All Data EDA

Description of data cleaning before this file:

1. Load and combine credit rating datasets to get unique ratings, company/tickers, and rating issue dates (unique key is credit rating, rating agency, and issuance date). Limit to S&P ratings from 2010-2016.
2. Load dataset of earnings call transcripts, dates, year + quarter of statement releases, companies, and sectors (unique key is company by earnings call date).
3. Load tabular financial statement datasets (unique key is company, year, quarter).
4. Use earnings call dates to transform dataset of credit ratings so there is one rating at each earnings call date. The key assumption is that a rating stays the same until a new rating is issued. Use leads to get rating at next earnings call date, date of next earnings call, rating 2 earnings call dates ahead, and date of that earnings call.
5. Inner join earnings call data with credit rating data. Use year and quarter from earnings calls to inner join with financial statement data. This produces the `all_data` dataset.

Setup - Sample Path and Packages

In [ ]:

```
# Flag for if you are running this on the sample dataset
# Sample comprises 100 earnings calls (transcripts included)
# Full data comprises 4532 earnings calls (transcripts included)
sample = False
# Modify this path as needed to run on your machine
sample_path = r'all_data_sample.csv'
```

In [ ]:

```
# Packages
import pandas as pd
import matplotlib.pyplot as plt
```

Code

In [ ]:
```
# Load in sample csv, or full parquet file
# Use inputted sample path, or ~\Box\STAT 222 Capstone\Intermediate Data\all_data.parquet
if sample:
    df = pd.read_csv(sample_path)
else:
    df = pd.read_parquet(r'all_data.parquet')
df
```

Out[ ]:

	ticker	earnings_call_date	next_earnings_call_date	rating_on_next_earnings_call_date	days_until_next_earnings_call	Rating	Rating Agency Name	r
0	ABBV	2014-07-25	2014-10-31	A	98.0	A	Standard & Poor's Ratings Services	
1	ABBV	2014-10-31	2015-01-30	A	91.0	A	Standard & Poor's Ratings Services	
2	ABBV	2015-01-30	2015-04-23	A	83.0	A	Standard & Poor's Ratings Services	
3	ABBV	2015-04-23	2015-07-24	A	92.0	A	Standard & Poor's Ratings Services	
4	ABBV	2015-07-24	2015-10-30	A	98.0	A	Standard & Poor's Ratings Services	
...	...	...	...	...	...	...	...	
4527	ZTS	2015-11-03	2016-02-16	BBB	105.0	BBB	Standard & Poor's Ratings Services	
4528	ZTS	2016-02-16	2016-05-04	BBB	78.0	BBB	Standard & Poor's Ratings Services	
4529	ZTS	2016-05-04	2016-08-03	BBB	91.0	BBB	Standard & Poor's Ratings Services	
4530	ZTS	2016-08-03	2016-11-02	BBB	91.0	BBB	Standard & Poor's	

	ticker	earnings_call_date	next_earnings_call_date	rating_on_next_earnings_call_date	days_until_next_earnings_call	Rating	Rating Agency Name	r
							Ratings Services	
4531	ZTS	2016-11-02	None	None	NaN	BBB	Standard & Poor's Ratings Services	

4532 rows × 158 columns

In [ ]:
```
## summary of the raw data

# Summarize all numeric columns
# use describe method, transpose, and print all rows
# round to two decimal places, no scientific notation, commas for thousands
pd.options.display.float_format = '{:,.2f}'.format
# pandas setting to display all rows
pd.set_option('display.max_rows', None)
df.describe().T
```

Out[ ]:

	count	mean	std	min	25%	
days_until_next_earnings_call	4,213.00	93.27	24.64	26.00	88.00	
Rating Rank AAA is 10	4,532.00	6.75	1.28	2.00	6.00	
Change in Rating	3,587.00	0.02	0.53	-2.00	0.00	
Year	4,532.00	2,013.28	1.64	2,010.00	2,012.00	
year	4,532.00	2,014.29	1.57	2,010.00	2,013.00	
quarter	4,532.00	2.52	1.12	1.00	2.00	
cik	4,532.00	709,311.50	549,506.65	1,800.00	75,677.00	8
calendarYear	4,532.00	2,014.29	1.57	2,010.00	2,013.00	
period	4,532.00	2.52	1.12	1.00	2.00	
cashAndCashEquivalents	4,532.00	1,535,929,224.37	5,523,419,258.67	0.00	133,942,000.00	432,8
shortTermInvestments	4,532.00	4,988,623,266.19	210,069,282,993.89	-515,000,000.00	0.00	
cashAndShortTermInvestments	4,532.00	2,120,898,494.52	7,525,146,382.75	0.00	134,574,250.00	478,7
netReceivables	4,532.00	1,921,999,022.63	3,524,348,166.20	-4,199,600,000.00	293,770,250.00	703,7
inventory	4,532.00	1,435,297,437.67	2,598,581,290.63	-27,358,000,000.00	166,375,000.00	575,9
otherCurrentAssets	4,532.00	630,328,755.44	1,757,252,133.71	-98,000,000.00	51,598,750.00	146,8
totalCurrentAssets	4,532.00	6,088,233,743.85	11,751,660,628.07	8,000.00	1,053,683,250.00	2,397,3
propertyPlantEquipmentNet	4,532.00	6,728,974,362.94	21,093,978,673.90	-25,785,000.00	577,575,000.00	1,617,2
goodwill	4,532.00	3,039,435,731.93	8,921,591,906.45	-202,702,100,000.00	154,275,000.00	935,9
intangibleAssets	4,532.00	2,325,212,577.20	8,903,990,647.55	0.00	14,045,000.00	295,0
goodwillAndIntangibleAssets	4,532.00	4,969,093,794.34	27,449,282,248.43	-1,618,944,000,000.00	330,725,000.00	1,427,4
longTermInvestments	4,532.00	994,677,505.72	6,987,172,497.54	-61,869,718,000.00	0.00	4
taxAssets	4,532.00	-5,013,282,441.79	102,067,956,917.95	-2,310,712,000,000.00	0.00	51,6
otherNonCurrentAssets	4,532.00	6,587,453,343.67	99,099,039,111.32	-68,662,600,000.00	45,429,000.00	202,0
totalNonCurrentAssets	4,532.00	14,274,745,532.44	31,580,342,771.36	0.00	2,071,209,250.00	5,075,8
otherAssets	4,532.00	37,279,285.53	644,947,874.93	-3,918,000,000.00	0.00	

	count	mean	std	min	25%	
totalAssets	4,532.00	20,403,086,673.02	41,000,881,771.28	8,000.00	3,496,069,750.00	7,838,8
accountPayables	4,532.00	1,678,428,355.08	4,262,431,485.12	-97,284,000.00	149,200,000.00	433,4
shortTermDebt	4,532.00	676,755,089.19	1,787,885,632.54	-8,741,000,000.00	7,787,250.00	89,0
taxPayables	4,532.00	252,751,663.90	4,584,983,498.83	-84,089,000.00	0.00	2,0
deferredRevenue	4,532.00	282,976,647.86	2,494,964,471.94	-45,145,000,000.00	0.00	29,9
otherCurrentLiabilities	4,532.00	1,364,017,965.60	3,593,550,562.73	-18,528,000,000.00	143,771,500.00	390,8
totalCurrentLiabilities	4,532.00	4,012,051,700.27	7,756,553,938.35	0.00	540,770,500.00	1,345,3
longTermDebt	4,532.00	4,773,257,541.75	7,464,001,970.49	0.00	970,825,000.00	2,339,1
deferredRevenueNonCurrent	4,532.00	555,861,959.34	6,673,727,887.05	-43,874,000,000.00	0.00	
deferredTaxLiabilitiesNonCurrent	4,532.00	1,146,091,788.93	3,917,907,632.54	-1,300,000,000.00	0.00	143,6
otherNonCurrentLiabilities	4,532.00	1,033,027,893.29	6,811,650,367.74	-233,364,494,000.00	111,630,500.00	382,5
totalNonCurrentLiabilities	4,532.00	7,496,271,646.01	13,084,046,755.07	0.00	1,437,500,000.00	3,323,4
otherLiabilities	4,532.00	234,123,409.75	3,268,113,305.73	-2,000,224,000.00	0.00	
capitalLeaseObligations	4,532.00	3,930,590,347.10	181,765,681,032.53	0.00	0.00	
totalLiabilities	4,532.00	11,755,157,553.91	20,449,574,343.05	0.00	2,170,291,750.00	4,967,8
preferredStock	4,532.00	30,252,917.70	236,416,285.14	0.00	0.00	
commonStock	4,532.00	3,329,870,100.17	148,117,367,392.35	-725,200,000.00	849,750.00	4,7
retainedEarnings	4,532.00	8,065,496,764.51	30,540,236,426.00	-17,049,000,000.00	324,558,000.00	1,691,5
accumulatedOtherComprehensiveIncomeLoss	4,532.00	-664,485,264.58	2,009,909,149.91	-24,336,000,000.00	-502,925,000.00	-133,0
othertotalStockholdersEquity	4,532.00	-316,108,664.45	152,960,148,216.58	-9,811,004,344,000.00	-110,156,250.00	418,8
totalStockholdersEquity	4,532.00	8,419,207,530.98	22,006,447,751.14	-13,616,000,000.00	1,045,520,500.00	2,442,3
totalEquity	4,532.00	8,442,168,854.46	22,015,882,160.19	-13,616,000,000.00	1,045,520,500.00	2,442,3
totalLiabilitiesAndStockholdersEquity	4,532.00	20,393,655,691.78	41,003,408,217.68	8,000.00	3,496,069,750.00	7,838,8
minorityInterest	4,532.00	419,426,378.95	2,598,610,015.15	-286,000,000.00	0.00	4,8
totalLiabilitiesAndTotalEquity	4,532.00	20,393,655,691.78	41,003,408,217.68	8,000.00	3,496,069,750.00	7,838,8

	count	mean	std	min	25%	
totalInvestments	4,532.00	5,993,122,725.57	210,191,911,022.34	-27,337,000,000.00	0.00	24,0
totalDebt	4,532.00	5,405,676,109.32	8,628,977,423.40	0.00	1,068,780,250.00	2,550,0
netDebt	4,532.00	3,869,583,041.62	7,424,227,884.30	-53,785,000,000.00	573,851,405.07	1,687,4
cik_cash_flow_statement	4,532.00	709,311.50	549,506.65	1,800.00	75,677.00	8
calendarYear_cash_flow_statement	4,532.00	2,014.29	1.57	2,010.00	2,013.00	
period_cash_flow_statement	4,532.00	2.52	1.12	1.00	2.00	
netIncome	4,532.00	302,995,573.48	1,041,969,993.59	-7,870,000,000.00	22,000,000.00	84,6
depreciationAndAmortization	4,532.00	223,932,147.44	584,681,733.18	-876,000,000.00	26,700,000.00	67,9
deferredIncomeTax	4,532.00	676,895,673.19	31,297,659,476.39	-53,034,000,000.00	-9,994,000.00	
stockBasedCompensation	4,532.00	32,500,847.64	383,138,471.64	-11,760,000,000.00	2,255,000.00	7,0
changeInWorkingCapital	4,532.00	-12,002,234.47	476,056,840.81	-5,424,000,000.00	-79,008,750.00	-4,0
accountsReceivables	4,532.00	-37,435,887.34	491,842,157.04	-11,553,000,000.00	-33,925,000.00	
inventory_cash_flow_statement	4,532.00	-10,036,094.41	307,450,762.43	-5,580,000,000.00	-24,003,250.00	
accountsPayables	4,532.00	321,600,846.26	27,470,206,813.37	-252,402,000,000.00	-14,299,250.00	
otherWorkingCapital	4,532.00	-285,813,132.41	27,462,190,803.97	-1,788,851,160,000.00	-29,934,750.00	
otherNonCashItems	4,532.00	-655,251,768.40	31,284,658,871.87	-1,848,719,007,000.00	-8,751,750.00	2,5
netCashProvidedByOperatingActivities	4,532.00	567,951,074.49	1,494,377,506.79	-3,632,000,000.00	56,124,750.00	175,6
investmentsInPropertyPlantAndEquipment	4,532.00	-298,219,160.97	964,072,855.03	-29,876,000,000.00	-222,250,000.00	-65,2
acquisitionsNet	4,532.00	-250,093,403.38	10,158,886,497.41	-671,827,500,000.00	-18,764,750.00	
purchasesOfInvestments	4,532.00	-3,984,689,962.39	193,699,941,835.12	-11,997,654,000,000.00	-9,925,000.00	
salesMaturitiesOfInvestments	4,532.00	3,455,908,870.96	157,025,636,938.15	-3,419,000,000.00	0.00	
otherInvestingActivites	4,532.00	674,424,669.55	45,541,049,692.38	-61,468,000,000.00	-4,032,500.00	1
netCashUsedForInvestingActivites	4,532.00	-417,764,666.07	1,647,947,792.53	-61,468,000,000.00	-301,550,000.00	-77,6
debtRepayment	4,532.00	-639,091,462.99	13,782,655,127.75	-852,427,427,000.00	-309,500,000.00	-41,3
commonStockIssued	4,532.00	466,715,172.07	16,663,884,471.91	-3,572,000,000.00	0.00	

	count	mean	std	min	25%	
commonStockRepurchased	4,532.00	-542,098,530.08	31,403,806,519.85	-2,086,545,366,000.00	-94,925,000.00	-2,2
dividendsPaid	4,532.00	-211,122,312.78	2,042,618,962.79	-98,706,821,000.00	-98,350,000.00	-24,6
otherFinancingActivites	4,532.00	793,501,339.24	27,642,923,224.07	-67,428,702,000.00	-2,000,000.00	8,4
netCashUsedProvidedByFinancingActivities	4,532.00	-125,704,187.53	1,226,280,898.58	-17,235,000,000.00	-235,121,750.00	-50,3
effectOfForexChangesOnCash	4,532.00	-1,860,099.42	298,283,302.45	-1,201,000,000.00	-4,466,500.00	
netChangeInCash	4,532.00	25,396,807.52	1,229,136,221.02	-24,572,000,000.00	-74,231,000.00	8
cashAtEndOfPeriod	4,532.00	1,539,855,313.77	5,525,465,099.40	18,000.00	133,696,250.00	436,1
cashAtBeginningOfPeriod	4,532.00	1,519,102,350.03	5,311,680,525.31	-2,556,000,000.00	132,530,000.00	431,6
operatingCashFlow	4,532.00	567,951,074.49	1,494,377,506.79	-3,632,000,000.00	56,124,750.00	175,6
capitalExpenditure	4,532.00	-298,219,160.97	964,072,855.03	-29,876,000,000.00	-222,250,000.00	-65,2
freeCashFlow	4,532.00	269,731,913.52	1,253,828,258.94	-28,286,000,000.00	-6,923,000.00	76,0
cik_income_statement	4,532.00	709,311.50	549,506.65	1,800.00	75,677.00	8
calendarYear_income_statement	4,532.00	2,014.29	1.57	2,010.00	2,013.00	
period_income_statement	4,532.00	2.52	1.12	1.00	2.00	
revenue	4,532.00	4,204,178,860.63	8,899,635,853.87	-6,646,000,000.00	632,090,750.00	1,539,9
costOfRevenue	4,532.00	2,855,652,233.55	6,575,020,857.80	-4,822,000,000.00	361,758,500.00	933,3
grossProfit	4,532.00	1,350,604,620.01	3,066,968,631.90	-7,195,000,000.00	187,950,000.00	434,8
grossProfitRatio	4,532.00	0.35	0.39	-17.88	0.20	
researchAndDevelopmentExpenses	4,532.00	87,732,766.97	362,379,838.43	-214,000,000.00	0.00	
generalAndAdministrativeExpenses	4,532.00	292,200,330.66	775,412,436.44	-2,738,500,000.00	0.00	53,7
sellingAndMarketingExpenses	4,532.00	96,685,322.94	387,151,964.22	-3,003,000,000.00	0.00	
sellingGeneralAndAdministrativeExpenses	4,532.00	542,701,563.99	1,071,459,892.15	-11,283,000,000.00	67,340,500.00	186,1
otherExpenses	4,532.00	7,126,960,541.04	331,745,140,306.39	-3,318,538,000.00	-536,500.00	3
operatingExpenses	4,532.00	873,182,276.73	1,967,764,774.15	-2,976,356,000.00	105,150,000.00	253,8
costAndExpenses	4,532.00	3,722,391,730.50	7,986,868,404.87	-5,794,000,000.00	539,946,000.00	1,354,5

	count	mean	std	min	25%	
interestIncome	4,532.00	12,209,526.00	297,194,211.59	-204,000,000.00	0.00	
interestExpense	4,532.00	51,561,580.00	94,690,315.36	-669,000,000.00	11,506,500.00	24,9
depreciationAndAmortization_income_statement	4,532.00	214,845,085.78	550,674,715.81	-1,550,000,000.00	27,000,000.00	68,3
ebitda	4,532.00	678,592,874.14	1,634,603,182.87	-7,390,000,000.00	92,460,000.00	232,0
ebitdaratio	4,532.00	0.18	0.71	-44.29	0.10	
operatingIncome	4,532.00	450,015,000.41	1,240,158,158.24	-7,390,000,000.00	54,515,750.00	155,
operatingIncomeRatio	4,532.00	0.09	0.50	-20.35	0.05	
totalOtherIncomeExpensesNet	4,532.00	-13,008,621.64	504,315,281.82	-6,511,000,000.00	-24,058,750.00	-2,3
incomeBeforeTax	4,532.00	426,471,164.85	1,479,028,977.89	-8,668,000,000.00	32,185,250.00	116,3
incomeBeforeTaxRatio	4,532.00	0.07	0.42	-9.38	0.03	
incomeTaxExpense	4,532.00	123,259,510.94	485,200,991.80	-3,179,000,000.00	6,321,750.00	33,0
netIncome_income_statement	4,532.00	303,307,353.12	1,038,326,149.05	-7,213,000,000.00	22,461,500.00	86,0
netIncomeRatio	4,532.00	0.05	0.54	-8.88	0.02	
eps	4,532.00	0.54	3.99	-156.36	0.21	
epsdiluted	4,532.00	0.55	3.81	-156.36	0.21	
weightedAverageShsOut	4,532.00	470,251,451.86	1,157,062,911.64	0.00	75,697,500.00	152,4
weightedAverageShsOutDil	4,532.00	477,493,023.07	1,175,446,913.33	0.00	76,834,750.00	156,4

Note: currently investigating issues with year variables being outside of the desired range of the data.

Also need to investigate financial statement variables in quadrillions, trillions, etc. and how to handle them.

In [ ]:
```
# Revert to default settings
pd.reset_option('display.float_format')
pd.reset_option('display.max_rows')
```

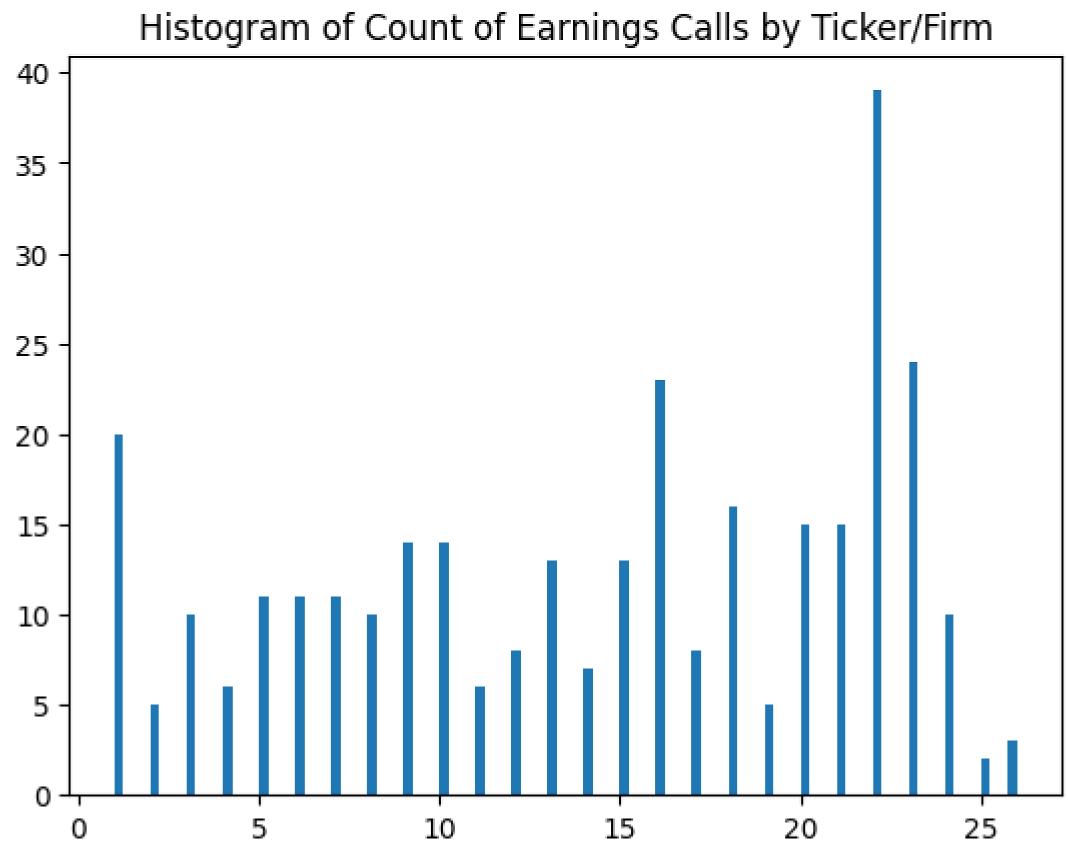
In [ ]:
```
# Check data is unique on ticker by earnings_call_date
df['ticker_earnings_date'] = df['ticker'] + '_' + df['earnings_call_date'].astype(str)
df['ticker_earnings_date'].value_counts().max()
```

Out[ ]: 1

In [ ]:
```
# Number of unique firms (identified by ticker)
df['ticker'].nunique()
```

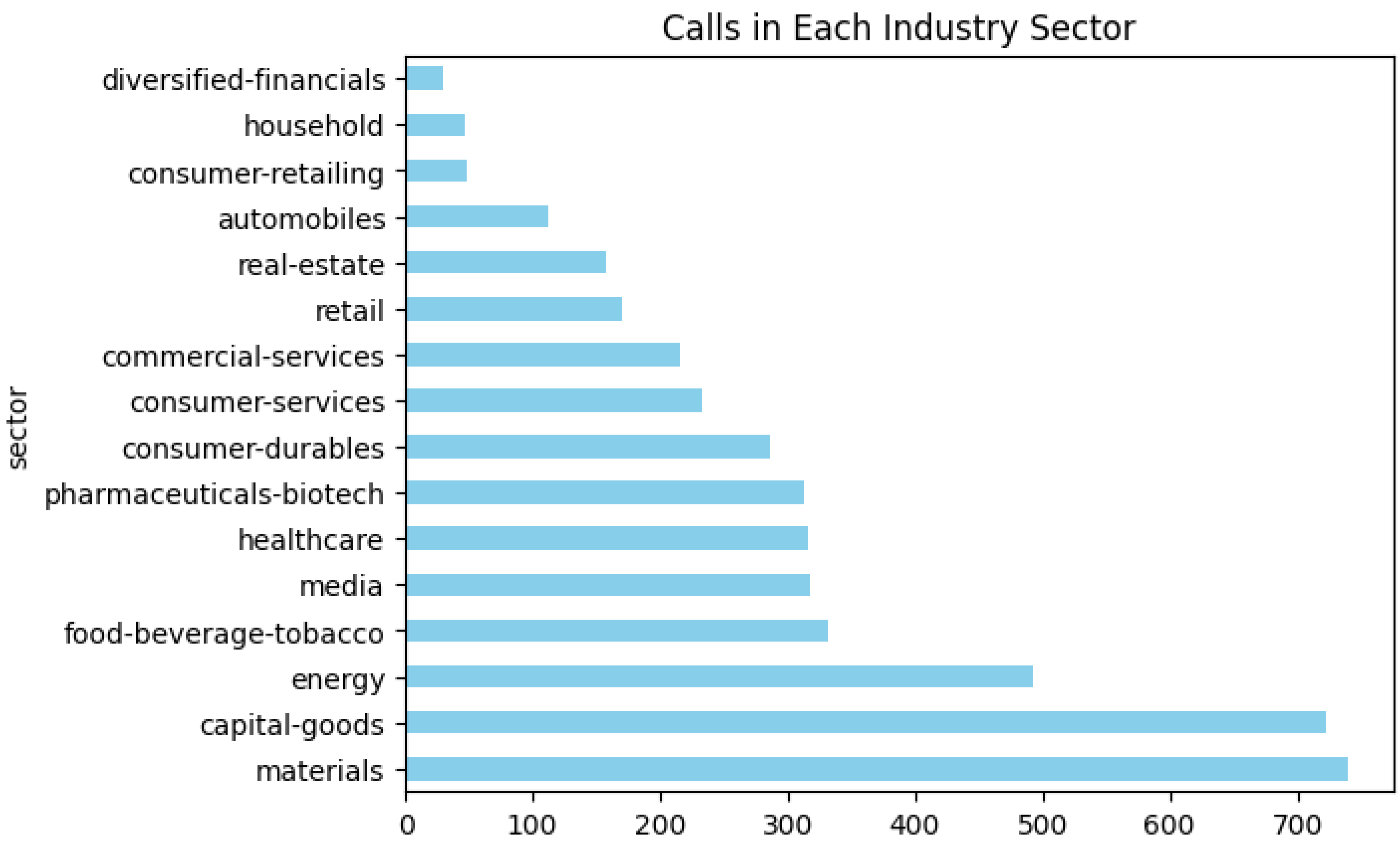
Out[ ]: 319

In [ ]:
```
# Histogram of count of earnings calls by ticker/firm
# Title: Histogram of Count of Earnings Calls by Ticker
plt.hist(df['ticker'].value_counts(), bins = 100)
plt.title('Histogram of Count of Earnings Calls by Ticker/Firm')
plt.show()
```

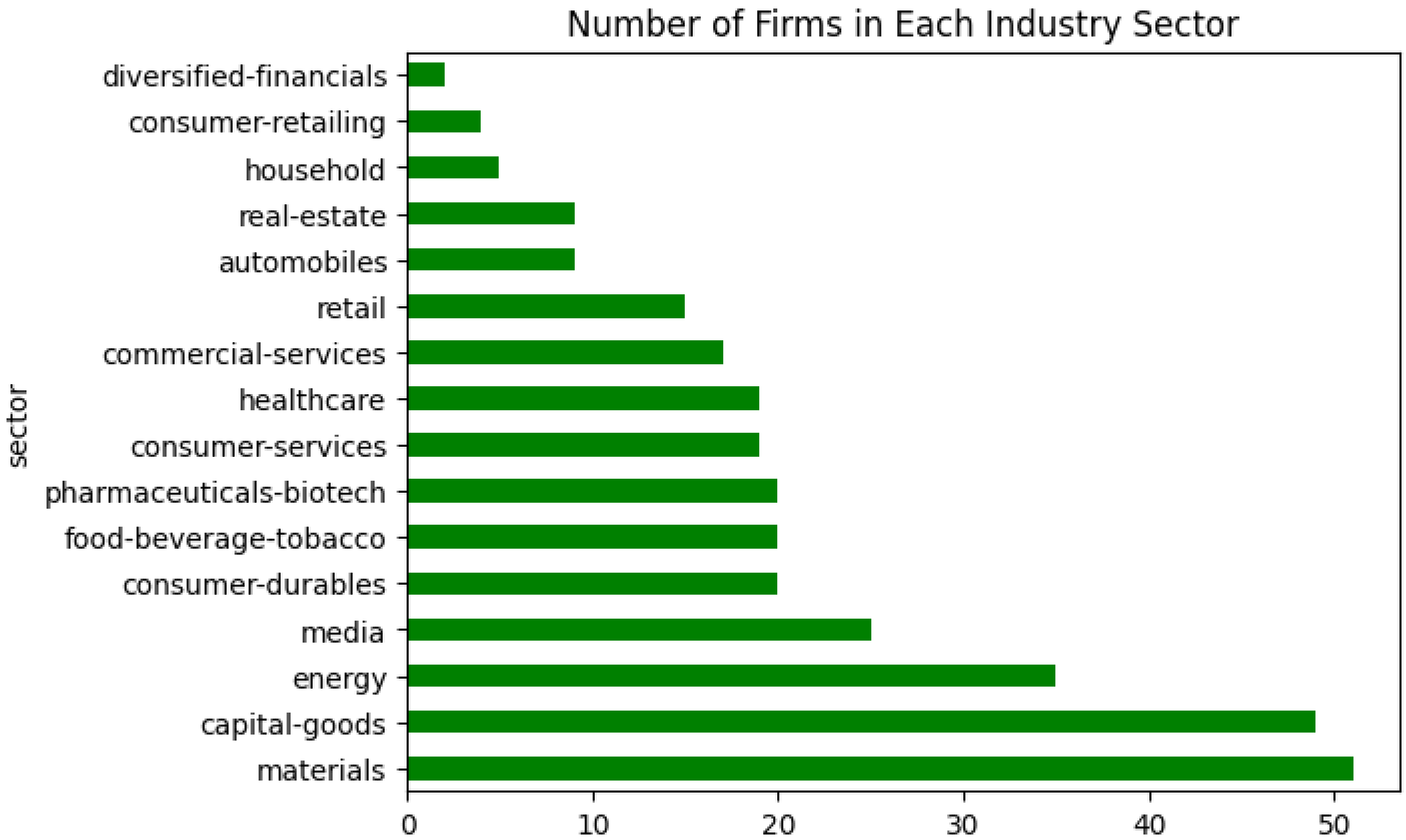


In [ ]:
```
# Distribution of industry sectors
# variable sector
```

```
# Title: Calls in Each Industry Sector
df['sector'].value_counts().plot(kind = 'barh', color = 'skyblue')
plt.title('Calls in Each Industry Sector')
plt.show()
```

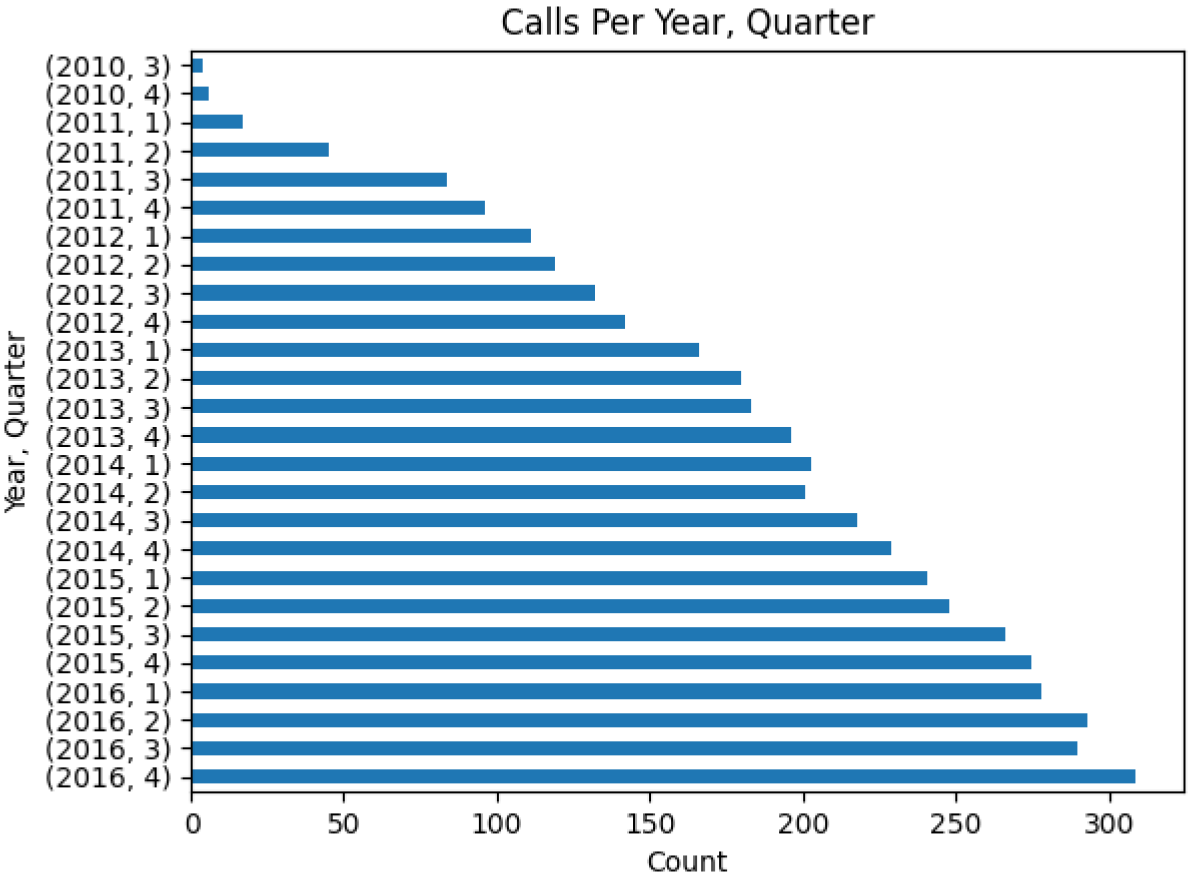


In [ ]:
```
# Number of firms in each industry sector
# Title: Number of Firms in Each Industry Sector
# Unique ticker by sector, sort by number of firms
df.groupby('sector')['ticker'].nunique().sort_values(ascending=False).plot(kind = 'barh', color = 'green')
plt.title('Number of Firms in Each Industry Sector')
plt.show()
```



In [ ]:

```
# Distribution of earnings_call_date
# Create call_year and call_quarter columns after converting earnings_call_date to datetime
df['call_year'] = pd.to_datetime(df['earnings_call_date']).dt.year
df['call_quarter'] = pd.to_datetime(df['earnings_call_date']).dt.quarter
# Group by year and quarter
data_grouped = df.groupby([df.call_year, df.call_quarter]).size().sort_index(ascending=False)
# Plot horizontal bar chart
# 2010 at the top
data_grouped.plot(kind='barh')
plt.title('Calls Per Year, Quarter')
plt.xlabel('Count')
plt.ylabel('Year, Quarter')
plt.show()
```



In [ ]:
```
# Company dropout
# For each ticker, get the max value of earnings_call_date, and print out items where it is not in the last quarter o
# Convert earnings_call_date to datetime
df['earnings_call_date_dt'] = pd.to_datetime(df['earnings_call_date'])
# Add column max_date to df
df['max_date'] = df.groupby('ticker')['earnings_call_date_dt'].transform('max')
# Display rows where max_date is not in the last quarter of 2016
df[df['max_date'] < '2016-10-01']
```

Out[ ]:

	ticker	earnings_call_date	next_earnings_call_date	rating_on_next_earnings_call_date	days_until_next_earnings_call	Rating	Rating Agency Name
143	AMCR	2016-08-25	None	None	NaN	BBB	Standard & Poor's Ratings Services
495	BTU	2011-10-25	2012-04-19	BB	177.0	BB	Standard & Poor's Ratings Services
496	BTU	2012-04-19	2012-07-24	BB	96.0	BB	Standard & Poor's Ratings Services
497	BTU	2012-07-24	2012-10-22	B	90.0	BB	Standard & Poor's Ratings Services
498	BTU	2012-10-22	2013-01-29	B	99.0	B	Standard & Poor's Ratings Services
...	...	...	...	...	...	...	...
3887	STON	2013-08-07	2013-11-08	B	93.0	B	Standard & Poor's Ratings Services
3888	STON	2013-11-08	2014-03-14	B	126.0	B	Standard & Poor's Ratings Services
3889	STON	2014-03-14	2014-05-08	B	55.0	B	Standard & Poor's Ratings Services
3890	STON	2014-05-08	2015-05-08	B	365.0	B	Standard & Poor's

	ticker	earnings_call_date	next_earnings_call_date	rating_on_next_earnings_call_date	days_until_next_earnings_call	Rating	Rating Agency Name
							Ratings Services
3891	STON	2015-05-08	None	None	NaN	B	Standard & Poor's Ratings Services

177 rows × 163 columns

In [ ]:
```
# Unique firms where this is the case
df[df['max_date'] < '2016-10-01'][['ticker', 'max_date']].drop_duplicates()
```

Out[ ]:

	ticker	max_date
143	AMCR	2016-08-25
495	BTU	2016-02-11
1598	FTI	2016-04-30
2257	KBH	2016-09-21
2687	MKC	2016-09-30
2742	MOS	2016-08-02
3056	NUE	2016-07-22
3267	PEP	2016-09-29
3318	PKG	2016-07-21
3496	RFP	2016-08-06
3887	STON	2015-05-08

AMCR is amcor, should still exist but it's date is kind of close to the end of 2016

BTU, peabody energy, seems to have gone bankrupt April 13, 2016

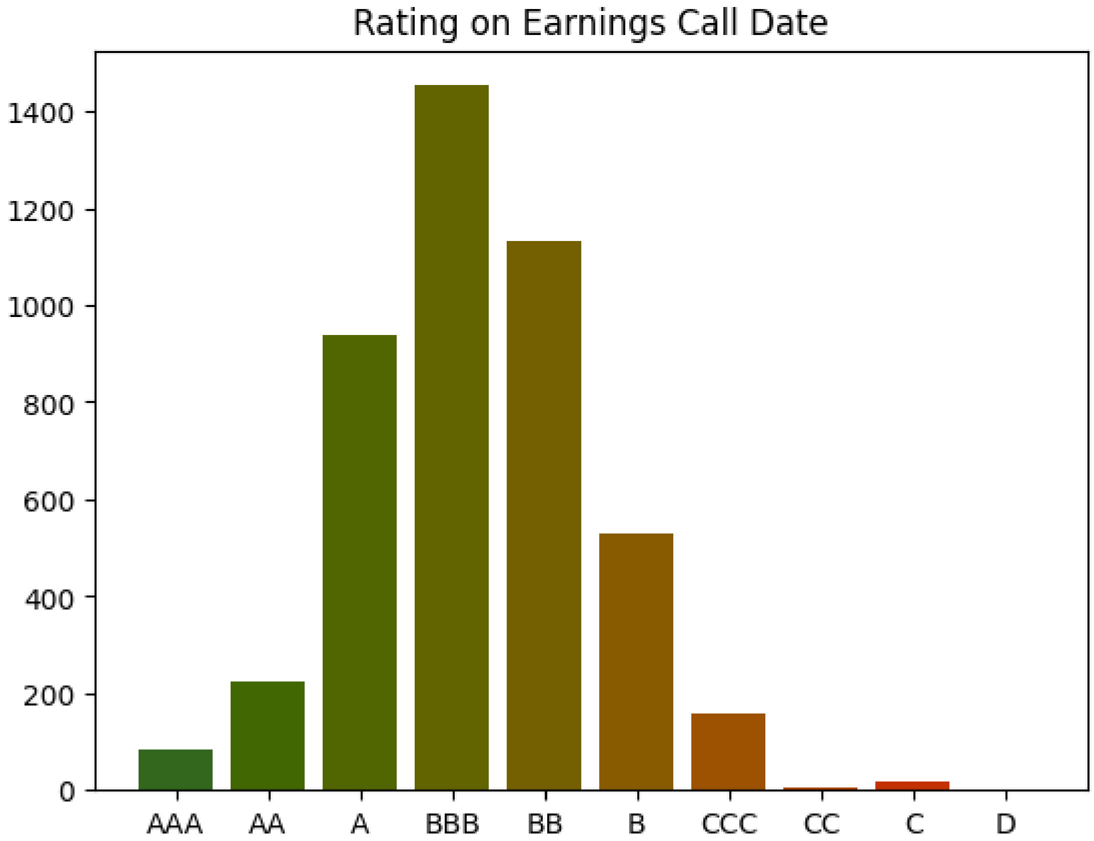
FTI underwent a merger in 2016-2017

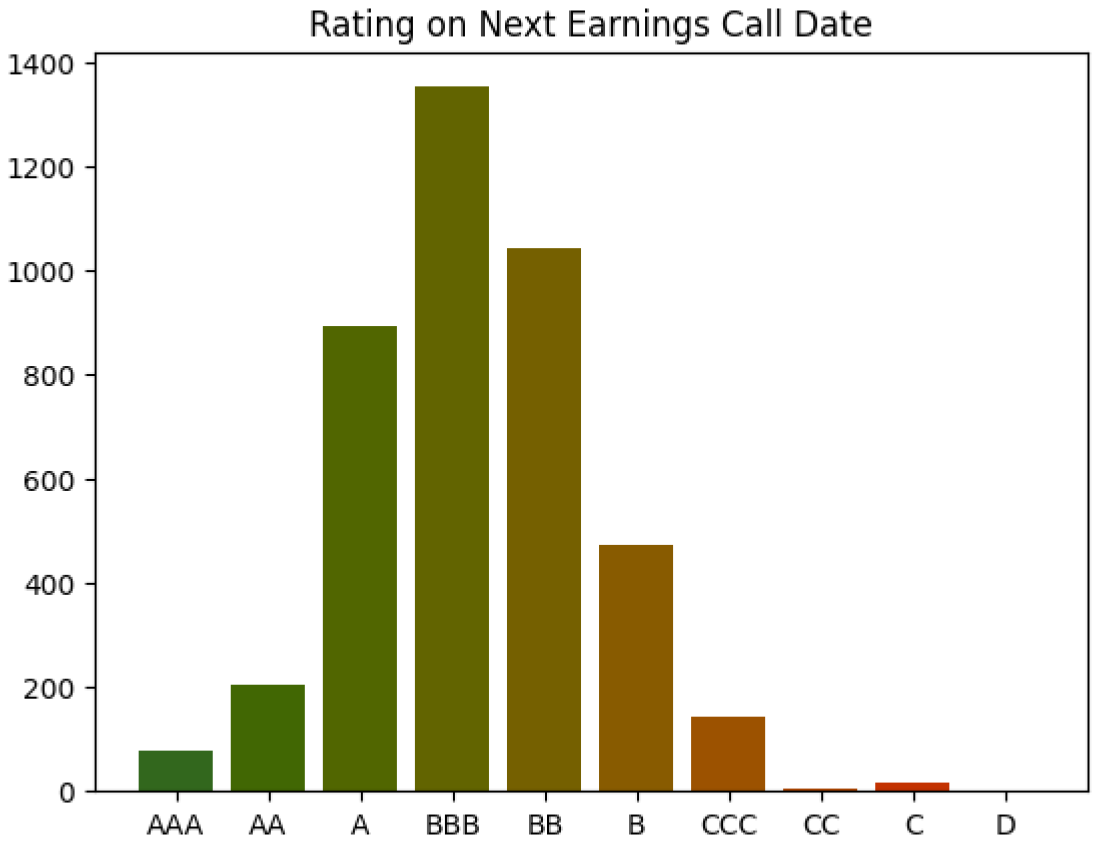
KBH still exists, but again the date is pretty close...

The other non-close one is STON. Notably, StonMor Partners had some issues with delayed SEC filings https://seekingalpha.com/article/4056108-prelude-to-bankruptcy-saving-grace-stonemor-partners-delays-10-k-again

In [ ]:
```
# Distribution of Rating and rating_on_next_earnings_call_date

# Colored with gradient and ordered

# Colors AAA through D
# Used https://colordesigner.io/gradient-generator#google_vignette
# Assign hex codes from green to red
#32671d
#416703
#516600
#626400
#756000
#885b00
#9c5200
#af4500
#c33200
#d60000
hex_code_mapper = {'AAA': '#32671d', 'AA': '#416703', 'A': '#516600', 'BBB': '#626400', 'BB': '#756000', 'B': '#885b00

# Ordering of bars - keys from hex_code_mapper
bar_order = ['AAA', 'AA', 'A', 'BBB', 'BB', 'B', 'CCC', 'CC', 'C', 'D']
# Assign values of Rating to this ordering
df['Rating'] = pd.Categorical(df['Rating'], categories=bar_order, ordered=True)

# Create plot
# Save to "../Output/Distribution of Ratings.png"
plt.bar(df['Rating'].value_counts().sort_index().index, df['Rating'].value_counts().sort_index(), color=[hex_code_mapp
plt.title('Rating on Earnings Call Date')
#plt.savefig('../../Output/Distribution of Rating Issuances.png')
plt.show()

# Rating on next earnings call date
df['rating_on_next_earnings_call_date'] = pd.Categorical(df['rating_on_next_earnings_call_date'], categories=bar_order
plt.bar(df['rating_on_next_earnings_call_date'].value_counts().sort_index().index, df['rating_on_next_earnings_call_da
plt.title('Rating on Next Earnings Call Date')
```

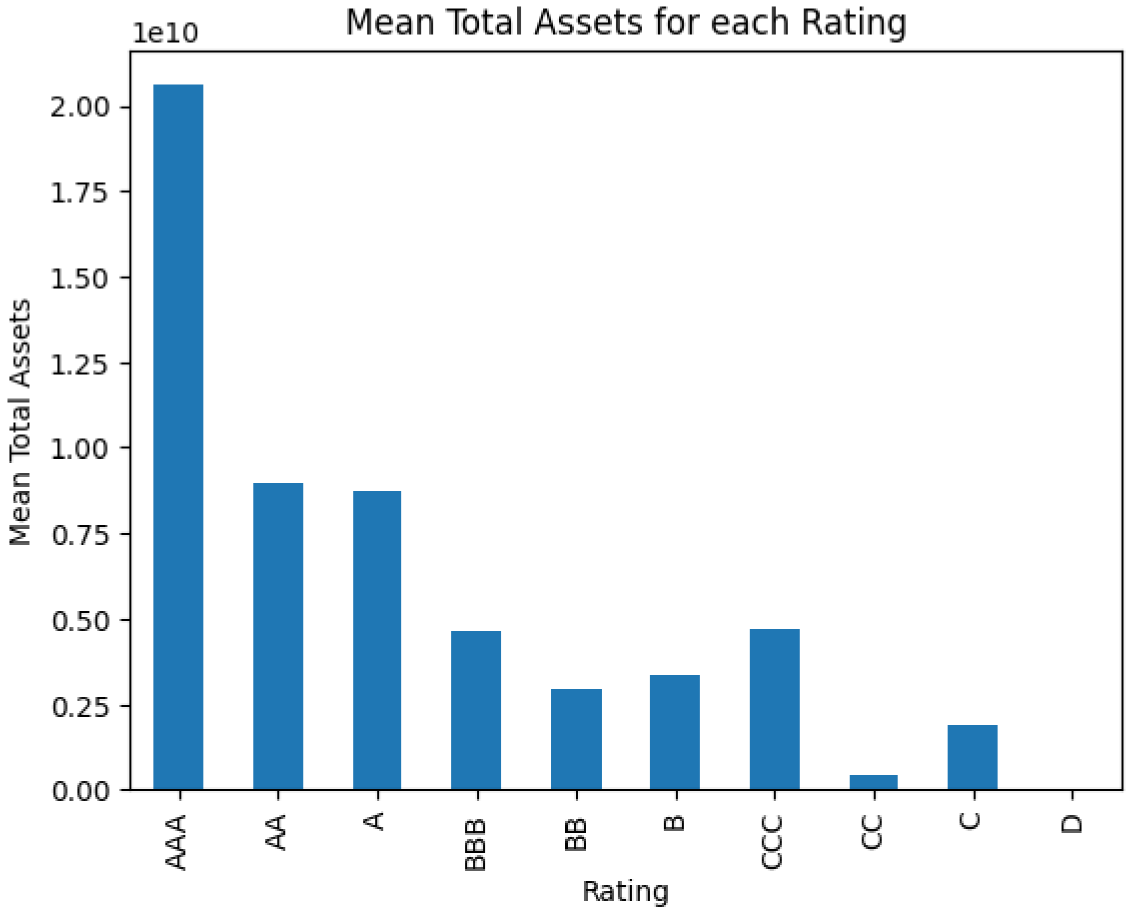
```
#plt.savefig('../../Output/Distribution of Rating on Next Earnings Call Date.png')
plt.show()
```





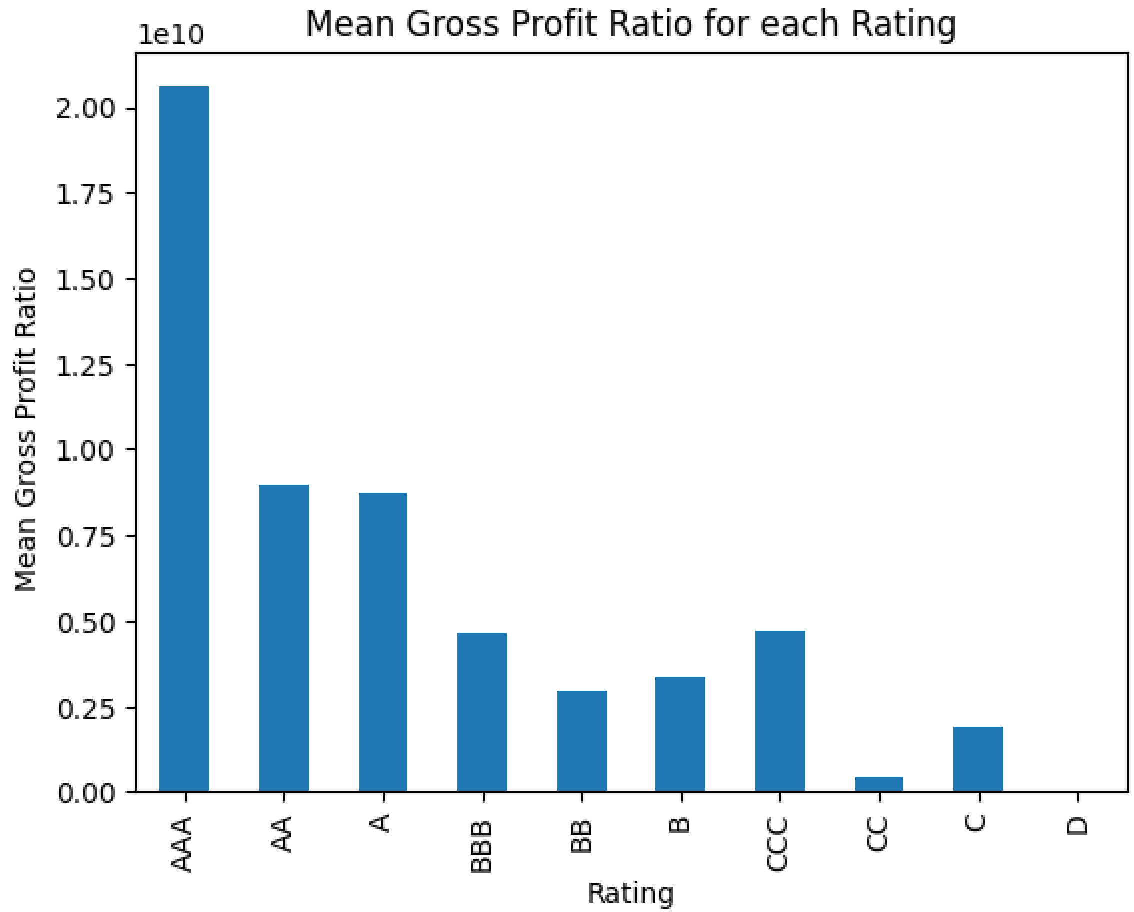
In [ ]:
```
# Calculate the mean of "totalAssets" for each kind of "Rating"
mean_assets_by_rating = df.groupby('Rating')['totalDebt'].mean()

# Plotting
mean_assets_by_rating.plot(kind='bar')
plt.title('Mean Total Assets for each Rating')
plt.xlabel('Rating')
plt.ylabel('Mean Total Assets')
plt.show()
```

```
/var/folders/2v/664l7ccj4vn7kztqkgv6y7mh0000gn/T/ipykernel_96962/3119831694.py:2: FutureWarning: The default of obser
ved=False is deprecated and will be changed to True in a future version of pandas. Pass observed=False to retain curr
ent behavior or observed=True to adopt the future default and silence this warning.
  mean_assets_by_rating = df.groupby('Rating')['totalDebt'].mean()
```



In [ ]:
```python
# Calculate the mean of "grossProfitRatio" for each kind of "Rating"
mean_gross_profit_by_rating = df.groupby('Rating')["grossProfitRatio"].mean()

# Plotting
mean_assets_by_rating.plot(kind='bar')
plt.title('Mean Gross Profit Ratio for each Rating')
plt.xlabel('Rating')
plt.ylabel('Mean Gross Profit Ratio ')
plt.show()
```

```
/var/folders/2v/664l7ccj4vn7kztqkgv6y7mh0000gn/T/ipykernel_96962/796684613.py:2: FutureWarning: The default of observ
ed=False is deprecated and will be changed to True in a future version of pandas. Pass observed=False to retain curre
nt behavior or observed=True to adopt the future default and silence this warning.
  mean_gross_profit_by_rating = df.groupby('Rating')["grossProfitRatio"].mean()
```



We can see the relationship between rating and total assets; rating and gross profit ratio. High rating like AAA will have high mean total assets and high mean gross profit ratio.

NLP EDA

In [ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.probability import FreqDist
```

```python
# Basic Statistics
num_records = len(df)
avg_length = df['transcript'].str.len().mean()

print(f"Number of records: {num_records}")
print(f"Average transcript length: {avg_length:.2f} characters")
```

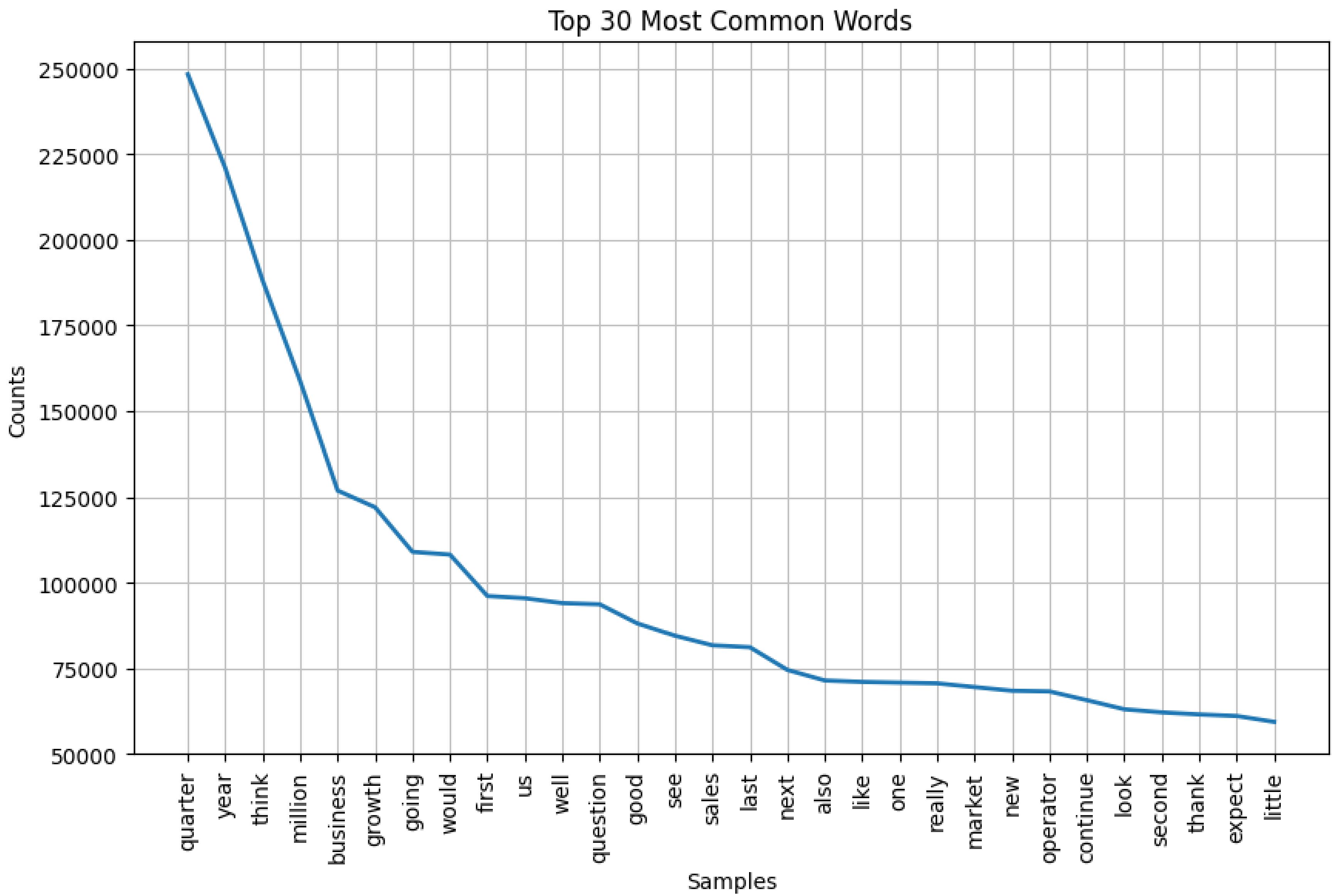
```
Number of records: 4532
Average transcript length: 51412.94 characters
```

In [ ]:

```python
# Tokenize the text
nltk.download('punkt')  # Download NLTK tokenizer data
tokens = df['transcript'].apply(word_tokenize)

# Remove stop words
nltk.download('stopwords')  # Download NLTK stop words data
stop_words = set(stopwords.words('english'))
tokens = tokens.apply(lambda tokens: [word for word in tokens if word.lower() not in stop_words and word.isalpha()])

# Word Frequency Analysis
all_words = [word.lower() for token_list in tokens for word in token_list]
fdist = FreqDist(all_words)
top_words = fdist.most_common(10)
print("Top 10 most common words:")
for word, freq in top_words:
    print(f"{word}: {freq}")

# Plot Word Frequency Distribution
plt.figure(figsize=(10, 6))
fdist.plot(30, title='Top 30 Most Common Words')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/chengzhengxing/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/chengzhengxing/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

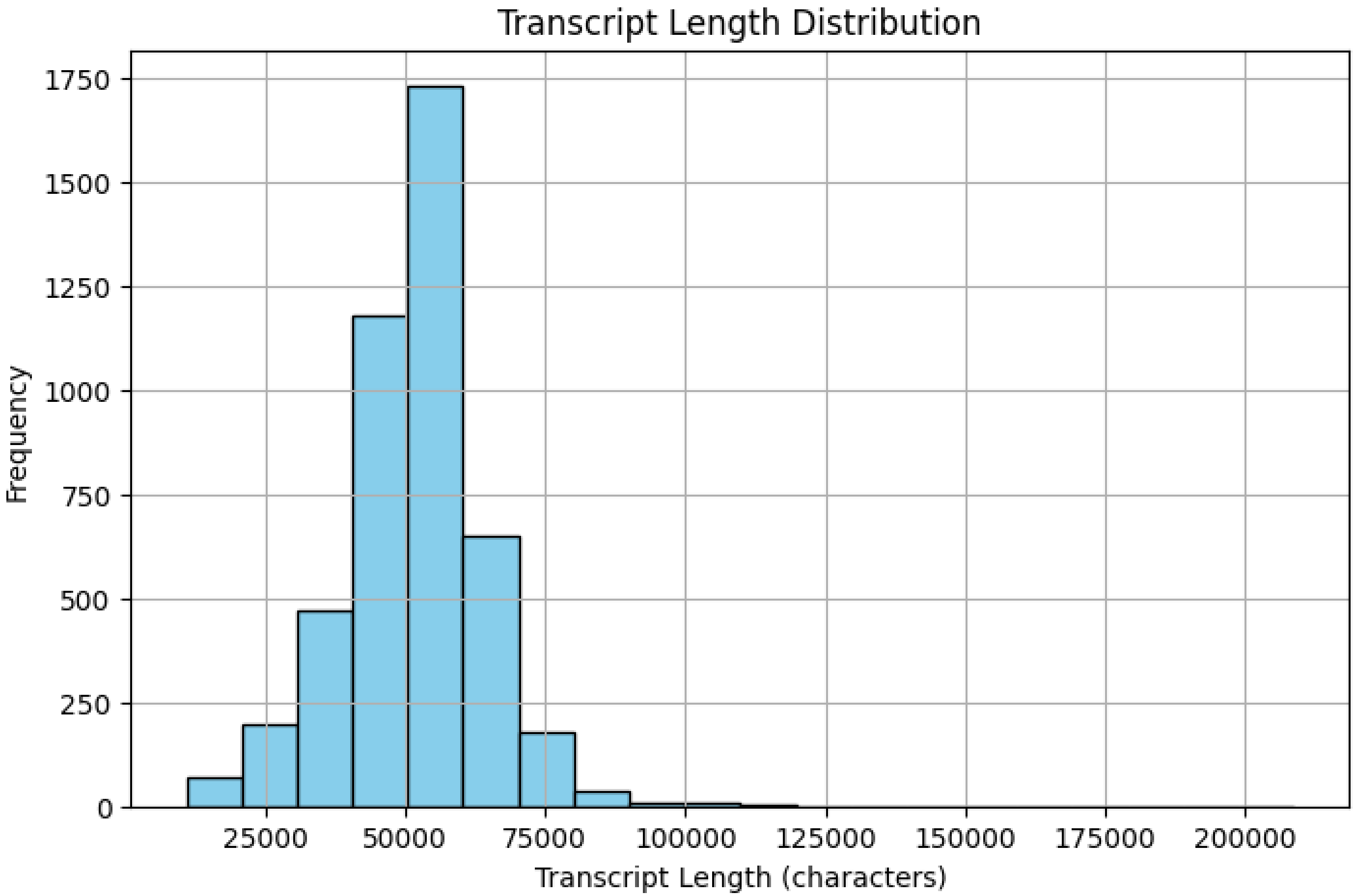
```
Top 10 most common words:
quarter: 248253
year: 220870
think: 187990
million: 158704
business: 126889
growth: 122016
going: 108995
would: 108204
first: 96107
us: 95476
```



Out[ ]: <AxesSubplot: title={'center': 'Top 30 Most Common Words'}, xlabel='Samples', ylabel='Counts'>

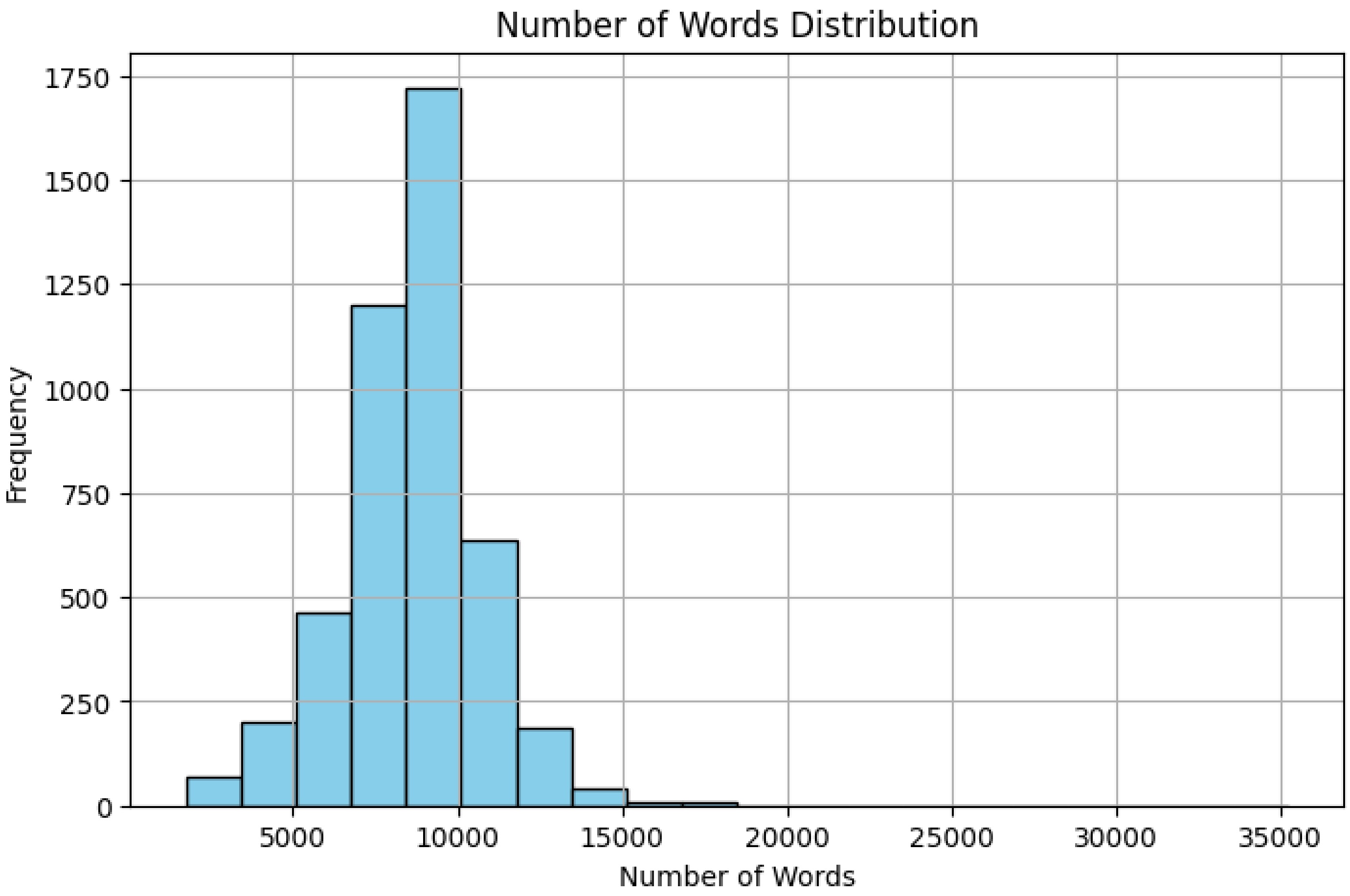
In [ ]:
```
# Text Length Distribution
transcript_lengths = df['transcript'].str.len()
plt.figure(figsize=(8, 5))
plt.hist(transcript_lengths, bins=20, color='skyblue', edgecolor='black')
```

```
plt.title('Transcript Length Distribution')
plt.xlabel('Transcript Length (characters)')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```



In [ ]:
```
# Distribution of Number of Words in earning calls
def count_words(tokens):
    tokens = [word.lower() for word in tokens if word.isalpha()]
    return len(tokens)

word_tokens = df['transcript'].apply(word_tokenize)
num_words = word_tokens.apply(count_words)
plt.figure(figsize=(8, 5))
```

```
plt.hist(num_words, bins=20, color='skyblue', edgecolor='black')
plt.title('Number of Words Distribution')
plt.xlabel('Number of Words')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```

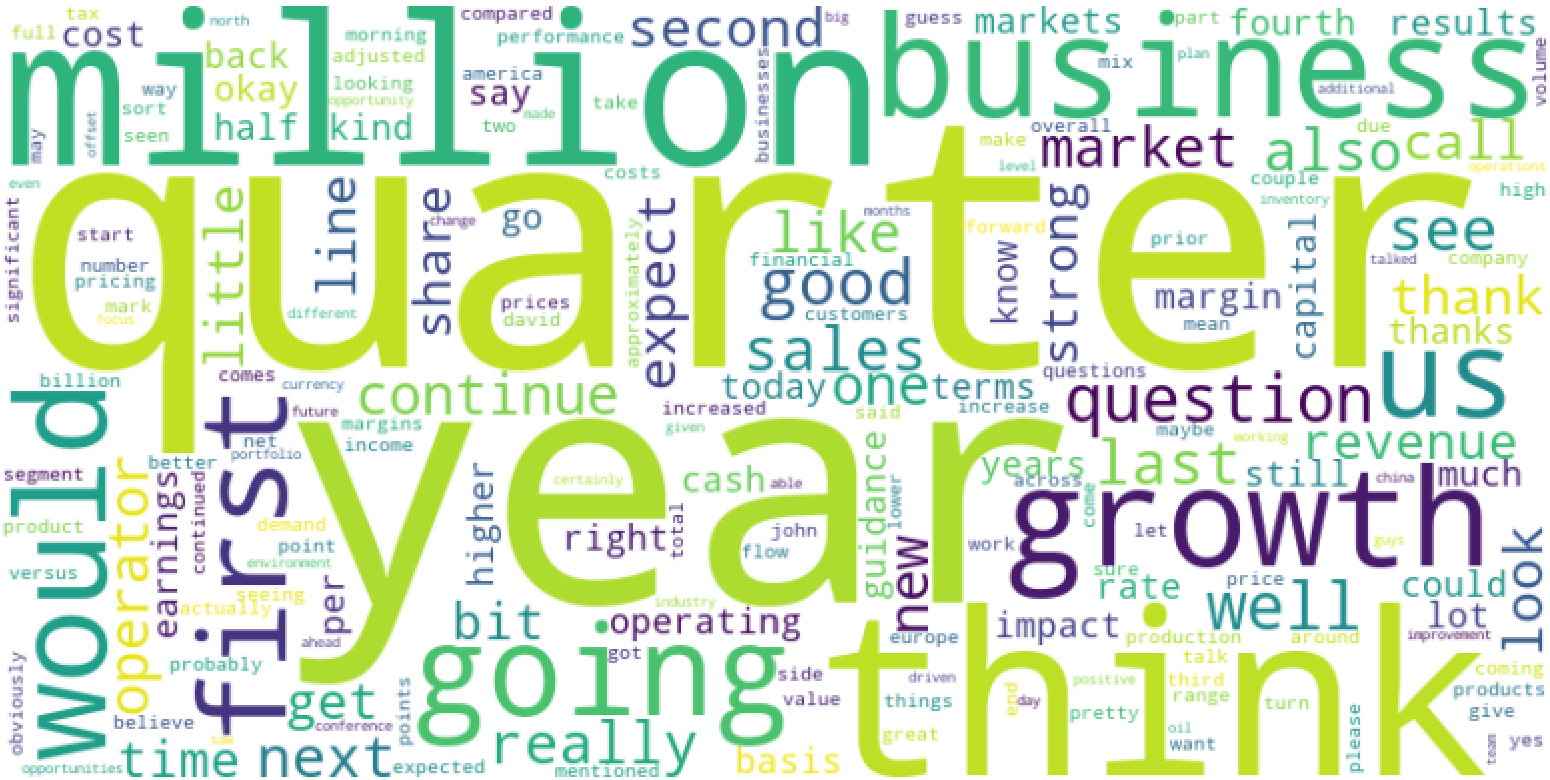


In [ ]:
```
from wordcloud import WordCloud

# Generate word clouds for each credit rating
ratings = ["AA","BBB","CCC"]
for rating in ratings:

    rating_df = df[df['Rating'] == rating]
```

```
    tokens = rating_df['transcript'].apply(word_tokenize)
    stop_words = set(stopwords.words('english'))
    tokens = tokens.apply(lambda tokens: [word for word in tokens if word.lower() not in stop_words and word.isalpha(

    all_words = [word.lower() for token_list in tokens for word in token_list]
    fdist = FreqDist(all_words)

    # Generate Word Cloud
    wordcloud = WordCloud(width=800, height=400, background_color='white').generate_from_frequencies(fdist)

    # Plot Word Cloud
    plt.figure(figsize=(10, 6))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.title(f'Word Cloud for Credit Rating {rating}')
    plt.show()
```

Word Cloud for Credit Rating AA

Word Cloud for Credit Rating BBB



Word Cloud for Credit Rating CCC

We can see from word cloud for credit rating AA, postive words like "growth" are larger when compared to credit rating CCC.