For this assignment you will use the *crime\_and\_incarceration\_altered.csv* data set that is provided on Bb. This data comes from Kaggle (and was altered for this assignment). You will perform the tasks and answer the questions below to practice pre-processing data. This assignment will require you to use *dplyr*, *tidyr*, and *ggplot2* which are all inside of our *tidyverse* package. So make sure to load this library first!

When finished you will submit this Word document containing your answers, R code, and copies of graphs. You will also submit your working R script file.

Information About the Kaggle Data Set:

https://www.kaggle.com/christophercorrea/prisoners-and-crime-in-united-states

The website says:

*In 1975, the United States set a new record with 240,593 prisoners incarcerated by state or federal agencies. The United States achieved new record totals during each of the next 34 years. Today, there are over 1,500,000 prisoners in the United States. Over one quarter of the world's entire population of prisoners is located in the United States.*

*The U.S. Education department*[*reports*](https://www2.ed.gov/rschstat/eval/other/expenditures-corrections-education/brief.pdf)*state and local government expenditures on prisons (not reflected in this dataset) have increased about three times as fast as spending on elementary and secondary education during this time period.* ***Does this significant investment into imprisonment improve public safety?***

*The*[*Bureau of Justice Statistics*](http://bjs.gov/)*administers the*[*National Prisoners Statistics Program (NPS)*](https://www.bjs.gov/index.cfm?ty=dcdetail&iid=269)*, an annual data collection effort that began in response to a 1926 congressional mandate. The Uniform Crime Report (UCR) has served as the FBI's primary national data collection tool since a 1930 congressional mandate directed the Attorney General to "acquire, collect, classify, and preserve identification, criminal identification, crime, and other records."*

library(tidyverse)

1. Bring the data set into R. Then create a subset of the data containing only the variables listed below for the states in the Midwest: Iowa, Illinois, North Dakota, South Dakota, Minnesota, Wisconsin, Nebraska, Kansas, Missouri, Indiana, Michigan, and Ohio. Is the data, by definition, “tidy”? Explain why or why not.

Variables of interest:

* jurisdiction
* year
* crimes\_estimated
* state\_population
* prisoner\_count
* murder\_manslaughter
* violent\_crime\_total
* robbery
* agg\_assault
* vehicle\_theft
* property\_crime\_total

CrimeAndInc<-read.csv(file='crime\_and\_incarceration\_by\_state\_altered.csv')

CrimeAndInc.1<-CrimeAndInc %>%

filter(jurisdiction %in% c('IOWA','ILLINOIS','NORTH DAKOTA','SOUTH DAKOTA','MINNESOTA','WISCONSIN','NEBRASKA','KANSAS','MISSOURI','INDIANA','MICHIGAN','OHIO')) %>%

select(jurisdiction,year,crimes\_estimated,state\_population,prisoner\_count,murder\_manslaughter,

violent\_crime\_total,robbery,agg\_assault,vehicle\_theft,property\_crime\_total)

CrimeAndInc.1

I would say the data is tidy because each variable is in a column, each observation is a row, and each value is a cell.

1. Look at the prisoner\_count over the years for each Midwestern state by creating a new data set containing this information (jurisdiction, year, prisoner\_count).

CrimeAndInc.2<-CrimeAndInc.1 %>%

select(jurisdiction, year, prisoner\_count)

* 1. Change the data format so that you have “wide” data, i.e. data in a spreadsheet style. Do this by using the spread() command and then View() the data. [Hint: There should be a column for each year]

CrimeAndInc.2Wide<-spread(CrimeAndInc.2, key=year, value=prisoner\_count)

view(CrimeAndInc.2Wide)

* 1. Put your data back into the “long” format (tidy) using gather() and then View() the data.

CrimeAndInc.2Long<-gather(CrimeAndInc.2Wide, key=year, value=prisoner\_count,-jurisdiction)

view(CrimeAndInc.2Long)

1. Start looking for outliers and/or other “problems” in the data set. Do this by creating a histogram for each variable in the data set where this would be appropriate, i.e. for all quantitative variables except year. What do you see? Describe any possible errors that you find. You can make histograms very quickly using *hist().* You are also welcome to use the ggplot2 package.

hist(CrimeAndInc.1$state\_population)

hist(CrimeAndInc.1$prisoner\_count)

hist(CrimeAndInc.1$murder\_manslaughter)

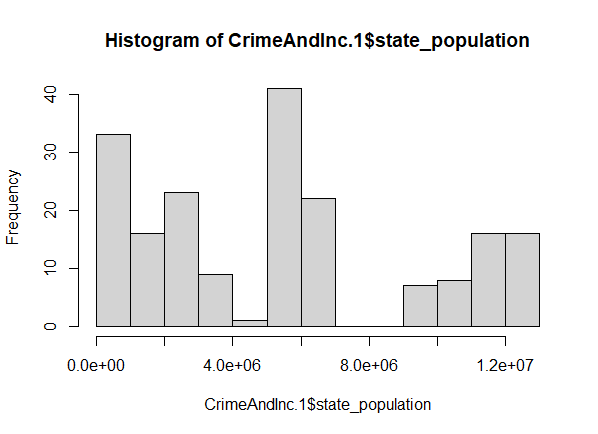
hist(CrimeAndInc.1$violent\_crime\_total)

hist(CrimeAndInc.1$robbery)

hist(CrimeAndInc.1$agg\_assault)

hist(CrimeAndInc.1$vehicle\_theft)

hist(CrimeAndInc.1$property\_crime\_total)

 Chart, histogram

Description automatically generated

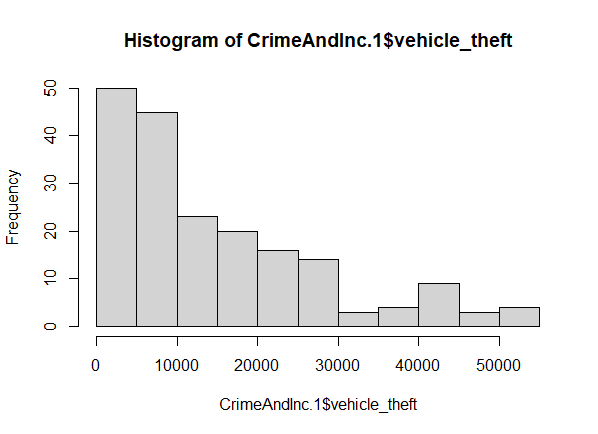
Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated  Chart, histogram

Description automatically generated

The histograms look pretty normal to me. There aren’t any points that seem super out of place for the histograms. Different populations are going to play some role the amount of crime. Though, there could be some errors that are going unnoticed because values are lumped from 0 to some number for the first bin. It is possible that there are some small numbers that the histogram prevents me from seeing.

By just clicking through the table and sorting the variables by high to low and vice versa I’ve found that Nebraska has three zeros in the property\_crime\_total while the next lowest is 12000. I think its safe to say that these are outliers.

1. Due to the type of data we have, the histograms were not super helpful. Since the data was recorded over time, we should try looking at some time series plots. Use the code below to make the time series plot that shows how the number of *murder\_manslaughter* crimes have changed over time for each state. What do you see? Describe any possible errors that you find.

*ggplot(data set, mapping=aes(x=year, y= murder\_manslaughter, color=jurisdiction))+*

*geom\_point()+*

*geom\_line()*

ggplot(CrimeAndInc.1, mapping=aes(x=year, y= murder\_manslaughter, color=jurisdiction))+

geom\_point()+

geom\_line()

Chart, line chart

Description automatically generated

I see a gap in the line for Wisconsin where 2008 is. I’m assuming that it had an NA for for that point in time.

Also, there is a chance that Illinois has an outlier in 2016. It looked like it was on a decreasing trajectory before setting a new high. Though it was at a time of political unrest and other states had similar (but less extreme) jumps that year.

Other than those two things I’m not sure if I see anything else.

1. Create similar time series plots for the remaining quantitative variables. What do you see? Describe any possible errors that you find.

ggplot(CrimeAndInc.1, mapping=aes(x=year, y= state\_population, color=jurisdiction))+

geom\_point()+

geom\_line()

Chart

Description automatically generated

There’s a pretty obvious error with Michigan 2014 for population.

ggplot(CrimeAndInc.1, mapping=aes(x=year, y= prisoner\_count, color=jurisdiction))+

geom\_point()+

geom\_line()

Chart, line chart

Description automatically generated

There is no obvious errors in the plot. Though Indiana’s jump in 2014 looks like there could be something. However, I don’t think it is extreme enough to cause concern.

ggplot(CrimeAndInc.1, mapping=aes(x=year, y= violent\_crime\_total, color=jurisdiction))+

geom\_point()+

geom\_line()

Chart, line chart

Description automatically generated

This seems alright. I don’t see anything obvious.

ggplot(CrimeAndInc.1, mapping=aes(x=year, y= robbery, color=jurisdiction))+

geom\_point()+

geom\_line()

Chart, line chart

Description automatically generated

There’s a large spike for Indiana 2011. Definitely an error.

ggplot(CrimeAndInc.1, mapping=aes(x=year, y= agg\_assault, color=jurisdiction))+

geom\_point()+

geom\_line()

Chart, line chart

Description automatically generated

There are no obvious errors for this variable.

ggplot(CrimeAndInc.1, mapping=aes(x=year, y= vehicle\_theft, color=jurisdiction))+

geom\_point()+

geom\_line()

Chart, line chart

Description automatically generated

I see a gap in the line for Wisconsin where 2008 is. There is most likely an NA there.

ggplot(CrimeAndInc.1, mapping=aes(x=year, y= property\_crime\_total, color=jurisdiction))+

geom\_point()+

geom\_line()

Chart, line chart

Description automatically generated

Nebraska has three zeros from 2001-2003. Definitely an error in the data.

1. You should have found several “problems” in the data. You will now “fix” the more obvious ones by doing the following:
   1. Fill in the missing value that you saw in the murder\_manslaughter data using either the mean or median of the data for that state. Explain how you chose which one to use. Show your work.

Wisconsin.MM<-CrimeAndInc.1 %>%

filter(CrimeAndInc.1$jurisdiction=='WISCONSIN') %>%

select(jurisdiction,year,murder\_manslaughter)

mean(Wisconsin.MM$murder\_manslaughter,na.rm=TRUE) #176.4667

median(Wisconsin.MM$murder\_manslaughter,na.rm=TRUE) #165

CrimeAndInc.1[96,6]=165 #replace with median

I replaced the NA with the median because it fit the data better. In 2007 the number was

around 185 and in 2009 the number was around 145. So, naturally, 165 looked like the

better fit. Also, the numbers of murder manslaughter really exploded at the end of the

graph which probably played a role in the average being higher.

* 1. The same state had a missing value for vehicle\_theft. The trend for this variable seems to closely follow that of another state. So it seems reasonable to fill in the missing value with the same value from the other state. Do this.

WiscMinn.Veh<-CrimeAndInc.1 %>%

filter(jurisdiction %in% c('MINNESOTA',"WISCONSIN"))%>%

select(jurisdiction,year,vehicle\_theft) #just used to easily compare the states

CrimeAndInc.1[96,10]=CrimeAndInc.1[90,10] #replace NA with 10195

Replaced the na for 2008 Wisconsin with the value of Minnesota from 2008 since they followed similar trends.

* 1. There was most likely a typo in the robbery data. What do you think happened, i.e. what was the typo? Take your best guess and then make the change.

Based on Indiana’s other data, (roughly 6000-7000), I would say that the 16997 has a typo. I would guess that the typo is that they accidentally added a one to the front. Instead, I think the data should be 6997.

CrimeAndInc.1[122,8] #indiana 2011 - the big jump on the graph - 16977

CrimeAndInc.1[122,8]=6977

* 1. Another fairly obvious typo can be seen in the state\_population data. Since state populations are easy to look up, find and replace the typo with the correct value.

The error is Michigan 2014’s population. The actual amount should be 9,929,848 according to <https://www.senate.michigan.gov/sfa/Economics/MichiganPopulationByCounty.PDF>.

CrimeAndInc.1[161,4] #999877

CrimeAndInc.1[161,4]=9929848

* 1. There also seem to be missing values in the property\_crime\_total data. A few values were coded as a zero. It’s possible that there were no crimes of this nature, but it’s more likely that the 0’s represent missing values. Replace this data with the first non-missing value for the state (2004). This may not be a good replacement, but the downward trend in the data for this state is fairly slow.

CrimeAndInc.1[c(8,20,32),11] #our zeros – Nebraska 2001-2003

CrimeAndInc.1[44,11] #2004 Nebraska - 61512

CrimeAndInc.1[c(8,20,32),11]=61512

1. Another graphical option is to use scatterplots that display the relationship between two variables (rather than a time series plot). For example, plot prisoner\_count and vehicle\_theft together. Do you think this type of graph is helpful? You can do this using the *plot()* function or use ggplot2 functions.

ggplot(CrimeAndInc.1, mapping=aes(x=prisoner\_count, y=vehicle\_theft))+

geom\_point()+

geom\_smooth()

Chart, scatter chart

Description automatically generated

I think this type of graph could be helpful. It appears there is a relationship between the number of prisoners and the number incarcerated for vehicle theft.

1. What if we add a third variable to the scatter plot? Try mapping the year variable to color. Use as.factor(year) to make the color-coding easier to read. Then try mapping jurisdiction to color. Do either of these present the data in a way that’s easier to identify possible outliers, typos, etc? Explain.

ggplot(CrimeAndInc.1, mapping=aes(x=prisoner\_count, y=vehicle\_theft, color=as.factor(year)))+

geom\_point()

Chart, scatter chart

Description automatically generated

In addition to the information discovered earlier, this graph shows us that there have been less vehicle theft related incarcerations as time has gone on.

ggplot(CrimeAndInc.1, mapping=aes(x=prisoner\_count, y=vehicle\_theft, color=as.factor(jurisdiction)))+

geom\_point()

Chart, scatter chart

Description automatically generated

I think this plot allows you to see how little clusters of each state’s data. Some are more spread out than others, but you can notice that some are clumped together with it’s own data from other years.

I would say that neither are incredibly helpful in identifying outliers. Though, mapping by jurisdiction could allow you to see a colored dot that is far from its state cluster. Because of this, I would say that this is the most helpful of the two. However, I’m colorblind with reds and greens so these dots are harder for me to find which makes this harder to identify when one color is farther from it’s group.

1. We still need to check the crimes\_estimated variable. Create a table for this categorical variable using *table()*. What do you see? Describe any possible errors that you find.

table(CrimeAndInc.1$crimes\_estimated)

I see that there is one empty cell, which means there is a missing value. I also see that there are some parts of the data that use 1 and 0 instead of true and false.

1. You should have found two problems with the crimes\_estimated data. You will now “fix” them by doing the following:
   1. The years 2015 and 2016 were coded differently. Rather than True/False, we have 1/0 (where 1 = true). Change this so that all of the data is recorded in the same manner.

str(CrimeAndInc.1) #character data for crimes\_esimated

crimes<-CrimeAndInc.1 #just incase I mess up

ii<-CrimeAndInc.1$crimes\_estimated=='FALSE'

jj<-CrimeAndInc.1$crimes\_estimated=='TRUE'

CrimeAndInc.1[jj,3]=1

CrimeAndInc.1[ii,3]=0

CrimeAndInc.1$crimes\_estimated=as.numeric(CrimeAndInc.1$crimes\_estimated)

* 1. Replace the missing value in this data with the mode for this variable.

table(CrimeAndInc.1$crimes\_estimated) #mode is 0 since it occurs the most

CrimeAndInc.1[190,3] #the NA

CrimeAndInc.1[190,3]=0 #change to 0

CrimeAndInc.1$crimes\_estimated #No more NA

* 1. Check and make sure that the above replacement makes sense. This is easiest to do by looking at the crimes\_estimated data in wide/spreadsheet format. What do you see?

CrimeAndInc.1Wide<-CrimeAndInc.1 %>%

select(jurisdiction,year,crimes\_estimated)%>%

spread(key=year, value=crimes\_estimated)

Yes it makes sense, the only jurisdiction for crimes\_estimated with “true” or 1 is Illinois. The rest are 0’s. Ohio, the jurisdiction that we fixed the NA for 2016 crimes\_estimated, has all 0’s so it makes sense to also be 0.

1. Normally this type of data would be scaled based on population. This enables us to compare states and makes more sense since population changes over time. Use *mutate()* to create and add two new variables to the data set: total violent crimes per 100,000 population and total property crime per 100,000 population. [Hint: Divide by state\_population then multiply by 100,000.]

CrimeAndInc.p100k<-CrimeAndInc.1 %>%

mutate(VCT.p100k=(violent\_crime\_total/state\_population)\*100000,

PCT.p100k=(property\_crime\_total/state\_population)\*100000)

CrimeAndInc.p100k

1. Compare the time series graph for violent crimes for each state to the time series graph for violent crimes per 100,000 population over time for each state. What do you see? How does the story change?

ggplot(CrimeAndInc.p100k, mapping=aes(x=year, y= VCT.p100k, color=jurisdiction))+

geom\_point()+

geom\_line()

ggplot(CrimeAndInc.p100k, mapping=aes(x=year, y= violent\_crime\_total, color=jurisdiction))+

geom\_point()+

geom\_line()

Chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

The original graph (with violent\_crime\_total), shows that most of the totals stay roughly the same but increase slightly. The obvious exceptions are Illinois and Michigan. These two states are also seen as more violent places. Specifically, Chicago and Detroit.

However, the new graph shows that violent crimes per 100k are changing more than they appear to in the original graph. Also, Illinois and Michigan aren’t the only states with a lot of violent crimes. According to the new graph, Missouri has a comparable violent crime total per 100k people. This paints a new story for Missouri which should be seen as a higher crime state like Michigan and Illinois.

1. Practice doing another data transformation by creating and adding a new variable to the data set: the min-max normalization for prisoner\_count. Then create a histogram to verify that all values fall between 0 and 1. [Hint: Subtract the minimum prisoner\_count and then divide by its range.]

min(CrimeAndInc.p100k$prisoner\_count) #1088 min

max(CrimeAndInc.p100k$prisoner\_count) #52240 max

52240-1088 #51152 range

prisoner\_count.MinMax<-(CrimeAndInc.p100k$prisoner\_count-1088)/51152

prisoner\_count.MinMax

hist(prisoner\_count.MinMax)

CrimeAndInc.p100k<-mutate(CrimeAndInc.p100k,

prisoner\_count.MinMax=prisoner\_count.MinMax)

CrimeAndInc.p100k

Chart, histogram

Description automatically generated