CUNY School of Professional Studies

DATA 607: Data Acquisition and Management, Final Project

Michael O'Donnell, Spring 2019

Proposal (motivation for analysis):

In today's 24/7 news cycle we are bombarded with current events. We are notified about local, national, and international stories minute-by-minute on our phones and computers. This massive quantity of news often dilutes our attention, such that even major news stories may only hold public attention for a few hours. Then, we're onto the next.

Unfortunately, this news overload leaves us little time to revisit past events or explore history. Yet, history is critically important and tied to the present. So, for my DATA607 final project I chose to analyze historical data that has value today.

Thus, my final project is a study of the U.S. Senate. The Senate is one of our country's most important governing bodies, as old as the Constitution, reflects the political landscape, entails many forgotten details, and has simple data (two senators per state, six-year terms). Hence, my analysis will (hopefully) provide both an interesting historical overview and a renewed point of view on today's Senate.

My analysis of the Senate focuses on political party trends by time and U.S. region. My analysis is thorough, but there are many more insights in the dataset if you choose to dig in.

Largest Challenge:

The largest, and most unanticipated challenge of this project was aggregating US Senate data. Although the history of US Senate is documented, it is not kept in clean, tabular form. Rather, it is kept in text or scattered lists. Thus, it is easy to find the US Senators from Washington in 1905. But it is difficult to aggregate the number of years the US Senators from Washington were Democrats from 1905-2000. Hence, although the below analysis is interesting, the real value in this project is the raw, clean datasets used to drive the analysis.

Context:

The primary responsibilities of the Senate are:

- To pass legislation (in tandem with the House)
- To review, approve, and/or reject presidential appointees to executive and judicial branch posts
    - Example: The Senate can vote to approve or reject the presidential nomination of a supreme court justice
- To vote on impeachment of government officials (the impeachment process is initiated by the House of Representatives)
- To approve treaties made by the executive branch

Senate elections:

- Each state elects 2 Senators (currently 100 senators total)
- Each Senate term is 6 years
- There is no limit for number of times a Senator can be re-elected

Overview of the Senate, 1789-2019:

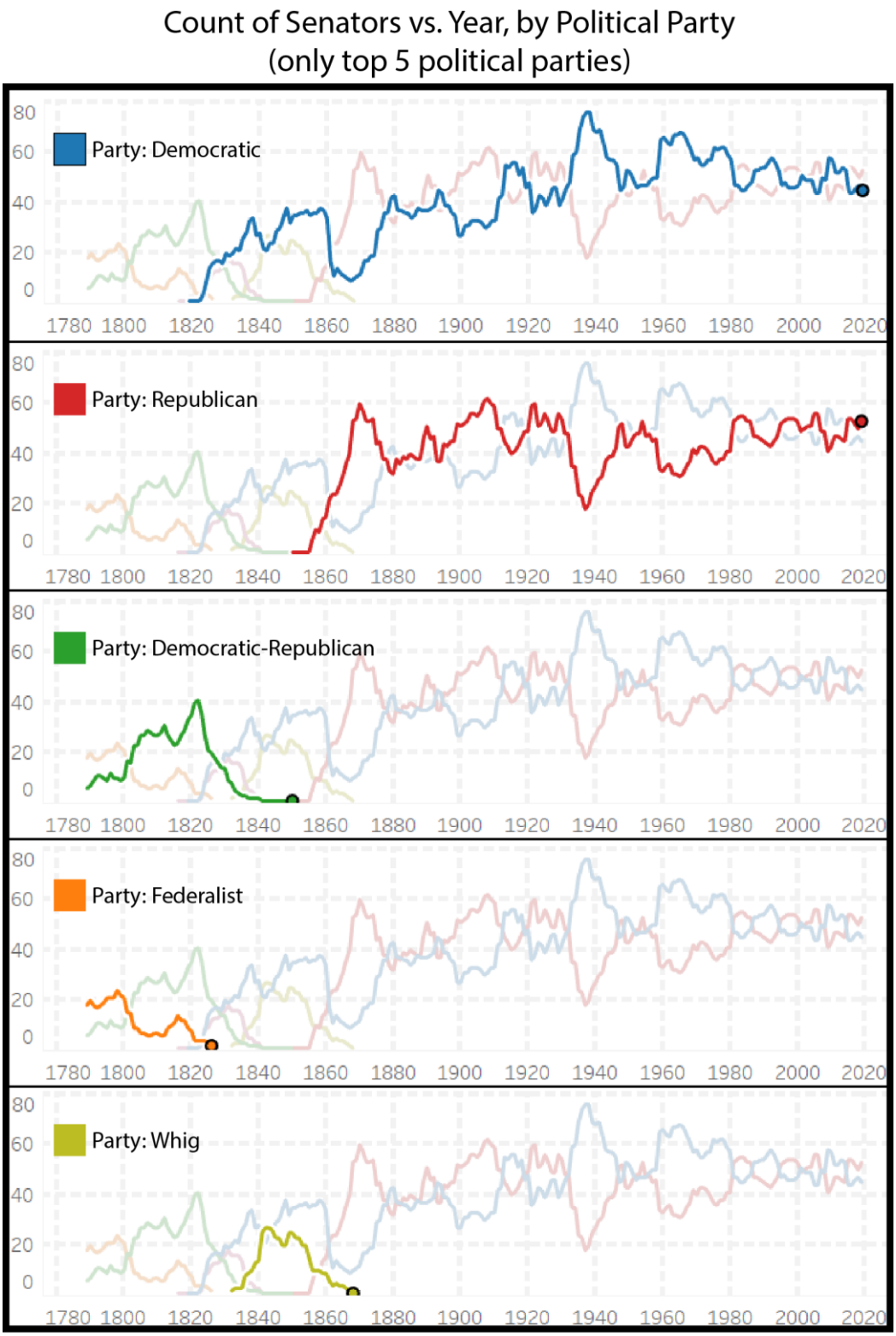
- Total Senators in U.S. history: 1,989
- Longest serving Senator: Robert C. Byrd – 51 years, West Virginia (1959-2010)
- Average Senate career: 8.97 years

Count of Senators by Political Party

Frequency	Party	# of Senators	% of Senators	First Year - Last Year
1	Democratic	879	44.20%	1819-2019
2	Republican	719	36.15%	1850-2019
3	Democratic-Republican	163	8.20%	1789-1850
4	Federalist	92	4.63%	1789-1826
5	Whig	65	3.27%	1832-1868
6	National Republican	29	1.46%	1816-1848
7	Free Soiler	6	0.30%	1849-1871
7	Populist	6	0.30%	1891-1902
7	Unionist	6	0.30%	1861-1870
10	Farmer-Laborite	4	0.20%	1923-1946
10	Independant	4	0.20%	1847-2019
12	Nullifier	3	0.15%	1831-1842
13	Oppositionist	2	0.10%	1855-1864
13	Readjuster	2	0.10%	1881-1888
13	Silver Republican	2	0.10%	1893-1902
16	American	1	0.05%	1855-1862
16	Anti-Monopolist	1	0.05%	1875-1880
16	Conservative	1	0.05%	1971-1976
16	Independance	1	0.05%	2002-2002
16	Law and Order	1	0.05%	1844-1844
16	Liberty	1	0.05%	1846-1846
16	Unconditional Unionist	1	0.05%	1863-1868

Today, we rarely reference Political Parties other than our two most popular, Democratic and Republican. But from the above table, we have elected Senators from 22 different political parties. The stacked line graph below displays the

frequency of senators by year for the top 5 political parties in the Senate. As you can see, our first 100 years were not all Democratic and Republican:

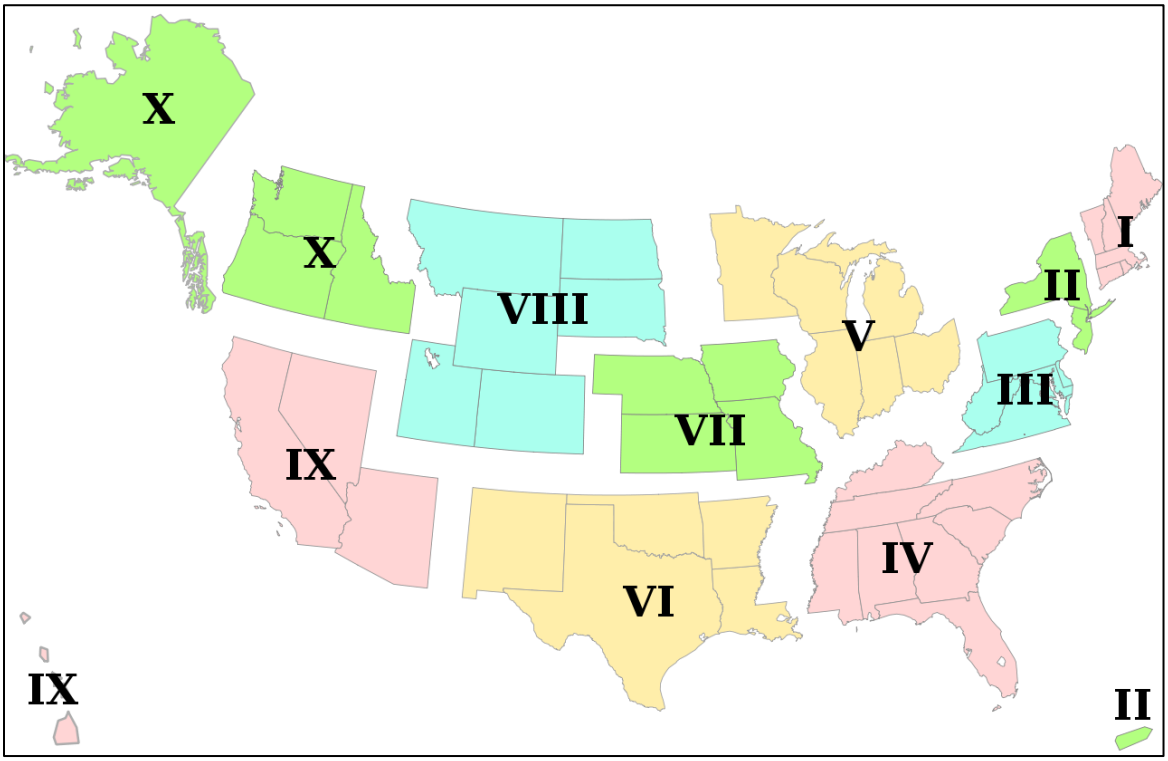


The US Senate by Federal Region:

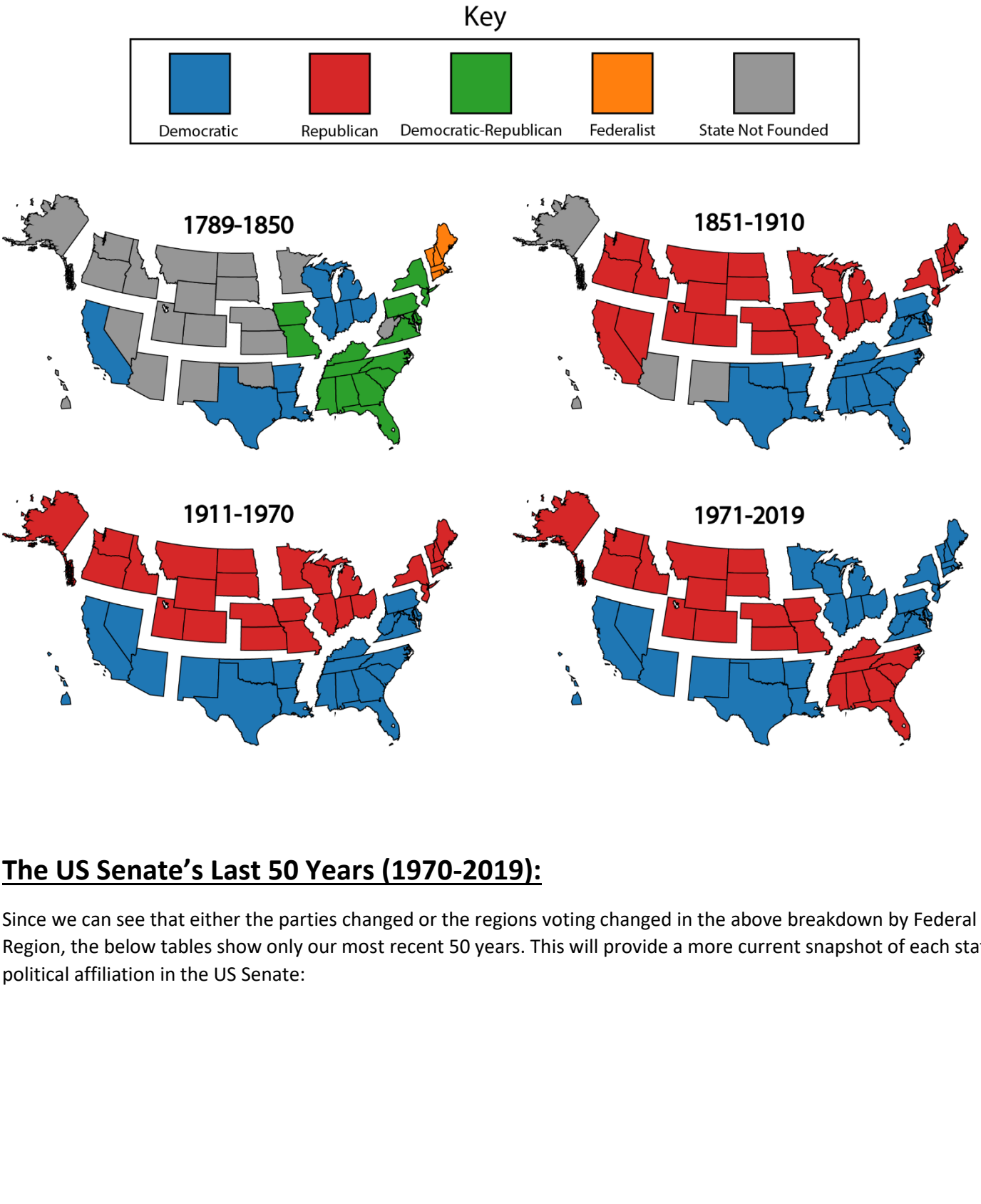
Another way to view the US Senate is by Federal Region. This showcases that regions of the United States have changed the parties they elect to the Senate over time. So, have the regions political opinions changed or have the parties changed? That will take further research, but the below will show change regardless.

Standard Federal Regions: The ten standard federal regions were established by the office of Management and Budget in April 1974, and required for all executive agencies.

- Region I: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
- Region II: New Jersey, New York, Puerto Rico, US Virgin Islands
- Region III: Delaware, District of Columbia, Maryland, Pennsylvania, Virginia, West Virginia
- Region IV: Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee
- Region V: Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin
- Region VI: Arkansas, Louisiana, New Mexico, Oklahoma, Texas
- Region VII: Iowa, Kansas, Missouri, Nebraska
- Region VIII: Colorado, Montana, North Dakota, South Dakota, Utah, Wyoming
- Region IX: Arizona, California, Hawaii, Nevada, American Samoa, Guam, Northern Mariana Islands
- Region X: Alaska, Idaho, Oregon, Washington



US Federal Region by Senate Majority



The US Senate's Last 50 Years (1970-2019):

Since we can see that either the parties changed or the regions voting changed in the above breakdown by Federal Region, the below tables show only our most recent 50 years. This will provide a more current snapshot of each state's political affiliation in the US Senate:

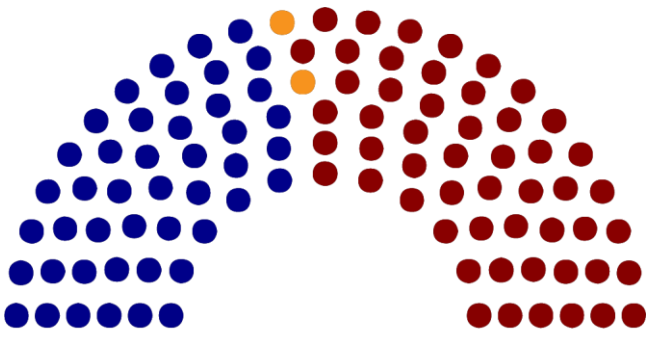
US Senate, Party by State from 1970-2019

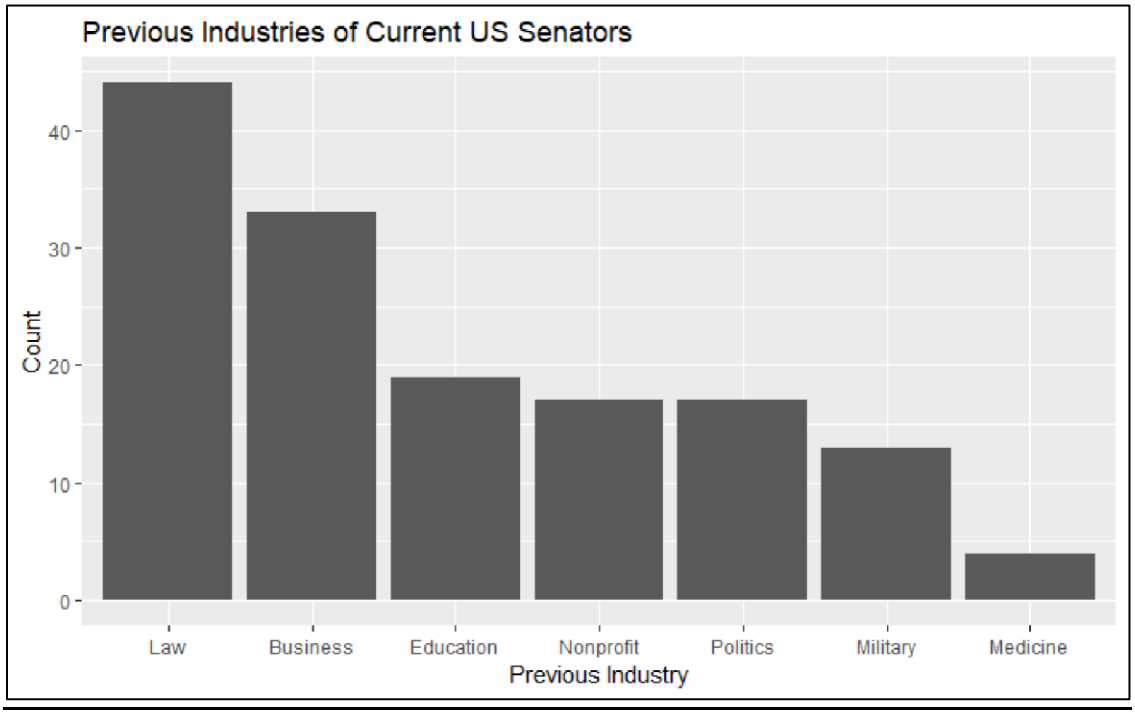
States with Democratic Majority, 1970-2019		
State	Democratic	Republican
West Virginia	96%	4%
Hawaii	93%	7%
New Jersey	91%	9%
Massachusetts	89%	11%
California	84%	16%
Michigan	84%	15%
Connecticut	82%	18%
Louisiana	81%	20%
Arkansas	79%	20%
Wisconsin	79%	21%
Montana	78%	23%
Maryland	77%	23%
Washington	76%	24%
North Dakota	74%	27%
Illinois	71%	28%
Rhode Island	69%	31%
Delaware	66%	33%
New York	64%	30%
Nevada	63%	37%
Florida	60%	38%
Ohio	59%	42%
Minnesota	58%	41%
New Mexico	58%	42%
Colorado	57%	43%
Georgia	54%	45%
Nebraska	53%	47%
South Dakota	53%	47%
Virginia	52%	48%

States with Republican Majority, 1970-2019		
State	Democratic	Republican
Kansas	0%	100%
Utah	7%	93%
Wyoming	6%	92%
Idaho	11%	89%
Alaska	17%	83%
Pennsylvania	17%	83%
Oklahoma	18%	82%
Arizona	19%	80%
Texas	24%	77%
Mississippi	27%	73%
Tennessee	28%	72%
New Hampshire	28%	70%
North Carolina	32%	68%
South Carolina	35%	65%
Indiana	36%	64%
Oregon	35%	64%
Missouri	36%	63%
Maine	31%	62%
Alabama	41%	60%
Kentucky	37%	60%
Iowa	47%	53%
Vermont	45%	42%

Current State of the US Senate, 2019:

Finally, a quick look at today's Senate by political party. Then for fun, a breakdown of our current Senate members' previous professions.





Data Sources (2 different types):

1. Scraped html tables, https://en.wikipedia.org/wiki/List_of_former_United_States_Senators
    a. All former US Senators
    b. Source: Wikipedia, validated against senate.gov
2. CSV file, "data/add_current_senators.csv"
    a. All current US Senators
    b. Source, Wikipedia, validated against senate.gov

Data Transformation:

In the code, it is clear the data goes through many transformations. Primarily, from its original scraped form to its final two data frames ("senator.by.year" and "us.senate.final"). This was difficult, but vital for aggregation and analysis. I needed one data frame with a row for each US Senator, and a second with a row for each year of each Senator. To do so, I used stringr, dyplr, and many for loops to transform rows like this:

"1878-18811896-1900"

To rows like this:

"1878, 1879, 1880, 1881, 1896, 1897, 1898, 1899, 1900"

Further transformation was done on these rows to associate each year with a political party. I found many Senators served in the Senate for multiple political parties.

Statistical Analysis to validate data:

I used dyplr to validate my data. I validated the number of Senators I had in my data frame per year:

Year <chr>	Count <int>
2019	100
2018	100
2017	100
2016	100
2015	100
2014	100

After over 600 lines of R to scrape and transform my data into a data frame that had an individual row for each year or each US Senator, this validation was critical. Note: this was difficult because Senators serve in 2 calendar years during each year they serve (January 3rd to January 3rd), so I had to programmatically remove the second January dates of each year.

One feature (package, etc) we did not cover in class:

I used ggplot to create a bar graph of the current Senators' previous professions. I also exported my data frames into Tableau to create the line graphs featured in the "Overview of the US Senate" section above.

One challenge I encountered:

During this project, the largest challenge I encountered was creating a data frame with an individual row for each year of each US Senator with associated State and Political Party. From my research, this type of aggregation is not public on the internet today. Thus, I had no dataset to easily validate against nor precedent to follow.

More specifically, while creating this data frame I found about 8% of Senators in U.S. history have served in the Senate for more than one political party. Additionally, over 10% of Senators have a gap between time they served in the Senate (ex: served from 1900-1906, then again from 1912-1918). Thus, identifying the political party that each Senator associated with for each year they were in office was challenging. For example, one Senator served four non-consecutive terms in the Senate under three different political parties.

Conclusion:

I am proud of the effort this project took to reach its conclusions. Although the analysis is interesting, the raw data frames I created were the most time consuming and valuable part of this project.