Analyzing Seasonal Trends and Weather Impacts on AQI in Stanislaus, CA

Ellie Harrigan, Rosalind Hu, Hanbin Lyu, Lauren Ng

December 12, 2024

Contents

List of Figures

#Load Datasets for Air Quality and Weather

```
# Set your working directory
getwd()
```

```
## [1] "/home/guest/Team_Project/EDE Final Project"
```

```
# Load your packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
##
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
##
## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(lubridate)
library(trend)
library(here)
```

```
## here() starts at /home/guest/Team_Project/EDE Final Project
##
## Attaching package: 'here'
##
## The following object is masked from 'package:plyr':
##
##     here
```

```r
#install.packages("openmeteo")
library(openmeteo)

#Load air datasets
Stanislaus_Air_Files <- list.files(path = "./Data/Raw/Stanislaus_EPAair/",
                                   pattern = "*.csv",
                                   full.names = TRUE)
#Combine air datasets
Stanislaus_Air_Data <- Stanislaus_Air_Files %>%
  plyr::ldply(read.csv)


#Load weather data
weather_data <- weather_history(
 location = c(37.6393,120.9970), #Coordinates for Modesto, CA
 start = "2013-01-01",
 end = "2023-12-31",
 hourly = "temperature_2m"
)
```

```r
#Ensure plyr is detached to avoid conflicts
if ("package:plyr" %in% search()) {
  detach("package:plyr", unload = TRUE)
}

#Ensure datetime is in proper format
weather_data <- weather_data %>%
  mutate(datetime = as.POSIXct(datetime, format = "%Y-%m-%d %H:%M:%S"))

#Combine rows by month
weather_data_processed <- weather_data %>%
  mutate(
    year_month = format(as.Date(datetime), "%Y-%m"),
  ) %>%
  group_by(year_month) %>%
  summarise(
    avg_temperature = mean(hourly_temperature_2m, na.rm = TRUE),
    max_temperature = max(hourly_temperature_2m, na.rm = TRUE),
    min_temperature = min(hourly_temperature_2m, na.rm = TRUE),
    total_records = n()
  )
```

Rationale and Research Questions

The primary focus is to identify patterns or correlations between weather conditions and changes in air quality metrics, such as AQI (Air Quality Index) and PM2.5 concentrations. The goal of this project is to understand how specific weather variables, such as precipitation and temperature, impact air quality. This research will provide insights into seasonal and weather-related factors contributing to air pollution, offering a some insight for future mitigation strategies, public health advisories, or policy recommendations.

```r
# Process the data: Combine rows by month
Stanislaus_Air_Data_Processed <- Stanislaus_Air_Data %>%
  mutate(
    Date = as.Date(Date, format = "%m/%d/%Y"),
    year_month = format(Date, "%Y-%m")
  ) %>%
  group_by(year_month) %>%
  summarise(
    avg_PM25 = mean(Daily.Mean.PM2.5.Concentration, na.rm = TRUE),
    max_PM25 = max(Daily.Mean.PM2.5.Concentration, na.rm = TRUE),
    min_PM25 = min(Daily.Mean.PM2.5.Concentration, na.rm = TRUE),
    total_records = n()
  )


Stanislaus_Air_Data_Processed
```

```
## # A tibble: 132 x 5
##    year_month avg_PM25 max_PM25 min_PM25 total_records
##    <chr>         <dbl>    <dbl>    <dbl>         <int>
##  1 2013-01       27.0      57.7      1.8            67
##  2 2013-02       16.9      41.3      3.1            60
##  3 2013-03        8.94     24.7      1.9            67
##  4 2013-04        5.74     12.6     -0.5            64
##  5 2013-05        8.50     24.5      2.1            70
##  6 2013-06        6.44     12.1      2              65
##  7 2013-07       10.6      33.4      5              69
##  8 2013-08        7.22     20.4      0.2            65
##  9 2013-09        5.73     10.9     -1.2            63
## 10 2013-10       12.3      21.9      4.2            65
## # i 122 more rows
```

Dataset Information

We downloaded two datasets, one from EPA Stanislaus_Air_Data to present the air quality(include two values we want to examine PM2.5 and AQI values), another is weather data from NOAA, we installed "openmeteo" package to process data. Since two datasets have different format of dates, we first use mutate function and group _by month to combine the rows for ensuring the consistency of "yyyy%-mm%" format. However, each month has different numbers of observations, we then combine all the rows happen in the same month so that we can easier to analyze the data.

Exploratory Analysis

Are there any noticeable trends in temperature or air pollution?

```
#Explore data set
class(weather_data_processed$year_month)
```

```
## [1] "character"
```

```
#Create new column that is a date
weather_data_processed <- weather_data_processed %>%
  mutate(month_year_date = as.Date(paste0(year_month, "-01")))

Stanislaus_Air_Data_Processed<- Stanislaus_Air_Data_Processed %>%
  mutate(month_year_date = as.Date(paste0(year_month, "-01")))

#Verify
str(weather_data_processed)
```

```
## tibble [133 x 6] (S3: tbl_df/tbl/data.frame)
##  $ year_month     : chr [1:133] "2012-12" "2013-01" "2013-02" "2013-03" ...
##  $ avg_temperature: num [1:133] -4.34 -2.68 -1.43 3.9 8.96 ...
##  $ max_temperature: num [1:133] -4 7.9 10.4 21.4 24.5 29.7 29.6 32.3 34.2 29.8 ...
##  $ min_temperature: num [1:133] -5.1 -12.6 -11.7 -4.7 1 6.6 15.4 19.7 16.8 14 ...
##  $ total_records  : int [1:133] 8 744 672 744 720 744 720 744 744 720 ...
##  $ month_year_date: Date[1:133], format: "2012-12-01" "2013-01-01" ...
```

```
str(Stanislaus_Air_Data_Processed)
```

```
## tibble [132 x 6] (S3: tbl_df/tbl/data.frame)
##  $ year_month     : chr [1:132] "2013-01" "2013-02" "2013-03" "2013-04" ...
##  $ avg_PM25       : num [1:132] 27 16.85 8.94 5.74 8.5 ...
##  $ max_PM25       : num [1:132] 57.7 41.3 24.7 12.6 24.5 12.1 33.4 20.4 10.9 21.9 ...
##  $ min_PM25       : num [1:132] 1.8 3.1 1.9 -0.5 2.1 2 5 0.2 -1.2 4.2 ...
##  $ total_records  : int [1:132] 67 60 67 64 70 65 69 65 63 65 ...
##  $ month_year_date: Date[1:132], format: "2013-01-01" "2013-02-01" ...
```

```
#Create a joined data frame with air and temperature data
combined_air_weather <- left_join(weather_data_processed, Stanislaus_Air_Data_Processed,
                                  by = "month_year_date")%>%
  slice(-1) #remove first column that contained NA values for Dec 2012

combined_data_long <- combined_air_weather %>%
  pivot_longer(cols = c(avg_temperature, avg_PM25), names_to = "variable",
               values_to = "value")

#Create a faceted plot of average temperature and average PM2.5
ggplot(combined_data_long, aes(x = month_year_date, y = value,
                               color = variable)) +
  geom_line(size = 1) +
```

```
  scale_x_date(date_breaks = "1 year", date_labels = "%Y")+  # Adjust interval
  facet_wrap(~ variable, scales = "free_y", ncol = 1,
             labeller = labeller(variable = c
                                 ("avg_temperature" = "Average Temperature (°C)",
                                   "avg_pm25" = "Average PM2.5 (µg/m³)"))) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    legend.position = "none"
  ) +
  labs(
    title = "Monthly Trends of Temperature and PM2.5")
```

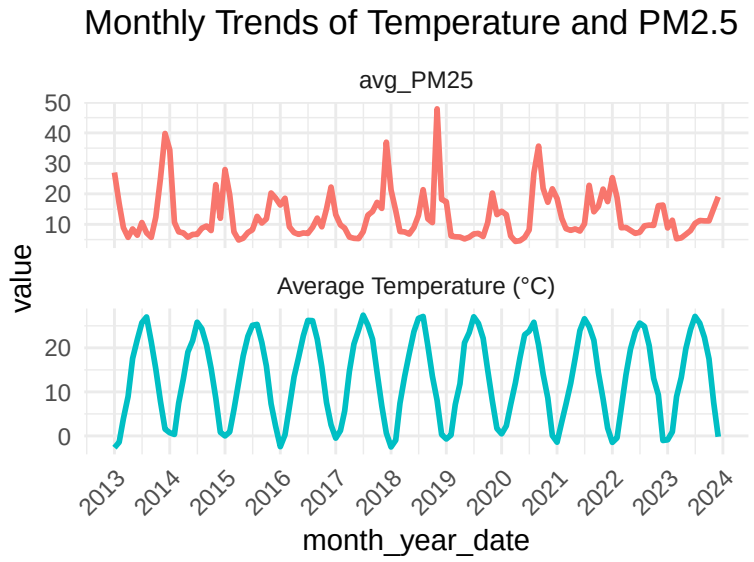

Figure 1: Monthly Trends of Temperature and PM2.5

Looking at this graph, it appears that in the last decade, temperatures in Stanislaus County have followed a regular seasonal pattern. It does not show any type of trend year over year. On the other hand, PM2.5 shows a large spike between 2018-2019 and some other smaller spikes which appear to occur at somewhat regular intervals.

How do PM2.5 concentrations vary over the course of a year?

```
#Create a new column with month and year
Stanislaus_Air_Data_Processed <- Stanislaus_Air_Data_Processed %>%
  mutate(
    year = format(as.Date(month_year_date), "%Y"),
    month = format(as.Date(month_year_date), "%m")
  )

#Plot PM2.5 levels by year
```

```r
ggplot(Stanislaus_Air_Data_Processed,
       aes(x = as.numeric(month),
           y = avg_PM25,
           color = year,
           group = year)) +
  geom_line() +
  scale_x_continuous(breaks = 1:12, labels = month.abb) +  #Show months as x-axis labels
  labs(
    title = "Monthly PM2.5 Concentrations",
    x = "Month",
    y = "PM2.5 Concentrations (µg/m³)"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    legend.title = element_blank()
  )
```

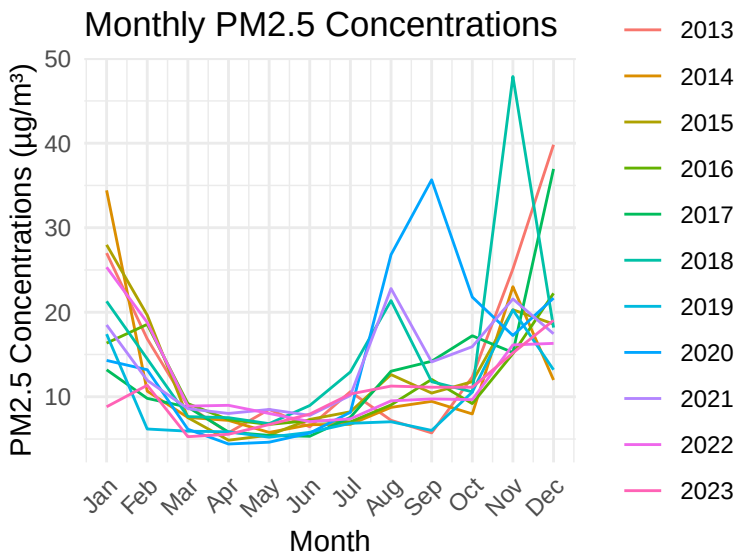

Figure 2: Monthly PM2.5 Concentrations 2013-2023

Summary: It looks like from 2013-2023, there is generally an increase in PM2.5 from July onwards, then it dips back down during the winter months.

It is possible that there could be a seasonal component to particulate matter. We know that anthropogenic sources of PM 2.5 include combustion from motor vehicles, smelters, power plants, industrial facilities, residential fireplaces/wood stoves, agricultural burning, and forest fires. Thus, it is possible that in California, there could be higher PM 2.5 in the summer, when the risk of forest fires is higher and power plants are burning more fossil fuels to meet peak demand for air conditioning during hot summer months.

Time Series Analysis

```
#install.packages("trend")
library(trend)
#install.packages("zoo")
library(zoo)
```

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

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

```
#install.packages("Kendall")
library(Kendall)
#install.packages("tseries")
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

Do concentrations of average PM2.5 correlate over time from 2013-2023?

```
#Generate time series
f_month <- month(first(Stanislaus_Air_Data_Processed$month_year_date))
f_year <- year(first(Stanislaus_Air_Data_Processed$month_year_date))

Stanislaus_Air_Data_ts <- ts(Stanislaus_Air_Data_Processed$avg_PM25,
                             start = c(f_year, f_month),
                             frequency = 12)

#Generate the Decomposition
Stanislaus_Air_Data_Decomposed <- stl(Stanislaus_Air_Data_ts,
                                      s.window = "periodic")

#Visualize how the trend maps onto the data
Stanislaus_Air_Data_components <- as.data.frame(
  Stanislaus_Air_Data_Decomposed$time.series)

#Visualize decomposed series
plot(Stanislaus_Air_Data_Decomposed)
```

```
#Run Seasonal Mann Kendall Test
stanislaus_air_data_trend <-
  Kendall::SeasonalMannKendall(Stanislaus_Air_Data_ts)

#Inspect Results
stanislaus_air_data_trend
```

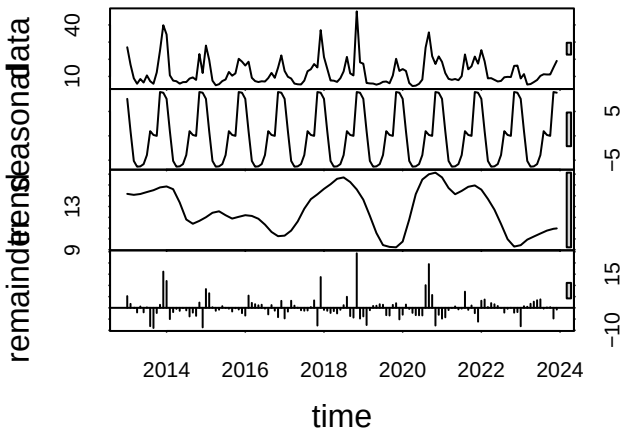

Figure 3: Decomposed Air Quality Data

```
## tau = -0.0182, 2-sided pvalue =0.78741
```

```
summary(stanislaus_air_data_trend)
```

```
## Score =  -12 , Var(Score) = 1980
## denominator =  660
## tau = -0.0182, 2-sided pvalue =0.78741
```

```
#Visualization
stanislaus_data_plot <-
  ggplot(Stanislaus_Air_Data_Processed, aes( x= month_year_date,
                                             y= avg_PM25)) +
  geom_point() +
  geom_line() +
  ylab("Avg Pm2.5") +
  xlab("2013-2023") +
  geom_smooth(method = lm)
print(stanislaus_data_plot)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Kendall's Tau and P-Value The Kendall's tau coefficient was calculated as -0.0182, which is very close to zero. This means there's little to no consistent pattern in how the average PM 2.5 changes over time. There is no clear tendency for one to increase or decrease as the other changes.

The p-value was 0.78741, which is much higher than the typical threshold of 0.05. This means we didn't find any statistically significant relationship between the avergae PM 2.5 over time To summarize, the analysis shows that the two variables don't have a strong or consistent relationship in this dataset.

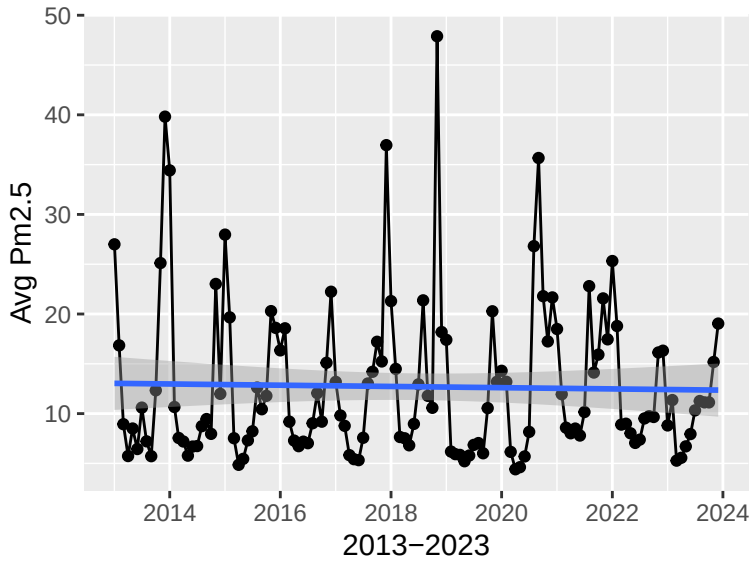

Figure 4: Average PM2.5 Concentrations 2013-2023

General Linear Model Analysis

What is the relationship between Average PM2.5 and Average Temperature

```
#Test PM2.5 and Temperature in Regression Analysis
cor.test(combined_air_weather$avg_PM25, combined_air_weather$avg_temperature,
         method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  combined_air_weather$avg_PM25 and combined_air_weather$avg_temperature
## t = -5.2575, df = 130, p-value = 5.829e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5502433 -0.2669713
## sample estimates:
##        cor
## -0.4187415
```

```
model <- lm(avg_PM25 ~ avg_temperature, data = combined_air_weather)
summary(model)
```

```
##
## Call:
## lm(formula = avg_PM25 ~ avg_temperature, data = combined_air_weather)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -10.856  -4.267  -1.161   2.174  33.487
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     17.10966    1.04043  16.445  < 2e-16 ***
## avg_temperature -0.33512    0.06374  -5.258 5.83e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.071 on 130 degrees of freedom
## Multiple R-squared:  0.1753, Adjusted R-squared:  0.169
## F-statistic: 27.64 on 1 and 130 DF,  p-value: 5.829e-07
```

```
#Visualize
stanislaus_data_plot1 <-
  ggplot(combined_air_weather, aes( x= avg_PM25,
                                    y= avg_temperature)) +
  geom_point() +
  geom_line() +
  ylab("Avg Pm2.5") +
  xlab("Temperature") +
  geom_smooth(method = lm, color = "blue")
print(stanislaus_data_plot1)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

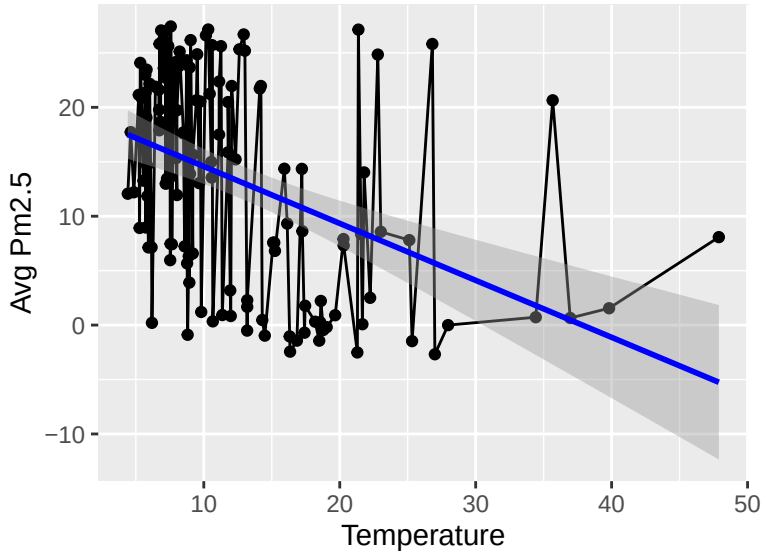

Figure 5: Average PM2.5 and Temperature 2013-2023

Results of the Regression Analysis Performed an analysis to see if there's a relationship between average temperature and average PM2.5 levels in Stanislaus County.

Relationship Between Temperature and PM2.5: There is a negative relationship between temperature and PM2.5. This means that as the temperature increases, the PM2.5 levels tend to

decrease. However, the effect is small, for each 1°C increase in temperature, PM2.5 drops by about 0.34 µg/m³.

Model Performance: The model does a fair job of showing the relationship, but it only explains about 17% of the variation in PM2.5 levels. This means that temperature is just one factor influencing PM2.5, and there are likely other important factors (like wind or pollution sources) that play a bigger role.

Significance: The relationship we found between temperature and PM2.5 is statistically significant, meaning it's unlikely to be due to random chance and that temperature does effect on PM2.5. However, temperature only explains a small part of the variation in air quality. Other factors like motor vehicles, smelters, power plants, industrial facilities, residential fireplaces/wood stoves, agricultural burning, and forest fires.

Analysis of Variance

Using a One-Way Anova to test the correlation between temperature and PM2.5.

```
#load package
library(agricolae)

#Get the quantile of the average temperatures
quantile = summary(combined_air_weather$avg_temperature)
quantile
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -2.681   3.723  13.727  13.161  22.014  27.429
```

```
#Divide temperatures into four categories very low, low, median, high based on quantile
combined_air_weather_groups = combined_air_weather %>%
  mutate(
    temp_category = cut(
      avg_temperature,
      breaks = c(-2.682, 3.723, 13.727, 22.014, 27.429),
      labels = c("Very Low", "Low", "Medium", "High")
      ))

#Test PM2.5 and Temperature in one-way anova
anova = aov(data = combined_air_weather_groups, avg_PM25 ~ temp_category)
summary(anova)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## temp_category   3   1586   528.6   10.75 2.39e-06 ***
## Residuals     128   6296    49.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Post-hoc test: Tukey HSD
tukey = HSD.test(anova, "temp_category", group = TRUE)
tukey
```

```
## $statistics
##    MSerror  Df     Mean       CV      MSD
##   49.18898 128 12.69926 55.22752 4.494519
##
## $parameters
##    test        name.t ntr StudentizedRange alpha
##   Tukey temp_category   4         3.681343  0.05
##
## $means
##           avg_PM25      std  r       se      Min      Max       Q25       Q50
## High      9.868914 4.934042 33 1.220891 5.326866 26.81385  7.058462  8.167692
## Low      12.068622 8.704286 33 1.220891 4.411290 47.88939  7.201538  8.773134
## Medium   10.327638 5.996313 33 1.220891 4.622581 35.66721  6.693750  8.986957
## Very Low 18.531861 7.790430 33 1.220891 6.185000 39.81642 13.193750 17.455224
##               Q75
## High     10.60290
## Low      15.23529
## Medium   11.78358
## Very Low 21.29571
##
## $comparison
## NULL
##
## $groups
##           avg_PM25 groups
## Very Low 18.531861      a
## Low      12.068622      b
## Medium   10.327638      b
## High      9.868914      b
##
## attr(,"class")
## [1] "group"
```

Summary Based on One-way Anova and Post-hoc Test: The analysis suggests a significant relationship between temperature categories and PM2.5 concentrations. PM2.5 levels are highest in the Very Low Temperature category and decrease as temperatures increase, with the High Temperature category having the lowest PM2.5 levels. This indicates that lower temperatures are associated with higher PM2.5 levels, likely due to reduced atmospheric dispersion and increased emissions from heating sources. However, no significant differences were observed among the Low, Medium, and High temperature categories. Further research is recommended to explore additional factors influencing PM2.5 levels, such as wind speed, humidity, and pollution sources.

Saving Processed Datasets

```
write.csv(
  Stanislaus_Air_Data_Processed,
  file = here("./Data/Processed/Stanislaus_Air_Data_Processed"),
  row.names = FALSE)

write.csv(
  weather_data_processed,
  file = here("./Data/Processed/Weather_Data_Processed"),
  row.names = FALSE)
```

GitHub Link

https://github.com/Rosalind1218/EDE-Team-Final-Project