Appendix A: Logistic Model Variables

N = 4,951,648

1. **Total Vote**: The total number of times a voter had voted before the 2016 presidential election (which includes the 2016 primaries).
    a. **Total Vote**: The buckets for the revised **Total Vote** variable are the following: 1, 2, 3, (4 or 5), 5+.
2. **Midterm Percentage**: **Midterm Voter**/**Total Vote**.
    a. Note: The **Midterm Vote** is the total number of times a voter had voted in an off-presidential year election (i.e. years [2007,2016] excluding 2008, 2012, and 2016).

    b. **Midterm_Pct_Over25**: A boolean variable, 1 if the **Midterm Percentage** is greater than 25%, 0 otherwise.
3. **Preceding Presidential**: A boolean variable with 1 indicating that the voter had voted in the 2012 presidential election and a 0 otherwise.
4. **Race**: Asian, African-American, Indian-American, Two or More, Other, Undesignated, and White.
5. **Party**: Democrat, Libertarian, Republican, and Unaffiliated.
6. **Gender**: Female, Male, Undesignated.
7. **Age Group**: 27-34, 35-44, 45-54, 55-64, 65+.

    The age groups were determined by suggested survey classifications (link: http://www.pgagroup.com/standardized-survey-classifications.html).
8. **Ethnicity:** Hispanic, not-Hispanic, and Unknown.
9. **Pct_Under20**: The percentage of population younger than 20 years of age.
10. **Pct_Under40**: The percentage of population between the age of 20 and 39, inclusive.
11. **Pct_Under60**: The percentage of population between the age of 40 and 59, inclusive.
12. **Pct_Above60**: The percentage of population 60 years of age or older.
13. **Pct_Hisp**: The percentage of Hispanic population.
14. **Pct_White**: The percentage of White population.
15. **Pct_Black**: The percentage of Black population.
16. **Pct_Asian**: The percentage of Asian population.
17. **Pct_Other**: The percentage of another identified ethnicity population.
18. **Pct_Renter**: The percentage of individuals renting their houses.
19. **Pct_Owner**: The percentage of individuals owning their houses.
20. **House_Incom_Below55k**: A boolean variable: 1 if the average household income for the zip code is below 55k, 0 otherwise.
21. **Avg_House_Size**: The average household size.
22. **Pct_Blue_Collar**: The percentage of blue-collar workers (population = civilian employed population who are 16 years and older).
23. **Pct_White_Collar**: The percentage of white-collar workers (population = civilian employed population who are 16 years and older).

Appendix B: The Interactive Model

The model:

The probability model uses the coefficients from the logistic output to calculate probabilities based on the standard formula:

$$\hat{p} = \frac{\exp(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}{1 + \exp(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}$$

Where $\hat{p}$ is the expected probability, b_0 is the intercept, b_p are the variable coefficients, and X_p are the variable inputs.

Using the model:

The model can be found at: https://github.com/richardtwang/HVT. Columns G and H represent the model input of before and after for a particular voter. Users may modify any X_p in the model to see the change in probability of that voter turning out. For example, setting cell G5 = 0 and H5 = 5, one may observe the change in probability of a voter turning out when his or her total vote changes from 0 to 5 (assuming every other variable is identical across the columns G and H). The probability is calculated and shown at the bottom of the sheet, in cells G37 and H37 respectively, with the change in probability shown in H38. The 'Reset' button at the top will reset the inputs for all the variables under column G that were arbitrarily picked to be reset to, *except for the census variables (rows 24-34 inclusive)*: these variables are reset to the state averages.

There are restrictions to the inputs one may provide for X_p, which is built in the model and listed as follows:

1. The total vote is bucketed at five levels; the valid input value range is [0, 5]
2. For all the binary variables, the valid input value is 0 or 1
3. For the percentage variables, the valid input value is between 0 and 100% (The model estimates in the interactive model have been adjusted for the input variable scale – from the percentage units to fractions of 1).
4. User must not enter multiple 1s for each category "block" as a voter cannot be registered, for example, as a male AND a female in the voter files. Same rule applies to all the census variables – the percentages for each category block cannot exceed 100%. If there is 60% of the population between the ages of 20 and 40 in a neighborhood, there cannot be 70% of the population between the ages of 40 and 60 in the same neighborhood.

Rules # 1 – 3 are controlled by data validation and rule #4 is controlled by the "Calculation" button in the model.

Users must also keep in mind that the reference variables are still integrated in the model even though they are not shown. For example, if Libertarian, Republican, and Affiliated are all 0, then Democrat (which was the reference variable for party, as mentioned above) is automatically treated as a 1.

Limitations:

There are several limitations to the model. Firstly, while the model uses the habitual voter theorem to model voter behavior, it does not (nor does it attempt to) measure the exact effects of habit on voting. The model only uses past voting behavior as a predictor of future voting behavior. As previously mentioned, Gerber et al. have argued that a simple regression on lagged voting is insufficient to determine the influence of habit on voting due to the possibility of external variables effecting both past and future turnout. Consequently, users should not interpret our model as if it were proving the concept of habitual voting itself.

Other limitations exist within the North Carolina voter files. The data does not include crucial information such as voters' education and income levels. I was able to mitigate this issue by utilizing census data offered at the zip code level; an improvement would be to rerun the model with individual level data, along with other information such as

residential mobility, exposure to campaign ads, political interest, weather, etc. (i.e. survey data). Furthermore, the data does not include the nonvoter population – individuals who have never voted in any election (including the 2016 presidential election). Consequently, we had to remove the population of voters with Total Vote = 0 from our analysis. Although the issue is mitigated by the large population size of the remaining voter pool, it remains as a limitation nonetheless.

Lastly, North Carolina's status as a swing state is also likely to introduce concerns to our model. Since the election is considered to be more competitive, it is more likely to receive a greater amount of media coverage. Furthermore, individuals may be more likely to turnout merely because they perceive their votes to count more. While this does not pose a limitation to our model, since all subjects observed (North Carolina voters) are exposed to the influences, it does pose as a challenge to any attempt to generalize the model's results to extra-state elections and extra-state voters.

Appendix C: Logistic regression output

Deviance Residuals:	**Min**: -2.8114	**1Q**: 0.2718	**Median**: 0.3566	**3Q**: 0.5415	**Max**: 1.8672
Dependent Variable					
2016 General					
Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.668638	0.18274	-20.076	< 2e-16	***
Predictor Variables:					
Total Vote	0.624869	0.002197	284.417	< 2e-16	***
Preceding Presidential	0.401873	0.005766	69.694	< 2e-16	***
Midterm_Pct_Over25	0.227919	0.005611	40.623	< 2e-16	***
Race (with 'White' as reference):					
Asian	0.126007	0.021076	5.979	2.25E-09	***
African-American	-0.147048	0.006584	-22.333	< 2e-16	***
Indian-American	-0.296352	0.02645	-11.204	< 2e-16	***
Two or More	-0.295859	0.029507	-10.027	< 2e-16	***
Other	-0.043429	0.017136	-2.534	0.011265	*
Undesignated	0.026339	0.01577	1.67	0.094867	.
Party (with 'Democrat' as reference):					
Libertarian	0.104778	0.034481	3.039	0.002376	**
Republican	0.223923	0.006114	36.625	< 2e-16	***
Unaffiliated	0.150605	0.00592	25.44	< 2e-16	***
Gender (with 'Male' as reference):					
Female	0.046913	0.004479	10.474	< 2e-16	***
Undesignated	0.181087	0.020086	9.015	< 2e-16	***
Age Group (with '27-34' as reference):					
34-44	0.206618	0.007199	28.699	< 2e-16	***
45-54	0.332455	0.0073	45.541	< 2e-16	***
55-64	0.408585	0.007646	53.437	< 2e-16	***
65+	-0.06172	0.007367	-8.378	< 2e-16	***
Ethnicity (with 'not-Hispanic' as reference):					
Hispanic	0.203101	0.018094	11.224	< 2e-16	***
Unknown	-0.061083	0.005708	-10.702	< 2e-16	***
Zip Code Age (with 'Pct_Under20' as complement):					
Pct_Under40	0.006477	0.001650	3.925	8.67E-05	***
Pct_Under60	0.020495	0.001791	11.443	< 2e-16	***
Pct_Above60	0.005355	0.001513	3.539	0.000402	***
Zip Code Ethnicity (with 'Pct_Other' as complement):					
Pct_Hisp	0.009564	0.000693	13.806	< 2e-16	***
Pct_White	0.008784	0.000464	18.919	< 2e-16	***
Pct_Black	0.008382	0.000486	17.241	< 2e-16	***
Pct_Asian	0.011895	0.001121	10.614	< 2e-16	***
Zip Code Housing (with 'Pct_Renter' as complement):					
Pct_Owner	0.004546	0.000450	10.098	< 2e-16	***

Appendix C: Logistic regression output (*continued*)

Zip Code Household and Income:					
Avg_House_Size	0.174605	0.030613	5.704	1.17E-08	***
House_Income_Below55k	-0.077051	0.006564	-11.738	< 2e-16	***
Zip Code Occupation (with 'Pct_Blue_Collar' as complement):					
Pct_White_Collar	0.003645	0.000359	10.153	< 2e-16	***

Signif. codes: ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 1637072 on 1809970 degrees of freedom					
Residual deviance: 1340913 on 1809939 degrees of freedom					
AIC: 1340977					

Appendix D: Logistic regression output with odds ratio*

	Confidence Interval		
	Odds Ratio	lower	upper
Predictor Variables			
Total Vote	1.868001	1.859975	1.876062
Preceding Presidential	1.494621	1.477824	1.511609
Midterm_Pct_Over25	1.255984	1.242247	1.269872
Race (with 'White' as reference):			
Asian	1.134290	1.088387	1.182128
Black	0.863253	0.852184	0.874465
Indian-American	0.743526	0.705962	0.783088
Two or More	0.743892	0.702091	0.788182
Other	0.957501	0.925875	0.990206
Party (with 'Democrat' as reference):			
Libertarian	1.110465	1.037895	1.188108
Republican	1.250975	1.236074	1.266056
Unaffiliated	1.162537	1.149126	1.176105
Gender (with 'Male' as reference):			
Female	1.048031	1.038870	1.057271
Undesignated	1.198519	1.152251	1.246645
Age Group (with '27-34' as reference):			
35-44	1.229513	1.212285	1.246985
45-54	1.394387	1.374578	1.414481
55-64	1.504687	1.482305	1.527406
65+	0.940146	0.926668	0.953820
Ethnicity (with 'not-Hispanic' as reference):			
Hispanic	1.225196	1.182505	1.269427
Unknown	0.940746	0.930280	0.951329
Zip Code Age (with 'Pct_Under20' as complement):			
Pct_Under40	1.006498	1.003248	1.009759
Pct_Under60	1.020707	1.017130	1.024296
Pct_Above60	1.005369	1.002392	1.008355
Zip Code Ethnicity (with 'Pct_Other' as complement):			
Pct_Hisp	1.009610	1.008240	1.010982
Pct_White	1.008822	1.007905	1.009741
Pct_Black	1.008417	1.007457	1.009379
Pct_Asian	1.011966	1.009746	1.014192

Appendix D: Logistic regression output with odds ratio (*continued*)

Zip Code Housing (with 'Pct_Renter' as complement):			
Pct_Owner	1.004557	1.003671	1.005443
Zip Code Household and Income:			
Avg_House_Size	1.190776	1.121429	1.264411
House_Income_Below55k	0.925843	0.914008	0.937831
Zip Code Occupation (with 'Pct_Blue_Collar' as complement):			
Pct_White_Collar	1.003652	1.002946	1.004359

***** The percentage variables are in percent units. The odds ratios for them apply to each increase in percent units.

Appendix E: Interpreting the Model Results

Quickly summarizing some of the results of our control variables, we see that individuals of every race are less likely to vote than White voters except for Asian voters (who are more likely to turnout) and racially-undesignated voters (who turnout at essentially the same rate compared to Whites); voters registered as Libertarian, Republican, or Unaffiliated were all more likely to turnout than those registered as Democrat; male voters were more likely to turnout than female and undesignated voters; voters in the age groups 34-44, 45-54, and 55-64 were more likely to turnout than voters in the age group 27-34 while voters in the age group 65+ were less likely; Hispanic voters were more likely to vote than non-Hispanic voters, while ethnically-unknown were less likely; and so forth.

The coefficients for our predictor variables (Total Vote, Midterm Percentage, and Preceding Presidential) were all statistically significant at the $p < 0.001$ level, and in the correct (positive) direction. For Total Vote, an increase from 0 to 1, 1 to 2, 2 to 3, 3 to (4 or 5), and 5+ all each increased the likelihood of that voter turning out in the 2016 presidential election by 1.868 times. For Preceding Presidential, we see that if a voter participates in the 2012 presidential election, the chances of his or her voting in the 2016 presidential election increases 1.494 times. Lastly, if the Midterm Percentage (Midterm Vote / Total Vote) is greater than 25%, then that voter is 1.255 times more likely to turnout in the 2016 presidential election. All of these interpretations involve the odds ratios.