INTELIGENCIA DE NEGOCIOS

Proyecto 2 – Análisis sobre la encuesta multipropósito

PRESENTADO POR:

Paula Daza Díaz – 202111276
Sofía Torres Ramírez – 202014872
Juan Camilo Reyes - 201922989

2 de Diciembre del 2023

PROFESOR:

Fabián Peña Lozano

UNIVERSIDAD DE LOS ANDES
DPTO. INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
INTELIGENCIA DE NEGOCIOS
BOGOTÁ D.C

INDICE

1. Identificar necesidades analíticas.

Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - Tablero de control, análisis OLAP, Minería de datos	Procesos de negocio	Fuentes de datos y datos
Impacto de factores ambientales en el desarrollo de enfermedades respiratorias	Existe contaminación en el aire en la zona de vivienda	Tablero de control	Identificación causas para medicina preventiva	Encuesta multiproposito Capitulo B: Factores ambientales
	Existe generación y manejo inadecuado de basuras en la zona de vivienda			Encuesta multiproposito Capitulo B: Factores Ambientales y Condiciones del hogar
	Existe servicio de alcantarillado en la zona de vivienda			Encuesta multiproposito Capitulo B: Factores Ambientales y Condiciones del hogar
Impacto de las condiciones de la zona de vivienda en el desarrollo de enfermedades respiratorias	Upz en la que se encuentra la vivienda	Tablero de control	Identificación causas para medicina preventiva	Encuesta multiproposito Capitulo A: Codigos_UPZ
	La zona de vivienda queda cerca a fabricas o zonas industriales			Encuesta multiproposito Capitulo B: Condiciones de la zona de la vivienda
	Hay presencia de insectos en la zona de vivienda			Encuesta multiproposito Capitulo B: Condiciones de la zona de la vivienda
	La vivienda queda cerca a basureros			Encuesta multiproposito Capitulo B: Factores Ambientales y Condiciones del hogar
	Estado de la via para acceder a la vivienda			Encuesta multiproposito Capitulo B: Condiciones de la zona de la vivienda
Impacto de las condiciones fisicas de la vivienda en el desarrollo de enfermedades respiratorias	Estrato de la vivienda	Tablero de control	Identificación causas para medicina preventiva	Encuesta multiproposito Capitulo B: Condiciones del hogar
	Existe humedad en la vivienda			Encuesta multiproposito Capitulo B: Condiciones del hogar
	Hay ventilación adecuada en la vivienda			Encuesta multiproposito Capitulo B: Condiciones del hogar
	Hay servicio de acueducto para la vivienda			Encuesta multiproposito Capitulo B: Condiciones del hogar
	Clase a la que pertenece la vivienda			Encuesta multiproposito Capitulo A: Clase
	Existen insectos en la zona rural			Encuesta multiproposito Capitulo B: Condiciones de la zona de la vivienda
	Codigo UPZ de la vivienda			Encuesta multiproposito Capitulo A: Codigos_UPZ
Impacto de la frecuencia de las visitas al médico en el desarrollo de enfermedades respiratorias	Por lo menos una vez al año va al médico	Tablero de control	Identificación causas para medicina preventiva	Encuesta multiproposito Capitulo F: Salud

Después de discutir con el cliente de medicina se llegaron a las siguientes necesidades analíticas:

- Impacto de factores ambientales en el desarrollo de enfermedades respiratorias: Analizar la contaminación del aire y el manejo de basuras proporciona información crucial para implementar medidas preventivas en salud pública, abordando las causas ambientales de enfermedades respiratorias y mejorando la calidad de vida.

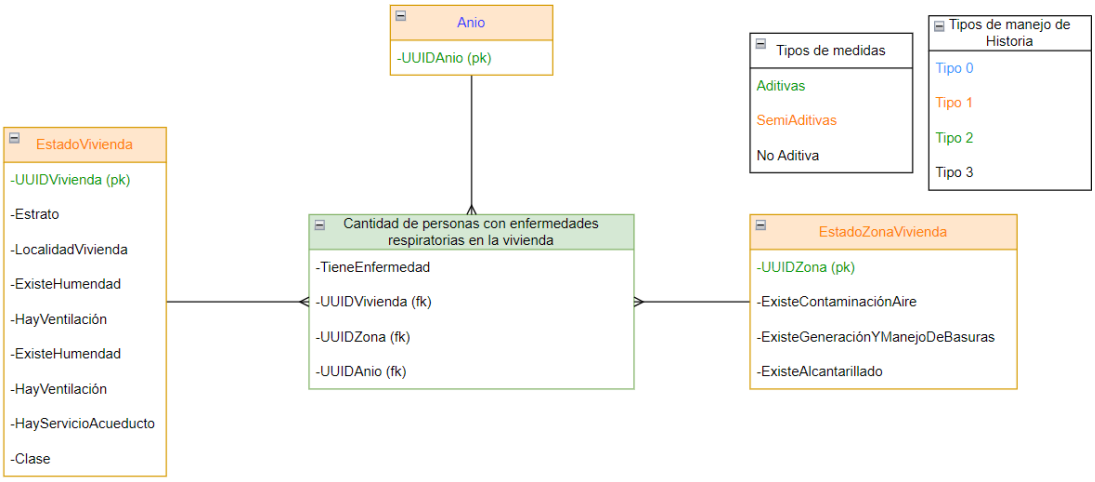
- Impacto de las condiciones de la zona de vivienda en el desarrollo de enfermedades respiratorias: Entender cómo la ubicación de la vivienda y las condiciones del entorno afectan la salud respiratoria permite identificar áreas de

riesgo y desarrollar estrategias para mejorar las condiciones habitacionales, reduciendo así la incidencia de enfermedades respiratorias.

- Impacto de las condiciones físicas de la vivienda en el desarrollo de enfermedades respiratorias: Analizar aspectos como el estrato, la humedad, la ventilación y otros factores de la vivienda es crucial para personalizar intervenciones preventivas, ya que estas condiciones pueden influir directamente en la salud respiratoria de los habitantes.

- Impacto de la frecuencia de las visitas al médico en el desarrollo de enfermedades respiratorias: Entender la relación entre la frecuencia de visitas al médico y las enfermedades respiratorias permite evaluar la efectividad de las medidas de prevención, identificar patrones de atención médica y adaptar estrategias para fomentar un cuidado de la salud más proactivo.

2. Modelar Data Marts:
   a. Modelo multidimensional



   b. Justificación del modelo

Granularidad: Se observa que el detalle en el cual se analizaran los datos es a nivel de las viviendas individuales esto dado a la cantidad de datos nulos a nivel de hogar.

Hecho: se toma como hecho la cantidad de personas con enfermedades respiratorias en una vivienda, tieniendo en cuenta que resolver esta pregunta puede traer información valiosa sobre las encuestas multidimensionales

Dimensiones: se describen 3 dimensiones

- Año: la cual describe el año en el cual se realizó la encuesta
- EstadoVivienda: que describe las condiciones de la vivienda por medio de las medidas
  - Estrato
  - LocalidadVivienda
  - ExisteHumedad

    - HayVentilación
    - ExisteHumedad
    - HayServicioAcueducto
    - Clase
- EstadoZona: donde se modela las condiciones de la zona donde se ubica la vivienda
    - ExisteContaminaciónAire
    - ExisteGeneraciónYManejoBasuras
    - ExisteAlcantarillado

Tipos de medidas: Tomando en cuenta las necesidades del modelo se definen los tipos de las medidas:

Aditiva	No aditiva
Todas las Pk de las dimensiones, Ya que deben ser Incluidas en todas las dimensiones	Todas las medidas restantes las cuales son características específicas de cada dimensión (zona,Vivienda)

Tipos de Historia: dadas las dimensiones planteadas se plantean los siguientes tipos de manejo de historia:

- Año: se utilizará un manejo tipo 0. Es decir, se eliminarán todas las nuevas entradas ya que los años de las encuestas son fijos por lo que al intentar sobre escribirlo no se agregara información útil o incluso de podrían cometer errores.
- EstadoVivienda: se utilizara un manejo tipo 1, donde se eliminan las entradas anteriores sin llevar un registro de ello. Esto dado que la información de las encuestas es estática y tomada en un momento exacto en la historia, toda información que sea reemplazada se tomará como la muestra que se decidió guardar para futuro estudio.
- EstadoZons: se utiliza un manejo tipo 1, esto dado por las mismas razones por las que se utiliza tipo 1 en la dimensión EstadoVivienda.

3. Entendimiento de los datos, creación de Data mart y proceso ETL:
    a. Entender las fuentes de datos

Los datos que hemos obtenido proceden de una encuesta multipropósito realizada en los años 2017 y 2021. Esta encuesta tiene como propósito recopilar información variada sobre diversos temas para ofrecer una visión completa de la situación socioeconómica y demográfica de la población colombiana. Se ha seleccionado una muestra representativa para asegurar que los resultados reflejen de manera precisa la realidad de toda la población. Estos datos recopilados se utilizan para crear indicadores y estadísticas que permiten comprender y analizar varios aspectos de la sociedad. En nuestro caso, nos interesa analizar indicadores que puedan tener impacto en enfermedades respiratorias.

Para llevar a cabo nuestro proyecto, contamos con cuatro fuentes de datos correspondientes a los capítulos A, B, C y F de la encuesta en cada año. Tenemos

respuestas de la encuesta de personas con y sin enfermedades respiratorias correspondientes a los años 2017 y 2021. En ambos casos, disponemos de una cantidad considerable de respuestas, lo que indica que esto no representa una limitación al presentar nuestras conclusiones.

La encuesta abarca una amplia gama de preguntas sobre diversos aspectos sociales y socioeconómicos. Se divide en 14 categorías de variables, cada una con entre 10 y 250 preguntas. Algunas de estas variables contienen muchos valores nulos, por lo que hemos procurado no basarnos en ese tipo de variables. En total, había más de 550 variables, de las cuales hemos seleccionado 12 para enfocar nuestro análisis:

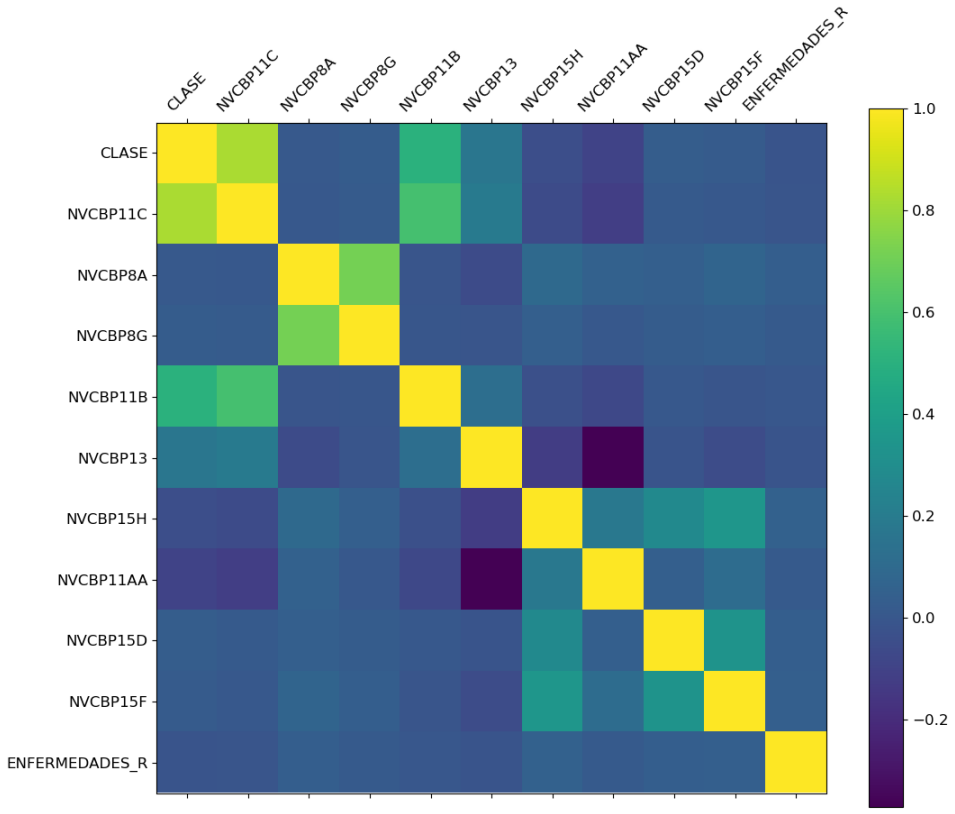
- DIRECTORIO: Identificador único de cada vivienda
- LOCALIDAD: Localidad de la vivienda
- CLASE: Área de la vivienda
- NVCBP15D: Contaminación del aire
- NVCBP15F: Disposición inadecuada de las basuras
- NVCBP11C: ¿Dispone de alcantarillado?
- NVCBP11AA: Estrato
- NVCBP8A: Humedades en el techo o paredes
- NVCBP8G: Escasa ventilación
- NVCBP11B: ¿Dispone de acueducto?
- NVCBP15H: Presencia de insectos, roedores o animales molestos
- ENFERMEDADES_R: ¿Presenta enfermedades respiratorias?

A continuación, se muestra un sample de la representación de las variables en los dataframes:

DIRECTORIO	NVCBP15D	NVCBP15F	NVCBP11C	NVCBP11AA	NVCBP8A	NVCBP8G	NVCBP11B	NVCBP13	NVCBP15H
1028887.0	2	2	1	2.0	2	2	1	4	2
1105395.0	1	1	1	2.0	1	2	1	4	2
1001469.0	2	2	1	3.0	2	2	1	4	1
1026321.0	2	2	2	1.0	1	2	1	6	2
346046.0	2	2	1	3.0	2	2	1	4	2

Durante el análisis exploratorio de los datos, se evidenció una gran cantidad de duplicados, por lo que se llevó a una transformación de estos datos donde se corrigió este porcentaje de duplicados y adicionalmente, se transformaron otros aspectos a tener cuenta para obtener una mayor calidad de la fuente de datos y posteriormente un mejor manejo de estos mismos. Entre estas transformaciones están: estandarizar formatos de los valores de las variables, para las localidades con valores nulos se define una fila con nombre ‘desconocido’ para agrupar a estas viviendas, eliminar columnas vacías, entre otros.

Además de esto, se tuvo en cuenta la correlación de estas variables para tener evitar redundancias y un mejor análisis:



Se puede evidenciar que hay baja correlación entre la mayoría de las variables a analizar.

b. Diseñar e implementar el proceso de ETL

A continuación, se presenta una explicación detallada del proceso ETL desarrollado para el proyecto, teniendo en cuenta las dos fuentes de datos resultantes de la limpieza de datos: datos2017.csv y datos2021.csv. Se analizarán cada una de las tres dimensiones creadas:

1. **Filtrar Columnas de las Dimensiones:**

Para cada fuente de datos se filtraron las columnas correspondientes a cada dimensión:

- ViviendaDim: Comienza con el filtro de seleccionar las columnas DIRECTORIO, LOCALIDAD, CLASE, NVCBP11AA-Estrato, NVCBP8A-Humedades, NVCBP8G-Escasa ventilación, NVCBP11B-Acueducto, NVCBP15H-Presencia de insectos, roedores o animales molestos, ENFERMEDADES_R. Se genera un UUID para cada registro de la dimensión y finalmente se carga la tabla.

- AnioDim: Comienza con el filtro de seleccionar las columnas DIRECTORIO, Anio y ENFERMADES_R. Se genera un UUID para cada registro de la dimensión y finalmente se carga la tabla.

- ZonaDim: Comienza con el filtro de seleccionar las columnas NVCBP15D-Contaminación del aire, NVCBP15F-Disposición inadecuada de las basuras, NVCBP11C-Alcantarillado, DIRECTORIO y ENFERMEDADES_R.

2. Creación de la tabla de hechos.

- Se realizó un inner join entre las dimensiones de ZonaDim y ViviendaDim por medio de la variable DIRECTORIO, para luego realizar otro inner join con la dimensión AnioDim por medio del directorio de igual manera. Se eliminaron las columnas sobrantes y se extrajeron los UUID de cada dimensión junto con el valor de de ENFERMEDADES_R la cual se renombró como 'enfermedad', la cual representa si ese registro presentó o no enfermedad respiratoria. Finalmente, se cargó la tabla con las columnas de UUIDZona, UUIDVivienda, UUIDAnio y Enfermedad.

- Diagramas de bloques final:

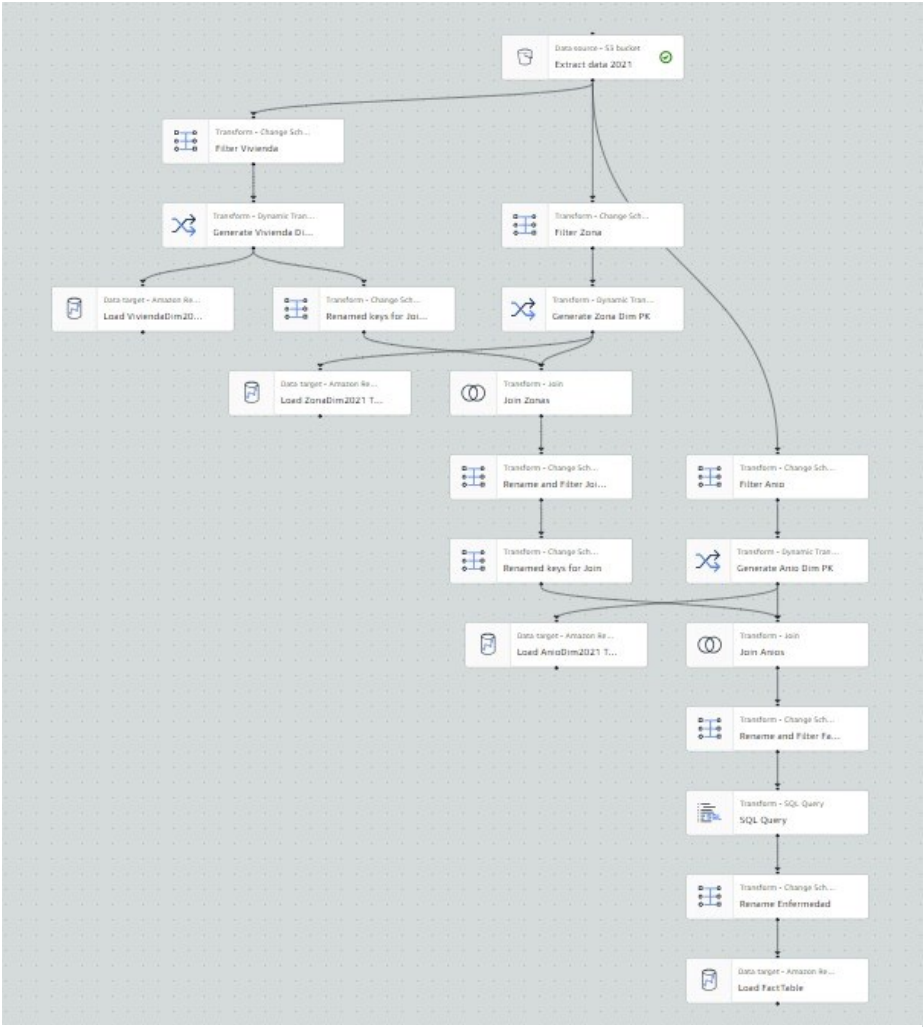


Ilustración 1 Proceso ETL Datos 2021

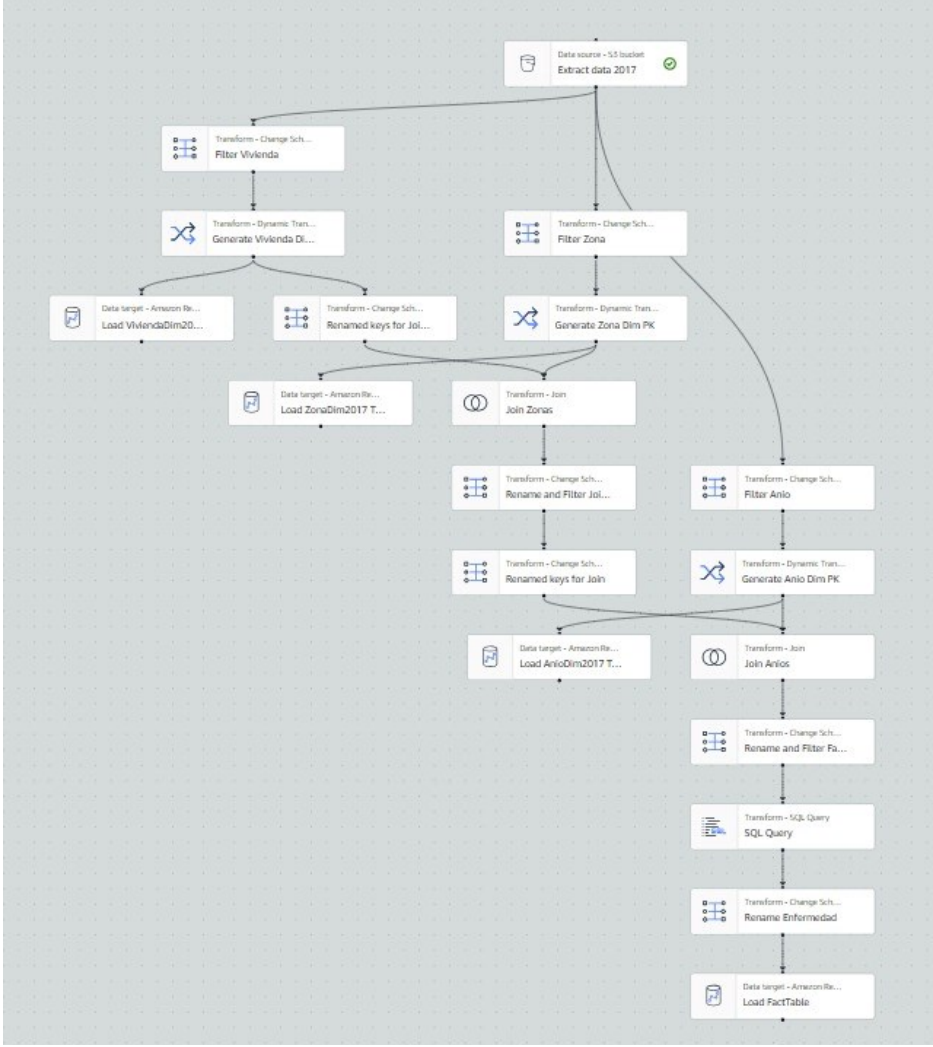


Ilustración 2 Proceso ETL Datos 2017

4. Arquitectura de solución
   a. Proponer arquitectura de solución

La pregunta planteada, "¿Cuál es el Impacto de factores ambientales y condiciones físicas del hogar en el desarrollo de enfermedades respiratorias a través de los años 2017 y 2021?", requiere una solución analítica que permita analizar diferentes aspectos relacionados con las condiciones de vivienda, las condiciones físicas del hogar y su impacto en el desarrollo de enfermedades respiratorias, adicional a la evaluación de estas variables en dos años particulares, 2017 y 2021. Para abordar esta pregunta de manera integral, se propone utilizar un proceso de Extracción, Transformación y Carga (ETL) para construir dos Data Mart, uno para cada año.

Los ETL permitirá extraer los datos relevantes de diferentes fuentes, como los datos de condiciones de vivienda y su entorno y los datos de salud de las diferentes viviendas analizadas, provenientes de la encuesta multipropósito. Luego, se realizará una transformación de los datos para estructurarlos de manera adecuada (obtener los campos necesarios, cerciorarse de que las dimensiones de estos campos concuerden, etc.) y se cargarán en un Data Mart. Como se mencionaba anteriormente es importante tener en cuenta que al estar evaluando estas condiciones de salud en dos años diferentes, se generan dos ETLs exactamente iguales pero manejando datos distintos

En los Data Mart, se organizarán las dimensiones y medidas necesarias para abordar cada uno de los requerimientos analíticos planteados. Cada dimensión de los Data Marts representarán un requerimiento analítico específico, como las condiciones de vivienda y los factores ambientales. Las medidas estarán relacionadas con la prevalencia de enfermedades respiratorias y otros indicadores relevantes.

Al construir los Data Marts de esta manera, se establece una estructura que permite analizar los datos de manera eficiente y responder a las preguntas planteadas en cada uno de los requerimientos analíticos. La organización por dimensiones y medidas facilitará la exploración de los datos y la generación de conclusiones.

Posteriormente, utilizando la herramienta de Business Intelligence seleccionada, se podrán construir tres dashboards que aborden los requerimientos analíticos y, en última instancia, respondan a la pregunta general planteada. Cada dashboard estará compuesto por diferentes visualizaciones, gráficos y métricas relevantes, que permitirán analizar y presentar los resultados obtenidos.

En resumen, mediante el uso de un proceso de ETL y la construcción de dos Data Marts que reflejen los requerimientos analíticos, se busca obtener una solución que facilite el análisis y la visualización de los datos relacionados con las condiciones de vivienda, su entorno y la prevalencia de asma. Los dashboards resultantes proporcionarán una visión completa y comprensible de los hallazgos, permitiendo concluir si existe o no una relación entre las condiciones de la vivienda y el asma en los ocupantes.

Para generar un mayor entendimiento de este proceso y arquitectura se resumen los pasos ejecutados en el siguiente diagrama:

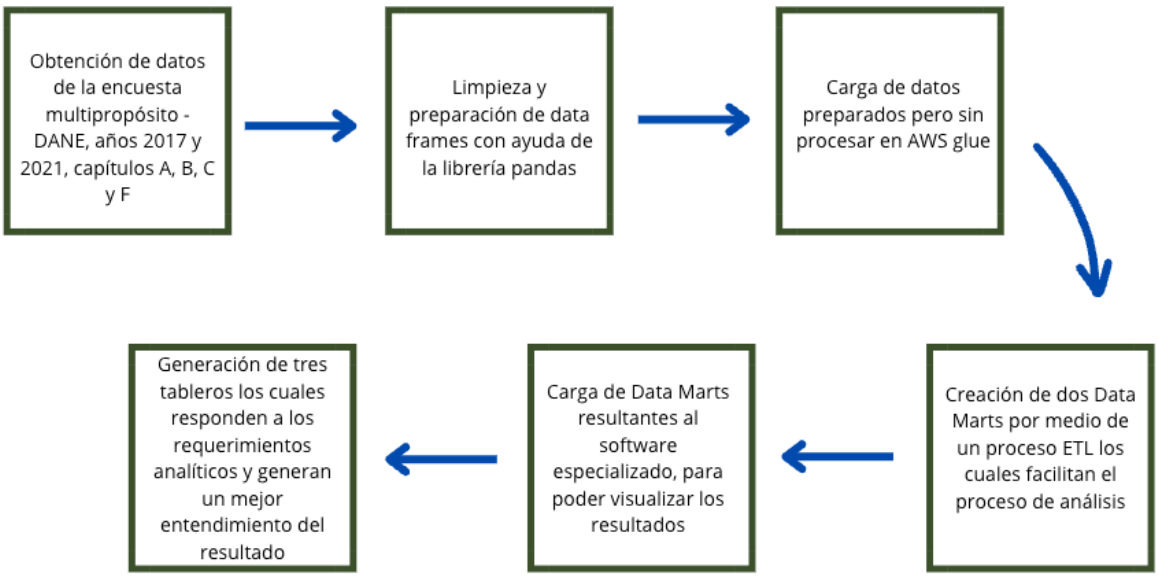


Ilustración 3 proceso de solución

b. Implementar tablero de control

Se realizaron tres dashboards para agrupar los requerimientos analíticos de manera estratégica con el fin de brindar una visión más clara y específica de los análisis relacionados con enfermedades respiratorias y sus factores de influencia.

En el primer dashboard se encuentra la visualización del primer requerimiento analítico “Relación de factores ambientales con el desarrollo de enfermedades respiratorias a través de los años 2017 y 2021”. Por medio de este se busca explorar la relación entre los factores ambientales que afectan cada vivienda y el desarrollo de enfermedades respiratorias, esto comparando los resultados generados en 2 años diferentes. Al representar gráficamente estos resultados, se puede analizar conjuntamente la importancia de cada factor (Disposición de basuras, contaminación del aire y alcantarillado) en ambos años y su impacto en la prevalencia de enfermedades respiratorias en las viviendas. Al agrupar estos requerimientos, se proporcionará una comprensión más completa de cómo estos factores ambientales pueden estar asociados con el riesgo y la gravedad de enfermedades respiratorias.

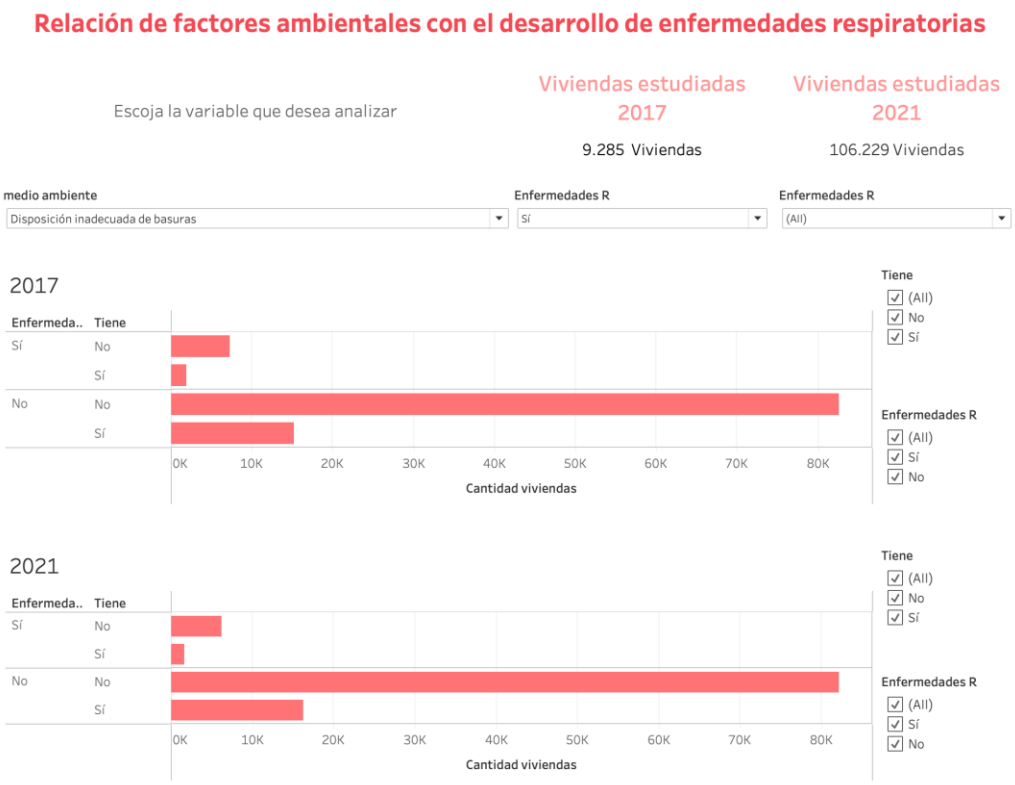


Ilustración 4 Dashboard: Realción de factores ambientales con el desarrollo de enfermedades respiratorias a través de los años 2017 y 2021

En el segundo dashboard se encuentra la visualización del segundo requerimiento analítico “Relación de condiciones de la vivienda con el desarrollo de enfermedades respiratorias a través de los años 2017 y 2021”. Por medio de este se busca explorar la relación entre las condiciones de vivienda y el desarrollo de enfermedades respiratorias, esto comparando los resultados generados en 2 años diferentes. Al representar gráficamente estos resultados, se puede analizar conjuntamente la importancia de cada condición (Clase, estrato, localidad, humedad, humedad, ventilación, etc.) en ambos años y su impacto en la prevalencia de enfermedades respiratorias en las viviendas. Esto proporcionará una visión holística de cómo las condiciones de vivienda y la ubicación geográfica pueden estar interrelacionadas y afectar la salud respiratoria de las personas.

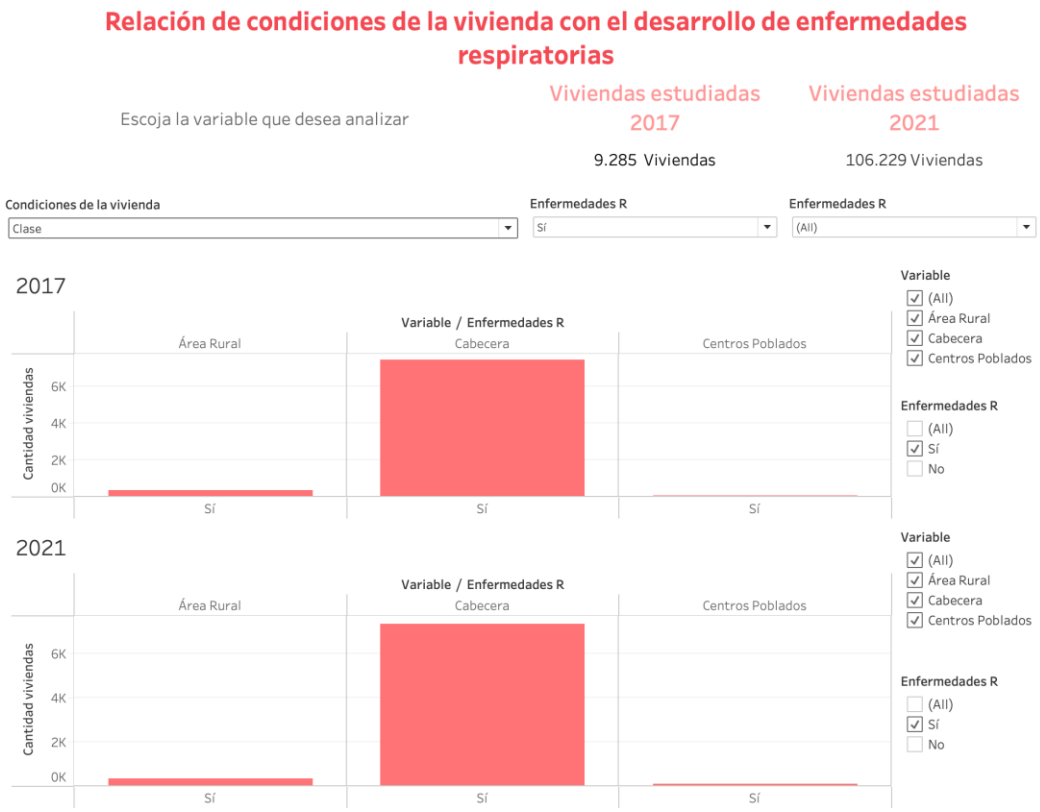


Ilustración 5 Dashboard: Realción de condiciones físicas de la vivienda con el desarrollo de enfermedades respiratorias a través de los años 2017 y 2021

Por último, en el tercer dashboard se busca relaciona el primer y segundo requerimiento analítico al querer identificar la incidencia de factores ambientales en viviendas que presentan enfermedades respiratorias por localidad y a su vez poder comparar esta incidencia entre los años 2017 y 2021. Sin embargo se buscó ir un poco más allá y no solo representar estos datos en forma de gráficas sino con ayuda de mapas de la ciudad para así poder identificar estas incidencias de forma más clara.

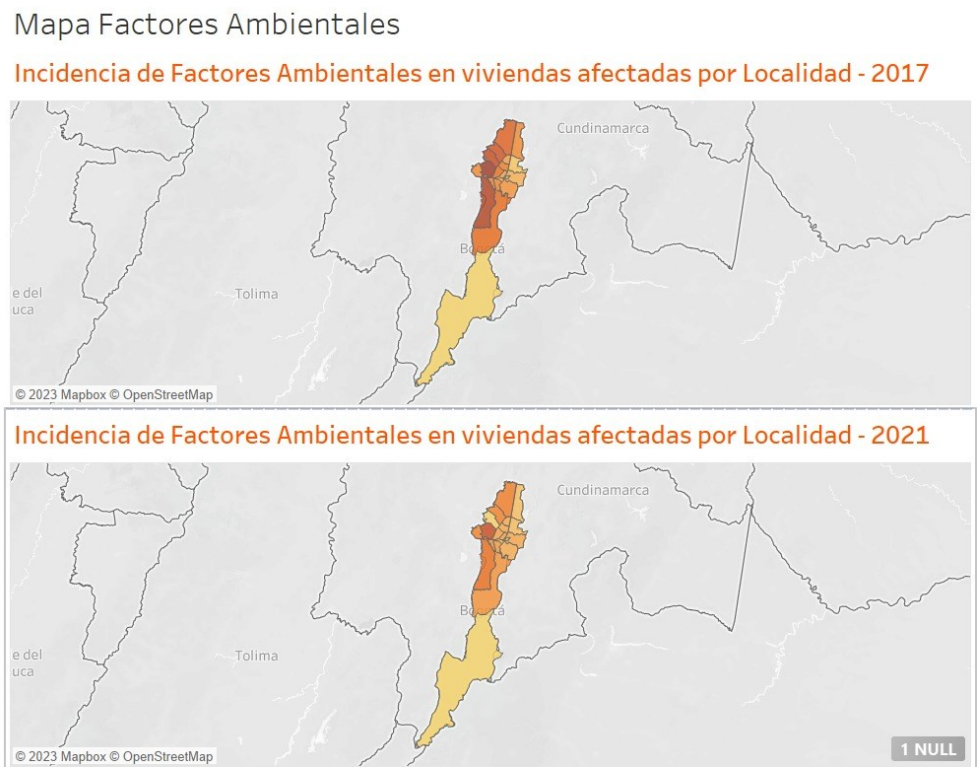


Ilustración 6 Dashboard: Incidencia de factores ambientables en viviendas con enfermedades respiratorias por localidad a través de los años

Finalmente, al dividir los requerimientos analíticos en dos dashboards, se facilita la exploración y el análisis de los datos de manera más enfocada y específica, brindando una visión detallada de la relación entre las condiciones de vivienda, y sus factores ambientales en la generación de enfermedades respiratorias.

5. Link del video

   https://youtu.be/ZO7158a4y9s

6. Actividades realizadas

A pesar de que todos contribuimos en cada entregable del proyecto, de manera general, se podría decir que hubo un líder encargado de asegurarse que se cumplieran con los siguientes tres aspectos indispensables en el proyecto:

- - Diseño del modelo multidimensional y documentación del proyecto: Juan Camilo Reyes
- - Construcción del ETL: Paula Daza.
- - Construcción de los dos dashboards solicitados: Sofia Torres

Por lo tanto, los 100 puntos de contribución fueron repartidos de manera igualitaria para cada integrante:

- - Juan Camilo Reyes: 33 puntos.
- - Sofia Torres: 33 puntos.
- - Paula Daza: 33 puntos.