National Park Mini-Project

Meredith Sliger

Importing and cleaning data

To begin this project, I collected and organized annual visitation data from eight U.S. national parks. The goal was to bring these separate files into a single clean dataset that could be used for analysis. Since the raw CSVs contained metadata rows and varying formats, I used R to standardize them, filter by year, and ensure all numeric columns were correctly formatted.

```
library(tidyverse)

# Set your working directory
setwd("C:/Users/mered/OneDrive/Documents/miniproject1natlparks")

# Park names and file names
parks <- tibble(
  park_name = c("Yosemite", "Olympic", "Everglades", "Big Bend",
                "Glacier", "Voyageurs", "Great Smoky Mountains", "Cuyahoga Valley"),
  file = c("yosemite.csv", "olympic.csv", "everglades.csv", "bigbend.csv",
           "glacier.csv", "voyageurs.csv", "greatsmokymountains.csv", "cuyahogavalley.csv")
)

# Load, clean, and combine
all_parks <- parks %>%
  rowwise() %>%
  mutate(data = list(
    read_csv(file, skip = 3, show_col_types = FALSE) %>%  # Skip metadata rows
      select(Year, Visits = RecreationVisitors) %>%
      filter(!is.na(Year), Year >= 1990) %>%
      mutate(
        Visits = as.numeric(gsub(",", "", Visits)),  # Remove commas, convert to numeric
        Park = park_name
      )
  )) %>%
  pull(data) %>%
  bind_rows()

# Preview
glimpse(all_parks)
```

```
## Rows: 280
## Columns: 3
## $ Year   <dbl> 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 200~
## $ Visits <dbl> 3124939, 3423101, 3819518, 3839645, 3962117, 3958406, 4046207, ~
## $ Park   <chr> "Yosemite", "Yosemite", "Yosemite", "Yosemite", "Yosemite", "Yo~
```

```r
# Save cleaned file (optional)
write_csv(all_parks, "combined_parks_data.csv")
```

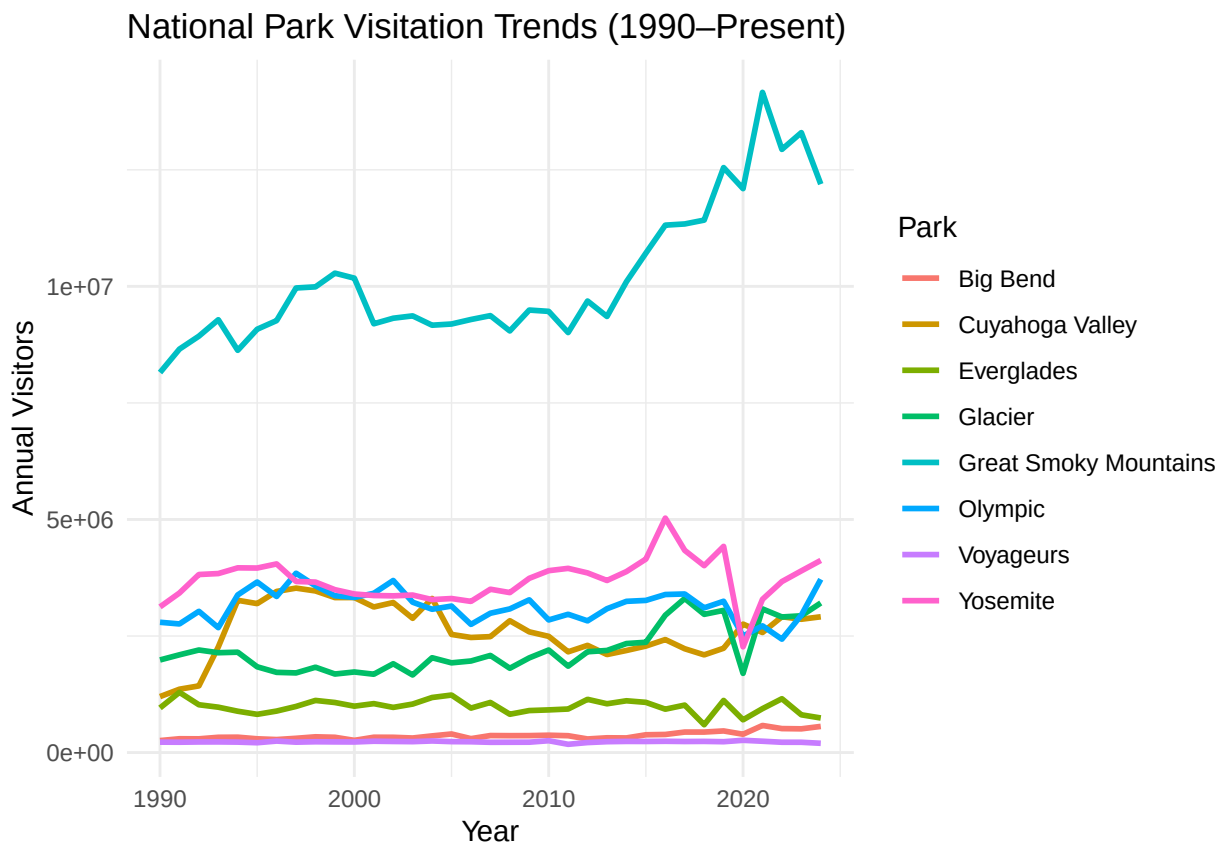
The final dataset includes 280 rows of cleaned annual data from 1990 onward. This ensures consistency across all parks and provides a solid foundation for analysis.

Plotting Trends

Now that the data is cleaned and combined, I created a line plot to visualize how visitation has changed over time for each park. This gives a broad look at growth patterns, stability, or any noticeable shifts in usage.

```r
library(ggplot2)

# Historical visitation plot
print(
historical_plot <- ggplot(all_parks, aes(x = Year, y = Visits, color = Park)) +
  geom_line(linewidth = 1) +
  labs(title = "National Park Visitation Trends (1990-Present)",
       x = "Year", y = "Annual Visitors",
       color = "Park") +
  theme_minimal()
)
```



```
ggsave("historic_trends.png", plot = historical_plot, width = 10, height = 6, dpi = 300)
```

The plot shows that some parks like Great Smoky Mountains and Cuyahoga Valley have seen strong growth, while others such as Voyageurs and Big Bend have maintained relatively steady visitation. Spikes and dips can also be spotted, such as the decline in 2020 likely tied to the COVID-19 pandemic.

Forecasting

To explore how visitation might change in the future, I applied time series forecasting using ARIMA models. Each park was modeled individually to capture its unique visitation pattern, and forecasts were generated for the next five years.

```
library(forecast)

# Create a list to store forecasts
forecasts <- list()

# Loop through each park
for (park in unique(all_parks$Park)) {
  park_data <- all_parks %>%
    filter(Park == park) %>%
    arrange(Year)

  park_ts <- ts(park_data$Visits, start = min(park_data$Year), frequency = 1)

  model <- auto.arima(park_ts)
  fc <- forecast(model, h = 5)

  fc_tibble <- as_tibble(fc) %>%
    mutate(
      Year = seq(max(park_data$Year) + 1, by = 1, length.out = 5),
      Park = park
    ) %>%
    select(Park, Year, PointForecast = `Point Forecast`, Lo95 = `Lo 95`, Hi95 = `Hi 95`)

  forecasts[[park]] <- fc_tibble
}

# Combine all forecasts into one dataframe
all_forecasts <- bind_rows(forecasts)

# Preview
head(all_forecasts)
```

```
## # A tibble: 6 x 5
##   Park      Year PointForecast     Lo95     Hi95
##   <chr>    <dbl>         <dbl>    <dbl>    <dbl>
## 1 Yosemite  2025      3877045. 3035620. 4718470.
## 2 Yosemite  2026      3773052. 2858831. 4687273.
## 3 Yosemite  2027      3728868. 2802115. 4655622.
## 4 Yosemite  2028      3710096. 2781099. 4639093.
```

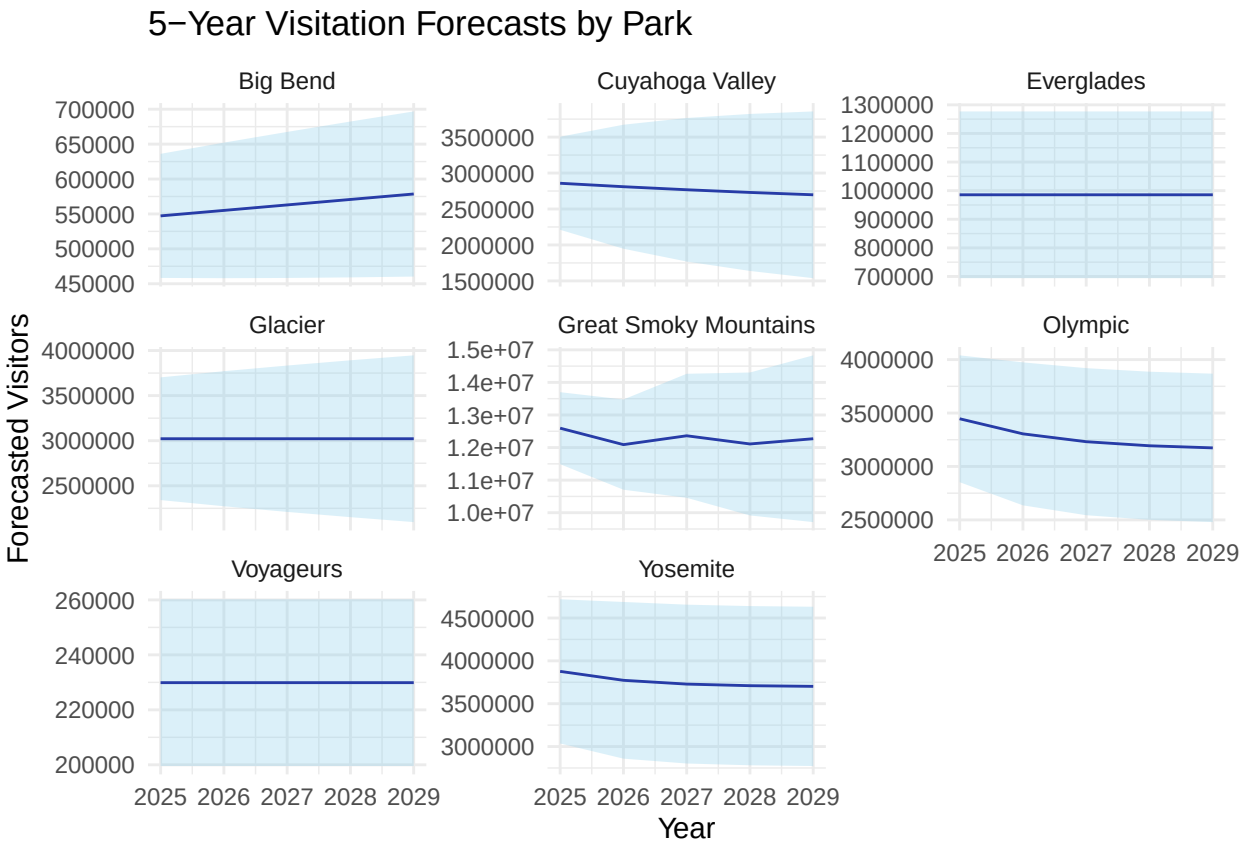
```
## 5 Yosemite  2029      3702120. 2772718. 4631522.
## 6 Olympic   2025      3446169. 2852598. 4039741.
```

Each park now has a five-year forecast with associated confidence intervals. This will help visualize not only expected trends but also the uncertainty in those predictions.

Forecast Plotting

The following plot presents the forecasts for each park on separate axes. This allows for a clear comparison of projected growth or stabilization over time.

```
ggplot(all_forecasts, aes(x = Year, y = PointForecast)) +
  geom_line(color = "darkblue") +
  geom_ribbon(aes(ymin = Lo95, ymax = Hi95), fill = "skyblue", alpha = 0.3) +
  facet_wrap(~Park, scales = "free_y") +
  labs(title = "5-Year Visitation Forecasts by Park",
       y = "Forecasted Visitors", x = "Year") +
  theme_minimal()
```



Some parks, such as Yosemite and Cuyahoga Valley, are forecasted to maintain high or growing visitation. Others, like Voyageurs or Big Bend, are expected to remain relatively steady, with broader uncertainty bands reflecting more variable historical patterns.

Historical and Forecast Data combined

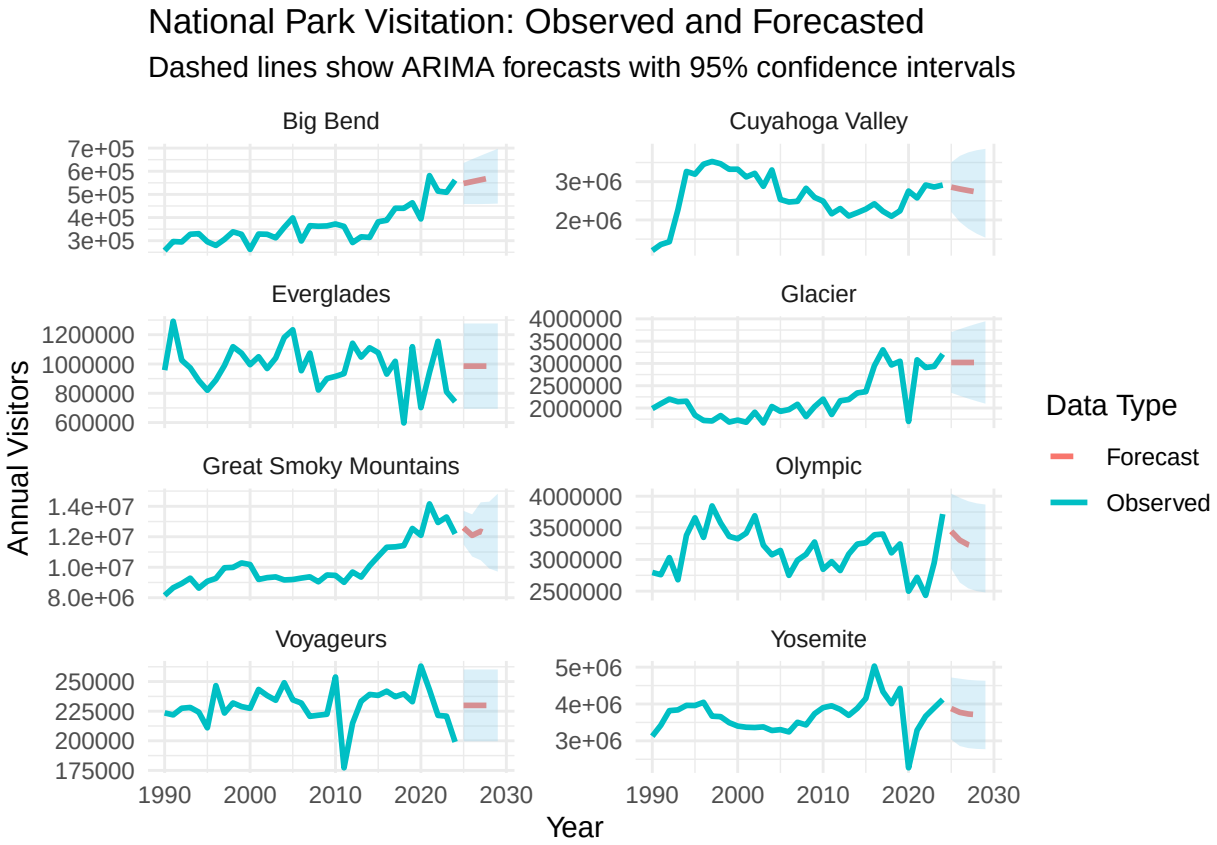
To make comparisons easier, I combined both observed and forecasted values into a single plot. Dashed lines indicate predicted values, and shaded bands represent 95% confidence intervals.

```
# Prepare historical data for plotting
historical_data <- all_parks %>%
  select(Park, Year, Visitors = Visits) %>%
  mutate(Type = "Observed")

# Prepare forecast data
forecast_data <- all_forecasts %>%
  select(Park, Year, Visitors = PointForecast, Lo95, Hi95) %>%
  mutate(Type = "Forecast")

# Combine both
combined_data <- bind_rows(historical_data, forecast_data)
```

```
print(
# Combined historical + forecast plot
final_plot <- ggplot(combined_data, aes(x = Year, y = Visitors, color = Type)) +
  geom_line(data = subset(combined_data, Type == "Observed"), linewidth = 1) +
  geom_line(data = subset(combined_data, Type == "Forecast"), linetype = "dashed", linewidth = 1) +
  geom_ribbon(
    data = forecast_data,
    aes(x = Year, ymin = Lo95, ymax = Hi95),
    fill = "skyblue", alpha = 0.3, inherit.aes = FALSE
  ) +
  facet_wrap(~Park, ncol = 2, scales = "free_y") +
  labs(
    title = "National Park Visitation: Observed and Forecasted",
    subtitle = "Dashed lines show ARIMA forecasts with 95% confidence intervals",
    x = "Year", y = "Annual Visitors", color = "Data Type"
  ) +
  theme_minimal()
)
```




```
# Save as high-resolution PNG for GitHub
ggsave("forecast_combined.png", plot = final_plot, width = 10, height = 8, dpi = 300)
```

This combined plot highlights each park's historical trend along with the projected future. The use of confidence intervals helps express the uncertainty in these forecasts, particularly for parks with more variability or fewer visitors historically.

Reflection

This mini-project gave me the opportunity to combine personal interest with analytical rigor. Working with real visitation data from the National Park Service allowed me to explore not only historical usage patterns but also apply time series forecasting to generate forward-looking insights. One of the key takeaways was recognizing how different parks exhibit distinct trends—some with rapid growth, others with seasonal or steady visitation—which required flexible modeling approaches.

On a technical level, I practiced data cleaning across multiple files, applied `tidyverse` for structuring, and used ARIMA models to forecast future trends. Visually, I learned how to effectively communicate both observed data and forecast uncertainty. This project showed me how even a single question—like how park attendance changes over time—can be broken down into data, method, visualization, and insight. I'm excited to build on this foundation with more layered analyses in future projects.