Alfredo Rojas - BIS Data Analyst - Visualization Demo

Visualization demo

Started by importing libraries

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(plotly)
library(stringr)
library(DataExplorer)
```

And loading the dataset

```
setwd("C:/Users/alfrs/Documents/git/RProjects/Experian")
dataset <- readxl::read_xlsx("DATA_FILE_FOR_INTERVIEW.xlsx")
head(dataset)
```

```
## # A tibble: 6 x 10
##   COMPANY_NAME CITY  STATE ZIP   COUNTRY PHONE YEAR_INCORP ANNUAL_SALES
##   <chr>        <chr> <chr> <chr> <chr>   <chr> <chr>              <dbl>
## 1 AAR Corp     Wood~ IL    60191 USA     630 ~ 1955          2051800000
## 2 AFA Protect~ Syos~ NY    11791 USA     516 ~ 1873            73220115
## 3 American Lo~ DFW ~ TX    75261 USA     817 ~ 1898            14625889
## 4 Abbott Labo~ Abbo~ IL    60064 USA     224 ~ 1900         30578000000
## 5 ACMAT Corp.  Farm~ CT    06032 USA     860 ~ 1951             2750729
## 6 Acme United~ Fair~ CT    06824 USA     203 ~ 1867           137321395
## # ... with 2 more variables: EMPLOYEE_COUNT <dbl>, NET_INCOME <dbl>
```

Then I checked column types and reassigned those I thought needed reassignment

```
str(dataset)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    8382 obs. of  10 variables:
##  $ COMPANY_NAME  : chr  "AAR Corp" "AFA Protective Systems, Inc." "American Locker Group, Inc." "Abb
##  $ CITY          : chr  "Wood Dale" "Syosset" "DFW Airport" "Abbott Park" ...
##  $ STATE         : chr  "IL" "NY" "TX" "IL" ...
##  $ ZIP           : chr  "60191" "11791" "75261" "60064" ...
##  $ COUNTRY       : chr  "USA" "USA" "USA" "USA" ...
##  $ PHONE         : chr  "630 227-2000" "516 496-2322" "817 329-1600" "224 667-6100" ...
##  $ YEAR_INCORP   : chr  "1955" "1873" "1898" "1900" ...
##  $ ANNUAL_SALES  : num  2.05e+09 7.32e+07 1.46e+07 3.06e+10 2.75e+06 ...
##  $ EMPLOYEE_COUNT: num  6550 0 120 103000 NA 435 703 10100 16300 570 ...
##  $ NET_INCOME    : num  7.50e+06 2.60e+05 -2.82e+06 2.37e+09 7.44e+05 ...
```

```r
dataset$COMPANY_NAME <- as.factor(dataset$COMPANY_NAME)
dataset$CITY <- as.factor(dataset$CITY)
dataset$STATE <- as.factor(dataset$STATE)
dataset$ZIP <- as.factor(dataset$ZIP)
dataset$COUNTRY <- as.factor(dataset$COUNTRY)
dataset$PHONE <- as.factor(dataset$PHONE)
dataset$YEAR_INCORP <- as.numeric(dataset$YEAR_INCORP)
dataset$ANNUAL_SALES <- as.double(dataset$ANNUAL_SALES)
dataset$EMPLOYEE_COUNT <- as.double(dataset$EMPLOYEE_COUNT)
dataset$NET_INCOME <- as.double(dataset$NET_INCOME)

summary(dataset)
```

```
##                                 COMPANY_NAME         CITY
##  024 Pharma Inc                      :   1   New York : 473
##  1-800 Flowers.com, Inc.             :   1   Houston  : 262
##  10x Genomics Inc                    :   1   Las Vegas: 188
##  11 Good Energy Inc                  :   1   Dallas   : 130
##  1347 Property Insurance Holdings Inc:   1   San Diego: 109
##  180 Degree Capital Corp             :   1   (Other)  :7117
##  (Other)                             :8376   NA's     : 103
##      STATE           ZIP          COUNTRY              PHONE
##  CA     :1280   10022  :  84   USA    :7283   800 983-0903:  11
##  NY     : 739   77002  :  67   CHN    : 308   855 588-7839:   8
##  TX     : 733   92121  :  47   CAN    : 205   510 522-9600:   7
##  FL     : 553   80202  :  43   HKG    :  91   512 236-6555:   6
##  NV     : 315   10019  :  36   ISR    :  80   800 736-3402:   6
##  (Other):4175   (Other):7968   (Other): 414   (Other)     :8310
##  NA's   : 587   NA's   : 137   NA's   :   1   NA's        :  34
##   YEAR_INCORP    ANNUAL_SALES        EMPLOYEE_COUNT
##  Min.   :1784   Min.   :-2.781e+08   Min.   :      0
##  1st Qu.:1986   1st Qu.: 4.095e+06   1st Qu.:     13
##  Median :1999   Median : 8.760e+07   Median :    187
##  Mean   :1991   Mean   : 2.709e+09   Mean   :   6671
##  3rd Qu.:2008   3rd Qu.: 9.903e+08   3rd Qu.:   2228
##  Max.   :2019   Max.   : 5.144e+11   Max.   :2200000
##  NA's   :157    NA's   :1625         NA's   :1668
##    NET_INCOME
##  Min.   :-2.244e+10
##  1st Qu.:-5.216e+06
##  Median :-7.610e+04
##  Mean   : 1.758e+08
##  3rd Qu.: 2.798e+07
##  Max.   : 5.953e+10
##  NA's   :25
```

From this summary, I can see that:

A. We have NAs in Year, however, it does not makes sense to change them as Year is a very specific column.

B. We have NAs in Employee Count. These can be replaced, we'll need to analyze to determine with what.

C. We have NAs in COuntry. Again, this is very specific, so we can just exclude it.

D. We have NAs in Annual Sales. This can also be replaced, so need to analyze this.

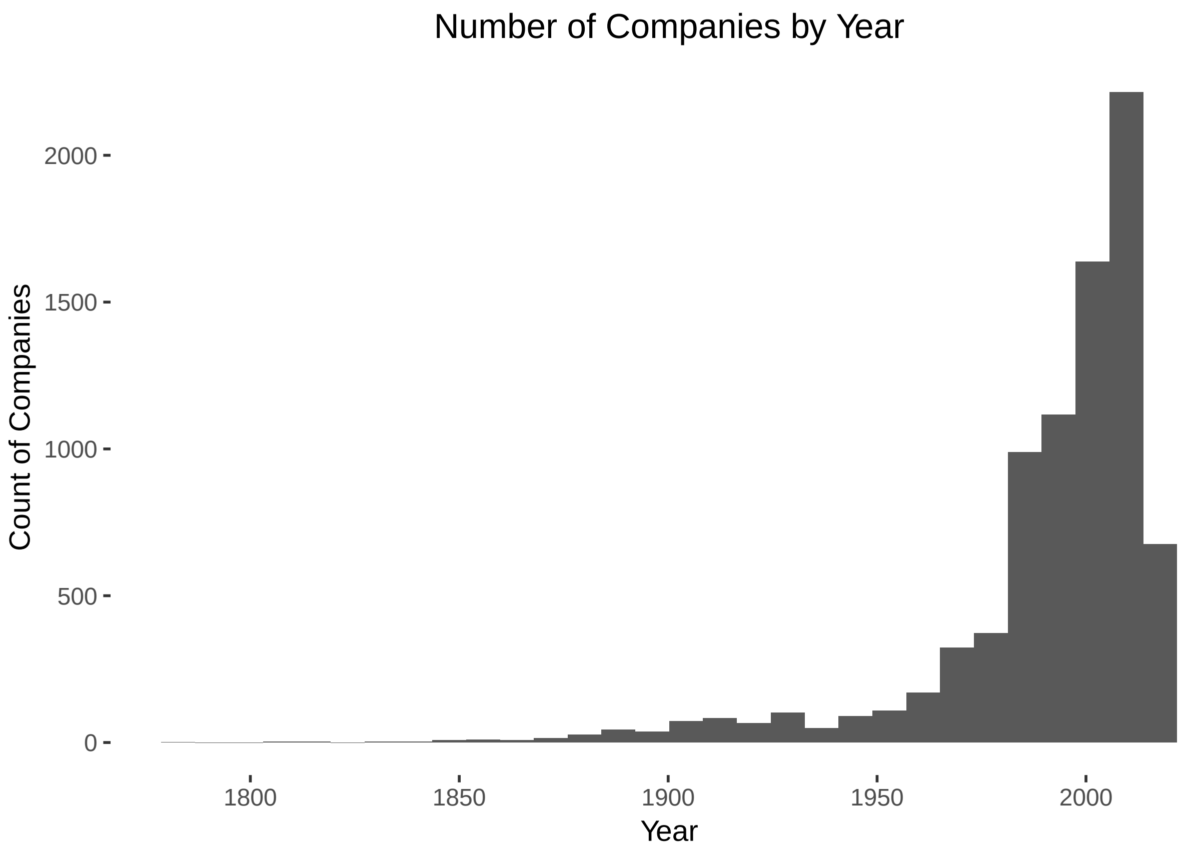
I started working on the first point. A distribution of companies by year sounded very easy, but the graph said otherwise.

As a matter of fact, the graph was so big that it can't really show in the document

```r
year_graph_1 <-
ggplot(dataset %>%
         count(YEAR_INCORP),
       aes(x = YEAR_INCORP,
           y = n,
           fill = as.factor(YEAR_INCORP)))+
  geom_col()+
  #geom_text(aes(label = n),position = position_stack(vjust = .5)) +
  ggtitle("Count of Companies by Year Bucket")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank())+
  ylab("Count of Companies by Year") +
  labs(fill = "Year Incorporated") +
  coord_flip()
```

So then tried a histogram, since we're using years, maybe I can see which are the most valuable, however…

```r
ggplot(dataset,aes(x = YEAR_INCORP))+
  geom_histogram() +
  ggtitle("Number of Companies by Year")+
  theme(panel.background = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  xlab("Year") +
  ylab("Count of Companies")
```

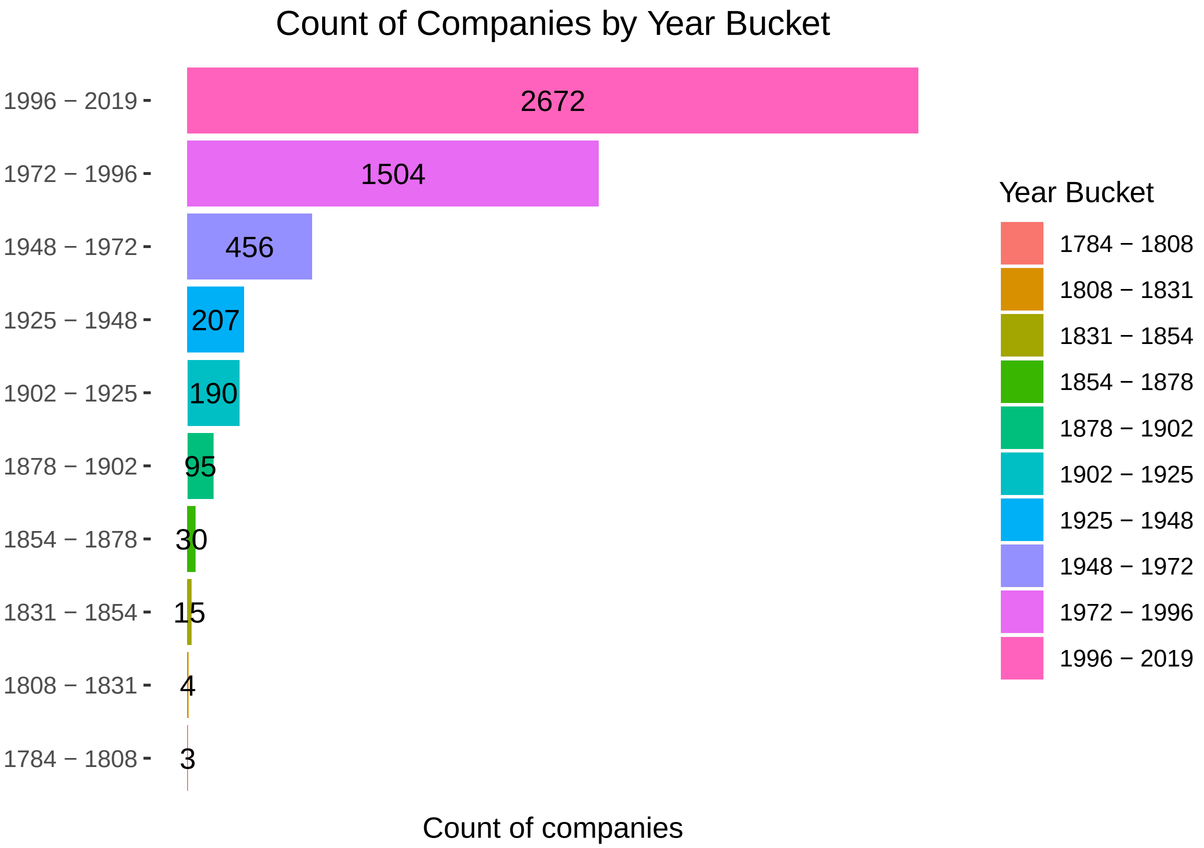


The histogram did show me that most of my companies are arround the 2000's

But I can't really see which year is the most valuable

So I did year buckets, that would allow me to graph and see the data in a more manageable way

```
dataset$YEAR_BUCKET <- cut(dataset$YEAR_INCORP,dig.lab=4,breaks=10)
dataset$YEAR_BUCKET <- str_replace(dataset$YEAR_BUCKET, "\\(", "")
dataset$YEAR_BUCKET <- str_replace(dataset$YEAR_BUCKET, "]", "")
dataset$YEAR_BUCKET <- str_replace(dataset$YEAR_BUCKET, ",", " - ")

dataset$YEAR_BUCKET <- as.factor(dataset$YEAR_BUCKET)

ggplot(na.exclude(dataset) %>%
         count(YEAR_BUCKET),
       aes(x = YEAR_BUCKET,
           y = n,
           fill = as.factor(YEAR_BUCKET)))+
  geom_col()+
  geom_text(aes(label = n),position = position_stack(vjust = .5)) +
  ggtitle("Count of Companies by Year Bucket")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        panel.background = element_blank())+
  ylab("Count of companies") +
  xlab(element_blank()) +
  labs(fill = "Year Bucket") +
  coord_flip()
```



Thanks to the buckets, I saw that most of my data is from 1972 going forward. So I did buckets again, but only with these years.

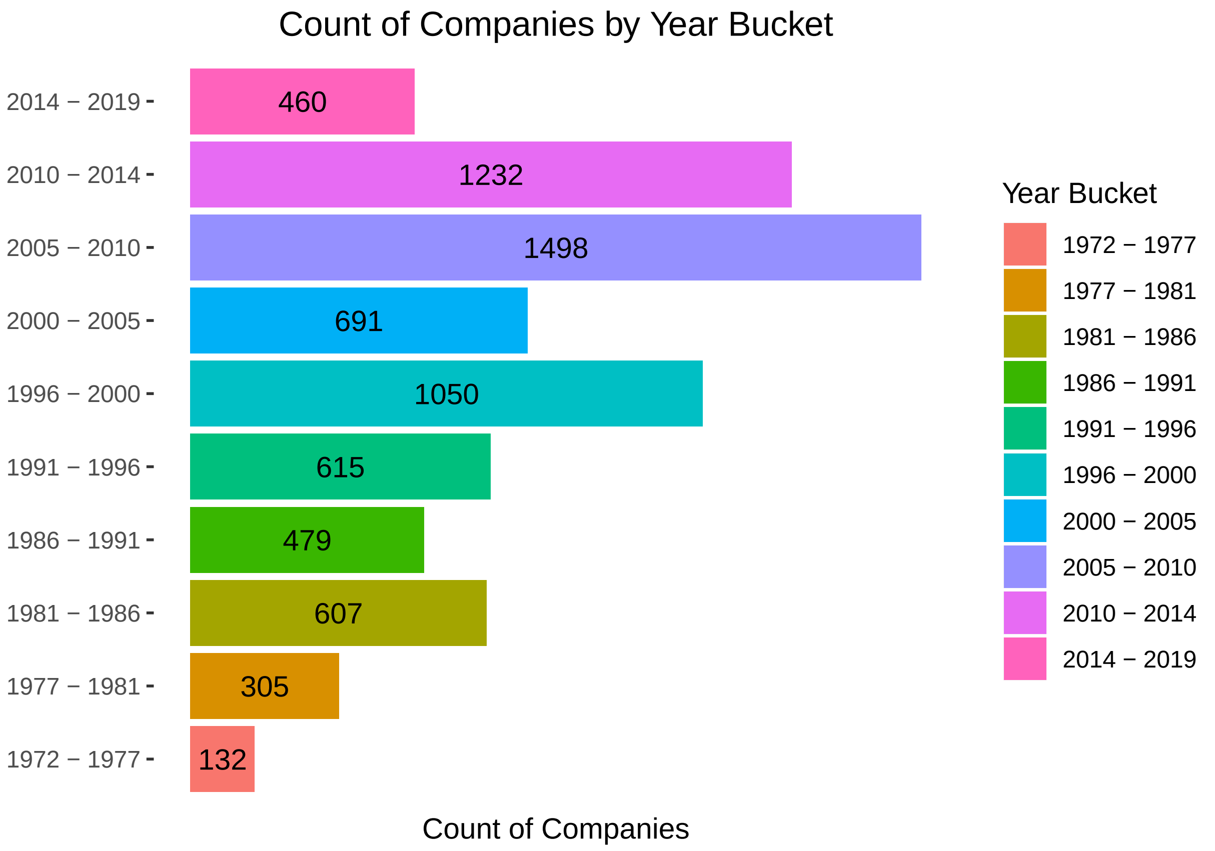
```
dataset_filtered <- dataset %>% filter(YEAR_INCORP >= 1972)
dataset_filtered$YEAR_BUCKET <- cut(dataset_filtered$YEAR_INCORP,dig.lab=4,breaks=10)
dataset_filtered$YEAR_BUCKET <- str_replace(dataset_filtered$YEAR_BUCKET, "\\(", "")
dataset_filtered$YEAR_BUCKET <- str_replace(dataset_filtered$YEAR_BUCKET, "]", "")
dataset_filtered$YEAR_BUCKET <- str_replace(dataset_filtered$YEAR_BUCKET, ",", " - ")

dataset_filtered$YEAR_BUCKET <- as.factor(dataset_filtered$YEAR_BUCKET)

summary(dataset_filtered)
```

```
##                                 COMPANY_NAME         CITY
##  024 Pharma Inc                      :   1   New York : 400
##  1-800 Flowers.com, Inc.             :   1   Houston  : 223
##  10x Genomics Inc                    :   1   Las Vegas: 181
##  11 Good Energy Inc                  :   1   Dallas   : 106
##  1347 Property Insurance Holdings Inc:   1   San Diego: 104
##  180 Degree Capital Corp             :   1   (Other)  :5956
##  (Other)                             :7063   NA's     :  99
##      STATE           ZIP          COUNTRY             PHONE
##  CA     :1175   10022  :  73   USA    :6021   800 983-0903:  11
##  TX     : 622   77002  :  57   CHN    : 298   510 522-9600:   7
##  NY     : 599   92121  :  46   CAN    : 196   512 236-6555:   6
##  FL     : 487   80202  :  42   HKG    :  91   855 588-7839:   6
##  NV     : 298   94080  :  31   ISR    :  76   214 981-0700:   4
##  (Other):3331   (Other):6685   (Other): 386   (Other)     :7001
##  NA's   : 557   NA's   : 135   NA's   :   1   NA's        :  34
##   YEAR_INCORP    ANNUAL_SALES        EMPLOYEE_COUNT     NET_INCOME
##  Min.   :1972   Min.   :-2.781e+08   Min.   :     0   Min.   :-5.086e+09
##  1st Qu.:1993   1st Qu.: 2.045e+06   1st Qu.:     9   1st Qu.:-6.497e+06
##  Median :2003   Median : 4.817e+07   Median :   107   Median :-2.498e+05
##  Mean   :2000   Mean   : 1.781e+09   Mean   :  4014   Mean   : 1.167e+08
##  3rd Qu.:2009   3rd Qu.: 5.341e+08   3rd Qu.:  1122   3rd Qu.: 1.011e+07
##  Max.   :2019   Max.   : 2.656e+11   Max.   :647500   Max.   : 5.953e+10
##                 NA's   :1572         NA's   :1494     NA's   :20
##       YEAR_BUCKET
##  2005 - 2010:1498
##  2010 - 2014:1232
##  1996 - 2000:1050
##  2000 - 2005: 691
##  1991 - 1996: 615
##  1981 - 1986: 607
##  (Other)    :1376
```

With this summary I wanted to check the distribution of the Year Bucket

This graph looks better, now I can see that most of the companies got incorporated between 2005 and 2010.

```
ggplot(dataset_filtered %>%
         count(YEAR_BUCKET),
       aes(x = YEAR_BUCKET,
           y = n,
           fill = as.factor(YEAR_BUCKET)
           )
       )+
  geom_col()+
  geom_text(aes(label = n),
            position = position_stack(vjust = .5)) +
  ggtitle("Count of Companies by Year Bucket")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        panel.background = element_blank())+
  ylab("Count of Companies") +
  xlab(element_blank()) +
  labs(fill = "Year Bucket") +
  coord_flip()
```



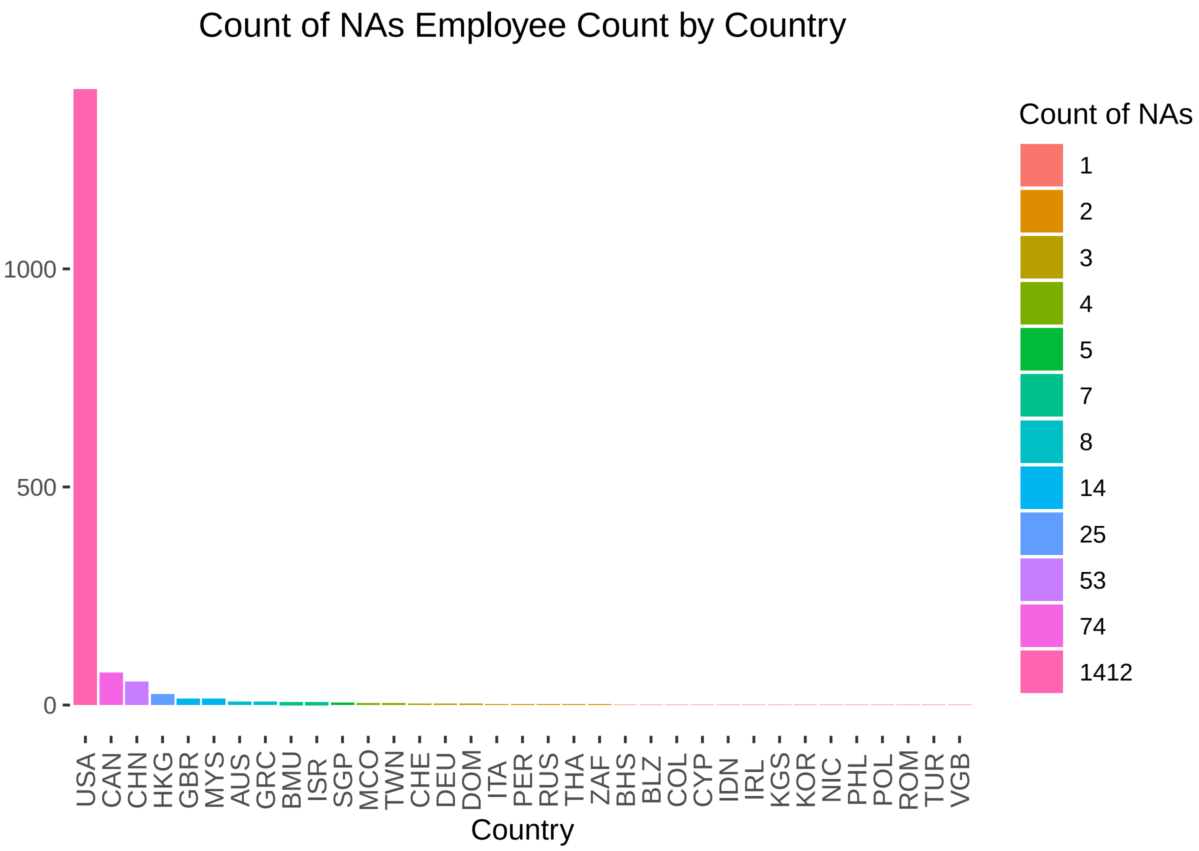
And so, now I know that most of my value is in the time period that goes from 2005 to 2010.

I then wanted to see a distribution of companies by employee count.

But first, I wanted to see if we had any NAs in the data, and how can we replace them.

```
dataset_emp <- dataset %>%
  select(EMPLOYEE_COUNT,COUNTRY) %>%
  filter(is.na(EMPLOYEE_COUNT)) %>%
  count(COUNTRY)

dataset_emp$COUNTRY <-
  factor(dataset_emp$COUNTRY,
         levels = dataset_emp$COUNTRY[order(dataset_emp$n,
                                            decreasing = TRUE)])
```

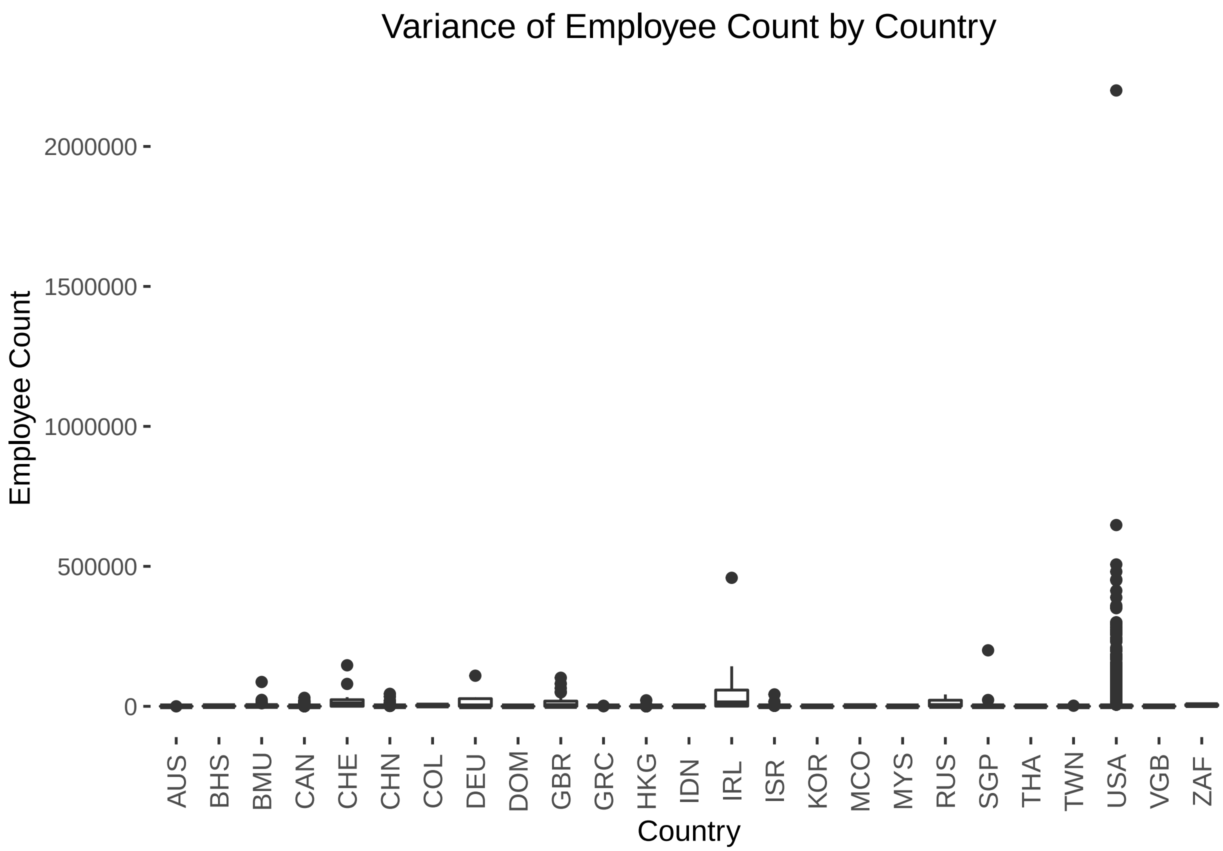
```
dataset_emp %>%
  ggplot(aes(x=COUNTRY,
             y=n,
             fill=as.factor(n))
         )+
  geom_col()+
  ggtitle("Count of NAs Employee Count by Country")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(size = 10,
                                   angle = 90,
                                   hjust = .5,
                                   vjust = .5),
        panel.background = element_blank()) +
  ylab(element_blank()) +
  xlab("Country") +
  labs(fill = "Count of NAs")
```



Unsurprisingly USA has the highest amount of NAs, which correlates with it having the highest amount of companies

Then I wanted to see the variance of Employee Counts, of all countries where I had NAs, but where I had no NA values

```r
dataset_emp_noNA <-
dataset %>%
  select(EMPLOYEE_COUNT,COUNTRY) %>%
  filter(!is.na(EMPLOYEE_COUNT))

dataset_emp_noNA <-
  merge(dataset_emp_noNA %>%
          select(COUNTRY,EMPLOYEE_COUNT),
        dataset_emp %>%
          select(COUNTRY),all=FALSE)

dataset_emp_noNA %>%
  ggplot(aes(x=COUNTRY,y=EMPLOYEE_COUNT))+
  geom_boxplot()+
  ggtitle("Variance of Employee Count by Country")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(size = 10,
                                   angle = 90,
                                   hjust = .5,
                                   vjust = .5),
        panel.background = element_blank()) +
  ylab("Employee Count") +
  xlab("Country")
```



With this data, I decided that the best way to replace the NAs, was to use the median by country.

The reason behind this is because of the outliers in the data, means we can use the median as better measure.

```
for (country in unique(dataset$COUNTRY)){

  dataset_fil <-
    dataset %>%
    filter(!is.na(EMPLOYEE_COUNT)) %>%
    filter(COUNTRY == country)

  dataset$EMPLOYEE_COUNT <-
    replace_na(dataset$EMPLOYEE_COUNT,
               quantile(na.exclude(dataset_fil$EMPLOYEE_COUNT),
                        probs=0.5))
}

summary(dataset)
```

```
##                                 COMPANY_NAME         CITY
##  024 Pharma Inc                       :   1   New York : 473
##  1-800 Flowers.com, Inc.             :   1   Houston  : 262
##  10x Genomics Inc                    :   1   Las Vegas: 188
##  11 Good Energy Inc                  :   1   Dallas   : 130
##  1347 Property Insurance Holdings Inc:   1   San Diego: 109
##  180 Degree Capital Corp             :   1   (Other)  :7117
##  (Other)                             :8376   NA's     : 103
##      STATE           ZIP          COUNTRY              PHONE
##  CA      :1280   10022  :  84   USA    :7283   800 983-0903:  11
##  NY      : 739   77002  :  67   CHN    : 308   855 588-7839:   8
##  TX      : 733   92121  :  47   CAN    : 205   510 522-9600:   7
##  FL      : 553   80202  :  43   HKG    :  91   512 236-6555:   6
##  NV      : 315   10019  :  36   ISR    :  80   800 736-3402:   6
##  (Other):4175   (Other):7968   (Other): 414   (Other)     :8310
##  NA's    : 587   NA's   : 137   NA's   :   1   NA's        :  34
##   YEAR_INCORP    ANNUAL_SALES        EMPLOYEE_COUNT
##  Min.   :1784   Min.   :-2.781e+08   Min.   :      0
##  1st Qu.:1986   1st Qu.: 4.095e+06   1st Qu.:     28
##  Median :1999   Median : 8.760e+07   Median :    207
##  Mean   :1991   Mean   : 2.709e+09   Mean   :   5385
##  3rd Qu.:2008   3rd Qu.: 9.903e+08   3rd Qu.:   1177
##  Max.   :2019   Max.   : 5.144e+11   Max.   :2200000
##  NA's   :157    NA's   :1625
##    NET_INCOME              YEAR_BUCKET
##  Min.   :-2.244e+10   1996 - 2019:4931
##  1st Qu.:-5.216e+06   1972 - 1996:2107
##  Median :-7.610e+04   1948 - 1972: 567
##  Mean   : 1.758e+08   1925 - 1948: 229
##  3rd Qu.: 2.798e+07   1902 - 1925: 221
##  Max.   : 5.953e+10   (Other)    : 170
##  NA's   :25           NA's       : 157
```

With this summary I wanted to check the new values of the Employee Count column

I decided to use the same approach as before, where I created buckets to see the distribution

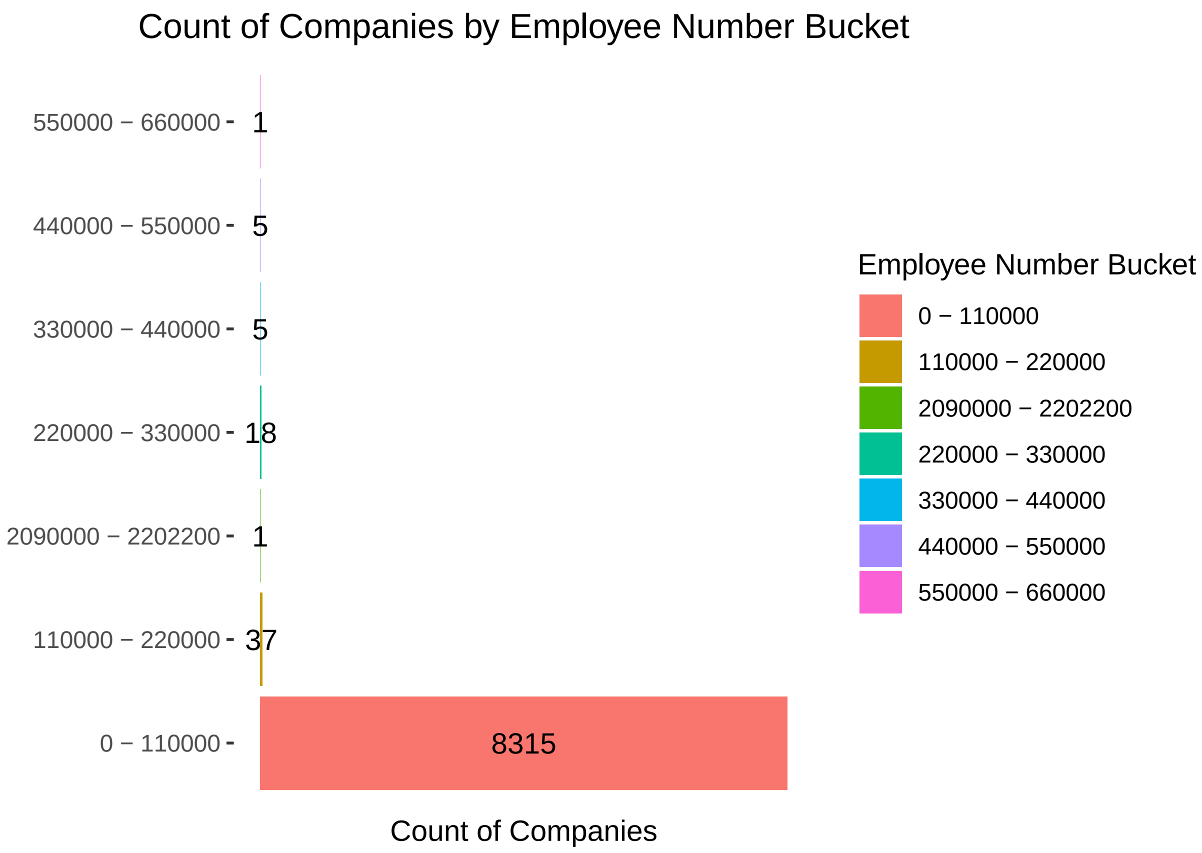
```
dataset$EMPLOYEE_COUNT_BUCKET <- cut(dataset$EMPLOYEE_COUNT,breaks = 20,dig.lab = 10)
dataset$EMPLOYEE_COUNT_BUCKET <- str_replace(dataset$EMPLOYEE_COUNT_BUCKET, "\\(", "")
dataset$EMPLOYEE_COUNT_BUCKET <- str_replace(dataset$EMPLOYEE_COUNT_BUCKET, "]", "")
dataset$EMPLOYEE_COUNT_BUCKET <- str_replace(dataset$EMPLOYEE_COUNT_BUCKET, ",", " - ")
dataset$EMPLOYEE_COUNT_BUCKET <- str_replace(dataset$EMPLOYEE_COUNT_BUCKET, "-2200", "0")

dataset$EMPLOYEE_COUNT_BUCKET <- as.factor(dataset$EMPLOYEE_COUNT_BUCKET)

summary(dataset)
```

```
##                                COMPANY_NAME         CITY
##  024 Pharma Inc                      :   1   New York : 473
##  1-800 Flowers.com, Inc.             :   1   Houston  : 262
##  10x Genomics Inc                    :   1   Las Vegas: 188
##  11 Good Energy Inc                    :   1   Dallas   : 130
##  1347 Property Insurance Holdings Inc:   1   San Diego: 109
##  180 Degree Capital Corp             :   1   (Other)  :7117
##  (Other)                             :8376   NA's     : 103
##      STATE           ZIP          COUNTRY              PHONE
##  CA     :1280   10022  :   84   USA    :7283   800 983-0903:  11
##  NY     : 739   77002  :   67   CHN    : 308   855 588-7839:   8
##  TX     : 733   92121  :   47   CAN    : 205   510 522-9600:   7
##  FL     : 553   80202  :   43   HKG    :  91   512 236-6555:   6
##  NV     : 315   10019  :   36   ISR    :  80   800 736-3402:   6
##  (Other):4175   (Other):7968   (Other): 414   (Other)     :8310
##  NA's   : 587   NA's   : 137   NA's   :   1   NA's        :  34
##   YEAR_INCORP    ANNUAL_SALES        EMPLOYEE_COUNT
##  Min.   :1784   Min.   :-2.781e+08   Min.   :      0
##  1st Qu.:1986   1st Qu.: 4.095e+06   1st Qu.:     28
##  Median :1999   Median : 8.760e+07   Median :    207
##  Mean   :1991   Mean   : 2.709e+09   Mean   :   5385
##  3rd Qu.:2008   3rd Qu.: 9.903e+08   3rd Qu.:   1177
##  Max.   :2019   Max.   : 5.144e+11   Max.   :2200000
##  NA's   :157    NA's   :1625
##    NET_INCOME              YEAR_BUCKET          EMPLOYEE_COUNT_BUCKET
##  Min.   :-2.244e+10   1996 - 2019:4931   0 - 110000       :8315
##  1st Qu.:-5.216e+06   1972 - 1996:2107   110000 - 220000  :  37
##  Median :-7.610e+04   1948 - 1972: 567   2090000 - 2202200:   1
##  Mean   : 1.758e+08   1925 - 1948: 229   220000 - 330000  :  18
##  3rd Qu.: 2.798e+07   1902 - 1925: 221   330000 - 440000  :   5
##  Max.   : 5.953e+10   (Other)    : 170   440000 - 550000  :   5
##  NA's   :25           NA's       : 157   550000 - 660000  :   1
```

With this summary I wanted to check the Employee Count Bucket column

And with the graphic, I realized that most companies have between 0 and 110,000 employees.

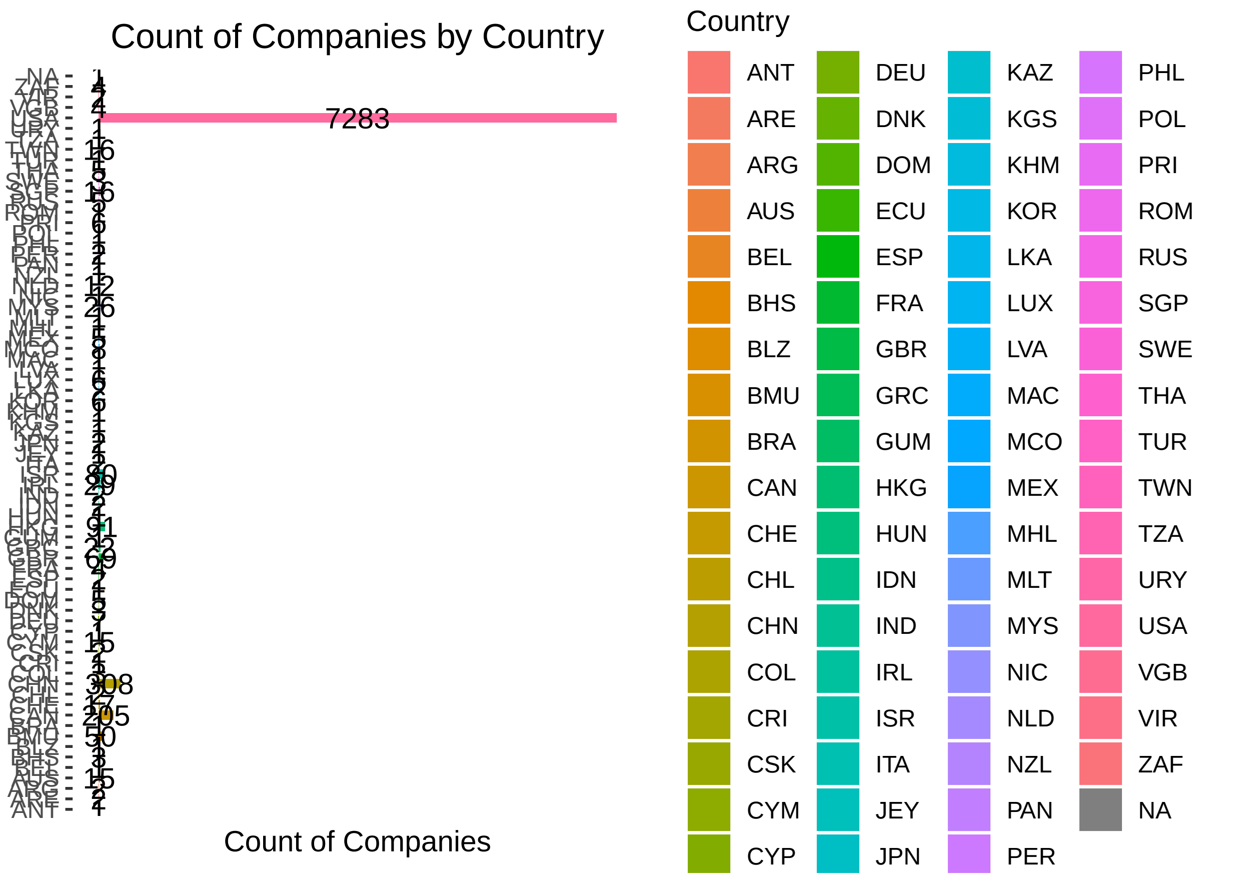
```r
ggplot(dataset %>%
  count(EMPLOYEE_COUNT_BUCKET),
  aes(x = EMPLOYEE_COUNT_BUCKET,
      y=n,
      fill=EMPLOYEE_COUNT_BUCKET))+
  geom_col()+
  geom_text(aes(label = n),position = position_stack(vjust = .5)) +
  ggtitle("Count of Companies by Employee Number Bucket")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        panel.background = element_blank())+
  ylab("Count of Companies") +
  xlab(element_blank()) +
  labs(fill = "Employee Number Bucket") +
  coord_flip()
```



As a third point, I wanted to see the distribution by Country.

This distribution sounds easy enough, but again, the graphic shows otherwise.

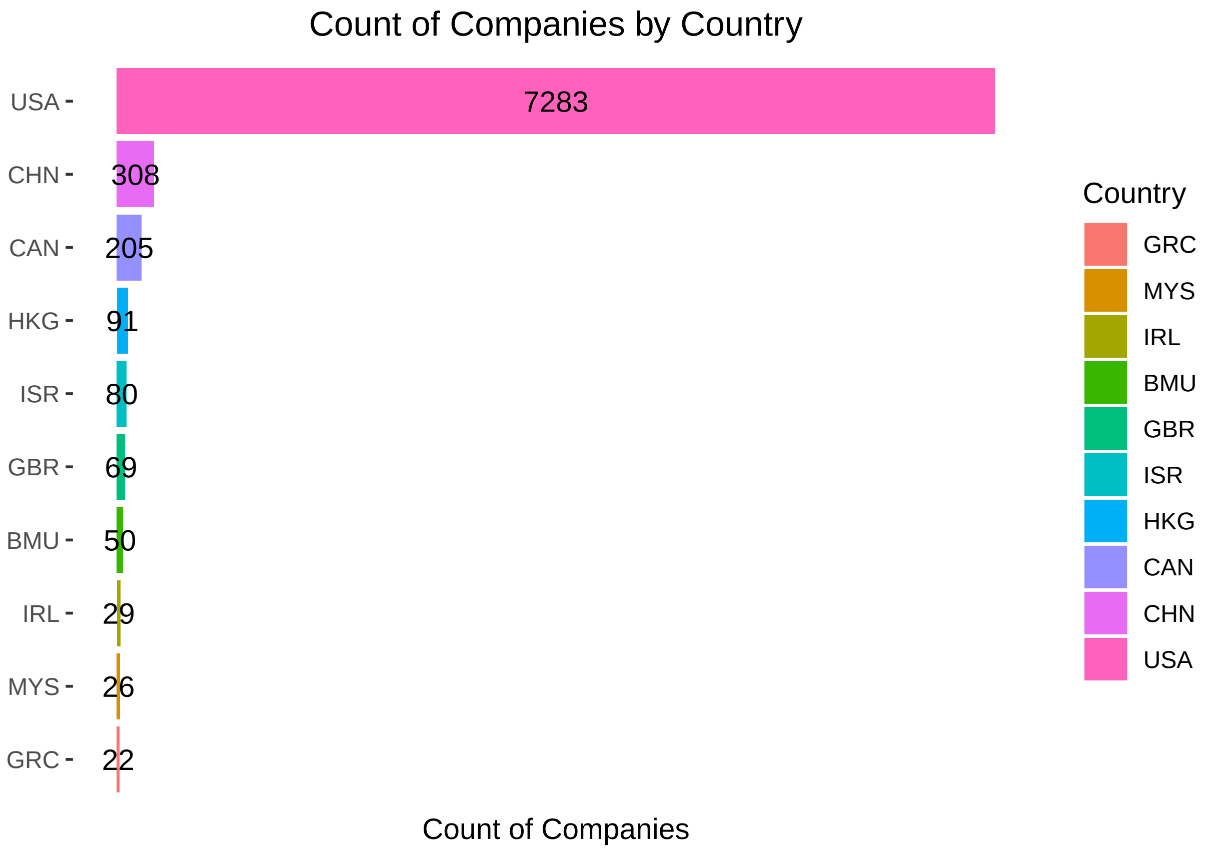
```
ggplot(dataset %>%
         count(COUNTRY),
       aes(x=COUNTRY,
           y=n,
           fill=COUNTRY))+
  geom_col()+
  geom_text(aes(label = n),position = position_stack(vjust = .5)) +
  ggtitle("Count of Companies by Country")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        panel.background = element_blank())+
  ylab("Count of Companies") +
  xlab(element_blank()) +
  labs(fill = "Country") +
  coord_flip()
```



What I decided for this distribution, and since we have way too many countries, was that I wanted to see the top countries

So I did the distribution, ordered the results by number of companies, and then took the top 10 companies

```
dataset_country <- dataset %>% filter(!is.na(COUNTRY))

dataset_country_2 <- dataset_country %>% count(COUNTRY)

dataset_country_f <- tail(dataset_country_2[order(dataset_country_2$n),],10)

dataset_country_f$COUNTRY <-
  factor(dataset_country_f$COUNTRY,
         levels = dataset_country_f$COUNTRY[order(dataset_country_f$n)])

ggplot(dataset_country_f,aes(x=COUNTRY,y=n,fill=COUNTRY))+
  geom_col()+
  geom_text(aes(label = n),position = position_stack(vjust = .5)) +
  ggtitle("Count of Companies by Country")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        panel.background = element_blank())+
  ylab("Count of Companies") +
  xlab(element_blank()) +
  labs(fill = "Country") +
  coord_flip()
```

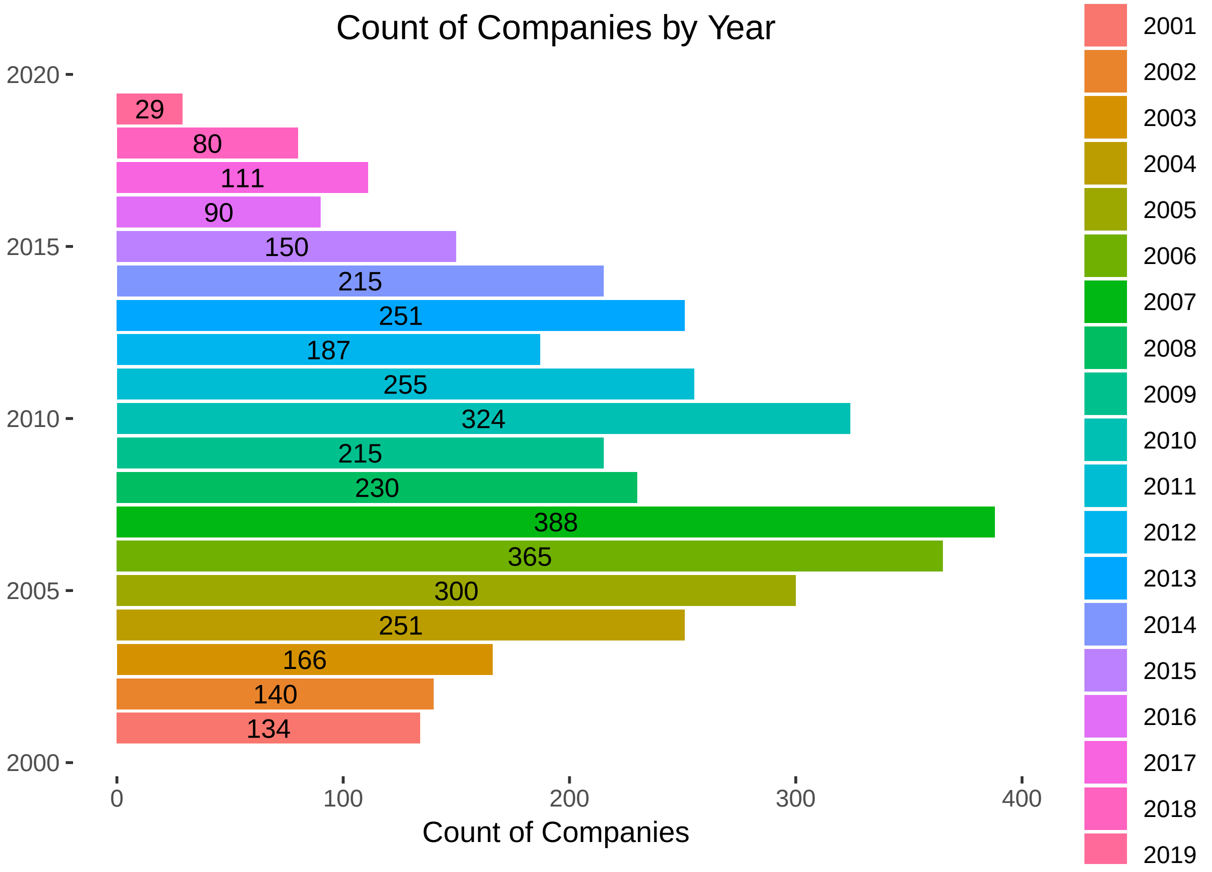


Unsurprinsingly, USA is the top country.

What is noteworthy, is that the rest of the coutries have less than 5% of the companies that USA has.

Now I want to see the annual sales during the year where most companies where incorporated.

```
ggplot(dataset %>%
         filter(YEAR_INCORP > 2000) %>%
         count(YEAR_INCORP),
       aes(x=YEAR_INCORP,
           y=n,
           fill=as.factor(YEAR_INCORP))) +
  geom_col()+
  geom_text(aes(label = n),position = position_stack(vjust = .5), size = 3.5) +
  ggtitle("Count of Companies by Year")+
  theme(plot.title = element_text(hjust = 0.5),
        #axis.text.x = element_blank(),
        #axis.ticks.x = element_blank(),
        panel.background = element_blank())+
  ylab("Count of Companies") +
  xlab(element_blank()) +
  labs(fill = "Year") +
  coord_flip()
```



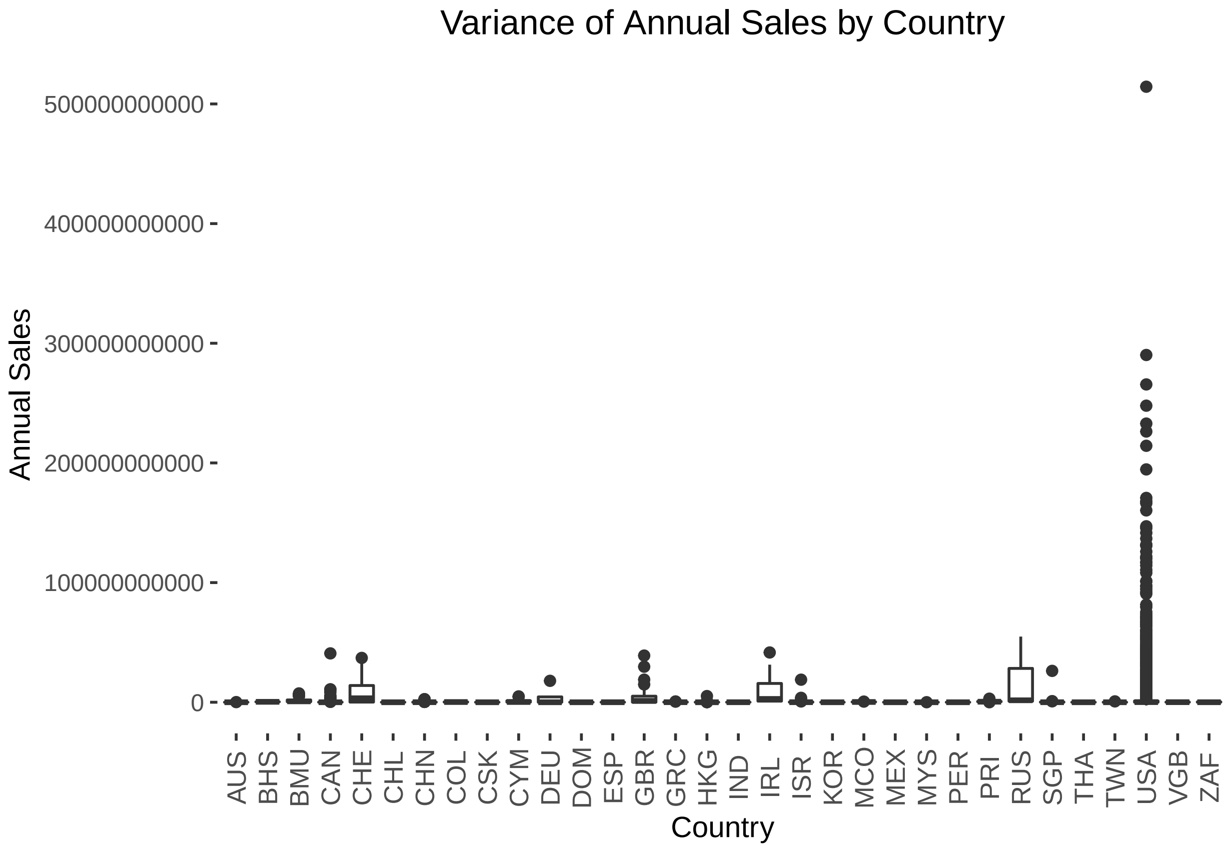
With this, I can see that 2007 is the year where the most companies where incorporated.

Ww knew to look into this time period because of the previous bucket analysis

However, before doing the analysis, we need to replace the NAs in the dataset

We can check the variance of the countries, in the same way we did before.

```
options(scipen = 999)

dataset_annsl <- dataset %>%
  select(ANNUAL_SALES,COUNTRY) %>%
  filter(is.na(ANNUAL_SALES)) %>%
  count(COUNTRY)

dataset_annsl_noNA<-dataset %>%
  select(ANNUAL_SALES,COUNTRY) %>%
  filter(!is.na(ANNUAL_SALES))

dataset_annsl_noNA <- merge(dataset_annsl_noNA %>%
                              select(COUNTRY,ANNUAL_SALES),
                            dataset_annsl %>%
                              select(COUNTRY),all=FALSE)

dataset_annsl_noNA %>% ggplot(aes(x=COUNTRY,
                                  y=ANNUAL_SALES))+
  geom_boxplot()+
  ggtitle("Variance of Annual Sales by Country")+
  theme(axis.text.x = element_text(size = 10, angle = 90, hjust = .5, vjust = .5),
        panel.background = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  ylab("Annual Sales") +
  xlab("Country")
```



And, as before, I'm using the median for each country

```
for (country in unique(dataset$COUNTRY)){
  dataset_fil <- dataset %>%
    filter(!is.na(ANNUAL_SALES)) %>%
    filter(COUNTRY == country)
  dataset$ANNUAL_SALES <- replace_na(dataset$ANNUAL_SALES,
                                     quantile(na.exclude(dataset_fil$ANNUAL_SALES),
                                              probs=0.5))
}

summary(dataset)
```

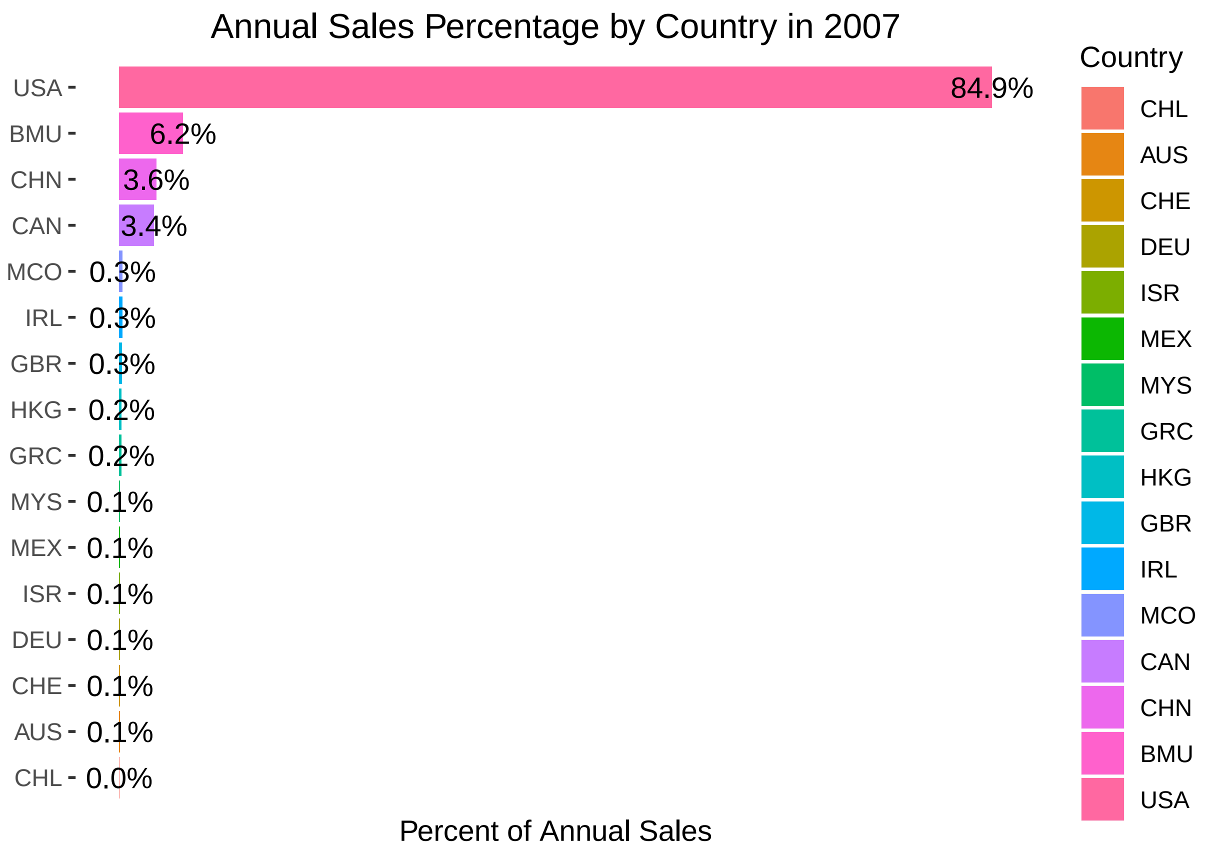
```
##                                COMPANY_NAME         CITY
##  024 Pharma Inc                      :   1   New York : 473
##  1-800 Flowers.com, Inc.             :   1   Houston  : 262
##  10x Genomics Inc                    :   1   Las Vegas: 188
##  11 Good Energy Inc                  :   1   Dallas   : 130
##  1347 Property Insurance Holdings Inc:   1   San Diego: 109
##  180 Degree Capital Corp             :   1   (Other)  :7117
##  (Other)                             :8376   NA's     : 103
##      STATE           ZIP          COUNTRY             PHONE
##  CA     :1280   10022  :  84   USA    :7283   800 983-0903:  11
##  NY     : 739   77002  :  67   CHN    : 308   855 588-7839:   8
##  TX     : 733   92121  :  47   CAN    : 205   510 522-9600:   7
##  FL     : 553   80202  :  43   HKG    :  91   512 236-6555:   6
##  NV     : 315   10019  :  36   ISR    :  80   800 736-3402:   6
##  (Other):4175   (Other):7968   (Other): 414   (Other)     :8310
##  NA's   : 587   NA's   : 137   NA's   :   1   NA's        :  34
##   YEAR_INCORP    ANNUAL_SALES          EMPLOYEE_COUNT
##  Min.   :1784   Min.   :  -278112421   Min.   :      0
##  1st Qu.:1986   1st Qu.:    10278500   1st Qu.:     28
##  Median :1999   Median :    99560721   Median :    207
##  Mean   :1991   Mean   :  2203257325   Mean   :   5385
##  3rd Qu.:2008   3rd Qu.:   579714000   3rd Qu.:   1177
##  Max.   :2019   Max.   :514405000000   Max.   :2200000
##  NA's   :157
##    NET_INCOME                YEAR_BUCKET         EMPLOYEE_COUNT_BUCKET
##  Min.   :-22443000000   1996 - 2019:4931   0 - 110000       :8315
##  1st Qu.:    -5216000   1972 - 1996:2107   110000 - 220000  :   37
##  Median :      -76102   1948 - 1972: 567   2090000 - 2202200:    1
##  Mean   :   175753539   1925 - 1948: 229   220000 - 330000  :   18
##  3rd Qu.:    27982000   1902 - 1925: 221   330000 - 440000  :    5
##  Max.   : 59531000000   (Other)    : 170   440000 - 550000  :    5
##  NA's   :25             NA's       : 157   550000 - 660000  :    1
```

With this summary we can check the values for the Annual Sales column

And now we can do the analysis of annual sales.

However, since the numbers are way too big, I'm using percentages in the graph

```
dataset_annsl_f <-
dataset %>%
  filter(YEAR_INCORP == "2007") %>%
  group_by(COUNTRY,YEAR_INCORP) %>%
  summarise(ANNUAL_SALES = sum(ANNUAL_SALES))

dataset_annsl_f$COUNTRY <-
  factor(dataset_annsl_f$COUNTRY,
         levels = dataset_annsl_f$COUNTRY[order(dataset_annsl_f$ANNUAL_SALES)])

ggplot(dataset_annsl_f,aes(x=COUNTRY,y=ANNUAL_SALES,fill=COUNTRY))+
  geom_col()+
  geom_text(aes(label = scales::percent(ANNUAL_SALES/sum(ANNUAL_SALES))),
            position = "dodge") +
  ggtitle("Annual Sales Percentage by Country in 2007")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        panel.background = element_blank())+
  ylab("Percent of Annual Sales") +
  xlab(element_blank()) +
  labs(fill = "Country") +
  coord_flip()
```


Annual Sales Percentage by Country in 2007
USA
BMU
CHN
CAN
MCO
IRL
GBR
HKG
GRC
MYS
MEX
ISR
DEU
CHE
AUS
CHL
84.9%
6.2%
3.6%
3.4%
0.3%
0.3%
0.3%
0.2%
0.2%
0.1%
0.1%
0.1%
0.1%
0.1%
0.1%
0.0%
Percent of Annual Sales
Country
CHL
AUS
CHE
DEU
ISR
MEX
MYS
GRC
HKG
GBR
IRL
MCO
CAN
CHN
BMU
USA