DS 4A COLOMBIA

Colombia Compra Eficiente

SECOP TOPICS

PUBLIC PROCUREMENT TOPIC CLUSTERING ON KEYWORDS FOR CONTRACTUAL OBJECTIVES

Colombia Compra Eficiente - Agencia Nacional De Contratación

Team 82

Camilo Cabrera
Nicolas Casas
Jorge Enciso
Samuel Pérez
Cindy Ramirez
Karina Mesa
Yasmin Moya

Abstract

Public procurement refers to the process by which public authorities, such as government departments or local authorities, purchase work, goods or services from companies. In Colombia, *Colombia Compra Eficiente* is the agency whose main role is to supervise public procurement, and aims to maximize its impact and efficiency, by knowing what is being purchased by local governments and state owned enterprises, and why it is being purchased. SECOP TOPICS is a dashboard that offers a clean and simple way to show which are the main reasons of public procurement in Colombia, providing insights on each topic though Latent Dirichlet Allocation machine learning algorithm (LDA). It also offers a clean analysis on where the money is being spent geographically and how the procurement varies with time. With SECOP TOPICS, *Colombia Compra Eficiente* will have a clear view of the different topics the colombian government and institutions are purchasing.

Content:

1. Context

Public procurement refers to the process by which public authorities, such as government departments or local authorities, purchase work, goods or services from companies[1]. As public procurement accounts for a substantial portion of the taxpayers' money, governments are expected to carry it out efficiently and with high standards of conduct in order to ensure high quality of service delivery and safeguard the public interest[2].

Public procurement can be a key tool in driving the development of key sectors in different regions, by acquiring the adequate goods and services on the Colombian territory. As the local entity which supervises the public procurement, Colombia Compra Eficiente looks to maximize the value generated by acquisitions, and improve its efficiency.

The *Agencia Nacional de Contratación Pública - Colombia Compra Eficiente (ANCPCCE)*[3] , as a regulatory entity, aims to develop and promote public policies and tools for organization and articulation of the stakeholders in purchase processes and public procurement, in order to improve efficiency, transparency, and optimization of the public resources.

The data on public procurement in Colombia (contracts, process milestones, budget, etc.) are recorded in the web applications known as "SECOP I" and "SECOP II" supervised by Colombia Compra Eficiente.

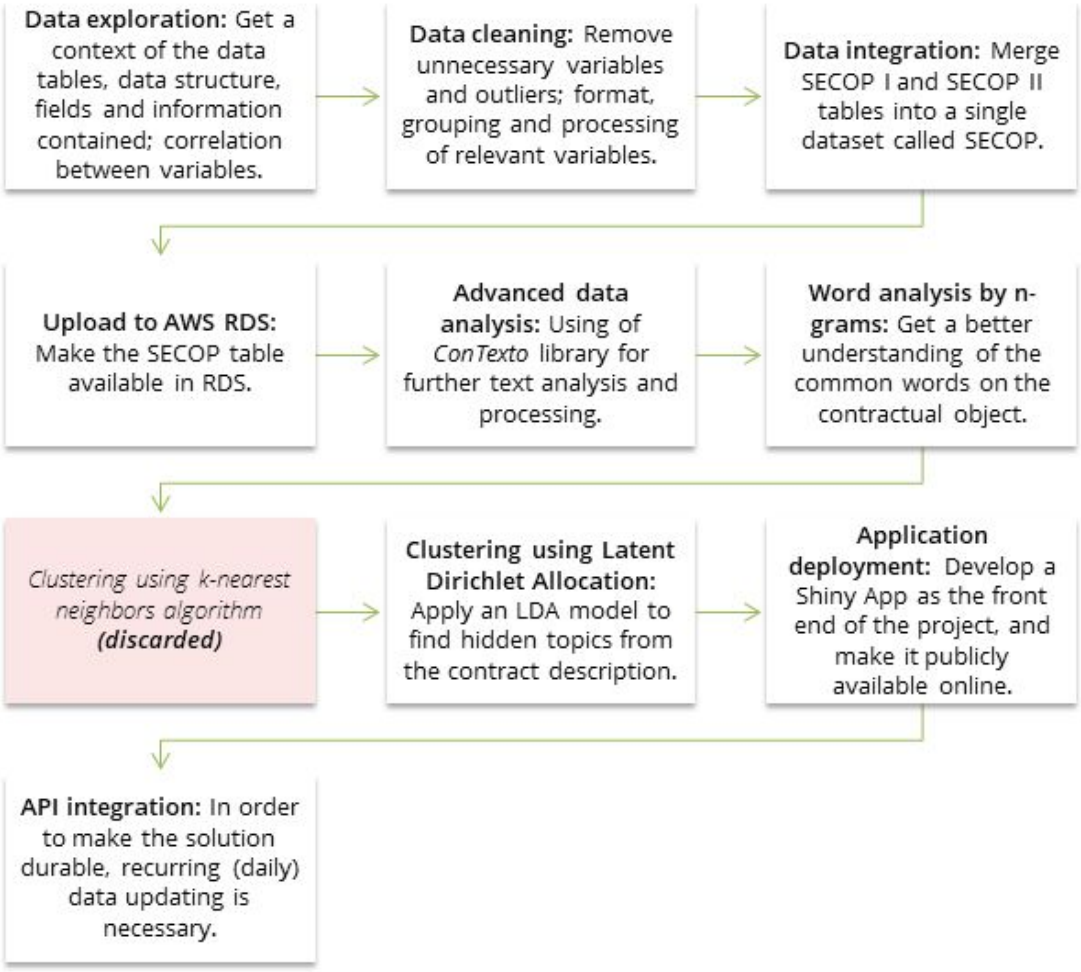
The organization requires to group contracts that are related by identifying common words in the contractual object in an automatic way. Most of the time, the organization does it manually. This activity is cumbersome and rudimentary.

Thus *Colombia Compra Eficiente* needs an analytical model that allows forming groups of contracts automatically by analyzing the contractual object. In this way, the organization will be able to generate insights and optimize its process, and finally help public entities to improve purchasing.

1.1. Why does this problem matter?

Grouping contracts by keywords provides a broader and more realistic overview of public procurement, by having a greater understanding of the goods or services that are being used by public entities. This is a task that, if done manually, would require dedicated personnel to read all the contracts and classify them, and some sort of process and procedure update to classify all the new contracts. The actual solution Colombia Compra Eficiente has implemented, that is grouping by keywords in raw data, has proven to be inefficient, inconvenient and to drive incomplete results. Having a ML algorithm to cluster contracts by their contractual object, and analyze the commod terms among them would allow *Colombia Compra Eficiente* to have clearer and more useful information to generate real value and savings in the country's public procurement.

2. Workflow

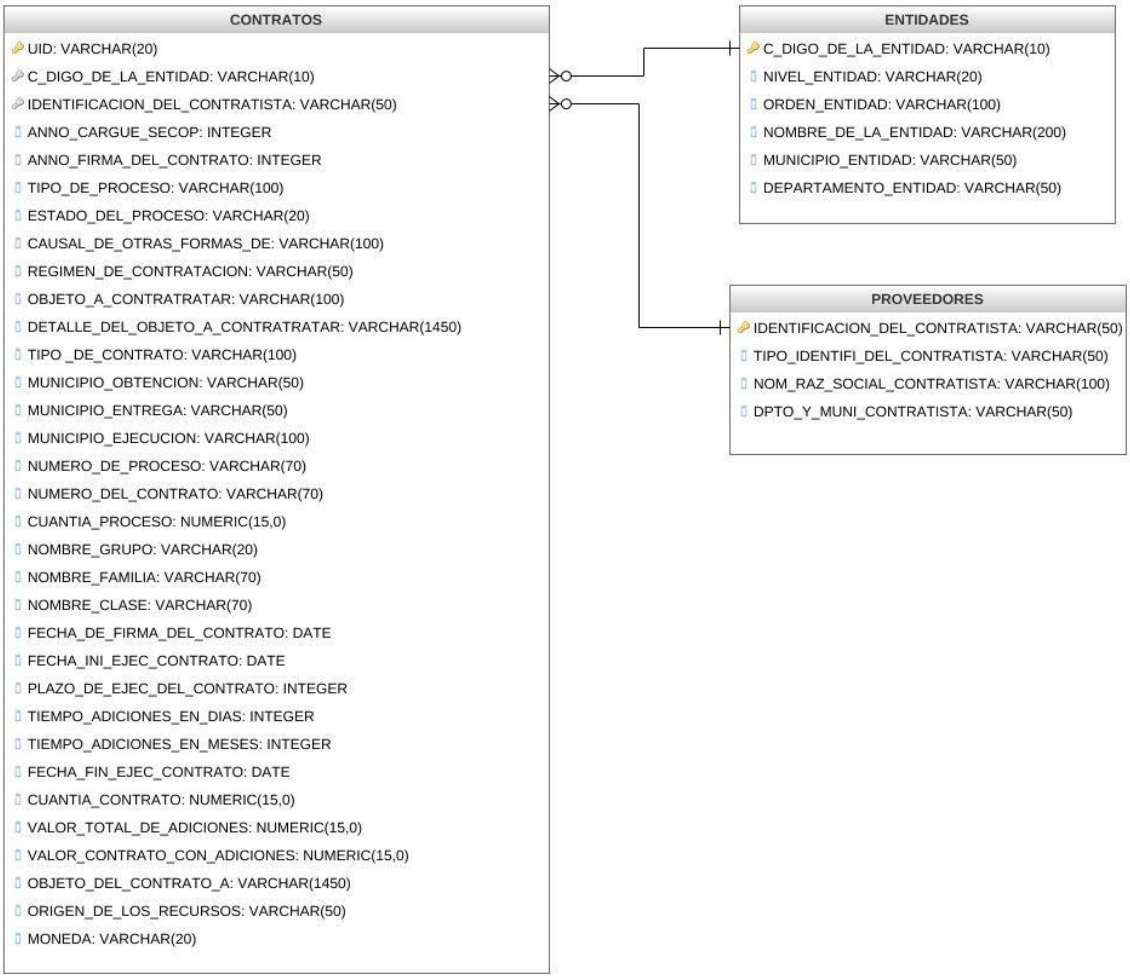


Graph 1. Workflow. Own source

3. Information Sources

SECOP - I

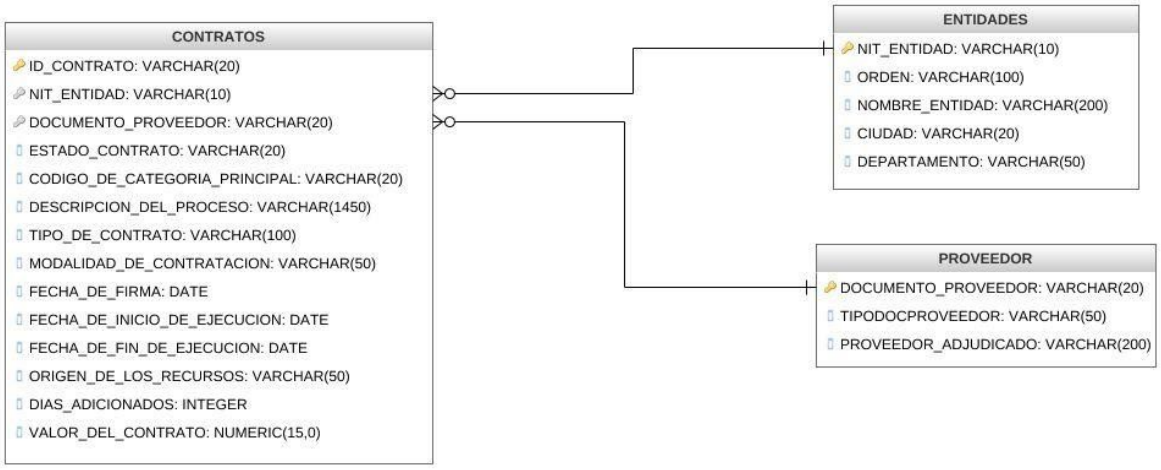
The information is self-documented by the public entities of the country. Each row of the database corresponds to a contract, there are more than 527 thousand rows. It has 54 columns in which each column corresponds to information about the process of procurement[4].



Graph 2. SECOP I Diagram. Own source, based on [4]

SECOP - II

The information is self-documented by the public entities of the country. Each row of the database corresponds to a contract, there are more than 9 million rows. It has 72 columns in which each column corresponds to information about the process of procurement[4].



Graph 3. SECOP II Diagram. Own source, based on [4]

4. EDA

The Exploratory Data Analysis (EDA) is divided into two sections. The first one comprises the basic analyses over numerical and categorical variables. The second one performs basic statistics over the contractual object. This is a text variable and is the focus of our analysis.

The name, description, data type, cleaning method and relevance of each variable is in the appendix.

Below are some graphs referring to the contracts, taking into account the data contained in the SECOP I and SECOP II databases of the company *Colombia Compra Eficiente*.

For the purposes of a better visualization, the first 20 categories are taken into account if there are more than a thousand categories.

4.1. SECOP I

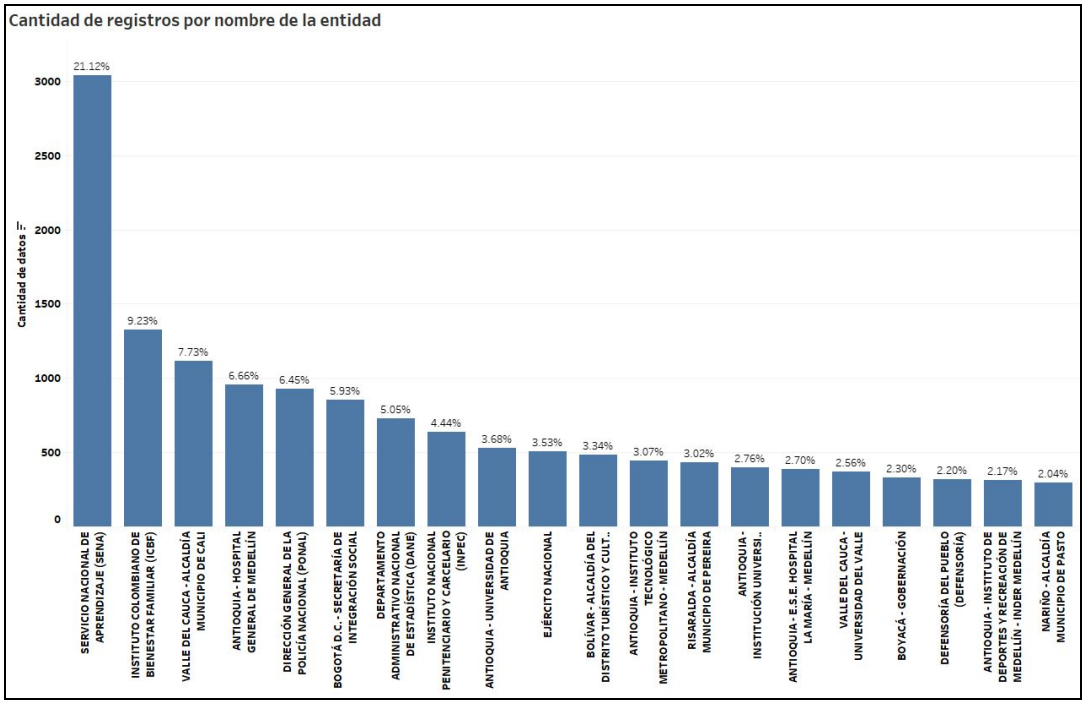
Objeto A Contratal.	Valor promedio del contrato ya pagado	Valor del contrato promedio	Valor Adicional promedio	Tiempo adicional en días	Plazo para pagar	Cantidad de datos
Servicios de Edificación, Construcción de Instalaciones y Mantenimiento	$ 763,222,736.68	$ 751,279,833.24	$ 11,942,903.43	3	35	3494
Terrenos, Edificios, Estructuras y Vías	$ 640,378,149.20	$ 611,668,351.06	$ 28,709,798.14	6	46	996
Servicios Financieros y de Seguros	$ 358,640,521.61	$ 353,144,784.68	$ 5,495,736.93	3	72	996
Servicios Políticos y de Asuntos Cívicos	$ 194,160,903.74	$ 179,073,666.58	$ 15,087,237.16	4	76	1760
Servicios Basados en Ingeniería, Investigación y Tecnología	$ 181,768,586.61	$ 174,745,734.16	$ 7,022,852.46	5	57	1920
Servicios Públicos y Servicios Relacionados con el Sector Público	$ 132,808,953.55	$ 117,772,408.91	$ 15,036,544.64	2	66	1477
Servicios Educativos y de Formación	$ 131,530,212.19	$ 125,220,138.50	$ 6,310,073.69	4	88	3217
Otros	$ 113,883,588.55	$ 109,525,554.53	$ 4,358,034.03	2	33	12650
Servicios de Transporte, Almacenaje y Correo	$ 103,489,538.23	$ 94,223,765.40	$ 9,265,772.83	2	46	1745
Servicios de Viajes, Alimentación, Alojamiento y Entretenimiento	$ 99,790,541.86	$ 88,726,336.70	$ 11,064,205.16	2	35	1497
Alimentos, Bebidas y Tabaco	$ 93,478,635.44	$ 76,790,560.15	$ 16,688,075.29	1	35	1099
Servicios de Salud	$ 54,641,843.97	$ 50,214,614.19	$ 4,427,229.78	4	40	9594
Medicamentos y Productos Farmacéuticos	$ 39,891,468.31	$ 33,460,968.77	$ 6,430,499.54	1	27	1054
Equipo Médico, Accesorios y Suministros	$ 35,487,044.67	$ 33,405,226.75	$ 2,081,817.92	2	26	1790
Servicios Editoriales, de Diseño, de Artes Graficas y Bellas Artes	$ 30,996,677.04	$ 29,191,557.94	$ 1,805,119.10	2	72	1316
Equipos de Oficina, Accesorios y Suministros	$ 20,803,651.36	$ 18,898,923.78	$ 1,904,727.59	1	21	1364
Servicios Personales y Domésticos	$ 16,859,626.00	$ 16,143,047.84	$ 716,578.16	3	66	1891

Graph 4. Quantitative variables heat map by the contractual subject. Own source, based on [4]

We can see that of the five most important variables to consider in the exploratory data analysis, regarding the object to be hired, are construction and building services, land, buildings and infrastructure in general, followed by financial and insurance activities. By the other hand, the least subjects that are invested are the domestic services and the equipment per office with an average of 20 million per contract.

In terms of time limit, educational and formative including art contracts have a larger slack to be paid rather than office equipment or pharmaceutical contracts. On other hand, the quantity has a lot of concentration in the health contracts and in others contracts. This means that a considerable amount of contracts have not defined subject.

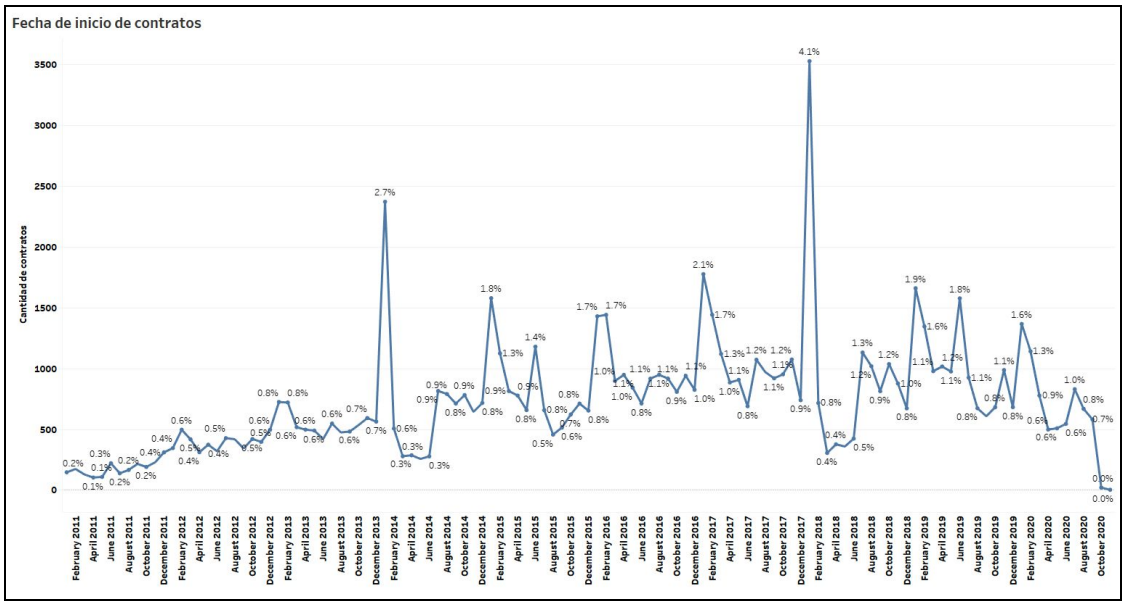
Number of contracts per entity



Graph 5. Number of contracts per entity. Own source, based on [4]

From this graph, SENA is the entity with the highest number of contracts followed by the ICBF, the Mayoralty of Cali, the General Hospital of Medellin and the General Directorate of the National Police.

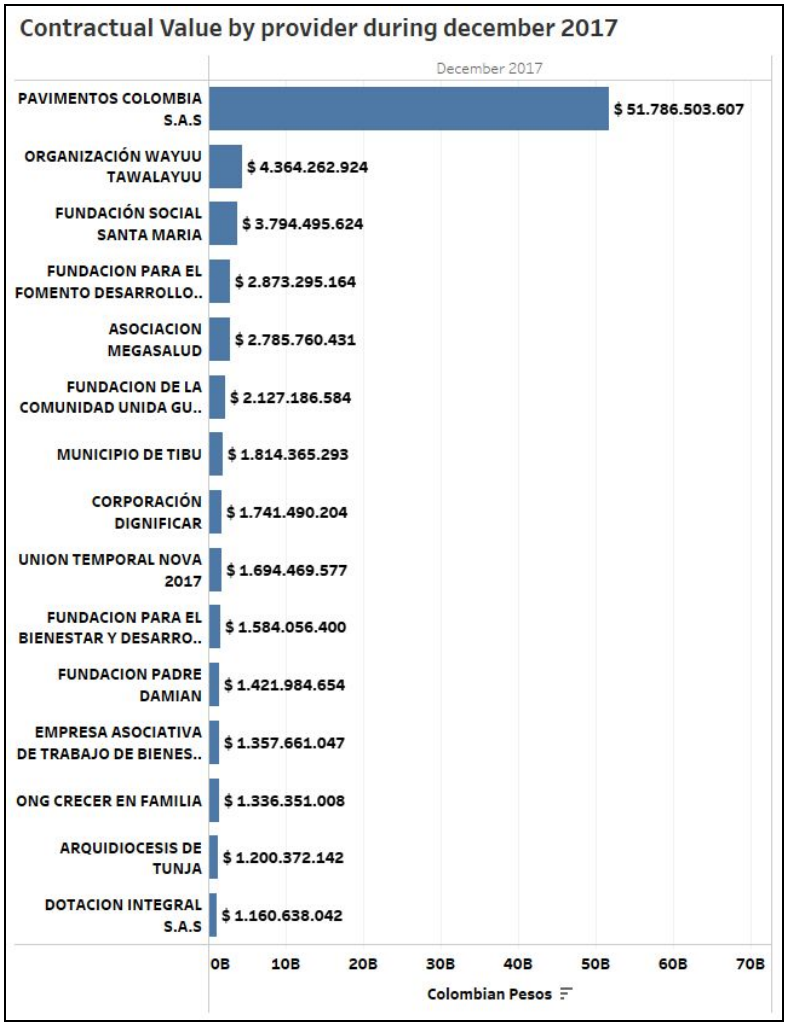
Contract start date



Graph 6. Initial date of contracts. Own source, based on [4]

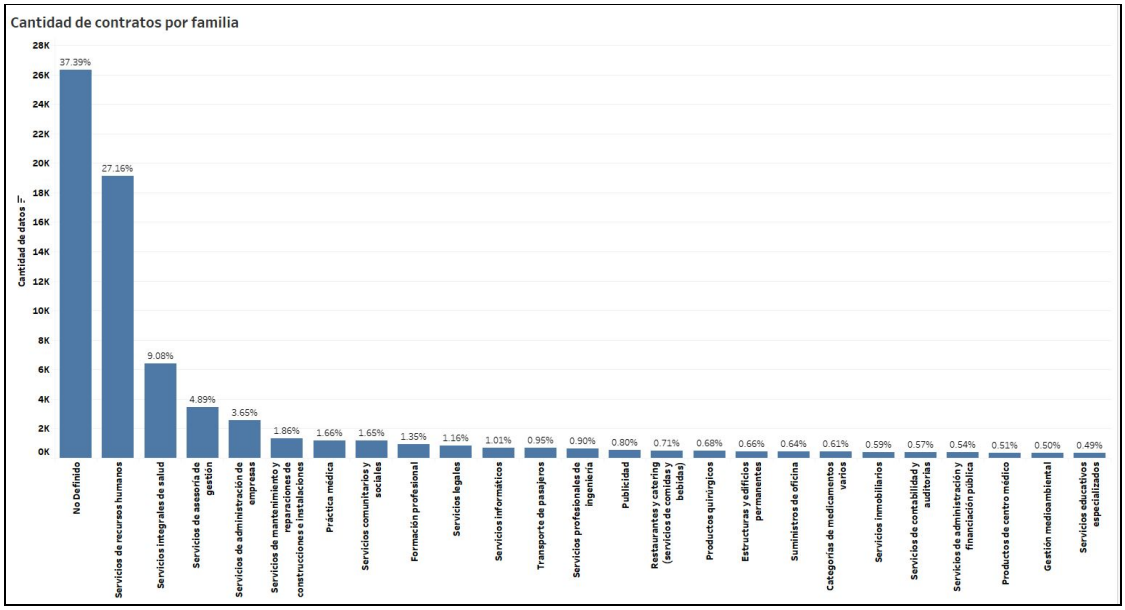
Main insights:

1. There are hiring peaks at the end of the year and in the middle. Lowest hiring peak during August 2020.
2. Maximum value in December 2017 and December 2013 due to “Ley de Garantías” (a special law that restricts contract signs before elections). During December 2017, the supplier that billed the most was Pavimentos de Colombia with a contract for 51 billion COP and the next supplier was Organizacion Wayuu Tawayuluu with 4.3 billion.



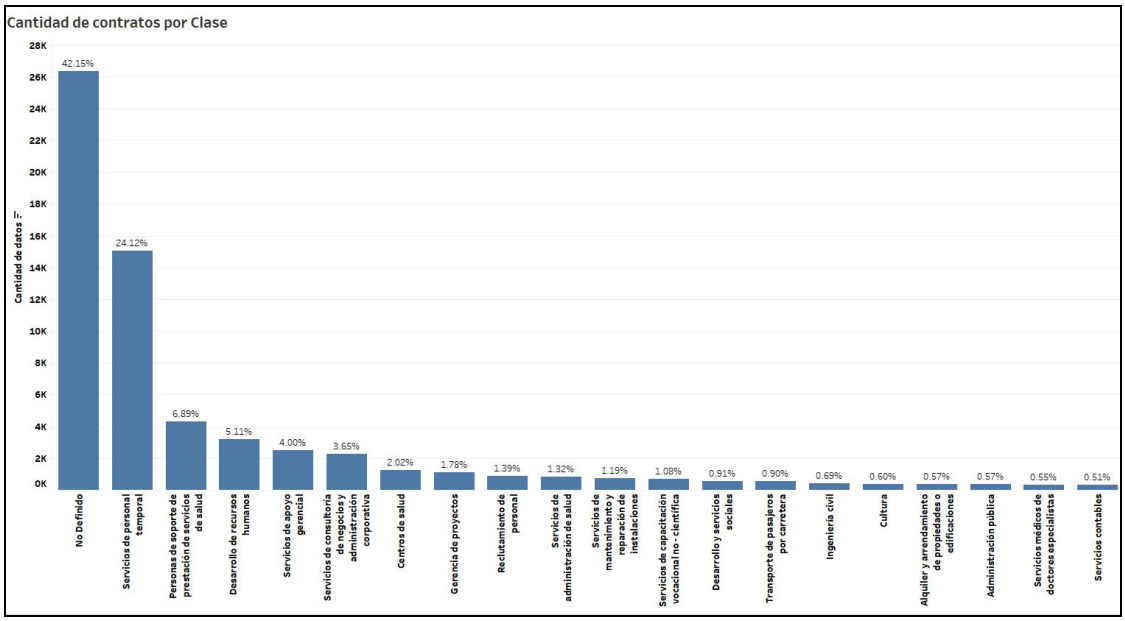
Graph 7. Contractual Value by provider during December 2017. Own source, based on [4]

Number of contracts per family



Graph 8. Number of contracts per family. Own source, based on [4]

Number of contracts per class



Graph 9. Number of contracts per class. Own source, based on [4]

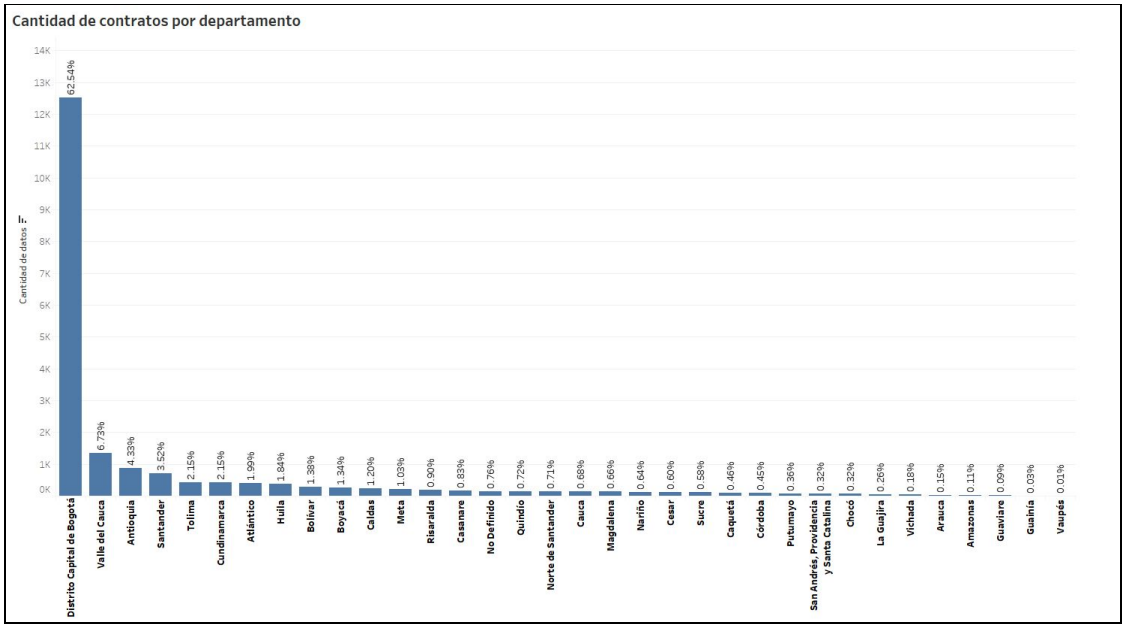
These two graphs show the family and the class for each procurement. There is a lot of variety in these two graphs. As it is very hard to conclude a trend between these classes,

the only conclusion is that most of the values have not a family or class. Also the temporary services and the human resources contracts are the most common among all the families and contracts.

4.2. SECOP II

Each row of the database corresponds to a contract, there are more than 9,23 million rows. It has 72 columns in which each column corresponds to information about the process of procurement. However we are connected to the Official API from SECOP, so we are analyzing a sample of 20.000 records.

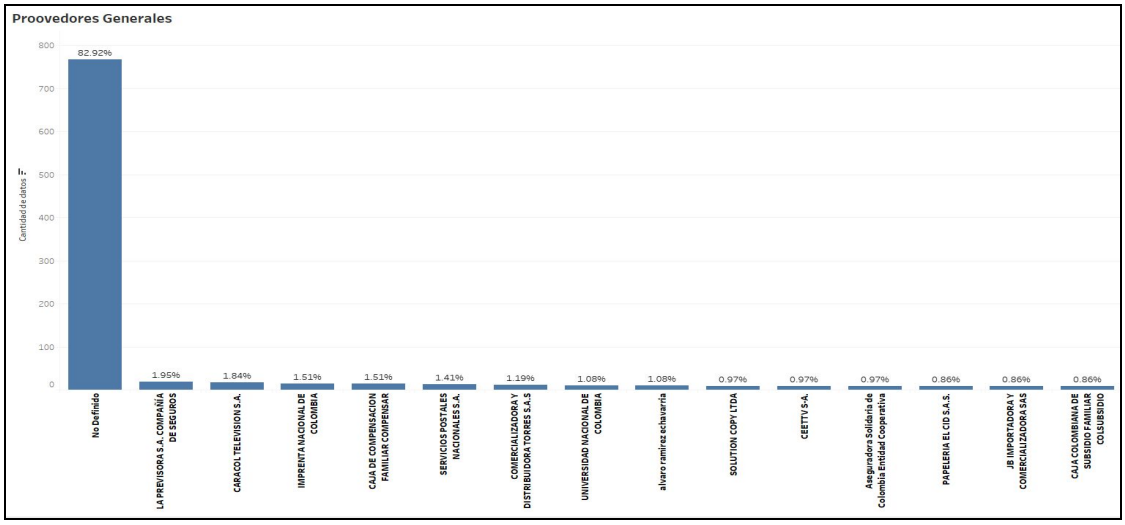
Number of contracts by department



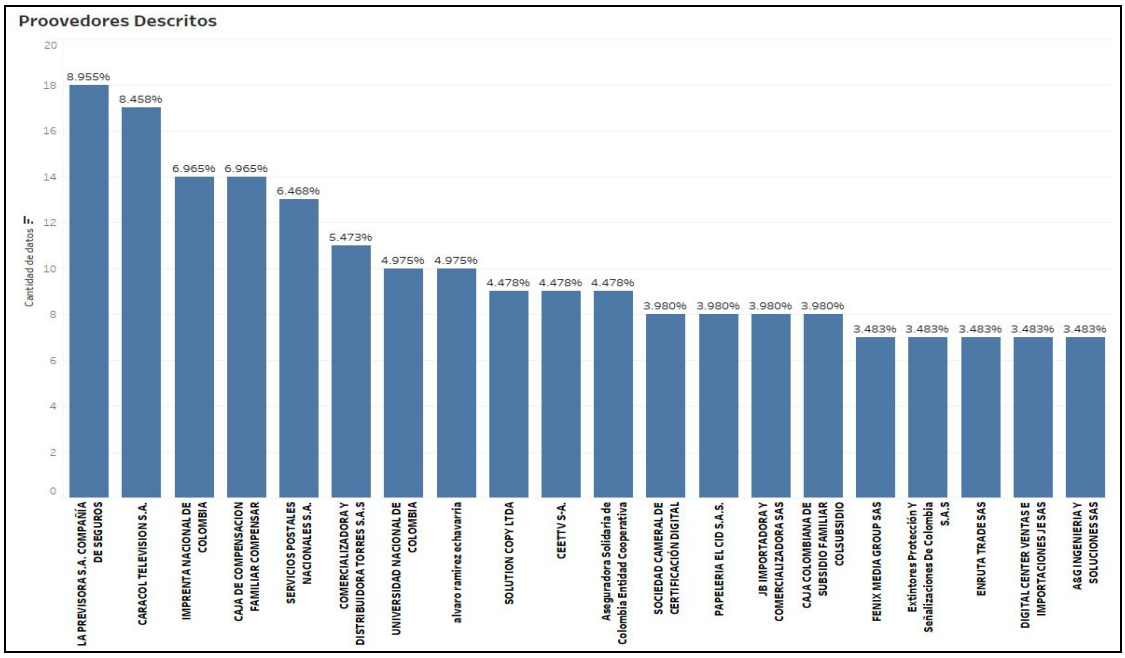
Graph 10. Number of contracts by department. Own source, based on [4]

In graph 7, it can be seen that the vast majority of contracts were negotiated in Bogotá, the Capital District, this may occur because the companies that contract with the state are based in this city.

Suppliers



Graph 11. General Supplies. Own source, based on [4]



Graph 12. Define Suppliers. Own source, based on [4]

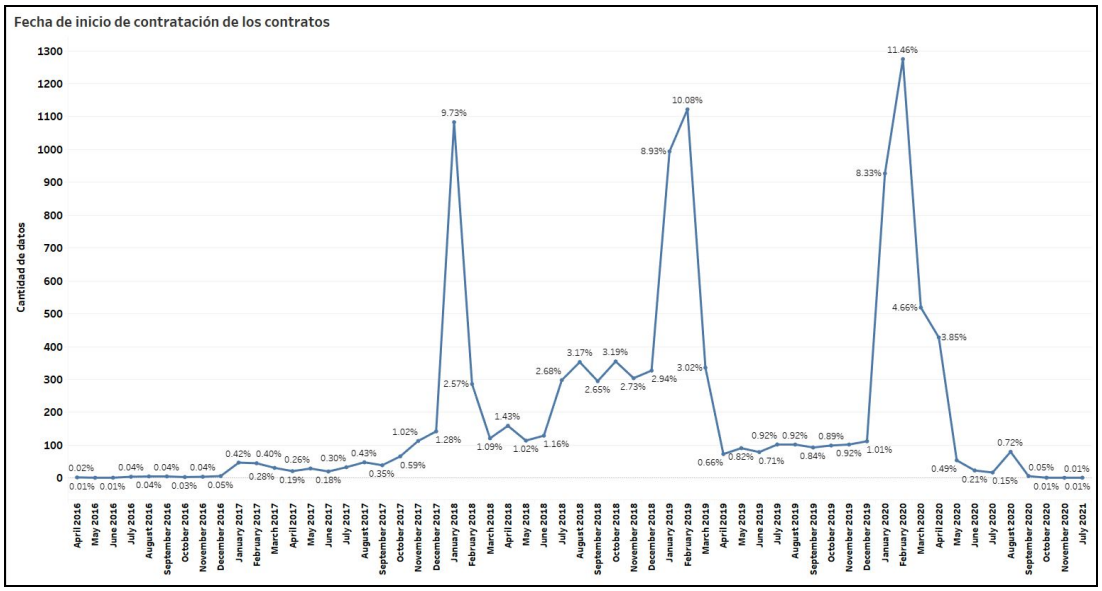
In the graph of General Suppliers, it can be observed that more than 80% of the data is concentrated in a supplier NOT defined. This may be due to the fact that when entering the information into the SECOP II database, this field is not mandatory and they omit it.

By removing the undefined data, you can see the graph 9, in which it is observed:

The top 5 governmental service suppliers, based on the number of contracts, are below. Each one has next to it the number of contracts signed historically:

- La Previsora S.A. Compañía de Seguros (18 contracts). Company specialized in all types of insurances which might be related to public car insurances, public building and infrastructure insurances and public employee's insurances.
- Caracol Televisión S.A (17 contracts). This national company is specialized in propaganda and paid TV and radio media. Most of the contracts are related to paid TV commercials and diffusion of relevant information to the citizens.
- Imprenta Nacional de Colombia (14 contracts). Its functions are to direct, edit, print, disseminate and market the Official Gazette, in accordance with current legal provisions. It must also edit and publish the Congress Gazette, the Judicial Gazette, the Constitutional Gazette and other publications of the Judicial Branch.
- Caja de Compensación Familiar Compensar (14 contracts). Its functions are to bring social wealth to institutional employees throughout health, pension, layoffs and other services valuable to the public institutional payroll.
- Servicios Postales Nacionales (13 contracts). It is the official postal company of Colombia, operating under its brand 4-72. It is aimed at offering all citizens a universal postal service. 4-72 has a wide portfolio of express, physical mail, electronic and virtual messaging services and Postal Payment Services.
- Brings the attention that the only person present in the top 20 providers is Álvaro Ramírez Echavarría. This person has earned 10 national contracts which some of his clients are the INPEC (National Penitentiary and Prison Institute) and the Hospital San Vicente de Paula in Colombia.

Contract start date

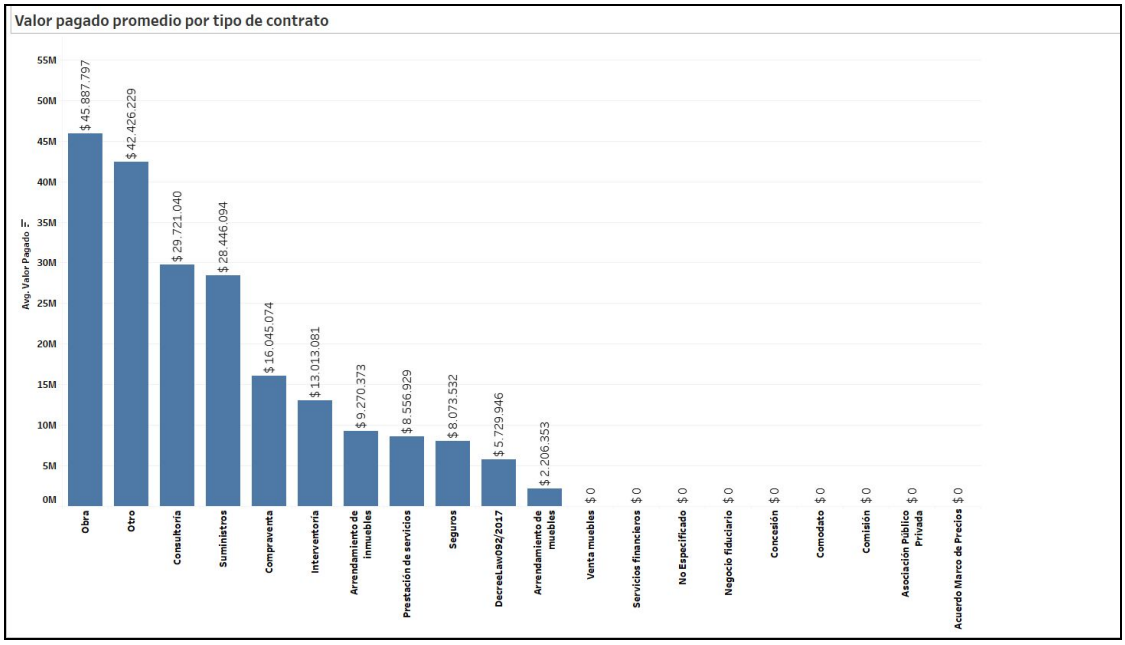


Graph 13. Contract Start date. Own source, based on [4]

As can be seen from the timeline graph above, the start of the procurement process is seasonal and presents peaks at the beginning of the 2018, 2019 and 2020 years. This is consistent with the budget allocation of the national government in January.

There are other peaks in August and October. To understand this seasonality, the analysis will have to include more variables like the kind of contract that usually starts in those months.

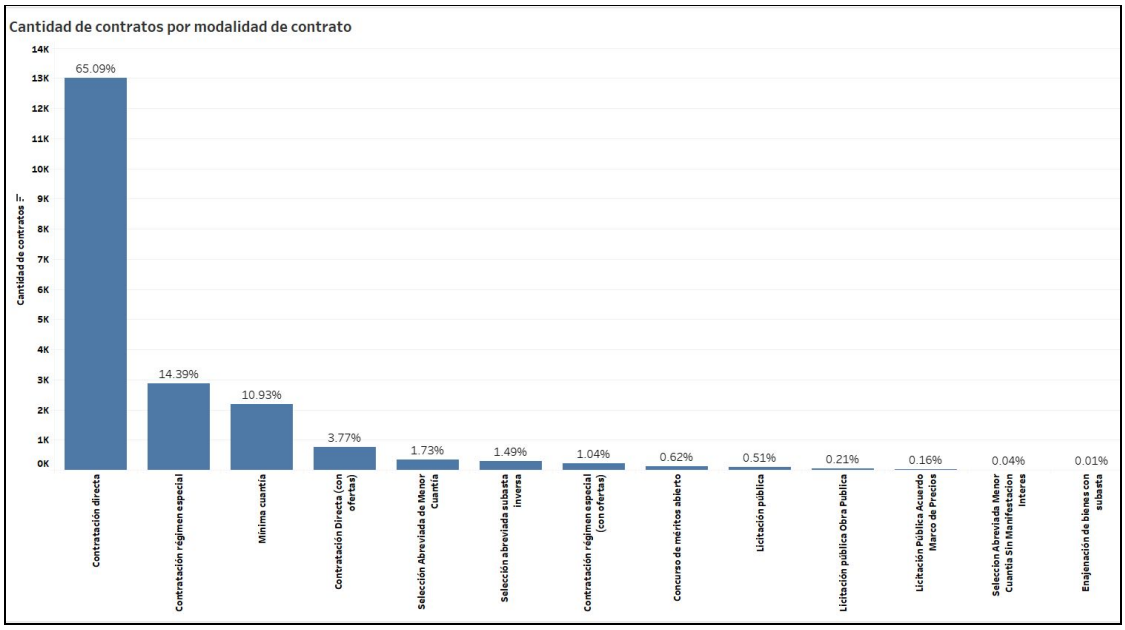
Average paid value by type of contract



Graph 14. Average value paid by type of contract. Own source, based on [4]

"Obra" is the contract type with the highest average paid value (45.9 M COP), followed by "Otro" (42.4 M COP), "Consultoría" (29.7 M COP), "Suministros" (28.4 M COP) and "Compraventa" (16 M COP).

Number of contracts by type of contract

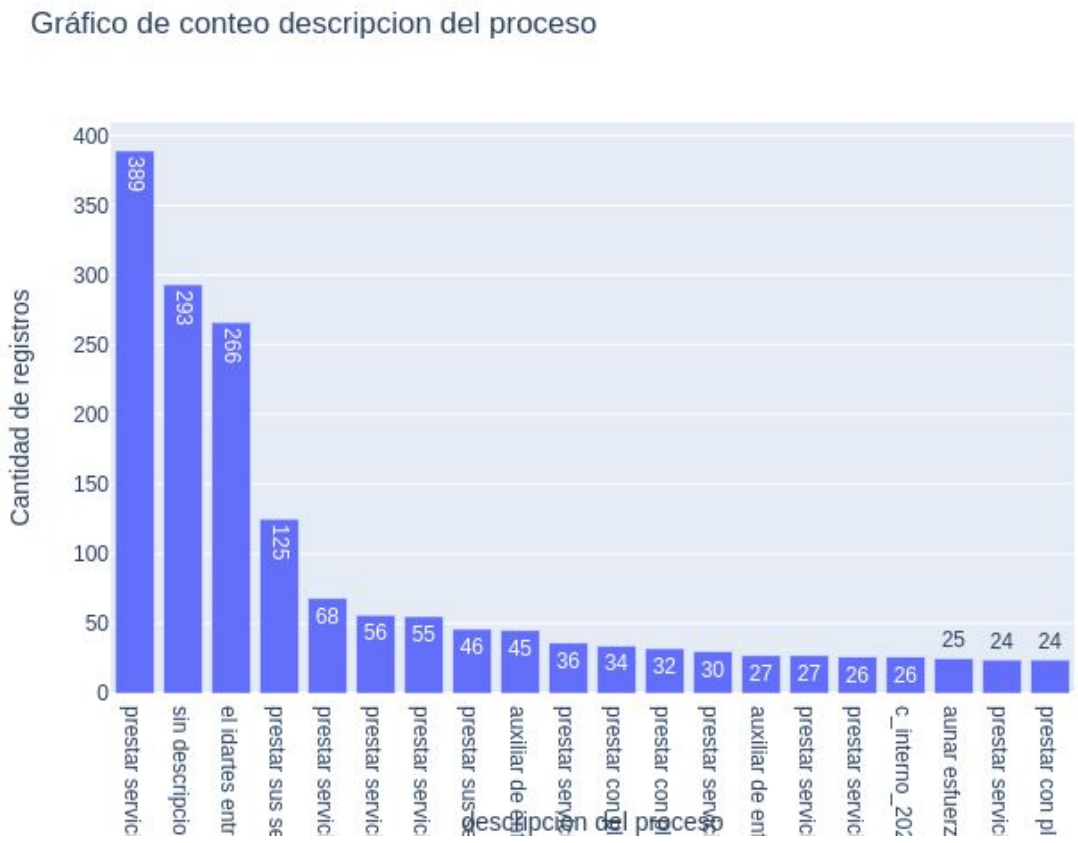


Graph 15. Number of contracts by type of contract. Own source, based on [4]

After the analysis of the type of contracts, around 80% of the records are distributed between "direct contracting" and "contracting special regime". Adding the “minimum amount” category the concentration raises to 90% of the records.

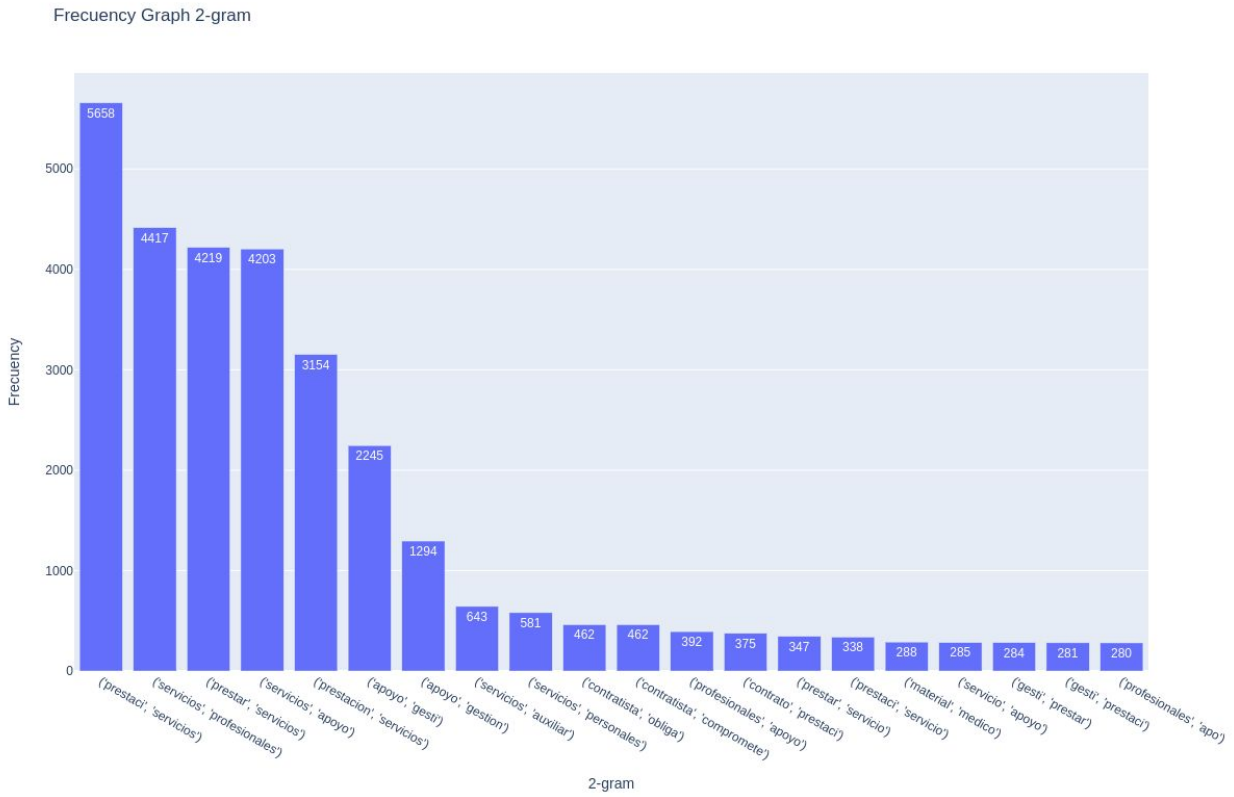
4.3. Advanced EDA over "Contractual Object"

SECOP I



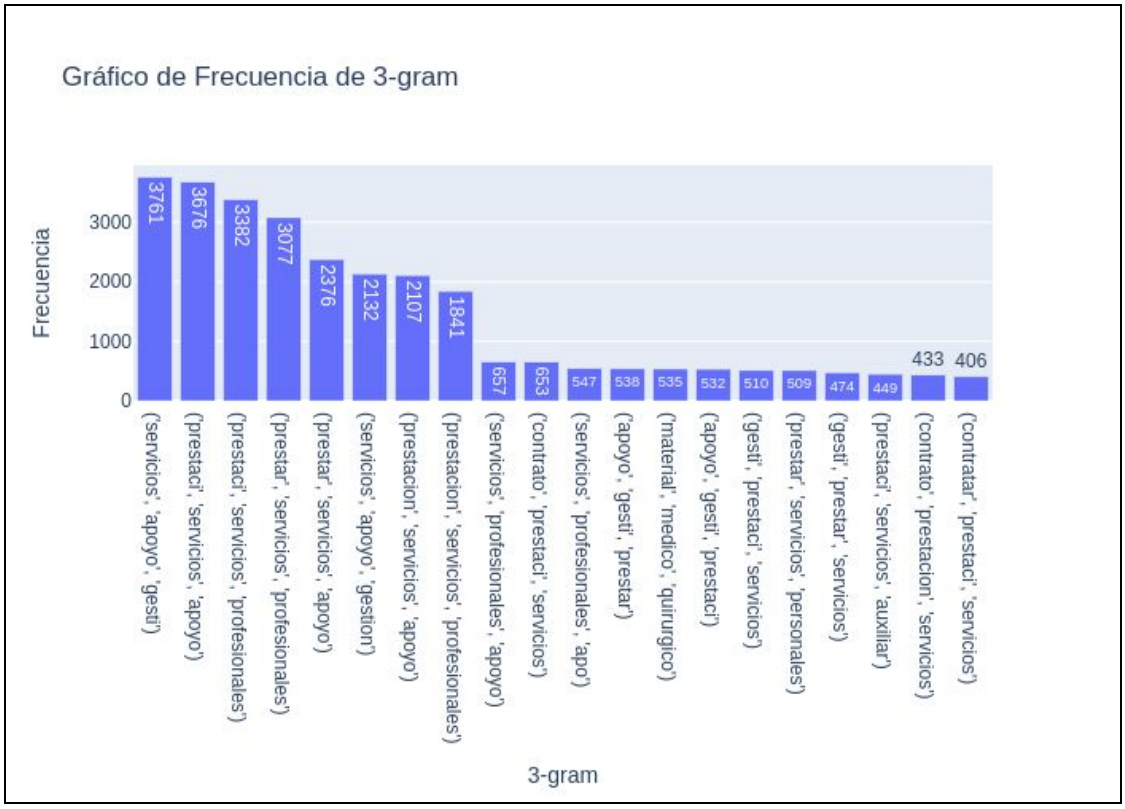
Graph 16. Top 20 words from the contractual object of SECOP I. Own source, based on [4]

Word Frequency of the SECOP I. In this graph the word service, support and present or lending are the ones that appear more frequent. This could mean that most of the contracts in the database are related to support or lend services to the entities that need them.



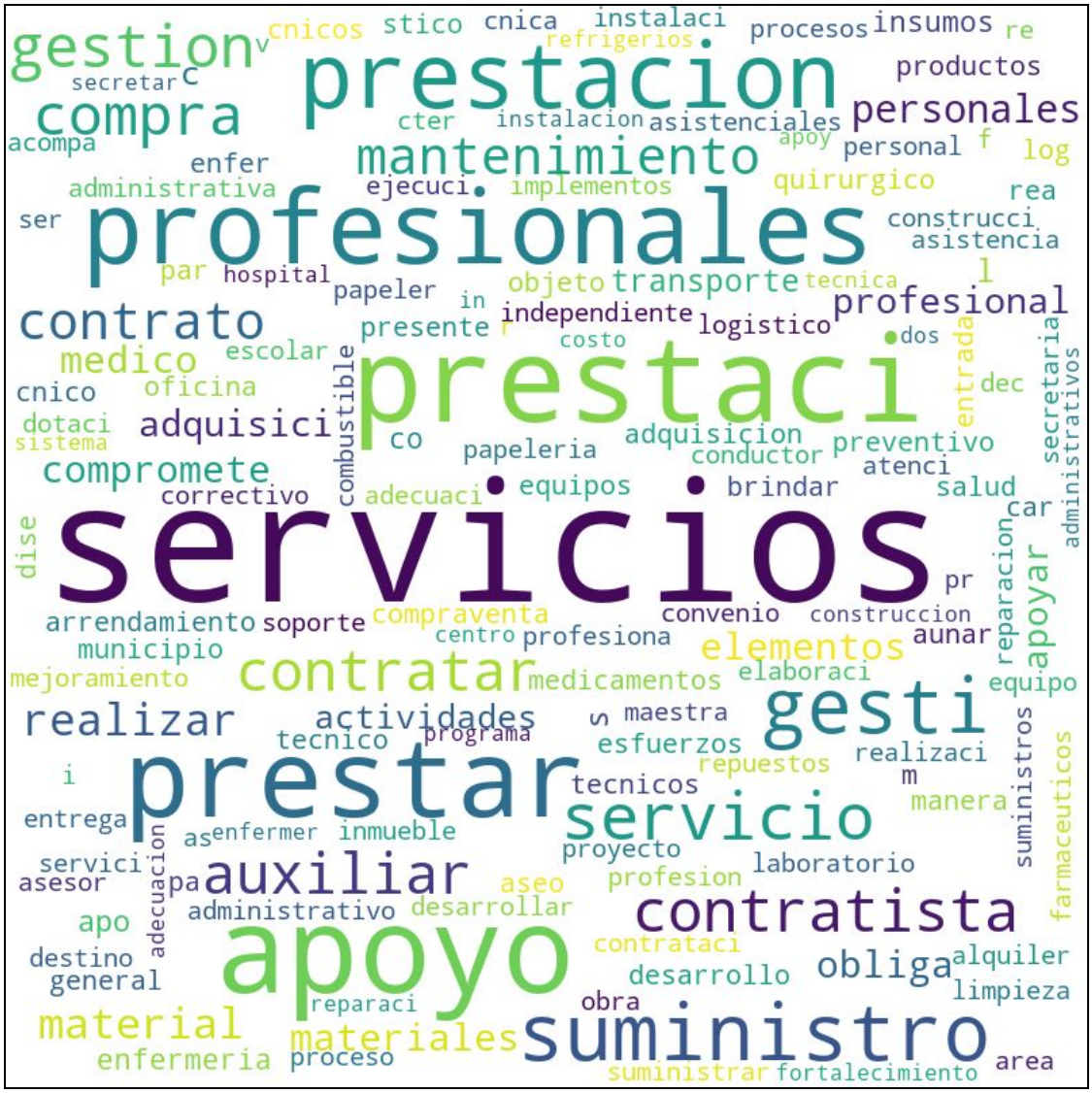
Graph 17. Top 20 bigrams from the contractual object of SECOP I. Own source, based on [4]

In this graph, we find that the bigrams “prestaci’, servicios”, “prestar’, servicios”, “prestacion, servicios”, “prestaci’, servicio” and “prestar’, servicio” are the most frequent. This may represent that most contracts are related to the provision of services. However, we should analyze the contractual object as a whole to achieve an adequate clustering of contracts.



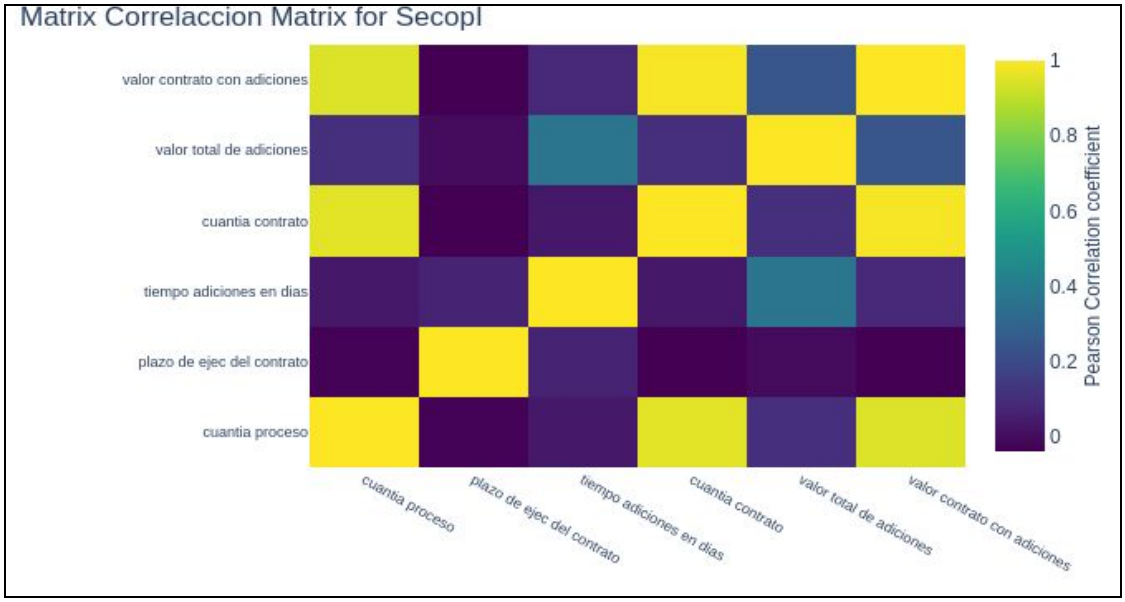
Graph 18. Top 20 trigrams from the contractual object of SECOP I. Own source, based on [4]

In this graph the trigrams that are shown are the combinations of the top 5 words that were at the bigrams and the words alone. This means that the contracts that have these words are related among them.



Graph 19. WordCloud from SECOP I. Own source, based on [4]

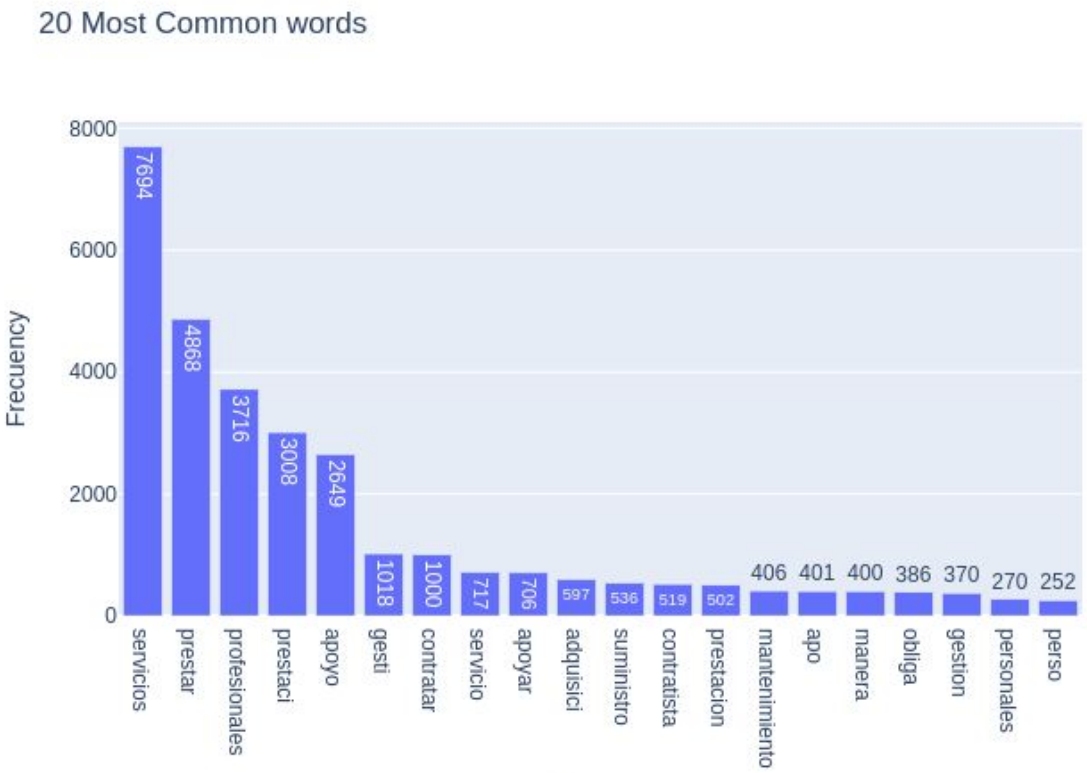
The graphs above show that the most common words are “servicios”, “apoyo” and “prestasi”. For bigrams the most common pairs of words are: “prestaci, servicios” and “servicios, profesionales”. Finally the most common trigrams in SECOP I are “servicios, apoyo gesti” and “prestaci, servicios, apoyo”. This suggests that it would be necessary to perform some specialized cleaning and manipulation methods over the words in order to group them better.



Graph 20. Correlation matrix of SECOP I. Own source, based on [4]

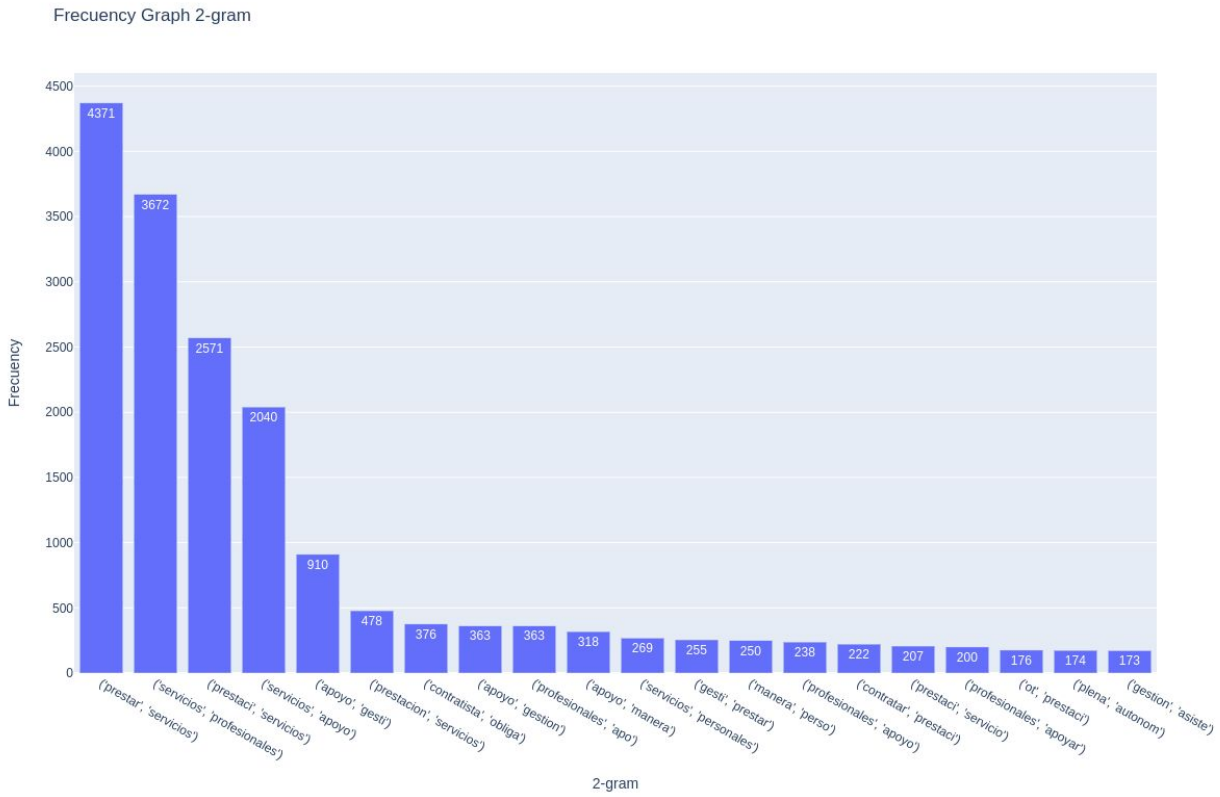
The correlation matrix shows natural dependency between the variables the total value pay and the value of the process as well as value of the extra fees with the contract extra time. It seems that the duration of the contract is independent from the other variables.

SECOP II



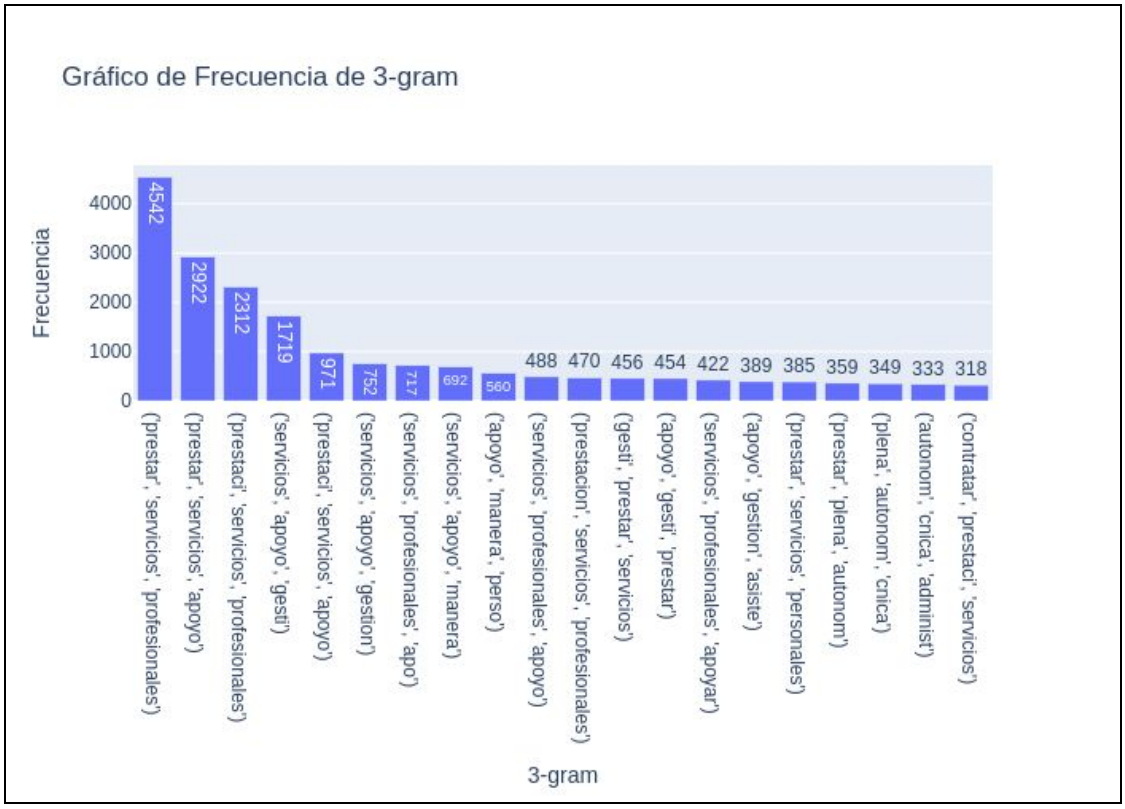
Graph 21. Top 20 words from the contractual object of SECOP II. Own source, based on [4]

The word frequency graph indicates that the words of the contractual objects of SECOP II (20), words such as “servicios”, “prestacion”, “profesionales”, “apoyo”, “gestion” are used. Therefore, it follows that the country has a deficit of labor contracts.



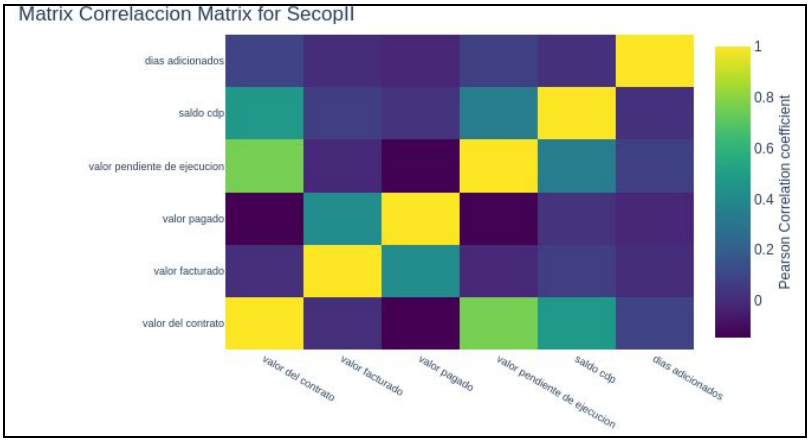
Graph 22. Top 20 bigrams from the contractual object of SECOP II. Own source, based on [4]

In the same way, the information added with the most mentioned variables is displayed through the biogram. The combination of the words that are most repeated are the ones that appear at top also in the bigram. However contracts related to management are less common rather than the ones related to lend services towards others.



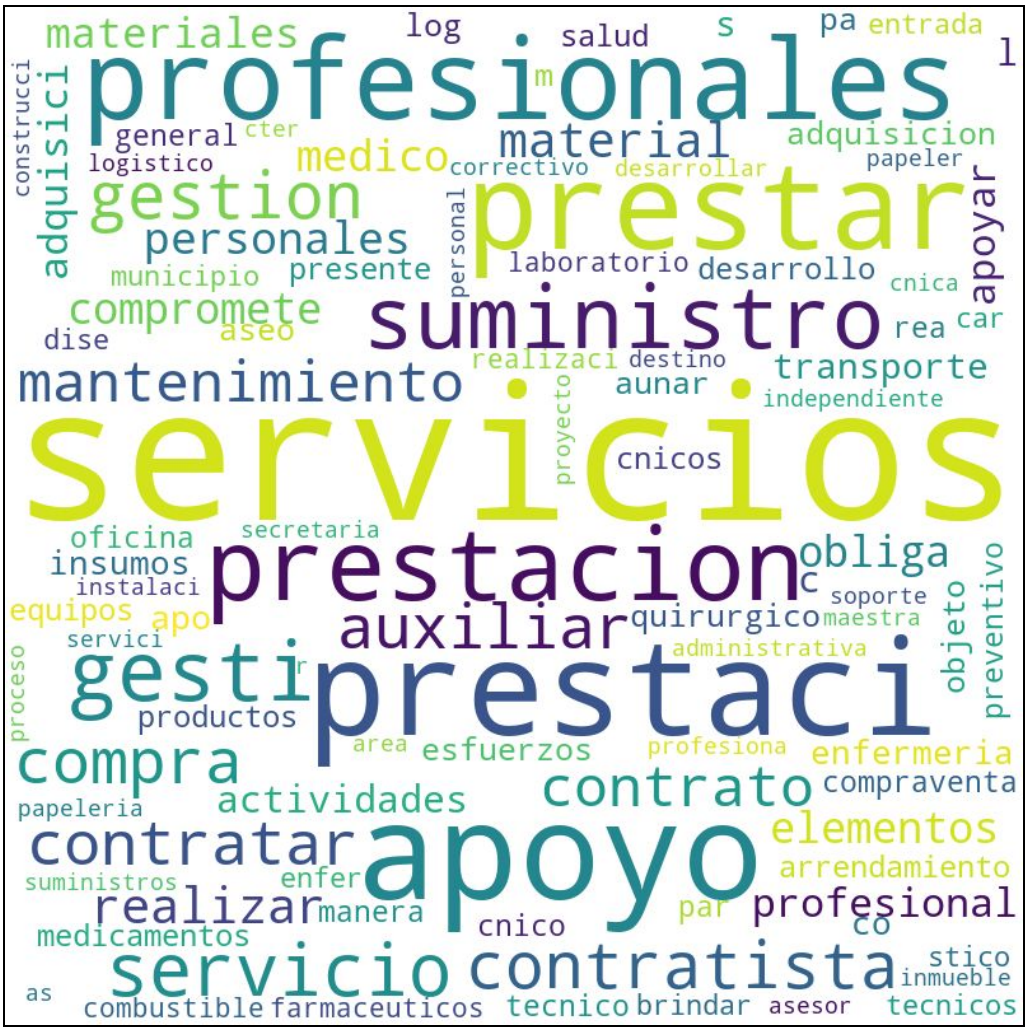
Graph 23. Top 20 trigrams from the contractual object of SECOP II. Own source, based on [4]

From the object contractual in SECOP II the most common words are “servicios”, “prestar” and “profesionales”. The most popular bigrams are “prestar, servicios”, “servicios, profesionales”, and “prestasi, servicios”. And finally the most common trigrams are “Presta, servicios, profesionales”, “prestar, servicios, apoyo” and “prestaci, servicios, profesionales”. As in the case of SECOP I, there is a need to perform cleaning and standardization of words to discard some typos as “prestaci” instead of “prestación”.



Graph 24. Correlation matrix of SECOP II. Own source, based on [4]

From the correlation matrix with SECOP II variables, we see a relationship between the “valor pendiente de ejecución” and “valor del contrato”. There is also an obvious relation between “valor pagado” and “valor facturado”.



Graph 25. Word Cloud from SECOP II. Own source, based on [4]

In the world cloud graph (Graph 25), it can be seen that the words that stand out the most are ‘service’, ‘provision’, ‘support’, which coincides with the results shown in the graph of Top 20 trigrams from the contractual object.

5. Data Cleaning [5]

Format

Change the format of the variables to their nature, this means that the columns that were loaded as strings were changed to numbers, if the value of a variable was a date it changes to have a specific format, in this case the format selected was Day/Month/Year.

Outliers

After detecting the nature of the variable, it follows an analysis to detect the outliers. In this case, the anomaly analysis was over the numeric variables. In the datasets there were values very different from the rest of the data. The outliers have humongous values that increase artificially the mean of contract values. These values can be explained by two possible reasons. First one, a bad imputation of the people to fill the contract or the second can be that it was a group of contracts that were meant to be separate.

In addition, in both dataset they had many registers that don't have any date or department. To have a consistent dataset, the option that was chosen was to impute these values, due to the fact that these values could not be attributed to a specific department or to a possible date.

Integration

There were two databases that were meant to be analysis. SECOP I and SECOP II. These dataset had different variables and with different names. In consequence, the modeling had many difficulties, because there was no point in comparison among them. Therefore, a consolidated dataset was created. This dataset had the common variables both dataset shared and the names of these variables were changed. This can be seen in the graph 23 that shows names in the two tables and in the final one.

Name	Name_DB_SECOP_I	Name_DB_SECOP_II
Nombre Entidad	NOMBRE_DE_LA_ENTIDAD	NOMBRE_ENTIDAD
Nit Entidad	NIT_DE_LA_ENTIDAD	NIT_ENTIDAD
Departamento	DEPARTAMENTO_ENTIDAD	DEPARTAMENTO
Ciudad	MUNICIPIO_ENTIDAD	CIUDAD
Orden	NIVEL_ENTIDAD	ORDEN
ID Contrato	NUMERO_DEL_CONTRATO	ID_CONTRATO
Estado Contrato	ESTADO_DEL_PROCESO	ESTADO_CONTRATO
Codigo de Categoria Principal	ID_OBJETO_A_CONTRATAR	CODIGO_DE_CATEGORIA_PRINCIPAL
Descripcion del Proceso	DETALLE_DEL_OBJETO_A_CONTRATAR	DESCRIPCION_DEL_PROCESO
Tipo de Contrato	TIPO_DE_CONTRATO	TIPO_DE_CONTRATO
Modalidad de Contratacion	REGIMEN_DE_CONTRATACION	MODALIDAD_DE_CONTRATACION
Fecha de Firma	FECHA_DE_FIRMA_DEL_CONTRATO	FECHA_DE_FIRMA
Fecha de Inicio de Ejecucion	FECHA_INI_EJEC_CONTRATO	FECHA_DE_INICIO_DE_EJECUCION
Fecha de Fin de Ejecucion	FECHA_FIN_EJEC_CONTRATO	FECHA_DE_FIN_DE_EJECUCION
TipoDocProveedor	TIPO_IDENTIFI_DEL_CONTRATISTA	TIPODOCPROVEEDOR
Documento Proveedor	IDENTIFICACION_DEL_CONTRATISTA	DOCUMENTO_PROVEEDOR
Proveedor Adjudicado	NOM_RAZ_SOCIAL_CONTRATISTA	PROVEEDOR_ADJUDICADO
Valor del Contrato	VALOR_CONTRATO_CON_ADICIONES	VALOR_DEL_CONTRATO
Origen de los Recursos	ORIGEN_DE_LOS_RECURSOS	ORIGEN_DE_LOS_RECURSOS
Días Adicionados	TIEMPO_ADICIONES_EN_DIAS	DIAS_ADICIONADOS

Graph 26. Configuration in the main table. Own source, based on [4]

Cleaning focus variable

Also there was a rigorous cleanness process in the detail of each contract, due to the fact that this was the main variable in the analysis. This process consists in changing all the special characters like ñ or vocals with accents to common characters that can be read with encoders like ANSI or UTF-8. After this process, all the words were lower for making them comparable between them. Finally, the words that had less than 4 characters and stopwords at spanish were imputed, this happened because these words do not generate value to the variable itself.

6. Statistical Models

6.1. Supervised methods

The nature of the problem does not have an independent variable to predict. The scope and the company requires to have a classification among the contracts that allows a classification between them. This means that the problem is more oriented to be an unsupervised modeling where the output variable is the cluster itself. In conclusion any supervised method such as linear regression or logistic regression is discarded.

6.2. Vectorized words

To vectorize the words to numbers it is used the bag of words approach. This approach consists in vectorizing the corpus of all the text. There are different ways to encode the text such as one-hot encoding, encoding frequency and TF-IDF. In these cases both of the models use the TF-IDF, because the time that it takes to encode all the words was less demanding than the other techniques, for example the one hot encoding technique consumes all the available memory of the computer and it didn't take into account the less frequent terms[6] .

6.3. No supervised methods - K means

Due to the fact that the data had more than 1 million registers almost all kinds of algorithms are going to have problems because of the number of data. In consequence different kinds of clustering algorithms are going to consume more memory than the capable and will not give a result without consuming all memory of the computer. Algorithms such as Agglomerative clusters [7] or DBSCAN that are based on distances and density of data are exhaustive algorithms that reach their limit around 20 thousand of data (12 gb of memory). Trying a batch approach or an approximation approach would not be enough to describe the data because the sample available represents less than the 2% of original dataset.

There is only one algorithm that allows the batch approach easily and that the results were relative fast due to the amount of information. This algorithm was the Kmeans, it has a derivative that is called MinibatchKmeans from the library Sklearn[6]. It allows us to work with big datasets and also be efficient in terms of computational resources and time. However, the K means algorithm has a hyperparameter that it is needed to define. The number of clusters.

There are two approaches to define the number of clusters, the first one and the traditional one is the elbow technique. This technique measures the distortion of the cluster to all the points. This is calculated by calculating the sum of distances squares from each point to the center of their respective cluster. The number of clusters is chosen when the slope of two points in the graph of distortion versus the number of clusters is different from the rest of the information. This technique is not recommended at all, because it relies a lot on the perception of the person and the business knowledge.

The second technique used to determine the cluster number is the silhouette measure. The silhouette analysis can be used to determine the number of clusters by calculating the distance between the points from a cluster to the center of another cluster. The range of this measure is from (-1,1), where the objective is to maximize this measure.

6.4. LDA

"In natural language processing, the latent Dirichlet allocation (LDA)[9] *is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is an example of a topic model and belongs to the machine learning toolbox and in a wider sense to the artificial intelligence toolbox."*[1]

In our context, the main stakeholder (*Colombia Compra Eficiente*) is concerned on the main clusters created by the interaction between words in contract description. As it is a text analytics clustering problem, we chose LDA because it uses machine learning algorithms to identify those clusters (or topics).

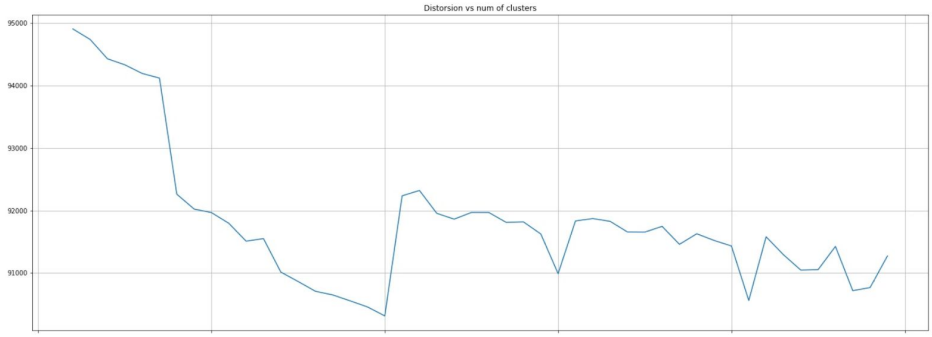
"In LDA, each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA". For our case, each document (in LDA notation) is the contract description and the hidden topics are the clusters we want to find.[8]

[1] Wikipedia contributors, "Latent Dirichlet allocation," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Latent_Dirichlet_allocation&oldid=976680107 (accessed October 25, 2020).

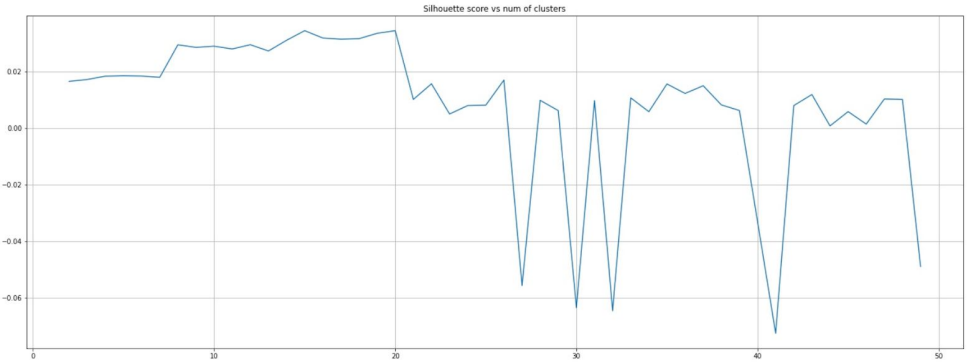
6.5. Model

6.5.1. K means

The results of the models can be divided in two, the first part the results using the K means algorithm and the second part using the LDA model.



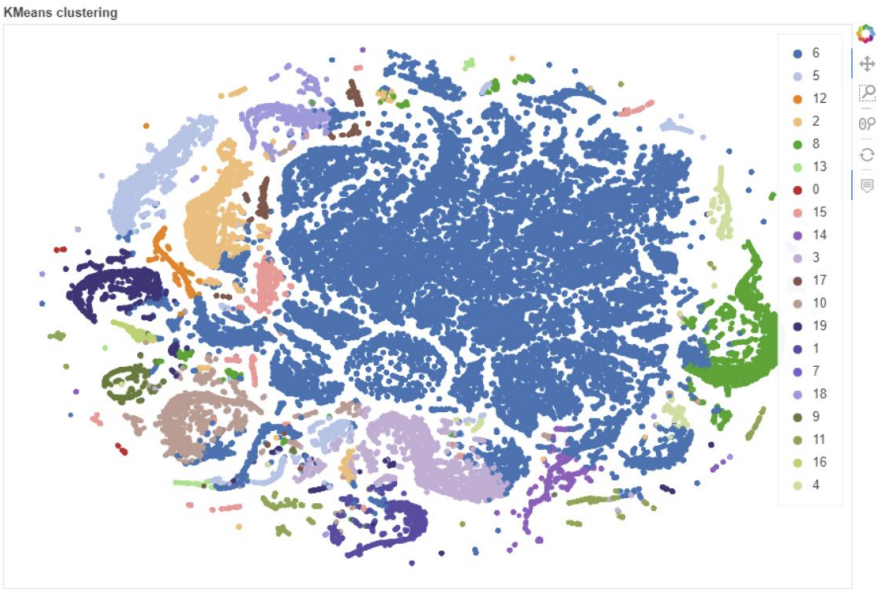
Graph 27. Distortion Graph. Own source, based on [4]



Graph 28. Silhouette score. Own source, based on [4]

At graph 27 the distortion of the graph shows the best number of clusters that minimizes the measure at 20 clusters, after this the distance among the clusters starts to increase drastically.

The nature of the problem doesn't allow the silhouette score to be higher, this occurs because the vector used makes the codify variables similar among them. So the scores shown in graph 28 are around to 0. However, the silhouette score that maximizes the number measure is 20 clusters. Because both measures coincide in the number of clusters. The optimal cluster is going to be 20.



Graph 29. TSNE Kmeans results. Own source, based on [4]

To graph the results of the clustering, it is needed to make a reduction of dimensionality. In this case, the technique to graph and have reduction of dimensionality is the TSNE. This technique uses an Student T distribution to graph all the non linear relationships the data have. This technique is recommended only to graph due to the complexity of itself. Because of this, the graph 29 only shows the distribution between the clusters among them.

Cluster	Percentage of dataset
6	65
5	9
8	6

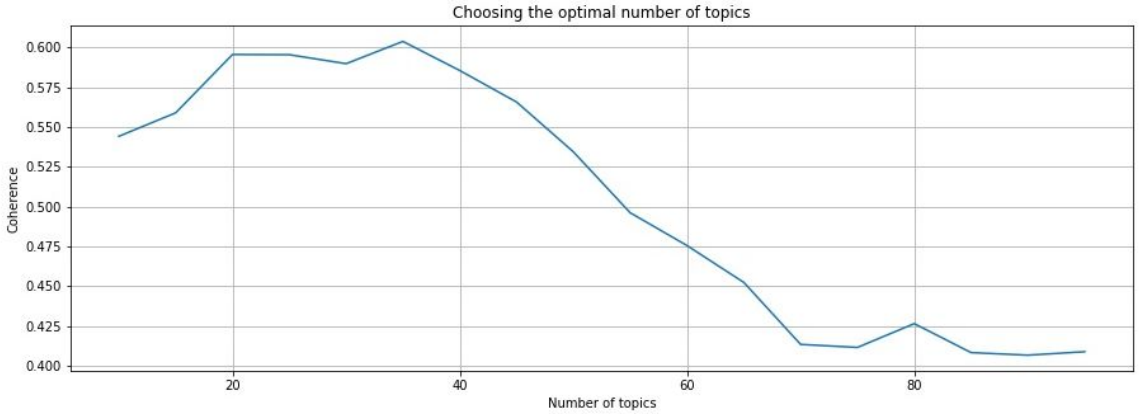
Graph 30. Top 3 Clusters. Own source, based on [4]

Graph 30 shows the top 3 clusters with the highest density. Cluster 6 is very dense, as it has 65 % of all the data classified in it. This cluster is about castral information, contracts related to evaluation of properties at municipal and departmental level. The cluster 5 that is the next most dense has a share of 9% of all the dataset, it contains contracts that are related to monitoring and surveillance in the superintendence of projects. Finally, cluster number 8 is the third with highest density with a share of 6% of the data, and it is related to all the contracts related to health and environment.

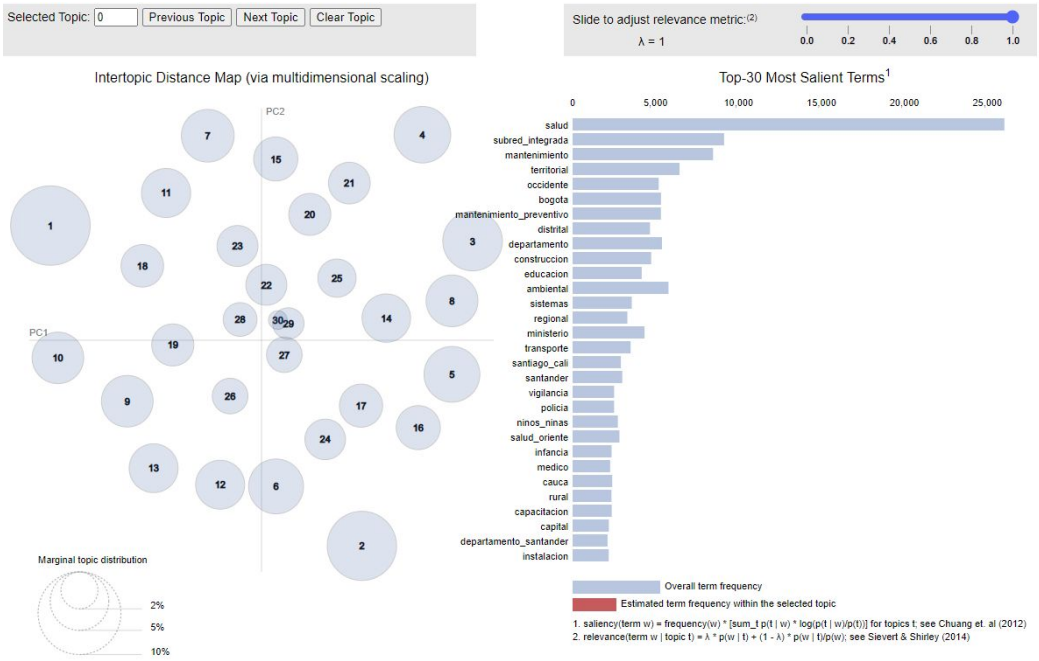
The results of the K means shows that the text data is hard to separate in unique groups that only share a similar word. In consequence, there are algorithms that are more sophisticated to allocate the common topics that the contracts may share among them.

6.5.2. LDA Model

At first, similar to the cluster algorithms it is needed to decide the number of clusters or in this case the number of topics to classify the contracts. To determine this, the measure used is called the coherence and the perplexity. Coherence measures how good is the model to classify new information that has not been seen before in the topics proposed and the perpexility is how the models classify information that have been seen before. The objective is to maximize the coherence and minimize the perplexity. In the graph 31, it is shown that the optimal number of around the 30 and the 40 topics.



Graph 31. LDA Coherence . Own source, based on [4]



Graph 32. LDA Results . Own source, based on [4]

To achieve coherent results, a manual iterative process should be done to clean all the terms that added no value to the information. A dictionary of over 1000 words was created to clean the contractual object in a way the LDA could cluster terms that at their own could explain the objective or the purpose of the contract (e.g. conflicto, educacion, recurso_hidrico).

The results of the LDA shows that the most common terms (without all the slag of the contracts) are related to health, maintenance and territorial contracts. These terms are the most common ones, however this doesn't mean that they are the most common ones for all the contracts. Each topic shown in the interface refers to a different kind of contract. The advantage of this model is that a contract can be classified in two or more topics rather than one like K means result.

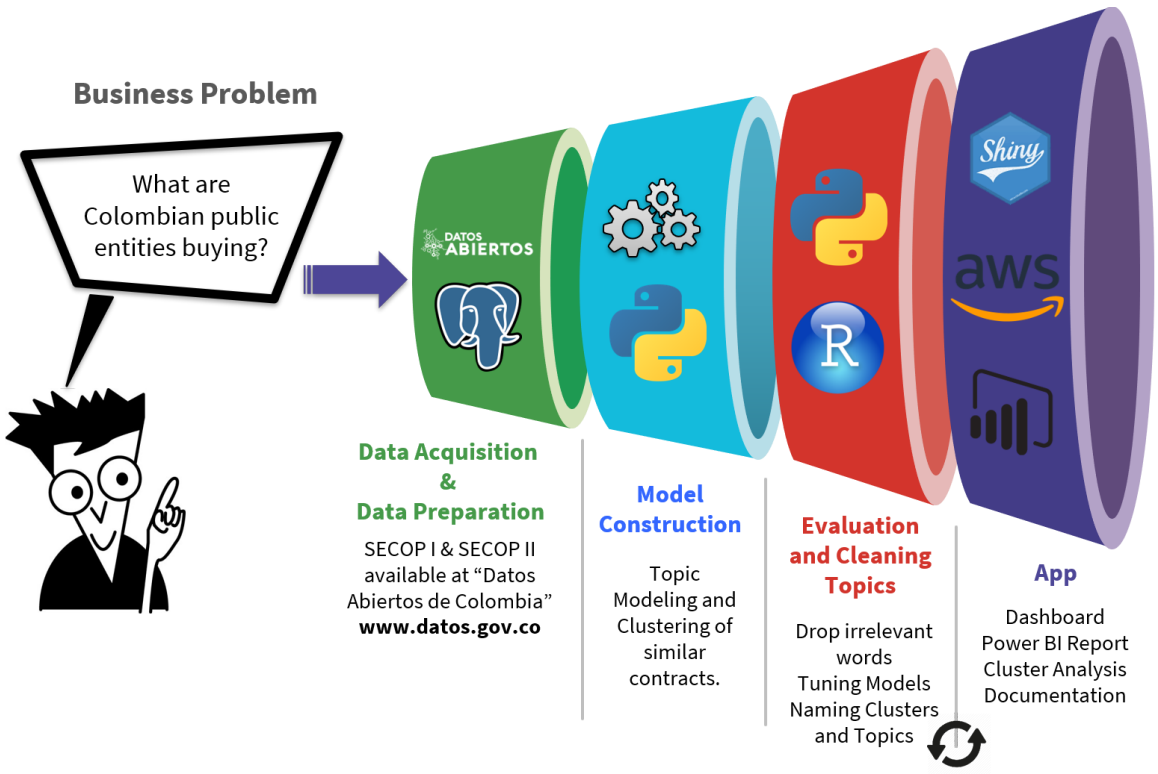
There is a clear variety of contracts that does not allow a homogeneous classification among them. The common trends in the topics proposed at the LDA are differences at the 5 first words in each cluster. However, the last words in each cluster are names from hospitals, territorial names such as departments and municipalities or educational centers. So, it is impossible to leave them out because they are names of the entities.

In addition, there are Topics highly defined in the following topics : Health, education, environment, construction.

7. Backend

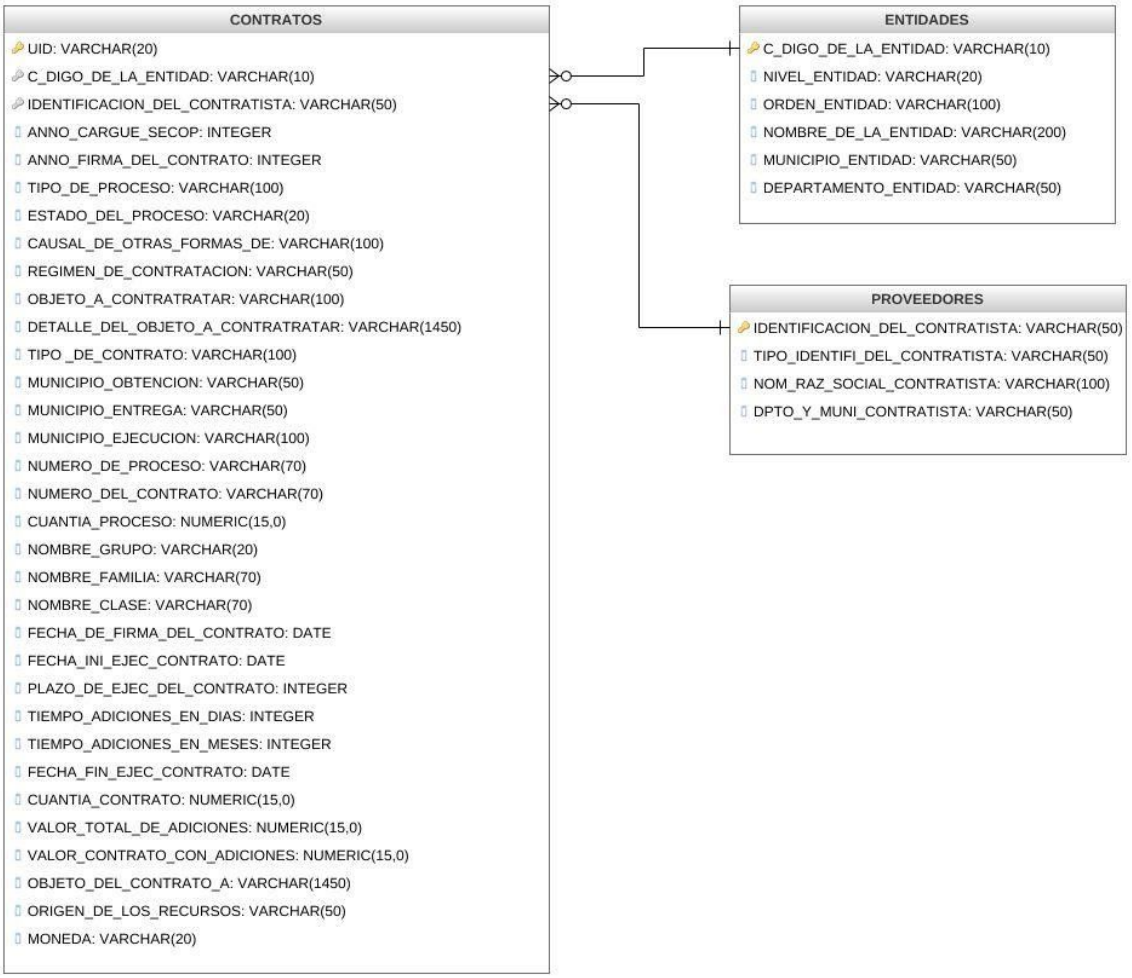
This is the data funnel that the project follows. It is divided by 4 steps, first the data acquisition and the data preparation. The raw data came from the recompilation of data from “Datos Abiertos”[4]. It is searched for information related to SECOP I and SECOP II , after having an extract of this data, it is required to clean the information. For this it runs a script that codifies the information, selects the information from both databases and consolidates into one. In this step it is used the techniques explained in section 6 [5].

The second step is to process the information to create the model. In this step the model takes the description of the contracts and vectorized it, after this it is run the clustering algorithms to classify the contracts depending on their object contract. Then, it visualizes the results from the algorithms. In the case of the K means it shows a TSNE scatter plot and in LDA it shows a graphic representation of it. After running the model it is needed to prove the quality of the results.For this it is run several times the selected model to improve the results in terms of impute words that do not generate value to the classification of contracts. Finally the information is visualized using a dashboard in Shiny, and Power BI.

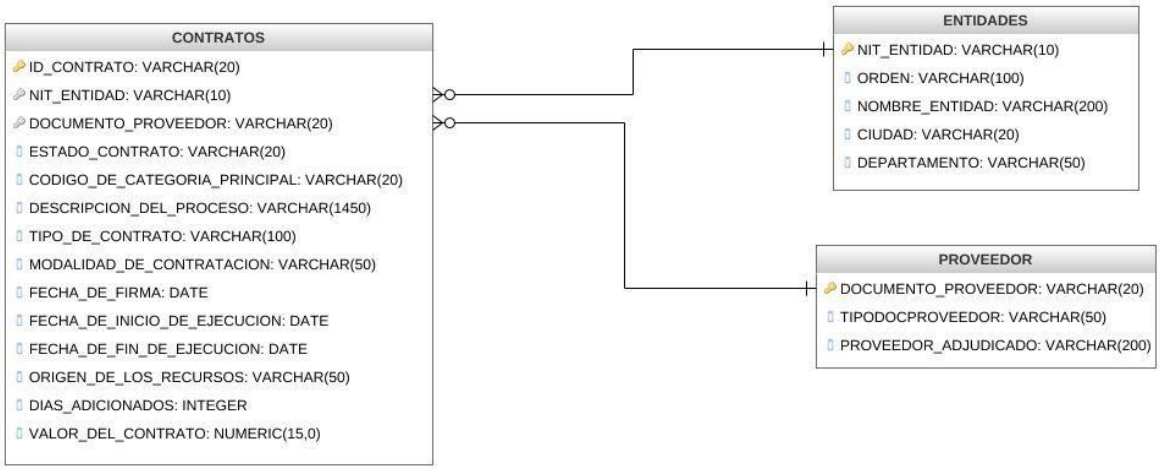


Graph 33. Backend Workflow. Own source

The diagram of the SECOP I and II databases consists of 3 tables: *Contratos*, *Entidades* and *Proveedores*. The *Contratos* table contains the detailed information of the contracts signed by the entities, whose information is found in the *Entidades* table. In the *Proveedores* table is the information of the suppliers to whom the contracts are awarded. Below is the diagram:



Graph 33. Diagram SECOP I. Own source, based on [4]

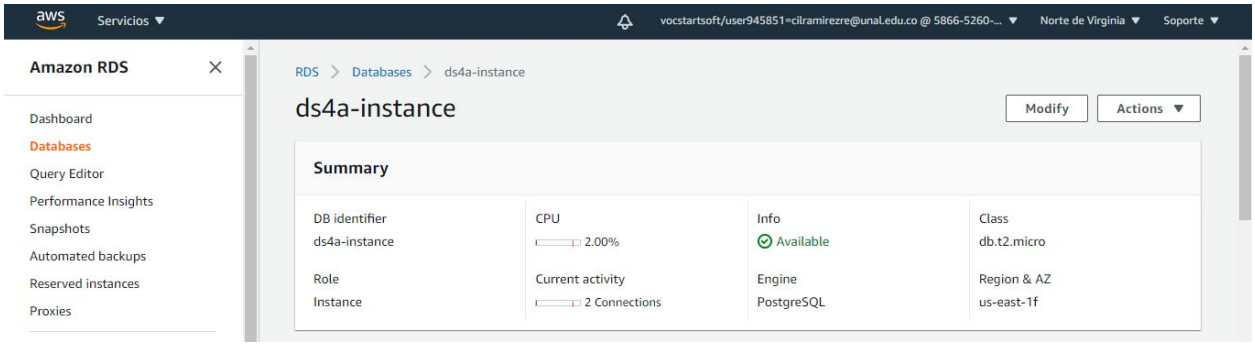


Graph 34. Diagram SECOP II. Own source, based on [4]

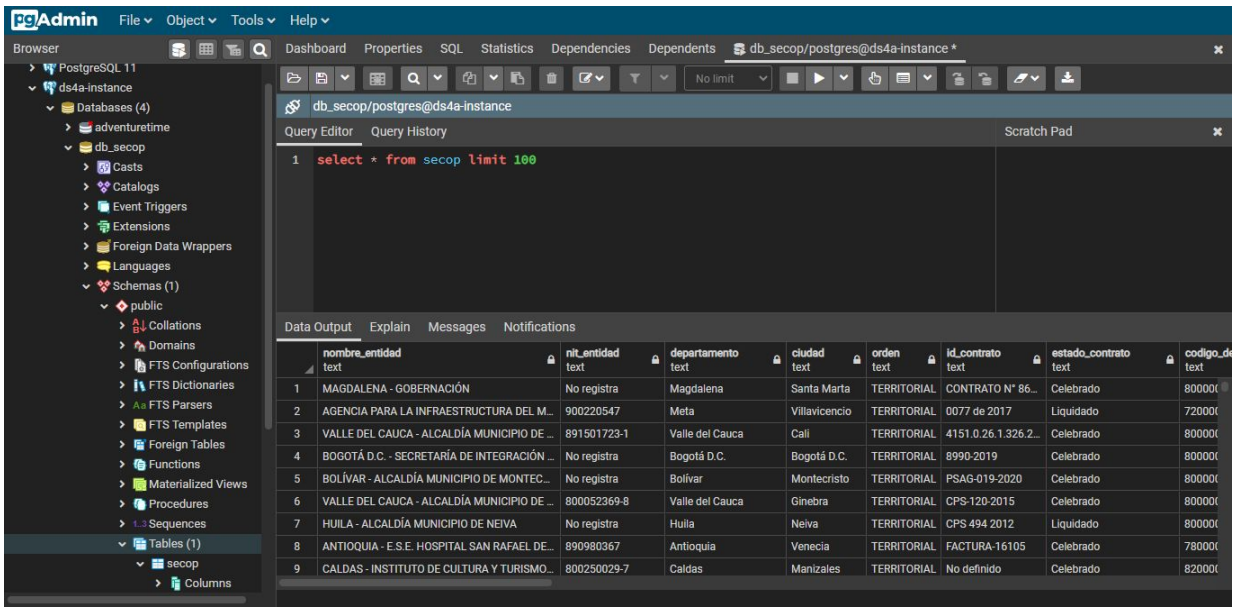
From the results of the EDA, the relevant fields were selected for the analysis, the cleaning was carried out, column names were matched. The *FUENTE* column was added to differentiate the data source (SECOP I or SECOP II) and the two databases were merged in the *SECOP* table. Graph 26 shows the table schema:

Graph 35. Diagram table SECOP . Own source, based on [4]

Finally, *ds4a-instance* was created in RDS[10][11] and the *db_secop* database where the information from the SECOP table was loaded as shown below (Graph 27 & 28):



Graph 36. Creation of ds4a-instance on RDS. Own source, based on https://aws.amazon.com/es/



	nombre_entidad text	nit_entidad text	departamento text	ciudad text	orden text	id_contrato text	estado_contrato text	codigo_de text
1	MAGDALENA - GOBERNACIÓN	No registra	Magdalena	Santa Marta	TERRITORIAL	CONTRATO N° 86...	Celebrado	80000(
2	AGENCIA PARA LA INFRAESTRUCTURA DEL M...	900220547	Meta	Villavicencio	TERRITORIAL	0077 de 2017	Liquidado	72000(
3	VALLE DEL CAUCA - ALCALDÍA MUNICIPIO DE ...	891501723-1	Valle del Cauca	Cali	TERRITORIAL	4151.0.26.1.326.2...	Celebrado	80000(
4	BOGOTÁ D.C. - SECRETARÍA DE INTEGRACIÓN ...	No registra	Bogotá D.C.	Bogotá D.C.	TERRITORIAL	8990-2019	Celebrado	80000(
5	BOLÍVAR - ALCALDÍA MUNICIPIO DE MONTEC...	No registra	Bolívar	Montecristo	TERRITORIAL	PSAG-019-2020	Celebrado	80000(
6	VALLE DEL CAUCA - ALCALDÍA MUNICIPIO DE ...	800052369-8	Valle del Cauca	Ginebra	TERRITORIAL	CPS-120-2015	Celebrado	80000(
7	HUILA - ALCALDÍA MUNICIPIO DE NEIVA	No registra	Huila	Neiva	TERRITORIAL	CPS 494 2012	Liquidado	80000(
8	ANTIOQUIA - E.S.E. HOSPITAL SAN RAFAEL DE...	890980367	Antioquia	Venecia	TERRITORIAL	FACTURA-16105	Celebrado	78000(
9	CALDAS - INSTITUTO DE CULTURA Y TURISMO...	800250029-7	Caldas	Manizales	TERRITORIAL	No definido	Celebrado	82000(

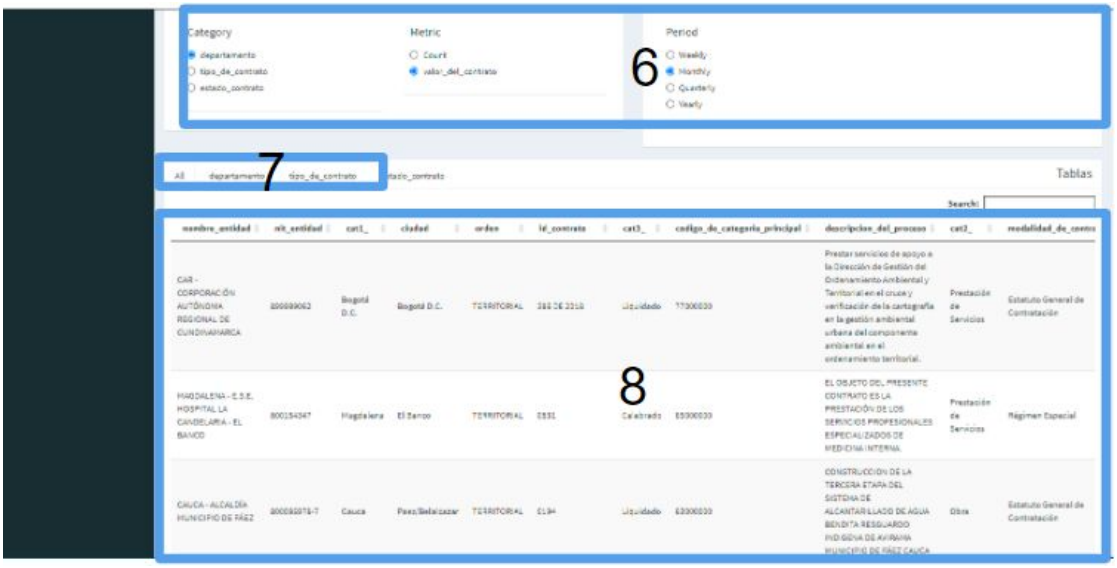
Graph 37. SECOP table on RDS. Own source, based on https://www.pgadmin.org/

8. Front End

8.1. Exploratory

Graph 38. Exploratory and Descriptive View 1. Own source, based on [4] and https://rstudio.com/

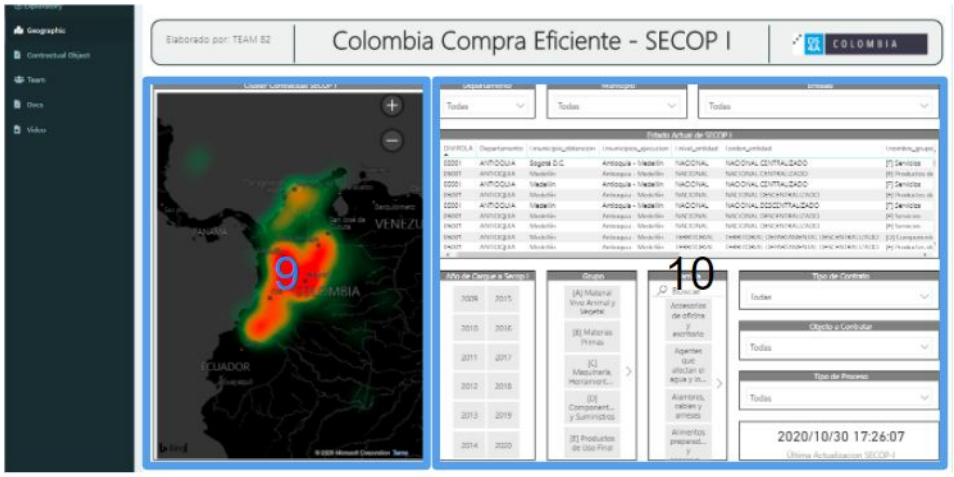
1. 5-tab menu:
    - Exploratory Analysis: Data analysis
    - Geographic: Geographical analysis
    - Contractual Object: text analytics to contractual objects
    - Team: Team contact and description
    - Docs: Project documentation
    - Video: Video summary
2. Filters to analyze the data
3. Main insights
4. Dynamic Pie chart
5. Amount spent dynamic bar chart



Graph 39. Exploratory and Descriptive View 2. Own source, based on [4] and https://rstudio.com/

6. Select boxes for pie chart and barchart
7. Multiple tables for analysis
8. Interactive table

8.2. Geographic

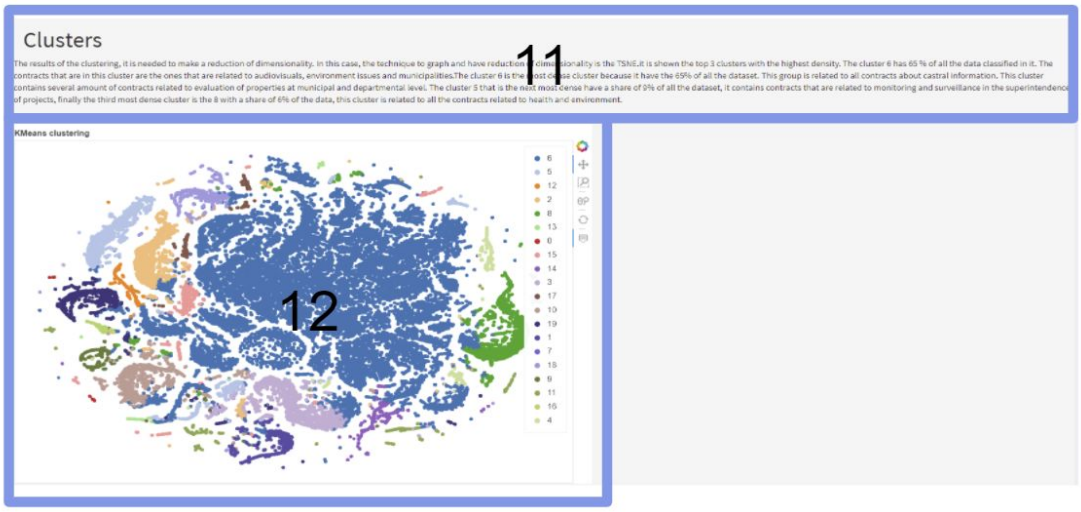


Graph 40. Geographic View. Own source, based on [4] and https://rstudio.com/

9. Interactive Map
10. Filters and other geographical insights

8.3. Word Analysis

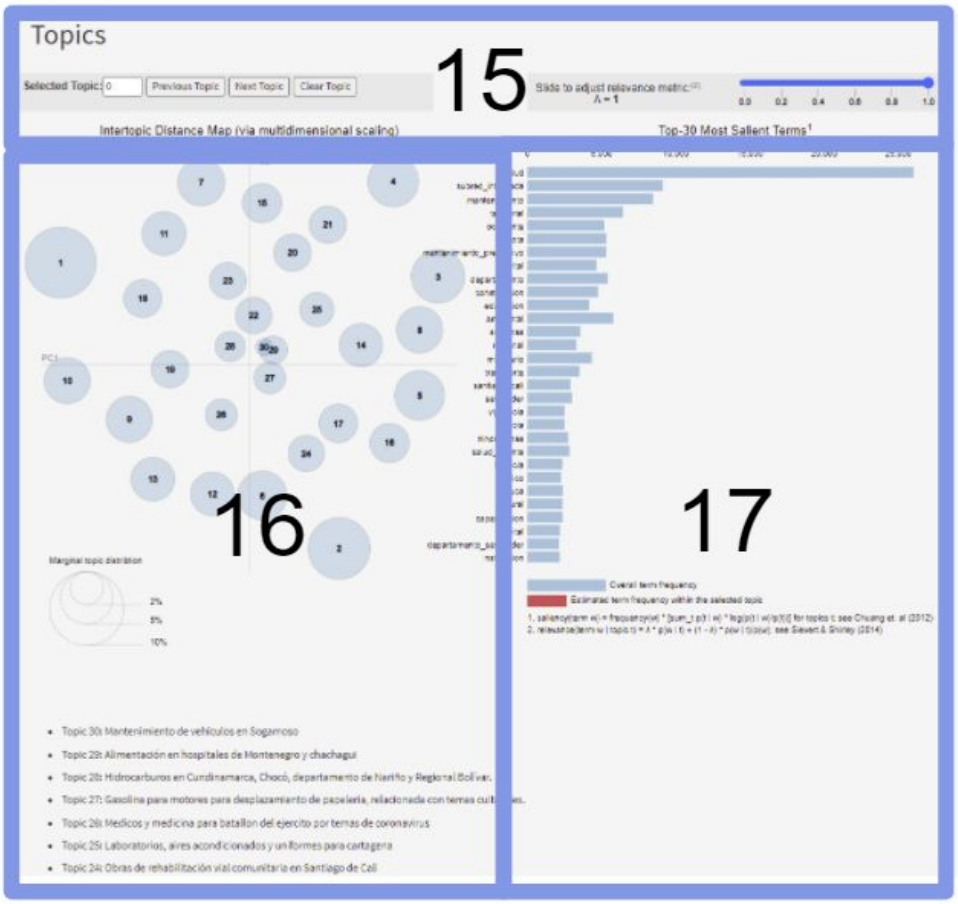
The section of the top will contain the 1 word clouds of the datasets. The two images that are located in this zone will be static images.



Graph 42. Clustering View 1. Own source, based on [4] and https://rstudio.com/

11. Clustering representation
12. Clustering analysis content

The section in the middle is going to have an analysis of the cluster and a graph with a transformation of tSNE that makes it easier to watch the data in two dimensions. This graph will also be static, due to the fact that tSNE graph is mainly used at visualizing data and doesn't give a data interpretation.



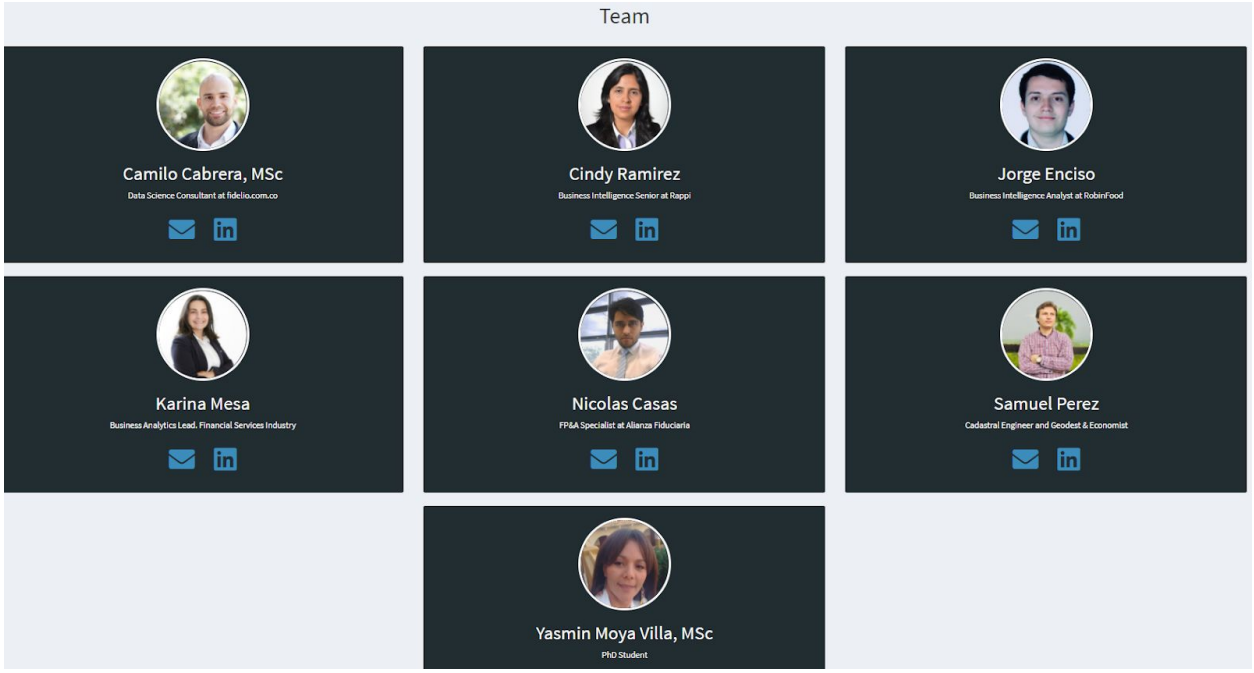
Graph 43. Clustering View 2. Own source, based on [4] and https://rstudio.com/

15. Latent Dirichlet Allocation Analysis: In this part is going to have an explanation about what is Latent Dirichlet Allocation (LDA). It will also have the reasons that it is used in this problem and the main insights found during the analysis.
16. Latent Dirichlet Allocation topic visualization: This bubble graph is a representation using PCA with two components where the size of the bubble represents the quantity of contracts associated with their cluster. This graph has 4 buttons, one text button that allows to search the main topic name of the cluster and three multiple drop list buttons that allows to filter the clusters in the graph.
17. Latent Dirichlet Allocation word distribution: At last this bar graph shows the most common terms according to their respective cluster, and the relevance according to the hyperparameters chosen. At the top it has a list with the lambda parameter. This parameter means the overall fitting of the words with each cluster.

In this section there will be 3 parts.

The second part of the section is going to be the LDA of the interactive graph. This graph is allowed to watch the concentration of words according to their respective clusters. It will have the button that selects the cluster.

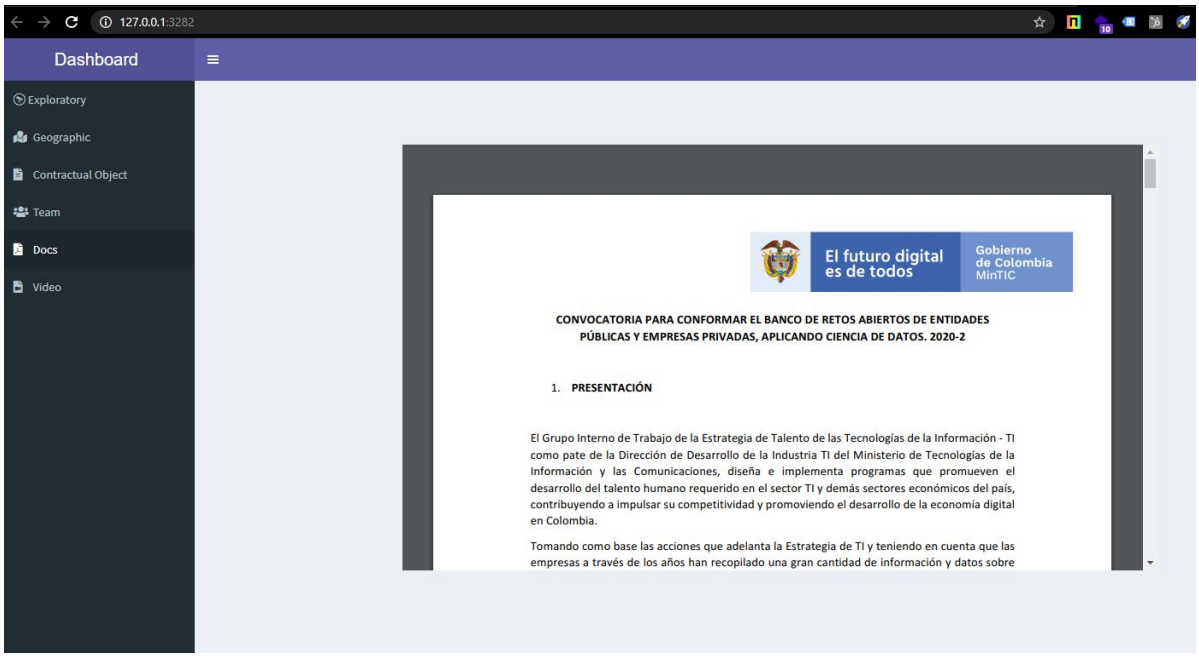
8.4. Team



Graph 44. Developers View. Own source, based on [4] and https://rstudio.com/

In this section going to have the information of the people that help to build this project. It will have the contact email and their link to LinkedIn.

8.5. Docs



Graph 45. Documentation View. Own source, based on [4] and https://rstudio.com/

In this section it is going to be located in the final document of the project. It will also include the link to the GitHub repository where all the files used in the project are going to be located.

8.6. Video

Graph 46. Video. Own source, based on [4], https://rstudio.com/ and https://www.veed.io/

In this section is going to be the presentation video used to explain the project.

9. Dashboard

To explore the front-end proposed for the final deliverable visit this link:

http://app.camicabrera.com:3838/contract-grouping/

The final version is going to have a personalized domain related to Colombia Compra Eficiente. The final version is going to include a QR code to access the site using the camera.

10. Conclusions

Machine Learning is an extension to help people make decisions. It allows the people to view in a more analytical, how to manage a problem using data. Nowadays with the rise of big data, the decisions need to be supported with data. The decisions taken need to be taken data driven rather than just business knowledge. For example in this process around two people and a day of complete work are needed to classify and obtain the analysis of the contractual objects . Only the extraction and the cleaning of them is exhaustive and as humans mistakes are made. Using machine learning to help the process the time of extraction, cleaning and classification can be saved and invested to improve the analysis made and obtain more value for this.

However, ML is not always the solution, a person has the perception that with machine learning their life is going to be solved. The algorithms will try to solve problems but it is almost impossible to obtain an accuracy of 100%. In text analytics is more difficult to obtain good results, because first most of the algorithms are meant to be used in english, and the documentation or APIs that works with spanish texts are barely seen or not optimized to work with Big Data, and second is the way people write and express their feelings and objective. People like to decorate their texts, they use more sophisticated words to explain what they meant. For a human it is easy to recognize these patterns and to extract the main idea of a text. Nevertheless, the machine can not do this easily. There are themes such as lemmanitazion, steaming, context , sarcasm that doesn't allow the machine to extract the real meaning of it. In consequence the process of cleaning the texts of all the words that don't mean nothing is an essential step to work with text data. Without this step, the ML algorithm will dirt information and will not obtain the desired results.

11. Recommendations and Next Steps

11.1. In-Database Algorithms

Every time the dashboard runs, it loads a local '.RDS' (R Object file) containing 100,000 random contracts read from a PostgreSQL database on an AWS server. The next step with databases is to take the command controls from the panels and connect them directly to the database to reduce load time.

This approach opens up the opportunity to run an integrated data pipeline where data is uploaded to AWS Postgres incrementally using the ETL code, but also looking for ways to optimize compute power or implement algorithms in databases using libraries like Apache MadLib[9]. Apache MadLib offers a set of supervised and unsupervised machine learning algorithms, including the implementation of the LDA topic modeling algorithm.

11.2. On Premises Deploy

For replicability and improvement purposes the solution resides in Github. If decided by CCE, the deployment of the solution could be done on premises using CCE infrastructure and GNU software (Python, R and Postgres mainly).

11.3. Semi-Automatically Categorize

There are several main categories already identified by the Colombia Compra Eficiente (CCE) team. Some of them involve "health", "construction", "infrastructure" and "technology" among many others. The next step would be to list the most relevant categories and semi-automatically categorize all the contracts based on the keywords in their description. For example the word "hospital" can be matched within the Health category. After processing all the categories, Clustering and LDA topic modelling analysis can be re-run into each category to analyze their main topics and take decisions regarding national budget spending.

12. References

[1] European Commision (September 9, 2020). Public Procurement. Recovered from https://ec.europa.eu/growth/single-market/public-procurement_en

[2] OECD (September 9, 2020). Public Procurement. Recovered from https://www.oecd.org/gov/public-procurement/

[3] Colombia Compra Eficiente (September 13, 2020). Recovered from https://www.colombiacompra.gov.co/

[4] Datos abiertos (September 13, 2020). Recovered from https://www.datos.gov.co/Gastos-Gubernamentales/SECOP-II-Procesos-de-Contrataci-n/p6dx-8zbt

[5] Selecting the number of clusters with silhouette analysis on KMeans clustering. (September 13, 2020) Recovered from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

[6] MiniBatchKMeans. (September 13, 2020). Recovered from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html?highlight=minibatch#sklearn.cluster.MiniBatchKMeans

[7] Agglomerative clusters. (September 19, 2020). Recovered from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[8] Topic Modeling with Gensim (Python) (September 19, 2020) Recovered from www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

[9] Latent Dirichlet Allocation implementation on Apache MadLib for Postgres (October 2, 2020) https://madlib.apache.org/docs/latest/group__grp__lda.html

[10] Amazon Relational Database Service (October 2, 2020). Configuración de Amazon RDS. Recovered from https://docs.aws.amazon.com/es_es/AmazonRDS/latest/UserGuide/CHAP_SettingUp.html

[11] DS4A Colombia 3.0 (September 11, 2020). Week 2 - AWS-Bootcamp and RDS-Bootcam

13. Appendix A

Table 1. SECOP I

Name	Description	Type	Cleaning Method	Relevant
UID	Compound value to individually identify each record	Text	NA	Yes
Anno Cargue SECOP	Year in which the process was registered on the platform	Number	Cast to number	No
Anno Firma del Contrato	If it is a signed contract, the date this signature was made	Text	Convert to integer and replace 0 with NaN	Yes
Nivel Entidad	Determines the first degree of characterization of the entity according to its order: National or Territorial	Text	NA	Yes
Orden Entidad	It details the order of the Entity, defining the type of National or Territorial Entity according to its degree of centralization	Text	Treat the null values	Yes
Nombre de la Entidad	Name of the State Entity to which the process corresponds	Text	NA	Yes
NIT de la Entidad	NIT of the Entity, as registered on the platform	Text	Treat the null values	Yes
Código de la Entidad	Entity Code, used as a unique identifier on the SECOP I platform	Text	NA	Yes
ID Tipo de Proceso	The ID and Type of Process describe the modality through which the purchase process was developed	Text	Convert them into key dictionaries, it can work if the data is decodified.	No
Tipo de Proceso	The Type of Process describes the modality through which the purchase process was developed	Text	NA	Yes
Estado del Proceso	The status of the process as of the publication date	Text	Group the values that have less than	Yes

			1%	
Causal de Otras Formas de Contratacion Directa	In the case of being a process developed under the direct contracting modality, this field describes the cause by which this type of contracting is determined	Text	Group the values that have less than 1%	Yes
ID Regimen de Contratacion	ID of the regime under which the entity develops the public purchase process	Text	Convert them into key dictionaries, it can work if the data is decodify.	No
Regimen de Contratacion	Description of the regime under which the entity develops the public purchasing process	Text	NA	Yes
ID Objeto a Contratar	Contract Object ID, based on the UNSPSC catalog of goods and services, viewable from https://www.colombiacompra.gov.co/clasificador-de-bienes-y-servicios	Text	Convert them into key dictionaries, it can work if the data is decodify.	No
Objeto a Contratar	Description of the Purpose of the contract, based on the UNSPSC catalog of goods and services, available from https://www.colombiacompra.gov.co/clasificador-de-bienes-y-servicios	Text	NA	Yes
Detalle del Objeto a Contratar	Detailed definition of the good or service to be acquired within the process.	Text	Change format all lower or upper, but not mix.	Yes
Tipo de Contrato	Type of contract to be carried out	Text	NA	Yes
Municipio Obtencion	Municipality in which the public purchasing process takes place	Text	DANE (National Department of Statistics) coding matching	Yes
Municipio Entrega	Municipality in which the good or service is delivered	Text	DANE (National Department of Statistics) coding matching	Yes

Municipios Ejecucion	Municipalities in which the object of the public purchase process will be developed	Text	DANE (National Department of Statistics) coding matching	Yes
Fecha de Cargue en el SECOP	Date on which the registration was made on the platform	Date	NA	No
Numero de Constancia	Identifier of the purchase process, generated by SECOP I	Text	High Cardinality	No
Numero de Proceso	Process identifier, according to the entity's nomenclature	Text	Change codification to a uniform one. If not a number create a category	Yes
Numero del Contrato	Contract identifier, according to the entity's nomenclature	Text	Standardize codification	Yes
Cuantia Proceso	In addition to the code that defines the object of the contract, a detail of the definition of the good or service to be acquired within the process is recorded.	Number	Cast to number, take care with the 0	Yes
ID Grupo	Initial categorization of the good or service defined in the purchase process, according to its main characteristics	Text	Convert them into key dictionaries, it can work if the data is decodify.	No
Nombre Grupo	Initial categorization of the good or service defined in the purchase process, according to its main characteristics	Text	Treat the null values	Yes
ID Familia	Second level of detail within the characterization of the good or service	Text	Convert them into key dictionaries, it can work if the data is decodify.	No
Nombre Familia	Second level of detail within the characterization of the good or service	Text	Treat the null values	Yes
ID Clase	Third level of detail within the characterization of the good or	Text	Convert them into key dictionaries, it	No

	service		can work if the data is decodify.	
Nombre Clase	Third level of detail within the characterization of the good or service	Text	Treat the null values	Yes
ID Ajudicacion	Identifier of the award or awards made in the purchase process	Text	Convert them into key dictionaries, it can work if the data is decodify.	No
Tipo Identifi del Contratista	Type of Identification of the contractor selected in the award	Text	NA	Yes
Identificacion del Contratista	Identification of the contractor selected in the award	Text	NA	Yes
Nom Raz Social Contratista	Name or Company Name of the contractor selected in the award	Text	NA	Yes
Dpto y Muni Contratista	Department and Municipality in which the contractor selected in the award operates	Text	NA	Yes
Tipo Doc Representante Legal	In case of being a company, the type of identification of the legal representative of the company selected in the award	Text	NA	No
Identific del Represen Legal	In case of being a company, identification of the legal representative of the company selected in the award	Text	Most values are not defined or an id	No
Nombre del Represen Legal	In case of being a company, Name of the legal representative of the company selected in the award	Text	NA	No
Fecha de Firma del Contrato	Date on which the contract corresponding to the award of the registry is signed	Date	Cast to date	Yes
Fecha Ini Ejec Contrato	Date on which the execution of the contract corresponding to the award of the registry begins	Date	Cast to date	Yes

Plazo de Ejec del Contrato	Value and unit in which the execution time of the contract is measured, be it days or months	Date	Cast to number and convert to days in order to have only one unit of time.	Yes
Rango de Ejec del Contrato	Value and unit in which the execution time of the contract is measured, be it days or months	Text	NA	No
Tiempo Adiciones en Dias	Contract extension, outside the initial definition, in days	Number	Cast to number	Yes
Tiempo Adiciones en Meses	Contract extension, outside the initial definition, in days	Number	Cast to number	Yes
Fecha Fin Ejec Contrato	Completion date of contract performance	Date	Cast to date	Yes
Compromiso Presupuestal	In case of having a budget record, the field shows the corresponding code	Text	Most not define (94%)	No
Cuantia Contrato	Value for which the contract is signed	Number	Cast to number, treat null values	Yes
Valor Total de Adiciones	Value of the sum of the additions made to the contract	Number	Cast to number, treat null values	Yes
Valor Contrato con Adiciones	Total value of the contract, including additions	Number	Cast to number, treat null values	Yes
Objeto del Contrato a la Firma	Procurement's topic, registered until the moment of signed	Text	Treat the null values	Yes
ID Origen de los Recursos	Identifier of the way in which the resources with which the contract will be paid are obtained	Text	Convert them into key dictionaries, it can work if the data is decodify.	No
Origen de los Recursos	The way in which the resources with which the contract will be paid are obtained	Text	Standardize the values. Lower the values and make them more short	Yes

Codigo BPIN	If it corresponds to a process financed by the National Investment Programs and Projects Bank - DNP, the code is recorded here	Text	Transform variable to be able to identify if the project is in the DNP or not.	Yes
Proponentes Seleccionados	List of the selected bidders within the purchase process	Text	Most not defined (96%). High Cardinality	No
Calificacion Definitiva	Final qualification of the bidders within the purchase process	Text	Most not define (96%), Most values are strings with the word "Puntos"	No
ID Sub Unidad Ejecutora	ID of the budget execution Sub Unit assigned to the purchasing process	Text	Most not define (93%)	No
Nombre Sub Unidad Ejecutora	ID and name of the budget execution Sub Unit assigned to the purchasing process	Text	Most not define (95%)	No
Ruta Proceso en SECOP I	Process path in SECOP, to find detailed information about the purchase process	Text	Is URL	No
Moneda	Currency on which the purchases are registered in	Text	NA	Yes
EsPostConflicto	Flags if the purchase process is related to any of the actions relative to the 2017 peace process agreement	Text	Almost all the values are 1 value	No
Marcacion Adiciones	Flags if the purchase process has any registered additions	Text	Almost all the values are 1 value	No
Posicion Rubro	ID of the budget line	Text	Most not define (99%)	No
Nombre Rubro	Budget line	Text	Most not define (99%)	No
Valor Rubro	Value of the budget line	Number	Cast to number. Most of data is equal to 0	No
Sexo RepLegal Entidad	Entity's legal representative gender	Text	Most not define (79%)	No

Pilar Acuerdo Paz	If the purchase is related to the 2016 peace agreement, pillar or foundation of the agreement the purchase aims.	Text	Most not define (99%)	No
Punto Acuerdo Paz	If the purchase is related to the 2016 peace agreement, item of the agreement the purchase aims.	Text	Most not define (99%)	No
Municipio Entidad	Municipality to which belongs the purchaser entity	Text	None	Yes
Departamento Entidad	Department to which belongs the purchaser entity	Text	None	Yes
Ultima Actualizacion	Date and time the record was last updated	Text	High Cardinality	No

Table 2. SECOP II

Name	Description	Type	Cleaning Method	Relevant
Nombre Entidad	Name of the state entity that publishes the contract	Text	NA	Yes
Nit Entidad	NIT of the state entity that publishes the contract	Number	Cast to number	No
Departamento	Department in which the state entity that publishes the contract was registered	Text	DANE (National Department of Statistics) coding matching	Yes
Ciudad	City in which the state entity that publishes the contract was registered	Text	DANE (National Department of Statistics) coding matching	Yes
Localización	Full location of the state entity that publishes the contract	Text	DANE (National Department of Statistics) coding matching	Yes
Orden	Order of the state entity that	Text	Cast to	No

	publishes the contract		categorical variable	
Sector	Sector entity of the state that publishes the contract	Text	Cast to categorical variable	Yes
Rama	Branch of the state of the entity that publishes the contract	Text	Cast to categorical variable	Yes
Entidad Centralizada	Defines if the entity is decentralized or centralized	Text	Cast to boolean	Yes
Proceso de Compra	Identifier of the published purchase process	Text	Cast to categorical variable	Yes
ID Contrato	Identifier of the signed contract, generated by the platform	Text	NA	Yes
Referencia del Contrato	Identifier of the signed contract, generated by the state entity	Text	NA	No
Estado Contrato	Status of the contract, compared to its execution, signature or settlement	Text	Cast to categorical variable	Yes
Codigo de Categoria Principal	UNSPSC code of the main category for the contract	Text	Cast to categorical variable	Yes
Descripcion del Proceso	Description of the object of the purchase process	Text	NA	Yes
Tipo de Contrato	Type of contract according to its legal framework	Text	Cast to categorical variable	Yes
Modalidad de Contratacion	Contracting modality according to the selection model	Text	Cast to categorical variable	Yes
Justificacion Modalidad de Contratacion	Justification of the modality, the scenario under which the decision to define one or another contracting modality is made	Text	Cast to categorical variable	Yes
Fecha de	Date the contract was digitally signed	Datetime	Cast to date	Yes

Firma				
Fecha de Inicio del Contrato	Start date of contractual responsibilities	Datetime	Cast to date	Yes
Fecha de Fin del Contrato	End date of contractual responsibilities	Datetime	Cast to date	Yes
Fecha de Inicio de Ejecucion	Start date of the execution of the contract activities	Datetime	Cast to date	Yes
Fecha de Fin de Ejecucion	End date of the execution of the contract activities	Datetime	Cast to date	Yes
Condiciones de Entrega	Conditions under which the product or service is delivered	Text	NA	No
TipoDocProveedor	Type of document of the awarded supplier	Text	NA	No
Documento Proveedor	Document number of the awarded supplier	Text	NA	Yes
Proveedor Adjudicado	Name of the awarded provider	Text	NA	Yes
Es Grupo	Determines the provider is a group of entities, there is a CCE data set that contains the conformation of the groups	Text	Cast to categorical variable	Yes
Es Pyme	Determine if the company is an SME	Text	Cast to categorical variable	Yes
Habilita Pago Adelantado	Determine if the contract has the advance payment option enabled	Text	Cast to boolean	Yes
Liquidación	Determine if the contract has been settled	Text	Cast to boolean	Yes
Obligación Ambiental	Determine if the contract has commitments to comply with environmental obligations	Text	NA	No
Obligaciones Postconsumo	Determines if the contract has commitments to fulfill obligations subsequent to the delivery of the	Text	Cast to categorical variable	Yes

	product or provision of the service			
Reversion	Determine if the contract has been reversed	Text	NA	Yes
Valor del Contrato	Total amount of the contract until date	Number	Cast to number	Yes
Valor de pago adelantado	Anticipated contract value	Number	Cast to number	No
Valor Facturado	Total amount of the contract until date	Number	Cast to number	Yes
Valor Pendiente de Pago	Total amount not paid yet.	Number	Cast to number	No
Valor Pagado	Total amount paid.	Number	Cast to number	Yes
Valor Amortizado	Amortized value to date	Number	Cast to number	No
Valor Pendiente de Amortizacion	Amortized value not pay yet to date	Number	Cast to number	No
Valor Pendiente de Ejecucion	Total amount not paid yet in execution.	Number	Cast to number	Yes
Estado BPIN	Status of assignment of the Investment Projects Bank code	Text	Cast to categorical variable	Yes
Código BPIN	Code associated with the Investment Projects Bank	Text	NA	Yes
Anno BPIN	Year of allocation of the Investment Projects Bank code	Text	NA	No
Saldo CDP	CDP balance assigned to process and contract	Number	Cast to number	Yes
Saldo Vigencia	Current balance for the term of the CDP assigned to the process and the contract	Number	Cast to number	Yes
EsPostConflicto	Determine if the process is associated with any peace agreement event	Text	Cast to boolean	Yes

URLProceso	URL of the purchase process on the SECOP II platform	Text	NA	No
Destino Gasto	Destination of expenditure, at the budget level	Text	Cast to categorical variable	Yes
Origen de los Recursos	Origin of resources, at the budget level	Text	Cast to categorical variable	Yes
Días Adicionados	Number of days the contract has been added	Number	Cast to number	Yes
Puntos del Acuerdo	In case of being a process that fulfills commitments in the peace agreement, it determines to which points it gives conformity	Text	NA	Yes
Pilares del Acuerdo	In case of being a process derived from peace agreement commitments, define the pillar of the peace agreement to which it corresponds	Text	NA	Yes

14. Appendix B- Team Contribution