Data Breach Analysis

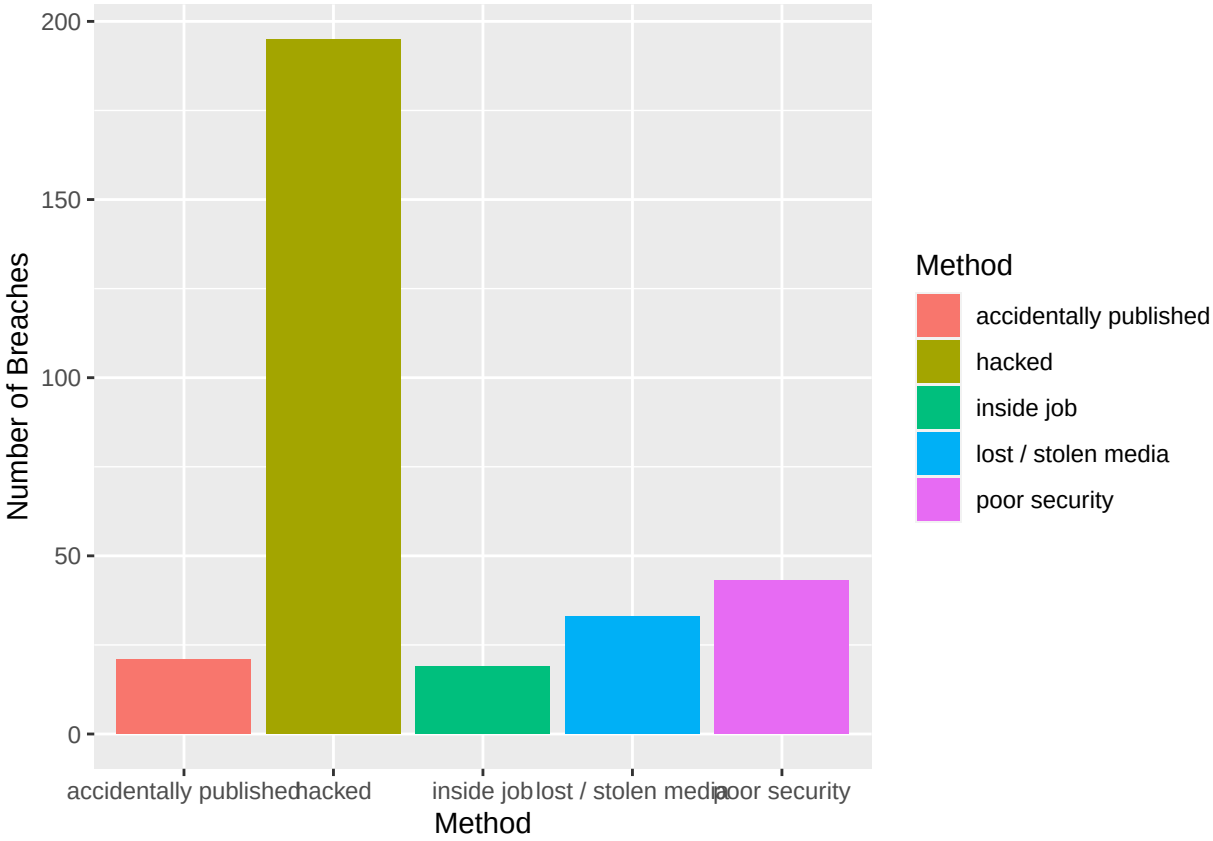
Cicely Karas

2023-07-08

1.

a. Show the top 5 Methods of breaches by number of breaches.

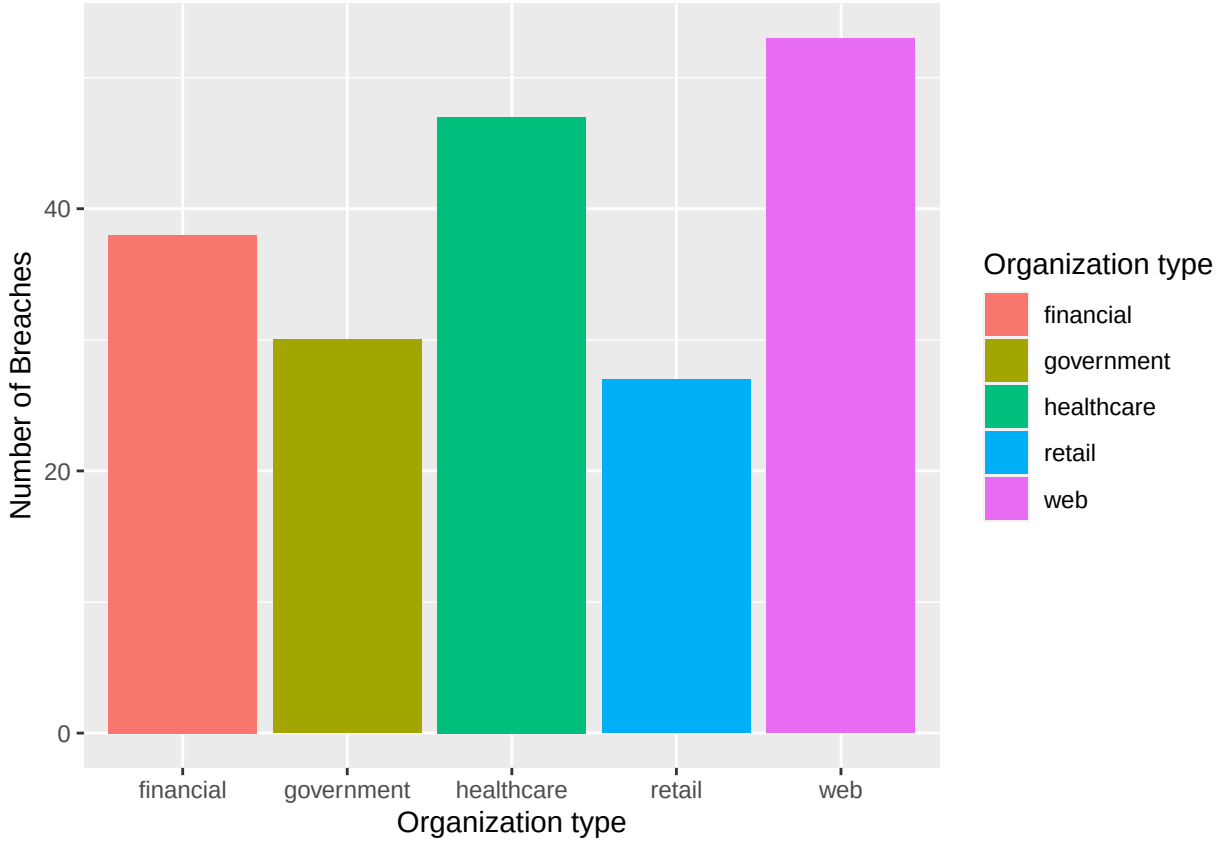
```
DataBreaches%>%
  group_by(Method)%>%
  summarize(count= n())%>%
  top_n(5)%>%
  ggplot(aes(x= Method, y= count, fill = Method))+
  geom_bar(stat="identity",position = "dodge")+
  ylab("Number of Breaches")
```



b. Show the top 5 Organization types that is breached by number of breaches.

```
DataBreaches%>%
  group_by(`Organization type`)%>%
  summarize(count= n())%>%
  top_n(5)%>%
```

```
  ggplot(aes(x= `Organization type`, y= count, fill = `Organization type`))+
  geom_bar(stat="identity",position = "dodge")+
  ylab("Number of Breaches")
```



2. You are interested in number of Breaches that happened Pre-COVID versus Post-COVID. Show the years that had the top 5 number of breaches. Include a column that says if it was Pre or During/Post-COVID (2020 and later).

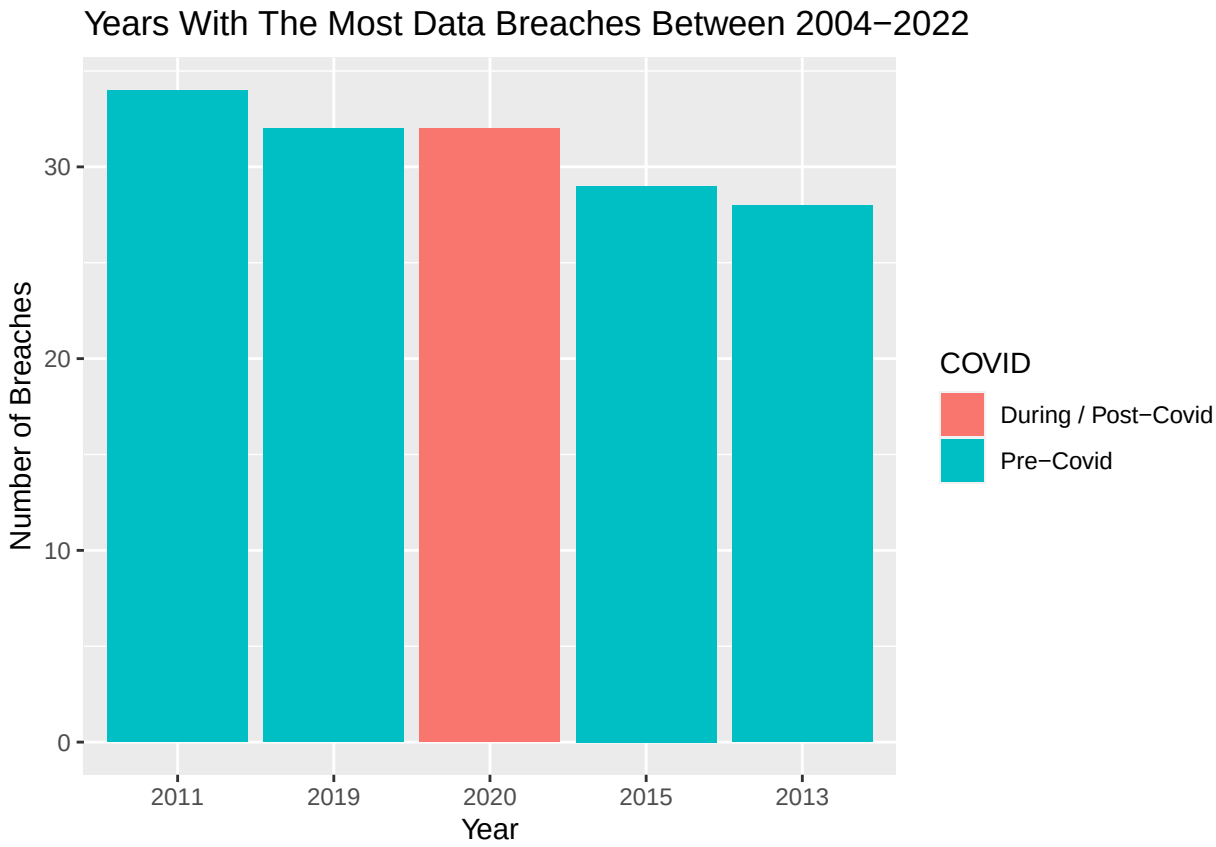
```
data = DataBreaches%>%
  group_by(Year)%>%
  summarize(count= n())%>%
  top_n(5)%>%
  mutate(COVID = ifelse(Year == 2020, "During / Post-Covid", "Pre-Covid"))
data
```

```
## # A tibble: 5 x 3
##    Year count COVID
##   <dbl> <int> <chr>
## 1  2011    34 Pre-Covid
## 2  2013    28 Pre-Covid
## 3  2015    29 Pre-Covid
## 4  2019    32 Pre-Covid
## 5  2020    32 During / Post-Covid
```

3. Using the table in (2), create the following barplot:

```
data%>%
  mutate(Year = as.factor(Year))%>%
  ggplot(aes(x= reorder(Year, desc(count)), y= count, fill = COVID))+
```

```
  geom_bar(stat="identity",position = "dodge")+
  ylab("Number of Breaches")+
  xlab("Year")+
  ggtitle("Years With The Most Data Breaches Between 2004-2022")
```



Main takeaway: The top years for data breaches between 2004-2022 are all pretty close in range, with the max. being 34 and the min. being 28, the count of data breaches is pretty close between these years. There's not one year with a significant amount of more data breaches than the others.

4. Use Organization and Method to predict Records. For Organization, focus on web, healthcare and financial and for Method focus on hacked and lost/stolen media. Include only records that are less than 25 million (More than that I consider an outlier)

(using linear regression)

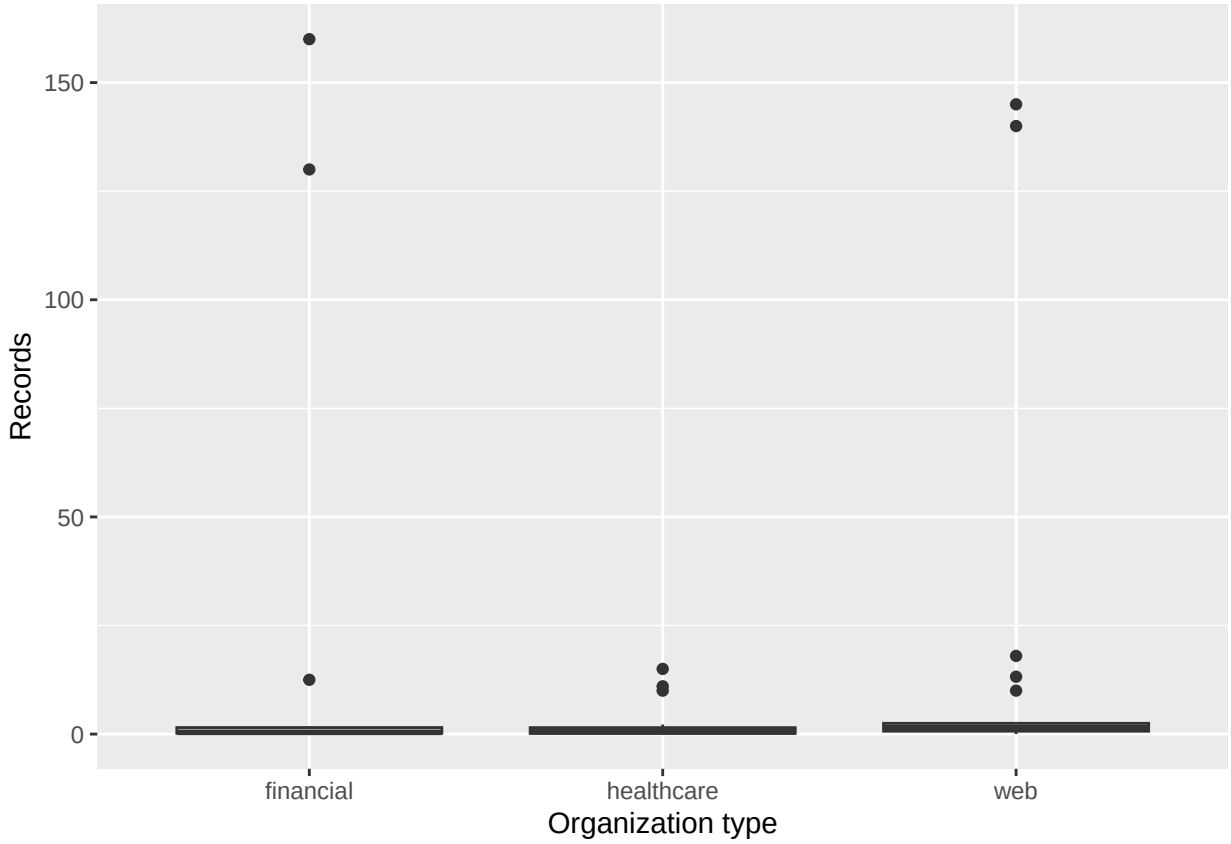
```
# Replace "#VALUE!" with NA
DataBreaches[DataBreaches == "#VALUE!"] <- NA

# Remove any rows with missing values
clean_data <- na.omit(DataBreaches)
```

```
data2 = clean_data%>%
  filter(`Organization type` %in% c("web", "healthcare", "financial"),
         Method %in% c("hacked", "lost / stolen media"),
         Records <= 25000000, Na.rm = TRUE)%>%
  na.omit(Records)
```

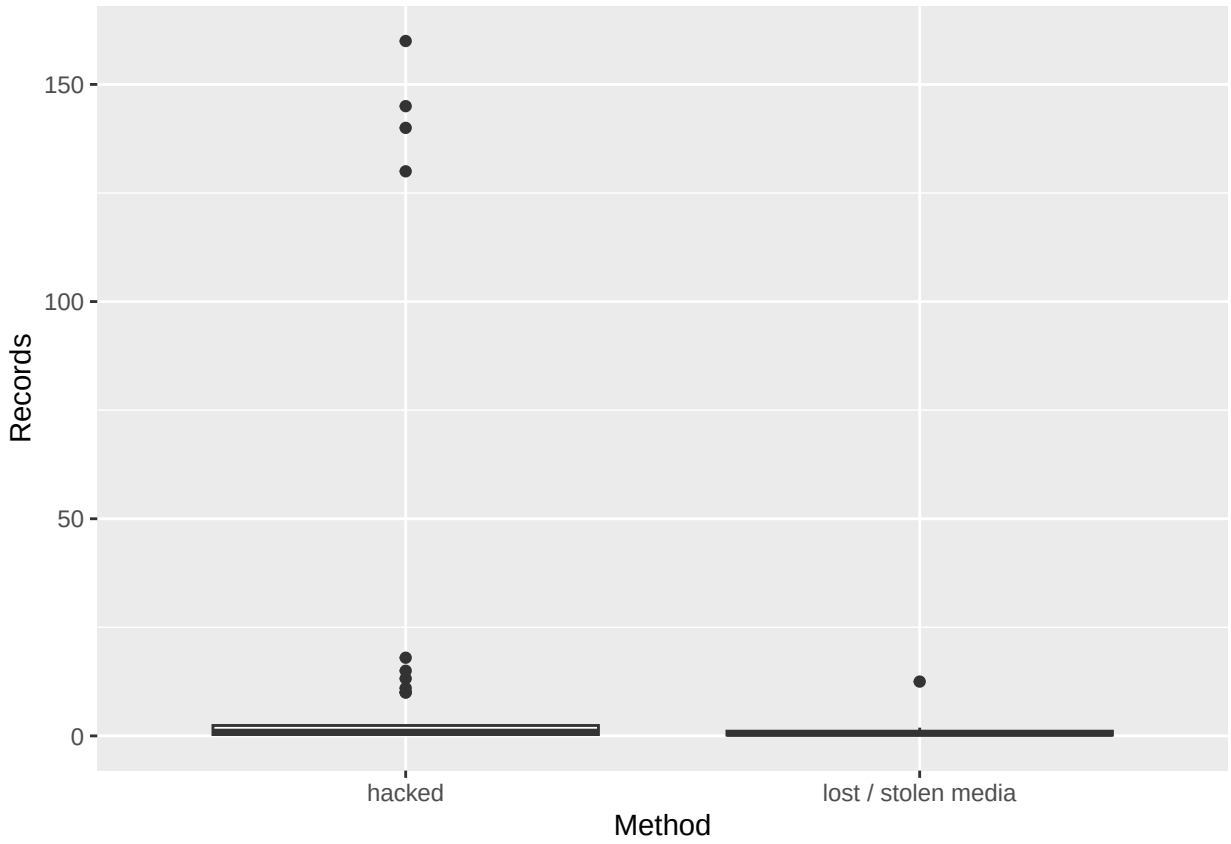
a. Is there a relationship between the predictors and the response?

```
data2%>%
  mutate(Records = as.numeric(Records))%>%
  ggplot(aes(y= Records, x= `Organization type`))+
  geom_boxplot()
```



Its hard to determine if there's a relationship between the predictor organization type and the response records. All 3 of th box plots on the graph above have a very slim interquartile range, this means that the minimum and maximum values are very close to each other, and so are the first and third quartile. This indicates that the data has very little variation, and that most of the values in the data set are similar to each other. And therefore, hard to determine a relationship.

```
data2%>%
  mutate(Records = as.numeric(Records))%>%
  ggplot(aes(x= Method, y= Records))+
  geom_boxplot()
```



Its hard to determine if there's a relationship between the predictor method and the response records. This box plot has a similar range to the one above, meaning the max. and min. are very close together, therefore, hard to determine if there is a relationship.

b. How strong are the relationships?

```
MLR= lm(Records ~ Method + `Organization type`, data = data2)
summary(MLR)
```

```
##
## Call:
## lm(formula = Records ~ Method + `Organization type`, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.779 -15.123  -6.609   3.789 133.161
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      26.84      10.39   2.583   0.0124 *
## Methodlost / stolen media       -10.39      11.58  -0.897   0.3733
## `Organization type`healthcare   -20.07      12.09  -1.661   0.1022
## `Organization type`web          -11.12      12.82  -0.867   0.3894
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.19 on 57 degrees of freedom
## Multiple R-squared:  0.07434,    Adjusted R-squared:  0.02562
```

```
## F-statistic: 1.526 on 3 and 57 DF,  p-value: 0.2176
```

The R-squared is 0.07434, so altogether, the predictors do not have a very strong relationship with the response.

c. What are the effect of the predictors on the response?

- baseline (hacked) = 26.84 records on average get stolen through hacked
- lost / stolen media: for the method lost / stolen media, on average has -10.39 records less than the baseline
- healthcare: for the `Organization type` healthcare, on average has -20.07 less records than the baseline
- web: for the `Organization type` web, on average has -11.12 less record than the baseline

d. What is the value that represents the predictive accuracy of the model (no need to interpret it.)

The RSE is 35.19.

5. You are reporting the results to your organization. Talk about the relationship between Organization Type and Methods when trying to predict the number of records breached. (i.e. Can you tell if a specific type of organization or Method will result in more or less records being stolen? How and why?)

Note: Use 100 words to write a non-technical summary.

Based on my results, we can determine on average what kind of organizations are more likely to be breached out of financial, healthcare, and web organizations. And what method the records will be breached out of hacked and lost/stolen media methods. Hacked methods on average steal more than lost/stolen media methods. Web organizations are most likely to get the most records stolen, then financial organizations, then healthcare organizations. For example, someone stealing records from a web organization through hacking is likely to steal the most records on average compared to any other combination of method and organization type. So it's important to be cautious over which methods tend to steal more records, and which organization types, so we can take preventative measures to protect our organization.