EDA - Corporate Governance and Company Performance

Conor Reid

23/5/2017

Set the working directory (not shown) and load in the required packages.

```
library(reshape)
library(stringr)
library(rpart)
library(lattice)
library(ggplot2)
```

S&P 500

The S&P 500 is an American stock market index including market capitalizations of 500 companies, listed on the NYSE or the NASDAQ. It is distinct from other indexs such as the Dow Jones etc due to its diverse constituency and weighting methodology. *Would it be useful to look at other index's?*

Read in full data as below, and carry out some subsetting. Set NA values.

```
spx=read.csv("spx.csv", sep=",", na.strings="#N/A N/A")
spx_EDA = spx[c("Ticker", "Tobin.s.Q", "P.E", "EPS", "P.B", "P.EBITDA",
                "Board.Size", "CEO.Duality", "X..Feml.Execs",
                "X..Feml.Execs.1", "Bd.Avg.Age", "Board.Mtg.Att..")]
spx_EDA[ spx_EDA == "#N/A Field Not Applicable" ] <- NA
```

View data types. Convert to numeric columns that need to be numeric. Parse out the actual ticker string (output hidden for clarity).

```
sapply(spx_EDA,class)
cols.num <- c("P.B","EPS","P.E","P.EBITDA")
spx_EDA[cols.num] <- sapply(spx_EDA[cols.num],as.numeric)
sapply(spx_EDA, class)
spx_EDA$TickerID = str_split_fixed(spx_EDA$Ticker, " ", 2)[,1] #parse the ticker itself
```

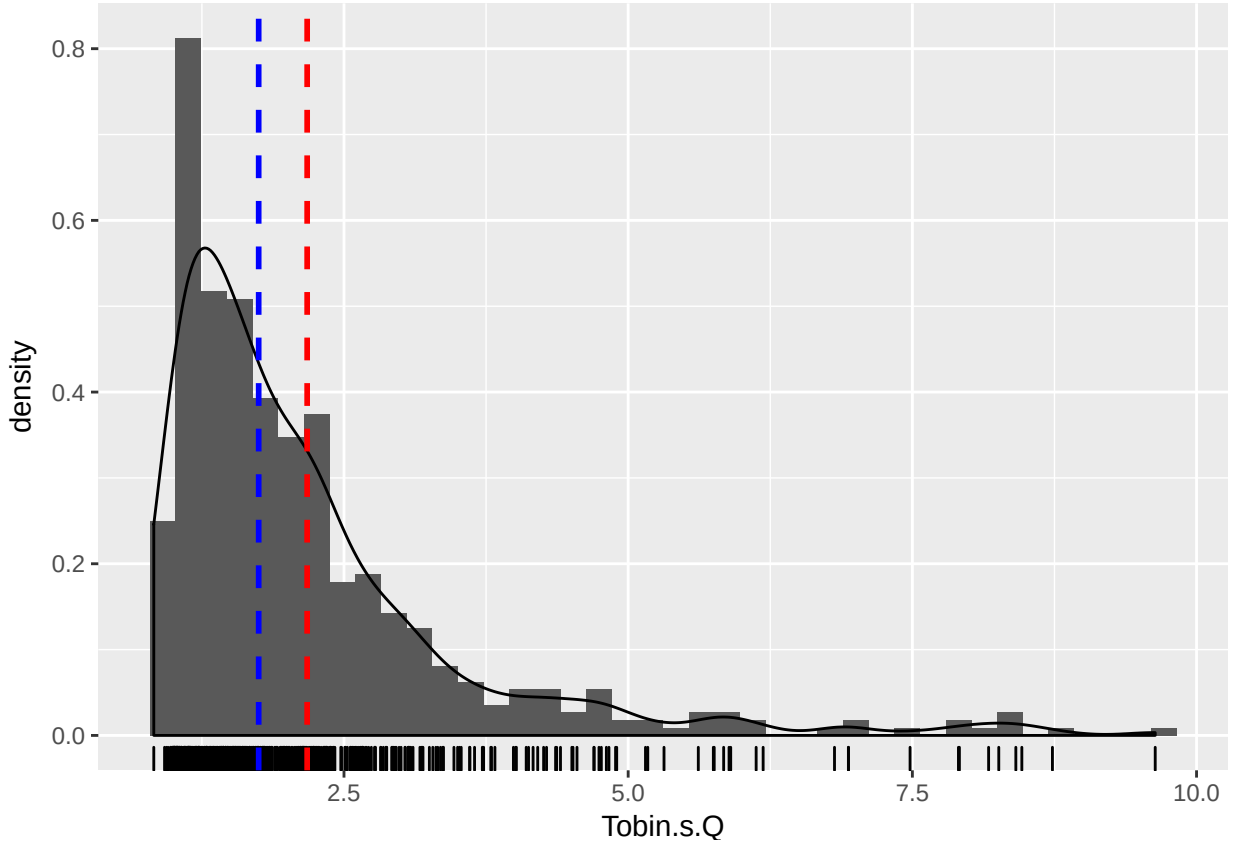
Tobins Q

Histogram of Tobins Q over the entire dataset. Heavily skewed right, most companies have low scores. It is going to be difficult to learn rules for high performing companies? From M&M:

> As suggested by Creamer [14], we discretized Tobin's Q ratio in order to obtain two classes, dividing each dataset according to its median value. In this way, a company that lies in the upper side of the median will be looked positively by the machine learning algorithms.

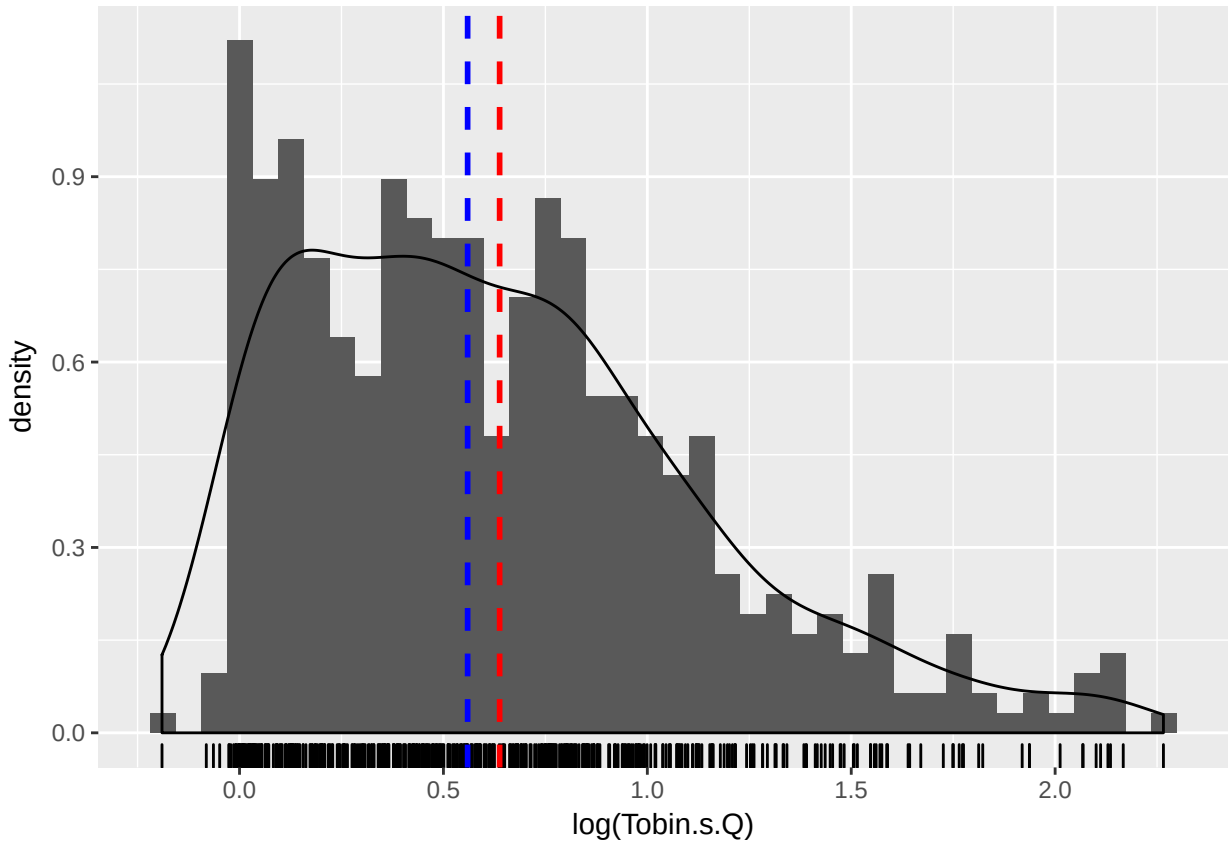
Median (blue) and mean (red) of Q score is shown below. Regression may not be suitable for data that has such a skewness.

```
#distrib
ggplot(data=spx_EDA) +
  geom_histogram( aes(Tobin.s.Q, ..density..), bins=40 ) +
  geom_density( aes(Tobin.s.Q, ..density..) ) +
  geom_rug( aes(Tobin.s.Q) ) +
  geom_vline(aes(xintercept=mean(Tobin.s.Q, na.rm=T)),  # red line is the mean
             color="red", linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(Tobin.s.Q, na.rm=T)),  # blue line is the median
             color="blue", linetype="dashed", size=1)
```



Try taking the log of Tobins Q instead, histogram shown below. The code is emitted since its the same as previous page but used log(Tobin.s.Q) instead. Much less skewed, making regression more applicable?

```
ggplot(data=spx_EDA) +
  geom_histogram( aes(log(Tobin.s.Q), ..density..), bins=40 ) +
  geom_density( aes(log(Tobin.s.Q), ..density..) ) +
  geom_rug( aes(log(Tobin.s.Q)) ) +
  geom_vline(aes(xintercept=mean(log(Tobin.s.Q), na.rm=T)),  # red line is the mean
             color="red", linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(log(Tobin.s.Q), na.rm=T)),  # blue line is the median
             color="blue", linetype="dashed", size=1)
```



Poisson regression could be a good alternative to having to process the Q score directly, obviously assumes the dependant variable has a Poisson distribution.

See https://en.wikipedia.org/wiki/Poisson_distribution

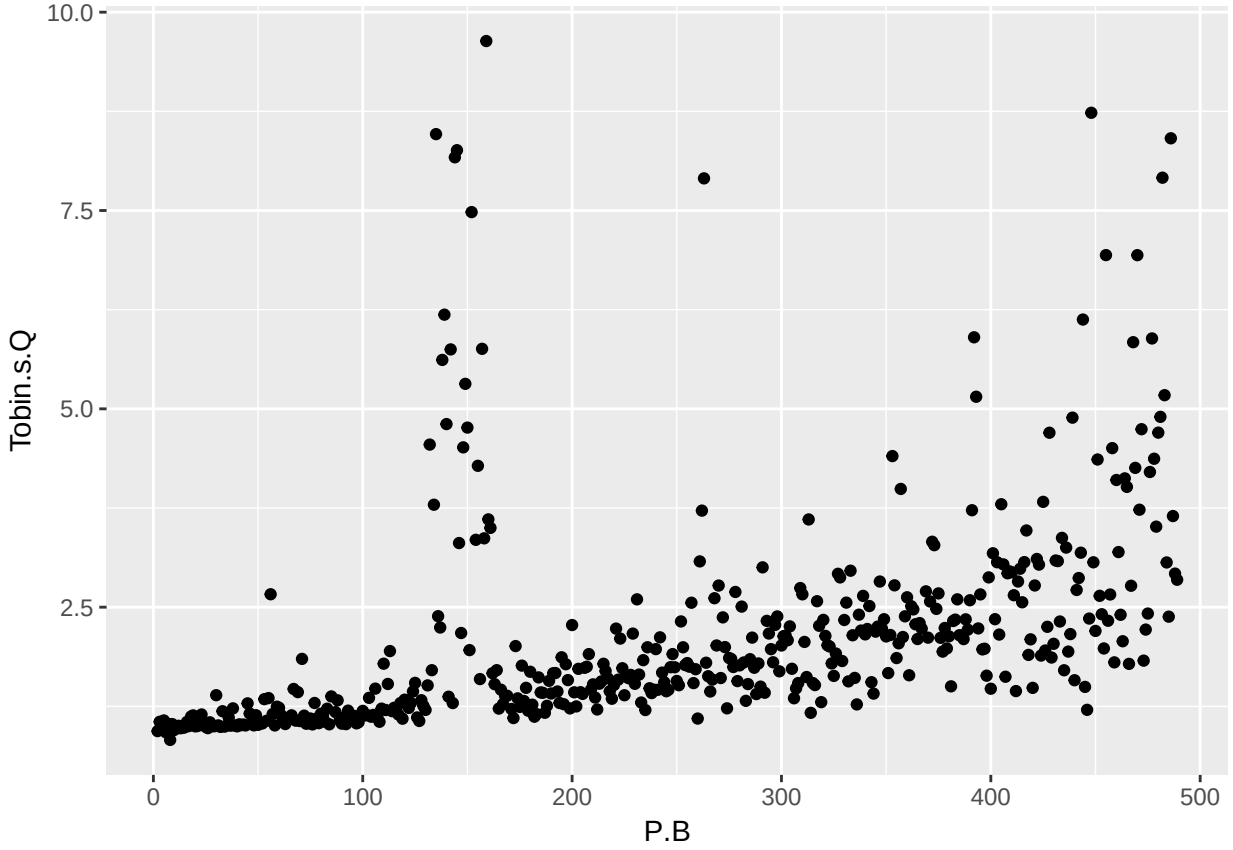
I dont think this it actually does, but Poisson regression may be more applicable here that OLS for example.

P.B (price to book ratio)

A high P.B means that investors see the company as growing, low means they think the companys book value is exaggerated (http://www.investopedia.com/terms/p/price-to-bookratio.asp).

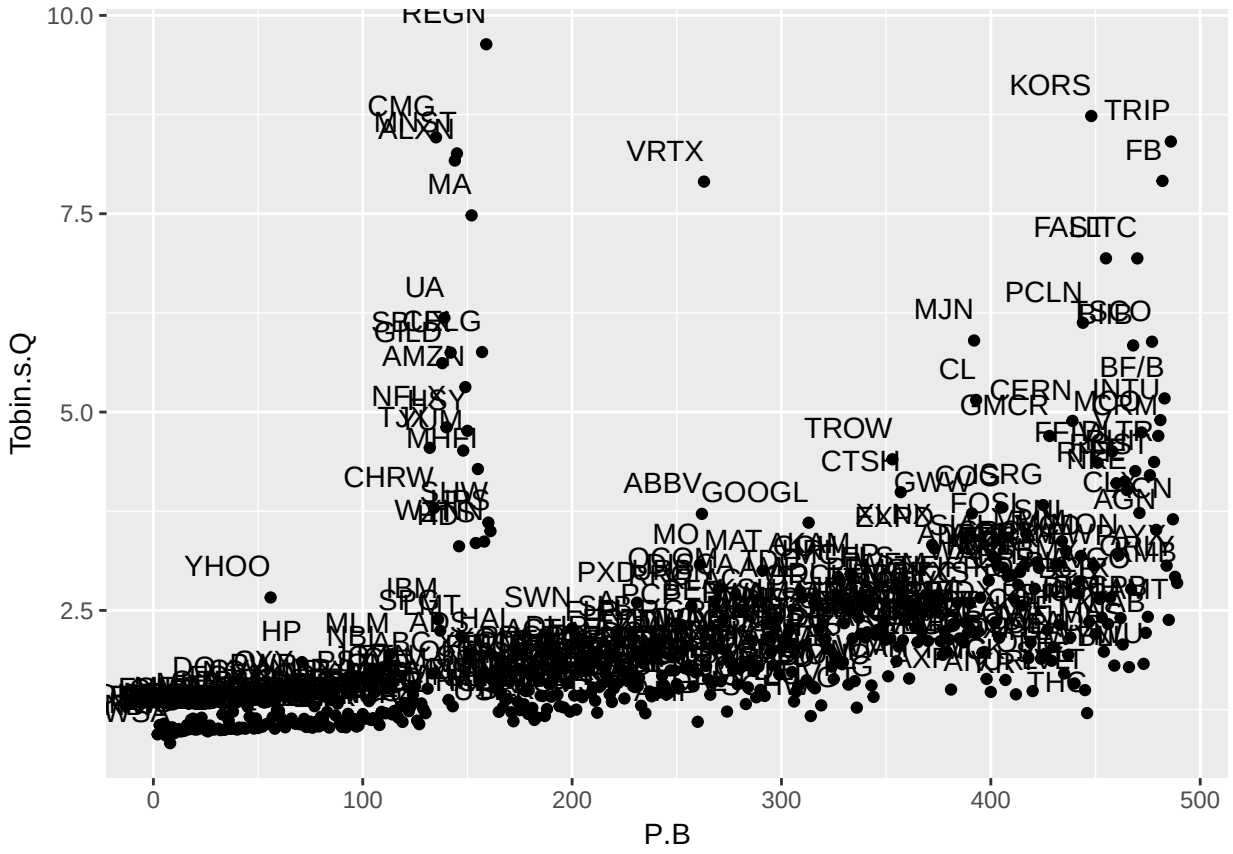
Interesting spike in Q score around the 150 P.B mark. Maybe worth further investigation?

```
##P.B
ggplot(spx_EDA, aes(P.B, Tobin.s.Q) ) +
  geom_point()
```

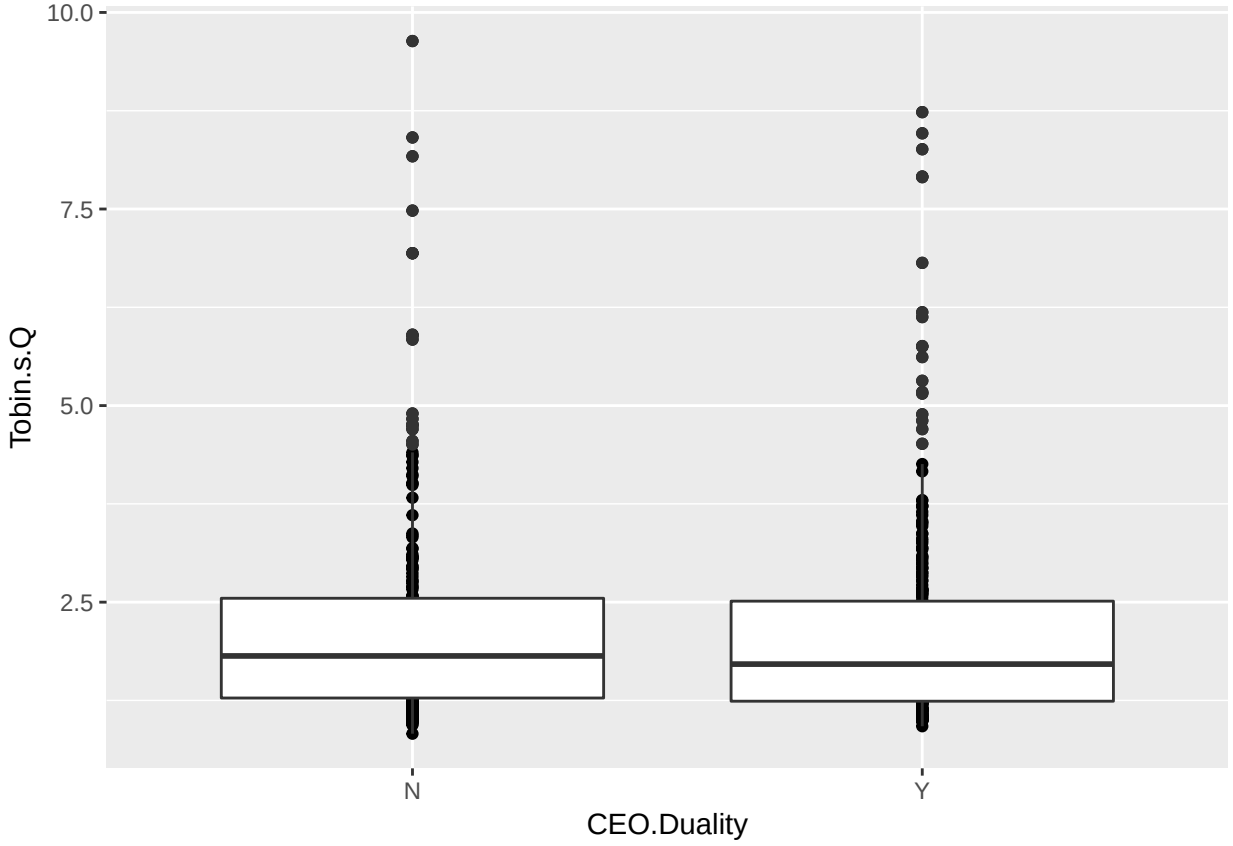


Same as previous page, but showing the tickers for each data point. VRTX is a particular outler (Vertex Pharmaceuticals Incorporated).

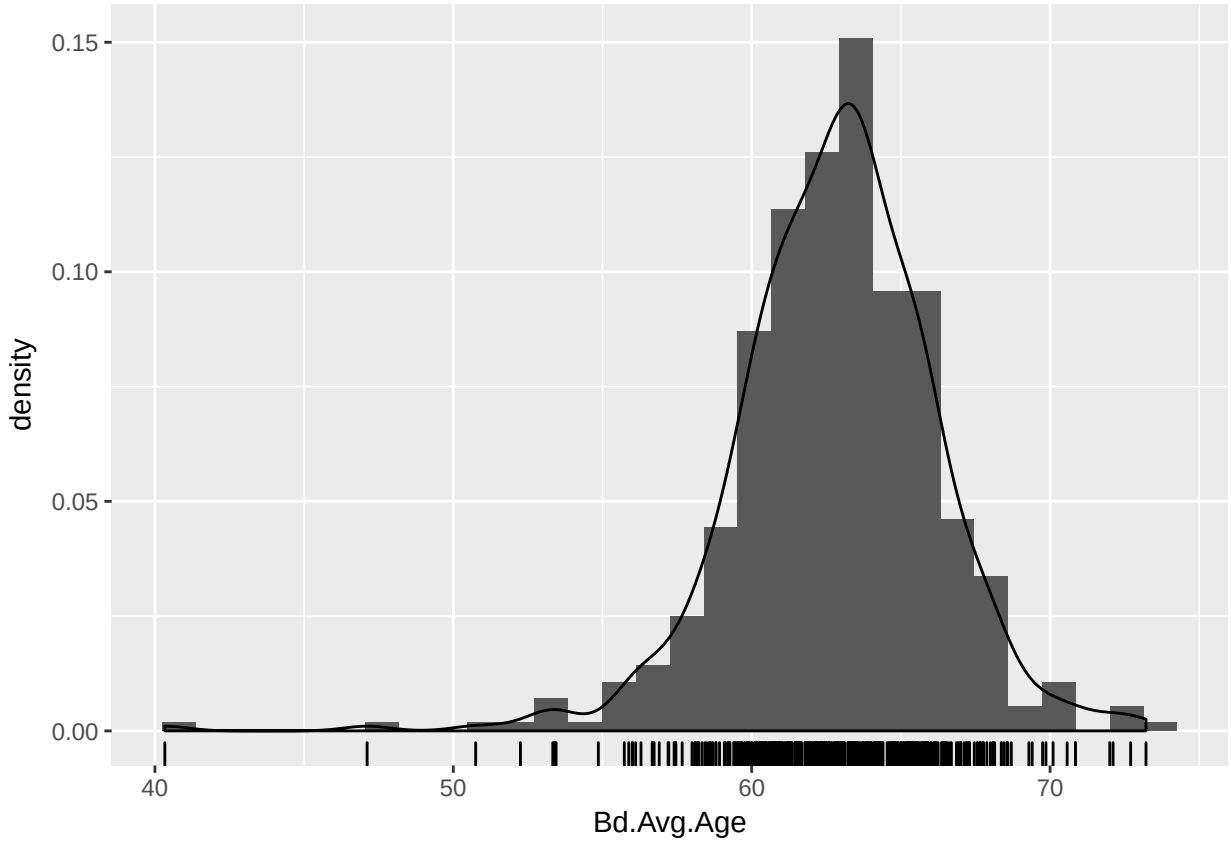
```
#with labels
 ggplot(spx_EDA, aes(P.B, Tobin.s.Q) ) +
    geom_point() +
    geom_text(aes(label=TickerID),hjust=1, vjust=-1)
```



```r
#no difference
ggplot(data=subset(spx_EDA, !is.na(CEO.Duality)),
  aes(CEO.Duality, Tobin.s.Q )) +
  geom_point() + geom_boxplot()
```

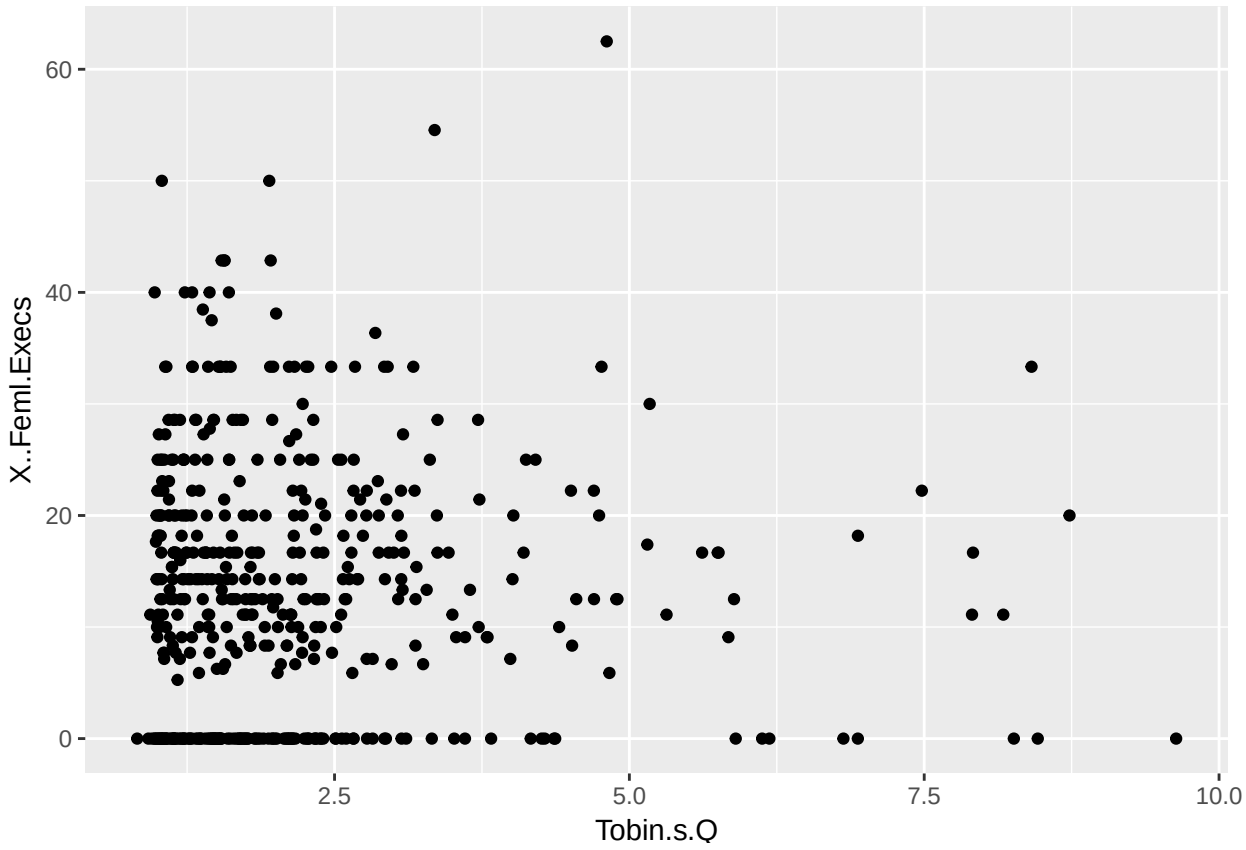


```
#board average age
ggplot(data=spx_EDA) +
  geom_histogram( aes(Bd.Avg.Age, ..density..) ) +
  geom_density( aes(Bd.Avg.Age, ..density..) ) +
  geom_rug( aes(Bd.Avg.Age) )
```



Tobins Q and the number of female executives. Seems to cluster around the bottom left hand courner at low numbers and low performance, but cant see much to the upper right. Doesnt look like a strong relationship here anyway.

```
#female presence on board and success?
ggplot(spx_EDA, aes(Tobin.s.Q, X..Feml.Execs) ) +
  geom_point()
```



STOXX® Europe 600

STOXX Eastern Europe 300