ds3

Matthew Robinson

library(ggplot2)  
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  
## ✔ lubridate 1.9.3 ✔ tibble 3.2.1  
## ✔ purrr 1.0.2 ✔ tidyr 1.3.1  
## ── 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(ggstance)

##   
## Attaching package: 'ggstance'  
##   
## The following objects are masked from 'package:ggplot2':  
##   
## geom\_errorbarh, GeomErrorbarh

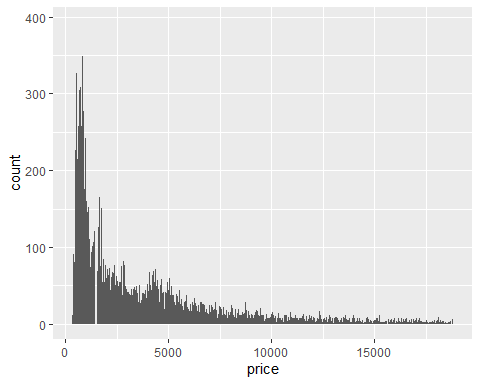
library(lvplot)  
library(ggbeeswarm)  
library(nycflights13)

# Section 7.3.4

# Exercise 2

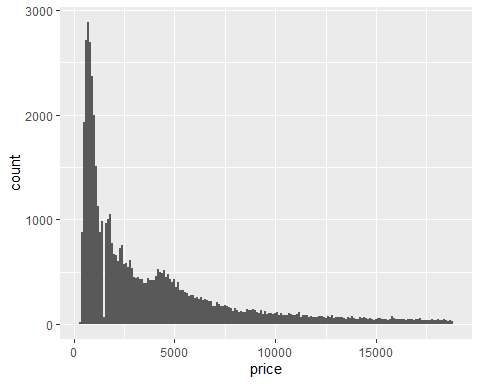
Explore the distribution of price. Do you discover anything unusual or surprising?

ggplot(data = diamonds) + geom\_histogram(mapping = aes(x = price), binwidth = 10)



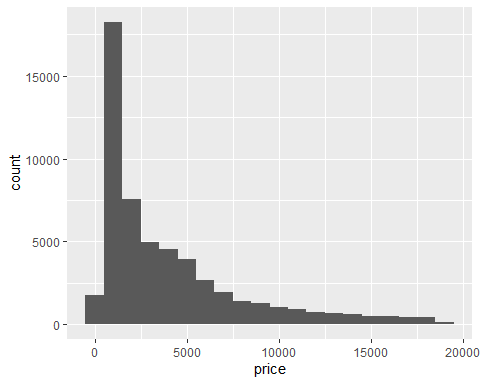
This histogram uses a bin width of 10. As shown in this histogram, there is a gap around the $1500 mark.

ggplot(data = diamonds) + geom\_histogram(mapping = aes(x = price), binwidth = 100)



This histogram uses a bin width of 100. There is still a clear drop around the $1500 mark, but it is not a complete gap.

ggplot(data = diamonds) + geom\_histogram(mapping = aes(x = price), binwidth = 1000)



This histogram uses a bin width of 100. No drop can be seen in this histogram where there should be one.

# Exercise 3

How many diamonds are 0.99 carat? How many are 1 carat? What do you think is the cause of the difference?

filter(diamonds, carat == .99)

## # A tibble: 23 × 10  
## carat cut color clarity depth table price x y z  
## <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 0.99 Fair I SI2 68.1 56 2811 6.21 6.06 4.18  
## 2 0.99 Fair J SI1 55 61 2812 6.72 6.67 3.68  
## 3 0.99 Fair J SI1 58 67 2949 6.57 6.5 3.79  
## 4 0.99 Fair I SI1 60.7 66 3337 6.42 6.34 3.87  
## 5 0.99 Fair H VS2 71.6 57 3593 5.94 5.8 4.2   
## 6 0.99 Very Good J SI1 60.3 57 4002 6.44 6.49 3.9   
## 7 0.99 Good F SI2 63.3 54 4052 6.36 6.43 4.05  
## 8 0.99 Premium F SI2 60.6 61 4075 6.45 6.38 3.89  
## 9 0.99 Ideal I SI1 61.8 57 4763 6.4 6.42 3.96  
## 10 0.99 Very Good E SI2 61.8 59 4780 6.3 6.33 3.9   
## # ℹ 13 more rows

There are only 23 .99 carat diamonds while there 1,558 1 carat diamonds.

filter(diamonds, carat == 1)

## # A tibble: 1,558 × 10  
## carat cut color clarity depth table price x y z  
## <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 1 Premium I SI2 58.2 60 2795 6.61 6.55 3.83  
## 2 1 Premium J SI2 62.3 58 2801 6.45 6.34 3.98  
## 3 1 Fair G I1 66.4 59 2808 6.16 6.09 4.07  
## 4 1 Fair J VS2 65.7 59 2811 6.14 6.07 4.01  
## 5 1 Premium H I1 61.3 60 2818 6.43 6.39 3.93  
## 6 1 Fair H SI2 65.3 62 2818 6.34 6.12 4.08  
## 7 1 Premium F I1 58.9 60 2841 6.6 6.55 3.87  
## 8 1 Fair G SI2 67.8 61 2856 5.96 5.9 4.02  
## 9 1 Fair I SI1 67.9 62 2856 6.19 6.03 4.15  
## 10 1 Fair H SI2 66.1 56 2856 6.21 5.97 4.04  
## # ℹ 1,548 more rows

It makes sense that there would be many more 1 carat diamonds than .99 carat diamonds because a diamond that close to 1 carat could be rounded up to 1 because it is more marketable.

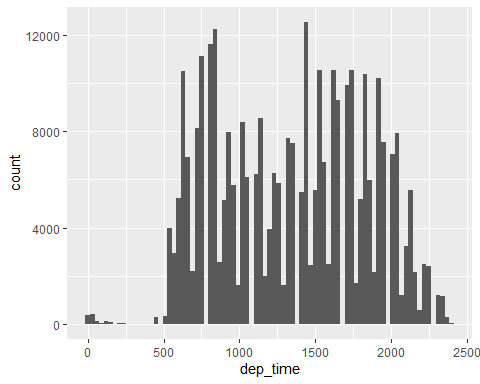
# Section 7.4.1

# Exercise 1

What happens to missing values in a histogram?

ggplot(data = flights) + geom\_histogram(mapping = aes(x = dep\_time), binwidth = 30)

## Warning: Removed 8255 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

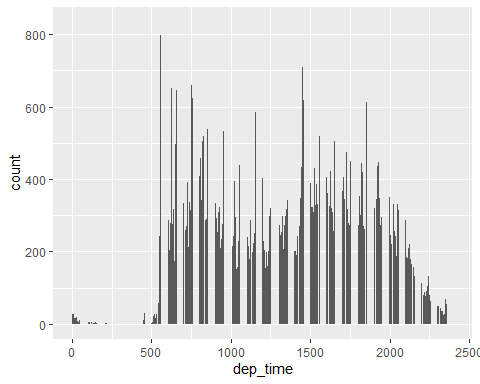


Missing values in a histogram are removed because they are outside the scale range of the stat bin function.

What happens to missing values in a bar chart?

ggplot(data = flights) + geom\_bar(mapping = aes(x = dep\_time))

## Warning: Removed 8255 rows containing non-finite outside the scale range  
## (`stat\_count()`).



Missing values in a bar chart are removed because they are outside the scale range of stat count function.  
Why is there a difference?  
The x axis of a histogram is based on bins while the x axis of a bar chart is based on counts.

# Exercise 2

What does na.rm = TRUE do in mean() and sum()?

mean(flights$dep\_time)

## [1] NA

sum(flights$dep\_time)

## [1] NA

Finding the mean and sum do not work if there are missing values.

mean(flights$dep\_time, na.rm = TRUE)

## [1] 1349.11

sum(flights$dep\_time, na.rm = TRUE)

## [1] 443210949

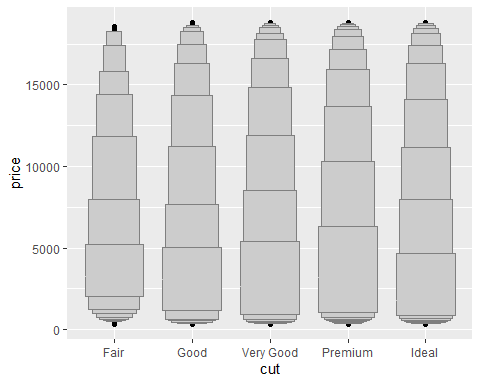
The argument “na.rm = TRUE” allows you to find the mean and sum by ignoring the missing values.

# Section 7.5.1.1

# Exercise 4

One problem with boxplots is that they were developed in an era of much smaller datasets and tend to display a prohibitively large number of “outlying values”. One approach to remedy this problem is the letter value plot. Install the lvplot package, and try using geom\_lv() to display the distribution of price vs cut. What do you learn? How do you interpret the plots?

ggplot(data = diamonds) + geom\_lv(mapping = aes(x = cut, y = price))

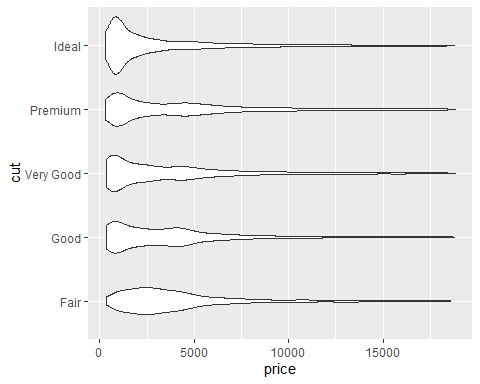


Based on this plot, premium diamonds seem to have the largest variation in price while fair diamonds have the smallest variation in price.

# Exercise 5

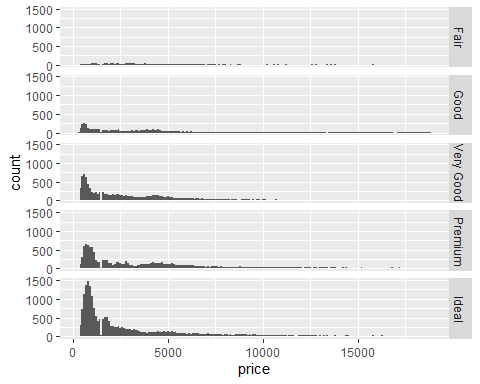
Compare and contrast geom\_violin() with a facetted geom\_histogram(), or a coloured geom\_freqpoly(). What are the pros and cons of each method?

ggplot(data = diamonds) + geom\_violin(mapping = aes(x = price, y = cut))



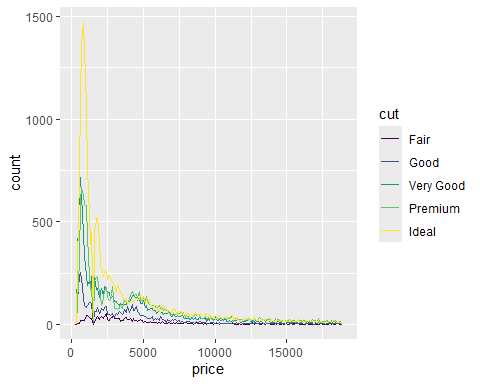
The geom\_violin function gives a good view of how common different prices are for each cut. However, there is no way to see exact counts.

ggplot(data = diamonds) + geom\_histogram(mapping = aes(x = price), binwidth = 100) + facet\_grid(cut ~ .)



The geom\_histogram with the facet\_grid function gives a good view of how many diamonds of each cut fall in different price ranges. However, it is hard to compare between the cuts.

ggplot(data = diamonds) + geom\_freqpoly(mapping = aes(x = price, colour = cut), binwidth = 100)



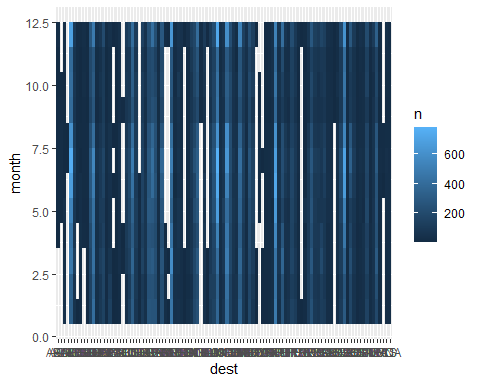
The geom\_freqpoly function gives a good view of how prices compare between different cuts. However, it can be a bit difficult to read.

# Section 7.5.2.1

# Exercise 2

Use geom\_tile() together with dplyr to explore how average flight delays vary by destination and month of year. What makes the plot difficult to read?

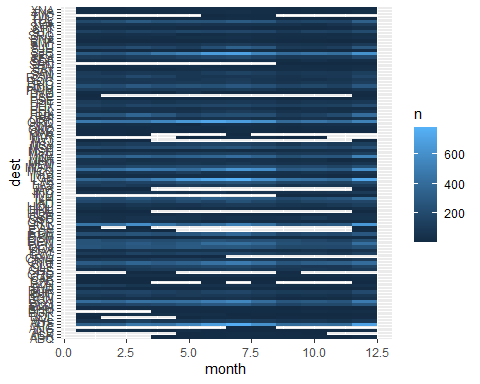
filter(flights, dep\_delay > 0) %>% count(dest, month) %>% ggplot(mapping = aes(x = dest, y = month)) + geom\_tile(mapping = aes(fill = n))



This plot is difficult to read because you cannot read the destinations on the x axis.

How could you improve it?

filter(flights, dep\_delay > 0) %>% count(month, dest) %>% ggplot(mapping = aes(x = month, y = dest)) + geom\_tile(mapping = aes(fill = n))

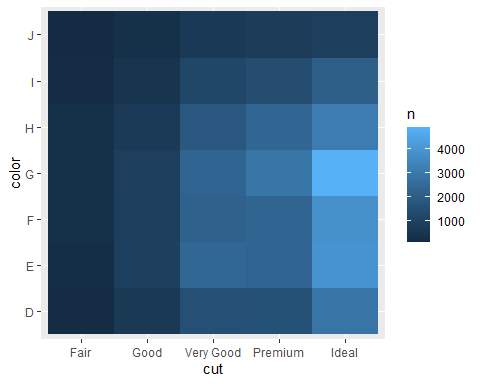


Putting the destinations on the y axis makes it a little easier to read, but it is still not perfect.

# Exercise 3

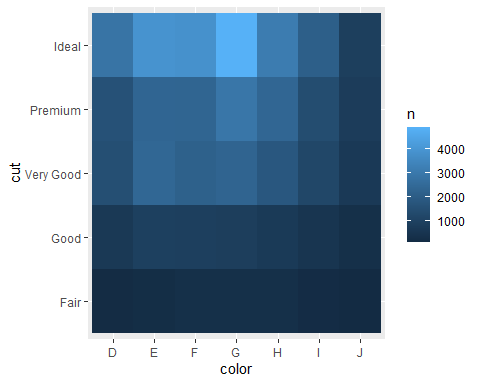
Why is it slightly better to use aes(x = color, y = cut) rather than aes(x = cut, y = color) in the example above?

diamonds %>% count(cut, color) %>% ggplot(mapping = aes(x = cut, y = color)) + geom\_tile(mapping = aes(fill = n))



It is usually better to have an ordered variable on the y axis. Color is not an ordered variable, so it would be better to put it on the x axis.

diamonds %>% count(color, cut) %>% ggplot(mapping = aes(x = color, y = cut)) + geom\_tile(mapping = aes(fill = n))



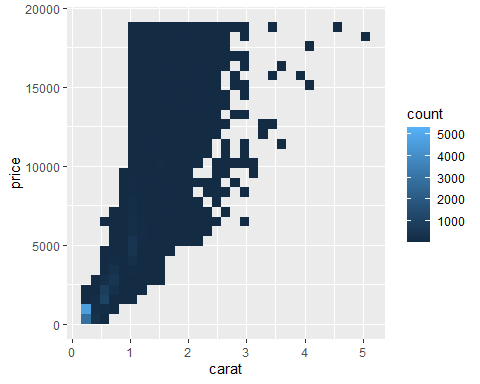
Putting an ordered variable such as cut on the y axis makes it easier to see trends in the data as seen above.

# Section 7.5.3.1

# Exercise 1

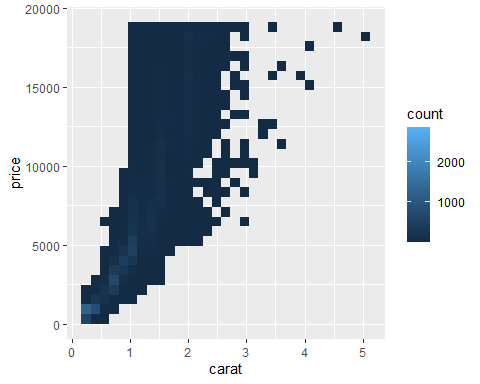
Instead of summarising the conditional distribution with a boxplot, you could use a frequency polygon. What do you need to consider when using cut\_width() vs cut\_number()? How does that impact a visualisation of the 2d distribution of carat and price?

ggplot(data = diamonds) + geom\_bin2d(mapping = aes(x = carat, y = price, group = cut\_width(carat, .1)))



The cut\_width function divides the data into bins of a certain width. When this is used, the count goes all the way up to around 5000.

ggplot(data = diamonds) + geom\_bin2d(mapping = aes(x = carat, y = price, group = cut\_number(carat, 20)))



The cut\_number function divides the data into bins with the same number of points in each bin. When this is used, the count only goes up to around 2000.

# Exercise 5

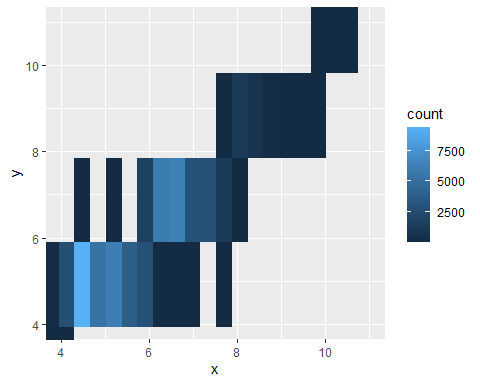
Two dimensional plots reveal outliers that are not visible in one dimensional plots. For example, some points in the plot below have an unusual combination of x and y values, which makes the points outliers even though their x and y values appear normal when examined separately.

ggplot(data = diamonds) + geom\_point(mapping = aes(x = x, y = y)) + coord\_cartesian(xlim = c(4, 11), ylim = c(4, 11))



Why is a scatterplot a better display than a binned plot for this case?

ggplot(data = diamonds) + geom\_bin2d(mapping = aes(x = x, y = y)) + coord\_cartesian(xlim = c(4, 11), ylim = c(4, 11))

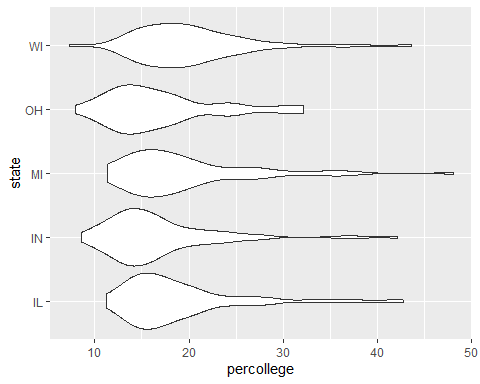


There is no way to see outliers from a binned plot like there is in a scatterplot.

# Midwest Demographics

This data set contains data on the demographics of each county in five midwestern states: Illinois, Indiana, Michigan, Ohio, and Wisconsin. The following plot shows the percent of people who are college educated in each county divided by state.

ggplot(data = midwest) + geom\_violin(mapping = aes(x = percollege, y = state))



From this data, there are a few observations that can be made. Wisconsin has the highest average percent with Illinois and Ohio being practically tied for last. However, Wisconsin does have the lowest floor. Michigan has the highest ceiling and highest floor while Ohio has the lowest ceiling. The following table shows the ten counties with the hightest percent of people who are college educated.

select(arrange(midwest, desc(percollege)), county, state, percollege)

## # A tibble: 437 × 3  
## county state percollege  
## <chr> <chr> <dbl>  
## 1 WASHTENAW MI 48.1  
## 2 DANE WI 43.6  
## 3 DU PAGE IL 42.8  
## 4 HAMILTON IN 42.1  
## 5 CHAMPAIGN IL 41.3  
## 6 LAKE IL 37.8  
## 7 MONROE IN 37.7  
## 8 OZAUKEE WI 37.4  
## 9 OAKLAND MI 37.0  
## 10 INGHAM MI 36.8  
## # ℹ 427 more rows

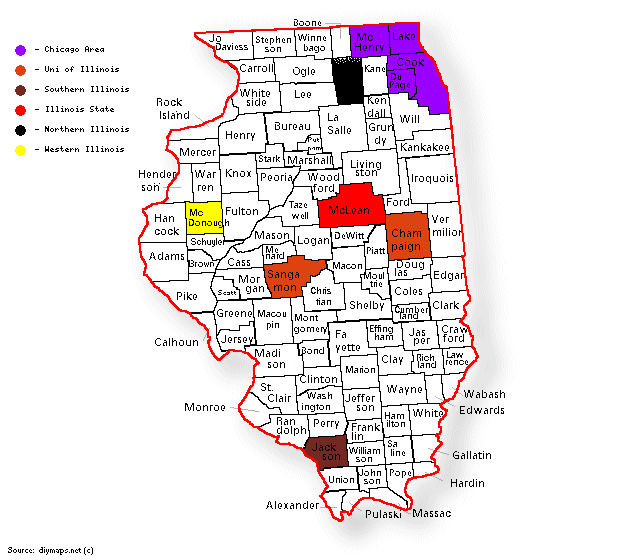
Unsurprisingly, all of these counties are either home to or near major universities. Washtenaw, Dane, Champaign, and Monroe counties are each home to thier state’s respective flagship universities. In fact, Ohio is the only state whose flagship university is not included in this list. Oakland and Ingham counties are also home to major universities. Meanwhile, Du Page, Hamilton, Lake, and Ozaukee counties are all suburbs of their state’s respective largest cities which are home to several major universities. To get a closer look at this, we will look at each of these five states individually.

# Illinois

select(arrange(filter(midwest, state == "IL"), desc(percollege)), county, percollege)

## # A tibble: 102 × 2  
## county percollege  
## <chr> <dbl>  
## 1 DU PAGE 42.8  
## 2 CHAMPAIGN 41.3  
## 3 LAKE 37.8  
## 4 JACKSON 36.6  
## 5 MCLEAN 33.8  
## 6 DE KALB 32.8  
## 7 SANGAMON 29.0  
## 8 MCHENRY 28.1  
## 9 COOK 28.0  
## 10 MCDONOUGH 27.9  
## # ℹ 92 more rows

As mentioned earlier, Champaign County is home to the University of Illinois. Sangamon County is also home to a campus of the University of Illinois in Springfield. Cook County is also home to a campus of the University of Illinois in Chicago along with Chicago State University and several private universities including Northwestern University, DePaul University, and the University of Chicago. The surrounding counties, including Du Page, Lake and McHenry counties benefit from this large concentration of universities. Other major universities in counties on this list include Southern Illinois University in Jackson County, Illinois State University in McLean County, Northern Illinois University in DeKalb County, and Western Illinois University in McDonough County.



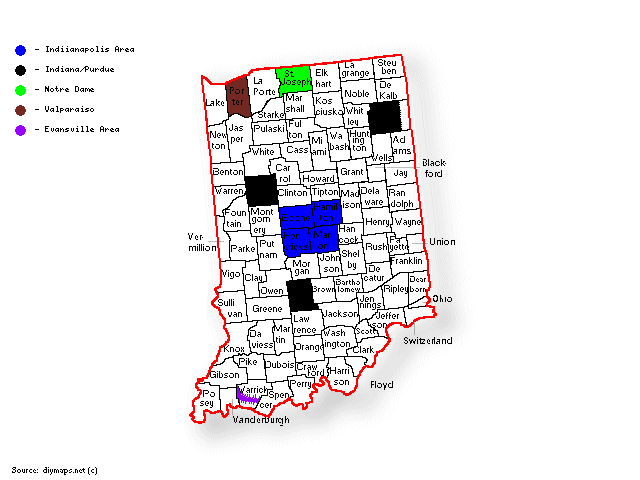
Illinois Counties

# Indiana

select(arrange(filter(midwest, state == "IN"), desc(percollege)), county, percollege)

## # A tibble: 92 × 2  
## county percollege  
## <chr> <dbl>  
## 1 HAMILTON 42.1  
## 2 MONROE 37.7  
## 3 TIPPECANOE 36.2  
## 4 BOONE 27.8  
## 5 ALLEN 27.4  
## 6 MARION 26.7  
## 7 ST JOSEPH 24.6  
## 8 PORTER 24.5  
## 9 HENDRICKS 24.2  
## 10 WARRICK 23.8  
## # ℹ 82 more rows

As mentioned earlier, Monroe County is home to Indiana University. Another major public university in the state is Purdue University in Tippecanoe County. Allen County is home to campuses for both of these universities in Fort Wayne. Marion County is also home to campuses for both of these universities in Indianapolis along with several private universities including Butler University and the University of Indianapolis. The surrounding counties, including Hamilton, Boone, and Hendricks counties, benefit from this large concentration of universities. Warrick County is a suburb of Evansville which is home to the University of Southern Indiana and the University of Evansville. The other two counties on this list are home to major private universities with the University of Notre Dame in St. Joseph County and Valparaiso University in Porter County.



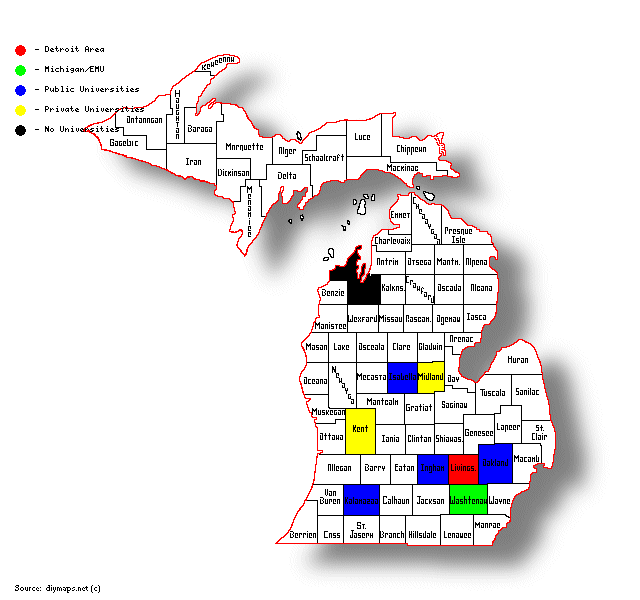
Indiana Counties

# Michigan

select(arrange(filter(midwest, state == "MI"), desc(percollege)), county, percollege)

## # A tibble: 83 × 2  
## county percollege  
## <chr> <dbl>  
## 1 WASHTENAW 48.1  
## 2 OAKLAND 37.0  
## 3 INGHAM 36.8  
## 4 MIDLAND 35.6  
## 5 KALAMAZOO 34.6  
## 6 LEELANAU 32.6  
## 7 GRAND TRAVERSE 31.0  
## 8 KENT 28.6  
## 9 LIVINGSTON 27.6  
## 10 ISABELLA 27.2  
## # ℹ 73 more rows

As mentioned earlier, Washtenaw County is home to the University of Michigan. It is also home to Eastern Michigan University. Also mentioned earlier were Oakland County, home to Oakland University, and Ingham University, home to Michigan State University. Livingston County is a suburb of Detroit which is home to Wayne State University and the University of Detroit Mercy. Other major universities in counties on this list include Northwood University in Midland County, Western Michigan University in Kalamazoo County, Davenport University in Kent County, and Central Michigan University in Isabella County. Interestingly, Leelanau and Grand Traverse counties are not home to or near any major universities. The only college in the area is Northwestern Michigan College.



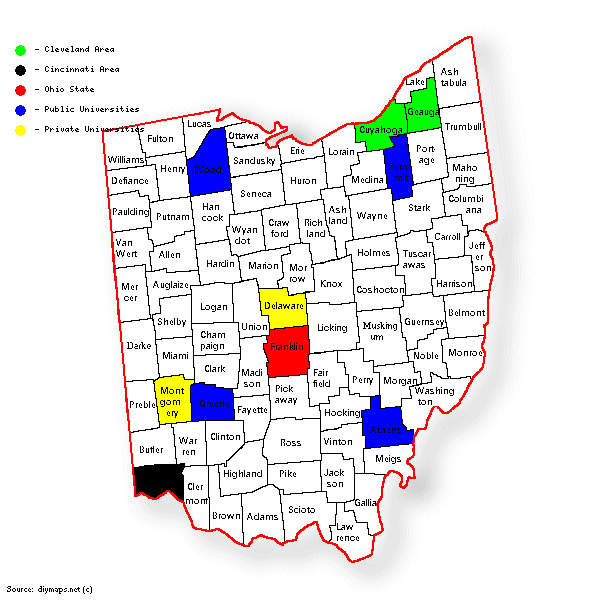
Michigan Counties

# Ohio

select(arrange(filter(midwest, state == "OH"), desc(percollege)), county, percollege)

## # A tibble: 88 × 2  
## county percollege  
## <chr> <dbl>  
## 1 FRANKLIN 32.2  
## 2 GREENE 32.0  
## 3 GEAUGA 31.6  
## 4 DELAWARE 31.6  
## 5 HAMILTON 29.8  
## 6 ATHENS 29.1  
## 7 WOOD 29.1  
## 8 MONTGOMERY 26.6  
## 9 CUYAHOGA 25.1  
## 10 SUMMIT 24.7  
## # ℹ 78 more rows

Franklin County is home to Ohio State University and is the only county home to a flagship university not included in the earlier list. Cuyahoga County is home to several universities including Cleveland State University and Case Western Reserve University. The surrounding counties, including Geauga County, benefit from this large concentration of universities. Hamilton County is also home to several universities including the University of Cincinnati and Xavier University. Other major universities in counties on this list include Wright State University in Greene County, Ohio Wesleyan University in Delaware County, Ohio University in Athens County, Bowling Green State University in Wood County, the University of Dayton in Montgomery County, and the University of Akron in Summit County.



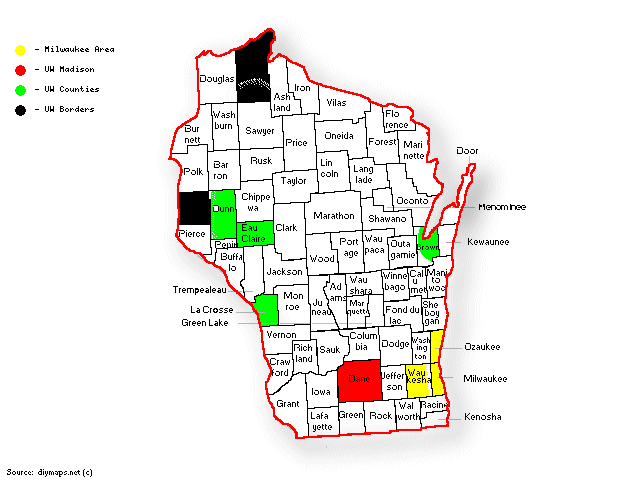
Ohio Counties

# Wisconsin

select(arrange(filter(midwest, state == "WI"), desc(percollege)), county, percollege)

## # A tibble: 72 × 2  
## county percollege  
## <chr> <dbl>  
## 1 DANE 43.6  
## 2 OZAUKEE 37.4  
## 3 WAUKESHA 35.4  
## 4 LA CROSSE 30.5  
## 5 EAU CLAIRE 29.9  
## 6 ST CROIX 28.6  
## 7 BROWN 26.3  
## 8 DUNN 26.3  
## 9 BAYFIELD 25.8  
## 10 MILWAUKEE 25.4  
## # ℹ 62 more rows

As mentioned earlier, Dane County is home to the University of Wisconsin. Milwaukee County is also home to a campus of the University of Wisconsin along with several private universities including Maquette University. The surrounding counties, including Ozaukee and Waukesha counties, benefit from this large concentration of universities. St. Croix County borders Pierce County which is home to a campus of the University of Wiscosnin in River Falls while Bayfield County borders Douglas County which is also home to a campus of the University of Wisconsin in Superior. The rest of the counties on this list are also home to campuses of the University of Wisconsin.



Wisconsin Counties

# Conclusion

Based on this data, it is clear that the counties with the highest percentage of people who are college educated are those which are home to or near major universities. Out of the fifty counties listed, only two of them do not fit this criteria. This makes sense as most of the people living near a major university are likely to have some affiliation with that university. While this conclusion is not surprising, it is interesting to see how the data supports it.