Project Part 1

Question 1: Is the average daily PM2.5 in Los Angeles County during the summer of 2024 (June to August) significantly lower than that in the winter (December to February)?

Introduction

Air pollution remains a major public health concern in large metropolitan regions, particularly in Southern California. Fine particulate matter (PM2.5) is of special interest because particles smaller than 2.5 micrometers can penetrate deep into the lungs and bloodstream, contributing to respiratory and cardiovascular risks.

Seasonal variation is commonly observed in PM2.5 levels, with winter often associated with higher concentrations due to weaker atmospheric mixing and increased heating-related emissions. To explore this pattern in Los Angeles County, we compare daily PM2.5 levels between summer (June–August 2024) and winter (December 2023–February 2024). Specifically, we ask: Is the daily average PM2.5 level in Los Angeles County lower in summer than in winter?

To answer this question, we use publicly available EPA AirData daily PM2.5 measurements and summarize daily county-level averages across monitoring sites.

Data Summary

Source: EPA AirData
Time span: 2023-12–2024-08
Measure: Daily mean PM2.5 ($\mu g/m^3$)
Aggregation: county-day mean across sites

```
##
## Summer Winter
##   1203   1190
```

```
##         Date pm25      County month season
## 1 2024-01-01  2.8 Los Angeles     1 Winter
## 2 2024-01-02  6.5 Los Angeles     1 Winter
## 3 2024-01-03  6.2 Los Angeles     1 Winter
## 4 2024-01-04  2.9 Los Angeles     1 Winter
## 5 2024-01-05  4.0 Los Angeles     1 Winter
## 6 2024-01-06  6.7 Los Angeles     1 Winter
```

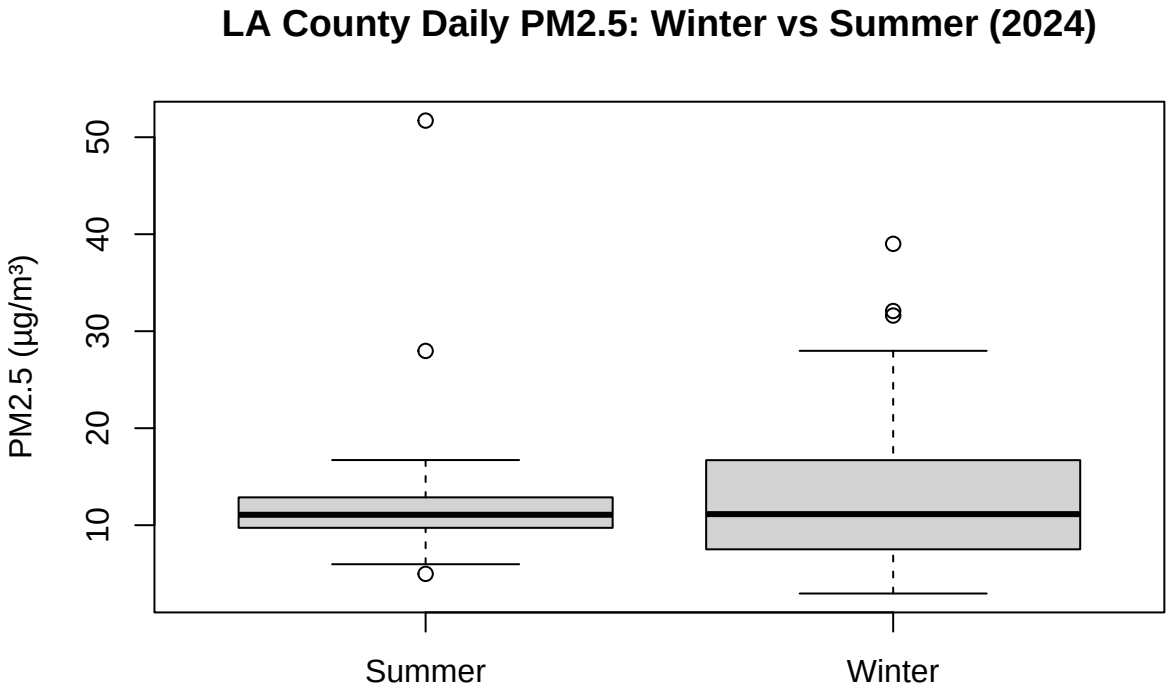
```
##
## Summer Winter
##     92     91
```

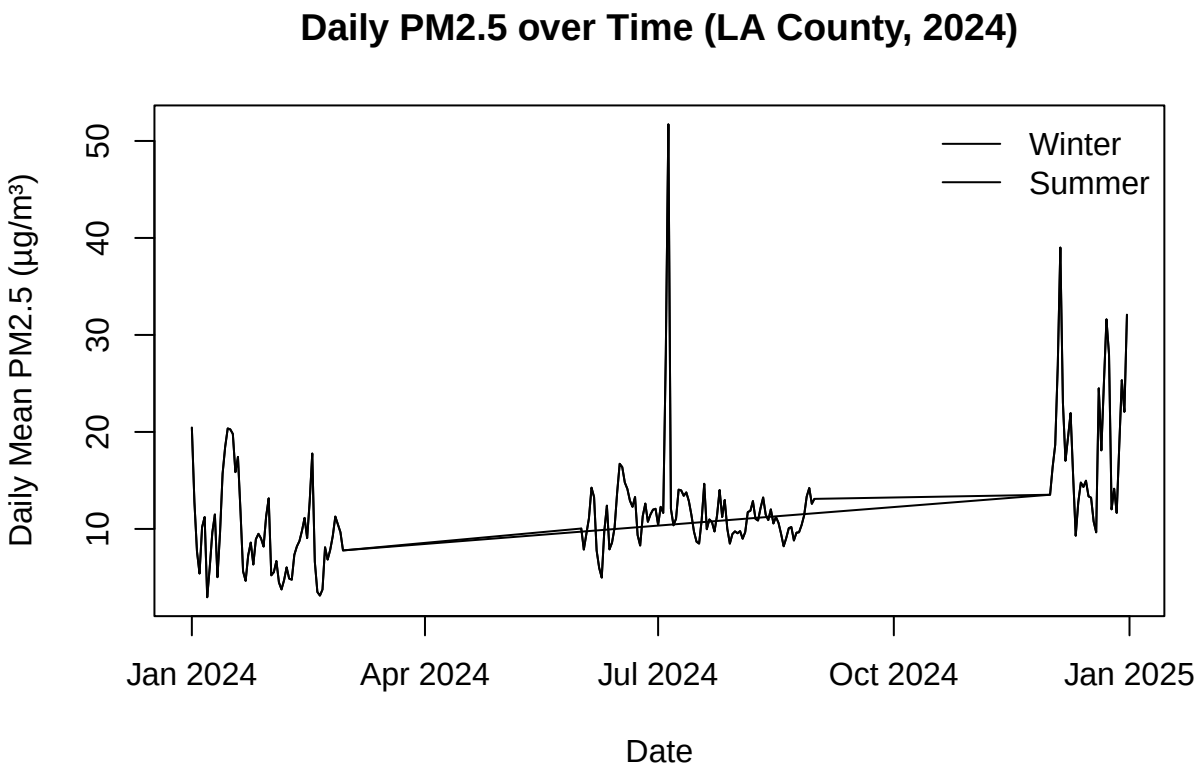
```
##         Date season      pm25
## 1 2024-06-01 Summer 10.041667
## 2 2024-06-02 Summer  7.873333
## 3 2024-06-03 Summer  9.500000
## 4 2024-06-04 Summer 11.100000
## 5 2024-06-05 Summer 14.252632
## 6 2024-06-06 Summer 13.290909
```

```
##   Summer   Winter
## 11.76828 12.65374
```

```
##   Summer   Winter
## 5.019253 7.262527
```

Exploratory Analysis





We computed county-level daily means of PM2.5 by averaging across monitoring sites for each date. Comparing Summer (June–August) and Winter (December–February) days in 2024, the summer mean PM2.5 was 11.77 µg/m^3 while the winter mean was 12.65 µg/m^3, suggesting higher daily concentrations in winter on average. The standard deviation was 5.02 (summer) vs 7.26 (winter), indicating greater day-to-day variability in winter. Boxplots show a generally higher center and a wider spread for winter. The time-series plot reveals elevated winter levels and several winter peaks, whereas summer tends to be lower and more stable. These summaries are descriptive only; no statistical inference is conducted in Part 1.

Conclusions

Based on descriptive summaries, daily PM2.5 levels in Los Angeles County appear lower in summer (mean = 11.77 µg/m^3) than in winter (mean = 12.65 µg/m^3). Winter days also show greater variability (SD = 7.26 vs. 5.02), suggesting more frequent short-term pollution spikes during colder months.

Boxplots illustrate a higher center and wider spread for winter PM2.5 values, while the time-series visualization shows elevated concentrations and several distinct peaks during winter

months. Summer PM2.5 levels are generally lower and more stable, consistent with stronger atmospheric mixing and fewer seasonal emission sources.

These results align with typical seasonal air quality patterns in Southern California. However, this analysis is descriptive only. In Part 2, we will apply formal statistical inference to determine whether the observed difference is statistically significant and quantify uncertainty around seasonal mean estimates.

Question 2: Did NBA teams have "home-court advantage" during the period from 2014-10-04 to 2022-12-22?

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- 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
##
## Attaching package: 'janitor'
##
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
##
##
##
## Attaching package: 'scales'
##
##
## The following object is masked from 'package:purrr':
##
##     discard
##
##
## The following object is masked from 'package:readr':
##
##     col_factor

## # A tibble: 1 x 7
```

```
##   rows_all rows_range used_rows date_min_available date_max_available mean_diff
##      <int>      <int>     <int> <date>             <date>                   <dbl>
## 1    26552      11420     11420 2014-10-04         2022-12-22                2.35
## # i 1 more variable: home_win_rate <dbl>
```

Introduction

Basketball fans often debate whether home teams really perform better. This report studies the home-court advantage in the NBA from 2014-10-04 to 2022-12-22, asking: Do home teams, on average, score more points and win more often than away teams? **Research Question:**
Does the NBA exhibit a statistically measurable *home-court advantage* between October 4 2014 and December 22 2022?
Specifically, do home teams on average score more points than away teams?

This question is explored using game-level data that include both team scores and win/loss indicators. ### Data Summary

Data Source.
The dataset contains results of NBA games from *October 4 2014 through December 22 2022*. These data were originally collected from official NBA box scores and provided in a structured CSV file containing each game's date, home and away team identifiers, total points, and a binary indicator of whether the home team won.

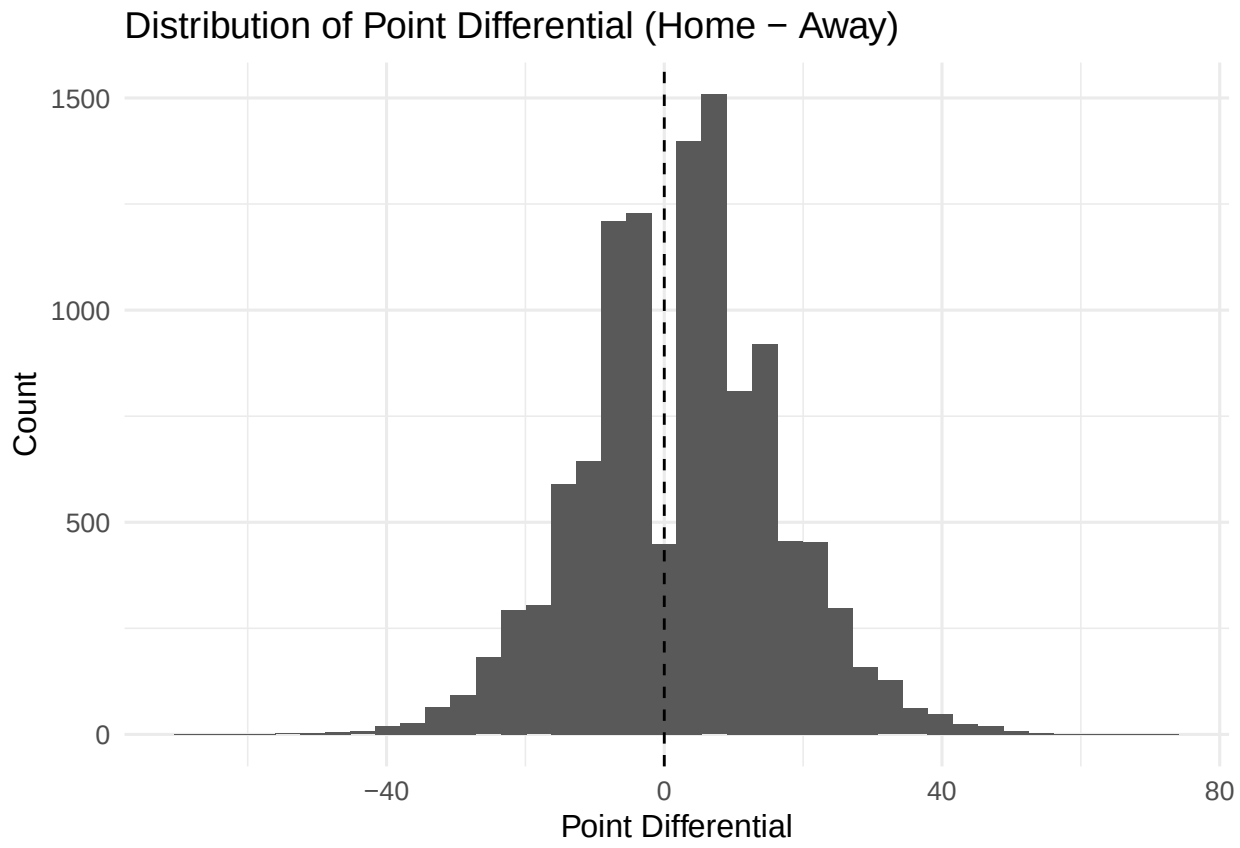
Population or Sample.
The file represents a large sample of NBA games during this period rather than a complete population of all historical games. The sample covers 11,420 valid games after cleaning.

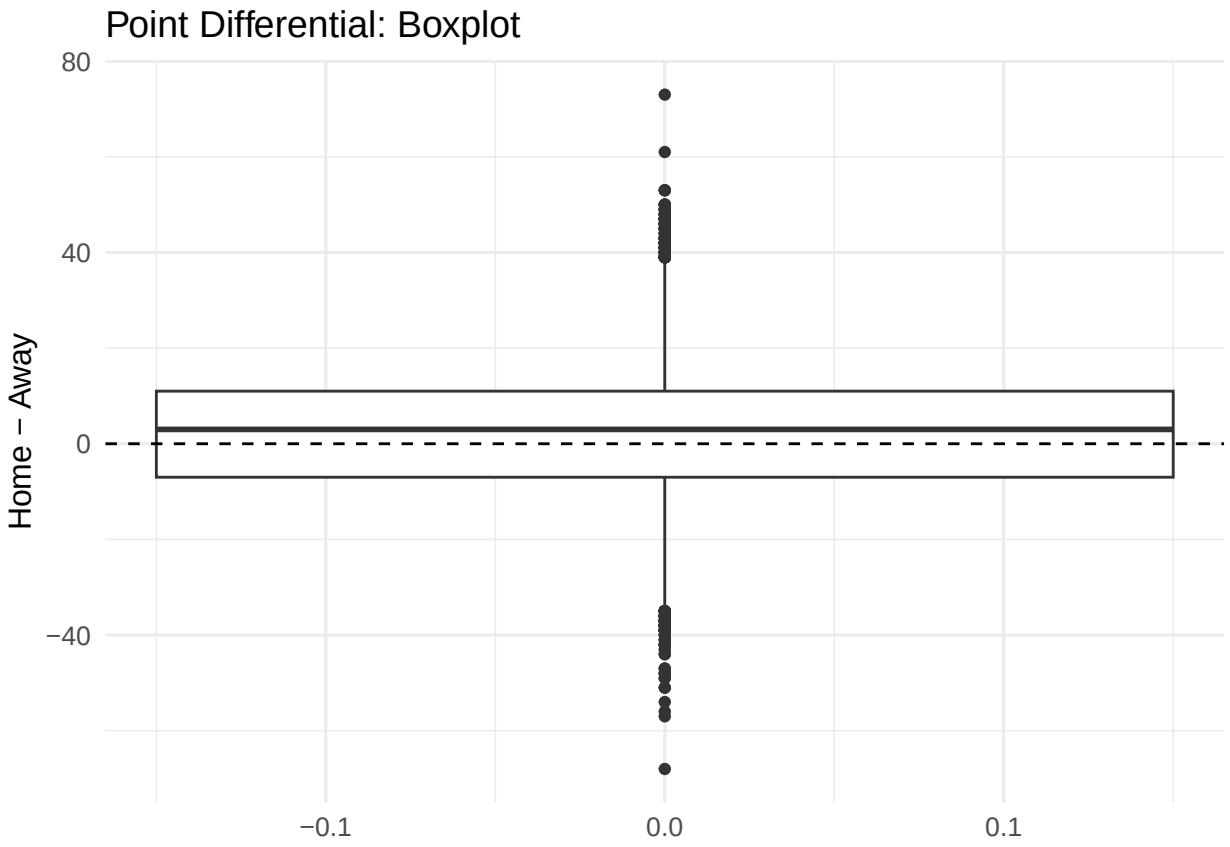
Data Modifications.
After importing the raw CSV, column names were standardized, dates were parsed into `Date` format, and missing or invalid scores were removed.
A new variable `diff = PTS_home - PTS_away` was created to measure the home–away point differential, and a binary variable `home_win` = 1 if the home team won.
Only games with complete score information within the target date range were retained.

Potential Issues.
Some seasons include playoff or preseason games, and rule changes (e.g., shot clock or foul interpretations) may influence scoring. Additionally, team composition varies over years, which introduces unobserved heterogeneity. These factors may slightly bias the magnitude of the estimated advantage but not the direction.

Appropriateness of the Data.
Because each row represents one game with independent outcomes and contains both home and away points, the dataset is directly suited to quantify the home-court advantage through summary comparisons of point differentials and win percentages.

```
# ---- P1-EDA ----
p1 <- ggplot(games, aes(diff)) +
  geom_histogram(bins = 40) +
  geom_vline(xintercept = 0, linetype = 2) +
  labs(title = "Distribution of Point Differential (Home - Away)",
       x = "Point Differential", y = "Count")
p2 <- ggplot(games, aes(y = diff)) +
  geom_boxplot(width = .3) +
  geom_hline(yintercept = 0, linetype = 2) +
  labs(title = "Point Differential: Boxplot", x = NULL, y = "Home - Away")
p1; p2
```






```
games |>
  summarise(
    n_games = n(),
    mean_diff = mean(diff),
    sd_diff = sd(diff),
    home_win_rate = if (!all(is.na(home_win))) scales::percent(mean(home_win)) else NA_c
  )
```

```
## # A tibble: 1 x 4
##   n_games mean_diff sd_diff home_win_rate
##     <int>     <dbl>   <dbl> <chr>
## 1   11420      2.35    14.2 57%
```

Exploratory Analysis

Numerical Summaries.

- Number of games: 11,420

- Mean point differential (`PTS_home - PTS_away`): 2.35 points
- Standard deviation: 14.25 points
- Home win rate: 57.1 %

Graphical Summaries.
Two plots illustrate the distribution of point differentials:
1. *Histogram of Point Differential (Figure 1)* – shows a nearly symmetric distribution centered slightly above 0.
2. *Boxplot of Point Differential (Figure 2)* – median and interquartile range lie on the positive side, confirming a consistent upward shift for home teams.

Interpretation of Summaries.
Most games cluster within ±20 points, with a slight right shift of the distribution and median above 0. The average differential of +2.35 points suggests that home teams outscore visiting teams by roughly two to three points on average.
The home win rate of 57 % further supports a moderate but consistent home advantage.

Conclusions

From the exploratory summaries, both numerical and graphical evidence indicate the presence of a meaningful home-court advantage in the NBA between 2014 and 2022.

- The mean score differential of +2.35 points implies that home teams typically score slightly more than their opponents.

- The distribution of differentials centers on the positive side, suggesting the effect is widespread rather than driven by outliers.

- A home win rate near 57 % confirms that the majority of games are won by home teams, consistent with the positive mean differential.

While these results describe a clear pattern, they do not yet constitute formal statistical inference. Part 2 of the project will test whether the observed mean difference and win proportion are statistically greater than zero and 0.5, respectively.

In context, the findings align with sports psychology research suggesting that fan support, reduced travel stress, and familiarity with local conditions contribute to better home performance.

```
# ---- P2-Inference ----
if (sum(!is.na(games$diff)) >= 2) {
  t_res <- t.test(games$diff, mu = 0, alternative = "greater")
  t_res
  broom::tidy(t_res)
} else {
  message("Not enough non-missing values in 'diff' to conduct a t-test (need at least 2
}
```

```
## # A tibble: 1 x 8
##   estimate statistic  p.value parameter conf.low conf.high method    alternative
##      <dbl>     <dbl>    <dbl>     <dbl>    <dbl>     <dbl> <chr>     <chr>
## 1     2.35      17.6 1.13e-68     11419     2.13       Inf One Samp~ greater
```

```
if (!all(is.na(games$home_win))) {
  wins <- sum(games$home_win, na.rm = TRUE)
  n    <- sum(!is.na(games$home_win))
  if (n >= 1) {
    prop_res <- prop.test(x = wins, n = n, p = 0.5, alternative = "greater", correct = F
    prop_res
    broom::tidy(prop_res)
  } else {
    message("There are no valid 'home_win' observations available for the proportion tes
  }
} else {
  message("No 'home_win' column found (or all values missing), skipping the proportion t
}
```

```
## # A tibble: 1 x 8
##   estimate statistic  p.value parameter conf.low conf.high method    alternative
##      <dbl>     <dbl>    <dbl>     <int>    <dbl>     <dbl> <chr>     <chr>
## 1    0.571      231. 1.86e-52         1    0.563         1 1-sample~ greater
```

References

- https://www.epa.gov/outdoor-air-quality-data/download-daily-data
- National Basketball Association. *Game Box Score Statistics.* Data compiled from official NBA results, 2014–2022.
- R Core Team (2025). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.
- Wickham, H. (2019). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag, New York.
- UCLA Statistical Consulting Group. "T-Tests and Proportion Tests in R." (Accessed 2025).
- https://www.kaggle.com/datasets/nathanlauga/nba-games


```
## R code to conduct the test in part 2.
## This R code will be printed in the PDF.
## Remove the ## before all code.
```