Final Project Code and Interpretations

Mia Weathersby

2024-12-03

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

```
## Rows: 28531 Columns: 33

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): Date, Agency, State, Agency_State, Source.Link, Source.Type, Sourc...
## dbl (22): Month, Year, Murder, Rape, Robbery, Aggravated Assault, Burglary, ...
## lgl  (2): Latitude, Longitude
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# 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.
```

```r
# 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))
```

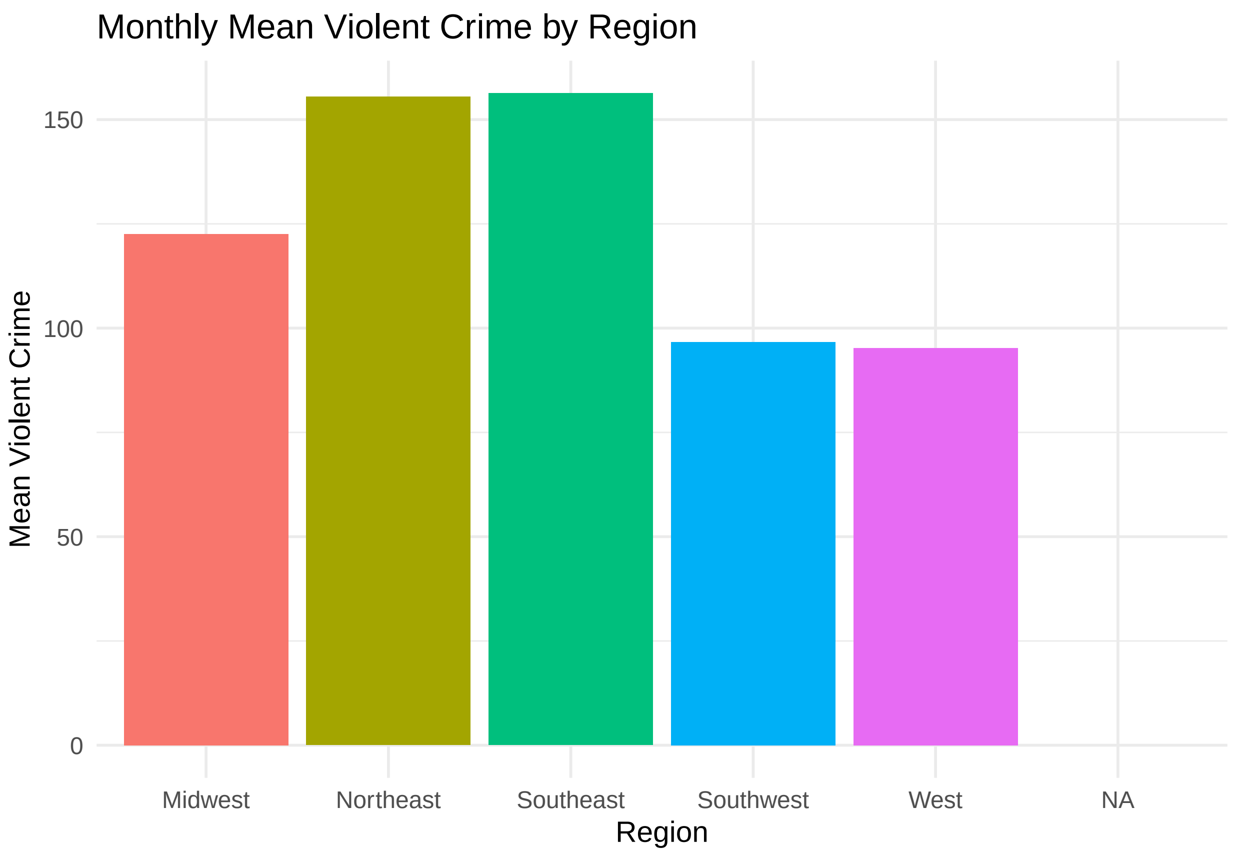
```
## [1] 276
```

```r
# 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
```

```r
# 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")
```

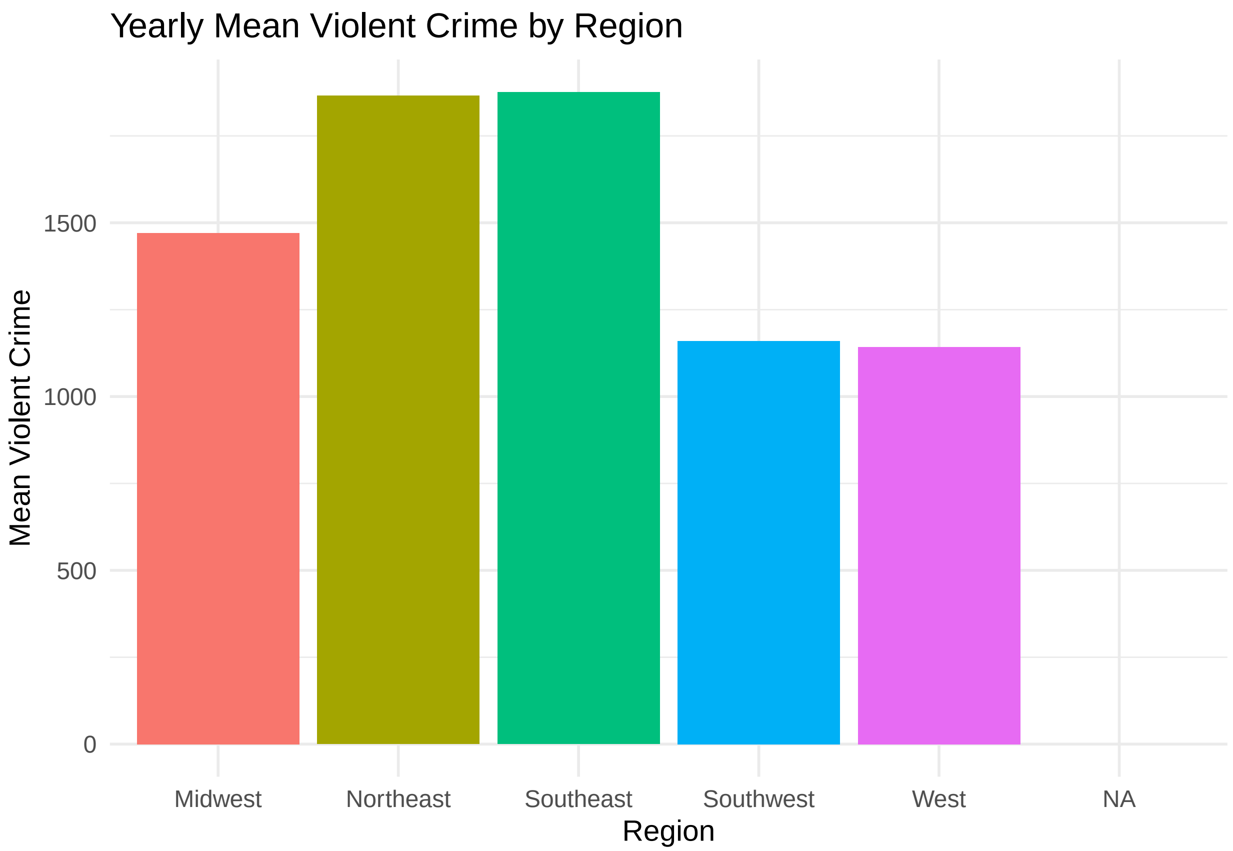
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_bar()‘).
```



- Northeast and Southeast: These regions have the highest average monthly violent crime rates, with the Northeast slightly exceeding the Southeast.
- Midwest and Southwest: These regions have lower average monthly violent crime rates compared to the Northeast and Southeast.
- 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")
```

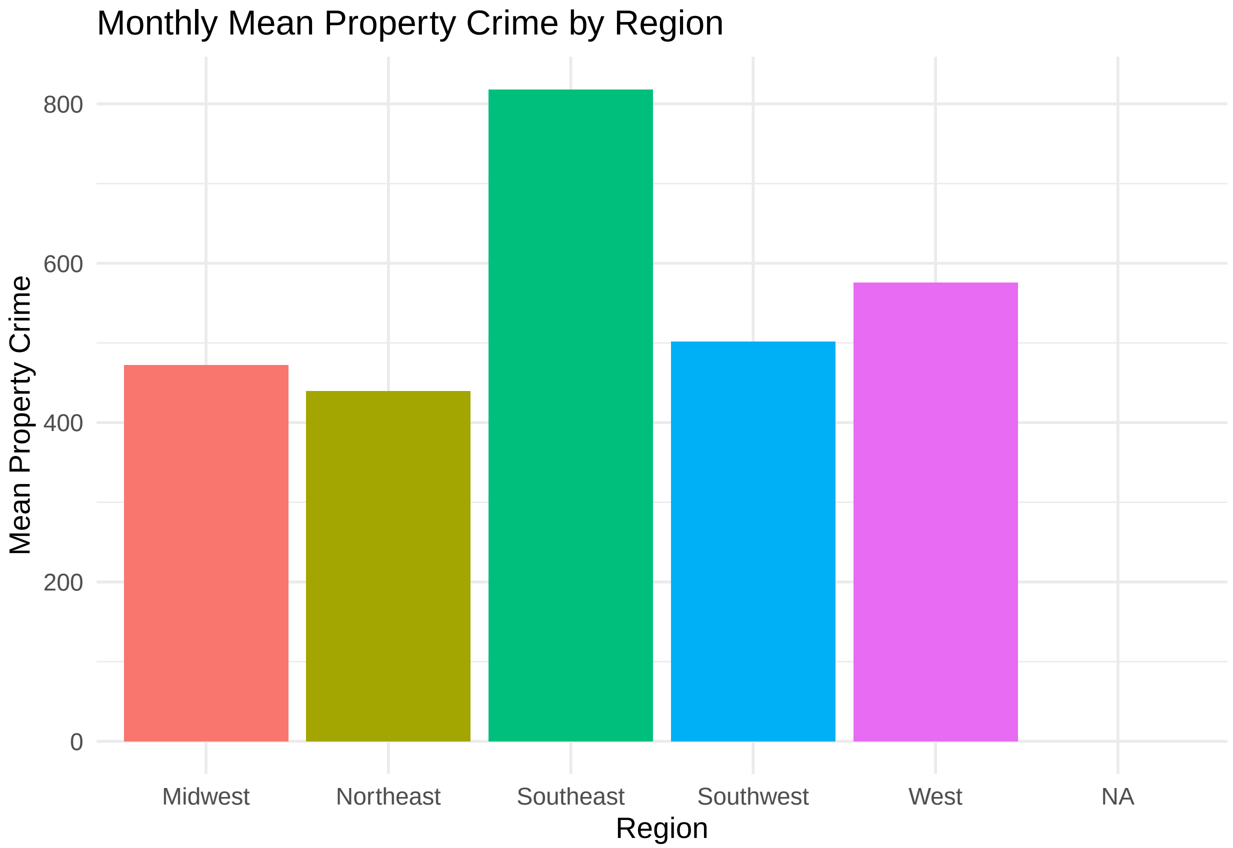
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_bar()‘).
```



- Northeast and Southeast: These regions have the highest average yearly violent crime rates, with the Northeast slightly exceeding the Southeast.
- Midwest and Southwest: These regions have lower average yearly violent crime rates compared to the Northeast and Southeast.
- 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")
```

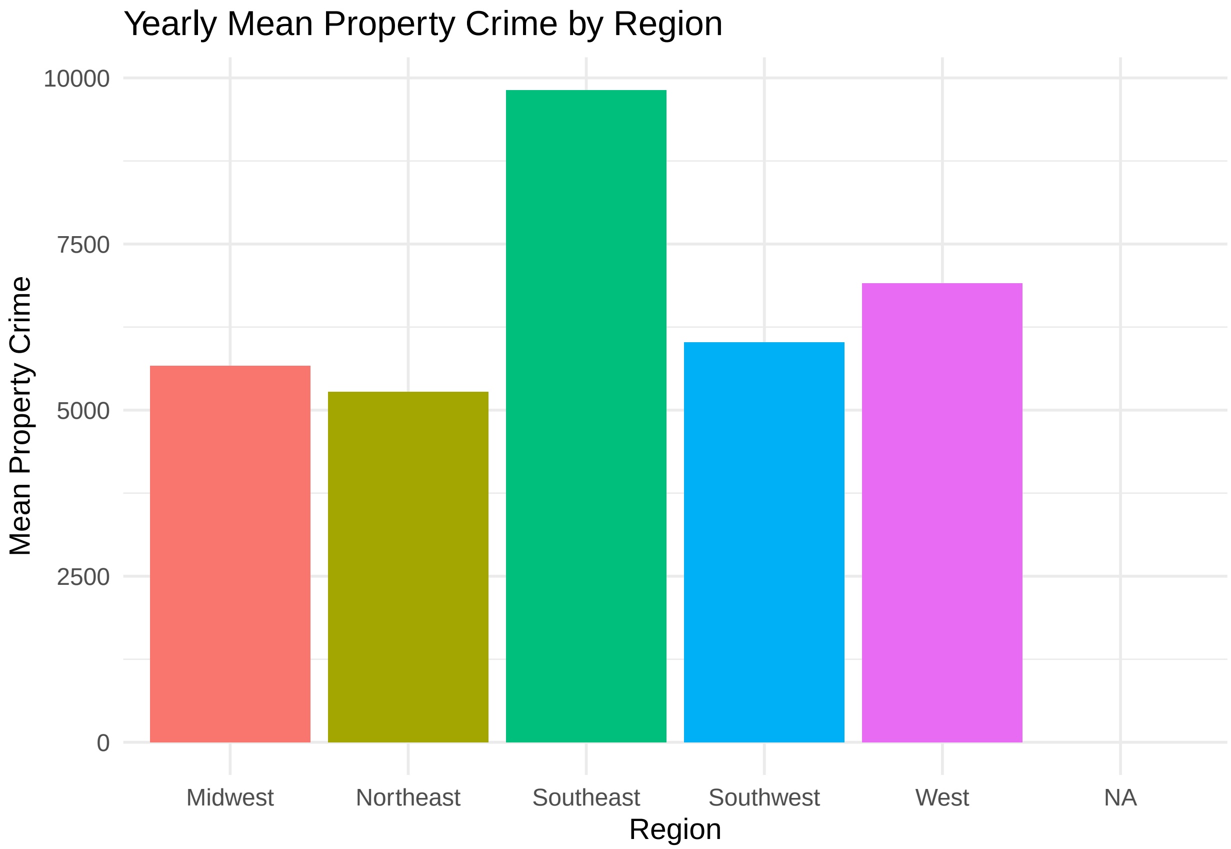
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_bar()‘).
```



- Southeast: This region has the highest average monthly property crime rate.
- West: The West region has the second-highest average monthly property crime rate.
- Midwest and Northeast: These regions have similar and lower average monthly property crime rates compared to the Southeast and West.
- Southwest: The Southwest 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()‘).
```



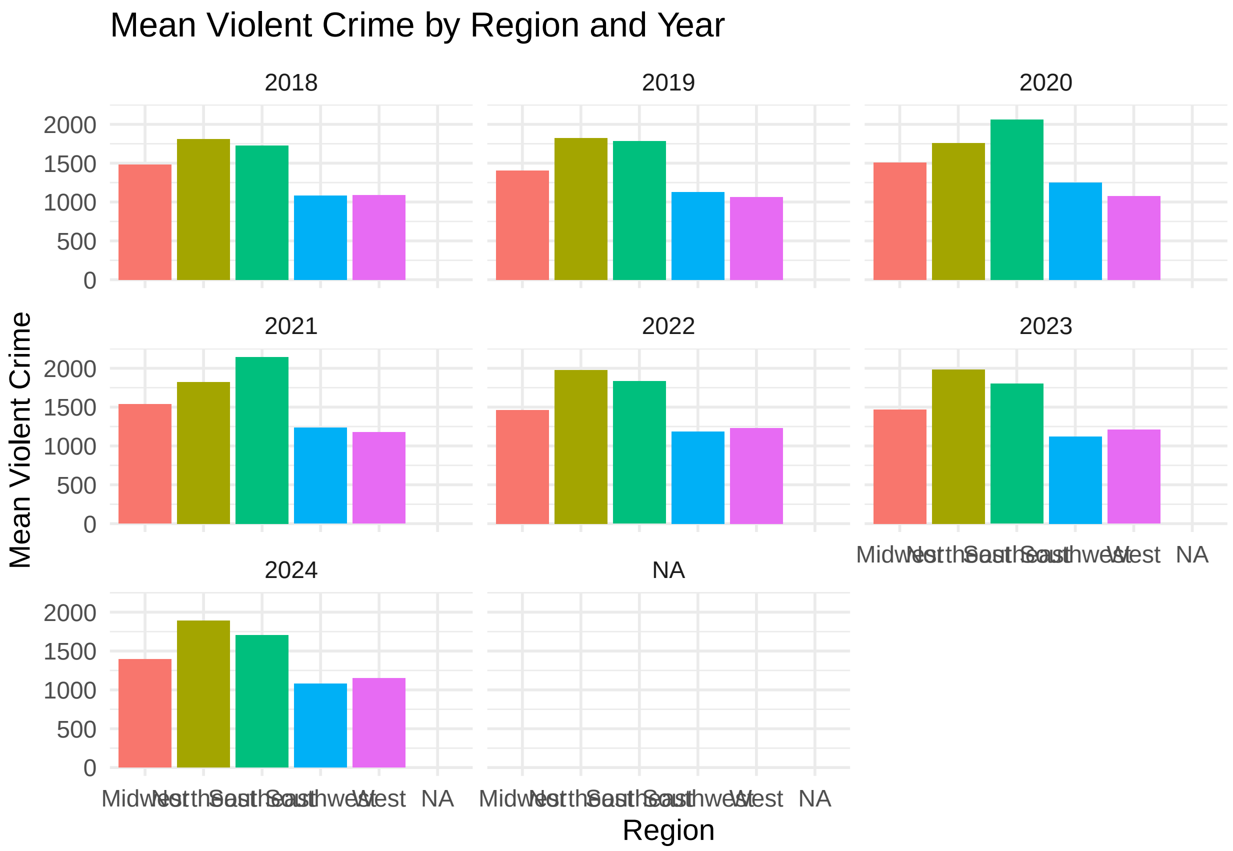
- Southeast: This region has the highest average yearly property crime rate.
- West: The West region has the second-highest average yearly property crime rate.
- Midwest and Northeast: These regions have similar and lower average yearly property crime rates compared to the Southeast and West.
- Southwest: The Southwest 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)))
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
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()`).
```



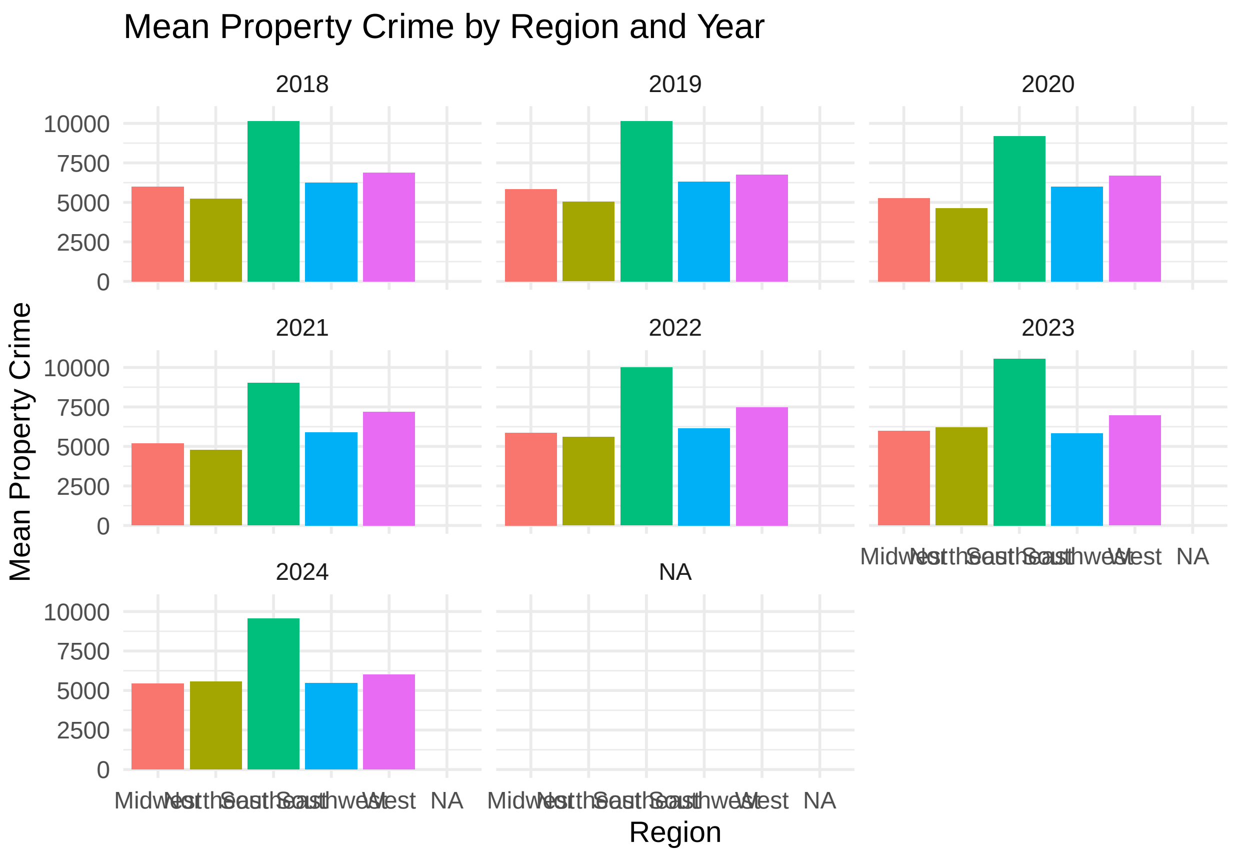
- Northeast and Southeast consistently have the highest mean violent crime rates across all years.
- Midwest, Southwest, and West generally have lower mean violent crime rates compared to the Northeast and Southeast.
- 2018: The Northeast and Southeast have significantly higher mean violent crime rates than other regions.
- 2019: Similar to 2018, the Northeast and Southeast dominate with high mean violent crime rates.
- 2020: The Southeast has the highest mean violent crime rate, followed closely by the Northeast.
- 2021: The Northeast and Southeast remain dominant, with the Southeast slightly edging out the Northeast.
- 2022: The Northeast and Southeast continue to have the highest mean violent crime rates.
- 2023: The Southeast has the highest mean violent crime rate, followed by the Northeast.
- 2024: The Northeast and Southeast still have the highest mean violent crime rates.

```
# 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)))
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
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, followed by the West.
- 2019: The Southeast continues to dominate, with the West again in second place.
- 2020: The Southeast remains the highest, and the West follows closely.
- 2021: The Southeast still leads, with the West in second place.
- 2022: The Southeast maintains its top position, followed by the West.
- 2023: The Southeast continues to have the highest mean property crime rate, with the West in second.
- 2024: The Southeast still leads, with the West in second place.

https://github.com/miaw06/stats15final