RICE UNIVERSITY

9/25/2020

Financial Data Distribution

Group 3: Brandon Gan, Giovanni Catania, Jason Guan, Ningning Zhu

Introduction

We explored the major factors for evaluating different companies' financial status. Some of the factors include: revenues, EPS, P/E ratio, and dividend yield. We explored the market's performance by analyzing some of the important financial data for publicly traded companies. We collected the data from 1970 to 2019 and drew distributions for each of the variables that were chosen. Some of the variables included market capitalization, book value, along with other important financial indicators such as, the price to earnings ratio and the price to book ratio. We also analyzed net income, because it is influenced by extraordinary items and probability of extraordinary items that were estimated. We then looked at the end of the relationship between total book value and common equity liquidation.

Methodology and Data Gathering

The data for the financial metrics was gathered from the Center for Research in Security Pricing (CRSP) and Compustat Merged database from Wharton Research Data Services (WRDS). This data consisted of annual variables from 1970-2019. After cleaning the dataset, it was used to calculate other variables for our analysis, such as market capitalization.

Coverage Analysis and Variable Creation

Table 1. Variable Coverage

Variable Code	Variable Description	Coverage
GVKEY	Global Company Key	1
datadate	Data Year -Fiscal	1
fyear	Fiscal Year	0.999993
indfmt	Industry Formats	1
consol	Consolidated Information	1
popsrc	Population Source of the data as Domestic (D) or International (I)	1
datafmt	Data Format	1
tic	Ticker Symbol	0.999984
conm	Company Name	1
curcd	Currency of Data	1
scf	Postal code	0.816395
act	Total Current Asset	0.781164
at	Total Assets	0.933912
bkvlps	Book Value Per Share	0.917933
ceql	Common Equity Liquidation Value	0.919281
csho	Common Shares Outstanding	0.99414
dltt	Total Long-Term De	0.929807
dp	Depreciation and Amortization	0.890821
dvc	Dividends Common/Ordinary	0.929049
dvt	Dividends - Total	0.929033
ebitda	Earnings Before Interest, Taxes, D &A	0.908273
epspi	Basic EPS Including Extraordinary Items	0.905453

epspx	Basic EPS Excluding Extraordinary Items	0.927653
gp	Gross Profit (Net Sales - COG	0.925649
ib	Income Before Extraordinary Items	0.932522
lct	Current Liabilities - Tot	0.793005
ni	Net Income (Loss)	0.930273
revt	Revenue - Tota	0.930286
sale	Gross Sales reduced by cash discounts, trade discounts, and returned sales and allowances	0.930286
exchg	Stock Exchange Code	0.999984
costat	Postal code	1
dvpsp_c	Dividends Per Share - Pay Date - Calendar	0.930758
prcc_c	Price Close - Annual - Calendar	0.987044
adjex_c	Cumulative Adjustment Factor by Ex-Date - Calendar	0.994425
dvpsp_f	Dividends per Share - Pay Date - Fisc	0.935486
mkvalt	Market Value - Total - Fis	0.361921
prcc_f	Price Close - Annual - Fisc	0.961411
adjex_f	Cumulative Adjustment Factor by Ex-Date - Fiscal	0.997695
gsector	GIC Sectors	0.878938
mktcap	Market Capitalization (csho*prcc_f)	0.956394
bookvalue	Book Value (csho*bkvlps)	0.917933
bv	Book Value (csho*bkvlps)	0.917933
pb	Price to Book Ratio (mktcap/bookvalue)	0.885926
pe	Price to Earnings Ratio (mktcap/ib)	0.893297
cr	Current Ratio (act/lct)	0.780331
leverage	Leverage (dltt/at)	0.92963

turnover	Turnover (revt/at)	0.929836
grossMarginRatio	Gross Margin Ratio (gp/sale)	0.913867
ROA	Return on Assets (ni/at)	0.929958
oshaughnessy_cfo	ib+dp	0.890798
cfo_assetsratio	Cash Flow from Operations to Total Assets Rat	0.890605
divYield	Dividend Yield (dvpsp_c/prcc_c)	0.923645
Decade	Text of Decade of Dat	1
Sector	GIC Sectors	0.878938

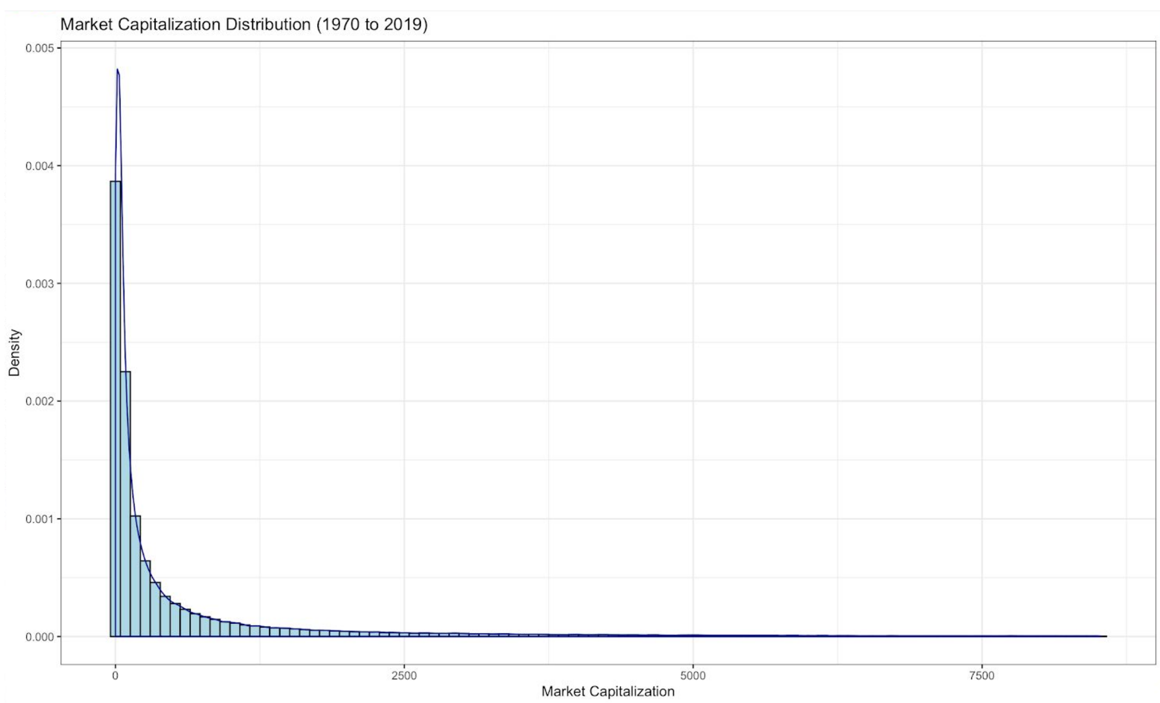
Financial Data Distributions

This section is split into several subsections for each of the financial metrics that has been analyzed. The distributions are split over time and/or sector to highlight key data.

Market Capitalization

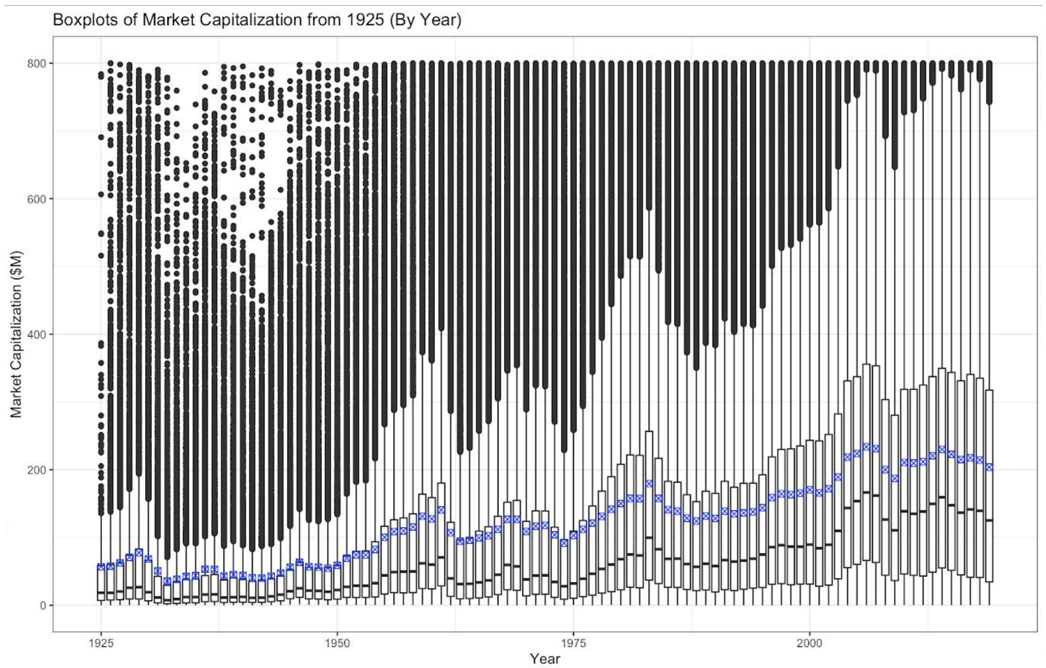
Market capitalization (often called "market cap") is the total dollar market value of a company's outstanding shares. It is the product of the company's outstanding shares and current price of one share. Market cap measures what a company is worth and allows investors to understand the size of the company compared to others on the market. Today, the company with the highest market cap is Apple, with a market cap of about $1.94 trillion. They had briefly surpassed the $2 trillion mark in late August, but fell in early September. A company is typically placed into one of four categories based on their market cap: micro-cap ($50-300MM), small-cap ($300MM-2B), mid-cap ($2-10B), and large-cap (>$10B).

Figure 1. Distribution of Market Capitalization



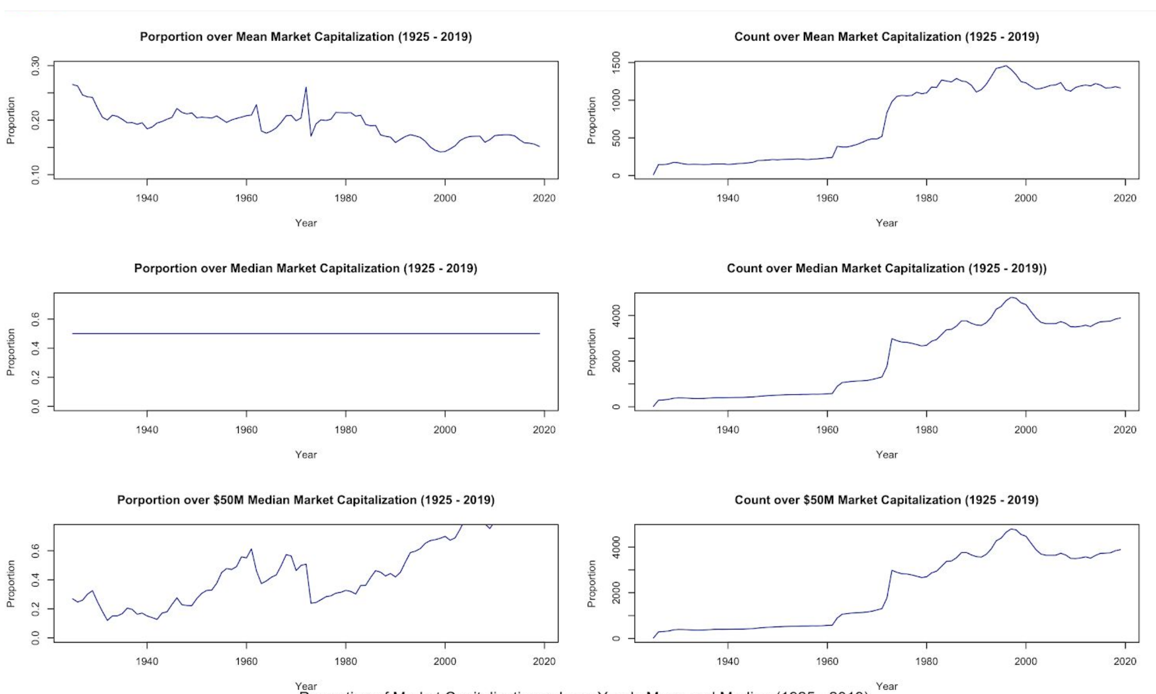
The market capitalization is positively skewed, which indicates that the market is dominated by small- to mid- cap stocks with few reaching into the large-cap range.

Figure 2. Distribution of Market Capitalization by Time



In temporal distributions, the blue square represents the mean and the black bar represents the median. Since 1925, market cap has gradually increased with only a few drops, such as in 2008.

Figure Set 3. Proportions of Market Cap Above Yearly Mean/Median

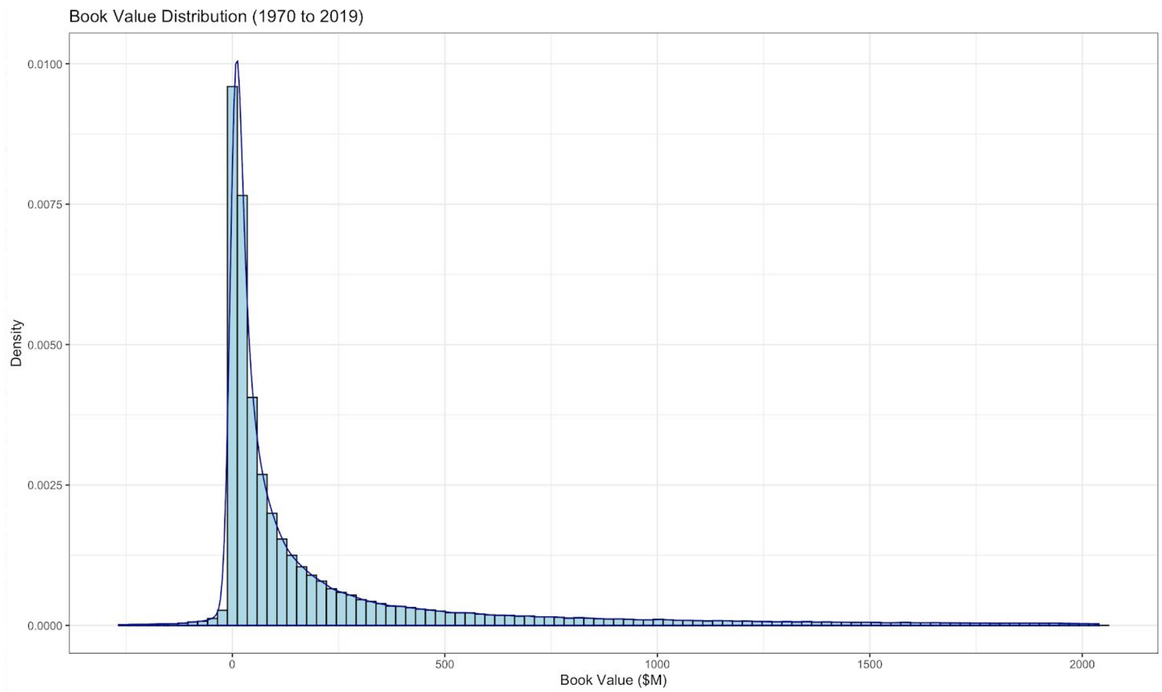


Analyzing the proportions and counts of market caps above the mean and median allow us to observe trends in the spread of market caps and number of companies. We can see that the proportion of mean market cap has generally decreased, while the proportion over median has increased. This is due to the mean being lifted by exceedingly large-cap stocks, such as Apple, Microsoft, and Amazon.

Company Book Value

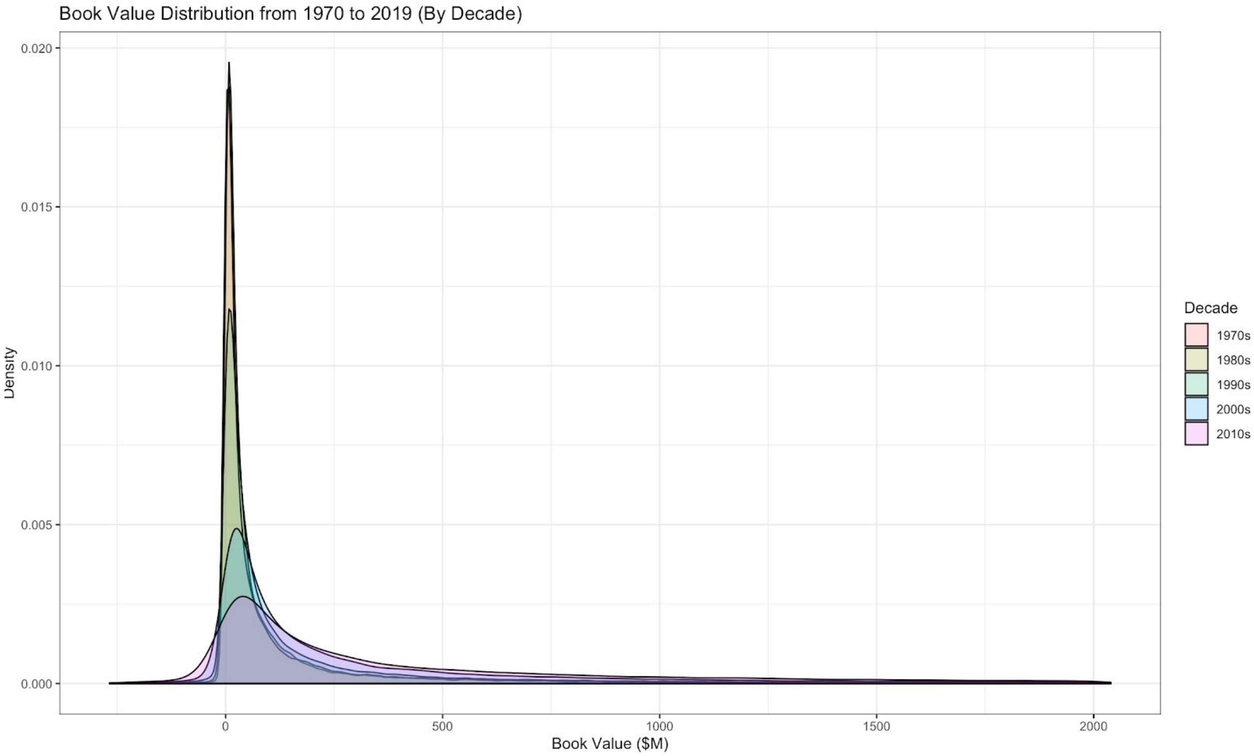
A company's book value is the difference between its assets and liabilities. It gives a good indication to investors of a stock's valuation. We calculated this value using the data.

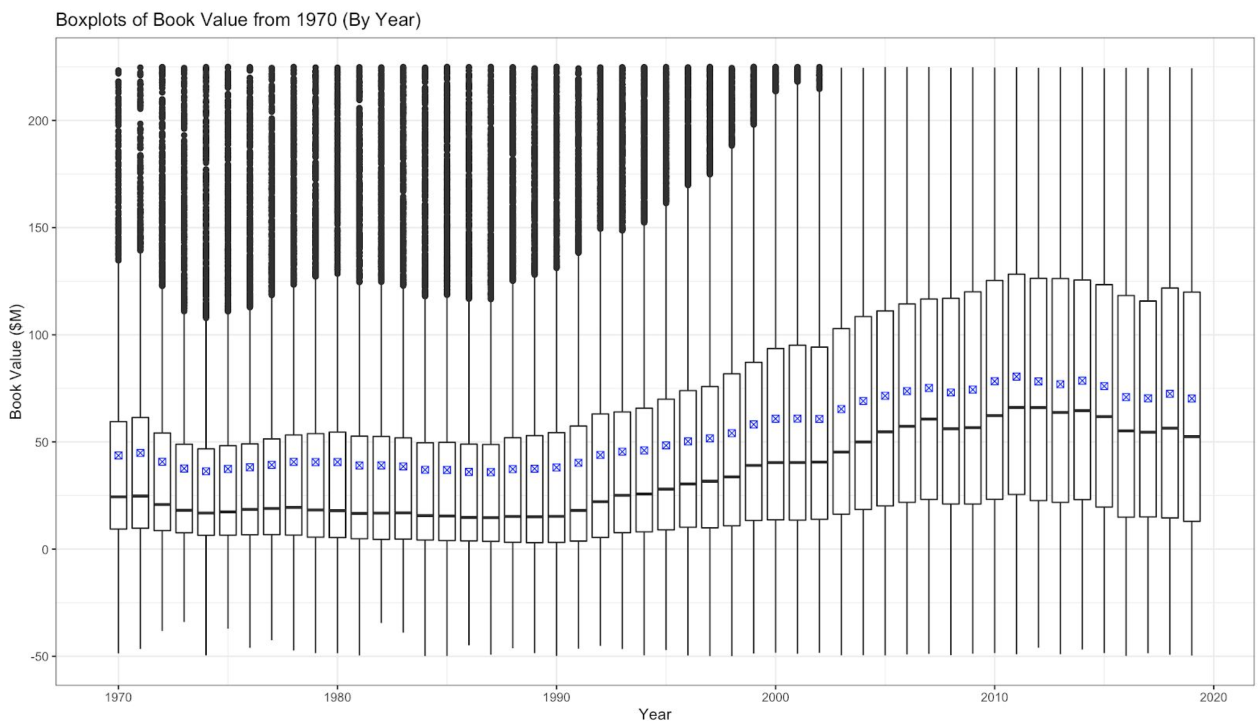
Figure 4. Distribution of Book Value



As shown in the graph, book values for major publicly traded companies trends low overall, leading to the conclusion that it is more optimal to keep assets and liabilities relatively near each other in value.

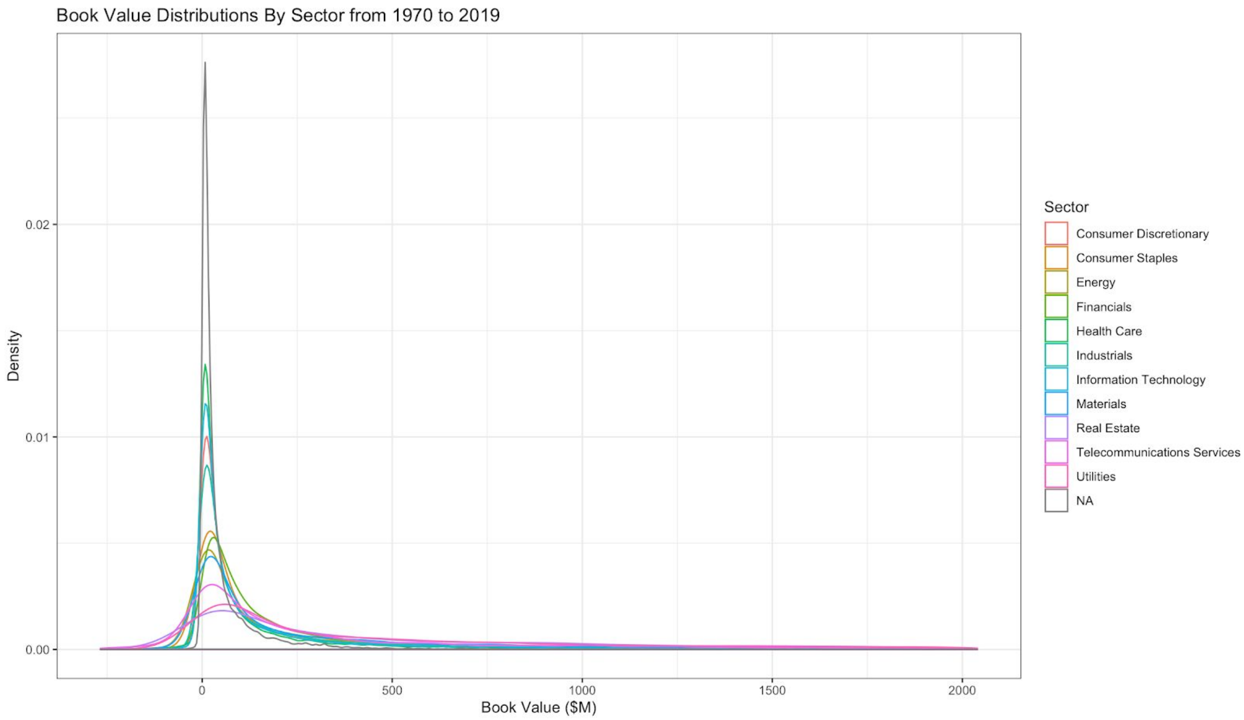
Figure Set 5. Distribution of Book Value by Time





Over time, however, book value curves have flattened due to the rise in use of more aggressive business practices that create more disparity between assets and liabilities.

Figure Set 6. Distribution of Book Value by Sector



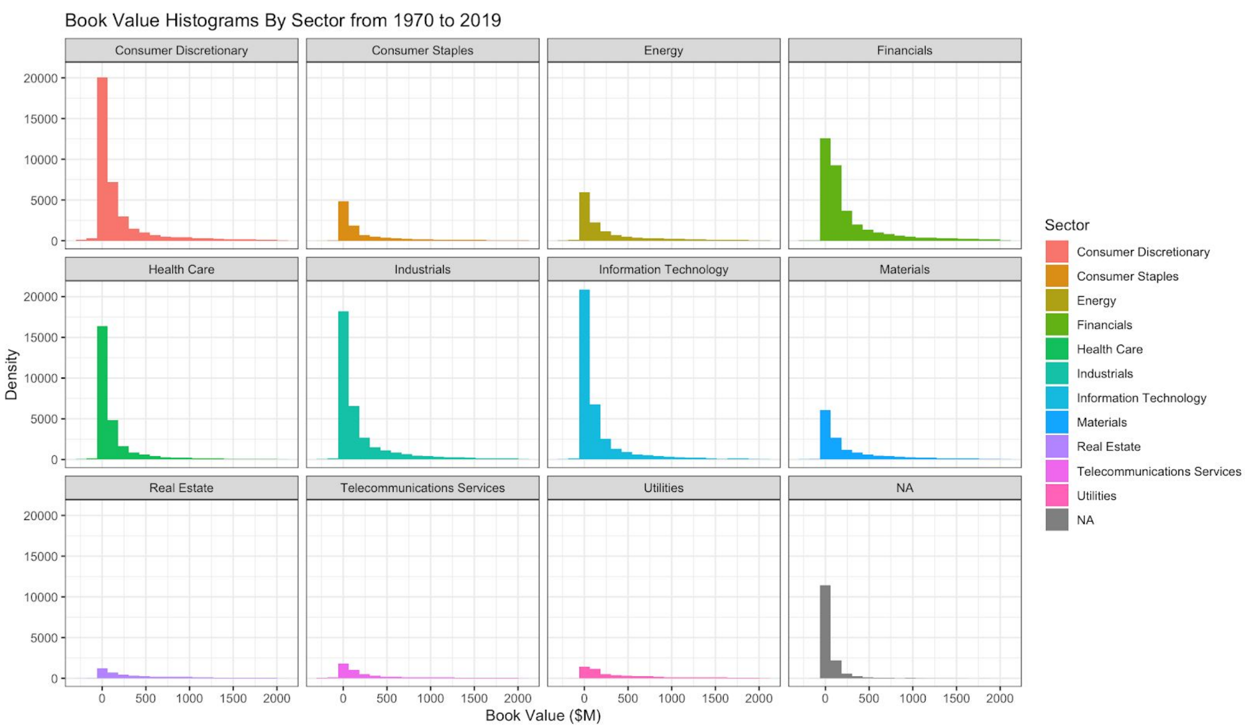


Table 1. Statistics for Company Book Value

	Sector	Mean	Minimum	Q1	Median	Q3	Maximum	Standard Deviation
1	Energy	2559.847	-3704.995	15.046	131.032	806.355	198646.082	12262.367
2	Materials	1029.742	-5510.536	21.218	124.211	647.921	100329.886	3888.412
3	Industrials	567.443	-8617.011	11.969	58.838	287.677	130566.089	2921.799
4	Consumer Discretionary	617.060	-86153.986	11.104	50.325	223.210	176421.999	3946.350
5	ConsumerStaples	1132.000	-13243.947	17.530	85.139	519.400	84631.830	4987.120
6	Health Care	724.212	-9659.987	8.046	36.543	163.671	89953.062	4170.177
7	Financials	1963.771	-96619.740	40.285	128.494	515.010	424790.914	11190.600
8	Information Technology	650.363	-5765.014	9.905	45.032	203.379	134047.080	3663.242
9	Telecommunications Services	4110.905	-10505.988	25.922	232.342	1662.359	224234.147	14672.413
10	Utilities	1829.243	-5992.880	80.105	402.639	1594.001	67153.928	4498.998
11	Real Estate	919.732	-420301.987	57.815	337.387	1163.140	22584.152	6219.975

Revenue

Revenue is the income generated from normal business activities. It is referred to as sales on the income statement and is the top line from which costs are subtracted to get net income.

Figure Set 7. Distribution of Revenue

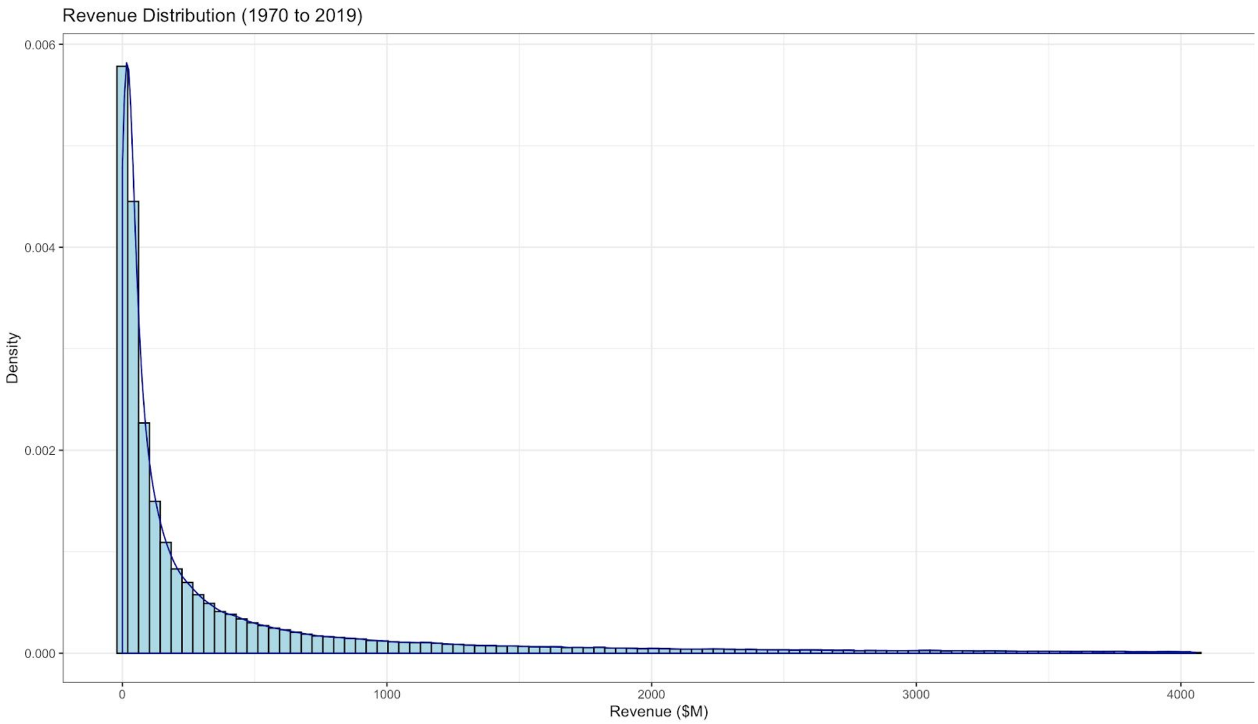
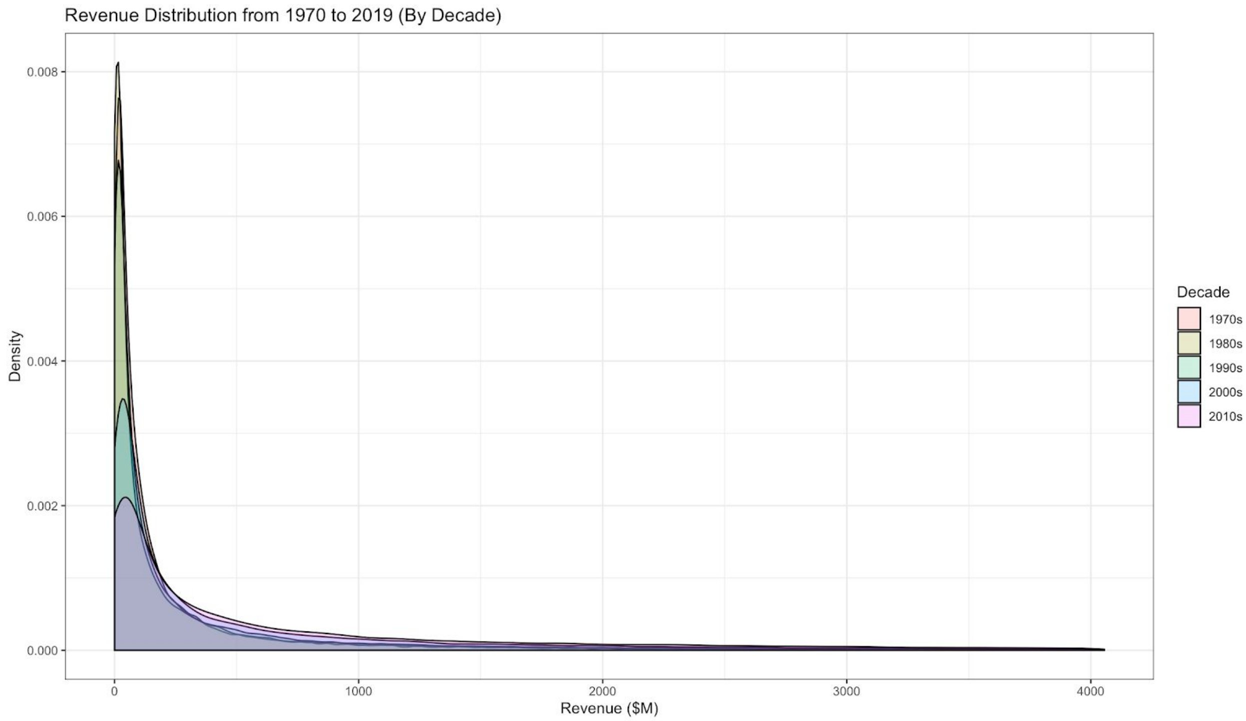


Figure Set 8. Distribution of Revenue by Time



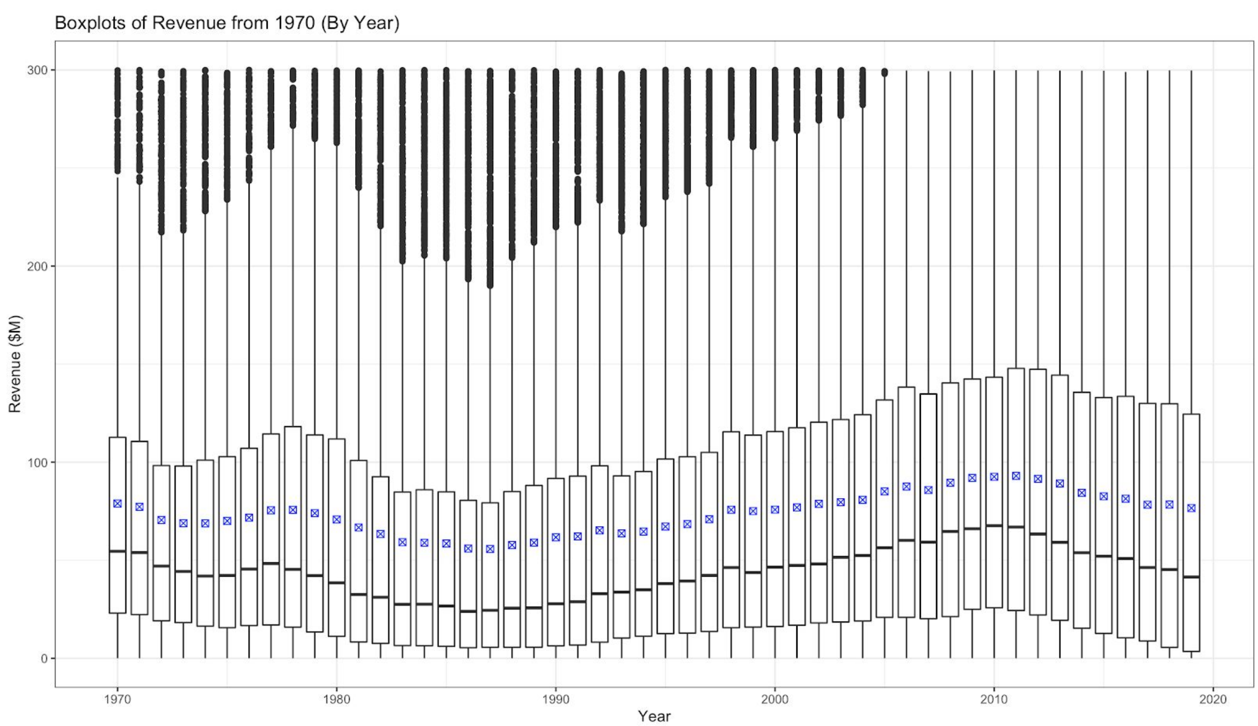
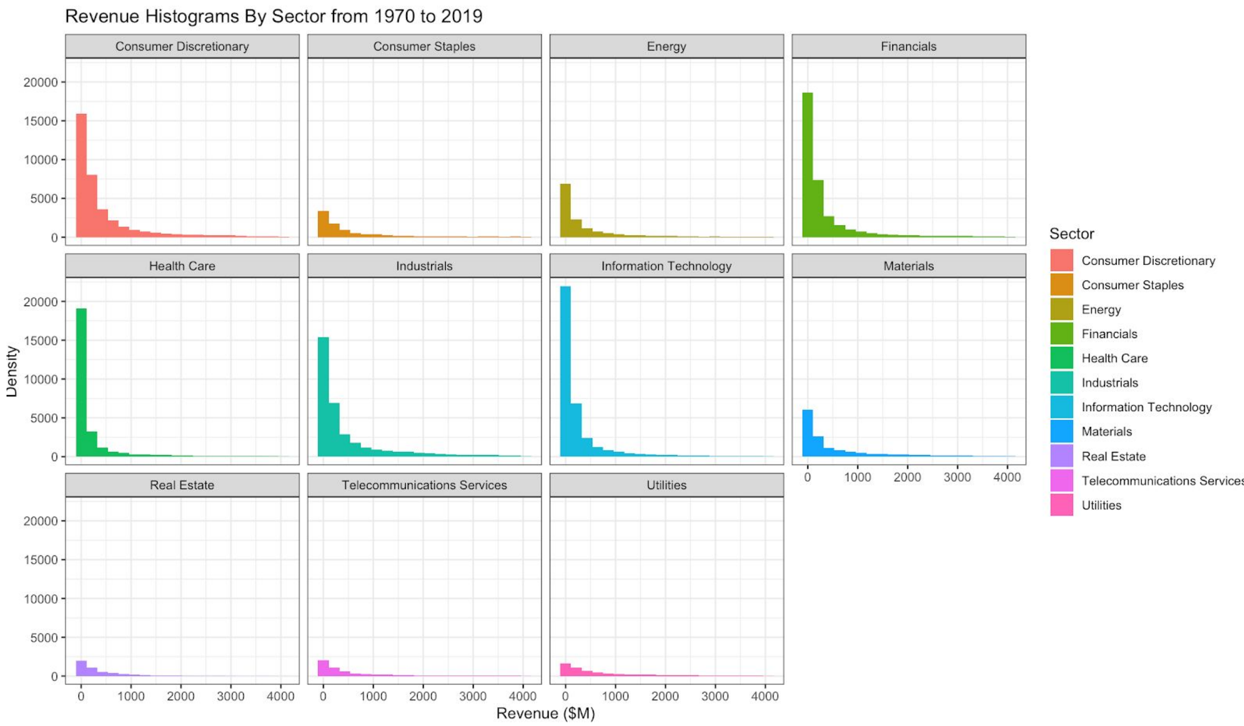
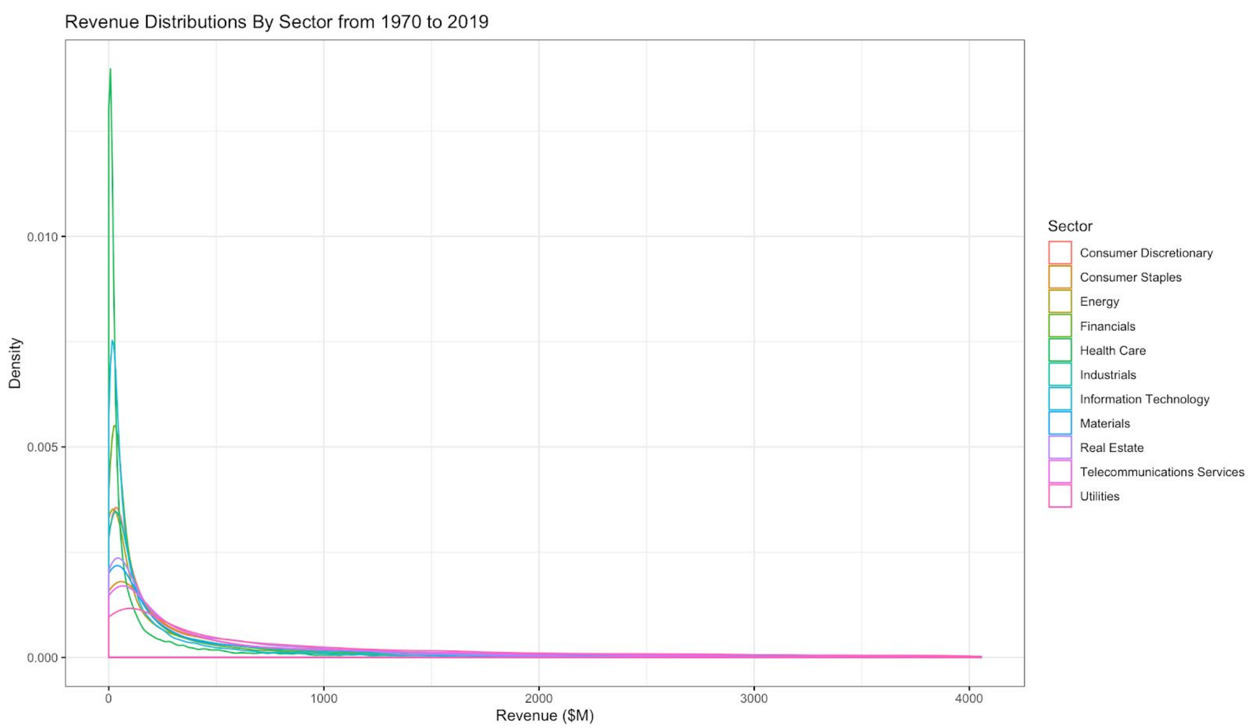


Figure Set 9. Distribution of Revenue by Sector

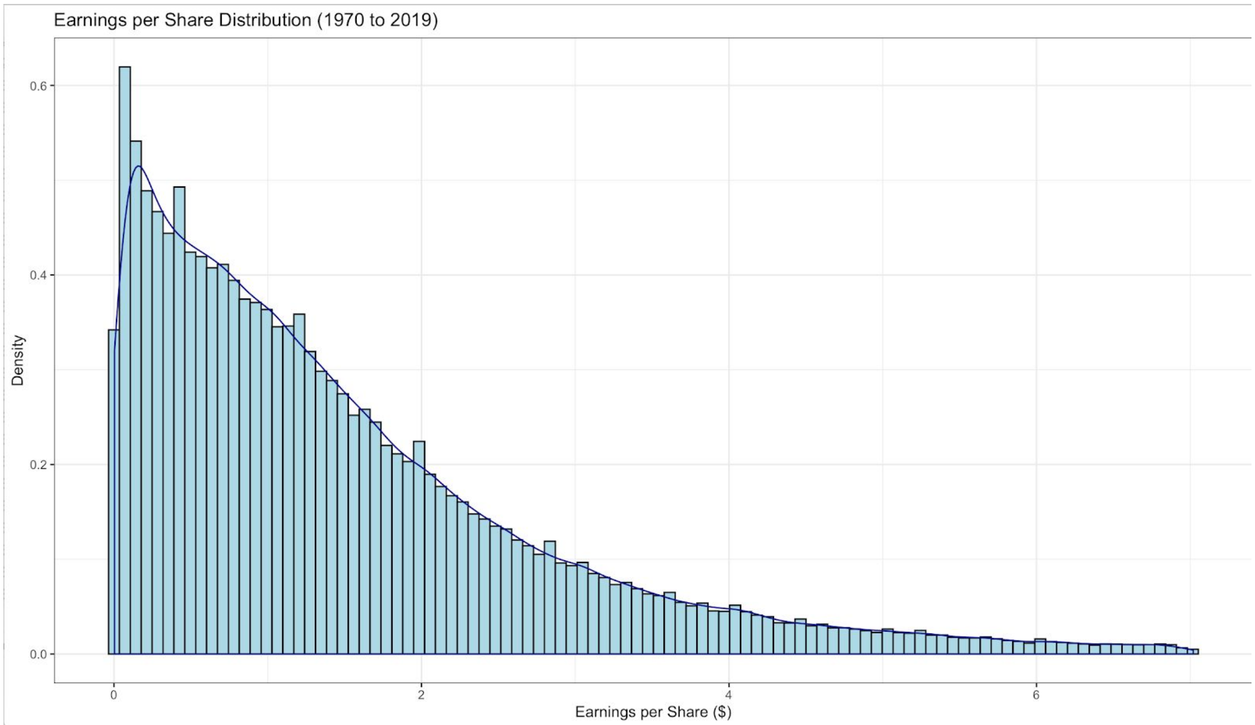




Earnings Per Share

Earnings per share (EPS) is the ratio of net income to outstanding shares. It is an important measure for investors as it represents a company's profitability. It also plays an essential role in stock pricing for investors. Negative EPS values were excluded from our analysis.

Figure 10. Distribution of EPS



As shown in Figure 10, earnings per share generally trends on the lower end for publicly traded companies as issuing larger amounts of shares for lower prices serves as a better capital raising system to bring in more investors and raise capital more quickly.

Figure Set 11. Distribution of EPS by Time

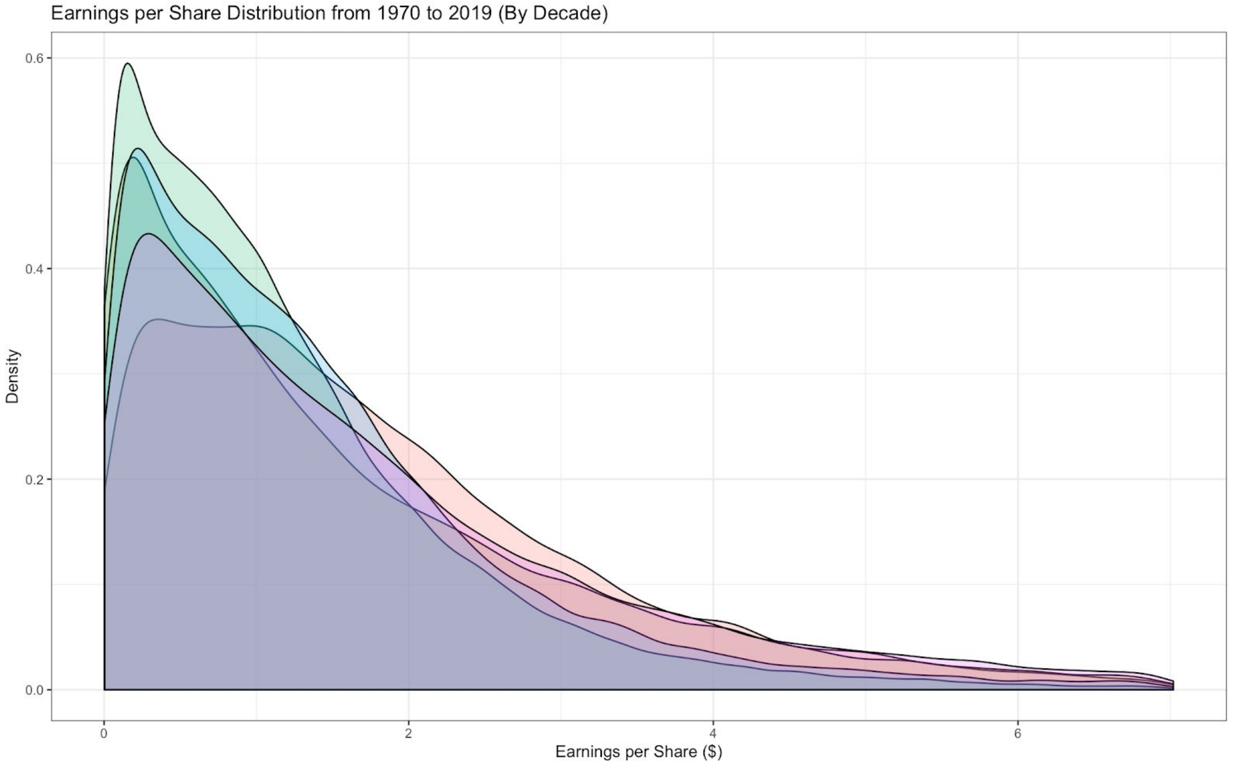
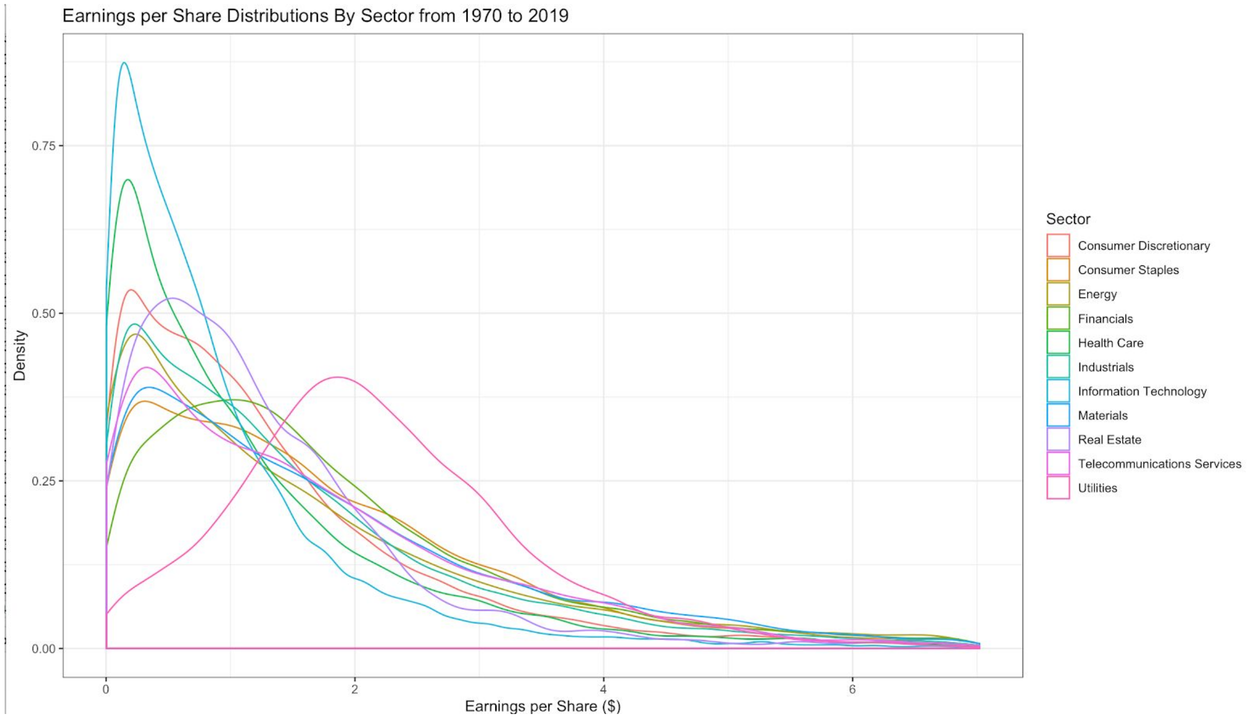
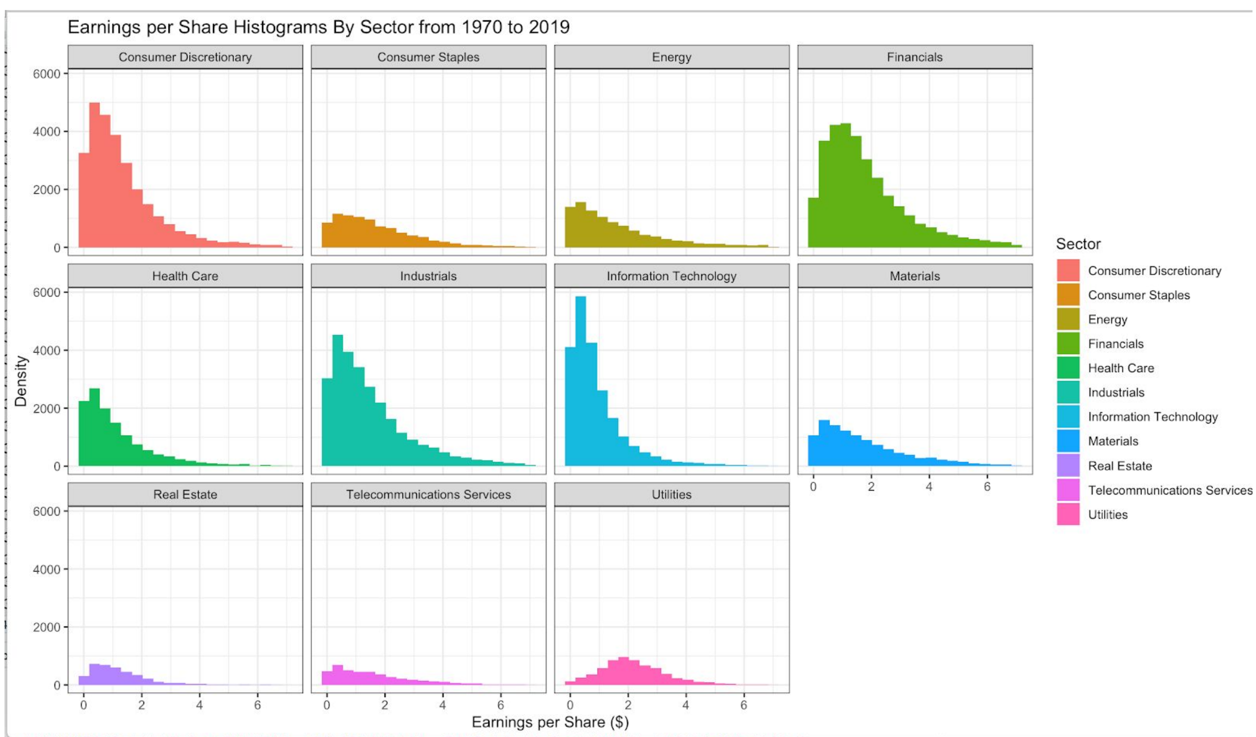


Figure Set 12. Distribution of EPS by Sector





Net Income

Net income is the total profit of a company during the period. It is calculated as revenue minus all expenses including costs, depreciation, taxes, etc. It is used to calculate many key ratios, such as EPS.

Figure 13. Distribution of Net Income

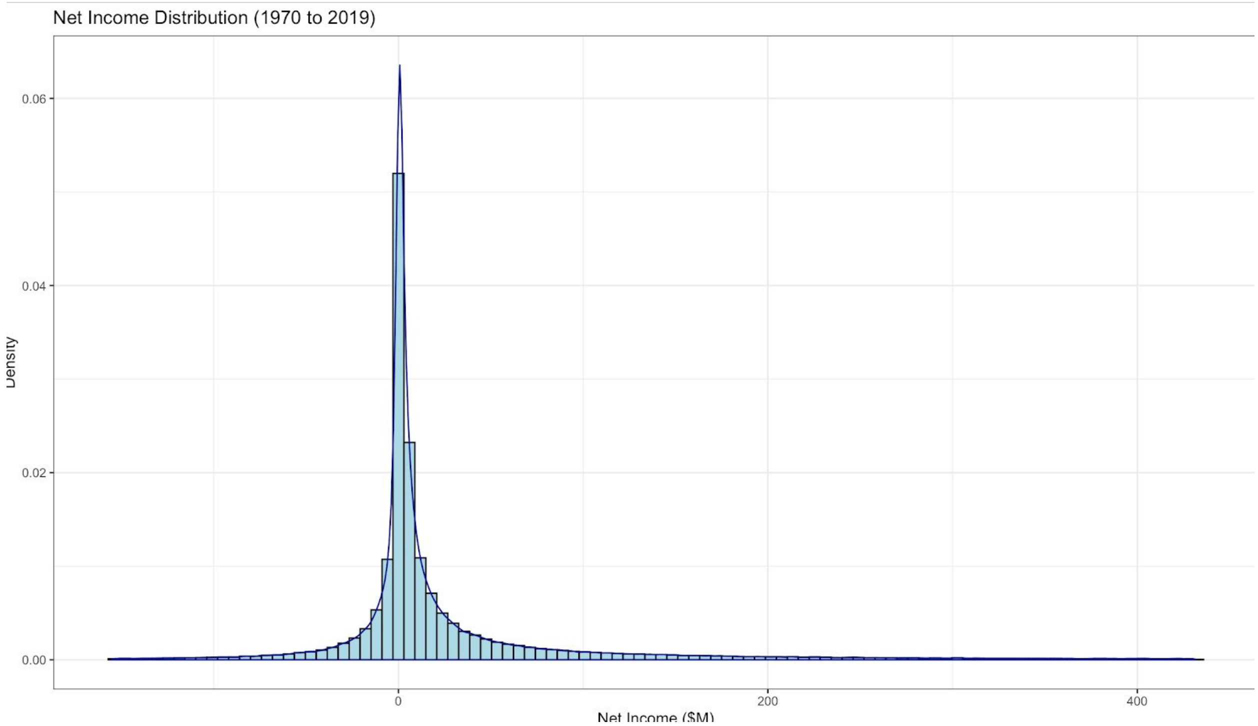
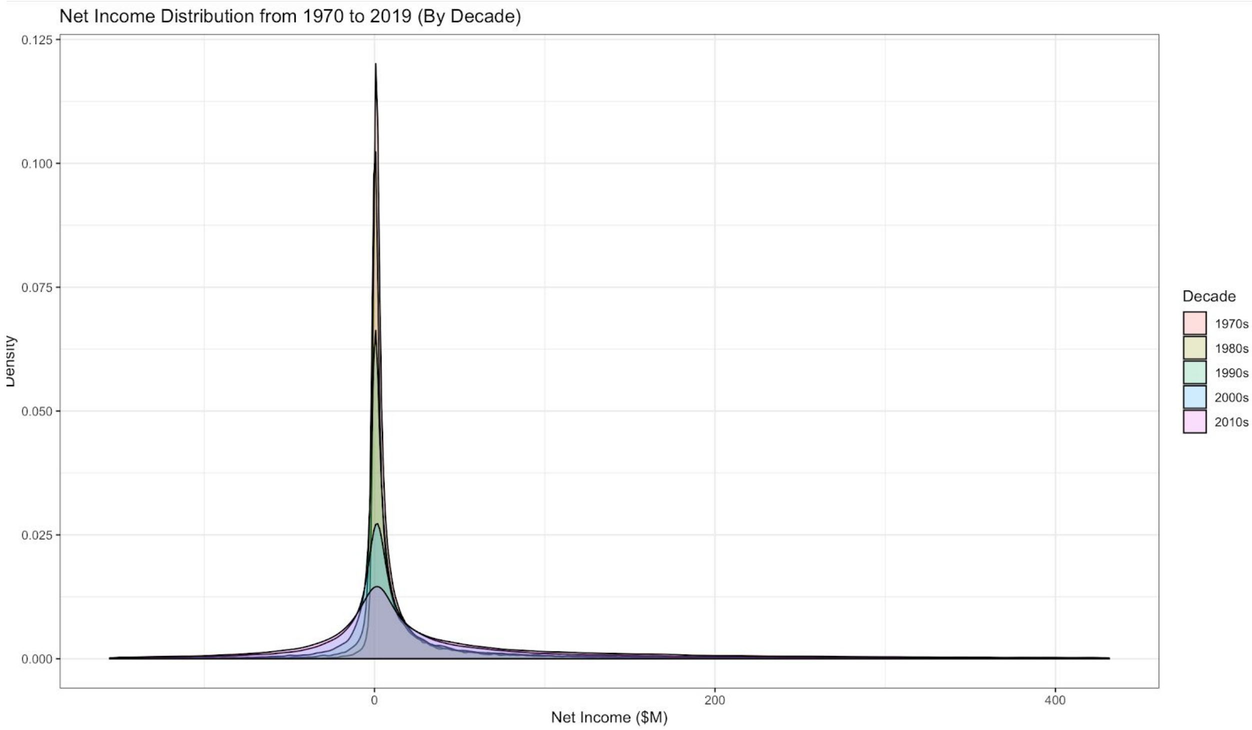


Figure Set 14. Distribution of Net Income by Time



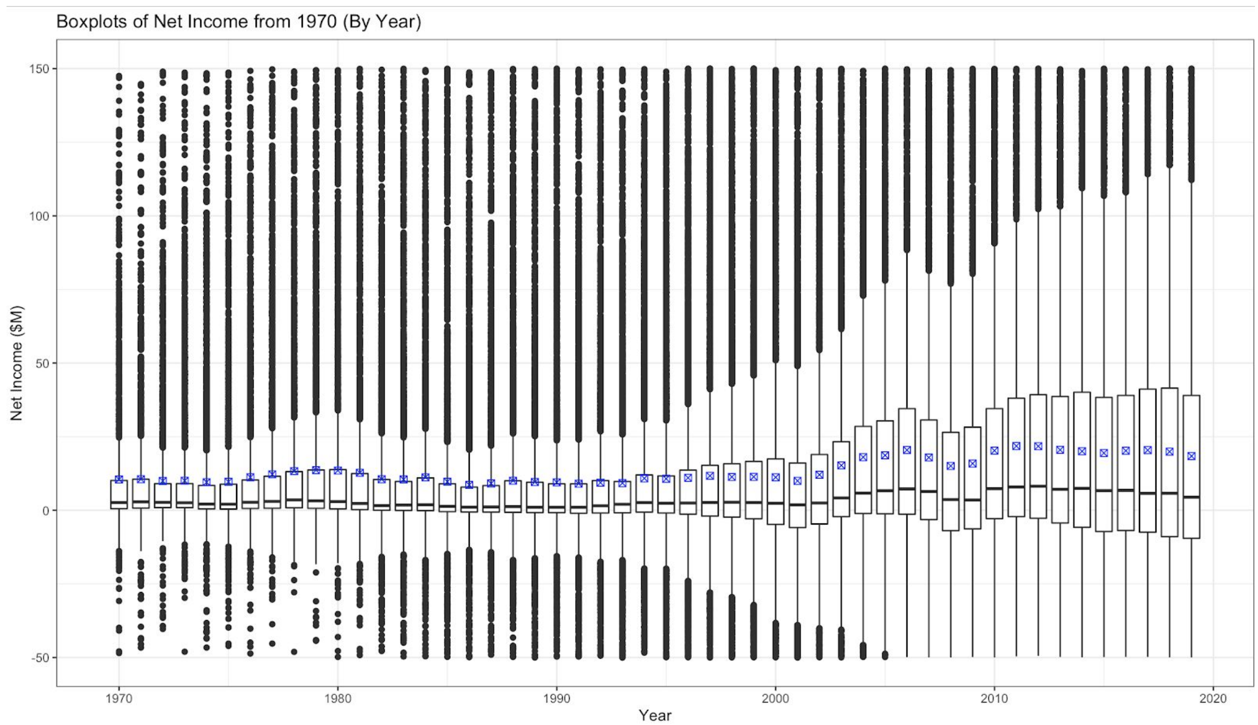
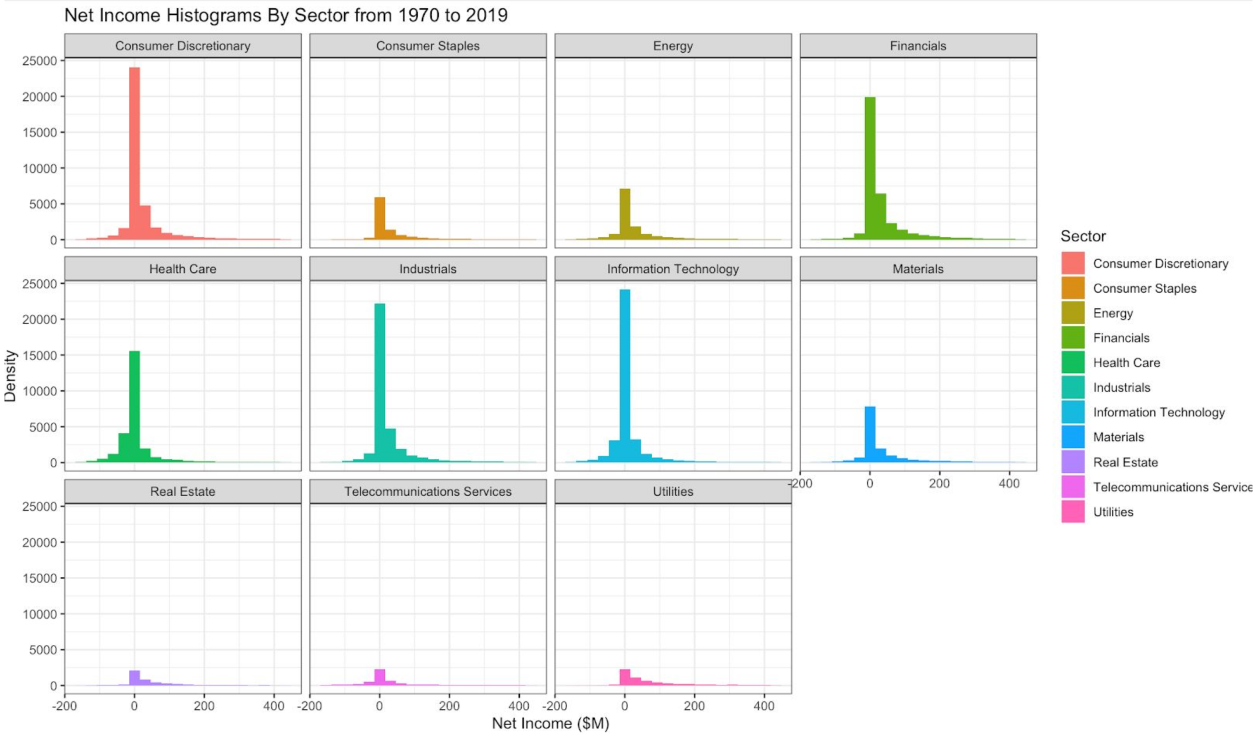
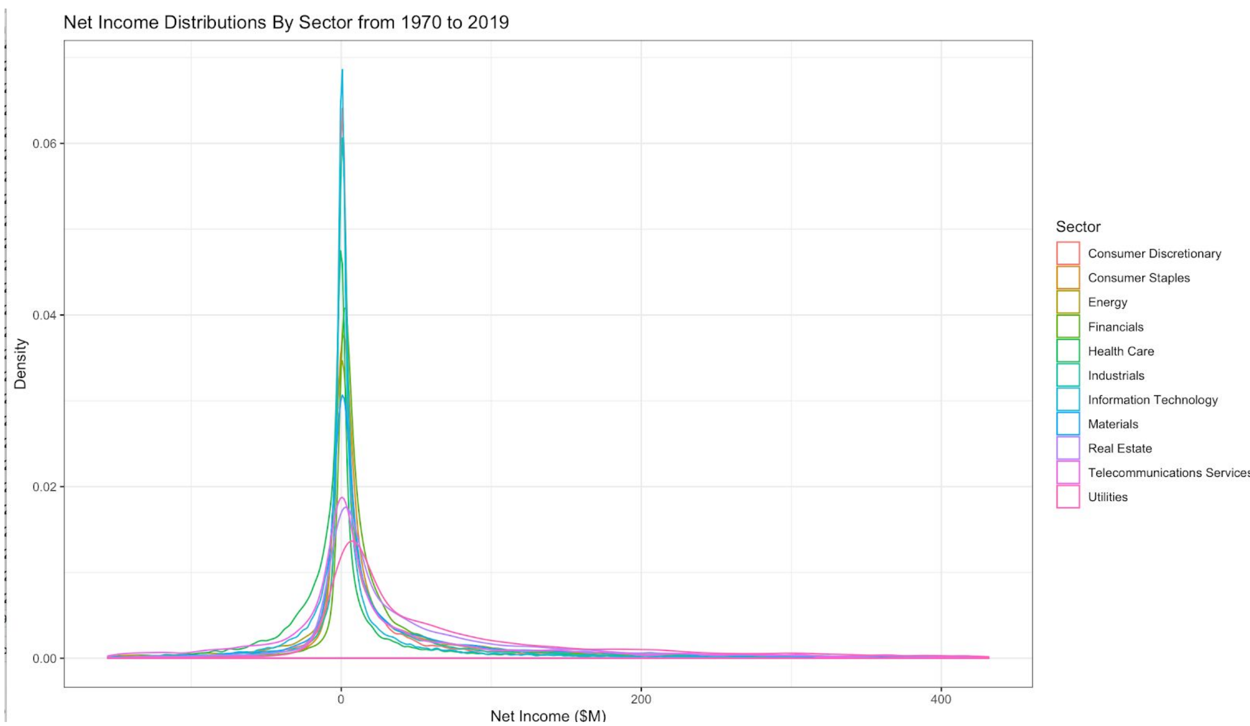


Figure Set 15. Distribution of Net Income by Sector


Net Income Histograms By Sector from 1970 to 2019
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care
Industrials
Information Technology
Materials
Real Estate
Telecommunications Services
Utilities
Density
Net Income ($M)
Sector


Net Income Distributions By Sector from 1970 to 2019
Density
Net Income ($M)
Sector
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care
Industrials
Information Technology
Materials
Real Estate
Telecommunications Services
Utilities

Price to Book Ratio

Price to book ratio (P/B ratio) is determined by dividing a company's share price by its book value per share. Book value per share is total assets minus liabilities. Investors understand this ratio is the price one pays for one dollar of the company's assets. Lower P/B ratios may indicate an undervalued stock and vice versa. P/B ratio is especially useful as a comparison tool for determining which of two stocks might be a better value at the given time.

Figure 16. Distribution of P/B Ratio

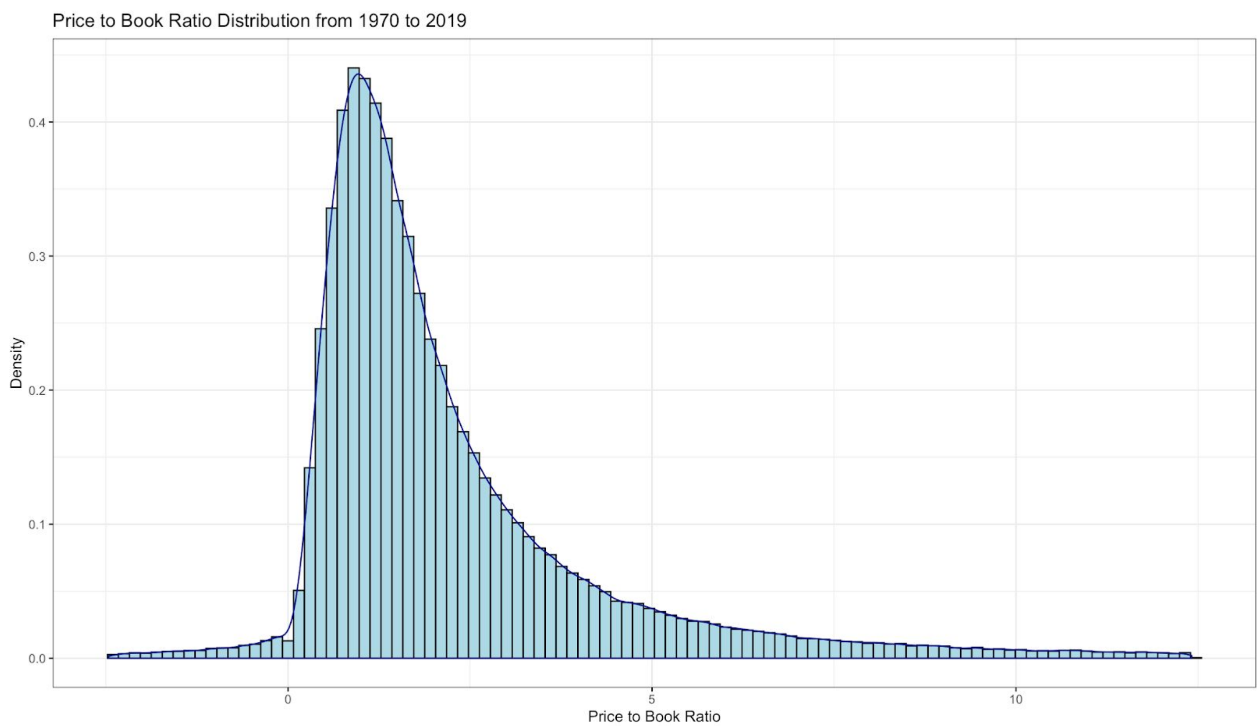
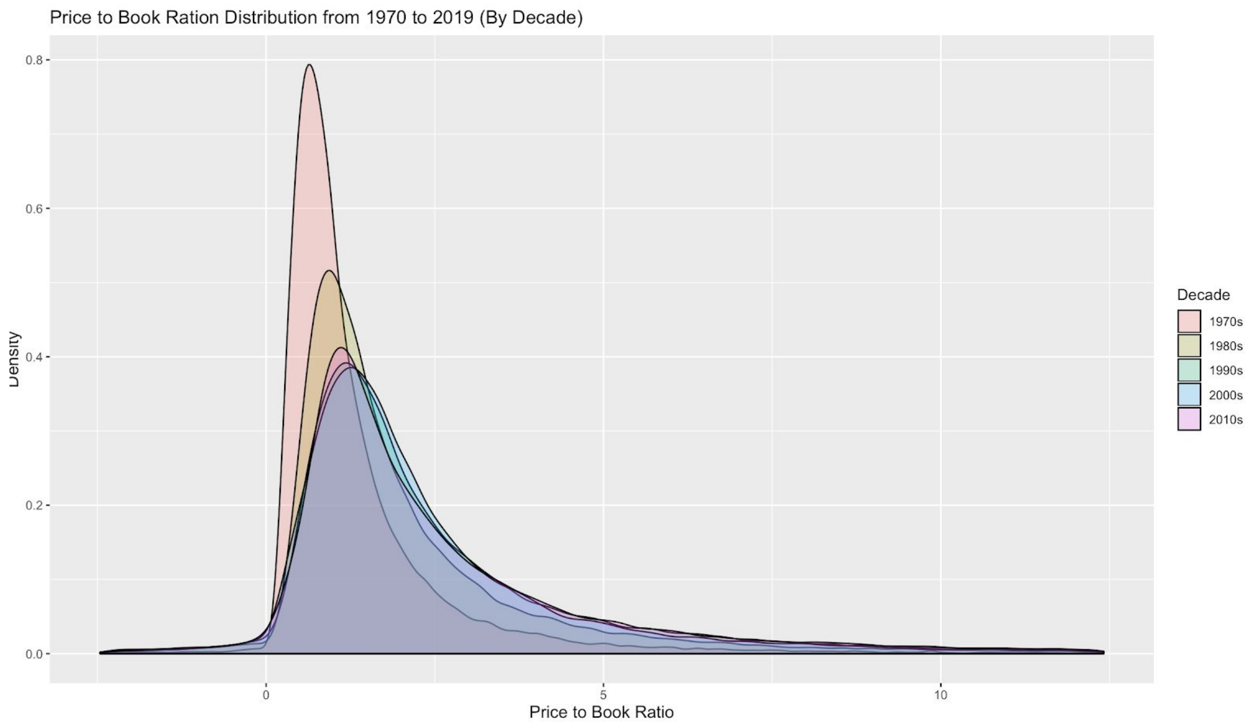


Figure Set 17. Distribution of P/B Ratio by Time



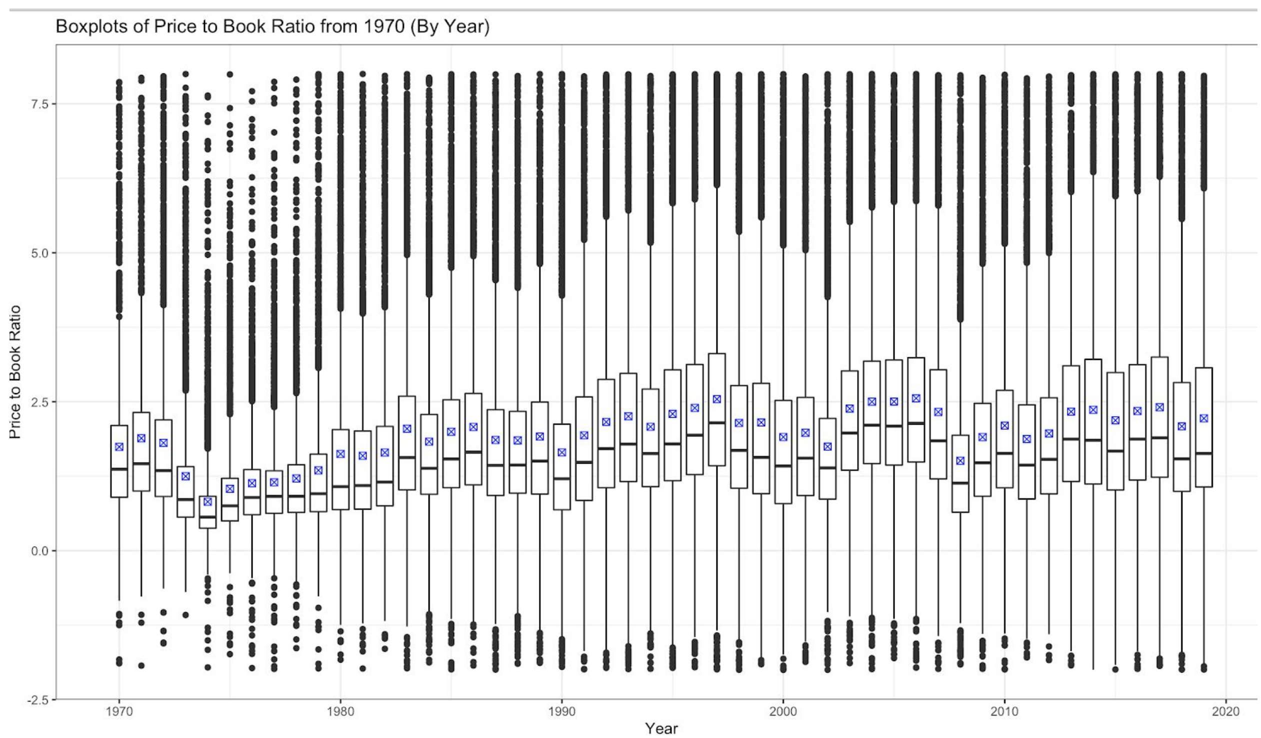
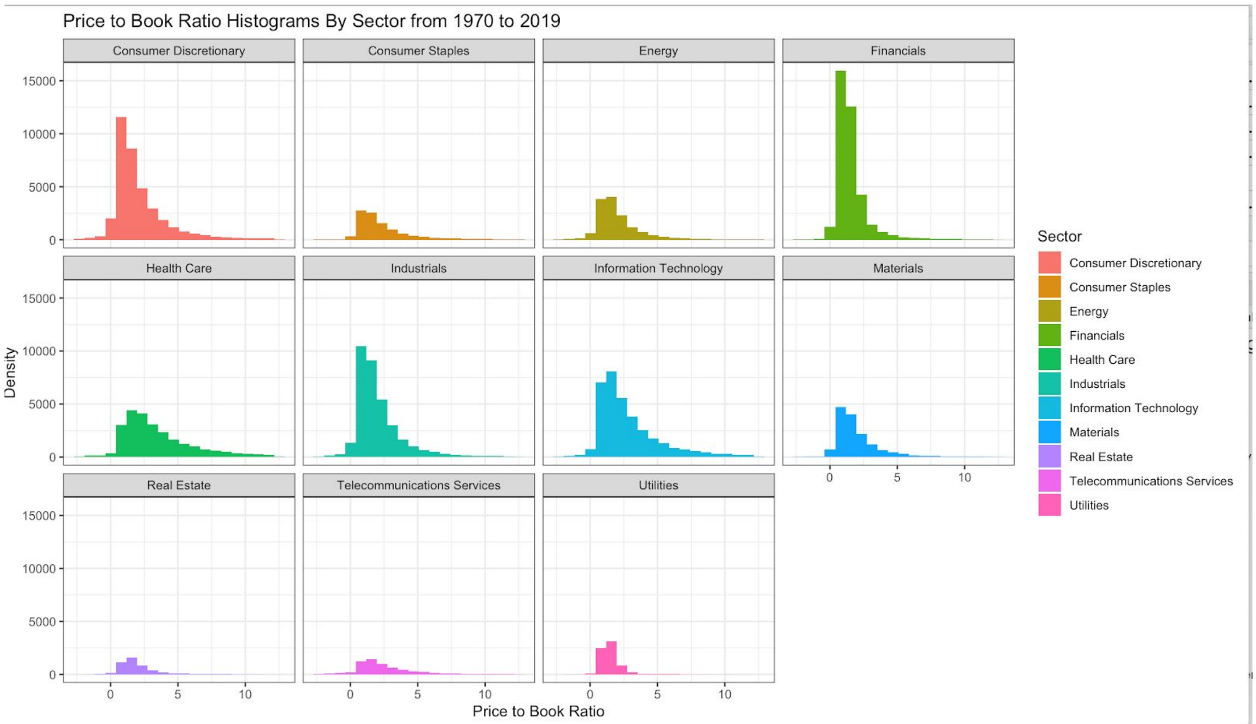
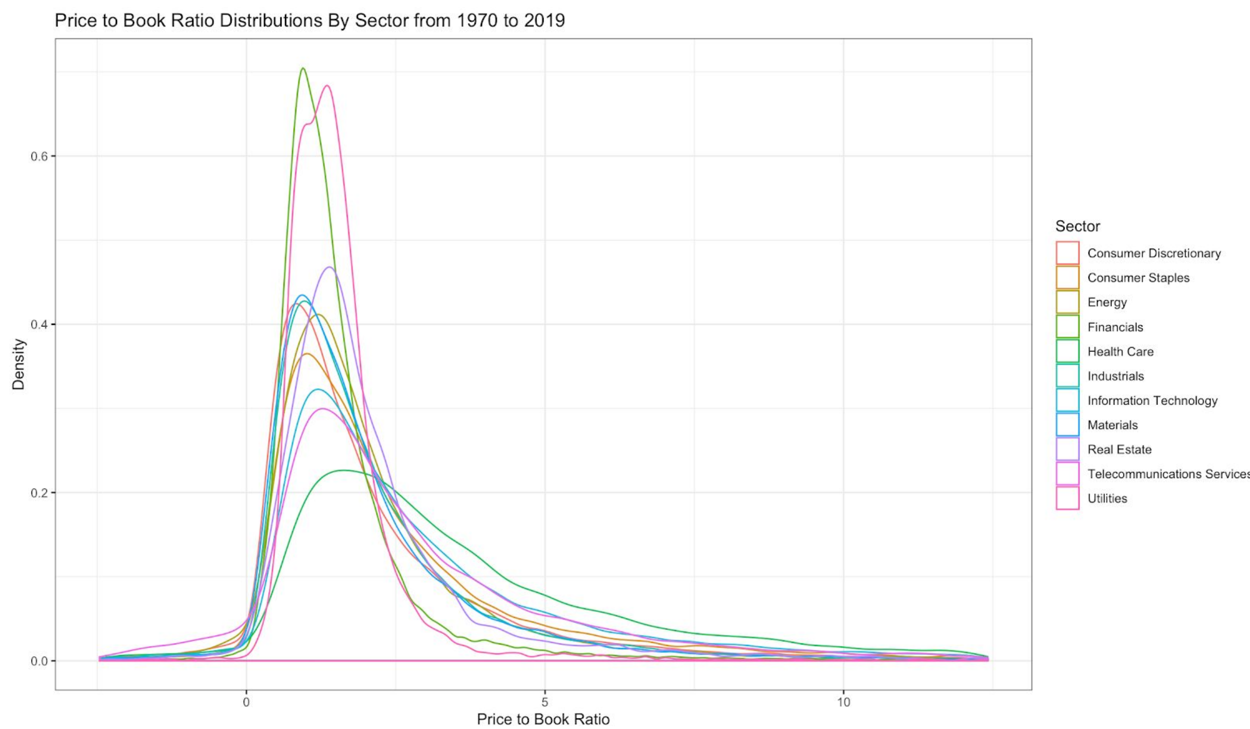


Figure Set 18. Distribution of P/B Ratio by Sector


Price to Book Ratio Histograms By Sector from 1970 to 2019
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care
Industrials
Information Technology
Materials
Real Estate
Telecommunications Services
Utilities
Density
Price to Book Ratio
Sector


Price to Book Ratio Distributions By Sector from 1970 to 2019
Density
Price to Book Ratio
Sector
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care
Industrials
Information Technology
Materials
Real Estate
Telecommunications Services
Utilities

Price to Earnings Ratio

Price to earnings ratio (P/E ratio) represents the share price over the earnings per share, thus giving a quick measure of whether a stock is overvalued, undervalued, or fairly valued.

Figure 19. Distribution of P/E Ratio

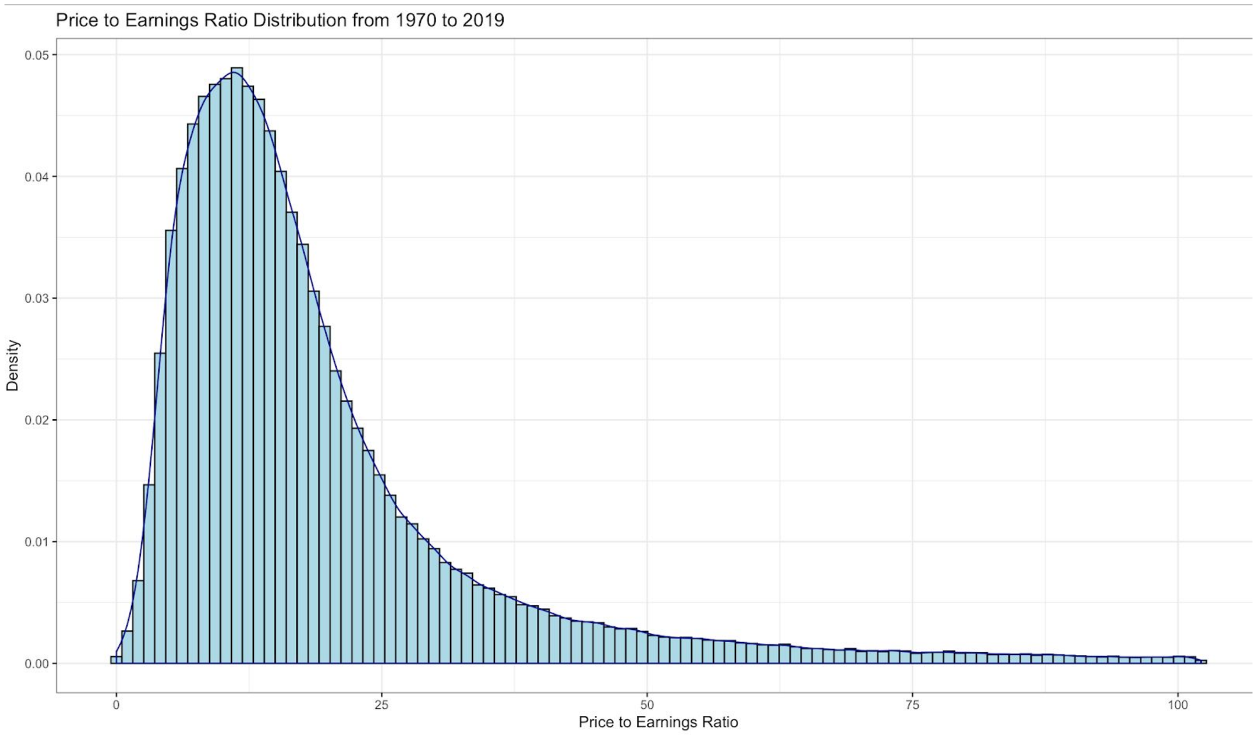
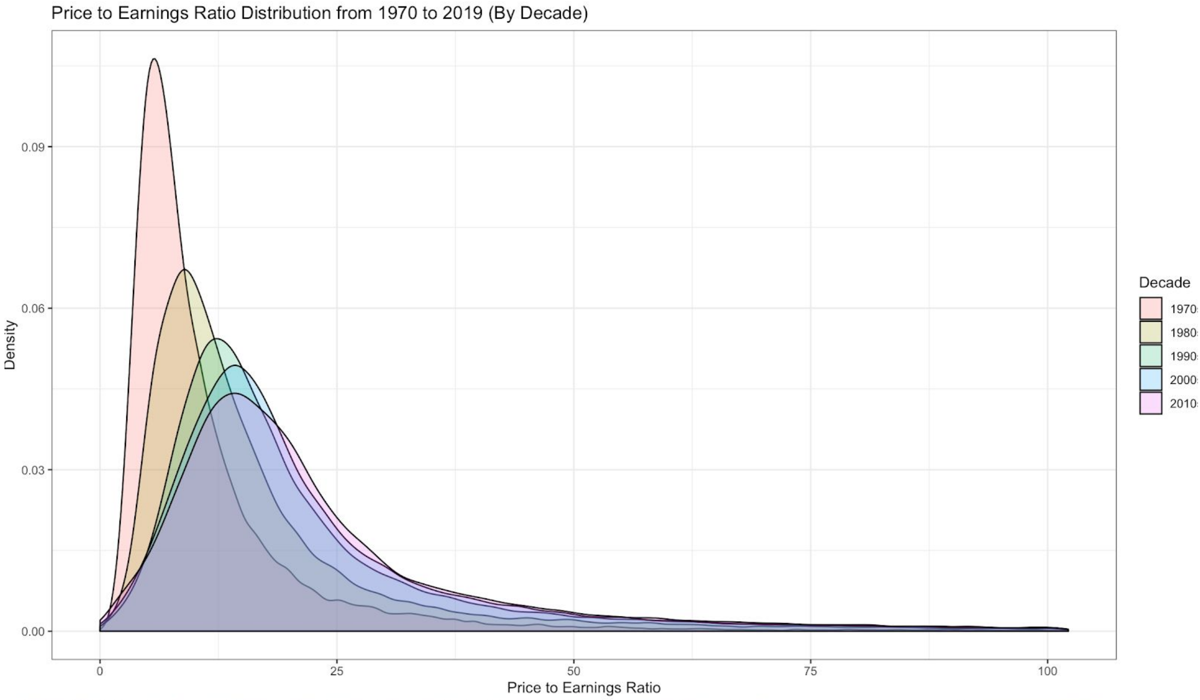


Figure Set 20. Distribution of P/E Ratio by Time



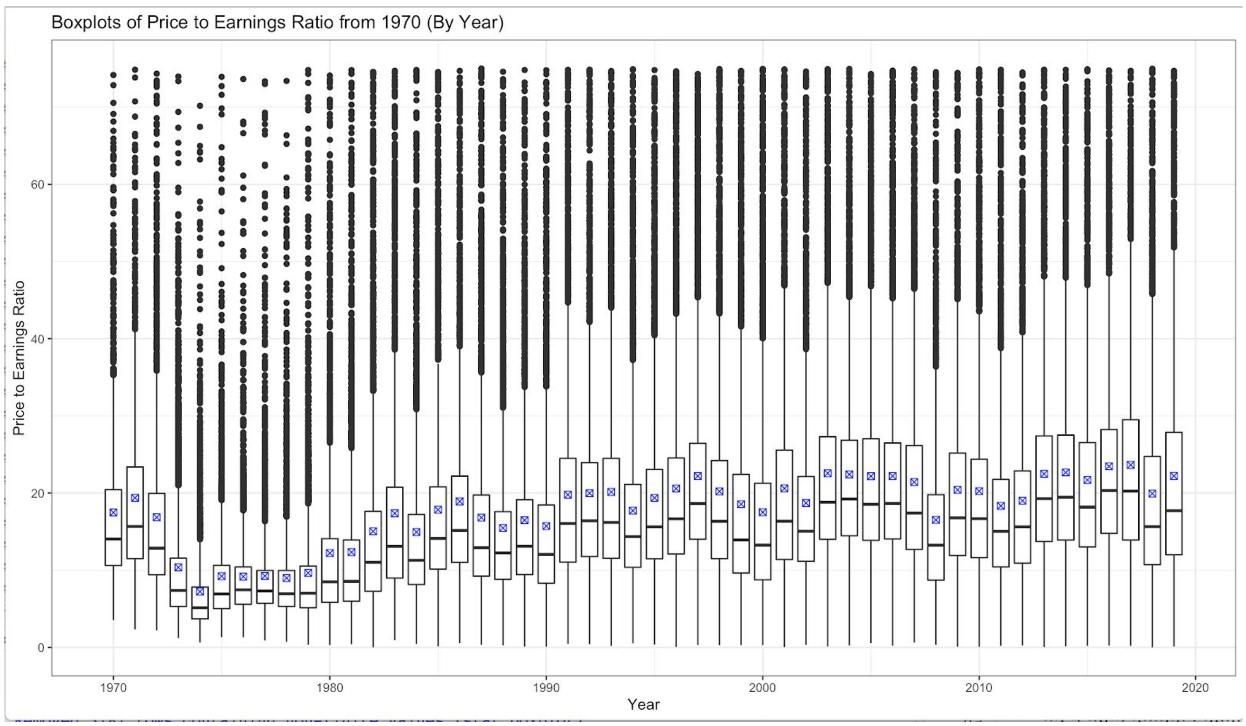
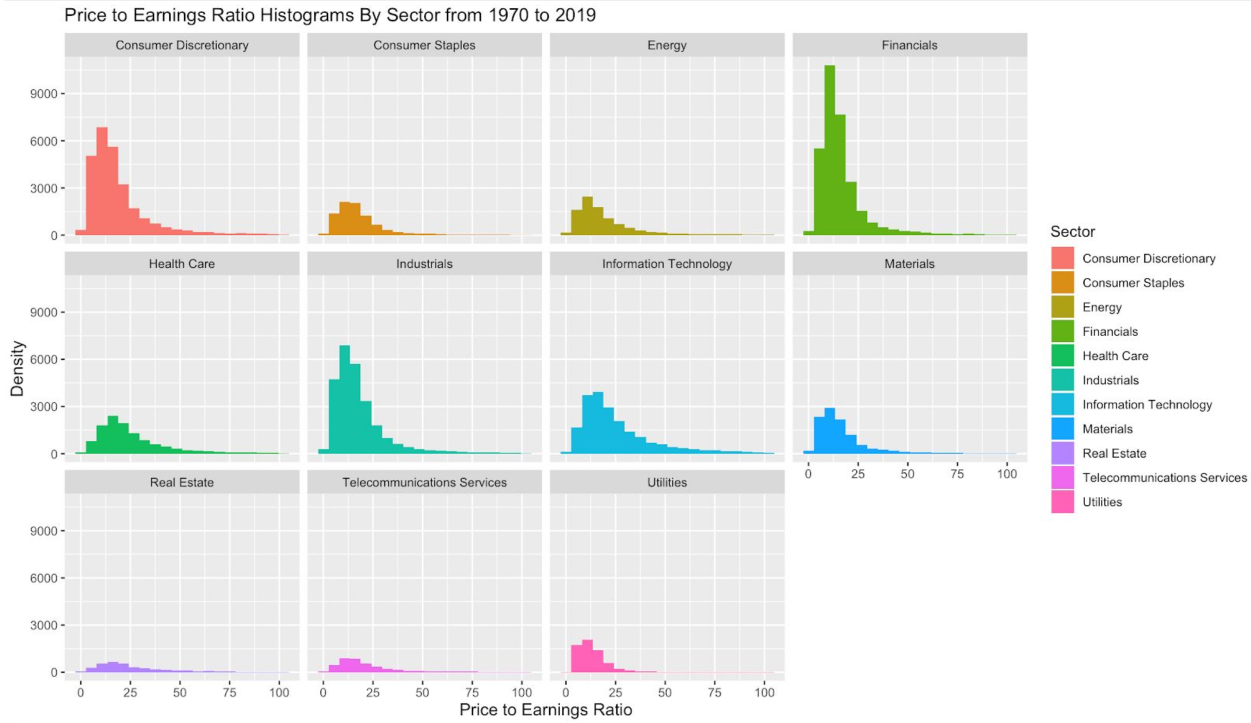
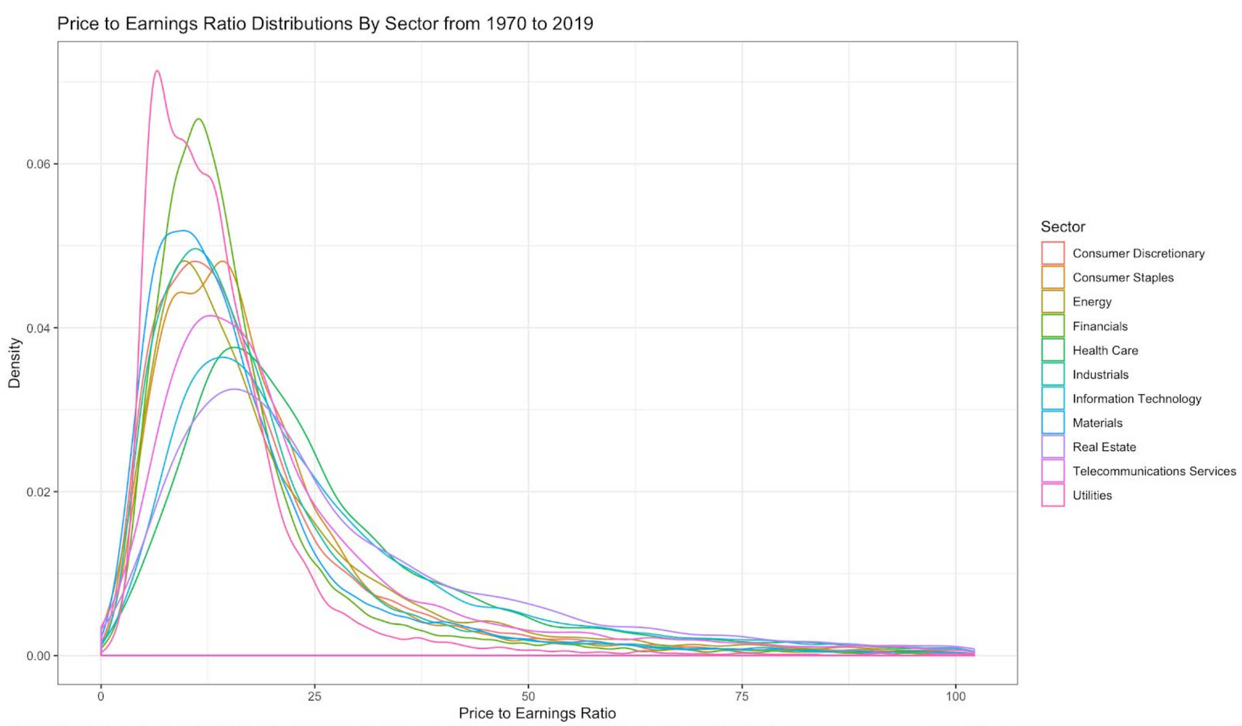


Figure Set 21. Distribution of P/E Ratio by Sector





Dividend Yield

Dividend yield is the total annual dividends divided by the market capitalization, serving as a measure of earning on investment.

Figure 22. Distribution of Dividend Yield

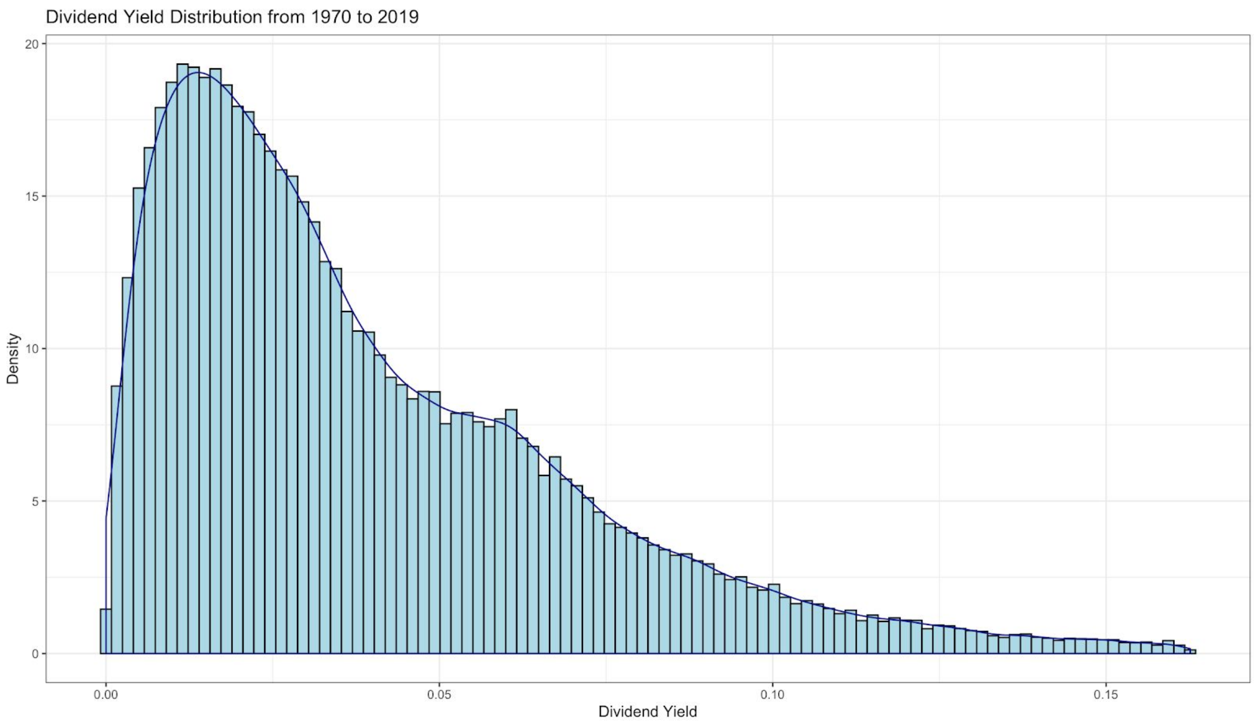
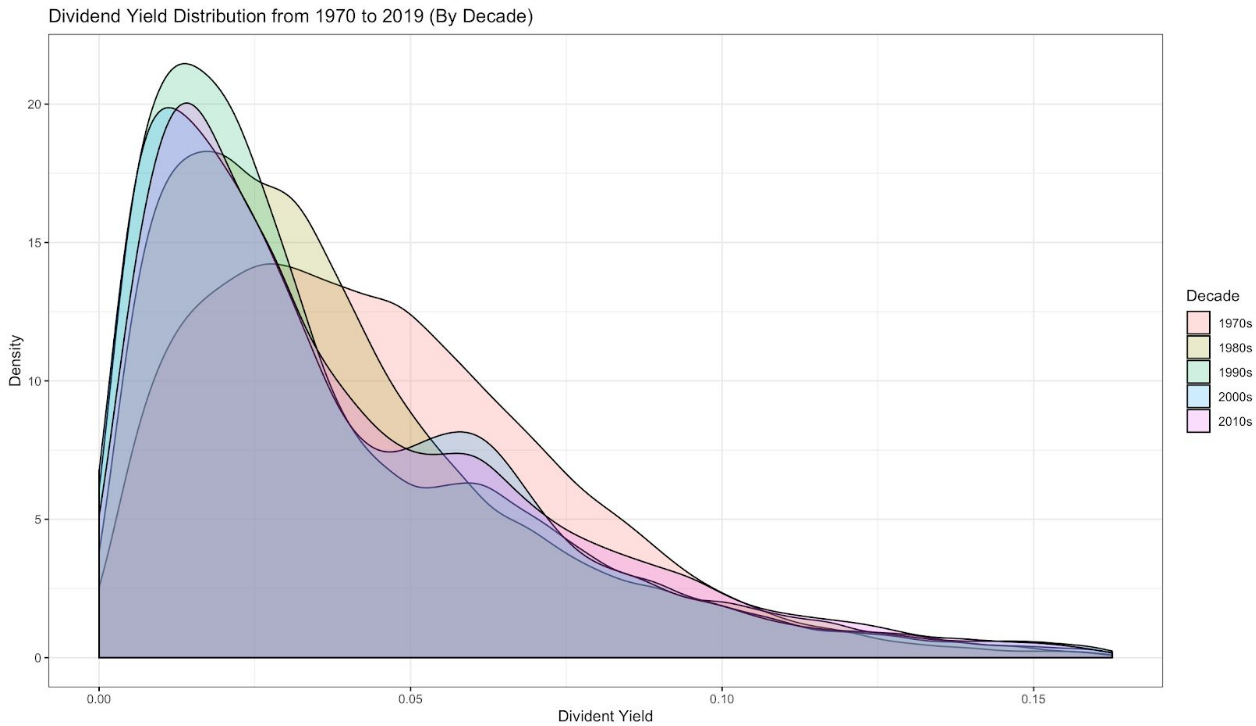


Figure Set 23. Distribution of Dividend Yield by Time



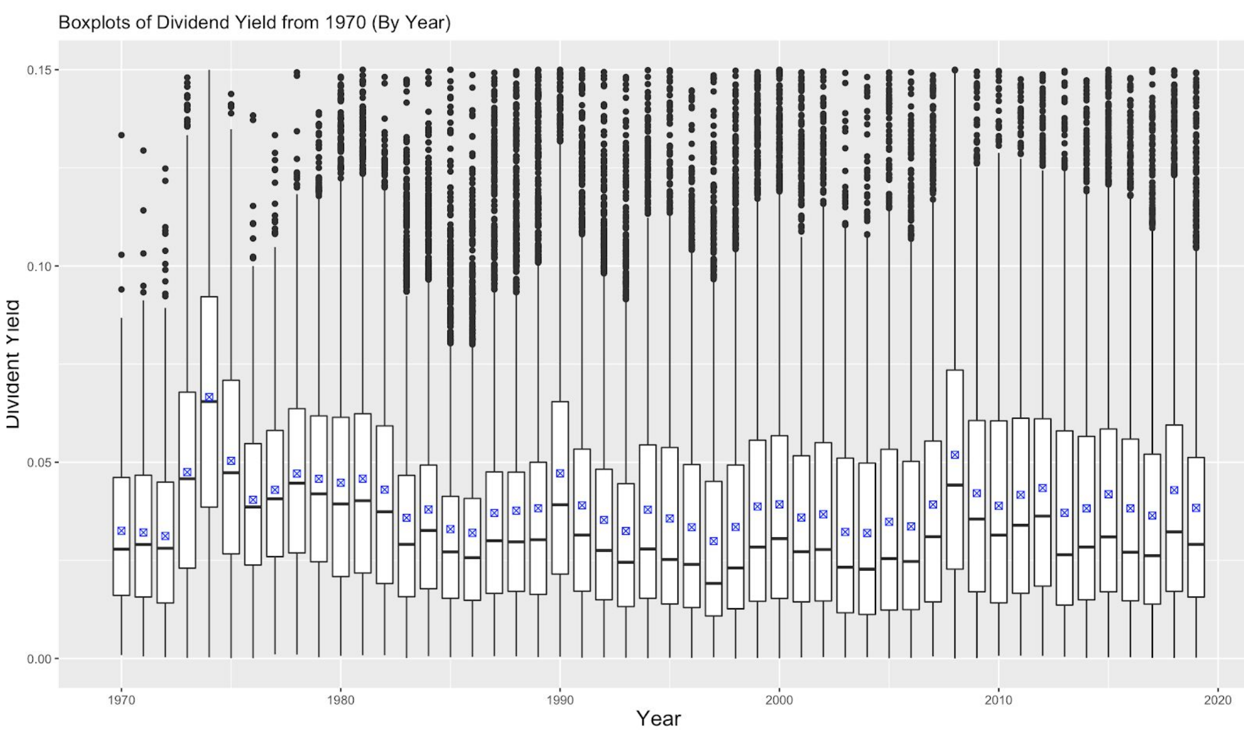
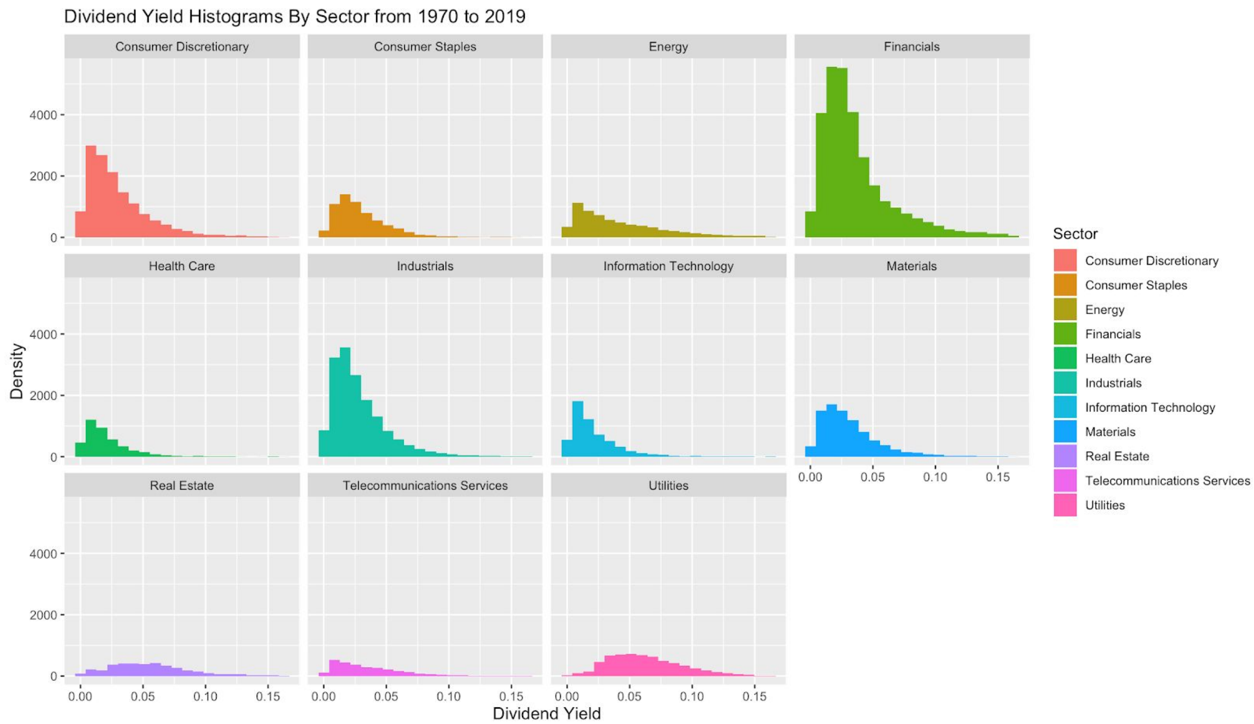


Figure Set 24. Distribution of Dividend Yield by Sector



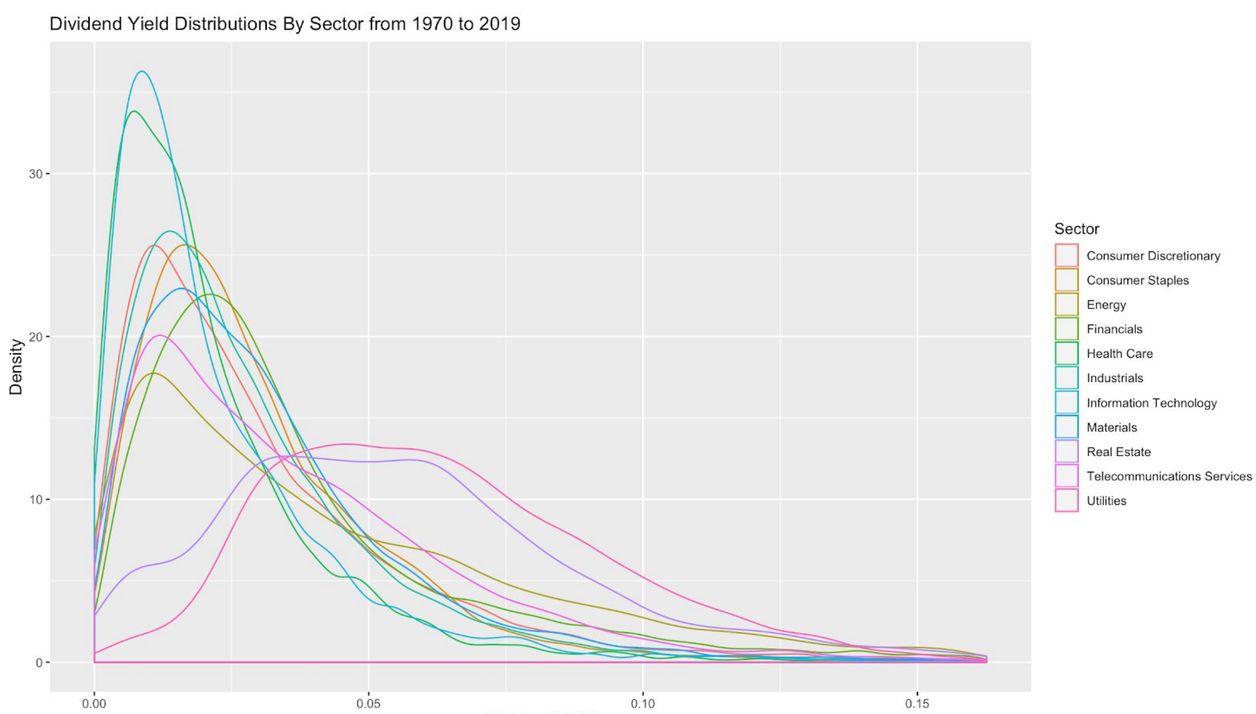


Table 2. Statistics for Dividend Yield

	Sector	Mean	Minimum	Q1	Median	Q3	Maximum	Standard Deviation
1	Energy	3.093	0	0.924	1.349	2.234	2348	26.455
2	Materials	3.614	0	1.474	2.047	2.95	1110	12.852
3	Industrials	2.649	0	1.368	1.961	2.817	664	6.585
4	Consumer Discretionary	2.764	0	1.259	1.913	2.842	1719.25	15.407
5	ConsumerStaples	2.479	0	1.26	1.733	2.542	587.333	6.804
6	Health Care	5.305	0	1.735	2.952	5.753	812	11.025
7	Financials	5.171	0	1.193	1.82	3.304	1354.333	30.933
8	Information Technology	3.631	0	1.638	2.502	4.024	4928.571	25.689
9	Telecommunications Services	2.179	0	0.846	1.286	2.192	411.8	6.955
10	Utilities	1.195	0	0.699	0.913	1.231	141.2	2.592
11	Real Estate	3.5	0.014	1.039	1.569	2.479	206.994	11.5

Current Ratio

Current ratio represents current assets over current liabilities, serving as a measure of liquidity that signals whether a company has enough resources to pay off short term debts.

Figure 25. Distribution of Current Ratio

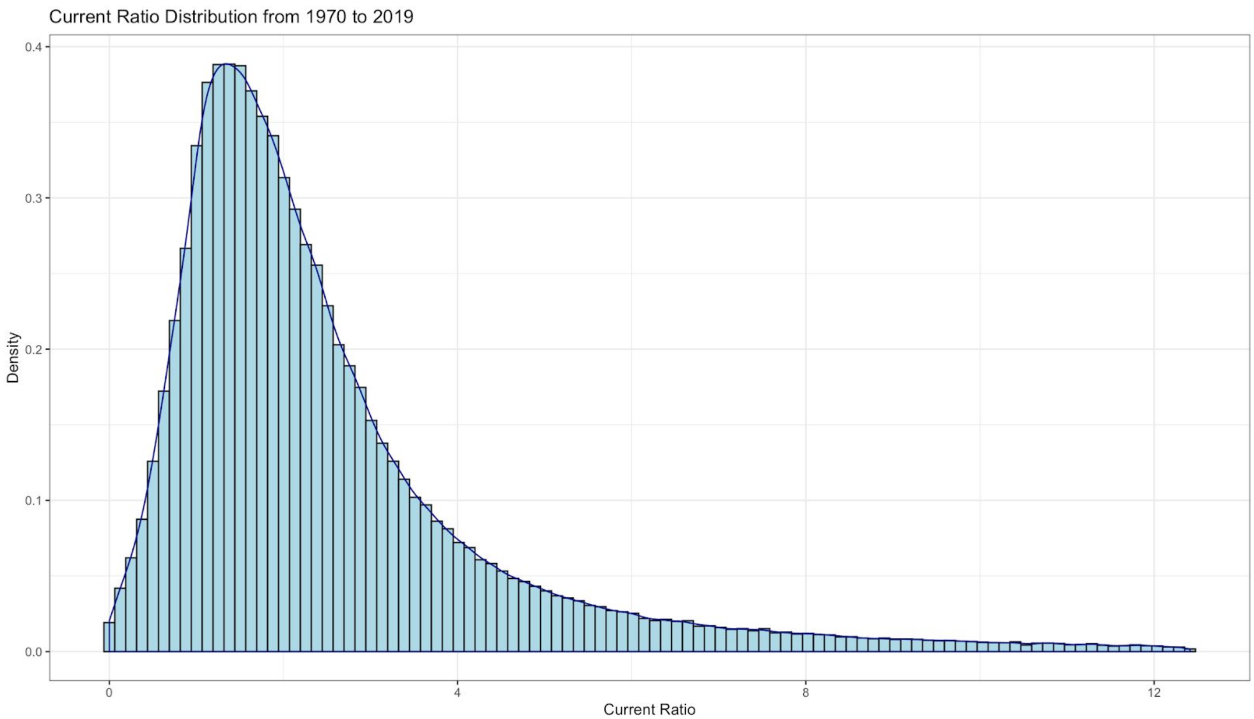
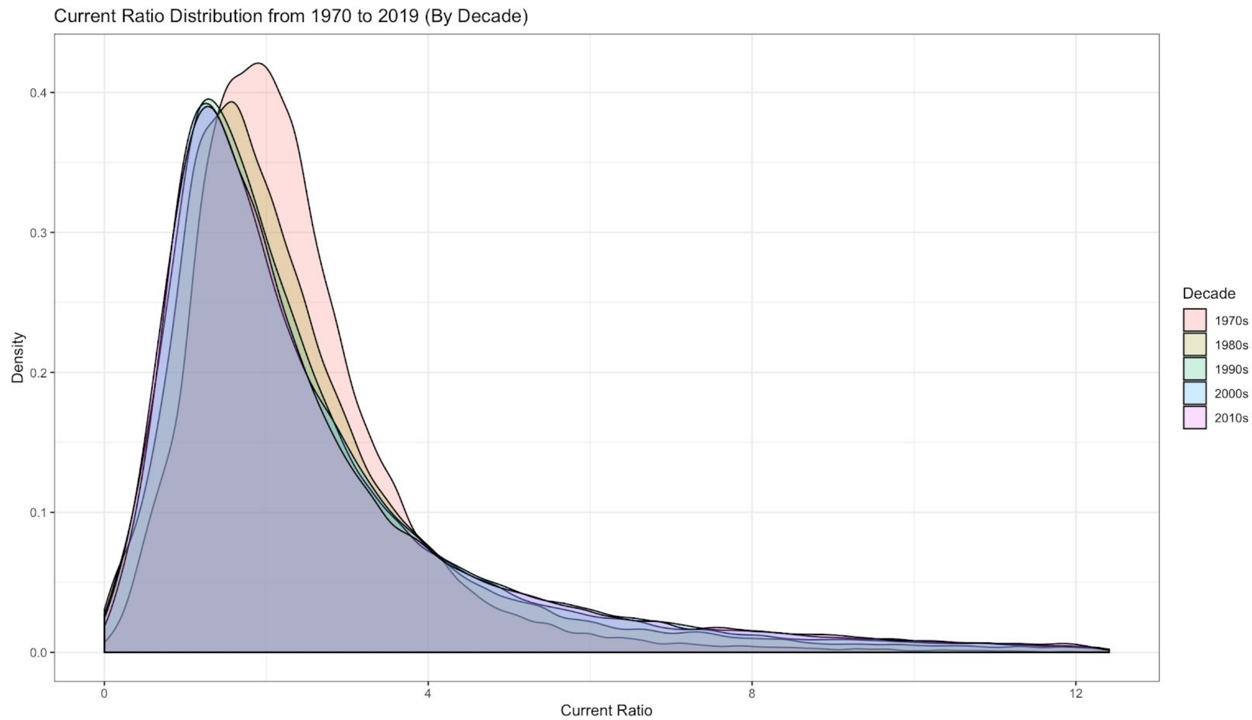


Figure Set 26. Distribution of Current Ratio by Time



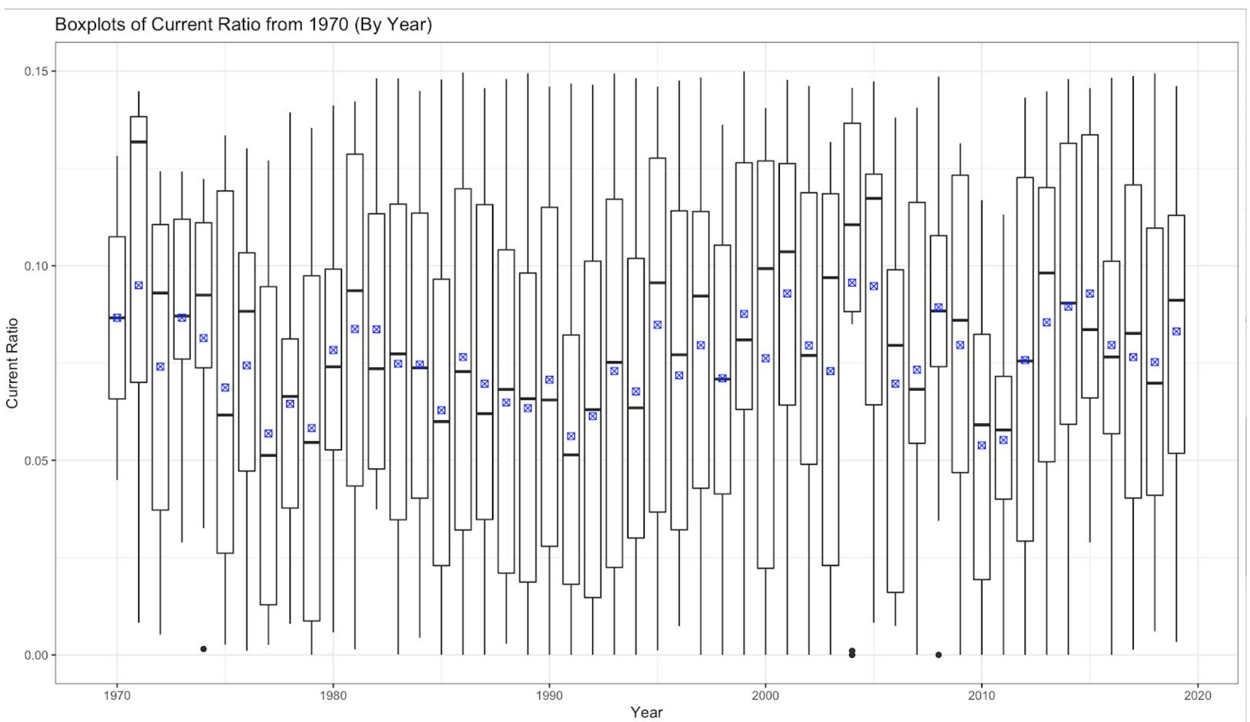
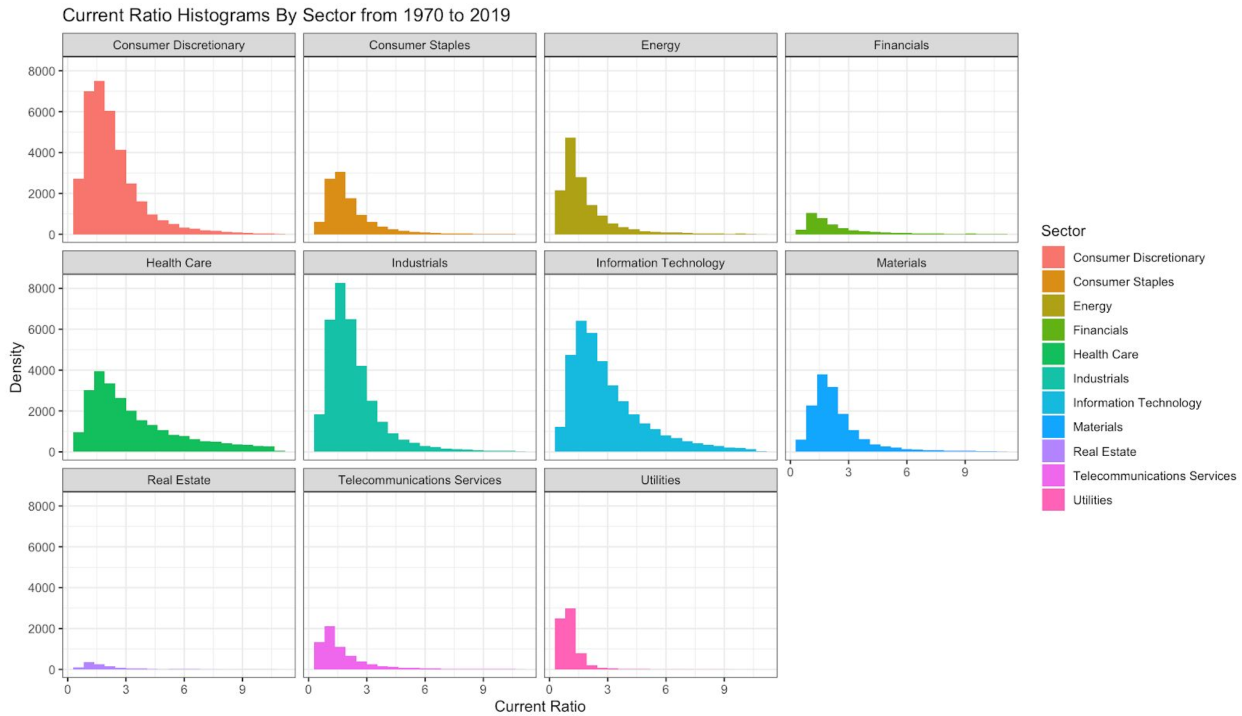


Figure Set 27. Distribution of Current Ratio by Sector



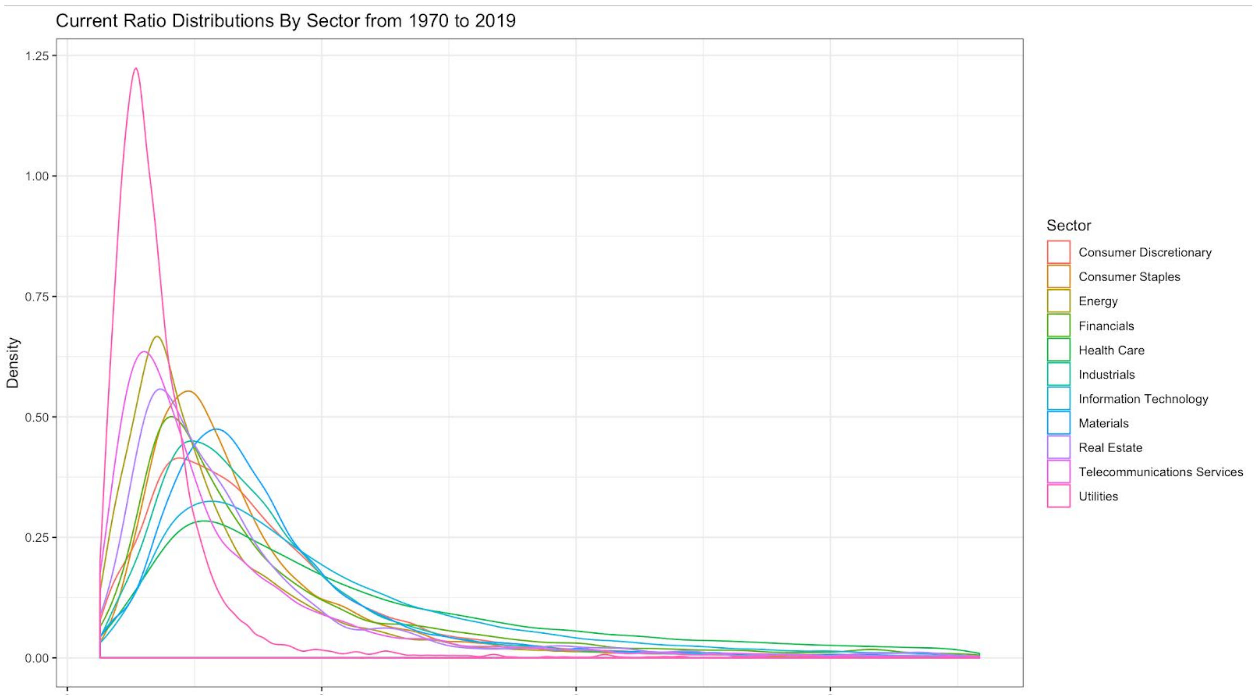


Table 3. Statistics for ROA

	Sector	Mean	Minimum	Q1	Median	Q3	Maximum	Standard Deviation
1	Energy	3.093	0	0.924	1.349	2.234	2348	26.455
2	Materials	3.614	0	1.474	2.047	2.95	1110	12.852
3	Industrials	2.649	0	1.368	1.961	2.817	664	6.585
4	Consumer Discretionary	2.764	0	1.259	1.913	2.842	1719.25	15.407
5	ConsumerStaples	2.479	0	1.26	1.733	2.542	587.333	6.804
6	Health Care	5.305	0	1.735	2.952	5.753	812	11.025
7	Financials	5.171	0	1.193	1.82	3.304	1354.333	30.933
8	Information Technology	3.631	0	1.638	2.502	4.024	4928.571	25.689
9	Telecommunications Services	2.179	0	0.846	1.286	2.192	411.8	6.955
10	Utilities	1.195	0	0.699	0.913	1.231	141.2	2.592
11	Real Estate	3.5	0.014	1.039	1.569	2.479	206.994	11.5

Leverage

Leverage is debt used to finance assets. This serves to boost returns, but at a greater downside risk. Looking at leverage tells investors a generalized sense of both potential returns and potential risk levels.

Figure 28. Distribution of Leverage

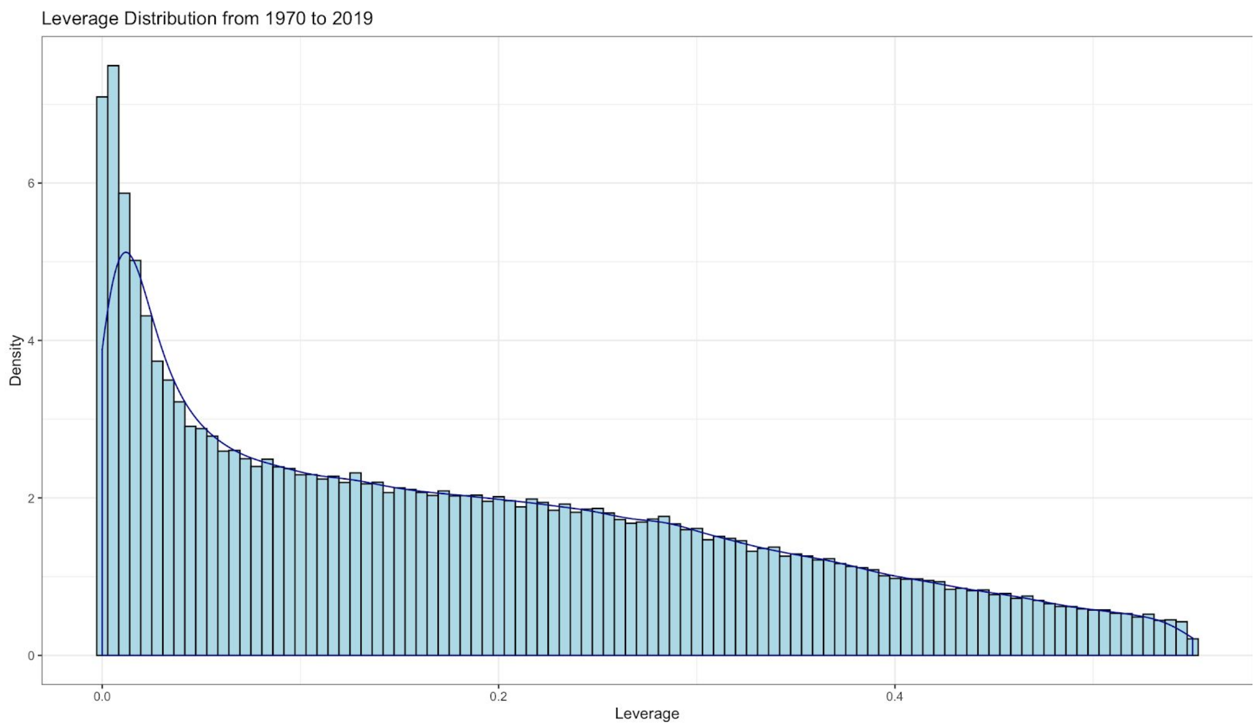
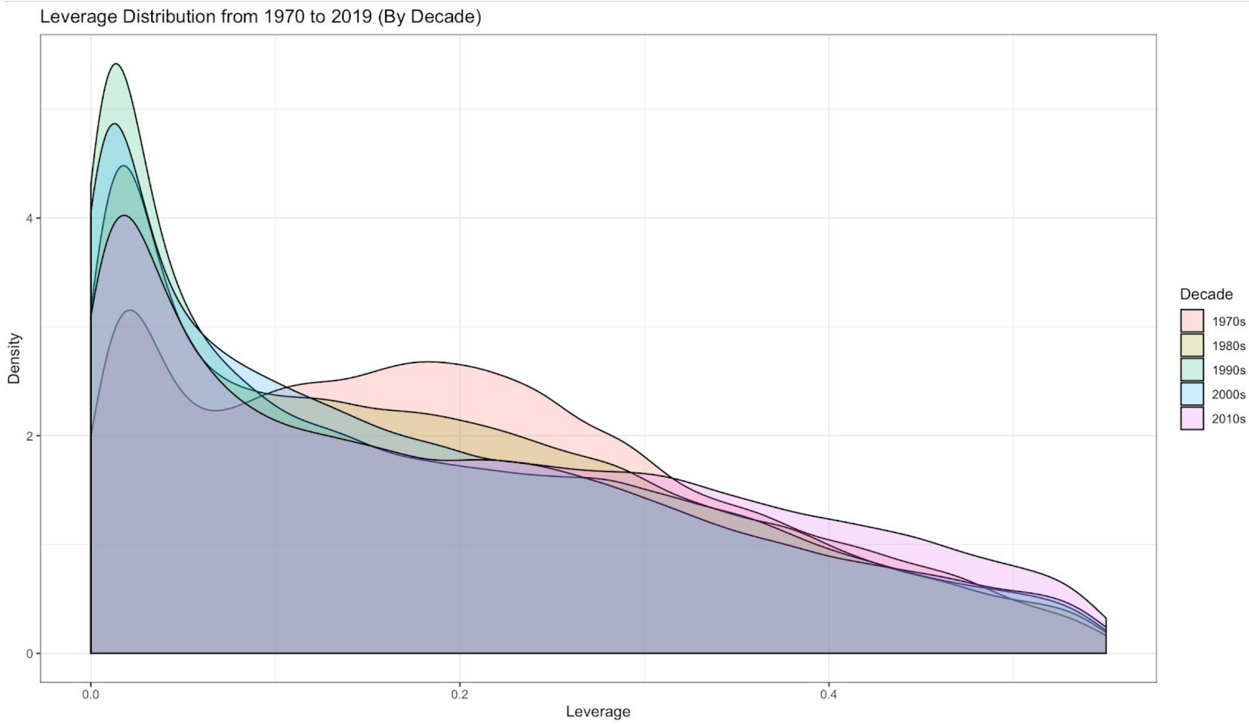


Figure Set 29. Distribution of Leverage by Time



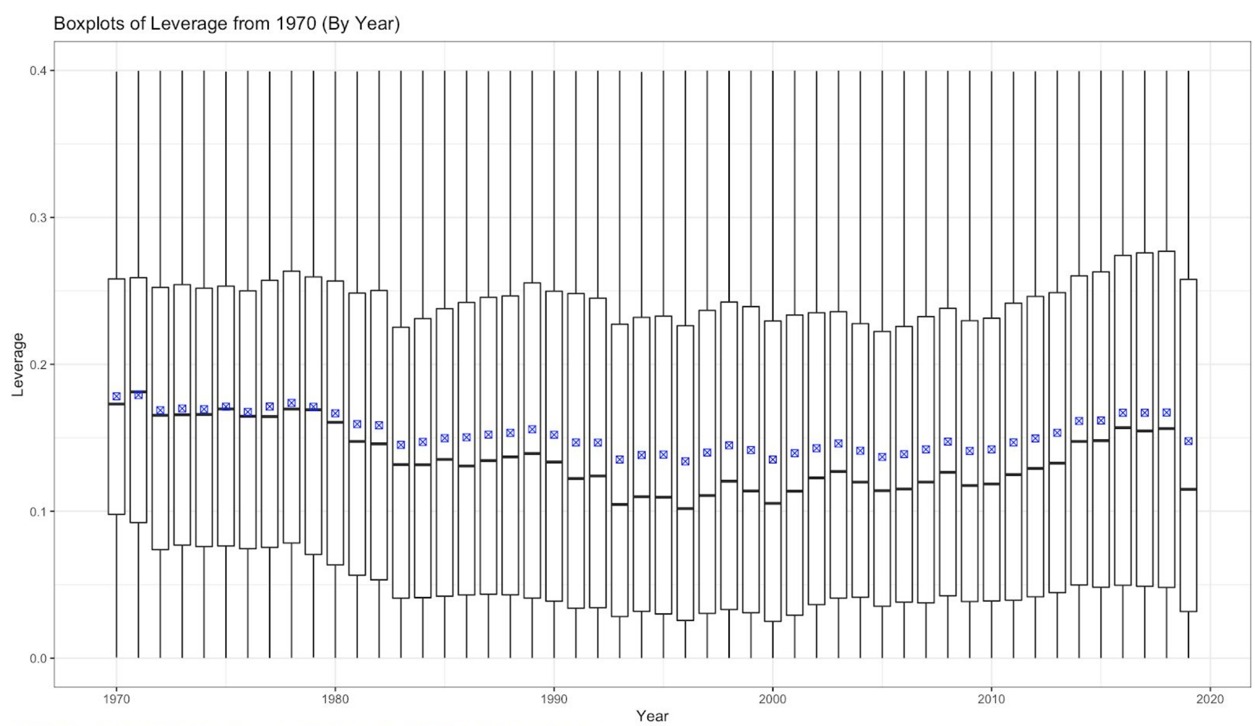
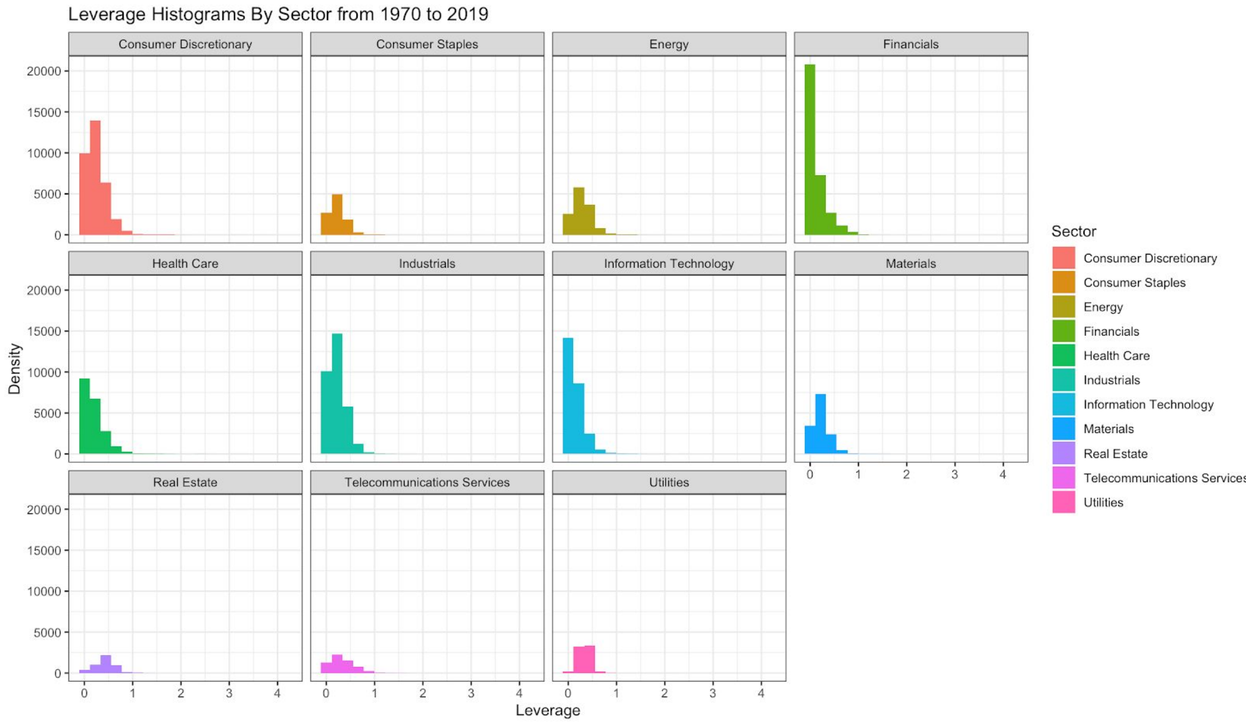
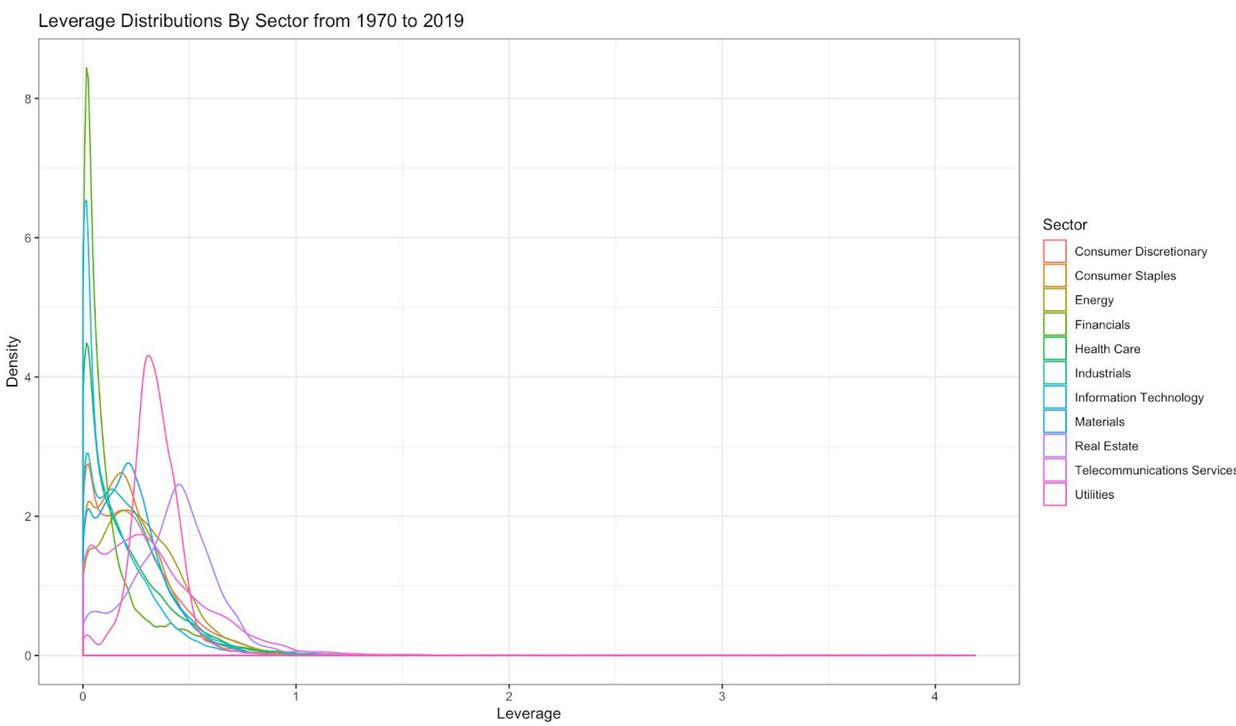


Figure Set 30. Distribution of Leverage by Sector


Leverage Histograms By Sector from 1970 to 2019
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care
Industrials
Information Technology
Materials
Real Estate
Telecommunications Services
Utilities
Density
Leverage
Sector


Leverage Distributions By Sector from 1970 to 2019
Density
Leverage
Sector
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care
Industrials
Information Technology
Materials
Real Estate
Telecommunications Services
Utilities

Turnover

Turnover is the ratio of the value of a company's revenues divided by the value of its assets. It can be used to understand a company's efficiency in using its assets to generate revenues.

Figure 31. Distribution of Turnover

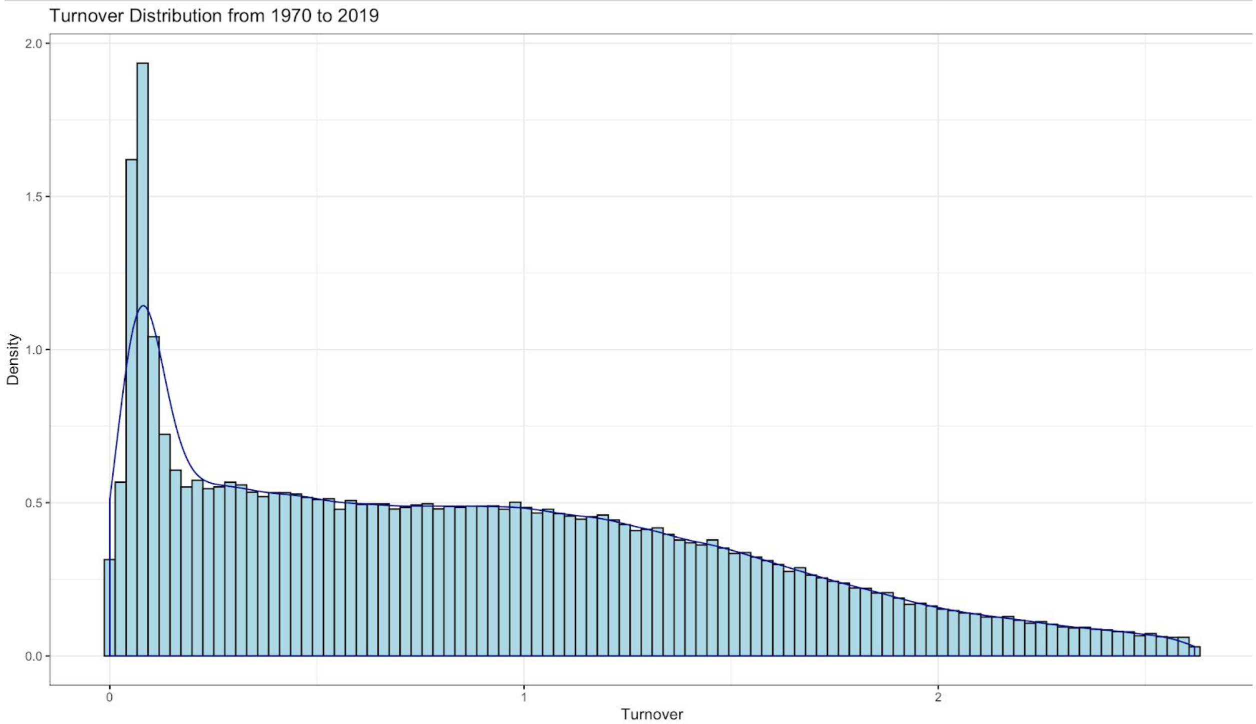
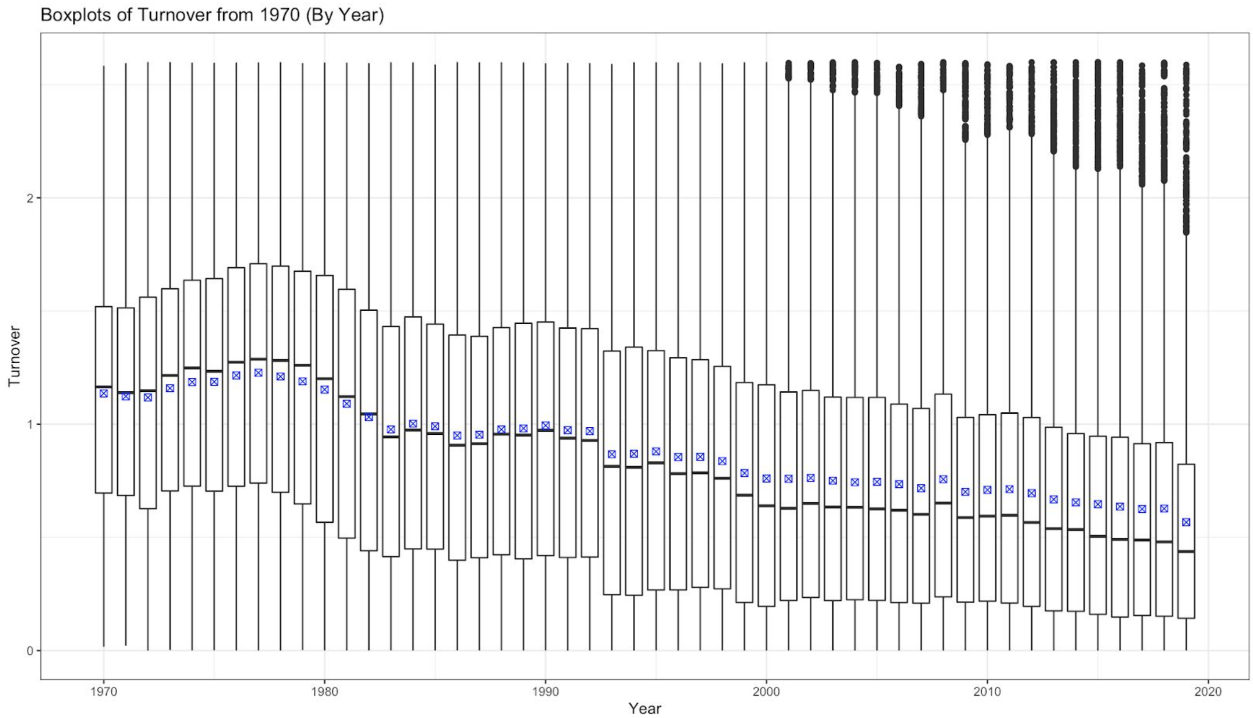


Figure Set 32. Distribution of Turnover by Time



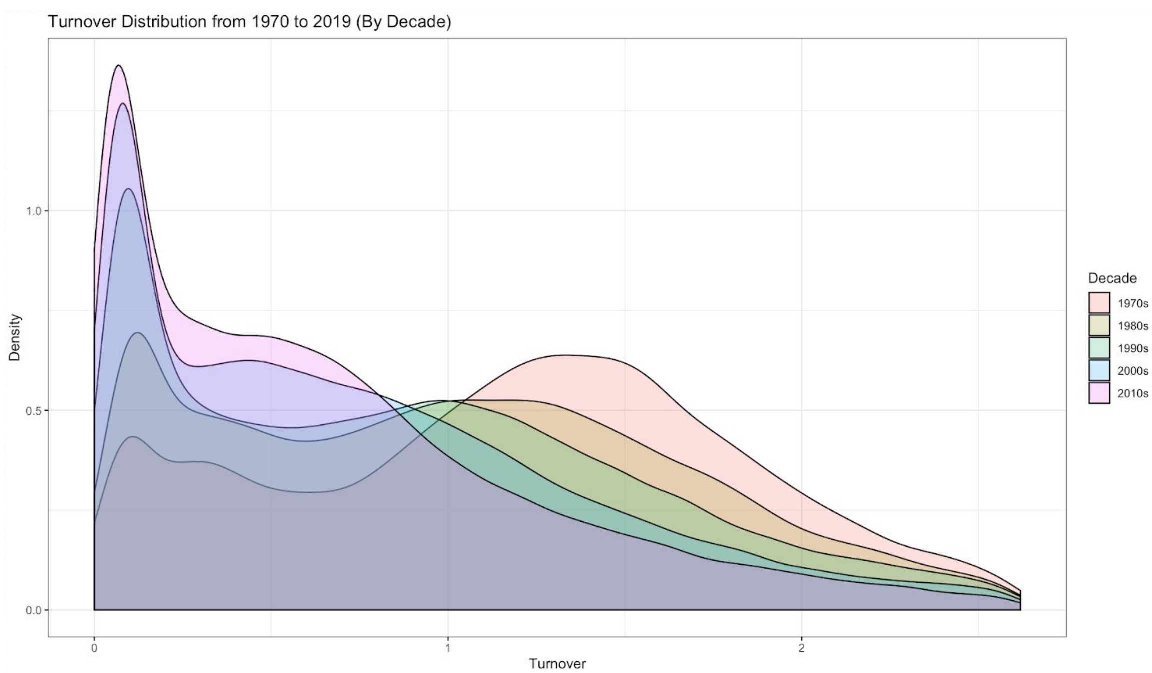
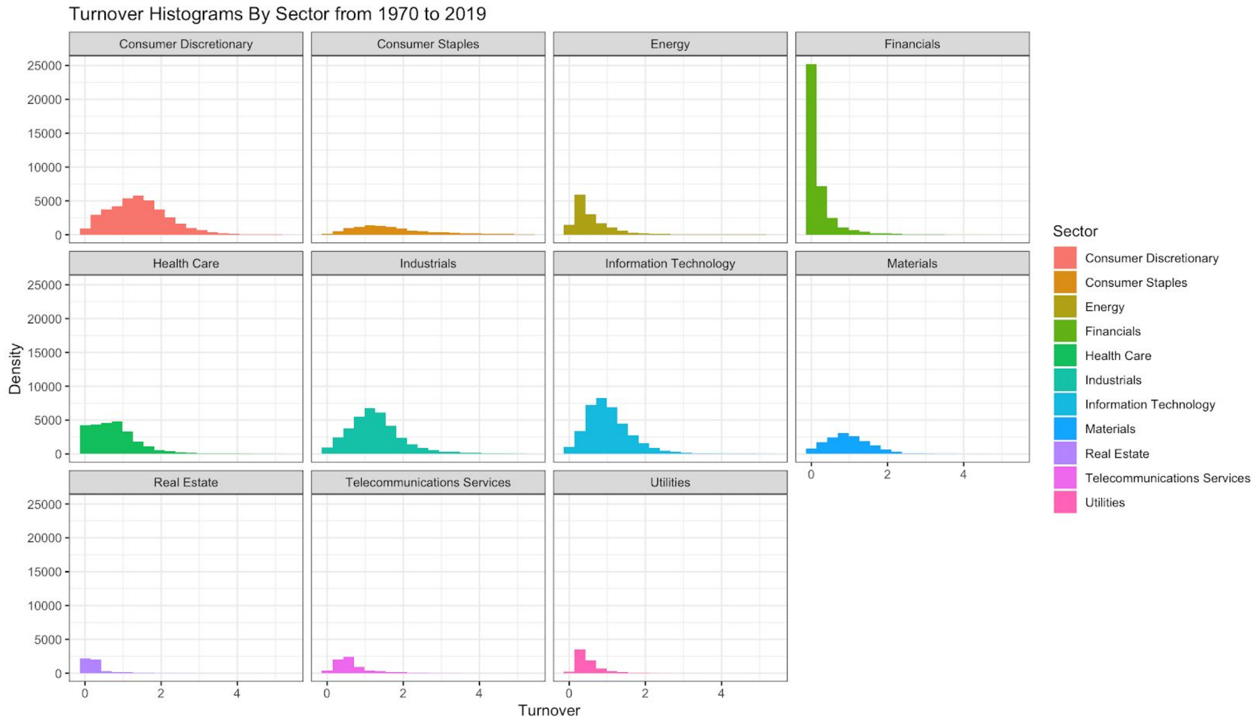
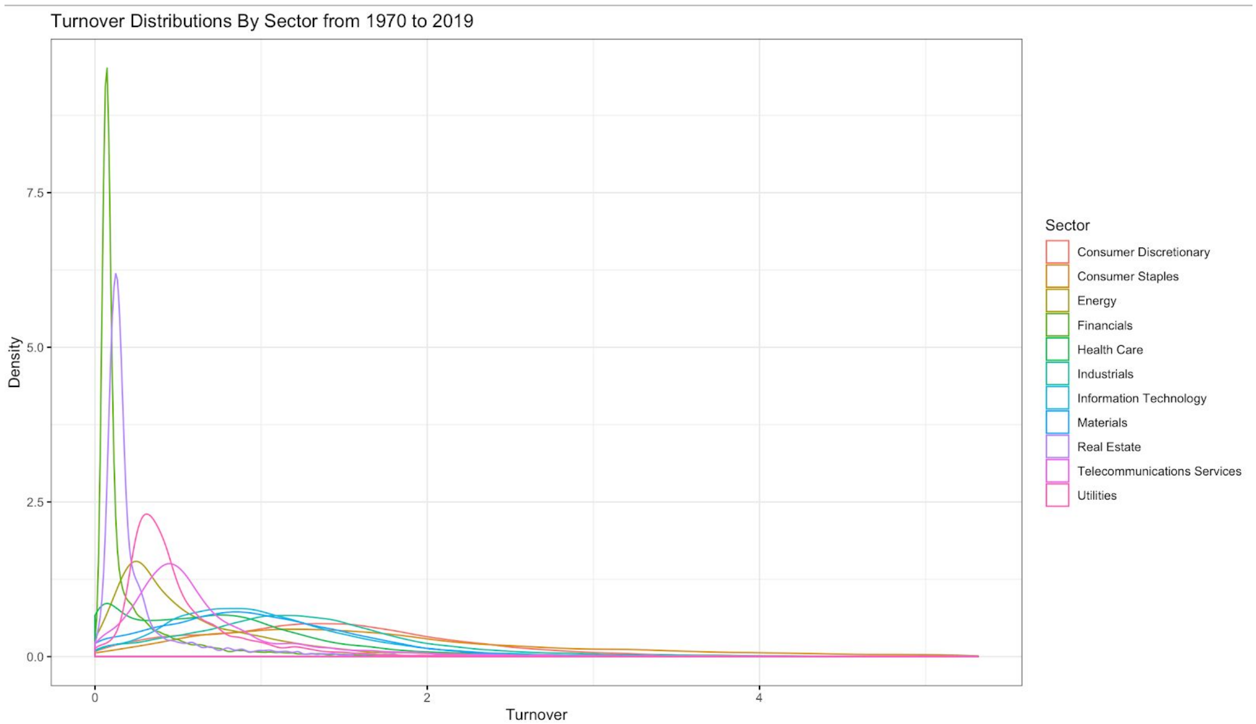


Figure Set 33. Distribution of Turnover by Sector


Turnover Histograms By Sector from 1970 to 2019
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care
Industrials
Information Technology
Materials
Real Estate
Telecommunications Services
Utilities
Density
Turnover
Sector


Turnover Distributions By Sector from 1970 to 2019
Density
Turnover
Sector
Consumer Discretionary
Consumer Staples
Energy
Financials
Health Care
Industrials
Information Technology
Materials
Real Estate
Telecommunications Services
Utilities

Return on Assets

Return on assets (ROA) is defined as the operating income divided by the total assets. This shows the company's efficiency of asset utilization to generate returns.

Figure 34. Distribution of ROA

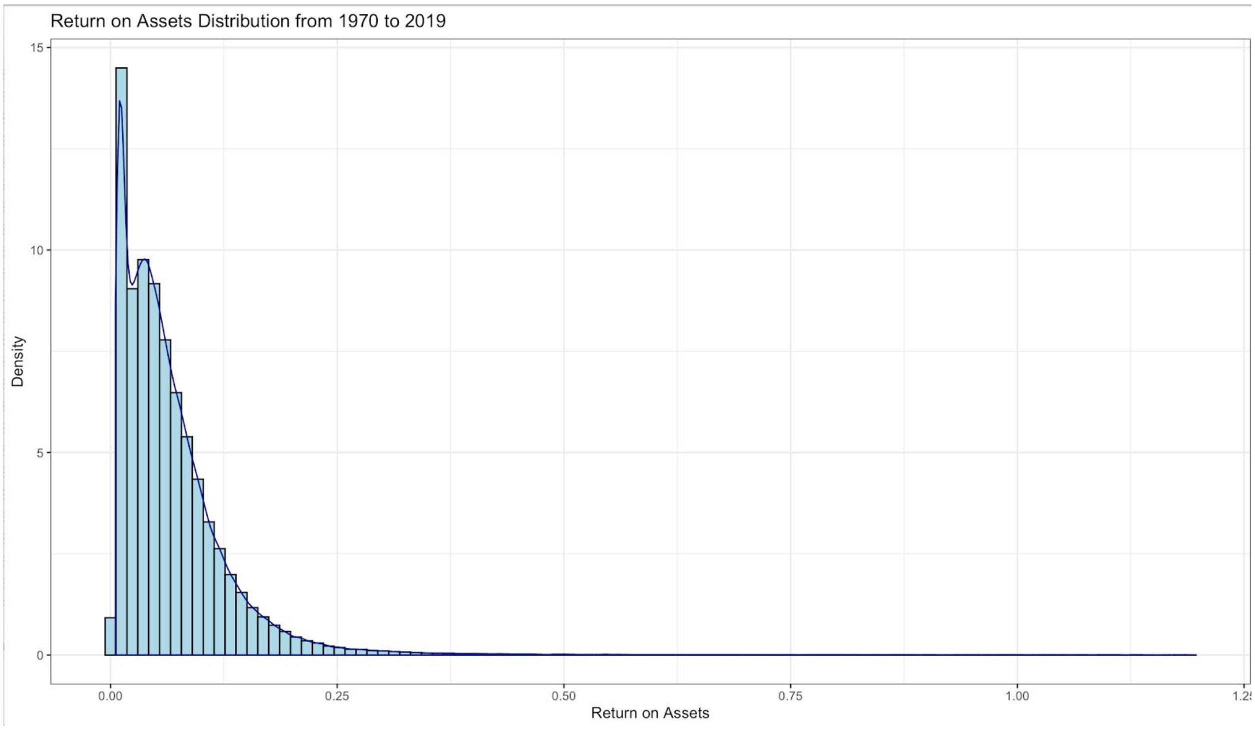
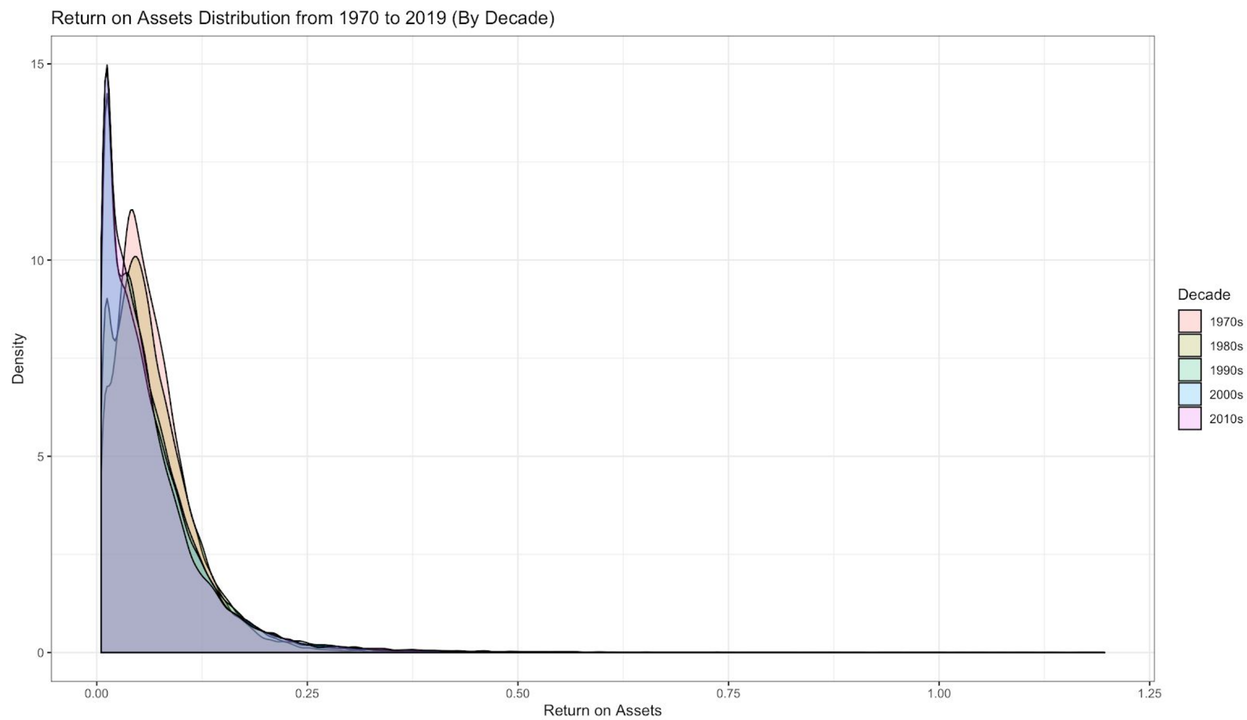


Figure Set 35. Distribution of ROA by Time



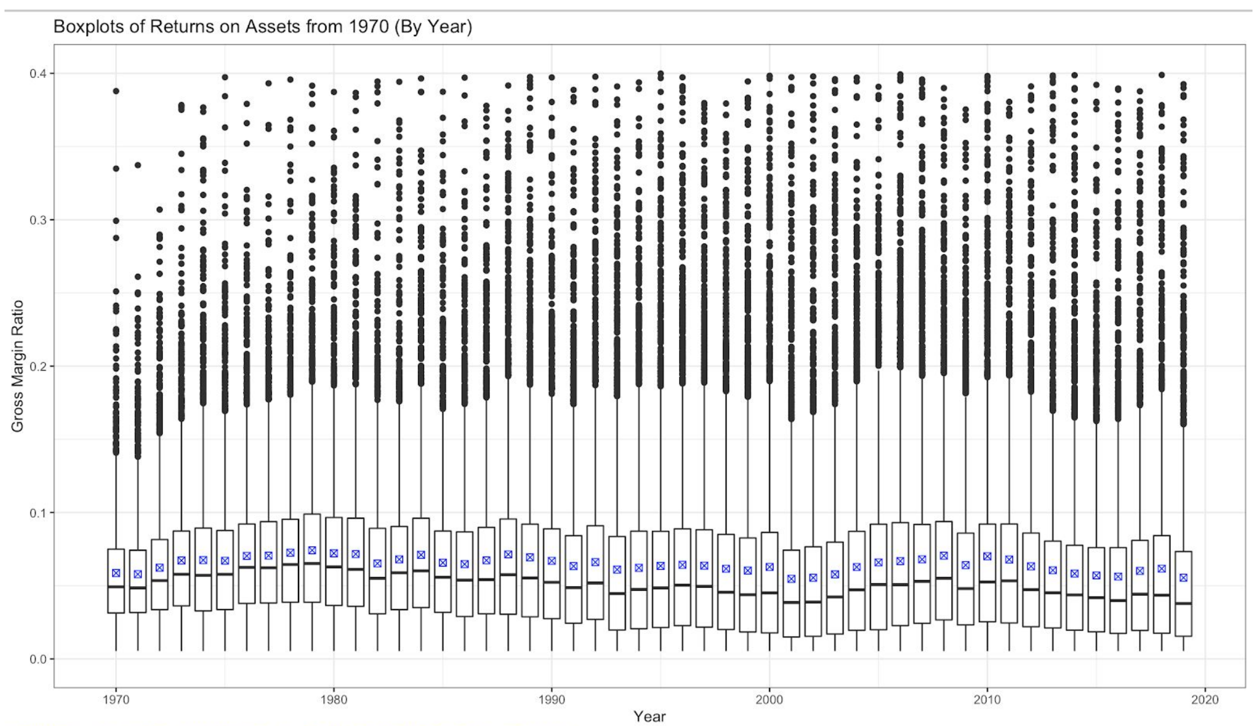
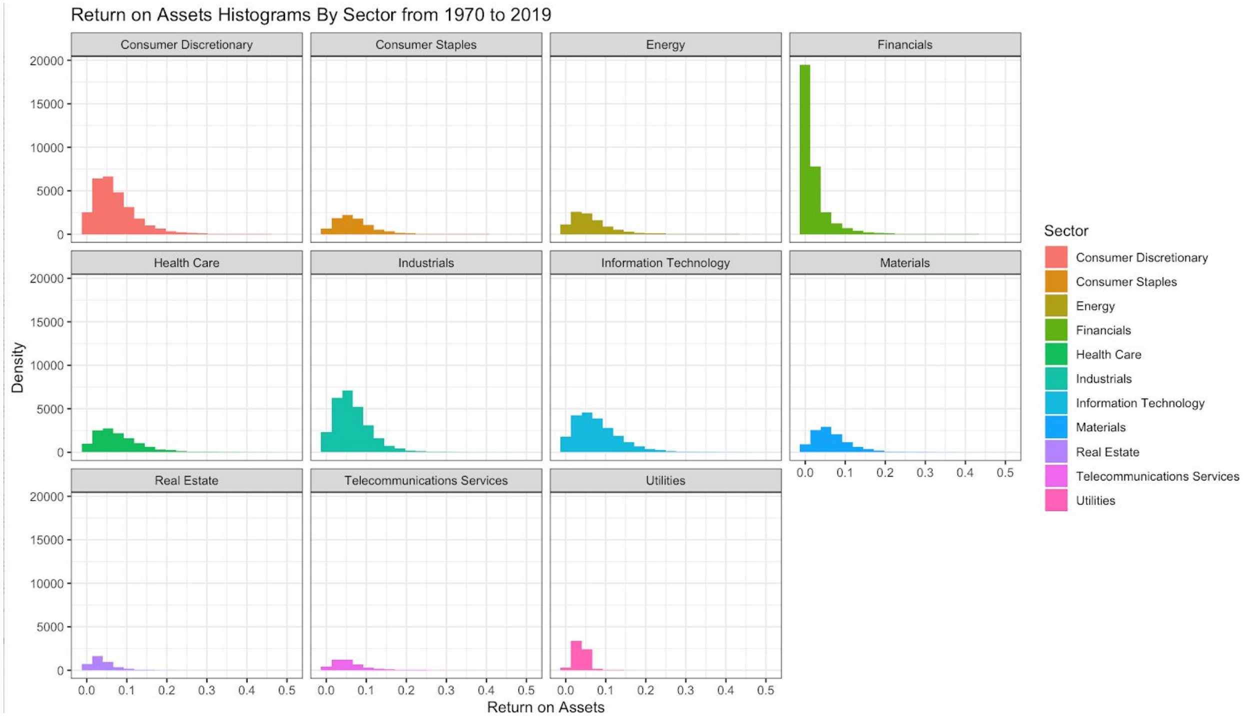
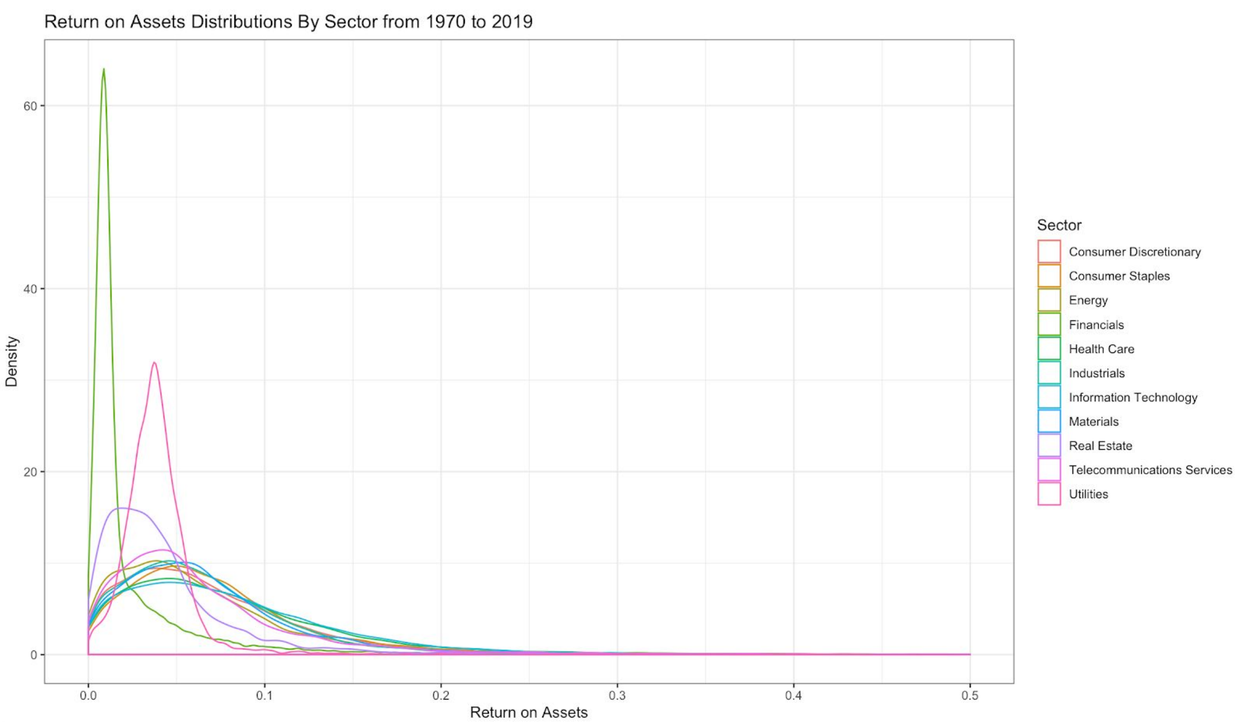


Figure Set 36. Distribution of ROA by Sector





Cash Flow from Operations over Total Assets

Revenue generated from operations divided by total assets, showing the efficiency and state of asset utilization to generate cash flows.

Table 4. Statistics for CF

	Sector	Mean	Minimum	Q1	Median	Q3	Maximum	StandardDeviation
1	Energy	0.2	-150.125	0.03	0.086	0.138	226.311	4.406
2	Materials	0.029	-107.929	0.024	0.081	0.121	6.671	0.989
3	Industrials	0.027	-218	0.036	0.079	0.119	24.587	1.483
4	Consumer Discretionary	0.023	-172.69	0.026	0.075	0.121	2.231	1.147
5	ConsumerStaples	0.054	-11.627	0.049	0.089	0.127	1.388	0.257
6	Health Care	-0.217	-99.855	-0.331	0.003	0.095	8.261	0.945
7	Financials	0.008	-98.903	0.007	0.012	0.028	5.277	0.606
8	Information Technology	-0.054	-62.198	-0.058	0.061	0.119	3.562	0.651
9	Telecommunications Services	0.004	-49.181	-0.002	0.077	0.127	2.951	0.805
10	Utilities	0.057	-6.177	0.053	0.067	0.079	0.861	0.124
11	Real Estate	0.031	-3.148	0.011	0.043	0.079	1.398	0.176

Conclusion

In this project, we looked at the distribution of the fundamental indicators of the companies' performance. The Price Earning and Price of Book ratio have an important measure on whether a company is worth investing in. After looking at the distributions, there is a basic understanding of how the measures are distributed historically. The dividend yield and earnings per share is important because it helps to evaluate the company's financial performance/well-being. The financial industry usually has more firms with a good P/E ratio. However, in the Health Care sector, there is a tendency to have a lower EPS and P/E ratio. The IT industry has a high EPS with a lower P/E ratio. The sector has more tangible assets lower of P/B ratio.

We also looked at the temporal distribution where it allowed us to look for historical trends for variables within each industry. Many of the variables showed very similar trends in distribution over the past three decades. The price of earnings ratios shows the upward trend. This suggests the market expectation for companies in recent years is increasing. The Earnings Per Share and Net Income have increased over the years. Inflation is also needed to be taken into consideration because there is a discount we wanted to make interpretation of the trend.

Lastly, we explored the effect of extraordinary items when calculating Earnings Per Share. There are Extraordinary items that could change incomes in either direction. The change in percentage for either direction is really close. After exploring the relationship between book values and community Equity Liquidation Values. There is a perfect linear relationship where it makes sense since the book value has assets valued at fair-market price. This is the price of the asset if it was sold today.

Appendix - R Code

```
library(ggplot2)
mkt1925df <- read.csv("special.csv")

mkt1925df$mktcap <-mkt1925df$PRC*mkt1925df$SHROUT*1000/1000000
mkt1925df$year
<-as.numeric(substring(as.character(mkt1925df$date), 7, 10))

ggplot(data = mkt1925df, aes(x = year, y = mktcap,group = year))
+
  geom_boxplot() + stat_summary(fun.y=mean, colour="blue",
geom="point", shape=7,size=2) +
  ylim(0, 800) + xlab("Year") +  ylab("Market Capitalization
($M)") +
  labs(title = "Boxplots of Market Capitalization from 1925 (By
Year)") +
  theme(axis.title.x = element_text(size=15), axis.title.y
=element_text(size=15)) +
  theme_bw()


countYear <- 2019-1925+1
over50 <- numeric(countYear)
overMean <- numeric(countYear)
overMedian <- numeric(countYear)
propOver50 <- numeric(countYear)
mktcapMeans <- numeric(countYear)
mktcapMedians <- numeric(countYear)
propMeans <- numeric(countYear)
propMedians <- numeric(countYear)

k <- 1
for(val in seq(1925, 2019, 1)){
  temp <- mkt1925df[which(mkt1925df$year == val),]
  mktcapMeans[k] <- mean(temp$mktcap, na.rm = T)
  propMeans[k]
<-nrow(temp[temp$mktcap>mktcapMeans[k],])/nrow(temp)
```

```
  mktcapMedians[k] <- median(temp$mktcap, na.rm =TRUE)
  propMedians[k]
<-nrow(temp[temp$mktcap>mktcapMedians[k],])/nrow(temp)
  over50[k] <- nrow(temp[temp$mktcap>50,])
  overMean[k] <-nrow(temp[temp$mktcap>mktcapMeans[k],])
  overMedian[k] <-nrow(temp[temp$mktcap>mktcapMedians[k],])
  propOver50[k] <-nrow(temp[temp$mktcap>50,])/nrow(temp)
  k <- k+1
}

par(mfrow=c(3,2))
plot(seq(1925, 2019, 1),
     propMeans, type = "l",
     col = "darkblue", xlab = "Year",
     ylab = "Proportion",
     main ="Porportion over Mean Market Capitalization (1925 -
2019)", ylim =c(.1,.3))

plot(seq(1925, 2019, 1),
     overMean/12, type = "l",
     col = "darkblue", xlab = "Year",
     ylab = "Proportion",
     main ="Count over Mean Market Capitalization (1925 -
2019)")

plot(seq(1925, 2019, 1),
     rep(.5, 95), type = "l",
     col = "darkblue", xlab = "Year",
     ylab = "Proportion",
     main ="Porportion over Median Market Capitalization (1925 -
2019)", ylim = c(0,.75))

plot(seq(1925, 2019, 1),
     overMedian/12, type = "l",
     col = "darkblue", xlab = "Year",
     ylab = "Proportion",
     main ="Count over Median Market Capitalization (1925 -
2019))")
```

```
plot(seq(1925, 2019, 1),
     propOver50, type = "l",
     col = "darkblue", xlab = "Year",
     ylab = "Proportion",
     main ="Porportion over $50M Median Market Capitalization
(1925 - 2019)", ylim = c(0,.75))

plot(seq(1925, 2019, 1),
     overMedian/12, type = "l",
     col = "darkblue", xlab = "Year",
     ylab = "Proportion", main ="Count over $50M Market
Capitalization (1925 - 2019)")
mtext("Proportion of Market Capitalizations above Yearly Mean
and Median (1925 - 2019)", side = 1, line = -1, outer = TRUE)

origin_data <-read.csv("origin_data_mini1.csv")


#mktcap - Market Capitalization (csho*prcc_f)
origin_data$mktcap <- origin_data$csho*origin_data$prcc_f
#bookvalue - Book Value (csho*bkvlps)
origin_data$bv <- origin_data$bkvlps*origin_data$csho
#pb - Price to Book Ratio (mktcap/bookvalue)
origin_data$pb <- origin_data$mktcap/origin_data$bv
#pe - Price to Earnings Ratio (mktcap/ib) (O'Shaughnessy UsesIB)
origin_data$pe <- origin_data$mktcap/origin_data$ib
#cr - Current Ratio (act/lct)
origin_data$cr <- origin_data$act/origin_data$lct
#leverage - Leverage (dltt/at)
origin_data$leverage <- origin_data$dltt/origin_data$at
#turnover - Turnover (revt/at)
origin_data$turnover <- origin_data$sale/origin_data$at
#gross_margin - Gross Margin Ratio (gp/sale)
origin_data$grossMarginRatio <- origin_data$gp/origin_data$sale
#roa <- Return on Assets (ni/ta)
origin_data$ROA <- origin_data$ni/origin_data$at
#oshaughnessy_cfo (ib+db)
origin_data$oshaughnessy_cfo <- origin_data$ib+origin_data$dp
```

```
#cfo_assetsratio <- Cash Flow from Operations to Total
AssetsRatio ()
origin_data$cfo_assetsratio <-
origin_data$oshaughnessy_cfo/origin_data$at
#divyield <- Dividend Yield (dvpsp_c/prcc_c)
origin_data$divYield <- origin_data$dvpsp_c/origin_data$prcc_c

#Decade
origin_data$Decade <- 0
origin_data$Decade[origin_data$fyear >= 1970 & origin_data$fyear
< 1980] <- "1970s"
origin_data$Decade[origin_data$fyear >= 1980 & origin_data$fyear
< 1990] <- "1980s"
origin_data$Decade[origin_data$fyear >= 1990 & origin_data$fyear
< 2000] <- "1990s"
origin_data$Decade[origin_data$fyear >= 2000 & origin_data$fyear
< 2010] <- "2000s"
origin_data$Decade[origin_data$fyear >= 2010] <- "2010s"

#Sector
origin_data$Sector[origin_data$gsector == 10] <- "Energy"
origin_data$Sector[origin_data$gsector == 15] <- "Materials"
origin_data$Sector[origin_data$gsector == 20] <- "Industrials"
origin_data$Sector[origin_data$gsector == 25] <- "Consumer
Discretionary"
origin_data$Sector[origin_data$gsector == 30] <- "Consumer
Staples"
origin_data$Sector[origin_data$gsector == 35] <- "Health Care"
origin_data$Sector[origin_data$gsector == 40] <- "Financials"
origin_data$Sector[origin_data$gsector == 45] <- "Information
Technology"
origin_data$Sector[origin_data$gsector == 50] <-
"Telecommunications Services"
origin_data$Sector[origin_data$gsector == 55] <- "Utilities"
origin_data$Sector[origin_data$gsector == 60] <- "Real Estate"

data <- origin_data[ origin_data$fyear >= 1970 &
origin_data$exchg != 0 & origin_data$exchg != 1 &
origin_data$exchg != 3,]
```

```
#market cap
df <- data[!is.na(data$mktcap) & data$mktcap<=
quantile(data$mktcap, .95, na.rm=TRUE),]

ggplot(data = df, aes(x = mktcap)) +
  geom_histogram(aes(x = mktcap, y = ..density..), bins =
100,col = "black", fill = "lightblue") +
  geom_density(aes(x = mktcap), col = "darkblue") +
xlab("Market Capitalization") +
  ylab("Density") +  labs(title = "Market Capitalization
Distribution (1970 to 2019)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

#bookvalue
df <- data[!is.na(data$bookvalue) &!is.na(data$Sector),]
#Statistics Summary Table
Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$bv[df$gsector == val], na.rm= TRUE)
  Minimum[i] <- min(df$bv[df$gsector == val],na.rm = TRUE)
  Q1[i] <- quantile(df$bv[df$gsector == val],.25, na.rm = TRUE)
  Median[i] <- median(df$bv[df$gsector == val],na.rm = TRUE)
  Q3[i] <- quantile(df$bv[df$gsector == val],.75, na.rm = TRUE)
  Maximum[i] <- max(df$bv[df$gsector == val],na.rm = TRUE)
  StandardDeviation[i] <- sd(df$bv[df$gsector == val],na.rm =
TRUE)
  i <- i + 1}
```

```
sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3, Maximum,
StandardDeviation)
write.csv(sectors,"book_table")

df <- data[!is.na(data$bv) &data$bv <= quantile(origin_data$bv,
.925,na.rm=TRUE) & data$bv >=
quantile(origin_data$bookvalue,.005, na.rm=TRUE),]
ggplot(data = df, aes(x = bv)) +  geom_histogram(aes(x = bv, y =
..density..), bins =100, col = "black", fill = "lightblue") +
  geom_density(aes(x = bv), col = "darkblue") +  xlab("Book
Value ($M)") +  ylab("Density") +
  labs(title = "Book Value Distribution (1970 to 2019)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()
#decade
ggplot(data = df, aes(x = bv, fill = Decade)) +
geom_density(aes(x = bv), alpha = .25) +  xlab("Book Value
($M)") +  ylab("Density") +  labs(title = "Book Value
Distribution from 1970 to 2019 (By Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

#Boxplot for Each Year
ggplot(data = df, aes(x = fyear, y = bv, group =fyear)) +
geom_boxplot() +
  stat_summary(fun.y=mean, colour="blue", geom="point",
shape=7,size=2) +
  ylim(-50, 225) +  xlab("Year") +  ylab("Book Value ($M)") +
  labs(title = "Boxplots of Book Value from 1970 (By Year)") +
  theme(axis.title.x = element_text(size=15), axis.title.y
=element_text(size=15))+
  theme_bw()
```

```
#bv by sector
data_i <- data[!is.na(data$bv) &!is.na(data$Sector) & data$bv
<=quantile(origin_data$bv, .925, na.rm=TRUE) & data$bv>=
quantile(origin_data$bv, .005, na.rm=TRUE),]
ggplot(data = df, aes(x = bv, fill = Sector)) +
geom_histogram(aes(x = bv), bins = 20) +
  facet_wrap(~ Sector) +  xlab("Book Value ($M)") +
ylab("Density") +
  labs(title = "Book Value Histograms By Sector from 1970 to
2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = bv, color = Sector)) +  geom_density()
+  xlab("Book Value ($M)") +
  ylab("Density") +  labs(title = "Book Value Distributions By
Sector from 1970 to 2019") +  theme(axis.title.x =
element_text(size=12), axis.title.y =element_text(size=12)) +
  theme_bw()


###revt
df <- data[!is.na(data$revt) &!is.na(data$Sector),]
#Statistics Summary Table
Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$revt[df$gsector == val], na.rm =TRUE)
  Minimum[i] <- min(df$revt[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$revt[df$gsector == val], .25,na.rm =
TRUE)
```

```
  Median[i] <- median(df$revt[df$gsector == val], na.rm= TRUE)
  Q3[i] <- quantile(df$revt[df$gsector == val], .75,na.rm =
TRUE)
  Maximum[i] <- max(df$revt[df$gsector == val], na.rm =TRUE)
  StandardDeviation[i] <- sd(df$revt[df$gsector == val], na.rm
=TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3, Maximum,
StandardDeviation)
write.csv(sectors,"revt_table")

df <- data[!is.na(data$revt) & data$revt
<=quantile(origin_data$revt, .925, na.rm=TRUE) & data$revt
>=quantile(origin_data$revt, .005, na.rm=TRUE),]

ggplot(data = df, aes(x = revt)) +  geom_histogram(aes(x = revt,
y = ..density..), bins = 100, col= "black", fill = "lightblue")
+  geom_density(aes(x = revt), col = "darkblue") +
xlab("Revenue ($M)") +  ylab("Density") +  labs(title = "Revenue
Distribution (1970 to 2019)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = revt, fill = Decade)) +
geom_density(aes(x = revt), alpha = .25) +  xlab("Revenue ($M)")
+  ylab("Density") +  labs(title = "Revenue Distribution from
1970 to 2019 (By Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = fyear, y = revt, group = fyear)) +
geom_boxplot() +  stat_summary(fun.y=mean, colour="blue",
```

```
geom="point", shape=7,size=2) +  ylim(0, 300) +  xlab("Year") +
ylab("Revenue ($M)") +  labs(title = "Boxplots of Revenue from
1970 (By Year)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12))+
  theme_bw()

df <- data[!is.na(data$revt) &!is.na(data$Sector) & data$revt <=
quantile(origin_data$revt,.925, na.rm=TRUE) & data$revt >=
quantile(origin_data$revt, .005,na.rm=TRUE),]

ggplot(data = df, aes(x = revt, fill = Sector)) +
  geom_histogram(aes(x = revt), bins = 20) +
  facet_wrap(~ Sector) +  xlab("Revenue ($M)") +
  ylab("Density") +  labs(title = "Revenue Histograms By Sector
from 1970 to 2019") +  theme(axis.title.x =
element_text(size=12), axis.title.y =element_text(size=12)) +
theme_bw()

ggplot(data = df, aes(x = revt, color = Sector)) +
geom_density() +  xlab("Revenue ($M)") +  ylab("Density") +
labs(title = "Revenue Distributions By Sector from 1970 to
2019") +  theme(axis.title.x = element_text(size=12),
axis.title.y =element_text(size=12)) + theme_bw()


#ni
df <- data[!is.na(data$ni) &!is.na(data$Sector),]
Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$ni[df$gsector == val], na.rm =TRUE)
```

```
  Minimum[i] <- min(df$ni[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$ni[df$gsector == val], .25, na.rm= TRUE)
  Median[i] <- median(df$ni[df$gsector == val], na.rm =TRUE)
  Q3[i] <- quantile(df$ni[df$gsector == val], .75, na.rm= TRUE)
  Maximum[i] <- max(df$ni[df$gsector == val], na.rm =TRUE)
  StandardDeviation[i] <- sd(df$ni[df$gsector == val], na.rm
=TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3, Maximum,
StandardDeviation)
write.csv(sectors,"ni_table")

df <- data[!is.na(data$ni) & data$ni <=quantile(origin_data$ni,
.95, na.rm=TRUE) & data$ni >=quantile(origin_data$ni, .02,
na.rm=TRUE),]

ggplot(data = df, aes(x = ni)) +  geom_histogram(aes(x = ni, y =
..density..), bins = 100, col ="black", fill = "lightblue") +
geom_density(aes(x = ni), col = "darkblue") +  xlab("Net Income
($M)") +  ylab("Density") +  labs(title = "Net Income
Distribution (1970 to 2019)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) + theme_bw()

ggplot(data = df, aes(x = ni, fill = Decade)) +
geom_density(aes(x = ni), alpha = .25) +  xlab("Net Income
($M)") +  ylab("Density") +  labs(title = "Net Income
Distribution from 1970 to 2019 (By Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = fyear, y = ni, group = fyear)) +
geom_boxplot() +
```

```
  stat_summary(fun.y=mean, colour="blue", geom="point",
shape=7,size=2) +
  ylim(-50, 150) +  xlab("Year") +  ylab("Net Income ($M)") +
  labs(title = "Boxplots of Net Income from 1970 (By Year)") +
  theme(axis.title.x = element_text(size=15), axis.title.y
=element_text(size=15)) +
  theme_bw()

df <- data[!is.na(data$ni) &!is.na(data$Sector) & data$ni <=
quantile(origin_data$ni,.95, na.rm=TRUE) & data$ni >=
quantile(origin_data$ni, .02,na.rm=TRUE),]

ggplot(data = df, aes(x = ni, fill = Sector)) +
geom_histogram(aes(x = ni), bins = 20) +  facet_wrap(~ Sector) +
xlab("Net Income ($M)") +  ylab("Density") +  labs(title = "Net
Income Histograms By Sector from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = ni, color = Sector)) +  geom_density()
+  xlab("Net Income ($M)") +  ylab("Density") +  labs(title =
"Net Income Distributions By Sector from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

#epspx
df <- data[!is.na(data$epspx) &!is.na(data$Sector) & data$epspx
> 0,]
Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
```

```
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$epspx[df$gsector == val], na.rm =TRUE)
  Minimum[i] <- min(df$epspx[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$epspx[df$gsector == val], .25,na.rm =
TRUE)
  Median[i] <- median(df$epspx[df$gsector == val], na.rm= TRUE)
  Q3[i] <- quantile(df$epspx[df$gsector == val], .75,na.rm =
TRUE)
  Maximum[i] <- max(df$epspx[df$gsector == val], na.rm =TRUE)
  StandardDeviation[i] <- sd(df$epspx[df$gsector == val], na.rm
=TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"epspx_table")

df <- data[!is.na(data$epspx) & data$epspx
<=quantile(origin_data$epspx, .985, na.rm=TRUE) & data$epspx >
0,]

ggplot(data = df, aes(x = epspx)) +  geom_histogram(aes(x =
epspx, y = ..density..), bins = 100,col = "black", fill =
"lightblue") +  geom_density(aes(x = epspx), col = "darkblue") +
xlab("Earnings per Share ($)") +  ylab("Density") +  labs(title
= "Earnings per Share Distribution (1970 to 2019)")+
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()


ggplot(data = df, aes(x = epspx, fill = Decade)) +
geom_density(aes(x = epspx), alpha = .25) +  xlab("Earnings per
Share ($)") +  ylab("Density") +  labs(title = "Earnings per
Share Distribution from 1970 to 2019 (By Decade)") +
```

```
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = fyear, y = epspx, group = fyear))+
geom_boxplot() +
  stat_summary(fun.y=mean, colour="blue", geom="point",
shape=7,size=2) +
  ylim(0, 6) +  xlab("Year") +  ylab("Earnings per Share ($)") +
  labs(title = "Boxplots of Earnings per Share from 1970 (By
Year)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

df <- data[!is.na(data$epspx) &!is.na(data$Sector) & data$epspx
<=quantile(origin_data$epspx, .985, na.rm=TRUE) & data$epspx >
0,]

ggplot(data = df, aes(x = epspx, fill = Sector)) +
  geom_histogram(aes(x = epspx), bins = 20) +
  facet_wrap(~ Sector) +  xlab("Earnings per Share ($)") +
  ylab("Density") +
  labs(title = "Earnings per Share Histograms By Sector from
1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = epspx, color = Sector)) +
geom_density() +  xlab("Earnings per Share ($)") +
ylab("Density") +  labs(title = "Earnings per Share
Distributions By Sector from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

#pb
```

```
df <- data[!is.na(data$pb) &!is.na(data$Sector) & data$pb > 0 &
data$pb <10000,]
Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$pb[df$gsector == val], na.rm =TRUE)
  Minimum[i] <- min(df$pb[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$pb[df$gsector == val], .25,na.rm = TRUE)
  Median[i] <- median(df$pb[df$gsector == val], na.rm= TRUE)
  Q3[i] <- quantile(df$pb[df$gsector == val], .75,na.rm = TRUE)
  Maximum[i] <- max(df$pb[df$gsector == val], na.rm =TRUE)
  StandardDeviation[i] <- sd(df$pb[df$gsector == val], na.rm
=TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"pb_table")

df <- data[!is.na(data$pb) & data$pb <=quantile(origin_data$pb,
.97, na.rm=TRUE) & data$pb >=quantile(origin_data$pb, .02,
na.rm=TRUE),]

ggplot(data = df, aes(x = pb)) +
  geom_histogram(aes(x = pb, y = ..density..), bins = 100, col
="black", fill = "lightblue") +
  geom_density(aes(x = pb), col = "darkblue") +  xlab("Price to
Book Ratio") +
```

```
  ylab("Density") +  labs(title = "Price to Book Ratio
Distribution from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12))+
  theme_bw()

ggplot(data = df, aes(x = pb, fill = Decade)) +
  geom_density(aes(x = pb), alpha = .25) +  xlab("Price to Book
Ratio") +
  ylab("Density") +  labs(title = "Price to Book Ration
Distribution from 1970 to 2019 (By Decade)")+
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12))


ggplot(data = df, aes(x = fyear, y = pb, group = fyear)) +
  geom_boxplot() +
  stat_summary(fun.y=mean, colour="blue", geom="point",
shape=7,size=2) +
  ylim(-2, 8) +  xlab("Year") +  ylab("Price to Book Ratio") +
  labs(title = "Boxplots of Price to Book Ratio from 1970 (By
Year)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

df <- data[!is.na(data$pb) &!is.na(data$Sector) & data$pb <=
quantile(origin_data$pb,.97, na.rm=TRUE) & data$pb >=
quantile(origin_data$pb, .02,na.rm=TRUE),]

ggplot(data = df, aes(x = pb, fill = Sector)) +
  geom_histogram(aes(x = pb), bins = 20) +
  facet_wrap(~ Sector) +  xlab("Price to Book Ratio") +
  ylab("Density") +  labs(title = "Price to Book Ratio
Histograms By Sector from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()
```

```
ggplot(data = df, aes(x = pb, color = Sector)) +  geom_density()
+  xlab("Price to Book Ratio") +  ylab("Density") +
  labs(title = "Price to Book Ratio Distributions By Sector from
1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

#pe
df <- data[!is.na(data$pe) &!is.na(data$Sector) & data$pe > 0 &
data$pe <10000,]
Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$pe[df$gsector == val], na.rm =TRUE)
  Minimum[i] <- min(df$pe[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$pe[df$gsector == val], .25,na.rm = TRUE)
  Median[i] <- median(df$pe[df$gsector == val], na.rm= TRUE)
  Q3[i] <- quantile(df$pe[df$gsector == val], .75,na.rm = TRUE)
  Maximum[i] <- max(df$pe[df$gsector == val], na.rm =TRUE)
  StandardDeviation[i] <- sd(df$pe[df$gsector == val], na.rm
=TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"pe_table")
```

```
df <- data[!is.na(data$pe) & data$pe > 0 &data$pe <=
quantile(origin_data$pe, .97, na.rm=TRUE),]
ggplot(data = df, aes(x = pe)) +
  geom_histogram(aes(x = pe, y = ..density..), bins = 100, col
="black", fill = "lightblue") +
  geom_density(aes(x = pe), col = "darkblue") +  xlab("Price to
Earnings Ratio") +
  ylab("Density") +  labs(title = "Price to Earnings Ratio
Distribution from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()


ggplot(data = df, aes(x = pe, fill = Decade)) +
  geom_density(aes(x = pe), alpha = .25) +
  xlab("Price to Earnings Ratio") +
  ylab("Density") +
  labs(title = "Price to Earnings Ratio Distribution from 1970
to 2019 (By Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = fyear, y = pe, group = fyear)) +
  geom_boxplot() +  stat_summary(fun.y=mean, colour="blue",
geom="point", shape=7,size=2) +
  ylim(0, 75) +  xlab("Year") +  ylab("Price to Earnings Ratio")
+
  labs(title = "Boxplots of Price to Earnings Ratio from 1970
(By Year)")+
  theme(axis.title.x = element_text(size=15), axis.title.y
=element_text(size=15)) +
  theme_bw()

df <- data[!is.na(data$pe) &!is.na(data$Sector) & data$pe <=
quantile(origin_data$pe,.97, na.rm=TRUE) & data$pe > 0,]
ggplot(data = df, aes(x = pe, fill = Sector)) +
```

```
  geom_histogram(aes(x = pe), bins = 20) +  facet_wrap(~ Sector)
+
  xlab("Price to Earnings Ratio") +  ylab("Density") +
labs(title = "Price to Earnings Ratio Histograms By Sector from
1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12))

ggplot(data = df, aes(x = pe, color = Sector)) +  geom_density()
+  xlab("Price to Earnings Ratio") +  ylab("Density") +
labs(title = "Price to Earnings Ratio Distributions By Sector
from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

#divYield
df <- data[!is.na(data$divYield) &!is.na(data$Sector) &
data$dvpsp_c > 0,]

Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$divYield[df$gsector == val], na.rm =TRUE)
  Minimum[i] <- min(df$divYield[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$divYield[df$gsector == val], .25,na.rm =
TRUE)
  Median[i] <- median(df$divYield[df$gsector == val], na.rm=
TRUE)
  Q3[i] <- quantile(df$divYield[df$gsector == val], .75,na.rm =
TRUE)
  Maximum[i] <- max(df$divYield[df$gsector == val], na.rm =TRUE)
```

```
  StandardDeviation[i] <- sd(df$divYield[df$gsector == val],
na.rm =TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"divYield_table.csv")

df <- data[!is.na(data$divYield) &data$dvpsp_c > 0 &
data$divYield <=quantile(origin_data$divYield, .99,
na.rm=TRUE),]

ggplot(data = df, aes(x = divYield)) +
  geom_histogram(aes(x = divYield, y = ..density..), bins =
100,col = "black", fill = "lightblue") +
  geom_density(aes(x = divYield), col = "darkblue") +
xlab("Dividend Yield") +  ylab("Density") +
  labs(title = "Dividend Yield Distribution from 1970 to 2019")
+
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = divYield, fill = Decade)) +
  geom_density(aes(x = divYield), alpha = .25) +
  xlab("Divident Yield") +  ylab("Density") +
  labs(title = "Dividend Yield Distribution from 1970 to 2019
(By Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()


ggplot(data = df, aes(x = fyear, y = divYield, group =fyear)) +
```

```
  geom_boxplot() +  stat_summary(fun.y=mean, colour="blue",
geom="point", shape=7,size=2) +
  ylim(0, .15) +  xlab("Year") +  ylab("Divident Yield") +
  labs(title = "Boxplots of Dividend Yield from 1970 (By Year)")
+
  theme(axis.title.x = element_text(size=15), axis.title.y
=element_text(size=15))

df <- data[!is.na(data$divYield) &!is.na(data$Sector) &
data$divYield <=quantile(origin_data$divYield, .99, na.rm=TRUE)
& data$divYield >0,]

ggplot(data = df, aes(x = divYield, fill = Sector)) +
  geom_histogram(aes(x = divYield), bins = 20) +
  facet_wrap(~ Sector) +
  xlab("Dividend Yield") +
  ylab("Density") +
  labs(title = "Dividend Yield Histograms By Sector from 1970 to
2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12))

ggplot(data = df, aes(x = divYield, color = Sector)) +
  geom_density() +  xlab("Dividend Yield") +
  ylab("Density") +
  labs(title = "Dividend Yield Distributions By Sector from 1970
to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12))

#cr

df <- data[!is.na(data$cr) &!is.na(data$Sector) & data$lct !=
0,]

Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
```

```
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$cr[df$gsector == val], na.rm =TRUE)
  Minimum[i] <- min(df$cr[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$cr[df$gsector == val], .25,na.rm = TRUE)
  Median[i] <- median(df$cr[df$gsector == val], na.rm= TRUE)
  Q3[i] <- quantile(df$cr[df$gsector == val], .75,na.rm = TRUE)
  Maximum[i] <- max(df$cr[df$gsector == val], na.rm =TRUE)
  StandardDeviation[i] <- sd(df$cr[df$gsector == val], na.rm
=TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"cr.csv")


df <- data[!is.na(data$cr) & data$cr <=quantile(origin_data$pb,
.97, na.rm=TRUE) & data$cr >=quantile(origin_data$pb, .02,
na.rm=TRUE),]

ggplot(data = df, aes(x = cr)) +
  geom_histogram(aes(x = cr, y = ..density..), bins = 100,col =
"black", fill = "lightblue") +
  geom_density(aes(x = cr), col = "darkblue") +  xlab("Current
Ratio") +  ylab("Density") +
  labs(title = "Current Ratio Distribution from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()
```

```
ggplot(data = df, aes(x = cr, fill = Decade)) +
  geom_density(aes(x = cr), alpha = .25) +
  xlab("Current Ratio") +  ylab("Density") +
  labs(title = "Current Ratio Distribution from 1970 to 2019 (By
Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()


ggplot(data = df, aes(x = fyear, y = cr, group =fyear)) +
  geom_boxplot() +  stat_summary(fun.y=mean, colour="blue",
geom="point", shape=7,size=2) +
  ylim(0, .15) +  xlab("Year") +  ylab("Current Ratio") +
  labs(title = "Boxplots of Current Ratio from 1970 (By Year)")
+
  theme(axis.title.x = element_text(size=15), axis.title.y
=element_text(size=15)) +
  theme_bw()

df <- data[!is.na(data$cr) &!is.na(data$Sector) & data$cr <=
quantile(origin_data$cr,.97, na.rm=TRUE) & data$cr >=
quantile(origin_data$cr, .02,na.rm=TRUE),]
ggplot(data = df, aes(x = cr, fill = Sector)) +
  geom_histogram(aes(x = cr), bins = 20) +
  facet_wrap(~ Sector) +
  xlab("Current Ratio") +
  ylab("Density") +
  labs(title = "Current Ratio Histograms By Sector from 1970 to
2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) + theme_bw()

ggplot(data = df, aes(x = cr, color = Sector)) +
  geom_density() +  xlab("Current Ratio") +
  ylab("Density") +
  labs(title = "Current Ratio Distributions By Sector from 1970
to 2019") +
```

```
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

#leverage
df <- data[!is.na(data$leverage) &!is.na(data$Sector) & data$at
!= 0 & data$dltt >0,]

Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$leverage[df$gsector == val], na.rm =TRUE)
  Minimum[i] <- min(df$leverage[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$leverage[df$gsector == val], .25,na.rm =
TRUE)
  Median[i] <- median(df$leverage[df$gsector == val], na.rm=
TRUE)
  Q3[i] <- quantile(df$leverage[df$gsector == val], .75,na.rm =
TRUE)
  Maximum[i] <- max(df$leverage[df$gsector == val], na.rm =TRUE)
  StandardDeviation[i] <- sd(df$leverage[df$gsector == val],
na.rm =TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"leverage.csv")
```

```
df <- data[!is.na(data$leverage) &data$leverage <=
quantile(origin_data$leverage, .95, na.rm=TRUE) &data$dltt > 0,]

ggplot(data = df, aes(x = leverage)) +
  geom_histogram(aes(x = leverage, y = ..density..), bins =
100,col = "black", fill = "lightblue") +
  geom_density(aes(x = leverage), col = "darkblue") +
  xlab("Leverage") +  ylab("Density") +
  labs(title = "Leverage Distribution from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = leverage, fill = Decade)) +
  geom_density(aes(x = leverage), alpha = .25) +
  xlab("Leverage") +  ylab("Density") +
  labs(title = "Leverage Distribution from 1970 to 2019 (By
Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()


ggplot(data = df, aes(x = fyear, y = leverage, group =fyear)) +
  geom_boxplot() +  stat_summary(fun.y=mean, colour="blue",
geom="point", shape=7,size=2) +
  ylim(0, .4) +  xlab("Year") +
  ylab("Leverage") +  labs(title = "Boxplots of Leverage from
1970 (By Year)") +
  theme(axis.title.x = element_text(size=15), axis.title.y
=element_text(size=15)) +
  theme_bw()

df <- data[!is.na(data$leverage) &!is.na(data$Sector) &
data$leverage <=quantile(origin_data$cr, .85, na.rm=TRUE) &
data$dltt > 0,]


ggplot(data = df, aes(x = leverage, fill = Sector)) +
```

```
  geom_histogram(aes(x = leverage), bins = 20) +
  facet_wrap(~ Sector) +  xlab("Leverage") +
  ylab("Density") +  labs(title = "Leverage Histograms By Sector
from 1970 to 2019")+
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = leverage, color = Sector)) +
  geom_density() +  xlab("Leverage") +
  ylab("Density") +
  labs(title = "Leverage Distributions By Sector from 1970 to
2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()


#turnover
df <- data[!is.na(data$turnover) &!is.na(data$Sector) & data$at
!= 0 & data$sale >0,]


Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$turnover[df$gsector == val], na.rm =TRUE)
  Minimum[i] <- min(df$turnover[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$turnover[df$gsector == val], .25,na.rm =
TRUE)
  Median[i] <- median(df$turnover[df$gsector == val], na.rm=
TRUE)
```

```
  Q3[i] <- quantile(df$turnover[df$gsector == val], .75,na.rm =
TRUE)
  Maximum[i] <- max(df$turnover[df$gsector == val], na.rm =TRUE)
  StandardDeviation[i] <- sd(df$turnover[df$gsector == val],
na.rm =TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"turnover.csv")

df <- data[!is.na(data$turnover) &data$turnover <=
quantile(origin_data$turnover, .95, na.rm=TRUE) &data$sale > 0,]


ggplot(data = df, aes(x = turnover)) +
  geom_histogram(aes(x = turnover, y = ..density..), bins =
100,col = "black", fill = "lightblue") +
  geom_density(aes(x = turnover), col = "darkblue") +
  xlab("Turnover") +  ylab("Density") +
  labs(title = "Turnover Distribution from 1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = turnover, fill = Decade)) +
  geom_density(aes(x = turnover), alpha = .25) +
  xlab("Turnover") +  ylab("Density") +
  labs(title = "Turnover Distribution from 1970 to 2019 (By
Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = fyear, y = turnover, group =fyear)) +
```

```
  geom_boxplot() +  stat_summary(fun.y=mean, colour="blue",
geom="point", shape=7,size=2) +
  ylim(0, 2.6) +  xlab("Year") +  ylab("Turnover") +
  labs(title = "Boxplots of Turnover from 1970 (By Year)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

df <- data[!is.na(data$turnover) &!is.na(data$Sector) &
data$turnover <=quantile(origin_data$cr, .90, na.rm=TRUE) &
data$sale > 0,]
ggplot(data = df, aes(x = turnover, fill = Sector)) +
  geom_histogram(aes(x = turnover), bins = 20) +  facet_wrap(~
Sector) +
  xlab("Turnover") +  ylab("Density") +
  labs(title = "Turnover Histograms By Sector from 1970 to
2019")+
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = turnover, color = Sector)) +
  geom_density() +  xlab("Turnover") +  ylab("Density") +
  labs(title = "Turnover Distributions By Sector from 1970 to
2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
   theme_bw()

#gross margin ratio
df <- data[!is.na(data$grossMarginRatio) &!is.na(data$Sector) &
data$sale != 0,]


Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
```

```
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$grossMarginRatio[df$gsector == val], na.rm
=TRUE)
  Minimum[i] <- min(df$grossMarginRatio[df$gsector == val],
na.rm =TRUE)
  Q1[i] <- quantile(df$grossMarginRatio[df$gsector == val],
.25,na.rm = TRUE)
  Median[i] <- median(df$grossMarginRatio[df$gsector == val],
na.rm= TRUE)
  Q3[i] <- quantile(df$grossMarginRatio[df$gsector == val],
.75,na.rm = TRUE)
  Maximum[i] <- max(df$grossMarginRatio[df$gsector == val],
na.rm =TRUE)
  StandardDeviation[i] <- sd(df$grossMarginRatio[df$gsector ==
val], na.rm =TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"grossMargin.csv")


df <- data[!is.na(data$grossMarginRatio) &data$grossMarginRatio
<= 1 & data$grossMarginRatio > 0,]

ggplot(data = df, aes(x = grossMarginRatio)) +
  geom_histogram(aes(x = grossMarginRatio, y = ..density..),
bins =100, col = "black", fill = "lightblue") +
  geom_density(aes(x = grossMarginRatio), col = "darkblue") +
  xlab("Gross Margin Ratio") +  ylab("Density") +
```

```
  labs(title = "Gross Margin Ratio Distribution from 1970 to 
2019")+
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = grossMarginRatio, fill = Decade)) +
  geom_density(aes(x = grossMarginRatio), alpha = .25) +
  xlab("Gross Margin Ratio") +
  ylab("Density") +
  labs(title = "Gross Margin Ratio Distribution from 1970 to 
2019 (By Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = fyear, y = grossMarginRatio, group
=fyear)) +
  geom_boxplot() +
  stat_summary(fun.y=mean, colour="blue", geom="point",
shape=7,size=2) +
  ylim(0, 1) +  xlab("Year") +  ylab("Gross Margin Ratio") +
  labs(title = "Boxplots of Gross Margin Ratios from 1970 (By
Year)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

df <- data[!is.na(data$grossMarginRatio) &!is.na(data$Sector) &
data$grossMarginRatio <= 1 &data$grossMarginRatio > 0,]

ggplot(data = df, aes(x = grossMarginRatio, fill = Sector)) +
  geom_histogram(aes(x = grossMarginRatio), bins = 20) +
  facet_wrap(~ Sector) +
  xlab("Gross Margin Ratio") +
  ylab("Density") +
  labs(title = "Gross Margin Ratio Histograms By Sector from
1970 to 2019") +
```

```
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = grossMarginRatio, color = Sector)) +
  geom_density() +  xlab("Gross Margin Ratio") +
  ylab("Density") +
  labs(title = "Gross Margin Ratio Distributions By Sector from
1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

#ROA
df <- data[!is.na(data$ROA) &!is.na(data$Sector) & data$at !=
0,]


Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$ROA[df$gsector == val], na.rm =TRUE)
  Minimum[i] <- min(df$ROA[df$gsector == val], na.rm =TRUE)
  Q1[i] <- quantile(df$ROA[df$gsector == val], .25,na.rm = TRUE)
  Median[i] <- median(df$ROA[df$gsector == val], na.rm= TRUE)
  Q3[i] <- quantile(df$ROA[df$gsector == val], .75,na.rm = TRUE)
  Maximum[i] <- max(df$ROA[df$gsector == val], na.rm =TRUE)
  StandardDeviation[i] <- sd(df$ROA[df$gsector == val], na.rm
=TRUE)
  i <- i + 1}
```

```
sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"ROA.csv")

df <- data[!is.na(data$ROA) & data$at != 0 &data$ROA <= 1.2 &
data$ROA >= quantile(origin_data$turnover,.02, na.rm=TRUE),]

ggplot(data = df, aes(x = ROA)) +
  geom_histogram(aes(x = ROA, y = ..density..), bins = 100, col=
"black", fill = "lightblue") +
  geom_density(aes(x = ROA), col = "darkblue") +
  xlab("Return on Assets") +  ylab("Density") +
  labs(title = "Return on Assets Distribution from 1970 to
2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = ROA, fill = Decade)) +
  geom_density(aes(x = ROA), alpha = .25) +
  xlab("Return on Assets") +  ylab("Density") +
  labs(title = "Return on Assets Distribution from 1970 to 2019
(By Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()


ggplot(data = df, aes(x = fyear, y = ROA, group = fyear)) +
  geom_boxplot() +  stat_summary(fun.y=mean, colour="blue",
geom="point", shape=7,size=2) +
  ylim(0,.4) +  xlab("Year") +  ylab("Gross Margin Ratio") +
  labs(title = "Boxplots of Returns on Assets from 1970 (By
Year)") +
```

```
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

df <- data[!is.na(data$ROA) &!is.na(data$Sector) & data$ROA <=
.5 & data$ROA>= 0,]

ggplot(data = df, aes(x = ROA, fill = Sector)) +
  geom_histogram(aes(x = ROA), bins = 20) +
  facet_wrap(~ Sector) +  xlab("Return on Assets") +
  ylab("Density") +
  labs(title = "Return on Assets Histograms By Sector from 1970
to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = ROA, color = Sector)) +
  geom_density() +  xlab("Return on Assets") +  ylab("Density")
+
  labs(title = "Return on Assets Distributions By Sector from
1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

#cfo_assetsratio
df <- data[!is.na(data$cfo_assetsratio) &!is.na(data$Sector) &
data$at != 0,]

Mean <- numeric(11)
Minimum <- numeric(11)
Q1 <- numeric(11)
Median <- numeric(11)
Q3 <- numeric(11)
Maximum <- numeric(11)
StandardDeviation <- numeric(11)

i <- 1
```

```
for(val in seq(10,60,5)) {
  Mean[i] <- mean(df$cfo_assetsratio[df$gsector == val], na.rm
=TRUE)
  Minimum[i] <- min(df$cfo_assetsratio[df$gsector == val], na.rm
=TRUE)
  Q1[i] <- quantile(df$cfo_assetsratio[df$gsector == val],
.25,na.rm = TRUE)
  Median[i] <- median(df$cfo_assetsratio[df$gsector == val],
na.rm= TRUE)
  Q3[i] <- quantile(df$cfo_assetsratio[df$gsector == val],
.75,na.rm = TRUE)
  Maximum[i] <- max(df$cfo_assetsratio[df$gsector == val], na.rm
=TRUE)
  StandardDeviation[i] <- sd(df$cfo_assetsratio[df$gsector ==
val], na.rm =TRUE)
  i <- i + 1}

sectors <- data.frame(Sector = c("Energy","Materials",
"Industrials", "Consumer Discretionary", "ConsumerStaples",
"Health Care", "Financials", "Information
Technology","Telecommunications Services", "Utilities", "Real
Estate"),Mean, Minimum, Q1, Median, Q3,
Maximum,StandardDeviation)
write.csv(sectors,"cfo_assetsratio.csv")

df <- data[!is.na(data$cfo_assetsratio) &data$cfo_assetsratio <=
quantile(origin_data$cfo_assetsratio, .98,na.rm=TRUE) &
data$cfo_assetsratio >quantile(origin_data$cfo_assetsratio, .05,
na.rm=TRUE),]

ggplot(data = df, aes(x = cfo_assetsratio)) +
  geom_histogram(aes(x = cfo_assetsratio, y = ..density..),
bins= 100, col = "black", fill = "lightblue") +
  geom_density(aes(x = cfo_assetsratio), col = "darkblue") +
  xlab("CFO/Assets Ratio") +  ylab("Density") +
  labs(title = "CFO/Assets Ratio Distribution from 1970 to
2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
```

```
  theme_bw()

ggplot(data = df, aes(x = cfo_assetsratio, fill = Decade)) +
  geom_density(aes(x = cfo_assetsratio), alpha = .25) +
  xlab("CFO/Assets Ratio") +  ylab("Density") +
  labs(title = "CFO/Assets Ratio Distribution from 1970 to 2019
(By Decade)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

ggplot(data = df, aes(x = fyear, y = cfo_assetsratio, group=
fyear)) +
  geom_boxplot() +  stat_summary(fun.y=mean, colour="blue",
geom="point", shape=7,size=2) +
  ylim(0, .25) +  xlab("Year") +  ylab("CFO/Assets Ratio") +
labs(title = "Boxplots of CFO/Assets Ratios from 1970 (By
Year)") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()

df <- data[!is.na(data$cfo_assetsratio) &!is.na(data$Sector) &
data$cfo_assetsratio <=quantile(origin_data$cfo_assetsratio,
.98, na.rm=TRUE) &data$cfo_assetsratio >
quantile(origin_data$cfo_assetsratio, .05,na.rm=TRUE),]

ggplot(data = df, aes(x = cfo_assetsratio, fill = Sector)) +
  geom_histogram(aes(x = cfo_assetsratio), bins = 20) +
  facet_wrap(~ Sector) +  xlab("CFO/Assets Ratio") +
  ylab("Density") +
  labs(title = "CFO/Assets Ratio Histograms By Sector from 1970
to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()


ggplot(data = df, aes(x = cfo_assetsratio, color = Sector))+
```

```
  geom_density() +  xlab("CFO/Assets Ratio") +  ylab("Density")
+
  labs(title = "CFO/Assets Ratio Distributions By Sector from
1970 to 2019") +
  theme(axis.title.x = element_text(size=12), axis.title.y
=element_text(size=12)) +
  theme_bw()
```

References

[1] Barone, Adam. "Small Cap." *Investopedia*, Investopedia, 16 Sept. 2020, www.investopedia.com/terms/s/small-cap.asp.

[2] Wharton Research Data Services (WRDS), The University of Pennsylvania Wharton School of Business. September 13, 2019.