Final Capstone Project for

Springboard Foundations of Data Science

Cyber Attack Survey Project

C Halsell

# Project Overview

I reviewed many data sets before selecting survey data from Pew Research Center for the People & the Press. I wanted to select data that I normally do not use in my supply chain management activities and include a current topic. Pew conducted a phone survey in March 2016 that investigated cyber security issues. Of the 2,254 people called during their survey period, 1,040 observations were included in the data pulled from their website. The data started with 104 variables.

**The goal of this project is to identify a potential target audience for a secure password application.**

# Data Wrangling

Initially, I imported it into Excel to get an idea of the column headers. They were very long, so I shortened them into R-friendly formats prior to importing the dataset into R. The majority of the variables were categorical or ordinal, and there were a few numerical and interval variables as well. I pasted the initial variable string below:

> str(Cybersecuritymeans)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 1040 obs. of 105 variables:

$ psraid : num 101100 100762 201683 100776 200149 ...

$ sample : num 1 1 1 1 1 1 1 1 1 1 ...

$ sex : num 1 2 2 2 1 1 2 2 2 1 ...

$ age : num 60 88 90 76 58 67 85 83 84 53 ...

$ agegroup : chr "Young\_Boomer" "GI\_Gen" "GI\_Gen" "Silent\_Gen" ...

$ educ2 : num 3 2 2 3 1 2 8 8 3 2 ...

$ marital : num 6 5 5 1 3 5 3 5 5 1 ...

$ par : num 2 2 2 2 2 2 2 2 2 2 ...

$ ideo : num NA 5 3 3 5 NA 2 2 2 1 ...

$ hisp : num 2 2 2 2 1 2 2 2 2 2 ...

$ race3m1 : num 1 1 1 1 7 NA NA 2 1 1 ...

$ emplnw3 : num 3 3 3 3 3 3 3 3 3 1 ...

$ party : num 1 2 2 2 2 2 2 1 3 1 ...

$ partyln : num NA NA NA NA NA NA NA NA 1 NA ...

$ inc : num 1 NA 1 3 2 2 NA NA NA 3 ...

$ hh1 : num 1 1 2 1 3 1 8 1 1 2 ...

$ hh3 : num NA NA 2 NA 3 NA 1 NA NA 2 ...

$ q1 : num 1 2 2 2 2 2 2 1 1 2 ...

$ qc1 : num NA NA 1 NA 2 2 NA NA NA 2 ...

$ eminuse : num 2 2 2 2 2 2 2 2 2 2 ...

$ intmob : num 2 2 2 2 2 2 2 2 2 2 ...

$ intfreq : num NA NA NA NA NA NA NA NA NA NA ...

$ device1a : num 2 2 NA 2 NA NA 1 2 1 NA ...

$ smart1 : num NA NA 2 NA 8 2 8 NA 2 2 ...

$ snsint2 : num NA NA NA NA NA NA NA NA NA NA ...

$ meanacct : num 5.33 6 5.6 4 3.2 ...

$ meansecur : num 3 1.83 2 2 5 ...

$ meanhabit : num NA NA NA NA NA NA NA NA NA NA ...

$ meanpolicy : num 5.73 6.18 4.64 5.73 3.45 ...

$ meanactivity: num 4.69 4.67 4.08 3.91 3.88 ...

$ usr : chr "Urban" "Urban" "Urban" "Rural" ...

$ cregion : chr "West" "South" "West" "South" ...

$ state : num 6 12 53 5 48 51 5 13 55 55 ...

$ density : num 4 3 2 1 5 5 2 1 3 1 ...

$ acct1a : num NA NA NA NA NA NA NA NA NA NA ...

$ acct1b : num NA NA NA NA NA NA NA NA NA NA ...

$ acct1c : num NA NA NA NA NA NA NA NA NA NA ...

$ acct1d : num NA NA NA NA NA NA NA NA NA NA ...

$ acct2 : num NA NA NA NA NA NA NA NA NA NA ...

$ acct3a : num NA NA 8 NA 2 3 2 NA 2 4 ...

$ acct3b : num NA NA 5 NA 4 3 3 NA 2 8 ...

$ acct3c : num NA NA NA NA NA NA NA NA NA NA ...

$ acct3d : num NA NA NA NA NA NA NA NA NA NA ...

$ acct3e : num 3 5 2 4 4 8 2 2 2 4 ...

$ acct3f : num 5 5 5 4 4 3 5 5 1 4 ...

$ acct3g : num 8 8 8 4 2 3 5 5 2 3 ...

$ secur1 : num 8 1 3 2 2 2 3 2 3 2 ...

$ secur2a : num 2 2 2 2 8 2 2 2 2 2 ...

$ secur2b : num 2 2 1 2 8 8 2 2 2 2 ...

$ secur2c : num 2 2 2 2 2 2 2 2 2 2 ...

$ secur2d : num NA NA NA NA NA NA NA NA NA NA ...

$ secur2e : num NA NA NA NA NA NA NA NA NA NA ...

$ secur2f : num 2 2 2 2 2 8 2 2 8 2 ...

$ secur2g : num 2 2 2 2 8 8 2 2 2 2 ...

$ habits1a : num NA NA NA NA NA NA NA NA NA NA ...

$ habits1b : num NA NA NA NA NA NA NA NA NA NA ...

$ habits1c : num NA NA NA NA NA NA NA NA NA NA ...

$ habits1d : num NA NA NA NA NA NA NA NA NA NA ...

$ habits1e : num NA NA NA NA NA NA NA NA NA NA ...

$ habits1f : num NA NA NA NA NA NA NA NA NA NA ...

$ habits2 : num NA NA NA NA NA NA NA NA NA NA ...

$ habits3 : num NA NA NA NA NA NA NA NA NA NA ...

$ habits4a : num NA NA NA NA NA NA NA NA NA NA ...

$ habits4b : num NA NA NA NA NA NA NA NA NA NA ...

$ habits4c : num NA NA NA NA NA NA NA NA NA NA ...

$ habits5 : num NA NA NA NA NA NA NA NA NA NA ...

$ habits6 : num NA NA NA NA NA NA NA NA NA NA ...

$ habits7 : num NA NA NA NA NA NA NA NA NA NA ...

$ habits8 : num NA NA NA NA NA NA NA NA NA NA ...

$ habits9 : num NA NA NA NA NA NA NA NA NA NA ...

$ habits10 : num NA NA NA NA NA NA NA NA NA NA ...

$ habits11 : num NA NA NA NA NA NA NA NA NA NA ...

$ habits12 : num NA NA NA NA NA NA NA NA NA NA ...

$ meanwifi : num NA NA NA NA NA NA NA NA NA NA ...

$ wifi1 : num NA NA NA NA NA NA NA NA NA NA ...

$ wifi2a : num NA NA NA NA NA NA NA NA NA NA ...

$ wifi2b : num NA NA NA NA NA NA NA NA NA NA ...

$ wifi2c : num NA NA NA NA NA NA NA NA NA NA ...

$ wifi2d : num NA NA NA NA NA NA NA NA NA NA ...

$ policy1 : num 8 8 8 8 2 2 8 8 8 2 ...

$ policy2a : num 8 8 8 8 1 1 8 2 8 2 ...

$ policy2b : num 8 8 3 8 1 1 8 8 8 2 ...

$ policy3 : num 8 8 3 8 4 3 8 8 8 4 ...

$ policy4 : num 8 8 8 8 3 2 8 8 8 4 ...

$ policy5 : num 8 8 8 8 3 4 8 8 8 4 ...

$ policy6a : num 3 3 3 3 3 3 3 3 3 8 ...

$ policy6b : num 3 3 2 3 8 3 3 3 3 8 ...

$ policy6c : num 3 8 2 3 2 2 3 3 2 2 ...

$ policy6d : num 3 3 3 3 3 3 3 3 3 3 ...

$ policy6e : num 3 3 3 3 8 3 3 3 2 3 ...

$ habits1cnt : num NA NA NA NA NA NA NA NA NA NA ...

$ iphoneuse : num 1 1 2 1 3 3 2 1 2 3 ...

$ hphoneuse : num 1 1 2 1 3 3 2 1 2 3 ...

$ recage : num 5 6 6 6 5 6 6 6 6 4 ...

$ receduc : num 1 1 1 1 1 1 3 3 1 1 ...

$ adults : num 1 1 2 1 3 1 1 1 1 2 ...

$ weight : num 5.91 5.24 2.47 4.47 3.53 ...

$ standwt : num 2.064 1.828 0.863 1.561 1.232 ...

$ sample\_\_1 : num 1 1 2 1 2 2 1 1 1 2 ...

The number of non-responses included in the dataset was concerning, so I researched several survey analysis websites to get ideas about how to deal with them. Initially, I thought of filling them by imputing the mean of the existing responses; however, I learned that in surveys a non-response can be as meaningful as a response. I kept them and noted how many missing observations were in various subsets. Considering that the purpose of the survey was to learn about cyber security, 4 of the 5 public WiFi questions contained more than 538 non-responses.

# Beginning Analysis

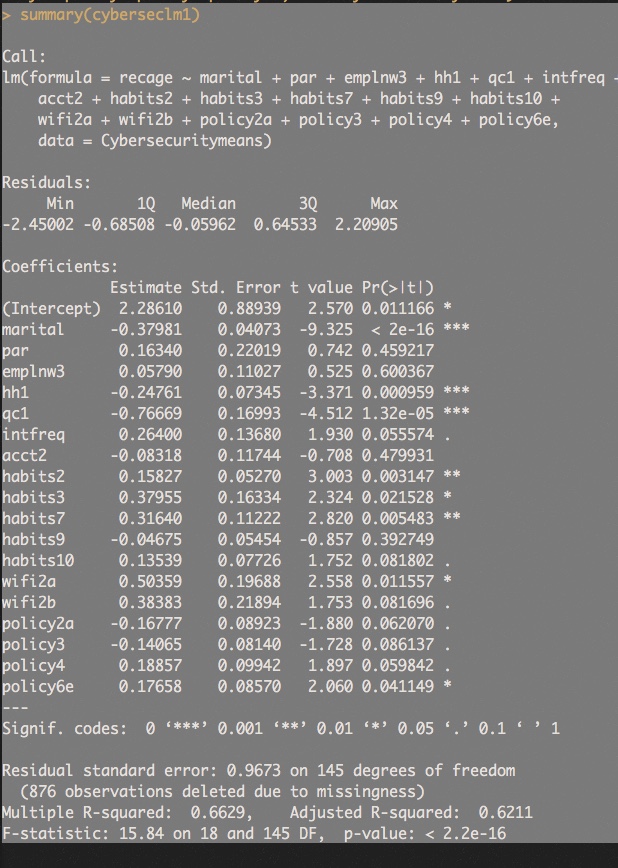
Demographic variables include sex, age, education, marital status, parental status, ideology, political affiliation, employment, race, income, number of people in the household, state, density and region.

Questions included:

* Type of mobile device used?
* Do you use social media like Facebook or Twitter?
* Do you bank online?
* Do you have online health care accounts?
* Do you have online utility bill pay?
* Do you have any type of online bill pay accounts?
* Have you chosen to not create an online account out of fear of its getting hacked?
* How confident are you that cell phone service companies are safe?
* How confident are you that cell phone manufacturers are safe?
* How confident are you that your email provider keeps your information safe?
* How confident are you that your social media sites are keeping your information safe?
* How confident are you that the federal government is keeping your information safe?
* How confident are you that your credit card companies are keeping your information safe?
* How confident are you that retailers and other companies keep your information safe?
* How safe is your information now as compared to 5 years ago?
* Have you received a notice that your social security number has been compromised?
* Have you received a notice that your personal information has been compromised?
* Have you noticed fraudulent charges on your credit card?
* Has someone taken over your email account?
* Has someone attempted to open a credit line using your name?
* Has someone taken over your social media accounts without permission?
* Has someone taken over your email account without your permission?
* Has someone tried to get a federal tax refund using your name?
* Do you keep track of passwords by memorizing them?
* Do you keep track of passwords by writing them on a piece of paper?
* Do you use a password management program?
* Do you save your passwords in a note or document on your pc or mobile device?
* Do you save your passwords in your internet browser?
* Do you keep track of passwords using another method?
* Do you have a hard time keeping track of your passwords?
* Do you worry about how secure your passwords are?
* Do you use less secure passwords because complicated ones are too difficult?
* Have you shared a password with a friend or family member?
* Do you use a two-factor or two-step authentication for any accounts?
* Do you have to use a code or password or other security feature to access your phone?
* What kind of security feature do you use to access your phone?
* How frequently do you update your os on your phone?
* Have you installed any virus protection apps on your smartphone?
* Do you ever access public WiFi in places like airports, cafes or hotels?
* Do you use social media while connected to public WiFi?
* Do you use email while connected to public WiFi?
* Should tech companies be able to use encryption that is unbreakable even to law enforcement?
* How likely do you think the US will experience a significant cyberattack on the banking or financial systems?
* How well-prepared do you think US businesses are to prevent cyberattacks?
* Have you heard about the publication of company emails at Sony?
* Have you heard exposure of government security clearance information at the Office of Personnel Management?
* Have you heard about the exposure of credit card data of customers that shopped at Target?
* Have you heard about the disruption of the power grid in the Ukraine?
* Which way do you do the most to keep track of your passwords?
* Are most of your passwords similar to each other or very different?
* Have you used your social media account information to log into another website?
* How frequently for you update your smartphone apps?
* Do you make online purchases while connected to public WiFi?
* Do you every do online banking or financial transactions while on public WiFi?
* How likely during the next 5 years do you think we will have a significant cyberattack on our public infrastructure?
* How well-prepared do you think the US government is to prevent cyberattacks on our government agencies?
* Have you heard about the publishing of identities of AshleyMadison.com customers?
* How many people live in your house?
* Do you make online purchases while connected to public WiFi?
* How often are you online?
* Land line working in home?

Significant variables are highlighted above in red. In the many models I tested I decided to add a variable that grouped ages into generations, i.e. Millenial, Gen-X, Young Boomer, Old Boomer, Silent Generation and GI Generation, to make graphics less cluttered since age ended up anchoring the final model. Although some of the variables appear to be insignificant, their removal resulted in lower results. (Below)

**Complete Dataset Summary**



# Predictions

>cyberseclm1 <- lm(recage~marital+par+emplnw3+hh1+qc1+intfreq+acct2+habits2+habits3+habits7+habits9+habits10+wifi2a+wifi2b+policy2a+policy3+policy4+policy6e, data=Cybersecuritymeans)

To start the final model analysis, the dataset was split into Train (708 observations) and Test (332 observations); it was a 70/30% split. In an effort to maximize the accuracy some of the variables were changed.

>train1 <- lm(recage~marital+hh1+qc1+intfreq+habits2+habits3+habits7+habits9+habits10+wifi2a+wifi2b+policy2a+policy3+policy4+policy6e, data=Train)

The R-squared is 0.6767 and adjusted R-squared is 0.6216. There are also 604 deleted observations due to missingness in our Train dataset.

We ran the predictions and applied them to the Test dataset. The R-squared is 0.7169 and adjusted R-squared is 0.6204. There are 272 deleted observations due to missingness (332 observations are in Test). The significance of some of the variables changed.

I ran multiple data checks including cor, Desc, RMSE, MAE, influence plots, outlier plots, leverage plots, added variable plots, Cook Level plots, QQ plots, ncv, chisq.test, heat tables, boxplots and histograms. I included the results in the accompanying presentation.

Next I tested the model using randomForest to check the variable importance. Marital status is the most important variable, followed by wifi2b. All of the variables are in the table below.

> train.rf <- randomForest(recage~marital+hh1+qc1+intfreq+habits2+habits3+habits7+habits9+habits10+wifi2a+wifi2b+policy2a+policy3+policy4+policy6e, data=Train, importance=TRUE, na.action=na.omit)

> importance(train.rf)

# %IncMSE IncNodePurity

#marital 33.698408 65.055547

#hh1 6.569315 14.208142

#qc1 5.240683 6.047324

#intfreq 4.862202 10.355221

#habits2 2.344607 9.907609

#habits3 3.292992 5.952409

#habits7 5.299240 14.235356

#habits9 2.283661 7.720894

#habits10 8.510213 16.963981

#wifi2a 3.886814 5.156294

#wifi2b 9.396419 7.356438

#policy2a 2.142870 7.764194

#policy3 3.128466 6.964825

#policy4 3.977840 11.957076

#policy6e 4.077428 10.107273

The randomForest variable importance was run. Note the significant changes in the percentages. Please remember that 272 of the 332 observations were removed due to missingness. To compare the two datasets the Train dataset has a Mean of Squared Residuals of 0.9826042 (Test = 1.420179), and 53.31% (Test=49.53%) of the variance was explained.

> test.rf <- randomForest(recage~marital+hh1+qc1+intfreq+habits2+habits3+habits7+habits9+habits10+wifi2a+wifi2b+policy2a+policy3+policy4+policy6e, data=Test, importance=TRUE, na.action=na.omit)

> importance(test.rf)

%IncMSE IncNodePurity

marital 16.3195228 24.274281

hh1 7.1841382 15.605476

qc1 12.5571876 18.781811

intfreq 5.9314829 6.800503

habits2 0.9483848 7.207831

habits3 1.7274607 6.293896

habits7 16.1477941 24.721924

habits9 -2.1308427 4.362490

habits10 8.4692570 13.607269

wifi2a -0.7367803 1.744641

wifi2b -0.8075705 1.688217

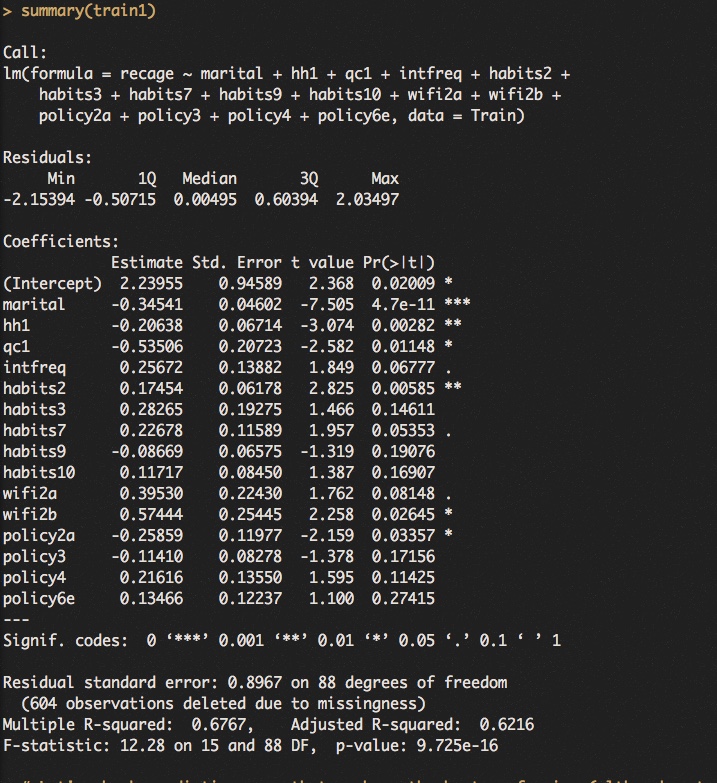
policy2a 1.5713233 5.631848

policy3 2.6335833 6.522630

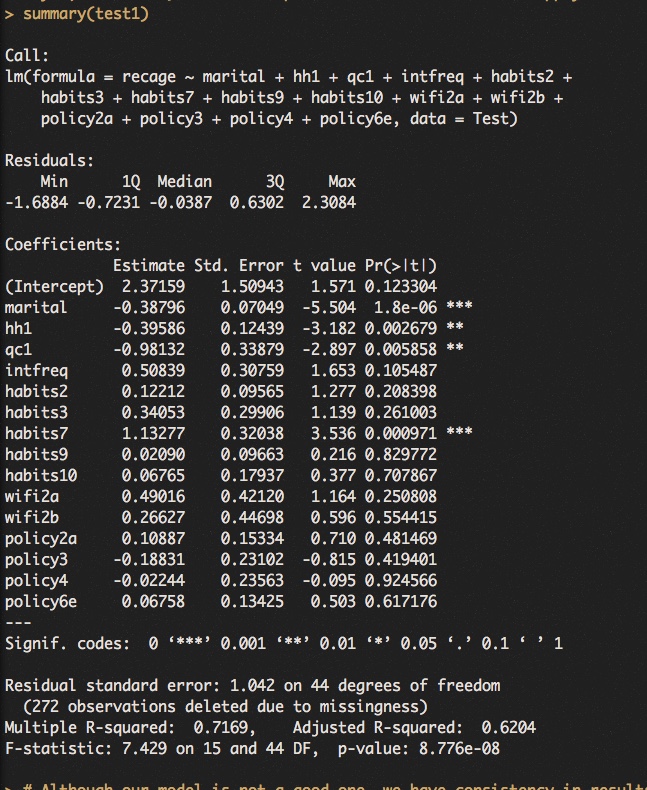
policy4 5.1734003 7.954957

policy6e 0.7300359 6.337147

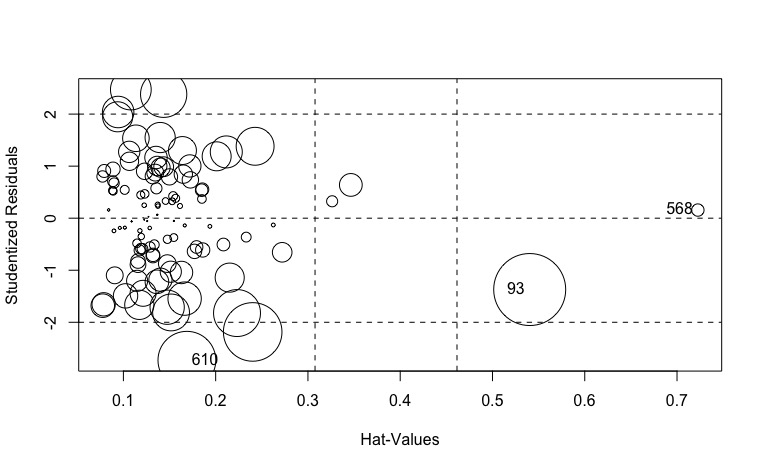
**Train Dataset Summary**



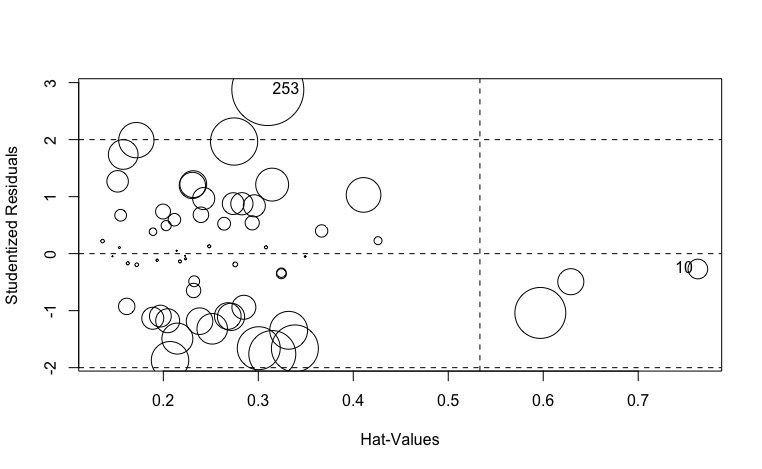
**Test Dataset Summary**



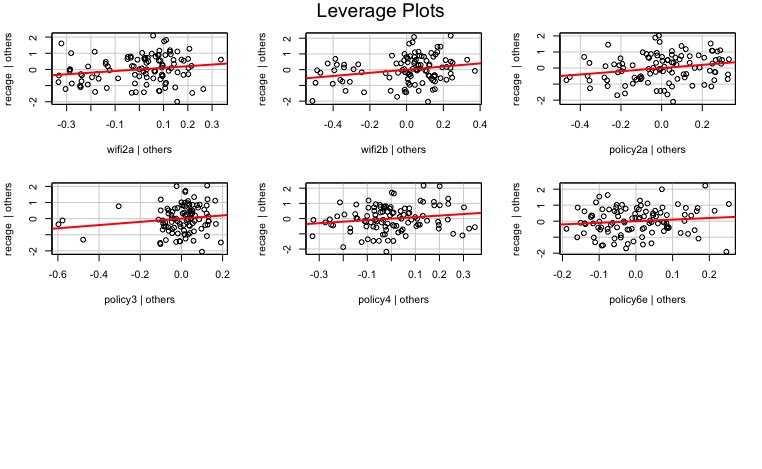
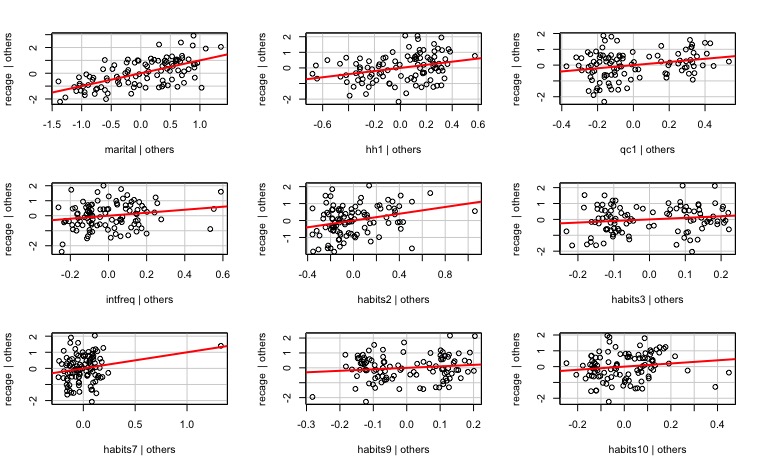
**Train Dataset Influence Plot**



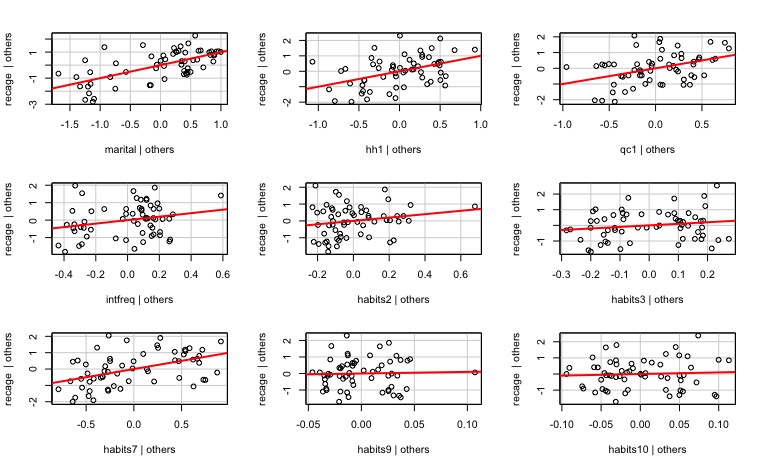
**Test Dataset Influence Plot**

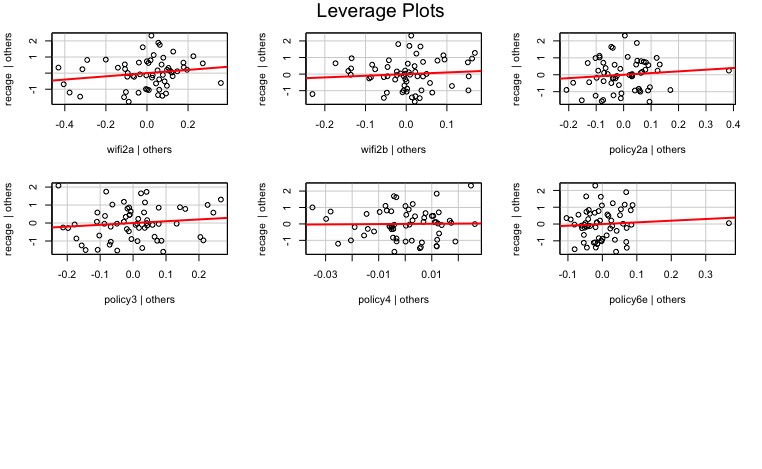


**Train Dataset Leverage Plots**



**Test Dataset Leverage Plots**





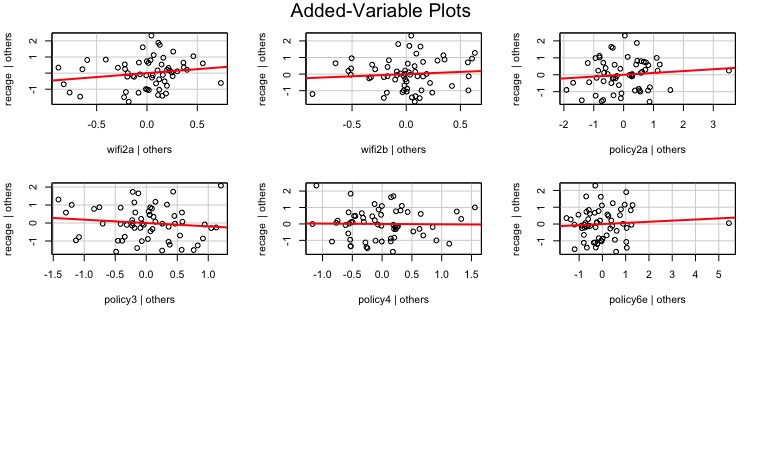
**Train Dataset Added Variables Plots**

# 

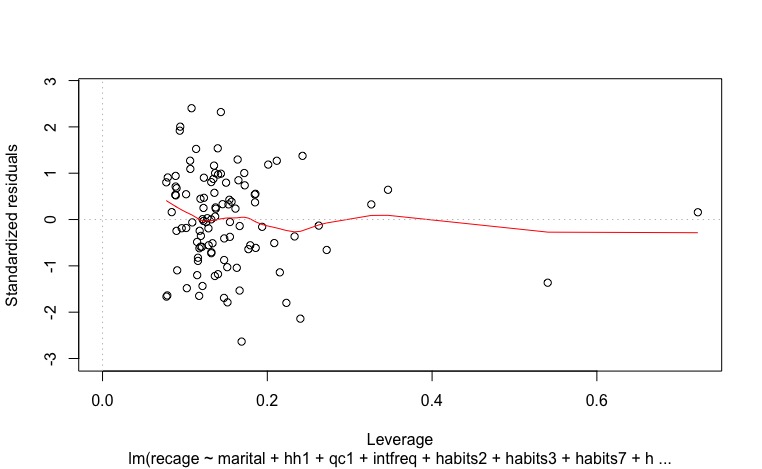
# 

**Test Dataset Added Variables Plots**

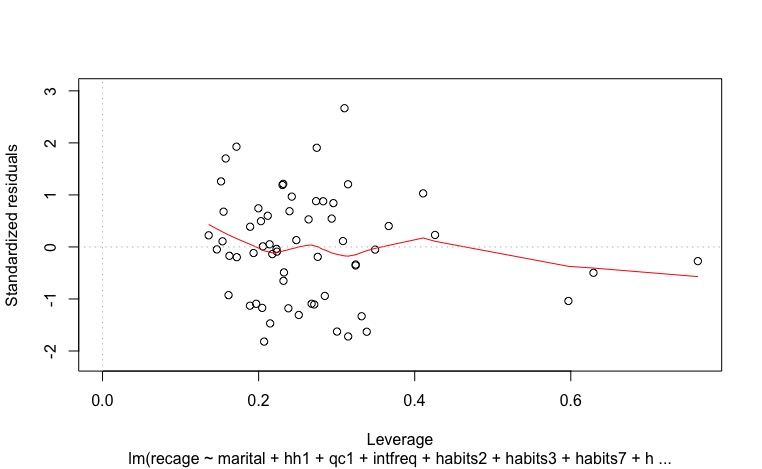
# 



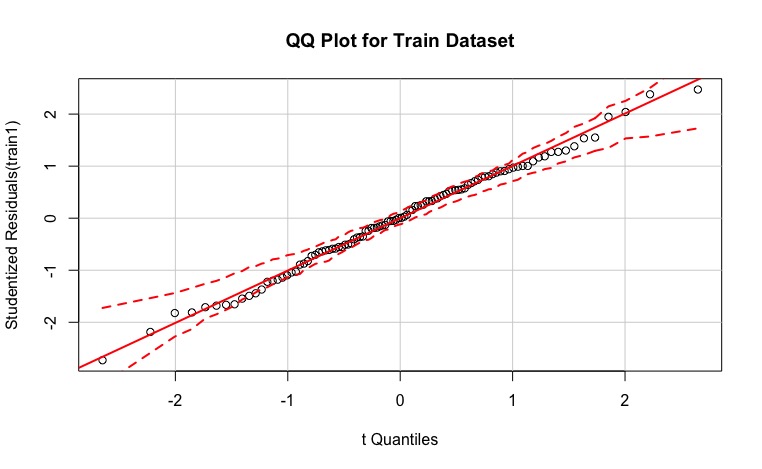
**Train Dataset Cook’s Test Plot**



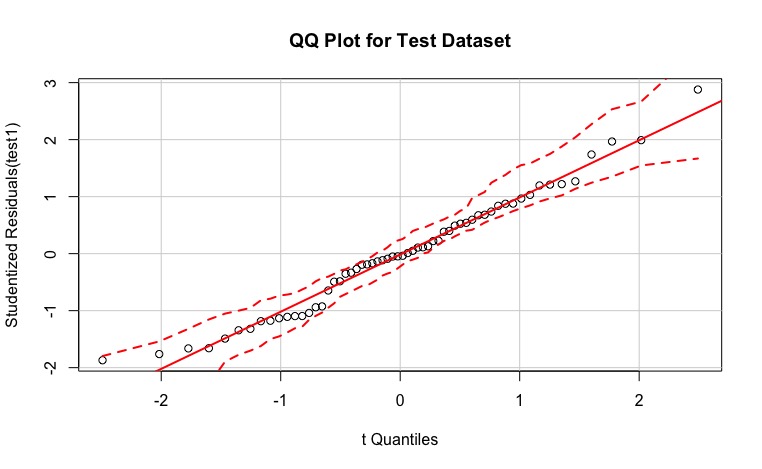
**Test Dataset Cook’s Test Plot**



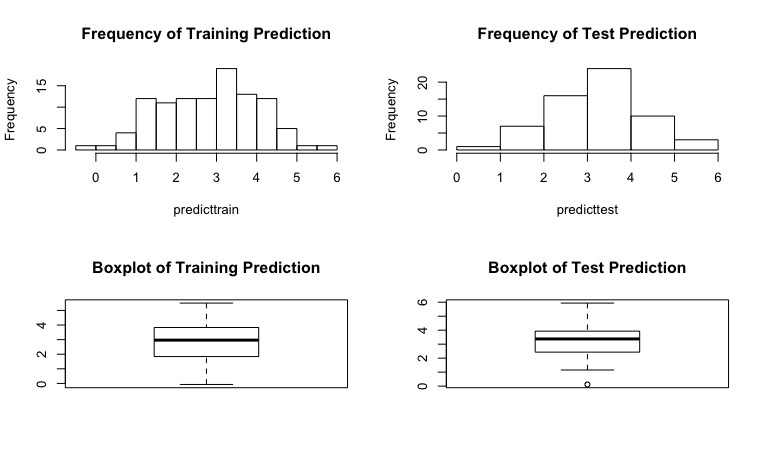
**Train Dataset QQ Plot**



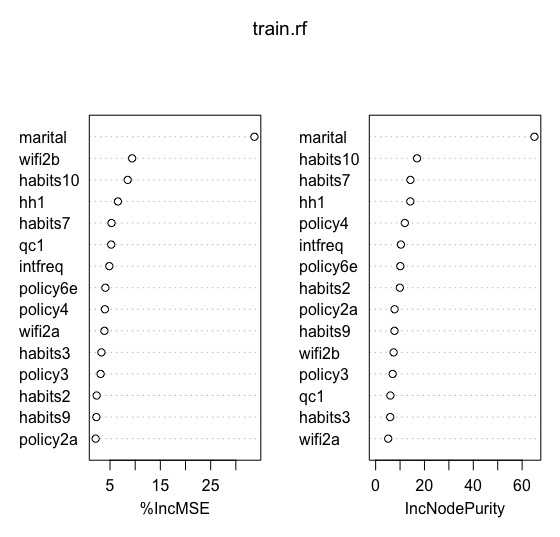
**Test Dataset QQ Plot**



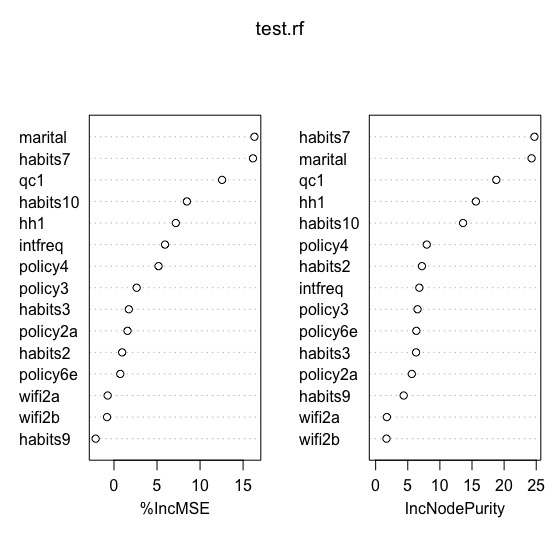
**Train and Test Prediction Plots**



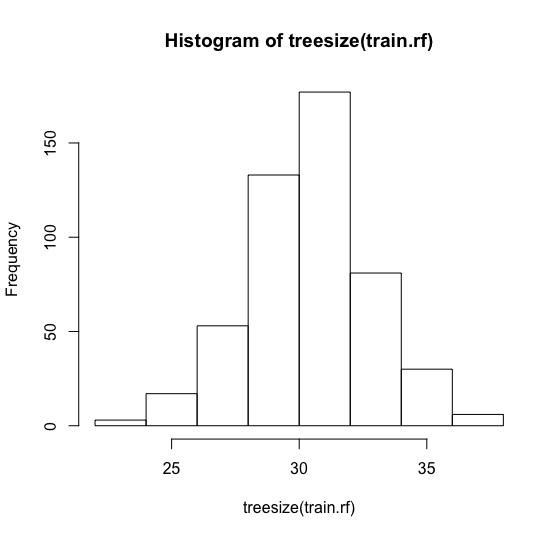
**Train RandomForest Plots**



**Test RandomForest Plots**



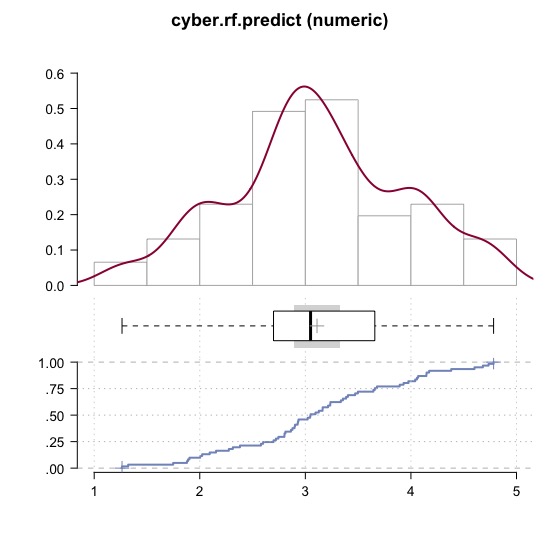
**Train Dataset Histogram of RandomForest Results**



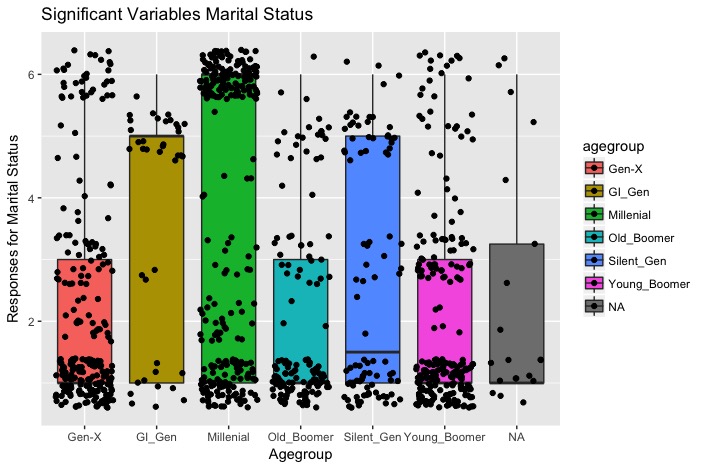
**Test Dataset Histogram of RandomForest Results**

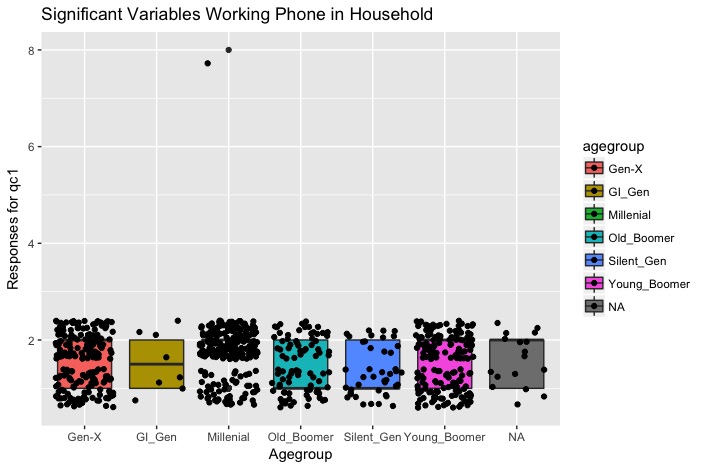


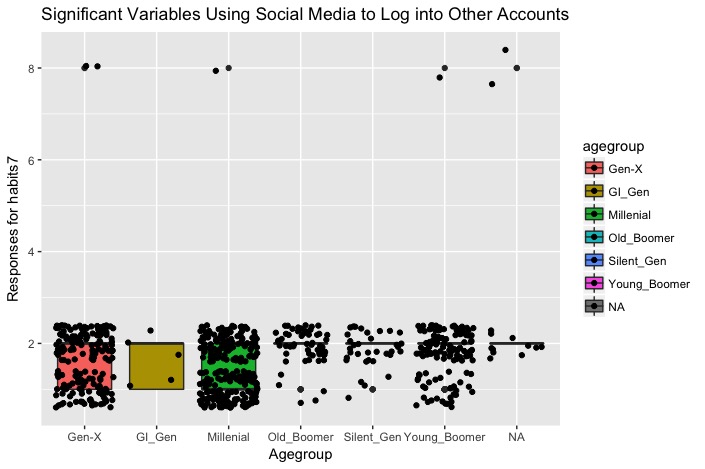
**RandomForest Prediction Plots**

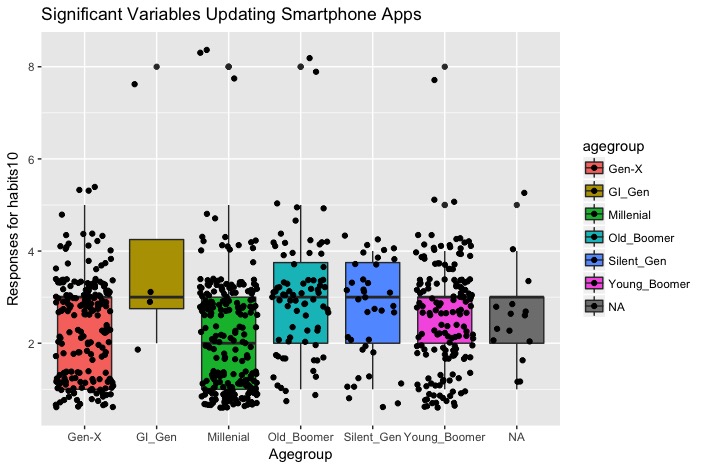


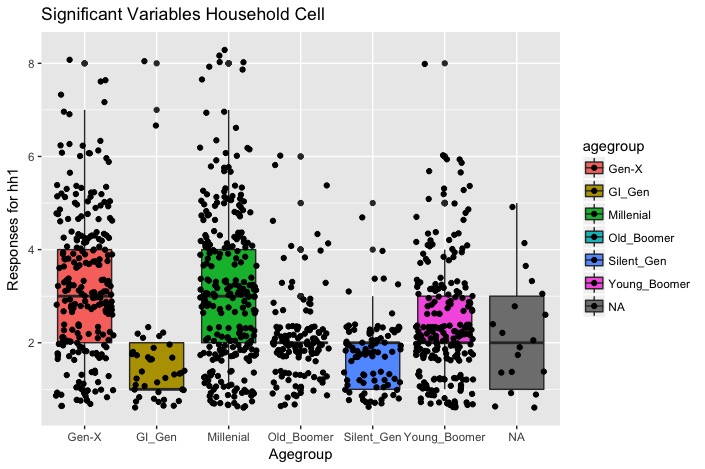
# Variable Level Plots by Age Ranges

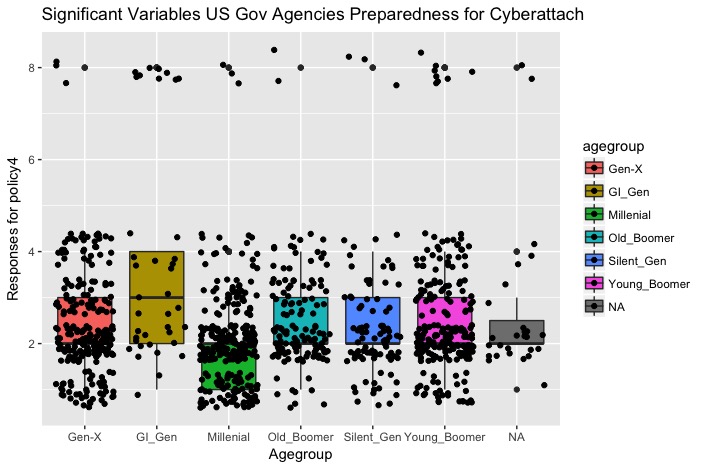


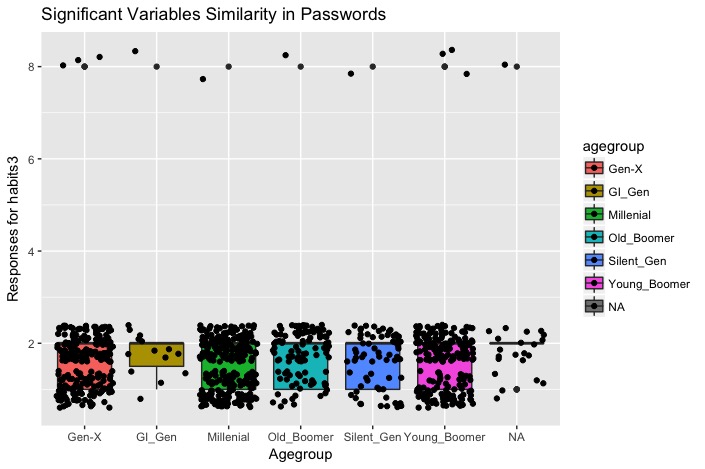












# 

# 

# Recommendations / Final Summary

I found two demographic variables that are not aligned with the US population – Race and Region. The South region is 39.2% of the total, which is higher than the last census. The other variable (Race) has 77% white participants, which is also higher than US population estimates. I cannot say that it affected the results or how much due to the number of non-responses in the dataset.

The result of my analysis shows that we need to focus on Millenial’s and Gen-X’ers since they trended together. If possible we can add Young Boomer’s as they trended with them but not with the same degree of frequency. These people are more active online. The password security questions show that approximately 3% of them use a password security application as of 2016, so there is a large market opportunity. These age groups have a range of income that can support a small payment for a secure application protecting their information. Very few of the subjects (<5%) have experienced a cyber leak, and many have not heard about breaches of some large corporations and governments.

After reviewing the important variables in the models that are more actionable, we can focus on the following traits:

* Marital Status – Divorced and Living with a Partner
* Working Land Line in Home – No
* Using Social Media Password to Access Another Website – No
* Smartphone App Update Preference – People updating their apps as soon as they become available.
* Number of People in Household – Between 2 and 3
* Commonality of Passwords – People using very different passwords between websites
* Online frequency – People that check the internet multiple times per day.
* Public Wifi Activities – People that do not make purchases or bank online while using public WiFi

I would recommend another survey focusing more on Millenial, Gen-X and Young Boomer generations. No questions in this survey asked about preferences, which should be included in the next survey. The current model is mediocre and needs to be improved with planned questions that will not allow as many non-responses as this survey includes. Adding numerical data will help with further quantitative analysis. We can focus on the more important variables listed above to get more detailed information.