A Look at Health Care in Sogamoso, Boyacá

Team 222

PQRSDF Sogamoso
Sector Salud

DS4A Colombia

Aura Ramírez Arévalo ✦ David Vergara Agudelo ✦ Fabio Calderón Mateus
Katherin Del Risco Serje ✦ Laura Daniela Quiza ✦ Sandra Rivera Torres

Contents

1. Acronyms

DANE	Departamento Administrativo Nacional de Estadística *National Administrative Department of Statistics*
EAPB	Entidades Administradoras de Planes de Beneficios *Benefit Plan Administration Entities*
EPS	Entidades Prestadora de Salud *Health Provider Entities*
MinSalud	Ministerio de Salud de Colombia *Ministry of Health of Colombia*
PQRSDF	Peticiones, quejas, reclamos, sugerencias, denuncias y felicitaciones *Petitions, complaints, claims, suggestions, denunciations, and compliments*
SMS	Secretaría Municipal de Salud *Municipal Health Secretariat*
SIAU	Sistema Integral de Atención al Usuario *Integral system of User Support*
UNDP	*United Nations Development Programme*

2. The problem

OVERVIEW

Sogamoso is a city located in Boyacá, Colombia, which according to DANE projections, in 2022 has 132,985 inhabitants. Its economy is mainly based on the steel industry, construction materials production, coal mining, and agriculture. The hospital infrastructure of Sogamoso has three levels of attention through eight institutions.

132,985 inhabitants & 8 hospitals

Reception of PQRSDF through Orfeo

One of the branches of the Municipal Government is the SMS, which has implemented SIAU. Its main objective is the reception of PQRSDF, especially those related to the EAPB operation. This is done through Orfeo, which is a digital platform that handles all the PQRSDF of the city, including those directed to the SMS.

Requests are submitted either virtually (Orfeo) or in person/by telephone (SIAU Office). In the second case, the physical document is digitalized and attached to an online requirement, so Orfeo handles both methods.

The mission of SIAU is to address complaints and requirements filed by affiliates of the Subsidized and Contributory Regime, so the risk of diseases is mitigated and health prevention is performed. The implementation of the system follows the Quality Management System, governed by guidelines of MinSalud and UNDP.

Mitigation of diseases and health prevention

OPPORTUNITY

The SMS has detected that some of the EAPB do not guarantee proper attention to their affiliates, and thus users constantly fill PQRSDF. Currently, the city does not have a tool that can be used to prioritize the urgency and relevance of these petitions. Also, there is no analysis or study on this data at all.

Users fill PQRSDF about EAPB attention

Detect, model and predict

Hence, the main purpose of our Team is to use Data Science techniques to analyze the PQRSDF of Sogamoso in order to detect and explore the aspects of the Health System (procedures, medical dependencies, EAPB, etc.) that most affect the population of the city and that therefore require greater attention from the institutions. Also, to model and predict the type of PQRSDF, since Orfeo requests do not have this feature.

IMPACT

Healthcare is a Fundamental Right regulated by Colombian Law 1751 of 2015. Its objective is to regulate the Health System, and establish its protection mechanisms, via prevention of diseases, promotion of healthcare, and improvement of health indicators. Hence, raising the efficiency of the PQRSDF resolution process could deeply impact the quality of life of the population.

Healthcare is a Fundamental Right

Improving the Health System

These improvements also impact the trust of the people in public institutions, the Health System, the health management indicators, the quality of health services, and the efficiency of the institutional procedures of the quality system.

3. The data

The following are the datasets that the Municipal Government committed to disclose:

- PQRSDF statistics consolidated from the Orfeo platform.
- List of users that attended.
- Monthly ranking of complaints.
- Tabulation of satisfaction surveys.

However, it was later revealed that such information was not available and instead other datasets would be provided. Namely, the following data was disclosed:

1. A `.xlsx` file, called `aseguramiento`, with information of around 760 PQRSDF filled virtually using Orfeo.
2. Seven `.xlsx` files from SIAU Office with information of almost 2800 PQRSDF filled in-person and not digitized into Orfeo.
3. Around 760 `.pdf` files corresponding to PQRSDF filled in-person but digitized into Orfeo by an employee.
4. A `.sql` file, called `orfeomunsog-Abril`, containing the whole Orfeo database.

These datasets were provided as a result of several meetings between the Municipal Government and the Team, where it was agreed that new data ought to be provided so the project would be fulfilled, even if its scope had to change drastically. This document already reflects that turn in the main objective of the project (see the Business Problem above).

All the files were uploaded to a shared Google Drive folder, and using a Deepnote workspace they were imported to collaborative Python notebooks where the cleaning, exploratory data analysis and feature engineering were performed.

3.1. Cleaning the data

Each dataset provided by the Municipal Government was treated as follows:

1. `aseguramiento.xlsx` file. The dataset has four columns:
    - `radicado`: unique ID associated with each PQRSDF.
    - `fecha_entrada`: The date in which the PQRSDF was filed.
    - `asunto`: Brief subject of the PQRSDF. In general, it does not provide enough information to completely understand the request.

- `direccion`: Address of the person that fills the PQRSDF.

The file does not indicate the EAPB associated with each request nor contains the subject of the PQRSDF itself. Also, this dataset is a subset of the `orfeomunsog-Abril.sql` database discussed below. We have thus cross-referenced the information of that file with this one in order to add three new columns: `municipio`, `departamento` and `pais`. These indicate, respectively, the city, department, and country of the PQRSDF. It was worth noticing that not all of these PQRSDF were filled in Sogamoso, but rather some come from Bogotá or Tunja.

2. `.xlsx` files from SIAU. We received seven files, one for each month between October 2021 to April 2022. They contain in-person requests filled in the SIAU Office that were not processed through Orfeo. The data available in these files is not consistent in format and features since it was manually inputted.

   Since the files from Oct 2021 to Jan 2022 had similar structure, they were concatenated. The same was done, separately, with the corresponding files from Feb 2022 to Apr 2022. This way, the seven `.xlsx` files were reduced to two different `pandas` dataframes.

   We noticed some atypical values in the columns of both datasets. For example, some entries that should go in one column, were in a completely different one. These cases had to be addressed individually, but fortunately they were not that many. After some more cleaning and standardization of variables for the relevant columns both of the datasets were concatenated, thus obtaining a table with 1601 rows, each one corresponding to a different PQRSDF.

   The final dataframe had the following columns: `FECHA`, `NOMBRES`, `APELLIDOS`, `No. DOCUMENTO`, `CONTACTO`, `SOLICITUD`, `AREA O DEPENDENCIA`, `RESULTADO`, `CANAL DE COMUNICACIÓN`, `ESTADO`, `MONITOREO`, `EAPB`, `PQRSDF`, `GENERO`. Some of these were discarded and renamed when joining with the other dataframe below.

3. `.pdf` files. Using the libraries `pdfminer.high_level` and `os`, and several functions based in `for` loops (one for each folder containing the `.pdf` files), the text contained in the files was parsed into strings that then were splitted into rows for a `pandas` dataframe. This led to a dataset with 761 entries and having the columns `ID`, `DATE`, `SUBJECT`, `COUNTRY`, `ADDRESS`, `PQRSDF`.

   As with the first file, these PQRSDF are also registered in the `orfeomunsog-Abril.sql file`, so by cross-referencing with that information we were able to determine that all of these requests were submitted in Sogamoso. Hence, we added columns `MUNICIPIO` and `DEPARTAMENTO` with all entries being Sogamoso and Boyacá, respectively.

However, we also noticed that some addresses and the text of the PQRSDF were not registered in the `.sql` file, so the parsing of the `.pdf` files was indeed fruitful since it brought more information.

4. `orfeomunsog-Abril.sql` file. This database was huge (2.6Gb size) and had a lot of non-useful information. We had to explore for a while all the information (with queries ran in PostgreSQL) in order to get useful tables. After some exploration of that data, we realized that only a small percentage of the entries corresponded to PQRSDF for the SMS (that is, related to the Health System) and those were already included either in the `.pdf` files or the `aseguramiento.xlsx` file. Furthermore, the request itself (that is, the text) was not included in this database, so actually the `.pdf` files contained more information, as previously mentioned. Thus, we decided to only recover the information regarding location (city, department and country) and discard the rest of the file.

Hence, after deciding that only three of the four datasets were worthy of our attention and having performed some data cleaning, we joined them. This led to a `pandas` dataframe with 4324 rows.

This dataset contains the following columns:

- `RADICADO`: Unique ID of the request. The PQRSDF from SIAU Office do not have this ID. Type: `int`.
- `FECHA`: Date of creation of the request. Type: `string`.
- `ASUNTO`: General description of the PQRSDF. Type: `string`.
- `DIRECCION`: Address of the person that filled the request. Type: `string`.
- `MUNICIPIO`: City from which the person filled the request. Type: `string`.
- `DEPARTAMENTO`: Department, region or state from which the person filled the request. Type: `string`.
- `PAIS`: Country from which the person filled the request. Type: `string`.
- `TEXTO_PQRSDF`: Text with the PQRSDF. Type: `string`.
- `CANAL_DE_COMUNICACION`: Way in which the person filled the request ("Orfeo" for virtual or `.pdf` file, and "Presencial" or "Telefonico" for some SIAU Office – the remaining are `NaN`). Type: `string`.
- `GENERO`: Gender of the user. Type: `string`.
- `AREA_O_DEPENDENCIA`: Specific area within the SMS that should attend the request. Type: `string`.
- `RESULTADO`: Answer to the request. Type: `string`.
- `ESTADO`: Status of the request (finished, in progress, etc.) Type: `string`.
- `MONITOREO`: When the request is in progress, it indicates the current status. Type: `string`.
- `EAPB`: Benefit Plan Administration Entities. Type: `string`.
- `TIPO_DE_PQRSDF`: Whether the PQRSDF is a petition, complaint, claim, suggestion, denunciation, or compliment. Type: `string`.
- `EDAD`: Age of the person that filled the request. Type: `int`.

- `PREFERENCIAL`: Whether the person is an adult, child, or is disabled. Type: `string`.
- `NACIONALIDAD`: Whether the person is a migrant or a national citizen. Type: `string`.

In Diagram 1 below we present a summary of how the final unified dataset was built.

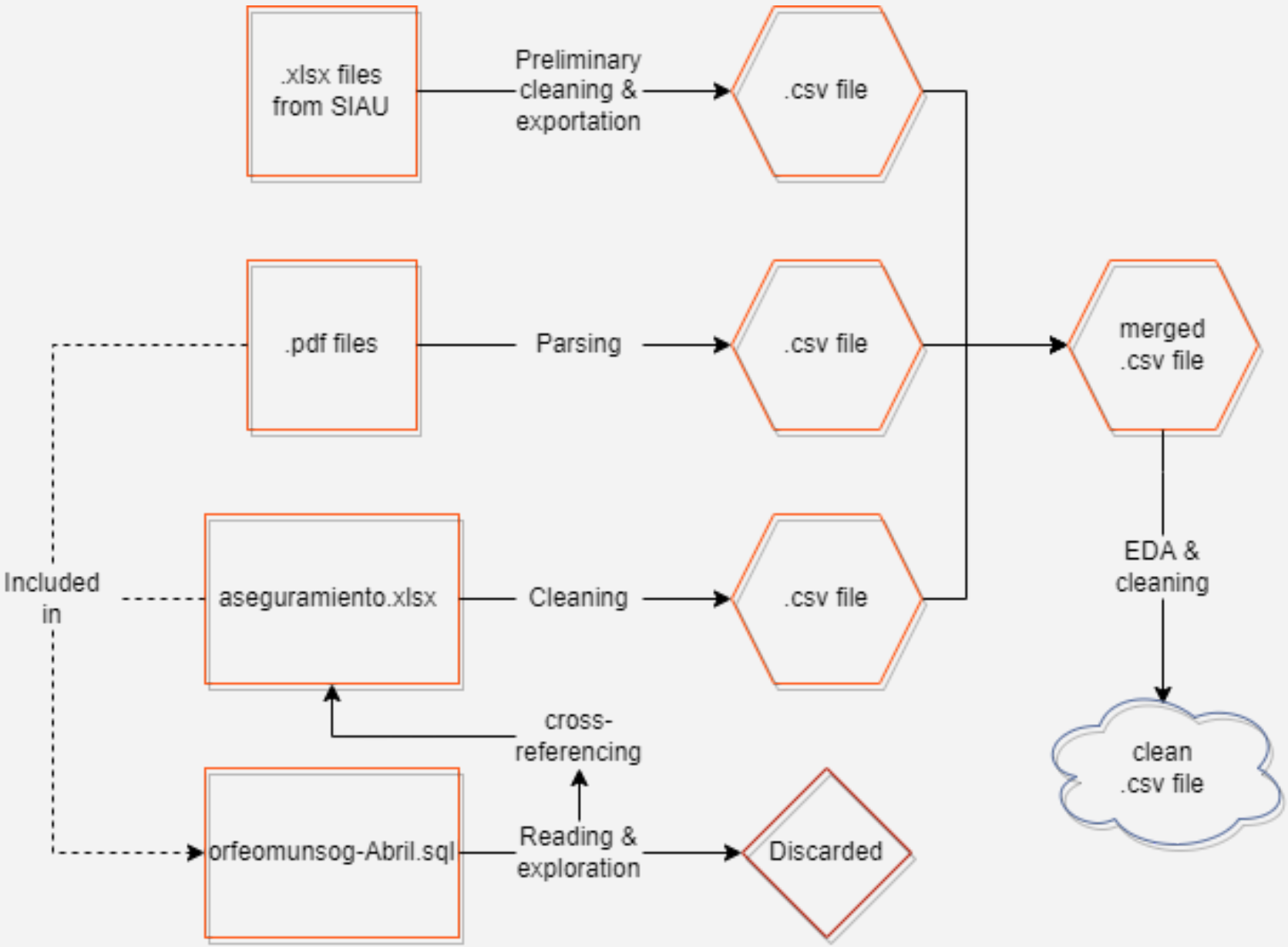


Diagram 1. Consolidation of the unified dataset.

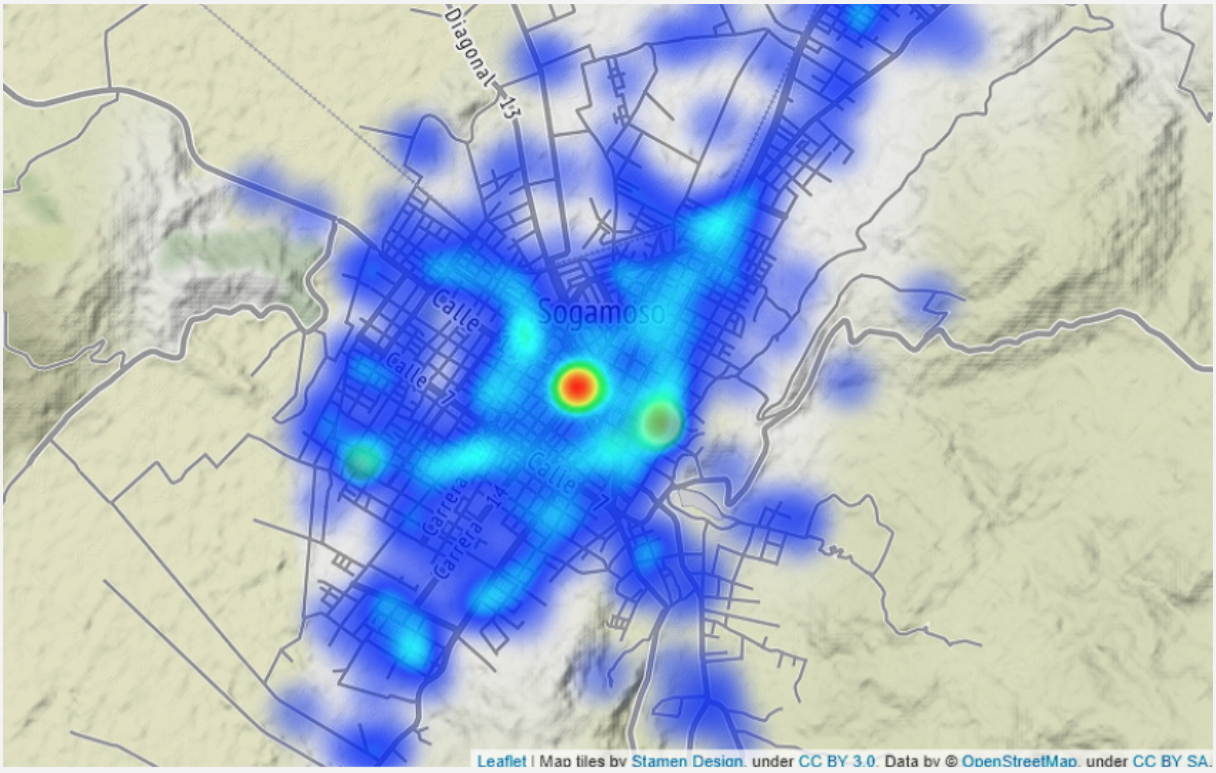
3.2. Improving the data

After merging all our datasets, we realized that the info could be improved in several ways before the exploratory analysis.

3.2.1. Georeferencing

Using the `googlemaps` library we connected Google Maps' API for georeferencing to our workspace so we could get coordinates for those PQRSDF having a non-NaN address associated with them. Thus, around 55% of the requests now have latitude/longitude coordinates.

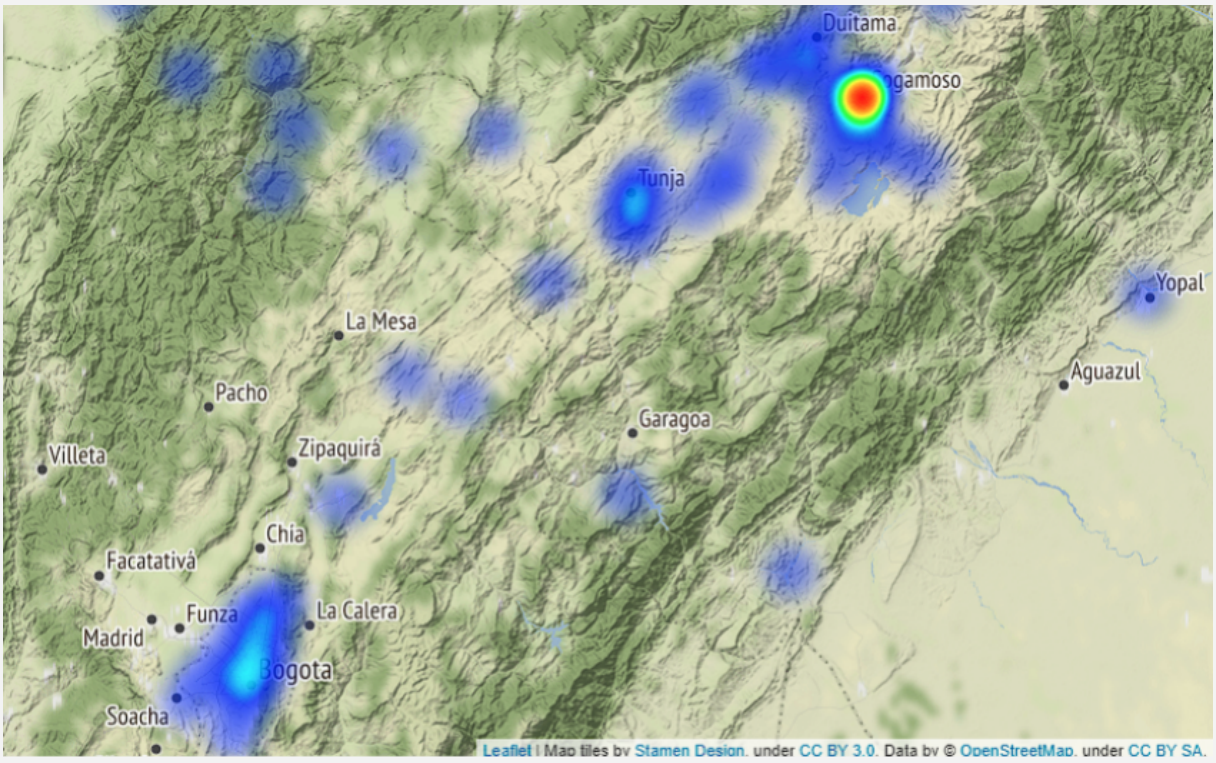
In Map 1 below we used this new data to visualize the geographical distribution of PQRSDF in Sogamoso. This map was done with the `folium` library.



Map 1. Distribution of PQRSDF in Sogamoso, Boyacá.

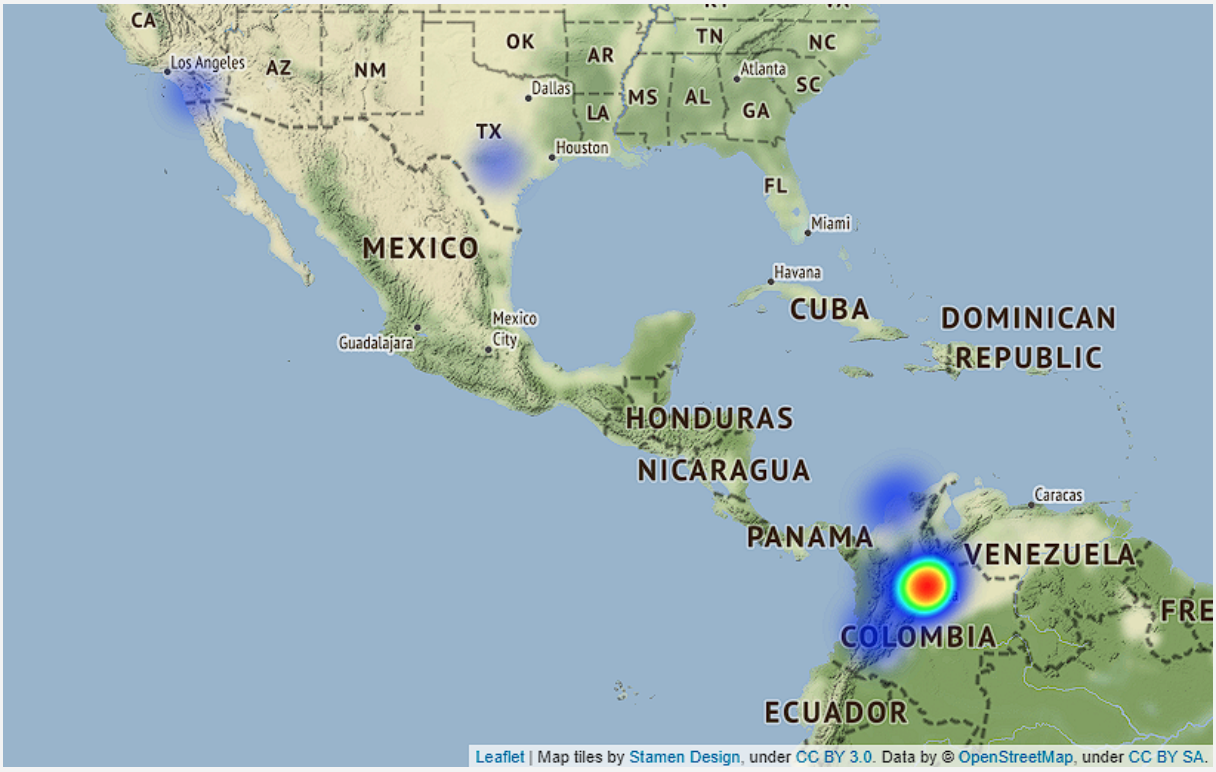
It is worth remarking that the big red point in the middle corresponds to those PQRSDF having only "Sogamoso, Boyacá" as address (so Google Maps assigned by default the coordinates of the city itself).

Also, as previously mentioned, not all PQRSDF were filled in Sogamoso, but rather some requests come from other cities/towns in the region (e.g. Bogotá, Tunja, Yopal), as we can see in Map 2.



Map 2. Distribution of PQRSDF in the Cundiboyacense region of Colombia.

Furthermore, there are some PQRSDF that were filled from California and Texas, in the USA.



Map 3. Distribution of PQRSDF in the American continent.

3.2.2. Filling empty entries and correcting corrupted values

In order to have as many non-NaN values as possible for this final database, we performed the following tasks:

1. Correcting and filling the DATE column: This column required laborious work, since dates through all files had different formats and thus, when merging, they were not standardized. To fix this, we had to modify dates directly in some of the .xlsx files, and re-concatenate all files. The final formatting decided for this column was YYYY-MM-DD HH:MM:SS.

   Furthermore, in order to fill the empty values (again, coming from the .xlsx files) in this column, we made some plots to visualize the behavior of the PQRSDF across time and then performed a simple interpolation that filled each empty entry with the last non-null date. This made sense, since after analyzing the .xlsx files we realized that these empty entries corresponded to the following situation: the person(s) filling these tables manually wanted to save some time, so they only introduced the date for the first PQRSDF of the day, leaving all others empty. Hence, all empty values corresponded to the last non-null value registered above them.

2. Filling the EAPB column: Since most PQRSDF do not have an EAPB associated, we consulted an official open-data list of all EAPB that have operated in the city. Then we ran a matching function through the TEXTO PQRSDF column (which contains the request itself) to detect which entries mentioned the EAPB in them. Thus, if the EAPB column had a NaN value on it, it was replaced by the matched EAPB. Prior to this process 87% of the requests did not have an EAPB associated. After performing it, only 70% have NaN values in this column, which means almost 310 entries were replaced with the matched EAPB.

3. Filling the GENERO column: Similar to the previous situation, we ran a matching function throughout the TEXTO PEQRSDF column detecting the expressions "usuario" and "usuaria" for male and female users, respectively. Thus, if the GENERO column had a NaN value on it, it was replaced by the matched gender. Prior to this process 80% of the requests did not have a gender associated with the person filing the request. After it, only 66% have NaN values in this column, which means almost 600 entries were replaced with the matched gender.

3.3. Exploring the data

To understand the behavior of the data and draw some early conclusions we made different types of plots. Firstly, we performed visualizations for some univariate distributions, then we plotted the amount of PQRSDF filled across time, and finally we made some natural language processing to get word-clouds.

3.3.1. Behavior of variables

Type of PQRSF. In Figure 1 it can be seen that most PQRSDF correspond to suggestions (S), followed by petitions (P). However we strongly suspect that some of these supposed suggestions are actually petitions, and that the information might have been manually entered as "solicitudes" (requests). We can also see that there are a lot of complaints (Q).

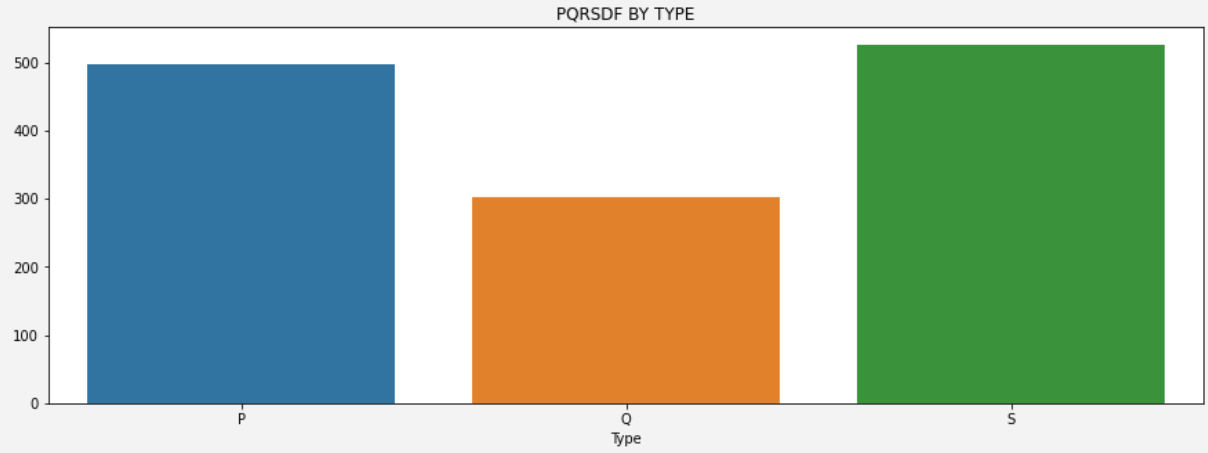


Figure 1. Type of PQRSDF.

Communication channel. In Figure 2 we can see that most PQRSDF are managed in person, some others are managed virtually on the Orfeo platform, and just a couple of them are submitted by phone.

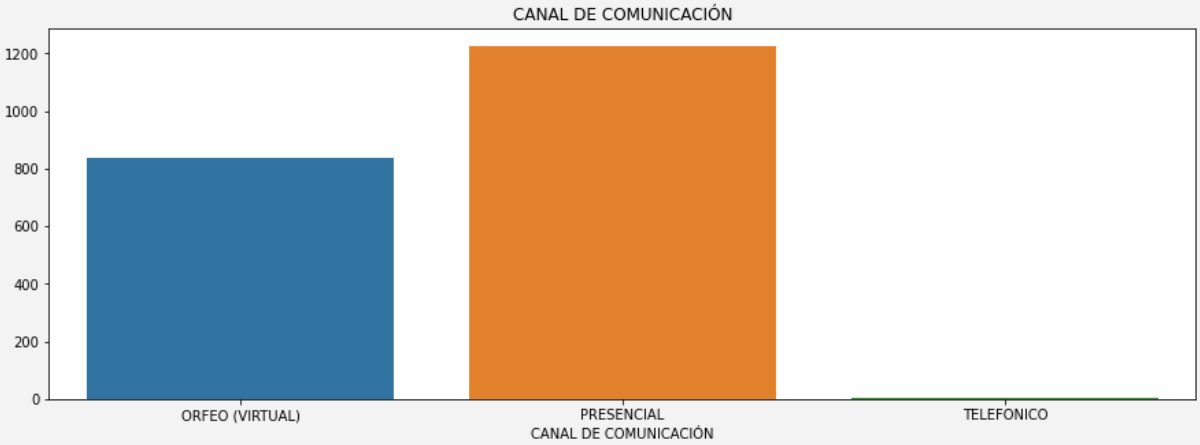


Figure 2. Communication channel.

Status of the PQRSDF. As might be expected, Figure 3 shows that most PQRSDF have been completely solved, and as of April 2022 a bit more than 200 PQRSDF had a follow-up.

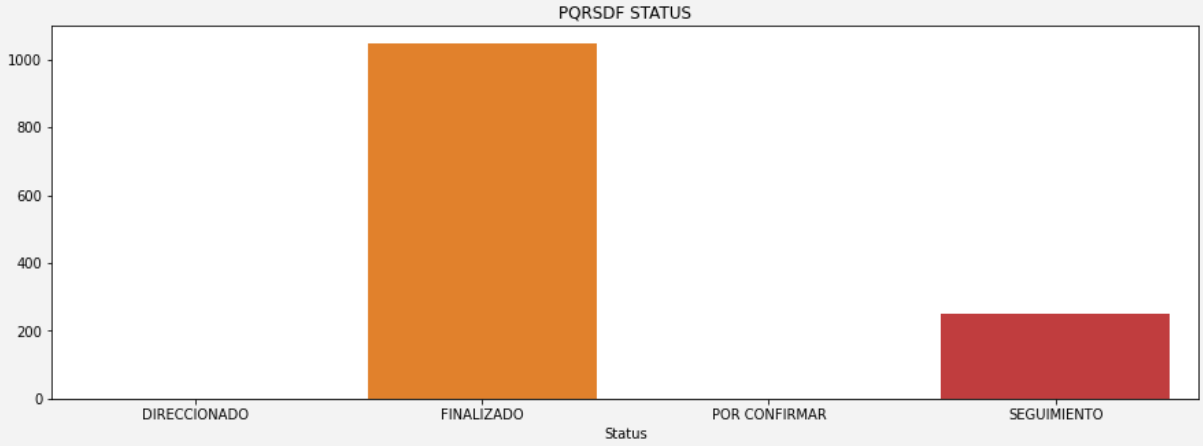


Figure 3. Status of the PQRSDF.

User gender. Figure 4 shows how the PQRSDF are distributed by gender. In spite of having a lot of people who do not declare their gender (NaN values), it would be interesting to find out why women seem to make more requests in Sogamoso. Maybe they get sick more often or they have certain needs different from those of the male population (e.g. pregnancy).

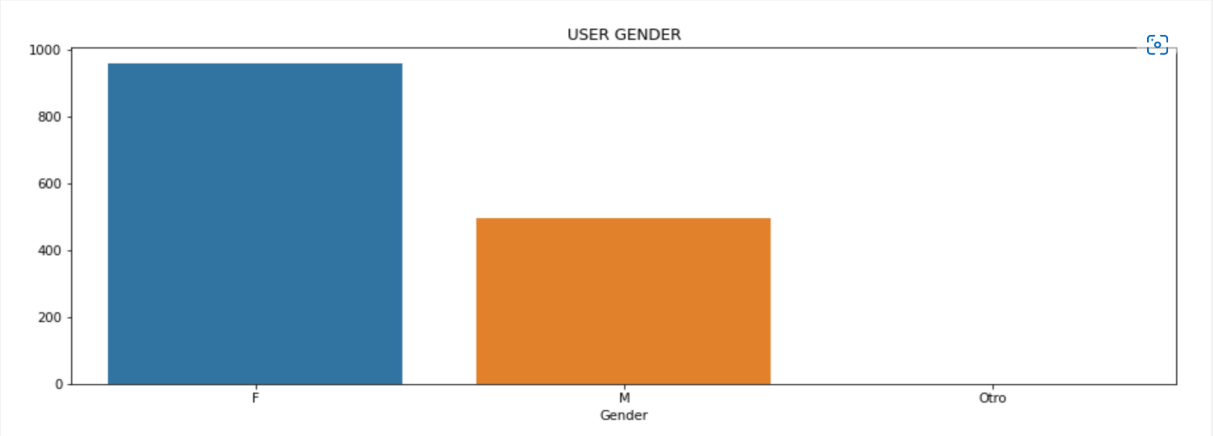


Figure 4. User gender.

Requests by EAPB. Figure 5 is probably the most relevant plot we have. It shows how the PQRSDFs are distributed by EAPB. Out of 1601 requests, "NUEVA EPS" has about 300 PQRSDFs.

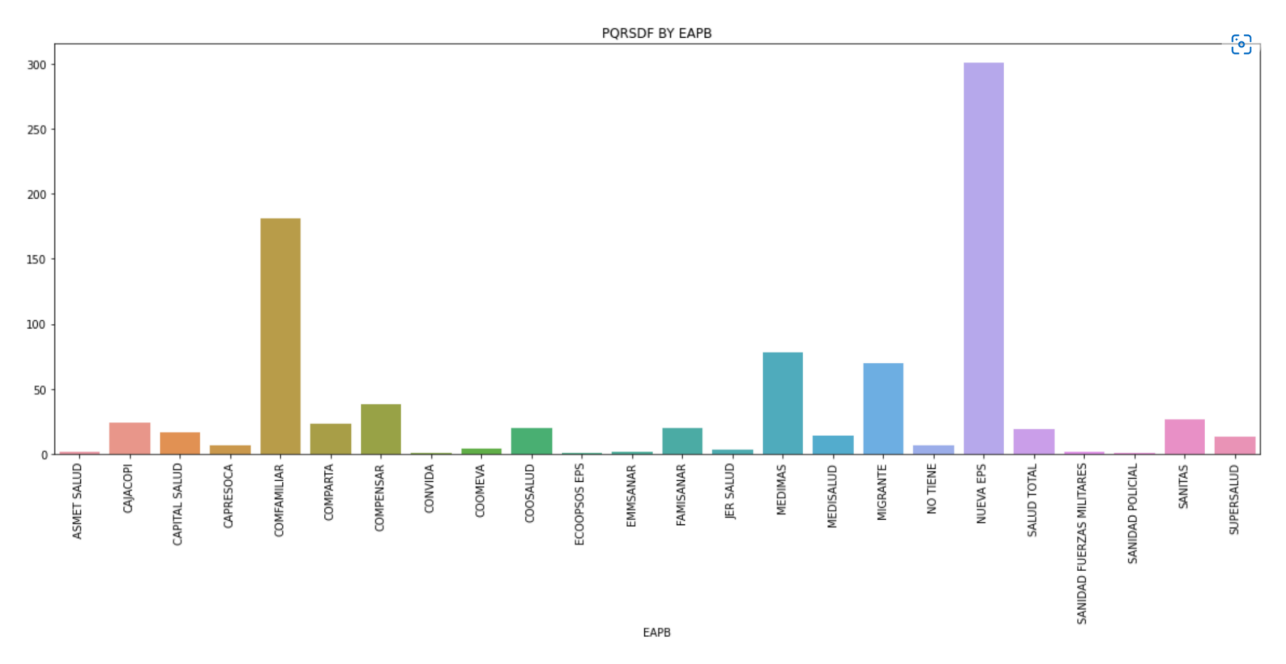


Figure 5. PQRSDF by EAPB.

Also, Figure 6 below indicates how the type of PQRSDF is distributed for each EAPB. It can be seen that COMFAMILIAR has a lot more complaints (Q) than petitions (P), and NUEVA EPS has a similar number for both of them. These two EAPB together with MEDIMAS seem to be the ones that fail the most at providing good health care services. Although the reason for this might just be that these are the EAPBs with the highest proportion of users.

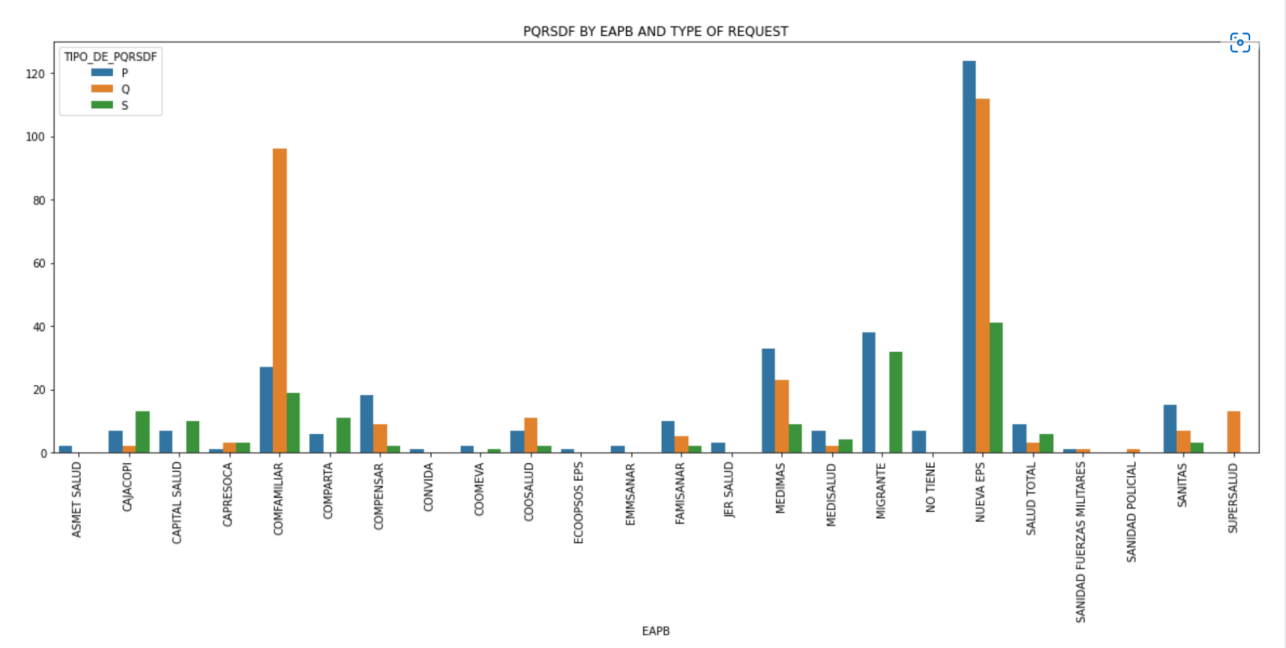


Figure 6. Type of PQRSDF by EAPB.

3.3.2. Behavior across time

In Figure 7 we can clearly see a reduction in the total number of PQRSDFs in 2020 and a quite steep increase for 2021. This is certainly due to the COVID-19 pandemic: In 2020 people were afraid of going to the hospital and using the Health System for non-urgent matters, so probably medical appointments together with petitions and complaints to the EAPBs piled up to the following year.

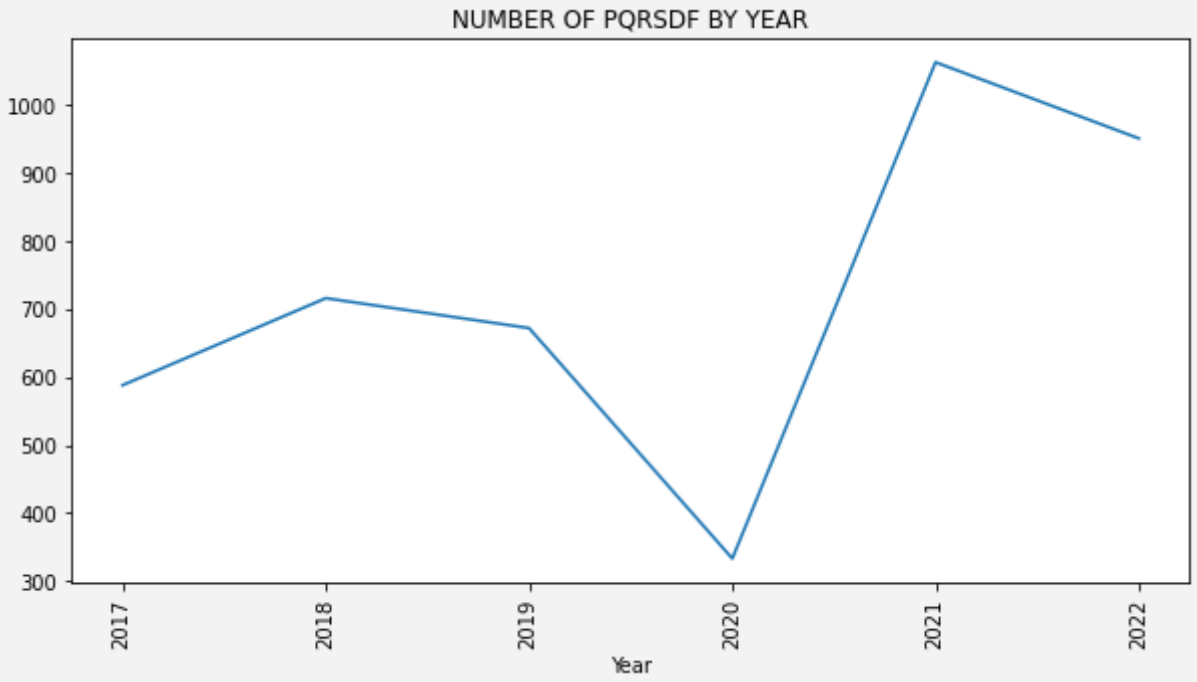


Figure 7. PQRSDF by year.

Now, in Figure 8 below we can see some interesting trends:

- In general, the periods March-April and October-December have more PQRSDF filled, while there are valleys mostly in June-Aug. This might be due to the fact that during holidays at the start and middle of the year people might be traveling or resting, and thus they tend to use the Health System less.
- There is a valley of almost no PQRSDF in the first months of 2020 and a huge increase in requests for August of 2020. As said previously, this is heavily related to the COVID-19 pandemic. The spike in August might be due to the fact that by that time people were not that scared of the virus and started using the Health System again for non-urgent matters.
- There are some PQRSDFs reported in the future (e.g. December 2022). This might be caused by typos in the `date` column. They were not corrected throughout the cleaning process, because we do not have a way of knowing what the actual dates are.

It is worth remarking that the huge spikes from October 2021 to mid 2022 correspond to the PQRSDF coming from SIAU files, which decompensate the distribution. That is why Figure 9 is useful, since most of the PQRSDF labeled as “in-person” and “telephonic” are those filled at the SIAU Office. Hence we can see that Orfeo requests (i.e., those labeled as “virtual”) have a more regular trend during this period.

Finally, Figure 10 confirms the behavior by months described above.

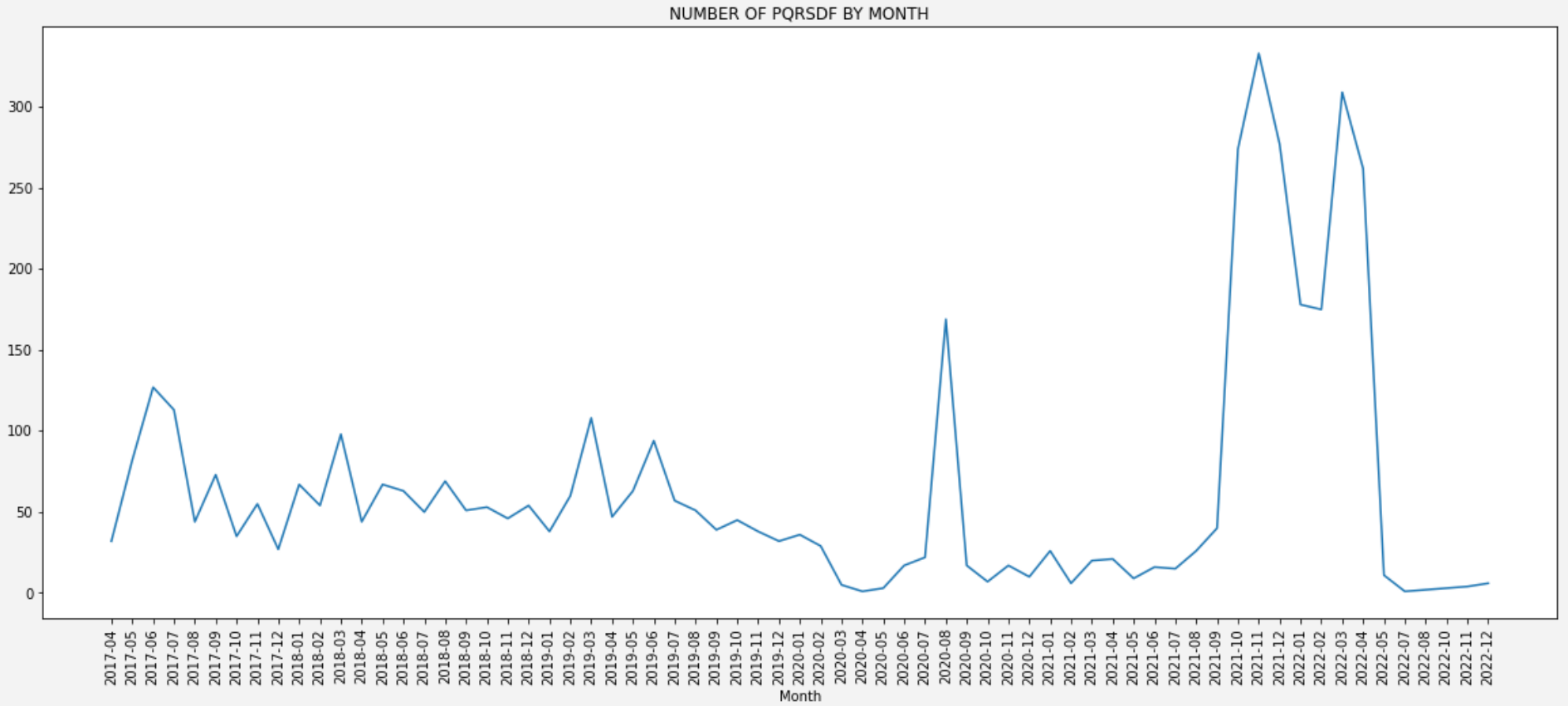


Figure 8. PQRSDF by month.

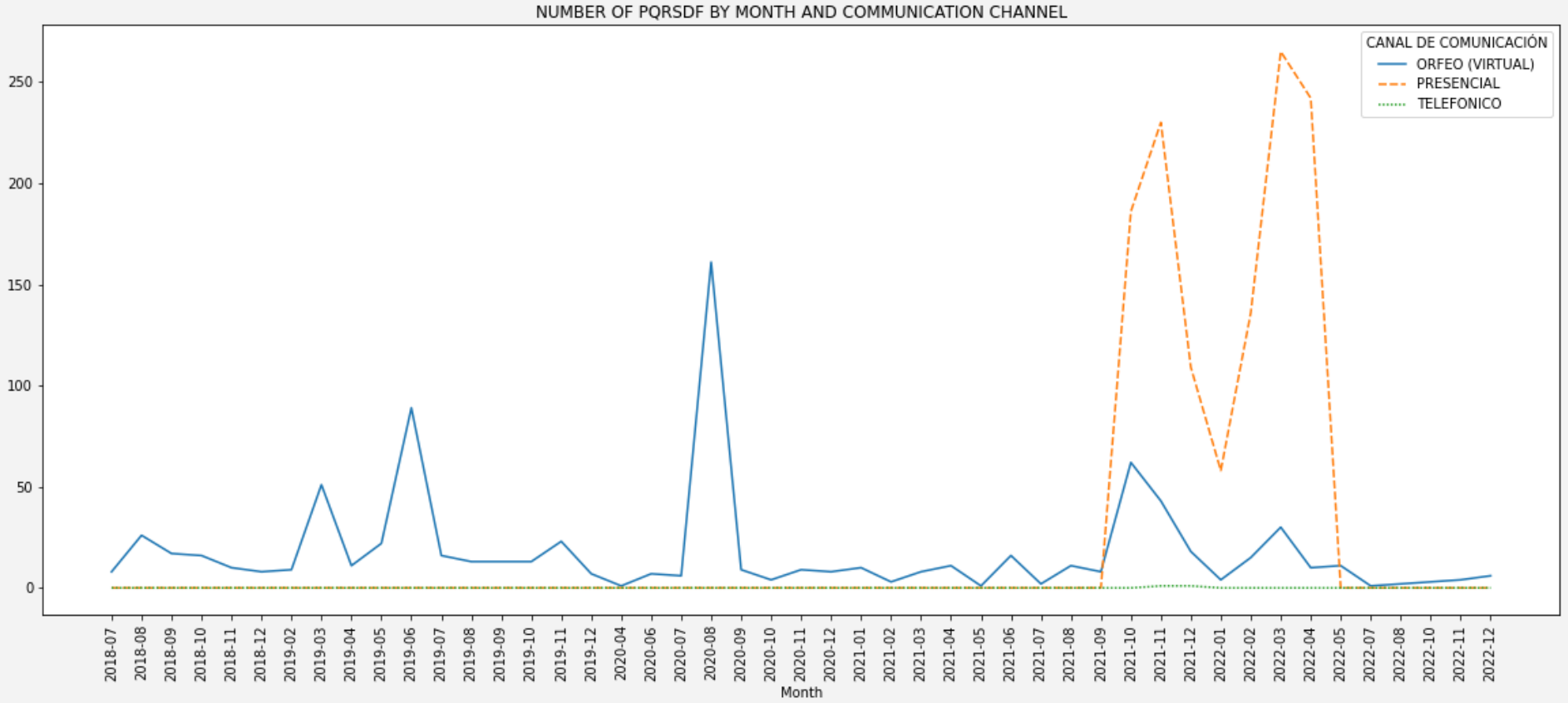


Figure 9. PQRSDF by month and communication channel.

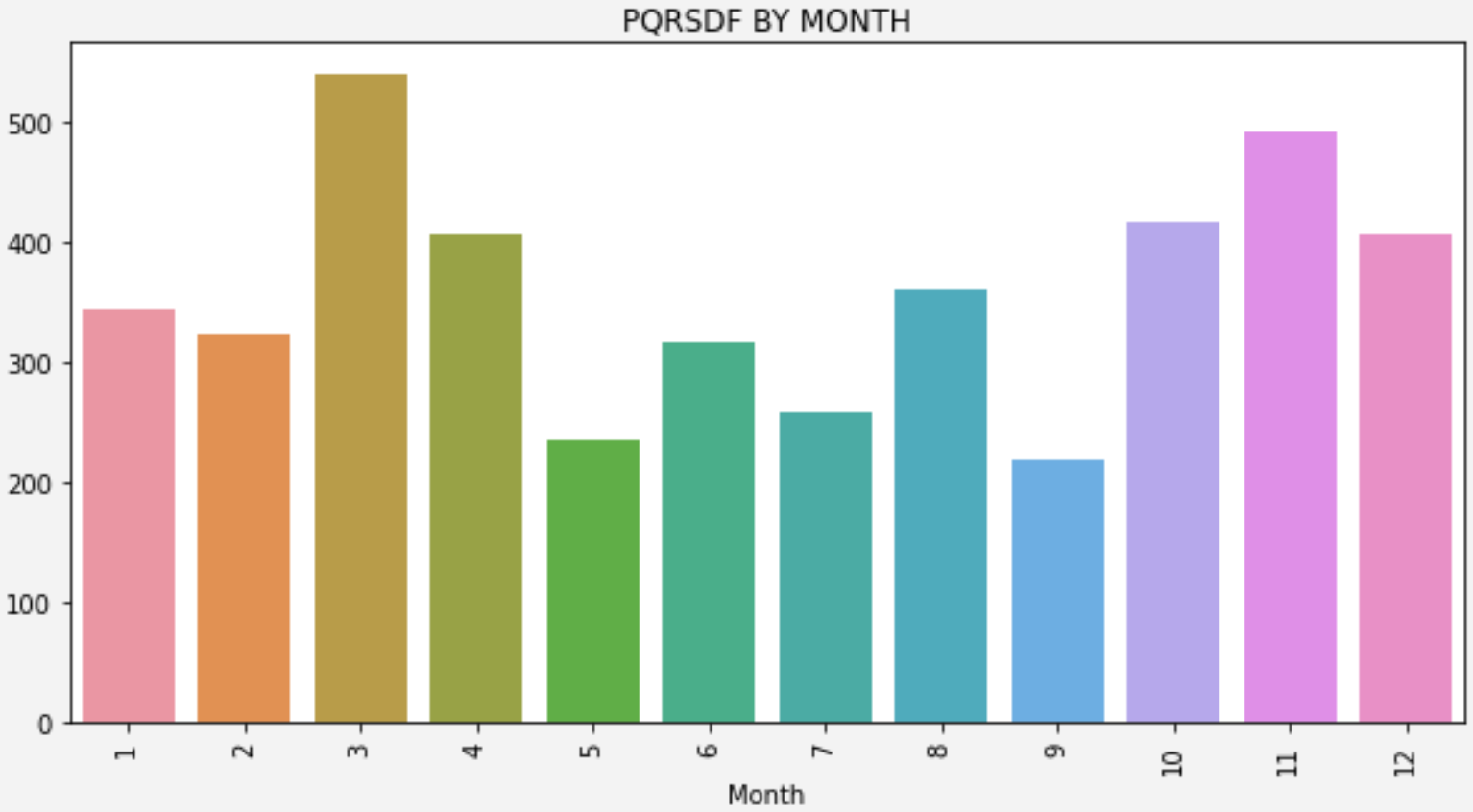


Figure 10. PQRSDF by months of the year.

3.3.3. Text processing

Some basic text analysis was done through word-clouds, by using the library `wordcloud`. But before making the graphs, some non-useful words were dropped (eg. *sincerely, date, month, day, sign,* etc.). In addition, stopwords that are widely used in Spanish were removed. The general result is Figure 11.

From this graph we see that most PQRSDF have to do with medicaments, requests for visits, authorizations, the EAPB "Nueva EPS", "Comfamiliar" and "Medimás" (which coincides with our findings in Figure 5 above), dissatisfaction with the service, and information requests.

Figure 11. Word-cloud of text of PQRSDF.

We also made word-clouds for PQRSDF submitted by women and men separately to see if we could detect something that would help explain why women make more requests (see Figure 4). In particular, we were looking for diseases and medical specialties specific to women that could require improvements from the health system in Sogamoso. However, Figure 12 and Figure 13 do not show important differences, and as in the one from Figure 11, PQRSDF related to medicaments, requests for medical appointments, procedures and authorizations, dissatisfaction, information requests, and some EAPB are the ones that stand out the most.

Figure 12. Word-cloud of text of PQRSDF submitted by women.

Figure 13. Word-cloud of text of PQRSDF submitted by men.

4. The model

As we previously stated, in the consolidated dataset there were a lot of `NaN` in most columns, and even though we filled some of those entries, one of the main features that we could not fill through feature engineering was the type of the PQRSDF (that is, if a request is P, Q, R, etc.). For this reason, we decided to try out a classification model, meaning that we wanted to take any PQRSDF text and predict its type (column `TIPO_DE_PQRSDF`). However, since we only had values P, Q, and S in this column, any model will only classify into these types.

To do so, we took the column `TEXTO_PQRSDF` and started by removing unwanted characters and standardizing the text. Specifically, the text was set to lowercase and every symbol that was not a letter or a space was removed. Then, we vectorized this column to convert the text into a numerical form that could be modeled.

4.1. Logistic regression

First we tried a logistic regression model. A training and testing process was performed by separating the data in two subsets. First, we tried out a non-stratified standard 80%-20% split with default parameters ($C = 1$ and 'penalty'='L2'), obtaining an accuracy of 66.03%. However, since the objective was to perform a classification of categories, we thought that it was better to split the data with stratification, so both the training and testing sets would have their categories in the same proportions. As expected, we obtained a slight increase in the accuracy: 67.56%.

Then, we moved on to real modeling by applying cross-validation and hyperparameter tuning. From the same 80%-20% split we used before, we took the training set and applied k-fold cross-validation with different values for k, deciding that our best choice was $k = 5$. At the same time, by means of scikit-learn's `GridSearchCV`, we build a grid of different values of the hyperparameters C and 'penalty', so we could fit the model for several combinations of them and get the best possible accuracy. We obtained that the best model had the parameters $C = 0.4394$ and 'penalty'='L2', and would give an accuracy of 66.89%.

After this training process, we applied the model to the test set and predicted the type of PQRSDF. With this data set, that had not been previously seen by the model, we obtained an accuracy of 67.56%, which is not very different from the one obtained during the training, so we could say that our model

performs in a consistent way and would give an accuracy of about 67% with new data.

Subsequently, we used the confusion matrix to better understand the behavior of our model. We obtained the following matrix:

	P	Q	S
P	70	9	21
Q	10	41	9
S	27	9	66

What this shows is that:

- 9 PQRSDF were predicted as Q and 21 as S, but they should have been predicted as P,
- 10 PQRSDF were predicted as P and 9 as S, but they should have been predicted as Q,
- 27 were predicted as P and 9 as Q, but they should have been predicted as S, and
- 70 were predicted correctly as P, 41 as Q and 66 as S.

So, it can be seen that the biggest confusion happens when classifying P and S, which makes sense because, as we mentioned above, we strongly believe that some PQRSDF classified as S (suggestions) are actually P (petitions) and this misclassification corresponds to a human error. This is definitely a problem that the SMS needs to tackle when it comes to data entry.

4.2. Random forest

Secondly, we created a random forest to see if this model could provide better results. With the same stratified train-test split that we used with the logistic model, and having vectorized the PQRSDF texts as well, we first performed a random forest model with no cross-validation and no hyperparameter tuning. This initial model achieved an accuracy of 62.98%.

Then, by using scikit-learn's `RandomizedSearchCV` (because for a random forest using `GridSearchCV` would require a lot of computational power), we trained a 3-fold cross-validation model and found that the best hyperparameters are the following: `'N_estimators': 400, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 90, 'bootstrap': False`.

With these best parameters, we obtained an accuracy of 68.04%, improving our initial random forest model by 5%, and attaining an accuracy that is a bit higher than the one of the logistic regression model. Then, we applied this tuned cross-validated model to the test set and obtained an accuracy of 67.18%.

Although this results are better than the ones obtained with the logistic model, the difference of 0.86% between the training and testing accuracy could mean that this random forest model would not be as consistent with new unseen data as the logistic model, for which we obtained a difference of only 0.67% between training and testing accuracy.

We also calculated the confusion matrix for this model:

	P	**Q**	**S**
P	67	10	23
Q	6	46	8
S	28	11	63

We can see that the random forest model predicts Q a little better, however the logistic model is better at predicting P and S. This fact, together with the consistency in accuracy, led us to choose the logistic model as the most optimal one.

Nevertheless, the random forest helped us obtain the n-grams that were most important for the predictions, and this gave us an idea, not only of the most used words that we saw in the world-clouds above (Figures 11-13), but also of those that contribute the most to the predictive capacity of the model: '*Busca*', '*inconformidad*', '*manifiesta*', '*informacion*', '*solicita*', and '*medicamentos*' are the most important ones (see Figure 14 below). In other words, we are able to say that these n-grams had the most impact on the classification of the requests, P, Q or S.

	importance
busca	0.070131
inconformidad	0.063112
manifiesta	0.054122
informacion	0.043597
solicita	0.026106
medicamentos	0.020318
parte	0.018963
2021	0.016730
asignada	0.016457
entrega	0.013986
examenes	0.013623
agilidad	0.013296
agendamiento	0.013211
cita	0.011275
usuaria	0.011053
asignacion	0.010783
covid	0.010714

Figure 14. Most relevant n-grams from the random forest.

4.3. Predictive tool

Finally, and as we mentioned before, we chose the logistic regression model as our best model. So we used it on our original data set to predict the missing values in the column `TIPO_DE_PQRSDF`. We managed to predict on 766 PQRSDF

that were not originally classified, and by filling all of those null cells, we obtained the following distribution of PQRSDF:

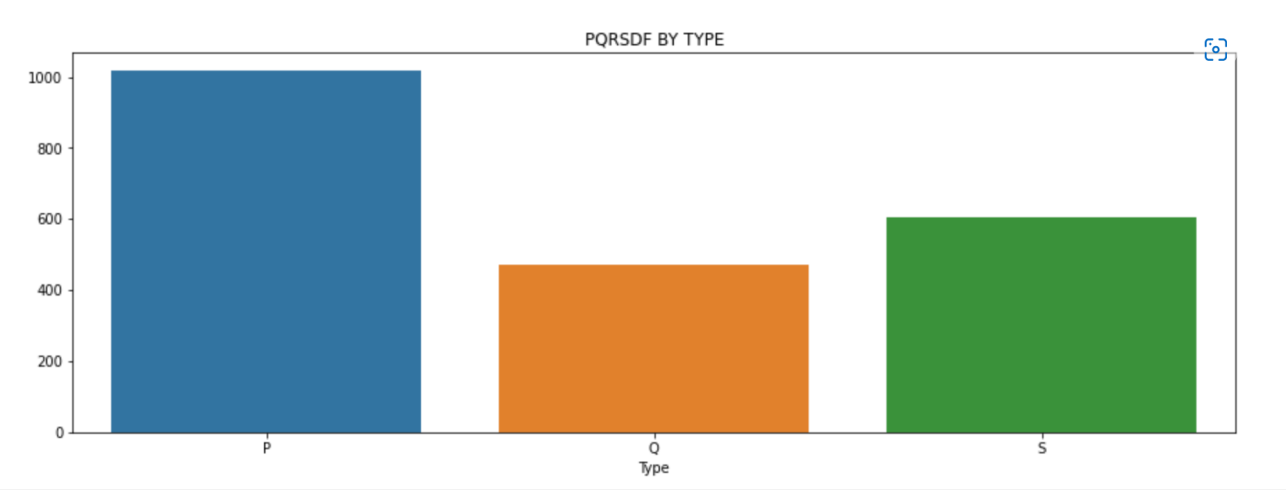


Figure 15. Type of PQRSDF after predicting null values.

If we compare Figure 15 to Figure 1, we can see that the proportion of P is much higher, while Q continues to be the least common type of PQRSDF.

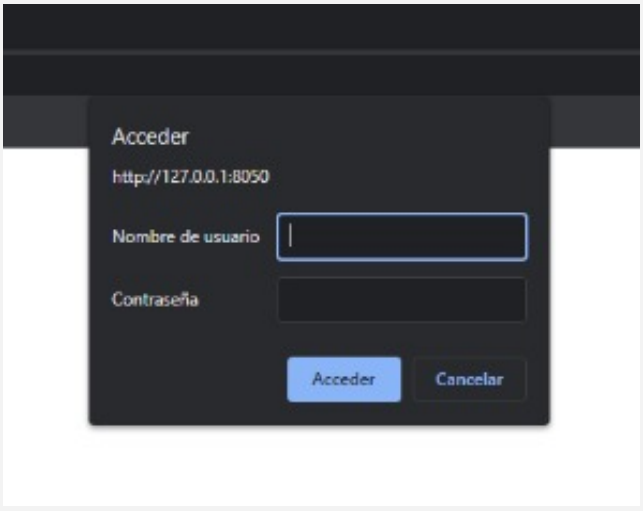
Additionally, with the logistic regression model and the most optimal vector conversion parameters we also created a tool that allows us to make up our own PQS and predict their type. For this we used a `for` loop that allows the user to type the desired number of PQS (even if it is just one) and uses the trained model to build a dataframe with the predicted classification for the PQS.

5. The application

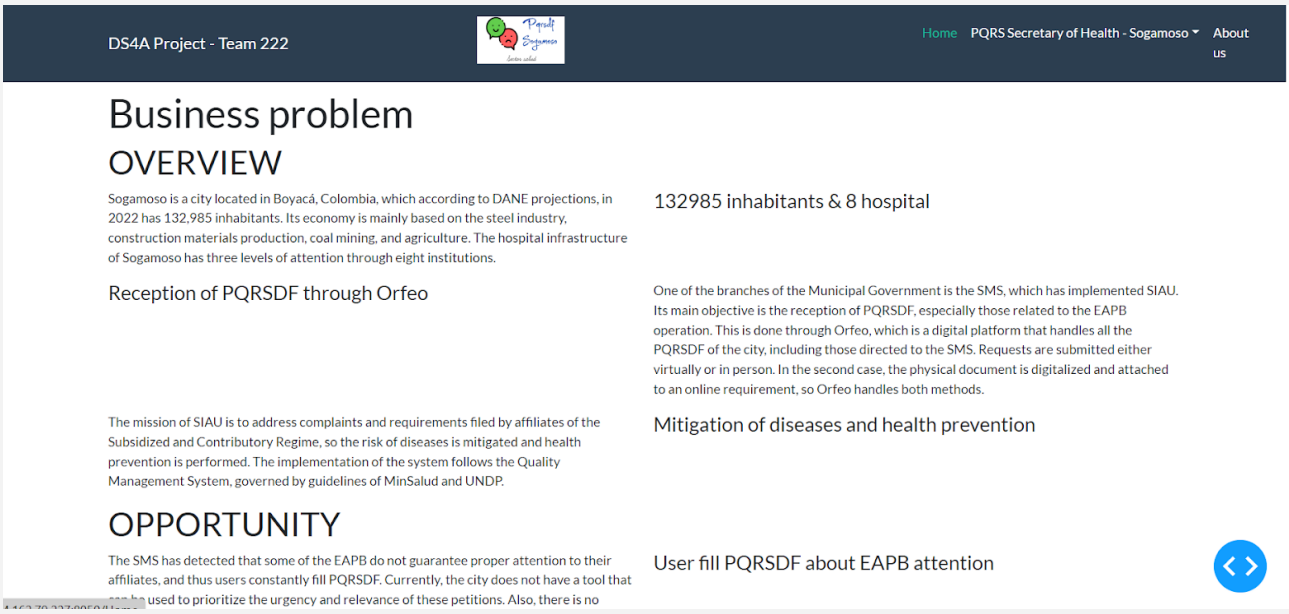
The front-end part of our project was built using VSC, where the libraries `dash`, `dash core`, `dash html`, `plotly` and `pandas` were imported and used. The final version of the dashboard has been deployed on the AWS cloud service, at the following link: http://54.162.79.227:8050/.

5.1. Dashboard

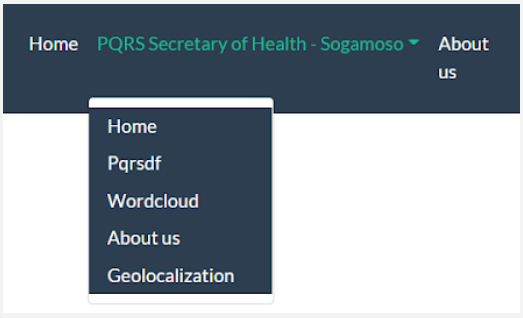
A user login is necessary to enter the website (User: user1 - Password: test1). This was implemented to give more security to our dashboard, due to the sensible information displayed.



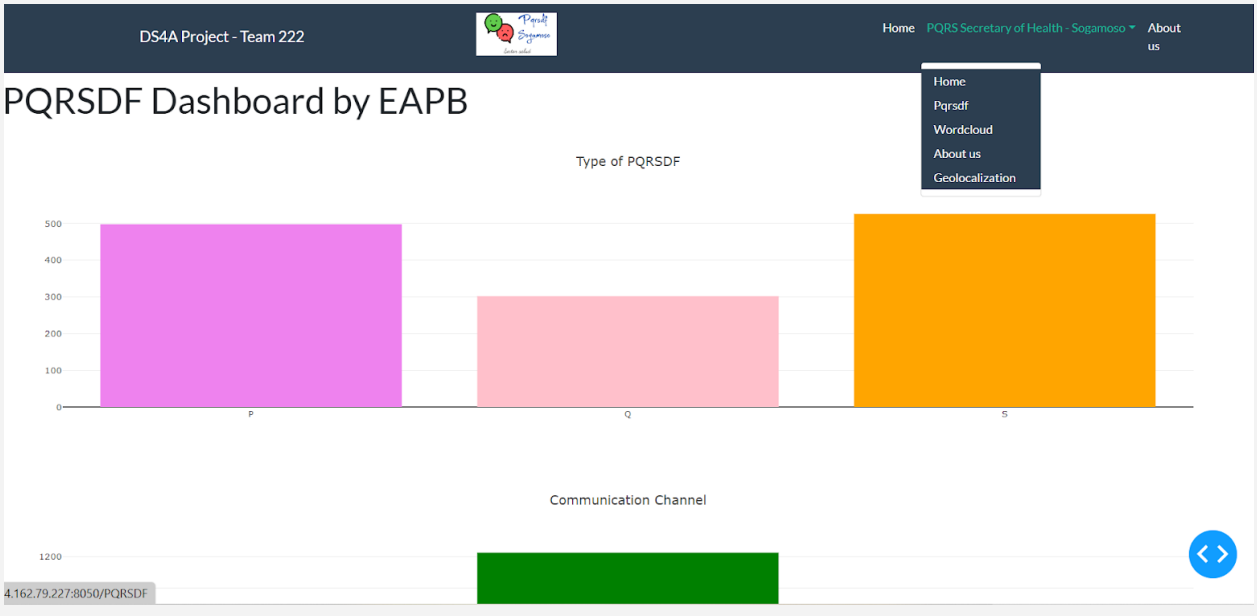
The website displays in its main home page (which can also be accessed through the button `Home`) relevant information about Sogamoso, the description of the PQRSDF attention channels, our scope in the project and finally the impact of our work (i.e., the information of Section 2 above).



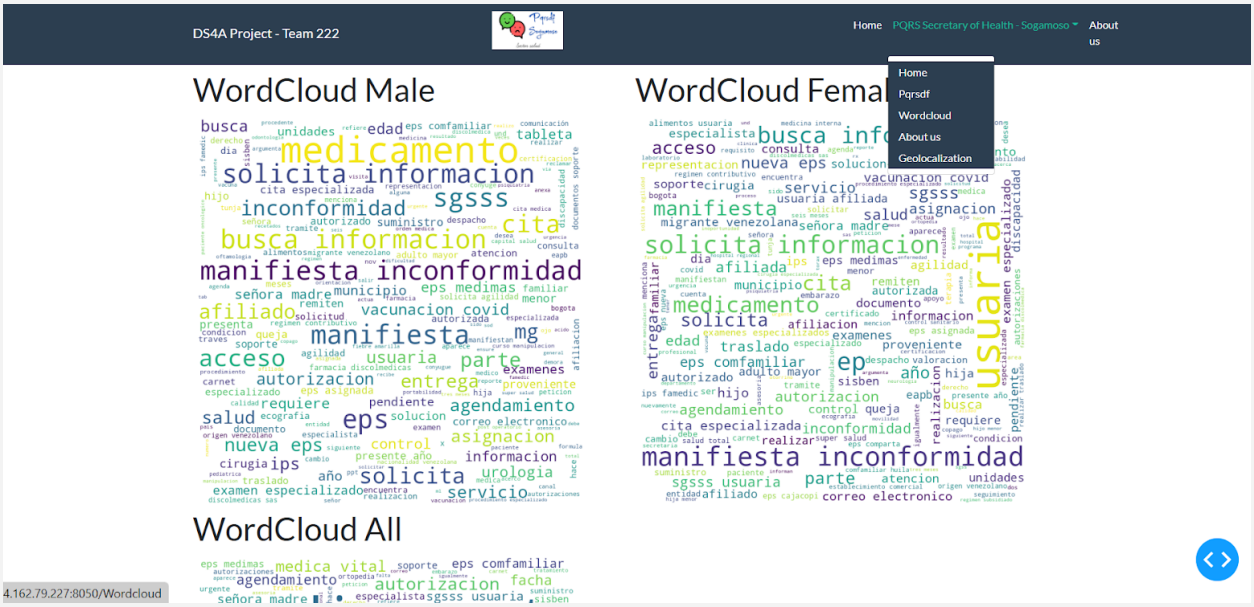
The `PQRS Secretaria de Salud Sogamoso` button has a drop-down menu that shows five more items, which can be used to navigate through the dashboard.



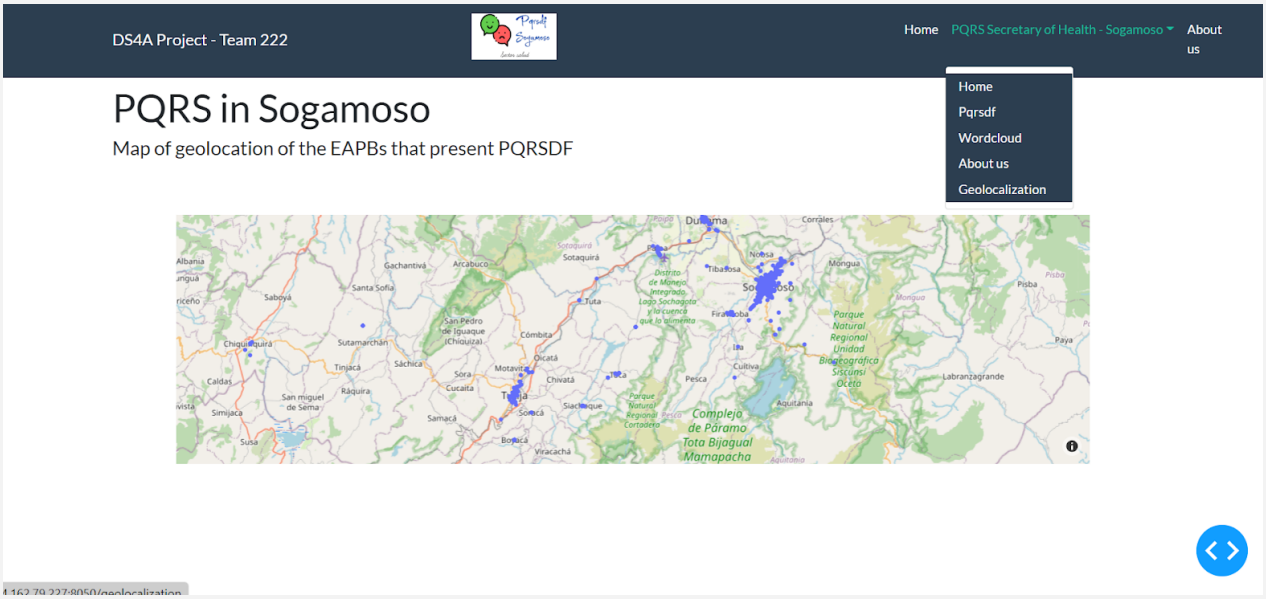
The `PQRSDF` page shows six graphs that describe the type of PQRSDF, the communication channel, gender, EAPB and finally the number of PQRSDF by EAPB. As said before, all these results have been already presented here (see Sections 3.2-3.3) and were obtained from our unified database.



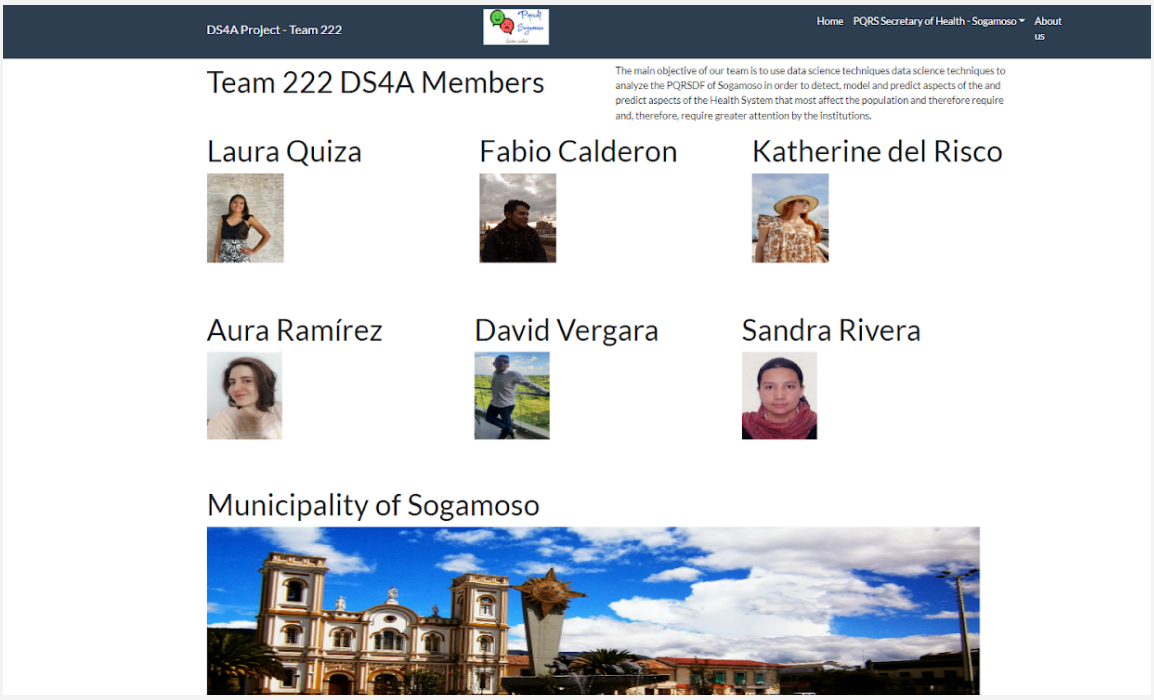
Also, inside the drop-down bar we can find a button called `Word Cloud`, which shows the three word-clouds of Figures 11-13.



As a final item in the drop-down bar, the `Geolocation` button leads to a map of Sogamoso with the coordinate points associated to each PQRSDF in Section 3.2.1.



Finally, our website also has an `About us` button, that shows a page with information of our Team and a nice photo of Sogamoso.


DS4A Project - Team 222
Home
PQRS Secretary of Health - Sogamoso
About us
Team 222 DS4A Members
The main objective of our team is to use data science techniques data science techniques to analyze the PQRSDF of Sogamoso in order to detect, model and predict aspects of the and predict aspects of the Health System that most affect the population and therefore require and, therefore, require greater attention by the institutions.
Laura Quiza
Fabio Calderon
Katherine del Risco
Aura Ramírez
David Vergara
Sandra Rivera
Municipality of Sogamoso

6. Conclusions & future work

From our work, we can draw the following conclusions and opportunities for a continuation of this project:

- The Municipal Health Secretariat (SMS) needs to improve its way of imputing the PQRSDF' information filled by the users, so the relevant data can be recorded for all requests. With this we are not saying that the Orfeo platform is not useful, but rather, that it could be improved so that the specific needs of the healthcare-related PQRSDF can be recorded. Furthemore, it is a redundant work to type manually the PQRSDF of users that contact or go to the SIAU Office into `.xlsx` files, when the entire city precisely has the Orfeo platform in place for that same purpose.
- Although we managed to fill some of the `NaN` values in most columns, this work can be pushed further. For example, the techniques used for filling the `GENERO`, `EAPB`, `TIPO_DE_PQRSDF` could be improved by using more advanced algorithms or models. Also, these can be extended to fill missing values in other columns, such as `AREA_O_DEPENDENCIA` or `CANAL_DE_COMUNICACION`.
- As with any data science project, the accuracy of the models and predictive tools strongly depends on the integrity of the dataset provided. So, even though our models did perform good with a dataset having so many `NaN` values, if the SMS had provided a more solid dataset, a higher accuracy might be reached. Also, the classification process could have been performed for all the letters (types) in the acronym PQRSDF, and not only for P, Q and S.
- Our models are not subject to the SMS, meaning that they use the information of the PQRSDF provided, but it does not matter that they are related to the Health System. Thus, any other entity within the Municipal Government of Sogamoso can use these tools to predict the type of PQRSDF received.

7. Credits

Project made using Python 3.7 in VSC, Deepnote and the DS4A Workspace. Team 222 was distributed as follows:

- Data processing & data cleaning: Aura Ramírez, Fabio Calderón, Katherin del Risco, Sandra Rivera.
- Exploratory data analysis: Aura Ramírez, David Vergara, Fabio Calderón.
- Feature engineering - parsing .pdf files: Fabio Calderón, Sandra Rivera.
- Feature engineering - Natural Language Processing: Aura Ramírez, Katherin del Risco, David Vergara.
- Feature engineering - georeferencing: Fabio Calderón.
- Modeling: David Vergara, Katherin del Risco, Aura Ramírez.
- Backend & AWS: Laura Quiza, Sandra Rivera.
- Frontend: Laura Quiza, Sandra Rivera.
- Datafolio: Katherin del Risco
- Video: Aura Ramírez, Fabio Calderón, Katherin del Risco.

Front page image extracted from:
https://commons.wikimedia.org/wiki/File:Panoramica_Sogamoso.JPG

7.1. Acknowledgements

We thank our TAs Miguel Pire and Carol Chamorro for their support and insightful comments. Also, we are thankful to Natalia Barrera for being the contact point at Sogamoso's Municipal Government and helping us overcome the difficulties with the datasets and the scope of the project.