Question2

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Please load the nycflights13 package and other necessary packages and follow the following steps.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Step 1. Describe the main question you are testing.

How does the type of breach correlate with the number of individuals affected?

Step 2. Identify the variables that are relevant to the question.

- What are the types of those variables? How do you determine that?

```
breach <- read.csv(file = 'Cyber Security Breaches.csv')
str(breach)
```

```
## 'data.frame':    1055 obs. of  14 variables:
##  $ X                               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Number                          : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Name_of_Covered_Entity          : chr  "Brooke Army Medical Center" "Mid America Kidney Stone Ass
##  $ State                           : chr  "TX" "MO" "AK" "DC" ...
##  $ Business_Associate_Involved     : chr  "" "" "" "" ...
##  $ Individuals_Affected            : int  1000 1000 501 3800 5257 857 6145 952 5166 5900 ...
##  $ Date_of_Breach                  : chr  "10/16/2009" "9/22/2009" "10/12/2009" "10/9/2009" ...
##  $ Type_of_Breach                  : chr  "Theft" "Theft" "Theft" "Loss" ...
##  $ Location_of_Breached_Information: chr  "Paper" "Network Server" "Other Portable Electronic Device
##  $ Date_Posted_or_Updated          : chr  "2014-06-30" "2014-05-30" "2014-01-23" "2014-01-23" ...
##  $ Summary                         : chr  "A binder containing the protected health information (PHI
##  $ breach_start                    : chr  "2009-10-16" "2009-09-22" "2009-10-12" "2009-10-09" ...
##  $ breach_end                      : chr  NA NA NA NA ...
##  $ year                            : int  2009 2009 2009 2009 2009 2009 2009 2009 2009 2009 ...
```

Type_of_Breach: discrete (chr) will change to factor Individuals_Affected: continuous(int)

- Describe why those variables may be relevant to this question and why other variables are not relevant
  The origin is relevant to the question because we are testing to see which type of breach affects more people. The individuals affected is relevant to the question because it is the measure we will use for comparisons between the type of breaches.

Step 3. Search for evidence by visualising, transforming, and modeling your data

(Check RDS 3, 5, 7.3, 7.4, 7.5, 7.6 for ideas and inspiration)

3.1 What type of variation occurs within each variable?

Pick one variable and test the following:

3.1.1 Variable 1 (dep_delay, arr_delay)

```
breach1 <- breach %>%
    group_by(Type_of_Breach) %>% summarize(mean_indiv_affected = mean(Individuals_Affected,na.rm=TRUE),m
breach1
```

```
## # A tibble: 29 x 3
##    Type_of_Breach                       mean_indiv_affect~ med_indiv_affect~
##  * <chr>                                           <dbl>              <dbl>
##  1 Hacking/IT Incident                            25052.              2723
##  2 Hacking/IT Incident, Other                      3200               3200
##  3 Improper Disposal                              17674.              1954.
```

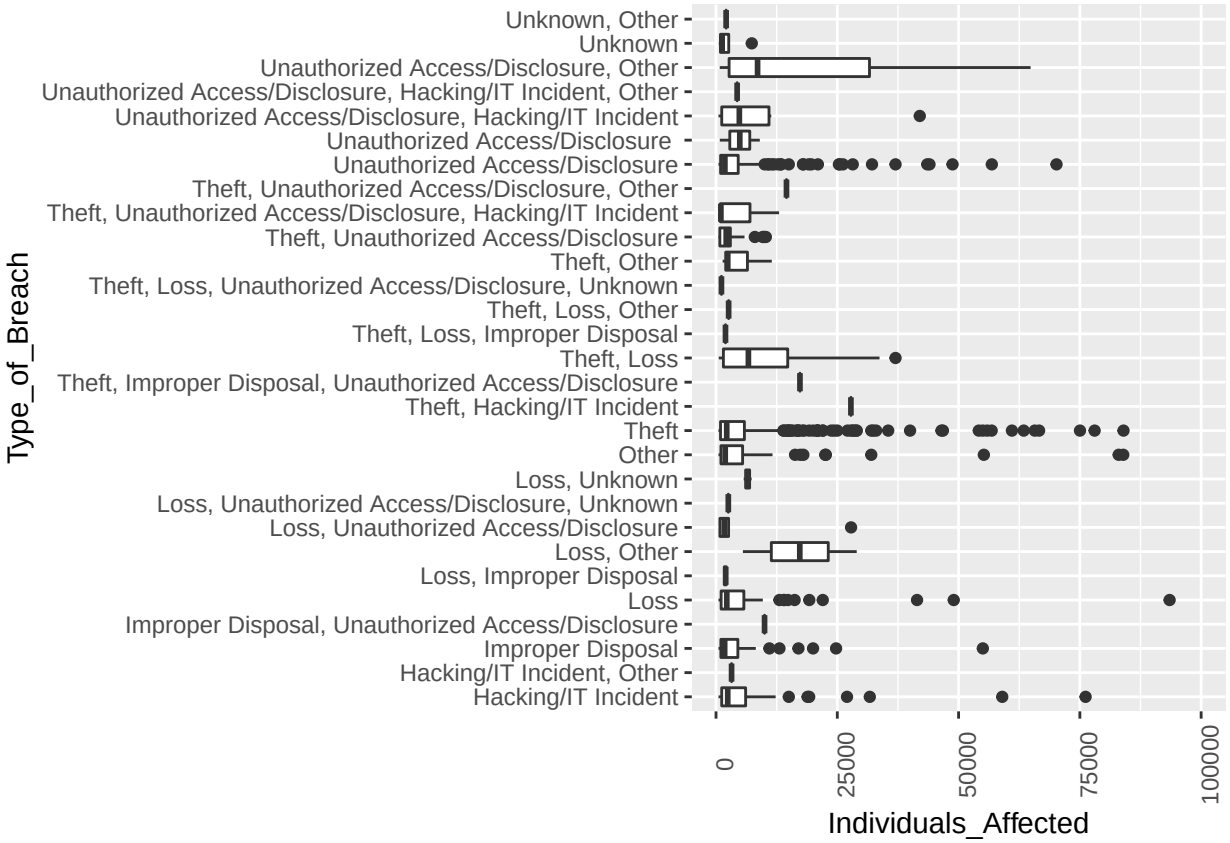
```
##  4 Improper Disposal, Unauthorized Access/~            10000            10000
##  5 Loss                                                85345.            2279
##  6 Loss, Improper Disposal                              1897.            2000
##  7 Loss, Other                                         17267            17267
##  8 Loss, Unauthorized Access/Disclosure                 6728.            1699
##  9 Loss, Unauthorized Access/Disclosure, U~             2533             2533
## 10 Loss, Unknown                                        6518.            6518.
## # ... with 19 more rows
```

3.1.1.1 Visualising distributions (Barcharts, Histograms, etc.) (RDS 7, RDS 3) -Qinyuan

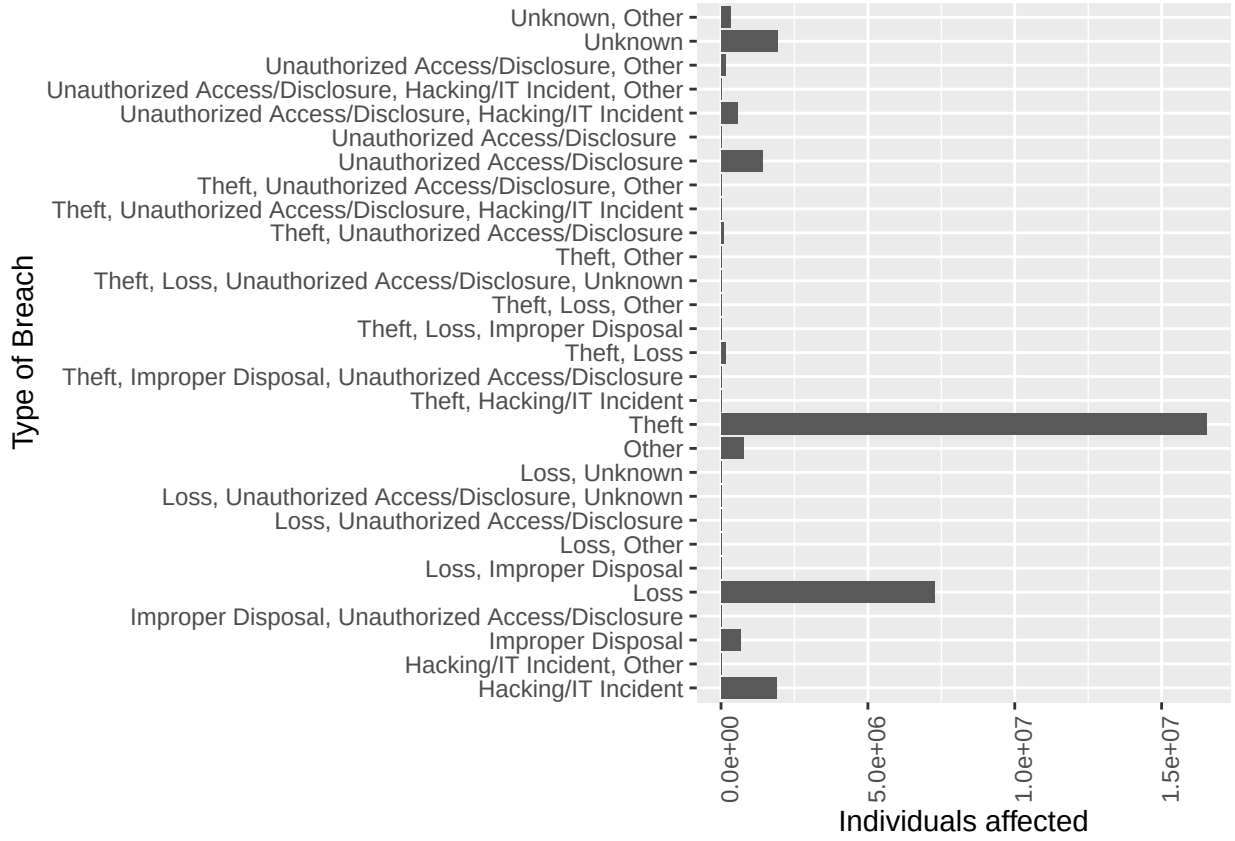
- What chart is appropriate for this variable? Why? A box plot is the most appropriate for these variables so that you
- Which values are the most common? Why?
- Which values are rare? Why? Does that match your expectations?
- Can you see any unusual patterns? What might explain them?

```
ggplot(data = breach, mapping = aes(x = Type_of_Breach, y = Individuals_Affected)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90)) +
  coord_flip() +
  ylim(0,100000)
```

```
## Warning: Removed 35 rows containing non-finite values (stat_boxplot).
```



```
ggplot(breach, aes(x=Type_of_Breach, y=Individuals_Affected)) +
  geom_bar(stat="identity") +
  labs(x= 'Type of Breach', y= 'Individuals affected') +
  theme(axis.text.x = element_text(angle = 90)) +
  coord_flip()
```



3.1.1.2 Unusual values (RDS 7, 5.2)

- Describe and demonstrate how you determine if there are unusual values in the data. E.g. too large, too small, negative, etc.
- Describe and demonstrate how you determine if they are outliers.
- Show how do your distributions look like with and without the unusual values.
- Discuss whether or not you need to remove unusual values and why.

Boxplots are very useful to determine these kinds of values; you can see the minimum and maximum of the delaysfor each airport and which values do not lie within the interquartile by inspecting the boxes. When a data point is unreasonably large or small compared to box's range, you can conclude it is an outlier. Again, you can determine outliers by seeing which data points do not lie within the max, min, and interquartile range of the box plot. It depends on the data, however since standard deviation and mean are very sensitive to unusual values, it is typically good practice to remove unusual values if they will negatively affect your data and goal during data analysis.

3.1.1.3 Missing values (RDS 5.2.3)

- Does this variable include missing values? Demonstrate how you determine that.
- Demonstrate and discuss how you handle the missing values. E.g., removing, replacing with a constant value, or a value based on the distribution, etc.
- Show how your data looks in each case after handling missing values. Describe and discuss the distribution.

3.1.1.4 Does converting the type of this variable help exploring the distribution of its values or identifying outliers or missing values? (RDS 7)

- What type can the variable be converted to?
- How will the distribution look? Please demonstrate with appropriate plots.

3.1.1.5 What new variables do you need to create from this? (RDS 5.5, 5.6, 5.7)

- List the variables
- Describe and discuss why they are needed and how you plan to use them.

3.2. What type of covariation occurs between the two variables? (RDS 7)

3.2.1 Between a categorical and continuous variable or between two categorical variables or between two continuous variables -Julia

- Describe what type of visualization you can use and why. It is a categorical(airports) vs continuous(delay time)
- Describe the patterns and relationships you observe. Could the identified patterns be due to coincidence (i.e. random chance)?
- Describe the relationship implied by the pattern? (e.g., positive or negative correlation)
- Calculate the strength of the relationship implied by the pattern (e.g., correlation)
- Discuss how the observed patterns support/reject your hypotheses or answer your questions.

A histogram should be used to visualize the data - can handle continuous data types the data is skewed to the right - there are a large number of flights with small delays, shifting down to a lower delay time

Step 4. Summarize your findings

- Summarize your findings about the questions you asked at the beginning.
- Discuss if you have enough evidence to make a conclusion about your analysis.

```
library(modelr)
mod <- lm(Individuals_Affected ~ Type_of_Breach, data = breach)
summary(mod)
```

```
##
## Call:
## lm(formula = Individuals_Affected ~ Type_of_Breach, data = breach)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -191280  -30807  -23691   -7274 4814655
##
## Coefficients:
##                                                                    Estimate
## (Intercept)                                                           25052
## Type_of_BreachHacking/IT Incident, Other                             -21852
## Type_of_BreachImproper Disposal                                       -7378
## Type_of_BreachImproper Disposal, Unauthorized Access/Disclosure      -15052
## Type_of_BreachLoss                                                    60293
## Type_of_BreachLoss, Improper Disposal                                -23155
## Type_of_BreachLoss, Other                                             -7785
## Type_of_BreachLoss, Unauthorized Access/Disclosure                   -18324
## Type_of_BreachLoss, Unauthorized Access/Disclosure, Unknown          -22519
## Type_of_BreachLoss, Unknown                                          -18534
## Type_of_BreachOther                                                  -16563
## Type_of_BreachTheft                                                    6955
## Type_of_BreachTheft, Hacking/IT Incident                               2748
## Type_of_BreachTheft, Improper Disposal, Unauthorized Access/Disclosure -7752
## Type_of_BreachTheft, Loss                                            -14420
## Type_of_BreachTheft, Loss, Improper Disposal                         -23102
## Type_of_BreachTheft, Loss, Other                                     -22452
## Type_of_BreachTheft, Loss, Unauthorized Access/Disclosure, Unknown   -23940
## Type_of_BreachTheft, Other                                           -20300
## Type_of_BreachTheft, Unauthorized Access/Disclosure                  -21983
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Hacking/IT Incident -20118
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Other           -10533
## Type_of_BreachUnauthorized Access/Disclosure                         -15428
## Type_of_BreachUnauthorized Access/Disclosure                         -20156
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident     36210
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident, Other -20698
## Type_of_BreachUnauthorized Access/Disclosure, Other                   -4704
## Type_of_BreachUnknown                                                166780
## Type_of_BreachUnknown, Other                                         133489
##                                                                   Std. Error
## (Intercept)                                                            26491
## Type_of_BreachHacking/IT Incident, Other                              230947
## Type_of_BreachImproper Disposal                                        45683
## Type_of_BreachImproper Disposal, Unauthorized Access/Disclosure       230947
## Type_of_BreachLoss                                                     36346
## Type_of_BreachLoss, Improper Disposal                                 135080
## Type_of_BreachLoss, Other                                             164375
## Type_of_BreachLoss, Unauthorized Access/Disclosure                    105966
## Type_of_BreachLoss, Unauthorized Access/Disclosure, Unknown           230947
## Type_of_BreachLoss, Unknown                                           164375
## Type_of_BreachOther                                                    35780
## Type_of_BreachTheft                                                    28351
## Type_of_BreachTheft, Hacking/IT Incident                              230947
## Type_of_BreachTheft, Improper Disposal, Unauthorized Access/Disclosure 230947
```

```
## Type_of_BreachTheft, Loss                                                  64891
## Type_of_BreachTheft, Loss, Improper Disposal                              230947
## Type_of_BreachTheft, Loss, Other                                          230947
## Type_of_BreachTheft, Loss, Unauthorized Access/Disclosure, Unknown        230947
## Type_of_BreachTheft, Other                                                105966
## Type_of_BreachTheft, Unauthorized Access/Disclosure                        52213
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Hacking/IT Incident  135080
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Other                230947
## Type_of_BreachUnauthorized Access/Disclosure                              32518
## Type_of_BreachUnauthorized Access/Disclosure                              164375
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident         80933
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident, Other  230947
## Type_of_BreachUnauthorized Access/Disclosure, Other                       85330
## Type_of_BreachUnknown                                                     77235
## Type_of_BreachUnknown, Other                                             164375
##                                                                         t value
## (Intercept)                                                               0.946
## Type_of_BreachHacking/IT Incident, Other                                 -0.095
## Type_of_BreachImproper Disposal                                          -0.162
## Type_of_BreachImproper Disposal, Unauthorized Access/Disclosure          -0.065
## Type_of_BreachLoss                                                        1.659
## Type_of_BreachLoss, Improper Disposal                                    -0.171
## Type_of_BreachLoss, Other                                                -0.047
## Type_of_BreachLoss, Unauthorized Access/Disclosure                       -0.173
## Type_of_BreachLoss, Unauthorized Access/Disclosure, Unknown              -0.098
## Type_of_BreachLoss, Unknown                                              -0.113
## Type_of_BreachOther                                                      -0.463
## Type_of_BreachTheft                                                       0.245
## Type_of_BreachTheft, Hacking/IT Incident                                  0.012
## Type_of_BreachTheft, Improper Disposal, Unauthorized Access/Disclosure   -0.034
## Type_of_BreachTheft, Loss                                                -0.222
## Type_of_BreachTheft, Loss, Improper Disposal                             -0.100
## Type_of_BreachTheft, Loss, Other                                         -0.097
## Type_of_BreachTheft, Loss, Unauthorized Access/Disclosure, Unknown       -0.104
## Type_of_BreachTheft, Other                                               -0.192
## Type_of_BreachTheft, Unauthorized Access/Disclosure                      -0.421
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Hacking/IT Incident -0.149
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Other               -0.046
## Type_of_BreachUnauthorized Access/Disclosure                             -0.474
## Type_of_BreachUnauthorized Access/Disclosure                             -0.123
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident         0.447
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident, Other -0.090
## Type_of_BreachUnauthorized Access/Disclosure, Other                      -0.055
## Type_of_BreachUnknown                                                     2.159
## Type_of_BreachUnknown, Other                                              0.812
##                                                                        Pr(>|t|)
## (Intercept)                                                              0.3445
## Type_of_BreachHacking/IT Incident, Other                                 0.9246
## Type_of_BreachImproper Disposal                                          0.8717
## Type_of_BreachImproper Disposal, Unauthorized Access/Disclosure          0.9480
## Type_of_BreachLoss                                                       0.0974
## Type_of_BreachLoss, Improper Disposal                                    0.8639
## Type_of_BreachLoss, Other                                                0.9622
## Type_of_BreachLoss, Unauthorized Access/Disclosure                       0.8627
```

```
## Type_of_BreachLoss, Unauthorized Access/Disclosure, Unknown                0.9223
## Type_of_BreachLoss, Unknown                                                0.9102
## Type_of_BreachOther                                                        0.6435
## Type_of_BreachTheft                                                        0.8063
## Type_of_BreachTheft, Hacking/IT Incident                                   0.9905
## Type_of_BreachTheft, Improper Disposal, Unauthorized Access/Disclosure     0.9732
## Type_of_BreachTheft, Loss                                                  0.8242
## Type_of_BreachTheft, Loss, Improper Disposal                               0.9203
## Type_of_BreachTheft, Loss, Other                                           0.9226
## Type_of_BreachTheft, Loss, Unauthorized Access/Disclosure, Unknown         0.9175
## Type_of_BreachTheft, Other                                                 0.8481
## Type_of_BreachTheft, Unauthorized Access/Disclosure                        0.6738
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Hacking/IT Incident   0.8816
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Other                 0.9636
## Type_of_BreachUnauthorized Access/Disclosure                               0.6353
## Type_of_BreachUnauthorized Access/Disclosure                               0.9024
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident          0.6547
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident, Other   0.9286
## Type_of_BreachUnauthorized Access/Disclosure, Other                        0.9560
## Type_of_BreachUnknown                                                      0.0311
## Type_of_BreachUnknown, Other                                               0.4169
##
## (Intercept)
## Type_of_BreachHacking/IT Incident, Other
## Type_of_BreachImproper Disposal
## Type_of_BreachImproper Disposal, Unauthorized Access/Disclosure
## Type_of_BreachLoss                                                          .
## Type_of_BreachLoss, Improper Disposal
## Type_of_BreachLoss, Other
## Type_of_BreachLoss, Unauthorized Access/Disclosure
## Type_of_BreachLoss, Unauthorized Access/Disclosure, Unknown
## Type_of_BreachLoss, Unknown
## Type_of_BreachOther
## Type_of_BreachTheft
## Type_of_BreachTheft, Hacking/IT Incident
## Type_of_BreachTheft, Improper Disposal, Unauthorized Access/Disclosure
## Type_of_BreachTheft, Loss
## Type_of_BreachTheft, Loss, Improper Disposal
## Type_of_BreachTheft, Loss, Other
## Type_of_BreachTheft, Loss, Unauthorized Access/Disclosure, Unknown
## Type_of_BreachTheft, Other
## Type_of_BreachTheft, Unauthorized Access/Disclosure
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Hacking/IT Incident
## Type_of_BreachTheft, Unauthorized Access/Disclosure, Other
## Type_of_BreachUnauthorized Access/Disclosure
## Type_of_BreachUnauthorized Access/Disclosure
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident
## Type_of_BreachUnauthorized Access/Disclosure, Hacking/IT Incident, Other
## Type_of_BreachUnauthorized Access/Disclosure, Other
## Type_of_BreachUnknown                                                      *
## Type_of_BreachUnknown, Other
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 229400 on 1026 degrees of freedom
## Multiple R-squared:  0.01317,    Adjusted R-squared:  -0.01376
## F-statistic: 0.489 on 28 and 1026 DF,  p-value: 0.9887
```

```
grid <- breach %>%
  data_grid(Type_of_Breach) %>%
  add_predictions(mod, "Individuals_Affected")

ggplot(breach, aes(Type_of_Breach, Individuals_Affected)) +
  geom_boxplot() +
  geom_point(data = grid, colour = "red", size = 4) +
  theme(axis.text.x = element_text(angle = 90)) +
  coord_flip()+
  ylim(0,100000)
```

```
## Warning: Removed 35 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```