DS 4A COLOMBIA

Colombia Compra Eficiente

# WEEK 7 PROJECT:
# IN-DEPTH EDA AND FE MOCKUP

## Colombia Compra Eficiente - Agencia Nacional De Contratación

### Team 82


Cindy Ramirez  Pedro Casas

Karina Mesa  Jorge Enciso

Camilo Cabrera  Samuel Pérez

Yasmin Moya


---

The Agencia Nacional de Contratación Pública - Colombia Compra Eficiente (ANCPCCE), 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" administered by Colombia Compra Eficiente.

The organization requires form groups of contracts that are related by identifying common words in the contractual object un automatically 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. 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. Therefore, Colombia Compra Eficiente would have more clearer and more useful information to generate real value and savings in the country's public procurement.

## 2. 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.

**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.

## 3. 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.

### 3.1. SECOP I

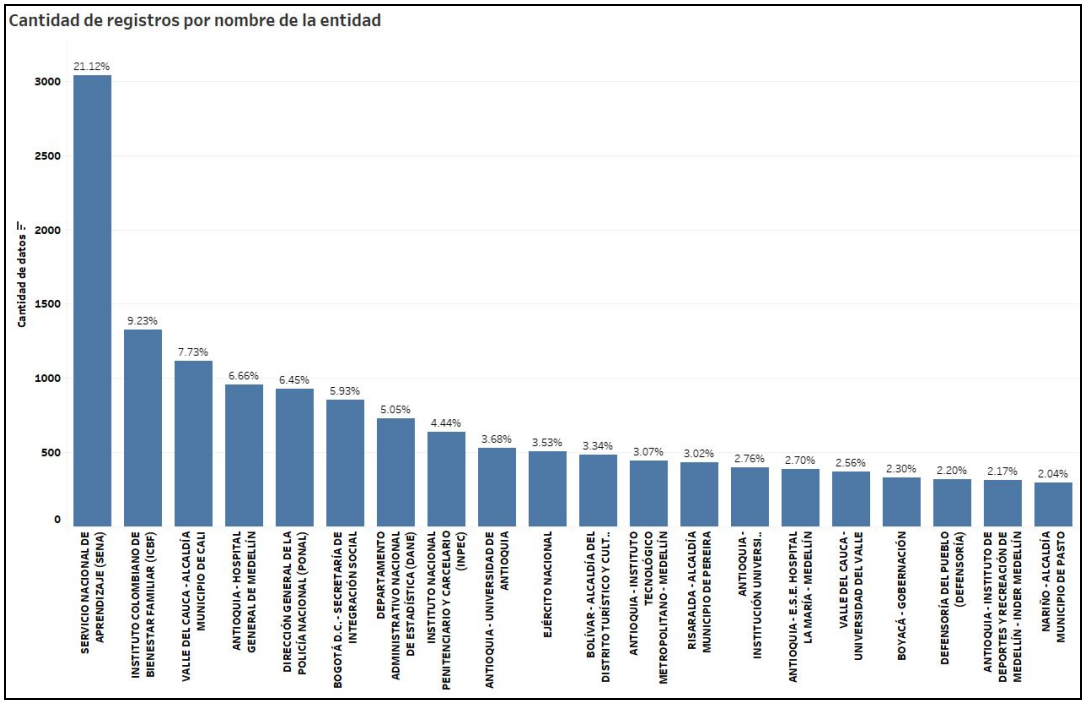
| Objeto A Contratal. | Valor promedio del contrato ya pagado | Valor del contrato promedio | Valor Adicional promedio | Tiempo adicional en días | Plazo para pagar | Cantidad de datos |
|---|---|---|---|---|---|---|
| Servicios de Edificación, Construcción de Instalaciones y Mantenimiento | $ 763,222,736.68 | $ 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 1. Heatmap quantitative variables by the contractual subject*

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 Millions 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. By 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.

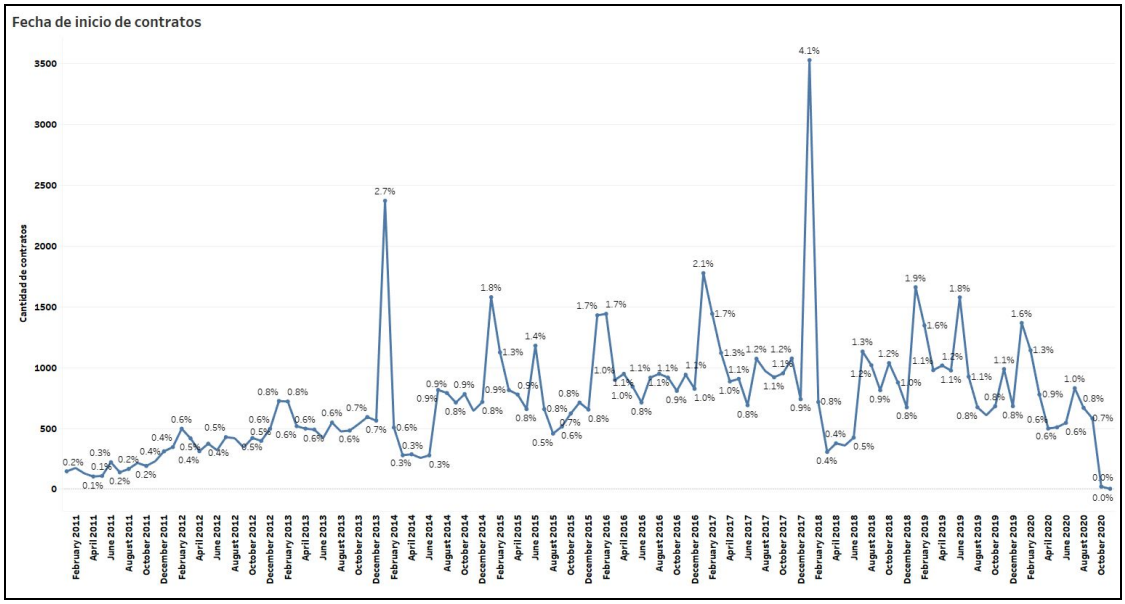
## Number of contracts per entity




*Graph 2. Number of contracts per entity*

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.

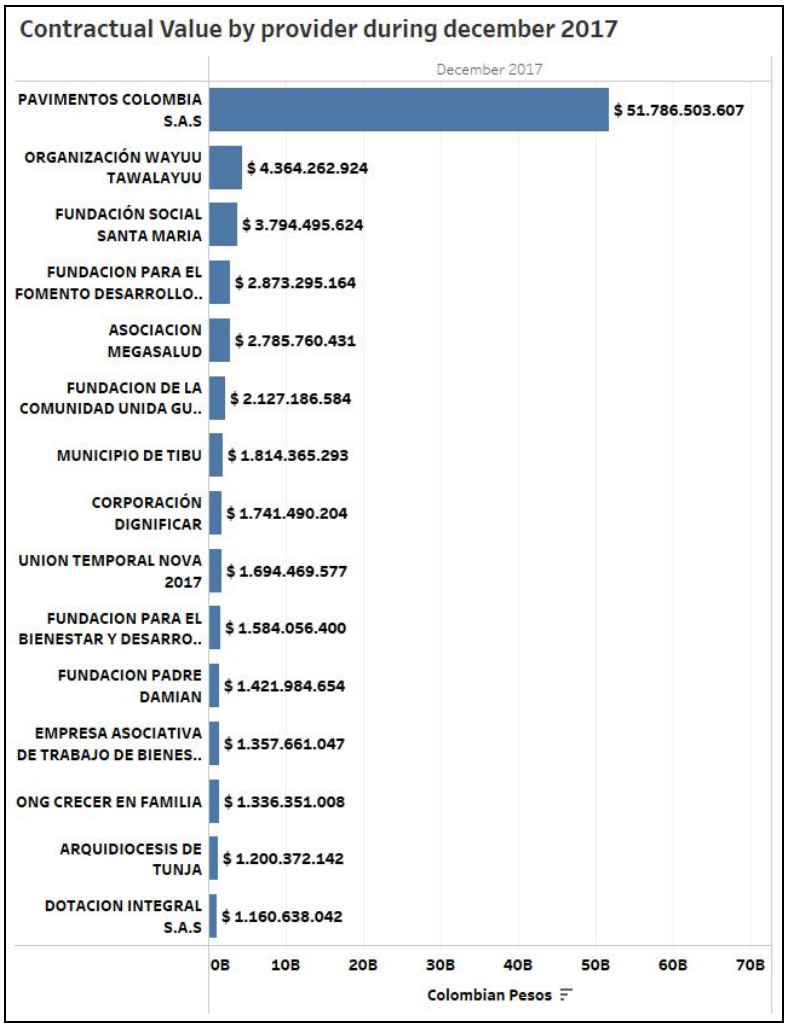
## Contract start date




*Graph 3. Initial date of contracts.*

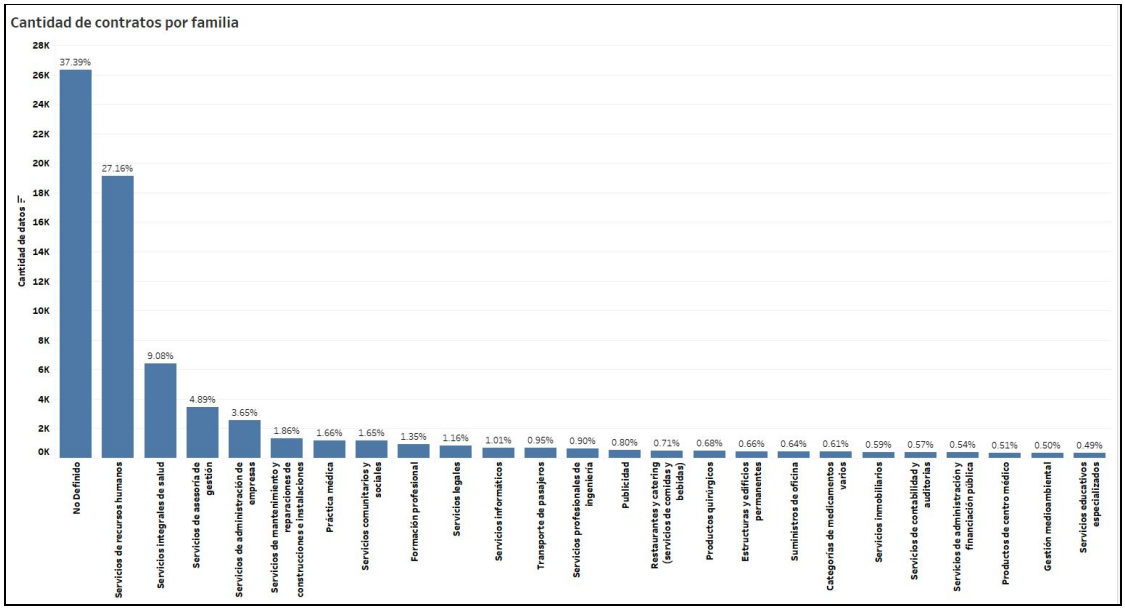
Main insights:

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.



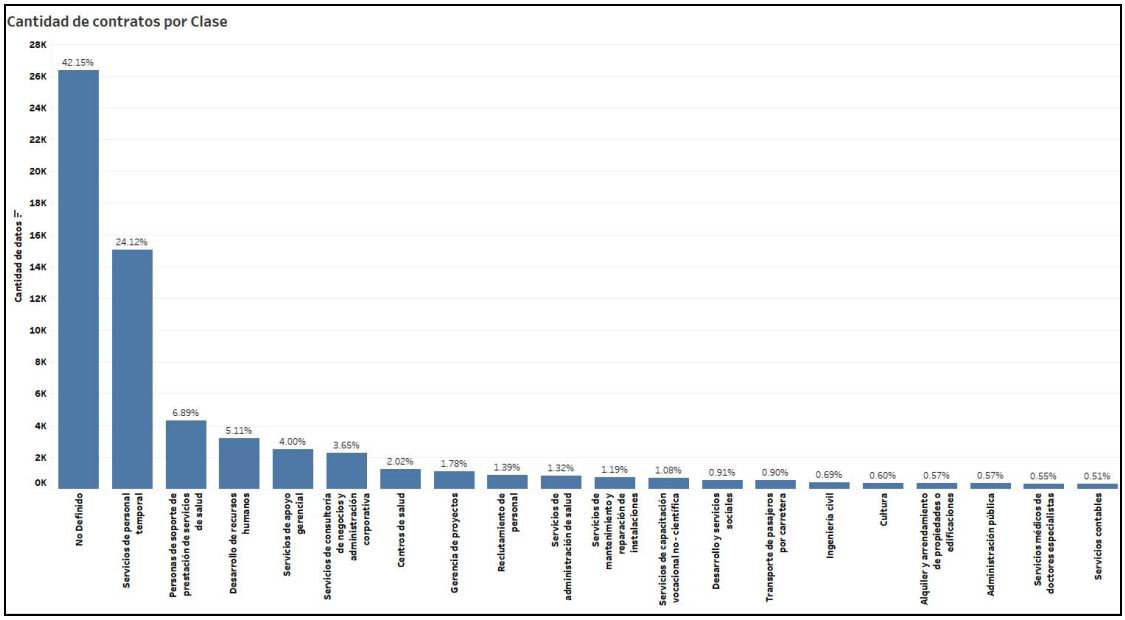

*Graph 4. Contractual Value by provider during december 2017*

## Number of contracts per family




*Graph 5. Number of contracts per family.*

## Number of contracts per class




*Graph 6. Number of contracts per class.*

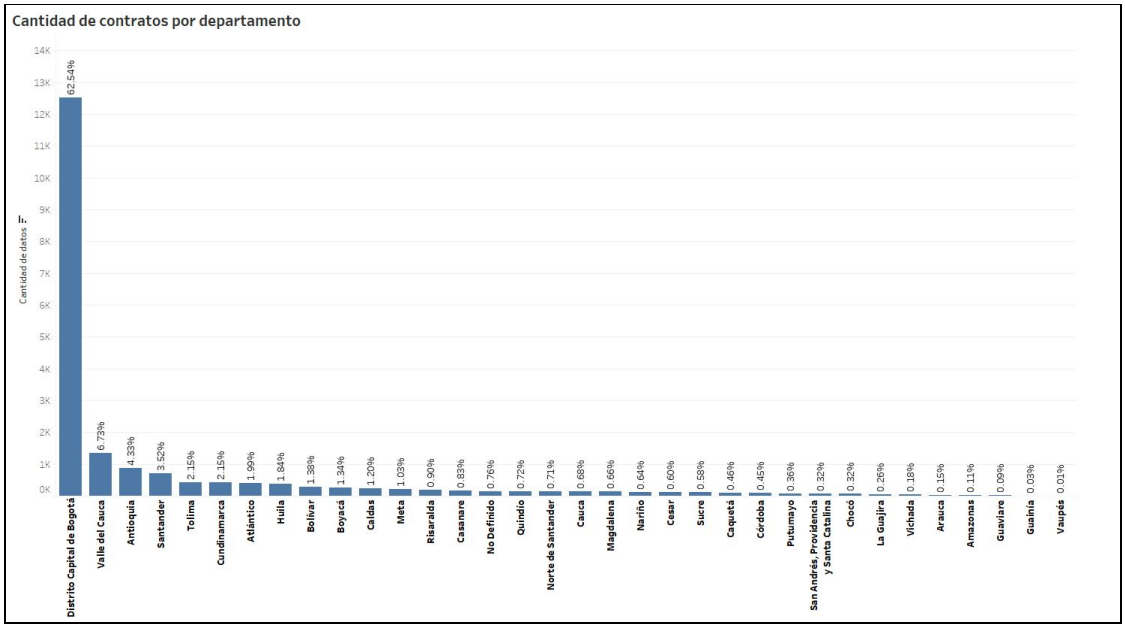
These two graphs show the family and the class for each procurement. There is a lot of variety in these two graphs. As it is very hard to conclude a trend between these clases, the only conclusion is that most of the values have not a family or class. Also the temporary

services and the human resources contracts are the most common among all the families and contracts.

### 3.2. SECOP II

Each row of the database corresponds to a contract, there are more than 9,23 millions 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.

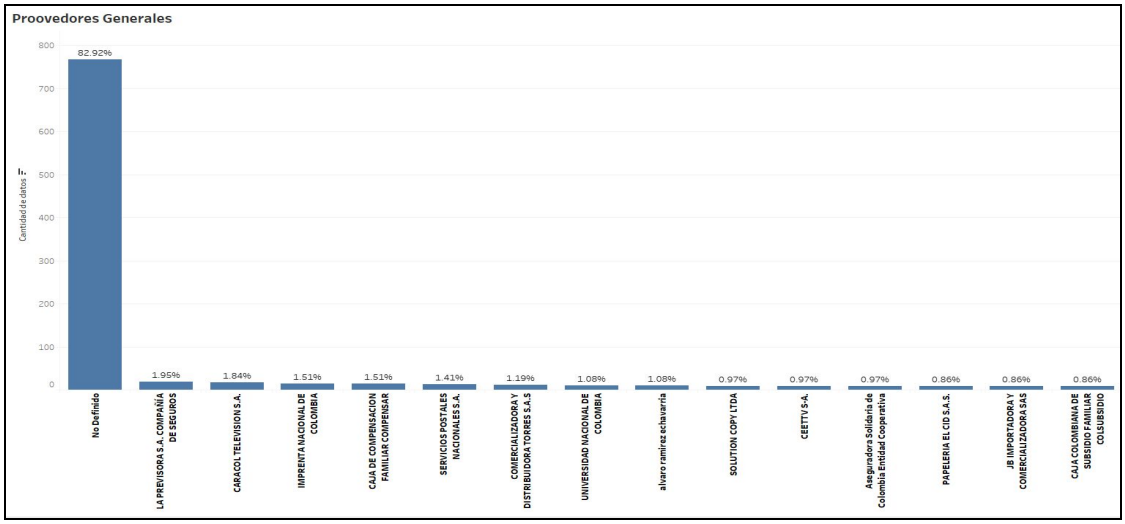
**Number of contracts by department**



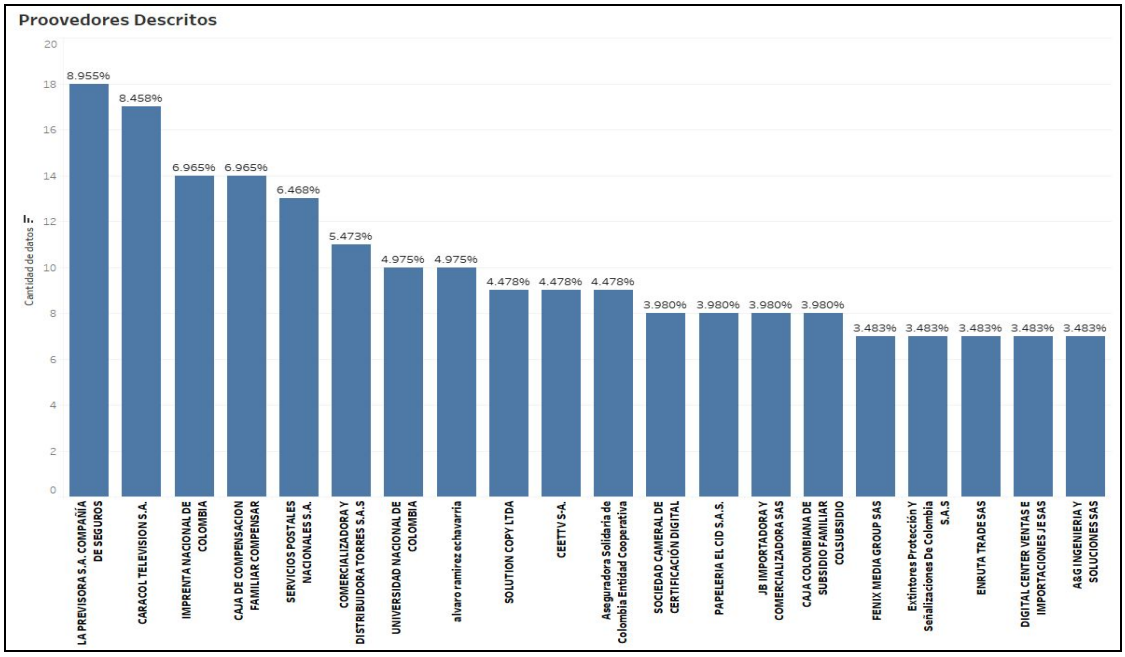

*Graph 7. Number of contracts by department*

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 8. General Supplies*




*Graph 9. Define Suppliers*

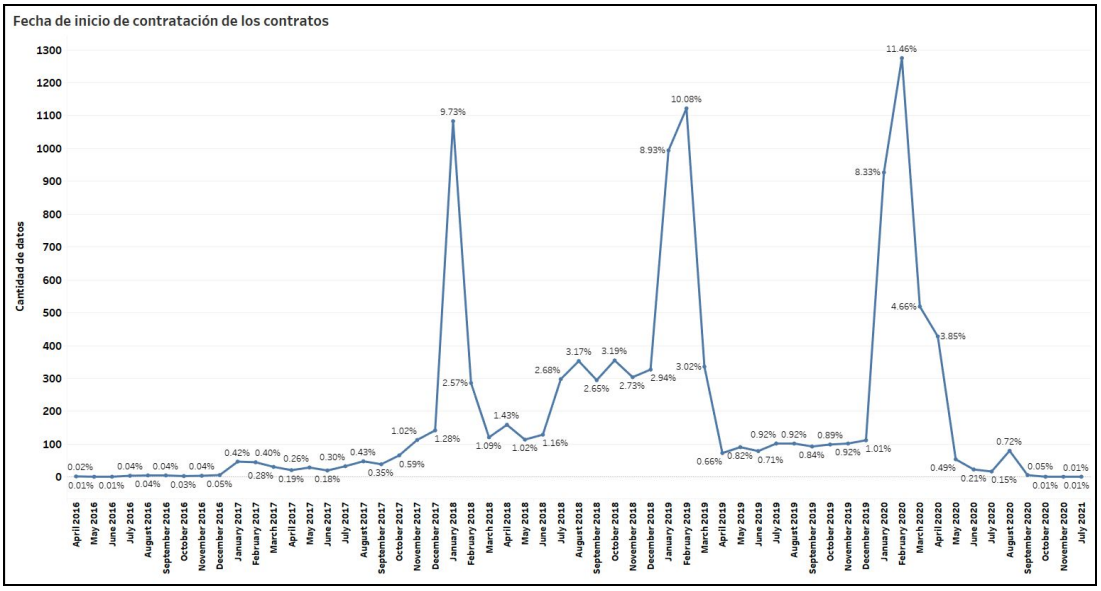
In the graph of General Suppliers, it can be observed that more than 80% of the data is concentrated in a supplier NOT defined. This may be due to the fact that when entering the information into the SECOP II database, this field is not mandatory and they omit it.

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**

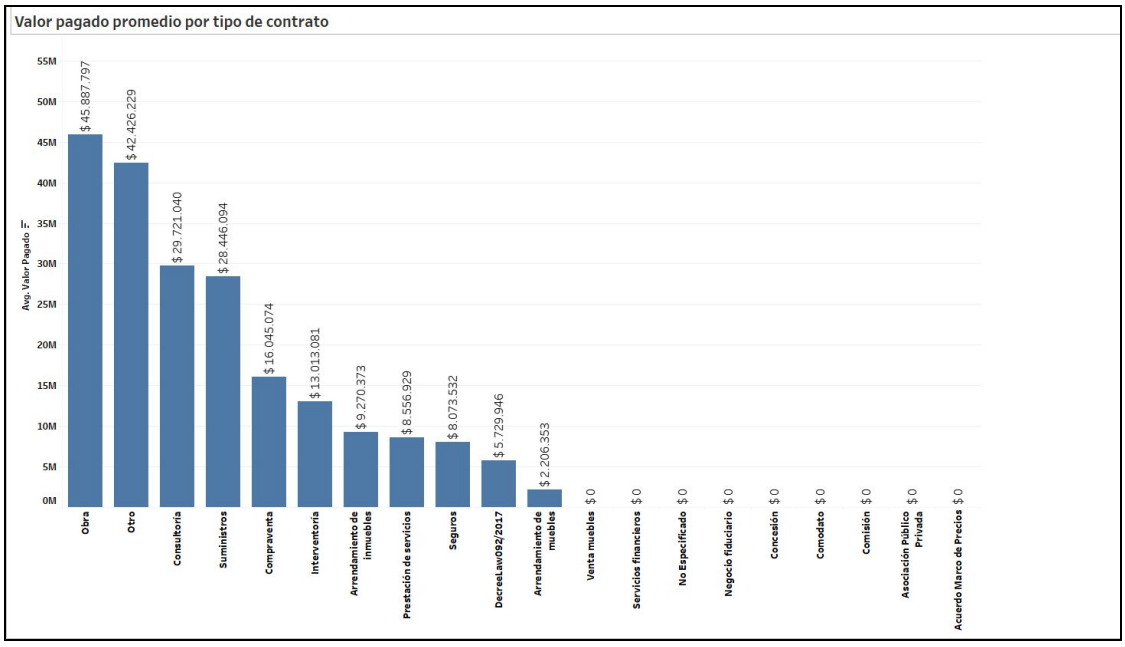



*Graph 10. Contract Start date*

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.

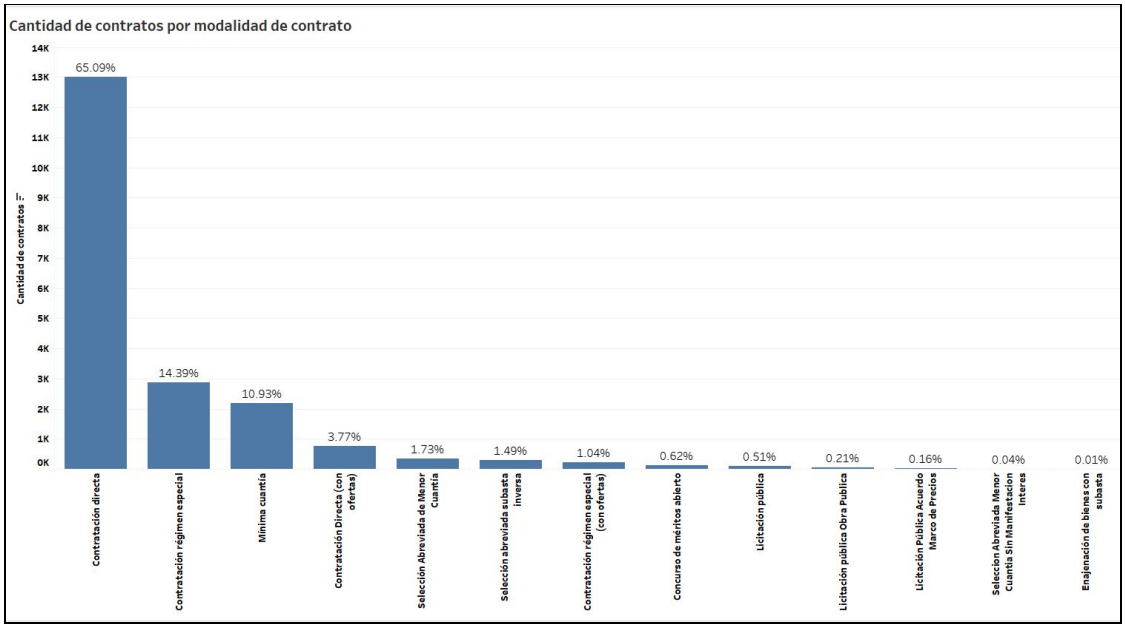
**Average paid value by type of contract**




*Graph 11. Average value paid by type of contract*

"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 12. Number of contracts by type of contract*

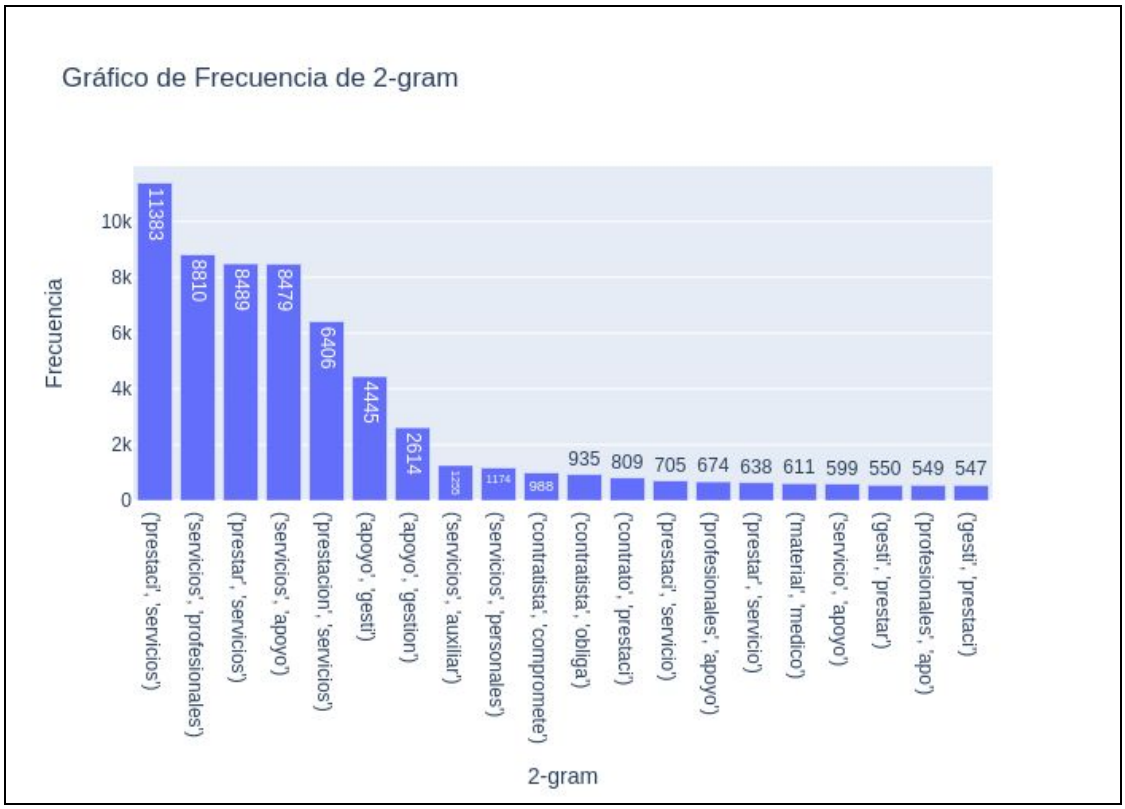
After the analysis of the type of contracts, around 80% of the records are distributed between "direct contracting" and "contracting special regime". Adding the “minimum amount” category the concentration raises to 90% of the records.

### 3.3. Advanced EDA over “Contractual Object”

**SECOP I**

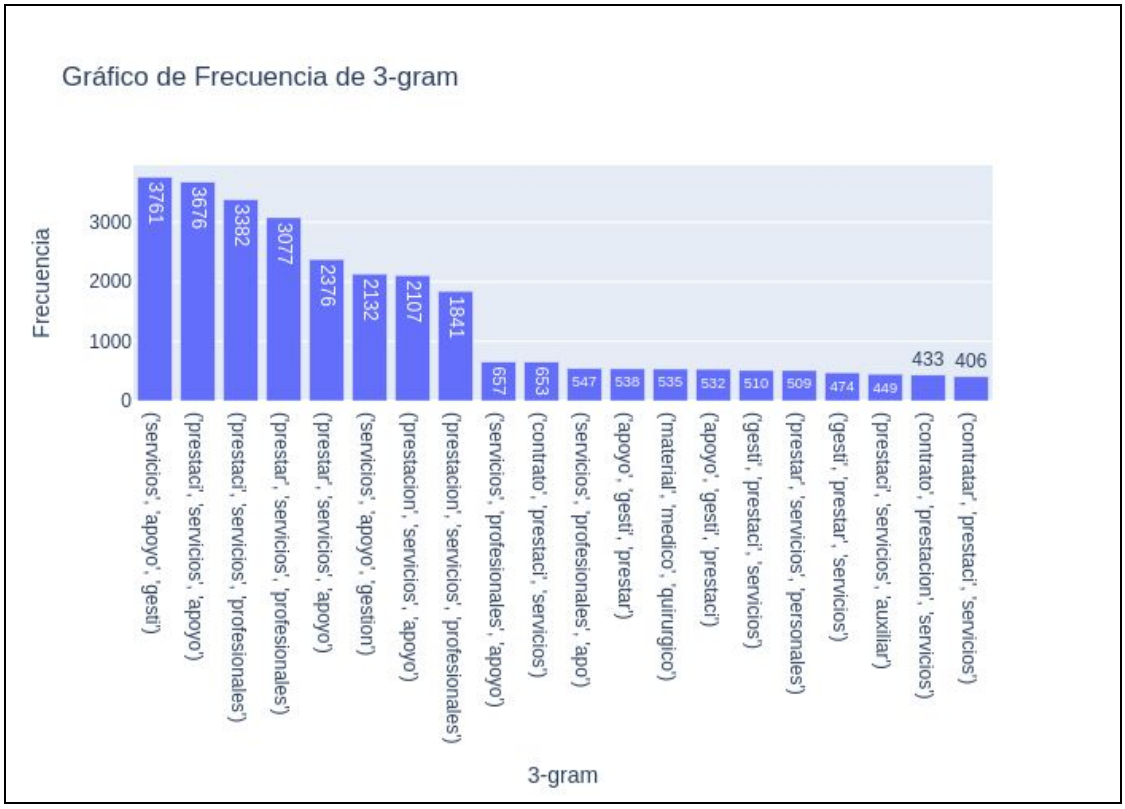
*Graph 13. Top 20 words from the contractual object of SECOP I .*

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.



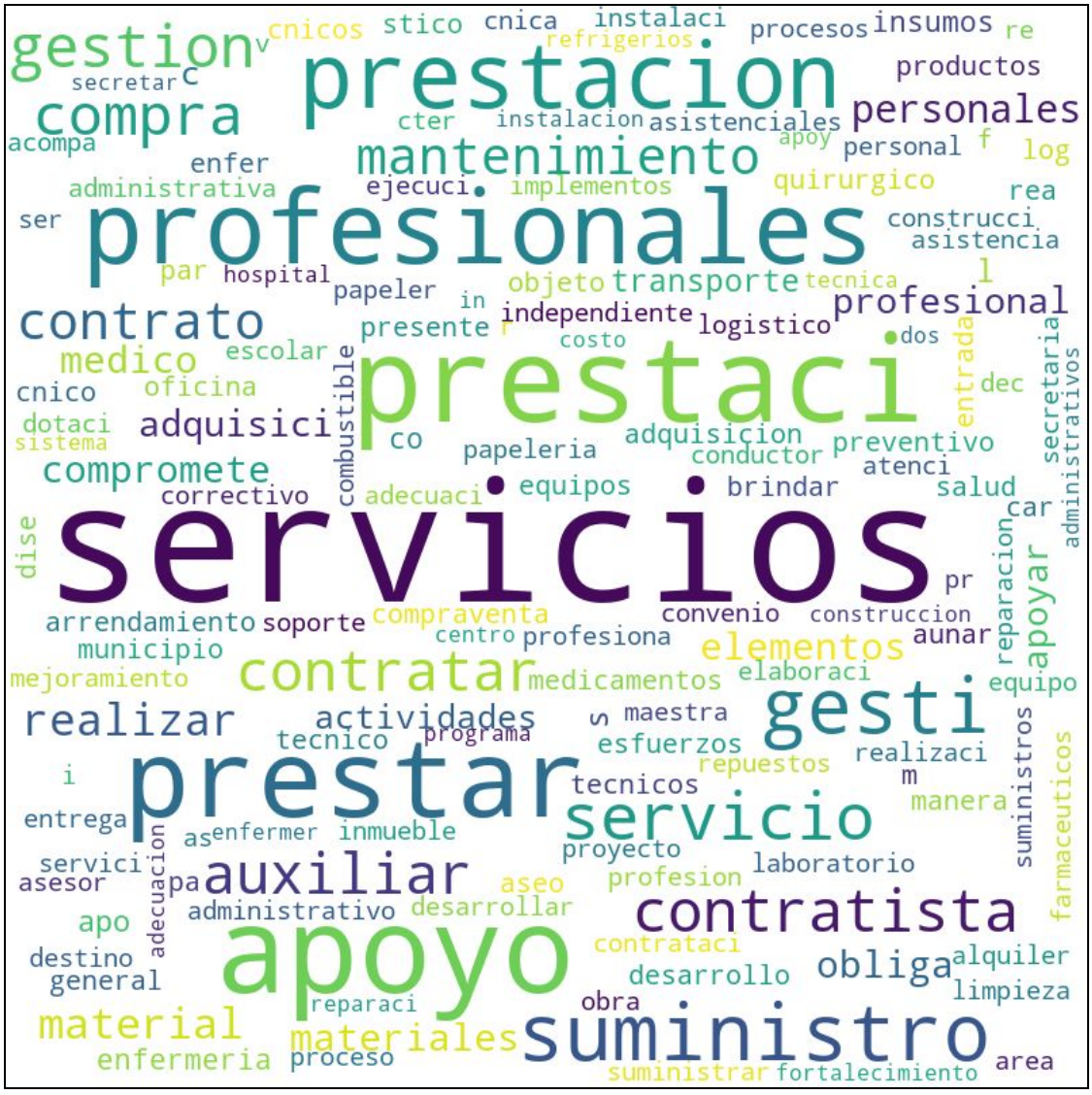

*Graph 14. Top 20 bigrams from the contractual object of SECOP I .*

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.



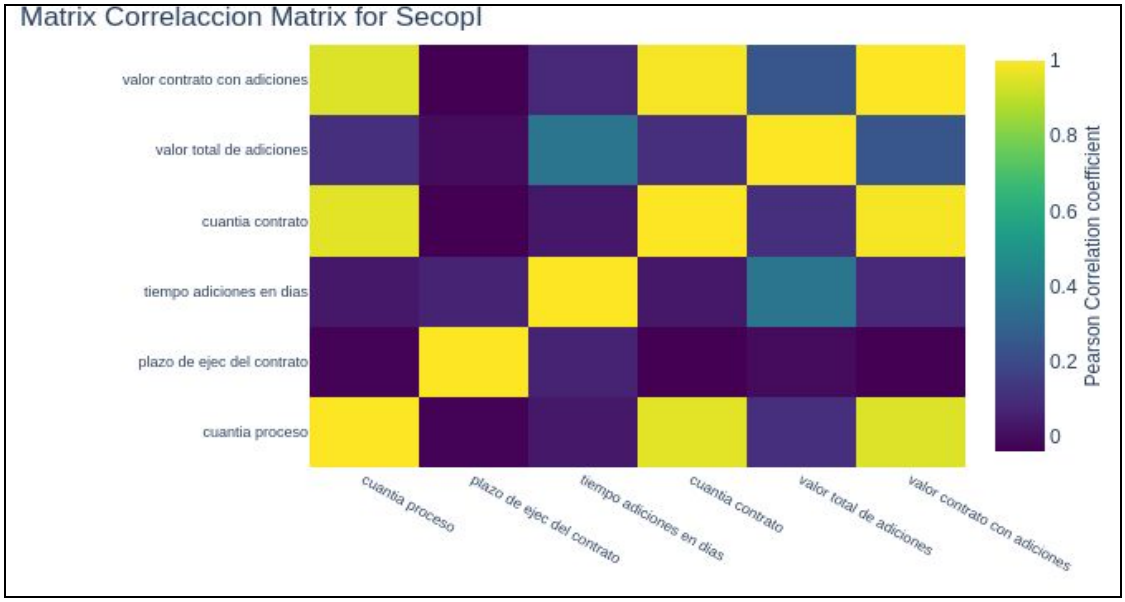

*Graph 15. Top 20 trigrams from the contractual object of SECOP I .*

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 16 . WordCloud from SECOP I.*

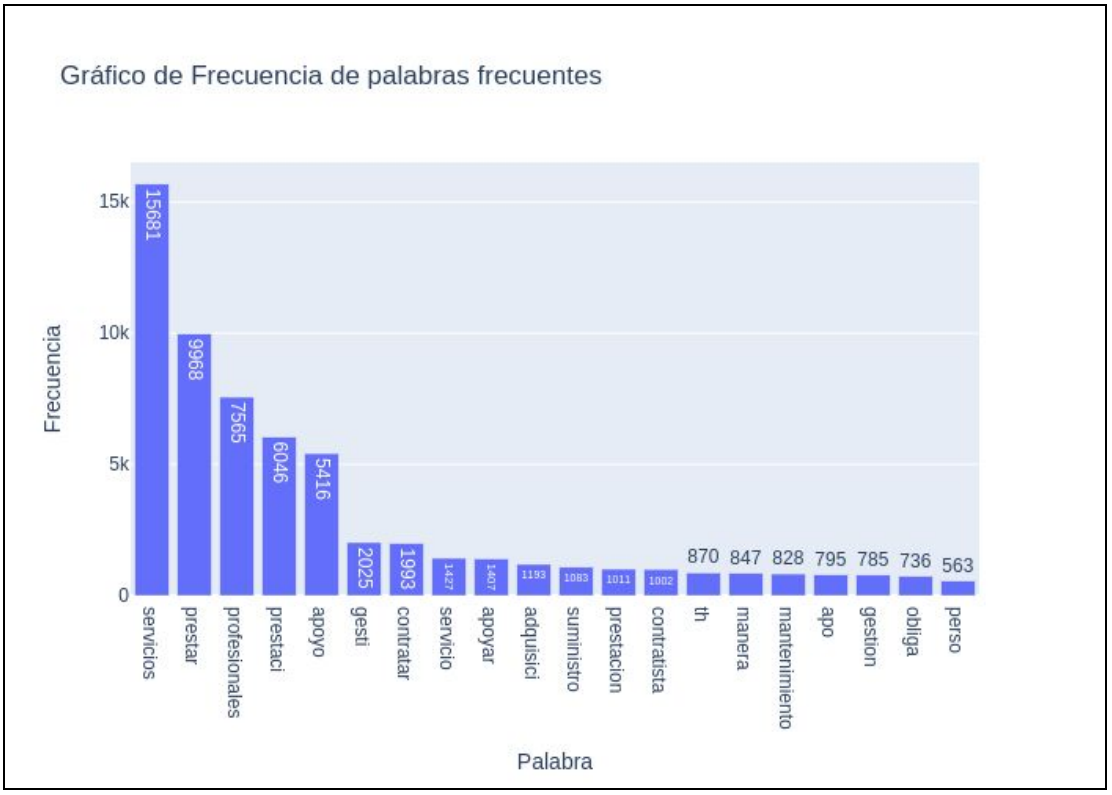
The graphs above show that the most common words are “servicios”, “apoyo” and “prestasi”. For bigrams the most common pairs of words are: “prestaci, servicios” and “servicios, profesionales”. Finally the most common trigrams in SECOP I are “servicios, apoyo gesti” and “prestaci, servicios, apoyo”. This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




*Graph 17. Correlation matrix of SECOP I .*

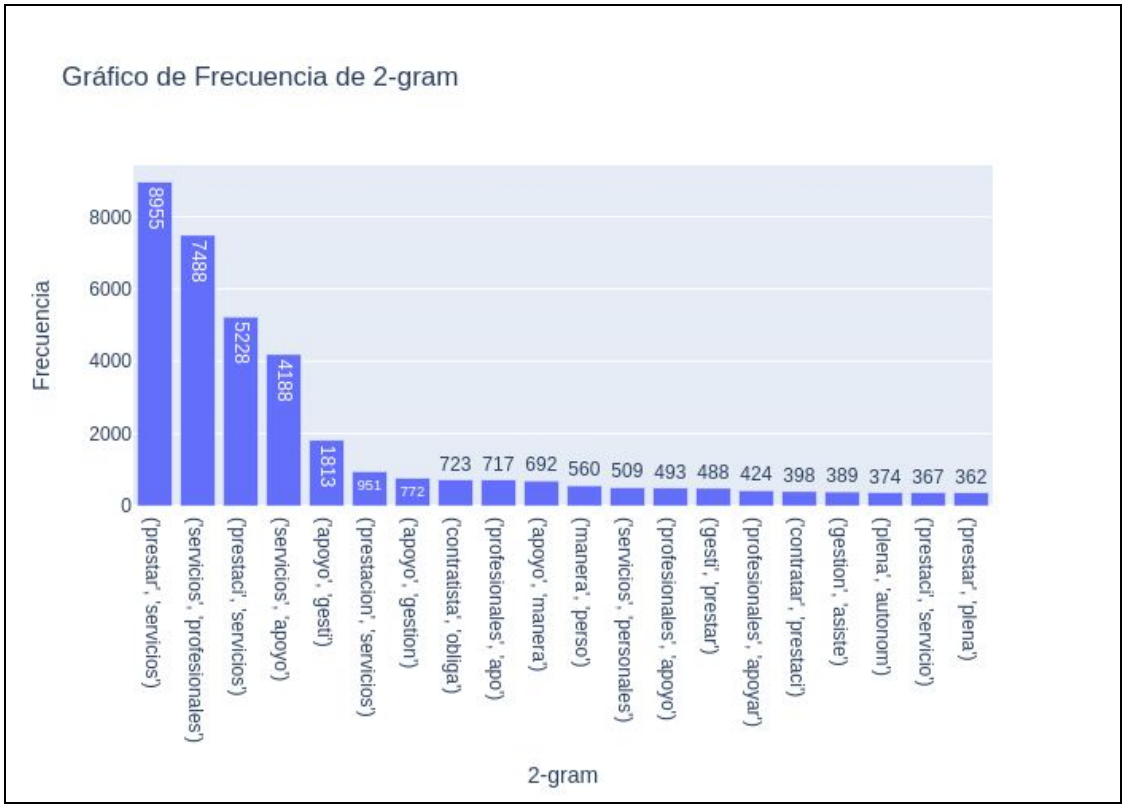
The correlation matrix shows natural dependency between the variables the total value pay and the value of the process as well as value of the extra fees with the contract extra time. It seems that the duration of the contract is independent from the other variables.

**SECOP II**



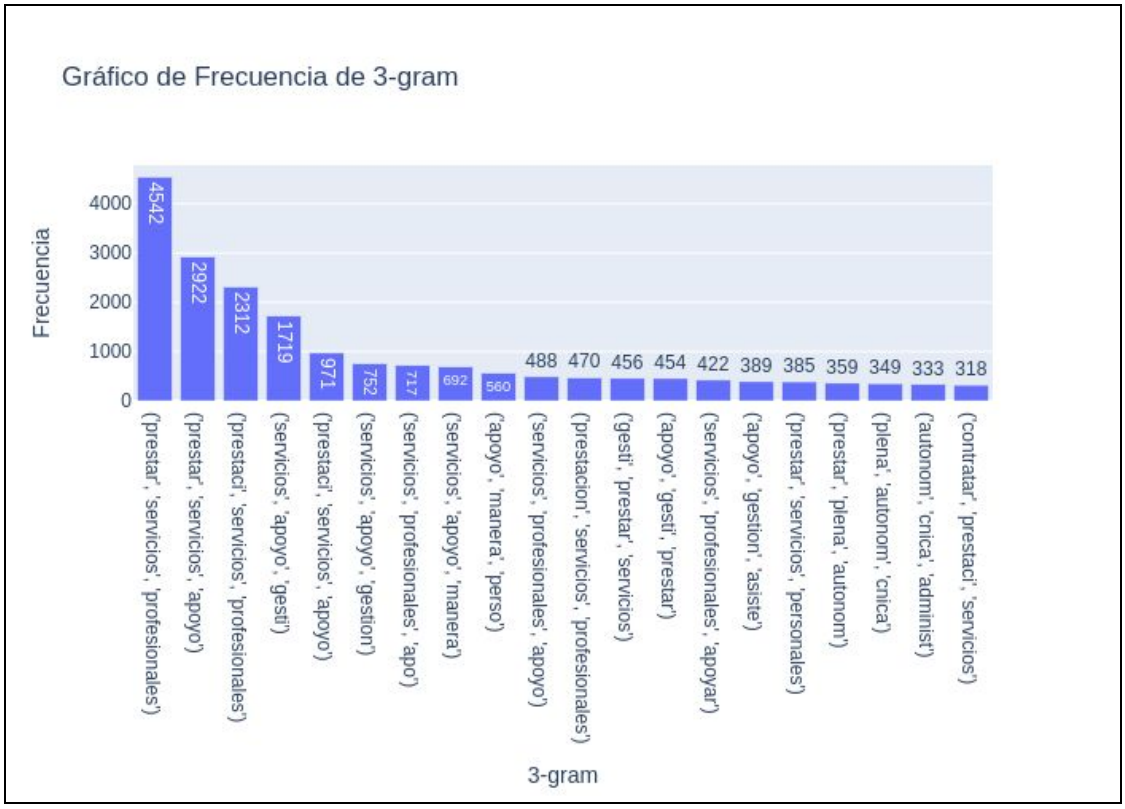

*Graph 18. Top 20 words from the contractual object of SECOP II .*

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.



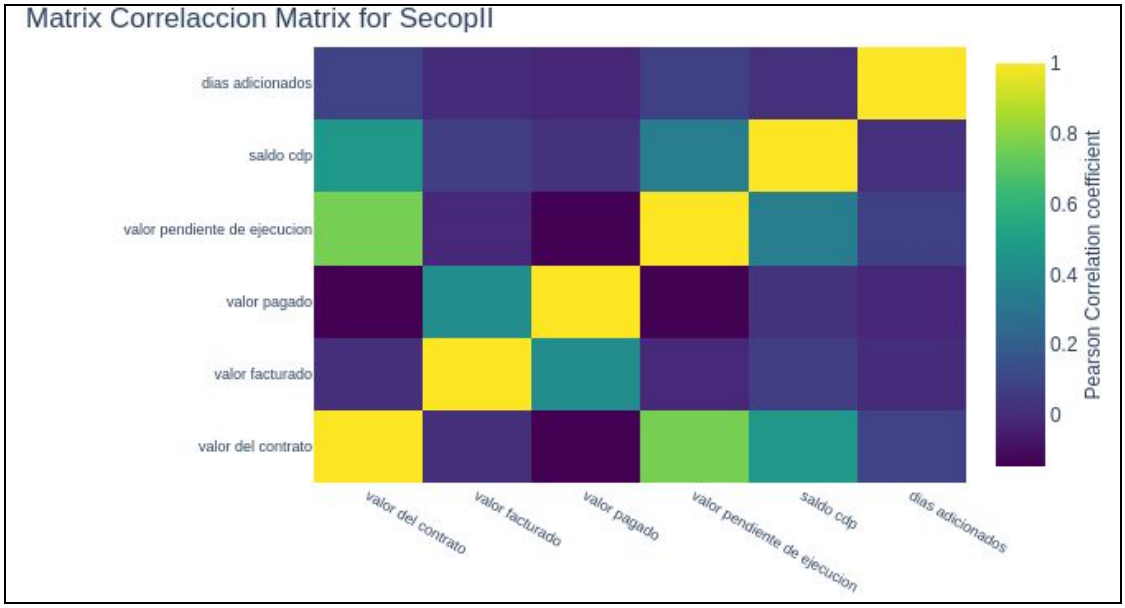

*Graph 19. Top 20 bigrams from the contractual object of SECOP II .*

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.



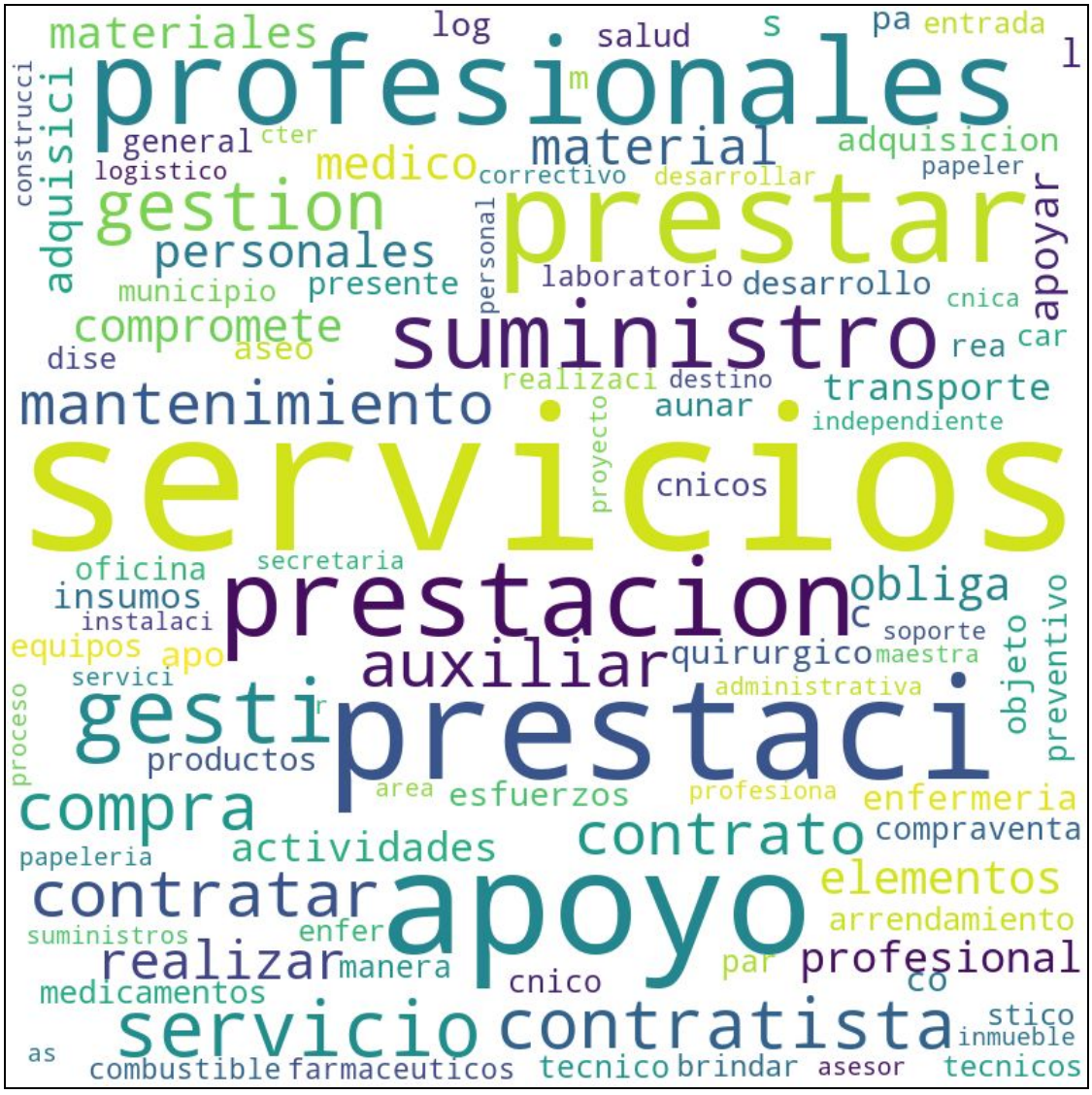

*Graph 20. Top 20 trigrams from the contractual object of SECOP II .*

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 21. Correlation matrix of SECOP II.*

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”.

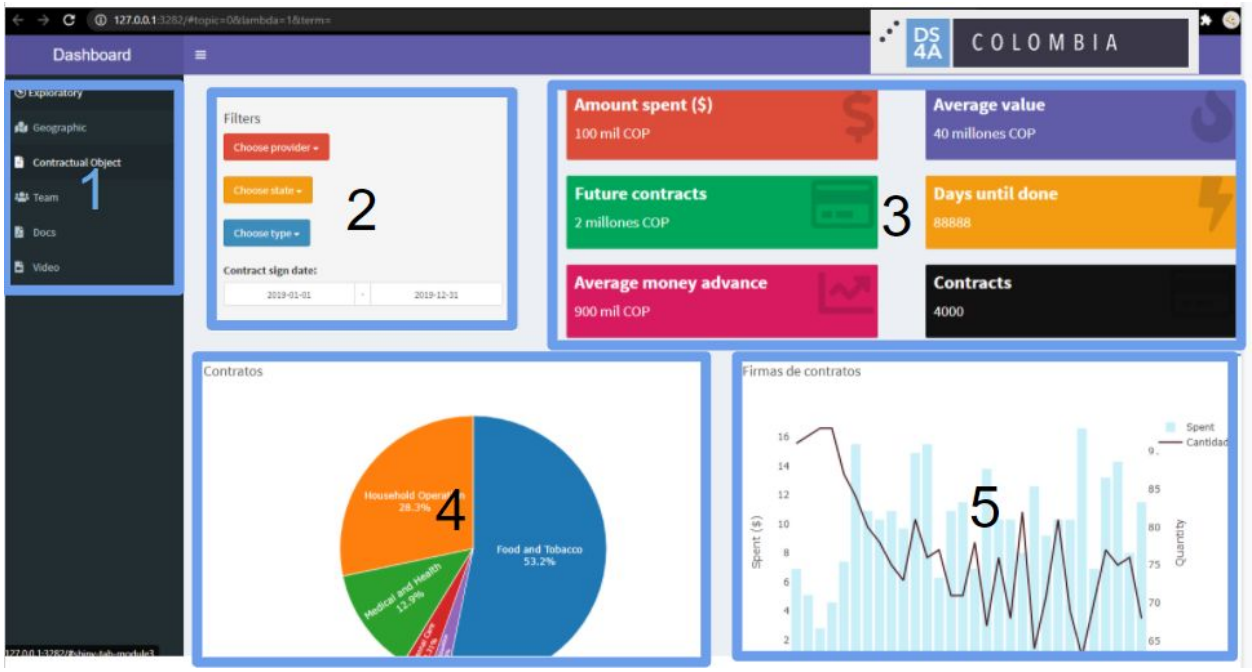



*Graph 22. Word Cloud from SECOP II.*

In the world cloud graph (Graph 22), 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.

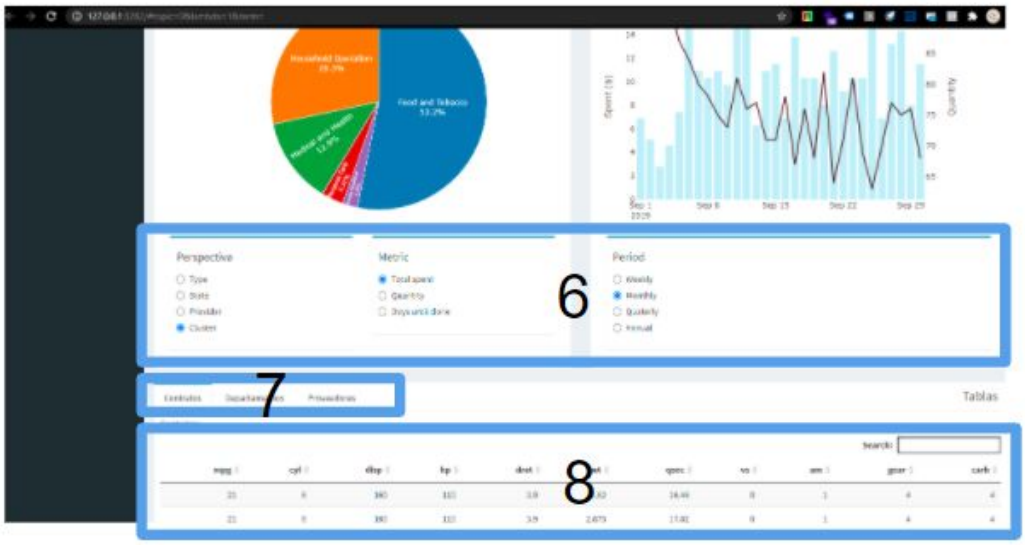
## 4. Mock Up

### Exploratory




*Graph 23. Exploratory and Descriptive View 1.*

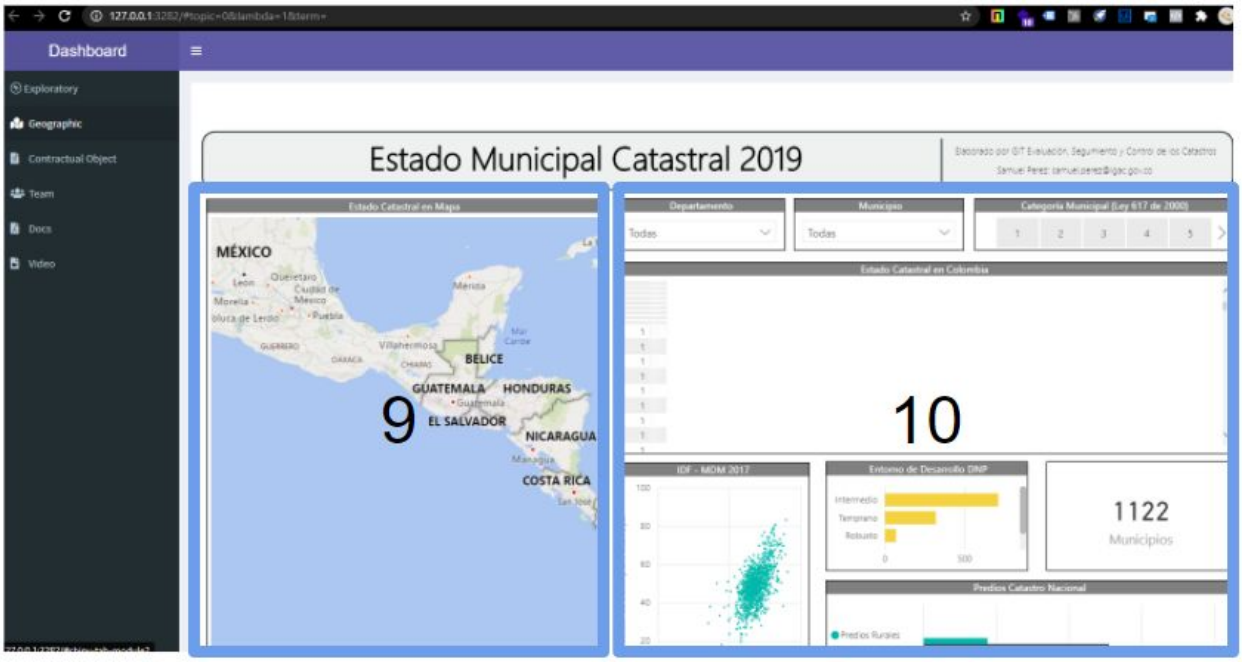
1. 5-tab menu:
    - Exploratory Analysis: Data analysis
    - Geographic: Geographical analysis
    - Contractual Object: text analytics to contractual objects
    - Team: Team contact and description
    - Docs: Project documentation
    - Video: Video summary
2. Filters to analyze the data
3. Main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




*Graph 24. Exploratory and Descriptive View 2.*

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

## Geographic

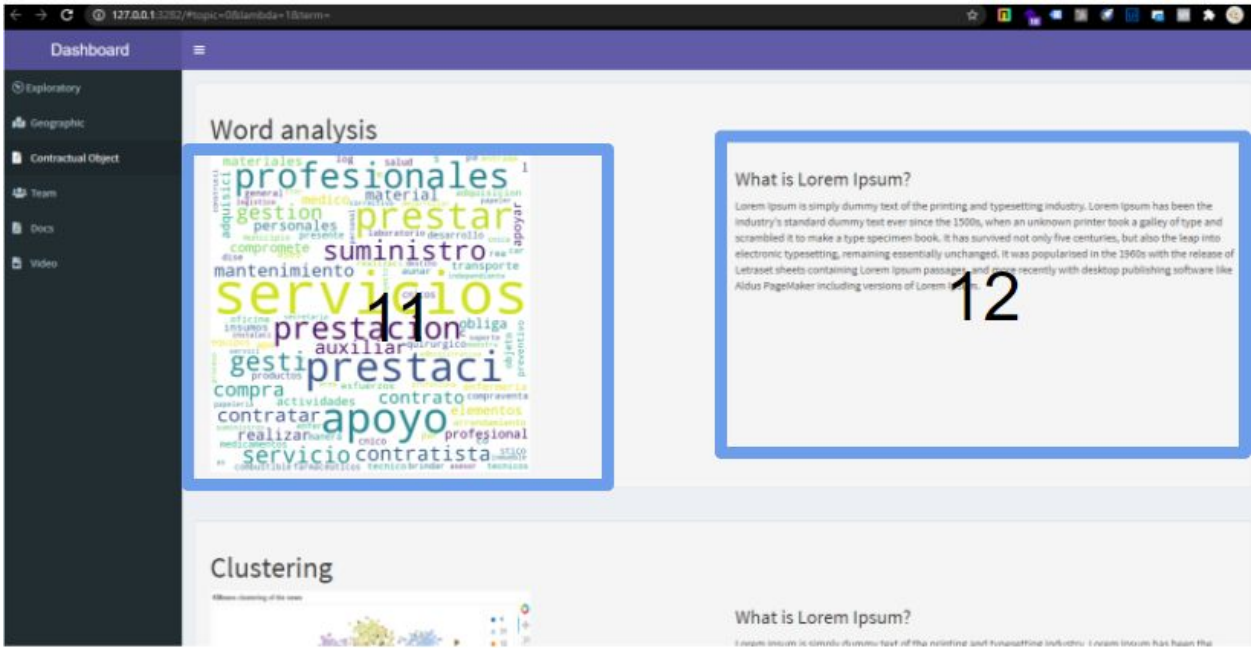



*Graph 25. Geographic View.*

9. Interactive Map
10. Filters and other geographical insights

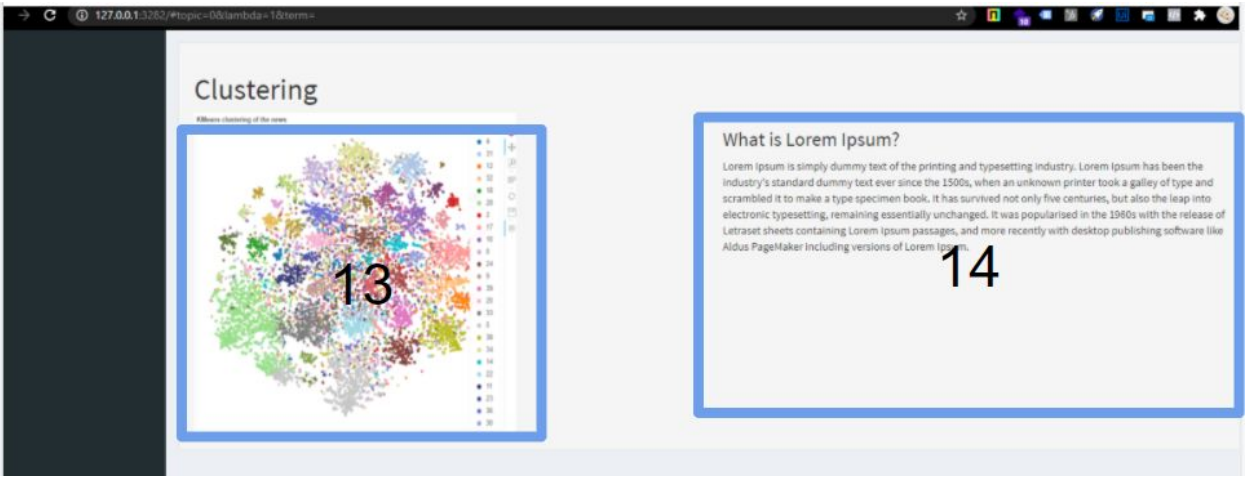
## Word Analysis

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 26. Word Analysis*

11. Words representation
12. Word analysis content




*Graph 27. Clustering View 1.*

13. Clustering representation
14. 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.

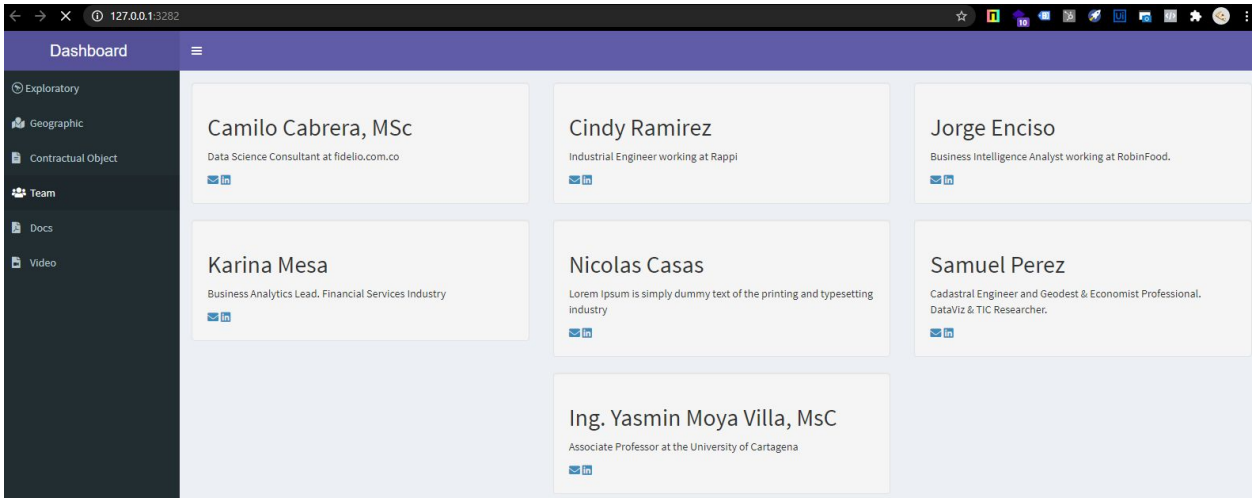

*Graph 28. Clustering View 2*

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 buble 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 section 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 .

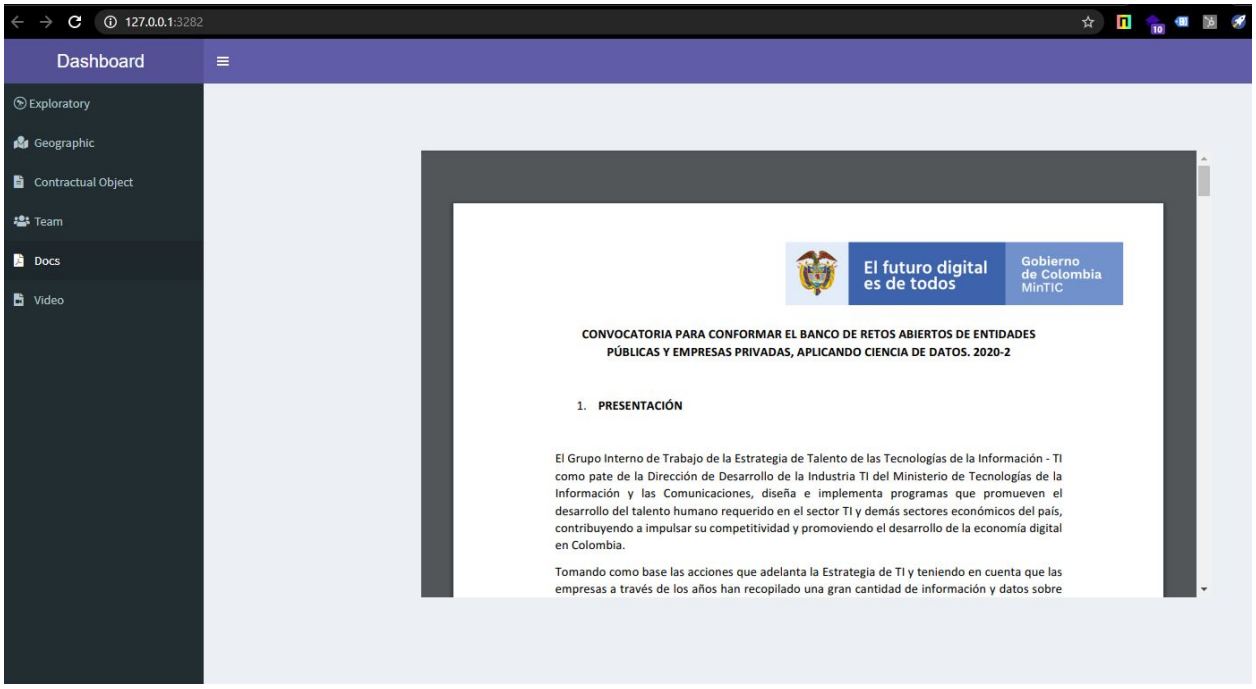
## Team




*Graph 29. Developers View*

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.

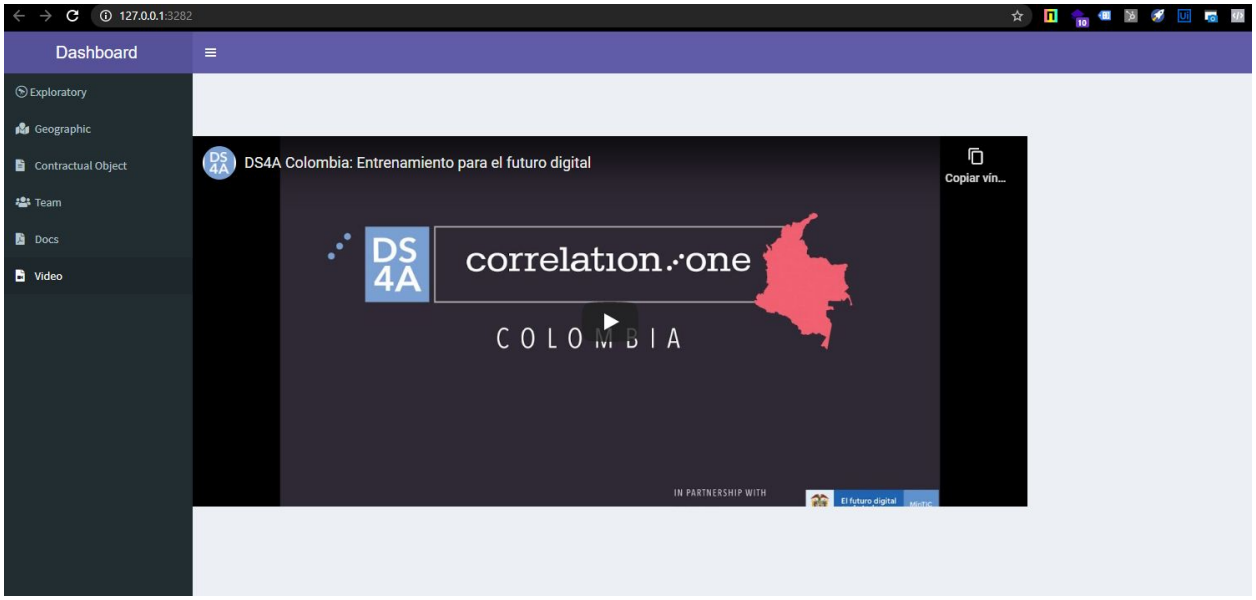
## Docs



*Graph 30. Documentation View*

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.

## Video




*Graph 31. Video*

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

## 5. Appendix

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 decodify. | 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 1% | Yes |
| **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 service | Text | Convert them into key dictionaries, it can work if the data is decodify. | No |
| **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** | In case of being a company, the type of identification of the legal | Text | NA | No |

| | | | | |
|---|---|---|---|---|
| **Legal** | representative of the company selected in the award | | | |
| **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%).<br>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** | Entity's legal representative gender | Text | Most not define | No |

| | | | | |
|---|---|---|---|---|
| **Entidad** | | | (79%) | |
| **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 publishes the contract | Text | Cast to categorical variable | No |
| **Sector** | Sector entity of the state that publishes the contract | Text | Cast to categorical | Yes |

| | | | | |
|---|---|---|---|---|
| | | | variable | |
| **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 Firma** | Date the contract was digitally signed | Datetime | Cast to date | Yes |
| **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 product or provision of the service | Text | Cast to categorical variable | Yes |
| **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** | Amortized value not pay yet to date | Number | Cast to number | No |

| | | | | |
|---|---|---|---|---|
| **Amortizacion** | | | | |
| **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 |