AirQuality

2024-12-9

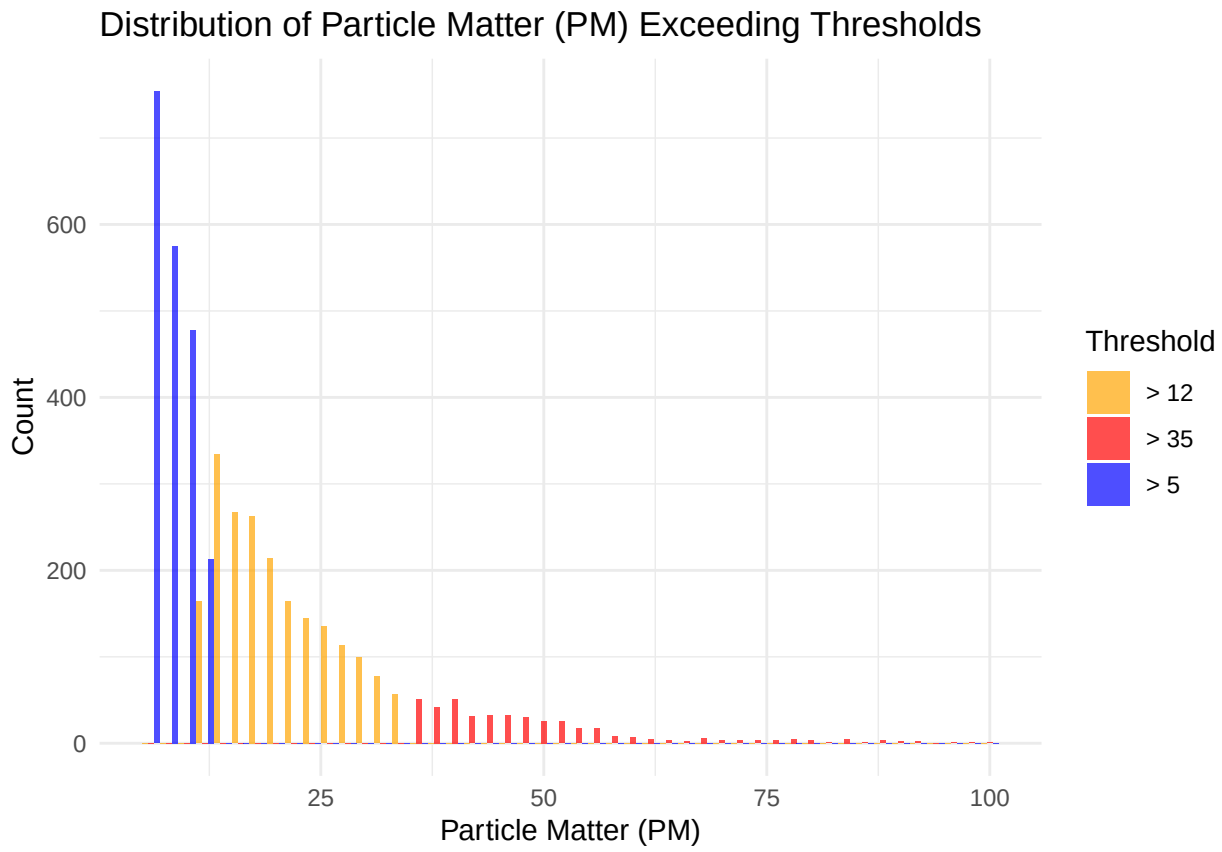
```
air_quality <- read.csv("/Users/nikkifuller/Library/CloudStorage/OneDrive-WashingtonStateUniversity(emai

demographics <- read.csv("/Users/nikkifuller/Library/CloudStorage/OneDrive-WashingtonStateUniversity(ema

geography <- read.csv("/Users/nikkifuller/Library/CloudStorage/OneDrive-WashingtonStateUniversity(email
```

```
air_quality_longer <- air_quality %>%
  pivot_longer(
    cols = starts_with("ID_"),
    names_to = "ID",
    names_prefix = "ID_",
    values_to = "Particle_Matter"
  ) |>
  mutate(
    ID = as.numeric(ID)
  )

full_aq_stats <- left_join(air_quality_longer, demographics, by = "ID")
full_aq_stats <- left_join(full_aq_stats, geography, by = "ID")

full_aq_stats <- full_aq_stats %>%
  mutate(PM_greater_than_5 = Particle_Matter > 5) %>%
  mutate(PM_greater_than_12 = Particle_Matter > 12) %>%
  mutate(PM_greater_than_35 = Particle_Matter > 35)

str(full_aq_stats)
```

```
## tibble [8,500 x 21] (S3: tbl_df/tbl/data.frame)
##  $ Time..UTC.00.00.   : chr [1:8500] "2019-09-01 00:00:00+00:00" "2019-09-01 00:00:00+00:00" "2019-09
##  $ ID                 : num [1:8500] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Particle_Matter    : num [1:8500] 1.55 4.76 3.73 3.62 2.81 2.54 4.4 3.02 3.94 4.78 ...
##  $ NativeAmerican     : num [1:8500] 0.02 0.01 0 0.01 0.02 0 0.01 0.02 0.01 0.02 ...
##  $ Asian              : num [1:8500] 0.03 0.01 0.01 0 0.03 0 0.03 0.01 0.03 0.03 ...
##  $ Black              : num [1:8500] 0.04 0.03 0.01 0.01 0.11 0.02 0.05 0.03 0.04 0.05 ...
##  $ Hispanic           : num [1:8500] 0.09 0.13 0.09 0.12 0.16 0.11 0.17 0.11 0.13 0.16 ...
##  $ Multiracial        : num [1:8500] 0.16 0.15 0.11 0.13 0.18 0.11 0.19 0.15 0.17 0.19 ...
##  $ PacificIslander    : num [1:8500] 0 0.05 0 0 0.07 0 0.05 0.05 0.05 0.08 ...
##  $ unknown            : int [1:8500] 0 0 0 0 0 0 0 0 0 0 ...
##  $ White              : num [1:8500] 0.64 0.61 0.79 0.72 0.44 0.75 0.5 0.63 0.57 0.48 ...
##  $ LowIncome          : num [1:8500] 0.7 0.84 0.45 0.62 0.83 0.38 0.91 0.85 0.84 0.89 ...
##  $ layout_type        : int [1:8500] 0 1 0 0 1 0 0 1 1 1 ...
##  $ Distance_to_freeway: num [1:8500] 3665 2328 272 1241 561 ...
##  $ Street_100         : num [1:8500] 1013 687 349 438 406 ...
##  $ Street_200         : num [1:8500] 2724 3006 1394 1912 1882 ...
```

```
##  $ Street_300m        : num [1:8500] 5550 6682 4108 4566 4244 ...
##  $ PercentTree        : num [1:8500] 31 69.6 84.4 84.1 73.8 89.1 47.2 38.3 32.2 51.5 ...
##  $ PM_greater_than_5  : logi [1:8500] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ PM_greater_than_12 : logi [1:8500] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ PM_greater_than_35 : logi [1:8500] FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```r
### histogram
pm_filtered_threshold <- full_aq_stats %>%
  filter(Particle_Matter < 100) %>%
  filter(PM_greater_than_5 | PM_greater_than_12 | PM_greater_than_35) %>%
  mutate(Threshold = case_when(PM_greater_than_35 ~ "> 35", PM_greater_than_12 ~ "> 12", PM_greater_than

ggplot(pm_filtered_threshold, aes(x = Particle_Matter, fill = Threshold)) +
  geom_histogram(binwidth = 2, position = "dodge", alpha = 0.7) +
  scale_fill_manual(values = c("> 5" = "blue", "> 12" = "orange", "> 35" = "red")) +
  labs(title = "Distribution of Particle Matter (PM) Exceeding Thresholds", x = "Particle Matter (PM)",
  theme_minimal()
```




```r
aq_5_lm <- lm(Particle_Matter ~ PM_greater_than_5, data = full_aq_stats)
summary(aq_5_lm)
```

```
##
## Call:
## lm(formula = Particle_Matter ~ PM_greater_than_5, data = full_aq_stats)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.89  -9.34  -1.20   1.14 542.23
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.1479     0.4206   5.107 3.35e-07 ***
## PM_greater_than_5TRUE  19.7482     0.5700  34.644  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.05 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.1248, Adjusted R-squared:  0.1247
## F-statistic:  1200 on 1 and 8418 DF,  p-value: < 2.2e-16
```

```
aq_12_lm <- lm(Particle_Matter ~ PM_greater_than_12, data = full_aq_stats)
summary(aq_12_lm)
```

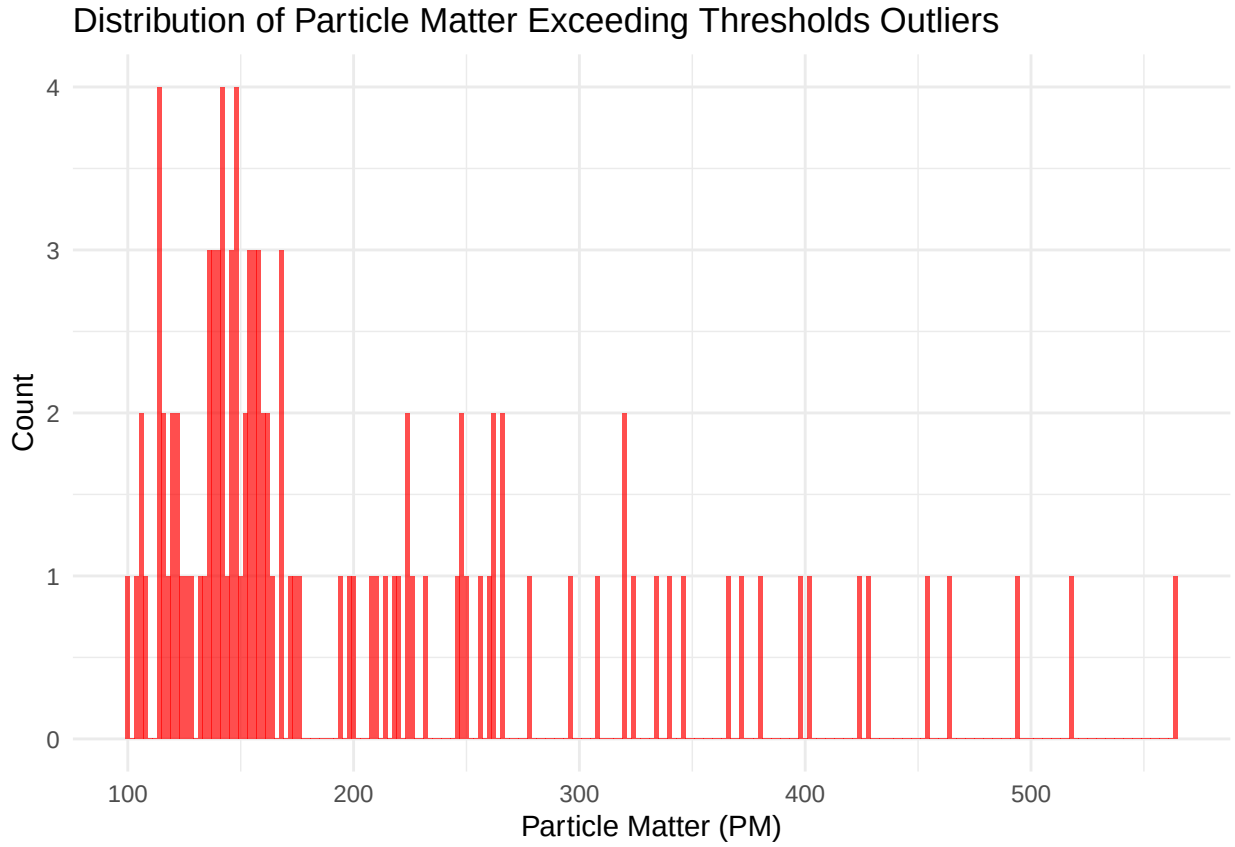
```
##
## Call:
## lm(formula = Particle_Matter ~ PM_greater_than_12, data = full_aq_stats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20.79  -3.86  -1.71   1.94 531.33
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.1839     0.3206   13.05   <2e-16 ***
## PM_greater_than_12TRUE  28.6207     0.5809   49.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.53 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.2238, Adjusted R-squared:  0.2237
## F-statistic:  2427 on 1 and 8418 DF,  p-value: < 2.2e-16
```

```
aq_35_lm <- lm(Particle_Matter ~ PM_greater_than_35, data = full_aq_stats)
summary(aq_35_lm)
```

```
##
## Call:
## lm(formula = Particle_Matter ~ PM_greater_than_35, data = full_aq_stats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45.89  -6.61  -3.37   4.03 483.23
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              8.3128     0.2424   34.30   <2e-16 ***
```

```
## PM_greater_than_35TRUE  72.5900     0.9642   75.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.53 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.4024, Adjusted R-squared:  0.4023
## F-statistic:  5667 on 1 and 8418 DF,  p-value: < 2.2e-16
```

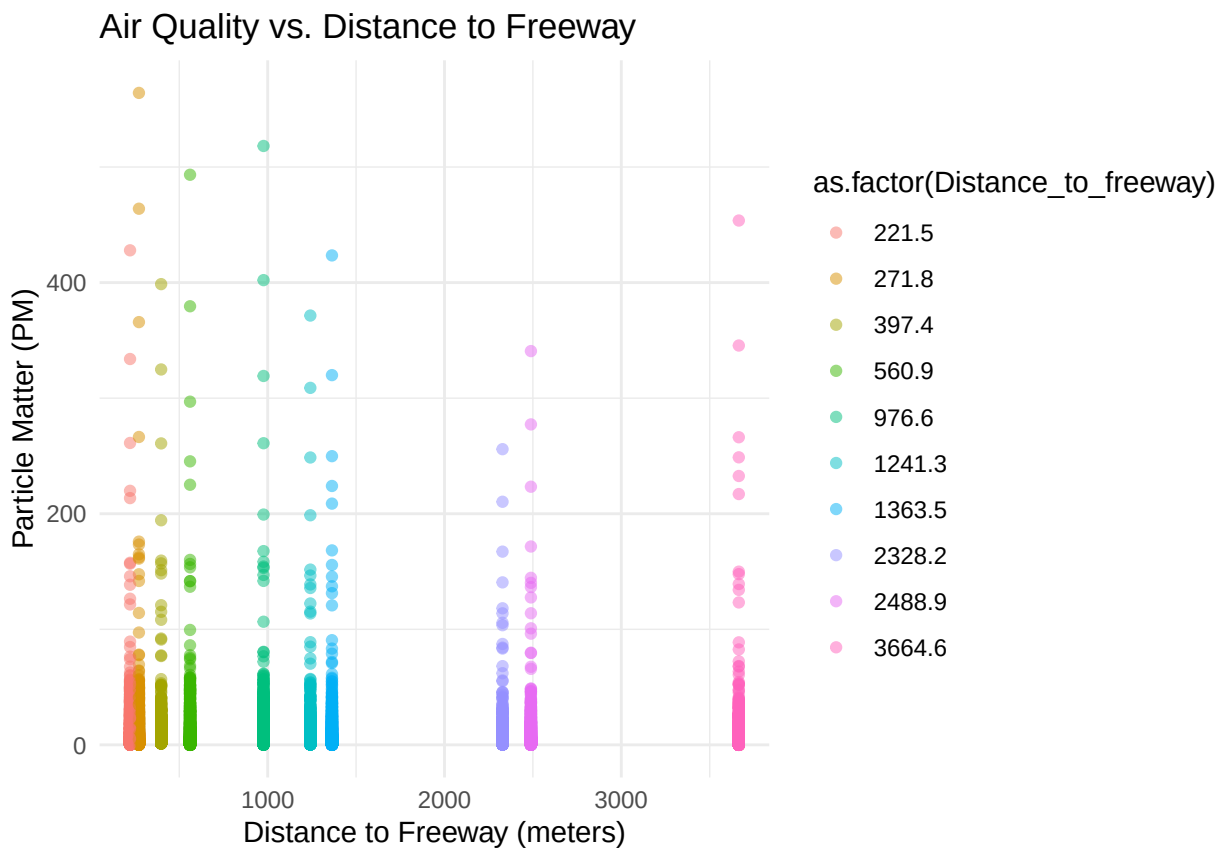
```r
### outliers
pm_filtered_outliers <- full_aq_stats %>%
  filter(Particle_Matter >= 100)

ggplot(pm_filtered_outliers, aes(x = Particle_Matter)) +
  geom_histogram(binwidth = 2, position = "dodge", alpha = 0.7, fill = "red") +
  labs(title = "Distribution of Particle Matter Exceeding Thresholds Outliers",
       x = "Particle Matter (PM)",
       y = "Count") +
  theme_minimal()
```




```r
ggplot(full_aq_stats, aes(x = Distance_to_freeway, y = Particle_Matter, color = as.factor(Distance_to_fi
  geom_point(alpha = 0.5) +
  labs(title = "Air Quality vs. Distance to Freeway", x = "Distance to Freeway (meters)", y = "Particle
  theme_minimal()
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
```

```
## (‘geom_point()‘).
```

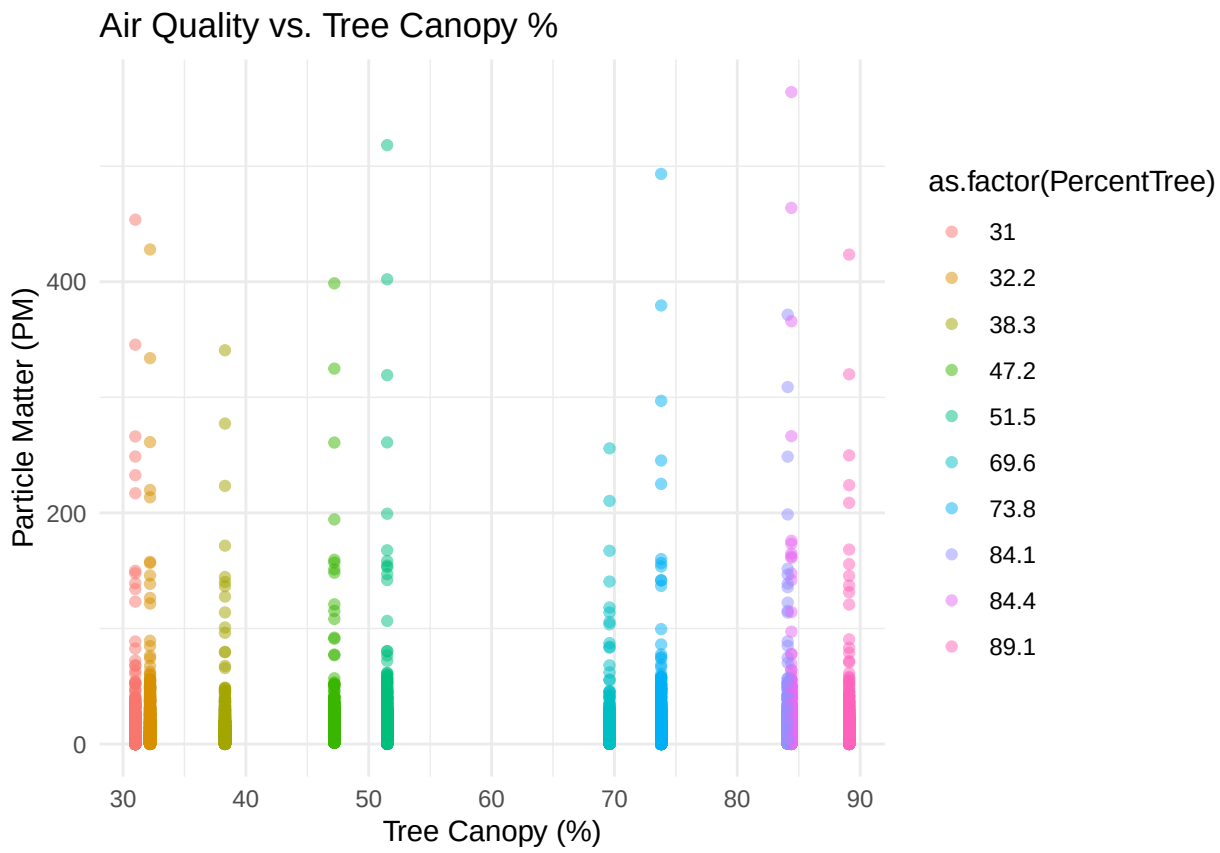



```r
aq_freeway_lm <- lm(Particle_Matter ~ Distance_to_freeway, data = full_aq_stats)
summary(aq_freeway_lm)
```

```
##
## Call:
## lm(formula = Particle_Matter ~ Distance_to_freeway, data = full_aq_stats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.62 -10.43  -6.85   1.86 549.58
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          14.9617167  0.4833746  30.953  < 2e-16 ***
## Distance_to_freeway -0.0015249  0.0002785  -5.475 4.49e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 27.8 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.003549,   Adjusted R-squared:  0.00343
## F-statistic: 29.98 on 1 and 8418 DF,  p-value: 4.494e-08
```

```
ggplot(full_aq_stats, aes(x = PercentTree, y = Particle_Matter, color = as.factor(PercentTree))) +
  geom_point(alpha = 0.5) +
  labs(title = "Air Quality vs. Tree Canopy %", x = "Tree Canopy (%)", y = "Particle Matter (PM)") +
  theme_minimal()
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



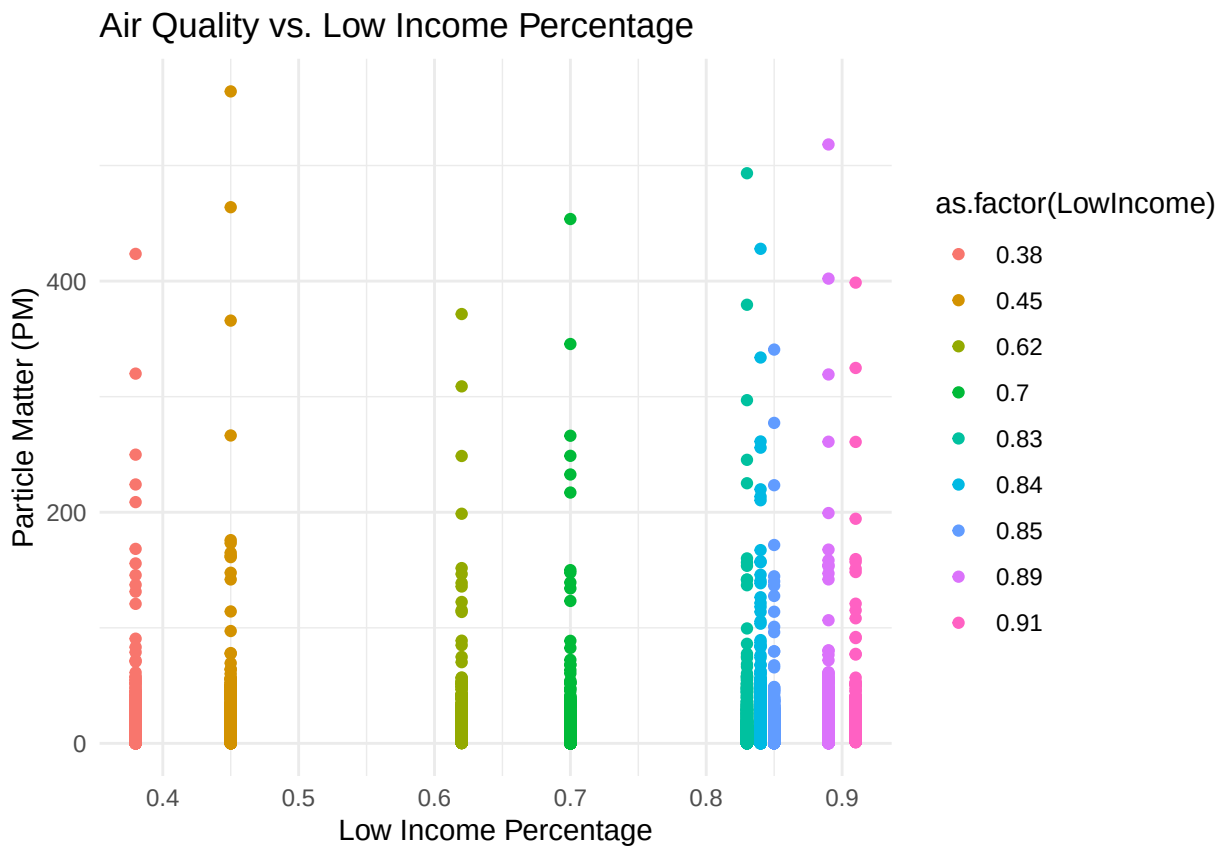

```
aq_tree_lm <- lm(Particle_Matter ~ PercentTree, data = full_aq_stats)
summary(aq_tree_lm)
```

```
##
## Call:
## lm(formula = Particle_Matter ~ PercentTree, data = full_aq_stats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.91 -10.70  -6.99   1.82 551.24
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.9298248  0.8987943  14.386   <2e-16 ***
## PercentTree -0.0005106  0.0141231  -0.036    0.971
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.84 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  1.553e-07,  Adjusted R-squared:  -0.0001186
## F-statistic: 0.001307 on 1 and 8418 DF,  p-value: 0.9712
```

```
ggplot(full_aq_stats, aes(x = LowIncome, y = Particle_Matter, color = as.factor(LowIncome))) +
  geom_point() +
  labs(title = "Air Quality vs. Low Income Percentage", x = "Low Income Percentage", y = "Particle Matte
  theme_minimal()
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



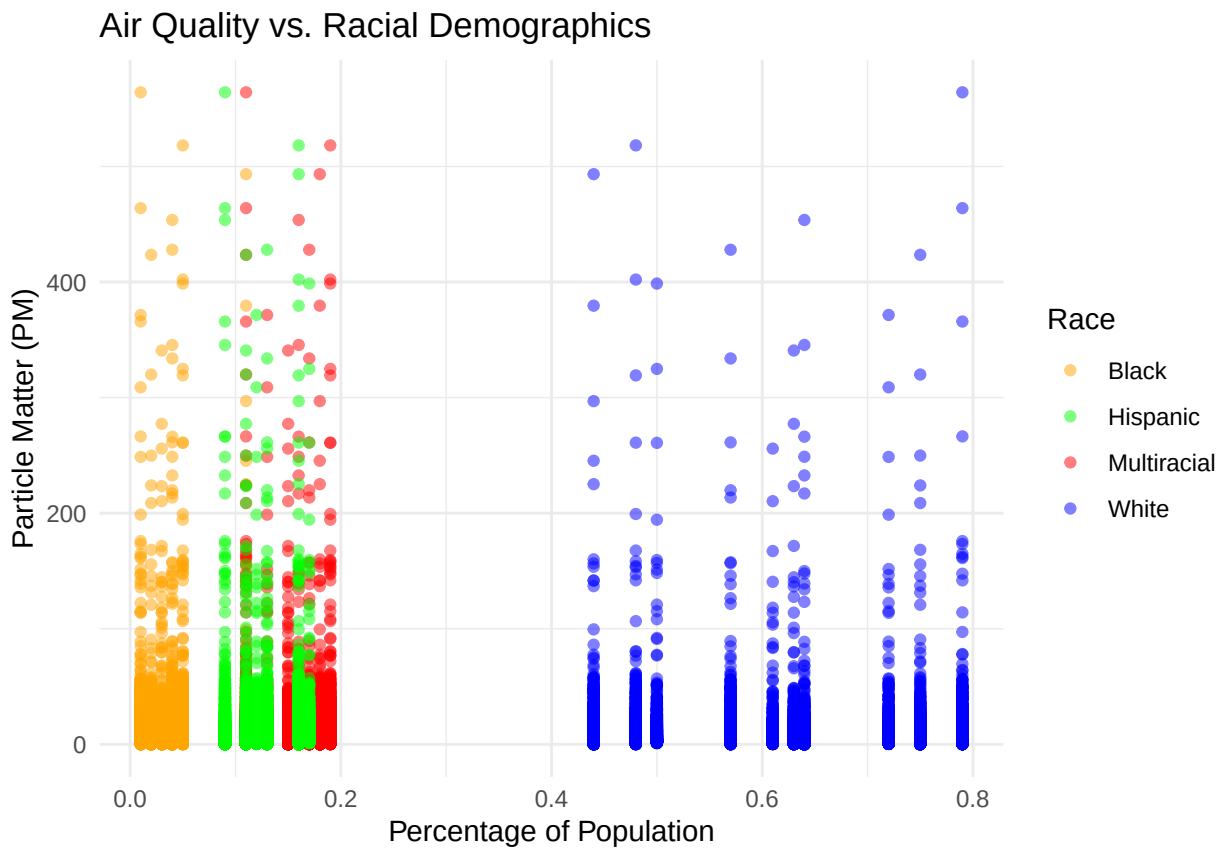

```
aq_income_lm <- lm(Particle_Matter ~ LowIncome, data = full_aq_stats)
summary(aq_income_lm)
```

```
##
## Call:
## lm(formula = Particle_Matter ~ LowIncome, data = full_aq_stats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -12.91 -10.71  -6.98   1.81 551.25
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.84374    1.27372  10.084   <2e-16 ***
## LowIncome    0.07582    1.69008   0.045    0.964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.84 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  2.391e-07,  Adjusted R-squared:  -0.0001186
## F-statistic: 0.002013 on 1 and 8418 DF,  p-value: 0.9642
```

```
ggplot() +
  # white
  geom_point(data = full_aq_stats, aes(x = White, y = Particle_Matter, color = "White"), alpha = 0.5) +

  # black
  geom_point(data = full_aq_stats, aes(x = Black, y = Particle_Matter, color = "Black"), alpha = 0.5) +

  # multiracial
  geom_point(data = full_aq_stats, aes(x = Multiracial, y = Particle_Matter, color = "Multiracial"), al

    # hispanic
  geom_point(data = full_aq_stats, aes(x = Hispanic, y = Particle_Matter, color = "Hispanic"), alpha =

  # Labels and formatting
  labs(title = "Air Quality vs. Racial Demographics",
       x = "Percentage of Population",
       y = "Particle Matter (PM)",
       color = "Race") +
  scale_color_manual(values = c("White" = "blue", "Black" = "orange", "Multiracial" = "red", "Hispanic"
  theme_minimal()
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
## Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
## Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
## Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
aq_white_lm <- lm(Particle_Matter ~ White, data = full_aq_stats)
summary(aq_white_lm)
```

```
##
## Call:
## lm(formula = Particle_Matter ~ White, data = full_aq_stats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.76 -10.70  -6.98   1.80 552.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   15.958      1.690   9.444   <2e-16 ***
## White         -4.998      2.716  -1.840   0.0658 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.84 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.0004022,  Adjusted R-squared:  0.0002834
## F-statistic: 3.387 on 1 and 8418 DF,  p-value: 0.06575
```

```
aq_black_lm <- lm(Particle_Matter ~ Black, data = full_aq_stats)
summary(aq_black_lm)
```

```
##
## Call:
## lm(formula = Particle_Matter ~ Black, data = full_aq_stats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.46 -10.69  -6.95   1.79 551.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.0361     0.5309  22.673   <2e-16 ***
## Black        22.0043    11.1063   1.981   0.0476 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.84 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.0004661,  Adjusted R-squared:  0.0003474
## F-statistic: 3.925 on 1 and 8418 DF,  p-value: 0.0476
```

```
aq_multiracial_lm <- lm(Particle_Matter ~ Multiracial, data = full_aq_stats)
summary(aq_multiracial_lm)
```

```
##
## Call:
## lm(formula = Particle_Matter ~ Multiracial, data = full_aq_stats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.50 -10.68  -6.98   1.78 552.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.950      1.675   5.939 2.98e-09 ***
## Multiracial   19.124     10.686   1.790   0.0735 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.84 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.0003803,  Adjusted R-squared:  0.0002616
## F-statistic: 3.203 on 1 and 8418 DF,  p-value: 0.07355
```

```
aq_hispanic_lm <- lm(Particle_Matter ~ Hispanic, data = full_aq_stats)
summary(aq_hispanic_lm)
```

```
##
## Call:
## lm(formula = Particle_Matter ~ Hispanic, data = full_aq_stats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.74 -10.68  -7.00   1.76 552.17
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.673      1.443   6.705 2.14e-11 ***
## Hispanic      25.396     11.100   2.288   0.0222 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.84 on 8418 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.0006215,  Adjusted R-squared:  0.0005027
## F-statistic: 5.235 on 1 and 8418 DF,  p-value: 0.02216
```