EDA - Corporate Governance and Company Performance

Conor Reid

28/5/2017

Set the working directory and read in data as "spx" (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 capitalization of 500 companies, listed on the NYSE or the NASDAQ. It is distinct from other indexes 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 sub-setting. Set NA values.

```
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..",
                "Asset", "Interest", "Tax", "ROE", "ROC",
                "Indep.Directors")]
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
```

Missing Data

Cases with missing values. Seems high (73 / 500 = 14.6% loss if we discount them).

```
sum(!complete.cases(spx_EDA))
```

```
## [1] 250
```

```
summary(spx_EDA)
```

```
##                 Ticker      Tobin.s.Q           P.E
##  A        UN Equity:  1   Min.   :0.8268   Min.   :  2.0
##  AA       UN Equity:  1   1st Qu.:1.2672   1st Qu.:124.2
##  AAPL     UW Equity:  1   Median :1.7495   Median :246.5
##  ABBV     UN Equity:  1   Mean   :2.1761   Mean   :246.5
##  ABC      UN Equity:  1   3rd Qu.:2.5184   3rd Qu.:368.8
##  ABT      UN Equity:  1   Max.   :9.6363   Max.   :491.0
##  (Other)           :494   NA's   :4        NA's   :10
##       EPS              P.B           P.EBITDA       Board.Size
##  Min.   :  1.00   Min.   :  2.0   Min.   :  2.0   Min.   : 5.00
##  1st Qu.: 98.75   1st Qu.:123.8   1st Qu.:115.5   1st Qu.:10.00
##  Median :181.50   Median :245.5   Median :229.0   Median :11.00
##  Mean   :183.82   Mean   :245.5   Mean   :229.0   Mean   :10.98
##  3rd Qu.:266.25   3rd Qu.:367.2   3rd Qu.:342.5   3rd Qu.:12.00
##  Max.   :374.00   Max.   :489.0   Max.   :456.0   Max.   :29.00
##  NA's   :4        NA's   :12      NA's   :45      NA's   :3
##  CEO.Duality X..Feml.Execs   X..Feml.Execs.1   Bd.Avg.Age
##  N   :225    Min.   : 0.00   Min.   :0.00    Min.   :40.33
##  Y   :272    1st Qu.: 0.00   1st Qu.:0.00    1st Qu.:60.85
##  NA's:  3    Median :12.50   Median :1.00    Median :63.00
##              Mean   :13.53   Mean   :1.33    Mean   :62.80
##              3rd Qu.:20.00   3rd Qu.:2.00    3rd Qu.:64.91
##              Max.   :62.50   Max.   :8.00    Max.   :73.21
##              NA's   :3       NA's   :3       NA's   :3
##  Board.Mtg.Att..      Asset            Interest            Tax
##  Min.   : 72.15   Min.   :0.03398   Min.   : -52.96   Min.   :-498.04
##  1st Qu.: 75.00   1st Qu.:0.34901   1st Qu.:  84.63   1st Qu.:  63.23
##  Median : 75.00   Median :0.61024   Median :  92.21   Median :  69.02
##  Mean   : 79.88   Mean   :0.79602   Mean   :  99.11   Mean   :  86.91
##  3rd Qu.: 80.00   3rd Qu.:0.94005   3rd Qu.:  98.89   3rd Qu.:  77.82
##  Max.   :100.00   Max.   :5.91183   Max.   :3706.66   Max.   :3127.97
##  NA's   :14       NA's   :3         NA's   :73        NA's   :2
##       ROE                ROC          Indep.Directors    TickerID
##  Min.   : -42.764   Min.   :-17.959   Min.   : 4.000   Length:500
##  1st Qu.:   9.661   1st Qu.:  6.425   1st Qu.: 8.000   Class :character
##  Median :  15.716   Median : 10.343   Median : 9.000   Mode  :character
##  Mean   :  25.190   Mean   : 12.103   Mean   : 9.302
##  3rd Qu.:  24.001   3rd Qu.: 14.813   3rd Qu.:11.000
##  Max.   :1741.641   Max.   : 86.278   Max.   :23.000
##  NA's   :12         NA's   :167       NA's   :4
```

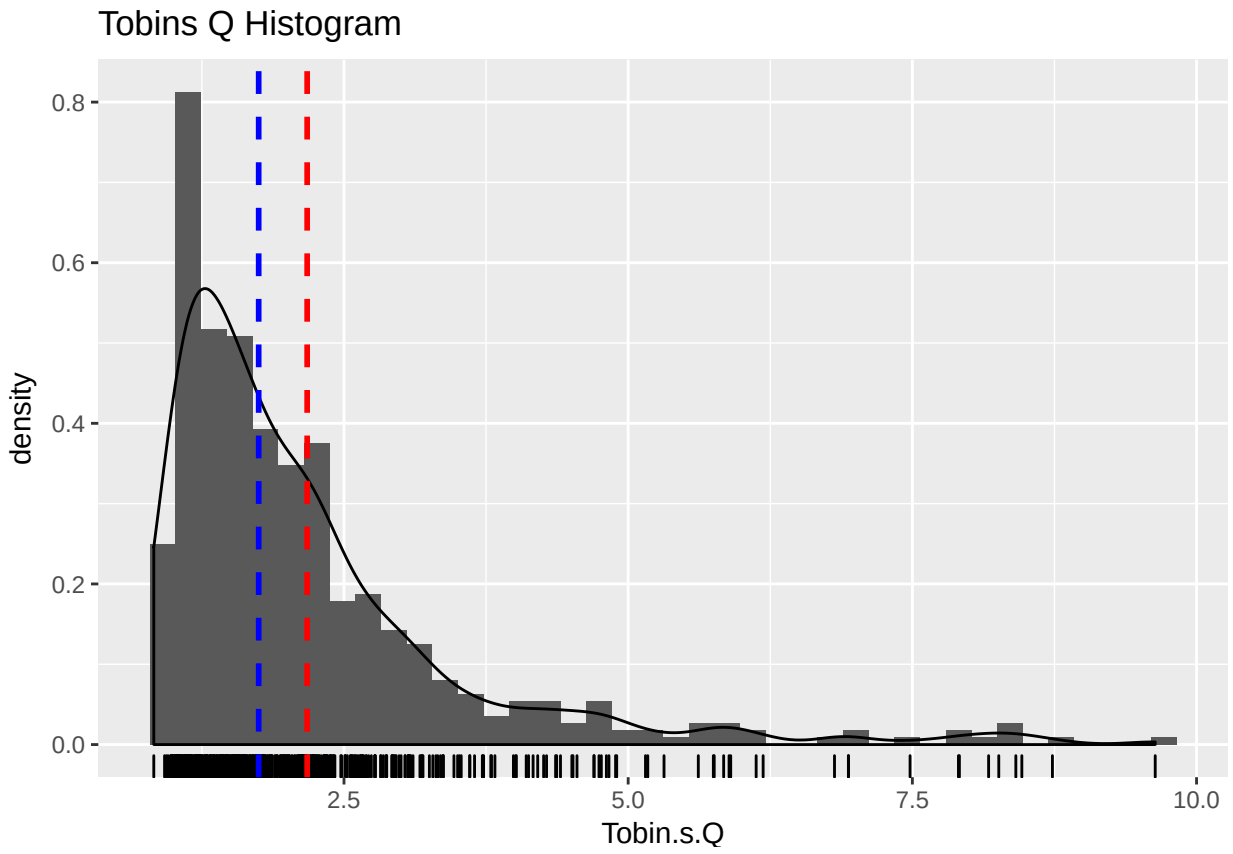
Tobins Q

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.

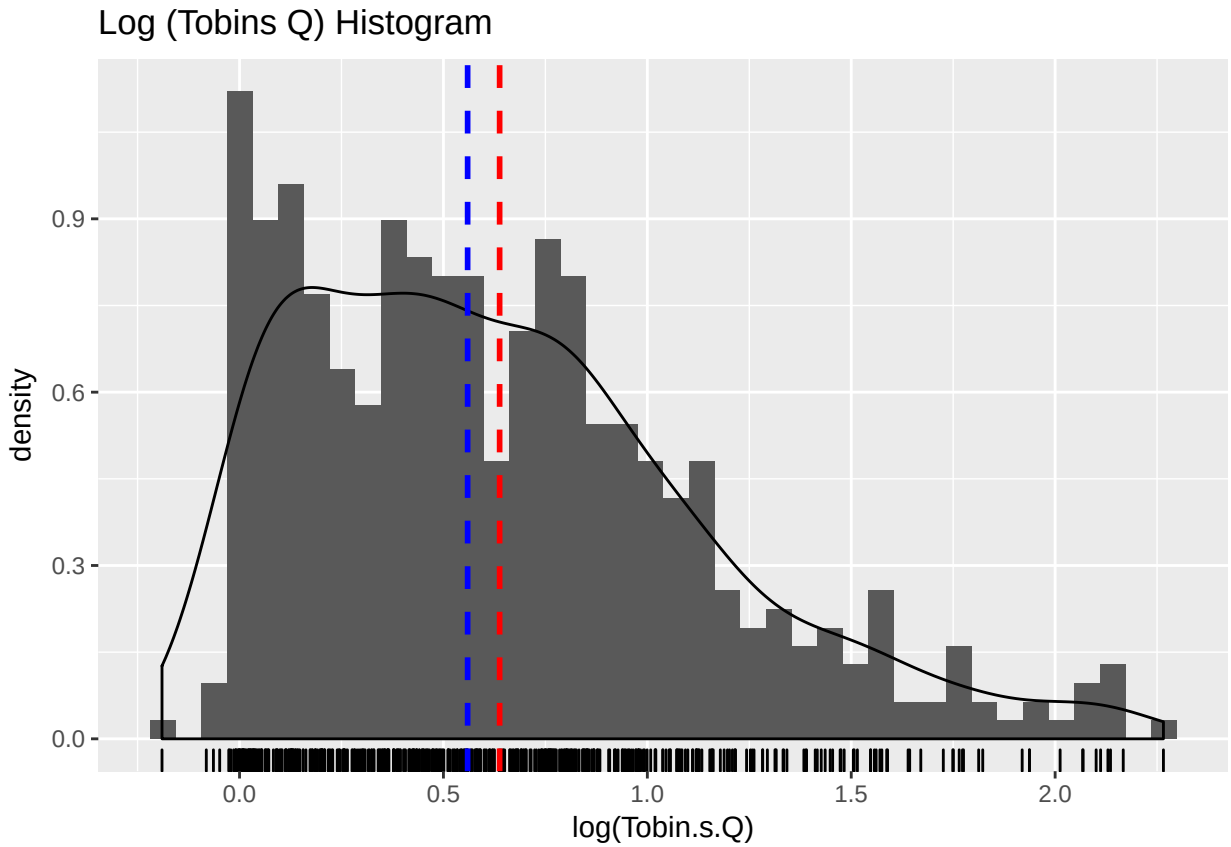
Median (blue) and mean (red) of Q score is shown below. Regression may not be suitable for data that has such a skewness.

```
#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) +
  ggtitle("Tobins Q Histogram")
```



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? Is it OK to process the dependent variable like this?

```
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) +
  ggtitle("Log (Tobins Q) Histogram")
```



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

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

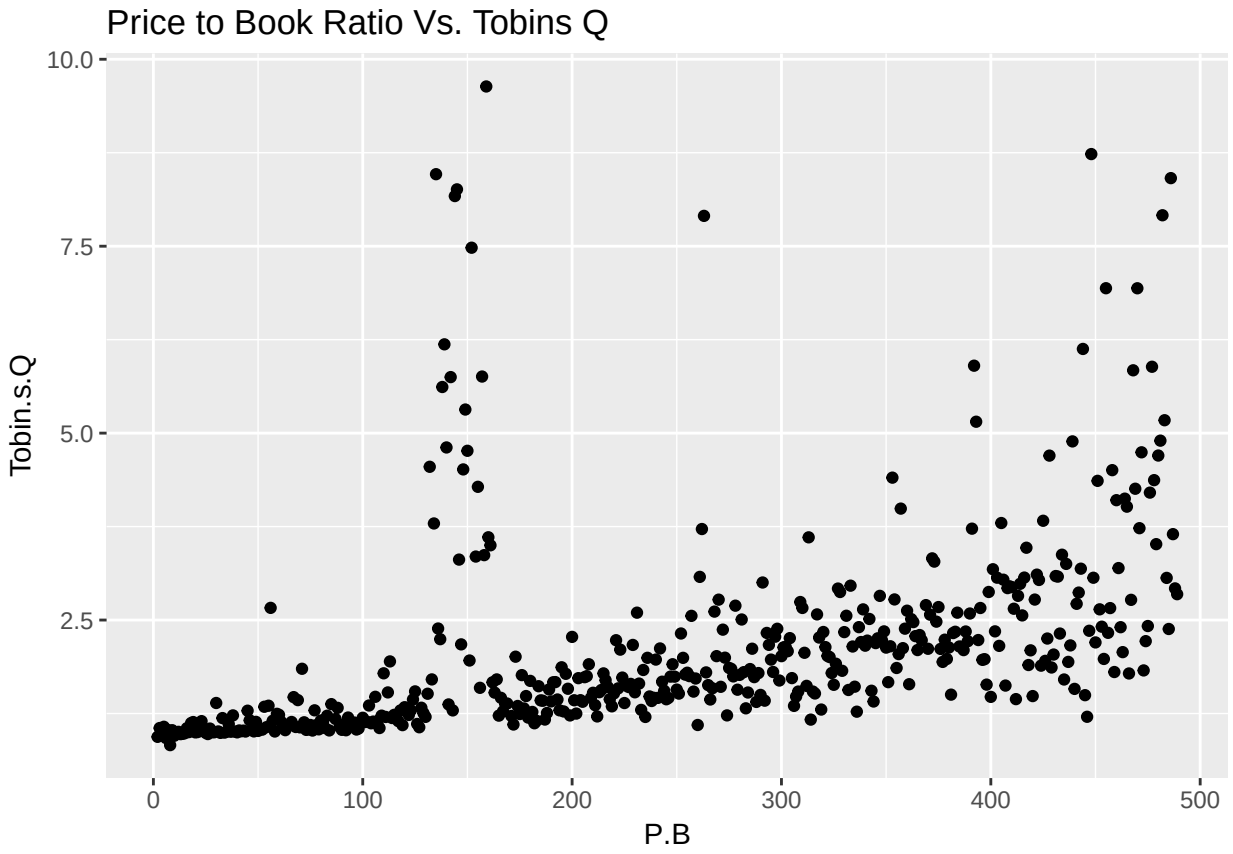
I don't think this it actually does, but Poisson regression may be more applicable here that OLS for example.

P.B (price to book ratio) and Q Score

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

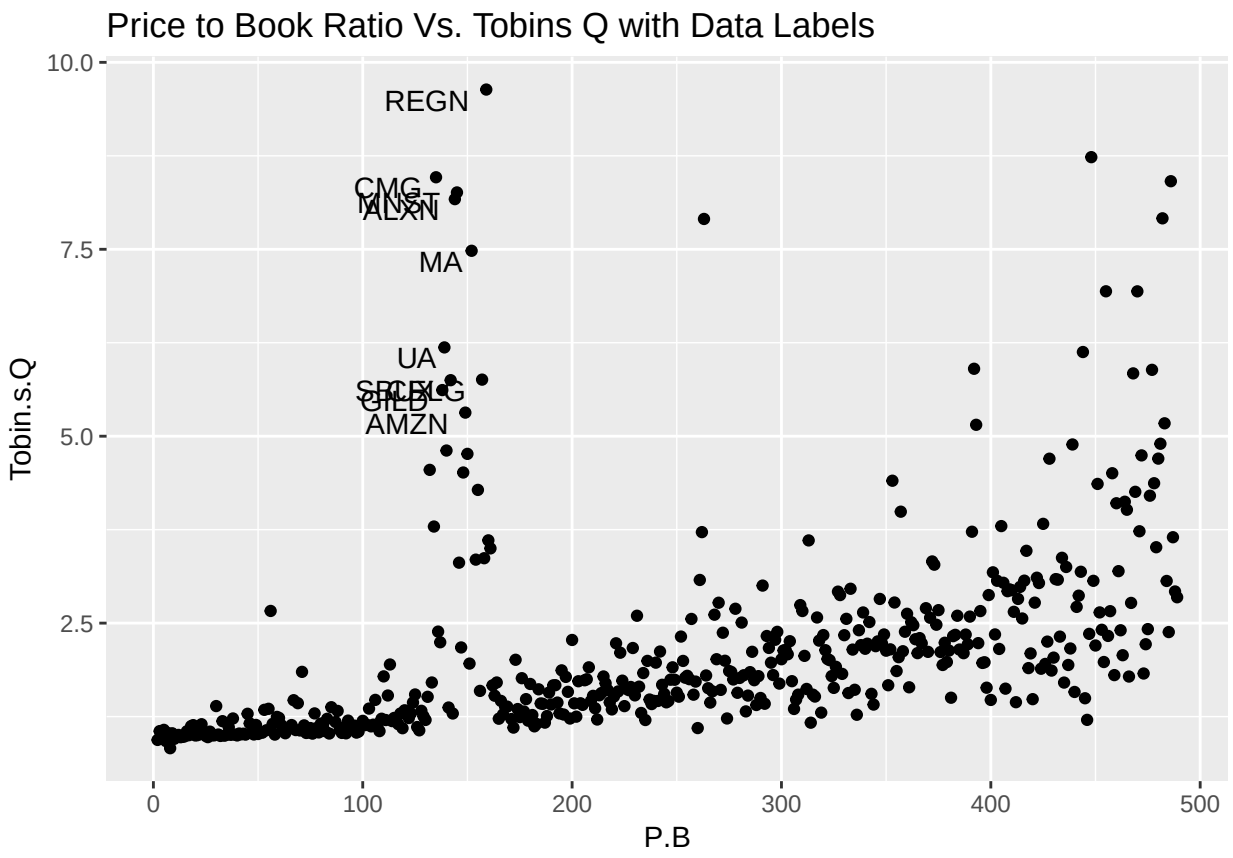
Interesting spike in Q score around the 150 P.B mark. Maybe worth further investigation?

```
##P.B
ggplot(spx_EDA, aes(P.B, Tobin.s.Q) ) +
  geom_point() +
  ggtitle("Price to Book Ratio Vs. Tobins Q")
```



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

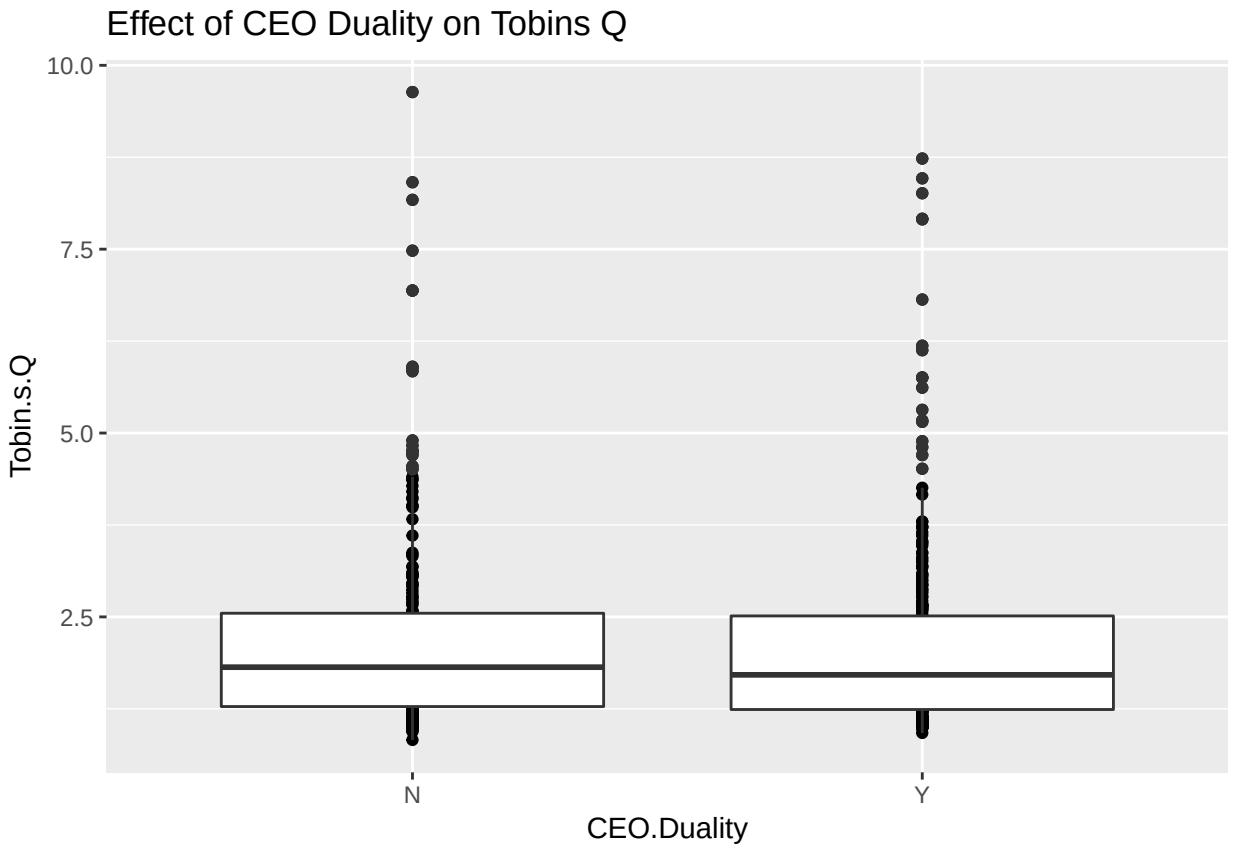
```
#with labels
 ggplot(spx_EDA, aes(P.B, Tobin.s.Q) ) +
    geom_point() +
    geom_text(aes(label=ifelse(P.B>100 & P.B<200 & Tobin.s.Q>5,as.character(TickerID),'')),hjust=1.2, v
  ggtitle("Price to Book Ratio Vs. Tobins Q with Data Labels")
```



CEO Duality and Q Score

Research suggests the CEO Duality can be either a positive or negative influence on the success of a company, looks from the below that perhaps its not that big of an influence at all. No deviation in Q score based on duality.

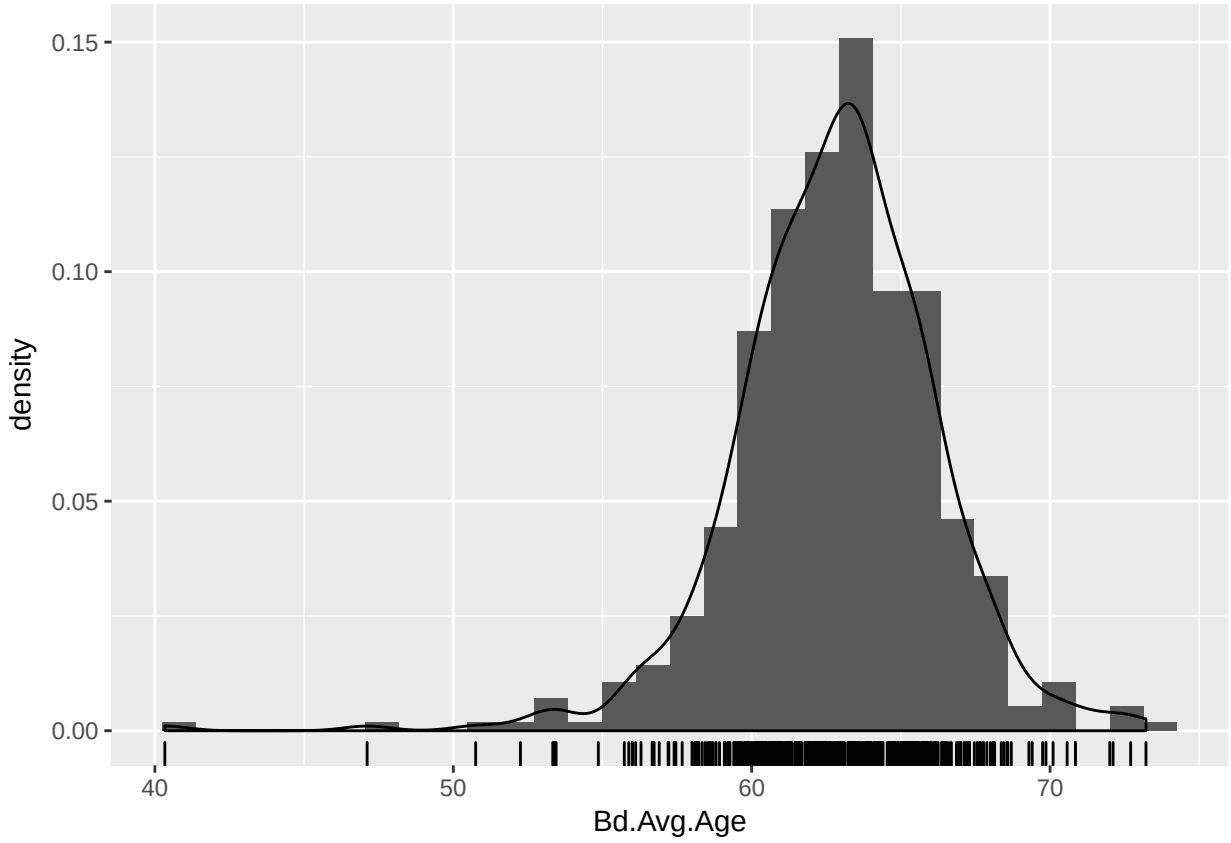
```
#no difference
ggplot(data=subset(spx_EDA, !is.na(CEO.Duality)),
  aes(CEO.Duality, Tobin.s.Q )) +
  geom_point() + geom_boxplot()+
  ggtitle("Effect of CEO Duality on Tobins Q")
```



Board Average Age and Q Score

Seems to be pretty normally distributed, skewed left due to some very young boards.

```
#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) )
```



Number of Female Executives and Q Score

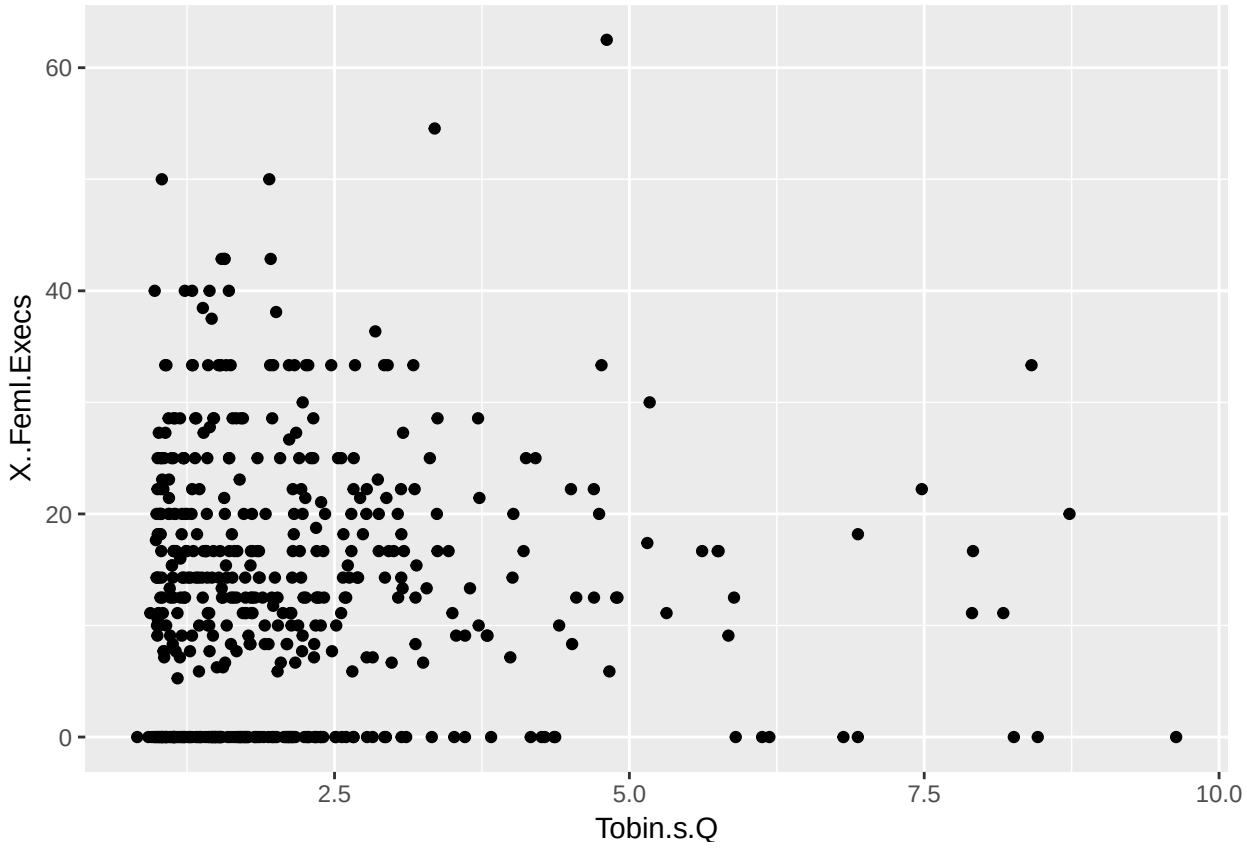
Tobins Q and the number of female executives. Seems to cluster around the bottom left hand corner at low numbers and low performance, but cant see much to the upper right. Doesn't look like a strong relationship here anyway.

According to M&M:

> Tobin's Q ratio is positively influenced by the percentage women in the board

May need to look into this further.

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

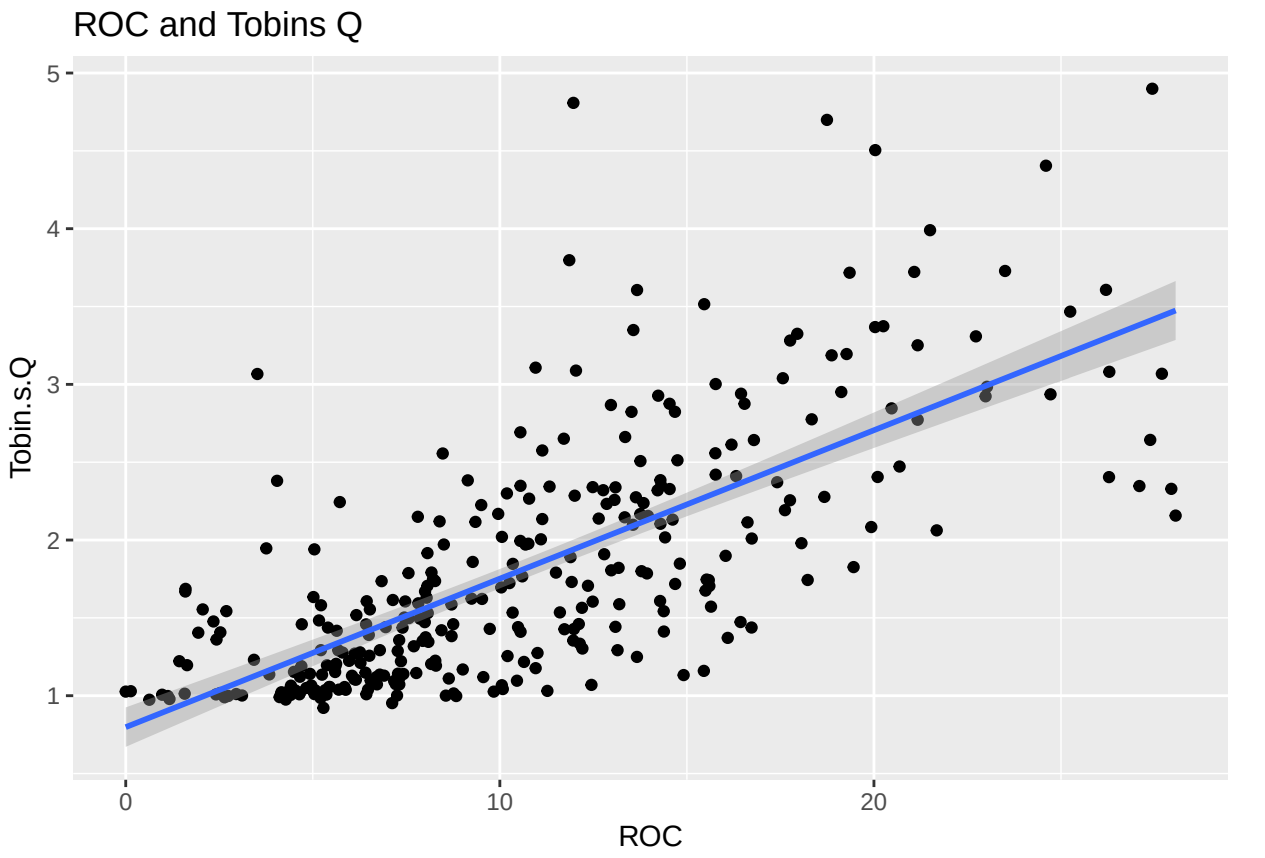


ROC (Return on Capital) and Q Score

Below shows Q score against ROC (return on capital), subsetted to remove extreme values in both variables. Thus, below is a representation of the main cluster in the data.

Looks to be a positive relationship between the two. Interesting, since ROC (may be) considered a bad strategy in the long term for companies (requires further investigation).

```
ggplot(data=subset(spx_EDA, !is.na(ROC) & ROC<30 & ROC>0 & Tobin.s.Q < 5),
  aes(ROC, Tobin.s.Q )) +
  geom_point() +
  ggtitle("ROC and Tobins Q") +
  geom_smooth(method = "lm")
```



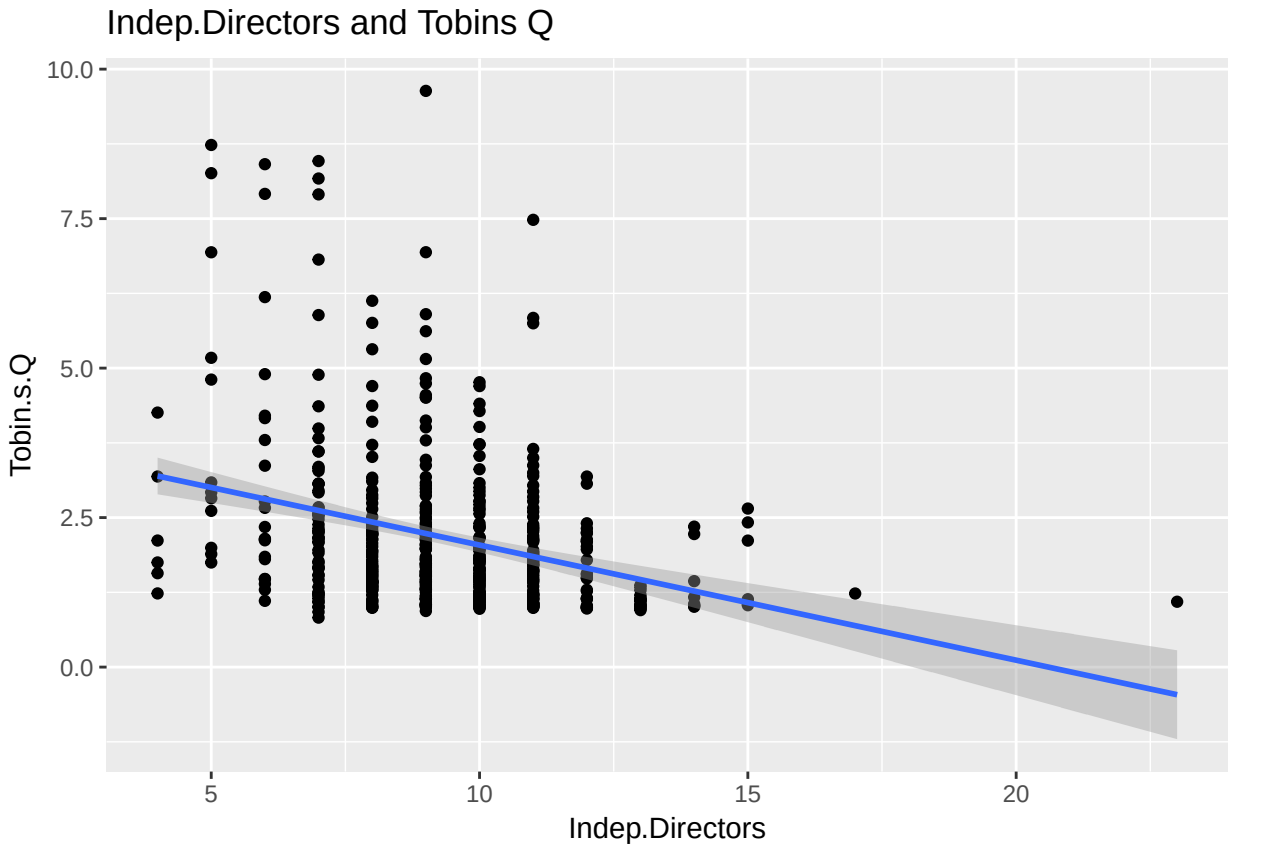
Independant Directors and Q Score

According to M&M;

> . . . while an independent lead director and a financial leverage higher than 2.5 generates a higher risk of bankruptcy.

Plot below seems to support this, is it strong enough though? R squared value likely to be pretty small when a proper regression is done.

```
ggplot(data=subset(spx_EDA, !is.na(Indep.Directors)),
  aes(Indep.Directors, Tobin.s.Q )) +
  geom_point() +
  ggtitle("Indep.Directors and Tobins Q") +
  geom_smooth(method = "lm")
```



STOXX® Europe 600

STOXX Eastern Europe 300