SVKM's Narsee Monjee Institute of Management Studies
Mukesh Patel School of Technology
Management and Engineering

FINAL REPORT
ON

NETWORK COLLUSION ANALYSIS
And
PUBLIC DISCLOSURE DATA EXTRACTION

By

AMIT PRAJAPATI
Roll Number: J075
SAP Number: 70041019060

Faculty Mentor
PROF. SIBA PANDA

Industry Mentor
MS. KRISHNA DALAL

A Report On

NETWORK COLLUSION ANALYSIS
And
PUBLIC DISCLOSURE DATA EXTRACTION

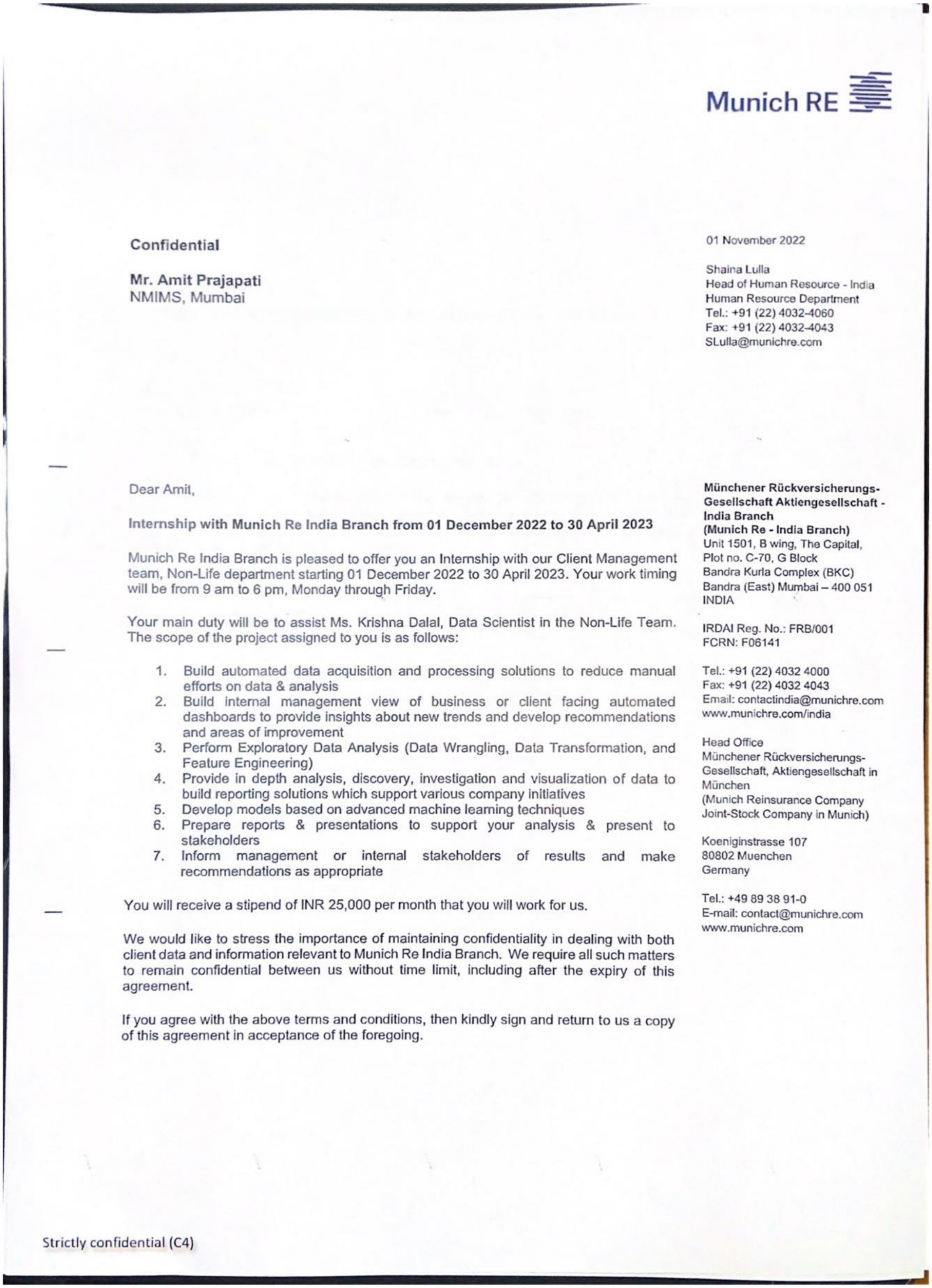
BY

AMIT PRAJAPATI
Roll Number: J075
SAP Number: 70041019060

A report submitted to the NMIMS University in partial fulfilment of the requirement for the award of the degree of “Bachelor of Technology” in “Data Science”.

Department of Data Science
NMIMS University
Mumbai
April 2023

OFFER LETTER



Munich RE

Confidential

Mr. Amit Prajapati
NMIMS, Mumbai

01 November 2022

Shaina Lulla
Head of Human Resource - India
Human Resource Department
Tel.: +91 (22) 4032-4060
Fax: +91 (22) 4032-4043
SLulla@munichre.com

Dear Amit,

Internship with Munich Re India Branch from 01 December 2022 to 30 April 2023

Munich Re India Branch is pleased to offer you an Internship with our Client Management team, Non-Life department starting 01 December 2022 to 30 April 2023. Your work timing will be from 9 am to 6 pm, Monday through Friday.

Your main duty will be to assist Ms. Krishna Dalal, Data Scientist in the Non-Life Team. The scope of the project assigned to you is as follows:

1. Build automated data acquisition and processing solutions to reduce manual efforts on data & analysis
2. Build internal management view of business or client facing automated dashboards to provide insights about new trends and develop recommendations and areas of improvement
3. Perform Exploratory Data Analysis (Data Wrangling, Data Transformation, and Feature Engineering)
4. Provide in depth analysis, discovery, investigation and visualization of data to build reporting solutions which support various company initiatives
5. Develop models based on advanced machine learning techniques
6. Prepare reports & presentations to support your analysis & present to stakeholders
7. Inform management or internal stakeholders of results and make recommendations as appropriate

You will receive a stipend of INR 25,000 per month that you will work for us.

We would like to stress the importance of maintaining confidentiality in dealing with both client data and information relevant to Munich Re India Branch. We require all such matters to remain confidential between us without time limit, including after the expiry of this agreement.

If you agree with the above terms and conditions, then kindly sign and return to us a copy of this agreement in acceptance of the foregoing.

Münchener Rückversicherungs-
Gesellschaft Aktiengesellschaft -
India Branch
(Munich Re - India Branch)
Unit 1501, B wing, The Capital,
Plot no. C-70, G Block
Bandra Kurla Complex (BKC)
Bandra (East) Mumbai – 400 051
INDIA

IRDAI Reg. No.: FRB/001
FCRN: F06141

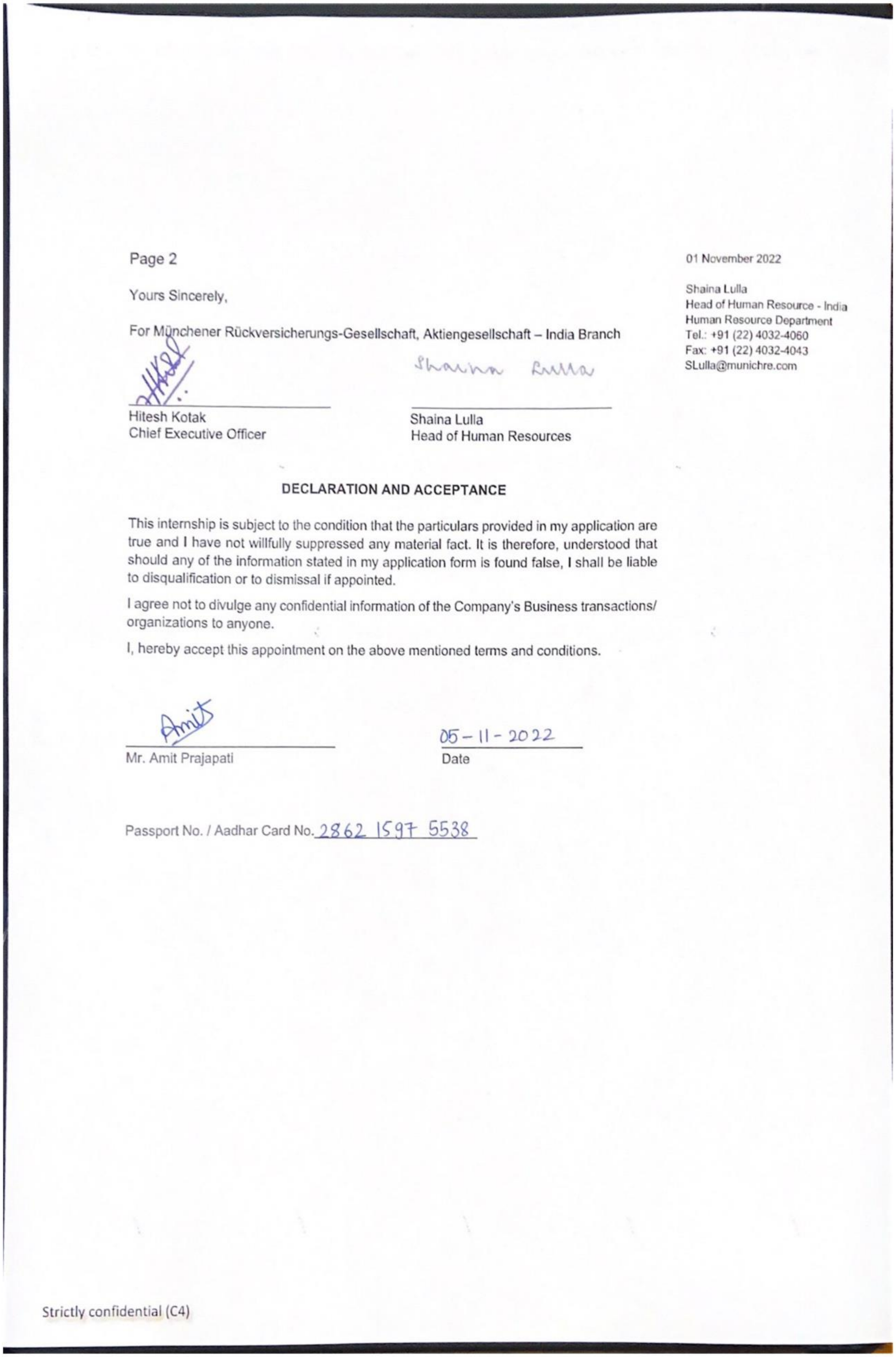
Tel.: +91 (22) 4032 4000
Fax: +91 (22) 4032 4043
Email: contactindia@munichre.com
www.munichre.com/india

Head Office
Münchener Rückversicherungs-
Gesellschaft, Aktiengesellschaft in
München
(Munich Reinsurance Company
Joint-Stock Company in Munich)

Koeniginstrasse 107
80802 Muenchen
Germany

Tel.: +49 89 38 91-0
E-mail: contact@munichre.com
www.munichre.com

Strictly confidential (C4)



Page 2

01 November 2022

Shaina Lulla
Head of Human Resource - India
Human Resource Department
Tel.: +91 (22) 4032-4060
Fax: +91 (22) 4032-4043
SLulla@munichre.com

Yours Sincerely,

For Münchener Rückversicherungs-Gesellschaft, Aktiengesellschaft – India Branch

Shaina Lulla

Hitesh Kotak
Chief Executive Officer

Shaina Lulla
Head of Human Resources

DECLARATION AND ACCEPTANCE

This internship is subject to the condition that the particulars provided in my application are true and I have not willfully suppressed any material fact. It is therefore, understood that should any of the information stated in my application form is found false, I shall be liable to disqualification or to dismissal if appointed.

I agree not to divulge any confidential information of the Company's Business transactions/ organizations to anyone.

I, hereby accept this appointment on the above mentioned terms and conditions.

Amit

Mr. Amit Prajapati

05 - 11 - 2022

Date

Passport No. / Aadhar Card No. 2862 1597 5538

Strictly confidential (C4)

CERTIFICATE

This is to certify that the thesis entitled **“Network Collusion Analysis”** and **“Public Disclosure Data Extraction”** is a Bonafede work of **“Amit Prajapati (SAP ID: 70041019060)”** submitted to the NMIMS University in partial fulfilment of the requirement for the award of the degree of **“Bachelor of Technology”** in **“Data Science”**.

ACKNOWLEDGEMENT

It is with a feeling of great pleasure that I would express my most sincere heartfelt gratitude to **Prof. Siba Panda** and **Ms. Krishna Dalal** for their steady and able guidance throughout my internship tenure.

I am filled with immense pleasure and gratitude as I express my sincere appreciation to
Prof. Siba Panda and **Ms. Krishna Dalal** for their unwavering guidance throughout my internship tenure. Their mentorship has been an invaluable asset to my personal and professional growth. Professor Panda's expertise and passion for teaching have challenged me to think critically and creatively, while Ms. Dalal's industry experience and practical knowledge have provided me with valuable insights. Their constructive feedback and encouragement have fostered a supportive and inclusive environment that has helped me to grow as a professional. I am truly grateful for the opportunity to work under their tutelage, and I will always treasure the knowledge and skills that I have acquired as a result.

INTRODUCTION OF MUNICH RE

Munich is the location of a significant reinsurance corporation called Munich Re. Since 1880, the company has provided forward-thinking risk management solutions. Approximately 40,000 employees are employed across 30 countries by Munich Re's numerous operations.

The insurance industry includes both primary and reinsurance. Munich Re's reinsurance division helps insurance companies mitigate the effects of natural disasters, pandemics, and cyber threats. Its primary insurance segment protects individuals and businesses against respective health, liability, and property risks.

It is common knowledge that Munich Re can manage both complex and extensive risks. By employing sophisticated analytics and data-driven insights, the company aids its customers in making risk reduction and management decisions.

Sustainability and corporate social responsibility are also among Munich Re's top priorities. The company is committed to attaining substantial carbon emission reductions and actively promotes environmentally responsible business practises. In addition, Munich Re provides funding for disaster relief and community development initiatives.

Munich Re is a dependable partner for devising innovative and effective risk management solutions. Munich Re possesses the knowledge, global presence, and commitment to sustainability and social responsibility required to meet the changing needs of its clients in a world enduring rapid transformation.

Table of Contents

ABSTRACT

I worked as a data analyst at Munich Re from December 1, 2022 till now. Network Collusion Analysis and Public Disclosure Data Extraction were client and internal projects I worked on. I used Python and Pyspark to clean, analyse, and build association models to identify network collusion as part of the Network Collusion project. I used association algorithms like Apriori and FP-growth to find network patterns.

I worked on Public Disclosure Data Extraction Automation and Network Collusion. Python was used to pre-process and clean PDFs. I used the Camelot library and other data extraction methods to develop a pipeline to extract relevant information from raw PDF documents. I improved my data analyst abilities by learning end-to-end data pipeline construction, Camelot PDF extraction, and data cleaning. eliminating manual processing and boosting disclosure accuracy.

My internship taught me data analytics and visualisation. Working with big datasets has improved my technical and problem-solving capabilities. The Internship helped me learn big data technologies.

PROJECT-1

NETWORK COLLUSION ANALYSIS

❖ Network Collusion.

- The term "network collusion" is used to describe situations in which multiple actors in a network, such as businesses or people, work together for the benefit of one another. Price fixing, bid rigging, and improper allocation of markets are just a few examples of how collusion can negatively affect businesses and customers.

- The "Libor scandal," which surfaced in 2012, is an illustration of network collusion. Banks use the London Interbank Offered Rate (Libor) as a standard when determining the interest rates for various loans and credit cards. The rate is set by polling a group of financial institutions and is meant to approximate the average interest rate at which financial institutions lend to one another.

- Several institutions were found to have worked together in the Libor scandal to artificially inflate the Libor rate to their benefit. The colluding banks would send misleading data to the survey to artificially raise the rate, which they would then use to justify charging their customers higher interest rates for various financial services. The banks' reputations were severely damaged, and they had to pay billions in penalties and legal settlements because of the scandal.

- Consumers and the economy can both suffer from network collusion, so it's crucial that authorities remain on high alert for signs of it and take swift action when they do.

❖ Network Collusion Analysis.

- Analysis of network collusion is the process of identifying and investigating instances of collusion among actors in a network, which may include individuals, businesses, or other organizations. The analysis typically includes locating patterns of behavior or communication that are indicative of collusion, as well as collecting and analyzing data on market trends, pricing, and other aspects of the business environment that may be indicative of collusion.

- Analysis of network collusion can be accomplished with the help of a number of different strategies and instruments. Methods of network analysis, such as social network analysis and graph theory, are two examples of approaches that can be taken to determine the connections and exchanges that take place between the various players in a network. These techniques can be helpful in locating previously unknown relationships as well as patterns of behavior that may point to the existence of collusion.

- Analyzing large datasets of transactional data, such as financial records or interactions that take place online, can also be done by employing data analytics and machine learning algorithms. This is yet another strategy. These methods can help to identify anomalies or patterns of behavior that are suggestive of collusion, and they can also be used to forecast the likelihood of future instances of collusion happening.

❖ Types of Insurance Frauds.

- The insurance business is susceptible to being victimized by a wide variety of different kinds of fraudulent activity. The following are some of the most widespread kinds of fraudulent insurance claims:

i. Staged Accidents: One form of insurance fraud is known as "staged accidents," and it includes deliberately causing an accident to file a false insurance claim. To submitting a compensation, claim to an insurance provider, a con artist might, for instance, stage an automobile accident or a slip-and-fall incident.

ii. Inflating the Value of a Legitimate Insurance Claim in Order to Receive a Larger payment from an Insurance Company Exaggerated claims involve inflating the value of a legitimate insurance claim in order to receive a larger payment from an insurance company. For instance, a con artist might assert that an object had a higher value than it did, or they might exaggerate how much it would cost to repair damage to a piece of property.

iii. False Claims In the context of this article, "false claims" refers to the practice of making a claim for an occurrence or item that did not actually take place or exist. A fraudster might, for instance, file a claim for the recovery of a stolen object that the victim never possessed or for compensation for an injury that never took place.

iv. Identity Theft: Using the personal information of another individual to submit a fraudulent insurance claim is an example of identity theft. For instance, a con artist may apply for an insurance policy or a claim using the identity of another person, including the person's name and confidential information.

v. Premium Fraud: Premium fraud refers to the practice of making false statements on an application for an insurance policy to obtain more affordable payments. For obtaining a reduced premium, for instance, a crook might give false information about their driving history or their medical history.

vi. Fraud Committed by Insurance Professionals In this form of insurance fraud, insurance professionals such as agents, brokers, and adjusters commit fraud against either an insurance company or the customers of the insurance company. For instance, an insurance agent may offer fraudulent policies to customers or may inflate the value of a client's claim to receive a larger commission. Both practices are done to increase the agent's earnings.

❖ Network Collusion in Insurance.

- When two or more individuals or organizations conspire to commit fraud against an insurance company, this type of scheme is known as network collusion in insurance fraud. The individuals may be insurance professionals like brokers or adjusters, or they may be policyholders who share a common interest in engaging in fraudulent activity. Either way, they may be involved in the scheme.

- This type of fraud can manifest itself in a variety of ways, including the staging of accidents, the exaggeration of claims, or the making of fraudulent claims. The policyholder, the insurance agent, the medical practitioner, and the attorney are all examples of potential participants in the illegal activity of collusion.

- Because the parties involved may take measures to conceal their activities and coordinate their fraudulent schemes, network collusion is particularly difficult to discover and investigate. This is because of the nature of the situation. When looking for signs of possible deception, insurance companies may use a variety of methods, including data analytics, analysis of social network data, and investigative methods, such as surveillance.

- If you are found guilty of insurance fraud, you could face serious consequences such as a monetary fine, time in prison, and the loss of your professional license. In addition to taking criminal action, insurance companies have the option of pursuing civil litigation against individuals who perpetrate fraud against them.

How beneficial can it be?

- An essential part of investigating insurance fraud is conducting network collusion analysis because it enables investigators to discover intricate fraud schemes that involve several different parties. Individuals who commit fraud independently of one another are much less likely to be successful than those who work together to commit fraud, which can result in substantial financial losses for insurance companies when fraudsters collaborate.

- Investigators can identify patterns and connections that would be difficult to detect using conventional investigative methods by analyzing the relationships between the various individuals and entities involved in a fraud scheme. Analysis of network cooperation can assist investigators in the following ways:

  - ➢ Identify key players Investigators can identify the key players in a fraud scheme by conducting an analysis of the network of relationships that exists between the various people and organizations under investigation. This may assist investigators in concentrating their efforts on the most significant targets and developing a case that is more compelling as a result.

  - ➢ Uncover previously unknown connections an investigation into network collusion can unearth previously unknown connections between the various people and organizations participating in a fraudulent scheme. For instance, it can assist investigators in determining whether there is a connection between a policyholder, a medical provider, and an attorney who are all working together to perpetrate fraud.

- ➢ Predict future fraudulent activity Investigators can recognize patterns and trends that may indicate future fraudulent activity by conducting an analysis of previous fraudulent schemes. Because of this, insurance companies may be able to take precautions that will decrease the likelihood of fraud occurring in the future.

❖ Methodology.

▪ Understanding the Data.

- Data understanding was a critical step in network collusion analysis for insurance fraud detection. Without a deep understanding of the data, it was difficult to identify patterns, relationships, and anomalies that may indicate fraudulent behavior.

  - ♦ Identify data quality issues: Data quality issues, such as missing data or inconsistent data, affect the accuracy of network collusion analysis. By understanding the data, it was possible to identify and address these issues.

  - ♦ Discover patterns and relationships: Understanding the data helped us to identify patterns and relationships that might indicate fraudulent activity. For example, if two or more individuals frequently submit claims together, it may suggest collusion.

  - ♦ Detect anomalies: Anomalies, such as unusual claims or claim amounts, may indicate fraudulent behavior. By understanding the data, it is possible to identify these anomalies and investigate further.

  - ♦ Optimize feature selection: Understanding the data can help optimize feature selection, which involves selecting the most relevant variables for the analysis. By selecting the right features, it is possible to improve the accuracy of the network collusion analysis. In our cases features can be Policy-Agent, Customer etc.

▪ Data Pre-Processing.

- When analyzing network collusion, data pre-processing is an essential step to ensure that the data is cleaned, organized, and properly formatted for further analysis. Here are some pre-processing steps that are typically necessary for network collusion analysis:

- Data cleaning: Remove irrelevant or duplicate data, like incomplete or inconsistent records, to ensure the accuracy of the analysis.

- Data normalization: Normalize the data to ensure that all data points are on the same scale. This can involve standardizing numerical data or converting categorical data into numerical data.

- Data discretization: Convert continuous numerical data into categorical data, which can help identify patterns and relationships in the data.

- Data sampling: Depending on the size of the dataset, data sampling may be necessary to select a representative subset of the data for analysis.

- Data encoding: Convert textual or categorical data into numerical data using techniques such as one-hot encoding or label encoding.

- Data aggregation: Group data together to create more meaningful data subsets that can be analyzed together.

  - For example, we can group the data State wise to see an overview, we can granularize the data as much as we want.

Problem Faced During Pre-Processing.

- As someone who has worked with pre-processing huge data for data mining and analysis, I have faced several challenges that can make it difficult and time-consuming. One of the major challenges is ensuring data quality and completeness. I have encountered data with missing values, outliers, and inconsistencies that can affect the analysis. We cannot depend on antiquated approaches to fill in the blanks because this data is related to insurance; rather, we need to identify the fundamental cause of the issue to develop solutions that are going to be successful.

- Another challenge I have faced is scalability and performance. Pre-processing huge data requires significant computational resources and can take a long time to complete. Traditional pre-processing techniques such as sorting, filtering, and joining may not be scalable and efficient for large datasets. To overcome this issue, I have used distributed computing frameworks such as Databricks and Azure Data Lake Storage.

- Data integration and fusion is another challenge that I have encountered while pre-processing huge data. Integrating and fusing data from different sources can be challenging, especially when the data is stored in different formats and has different structures. I have used techniques such as data normalization, aggregation, and schema mapping to address these issues.

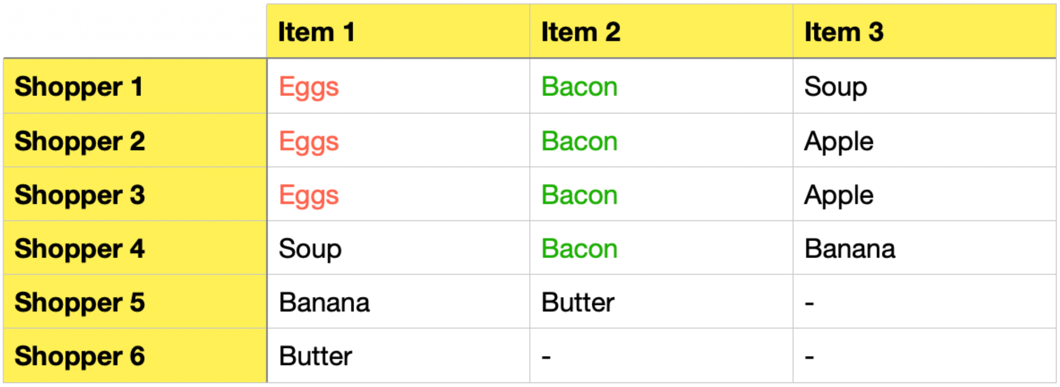
- Data privacy and security is also a major concern while pre-processing huge data. Handling sensitive information such as personal data and financial information requires ensuring that data privacy and security are maintained throughout the pre-processing process.

- Lastly, I have faced challenges in feature selection and dimensionality reduction. Pre-processing huge data often involves selecting relevant features and reducing the dimensionality of the dataset. Traditional feature selection techniques such as correlation analysis and principal component analysis may not be efficient for large datasets. To address these issues, we had to understand the data and figure out what does the data represents and then move on to the further steps.

Working of Association Algorithms

- Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

Association Algorithm Working



	Item 1	Item 2	Item 3
Shopper 1	Eggs	Bacon	Soup
Shopper 2	Eggs	Bacon	Apple
Shopper 3	Eggs	Bacon	Apple
Shopper 4	Soup	Bacon	Banana
Shopper 5	Banana	Butter	-
Shopper 6	Butter	-	-

- Assume we analyze the above transaction data to find frequently bought items and determine if they are often purchased together. To help us find the answers, we will make use of the following 4 metrics:

  - ***Support***

    - The first step for us and the algorithm is to find frequently bought items. It is a straightforward calculation that is based on frequency:

      Support(A) = Transactions(A) / Total Transactions

- Here we can set our first constraint by telling the algorithm the minimum support level we want to explore, which is useful when working with large datasets. We typically want to focus computing resources to search for associations between frequently bought items while discounting the infrequent ones.

- For the sake of our example, let's set minimum support to 0.5, which leaves us to work with Eggs and Bacon for the rest of this example.

- Important: while Support (Eggs) and Support (Bacon) individually satisfy our minimum support constraint, it is crucial to understand that we also need the combination of them (Eggs and Bacon) to pass this constraint. Otherwise, we would not have a single item pairing to progress forward to create association rules.

➢ *Confidence*

- Now that we have identified frequently bought items let's calculate confidence. This will tell us how confident (based on our data) we can be that an item will be purchased, given that another item has been purchased.

  Confidence(A→B) = Probability (A & B) / Support(A)

  Note, confidence is the same as what is also known as conditional probability in statistics:
  P(B|A) = P(A & B) / P(A) Please beware of the notation. The above two equations are equivalent, although the notations are in different order: (A→B) is the same as (B|A).

- In our case these values will be

  Confidence (Eggs→Bacon) **= P(Eggs & Bacon) / Support(Eggs) = (3/6) / (3/6) =**
  Confidence (Bacon→Eggs) **= P(Eggs & Bacon) / Support(Bacon) = (3/6) / (2/3) = 3/4 = 0.75.**

- **Insights:**
  - The above tells us that whenever eggs are bought, bacon is also bought 100% of the time. Also, whenever bacon is bought, eggs are bought 75% of the time.

- ***Lift***

  - Given that different items are bought at different frequencies, how do we know that eggs and bacon really do have a strong association, and how do we measure it? You will be glad to hear that we have a way to evaluate this objectively using **lift**.
  - There are multiple ways to express the formula to calculate lift. Let me first show what the formulas look like, and then I will describe an intuitive way for you to think about it.

    Lift(A→B) **= Probability (A & B) / (Support(A) * Support(B))**

    You should be able to spot that we can simplify this formula by replacing P(A&B)/Sup(A) with Confidence(A→B). Hence, we have:

    Lift(A→B) **= Confidence(A→B) / Support(B)**

  - In our case these values will be

    Lift (Eggs→Bacon) **= Confidence (Eggs→Bacon) / Support (Bacon) = 1 / (2/3) = 1.5**
    Lift (Bacon→Eggs) **= Confidence (Bacon→Eggs) / Support (Eggs) = (3/4) / (1/2) = 1.5**

  - **Insights:**
    - Lift for the two items is equal to 1.5. Note, lift>1 means that the two items are more likely to be bought together, while lift<1 means that the two items are more likely to be bought separately. Finally, lift=1 means that there is no association between the two items.

    - An intuitive way to understand this would be to first think about the probability of eggs being bought P(Eggs)=Support (Eggs)=0.5 as 3 out of 6 shoppers bought eggs.

    - Then think about the probability of eggs being bought whenever bacon was bought: P(Eggs | Bacon)=Confidence(Bacon->Eggs)=0.75 since out of the 4 shoppers that bought bacon, 3 of them also bought eggs.

    - Now, lift is simply a measure that tells us whether the probability of buying eggs increases or decreases given the purchase of bacon. Since the probability of buying eggs in such a scenario goes up from 0.5 to 0.75, we see a positive lift of 1.5 times (0.75/0.5=1.5). This means you are 1.5 times (i.e., 50%) more likely to buy eggs if you have already put bacon into your basket.

Model Building.

Apriori

- The Apriori algorithm works by generating a list of all possible itemsets and then pruning the list based on their frequency in the database. The algorithm starts with frequent itemsets of size 1, and then iteratively generates candidate itemsets of size k+1 from the frequent itemsets of size k. The support of each candidate itemset is then calculated, and those that have a support above a predefined minimum threshold (minus) are added to the list of frequent itemset. This process continues until no more frequent itemsets can be generated.

- Once the frequent itemsets are identified, association rules can be derived from them. Association rules express the relationship between different items in the database. For example, if customers who buy bread also tend to buy butter, an association rule might be "if a customer buys bread, then they are likely to buy butter". Association rules are usually expressed in the form of if-then statements, where the antecedent is the set of items on the left-hand side of the rule and the consequent is the set of items on the right-hand side of the rule.

- The Apriori algorithm is efficient because it reduces the search space by eliminating infrequent itemsets early in the process. However, it can still be computationally expensive for large databases, especially if the minimum support threshold is set too low. There have been several improvements and extensions to the Apriori algorithm, including the use of parallel computing and tree-based data structures, to overcome these limitations.

- ***Advantages:***
  - Easy to understand: The Apriori algorithm is simple and easy to understand. It is widely used in industry and research due to its simplicity.

  - Scalable: The Apriori algorithm can handle large datasets and is efficient for mining frequent itemsets.

  - Widely applicable: The Apriori algorithm can be applied to a wide range of applications, such as market basket analysis, web log analysis, and bioinformatics.

- ***Disadvantages:***
  - High memory usage: The Apriori algorithm requires a large amount of memory to store the itemsets and candidate itemsets. As the dataset size increases, the memory usage also increases.

  - Computationally intensive: The Apriori algorithm can be computationally intensive, especially when dealing with large datasets. It requires multiple passes over the data to generate candidate itemsets and determine their support.

- Low efficiency for low support values: The Apriori algorithm is not efficient for mining infrequent itemsets, as it generates a large number of candidates itemsets that do not meet the minimum support threshold.

♦ F-P Growth

- Frequent Pattern growth (F-P growth) is an algorithm for mining frequent itemsets in large datasets, introduced by Han et al. in 2000. F-P growth is an improvement over the classic Apriori algorithm in terms of efficiency and scalability.

- The F-P growth algorithm is based on a data structure called the FP-tree (Frequent Pattern tree), which stores the itemsets and their frequencies in a compact way. The FP-tree is constructed by scanning the dataset only once and grouping together transactions that share common items. Each group of transactions is then represented as a single path in the tree, where the nodes represent the unique items in the transactions and the edges represent the order in which the items appear.

- After constructing the FP-tree, the algorithm recursively mines frequent itemsets from the tree using a divide-and-conquer strategy. The algorithm begins by finding the frequent items in the tree, which are the ones that appear in transactions with a frequency greater than or equal to the minimum support threshold. The frequent items are then used to generate conditional FP-trees, which are smaller sub-trees that contain only the frequent items from the original tree.

- The algorithm then recursively applies the same procedure to each conditional FP-tree until no more frequent itemsets can be found. The frequent itemsets are then combined to generate association rules, as in the Apriori algorithm.

- The F-P growth algorithm has several advantages over the Apriori algorithm. Firstly, it requires only one pass over the dataset, which reduces the I/O cost and memory requirements. Secondly, it uses a compressed representation of the dataset, which reduces the size of the data and the search space. Finally, it does not generate candidate itemsets explicitly, which reduces the computational overhead.

- However, the F-P growth algorithm may not perform well when the dataset has many distinct items, or when the frequent itemsets have a low support.

- ***Advantages:***
  - Efficient for large datasets: The F-P algorithm is very efficient for mining frequent itemset from large datasets, as it only requires one scan of the dataset to generate the frequent patterns.

  - Low memory usage: The F-P algorithm uses a compact data structure called a frequent pattern tree (FP-tree) to store the itemset and their support, which can significantly reduce the memory usage compared to the Apriori algorithm.

- Fast: The F-P algorithm is fast and can handle large datasets with high efficiency, which makes it a suitable choice for mining frequent patterns in real-world applications.

- ***Disadvantages:***
  - Limited applicability: The F-P algorithm is designed for mining frequent itemsets only, and it cannot be used for other data mining tasks such as clustering, classification, or regression.

  - High preprocessing overhead: The F-P algorithm requires preprocessing the dataset to build the FP-tree, which can be time-consuming for large datasets.

  - Difficulty handling sparse datasets: The F-P algorithm can struggle with sparse datasets, where the support values for most itemsets are very low, which can lead to the generation of many candidates itemset.

♦ Difference between Apriori and FP Growth Algorithm

APRIORI	FP GROWTH
• Apriori generates frequent patterns by making the itemset using pairings such as single item set, double itemset, and triple itemset.	• FP Growth generates an FP-Tree for making frequent patterns.
• Apriori uses candidate generation where frequent subsets are extended one item at a time.	• FP-growth generates a conditional FP-Tree for every item in the data.
• Since apriori scans the database in each step, it becomes time-consuming for data where the number of items is larger.	• FP-tree requires only one database scan in its beginning steps, so it consumes less time.
• A converted version of the database is saved in the memory	• A set of conditional FP-tree for every item is saved in the memory
• It uses a breadth-first search	• It uses a depth-first search.

PROJECT-2

PUBLIC DISCLOSURE DATA EXTRACTION

❖What is Public Disclosure.

- The responsibility of an insured person to provide accurate and complete information about their risk profile and any relevant factors that may impact their insurability, or the cost of their insurance policy is what is meant by the term "public disclosure" in the context of the insurance industry. This information is generally supplied by the policyholder after they have filled out an application or questionnaire.

- Disclosure to the public serves the purpose of facilitating an accurate risk assessment on the part of the insurer in relation to the provision of coverage to the insured individual and the establishment of appropriate premiums. It is possible for the insurer to have grounds for cancelling the policy or denying a claim in the event of a loss if the insured person fails to disclose essential information or provides false information.

- Disclosure to the public is an essential part of the insurance business, as well as a legal requirement in many countries and regions. Insured individuals have a responsibility to provide information that is both truthful and accurate to maximize the likelihood that their insurance claims will be honored if they suffer a loss protected by their policies. Failure to do so can result in serious financial repercussions and legal liabilities.

❖ Why is Public Disclosure useful for Re-Insurance.

- Disclosure to the public is also essential for the practice of reinsurance, which refers to the transfer of a portion of an insurance company's risk to another insurance company for the purpose of lowering the insurance company's overall risk of experiencing a loss. To evaluating the risk connected with the policy, as well as determining the suitable amount of re-insurance coverage and premiums, re-insurance companies depend on public disclosures made by the primary insurer that are accurate and comprehensive.

- Reinsurance companies need to comprehend the risk profile of the primary insurer's policyholders as well as the potential losses that may result from claims in order to determine the appropriate level of reinsurance coverage and premiums. This is because the risk profile of the primary insurer's policyholders is directly related to the potential losses that may result from claims. Additionally, it is necessary for them to make certain that the primary insurer has accurately evaluated and priced the risk that is affiliated with their policyholders.

- It is possible that re-insurance companies will not have a clear comprehension of the risks associated with the policies they are covering if there is not accurate and complete public disclosure. This could lead to an insufficient amount of reinsurance coverage or an overcharging of premiums, both of which could eventually have an impact on the financial stability of both the primary insurer and the reinsurance company.

- In conclusion, public disclosure is extremely important for the reinsurance industry because it serves to ensure that reinsurance companies have a clear understanding of the risks they are covering and can set appropriate premiums to provide adequate coverage. In other words, public disclosure helps to ensure that reinsurance companies have adequate coverage. Because it helps to reduce the financial risks associated with insurance coverage, this ultimately is beneficial for both the original insurer and the re-insurance company.

❖ Problem Statement.

- The process of the client manager manually downloading PDFs and typing data into Excel produces several roadblocks that hinder the organization’s ability to function efficiently and effectively.

- To begin, the manual process takes an extremely long time, and because of the amount of time that is spent on data entry, the client manager is unable to prioritise other tasks that are equally or more essential. The monotony and lack of motivation that result from the repetitive nature of this activity can ultimately have a negative impact on overall productivity.

- Second, the manual process is prone to errors, and the risk of missing or inaccurate data increases with the amount of data that is processed. This is because the manual process is performed by human hands. Because of this, the research and decisions that are made might not be accurate, which could have significant repercussions for the company.

- Thirdly, the manual process is not scalable, and as the amount of data increases, it becomes increasingly difficult to handle. This is since the manual process is not scalable. The process also has the potential to become a bottleneck for the organization, which will result in delays in decision-making and will cause frustration among the other employees.

- Lastly, ever company publishes around 40-50 pdf and it is get very tedious for the client manager to work and manage the data.

❖ Solution Proposed.

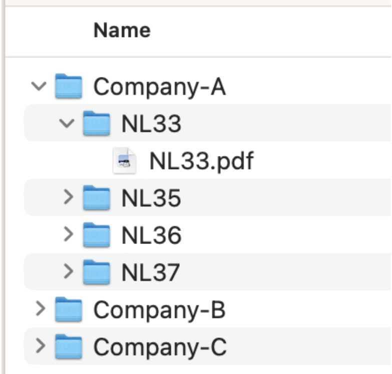
- Utilizing Python and the Camelot framework that it provides is one approach that could be taken to address the issue of the client manager having to manually enter data. Camelot is a Python library that facilitates the automated extraction of data from PDFs. As a result, the amount of time and effort required for data entry can be substantially reduced.

- The organization can extract data from PDFs and transform it into structured data that can then be easily input into Excel or other databases by utilizing the Camelot software. The need for the client manager to physically download and enter the data into Excel would be eliminated because of this, which would result in a reduction in the amount of time and effort required for data entry as well as a reduction in the risk of errors.

- Using Python and Camelot has several benefits, one of which is the reduction of wasted time, but there are also other benefits. To begin, Python is a very popular programming language, and there is a sizable community of programmers who can offer assistance and resources for those who choose to work with it. Camelot, on the other hand, is what is known as an open-source toolkit. This means that anyone is free to use it and that it can be modified to be tailored to the requirements of a particular organization.

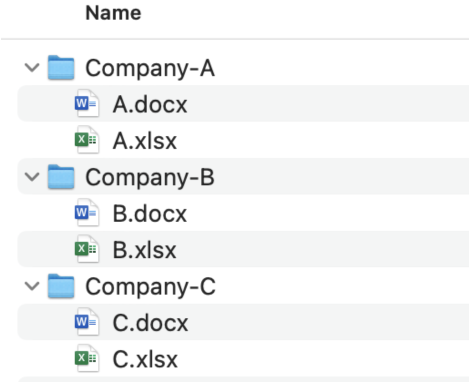
❖ Workflow.

- ### Data
  - Input Data:
    - ♦ So, to overcome the issue of large number of pdfs generated we had to create a hierarchy so how should the files be stored so that I become easy for the script to access it and create the descried output.
    - ♦ Here is a simple example of how the file-structure should look like.



```
Name
Company-A
  NL33
    NL33.pdf
  NL35
  NL36
  NL37
Company-B
Company-C
```

- Output Data:
    - So, after running the python script, we get this folder with the output file in it.



Name

- Company-A
    - A.docx
    - A.xlsx
- Company-B
    - B.docx
    - B.xlsx
- Company-C
    - C.docx
    - C.xlsx

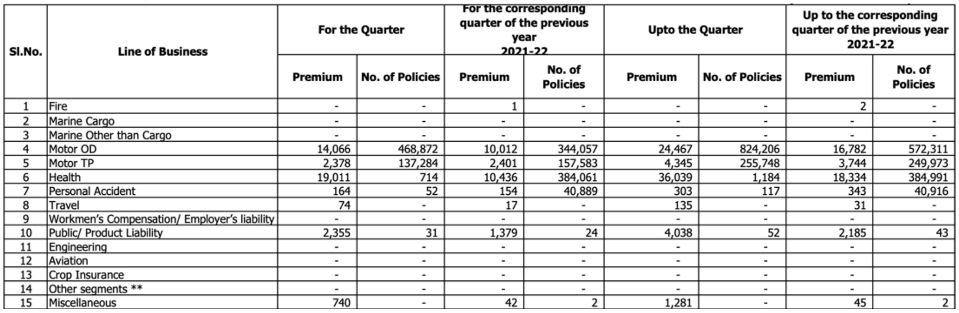
        - A excel file which has all the scrapped data.
        - A doc file which contains some quality check measure and tell us the if the file has any error.

Problems Faced

- Structural Dissimilarity:
    - The main challenge was that the structure of the PDFs varied significantly, depending on the source of the document. This meant that the same type of information was structured differently in different PDFs, making it difficult to extract the relevant data accurately.
        - To overcome the problem, I had to code the preprocessing script in such a way that it can overcome the structural dissimilarities.
- Generalization of the code:
    - The base of this solution was the next quarter's pdf should have the similar structure as the ones that were published in the previous quarter of the same year.
        - For this I wrote the script as per the pdf that were published in the Q1 of financial year 2021-2022 and then tested the code for the ones published in Q2.

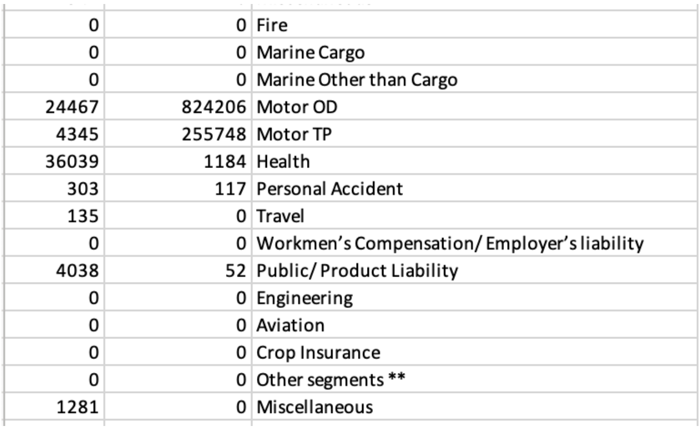
Workflow

- So, for example this is how our data pdf would look.



Sl.No.	Line of Business	For the Quarter		For the corresponding quarter of the previous year 2021-22		Upto the Quarter		Up to the corresponding quarter of the previous year 2021-22	
		Premium	No. of Policies	Premium	No. of Policies	Premium	No. of Policies	Premium	No. of Policies
1	Fire	-	-	1	-	-	-	2	-
2	Marine Cargo	-	-	-	-	-	-	-	-
3	Marine Other than Cargo	-	-	-	-	-	-	-	-
4	Motor OD	14,066	468,872	10,012	344,057	24,467	824,206	16,782	572,311
5	Motor TP	2,378	137,284	2,401	157,583	4,345	255,748	3,744	249,973
6	Health	19,011	714	10,436	384,061	36,039	1,184	18,334	384,991
7	Personal Accident	164	52	154	40,889	303	117	343	40,916
8	Travel	74	-	17	-	135	-	31	-
9	Workmen's Compensation/ Employer's liability	-	-	-	-	-	-	-	-
10	Public/ Product Liability	2,355	31	1,379	24	4,038	52	2,185	43
11	Engineering	-	-	-	-	-	-	-	-
12	Aviation	-	-	-	-	-	-	-	-
13	Crop Insurance	-	-	-	-	-	-	-	-
14	Other segments **	-	-	-	-	-	-	-	-
15	Miscellaneous	740	-	42	2	1,281	-	45	2

- Once we run the pipeline and we will get this output.



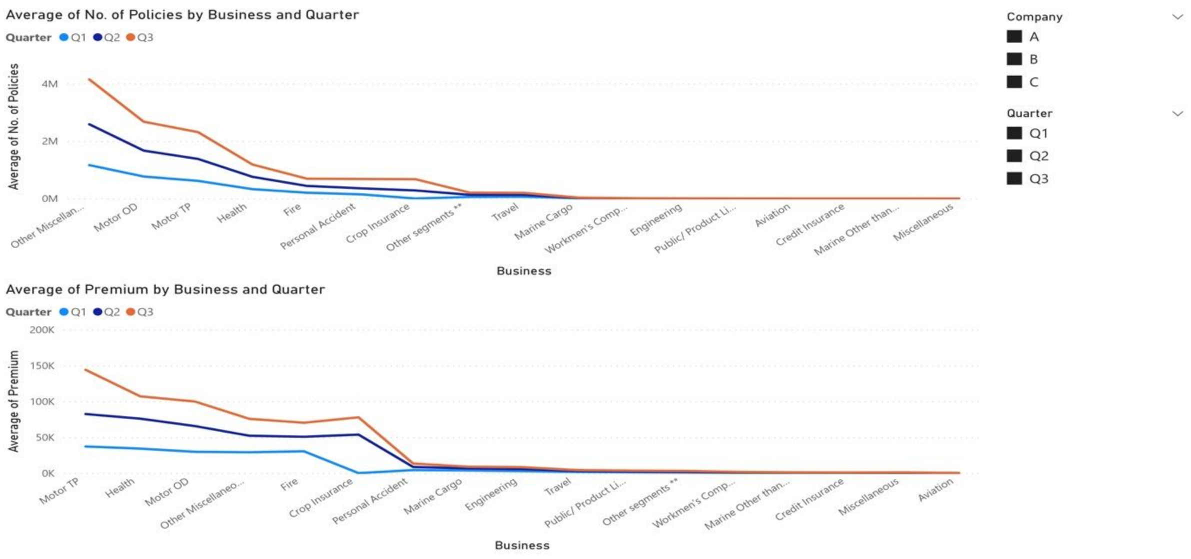
0	0	Fire
0	0	Marine Cargo
0	0	Marine Other than Cargo
24467	824206	Motor OD
4345	255748	Motor TP
36039	1184	Health
303	117	Personal Accident
135	0	Travel
0	0	Workmen's Compensation/ Employer's liability
4038	52	Public/ Product Liability
0	0	Engineering
0	0	Aviation
0	0	Crop Insurance
0	0	Other segments **
1281	0	Miscellaneous

- The code will collect the data and clean it and convert it in excel format so that we can use for visualization.

Useful Insights

- Re-insurance companies can make great use of this information to determine which of their competitors is more successful in each field by comparing it to the data presented here. The scraping and processing of the pdfs resulted in the creation of the sample data visualizations that are presented here.

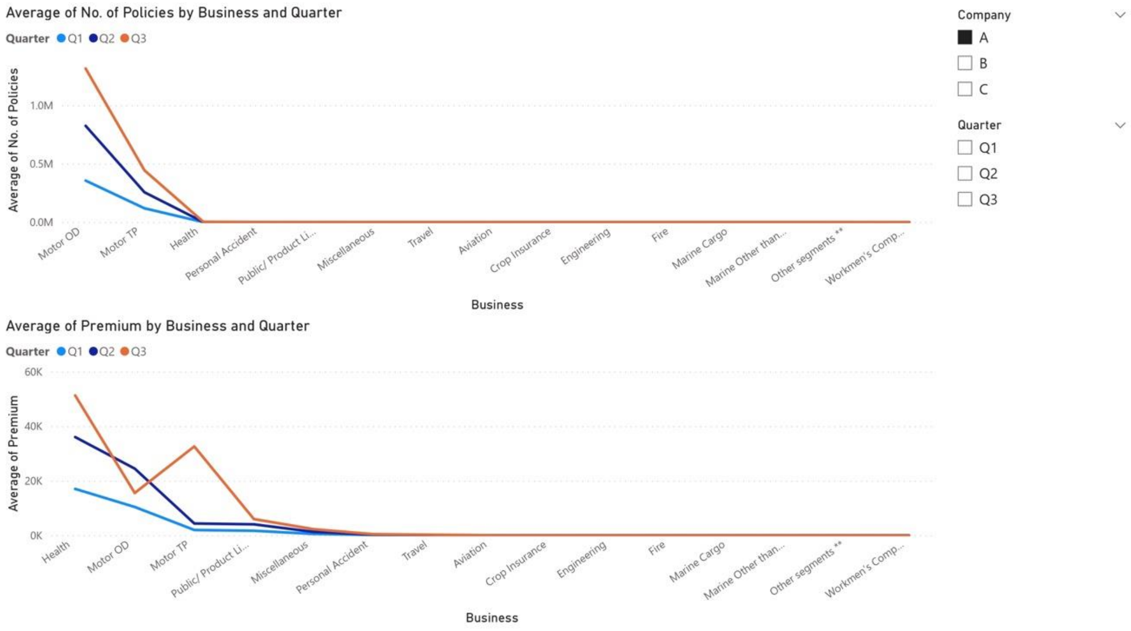
  - ♦ With the assistance of this graph, we can get a better understanding of how well all the companies have performed over the past three quarters.



Company
■ A
■ B
■ C

Quarter
■ Q1
■ Q2
■ Q3

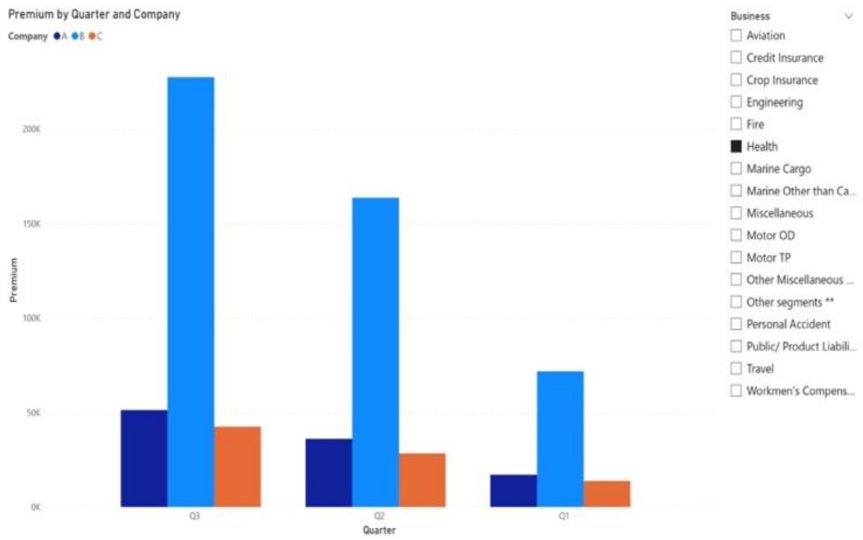
- ♦ We can even internally compare each company's quarterly performance using Power BI's features.



Company
■ A
□ B
□ C

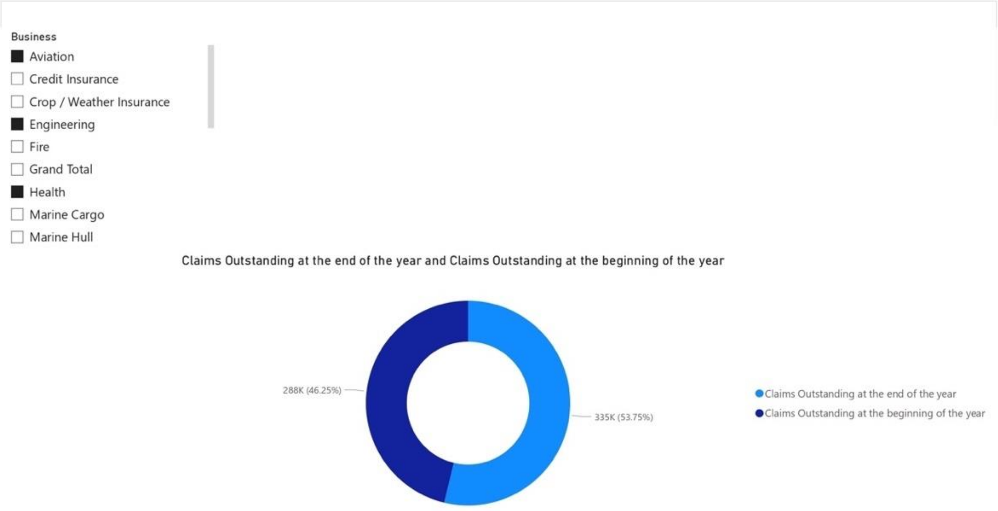
Quarter
□ Q1
□ Q2
□ Q3

SOME MORE VISUALISATION FROM THE SCRAPPED DATA



Business
□ Aviation
□ Credit Insurance
□ Crop Insurance
□ Engineering
□ Fire
■ Health
□ Marine Cargo
□ Marine Other than Ca...
□ Miscellaneous
□ Motor OD
□ Motor TP
□ Other Miscellaneous ...
□ Other segments **
□ Personal Accident
□ Public/ Product Liabili...
□ Travel
□ Workmen's Compens...

(graph showcases quarterly-business filtered premium earned)



Business
■ Aviation
□ Credit Insurance
□ Crop / Weather Insurance
■ Engineering
□ Fire
□ Grand Total
■ Health
□ Marine Cargo
□ Marine Hull

(graph showcases Proportion of Claims Out standing at the beginning and the ending of the year)

▪ Advantages and Disadvantages

ADVANTAGES	DISADVANTAGES
• Earlier it used to take 8-9 hours to manually type the data but now it just takes around 9mins via pipeline	• The Base rule for the code to run without errors is that PDF format should be same or-else we will lose some data.
• We get a doc file which has the error mentioned so we can easily navigate to the mentioned file and investigate it.	• There is a constant need of supervision as if there are any error, we will need to figure the possible outcome

❖ Future Scope.

- As we know now, we must still manually download those pdf and store all of them in a pre-defined manner, but we can even automate the process by using Scrapping API's and directly store them into the folder we want and further save time.
- As we are working on multiple companies, not all companies publish their pdf on the same day so we need to daily check on their website for the latest upload so we even created a basic telegram bot which when given input about a company send the latest quarters update so we can further investigate deploying this.