* 1. Executive Summary - Brief overview of the report's content and key findings
  2. Introduction
     + Background and context of the project
     + Purpose and objectives of the report
     + Outline of the report structure
  3. Data Acquisition and Profiling
     + Description of the data sources and collection methods
     + Summary of the data profiling process
     + Overview of the data quality and any issues encountered
  4. Data Wrangling and Preparation
     + Description of the data cleaning and transformation processes
     + Overview of the tools and libraries used in Python
  5. Data Analysis and Results
     + Presentation of the research questions
     + Description of the analysis methods and techniques applied
     + Summary of the findings and results
  6. Reflections and Recommendations
     + Discussion of the privacy and ethical considerations faced during the project, and the steps taken to address these considerations
     + Evaluation of the overall data acquisition and wrangling process
     + Suggestions for future research within the dataset
  7. Conclusion
  8. References - List of sources cited throughout the report

Assessment 2.2 - Analysis & Summary Report

Data Acquisition and Wrangling

Daniel Thompson,   
1103822831

Bachelor of Data Analytics

Contents

* 1. Executive Summary - Brief overview of the report's content and key findings
  2. Introduction
     + Background and context of the project
     + Purpose and objectives of the report

The Enron scandal marked a significant failure in corporate governance and ethics, resulting in one of the largest bankruptcies in history. Enron's collapse in 2001 was fuelled by deceptive accounting practices orchestrated by top executives like CEO Jeffrey Skilling and Chairman Kenneth Lay. They employed complex financial schemes, notably off-balance-sheet partnerships like Special Purpose Entities (SPEs), to conceal debts and inflate profits, misleading investors and regulators.

Enron's demise led to widespread financial losses for investors and layoffs for employees, exposing flaws in regulatory oversight and corporate governance. The scandal prompted regulatory reforms, including the Sarbanes-Oxley Act, aimed at enhancing transparency and accountability in financial reporting.

As an analyst for the Federal Energy Regulatory Commission (FERC), I was tasked with investigating the collapse of Enron. Supplied with approximately 500,000 emails, I have utilised many data wrangling, cleansing and analysis techniques to answer two questions. By answering these questions inside this report, I hope to gain an insight of the internal dynamics of the company during its collapse.

The two questions I aim to answer in this report are the following:

1. Were Enron employees that were employed in 2001 involved in email exchanges containing any of the listed keywords (Special Purpose Entities, Market-to-market, Off-balance, Fraud, Insider Trading), and were these individuals also among the most active in email communication during that year?

1. Did the word count and number of emails sent or received by employees of Enron drop from 2001 to 2002?
   * + Outline of the report structure
   1. Data Acquisition and Profiling
      * Description of the data sources and collection methods
      * Summary of the data profiling process
      * Overview of the data quality and any issues encountered

The main data source was provided by the University of South Australia (2024), although it was originally bought by Andrew McCallum for $10,000 USD from the FERC (Markoff, 2011). From examining the schema used throughout the emails, it was determined that plain text files of data would be acquired utilising command-line text processing utilities such as *grep*, *wc* and *sed* to filter through the dataset and extract data that met certain criteria.

Problems were encountered when it was noted that the data extracted was not structured that same throughout and that certain python methods (strip(), for example) would need to be depended on to see that correct snippets of data were being utilised.

Another problem of note was when the word count was extracted for each email. Although attempts were made to extract word counts from only the bodies of the email. When these counts were spot-checked, it was determined that the word counts were generally not within 10% of the real count. Therefore, it was determined to count the metadata in the word counts as this was commonly the same length within 2 words throughout the emails, or less than 10% of the word count.

* 1. Data Analysis and Results
     + Presentation of the research questions
     + Description of the analysis methods and techniques applied
     + Summary of the findings and results

Question 1

Were Enron employees that were employed in 2001 involved in email exchanges containing any of the listed keywords (Special Purpose Entities, Market-to-market, Off-balance, Fraud, Insider Trading), and were these individuals also among the most active in email communication during that year?

The first question’s text files extracted each keyword utilising *grep.* A text file was also extracted which contained a list of emails sent or received in 2001 which was used to correlate the list of employees who were working under Enron’s employ at this time.

From this, I was able to count the number of times each of the keywords listed above were used in each person’s email account. As seen below, 59% of employees utilised more than 3 of these words in their email conversations.

A pie chart with numbers and numbers

Description automatically generated

Further calculations found the following per employee:

Mean: 2.69

Mode: 3

Median: 3

From here, the question raised was ‘Was this a coincidence that all of these employees utilised the keywords on just one occurrence, or did the employees utilise these keywords on multiple occasions?’

A graph of a number of employees

Description automatically generated

From the above chart, marked in red are the employees who utilised at least 3 of the 5 keywords marked against their total occurrences of using the keywords. Out of these top 50 employees, 86% of them made use of at least 3 keywords.

A graph of numbers and a number of people

Description automatically generated with medium confidence

Here is the same chart with 4 or 5 occurrences. The number of employees drops to 44%.

Across all 149 employees, the following statistics were found for the number of uses of the keywords:

Mean: 1929.75

Median: 791

Mode: 1386

A graph of a number of keyword occurrence

Description automatically generated

The above charts show that although there is a slightly positive skew, the presence of the mode being in between the mean and median suggests that there is a degree of clustering occurring at the lower end, possibly rendering the high-end entries as outliers when calculating statistics on the set as a whole.

In summary, while 59% of employees utilised 3 or more of the keywords, this group made up 86% of the top 50 employees when measured against how many times they used the keywords. Although the outliers heavily skew the statistics on use of these keywords, the over-use of these keywords has definitely been noticed.

 Did the word count and number of emails sent or received by employees of Enron drop from 2001 to 2002?

* 1. Reflections and Recommendations
     + Discussion of the privacy and ethical considerations faced during the project, and the steps taken to address these considerations
     + Evaluation of the overall data acquisition and wrangling process
     + Suggestions for future research within the dataset
  2. Conclusion
  3. References - List of sources cited throughout the report