The relationship between location of breach and the number of individuals affected

Tanu Roy

5/3/2021

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5

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

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.0.5
```

```
library(forcats)
```

```
breach_data <- read_csv('C:/Users/tanu roy/OneDrive/Desktop/SYS2202/cyber-security-final/Tanu-Question1/
```

```
## 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()
## )
```

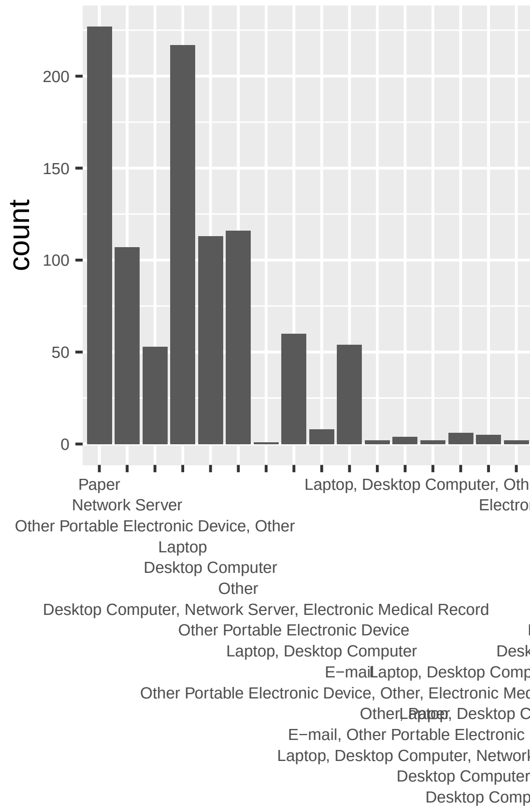
```r
levels <- unique(breach_data$Location_of_Breached_Information, incompareables= FALSE)
location_levels <- c(levels)
location <- factor(breach_data$Location_of_Breached_Information, levels = location_levels)
```

```r
head(breach_data)
```

```
## # A tibble: 6 x 14
##      X1 Number Name_of_Covered_Entity  State Business_Associat~ Individuals_Aff~
##   <dbl>  <dbl> <chr>                   <chr> <chr>                         <dbl>
## 1     1      0 Brooke Army Medical Ce~ TX    <NA>                           1000
## 2     2      1 Mid America Kidney Sto~ MO    <NA>                           1000
## 3     3      2 Alaska Department of H~ AK    <NA>                            501
## 4     4      3 Health Services for Ch~ DC    <NA>                           3800
## 5     5      4 L. Douglas Carlson, M.~ CA    <NA>                           5257
## 6     6      5 David I. Cohen, MD      CA    <NA>                            857
## # ... with 8 more variables: 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>
```

3.1.3 Location of Breached Information Variable

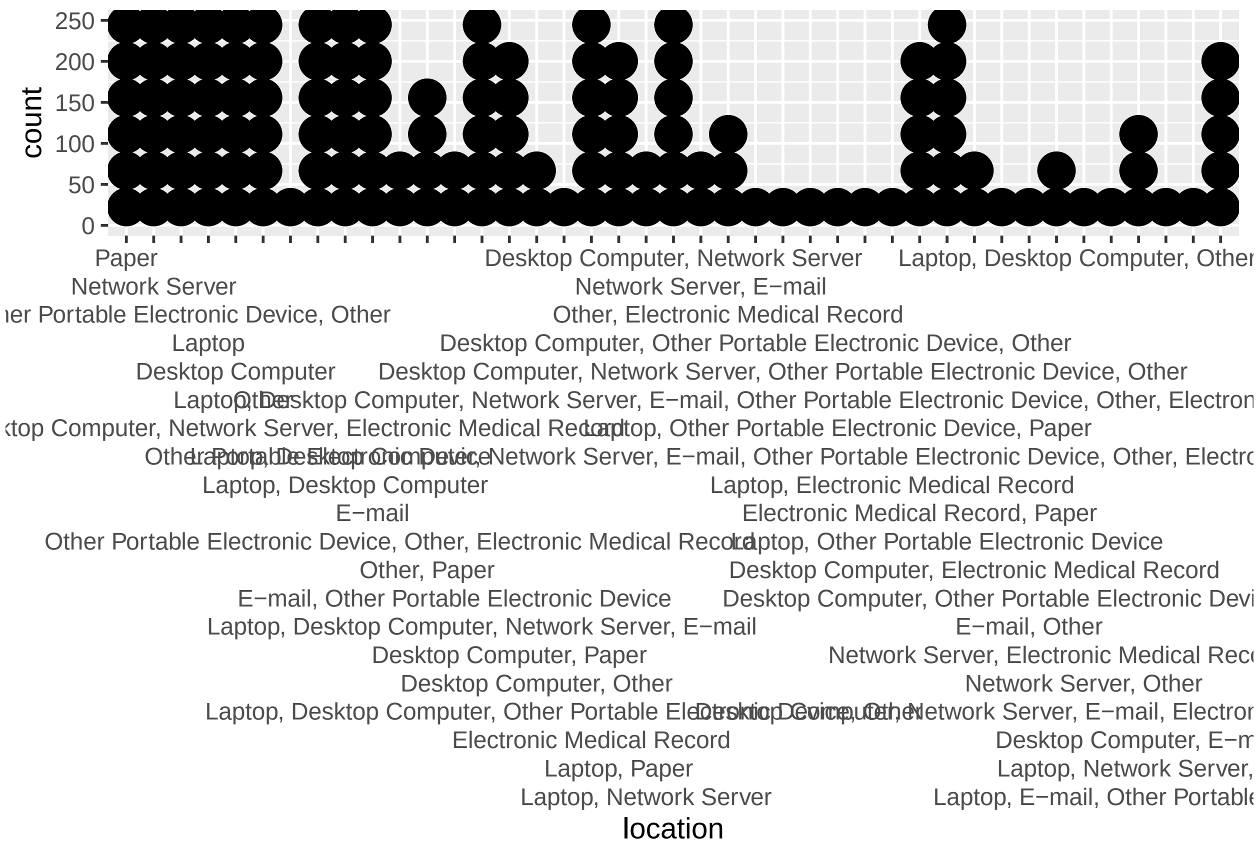
```r
col_types = cols(
  Individuals_Affected = col_integer(),
  Location_of_Breached_Information = col_factor()
  )
location_bargraph <- breach_data %>%
  ggplot(aes(x = location)) + scale_x_discrete(guide = guide_axis(n.dodge=16)) + geom_bar() + theme(axis
location_bargraph
```



3.1.2.1 Visualising distributions (Barcharts, Histograms) (5 points)

```
location_scatter <- breach_data %>%
  ggplot(aes(x=location)) + geom_dotplot() +  scale_x_discrete(guide = guide_axis(n.dodge=20)) + ylim(0
location_scatter
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



- Which values are the most common? Why? The most common location of breached information is those stolen in a paper format.The second most common are laptops. This can be explained by the fact that paper is hard to secure. It can be left lying around or it could be in an easy to breach storage system like an unlocked file cabinet. Compared to locations like electronic medical records which are most likely encrypted and stored in difficult areas, paper is the most definitely the easiest to breach.

- Which values are rare? Why? Does that match your expectations? In this data set, the rarest values are electronic medical devices and breaches where several types of information are located (such as a breach where a desktop and several laptops were stolen). This matches expectations as hospitals and other areas that utilize medical devices are under oath to protect patient confidentiality (HIPAA). The other type, where several mediums of information are stolen, also is not surprising. Breaching one type of location seems like an onerous task, so multiple locations in one breach must be rare.

- Can you see any unusual patterns? What might explain them? There are no unusual patterns in this column.

- Are there clusters in the data? If so, Since this a bar graph of independent factors within our variable, any clusters are due to how the graph has been ordered and not due to a specific reason causing said clusters.

- How are the observations within each cluster similar to or different from each other? See above

- How can you explain or describe the clusters? See above

3.1.2.2 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.

```r
location_bargraph <- breach_data %>%
  ggplot(aes(x = location))  + geom_bar() +  scale_fill_brewer(palette = "Dark2") +coord_flip()

location_bargraph
```

Laptop, Desktop Computer, Other Portable Electronic D

Laptop, E−mail, Other Portable Electronic D

Laptop, Network Server, E

Desktop Computer, E

Desktop Computer, Network Server, E−mail, Electronic Medical Record,

Network Server,

Network Server, Electronic Medical R

E−mail,

Desktop Computer, Other Portable Electronic D

Desktop Computer, Electronic Medical R

Laptop, Other Portable Electronic D

Electronic Medical Record,

Laptop, Electronic Medical R

top, Desktop Computer, Network Server, E−mail, Other Portable Electronic Device, Other, Electronic Medical Record,

Laptop, Other Portable Electronic Device,

Laptop, Desktop Computer, Network Server, E−mail, Other Portable Electronic Device, Other, Electronic Medical R

Desktop Computer, Network Server, Other Portable Electronic Device,

Desktop Computer, Other Portable Electronic Device,

Other, Electronic Medical R

Network Server, E

Desktop Computer, Network S

Laptop, Network S

Laptop,

Electronic Medical R

Laptop, Desktop Computer, Other Portable Electronic Device,

Desktop Computer,

Desktop Computer,

Laptop, Desktop Computer, Network Server, E

E−mail, Other Portable Electronic D

Other,

Other Portable Electronic Device, Other, Electronic Medical R

E

Laptop, Desktop Com

Other Portable Electronic D

Desktop Computer, Network Server, Electronic Medical R

Desktop Com

L

Other Portable Electronic Device,

Network S

There are no unusual values such as negatives or values that are exponentially (or some other factor) larger than the rest.

- Describe and demonstrate how you determine if they are outliers.

```
location_count <- breach_data %>%
  group_by(Location_of_Breached_Information) %>%
  count()
location_count
```

```
## # A tibble: 41 x 2
## # Groups:   Location_of_Breached_Information [41]
##    Location_of_Breached_Information                                         n
##    <chr>                                                                <int>
##  1 Desktop Computer                                                       113
##  2 Desktop Computer, E-mail                                                 3
##  3 Desktop Computer, Electronic Medical Record                              2
##  4 Desktop Computer, Network Server                                         8
##  5 Desktop Computer, Network Server, E-mail, Electronic Medical Record, P~     1
##  6 Desktop Computer, Network Server, Electronic Medical Record               1
##  7 Desktop Computer, Network Server, Other Portable Electronic Device, Ot~     1
##  8 Desktop Computer, Other                                                  2
##  9 Desktop Computer, Other Portable Electronic Device                       1
## 10 Desktop Computer, Other Portable Electronic Device, Other                1
## # ... with 31 more rows
```

```
#boxplot.stats(location_count)$out
```

- Show how do your distributions look like with and without the unusual values. The distribution looks the same since there are no unusual values.

- Discuss whether or not you need to remove unusual values and why. NO values need to be removed since none are unusual.

3.1.2.3 Missing values (2 points) - Does this variable include missing values? Demonstrate how you determine that. This variable has no missing values. This is determined by using a data frame that can provide a number of how many NA values there are as well as what type (TRUE OR FALSE). The data frame says that there are 1055 values in this variable and all are FALSe, meaning there are no missing values.

```
missing_values <- is.na(breach_data$Location_of_Breached_Information)

number_of_missing_values <- data.frame(table(missing_values))
number_of_missing_values
```

```
##   missing_values Freq
## 1          FALSE 1055
```

- 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 is unchanged since there are no missing values.

3.1.2.4 Does converting the type of this variable help exploring the distribution of its values or identifying outliers or missing values? (3) - What type can the variable be converted to?

The variable is already converted to a factor with levels.

3.1.2.5 What new variables do you need to create? (3)

I would need to create new variables that allow for a separation between those that are grouped together already, those that are "other" and combined with another type of location.

- List the variables Upon analysis, creating just one variable that allows us to just see the single location types such as just paper, or just laptops was sufficient. Fortunately, I explored this variable first and it allowed me to understand that the remaining breaches that were located in more that one place such as a breach where paper, laptops, and e-mails were breached all together was labeled as "NA" in my first bar graph (seen below). There is no need to create anymore variables. I decided to single out the singles as these types were the top most common types

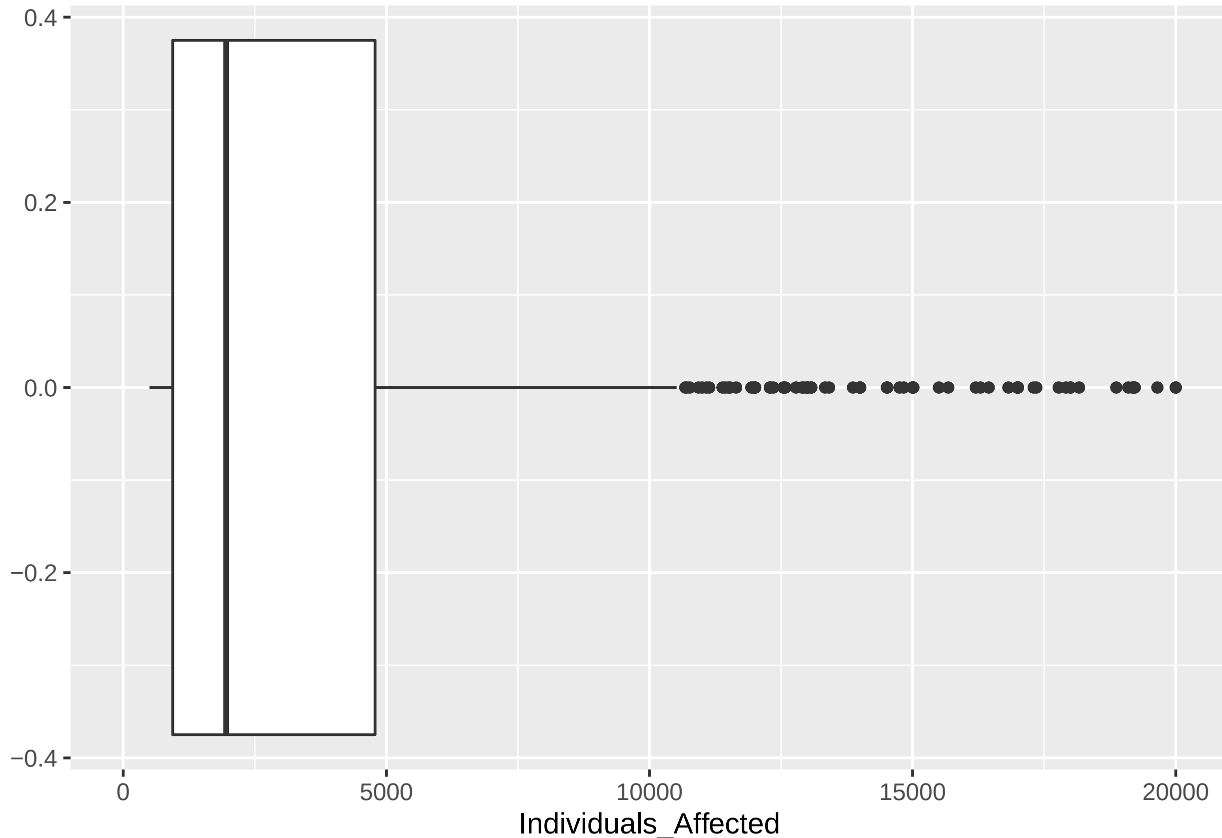
- Describe and discuss why they are needed and how you plan to use them. Explained above.

3.1.3 Individuals_affected Variable

```
#histogram <- breach_data %>%
  #ggplot(aes(x=Individuals_Affected)) + geom_boxplot()
#histogram

boxplot <- breach_data %>%
  ggplot(aes(x=Individuals_Affected)) +geom_boxplot() + xlim(0, 20000)
boxplot
```

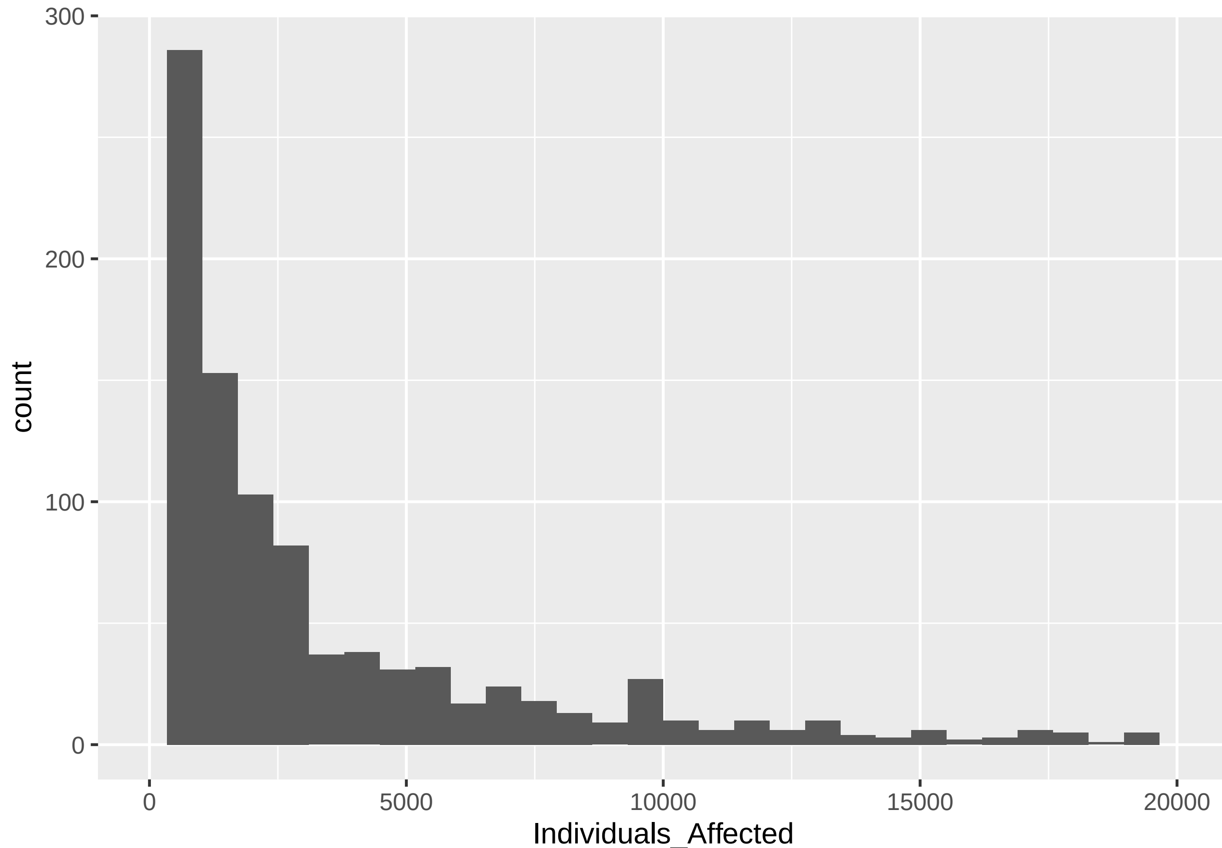
3.1.2.1 Visualising distributions (Barcharts, Histograms) (5 points)

```
## Warning: Removed 106 rows containing non-finite values (stat_boxplot).
```




```
histogram <- breach_data %>%
  ggplot(aes(x=Individuals_Affected)) + geom_histogram() + xlim(0, 20000)
histogram
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 106 rows containing non-finite values (stat_bin).

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



- Which values are the most common? Why?

```
summary(breach_data$Individuals_Affected)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     500    1000    2300   30262    6941 4900000
```

```
quantile(breach_data$Individuals_Affected)
```

```
##      0%     25%     50%     75%    100%
##     500    1000    2300    6941 4900000
```

```
sd(breach_data$Individuals_Affected)
```

```
## [1] 227859.8
```

```
30262 +  227859.8
```

```
## [1] 258121.8
```

The least is 500 number of individuals affected. **- Which values are rare? Why? Does that match your expectations?** The maximum number of individuals affected is 4,900,000. This matches my expectations as a breach of this many affected individuals must be hard to achieve.

- Can you see any unusual patterns? What might explain them? I see a spike around 10,000 individuals affected. There is not enough data in the dataset that could explain why a breach of 10k affected individuals is common.

- Are there clusters in the data? If so, There are no clusters.

- How are the observations within each cluster similar to or different from each other? No clusters observed

- How can you explain or describe the clusters? No clusters observed

3.1.2.2 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.

There are no unusual values, all look like they could be explained given a summary of what happened during the breach.

- Describe and demonstrate how you determine if they are outliers.

An outlier can be determined using the following method, which is a function that highlights which data points are outliers (using the interval provided by the interquartile range formula). Using this combined with changing it into a data frame allows us to see which values are outliers.

```
outliers_indivduals_affected <- boxplot.stats(breach_data$Individuals_Affected)$out
outliers_indivduals_affected_df <- data.frame(outliers_indivduals_affected)
outliers_indivduals_affected_df
```

```
##    outliers_indivduals_affected
## 1                         83000
## 2                         21000
## 3                         83945
## 4                        344579
## 5                         54165
## 6                         22012
## 7                        180111
## 8                         40000
## 9                         60998
## 10                      1220000
## 11                        16291
## 12                       130495
## 13                        29000
## 14                       105470
## 15                       800000
## 16                        16820
## 17                        23753
## 18                        27000
## 19                        31700
## 20                        25000
## 21                        21000
## 22                        24750
## 23                        22642
## 24                        19222
## 25                        33000
## 26                        20000
## 27                        19200
## 28                      1023209
```

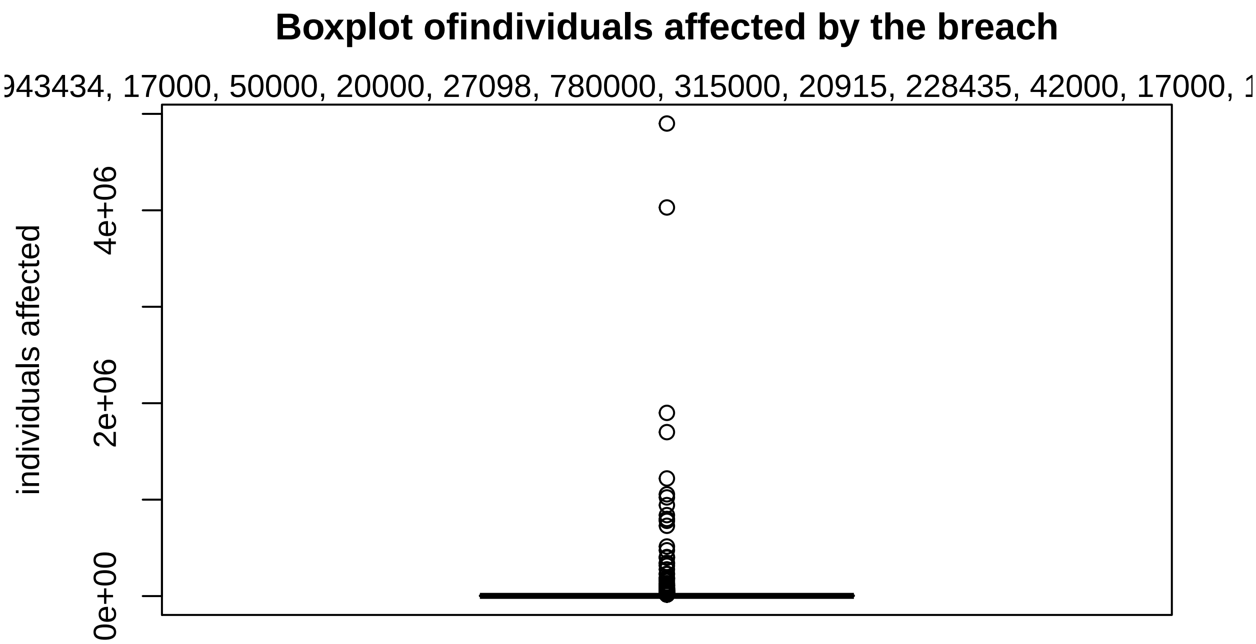
```
## 29                               475000
## 30                               115000
## 31                                24600
## 32                               156000
## 33                               231400
## 34                                16200
## 35                                18871
## 36                                37000
## 37                              1700000
## 38                                20744
## 39                               514330
## 40                                93500
## 41                                84000
## 42                               132940
## 43                              1900000
## 44                                22001
## 45                                32390
## 46                                24361
## 47                               400000
## 48                                17000
## 49                               175350
## 50                                78042
## 51                                25330
## 52                                63425
## 53                                32008
## 54                                19651
## 55                              1055489
## 56                                55000
## 57                              4900000
## 58                               943434
## 59                                17000
## 60                                50000
## 61                                20000
## 62                                27098
## 63                               780000
## 64                               315000
## 65                                20915
## 66                               228435
## 67                                42000
## 68                                17000
## 69                                19100
## 70                                66601
## 71                               105646
## 72                                55000
## 73                                64846
## 74                                65700
## 75                                27799
## 76                               116506
## 77                                18000
## 78                                28187
## 79                                35488
## 80                                28893
## 81                                27800
## 82                                56820
```

```
## 83                           19178
## 84                           29021
## 85                           16988
## 86                           43549
## 87                           18000
## 88                          109000
## 89                           28187
## 90                           18162
## 91                           17300
## 92                           22000
## 93                          189489
## 94                          187533
## 95                          277014
## 96                           21000
## 97                           32151
## 98                         4029530
## 99                           32000
## 100                          25461
## 101                          37000
## 102                         729000
## 103                          70189
## 104                          49000
## 105                          76183
## 106                          17350
## 107                          44000
## 108                          59000
## 109                         839711
## 110                          48752
## 111                          25513
## 112                          22511
## 113                         398000
## 114                          41437
## 115                         405000
## 116                          16446
## 117                          55207
## 118                          27839
## 119                          55900
## 120                          17776
## 121                         338700
## 122                          75026
## 123                          46473
## 124                          46771
## 125                          17914
## 126                          26162
## 127                          56853
## 128                          28413
## 129                          33702
```

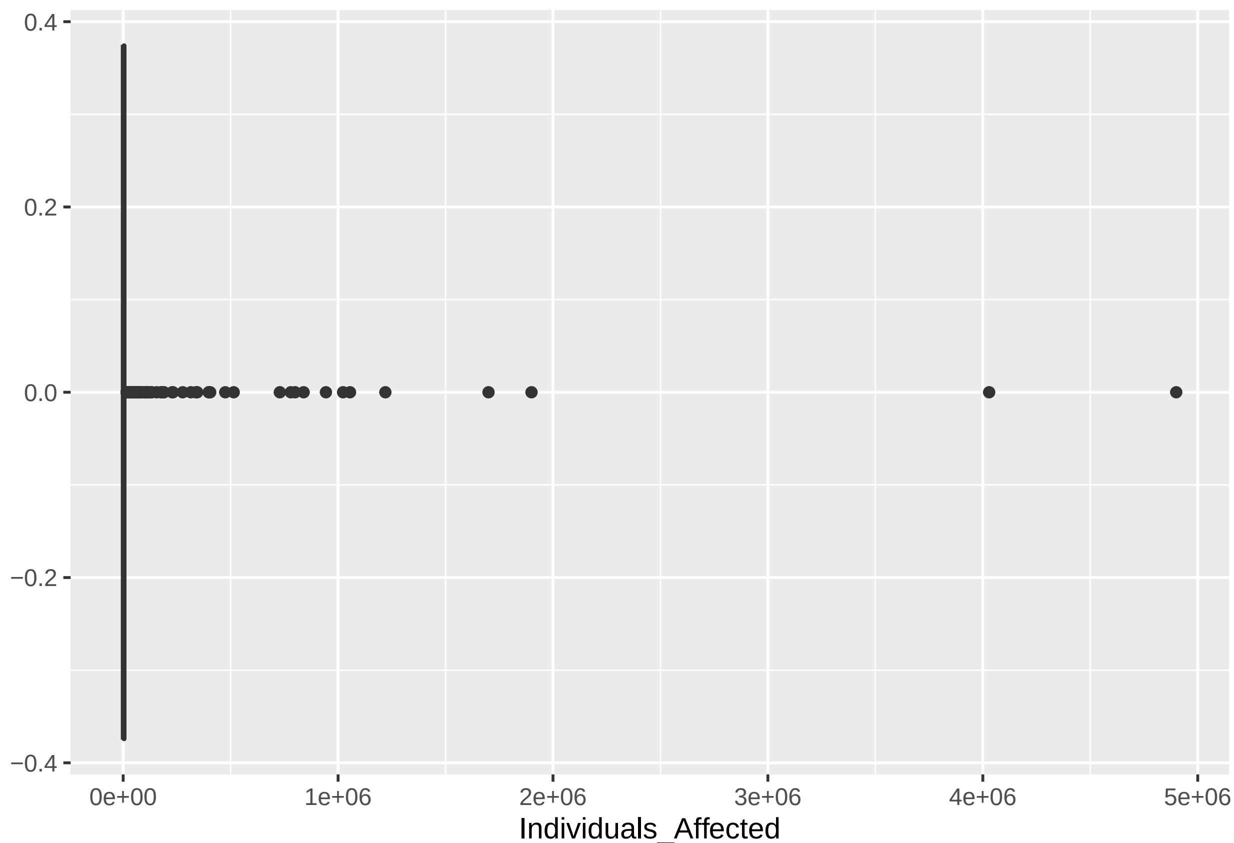
- Show how do your distributions look like with and without the unusual values.

```
boxplot(breach_data$Individuals_Affected,
  ylab = "individuals affected",
  main = "Boxplot ofindividuals affected by the breach"
)
```

```
mtext(paste("Outliers: ", paste(outliers_indivduals_affected, collapse = ", ")))
```




```
breach_data %>%
  ggplot(aes(x= Individuals_Affected)) + geom_boxplot()
```



- Discuss whether or not you need to remove unusual values and why. These values are important as they could help us understand why certain breaches affect so many individuals. #### 3.1.2.3 Missing values (2 points)

- Does this variable include missing values? Demonstrate how you determine that. There are no missing values

```
missing_values <- is.na(breach_data$Individuals_Affected)

number_of_missing_values <- data.frame(table(missing_values))
number_of_missing_values
```

```
##   missing_values Freq
## 1          FALSE 1055
```

- 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. No missing values as all 1055 values are FALSE. **- Show how your data looks in each case after handling missing values.Describe and discuss the distribution.**

No difference

3.1.2.4 Does converting the type of this variable help exploring the distribution of its values or identifying outliers or missing values? (3)

Since it is an integer, keeping it as a continuous variable is useful.

- What type can the variable be converted to? None that could help with my analysis.

- How will the distribution look? Please demonstrate with appropriate plots.

See above

3.1.2.5 What new variables do you need to create? (3) - List the variables individuals_affected_below_mean individuals_affected_above_mean

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

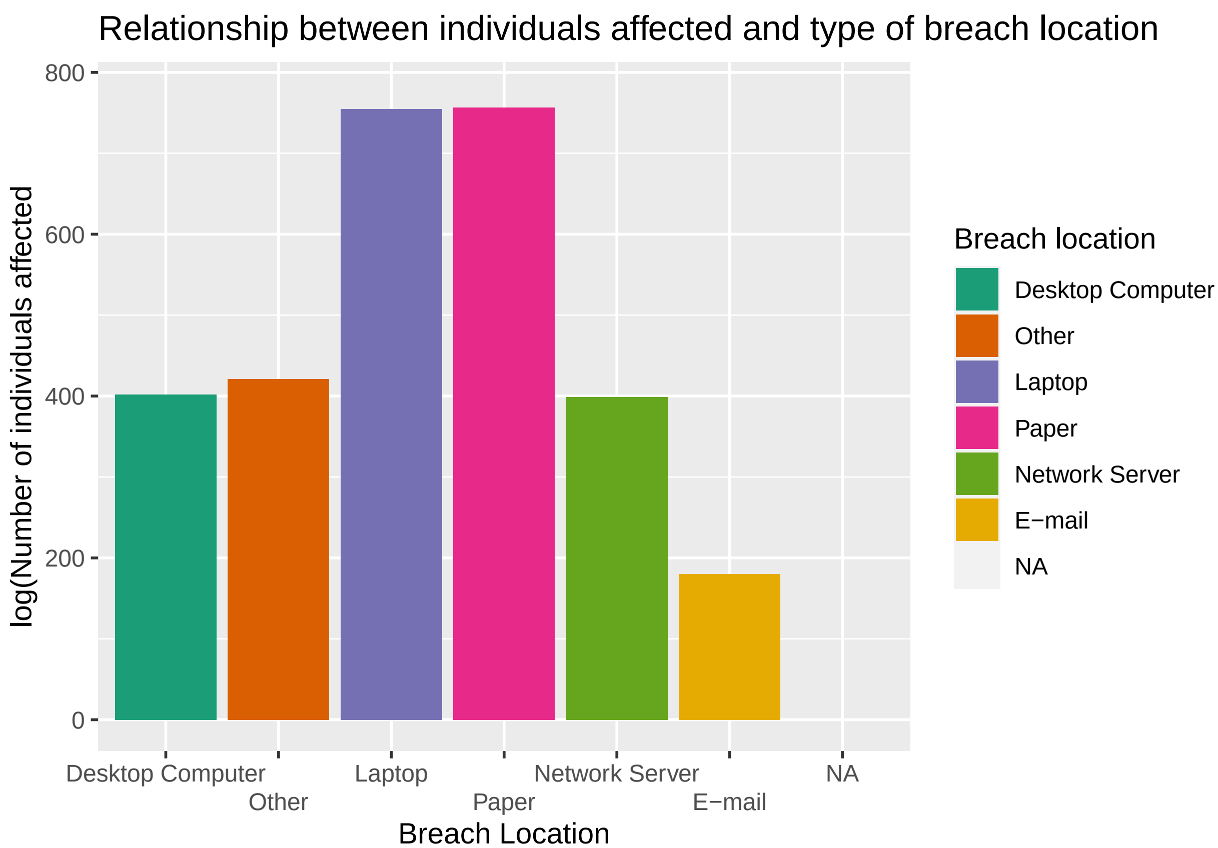
This separates the individuals_affected variable into two, a split at its mean plus one standard deviation. Looking at the data from two different perspectives might help us see later on if a certain type of location of breach is associated with a number of individuals affected that is higher than the at the split as well as lower than the split value. I chose this value to split at as I did not want to let the outliers completely skew my data.

3.2. What type of covariation occurs between the variables? (30 points)

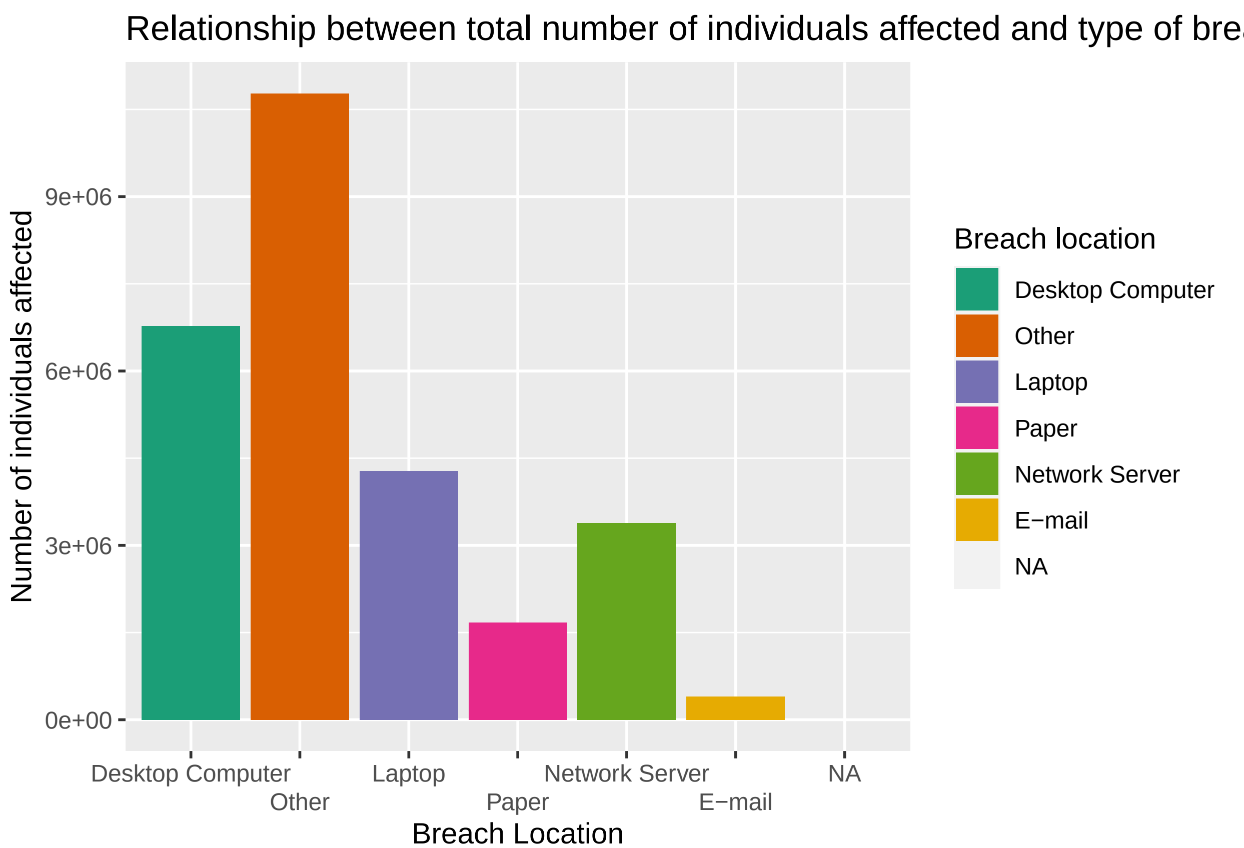
If you don't have variables of a certain type in the original dataset or among the created variables (features), you can further create them from the existing variables. See RDS chap. 5, 7.5 and 7.6.

3.2.1 Between a categorical and continuous variable (10 points)

```
singles_levels = c("Desktop Computer" , "Other", "Laptop" , "Paper", "Network Server","E-mail")
location_singles <- factor(breach_data$Location_of_Breached_Information, levels = singles_levels)


breach_data %>%
  ggplot(aes(x=location_singles, y=log10(Individuals_Affected), fill =location_singles)) +geom_col() + g
  xlab("Breach Location") + ylab("log(Number of individuals affected") + labs(fill = "Breach location")
```




```
singles_levels = c("Desktop Computer" , "Other", "Laptop" , "Paper", "Network Server","E-mail")
location_singles <- factor(breach_data$Location_of_Breached_Information, levels = singles_levels)


breach_data %>%
  ggplot(aes(x=location_singles, y=Individuals_Affected, fill =location_singles)) +geom_col() + ggtitle
```



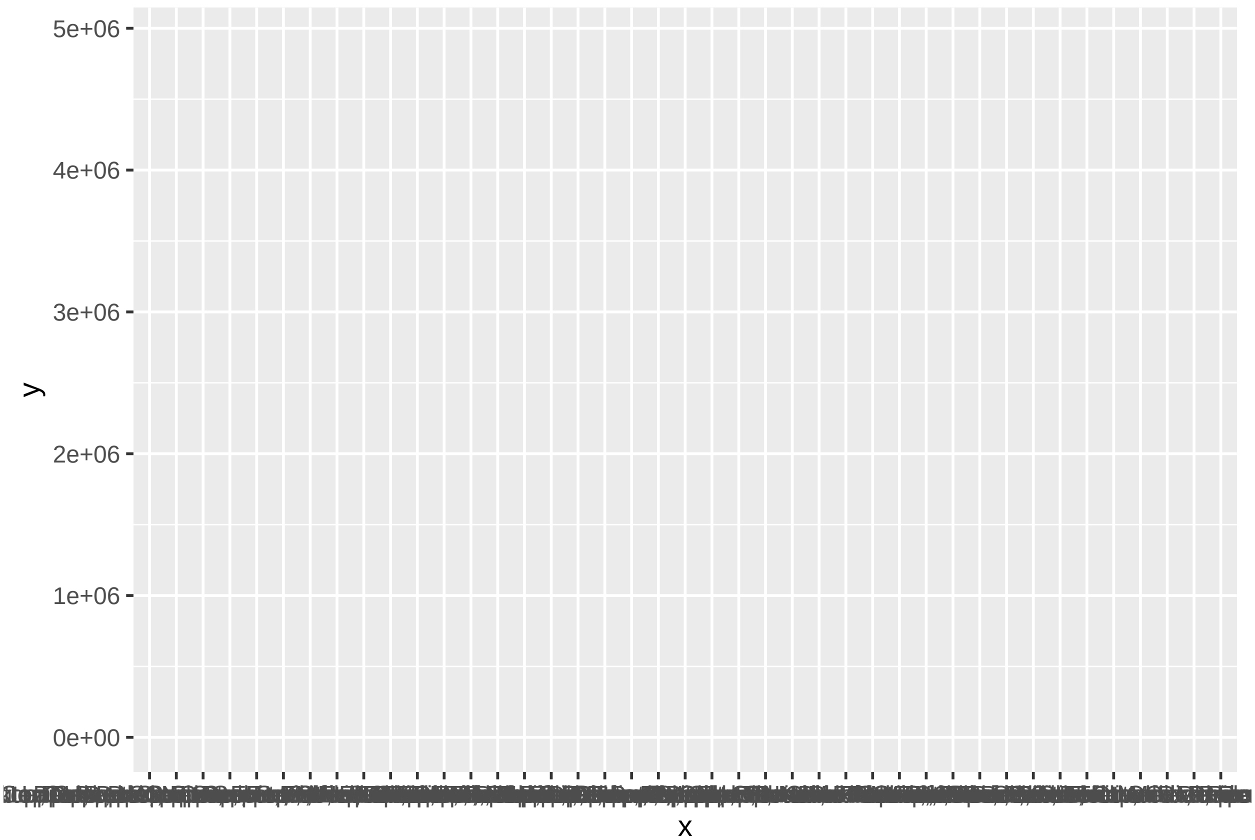

```
#mutate(breach_data, below = breach_data$Individuals_Affected <= 258121.8)

singles_levels = c("Desktop Computer" , "Other", "Laptop" , "Paper", "Network Server","E-mail")
location_singles <- factor(breach_data$Location_of_Breached_Information, levels = singles_levels)
```

```
#below_data %>%
 # filter(below == FALSE) %>%
  #ggplot(breach_data, mapping= aes(x=location_singles, y= Individuals_Affected)) +geom_col()
head(breach_data)
```

```
## # A tibble: 6 x 14
##      X1 Number Name_of_Covered_Entity  State Business_Associat~ Individuals_Aff~
##   <dbl>  <dbl> <chr>                   <chr> <chr>                         <dbl>
## 1     1      0 Brooke Army Medical Ce~ TX    <NA>                           1000
## 2     2      1 Mid America Kidney Sto~ MO    <NA>                           1000
## 3     3      2 Alaska Department of H~ AK    <NA>                            501
## 4     4      3 Health Services for Ch~ DC    <NA>                           3800
## 5     5      4 L. Douglas Carlson, M.~ CA    <NA>                           5257
## 6     6      5 David I. Cohen, MD      CA    <NA>                            857
## # ... with 8 more variables: 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>
```

- Calculate the strength of the relationship implied by the pattern (e.g., correlation)

```
x <- breach_data$Location_of_Breached_Information
y <- breach_data$Individuals_Affected
ggplot(breach_data, aes(x,y))
```



- Discuss what other variables might affect the relationship

Other variables that could affect this relationship could be state. Wealthier areas have access to more secure measures that affect less individuals compared to areas that are less wealthy and have less secure measures of storing data.

- Does the relationship change if you look at individual subgroups of the data? Please discuss and demonstrate.

By looking at subgroups where we can just observe singular instances of location versus breaches where multiple locations were involved, we can easily understand the relationship between these two variables.

- Discuss how the observed patterns support/reject your hypotheses or answer your questions.
Upon graphing just the first variable, location__of__breached__information, I assumed paper would be the type of breach location that affected individuals the most.We can also see that paper and laptop breaches are the most common types reported. Thus, it would make sense that the two types of breaches that are most common are also the two types that impact the most number of individuals. As to why paper and laptops are the most common, this can be attributed to the fact that paper can be easily absconded with as there are less robust security measures. Laptops, while they do have significantly more secure protocols for protecting information, can still be breached due to weak passwords, the fact that they are easily portable, and a variety of other reasons. Breaching the data in a laptop is still significantly easier than hacking a network server for example. The impact of network servers on the number of individuals is among the least relatively. When considering why these two types of breaches affect the most number of individuals, it's important to understand that most data is stored in either of these two formats. Companies and organizations still record

social security numbers, customer images and other information in files on their laptops, paper files, paper binders, et cetera.