En este informe, se presenta un análisis detallado del modelo de regresión lineal estimado para predecir el número de nacimientos en el año 2022, que depende de las variables: prestadores de nivel 2 (nivel_2) y población. El objetivo es construir un modelo de regresión lineal a partir de tablas importadas en SQL evaluando la validez y fiabilidad del mismo.

Tratamiento de datos:

El proceso de carga y tratamiento de datos. Inicialmente, se cargan datos desde archivos, incluyendo "DB_address_IE_escobar.db" con bases de datos de municipios y prestadores en formato SQL, junto con la tabla de nacimientos del 2022 del DANE. Se seleccionan datos específicos de la tabla de nacimientos y se convierten en data frames. También se importan datos DIVIPOLA desde www.datos.gov.co. Luego, se verifica la coherencia de variables cuantitativas y cualitativas, identificando y resolviendo caracteres especiales en la tabla de municipios mediante expresiones regulares. Se plantea un flujo de trabajo para abordar desafíos en la base de datos de prestadores, especialmente en la estandarización de nombres de municipios. Se realiza un cruce entre las tablas de municipios y prestadores, identificando una llave para su desarrollo. Variables cualitativas relevantes se convierten en cuantitativas. Finalmente, se importa la base de datos de nacimientos del DANE, consolidando un data frame con 1096 filas y 31 columnas. Este proceso garantiza la integridad y coherencia de los datos para su análisis posterior.

Análisis descriptivos:

- **Análisis exploratorio de datos:** se realizó un diagnóstico del estado de las variables iniciales, valores perdidos, su distribución y la visualización de la matriz de correlación.

- **Medidas de Tendencia Central:**

- La media de los prestadores de nivel 2 es de aproximadamente 0.104.
- La mediana de los prestadores de nivel 2 es 0.
- La moda de los prestadores de nivel 1 es "1".

- **Medidas de Dispersión:**

- La desviación estándar de los prestadores de nivel 1 es aproximadamente 0.921.
- El rango intercuartílico de los prestadores de nivel 1 va de 0 a 14.

Se generó un histograma y una tabla de frecuencias para los datos de nivel 1, lo que permite visualizar la distribución de los valores.

Modelo de Regresión:

A partir de la base de datos con 1096 filas y 31 columnas, se procedió a construir el modelo de regresión lineal, en el cual la variable dependiente son los nacimientos del año 2022. El proceso de construcción del modelo sigue la metodología Stepwise Backward la cual consiste en construir el modelo con todas las variables independientes de interés e ir descartando paso a paso variables que no son explicativas o aquellas que tienen valor-p mayor al 5% hasta lograr la parsimonia (el menor número de variables con el R^2 ajustado más alto). En ese sentido, luego de nueve (9) pasos, se logró llegar al siguiente modelo:

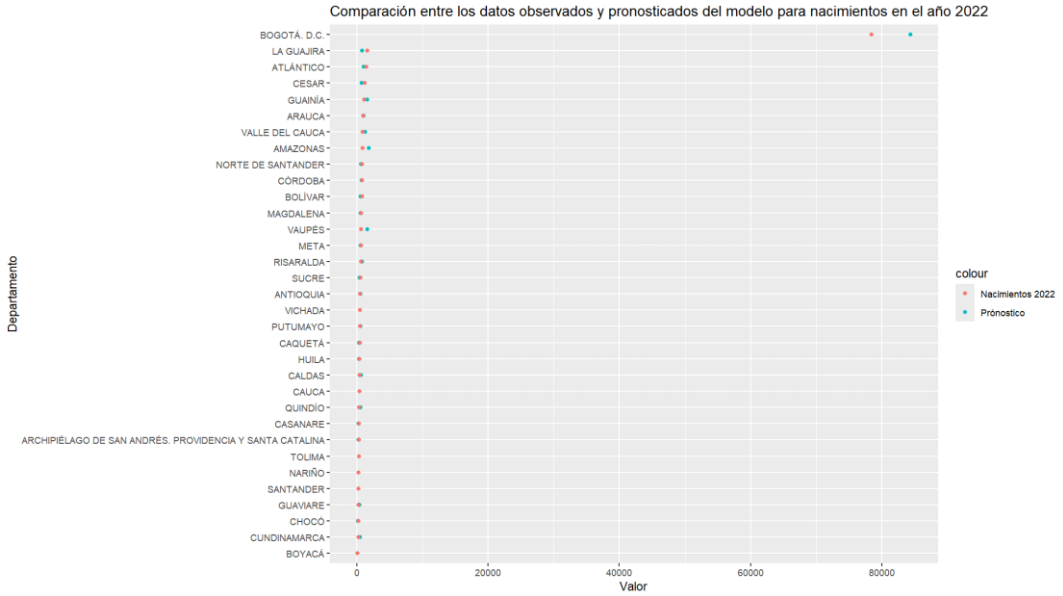
$$\hat{nacimientos\_2022} = -122.46757 + (1400.09942 * nivel\_2) + (0.01112 * Poblacion)$$ donde,

nacimientos_2022: Nacimientos ocurridos durante el año 2022.
nivel_2: Prestadores de servicios de salud de nivel 2.
Poblacion: Número de habitantes de los municipios.

El modelo presenta un R^2 ajustado de 0.9075, lo que indica que aproximadamente el 90.75% de la variabilidad en la variable dependiente puede ser explicada por las variables independientes.

Análisis de Supuestos:

- **Linealidad:** Se realizó una prueba de hipótesis para verificar la linealidad del modelo y estimó la media de los residuos cuyo resultado es cero concluyendo que hay linealidad en el modelo.
- **Normalidad:** Se evaluó la normalidad de los residuos mediante un gráfico Q-Q, concluyendo que los residuos siguen una distribución normal.
- **Homocedasticidad:** Se confirmó la homocedasticidad del modelo, indicando que hay variabilidad de los residuos.
- **Independencia:** Se confirmó la independencia de los residuos del modelo que no hay correlación entre ellos.
- **Multicolinealidad:** Se verificó la multicolinealidad entre las variables independientes mediante la prueba de Factor de Inflación de la Varianza (VIF) cuyo valor es inferior a cinco (5) para cada una de las variables independientes.



Conclusiones:

- Dentro de los diversos procesos realizados para alcanzar este producto, el más largo en términos de tiempos fue el tratamiento de datos toda vez que la base de datos de prestadores no contaba con una llave estandarizada que permitiera cruzar de manera óptima la base de datos de municipios y demás que se pudieran requerir.
- Según el análisis realizado, el modelo de regresión propuesto cumple con los supuestos necesarios para realizar inferencias válidas.
- Estos resultados respaldan la validez y fiabilidad del modelo de regresión propuesto para predecir el número de nacimientos en función de las variables nivel_2 y población en el año 2022.