

**MDS** Master of  
Data Science  
Universidad de Chile

PRESENTACIÓN 2 MDS7201

# ANÁLISIS EXPLORATORIO DE DATOS ENTIDADES MINSAL

DANIEL CARMONA, MARTÍN SEPÚLVEDA,  
MONSERRAT PRADO, CAMILO CARVAJAL

ENTIDADES MINSAL

# TABLA DE CONTENIDO

**/1**

INTRODUCCIÓN

**/2**

DATOS

**/3**

ANÁLISIS EXPLORATORIO DE  
DATOS

**/4**

CONCLUSIONES Y TRABAJO  
FUTURO

**/5**

REFERENCIAS

# INTRODUCCIÓN

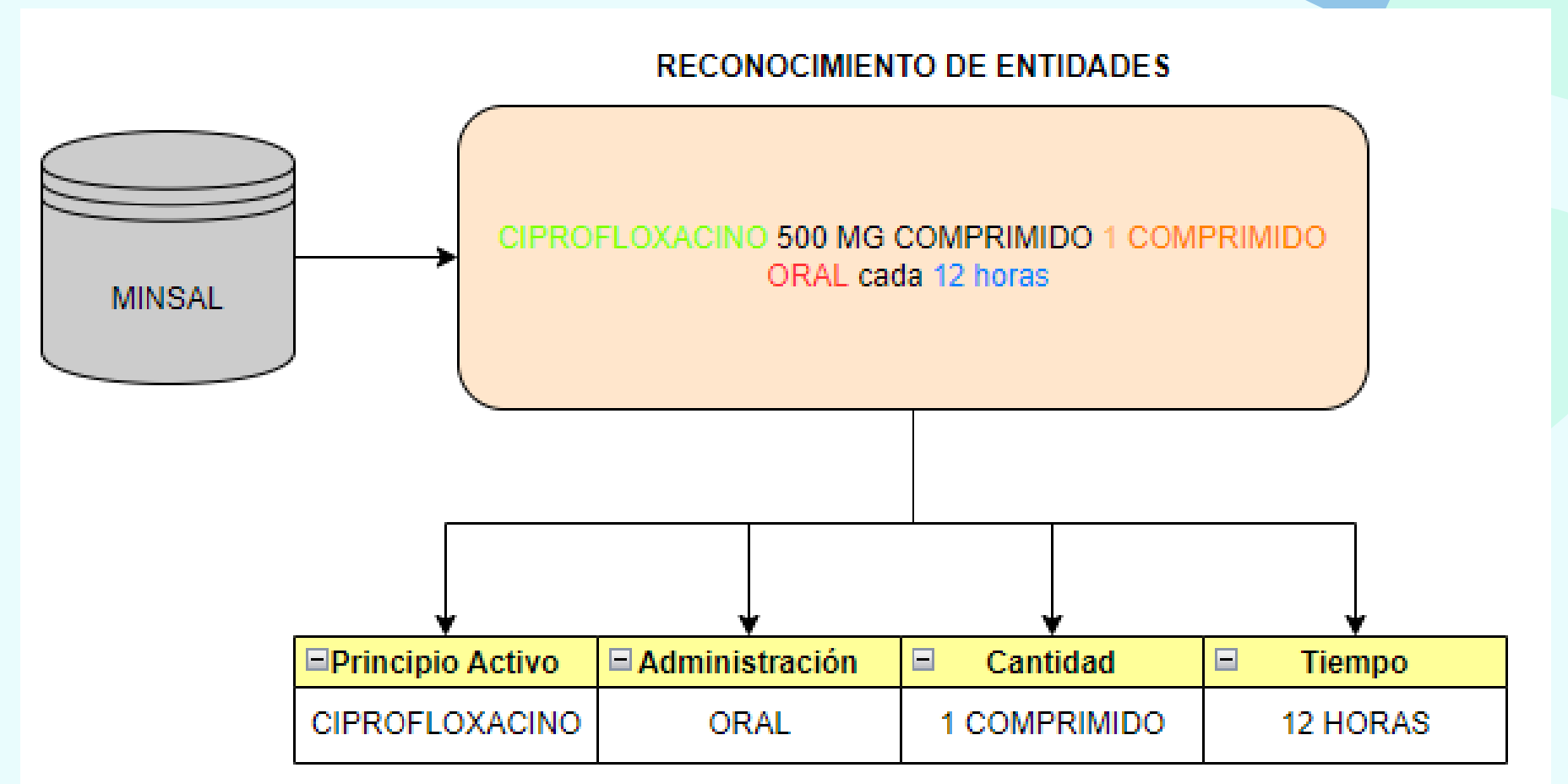
- Existen recetas médicas que carecen de cierta información importante.
- Esto puede llevar a errores de medicación y a un empeoramiento en el estado del paciente.
- Las recetas electrónicas pueden contener campos de texto libre.
- Esto dificulta la verificación de la completitud de la prescripción.



# INTRODUCCIÓN

## DESCRIPCIÓN DEL PROYECTO

- Dado un campo de texto libre, utilizar algoritmos de NLP para reconocer entidades y de esta manera completar columnas de manera automática en los datos de un paciente.
- Detectar errores de completitud o gramática en las indicaciones.
- Refraseo de la información para evitar errores de administración de medicamentos.



# DATOS

- 1.5 [M] de recetas médicas, con un total de 20 atributos por cada una, tales como:
  - fecha prescripción
  - id paciente
  - especialidad
  - diagnostico
  - medicamentos
  - indicaciones de administración
- En ciertos atributos se cuenta con un gran porcentaje de valores vacíos o NaN.
- Estos no se consideran relevantes para el entrenamiento del modelo.

# ANÁLISIS EXPLORATORIO DE DATOS



/1

DATOS FALTANTES Y  
PROBLEMAS DE ETIQUETAS



/3

PALABRAS COMUNES EN  
TEXTO LIBRE



/2

VISUALIZACION DE DATOS



/4

EXPLORACIÓN BASADA EN  
MODELOS DE LENGUAJE

# DATOS FALTANTES

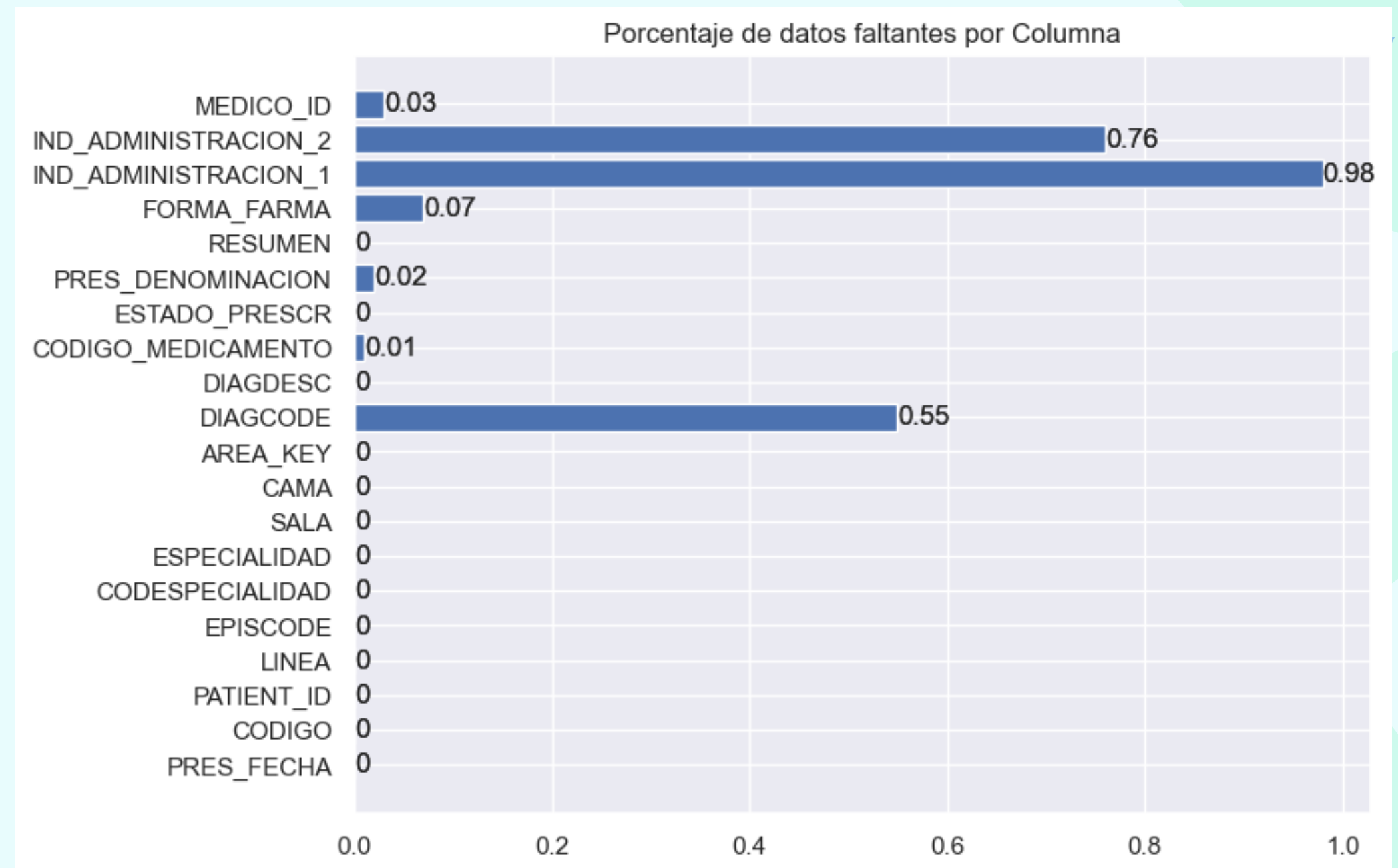
Gran parte de los datos faltantes corresponden a los atributos Indicación de Administración 1 y 2, los cuales son utilizados para casos especiales de administración.

Ejemplo:

- GLUCOSA 5 % SOLUCIÓN  
INYECTABLE UNIDAD 1000 ML
- 1000 ML PARENTERAL cada 12 horas

Forma Farma: solución inyectable

Ind Ad: 40cc/hora + 2g kcl



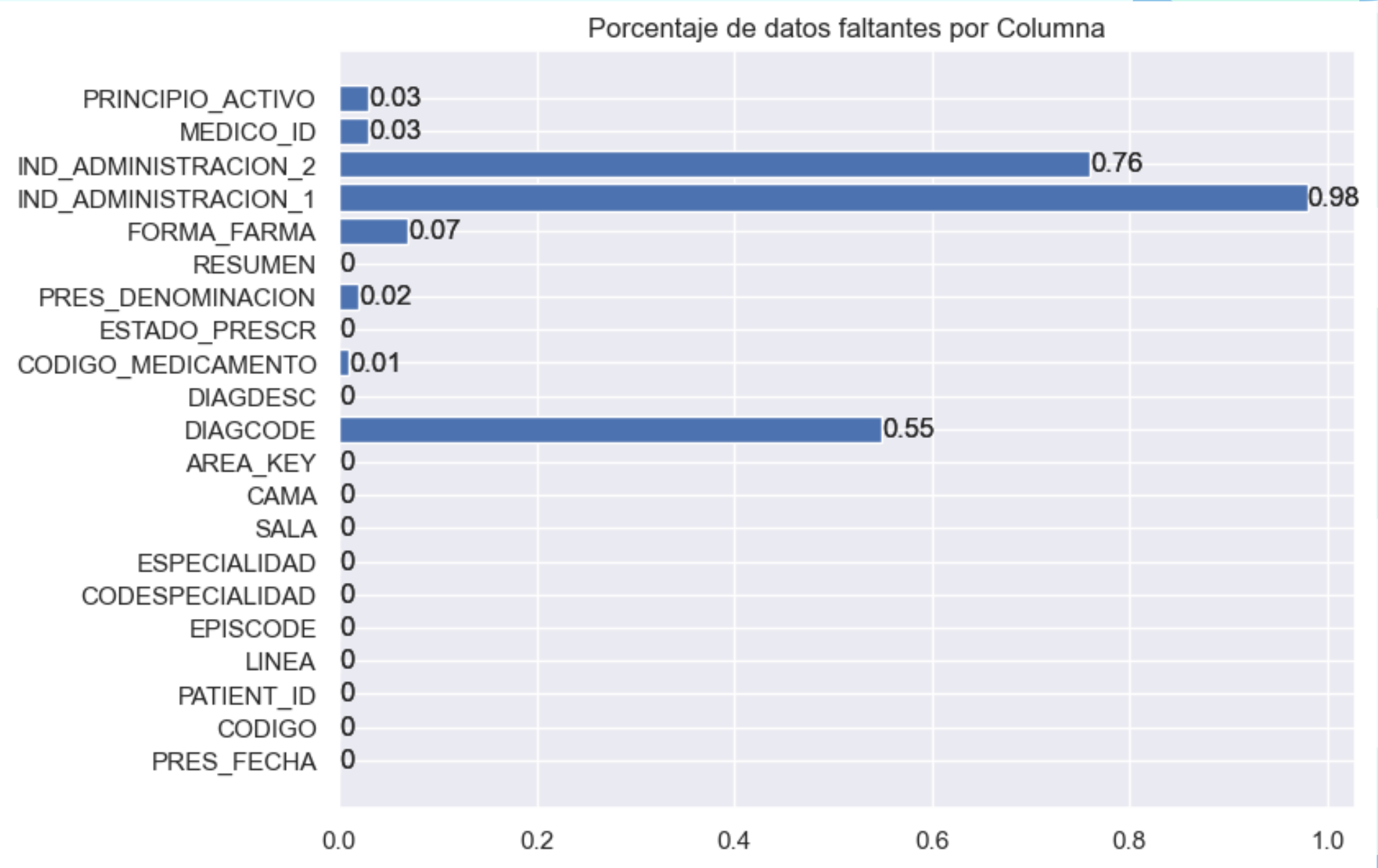
# DATA ADICIONAL, PRINCIPIOS ACTIVOS

Fue posible agregar una nueva columna de Principios activos respecto a el código HLF.

Datos faltantes en códigos de medicamentos: 18379

Datos faltantes en Principio activo: 50437

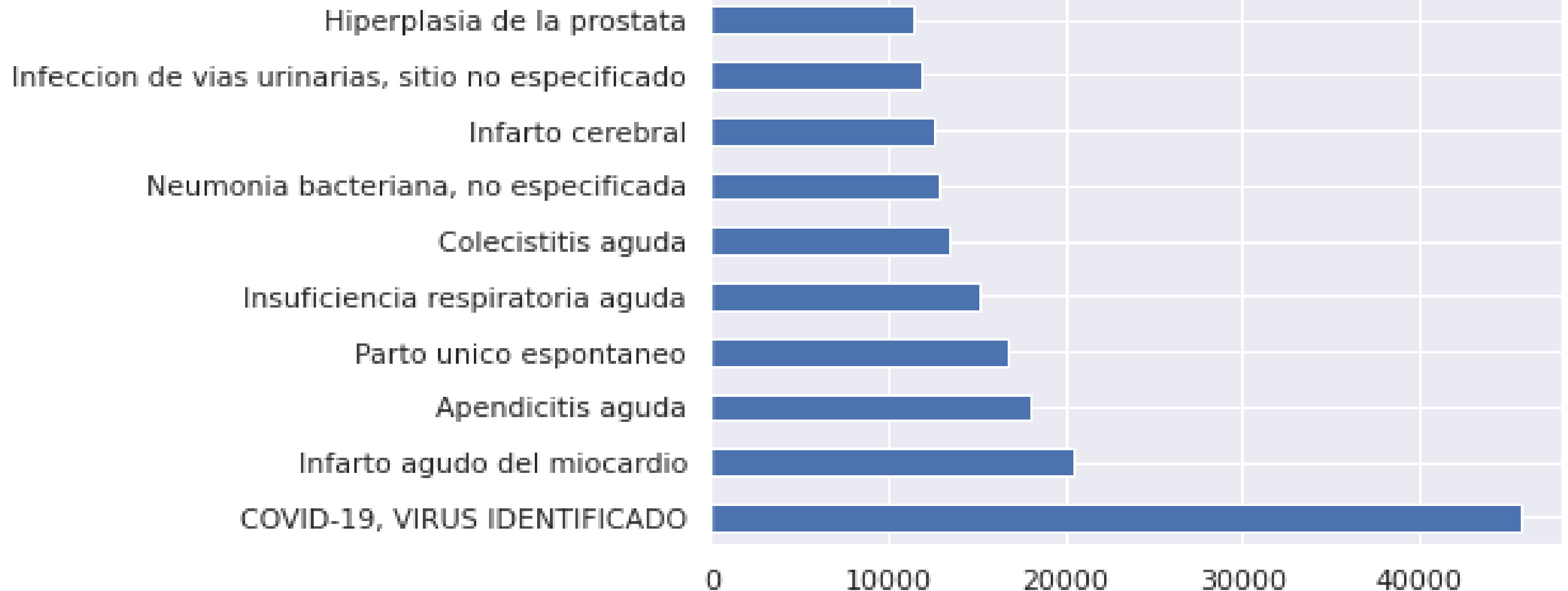
No todos los códigos de medicamento en las prescripciones tienen código HLF asociado.



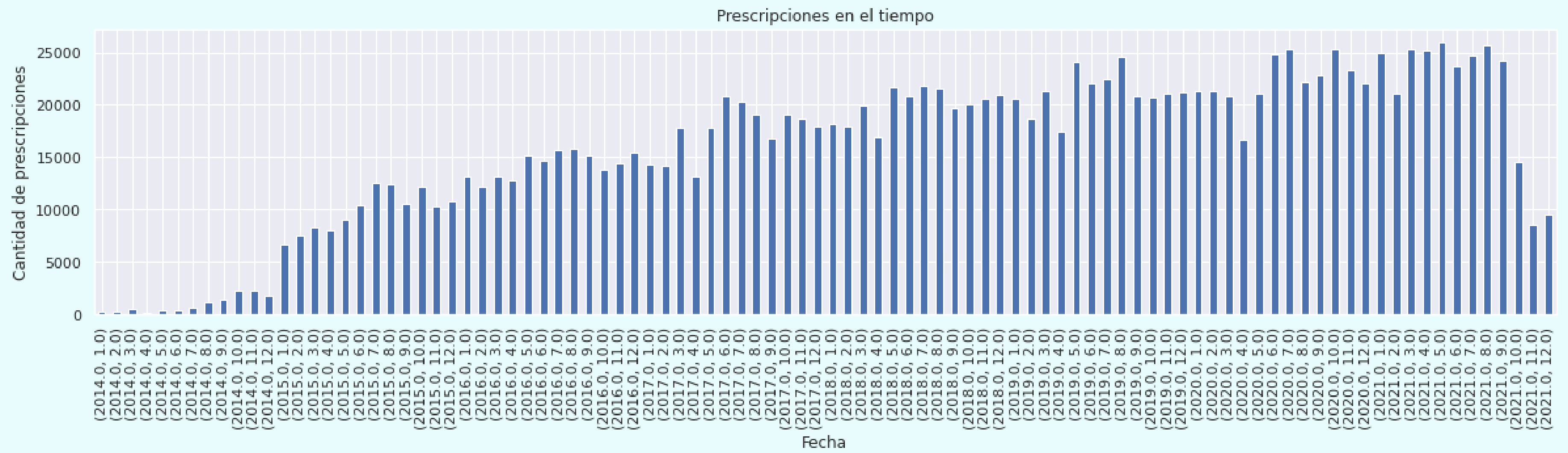


# VISUALIZACIÓN DE DATOS

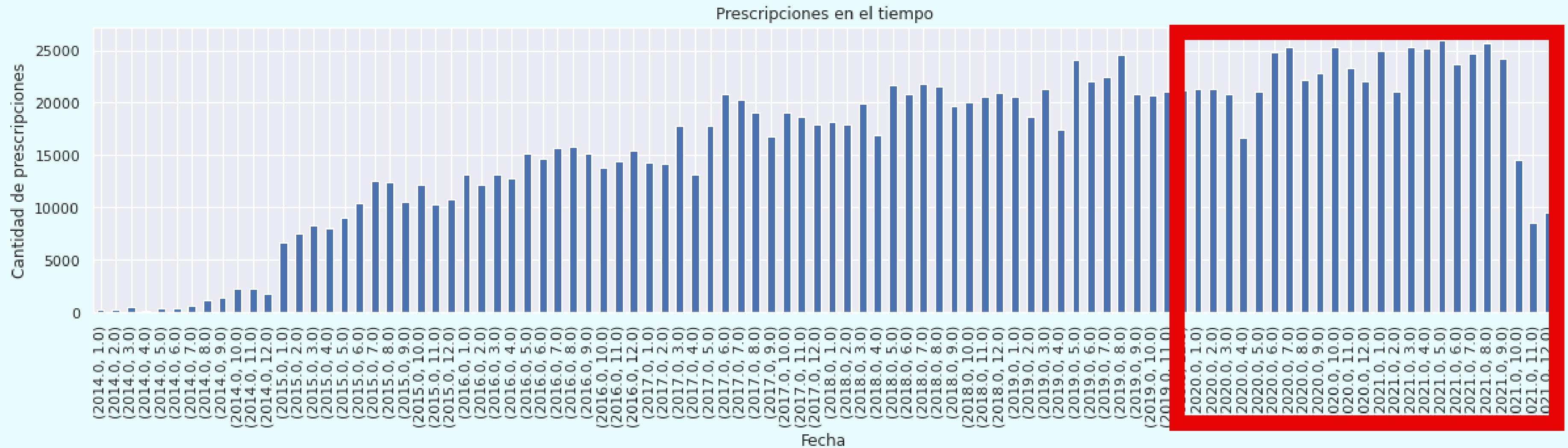
Diagnosticos más frecuentes



# VISUALIZACIÓN DE DATOS



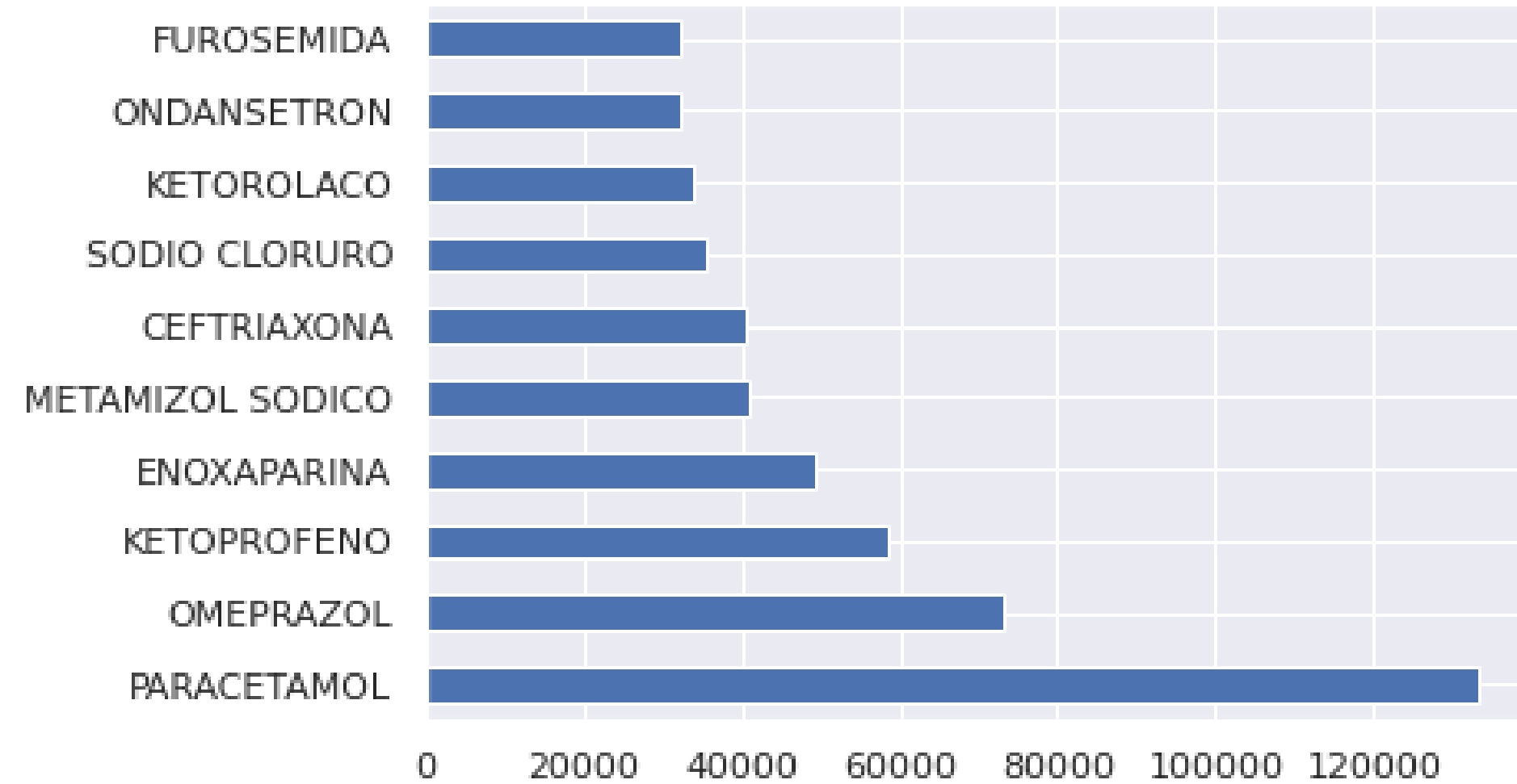
# VISUALIZACIÓN DE DATOS



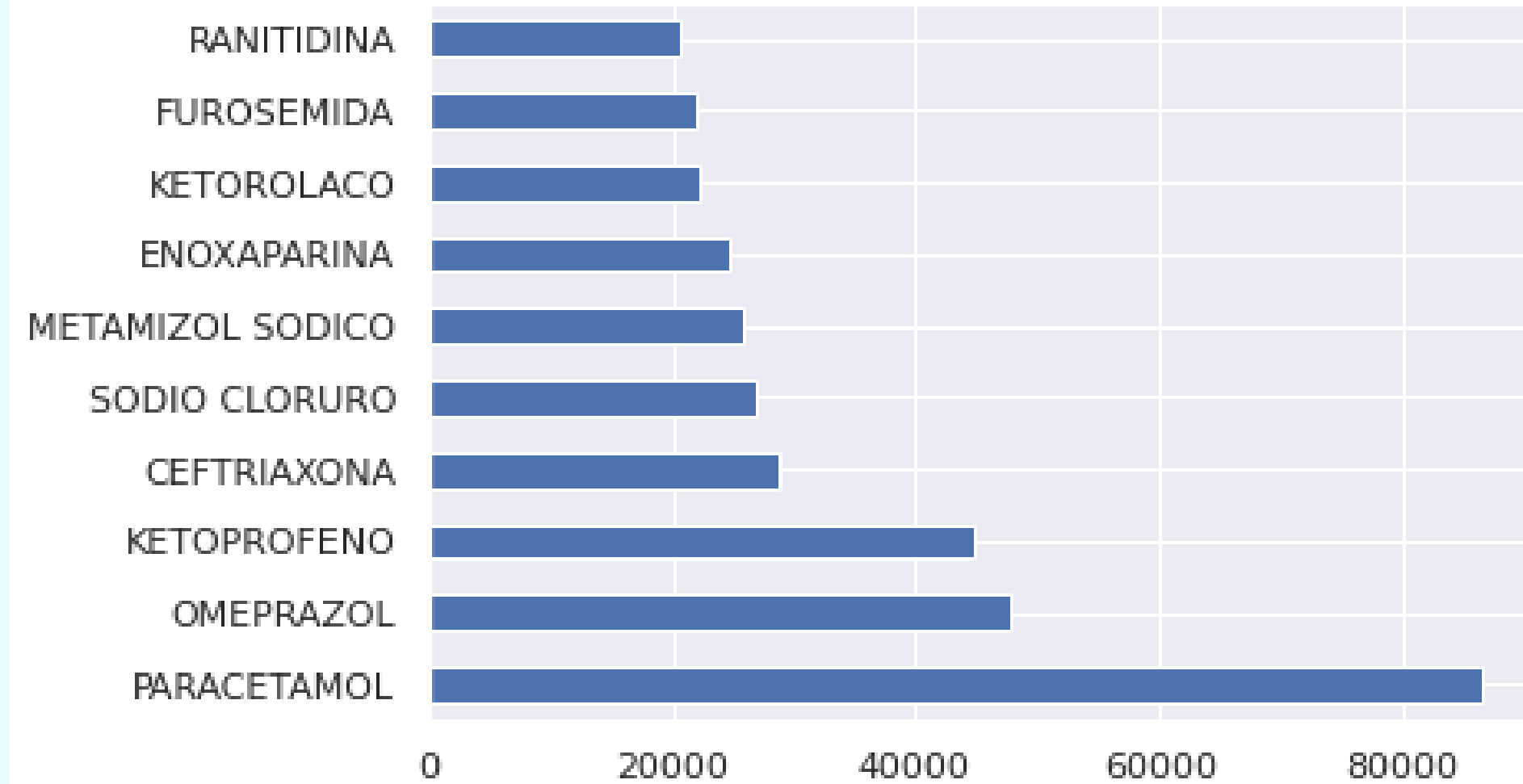
\*Aprox. un 40% de los datos corresponde periodo de pandemia.

# VISUALIZACIÓN DE DATOS

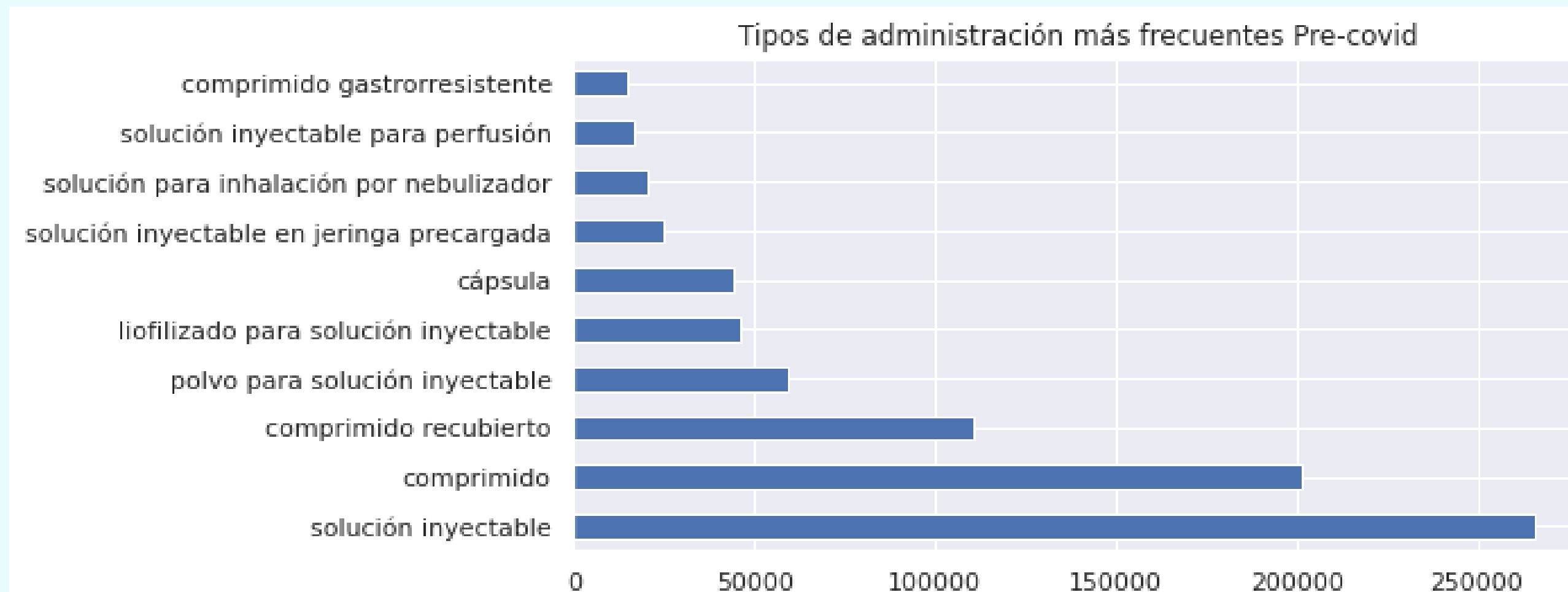
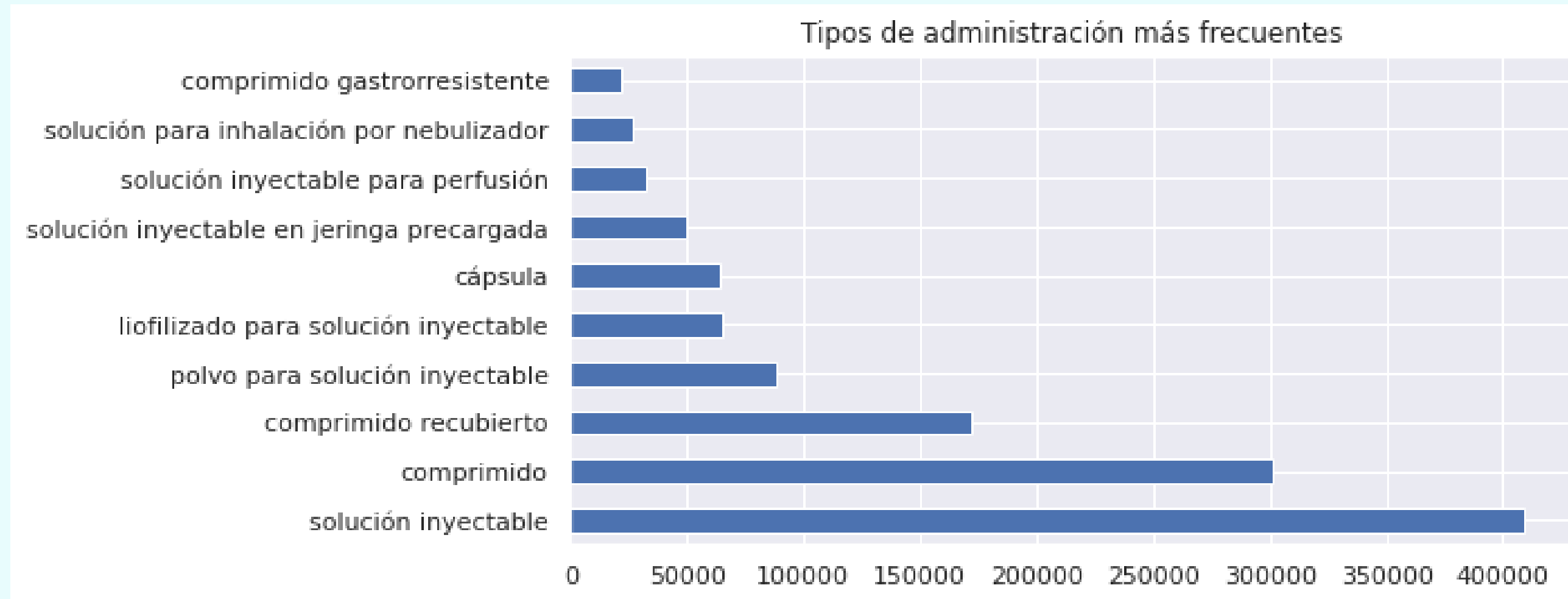
Principios Activos más frecuentes



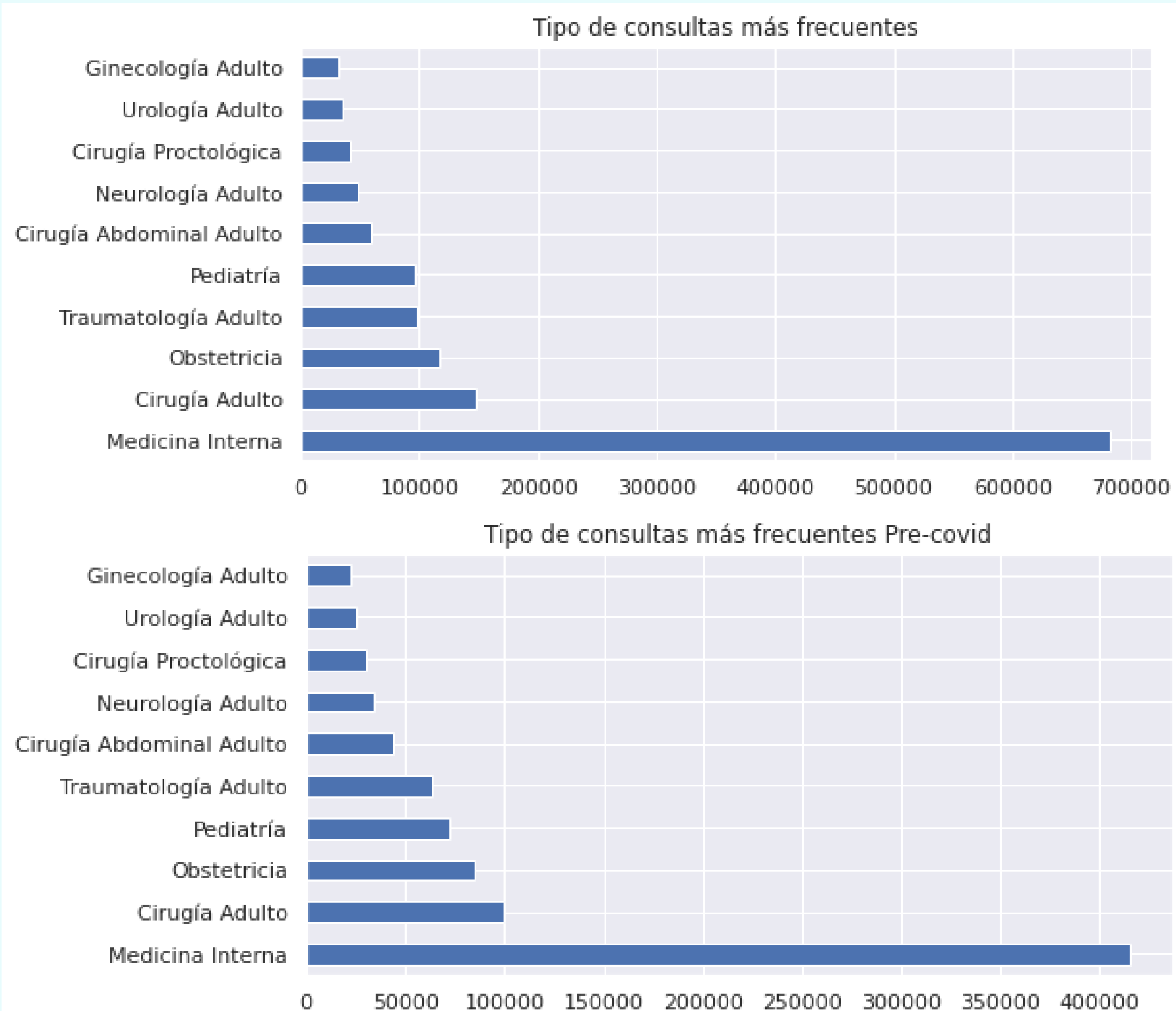
Principios Activos más frecuentes Pre-covid



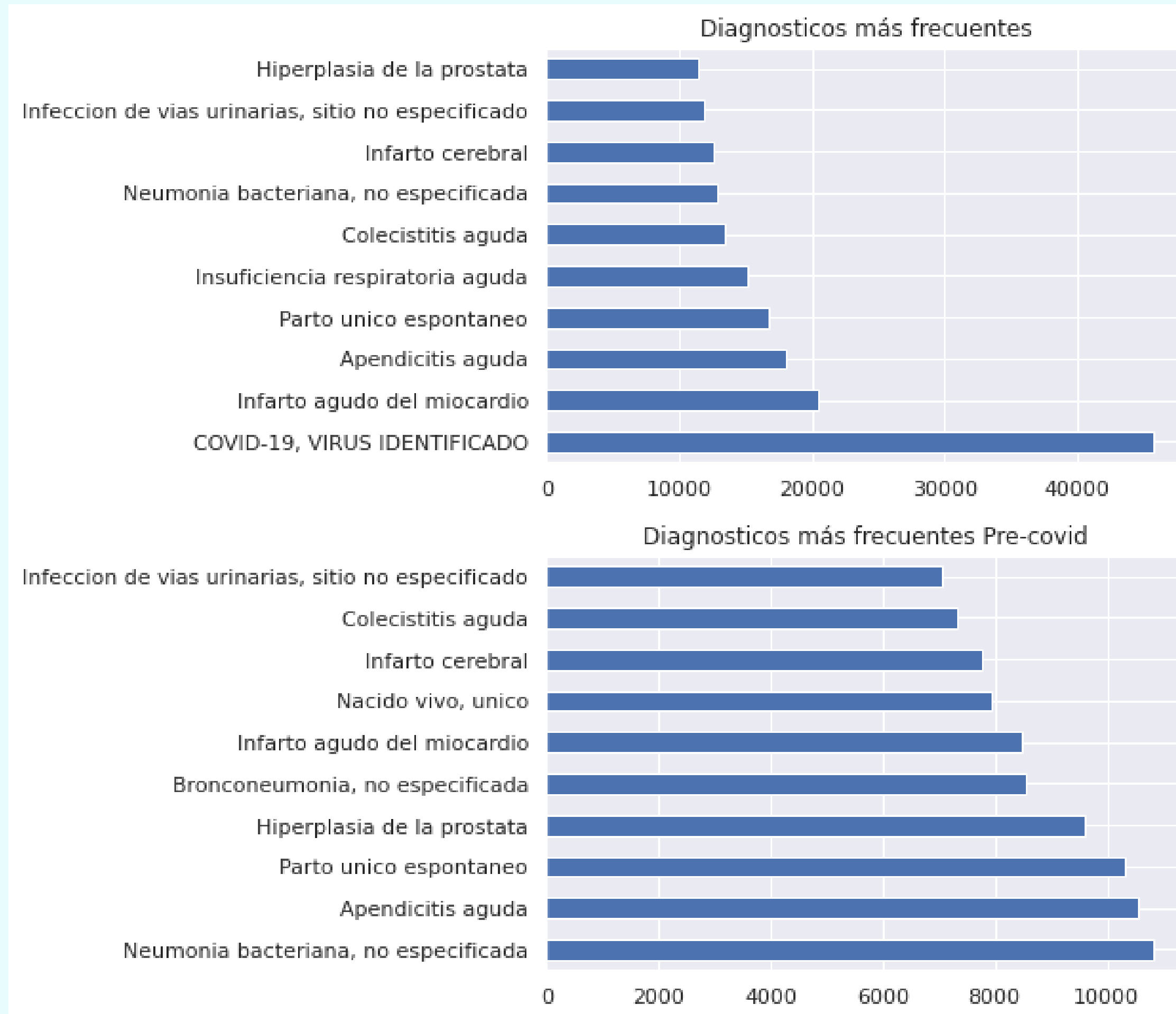
# VISUALIZACIÓN DE DATOS



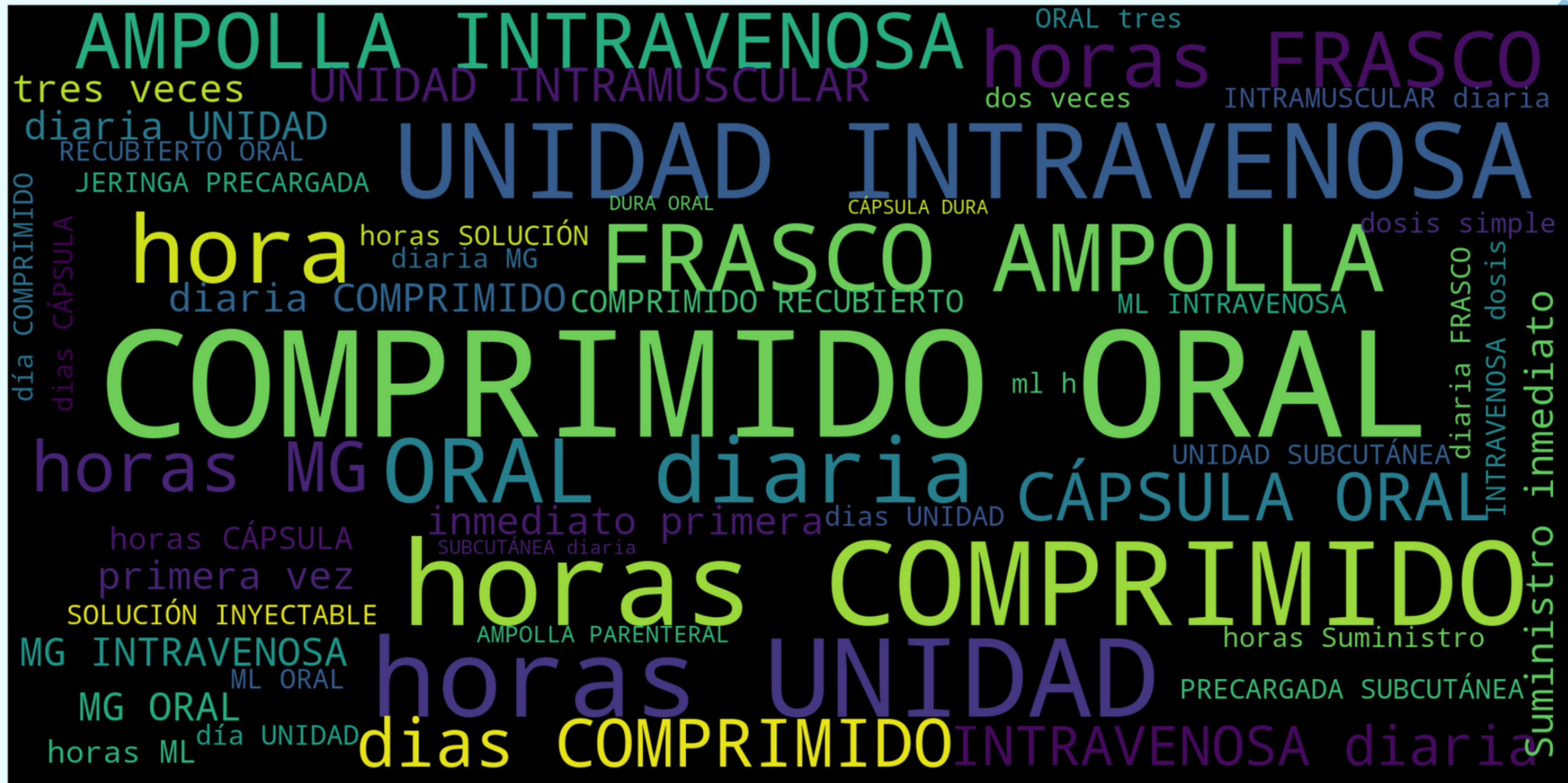
# VISUALIZACIÓN DE DATOS



# VISUALIZACIÓN DE DATOS



# PALABRAS COMUNES EN TEXTO LIBRE: RESUMEN DE LA PRESCRIPCIÓN





# PALABRAS COMUNES EN TEXTO LIBRE: DESCRIPCIÓN DE DIAGNÓSTICO



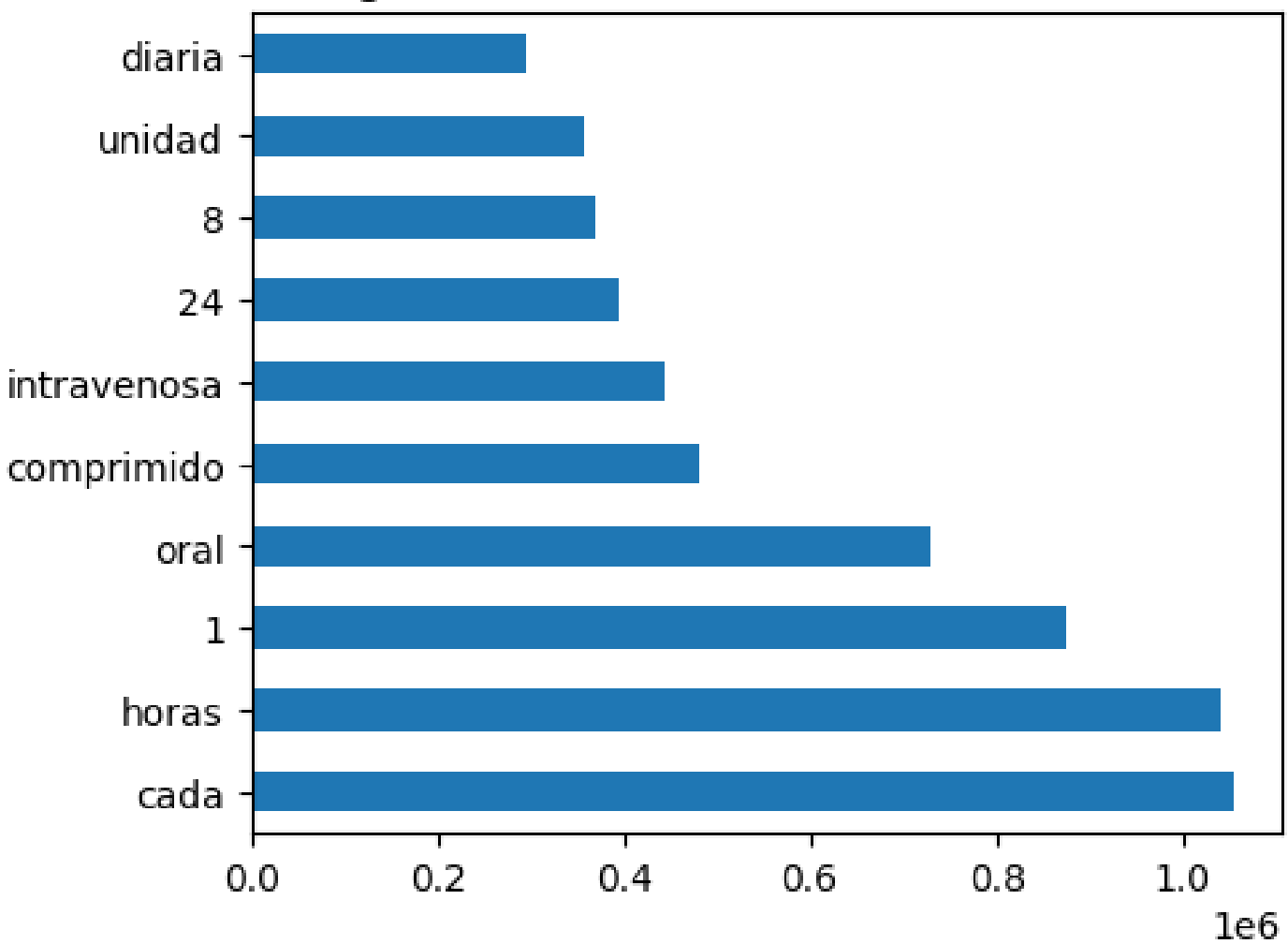
# PALABRAS COMUNES EN TEXTO LIBRE: RESUMEN DE LA PRESCRIPCIÓN

Cantidad de valores nulos en columna: 1

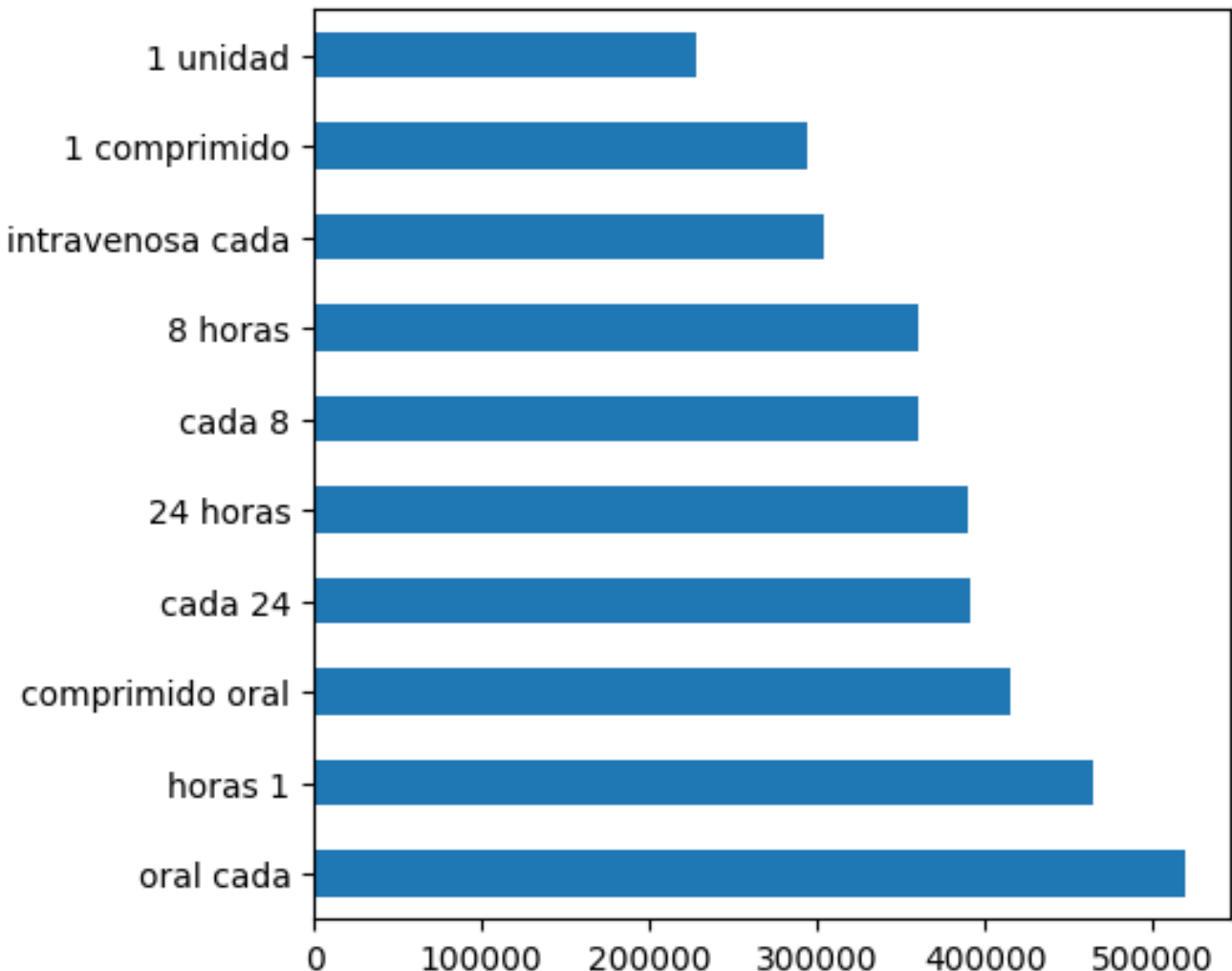
Total de 1473779 filas duplicadas (96.5%)

Cantidad de valores únicos: 52777 (3.5%)

Tokens (1-gramas) más frecuentes en columna RESUMEN



2-gramas más frecuentes en columna RESUMEN



# PALABRAS COMUNES EN TEXTO LIBRE: INDICACIONES DE ADMINISTRACIÓN

Cantidad de valores nulos en columna : **1495031**

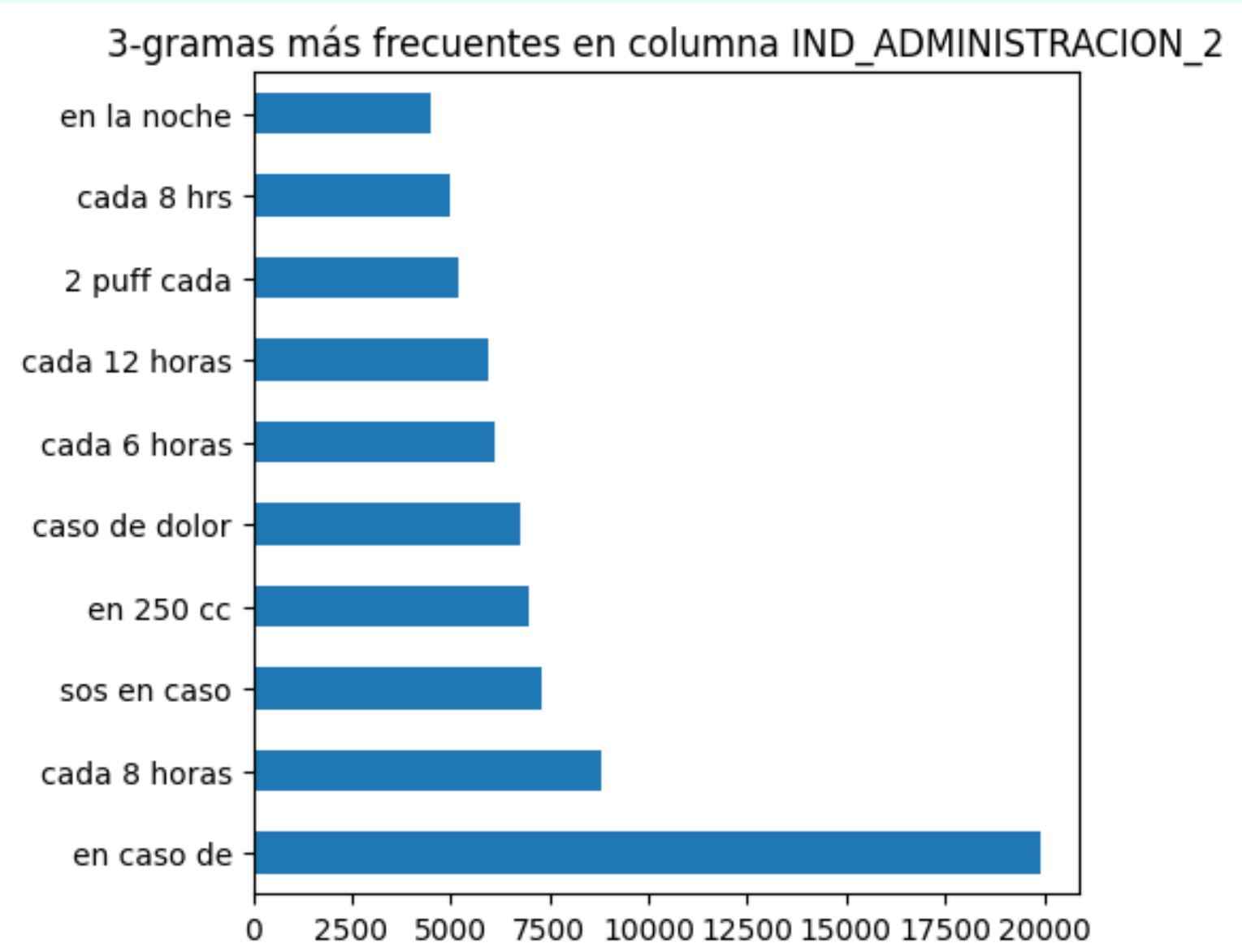
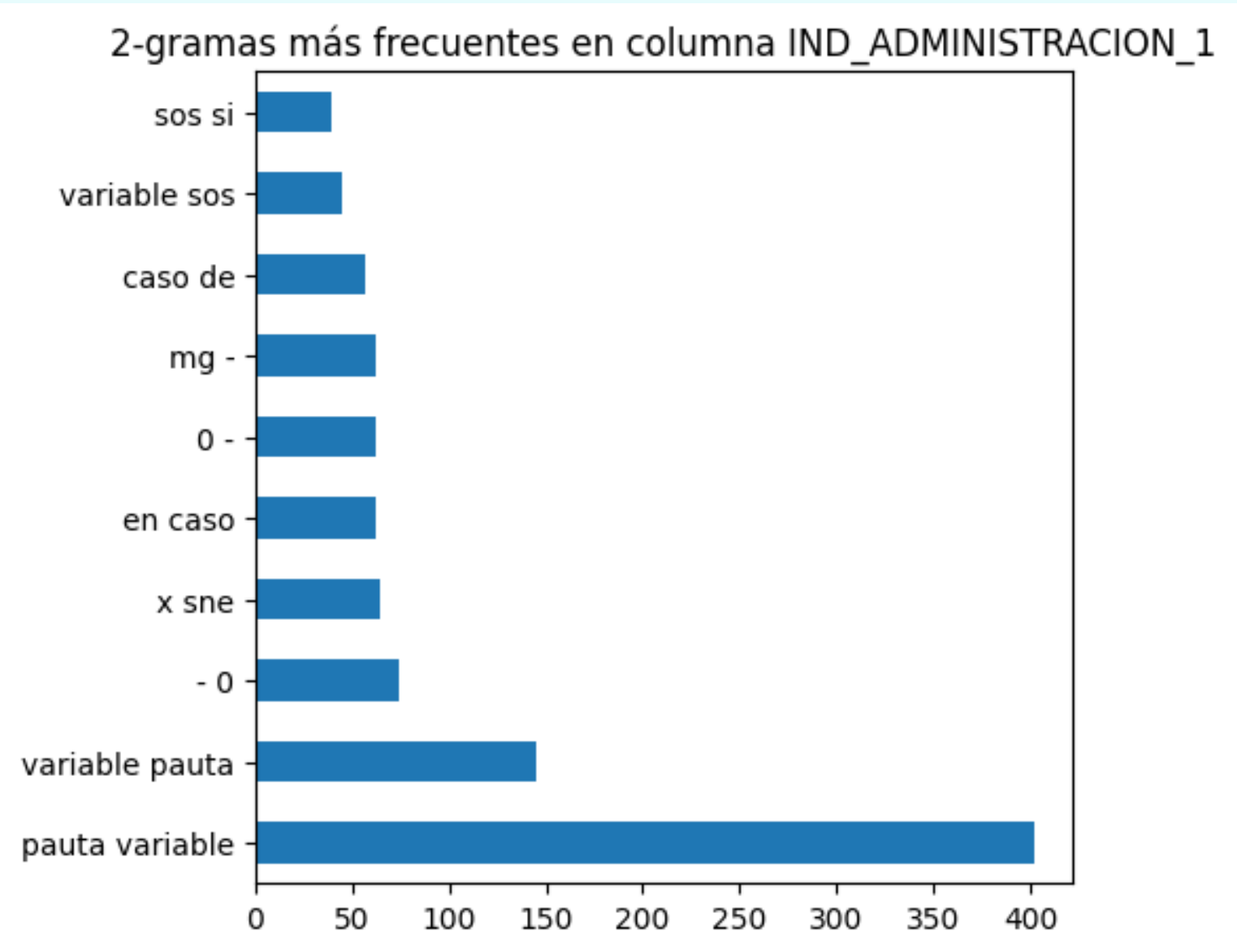
Total de **23208** filas duplicadas (**1.5%**)

Cantidad de valores únicos : **8317** (**0.5%**)

Cantidad de valores nulos en columna : **1162878**

Total de **188375** filas duplicadas (**12.3%**)

Cantidad de valores únicos : **175303** (**11.5%**)



# VECTORIZACIÓN DE TEXTO DE PRESCRIPCIONES USANDO ML

 [plncmm/bert-clinical-scratch-wl-es](#) 

 like 0



Fill-Mask



PyTorch



Transformers

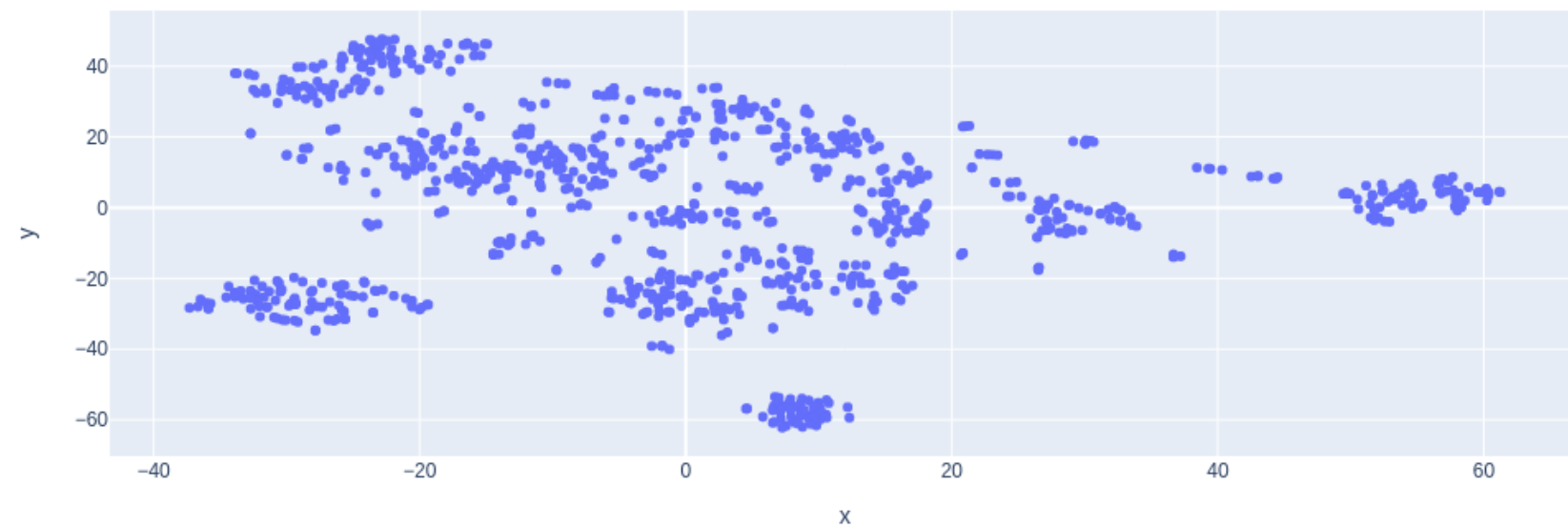
bert

generated\_from\_trainer

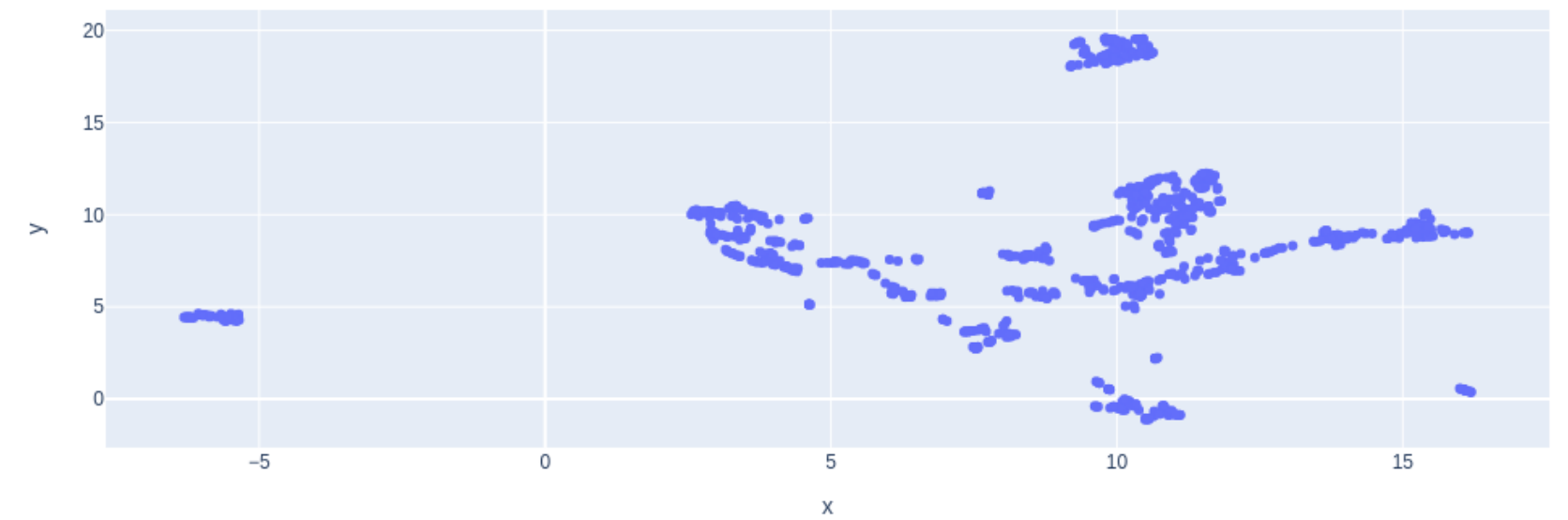


AutoTrain Compatible

Visualización con t-SNE, vectores suma



Visualización con UMAP, vectores suma



# LITERATURA

## NER: Named Entity Recognition

### Texto Clínico

Sang Meulder 2003  
↳ 2419

**Introduction to the CoNLL-2003 shared task: language-independent named entity recognition**  
CoNLL

Bose ... Ghosh 2021  
↳ 4

**A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts**  
Applied Sciences

Báez ... Dunstan 2020  
↳

**The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish**  
Association for Computational Linguistics

Báez ... Villena 2022  
↳ 0

**Automatic Extraction of Nested Entities in Clinical Referrals in Spanish**  
ACM transactions on computing for healthcare

Dunstan Villena 2022  
↳ 0

**Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing**  
Association for Computational Linguistics

### Contexto Chileno

# LITERATURA

## Texto Clínico

Báez ... Dunstan

2020  
↳

**The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish**

Association for Computational Linguistics

Báez ... Villena

2022  
↳ 0

**Automatic Extraction of Nested Entities in Clinical Referrals in Spanish**

ACM transactions on computing for healthcare

Dunstan Villena

2022  
↳ 0

**Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing**

Association for Computational Linguistics

## Conocimiento previo

Jiang ... Liu

2019  
↳

**Combining Contextualized Embeddings and Prior Knowledge for Clinical Named Entity Recognition: Evaluation Study.**

JMIR medical informatics

Akbik ... Vollgraf

2019  
↳

**FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP**

Association for Computational Linguistics

Kazama Torisawa

2007  
↳ 266

**Exploiting Wikipedia as External Knowledge for Named Entity Recognition**

EMNLP

Devlin ... Toutanova

2019  
↳

**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

Association for Computational Linguistics

## Modelos de lenguaje

## /4

# CONCLUSIONES Y TRABAJO FUTURO

- El efecto covid no influye de manera significativa para el reconocimiento de entidades (a priori). Sin embargo, la columna de diagnositcos, expuesta en texto libre, se ve sesgada por el efecto covid.
- En primer instancia, seria posible utilizar reglas para la detección de entidades, esto al comparar palabras frecuentes en el texto libre con otras columnas de la data.
- Las características de las columnas de texto libre nos hace pensar que la utilización de representaciones de lenguaje puede contribuir al objetivo de detectar entidades en esta.



**/4**

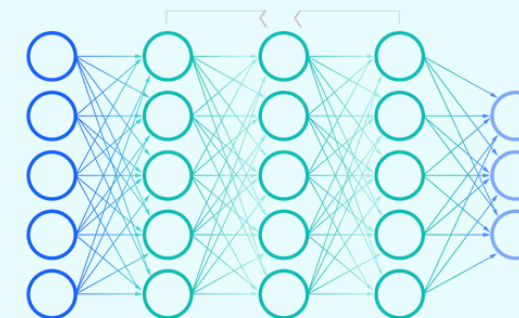
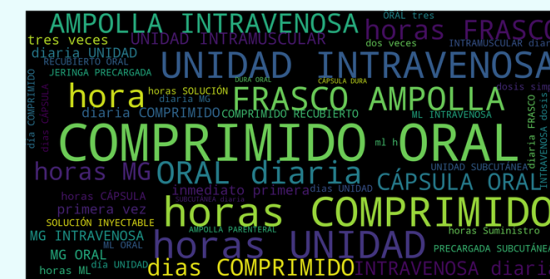
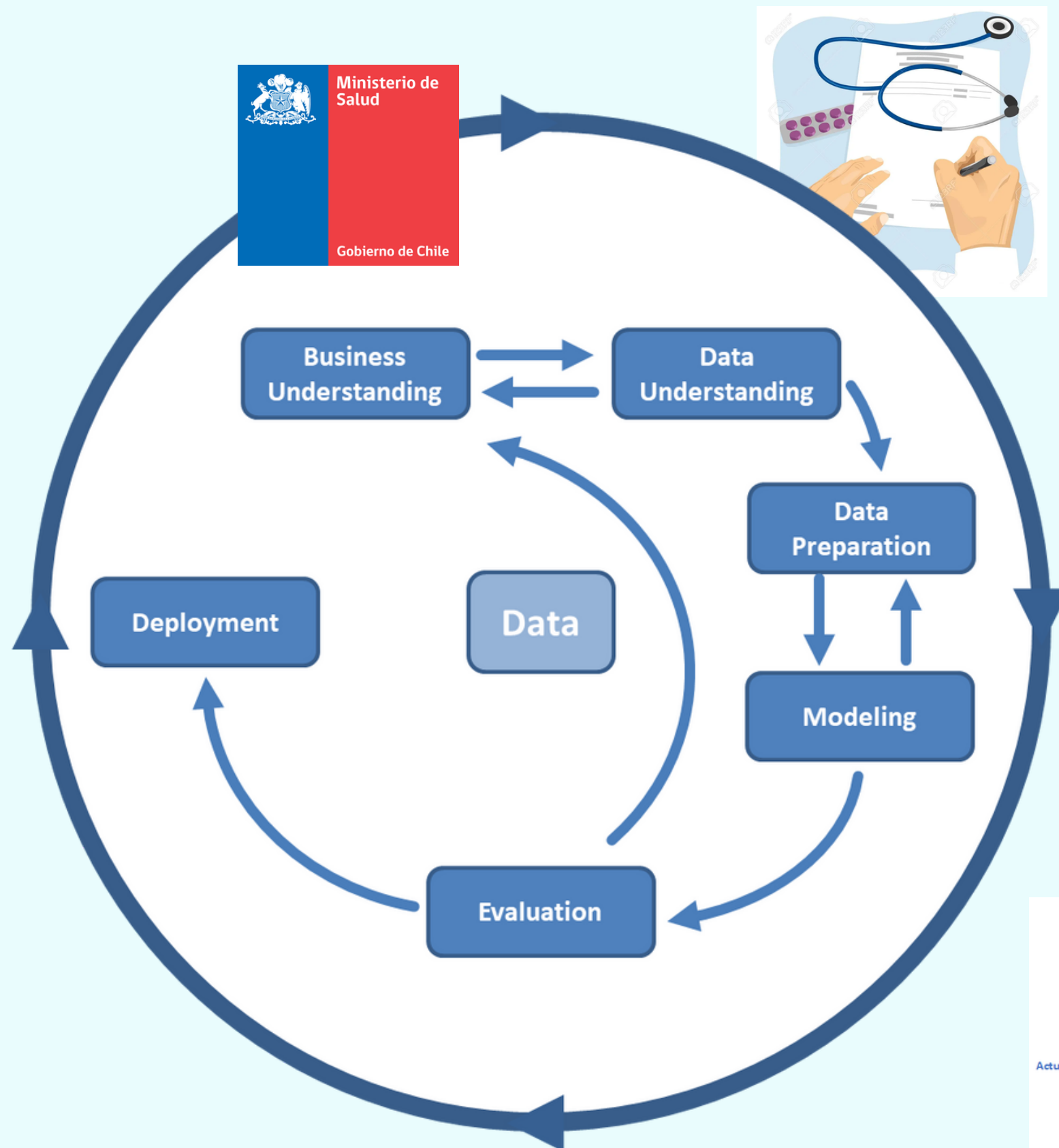
## **PLAN DE TRABAJO**

- Como primera iteración utilización de los modelos pre-entrenados "off the shelf".
- Ajuste de los modelos para nuestra tarea en específico, dados los resultados y conversaciones con la contraparte.
- Reunión con un experto del area de la salud
- Evaluar los resultados para futuras iteraciones.



/4

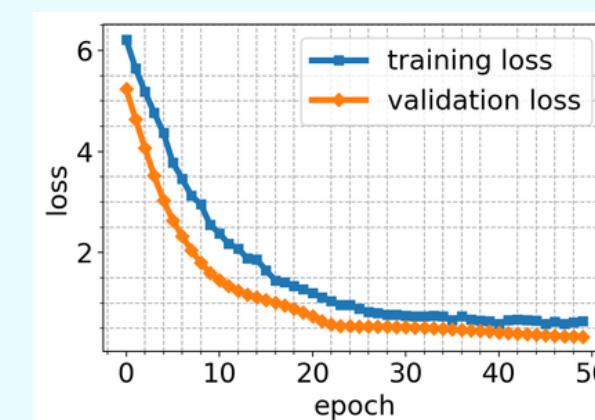
# PLAN DE TRABAJO

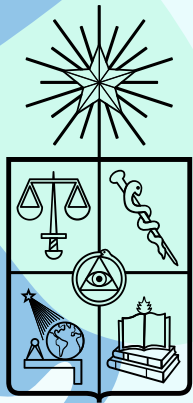


NER DEFINITION

Luke Rawlence **PERSON** joined Aimi **ORG** as a data scientist in Milton Keynes **PLACE**, after finishing his computer science degree at the University of Lincoln **ORG**.

		Predicted	
		Positive (+)	Negative (-)
Actual	Positive (+)	True Positive (TP)	False Negative (FN)
	Negative (-)	False Positive (FP)	True Negative (TN)





**MDS** Master of  
Data Science  
Universidad de Chile

PRESENTACIÓN 2 MDS7201

# ANÁLISIS EXPLORATORIO DE DATOS ENTIDADES MINSAL

DANIEL CARMONA, MARTÍN SEPÚLVEDA,  
MONSERRAT PRADO, CAMILO CARVAJA

- Sang, E. F., De Meulder, F.

Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition.

In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (142-147), 2003.

- Bose, P., Srinivasan, S., Sleeman IV, W. C., Palta, J., Kapoor, R., Ghosh, P.

A survey on recent named entity recognition and relationship extraction techniques on clinical texts.

In Applied Sciences (11(18), 8319.), 2021.

- Báez, P., Villena, F., Rojas, M., Durán, M., Dunstan, J. (2020, November).

The Chilean Waiting List Corpus: a new resource for clinical named entity recognition in Spanish.

In Proceedings of the 3rd clinical natural language processing workshop (pp. 291-300)., 2020.

- Báez, P., Bravo-Marquez, F., Dunstan, J., Rojas, M., Villena, F.

Automatic Extraction of Nested Entities in Clinical Referrals in Spanish.

In ACM Transactions on Computing for Healthcare, (3(3), 1-22.) - 2022.

- Rojas, M., Dunstan, J., Villena, F.

Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing.

In Proceedings of the 4th Clinical Natural Language Processing Workshop, (pp. 87-92)., 2022.

- Jiang, M., Sanger, T., Liu, X.

Combining contextualized embeddings and prior knowledge for clinical named entity recognition: evaluation study.

In JMIR medical informatics, (7(4), e14850.) - 2019