Comparing Texas Emissions and Rainfall with Air Quality

https://github.com/maevegr/BassettGRHowey

Sara Bassett, Maeve Gualtieri-Reed, Xia Meng Howey

December 12, 2025

Contents

List of Tables

List of Figures

Rationale and Research Questions

This report examines how local conditions affect local air quality, focusing on power plant emissions and precipitation levels. While pollution is often studied at a global level, examining its contribution to climate change and warming temperatures, its affect on communities local to the emissions source is studied less frequently. This analysis considers other factors (precipitation) that may influence air quality and explores how air quality levels vary day-to-day. Texas was chosen as the study area for this analysis because of its number of power plants and counties. This report investigates the following three questions:

1) Is there a correlation between county level emission levels from power plants and county air quality in Texas?
2) Is there a trend in air quality values and emissions over time in Harris county, Texas?
3) Is there a relationship between precipitation and air quality in Harris county Texas?

Dataset Information

This report analyzes three sources of data: EPA Air Quality, Emissions & Generation Resource Integrated Database (eGRID), and *INSERT PRECIPITATION DATA*

EPA Air Quality Data is collected at local monitors across the United States. It can be downloaded by pollutant type, region, and year. This report uses data for Texas between 2018 and 2023 for PM2.5 and Ozone. PM2.5 and Ozone are two criteria pollutants that can cause respiratory issues, especially for people with pre-existing health conditions, children, and seniors. The AQI based on PM2.5 and Ozone are recorded separately and range on a scale from 0-500, with higher values being more dangerous. Values 0-50 indicate "Good" AQ, 50 - 100 "Moderate", and increasingly poor from there. For more information about Air Quality Index and criteria pollutants, visit https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information

eGRID data is released annually and contains information for all power plants across the United States. It contains information about generation, age of the plant, fuel type, and emissions, including CO2 Equivalent emissions. This report uses the eGRID 2018 - 2023 annual reports and focuses only on Texas power plants.

INSERT SOMETHING ABOUT PRECIPITATION DATA

Exploratory Analysis

This report analyzes air quality data from 2019 - 2023 in Texas for both PM2.5 and Ozone. For each year, there are separate PM2.5 and Ozone files, but the structure is the same for all. Below is the structure of the 2019 Texas PM2.5 data.

```
## 'data.frame':    18138 obs. of  22 variables:
##  $ Date                          : chr  "03/16/2019" "03/17/2019" "03/18/2019" "03/19/2019" ...
##  $ Source                        : chr  "AQS" "AQS" "AQS" "AQS" ...
##  $ Site.ID                       : int  480271045 480271045 480271045 480271045 480271045 480271045 4
##  $ POC                           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily.Mean.PM2.5.Concentration: num  4.4 6.3 7 11.1 12.7 5.6 7.5 8.4 7.7 14.3 ...
##  $ Units                         : chr  "ug/m3 LC" "ug/m3 LC" "ug/m3 LC" "ug/m3 LC" ...
##  $ Daily.AQI.Value               : int  24 35 39 55 58 31 42 47 43 61 ...
##  $ Local.Site.Name               : chr  "Temple Georgia" "Temple Georgia" "Temple Georgia" "Temple Ge
##  $ 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  88101 88101 88101 88101 88101 88101 88101 88101 88101 88101
##  $ AQS.Parameter.Description     : chr  "PM2.5 - Local Conditions" "PM2.5 - Local Conditions" "PM2.5
##  $ Method.Code                   : int  209 209 209 209 209 209 209 209 209 209 ...
##  $ Method.Description            : chr  "Met One BAM-1022 Mass Monitor w/ VSCC or TE-PM2.5C" "Met One
##  $ CBSA.Code                     : int  28660 28660 28660 28660 28660 28660 28660 28660 28660 28660
##  $ CBSA.Name                     : chr  "Killeen-Temple, TX" "Killeen-Temple, TX" "Killeen-Temple, TX
##  $ State.FIPS.Code               : int  48 48 48 48 48 48 48 48 48 48 ...
##  $ State                         : chr  "Texas" "Texas" "Texas" "Texas" ...
##  $ County.FIPS.Code              : int  27 27 27 27 27 27 27 27 27 27 ...
##  $ County                        : chr  "Bell" "Bell" "Bell" "Bell" ...
##  $ Site.Latitude                 : num  31.1 31.1 31.1 31.1 31.1 ...
##  $ Site.Longitude                : num  -97.4 -97.4 -97.4 -97.4 -97.4 ...
```

Analysis

Question 1: Is there a correlation between county level emission levels from power plants and county air quality?

The eGRID data for each year was adjusted to ensure all numeric columns were imported as numeric for quantitative analysis. The year, state, plant name, county, latitude, longitude, and emissions data for all types of recorded emissions (including CO2 equivalent emissions) were selected to a new data frame. These data frames were combined and further manipulated to have a data frame with the total county emissions for each year.

Relevant columns were selected for each of the PM2.5 and Ozone data frames, including date AQI, site name, AQ type (PM2.5 or Ozone), State, County, longitude and latitude. This was combined into two dataframes, one for PM2.5 and one for Ozone and further manipulated to have data frames with the average air quality value for each county each year.

The data frames were joined to have a data frame with, for each year and each county, total emissions and average AQI.

How often is emissions data reported?

The following map displays all the plants referenced in at least one of the eGRID reports between 2019 and 2023. Not all plants listed in eGRID report CO2 equivalence. The plants that reported at some point in the 5 years are in blue. Plants that are listed in a least one of the eGRID reports but never report CO2 emissions (or have zero emissions) are in black.

How are air quality stations spatially located compared to emissions reporting power plants?

The following map shows how plants and EPA air quality monitors are spatially related. Not all counties have both an air quality monitor and at least one CO2 equivalence reporting plant. Only counties reporting both at some point between 2019 and 2023 are included in the statistical analysis below.

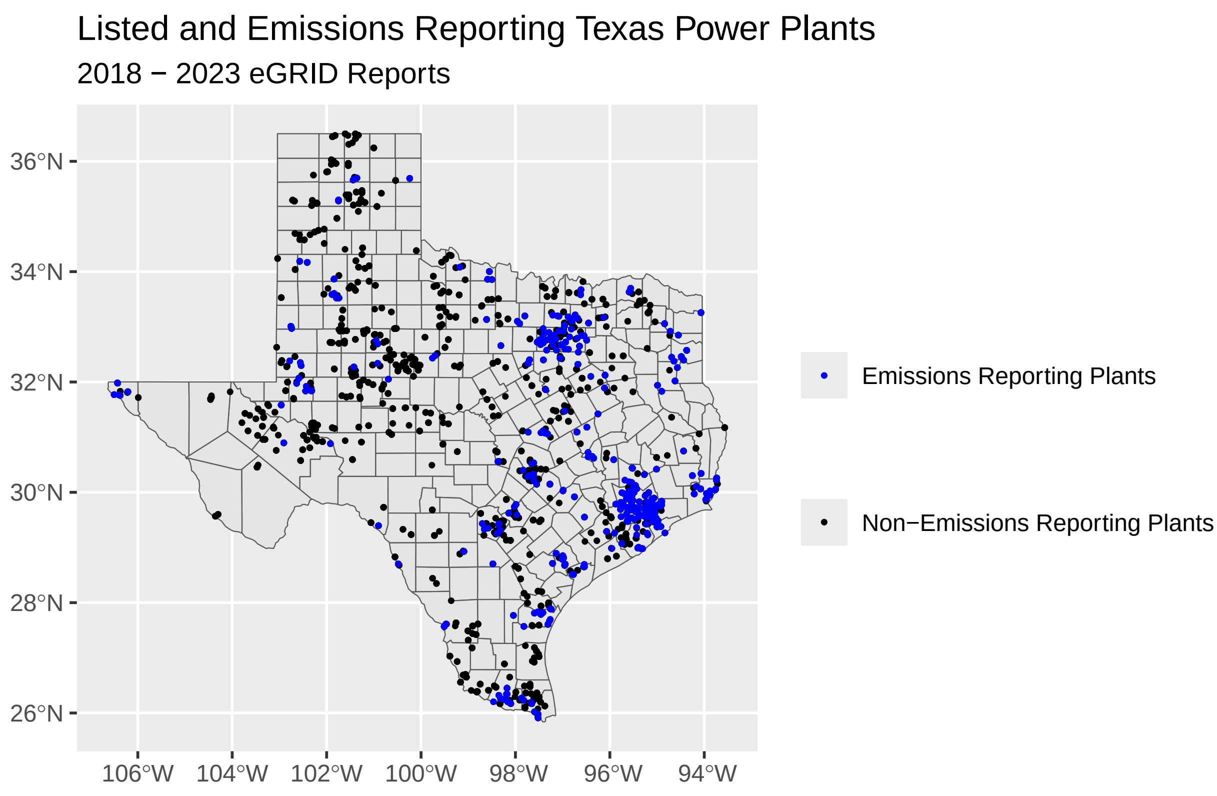


Figure 1: Listed and Emissions Reporting Texas Power Plants

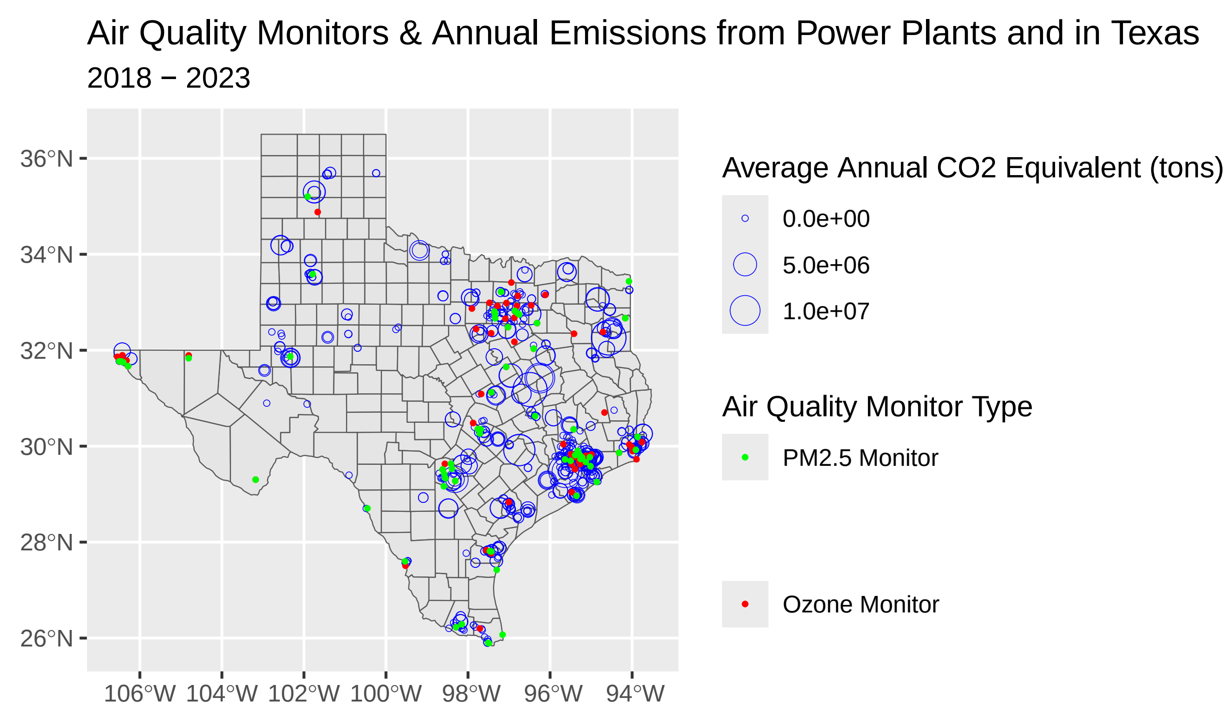


Figure 2: Air Quality Monitors & Annual Emissions from Power Plants and in Texas

Are annual county emissions correlated with annual county air quality?

The following graphs and tables show the relationship between total annual county emissions and average annual county air quality. A linear regression was run on the data to determine if the relationship has a significant correlation. Harris County data points are highlighted.

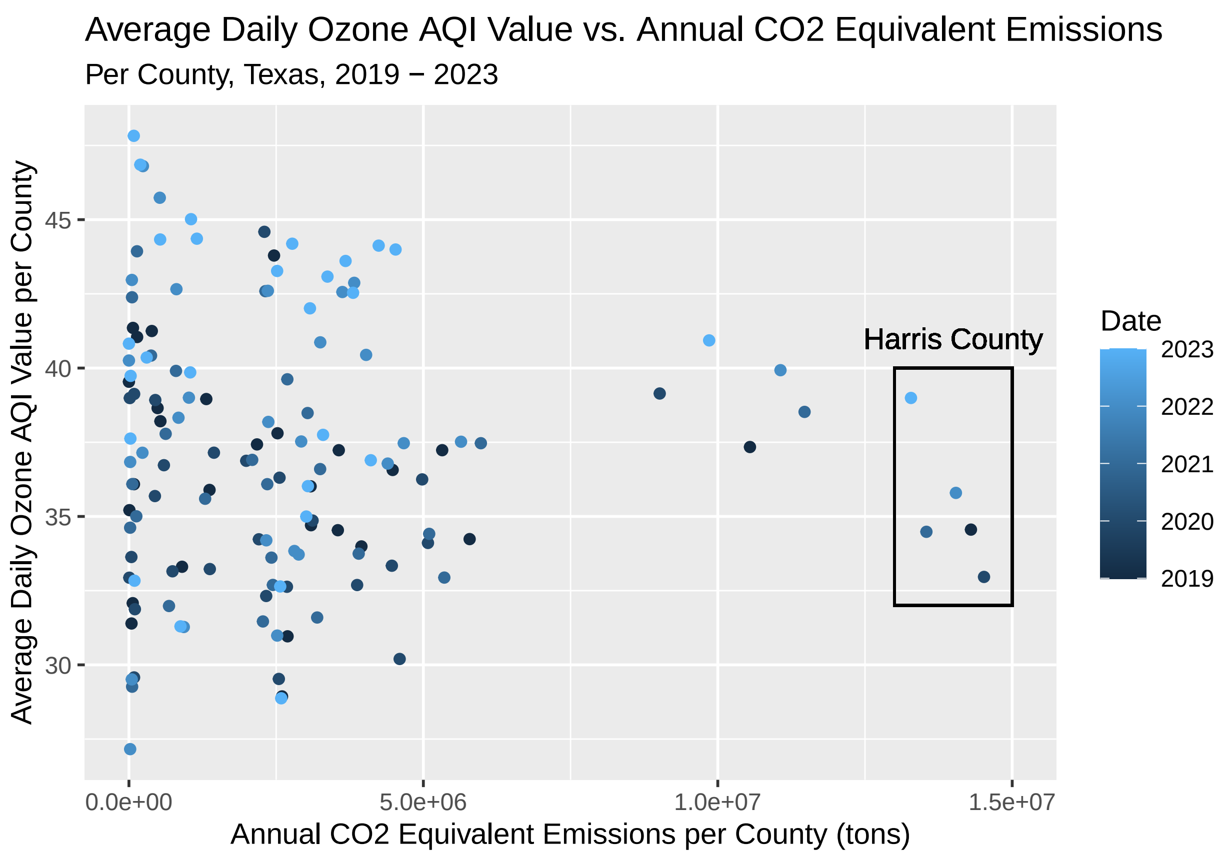


Figure 3: Average Daily Ozone AQI Value vs. Annual CO2 Equivalent Emissions

Table 1: Linear Regression of Emissions vs. Ozone Air Quality

Item	Value
P-value	0.5239
R-squared	-0.0043
Degrees of Freedom	136

Table 2: Linear Regression of Emissions vs. PM2.5 Air Quality

Item	Value
P-value	0.4074
R-squared	-0.0025
Degrees of Freedom	125

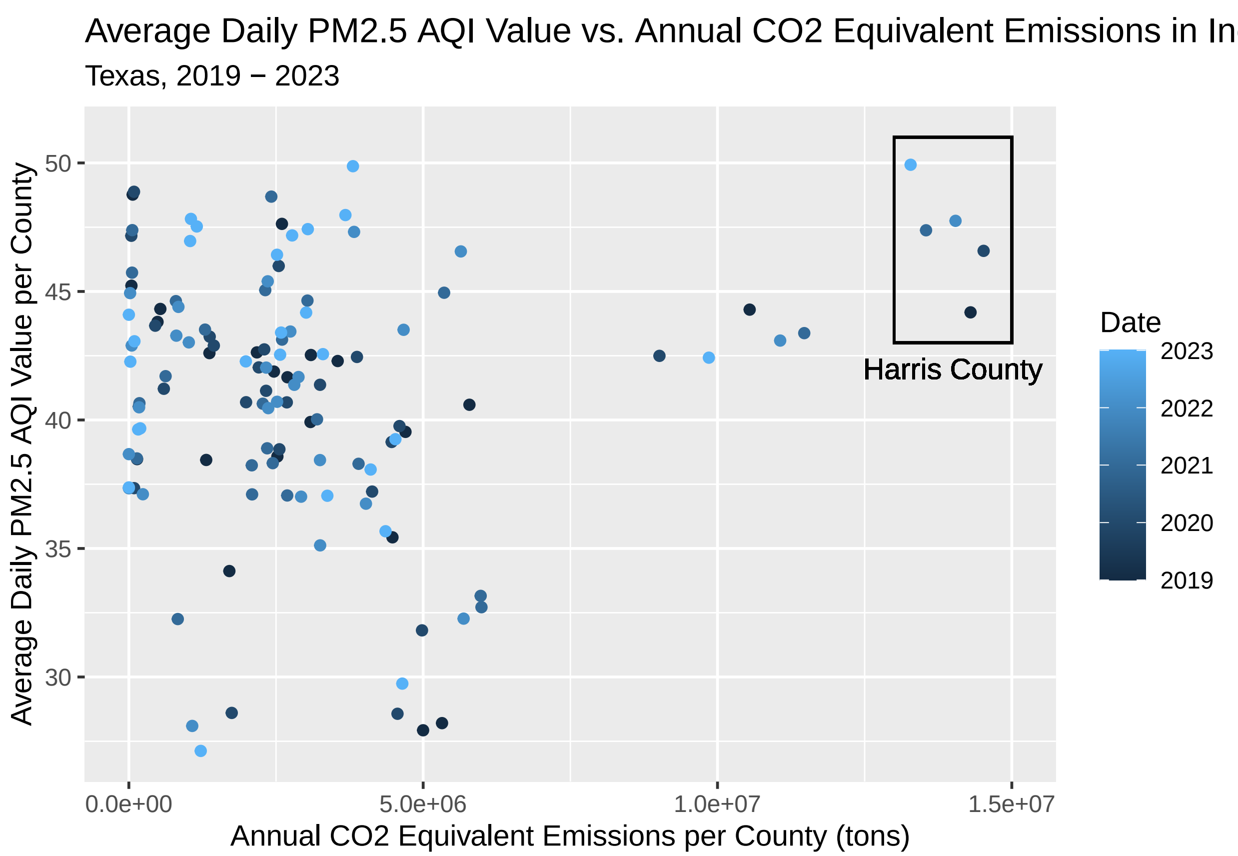


Figure 4: Average Daily PM2.5 AQI Value vs. Annual CO2 Equivalent Emissions in Individual Counties

Question 2: Is there a trend in air quality values and emissions over time in Harris county, Texas?

After analyzing the relationship between total CO2 equivalent emissions and AQI in PM2.5 and Ozone, we found no statistical correlation between either. To further investigate from a different approach and look at temporal trends, we focused on one area, Harris county, due to its high number of both emissions and air quality monitor data. This is shown in the Figure 2 map as the concentrated area in the southeastern part of Texas. In addition to analyzing CO2 equivalent emissions, methane emissions were analyzed. Unlike CO2, SO2, N2O which are all summed to calculate the total CO2 equivalent emissions, methane is a volatile organic compound (VOC) which can react with nitrogen oxides and sunlight to form tropospheric ozone. We wanted to test among these greenhouse gases, if methane specifically has a relationship with AQI.

```
## `summarise()` has grouped output by 'Plant.state.abbreviation',
## 'Plant.county.name'. You can override using the `.groups` argument.
```

```
## `summarise()` has grouped output by 'Date'. You can override using the
## `.groups` argument.
```

```
## `summarise()` has grouped output by 'Date'. You can override using the
## `.groups` argument.
```

The following plot shows the PM2.5 AQI over time, in Harris County, Texas. There is an overall trend of increasing PM2.5 AQI over the five-year period.

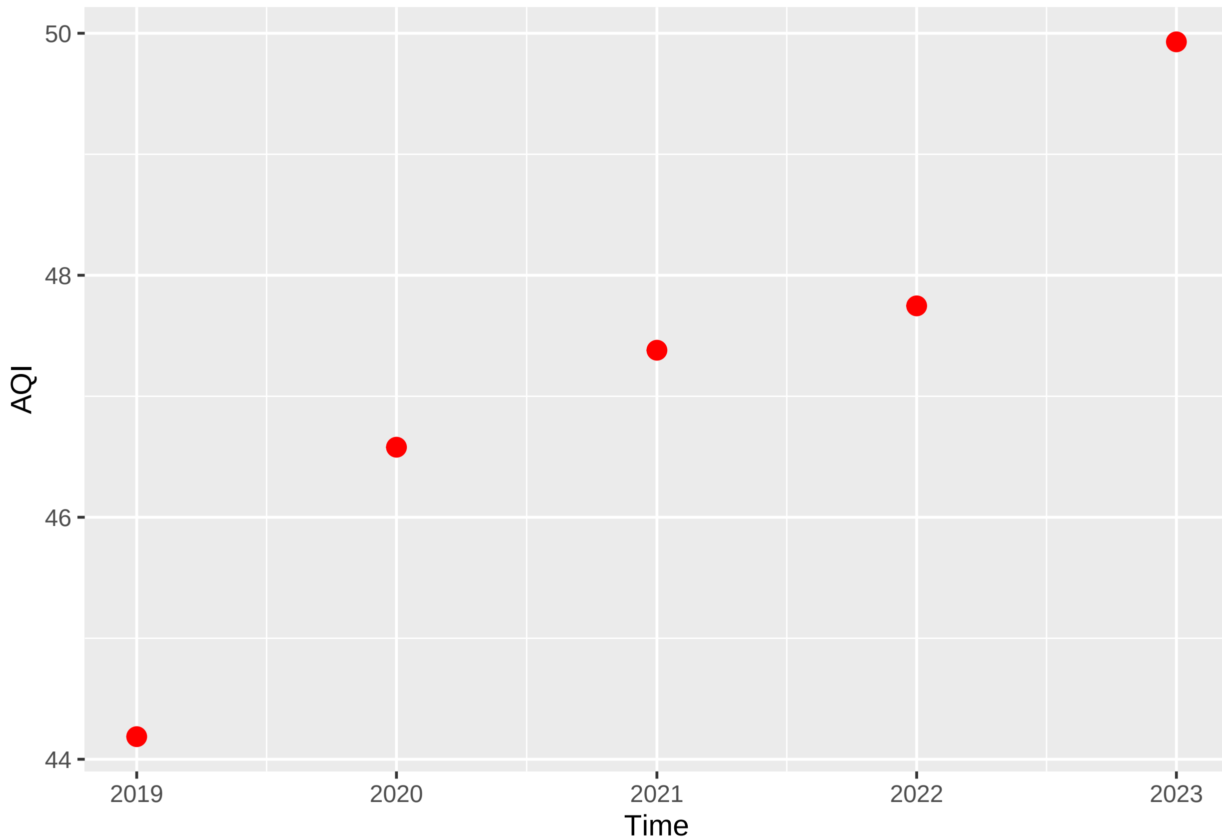


Figure 5: Average daily PM2.5 AQI in Harris County, Texas, 2019-2023

Figure 6 shows the ozone AQI over time in Harris County, Texas. There is an overall trend of increasing ozone AQI over the five-year period.

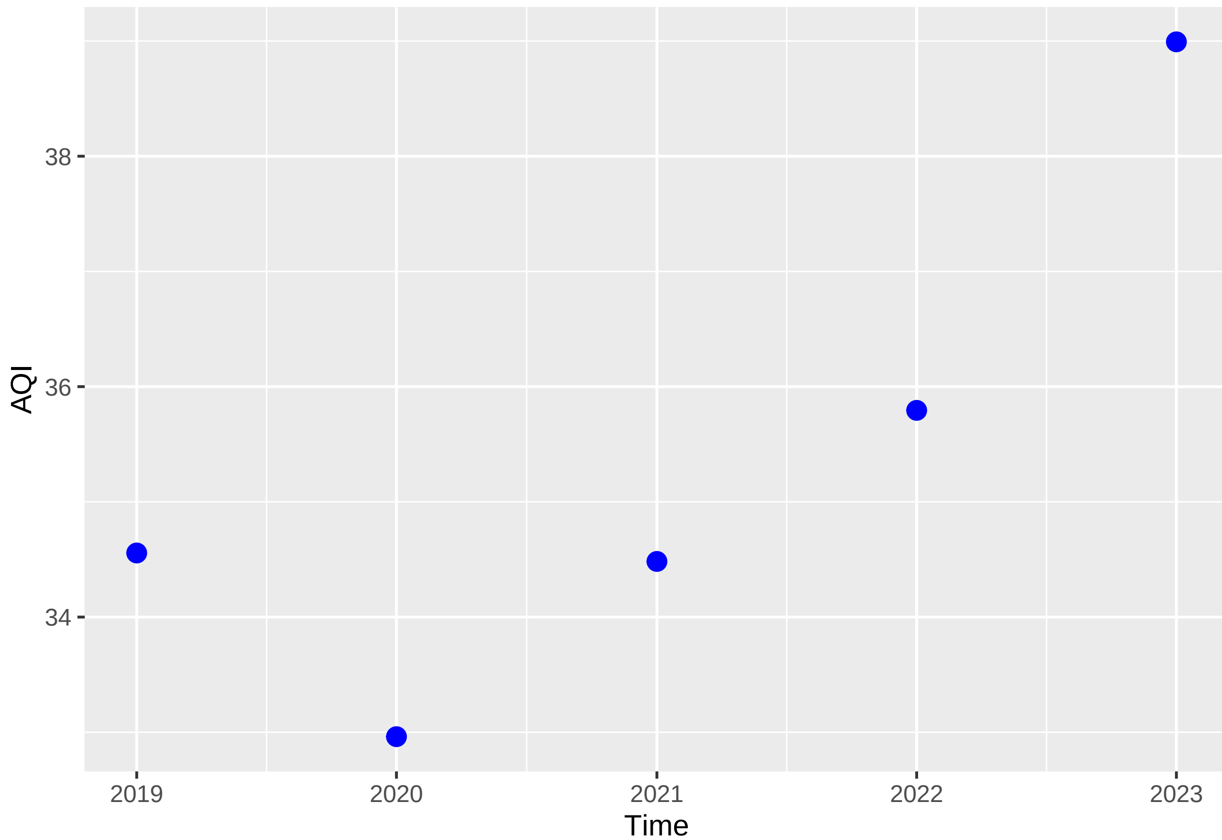


Figure 6: Average daily ozone AQI in Harris County, Texas, 2019-2023

Figure 7 shows the CO2 equivalent emissions over time in Harris County, Texas. There is an overall trend of decreasing CO2 equivalent emissions over the five-year period. The opposite trend of what occurred for PM2.5 and ozone AQI.

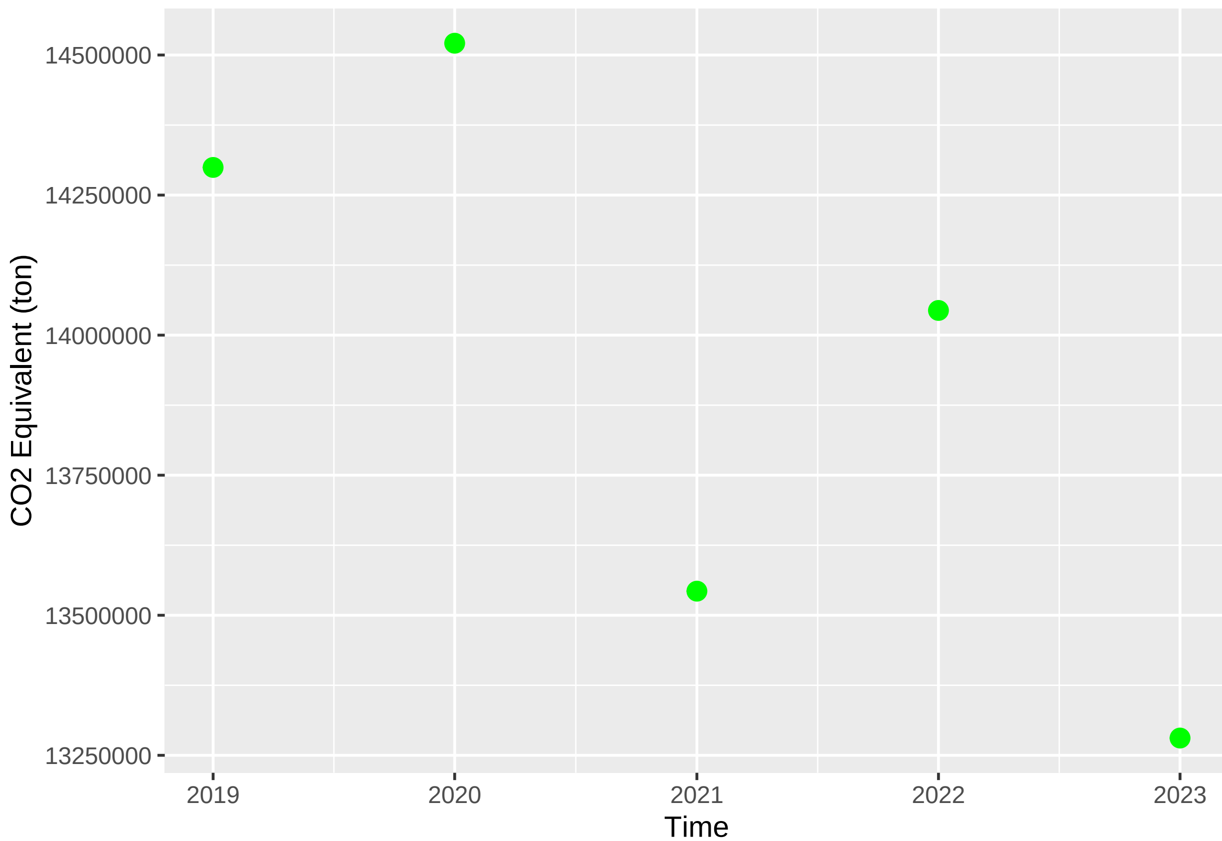


Figure 7: Annual CO2 equivalent in Harris County, Texas, 2019-2023

Figure 8 shows the methane equivalent emissions over time in Harris County, Texas. There is an overall trend of decreasing methane emissions over the five-year period. The opposite trend of what occurred for PM2.5 and ozone AQI.

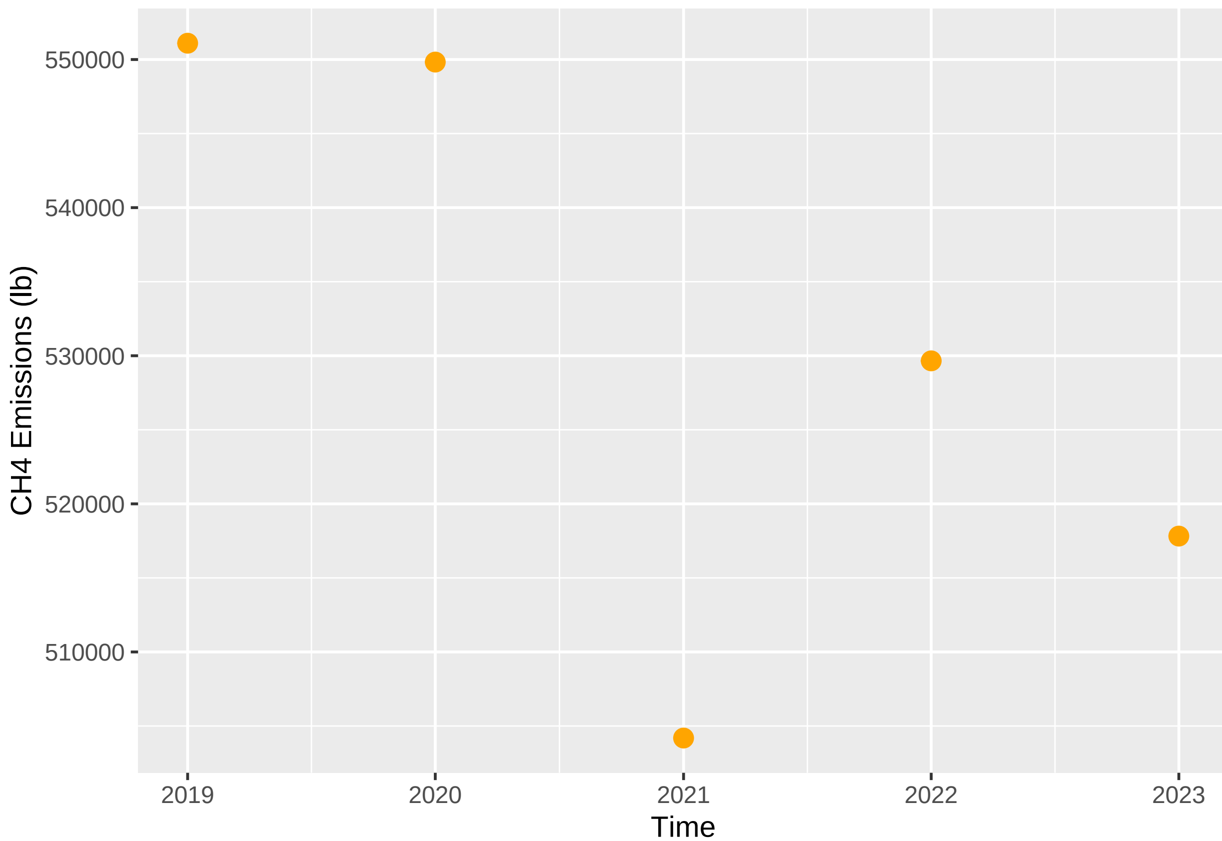


Figure 8: Annual methane emissions in Harris County, Texas, 2019-2023

Question 3:

```
#TX only PM2.5
TXPM2523_SB <- read.csv(here("Data/TXPM2523.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

TXPM2522_SB <- read.csv(here("Data/TXPM2522.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

TXPM2521_SB <- read.csv(here("Data/TXPM2521.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

TXPM2520_SB <- read.csv(here("Data/TXPM2520.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

TXPM2519_SB <- read.csv(here("Data/TXPM2519.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

#All Texas PM2.5 2018 - 2023 by County
TXPM25_SB <- rbind(TXPM2519_SB, TXPM2520_SB, TXPM2521_SB, TXPM2522_SB, TXPM2523_SB)

#TX only Ozone
TXOzone23_SB <- read.csv(here("Data/TXOzone23.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

TXOzone22_SB <- read.csv(here("Data/TXOzone22.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

TXOzone21_SB <- read.csv(here("Data/TXOzone21.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

TXOzone20_SB <- read.csv(here("Data/TXOzone20.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

TXOzone19_SB <- read.csv(here("Data/TXOzone19.csv")) %>%
  mutate(Date=mdy(Date)) %>%
  select("Date", "Daily.AQI.Value", "Local.Site.Name", "AQS.Parameter.Description", "State", "County",

#All Texas Ozone 2018 - 2023 by County
TXOzone_SB <- rbind(TXOzone19_SB, TXOzone20_SB, TXOzone21_SB, TXOzone22_SB, TXOzone23_SB)
```

```
TXOzoneHarris <- TXOzone_SB %>%
  filter(County == 'Harris')%>%
  group_by(Date) %>%
```

```
  summarise(Avg_AQI_Ozone = mean(Daily.AQI.Value),County='Harris')

TXPM25Harris <- TXPM25_SB %>%
  filter(County == 'Harris')%>%
  group_by(Date) %>%
  summarise(Avg_AQI_PM = mean(Daily.AQI.Value),County='Harris')

HarrisCoRain <- read.csv(here("Data/HarrisCoRain.csv"))%>%
  na.omit(PRCP)%>%
  group_by(DATE) %>%
  summarise(Avg_Rain = mean(PRCP))%>%
  mutate(Date=mdy(DATE))%>%
  select("Date","Avg_Rain")

HarrisCoRain_AQ1 <- left_join(TXOzoneHarris, HarrisCoRain)
```

```
## Joining with `by = join_by(Date)`
```

```
                              by = c("Date")

HarrisCoRain_AQ <- left_join(HarrisCoRain_AQ1, TXPM25Harris)
```

```
## Joining with `by = join_by(Date, County)`
```

```
                              by = c("Date")

HarrisRainOzone.regression <- lm(
  data = HarrisCoRain_AQ,
  Avg_Rain ~ Avg_AQI_Ozone)
summary(HarrisRainOzone.regression)
```

```
##
## Call:
## lm(formula = Avg_Rain ~ Avg_AQI_Ozone, data = HarrisCoRain_AQ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2110 -0.1475 -0.1149 -0.0268  4.6236
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.2297631  0.0195389  11.759  < 2e-16 ***
## Avg_AQI_Ozone -0.0025731  0.0004974  -5.173 2.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3642 on 1819 degrees of freedom
## Multiple R-squared:  0.0145, Adjusted R-squared:  0.01396
## F-statistic: 26.76 on 1 and 1819 DF,  p-value: 2.557e-07
```

```
HarrisRainPM.regression <- lm(
  data = HarrisCoRain_AQ,
  Avg_Rain ~ Avg_AQI_PM)
summary(HarrisRainPM.regression)
```

```
##
## Call:
## lm(formula = Avg_Rain ~ Avg_AQI_PM, data = HarrisCoRain_AQ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3520 -0.1483 -0.0873 -0.0071  4.4733
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4599711  0.0306523   15.01   <2e-16 ***
## Avg_AQI_PM  -0.0068218  0.0006266  -10.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3555 on 1819 degrees of freedom
## Multiple R-squared:  0.06117,    Adjusted R-squared:  0.06065
## F-statistic: 118.5 on 1 and 1819 DF,  p-value: < 2.2e-16
```

```
TXOzoneDallas <- TXOzone_SB %>%
  filter(County == 'Dallas')%>%
  group_by(Date) %>%
  summarise(Avg_AQI_Ozone = mean(Daily.AQI.Value),County='Dallas')

TXPM25Dallas <- TXPM25_SB %>%
  filter(County == 'Dallas')%>%
  group_by(Date) %>%
  summarise(Avg_AQI_PM = mean(Daily.AQI.Value),County='Dallas')

DallasCoRain <- read.csv(here("Data/DallasCoRain.csv"))%>%
  na.omit(PRCP)%>%
  group_by(DATE) %>%
  summarise(Avg_Rain = mean(PRCP))%>%
  mutate(Date=ymd(DATE))%>%
  select("Date","Avg_Rain")

DallasCoRain_AQ1 <- left_join(TXOzoneDallas, DallasCoRain)
```

```
## Joining with `by = join_by(Date)`
```

```
                              by = c("Date")

DallasCoRain_AQ <- left_join(DallasCoRain_AQ1, TXPM25Dallas)
```

```
## Joining with `by = join_by(Date, County)`
```

```r
                              by = c("Date")

DallasRainOzone.regression <- lm(
  data = DallasCoRain_AQ,
  Avg_Rain ~ Avg_AQI_Ozone)
summary(DallasRainOzone.regression)
```

```
##
## Call:
## lm(formula = Avg_Rain ~ Avg_AQI_Ozone, data = DallasCoRain_AQ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1785 -0.1261 -0.1016 -0.0335  4.1741
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.2017347  0.0187384  10.766  < 2e-16 ***
## Avg_AQI_Ozone -0.0022472  0.0004366  -5.147 2.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3328 on 1818 degrees of freedom
## Multiple R-squared:  0.01436,    Adjusted R-squared:  0.01382
## F-statistic: 26.49 on 1 and 1818 DF,  p-value: 2.933e-07
```

```r
DallasRainPM.regression <- lm(
  data = DallasCoRain_AQ,
  Avg_Rain ~ Avg_AQI_PM)
summary(DallasRainPM.regression)
```

```
##
## Call:
## lm(formula = Avg_Rain ~ Avg_AQI_PM, data = DallasCoRain_AQ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3467 -0.1386 -0.0574  0.0102  4.0097
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3813303  0.0219118   17.40   <2e-16 ***
## Avg_AQI_PM  -0.0061621  0.0004734  -13.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3164 on 1778 degrees of freedom
##   (40 observations deleted due to missingness)
## Multiple R-squared:  0.08701,    Adjusted R-squared:  0.0865
## F-statistic: 169.5 on 1 and 1778 DF,  p-value: < 2.2e-16
```

```r
HarrisDallas <- rbind(HarrisCoRain_AQ, DallasCoRain_AQ)

HarrisDallasRain.anova <- aov(data = HarrisDallas, Avg_Rain ~ County)
summary(HarrisDallasRain.anova)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)  
## County          1    0.6  0.5598   4.536 0.0333 *
## Residuals    3639  449.1  0.1234                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

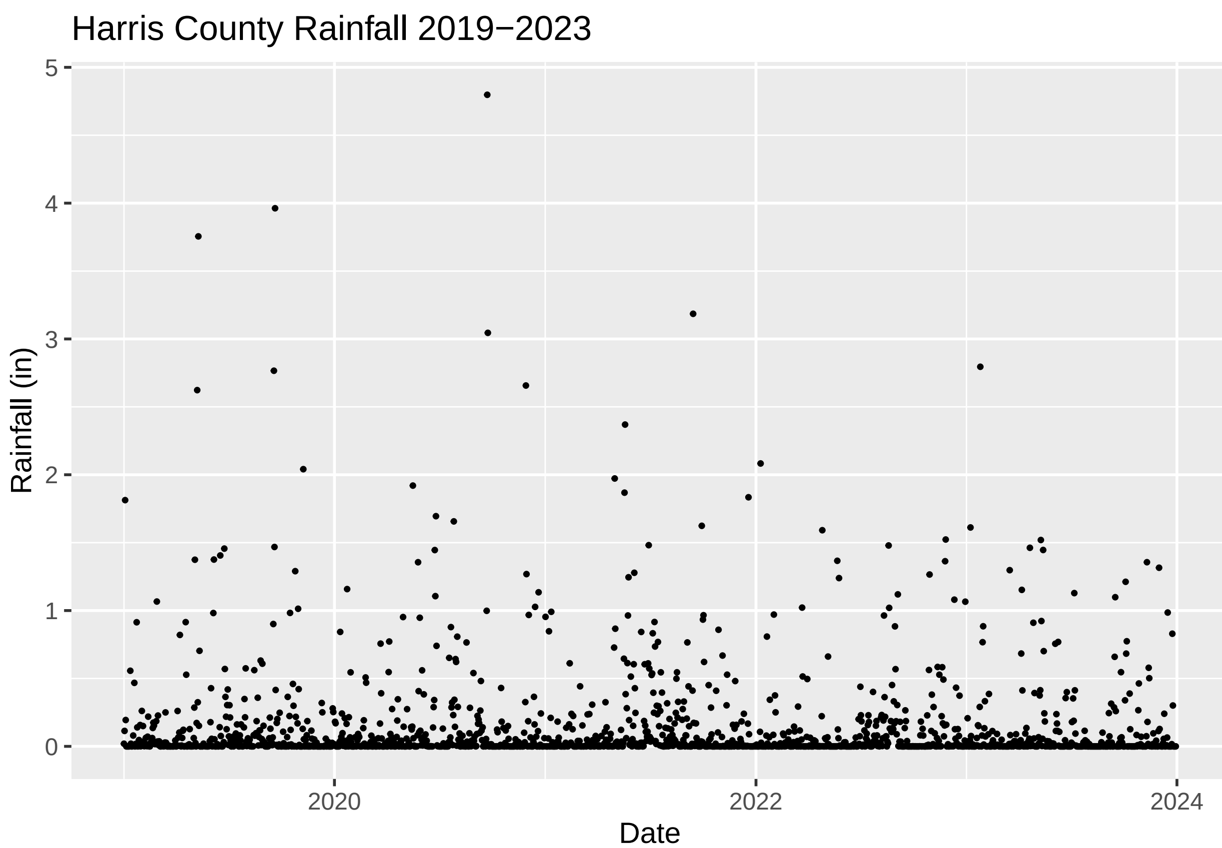
```r
HarrisDallasOzone.anova <- aov(data = HarrisDallas, Avg_AQI_Ozone ~ County)
summary(HarrisDallasOzone.anova)
```

```
##                Df  Sum Sq Mean Sq F value   Pr(>F)    
## County          1   12377   12377   40.31 2.43e-10 ***
## Residuals    3639 1117149     307                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

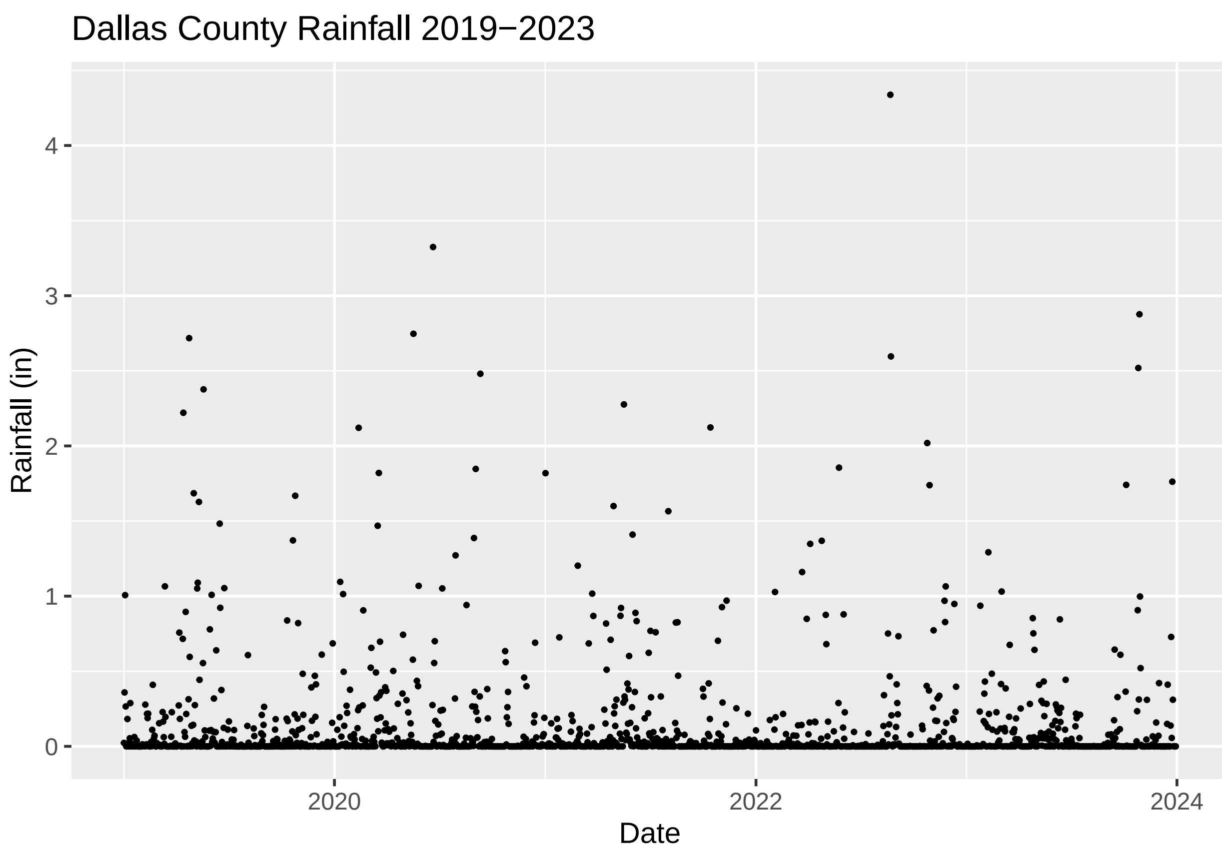
```r
HarrisDallasPM.anova <- aov(data = HarrisDallas, Avg_AQI_PM ~ County)
summary(HarrisDallasPM.anova)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## County          1  11549   11549   54.07 2.38e-13 ***
## Residuals    3599 768715     214                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 40 observations deleted due to missingness
```

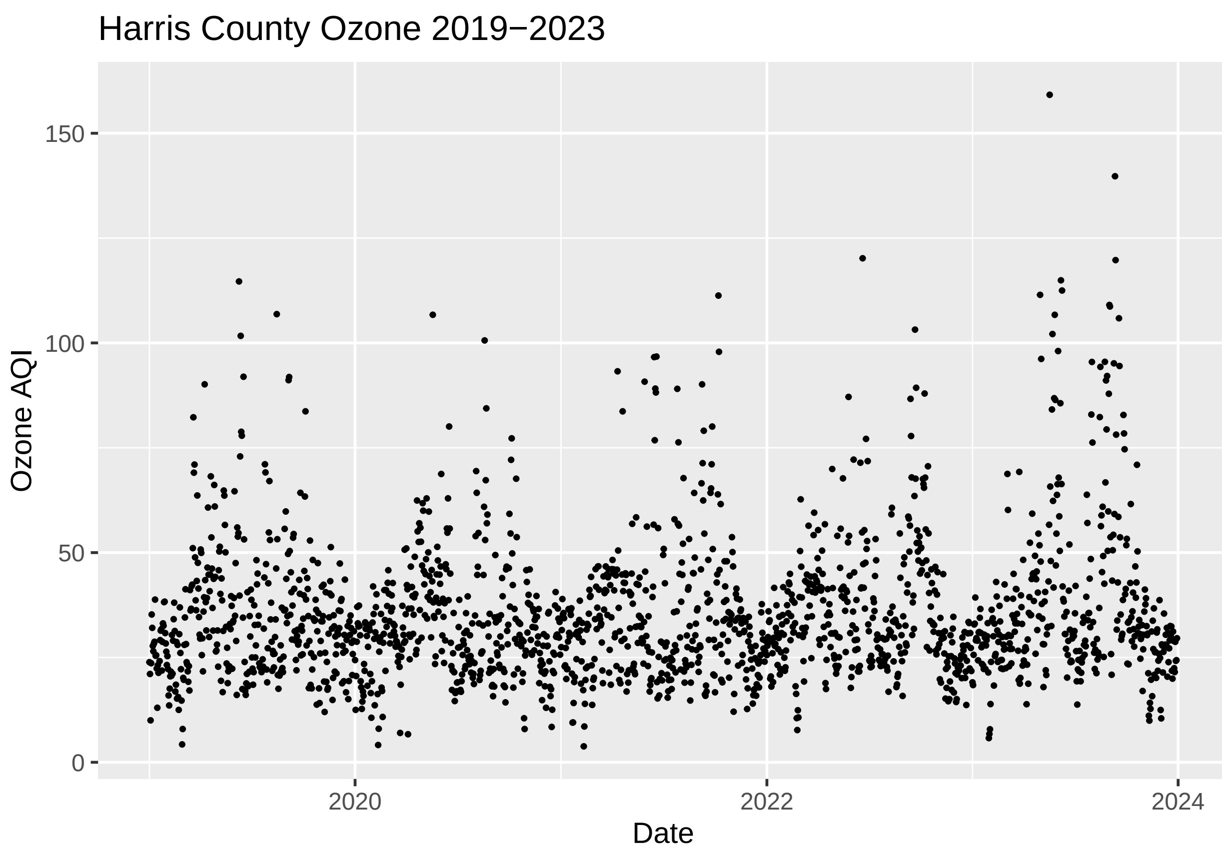
```r
Rainplot<-HarrisCoRain_AQ%>%
  ggplot(aes(
    x=Date,
    y=Avg_Rain))+
  labs(
    x="Date",
    y="Rainfall (in)")+
  geom_point(size=0.5)+
  ggtitle('Harris County Rainfall 2019-2023')
print(Rainplot)
```



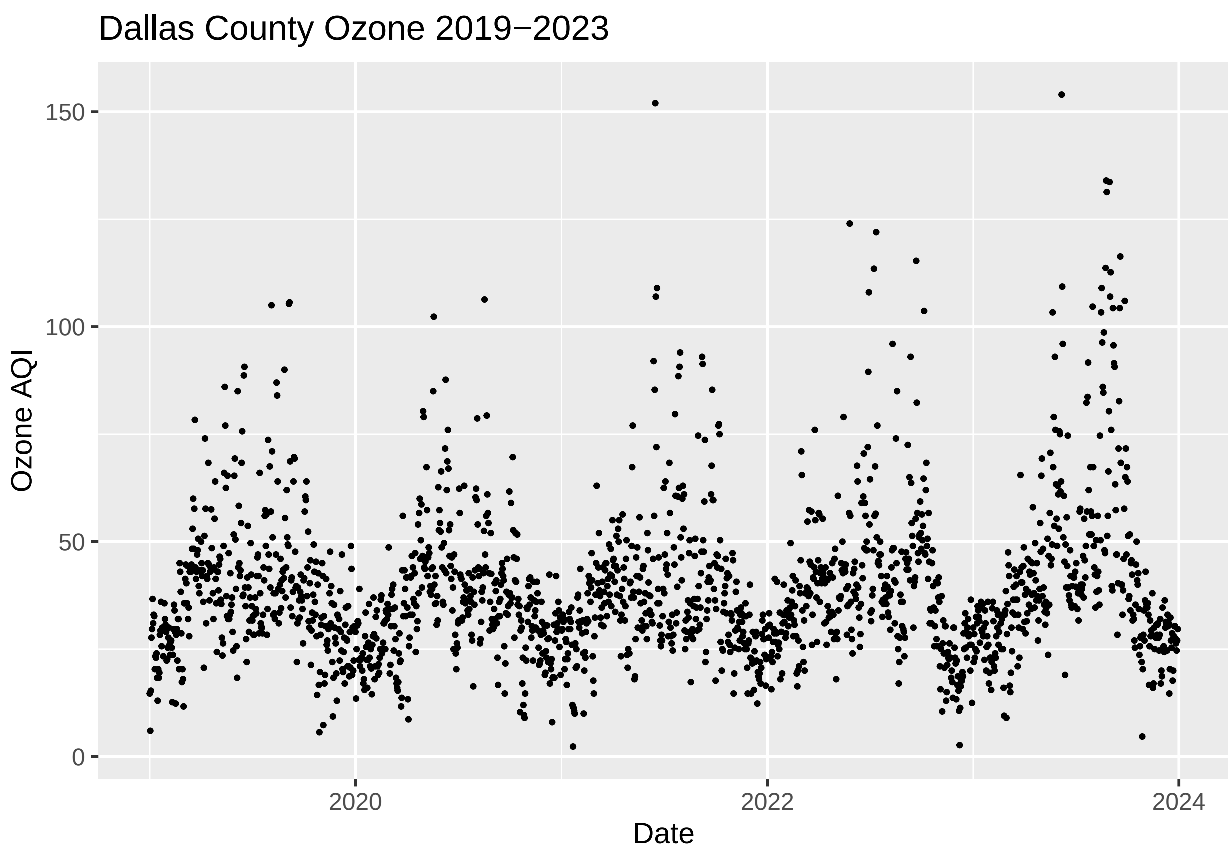

```r
Rainplot2<-DallasCoRain_AQ%>%
  ggplot(aes(
    x=Date,
    y=Avg_Rain))+
  labs(
    x="Date",
    y="Rainfall (in)")+
  geom_point(size=0.5)+
  ggtitle('Dallas County Rainfall 2019-2023')
print(Rainplot2)
```



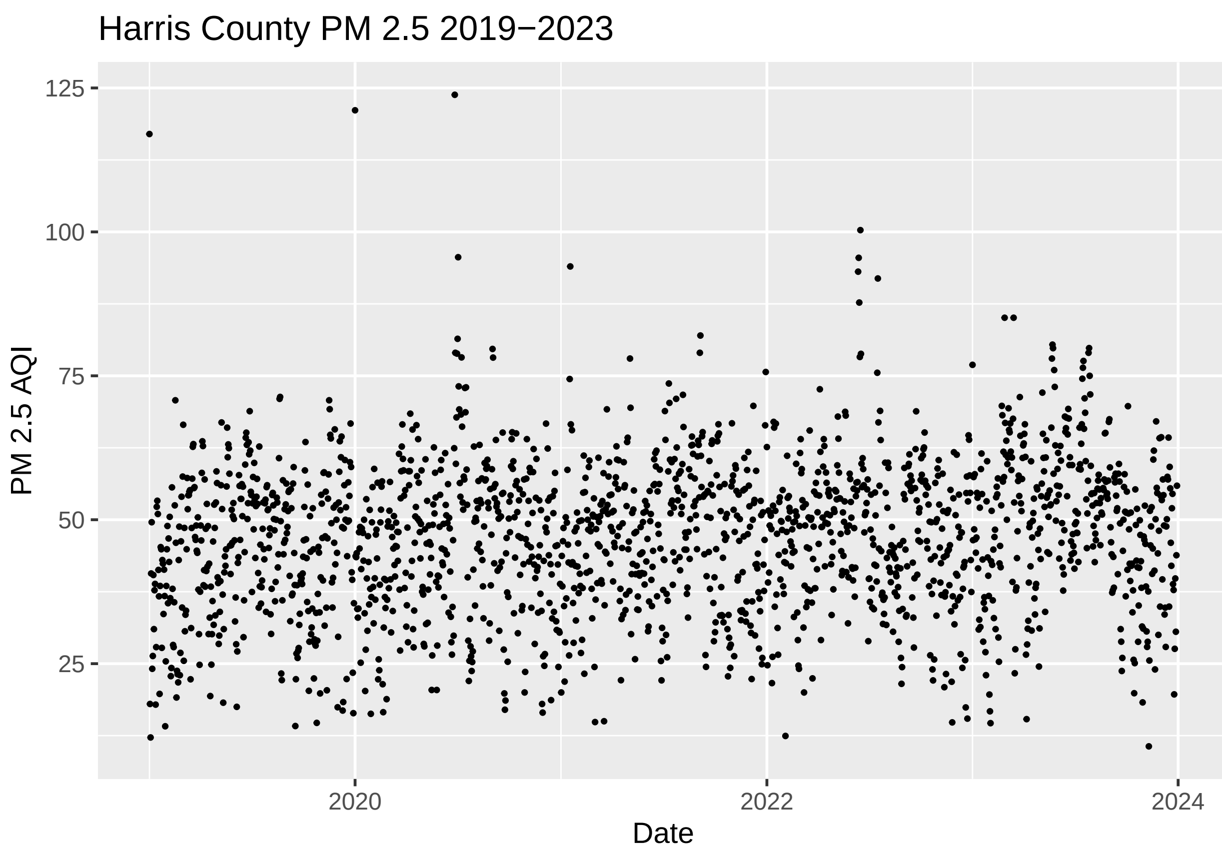

```
Ozoneplot<-HarrisCoRain_AQ%>%
  ggplot(aes(
    x=Date,
    y=Avg_AQI_Ozone))+
  labs(
    x="Date",
    y="Ozone AQI")+
  geom_point(size=0.5)+
  ggtitle('Harris County Ozone 2019-2023')
print(Ozoneplot)
```




```r
Ozoneplot2<-DallasCoRain_AQ%>%
  ggplot(aes(
    x=Date,
    y=Avg_AQI_Ozone))+
  labs(
    x="Date",
    y="Ozone AQI")+
  geom_point(size=0.5)+
  ggtitle('Dallas County Ozone 2019-2023')
print(Ozoneplot2)
```

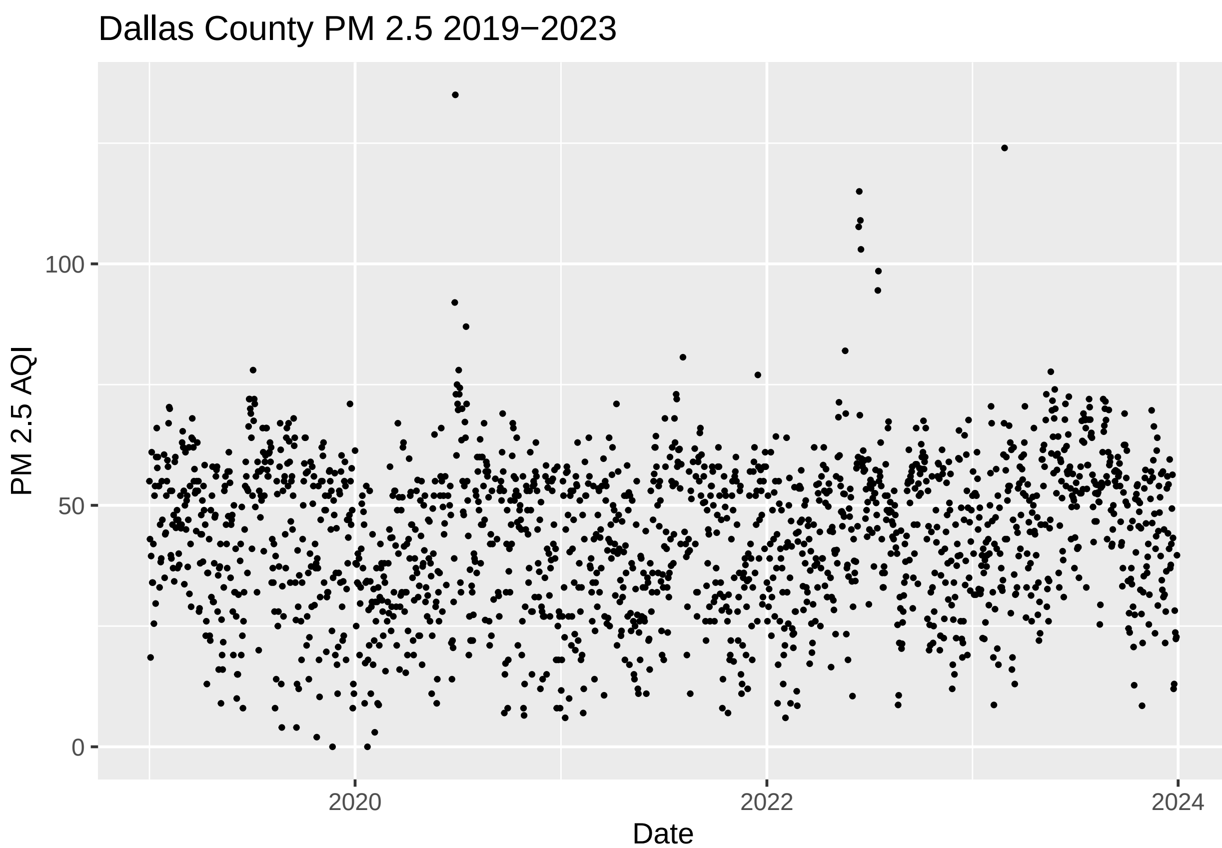



```r
PMplot<-HarrisCoRain_AQ%>%
  ggplot(aes(
    x=Date,
    y=Avg_AQI_PM))+
  labs(
    x="Date",
    y="PM 2.5 AQI")+
  geom_point(size=0.5)+
  ggtitle('Harris County PM 2.5 2019-2023')
print(PMplot)
```




```r
PMplot2<-DallasCoRain_AQ%>%
  ggplot(aes(
    x=Date,
    y=Avg_AQI_PM))+
  labs(
    x="Date",
    y="PM 2.5 AQI")+
  geom_point(size=0.5)+
  ggtitle('Dallas County PM 2.5 2019-2023')
print(PMplot2)
```

```
## Warning: Removed 40 rows containing missing values or values outside the scale range
## (`geom_point()`).
```


Dallas County PM 2.5 2019−2023
PM 2.5 AQI
100
50
0
2020
2022
2024
Date

Summary and Conclusions

The analysis of correlation between air quality and emissions at the county level per year did not result in significant results. There is not a significant correlation between annual county emission from power plants and average annual county AQI score from Ozone or PM2.5. Additionally, no county on average saw any poor or unsafe Air Quality Index levels. There are three potential factors influencing this lack of correlation. One is that air quality and emissions cannot be studied at localized scale. Likely emissions disperse out of the county it is in based on local weather patterns. Conducting analysis of local air quality based on power plant emissions might be more accurate if wind patterns were taken into consideration to estimate where emissions may be concentrating. Another reason is that topography plays a role in where emissions and particulates accumulate. For example, pollution commonly gets trapped in valleys. Therefore, including topography into the analysis might elucidate more clear patterns. Lastly, there may not be a correlation is because there are not enough data points in this analysis to see a trend. All plants do not report emissions and there are not air quality stations in all counties. To perform this analysis more accurately, data should be looked at country wide and there should be better standards for ensuring all active power plants report their emissions. Ground level ozone is more likely to form in high heat and sunlight, when nitrogen oxides react with VOCs. This report mainly focused on total CO_2 equivalent emissions. The emissions that are summed include SO_2, CO_2, CH_4, and N_2O. Since all but methane are not VOCs, we thought it would be interesting to explore the relationship specifically between methane and ozone over time in one isolated location with abundant emissions and air quality data. As shown in Figure 6 and Figure 8, analysis of ozone AQI showed an increase from 2019-2023 in Harris county, Texas, while methane emissions decreased in that same time period. Our results did not indicate a positive relationship between methane and ozone.

References

<add references here if relevant, otherwise delete this section>