Supplier Frequency Analysis &
Exploration of the Potential of LLMs for Data Preprocessing

An Analysis of Greek Government Procurement Data through Diavgeia

International Hellenic University

Vasileios Kalfopoulos

Vasileios Kesopoulos

Evangelos Ziliachovinos

Nikolaos Laoutaris

Data Science for Business

Dr. Vassilios Peristeras

February 4, 2025

Abstract

This study explores the application of frequency analysis to identify suppliers disproportionately winning bids by making use of procurement decisions data that were collected from Diavgeia, which is a portal database that stores administrative records maintained by Greek government. Data retrieval and part of preprocessing were conducted using KNIME, while frequency analysis, as well the exploratory data analysis was implemented in Power BI. A significant portion of the dataset had to be discarded due to missing values in variables such as Tax Identification Number (TIN) and award amount, information of high importance for the frequency analysis objective. To address this challenge, we investigated the potential of Large Language Models (LLMs) for extracting TIN values from a sample of administrative decision documents that were in PDF format. The selected models—Llama 3B, Llama 8B, and Qwen-2-5-coder-7B—were implemented in KNIME and ran locally using GPT4All. These models achieved a medium performance, with the Llama 8B having 50% accuracy and the rest of the models being comparatively worse. Error analysis revealed that performance limitations come essentially from prompt size constraints and incorrect formatting of LLMs output responses, which deviated from the JSON structure that was requested as a rule in the prompt. These findings highlight the challenges and potential of leveraging LLMs for data imputation in administrative document data related to procurement decisions or in several other categories in general.

Keywords

Data Extraction · Large Language Models · Structured Data · Diavgeia Portal · Text Processing · Data Transformation · Administrative Documents · Power BI · KNIME · Document Analysis · Public Administration · Real-World Data Challenges · Data Consistency · Natural Language Processing · Unstructured Data Processing · Frequency Analysis

Table of Contents

Table of Figures

1 Introduction

In today's digital age, public administration generates a vast amount of unstructured data in the form of documents, reports and decisions. The Greek Diavgeia portal, which serves as a repository of public administrative documents, is a prime example of such a system. While it provides transparency and accessibility to citizens, the unstructured nature of the data poses significant challenges for analysis and decision making.

This assignment focuses on addressing the challenge of making use of information from the Diavgeia portal. The key objective is to preprocess and analyze procurement decisions, performing frequency analysis of the suppliers in order to identify patterns and inconsistencies. To streamline the extraction and transformation process, we utilize KNIME as the workflow automation tool, enabling efficient data preprocessing. Additionally**,** Power BI is employed to perform the analysis itself as well as visualize the extracted data, providing insights that enhance the interpretability and usability of the structured information. Together, these tools form a robust pipeline for transforming unstructured documents into actionable, data insights.

However, this process is fraught with challenges. Unstructured documents often contain inconsistent formatting, missing values, and domain specific terminology, making it difficult for traditional rule-based methods to achieve accurate and consistent results. Several specific issues arise during extraction such as critical information may be absent or incorrect. Those issues hinder automation and necessitate robust preprocessing techniques. While LLMs show promise for the task of information extraction, they come with their own set of limitations and computational requirements. These challenges highlight the complexity of structuring government data and the need for continuous methodology adaptations.

1.1 Why frequency analysis?

Frequency analysis is a fundamental methodological approach employed to uncover patterns and structures within datasets, serving as a critical step in data exploration and interpretation. By identifying dominant categories, this technique allows researchers to determine which elements are most or least prevalent, thereby establishing a foundational understanding of the data's composition. Additionally, frequency analysis facilitates the detection of outliers or anomalies, highlighting unusual or unexpected values that may indicate errors, unique cases, or areas warranting further investigation.

Beyond anomaly detection, this method provides valuable insights into the distribution of data, revealing trends, clusters, or gaps that inform subsequent analytical decisions. Furthermore, frequency analysis serves as a preparatory step for advanced analytical techniques, as the resulting frequency counts can be utilized for filtering, grouping, or segmenting data in subsequent stages of analysis. In this study, frequency analysis is employed to ensure a comprehensive understanding of the dataset, laying the groundwork for robust and informed conclusions.

1.2 Tools and Platform

1.2.1 KNIME

KNIME, which stands for Konstanz Information Miner, is an open-source data analytics, reporting, and integration platform. It provides a user-friendly, visual interface that allows users to design data workflows without the need for extensive programming knowledge. KNIME integrates various components for machine learning, data mining, and data analysis through a modular data pipelining concept. Users can drag and drop nodes to create workflows that manipulate, analyze, and visualize data, making it a versatile tool for both beginners and advanced data scientists. Its extensibility through plugins and compatibility with numerous data formats and programming languages, such as Python and R, further enhances its functionality. KNIME is widely used in industries for tasks like predictive analytics, business intelligence, and research, offering a flexible and scalable solution for data-driven decision-making.

1.2.2 Power BI

Power BI is a business analytics tool developed by Microsoft that enables users to visualize and analyze data with greater efficiency and insight. It provides a suite of tools for data visualization, reporting, and sharing insights across an organization or with external stakeholders. Power BI allows users to connect to a wide range of data sources, transform raw data into meaningful information, and create interactive dashboards and reports that can be accessed on multiple devices.

1.2.3 Diavgeia

The "Diavgeia" Program (Law 3861/2010) represents a pivotal initiative aimed at promoting transparency in government policy and administrative operations. Its primary objectives include maximizing public awareness, ensuring accountability, and fostering responsibility among public authorities. Introduced in October 2010, the program mandates the publication of all decisions and actions by governmental and administrative bodies on a centralized online platform. Particular care is taken to protect national defense interests and sensitive personal data.

Currently Diavgeia has 66.1 million document entries, submitted by 5330 organizations.

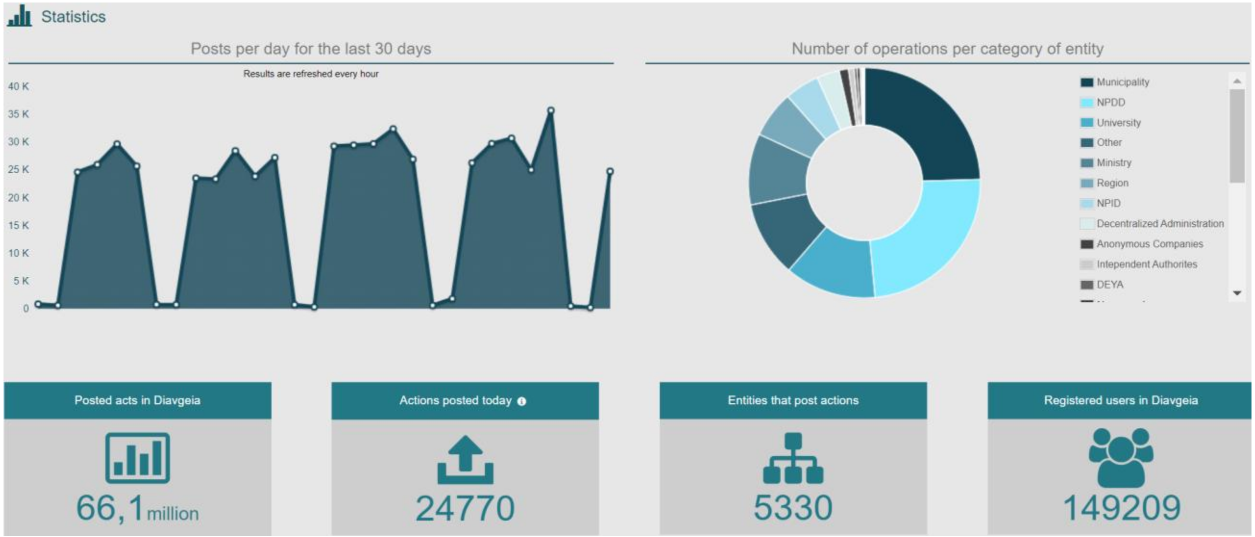
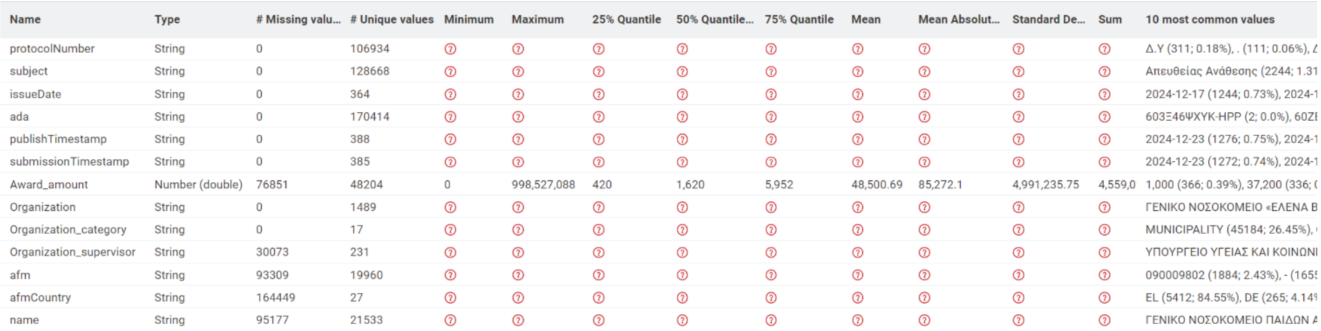


Figure 1.1: Diavgeia statistics

2 Dataset Description

1. **ADA (string):** A unique identifier or code, assigned to each record uploaded to Diavgeia. It follows an alphanumeric format, including Greek characters.
2. **AFM (string):** Refers to the Greek Tax Identification Number (TIN) of the beneficiary.
3. **Subject (string)**: The title of the decision.
4. **afmCountry (string)**: The country associated with the **AFM** (TIN).
5. **Name (string)**: The name of the beneficiary (person or organization).
6. **Organization (string)**: The entity responsible for the decision.
7. **Organization_category (string):** Categorizes the issuing organization involved. It provides a high-level classification of the entities (e.g., municipalities, ministries, companies).
8. **Organization_supervisor (string)**: The supervising authority of the organization.
9. **Award_amount (double):** Represents the monetary value of the contract.
10. **Signer_ID (string):** It refers to the individual who approved or signed the document associated with the record (e.g., the signatory or a responsible party).
11. **issueDate (date):** The date (YYYY-MM-DD) when the document was issued.
12. **PublishTimestamp (datetime):** The exact timestamp when the decision was published.
13. **submissionTimestamp (datetime)**: The exact timestamp when the decision was submitted.
14. **ProtocolNumber (string)**: A unique reference number for the decision.



Name	Type	# Missing valu...	# Unique values	Minimum	Maximum	25% Quantile	50% Quantile...	75% Quantile	Mean	Mean Absolut...	Standard De...	Sum	10 most common values
protocolNumber	String	0	106934	?	?	?	?	?	?	?	?	?	Δ.Υ (311; 0.18%), . (111; 0.06%),
subject	String	0	128668	?	?	?	?	?	?	?	?	?	Απευθείας Ανάθεσης (2244; 1.31
issueDate	String	0	364	?	?	?	?	?	?	?	?	?	2024-12-17 (1244; 0.73%), 2024-1
ada	String	0	170414	?	?	?	?	?	?	?	?	?	603Ξ46ΨΧΥΚ-ΗΡΡ (2; 0.0%), 60Ζ
publishTimestamp	String	0	388	?	?	?	?	?	?	?	?	?	2024-12-23 (1276; 0.75%), 2024-1
submissionTimestamp	String	0	385	?	?	?	?	?	?	?	?	?	2024-12-23 (1272; 0.74%), 2024-1
Award_amount	Number (double)	76851	48204	0	998,527,088	420	1,620	5,952	48,500.69	85,272.1	4,991,235.75	4,559,0	1,000 (366; 0.39%), 37,200 (336;
Organization	String	0	1489	?	?	?	?	?	?	?	?	?	ΓΕΝΙΚΟ ΝΟΣΟΚΟΜΕΙΟ «ΕΛΕΝΑ Β
Organization_category	String	0	17	?	?	?	?	?	?	?	?	?	MUNICIPALITY (45184; 26.45%),
Organization_supervisor	String	30073	231	?	?	?	?	?	?	?	?	?	ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ ΚΑΙ ΚΟΙΝΩΝΙ
afm	String	93309	19960	?	?	?	?	?	?	?	?	?	090009802 (1884; 2.43%), - (165
afmCountry	String	164449	27	?	?	?	?	?	?	?	?	?	EL (5412; 84.55%), DE (265; 4.14
name	String	95177	21533	?	?	?	?	?	?	?	?	?	ΓΕΝΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΠΑΙΔΩΝ

Figure 2.1: Dataset Statistics

3 Implementation and Methods

3.1 Tools and Technologies

Implementing our pipeline was made possible by utilizing a variety of tools and technologies. We list them below, along with their roles and the reason we chose them.

- KNIME Analytics Platform

  An open-source data analytics platform that served as our primary means of designing and executing the pipeline. KNIME's intuitive interface helped us visualize the flow of data despite our limited experience with similar projects.

- Power BI

  An interactive data visualization tool we used to create visuals for exploratory data analysis as well as presenting our final findings. Power BI's integration with KNIME allowed for a streamlined workflow.

- Large Language Models (via GPT4All)

  Integrated into KNIME, GPT4All was our connector to harnessing the power of LLMs to extract structured information from unstructured text. It provided a local, free alternative to paid services, making it ideal given our resource constraints.

- Diavgeia

  Our source of Greek government documents, accessed via API calls and additional parsing of downloaded documents.

For reference, we mention the hardware we used for this pipeline, which greatly influenced performance. Note that LLMs ran with CUDA support.

- CPU: Intel® Core™ i5-13600K (24M Cache, up to 5.10 GHz)
- RAM: 32 GB DDR5 at 6000MT/s
- GPU: NVIDIA GeForce RTX 4070 (5888 CUDA cores, up to 2.52 GHz, 12 GB GDDR6X)

3.2 Overview of the Pipeline

We will now give a high-level description of the whole process.

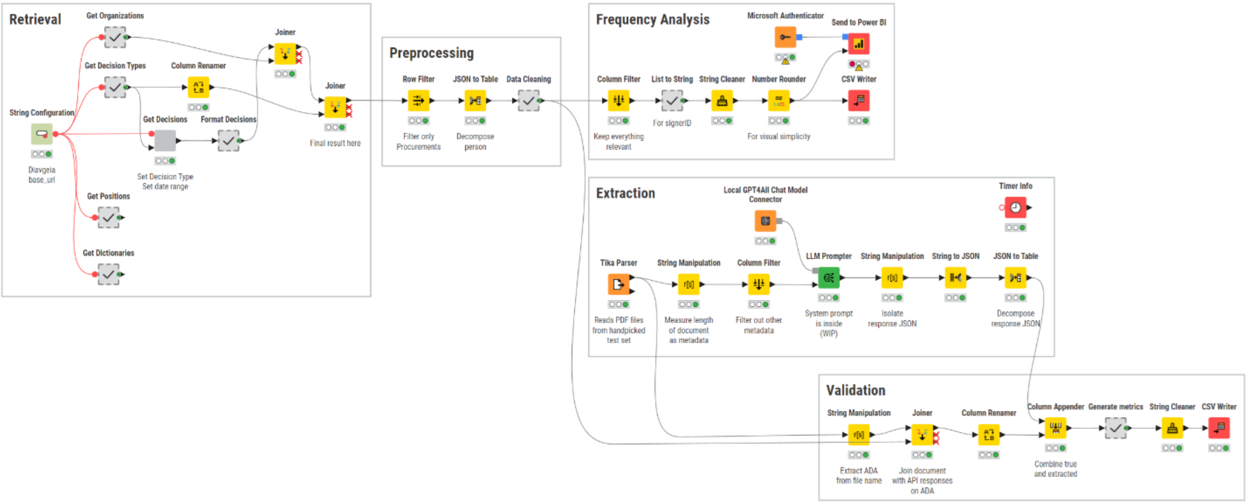


Figure 3.1: Data pipeline overview

Our workflow consists of retrieving data via API calls from Diavgeia, preprocessing it iteratively to handle missing and inconsistent entries and further focus our scope, and conducting frequency analysis on contractor counts. After identifying significant gaps and flaws in our analysis, we explored the possibility of using LLMs to extract structured information from unstructured documents to supplement or correct the API responses. For this, we constructed a validation scheme to evaluate the feasibility of this approach. Finally, we visualized our final findings using Power BI.

3.3 Data Retrieval

This part of our workflow was based on a pre-compiled set provided to us, which we expanded upon and customized for our specific project.

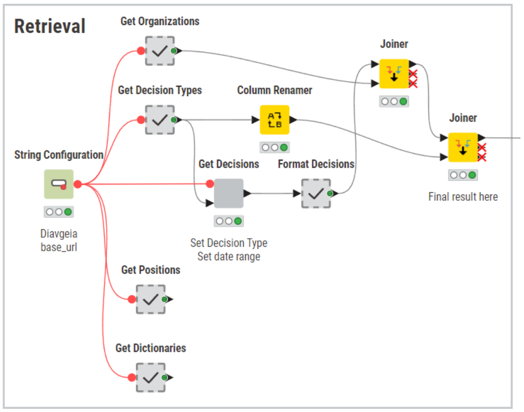


Figure 3.2: Data Retrieval workflow section

Our source of data was the Diavgeia API, which allowed us access to already structured records. Each one corresponds to a specific document and comes with extensive features and metadata, along with the link to the unstructured document itself in PDF format. We restricted our API queries to documents published in 2024, as well as a specific classification of decision types “Assignment of Projects / Procurements / Services / Studies”. We did an initial parsing and structuring of the data to properly format the responses, removing redundant features and metadata in the process.

Working with the API quickly revealed that a significant portion of the data received contained missing or incorrect entries, pertaining to features too important for our analysis. We address this issue in the following section.

3.4 Data Preprocessing

The preprocessing phase involved steps to try and improve the data’s quality and consistency, as well as further refine our scope of the analysis.

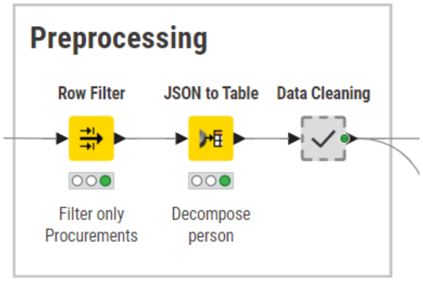


Figure 3.3: Preprocessing workflow section

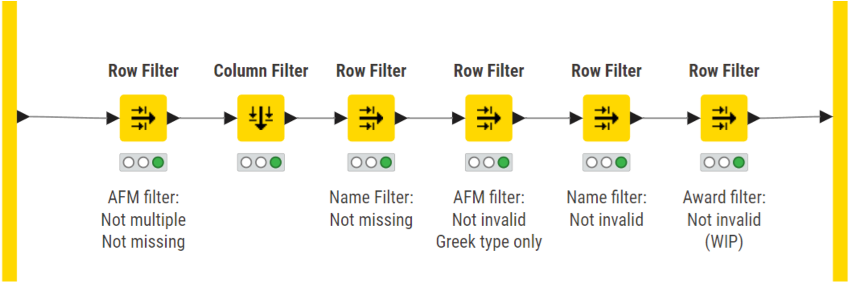
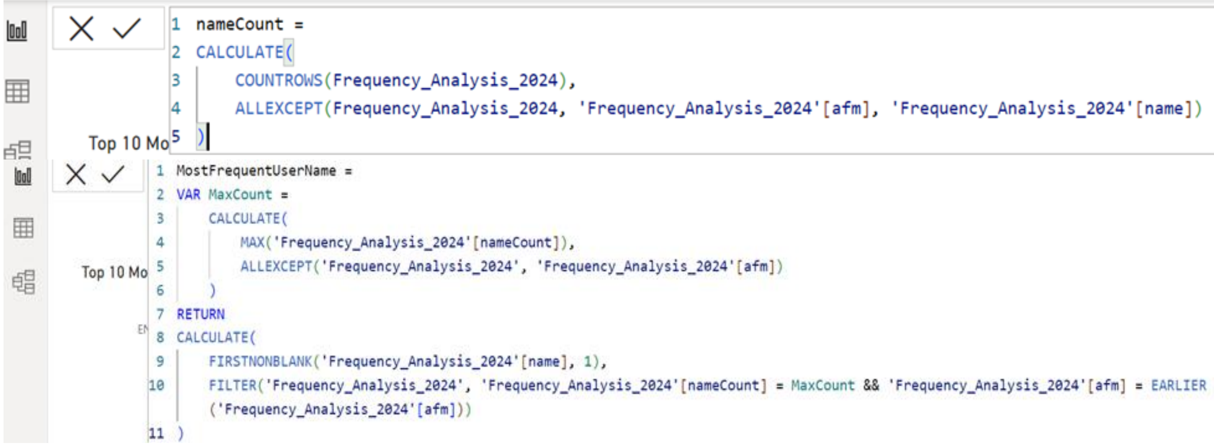


Figure 3.4: The Data Cleaning metanode

To narrow our focus, we decided to keep only Procurement documents from the extensive list retrieved from the API, to be able to compare a homogenous sample. For simplicity, we only kept Greek type TINs and documents with a single beneficiary, as these constituted the majority of cases.

With regards to preprocessing the name column which stores the supplier’s full name, we noticed how multiple names could be associated with the same TIN value due to inappropriate entries. For that reason, we decided to count the frequency of each unique string value that was linked to each supplier, and then selected the most frequent name, and finally, to assign it to all rows. To make this possible, DAX formulas had to be created. Also, in the scenario where more than one name had equal frequencies, we decided to keep the first one according to its internal order.

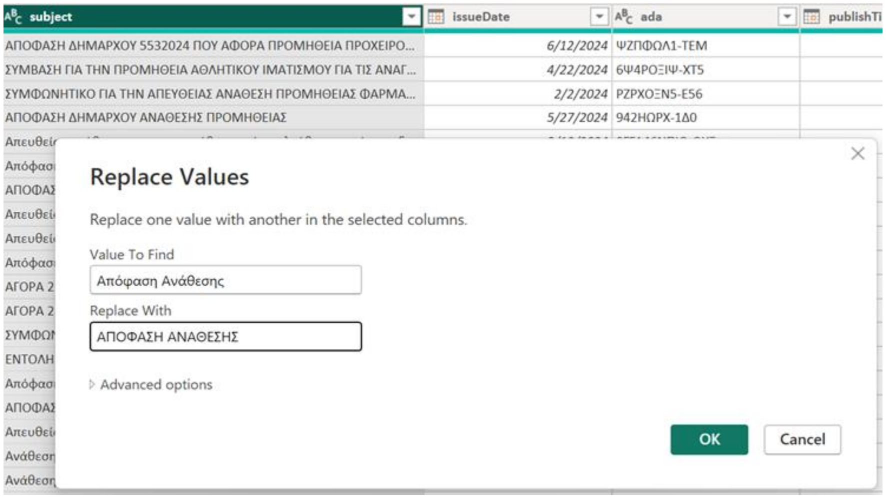


```
nameCount =
CALCULATE(
    COUNTROWS(Frequency_Analysis_2024),
    ALLEXCEPT(Frequency_Analysis_2024, 'Frequency_Analysis_2024'[afm], 'Frequency_Analysis_2024'[name])
)
```

```
MostFrequentUserName =
VAR MaxCount =
    CALCULATE(
        MAX('Frequency_Analysis_2024'[nameCount]),
        ALLEXCEPT('Frequency_Analysis_2024', 'Frequency_Analysis_2024'[afm])
    )
RETURN
CALCULATE(
    FIRSTNONBLANK('Frequency_Analysis_2024'[name], 1),
    FILTER('Frequency_Analysis_2024', 'Frequency_Analysis_2024'[nameCount] = MaxCount && 'Frequency_Analysis_2024'[afm] = EARLIER
    ('Frequency_Analysis_2024'[afm]))
)
```

Figure 3.5: Data cleaning on the “Name” key field (DAX formulas)

Furthermore, we realized through inspecting the most common values for the subject column, that many of them were stored in lower cases and many in upper case. So, we used the Power Query Editor to transform these values and implemented this method also on other columns as well. This preprocessing part was important for ensuring data consistency.



subject	issueDate	ada	publishTi
ΑΠΟΦΑΣΗ ΔΗΜΑΡΧΟΥ 5532024 ΠΟΥ ΑΦΟΡΑ ΠΡΟΜΗΘΕΙΑ ΠΡΟΧΕΙΡΟ...	6/12/2024	ΨΖΠΦΩΛ1-ΤΕΜ	
ΣΥΜΒΑΣΗ ΓΙΑ ΤΗΝ ΠΡΟΜΗΘΕΙΑ ΑΘΛΗΤΙΚΟΥ ΙΜΑΤΙΣΜΟΥ ΓΙΑ ΤΙΣ ΑΝΑΓ...	4/22/2024	6Ψ4ΡΟΞΙΨ-ΧΤ5	
ΣΥΜΦΩΝΗΤΙΚΟ ΓΙΑ ΤΗΝ ΑΠΕΥΘΕΙΑΣ ΑΝΑΘΕΣΗ ΠΡΟΜΗΘΕΙΑΣ ΦΑΡΜΑ...	2/2/2024	ΡΖΡΧΟΞΝ5-Ε56	
ΑΠΟΦΑΣΗ ΔΗΜΑΡΧΟΥ ΑΝΑΘΕΣΗΣ ΠΡΟΜΗΘΕΙΑΣ	5/27/2024	942ΗΩΡΧ-1Δ0	
Απευθεί			
Απόφασ			
ΑΠΟΦΑΣ			
Απευθεί			
Απευθεί			
Απόφασ			
ΑΓΟΡΑ 2			
ΑΓΟΡΑ 2			
ΣΥΜΦΩΝ			
ΕΝΤΟΛΗ			
Απόφασ			
ΑΠΟΦΑΣ			
Απευθεί			
Ανάθεσ			
Ανάθεσ			

Replace Values

Replace one value with another in the selected columns.

Value To Find: Απόφαση Ανάθεσης

Replace With: ΑΠΟΦΑΣΗ ΑΝΑΘΕΣΗΣ

Figure 3.6: Replacing lower-case values to upper-case values for data consistency

For missing values, we were forced to drop entries that lacked a TIN and name, after our efforts of using other available data points were unsuccessful. A visual inspection of tabular data easily revealed some invalid entries, which we also dropped.

It is important to note that this was an iterative process. After visualizing some preliminary results, we were able to perform exploratory data analysis which revealed incorrectly entered data, especially award amounts. This prompted us to revisit our preprocessing strategy to further improve upon it by applying new filters to correct as many of those issues as possible.

3.5 Frequency Analysis

The frequency analysis was performed by calculating the number of appearances of each TIN across the cleaned data, to highlight any potential disproportionalities.

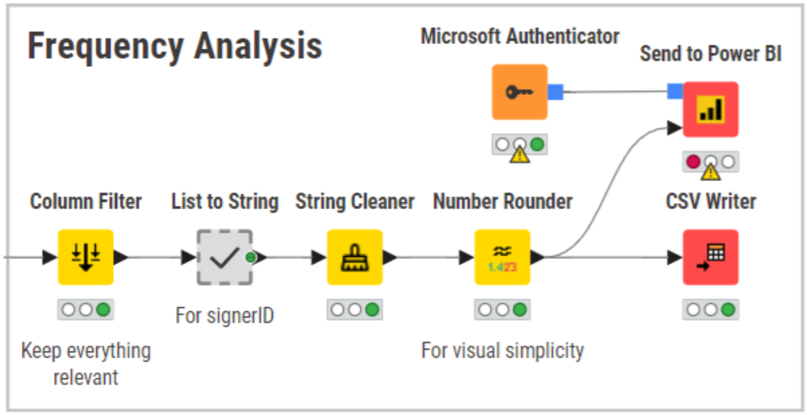
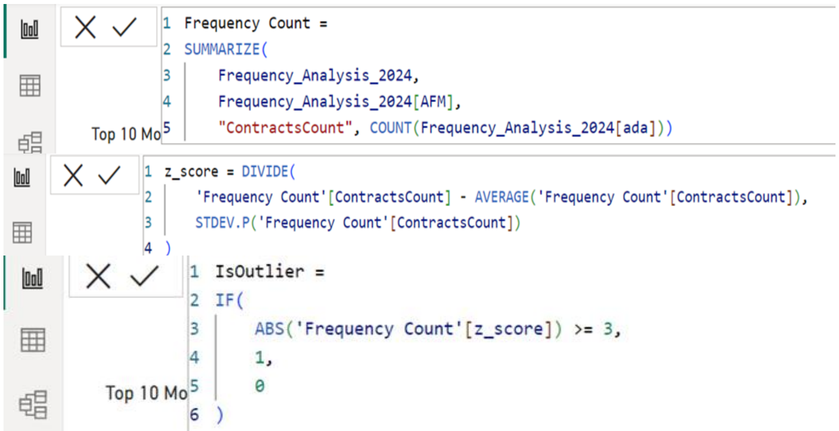


Figure 3.7: Frequency analysis workflow section

In KNIME, we focused on preparing the cleaned dataset for export for subsequent analysis in Power BI. At this stage, we chose to keep as many features as possible to be sure that no potentially valuable data was excluded prematurely, preserving the full richness of the dataset. This approach allowed us to decide, after seeing the graphs and performing further analysis in Power BI, which data to keep or discard based on its value of visualization and data integrity.

In Power BI, in order to spot the outliers for the task of performing frequency analysis, we created a separate table and we stored the TIN(AFM) column, which uniquely identifies a supplier, a frequency column(Frequency Count), a z-score column and finally, a numeric column named IsOutlier that flags every supplier(IsOutlier value to be equal to 1) where its calculated z-score is equal or greater than 3. The rest of the suppliers were considered regular ones and the assigned IsOutlier value were set as 0.



```
Frequency Count =
SUMMARIZE(
    Frequency_Analysis_2024,
    Frequency_Analysis_2024[AFM],
    "ContractsCount", COUNT(Frequency_Analysis_2024[ada]))
```

```
z_score = DIVIDE(
    'Frequency Count'[ContractsCount] - AVERAGE('Frequency Count'[ContractsCount]),
    STDEV.P('Frequency Count'[ContractsCount])
)
```

```
IsOutlier =
IF(
    ABS('Frequency Count'[z_score]) >= 3,
    1,
    0
)
```

Figure 3.8: Spotting the highly-frequent suppliers (DAX formulas)

As far as the exploratory data analysis step, several visualizations were created in order to discover to some extent what were the conditions of the winning bids with regards to the anomalously frequent suppliers, and thus to provide some basic insights. This analysis was made possible through the Power BI application which enabled the creation of clustered bar charts, histograms, box plots and many more.

At this point, the frequency analysis was complete. However, we had to acknowledge that it was based on only a part of the whole dataset, due to the necessary harsh removal and filtering. More importantly, we lacked confidence in the integrity of results, as our intuition suggested that some inconsistencies that escaped our detection were detrimental in shaping them. In our attempts to solve this problem, we explored the use of LLMs, as detailed in the next section.

3.6 LLM Implementation for Information Extraction

As our frequency analysis relied heavily on TINs as a primary component, we focused our extraction on this feature.

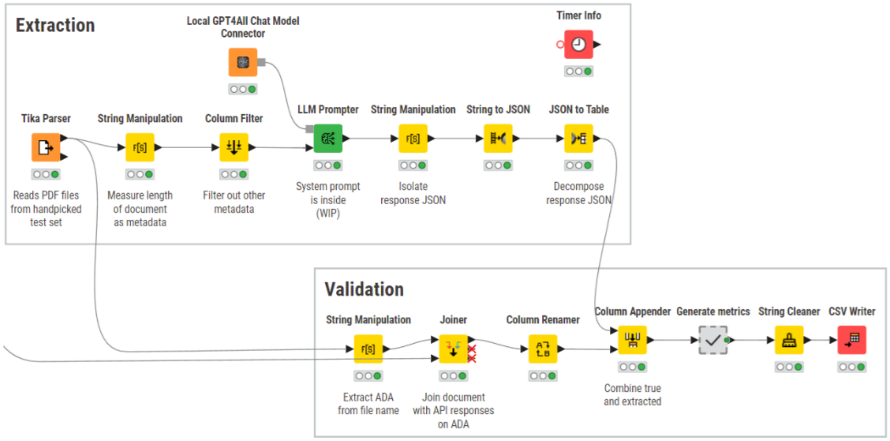


Figure 3.9: LLM Information extraction workflow section

Our idea was to use LLMs for imputing missing TINs. However, we were unsure of the feasibility of this strategy, so we decided to first test its effectiveness by constructing a "validation test" based on Machine Learning principles. For this controlled experiment we hand-picked a sample of 30 documents from our original dataset, ensuring that each contained TINs and names previously retrieved from the API, which we manually verified by reading the documents ourselves. To maximize sample diversity to the best of our ability, we tried to include documents of varying lengths and file sizes, with and without tabular data, featuring different fonts, layouts, and even embedded images. The LLM was then tasked with extracting the TIN and name from each document. The goal was to evaluate whether the LLM would accurately extract the same information from the documents, thus validating its potential for data imputation.

We chose GPT4All as the model connector and tried various popular models to compare their performance. The documents were parsed into a string, which was fed into the model with a system prompt we refined through several trial-and-error iterations:

```
You are a document analyzer for Greek Government procurements. Your task is to extract
information about the supplier from the provided text documents.
*Input:
The input will be a plain text string representing the content of a procurement document.
*Output:
The output should be a JSON representation of the gathered information and must only contain
the following:
{
"afm": The supplier's TIN (ΑΦΜ) as a 9-digit string.
"name": The full name of the supplier as a string.
}
*Data Extraction:
Look for keywords or phrases like "ΑΦΜ", "Αριθμός Φορολογικού Μητρώου", "Επωνυμία", "Ονομα
Εταιρείας" to locate relevant information.
Consider variations in formatting (e.g., different separators, capitalization).
*Data Validation:
Ensure the extracted AFM is a valid 9-digit number.
*Error Handling:
If the AFM or name cannot be extracted reliably or if the document does not contain any supplier
information, return empty values in the JSON format.
```

Figure 3.10: LLM system prompt

We evaluated the results using a set of metrics: whether the LLM extracted the correct or incorrect TIN, if the prompt exceeded the model's context window, and if the response was in an incorrect format. The outcomes of this validation as well as the reasons why our experiment did not go beyond this stage are discussed in the following two chapters.

4 Results

In this section, we will visualize the findings on the frequency analysis objective, which refers to the total contracts each supplier won, as well as the performance of the LLM models that were utilized for extracting information from the administrative documents, and finally, we will also provide some interpretations.

4.1 Frequency Analysis

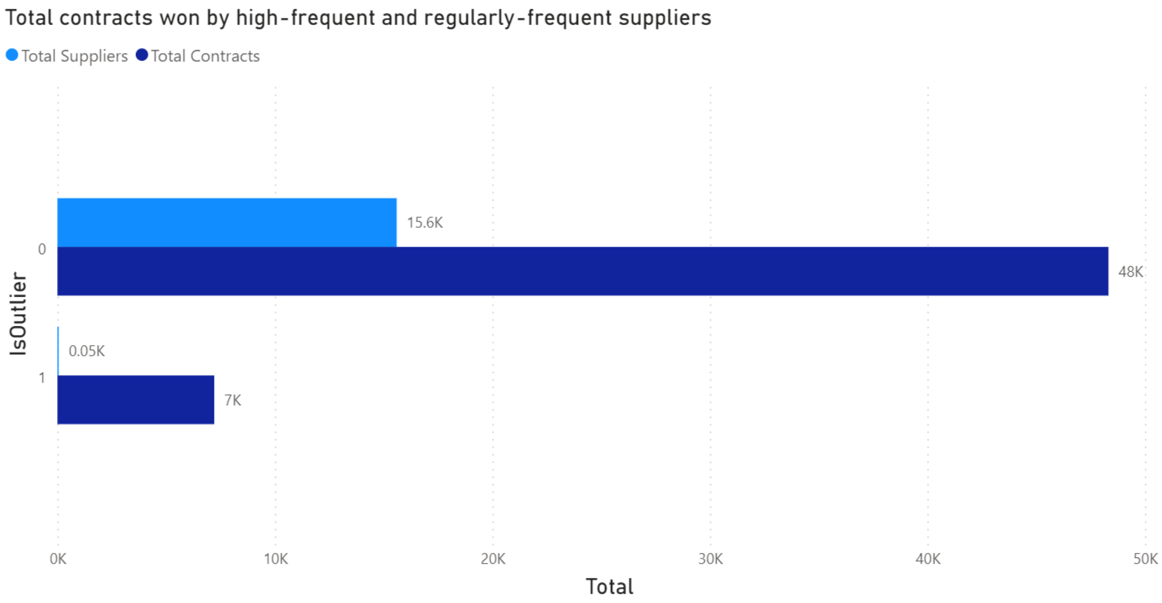


Figure 4.1: Total contract won for frequent and non-frequent suppliers

The dataset consisted of 55.529K administrative documents, where 11.5% of the total awarded contracts are associated with the anomalous suppliers. Also, the contract winning suppliers were approximately 15K in total while the outlier suppliers, regarding the contract winning frequency, were just 50 in number winning more than 7K contracts in total.

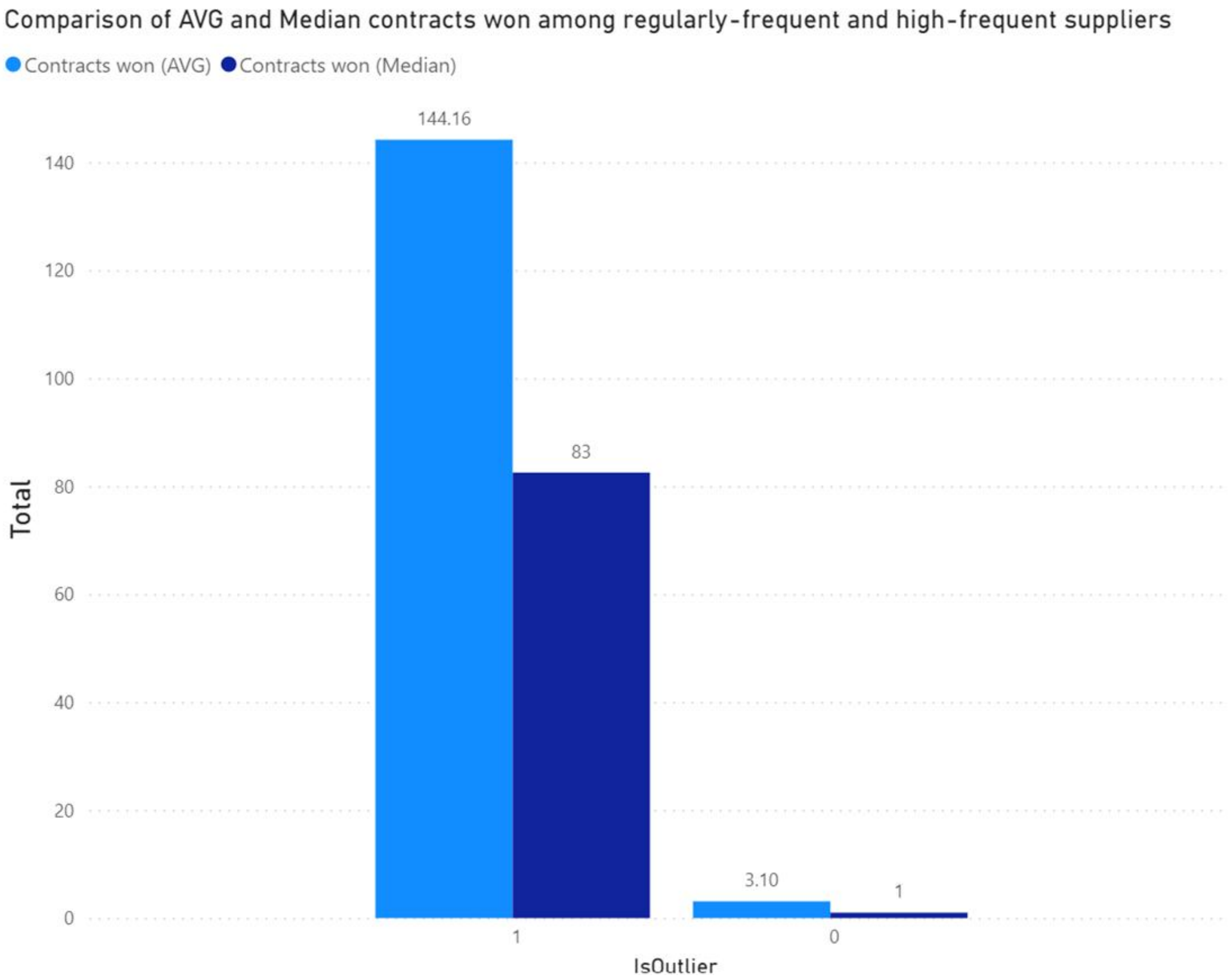


Figure 4.2: Comparison of the AVG and Median Frequency among regular and high frequent suppliers

In figure 4-2, we compare the mean and median values between the regular suppliers and the outliers. We observe that most suppliers won just only 1 contract on average and the middle value is approximately 3 contracts. On the other hand, the anomalously frequent suppliers won 144 on average and a median value of 83, which is far greater compared to the usual suppliers.

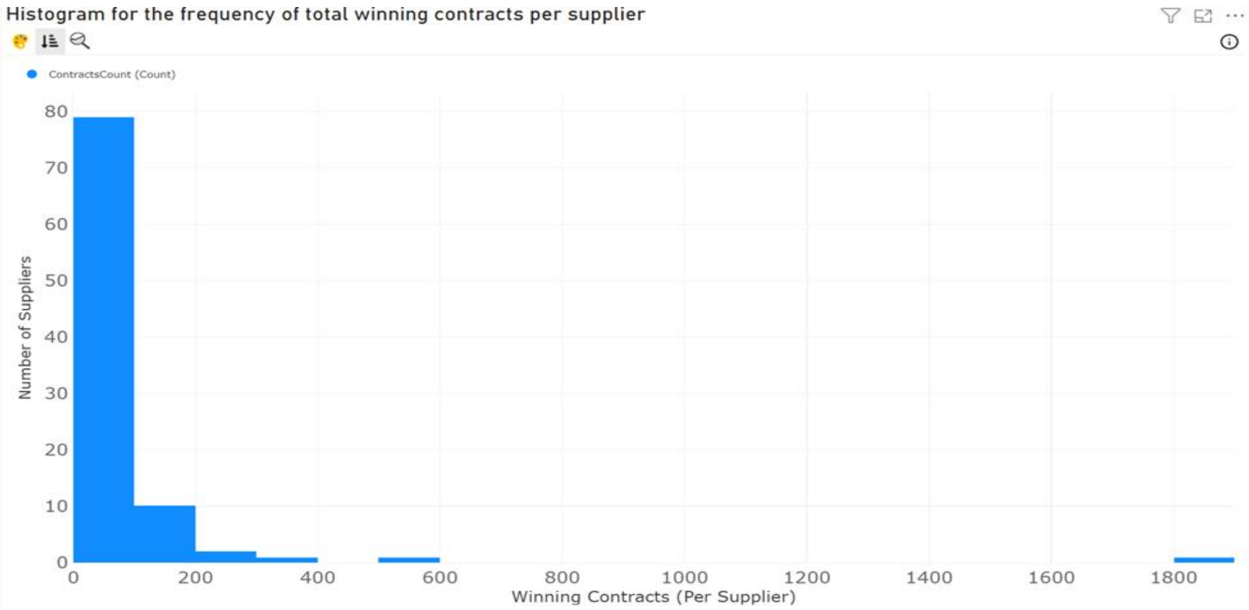


Figure 4.3: Histogram for the frequency of total contracts won per suppliers

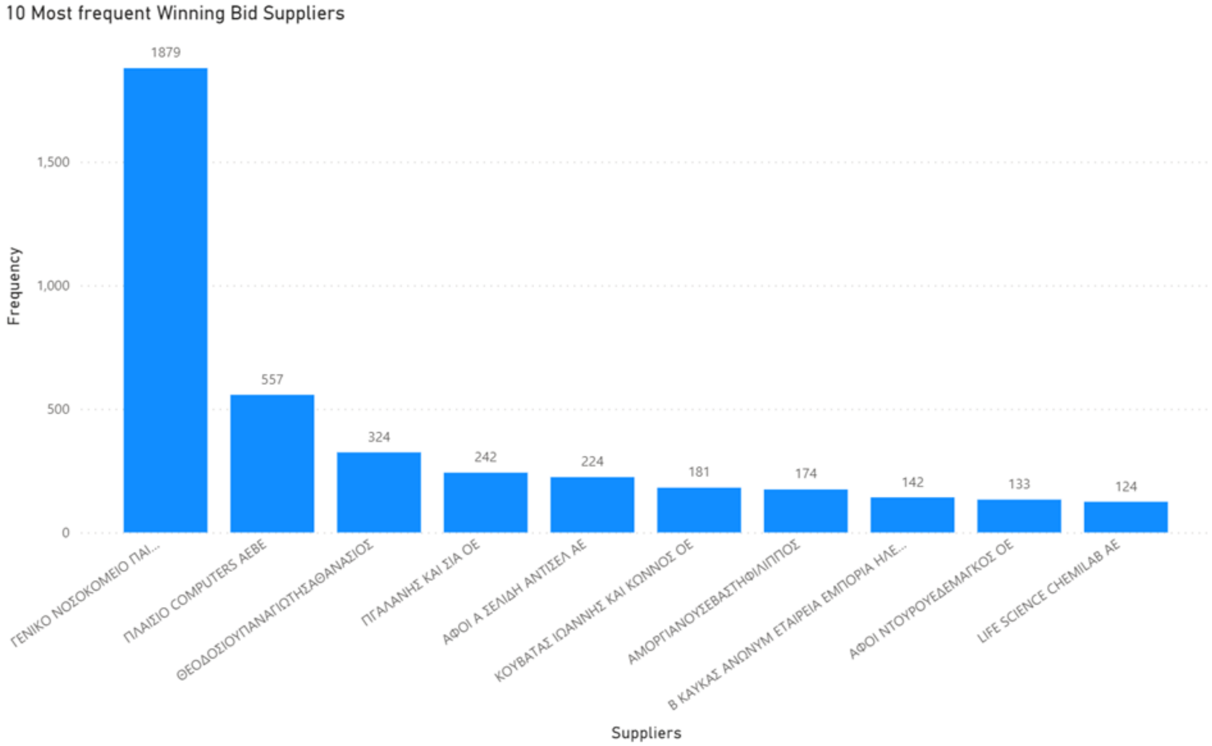


Figure 4.4: 10 most frequent suppliers grouped by name

In figure 4-4, we visualize the 10 highest in frequency suppliers grouped by their names. The distribution is highly skewed to the right and we observe that the top 1 supplier has three times higher frequency from the 2nd in order.

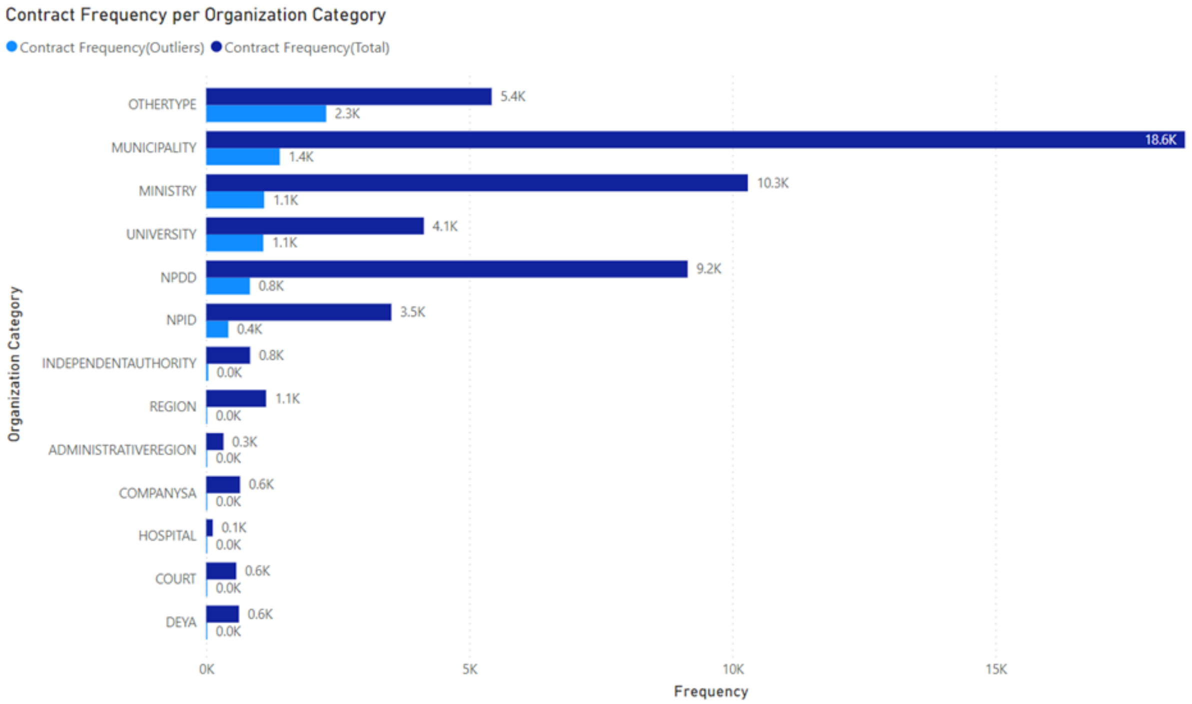


Figure 4.5: Contract frequency per organization category

In figure 4-5 we observe how the frequency of contracts won are distributed per organization category, for regular and for high-frequent suppliers. In general, the distribution among the comparison groups is right-skewed. Municipality is the most popular organization category for the usual suppliers, but it is not obvious which one is for the outliers since the assigned category value (OTHERTYPE) is not informative. However, the 2nd most popular category that is associated with the outliers seems to be the municipalities.

4.2 LLM evaluation

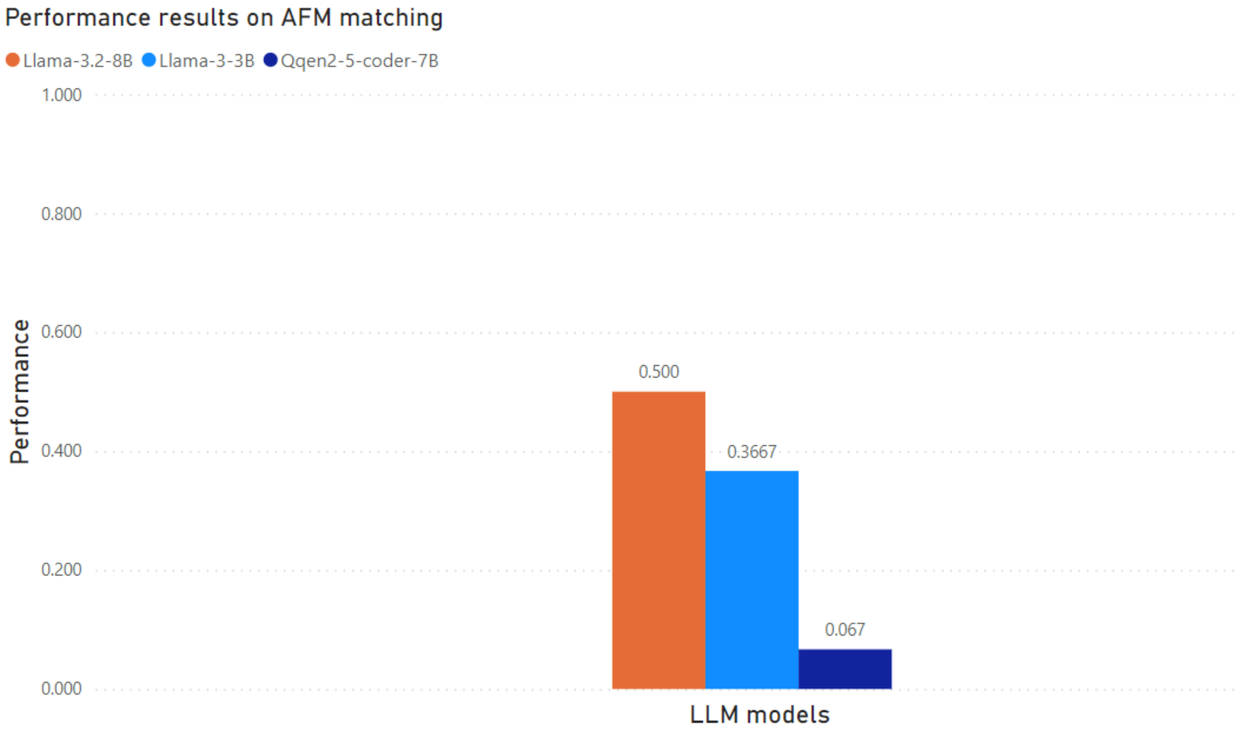


Figure 4.6 Performance results AFM matching

In figure 4-6, we present the performance of our selected models in their ability to correctly extract the AFM(TIN-Tax Identification Number) characteristic that is displayed in the selected pdf documents dataset and this characteristic specifically relates to the supplier tax number. The Llama-3.2 8B LLM model had an 50% accuracy on the task, with the Llama-3-3B having approximately 37% accuracy and Qquen-2-5-coder-7B having extremely poor performance (just 7% accuracy). In the following figure (figure 4-7) we will provide some insights to justify the poor accuracy performance of those models.

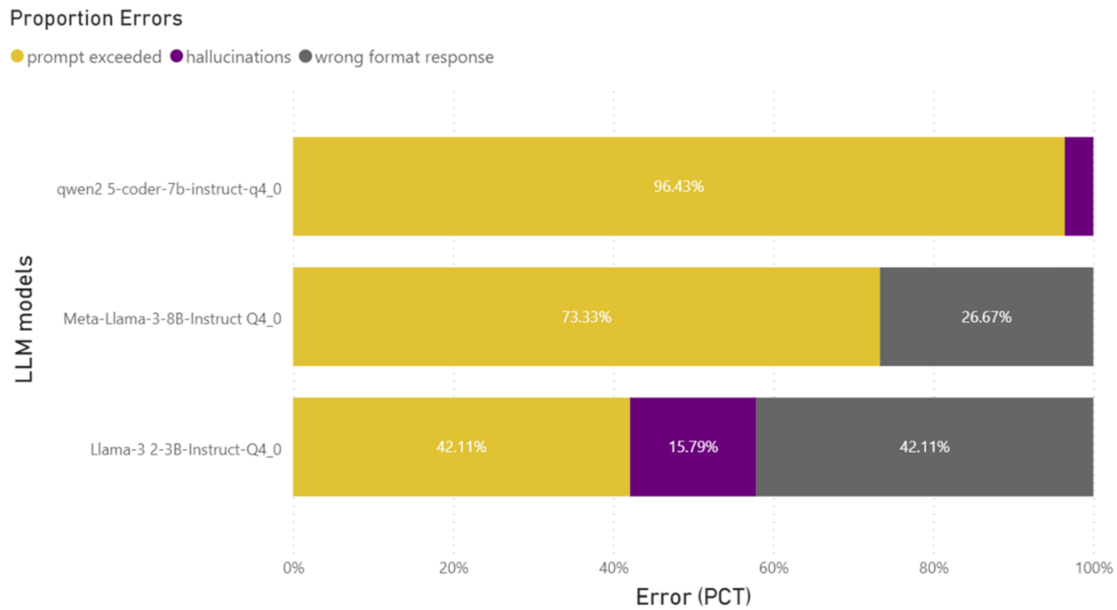


Figure 4.7: LLM proportion errors due to prompt exceeded, hallucinations, and wrong format responses

As we observe on the bar charts in figure 4-7, high proportions of errors were caused due to the limitation in the prompt size window and due to the wrong formatting of the LLM responses (not JSON). In the left clustered bar chart, we see that almost all extraction failures that occurred on the Qquen2-5-coder-7B were due to its limitation on the prompt size window. This type of error was very frequent also for the rest of the tested LLM models. Moreover, a high proportion of the total error on the Llama models, such as 42% and 27%, occurred due to the wrong formatting of the responses.

5 Discussion

5.1 Key Takeaways

Our work highlighted several key challenges and insights regarding taking advantage of government documents.

Missing and incorrect data were confirmed to be a major concern, particularly with categorical values. Traditional methods of imputation are not always applicable and adaptable approaches are necessary.

Frequency analysis is an effective way to identify patterns and anomalies, but its effectiveness is hindered by lack of data completeness and integrity. The lack of confidence in the dataset meant that even well-structured analytical methods risked producing misleading results.

Preprocessing proved to be an iterative necessity as new challenges emerged throughout. The need for additional cleaning and transforming of data at a later time is in the nature of most real datasets. We experienced firsthand that using a validation scheme before applying an idea at scale is the right approach. This prevents the risk of blindly integrating an unverified method into the pipeline, which can make things worse.

LLMs demonstrated significant potential for extracting structured information. However, model performance is indeed constrained by infrastructure limitations. Also, being inherently probabilistic, LLMs can generate errors with high confidence (“hallucinations”) so the human-in-the-loop fact/inference check cannot be bypassed. The highlight is that LLM adaptation must be carefully designed, validated, and supplemented with quality control measures.

Finally, human oversight remains essential. Even with powerful AI and automation tools, human verification is often crucial for resolving ambiguities and ensuring data integrity.

5.2 Challenges and Adaptations

Throughout the implementation of our pipeline, we encountered several challenges that made us shift our perspective and apply iterative refinements and adaptations to our methodology. A significant pivot was made when we shifted focus away from monetary analysis due to incomplete and incorrect data, which we were unsure we could reliably pinpoint and correct using known techniques.

Another challenge we encountered working with the Diavgeia API was the slow response time, requiring hours to retrieve a whole year’s worth of data.

These setbacks notwithstanding, the iterative nature of our work allowed us to refine our approach and focus on the aspects that were most feasible, such as validating the model's effectiveness for smaller datasets and addressing the data quality issues as best as possible within the given constraints.

5.3 Limitations

Despite our best efforts, fixing all inconsistencies in our dataset proved impossible. There was simply no viable way of imputing categorical features like TINs and names using the available data. A particularly problematic issue arose with reward amounts, where we identified a large number of API responses lacking a decimal point - potentially inflating values by a factor of up to 100.

Multiple attempts were made to detect a pattern and correct these issues, like using rule-based logic, code, RegEx, statistical analysis, visualization, and the use of AI, but we concluded that none provided a sufficient level of confidence for reliable correction. As a result, we proceeded with the frequency analysis using far fewer features and entries than we had hoped.

The insights gained from tackling these limitations were detrimental in shaping our decision to use LLMs, especially regarding the potential for imputing missing TINs. Ideally, our goal was to use LLMs to address the gaps identified in the frequency analysis. However, the results of our test case revealed that local models struggled. Context window size was often not enough when processing the documents, which were often large. Moreover, given our home infrastructure's resource limitations, we found that even small-scale processing took considerable time, and the scalability of our approach became a concern. The processing time and costs needed for tens of thousands of documents - enough to cover a full year of procurement data - was simply too long to be practical. Considering these limitations, we decided not to proceed with the full imputation of missing TINs. Consequently, we simply concluded our experiment with the validation of the idea.

6 Conclusion

6.1 Summary of the Study

This study investigated the application of frequency analysis to identify suppliers disproportionately winning bids in Greek government procurement, for data analytics purposes. Using data from the Diavgeia portal, we aimed to reveal potential irregularities in procurement decisions. Data retrieval and initial preprocessing were performed using KNIME, while frequency and exploratory data analysis were conducted in Power BI. A significant challenge arose from missing and incorrect data, particularly Tax Identification Numbers (TINs) and award amounts, which were crucial for our analysis. There were no reliable data imputing methods to be used, so, to address this challenge, we explored the use of Large Language Models (LLMs) for extracting TINs from PDF documents containing administrative decisions. Three models, Llama 3B, Llama 8B, and Qwen-2-5-coder-7B, were implemented locally in KNIME using GPT4All. While the Llama 8B model achieved 50% accuracy in TIN extraction, the other models performed comparatively worse. Error analysis indicated that performance limitations come primarily from prompt size constraints and secondly, from inconsistencies in the LLM output format, which deviated from the specified JSON structure. Moreover, we concluded that hardware limitations need to be considered, due to the high computing power the LLM models require which consequently affects the execution time as well. The findings highlight both the challenges and the potential of using LLMs for data imputation in administrative documents, specifically within the context of procurement decisions, but also potentially applicable to other categories as well.

6.2 Possible Research Direction

We feel that our work demonstrated the potential of LLMs for extracting structured data from unstructured government procurement documents. Future research could address our challenges through more powerful computational resources, improved methodologies, and novel approaches to automation.

The scalability limitations we encountered stemmed largely from the constraints of local hardware. Transitioning to cloud-based solutions could enable batch processing of thousands of documents at speeds that would make real-world application feasible. Furthermore, we could integrate paid APIs from advanced LLMs, such as GPT-4o, Copilot, DeepSeek-V3 and others.

Beyond infrastructure, the methodology itself could be expanded in promising ways. One potential future direction is the validation of API-retrieved data across the entire corpus of Diavgeia documents through LLM-powered document parsing. This process could incorporate a human-in-the-loop approach, allowing experts to manually review cases where the AI expresses uncertainty. The result would be an amazing level of data completeness, which was never before possible.

This result would in turn ensure the highest possible accuracy of tasks like frequency analysis, but why stop there? Frankly, this work could extend far beyond, into more sophisticated forms of data-driven transparency. It would become possible to explore trends in procurement practices, detect anomalies, and even apply predictive modeling techniques to forecast spending behaviors. These applications would demonstrate the broader impact of improving government data accessibility through AI, while of course supporting both research and policy making in new and exciting ways.

Finally, a rather highly ambitious but potentially transformative direction is the development of our own domain-specific LLM, trained explicitly on Greek government documents. Such a model could, in theory, overshadow the currently most impressive general-purpose models with its contextual understanding and specialized knowledge of Diavgeia's structure, terminology, and inconsistencies. If successful, such a system could serve as an "intelligent" intermediary between raw government documents and structured analytical workflows, enabling researchers and policymakers to extract insights with unprecedented precision.

We would like to believe that our work could lay the foundation for future research that could revolutionize the way government data is utilized.

7 Resources

7.1 Acknowledgements

We would like to express our gratitude to the lab and teaching staff of this class for their support.

7.2 Tools and Software

- KNIME Analytics Platform (Version 5.4.0)
- Power BI (Version)
- GPT4All (Version 3.8.0)
- Diavgeia [https://diavgeia.gov.gr/]

7.3 References

- Serderidis, K., Konstantinidis, I., Bassiliades, N., Meditskos, G., & Peristeras, V. (2023). D2KG: An Integrated Ontology for Knowledge Graph-based Representation of Government Decisions and Acts. Semantic Web Journal. Submitted 09/02/2023. Available at: https://www.semantic-web-journal.net/content/d2kg-integrated-ontology-knowledge-graph-based-representation-government-decisions-and-1

7.4 Additional Resources

- Diavgeia API Documentation [https://diavgeia.gov.gr/api/help]
- KNIME Documentation [https://docs.knime.com/]
- Power BI Documentation [https://powerbidocs.com/]