

# Impact of PM2.5 Exposure on Low Birth Weight in Ulaanbaatar (2016–2025)

Ahmedi, Barua, Bhuyan, Karayel

2025-04-29

## Contents

```
knitr::opts_knit$set(root.dir = here::here())
```

## 1 1. Load and Clean Data

```
# Read birth weight and live births data
birth_weight_low <- read.csv(here("Data/Raw/BIRTH WEIGHT LOWER THAN 2500 GRAMS.csv"), stringsAsFactors = TRUE)

live_births <- read.csv(here("./Data/Raw/LIVE BIRTHS.csv"), stringsAsFactors = TRUE)

# Clean live births: remove commas (if any) and converting to numeric()
# Might not need this as a visula check

live_births_clean <- live_births

for (col in names(live_births_clean)[-1]) {
  live_births_clean[[col]] <- as.numeric(gsub(",", "", live_births_clean[[col]]))
}

# The data is wide, need to convert the data to long format
birth_weight_low_long <- birth_weight_low %>%
  pivot_longer(-Aimag,
               names_to = "Month",
               values_to = "Low_Birth_Weight")

live_births_long <- live_births_clean %>%
  pivot_longer(-Aimag,
               names_to = "Month",
               values_to = "Live_Births")

# Merge two datasets
births_merged <- left_join(birth_weight_low_long,
                           live_births_long,
                           by = c("Aimag", "Month"))

# Removing "X" from month names
```

```

birth_weight_low_long <- birth_weight_low_long %>%
  mutate(Month = gsub("^X", "", Month))

live_births_long <- live_births_long %>%
  mutate(Month = gsub("^X", "", Month))

# Merging two datasets
births_merged <- left_join(birth_weight_low_long, live_births_long, by = c("Aimag", "Month"))

# Creating Date column
births_merged <- births_merged %>%
  mutate(Date = ym(Month)) %>%
  select(Aimag, Date, Low_Birth_Weight, Live_Births)

# # Quick checks
# str(births_merged)
# colSums(is.na(births_merged))
# summary(births_merged)
# class(births_merged)

```

```

# 1. Read and Combine All PM2.5 Files
years <- 2015:2025
pm25_files <- paste0(
  here("Data","Raw"),
  "/Ulaanbaatar_PM2.5_", years, "_YTD.csv"
)

names(pm25_files) <- years

# Read and bind all
pm25_all <- map_dfr(pm25_files, read_csv, show_col_types = FALSE)

# 2. Convert all -999 to NA across numeric columns only
pm25_all <- pm25_all %>%
  mutate(across(where(is.numeric), ~ na_if(., -999)))

# Now I need to convert all the hourly, daily, monthly and yearly data into a DateTime object. The plan

pm25_all <- pm25_all %>%
  clean_names() # Date(LT) was giving all troubles. used janitor package to rename to clean

pm25_all <- pm25_all %>%
  rename(DateTime = date_lt) %>%
  mutate(
    DateTime = parse_date_time(DateTime, orders = "ymd IMp"),
    Date = date(DateTime)
  )

# Now I will create 3 dataset, hourly, daily and monthly just to make sure if anything goes wrong I can
pm25_hourly <- pm25_all

```

```

# DAILY aggregation
# Did not use nowcast as it is smoothed data. used raw concentration

pm25_daily <- pm25_hourly %>%
  mutate(
    Date = date(DateTime)          # extract YYYY-MM-DD
  ) %>%
  group_by(Date) %>%
  summarize(
    raw_conc_daily = mean(raw_conc, na.rm = TRUE),
    aqi_daily      = mean(aqi,      na.rm = TRUE),
    hours_reported = n(),
    hours_missing_raw = sum(is.na(raw_conc)),
    hours_missing_aqi = sum(is.na(aqi)),
    .groups = "drop"
  ) %>%
  mutate(
    DateTime = as_datetime(Date)    # midnight timestamps
  )

# MONTHLY aggregation
pm25_monthly <- pm25_daily %>%
  mutate(
    Month = floor_date(Date, "month") # first day of each month
  ) %>%
  group_by(Month) %>%
  summarize(
    raw_conc_monthly = mean(raw_conc_daily, na.rm = TRUE),
    aqi_monthly      = mean(aqi_daily,      na.rm = TRUE),
    days_reported    = n(),
    days_missing_raw = sum(is.na(raw_conc_daily)),
    days_missing_aqi = sum(is.na(aqi_daily)),
    .groups = "drop"
  ) %>%
  mutate(
    DateTime = as_datetime(Month)    # first-of-month timestamps
  )

# YEARLY aggregation
pm25_yearly <- pm25_monthly %>%
  mutate(
    Year = year(Month)
  ) %>%
  group_by(Year) %>%
  summarize(
    raw_conc_yearly = mean(raw_conc_monthly, na.rm = TRUE),
    aqi_yearly      = mean(aqi_monthly,      na.rm = TRUE),
    months_reported = n(),
    months_missing_raw = sum(days_missing_raw > 0),
    months_missing_aqi = sum(days_missing_aqi > 0),
    .groups = "drop"
  ) %>%
  mutate(

```

```

    DateTime = ymd(paste0(Year, "-01-01"))    # Jan 1 of each year
  )

# Visualizing pattern

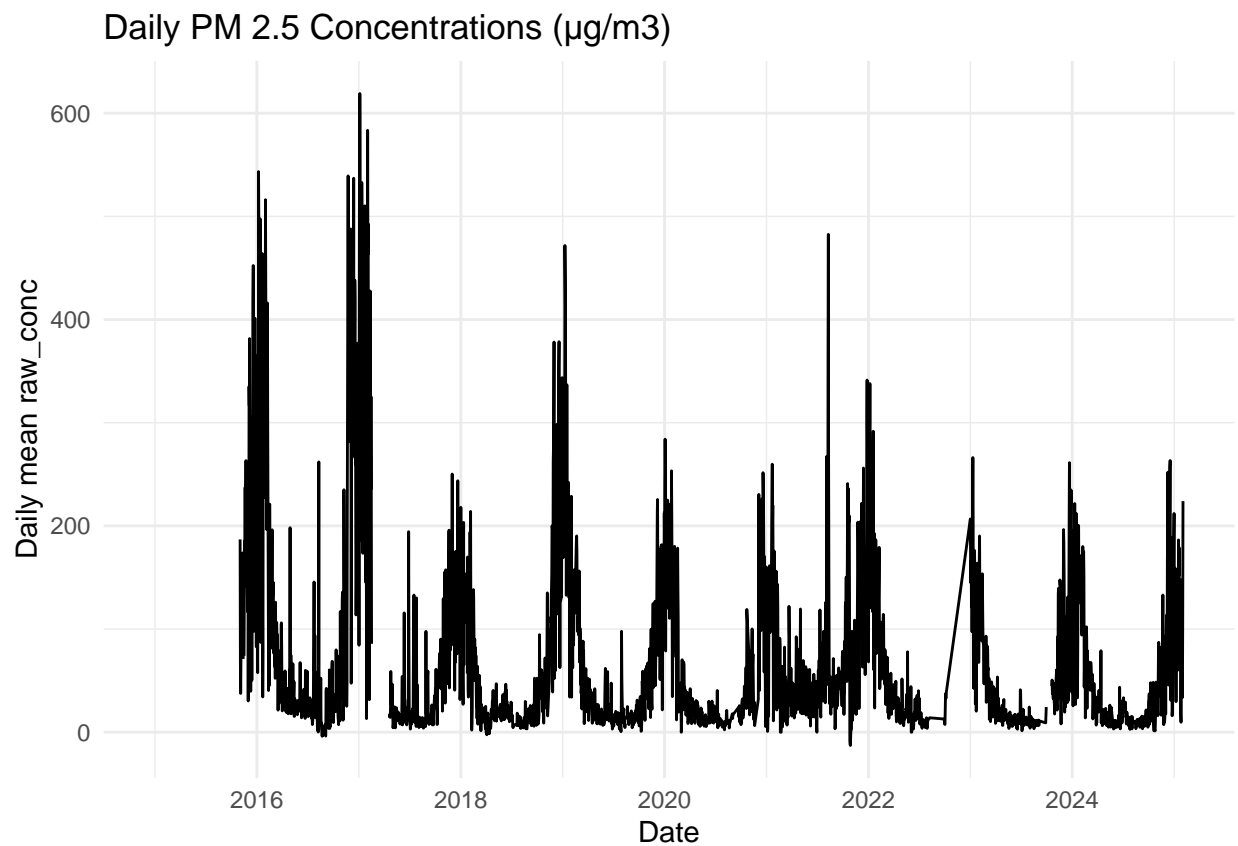
ggplot(pm25_daily, aes(x = Date, y = raw_conc_daily)) +
  geom_line() +
  labs(
    title = "Daily PM 2.5 Concentrations (µg/m3)",
    x      = "Date",
    y      = "Daily mean raw_conc"
  ) +
  theme_minimal()

```

```

## Warning: Removed 305 rows containing missing values or values outside the scale range
## (`geom_line()`).

```



```

# Treating missing value:
# visualizing missing value

# Bar chart: number of months with 1 missing day each year
# Create the pm25_yearly_missing summary
pm25_yearly_missing <- pm25_monthly %>%

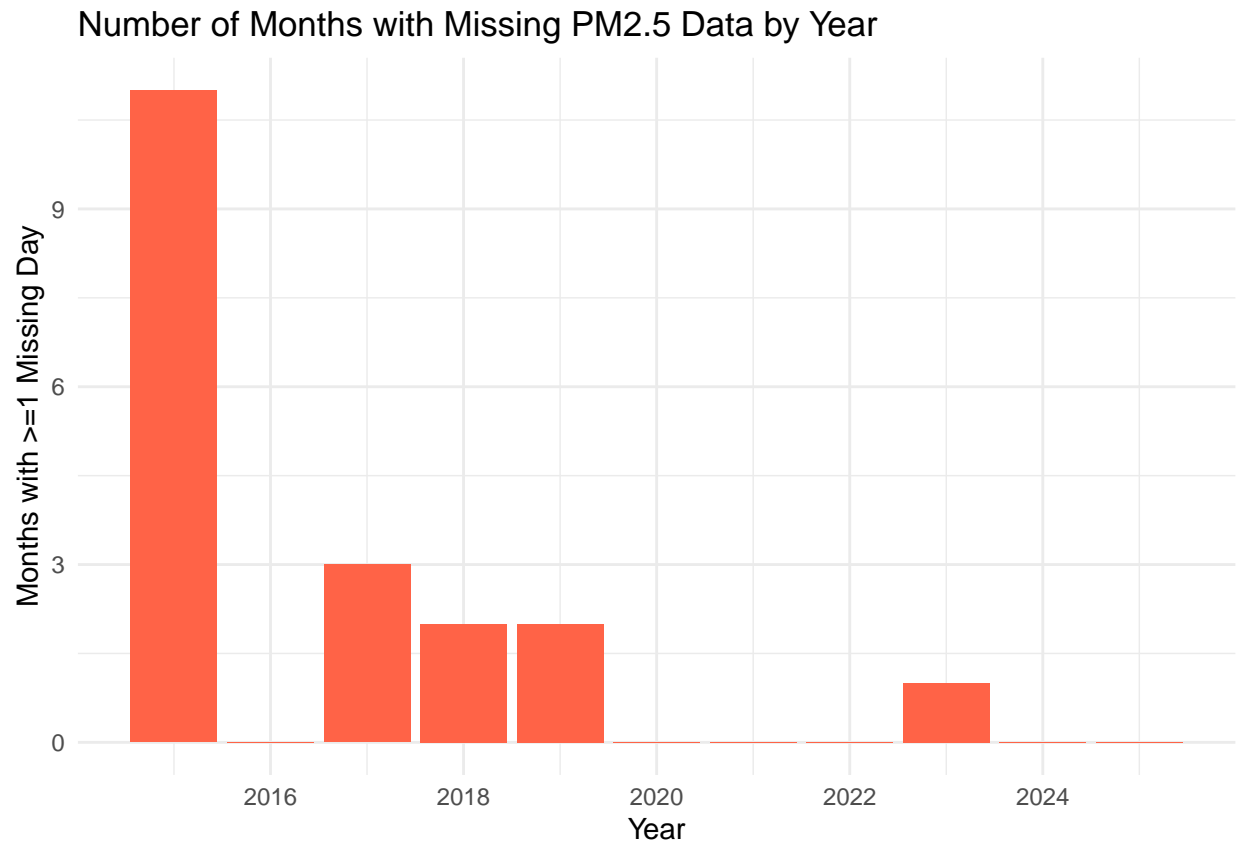
```

```

mutate(Year = year(Month)) %>%
group_by(Year) %>%
summarize(
  total_months          = n(),
  months_with_missing_days = sum(days_missing_raw > 0),
  total_missing_days     = sum(days_missing_raw),
  .groups = "drop"
)

# Plot
ggplot(pm25_yearly_missing, aes(x = Year, y = months_with_missing_days)) +
  geom_col(fill = "tomato") +
  labs(
    title = "Number of Months with Missing PM2.5 Data by Year",
    x      = "Year",
    y      = "Months with 1 Missing Day"
  ) +
  theme_minimal()

```



```

# It looks like there is 9 month of missing data in 2015. We need to consider it afterwards.

# Boxplot of monthly series to spot outliers
ggplot(pm25_monthly, aes(y = raw_conc_monthly)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 1) +
  labs(

```

```

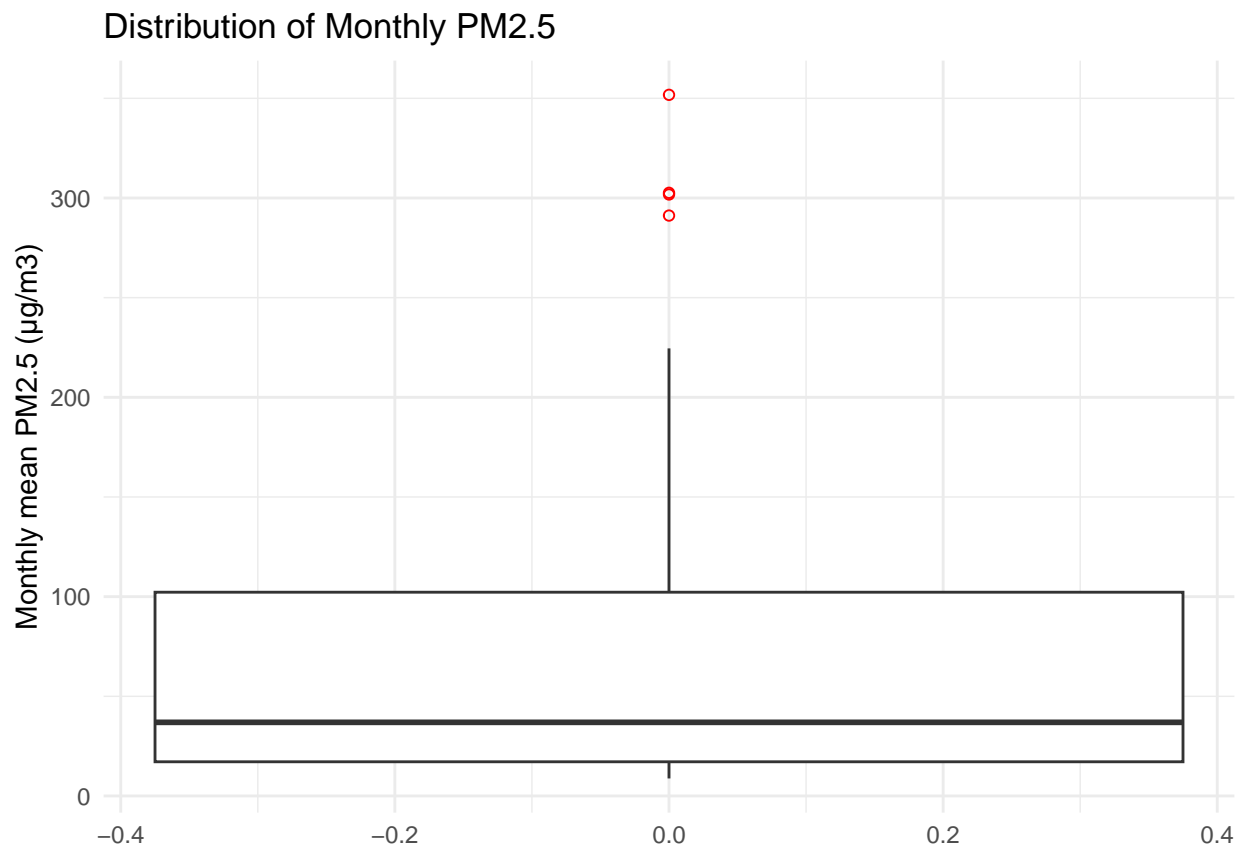
title = "Distribution of Monthly PM2.5",
y      = "Monthly mean PM2.5 (µg/m3)"
) +
theme_minimal()

```

```

## Warning: Removed 11 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

```



```

# Looks like does not have a lot of outliers. we can ignore

```

```

#write_csv(pm25_all, here("Data", "Processed", "pm25_all.csv"))

```

```

# Final NA check
summary(pm25_monthly)

```

```

##      Month      raw_conc_monthly  aqi_monthly  days_reported
##  Min.   :2015-01-01  Min.   :  8.796  Min.   : 31.73  Min.   : 1.00
## 1st Qu.:2017-06-23  1st Qu.: 17.157  1st Qu.: 55.17  1st Qu.:28.00
## Median :2019-12-16  Median : 36.937  Median : 93.02  Median :30.00
## Mean   :2019-12-29  Mean   : 69.419  Mean   :114.03  Mean   :28.19
## 3rd Qu.:2022-06-08  3rd Qu.:102.232  3rd Qu.:173.67  3rd Qu.:31.00

```

```
## Max. :2025-02-01 Max. :351.760 Max. :274.00 Max. :31.00
## NA's :11 NA's :11
## days_missing_raw days_missing_aqi DateTime
## Min. : 0.000 Min. : 0.000 Min. :2015-01-01 00:00:00
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.:2017-06-23 12:00:00
## Median : 0.000 Median : 0.000 Median :2019-12-16 12:00:00
## Mean : 3.125 Mean : 3.225 Mean :2019-12-29 15:48:00
## 3rd Qu.: 0.000 3rd Qu.: 0.000 3rd Qu.:2022-06-08 12:00:00
## Max. :31.000 Max. :31.000 Max. :2025-02-01 00:00:00
##
```

```
colSums(is.na(pm25_monthly))
```

```
## Month raw_conc_monthly aqi_monthly days_reported
## 0 11 11 0
## days_missing_raw days_missing_aqi DateTime
## 0 0 0
```

```
# Merge PM2.5 with births
```

```
full_data <- births_merged %>%
  left_join(
    pm25_monthly,
    by = c("Date" = "Month")
  ) %>%
  arrange(Date)
```

```
full_data %>%
  select(Date, Aimag, Low_Birth_Weight, Live_Births, raw_conc_monthly, aqi_monthly)
```

```
## # A tibble: 111 x 6
## Date Aimag Low_Birth_Weight Live_Births raw_conc_monthly aqi_monthly
## <date> <fct> <int> <dbl> <dbl> <dbl>
## 1 2016-01-01 Ulaanba~ 168 3221 291. 236.
## 2 2016-02-01 Ulaanba~ 152 3158 197. 208.
## 3 2016-03-01 Ulaanba~ 162 3401 73.6 131.
## 4 2016-04-01 Ulaanba~ 132 3229 39.6 87.1
## 5 2016-05-01 Ulaanba~ 130 3546 30.5 81.1
## 6 2016-06-01 Ulaanba~ 146 3450 29.3 78.4
## 7 2016-07-01 Ulaanba~ 179 3696 33.6 77.9
## 8 2016-08-01 Ulaanba~ 172 3556 17.2 31.7
## 9 2016-09-01 Ulaanba~ 148 3421 22.5 51.5
## 10 2016-10-01 Ulaanba~ 159 3566 36.9 83.7
## # i 101 more rows
```

```
# Summary for birth outcomes
```

```
births_summary <- full_data %>%
  summarise(
    Mean_LBW = mean(Low_Birth_Weight, na.rm = TRUE),
    Median_LBW = median(Low_Birth_Weight, na.rm = TRUE),
    Min_LBW = min(Low_Birth_Weight, na.rm = TRUE),
    Max_LBW = max(Low_Birth_Weight, na.rm = TRUE),
```





```
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
## Warning in attr(.knitEnv$meta, "knit_meta_id"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
## Warning in attr(.knitEnv$meta, "knit_meta_id"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
## Warning in attr(.knitEnv$meta, "knit_meta_id"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

Table 1: Summary of Birth Outcomes

Statistic	Value
Mean_LBW	155.97
Median_LBW	153.00
Min_LBW	88.00
Max_LBW	214.00
SD_LBW	21.87
N_LBW	111.00
Mean_Live	3110.86
Median_Live	3187.00
Min_Live	1934.00
Max_Live	3737.00
SD_Live	360.42
N_Live	111.00

```
# 2. Summary for PM2.5 exposure
pm25_summary <- full_data %>%
  summarise(
    Mean_PM25 = mean(raw_conc_monthly, na.rm = TRUE),
    Median_PM25 = median(raw_conc_monthly, na.rm = TRUE),
    Min_PM25 = min(raw_conc_monthly, na.rm = TRUE),
    Max_PM25 = max(raw_conc_monthly, na.rm = TRUE),
    SD_PM25 = sd(raw_conc_monthly, na.rm = TRUE),
    N_PM25 = sum(!is.na(raw_conc_monthly)),

    Mean_AQI = mean(aqi_monthly, na.rm = TRUE),
    Median_AQI = median(aqi_monthly, na.rm = TRUE),
    Min_AQI = min(aqi_monthly, na.rm = TRUE),
    Max_AQI = max(aqi_monthly, na.rm = TRUE),
    SD_AQI = sd(aqi_monthly, na.rm = TRUE),
```

[illegible]

```
## Warning in attr(.knitEnv$meta, "knit_meta_id"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

Table 2: Summary of Monthly PM2.5 Exposure

Statistic	Value
Mean_PM25	67.28
Median_PM25	35.22
Min_PM25	8.80
Max_PM25	351.76
SD_PM25	72.90
N_PM25	107.00
Mean_AQI	111.86
Median_AQI	92.67
Min_AQI	31.73
Max_AQI	274.00
SD_AQI	66.15
N_AQI	107.00

## 2 Descriptive Statistics

```
# Compute low birth weight rate (Percentage)
full_data <- full_data %>%
  mutate(
    LBW_rate = 100 * Low_Birth_Weight / Live_Births
  )

# Summary table of exposure and outcome
summary_tbl <- full_data %>%
  summarise(
    Mean_PM25      = mean(raw_conc_monthly, na.rm = TRUE),
    SD_PM25        = sd(raw_conc_monthly, na.rm = TRUE),
    Mean_LBWrate    = mean(LBW_rate, na.rm = TRUE),
    SD_LBWrate      = sd(LBW_rate, na.rm = TRUE),
    N               = n()
  ) %>%
  pivot_longer(everything(), names_to="Metric", values_to="Value")

summary_tbl %>%
  kable(caption="Summary of PM2.5 and LBW Rate", digits=2) %>%
  kable_styling(full_width=FALSE)
```



Table 3: Summary of PM2.5 and LBW Rate

Metric	Value
Mean_PM25	67.28
SD_PM25	72.90
Mean_LBWrate	5.03
SD_LBWrate	0.57
N	111.00

```
# Scatter + trend line
ggplot(full_data, aes(x = raw_conc_monthly, y = LBW_rate)) +
  geom_point() +
  geom_smooth(method="lm", se=TRUE, color="blue") +
  labs(
    title = "Low Birth Weight Rate vs. Monthly PM2.5",
    x      = "PM2.5 (µg/m3)",
    y      = "LBW Rate (Percentage)"
  ) +
  theme_minimal()
```

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

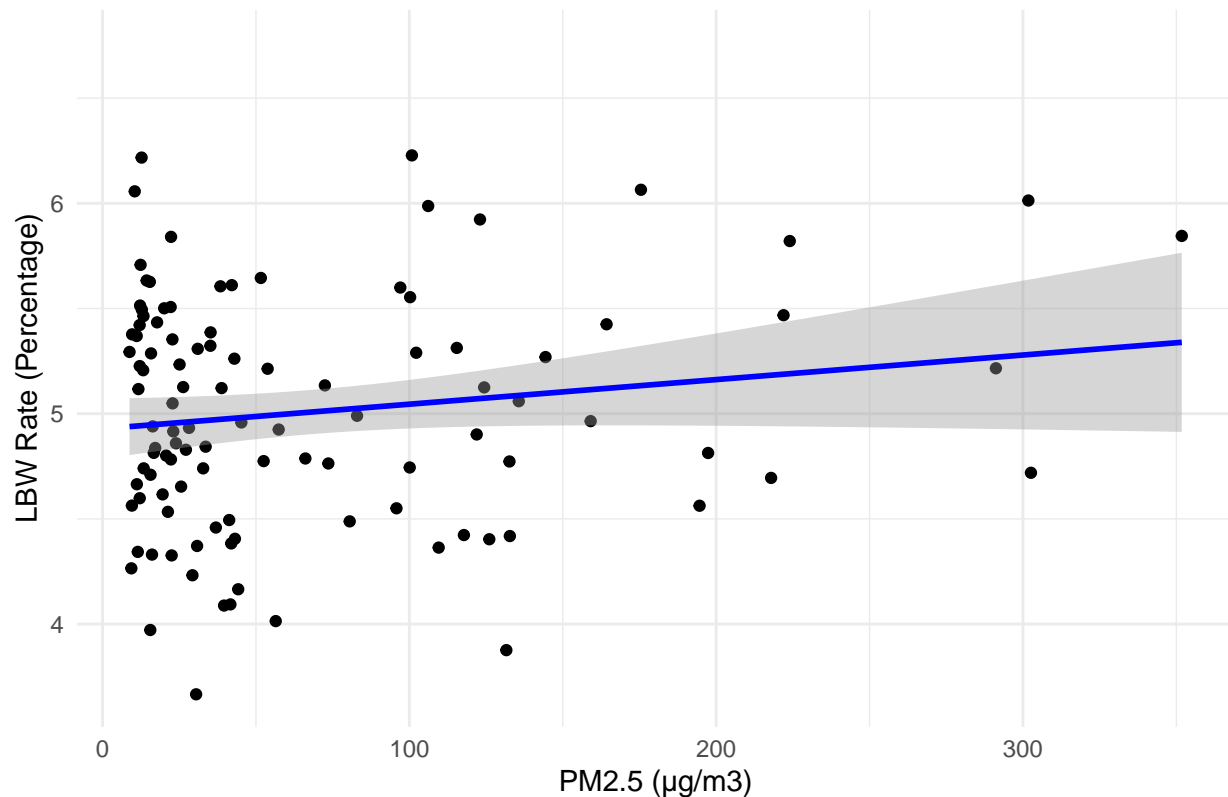
```
## Warning: Removed 4 rows containing non-finite outside the scale range
```

```
## (`stat_smooth()`).
```

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
```

```
## (`geom_point()`).
```

## Low Birth Weight Rate vs. Monthly PM2.5



# Linear regression

```
model <- lm(LBW_rate ~ raw_conc_monthly, data = full_data)
tidy(model) %>%
  kable(caption="Regression of LBW Rate on PM2.5", digits=3) %>%
  kable_styling(full_width=FALSE)
```

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

```
## Warning in attr(.knitEnv$meta, "knit_meta_id"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
## Warning in attr(.knitEnv$meta, "knit_meta_id"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
## Warning in attr(.knitEnv$meta, "knit_meta_id"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
## Warning in attr(.knitEnv$meta, "knit_meta_id"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
## Warning in attr(.knitEnv$meta, "knit_meta_id"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```