The Ground Game

My contributions to the 84th District State House Race

Drew Jones

### Introduction

In this project I analyze a dataset I created out of records from my canvassing in Michigan’s 84th State House District during the 2024 Election. To create the dataset, I logged the results of my individual door knock lists and activity information from my phone’s pedometer.

Common belief, and something I have oft-repeated, is that one should expect responses on approximatey 10% of doors. As a result, I examine if that was the case for me. Additionally, it is expected that weekends will see a higher response\_rate due to a greater portion of people being off work on those days. I test this as well. I also seek out which individual day saw the highest response\_rate. Finally, I model the data using Analysis of Variance and Multiple Linear Regression in an attempt to predict response\_rate.

### Making Additional Variables

My primary interest here is to determine what impacted the share of doors that I knocked that led to an interaction with a voter. As a result, my response variable will be response\_rate, measured in percentage of doors knocked that saw a response.

Steps was taken directly from my phone’s pedometer function. On days when I attended class, I took approximately 5000 additional steps while on-campus. On days when I did not attend class, I expect the vast majority of my step count to come directly from canvassing. I create the adjusted steps variable, adj\_steps to account for those additional steps.

### Explanatory Data Analysis

An observation is a single day that I canvassed, and are named based on the Date. The variable Doors indicates the total number of doors that I knocked that day. Attempts is the number of people I attempted to reach. Steps is the total number of steps I took that day. Lists is the number of lists of doors I was assigned on a given day. Class is a dummy variable indicating if I attended class that day on-campus at Grand Valley State University in Allendale, MI. Weekend is a dummy variable indicating if that day was a weekend day or not.

For example, a single address with 2 voters that I was reaching out to would have a Attempts value of 2, and a Doors value of 1. On this campaign, lists varied from ~30-90 doors, and were usually completed fully. The date variable is already stored in POSIXct format, but I will apply lubridate:ymd to remove the unused time codes.

My response variable response\_rate is comprised of 100 (Doors Responses) to give a percentage of Doors that saw a response.

To begin, I briefly examine the dataset with skim:

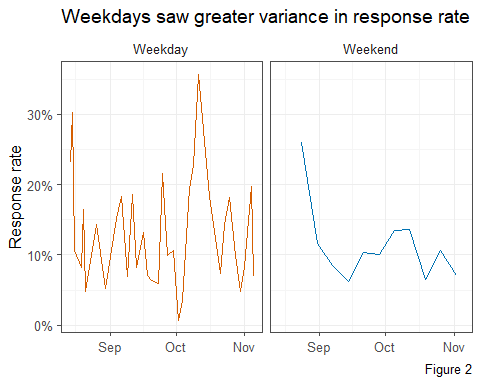
Figure 1: Basic Summary

| **Variable** | **Distribution** | **Mean Value** | **Standard Deviation** | **Minimum** | **First Quartile** | **Median** | **Third Quartile** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Doors | ▃▇▂▃▁ | 93.47 | 49.78 | 14.00 | 60.00 | 80.00 | 122.00 | 235.00 |
| Attempts | ▅▇▂▂▁ | 176.29 | 93.84 | 36.00 | 113.00 | 156.00 | 249.00 | 418.00 |
| Responses | ▆▇▇▅▂ | 10.04 | 5.09 | 1.00 | 6.00 | 9.00 | 13.00 | 23.00 |
| Lists | ▅▇▁▃▁ | 2.02 | 0.87 | 1.00 | 1.00 | 2.00 | 2.00 | 4.00 |
| Steps | ▇▆▇▅▂ | 10,726.62 | 2,861.87 | 5,901.00 | 8,316.00 | 10,534.00 | 12,631.00 | 17,385.00 |
| adj\_steps | ▇▃▇▇▃ | 8,615.51 | 4,616.41 | 901.00 | 4,001.00 | 9,487.00 | 12,060.00 | 17,385.00 |
| response\_rate | ▆▇▅▂▁ | 12.65 | 7.39 | 0.67 | 7.14 | 10.61 | 18.18 | 35.71 |

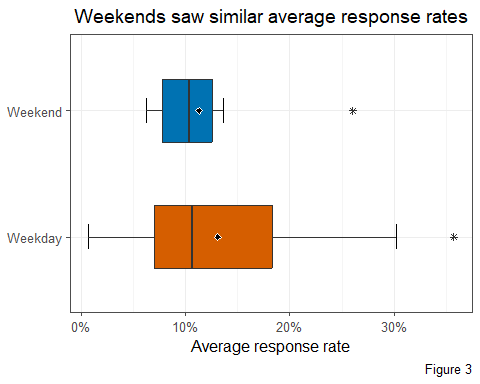
There are 45 total observations. I assembled this dataset by hand in Microsoft Excel, and as a result there are no missing values. Interestingly, most of the variables appear to have a slightly right-tailed distribution. This is perhaps most visible in repsonse\_rate, but one could argue it is the case for each variable except adj\_steps. The median response\_rate is 10.61, which suggets that the common rule of thumb of 10% is broadly accurate.

### Basic Analysis

Now I examine the response\_rate over time.



In *Figure 2*, I see that response\_rate varied more day-to-day during the week than it did between weekend days. However, both appear to have followed a similar trend of peaks in late August and in mid-October.



In *Figure 3*, I see that the mean and median of response\_rate are similar for weekdays and weekends alike. Once again though, I see that rates varied greatly on weekdays. Both also appear to have a single outlier of especially high response\_rate.

Figure 4: Summary of Weekend

| **Weekend** | **Range** | **Min** | **Q1** | **Median** | **Q3** | **Max** |
| --- | --- | --- | --- | --- | --- | --- |
| Weekday | 35.04 | 0.67 | 7.02 | 10.64 | 18.33 | 35.71 |
| Weekend | 19.78 | 6.25 | 7.81 | 10.34 | 12.53 | 26.03 |

This reflects the findings of EDA in *Figure 4*, which show weekdays saw response\_rate of <1% all the way up to >35%, while Weekends ranged only from 6.25% to 26%.

Figure 5: Biggest response days

| **Date** | **Response Rate** | **Doors** | **Weekend** | **Adjusted Steps** |
| --- | --- | --- | --- | --- |
| 2024-10-11 | 35.71 | 14 | no | 9,487 |
| 2024-08-15 | 30.26 | 76 | no | 13,680 |
| 2024-08-24 | 26.03 | 73 | yes | 11,660 |

In *Figure 5* the biggest values of response\_rate are shown. This includes the pair of outliers I noted in *Figure 3*, which are October 11th and August 24th. It is worth observing that October 11th only saw 14 Doors knocked, so that day appears to be an outlier in terms of my actual activity as well. August 15th therefore appears to be the largest “true” response\_rate I had on a day.

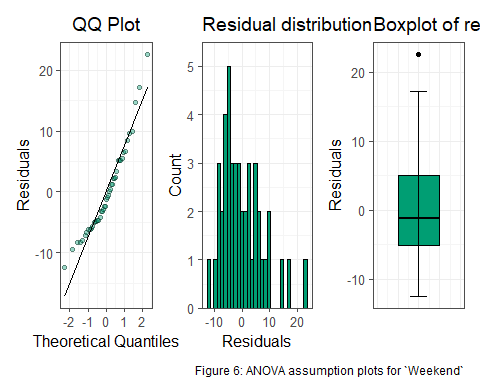
### ANOVA: Weekend

Next, I investigate whether Weekend is useful in helping to predict response\_rate. What follows is an ANOVA F-Test.

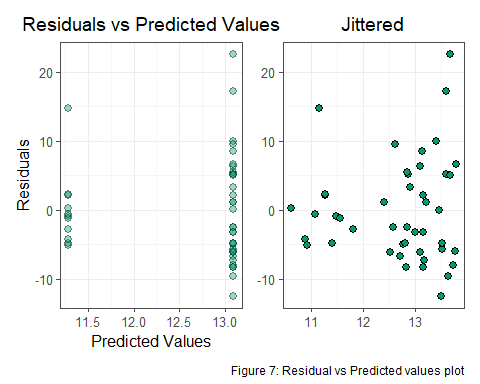
My response variable is response\_rate and the factor variable is Weekend. The population for this is all canvassing days and the sample is the 45 days that I have observatons for.

The null hypothesis for the F-test is .

The alternative hypothesis for the F-test is .



The first condition is normal distribution of residuals, which I examine using the QQ plot, histogram of residuals, and boxplot of residuals in *Figure 6*. The data is left-tailed but not egregiously so. I consider the residual distribution assumption met here.



Next, I examine the residuals vs. predicted value plot in *Figure 7*, and see that the residuals do not vary constantly. The residuals appear to increase as the predicted values grow larger. The assumption of constant variance is not met.

It is safe to assume that the residuals of one observation do not impact the residuals of any other observations, so I consider the residual independence assumption to not be met.

Figure 8: ANOVA Table for `Weekend`

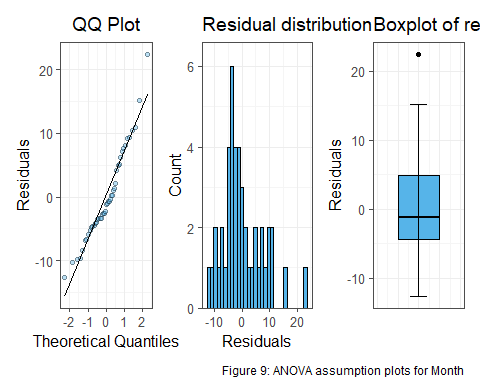
| **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
| --- | --- | --- | --- | --- |
| 1 | 26.90301 | 26.90301 | 0.4869471 | 0.4890487 |
| 43 | 2,375.67793 | 55.24832 |  |  |

As the assumption of constant variance is not met, it is not neccesarily useful to complete the process of ANOVA, but I show the table in *Figure 8* to simply share the p-value. This is significantly above the level of analysis 0.05 and reinforces my decision here to fail to reject the null hypotheses.

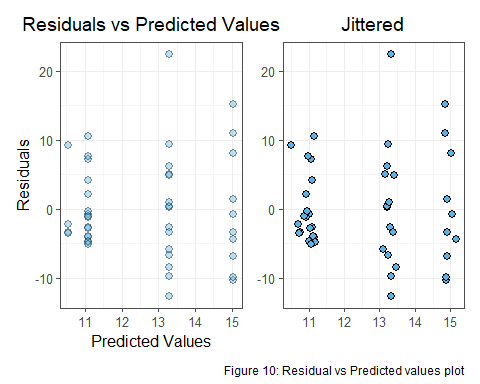
The difference in mean response\_rate for the two levels of Weekend is not statistically significant.

### ANOVA: Month

Month of the year may also be useful in predicting repsonse\_rate. I saw earlier that response\_rate varied over time, peaking in August and October. Therefore, I create the variable Month, which is categorical with levels for August, September, October, and November.



For the Month model, the normal distribution of residuals assumption appears to be met, as I see in *Figure 9*. There is an outlier point which makes the distribution slightly left tailed but, as with Weekend, not severely.



Some clustering is present in *Figure 10*. However, looking past the lone outlier with a residual above 20, the constant variance assumption is met. The residuals are relatively constant across the predicted values.

It is fair to assume that the residual of an observation does not impact the residuals of any other observations, so the independence of residuals assumption is met.

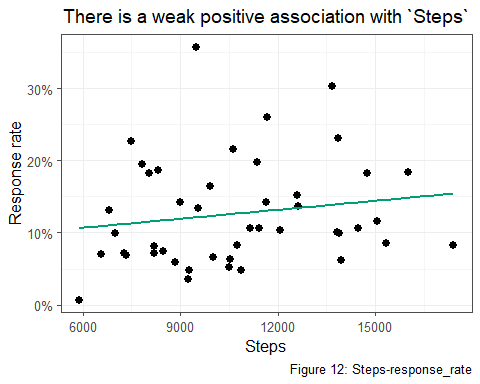
Figure 11: ANOVA Table for `Month`

| **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
| --- | --- | --- | --- | --- |
| 3 | 120.3049 | 40.10164 | 0.7204069 | 0.545566 |
| 41 | 2,282.2760 | 55.66527 |  |  |

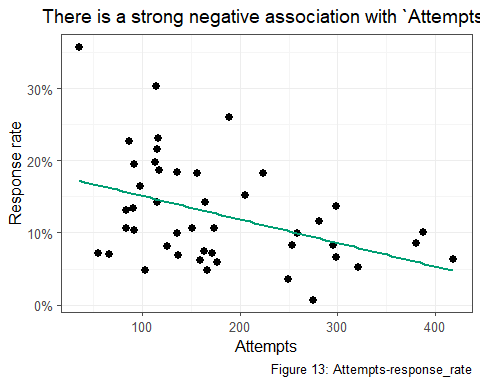
Based on the P-Value 0.54 seen in *Figure 11*, I conclude there is not statistically significant evidence that Month is useful for predicting the mean response\_rate.

### Linear Regression Model Term Selection

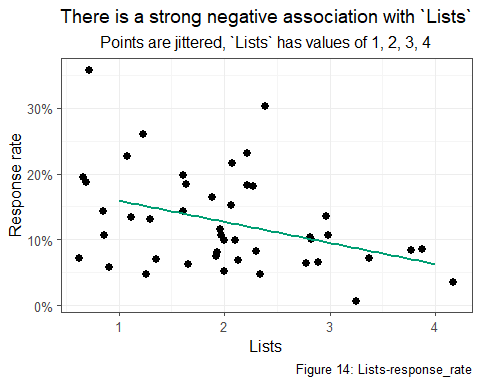
In this section I will construct a linear regression model to predict response\_rate. The first step in this is determining which variables are useful among Attempts, Lists, and Steps. This is done through the scatterplots that follow.



In *Figure 12* I see a weak positive relationship between Steps and response\_rate. As to why this particular relationship exists, I would theorize that higher steps totals are associated with areas with less-dense housing. That sort of area may be populated more by adult families than young people, who would live with roommates and drive up the Attempts per door.

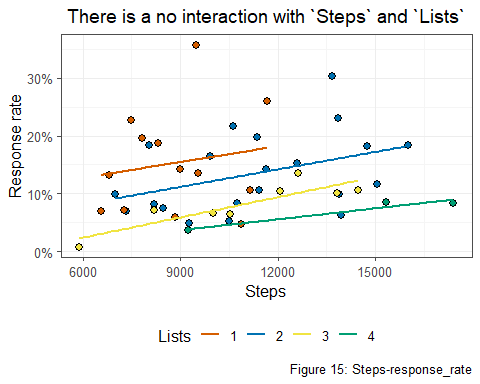


In *Figure 13*, there is a strong negative relationship between Attempts and response\_rate. This makes sense, as a the observations with the largest number of Attempts are likely households with multiple voters, such as one containing parents and adult children. A non-response to one of these doors would mean a high number of attempts.

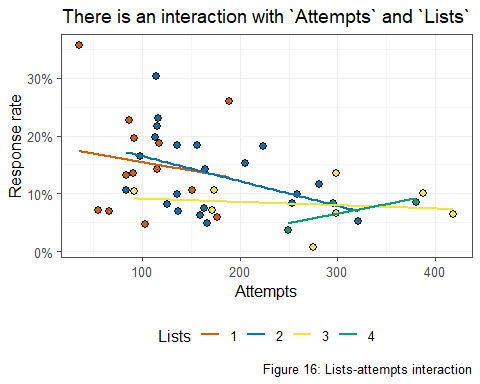


In *Figure 14*, the relationship between Lists and response\_rate is clearly negative. This makes sense as well. I would consider a standard day to have 1 or 2 Lists, maybe 3 on a bigger one. However, days with 3 or 4 Lists may have been areas with very low response\_rate, hence the ability to complete a high number of Lists.

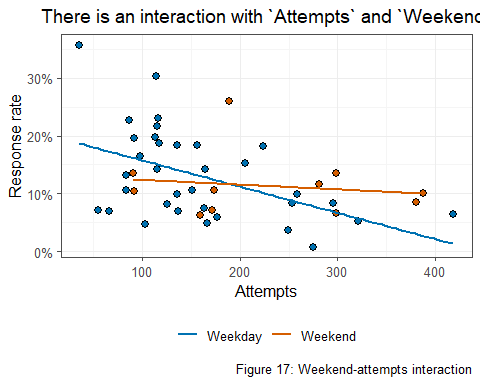
There may also be interactions between some variables.



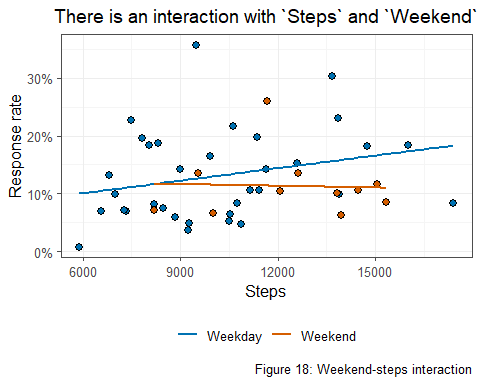
There could be an interaction between Steps and Lists. I would expect days where I did multiple Lists to be days where I worked a greater amount of time. As a result, I would expect those days to have a greater number of Steps as well. In *Figure 15*, that interaction is examined, and the plot suggests that an interaction effect is not present. The regression lines for each level of Lists are approximately parallel.



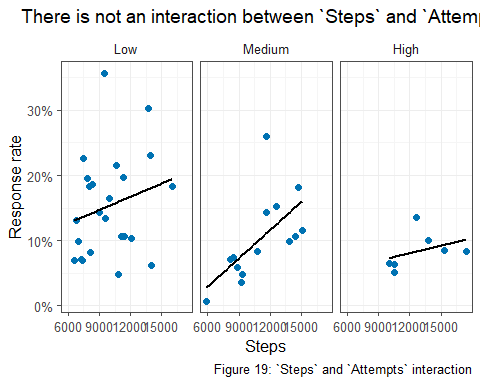
There may be an interaction between Attempts and Lists for a similar reason. Days with a greater number of Lists may have also seen a greater number of Attempts simply due to the amount of time spent canvassing. In *Figure 16*, I do appear to see an interaction effect between Attempts and Lists.



I would anticipate and interaction between Attempts and Weekend due to weekend days seeing a greater number of average hours worked than weekdays. In *Figure 17*, an apparent interaction between Attempts and Weekend is observed.



An interaction between Steps and Weekend may also be expected for the reason of time worked on a weekend day. In *Figure 18* and interaction between Steps and Weekend is observed.



Steps and Attempts may interact as well. Logically, Attempts would increase as Steps did. However, in *Figure 19*, there does not seem to be an effect, or at least not a strong one. *Figure 19* is faceted on a binned version of the Attempts variable, and although the regression lines are not parallel, the three facets show positive relationships, and the low number of observations in the *High* plot leads me to believe that the trend would be similarly positive if more observations were present. Overall, this interaction effect is not clear enough to be included in the model.

I have determined that the plausible interactions between variables here are Attempts\*Lists, Attempts\*Weekend, and Steps\*Weekend. These interactions will be used to fit the linear regression model.

### Fitting linear Regression Model

Figure 20: Models sorted by Adjusted R-Squared

| **n** | **predictors** | **adjr** | **aic** |
| --- | --- | --- | --- |
| 3 | Lists Weekend:Attempts Weekend:Steps | 0.3274997 | 297.4160 |
| 5 | Attempts Lists Steps Weekend:Attempts Weekend:Steps | 0.3274997 | 297.4160 |
| 4 | Attempts Lists Weekend:Attempts Weekend:Steps | 0.3274997 | 297.4160 |
| 4 | Lists Steps Weekend:Attempts Weekend:Steps | 0.3274997 | 297.4160 |
| 6 | Weekend Attempts Lists Steps Weekend:Attempts Weekend:Steps | 0.3131622 | 299.1964 |

Fitting all possible models from the variables shows that the 3 term model with Lists, the interaction Weekend\*Attempts, and the interaction Weekend\*Steps is the best, as can be seen in *Figure 20*. It features identical adjusted r-squared and AIC as bigger models that includes Attempts and Steps.

I now recreate the chosen model to examine some of the effects closer. Using lm() to create it will provide slighly different values than ols\_steps\_all\_possible(), which was used to fit all possible models and is the cource of the values in *Figure 20*

Figure 21: Model terms

| **term** | **estimate** |
| --- | --- |
| (Intercept) | 10.564096494 |
| Lists | -2.305580646 |
| Weekendyes | 4.811521266 |
| Attempts | -0.044435411 |
| Steps | 0.001362465 |
| Weekendyes:Attempts | 0.046011842 |
| Weekendyes:Steps | -0.001248254 |

Looking at *Figure 21*, I see that the most powerful piece of the model is Weekend, with Weekend days leading to an approximate increase in response\_rate of 4.8%.Lists is also somewhat more powerful, as each additional list is associated with a decrease of 2.3% in response\_rate. The full equation for the model is shown below:

Ultimately, my chosen model features an adjusted r-squared value lower than 0.4. This metric captures the amount of variance in the response variable, in this case response\_rate, explained by the model, and a value that low indicates that this model is not realistically useful in predicting respons\_rate, nor is any possible model with these variables.

### Conclusions

From all of this, answers to my original questions have emerged. The expected ~10% response\_rate held true with a median of 10.61. My guess that weekends would see a higher response\_rate was not reflected in reality. Weekends actually saw a slightly lower response\_rate than weekdays. Interestingly, weekdays saw significantly lower variance than weekdays. Other patterns are present as well. The two groups also saw similar peaks, with high response rates recorded in late August and in mid-October for both. This includes the highest individuals day for response\_rate, August 15th.

ANOVA models did not reveal any statistically significant relationships between the variables Weekend and response\_rate, or Month and response\_rate. My efforts to build a linear regression model to predict response\_rate were also not fruitful, as the best models I could develop could not surpass an adjusted r-squared value of 0.327. This dataset’s resistance to being modeled effectively is certainly due to the limited nature of the variables I had at my disposable. Lists, Steps, Attempts, nor any other I have here are likely to be found in proven voter outreach models.