The Ideological Bias of Newspapers

Text analysis gives researchers a powerful set of tools for extracting general information from a large body of documents.

This exercise is based on Gentzkow, M. and Shapiro, J. M. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78(1): 35–71.

We will analyze data from newspapers across the country to see what topics they cover and how those topics are related to their ideological bias. The authors computed a measure of a newspaper's "slant" by comparing its language to speeches made by Democrats and Republicans in the U.S. Congress.

You will use three data sources for this analysis. The first, `dtm`, is a document term matrix with one row per newspaper, containing the 1000 phrases – stemmed and processed – that do the best job of identifying the speaker as a Republican or a Democrat. For example, "living in poverty" is a phrase most frequently spoken by Democrats, while "global war on terror" is a phrase most frequently spoken by Republicans; a phrase like "exchange rate" would not be included in this dataset, as it is used often by members of both parties and is thus a poor indicator of ideology.

The second object, `papers`, contains some data on the newspapers on which `dtm` is based. The row names in `dtm` correspond to the `newsid` variable in `papers`. The variables are:

Name	Description
`newsid`	The newspaper ID
`paper`	The newspaper name
`city`	The city in which the newspaper is based
`state`	The state in which the newspaper is based
`district`	Congressional district where the newspaper is based (data for Texas only)
`nslant`	The "ideological slant" (lower numbers mean more Democratic)

The third object, `cong`, contains data on members of Congress based on their political speech, which we will compare to the ideological slant of newspapers from the areas that these legislators represent. The variables are:

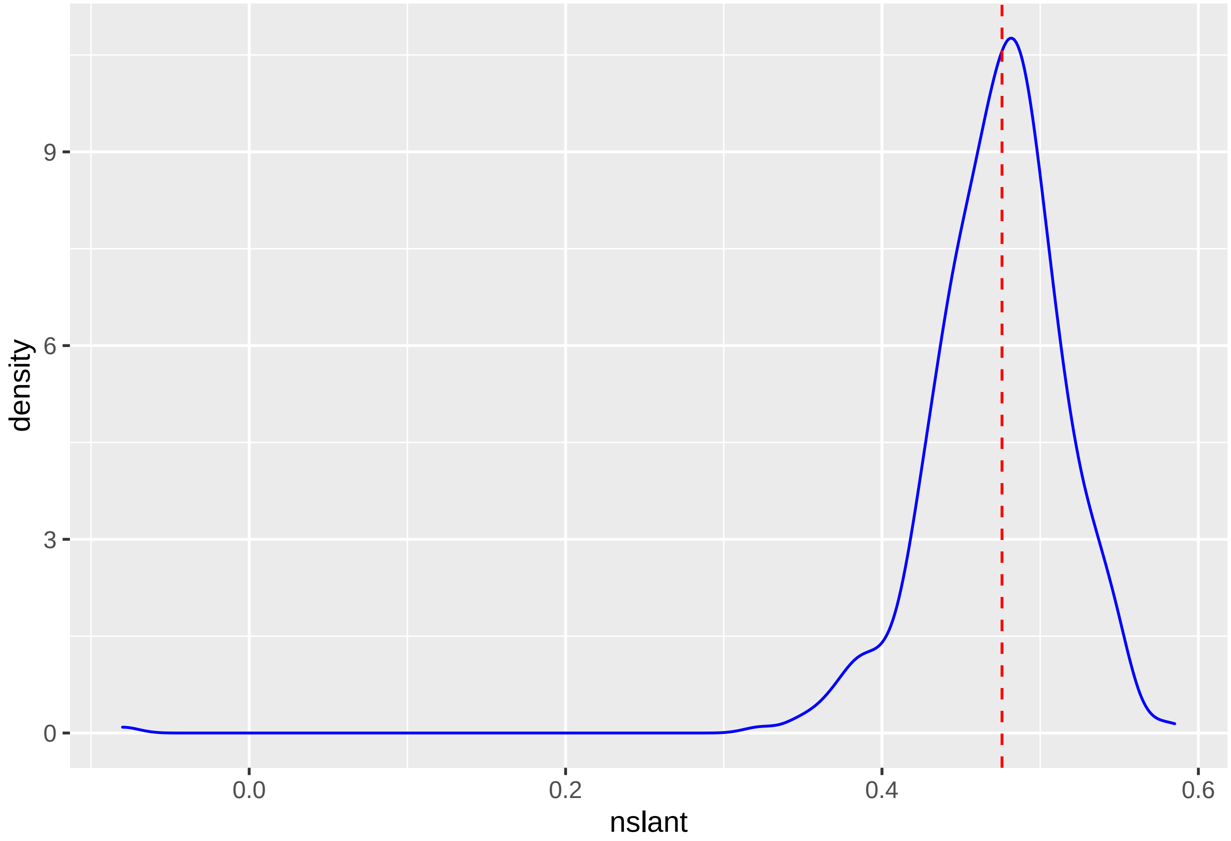
Name	Description
`legname`	Legislator's name
`state`	Legislator's state
`district`	Legislator's Congressional district
`chamber`	Chamber in which legislator serves (House or Senate)
`party`	Legislator's party
`cslant`	Ideological slant based on legislator's speech (lower numbers mean more Democratic)

Question 1

We will first focus on the slant of newspapers, which the authors define as the tendency to use language that would sway readers to the political left or right. Load the data and plot the distribution of `nslant` in the papers data frame, with a vertical line at the median. Which newspaper in the country has the largest left-wing slant? What about right?

Answer 1

```r
load("newspapers.RData")
#Import ggplot2 library for plotting purposes
library(ggplot2)

#Plot the nslant distribution with a vertical line at the median
ggplot(papers) +
  geom_density(aes(nslant),color = "blue") +
  geom_vline(aes(xintercept = median(nslant)),color = "red",linetype = "dashed")
```



```r
#Sort papers by nslant, take the first and last values to find left and rightmost papers
sorted_papers <- papers[order(papers$nslant),]
leftmostpaper <- sorted_papers$paper[1]
rightmostpaper <- sorted_papers$paper[length(sorted_papers)]

#Calculate the median and the mean of all the papers' nslants.
median_nslant <- median(papers$nslant)
mean_nslant <- mean(papers$nslant)
```

```
#The median of the nslant distribution was 0.4758851, which is slightly above the mean of 0.4721112.

#This implies that there are slightly more extremely Democratic papers (low nslant) than extremely Repu

#The Chicago Defender has the largest left-wing slant in the country

#The Daily News has the largest right-wing slant in the country.
```

Question 2

We will explore the content of these newspapers using the `wordcloud` package.

First load the `wordcloud` package. Make a word cloud of the top words (at most 20) in the `dtm` object. What were the biggest topics in the news in 2005 when these data were collected? Hint: first convert `dtm` into a `matrix`.

Now subset the data to the tenth of newspapers with the leftmost (lowest) political slant and the rightmost (highest) political slant. Make two word clouds showing the words most commonly used by each group of newspapers (again, at most 20 words). How does their language differ? Do they have anything in common? Hint: to use your usual subsetting/indexing tools, convert your dtm matrix into a data frame using the `data.frame` function.

Answer 2

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

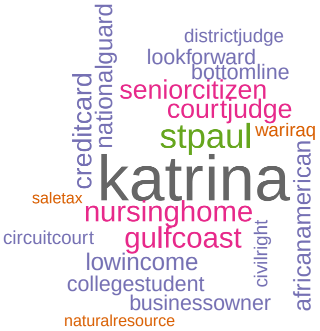
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(slam)
library(RColorBrewer)
library(tm)
```

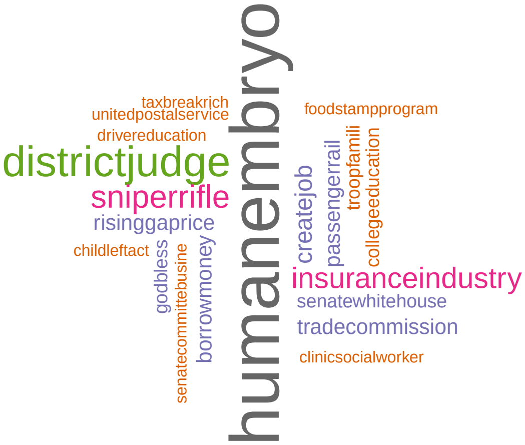
```
## Loading required package: NLP

##
## Attaching package: 'NLP'
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
#Convert dtm to matrix
dtm_matrix <- as.matrix(dtm)

#Sort the 1000 phrases by their frequency
words <- sort(colSums(dtm_matrix), decreasing = TRUE)

#Create wordcloud for all data
wordcloud(names(words), freq = words, min.freq = 1, max.words = 20, random.order = FALSE, rot.per = 0.35
```



```
#The top 20 words here reflect the most commonly discussed topics across all the newspapers.
#As expected, Hurricane Katrina, the War in Iraq, and low income are commonly mentioned.


#Convert dtm matrix to data frame
dtm_df <- data.frame(dtm_matrix)

#Find the IDs of the tenth most left/right newspapers. Since there are 434 newspapers in total, these s
left10thpaperids <- head(papers[order(papers$nslant),],43)$newsid
right10thpaperids <- head(papers[order(-papers$nslant),],43)$newsid

#Subset dtm_df into newspapers with the same IDs as above. Then, sums the columns to obtain the frequen
```

```
left10thwords <- dtm_df %>% filter(row.names(dtm_df) %in% left10thpaperids) %>% colSums()
right10thwords <- dtm_df %>% filter(row.names(dtm_df) %in% right10thpaperids) %>% colSums()

#Create wordcloud using the filtered data. Note that names(words) is repeated since the 1000 phrases ar
wordcloud(names(words), freq = left10thwords, min.freq = 1, max.words = 20, random.order = FALSE, rot.p
```



```
wordcloud(names(words), freq = right10thwords, min.freq = 1, max.words = 20, random.order = FALSE, rot.
```

insuranceindustry
passengerrail
lifelibertypursuit
unitedairf
troopfan
collegeeduca
godbless
humanembryo
axcutpermanent
taxbreakrich
arcticnationalwildli
senatewhitehouse
foodstampprogram
budgetdeficit
childleftact

```
#The wordcloud for the left-leaning newspapers discuss topics including clinicsocialworker, passengerra

#The wordcloud for the right-leaning newspapers discuss topics including taxbreakrich, unitedairforce, 

#Common topics include humanembryo, insuranceindustry, collegeeducation, unitedstatespostalservice, and
```

Question 3

We will now explore the relationship between the political slant of newspapers and the language used by members of Congress.

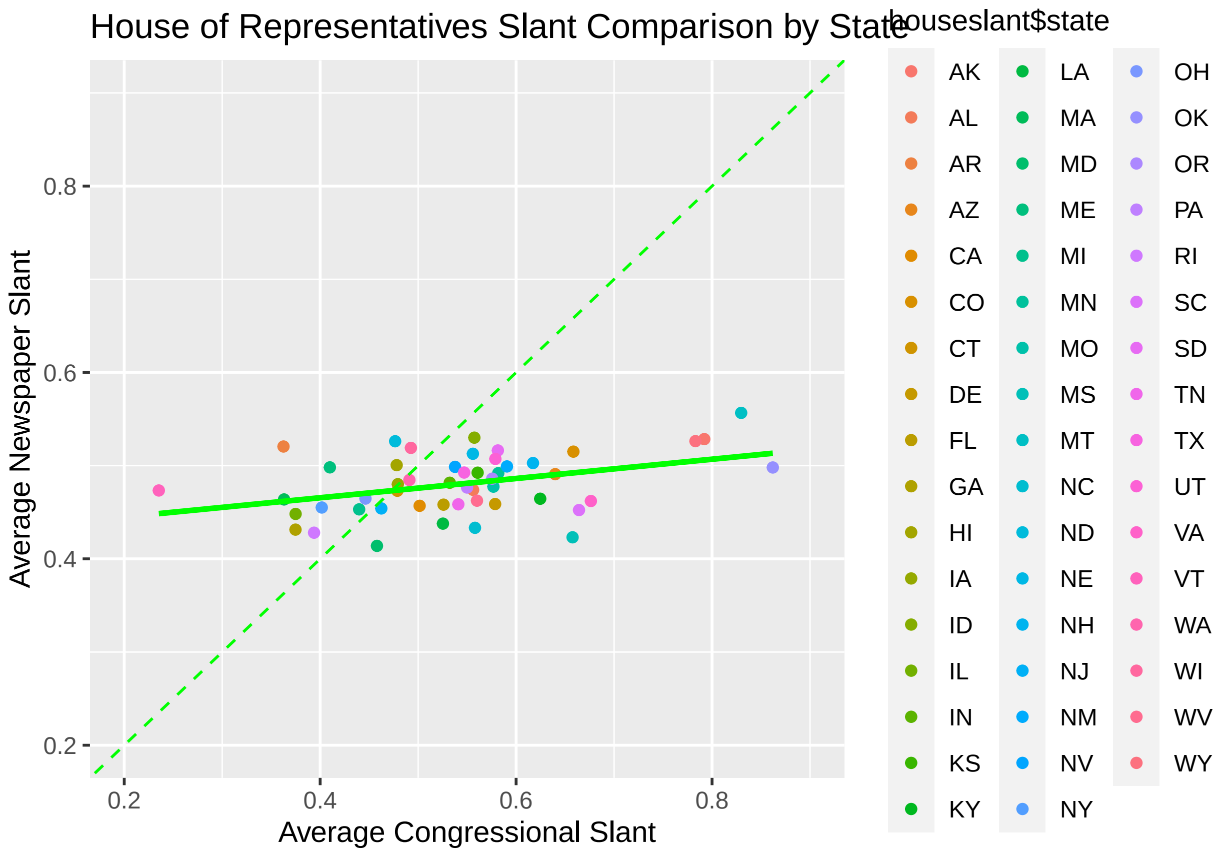
Using the dataset `cong`, compute average slant by state separately for the House and Senate. Now use `papers` to compute the average newspaper slant by state. Make two plots with Congessional slant on the x-axis and newspaper slant on the y-axis – one for the House, one for the Senate. Include a best-fit line in each plot – a red one for the Senate and a green one for the House. Label your axes, title your plots, and make sure the axes are the same for comparability. Can you conclude that newspapers are influenced by the political language of elected officials? How else can you interpret the results?

Answer 3

```
library(ggplot2)

#Use group_by and summarise to filter cong into House and Senate Slant by State.
```

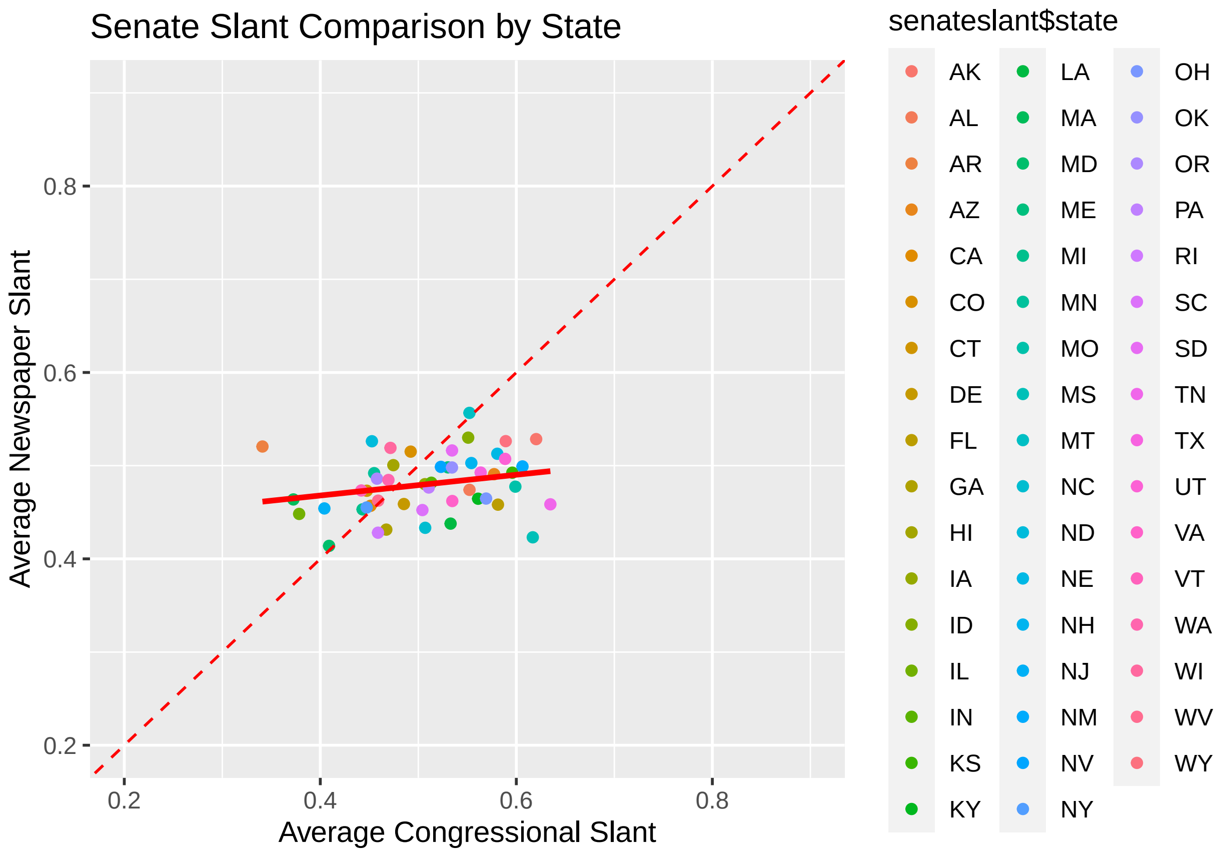
```
 houseslant <- cong %>%
  filter(chamber == "H") %>%
  group_by(state) %>%
  summarise(avgslant = mean(cslant))
senateslant <- cong %>%
  filter(chamber == "S") %>%
  group_by(state) %>%
  summarise(avgslant = mean(cslant))

#Use group_by and summarise to filter papers into Newspaper Average Slant by State
paperslant <- papers %>%
  group_by(state) %>%
  summarize(avgslant = mean(nslant))

#Remove DC Data, no corresponding cslant.
paperslant <- paperslant[-8,]

#Create dataframe with all the slant data for plotting purposes
slantdf <- data.frame(houseslant, senateslant, paperslant)

#Plot House Slants. Color points by state, label axes and title.
#Add line of best fit and y = x for comparison
#Matched x-axis scale with following Senate plot
ggplot(slantdf, aes(houseslant$avgslant,paperslant$avgslant)) +
  geom_point(aes(color = houseslant$state)) +
  geom_smooth(method = "lm", se = FALSE, color = "green") +
  geom_abline(intercept = 0, slope = 1, color = "green", linetype = "dashed") +
  lims(x = c(0.2,0.9), y = c(0.2,0.9)) +
  labs(x = "Average Congressional Slant", y = "Average Newspaper Slant", title = "House of Representati
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
#Plot Senate Slants. Color points by state, label axes and title.
#Add line of best fit and y = x for comparison
#Matched x-axis scale with previous House plot
ggplot(slantdf, aes(senateslant$avgslant,paperslant$avgslant)) +
  geom_point(aes(color = senateslant$state)) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  lims(x = c(0.2,0.9), y = c(0.2,0.9)) +
  labs(x = "Average Congressional Slant", y = "Average Newspaper Slant", title = "Senate Slant Comparis
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
#As shown by the slope of the solid lines, there is a very slight positive correlation between Newspape
#The dashed lines demonstrate how small this correlation is compared to a true 1-to-1 relation.
#From this small slope, I can conclude that the slant of newspapers is mostly independent of congressio
#In the House, even at cslant values above 0.6, there are nslant values ranging from 0.42 to 0.56.
#The same nslant range can be observed at cslant values below 0.4, further confirming this independence

#Alternatively, one could argue that the slightly positive slope indicates that nslant IS dependent on
#There are high nslant states with a cslant > 0.5 than high nslant states with a cslant < 0.5.
#One could interpret this as the Media conforming to Congressional expectations,
#or one could argue that Congressmen are doing a good job representing their district's biases.
```

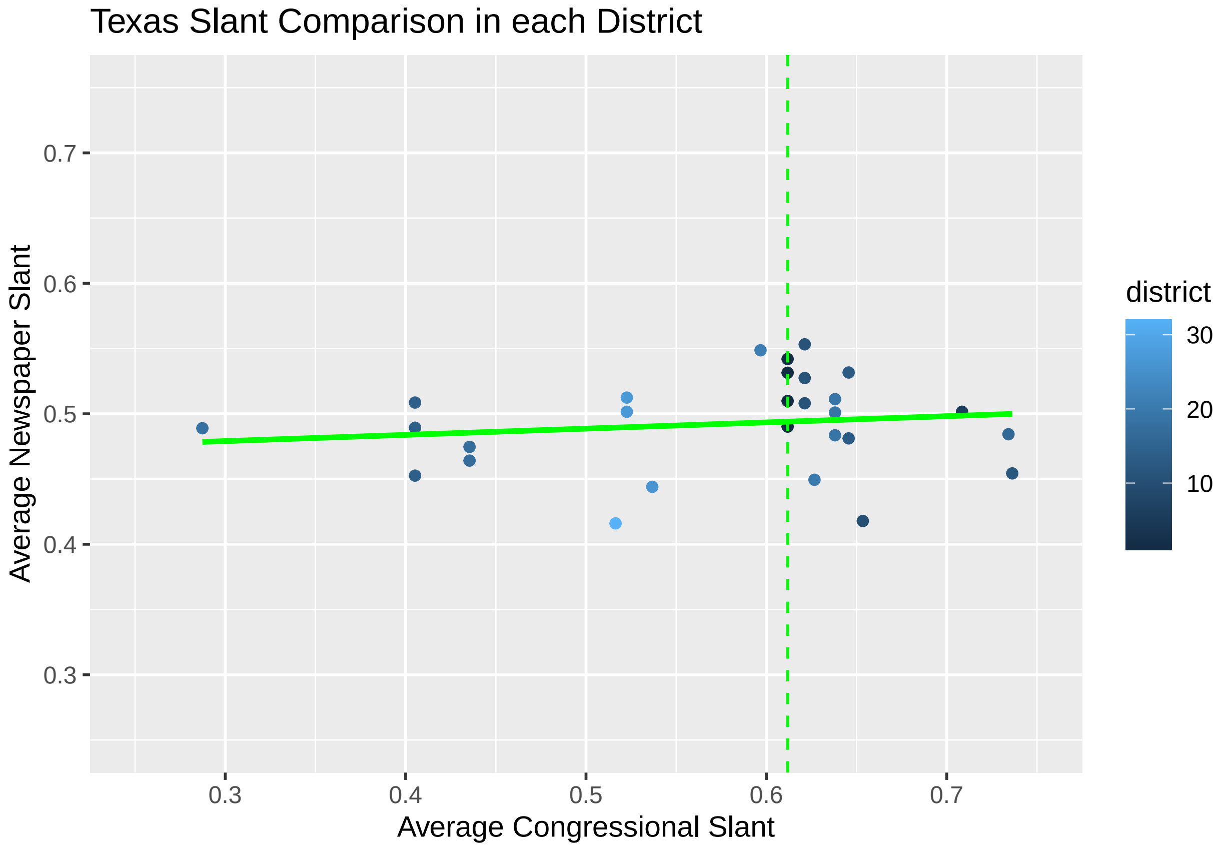
Question 4

We will now take a closer look at the relationship between congressional and media slant at the district level, for one particular state – Texas. To do so, subset the two datasets to Texas alone, then merge them by district and state, keeping only the observations that appear in both datasets. Then, produce the same plot as in question 3 above, but at the district level (just for the House). What do you find? Which results do you think are more informative, and why?

Answer 4

```
#Filter data for Texas and the House only. Merge datasets by district and state.
texascslant <- cong %>%
  filter(state == "TX", chamber == "H")
texasnslant <- papers %>%
  filter(state == "TX")
texasslant <- merge(texascslant, texasnslant, by = c("district","state"))

#Plot data with x = cslant, y =nslant. Include labels for axes and title
#Added line of best-fit, and dashed vertical line for District 1
ggplot(texasslant, aes(cslant,nslant)) +
  geom_point(aes(color = district)) +
  geom_smooth(method = "lm", se = FALSE, color = "green") +
  geom_vline(aes(xintercept = 0.6118010), color = "green", linetype = "dashed") +
  lims(x = c(0.25,0.75), y = c(0.25,0.75)) +
  labs(x = "Average Congressional Slant", y = "Average Newspaper Slant", title = "Texas Slant Compariso
```

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



```
#Here, the line of best-fit is similar to the slopes of the graphs shown in Question 3.
#I find that there is little correlation between the cslant and nslant of a district.
#However, I believe the data visualized here is more informative than the graphs above.
#Each district can have multiple newspapers (thus multiple nslant values), whereas each State in the pr

#In Texas, there are several district with nslants differing from their cslants.
#Most notably, District 1 has 4 distinct nslant values that share a single cslant value (shown along th
#The cslant is 0.61, while the nslants range from 0.49 to 0.54.
#This disparity demonstrates that the Newspaper Slant of a District is independent from its Congression
```