Cybersecurity *with* Statistical Programming

By: Camilla Ma, Haiman Wong, and Chen Zhang

April 29th, 2021

TABLE OF CONTENTS

1

PROBLEM + MOTIVATION

Employ statistical methodology and programming techniques to identify observable trends and vulnerabilities that organizations and nation-states face during cyber-attacks

Cyberincidentnum	Dyadpair	StateA	StateB	Name	interactionstartdate	interactionenddate	interactiontype	method	APT	targettype	initiator	cyber_objective
1	2365	US	Russia	Regin malware campaign	2/1/08	3/1/11	3	3	1	2	2	3
2	2365	US	Russia	QWERTY keystroke log	2/1/08	3/11/11	3	4.4	1	2	2	2
3	2365	US	Russia	Duke Series	4/8/08	9/17/15	3	4.2	1	2	365	3
4	2365	US	Russia	US govt employee in Georgia hacked	8/6/2008	8/12/2008	1	4.2	0	2	365	1
5	2365	US	Russia	Agent.bz/CENTCOM (linked to APT 28)	10/1/08	10/15/08	3	3	0	3	365	3
6	2365	US	Russia	Buckshot Yankee	11/26/08	11/28/08	2	4.2	0	2	2	4
7	2365	US	Russia	Sandworm	1/1/09	10/14/14	3	3	0	2	365	3
8	2365	US	Russia	Power grid hacked, traced to Russia	8/24/2009	8/24/2009	1	4.2	0	1	365	4
9	2365	US	Russia	Energetic Bear/Dragonfly/Crouching Yeti	1/1/11	7/1/14	3	3	1	1	365	2
10	2365	US	Russia	Yahoo breach 2	8/1/13	12/15/16	1	3	0	1	365	3
11	2365	US	Russia	Operation Pawn Storm/World Doping Agency	9/30/13	10/22/14	1	3	1	2	365	2
12	2365	US	Russia	CyberBerkut NATO Websites	3/15/14	3/26/14	1	2	1	3	365	1
13	2365	US	Russia	Operation Pawn Storm: military networks (fake OWA)	6/2/14	2/1/15	3	3	1	3	365	2
14	2365	US	Russia	Operation Pawn Storm: Nuclear power plants, newspapers	6/3/14	12/1/14	3	3	1	1	365	2
15	2365	US	Russia	US Banks hacked	6/4/14	7/8/14	1	3	1	1	365	1
16	2365	US	Russia	White House hack	10/26/14	10/28/14	1	3	0	2	365	2
17	2365	US	Russia	State Dept hack	11/15/14	11/17/14	1	3	0	2	365	2
18	2365	US	Russia	Yahoo breach 1	11/22/14	9/22/16	1	3	0	1	365	3
19	2365	US	Russia	DoD breach	3/1/15	3/15/15	1	3	0	3	365	3
20	2365	US	Russia	2016 Presidential Election/FSB/APT29	6/15/15	11/8/16	3	3	1	1	365	2
21	2365	US	Russia	JCS network breach	6/26/15	6/28/15	1	3	0	3	365	3
22	2365	US	Russia	ProjectSauron	9/2/15	8/9/16	3	3	1	1	2	3
23	2365	US	Russia	ProjectSauron	9/2/15	8/9/16	3	3	1	3	2	3
24	2365	US	Russia	2016 Presidential Election/GRU/APT28/Guccifer 2.0	4/3/16	11/8/16	3	3	1	1	365	2

02. THE DATA


ddos
launches
chinese
hacks
response
south
japanese
companies
energy
breached
contractors
breaches
attacks
korea
banks
sk
dept
info
scs
oil
action
cyber
hack
access
china
backdoor
oilrig
espionage
camaping
apt
botnet
company
disruptive
including
govt
breach
kim
security
defacement
hackers
korean
iran
sector

03. STATISTICAL METHODOLOGY

DATA VISUALIZATION

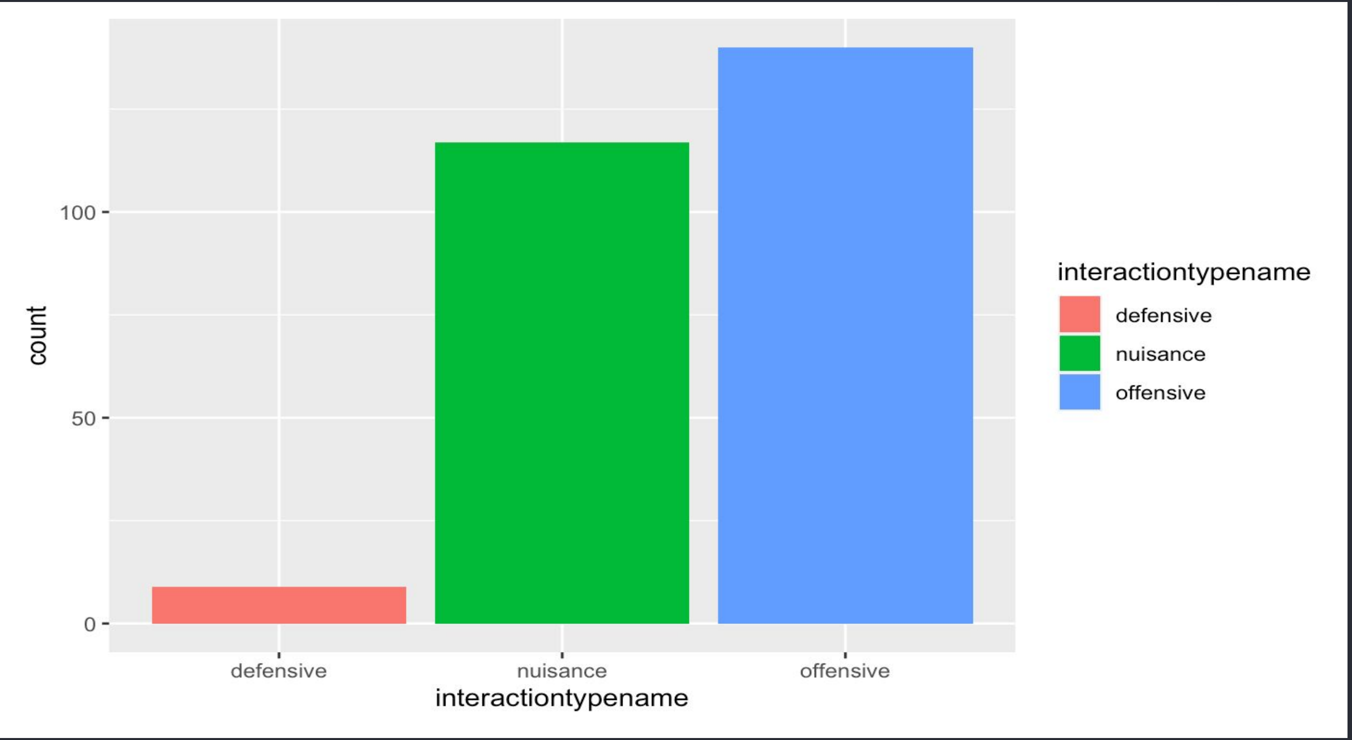
Bar Graphs, Pie Charts, Facet Plots, and Scatter Plots

REGRESSION

Multiple and Linear Regression

ANOVA

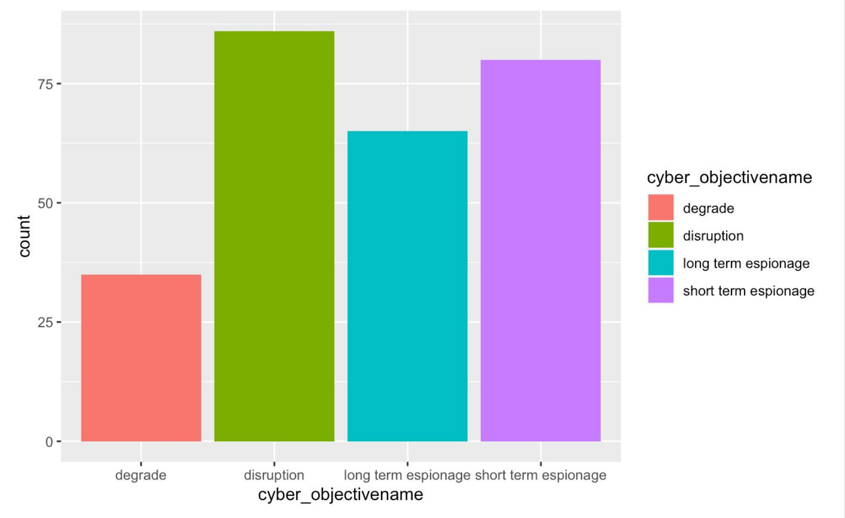
Check whether there are differences between two models



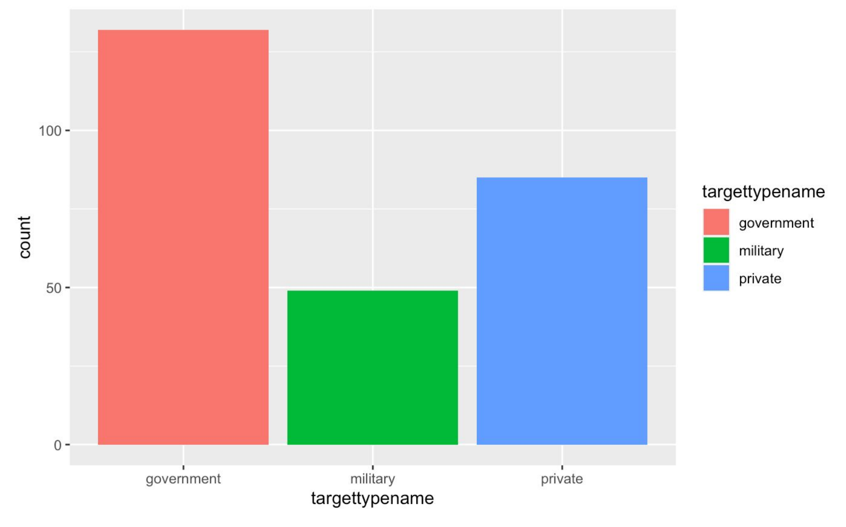
100
50
0
count
defensive
nuisance
offensive
interactiontypename
interactiontypename
defensive
nuisance
offensive

04. RESULTS

04. RESULTS

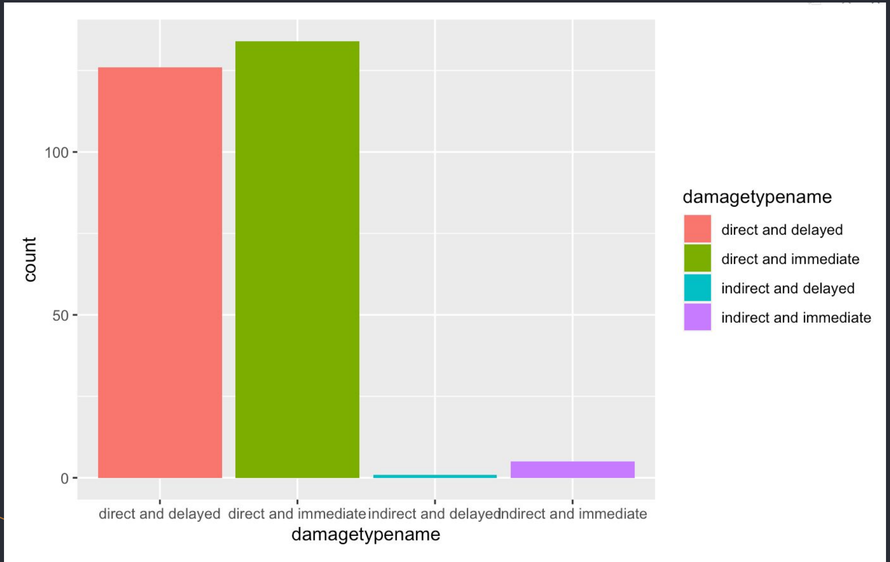


count
75
50
25
0
degrade
disruption
long term espionage
short term espionage
cyber_objectivename
cyber_objectivename
degrade
disruption
long term espionage
short term espionage

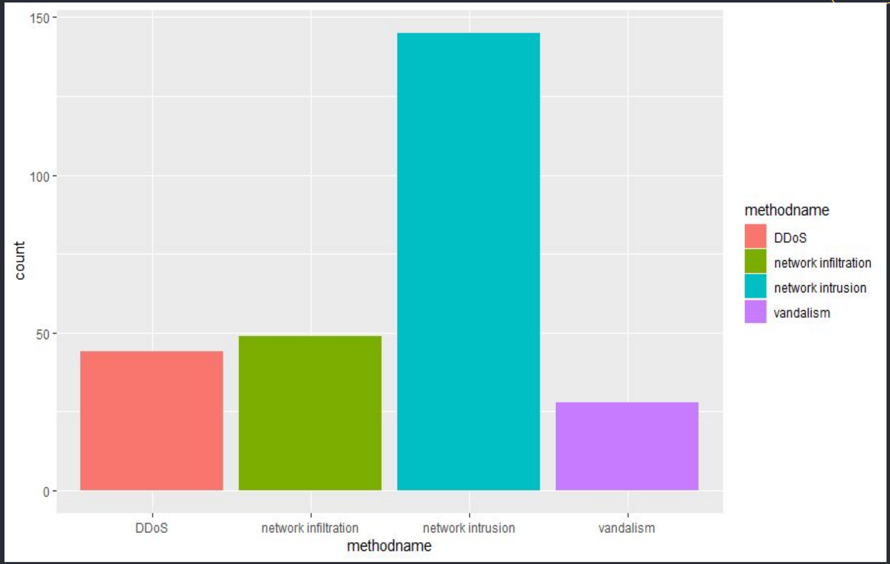


count
100
50
0
government
military
private
targettypename
targettypename
government
military
private

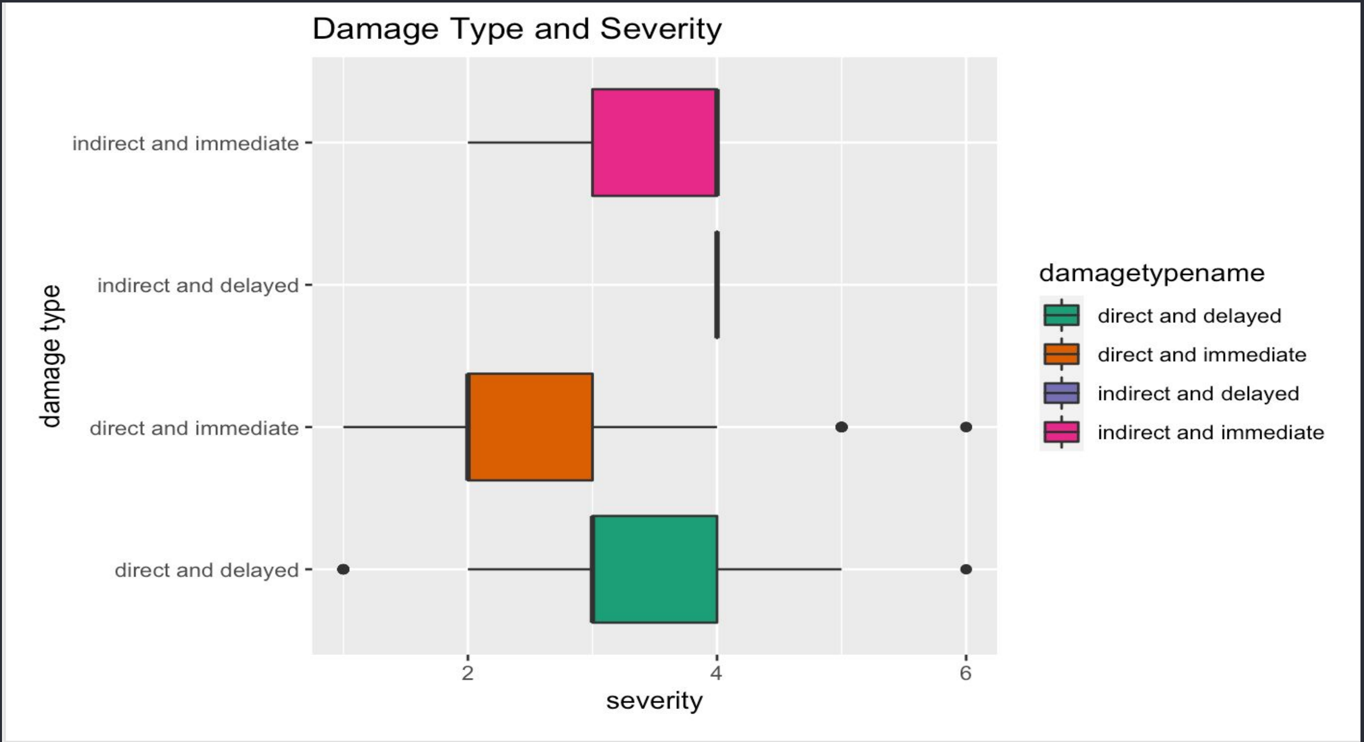
04. RESULTS



100
50
0
count
direct and delayed
direct and immediate
indirect and delayed
indirect and immediate
damagetypename
damagetypename
direct and delayed
direct and immediate
indirect and delayed
indirect and immediate



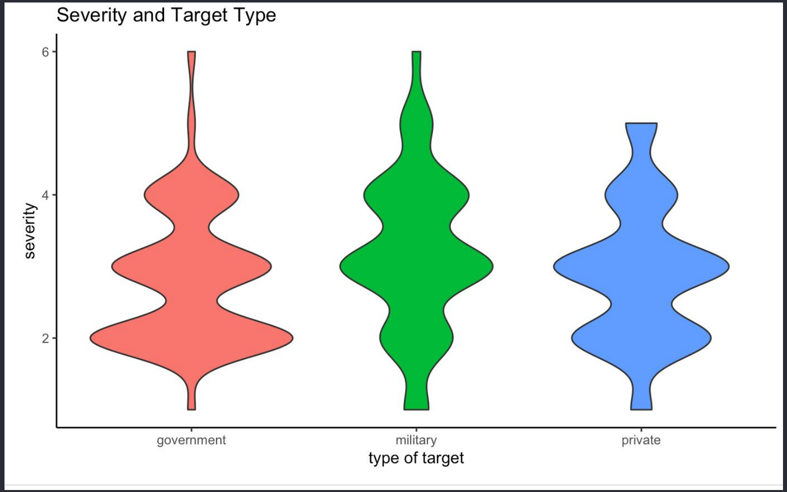
150
100
50
0
count
DDoS
network infiltration
network intrusion
vandalism
methodname
methodname
DDoS
network infiltration
network intrusion
vandalism

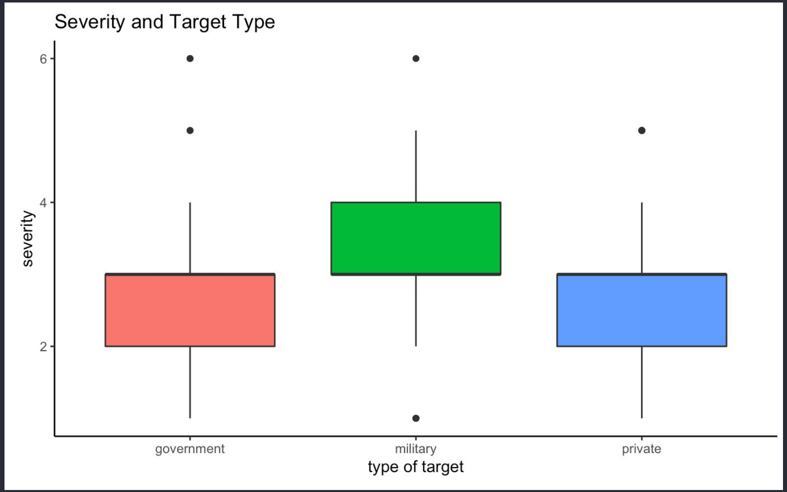


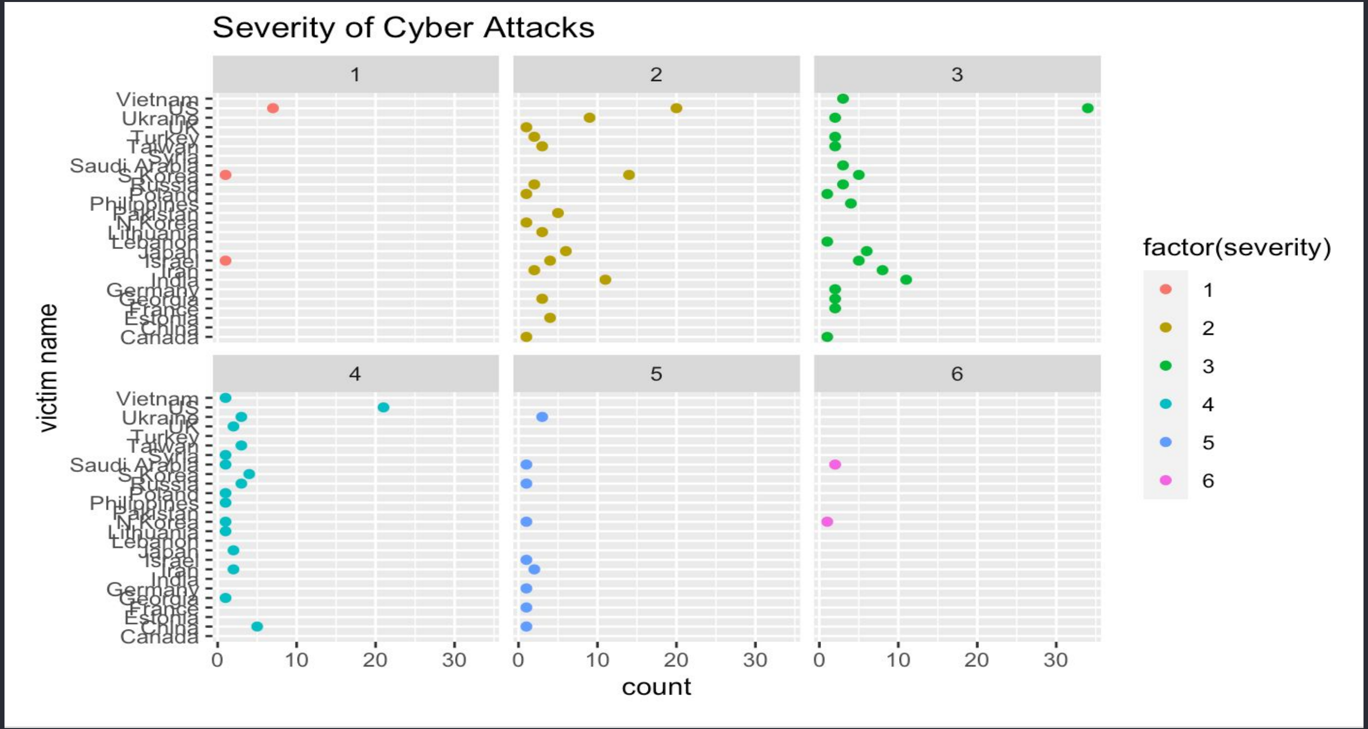
Damage Type and Severity
indirect and immediate
indirect and delayed
direct and immediate
direct and delayed
damage type
2
4
6
severity
damagetypename
direct and delayed
direct and immediate
indirect and delayed
indirect and immediate

04. RESULTS

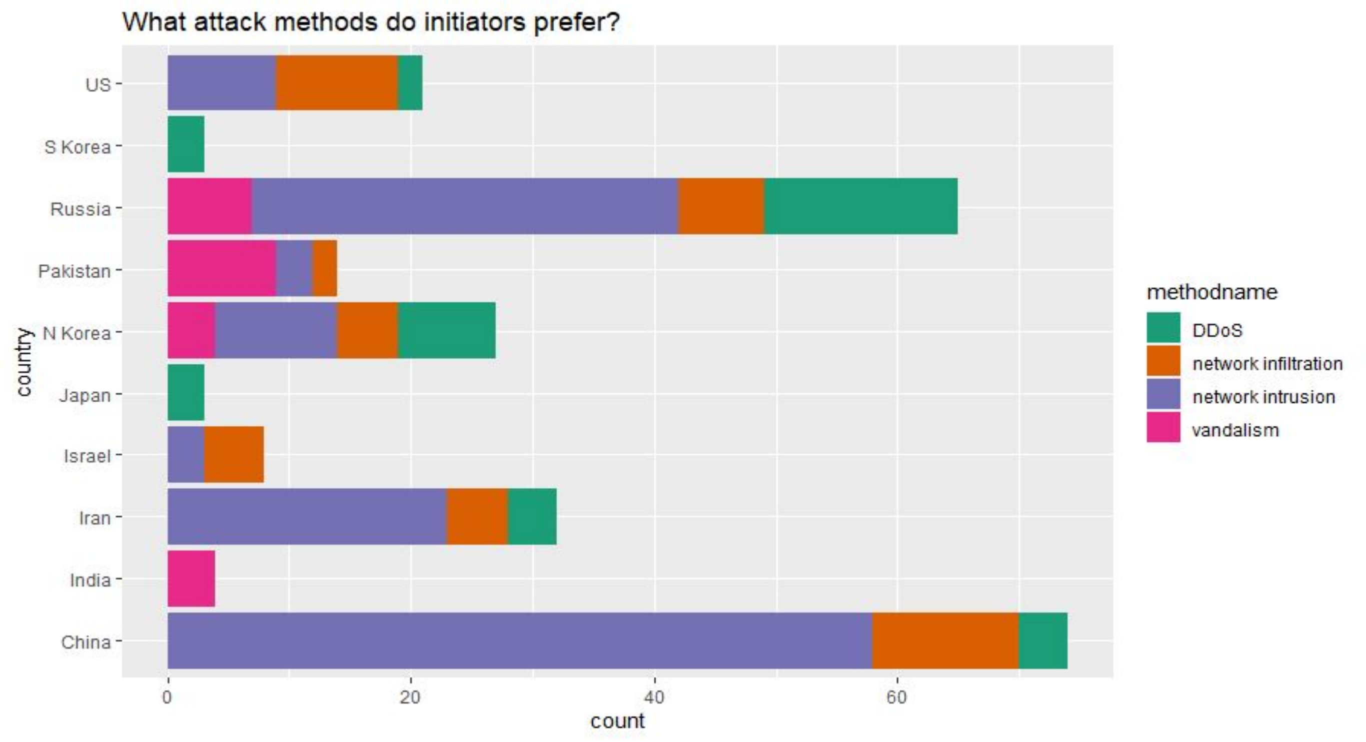
04. RESULTS


Severity and Target Type
6
4
2
severity
government
military
private
type of target


Severity and Target Type
6
4
2
severity
government
military
private
type of target

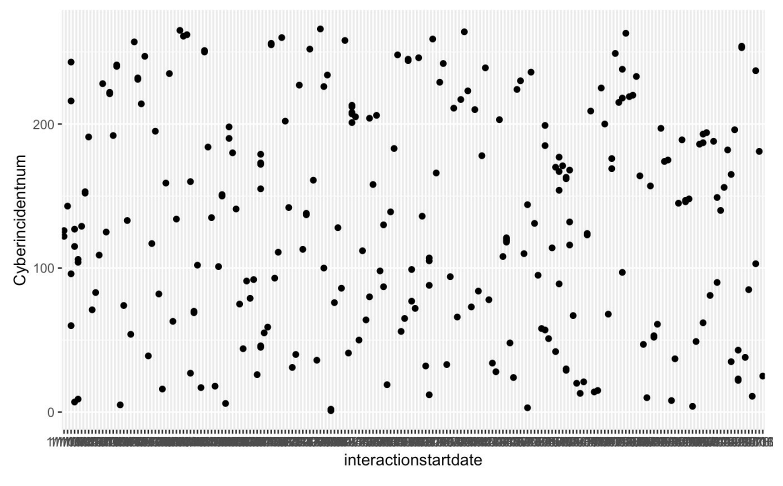

Severity of Cyber Attacks
1
2
3
4
5
6
victim name
Vietnam
US
Ukraine
UK
Turkey
Taiwan
Syria
Saudi Arabia
S.Korea
Russia
Poland
Philippines
Pakistan
N.Korea
Lithuania
Lebanon
Japan
Israel
Iran
India
Germany
Georgia
France
Estonia
China
Canada
0
10
20
30
count
factor(severity)
1
2
3
4
5
6

04. RESULTS

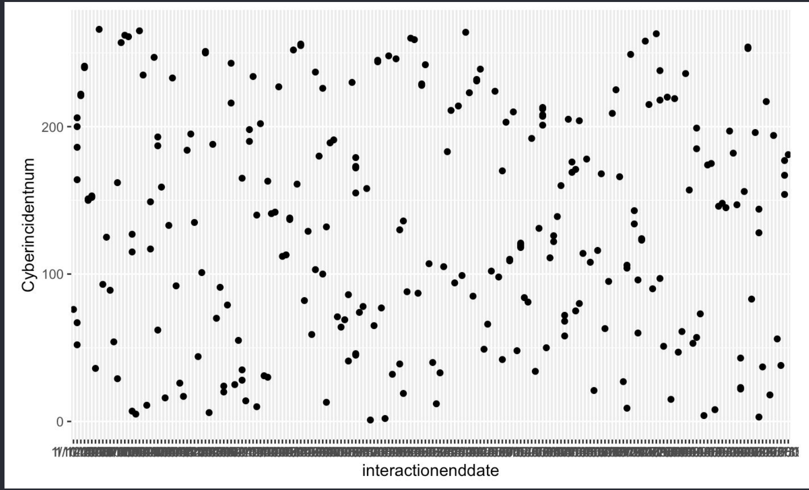

What attack methods do initiators prefer?
US
S Korea
Russia
Pakistan
N Korea
Japan
Israel
Iran
India
China
country
0
20
40
60
count
methodname
DDoS
network infiltration
network intrusion
vandalism

04. RESULTS

04. RESULTS

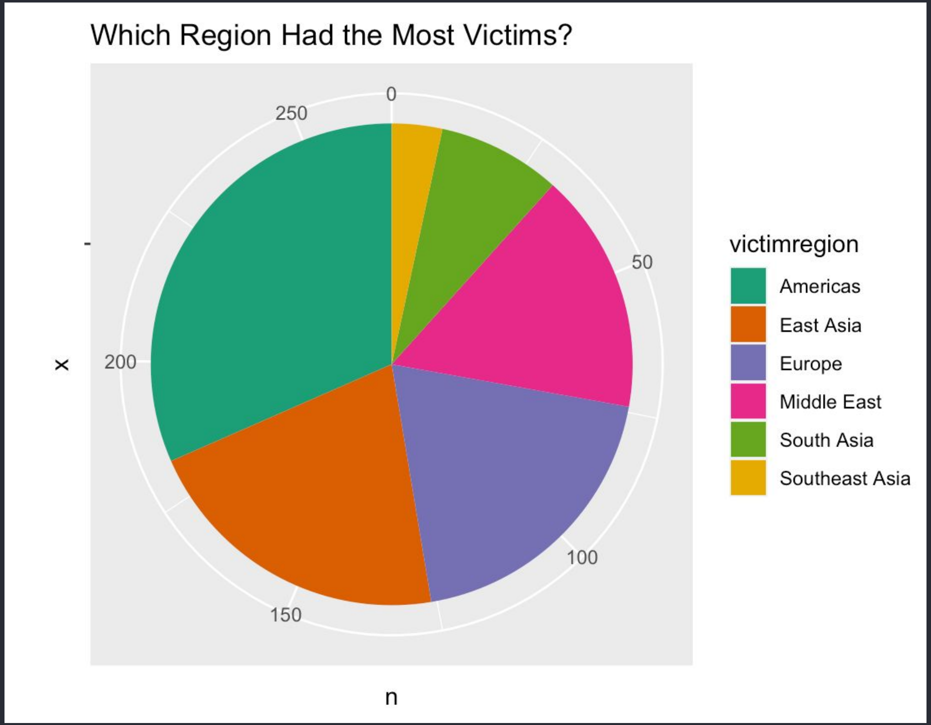


Cyberincidentnum
200
100
0
interactionstartdate



Cyberincidentnum
200
100
0
interactionenddate

04. RESULTS



Which Region Had the Most Victims?
0
50
100
150
200
250
x
n
victimregion
Americas
East Asia
Europe
Middle East
South Asia
Southeast Asia

04. RESULTS

Simple Linear Regression Model 1:

Y(hat) =
104.5+94.5(China)+5.5(France)+8.5(Germany)+152(India)+62.5(Iran)+77.5(Lebanon)+122.5(NKorea)+11.5(Poland)+31(Russia)+138(SKorea)+76.5(Syria)+2.5(UK)+(-52.5)(US)

Coefficients	Estimate	Std. Error	P-Value
Intercept	104.50	14.18	2.42e-12
N Korea	122.50	14.78	6.90e-15
China	94.50	14.60	4.99e-10
S Korea	138.00	15.85	4.19e-16
Iran	62.50	14.69	2.96e-05

04. RESULTS

Simple Linear Regression Model 2:

Y(hat) = 1.0775 + 0.1568(severity)

Coefficients	**Estimate**	**Std. Error**	**P-Value**
Intercept	1.07746	0.10734	<2e-16
Severity	0.15684	0.03497	1.09e-05

04. RESULTS

Simple Linear Regression Model 3: **Y(hat) = 730.06 + (-77.32)(cyber_objective)**			
Coefficients	**Estimate**	**Std. Error**	**P-Value**
Intercept	730.06	29.83	<2e-16
Severity	-77.32	12.35	1.56e-09

04. RESULTS

Multiple Linear Regression Model 1:

Y(hat) = 45.8066+0.1868(initiator)+ 0.5992(cyber_objective)+ (-12.4713)(damagetype)+0.2406(severity)

Coefficients	**Estimate**	**Std. Error**	**P-Value**
Intercept	45.8066	21.2165	0.0318
Initiator	0.1868	0.0192	<2e-16
Cyber_Objective	0.5992	4.8096	0.9010
DamageType	-12.4713	7.4133	0.0937
Severity	0.2406	4.7666	0.9598

04. RESULTS

Multiple Linear Regression Model 2: Y(hat) = 808.92+(-121.00)(military)+(-49.67)(private)+(-42.42)(damagetype)+(-49.67)(severity)			
Coefficients	**Estimate**	**Std. Error**	**P-Value**
Intercept	808.92	48.52	<2e-16
Military	-121.00	35.48	0.000751
Private	-49.67	29.00	0.087940
DamageType	-42.42	23.30	0.069815
Severity	-49.67	13.63	0.000322

04. RESULTS

Multiple Linear Regression Model 3: Y(hat) = 1.6340954+(-0.0008396)(initiator)+0.4420261(method)+0.2355995(interactiontype)			
Coefficients	**Estimate**	**Std. Error**	**P-Value**
Intercept	1.6340954	0.2214429	<2.11e-12
Military	-0.0008396	0.0002196	0.000164
Private	0.4420261	0.0560545	8.47e-14
DamageType	0.2355995	0.0525453	1.10e-05

04. RESULTS

Suppose APT (Whether or not the incident is considered an advanced persistent threat) is Y variable
(The dependent variable of logistic regression model is the variable of yes or no)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.25025	1.27984	-4.884	1.04e-06	***
interactiontype	0.67807	0.21076	3.217	0.00129	**
severity	1.41664	0.26682	5.309	1.10e-07	***
targettype	-0.35124	0.26001	-1.351	0.17673	
cyber_objective	-0.08567	0.24162	-0.355	0.72290	
information_operation	1.10474	0.53991	2.046	0.04074	*
objective_achievement	0.02789	0.69673	0.040	0.96807	
Concession	-2.36964	0.94667	-2.503	0.01231	*
X3rdpartyinitator	-1.59790	0.69667	-2.294	0.02181	*
X3rdparty.target	2.29770	0.47835	4.803	1.56e-06	***
govtstatement	-0.77826	0.28797	-2.703	0.00688	**
damage.type	1.06950	0.38523	2.776	0.00550	**

04. RESULTS

Optimization this model

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -6.7645     1.0846  -6.237 4.46e-10 ***
interactiontype           0.6284     0.2006   3.132  0.00174 ** 
severity                  1.3876     0.2503   5.544 2.95e-08 ***
information_operation     1.0855     0.5052   2.149  0.03166 *  
Concession               -2.5646     0.9245  -2.774  0.00554 ** 
X3rdpartyinitator        -1.7259     0.6936  -2.488  0.01283 *  
X3rdparty.target          2.1866     0.4637   4.716 2.41e-06 ***
govtstatement            -0.7582     0.2845  -2.665  0.00769 ** 
damage.type               1.0164     0.3812   2.667  0.00766 ** 
---
```

04. RESULTS

Using the anova test: Check whether there are differences between the two models

The anova test of the two models before and after optimization, to see whether there is a difference between the two models before and after optimization, because P = 0.5918 > 0.05, so the two models are still different(Chi sq test for logistic regression)

anova(cs.ful, cs.reduced, test = "Chisq")

```
Analysis of Deviance Table

Model 1: APT ~ interactiontype + severity + targettype + cyber_objective +
    information_operation + objective_achievement + Concession +
    X3rdpartyinitator + X3rdparty.target + govtstatement + damage.type
Model 2: APT ~ interactiontype + severity + information_operation + Concession +
    X3rdpartyinitator + X3rdparty.target + govtstatement + damage.type
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       254     206.93
2       257     208.83 -3  -1.9075   0.5918
```

04. RESULTS

Test possibility with control variable method

severity <dbl>	prob <dbl>
0	0.02080161
1	0.07841111
2	0.25415696
3	0.57713111
4	0.84534889
5	0.95631774
6	0.98872378
7	0.99716051
8	0.99928952
9	0.99982251

```
predict(cs.reduced,newdata=cs,type="response")

data.frame(interactiontype=mean(cs$interactiontype),
severity=seq(0,10,1),information_operation=mean(cs$infor
mation_operation),Concession=mean(cs$Concession),
X3rdpartyinitator=mean(cs$X3rdpartyinitator),
X3rdparty.target=mean(cs$X3rdparty.target),
govtstatement=mean(cs$govtstatement),
damage.type=mean(cs$damage.type)) -> testdata2

predict(cs.reduced,newdata=testdata2,type="response")
 -> testdata2$prob
```

05. DISCUSSION

CONCLUSION

- When the severity level of network attack increases, the probability that the incident is considered an advanced persistent threat also increases.
- Simple Linear Regression Model 1 was the strongest due to its high R-squared values and low p-value
- Results from our data visualization also indicate that the Americas were the greatest victim to cyber-attacks, while China and Russia were the top two initiators.

FUTURE WORK

- Develop our own original dataset from scratch or through synthesis across existing datasets
- Conduct an independent research and statistical study on why nation-state initiators prefer certain cyber-attack methods over others
- Connect the observable trends and vulnerabilities that organizations and nation-states face identified here to existing studies that can be leveraged to develop original research aimed at proposing new mitigation strategies

THANKS!

DO YOU HAVE ANY QUESTION?

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik.

Please keep this slide for attribution.

APPENDIX

Model 2: Bar Graph of Interaction Types Against Method Type**

```{r}
ggplot(data = cyberincidentsdata) +
  geom_bar(mapping = aes(x = interactiontypename, y = methodname, fill = interactiontypename),
stat = "identity")
```

Model 3: Bar Graph of Interaction Types**

```{r}
ggplot(data = cyberincidentsdata) +
  geom_bar(mapping = aes(x = interactiontypename, fill = interactiontypename))
```

Model 4: Bar Graph of Method Types**

```{r}
ggplot(data = cyberincidentsdata) +
  geom_bar(mapping = aes(x = methodname, fill = methodname))
```

APPENDIX

Model 8: Scatter Plot of Interaction Start Date vs Cyber Incident Num**

```{r}
ggplot(data=cyberincidentsdata) +
  geom_point(mapping = aes(x = interactionstartdate, y= Cyberincidentnum))
```

Model 9: Scatter Plot of Interaction End Date vs Cyber Incident Num**

```{r}
ggplot(data=cyberincidentsdata) +
  geom_point(mapping = aes(x = interactionenddate, y= Cyberincidentnum))
```

Simple Linear Regression

```{r}
lm(Cyberincidentnum~StateA, data = cyberincidentsdata) -> lm1
lm1
```

```
Call:
lm(formula = Cyberincidentnum ~ StateA, data = cyberincidentsdata)

Coefficients:
  (Intercept)     StateAChina    StateAFrance   StateAGermany     StateAIndia      StateAIran
        104.5            94.5             5.5             8.5           152.0            62.5
StateALebanon   StateAN Korea    StateAPoland    StateARussia   StateAS Korea     StateASyria
         77.5           122.5            11.5            31.0           138.0            76.5
     StateAUK        StateAUS
          2.5           -52.5
```

```{r}
summary(lm1)
```

```
Call:
lm(formula = Cyberincidentnum ~ StateA, data = cyberincidentsdata)

Residuals:
   Min     1Q Median     3Q    Max
   -51     -9      0      9     51

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     104.50      14.18   7.371 2.42e-12 ***
StateAChina      94.50      14.60   6.473 4.99e-10 ***
StateAFrance      5.50      18.30   0.301 0.764035
StateAGermany     8.50      18.30   0.464 0.642743
StateAIndia     152.00      14.87  10.223  < 2e-16 ***
StateAIran       62.50      14.69   4.254 2.96e-05 ***
StateALebanon    77.50      24.55   3.156 0.001793 **
StateAN Korea   122.50      14.78   8.288 6.90e-15 ***
StateAPoland     11.50      18.30   0.628 0.530349
StateARussia     31.00      14.56   2.128 0.034278 *
StateAS Korea   138.00      15.85   8.707 4.19e-16 ***
StateASyria      76.50      24.55   3.115 0.002049 **
StateAUK          2.50      18.30   0.137 0.891459
StateAUS        -52.50      14.31  -3.668 0.000298 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.05 on 252 degrees of freedom
Multiple R-squared:  0.9354,	Adjusted R-squared:  0.9321
F-statistic: 280.8 on 13 and 252 DF,  p-value: < 2.2e-16
```

APPENDIX

```r
#wordcloud
wordc<-cyberstop %>%
  count(word)%>%
  filter(n>5)


wordcloud2(data=wordc,size = 1.5,color = 'random-light',backgroundColor = 'Black')
```

Multiple Linear Regression

```r
```{r}
lm(Cyberincidentnum~initiator+cyber_objective+damagetype+severity, data = cyberincidentsdata) ->
mlr1
mlr1
```
```

```
Call:
lm(formula = Cyberincidentnum ~ initiator + cyber_objective +
    damagetype + severity, data = cyberincidentsdata)

Coefficients:
    (Intercept)        initiator  cyber_objective       damagetype         severity
        45.8066           0.1868           0.5992         -12.4713           0.2406
```

```r
```{r}
summary(mlr1)
```
```

```
Call:
lm(formula = Cyberincidentnum ~ initiator + cyber_objective +
    damagetype + severity, data = cyberincidentsdata)

Residuals:
    Min      1Q  Median      3Q     Max
-121.08  -72.16   21.29   47.84   92.80

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      45.8066    21.2165   2.159   0.0318 *
initiator         0.1868     0.0192   9.727   <2e-16 ***
cyber_objective   0.5992     4.8096   0.125   0.9010
damagetype      -12.4713     7.4133  -1.682   0.0937 .
severity          0.2406     4.7666   0.050   0.9598
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 64.19 on 261 degrees of freedom
Multiple R-squared:  0.3142,	Adjusted R-squared:  0.3037
F-statistic:  29.9 on 4 and 261 DF,  p-value: < 2.2e-16
```

APPENDIX

```r
#target type, victim, and number of times
cyber %>%
  group_by(targettypename)%>%
  count(victimname)%>%
  arrange(desc(n))%>%
  ggplot(aes(victimname,n))+
  geom_point(aes(fill=targettypename,color=targettypename))+
  theme(axis.text.x = element_text(angle = 90))+
  scale_color_brewer(palette="Dark2")+
  labs(x='victim name', y= 'count', title='Countries and Targets Attacked')
```

```r
cyber%>%
  select(targettypename,severity)%>%
  group_by(targettypename)%>%
  ggplot(aes(targettypename,severity))+
  geom_boxplot(aes(group=targettypename,fill=targettypename),show.legend = F)+
  theme_classic()+
  labs(x='type of target',title='Severity and Target Type')
```

```r
#damagetypename and severity
#there is only one indirect and delayed
ggplot(cyber,(aes(damagetypename,severity)))+
  geom_boxplot(aes(fill=damagetypename))+
  scale_fill_brewer(palette="Dark2")+
  labs(x='damage type',y='severity',title='Damage Type and Severity')+
  coord_flip()
```

```r
```{r}
cyberregion<-cyber %>%
  group_by(victimregion)%>%
  count()
cyberregion
cd<- ggplot(cyberregion, aes(x="", y = n, fill=victimregion))+
  geom_bar(width = .5, stat = "identity")
cd

pie <- cd + coord_polar("y", start=0)
pie+scale_fill_brewer(palette="Dark2")+
  labs(title='Which Region Had the Most Victims?')
```

```r
#who initiated the most attacks?
cyber %>%
  group_by(methodname)%>%
  count(initiatorname)%>%
  arrange(desc(n))%>%
  filter(n>1)%>%
  ggplot(aes(initiatorname,n))+
  geom_col(aes(fill=methodname))+
  scale_fill_brewer(palette="Dark2")+
  coord_flip()+
  labs(x='count',y='country',title='What attack methods do initiators prefer?')
```
```