**PRUEBA ANALISTA DE DATOS**

**Contexto**

On Daxelta es una empresa de gas vehicular que premia a sus usuarios por consumos en sus estaciones de servicio, el sistema lleva algunos años funcionando, pero para la construcción de la estrategia del siguiente año se debe analizar la data histórica de la marca.

Su tarea es gestionar los datos presentes en las bases Customers, Estaciones, Geo, Trans\_sem, Trans\_2\_Sem y generar hallazgos de valor sobre el comportamiento del negocio, para ello debe realizar una presentación y dashboards que apoyen el análisis de la data.

**Dentro del desarrollo del caso debe, como mínimo:**

1. Analizar la volumetría del galonaje por día de la semana.
2. Realizar el pronóstico del valor venta para los siguientes días martes y miércoles.
3. Analizar el comportamiento por Regional
4. Analizar el comportamiento por Estación de servicio
5. Haga una segmentación para conocer a nuestros clientes y genere gráficas que permitan entender cómo se caracteriza cada segmento.
6. Identifique los clientes de mayor valor para la marca.

Documente el proceso realizado y las herramientas utilizadas en la construcción del caso.

Genere recomendaciones sobre la estandarización de la información y reporte de inconsistencias en caso de encontrar problemas con la data entregada para el análisis.

**DOCUMENTACIÓN**

-Se reconoce que las bases “Trans\_Sem”, “Trans\_2\_Sem” y “Estaciones” no tienen un tipo de archivo útil. Por tanto, se guardan como archivo de texto, que es un formato compatible con el software que se va a utilizar (Power BI).

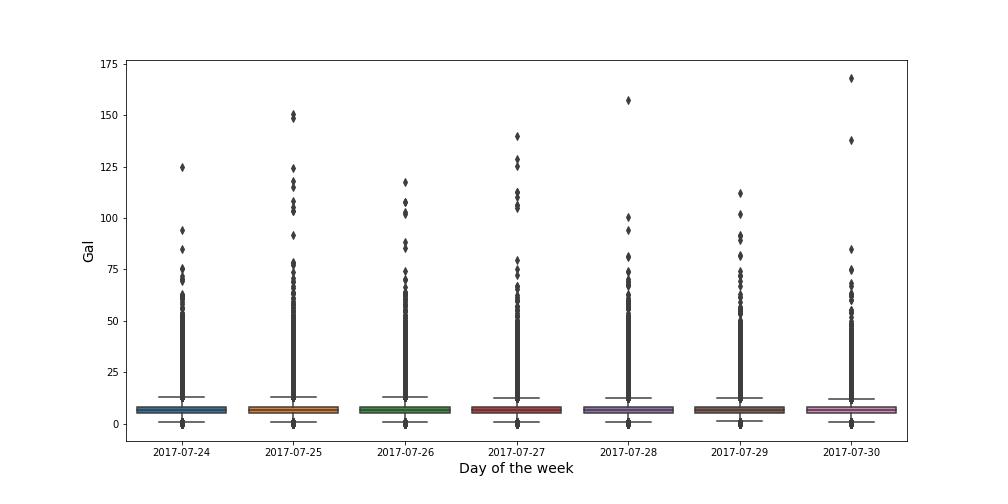
-Se reconoce que en la base “Trans\_2\_Sem” los nombres de las columnas están desfasados por uno y se procede a corregir los nombres manualmente.

-Se unen las tablas “Trans\_Sem” y “Trans\_2\_Sem” en una sola tabla “Combine”.

-Se limpian los espacios en blanco del principio y final, se eliminan las filas duplicadas, se verifica que las etiquetas de tipos de datos correspondan con las esperadas.

-Se verifica que no hay datos faltantes ni fuera del formato (recortar la imagen siguiente).

-Luego de hacer la limpieza de los datos, se procede con análisis exploratorio de los datos. Se toman las estadísticas de resumen y se hacen diagramas de caja y bigotes. Es notable que hay varios outliers, pero que parecen formar un patrón y no son poco comunes. Por lo tanto, se decide conservarlos. Se sugiere hacer una evaluación de la normalidad de los datos cuando haya tiempo.



- Se hacen diagramas de barras para saber el volumen de gas total dispensado por día de la semana. Se obtiene que los días en que más volumen se dispensa son viernes y sábados, y el que menos el domingo.

- Desde luego, esto se corresponde uno a uno con las ventas. El orden de magnitud del total vendido por día de la semana está en los miles de millones y, analizando el patrón de ventas de la semana, se determina que el pronóstico de ventas de los martes y miércoles es aproximadamente 117 mil millones.

-En cuanto a los datos de regionales y localización, se siguen los mismos pasos de limpieza que para los datos anteriores. Se encuentran 5 valores faltantes, pero no serán importantes, de modo que no se eliminan. Adicionalmente, mediante inspección visual, en la base “Geo” se encuentra que hay 5 observaciones que no se conforman al formato. Se procede a eliminar dichas filas. Se recomienda usar la herramienta de formato condicional cuando haya tiempo.

-Se hace una consulta en SQL en donde se unen las tablas “Trans\_Sem”, “Trans\_2\_Sem”, “Geo” y “Estaciones” para saber cuántas ventas produjo cada regional en la semana analizada. Boyacá fue la regional que menos produjo, mientras que el Valle fue la que más.

-Adicionalmente, se analiza cuántas ventas produjo cada uno de las 100 estaciones que más valor generaron con un diagrama de barras. Se encuentra que la estación que más vendió está en Cali, y vendió considerablemente más que la segunda, que está en Bogotá. Algunas preguntas para análisis en el futuro: de las 100 estaciones más lucrativas, ¿cuántas hay en cada departamento? ¿Cuánto del total de ventas representan estas 100 estaciones? ¿qué tipo de estación son estas? El dinero generado, ¿es por muchas ventas pequeñas o unas pocas grandes?

-En cuanto a la base “Customers”, se hace la misma limpieza que a los datos anteriores. Se encuentra que la variable “FechaNacimiento” tiene aproximadamente el 2.5% de sus valores faltantes. Se decide no utilizar estas observaciones en los análisis que involucren dicha variable, debido al pequeño número de estas y a que esta es una variable categórica. Sin embargo, un análisis de valores faltantes para determinar el tipo de faltantes (MCAR, MAR, MNAR) mejoraría el análisis.

-Se hace la segmentación de los clientes, de acuerdo a la variable “Segmento”, y se encuentra un error de formato. Se hace un preprocesamiento en excel, en donde se convierten todos los valores de esta variable a “plain text”, lo que solucionas el problema. Adicionalmente, se hace una segunda segmentación de acuerdo a si el cliente es un posible conductor (mayor de 16 años) o no.

-Se grafica de acuerdo a las dos segmentaciones, y se encuentra que la gran mayoría de los clientes consumen para sustento, pero el número de compradores no es tan grande como inicialmente parece. Bajo la asunción de que solo los conductores compran el producto, el número es cercano a los 100.00 cliente, como se puede ver en la gráfica.

-Se agrupan las bases que contienen los datos de ventas de acuerdo a la placa del auto que compró y se obtienen los autos que más compraron en la semana.

**RECOMENDACIONES**

- Establecer un solo formato para almacenar los datos (separados por comas, archivo xlms, separados por tab, etc.). Que este formato, además, sea legible por la mayoría de los programas.

- Recolectar más datos históricos. Hacer pronósticos con las observaciones de una sola semana no conlleva resultados acertados.

- Agregar una descripción de los campos. Lo que significan y el tipo de dato esperado.

**INCONSISTENCIAS**

- En Trans\_2\_Sem es claro que la primera columna debería ser “Placa” y la última “Fecha venta”.

- En la base Customers, campo “Segmento”, algunas observaciones están dadas como texto plano y otras como función.

-El campo “IdCliente” en Trans\_Sem y Trans\_2\_Sem varía para una misma placa. Esto lo hace irrelevante, ya que no aporta información adicional acerca del cliente o su status de afiliación.