MINI Project 3

Data Collection and Cleaning

STAT 686 Group 4

Christina Wang, Oliver (Ran) Jin, Rui Qin, Xingyue Zhang

2 March, 2018

Contents

Introduction

We will look at three key factors in financial analysis:

- Market capitalization: market value of the company's outstanding shares.
- Dividend yield: a ratio that represents how much a company pays out in dividends each year relative to its share price
- EBITDA/EV: EBITDA is an acronym that stands for Earnings before Interest, Taxes, Depreciation, and Amortization. EV is an acronym that stands for Enterprise Value.

The three factors are all key indicators of a company's current and potential future finance, and are widely used by investors. In this report, we will first display how the data was obtained and cleaned. Then we will present a table showing the data coverage, as well as the distributions of the key financial factors.

In addition, we will plot and analyze the distribution for 1965 ~ 1975, and 2005 ~ 2015 for the above three factors. We will also compare the distribution by industry (Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Telecommunication Services, and Utilities) for the above two decades and the above three factors.

Part I Data Description

Data Cleaning & Coverage

The data was obtained from WRDS CRSP-COMPUSTAT Merged (CCM), ranging from 1925 to 2017 for market cap, and from 1950 to 2014 for all other factors. However, the data was incomplete. Data labeled as NA, non-existing, or are negative for some variables (prcc_f and che) are considered as uncovered and are removed. The data coverage for each variable is presented below:

Table 1. Data Coverage

Variable Acronym	Full Variable Name	coverage
che	Cash and Cash Equivalents	92.45%
csho	Common Shares Outstanding	99.40%
dlc	Total Debt in Current Liabilities	92.00%
dltt	Total Long Term Debt	92.81%
dvc	Dividends (common)	92.58%
ebitda	Earnings before Interest, Taxes, Depreciation, and Amortization	90.71%
pstk	Preferred Stock	91.47%
exchg	Stock Exchange Code	100.00%
dvpsp_f	Dividends Paid Per Share at Fiscal Year End	93.72%
prcc_f	Fiscal Year End Price Close	96.33%
gsector	General Sector	87.48%

According to Table 1, the data coverage is not bad overall. Disregarding EXCHG, the variable that contains the least missing data is CSHO (99.40%), and the variable that contains the most missing data is EBITDA (90.71%).

Because Market Capitalization is calculated based on PRCC_Fand CSHO, the coverage for Market Capitalization is: **96.33%**. Because Dividend Yield is calculated using DVPSP_F and PRCC_F, the coverage for Dividend Yield is **93.72%**. Because EBITDA/EV is calculated by EBITDA, CHSO, PRCC, DLTT, DLC, PSTK, and CHE, the coverage for EBITDA/EV is **90.71%**.

Variable Calculation Methodology

For Market Capitalization:

$$\text{Market Capitalization} = \text{prcc\_f} * \text{csho}$$

$$\text{Inflation Adjusted Market Cap} = \frac{\text{Market Capitalization}}{\text{CPI}}$$

The base year is chained, with 1982 – 1984 = 100%

For Dividend Yield:

$$\text{Dividend Yield} = \frac{\text{dvpsp\_f}}{\text{prcc\_f}}$$

For EBITDA/EV:

$$\frac{\text{EBITDA}}{\text{EV}} = \frac{\text{EBITDA}}{\text{MKTCAP (CHSO}*\text{PRCC) + Debt (DLTT + DLC) +(PSTK) − (CHE)}}$$

Part II Market Capitalization

Overview

Market Capitalization, as mentioned in the Introduction, indicates the total dollar market value of a company's outstanding shares. Investors use this factor to determine a company's size. The size of a company is important because it is an essential factor for risk assessment as well as a good estimation for profit return. Companies with large market-cap are typically of $10 billion or more. Investing in large-cap companies does not usually bring in dramatic returns in short term, but usually gets consistent return over the long term. Companies with mid-capitalizations are usually worth between $2 billion and $10 billion. Mid-cap companies are in the process of expanding. They are riskier than the large-cap companies, but they also attract greater possible returns. Companies that are of higher risk, but also of higher possible returns, are the small-cap ones. They have the market capitalization between $300 million and $200 billion.

The two charts below show the annual trend from 1925 to 2017 for both the number of companies as well as the proportion of companies which have the market capitalization over $50 million, over annual average, and over universal median.

Annual Analysis

The table below shows presents the total count for the number of companies each year, the number of companies over 50 million each year, the number of companies over the universal median each year, the number of companies over the annual mean each year, and the corresponding proportion of companies over 50 million each year, the proportion of companies over universal median each year, and the proportion of companies over annual mean each year.

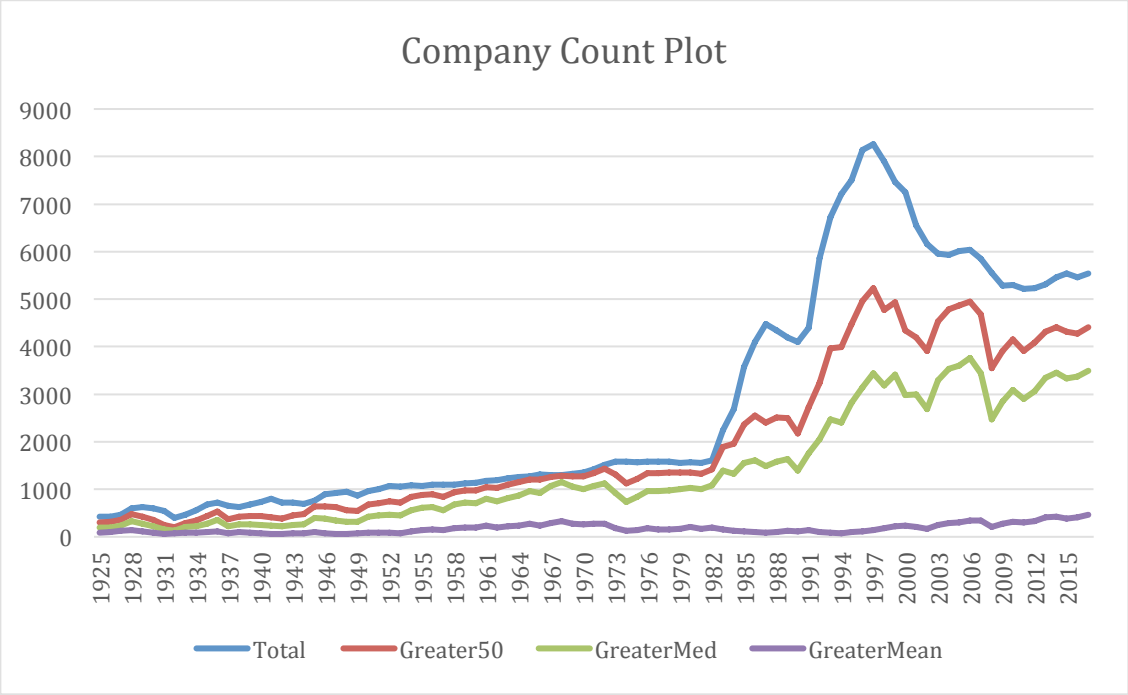
Table 2

Year	YearlyAvg	Total	Greater50	GreaterMed	GreaterMean	Prop50	PropMedian	PropMean
1925	354.71	426	304	189	88	71.36%	44.37%	20.66%
1926	414.21	421	317	200	97	75.30%	47.51%	23.04%
1927	512.81	463	356	241	121	76.89%	52.05%	26.13%
1928	634.10	593	472	334	141	79.60%	56.32%	23.78%
1929	522.19	625	429	273	111	68.64%	43.68%	17.76%
1930	400.31	597	350	219	89	58.63%	36.68%	14.91%
1931	256.85	538	252	159	56	46.84%	29.55%	10.41%
1932	308.85	398	201	130	68	50.50%	32.66%	17.09%
1933	430.63	464	291	210	88	62.72%	45.26%	18.97%
1934	365.88	564	339	226	81	60.11%	40.07%	14.36%
1935	435.95	677	443	281	97	65.44%	41.51%	14.33%
1936	547.05	718	527	354	115	73.40%	49.30%	16.02%
1937	347.28	654	376	221	79	57.49%	33.79%	12.08%
1938	453.42	621	422	256	98	67.95%	41.22%	15.78%
1939	430.31	676	443	267	91	65.53%	39.50%	13.46%
1940	350.61	727	439	251	75	60.39%	34.53%	10.32%
1941	268.93	796	416	235	56	52.26%	29.52%	7.04%
1942	277.85	722	386	226	57	53.46%	31.30%	7.89%
1943	320.88	716	450	254	67	62.85%	35.47%	9.36%
1944	368.66	688	483	268	79	70.20%	38.95%	11.48%
1945	450.50	756	634	395	97	83.86%	52.25%	12.83%
1946	353.34	898	644	377	77	71.71%	41.98%	8.57%
1947	296.21	924	624	344	55	67.53%	37.23%	5.95%

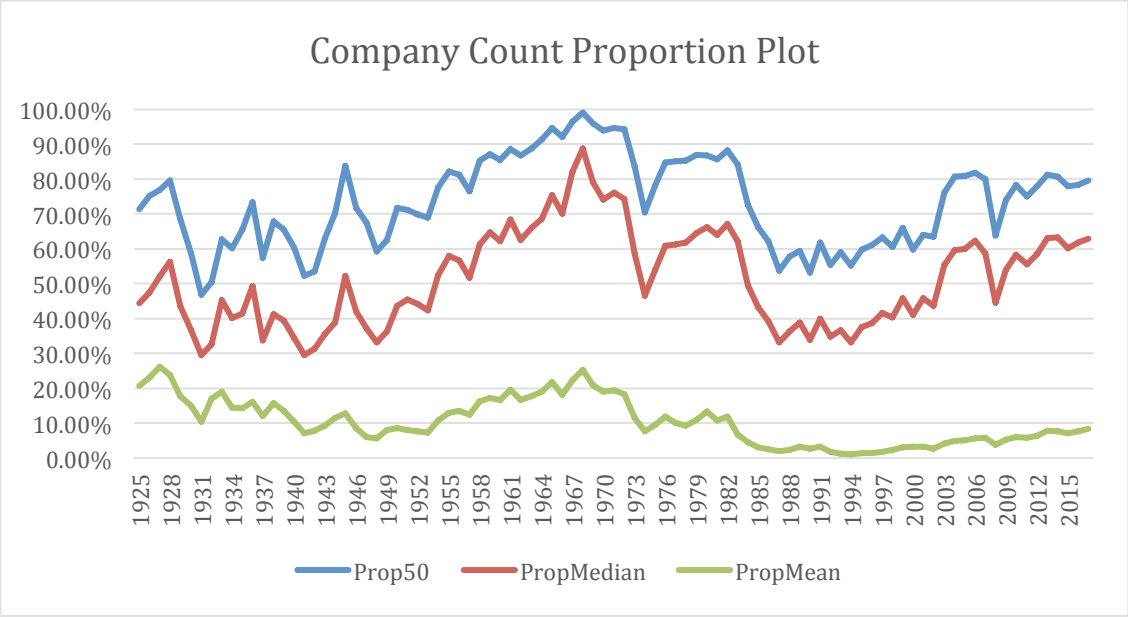
Year	YearlyAvg	Total	Greater50	GreaterMed	GreaterMean	Prop50	PropMedian	PropMean
1948	267.81	952	564	315	54	59.24%	33.09%	5.67%
1949	320.37	869	544	315	69	62.60%	36.25%	7.94%
1950	381.96	956	685	418	82	71.65%	43.72%	8.58%
1951	408.54	999	711	455	81	71.17%	45.55%	8.11%
1952	411.63	1063	743	470	82	69.90%	44.21%	7.71%
1953	401.90	1050	725	445	77	69.05%	42.38%	7.33%
1954	611.67	1077	835	564	116	77.53%	52.37%	10.77%
1955	723.18	1064	875	616	139	82.24%	57.89%	13.06%
1956	748.10	1103	896	625	150	81.23%	56.66%	13.60%
1957	651.74	1092	836	564	136	76.56%	51.65%	12.45%
1958	903.80	1101	939	674	179	85.29%	61.22%	16.26%
1959	981.83	1117	974	724	193	87.20%	64.82%	17.28%
1960	928.25	1135	970	706	189	85.46%	62.20%	16.65%
1961	1122.74	1174	1041	804	231	88.67%	68.48%	19.68%
1962	963.59	1192	1035	745	199	86.83%	62.50%	16.69%
1963	1087.09	1228	1090	811	218	88.76%	66.04%	17.75%
1964	1213.94	1258	1149	862	239	91.34%	68.52%	19.00%
1965	1326.51	1272	1204	959	277	94.65%	75.39%	21.78%
1966	1131.16	1310	1206	917	238	92.06%	70.00%	18.17%
1967	1390.16	1300	1255	1069	290	96.54%	82.23%	22.31%
1968	1515.77	1294	1281	1148	327	99.00%	88.72%	25.27%
1969	1268.04	1328	1274	1050	277	95.93%	79.07%	20.86%
1970	1173.73	1355	1274	1003	259	94.02%	74.02%	19.11%
1971	1253.53	1413	1339	1075	273	94.76%	76.08%	19.32%

Year	YearlyAvg	Total	Greater50	GreaterMed	GreaterMean	Prop50	PropMedian	PropMean
1972	1342.91	1518	1431	1128	277	94.27%	74.31%	18.25%
1973	985.30	1576	1318	924	180	83.63%	58.63%	11.42%
1974	623.90	1587	1119	738	121	70.51%	46.50%	7.62%
1975	780.04	1564	1222	842	147	78.13%	53.84%	9.40%
1976	914.81	1575	1336	959	186	84.83%	60.89%	11.81%
1977	805.07	1576	1343	966	157	85.22%	61.29%	9.96%
1978	777.21	1581	1349	977	148	85.33%	61.80%	9.36%
1979	820.53	1560	1357	1007	172	86.99%	64.55%	11.03%
1980	941.61	1563	1357	1034	208	86.82%	66.15%	13.31%
1981	789.77	1554	1332	996	168	85.71%	64.09%	10.81%
1982	848.10	1610	1420	1080	191	88.20%	67.08%	11.86%
1983	750.51	2242	1886	1392	149	84.12%	62.09%	6.65%
1984	592.23	2691	1954	1331	124	72.61%	49.46%	4.61%
1985	553.98	3574	2368	1548	108	66.26%	43.31%	3.02%
1986	536.77	4095	2547	1605	102	62.20%	39.19%	2.49%
1987	475.47	4471	2402	1485	91	53.72%	33.21%	2.04%
1988	513.39	4345	2513	1576	99	57.84%	36.27%	2.28%
1989	628.50	4191	2493	1631	133	59.48%	38.92%	3.17%
1990	544.25	4103	2180	1388	110	53.13%	33.83%	2.68%
1991	659.98	4396	2721	1759	139	61.90%	40.01%	3.16%
1992	536.92	5866	3244	2048	100	55.30%	34.91%	1.70%
1993	526.08	6730	3969	2471	85	58.97%	36.72%	1.26%
1994	471.72	7211	3985	2396	72	55.26%	33.23%	1.00%
1995	603.24	7502	4479	2825	105	59.70%	37.66%	1.40%

Year	YearlyAvg	Total	Greater50	GreaterMed	GreaterMean	Prop50	PropMedian	PropMean
1996	660.70	8134	4966	3148	111	61.05%	38.70%	1.36%
1997	829.35	8263	5230	3445	147	63.29%	41.69%	1.78%
1998	1050.00	7893	4775	3189	179	60.50%	40.40%	2.27%
1999	1411.14	7460	4931	3419	223	66.10%	45.83%	2.99%
2000	1280.85	7256	4340	2983	229	59.81%	41.11%	3.16%
2001	1217.52	6552	4191	3001	207	63.97%	45.80%	3.16%
2002	1006.97	6161	3905	2687	170	63.38%	43.61%	2.76%
2003	1351.51	5953	4528	3294	244	76.06%	55.33%	4.10%
2004	1497.02	5935	4786	3540	287	80.64%	59.65%	4.84%
2005	1518.25	6011	4867	3602	302	80.97%	59.92%	5.02%
2006	1647.72	6044	4950	3766	336	81.90%	62.31%	5.56%
2007	1696.63	5856	4684	3435	337	79.99%	58.66%	5.75%
2008	1009.80	5556	3548	2476	208	63.86%	44.56%	3.74%
2009	1396.93	5286	3908	2846	275	73.93%	53.84%	5.20%
2010	1585.17	5302	4153	3088	315	78.33%	58.24%	5.94%
2011	1501.52	5215	3916	2898	301	75.09%	55.57%	5.77%
2012	1678.22	5235	4086	3060	333	78.05%	58.45%	6.36%
2013	2094.97	5307	4316	3349	411	81.33%	63.11%	7.74%
2014	2191.69	5460	4403	3456	421	80.64%	63.30%	7.71%
2015	2053.34	5537	4314	3333	389	77.91%	60.20%	7.03%
2016	2212.29	5458	4274	3374	415	78.31%	61.82%	7.60%
2017	2551.96	5543	4415	3492	468	79.65%	63.00%	8.44%



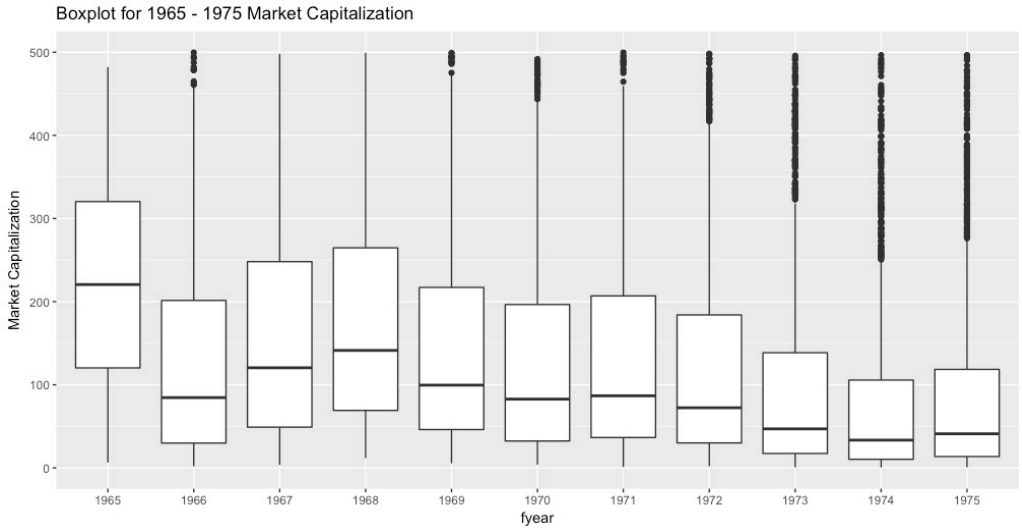
The company count plot shows that the number of public companies overall, greater than $ 50 million, and greater than universal median are all increasing. The number of companies above the annual mean is quite stable throughout the years. The time series trend reaches the maximum in the 1990s and the 2000s.

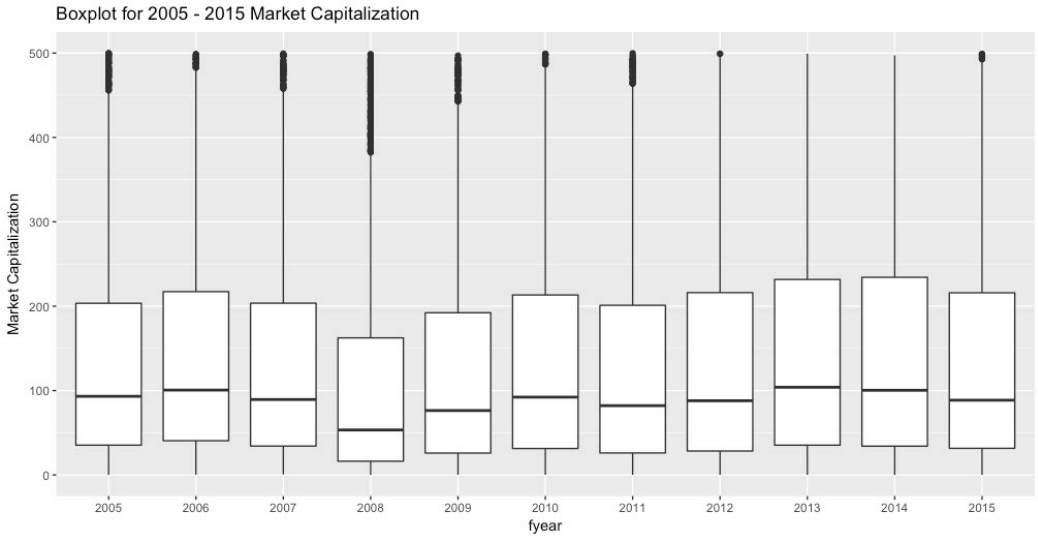


Unlike the count trend plot above, the count proportion plot is not in an overall increasing trend. The proportion of companies above annual mean is even decreasing over the years, indicating that the distribution of company market cap is becoming more and more rightly skewed. The proportion trend reaches its maximum in the 1960s for all three factors, indicating that the company market caps are generally larger in this period.

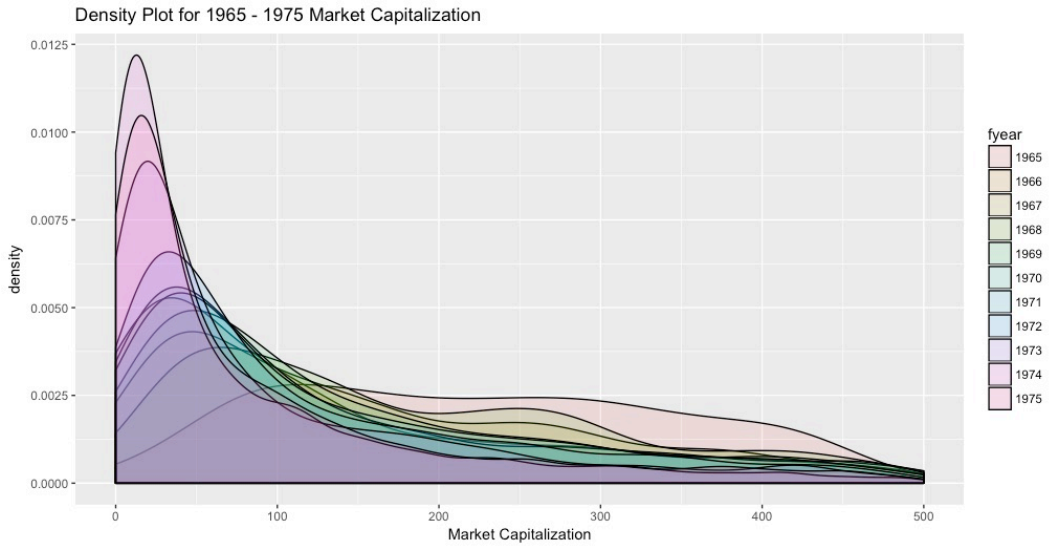
Part III Distribution Plot Comparison for 1965 ~ 1975, and 2005 ~ 2015

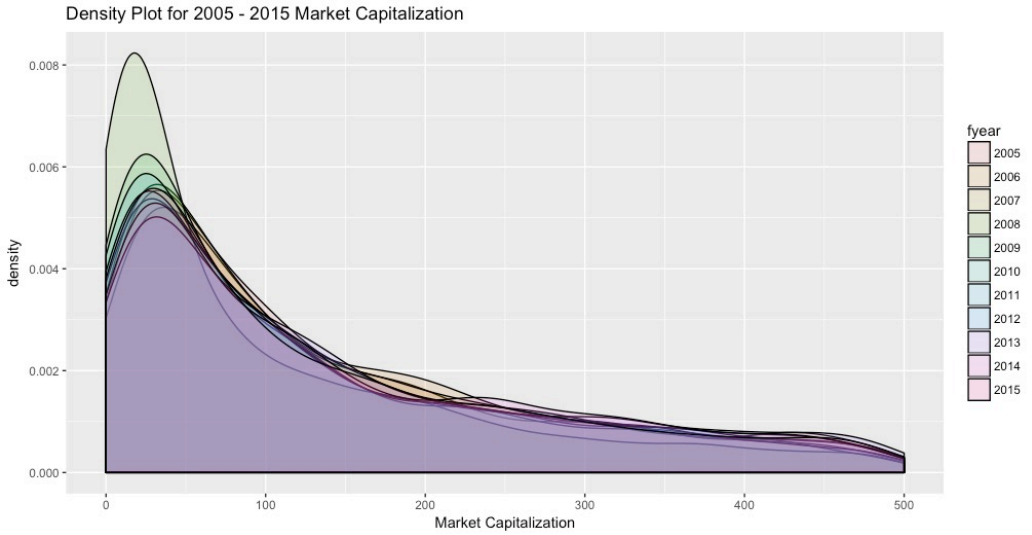
Market Capitalization



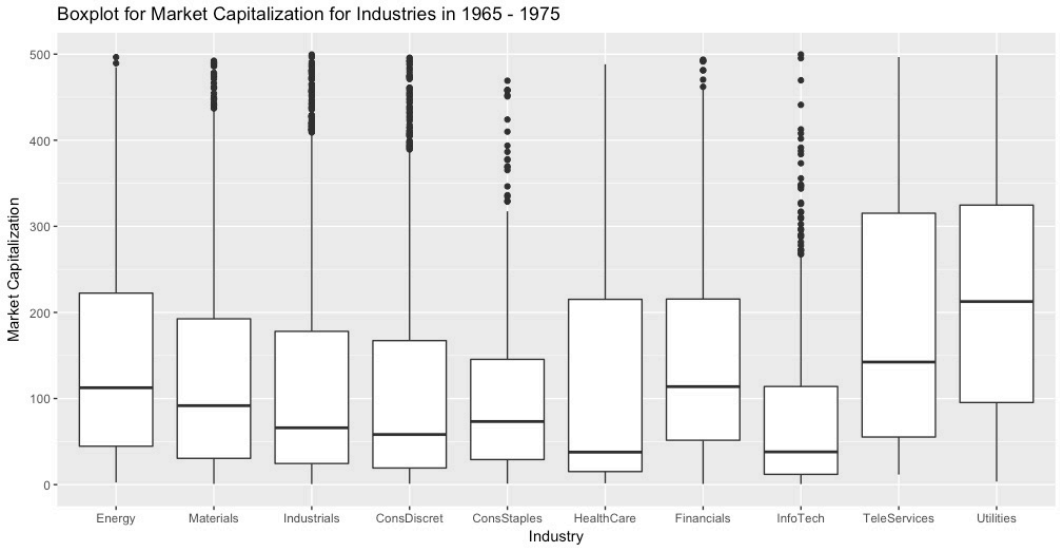


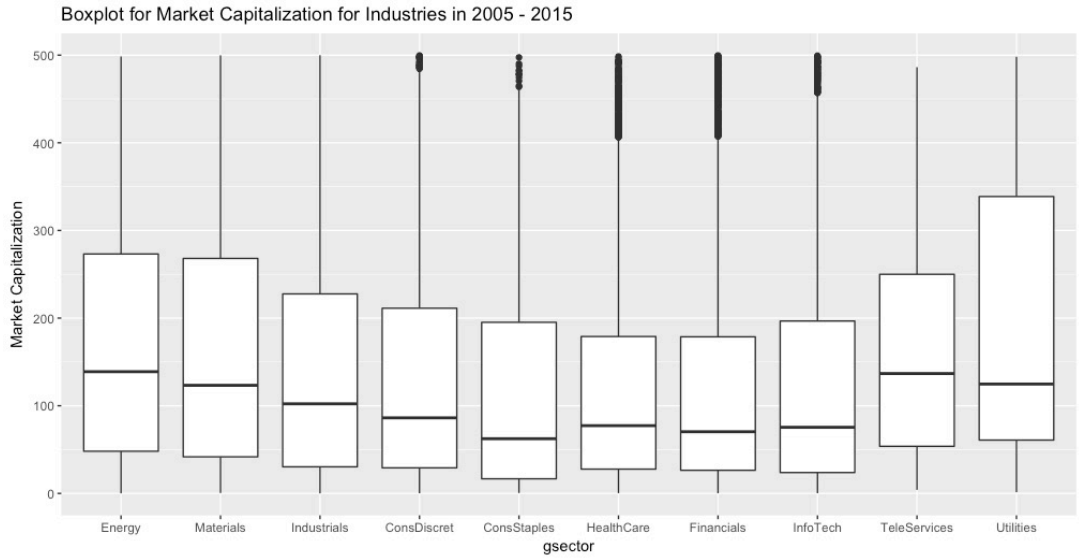
The market capitalization experiences more volatility in the decade of 1965 – 1975 than the decade of 2005 – 2015. The lowest point for the decade of 1965 – 1975 is 1966, and the highest point for the same decade is 1965. The lowest point for the decade of 2005 – 2015 is 2008, and the highest point for the same decade is 2013. Comparing the medians, the market capitalization in the decade of 1965 – 1975 is in general greater than that of 2005 – 2015.



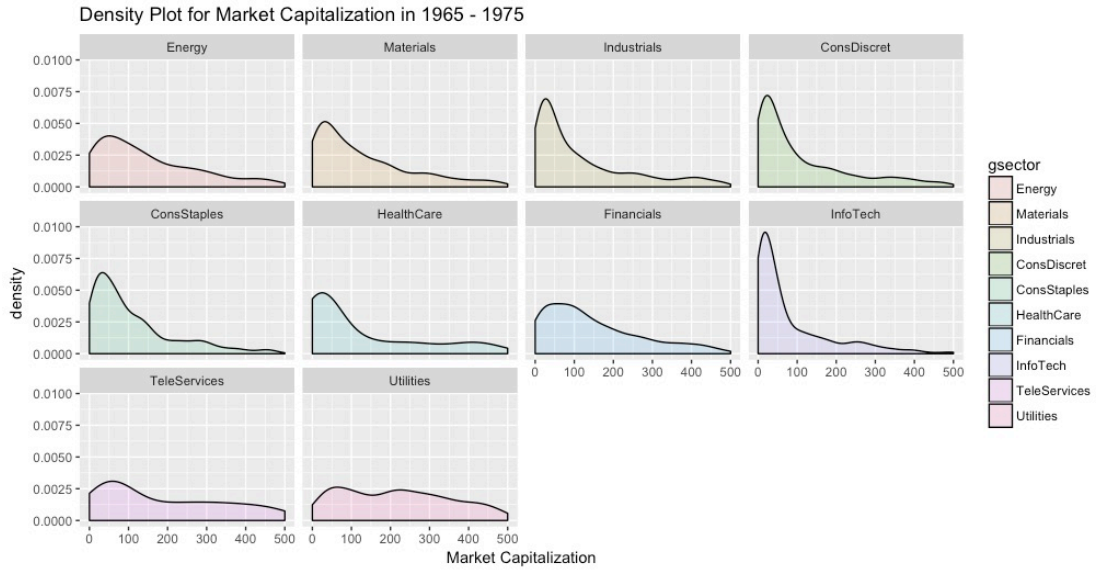


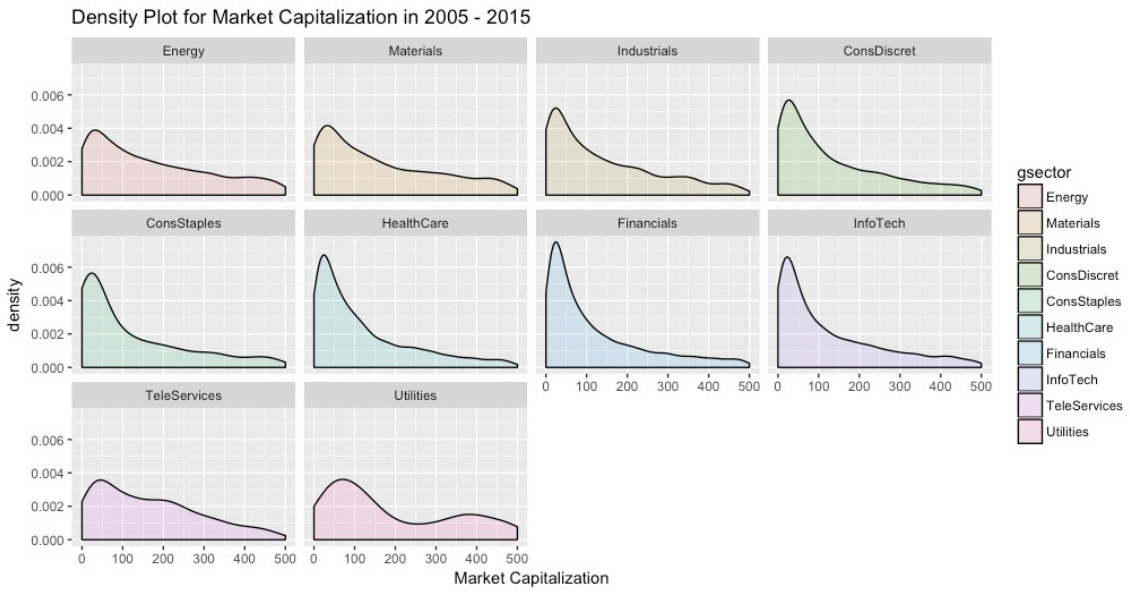
The density plot corresponds to the boxplot: more volatility is experienced in the decade of 1965 – 1975 compared to the decade of 2005 – 2015, and the market cap is overall greater in the decade of 1965 – 1975 than that of 2005 – 2015.





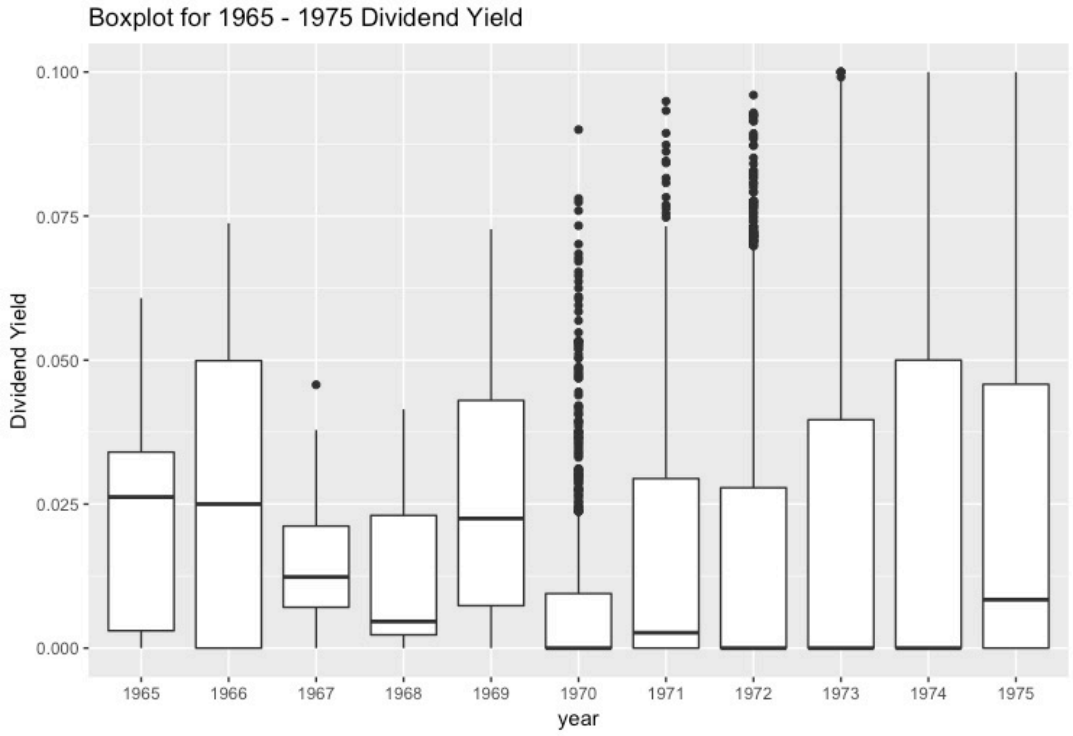
The market capitalizations differ by industry for both decades. For the decade of 1965 – 1975, Utilities, Telecommunication Services, and Financials obtain higher medians, but they also experience greater volatility. For the decade of 2005 – 2015, Energy, Materials, Telecommunication Services, and Utilities also obtain higher medians, and they also experience greater volatility. The industries that are of the lowest median in the decade of 1965 – 1975 are Health Care and Information Technologies, and the industries that are of the lowest median in the decade of 2005 – 2015.

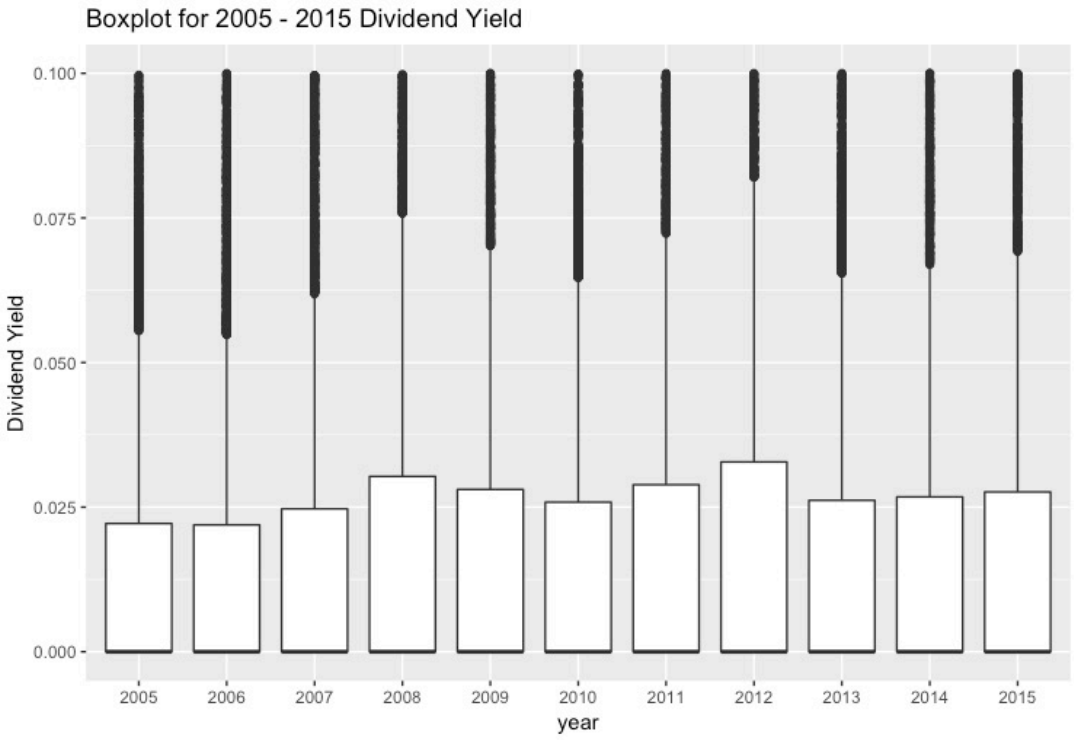




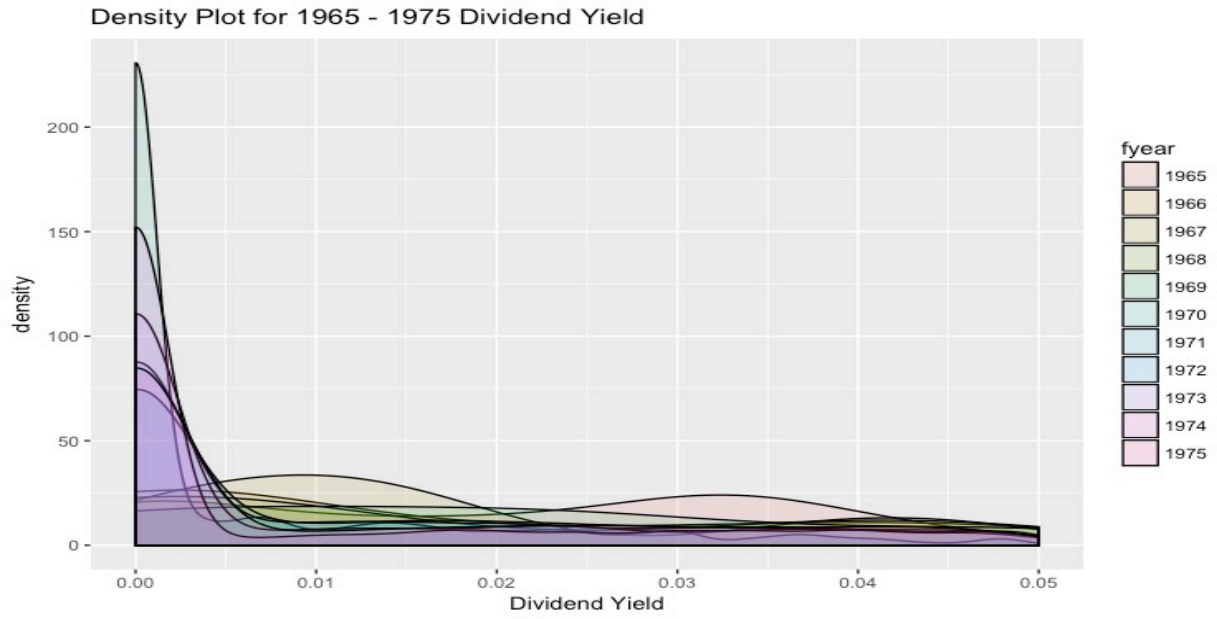
The trend shown in the density plot corresponds to the trend shown in the boxplot. In the decade of 1965 – 1975, Utilities, Telecommunications Services, and Financials are more widely spread, while Information Technologies, Consumer Discretionary, and Industrials are more concentrated on the left. In the decade of 2005 – 2015, Utilities and Telecommunication Services, Energy, and Materials are more widely spread, while all other sectors are more concentrated on the left.

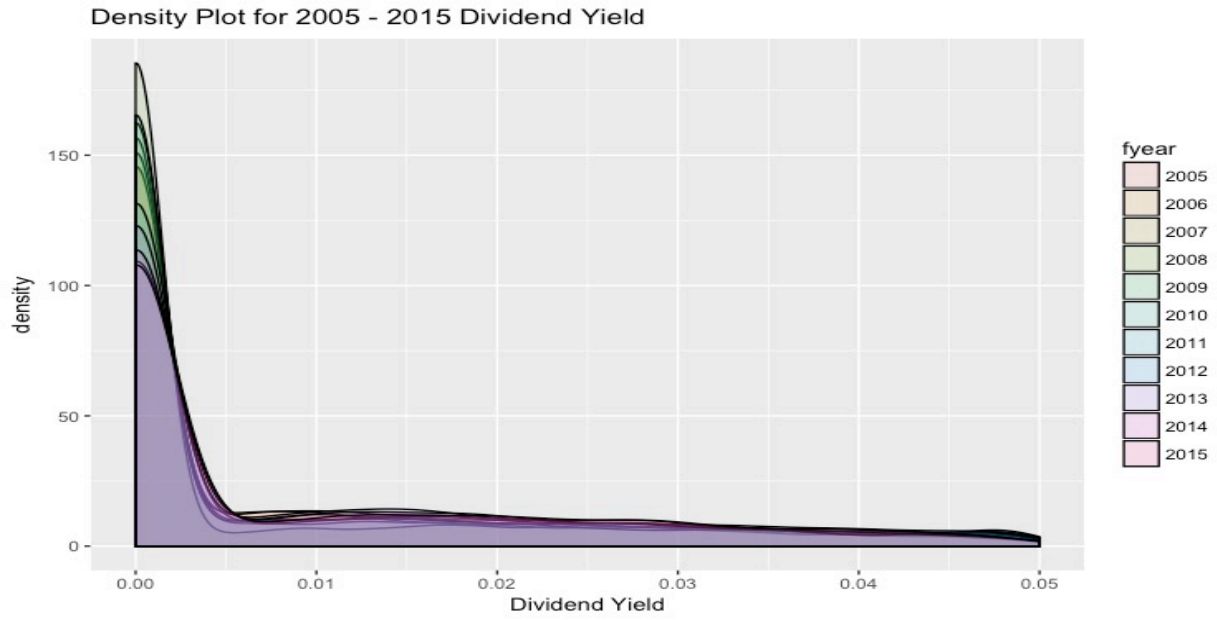
Dividend Yield



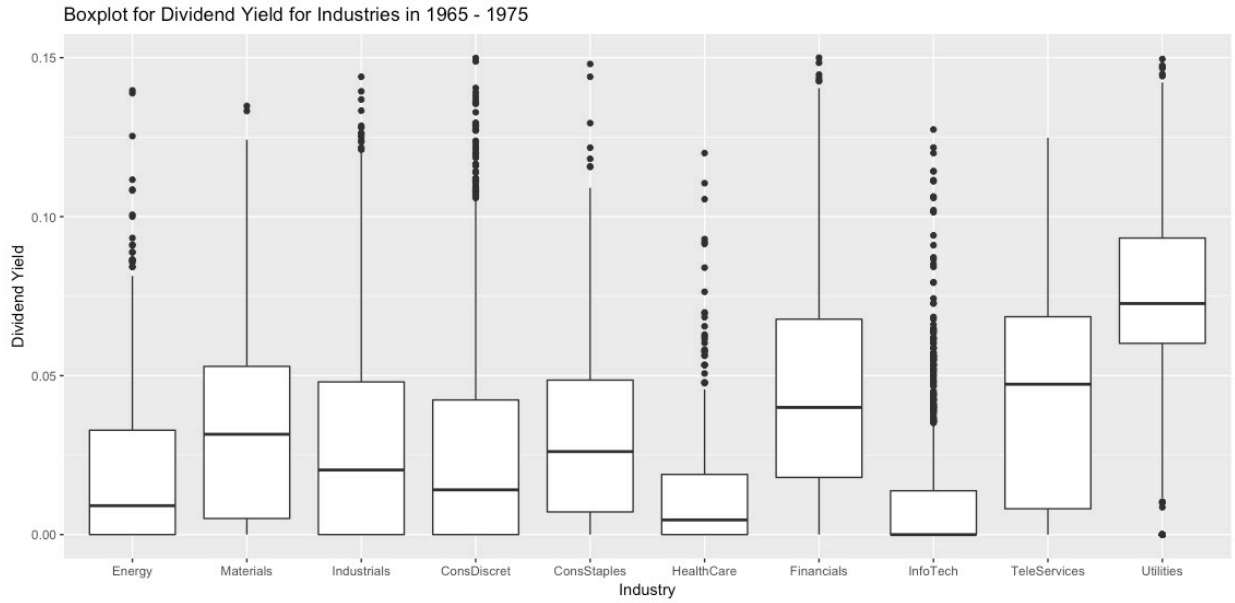


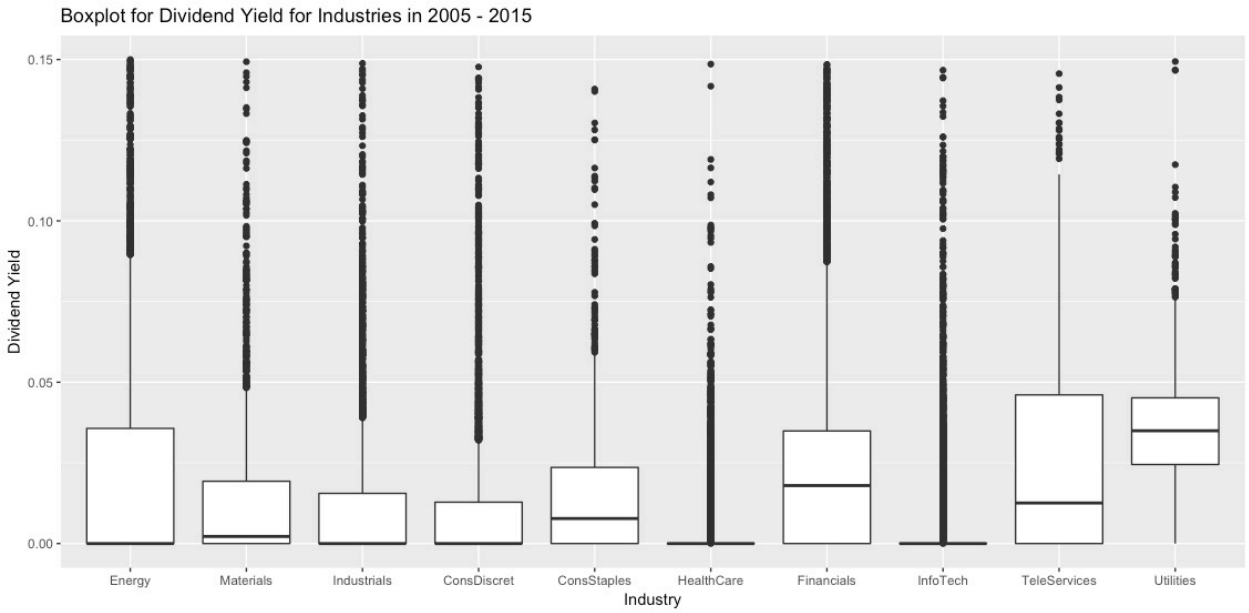
The volatility of dividend yield is greater in the decade of 1965 – 1975 than the decade of 2005 – 2015. The companies are much more likely to pay dividend in 1965 – 1975 than in 2005 – 2015. The economy is better in 1965 – 1969 compared to the economy in 1970 – 1974, since there are more companies paying dividend yield. However, median dividend yield values for 2005 – 2015 are overall 0, and their 25% percentile and 75% percentile are about the same.



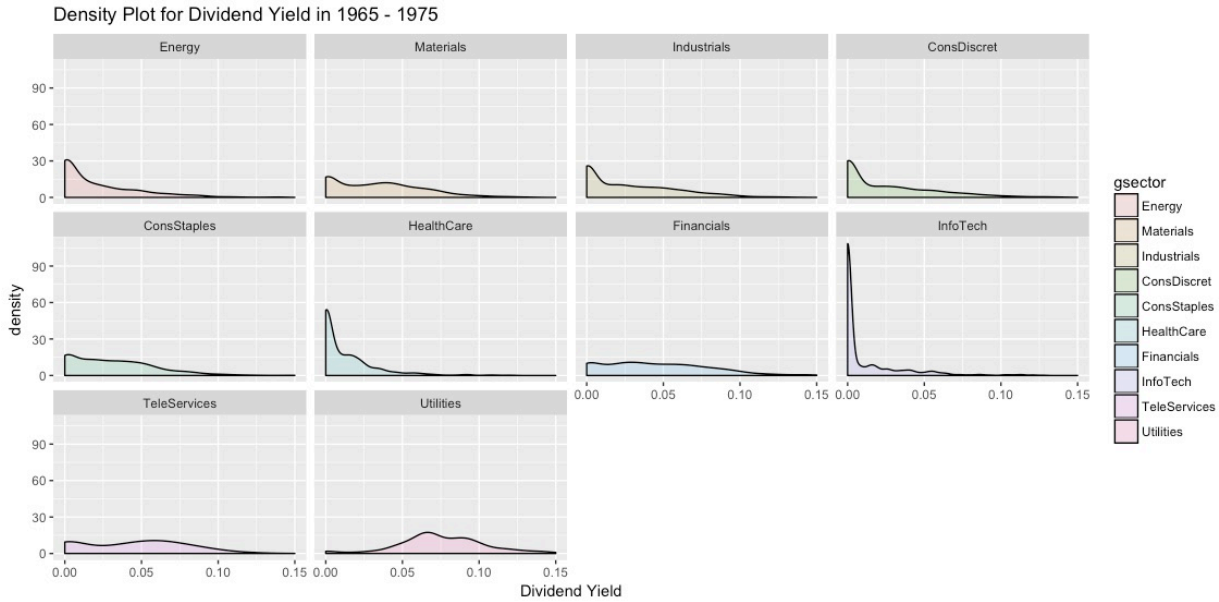


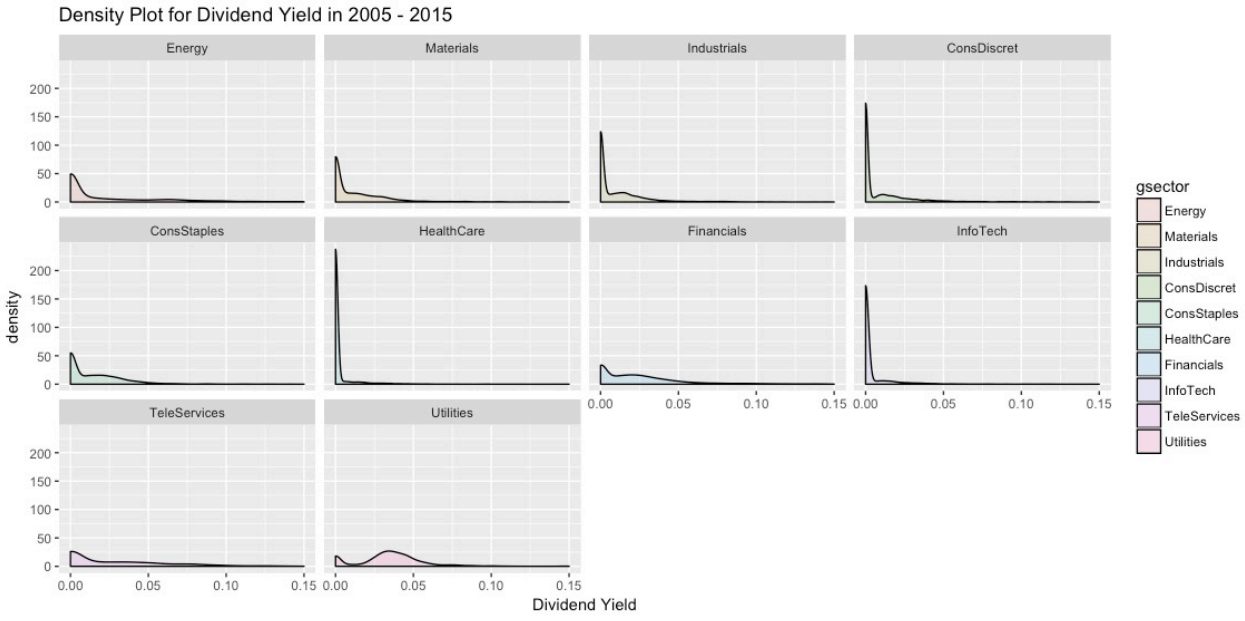
The density plots above correspond to the boxplots. The density distributions vary in greater extent in the decade of 1965 – 1975, than the decade in 2005 – 2015. Companies are less likely to pay dividend in 2005 – 2015 than in 1965 – 1975.





The boxplots for dividend yield by industry shows that Utilities, Telecommunication Services, and Financials are the sectors that yields greater dividend, while Health Care and Information Technologies are the sectors that yields lesser dividend. Overall the dividend yield is greater in 1965 – 1975 than in 2005 - 2015 throughout the industries.

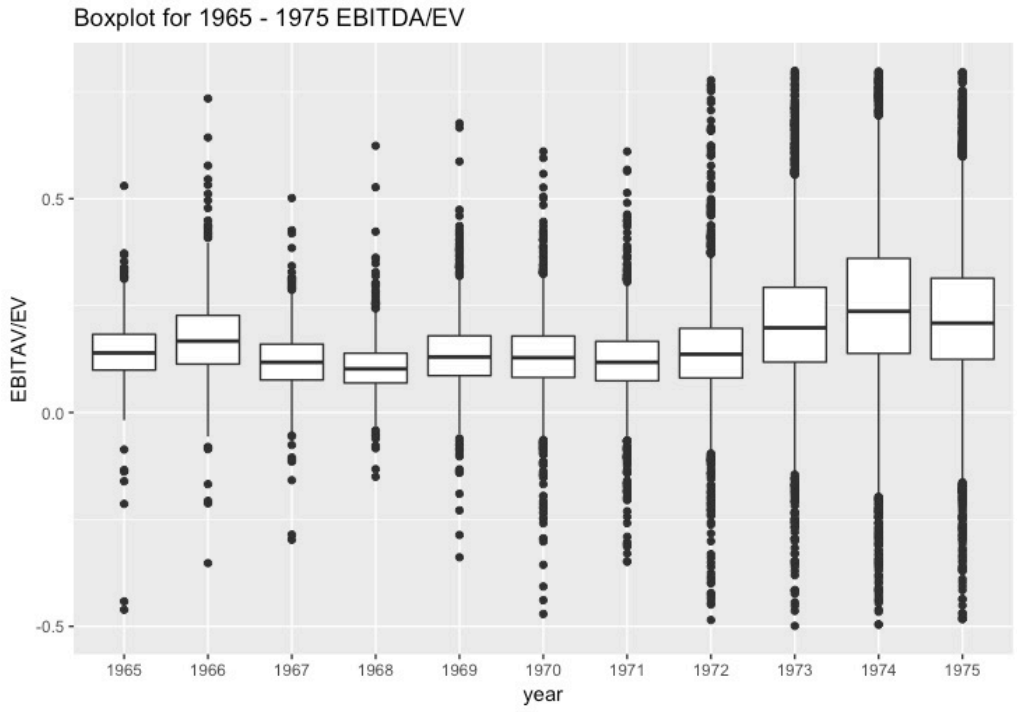


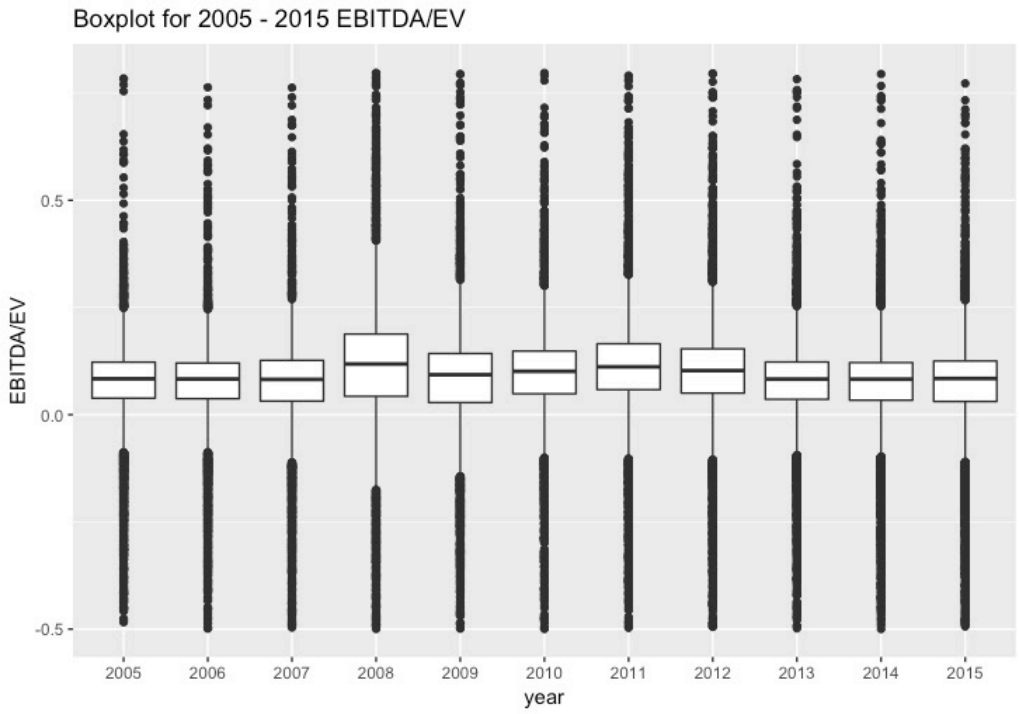


Looking at the density distributions, Health Care, Information Technologies, and Consumer Discretionary are the industries that most unlikely to yield dividend. In contrast, Utilities and

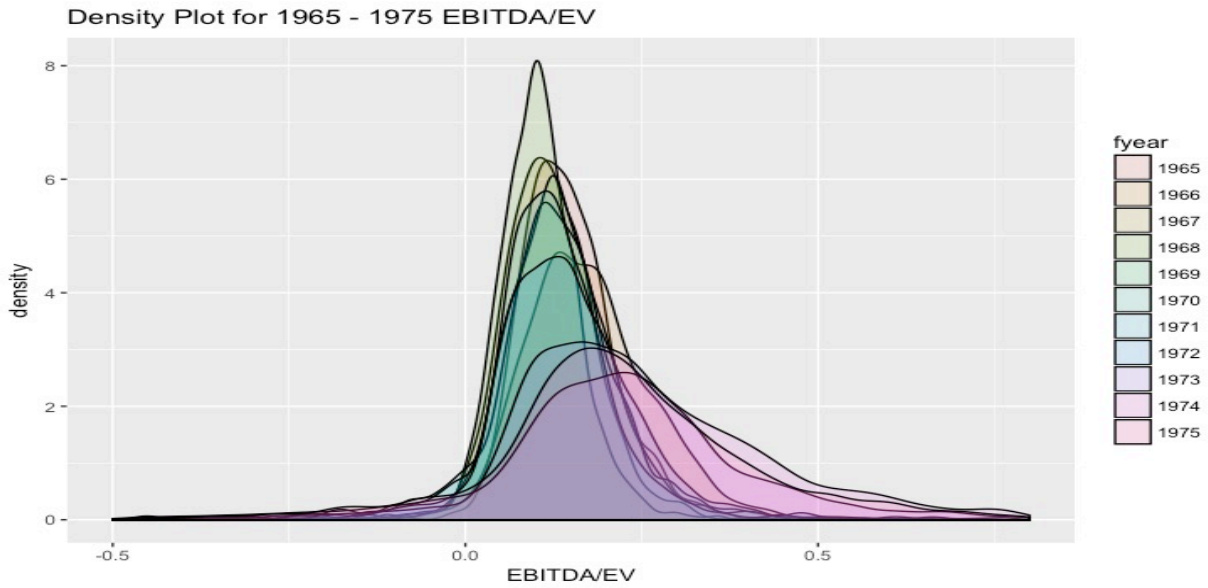
Telecommunication Services are the industries that are most likely to yield dividend. The companies are overall less likely to yield dividend in the decade 2005 – 2015 compared to the decade in 1965 – 1975.

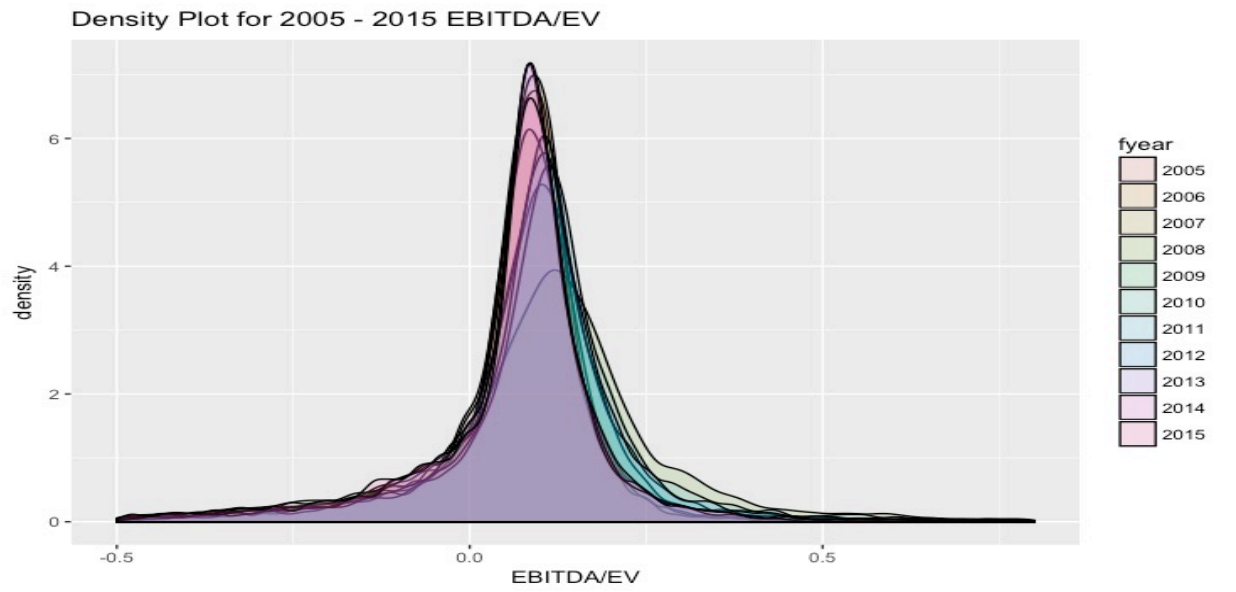
EBITDA/EV



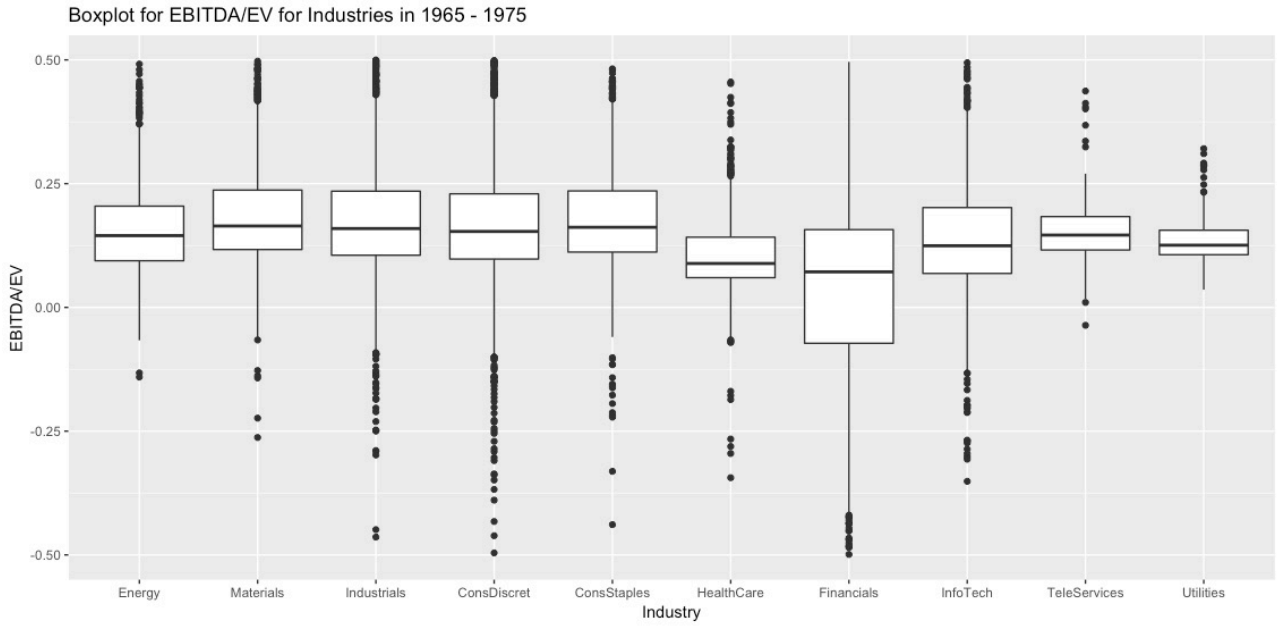


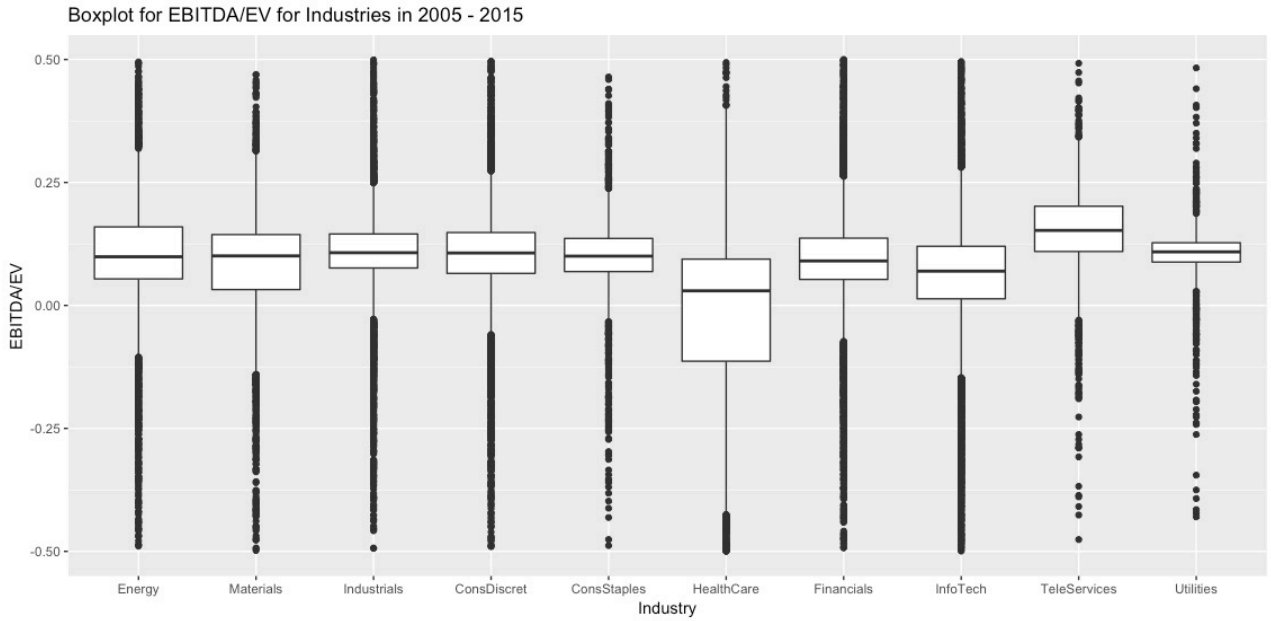
The boxplot trend for EBITDA/EV corresponds to the trend shown in Market Capitalization and Dividend Yield: the economy is more volatile in the decade 1965 – 1975 than the decade in 2005 – 2015. Interestingly, EBITDA/EV is highest for 2008 in the decade of 2005 – 2015, and highest for 1974 in the decade of 1965 - 1975. The year 2008 and the year 1974 are both the years of recessions.



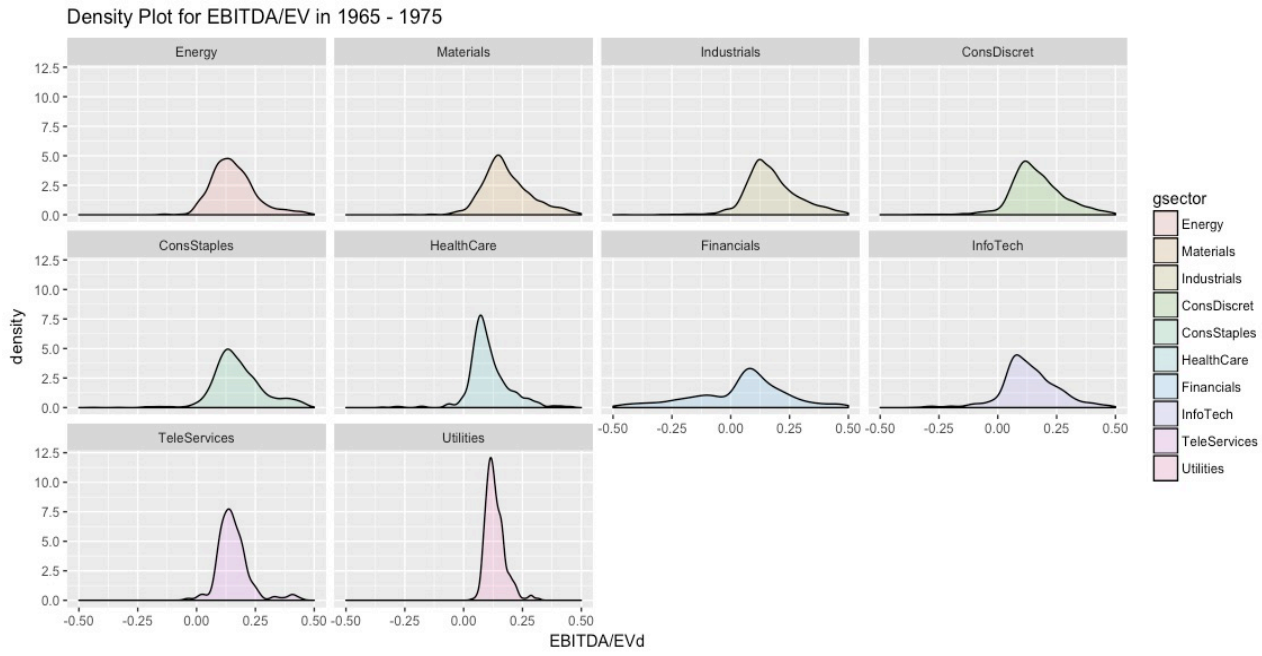


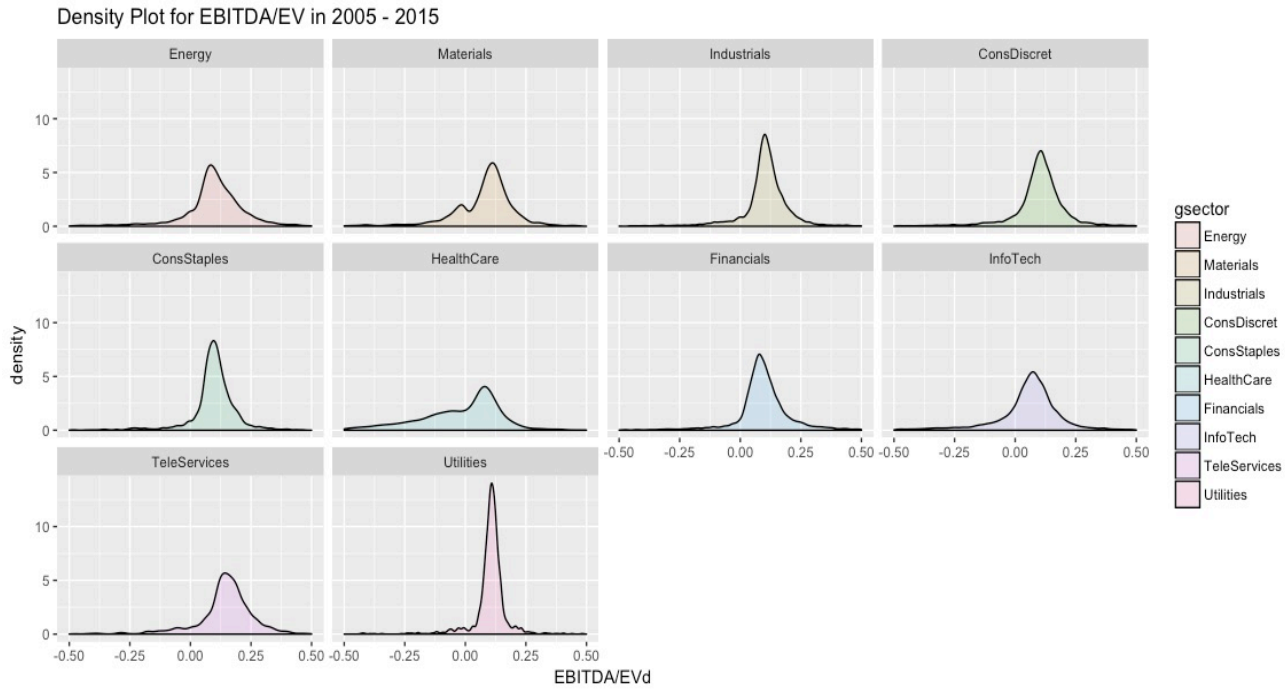
The density plots correspond to the volatility comparison shown in all other plots. EBITDA/EV is overall smaller in the decade of 2005 – 2015 than in the decade of 1965 – 1975.





The EBITDA/EV is in general higher in the decade 1965 – 1975 than in the decade 2005 – 2015. In the decade of 1965 – 1975, the volatility for the Financials sector is of the greatest, and its median is of the lowest. In the decade of 2005 – 2015, the volatility for the Health Care industry is of the greatest, and its median is also of the lowest. The Telecommunication Services sector performs the best in the decade of 2005 – 2015 compared to other industries. The volatility for the Utilities sector is always the smallest in both decades.





The trend shown in the density plots correspond to the trend shown in the other plots. One thing worth noticing is that the Materials sector in the decade 2005 - 2015 has two modes. The Financials sector is spread most widely in the decade 1965 – 1975, and the Health Care sector is spread most widely in the decade 2005 – 2015.

Conclusion

In this report, we cleaned the data and calculated three variables: Market Capitalization, Dividend Yield, and EBITDA/EV. We analyzed the annual Market Capitalization from 1925 to 2017, and we also compared the yearly trend between two decades (1965 – 1975, 2005 - 2015) for the above three factors both by year and by industry.

The market capitalization has grown over the years from 1925 to 2017, and for all years the distributions are dramatically rightly skewed. Most of the companies have the market cap over $ 50 million. Market Capitalization, Dividend Yield, and EBITDA/EV are three important indicators for the economy and finance. Comparing the factors for the two decades, the decade 1965 – 1975 in general experience more volatility in finance, while the decade 2005 – 2015

seems to be more stable in economy. Utilities usually perform better in all three factors, with higher median and lower volatility.

Reference

[1] Services, W. R. (n.d.). Retrieved February 28, 2018, from https://wrds-web.wharton.upenn.edu/wrds/query_forms/navigation.cfm?navId=118 Accessed Mar 2, 2018

[2] Staff, I. (2018, January 16). Market Capitalization Defined. Retrieved February 28, 2018, from https://www.investopedia.com/articles/basics/03/031703.asp Accessed Mar 2, 2018

[3] Staff, I. (2018, February 27). Market Capitalization. Retrieved March 02, 2018, from https://www.investopedia.com/terms/m/marketcapitalization.asp Accessed Mar 2, 2018

[4] Staff, I. (2018, February 27). Market Capitalization. Retrieved March 02, 2018, from https://www.investopedia.com/terms/m/marketcapitalization.asp Accessed Mar 2, 2018

[5] Shmoop Editorial Team. (2008, November 11). Economy in The 1960s. Retrieved March 02, 2018, from https://www.shmoop.com/1960s/economy.html Accessed Mar 2, 2018

[6] Forget the Algorithms and Start Cleaning Your Data. (2016, January 26). Retrieved March 02, 2018, from https://www.datanami.com/2014/03/26/forget_the_algorithms_and_start_cleaning_your_data/ Accessed Mar 2, 2018

Appendix (Written in R)

```
library(ggplot2)
library(plyr)
library(fBasics)

setwd("/Users/alisonzhang/Desktop/2018 Spring/STAT 686/Project 3")
data = read.csv("project3new.csv")

## Remove exchange code 0, 1, 3

data1 <- data[data$exchg != 0 & data$exchg != 1 & data$exchg != 3,]

## Calculate values: Div_y = dividend yield, ev = enterprise value,
## eb_v = EBITA/EV, mc = calculated market cap

data1$mc <- data1$csho * data1$prcc_f
data1$ev <- with(data1, mc + dltt + dlc + pstk - che)
data1$Div_y <- with(data1, dvpsp_f / prcc_f)
data1$eb_v <- with(data1, ebitda / ev)
factors <- c('che', 'csho', 'dlc', 'dltt', 'dvc', 'Div_y', 'ebitda',
             'pstk', 'exchg', 'prcc_c', 'dvpsp_f', 'prcc_f', 'mc', 'ev', 'Div_y', 'eb_v')
a <- length(factors)

## Find negative values

negative <- vector(mode = "numeric", length = a)

for (i in 1:a) {
  if (length(which(data1[,factors[i]] < 0)) > 0) {
    negative[i] = 'Neg'
  }
  else {
    negative[i] = 'No-Neg'
  }
}

## Show factors that have negative values

(t <- cbind(factors, negative))

## pstk (preferred stock) and dlc should not have negative values

(t <- which(data1$pstk < 0))
data1 <- data1[-t,]
```

```
(t <- which(data1$dlc < 0))
data1 <- data1[-t,]

## Coverage

coverage <- vector(mode = 'numeric', length = a)
for (i in 1:a) {
  coverage[i] <- 1-length(which(is.na(data1[,factors[i]])))/length(data1[,factors[i]])
}
(t <- cbind(coverage,factors))

## 50m mktcap analysis

## adding in year values for 1926 data
data2 = read.csv(file.choose())  #minipro3mktval.csv

View(data2)

data2$month = substr(data2$date,5,6)
data2$year = substr(data2$date, 1,4)

#removing negative price values
t <- which(data2$PRC < 0)
data2 <- data2[-t,]
# No negative value

###calculating mktcap for separate data source
data2$mc = data2$PRC * data2$SHROUT

###subsetting to december month values only
data2 <- data2[data2$month %in% 12,]
###removing NAs
data2 <- data2[-which(is.na(data2$mc)),]

## Count for inflation

data3 <- data2
cpi = read.csv("CPI.csv") # The base year is chained 1982-1984 = 100
cpi$AnnualAvg = cpi$AnnualAvg/100

dat_init = data.frame()
for(i in 1:(2017 - 1925+1)){
  dat1 = subset(data3, year == i + 1924)
  inflation = cpi[i, 2]
  dat1$mc = dat1$mc/inflation
```

```
  dat_init = rbind(dat_init, dat1)
}

dat_init$TICKER = NULL
dat_init$DIVAMT = NULL

dat_init$fyear = as.factor(as.character(dat_init$fyear))

gsecmc = completeFun(dat_init, 'gsector')

gsecmc$gsector = as.factor(gsecmc$gsector)
gsecmc = subset(gsecmc, gsector != '60')

gsecmc$gsector =  mapvalues(gsecmc$gsector,
                      from = c("10", "15", "20", "25", "30", "35", "40", "45",
                               "50", "55"),
                      to = c("Energy", 'Materials', 'Industrials', 'ConsDiscret',
                           'ConsStaples', 'HealthCare', 'Financials',
                           'InfoTech', 'TeleServices',
                           'Utilities'))

########## 1965 - 1975 MC ##########
datmc6575 = subset(dat_init, fyear %in% 1965:1975)
## Boxplot
ggplot(datmc6575, aes(x = fyear, y = mc)) + geom_boxplot() +
  ggtitle('Boxplot for 1965 - 1975 Market Capitalization') +
  ylab("Market Capitalization") + scale_y_continuous(limits = c(0, 500))

## Density plot
ggplot(datmc6575, aes(mc, fill = fyear)) + geom_density(alpha = 0.15) +
  ggtitle('Density Plot for 1965 - 1975 Market Capitalization') +
  scale_x_continuous(limits = c(0, 500)) + xlab("Market Capitalization")

gsecmc6575 = subset(gsecmc, fyear %in% 1965:1975)
##### By industry
ggplot(gsecmc6575, aes(x = gsector, y = mc)) + geom_boxplot() +
  ggtitle('Boxplot for Market Capitalization for Industries in 1965 - 1975') +
  ylab("Market Capitalization") + scale_y_continuous(limits = c(0, 500)) + xlab("Industry")

ggplot(gsecmc6575, aes(mc, fill = gsector)) + facet_wrap(~gsector) +
  geom_density(alpha = 0.15) +
  ggtitle('Density Plot for Market Capitalization in 1965 - 1975') +
  scale_x_continuous(limits = c(0, 500)) + xlab("Market Capitalization")

########## 2005 - 2015  MC ##########
```

```
datmc0515 = subset(dat_init, fyear %in% 2005:2015)
## Boxplot
ggplot(datmc0515, aes(x = fyear, y = mc)) + geom_boxplot() +
  ggtitle('Boxplot for 2005 - 2015 Market Capitalization') +
  ylab("Market Capitalization") + scale_y_continuous(limits = c(0, 500))

## Density plot
ggplot(gsecmc6575, aes(mc, fill = gsector)) + geom_density(alpha = 0.15) +
  ggtitle('Density Plot for Market Capitalization for Industries in 2005 - 2015') +
  scale_x_continuous(limits = c(0, 500)) + xlab("Market Capitalization")

gsecmc0515 = subset(gsecmc, fyear %in% 2005:2015)
##### By industry
ggplot(gsecmc0515, aes(x = gsector, y = mc)) + geom_boxplot() +
  ggtitle('Boxplot for Market Capitalization for Industries in 2005 - 2015') +
  ylab("Market Capitalization") + scale_y_continuous(limits = c(0, 500))

## Density plot
ggplot(gsecmc0515, aes(mc, fill = gsector)) + facet_wrap(~gsector) +
  geom_density(alpha = 0.15) +
  ggtitle('Density Plot for Market Capitalization in 2005 - 2015') +
  scale_x_continuous(limits = c(0, 500)) + xlab("Market Capitalization")

########## 1965 - 1975 dividend yield, EBITAV/EV ##########
completeFun <- function(data, desiredCols) {
  completeVec <- complete.cases(data[, desiredCols])
  return(data[completeVec, ])
}

data1$fyear = as.factor(data1$fyear)
datebv = completeFun(data1, 'eb_v')
datdy = completeFun(data1, 'Div_y')
gsec = completeFun(data1, 'gsector')
dim(gsec)[1]/dim(data1)[1]

Div_y

dat16575 = subset(datebv, fyear %in% 1965:1975)
############## EBITDA/EV
## Boxplot
ggplot(dat16575, aes(x = fyear, y = eb_v)) + geom_boxplot() +
  ggtitle('Boxplot for 1965 - 1975 EBITDA/EV') +
  ylab("EBITAV/EV") + scale_y_continuous(limits = c(-.5, .8)) + xlab("year")

## Density plot
```

```
ggplot(dat16575, aes(eb_v, fill = fyear)) + geom_density(alpha = 0.15) +
  ggtitle('Density Plot for 1965 - 1975 EBITDA/EV') +
  scale_x_continuous(limits = c(-0.5, .8)) + xlab("EBITDA/EV")

############## Dividend Yield
dat16575dy = subset(datdy, fyear %in% 1965:1975)
## Boxplot
ggplot(dat16575dy, aes(x = fyear, y = Div_y)) + geom_boxplot() +
  ggtitle('Boxplot for 1965 - 1975 Dividend Yield') +
  ylab("Dividend Yield") + scale_y_continuous(limits = c(0, 0.1)) + xlab("year")

## Density plot
ggplot(dat16575dy, aes(Div_y, fill = fyear)) + geom_density(alpha = 0.15) +
  ggtitle('Density Plot for 1965 - 1975 Dividend Yield') +
  scale_x_continuous(limits = c(0, 0.05)) + xlab("Dividend Yield")

########## 2005 - 2015  dividend yield, EBITAV ##########
dat10515 = subset(datebv, fyear %in% 2005:2015)
dat10515dy = subset(datdy, fyear %in% 2005:2015)
############## EBITDA/EV
## Boxplot
ggplot(dat10515, aes(x = fyear, y = eb_v)) + geom_boxplot() +
  ggtitle('Boxplot for 2005 - 2015 EBITDA/EV') +
  ylab("EBITDA/EV") + scale_y_continuous(limits = c(-.5, 0.8)) + xlab("year")

## Density plot
ggplot(dat10515, aes(eb_v, fill = fyear)) + geom_density(alpha = 0.15) +
  ggtitle('Density Plot for 2005 - 2015 EBITDA/EV') +
  scale_x_continuous(limits = c(-0.5, .8)) + xlab("EBITDA/EV")

############## Divident Yield
## Boxplot
ggplot(dat10515dy, aes(x = fyear, y = Div_y)) + geom_boxplot() +
  ggtitle('Boxplot for 2005 - 2015 Dividend Yield') +
  ylab("Dividend Yield") + scale_y_continuous(limits = c(0, .1)) + xlab("year")

## Density plot
ggplot(dat10515dy, aes(Div_y, fill = fyear)) + geom_density(alpha = 0.15) +
  ggtitle('Density Plot for 2005 - 2015 Dividend Yield') +
  scale_x_continuous(limits = c(0, .05)) + xlab("Dividend Yield")

########## By Industry ##########
gsec$gsector = as.factor(gsec$gsector)
gsec = subset(gsec, gsector != '60')
```

```
gsec$gsector =  mapvalues(gsec$gsector,
        from = c("10", "15", "20", "25", "30", "35", "40", "45",
                 "50", "55"),
        to = c("Energy", 'Materials', 'Industrials', 'ConsDiscret',
               'ConsStaples', 'HealthCare', 'Financials',
               'InfoTech', 'TeleServices',
               'Utilities'))

## Count for inflation
industry6575 = subset(gsec, fyear %in% 1965:1975)
industry0515 = subset(gsec, fyear %in% 2005:2015)
## EBITDA/EV
ggplot(industry6575, aes(x = gsector, y = eb_v)) + geom_boxplot() +
  ggtitle('Boxplot for EBITDA/EV for Industries in 1965 - 1975') +
  ylab("EBITDA/EV") + scale_y_continuous(limits = c(-0.5, 0.5)) + xlab("Industry")
ggplot(industry6575, aes(eb_v, fill = gsector)) + facet_wrap(~gsector) +
  geom_density(alpha = 0.15) +
  ggtitle('Density Plot for EBITDA/EV in 1965 - 1975') +
  scale_x_continuous(limits = c(-0.5, .5)) + xlab("EBITDA/EVd")

ggplot(industry0515, aes(x = gsector, y = eb_v)) + geom_boxplot() +
  ggtitle('Boxplot for EBITDA/EV for Industries in 2005 - 2015') +
  ylab("EBITDA/EV") + scale_y_continuous(limits = c(-0.5, 0.5)) + xlab("Industry")
ggplot(industry0515, aes(eb_v, fill = gsector)) + facet_wrap(~gsector) +
  geom_density(alpha = 0.15) +
  ggtitle('Density Plot for EBITDA/EV in 2005 - 2015') +
  scale_x_continuous(limits = c(-0.5, .5)) + xlab("EBITDA/EVd")

## Dividend Yield
ggplot(industry6575, aes(x = gsector, y = Div_y)) + geom_boxplot() +
  ggtitle('Boxplot for Dividend Yield for Industries in 1965 - 1975') +
  ylab("Dividend Yield") + scale_y_continuous(limits = c(0, 0.15)) + xlab("Industry")
ggplot(industry6575, aes(Div_y, fill = gsector)) + facet_wrap(~gsector) +
  geom_density(alpha = 0.15) +
  ggtitle('Density Plot for Dividend Yield in 1965 - 1975') +
  scale_x_continuous(limits = c(0, .15)) + xlab("Dividend Yield")

ggplot(industry0515, aes(x = gsector, y = Div_y)) + geom_boxplot() +
  ggtitle('Boxplot for Dividend Yield for Industries in 2005 - 2015') +
  ylab("Dividend Yield") + scale_y_continuous(limits = c(0, 0.15)) + xlab("Industry")
ggplot(industry0515, aes(Div_y, fill = gsector)) + facet_wrap(~gsector) +
  geom_density(alpha = 0.15) +
  ggtitle('Density Plot for Dividend Yield in 2005 - 2015') +
  scale_x_continuous(limits = c(0, .15)) + xlab("Dividend Yield")
```