Experian\_Excercise

# Visualization demo

## Started by importing libraries

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(ggplot2)

## Registered S3 methods overwritten by 'ggplot2':  
## method from   
## [.quosures rlang  
## c.quosures rlang  
## print.quosures rlang

library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

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." "Abbott Laboratories" ...  
## $ 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 ...

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

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

ggplotly(  
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()  
)

## Warning: Removed 1 rows containing missing values (position\_stack).

## I then tried a histogram, sice we’re using years, maybe I can see which are the most valuable, however…

ggplotly(  
ggplot(dataset,aes(x = YEAR\_INCORP))+  
 geom\_histogram()  
)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 157 rows containing non-finite values (stat\_bin).

## 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)  
  
ggplotly(  
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())+  
 ylab("Count of companies") +  
 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

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

ggplotly(  
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())+  
 ylab("Count of passengers") +  
 labs(fill = "Year Bucket") +  
 coord\_flip()  
)

## 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)])  
  
ggplotly(  
dataset\_emp %>% ggplot(aes(x=COUNTRY,y=n,fill=as.factor(n)))+geom\_col()+  
 theme(axis.text.x = element\_text(size = 10, angle = 90, hjust = .5, vjust = .5),  
 panel.background = element\_blank(),  
 plot.title = element\_text(size = 25,face = "bold")) +   
 ylab("Count of NAs") +  
 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

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)  
  
ggplotly(  
dataset\_emp\_noNA %>% ggplot(aes(x=COUNTRY,y=EMPLOYEE\_COUNT))+geom\_boxplot()+  
 theme(axis.text.x = element\_text(size = 10, angle = 90, hjust = .5, vjust = .5),  
 panel.background = element\_blank(),  
 plot.title = element\_text(size = 25,face = "bold")) +  
 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 that the median doesn’t look weird for the countries, it actually sits very well with the data

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

## 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

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

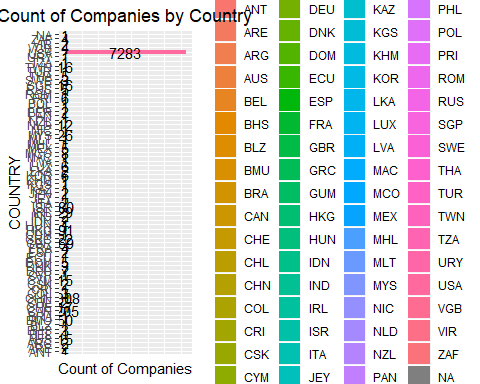
ggplotly(  
ggplot(dataset %>%  
 count(EMPLOYEE\_COUNT\_BUCKET),aes(x = EMPLOYEE\_COUNT\_BUCKET,y=n,fill=EMPLOYEE\_COUNT\_BUCKET))+  
 geom\_col()+  
 geom\_text(aes(label = n),position = position\_stack(vjust = .5)) +  
 ggtitle("Count of Companies by Employee Number Bucket")+  
 theme(plot.title = element\_text(hjust = 0.5),  
 axis.text.x = element\_blank(),  
 axis.ticks.x = element\_blank())+  
 ylab("Count of Companies") +  
 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())+  
 ylab("Count of Companies") +  
 coord\_flip()

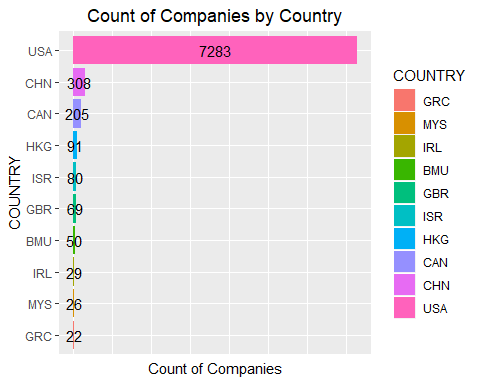
## Warning: Factor `COUNTRY` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`



## 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())+  
 ylab("Count of Companies") +  
 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