Is Solar Energy A Political Issue

Tuckman, Jonah

Introduction:

My research question is whether or not the solar output of each state can be dictated by political afiliation. This question pivoted from the original based on a poor representation of overal data. My initial question that I hoped to explore was whether headquarter location of top 10 (in valuation) american oil companies affected state wide solar output. The hiccup I ran into when exploring this was that over 70% of these companies had their headquarters in Texas. This lead to a poor dataset as there was a low representation in other states. After I realized this I pivoted my question to explore whether solar output state wide can see a correlation between positive output and political voting patterns. This is an interesting question because the idea of solar energy is one that is globaly impactful for all states and people, independent of political affiliation. Positive aspects include limited reliance on oil (and thus foreign policy implications as the US needs to send less soldiers to wars fought over dictating oil rich regions), climate change reduction (the fossil fuel emissions would be limited if cars, trains, planes and more were running off of solar power rather than currently harmful tendencies), fiscal advantages (government stipends for those who opt into a switch into renewable energy) and many more. I found interest in this question after deciding to work for a company over the summer who works towards using Artificial Inteligence to advance the output and profitability of solar fields. Once understanding this aspect of solar energy I began to wonder whether the aspects still holding us back could be political.

Data:

Original Data Set: https://catalog.data.gov/dataset/energy-generation-by-state-and-technology-2009/resource/bb0da868-f498-4c47-b61f-da217897198f

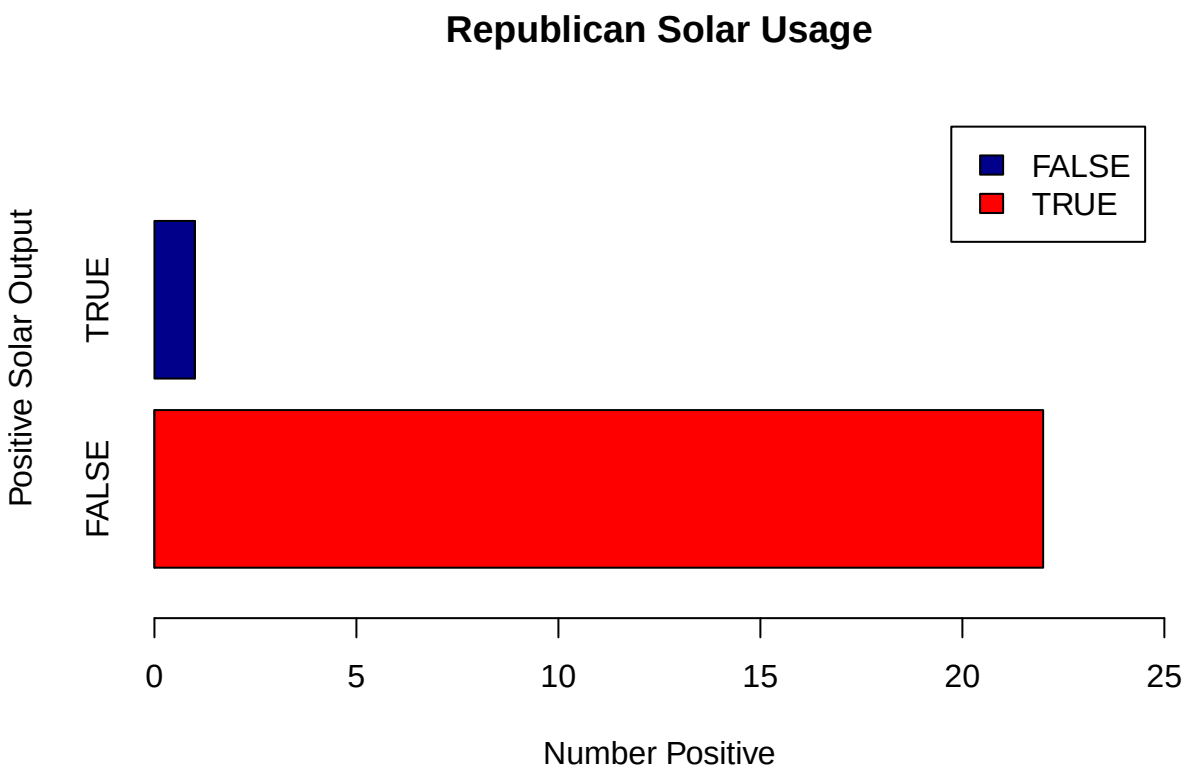
My initial data was from a 2009 study of Energy Generation by Source and State alongside the results from the 2008 general election by state. I initially sliced this Energy Generation csv in order to solely include State and Solar output. This dataset includes all types of energy output (Coal, petrolium, solar, wind, etc) but I decided to limit this study to solely solar output. A thought I had for a later extension of this project would be to broaden the exploration and create a boolean variable for all renewable energy output. This would encompass states who have a positive renewable output that does not include solar, thus making our exploration more accurate. Once I had a set of States and Solar output, I pulled data of voting records and merged this into the current dataset by state name. Now my data includes State, Voting History, and Solar output. From here I did a variety of things to be used later including splitting into seperate Democratic and Republican sets, adding a boolean variable of Solar Positivity to the original set and adding a boolean variable of Republican to the original set.

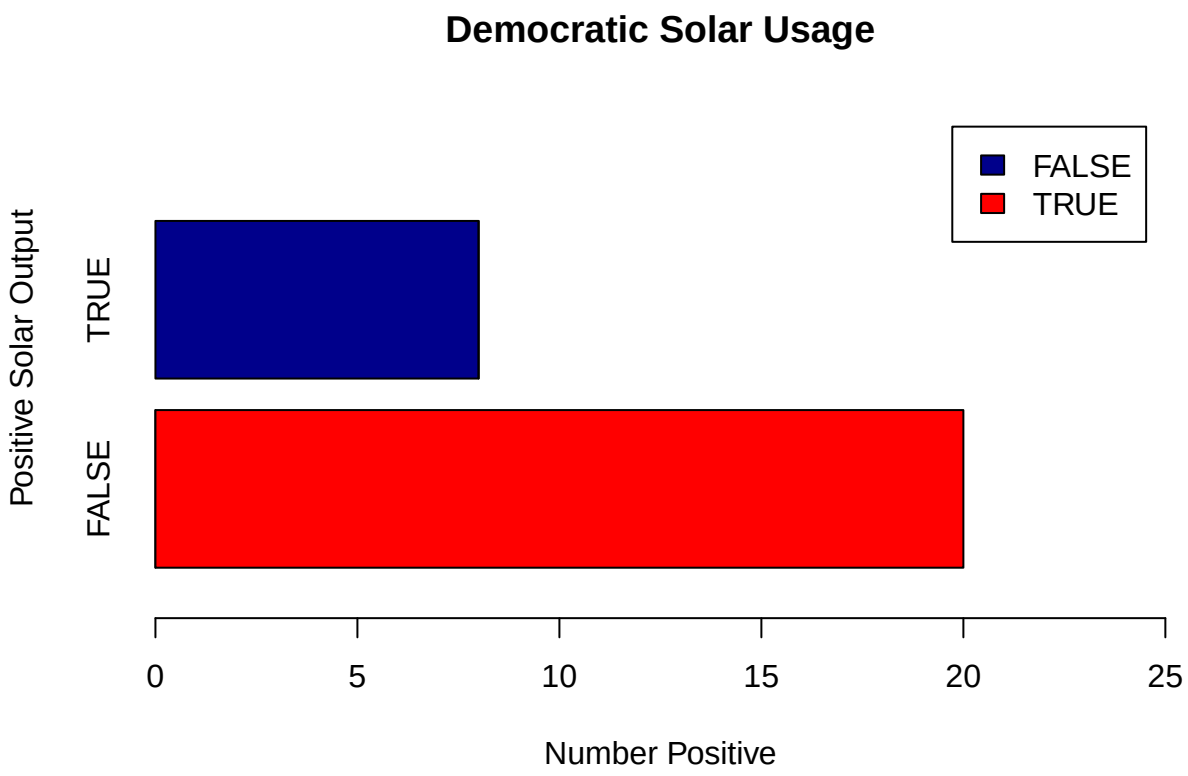
Exploratory data analysis:

```
## [1] "Democrat Positive Solar Output Percentage:  28.571429"
```

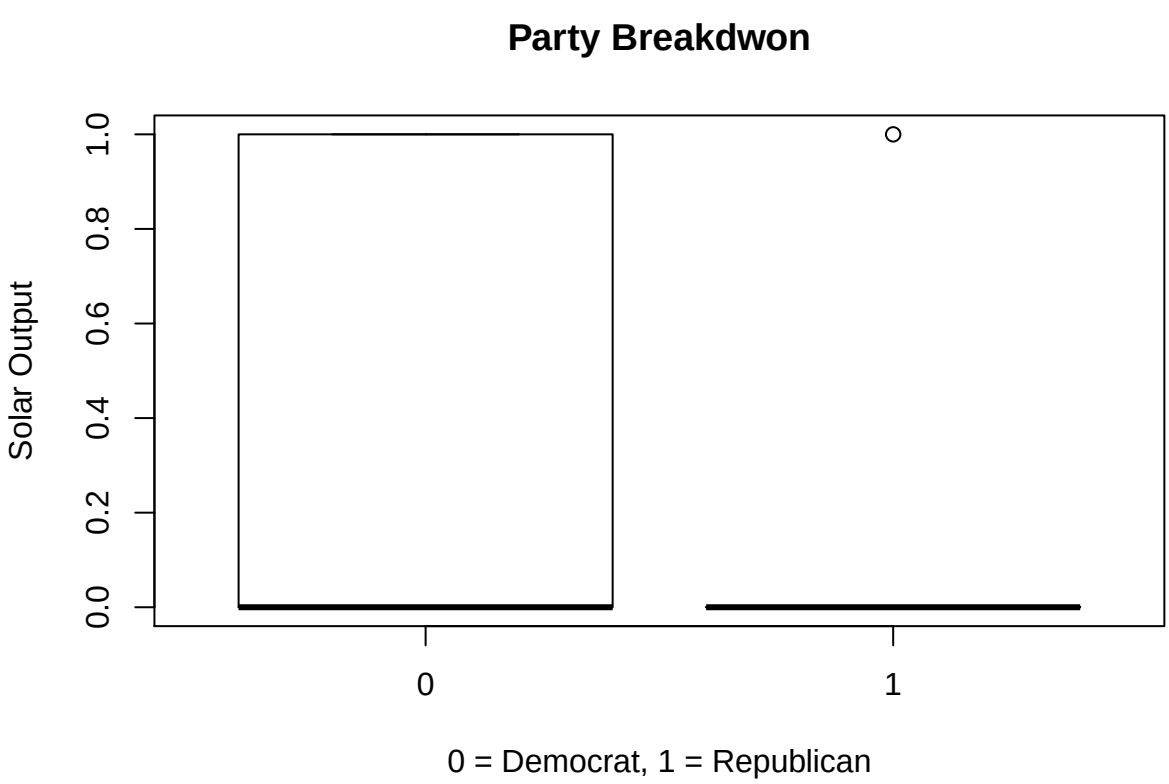
```
## [1] "Republican Positive Solar Output Percentage:  4.347826"
```

This shows the percentage of each political party affiliated state that has a positive solar output. 28.57% of the democratic states have a positive solar output and only 4.35% of the republican states have a positive solar output. This shows a large difference in output based on broken groups.


Republican Solar Usage
FALSE
TRUE
Positive Solar Output
TRUE
FALSE
0
5
10
15
20
25
Number Positive



In these plots we see the republican portion alongisde the democratic portion on the same horizontal range. This allows us to see the difference in internal to party breakdowns of output. We can see that within each party, a higher portion of democratic states have a positive solar output when compared to that of republican states.

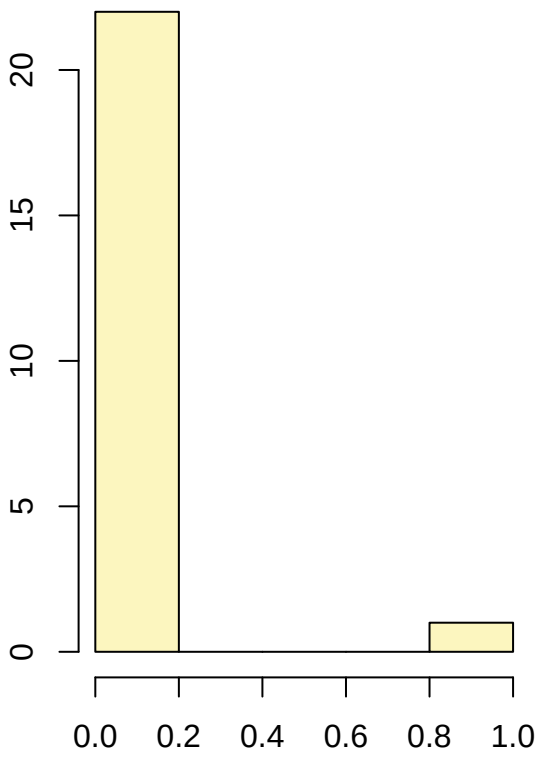


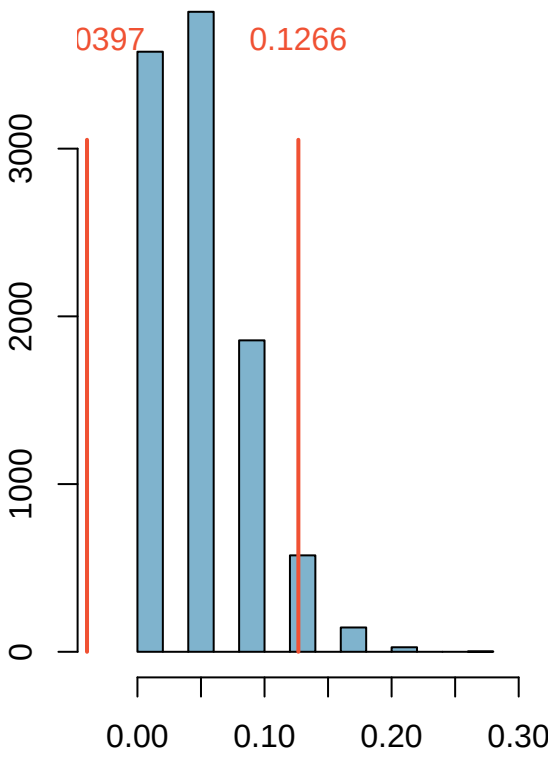
Although the above boxplot is not the prettiest representation it does share something important. As we know, in a box plot the lower quartile to the upper quartile (the box) show us a range of values that encompas the 25th to the 75th quartiles. Thus this includes the median as well as the middle 50% of values.
In the boxplot shown above we can see that within the republican portion of this, this middle 50% and the bottom 25% (the bottom 75%) is all at the value 0. The maximum is shown at 1 which indicates that there are indeed republican states that have a positive solar output, but this also indicates that at least 75% of these republican states have no solar output.
In the democrat boxplot, we see a median at 0, which means that less than 50% of the democrat states have positive solar output, but we see the upper quartile at 1. This indicates that more than 25% of democrat aligned states have a positive solar output.

```
## Warning: Ignoring success since y are numerical.
```

```
## Single mean
## Summary statistics:
```

```
## mean = 0.0435 ;  sd = 0.2085 ;  n = 23
```



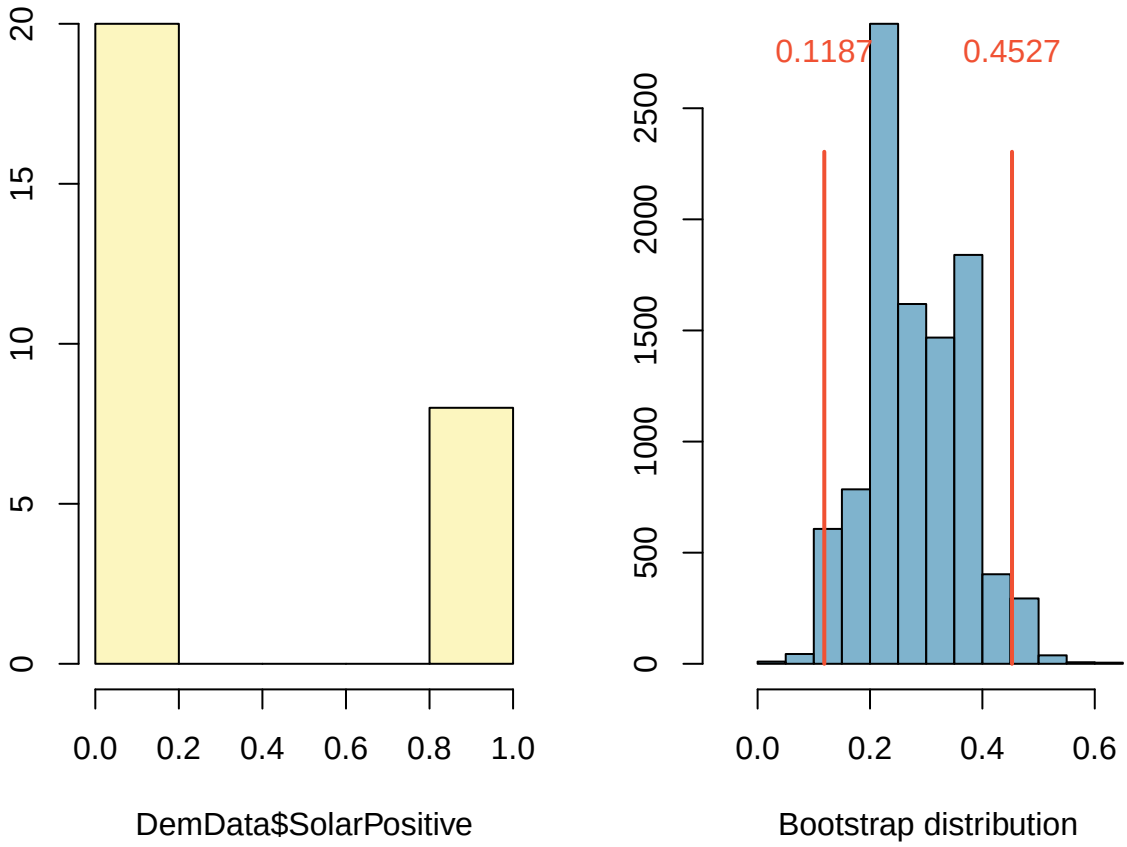



```
## Bootstrap method: Standard error; Boot. SE = 0.0424
## 95 % Bootstrap interval = ( -0.0397 , 0.1266 )

## Warning: Ignoring success since y are numerical.

## Single mean
## Summary statistics:

## mean = 0.2857 ;  sd = 0.46 ;  n = 28
```




```
## Bootstrap method: Standard error; Boot. SE = 0.0852
## 95 % Bootstrap interval = ( 0.1187 , 0.4527 )
```

```
## [1] "Condifence intervals: "
```

```
## [1] "Republican: 95 % Bootstrap interval = ( -0.0422 , 0.1331 )"
```

```
## [1] "Democrat: 95 % Bootstrap interval = ( 0.1195 , 0.4519 )"
```

Printing bootstrap confidence intervals because the knit is having an issue with it right now. Code works but is commented out because it will not convert to knit.
In this interval we see something very important. In the republican 95% confidence bootstrap interval the confidence range has a negative minumum and includes 0 in the range. Clearly our data could not be negative considering they are boolean responses, but the fact that 0 is included is very important as it shows that we are confident that 0 is in the range of possible values.
In the democrat data this is not the case, 0 is not in the range thus we are 95% confident that the response is not that there are 0 states with positive output. Although the range also does not include 1, the data not including 0 means that the distribution is high enough that the posibility on none containing positive solar output is not included.

```
##
##  Welch Two Sample t-test
##
## data:  statsData$SolarPositive by statsData$Republican
```

```
## t = 2.492, df = 39.186, p-value = 0.01704
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.04564886 0.43882319
## sample estimates:
## mean in group 0 mean in group 1
##      0.28571429      0.04347826
```

Conditions that need to be met to perform this T-Test. 1. The data is a random sample from the population. -> In this case the data is taken from the general population. The Solar output and voting records are not randomly collected but are both valid in their sources of validity. 2. The sampling distribution is nearly normal -> It is assumed that the sampling here is nearly normal. This can be seen in the prior tests bootstrapping plots (which have not been shown based on importing error). 3. Individual observations are independent. -> in this case I am under the assumption that states do not coordinate voting patterns (strong assumption) and that they do not coordinate solar output (medium strength assumption).

This T Test is built by looking at the solar positivite and the Rebulcican boolean variables. The two means we are testing is the mean which indicates the probability that a solar positive value of 1 is equal to a republican value of one, in simpler terms if a state has a positive solar output, is this state republican aligned?
I ran this this t - test to determine if we can conclude that there is significant evidence of a factor resulting in this data, or if this output discrepincy is purely due to chance. The data resulted in a t value of 2.492 and a p-value of 0.0174. This p-value is less than 0.05 thus we are able to reject the null. This means that we are able to understand that we would not be able to obtain this data if there was not some influence based on political affiliation.

Inference:

Goal: To simulate the Republican portion of the data (23 states) and use null and alternative probabilities and count the number of solar positive occurences.

Null Hypothesis: H0 <- The probability that a state in our republican simulation will have a positive solar output is 9/51 (the overall probability of all states excluding political factors). This would mean that we could reasonably get this data with no information concerning the states political history.
Alternative Hypothesis: HA <- The probability that a state has a positive solar output is 4.25% which is the probability that a republican state would have positive solar output.
When recreating vectors the size of the total republican states, we are simulating our data vs the null hypothesis.

```
##
##  Welch Two Sample t-test
##
## data:  newFrame$nullValsVec and newFrame$testValsVec
## t = 48.382, df = 1501.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.039559 3.296441
## sample estimates:
## mean of x mean of y
##     4.132     0.964
```

After simulating two vectors 1000 times with the null probability (the probability a republican state would have positive solar output independent of political alignment) and the alternative probability (the probability that a republican state would have positive solar output given it's political view), I ran a T-Test on these

two vectors to find whether or not the null can be accepted.
After running this test I found a p-value of 2.2e-16 which is incredibly small. This would lead me to reject the null hypothesis and state that we are able to understand that this data could not be found solely based on the probability of positve solar output state wide. It is not enough to know the number of states who have solar output and use this probability in recreating the data, we also have to understand more factors, which in this case, is the political alignment.
When recreating the republican vector in my simulation, when using the probability of a republican state to have solar output (4.35%), the simulation found a mean of 1.054 states with positive solar output, which is close to the true value in the data of 1. When using the probability that any state has positive solar output, the simulation found a mean of 4.090 which is much higher. This is based on the overal nation wide probability of 17% chance of finding a state with positive solar output.

Thus a further look at this data indicates that there is more to the eye than just nation wide probability of a state having positive solar output and thus we reject the null.

Conclusion:

Write a brief summary of your findings without repeating your statements from earlier. Also include a discussion of what you have learned about your research question and the data you collected. You should also acknowledge limitations of your study and include ideas for possible future research.

When looking at this data, I was able to conclude that political factors do indeed affect the amount of solar output that a state sees. I can even go further and say that, based on probabilities of states containing a solar positive energy output within the groups, democrat leaning states are more likely to have this positive ouput than republican leaning states.