

# Appellate Data Exploratory Analysis

Michael Sullivan

2025-01-29

The below code consists of initial exploratory data analysis I am pursuing for a database of over 15,000 randomly sampled federal appellate court decisions between 1925 and 1996.

```
library(here)
library(ggplot2)
library(dplyr)
```

```
data_path <- here("cta96.csv")
data <- read.csv(data_path)

#Data exploration
duplicate_casenums <- data$casenum[duplicated(data$casenum)]
print(duplicate_casenums)
```

```
## integer(0)
```

```
years <- sort(unique(data$year))
num_years <- length(years)
year_counts <- rep(0,num_years)
for (year in years) {
  print(paste(year, sum(data$year == year), sep = ":"))
}
```

```
## [1] "1925:150"
## [1] "1926:150"
## [1] "1927:150"
## [1] "1928:150"
## [1] "1929:150"
## [1] "1930:165"
## [1] "1931:165"
## [1] "1932:165"
## [1] "1933:165"
## [1] "1934:165"
## [1] "1935:165"
## [1] "1936:165"
## [1] "1937:165"
## [1] "1938:165"
## [1] "1939:165"
## [1] "1940:165"
## [1] "1941:165"
```

## [1] "1942:165"  
## [1] "1943:165"  
## [1] "1944:165"  
## [1] "1945:165"  
## [1] "1946:165"  
## [1] "1947:165"  
## [1] "1948:165"  
## [1] "1949:165"  
## [1] "1950:165"  
## [1] "1951:165"  
## [1] "1952:165"  
## [1] "1953:165"  
## [1] "1954:165"  
## [1] "1955:165"  
## [1] "1956:165"  
## [1] "1957:165"  
## [1] "1958:165"  
## [1] "1959:165"  
## [1] "1960:165"  
## [1] "1961:330"  
## [1] "1962:330"  
## [1] "1963:330"  
## [1] "1964:330"  
## [1] "1965:330"  
## [1] "1966:330"  
## [1] "1967:330"  
## [1] "1968:331"  
## [1] "1969:329"  
## [1] "1970:330"  
## [1] "1971:330"  
## [1] "1972:330"  
## [1] "1973:330"  
## [1] "1974:330"  
## [1] "1975:330"  
## [1] "1976:330"  
## [1] "1977:330"  
## [1] "1978:330"  
## [1] "1979:330"  
## [1] "1980:330"  
## [1] "1981:330"  
## [1] "1982:360"  
## [1] "1983:360"  
## [1] "1984:360"  
## [1] "1985:360"  
## [1] "1986:360"  
## [1] "1987:360"  
## [1] "1988:360"  
## [1] "1989:360"  
## [1] "1990:360"  
## [1] "1991:360"  
## [1] "1992:360"  
## [1] "1993:360"  
## [1] "1994:360"  
## [1] "1995:360"

```
## [1] "1996:360"
```

```
months <- c(1:12)
for (month in months) {
  print(paste(month, sum(data$month == month), sep = ":"))
}
```

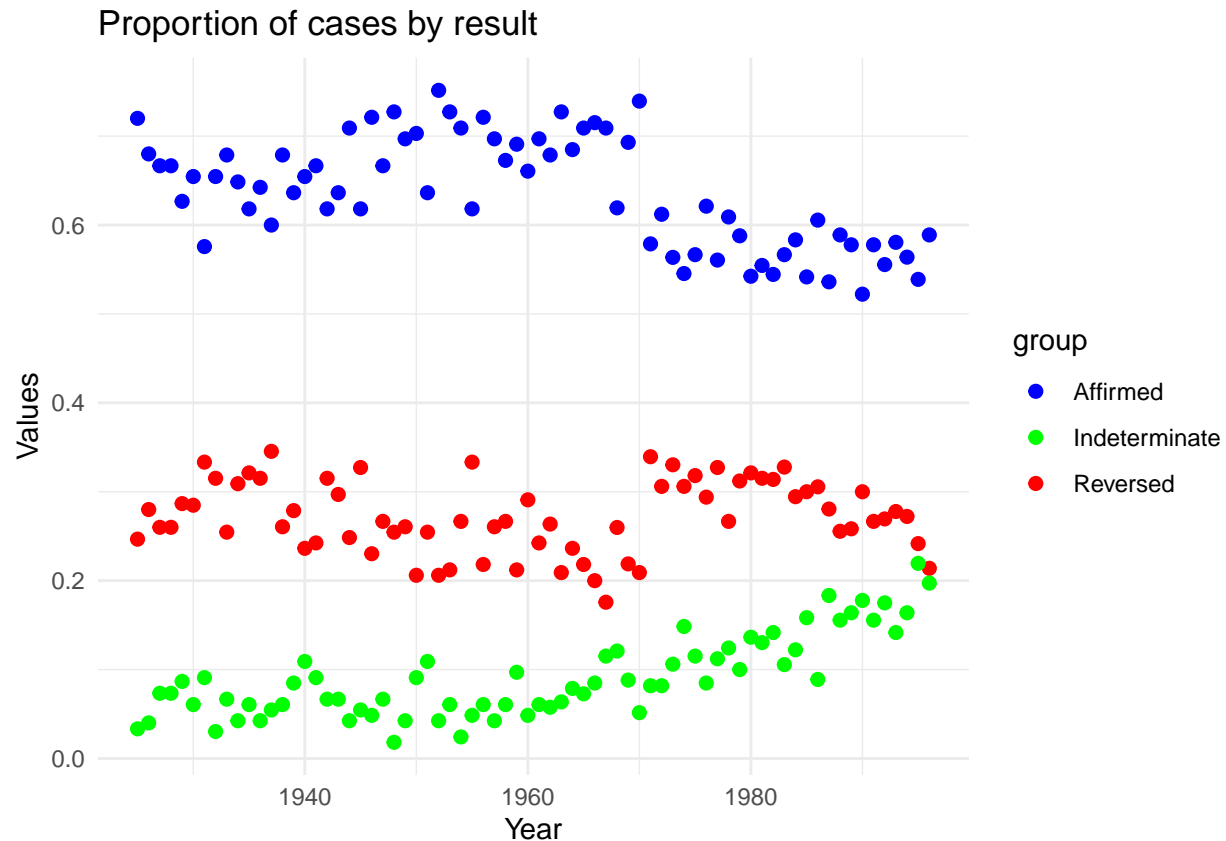
```
## [1] "1:1635"
## [1] "2:1507"
## [1] "3:1624"
## [1] "4:1727"
## [1] "5:1623"
## [1] "6:1886"
## [1] "7:1398"
## [1] "8:1260"
## [1] "9:1055"
## [1] "10:1274"
## [1] "11:1512"
## [1] "12:1694"
```

```
#Coding the various outcomes for the decision of the appellate court into groups
affirmed <- c(1,8)
reversed <- c(2,3,4,7)
indeterminate <- c(0,5,6,9,10)
```

```
#Iterating over every year to determine what proportion of decisions in each year affirmed, reversed or
year_percent_affirmed <- rep(0,num_years)
year_percent_reversed <- rep(0,num_years)
year_percent_indeterminate <- rep(0,num_years)
for (i in seq_along(years)) {
  year_percent_affirmed[i] <- nrow(data[data$year == years[i] & data$treat %in% affirmed, ]) / sum(data$year == years[i])
  year_percent_reversed[i] <- nrow(data[data$year == years[i] & data$treat %in% reversed, ]) / sum(data$year == years[i])
  year_percent_indeterminate[i] <- nrow(data[data$year == years[i] & data$treat %in% indeterminate, ]) / sum(data$year == years[i])
}
```

```
grouped_treat_data_long <- data.frame(
  years = rep(years, 3),
  values = c(year_percent_affirmed, year_percent_reversed, year_percent_indeterminate),
  group = rep(c("Affirmed", "Reversed", "Indeterminate"), each = length(years))
)
```

```
ggplot(grouped_treat_data_long, aes(x = years, y = values, color = group)) +
  geom_point(size = 2) +
  ggtitle("Proportion of cases by result") +
  xlab("Year") +
  ylab("Values") +
  scale_color_manual(values = c("blue", "green", "red")) +
  theme_minimal()
```



With the above plot I observed that there is a sharp change in the Affirm and Reverse data right around 1970. It is not immediately clear to me what is causing this change. It does not coincide with an obvious change in the courts. New circuits are added to the dataset in 1930 and 1982 but not around 1970. It also does not correspond to a change in procedure for the data set. The number of observations drawn from each circuit per year is doubled in 1961, but this change is well before the disjunction seen above. I continue below by breaking down the groups of results I have made (Affirm, Reverse, Indeterminate) into the exact codings in the data set to understand if this pattern is caused by a particular type of decision.

```
#Breaking decision types down into each logged category instead of the groups I created above
individual_treat_data <- data.frame(
  years = years,
  matrix(ncol = 11, nrow = length(years))
)
colnames(individual_treat_data)[2:12] <- 0:10

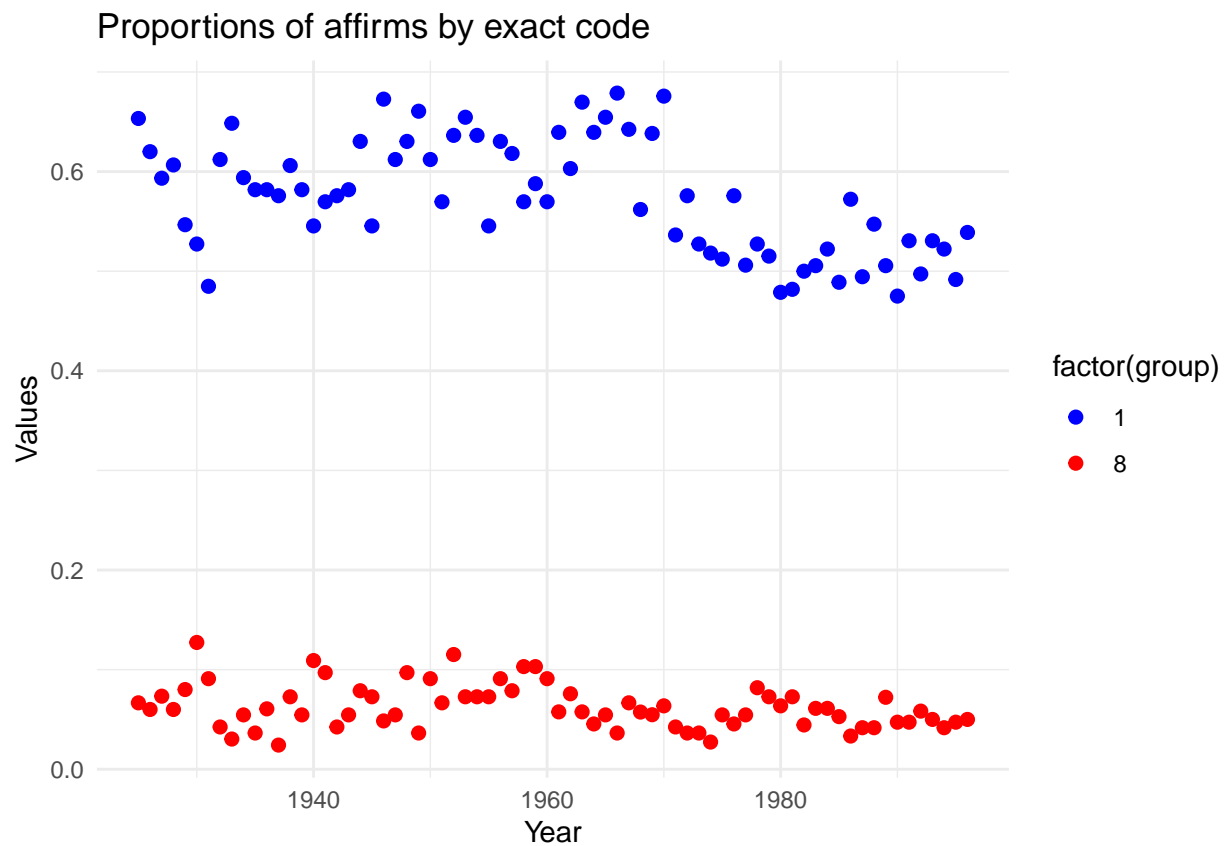
for (i in 0:10) {
  for (j in seq_along(individual_treat_data$years)) {
    year <- individual_treat_data$years[j]
    individual_treat_data[j,as.character(i)] <- nrow(data[data$year == year & data$treat == i, ]) / sum
  }
}

individual_treat_data_long <- data.frame(
  years = rep(years, 11),
  values = unlist(individual_treat_data[,2:12]),
  group = rep(c(0:10), each = length(years))
)
```

```
)

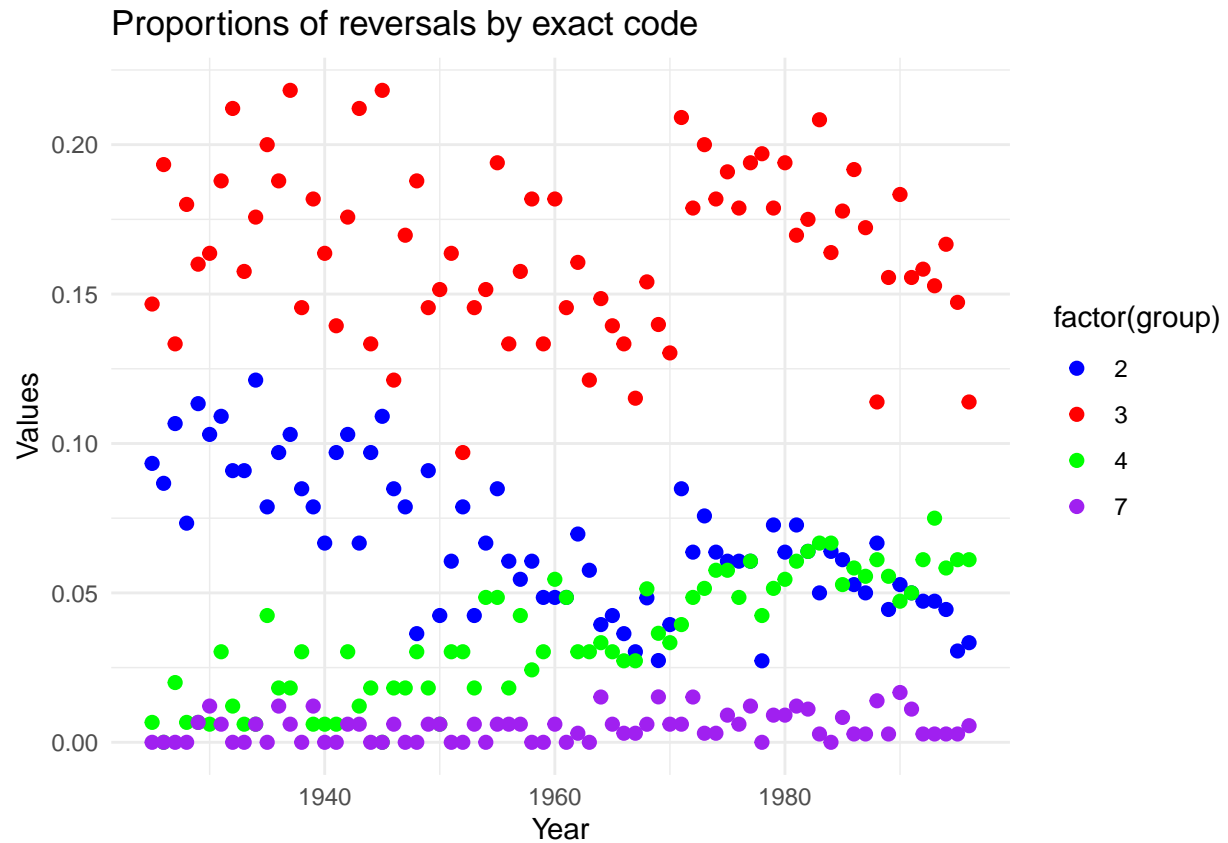
filtered_data <- individual_treat_data_long %>%
  filter(group == 1 | group == 8)

ggplot(filtered_data, aes(x = years, y = values, color = factor(group))) +
  geom_point(size = 2) +
  ggtitle("Proportions of affirms by exact code") +
  xlab("Year") +
  ylab("Values") +
  scale_color_manual(values = c("blue", "red")) +
  theme_minimal()
```



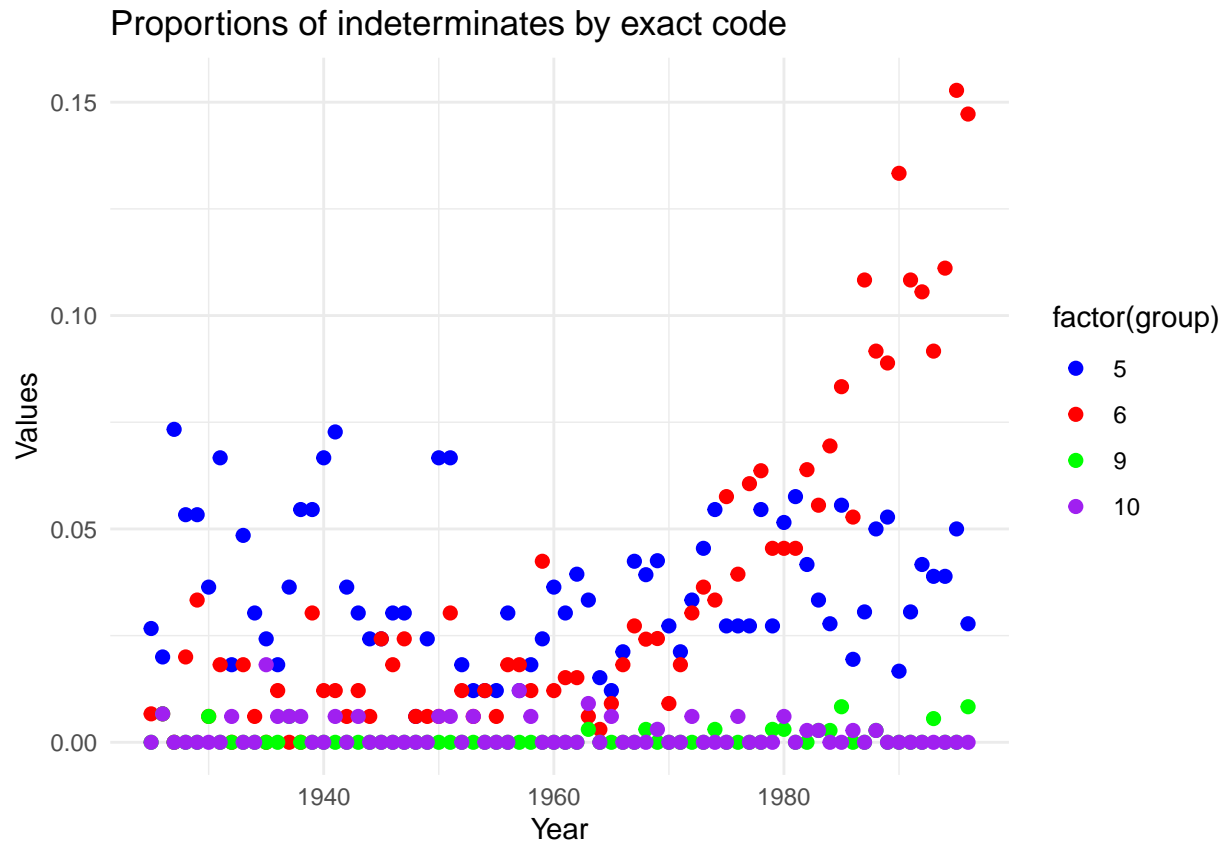
```
filtered_data <- individual_treat_data_long %>%
  filter(group == 2 | group == 3 | group == 4 | group == 7)

ggplot(filtered_data, aes(x = years, y = values, color = factor(group))) +
  geom_point(size = 2) +
  ggtitle("Proportions of reversals by exact code") +
  xlab("Year") +
  ylab("Values") +
  scale_color_manual(values = c("blue", "red", "green", "purple")) +
  theme_minimal()
```



```
filtered_data <- individual_treat_data_long %>%
  filter(group == 5 | group == 6 | group == 9 | group == 10)

ggplot(filtered_data, aes(x = years, y = values, color = factor(group))) +
  geom_point(size = 2) +
  ggtitle("Proportions of indeterminates by exact code") +
  xlab("Year") +
  ylab("Values") +
  scale_color_manual(values = c("blue", "red", "green", "purple")) +
  theme_minimal()
```



The above plots indicate that the sharp breaks are actually occurring in relevant categories for affirm and reverse. We see the pattern seems to be primarily arising out of movement in codes 1, 2, and 3. Code '1' indicates "affirmed" while codes 2 and 3 indicate "reversed" and "reversed and remanded" respectively. These are some of the most straightforward possible outcomes for these cases. Given that this sharp break does not seem to be an artifact of how I coded the decision groups, I continue below by exploring if the break is a pattern driven by particular circuit courts and is not generally occurring across the country.

```
#Analyzing the grouped results by year and appellate circuit
grouped_circuit_treat_data_long <- data.frame(
  year = rep(years, each = 3*12),
  circuit = rep(c(0:11), length(years)*3),
  group = rep(c("Affirmed", "Reversed", "Indeterminate"), each = 12, times = length(years)),
  treat_percent = rep(0, 3*12*length(years))
)

data_list <- list(
  "Affirmed" = affirmed,
  "Reversed" = reversed,
  "Indeterminate" = indeterminate
)

# Function to retrieve the vector of codes dynamically
get_vector <- function(key) {
  if (key %in% names(data_list)) {
    return(data_list[[key]])
  } else {

```

```

    stop("Key not found in the list.")
  }
}

for (i in seq_len(nrow(grouped_circuit_treat_data_long))) {
  circuit_year_total = nrow(data[data$year == grouped_circuit_treat_data_long$year[i]
                                & data$circuit == grouped_circuit_treat_data_long$circuit[i], ])
  circuit_year_group = nrow(data[data$treat %in% get_vector(grouped_circuit_treat_data_long$group[i])
                                & data$year == grouped_circuit_treat_data_long$year[i]
                                & data$circuit == grouped_circuit_treat_data_long$circuit[i], ])
  grouped_circuit_treat_data_long$treat_percent[i] <- circuit_year_group / circuit_year_total
}

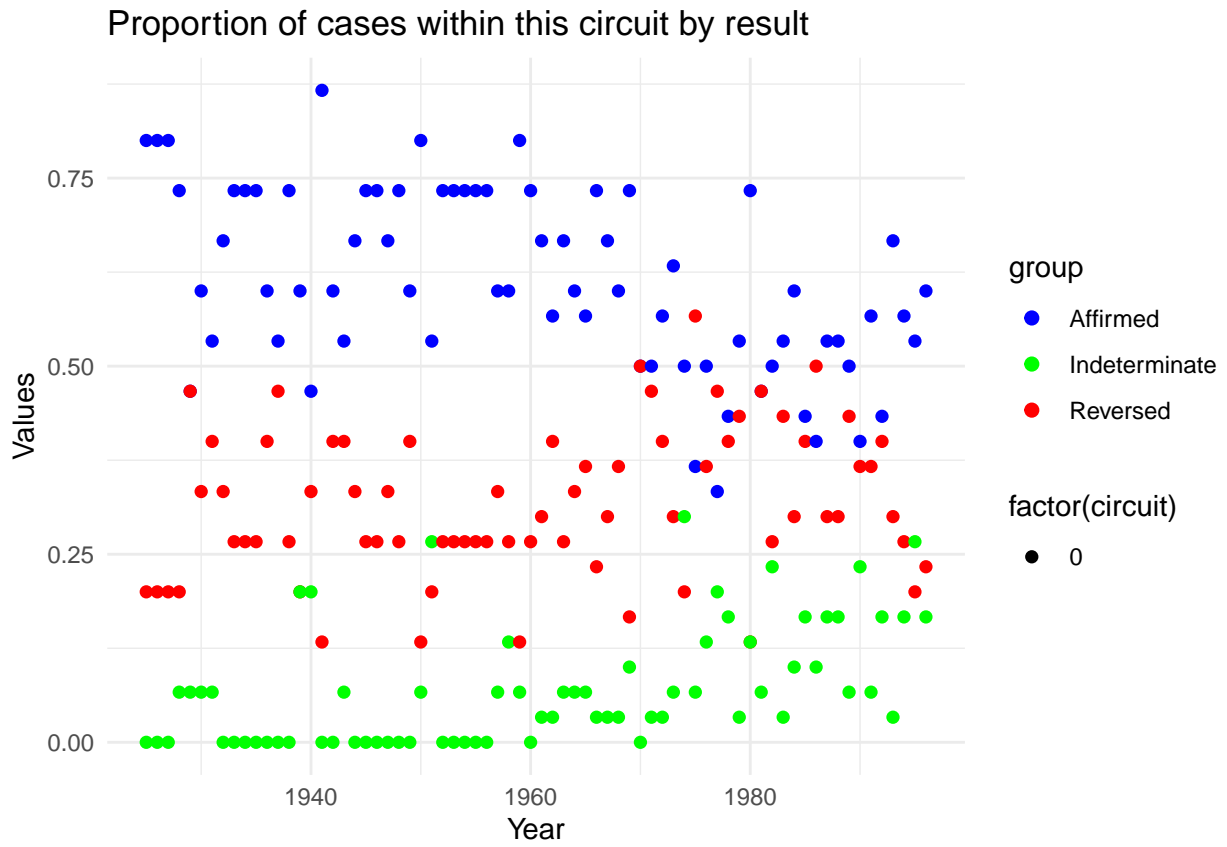
for (i in 0:11) {
  filtered_data <- grouped_circuit_treat_data_long %>%
    filter(circuit == i)

  p <- ggplot(filtered_data, aes(x = year, y = treat_percent, color = group, shape = factor(circuit))) +
    geom_point(size = 2) +
    ggtitle("Proportion of cases within this circuit by result") +
    xlab("Year") +
    ylab("Values") +
    scale_color_manual(values = c("blue", "green", "red")) +
    theme_minimal() +
    guides(color = guide_legend(order = 1),
           shape = guide_legend(order = 2))

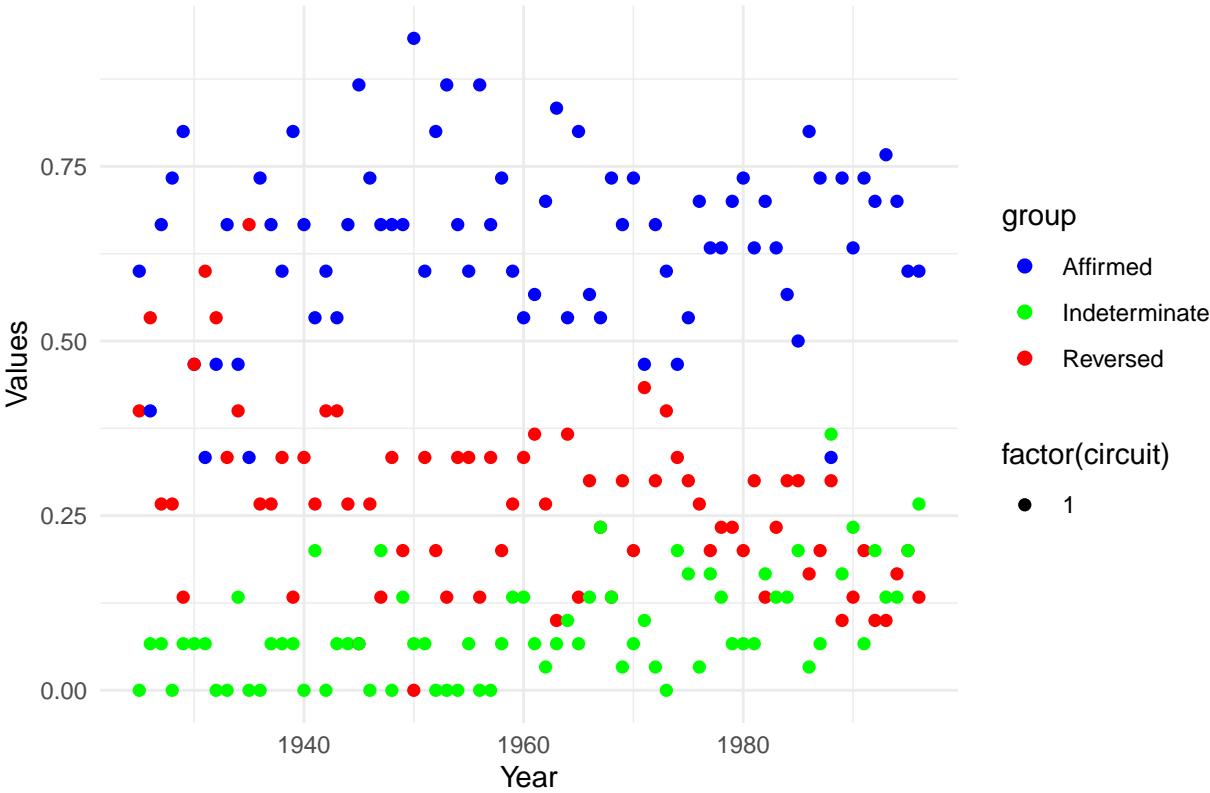
  print(p)
}

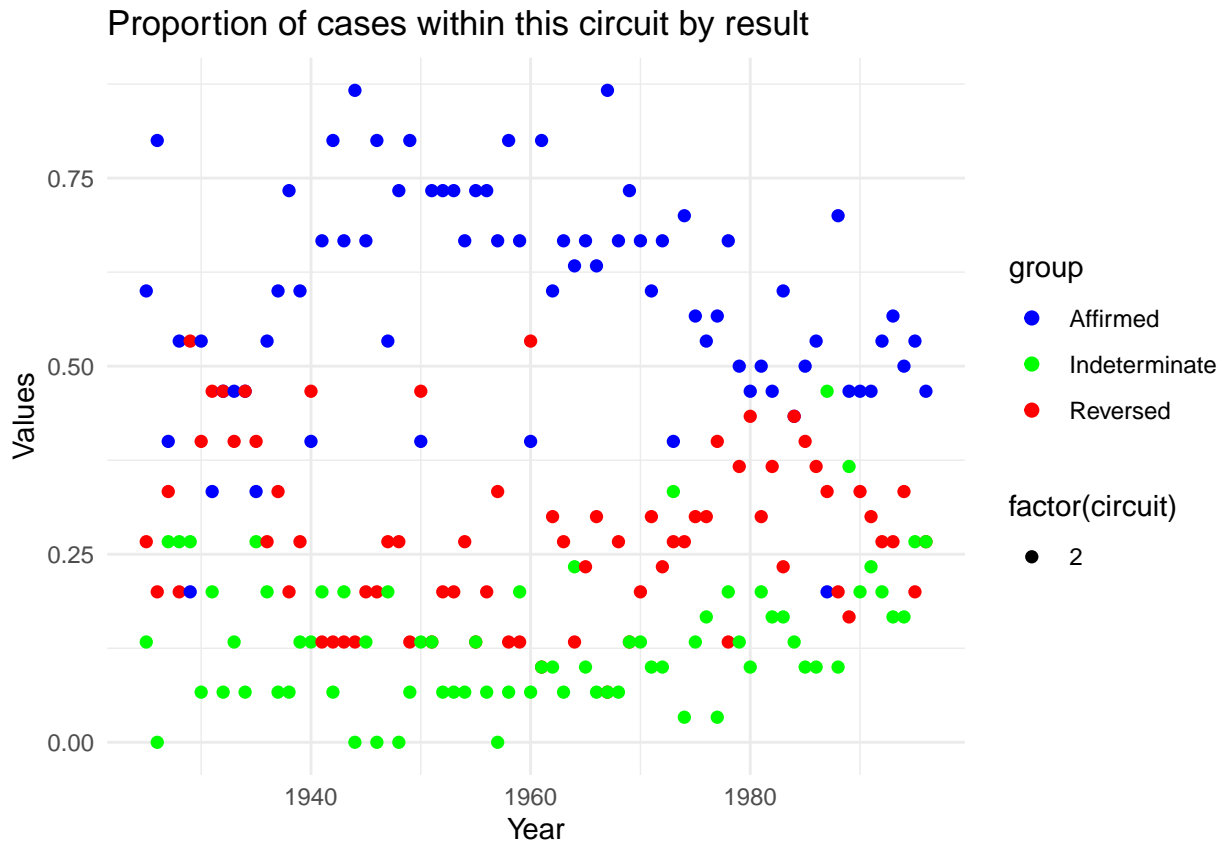
```

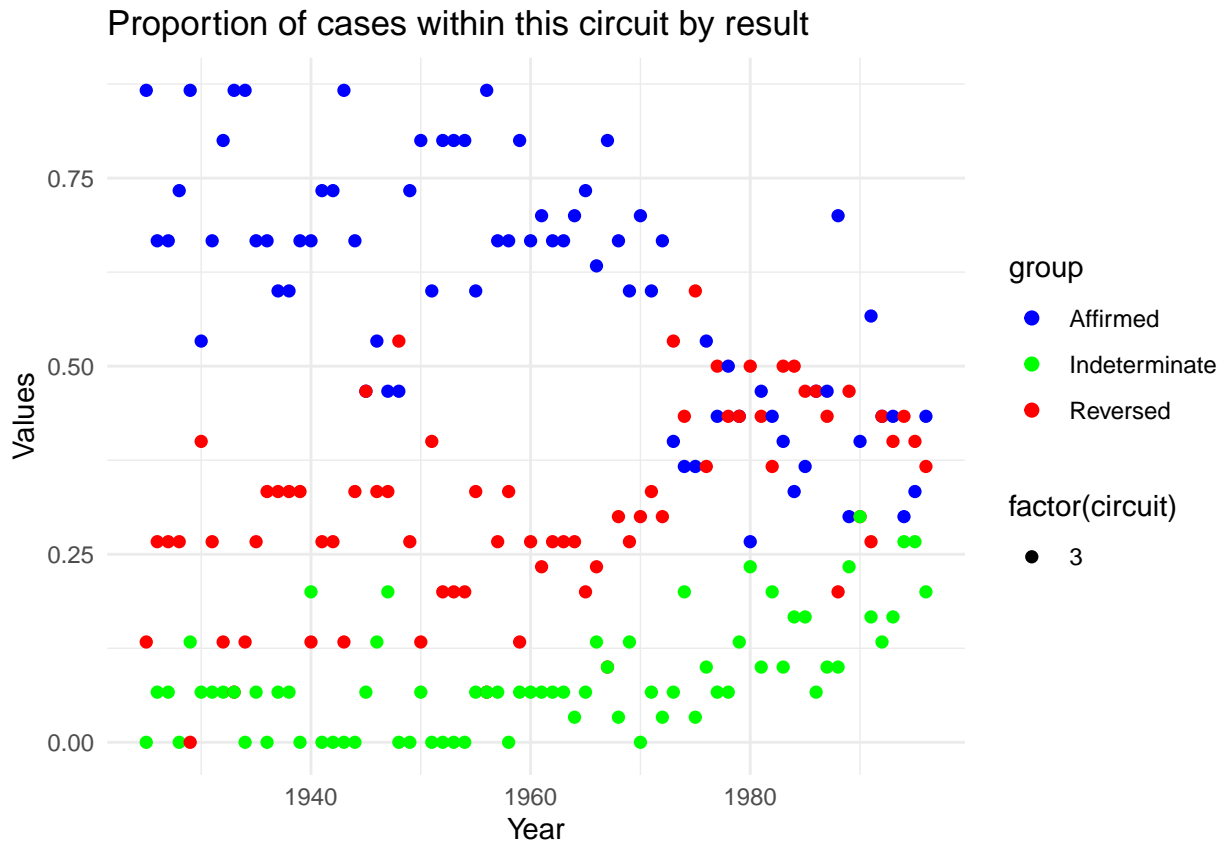


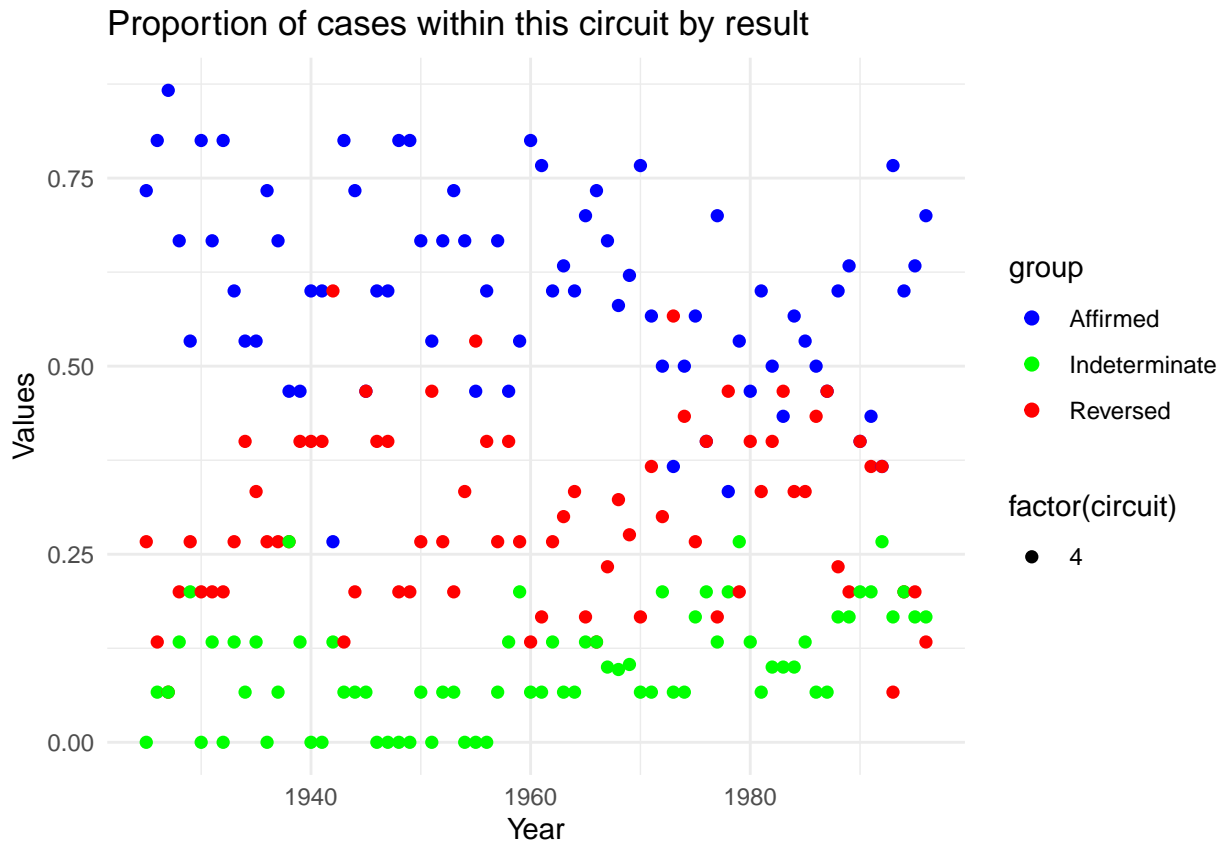


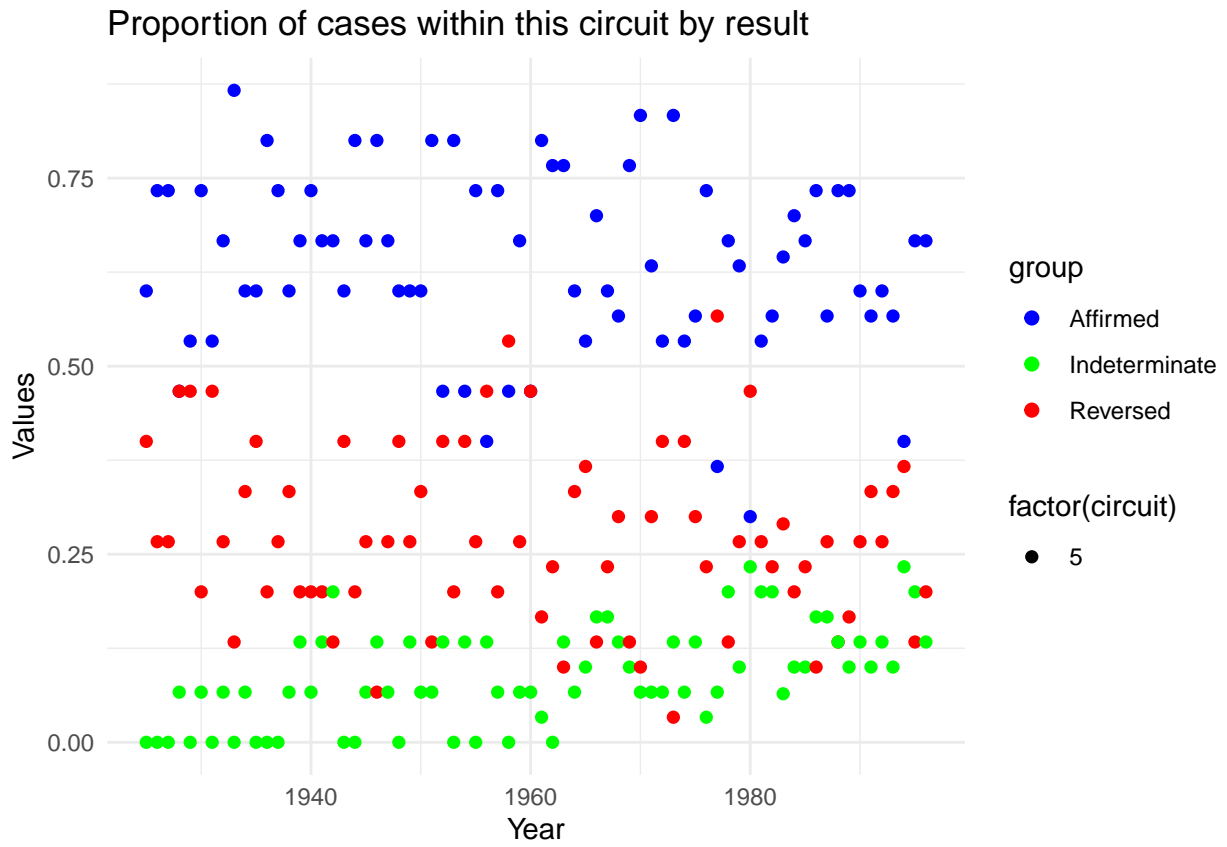
Proportion of cases within this circuit by result

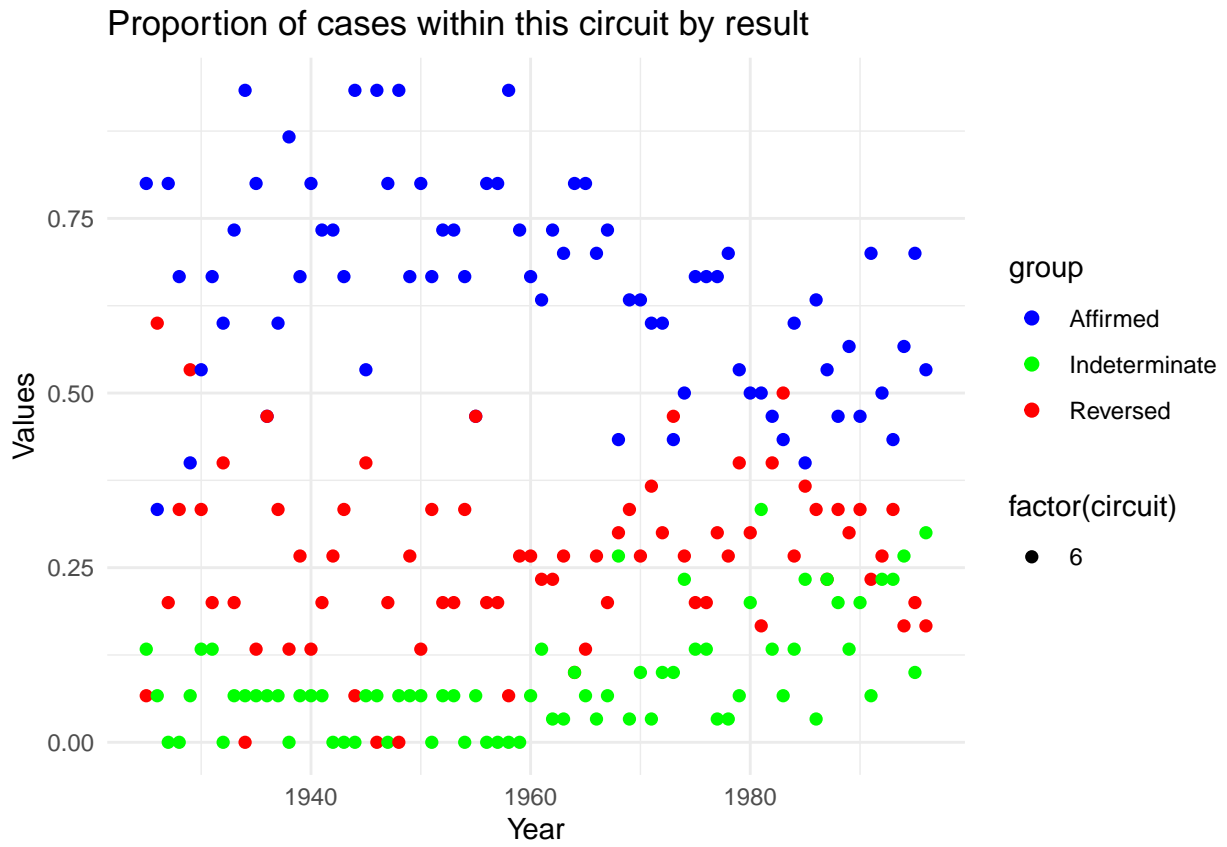


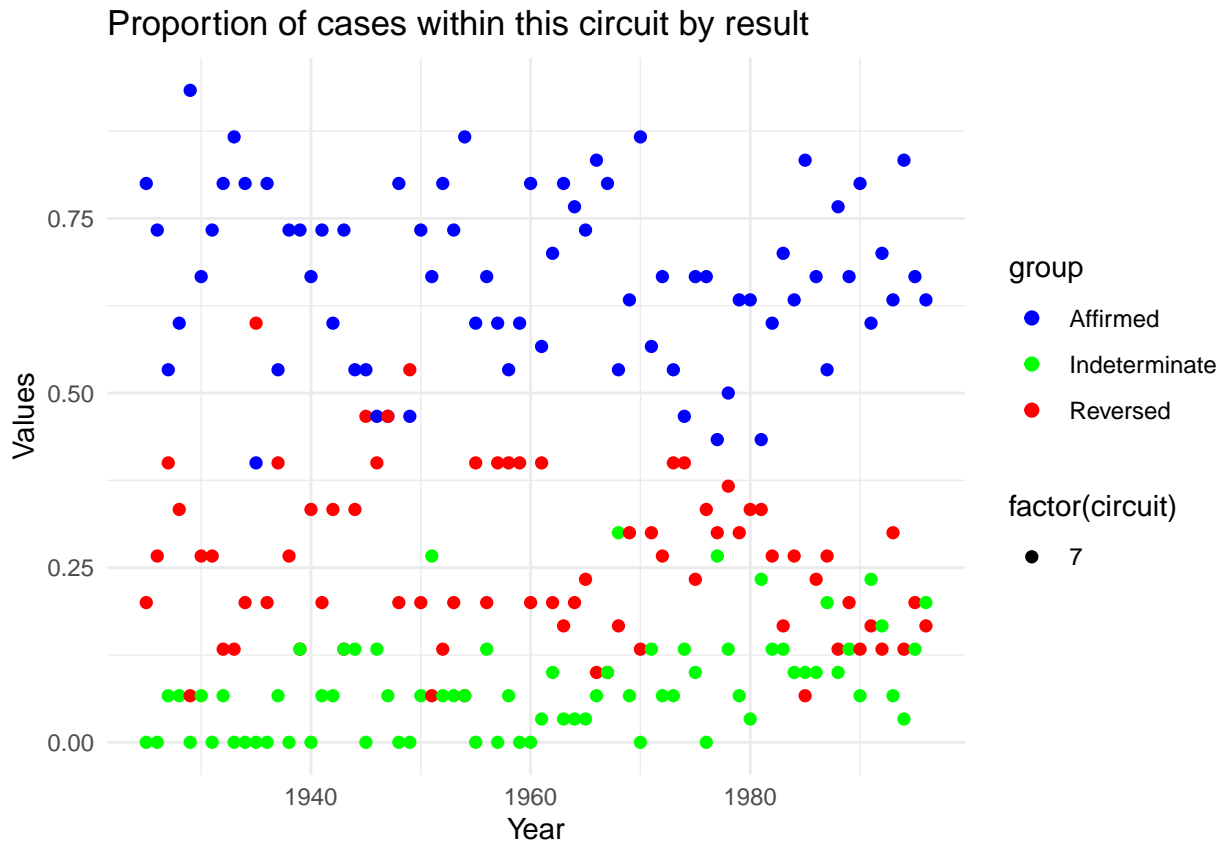




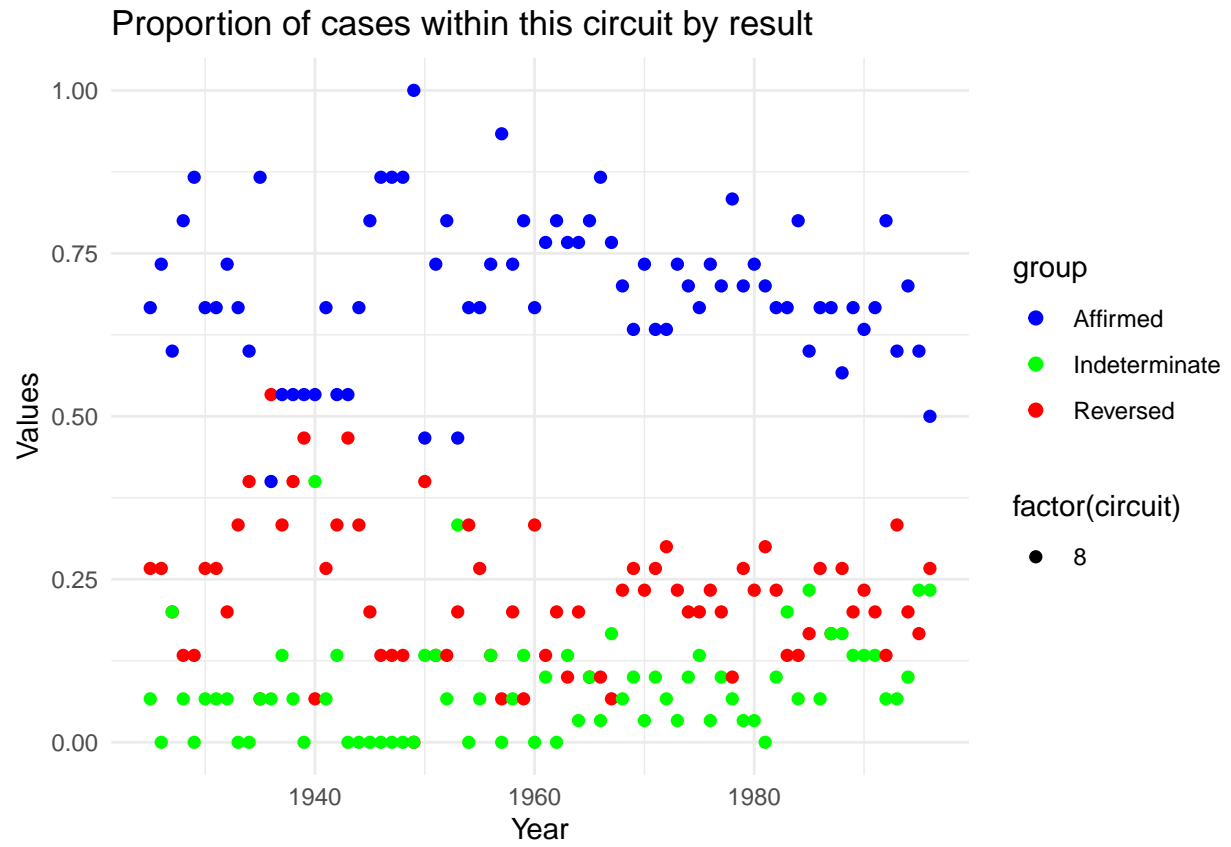


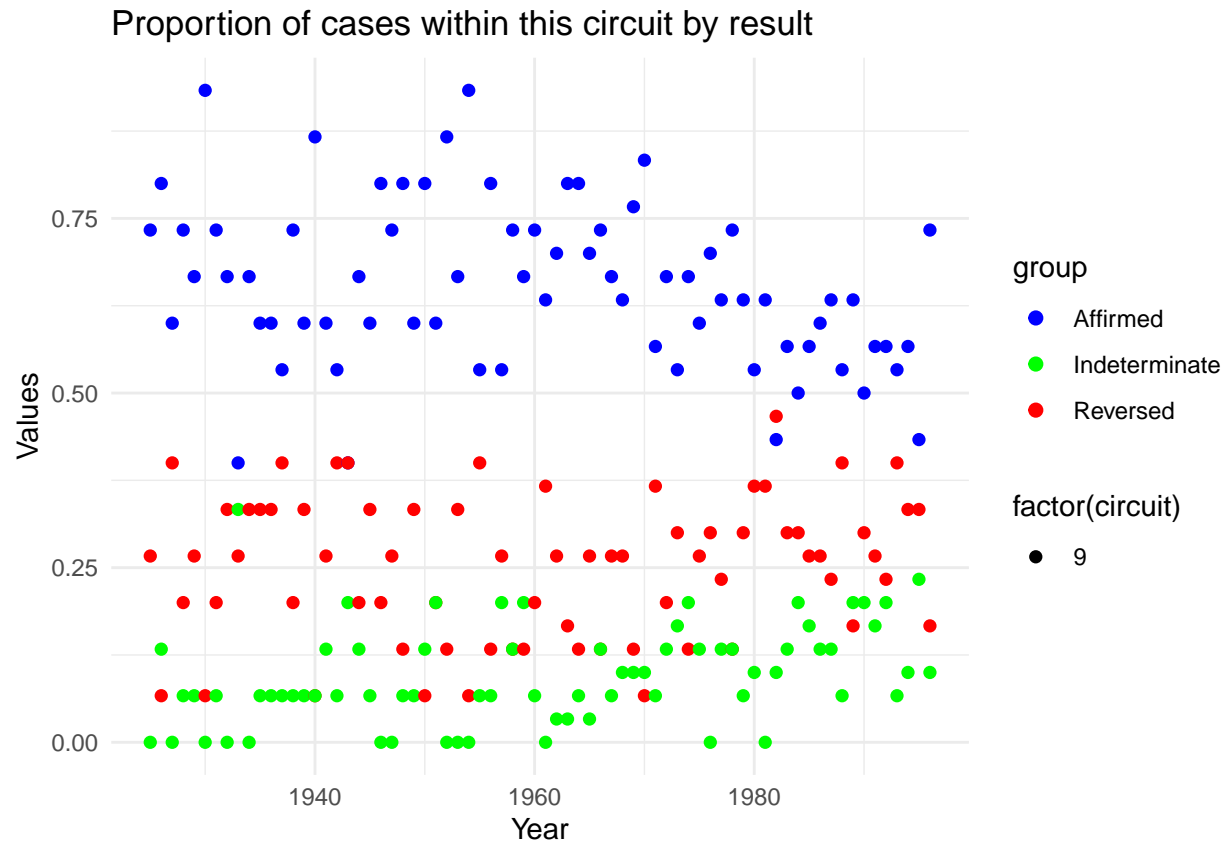




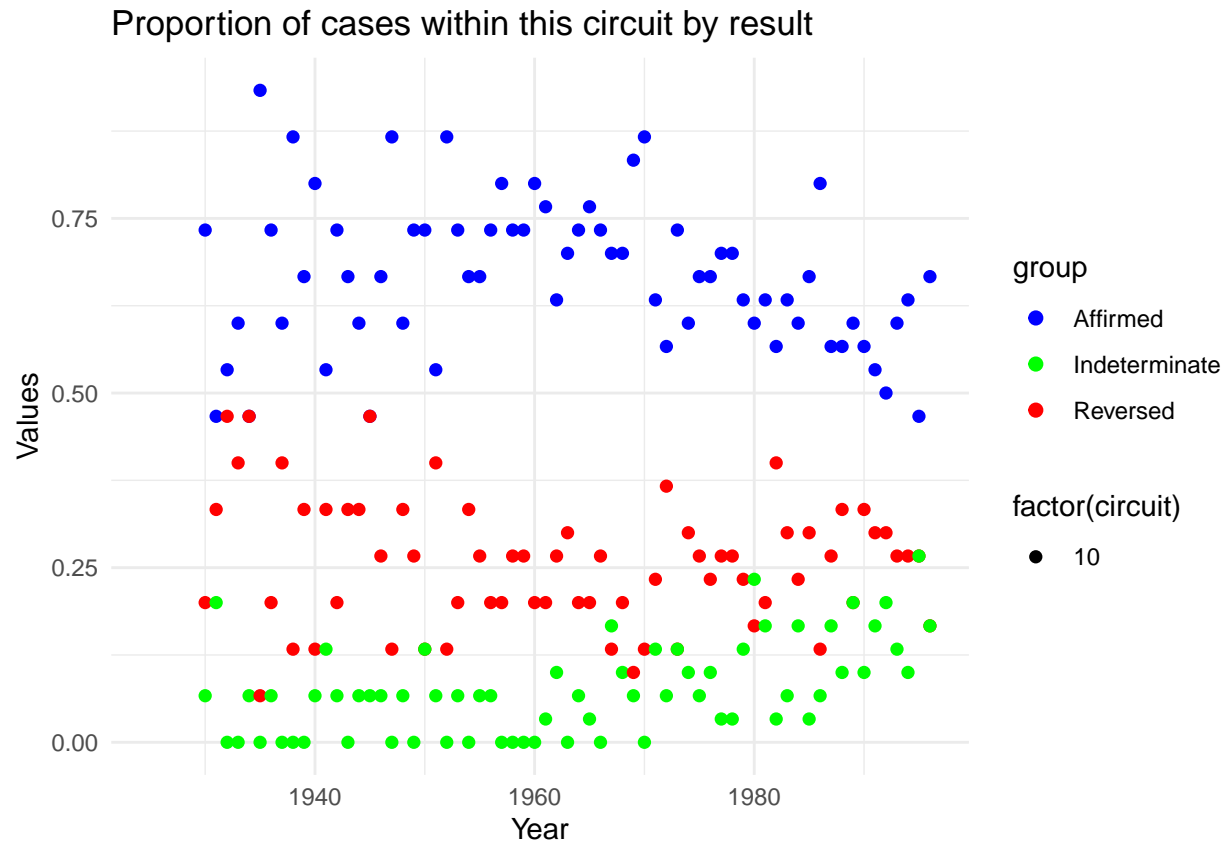




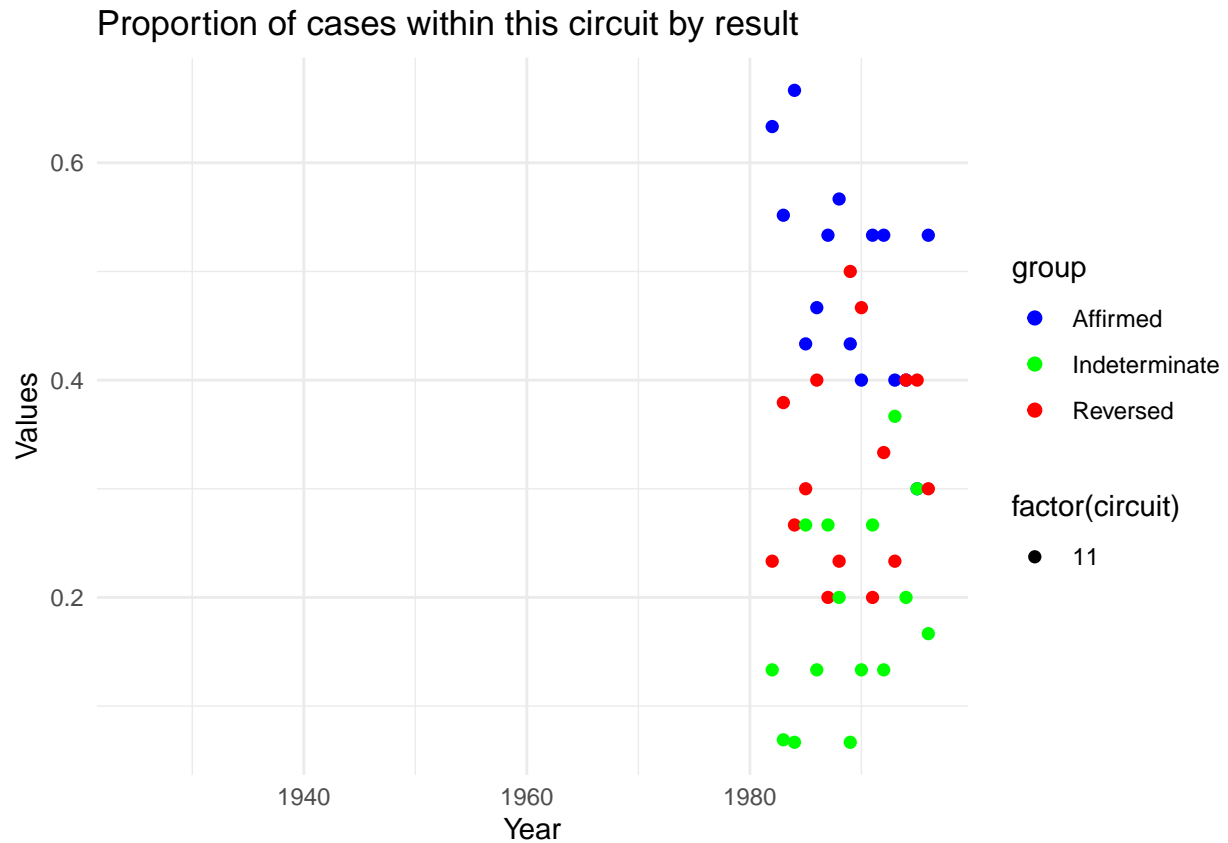




```
## Warning: Removed 15 rows containing missing values or values outside the scale range
## ('geom_point()').
```



```
## Warning: Removed 171 rows containing missing values or values outside the scale range
## ('geom_point()').
```



In the above charts I look circuit by circuit at the proportions of various decisions to see if a peculiarity of a single circuit or small group is causing the observed pattern. I find that there are some circuits that do not seem to have the pattern at all (e.g. 5, 7), some that exhibit sharp jumps similar to the aggregate data (0, 3, 6) and some that exhibit suggestive trends but do not clearly have a jump at 1970 (2, 4, 9). At this point I have stepped back from the data exploration to pursue further subject-matter research and see if there is some set of legal or historical facts that could explain this shift or further inform my continued exploration of this data.