MSc Data Science

Coursework
Data Science for Business

Georgios Diamantis

Christos Galanis

Ermioni Traka

October 2023
Thessaloniki – Greece

Abstract

This paper conducts a Geographic Analysis of Government Spending in the Greek public sector, with a specific focus on the allocation of funds and contracts related to the Assignment of Works / Supplies / Services / Studies ("ΑΝΑΘΕΣΗ ΕΡΓΩΝ / ΠΡΟΜΗΘΕΙΩΝ / ΥΠΗΡΕΣΙΩΝ / ΜΕΛΕΤΩΝ") within the hospital domain during the pivotal years of 2020-2021 and 2021-2022, marked by the influence of the COVID-19 pandemic. The study encompasses a rigorous data validation and cleansing process, complemented by Geospatial Analysis, yielding significant insights beneficial for researchers, decision-makers, companies, and government departments. It is imperative to consider the limitations outlined in this paper to grasp a comprehensive understanding of the findings. Notably, a major portion of the expenses was allocated to the administrative regions of Athens and Larissa. However, it is crucial to highlight the inherent challenges posed by the poor quality of metadata, obscuring the overall transparency of actual expenses and their distribution across the Hellenic country.

Table of Contents

Abstract....2
Introduction....4
Objective....4
Dataset Preprocessing....4
  Data Validation....10
  Data Cleansing....12
Final Dataset....13
GeoSpatial Analysis and Results....16
Conclusion....23
Limitations....23
Future work....23
Appendix....24

Introduction

In an age dominated by digital governance, ensuring transparency and accountability within the public sector has become increasingly vital. A noteworthy stride in this direction is the Diavgeia platform, a dedicated website by the Greek government serving as a comprehensive repository for laws and actions executed across diverse public sector departments. This platform stands as a testament to the commitment to transparency, offering the public invaluable insights into government expenditures and activities. This assignment undertakes a focused exploration—the Geographic Analysis of Government Spending in the Greek public sector, with a specific emphasis on the allocation of funds and contracts related to the Assignment of Works, Supplies, Services, and Studies ("ΑΝΑΘΕΣΗ ΕΡΓΩΝ / ΠΡΟΜΗΘΕΙΩΝ / ΥΠΗΡΕΣΙΩΝ / ΜΕΛΕΤΩΝ") within the hospital domain.

Objective

The goal of this study is to evaluate the geographical distribution of government funds and contracts allocated to hospitals in Greece during the two-year period spanning 2020-2021 and 2021-2022, influenced by the backdrop of the COVID-19 pandemic. Utilizing metadata sourced from Diavgeia, the study aims to chart the geographic concentration of financial resources, offering insights into regional variations, anomalies, and trends in resource allocation within the healthcare sector. This analysis goes beyond merely quantifying the number of funds and delves into the cumulative sum, providing valuable perspectives on how government funding for hospitals is distributed across diverse regions of Greece throughout the specified biennial periods. Ultimately, this research endeavors to enhance transparency and facilitate informed decision-making in the domain of public sector resource allocation.

Dataset Preprocessing

The data were taken from Diavgeia which is a Greek public platform storing information about laws and actions from various public sectors. More specifically, Diavgeia's API was used in order to get the metadata needed. Documentation about the API can be found here https://diavgeia.gov.gr/api/help.

Knime was the primary tool for accessing, gathering and preprocessing the data. This is "a complete platform for end-to-end data science, from creating analytic models, to deploying them and sharing insights within the organization, through to data apps and services" (https://www.knime.com/software-overview).

Knime Workflow

This section provides a comprehensive overview of the KNIME workflow, specifically designed to address the primary focus of this paper — the allocation of government funds and contracts within the hospital sector pertaining to the Assignment of Works / Supplies / Services / Studies ("ΑΝΑΘΕΣΗ ΕΡΓΩΝ / ΠΡΟΜΗΘΕΙΩΝ / ΥΠΗΡΕΣΙΩΝ / ΜΕΛΕΤΩΝ").

To attain the desired data, the diavgeia_methods.knwf file served as the foundation. This is a file given by the professor and his team containing a workflow on Knime platform. Modifications were applied to this file to extract the information needed and they will be described thoroughly later.

An overall view of the workflow can be seen on the figure below:

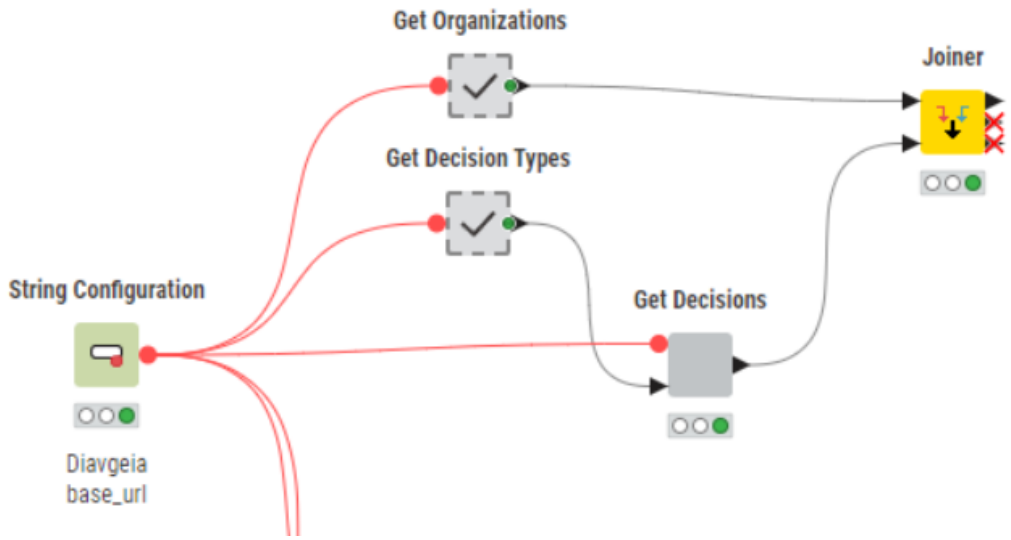


Figure 1: Overall KNIME workflow.

Node Settings Overview:

Get Organizations Node

An important node of the workflow is the so-called Get Organizations which includes the process of acquiring information about the hospitals.

In general, this workflow creates a connection between Knime and Diavgeia's API and after careful manipulations information about organizations is taken. Organizations can be hospitals, schools etc and the supervisors of them can vary from ministries, town halls, and so on.

In our case, in order to get data about all hospitals in Greece two additional Row Filters were applied in Get Organizations node:

- The first Row Filter node, leveraging regular expressions, is applied to filter hospitals independently of their supervisors.
- The second Row Filter node is utilized to exclude the Chest Disease Research Institute, categorized as an institute rather than a hospital.

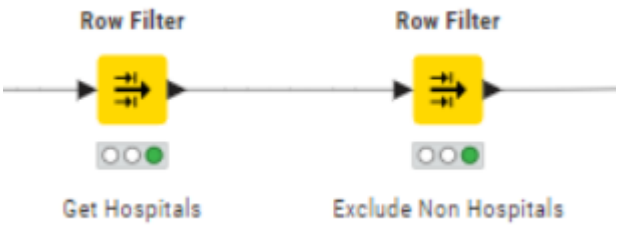


Figure 2: Additional Row Filter Nodes (Get Organizations Node)

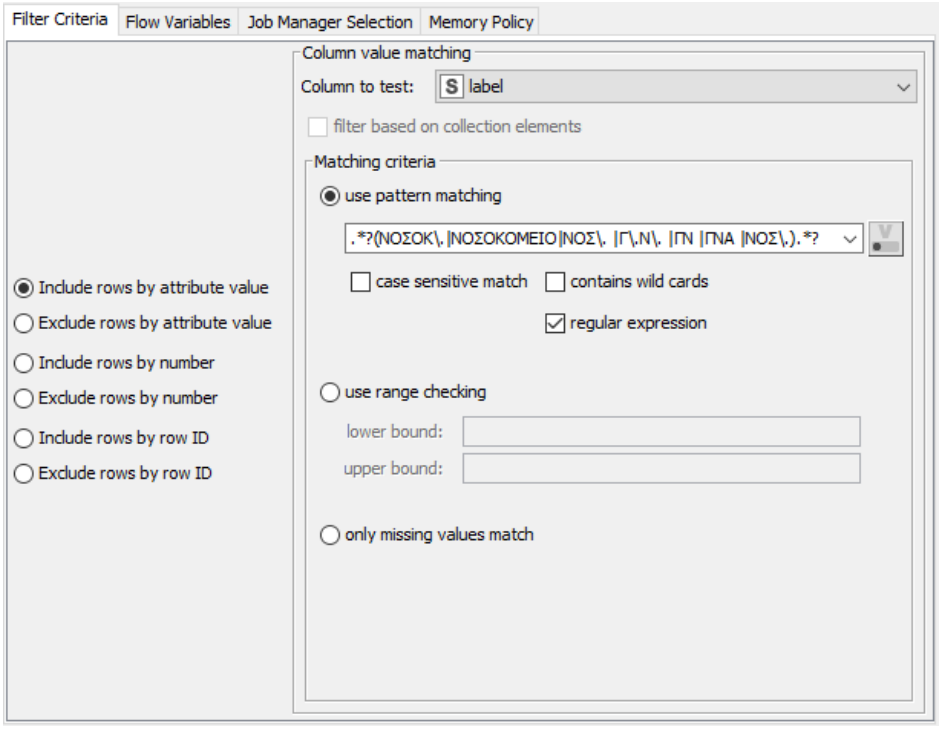


Figure 3: Hospitals Row Filter node configuration.

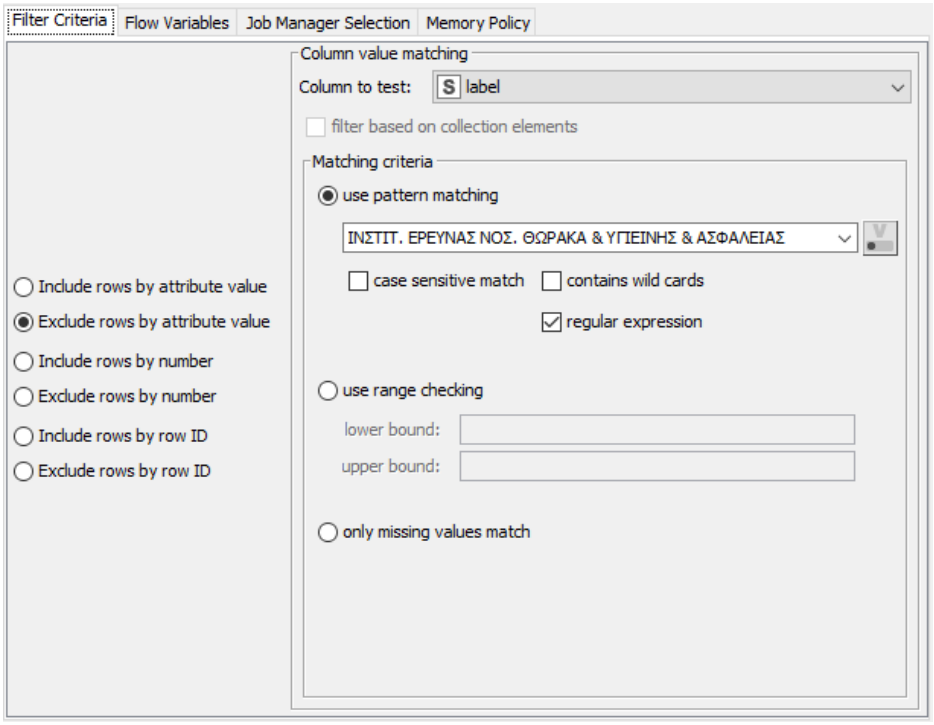
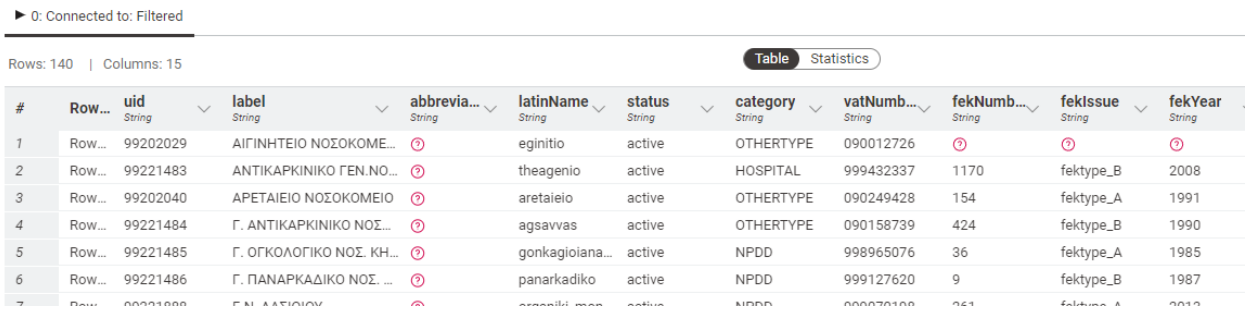


Figure 4: Exclusion of Non-Hospitals Row Filter node configuration.

The visual representation below showcases the results of our procedural efforts. A standout achievement is securing the "uid" column, a unique identifier assigned to each hospital. This identifier serves as a critical reference point for subsequent analysis. More specifically, the “uid” of each hospital will be mapped with every Assignment of Works / Supplies / Services / Studies within the two years duration.

Furthermore, the hospital labels were successfully retrieved, crucial for the upcoming task of mapping hospital positions in Greece geographical region.

Importantly, the dataset reveals a noteworthy statistic – a total of 140 hospitals are dispersed across Greece, each carrying its distinct identity and contributing to the intricate web of government funds and contracts within the healthcare sector.



▶ 0: Connected to: Filtered

Rows: 140 | Columns: 15 — Table | Statistics

#	Row...	uid String	label String	abbrevia... String	latinName String	status String	category String	vatNumb... String	fekNumb... String	fekIssue String	fekYear String
1	Row...	99202029	ΑΙΓΙΝΗΤΕΙΟ ΝΟΣΟΚΟΜΕ...		eginitio	active	OTHERTYPE	090012726			
2	Row...	99221483	ΑΝΤΙΚΑΡΚΙΝΙΚΟ ΓΕΝ.ΝΟ...		theagenio	active	HOSPITAL	999432337	1170	fektype_B	2008
3	Row...	99202040	ΑΡΕΤΑΙΕΙΟ ΝΟΣΟΚΟΜΕΙΟ		aretaieio	active	OTHERTYPE	090249428	154	fektype_A	1991
4	Row...	99221484	Γ. ΑΝΤΙΚΑΡΚΙΝΙΚΟ ΝΟΣ...		agsavvas	active	OTHERTYPE	090158739	424	fektype_B	1990
5	Row...	99221485	Γ. ΟΓΚΟΛΟΓΙΚΟ ΝΟΣ. ΚΗ...		gonkagioiana...	active	NPDD	998965076	36	fektype_A	1985
6	Row...	99221486	Γ. ΠΑΝΑΡΚΑΔΙΚΟ ΝΟΣ. ...		panarkadiko	active	NPDD	999127620	9	fektype_B	1987

Figure 5: Organizations table - 140 Hospitals in total

Get Decision Types and Get Decisions nodes

The Get Decision Types and Get Decisions nodes focus on getting only the Assignment of Works / Supplies / Services / Studies from a variety of other Decision Types within the public sector. The first gets the variety of decision types that exist on the API while the second returns data regarding each decision type in a time interval specified by the user.

In this paper, the goal was to focus on a dual-period analysis—2020-2021 and 2021-2022—which holds strategic importance, particularly as these years coincide with the onset of the COVID-19 pandemic.

To expedite data retrieval, we leverage the organization id sector, utilizing the unique identifier (uid) from the organization table. This pragmatic approach not only accelerates information extraction but aligns seamlessly with the overarching goal of comprehending the geographical distribution of financial allocations within the specified sector over the designated time frames.

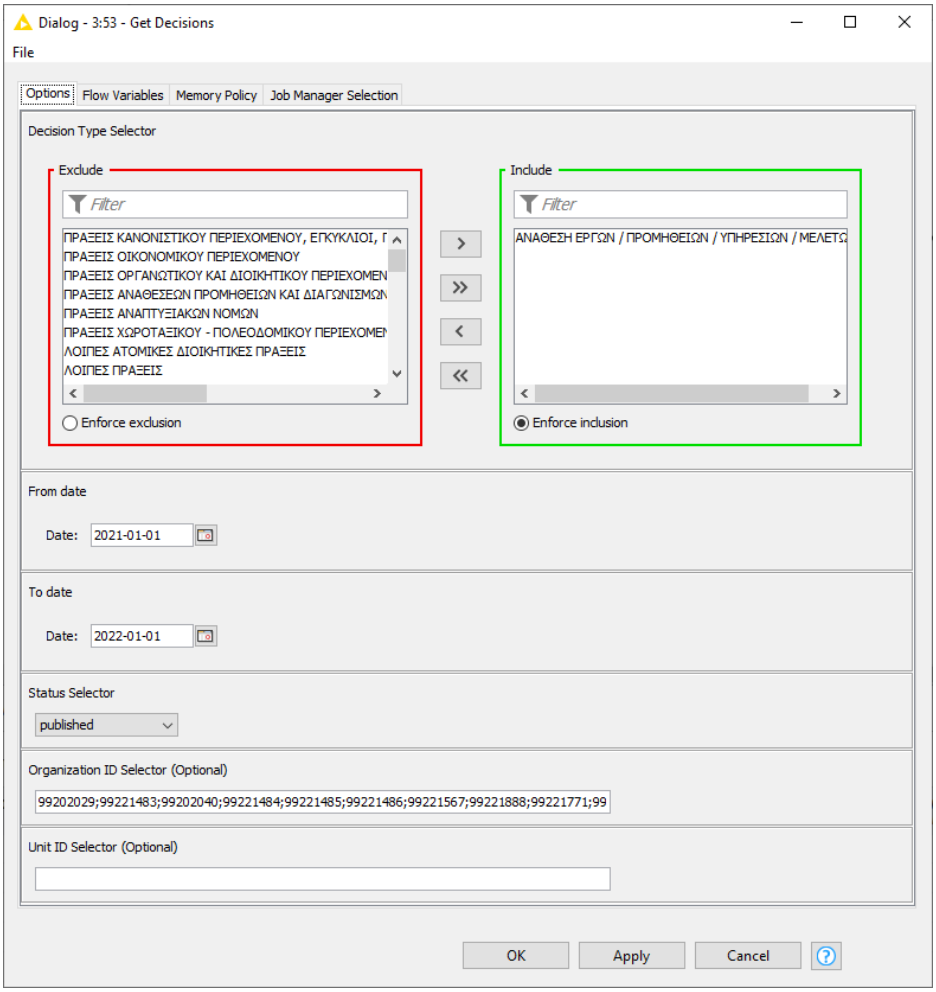


Figure 6: Get Decisions Node Configuration

Joiner Node

The concluding step in our workflow involves the utilization of the Joiner node, acting as the linchpin to map the hospitals with the output for the specified decisions for the time interval specified earlier. The result is a comprehensive table that serves as a consolidated reference point, offering a holistic view of the decision types alongside the respective organizations, providing a clear and structured understanding of the allocation of funds within the healthcare sector.

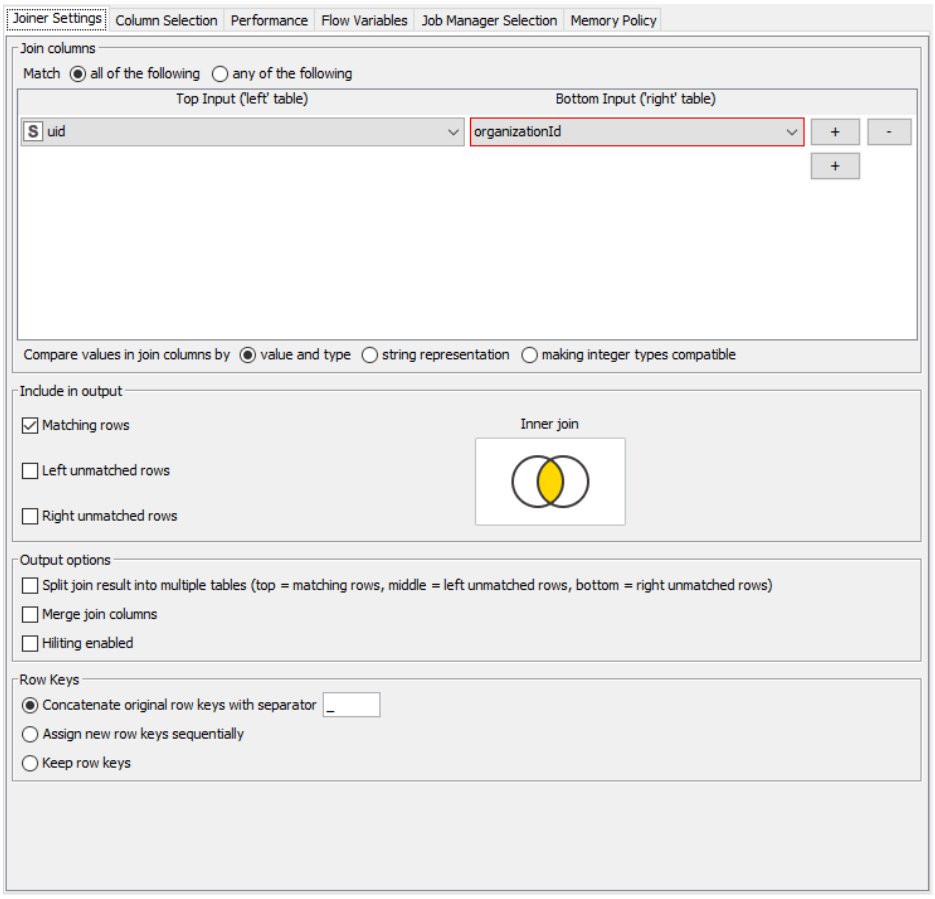


Figure 7: Joiner Node – configuration

Combining all the information

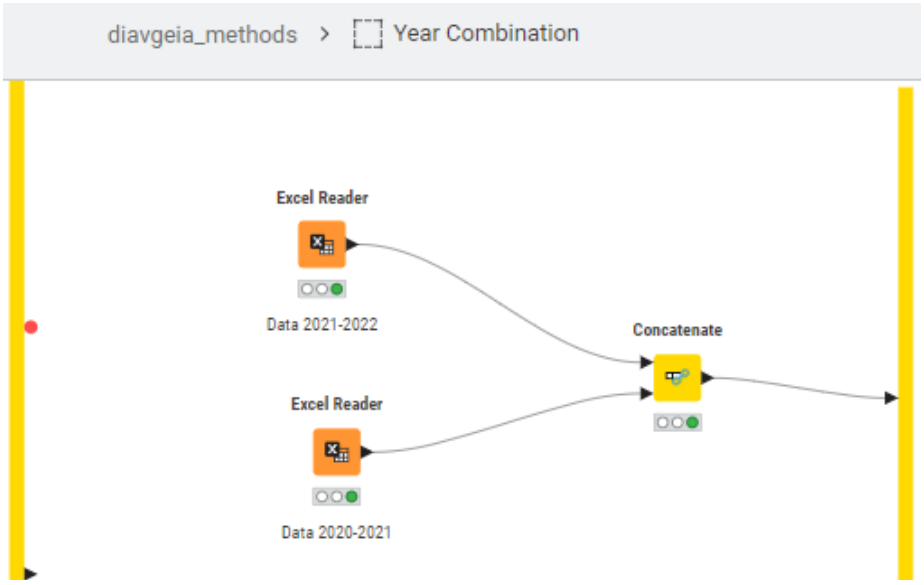


Figure 8: Combining information for the final dataset.

The integration process involves merging data from two Excel files derived from earlier steps, specifically the output of the "Joiner of Get Decisions" node and the "Get Organizations" node. This pivotal step consolidates information, combining the outcomes of decision-related data and organizational details into a unified

dataset. By bringing together these datasets, we create a comprehensive Excel file that serves as a consolidated and enriched source for subsequent analyses. This integrated dataset lays the groundwork for a more thorough exploration of government spending patterns within the specified time frame and organizational context.

To provide a clearer insight into the outcomes of the integration process, here's a concise overview detailing the various characteristics of the dataset:

- "uid": Unique identifier for each hospital.
- "label": Descriptive label for each hospital.
- "supervisorId": Unique identifier for each hospital supervisor.
- "supervisorLabel": Description of each hospital supervisor's role or entity (e.g., Ministry of Health).
- "subject": Topic or nature of each decision.
- "issueDate": Date when the decision was uploaded to the system.
- "awardAmount": Monetary value allocated for each decision.
- "person": Contains the receiver's AFM number and the company name.

Data Validation

This part focuses on assessing the quality of the extracted metadata, particularly gauging completeness and uncovering any disparities between the metadata and the information embedded within the PDF documents. Through a random selection of 10 documents, we scrutinize characteristics of the metadata such as DocumentUrl, label, subject, issueDate, person, awardAmount, and Supervisor since this information plays significant role in the purpose of this paper.

A thorough analysis took place and color-coded indicators have been employed to facilitate a visual understanding of the assessment results. The color scheme is described below:

- Orange: Indicates instances of false information. Entries marked in orange may contain inaccuracies or discrepancies between the extracted metadata and the actual content of the PDF documents.
- Green: Signifies correct information. Entries marked in green align accurately with the content found in the PDF documents, reflecting the reliability of the extracted metadata.
- Purple: Indicates information that is not clear or potentially misleading. Entries marked in purple may require further clarification or investigation, as they pose ambiguity or uncertainty regarding their accuracy or relevance.

Data Validation							
Samples	DocumentUrl	label	subject	issueDate	person	awardAmount	Supervisor
1	https://diavgei.	ΓΕΝΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΜΕΣΣΗΝΙΑΣ-	ΣΥΜΒΑΣΗ ΜΕ ΤΗΝ ΕΤΑΙΡΕ	2023-10-12	[ ]	{ "amount" : 5797, "currency" : "EUR"}	ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ
2	https://diavgei.	ΓΕΝΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΜΕΣΣΗΝΙΑΣ-	ΥΠΗΡΕΣΙΕΣ ΔΙΕΚΠΕΡΑΙΩΣΗ	2023-10-11	[ ]	{ "amount" : 5000, "currency" : "EUR"}	ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ
3	https://diavgei.	ΝΟΜ.ΓΕΝ. ΝΟΣΟΚΟΜΕΙΟ ΛΑΜΙΑΣ	Λήψη απόφασης για την	2023-10-24	[ ]	{ "amount" : 3965.52, "currency" : "EUR"	ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ
4	https://diavgei.	ΝΟΜ.ΓΕΝ. ΝΟΣΟΚΟΜΕΙΟ ΣΠΑΡΤΗΣ	ΠΡΟΜΗΘΕΙΑ ΥΛΙΚΩΝ	2023-10-23	[{ "afm" : "095029193", "name" : "ΠΑΠΑΠ	{ "amount" : 2872.8, "currency" : "EUR"}	ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ
5	https://diavgei.	ΠΕΡ.ΓΕΝ. ΝΟΣΟΚΟΜΕΙΟ ΠΑΙΔΩΝ 'Α	ΕΝΤΟΛΗ ΠΡΟΜΗΘΕΙΑΣ ΓΙ	2023-10-19	[{ "afm" : "090009802", "name" : "ΓΕΝΙΚΟ	{ "amount" : 992, "currency" : "EUR"}	ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ
6	https://diavgei.	ΓΕΝΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΘΕΣΣΑΛΟΝΙΙ	Αντικατάσταση λόγω διά	2023-10-11	[{ "afm" : "801347007", "name" : "PRO BC	{ "amount" : 4450, "currency" : "EUR"}	ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ
7	https://diavgei.	Γ. ΑΝΤΙΚΑΡΚΙΝΙΚΟ ΝΟΣΟΚΟΜΕΙΟ«	ΕΓΚΡΙΣΗ ΠΙΣΤΩΣΗΣ 6.200,	2023-10-11	[{ "afm" : "997808779", "name" : "ΣΑΒΒΑΣ	{ "amount" : 6200, "currency" : "EUR"}	ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ
8	https://diavgei.	ΘΕΡΑΠΕΥΤΙΚΟ ΚΑΤΑΣΤΗΜΑ ΚΡΑΤΗΣ	Προμήθεια γραφικής ύλι	23/10/2023	[ ]	{ "amount" : 1800, "currency" : "EUR"}	ΥΠΟΥΡΓΕΙΟ ΔΙΚΑΙΟ
9	https://diavgei.	ΑΡΕΤΑΙΕΙΟ ΝΟΣΟΚΟΜΕΙΟ	Περί έγκρισης της απευθ	2023-10-13	[{ "afm" : "997982917", "name" : "LIFE SC	{ "currency" : "EUR"}	ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑ
10	https://diavgei.	ΕΙΔΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΟΦΘΑΛΜΙΑΤ	Έγκριση : α) του υπ΄αρ 4	2023-10-05	[ ]	{ "amount" : 4870, "currency" : "EUR"}	ΔΥΠΕ Α' ΑΤΤΙΚΗΣ

Figure 9: Data Validation Results

Several noteworthy observations emerge from the data validation process on Figure 9 and are the following:

- Missing Values:
  In the "person" column, crucial information such as the Tax number aka ΑΦΜ and the name of the receiver is conspicuously absent. The "awardAmount" column exhibits missing entries, specifically lacking the amount of payment in certain instances
- Mistyped Inserts:
  Mistyped entries are discernible in the "label," "issueDate," and "awardAmount" columns, indicating potential errors or inconsistencies in the metadata
- Inconsistent Input Procedures:
  A lack of standardized input procedures becomes apparent, especially in the "awardAmount" column. Some entries include the amount with tax, while others omit tax, leading to discrepancies in the representation of financial figures

Important Update: Data Inconsistencies in Cost Variable

Following additional steps, the authors have unearthed new inconsistencies in the awardAmount variable. A subsequent sampling process has been executed, revealing discrepancies between the recorded costs and the actual values extracted from the PDF documents. Noteworthy anomalies include instances where the Tax Identification Number (TAX ID) has been erroneously input as the cost, leading to inaccuracies. Additionally, cases have been identified where the absence of commas results in the misrepresentation of decimal places, contributing to misleading cost values. The presented examples below illustrate these discrepancies, emphasizing the importance of meticulous scrutiny and further refinement in ensuring data accuracy.

subject	person	awardAmount
ΣΥΜΦΩΝΗΤΙΚΟ Αριθμός: 46/2021Για την προμήθεια του Γ.Ν. Ξάνθ	[{ "afm" : "095341314", "name" : "ΔΙΑΓΝΩΣΤΙΚΑ ΧΗΜΙΚΗ ΟΡΓΑΝ	{ "amount" : 95341314, "currency" : "EUR" }

Figure 10 Tax ID input as the cost value

E	F	G
subject	issueDate	awardAmount
Έγκριση υπογραφής ετήσιας σύμβασης συντήρησης μηχανη	2020-06-22	{ "amount" : 1866944, "currency" : "EUR"}

2.10. Έγκριση υπογραφής ετήσιας σύμβασης συντήρησης μηχανημάτων του οίκου Carestream – Kodac αντί του ποσού των 18.669,44 €

Figure 11 Absence of commas on Cost Value

Data Cleansing

A meticulous data cleaning process has been executed to ensure the dataset adheres to the desired characteristics. The key steps of this process are outlined below:

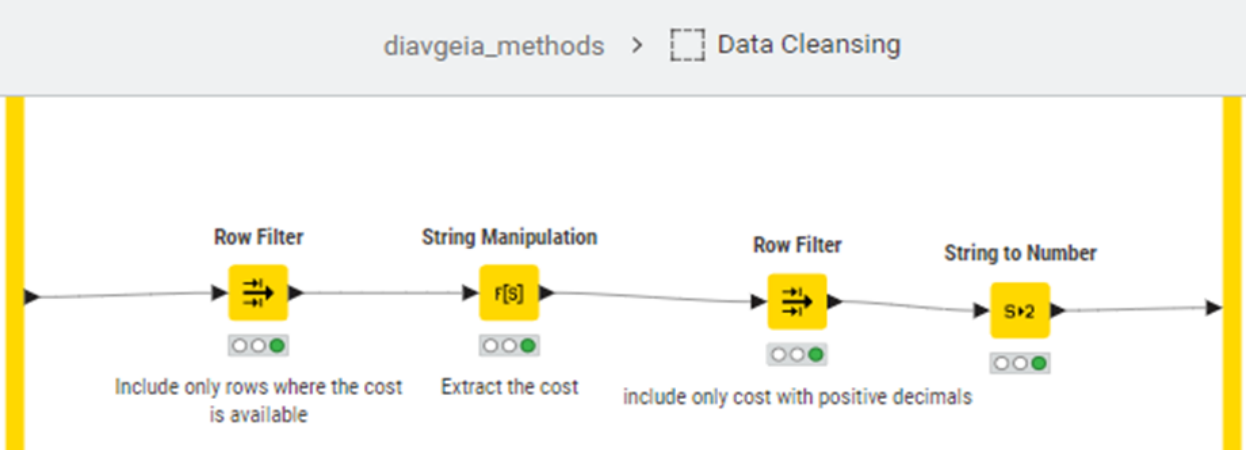


Figure 12 Data Cleansing Node

- Filtering Decisions with Available Cost:
  The initial dataset encompassed around 130,000 decisions for the specified two-year duration. However, a subset of decisions lacked cost information. To fulfill the paper's objective of illustrating government spending distribution in the healthcare sector, a decision was made to retain only rows where cost information is available in the final dataset.
- Cost Extraction with String Manipulation:
  Leveraging the string manipulation node, regular expressions were employed to extract and format cost values for each decision from the awardAmount column, creating a new column labeled "Cost". This extraction aids in creating a dedicated column for cost, facilitating ease in subsequent statistical and heatmap representations.
- Handling Cost Inconsistency:
  Using the row filter node, we mitigate the data inconsistency concerns that previously identified within the awardAmount variable, the extraction process deliberately concentrates on decimal values, ensuring a high level of accuracy and instilling resounding 100% confidence in the reliability of the extracted cost values.
- Handling Negative Cost:
  Some costs retrieved from the API were found to be negative. While the API did not provide specific information on this, upon scrutiny of PDF data and lacking domain knowledge by the authors, a decision was made to exclude these negative amounts. The Row Filter node, equipped with regular expressions, was used to selectively retain only decisions with positive costs.

In total, 26360 rows were retrieved with values in the awardamount column. From this amount, negative values, 0s and values without decimal points were excluded resulting in 17634 rows. That is

around one third of the data were excluded due to data validation reasons. This ensures for consistency and trustworthiness among the results presented in this paper.

Note: Through the assistance of Excel, we identified the number of instances where the AFM (Tax Identification Number) was erroneously imported instead of the actual amount. Remarkably, this discrepancy occurred only twice within the two-year timeframe.

Final Dataset

In the pursuit of a detailed analysis of government spending within the healthcare sector, the creation of the final dataset involved the following steps:

Mapping Hospital Locations:

The primary objective of this paper is the Geographic Analysis of Government Spending within the hospital sector. To map the hospitals into the Greek map their locations were needed. Unfortunately, the metadata lacked the information about the location. To address this, GIS software was employed to retrieve coordinates for the majority of hospitals. For the remaining hospitals, a more conventional method was employed. A manual search on a search engine was conducted using hospital names to acquire the necessary location data.

Counting Decisions for Each Hospital:

Employing the group-by method, the decisions were tallied for each hospital. This step is integral for statistical analyses and facilitates the creation of a heatmap illustrating decision counts across different hospitals. The insights derived from this count will contribute to a comprehensive understanding of decision distribution within the hospital sector.

Summing Costs for Each Hospital:

Utilizing the group-by method once again, the costs associated with each hospital were aggregated. This summation provides an overview of the total financial allocations for each hospital, offering valuable insights for statistical analyses. Additionally, this data will be used to generate a heatmap illustrating the distribution of financial resources across hospitals.

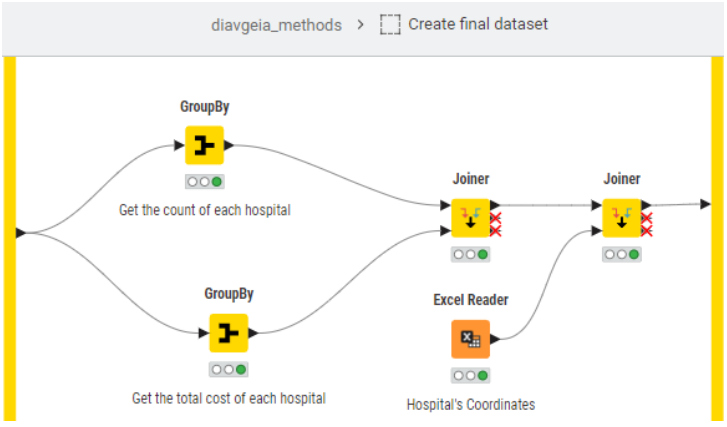


Figure 13 Final Dataset Node

This multi-faceted approach ensures the dataset is enriched with geographic coordinates, decision counts, and cumulative costs for each hospital. The resulting dataset forms a robust foundation for subsequent geographic and statistical analyses, shedding light on government spending patterns within the Greek healthcare sector.

An overview of the result can be seen below:

Rows: 83 | Columns: 6 Table Statistics

#	RowID	uid String	label String	Count*(Cost) Number (integer)	Sum(Cost) Number (double)	Longitute Number (double)	Latitute Number (double)
1	Row...	99202029	ΑΙΓΙΝΗΤΕΙΟ ΝΟΣΟΚΟΜΕΙΟ	26	72,875.27	23.754	37.979
2	Row...	99221483	ΑΝΤΙΚΑΡΚΙΝΙΚΟ ΓΕΝ.ΝΟΣΟΚΟΜΕΙ...	3	271,754.68	22.96	40.618
3	Row...	99202040	ΑΡΕΤΑΙΕΙΟ ΝΟΣΟΚΟΜΕΙΟ	20	806,327.97	23.755	37.979
4	Row...	99221484	Γ. ΑΝΤΙΚΑΡΚΙΝΙΚΟ ΝΟΣΟΚΟΜΕΙΟ«...	356	31,215,009.52	23.755	37.988
5	Row...	99221485	Γ. ΟΓΚΟΛΟΓΙΚΟ ΝΟΣ. ΚΗΦΙΣΙΑΣ «Ο...	14	62,188.86	23.77	38.08
6	Row...	99221486	Γ. ΠΑΝΑΡΚΑΔΙΚΟ ΝΟΣ. ΤΡΙΠΟΛΗΣ...	55	1,422,554.4	22.363	37.512

Figure 14 Final Dataset

Statistics and Visualizations of the dataset

We can identify from the below figure that median is 44 decisions and the 75% of the values are below 210 decisions. There are observed some outliers meaning extreme values. This can be investigated more supported by a domain expert.

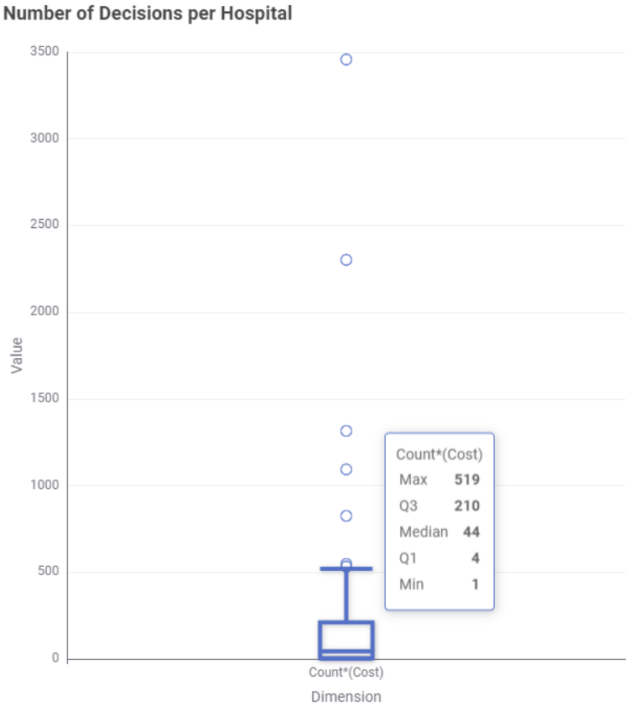


Figure 15: Box Plot of Decisions per Hospital

Below, the total percentage of the cost per hospital is depicted indicating that the biggest amount is taken by "ΓΕΝΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΛΑΡΙΣΑΣ ΚΟΥΤΛΙΜΠΑΝΕΙΟ ΚΑΙ ΤΡΙΑΝΤΑΦΥΛΛΕΙΟ".

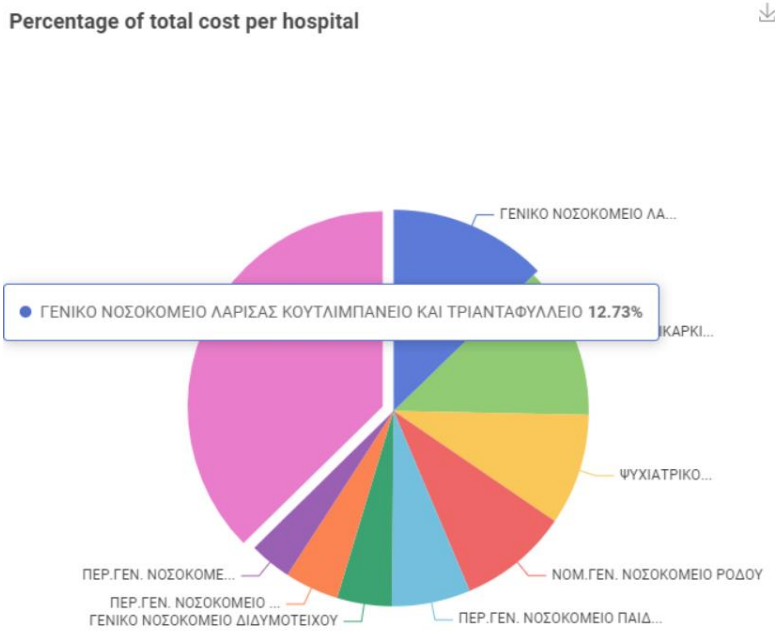


Figure 16: Percentage of total cost per hospital

The presented statistics table provides insights into the distribution of costs within the government funds allocated to hospitals during the examined two-year period:

Minimum cost: around 40 euros, indicating the least amount of financial allocations.

Mean cost: approximately 2,988,937 euros, offers a central measure reflecting the typical expenditure for the specified duration.

The cumulative expenditure across all hospitals during this period amounts: 248,081,805 euros, underscoring the considerable financial magnitude involved in government fund utilization within the healthcare sector.

Rows: 6 | Columns: 14

Table | Statistics

Name	Type	# Missing val...	# Unique val...	Minimum	Maximum	25% Quantile	50% Quantile...	75% Quantile	Mean	Mean Absolu...	Standard Dev...	Sum
uid	String	0	83									
label	String	0	83									
Count*(Cost)	Number (integer)	0	58	1	3,458	4	44	210	210.313	249.921	491.065	17,456
Sum(Cost)	Number (double)	0	83	40.94	31,591,309.82	70,200.19	432,271.86	2,669,867.68	2,988,937.422	3,653,409.696	6,229,576.393	248,081,805.99
Longitude	Number (double)	0	80	19.851	28.193	22.376	23.6	23.807	23.351	1.199	1.579	1,938.103
Latitude	Number (double)	0	80	35.194	41.352	37.979	38.17	40.385	38.688	1.27	1.525	3,211.073

Figure 17: Statistics Table

GeoSpatial Analysis and Results

By default, the KNIME installation does not support GeoSpatial nodes, the GeoSpatial extension was recovered and installed into our workstations. This extension provides numerous options and spatial algorithms for KNIME.

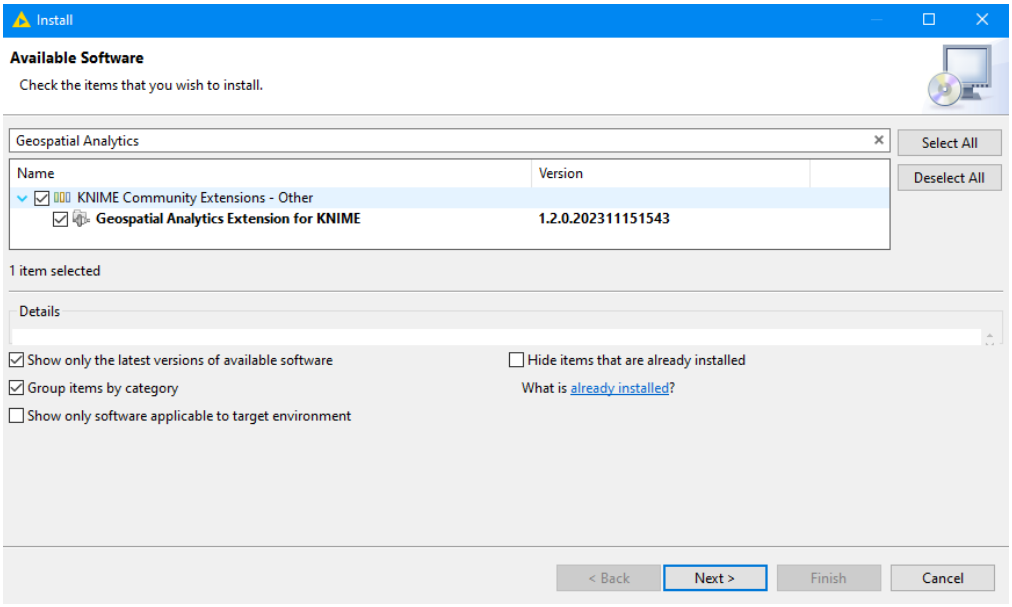


Figure 18: Installing Geospatial Analytics Extension

Getting the map layers containing EU Region and Population information from EuroStat

Getting the NUTS 2021[1] from NUTS - Eurostat (europa.eu) (https://gisco-services.ec.europa.eu/distribution/v2/nuts/download/). The NUTS[2] are a hierarchical system divided into 3 levels. NUTS 1: major socio-economic regions, NUTS 2: basic regions for the application of regional policies, NUTS 3: small regions for specific diagnoses. Additionally, a NUTS 0 level, usually co-incident with national boundaries are also available.

Additional LAU 2021 was downloaded from LAU - GISCO - Eurostat (europa.eu) https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/lau

LAUs are the building blocks of the NUTS (Nomenclature of territorial units for statistics) and statistical regions, and comprise the municipalities and communes of the European Statistical

[1] The GISCO statistical unit dataset represents the NUTS (Nomenclature of territorial units for statistics) and Statistical regions by means of multipart polygon, polyline and point topology. The NUTS geographical information is completed by attribute tables and a set of cartographic help lines to better visualize multipart polygonal regions.
[2] Eurostat europa eu. NUTS - GISCO - Eurostat (europa.eu) - https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts.

System (ESS). Data is available annually and also includes total resident population figures per LAU where available. Further information on LAUs can be found in the NUTS dedicated section.

Combining Nuts 2021 and LAU 2021

Combining LAU and NUTS Level 3 using publicly available datasets is not as straightforward as it may seem, due to small inconsistencies in the data, the cumbersome and internally inconsistent format chosen by GISCO to distribute the concordance tables, and first of all, the puzzling choice not to include reference to NUTS in the datasets with local administrative units to begin with[3].

Therefore, a csv table with optimal combination of NUTS and LAU was retrieved for this exercise from github [4]. Csv file name: lau_2020_nuts_2021_concordance_by_geo.csv.

Combining results to produce spatial and Heat maps.

The final results Diaygeia will be used combined with spatial data to produce heat map and spatial view of the Hospitals and their expenses. This will allow a better visualization and improve the understanding and explainability of our data.

Steps followed:

1. The final dataset contains Names of the Hospitals. The Hospitals should be georeferenced by enriching the final dataset with the relevant hospital coordinates (longitude and Latitude) . A map layer , polygon shape file, with Hospital coordinates was downloaded from Healthsites.io[5] , Building an open data commons of health facility data with OpenStreetMap.
2. Joining of the locations of Hospitals map layer with our final dataset was done based on Hospital Name. A number of Hospital names were matched manual due to slightly different spelling of the Hospital name.

Row ID	S uid	S label	I Count*...	D Sum(Cost)	D Longitude	D Latitude
Row0	99202029	ΑΙΓΙΝΗΤΕΙΟ ΝΟΣΟΚΟΜΕΙΟ	26	72,875.27	23.754	37.979
Row1	99221483	ΑΝΤΙΚΑΡΚΙΝΙΚΟ ΓΕΝ.ΝΟ...	3	271,754.68	22.96	40.618
Row2	99202040	ΑΡΕΤΑΙΕΙΟ ΝΟΣΟΚΟΜΕΙΟ	20	806,327.97	23.755	37.979
Row3	99221484	Γ. ΑΝΤΙΚΑΡΚΙΝΙΚΟ ΝΟΣΟ...	356	31,215,009.52	23.755	37.988
Row4	99221485	Γ. ΟΓΚΟΛΟΓΙΚΟ ΝΟΣ. Κ...	14	62,188.86	23.77	38.08
Row5	99221486	Γ. ΠΑΝΑΡΚΑΔΙΚΟ ΝΟΣ. Τ...	55	1,422,554.4	22.363	37.512

Figure 19: View of Hospitals and Locations Tables

[3] https://edjnet.github.io/lau_centres/lau_nuts.html

[4] https://edjnet.github.io/lau_centres/lau_nuts_concordance_by_geo/lau_2020_nuts_2021_concordance_by_geo.csv

[5] https://healthsites.io/map?country=Greece *Building an open data commons of health facility data with OpenStreetMap

3. Joining Nuts 3 level with LAU shape files.
   3.1 A joiner node was used to join the tables of LAU table (attribute Lau_Id) with the csv file containing the Lau_id of the Nuts_id. The result was filtered with a filter node and appears as following:

Table "default" - Rows: 6135 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	GISCO_ID	CNTR_...	LAU_ID	LAU_NAME	POP_2...	POP_D...	AREA_...	YEAR	nuts_3
Row10241_R...	EL_07030301	EL	07030301	Κοινότητα Νέας Απολλωνία	1922	26.928	71.375	2021	EL522
Row10242_R...	EL_07030302	EL	07030302	Κοινότητα Μελισσουργού	417	18.309	22.775	2021	EL522
Row10243_R...	EL_07030303	EL	07030303	Κοινότητα Νικομηδινού	519	43.149	12.028	2021	EL522

Figure 20: View of the table including LAU and Lau_id and Nuts_id

The result allows to a many (nuts_3) to many (population_2021) final table.

3.2 A group by node was used to group by nuts_3 and create the sum the population for each nuts_3. This results in having a table with the total number of population by each Greek small region (“Διοικητική Περιφέρεια” , or as formerly known “Νομός”).

3.3 A final joiner node between the 3.2 results and the NUTS map layer/spatial table, was done. This allowed the enrichment of the Nuts table with the population of 2021 within the boundaries of each Greek small region (“Διοικητική Περιφέρεια” , or as formerly known “Νομός”).

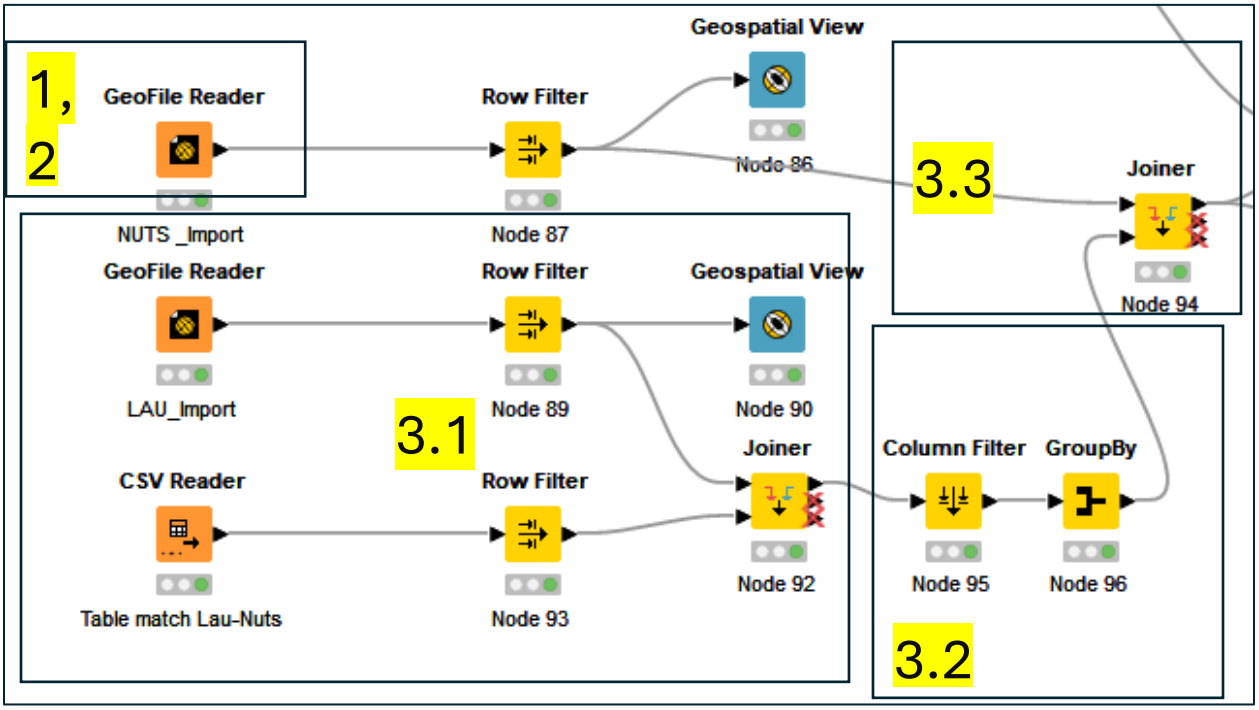


Figure 21: Geospatial Analytics Knime Workflow Steps 1 - 3.3

4. Final join between Nuts 3 level and our final data set.
   A spatial join node was used to join our final data set and Nuts_3 level. This allowed the table of Greek small regions to contain the total amount of expenses of the Hospitals to each Greek Small region to the final table. Additional the Names of the hospitals were enriched to the Joined layer.
5. Adding a node to fill the missing values with zero.
6. Adding attribute representing expenses by region / to population of the same region

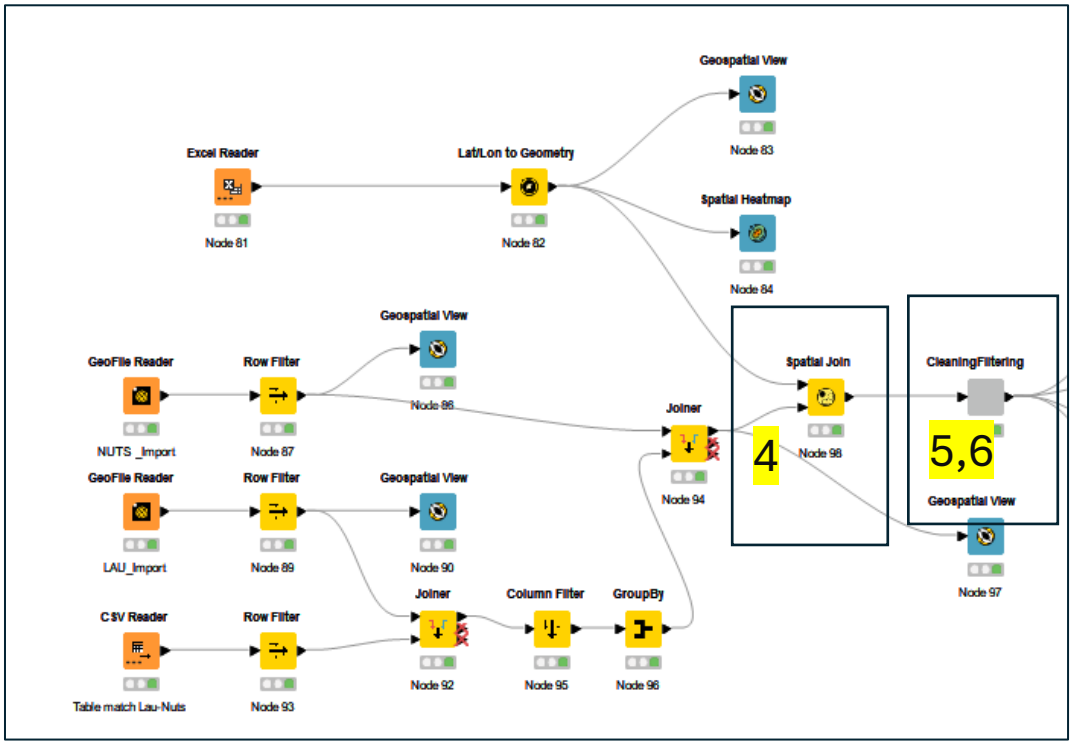


Figure 22: Geospatial Analytics Knime Workflow Steps 4 - 6

7. Geospatial representation and spatial Heat Map of the expenses during 2020 and 2021.

   7.1 Spatial Map Hospitals (points weighted by expenses- 5 classes)

   7.2 HeatMap by Hospital expenses

   7.3 Spatial Map Hospital expenses by Administration Area (points weighted by sum of expenses- 10 classes)

   7.4 Expenses ratio per person by administration area

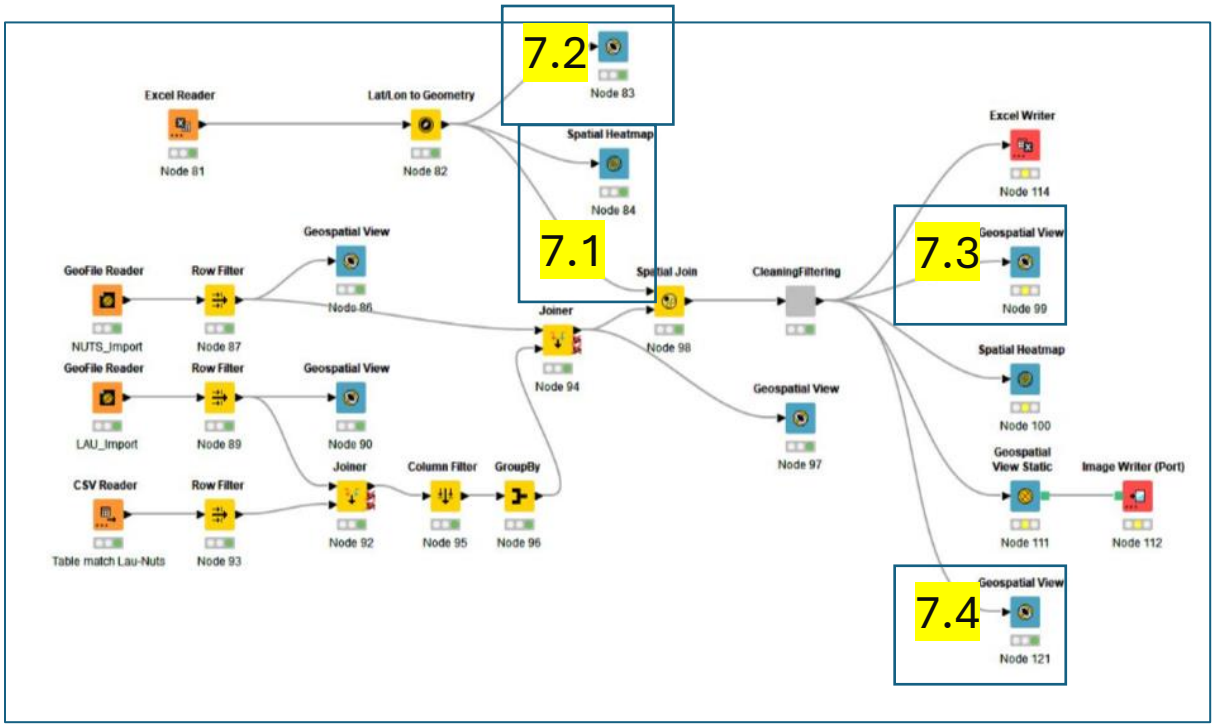


Figure 23: Geospatial Analytics Knime Workflow Steps 7.1 - 7.4

Results

Based on the above workflow the authors were able to create some visualizations in order to get important insights. The 7.1 figure shows a heatmap per Hospital with the corresponding expenses. In addition, the 7.2 figure illustrates a spatial map of hospitals categorized in five different classes. Furthermore, a spatial map hospitals expenses by administration Area is depicted in 7.3 figure. Lastly, figure 7.4 describes the ratio of expenses per hospital based on the hospital's administration area.

7.1 HeatMap by Hospital expenses

7.2 Spatial Map Hospitals (points weighted by expenses- 5 classes)

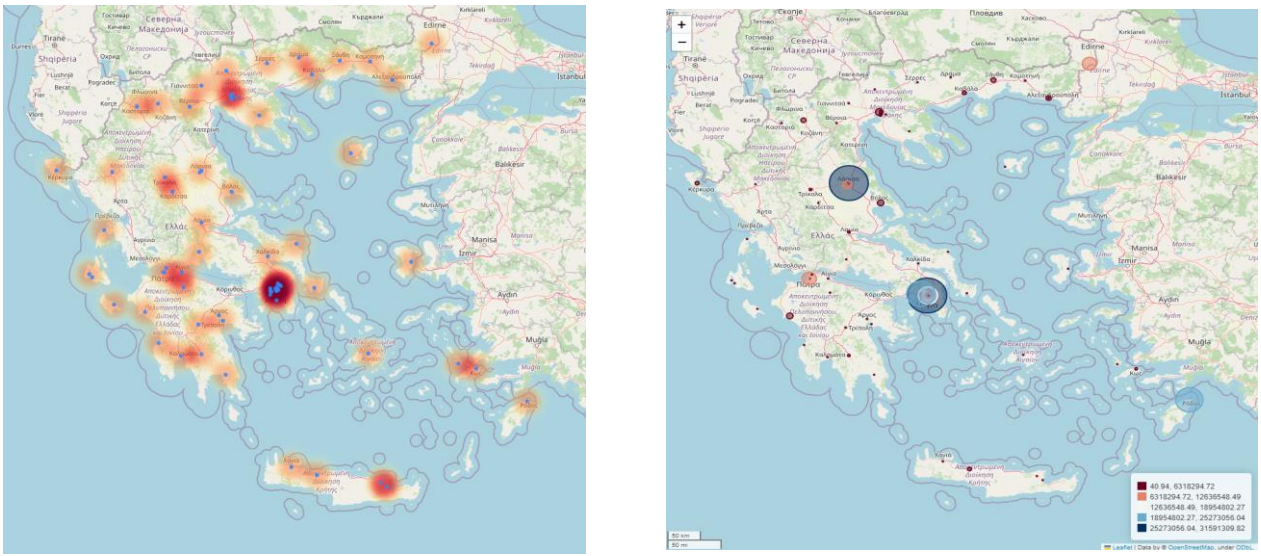


Figure 24: Heatmap and Map of Hospitals by expenses

Based on the figure 7.1 most expenses are observed in the administration area of Athens. Secondly, comes the administration area of Thessaloniki which has the next most expenses in the Greece country. This is logical compare to the size of population of each area. The rest of the areas follow this logical assumption.

On the figure 7.2 expenses per hospital are depicted. Blue colors describe large number of expenses whereas red colors illustrate small values of expenses. It can be seen that hospitals in Athens spent more expenses than Thessaloniki which is logical since the population of the region is larger. However, there are hospitals in Larissa where they have greater expenses than Thessaloniki or Patra which is important to be noted. Larissa although it has lower population than the previous referred cities it appears with larger expenses.

7.3 Spatial Map Hospital expenses by Administration Area (points weighted by sum of expenses- 10 classes

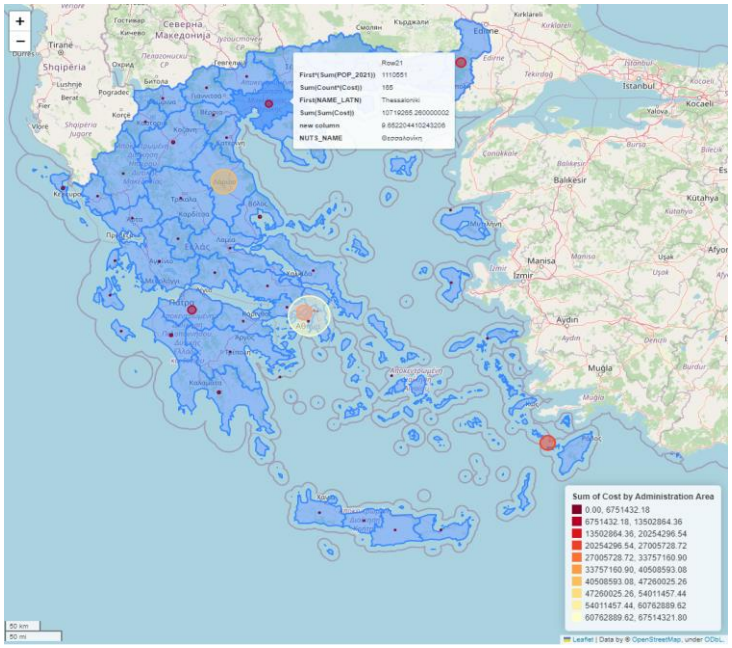


Figure 25: Spatial Map Hospitals expenses by Administration Area

7.4 Ratio of expenses per hospital based on the hospital's administration area

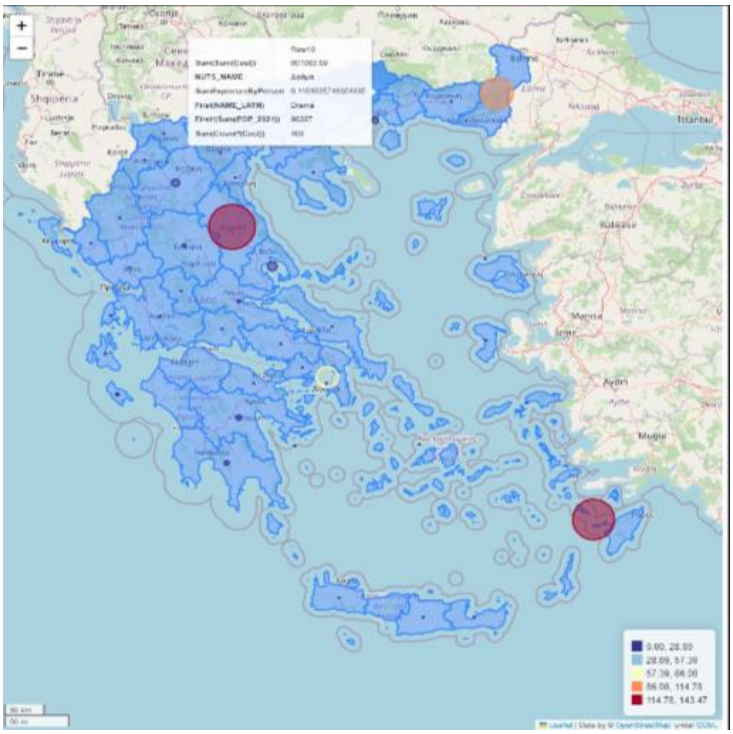


Figure 26: Ration of expenses per hospital based on the hospital's administration area.

The next figure 7.3 groups by administration area the expenses of hospitals inside that area. The lighter colors indicate larger amount of expenses whereas more red it is the lower the amount of expenses. Easily it can be seen that Athens area absorbs most of the expenses while secondly comes the administration area of Larissa city.

Last figure 7.4 highlights the ratio of hospitals expenses based on the population of each hospital's administration area. We can see easily that Larissa's and the South Dodecanese islands' hospitals spent more per person in comparison to other hospitals in other parts of Greece.

Conclusion

In conclusion, the objective of this paper is to assess the regional distribution of government funds and contracts allocated to hospitals in Greece for two years period 2020-2021 and 2021-2022 described by the COVID-19 pandemic. This was achieved by extracting metadata from Diavgeia's API and using Knime platform the authors utilized analytical and explanatory techniques. A thoroughly data validation process took place enhancing the accuracy and completeness of the extracted metadata. The cleaning process was vital due to the quality of metadata was low. After that, the data were ready for visualizations where the Geospatial Analytics Extension was used for that reason. Several figures created resulting in different conclusions depending on the angle of the view of the data. Noticeably, the biggest amount of the expenses was done by Athens and Larissa's administration areas. To summarize, since the overall quality of metadata was poor no clear picture of how the actual expenses got distributed across the Hellenic country.

Limitations

While the paper provides valuable insights, it is important to acknowledge certain limitations. The primary constraint lies in the poor quality of metadata, characterized by the absence of VAT, float indicators, and instances of negative values. These limitations may have led to the oversight of certain patterns in the distribution of government funds. Additionally, the varied PDF formats and lack of a standardized template across hospitals posed challenges in extracting information consistently, necessitating adaptive engineering techniques. These limitations underscore the need for cautious interpretation and consideration of potential gaps in the analysis.

Future work

Consideration should be given to implementing standardized data entry protocols and ensuring that all pertinent information is consistently captured to mitigate discrepancies and enhance the overall quality of the dataset. In addition, one can implement prompt engineering techniques in large scale in order to capture a larger number of samples through the data validation process. This can be scaled using new cutting-edge technologies such as LLM's.

Appendix

Figure 1: Overall KNIME workflow. 5
Figure 2: Additional Row Filter Nodes (Get Organizations Node) 5
Figure 3: Hospitals Row Filter node configuration. 6
Figure 4: Exclusion of Non-Hospitals Row Filter node configuration. 6
Figure 5: Organizations table - 140 Hospitals in total 7
Figure 6: Get Decisions Node Configuration 8
Figure 7: Joiner Node – configuration 9
Figure 8: Combining information for the final dataset. 9
Figure 9: Data Validation Results 10
Figure 10 Tax ID input as the cost value 11
Figure 11 Absence of commas on Cost Value 11
Figure 12 Data Cleansing Node 12
Figure 13 Final Dataset Node 13
Figure 14 Final Dataset 14
Figure 15: Box Plot of Decisions per Hospital 14
Figure 16: Percentage of total cost per hospital 15
Figure 17: Statistics Table 15
Figure 18: Installing Geospatial Analytics Extension 16
Figure 19: View of Hospitals and Locations Tables 17
Figure 20: View of the table including LAU and Lau_id and Nuts_id 18
Figure 21: Geospatial Analytics Knime Workflow Steps 1 - 3.3 18
Figure 22: Geospatial Analytics Knime Workflow Steps 4 - 6 19
Figure 23: Geospatial Analytics Knime Workflow Steps 7.1 - 7.4 20
Figure 24: Heatmap and Map of Hospitals by expenses 21
Figure 25: Spatial Map Hospitals expenses by Administration Area 22
Figure 26: Ration of expenses per hospital based on the hospital's administration area. 22