Exploratory Analysis with Visualization

Piyush Agrawal

Objective: Learn the basics of `ggplot2()’ by visualizing a sample data set

Learning Outcomes: Understand the purpose and effective use of the following tools –line chart, histogram, box plot, scatter plot, bubble chart, and bullet graph.

# Clear the workspace

Let’s start with cleaning the workspace and loading the required packages. Today, we will use a new package, ggplot2(), which is a powerful R graphic package. I highly recommend this site (and book) for more examples: <http://www.cookbook-r.com/Graphs/>

rm(list = ls())  
library(readxl)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2)

# Understanding the dataset

In this exercise, we will use the crime statistics in large U.S. cities for 2009-2014. This dataset is obtained from the Uniform Crime Reports (UCR) published by the FBI (<https://ucr.fbi.gov/crime-in-the-u.s/2014/crime-in-the-u.s.-2014>). This file includes the number of crime occurrences per population (crime rates) and the number of police officers killed or assaulted in the line of duty.

1. Download *Crimes 2009-2014.xlsx* from Canvas *Modules > Week 2*.
2. Open the file in Excel, and browse the data.
3. Take a look at Data Dictionary tab, and understand what each crime data attribute is for.

crimes <- read\_excel("Crimes 2009-2014.xlsx")  
  
crimes

## # A tibble: 1,546 × 10  
## City State Region Year Population `Murder Rate` `Violent Crime Rate`  
## <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Anchorage AK Pacific 12-31-… 257762 8.15 1829.  
## 2 Anchorage AK Pacific 12-31-… 260283 5.38 1714.  
## 3 Anchorage AK Pacific 12-31-… 263588 4.17 1864.  
## 4 Anchorage AK Pacific 12-31-… 267280 7.11 1807.  
## 5 Anchorage AK Pacific 12-31-… 271085 7.01 1688.  
## 6 Anchorage AK Pacific 12-31-… 273714 5.85 1594.  
## 7 Anchorage AK Pacific 12-31-… 276109 6.52 1654.  
## 8 Anchorage AK Pacific 12-31-… 277692 7.20 1913.  
## 9 Anchorage AK Pacific 12-31-… 284142 7.74 1961.  
## 10 Anchorage AK Pacific 12-31-… 280068 4.28 2268.  
## # ℹ 1,536 more rows  
## # ℹ 3 more variables: `Violent Crime Rate in Prior Year` <dbl>,  
## # `Property Crime Rate` <dbl>, `Officer Assault Rate` <dbl>

glimpse(crimes)

## Rows: 1,546  
## Columns: 10  
## $ City <chr> "Anchorage", "Anchorage", "Anchorag…  
## $ State <chr> "AK", "AK", "AK", "AK", "AK", "AK",…  
## $ Region <chr> "Pacific", "Pacific", "Pacific", "P…  
## $ Year <chr> "12-31-1999", "12-31-2000", "12-31-…  
## $ Population <dbl> 257762, 260283, 263588, 267280, 271…  
## $ `Murder Rate` <dbl> 8.147050, 5.378761, 4.173179, 7.108…  
## $ `Violent Crime Rate` <dbl> 1829.207, 1714.288, 1863.514, 1807.…  
## $ `Violent Crime Rate in Prior Year` <dbl> NA, 1829.207, 1714.288, 1863.514, 1…  
## $ `Property Crime Rate` <dbl> 4370.311, 4357.565, 4349.970, 4470.…  
## $ `Officer Assault Rate` <dbl> 30.64843, 24.58862, 32.62667, 31.80…

# Data Preparation

The names of variables contain spaces, which make it difficult to manage in R. Let’s rename variables.

# Change the variable names  
colnames(crimes)<-c("city","state","region","date","population","murder\_rate","violent\_crime\_rate","violent\_crime\_rate\_pr","property\_crime\_rate","officer\_assault\_rate")  
glimpse(crimes)

## Rows: 1,546  
## Columns: 10  
## $ city <chr> "Anchorage", "Anchorage", "Anchorage", "Anchorag…  
## $ state <chr> "AK", "AK", "AK", "AK", "AK", "AK", "AK", "AK", …  
## $ region <chr> "Pacific", "Pacific", "Pacific", "Pacific", "Pac…  
## $ date <chr> "12-31-1999", "12-31-2000", "12-31-2001", "12-31…  
## $ population <dbl> 257762, 260283, 263588, 267280, 271085, 273714, …  
## $ murder\_rate <dbl> 8.147050, 5.378761, 4.173179, 7.108650, 7.008872…  
## $ violent\_crime\_rate <dbl> 1829.207, 1714.288, 1863.514, 1807.468, 1687.663…  
## $ violent\_crime\_rate\_pr <dbl> NA, 1829.207, 1714.288, 1863.514, 1807.468, 1687…  
## $ property\_crime\_rate <dbl> 4370.311, 4357.565, 4349.970, 4470.593, 4318.203…  
## $ officer\_assault\_rate <dbl> 30.64843, 24.58862, 32.62667, 31.80186, 30.98659…

Also, change the variable type of date as a date variable and create year for further analysis.

# create year variable  
crimes$date<-as.Date(crimes$date,"%m-%d-%Y")  
crimes

## # A tibble: 1,546 × 10  
## city state region date population murder\_rate violent\_crime\_rate  
## <chr> <chr> <chr> <date> <dbl> <dbl> <dbl>  
## 1 Anchorage AK Pacific 1999-12-31 257762 8.15 1829.  
## 2 Anchorage AK Pacific 2000-12-31 260283 5.38 1714.  
## 3 Anchorage AK Pacific 2001-12-31 263588 4.17 1864.  
## 4 Anchorage AK Pacific 2002-12-31 267280 7.11 1807.  
## 5 Anchorage AK Pacific 2003-12-31 271085 7.01 1688.  
## 6 Anchorage AK Pacific 2004-12-31 273714 5.85 1594.  
## 7 Anchorage AK Pacific 2005-12-31 276109 6.52 1654.  
## 8 Anchorage AK Pacific 2006-12-31 277692 7.20 1913.  
## 9 Anchorage AK Pacific 2007-12-31 284142 7.74 1961.  
## 10 Anchorage AK Pacific 2008-12-31 280068 4.28 2268.  
## # ℹ 1,536 more rows  
## # ℹ 3 more variables: violent\_crime\_rate\_pr <dbl>, property\_crime\_rate <dbl>,  
## # officer\_assault\_rate <dbl>

crimes$year<-format(crimes$date,"%Y")  
glimpse(crimes)

## Rows: 1,546  
## Columns: 11  
## $ city <chr> "Anchorage", "Anchorage", "Anchorage", "Anchorag…  
## $ state <chr> "AK", "AK", "AK", "AK", "AK", "AK", "AK", "AK", …  
## $ region <chr> "Pacific", "Pacific", "Pacific", "Pacific", "Pac…  
## $ date <date> 1999-12-31, 2000-12-31, 2001-12-31, 2002-12-31,…  
## $ population <dbl> 257762, 260283, 263588, 267280, 271085, 273714, …  
## $ murder\_rate <dbl> 8.147050, 5.378761, 4.173179, 7.108650, 7.008872…  
## $ violent\_crime\_rate <dbl> 1829.207, 1714.288, 1863.514, 1807.468, 1687.663…  
## $ violent\_crime\_rate\_pr <dbl> NA, 1829.207, 1714.288, 1863.514, 1807.468, 1687…  
## $ property\_crime\_rate <dbl> 4370.311, 4357.565, 4349.970, 4470.593, 4318.203…  
## $ officer\_assault\_rate <dbl> 30.64843, 24.58862, 32.62667, 31.80186, 30.98659…  
## $ year <chr> "1999", "2000", "2001", "2002", "2003", "2004", …

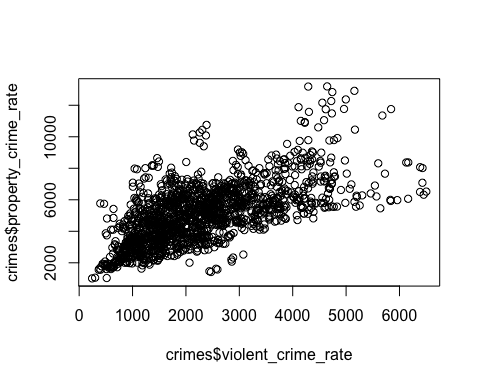
# Scatter plot

Let’s start by examining the relationship between the violent crime rate and the property crime rate in the major cities. You can visualize the relationship of two numeric variables by plotting a scatter plot where x- and y-axis represents each numeric variable.

## Base R plot function

Using base R plot function:

plot(crimes$violent\_crime\_rate, crimes$property\_crime\_rate)



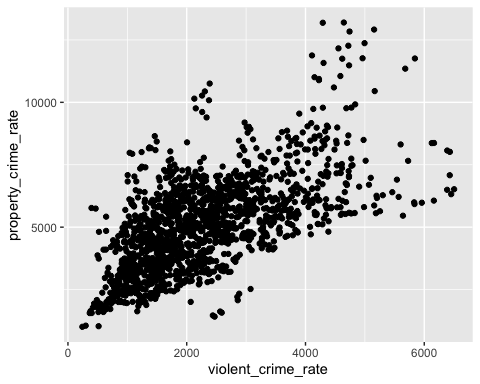
However, the combination of tidyverse and ggplot2 can help you create a nicer-looking chart in a more intrinsic way with more flexibility.

## ggplot: Intro

ggplot() takes a dataset as the first argument, and you specify aesthetics (e.g. axes, color, type, etc) with the parameters in aes().

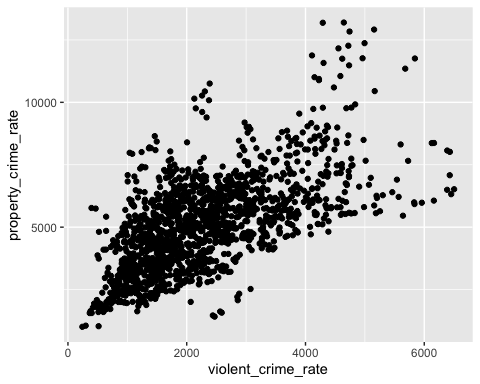
In ggplot(), you specify the dataset and aesthetics, and `geom\_point() indicates you’d like to create a scatter plot.

ggplot(crimes, aes(x=violent\_crime\_rate,y=property\_crime\_rate))+  
 geom\_point()



We can create the same plot with a pipe operator:

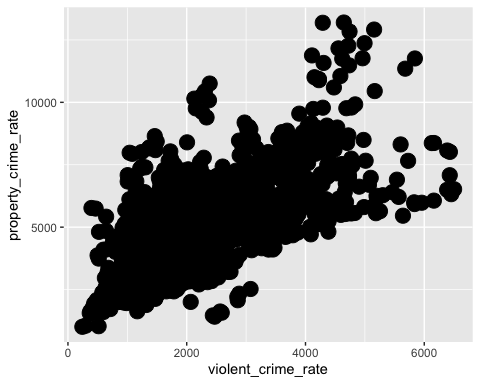
crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate))+  
 geom\_point()



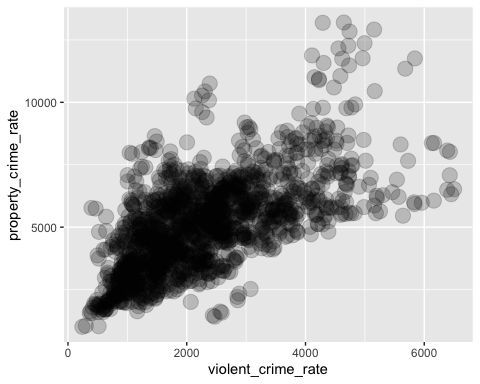
## Change colors, data point sizes, etc

If you want to apply the different color, shape, etc, for *all* points, you can do it by changing the parameter in `geom\_point()’.

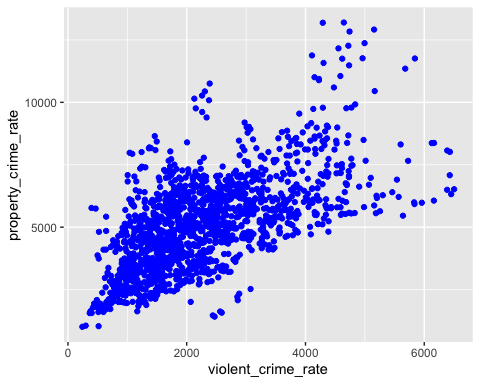
crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate))+geom\_point(size=5)



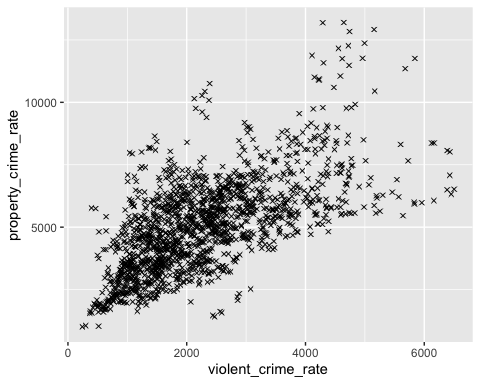
crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate))+geom\_point(size=5, alpha=0.2) #alpha for transparency: for densely populated data points



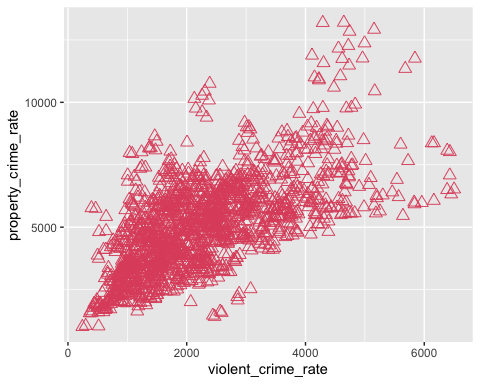
crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate))+geom\_point(col="blue")



crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate))+geom\_point(shape=4)



crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate))+geom\_point(col=2,shape=2,size =3)



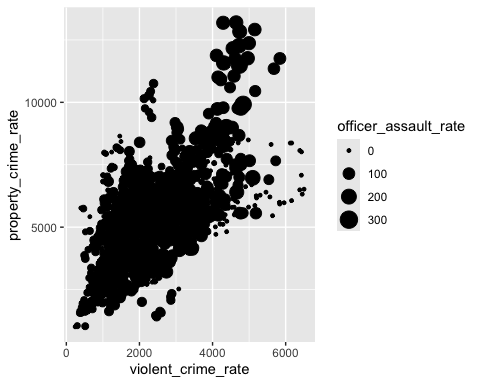
crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate))+  
 geom\_point(col="#86229A",shape=2,size =3)+  
 labs(title="Violent vs. Property Crime Rate", x="Violent Crime Rate", y="Property Crime Rate")



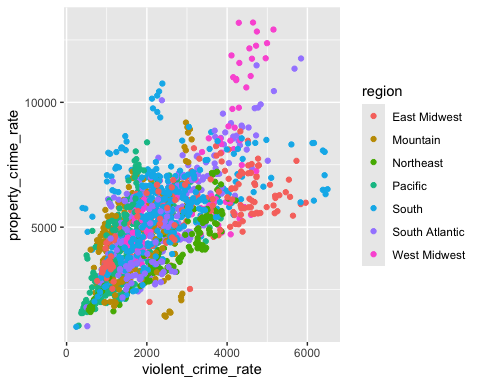
## Different colors, sizes, etc. by groups

If you want to use different colors by different groups, you can do it easily by setting it in aes()! You can change the aesthetics to deliver more information on a chart.

crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate, size=officer\_assault\_rate))+geom\_point()

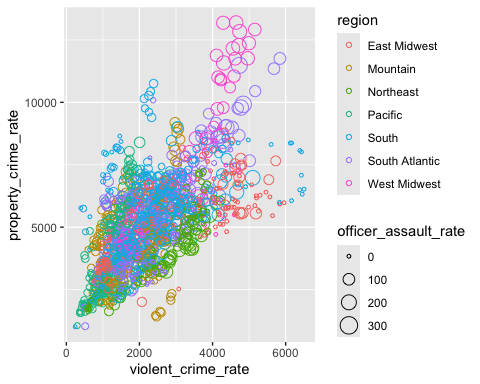


crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate, col=region))+geom\_point()



You can do it in combination!

crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate,size=officer\_assault\_rate,col=region))+  
 geom\_point(shape=1)

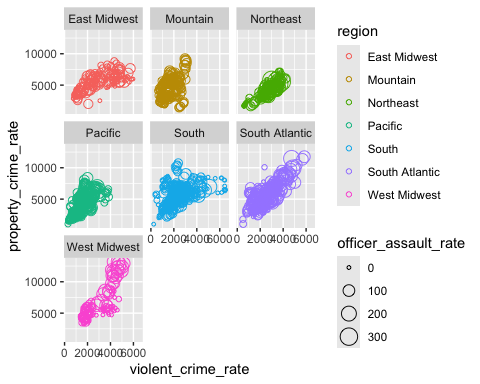


Note that Officer Assault Rate is the number of police officers killed or assaulted in the line of duty. This bubble chart now shows that police officers are in more danger (as shown in bigger circles) in cities with higher crime rates.

## Facet

You can also easily create separate charts by subgroups (i.e. regions in this case) using facet\_wrap(), which creates charts, which share the same axes, and displays them in a panel.

crimes%>%  
 ggplot(aes(y=property\_crime\_rate,x=violent\_crime\_rate,size=officer\_assault\_rate,colour=region))+  
 geom\_point(shape=1)+  
 facet\_wrap(~region)



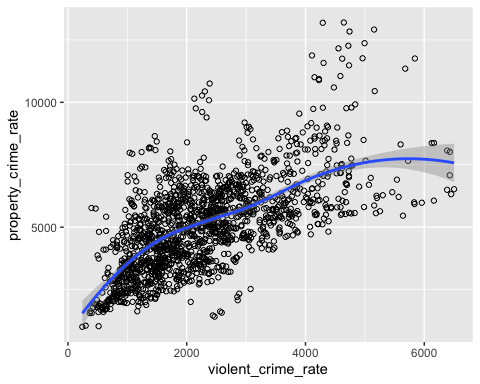
## Add a statistics layer

We can also add some statistical analysis results to the charts.

LOESS, which is a default smoothing method of geom\_smooth(), is a non-parametric form of regression that uses a weighted, sliding-window, average to calculate a line of best fit.

crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate))+  
 geom\_point(shape=1)+  
 geom\_smooth(method=loess)

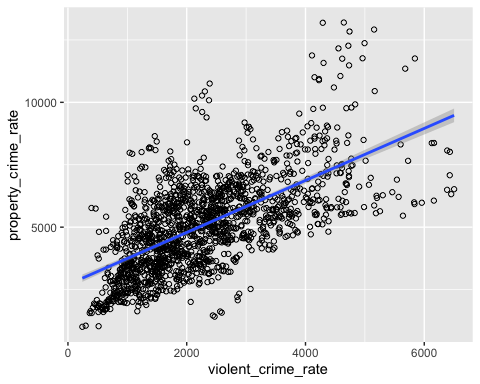
## `geom\_smooth()` using formula = 'y ~ x'



A linear regression model can be added into a scatter plot by adding specifying the smoothing method as “lm”.

crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate))+  
 geom\_point(shape=1)+  
 geom\_smooth(method=lm)

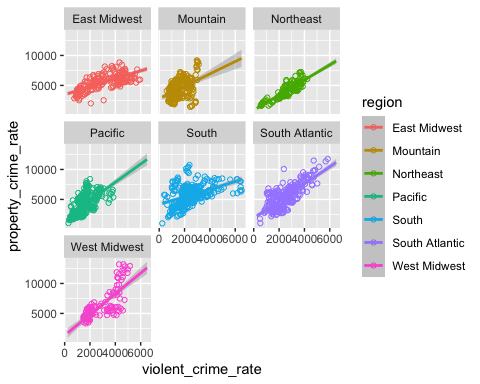
## `geom\_smooth()` using formula = 'y ~ x'



Using facet\_wrap(), you can compare the trend by different regions in the United States.fullrange=TRUE extends the span of the linear model to the full range.

crimes%>%  
 ggplot(aes(x=violent\_crime\_rate,y=property\_crime\_rate, col=region))+  
 geom\_point(shape=1)+  
 geom\_smooth(method=lm,fullrange=TRUE)+  
 facet\_wrap(~region)

## `geom\_smooth()` using formula = 'y ~ x'



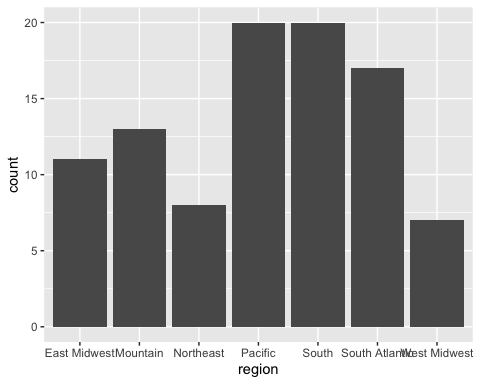
# Bar charts

A bar chart is a simple way to depict the frequencies of the values of a *categorical* attribute.

How many cities are contained in this dataset by region? In fact, the dataset is a panel dataset (i.e. repeated observations over time for the same entities) at a city-year level, where each city appears once per year. So, we can count the number of cities, if we set year to a specific year (e.g. year=2000), and count the number of observations.

Let’s filter data with year==2000 and then call ggplot() with geom\_bar() to create a bar chart.

crimes%>%  
 filter(year==2000)%>%  
 ggplot(aes(x=region))+  
 geom\_bar()



# Histogram

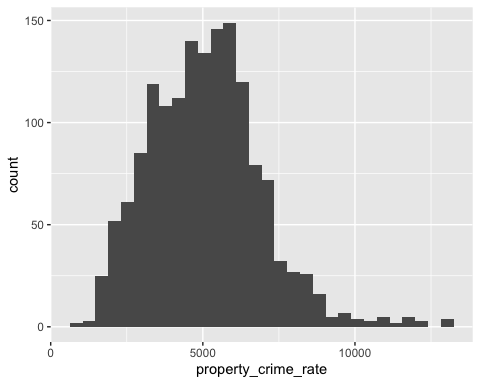
A histogram shows the frequency distribution for a numerical attribute.

The range of a numerical attribute is discretized into a fixed number of intervals (“bins”), usually of equal length. For each interval, the (absolute) frequency of values falling into each interval is indicated by the height of a bar.

Let’s examine the distribution of the property crime rate.

crimes%>%  
 ggplot(aes(x=property\_crime\_rate))+  
 geom\_histogram()

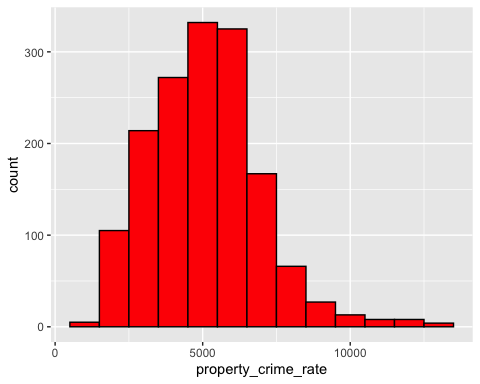
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## Change bin width, fill, color..

Try changing the value of binwidth. You will see the shape of the histogram changes. Also, as we did with a scatterplot, you can change the color of the histogram.

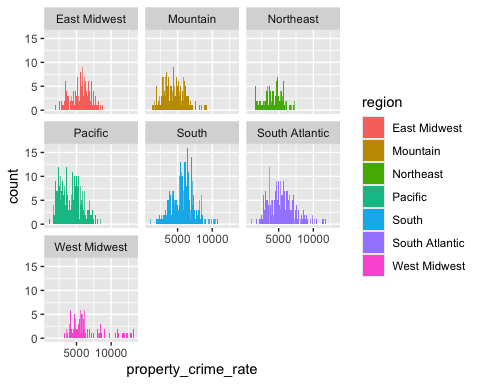
crimes%>%  
 ggplot(aes(x=property\_crime\_rate))+  
 geom\_histogram(binwidth=1000, col="black",fill="red")



## Overlapping histograms

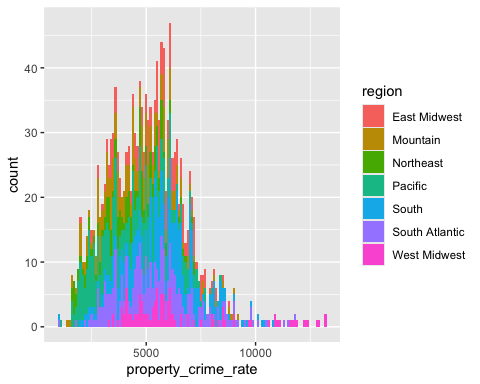
We can easily create histograms for different regions using facet\_wrap().

crimes%>%  
 ggplot(aes(x=property\_crime\_rate, fill=region))+  
 geom\_histogram(binwidth=100)+  
 facet\_wrap(~region)



But you can also display these histograms in one chart with various options.

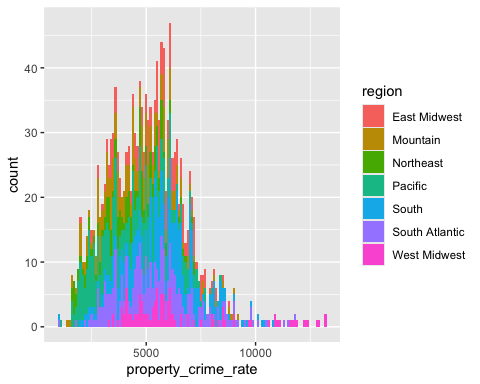
crimes%>%  
 ggplot(aes(x=property\_crime\_rate,fill=region))+  
 geom\_histogram(binwidth=100)



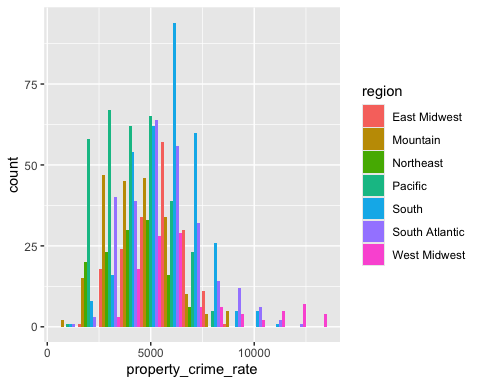
crimes%>%  
 ggplot(aes(x=property\_crime\_rate,fill=region))+  
 geom\_histogram(binwidth=100,alpha=0.3, position="identity")



crimes%>%  
 ggplot(aes(x=property\_crime\_rate,fill=region))+  
 geom\_histogram(binwidth=100,position="stack")

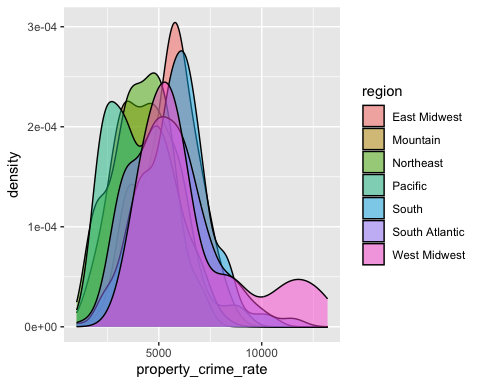


crimes%>%  
 ggplot(aes(x=property\_crime\_rate,fill=region))+  
 geom\_histogram(binwidth=1000,position="dodge")



Instead of a histogram, you can create a density function in a similar manner.

crimes%>%  
 ggplot(aes(x=property\_crime\_rate, fill=region))+  
 geom\_density(alpha=0.5)

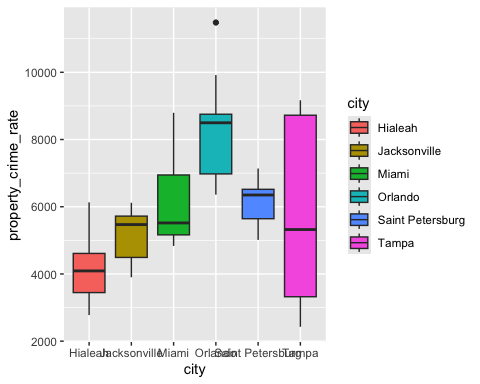


# Box plot

A boxplot is a very compact way to visualize and summarize the key statistics of a numeric attribute.

Let’s examine the distribution of the property crime rate (which will be displayed in y-axis) by cities (which will be x-axis). Because we have too many cities in our datasets, let’s focus on the cities in Florida.

crimes%>%  
 filter(state=="FL")%>%  
 ggplot(aes(x=city,y=property\_crime\_rate,fill=city))+  
 geom\_boxplot()



The upper whisker extends from the hinge to the largest value no further than 1.5 \* IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 \* IQR of the hinge. Data beyond the end of the whiskers are called “outlying” points and are plotted individually.

# Time Series with Line Graph

Let’s examine the trend of the violent crime rate of the cities in New York state. Let’s first examine the data.

crimes%>%  
 filter(state=="NY")%>%  
 distinct(city)

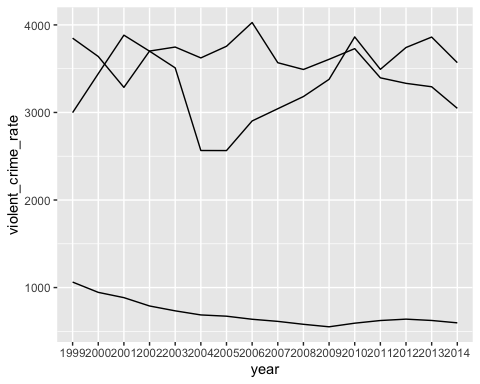
## # A tibble: 3 × 1  
## city   
## <chr>   
## 1 Buffalo   
## 2 New York   
## 3 Rochester

crimes%>%  
 filter(state=="NY")

## # A tibble: 48 × 11  
## city state region date population murder\_rate violent\_crime\_rate  
## <chr> <chr> <chr> <date> <dbl> <dbl> <dbl>  
## 1 Buffalo NY Northeast 1999-12-31 301069 10.6 2998.  
## 2 Buffalo NY Northeast 2000-12-31 292648 14.0 3443.  
## 3 Buffalo NY Northeast 2001-12-31 293187 22.5 3883.  
## 4 Buffalo NY Northeast 2002-12-31 295441 14.9 3700.  
## 5 Buffalo NY Northeast 2003-12-31 288187 22.9 3509.  
## 6 Buffalo NY Northeast 2004-12-31 285567 17.9 2565.  
## 7 Buffalo NY Northeast 2005-12-31 283269 20.1 2565.  
## 8 Buffalo NY Northeast 2006-12-31 280494 26.4 2903.  
## 9 Buffalo NY Northeast 2007-12-31 273832 20.1 3043.  
## 10 Buffalo NY Northeast 2008-12-31 270289 13.7 3182.  
## # ℹ 38 more rows  
## # ℹ 4 more variables: violent\_crime\_rate\_pr <dbl>, property\_crime\_rate <dbl>,  
## # officer\_assault\_rate <dbl>, year <chr>

For line graphs, the data points must be grouped so that it knows which points to connect. If you want to connect all points, set group=1. Because we want to connect the points by cities, we will set group=city.

crimes%>%  
 filter(state=="NY")%>%  
 ggplot(aes(x=year,y=violent\_crime\_rate,group=city))+geom\_line()

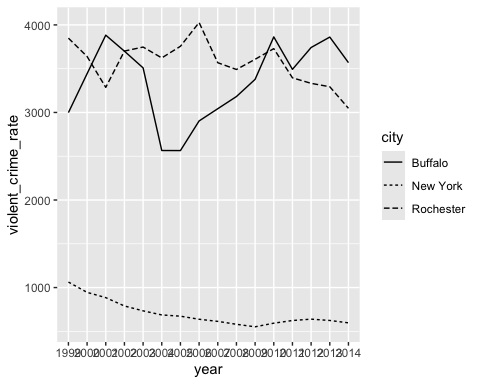


But we lost the information about which line is for which city. Let’s use different types or colors of lines, so that we can distinguish them.

crimes%>%  
 filter(state=="NY")%>%  
 ggplot(aes(x=year,y=violent\_crime\_rate, group=city,col=city))+geom\_line()



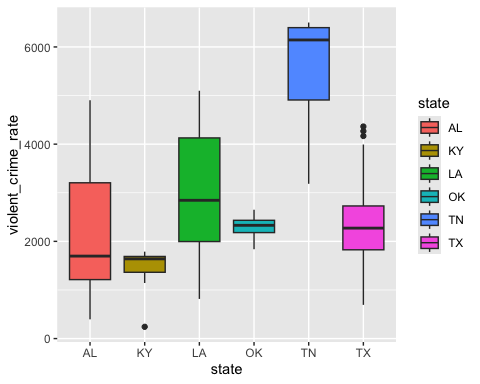
crimes%>%  
 filter(state=="NY")%>%  
 ggplot(aes(x=year,y=violent\_crime\_rate, group=city,linetype=city))+geom\_line()



# Assignment 2

By modifying the code below, create box plots that show the range of the violent crime rate for each state in the South region, where each box represents a state. FYI, six states are categorized as South in the dataset. After creating the chart, answer the following questions.

crimes %>%  
 filter(region == "South") %>%  
 ggplot(aes(x = state, y = violent\_crime\_rate, fill=state)) +  
 geom\_boxplot()



Q1. Which state has the city that has the lowest violent crime rate in the South region?

KY: Kentucky

Q2. Which state has the highest median value of the violent crime rate?

Tennessee (TN)

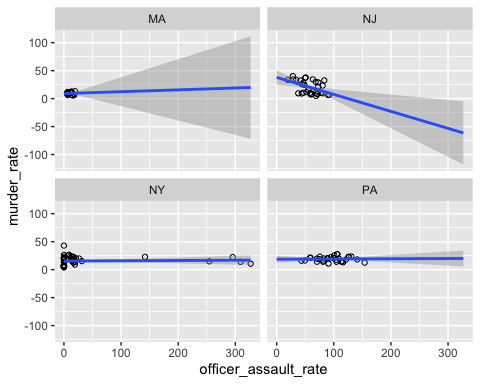
Q3. By modifying the code below, create scatterplots that show the relationship between the officer assault rate (x-axis) and the murder rate (y-axis) by the states in the Northeast region. It will produce four scatter plots with regression lines, one for each state.

Which state shows the strongest *negative* correlation (i.e. the steeper, downward slope of regression lines)?

NJ : New Jersey

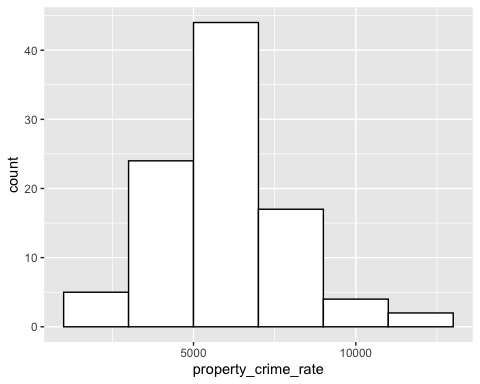
crimes %>%  
 filter(region == "Northeast") %>%  
 ggplot(aes(x = officer\_assault\_rate, y = murder\_rate)) +  
 geom\_point(shape = 1) +  
 geom\_smooth(method = lm, fullrange = TRUE) +  
 facet\_wrap(~ state)

## `geom\_smooth()` using formula = 'y ~ x'



Q4.By modifying the code below, create a histogram that shows the distribution of property crime rate in year 2002. Set the binwidth=2000. Which of the bins contains the largest number of cities?

crimes %>%  
 filter(year == "2002") %>%  
 ggplot(aes(x = property\_crime\_rate)) +  
 geom\_histogram(binwidth = 2000, colour = "black", fill = "white")



1. The bin whose x-axis values go from 0 to 2500.
2. The bin whose x-axis values go from 2500 to 5000.
3. The bin whose x-axis values go from 5000 to 7500.
4. The bin whose x-axis values go from 7500 to 10000.
5. The bin whose x-axis values go from 10000. to 12500

Answer: The bin whose x-axis values go from 5000 to 7500. This bin contains more than 40 cities, making it the one with the largest number of cities in 2002.

Q5. Similar to Assignment 1, generate a report from “Creating-Charts-Student.Rmd”. This time, please knit it in a Word document. Change the author name on the top of this R markdown file to your name. Compile this R markdown file into a Word document and submit it through the course Canvas. Your report should contain all the codes and the results of in-class exercises as well as the codes for the assignment questions.

Q6. Create a dashboard with Tableau by following the instruction in “Creating-Charts-Tableau.doc”. After creating a dashboard, export your Tableau dashboard view as PowerPoint. Click “File>Export As Powerpoint”, and include “This View”. It will create a PPT file that includes a snapshot of the dashboard. Submit this ppt file through Canvas as your answer for Question 6.