Son Bias in the US: Evidence from Business Names

Walter Guillioli and Gaurav Sood

20 January, 2020

Abstract

Do we want one? Maybe write at the end.

1. Introduction

NEEDS WORK. Will write at the end. We estimate preference for passing on businesses to sons by examining how common words son and sons are compared to daughter and daughters in the names of businesses. intro. = bias against women is common one way bias manifests itself is that people are less likely to transfer their businesses to their daughters we exploit a novel publicly available administrative source of data

2. Data and Methods Overview

Our data acquisision process is divided in two steps which are briefly explained below.

2.1 Acquire Business Names across US states

In the United States, businesses have to register with their state and all states provide a website to search for business names. The functionality of these websites vary by state which made the data acquisition harder. We began by searching for businesses with the words son(s) and daughter(s) on their names. These wasn't a trivial process for several reasons and some are worth highlighting here along with the solution.

a) Search results for son(s) are inflated

Search results for son(s) are inflated mainly for three reasons. (a) son is part of many English words, from names such as Jason and Robinson to ordinary English words like mason (which can also be a name), (b) son is a Korean name and (c) some businesses use the word son playfully; for instance, son is a homonym of sun and some people use that to create names like son of a beach. We address (a) by cleaning the data using regular expressions to only look for exact matches of son and sons.

b) Limits in the number of results shown

Some states limit the number of search results. For example, Alabama only displays up to 1,000 results. This is tricky because we know there are at least 1,000 companies with son(s) on their name but we don't know how many. In this case we can only derive a conservative estimate for the ratio of companies with son vs daughter and we note that on the results. In order to increase the number of samples in this case we do two searches, one for son and one for sons. we then combine the results knowing there might be some overlap but we dedup these before the analysis.

c) Technological challenges in data acquisition

The technologies used for these websites is different. In some cases we are able to simply copy and paste the results to our computer for analysis. But in other cases more sophisticated scrapping tools were built to parse and download the data using packages like rvest in R and selenium in Python.

After acquirign the number of companies with the word son(s) and daughter(s) on their names we calcualted the son/daughter ratio which is the estimate of most concern in this paper.

2.2 Additional State Information

Additionaly, we enriched our dataset by acquiring state data from other sources to profile the results. These new attributes include: US Region of the state, US Division of the state, population of the state, GPD of the state, political party of the state and number of establishments on each state. The sources are identified in the References section of this paper.

2.3 Final Dataset used in Analysis

Due to the challenges outlined above some care is needed when interpreting the results. All in all, we were able to acquire data for 36 states. Though not all 50 states were covered we believe we have a good representation of the United States since these 36 states represent 69.9% of the US population, 71.2% of the US GDP and 71.% of the registered establishments.

The data and scripts used are posted here: https://github.com/soodoku/sonny_side

A sample of the final dataset used for analysis is displayed here:

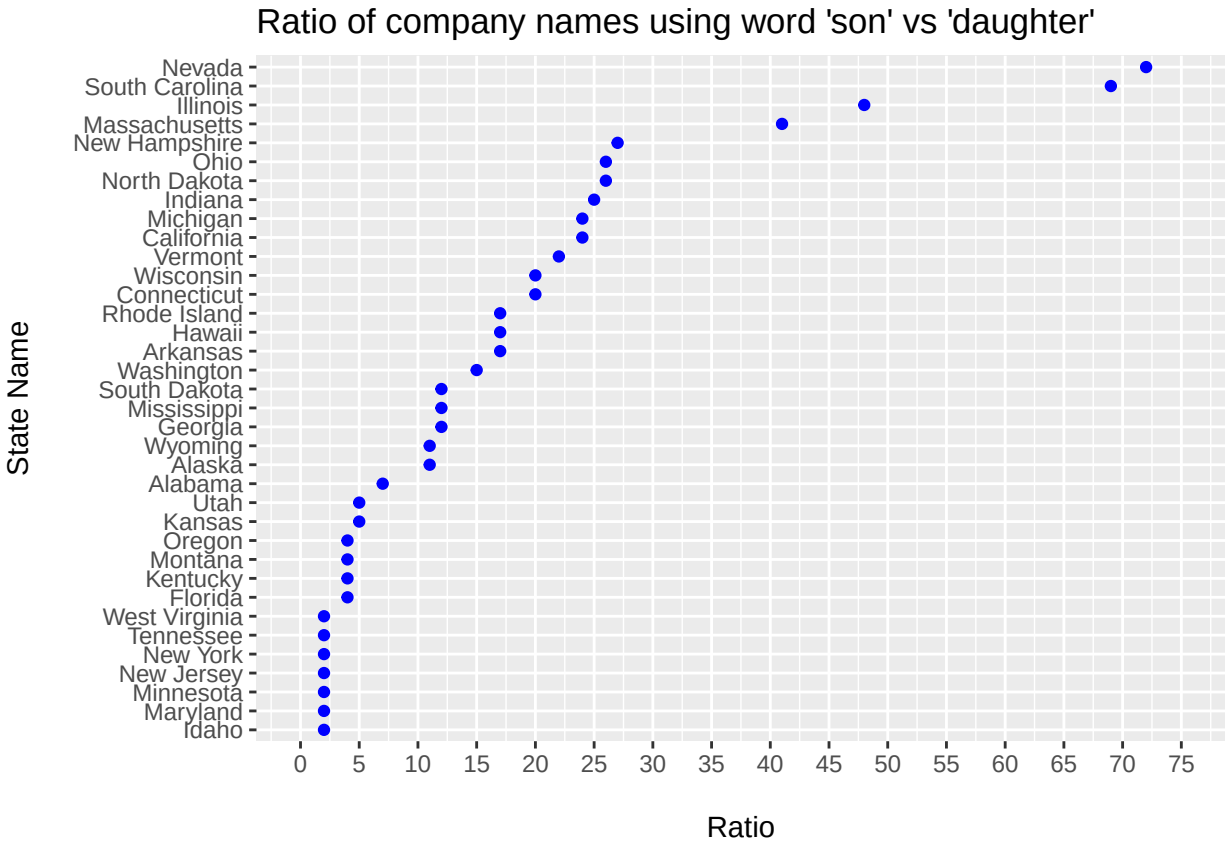
NOTE: This table looks horrible. See if I can do stargazer package here?

```
##          Name estimate  son daughter    Region           Division popestimate2019pct
## 1     Alabama        7  884      126     South East South Central        0.014937826
## 2      Alaska       11  246       22      West            Pacific        0.002228693
## 4    Arkansas       17 1482       87     South West South Central        0.009193908
## 5  California       24 3609      150      West            Pacific        0.120376189
## 7 Connecticut       20  875       43 Northeast        New England        0.010861846
## 9     Florida        4  729      176     South     South Atlantic        0.065433123
##   presidentialelection2016 gdp_pct establishments_pct
## 1               Republican   0.011        0.012836617
## 2               Republican   0.003        0.002716876
## 4               Republican   0.006        0.008457416
## 5               Democratic   0.145        0.118909506
## 7               Democratic   0.013        0.011525938
## 9               Republican   0.051        0.070408815
```

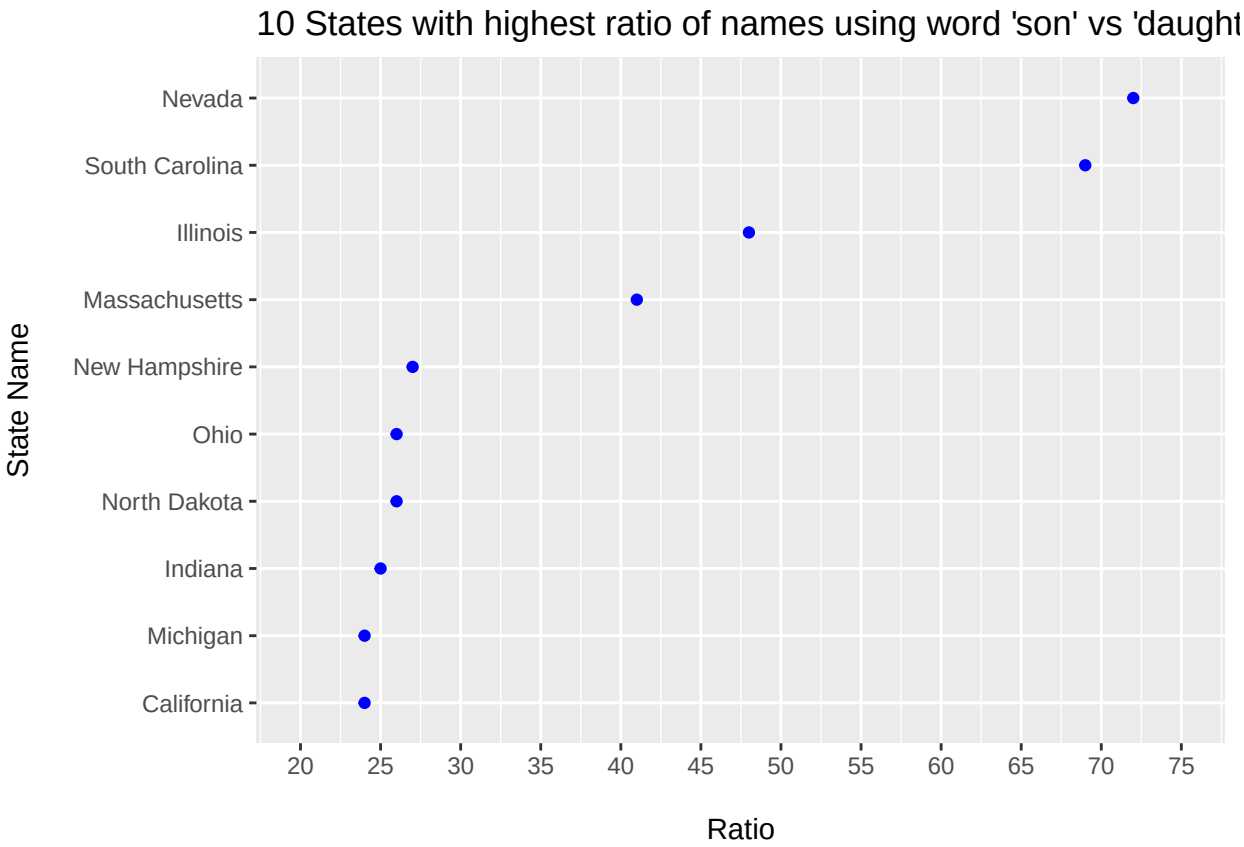
Question for Gaurav – do we want to do some univariate EDA and description? I doubt it.

3. Results

In all, we find that a conservative estimate of son to daughter ratio is between 2 to 1 to 72 to 1 across the 36 states where we have data with a median of 12 to 1. This is displayed in the figure below.



NOTE: If the above feels too busy we could show the top 10 only like this:

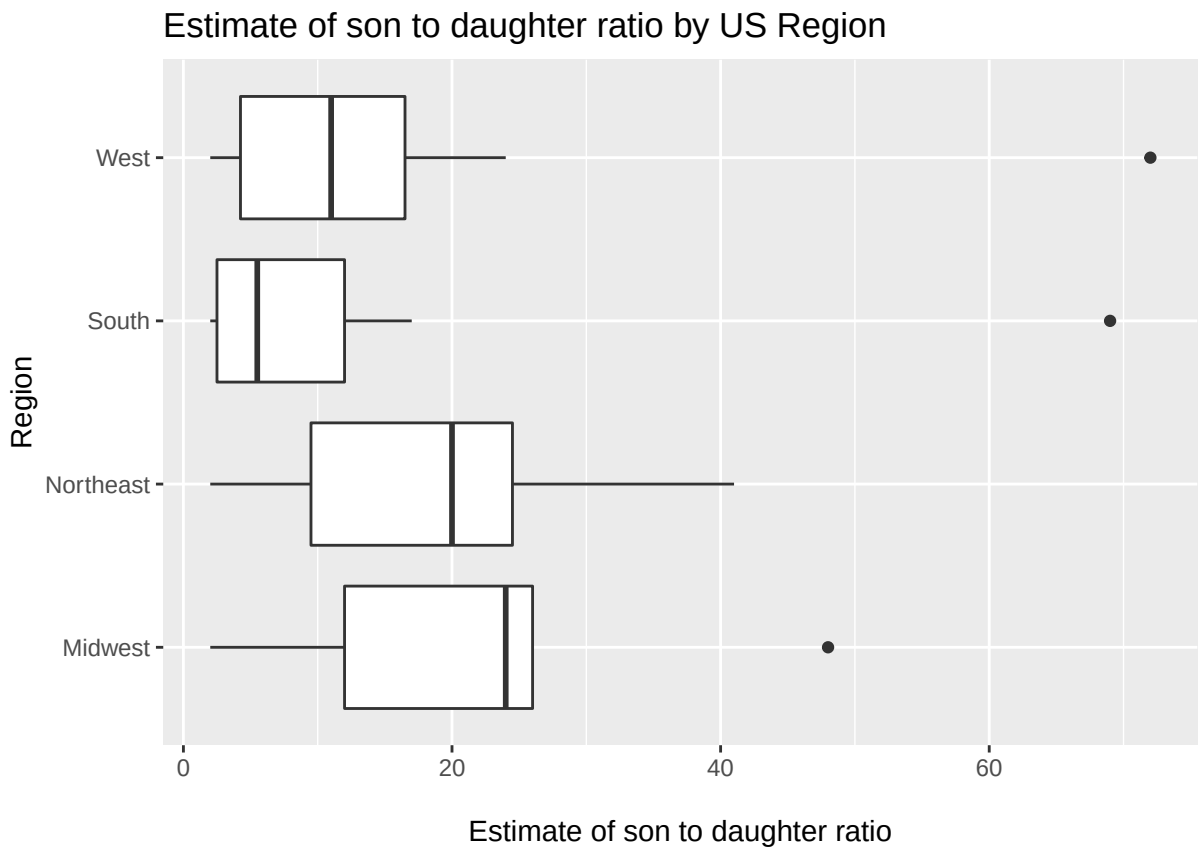


We know proceed to explore how these results vary by location of the state in the United States, by the state population and political party and by its GDP and number of bueinsss establishments.

NOTE: I CAN BRING % OF MAILES VS FEMALES IN STATE TO SEE IF ANYTHING THERE? OR OTHER VARIABLES TO CORRELATE. MAYBE LATER.

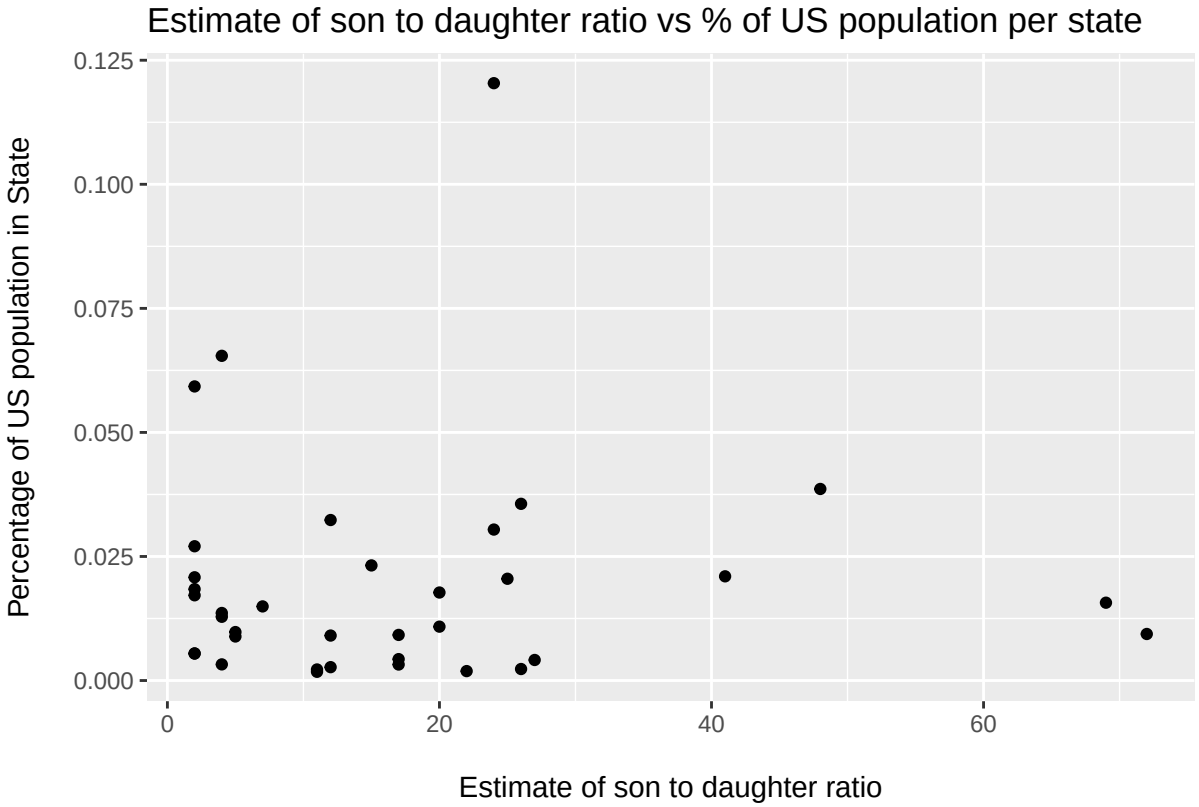
3.1 Differences by US Region

When we look at the estimate of son to daughter ratio by Region in USA we see states in the Midwest and Northeast with higher ratios when comparing to the West and particularly the South.The biggest gap is Midwest vs South with a median ratio of 24.0 vs 5.5 respectively.



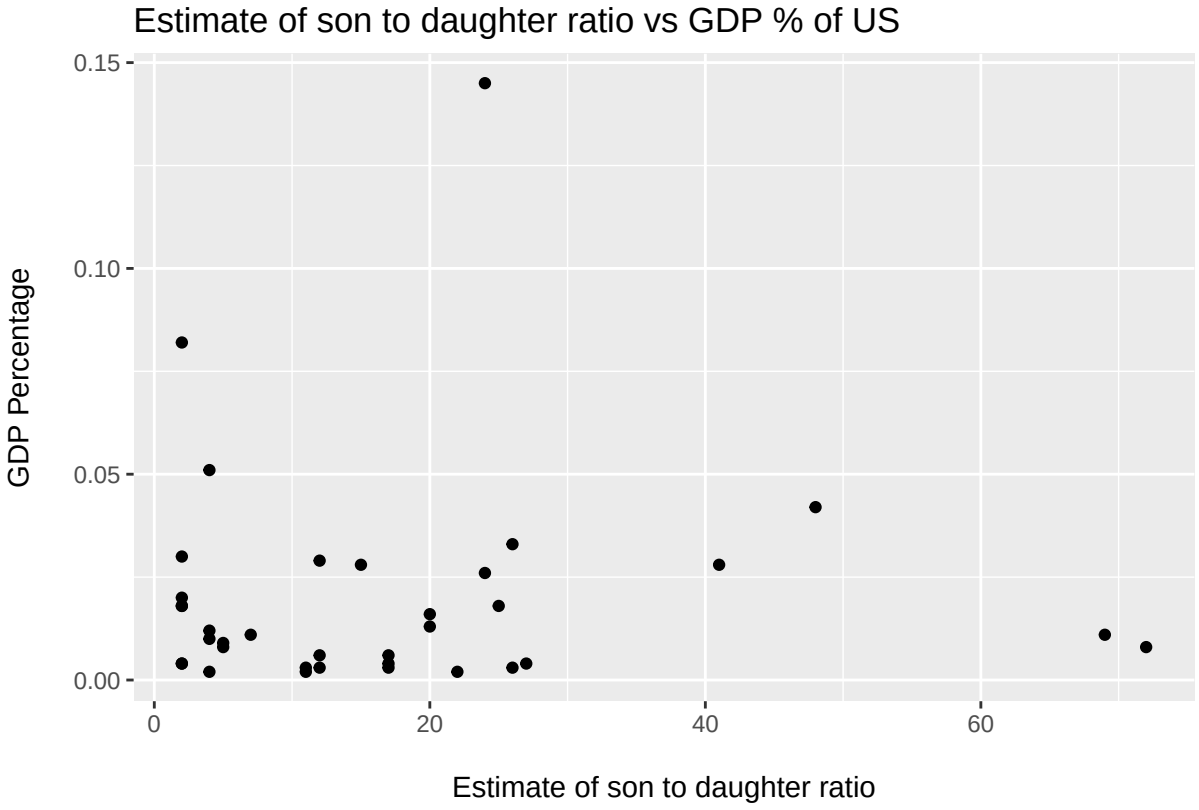
3.2 Relationship with the Population size of the State

When we look at the estimate ratio of son vs daughter with the population fo the state we don't see any relationship between these data points. In fact the correlation if basically zero as seen below.



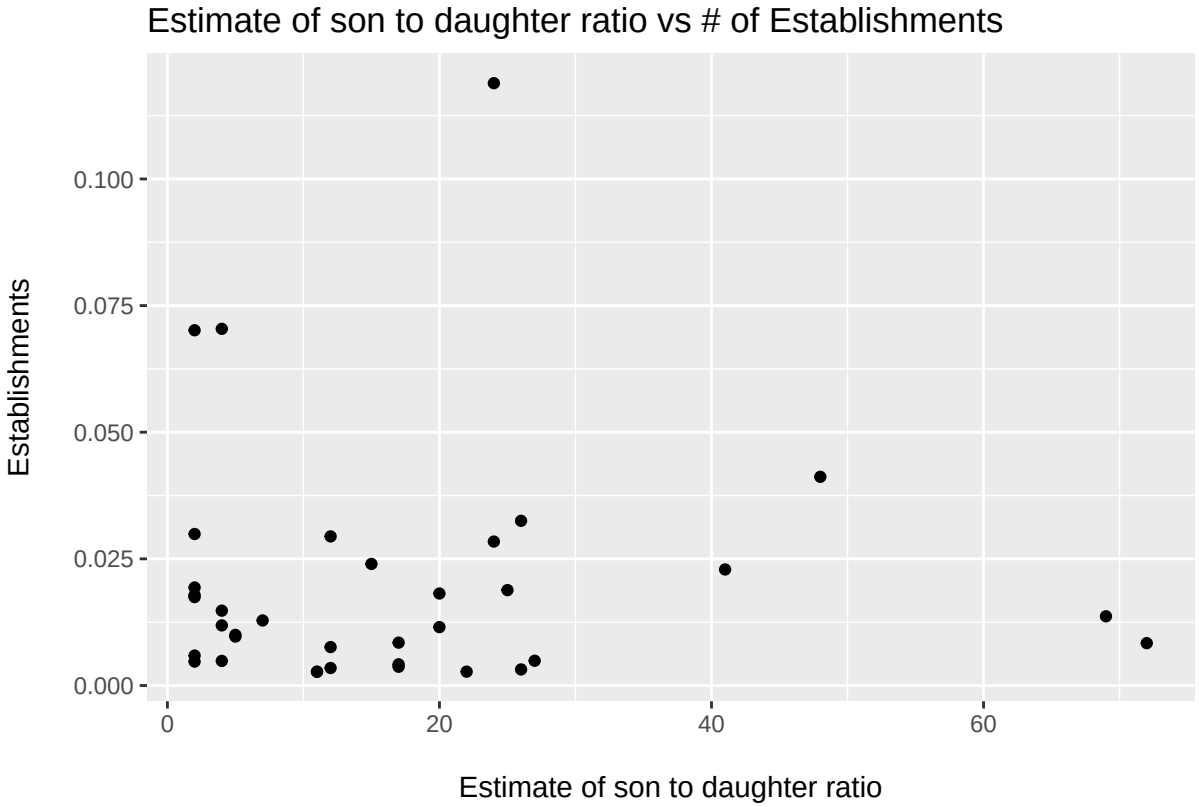
3.3 Relationship with the GDP of the State

We also looked at how the differences of the ratio of business names using son vs daughter could vary by state as it relates to the size of the state in terms of percentage of the gross domesic product (GDP) of the country. We didn't find any evidence of relationship between these two with a correlation of basically zero, as seen in this figure.



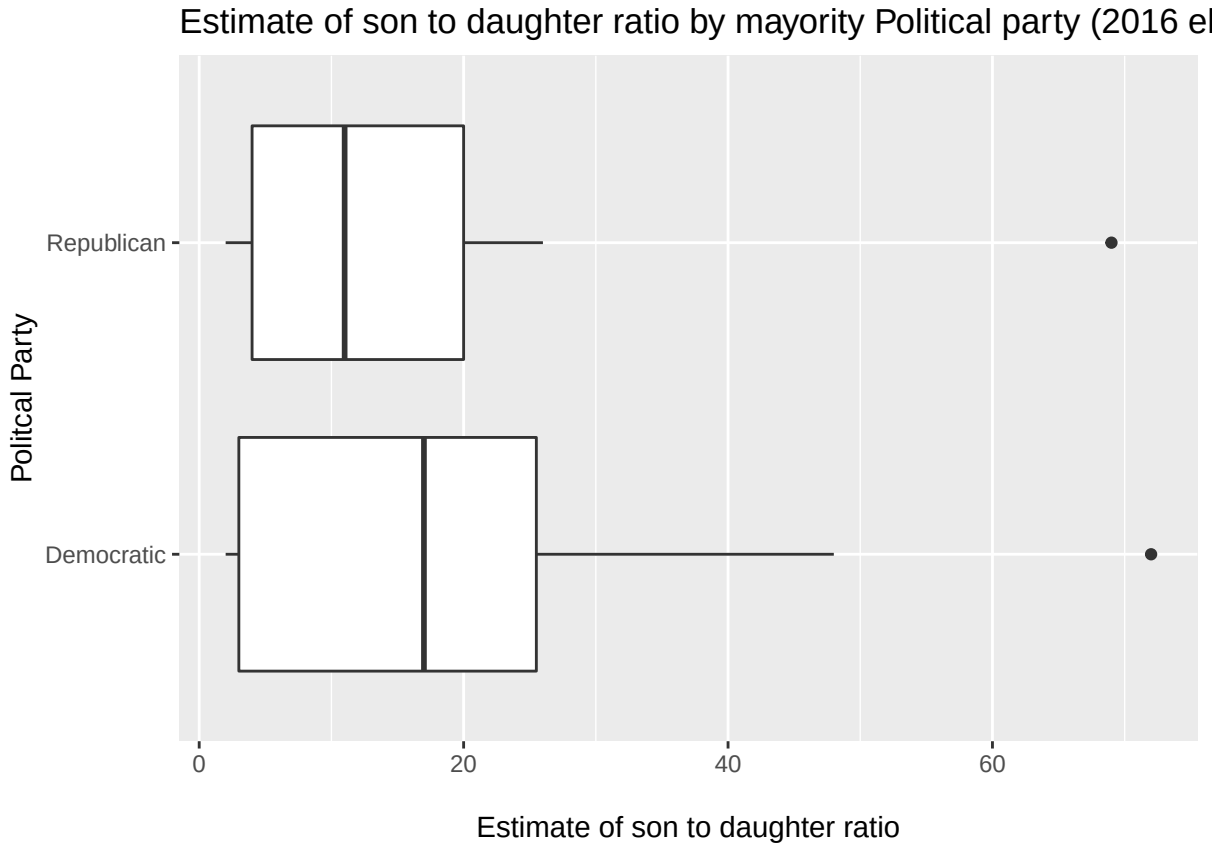
3.4 Relationship with the number of Establishments of the State

We also obtained data from the census organization from US that offers the number of registered establishments in the USA, this is not necesarily the same as the number of business companies registered per state but since obtaining that exact number wasn't not possible we use this as a proxy. Again, no evidence or correlation is seen per figure below.



3.5 Relationship with the major Political Party of the State

Finally we looked at the voting data from the 2016 elections and compare how the ratio of son vs daughter is when separating states with a majority of Democratics vs Republicans. As can be seen below the ratio tends to be higher on Democratics states with a median of 17 vs a median of 11 for Republican states.



4. Conclusion

There is clearly an inclination to name businesses including the word son(s) vs daughter(s). We found evidence for 36 states that that a conservative estimate of son to daughter ratio is between 2 to 1 to 72 to 1 across the 36 states where we have data with a median of 12 to 1. Despite not having data for 50 states we feel this is a good representation of the whole country since these 36 states represent 69.9% of the US population, 71.2% of the US GDP and 71.% of the registered establishments.

Although we didn't find any relatioship with the size of the states in terms of GDP, population or number of establishments we do see some differences across regions and political parties dominating the state. We cannot conclude any causality because of this but further exploration is recommended.

References

1. Kingl, Arvid. Web Scraping in R: rvest Tutorial. https://www.datacamp.com/community/tutorials/r-web-scraping-rvest
2. Halpert, Chris. US census bureau regions and divisions. https://github.com/cphalpert/census-regions/
3. United States Census Bureau. State Population Totals and Components of Change: 2010-2019. https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html
4. Wikipedia contributors. (2020, January 13). Political party strength in U.S. states. In Wikipedia, The Free Encyclopedia. Retrieved 03:11, January 20, 2020, from https://en.wikipedia.org/w/index.php?title=Political_party_strength_in_U.S._states&oldid=935536430
5. United States Census Bureau. SUSB Historical Data. https://www.census.gov/data/tables/time-series/econ/susb/susb-historical.html

6. Kaushik, Saurav. Beginner's Guide on Web Scraping in R (using rvest) with hands-on example. https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/

Appendix

Appendix A - States with number of companies found with word son and daughter and ratio

NOTE: add column indicating which is super conservative based on limit for search and why.

NOTE: table looks bad, stargazer package maybe?

	Name	estimate	son	daughter
1	Alabama	7	884.0	126
2	Alaska	11	246.0	22
4	Arkansas	17	1482.0	87
5	California	24	3609.0	150
7	Connecticut	20	875.0	43
9	Florida	4	729.0	176
10	Georgia	12	6002.0	497
11	Hawaii	17	1454.0	88
12	Idaho	2	60.0	39
13	Illinois	48	2324.0	48
14	Indiana	25	4928.0	195
16	Kansas	5	75.0	14
17	Kentucky	4	66.0	16
20	Maryland	2	128.0	82
21	Massachusetts	41	5979.0	147
22	Michigan	24	2265.0	93
23	Minnesota	2	392.0	213
24	Mississippi	12	1918.0	165
26	Montana	4	240.0	66
28	Nevada	72	1440.0	20
29	New Hampshire	27	3203.0	119
30	New Jersey	2	173.0	73
32	New York	2	1190.0	745
34	North Dakota	26	605.0	23
35	Ohio	26	2550.0	100
37	Oregon	4	1000.0	227
39	Rhode Island	17	206.0	12
40	South Carolina	69	4083.3	59
41	South Dakota	12	129.0	11
42	Tennessee	2	203.0	132
44	Utah	5	81.0	16
45	Vermont	22	1361.0	63
47	Washington	15	2424.0	161
48	West Virginia	2	128.0	72
49	Wisconsin	20	845.0	43
50	Wyoming	11	238.0	21