La base de datos seleccionada trata sobre el rendimiento de establecimientos de salud en la provincia de Buenos Aires entre el periodo 2018-2020. En el mismo se detalla el rendimiento con información sobre cantidad de consultas odontológicas, médicas y paramédicas, interconsultas, egresos, camas disponibles, días de estada, etc. Al ser un sistema de información empleado por los efectores de salud, permite obtener información sobre las variables de producción hospitalaria y la estructura prestacional de la totalidad de los efectores, sean de dependencia municipal, provincial o nacional. La carga de los datos se realiza de manera descentralizada es decir, se carga directamente desde el establecimiento obteniéndose los principales indicadores de coparticipación, teniendo periodos de carga y de cierre.

En la materia Taller de Proyectos I y II contempla el trabajo en terreno concurriendo a un efector de salud donde nos profundizamos en el manejo de estos datos a través de la carga de diversos formularios. Nos pareció oportuno la selección de dicho tema, ya que es de vital importancia para la toma de decisiones en las políticas públicas a implementar.

Para la selección de nuestro tema en cuestión, empleamos como fuente de información la página Datos Abiertos PBA (<https://catalogo.datos.gba.gob.ar/>) visualizando y descargando un archivo csv sobre el Rendimiento de Establecimientos de Salud 2018-2020. Al observar el archivo notamos que las variables con las que trabaja presentan una gran cantidad de datos faltantes (espacios en blanco).

Se utilizó Google Drive y Google Colaboratory para realizar un Análisis Exploratorio Descriptivo, permitiendo acceder a los datos de manera simultánea a cualquier integrante del grupo. Así también, empleamos como herramienta para realizar el análisis y los gráficos sobre los datos estudiados la plataforma ChatGPT, siendo muy útil para trabajar con código complementandose con Google Colab, haciendo posible trabajar con el programa Python.

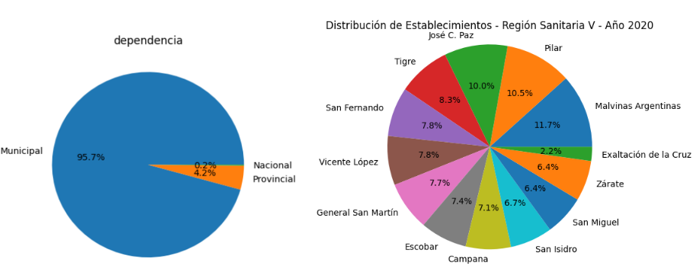
Como inicio, instalamos e importamos las librerías (pandas, matplotlib, seaborn) necesarias para trabajar con Python, mostrando la estructura de los datos. Para tener una idea general de la misma, nuestra primera acción es visualizar los datos.Se procede a obtener la totalidad de filas y columnas que presenta nuestra dataframe, con df.shape dando una totalidad de 78.290 filas y 23 columnas.

Luego, tratamos de generar una descripción más detallada de nuestra base de datos con información sobre las columnas y su estadística descriptiva.

Conforme a ello, continuamos a verificar los valores faltantes, más del 51% de nuestra base se encuentra sin datos, ya sea por omisión o por falta de carga. Como así también en la falla en el registro al cargar el dato con espacios entre los números de 4 dígitos.

Antes de proceder con la eliminación o reemplazo de los valores nulos, se trabajó con las variables que no presentaban NAN y se presentaban completos tales como las variables de año, mes, región sanitaria, municipio, dependencias. Con ellos graficamos el porcentaje de dependencias de los establecimientos sanitarios: nacional, provincial, municipal (figura 1).

Figura 1: Dependencias Figura 2: Porcentaje de establecimientos

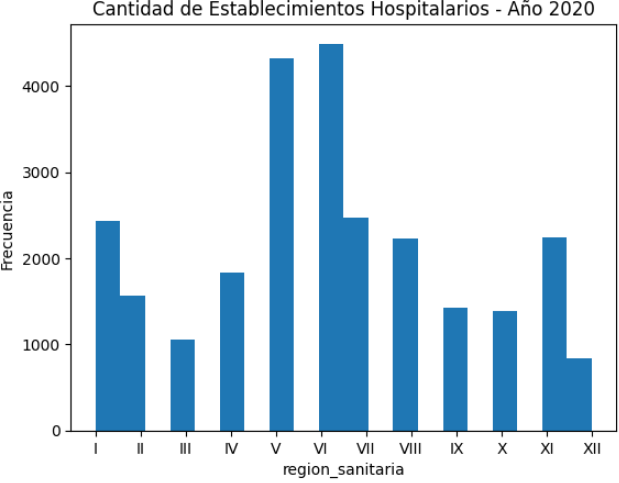


FUENTE: ELABORACIÓN PROPIA

Nos centramos en la región sanitaria V, para conocer qué municipios presentan mayor cantidad de efectores (figura 2).

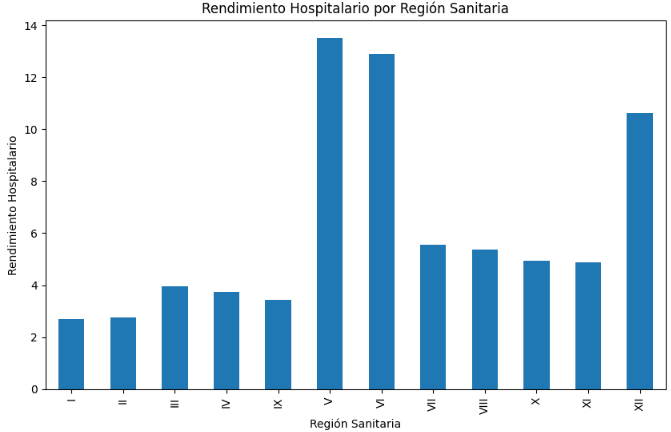
Creamos un histograma con la cantidad de establecimientos por región sanitaria, concluyendo que las regiones V y VI presentan mayor cantidad de efectores (figura 3). A su vez, se analizó el rendimiento hospitalario según la región sanitaria, que a simple vista se observa como en el gráfico anterior, las regiones V y VI son las que presentan mayor desarrollo (figura 4).

Figura 3:Cantidad de establecimientos



FUENTE: ELABORACIÓN PROPIA.

Figura 4: Rendimiento Hospitalario



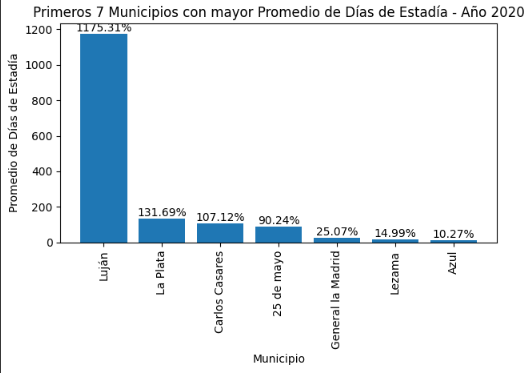
FUENTE: ELABORACIÓN PROPIA.

Con las siguientes variables que se describen en la dataframe sobre las diferentes consultas de atención a los pacientes (odontológicas, médicas, paramédicas e interconsultas), son las que presentan menor cantidad de valores nulos. Ante la falta de datos en dichas variables, las suposiciones planteadas como grupo fue que la ausencia de los mismos se debía a que las consideramos como valor 0 (cero) o que directamente no se cargaba en el registro. Por ello, al observar detenidamente la base de datos, pudimos visualizar que si se utiliza el valor 0 para el registro de algunos datos. Por ello, procedimos a eliminar de manera individual las filas que no poseían valores. Al modificar los valores nulos, notamos que influía en los NAN de las demás variables.

Como objetivo de nuestro análisis buscamos desarrollar un detector de anomalías especialmente en la estadía de camas para determinar cuál es el tiempo estándar de internación dentro de los efectores en la provincia de Buenos Aires, trabajando con la variable promedio de día de estada de nuestra base de datos seleccionada, para que a futuro localizar cuáles son dichos efectores diagnosticando el principal problema e intentando resolverlo según la complejidad del establecimiento.

Al igual que trabajamos de manera individual con cada una de las variables mencionadas con anterioridad, con el promedio de días de estada se procedió a eliminar los valores nulos por filas y se pudo representar en un gráfico de barras los municipios que presentaban mayor días de estada (figura 5).

Figura 5 :Promedio de días de estada



FUENTE: ELABORACIÓN PROPIA.

Notamos que en los gráficos realizados en cada una de las variables, mayormente se destacaron 2 municipios La Plata y Luján, planteandonos si esto se debe a que la carga del dato la realizan de manera errónea o directamente no se lo registra.

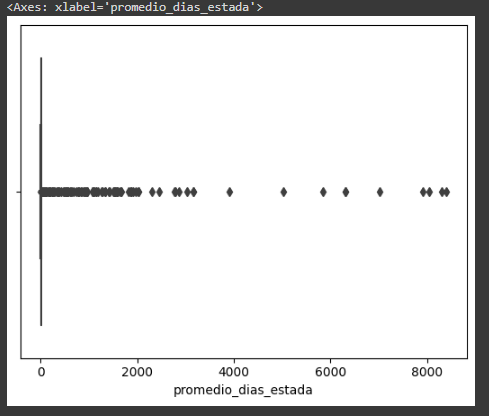
Enfocándonos en nuestro análisis, empleamos la utilización de un boxplot como una manera de visualizar la información donde la mayoría de los datos se concentran en la línea media (Q2), delimitados por el primer y tercer cuartil. Esto hace que los “bigotes” que conforman el diagrama sean los valores máximos y mínimos de los datos. Permitiendo identificar los valores atípicos y poder comparar distribuciones, estos valores atípicos u outliers son aquellos que están más allá del límite inferior o superior, y se visualizan rápidamente al observar el gráfico llamando la atención. Procedemos a obtener los valores de los cuartiles necesarios y para detectar valores atípicos usamos el criterio IQR 1.5, permitiendo eliminarlos.

Luego generamos un reporte en el que se describen los análisis fundamentales de las variables permitiendo tener una visualización mucho más gráfica de los datos. Este reporte lo generamos con un sub-paquete de pandas, utilizando la función ProfileReport. Además, permite guardarlo en la carpeta del Drive compartido.

En la figura 6, muestra el gráfico boxplot utilizado para detectar anomalías, en este caso sobre el promedio de días de estada, en él determinamos que el rango intercuartílico calculado es de 3.40, donde Q1 es 3.28 y Q3 corresponde a 6.82. Es decir, para definir los valores atípicos definimos los límites superior e inferior utilizando el criterio intercuartílico (IQR), quedando;

* Límite inferior: Q1-1.5\*IQR.
* Límite superior: Q3+1.5\*IQR.

Figura 6: Boxplot Promedio día de estada

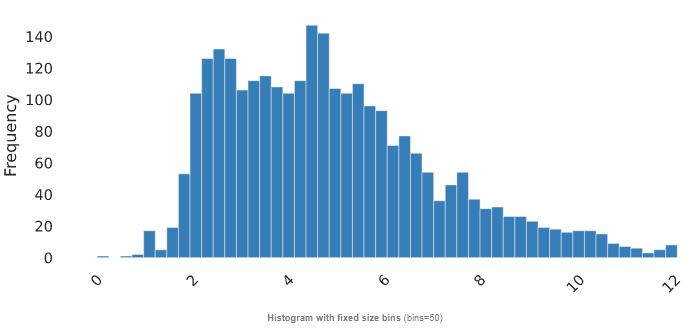


FUENTE: ELABORACIÓN PROPIA.

El análisis del diagrama de caja ayuda a comprender la distribución, la presencia de valores atípicos, la tendencia central y la variabilidad de los datos. Esto permite identificar patrones, detectar valores inusuales y obtener una visión general de la distribución de los datos.

Por último, en el histograma del promedio de día de estada (figura 7) podemos concluir que la mediana o segundo cuartil (Q2) ronda alrededor de 4.58 días, delimitado entre el primer cuartil (Q1) con 3,14 y el tercer cuartil (Q3) con 6,16. Es decir, que los valores de promedio de días de estada por encima y por debajo del Q1 y Q3 son considerados como valores atípicos, y a partir de allí poder realizar otros tipos de análisis con dichos valores para poder determinar a qué efectores pertenecen para mejorar su rendimiento.

Figura 7: Histograma promedio de días de estada



FUENTE: ELABORACIÓN PROPIA.