DV ASSIGN 3-5

Rachel Kaufman

2022-09-13

CHAPTER 3,4,5*

CHAPTER 3

loading it up!

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Read in the data
exercise_data <- read_csv("Data/visualize_data.csv")
```

```
## New names:
## Rows: 142 Columns: 4
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (4): ...1, ...2, Exercise, BMI
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘‘ -> ‘...1‘
## * ‘...1‘ -> ‘...2‘
```

```
glimpse(exercise_data)
```

```
## Rows: 142
## Columns: 4
## $ ...1     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ ...2     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ Exercise <dbl> 55.3846, 51.5385, 46.1538, 42.8205, 40.7692, 38.7179, 35.6410~
## $ BMI      <dbl> 1.8320590, 1.7892194, 1.7321050, 1.6178724, 1.5036362, 1.3751~
```

So, I would make the general assumption that the more you exercise the lower your BMI, but because BMI is known to suck, so I doubt there is much of a relationship.

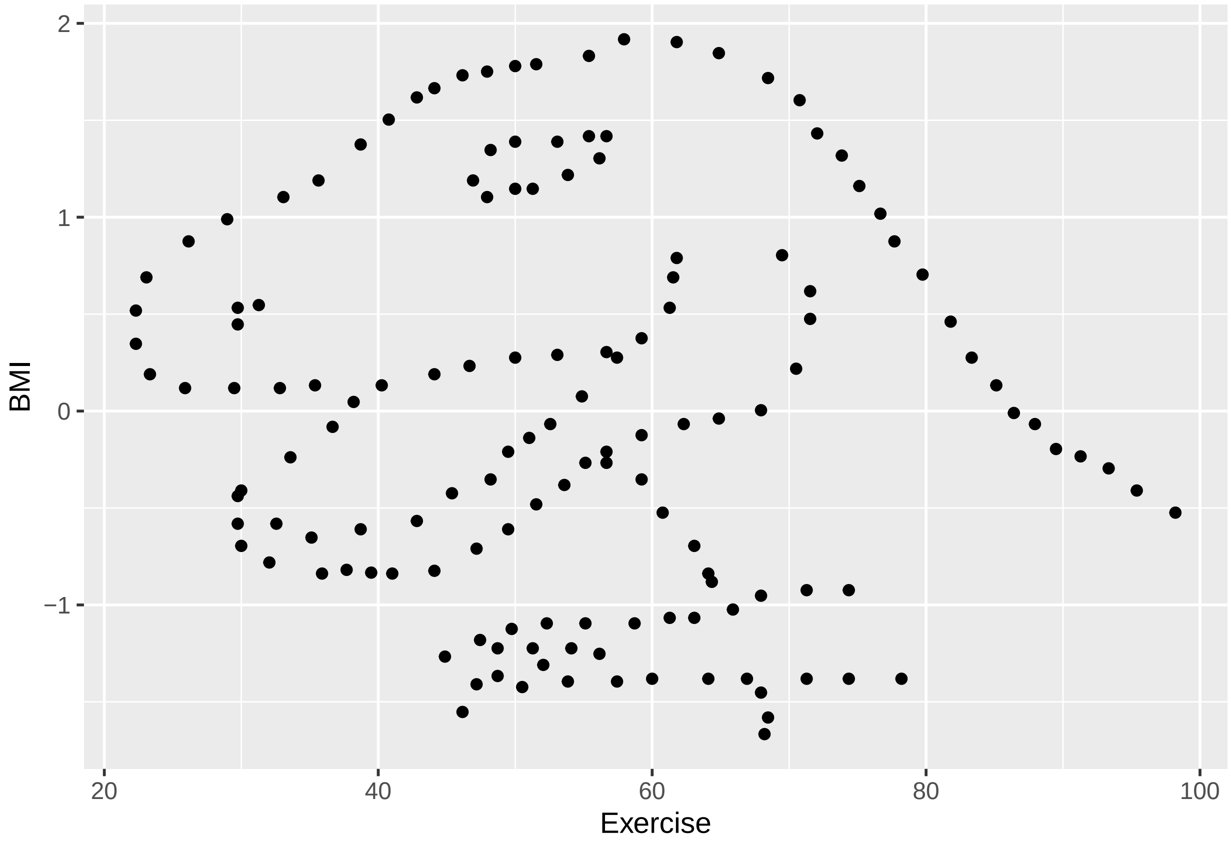
Question 1

```
cor(exercise_data$Exercise, exercise_data$BMI)
```

```
## [1] -0.06447185
```

This shows a correlation of -0.06, meaning their is little correlation between the two variables.

```
ggplot(exercise_data, aes(x = Exercise, y = BMI)) +
  geom_point()
```



Question 2

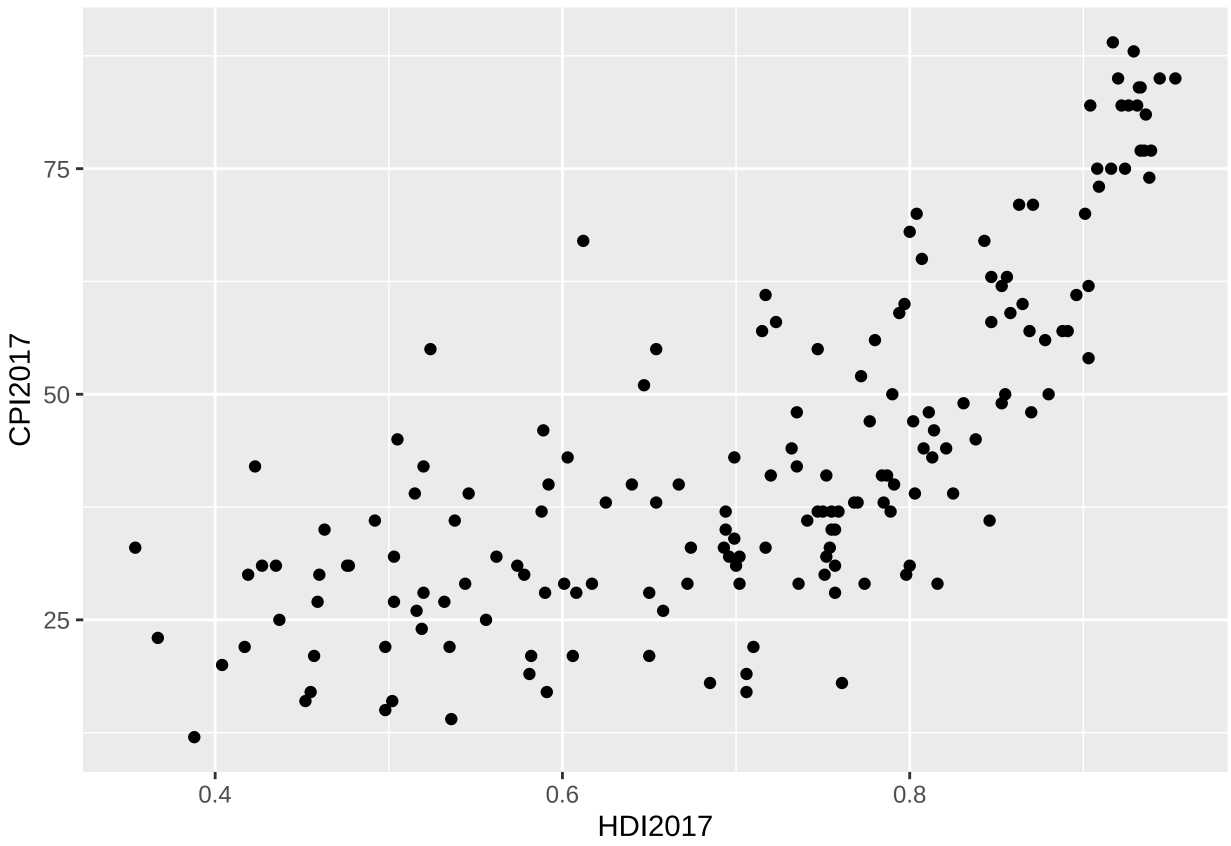
```
library(causact)
glimpse(corruptDF)
```

```
## Rows: 174
## Columns: 7
## $ country     <chr> "Afghanistan", "Albania", "Algeria", "Angola", "Argentina"~
## $ region      <chr> "Asia Pacific", "East EU Cemt Asia", "MENA", "SSA", "Ameri~
## $ countryCode <chr> "AFG", "ALB", "DZA", "AGO", "ARG", "ARM", "AUS", "AUT", "A~
## $ regionCode  <chr> "AP", "ECA", "MENA", "SSA", "AME", "ECA", "AP", "WE/EU", "~
## $ population  <int> 35530081, 2873457, 41318142, 29784193, 44271041, 2930450, ~
## $ CPI2017     <int> 15, 38, 33, 19, 39, 35, 77, 75, 31, 65, 36, 28, 68, 44, 75~
## $ HDI2017     <dbl> 0.498, 0.785, 0.754, 0.581, 0.825, 0.755, 0.939, 0.908, 0.~
```

The HDI is the human development index and CPI is the corruption perceptions index where each observation is a country. I'm going to assume 2017 is just the year the data is from.

Question 3

```
ggplot(corruptDF, aes(x = HDI2017, y = CPI2017)) +
  geom_point()
```

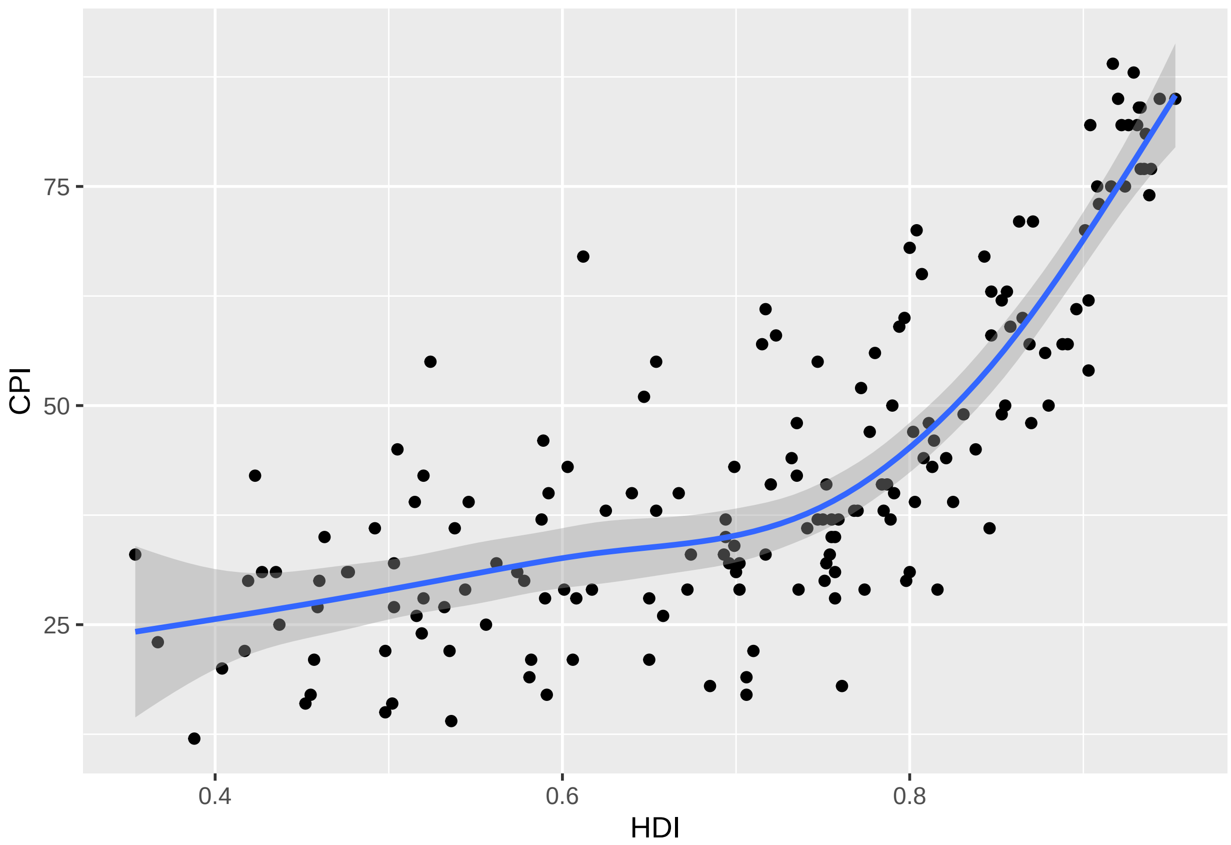


There definitely seems to be some sort of relationship between the CPI and HDI indices. There also seems to be non-linear curvature.

Question 4

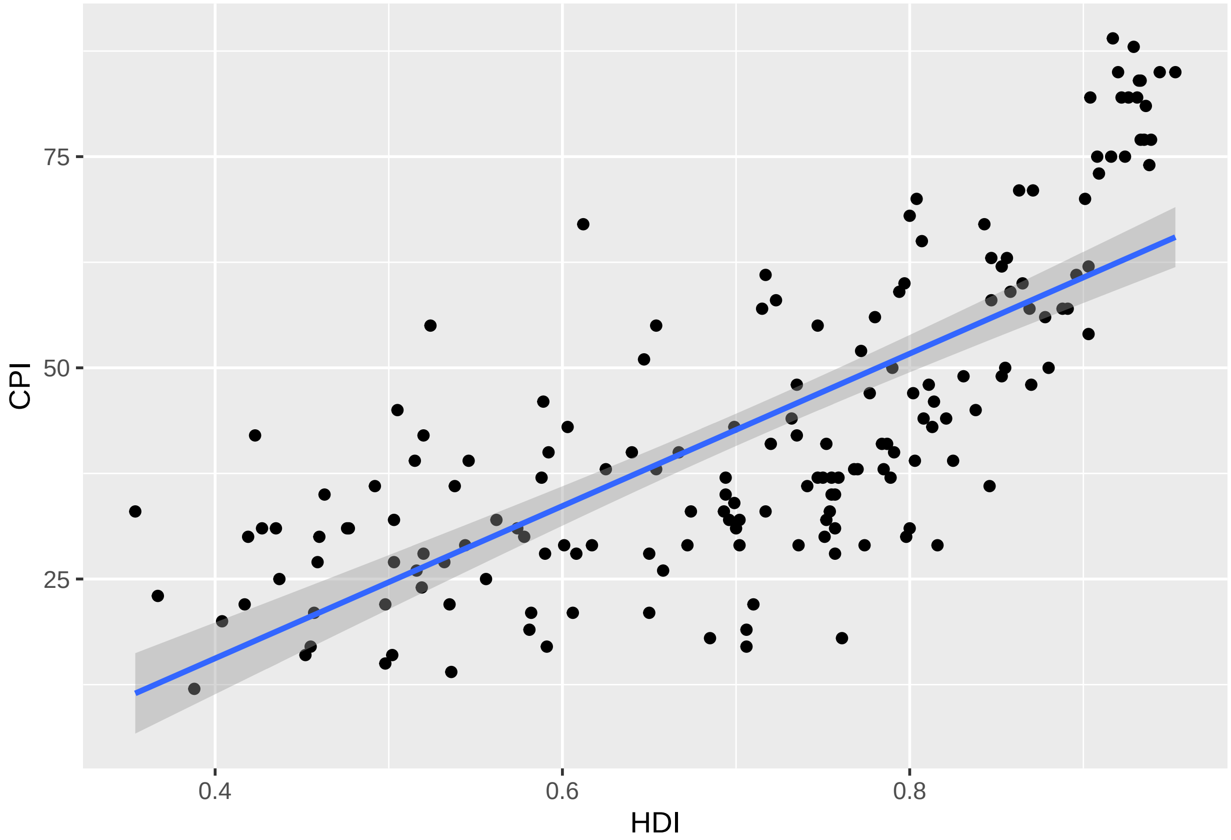
```
ggplot(corruptDF, aes(x = HDI2017, y = CPI2017)) +
  geom_point() +
  geom_smooth(method = "gam")+
  labs(x = "HDI",
       y = "CPI")
```

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```



```
?gam
ggplot(corruptDF, aes(x = HDI2017, y = CPI2017)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "HDI",
       y = "CPI")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

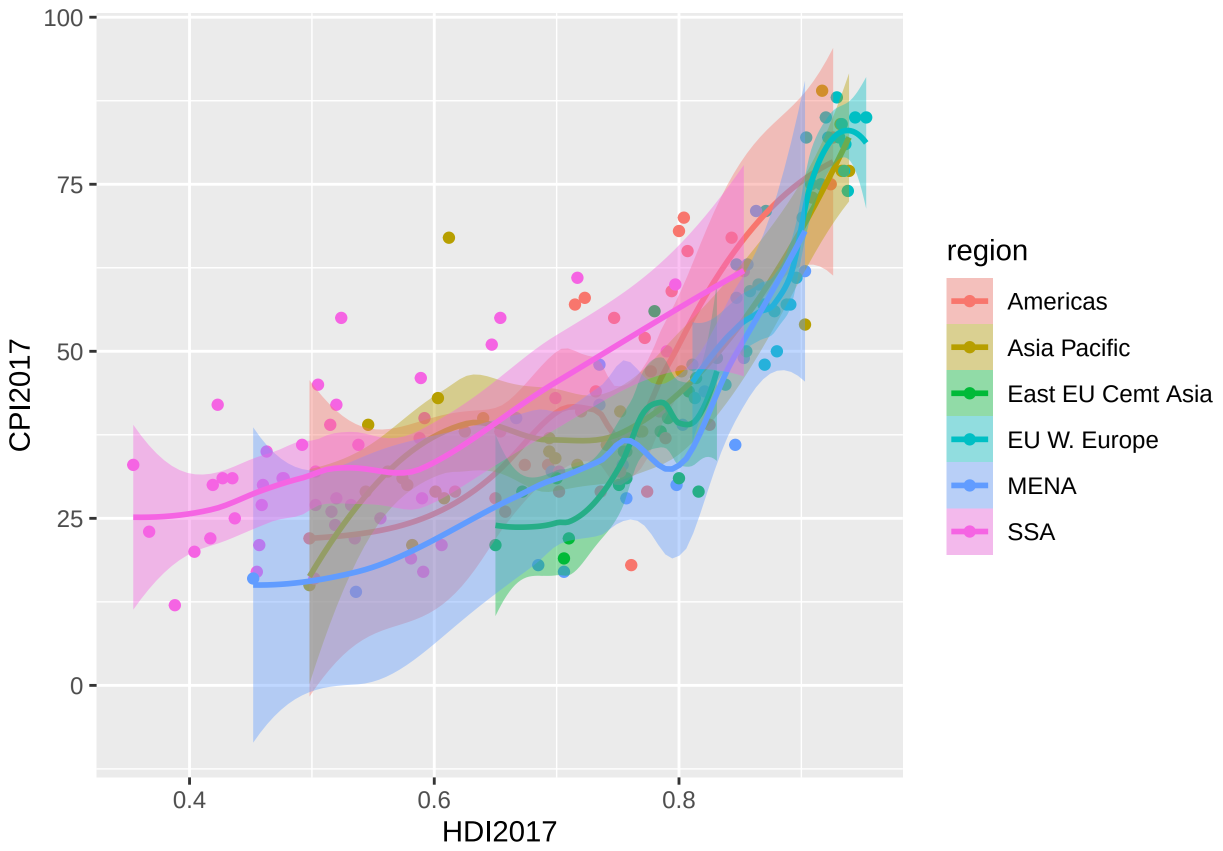


Granted I know nothing about GAM (general additive models?), I prefer the use of that function. By looking at the graph in comparison to the use of the lm method, it seems to not attribute equal weight in the way we see lm operating. I this this could be valuable just in looking at non-linearity, especially as there is a curvature to the scatter plot by itself. The differences visually are that lm uses "y~x" which produces a straight line and gam seems to look more like the scatter plot.

Question 5 It would be interesting to explore if this relationship varies by region. Add a fill and color aesthetic to the graph so that the lines and points are grouped by the variable region.

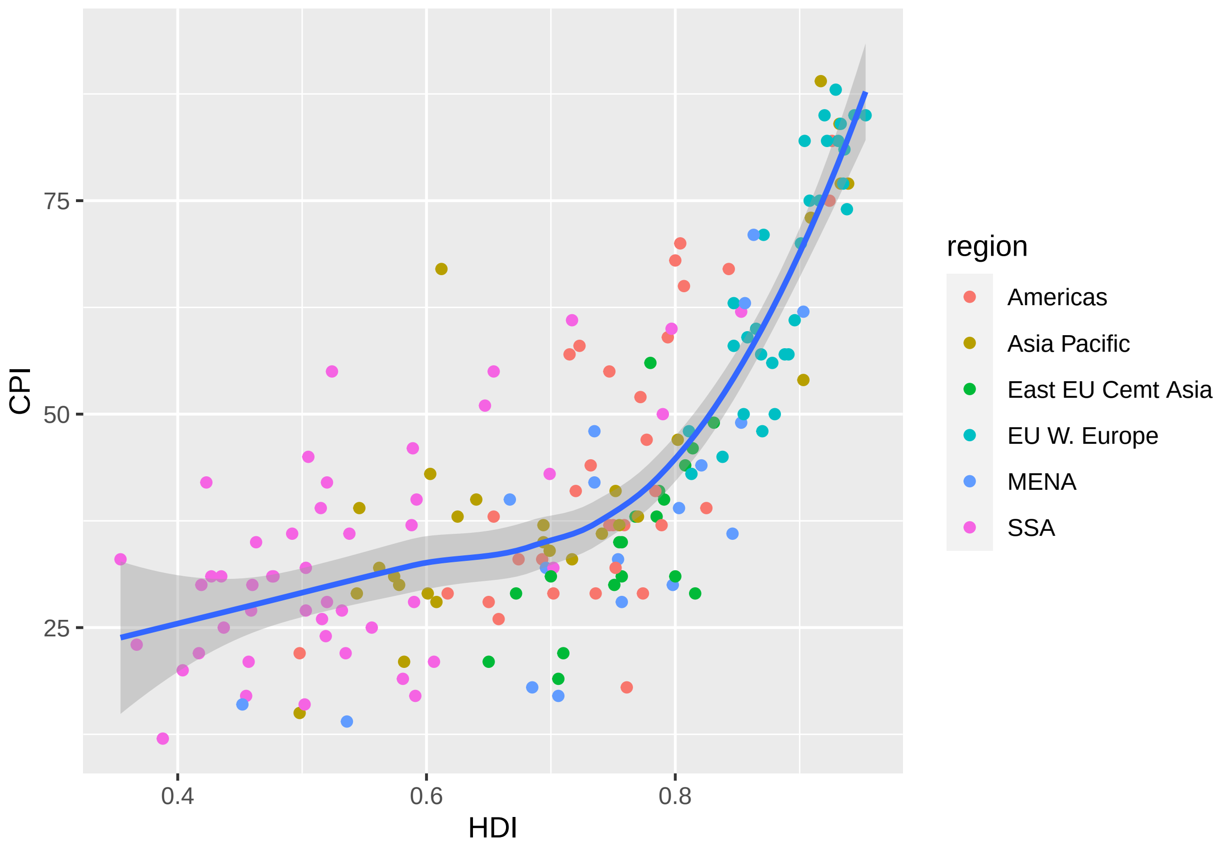
```
ggplot(corruptDF, aes(x = HDI2017, y = CPI2017, color = region, fill = region)) +
  geom_point() +
  geom_smooth() ##this looks veryyy messy, so i need to make the one line come back
```

```
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```



```
ggplot(corruptDF, aes(x = HDI2017, y = CPI2017)) +
  geom_point(mapping = aes(fill = region, color = region)) +
  geom_smooth() +
  labs(x = "HDI",
       y = "CPI")
```

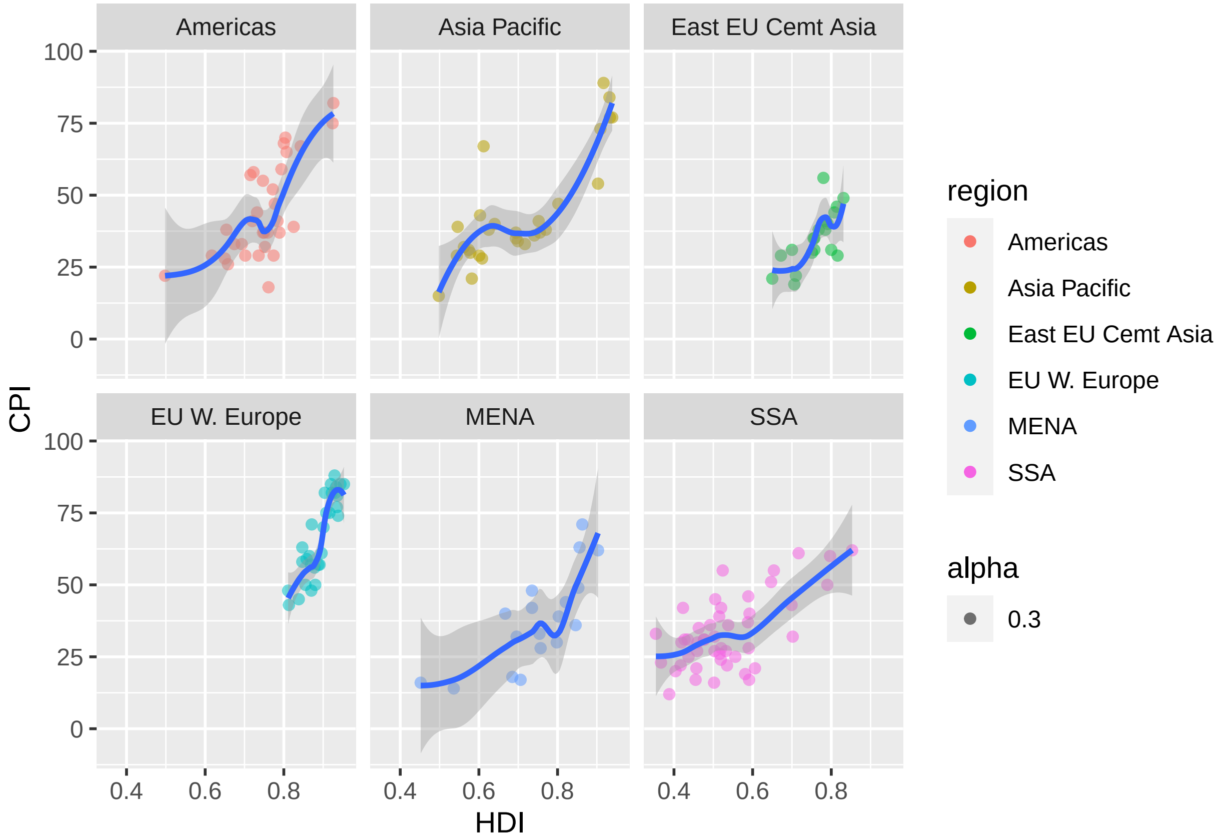
```
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```



I see that EU W. Europe countries are clustered towards the top end of CPI and HDI. In a similar sentiment, it seems that SSA counties are towards the bottom left of the graph. While splitting by region definitely allows me to see more, I don't super see anyother patterns that may of be importance at first glance.

```
ggplot(corruptDF, aes(x = HDI2017, y = CPI2017)) +
  geom_point(mapping = aes(color = region, fill = region, alpha = 0.3)) +
  geom_smooth() +
  facet_wrap("region") +
  labs(x = "HDI",
       y = "CPI")
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```

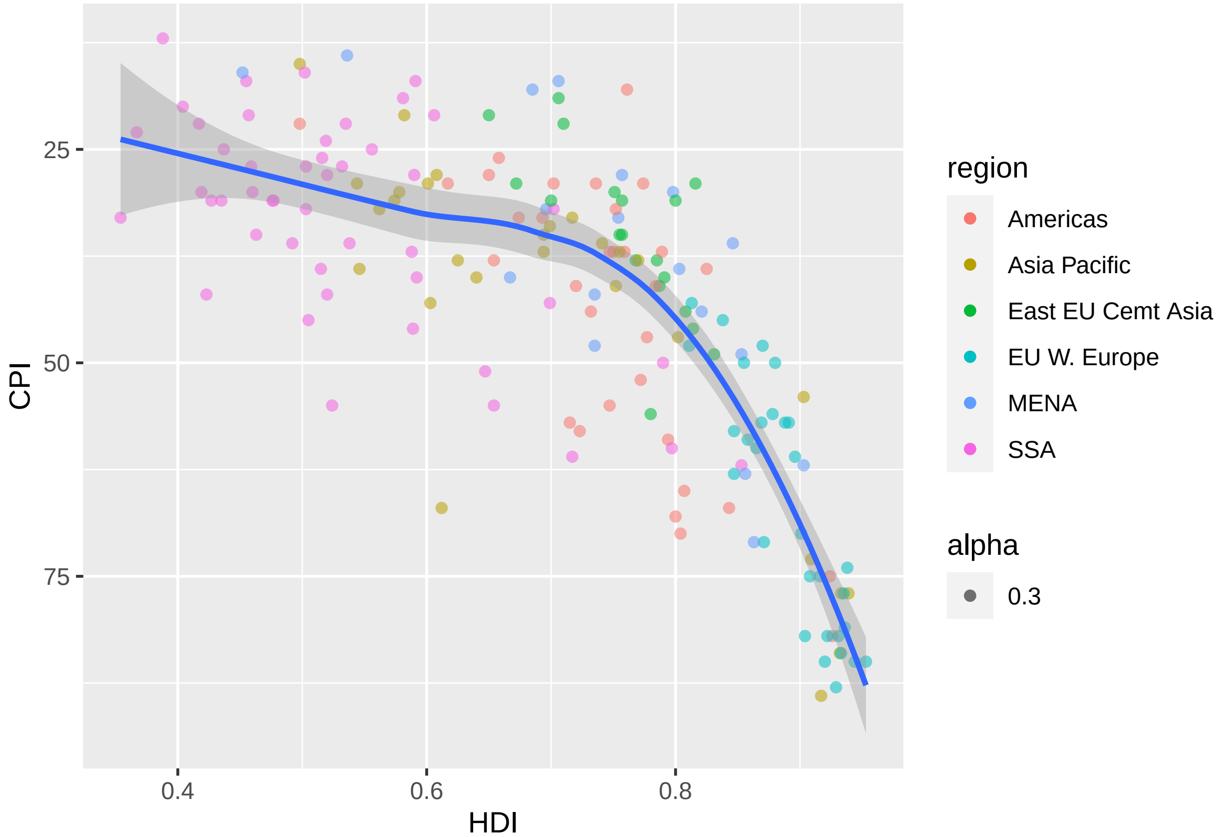


Region absolutely makes it look better, I think that making the scatter plot dots smaller and or more translucent could also potentially help with the cluttering.

Question 6

```
ggplot(corruptDF, aes(x = HDI2017, y = CPI2017)) +
  geom_point(mapping = aes(color = region, fill = region, alpha = 0.3)) +
  geom_smooth() +
  scale_y_reverse() +
  labs(x = "HDI",
       y = "CPI")
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



So I used scale_y_reverse to switch the the order of my Y-axis. It could be useful because I feel like it draws attention on CPI's with higher score and a higher HDI and makes like logical sense, as in damn the lower score means the more curropt, which could potentially help for reading comprehension.

Question 7

```
?corruptDF
ggplot_HDI_CPI <- ggplot(corruptDF, aes(x = HDI2017, y = CPI2017)) +
  geom_point(mapping = aes(color = region, fill = region, alpha = 0.3)) +
  geom_smooth() +
  scale_y_reverse() +
  labs(x = "HDI",
       y = "CPI",
       title = "The relationship between Human Deprevation Index (HDI) and
       Corruption Perception Index (CPI) indices",
       subtitle = "2017",
         caption = "Data: CPI available from Transparency International,
       HDI available from UN Development reports,
       Population data from World Bank.
       Accessed 2018.")
```

The title feels wordy, but eh here we are.

Question 8

```
##?ggsave
ggsave("my_ggplot.pdf", plot = ggplot_HDI_CPI)
```

```
## Saving 6.5 x 4.5 in image
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```

CHAPTER 4 Question 1

```
##tidyverse is already downloaded, see line 17.
tv_ratings <- read_csv("Data/tv_ratings.csv")
```

```
## Rows: 2266 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): titleId, title, genres
## dbl  (3): seasonNumber, av_rating, share
## date (1): date
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
glimpse(tv_ratings)
```

```
## Rows: 2,266
## Columns: 7
## $ titleId      <chr> "tt2879552", "tt3148266", "tt3148266", "tt3148266", "tt31~
## $ seasonNumber <dbl> 1, 1, 2, 3, 4, 1, 2, 1, 2, 3, 4, 5, 6, 7, 8, 1, 1, 1, 1, ~
## $ title        <chr> "11.22.63", "12 Monkeys", "12 Monkeys", "12 Monkeys", "12~
## $ date         <date> 2016-03-10, 2015-02-27, 2016-05-30, 2017-05-19, 2018-06-~
## $ av_rating    <dbl> 8.4890, 8.3407, 8.8196, 9.0369, 9.1363, 8.4370, 7.5089, 8~
## $ share        <dbl> 0.51, 0.46, 0.25, 0.19, 0.38, 2.38, 2.19, 6.67, 7.13, 5.8~
## $ genres       <chr> "Drama,Mystery,Sci-Fi", "Adventure,Drama,Mystery", "Adven~
```

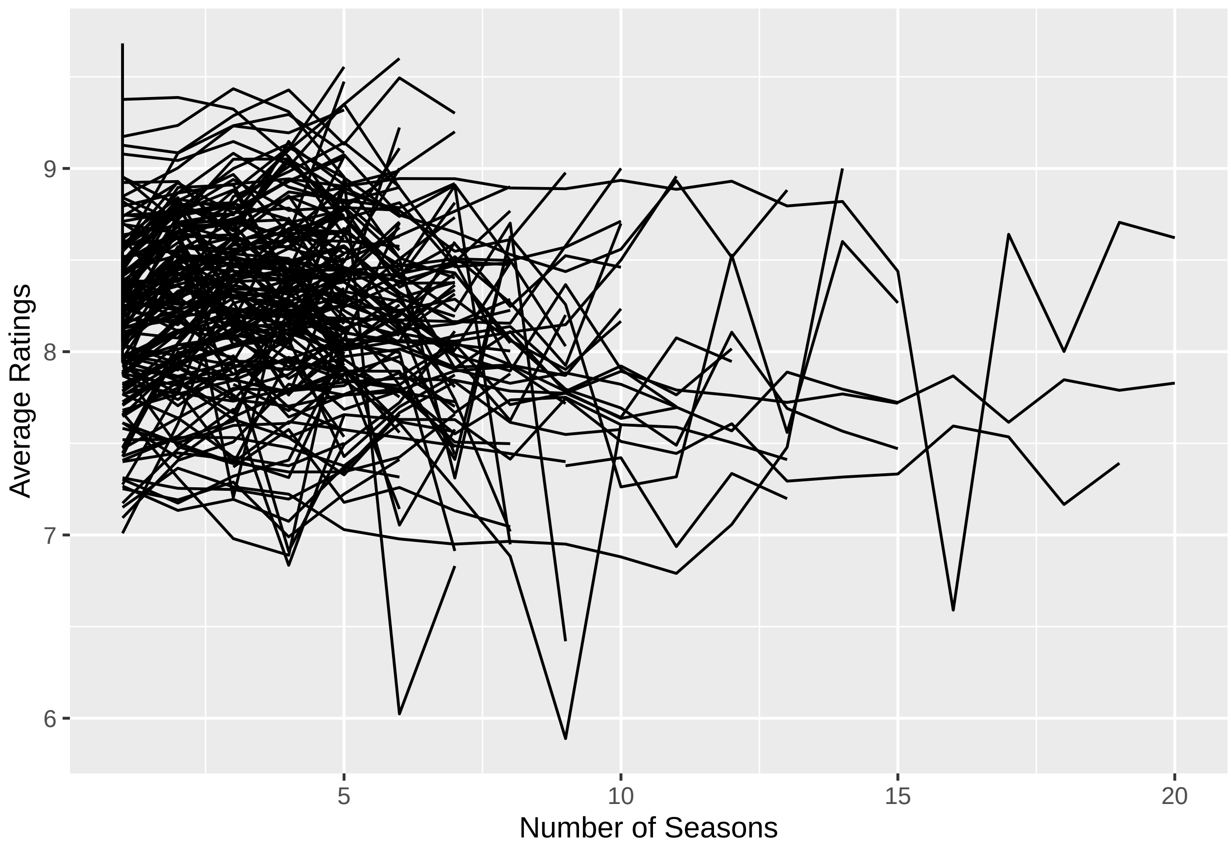
now i am fiddling with the data set, tidying her up

```
tv_long <- tv_ratings %>%
  group_by(title) %>%
  summarize(number_seasons = n()) %>%
  ungroup() %>%
  left_join(tv_ratings, by = "title")

tv_long <- tv_long %>% ##five or more seasons, this alters tv_long
  filter(number_seasons >= 5)
```

time to make a line plot for average ratings across seasons

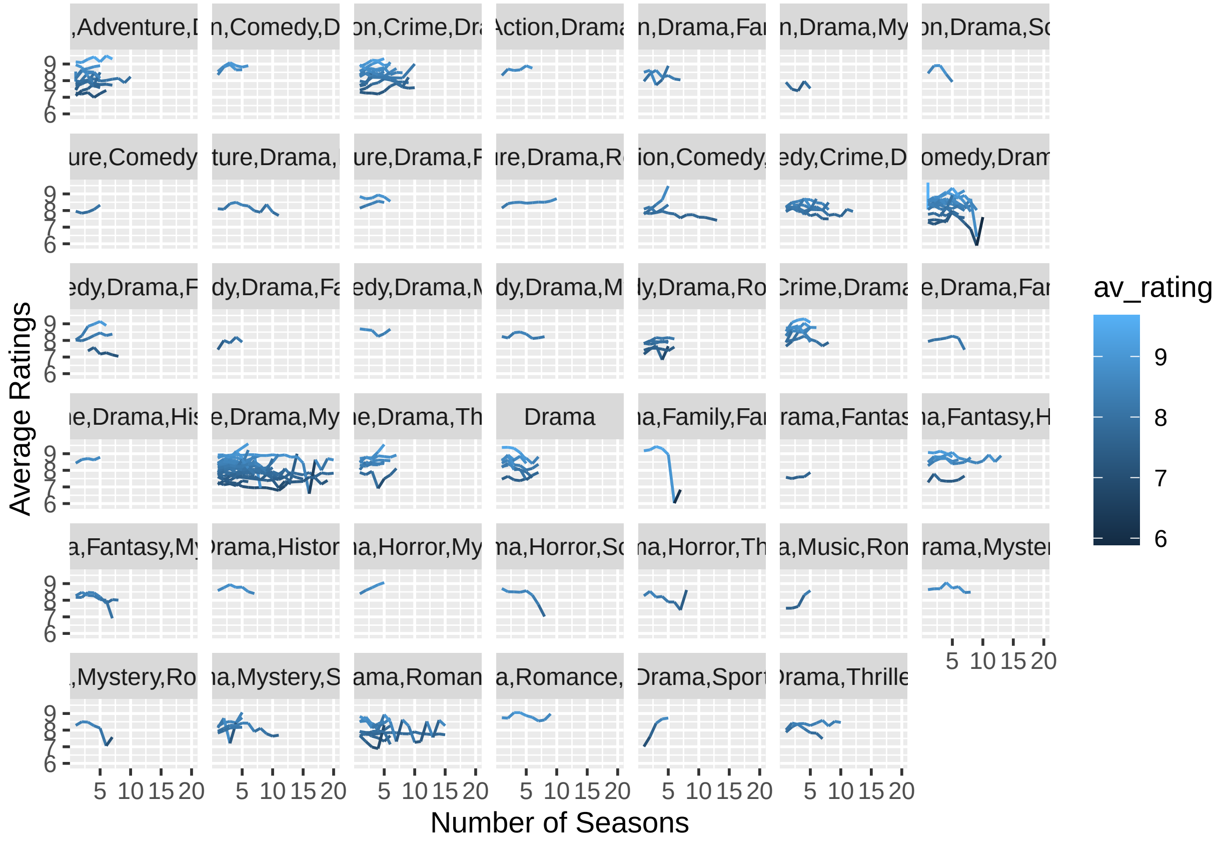
```
ggplot(tv_long, aes(x = seasonNumber, y = av_rating)) +
  geom_line(mapping = aes(group = title)) +
   labs(x = "Number of Seasons",
       y = "Average Ratings")
```



I'm going to say. . . . no I cannot come up with any conclusions of this ugly line plot.

Question 2

```
ggplot(tv_long, aes(x = seasonNumber, y = av_rating, color = av_rating)) +
  geom_line(mapping = aes(group = title)) +
  facet_wrap("genres") +
  labs(x = "Number of Seasons",
       y = "Average Ratings")
```



it seems that crime, drama, mystery and drama, romance last longer than other shows. There also seems to be a dip, then the show ends, in ratings around 5-7 seasons. That makes logisitcal sense, ya know? Show gets bad reviews, show ends. Interesting that drama sport eends on a high note, but there is only one plotted line so...

```
tv_long %>%
  filter(genres == "Drama,Family,Fantasy") %>%
  select(title)
```

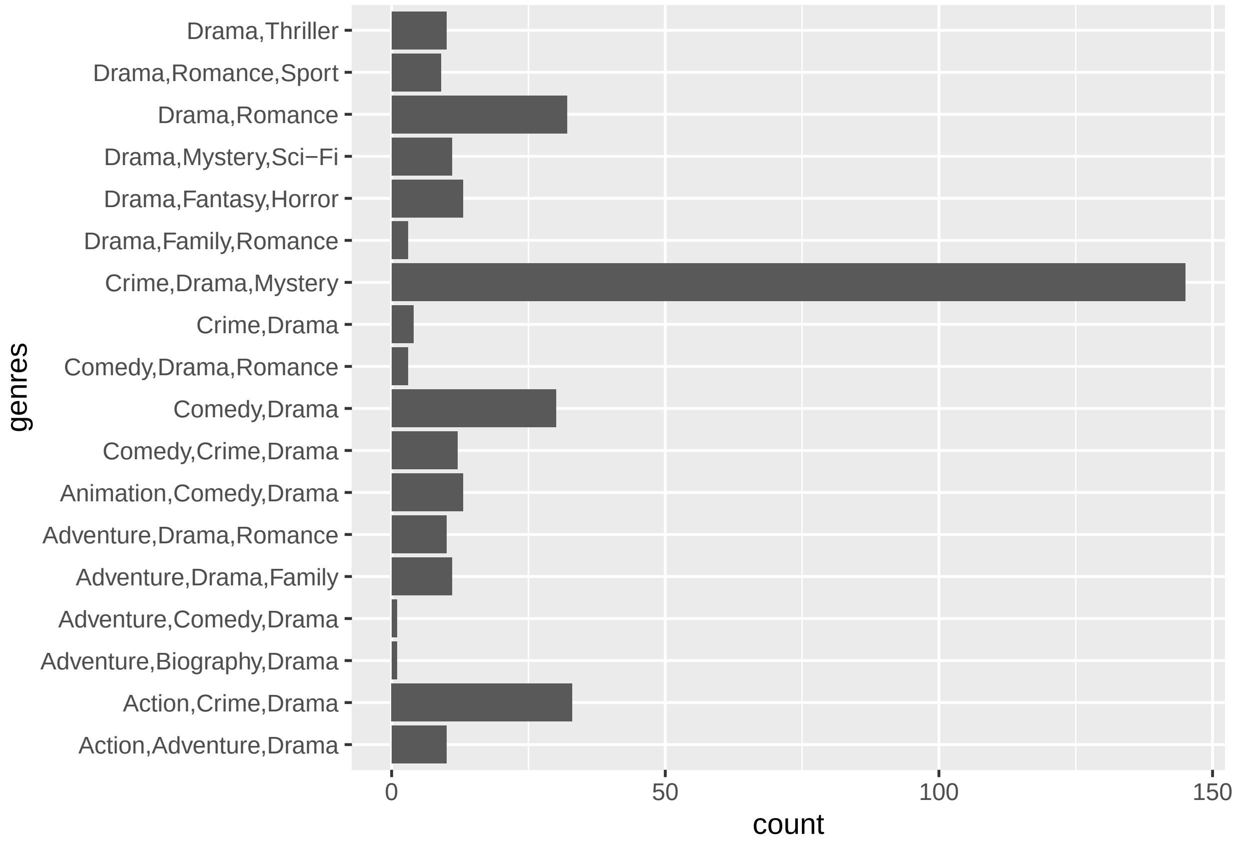
```
## # A tibble: 7 x 1
##   title
##   <chr>
## 1 Are You Afraid of the Dark?
## 2 Are You Afraid of the Dark?
## 3 Are You Afraid of the Dark?
## 4 Are You Afraid of the Dark?
## 5 Are You Afraid of the Dark?
## 6 Are You Afraid of the Dark?
## 7 Are You Afraid of the Dark?
```

It is the show, are you afraid of the dark? That had the plummeted ratings.

Question 3

```
top_tier_shows <- tv_ratings %>%
  group_by(title) %>%
  mutate(num_seasons = max(seasonNumber)) %>%
  filter(num_seasons >= 9) %>%
  ungroup()

ggplot(top_tier_shows, aes(x = genres)) +
  geom_bar() +
  coord_flip()
```



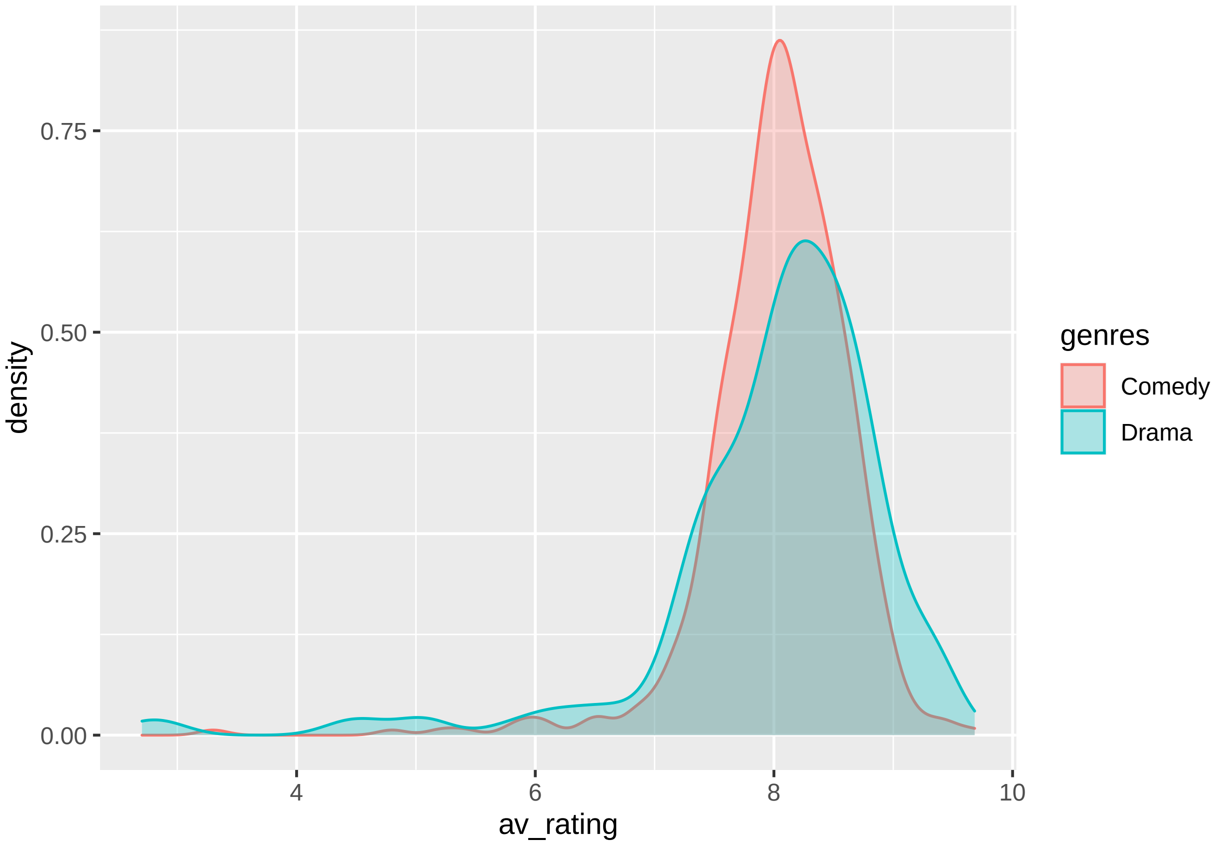
Crime, Drama, and Mystery!

Question 4

```
comedies_dramas <- tv_ratings %>%
  mutate(comedy_is = if_else(str_detect(genres, "Comedy"),
                              1, #the argument post commedy requires the next two entries, 1 for if it h
                              0)) %>%
           filter(comedy_is == 1 | genres == "Drama") %>%
           mutate(genres = if_else(genres == "Drama",
                                   "Drama",
                                   "Comedy"))
glimpse(comedies_dramas)
```

```
## Rows: 684
## Columns: 8
## $ titleId      <chr> "tt0312081", "tt0312081", "tt0312081", "tt1225901", "tt12~
## $ seasonNumber <dbl> 1, 2, 3, 1, 2, 3, 4, 5, 1, 2, 1, 25, 1, 1, 2, 3, 4, 5, 1,~
## $ title        <chr> "8 Simple Rules", "8 Simple Rules", "8 Simple Rules", "90~
## $ date         <date> 2002-09-17, 2003-11-04, 2004-11-12, 2009-01-03, 2009-11--~
## $ av_rating    <dbl> 7.5000, 8.6000, 8.4043, 7.1735, 7.4686, 7.6858, 6.8344, 7~
## $ share        <dbl> 0.03, 0.10, 0.06, 0.40, 0.14, 0.10, 0.04, 0.01, 0.48, 0.4~
## $ genres       <chr> "Comedy", "Comedy", "Comedy", "Comedy", "Comedy", "Comedy~
## $ comedy_is    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

```
ggplot(comedies_dramas, aes(x = av_rating, fill = genres, color = genres)) +
  geom_density(alpha = 0.3)
```

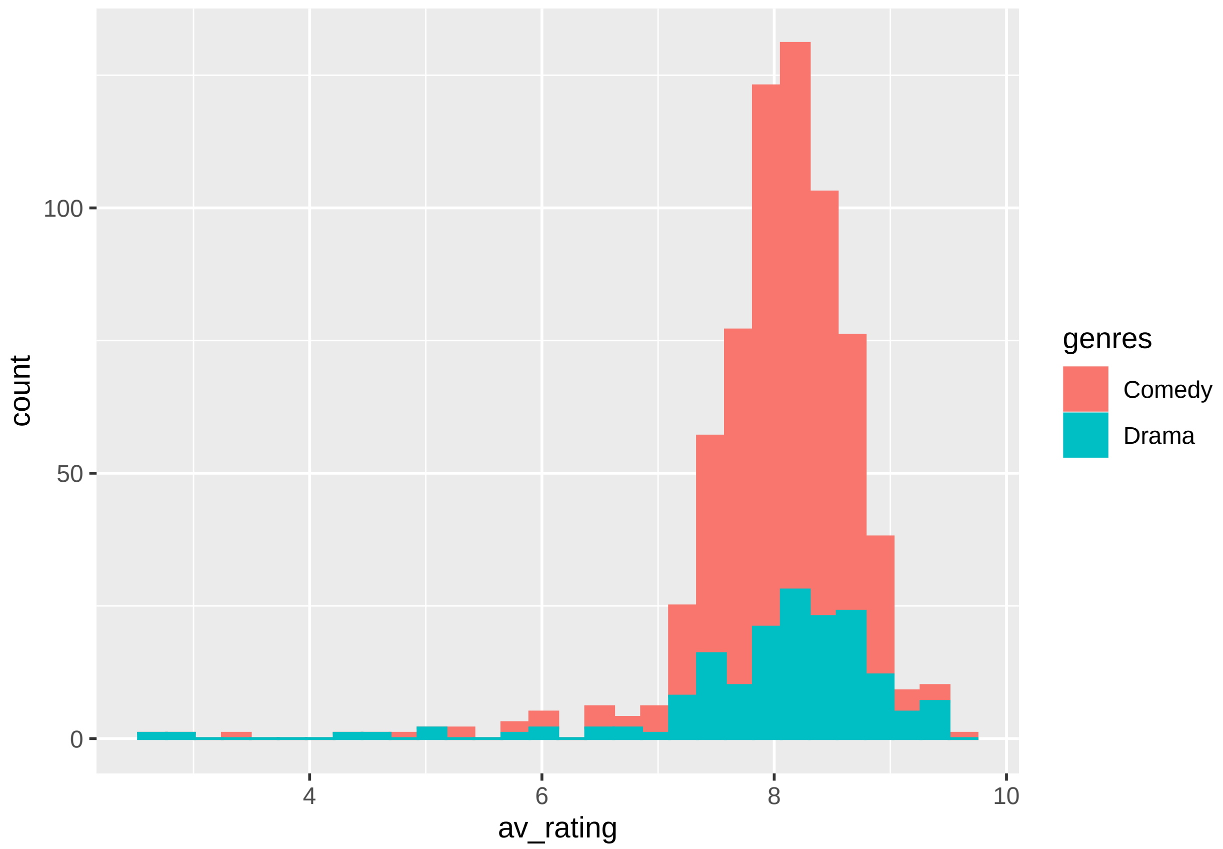


Comedy does have a greater density for shows with the average rating at 8 in comparison to dramas. I looks like dramas have more weight in the 9 and 10 areas of the scale in terms of having greater density though.

Question 5

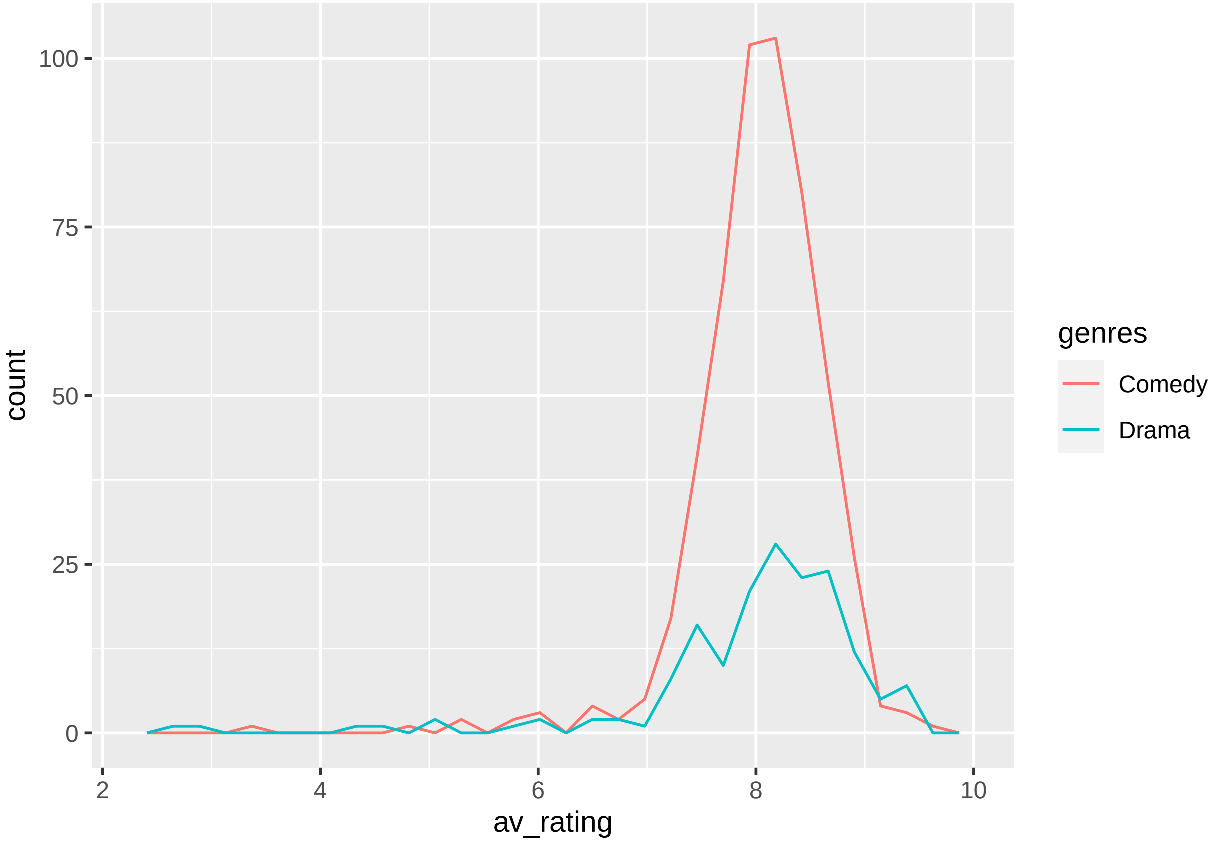
```
ggplot(comedies_dramas, aes(x = av_rating, fill = genres, color = genres)) +
  geom_histogram()
```

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



```
ggplot(comedies_dramas, aes(x = av_rating, fill = genres, color = genres)) +
  geom_freqpoly()
```

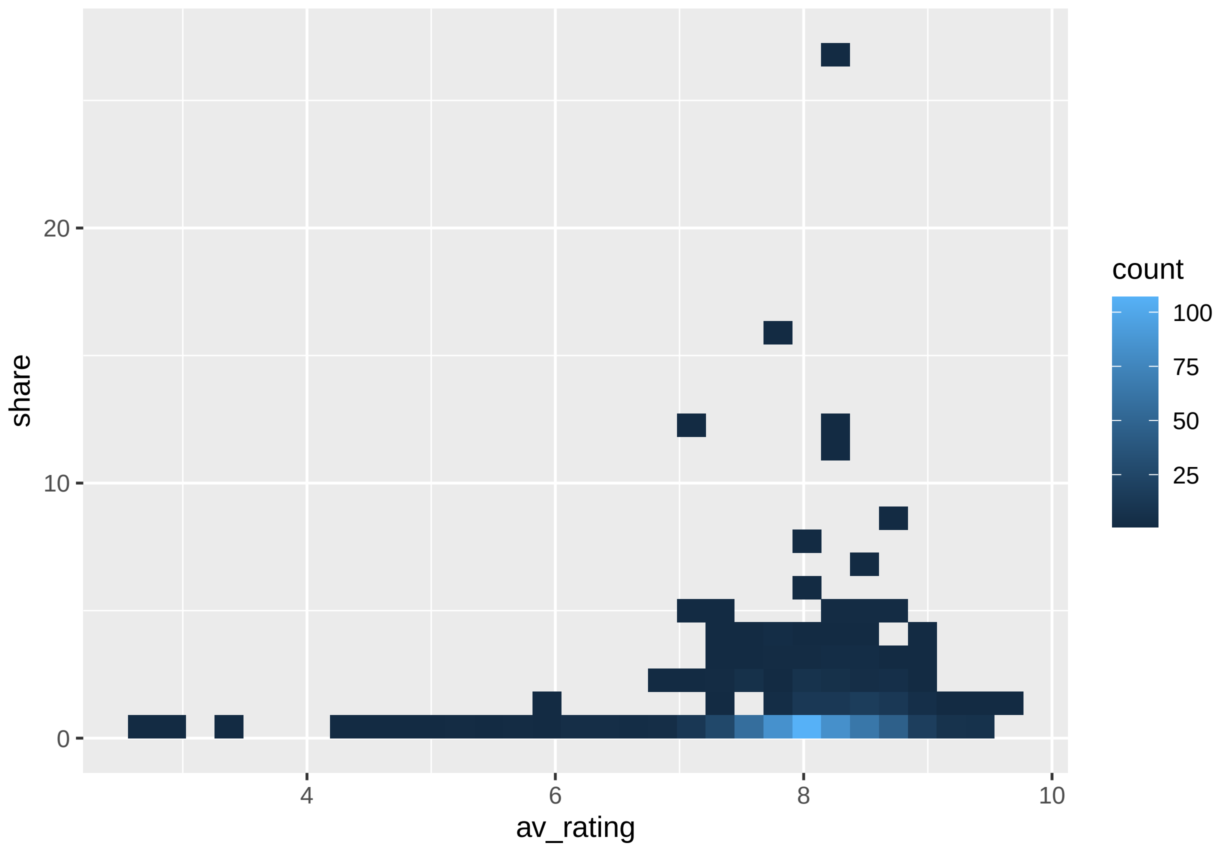
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



I think the density plot provides a better story because it shows the distribution of the ratings because it is a proptional esetimate in comparison to the other two giving count data.

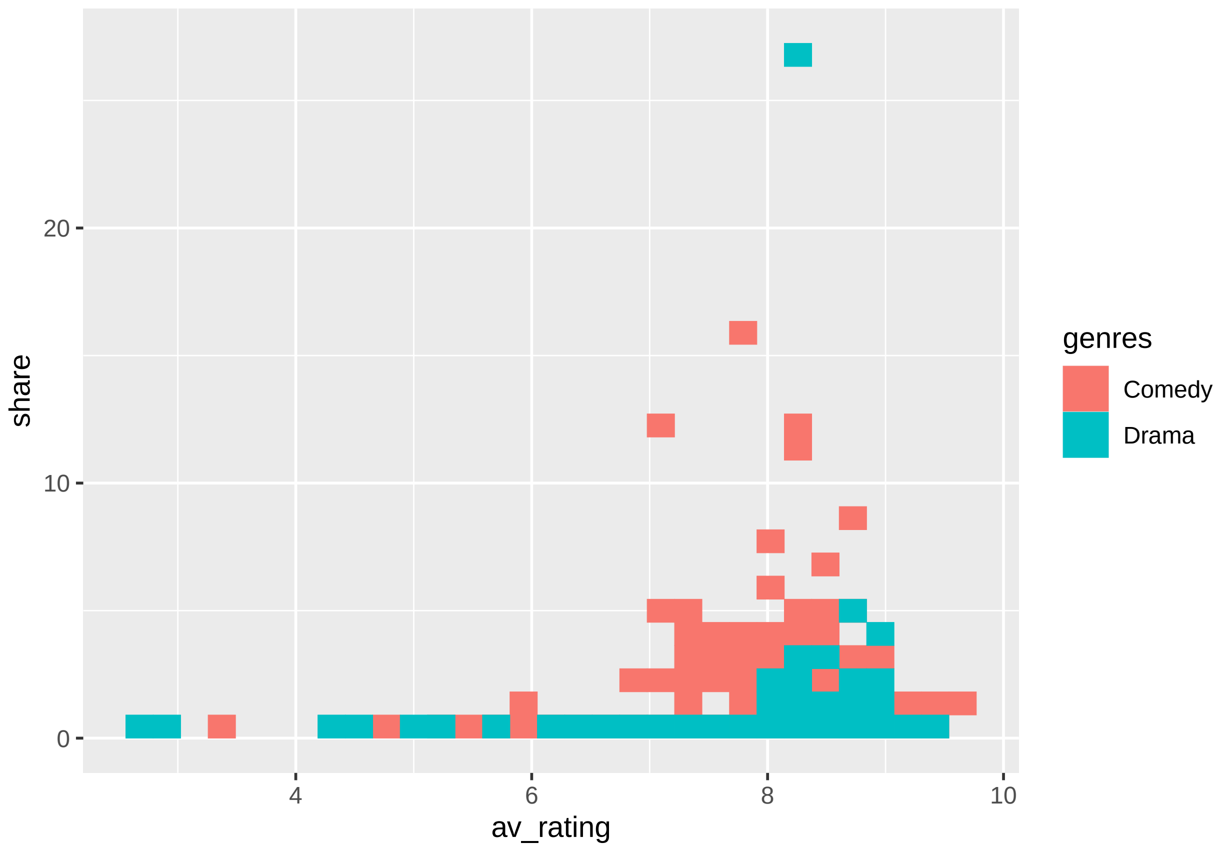
Question 6

```
ggplot(comedies_dramas, aes(x = av_rating, y = share)) +
  geom_bin_2d()
```



In this I see that there are a few shows that really capture attention, and then whoever has a 1 rating but has been on for 5+ seasons really must be bad.

```
ggplot(comedies_dramas, aes(x = av_rating, y = share, fill = genres, color = genres)) +
  geom_bin_2d(mapping = aes(fill = genres))
```



Now to find out what drama captured everyones attention… I just feel like it is grey's anatomy….

```
comedies_dramas %>%
  group_by(genres, title) %>%
  select(title, share) %>%
  arrange(desc(share))
```

```
## Adding missing grouping variables: `genres`
```

```
## # A tibble: 684 x 3
## # Groups:   genres, title [266]
##    genres title              share
##    <chr>  <chr>              <dbl>
##  1 Drama  Dekalog            27.2
##  2 Comedy Cheers             16.0
##  3 Comedy Full House         12.6
##  4 Comedy The Wonder Years   12.2
##  5 Comedy The Wonder Years   10.9
##  6 Comedy Freaks and Geeks    8.32
##  7 Comedy Cheers              7.59
##  8 Comedy Northern Exposure   6.85
##  9 Comedy Roseanne            5.91
## 10 Drama  The West Wing       5.43
## # ... with 674 more rows
```

Dekalog??? I guess that is the answer but i have NEVER heard of this show.

CHAPTER 5

```
wncaa <- read_csv("Data/wncaa.csv")
```

```
## Rows: 2092 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (6): school, conference, conf_place, how_qual, x1st_game_at_home, tourn...
## dbl (13): year, seed, conf_w, conf_l, conf_percent, reg_w, reg_l, reg_percen...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

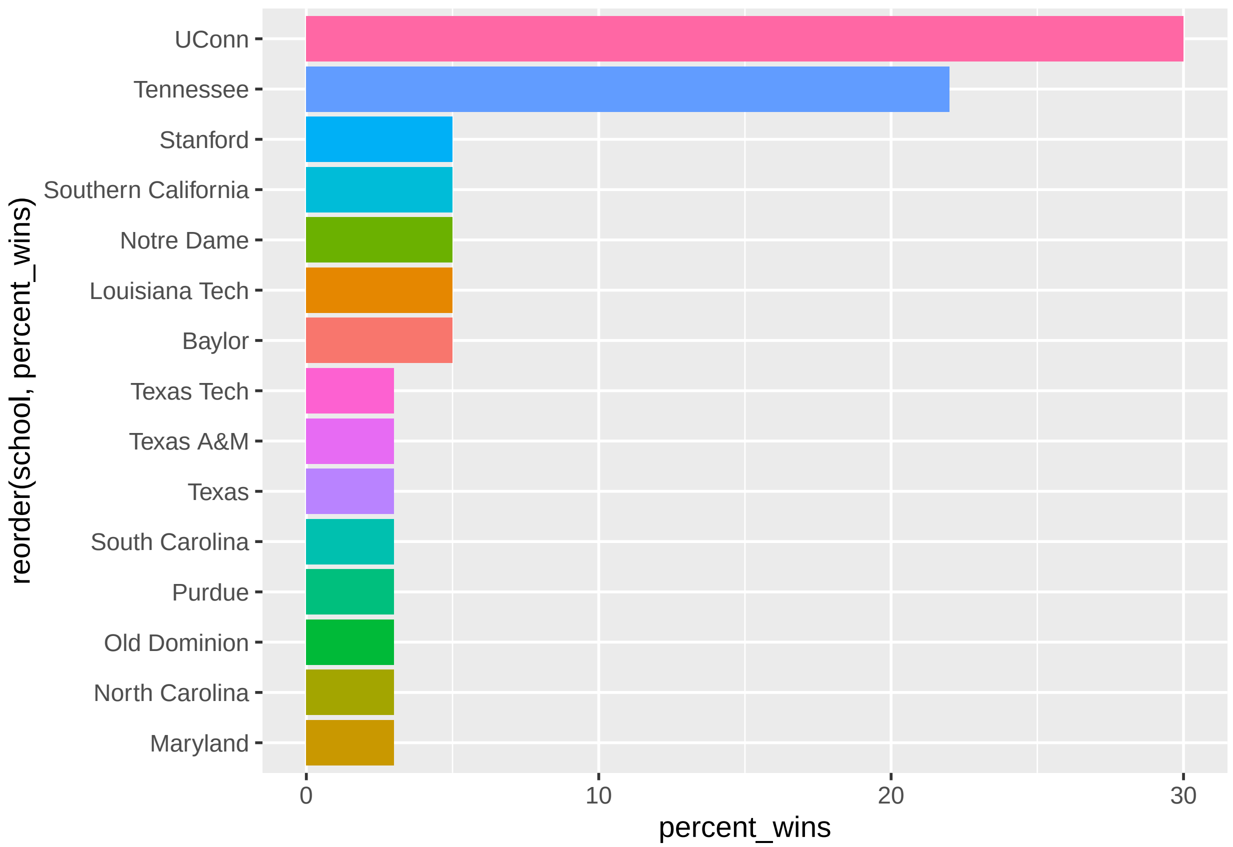
```
glimpse(wncaa)
```

```
## Rows: 2,092
## Columns: 19
## $ year              <dbl> 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982~
## $ school            <chr> "Arizona St.", "Auburn", "Cheyney", "Clemson", "Drak~
## $ seed              <dbl> 4, 7, 2, 5, 4, 6, 5, 8, 7, 7, 4, 8, 2, 1, 1, 2, 3, 6~
## $ conference        <chr> "Western Collegiate", "Southeastern", "Independent",~
## $ conf_w            <dbl> NA, NA, NA, 6, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ conf_l            <dbl> NA, NA, NA, 3, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ conf_percent      <dbl> NA, NA, NA, 66.7, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ conf_place        <chr> "-", "-", "-", "4th", "-", "-", "-", "-", "-", "-", ~
## $ reg_w             <dbl> 23, 24, 24, 20, 26, 19, 21, 14, 21, 28, 24, 17, 22, ~
## $ reg_l             <dbl> 6, 4, 2, 11, 6, 7, 8, 10, 8, 7, 5, 13, 7, 5, 1, 6, 4~
## $ reg_percent       <dbl> 79.3, 85.7, 92.3, 64.5, 81.3, 73.1, 72.4, 58.3, 72.4~
## $ how_qual          <chr> "at-large", "at-large", "at-large", "at-large", "aut~
## $ x1st_game_at_home <chr> "Y", "N", "Y", "N", "Y", "N", "N", "N", "N", "N", "Y~
## $ tourney_w         <dbl> 1, 0, 4, 0, 2, 0, 0, 0, 0, 0, 2, 0, 2, 1, 5, 3, 1, 1~
## $ tourney_l         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1~
## $ tourney_finish    <chr> "RSF", "1st", "N2nd", "1st", "RF", "1st", "1st", "1s~
## $ full_w            <dbl> 24, 24, 28, 20, 28, 19, 21, 14, 21, 28, 26, 17, 24, ~
## $ full_l            <dbl> 7, 5, 3, 12, 7, 8, 9, 11, 9, 8, 6, 14, 8, 6, 1, 7, 5~
## $ full_percent      <dbl> 77.4, 82.8, 90.3, 62.5, 80.0, 70.4, 70.0, 56.0, 70.0~
```

Question 1

```
champ_energy <- wncaa %>%
  filter(tourney_finish == "Champ") %>%
  group_by(school) %>% ##unsure about this groupby here
  summarize(N = n()) %>%
  mutate(freq = N / sum(N),
         percent_wins = round(freq*100,0)) %>%
  arrange(desc(percent_wins))

ggplot(champ_energy, aes(x = reorder(school, percent_wins), y = percent_wins, fill = school)) +
  geom_col(position = "dodge") +
  guides(fill = "none") +
  coord_flip()  ##why does it not want to do this
```

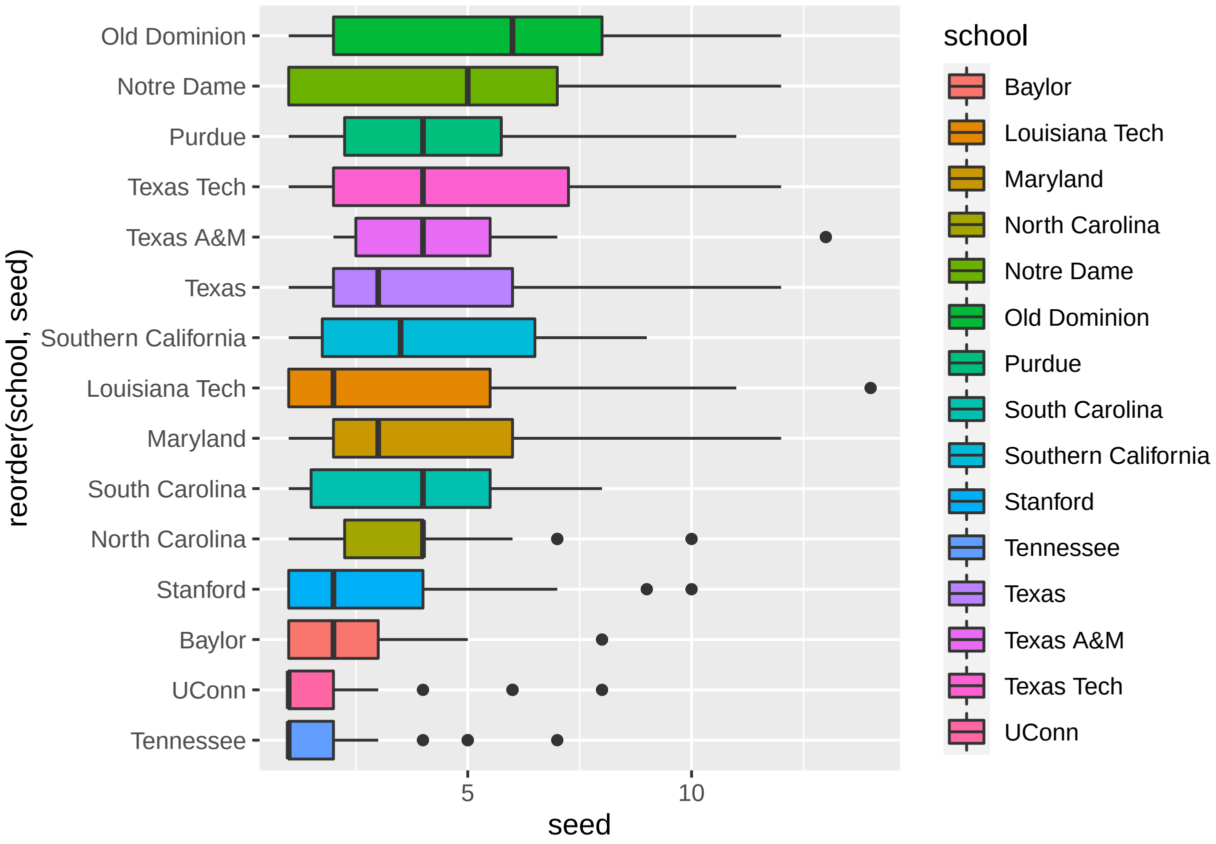


It looks that the two teams with the most wins (via %) are uconn and Tennessee

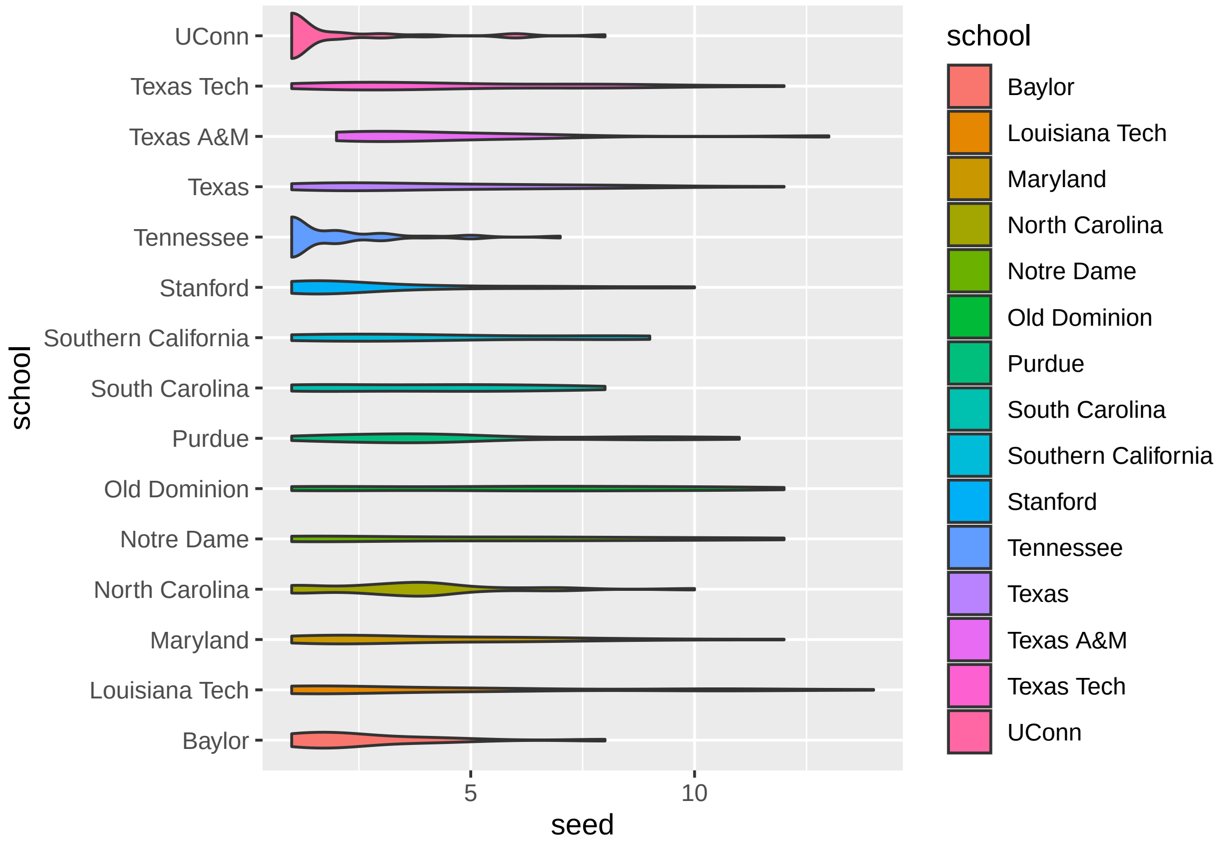
Question 2

```r
champ_names <- unique(champ_energy$school)
winners <- wncaa %>%
  filter(school %in% champ_names)

ggplot(winners, aes(y = seed,
                    x = reorder(school, seed),
                    fill = school
                    )) +
  geom_boxplot() +
  coord_flip()
```



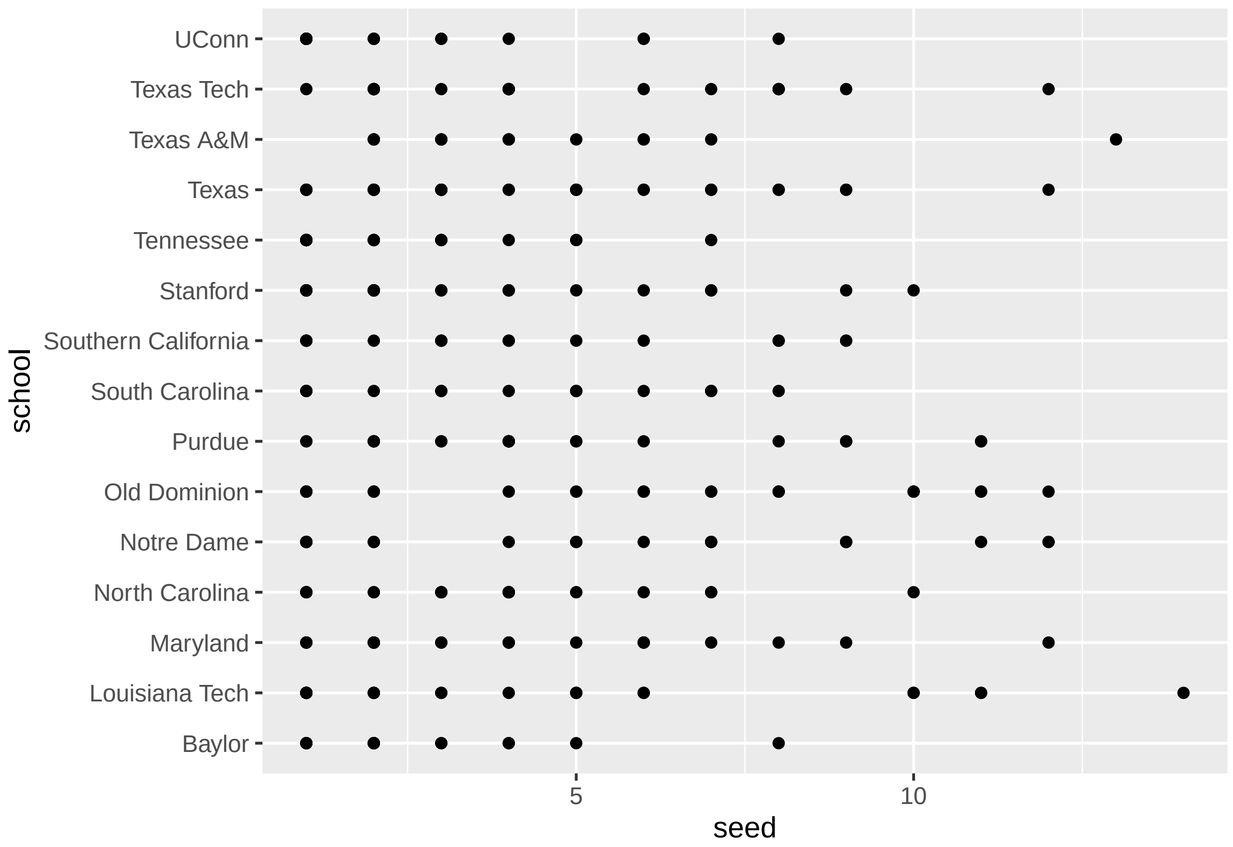
```
ggplot(winners, aes(y = seed,
                    x = school,
                    fill = school
                    )) +
  geom_violin() +
  coord_flip()
```



I think Box plots in this particular situation are more useful because the median tick mark helps the reader see where exactly they should be looking in comparison to the other schools.

Question 3

```
ggplot(winners, aes(y = seed, x = school,vfill = school)) +
  geom_point() +
  coord_flip()
```



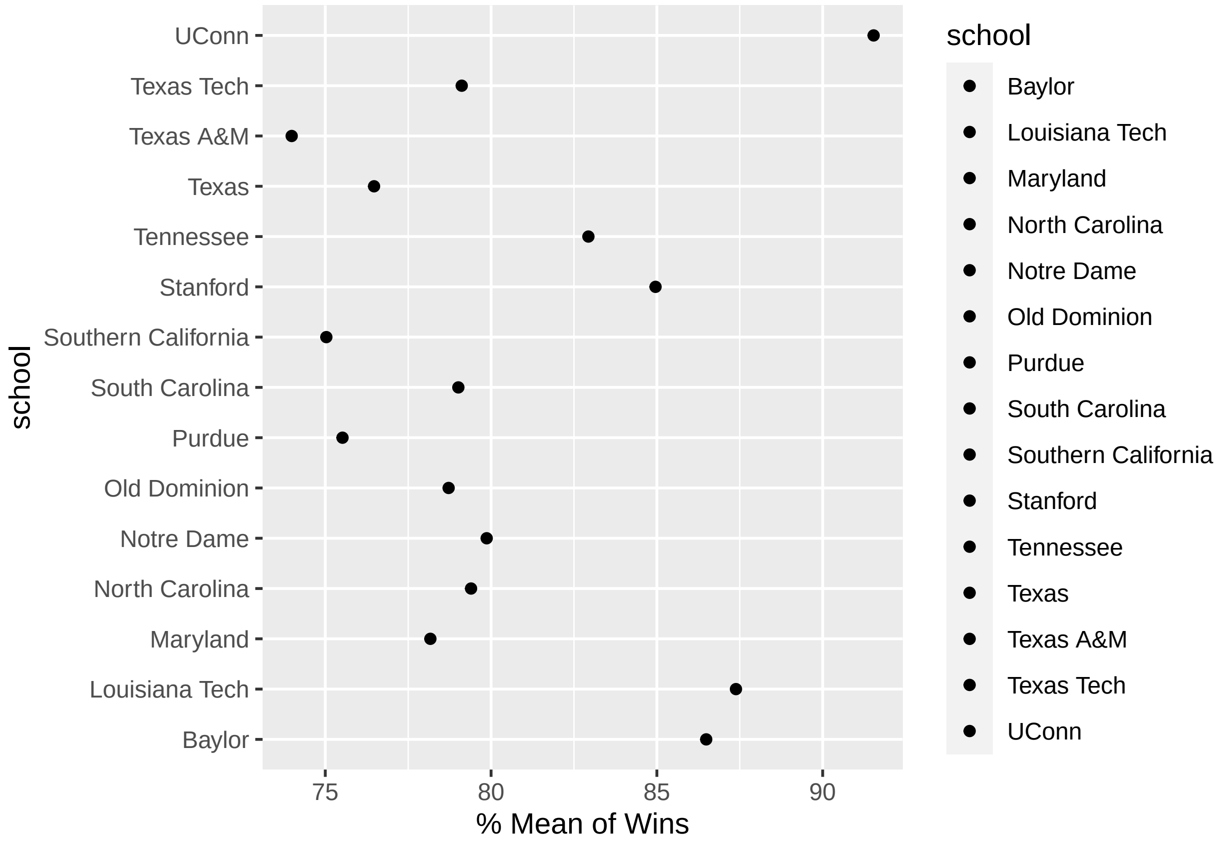
Geom point does not work as well because the seeds that are given are the changes over time (i think?) and these little dot guys do not really give any indication of what is more or less important to look at.

Question 4

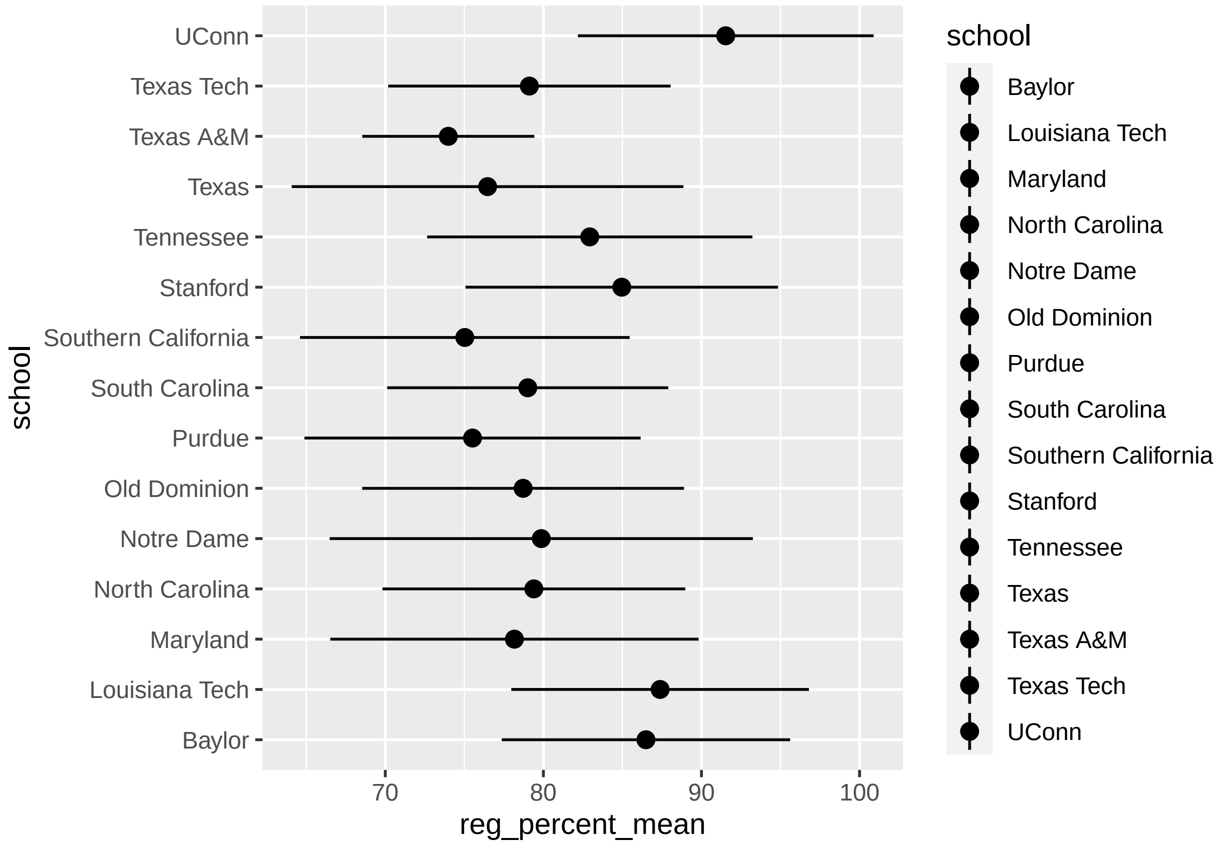
```
winners_sum <- winners %>%
  group_by(school) %>%
  summarise_if(is.numeric, funs(mean,sd), na.rm = TRUE) %>%
  ungroup()
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

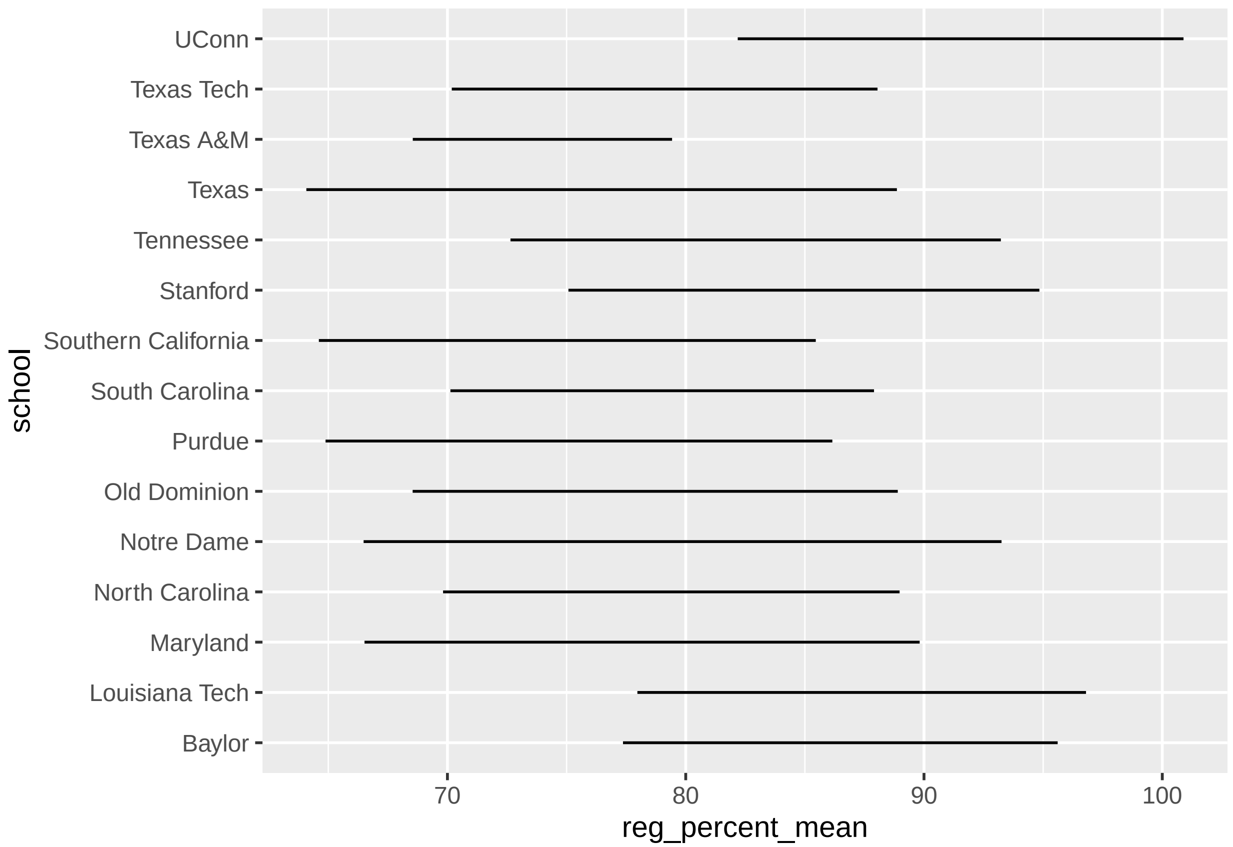
```
ggplot(winners_sum, aes(x = school, y = reg_percent_mean, fill = school)) +
  geom_point() +
  coord_flip() +
  labs(y = "% Mean of Wins")
```



```
ggplot(winners_sum, aes(x = school, y = reg_percent_mean, fill = school)) +
  geom_pointrange(mapping = aes(ymin = reg_percent_mean - reg_percent_sd, ymax = reg_percent_mean + reg_
  coord_flip()
```



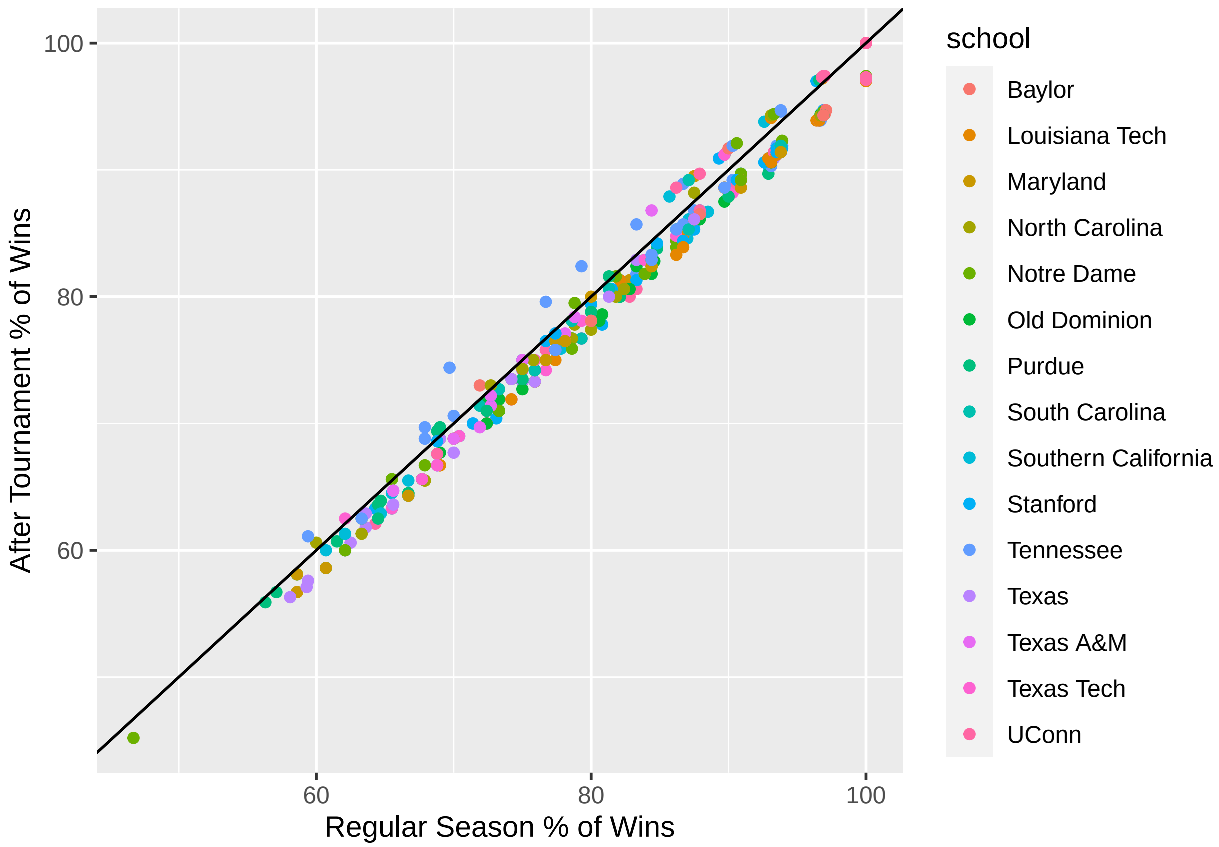
```
ggplot(winners_sum, aes(x = school, y = reg_percent_mean, fill = school)) +
  geom_linerange(mapping = aes(ymin = reg_percent_mean - reg_percent_sd, ymax = reg_percent_mean + reg_
  coord_flip() ##i prefer pointrange
```



The results indicate that UConn has the the largest average win percentage (not surprising, they have been RAW for a solid ten years), and A&M sucks which is great because Ally went UT hookem'! As for the team with the most narrow interval, it is A&M, again, WHICH YAY! Screw aggies.

Question 5

```
ggplot(winners, aes(x = reg_percent, y = full_percent, fill = school, color = school)) +
  geom_point(aes(fill = school)) +
  geom_abline() +
  labs(x = "Regular Season % of Wins",
       y = "After Tournament % of Wins",
       Title =
       "Performance of Schools' Wins")
```

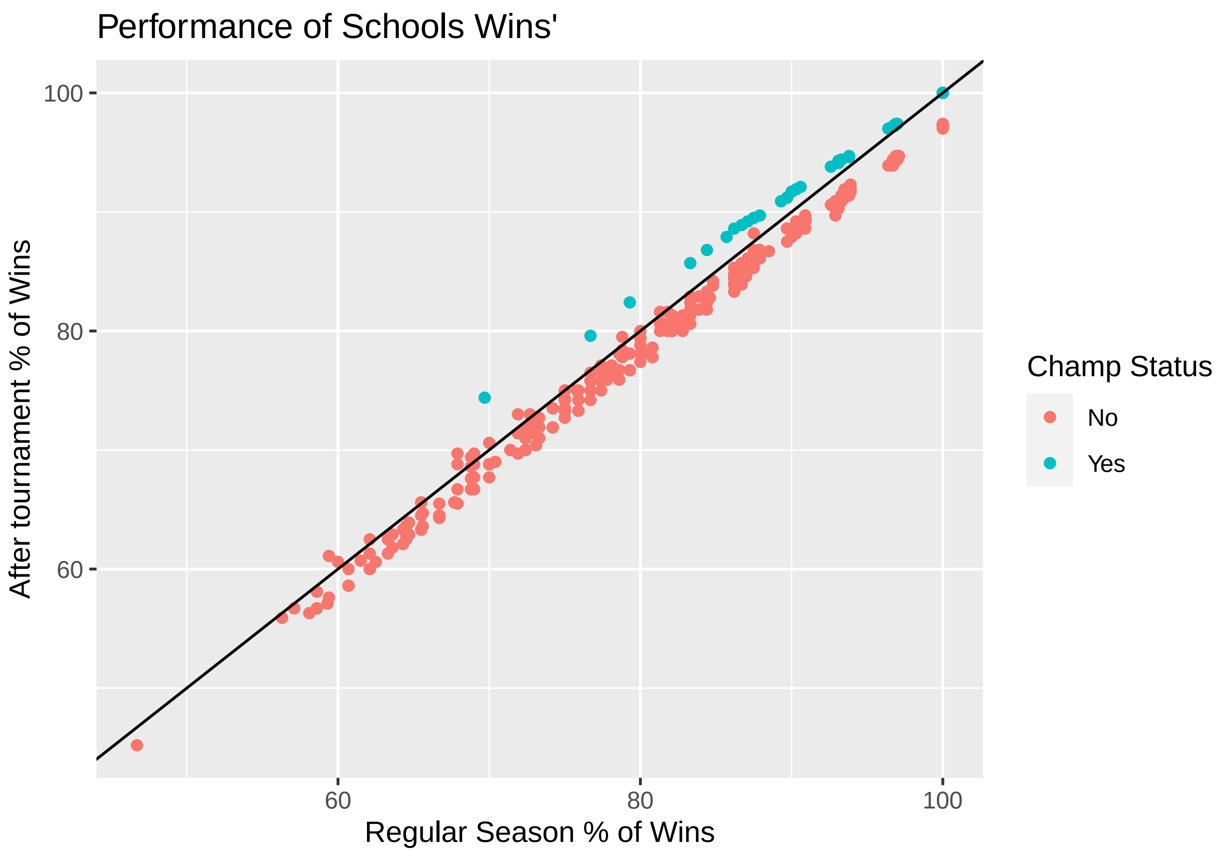


In terms of patterns it is interesting that both the full and reg percents stay above 60 (for the most part). When I added in color = school, i am able to better see which schools lang on top and bottom, which again shows UCONN dominating.

Question 6

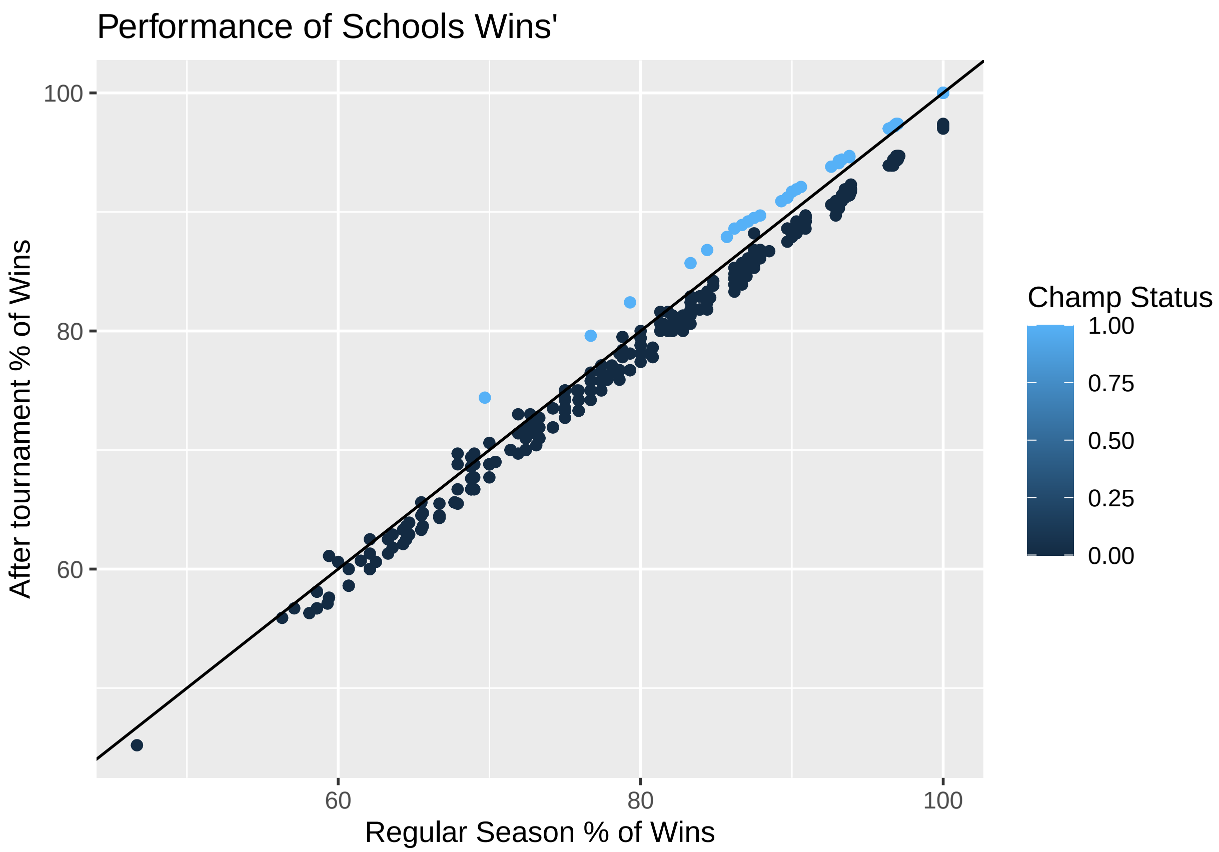
```r
#champs variable
winners <- winners %>%
  mutate(is_champ = if_else(tourney_finish == "Champ", 1, 0),
         is_champ = as.factor(is_champ)) ##mutate creates a new col!

ggplot(winners, aes(x = reg_percent, y = full_percent,
                                   color = is_champ)) +
  geom_point() +
  geom_abline() +
  labs(x = "Regular Season % of Wins",
      y = "After tournament % of Wins",
     title = "Performance of Schools Wins'",
     col = "Champ Status") +
  scale_colour_discrete(labels = c("No", "Yes"))
```



```
##what happens when you remove the as.factor argument
winners <- winners %>%
  mutate(is_champ_nofactor = if_else
         (tourney_finish == "Champ", 1, 0))

ggplot(winners, aes(x = reg_percent,y = full_percent,
                     color = is_champ_nofactor)) +
  geom_point() +
  geom_abline() +
  labs(x = "Regular Season % of Wins", y = "After tournament % of Wins",
       title = "Performance of Schools Wins'",
       col = "Champ Status")
```



You get numeric values instead of discrete values! Champs seem to have improvement over time, which makes sense because practice makes perfect.

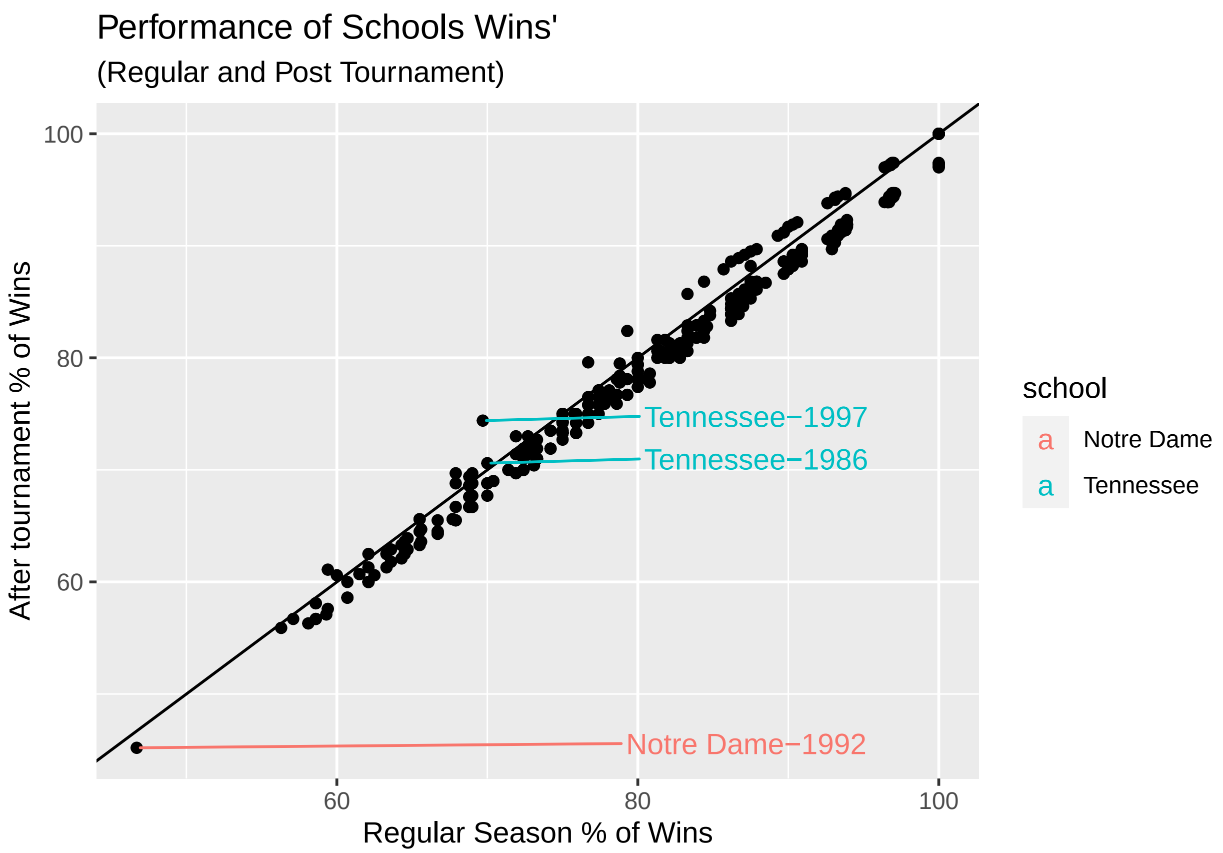
Question 7 Do you see anything interesting? I'll give you a hint: the school that has overperformed the most has been the same one, one decade apart.

```
winners <- winners %>%
  mutate(plot_label = paste(school, year, sep = "-"))
winners <- winners %>%
  mutate(difference = full_percent - reg_percent)
##time to label points of interest..... yikes!
winners_2 <- winners %>%
  filter(reg_percent < 50 | reg_percent <= 70 & full_percent >= 70)

#install.packages("ggrepel")
          ##bestie Healy used this for point finding
library(ggrepel)
```

Time to make my plot with labels now!

```
champ_labeling_plot <- ggplot(winners, aes(x = reg_percent,
                                   y = full_percent)) +
  geom_point() +
  geom_abline()
##breaking it up here to help with the cluttering
champ_labeling_plot +
  geom_text_repel(data = winners_2,
                  mapping = aes(label = plot_label, color = school),
                  hjust = -3.5, vjust = .4) +
  labs(x = "Regular Season % of Wins", y = "After tournament % of Wins",
     title = "Performance of Schools Wins'",
     subtitle = "(Regular and Post Tournament)")
```



```
##unsure how to remove Tennessee from being plotted twice
```

In terms of interesting, I am curious as to how Notre Dame can suck that bad in 92. Like what on earth was happening there? It is also interesting to see how Tennessee falls on this plot. **Question 8**

```
winners %>%
  group_by(school) %>%
  filter(full_percent == 100 & reg_percent == 100)
```

```
## # A tibble: 8 x 23
## # Groups:   school [3]
##     year school  seed confere~1 conf_w conf_l conf_~2 conf_~3 reg_w reg_l reg_p~4
##    <dbl> <chr>  <dbl> <chr>      <dbl>  <dbl>   <dbl> <chr>   <dbl> <dbl>   <dbl>
## 1  1986 Texas      1 Southwest     16      0     100 1st        29     0     100
## 2  1995 UConn      1 Big East      18      0     100 1st        29     0     100
## 3  2002 UConn      1 Big East      16      0     100 1st        33     0     100
## 4  2009 UConn      1 Big East      16      0     100 1st        33     0     100
## 5  2010 UConn      1 Big East      16      0     100 1st        33     0     100
## 6  2012 Baylor     1 Big 12        18      0     100 1st        34     0     100
## 7  2014 UConn      1 American~     18      0     100 1st        34     0     100
## 8  2016 UConn      1 American~     18      0     100 1st        32     0     100
## # ... with 12 more variables: how_qual <chr>, x1st_game_at_home <chr>,
## #   tourney_w <dbl>, tourney_l <dbl>, tourney_finish <chr>, full_w <dbl>,
## #   full_l <dbl>, full_percent <dbl>, is_champ <fct>, is_champ_nofactor <dbl>,
## #   plot_label <chr>, difference <dbl>, and abbreviated variable names
## #   1: conference, 2: conf_percent, 3: conf_place, 4: reg_percent
```

Me smiling for UT rn :') The undefeated teams are Baylor, UConn, and UT! In terms of being surprised, I don't really like Baylor so that was suprsing, but ya know. Is what it is. I think I was especially suprsied by UT, but that sort of makes sense as it wad 1986. Lastly, UConn just win after win for TWO DECADES! That is pretty insane.