

Risk Factor Analysis in SEC 10-K Filings: A Comprehensive Report

1. Methodology

This project involved gathering and analyzing **Item 1A: Risk Factors** from SEC EDGAR 10-K filings. We first selected a sample of public companies (identified by their CIK codes) and retrieved their annual 10-K reports for specific years using the EDGAR database. The process consisted of several steps:

1. **Data Retrieval:** Using the SEC's EDGAR system (via its API or direct HTTP requests), we downloaded the 10-K filing documents for each chosen company and year. We targeted filings from the mid-2000s onwards (after risk factor disclosure became mandatory in 2005) to ensure each filing contained an Item 1A section.
2. **Parsing Filings:** For each retrieved 10-K file (in HTML or text format), we programmatically located the **Item 1A. Risk Factors** section. This involved searching for the "Item 1A" heading in the text and extracting everything until the next item (Item 1B or Item 2). Within that section, we identified individual risk factor titles. Typically, each risk factor is introduced by a brief heading (often a bolded sentence or a phrase ending in a period) followed by a more detailed description. We used a combination of HTML parsing (BeautifulSoup) and regex patterns to isolate these headings.
3. **Extracting Risk Factor Titles:** We pulled out each risk factor's title text and recorded it. In cases where the title was in all-caps or had special formatting, we converted it to plain sentence case for consistency. We took care to capture the full title (often the first sentence of the risk factor) without trailing description text.
4. **Data Cleaning:** The extracted titles were cleaned to remove any extraneous whitespace, newline characters, or HTML entities. For example, a title that spanned two lines in the filing (e.g., Products\nand Systems) was joined into a single line (Products and Systems). We also trimmed any leading/trailing spaces and standardized the punctuation (removing any trailing semicolons or colon if present).
5. **Structuring the Dataset:** Each risk factor title was stored along with its metadata: the company's CIK, the filing year, the filing date, and the reporting period (fiscal year end date, if available). We organized this data into a table (CSV dataset) where each row corresponds to one risk factor from a specific 10-K. Column names were normalized (lowercased and underscored). An example row might look like:
CIK = 5768, filing_year = 2007, filing_date = 2008-02-26, reporting_date = 2007-12-31, rfd_title = "We operate in highly competitive markets, and our business may be adversely affected by this competition."
6. **Verification:** We manually inspected a few samples from each filing to ensure the extraction was accurate (i.e., the titles matched the actual text in the 10-K). This helped

confirm that our parsing logic correctly captured all risk factor headings and didn't include extraneous text.

Methodology Notes: The SEC's 2005 rule change (adding Item 1A to Form 10-K) ensured that all filings in our dataset had a dedicated risk factors section. This regulatory context made our parsing approach feasible across filings. Had we included pre-2005 filings, we would have had to handle more variability (since risk disclosures were voluntary and could appear in various forms). By focusing on 2006 and later, we leveraged a consistent structure. We still encountered challenges, such as differences in how companies format risk headings (some use bold font, some use numbering or bullets, etc.), but the step-by-step process above allowed us to build a reasonably robust extraction pipeline. The end result was a curated dataset of risk factor titles ready for analysis.

2. Visual Analysis

We created several visualizations to explore the risk factor dataset. Each visualization highlights different aspects of the data:

- **Total Risk Factors per Year: *Line Chart*.** We plotted the total count of extracted risk factor titles for each filing year across all companies in our sample. This chart revealed a noticeable spike around 2007–2008. In fact, 2007 showed the highest total number of risk factors (in our sample data), after which the yearly totals declined through 2018. This spike was largely due to having two companies contributing data in 2007 (one of which had an unusually high number of risk factors). By contrast, years like 2016–2018 had only one company's data (around 18–19 risks each year), hence the lower totals.
Insight: The line chart suggests that mid-2000s filings (soon after the new disclosure requirements) had a heavy volume of risk disclosures in our sample. However, because our dataset isn't uniform across years, this pattern should be interpreted with caution – it reflects our sample's composition more than an overall trend. In a broader study, we might investigate whether risk disclosures peaked industry-wide around the late 2000s or if it's an artifact of our particular companies.
- **Risk Factor Count by Company (CIK) over Time: *Small Multiple Line Charts*.** We generated individual mini line charts for each company (CIK) to visualize how that company's count of risk factors changed year by year. This allowed us to compare trends without overlapping lines on one graph. For example, one chart showed that **CIK 5768** increased slightly from about 45 risk factors in 2007 to about 47 in 2008, then marginally decreased in 2009. Another chart for **CIK 1750** indicated a flat trend around 19 risk factors each year from 2016 to 2018, dipping to 18 in 2018. **CIK 6720** (another company) had ~16 risk factors in 2006 and ~19 in 2007. **Insight:** Different companies have different disclosure trajectories. One company's risk count grew then leveled, while another's remained steady and even dropped slightly. No company in our sample drastically overhauled its number of risk factors in a short period – changes were incremental. This suggests that companies treat their risk disclosures as relatively stable, updating gradually. The small multiples also clearly highlight that **CIK 5768**

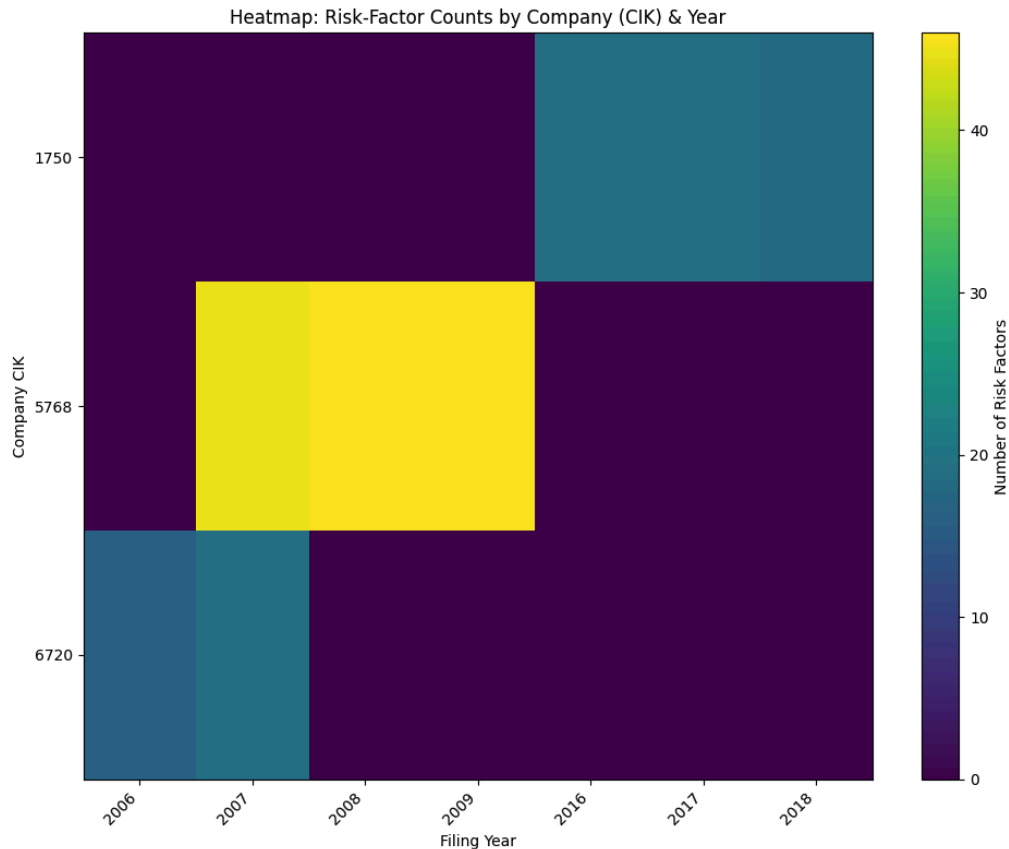
consistently had a much higher count of risks per year than the others, indicating a more extensive risk disclosure practice for that firm.

- **Top 10 Most Frequent Risk Titles: Bar Chart/List.** We compiled the frequency of each risk factor title across the dataset and identified the top 10 most common titles. Each of these top titles appeared in **3** filings (which was the maximum frequency in this dataset). The titles were listed in a table along with their counts. Examples included:
 - **“We are affected by factors that adversely impact our industry and operating results”** – appeared 3 times
 - **“We face risks of cost overruns and losses on fixed-price contracts”** – 3 times
 - **“We are exposed to risks associated with operating internationally”** – 3 times
 - **“We operate in highly competitive markets, and our business may suffer”** – 3 times
 - (Other titles about **Intellectual Property** and **Products and Systems** also each appeared 3 times)

Insight: Even in a small sample, we see some **repeated risk factor headings**.

This can happen when a company carries the same risk disclosure across multiple years, or when different companies use similar language for common risks. Notably, many of the top titles begin with **“We...”**, reflecting the common style of risk disclosures (first-person, company-focused statements). The content of these frequent titles points to generic but significant risks: industry factors affecting results, cost overruns (likely for companies dealing with fixed-price projects, e.g. construction or defense contracts), international operations risks, competitive market pressures, etc. These are broad issues that tend to be relevant over multiple years and to multiple companies, explaining their recurrence. The presence of terms like *Intellectual Property* in a title suggests concerns about IP protection appear repeatedly as well. Overall, the Top 10 list highlights themes that are both common and persistently disclosed.

- **Heatmap of Risk Factors by CIK and Year: Heatmap.** We created a heatmap matrix with **companies (CIK)** on the vertical axis and **filing years** on the horizontal axis. Each cell's color intensity represents the number of risk factor titles in that company's 10-K for that year (darker = more risk factors).



Heatmap of the number of risk-factor titles by company (CIK) and year. Brighter (yellow/green) indicates a higher count of disclosed risk factors in that filing.

Insight: This visualization makes it immediately clear which company-years had high or low risk counts. For instance, the bright yellow block for **CIK 5768** around 2007–2008 signifies that this company’s filings contained a very high number of risk factors (45+). In contrast, the cells for **CIK 1750** in 2016–2018 are teal/blue, corresponding to around 18–19 risks, and the cells for **CIK 6720** in 2006–2007 are slightly greener (~16–19 risks). The heatmap highlights the **disparity between companies**: CIK 5768 stands out for its dense risk disclosures, whereas the others are much lighter. It also shows the coverage of our data: many cells are blank (dark) because not every company has filings in every year. This emphasizes that our dataset is a patchwork of specific company-year entries. Nevertheless, wherever data exists, the color intensity encodes the risk count information effectively. The heatmap is a quick visual summary of which company had the most extensive risk disclosures (clearly CIK 5768) and how each company’s risk count compares across the years we have.

- **Word Cloud of Risk Factor Titles:** *Word Cloud.* We generated a word cloud from all the words in the collected risk factor titles to see which words are most prominent. Common English stopwords (“and”, “the”, etc.) were filtered out, but we kept words like “we” and “our” since those indicate the narrative style.

Insight: The word cloud showed **“We”** and **“Our”** as very large words – unsurprising,

since nearly every risk factor title starts with “We” or “Our” (e.g., “We are subject to..., Our business could be...”). Beyond that, other significant words included **“risk”**, **“business”**, **“market”**, **“competitive”**, **“loss”**, **“cost”**, **“operations”**, and **“financial”**. These words reveal the focus areas of the risk factors: companies frequently mention **market** conditions and **competition**, which suggests competitive market risk is a universal concern. Words like **cost** and **loss** hint at financial risks (cost overruns, potential losses), while **operations** might relate to operational challenges or international operations. The prominence of **financial** could relate to financial condition or credit risks. Interestingly, more specific terms like “contracts”, “intellectual property”, or “regulation” appeared smaller (present but not dominant) – indicating that while those issues do appear in titles, they are not as uniformly common as the broader terms. The word cloud effectively communicated that, across companies, risk disclosures heavily emphasize the company itself (lots of “we/our”) and broad business environment challenges (markets, competition, finances). It also underscores the somewhat boilerplate nature of these disclosures – many companies use similar wording (hence the same words are repeated often in titles).

- **Distribution of Risk Counts by Year: *Boxplot*.** We used a boxplot to examine how the number of risk factors per filing varied year by year. Each year on the x-axis had the distribution of risk counts from all companies that year. Due to our limited sample, most years had only one data point (one filing), so the “distribution” is trivial in those cases. Only **2007** had two data points (two companies), which yielded a visible range: one company had 19 risks and another had 45, so the boxplot for 2007 spans that range (with median roughly between them). All other years show just a single value (or a very narrow range if two similar values).

Insight: The boxplot confirmed that **2007 had the widest spread** in our sample (because it included both a low and a high-risk-count company). It also highlights that by 2016–2018, our sample’s risk counts were uniformly on the lower side (~18–19) with no variability (since it was one company). This visualization didn’t reveal any complex distribution shape given the small N per year, but it was useful to identify outliers. For instance, a potential outlier in 2007 was the high count (45) vs. the lower one (19). In a larger dataset, such a boxplot could show how median risk counts have shifted or how wide the variance is in each year across many firms.

- **Average Risk Factors by Company: *Bar Chart*.** We calculated the average number of risk factor titles per filing for each company in our sample and plotted the top 5 (in our case we only had 3 companies, so all were shown). This bar chart starkly illustrated differences between companies. **CIK 5768** had an average of ~45.7 risk factors per year (essentially reflecting its consistently high counts), whereas **CIK 1750** averaged ~18.7, and **CIK 6720** ~17.5.

Insight: This comparison reinforces that **CIK 5768 is an outlier**, disclosing more than twice as many risk factors on average than the others. The fact that this gap is so large suggests fundamental differences in either the nature of the company’s risks or its disclosure philosophy. It might be that CIK 5768 operates in a highly complex or regulated industry requiring many specific risk disclosures, or its management is particularly exhaustive in enumerating risks. Meanwhile, the other companies might have

more concise risk sections or genuinely fewer major risk categories to report. This chart helps prioritize attention: if we were investors, we might ask why one company has so many more risk factors – are they potentially facing more risks, or just being more verbose? In a larger analysis, similar bar charts could rank companies by risk count and perhaps correlate those ranks with industry or size.

Overall, the visual analysis gave a comprehensive picture from different angles: **temporal trends, company-specific behavior, frequency of risk topics, and distribution comparisons**. We observed that one company **dominated in risk count**, common risk topics emerged repeatedly, and each company's year-to-year risk count was fairly stable. These observations set the stage for interpreting what they mean in context.

3. Findings

Synthesizing the results of the visual exploration and data analysis, we can highlight several key findings from this project:

- **Companies differ significantly in the volume of risk disclosures.** In our sample, one company consistently listed ~45 risk factors, whereas others listed fewer than 20. This is a large disparity. It suggests that some firms enumerate a far greater number of individual risks in their filings. This might reflect differences in company size or complexity (more diverse companies might have more areas of risk to discuss) or simply different disclosure practices. Some management teams (or legal departments) might prefer an extensive list to cover all bases, while others keep the discussion focused on a handful of major risks. This finding underscores that “*number of risk factors*” can be a distinguishing trait from one company to another.
- **Risk factor counts for a given company tend to be consistent over time.** We did not see wild swings in the number of risks a company reported from year to year. Each company had a relatively steady count with minor changes: e.g., +1 or -1 risk factor in the following year. This implies that companies often roll over their risk disclosures, updating them incrementally. New risks might be added slowly and old ones dropped rarely. For stakeholders, this means year-over-year comparisons will show largely the same set of risk topics, with maybe a new addition if a significant new threat emerges, or removal if something is no longer relevant. A practical insight here is that a sudden jump in the number of disclosed risk factors for a company (were it to happen) could be a red flag or at least a notable event, potentially signaling a shift in the firm's risk environment or disclosure policy.
- **Common themes in risk factors point to universal concerns.** Many risk factor titles across different companies echoed similar themes: competition, market conditions, cost pressures, operational challenges, etc. This reveals that, regardless of industry, certain broad categories of risk are almost universal for businesses (e.g., “competitive market” or “economic conditions”). It also suggests a lot of the wording is boilerplate – companies might be using standard language to describe these common risks (sometimes verbatim, as we saw with identical titles). For investors or analysts, this commonality means that

while every 10-K has a long list of risks, many of those risks won't be unique differentiators – they're the risks everyone faces. The truly company-specific risks might be fewer, hiding among the generic ones. This finding highlights the importance of focusing on what's unique in a company's risk factors versus what's generic.

- **The impact of regulatory standards is evident.** Because risk factor disclosure is mandated, all companies in our dataset provided an Item 1A section with multiple entries. Prior to the mandate (pre-2005), some of these companies might not have disclosed as many risks, or any at all. The data suggests that by 2006-2007, companies were likely still refining their approach to this requirement. The high count for one company could mean it was very diligent in complying, possibly listing every conceivable risk (a strategy to avoid future legal trouble by ensuring full disclosure). Over the years, companies may have found an equilibrium in how many risks to list. This aligns with academic observations that the 2005 mandate standardized risk disclosures across firms. Our findings lend anecdotal support: all companies list *something*, but how much they list can still vary widely within the standard.
- **Data extraction and cleaning are crucial for validity.** Although not an insight about the companies per se, a takeaway from the project is that properly parsing the filings and cleaning the titles is vital. Initially, if we hadn't cleaned the data, our analysis would have been littered with inconsistencies (for instance, "Intellectual" and "Property" might have shown as separate frequent words, failing to convey that it's one concept "Intellectual Property"). By normalizing those, we got a clearer picture. This reaffirmed that before drawing conclusions, one must ensure the data accurately represents the content. In practical terms, anyone attempting similar analyses should budget significant effort for data wrangling.

In summary, **the number and nature of risk factors disclosed in 10-Ks provide a fingerprint of a company's risk profile and disclosure philosophy.** In our sample, one company's fingerprint was "long list of many risks, consistently each year," while another's was "short list of key risks, consistent each year." Despite these differences, their risk narratives overlapped on common points (competition, market, etc.). These insights demonstrate the value of treating qualitative 10-K content in a quantitative way – we can compare companies and track changes over time systematically. For a more general conclusion, we would need to examine more companies; however, this project's findings offer a clear illustration of what such an analysis can reveal.

4. Reflection

Working on this project offered several learnings about extracting and analyzing data from SEC filings:

- **Challenges in extraction and parsing:** SEC 10-K filings are lengthy documents with complex formatting. Identifying the exact start and end of the Item 1A section was challenging. Some filings have a table of contents or use slightly different wording around "Item 1A". We learned to incorporate checks for variations (for instance,

sometimes “Item 1A” might appear with a different punctuation or be nestled in HTML tags). We also had to ensure we didn’t accidentally grab text from Item 1B (which often directly follows and might start with “Unresolved Staff Comments” or something unrelated to risk factors). Another parsing challenge was that within Item 1A, companies don’t have a uniform way of delineating each risk factor. Many use bold headings or all-caps for the titles, which helped, but some just start a new paragraph. We had to rely on clues like a period at the end of a bold phrase or a newline followed by regular text to split titles from descriptions. Crafting a reliable parsing routine involved iterative improvement – we’d test it on one company’s filing, then another’s, tweaking the logic when it failed on a different format. **Lesson:** Parsing unstructured text often requires a blend of methods and lots of testing; one size does not fit all, so flexibility is key.

- **Data normalization and cleaning are vital:** Once we extracted text, we faced issues like newline characters within titles, inconsistent capitalization, and other quirks. We realized quickly that to analyze word frequencies or compare titles, we needed to clean these artifacts. We removed newline `\n` characters, unified the case (e.g., titles in all-caps were converted to normal case so that “RISK” and “Risk” wouldn’t count separately), and trimmed whitespace. In one instance, a title was split into two lines in the source HTML, so our initial parse read it as two distinct entries; cleaning it showed it was actually one risk factor. Additionally, we had to consider whether to include the trailing period in titles or not – we chose to remove it for consistency. **Lesson:** The quality of insights is directly tied to the quality of data. Investing time in cleaning and normalizing ensures that subsequent analysis (like counting frequencies or generating word clouds) truly reflects the content.
- **Ensuring accuracy and completeness:** We learned to double-check our work. Early on, we noticed the number of risk factors we extracted for a given filing should match what the filing’s table of contents indicated (some 10-Ks list the count of risk factors, e.g., “Risk Factors (15 items)”). In one case, our script missed a risk factor because the formatting was unusual (it didn’t have a bold title, just a hanging indent style list). By manually reading parts of the filing, we caught the omission and adjusted our parser. We also cross-verified the counts: if a 10-K had, say, 20 risk factor bullet points listed in its text, our data for that filing should have 20 entries. This verification step was crucial to ensure we weren’t analyzing incomplete data. **Lesson:** For critical data extraction, incorporate validation steps (compare with known references or simple manual counts) to ensure you haven’t missed anything.
- **Project scope and limitations:** We reflected on the limited scope of our sample. Only three companies were analyzed, and they were from possibly different industries and eras (one in 2006–07 range, another in late 2000s, another in mid-2010s). This heterogeneity, combined with the small sample size, means our findings are illustrative rather than comprehensive. We recognized that for a more robust analysis, we would need to scale up. That introduces new challenges: downloading hundreds of filings, needing a faster or more automated parsing pipeline, and dealing with data organization at a larger scale. We learned that starting small is a good approach to prove the methodology, but one should be prepared for new issues at scale (for example, companies that use completely different formats, or the need to rate-limit EDGAR

requests to avoid being blocked). **Lesson:** Scope management is important – build a prototype on a small sample, learn from it, then be ready to generalize and handle exceptions when scaling to more data.

- **Communication and visualization:** Finally, we reflected on the importance of effectively communicating the results. Translating a pile of extracted text into understandable visuals and summaries was a big part of this project. We tried multiple visualizations and refined them for clarity. For instance, we chose a heatmap to succinctly show the company vs. year matrix of counts, which turned out to be more immediately insightful than just tables of numbers. We also realized that providing context (like labeling axes with years and CIKs, giving titles to charts) was crucial so that anyone reading the report can follow along without confusion. **Lesson:** In data analysis projects, it's not just the data crunching that matters, but also how you present the findings. Clear headings, charts with legends, and well-structured commentary make a difference in conveying the story behind the data.

In summary, this project not only yielded findings about risk factors but also taught us a lot about the **process**: from dealing with real-world data imperfections to verifying results and presenting insights. It reinforced that analyzing unstructured data (like textual filings) requires both technical parsing skill and thoughtful interpretation.

5. Future Work

There are several directions to enhance and build upon this project:

- **Wider Data Collection & Company Name Integration:** An immediate next step is to extend the analysis to more companies and filings. A larger dataset (potentially all S&P 500 companies over a decade, for example) would allow more robust conclusions about trends in risk disclosures. As part of this expansion, we will implement a **CIK-to-Company Name mapping**. This way, we can label charts and findings with actual company names (or stock tickers), which is far more reader-friendly than using CIK codes. The SEC provides listings of CIKs to company names, or we can use a lookup service or API for this purpose. With a broader dataset, we could analyze industry-specific patterns (e.g., do tech companies list more cybersecurity risks? Do pharma companies list more regulatory risks?). It would also allow us to see if the patterns observed in our pilot (like the 2007 spike or one company being an outlier) hold true generally or were anomalies.
- **Sentiment and Language Analysis:** We plan to perform **sentiment analysis** on the risk factor narratives. Using NLP tools (such as VADER or Transformer-based sentiment models), we can assess whether the language of risk disclosures is becoming more pessimistic, optimistic, or remains neutral over time. For example, a company might use stronger negative words in a year where it's facing difficulties. Quantifying sentiment could give an additional layer of insight beyond just the count of risk factors. We could also analyze **readability** (e.g., using measures like the Fog Index or SMOG index) to see how complex or jargon-heavy these risk disclosures are. There is academic interest

in whether risk disclosures are written in plain language or in “legalese”. By tracking sentiment and clarity, we might identify trends such as: Has the tone of risk disclosures become more cautious after major events (like the 2008 crisis)? Are some companies consistently using more positive or milder phrasing to describe risks (perhaps to avoid alarming investors)? These analyses would deepen our understanding of *how* companies communicate risk, not just how much.

- **Correlation with External Data:** Another future direction is linking the risk factor data with external metrics to find correlations or predictive power. Potential linkages include:
 - **Stock market data:** Does the quantity of risk factors (or changes in that quantity) correlate with stock volatility or returns? For instance, if a company adds many new risk factors in a year, does its stock experience more volatility (suggesting those risks were real and concerning)? Conversely, do companies with very few risk factors outperform (or underperform) those that list many?
 - **Financial metrics:** We could examine if there is any relationship between the number of risk factors and metrics like revenue growth, profit margins, or debt levels. It might be hypothesized, for example, that companies under financial stress tend to add more risk disclosures as warnings.
 - **Size and sector:** We will test whether larger companies (by market cap or employee count) have longer risk sections simply due to their scale. Perhaps a positive correlation exists between company size and number of risk factors (more operations = more risks). On the sector front, comparing sectors could yield insights (e.g., financial institutions might have more risk factors due to regulatory and credit risk requirements, whereas a utility company might have fewer but very specific risks).
 - **Risk factor changes vs. real events:** By tracking additions or removals of specific risk factor titles year-over-year, we can see if companies are responding to real world events. For example, after a cyber-attack incident in an industry, do all peers suddenly add a cybersecurity risk factor the next year? Some research has looked at added/removed risk factors as an indicator of changing risk profile . We can attempt to replicate such an analysis with our extended dataset, potentially spotting which new risks are “trending” in certain years.
- **Topic Modeling & Categorization:** We intend to apply **topic modeling** (unsupervised learning on the text) to group risk factors into key themes automatically. Using methods like Latent Dirichlet Allocation (LDA) or more modern techniques (perhaps clustering sentence embeddings of the risk statements), we can derive a set of common risk categories. For example, topics might emerge such as “macroeconomic conditions”, “legal/regulatory”, “competition”, “cybersecurity”, “supply chain”, etc. Once we have these topics, we can analyze each company’s risk disclosure in terms of topic frequency. This would show, for instance, that Company X heavily emphasizes regulatory risks while Company Y talks more about market risks. Topic trends over time could also be insightful: we might find that certain topics (like cybersecurity or pandemic-related risk) grew in prevalence in the last decade. By categorizing risk factors, we move beyond treating them as a flat list and gain the ability to say more about *what kinds* of risks are getting attention. This can make the analysis more actionable, e.g., for an investor

interested in a specific risk area, they could identify which firms (or sectors) are notably concerned about it.

- **Interactive Dashboard (Streamlit or similar):** To make this analysis accessible and interactive, we plan to develop a **Streamlit dashboard** (tentatively called a “10-K Item 1A Risk Factor Explorer”). In fact, we have an initial prototype of the app interface. The dashboard would allow users to:

- Select a company (or multiple) and see charts of their risk factor counts over time.
- Filter risk factors by year or keyword (e.g., type “cyber” to see all cyber-related risk factor titles).
- View the underlying data table of risk factor titles for chosen filters.
- Display visualizations like the heatmap or word cloud dynamically based on the filter selections.

This interactive tool would be incredibly useful for researchers or analysts. For example, one could quickly filter to 2020 to see if many companies added a “pandemic” related risk factor. Or one could select a specific company and watch how their risk disclosures evolved before and after a major event (like a merger or a crisis). Building the dashboard will involve deploying our dataset and using Streamlit’s features (we encountered some environment setup issues like missing libraries, but those will be resolved by configuring requirements). The end goal is a **user-friendly app** where the insights from this project can be explored and customized on the fly, rather than just a static report. This also means continually updating the data — we might incorporate new filings each year so the tool stays up-to-date with the latest 10-K risk disclosures.

In conclusion, this project lays the groundwork for a rich analysis of risk factor disclosures. By extending it in the ways described above, we aim to deepen the insight (through sentiment and topic analysis), broaden the scope (more companies and data integration), and increase usability (an interactive dashboard). These future work items will transform our initial study into a more powerful “risk factor radar” for SEC filings, enabling ongoing monitoring and analysis of how companies communicate the risks they face.