EDAV final project fall 2018

Case.law

Benedikt Schifferer (bds2141), Patrick Kwon (yk2805), Shadi Fadaee (sf2917)

2018-12-10

```
WD = "~/Projects/Columbia/01_EDV/03b_Project/exploratory_law.case/02_r_report/"
setwd(WD)

# Load data and preprocess
dfb <- fread('../data/03_combined/df_combined_v2.csv', sep=',')
dfb$year <- as.numeric(substr(dfb$decision_date, 1, 4))

dfb <- dfb %>%
  mutate(state_tmp = str_replace(file, '/Projects/Columbia/01_EDV/03_Project/5702_Pro
ject/', ''))
dfb <- dfb %>%
  mutate(state = substr(state_tmp, 1, regexpr("-", dfb$state_tmp)-1))

dfb$no_case <- 1
dfb$cb_no_opinions <- as.numeric(dfb$cb_no_opinions)
```

```
## Warning: NAs introduced by coercion
```

```
dfb$cb_no_parties <- as.numeric(dfb$cb_no_parties)
dfb$cb_no_attorneys <- as.numeric(dfb$cb_no_attorneys)
dfb$cb_no_judges <- as.numeric(dfb$cb_no_judges)
```

0. Structure

The final project including all code and the final report is published on github: GitHub (https://github.com/bschifferer/exploratory_law.case/)

- 01_download_preprocess contains scripts for downloading and preprocessing the dataset
- 02_r_report contains the RMarkdown script for the project report
- 03_r_interactive contains the RShiny tool
- 04_r_spatial contains the interactive spatial script
- data is a directory for data exchange between the script. It contians some basic configuration files

  Live demos of interactive components are:

- Spatial plots for websites (https://yj7082126.github.io/patrick_kwon/)
- Interactive tool for fast data exploration (http://ec2-54-152-240-211.compute-1.amazonaws.com:3838/03_r_interactive/)

The datasets are only accessible for registered users. Therefore, all sensitive data and API KEYs are removed from the repository and report. The script contains placeholder API_KEY, which can be replaced with an own API_KEY. If there are any questions, partial data can may be provided.

1. Introduction

One of the most useful tools that we have for understanding society and its changes is data from law cases. Law cases are dependent on the trends of the times in which they are reported. Changes in political, societal and personal realms affect the attitudes of individuals towards issues of law and the enforcement of such laws. A direct indicator of these attitudes is the law cases reported in different categories over time.

In order to understand the trends in law cases we looked at a comprehensive dataset of law cases in the United States to analyze these trends and understand patterns.

Team members: Benedikt Schifferer (bds2141), Patrick Kwon (yk2805), Shadi Fadaee (sf2917)

Benedikt Schifferer: Focused on data preprocessing.
Patrick Kwon: Focused on visualization of spatial plots.
Shadi Fadaee: Focused on visualization of linear plots over time.

2. Description of datasets

2.1 Case.law dataset and access

```
theDate <- as.Date("2009-12-01",format="%Y-%m-%d")

api_key = '...'
search_term = 'dog'
url_law <- paste0("https://api.case.law/v1/cases/?cite=&full_case=true&search=", sear
ch_term)
query <- paste0(url_law, "&decision_date_min=", format(theDate,"%Y-%m-%d"), "&decisio
n_date_max=", format(theDate+6,"%Y-%m-%d"))
getdata <- GET(url = query, add_headers(Authorization=api_key))
df <- fromJSON(content(getdata, type='text'))$results
```

Our dataset (https://case.law/) is collected by Harvard Law School, which aimed to expand public access to U.S. case law, by making all published U.S. court decisions online & free to the general public. Case.law contains most case laws from 1658 to 2018, decisions from state courts, federal courts, and territorial courts (American Samoa, Dakota Territory, Guam, Native American Courts, Navajo Nation, Northern Mariana). There are 6.4 million unique cases, 627 reporters, and 40 million pages scanned.

Harvard Law School generated those data through machine OCR (Optical Character Recognition) and additional human review. Some part of the data might have higher quality than others due to human review.

Case.law provides an API for accessing data, which returns the data in JSON format. Only registered users who signed the research agreement are able to view the full case text. For this group, Patrick has signed the agreement and acquired the API_KEY necessary for the data collection. For security reasons, this API_KEY will not be included in this document or other resources. However, the code has a placeholder where one's own API_KEY can be inserted.

Each case contains the basic identification parameters, the reporter information, the jurisdiction information, court information, volume information, and the casebody:

- "id", "url": identifies each unique case.
- "name", "name_abbreviation": name & its abbreviation for each case.
- "decision_date": date of court decision
- "docket_number": docket identification for each case
- "first_page", "last_page": page information for case in printed form
- "citations": citation format for each case. Has paramter "type" and "cite"
- "reporter": contains information for the reporters. Has parameter "url" and "full_name"
- "jurisdiction": contains jurisdiction information for each case. Has parameter "id", "url", "slug", "name", "name_long", and "whitelisted"
- "court": contains court information for each case. Has parameter "id", "url", "slug", "name", and "name_abbreviation"
- "volume": contains volume information for each case. Has parameter "url", "volume_number"
- "casebody": contains the textual information for each case.

    - “opinions”: array containing each opinions for the case. Contains parameter “text”, “author”, and “type” for each opinion.
    - “judges”: name of judges relevant to the case.
    - “head_matter”: summary of case material.
    - “parties”: parties involved in the case.
    - “attorneys”: attorneys involved in the case.

For our project, we focused on the following variables:

- “decision_date”: time information for cases
- “name” (of “jurisdiction”): regional information for cases
- cb_0_(text name): number of appearances of certain category for different cases. It will be further explained in section 4 of this document. These variables are created during feature extraction and refers to the case categories.

2.2 Population data - American Community Survey (ACS)

To validate the completness of case.law, our team inspected the data with another dataset; the population census for U.S. states.

The American Community Survey (ACS), operated by the United States Census Bureau, collects data of United States, on country/state/county basis. ACS provides a R package, “tidycensus”, for efficient data downloading & analysis. A request must be sent to ACS in order to acquire the api key required for data access. For this group, Patrick has signed the agreement and acquired the API_KEY.

For this project, the Total Population Estimate for states (ACS table id: B00001_001) was used. Because the dataset only supports years from 2006-2010 to 2013-2017, the estimation data for 2006-2010 was only used to estimate the total population. While some level of error is expected, this project assumes that the population ratio between states are similar across the years.

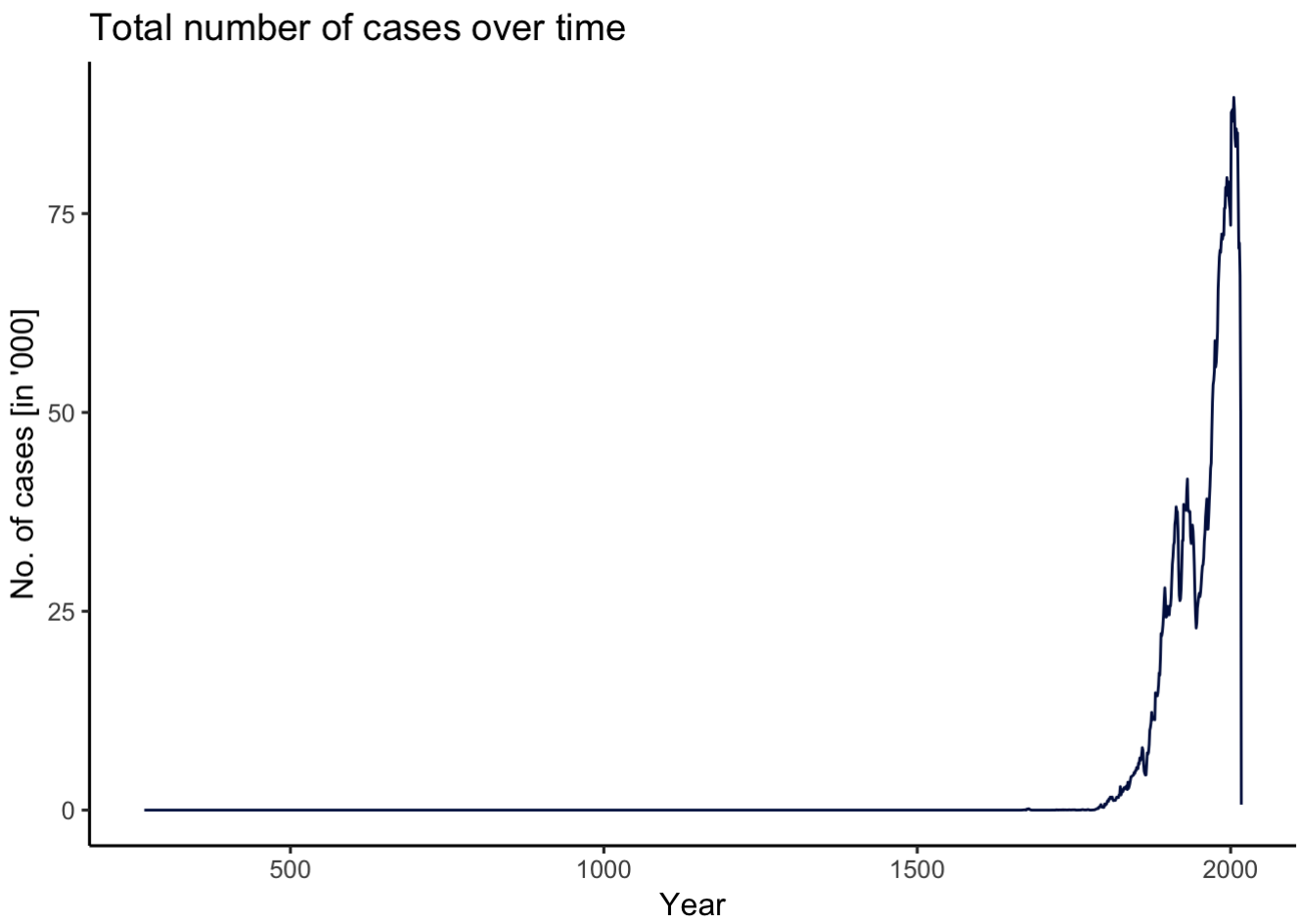
The access and usage of the dataset can be observed in 3.4

3. Data quality

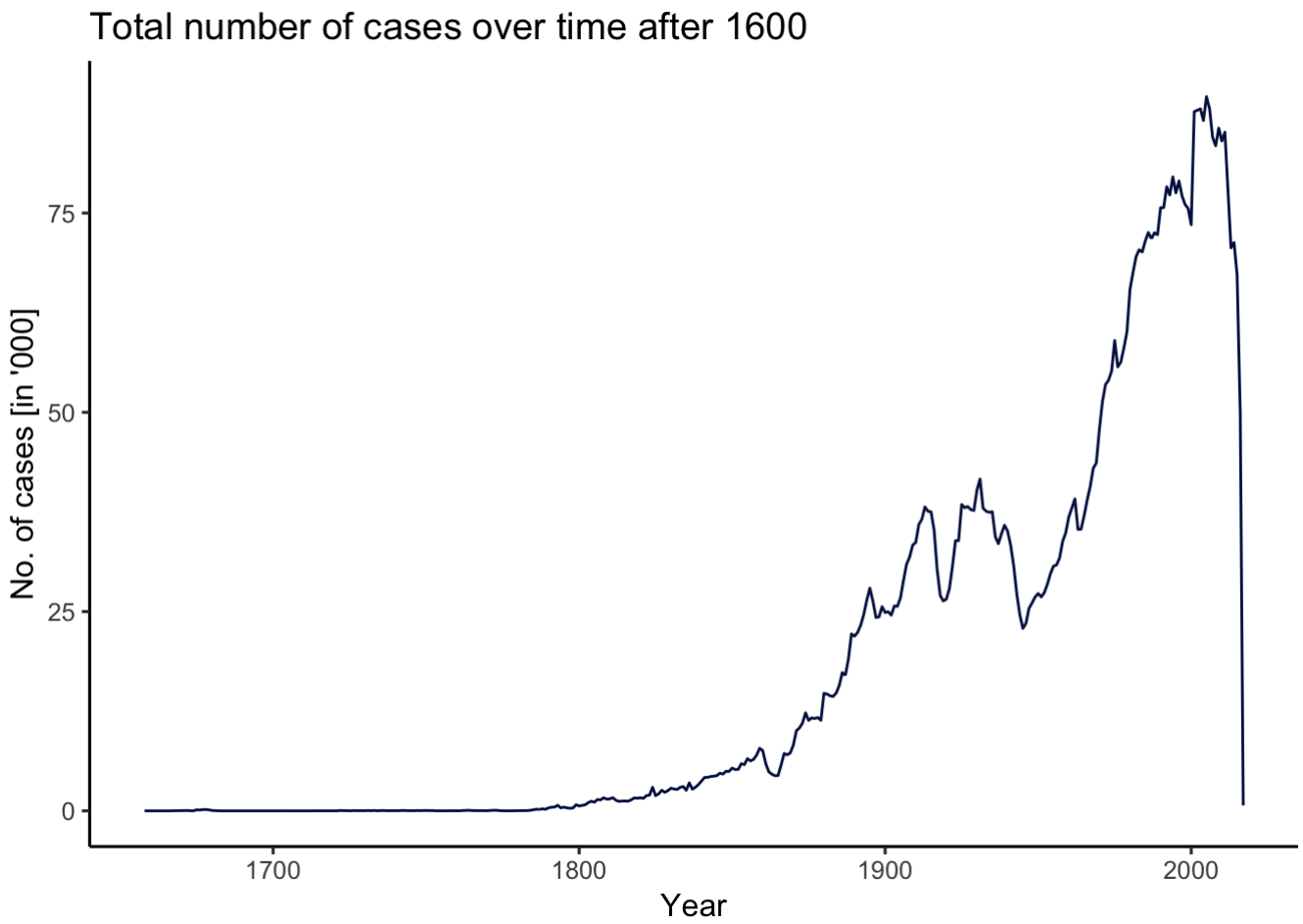
Working with our dataset was quite challenging, as it was comprehensive and it spanned over a long period of 360 years. In order to work with our dataset more efficiently, we made a series of decisions about the process that will be discussed in the following sections.

3.1 Number of cases over time

```
dfb_results <- dfb %>%
  group_by(year) %>%
  summarise(no_cases = sum(no_case))

ggplot(dfb_results, aes(year, no_cases/1000)) +
  geom_path(color="#01144d") +
  ggtitle('Total number of cases over time ') +
  xlab("Year") +
  ylab("No. of cases [in '000]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```




```
ggplot(dfb_results %>% filter(year>=1600), aes(year, no_cases/1000)) +
  geom_path(color="#01144d") +
  ggtitle('Total number of cases over time after 1600') +
  xlab("Year") +
  ylab("No. of cases [in '000]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



The plots above indicate the total number of all cases over time, plotted from the data for the first year that appeared in our dataset until the data for the last year. These plots help indicate if there is significant data over all the years or whether we should focus our analysis on a narrower window.

As the plots show, there are not many cases reported before 1900. Therefore, we will not lose information if we focus our analysis after January 1, 1900. In addition, we found that the data for the years 2016 and 2017 are incomplete, because the total number of cases goes down for these years.

In the process of exploring the number of cases per year, we found one invalid datapoint in the year 267.

```
# Removing outlier
dfb <- dfb %>%
  filter(year>=1600)
```

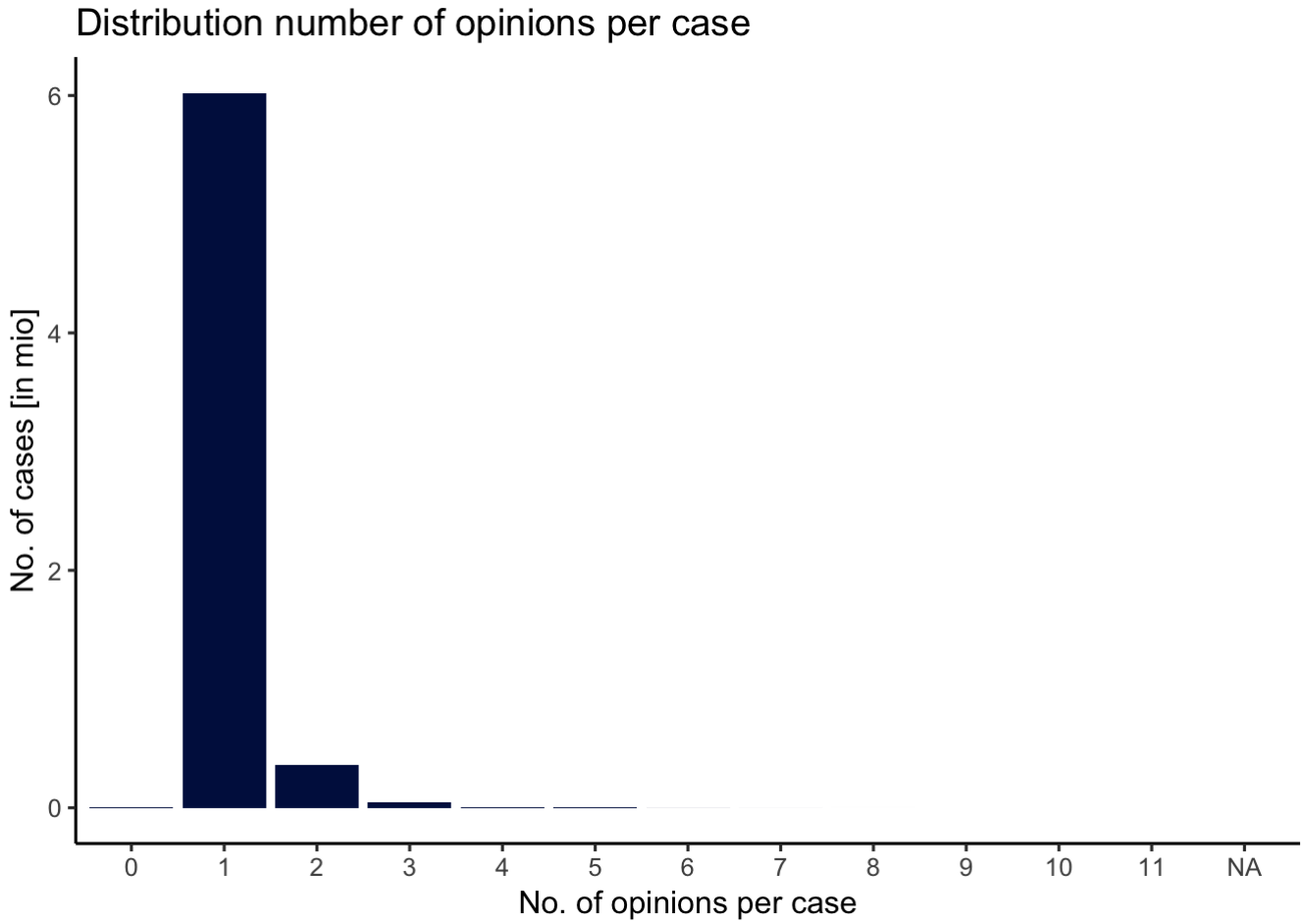
3.2 Array variables and their distriburtions

As the provided dataset from Harvard are JSON objects, a data field can be an array with multiple data objects. Therefore, we analyzed the distribution to understand the missing patterns and the necessity to extract multiple values from it.

In the process of exploring our dataset, we discover following patterns for some featuers:

3.2.1 Number of opinions per case

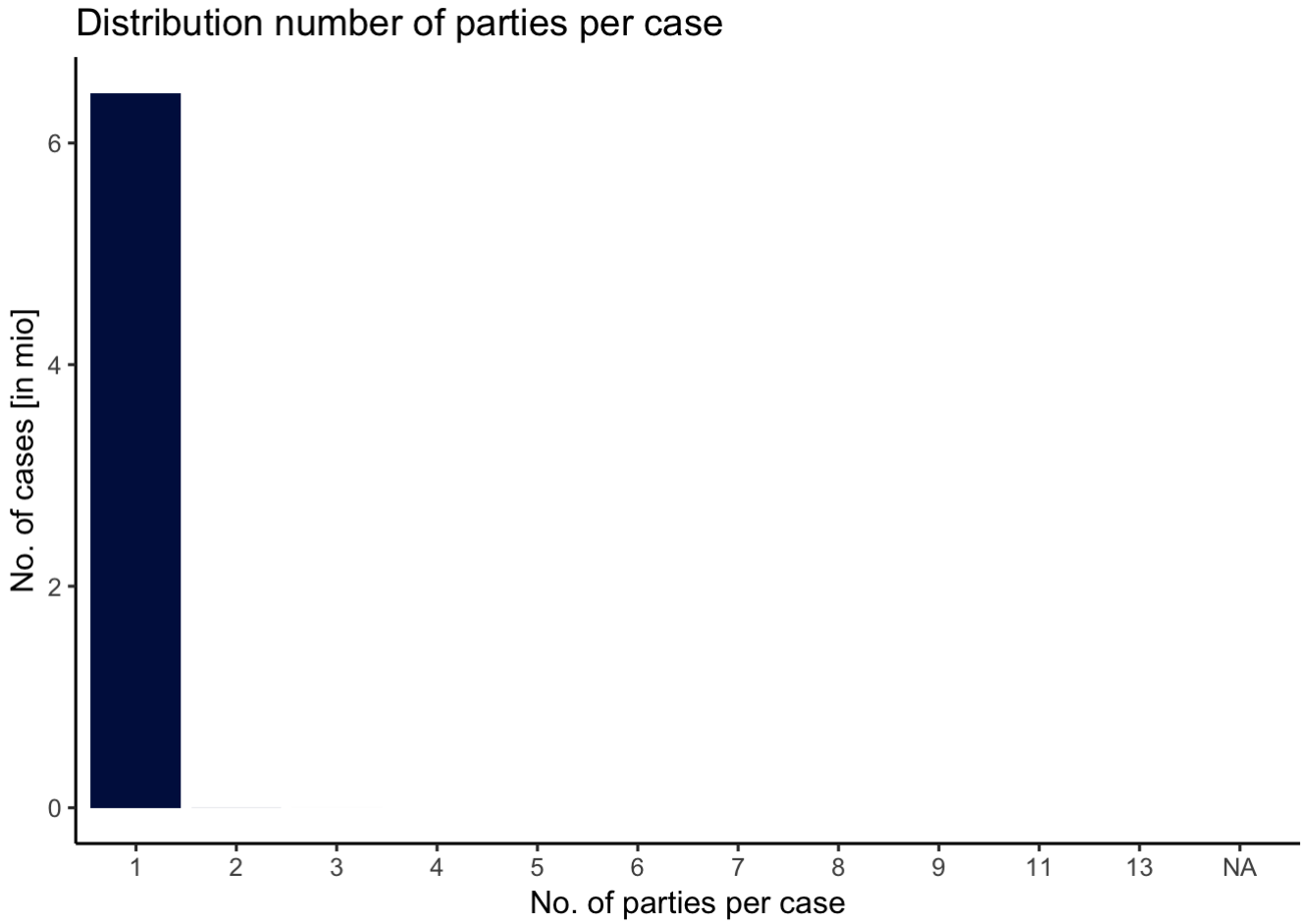
```
ggplot(dfb %>% group_by(cb_no_opinions) %>% summarise(no_cases = n()), aes(x=as.factor(cb_no_opinions), y=no_cases/1000000))+
  geom_bar(stat="identity", fill="#01144d") +
  ggtitle('Distribution number of opinions per case') +
  xlab("No. of opinions per case") +
  ylab("No. of cases [in mio]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



Most cases have only one opinion (93%) and less than 0.5% of cases has 0 opinions. Opinions contains a large corpus of unstructered text, which describes the content of a case. As we focused our analysis on the number of cases, we decided to limit our scope to the first texts in the case body. We extracted case categories from the first opinion with the assumption that different opinions should still document the same case category.

3.2.2 Number of parties per case

```
ggplot(dfb %>% group_by(cb_no_parties) %>% summarise(no_cases = n()), aes(x=as.factor
(cb_no_parties), y=no_cases/1000000))+
  geom_bar(stat="identity", fill="#01144d") +
  ggtitle('Distribution number of parties per case') +
  xlab("No. of parties per case") +
  ylab("No. of cases [in mio]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```

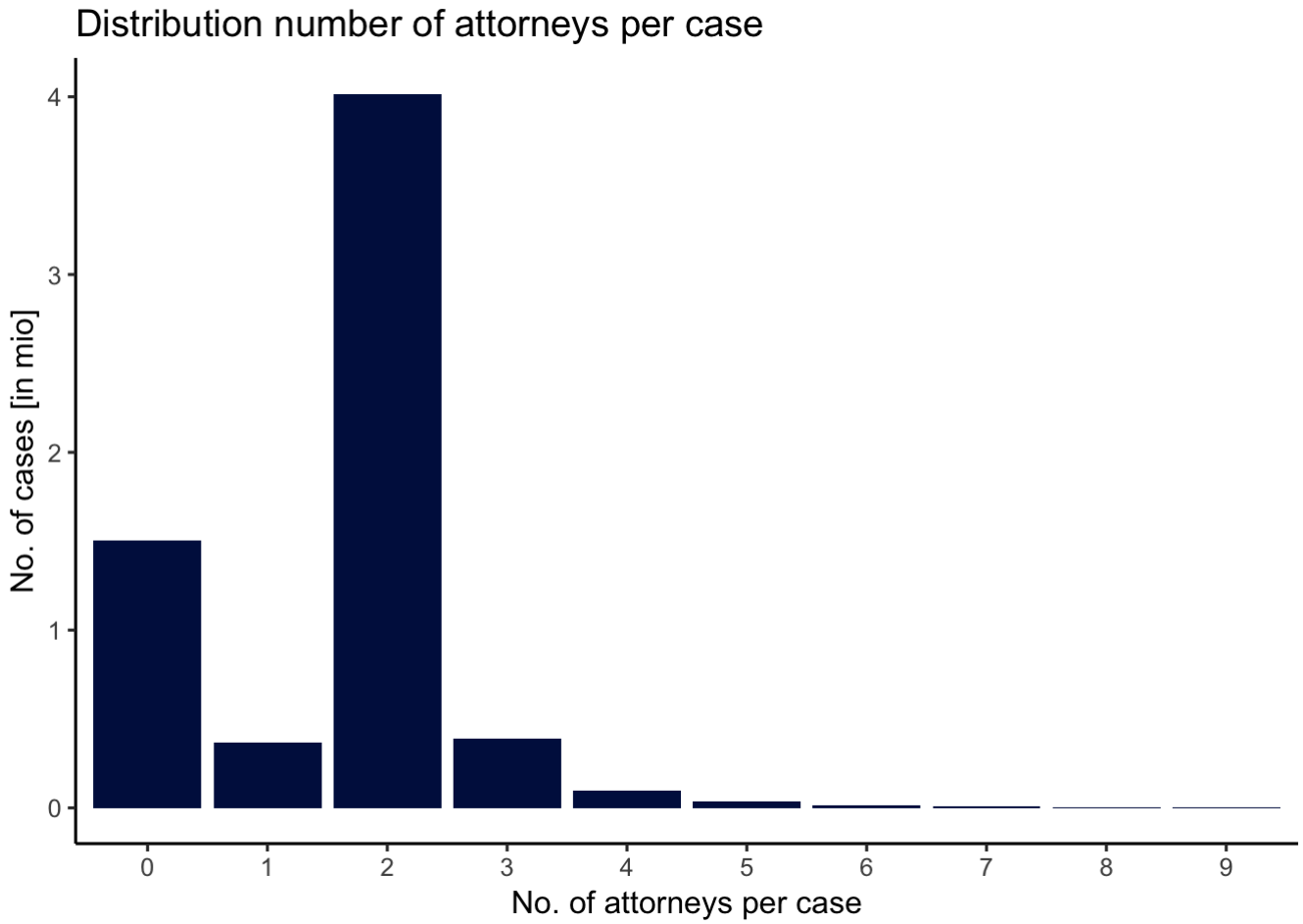


Most cases have only one party (99%).

3.2.3 Number of attorneys per case

```
ggplot(dfb %>% group_by(cb_no_attorneys) %>% summarise(no_cases = n()) %>% arrange(cb
_no_attorneys) %>% top_n(10), aes(x=as.factor(cb_no_attorneys), y=no_cases/1000000))+
  geom_bar(stat="identity", fill="#01144d") +
  ggtitle('Distribution number of attorneys per case') +
  xlab("No. of attorneys per case") +
  ylab("No. of cases [in mio]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```

```
## Selecting by no_cases
```

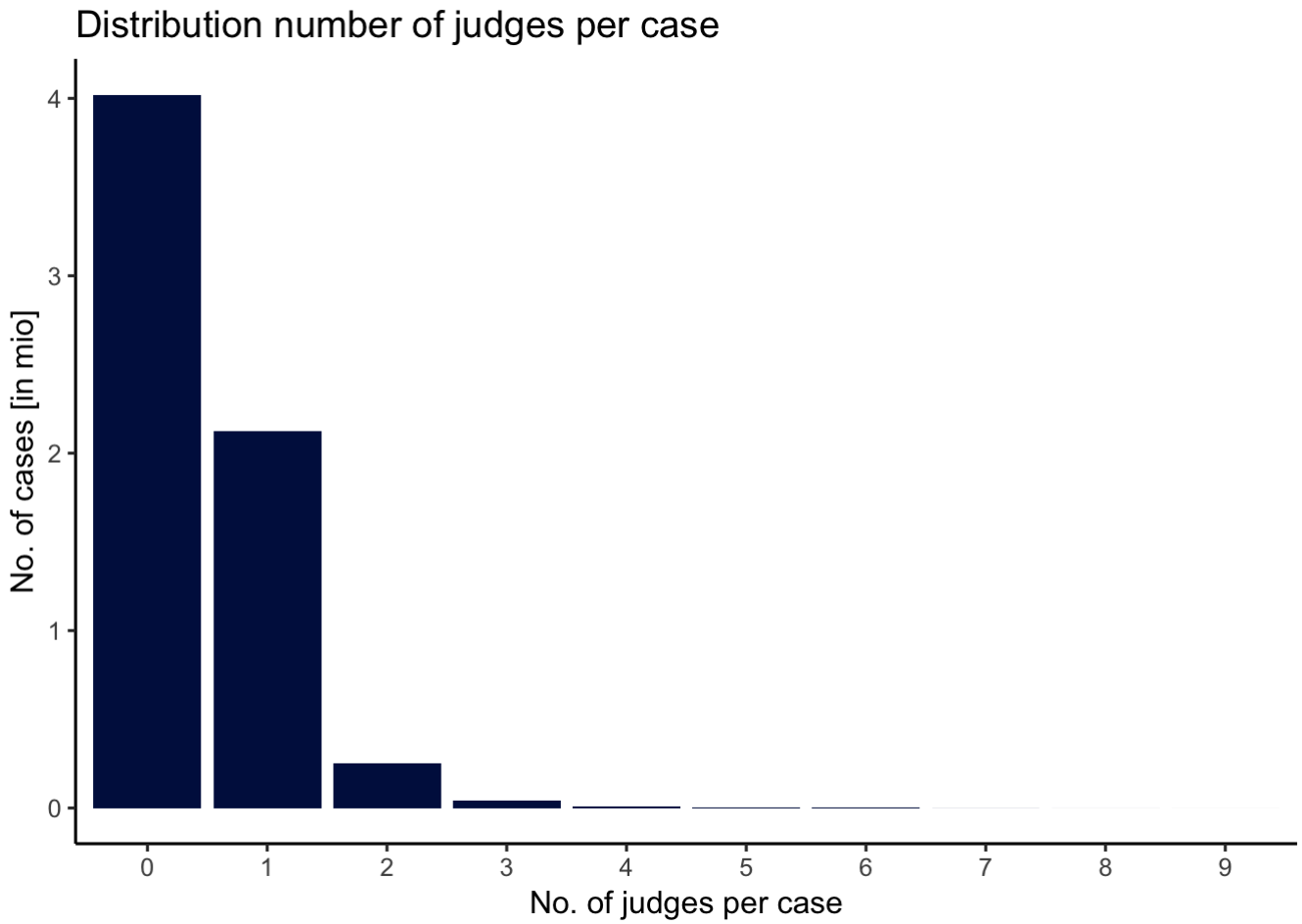


Most cases have two attorneys (62%), but over 1.5 mio cases have no attorney reported.

3.2.4 Number of judges per case

```
ggplot(dfb %>% group_by(cb_no_judges) %>% summarise(no_cases = n()) %>% arrange(cb_no
_judges) %>% top_n(10), aes(x=as.factor(cb_no_judges), y=no_cases/1000000))+
  geom_bar(stat="identity", fill="#01144d") +
  ggtitle('Distribution number of judges per case') +
  xlab("No. of judges per case") +
  ylab("No. of cases [in mio]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```

```
## Selecting by no_cases
```



Judges are not well documented in the case law dataset. Over 4 mio cases has no judge assigned (62%).

Interestingly, our dataset had no information about the outcome of the cases.

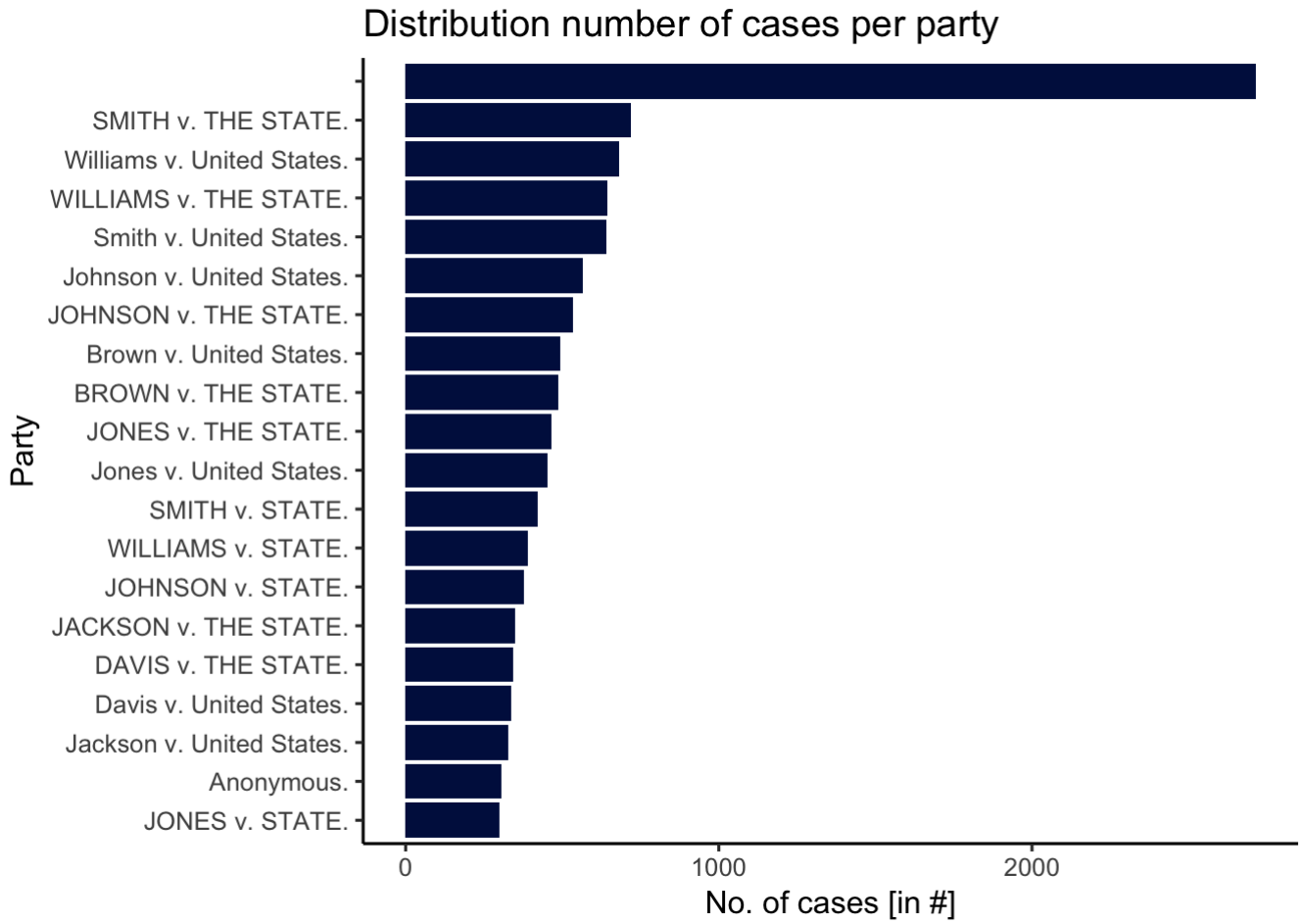
3.3 Pseudo categorical variables

From the array field “judge”, “atterney” and “party”, we extract always the first element and analyse their distribution. As there should be only a limited number of distinct values per field, we may can use the variables as categorical variables. For example, there should be only a limited number of different judges in the United States.

3.3.1 Number of cases per party

```
ggplot(dfb %>% group_by(cb_data_party1) %>% filter(cb_data_party1 != "") %>% summaris
e(no_cases = n()) %>% arrange(desc(no_cases)) %>% top_n(20),
       aes(x=reorder(cb_data_party1, sort(as.numeric(no_cases), decreasing = TRUE)),
 y=no_cases))+
  geom_bar(stat="identity", fill="#01144d") +
  ggtitle('Distribution number of cases per party') +
  xlab("Party") +
  ylab("No. of cases [in #]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black")) +
  coord_flip()
```

```
## Selecting by no_cases
```



Note: Only top 20 values are plotted Empty values are removed.

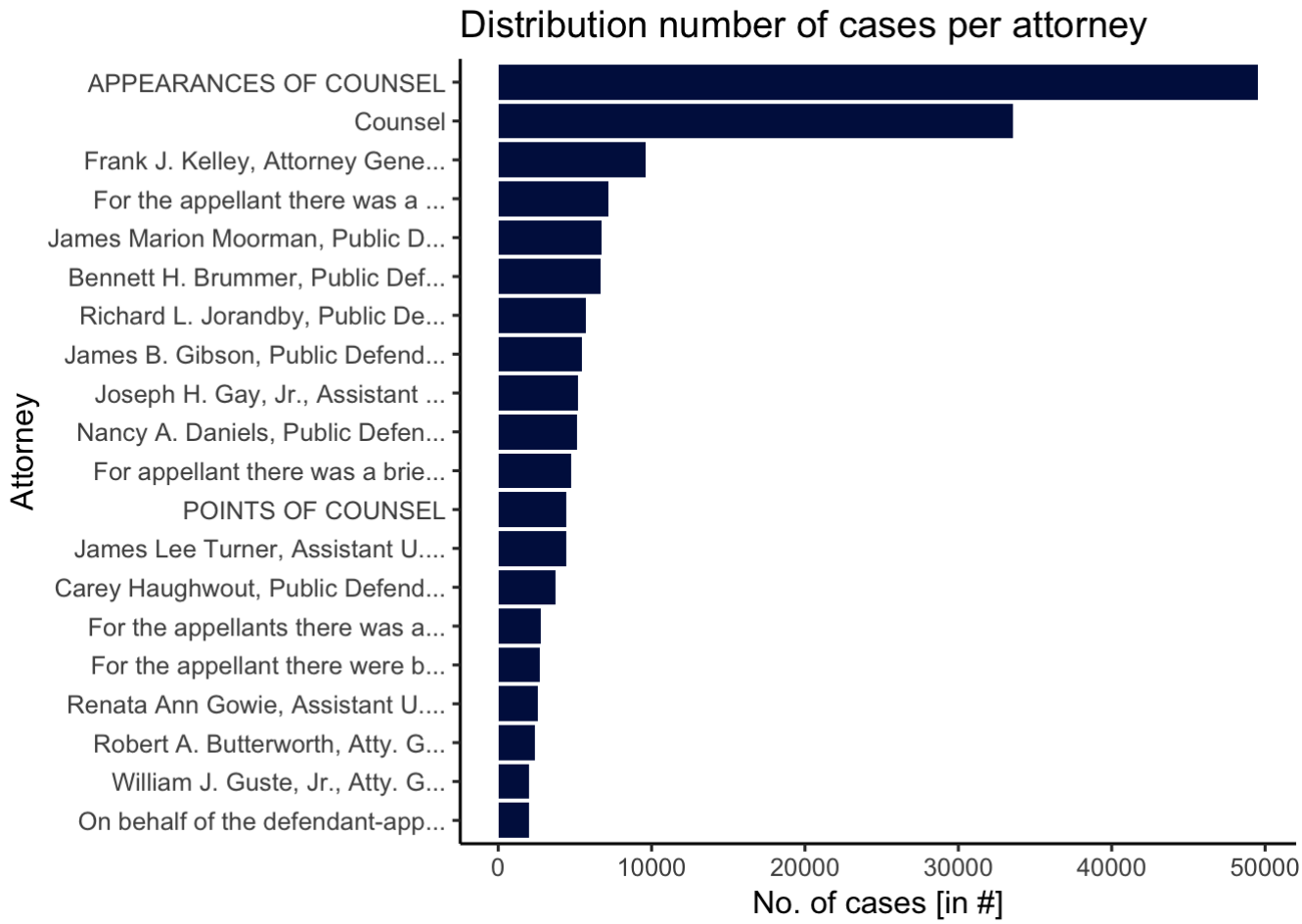
The most frequent, non-empty party is “SMITH v. THE STATE” with less than 1000 cases. A hypothesis is, that “SMITH” could be different people as the name is common in the US. We don’t continue to extract meaningful insights from this variable as there is no dominant party.

Possible next step: We could extract the text “STATES” as one party and analyse how often the “STATE” is represented.

3.3.2 Number of cases per attorney

```
ggplot(dfb %>% mutate(cb_data_attr1_first_attorneys = ifelse(nchar(cb_data_attr1_firs
t_attorneys)>30,
                                                    paste0(substr(cb_data_attr1_first_attor
neys, 1, 30),'...'),cb_data_attr1_first_attorneys)
                      ) %>% filter(cb_data_attr1_first_attorneys != "") %>%
         group_by(cb_data_attr1_first_attorneys) %>% summarise(no_cases = n()) %>% ar
range(desc(no_cases)) %>% top_n(20),
       aes(x=reorder(cb_data_attr1_first_attorneys, sort(as.numeric(no_cases), decrea
sing = TRUE)), y=no_cases))+
  geom_bar(stat="identity", fill="#01144d") +
  ggtitle('Distribution number of cases per attorney') +
  xlab("Attorney") +
  ylab("No. of cases [in #]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black")) +
  coord_flip()
```

```
## Selecting by no_cases
```



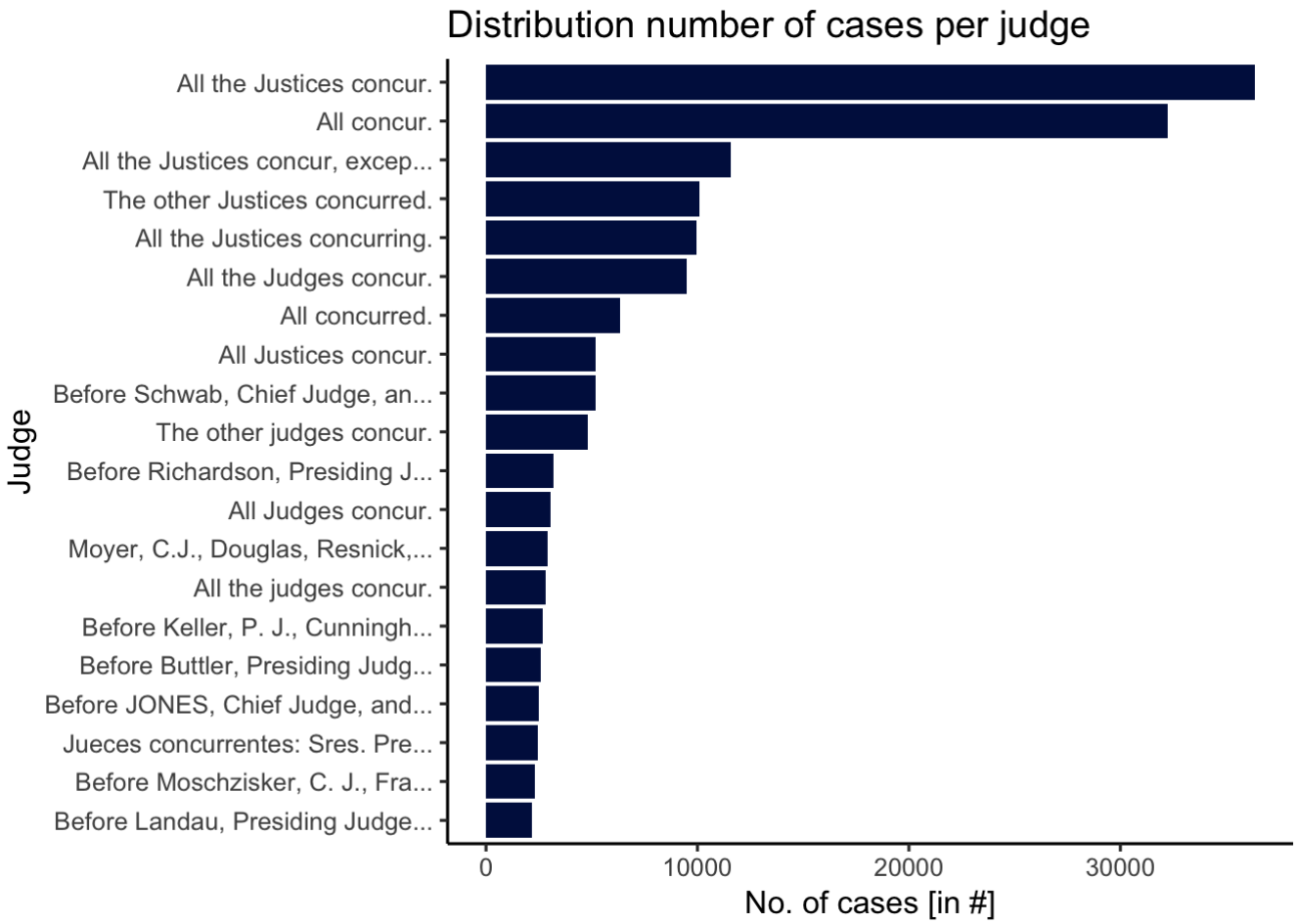
Note: Only top 20 values are plotted. Names are truncated after 30 characters. Empty values are removed.

There are two generic, non-empty attorneys related to “COUNSEL” with ~50,000 and ~33000 cases. The next attorneys are real people with names - and have between 700 and 4500 cases. In addition, we can see their role, such as “Attorney General” or “Public Defender”.
In some cases, there is a description instead of a name in the field.

3.3.3 Number of cases per judge

```
ggplot(dfb %>% mutate(cb_data_judge1 = ifelse(nchar(cb_data_judge1)>30,
                                              paste0(substr(cb_data_judge1, 1, 30),
'...'),cb_data_judge1)
                      ) %>% filter(cb_data_judge1 != "") %>%
         group_by(cb_data_judge1) %>%
         summarise(no_cases = n()) %>%
         arrange(desc(no_cases)) %>% top_n(20), aes(x=reorder(cb_data_judge1, sort(a
s.numeric(no_cases), decreasing = TRUE)), y=no_cases)) +
  geom_bar(stat="identity", fill="#01144d") +
  ggtitle('Distribution number of cases per judge') +
  xlab("Judge") +
  ylab("No. of cases [in #]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black")) +
  coord_flip()
```

```
## Selecting by no_cases
```



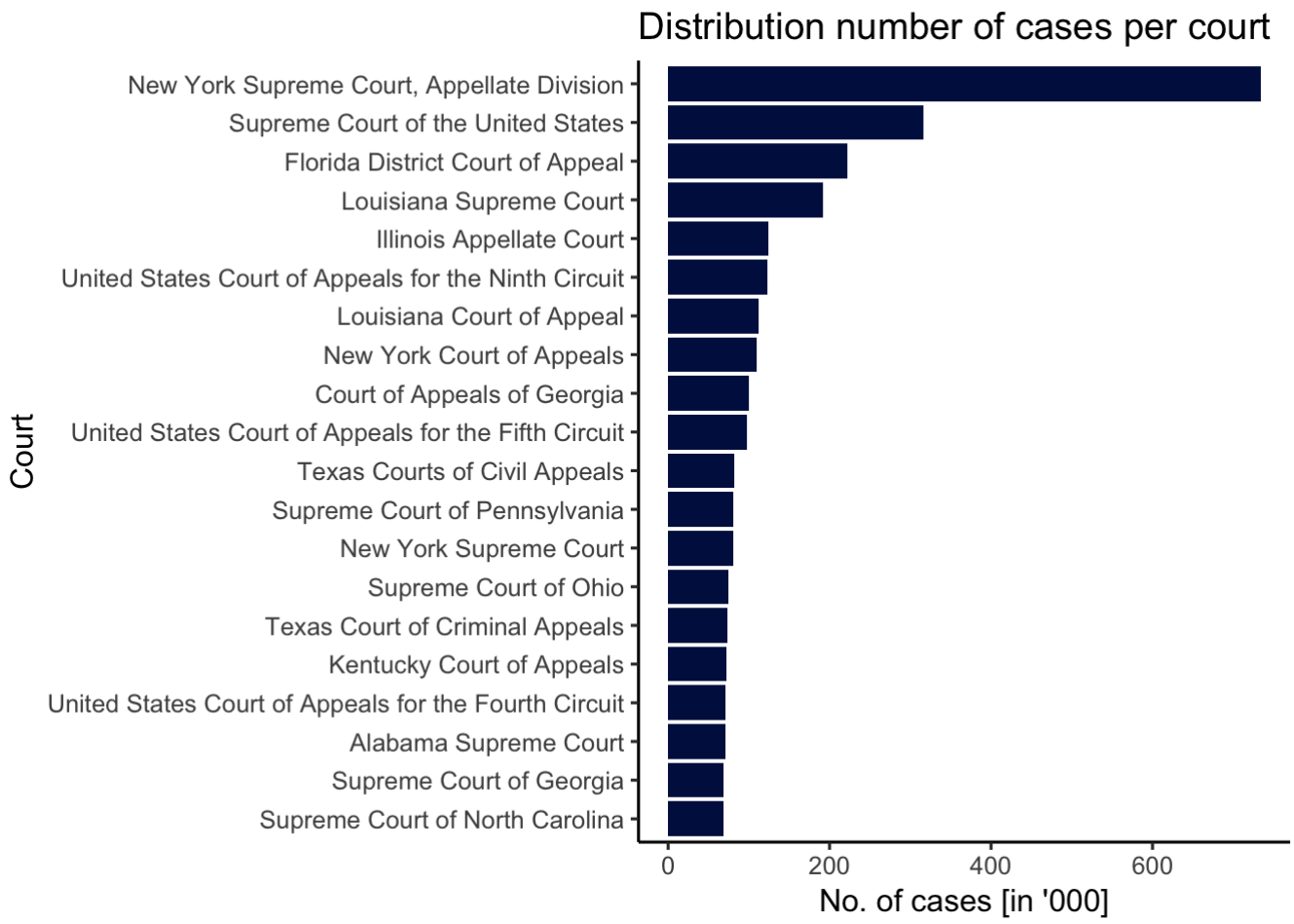
Note: Only top 20 values are plotted. Names are truncated after 30 characters. Empty values are removed.

The variable judge is to 62% empty and the next non-empty value is a generic description. The data quality of this variable seems low and we don't continue using it.

3.3.4 Number of cases per court

```
ggplot(dfb %>% group_by(court_name) %>% summarise(no_cases = n()) %>% arrange(desc(no
_cases)) %>% top_n(20),
       aes(x=reorder(court_name, sort(as.numeric(no_cases), decreasing = TRUE)), y=no
_cases/1000))+
  geom_bar(stat="identity", fill="#01144d") +
  ggtitle('Distribution number of cases per court') +
  xlab("Court") +
  ylab("No. of cases [in '000]") +
  theme_grey(12) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black")) +
  coord_flip()
```

```
## Selecting by no_cases
```



Note: Only top 20 values are plotted.

The plot shows the number of cases per court. The majority of the cases in our dataset are taken from the Appellate Division of the New York Supreme Court. There are 3600 unique court names in the dataset, with no missing data for courts.

3.4 A closer look: Number of cases per state vs. population

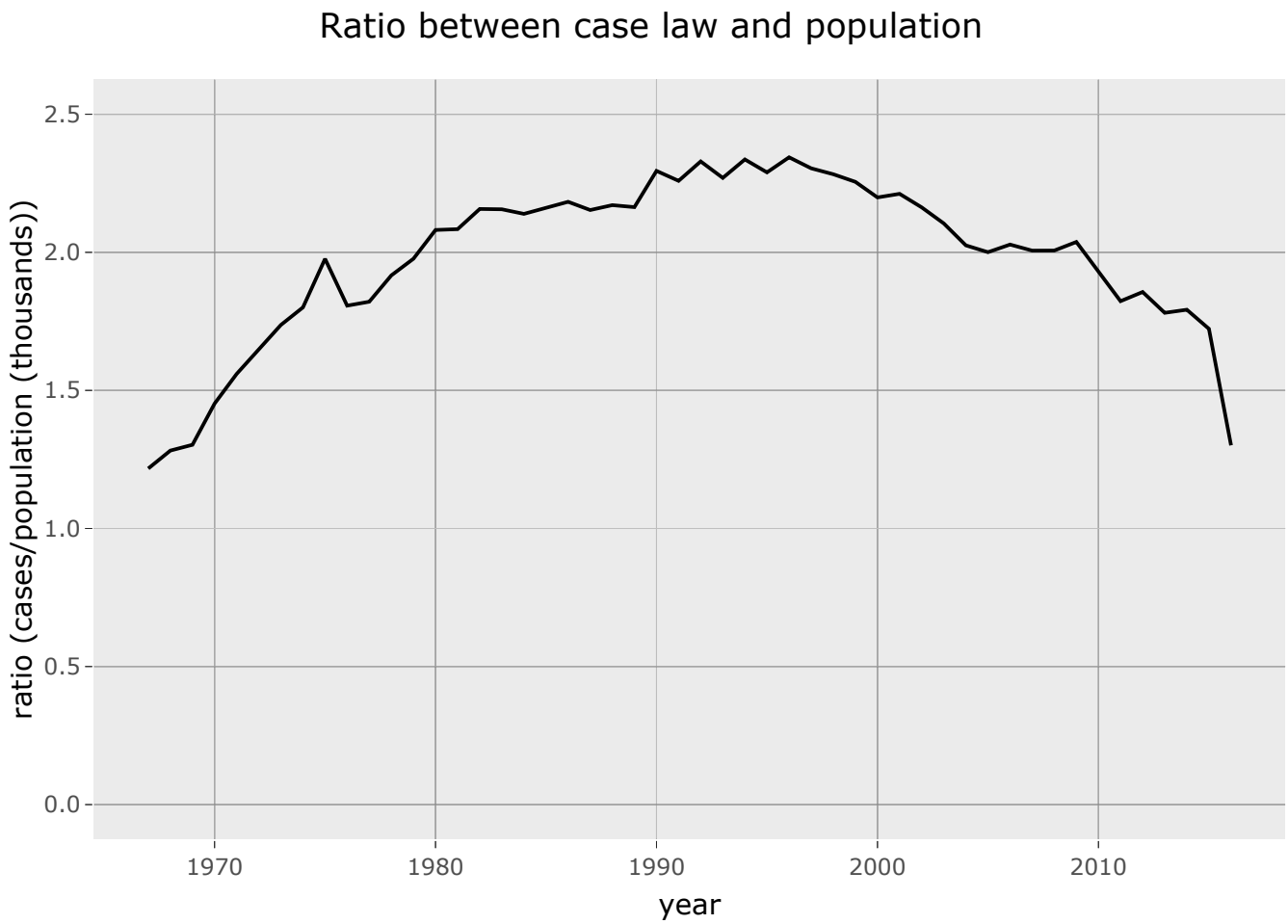
The number of court cases for each state were divided by the population estimate for 2006-2010 (by thousands) and was compared across different years, to see if there was a increase in court cases per population.

```
date_lim1 = "1967-01-01"
date_lim2 = "2017-01-01"

start_year = as.integer(strsplit(date_lim1, "-")[[1]][1])
end_year = as.integer(strsplit(date_lim2, "-")[[1]][1])-1

file_list <- list.files(path='../data/02_processed_csvs/', pattern="*text.csv")

totalList = data.frame(year=integer(), total=integer(), region=character())

for (name in file_list){
  region <- (strsplit(name, "-")[[1]][1])
  data <- fread(paste0('../data/02_processed_csvs/', name), sep=",")
  data <- data %>% group_by(decision_date) %>% summarise(total = n())

  data$decision_date <- as.Date(data$decision_date, format="%Y-%m-%d")
  data <- data %>% na.omit() %>% filter(decision_date >= as.Date(date_lim1) & decisio
n_date < as.Date(date_lim2))
  if(nrow(data)>0){
    data['year'] <- as.integer(format(data$decision_date, "%Y"))
    data <- data %>% group_by(year) %>% summarise(total = sum(total))
    data['region'] <- tolower(region)
    totalList <- rbind(totalList, data)
  }
}

pop <- get_acs(geography = "state", variables = c(popluation = "B00001_001"), year=20
10, key = "...")
```

```
## Getting data from the 2006-2010 5-year ACS
```

```
pop$NAME <- tolower(pop$NAME)
colnames(pop) <- c('GEOID', 'region', 'variable', 'estimate')
pop$estimate <- pop$estimate/1000
totalList <- merge(totalList, pop, by='region')

totalList$value = totalList$total / totalList$estimate
totalList <- totalList[, c("year", "region", "total", "estimate", "value")]

keyword='total'
dir.create(file.path("../data/05_finalreport_plots/", keyword), showWarnings = FALSE)
write.csv(totalList, file = paste0("../data/05_finalreport_plots/", keyword, "/", key
word, "_", toString(start_year), "_", toString(end_year), ".csv"))
```

```
totalList2 <- totalList %>% group_by(year) %>% summarise(total = sum(total), populati
on = sum(estimate))
totalList2$ratio <- totalList2$total / totalList2$population

p <- ggplot(data = totalList2, aes(x = year, y = ratio)) + ggtitle("Ratio between cas
e law and population") + geom_line() +
  scale_y_continuous(name="ratio (cases/population (thousands))", limits=c(0, 2.5))
(gg <- ggplotly(p))
```

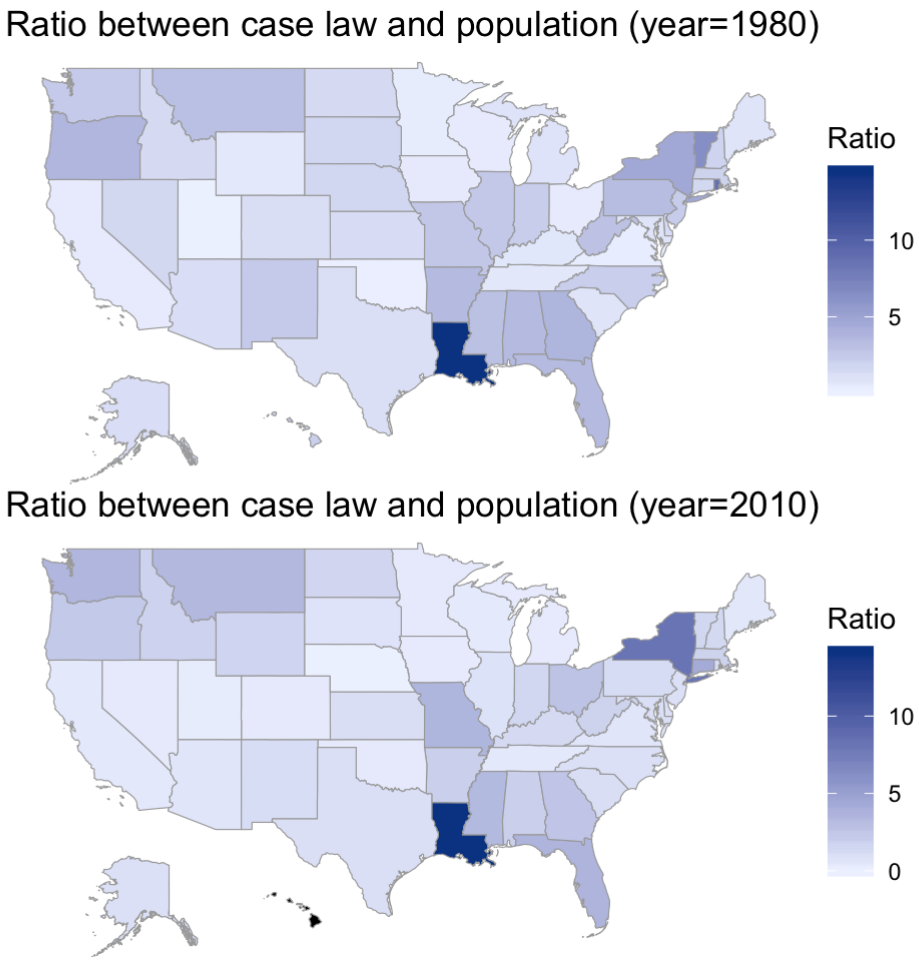


The total number of cases divided by population estimate shows a concave curve; a steadily increasing pattern from 1967 to the 1990s, then a decreasing pattern until 2016.

```
data(df_state_demographics)

mapData1 <- totalList %>% filter(year == 1980)
mapData1 <- mapData1[mapData1$region %in% df_state_demographics$region, c('region',
'value')]

c = StateChoropleth$new(mapData1)
c$title = "Ratio between case law and population (year=1980)"
c$legend = "Ratio"
c$set_num_colors(1)
c$set_zoom(NULL)
c$show_labels = FALSE
choropleth_state1 = c$render()

mapData2 <- totalList %>% filter(year == 2010)
mapData2 <- mapData2[mapData2$region %in% df_state_demographics$region, c('region',
'value')]
df<-data.frame("nebraska",0)
names(df)<-c("region","value")
mapData2 <- rbind(mapData2, df)

c = StateChoropleth$new(mapData2)
c$title = "Ratio between case law and population (year=2010)"
c$legend = "Ratio"
c$set_num_colors(1)
c$set_zoom(NULL)
c$show_labels = FALSE
choropleth_state2 = c$render()
```

```
## Warning in self$bind(): The following regions were missing and are being
## set to NA: hawaii
```

```
grid.arrange(choropleth_state1, choropleth_state2, nrow=2)
```



The influence of time on the ratio is relatively small. The ratio for each state seems to have are mostly maintained.

The case law ratio differs a lot for different states. States like Louisiana, District of Columbia, and New York seems to have very high ratio, while other states like Utah, Minnesota, and South Dakota are below 1. Among those states, Louisiana mostly ranks as the top state for the ratio of cases, which might result from a more diverse set of laws due to its base on French and Spanish law system.

4. Exploratory data analysis

4.1 Case category extraction

As it was analyzed in 3. Data Quality, almost all cases contain one opinion text (only few have no opinion text or more than 1). We assume that the case category should be consistent over multiple opinions - e.g. the category type “sexual harassment” should not change from opinion 1 to 2. Therefore, we consider only the first case opinion text.

As the scope of this projects does not allow modelling techniques, we restricted ourself to basic text mining and keyword extraction:

First, we conduct following preprocessing: 1. Lowercase all letters in the text 2. Replace all non-alpha characters and not “&” (regex: [^0-9a-zA-Z&]+) with a space 3. Replace multiple space with a single splace

Second, we count the number of appearance of a keyword in the text. For example, how often “sexual harassment” appeared in the opinion text.

One limitation is that multiple categories could appear in the same text. We solved this issue by discussing similar categories together (e.g. “insurance claim”" and “mal practice”) and at the same time having distinct categories (e.g. “sexual harassment” and “insurance claim”).

We would prefer more advanced methods, such as, clustering the cases with k-neigherst-neighbours first and extract common word similarity. With this method, we would find bigger cluster and more distinct categories.

4.2 Number of cases over time

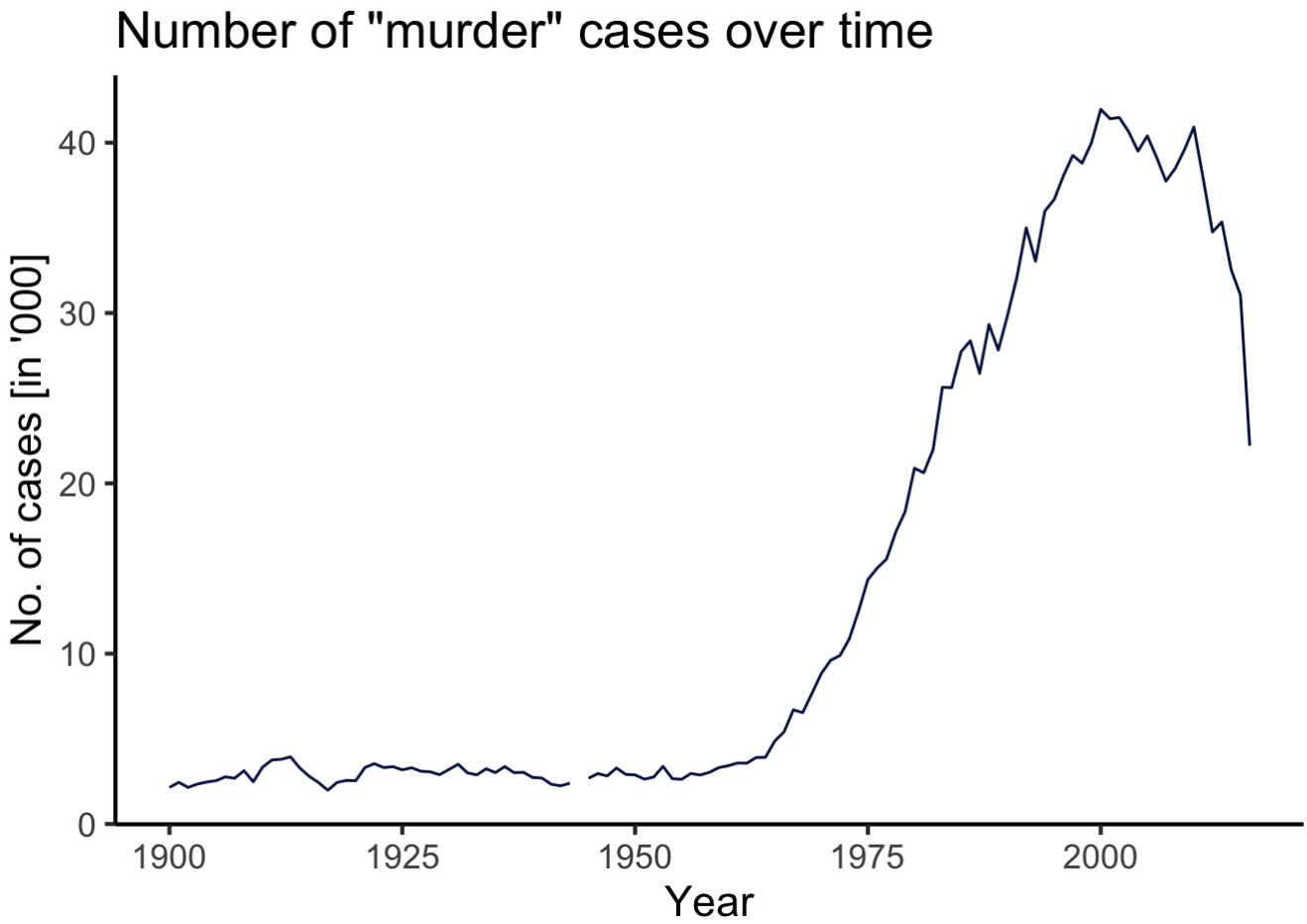
```
colnames(dfb) <- make.names(colnames(dfb))
rel_colnames <- c('cb_data_text_0_medical.malpractice',
                  'cb_data_text_0_drug.recall',
                  'cb_data_text_0_antitrust.prosecution',
                  'cb_data_text_0_asylum',
                  'cb_data_text_0_murder',
                  'cb_data_text_0_arson',
                  'cb_data_text_0_sexual.harassment',
                  'cb_data_text_0_divorce',
                  'cb_data_text_0_intellectual.property',
                  'cb_data_text_0_insurance.claims',
                  'cb_data_text_0_free.speech',
                  'cb_data_text_0_capital.murder',
                  'cb_data_text_0_patent')

dfb_grouped <- dfb %>%
  select_(.dots = c('year', rel_colnames)) %>%
  group_by(year) %>%
  summarise_all(funs(sum))

dfb_grouped_tidy <- dfb_grouped %>%
  gather('Case.Type', 'No', -year)
```

The number of cases that are filed in court of the United states depends likely on societal and political factors. Therefore, interesting patterns can be observed by analysing the number of cases per categories over time.

The below line plots show the trends in the number of cases for a categories over time. As discussed in previous sections, we focus our analysis on between 1900 and 2016. There are cases with significant historical trends reported in the following categories: murder, sexual harassment, medical malpractice, insurance claim

4.2.1 Murder

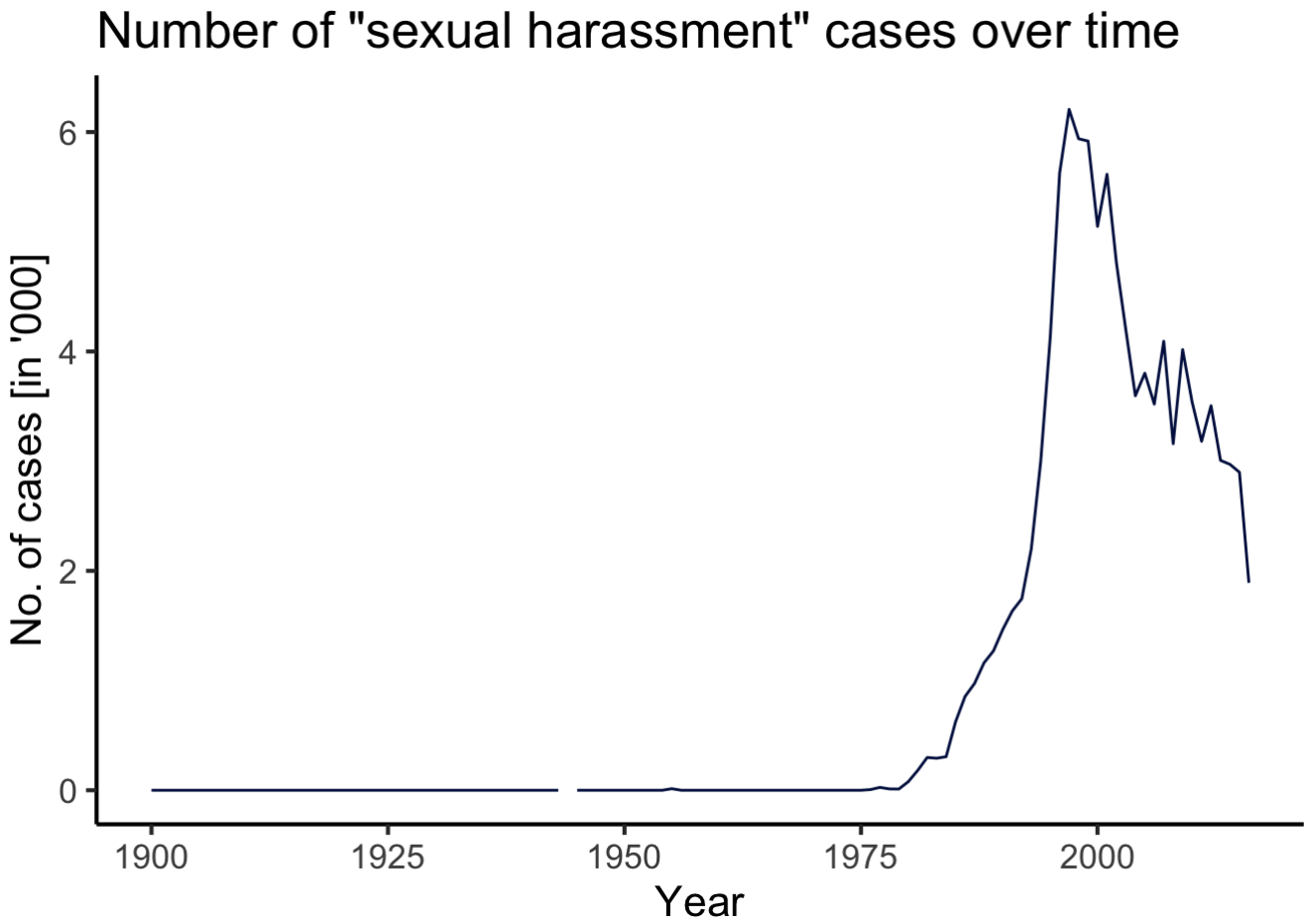
```
ggplot(dfb_grouped_tidy %>% filter(Case.Type == 'cb_data_text_0_murder') %>% filter(y
ear >= 1900 & year <= 2016), aes(year, No/1000))+
  geom_path(color="#01144d") +
  ggtitle('Number of "murder" cases over time ') +
  xlab("Year") +
  ylab("No. of cases [in '000]") +
  theme_grey(16) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



In the category of murder, there is a relatively small number of cases reported starting from 1900, reaching a maximum of ~2500 murder cases in years 1912 and 1925. There is significant increase in the rate of number of cases from around 1963 until 2000 with a maximum of around 40000 cases. After the year 2000 there is a decline in the number of cases.

4.2.2 Sexual harassment

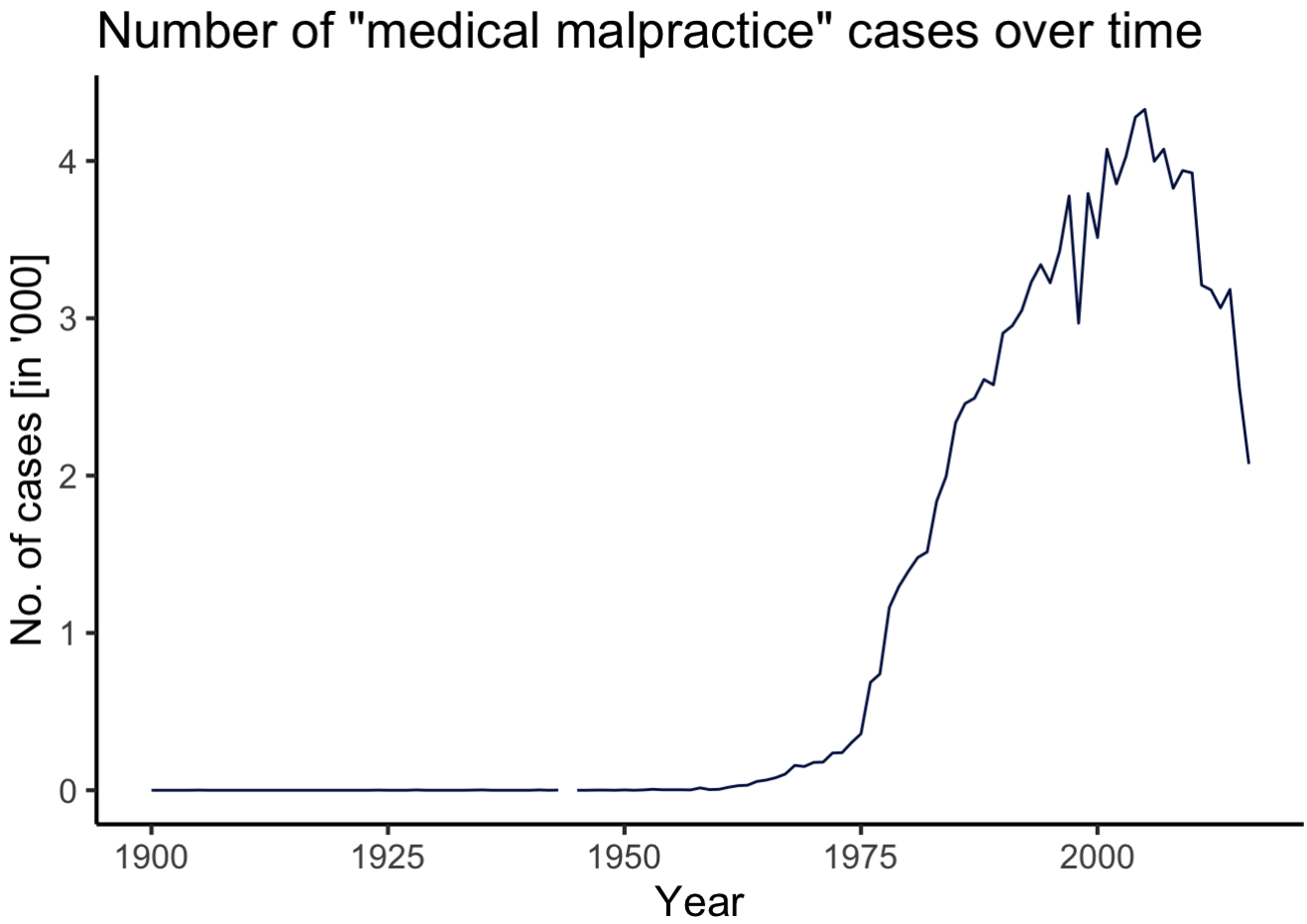
```
ggplot(dfb_grouped_tidy %>% filter(Case.Type == 'cb_data_text_0_sexual.harassment') %
>% filter(year >= 1900 & year <= 2016), aes(year, No/1000))+
  geom_path(color="#01144d") +
  ggtitle('Number of "sexual harassment" cases over time ') +
  xlab("Year") +
  ylab("No. of cases [in '000]") +
  theme_grey(16) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



In the category of sexual harassment cases, there are no cases before 1975. After this point there is a sharp increasing trend in the number of cases reported, until a maximum of around 6000 cases is reached around the year 2000.

4.2.3 Medical malpractice

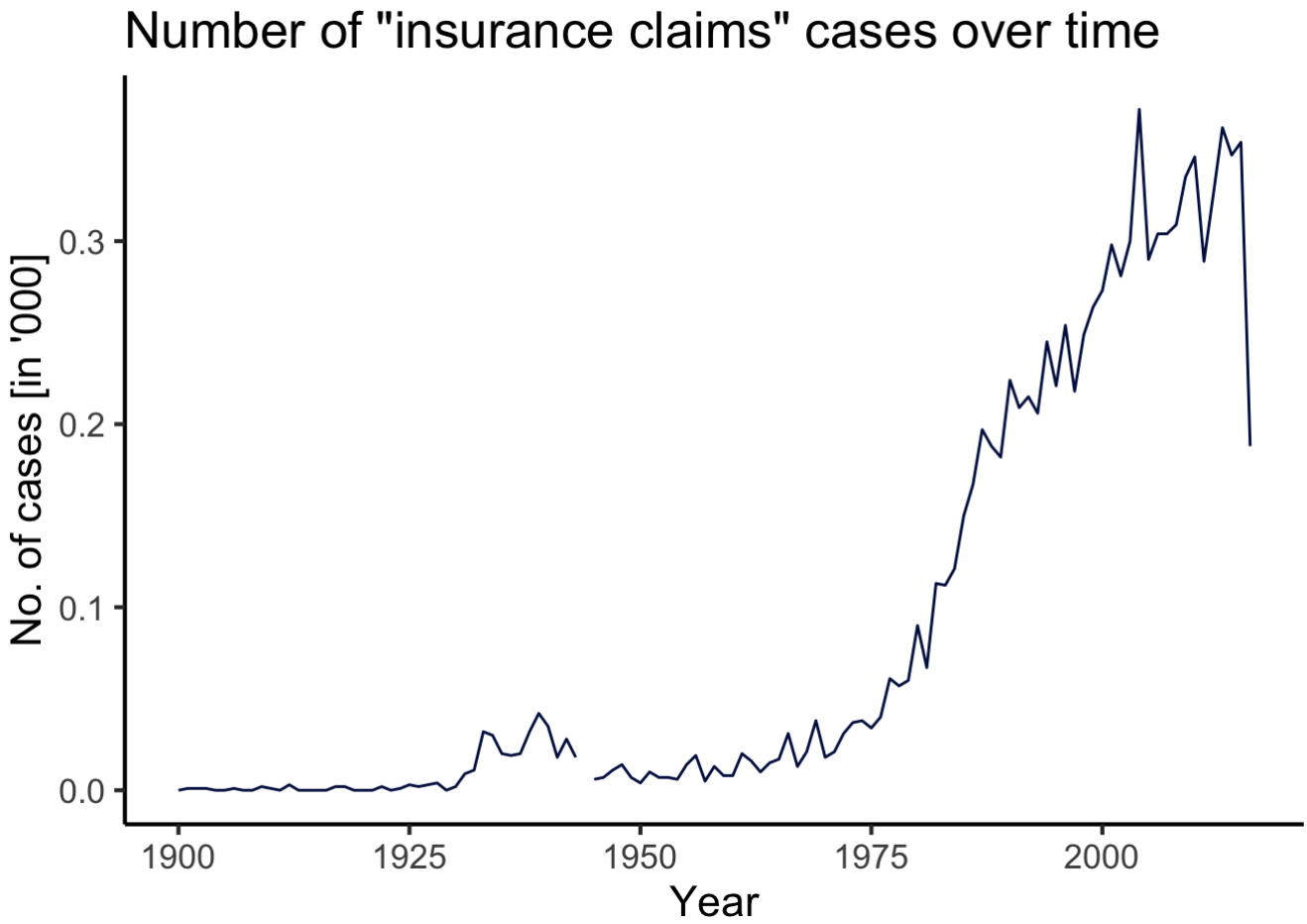
```
ggplot(dfb_grouped_tidy %>% filter(Case.Type == 'cb_data_text_0_medical.malpractice')
 %>% filter(year >= 1900 & year <= 2016), aes(year, No/1000))+
  geom_path(color="#01144d") +
  ggtitle('Number of "medical malpractice" cases over time ') +
  xlab("Year") +
  ylab("No. of cases [in '000]") +
  theme_grey(16) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



In the cases in category of medical malpractice, we see no occurrence of cases before the 1950's. This category of cases experiences a sharp increase starting from around 1975, reaching a maximum around the year 2000, after which there is a decrease in these cases per year.

4.2.4 Insurance claim

```
ggplot(dfb_grouped_tidy %>% filter(Case.Type == 'cb_data_text_0_insurance.claims') %
>% filter(year >= 1900 & year <= 2016), aes(year, No/1000))+
  geom_path(color="#01144d") +
  ggtitle('Number of "insurance claims" cases over time ') +
  xlab("Year") +
  ylab("No. of cases [in '000]") +
  theme_grey(16) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



In the category of insurance claims, there are a number cases that are reported starting in the year 1900. After 1975, the rate at which insurance claim cases are reported accelerates, reaching the maximum number of 350 cases in 2016.

4.3 Relative frequency of cases per states

In the following sections, we will dive deeper into the spatial patterns for different court cases, for years 1967-2016.

Our team used the choroplethr package for spatial visualization. For each state and a certain case type (ex: intellectual property), the statistic for each state is acquired by dividing the number of certain case with the number of total cases of that state, to negate the influence of population difference between states. Because of low population, alaska and hawaii were ignored.

One challenge our team met was the limited functionality of the choroplethr package. The package did not provided detailed adjustments to the plot, such as font size change, font location change, color plot change, and so on. To achieve the optimal visualization we had to neglect from using the main functions and instead declared choroplethr objects and changed their properties manually.

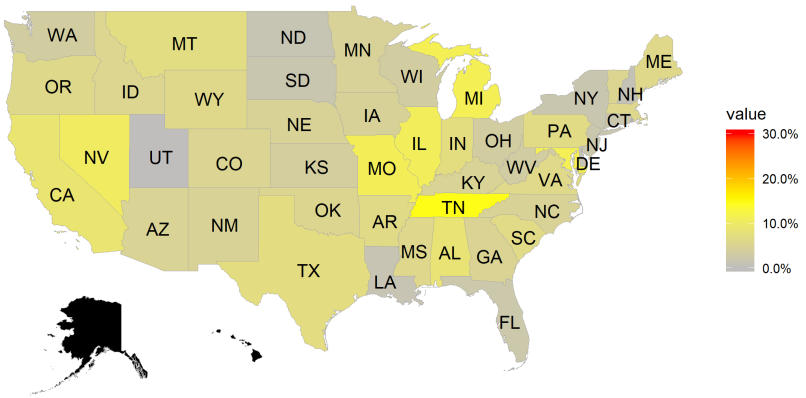
The R script used to create the following spatial plots is choro.R. On activation, the program will generate a csv file with the relevant data, 50 choroplethr images for the different years, and a html file for interactivity.

One disadvantage of our choroplethr plots was that it was not effective on showing the statistics for small states, such as District of Columbia. If our group had more time, we might have tried a different graphic format, such as a cartogram.

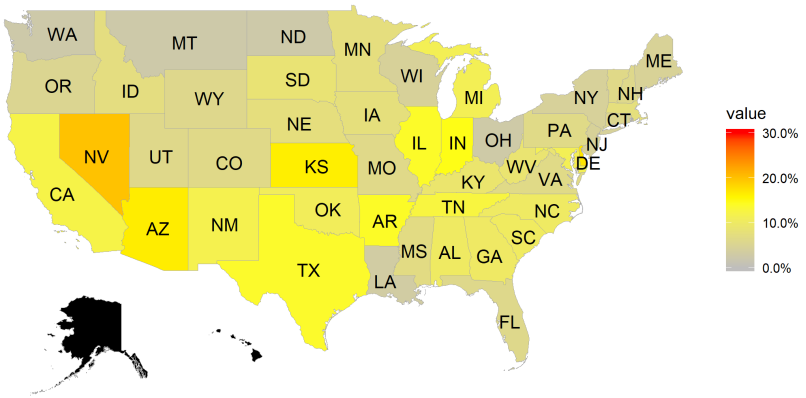
4.3.1 Murder

```
library(knitr)
include_graphics(c('../data/05_finalreport_plots/murder/choropleth_4.png',
                   '../data/05_finalreport_plots/murder/choropleth_30.png'))
```

Appearance of the word murder in US Case law, year=1970



Appearance of the word murder in US Case law, year=1996

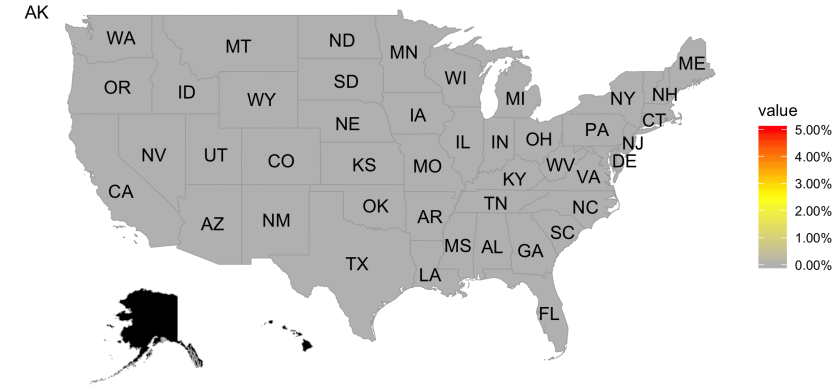


Murder seems to have increased between the years 1970 and 1996. The southern region of America seems to have higher law cases on murder.

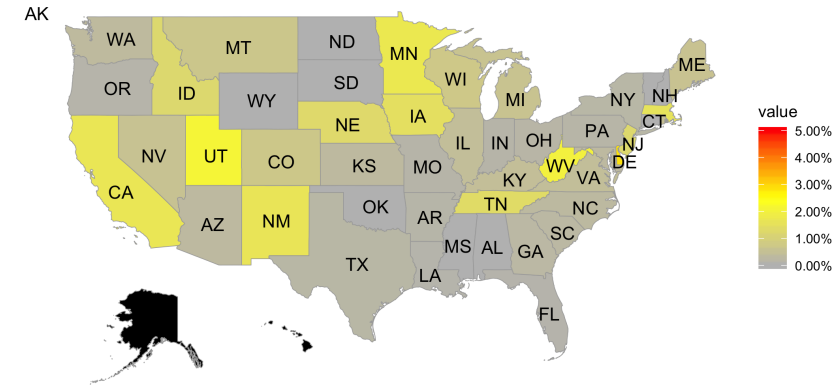
4.3.2 Sexual harassment

```
include_graphics(c('../data/05_finalreport_plots/sexual harassment/choropleth_4.png',
                   '../data/05_finalreport_plots/sexual harassment/choropleth_30.png'
))
```

Appearance of the word sexual harassment in US Case law, year=1970



Appearance of the word sexual harassment in US Case law, year=1996

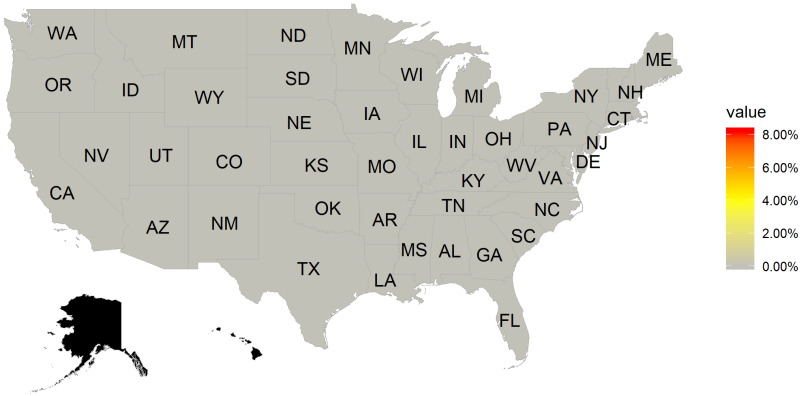


Before 1977, there was no discussion on the topic of “sexual harassment”, and then the attention on it started to increase. States like California, Utah, and Minesota seems to either exhibit high interest on the term, or suffered from it.

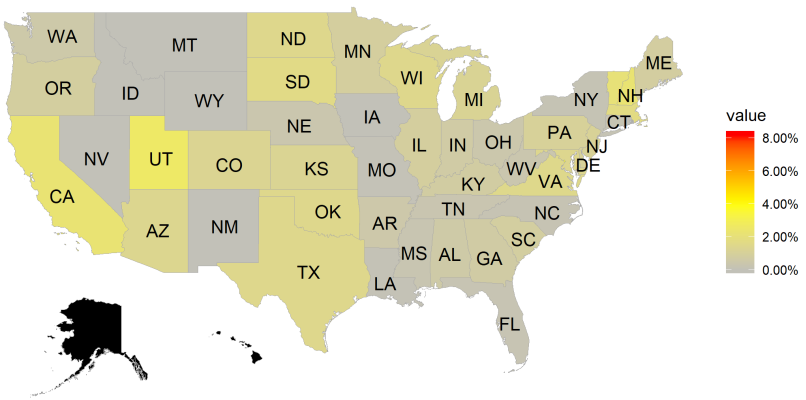
4.3.3 Internet

```
include_graphics(c('../data/05_finalreport_plots/internet/choropleth_4.png',
                   '../data/05_finalreport_plots/internet/choropleth_35.png'))
```

Appearance of the word internet in US Case law, year=1970



Appearance of the word internet in US Case law, year=2001

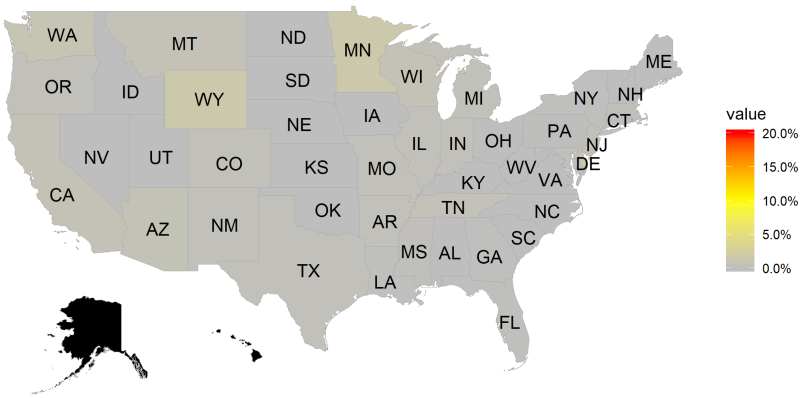


(For better comparison, internet was compared between 1970 and 2001) Like sexual harassment, internet was not a term aroung 1970, but after the evolution of internet, the interest on the term rapidly increased for most states.

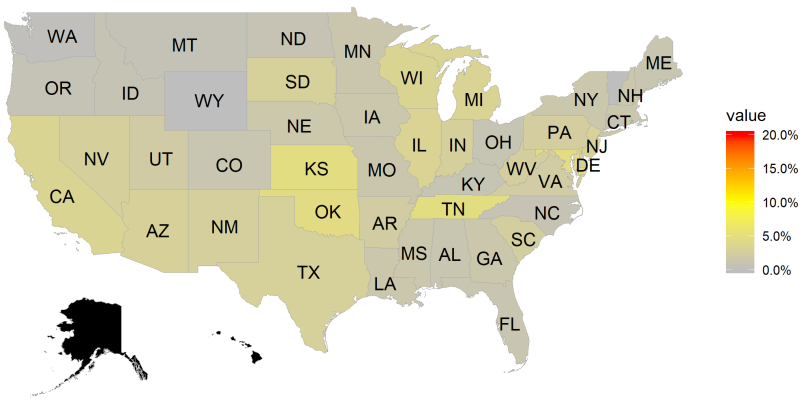
4.3.4 Medical malpractice

```
include_graphics(c('../data/05_finalreport_plots/medical malpractice/choropleth_4.pn
g',
                   '../data/05_finalreport_plots/medical malpractice/choropleth_30.pn
g'))
```

Appearance of the word medical malpractice in US Case law, year=1970



Appearance of the word medical malpractice in US Case law, year=1996

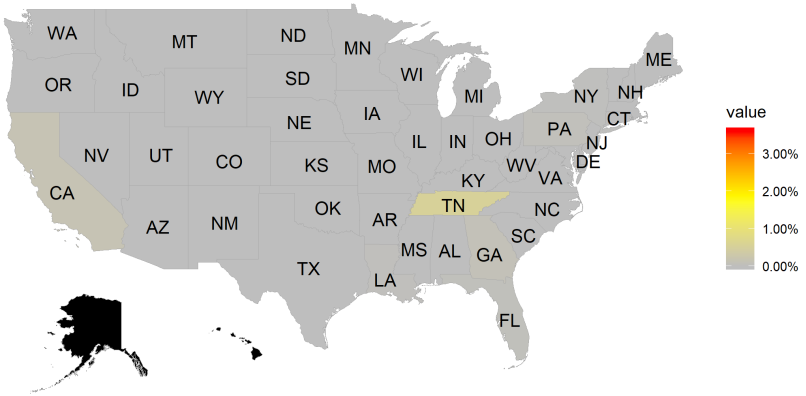


Medical malpractice cases increased from 1970 to 1996, as it mostly happened around the south-west coast (California, Arizona, Nevada, Texas).

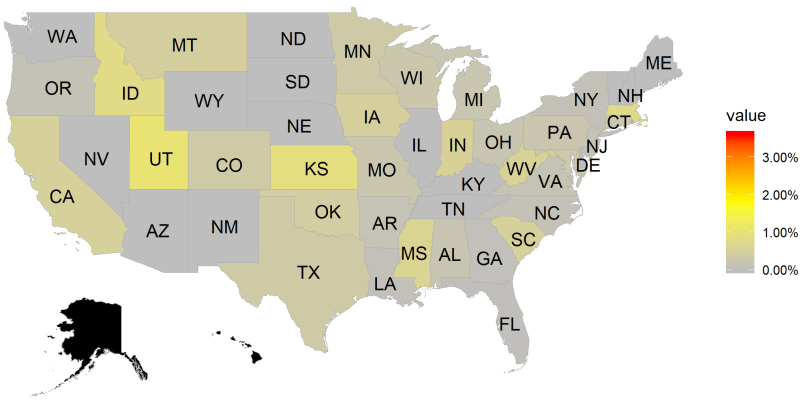
4.3.5 Insurance claim

```
include_graphics(c('../data/05_finalreport_plots/insurance claims/choropleth_4.png',

                   '../data/05_finalreport_plots/insurance claims/choropleth_30.png'
))
```

Appearance of the word insurance claims in US Case law, year=1970



Appearance of the word insurance claims in US Case law, year=1996

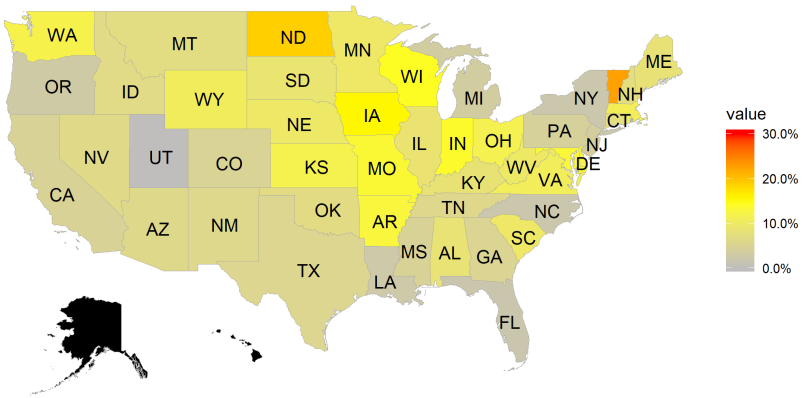


Insurance claims also increased, but it does not shows a clear spatial pattern.

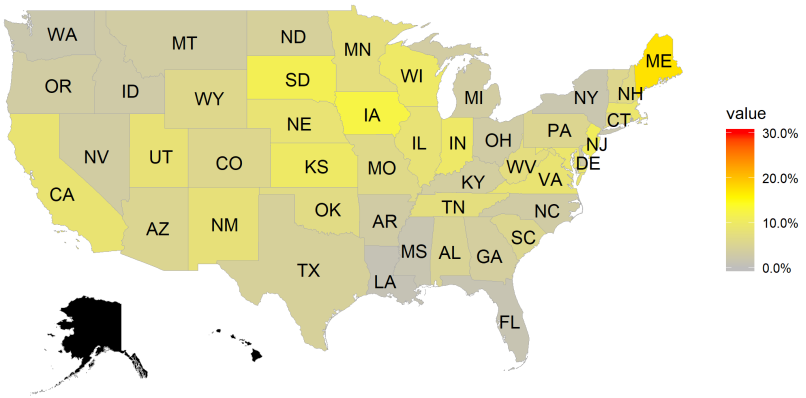
4.3.6 Real estate

```
include_graphics(c('../data/05_finalreport_plots/real estate/choropleth_4.png',
                   '../data/05_finalreport_plots/real estate/choropleth_30.png'))
```

Appearance of the word real estate in US Case law, year=1970



Appearance of the word real estate in US Case law, year=1996



Real Estate cases relatively decreased, from 1970 to 1996.

5. Executive summary

5.1 Data quality

Case.law datset is probably the most comprehensive, public available dataset in the domain of law with over 6 million cases since 1658.

The analysis shows, that most of the dataset is unstructured text with different level of quality. For example, “court” and “attorney” contain valid and useful values. These variables can be used for further analysis to receive meaningful insights. In contrast “judge” seems to be the only generic variables with no value.

Analysis should be considered carefully as the collection could be baised. A comparison of number of cases and population per state in 2010 showed, that there is difference for the ratio per state. A assumption is that the ratio should be more constant, as states with high population have more law cases.
The observation is supported by the cases per court: It seems that Appellate and Supreme courts are more frequently reported.

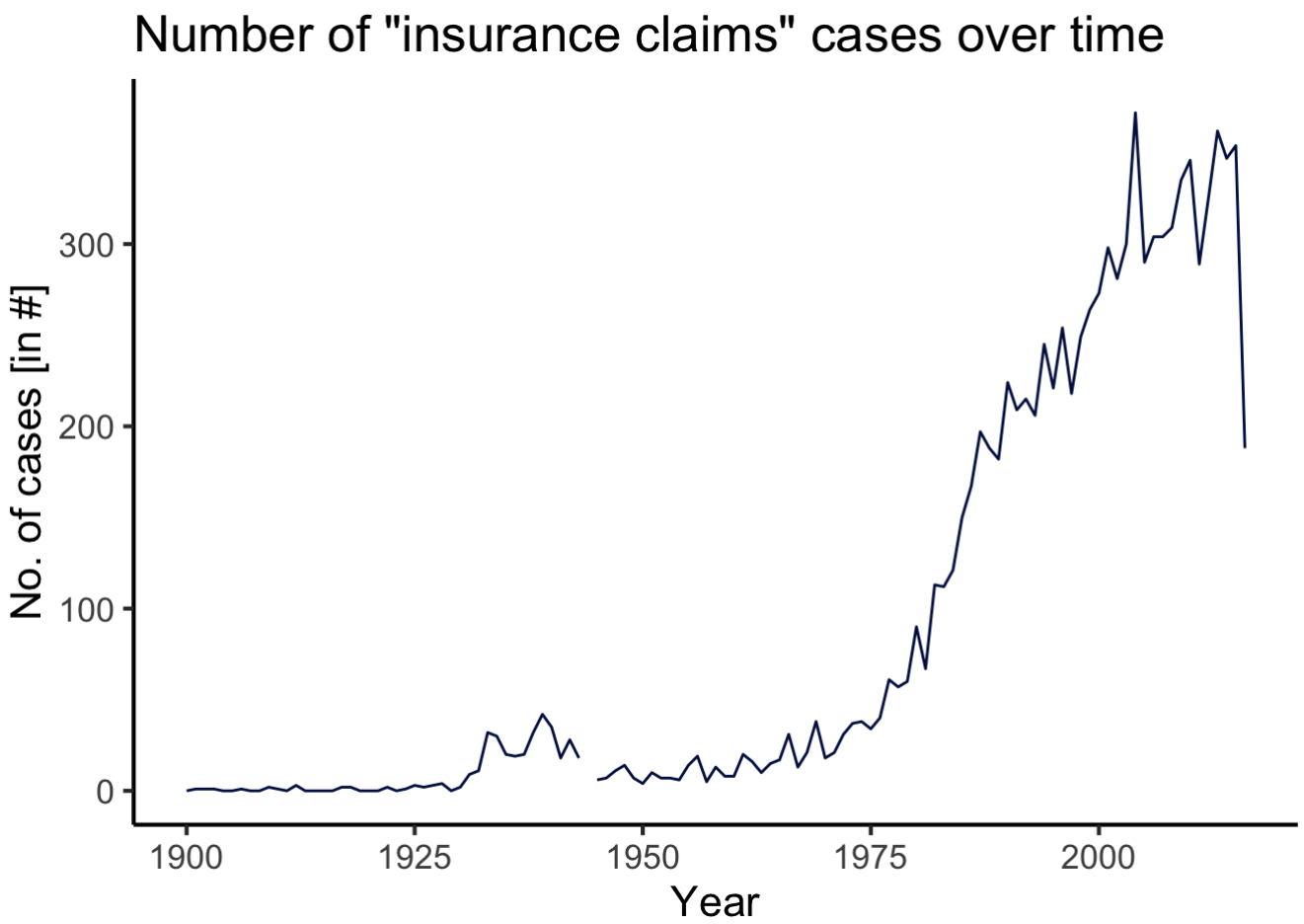
For analysis of trends, we assume that if there are some errors in the collection of data, then these error are a systematic.

5.2 Insights

5.2.1 Signficant trends of insurance claim cases, sexual harassment cases and medical malpractice over time

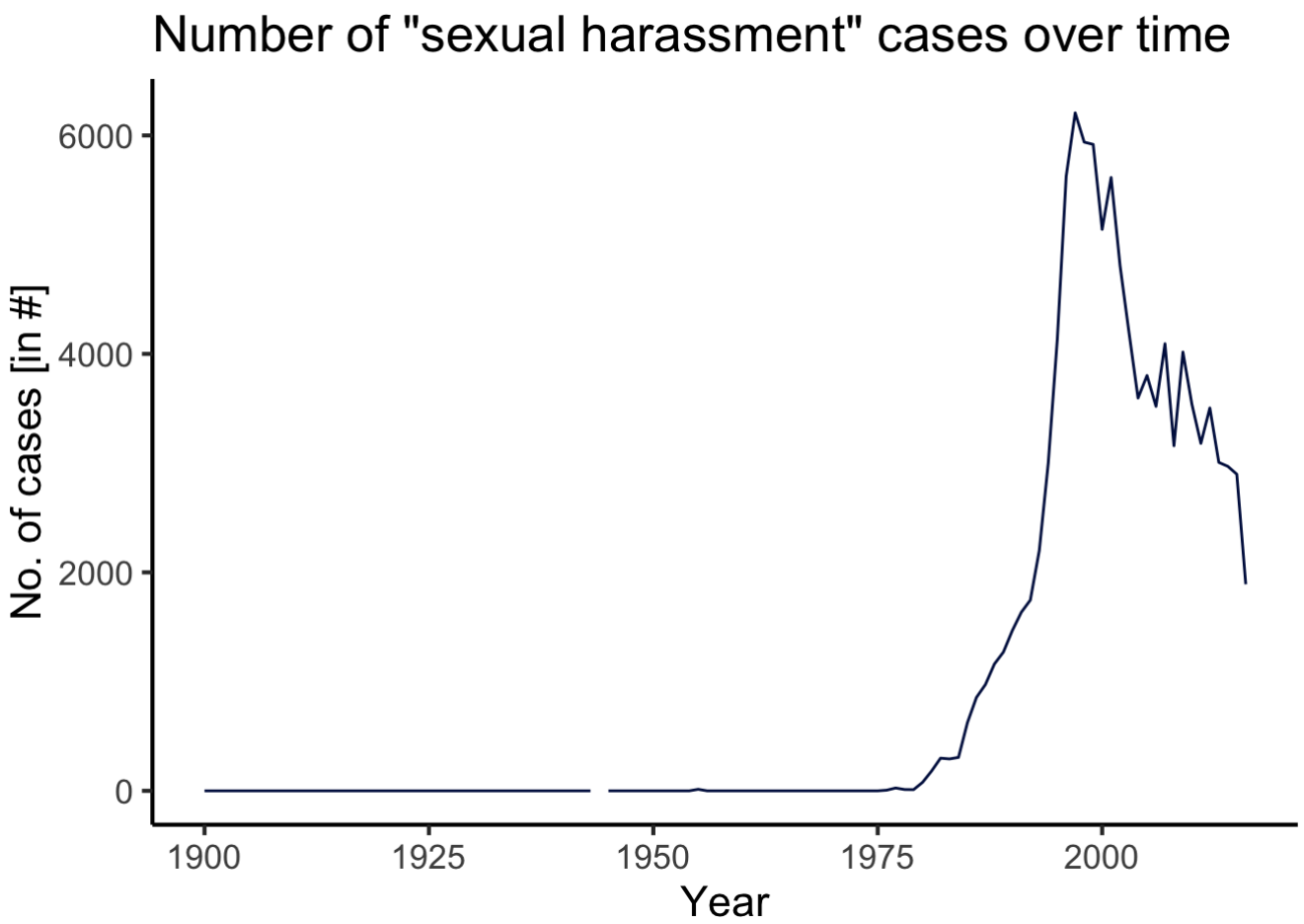
Some of the most interesting trends were found in the line plots that were created to see the progression of categories of cases over time, providing a simple yet comprehensive view of the trends in law cases. Some of the most interesting cases that we deemed significant were in categories of insurance claim cases, sexual harassment cases and medical malpractice cases.

```
ggplot(dfb_grouped_tidy %>% filter(Case.Type == 'cb_data_text_0_insurance.claims') %
>% filter(year >= 1900 & year <= 2016), aes(year, No))+
  geom_path(color="#01144d") +
  ggtitle('Number of "insurance claims" cases over time ') +
  xlab("Year") +
  ylab("No. of cases [in #]") +
  theme_grey(16) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



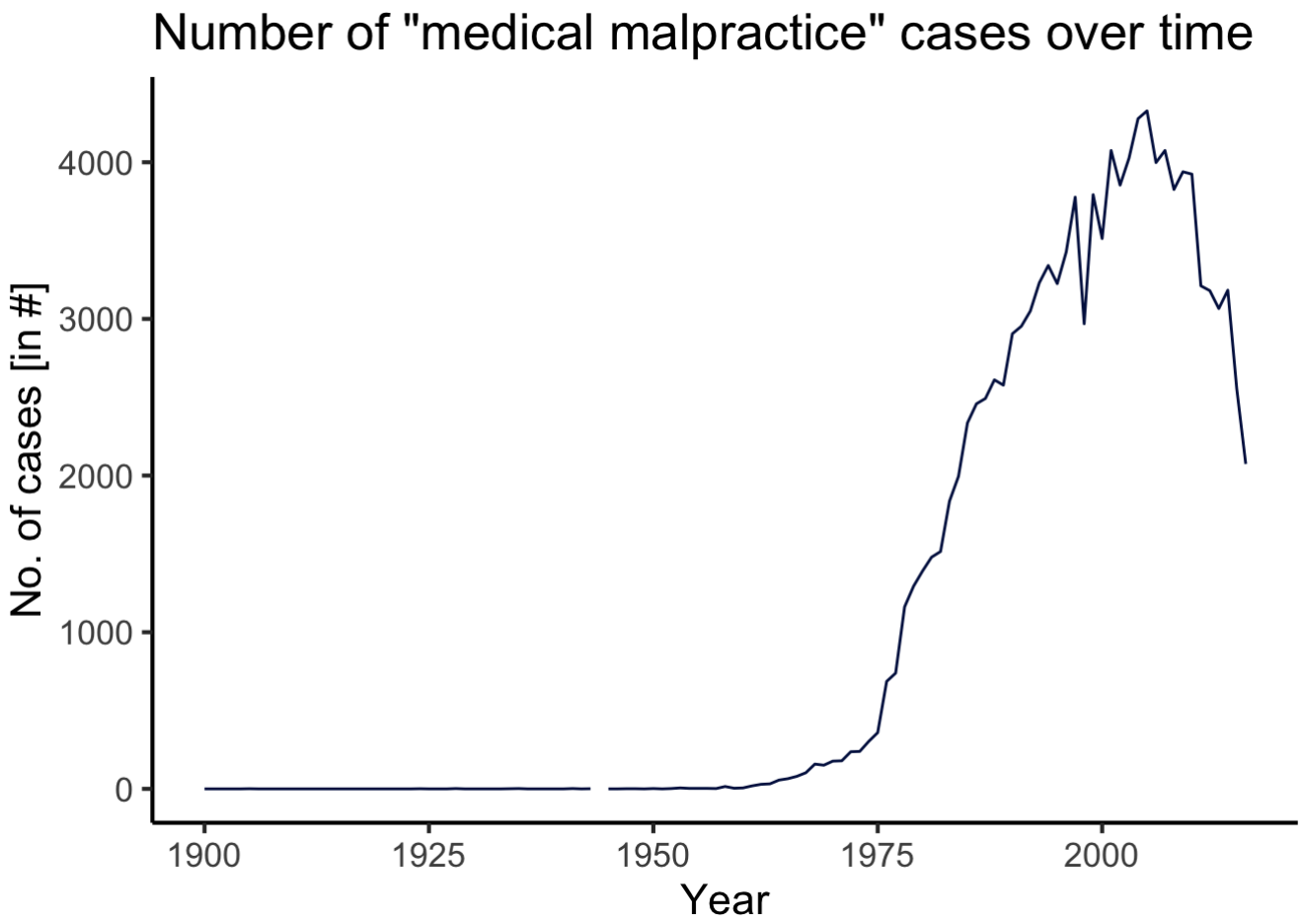
We can see that there are cases reported on insurance claims beginning from year 1900, as discussed before. However, there is a significant acceleration in number of cases brought to courts of the United States about insurance claims starting around 1975. This trend seems to continue, resulting in a consistent sharp increase in the number of cases from 1975 onwards. The number of insurance claims is one of the categories that seems to be increasing consistently over the course of the past 42 years.

```
ggplot(dfb_grouped_tidy %>% filter(Case.Type == 'cb_data_text_0_sexual.harassment') %
>% filter(year >= 1900 & year <= 2016), aes(year, No))+
  geom_path(color="#01144d") +
  ggtitle('Number of "sexual harassment" cases over time ') +
  xlab("Year") +
  ylab("No. of cases [in #]") +
  theme_grey(16) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



The trend in the number of cases on sexual harassment is notable as there are no cases before 1975. A year in which there are changes in the attitude of society towards sexual harassment, resulting from increased attention in media, with some famous cases brought to courts on sexual harassment and discrimination. In the category of sexual harassment, we can trace some of the increases in the number of cases with societal patterns such as the attention brought by the sexual harassment case of Anita Hill vs. Clarence Thomas in year 1991. The number of cases of sexual harassment reduce after around 2002. A more in depth look into the trends towards sexual harassment over the years will result into interesting patterns in the attitude of society and its attention to sexual harassment.

```
ggplot(dfb_grouped_tidy %>% filter(Case.Type == 'cb_data_text_0_medical.malpractice')
 %>% filter(year >= 1900 & year <= 2016), aes(year, No))+
  geom_path(color="#01144d") +
  ggtitle('Number of "medical malpractice" cases over time ') +
  xlab("Year") +
  ylab("No. of cases [in #]") +
  theme_grey(16) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



Another interesting category is the cases reported on medical malpractice. There are almost no cases before 1970 and in this year there is a sharp increase in the number of cases. An increase that continues to accelerate throughout the years. This might reflect the decrease of trust between individuals in the american society and the medical industry, hospitals and doctors. There needs to be a more in depth look for more precise insight into this odd trend.

5.2.2 Geographical differences for sexual harassment per state

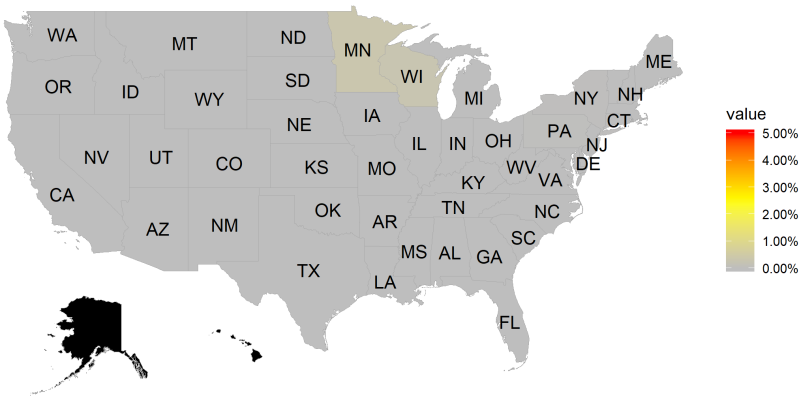
We plotted the spatial pattern of the proportion of sexual harassment cases for different years. In the spatial plots, the increase of sexual cases from 1980 to 2000 centered around the southern-west part of the country can be noted (California, Arizona). Then there is a decrease on the related cases nationwide around 2010, with the exceptions of states like California and Arizona. Such cases might either indicate the interest on the term and such cases, or the increase on actual related crimes.

```
### Part of the code from choro.R to produce the images of the next cell.
date_lim1 = "1967-01-01"
date_lim2 = "2017-01-01"

start_year = as.integer(strsplit(date_lim1, "-")[[1]][1])
end_year = as.integer(strsplit(date_lim2, "-")[[1]][1])-1

keyword = "sexual harassment"

data(df_state_demographics)

file_list <- list.files(path='../data/02_processed_csvs/', pattern="*text.csv")

totalList = data.frame(year=integer(), value1=integer(), region=character())

for (name in file_list){
  region <- (strsplit(name, "-")[[1]][1])
  print(region)
  data <- fread(paste0('../data/02_processed_csvs/', name), sep=",")
  data$cb_data_text_0 = ifelse(data[,c(47)] > 0, 1, 0)
  data <- data %>% group_by(decision_date) %>% summarise(total_cases = n(), value1 =
 sum(cb_data_text_0))

  data$decision_date <- as.Date(data$decision_date, format="%Y-%m-%d")
  data <- data %>% na.omit() %>% filter(decision_date >= as.Date(date_lim1) & decisio
n_date < as.Date(date_lim2))
  if(nrow(data)>0){
    data['year'] <- as.integer(format(data$decision_date, "%Y"))
    data <- data %>% group_by(year) %>% summarise(total_cases = sum(total_cases), val
ue1 = sum(value1))
    data['region'] <- tolower(region)
    totalList <- rbind(totalList, data)
  }
}
totalList$value <- totalList$value1/totalList$total_cases

choropleths = list()
index = 1
for (i in c(1980, 1990, 2000, 2010)){
  mapData <- totalList %>% filter(year == i)
  mapData <- mapData[,c('region', 'value')]

  for (state in setdiff(df_state_demographics$region, mapData$region)) {
    df<-data.frame(state,0)
    names(df)<-c("region","value")
    mapData <- rbind(mapData, df)
  }
  mapData = mapData[!mapData$region %in% c("alaska", "hawaii"), ]

  c = StateChoropleth$new(mapData)
  c$title = paste0("Appearance of the word ", keyword, " in US Case law, year=", i)
  c$set_num_colors(1)
  c$show_labels=FALSE
  c$clip()

  c$bind()
```

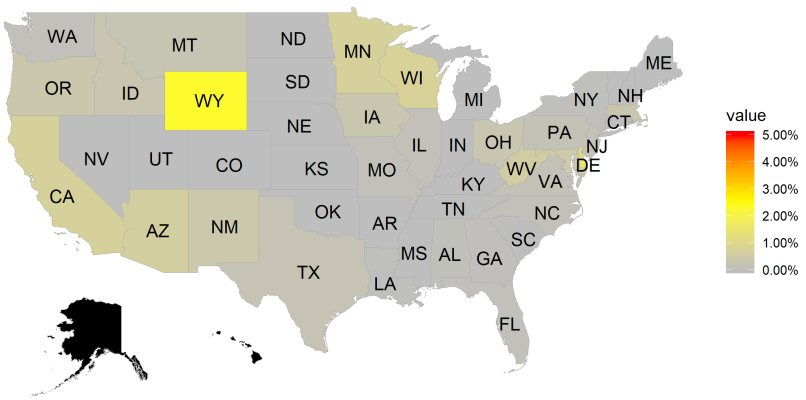
```
  choropleth <- c$render()

  df_state_labels = data.frame(long = state.center$x, lat = state.center$y, name=tolo
wer(state.name), label = state.abb)
  df_state_labels = df_state_labels[!df_state_labels$name %in% c("alaska", "hawaii"),
 ]

  choropleth = choropleth + geom_text(data = df_state_labels, aes(long, lat, label =
 label, group = NULL), color = 'black', check_overlap = TRUE)
  choropleth = choropleth + scale_fill_gradient2(low = "gray", mid = "yellow", high =
 "red", midpoint=0.025, labels = percent, limits=c(0, 0.05))
  choropleth = choropleth + theme(text = element_text(size=10))

  choropleths[[index]] = choropleth

  index = index + 1
}

grid.arrange(choropleths[[1]], choropleths[[2]], choropleths[[3]], choropleths[[4]],
 nrow=4)
```

```
include_graphics(c('../data/05_finalreport_plots/sexual harassment/choropleth_14.png'
,
                   '../data/05_finalreport_plots/sexual harassment/choropleth_24.png'
,
                   '../data/05_finalreport_plots/sexual harassment/choropleth_34.png'
,
                   '../data/05_finalreport_plots/sexual harassment/choropleth_44.png'
))
```

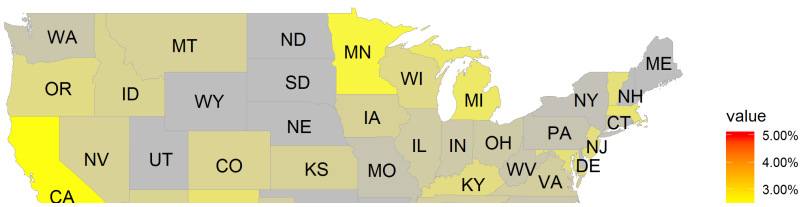
Appearance of the word sexual harassment in US Case law, year=1980

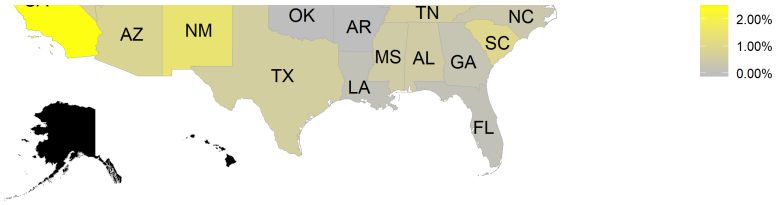


Appearance of the word sexual harassment in US Case law, year=1990

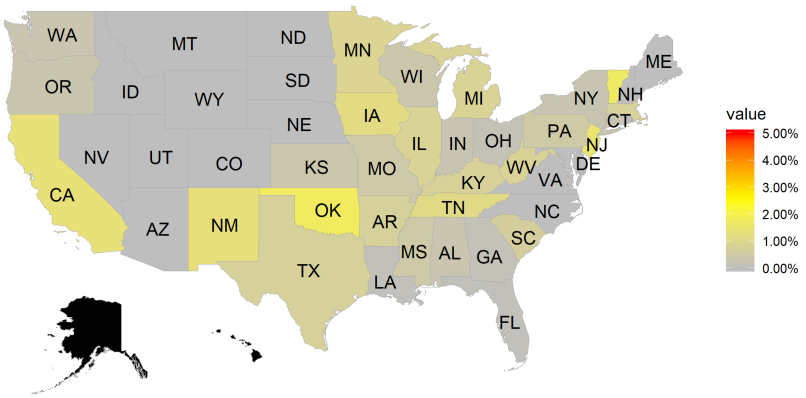


Appearance of the word sexual harassment in US Case law, year=2000





Appearance of the word sexual harassment in US Case law, year=2010



Note: The label “AK” should be removed and the legend should have the correct title. However, choropleth is limited in its configuation.

6. Interactive visualization

6.1 Spatial plots for websites

```
date_lim1 = "1967-01-01"
date_lim2 = "2017-01-01"

start_year = as.integer(strsplit(date_lim1, "-")[[1]][1])
end_year = as.integer(strsplit(date_lim2, "-")[[1]][1])-1

keyword = "sexual harassment"
col_num = 47

data(df_state_demographics)

file_list <- list.files(path='../data/02_processed_csvs/', pattern="*text.csv")

totalList = data.frame(year=integer(), value1=integer(), value2=integer(), value3=int
eger(), region=character())

for (name in file_list){
  region <- (strsplit(name, "-")[[1]][1])
  data <- fread(paste0('../data/02_processed_csvs/', name), sep=",")
  data$cb_data_text_0 = ifelse(data[,c(47)] > 0, 1, 0)
  data <- data %>% group_by(decision_date) %>% summarise(total_cases = n(), value1 =
 sum(cb_data_text_0))

  data$decision_date <- as.Date(data$decision_date, format="%Y-%m-%d")
  data <- data %>% na.omit() %>% filter(decision_date >= as.Date(date_lim1) & decisio
n_date < as.Date(date_lim2))
  if(nrow(data)>0){
    data['year'] <- as.integer(format(data$decision_date, "%Y"))
    data <- data %>% group_by(year) %>% summarise(total_cases = sum(total_cases), val
ue1 = sum(value1))
    data['region'] <- tolower(region)
    totalList <- rbind(totalList, data)
  }
}
totalList$value <- totalList$value1/totalList$total_cases
dir.create(file.path("../data/05_finalreport_plots/", keyword), showWarnings = FALSE)
write.csv(totalList, file = paste0("../data/05_finalreport_plots/", keyword, "/", key
word, "_", toString(start_year), "_", toString(end_year), ".csv"))

choropleths = list()
index = 1
for (i in seq(from=start_year, to=end_year)){
  mapData <- totalList %>% filter(year == i)
  mapData <- mapData[,c('region', 'value')]

  for (state in setdiff(df_state_demographics$region, mapData$region)) {
    df<-data.frame(state,0)
    names(df)<-c("region","value")
    mapData <- rbind(mapData, df)
  }
  mapData = mapData[!mapData$region %in% c("alaska", "hawaii"), ]

  c = StateChoropleth$new(mapData)
  c$title = paste0("Appearance of the word ", keyword, " in US Case law, year=", i)
  c$set_num_colors(1)
  c$show_labels=FALSE
  c$clip()
```

```
  c$bind()

  choropleth <- c$render()

  df_state_labels = data.frame(long = state.center$x, lat = state.center$y, name=tolo
wer(state.name), label = state.abb)
  df_state_labels = df_state_labels[!df_state_labels$name %in% c("hawaii"), ]

  choropleth = choropleth + geom_text(data = df_state_labels, aes(long, lat, label =
 label, group = NULL), color = 'black', check_overlap = TRUE)
  choropleth = choropleth + scale_fill_gradient2(low = "gray", mid = "yellow", high =
 "red", midpoint=0.025, labels = percent, limits=c(0, 0.05))
  choropleth = choropleth + theme(text = element_text(size=10))

  choropleths[[index]] = choropleth

  index = index + 1
}

setwd(paste0("../data/05_finalreport_plots/", keyword))
choroplethr_animate(choropleths)
setwd(WD)
```

One method this group used to develop interactive visualization was to create an html-based spatial interactive graph, which can be integrated in any website, such as newspaper article or blog post. The widget models the frequency of different court cases for 48 US states for different years. This interactive app was developed based on the choroplethr package, and was published on https://yj7082126.github.io/patrick_kwon/ (https://yj7082126.github.io/patrick_kwon/) (The link might not be correctly loaded due to unauthenticated script issue. Depending on the browser type, one should disable the “block unsafe script” option. For chrome users, click the “Insecure Content Blocked” on the url area).

The objective of this interactive visualization is to engage the readers of the article by showing the spatial distribution of the frequency of different case laws on US map data. We decided to utilize choroplethr package’s animation component for visualization. Because the package provided limited freedom on editing the visualization, we overrided some of the package’s functions for customized visualization (ex: changing the color of the plot, changing the font size of letters).

The interactive plot can be used by adjusting the control scheme on the top right section of the web page. By dragging the slider the viewer can adjust the views from year 1967 and 2016. When the play button is clicked, it automatically generates a animation from the current map to the last map (at year 2016), and can be cancelled by clicking the stop button.

Among the choroplethr maps generated, sexual harassment cases showed the clearest spatial pattern. While other animated choropletehr maps showed a increasing/decreasing pattern by year, sexual harassment cases showed a clear regional pattern with a steep increase around the 1975s. So we chose the sexual harassment animated choropleth for the presentation of this group project. The maximum value of sexual harassment ratio compared to the total number of cases was around 5%. Alaska and Hawaii, as adressed above, were ignored.

As adressed above, one key disadvantage for the choroplethr maps was that because it followed geographical regions, it was hard to effectively show the significance of smaller states, such as District of Columbia, which contains the most number of courts. If we had more time, a animated cartograph might also be used for display.

6.2 Interactive tool for fast data exploration

In addition, we developed another tool for fast data exploration and interaction with the dataset. Our RShiny tool recreates our visualizations and provides deep dive in the dataset. As the dataset was recently published and no other anlysis has been published for it, it is a best opportunity to get familiar with the data.

As the dataset is only available after registration, we hosted RShiny on our own AWS instance with RShiny server: Live Demo (http://ec2-54-152-240-211.compute-1.amazonaws.com:3838/03_r_interactive/). Therefore, only aggreated data is available to the end user.

The tool provides three views:
1. Time analysis - the number of cases over time
2. Geo analysis - spatial plots over time
3. Case deep dive - real case example for keywords

The first two views are animated and by pressing the “play” button, the plots draw overtime. User should be more engaged by animating the plots.
On the leftside navigation, there are different filter for keywords, states and year (depending on the selected view).

The source code can be found on our github repository - GitHub (https://github.com/bschifferer/exploratory_law.case/tree/master/03_r_interactive) - and can be easily hosted and extended.

By using rshiny server and plotly, our tool has some limitation:

- The free version of rshiny servrer supports only 1 concurrent user, but that should be enough for providing a demo.
- Spatial plots of plotly are limited in text and hovertext. The plot should display the abbreviations of a state on the plot. The hovertext could be better formatted (incl. the numbers)
- The deep dive view should provide forward/next buttons
- For some filter combination, rshiny has some bugs, which could not be 100% resolved

7. Conclusion

This report should give readers a better understanding on the history of different law cases and its spatial patterns.

For the project, the USA Caselaw data was analyzed, focusing specifically on different types of cases such as sexual harassment and medical malpractice. We focused on the appearance of certain case-specific words on the casebody of each cases to identify case types. The linear charts for each different cases mostly displayed a sharp increase in case numbers around 1975, especially for cases like insurance claim, medical malpractice and sexual harassment. The spatial choropleth graphs displayed increase on the ratio of such cases compared to the total number of cases, and also revealed a slightly higher trend of increase on the southern west states. Those analysis shows that many new types of law cases, like sexual harassment and intellectual property infringement, made its entrance around 1975~2000 and were widely noticed around the southern west states.

The sheer size of the data set and its lack of clear structure became significant problems for our group’s analysis. To counter these problems, our group preprocessed data on cloud systems and made basic text mining to get our variables. Our group also modified some basic functions of the choroplethr package to generate better view of spatial data. With extended research time, our group might be able to match additional data sources on political/economical issues to understand trends better, or deploy more sophisticated natural language models for a clear picture.