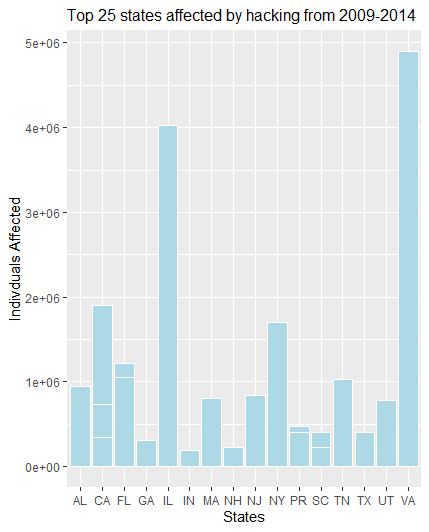
CIS 435 Project Question 3 of group 6

The ways I tried and analyzed the dataset is to construct a visual bar graph and base the top 25 states affected by hacking from 2009-2015 and in each bar for each state each hacking attempted is grouped for each individual incident. I then clean the data further by taking care of all things related to subcategories for theft improper disposal and so and changed the categories to just those types. Finally, I used the summary on the data for the type of breach to give a list of categories and how many of each of those cases occur in the time period of 2009 to 2014. The things I will still have to do is to figure out what the trend is for the entire dataset and not just the top 25 states and to see if there are different patterns in the states with lower affected individuals. According to the data seen in the graph there seems to be more individuals affected in a single event in the east coast and seem to have more individuals affected than any of the southern states such as New York state seems to have twice as many individuals affected by hacking than the highest southern state Tennessee. Another trend most of the states in the top 25 only are linked to a single hacking event which all individuals are affected by the only states not to a single event by hacking in the visualization below is California, Puerto Rico, South Carolina and Florida. Based on the graph below, these observations show exactly how hacking and other computer related incidents affect people in all different states from 2009 through 2014 and how they events could predicts what is to come in the united states.



Hacking/IT Incident

76

Improper Disposal

39

Loss

98

Other

91

Theft

571

Unauthorized Access/Disclosure

168

Unknown

12

top25 <-Projectdata[Projectdata$Individuals\_Affected >= 185000,]

ggplot(top25, aes(x=top25$State,y=top25$Individuals\_Affected)) + geom\_bar(stat = "identity",fill="lightblue",colour ="white",position = "dodge") + xlab("States") + ylab("Indivduals Affected") + ggtitle("Top 25 states affected by hacking from 2009-2014") + theme(plot.title = element\_text(size = 12))

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Loss, Unauthorized Access/Disclosure"] <- "Loss"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Loss, Improper Disposal"] <- "Loss

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach %in% c("Theft, Hacking/IT Incident","Theft, Improper Disposal, Unauthorized Access/Disclosure","Theft, Loss","Theft, Loss, Improper Disposal","Theft, Loss, Other","Theft, Loss, Unauthorized Access/Disclosure, Unknown","Theft, Other","Theft, Unauthorized Access/Disclosure","Theft, Unauthorized Access/Disclosure, Hacking/IT Incident"," Theft, Unauthorized Access/Disclosure, Other")] <- "Theft"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Theft, Unauthorized Access/Disclosure, Other"] <- "Theft"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach %in% c("Unauthorized Access/Disclosure, Hacking/IT Incident","Unauthorized Access/Disclosure, Hacking/IT Incident, Other","Unauthorized Access/Disclosure, Other")] <- "Unauthorized Access/Disclosure"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Unknown, Other"] <- "Unknown"

Projectdata$Type\_of\_Breach[Projectdata$Type\_of\_Breach == "Hacking/IT Incident, Other"] <- "Hacking/IT Incident"

"Improper Disposal, Unauthorized Access/Disclosure"] <- "Improper Disposal"

summary(Projectdata$Type\_of\_Breach)

AK AL AR AZ CA CO CT DC DE FL

187533 0 0 0 0 0 0 0 0 0 0

189489 0 0 0 0 1 0 0 0 0 0

228435 0 0 0 0 0 0 0 0 0 0

231400 0 0 0 0 0 0 0 0 0 0

277014 0 0 0 0 0 0 0 0 0 0

315000 0 0 0 0 0 0 0 0 0 0

338700 0 0 0 0 1 0 0 0 0 0

344579 0 0 0 0 0 0 0 0 0 0

398000 0 0 0 0 0 0 0 0 0 0

400000 0 0 0 0 0 0 0 0 0 0

405000 0 0 0 0 0 0 0 0 0 0

475000 0 0 0 0 0 0 0 0 0 0

514330 0 0 0 0 1 0 0 0 0 0

729000 0 0 0 0 1 0 0 0 0 0

780000 0 0 0 0 0 0 0 0 0 0

800000 0 0 0 0 0 0 0 0 0 0

839711 0 0 0 0 0 0 0 0 0 0

943434 0 1 0 0 0 0 0 0 0 0

10232090 0 0 0 0 0 0 0 0 0

GA HI IA ID IL IN KS KY LA MA

187533 0 0 0 0 0 1 0 0 0 0

189489 0 0 0 0 0 0 0 0 0 0

228435 0 0 0 0 0 0 0 0 0 0

231400 0 0 0 0 0 0 0 0 0 0

277014 0 0 0 0 0 0 0 0 0 0

315000 1 0 0 0 0 0 0 0 0 0

338700 0 0 0 0 0 0 0 0 0 0

344579 0 0 0 0 0 0 0 0 0 0

398000 0 0 0 0 0 0 0 0 0 0

400000 0 0 0 0 0 0 0 0 0 0

405000 0 0 0 0 0 0 0 0 0 0

475000 0 0 0 0 0 0 0 0 0 0

514330 0 0 0 0 0 0 0 0 0 0

729000 0 0 0 0 0 0 0 0 0 0

780000 0 0 0 0 0 0 0 0 0 0

800000 0 0 0 0 0 0 0 0 0 1

839711 0 0 0 0 0 0 0 0 0 0

943434 0 0 0 0 0 0 0 0 0 0

1023209 0 0 0 0 0 0 0 0 0 0

MD ME MI MN MO MS MT NC ND NE

187533 0 0 0 0 0 0 0 0 0 0

189489 0 0 0 0 0 0 0 0 0 0

228435 0 0 0 0 0 0 0 0 0 0

231400 0 0 0 0 0 0 0 0 0 0

277014 0 0 0 0 0 0 0 0 0 0

315000 0 0 0 0 0 0 0 0 0 0

338700 0 0 0 0 0 0 0 0 0 0

344579 0 0 0 0 0 0 0 0 0 0

398000 0 0 0 0 0 0 0 0 0 0

400000 0 0 0 0 0 0 0 0 0 0

405000 0 0 0 0 0 0 0 0 0 0

475000 0 0 0 0 0 0 0 0 0 0

514330 0 0 0 0 0 0 0 0 0 0

729000 0 0 0 0 0 0 0 0 0 0

780000 0 0 0 0 0 0 0 0 0 0

800000 0 0 0 0 0 0 0 0 0 0

839711 0 0 0 0 0 0 0 0 0 0

943434 0 0 0 0 0 0 0 0 0 0

1023209 0 0 0 0 0 0 0 0 0 0

NH NJ NM NV NY OH OK OR PA PR

187533 0 0 0 0 0 0 0 0 0 0

189489 0 0 0 0 0 0 0 0 0 0

228435 0 0 0 0 0 0 0 0 0 0

231400 1 0 0 0 0 0 0 0 0 0

277014 0 0 0 0 0 0 0 0 0 0

315000 0 0 0 0 0 0 0 0 0 0

338700 0 0 0 0 0 0 0 0 0 0

344579 0 0 0 0 1 0 0 0 0 0

398000 0 0 0 0 0 0 0 0 0 1

400000 0 0 0 0 0 0 0 0 0 0

405000 0 0 0 0 0 0 0 0 0 0

475000 0 0 0 0 0 0 0 0 0 1

514330 0 0 0 0 0 0 0 0 0 0

729000 0 0 0 0 0 0 0 0 0 0

780000 0 0 0 0 0 0 0 0 0 0

800000 0 0 0 0 0 0 0 0 0 0

839711 0 1 0 0 0 0 0 0 0 0

943434 0 0 0 0 0 0 0 0 0 0

1023209 0 0 0 0 0 0 0 0 0 0

RI SC SD TN TX UT VA VT WA WI

187533 0 0 0 0 0 0 0 0 0 0

189489 0 0 0 0 0 0 0 0 0 0

228435 0 1 0 0 0 0 0 0 0 0

231400 0 0 0 0 0 0 0 0 0 0

277014 0 0 0 0 1 0 0 0 0 0

315000 0 0 0 0 0 0 0 0 0 0

338700 0 0 0 0 0 0 0 0 0 0

344579 0 0 0 0 0 0 0 0 0 0

398000 0 0 0 0 0 0 0 0 0 0

400000 0 1 0 0 0 0 0 0 0 0

405000 0 0 0 0 1 0 0 0 0 0

475000 0 0 0 0 0 0 0 0 0 0

514330 0 0 0 0 0 0 0 0 0 0

729000 0 0 0 0 0 0 0 0 0 0

780000 0 0 0 0 0 1 0 0 0 0

800000 0 0 0 0 0 0 0 0 0 0

839711 0 0 0 0 0 0 0 0 0 0

943434 0 0 0 0 0 0 0 0 0 0

1023209 0 0 0 1 0 0 0 0 0 0

WV WY

187533 0 0

189489 0 0

228435 0 0

231400 0 0

277014 0 0

315000 0 0

338700 0 0

344579 0 0

398000 0 0

400000 0 0

405000 0 0

475000 0 0

514330 0 0

729000 0 0

780000 0 0

800000 0 0

839711 0 0

943434 0 0

1023209 0 0

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5372 0.2854 0.2871 0.2876 0.2880

Coefficients:

Estimate Std. Error z value

(Intercept) 3.158e+00 1.346e+00 2.346

top25$Individuals\_Affected 2.147e-08 9.290e-07 0.023

Pr(>|z|)

(Intercept) 0.019 \*

top25$Individuals\_Affected 0.982

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8.3972 on 24 degrees of freedom

Residual deviance: 8.3967 on 23 degrees of freedom

AIC: 12.397

Number of Fisher Scoring iterations: 6

One of the other ways which I tried to analyze the data was to compare the top 25 states which were affected by hacking from 2009-2014 to all of the states in our dataset and by using a confusion matrix determine if they are statistically significant and determine if data can be displayed in more effective way. By looking at the prediction model of the confusion matrix you can clearly see based on all the states in the original data set and the values of the top 25 dataset it can be shown where each time in the matrix shows when major events occur in each state based on the individuals affected compared between both datasets. For each state apart of the top 25 there will be a 1 in the matrix for those affected and those not in the list will show a zero based on the individuals affected from the data. One of the challenges which I had faced while creating this confusion matrix was determine how to set it up properly and if there was any statistical significance to our original data set. I have also used a generalized linear model to show the different statistical terms we have discussed to further interpret the data. Based on that the null and residual deviance are low values of 8.3972 and 8.3967 meaning they would have strong goodness of fit if this data were to be placed into a model as well as low standard error of 9.290e-07 which means high statistical significance for the individuals affected based on the given states. In conclusion, using these different models shows just how data analysis can be applied and can be used to represent the data different forms and give different conclusions based on model and methods used.

glm(formula = top25$State ~ top25$Individuals\_Affected, family = "binomial” data = top25)

summary(logistic)

predict <- predict(logistic,Projectdata,type = 'response')

table\_mat <- table(top25$Individuals\_Affected,top25$State,predict >0.5)

table\_mat