Report on Extracting Performance-Based Compensation Data from S&P 500 Companies (1994-2023)

Our research team was tasked with extracting and analyzing performance-based compensation data from S&P 500 companies over a span of nearly three decades, from 1994 to 2023. The objective was to compile a comprehensive dataset that reflects trends and patterns in executive compensation. The process involved multiple phases, including data acquisition, cleaning, extraction, and analysis. This report outlines each step of our methodology, the challenges encountered, and the solutions implemented, providing a detailed account of our research journey.

Step 1: Initial Data Acquisition

The first phase involved acquiring a list of S&P 500 companies for the years 1994-2023 using Wharton Research Data Services (WRDS). This data served as the foundation for our study, ensuring we covered the relevant companies that have been part of the S&P 500 index over the specified period. The data also helped us account for changes in the index, such as companies entering or exiting due to mergers, acquisitions, or other corporate actions.

Step 2: Extracting and Cleaning Ticker Symbols

From the list of companies, we extracted ticker symbols, a crucial step for querying subsequent data. This involved cleaning the data to remove special characters, suffixes, and any inconsistencies. Regular expressions were employed to standardize the ticker symbols, ensuring they were uniformly formatted. This cleaning process was essential to avoid errors in API queries and ensure accurate data extraction.

Step 3: Identifying Relevant SEC Filings

Our preliminary research indicated that the most relevant data for executive compensation, particularly performance-based incentives, was found in DEF 14A filings. These proxy statements provide detailed disclosures on compensation structures and incentives for top executives. This identification was pivotal, as it focused our data extraction efforts on a specific type of filing, streamlining our methodology.

Step 4: Acquiring Access to the SEC API

We faced significant challenges in obtaining access to the SEC API, which is necessary for querying filings from the EDGAR database. After overcoming initial technical issues and securing the API key, we thoroughly reviewed the API's documentation to understand its capabilities and limitations. We also reached out to the SEC API support team for guidance (huge thanks to Jan Schroeder, Founder and CEO of this API), particularly regarding the

extraction of DEF 14A text. Their support was invaluable in resolving early difficulties and establishing a reliable extraction pipeline.

Step 5: Automating the Retrieval of DEF 14A Filings

The next step involved automating the retrieval of DEF 14A filings using the SEC API. We developed a script that queried the API for each company's filings within the specified date range. The script had to handle pagination, comply with rate limits, and manage large volumes of data. Utilizing API documentation and AI tools like ChatGPT, we optimized the script to efficiently fetch filing URLs and other relevant details.

Step 6: Downloading and Organizing HTML Filings

After obtaining the URLs for the relevant filings, we downloaded the HTML content and organized the files systematically. This organization facilitated easy access and management of the large dataset. Each file was named and stored appropriately, ensuring that we could efficiently retrieve and process them in subsequent steps.

Step 7: Extracting Performance-Based Compensation Information

The core task was extracting performance-based compensation information from the downloaded HTML filings. We implemented a function that parsed the HTML content, identifying keywords and phrases associated with performance-based incentives. Regular expressions played a key role in isolating relevant text and extracting associated data, such as percentages indicating compensation. We iteratively refined our keyword list to enhance the accuracy and completeness of our data extraction.

Step 8: Addressing Image-Based Data Challenges

During the data extraction, we encountered instances where compensation details were embedded in images, such as in Walmart's filings. To extract text from these images, we integrated Optical Character Recognition (OCR) technology using the Tesseract library. This addition was crucial for capturing all pertinent data, regardless of its format. The OCR process involved extracting text from images and then parsing this text to identify relevant compensation details.

Step 9: Extracting Data from Tables

In addition to textual and image data, many filings contained structured data in tables. These tables often provided detailed breakdowns of compensation components, critical for our analysis. We developed a function to parse these tables, extracting and organizing the data into a usable format. This step ensured that no valuable data was missed and that the dataset was comprehensive.

Step 10: Sequential Processing and Error Handling

We implemented a sequential processing pipeline to manage the complexity of the extraction process. This pipeline systematically processed each filing, applying the appropriate extraction techniques in a logical order. Robust error-handling mechanisms

were included to manage issues such as incomplete downloads or parsing errors, ensuring data integrity and completeness.

Step 11: Data Management and Storage

Given the extensive nature of our dataset, we implemented a robust data management strategy. We saved extracted data incrementally, ensuring secure storage and backup. This strategy was critical to prevent data loss, especially during instances of API downtime or other technical disruptions. By saving data incrementally, we safeguarded our progress and maintained the dataset's integrity.

Step 12: Parallel Processing and Consolidation

To optimize the data extraction process, we divided the tickers into four groups based on their initial letters (A-F, G-L, M-R, S-Z) and processed them in parallel. This approach significantly reduced processing time. After completing the extraction, we consolidated the data into a single comprehensive dataset, ensuring that all relevant information was captured and systematically organized.

Step 13: Data Cleaning and Preliminary Analysis

The final step involved cleaning the extracted data and conducting preliminary analyses. We focused on identifying and removing duplicates, correcting inconsistencies, and ensuring data accuracy. We also performed initial analyses, such as creating histograms, to explore the distribution of performance-based compensation across different companies and years.

A significant limitation of our study arose from the availability of data through the SEC API. The API only provides access to filings from 2000 onwards. Consequently, we were unable to retrieve DEF 14A filings for the years 1994-1999, resulting in a gap in our dataset. This limitation was beyond our control, and we have acknowledged it in our analysis. Despite this gap, the data from 2000-2023 provides substantial insights into executive compensation trends.

Our research involved a comprehensive process to extract and analyze performance-based compensation data from DEF 14A filings of S&P 500 companies. Each phase, from initial data acquisition to the final analysis, presented unique challenges that required innovative solutions. We successfully navigated these challenges, employing advanced technologies such as OCR and leveraging APIs for data extraction. Despite the limitation of missing data from 1994-1999, our study offers valuable insights into executive compensation practices over the last two decades. This report provides a detailed account of our methodologies and findings, contributing to the broader understanding of corporate governance and compensation strategies.