PROYECTO ANALÍTICA DE NEGOCIO – DATASET INSURANCE

ESTEBAN CARO PELAEZ
Estudiante de Ingeniería Industrial
Universidad de Antioquia

JORGE LUIS RODRIGUEZ LOPEZ
Estudiante de Ingeniería Industrial
Universidad de Antioquia

Hipervínculo al repositorio del proyecto: Proyecto final

Resumen

En el presente informe se desarrollan el planteamiento de preguntas de negocio con relación a la base de datos de una aseguradora, en donde para dar respuestas a estas preguntas se parte de la limpieza de datos en conjunto con un análisis exploratorio y visualización de la información. En el desarrollo del proyecto se utiliza el lenguaje de consulta estructurado SQLite junto con el lenguaje de programación de alto nivel Python. El objetivo principal del presente estudio es analizar los costos y la distribución de los afiliados a la aseguradora a la luz de diferentes atributos como edad, género, región y enfermedades prevalecientes. Se ha descubierto mediante el análisis de los datos, que la aseguradora cuenta con un mayor número de hombres que de mujeres, y no por esto, son aquellos quienes más reclamaciones producen. Por otro lado, el regional centro cuenta con el mayor número de afiliados. Finalmente, la enfermedad más prevaleciente dentro de los afiliados es la hipertensión, representando a su vez, el mayor coste para la compañía.

Palabras Claves: Preguntas de negocio, SQLite, Analítica de negocio, Python.

1. INTRODUCCIÓN

En la actualidad la analítica de negocio tiene un papel fundamentan en la toma de decisiones para las organizaciones, ya que permite recopilar, procesar y analizar grandes cantidades de datos para obtener información de interés que pueda ser utilizada para el mejoramiento y continuidad de la actividad empresarial. Esto ayuda a los lideres empresariales a tomar decisiones más fundamentadas y respaldadas por datos, en lugar de depender de la intuición o la experiencia en negocios [1]. Al aprovechar la analítica de negocio, las organizaciones pueden lograr una mayor eficiencia operativa, mejorar la satisfacción al cliente y obtener una ventaja competitiva sostenible en mercado por lo que en este informe se presentan algunas preguntas de negocio que son de utilidad para la toma de decisiones de una aseguradora.

2. ANÁLISIS EXPLOTORIO Y ESTADISTICOS DE LOS DATOS

Para poder comprender de qué manera se relacionan los atributos de las entidades, se elabora un diagrama Entidad Relación (ER), en cual muestra cómo interactúan las entidades entre sí.

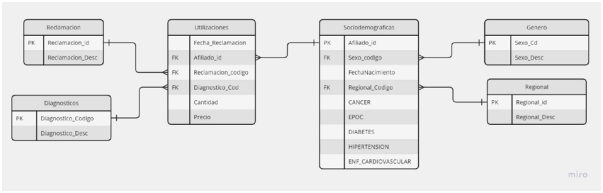


Diagrama 1. Diagrama entidad – relación.

En caso de visualizar más a detalle el diagrama puede ingresar al siguiente enlace: https://n9.cl/vw6ds

Una vez realizado el diagrama ER, se facilita el entendimiento y la elaboración de los queries para las distintas tablas a disposición.

La base de datos suministrada pertenece a los registros de los afiliados de una aseguradora con ciertas características como las enfermedades que puede tener estos afiliados, el género, región donde pertenecen, las reclamaciones que generan, los diagnósticos asignados y costos que representan para la aseguradora.

A continuación, se presentan algunas estadísticas para tener un primer bosquejo de la información.

A) Cantidad de hombres y mujeres afiliados.
Se tienen observaciones de 119.253 afiliados femeninos y 98.951 afiliados masculinos

B) Costo promedio de los afiliados para la aseguradora.
El costo promedio de los afiliados es de 289.604,6

C) Cantidad total de reclamaciones realizadas por los afiliados.
Se tiene registro de 595.571 reclamaciones

D) Costos máximos y mínimos en las reclamaciones hechas por los afiliados.
Los costos máximos son 541.000.682,5 y los mínimos son 2,48. El afiliado que representó los mayores costos fue debido a un tratamiento quirúrgico hospitalario y complicaciones, el cual es el diagnóstico 30 dentro de las identificaciones de reclamaciones de la compañía.

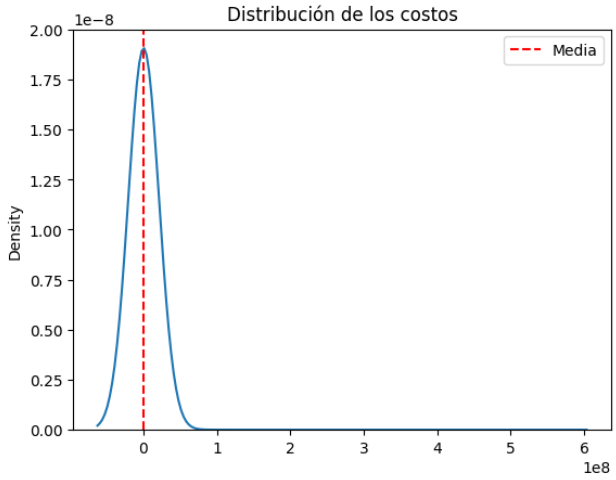
E) Afiliados que realizaron reclamaciones de DIÁLISIS.
Se tiene registro de 10 afiliados con reclamaciones por diálisis.

F) Tabla de estadísticas resumidas de la tabla utilizaciones
Analizar estadísticamente esta tabla permite tener un mayor entendimiento del dataset.

	Cantidad	Precio
count	395868.00	395868.00
mean	1.50	289604.65
std	2.34	2088675.78
min	0.00	2.48
25%	1.00	54150.80
50%	1.00	68102.66
75%	1.00	142325.03
max	210.00	541000682.50

Tabla1. Estadísticas resumidas utilizaciones

G) Distribución de los costos dentro de la compañía



El gráfico de densidad de la variable costos muestra una asimetría positiva, lo cual implica que hay una tendencia hacia valores más bajos, mientras que los costos más altos se dan con menor frecuencia.

3. HIPÓTESIS DE ANÁLISIS

Inicialmente se plantean siete hipótesis o preguntas de negocio, que por medio del lenguaje de consulta estructurado SQLite se dan respuesta. Con estas preguntas se espera encontrar patrones y comportamientos de los afiliados y sus características dentro de la aseguradora.

1. ¿Qué género presenta mayores reclamaciones para la aseguradora?
un estudio realizado por Fedesarrollo, la tendencia histórica mostraba que los hombres tenían una mayor tasa de afiliación que las mujeres, sin embargo, en el año 2011 la población femenina repunta y se ubica por encima de la masculina. Este fenómeno puede ser explicado por la inclusión de la mujer en el mercado laboral colombiano [2].

En la siguiente tabla se muestra los datos sobre las reclamaciones realizadas por los géneros.

Género	Reclamaciones	Proporción
F	291.752	60%
M	195.903	40%
Total	487.655	100%

Tabla 1. Reclamaciones basadas en género

Para entender la diferencia entre el porcentaje de hombre y mujeres se presentan los resultados en el siguiente diagrama de torta.

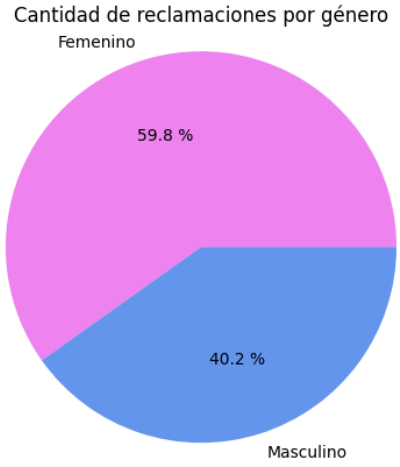
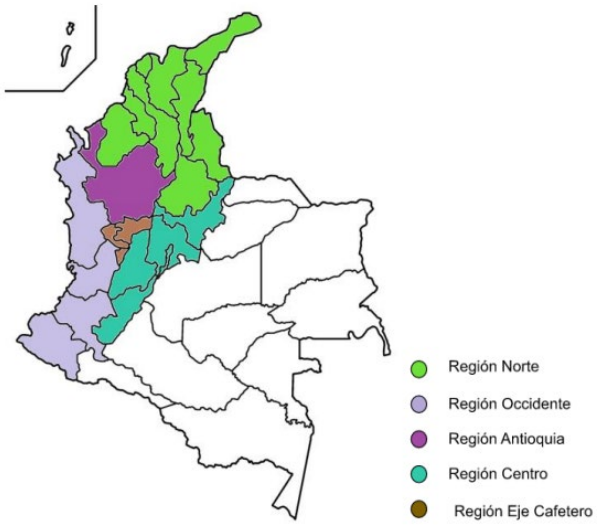


Diagrama de torta 1. Cantidad de reclamaciones por género.

Dadas las proporciones para las reclamaciones entre hombres y mujeres, se evidencia que la mayor cantidad de reclamaciones para la aseguradora son representadas por las mujeres afiliadas.

2. ¿En qué región se presentan la mayor cantidad de reclamaciones?

Se realiza una ubicación estimada de cada una de las regionales teniendo en cuenta diferentes mapas políticos y económicos como guía para clasificar las diferentes zonas de interés del país



Mapa 1. Mapa de Colombia por regiones

Se espera que, en la región norte al estar compuesto por una mayor cantidad de departamentos, en esta se presenten la mayor cantidad de reclamaciones. Al realizar la consulta se obtiene la siguiente información

Región	Reclamaciones
Antioquia	62.330
Centro	208.913
Norte	42.968
Occidente	65.656
Eje Cafetero	11.071
Sin información	54

Tabla 3. Reclamaciones por regiones

En el siguiente diagrama de barras se evidencia mejor la comparación entre estas cantidades de reclamaciones por región.

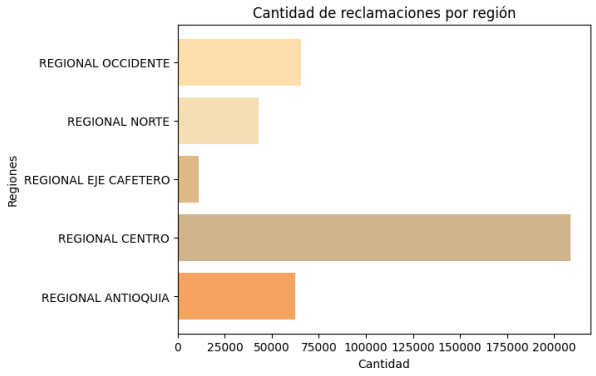


Diagrama de barras 1. Reclamaciones por regiones

A pesar de que la región Norte es la más grande en esta no se presentan la mayor cantidad de reclamaciones, en la región Centro del país es donde hay más reclamaciones.

3. ¿Los costos más elevados en las reclamaciones corresponden aquellas que son tratamientos?

Se entiende tratamiento como un conjunto de acciones, terapias y fármacos que buscan propiciar al paciente de manera curativa síntomas dados por una patología. En la siguiente tabla se muestra los costos por los tratamientos y los que no son tratamientos.

	Costos
Tratamientos	39.856.416.353,96
Otras reclamaciones	74.788.798.303,56
Total	114.645.214.657,52

Tabla 4. Costos para las reclamaciones según el diagnóstico requerido.

En el siguiente diagrama de barras se presenta la información en la tabla 3.

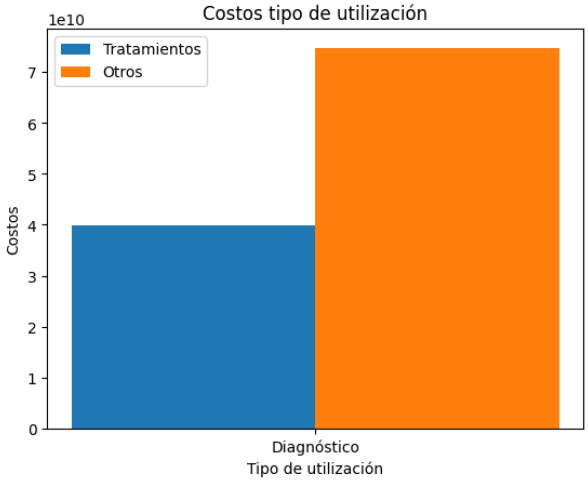


Diagrama de barras 2. Costos por tipo de utilización

Los tratamientos que corresponden a otras reclamaciones por fuera de los tratamientos suelen ser más costosos.

4. De los afiliados que tengan únicamente una enfermedad, aquellos que tienen enfermedades cardiovasculares ¿son quienes representan las reclamaciones con los costos más altos para la aseguradora?

Según un estudio realizado por el Ministerio de Salud, la principal causa de mortalidad en Colombia es la enfermedad isquémica del corazón, relacionada con enfermedades cardiovasculares [3]. En la consulta realizada se obtuvieron los siguientes datos.

	Costos
Enfermedades cardiovasculares	527.881.356
Cáncer	9.567.139.260
EPOC	429.694.086
Hipertensión	9.947.449.706
Diabetes	1.228.899.575

Tabla 5. Costos para las reclamaciones por enfermedad.

Veamos el comportamiento de estos costos en un diagrama

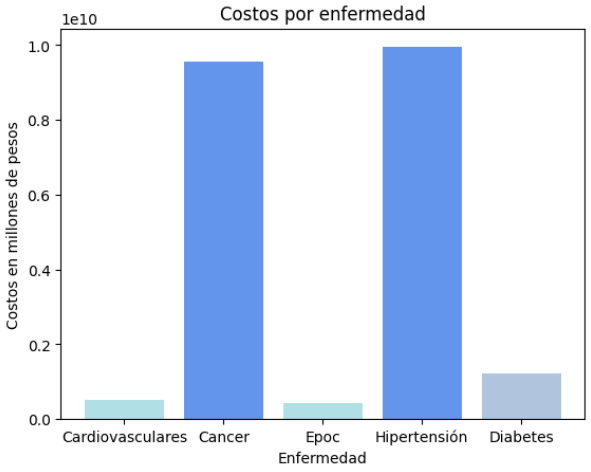


Diagrama de barras 3. Costos por tipo de enfermedad

Tanto en la tabla 4 como en el diagrama de barras 3, se observa que de aquellos pacientes que únicamente poseen una enfermedad, quienes representan los mayores costos para la aseguradora son los hipertensos.

5. ¿Los diagnósticos más recurrentes suelen contribuir en mayor cantidad en los costos para la aseguradora?

Para ejercicios de practicidad, se entiende como mayores recurrentes aquellos diagnósticos que sean solicitados superiores a mil veces. En la siguiente tabla se muestra la información de estas cantidades por diagnóstico.

Diagnóstico	Cantidad	Costo
Pendiente	343.263	139.772,8
Otros controles generales de salud	6.773	515.466,76
Examen de laboratorio	6.470	50.010,44
Hipotiroidismo	2.441	11.969,72
Diabetes Mellitus especificada	1.086	421.456,16

Tabla 6. Diagnósticos más frecuentes con sus costos.

En la tabla se muestra que no existe relación directa entre la cantidad de diagnóstico y la cantidad de costos que representan, el diagnóstico "Otros controles generales de salud" a pesar de no ser el más concurrido representa el mayor costo para la aseguradora.

6. ¿Las reclamaciones con relación a los tratamientos suelen ser más solicitadas en comparación con aquellas que no hacen parte de algún tipo de tratamiento?

Los tratamientos médicos suelen ser más costosos que otros tipos de reclamaciones de seguros, como daños por accidentes de tráfico o daños a la propiedad. Por lo tanto, los asegurados que buscan tratamiento médico para una enfermedad o lesión pueden presentar reclamaciones más altas que aquellos que presentan reclamaciones para otros tipos de daños [4]. En la siguiente tabla se muestran la cantidad de reclamaciones con respecto a los tratamientos y los que no representan tratamiento.

	Cantidad
Tratamientos	17.804
Otras reclamaciones	378.424
Total	395.868

Tabla 7. Recurrencia de reclamaciones entre tratamientos y no tratamientos.

Aquellas reclamaciones que no son tratamientos son las que mayor número de solicitudes presentan.

7. ¿La población con más de cincuenta años presenta un mayor costo para la aseguradora?

A medida que las personas envejecen, aumenta la probabilidad de que sufran enfermedades crónicas y requieran atención médica [5]. A continuación, se presentan los datos obtenidos a partir de la consulta efectuada.

	Costo
Mayores a 50	41.986.176.836,06
Menores a 50	70.298.764.183,7

Tabla 8. Costos asociados a afiliados por edad de 50 años

En el siguiente grafico de tortas se presenta visualmente la proporción de afiliados mayores y menores de 50 años.

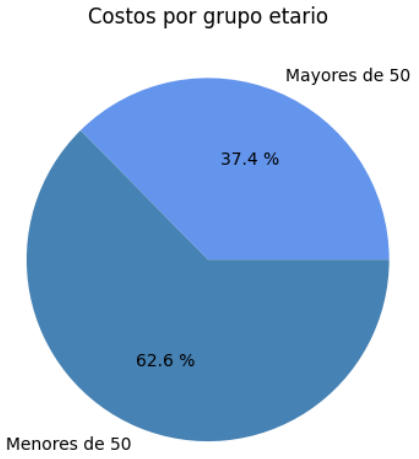


Diagrama de torta 2. Proporción de afiliados por edad de 50.

Los costos asociados a las personas menores a 50 años presentan un mayor valor frente a las personas mayores a 50 años

4. PREGUNTAS DE INVESTIGACIÓN

En este apartado, se plantean siete preguntas de negocios que son complementaria para los análisis propuestos en las hipótesis anteriores, e desarrollo de estas preguntas se realizó en el lenguaje de programación Python, con ayuda de librerías de tratamiento y visualización de datos como Pandas, Numpy y sqlite3.

1. ¿Cuáles son los diez afiliados que más reclamaciones solicitan?

Los afiliados con más reclamaciones se presentan en la siguiente tabla.

Afiliado	Reclamaciones
56682173	285
8133470	225
24695323	203
58181850	160
39327324	157
35423628	146
9732625	143

Tabla 9. Reclamaciones por afiliados

Las mayores cantidades de reclamaciones oscilan entre 143 y 285 unidades, las cuales son realizadas por las personas que mas recurren a esta actividad.

2. ¿Cuáles son los diez afiliados que más costos representan para la aseguradora?

Los afiliados con más costos se presentan en la siguiente tabla.

Afiliado	Costo
16211397	5.472391e+08
48112994	2.727043e+08
4696275	2.302710e+08
8046935	2.279883e+08
15520328	1.849156e+08
5055740	1.831462e+08
8001431	1.788804e+08
4219884	1.775988e+08
9819800	1.737690e+08
1762631	1.708372e+08

Tabla 10. Mayores costos por afiliados

Veamos cómo se presentan las proporciones de los costos de cada afiliado en un diagrama de tortas.

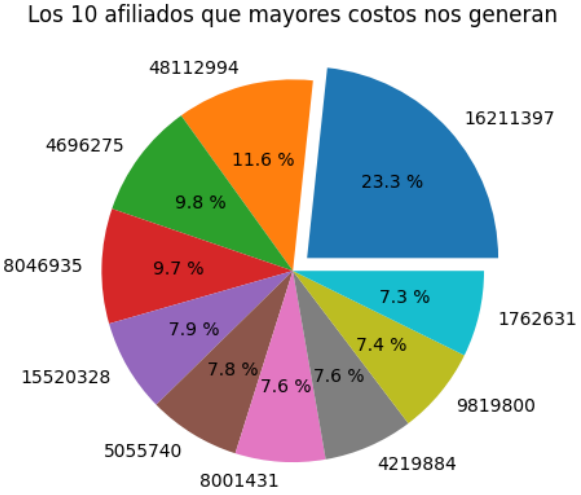


Diagrama de torta 3. Proporción de costos por afiliado.

Al comparar los afiliados en la tabla 8 y la tabla 9 se evidencia que ninguno se presenta en las dos tablas, por lo que los afiliados que más recurren a

reclamaciones no son los que representa mayores costos.

3. ¿Cuál es la enfermedad que más prevalece en los afiliados de la aseguradora?

En la siguiente tabla se muestra la cantidad de afiliados que presentan alguna de las enfermedades principales que trata la aseguradora.

Enfermedad	Presencia en afiliados
Cáncer	4149
EPOC	1116
Diabetes	3239
Hipertensión	15187
Enfermedades Cardiovasculares	883

Tabla 11. Presencia de enfermedades en afiliados.

En el siguiente diagrama de barras se presenta la información de la tabla 10.

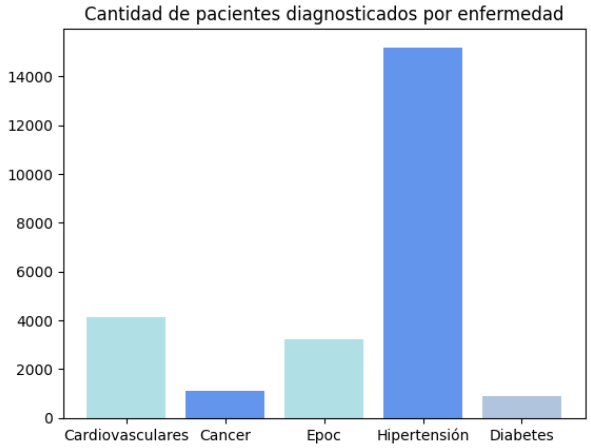


Diagrama de barras 4. Pacientes por enfermedad.

Cómo es evidente tanto en la tabla 10 como en el diagrama de barras 4, la enfermedad que mayor prevalencia tiene en los afiliados es la Hipertensión.

4. De las enfermedades cómo *Cáncer, EPOC, Diabetes, Hipertensión y Cardiovascular*. ¿Cuál es la proporción de afiliados que no poseen ninguna de estas?

La cantidad de afiliados que tiene la aseguradora son 218.205. la cantidad de afiliados que no poseen ninguna de estas enfermedades es de 198.143, quienes representa una proporción del 90.81% respecto al total de afiliados para la aseguradora.

5. ¿Los costos asignados al mayor número de reclamaciones se encuentran por encima del promedio de estos costos?

Los costos promedios de todas las reclamaciones realizadas por los afiliados corresponden a un valor de 298.604,65. La siguiente tabla muestra los costos asociados a la mayor cantidad de reclamaciones.

Afiliado	Reclamaciones	Costo
56682173	285	9130599,88
8133470	225	2056031,6
24695323	203	26668982,56
58181850	160	4525132
39327324	157	3424210,4
35423628	146	2931980
9732625	143	4593159,64

Tabla 12. Costos por mayores reclamaciones

Los costos asociados a la mayor cantidad de reclamaciones se encuentran pronunciadamente por encima del promedio.

6. ¿Cuál es el año en el que más reclamaciones se han realizado?

En la siguiente tabla muestra los años con sus respectivas cantidades de reclamaciones.

Año	Reclamaciones
2019	594.455
2018	1.046
2017	34
2016	18
2014	15
2015	3

Tabla 13. Reclamaciones por año

El comportamiento a lo largo de los años de estas reclamaciones se aprecia en el siguiente grafico de líneas.

Diagrama de líneas 1. Reclamaciones por año

Las reclamaciones han ido aumentando en proporciones desiguales a medida que aumentan los años. En los años más recientes es donde se presentan mayor número de reclamaciones.

7. ¿Cuál es la región con menor cantidad de reclamaciones?

En la siguiente tabla se muestra la cantidad de reclamaciones por regiones.

Región	Reclamaciones
Centro	302592
Occidente	106570
Antioquia	92157
Norte	72796
Eje Cafetero	14749

Tabla 14. Reclamaciones por regiones

Para mirar de mejor manera la proporción que representan las reclamaciones en cada región, se realiza el siguiente diagrama de torta.

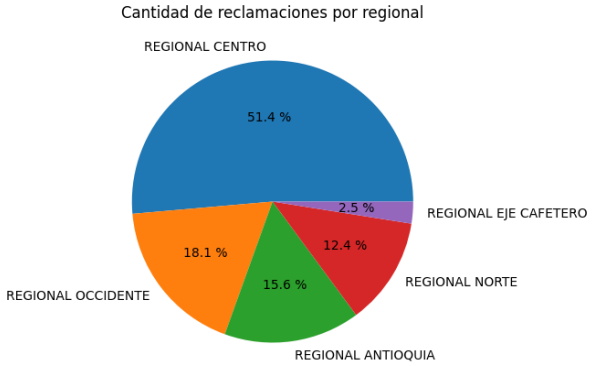


Diagrama de torta 4. Proporción de reclamaciones por región.

Se evidencia que la región con menor cantidad de reclamaciones es el Eje Cafetero, la cual es la región más pequeña de la zona de servicio de la aseguradora.

5. LIMPIEZA Y TRANSFORMACIÓN

Dado que en algunas observaciones en las tablas 'Sociodemográficas' y 'Utilizaciones' presentan algunos datos nulos y otros sin información, se procede hacer la limpieza de estos por medio de los métodos de eliminación de filas y nulos en Python, con el fin de obtener resultados de valor a partir de un buen tratamiento de la información inicial.

Debido a que es necesario hacer algunas modificaciones de agregación y conversión de la información para una mayor comprensión de estos, se procede hacer las siguientes transformaciones para las tablas iniciales y los resultados de las hipótesis plantadas.

-**Transformación 0:** Para el data frame 'Sociodemográficas' se generó una nueva tabla con el fin de mostrar el promedio de las personas que sufrían de cada una de las enfermedades presentes en los afiliados.

-**Transformación 1:** En la importación de la base de datos, el data frame 'Utilizaciones' en la variable 'Cantidad' no se visualiza el tipo de dato como un entero sino como una float, por lo que se hace la transformación de todas las observaciones de esta variable para que el tipo de dato sea entero.

-Transformación 2: Para los resultados de la pregunta de investigación 2, el formato de los valores de la variable costo estaban en notación científica de 1e08, por lo que se hace la transformación de esta columna para que el costo se presente en millones.

-Transformación 3: Con el propósito de comparar las proporciones de los afiliados con enfermedades en la pregunta de investigación 3, se agrega una nueva columna relacionando la cantidad de personas con cada enfermedad con el total de afiliados enfermos.

-Transformación 4: En la pregunta de investigación 5, se agrega una nueva columna en la tabla de resultado la cual alberga los precios unitarios por la cantidad de cada reclamación realizadas por los afiliados.

6. VISUALIZACIÓN DE DATOS

La visualización de datos es crucial para comprender y comunicar información compleja de manera clara y efectiva. Permite identificar patrones, tendencias y relaciones ocultas, facilitando la toma de decisiones informadas, por lo que a continuación se presentan algunas gráficas de la tabla de datos 'sociodemográficas'.

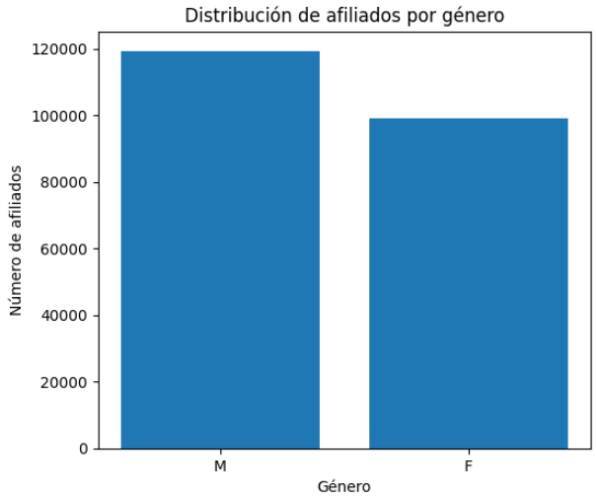


Diagrama de barras 5. Distribución de afiliados por género

Se presentan mayor cantidad de afiliados masculinos que afiliados femeninos, este comportamiento puede ayudar a entender las preguntas de hipótesis planteadas respecto a dichos géneros.

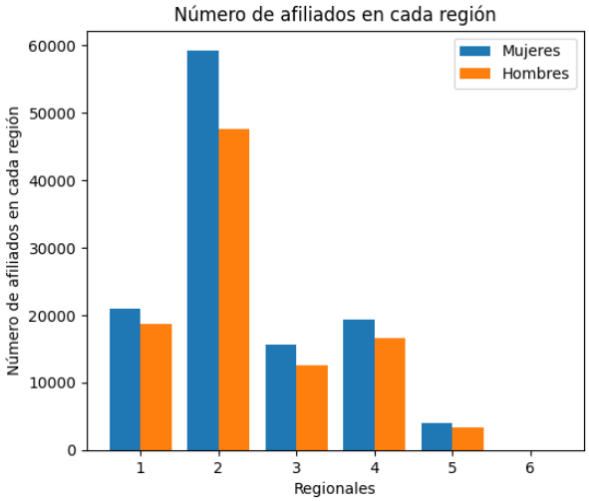


Diagrama de barras 6. Distribución de afiliados por género

Este diagrama de barras muestra cómo se distribuyen los géneros en las seis regiones donde hace presencia la aseguradora. Esta visualización es útil al momento de analizar las preguntas de investigación e hipótesis relacionadas con los géneros en las distintas regiones.

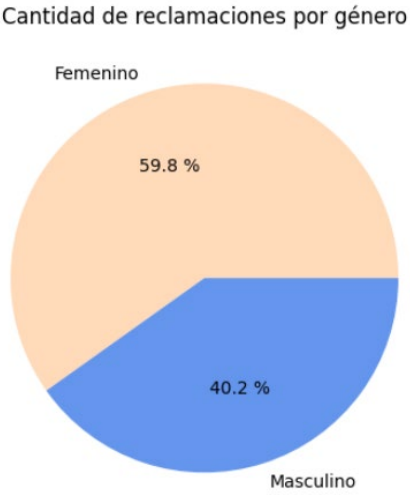


Diagrama de torta 5. Cantidad de reclamaciones por género

A pesar de que la mayoría de los afiliados en la aseguradora son masculinos, las mujeres representan mas de la mitad de las reclamaciones. Este gráfico de total es útil para contrastar posibles comportamientos y resultados de las hipótesis planteadas.

7. CONCLUSIONES

1. Las mujeres representan el 60% de las reclamaciones realizadas dentro de la aseguradora
2. El regional centro es donde hay mayor número de reclamaciones
3. La hipertensión y el cáncer son las enfermedades que mayores costos le generan a la compañía
4. Es mayor la proporción de afiliados menores de 50 años que la proporción de mayores de 50 años
5. Para el año 2019 hubo un incremento sustancial en el número de reclamaciones realizadas dentro de la compañía.
6. Hay una tendencia a reclamaciones con costos más bajos.

8. REFERENCIAS

[1] Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: The new science of winning. Harvard Business Press.

[2] Bardey, D., Zapata, J.G., Buitrago, G., Concha, T. (2013). Mercado de seguros voluntarios de salud en Colombia. Recuperado de Mercado de seguros voluntarios de salud en Colombia.

[3] Ministerio de Salud y Protección Social. Principales causas de mortalidad en Colombia. (2010). Recuperado de: https://www.minsalud.gov.co/salud/Paginas/Enfermedadescardiovasculares.aspx

[4] Newhouse, J. P., Manning, W. G., Morris, C. N., Orr, L. L., Duan, N., Keeler, E. B., & Leibowitz, A. (1981). Some interim results from a controlled trial of cost sharing in health insurance. New England Journal of Medicine, 305(25), 1501-1507. doi: 10.1056/NEJM198112173052504

[5] Schieber, G. J., Poullier, J.-P., Greenwald, L. M., Glickman, A., Ossowski, S., & Shoven, J. B. (1991). The aging of the population and the rising cost of health care: Demographic and macroeconomic effects. Urban Institute Press.