ds5

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(lubridate)  
library(nycflights13)  
library(maps)

##   
## Attaching package: 'maps'  
##   
## The following object is masked from 'package:purrr':  
##   
## map

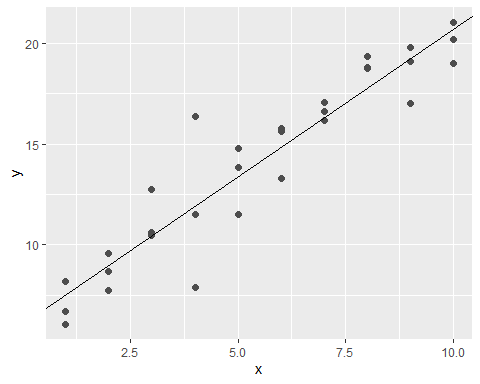
library(lubridate)  
library(hexbin)  
library(splines)  
library(modelr)  
options(na.action = na.warn)

# Section 23.2.1

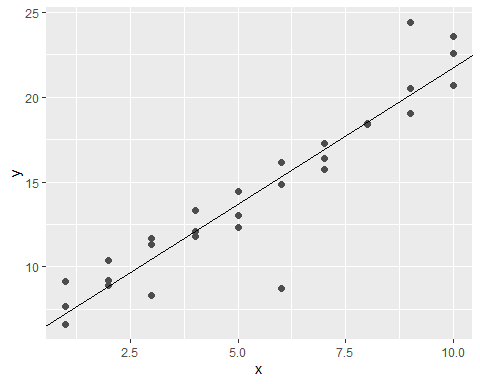
# Exercise 1

One downside of the linear model is that it is sensitive to unusual values because the distance incorporates a squared term. Fit a linear model to the simulated data below, and visualise the results. Rerun a few times to generate different simulated datasets. What do you notice about the model?

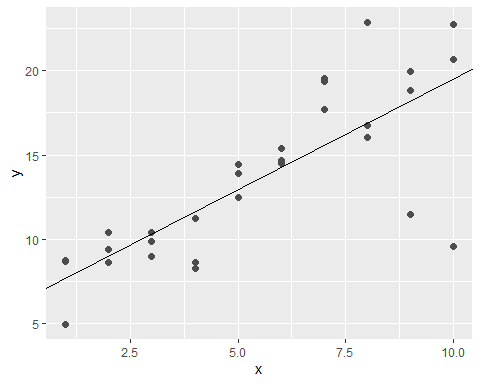
model1 <- function(a, data)   
{  
 a[1] + data$x \* a[2]  
}  
measure\_distance <- function(mod, data)   
{  
 diff <- data$y - model1(mod, data)  
 sqrt(mean(diff ^ 2))  
}  
sim1a <- tibble(x = rep(1:10, each = 3), y = x \* 1.5 + 6 + rt(length(x), df = 2))  
best1 <- optim(c(0, 0), measure\_distance, data = sim1a)  
ggplot(sim1a, aes(x, y)) + geom\_point(size = 2, colour = "grey30") + geom\_abline(intercept = best1$par[1], slope = best1$par[2])



sim1b <- tibble(x = rep(1:10, each = 3), y = x \* 1.5 + 6 + rt(length(x), df = 2))  
best2 <- optim(c(0, 0), measure\_distance, data = sim1b)  
ggplot(sim1b, aes(x, y)) + geom\_point(size = 2, colour = "grey30") + geom\_abline(intercept = best2$par[1], slope = best2$par[2])



sim1c <- tibble(x = rep(1:10, each = 3), y = x \* 1.5 + 6 + rt(length(x), df = 2))  
best3 <- optim(c(0, 0), measure\_distance, data = sim1c)  
ggplot(sim1c, aes(x, y)) + geom\_point(size = 2, colour = "grey30") + geom\_abline(intercept = best3$par[1], slope = best3$par[2])



The models for each simulation seem to be a bit off because they are not resistant to unusual values.

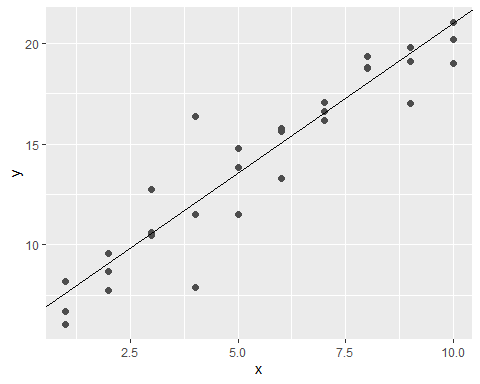
# Exercise 2

One way to make linear models more robust is to use a different distance measure. For example, instead of root-mean-squared distance, you could use mean-absolute distance:

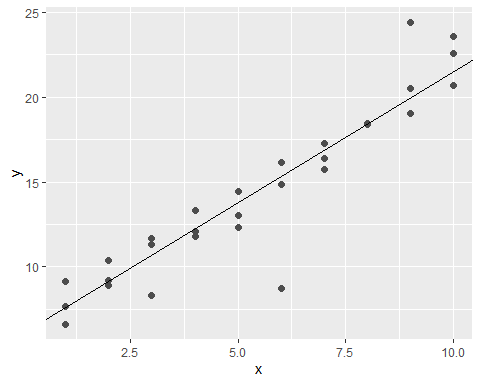
measure\_distance <- function(mod, data)   
{  
 diff <- data$y - model1(mod, data)  
 mean(abs(diff))  
}

Use optim() to fit this model to the simulated data above and compare it to the linear model.

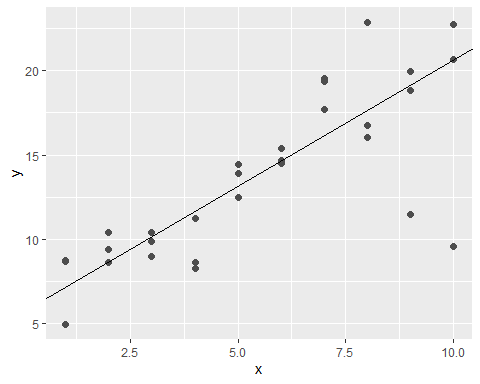
best1 <- optim(c(0, 0), measure\_distance, data = sim1a)  
ggplot(sim1a, aes(x, y)) + geom\_point(size = 2, colour = "grey30") + geom\_abline(intercept = best1$par[1], slope = best1$par[2])



best2 <- optim(c(0, 0), measure\_distance, data = sim1b)  
ggplot(sim1b, aes(x, y)) + geom\_point(size = 2, colour = "grey30") + geom\_abline(intercept = best2$par[1], slope = best2$par[2])



best3 <- optim(c(0, 0), measure\_distance, data = sim1c)  
ggplot(sim1c, aes(x, y)) + geom\_point(size = 2, colour = "grey30") + geom\_abline(intercept = best3$par[1], slope = best3$par[2])



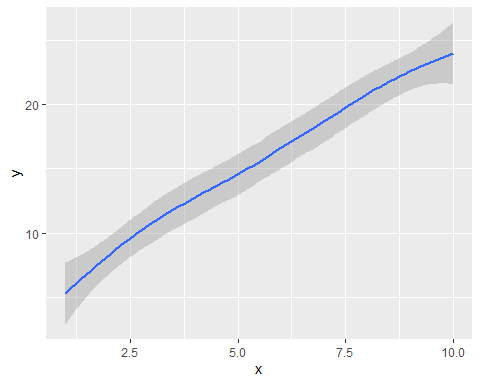
These models seem to be more accurate than the previous models and are much more resistant to unusual values.

# Section 23.3.3

# Exercise 1

Instead of using lm() to fit a straight line, you can use loess() to fit a smooth curve. Repeat the process of model fitting, grid generation, predictions, and visualisation on sim1 using loess() instead of lm(). How does the result compare to geom\_smooth()?

ggplot(sim1, aes(x, y)) + geom\_smooth(method = loess, formula = y ~ x)



The geom\_smooth function uses the loess function to create a smooth curve instead of a straight line.

# Exercise 2

add\_predictions() is paired with gather\_predictions() and spread\_predictions(). How do these three functions differ?

sim1\_mod <- lm(y ~ x, data = sim1)  
sim1 %>% data\_grid(x) %>% add\_predictions(sim1\_mod)

## # A tibble: 10 × 2  
## x pred  
## <int> <dbl>  
## 1 1 6.27  
## 2 2 8.32  
## 3 3 10.4   
## 4 4 12.4   
## 5 5 14.5   
## 6 6 16.5   
## 7 7 18.6   
## 8 8 20.6   
## 9 9 22.7   
## 10 10 24.7

sim1 %>% data\_grid(x) %>% gather\_predictions(sim1\_mod)

## # A tibble: 10 × 3  
## model x pred  
## <chr> <int> <dbl>  
## 1 sim1\_mod 1 6.27  
## 2 sim1\_mod 2 8.32  
## 3 sim1\_mod 3 10.4   
## 4 sim1\_mod 4 12.4   
## 5 sim1\_mod 5 14.5   
## 6 sim1\_mod 6 16.5   
## 7 sim1\_mod 7 18.6   
## 8 sim1\_mod 8 20.6   
## 9 sim1\_mod 9 22.7   
## 10 sim1\_mod 10 24.7

sim1 %>% data\_grid(x) %>% spread\_predictions(sim1\_mod)

## # A tibble: 10 × 2  
## x sim1\_mod  
## <int> <dbl>  
## 1 1 6.27  
## 2 2 8.32  
## 3 3 10.4   
## 4 4 12.4   
## 5 5 14.5   
## 6 6 16.5   
## 7 7 18.6   
## 8 8 20.6   
## 9 9 22.7   
## 10 10 24.7

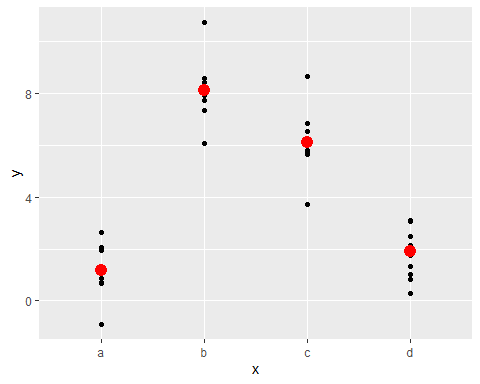
All three of these function do the same thing by giving a prediction to each x value. The only difference is that the gather\_predictions function also specifies what model is used.

# Section 23.4.5

# Exercise 1

What happens if you repeat the analysis of sim2 using a model without an intercept. What happens to the model equation?

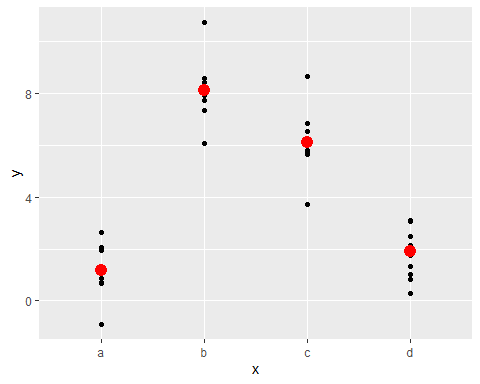
mod2 <- lm(y ~ 0 + x, data = sim2)  
grid <- sim2 %>% data\_grid(x) %>% add\_predictions(mod2)  
ggplot(sim2, aes(x)) + geom\_point(aes(y = y)) + geom\_point(data = grid, aes(y = pred), colour = "red", size = 4)



A 0 is added to the model equation to use a model without an intercept.

What happens to the predictions?

mod2 <- lm(y ~ x, data = sim2)  
grid <- sim2 %>% data\_grid(x) %>% add\_predictions(mod2)  
ggplot(sim2, aes(x)) + geom\_point(aes(y = y)) + geom\_point(data = grid, aes(y = pred), colour = "red", size = 4)

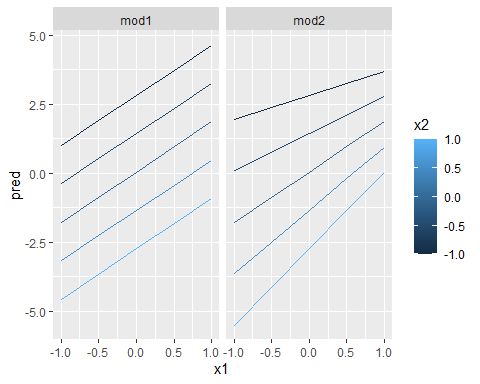


Using a model without an intercept for sim2 does not seem to affect the predictions.

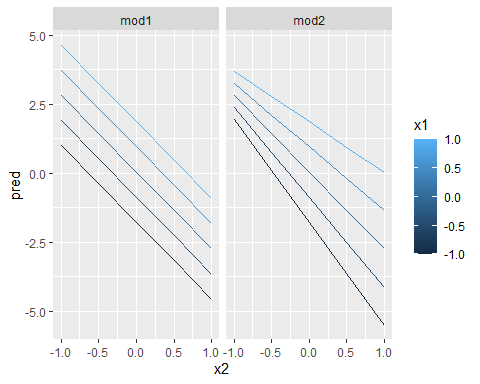
# Exercise 4

For sim4, which of mod1 and mod2 is better? I think mod2 does a slightly better job at removing patterns, but it’s pretty subtle. Can you come up with a plot to support my claim?

mod1 <- lm(y ~ x1 + x2, data = sim4)  
mod2 <- lm(y ~ x1 \* x2, data = sim4)  
grid <- sim4 %>% data\_grid(x1 = seq\_range(x1, 5), x2 = seq\_range(x2, 5)) %>% gather\_predictions(mod1, mod2)  
ggplot(grid, aes(x1, pred, colour = x2, group = x2)) + geom\_line() + facet\_wrap(~ model)



ggplot(grid, aes(x2, pred, colour = x1, group = x1)) + geom\_line() + facet\_wrap(~ model)



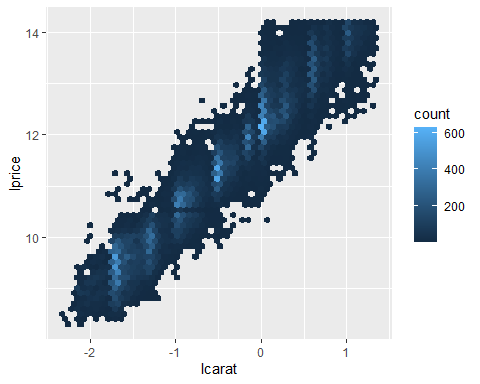
The claim that mod2 is better than mod1 is correct because the lines in mod1 always have the same slope while the lines for mod2 have different slopes which is more accurate.

# Section 24.2.3

# Exercise 1

In the plot of lcarat vs. lprice, there are some bright vertical strips. What do they represent?

diamonds2 <- diamonds %>% filter(carat <= 2.5) %>% mutate(lprice = log2(price), lcarat = log2(carat))  
ggplot(diamonds2, aes(lcarat, lprice)) + geom\_hex(bins = 50)

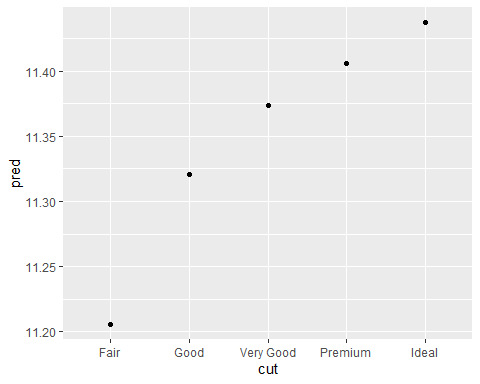


These vertical stripes represent the most common combinations of carat and price.

# Exercise 4

Does the final model, mod\_diamond2, do a good job of predicting diamond prices? Would you trust it to tell you how much to spend if you were buying a diamond?

mod\_diamond2 <- lm(lprice ~ lcarat + color + cut + clarity, data = diamonds2)  
grid <- diamonds2 %>% data\_grid(cut, .model = mod\_diamond2) %>% add\_predictions(mod\_diamond2)  
ggplot(grid, aes(cut, pred)) + geom\_point()



This model does seem to be accurate because better cuts should have higher prices. I do not think I would completely trust it though as I would want to do my own research.

# Section 24.3.5

# Exercise 1

Use your Google sleuthing skills to brainstorm why there were fewer than expected flights on Jan 20, May 26, and Sep 1. (Hint: they all have the same explanation.) How would these days generalise to another year?

daily <- flights %>% mutate(date = make\_date(year, month, day)) %>% group\_by(date) %>% summarise(n = n())  
daily <- daily %>% mutate(wday = wday(date, label = TRUE))  
mod <- lm(n ~ wday, data = daily)  
grid <- daily %>% data\_grid(wday) %>% add\_predictions(mod, "n")  
daily <- daily %>% add\_residuals(mod)  
filter(daily, date == "2013-01-20" | date == "2013-05-26" | date == "2013-09-01")

## # A tibble: 3 × 4  
## date n wday resid  
## <date> <int> <ord> <dbl>  
## 1 2013-01-20 786 Sun -105.  
## 2 2013-05-26 729 Sun -162.  
## 3 2013-09-01 718 Sun -173.

Each of these three dates were the Sundays before holidays. January 20 was the day before MLK Day, May 26 was the day before Memorial Day, and September 1 was the day before Labor Day. Since the dates of these holidays changes each year, these dates would also change.

# Exercise 2

What do the three days with high positive residuals represent? How would these days generalise to another year?

daily %>% slice\_max(n = 3, resid)

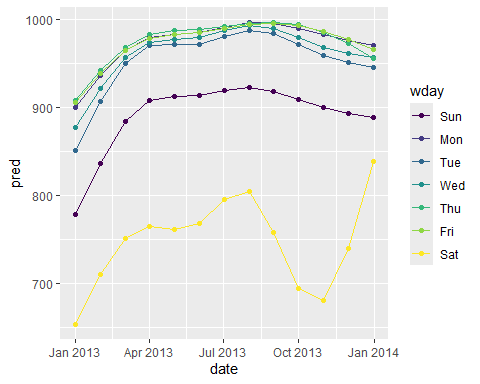
## # A tibble: 3 × 4  
## date n wday resid  
## <date> <int> <ord> <dbl>  
## 1 2013-11-30 857 Sat 112.   
## 2 2013-12-01 987 Sun 95.5  
## 3 2013-12-28 814 Sat 69.4

These are the days that have the most flights above expected. These days are the weekend after Thanksgiving and the Saturday after Christmas. The exact dates would probably change each year based on what date Thanksgiving falls on and what day of the week Christmas falls on.

# Exercise 6

What would you expect the model n ~ wday + ns(date, 5) to look like? Knowing what you know about the data, why would you expect it to be not particularly effective?

mod <- MASS::rlm(n ~ wday \* ns(date, 5), data = daily)  
daily %>% data\_grid(wday, date = seq\_range(date, n = 13)) %>% add\_predictions(mod) %>% ggplot(aes(date, pred, colour = wday)) + geom\_line() + geom\_point()

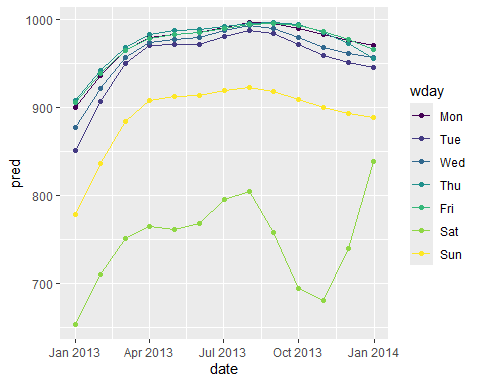


I expected this to show a strong pattern for Saturdays which turned out to be accurate. However, this model is not very effective because it is hard to differentiate between the weekdays.

# Exercise 8

It’s a little frustrating that Sunday and Saturday are on separate ends of the plot. Write a small function to set the levels of the factor so that the week starts on Monday.

daily %>% mutate(wday = factor(wday, levels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"))) %>% data\_grid(wday, date = seq\_range(date, n = 13)) %>% add\_predictions(mod) %>% ggplot(aes(date, pred, colour = wday)) + geom\_line() + geom\_point()

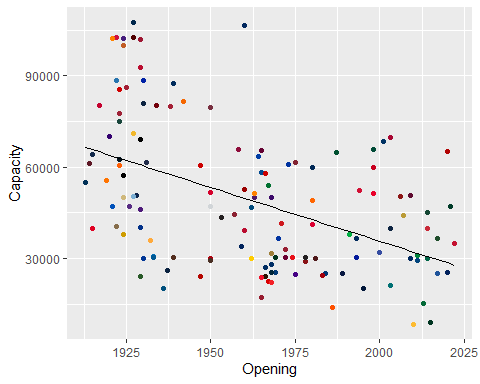


The mutate and factor functions can be used to move Sunday to the end, so now the week starts on Monday.

# FBS Stadiums

This data set contains data on all NCAA Division 1 Football Bowl Subdivision (FBS) schools during the 2024 college football season. The data set was originally downloaded from collegefootballdata.com as a csv file, and I tidied it up in assignment 4 and uploaded it as a new csv file. The read\_csv function is used to transfer this data set into a data frame. The following plot shows all FBS stadiums based on opening date and capacity with a linear model that shows the general trend of how opening date affects capacity.

fbs <- read\_csv("fbs2.csv", show\_col\_types = FALSE)  
mod <- lm(Capacity ~ Opening, data = fbs)  
grid <- fbs %>% data\_grid(Opening) %>% add\_predictions(mod)  
ggplot(data = fbs) + geom\_point(mapping = aes(x = Opening, y = Capacity, color = Color1)) + geom\_line(data = grid, aes(x = Opening, y = pred)) + scale\_color\_identity()

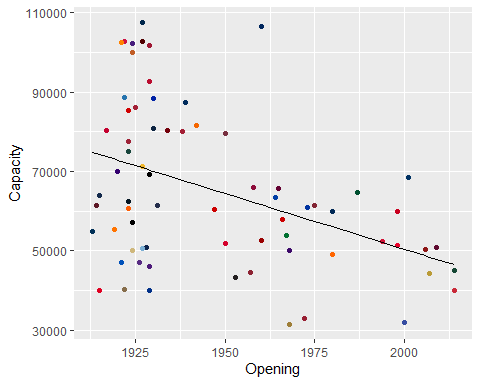


As can be seen, newer stadiums tend to have smaller capacities than older ones. This makes sense as older stadiums have had time to have several expansions throughout the years. Another factor is that older stadiums usually belong to the schools who have had football the longest and therefore have the largest established fan bases. However, the data is clearly very spread out since FBS stadiums capacities can range anywhere from 10,000 all the way to 110,000. This is partially because Power 5 schools tend to have larger stadiums than Group of 5 schools. To get a closer look at this trend, we will look at Power 5 and Group of 5 schools separately.

# Power 5

The Power 5 schools include those in the ACC, Big 12, Big Ten, Pac-12, SEC, and the independent Notre Dame. The Power 5 is considered to be the highest level of college football, so these schools tend to have larger stadiums. The following plot shows all Power 5 stadiums based on opening date and capacity with a linear model that shows the general trend of how opening date affects capacity.

p5 = filter(fbs, Conference == "ACC" | Conference == "Big 12" | Conference == "Big Ten" | Conference == "Pac-12" | Conference == "SEC" | School == "Notre Dame")  
mod <- lm(Capacity ~ Opening, data = p5)  
grid <- p5 %>% data\_grid(Opening) %>% add\_predictions(mod)  
ggplot(data = p5) + geom\_point(mapping = aes(x = Opening, y = Capacity, color = Color1)) + geom\_line(data = grid, aes(x = Opening, y = pred)) + scale\_color\_identity()



As can be seen, the Power 5 stadiums follow the general trend of newer stadiums being smaller than older ones. The only notable difference between this plot and the previous plot is that the average capacity is a bit higher. To get a closer look at Power 5 stadiums, we can look at the ten largest and smallest power 5 stadiums.

select(arrange(p5, desc(Capacity)), School, Conference, Stadium, Capacity, Opening)

## # A tibble: 70 × 5  
## School Conference Stadium Capacity Opening  
## <chr> <chr> <chr> <dbl> <dbl>  
## 1 Michigan Big Ten Michigan Stadium 107601 1927  
## 2 Penn State Big Ten Beaver Stadium 106572 1960  
## 3 Ohio State Big Ten Ohio Stadium 102780 1922  
## 4 Texas A&M SEC Kyle Field 102733 1927  
## 5 Tennessee SEC Neyland Stadium 102455 1921  
## 6 LSU SEC Tiger Stadium (LA) 102321 1924  
## 7 Alabama SEC Bryant-Denny Stadium 101821 1929  
## 8 Texas SEC DKR-Texas Memorial Stadium 100119 1924  
## 9 Georgia SEC Sanford Stadium 92746 1929  
## 10 UCLA Big Ten Rose Bowl 88565 1922  
## # ℹ 60 more rows

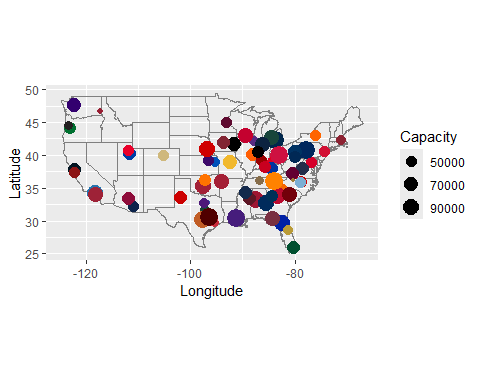
As shown in the above table, Beaver Stadium is the only one of the ten largest Power 5 stadiums to open since 1930. This makes it a bit of an outlier, but it still opened over 60 years ago. There are some other interesting things that can be seen from this data. First of all, all of these stadiums are in either the Big Ten or SEC. This makes sense as they are the two most prestigious conferences in college football. Speaking of prestige, these are ten of the most prestigious schools in college football. All of them have at least one national championship, and all of them except for UCLA have at least three. Another anomaly with UCLA is that they are the only school on this list that does not play on campus. Their home stadium, the Rose Bowl, is more notably known as the home of the famous bowl game of the same name.

select(arrange(p5, Capacity), School, Conference, Stadium, Capacity, Opening)

## # A tibble: 70 × 5  
## School Conference Stadium Capacity Opening  
## <chr> <chr> <chr> <dbl> <dbl>  
## 1 Wake Forest ACC Allegacy Federal Credit Union S… 31500 1968  
## 2 SMU ACC Gerald J. Ford Stadium 32000 2000  
## 3 Washington State Pac-12 Gesa Field 32952 1972  
## 4 Cincinnati Big 12 Nippert Stadium 40000 1915  
## 5 Houston Big 12 TDECU Stadium 40000 2014  
## 6 Duke ACC Wallace Wade Stadium 40004 1929  
## 7 Vanderbilt SEC FirstBank Stadium 40351 1922  
## 8 Oregon State Pac-12 Reser Stadium 43363 1953  
## 9 UCF Big 12 FBC Mortgage Stadium 44206 2007  
## 10 Boston College ACC Alumni Stadium (Chestnut Hill, … 44500 1957  
## # ℹ 60 more rows

As shown in the table above, the ten smallest Power 5 stadiums tend to be a bit newer with three of them opening since 2000 and only three of them opening before 1930. The conference makeup is also very different from the previous list. Vanderbilt is the only SEC school on the list while there are no Big Ten teams on the list. Both Pac-12 schools are on here which makes sense as they are among the least prestigious Power 5 schools and are on the verge of losing Power 5 status. Besides the two Pac-12 schools, the rest of the schools are either recent additions to Power 5 conferences, relatively small private schools, or both in the case of SMU. One thing these stadiums have in common with most of the largest stadiums is that they are all on campus stadiums.

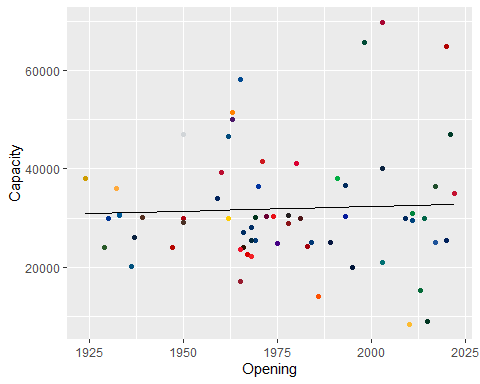
p5 %>% ggplot(aes(Longitude, Latitude, color = Color1, size = Capacity)) + borders("state") + geom\_point() + coord\_quickmap() + scale\_color\_identity()



# Group of 5

The Group of 5 schools include those in the American, Conference USA, MAC, MWC, Sun Belt, and the independents UCONN and UMASS. The Group of 5 is considered to be below the Power 5, so these schools tend to have smaller stadiums. The following plot shows all Group of 5 stadiums based on opening date and capacity with a linear model that shows the general trend of how opening date affects capacity.

go5 = filter(fbs, Conference == "American Athletic" | Conference == "Conference USA" | Conference == "Mid-American" | Conference == "Mountain West" | Conference == "Sun Belt East" | Conference == "Sun Belt West" | School == "Massachusetts" | School == "UConn")  
mod <- lm(Capacity ~ Opening, data = go5)  
grid <- go5 %>% data\_grid(Opening) %>% add\_predictions(mod)  
ggplot(data = go5) + geom\_point(mapping = aes(x = Opening, y = Capacity, color = Color1)) + geom\_line(data = grid, aes(x = Opening, y = pred)) + scale\_color\_identity()



Surprisingly, Group of 5 stadiums do not follow the same trend as the previous two plots. Instead, the capacities of Group of 5 stadiums has remained rather steady throughout the years with newer stadiums actually being slightly larger than older ones. This is certainly an interesting contrast, and it begs the question of why the trend is different. To get a closer look at why this is, we can look at the ten largest and smallest Group of 5 stadiums.

select(arrange(go5, desc(Capacity)), School, Conference, Stadium, Capacity, Opening)

## # A tibble: 64 × 5  
## School Conference Stadium Capacity Opening  
## <chr> <chr> <chr> <dbl> <dbl>  
## 1 Temple American Athletic Lincoln Financial Field 69879 2003  
## 2 South Florida American Athletic Raymond James Stadium 65828 1998  
## 3 UNLV Mountain West Allegiant Stadium 65000 2020  
## 4 Memphis American Athletic Simmons Bank Liberty Stadium 58325 1965  
## 5 UTEP Conference USA Sun Bowl 51500 1963  
## 6 East Carolina American Athletic Dowdy-Ficklen Stadium 50000 1963  
## 7 UAB American Athletic Protective Stadium 47100 2021  
## 8 Rice American Athletic Rice Stadium 47000 1950  
## 9 Air Force Mountain West Falcon Stadium 46692 1962  
## 10 Louisiana Sun Belt West Cajun Field 41426 1971  
## # ℹ 54 more rows

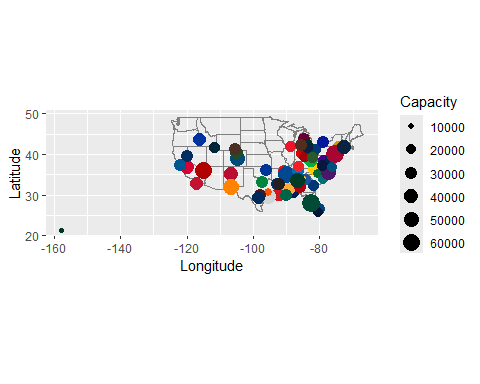
As shown in the table above, the ten largest Group of 5 stadiums are much newer than the ten largest Power 5 stadiums with three of them opening since 2000 and all of them opening since 1950. The reasoning behind this becomes clear with a bit of research. The three largest stadiums on here are all NFL stadiums with all of them opening since 1998. There is no doubt that these three stadiums skew the linear model quite a bit. Memphis and UAB also play in off campus stadiums with UAB having the newest stadium on this list. The on campus stadiums on here are much older by comparison although they are still not as old as the Power 5 stadiums. This makes sense as Group of 5 schools tend to have less established football programs.

select(arrange(go5, Capacity), School, Conference, Stadium, Capacity, Opening)

## # A tibble: 64 × 5  
## School Conference Stadium Capacity Opening  
## <chr> <chr> <chr> <dbl> <dbl>  
## 1 Kennesaw State Conference USA Fifth Third Stadium 8318 2010  
## 2 Hawai'i Mountain West Clarence T.C. Ching… 9000 2015  
## 3 Sam Houston Conference USA Elliott T. Bowers S… 14000 1986  
## 4 Charlotte American Athletic Jerry Richardson St… 15300 2013  
## 5 Massachusetts FBS Independents Warren McGuirk Alum… 17000 1965  
## 6 Florida International Conference USA FIU Stadium 20000 1995  
## 7 Old Dominion Sun Belt East S.B. Ballard Stadium 20118 1936  
## 8 Coastal Carolina Sun Belt East Brooks Stadium (SC) 21000 2003  
## 9 Western Kentucky Conference USA Houchens Industries… 22113 1968  
## 10 Ball State Mid-American Scheumann Stadium 22500 1967  
## # ℹ 54 more rows

As shown in the table above, S.B. Ballard Stadium is the only one of the ten smallest Group of 5 stadiums to open before 1965. This makes it a bit of an outlier especially because it is also older than all of the top ten largest Group of 5 stadiums. However, the rest of the stadiums are much newer with four of them opening since 2000. This makes sense as many of these schools are recent additions to the FBS. In fact, the smallest stadium on this list and, by extension, in the entire FBS belongs to Kennesaw State which is currently in its first year in the FBS. The second smallest stadium on this list is a temporary home for Hawaii while they try to build a new Aloha Stadium.

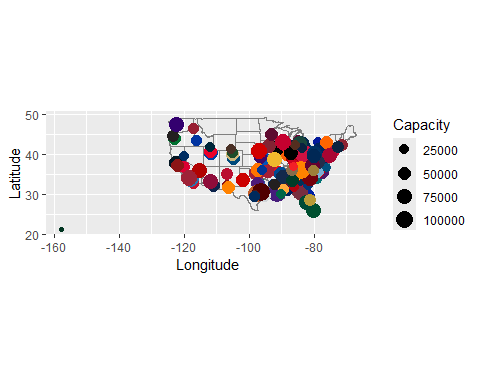
go5 %>% ggplot(aes(Longitude, Latitude, color = Color1, size = Capacity)) + borders("state") + geom\_point() + coord\_quickmap() + scale\_color\_identity()



# Conclusion

Based on this data, it is clear that newer FBS stadiums tend to be smaller than older ones. This trend holds true for Power 5 schools, but it is not so for Group of 5 schools. This contrast is the result of a few Group of 5 schools playing in relatively new NFL stadiums which are larger than other Group of 5 stadiums. It would certainly be interesting to see how the data trend would look if all off campus stadiums were removed. While it probably would not make the Group of 5 trend match the overall trend, it would most likely make it a bit closer.

fbs %>% ggplot(aes(Longitude, Latitude, color = Color1, size = Capacity)) + borders("state") + geom\_point() + coord\_quickmap() + scale\_color\_identity()



source: <https://collegefootballdata.com/exporter/teams/fbs?year=2024>