Son Bias in the US: Evidence from Business Names

Walter Guillioli and Gaurav Sood

23 January, 2020

1. Introduction

We explore bias and preference for business owners in the United States to inherit their businesses to their sons instead of their daughters. We do this by examining how common words "son(s)" compare to "daughter(s)" in the names of businesses.

To do this, we collect the information that states in the United States have avaialble online as it relates to businesses already registered in the state. This process is not trivial since all states have different ways and rules to allowing to search for this data and in some cases obtaining all the data is impossible.

Once we collect the companies using the words son(s) and daughter(s) we estimate a ratio of these two numbers to measure the son to daughter bias per state. Finally, these data is correlated with other metrics in the state like GDP, population, geograhic location and political majority.

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. This is not a trivial process for several reasons. The main challenges and nuances are:

a) **Search results for son(s) are inflated** . This is mainly for three reasons. First, 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). Second, son is a Korean name. Third, 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 the first issue by cleaning the data using regular expressions to only look for exact matches of son and sons. We do not deal with the other two issues but we believe the impact is minimal.

b) **Limits in the number of results shown**. Some states show all the results when doing a search but some states limit the number of search results shown. For example, Alabama only displays up to 1,000 results. This has significant impact since we only know if a particular search for son(s) has more than 1,000 results but we don't know how many - it could be 10,000 or 500,000. To deal with this challenge, we only derive a conservative estimate for the ratio of companies with son vs daughter and we note that on the results. In addition, in order to increase the number of samples in some cases we do multiple searches for son and sons and for business names starting and containg this text. We then combine the results knowing there might be some overlap but we dedup before the analysis.

c) **Technological challenges in data acquisition**. In some states we are able to simply copy and paste the results in tabular format 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

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, major 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. 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.

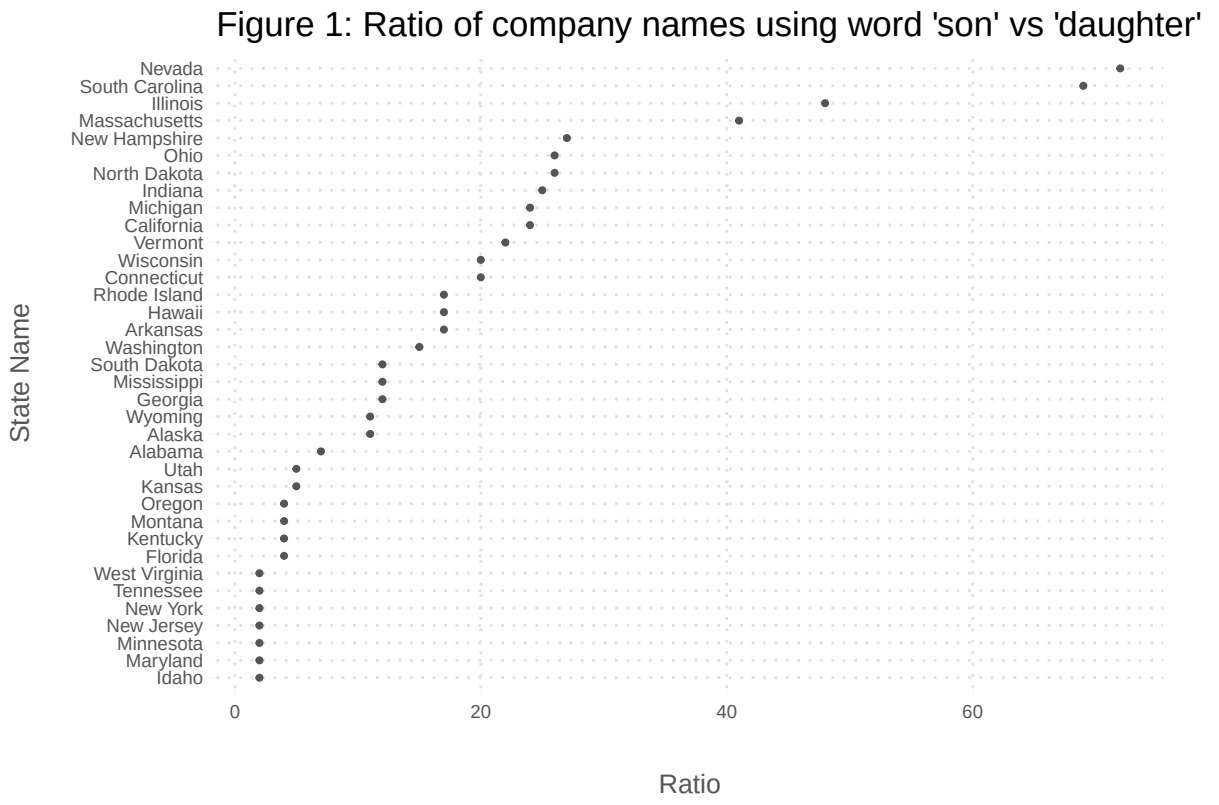
The data and scripts used are posted here: https://github.com/soodoku/sonny_side. Table 1 shows a sample of 10 states from the final dataset used for analysis.

Table 1: Table 1: Sample of Data Collected fro 10 States

State	Son/Daughter	US Region	Population %	GDP %	Establishments %	Political Party
Alabama	7	South	0.015	0.011	0.013	Republican
Alaska	11	West	0.002	0.003	0.003	Republican
Arkansas	17	South	0.009	0.006	0.008	Republican
California	24	West	0.120	0.145	0.119	Democratic
Connecticut	20	Northeast	0.011	0.013	0.012	Democratic
Florida	4	South	0.065	0.051	0.070	Republican
Georgia	12	South	0.032	0.029	0.029	Republican
Hawaii	17	West	0.004	0.004	0.004	Democratic
Idaho	2	West	0.005	0.004	0.006	Republican
Illinois	48	Midwest	0.039	0.042	0.041	Democratic
Indiana	25	Midwest	0.021	0.018	0.019	Republican
Kansas	5	Midwest	0.009	0.008	0.010	Republican
Kentucky	4	South	0.014	0.010	0.012	Republican
Maryland	2	South	0.018	0.020	0.018	Democratic
Massachusetts	41	Northeast	0.021	0.028	0.023	Democratic
Michigan	24	Midwest	0.030	0.026	0.028	Republican
Minnesota	2	Midwest	0.017	0.018	0.019	Democratic
Mississippi	12	South	0.009	0.006	0.008	Republican
Montana	4	West	0.003	0.002	0.005	Republican
Nevada	72	West	0.009	0.008	0.008	Democratic
New Hampshire	27	Northeast	0.004	0.004	0.005	Democratic
New Jersey	2	Northeast	0.027	0.030	0.030	Democratic
New York	2	Northeast	0.059	0.082	0.070	Democratic
North Dakota	26	Midwest	0.002	0.003	0.003	Republican
Ohio	26	Midwest	0.036	0.033	0.033	Republican
Oregon	4	West	0.013	0.012	0.015	Democratic
Rhode Island	17	Northeast	0.003	0.003	0.004	Democratic
South Carolina	69	South	0.016	0.011	0.014	Republican
South Dakota	12	Midwest	0.003	0.003	0.003	Republican
Tennessee	2	South	0.021	0.018	0.017	Republican

3. Results

In all, as shown in Figure 1 we find that a conservative estimate of son to daughter ratio is between 2:1 to 72:1 across the 36 states where we have data with a median of 12:1.



3.1 Profiling the Results with State Data

We now 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 busines establishments.

a) **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. Figure 2 shows the biggest gap is Midwest vs South with a median ratio of 24:1 vs 5:1 respectively.

b) **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 in Figure 3.

c) **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. The chart is on the appendix B.

d) **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 observed. The chart is on the appendix C.

e) **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. The ratio tends to be higher on Democratics states with a median of 17 vs a median of 11 for Republican states - see appendix D.

Figure 2: Estimate of son to daughter ratio by US Region

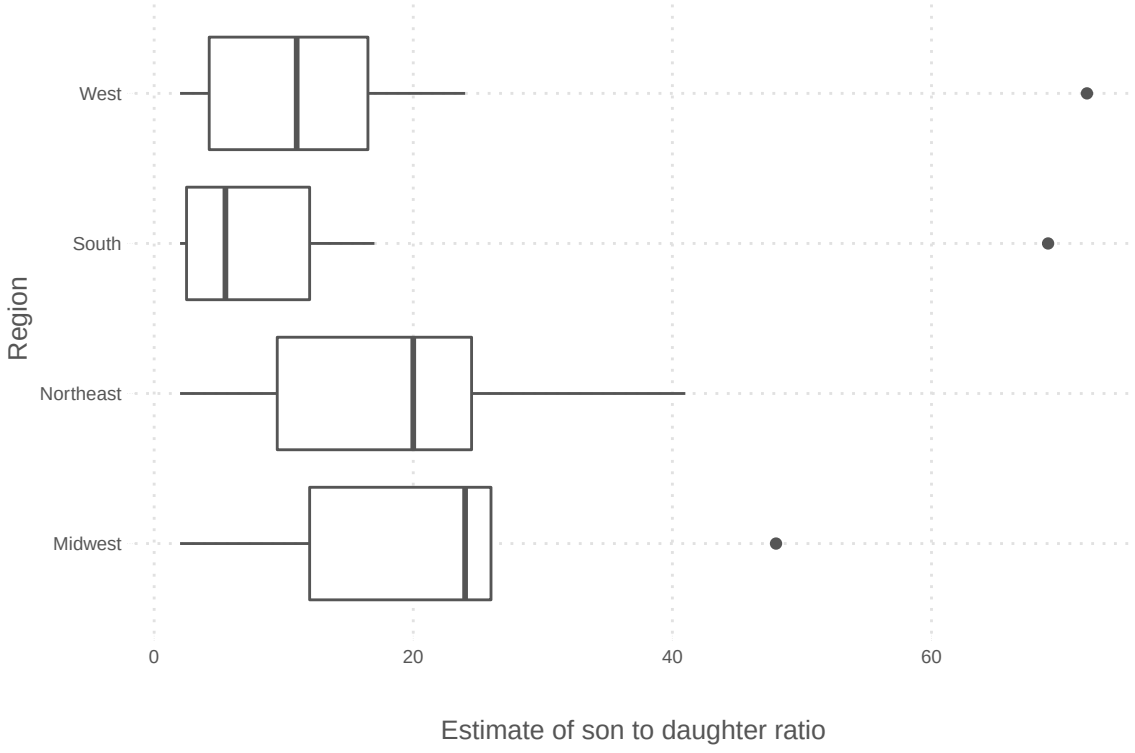
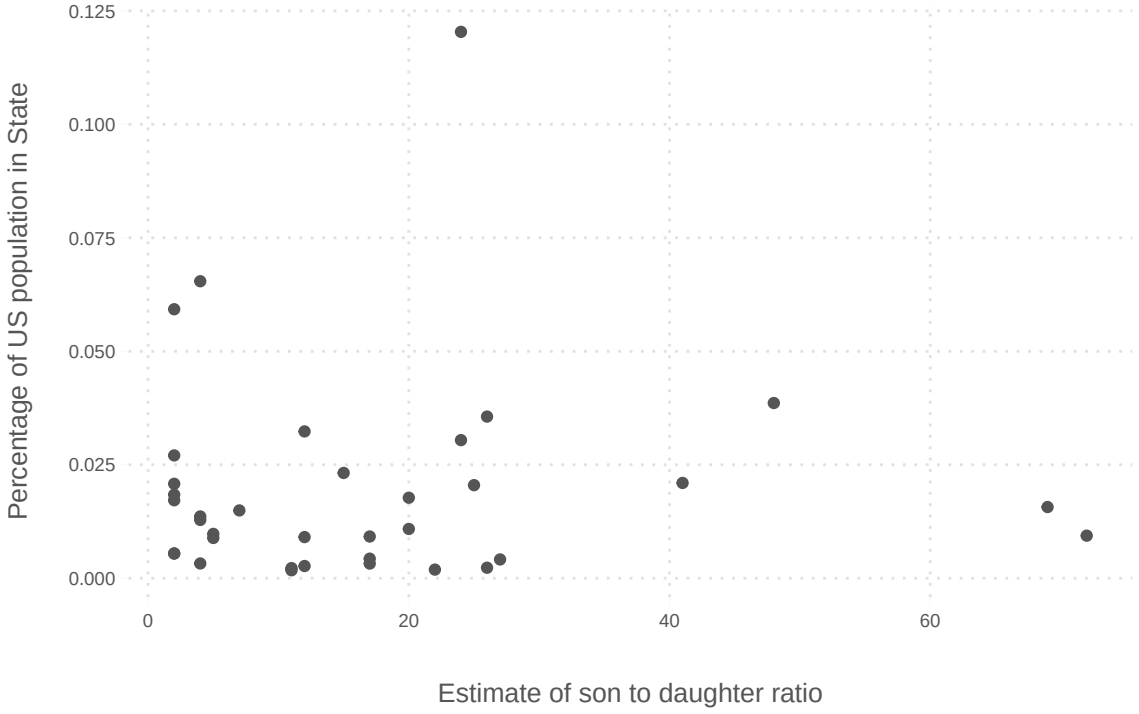


Figure 3: Estimate of son to daughter ratio vs % of US population per



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. Halpert, Chris. US census bureau regions and divisions.
   https://github.com/cphalpert/census-regions/

2. Kaushik, Saurav. Beginner's Guide on Web Scraping in R (using rvest) with hands-on example.
   https://bit.ly/2Gj0sF6

3. Kingl, Arvid. Web Scraping in R: rvest Tutorial.
   https://www.datacamp.com/community/tutorials/r-web-scraping-rvest

4. 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

5. United States Census Bureau. SUSB Historical Data.
   https://www.census.gov/data/tables/time-series/econ/susb/susb-historical.html

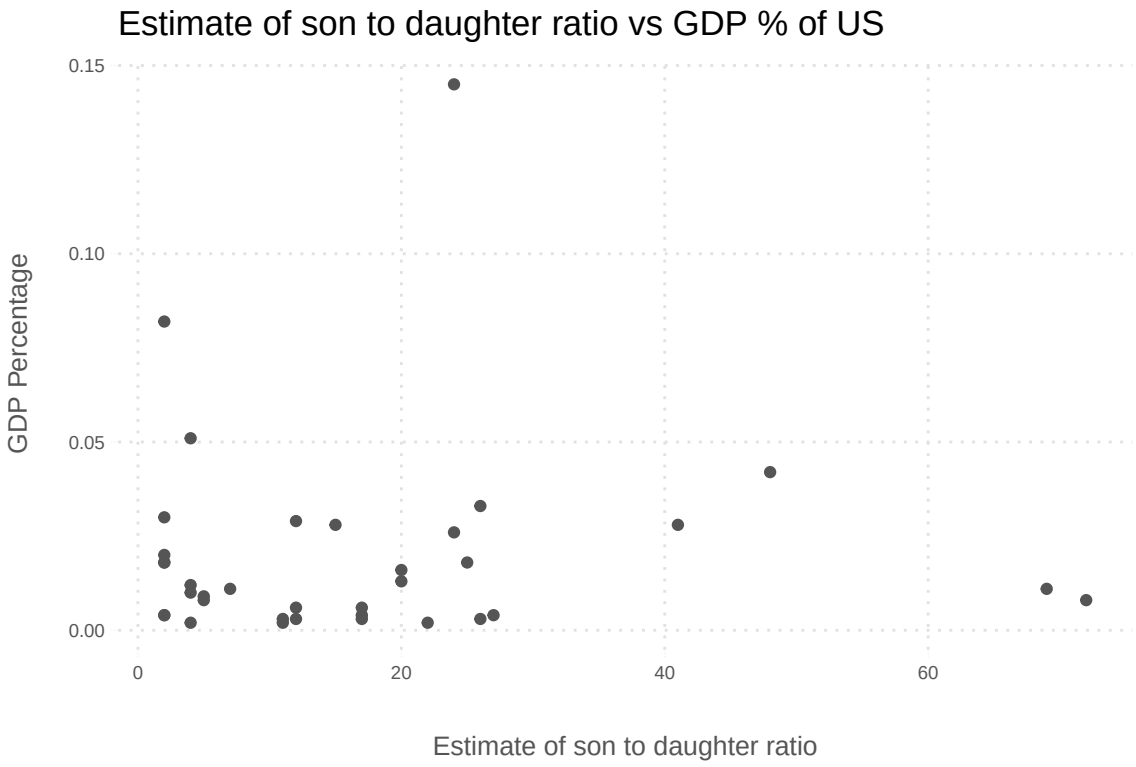
6. Wikipedia contributors. (2020, January 13). Political party strength in U.S. states. In Wikipedia, The Free Encyclopedia. Retrieved 03:11, January 20, 2020.
   https://bit.ly/37o6AYB

Appendix

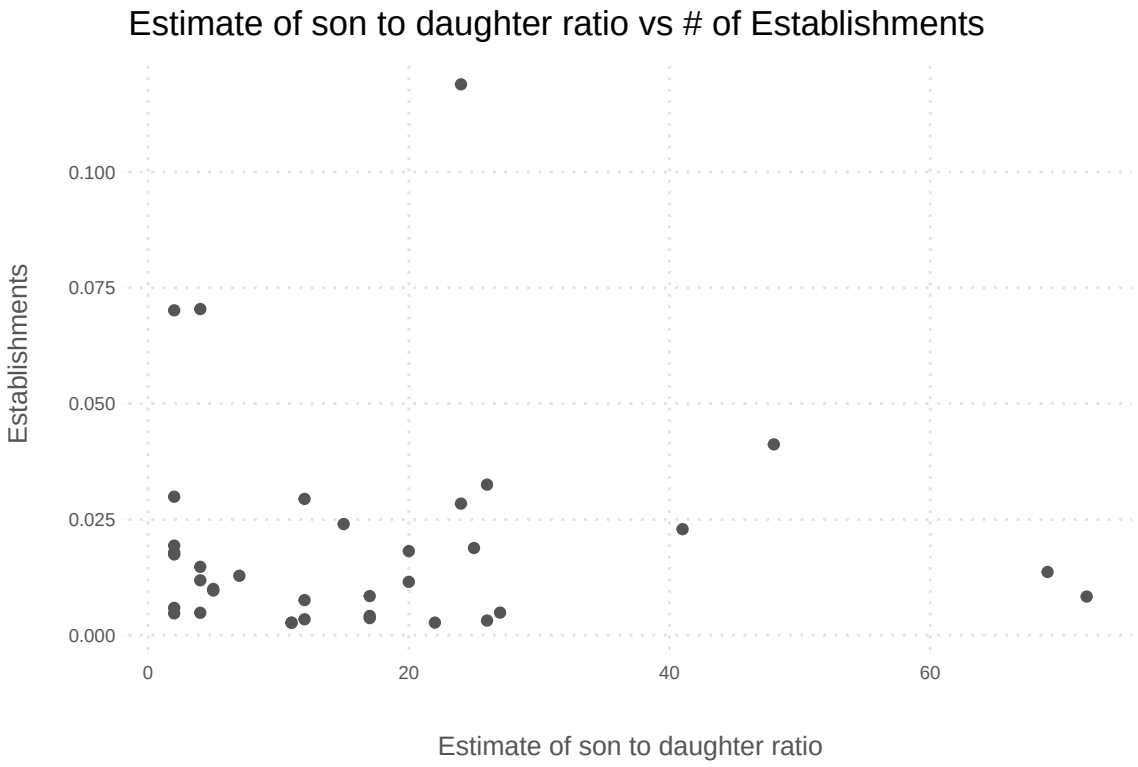
A - States with number of companies found with word son and daughter and ratio

State	Son/Daughter	# Companies 'son'	# Companies 'daughter'	Conservative?
Alabama	7	884	126	Yes
Alaska	11	246	22	
Arkansas	17	1482	87	
California	24	3609	150	Yes
Connecticut	20	875	43	
Florida	4	729	176	Yes
Georgia	12	6002	497	
Hawaii	17	1454	88	Yes
Idaho	2	60	39	
Illinois	48	2324	48	
Indiana	25	4928	195	
Kansas	5	75	14	Yes
Kentucky	4	66	16	Yes
Maryland	2	128	82	Yes
Massachusetts	41	5979	147	
Michigan	24	2265	93	
Minnesota	2	392	213	Yes
Mississippi	12	1918	165	
Montana	4	240	66	Yes
Nevada	72	1440	20	
New Hampshire	27	3203	119	
New Jersey	2	173	73	Yes
New York	2	1190	745	Yes
North Dakota	26	605	23	Yes
Ohio	26	2550	100	
Oregon	4	1000	227	Yes
Rhode Island	17	206	12	
South Carolina	69	4083	59	Yes
South Dakota	12	129	11	
Tennessee	2	203	132	Yes
Utah	5	81	16	Yes
Vermont	22	1361	63	
Washington	15	2424	161	
West Virginia	2	128	72	Yes
Wisconsin	20	845	43	
Wyoming	11	238	21	

B - Estimate of son to daughter ratio vs GDP % of US



C - Estimate of son to daughter ratio vs # of Establishments



D - Estimate of son to daughter ratio by mayority Political party (2016 elections)