Informe Solución Prueba Técnica – Data Scientist Altipal

11 FEBRERO

ALTIPAL

Creado por: Julian Andres Quimbayo Castro

Desarrollo de Solución

Punto 1: Indique y defina las funciones de agregación disponibles en SQL estándar

Las funciones estándar de agregación en SQL estándar son:

- Count: Valor de filas en total seleccionadas dentro de la consulta.
- Min: Trae el valor mínimo del campo que se especifique en la consulta.
- Max: Trae el valor máximo del campo que se especifique en la consulta.
- Sum: Trae la suma del campo que se especifique en la consulta. Allí solo se puede realizar con datos numéricos.
- Avg: Trae el valor promedio del campo que se especifique en la consulta. Allí solo se puede realizar con datos numéricos.

Punto 2: Defina los siguientes conceptos.

- ETL: Extraer, Transformar y Cargar, es una canalización de datos que se utiliza para recopilar datos de varias fuentes. A continuación, transforma los datos de acuerdo con las reglas de negocio y los carga en un almacén de datos de destino. El trabajo de transformación en ETL tiene lugar en un motor especializado, y a menudo implica el uso de tablas de preparación para mantener temporalmente los datos mientras se transforman y se cargan finalmente en su destino.[1]
- Data warehouse: es un sistema que agrega y combina información de diferentes fuentes en un almacén de datos único y centralizado; consistente para respaldar el análisis empresarial, la minería de datos, inteligencia artificial (IA) y Machine Learning. Data warehouse permite a una organización o empresa ejecutar análisis potentes en grandes volúmenes (petabytes y petabytes) de datos históricos de formas que una base de datos estándar simplemente no puede.[2]
- Data lake: es un repositorio de almacenamiento que contienen una gran cantidad de datos en bruto y que se mantienen allí hasta que sea necesario. A diferencia de un data warehouse jerárquico que almacena datos en ficheros o carpetas, un data lake utiliza una arquitectura plana para almacenar los datos.

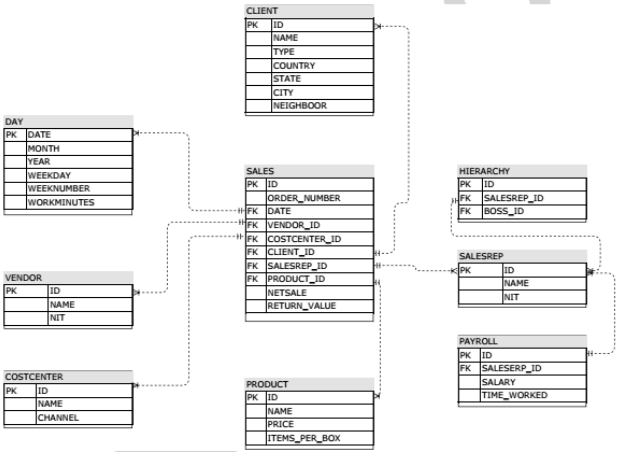
  A cada elemento de un data lake se le asigna un identificador único y se etiqueta con un conjunto de etiquetas de metadatos extendidas. Cuando se presenta una cuestión de negocios que debe ser resuelta, podemos solicitarle al data lake los datos que estén relacionados con esa cuestión. Una vez obtenidos podemos

[1] **Tomado de:** Extract, transform, and load (ETL) - Azure Architecture Center | Microsoft Docs
[2] **Tomado de:** ¿Qué es un Data Warehouse? - México | IBM

analizar ese conjunto de datos más pequeño para ayudar a obtener una respuesta.[3]

- Data mart: es una estructura de datos, construido dentro de un repositorio o base de datos. En esta estructura, se almacena información agregada o consolidada la cual será consumida por alguna herramienta de visualización o data analytics como Tableau. Usualmente un Data Mart se especializa; es decir, solo almacena información de un área de la empresa o de un flujo o proceso específico. Un conjunto de Data Marts vinculados entre sí, se denomina Data Warehouse.[4]

Punto 3: Responda las siguientes preguntas usando el siguiente modelo de datos.



Construya una consulta que contenga los campos:

a. ORDER_NUMBER
b. DATE
c. VENDOR_NAME
d. CLIENTE_NAME
e. CLIENT_TYPE
f. CLIENT_STATE
g. CLIENT_CITY
h. SALESREP_ID
i. PRODUCT_ID
j. SALESREP_BOSS_ID
k. NETSALE
l. Unidades vendidas
m.Número de cajas/BOXES vendidas
n. Total de venta por ORDER_NUMBER

[3] **Tomado de:** Data lake: definición, conceptos clave y mejores prácticas (powerdata.es)
[4] **Tomado de:** Data Mart: ¿Qué es y cómo funciona? | Tableau Perú (tableauperu.com)

Consulta:
Select s.order_number, d.date, v.name as vendor_name, c.name as cliente_name, c.type as cliente_name, c.state as client_state, c.city as client_city, sa.salesrep_id, p.id as product_id, h.boss_id as salesrep_boss_id, s.netsale, count(p.id) as unidades_vendidas, sum(p.items_per_box) as total_boxes, sum(s.order_number) as totalventas from sales s inner join day d on s.date=d.date inner join vendor v on s.vendor_id=v.id inner join client c on s.client_id=c.id inner join salesrep sa on s.salesrep_id=sa.id inner join product p on s.product_id=p.id order by unidades_vendidas desc.

Punto 4. Realice una consulta para mostrar:

- Nombre y ID del vendedor concatenado
- Antigüedad
- ID del jefe
- Salario
- Ventas totales (neto – devoluciones) durante los últimos 12 meses
- Número de clientes atendidos con ventas

Asumiendo que existiera un campo denominado año de ingreso a la compañía.

Select concat (v.name, ‘ ’,v.id) as nombreid_vendedor, TIMESTAMPDIFF(YEAR,d.year, CURDATE()) as antiguedad, h.boss_id, p.salary, sum(s.netsale- s.returnvalue) as total, count(c.id) as clientesventas from vendor v inner join sales s on v.id=s.vendor_id inner join day d on d.date = s.date inner join salesrep sa on sa.id = s.salesrep_id inner join hierachy h on h.salesrep_id = sa.id inner join product p on p.id=s.product_id group by clientesventas order by total desc

Punto 5: Indique para que se utilizan los siguientes gráficos y qué información se puede presentar en ellos:

Diagrama de Caja: Sirve para determinar en una distribución de datos cual es el valor minimo, máximo, mediana, cuartiles y outliers de nuestra información, en esta figura de debe de leer de izquiera a derecha siendo al linea inicial vertical el punto mínimo.

A.

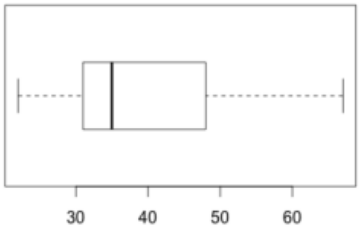


Diagrama de Barras (Stack), permite conocer la distribución de la información entre variables categóricas y numéricas, utilizando las variables categóricas como leyenda.

B.

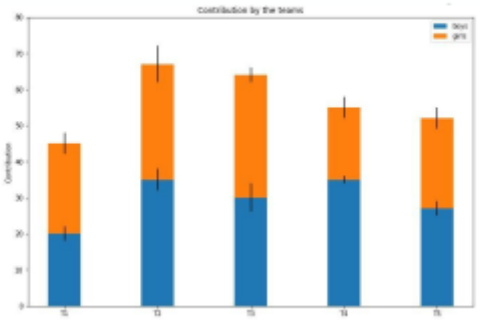


Diagrama Pie : muestra la proporción de información para cada una de la variables categóricas seleccionadas, esto permite saber con respecto a la totalidad cuantos datos están para cada etiqueta de una observación.

C.

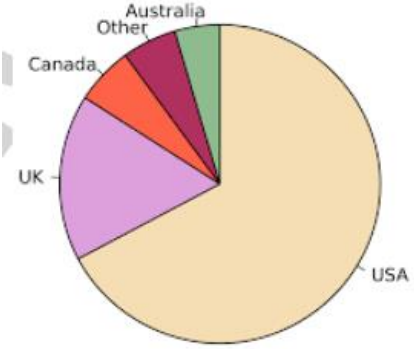


Diagrama histograma: permite conocer la distribución normal de los datos y asi determinar las tendencias dentro de una o varias observaciones y asi saber si los datos estan ajustados a la campana de gauss o no, esto ayuda a dilucidar mas adelante como podemos enfocar la validación de hipótesis tanto nula y/o alternativa al identificar los niveles de confianza.

D.

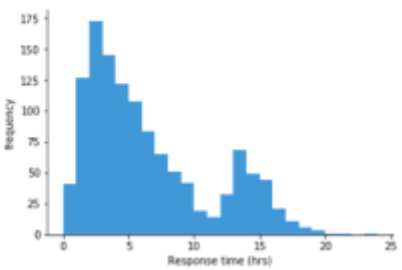


Grafico de areas apiladas: sirve para en un forecasting conocer el comportamiento en un espacio de tiempo de las variables a analizar.

E.

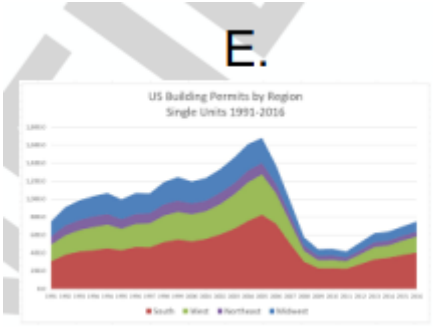
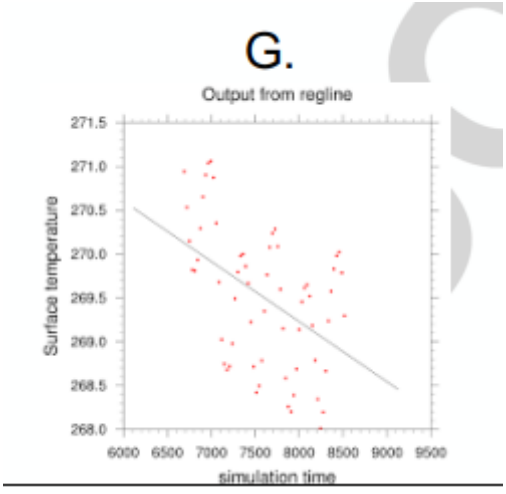


Gráfico de Líneas: sirve de igual manera para conocer el comportamiento de las variables dentro de un lapso de tiempo o fecha.

F.

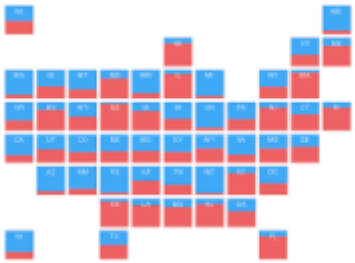
Diagrama de dispersión con modelo de regresión: permite conocer como al cruzar las variables estas se ajustan o no al modelo dibujado en una línea perpendicular. Vemos que la línea esta invertida por tal motivo sugiere una correlación negativa, quiere decir que las variable independiente no es capaz de definir la dependiente por tal motivo no puedo realizar un proceso predictivo correcto.

G.



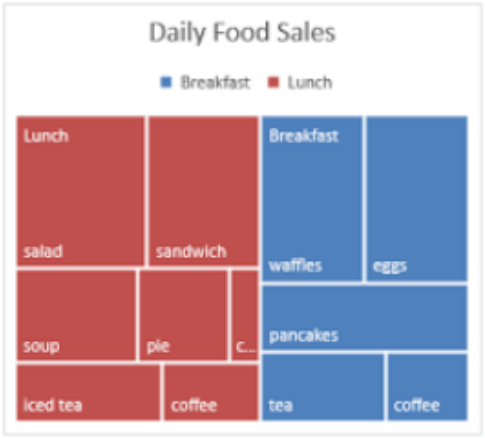
Mapa de Calor: permite conocer por porcentajes y colores los valores mínimos y máximos de una distribucion o sus proporciones.

H.



Treemap: Permite conocer cual es la proporcion de datos por cada una de las variables categóricas en la leyenda y asi mismo realizar comparativas.

I.



Punto 6: Desarrolle un modelo para la competencia de Kaggle Predict Future Sales(https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data). Reporte de forma autocontenida su exploración de los datos, el modelo seleccionado justificando su selección y los resultados obtenidos usando lasmétricas de la competencia. Determine si su modelo tiene un ajuste adecuado, si la calidad de los datos es suficiente y si variables adicionales pueden ayudar a mejorar su predicción.

Resultados en Google Colab:
https://colab.research.google.com/drive/19g1Yse_h0EcgELAO1mCfXluWbLKLTy7B?usp=sharing