MATH 130 FINAL Project

Sharon Mai

2023-09-25

Set Up

```r
knitr::opts_chunk$set(fig.width=6, fig.height=4)
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
```

```r
library(ggplot2)
```

```r
library(readxl)
CrimeData <- read_excel("/Users/Sharon/Desktop/MATH130/Data/Crime_Data.xlsx", sheet=1, col_names=TRUE)
dim(CrimeData)
```

```
## [1] 51 11
```

Introduction

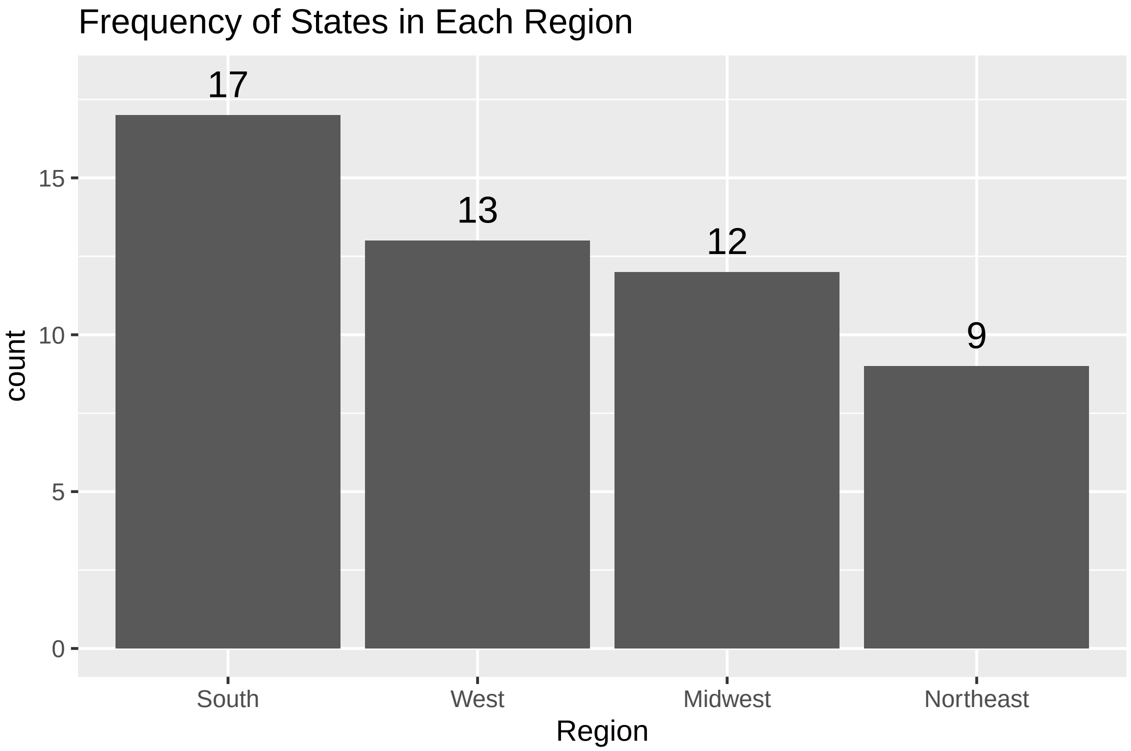
This data set is information of crime and murder rates in different regions in the United States. There are 51 observations of 11 variables in this data set. Using the variables 'crime' and 'region', I want to know how crime related to region. The hypothesis is that there will be more crimes in the "South" than there are in any other region. I expect this to be true because there is a larger population of people living in the South than in other regions which increases the chances of incidence for crime.

Univariable Exploration

Region

```
ggplot (CrimeData, aes(x=forcats::fct_infreq (region))) +
  geom_bar() + xlab('Region')+
  geom_bar(aes(y = ..count..)) + ggtitle("Frequency of States in Each Region") +
  geom_text(aes(y=..count.. + 1, label=..count..), stat='count', size = 5)
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



This bar chart shows how many states are within each region. Using this bar chart, it becomes clear the category ‘South’ has the highest frequency and ‘Northeast’ has the lowest. This continues to support the reasoning behind the hypothesis of the South having a higher crime because the South has the most states who are recording crime incidences.

Crime

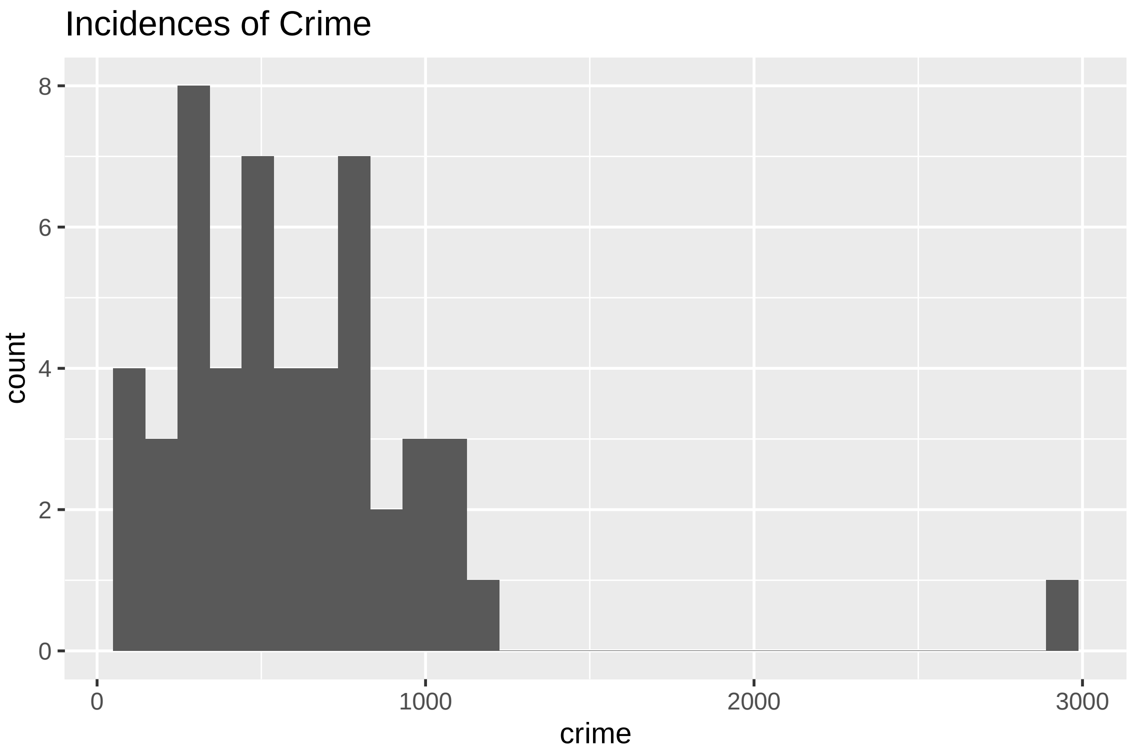
```
summary(CrimeData$crime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    82.0   326.5   515.0   612.8   773.0  2922.0
```

‘crime’ is the number of reported crime incidences per state. The lowest reported crime is 82 and the highest is 2922. However, it is not clear what state or which region the reported crimes are under. This table also shows that the 3rd quartile is 773 reported incidences of crime and the max is 2922 reported incidences of crime. Since the maximum reported crime is so much higher than the 3rd quartile, this may indicate that the maximum recorded crime is an outlier. Since ‘crime’ is a continuous variable, the following histogram can be used to see the individual reports for all reported crime and help identify outliers.

```
ggplot(CrimeData, aes(x=crime)) + geom_histogram() + ggtitle("Incidences of Crime")
```

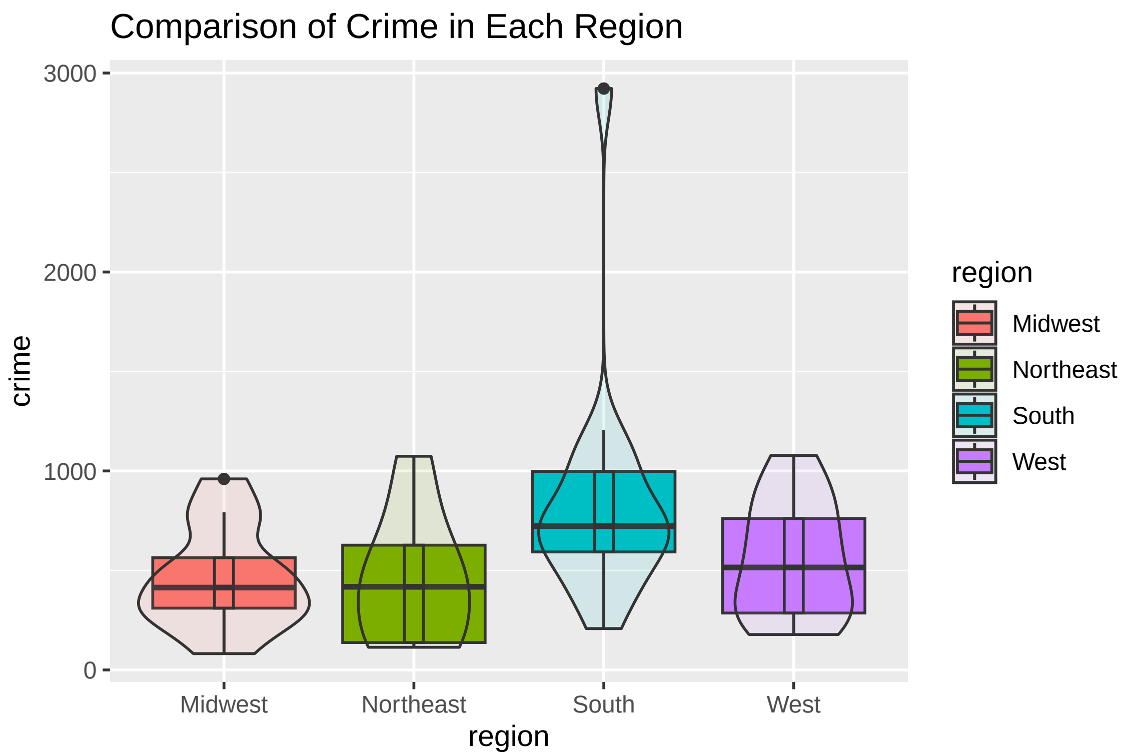
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



In this histogram, we can see the maximum reported crime is an outlier. The majority of reported crime is less than 1500. The average most states have is 700 reports of crime.

Bivariate Exploration

```
ggplot(CrimeData, aes(x=region, y=crime, fill=region)) + geom_boxplot() +
  geom_violin(alpha=.1) +
  geom_boxplot(alpha=.5, width=.1) +
  ggtitle("Comparison of Crime in Each Region")
```



This violin boxplot compares both the 'crime' variable with the 'region' variable. It shows that there are some outliers in the 'Midwest', and 'South' observations and how far they are from the data collected. The violin boxplot also shows where most of the crime distribution is. In all regions, the density is on the lower end of their respective boxplots. This shows that there are more states that have low crime compared to the amount with high crime in the region.

Conclusion

There are more states that fall under the category of 'South' than any other region. The South have more reported crime incidences because there are more states reporting. In the crime variable barchart, only one outlier was identified. However using the barchart did not identify what region the outlier was in. Using the violin boxplots to compare 'crime' against 'region', 2 outliers are identified along with which region they are in. The boxplot shows the median reported crime in the South is higher than all other median. From this data, it can be inferred it is due to the increased amount of states that fall under the 'South' observation compared to the other regions. The boxplot also shows that in the 'Midwest', 'Northeast', and 'West', the density of their crime is in the lower quartiles which means more states have low reported crime compared to the amount of high reported crime. The violin boxplot supports the hypothesis that there is more reported crime in the South.