Entendimiento de datos

Entregable 1 del proyecto
Tableros de control RaSa

Facultad de Ingeniería Industrial

Asignatura:

Modelado de datos y ETL

Presentado por:

Margarita Bain
Rodolfo Moreno
Luis Ortega
William Pinilla

Noviembre 2024

Universidad de los Andes

1 Conclusiones fuente área servicio

1.1 Conclusiones sobre el entendimiento general de los datos

Redundancia de Datos: Se observa una redundancia significativa en algunas columnas, como Población y Área, las cuales podrían ser movidas a una tabla separada para evitar la repetición innecesaria. Además, la columna Densidad es redundante, ya que se puede calcular directamente a partir de Población y Área mediante una consulta.

Estructura del Nombre del Área de Servicio: El campo de nombre del área de servicio contiene información compleja que incluye tanto un código como una descripción. En fases futuras, se recomienda extraer y estructurar estos elementos por separado para facilitar la búsqueda y el análisis.

Optimización de la Fecha: La columna Fecha es un excelente candidato para ser transformado en una dimensión independiente que se pueda vincular a una tabla de hechos, mejorando así la estructura del modelo de datos.

Inconsistencias y Calidad de los Datos: Se han identificado problemas de inconsistencia en los valores, junto con registros duplicados y valores vacíos que deben ser limpiados y validados para asegurar la integridad de los datos.

Presencia excesiva de Outliers: Existen demasiados valores atípicos en la tabla, por ejemplo, en Población Área y Densidad.

1.2 Conclusiones sobre las reglas de negocio

Validación de la fuente para responder la siguiente pregunta: ¿Han existido áreas de servicio que sean cubiertas, a nivel de planes, por menos de dos proveedores?

La información disponible en esta tabla es adecuada para establecer relaciones entre planes y áreas de servicio, permitiendo así responder a esta pregunta de negocio. A continuación, se detallan algunos puntos clave relacionados con la cobertura y la integridad de los datos:

1. ¿Las áreas de servicio reportadas cubren todos los condados del país?

Respuesta: Incorrecto. Como se ha observado, la cobertura no abarca todos los condados de los Alpes, dejando áreas sin representación en la base de datos.

2. ¿Las fuentes FuenteAreasDeServicio_Copia_E y FuenteTiposBeneficio_Copia_E comparten información de los años 2017 al 2019?

Respuesta: Incorrecto. FuenteAreasDeServicio_Copia_E solo incluye datos de los años 2017 y 2018. Además, se detecta un registro anómalo con una fecha del año 1800, lo cual parece ser un error en los datos.

3. ¿La empresa comparte 5409 áreas de servicio?

Respuesta: Correcto. Aunque la base de datos registra 5412 áreas de servicio, esta cifra cumple con la regla de negocio al incluir 3 áreas adicionales, lo cual es consistente con la información proporcionada.

1.3 Conclusiones sobre el análisis descriptivo

Columna	Media	Desviación Estándar	Rango	A Resaltar
IdGeografia_T	42,958.96	32,458.82	1,001 - 168,135	La variable tiene una distribución amplia con valores atípicos significativos en el extremo superior, indicando áreas geográficas con identificadores notablemente altos.
PoblacionAct	41.7 millones	59.9 millones	82 - 47.28 mil millones	Hay un valor extremo e irreal de 47 mil millones, lo que evidencia errores graves en los datos.
Área	770.85	1,303.66	-24,707 - 88,824	Se observan valores negativos de área, lo cual es ilógico y sugiere datos

				incorrectos. También hay una gran variación con valores atípicos elevados.
Densidad	276.28	857.47	0 - 14,946	La densidad presenta alta dispersión, con máximos que sugieren áreas densamente pobladas o datos posiblemente incorrectos. Una densidad no puede ser 0 o negativa
Fecha	2010.21	39.04	1800 - 2018	Un valor atípico de 1800 destaca como un error probable, afectando la precisión de las estadísticas temporales.

1.4 Conclusiones sobre la calidad de los datos

De los 188,815 registros totales, se identificaron 43,573 duplicados, dejando 145,242 registros únicos.

Vacíos Significativos

Las siguientes columnas tienen un número considerable de valores vacíos:

IdAreaDeServicio_T: 4,841 vacíos

IdGeografia_T: 4,841 vacíos

Área: 1,835 vacíos

Densidad: 1,835 vacíos

Falta de Información:

Estados Faltantes: RaSa no tiene datos sobre 15 estados, incluidos California, Nueva York, y Washington.

Cobertura de Condados: De los 3,243 condados en Alpes, solo se tiene información de 1,398.

Inconsistencias y Anomalías:

Área: Hay 6,177 registros con valores negativos, lo cual es ilógico.

Población Actual: Se encontraron cifras extremadamente altas, como 6.43 mil millones de habitantes, que no coinciden con datos reales. Se requiere validación histórica.

Fecha: Existen 6,282 registros con fechas menores a 1900, lo cual debe ser revisado.

Problemas de Consistencia:

Estructural:

PoblacionAct debería ser un entero, no un flotante.

NombreAreaDeServicio debe estructurarse mejor, separando el código y la descripción.

Densidad podría calcularse con Población y Área.

Semántica:

IdGeografia_T podría ser un identificador externo para evitar duplicar datos.

Contenido:

NombreAreaDeServicio contiene demasiada información no estructurada.

Área debe ajustarse, ya que no debería tener valores negativos.

Densidad tiene valores de cero cuando Población y Área son diferentes de cero, lo cual es inconsistente.

1.5 Conclusiones sobre las correlaciones

Relaciones Significativas:
La relación más fuerte observada es entre PoblacionAct y Densidad ($r \approx 0.90$). Esto es coherente, ya que la densidad depende directamente de la población y el área. Sin embargo, esta alta correlación indica un posible problema de colinealidad que debe considerarse en modelos de regresión, ya que podría afectar la precisión y estabilidad de las estimaciones.

Relaciones Insignificantes:

Variables como IdGeografia_T no presentan correlaciones significativas con otras variables numéricas, lo que refuerza la idea de que es un identificador categórico y no un factor influyente en análisis numéricos.

Distribución de Datos:

Las variables muestran asimetría y la presencia de outliers. Estos valores extremos pueden distorsionar análisis estadísticos y modelos predictivos, por lo que podría ser necesario aplicar transformaciones de datos (como escalado logarítmico) o analizar los datos en subconjuntos más representativos.

2 Conclusiones fuente condiciones pago

2.1 Conclusiones sobre el entendimiento general de los datos

La tabla `FuenteCondicionesDePago_Copia_E` contiene la información de las condiciones de pago asociadas a cada uno de los planes de beneficios de la seguridad social. Esta tabla es importante porque permite medir el costo de dichos planes de beneficios.

Descripción de la tabla.

Variable	Descripción
idCondicionesDePago_T	identificador de la condición de pago.
Descripcion	Descripción de la condición de pago como NoAplica, Después de deducible.
Tipo	Copago o Coseguro o Anticipado

La variable "idCondicionesDePago_T" es de tipo Integer, la variable "Descripcion" es de tipo String y la variable "Tipo" también es de tipo String.

Cada fila registra el tipo de condición de pago, copago, coseguro o pago anticipado, que se asocia con un plan de beneficios existente; además, incluye el nombre y el identificador.

La tabla está compuesta por 31 registros, de los cuales 24 son registros únicos. La llave primaria es la variable idCondicionesDePago_T.

2.2 Conclusiones sobre las reglas de negocio

La empresa nos informa que existen 15 condiciones de copago y 5 de coseguro, respectivamente al revisar la información de la tabla, se observa que este dato es correcto.

2.3 Conclusiones sobre el análisis descriptivo

Se destaca en la variable "idCondicionesDePago_T" el valor mínimo de 9 y máximo de 714, lo que muestra una dispersión de los datos que además no guarda una secuencia incremental, característica propia en el identificador de este tipo de tablas.

La frecuencia relativa muestra que, de los 23 registros únicos de condiciones de pago, 15 corresponden a copago, 5 a coseguro y 3 a otras condiciones de pago, además existen 2 registros NaN.

2.4 Conclusiones sobre la calidad de los datos

Teniendo en cuenta que es una tabla con pocos registros y solo tres variables, dos de tipo string y una de tipo integer que actúa como identificador, se obtienen los siguientes resultados:

Se registran únicamente dos valores vacíos en la columna Tipo.

Siete de los 31 registros son duplicados. Aunque es un porcentaje significativo, estos registros deben ser eliminados.

2.5 Conclusiones sobre las correlaciones

Las características de la tabla, que contiene dos variables de tipo *string* y una variable numérica como identificador, no permiten realizar un análisis de correlaciones, ya que este tipo de análisis no es aplicable a estos tipos de datos.

3 Conclusiones fuente planes beneficio

3.1 Conclusiones sobre el entendimiento general de los datos

1. De acuerdo con lo que evidencia el negocio, nos confirman que tienen 5.409 áreas de servicio; sin embargo, en la fuente evidenciamos 6.497, lo que podría indicar inconsistencias en la fuente o duplicados.

2. El negocio evidencia que las condiciones de copago son 15; sin embargo, se evidencian solo 14, por lo que se debe validar con el negocio si hace falta algún dato.

3. Para IdPlan_T, el negocio nos dice que hay 301 planes para 2017 y 422 para el año 2018, pero se identifican solo 393 valores únicos.

4. Para IdTipoBeneficio_T, el negocio nos dice que hay 170; sin embargo, la fuente muestra 286, por lo que hay que revisar si se tienen duplicados o hay inconsistencias en lo que dice el negocio.

5. Se evidencian solo 5 valores únicos para la fecha, pero en un periodo de tiempo de 2 años, por lo que se deben revisar esas fechas tanto en el formato como con el negocio para validar que sean correctas o si la información se sube en un mismo periodo de tiempo, y por eso tenemos tan pocas fechas.

3.2 Conclusiones sobre las reglas de negocio

Regla 1: Las áreas de servicios reportadas cubren todos los condados del país: IdAreaDeServicio_T se encuentran 2.041 valores nulos lo que hace que se deba validar la información proporcionada ya que debería ser un valor que este reportado dentro de la base.

Regla 2: Los tipos de beneficios con límite cuantitativo deben tener una cantidadLimite diferente de cero en los planes que los ofrecen: La cantidad de registros que no cumplen con la condición de tener un `cantidadLimite` diferente de cero es: 30571. Se evidencia en la base que cantidadLimite tiene nulos, lo que representa un problema para la completitud de la base y no cumple con lo definido por el negocio.

Regla 3: Las fuentes FuenteAreasDeServicio_Copia_E y FuenteTiposBeneficio_Copia_E comparten información de los años 2017 al 2019: Se evidencian que se encuentran fechas en null y años que aparecen como 1920 que no corresponde a los años reportados por el negocio. No se encuentra información de 2019 por lo que hay que revisar con el negocio si 1920 es un error de tipeo y validar los años que se encuentran en null.

Regla 4: La empresa comparte 5409 áreas de servicios y 170 tipos de beneficios: La cantidad de áreas de servicio es 6497, lo cual no coincide con las 5409 áreas esperadas. La cantidad de tipos de beneficios es 286, lo cual no coincide con los 170 tipos esperados: Se debe validar con el negocio la información proporcionada, ya que la cantidad de áreas y tipos de beneficios no coinciden. Por lo tanto, es necesario verificar que toda la información suministrada sea correcta o entender por qué puedo tener más cantidades.

Regla 5: El valor máximo Copago y Coseguro para el año 2018 es respectivamente 3300 y 100: Cantidad de registros con `valorCopago` > 3300 en 2018: 8 Cantidad de registros con `valorCoseguro` > 100 en 2018: 0 De acuerdo con la información proporcionada, tenemos solo 8 registros que no cumplen con las reglas del negocio para valorCopago. Se requiere validar con el negocio si esta información es correcta o si se deben hacer ajustes en la información.

Regla 6: Además, les comparte información de 301 planes para 2017 y de 422 para el año 2018: Planes únicos en 2017: 203
Planes únicos en 2018: 286 El número total de planes únicos en la tabla es: 393
La cantidad de planes por año reportados por el negocio no corresponde a la información proporcionada. Tenemos una menor cantidad para cada uno de los años mencionados.

Regla 7: Existen 15 y 5 diferentes condiciones de copago y coseguro respectivamente: La cantidad de condiciones de copago es 14, lo cual no coincide con las 15 condiciones esperadas. La cantidad de condiciones de coseguro es correcta: 5 condiciones únicas. Se debe validar con el negocio por qué nos hace falta un copago o por qué no se incluyó.

3.3 Conclusiones sobre el análisis descriptivo

1. IdTipoBeneficio: Se observa que existe diversidad de tipos de beneficios. Se deben validar los outliers que muestran valores altos para entender si son correctos o por qué presentan estos valores tan altos. ¿Son casos especiales?

2. IdAreaDeServicio_T: Se observa que existe diversidad en las áreas de servicio, con valores altos. Por lo tanto, los outliers que muestran valores altos se deben validar para entender por qué algunas áreas pueden tener valores altos.

3. IdCondicionDePagoCopago_T: Se observa que la mayoría de los valores se encuentran dentro del rango. No se observan outliers, lo que indicaría una buena consistencia en los valores de las condiciones de pago.

4. ValorCopago: Se observa que la mayoría de los datos se encuentran en la parte baja; sin embargo, se evidencian algunos outliers importantes que deben ser validados para determinar si muestran la cifra correcta, si de acuerdo a las reglas de negocio pueden tener ese valor o si corresponde a un error.

5. CantidadLimite: Tiene varios outliers a pesar de que la mayoría de sus datos se encuentran en la parte baja. Se deben validar si estos valores son correctos, por qué la cantidad límite puede tener valores tan dispersos y a qué hacen referencia. ¿Es alguna condición especial?

3.4 Conclusiones sobre la calidad de los datos

Completitud y Validez de los datos

1. CantidadLimite tiene un 80% de valores nulos, lo que representa un problema para el negocio, ya que ellos definen: "Los tipos de beneficios con límite cuantitativo deben tener una cantidadLimite diferente de cero en los planes que los ofrecen", lo que no permite validar que este límite se esté cumpliendo de acuerdo con el beneficio a evaluar, teniendo en cuenta que afecta la completitud. Se debe validar con el negocio si hay forma de obtener estos datos, ya que son clave para el análisis.

2. IdTipoBeneficio_T y IdAreaDeServicio_T tienen menos del 10% de datos nulos. Es muy relevante poder tener el 100% de los datos, ya que no se puede identificar el tipo de beneficio y no se puede especificar el área a la que pertenece. Estas son claves en los análisis y afectan la completitud.

Consistencia y Exactitud de los datos

Se deben validar 4.127 registros que tienen nulos en las llaves principales y que son esenciales para analizar la base de beneficios.

Consistencia en Valores Numéricos

1. No se encuentran valores negativos para las variables numéricas, por lo que se verifica que no hay inconsistencias en estas columnas.

2. Se encuentran 8 registros a validar con el negocio para entender por qué el valor del copago es superior a la información dada por el negocio.

3. Se presenta un ejemplo de cantidadLimite nulo relacionado con las llaves principales, con el fin de darle al negocio una idea de los registros que tienen esa condición para lograr la completitud de los datos.

Consistencia en Fechas

1. Se evidencian diferentes formatos de fecha dentro de la base, por lo que se deben normalizar en DD-MM-YYYY.

2. Se debe validar con el negocio por qué aparecen fechas de 1920, ya que no corresponden con el periodo de análisis y esto afecta la calidad de los datos.

3. Se debe validar por qué no aparece 2019 en la base.

Consistencia en Condiciones de Pago

Condiciones de Copago Únicas: 14 (esperado: 15)
Condiciones de Coseguro Únicas: 5 (esperado: 5)
Se debe validar con el negocio por qué nos falta 1 condición de copago.

Consistencia en registros duplicados
Se identifican 8,342 registros duplicados que coinciden exactamente en todas sus columnas, por lo que hay que considerar eliminar estos registros para reducir el ruido en la base.

3.5 Conclusiones sobre las correlaciones

1. IdNivelServicio_T y valorCoseguro: presentan una correlación positiva fuerte, lo que indica que a medida que aumenta el nivel de servicio, también aumenta el valor del coseguro. Se esperaría que esto fuera así.

2. IdCondicionDePagoCoseguro_T y valorCoseguro: presentan una correlación positiva fuerte, lo que indica que la condición de pago del coseguro influye en el valor del coseguro.

3. valorCopago y valorCoseguro: presentan una correlación negativa, lo que indicaría que los copagos más altos podrían tener una condición de coseguro más baja.

4 Conclusiones fuente tipo beneficio

4.1 Conclusiones sobre el entendimiento general de los datos

la tabla de beneficios contiene la siguiente información:

- **IdTipoBeneficio_T** que corresponde al identificador del identificador del tipo de beneficio.
- **Nombre** que corresponde a la descripción del tipo de beneficio.
- **UnidadDelLimite** Corresponde a la unidad en la que se expresa el límite del beneficio (si lo hay).
- **EsEHB** Indica si el beneficio es esencial de salud.
- **EstaCubiertaPorSeguro** Esta columna indica si el tipo de beneficio está cubierto por el seguro.
- **TieneLimiteCuantitativo** Esta columna señala si el tipo de beneficio tiene límite cuantitativo, toma valores Yes/No.
- **ExcluidoDelDesembolsoMaximoDentroDeLaRed** Esta columna indica si el tipo de beneficio está excluido del desembolso máximo dentro de la red, toma valores Yes/No.
- **ExcluidoDelDesembolsoMaximoFueraDeLaRed** Esta columna indica si el tipo de beneficio está excluido del desembolso máximo fuera de la red, toma valores Yes/No.
- **Fecha** Indica el año en que se define el tipo de beneficio.

La primera columna - **IdTipoBeneficio_T** -correspondería a la llave primaria de la tabla. Solo las columnas **IdTipoBeneficio_T** y **Fecha** son de tipo entero (INT); las demás columnas son de tipo texto (STRING)

¿Qué representa una fila de esta tabla?

Cada fila de esta tabla representa un tipo de beneficio en el sistema de salud vigilado por RaSA indicando las características generales de cada beneficio como su descripción; unidades del límite; si es un beneficio esencial; si lo cubre el seguro; si se excluye del desembolso máximo dentro y fuera de la red; y el año de definición del beneficio.

4.2 Conclusiones sobre las reglas de negocio

Hay 3 reglas de negocio que hacen referencia a los datos de esta fuente:

- **(REGLA 2)** Los tipos de beneficios con límite cuantitativo deben tener una cantidad límite diferente de cero en los planes que los ofrecen.
- **(REGLA 3)** Las fuentes **FuenteAreasDeServicio_Copia_E** y **FuenteTiposBeneficio_Copia_E** comparten información de los años 2017 al 2019
- **(REGLA 4)** La empresa comparte 5409 áreas de servicios y 170 tipos de beneficios.

A continuación, las conclusiones sobre el cumplimiento de estas reglas en la tabla revisada:

- Existen al menos 44 planes de beneficios para los que no se cumple la regla de negocio # 2, dado que a pesar de ser beneficios con limite cuantitativo en la tabla de planes aparecen con cantidad limite igual a cero o nula.
- Luego de esta breve inspección es posible constatar que la regla de negocio # 2 no se cumple en las tablas analizadas dado que se encuentran al menos 44 beneficios marcados con 'Yes' en la columna "**TieneLimiteCuantitativo**" en la "**FuenteTiposBeneficio**_Copia_E" que no reflejan un límite acorde (diferente de cero) en la "**FuentePlanesBeneficio_Copia_E**".
- No hay datos para el año 2019, lo cual podría indicar falta de datos ya que la regla de negocio # 3 indica que debería haber datos desde 2017 hasta 2019.
- Hay muchos más datos del año 2017 que del año 2018, siendo que aproximadamente el 79.2% de los datos corresponden a 2017.
- Hay un total de 849 filas en la tabla tipo de beneficio, pero solo 578 de dichas filas son únicas. Esto haría pensar que existen entonces 578 tipos de beneficios diferentes, sin embargo, solo hay 178 IDs de tipo de beneficio únicos lo que indica una gran variabilidad sobre los datos de las otras columnas.
- Existen diferentes valores de los atributos de la tabla para un mismo ID, esto hace que esta la columna ID de la tabla tipo de beneficio no pueda ser usada como llave primaria de la tabla.

- Se debe consultar con el cliente acerca de las razones por las que se encuentran registros distintos para un mismo ID de tipo de beneficio.

4.3 Conclusiones sobre el análisis descriptivo

- La columna **IdTipoBeneficio_T** parece estar uniformemente descrita entre 5 y 1055, tomando valores enteros y sin presentar valores atípicos. De esta columna, siendo un índice, no se espera poder llegar a conclusiones relevantes, pero es importante inspeccionarla para establecer si puede llegar a haber vacíos de información ya que los índices numéricos normalmente son consecutivos.
- Las columnas Nombre, **UnidadDelLimite**, **EsEHB**, **EstaCubiertaPorSeguro**, **TieneLimiteCuantitativo**, **ExcluidoDelDesembolsoMaximoDentroDeLaRed** y **ExcluidoDelDesembolsoMaximoFueraDeLaRed** son de tipo texto.
- La columna **Fecha** toma datos enteros entre 2017 y 2018.
- Las columnas **IdTipoBeneficio_T** y **Nombre** tienen 178 valores únicos y una cardinalidad de 0.20966 la más alta del dataframe.
- La columna **UnidadDelLimite** tiene 63 valores únicos y una cardinalidad de 0.07420. Se debe tener en cuenta que por definición esta columna puede estar vacía y de hecho lo está en la mayoría de los registros (559) representando un 65.8% del total (incluyendo los duplicados).
- La columna **EsEHB** cuenta con 3 valores únicos, siendo uno de ellos el valor 'True' que posiblemente puede imputarse como un 'Yes'. Sin embargo, esto debe consultarse con el cliente para definir reglas de imputación. La cardinalidad es de 0.00353.
- La columna **EstaCubiertaPorSeguro** cuenta con 3 valores únicos, siendo uno de ellos el valor 'False' que posiblemente puede imputarse como un 'No'. Sin embargo, esto debe consultarse con el cliente para definir reglas de imputación. La cardinalidad es de 0.00353.
- La columna **TieneLimiteCuantitativo** cuenta con 4 valores únicos, siendo uno de ellos el valor 'Nein' que posiblemente puede imputarse como un 'No' y el valor 'Si' que posiblemente puede imputarse como un 'Yes'. Sin embargo, esto debe consultarse con el cliente para definir reglas de imputación. La cardinalidad es de 0.00471.
- La columna **ExcluidoDelDesembolsoMaximoDentroDeLaRed** cuenta con 3 valores únicos, siendo uno de ellos el valor 'Algunas veces' apareciendo un total de 2 veces. Esto debe consultarse con el cliente para definir se dicho valor debe considerarse como valido para esta columna o si debe imputarse a 'Yes' o 'No'. La cardinalidad es de 0.00353.
- La columna **ExcluidoDelDesembolsoMaximoFueraDeLaRed** cuenta con 2 valores únicos 'Yes' y 'No'. La cardinalidad es de 0.00236.
- La mayoría de los beneficios esenciales de salud están cubiertos por los seguros (66.2%).
- Hay al menos 20 beneficios esenciales de salud no cubiertos por los seguros. Se debe ahondar mas en el entendimiento del significado de la categoría beneficios esenciales.
- La mayoría de los beneficios que NO están **ExcluidoDelDesembolsoMaximoDentroDeLaRed** tampoco están **ExcluidoDelDesembolsoMaximoFueraDeLaRed** (344).
- Se necesita comprender en mayor profundidad el significado de las columnas **ExcluidoDelDesembolsoMaximoDentroDeLaRed** y **ExcluidoDelDesembolsoMaximoFueraDeLaRed**.
- La mayoría de los beneficios cubiertos por los seguros no tienen limite cuantitativo (295).
- Existen al menos 219 tipos de beneficios cubiertos por los seguros que tienen limite cuantitativo.
- Hay solo 1 tipo de beneficio con limite cuantitativo no cubierto por el seguro.

4.4 Conclusiones sobre la calidad de los datos

4.4.1 Unicidad

- Hay un total de 849 filas en el dataframe
- Existen 271 filas duplicadas
- Solo hay 178 IDs unicos de beneficios

4.4.2 Validez

- Para cada ID existe un unico Nombre o descripción del beneficio
- Las columnas **EsEHB**, **EstaCubiertaPorSeguro**, **TieneLimiteCuantitativo**, **ExcluidoDelDesembolsoMaximoDentroDeLaRed** y **ExcluidoDelDesembolsoMaximoFueraDeLaRed** toman principalmente valores 'Yes' y 'No', lo que indica que podrían modelarse como tipo booleano.
- Se necesita una mejor descripción de los atributos de la tabla y un mayor entendimiento del negocio para llegar a mejores conclusiones frente a la validez de los datos registrados.

4.4.3 Completitud

- No existen celdas nulas
- Las celdas **Nombre** y **UnidadDelLimite** contienen espacios, esto puede ser normal dada la naturaleza del tipo de dato y la manera en que están definidos los valores permitidos de estas columnas.
- Hay 2 valores con espacios en la columna **ExcluidoDelDesembolsoMaximoDentroDeLaRed** lo cual no es lo esperado ya que principalmente esta columna toma valores 'Yes'/'No' (ver análisis descriptivo). Sin embargo, como ya se indicó hay 2 registros que toman el valor 'Algunas veces'.

- La columna **UnidadDelLimite** tiene 559 datos vacíos, esto parece estar dado por la naturaleza del atributo, no todos los tipos de beneficio tienen limite cuantitativo.
- No hay celdas con valor cero.

4.4.4 Consistencia

- Existen registros que tienen indicado **UnidadDelLimite** aun cuando indican que no tienen **TieneLimiteCuantitativo**.
- Existen registros que no tienen indicado **UnidadDelLimite** aun cuando indican que Si tienen **TieneLimiteCuantitativo**.
- Se tienen 33 casos como los apenas descritos.

5 Conclusiones de consultoría

Tema analítico	Análisis requeridos o inferidos	Fuentes de datos
Comportamiento desleal de proveedores	**Análisis 1.a:** Dado un proveedor o grupo de proveedores si tiene o ha tenido un comportamiento desleal? **Análisis 1.b:** Dado un rango de fechas se identifican proveedores con comportamientos desleales. Un comportamiento desleal corresponde a proveedores que brinden planes con el mismo tipo de beneficio cuyo valor de copago o coseguro evidencian diferencias mayores al 20%	F2. Beneficios (Tipos de beneficio y condiciones), F3. BeneficiosPlanes

- Para este análisis se tienen en cuenta las variables "IdProveedor_T", "IdTipoBeneficio_T", "valorCopago", "valorCoseguro" y “Fecha” para el mismo beneficio.
- Se denomina a un proveedor desleal cuando la diferencia entre los valores máximo y mínimo de copago o coseguro para un tipo de beneficio excede el 20%.
- Para este análisis se requiere que “IdTipoBeneficio_T” tenga completitud en los datos, por lo que hay que limpiarlo, imputarlo y asegurar que la tabla de dimensión que describe a los tipos de beneficios sea completa y consistente.
- No existen valores nulos o anormales en las columnas de "valorCopago" y "valorCoseguro" que no permitan realizar los cálculos de mínimos, máximos y diferencias porcentuales, los datos registrados son válidos para el análisis requerido.
- Para realizar el análisis con el campo de fecha se debe:
    - Estandarizar el formato de fecha a solo dd-MM-yyyy.
    - Existen 2,056 registros que no tienen fecha y que deben ser validados para el análisis, y se debe determinar si pueden ser imputados o si se pueden traer de otras tablas.
    - Se deben validar los registros que aparecen con fecha de 1920, ya que no parecen estar acordes a lo reportado por el negocio.

Recomendaciones:

- Examinar el comportamiento desleal de los proveedores a lo largo del tiempo es relevante; sin embargo, se recomienda validar el comportamiento por meses para identificar si hay temporadas en las que este comportamiento sea más recurrente y así poder tomar medidas.
- Se recomienda identificar y segmentar por tipo de proveedores para determinar si hay un grupo específico que presente comportamientos desleales. Asimismo, realizar el mismo ejercicio para identificar proveedores con buenas prácticas y aplicarlas a aquellos que no cumplen.
- Identificar si algún tipo de beneficio presenta en mayor medida prácticas desleales.
- Evaluar e identificar el impacto de eliminar a los proveedores desleales de los planes de beneficios.

Tema analítico	Análisis requeridos o inferidos	Fuentes de datos
Cobertura de planes	**Análisis 2.a:** Dado un rango de años, mostrar el nivel de cobertura de los planes con respecto a las áreas de servicio **Análisis 2.b** ¿Se ha logrado una cobertura total de los proveedores, cubriendo con sus planes TODAS las áreas de servicio? **Análisis 2.c:** ¿Han existido áreas de servicios que sean cubiertas, a nivel de planes, por menos de dos proveedores?	F1. GruposAreasdeServicio y areas de servicio, F3. BeneficiosPlanes

- Para este análisis se tienen en cuenta las variables "IdAreaDeServicio_T", "IdGeografia_T ", "Condado", "IdPlan_T", "IdProveedor_T" y "Fecha" para el mismo plan.
- Variables opcionales: "valorCopago" y "valorCoseguro" para el mismo plan.
- Para este análisis se requiere que "IdAreaDeServicio_T" y "Fecha" tengan completitud y consistencia en los datos, por lo que hay que limpiarlos y/o imputarlos para poder hacer un análisis de la cobertura de los planes versus las áreas de servicio.
- Se puede realizar el análisis de áreas de servicio con menos de 2 proveedores; sin embargo, se debe tener completitud en los datos, por lo que hay que limpiarlos y/o imputarlos.

Recomendaciones:

- Examinar el comportamiento de la cobertura de planes es relevante; sin embargo, se recomienda que antes de iniciar el análisis, se tenga calidad de datos para las variables "IdAreaDeServicio_T" y "Fecha".
- Analizar cuántos proveedores cubren cada una de las áreas de servicio para identificar si hay zonas con mucha densidad y otras con menos, con el fin de optimizar las áreas por planes.
- Se propone profundizar en las áreas de servicio con baja cobertura versus la cantidad de proveedores que se tienen para validar si se recomienda buscar nuevos proveedores para esas zonas o si se sugiere evaluarlas.
- Se propone generar un modelo que muestre las zonas que son consideradas en riesgo por no tener la cantidad de proveedores requeridos, lo que pone en riesgo la cobertura de beneficios.

Tema analítico	Análisis requeridos o inferidos	Fuentes de datos
Concentraciones de planes	**Análisis 3.a:** Dado un rango de años, identificar ¿cuántos y cuales planes hay por áreas de servicio? **Análisis 3.b:** ¿en qué áreas hay concentraciones de planes que no correspondan con la cantidad de habitantes del área?	F1. GruposAreasdeServicio y areas de servicio, F3. Beneficios planes

- Para este análisis se tienen en cuenta las variables "IdAreaDeServicio_T", "IdGeografia_T ", "Condado", "IdPlan_T", "IdProveedor_T", "Fecha" y "habitantes por área" para el mismo plan. Variables opcionales: "IdProveedor_T" para el mismo beneficio.
- Para este análisis se requiere que "IdAreaDeServicio_T" y "Fecha" tengan calidad en los datos; por lo que hay que limpiarlos y/o imputarlos para poder validar por año cuántos y cuáles planes hay por área de servicio.
- Para este análisis se debe crear una dimensión asociada a la Fecha, de igual manera se deben considerar las transformaciones para los valores inconsistentes como referencias a años como 1920.
- Para poder saber cuántos y cuáles planes hay por áreas de servicio, se necesita una calidad del 100% de los datos de área, ya que el análisis depende de este campo; de lo contrario, no se podría determinar cuántos planes hay por área.
- Para este análisis se requiere, además, poder tener información de datos de población; por lo que será importante revisar fuentes confiables de censos de población para poder realizar esta pregunta.
- La densidad reportada para las áreas de servicio no corresponde a los datos de área y población. Esta debería ser una medida calculada una vez asegurada la confiabilidad de los datos de área y población.
- Se propone adicionar al análisis 3.b la concentración por proveedor en las áreas de servicio.
- Se requiere tener definido cuántos planes hay por área para avanzar con el análisis del punto 3.b.

- Se puede trabajar con los datos de áreas y de "IdPlan_T", teniendo como requisito previo la calidad de la información.

Tema analítico	Análisis requeridos o inferidos	Fuentes de datos
Evolución de planes	**Análisis 4.a:** Dado un rango de fechas y tipos de beneficios ¿Cómo han evolucionado los costos y tipos de beneficios a lo largo del tiempo por tipo de beneficio, proveedor, fecha? **Análisis 4.b:** ¿Qué tipos de beneficios han aumentado o disminuido costos?	F2. Beneficios (Tipos de beneficio y condiciones), F3. Beneficios planes

- Para este análisis se requiere trabajar con los campos "IdTipoBeneficio_T", "IdCondicionDePagoCopago_T", "IdCondicionDePagoCoseguro_T", "IdProveedor_T", "Fecha", "valorCopago" y "valorCoseguro".
- Para el análisis de aumento o disminución de costos de los beneficios es necesario tener en cuenta condiciones de pago sobre los diferentes beneficios.
- Para este análisis se debe crear una dimensión asociada a la Fecha, de igual manera se deben considerar las transformaciones para los valores inconsistentes como referencias a años como 1920.

Se concluye que es posible realizar los análisis sugeridos una vez se aseguren las dimensiones de calidad de las tablas y un correcto diseño de ETL para el modelo de datos. No se evidencia un riesgo significativo de falta de datos de manera que posterior a la limpieza se imposibilite el análisis para responder las preguntas de negocio.