Final

Qinyuan Jiang

```
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()
```

```
library(ggplot2)
```

```
breaches <- read_csv("Cyber Security Breaches.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]

##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   Number = col_double(),
##   Name_of_Covered_Entity = col_character(),
##   State = col_character(),
##   Business_Associate_Involved = col_character(),
##   Individuals_Affected = col_double(),
##   Date_of_Breach = col_character(),
##   Type_of_Breach = col_character(),
##   Location_of_Breached_Information = col_character(),
##   Date_Posted_or_Updated = col_date(format = ""),
##   Summary = col_character(),
##   breach_start = col_date(format = ""),
##   breach_end = col_date(format = ""),
##   year = col_double()
## )
```

Step 1. Describe what question is being tested.

Question: Has the number of breaches increased overtime? Is there a trend?

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

Only the year column in the dataset is relevant because that can be used to determine the trend across the years and to create a separate column to count how many breaches there have been in that given year.

```
number_of_breaches <- breaches %>%
  group_by(year)%>%
  summarise(freq = n())
```

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?

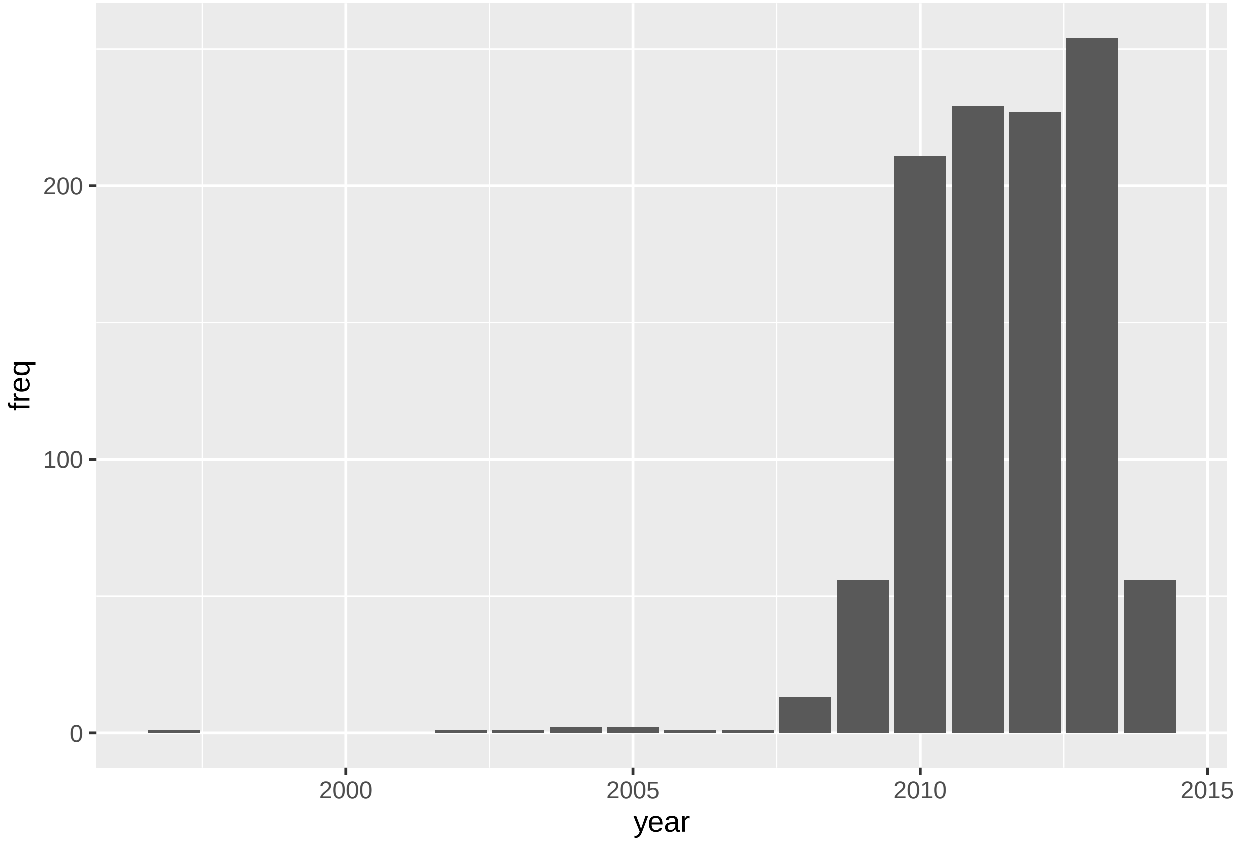
3.1.1 Variable

```


ggplot(number_of_breaches, aes(year, freq)) +
  geom_histogram(stat = "identity")
```

3.1.1.1 Visualising distributions (Barcharts, Histograms)

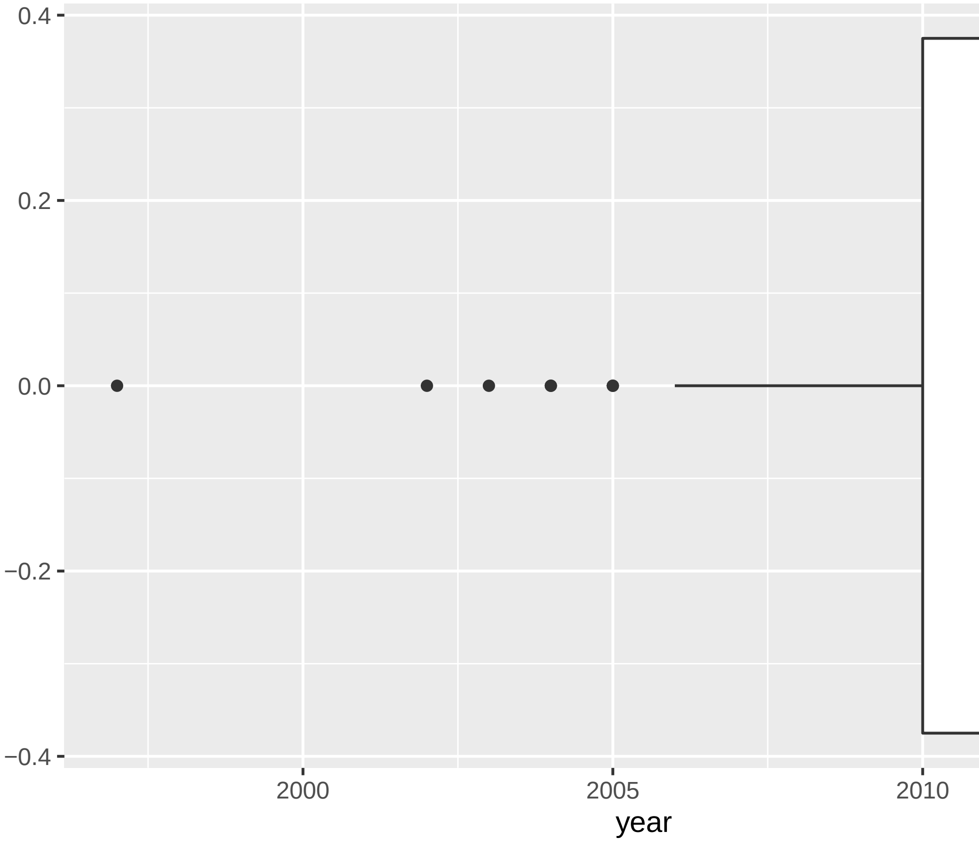
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



- Which values are the most common? Why? The most common values are around the year 2010-2013. This may be due to where the dataset is collected.

- Which values are rare? Why? Does that match your expectations? Values before 2007 was rare. This may be because the internet is not as relevant as it is later in the dataset than before.
- Can you see any unusual patterns? What might explain them? There is a increase up to the year 2014. This may be because the dataset is cut off in the year 2014.
- Are there clusters in the data? If so,
- How are the observations within each cluster similar to or different from each other?
- How can you explain or describe the clusters?

```
ggplot(breaches, aes(x = year)) +
  geom_boxplot()
```

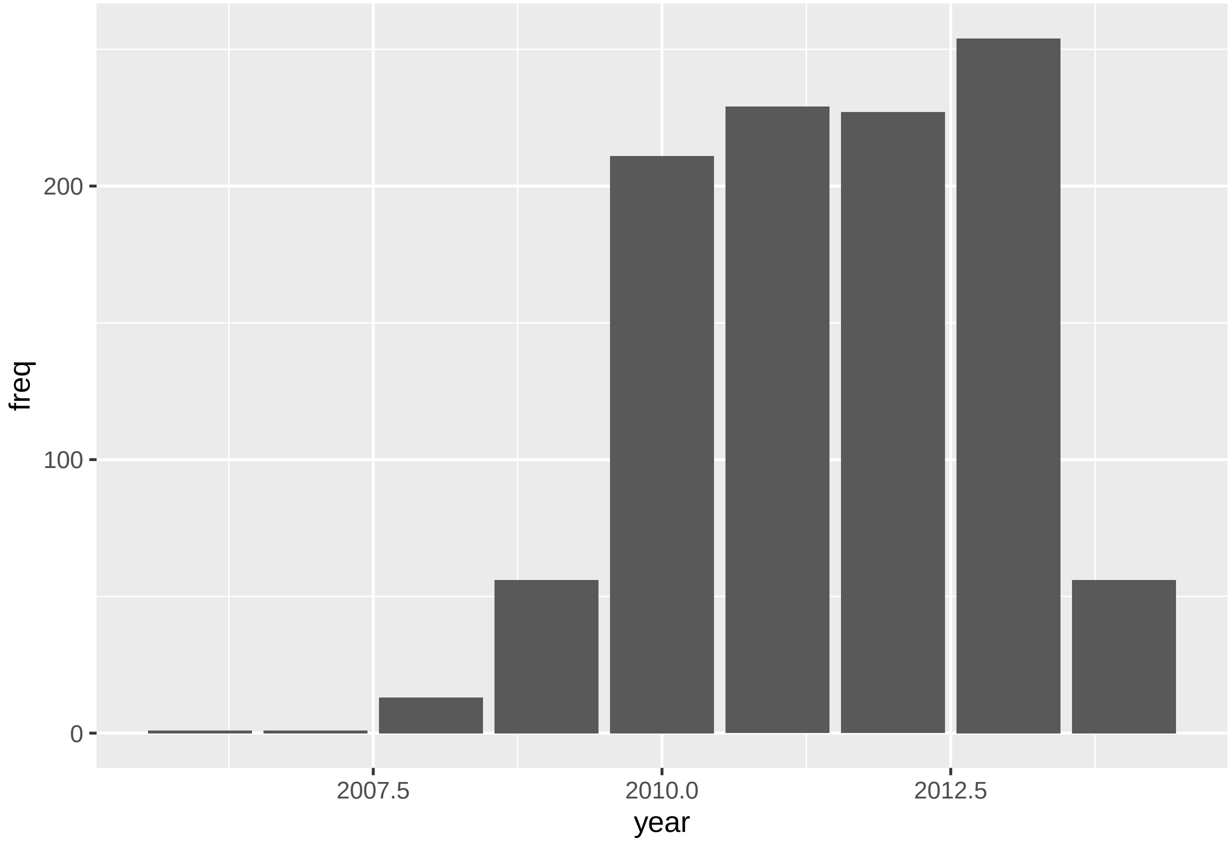


3.1.1.2 Unusual values (2 points)

- 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. I use a box plot to determine the outliers within this data.
- Show how do your distributions look like with and without the unusual values.

```
number_of_breaches %>%
  filter(year > 2005) %>%
  ggplot(aes(year, freq)) +
  geom_histogram(stat = "identity")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



- Discuss whether or not you need to remove unusual values and why. There is no need to remove unusual values because I am interested in the overall trend of the data.

3.1.1.3 Missing values

- Does this variable include missing values? Demonstrate how you determine that.

```
missing <- breaches %>% filter(is.na(year))
missing
```

```
## # A tibble: 0 x 14
## # ... with 14 variables: X1 <dbl>, Number <dbl>, Name_of_Covered_Entity <chr>,
## #   State <chr>, Business_Associate_Involved <chr>, Individuals_Affected <dbl>,
## #   Date_of_Breach <chr>, Type_of_Breach <chr>,
## #   Location_of_Breached_Information <chr>, Date_Posted_or_Updated <date>,
## #   Summary <chr>, breach_start <date>, breach_end <date>, year <dbl>
```

- 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. There are no missing values.
- Show how your data looks in each case after handling missing values.Describe and discuss the distribution. The data would look the same.

Step 4. Summarize your findings (20 points)

- Summarize your findings about the questions you asked at the beginning. (5 points) The number of breaches increases up to 2013 and ends at 2014. This may be because the data ends sometime in 2014.
- Describe and discuss how your observations support or reject your hypotheses or answer your questions. (5 points) There is an increasing trend for the number of breaches as the year increases.
- Describe what new questions your analysis may generate. (5 points) Does the trend continue past 2014?
- Discuss if you have enough evidence to make a conclusion about your analysis. (5 points) There is not enough data as the data stops at 2014 and starts roughly at 2005 with a few data points from prior to 2005.

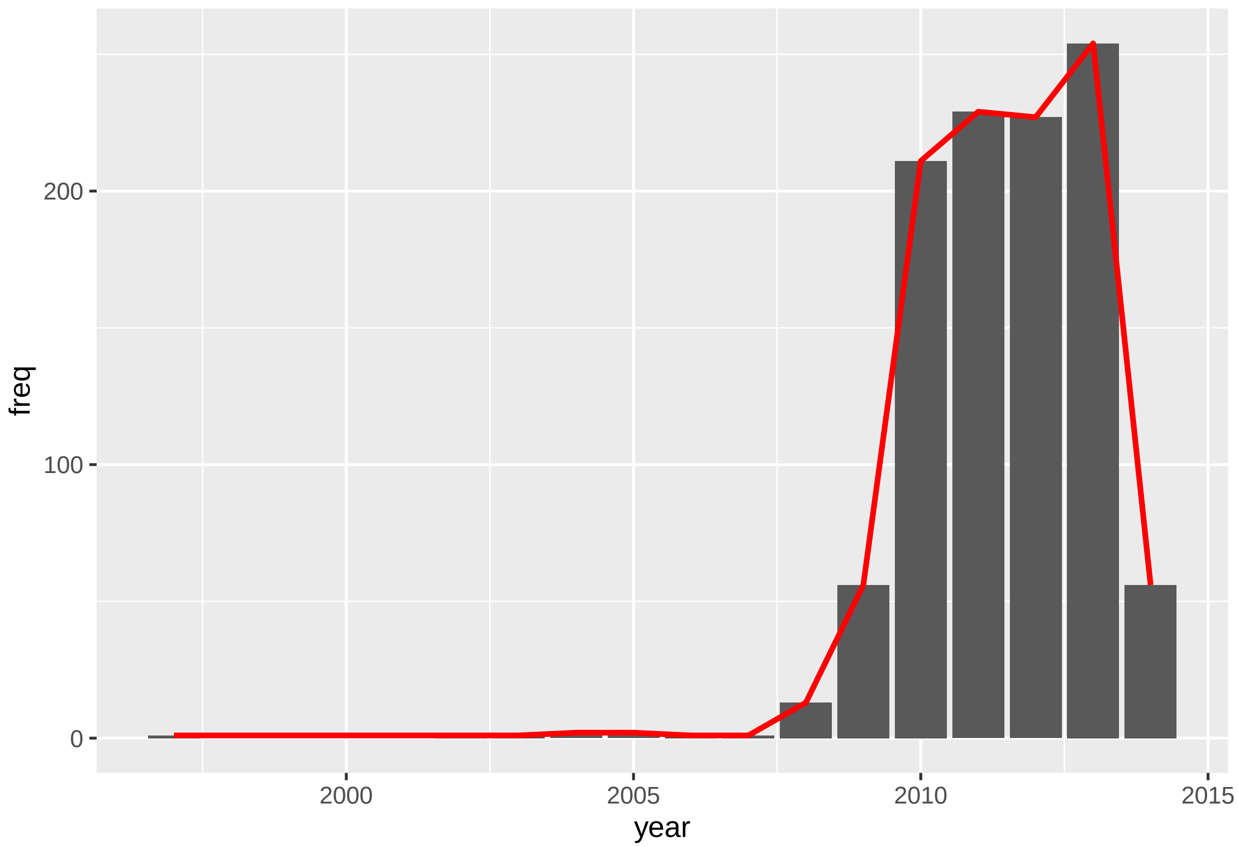
#Modeling

```
library(modelr)
mod <- lm(freq ~ year, data = number_of_breaches)
summary(mod)
```

```
##
## Call:
## lm(formula = freq ~ year, data = number_of_breaches)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -122.38  -53.56  -17.47   80.28   96.17
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -30398.159   9069.234  -3.352  0.00576 **
## year            15.182      4.518   3.360  0.00567 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.52 on 12 degrees of freedom
## Multiple R-squared:  0.4848, Adjusted R-squared:  0.4418
## F-statistic: 11.29 on 1 and 12 DF,  p-value: 0.005673
```

```
range <- number_of_breaches %>%
  add_predictions(mod, "freq1")

ggplot(number_of_breaches, aes(year, freq)) +
  geom_col() +
  geom_line(data = range, colour = "red", size = 1)
```


freq
200
100
0
2000
2005
2010
2015
year