Introduction

This report explores the relationship between voter turnout and three different explanatory variables, including a county's number of registered voters, median income, and bachelor’s degree or higher completion status. The purpose of this statistical analysis is to determine whether there is a correlation between these variables to see if one of these variables could be used to predict voter turnout. This would be valuable information for candidates and their campaign teams, as it would allow them to identify useful areas to focus their campaigning efforts.

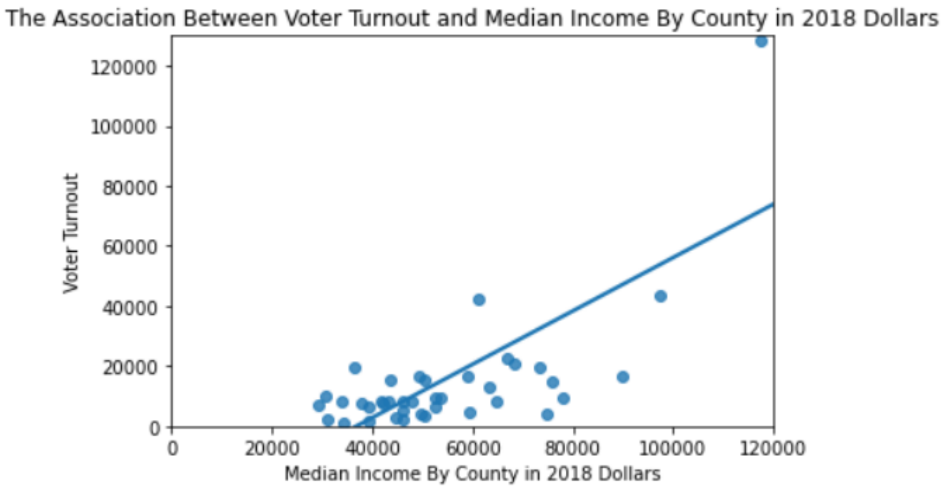
The response variable for this statistical analysis is voter turnout. Voter turnout is the amount of individuals per county that voted in a given election. The three initial explanatory variables that we chose were “Number of Registered Voters”, “Median Income by County in 2018 Dollars”, and “Bachelor’s Degree or High Percentage by County (%)”. These explanatory variables were chosen because we believed that they would have a significant impact on the response variable - voter turnout, as well as a relatively easy means of measurement compared to the other variables.

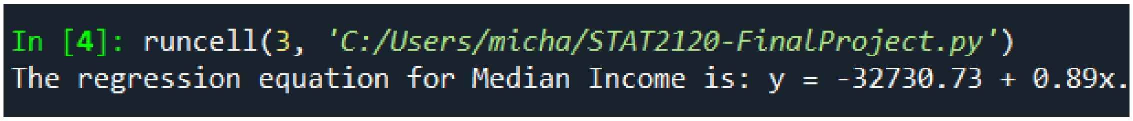
The use of “Number of Registered Voters” allows for the comparison between the amount of people that are actually registered to vote and the voter turnout for each county. We know that if there are more registered voters in a country, there will be more people able to vote, causing voter turnout to increase. Similarly, the use of “Median Income by County in 2018 Dollars” allows for the comparison between the level of income per county in 2018 and their respective voter turnouts. We believe that as the income of a household increases, the voters will be more able to make their way to the polls as well as have information that allows them to register to vote. This will cause voter turnout to increase, which indicates a possible positive relationship between the two variables. Finally, the use of “Bachelor’s Degree or High Percentage by County (%)” allows for the comparison between the proportion of individuals who have obtained their bachelor’s degree in each county and the voter turnout for said counties. We assumed that this variable would affect voter turnout in a positive way, since a higher level of education will allow people to know more about political issues and feel obligated to involve themselves in resolving these issues. The information used in this report can be obtained from publicly-available governmental sources on demographics such as census reports.

Analysis with simple linear regression

For our analysis, we first created visual summaries of the data and calculated the regression line to observe the relationship between the explanatory and response variables in order to choose which variable would be most apt for predicting voter turnout. We used both scatter plots with a regression line and the residual plots for each chosen explanatory variable to make this determination. It should be noted that the data from five of the counties was omitted when we were constructing the plots, as the dataset did not include voter turnout for these counties. We use a least squares regression model for our analysis: $y_i = b_0 + b_1 x + e_i$ for our single simple random sample including measurements for n cases in our sample denoted (x_i, y_i). In this model, b_0 is the estimated intercept, $b_1$is the estimated slope, and e_i is the estimated residual for a single case *i.* Using this model, we will conclude our analysis by returning to the five initially omitted counties and predicting their voter turnout based on the number of registered voters in the county.

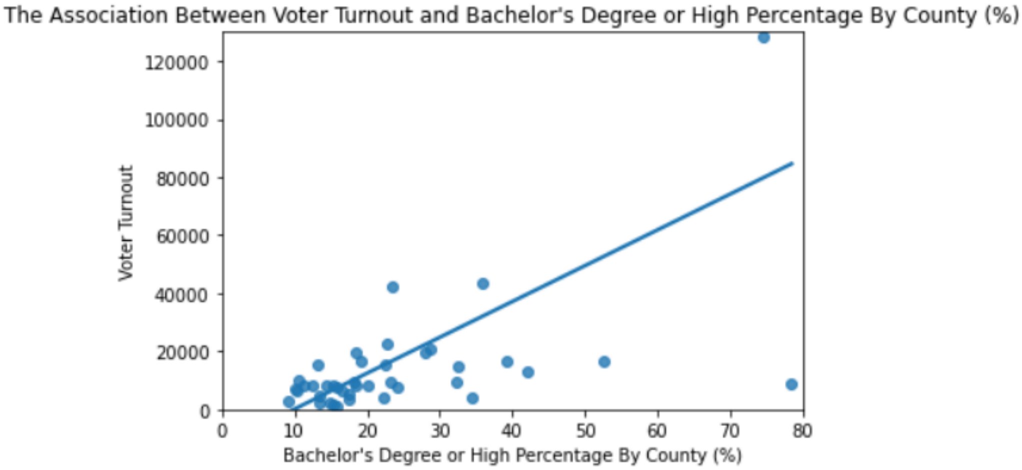
We calculated the least squares regression line of the form $\hat{y} = b_0 + b_1 x$ for all 3 variables. The least-squares regression line for voter turnout and median income by county shows a positive, weak, linear relationship with a few possible outliers for both variables. Its corresponding regression equation is $\hat{y}$ = 032730.73 + 0.89x.

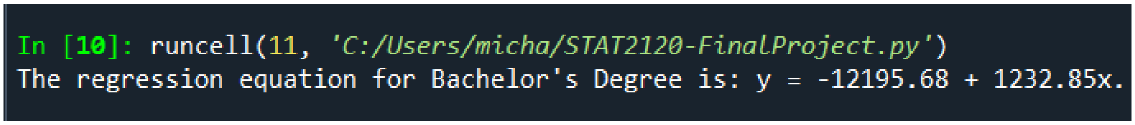




```
In [4]: runcell(3, 'C:/Users/micha/STAT2120-FinalProject.py')
The regression equation for Median Income is: y = -32730.73 + 0.89x.
```

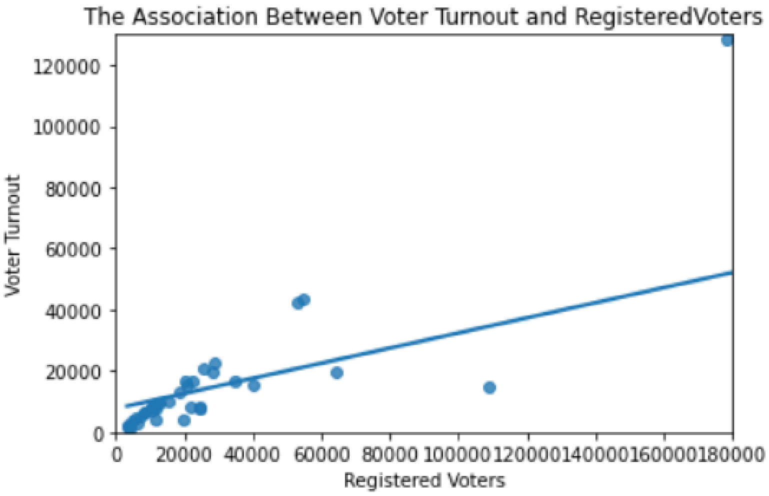
The least-squares regression line for voter turnout and Bachelor’s degree or higher shows a positive, weak, linear relationship with a few possible outliers for both variables. Its corresponding regression equation is $\hat{y}$ = -12,195.68 + 1,232.85x.

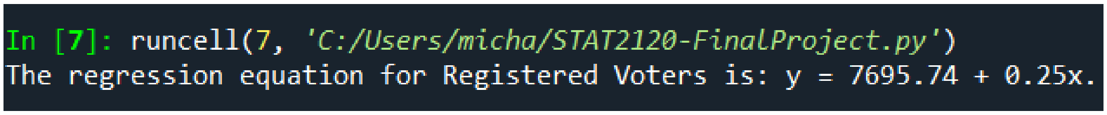




```
In [10]: runcell(11, 'C:/Users/micha/STAT2120-FinalProject.py')
The regression equation for Bachelor's Degree is: y = -12195.68 + 1232.85x.
```

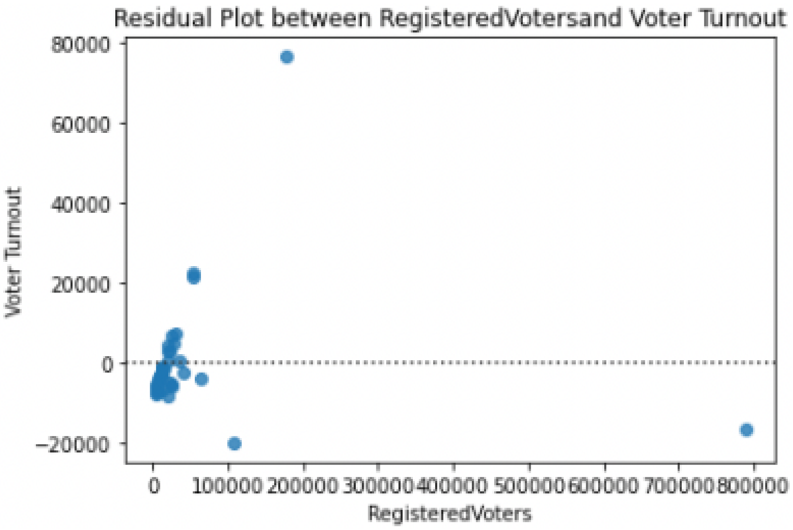
We selected the variable ‘registered voters’ as our variable of focus. Thus, for the regression line, b_0 is the y-intercept, b_1 is the slope, x is the number of registered voters in a county, and $\hat{y}$ is the predicted amount of voter turnout in that county. The corresponding regression equation is $\hat{y}$ = 7,695.74 + 0.25x. The slope of the regression line indicates that for every 1 unit increase in registered voters, there is a 0.25 unit increase for voter turnout. This suggests that a 4 person increase in the number of registered voters correlates to 1 more person actually voting. The y-intercept indicates that if there were 0 registered voters in a county, there would still be 7,696 voters in an election; however, this is not useful or relevant information, as it suggests an impossible situation: you cannot vote if you have not registered, so it is considered an extrapolation data point and largely disregarded for the analysis that follows. The least-squares regression line for voter turnout and registered voters shows a positive, moderately strong, linear relationship with only a couple possible outliers for both variables.





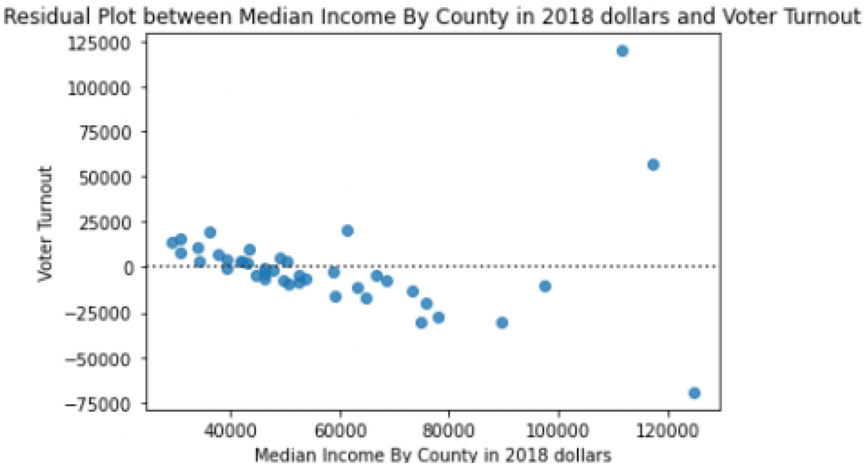
```
In [7]: runcell(7, 'C:/Users/micha/STAT2120-FinalProject.py')
The regression equation for Registered Voters is: y = 7695.74 + 0.25x.
```

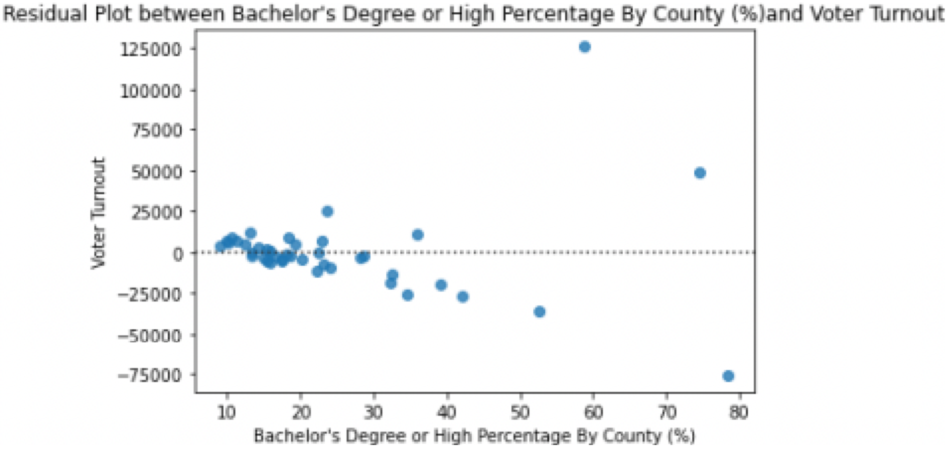
To address the existence of an apparent outlier on all of these scatter plots, we noted that this point represents a county that has a voter turnout of 128,167 with 178,427 registered voters, a median income of $117,374 in 2018 dollars, and 74.6% of its population has a bachelor's degree. Along with this, there is another apparent outlier on the scatter plot titled "The Association Between Voter Turnout and Bachelor's Degree or High Percentage By County (%)." This point represents a county that has a voter turnout of 8,886 and 78.5% of its population has a bachelor's degree, which is unusually high. We considered removing these outliers but ultimately, chose to keep them in our dataset. We found that even with the outliers, the regression line still closely follows the form of the points, so we concluded that the effect of the outliers was not great enough to pose significant issues. Given this, we deemed it better to keep the outliers than to make further manipulations to the dataset that were not clearly justifiable.



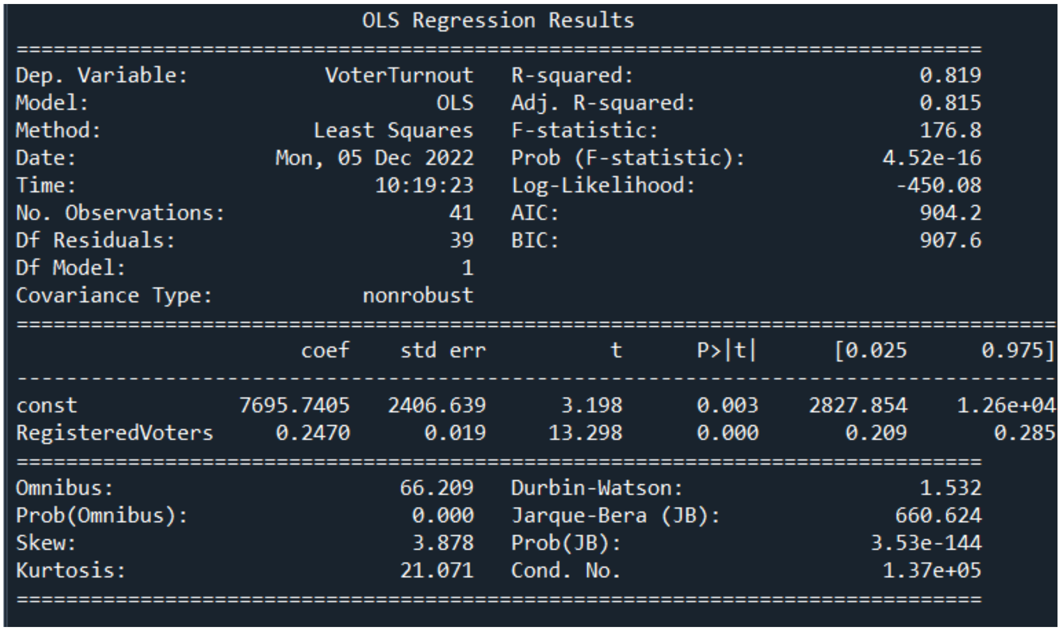
Based on the residual plots, we concluded that the variable ‘registered voters’ had the strongest correlation to voter turnout, so it would be the best explanatory variable for predicting voter turnout. We chose this variable because it best met the assumptions necessary for simple linear regression: this meant that the regression line shows independence, linearity, constant spread, and normality for each value of the explanatory variable. In the residual plot for registered votes, the points are clustered closely to zero on the residual plot, with fewer points farther away from zero and a similar number of points one either size of zero. This uniform variation of residuals centered on zero met the requirement, indicating linearity, constant spread, and normality.

As evidenced by the following residual plots, the other potential variables did not meet these requirements as desired. The residual plots for both median income and bachelors degree attainment both show a downward trend, suggesting that the model routinely tends to overestimate the voter turnout at higher income levels/bachelor’s degree attainment percentages. Thus, choosing either of these variables would be undesirable, severely limiting the predictive power of our model and our ability to provide a useful tool to campaign organizers.





As further support for our decision to take 'registered voters' as our main variable, we assessed its coefficient of determination, or R-squared value. This measurement tells us how much of the variation in the response variable can be explained by the regression line. For registered voters, this value was 0.819, which not only signifies a strong relationship, but is significantly higher than that of the other potential variables (median income at around 0.38 and bachelor's degree at around 0.357). Though we determined that using registered voters is optimal amongst the three options, we note that it does not provide a perfect representation of the population.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           VoterTurnout   R-squared:                       0.819
Model:                            OLS   Adj. R-squared:                  0.815
Method:                 Least Squares   F-statistic:                     176.8
Date:                Mon, 05 Dec 2022   Prob (F-statistic):           4.52e-16
Time:                        10:19:23   Log-Likelihood:                -450.08
No. Observations:                  41   AIC:                             904.2
Df Residuals:                      39   BIC:                             907.6
Df Model:                           1
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const            7695.7405   2406.639      3.198      0.003    2827.854    1.26e+04
RegisteredVoters    0.2470      0.019     13.298      0.000       0.209       0.285
==============================================================================
Omnibus:                       66.209   Durbin-Watson:                   1.532
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              660.624
Skew:                           3.878   Prob(JB):                    3.53e-144
Kurtosis:                      21.071   Cond. No.                     1.37e+05
==============================================================================
```

Regression Model Output Summary

Having selected our variable of choice, we proceeded by performing a hypothesis test for our regression model based on “Number of Registered Voters” to formally assess the relationship of the two variables. The null hypothesis in this t-test for slope is that there is no relationship between the two variables. In other words, β_1 (the population slope) is equal to 0. The alternative hypothesis is that there is a relationship between the two variables: $\beta_1 \neq 0$.We were able to do this test because our simple random sample’s size was sufficiently large and it fit our assumptions of linearity and constant spread, so we could consider the data approximately normal and robust. Our test statistic was t=13.298, based on the formula $t = (b_1 - \beta_{1\_0})/SE_{b1}$. The p-value for this test, $p = 2P(T \leq -|13.298|)$, was found to be 0.000 (an extremely small decimal number that rounds to zero at 4 significant figures). The small p-value indicates that there is significant statistical evidence to reject the null hypothesis. Therefore, we concluded that based on the data in our sample, there is statistical evidence to suggest that there is a correlation between the number of registered voters in a county and voter turnout.

We used our model to estimate voter turnout in the counties for which this data was missing (referred to below as counties A-E). We also established a range of possible values that are plausible for the mean voter turnout in each county based on the number of registered voters that take into account the margin of error for the estimate we have provided.

County	Registered voters	Estimate	Confidence Interval
County A	84,140	28,477	(23,639 , 33,316)
County B	26,078	14,137	(9,503 , 18,770)
County C	236,944	66,219	(57,616 , 74,821)
County D	1,851	8,153	(3,308 , 12,998)
County E	15,218	11,454	(6,746 , 16,162)

Based on the mean confidence intervals obtained for the voter turnout from the least-square regression line between registered voters and Voter turnout, it should be that in 95% of samples, if the process were repeated, the true population number of for voter turnout for counties with the given number of registered voters would will fall within the ranges listed above. It can be seen that the estimates given could easily be off by a few thousand people. Nevertheless, these ranges are small enough such that they are still useful in most cases for getting an idea of what the turnout could be. It’s worth nothing that for smaller numbers of registered voters (<2000) our model breaks down slightly, and will tend to overestimate the voter turnout that can be gained from this population. However, when dealing with larger numbers of registered voters, our model outputs plausible estimates. This trend is satisfactory as campaigns will wish to focus more on counties with these larger numbers, making the accuracy of these estimates especially important.

Conclusion

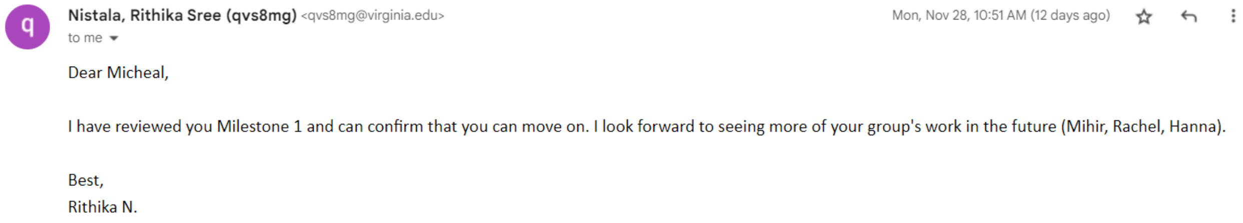
In our analysis, we assessed the relationship of registered voters, degree attainment, and median income to voter turnout, and determined that the best predictor of voter turnout is the variable exploring the number of registered voters in each county. Our analyses show that while this variable is unlikely to perfectly predict the voter turnout for a given county, there is still a significant relationship between registered voters and voter turnout, so the model can give a decent idea of what voter turnout will look like within a few thousand. We advise that our model is likely to be most accurate for counties with a moderate amount of registered voters. We caution that attempting to use this model to predict turnout for extreme highs or lows of registered voters may produce inaccurate results. Though it's important to take this into account, we note that our statistical inference confirmed the correlation between the variables. Therefore, the approximate predictions given by our model can nonetheless provide valuable information that can be used by governmental candidates and their election campaign teams to inform where they ought to direct their campaign efforts: focusing on counties with larger numbers of registered voters or even encouraging and facilitating voter registration to attempt to increase voter turnout numbers.

Reflection

If we were to complete this project again, we would explore additional and different variables for the sake of being more accurate with our conclusion and to understand the relationships that may better predict the response variable. It may be beneficial to create a regression model for more than just 3 variables, as this tool allows us to judge the predictive potential of the variables, so we might have found an even better predictor of voter turnout that we overlooked in this report due to our method of choosing our variables before analyzing the data itself.

Additionally, the use of different, or even multiple, variables would be useful in finding what variables correlate with voter turnout. We think that we could have created a model capable of more accurate predictions using multiple linear regressions, but this was ultimately outside the scope of this project. Again outside the scope of this project, but it may have also been worthwhile to explore other potential explanatory variables outside of what was included in the given dataset. In our initial phase we had brainstormed other factors including age, race, and past voting history; if we were seriously trying to accurately predict voter turnout, we would likely want to consider these variables, as they likely do correlate with voter turnout. These changes would ultimately assist in seeing where candidates should campaign and bolster previous conclusions about which counties would be best.

Appendix A



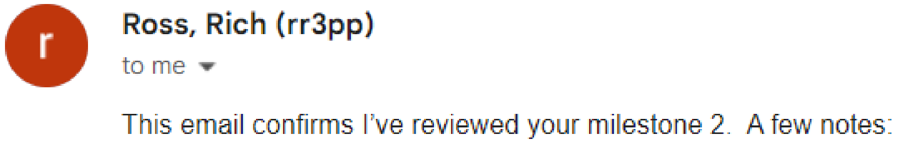
Nistala, Rithika Sree (qvs8mg) <qvs8mg@virginia.edu>
to me

Mon, Nov 28, 10:51 AM (12 days ago)

Dear Micheal,

I have reviewed you Milestone 1 and can confirm that you can move on. I look forward to seeing more of your group's work in the future (Mihir, Rachel, Hanna).

Best,
Rithika N.



Ross, Rich (rr3pp)
to me

This email confirms I've reviewed your milestone 2. A few notes:

Appendix B

Below is the complete set of data we used to perform our analysis. Each row represents the values for each variable pertaining to a given county.

VoterTurnout	RegisteredVoters	Median Income By County in 2018 dollars	Bachelor's Degree or High Percentage By County (%)
8468	24624	43,210	18.5
8178	10961	47,794	15.4
16356	22245	49,170	19.2
128167	178427	117,374	74.6
42132	52752	61,305	23.5

2505	3307	46,137	14.9
3490	4569	50,511	17.4
21005	25675	68,410	28.7
7826	11714	37,904	24.1
8034	10943	41,927	14.3
10040	15268	30,806	10.6
8141	10752	46,261	11.3
976	4263	34,273	15.9
15269	39877	50,258	22.5
8322	21743	64,715	20.1
15699	20701	43,532	13.1
4442	5651	59,192	13.3
6177	8393	39,212	10.2
16799	34964	58,933	52.6
14604	108695	75,790	32.5
VoterTurnout	RegisteredVoters	Median Income By County in 2018 dollars	Bachelor's Degree or High Percentage By County (%)
9374	11688	77,936	32.2
9176	12755	53,716	23.2
1493	3753	39,432	15.3
5337	7198	46,221	17.5
19568	28397	36,301	18.4
7339	10277	29,226	10
2399	3873	30,857	13.4
6192	8106	52,681	16.5
186244	789950	111,574	58.7

8886	11128	124,796	78.5
43552	54372	97,469	35.9
3886	11544	49,729	22.3
4177	19840	74,931	34.5
19636	64271	73,250	28.1
12919	18548	63,274	42.1
9200	11984	52,478	18.2
22326	28978	66,701	22.8
16847	19961	89,741	39.2
8150	10766	33,969	12.5
2576	6407	44,534	9
7756	24589	42,289	15.9
-	84140	88,652	39.2
-	26078	43,893	36
-	236944	68,572	42.9
VoterTurnout	RegisteredVoters	Median Income By County in 2018 dollars	Bachelor's Degree or High Percentage By County (%)
-	1851	46,147	23.1
-	15218	40,497	13.5

*The final 5 rows of the dataset represent the counties for which no data on voter turnout was available, hence the missing values - therefore, these values did not play a role in our analysis.