UBC SAUDER SCHOOL OF BUSINESS

ROBERT H. LEE GRADUATE SCHOOL

BAIT 508 Group Project Report

- **Project Title: Industry Analysis**
- **Section: BA1**

Team Members			
#	**Name**	**Student ID**	**Email**
1	Zhi (Krystal) Li	95680609	zhili01@student.ubc.ca
2	Yousef Jafarnia	37746765	Yousef.jafarnia@gmail.com
3	Pranav Mehta	19184282	pranavmehta98@gmail.com

Introduction

The goal of this project is to conduct an in-depth analysis of public U.S. firms within a selected industry sector using various data analytics and natural language processing (NLP) techniques. The analysis spans both quantitative and qualitative evaluations, which will be detailed in the following sections. We utilized multiple datasets, including numerical and textual data, to generate insights into the selected industry sector, identifying key trends, firm behaviors, and make strategic suggestions.

Methodology

In this project, the tasks were divided among the team members to ensure an efficient workflow and a comprehensive analysis. The methodology used for the project can be summarized as follows:

- **Initial Coding and Setup:**
  Krystal initiated the work by writing the first version of the code for the quantitative and the text analysis on the industry sector. This included data exploration, filtering, and cleaning, as well as setting up the framework for text processing using NLP techniques.
- **Code Review and Corrections:**
  Yousef and Pranav both reviewed the code Krystal developed for the first two parts. They identified and corrected errors, improving the accuracy of the analysis. In addition, they made several alterations to enhance the performance and structure of the code.
- **Completion of the Third Part and Draft Report:**
  Yousef was responsible for completing the comprehensive analysis of one sample firm. He also took the lead in drafting the first version of the project report, ensuring that the analysis was properly documented, and the insights were clearly explained.
- **Building on the Part 3 and Report Completion:**
  Pranav improved part 3 and ensured clean and clear formatting. He found our target firm's competitors using the trained word2vec model, added graphs, corrected remaining issues, added missing elements, and ensured the clarity of the report.
- **Final Review and Editing of the Report:**
  Krystal conducted final review of the report. She checked content consistency, corrected typos, and made proper citation of AI usage. She made sure all necessary components were addressed before submission.

Through this collaborative approach, the project tasks were distributed effectively, with each member contributing significantly to both the technical analysis and the documentation.

AI Citations:

- Part 1: we used Chat GPT 4.0 to assist with data visualization.
    - For specific tasks, such as **Part 1-B-4** and **Part 1-B-6**, we prompted ChatGPT with: *"Add a piece of code to annotate the data points in the graph."*
- Part 2: we used Professor Lee's Jupiter notebooks code to create the functions for text analytics.
- Part 3: we used Chat GPT 4.0 to create the main structures of the code and for visualization, and then we wrote the code with the insights of methods and functions.
    - Part 3-F-1:
        - To add a new column ("firm_level_embedding") to the dataframe ("selected_firms_10k_reports"), we used the prompt: *"Here is my code for generating the top 10 keywords based on the TF-IDF score and the code where I trained a Word2Vec model. I need your help to come up with code for firm-level embedding."*
        - To generate a new dataframe ('top_5_competitors_display') for storing the top 5 competitor firms based on similarity scores, our prompt was: *"Give me the top 5 competitor firms for Tenet Healthcare (gvkey = 7750) based on their firm-level embeddings and similarity to Tenet's score."*
    - Part 3-F-2:
        - To create a graph of the stock price trends for Tenet and its competitors, we used the prompt: *"Generate a graph to plot the stock price trend for all the firms in the dataframe 'new_selected_firms_df'."*
        - To generate a graph showing sales trends, we prompted: *"Generate a graph to plot the sales trend for all the firms in the dataframe 'new_selected_firms_df'."*
- Finally, we used Chat GPT for grammar and typo check of the written report.

Project Overview

Project Main Parts

The analysis consists of three main parts:

1. **Quantitative Analysis of the Industry Sector**
2. **Text Analysis on the Industry Sector**
3. **Comprehensive Analysis of One Sample Firm**

These analyses are conducted using Python-based data analysis tools and techniques, including pandas, matplotlib, natural language processing libraries, and word2vec.

Python Libraries and Packages

As it is demonstrated in the beginning of the code, here is the list of libraries and packages used in the project.

- **pandas**
- **matplotlib**
- **seaborn**
- **numpy**
- **string**
- **nltk**
- **sklearn**
- **collections**
- **wordcloud**
- **gensim**

Project Main Datasets

There are three main datasets as the input for the analysis.

1. **major_groups.csv:** this dataset contains the list of industries and their SIC code. The main purpose of this dataset is to choose the target industry of the analysis and use the related SIC code to filter the public firm in the selected industry.

2. **public_firms.csv:** this dataset represents financial data for a list of U.S. public firms named over multiple fiscal years, from 1994 to 2020. It is used to analyze the financial behavior of the selected industry and the sample firm.

3. **2020_10K_item1_full.csv:** This dataset contains 10K report of the year 2020 for U.S. public firms. It is used to perform text analysis and find insights about keywords about the firms of the selected industry and sample firm.

Part1- Quantitative Analysis of the Industry Sector

Initial Exploration

In the first step, we need to explore the input data.

For this part we use the "major_groups.csv" to find the SIC code of the selected industry. Based on the initial exploration, this data contains 83 rows and 2 columns described below:

- **major_group:** the first two digits of SIC code as an indicator of the industry of the firms
- **description:** a brief description of the industry sector related to the code

Next, we check for any missing values and duplicate records in the data set and we see that this dataset does not have any missing values or any duplicate records.

We repeat this analysis for the "public_firms.csv" as well to check the data. We can see that there are 209,212 records in this data set. The columns of this data set are:

- **gvkey:** A unique company identifier used to track the company across different fiscal years.
- **fyear:** The fiscal year corresponding to the financial data, ranging from 1994 to 2020.
- **location:** The country where the firm is based.
- **conm:** The company name.
- **ipodate:** The initial public offering (IPO) date of the company.
- **sic:** The Standard Industrial Classification (SIC) code, a four-digit code representing the company's primary industry sector.
- **prcc_c:** The closing stock price for the company in the respective fiscal year.
- **ch:** Represents cash holding for the company in the respective financial year.
- **ni:** The net income for the company in the respective fiscal year.
- **asset:** The total assets of the company in the respective fiscal year.
- **sale:** The total sales (or revenue) generated by the company.
- **roa:** The return on assets (ROA), a financial ratio calculated as net income divided by total assets.

Next, we check for any missing values and duplicate records in the data set and we see that this dataset does not have any duplicate records but there are missing values in some of the main columns like "sale", "prcc_c", "roa", and "ni"

For the last step of initial exploration, since we have numerical data in the second dataset, we use "describe" method to check the main statistics about the numeric columns. Based on this observation, we can see that for most of the firms in the dataset, the first three quantiles are very different than the maximum values for net income, sales, asset and stock prices, and this shows that the data contains outliers driving the mean statistics much higher than the median, so median would be a better statistic for these columns.

Also by checking the describe output for "sic" column maximum and minimum, we figured out that all of the firms have 4 digit SIC code and we can easily use the first two digits without worrying about SIC code with other lengths.

A. Industry Sector Selection and Data Filtering

For the purpose of this project, we selected the **"Health Services"** industry sector. Using the major_groups.csv dataset and searching for the Health Services in the description column, we find out that the first two digit of SIC code for this sector is "80".

Then, we filtered out firms based on the first two digits of their SIC codes that correspond to the major industry group. After filtering the dataset, first we examine the data by checking its first 10 records.

In the next step we performed a series of analyses on the public_firms.csv dataset to get these observations.

- **Unique Firm-Year Observations**: We identified **3064** unique firm-year observations in the dataset for the selected industry. To do so, we used column selection and the drop_duplicates method and counting the rows by using shape method.
- **Unique Firms**: A total of **358** unique firms was in the filtered dataset. For this part we used nunique method over the "gvkey" column.
- **Firms with Records Over 27 Years**: Out of all firms in the dataset, only **2** firms have complete records spanning 27 years (1994–2020). To calculate this number, we used groupby method over "gvkey" column and counted the records.

B. Preliminary Analysis

We conducted a series of preliminary analyses to understand the financial status and geographical distribution of firms in the sector:

1. **Top 10 Firms by Stock Price in 2020**: The top 10 firms with the highest stock prices in 2020 and their stock prices are:
    - CHEMED CORP 532.61
    - AMEDISYS INC 293.33
    - LHC GROUP INC 213.32
    - LABORATORY CP OF AMER HLDGS 203.55
    - TELADOC HEALTH INC 199.96
    - HCA HEALTHCARE INC 164.46
    - UNIVERSAL HEALTH SVCS INC 137.50
    - U S PHYSICAL THERAPY INC 120.25
    - QUEST DIAGNOSTICS INC 119.17
    - DAVITA INC 117.40

    To find these firms, we used filtering and then sort_values method.
2. **Top 10 Firms by Sales**: Across the entire filtered dataset, the top 10 firms with the highest sales and their sales are:
    - HCA HEALTHCARE INC 765445.000
    - TENET HEALTHCARE CORP 322940.000
    - FRESENIUS MEDICAL CARE AG&CO 270953.523
    - COMMUNITY HEALTH SYSTEMS INC 215654.733
    - DAVITA INC 157702.707

- QUEST DIAGNOSTICS INC 146895.716
- UNIVERSAL HEALTH SVCS INC 144328.980
- LABORATORY CP OF AMER HLDGS 134399.500
- KINDRED HEALTHCARE INC 96953.895
- ENCOMPASS HEALTH CORP 82040.121

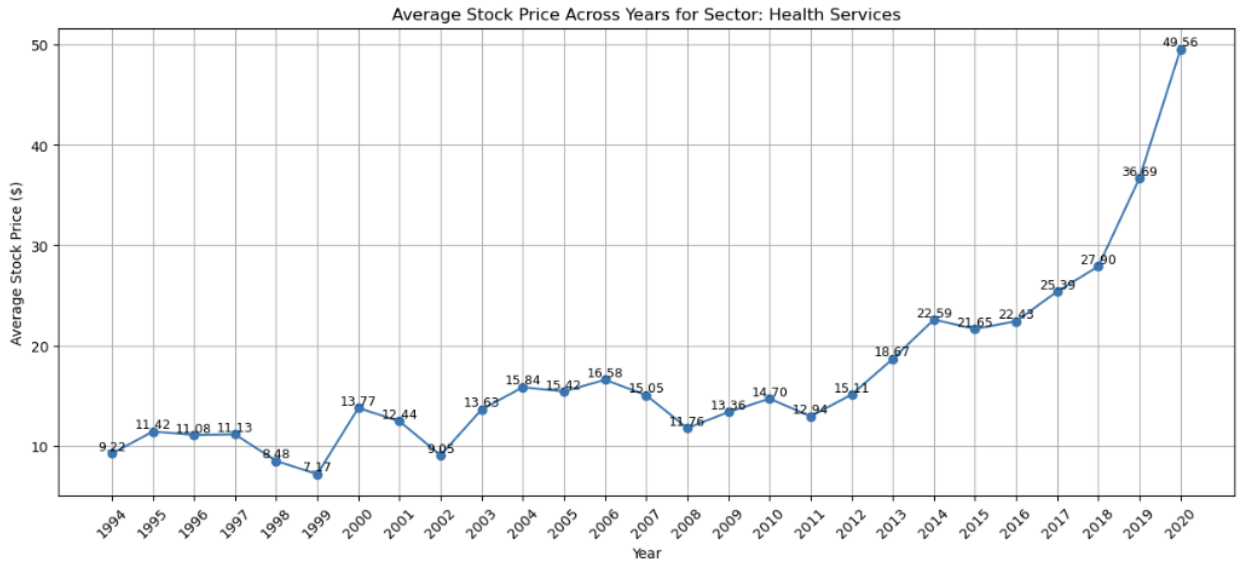
To find these firms, we used groupby, sum, and sort_values methods.

3. **Geographical Distribution**: The top 10 locations with the highest number of firms and the number of firms in those locations are as follows:

   - USA 344
   - CAN 5
   - CHN 5
   - HKG 2
   - AUS 1
   - DEU 1

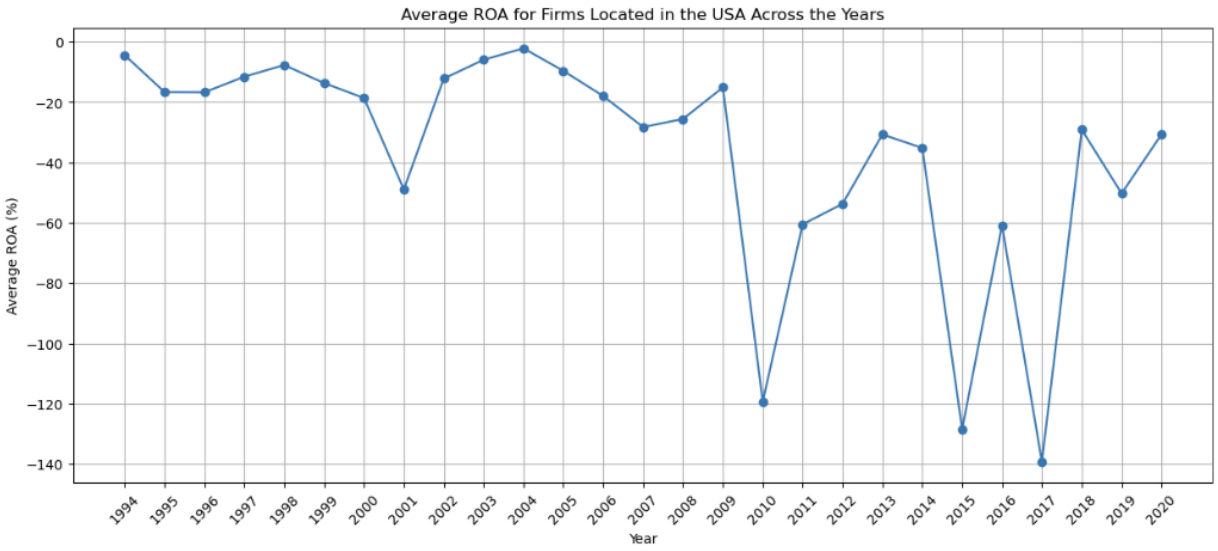
   The result contains only 6 locations because the entire filtered data only contains 6 locations. To get this result we first use drop_duplicates method over 'conm' and 'location' columns and then used groupby method for the location to count the firms.

4. **Visual Line Chart of the average stock price**: we created a line chart showing the average stock price across the years for the sector, allowing us to visualize sector trends over time.
   To create this chart, we used groupby and mean methods to calculate the average stock price for the all the firms in the selected sector and then used plot from matplotlib package to create the chart.



5. **Impact of the 2008 Financial Crisis**: We analyzed the firms most affected by the 2008 financial crisis by examining the percentage drop in stock prices from 2007 to 2008. To do so we used filtering, pivot table (to calculate the drop in two records), and idxmin and min methods to find the firm.
   The firm that experienced the most significant drop in stock price was INSIGHT HEALTH SVCS HLDG CP, with a decrease of -99.33%.
6. **Visual Line Chart of average ROA for the firms located in the USA**: we analyzed the data by using filters on the dataset to focus on firms located in the USA within the selected industry

sector and then calculated their average Return on Assets (ROA) across the years, and finally used groupby and mean methods to calculate the average. To plot the results, we used plot from matplotlib package.



Part2- Text Analysis on the Industry Sector

Initial Exploration

In the first step, we need to explore the input data.

For this part we use the "2020_10K_item1_full.csv" to access the 10K reports of the firms in year 2020. Based on the initial exploration, this data contains 5481 rows and 5 columns described below:

- **cik:** The Central Index Key (CIK) is a unique identifier assigned by the U.S. Securities and Exchange Commission (SEC) to identify corporations and individuals who file disclosures with the SEC.
- **year:** This represents the year of the filing or report associated with the firm which is 2020 for all the records.
- **name:** The company name, referring to the firm that submitted the filing.
- **item_1_text:** This column contains the textual content of 10K report.
- **gvkey:** A unique identifier assigned to companies in financial databases, used to track companies over time in datasets.

Next, we check for any missing values and duplicate records (based on the 10K report column) in the data set and we see that this dataset does not have any missing values but there are 962 duplicate records.

C. Text Cleaning

To analyze the data, we are taking some steps to clear the data. Here are the steps:

0. Removing duplicates: since we are going to perform keywords analytics techniques such as TF-IDF, duplicate records compromise the scores. So, we use drop_duplicates method over "**item_1_text" column** to do so.
1. Converting all words to lowercase.
2. Removing punctuation.
3. Removing stop words using NLTK.

To perform steps 1, 2, and 3, we used this code:

```
translator = str.maketrans('', '', string.punctuation)
sw = stopwords.words('english')

def clean_text(text):
    text_lower = text.lower()
    text_no_punctuation = text_lower.translate(translator)
    clean_words = [w for w in text_no_punctuation.split() if w not in sw]
    return ' '.join(clean_words)

all_10k_reports_cleaned.loc[:, 'item_1_cleaned'] = all_10k_reports_cleaned['item_1_text'].apply(clean_text)
```

Here is a brief explanation of this part:

- **Translator for Removing Punctuation**: This line creates a translation table that is used to remove all punctuation from the text. The str.maketrans() function creates a mapping from the characters we want to remove to None.
- **Stopwords Definition**: This defines the set of English stopwords using the stopwords.words('english') method from the NLTK library.
- **Defining the clean_text Function**: first converting the text to lowercase to make the cleaning case-insensitive. Then removing punctuation from the text using the translation table defined earlier. Next, split the text into words, filtering out stopwords, and return only the meaningful words. Finally joining the cleaned words back into a single string, separated by spaces.
- **Applying the clean_text Function to the DataFrame:** This applies the clean_text function to the item_1_text column of the all_10k_reports_cleaned DataFrame. The result is a new column, item_1_cleaned, containing the cleaned version of the text.

D. Keyword Analysis

Using two methods, word counts and TF-IDF scores, we generated the top 10 keywords for firms in the selected industry sector. These keywords provide insight into the main concepts within the firms of the selected sector. Then we visualized the keywords using word clouds, which highlight the key terms across the firms, based on both word counts and TF-IDF scores.

- Step 1: creating the data frame for the firms in the selected sector
  By using merge method, we join the two data frames from 10K reports and financial data for the firms in the selected industry.

- Step 2: generating top 10 keywords for each firm based on word counts and TF-IDF score
  - For the word count method, we define a function that gets text as input and then apply this function to each row by apply method. Here is the code:

```
def get_keywords_wordcounts(text):
    c = Counter(str(text).split())
    words = []
    for pair in c.most_common(10):
        words.append(pair[0])
    return ' '.join(words)

selected_firms_10k_reports['keywords_wordcounts'] = \
selected_firms_10k_reports['item_1_cleaned'].apply(get_keywords_wordcounts)
```

In this defined function, we use split method to break the text into words and then use most_common method to return the 10 most common words and their count. Finally, we concatenate the words into a single output.

- For the TF-IDF method, we define a function that get a list of documents as an input, then provides a list of keywords then we applied this function to all 10K reports in the selected industry and get the new column as the keywords. Here is the code:

```
def get_keywords_tfidf(documents):
    vectorizer = TfidfVectorizer()
    tfidf_matrix = vectorizer.fit_transform(documents)
    feature_names = vectorizer.get_feature_names_out()
    top_keywords = [] # accumulator
    for i in range(tfidf_matrix.shape[0]):
        feature_index = tfidf_matrix[i, :].nonzero()[1]
        tfidf_scores = zip(feature_index, [tfidf_matrix[i, x] for x in feature_index])
        sorted_tfidf_scores = sorted(tfidf_scores, key=lambda x: x[1], reverse=True)
        top_keywords.append(' '.join([feature_names[i] for i, _ in sorted_tfidf_scores[:10]]))

    return top_keywords

selected_firms_10k_reports['keywords_tfidf'] = \
get_keywords_tfidf(selected_firms_10k_reports['item_1_cleaned'])
```

The function get_keywords_tfidf takes a list of cleaned textual documents (one document per firm) and returns the top 10 keywords for each document based on their TF-IDF scores.
In this function, first we initialize a TfidfVectorizer, which converts the text into a matrix of TF-IDF features. Each row represents a document, and each column represents a word with its associated TF-IDF score. Then the fit_transform() method converts the documents (list of firm reports) into a sparse matrix where each row corresponds to a document and each column to a term, with the values representing the TF-IDF scores for each term. Next, we extract the names of the features (words) from the vectorizer, allowing access to the actual words corresponding to the TF-IDF scores. Finally, we are looping Through Documents and Getting Top Keywords by identifying the non-zero TF-IDF scores, zipping the feature index (word position) and its corresponding TF-IDF score, sorting the TF-IDF scores in descending order, and adding the top 10 keywords for each document to the top_keywords list.

- Step 3: creating two wordclouds to visualize the keywords of firms in the selected sector based on each method.
  To do so, we first need to create a string of all the keywords from each method and then use WordCloud function from wordcloud package to create the wordclouds.
  Then we use matplotlib package to visualize the wordclouds. Here are two wordclouds (1st is the wordcloud created based on word counts method and the 2nd one is the wordcloud created based on TF-IDF method).

E. Word embedding

We trained a word2vec model to analyze word similarities within the cleaned text data. To do so, first we convert the 10K reports into a list of lists of words by using split method and list comprehension. Then by using Word2Vec function from genism package, we train and save a word2vec model based on the data.

After manually inspecting the word clouds, we selected three representative keywords: “care”, “health”, and “services”. Now we used the trained word2vec model to provide 5 similar words by applying wv.most_similar method.

The word2vec model provided the five most relevant words for each of these keywords, to improve our understanding of industry-specific words.

- Most relevant words for **'care'**:
  - healthcare: 0.8272
  - postacute: 0.8132
  - nonacute: 0.8043
  - employerfunded: 0.7863
  - homebased: 0.7820
- Most relevant words for **'health'**:
  - canadians: 0.7877
  - health: 0.7777
  - care: 0.7696
  - dhhs: 0.7573
  - nonhealth: 0.7560
- Most relevant words for **'services'**:
  - softwarerelated: 0.8431
  - valueadded: 0.8169
  - education: 0.8084
  - atmrelated: 0.8078
  - aftersale: 0.8046

Part3- Comprehensive Analysis of One Sample Firm

F. Competitor Analysis/Historical Analysis

In this section, we focus on Tenet Healthcare Corporation (gvkey = 7750) from the health services industry. The reason we narrowed down to Tenet is that, in the raw data file we have, it is one of only two firms in the healthcare sector for which we have data for all the years, from 1994 to 2020. The analysis is divided into two parts. First, we conduct a competitor analysis, comparing Tenet Healthcare Corporation with key industry peers. In the second part, we perform a historical analysis of Tenet Healthcare Corporation itself to examine its performance over time.

For competitor analysis, we used ChatGPT-4 to generate firm-level embedding scores based on the TF-IDF scores and Word2Vec model that were generated in the earlier sections.

Below are three code snapshots for the process:

a. In the first snapshot, a new column ("firm_level_embedding") is added to the DataFrame "selected_firms_10k_reports.
b. In the second snapshot, a new DataFrame named "top_5_competitors" is generated to store the data of the top 5 competitor firms based on their firm-level embedding scores relative to Tenet Healthcare Corp.
c. In the final snapshot, a new DataFrame named "new_selected_firms_df" is created to store data for Tenet and its 5 competitor firms. This DataFrame is then used to generate graphs for the analysis.

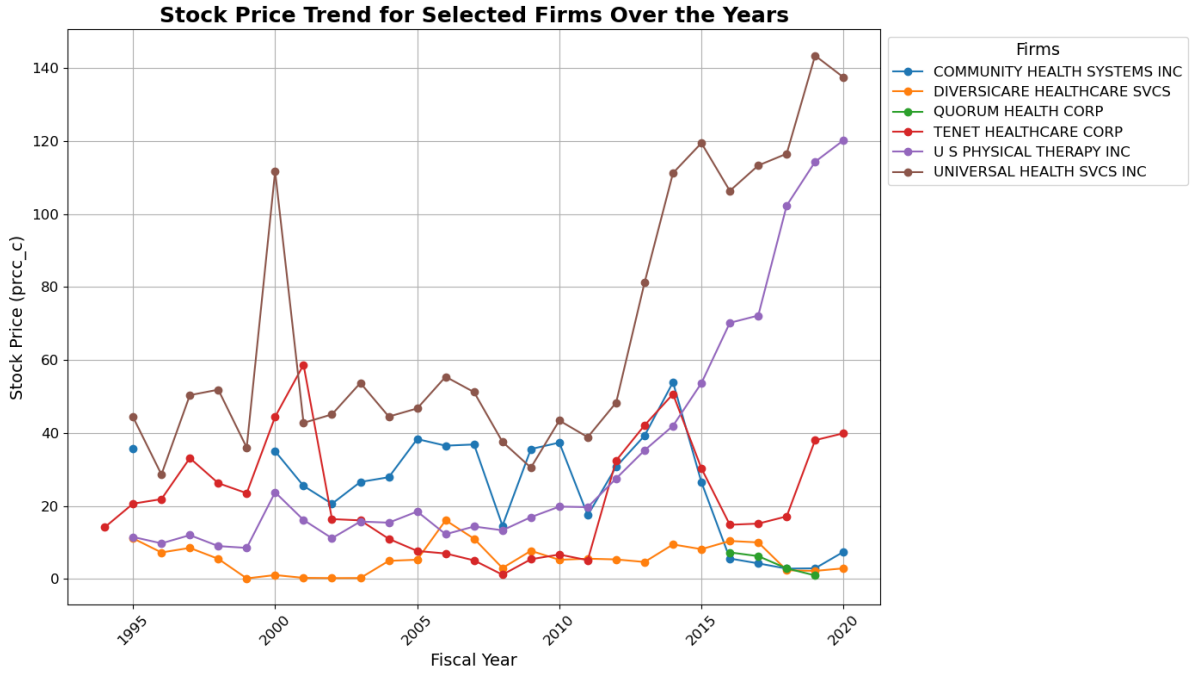
```
#I used ChatGPT 4 to add a new column ("firm_level_embedding") to the dataframe ("selected_firms_10k_reports"). My prompt was: 'Here is my code for g

import numpy as np
from gensim.models import Word2Vec
from sklearn.feature_extraction.text import TfidfVectorizer

# Convert keywords to embeddings and aggregate them
def get_firm_embeddings(keywords_list, model):
    firm_embeddings = []  # List to store firm-level embeddings
    vector_size = model.vector_size  # The size of word vectors in Word2Vec

    for keywords in keywords_list:  # Loop through each firm's keywords
        word_embeddings = []  # List to store word embeddings for current firm

        for word in keywords:  # Loop through each keyword
            if word in model.wv:  # Check if the keyword exists in Word2Vec vocabulary
                word_embeddings.append(model.wv[word])
            else:
                # If the word is not in the vocabulary, you can either skip it or use a zero vector
                word_embeddings.append(np.zeros(vector_size))

        # If we have any embeddings for the firm, aggregate them (e.g., by averaging)
        if word_embeddings:
            firm_embedding = np.mean(word_embeddings, axis=0)  # Aggregate by taking the mean
            firm_embeddings.append(firm_embedding)
        else:
            # If no embeddings were found (e.g., no valid keywords), append a zero vector for the firm
            firm_embeddings.append(np.zeros(vector_size))

    return firm_embeddings

# Get firm-level embeddings
firm_level_embeddings = get_firm_embeddings(selected_firms_10k_reports['keywords_tfidf'], w2v_model)

# Adding the firm-level embeddings to the DataFrame
selected_firms_10k_reports['firm_level_embedding'] = firm_level_embeddings

# Displaying the first few rows with firm-level embeddings
selected_firms_10k_reports.head()
```

```python
#I used ChatGPT-4 to generate a new dataframe, 'top_5_competitors_display,' to store the top 5 competitor firms based on their similarity scores. M

import numpy as np
from sklearn.metrics.pairwise import cosine_similarity

# Step 1: Get the embedding for "TENET HEALTHCARE CORP" (gvkey = 7750)
tenet_embedding = selected_firms_10k_reports.loc[selected_firms_10k_reports['gvkey'] == 7750, 'firm_level_embedding'].values[0]

# Step 2: Create a matrix of all firm embeddings
firm_embeddings_matrix = np.vstack(selected_firms_10k_reports['firm_level_embedding'].values)

# Step 3: Calculate cosine similarities between TENET HEALTHCARE CORP and all other firms
similarities = cosine_similarity([tenet_embedding], firm_embeddings_matrix)[0]

# Step 4: Add similarities to the dataframe
selected_firms_10k_reports['similarity_to_tenet'] = similarities

# Step 5: Find the top 5 competitors by sorting similarity scores (excluding TENET itself)
top_5_competitors = selected_firms_10k_reports[selected_firms_10k_reports['gvkey'] != 7750].nlargest(5, 'similarity_to_tenet')

# Display the top 5 competitors
top_5_competitors_display = top_5_competitors[['gvkey', 'similarity_to_tenet', 'name']]

top 5 competitors display
```

```python
# List of gvkeys from the top 5 competitors, including TENET HEALTHCARE CORP (gvkey = 7750)
selected_gvkeys = [11032, 26157, 25318, 30175, 23714, 7750]

# Filter the selected_sector_firms dataframe to include only these gvkeys
new_selected_firms_df = selected_sector_firms[selected_sector_firms['gvkey'].isin(selected_gvkeys)]
```

a. Competitor Analysis:

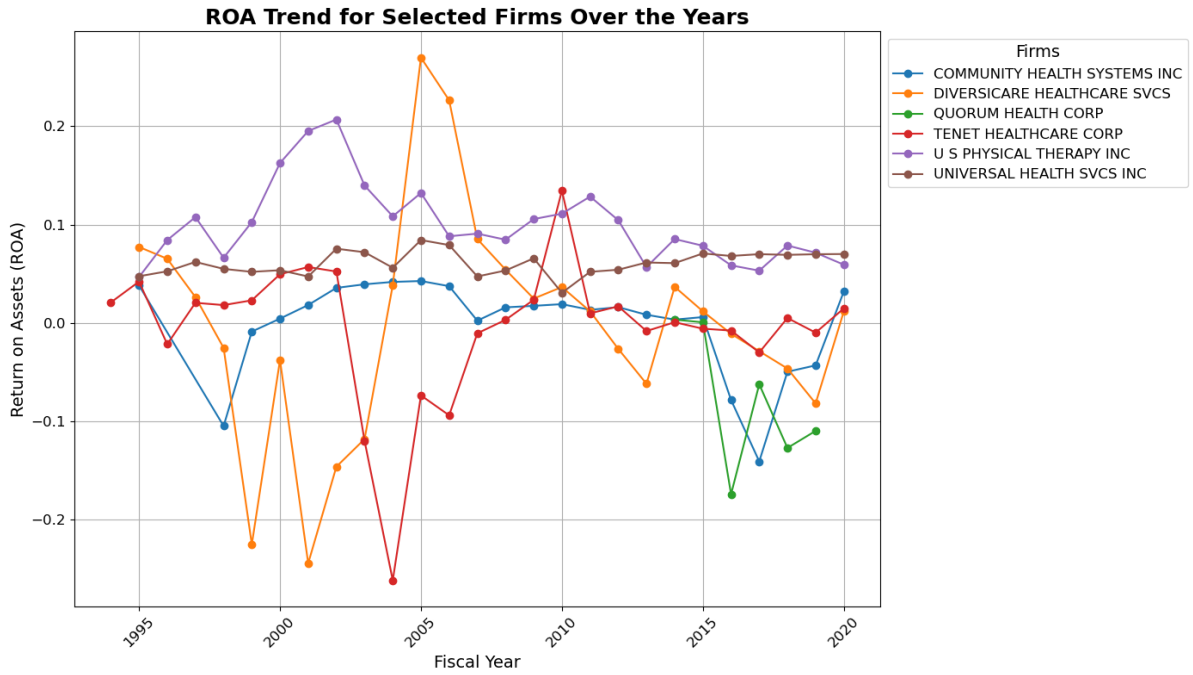
i. Stock Price Analysis:



Stock Price Trend (1994-2020):

Tenet's stock price showed modest growth after 2012, peaking at around $40 by 2020. Competitors like United Health Services and US Physical Therapy have a much better stock price ($120-$140), indicating stronger investor confidence. On the other hand, Community Health Systems, saw a decline in stock price after 2015, dropping below $10 in 2020.

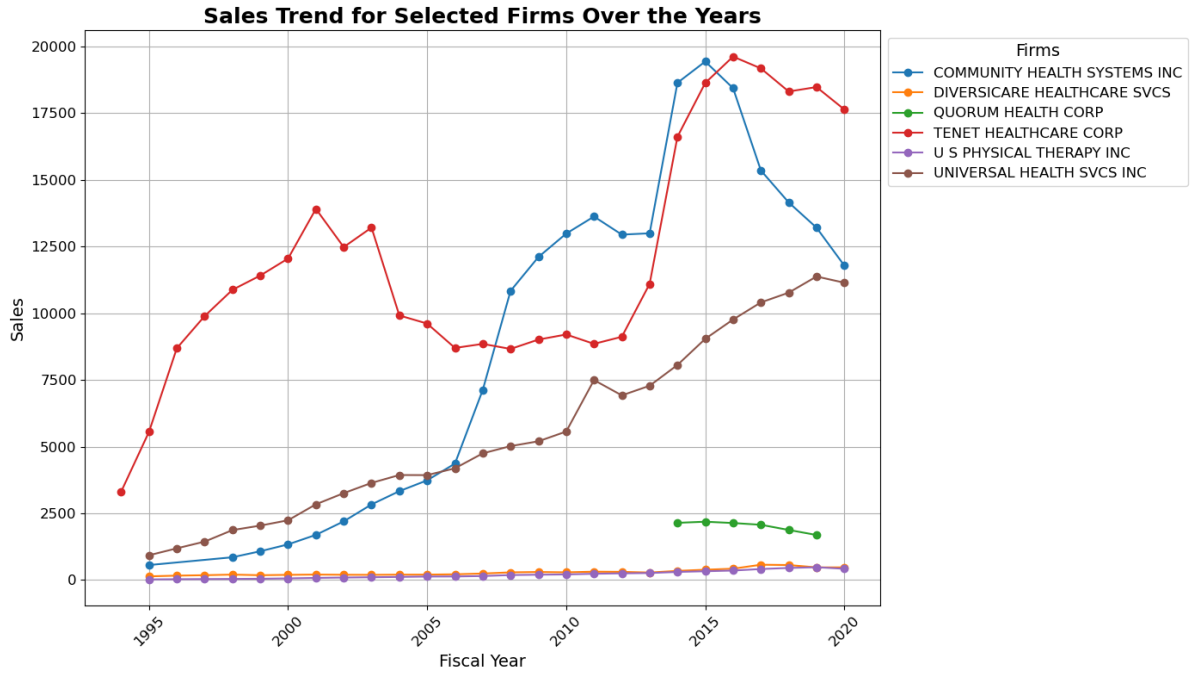
ii. ROA Analysis:



ROA Trend (1994-2020)

Tenet's ROA hovers around 0.05, reflecting moderate asset efficiency. The more or less flat ROA combined with a relatively low stock price indicates that investors perceive Tenet as underperforming. US Physical Therapy shows the strongest ROA, consistently hovering between 0.1 and 0.2. Community Health Systems ROA declined significantly after 2005, even dipping into negative territory. Diversicare Healthcare Services experienced volatile ROA with significant peaks and troughs.

iii. Sales Trend (1994-2020)



Tenet shows steady sales growth from 1994-2015, reaching over $17 billion before experiencing a decline post 2015. Despite the decline, Tenet remains one of the top performers in terms of sales volume. Community Health Systems exhibited something similar as well. Post 2015 it suffered a steep decline. Universal Health Service displays a steady sales growth reaching around $12 billion by 2020. US Physical therapy annual sales are relatively quite low at around $1 billion on an average.

iv. Correlation Across the three metrics:

- Tenet Healthcare Corp:

  Stock Price vs Sales: Despite high sales volume, Tenet's modest stock price growth reflects potential **operational inefficiencies or investor scepticism about the company's ability to convert sales into profit**. The flat ROA (indicating steady asset efficiency) further suggests that Tenet **may not be maximizing the profitability of its large sales base**.

  Sales Decline and Stock Price Stagnation: After 2015, Tenet's declining sales are mirrored by a lack of significant stock price growth, reinforcing concerns about the firm's ability to maintain long-term profitability.

- Universal Health Services:

  Sales, ROA, and Stock Price Alignment: Universal Health has achieved a balanced alignment between sales growth, ROA, and stock price. This indicates that **the company's operational efficiency and steady growth strategy are being rewarded by investors**, leading to strong stock performance.

  Investor Confidence: High stock prices indicate that investors have long-term confidence in Universal Health's ability to continue growing its market share and maintaining operational efficiency.

- U S Physical Therapy:

  High ROA and Stock Price Despite Low Sales: U S Physical Therapy's relatively low sales compared to Tenet or Universal Health have not hindered its stock price. Its high ROA reflects **efficient operations and profitability, which investors highly value, allowing the firm to achieve a high stock price relative to its sales**.

  Profitability Focus: **Investors are clearly focusing on profit margins and asset utilization over raw revenue**, suggesting that U S Physical Therapy's management is maximizing return on its smaller asset base.

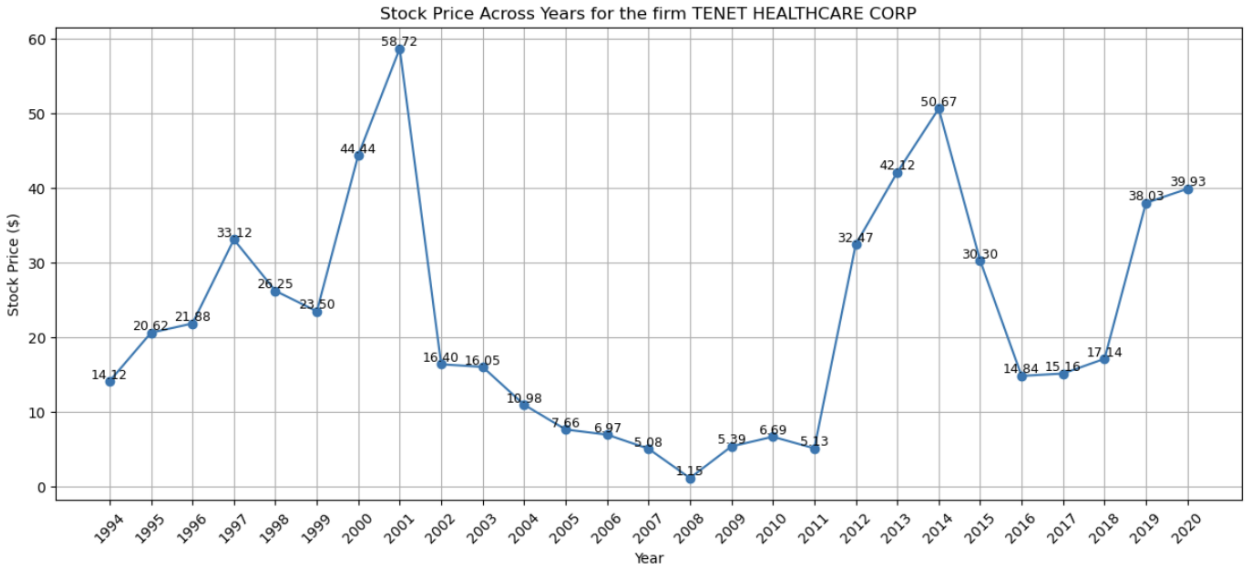
- Community Health Systems:

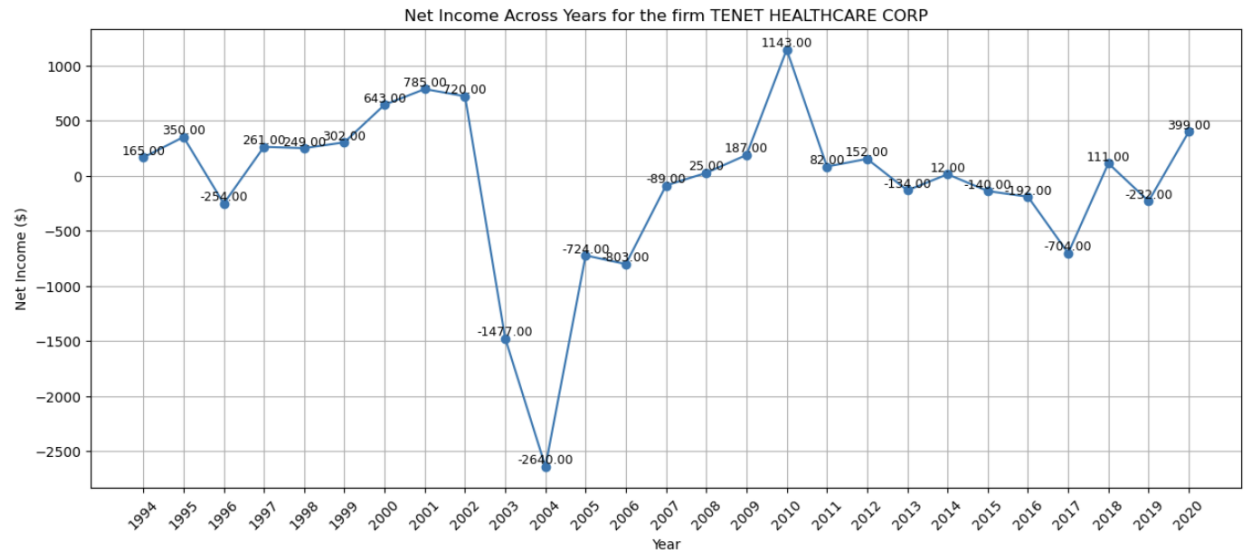
  Declining Across All Metrics: The company shows declining performance across sales, ROA, and stock price, reflecting financial distress. Its ROA dipping below zero, combined with falling sales and stock price, indicates fundamental issues in the firm

b. Historical Analysis

For this section, we focused on **Tenet Healthcare Corporation** with gvkey = 7750 from the health services industry. Our analysis included examining its financial data status through time analysis of its historical stock prices, net income, sale, and assets.

To do so, we use matplotlib package to plot line charts for stock price, net income, sales, and asset. Here are the results for the stock price and net income:





By analyzing these two charts, we can see that **the stock price and net income have some correlations**. This correlation is somehow intuitive since the pricing of the stock market is mainly based on the methods of evaluation based on the potential incomes (like NPV).

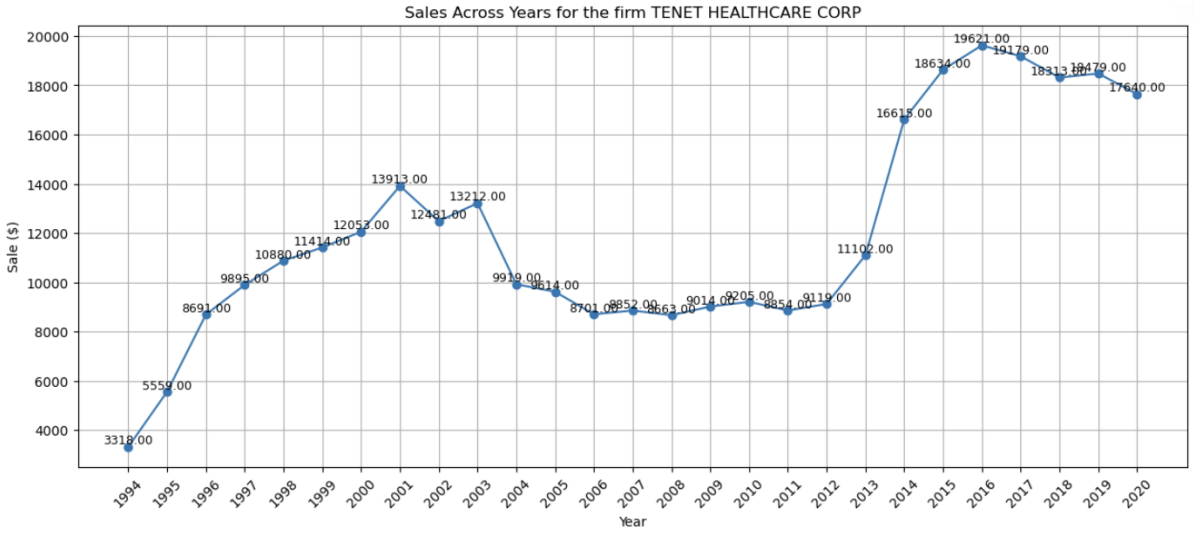
Reasons for Stock Price and Net Income Fluctuations:

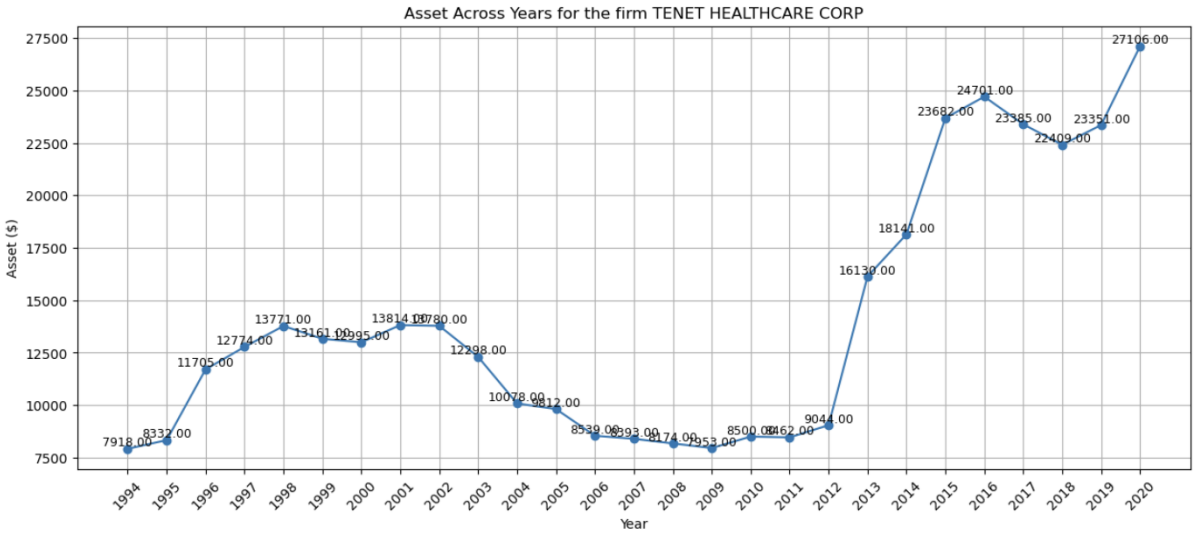
- Major economic events, such as **the dot-com bubble burst (2000-2002) and the global financial crisis (2007-2009), likely had an impact on Tenet's performance** which is visible in both the stock price and the net income.
- Tenet Healthcare has faced various **legal challenges and regulatory scrutiny** over the years. For example, there were **investigations into Medicare fraud around the early 2000s, which could have impacted the company's financial performance and investor confidence, contributing to the dip in stock price around that time**.

Correlation between the Stock Price and the Net Income:

- Based on observations from the graph, both net income and stock price experienced significant peaks during 2000-2001.
- In 2006-2007, the net income dropped to negative, whereas stock price was not as dramatically affected. This suggests that investors may have been anticipating a recovery.
- 2013-2014: Both net income and stock price show a peak. This could reflect a period of growth or strategic success for the company.
- Post 2014: We see a visible and steady recovery in both stock price and net income after some volatility in the early 2010s, indicating alignment between company performance and market valuation.

Now we will analyze the charts for sales and asset.





Reasons for Sales and Asset Fluctuations:

- **1990-2002, Expansion**: Both of them show a significant growth. This period likely reflects expansion and capital investments, where the increase in assets lead to revenue growth.
- **2003-2012, Scandal and Financial Crisis**: Both metrics stagnated, possibly due to the Medicare scandal and the financial crisis of 2007-08.
- **2012-2015, Rapid growth**: A sharp increase in both sales and assets occurred, pointing a significant investment in new assets. This indicates a period of rapid growth, where Tenet capitalized on healthcare demand (possibly influenced by the Affordable Care Act). The direct correlation here is strong, as more assets translated directly into higher revenues.
- **Post 2015, Stabilize:** Both the metrics have plateaued. This suggests that Tenet reached a stable phase, where further large investments in assets weren't necessary for maintaining its revenue levels.

Strategic Suggestions:

Based on competitor and historical analysis of Tenet Healthcare Corporation, here are two strategic suggestions that could be provided to the management:

i. Improving Operational Efficiency and Profitability:

The analysis shows that Tenet Healthcare Corporation's stock price and Return on Assets (ROA) remain modest compared to competitors like US Physical Therapy and Universal Health Services. **While Tenet has a large sales base, its relatively flat ROA and stock price suggest inefficiencies in converting revenue into profit.**

Suggested Action:

- **Cost optimization**: Conduct an in-depth analysis of operational expenses and implement cost saving initiatives. Streamline operations where needed.
- **Digitalization**: Leverage technology and digital health solutions to improve operational efficiency. Implement automation in administrative and clinical workflows to reduce costs and improve service delivery.

ii. Strengthening Investor Confidence and Market Position:

The competitor analysis indicates that Tenet has suffered from investor skepticism, which is reflected in its stagnant stock price and lower ROA compared to peers. Moreover, the decline in sales and assets post-2015 raises concerns about long-term growth prospects.

Suggested Action:

- **Improve Risk Management and Compliance:** As the historical analysis points out, Tenet has faced legal challenges and compliance issues in the past. Enhancing risk management processes and internal auditing could prevent future regulatory and legal setbacks. Establishing a proactive compliance framework would also strengthen Tenet's reputation and reduce risks that could negatively affect investor sentiment.

iii. Increasing internal regulatory auditing

There are three main approaches to increase regulatory auditing and we suggest that the company use all of them:

- **Enhance Knowledge on Regulation:** Improving knowledge of the managers and decision makers about the regulations and standards and the consequences of misconducts
- **Establish Risk Reporting Channels:** Providing feedback channels and whistle blower processes for patients and employees to share risky incidents
- **Leverage Data Science Tools for Risk Detection:** Using data analytics techniques such as machine learning, anomaly detection, classification and etc. to analyze patterns in the service data and find susceptible incidents

Conclusion

This project analyzed the health services sector with a focus on Tenet Healthcare Corporation, combining financial and text-based data. Our competitor analysis revealed that Tenet lags behind peers like US Physical Therapy and Universal Health Services in stock price growth and ROA, signaling **operational inefficiencies**. Historical analysis showed growth but highlighted challenges such **as legal issues and stagnant performance after 2015**.

To address these issues, we recommend: 1) **improving operational efficiency** through cost optimization and digital transformation, and 2) **enhancing investor confidence** by strengthening compliance, and 3) **strengthening internal training on risk auditing**. These actions could improve Tenet's market position and long-term healthy and sustainable growth.