On our honor, we did not give nor receive aid on this assignment.

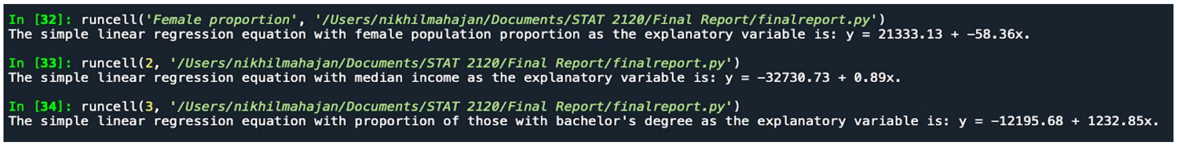
Nikhil Mahajan (nrm2rqx), Raghav Dalmia (rsd6amu), Taylor Musa (tbm9gss), Jack Lesemann (jwl4vg)

Voter Turnout Final Report

Introduction: The response variable in this data set will be the official voter turnout levels in localities of Virginia, found in column A of the dataset. This variable describes the resulting total number of voters who voted in a given county or locality during the 2020 election season. The number of votes for localities in Virginia can be found in our sources. Note that in our dataset shown in Appendix B, the last five rows of data are missing the voter turnout numbers, and thus for the purposes of our analysis, have excluded those entire five rows when analyzing our data seeing as that there is no response variable to report.

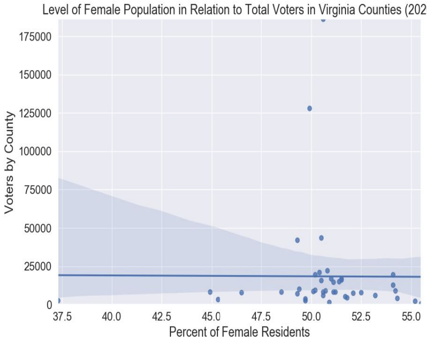
Based on our prior knowledge from government classes and outside research from our sources (2-4), we knew some of these explanatory variables were more likely to be correlated with voter turnout numbers than others, at least on the national level in the United States. Namely, female population proportion, median income, and the proportion of those with a bachelor's degree or higher in a given county struck out as viable candidates for explanatory variables, as these were known to be positively correlated with voter turnout. To be more specific, the female population proportion represents the proportion of the population of a given locality who is female, expressed as a percentage. Median income represents the median income of a household in a given locality. Lastly, the proportion of those with a bachelor's degree or higher is the proportion of the population in a given county with a higher education degree, expressed as a percentage. These three variables can be found in Appendix B with each of the mentioned explanatory variables corresponding with the logical matching column of the dataset.

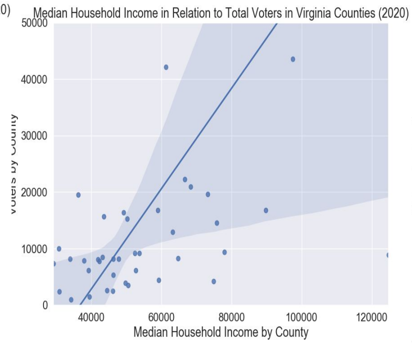
Analysis with SLR: Using the female population, median household income, and population of people with a Bachelor's degree or higher as the explanatory variables and the voter turnout levels as our response variable, we found the following models. For the female proportion explanatory variable, the model is $y_i = \beta_0 + \beta_f x_i + \varepsilon_i$. In context, β_0 represents the voter turnout when the proportion of females in a county is 0, while β_f represents how much the voter turnout change for a single percentage increase in the proportion of females in a county. Similarly, for the median income explanatory variable, the model is $y_i = \beta_0 + \beta_m x_i + \varepsilon_i$, where β_0 represents the voter count if the median income were 0, and β_m represents how much the voter turnout would change for a single dollar increase in median income. Lastly, the model for the proportion of those with a higher education degree is $y_i = \beta_0 + \beta_h x_i + \varepsilon_i$, where β_0 represents the voter turnout if the proportion of those with a higher education degree were 0, while β_h represents the amount of change in voter turnout if the percentage of those with a higher education degree increased by 1%. For the three models described, y_i describes the voter turnout, x_i describes the value of the explanatory variable, and ε_i describes the deviation in voter turnout from what it actually is (all of the i^{th} county). Using these models and our dataset, we can produce the following least squares regression equations.

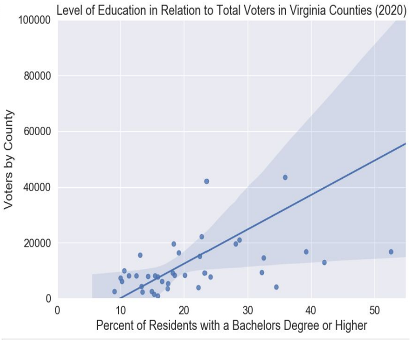


```
In [32]: runcell('Female proportion', '/Users/nikhilmahajan/Documents/STAT 2120/Final Report/finalreport.py')
The simple linear regression equation with female population proportion as the explanatory variable is: y = 21333.13 + -58.36x.

In [33]: runcell(2, '/Users/nikhilmahajan/Documents/STAT 2120/Final Report/finalreport.py')
The simple linear regression equation with median income as the explanatory variable is: y = -32730.73 + 0.89x.

In [34]: runcell(3, '/Users/nikhilmahajan/Documents/STAT 2120/Final Report/finalreport.py')
The simple linear regression equation with proportion of those with bachelor's degree as the explanatory variable is: y = -12195.68 + 1232.85x.
```

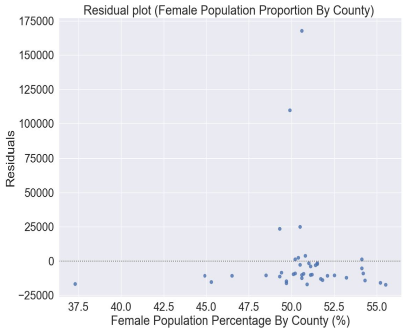
Looking at the regression plots using these SLR equations, it is apparent that the median income and education levels in relation to voter turnout have positive slopes while the female population graph contains a slightly negative slope. These slopes are also evident in the earlier regression equations.

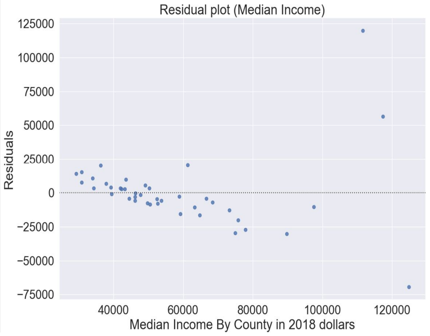


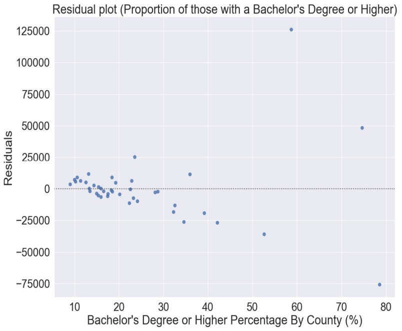




Moving to the residual plots, which show the distances of the predicted voter turnout levels from the observed levels, we assume that the normality assumption for these three variables should be satisfied as there are 41 different counties and the SLR model being generally robust against deviations from the Normal distribution.







When evaluating the residual plots, it seems that the regression assumptions for the models using female population proportion and median income as the explanatory variables (separately) are not met. Namely, the independent and constant variance assumption is not met, as the residuals are clearly not Normally distributed around 0 since there are significantly more negative residuals than positive residuals, with the negative residuals having notably more variance than the positive residuals. For the residual plot of median income as the explanatory variable, there appears to be a clear trend in the residuals, with a negative linear relationship. This indicates that the constant variance and independent assumptions are not met either.

Of all the residual plots, the one with the proportion of higher education degrees seems most promising. While there is some cause for concern with the initial residuals in the plot being largely positive, the later residuals after that initial portion are randomly and uniformly distributed around 0. The residuals do appear to be randomly and uniformly distributed around 0. When x is about 30-40%, the variance seems to deviate slightly but it is relatively slight. Thus, we can say that the linearity, independence, and constant variance assumptions seem to hold.

The explanatory variable that seems to be best is the proportion of those with a bachelor's degree or higher in a given county. Based on the residual plots, the assumptions for its SLR appear to hold the most with the least concern, while there are glaring issues in the residuals plots of the other explanatory variables (i.e. non random scattering and/or non uniform distribution). Using the scatter plots constructed for each explanatory variable as well as the regression equation, y = -12195.68 + 1232.85x, there does seem to be a positive linear relationship between proportion of those with a bachelor's degree or higher and the voter turnout in a given county, evident in the equations slope. In this context, the slope is significant because the slope of 1232.85 means that in each county, every increase in the percent of the population with a Bachelor's degree or higher, the total number of voters in that county rose 1232.85.

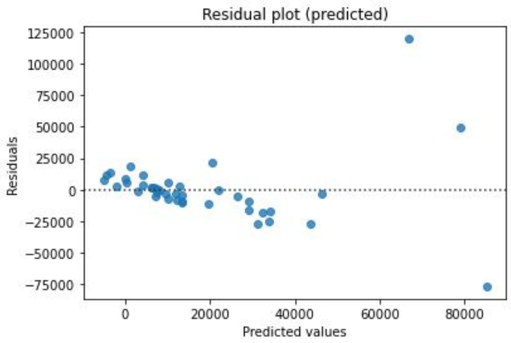
Using the bachelor's degree or higher variable, we conducted the following t-test: The H_0 for this test is that there is a zero slope ($\beta_1 = 0$) whereas H_a is that there is a non-zero slope of the regression line. After the test was conducted, we found a test statistic of 4.650 and a p-value of 0.000. Using the default significance level of 0.05, we have a p-value that is far less than 0.05 which indicates that there exists significant evidence against H_0 and we thus reject H_0. This means the proportion of those with a bachelor's degree or higher is useful in predicting the voter turnout in a given county.

Analysis with MLR: After performing the analysis on the simple linear regression model with the proportion of those with a higher education degree as the explanatory variable, we turn to our multiple linear regression model with the three explanatory variables outlined at the beginning.

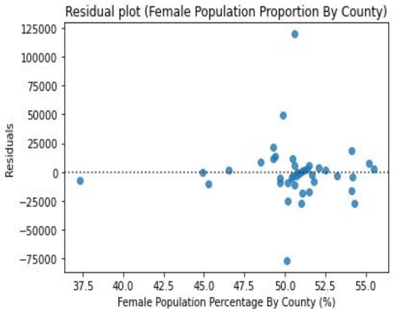
Due to the missing voter turnout values in the last five rows, we'd like to use our model to make some predictions about these values, with an according expression of confidence. Using the multiple linear regression model constructed and its prediction function in python, we see the following predictions for the missing rows (in the order they appear in the dataset and rounded to the nearest integer as we can't have a decimal amount of votes): 43071, 16381, 33373, 11465, 1018. Thus, in terms of voter turnout, we can rank the counties from highest to lowest in the following order: 43071, 33373, 16381, 11465, and 1018.

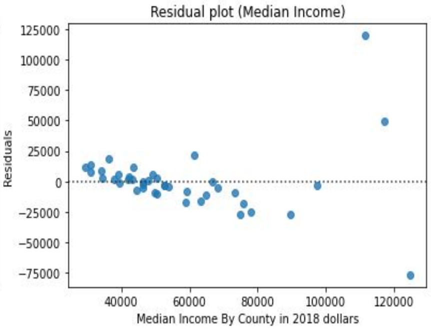
To assess our confidence in these predictions, we would like to evaluate the corresponding prediction intervals for each of the predictions. Accordingly, we compute the following prediction intervals for the voter turnout: (-14460, 100603), (-44245, 77007), (-24262, 91009), (-45082, 68013), and (-56148, 58185). In general, we can say we are rather unconfident in the predictions for voter turnout outlined earlier. The width for each of the prediction intervals are considerably wide and thus hint to notable imprecision and uncertainty, with a width of around 100,000-110,000 for each of the intervals. In addition, the lower bounds for each of the intervals are negative numbers (and thus include 0 in their interval as well) which are not feasible values for voter turnout. Due to so much uncertainty via the prediction intervals, we can not at all be confident in the predictions generated by the model, and by extension, are not confident in the ranking of the counties by voter turnout since we are unsure of the predictions themselves.

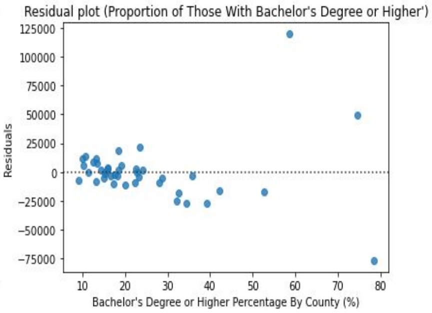
To ensure we can proceed normally with the analysis using this multiple linear regression model, we must first check our assumptions using the relevant residual plots. Due to having n=41 different counties and the MLR model being generally robust against deviations from the Normal distribution, the normality assumption should be satisfied for this model. The residual plots for this model are as follows.



The linearity, independence, and constant variance assumptions are not satisfied, because the residuals are not randomly and uniformly distributed around 0. Namely, the residuals follow a clear trend as seen in the plot (i.e. a negative linear relationship), where lower predicted values correspond to typically positive residuals. Likewise, higher predicted values typically correspond to negative residuals. Thus, there is a clear trend (that is non-uniform) in the distribution of the residuals and so the linearity, independence, and constant variance assumptions are not satisfied.







Again, the female population and median income residual plots fail to meet the linearity, constant variance, and independence assumptions. The female plot still has a large proportion of negative values. The median income does not meet any of the assumptions. There is a clear negative trend in the distribution of residuals so the linearity, independence, and constant variance assumptions are not satisfied; Because of this, we should proceed with caution in the following analyses using this model.

There is very little cause for concern with the education residual plot. Namely, aside from a few outliers on the right hand side of the plot, most of the residuals are randomly scattered around 0 with uniform distribution. There is no clear trend in the data, and the variance is rather constant aside from the outliers. Thus, it's reasonable to conclude that the linearity, independence, and constant variance assumptions appear to hold for this residual plot. T-tests for each of the explanatory variables will now be conducted using the MLR model.

For the proportion of females in a given county, we perform the following t-test:

The H_0 is $\beta_f = 0$, whereas the H_a is $\beta_f \neq 0$. When conducting the appropriate tests, we found the test statistic to be -0.390 and the corresponding p-value to be 0.699. Given that the significance level is unspecified, we assume an alpha of 0.05. Since the p-value is greater than the assumed alpha, we fail to reject H_0. Due to this, we can conclude that there is insufficient evidence in the data to state that the female population percentage by county is useful in predicting the voter turnout in a given county, when median income and the proportion of those with higher education degrees are already in the model.

For the median income explanatory variable, we perform the following t-test:

The H_0 is $\beta_m = 0$, whereas the H_a is $\beta_m \neq 0$. When conducting the appropriate tests, we found a test statistic of 1.322 and the corresponding p-value to be 0.194. Given that the significance level is unspecified, we assume an alpha of 0.05. Since the p-value is greater than the assumed alpha, we fail to reject H_0. Due to this, we can conclude that there is insufficient evidence in the data to state that median income is useful in predicting the voter turnout in a given county, when the proportion of females and those with higher education degrees are already in the model.

Lastly, for the proportion of those with a bachelor's degree or higher, we perform the following t-test: the H_0 is $\beta_h = 0$, whereas H_a is $\beta_h \neq 0$. When conducting the appropriate tests, we found the test statistic to be 0.940 and the corresponding p-value to be 0.353. Given that the significance level is unspecified, we assume an alpha of 0.05. Since the p-value is greater than the assumed alpha, we fail to reject H_0. Because we failed to reject H_0, that means there is insufficient evidence in the data to state that the proportion of those with a bachelor's degree or higher is useful in predicting the voter turnout in a given county, when median income and the proportion of females are already in the model. The values of the test statistics and p values for each of the t-tests performed can be found below.

We would now like to see if the model overall with all the explanatory variables is useful in predicting voter turnout. Consider the following ANOVA F test to test the usefulness of this model in predicting voter turnout. Our H_0 is that the model is not useful (all β are 0), and our H_a is that the model is useful (at least one β is not 0). When doing this test, we found a test statistic of 8.043, and a p value of 0.000299. The significance level is unspecified, so we assume an alpha of 0.05. Because the p value is less than alpha, we reject H_0. Based on the voting data from the 41 Virginia Counties and with the resulting p-value, there exists evidence to reject H_0 and to thus state that the model using the female population proportion of a county, median household income, and higher education proportion is meaningful in predicting voter turnout numbers. It seems that the model overall seems to be useful in predicting voter turnout (implicitly meaning at least one $\beta \neq 0$).

According to the ANOVA F-tests performed on both models computed, both models appear to be useful in predicting voter turnout. In that case, we would like to evaluate whether the additional explanatory variables used in the multiple linear regression model are useful when our "best" explanatory variable is already in the model. Thus, we perform the following partial ANOVA F test. Our H_0 is that the set of extending variables (proportion of females in the county and median income) is not useful (all $\beta = 0$), while the H_a is that the the set of extending variables (proportion of females in the county and median income) is useful (at least one $\beta \neq 0$). The test statistic derived from this test is 12.069, and the p value is 0.0001. Because the significance level is unspecified, we assume an alpha of 0.05. Since the p value is less than alpha, we reject H_0. Because we have rejected H_0, that means there is sufficient evidence in the data to state that the set of extending variables, namely the proportion of females in a county and median income, is useful in predicting voter turnout in a county, with the proportion of those with a bachelor's degree or higher in a county already in the model.

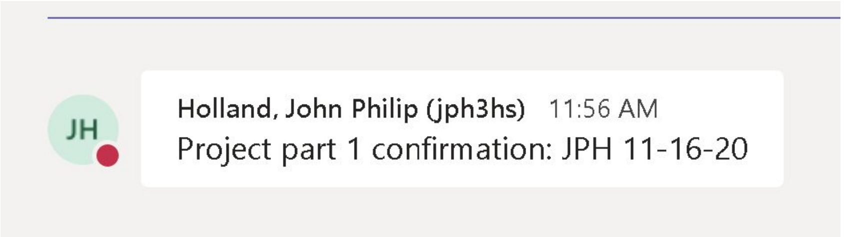
Conclusion: In this report, we conducted several analyses to try and find the "best" explanatory variable to predict voter turnout in a given county or locality in Virginia during the 2020 election. We performed analyses using a model with just one explanatory variable we thought best predicted voter turnout (the proportion of those with a higher education degree), and on another model that included the three explanatory variables outlined from the beginning.

The results of our report were somewhat contradictory. Namely, when evaluating the usefulness of each of the explanatory variables in our extended model (the model with three explanatory variables), it was found that there was insufficient evidence to support any of the explanatory variables being useful in the presence of the others in the model. On the other hand, a following test found that the extended model overall is useful in predicting voter turnout, thus implying that at least one of the explanatory variables is useful in the presence of the others. In addition, it seems that using another test we performed, the explanatory variables added to the extended model are useful in predicting voter turnout. However, the results of the tests using the extended model are negligible. Namely, to evaluate if the extended model was reliable or not, we found that the necessary assumptions to be confident in our extended model were not actually met. As a result, the uncertainty in the reliability of our extended model and the contradictory results of the tests lead us to believe that we cannot heed the extended model too closely.

As for our original model that used the proportion of those with a higher education degree as the sole explanatory variable, we found more concrete results. In this model, we found that the necessary assumptions were met for us to be confident in the results of the test using this model. In addition, we were able to see a linear relationship between the stated explanatory variable and voter turnout numbers. We also found in the relevant test that the model using the proportion of those with a higher education degree is indeed useful in predicting voter turnout. Because of the reliability and consistent evidence of our tests on the original model, we feel confident in stating that the best explanatory variable for predicting voter turnout is the proportion of those with a bachelor's degree.

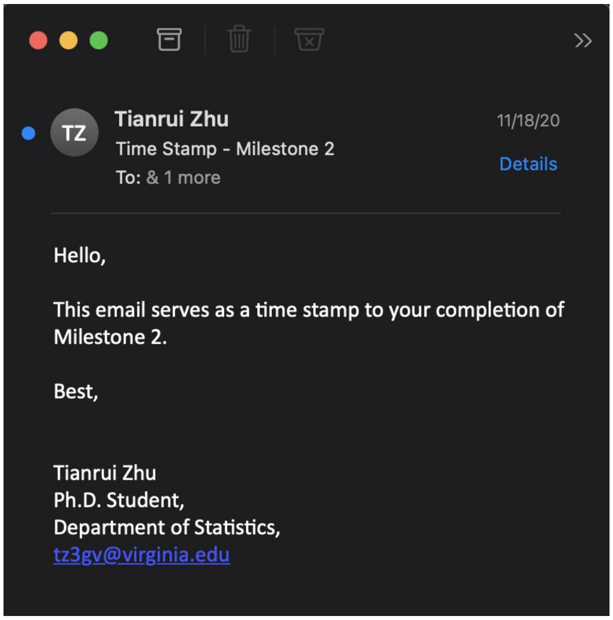
Reflection: After the completion of this project, we found a few changes that could be made to a future project to improve its robustness. We found that a larger data set would have been helpful in this statistical analysis project. A larger data set would allow us to execute more accurate t-tests and predictions. This would have allowed us to choose our "best" variable with more confidence. We believe that having a large data set is always helpful, especially when conducting the amount of tests we did. Since there was such variation between the 41 counties' voter turnout levels, adding another state(s) might be helpful in finding more accurate conclusions. Regarding our methodology, we think that one thing we could have changed was to examine more of the explanatory variables. A different variable might have provided a more accurate model and narrower prediction intervals but that is not something we had time nor space to investigate.

Appendix A:



Holland, John Philip (jph3hs) 11:56 AM

Project part 1 confirmation: JPH 11-16-20



Tianrui Zhu 11/18/20

Time Stamp - Milestone 2

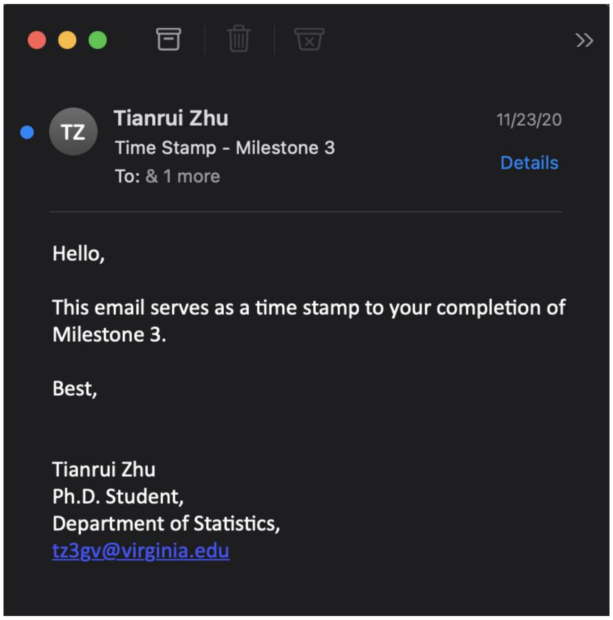
To: & 1 more

Hello,

This email serves as a time stamp to your completion of Milestone 2.

Best,

Tianrui Zhu
Ph.D. Student,
Department of Statistics,
tz3gv@virginia.edu



Tianrui Zhu 11/23/20

Time Stamp - Milestone 3

To: & 1 more

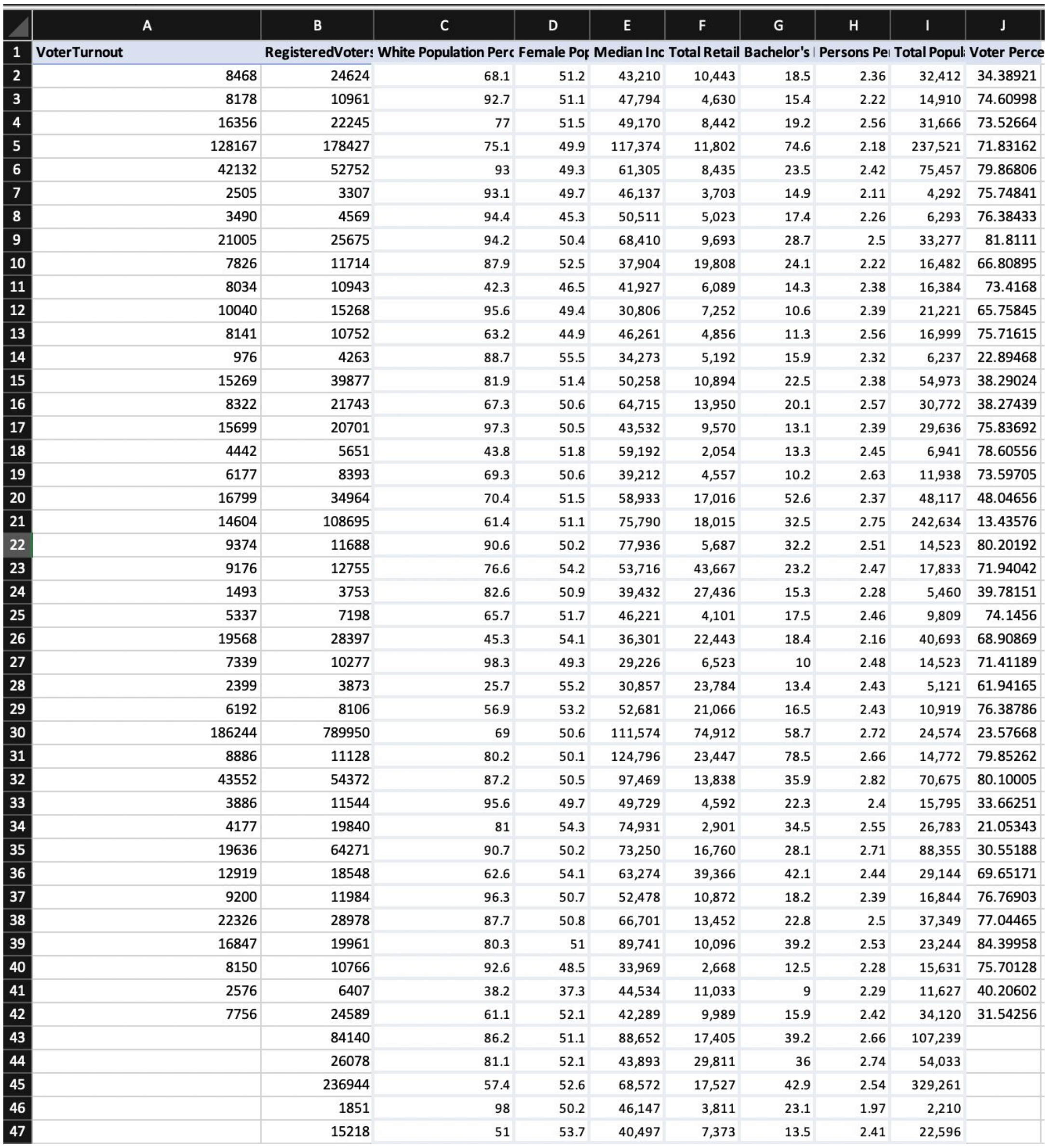
Hello,

This email serves as a time stamp to your completion of Milestone 3.

Best,

Tianrui Zhu
Ph.D. Student,
Department of Statistics,
tz3gv@virginia.edu

Appendix B:



	A	B	C	D	E	F	G	H	I	J
1	VoterTurnout	RegisteredVoters	White Population Perc	Female Pop	Median Inc	Total Retail	Bachelor's	Persons Pe	Total Popul	Voter Perce
2	8468	24624	68.1	51.2	43,210	10,443	18.5	2.36	32,412	34.38921
3	8178	10961	92.7	51.1	47,794	4,630	15.4	2.22	14,910	74.60998
4	16356	22245	77	51.5	49,170	8,442	19.2	2.56	31,666	73.52664
5	128167	178427	75.1	49.9	117,374	11,802	74.6	2.18	237,521	71.83162
6	42132	52752	93	49.3	61,305	8,435	23.5	2.42	75,457	79.86806
7	2505	3307	93.1	49.7	46,137	3,703	14.9	2.11	4,292	75.74841
8	3490	4569	94.4	45.3	50,511	5,023	17.4	2.26	6,293	76.38433
9	21005	25675	94.2	50.4	68,410	9,693	28.7	2.5	33,277	81.8111
10	7826	11714	87.9	52.5	37,904	19,808	24.1	2.22	16,482	66.80895
11	8034	10943	42.3	46.5	41,927	6,089	14.3	2.38	16,384	73.4168
12	10040	15268	95.6	49.4	30,806	7,252	10.6	2.39	21,221	65.75845
13	8141	10752	63.2	44.9	46,261	4,856	11.3	2.56	16,999	75.71615
14	976	4263	88.7	55.5	34,273	5,192	15.9	2.32	6,237	22.89468
15	15269	39877	81.9	51.4	50,258	10,894	22.5	2.38	54,973	38.29024
16	8322	21743	67.3	50.6	64,715	13,950	20.1	2.57	30,772	38.27439
17	15699	20701	97.3	50.5	43,532	9,570	13.1	2.39	29,636	75.83692
18	4442	5651	43.8	51.8	59,192	2,054	13.3	2.45	6,941	78.60556
19	6177	8393	69.3	50.6	39,212	4,557	10.2	2.63	11,938	73.59705
20	16799	34964	70.4	51.5	58,933	17,016	52.6	2.37	48,117	48.04656
21	14604	108695	61.4	51.1	75,790	18,015	32.5	2.75	242,634	13.43576
22	9374	11688	90.6	50.2	77,936	5,687	32.2	2.51	14,523	80.20192
23	9176	12755	76.6	54.2	53,716	43,667	23.2	2.47	17,833	71.94042
24	1493	3753	82.6	50.9	39,432	27,436	15.3	2.28	5,460	39.78151
25	5337	7198	65.7	51.7	46,221	4,101	17.5	2.46	9,809	74.1456
26	19568	28397	45.3	54.1	36,301	22,443	18.4	2.16	40,693	68.90869
27	7339	10277	98.3	49.3	29,226	6,523	10	2.48	14,523	71.41189
28	2399	3873	25.7	55.2	30,857	23,784	13.4	2.43	5,121	61.94165
29	6192	8106	56.9	53.2	52,681	21,066	16.5	2.43	10,919	76.38786
30	186244	789950	69	50.6	111,574	74,912	58.7	2.72	24,574	23.57668
31	8886	11128	80.2	50.1	124,796	23,447	78.5	2.66	14,772	79.85262
32	43552	54372	87.2	50.5	97,469	13,838	35.9	2.82	70,675	80.10005
33	3886	11544	95.6	49.7	49,729	4,592	22.3	2.4	15,795	33.66251
34	4177	19840	81	54.3	74,931	2,901	34.5	2.55	26,783	21.05343
35	19636	64271	90.7	50.2	73,250	16,760	28.1	2.71	88,355	30.55188
36	12919	18548	62.6	54.1	63,274	39,366	42.1	2.44	29,144	69.65171
37	9200	11984	96.3	50.7	52,478	10,872	18.2	2.39	16,844	76.76903
38	22326	28978	87.7	50.8	66,701	13,452	22.8	2.5	37,349	77.04465
39	16847	19961	80.3	51	89,741	10,096	39.2	2.53	23,244	84.39958
40	8150	10766	92.6	48.5	33,969	2,668	12.5	2.28	15,631	75.70128
41	2576	6407	38.2	37.3	44,534	11,033	9	2.29	11,627	40.20602
42	7756	24589	61.1	52.1	42,289	9,989	15.9	2.42	34,120	31.54256
43		84140	86.2	51.1	88,652	17,405	39.2	2.66	107,239	
44		26078	81.1	52.1	43,893	29,811	36	2.74	54,033	
45		236944	57.4	52.6	68,572	17,527	42.9	2.54	329,261	
46		1851	98	50.2	46,147	3,811	23.1	1.97	2,210	
47		15218	51	53.7	40,497	7,373	13.5	2.41	22,596	

Appendix C:

Sources:

1. https://results.elections.virginia.gov/vaelections/2020%20November%20General/Site/Statistics/Registration.html
2. https://www.pewresearch.org/fact-tank/2020/08/18/men-and-women-in-the-u-s-continue-to-differ-in-voter-turnout-rate-party-identification/
3. https://www.pewresearch.org/fact-tank/2020/08/18/men-and-women-in-the-u-s-continue-to-differ-in-voter-turnout-rate-party-identification/
4. https://econofact.org/voting-and-income