_LATITUDE	NAD83
SITE_LONGITUDE	NAD83

Table 2: Dataset content

2.2 EPA PM10 Dataset

This dataset contains data from air quality monitoring of PM10 in North Carolina in 2018, and it was obtained using the Download Daily Data Tool in the United States Environmental Protection Agency (EPA) webpage https://www.epa.gov/outdoor-air-quality-data/download-daily-data where the options showed in Table 3 were selected:

Option	Selection
Pollutant	PM10
Year	2018
Geographic Area	North Carolina
Monitor Site	All Sites
Download	Download CSV (spreadsheet)

Table 3: Selections

The downloaded file was saved in the project folder path ./Data/Raw/ as EPAair_PM10_NC2018_raw.csv on 2019-03-31.

2.2.1 Data Content Information

The dataset contains daily mean PM10 concentration in ug/m3 in 2018. Data from 9 stations in 8 different counties of North Carolina with their location in NAD83 lat/long coordinates.

The dataset contains 19 columns, which are shown in Table 4. Column names without description are self-explanatory.

Column	Description
Date	mm/dd/YY
Source	AQS (Air Quality System)
Site ID	A unique number identifying the site.
POC	"Parameter Occurrence Code", distinguishes different instruments that measure the same parameter at the same site.
Daily Mean PM10 Concentration	
Units	Concentration Units
DAILY_AQI_VALUE	AQI = Air quality index
Site Name	
DAILY_OBS_COUNT	
PERCENT_COMPLETE	
AQS_PARAMETER_CODE	
AQS_PARAMETER_DESC	
CBSA_CODE	
CBSA_NAME	
STATE_CODE	A unique number identifying the State.
STATE	
COUNTY CODE	A unique number identifying the County.
COUNTY	
SITE_LATITUDE	NAD83
SITE_LONGITUDE	NAD83

Table 4: Dataset content

2.3 NOAA Average Temperature Dataset

This dataset contains data from temperature monitoring in North Carolina in 2018, and it was obtained using the Search Tool in the National Center for Environmental Information of the National Oceanic and Atmospheric Administration (NOAA) Webpage https://www.ncdc.noaa.gov/cdo-web. Options showed in Table 5 were selected:

The downloaded file was saved in the project folder path ./Data/Raw/ as NOAA_TAVG_NC2018_raw.csv on 2019-03-28.

Option	Selection
Discover Data By	Search Tool
Weather Observation Dataset	Daily Summaries
Date Range	2018-01-01 to 2018-12-31
Search For	States
Search Term	North Carolina
Output Format	Custom GHCN-Daily CSV
Station Detail	Station Name & Geographic Location
Data Type	Average Temperature. (TAVG)

Table 5: Selections

2.3.1 Data Content Information

The dataset contains daily mean air temperature in Fahrenheit in 2018. Data from 39 stations in North Carolina with their location in NAD83 lat/long coordinates. No county information.

The dataset contains 7 columns, which are shown in Table 6. Column names without description are self-explanatory.

Column	Description
STATION	A unique code identifying the site.
NAME	Station Name
Site ID	A unique number identifying the site.
LATITUDE	NAD83
LONGITUDE	NAD83
DATE	dd/mm/YY
TAVG	Daily Average Temperature in °F

Table 6: Dataset content

2.4 US Census Bureau US counties shapefile

This dataset contains geographic and geometric information of all the counties of the US. The data is in NAD83 lat/long coordinates. The file was provided by John Fay in the Environmental Data Analytics (ENV 872L) course at Duke University, Spring 2019.

The files containing the information were saved in the project folder path ./Data/Spatial/ as cb_2017_us_county_20m on 2019-03-28.

2.4.1 Data Content Information

The dataset contains geographic and geometric information of all the counties of the US in NAD83 lat/long coordinates.

The dataset contains 10 columns, which are shown in Table 7. Column names without description are self-explanatory.

Column	Description
STATEFP	A unique number identifying the State.
COUNTYFP	County Federal Information Processing Standards (FIPS) Code
COUNTYNS	Provides the American National Standards Institute (ANSI) code for the county or equivalent entity, as used by GNIS.
AFFGEOID	AFF Summary Level Code
GEOID	NAD83
NAME	County Name
LSAD	Legal/statistical area description
ALAND	County Land Area in square meters
AWATER	County Water Area in square meters
Geometry	Geometry and geographic information

Table 7: Dataset content

2.5 EPA combustion points for electricity generation in the US Dataset

This dataset contains facility-level locations for combustion points for electricity generation in the US, and it was obtained from the United States Environmental Protection Agency (EPA) webpage https://www3.epa.gov/air/emissions/where.htm. The Top PM2.5 emitting sectors link was selected.

The downloaded file was saved in the project folder path ./Data/Raw/ as EPA_ElecGenComb_US_raw.kml on 2019-03-31.

2.5.1 Data Content Information

The dataset is a kml file that contains combustion points for electricity generation in the US. The data is in WGS84 lat/long coordinates.

All data sets, variable, and files are named according to the following naming convention: *databasename_datatype_details_stage.format*, where:

- databasename refers to the database from where the data originated
- datatype is a description of data
- details are additional descriptive details, particularly important for processed data

- stage refers to the stage in data management pipelines (e.g., raw, cleaned, temp or processed)
- format is a non-proprietary file format (e.g., .csv, .txt)

2.6 Analized data structure

With these datasets an exploratory data analysis was done for the study. The datasets were wrangled in a dataframe called PM2.5_Full_Elev_utm and stored in the project processed folder in a file called PM2.5_Full_Elev_utm.shp. This dataframe has the data structure shown in Table 8.

Variable	Units	N.Elements	Range	Source.File
Date	YY-mm-dd	343	From 2018-01-01 to 2018-12-09	EPAair_PM25_NC2018_raw.csv
Site_ID	-	24	-	EPAair_PM25_NC2018_raw.csv
COUNTY	-	21	-	EPAair_PM25_NC2018_raw.csv
Population	People	21	From 5,507 to 1,034,290	https://en.wikipedia.org/
Zone	-	3	Coastal, Piedmont, and Mountains	NC County Maps
PM2_5	ug/m3	6499	From -2.5 to 34.2	EPAair_PM25_NC2018_raw.csv
PM10	ug/m3	926	From 0 to 35	EPAair_PM10_NC2018_raw.csv
TAVG	Farenheit	4011	From 11 to 87	NOAA_TAVG_NC2018_raw.csv
Emiss_Dist	meters	24	From 813.5 to 81800.9	Self made
Elevation	meters	24	From 0.04 to 1418.8	Package elevatr

Table 8: Summary of data structure

3 Exploratory Data Analysis and Wrangling

3.1 EPA PM2.5 and PM10 Datasets

Uploading PM2.5 and PM10 2018 raw data files associated with EPA Air dataset and format date column.

```
EPA_AQPM25_NC2018_raw <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")
EPA_AQPM10_NC2018_raw <- read.csv("./Data/Raw/EPAair_PM10_NC2018_raw.csv")

#Formatting Dates
EPA_AQPM25_NC2018_raw$Date <- as.Date(EPA_AQPM25_NC2018_raw$Date,
                                      format = "%m/%d/%Y")
EPA_AQPM10_NC2018_raw$Date <- as.Date(EPA_AQPM10_NC2018_raw$Date,
                                      format = "%m/%d/%Y")
```

Data exploration of the PM2.5 and PM10 2018 raw data files associated with EPA Air dataset.

```
dim(EPA_AQPM25_NC2018_raw)
```

```
## [1] 9644   20
```

```
dim(EPA_AQPM10_NC2018_raw)
```

```
## [1] 2905   20
```

```
str(EPA_AQPM25_NC2018_raw)
```

```
## 'data.frame':    9644 obs. of  20 variables:
##  $ Date                          : Date, format: "2018-01-02" "2018-01-05" ...
##  $ Source                        : Factor w/ 2 levels "AirNow","AQS": 2 2 2 2 2 2 2 2
##  $ Site.ID                       : int  370110002 370110002 370110002 370110002 37011
##  $ POC                           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily.Mean.PM2.5.Concentration: num  2.9 3.7 5.3 0.8 2.5 4.5 1.8 2.5 4.2 1.7 ...
##  $ UNITS                         : Factor w/ 1 level "ug/m3 LC": 1 1 1 1 1 1 1 1 1 1
##  $ DAILY_AQI_VALUE               : int  12 15 22 3 10 19 8 10 18 7 ...
##  $ Site.Name                     : Factor w/ 25 levels "","Blackstone",..: 15 15 15 1
##  $ DAILY_OBS_COUNT               : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PERCENT_COMPLETE              : num  100 100 100 100 100 100 100 100 100 100 ...
##  $ AQS_PARAMETER_CODE            : int  88502 88502 88502 88502 88502 88502 88502 885
##  $ AQS_PARAMETER_DESC            : Factor w/ 2 levels "Acceptable PM2.5 AQI & Speciat
##  $ CBSA_CODE                     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ CBSA_NAME                     : Factor w/ 14 levels "","Asheville, NC",..: 1 1 1 1
##  $ STATE_CODE                    : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                         : Factor w/ 1 level "North Carolina": 1 1 1 1 1 1 1
##  $ COUNTY_CODE                   : int  11 11 11 11 11 11 11 11 11 11 ...
```

```
##  $ COUNTY                        : Factor w/ 21 levels "Avery","Buncombe",..: 1 1 1 1
##  $ SITE_LATITUDE                 : num  36 36 36 36 36 ...
##  $ SITE_LONGITUDE                : num  -81.9 -81.9 -81.9 -81.9 -81.9 ...
```

```
str(EPA_AQPM10_NC2018_raw)
```

```
## 'data.frame':    2905 obs. of  20 variables:
##  $ Date                          : Date, format: "2018-01-01" "2018-01-02" ...
##  $ Source                        : Factor w/ 1 level "AQS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Site.ID                       : int  370510009 370510009 370510009 370510009 370510
##  $ POC                           : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Daily.Mean.PM10.Concentration: int  3 9 15 12 12 10 9 15 22 15 ...
##  $ UNITS                         : Factor w/ 1 level "ug/m3 SC": 1 1 1 1 1 1 1 1 1 1 .
##  $ DAILY_AQI_VALUE               : int  3 8 14 11 11 9 8 14 20 14 ...
##  $ Site.Name                     : Factor w/ 9 levels "Durham Armory",..: 9 9 9 9 9 9
##  $ DAILY_OBS_COUNT               : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PERCENT_COMPLETE              : num  100 100 100 100 100 100 100 100 100 100 ...
##  $ AQS_PARAMETER_CODE            : int  81102 81102 81102 81102 81102 81102 81102 8110
##  $ AQS_PARAMETER_DESC            : Factor w/ 1 level "PM10 Total 0-10um STP": 1 1 1 1
##  $ CBSA_CODE                     : int  22180 22180 22180 22180 22180 22180 22180 2218
##  $ CBSA_NAME                     : Factor w/ 8 levels "","Charlotte-Concord-Gastonia,
##  $ STATE_CODE                    : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                         : Factor w/ 1 level "North Carolina": 1 1 1 1 1 1 1 1
##  $ COUNTY_CODE                   : int  51 51 51 51 51 51 51 51 51 51 ...
##  $ COUNTY                        : Factor w/ 8 levels "Cumberland","Durham",..: 1 1 1
##  $ SITE_LATITUDE                 : num  35 35 35 35 35 ...
##  $ SITE_LONGITUDE                : num  -79 -79 -79 -79 -79 ...
```

```
colnames(EPA_AQPM25_NC2018_raw)
```

```
##  [1] "Date"                           "Source"
##  [3] "Site.ID"                        "POC"
##  [5] "Daily.Mean.PM2.5.Concentration" "UNITS"
##  [7] "DAILY_AQI_VALUE"                "Site.Name"
##  [9] "DAILY_OBS_COUNT"                "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"             "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"                      "CBSA_NAME"
## [15] "STATE_CODE"                     "STATE"
## [17] "COUNTY_CODE"                    "COUNTY"
## [19] "SITE_LATITUDE"                  "SITE_LONGITUDE"
```

```
colnames(EPA_AQPM10_NC2018_raw)
```

```
##  [1] "Date"                           "Source"
##  [3] "Site.ID"                        "POC"
##  [5] "Daily.Mean.PM10.Concentration" "UNITS"
##  [7] "DAILY_AQI_VALUE"                "Site.Name"
```

```
##  [9] "DAILY_OBS_COUNT"               "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"            "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"                     "CBSA_NAME"
## [15] "STATE_CODE"                    "STATE"
## [17] "COUNTY_CODE"                   "COUNTY"
## [19] "SITE_LATITUDE"                 "SITE_LONGITUDE"
```

```
summary(EPA_AQPM25_NC2018_raw)
```

```
##       Date                Source        Site.ID               POC
##  Min.   :2018-01-01   AirNow: 873   Min.   :370110002   Min.   :1.000
##  1st Qu.:2018-04-04   AQS   :8771   1st Qu.:370650099   1st Qu.:3.000
##  Median :2018-06-27                 Median :371190041   Median :3.000
##  Mean   :2018-06-30                 Mean   :371023866   Mean   :2.948
##  3rd Qu.:2018-09-30                 3rd Qu.:371230001   3rd Qu.:3.000
##  Max.   :2018-12-31                 Max.   :371830021   Max.   :5.000
##
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-2.80                  ug/m3 LC:9644   Min.   : 0.00
##  1st Qu.: 5.00                                  1st Qu.:21.00
##  Median : 7.20                                  Median :30.00
##  Mean   : 7.61                                  Mean   :31.22
##  3rd Qu.: 9.80                                  3rd Qu.:41.00
##  Max.   :34.20                                  Max.   :97.00
##
##                          Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Garinger High School        : 732   Min.   :1       Min.   :100
##  Millbrook School            : 722   1st Qu.:1       1st Qu.:100
##  Remount                     : 668   Median :1       Median :100
##  Montclaire Elementary School: 648   Mean   :1       Mean   :100
##  Hattie Avenue               : 510   3rd Qu.:1       3rd Qu.:100
##  Board Of Ed. Bldg.          : 478   Max.   :1       Max.   :100
##  (Other)                     :5886
##  AQS_PARAMETER_CODE                                AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:2008
##  1st Qu.:88101      PM2.5 - Local Conditions              :7636
##  Median :88101
##  Mean   :88184
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                  CBSA_NAME      STATE_CODE
##  Min.   :11700   Charlotte-Concord-Gastonia, NC-SC:2048   Min.   :37
##  1st Qu.:16740   Raleigh, NC                     :1418   1st Qu.:37
##  Median :24780   Winston-Salem, NC               :1323   Median :37
```

```
##  Mean   :29881                                    :1165   Mean   :37
##  3rd Qu.:40580   Asheville, NC                    : 532   3rd Qu.:37
##  Max.   :49180   Durham-Chapel Hill, NC           : 469   Max.   :37
##  NA's   :1165    (Other)                          :2689
##             STATE       COUNTY_CODE            COUNTY     SITE_LATITUDE
##  North Carolina:9644   Min.   : 11.0   Mecklenburg:2048   Min.   :34.36
##                        1st Qu.: 65.0   Wake       :1069   1st Qu.:35.24
##                        Median :119.0   Forsyth    : 876   Median :35.64
##                        Mean   :102.4   Buncombe   : 478   Mean   :35.58
##                        3rd Qu.:123.0   Durham     : 469   3rd Qu.:35.91
##                        Max.   :183.0   Pitt       : 461   Max.   :36.11
##                                        (Other)    :4243
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-80.03
##  3rd Qu.:-78.82
##  Max.   :-76.21
##
```

```
summary(EPA_AQPM10_NC2018_raw)
```

```
##       Date            Source        Site.ID               POC
##  Min.   :2018-01-01   AQS:2905   Min.   :370510009   Min.   :1.000
##  1st Qu.:2018-04-07              1st Qu.:370670022   1st Qu.:3.000
##  Median :2018-07-02              Median :371170001   Median :3.000
##  Mean   :2018-07-03              Mean   :371072712   Mean   :3.172
##  3rd Qu.:2018-10-04              3rd Qu.:371190042   3rd Qu.:4.000
##  Max.   :2018-12-31              Max.   :371830014   Max.   :5.000
##
##  Daily.Mean.PM10.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   : 0.00                 ug/m3 SC:2905   Min.   : 0.00
##  1st Qu.:10.00                                 1st Qu.: 9.00
##  Median :13.00                                 Median :12.00
##  Mean   :13.72                                 Mean   :12.67
##  3rd Qu.:17.00                                 3rd Qu.:16.00
##  Max.   :64.00                                 Max.   :55.00
##
##                          Site.Name   DAILY_OBS_COUNT PERCENT_COMPLETE
##  Millbrook School              :588   Min.   :1       Min.   :100
##  Garinger High School          :351   1st Qu.:1       1st Qu.:100
##  Montclaire Elementary School:344   Median :1       Median :100
##  Hattie Avenue                 :342   Mean   :1       Mean   :100
##  Durham Armory                 :335   3rd Qu.:1       3rd Qu.:100
```

```
##  William Owen School         :321   Max.   :1       Max.   :100
##  (Other)                     :624
##  AQS_PARAMETER_CODE             AQS_PARAMETER_DESC   CBSA_CODE
##  Min.   :81102      PM10 Total 0-10um STP:2905     Min.   :16740
##  1st Qu.:81102                                     1st Qu.:16740
##  Median :81102                                     Median :22180
##  Mean   :81102                                     Mean   :28310
##  3rd Qu.:81102                                     3rd Qu.:39580
##  Max.   :81102                                     Max.   :49180
##                                                    NA's   :247
##                               CBSA_NAME     STATE_CODE
##  Charlotte-Concord-Gastonia, NC-SC:695   Min.   :37
##  Raleigh, NC                      :588   1st Qu.:37
##  Winston-Salem, NC                :342   Median :37
##  Durham-Chapel Hill, NC           :335   Mean   :37
##  Fayetteville, NC                 :321   3rd Qu.:37
##  Greensboro-High Point, NC        :320   Max.   :37
##  (Other)                          :304
##             STATE       COUNTY_CODE            COUNTY    SITE_LATITUDE
##  North Carolina:2905   Min.   : 51.0   Mecklenburg:695   Min.   :35.04
##                        1st Qu.: 67.0   Wake       :588   1st Qu.:35.24
##                        Median :117.0   Forsyth    :342   Median :35.86
##                        Mean   :107.3   Durham     :335   Mean   :35.67
##                        3rd Qu.:119.0   Cumberland :321   3rd Qu.:36.03
##                        Max.   :183.0   Guilford   :320   Max.   :36.11
##                                        (Other)    :304
##  SITE_LONGITUDE
##  Min.   :-80.87
##  1st Qu.:-80.23
##  Median :-78.95
##  Mean   :-79.36
##  3rd Qu.:-78.57
##  Max.   :-76.91
##
```

Visual data exploration of the PM2.5 2018 raw data file in Figure 1, Figure 2, and Figure 3.

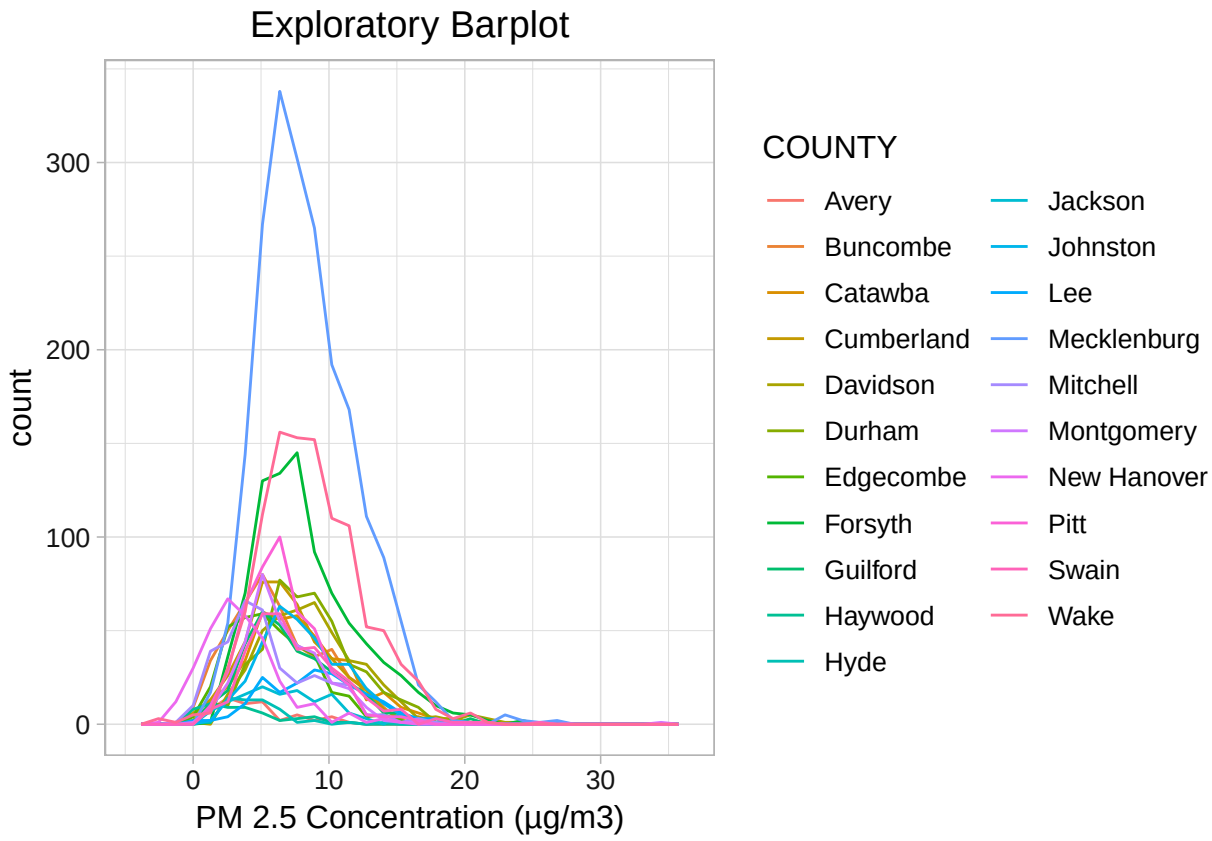

Figure 1: PM2.5 NC 2018 frequency polygon.

Visual data exploration of the PM10 2018 raw data file in Figure 4.

Data wrangling of the PM2.5 and PM10 2018 raw data files.

```
#Selecting Columns
EPA_AQ_PM25_NC2018_Temp <- select(EPA_AQPM25_NC2018_raw, Date, Site.ID,
                                  Daily.Mean.PM2.5.Concentration,
                                  AQS_PARAMETER_DESC,
                                  COUNTY:SITE_LONGITUDE)

#Changing column name
colnames(EPA_AQ_PM25_NC2018_Temp)[colnames(EPA_AQ_PM25_NC2018_Temp)
                                  =="Daily.Mean.PM2.5.Concentration"]<-"Daily.Mean.Concent

#Selecting Columns
EPA_AQ_PM10_NC2018_Temp <- select(EPA_AQPM10_NC2018_raw, Date, Site.ID,
                                  Daily.Mean.PM10.Concentration, AQS_PARAMETER_DESC,
                                  COUNTY:SITE_LONGITUDE)

#Changing column name
```

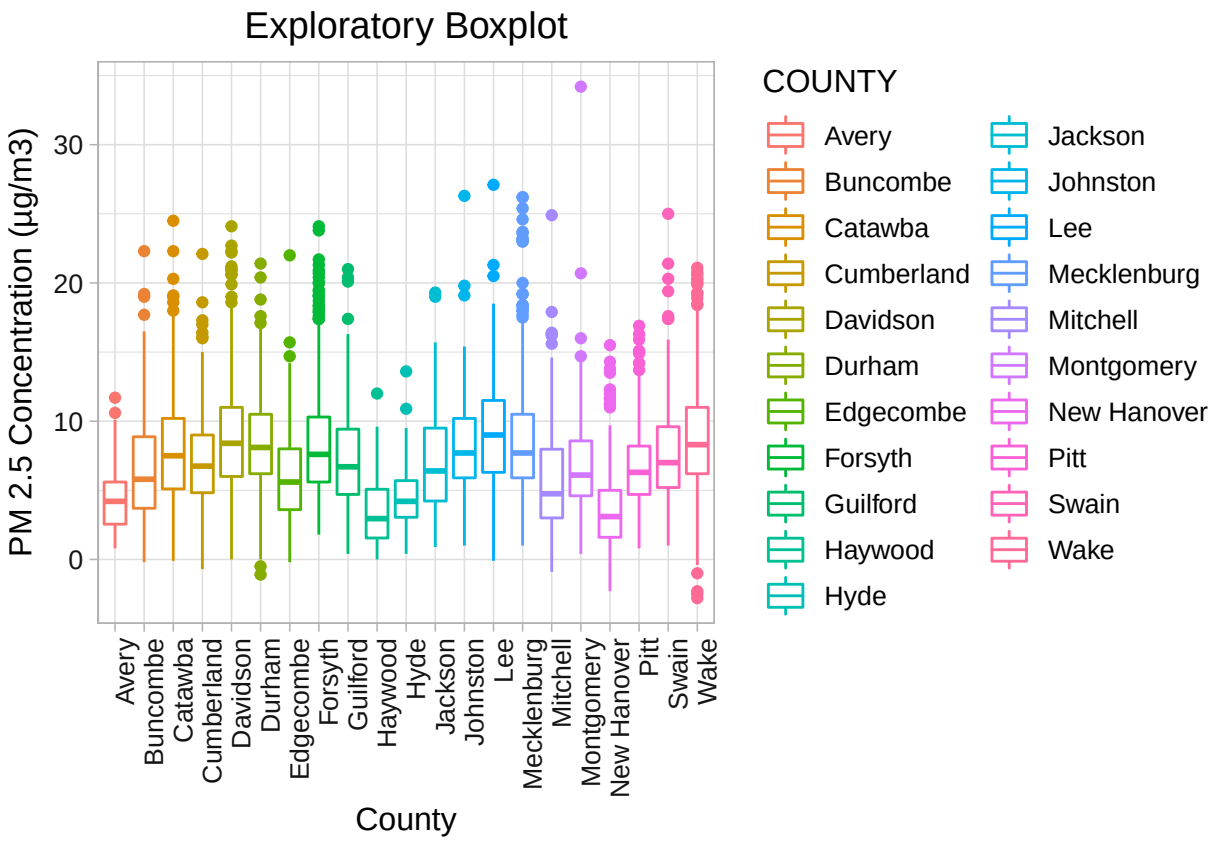

Figure 2: PM2.5 NC 2018 boxplot.

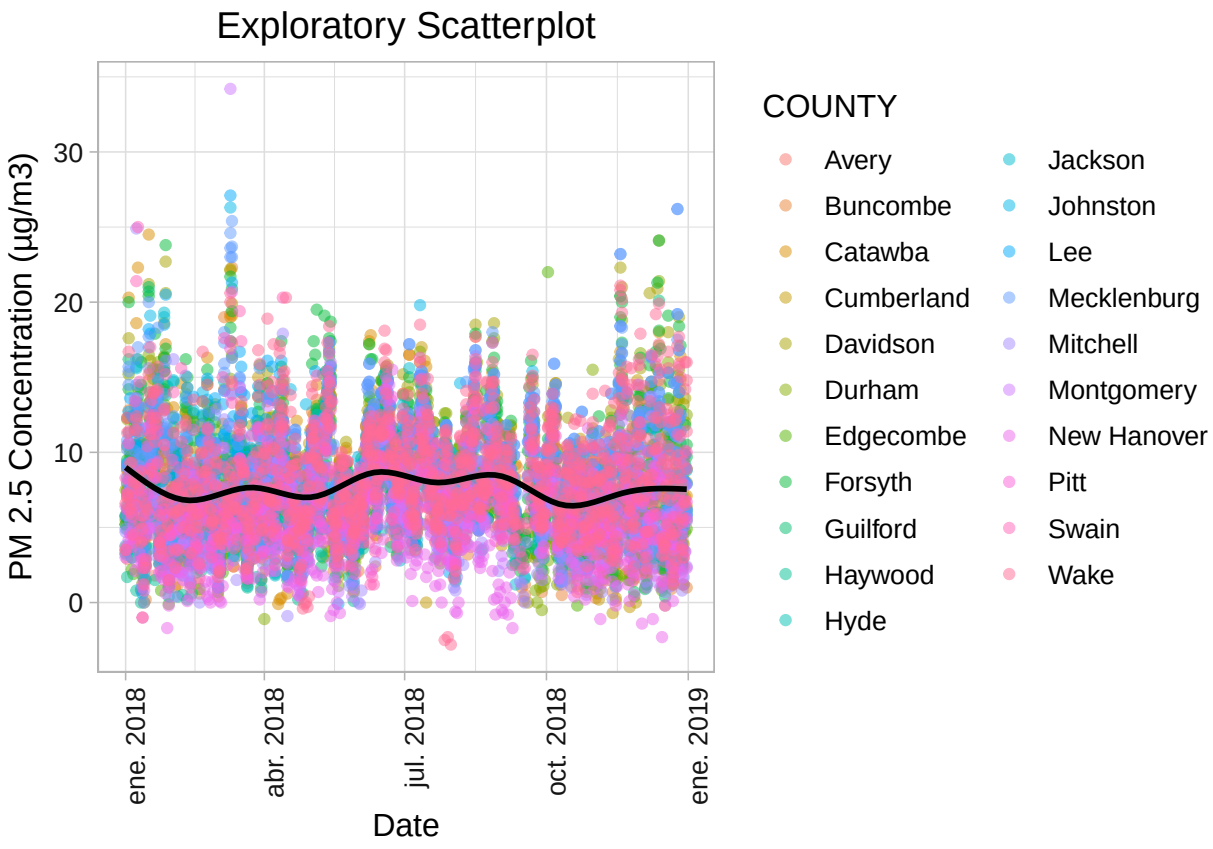

Figure 3: PM2.5 NC 2018 scatterplot.

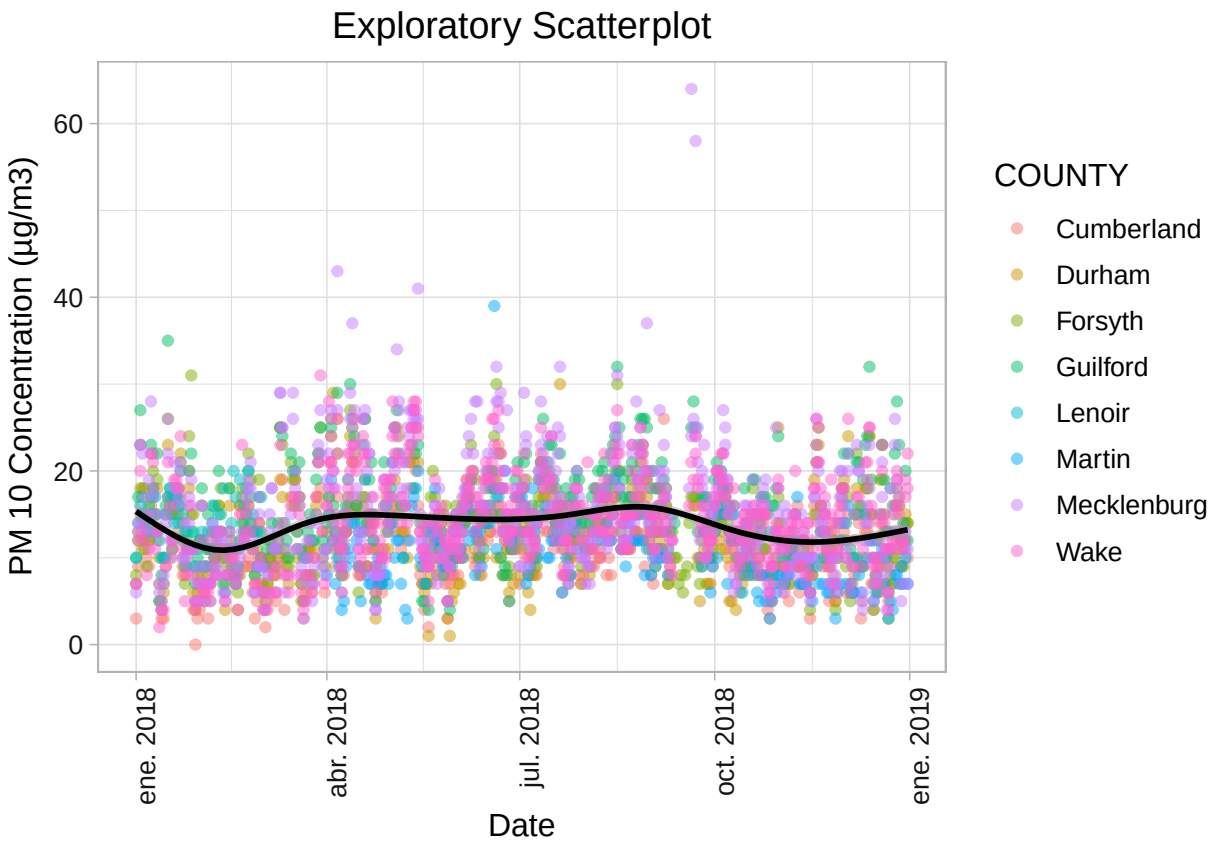

Figure 4: PM10 NC 2018 scatterplot.

```
colnames(EPA_AQ_PM10_NC2018_Temp)[colnames(EPA_AQ_PM10_NC2018_Temp)
                                  =="Daily.Mean.PM10.Concentration"]<-"Daily.Mean.Concen

#Create AQS_PARAMETER_DESC Column with Contaminant description.
EPA_AQ_PM25_NC2018_Temp$AQS_PARAMETER_DESC <- "PM2.5"
EPA_AQ_PM10_NC2018_Temp$AQS_PARAMETER_DESC <- "PM10"

#Eliminates duplicate dates
EPA_AQ_PM25_NC2018_Cleaned <-
  EPA_AQ_PM25_NC2018_Temp [!duplicated(EPA_AQ_PM25_NC2018_Temp[c(1,2)]),]

EPA_AQ_PM10_NC2018_Cleaned <-
  EPA_AQ_PM10_NC2018_Temp [!duplicated(EPA_AQ_PM10_NC2018_Temp[c(1,2)]),]

# Combine the data.
EPA_AQ_PM2.5PM10_NC2018_Cleaned <-
  rbind(EPA_AQ_PM25_NC2018_Cleaned, EPA_AQ_PM10_NC2018_Cleaned)

#Save the data in the processed folder
write.csv(EPA_AQ_PM2.5PM10_NC2018_Cleaned,
          "./Data/Processed/EPA_AQ_PM2.5PM10_NC2018_Cleaned.csv")

#Spread PM2.5 and PM10
EPA_AQ_PM2.5PM10_NC2018_Spread <-
  EPA_AQ_PM2.5PM10_NC2018_Cleaned %>%
  spread(AQS_PARAMETER_DESC, Daily.Mean.Concentration)

#Remove rows without PM2.5 data
EPA_AQ_PM2.5PM10_NC2018_Spread <-
  EPA_AQ_PM2.5PM10_NC2018_Spread[!is.na(EPA_AQ_PM2.5PM10_NC2018_Spread$PM2.5),]

#Convert the dataset to a spatially enabled "sf" data frame
PM2.5_PM10_sf <- st_as_sf(EPA_AQ_PM2.5PM10_NC2018_Spread,coords
                          = c('SITE_LONGITUDE','SITE_LATITUDE'),crs=4269)

#Convert all to UTM Zone 17 (crs = 26917)
PM2.5_PM10_sf_utm <- st_transform(PM2.5_PM10_sf, c=26917)
```

In Figure 5 is presented the locations of the PM2.5 monitoring stations.

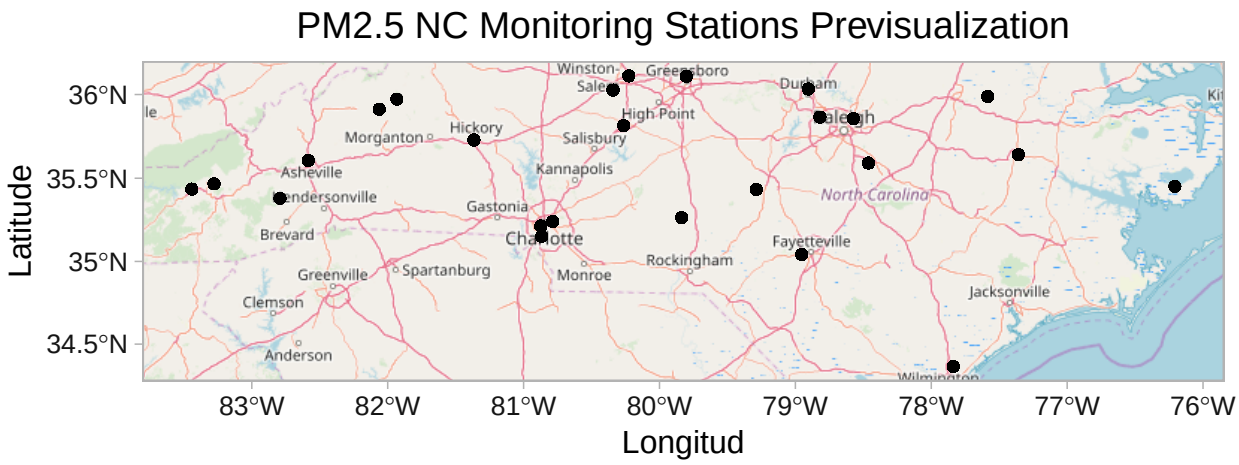

Figure 5: PM2.5 NC Monitoring Stations Previsualization.

3.2 North Carolina Counties Zoning, Geographic information, and Population Data

Downloading the list of North Carolina Counties and Population from a Wikipedia URL.

```
#North Carolina Counties
url <- "https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina"
webpage <- read_html(url)

County_Name <- webpage %>% html_nodes("th:nth-child(1)") %>% html_text()
County_Population <- webpage %>% html_nodes("tr :nth-child(7)") %>% html_text()

#Remove unwanted info and characters
County_Info <- data_frame(County = County_Name[9:108])
County_Info$County <- str_replace(County_Info$County, " County", "")
County_Info$County <- str_replace(County_Info$County, "\n", "")

Population <- data_frame(Population=County_Population[2:101])

County_Info <- cbind(County_Info, Population)

County_Info$Population <- str_replace(County_Info$Population,",","")
County_Info$Population <- str_replace(County_Info$Population,",","")

County_Info$Population <- as.numeric(County_Info$Population)
```

Assigning the corresponding zone to each county. Info from: Rudersdorf, Amy. 2010. "NC County Maps." Government & Heritage Library, State Library of North Carolina.

```
#North Carolina Zones
County_Info$Zone<-ifelse(County_Info$County == 'Ashe'
                         |County_Info$County =='Alleghany'
                         |County_Info$County =='Wilkes'
                         |County_Info$County =='Watauga'
                         |County_Info$County =='Avery'
                         |County_Info$County =='Caldwell'
                         |County_Info$County =='Mitchell'
                         |County_Info$County =='Burke'
                         |County_Info$County =='Yancey'
                         |County_Info$County =='McDowell'
                         |County_Info$County =='Rutherford'
                         |County_Info$County =='Madison'
                         |County_Info$County =='Buncombe'
                         |County_Info$County =='Polk'
                         |County_Info$County =='Henderson'
                         |County_Info$County =='Haywood'
```

```
                        |County_Info$County =='Transylvania'
                        |County_Info$County =='Swain'
                        |County_Info$County =='Jackson'
                        |County_Info$County =='Graham'
                        |County_Info$County =='Macon'
                        |County_Info$County =='Cherokee'
                        |County_Info$County =='Clay','Mountains',
                       ifelse(County_Info$County == 'Surry'
                              |County_Info$County =='Stokes'
                              |County_Info$County =='Rockingham'
                              |County_Info$County =='Caswell'
                              |County_Info$County =='Person'
                              |County_Info$County =='Granville'
                              |County_Info$County =='Vance'
                              |County_Info$County =='Warren'
                              |County_Info$County =='Yadkin'
                              |County_Info$County =='Forsyth'
                              |County_Info$County =='Guilford'
                              |County_Info$County =='Alamance'
                              |County_Info$County =='Orange'
                              |County_Info$County =='Durham'
                              |County_Info$County =='Franklin'
                              |County_Info$County =='Alexander'
                              |County_Info$County =='Iredell'
                              |County_Info$County =='Davie'
                              |County_Info$County =='Rowan'
                              |County_Info$County =='Davidson'
                              |County_Info$County =='Randolph'
                              |County_Info$County =='Chatham'
                              |County_Info$County =='Wake'
                              |County_Info$County =='Catawba'
                              |County_Info$County =='Cleveland'
                              |County_Info$County =='Lincoln'
                              |County_Info$County =='Gaston'
                              |County_Info$County =='Mecklenburg'
                              |County_Info$County =='Cabarrus'
                              |County_Info$County =='Stanly'
                              |County_Info$County =='Union'
                              |County_Info$County =='Montgomery'
                              |County_Info$County =='Anson'
                              |County_Info$County =='Moore'
                              |County_Info$County =='Lee'
                              |County_Info$County =='Richmond','Piedmont',
```

```
                                ifelse(County_Info$County == 'Scotland'
                                       |County_Info$County =='Hoke'
                                       |County_Info$County =='Harnett'
                                       |County_Info$County =='Johnston'
                                       |County_Info$County =='Nash'
                                       |County_Info$County =='Halifax'
                                       |County_Info$County =='Northampton'
                                       |County_Info$County =='Robeson'
                                       |County_Info$County =='Cumberland'
                                       |County_Info$County =='Sampson'
                                       |County_Info$County =='Wayne'
                                       |County_Info$County =='Wilson'
                                       |County_Info$County =='Edgecombe'
                                       |County_Info$County =='Columbus'
                                       |County_Info$County =='Bladen'
                                       |County_Info$County =='Brunswick'
                                       |County_Info$County =='New Hanover'
                                       |County_Info$County =='Pender'
                                       |County_Info$County =='Duplin'
                                       |County_Info$County =='Onslow'
                                       |County_Info$County =='Lenoir'
                                       |County_Info$County =='Jones'
                                       |County_Info$County =='Carteret'
                                       |County_Info$County =='Greene'
                                       |County_Info$County =='Craven'
                                       |County_Info$County =='Pitt'
                                       |County_Info$County =='Pamlico'
                                       |County_Info$County =='Beaufort'
                                       |County_Info$County =='Martin'
                                       |County_Info$County =='Bertie'
                                       |County_Info$County =='Hyde'
                                       |County_Info$County =='Dare'
                                       |County_Info$County =='Tyrrell'
                                       |County_Info$County =='Washington'
                                       |County_Info$County =='Hertford'
                                       |County_Info$County =='Gates'
                                       |County_Info$County =='Currituck'
                                       |County_Info$County =='Camden'
                                       |County_Info$County =='Pasquotank'
                                       |County_Info$County =='Perquimans'
                                       |County_Info$County =='Chowan','Coastal','NoInfo'
```

Data exploration of the County_Info dataframe.

```r
dim(County_Info)
```

```
## [1] 100   3
```

```r
str(County_Info)
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ County    : chr  "Alamance" "Alexander" "Alleghany" "Anson" ...
##  $ Population: num  157844 37159 10935 25531 26833 ...
##  $ Zone      : chr  "Piedmont" "Piedmont" "Mountains" "Piedmont" ...
```

```r
colnames(County_Info)
```

```
## [1] "County"     "Population" "Zone"
```

```r
summary(County_Info)
```

```
##     County            Population          Zone
##  Length:100         Min.   :   4090   Length:100
##  Class :character   1st Qu.:  25001   Class :character
##  Mode  :character   Median :  55311   Mode  :character
##                     Mean   : 100526
##                     3rd Qu.: 114764
##                     Max.   :1034290
```

```r
unique(County_Info$Zone)
```

```
## [1] "Piedmont"  "Mountains" "Coastal"
```

Adding the County Information to the PM2.5_PM10_sf_utm dataframe.

```r
PM2.5_PM10_Info_sf_utm <- PM2.5_PM10_sf_utm %>%
left_join(County_Info, by = c("COUNTY"="County"))
```

3.3 US Census Bureau US counties shapefile

Reading the USA county shapefile, sub-setting for NC.

```
counties_sf<- st_read('./Data/Spatial/cb_2017_us_county_20m.shp') %>%
  filter(STATEFP == 37) #Filter for just NC Counties
```

```
## Reading layer `cb_2017_us_county_20m' from data source `C:\Users\Felipe\OneDrive - Du
## Simple feature collection with 3220 features and 9 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -179.1743 ymin: 17.91377 xmax: 179.7739 ymax: 71.35256
## epsg (SRID):    4269
## proj4string:    +proj=longlat +datum=NAD83 +no_defs
```

```
#CRS
st_crs(counties_sf) #crs=4269 = NAD83.
```

```
## Coordinate Reference System:
##   EPSG: 4269
##   proj4string: "+proj=longlat +datum=NAD83 +no_defs"
```

Converting the counties_sf to UTM Zone 17

```
#Convert all to UTM Zone 17 (crs = 26917)
counties_sf_utm <- st_transform(counties_sf, c=26917)

#Adding the Zone Info
counties_sf_utm <- counties_sf_utm %>%
left_join(County_Info, by = c("NAME"="County"))
```

Data exploration of the County_Info dataframe.

```
dim(counties_sf_utm)
```

```
## [1] 100  12
```

```
str(counties_sf_utm)
```

```
## Classes 'sf' and 'data.frame':   100 obs. of  12 variables:
##  $ STATEFP   : Factor w/ 52 levels "01","02","04",..: 34 34 34 34 34 34 34 34 34 34 .
##  $ COUNTYFP  : Factor w/ 325 levels "001","003","005",..: 116 71 94 54 62 125 84 102
##  $ COUNTYNS  : Factor w/ 3220 levels "00023901","00025441",..: 1672 1649 1659 1639 16
##  $ AFFGEOID  : Factor w/ 3220 levels "0500000US01001",..: 1981 1944 1964 1931 1937 19
##  $ GEOID     : Factor w/ 3220 levels "01001","01003",..: 1981 1944 1964 1931 1937 198
##  $ NAME      : chr  "Vance" "Lenoir" "Pitt" "Guilford" ...
##  $ LSAD      : Factor w/ 9 levels "00","03","04",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ ALAND     : num  6.54e+08 1.03e+09 1.69e+09 1.67e+09 1.01e+09 ...
##  $ AWATER    : num  42187365 5900300 8248766 30723331 3981006 ...
```

```
##  $ Population: num  44420 57934 176484 517197 52571 ...
##  $ Zone      : chr  "Piedmont" "Coastal" "Coastal" "Piedmont" ...
##  $ geometry  :sfc_MULTIPOLYGON of length 100; first list element: List of 1
##   ..$ :List of 1
##   .. ..$ : num [1:10, 1:2] 724066 727615 739570 744478 741926 ...
##   ..- attr(*, "class")= chr  "XY" "MULTIPOLYGON" "sfg"
##  - attr(*, "sf_column")= chr "geometry"
##  - attr(*, "agr")= Factor w/ 3 levels "constant","aggregate",..: NA NA NA NA NA NA NA
##   ..- attr(*, "names")= chr  "STATEFP" "COUNTYFP" "COUNTYNS" "AFFGEOID" ...
```

```
colnames(counties_sf_utm)
```

```
##  [1] "STATEFP"    "COUNTYFP"   "COUNTYNS"   "AFFGEOID"   "GEOID"
##  [6] "NAME"       "LSAD"       "ALAND"      "AWATER"     "Population"
## [11] "Zone"       "geometry"
```

```
summary(counties_sf_utm)
```

```
##     STATEFP       COUNTYFP      COUNTYNS            AFFGEOID      GEOID
##  37     :100   001    : 1   01008531: 1   0500000US37001: 1   37001  : 1
##  01     :  0   003    : 1   01008532: 1   0500000US37003: 1   37003  : 1
##  02     :  0   005    : 1   01008533: 1   0500000US37005: 1   37005  : 1
##  04     :  0   007    : 1   01008534: 1   0500000US37007: 1   37007  : 1
##  05     :  0   009    : 1   01008535: 1   0500000US37009: 1   37009  : 1
##  06     :  0   011    : 1   01008536: 1   0500000US37011: 1   37011  : 1
##  (Other):  0   (Other):94   (Other) :94   (Other)       :94   (Other):94
##      NAME                LSAD         ALAND               AWATER
##  Length:100         06     :100   Min.   :4.472e+08   Min.   :4.453e+05
##  Class :character   00     :  0   1st Qu.:8.936e+08   1st Qu.:7.287e+06
##  Mode  :character   03     :  0   Median :1.192e+09   Median :1.595e+07
##                     04     :  0   Mean   :1.259e+09   Mean   :1.347e+08
##                     05     :  0   3rd Qu.:1.501e+09   3rd Qu.:3.831e+07
##                     12     :  0   Max.   :2.453e+09   Max.   :3.001e+09
##                     (Other):  0
##    Population          Zone                    geometry
##  Min.   :   4090   Length:100         MULTIPOLYGON :100
##  1st Qu.:  25001   Class :character   epsg:26917   :  0
##  Median :  55311   Mode  :character   +proj=utm ...:  0
##  Mean   : 100526
##  3rd Qu.: 114764
##  Max.   :1034290
##
```

Visual data exploration of the counties_sf_utm dataframe in Figure 6 and Figure 7.

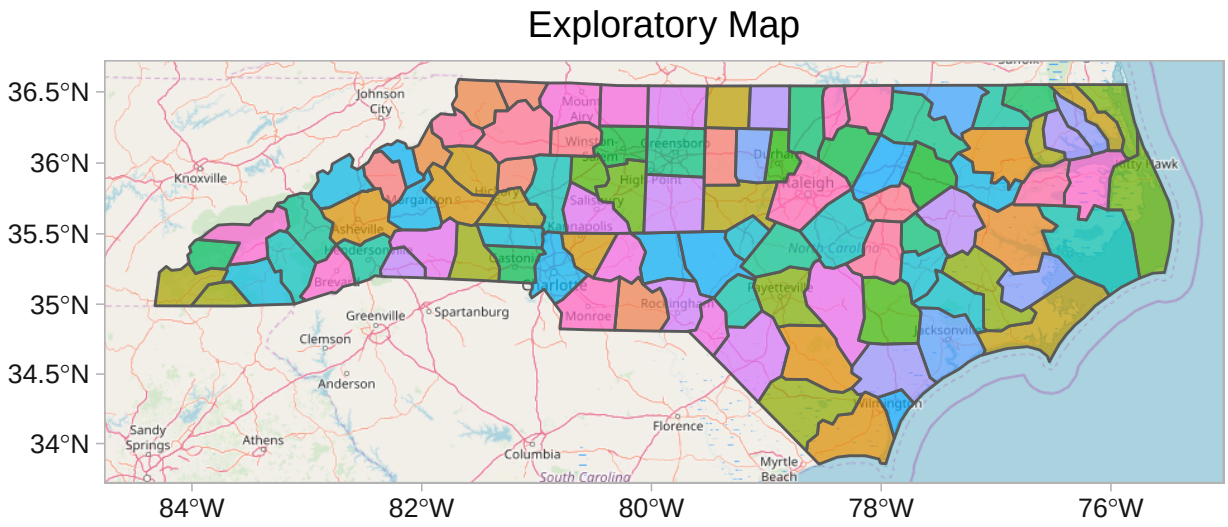

Figure 6: Counties exploratory map.

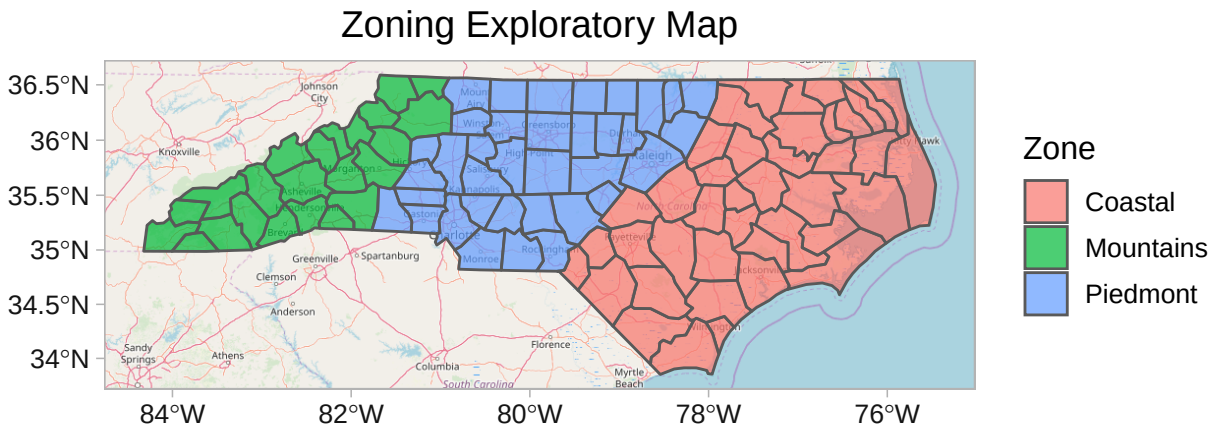

Figure 7: NC Zoning exploratory map.

3.4 NOAA Average Temperature Dataset

Reading the 2018 North Carolina Air Temperature data.

```
#Read the 2018 Air Temperature data
NOAA_DTAVG_NC2018_raw <- read.csv("./Data/Raw/NOAA_TAVG_NC2018_raw.csv")
```

Data exploration of the NOAA_DTAVG_NC2018_raw dataframe.

```
dim(NOAA_DTAVG_NC2018_raw)
```

```
## [1] 283423      7
```

```
str(NOAA_DTAVG_NC2018_raw)
```

```
## 'data.frame':    283423 obs. of  7 variables:
##  $ STATION  : Factor w/ 1066 levels "US1NCAG0001",..: 217 217 217 217 217 217 217 217
##  $ NAME     : Factor w/ 1063 levels "ABERDEEN 0.7 NW, NC US",..: 716 716 716 716 716
##  $ LATITUDE : num  34.8 34.8 34.8 34.8 34.8 ...
##  $ LONGITUDE: num  -76.9 -76.9 -76.9 -76.9 -76.9 ...
##  $ ELEVATION: num  8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 ...
##  $ DATE     : Factor w/ 365 levels "01/01/2018","01/02/2018",..: 37 133 145 205 265 2
##  $ TAVG     : int  NA NA NA NA NA NA NA NA NA NA ...
```

```
colnames(NOAA_DTAVG_NC2018_raw)
```

```
## [1] "STATION"   "NAME"      "LATITUDE"  "LONGITUDE" "ELEVATION" "DATE"
## [7] "TAVG"
```

```
summary(NOAA_DTAVG_NC2018_raw)
```

```
##         STATION                                  NAME           LATITUDE
##  US1NCBC0005:   365   SPARTA 3.5 SSW, NC US      :   545   Min.   :33.88
##  US1NCBC0041:   365   HILLSBOROUGH 5.6 NNW, NC US:   502   1st Qu.:35.16
##  US1NCBK0004:   365   ADVANCE 0.2 ESE, NC US     :   365   Median :35.56
##  US1NCCH0004:   365   ALBEMARLE, NC US           :   365   Mean   :35.49
##  US1NCCS0002:   365   ARDEN 1.6 ENE, NC US       :   365   3rd Qu.:35.90
##  US1NCCY0003:   365   ASHEBORO 1.3 SSE, NC US    :   365   Max.   :36.56
##  (Other)    :281233   (Other)                    :280916
##    LONGITUDE        ELEVATION              DATE             TAVG
##  Min.   :-84.30   Min.   :   0.0   16/04/2018:   887   Min.   : 9.00
##  1st Qu.:-81.67   1st Qu.:  29.3   17/05/2018:   874   1st Qu.:47.00
##  Median :-79.16   Median : 150.9   20/03/2018:   871   Median :63.00
##  Mean   :-79.70   Mean   : 279.9   12/06/2018:   870   Mean   :60.19
##  3rd Qu.:-78.01   3rd Qu.: 389.5   30/05/2018:   869   3rd Qu.:75.00
##  Max.   :-75.46   Max.   :1902.0   01/08/2018:   867   Max.   :87.00
##                                    (Other)   :278185   NA's   :269572
```

Data wrangling of the NOAA_DTAVG_NC2018_raw dataframe.

```
#Remove stations without Temperature information
NOAA_DTAVG_NC2018_Complete <- na.omit(NOAA_DTAVG_NC2018_raw)

#Convert the dataset to a spatially enabled "sf" data frame
NOAA_DTAVG_NC2018_sf <-
  st_as_sf(NOAA_DTAVG_NC2018_Complete,coords = c('LONGITUDE','LATITUDE'),crs=4269)

#Convert all to UTM Zone 17 (crs = 26917)
NOAA_DTAVG_NC2018_sf_utm <- st_transform(NOAA_DTAVG_NC2018_sf, c=26917)

#Formatting dates
NOAA_DTAVG_NC2018_sf_utm$DATE <-
  as.Date(NOAA_DTAVG_NC2018_sf_utm$DATE, format = "%d/%m/%Y")
```

The 2018 Air Temperature data does not have County information, so the location is used with the counties_sf_utm dataframe to locate the county of each station.

```
#Adding the county and zone information to the Temperature dataframe

#Index of the matching feature
county_index <- st_nearest_feature(NOAA_DTAVG_NC2018_sf_utm, counties_sf_utm)

#Eliminates geo info
aux1 <- st_set_geometry(counties_sf_utm[county_index,"NAME"], value=NULL)

#adds the columns
NOAA_DTAVG_NC2018_sf_utm$COUNTY <- aux1$NAME

#Reordering
NOAA_DTAVG_NC2018_sf_utm <- NOAA_DTAVG_NC2018_sf_utm[,c(1,2,3,4,5,7,6)]
```

Visual data exploration of the 2018 North Carolina Air Temperature data in Figure 8, Figure 9, Figure 10, and Figure 11..

Next, the temperature of the nearest Temperature Station is added to each PM2.5 Station in the PM2.5_PM10_Info_sf_utm dataframe.

```
#Create a Data frame with only the PM2.5 station info
PM2.5_Stations <- PM2.5_PM10_Info_sf_utm %>%
  select(Site.ID, geometry) %>%
  subset(!duplicated(Site.ID))

#Distances bewteen the PM2.5 stations and the Temperature Stations
Nearest <- st_nearest_feature(PM2.5_Stations, NOAA_DTAVG_NC2018_sf_utm)

a <- length(unique(PM2.5_Stations$Site.ID))
```

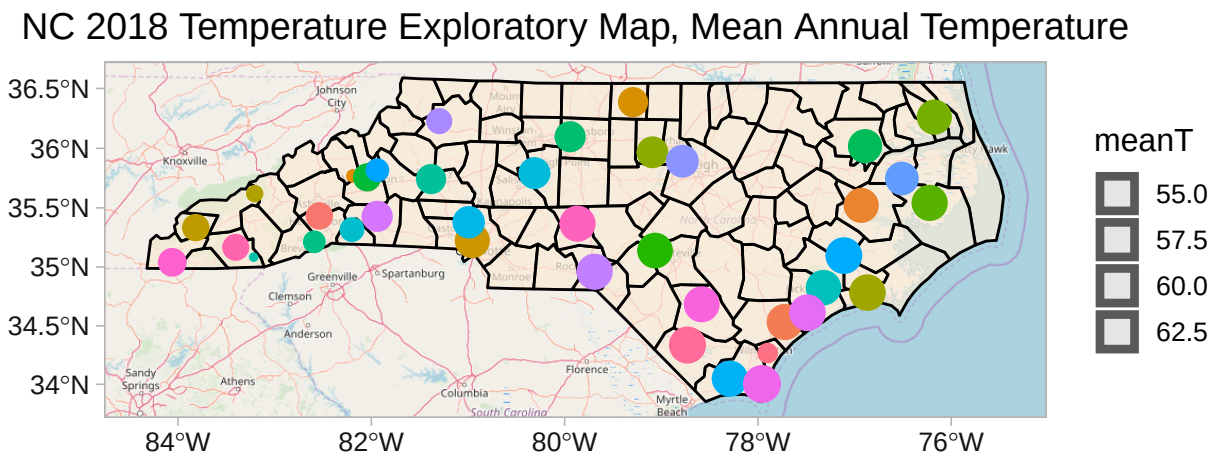

Figure 8: Mean Annual Temperature exploratory map.

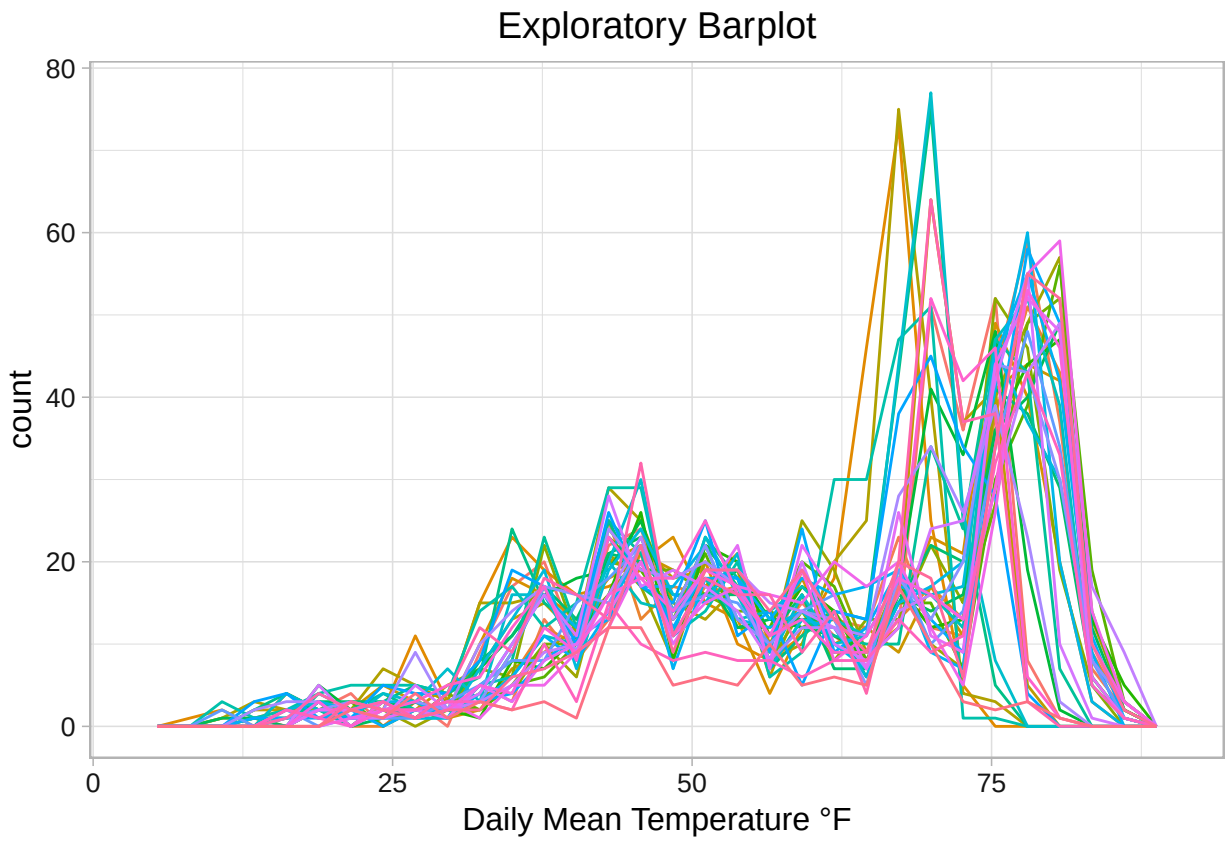

Figure 9: Daily Mean Temperature NC 2018 frequency polygon.

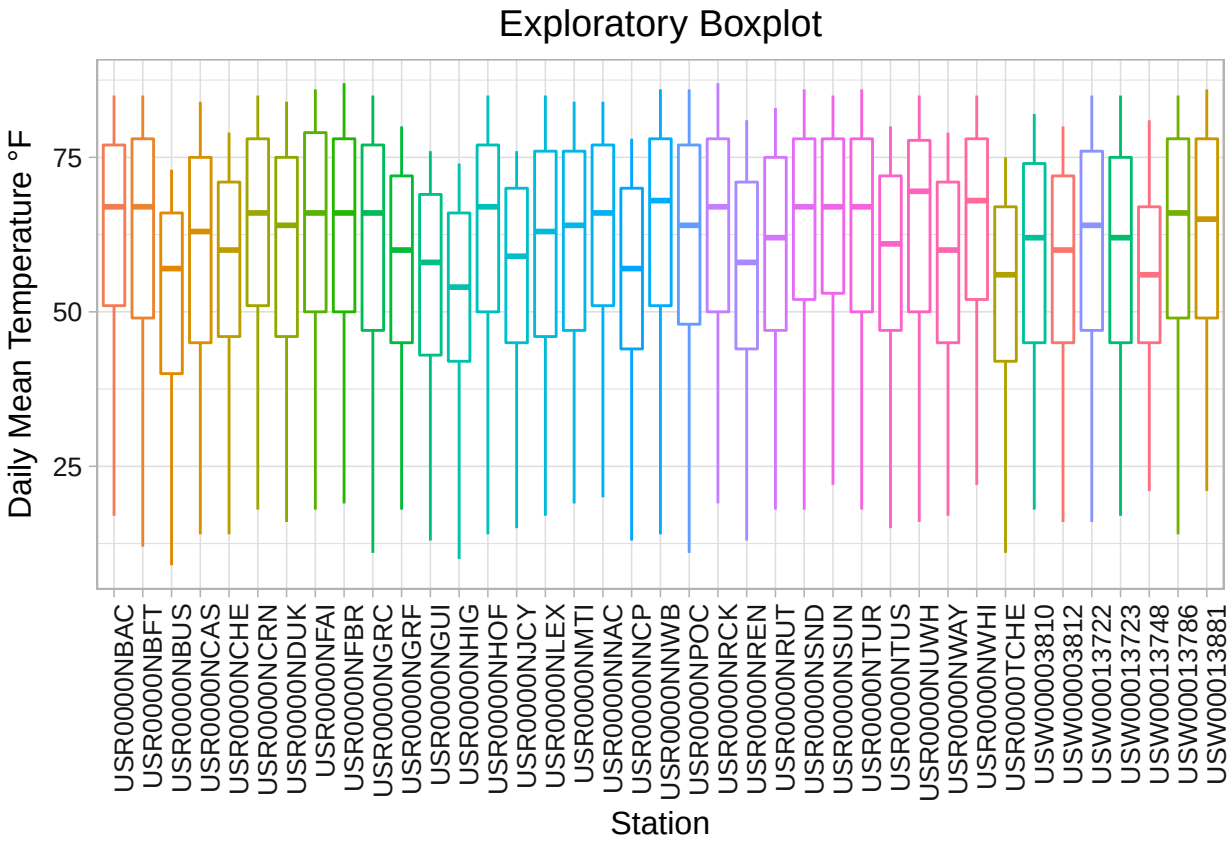

Figure 10: Daily Mean Temperature NC 2018 boxplot.

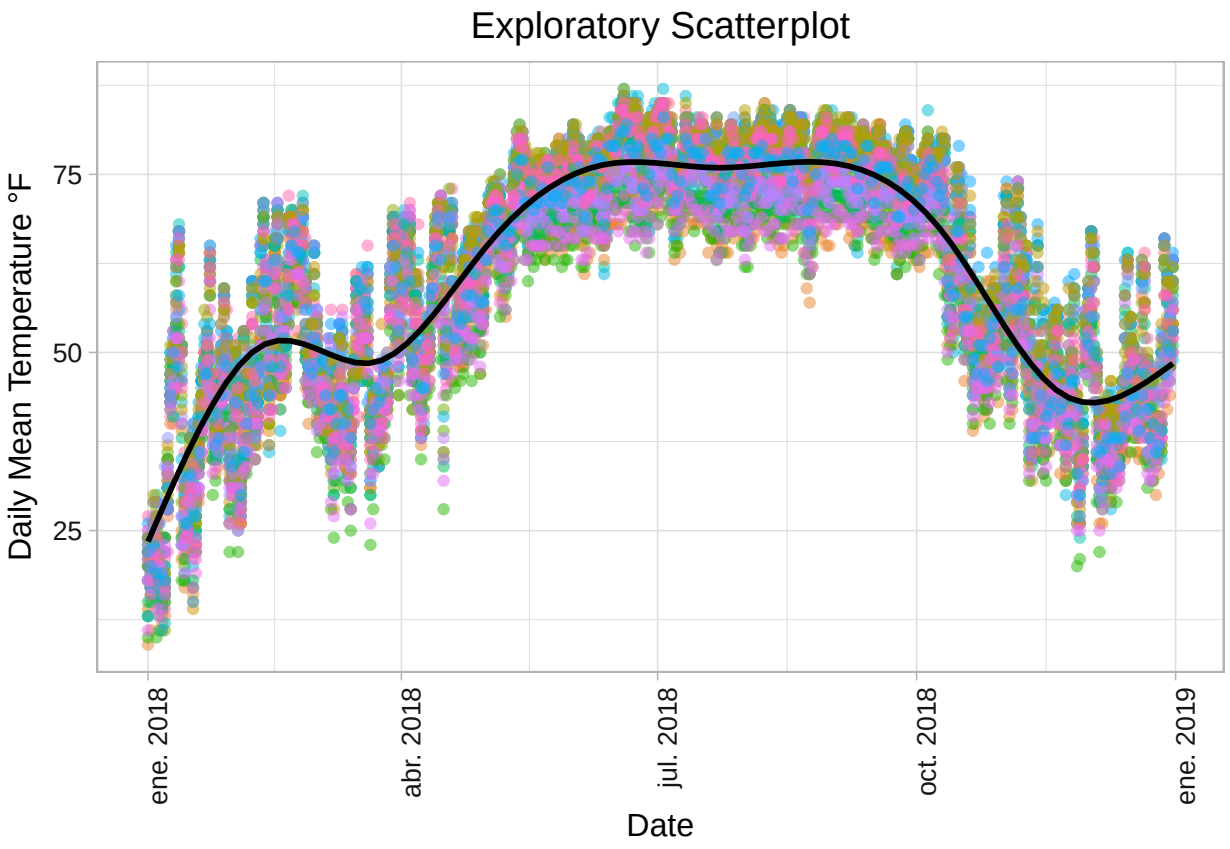

Figure 11: Daily Mean Temperature NC 2018 scatterplot.

```
NOAA_DTAVG_NC2018_sf_utm$NAME <- as.character(NOAA_DTAVG_NC2018_sf_utm$NAME)

#Assingning the nearest Temperature Station to each PM2.5 station.
for (i in 1:a){
  PM2.5_Stations$Temp_Est[i] <- NOAA_DTAVG_NC2018_sf_utm$NAME[Nearest[i]]
}

#Drop the geo data
aux2 <- st_set_geometry(PM2.5_Stations, value=NULL)

#Left_join the data
PM2.5_PM10_Temp_sf_utm <- PM2.5_PM10_Info_sf_utm %>%
left_join(aux2)

#Assingning the Temperature of the nearest Temperature Station to
#each PM2.5 station.

#Drops the geo data
aux3 <- st_set_geometry(NOAA_DTAVG_NC2018_sf_utm, value=NULL)

#Left_join the data
PM2.5_PM10_Temp_sf_utm <- PM2.5_PM10_Temp_sf_utm %>%
left_join(aux3, by = c("Temp_Est"="NAME", "Date"="DATE", "COUNTY")) %>%
  select(Date,Site.ID,COUNTY,Population,Zone,PM2.5,PM10,TAVG,geometry)
```

3.5 EPA combustion points for electricity generation in the US Dataset

Reading the Electricity Generation via Combustion data.

```
EPA_US_CombEmissions <- st_read("./Data/Raw/EPA_ElecGenComb_US_raw.kml")
```

```
## Reading layer `Electricity Generation via Combustion' from data source `C:\Users\Feli
## Simple feature collection with 2042 features and 2 fields
## geometry type:  POINT
## dimension:      XYZ
## bbox:           xmin: -176.6593 ymin: 19.63283 xmax: -67.00325 ymax: 71.29221
## epsg (SRID):    4326
## proj4string:    +proj=longlat +datum=WGS84 +no_defs
```

Wrangling the data

```
st_crs(EPA_US_CombEmissions) #crs=4326 = WGS 84
```

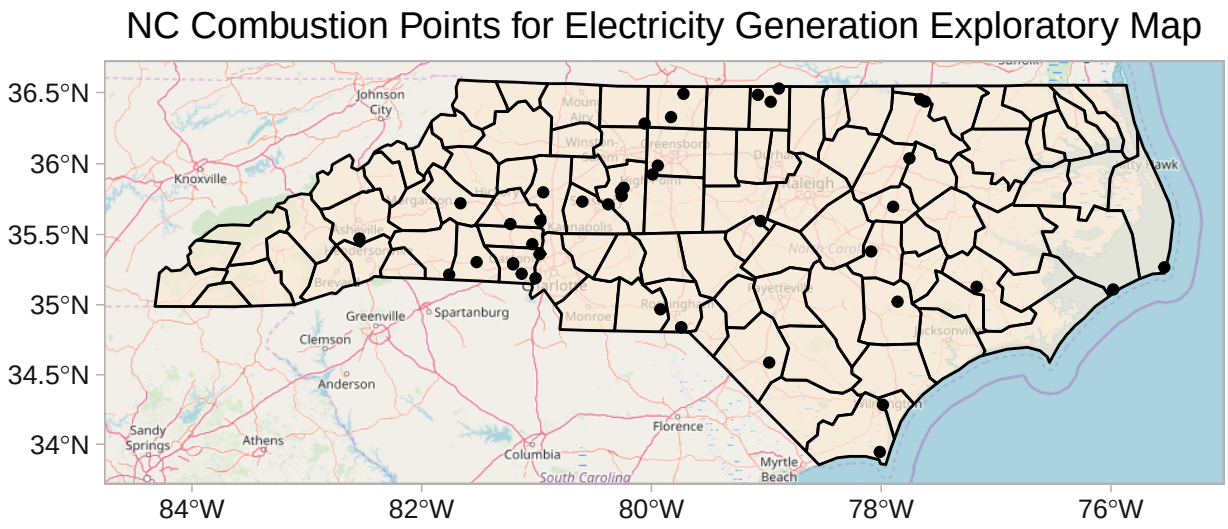

Figure 12: Combustion points for electricity generation in the North Carolina.

```
## Coordinate Reference System:
##   EPSG: 4326
##   proj4string: "+proj=longlat +datum=WGS84 +no_defs"
```

```
#Convert all to UTM Zone 17 (crs = 26917)
EPA_US_CombEmissions_utm <- st_transform(EPA_US_CombEmissions, c=26917)

#Clip the EPA_US_CombEmissions data set by the NC State boundary dataset

#First create a State_sf file
#Aggregate the data using group_by and summarize
state_sf_utm <- st_union(counties_sf_utm)

#Eliminate the emission points outside NC
EPA_NC_CombEmissions_utm <- st_intersection(EPA_US_CombEmissions_utm,state_sf_utm)
```

Visual data exploration of the EPA combustion points for electricity generation in the North Carolina in Figure 8, Figure 9, Figure 12, and Figure 11..

Now the distance between PM2.5 stations and Electricity Generation via Combustion points

is determined and added to the PM2.5_PM10_Temp_sf_utm dataframe.

```
#Distances between PM2.5 stations and Electricity Generation
#via Combustion points
Distances <- st_distance(PM2.5_Stations, EPA_NC_CombEmissions_utm)

a <- length(unique(PM2.5_Stations$Site.ID))

#Determining the minimum distance of each PM2.5 station to
#a combustion point in meters.
for (i in 1:a){
  PM2.5_Stations$Emiss_Dist[i] <- min(Distances[i,])
}

#Filling the PM2.5_PM10_Temp_sf_utm file with the distances

#Drops the geo data
aux4 <- PM2.5_Stations %>%
  st_set_geometry(value=NULL) %>%
  select(Site.ID,Emiss_Dist)

#Left_join the data
PM2.5_Full_utm <- PM2.5_PM10_Temp_sf_utm %>%
left_join(aux4, by = c("Site.ID")) %>%
  select(Date,Site.ID,COUNTY,Population,Zone,PM2.5,PM10,TAVG,Emiss_Dist,geometry)
```

Finally, using the elevatr package, elevation information is added to the PM 2.5 stations in the PM2.5_Full_utm dataframe, creating the PM2.5_Full_Elev_utm, which is saved in the Project folder ./Data/Processed.

Elevations for the PM2.5 Stations

```
prj_dd <- "+proj=utm +zone=17 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs"
PM2.5_Full_Elev_utm <- get_elev_point(PM2.5_Full_utm, prj = prj_dd, src = "epqs")

st_write(PM2.5_Full_Elev_utm,
         "./Data/Processed/PM2.5_Full_Elev_utm.shp", driver = "ESRI Shapefile")
```

```
PM2.5_Full_Elev_utm <- st_read('./Data/Processed/PM2.5_Full_Elev_utm.shp')
```

```
## Reading layer `PM2.5_Full_Elev_utm' from data source `C:\Users\Felipe\OneDrive - Duke
## Simple feature collection with 7460 features and 11 fields
## geometry type:  POINT
## dimension:      XY
## bbox:           xmin: 278314.3 ymin: 3807066 xmax: 935107.5 ymax: 3996703
## epsg (SRID):    NA
## proj4string:    +proj=utm +zone=17 +ellps=GRS80 +units=m +no_defs
```

3.6 Additional previsualization of the data

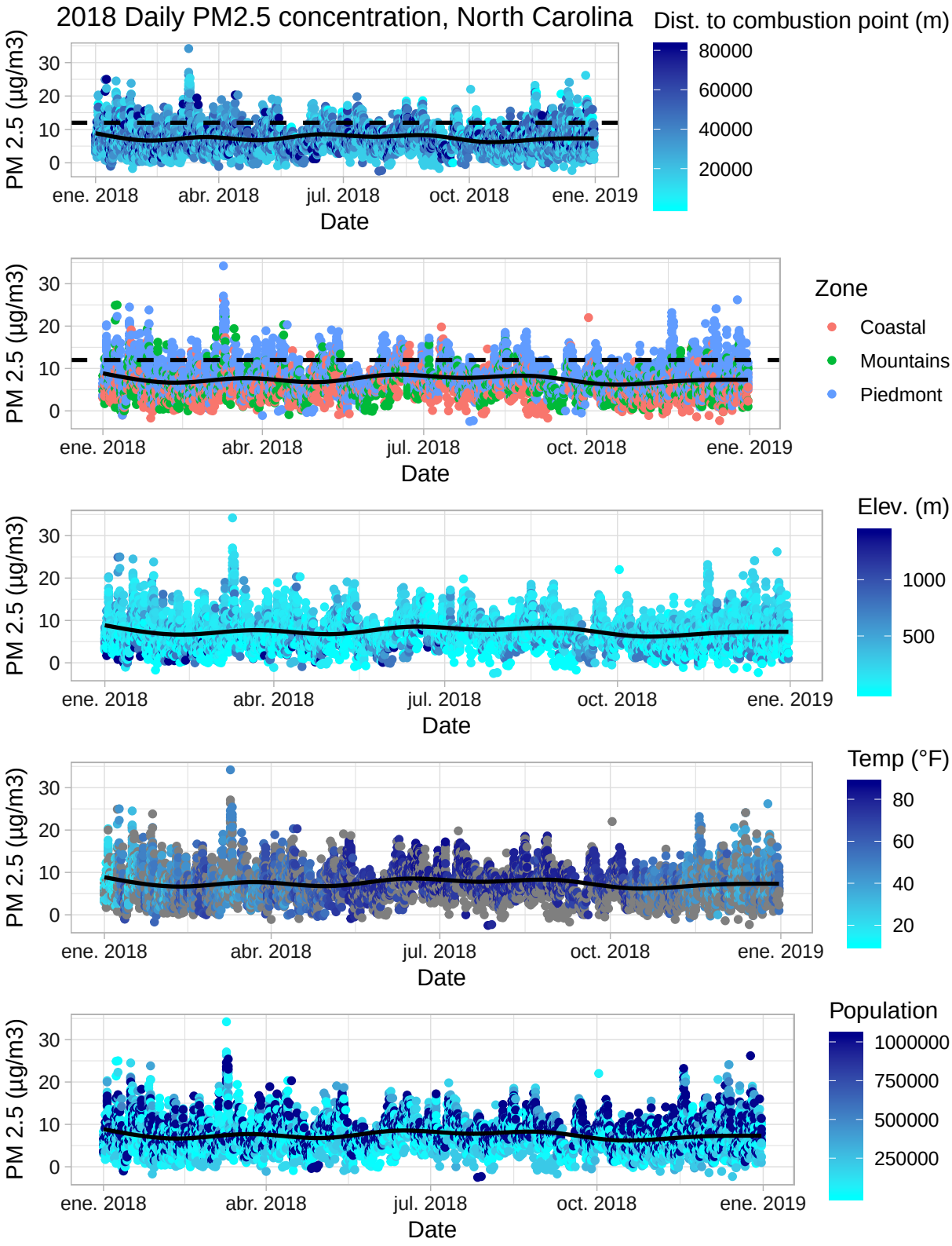

Figure 13: Daily PM2.5 Concentration.

4 Analysis

In 2012, the United States Environmental Protection Agency (USEPA) established a complementary primary regulatory standard for PM2.5 at a concentration of 12 micrograms per cubic meter, ug/m3< therefore, the first statistical test should look at the standard in each station.

4.1 One-sample t-test

The first statistical analysis will test the null hypothesis that the mean of the PM 2.5 concentrations in North Carolina are below the regulatory standard of 12 micrograms per cubic meter for the three geographical zones (Coastal, Piedmont, and Mountains).

First the assumption of normal distribution is evaluated.

```
PM2.5_Coastal <- PM2.5_Full_Elev_utm$PM2_5[PM2.5_Full_Elev_utm$Zone == "Coastal"]
PM2.5_Coastal <- as.data.frame(PM2.5_Coastal)

PM2.5_Piedmont <- PM2.5_Full_Elev_utm$PM2_5[PM2.5_Full_Elev_utm$Zone == "Piedmont"]
PM2.5_Piedmont <- as.data.frame(PM2.5_Piedmont)

PM2.5_Mountains <- PM2.5_Full_Elev_utm$PM2_5[PM2.5_Full_Elev_utm$Zone == "Mountains"]
PM2.5_Mountains <- as.data.frame(PM2.5_Mountains)
```

```
shapiro.test(PM2.5_Coastal$PM2.5_Coastal)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  PM2.5_Coastal$PM2.5_Coastal
## W = 0.9786, p-value = 1.564e-15
```

```
shapiro.test(PM2.5_Piedmont$PM2.5_Piedmont)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  PM2.5_Piedmont$PM2.5_Piedmont
## W = 0.96427, p-value < 2.2e-16
```

```
shapiro.test(PM2.5_Mountains$PM2.5_Mountains)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  PM2.5_Mountains$PM2.5_Mountains
```

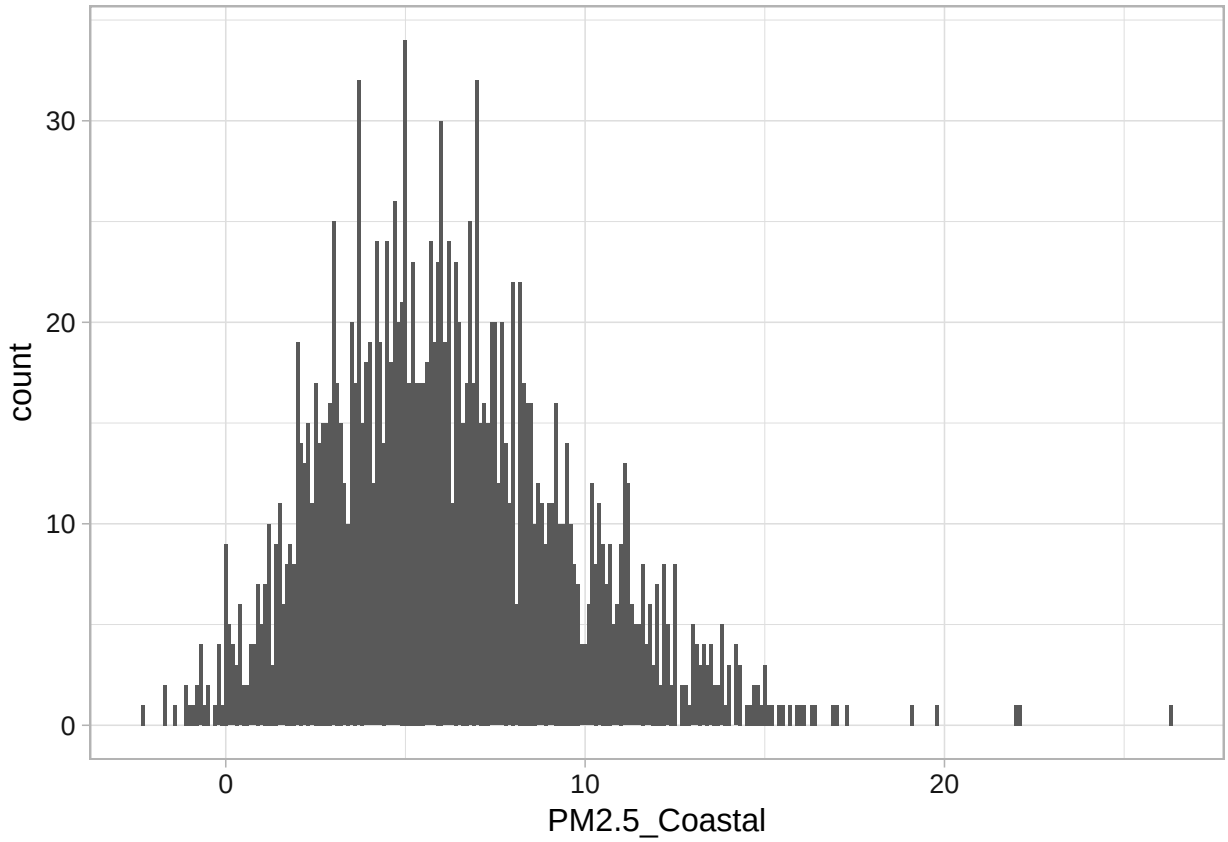

Figure 14: Daily PM2.5 Concentration Coastal Histogram.

```
## W = 0.95824, p-value < 2.2e-16
```

The Shapiro-Wilk normality test says that the PM 2.5 concentrations data in the three NC zones are significantly different from a normal distribution (Coastal: Shapiro-Wilk normality test, W = 0.9786, p-value < 0.0001; Piedmont: Shapiro-Wilk normality test, W = 0.96427, p-value < 0.0001; Mountain: Shapiro-Wilk normality test, W = 0.95824, p-value < 0.0001).

Next, a graphical analysis of the data is performed.

In Figure 14,Figure 15, Figure 16, Figure 17, Figure 18, and Figure 19 it can be seen that the data have more than one peak, heavier tails, and longer tail to the right than a normal distribution; nevertheless, environmental data often violate the assumptions of normality and the histograms fairly resembles a bell curve, so a t-test is performed anyway.

```
t.test(PM2.5_Coastal$PM2.5_Coastal, mu = 12, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  PM2.5_Coastal$PM2.5_Coastal
## t = -69.865, df = 1754, p-value < 2.2e-16
```

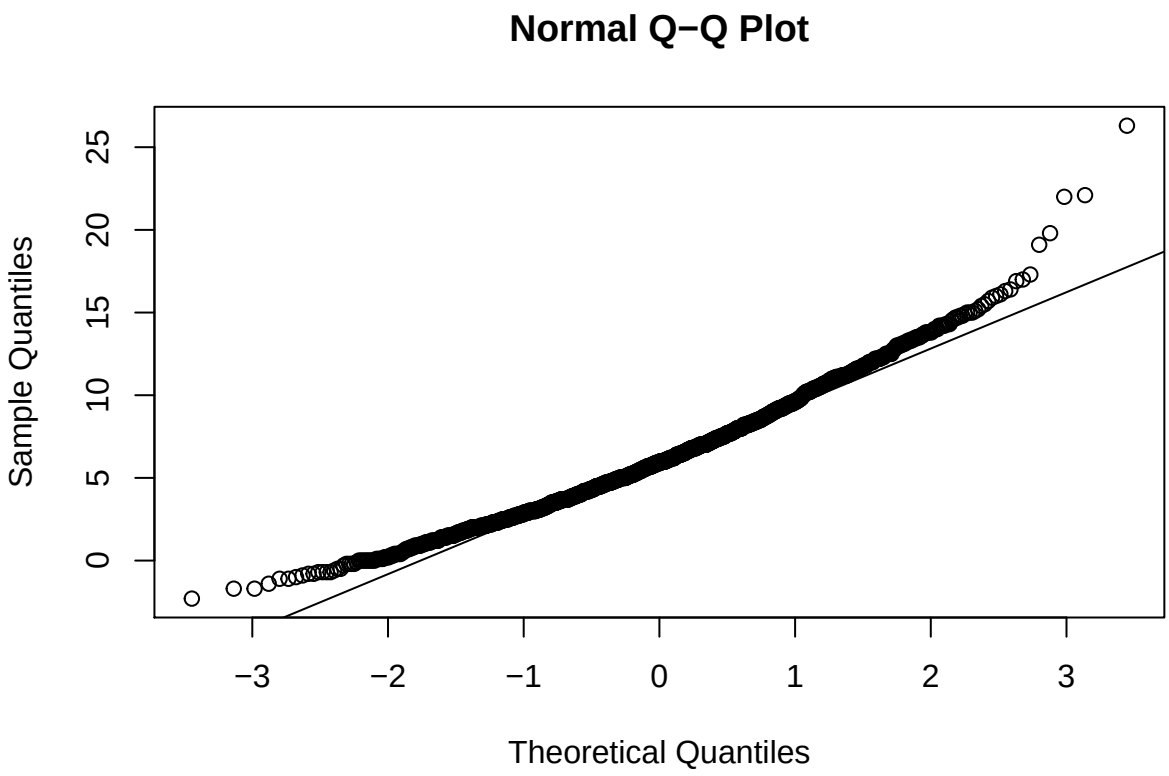

Figure 15: Daily PM2.5 Concentration Coastal qqplot.

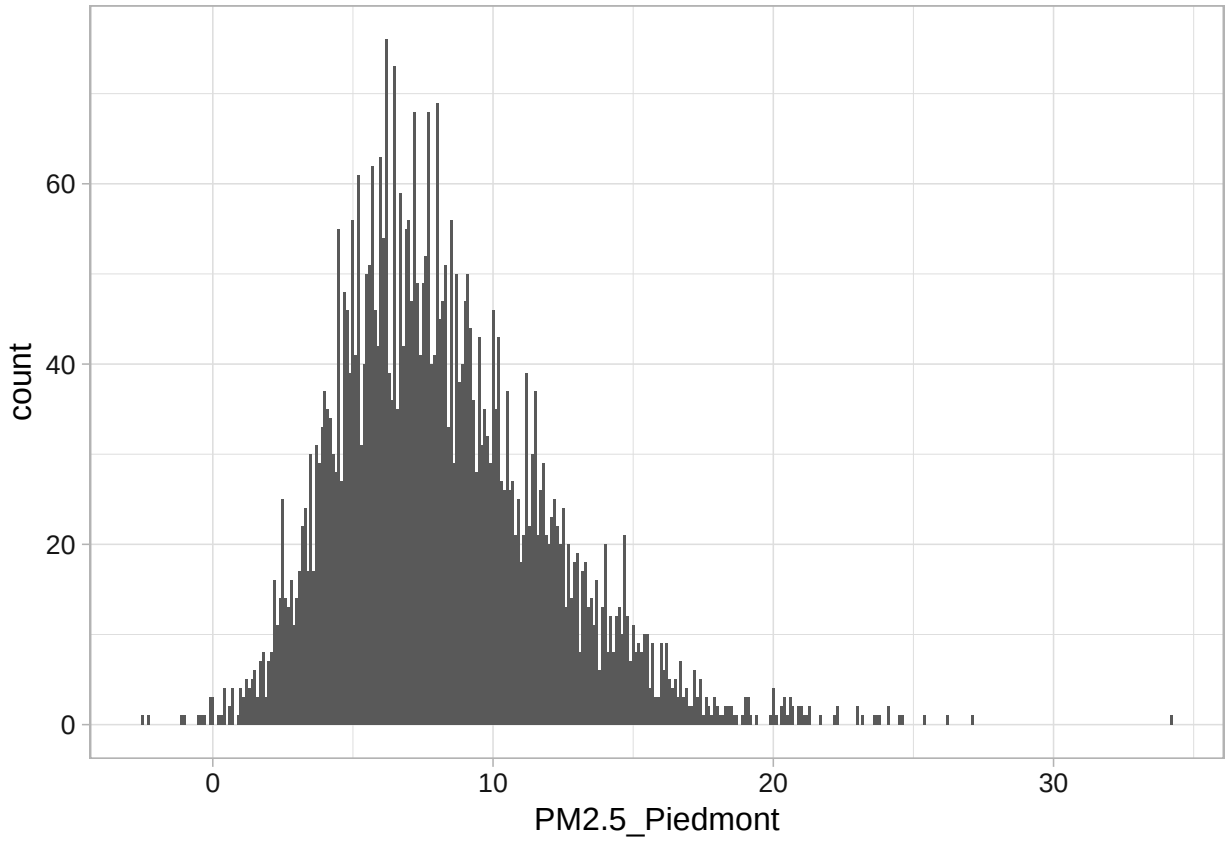

Figure 16: Daily PM2.5 Concentration Piedmont Histogram.

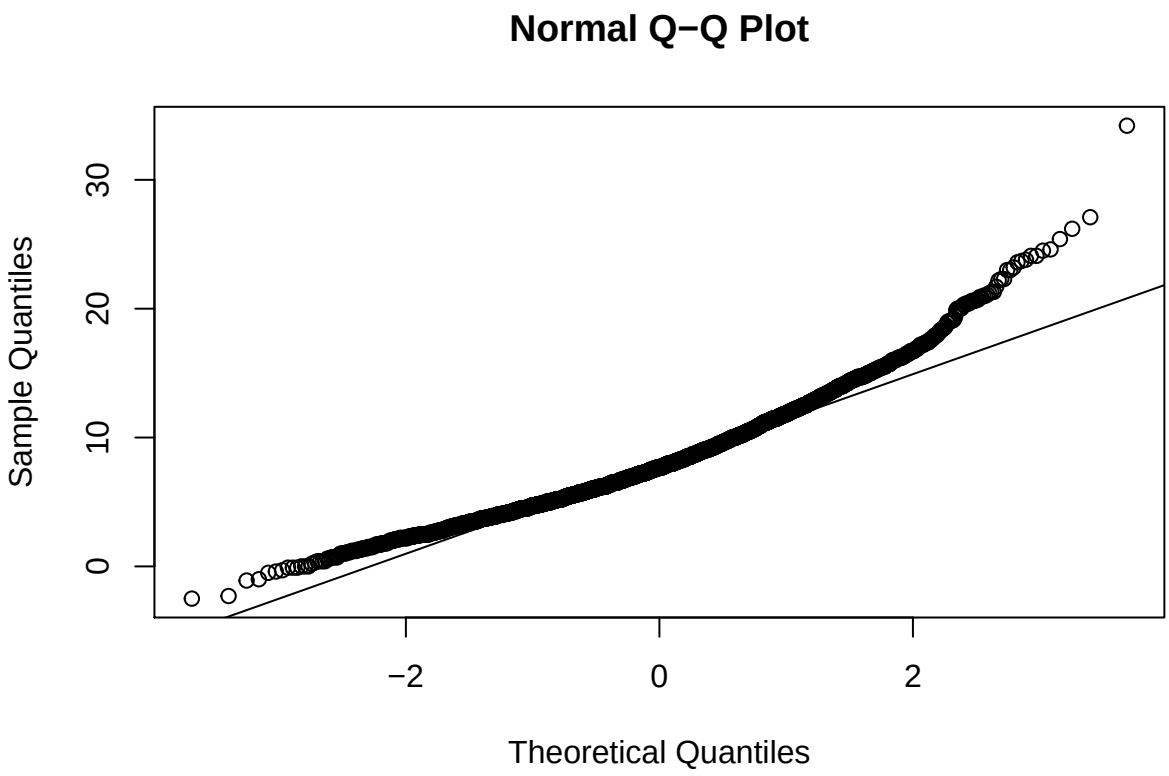

Figure 17: Daily PM2.5 Concentration Piedmont qqplot.

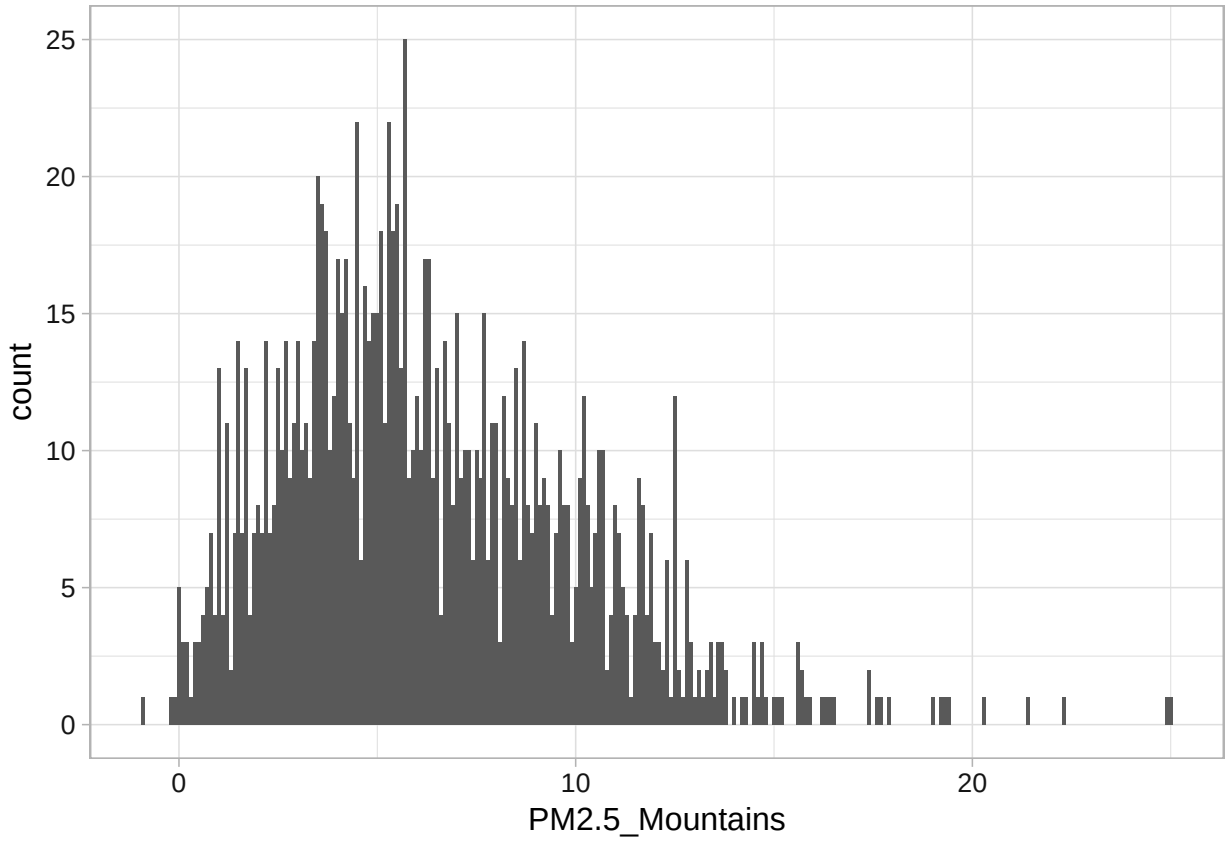

Figure 18: Daily PM2.5 Concentration Mountains Histogram.

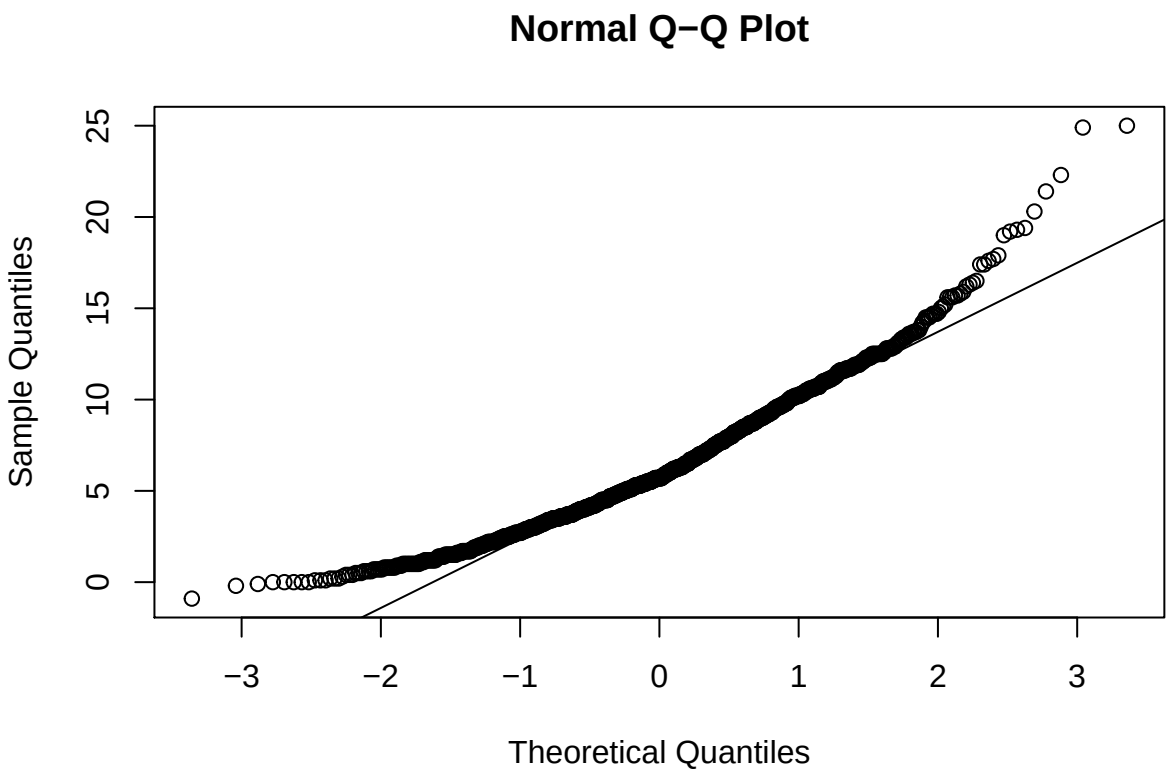

Figure 19: Daily PM2.5 Concentration Mountains qqplot.

```
## alternative hypothesis: true mean is less than 12
## 95 percent confidence interval:
##     -Inf 6.37724
## sample estimates:
## mean of x
##  6.241595
```

According to the One Sample t-test, NC Coastal PM 2.5 concentrations in 2018 were significantly lower than 12 ug/m3 (one sample t-test; t = –69.865, df =1754, p<0.0001).

```
t.test(PM2.5_Piedmont$PM2.5_Piedmont, mu = 12, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  PM2.5_Piedmont$PM2.5_Piedmont
## t = -67.406, df = 4433, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 12
## 95 percent confidence interval:
##      -Inf 8.292637
## sample estimates:
## mean of x
##  8.199887
```

According to the One Sample t-test, NC Piedmont PM 2.5 concentrations in 2018 were significantly lower than 12 ug/m3 (one sample t-test; t = -67.406, df = 4433, p-value < 0.0001).

```
t.test(PM2.5_Mountains$PM2.5_Mountains, mu = 12, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  PM2.5_Mountains$PM2.5_Mountains
## t = -53.529, df = 1270, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 12
## 95 percent confidence interval:
##      -Inf 6.563129
## sample estimates:
## mean of x
##  6.390637
```

According to the One Sample t-test, NC Mountains PM 2.5 concentrations in 2018 were significantly lower than 12 ug/m3 (one sample t-test; t = -53.529, df = 1270, p<0.0001).

According to the data and the One sample t-tests performed, the PM 2.5 concentrations in North Carolina are below the regulatory standard of 12 micrograms per cubic meter for the three geographical zones (Coastal, Piedmont, and Mountains).

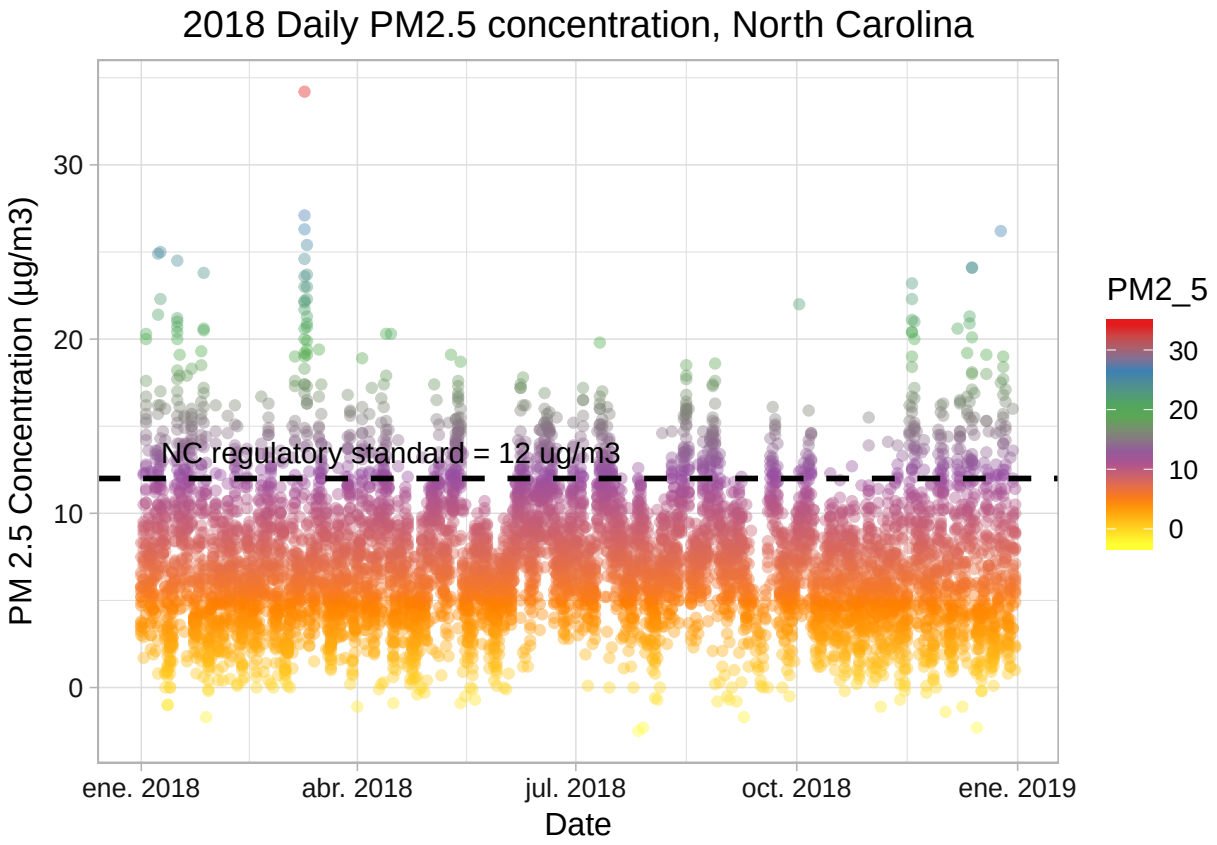

Figure 20: Daily PM2.5 Concentration, North Carolina.

In Figure 20 is presented a visualization of the PM 2.5 concentrations in North Carolina data in comparison with the North Carolina regulatory standard.

4.2 One-way ANOVA

Now that is known that the PM 2.5 concentrations in North Carolina in 2018 were below the regulatory standard of 12 micrograms per cubic meter for the three geographical zones, it is of interest to check if there are significant differences between the means of PM 2.5 concentrations for the same three geographical zones.

Therefore, the second statistical analysis will test the null hypothesis that the mean of the PM 2.5 concentrations in North Carolina are equal for the three geographical zones (Coastal, Piedmont, and Mountains).

For this, a One-way ANOVA test is performed. This test requires a second assumption to be complied, which is that the variance of the groups is equal across groups. Taking into account that the data are not perfectly normal, to test for the homogeneity of variance across groups a Fligner-Killeen test is used. This test is a non-parametric test, which is very robust against departures from normality.

```
fligner.test(PM2.5_Full_Elev_utm$PM2_5 ~ as.factor(PM2.5_Full_Elev_utm$Zone))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  PM2.5_Full_Elev_utm$PM2_5 by as.factor(PM2.5_Full_Elev_utm$Zone)
## Fligner-Killeen:med chi-squared = 7.6425, df = 2, p-value = 0.0219
```

The Fligner-Killeen test of homogeneity of variances says that the variance across groups is not homogeneous, but with a p-value close to 0.05 (med chi-squared = 7.6425, df = 2, p-value = 0.0219 < 0.05).

For this reason, for testing if there are significant differences between the means of PM 2.5 concentrations for the same three geographical zones, it is used a One-way ANOVA test and a Non-parametric equivalent of ANOVA, the Kruskal-Wallis Test.

```
PM2.5.anova <- lm(PM2.5_Full_Elev_utm$PM2_5 ~ as.factor(PM2.5_Full_Elev_utm$Zone))
anova(PM2.5.anova)
```

```
## Analysis of Variance Table
##
## Response: PM2.5_Full_Elev_utm$PM2_5
##                                      Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(PM2.5_Full_Elev_utm$Zone)    2   6480  3239.9  238.95 < 2.2e-16
## Residuals                           7457 101110    13.6
##
## as.factor(PM2.5_Full_Elev_utm$Zone) ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ktest_PM2.5 <- kruskal.test(PM2.5_Full_Elev_utm$PM2_5 ~
                            as.factor(PM2.5_Full_Elev_utm$Zone))
ktest_PM2.5
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  PM2.5_Full_Elev_utm$PM2_5 by as.factor(PM2.5_Full_Elev_utm$Zone)
## Kruskal-Wallis chi-squared = 471.69, df = 2, p-value < 2.2e-16
```

According to both test, there is a significant difference between the means of PM 2.5 concentrations for the same three geographical zones (ANOVA; F = 238.95, df = 7457, p<2.2e-16) and (Kruskal-Wallis chi-squared = 471.69, df = 2, p-value < 2.2e-16)

To analyze which zones are different, two *post hoc* tests were used, a Tukey multiple comparisons of means test for ANOVA and a Dunn's test for Kruskal-Wallis.

```r
# Run a post-hoc test for pairwise differences
TukeyHSD(aov(PM2.5_Full_Elev_utm$PM2_5 ~ as.factor(PM2.5_Full_Elev_utm$Zone)))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = PM2.5_Full_Elev_utm$PM2_5 ~ as.factor(PM2.5_Full_Elev_utm$Zone))
##
## $`as.factor(PM2.5_Full_Elev_utm$Zone)`
##                           diff        lwr       upr     p adj
## Mountains-Coastal  0.1490419 -0.1688845 0.4669682 0.5147468
## Piedmont-Coastal   1.9582918  1.7148601 2.2017235 0.0000000
## Piedmont-Mountains 1.8092499  1.5346120 2.0838879 0.0000000
```

```r
dunnTest(PM2.5_Full_Elev_utm$PM2_5 ~ as.factor(PM2.5_Full_Elev_utm$Zone))
```

```
##              Comparison           Z      P.unadj        P.adj
## 1  Coastal - Mountains  -0.5091436 6.106516e-01 6.106516e-01
## 2   Coastal - Piedmont -18.4333092 7.100397e-76 2.130119e-75
## 3 Mountains - Piedmont -15.7493971 6.933918e-56 1.386784e-55
```

Both test give as a result that there is a significant difference between the means of PM 2.5 concentrations between the Piedmont and both the Mountains and the Coastal zone (TukeHSD; p-value<0.0001) and (Dunn; p-value<0.0001). There is no significant difference between the means of PM 2.5 concentrations between the Mountains and the Coastal zone (TukeHSD; p-value>0.05) and (Dunn; p-value>0.05).

To explore graphically these results and present a visualization of the results of this test, in Figure 21} is presented a Boxplot for the Daily PM2.5 Concentration for the three North Carolina Zones.

4.3 Linear model

Finally, to answer the specific research question of this study (What are the effects of temperature, PM10 concentration, zoning (piedmont, coastal, mountain), population, elevation, distance to combustion points for electricity generation, in PM2.5 concentrations within North Carolina in the year 2018?), an analysis of covariance (ANCOVA) test was performed.

In this study multiple multiple explanatory variables are being considered at a time in the model, so to check the model is not over-parameterized the Akaike's Information Criterion (AIC) was used.

```r
PM2.5_Full_for.model <-
PM2.5_Full_Elev_utm %>%
na.exclude()
```

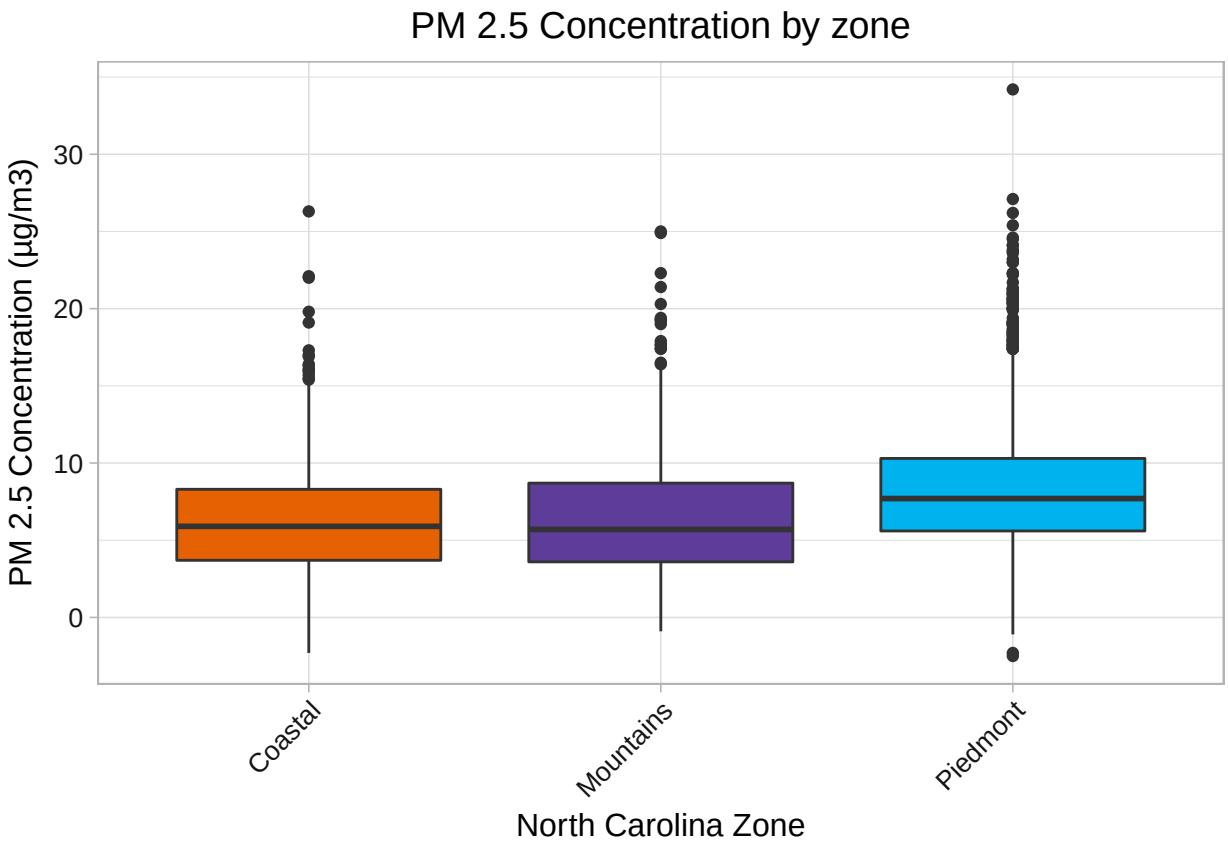

Figure 21: Daily PM2.5 Concentration Zones Boxplot.

```r
PM252018.model <-
  lm(data = PM2.5_Full_for.model, PM2_5 ~
       Population + as.factor(Zone) + PM10 + TAVG + Emiss_Dist + elevation)

step(PM252018.model) # the lower AIC value the better
```

```
## Start:  AIC=2904.83
## PM2_5 ~ Population + as.factor(Zone) + PM10 + TAVG + Emiss_Dist +
##     elevation
##
##                   Df Sum of Sq     RSS    AIC
## - Emiss_Dist       1       0.1  9361.0 2902.9
## - as.factor(Zone)  1       9.4  9370.3 2904.5
## <none>                          9360.9 2904.8
## - elevation        1      43.5  9404.5 2910.7
## - Population       1     134.6  9495.5 2926.9
## - TAVG             1     415.7  9776.6 2976.1
## - PM10             1   11848.9 21209.8 4282.7
##
## Step:  AIC=2902.85
## PM2_5 ~ Population + as.factor(Zone) + PM10 + TAVG + elevation
##
##                   Df Sum of Sq     RSS    AIC
## <none>                          9361.0 2902.9
## - as.factor(Zone)  1      13.5  9374.5 2903.3
## - Population       1     172.9  9533.9 2931.7
## - TAVG             1     415.8  9776.9 2974.2
## - elevation        1     429.5  9790.5 2976.5
## - PM10             1   11851.1 21212.1 4280.8
##
## Call:
## lm(formula = PM2_5 ~ Population + as.factor(Zone) + PM10 + TAVG +
##     elevation, data = PM2.5_Full_for.model)
##
## Coefficients:
##             (Intercept)               Population  as.factor(Zone)Piedmont
##               4.021e+00                1.862e-06                -5.372e-01
##                    PM10                     TAVG                  elevation
##               4.728e-01               -3.138e-02                -1.133e-02
```

The variable distance to combustion emission point was dropped by the Akaike's Information Criterion (AIC). According to this result, the set of explanatory variables that are best suited to predict PM 2.5 concentration are Population, Zone, PM10, TAVG, and elevation. Next, a multiple regression is performed on the recommended set of variables.

```
PM252018.model <-
  lm(data = PM2.5_Full_for.model, PM2_5 ~
       Population + as.factor(Zone) + PM10 + TAVG + elevation)
anova(PM252018.model)
```

```
## Analysis of Variance Table
##
## Response: PM2_5
##                    Df  Sum Sq Mean Sq   F value Pr(>F)
## Population          1   417.0   417.0   74.8771 <2e-16 ***
## as.factor(Zone)     1     0.4     0.4    0.0643 0.7999
## PM10                1 11194.9 11194.9 2010.3200 <2e-16 ***
## TAVG                1   428.9   428.9   77.0236 <2e-16 ***
## elevation           1   429.5   429.5   77.1205 <2e-16 ***
## Residuals        1681  9361.0     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the full model, only Zone is not statistically significant. The model is run without that variable.

```
PM252018.model <-
  lm(data = PM2.5_Full_for.model, PM2_5 ~
       Population + PM10 + TAVG + elevation)
summary(PM252018.model)
```

```
##
## Call:
## lm(formula = PM2_5 ~ Population + PM10 + TAVG + elevation, data = PM2.5_Full_for.mode
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7054  -1.2551   0.0033   1.2707  12.1720
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.168e+00  2.891e-01  14.414  < 2e-16 ***
## Population   1.450e-06  2.039e-07   7.111  1.7e-12 ***
## PM10         4.709e-01  1.018e-02  46.246  < 2e-16 ***
## TAVG        -3.096e-02  3.623e-03  -8.546  < 2e-16 ***
## elevation   -1.281e-02  8.719e-04 -14.691  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.361 on 1682 degrees of freedom
## Multiple R-squared:  0.5706, Adjusted R-squared:  0.5696
```

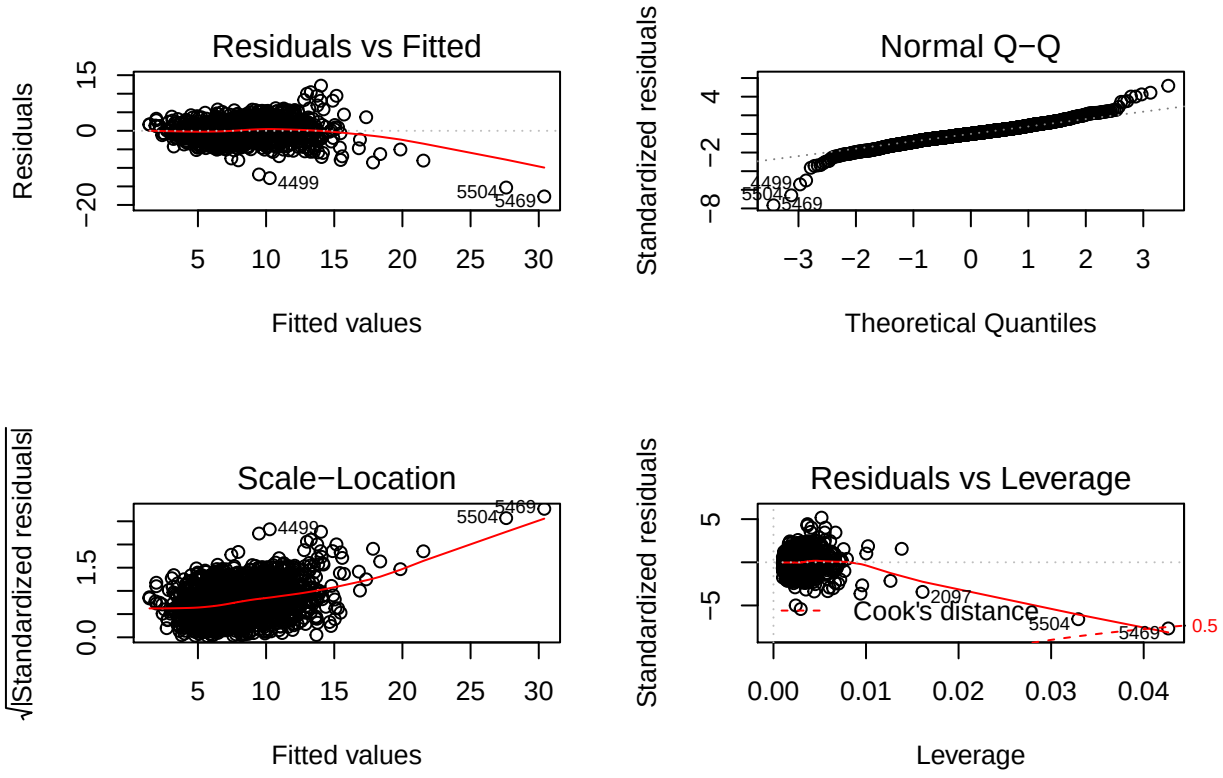

Figure 22: Model diagnostic plots.

```
## F-statistic: 558.8 on 4 and 1682 DF,  p-value: < 2.2e-16
```

Plotting the diagnostic plots of the model in Figure 22} to check the assumptions.

In Figure 22}, in the residuals vs fitted plot it can be seen an slightly increasing variance pattern but overall we can say that the points appear to be randomly scattered around the centerline. Also, in the normal Q-Q plot the points appear to follow an straight line (except for the extremes which is common for environmental data), so the graph is accepted. In the Scale- Location plot the points appear to be randomly scattered, so there is no evidence of heteroscedasticity and the homogeneous variance assumption is met. In the Residuals versus Leverage plots we check for outliers to verify that no single data point is so influential that leaving it out changes the structure of the model.

To make the coefficients easier to manipulate, let's first save the regression coefficients from our model model4 to a variable.

From the model we got that the intercept, 4.168, is the PM 2.5 mean daily concentration (ug/m3) when all the other variables are zero. We reject the null hypothesis of no effect of the explanatory variables on the response variable.

Population significantly increases the PM 2.5 mean daily concentration (ug/m3). With an

increase of population by 1 the PM 2.5 concentration is increased by 1.450e-06 ug/m3 (t = 7.111, df = 1682, p < 0.001).

PM10 daily mean concentration (ug/m3) significantly increases the PM 2.5 mean daily concentration (ug/m3). With an increase of PM10 daily mean concentration by 1 ug/m3 the PM 2.5 concentration is increased by 4.709e-01 ug/m3 (t = 46.246, df = 1682, p < 0.001).

Daily Average Temperature (°F) significantly decreases the PM 2.5 mean daily concentration (ug/m3). With an increase of daily average temperature by 1 °F the PM 2.5 concentration is decreased by 3.096e-02 ug/m3 (t = -8.546, df = 1682, p < 0.001).

Finally, elevation (meters) significantly decreases the PM 2.5 mean daily concentration (ug/m3). With an increase of elevation by 1 meter the PM 2.5 concentration is decreased by -1.281e-02 ug/m3 (t = -14.691, df = 1682, p < 0.001). The Adjusted R-squared = 0.5696, which is the fraction of total variance explained by the model.

The model explain 56.96% of the observed variance. The final linear equation to predict PM 2.5 mean daily concentration (ug/m3) from the explanatory variables is:

$$PM2.5 = 4.168 + 0.00000145 * Population + 0.471 * PM10 - 0.03096 * Temp - 0.0128 * Elev + \epsilon$$

In Figure 23} is presented the Daily PM2.5 Concentration vs Model Variables.

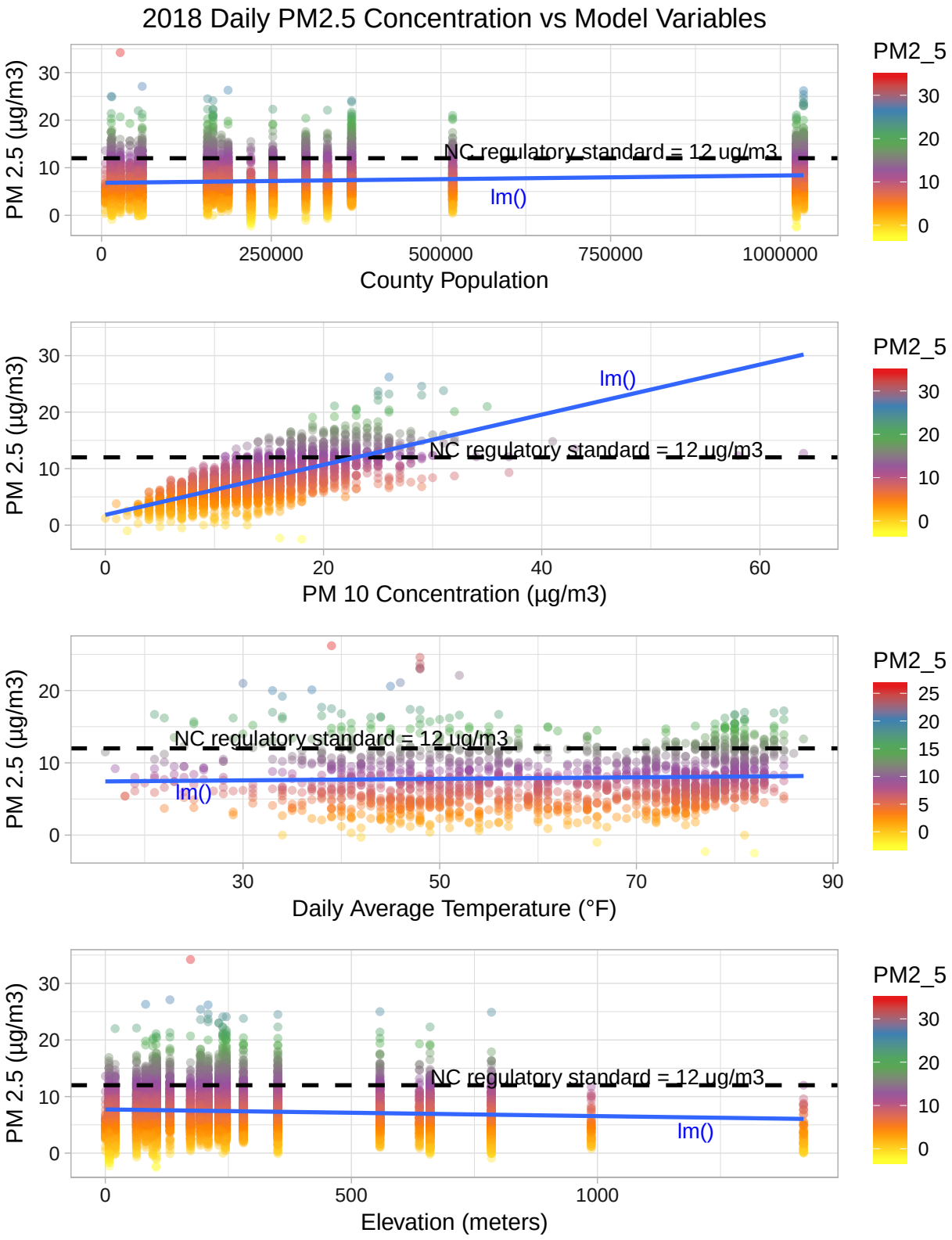

Figure 23: Daily PM2.5 Concentration vs Model Variables.

5 Summary and Conclusions

This study focused on trying to understand how PM2.5 concentration in North Carolina vary with temperature, PM10 concentration, zoning (piedmont, coastal, mountain), population, elevation, and distance to combustion points for electricity generation. The analysis performed indicated that during the year 2018, North Carolina had a mean PM 2.5 concentration below the regulatory standard of 12 micrograms per cubic meter for the three geographical zones (Coastal, Piedmont, and Mountains). Moreover, it was found that there is no significant difference between the mean PM 2.5 concentration of the Mountains and Coastal zones, yet there is a significant difference between the mean PM 2.5 concentration of these two zones and the Piedmont zone, which had a higher mean PM 2.5 concentration. This fact could be due to the fact that the piedmont area has the highest population, the biggest urban areas, and the most economic activity in North Carolina.

The final model indicates that there are multiple independent variables that explain a significant amount of the variation of PM 2.5 concentration in North Carolina. The data of 2018 was used to obtain the best fit linear models with the considered variable. The results showed that concentrations of PM2.5 and PM10, Temperature , Elevation, and county populations have a significant linear relationship during 2018. The model suggest that PM 2.5 concentrations increases with increase in county population and PM10 concentrations, and decreases with increases in Temperature and elevation.

Variables such as population, elevation, temperature, and PM 10 can be used to estimate PM 2.5 concentration levels in North Carolina, which can be an important tool in health management and disease prevention. With proper information managers can prepare for high concentration events, warn the population, take measures to lower concentrations, and mitigate the harmful effects that high PM2.5 concentrations have on human health.