Final Project Code and Interpretations

Mia Weathersby

2024-12-03

If you don't want to look at all of the code and visualizations, you can scroll down to the bottom to see a summary of my findings.

```
library(dplyr)
library(ggplot2)
library(readr)
final_sample <- read_csv("final_sample.csv")
```

```
# Selects all rows up to "Property Crime" variable and removes rest
cleaned_data <- final_sample %>%
  select(1:which(names(final_sample) == "Property Crime"))%>%
  filter(if_all(everything(), ~ . >= 0))
dim(final_sample)
```

```
## [1] 28531    33
```

```
#Finding each NA value left and cleaning
colSums(is.na(cleaned_data))
```

```
##                Month                 Year                 Date               Agency
##                    0                    0                    0                    0
##                State          Agency_State               Murder                 Rape
##                    0                    0                    0                    0
##              Robbery    Aggravated Assault             Burglary                Theft
##                    0                    0                    0                    0
## Motor Vehicle Theft         Violent Crime       Property Crime
##                    0                    0                    0
```

```
# (Alternatively) To avoid getting N/A in the "Violent Crime" or
# "Property Crime" columns, you can create a new column that is the sum of what
# is considered violent crime and property crime and after this, you can get
# rid of the original "Violent Crime" and "Property Crime" columns.
# This will help us avoid having to delete whole rows
# just because they have N/A in that category.
# cleaned_data <- final_sample %>% mutate("Total Violent Crime" = Murder + Rape +
#                                         Robbery + `Aggravated Assault`)
# cleaned_data <- cleaned_data %>% mutate("Total Property Crime" = Burglary +
```

```
#                                         Theft + `Motor Vehicle Theft`)
# However, in this case, we will be using the first method of getting rid rows
# with N/A values entirely.


# Moving on to organizing the states by region.
names(cleaned_data$State)
```

```
## NULL
```

```
length(unique(cleaned_data$State))
```

```
## [1] 35
```

```
# There are 39 states (including Nationwide) used in this data

table(cleaned_data$State)
```

```
##
##          AR          AZ          CA          CO          CT          DC          FL
##          79        1261        1487        1264         869          79          79
##          GA          ID          IL          IN          KY          LA          MA
##         156         474        1027         236         158          79        1027
##          MD          MI          MN          MO          MS Nationwide          NC
##          79         237        1501         947          79         395         316
##          NE          NH          NJ          NV          NY          OH          OR
##         316          79         158         315         237        1106         632
##          PA          RI          TN          TX          UT          VA          WA
##        1185         158         237        5054          79         869          79
```

```
# This shows how many times a state shows up in the data.
# This shows that are actually 38 states used in the data since "Nationwide"
# isn't a state.
# We can use this to organize the states into regions

# WEST ~> OR, HI, WA, CA, ID, CO, WY, NV, UT (9)
# SOUTHWEST ~> AZ, TX (2)
# MIDWEST ~> SD, NE, MN, MO, WI, IL, IN, MI, OH (9)
# SOUTHEAST ~> AR, LA, MS, TN, KY, VA, NC, FL, GA (9)
# NORTHEAST ~> PA, MD, DC, NJ, NY, CT, MA, NH, RI (9)
# NATIONWIDE ~> Nationwide (1)

length(unique(cleaned_data$Agency))
```

```
## [1] 276
```

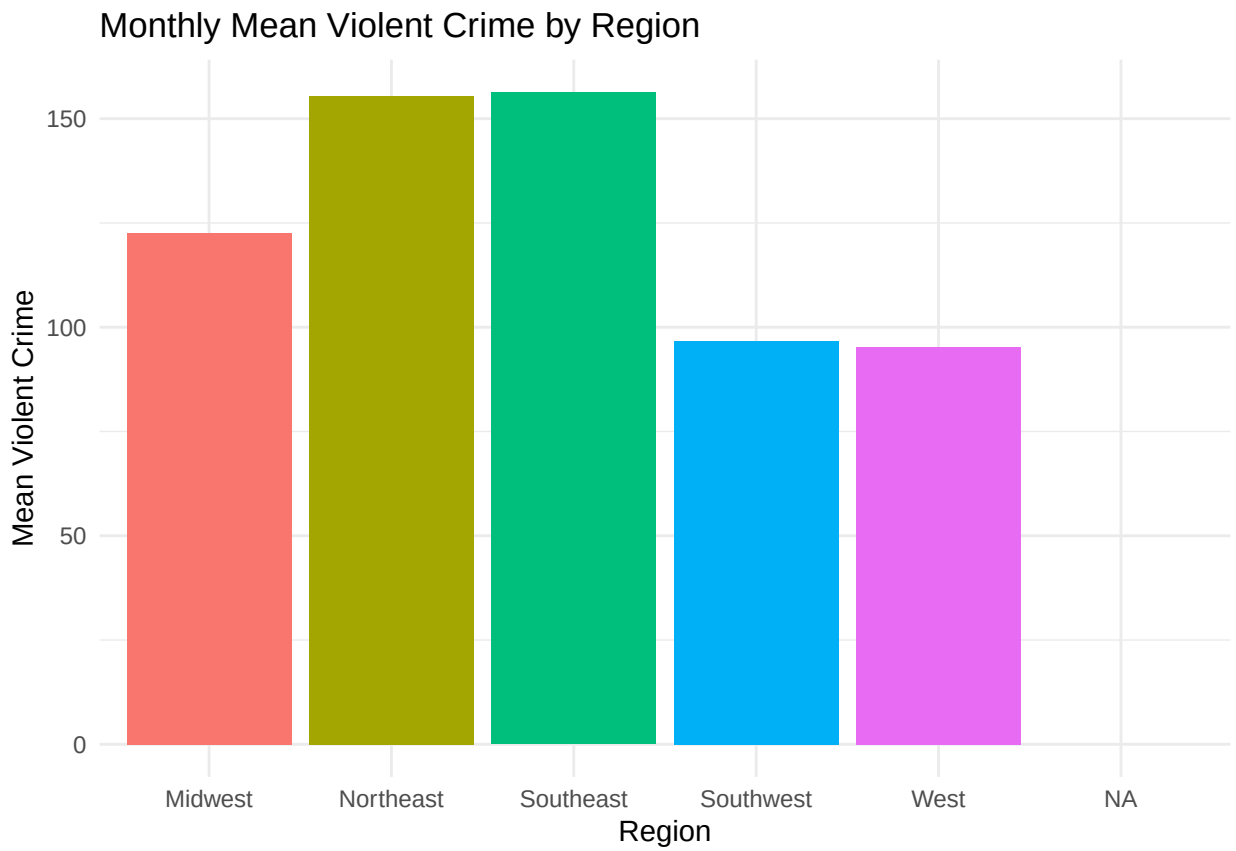
```
# Creates a data frame that matches states to their regions
# (there was probably a faster way to do this but it's whatever)
region_lookup <- data.frame(State = c("OR", "HI", "WA", "CA", "ID", "CO", "WY",
                                      "NV", "UT", "AZ", "TX", "SD", "NE", "MN",
```

```
                                      "MO", "WI", "IL", "IN", "MI", "OH", "AR",
                                      "LA", "MS", "TN", "KY", "VA", "NC", "FL",
                                      "GA", "PA", "MD", "DC", "NJ", "NY", "CT",
                                      "MA", "NH", "RI"),
                            Region = c(rep("West", 9), rep("Southwest", 2),
                                       rep("Midwest", 9), rep("Southeast", 9),
                                       rep("Northeast", 9))
)

# This adds a region column to the data (after you input the code above)
cleaned_data <- cleaned_data %>% left_join(region_lookup, by = "State")

# Shows how many times a region appears in the data set
# Although this does not show how many crimes each region has,
# it can lead us to infer which region has the most crime.
table(cleaned_data$Region)
```

```
##
##   Midwest Northeast Southeast Southwest      West
##      5370      3871      2052      6315      4330
```

```
# For the sake of the data visualizations, we're going to get rid of the
# Nationwide row.
cleaned_data <- cleaned_data[cleaned_data$Region != "N/A", ]

# Average Violent Crime
monthly_mean_violent_crime <- cleaned_data %>%
  group_by(Region) %>%
  summarise(monthly_mean_violent_crime = mean(`Violent Crime`, na.rm = FALSE))
yearly_mean_violent_crime <- cleaned_data %>%
  group_by(Region) %>%
  summarise(yearly_mean_violent_crime = 12*(mean(`Violent Crime`,
                                                 na.rm = FALSE)))
# Violent crime by region data visualizations
ggplot(monthly_mean_violent_crime, aes(x = Region,
                                       y = monthly_mean_violent_crime,
                                       fill = Region)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Monthly Mean Violent Crime by Region",
    x = "Region",
    y = "Mean Violent Crime"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

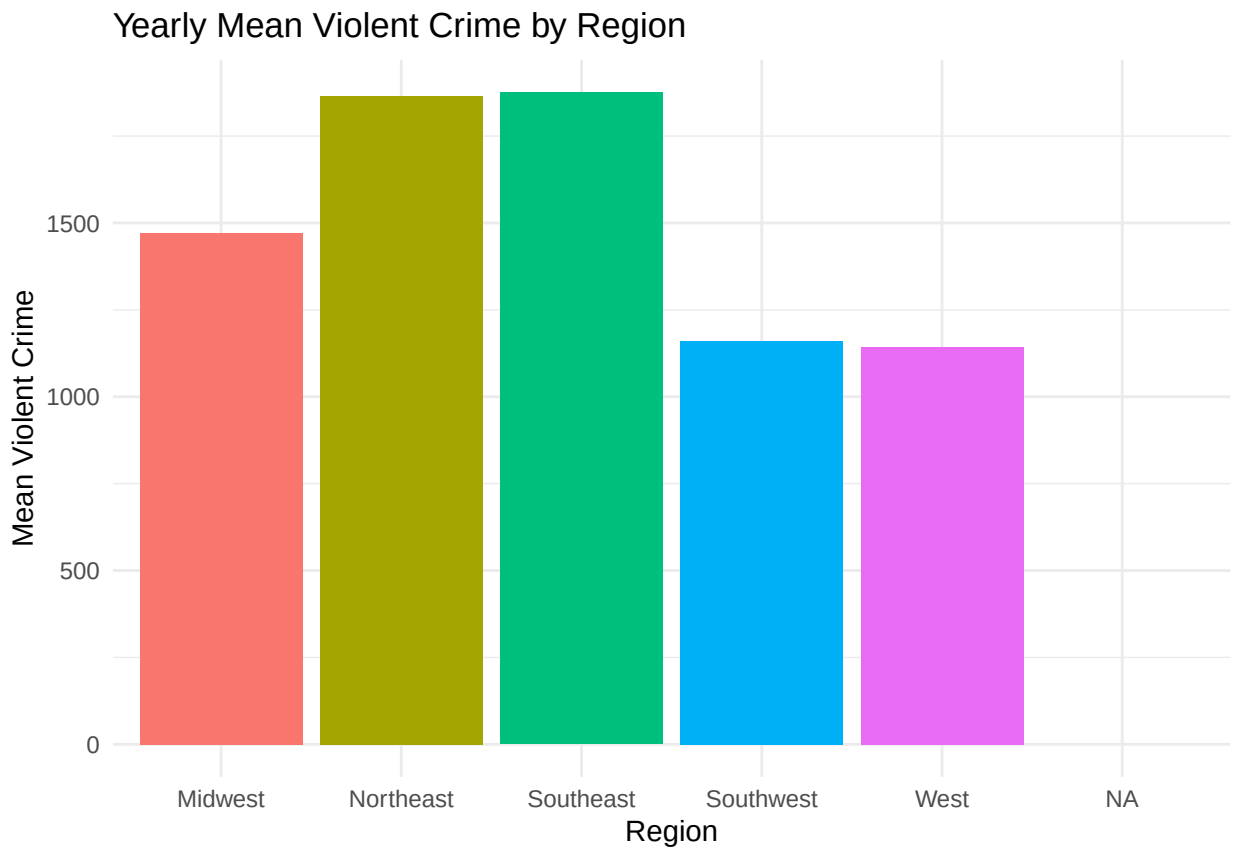
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



- Northeast and Southeast: These regions have the highest average monthly violent crime rates, with the Southeast slightly exceeding the Northeast
- West: The West region has the lowest average monthly violent crime rate among the displayed regions.

```
ggplot(yearly_mean_violent_crime, aes(x = Region, y = yearly_mean_violent_crime,
                                      fill = Region)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Yearly Mean Violent Crime by Region",
    x = "Region",
    y = "Mean Violent Crime"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

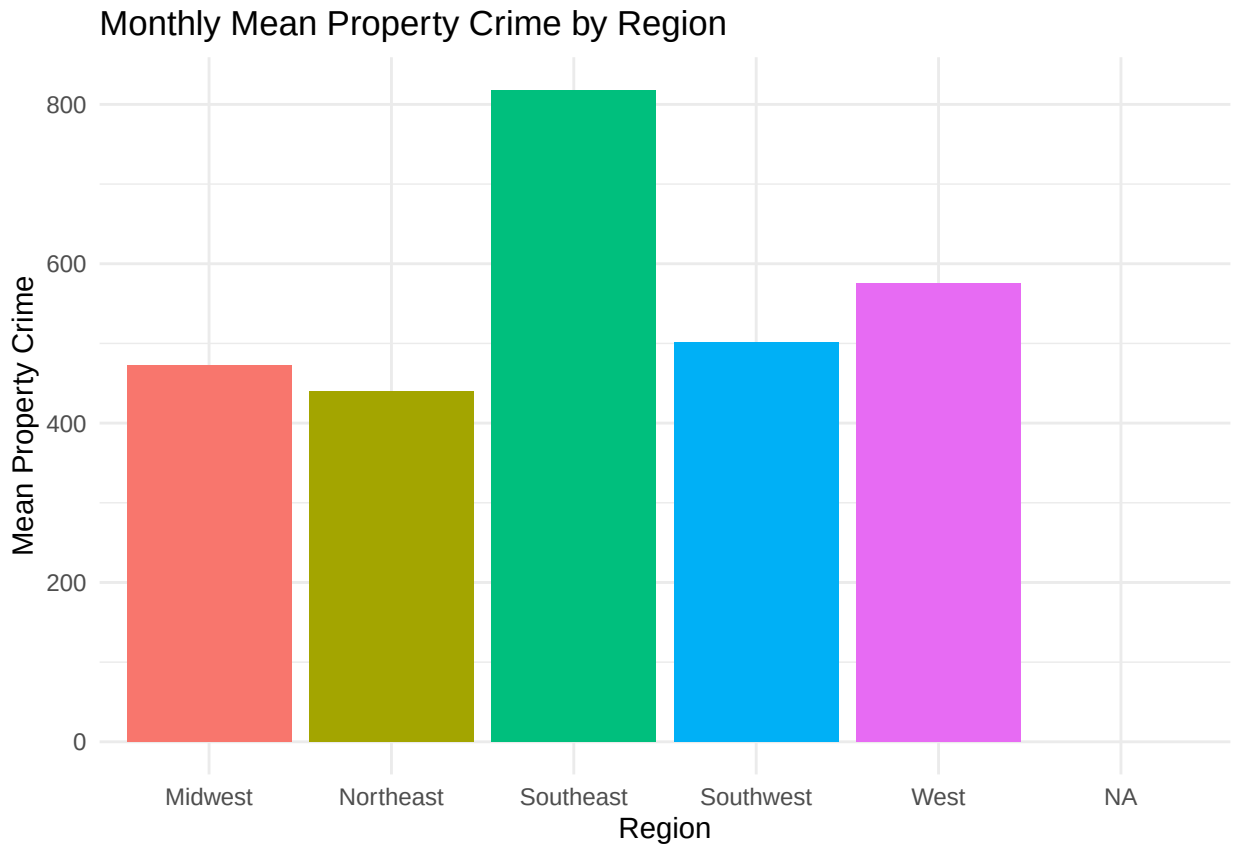
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



- Northeast and Southeast: These regions have the highest average yearly violent crime rates, with the Southeast slightly exceeding the Northeast
- West: The West region has the lowest average yearly violent crime rate among the displayed regions.

```
# Average Property Crime
monthly_mean_property_crime <- cleaned_data %>%
  group_by(Region) %>%
  summarise(monthly_mean_property_crime = mean(`Property Crime`, na.rm = FALSE))
yearly_mean_property_crime <- cleaned_data %>%
  group_by(Region) %>%
  summarise(yearly_mean_property_crime = 12*(mean(`Property Crime`,
                                                  na.rm = FALSE)))
# Property Crime by region data visualization
ggplot(monthly_mean_property_crime, aes(x = Region,
                                        y = monthly_mean_property_crime,
                                        fill = Region)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Monthly Mean Property Crime by Region",
    x = "Region",
    y = "Mean Property Crime"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

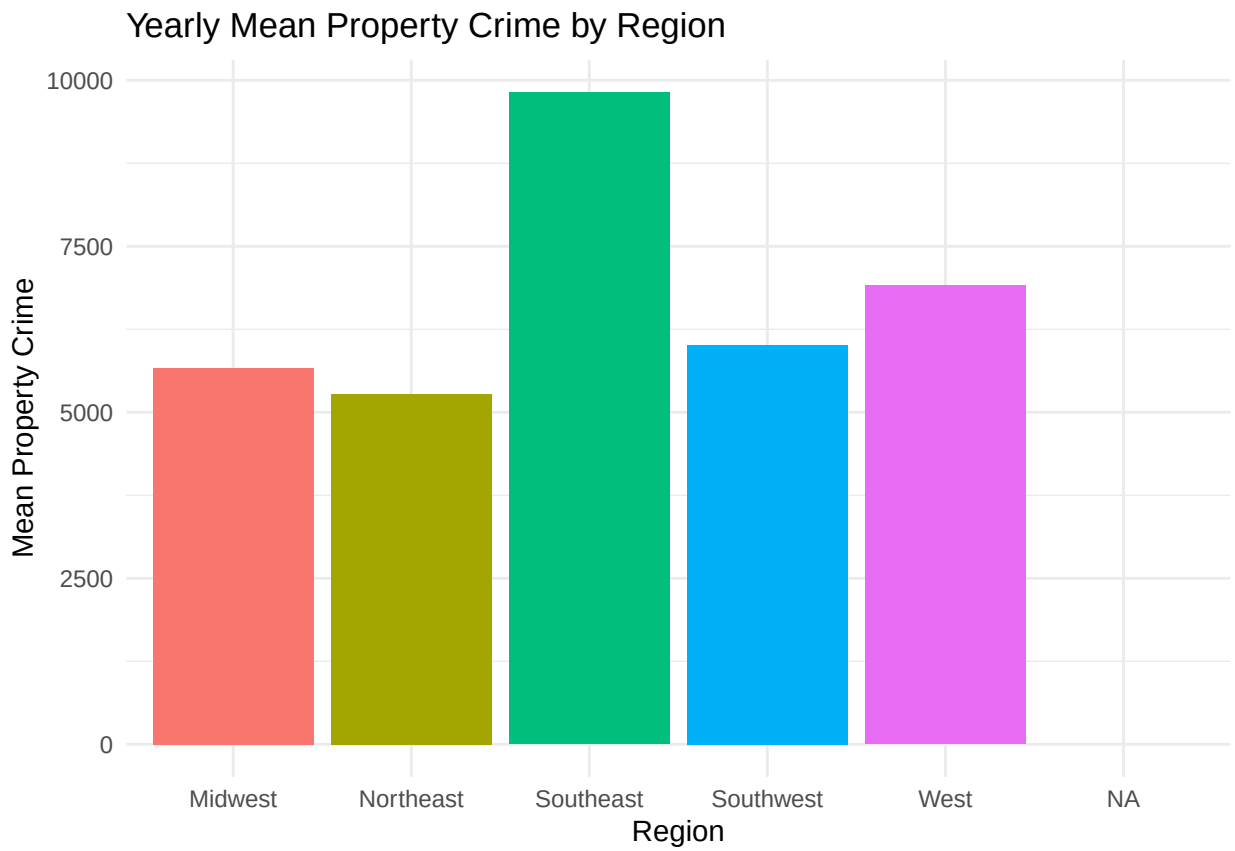
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



- Southeast: This region has the highest average monthly property crime rate.
- Northeast: The Northeast region has the lowest average monthly property crime rate.

```
ggplot(yearly_mean_property_crime, aes(x = Region,
                                       y = yearly_mean_property_crime,
                                       fill = Region)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Yearly Mean Property Crime by Region",
    x = "Region",
    y = "Mean Property Crime"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

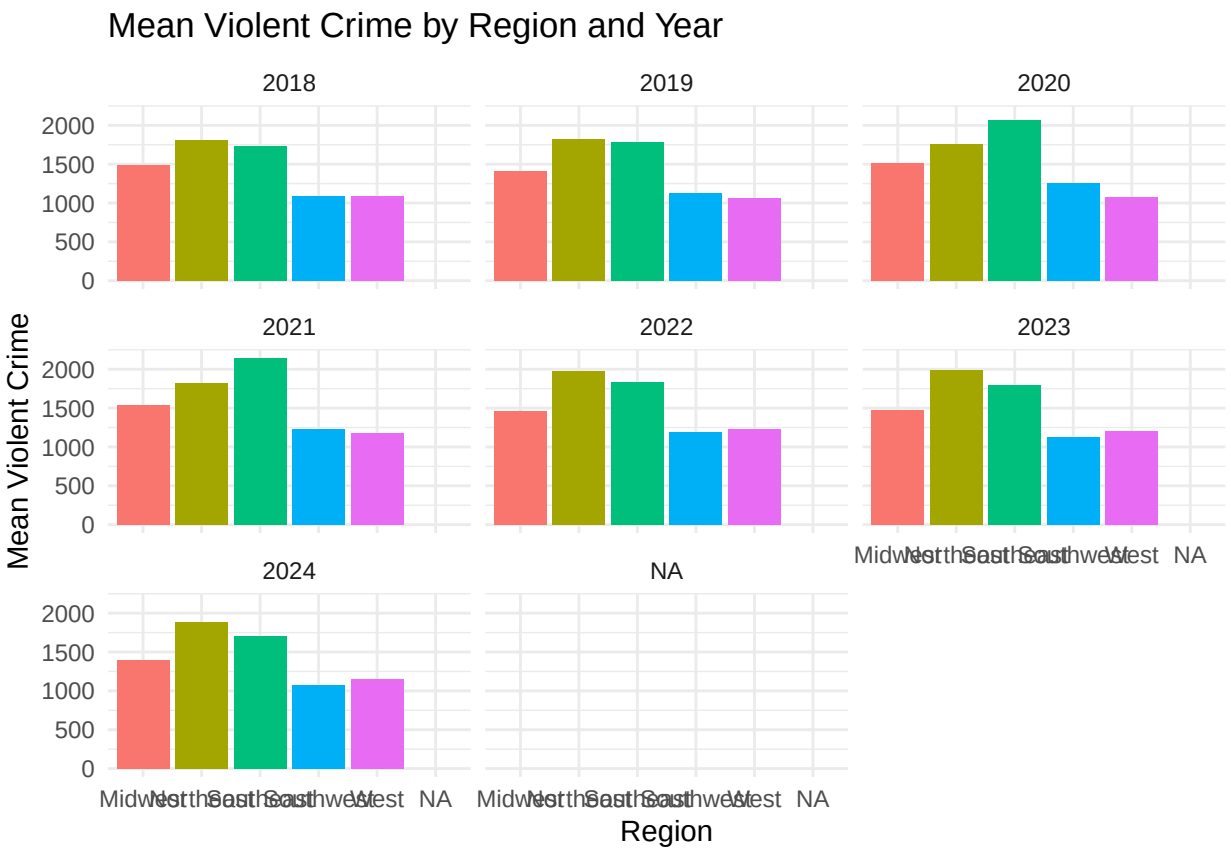
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



- Southeast: This region has the highest average yearly property crime rate.
- Northeast: The Northeast region has the lowest average yearly property crime rate.

```
# Here's how to compare by year
# Violent Crime
mean_violent_crime_by_region_year <- cleaned_data %>%
  group_by(Region, Year) %>%
  summarize(yearly_mean_violent_crime = 12*(mean(`Violent Crime`,
                                                 na.rm = FALSE)))

ggplot(mean_violent_crime_by_region_year, aes(x = Region,
                                              y = yearly_mean_violent_crime,
                                              fill = Region)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Mean Violent Crime by Region and Year",
    x = "Region",
    y = "Mean Violent Crime"
  ) +
  theme_minimal() +
  theme(legend.position = "none") +
  facet_wrap(~ Year)
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```

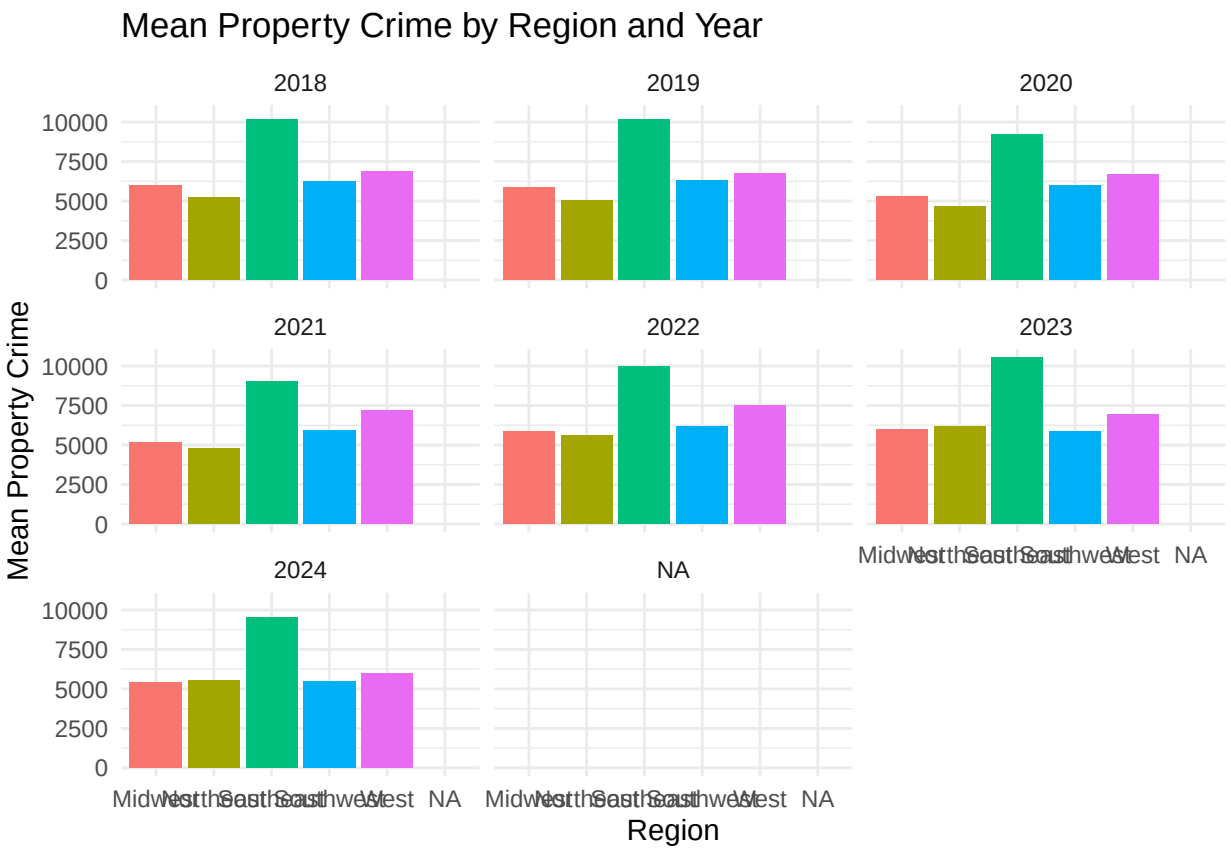


- Northeast and Southeast consistently have the highest mean violent crime rates across all years.
- Midwest, Southwest, and West generally have lower mean violent crime rates compared to the Northeast and Southeast.
- 2018: The Northeast has the highest average and the Southwest and West have the lowest averages.
- 2019: The Northeast has the highest average and the West has the lowest averages.
- 2020: The Southeast has the highest mean violent crime rate and the West has the lowest average.
- 2021: The Southeast has the highest mean violent crime rate and the West has the lowest average.
- 2022: The Northeast has the highest average and the Southwest has the lowest average.
- 2023: The Northeast has the highest average and the Southwest has the lowest average.
- 2024: The Northeast has the highest average and the Southwest has the lowest average.

```
# Property Crime
mean_property_crime_by_region_year <- cleaned_data %>%
  group_by(Region, Year) %>%
  summarize(yearly_mean_property_crime = 12*(mean(`Property Crime`,
                                                  na.rm = FALSE)))

ggplot(mean_property_crime_by_region_year, aes(x = Region,
                                               y = yearly_mean_property_crime,
                                               fill = Region)) +
```

```
  geom_bar(stat = "identity") +
  labs(
    title = "Mean Property Crime by Region and Year",
    x = "Region",
    y = "Mean Property Crime"
  ) +
  theme_minimal() +
  theme(legend.position = "none") +
  facet_wrap(~ Year)
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



- Southeast: This region consistently shows the highest mean property crime rates across all years.
- West: The West region generally has the second-highest mean property crime rates.
- Midwest and Northeast: These regions often have similar and lower mean property crime rates compared to the Southeast and West.
- 2018: The Southeast has the highest mean property crime rate and the Northeast has the lowest mean rate.
- 2019: The Southeast has the highest mean property crime rate and the Northeast has the lowest mean rate.

- 2020: The Southeast has the highest mean property crime rate and the Northeast has the lowest mean rate.
- 2021: The Southeast has the highest mean property crime rate and the Northeast has the lowest mean rate.
- 2022: The Southeast has the highest mean property crime rate and the Northeast has the lowest mean rate.
- 2023: The Southeast has the highest mean property crime rate and the Southwest has the lowest mean rate.
- 2024: The Southeast has the highest mean property crime rate and the Southwest and the Midwest have the lowest mean rates.

Overall Conclusion from Visualizations and Data:

The Southeast region generally has the highest average for most of the years and categories. This region has the highest monthly and yearly average for property and violent crime. However, when looking at each year, the region with the highest mean can vary. In 2018, 2019, 2022, 2023, and 2024, the Northeast has the highest mean for violent crime, but overall the Southeast region has a higher yearly violent crime mean. This is because in the years where the Southeast violent crime mean is the highest (2020 and 2021), the means are significantly higher than the means of the other years. For the property crime though, the Southeast region is the highest for every year. The mean property crime was the highest for the Southeast in the year of 2023 and the lowest in 2020 and 2021.

Since about 40% of the US population lives in the South, in this case the Southeast region, it's more likely the crime will happen there. If we were to do a prediction test to predict where a crime would occur, it would most likely occur in the Southeast region since that's where, on average, both types of crime happens the most.

However, specific different kinds of crime, such as burglary when discussing property crimes, may be more likely to occur in a region different than the Southeast.

https://github.com/miaw06/stats15final