Descripción del Proyecto Final – Bootcamp: Data Analyst Jr.

Indicaciones para preparar la base de datos

La base de datos llamada “custshop” se utilizará como insumo para todo el proyecto. Para prepararla en su servidor local de PostgreSQL, siga estas indicaciones:

1. Ingrese a PostgreSQL utilizando la aplicación pgAdmin, tal como lo hacíamos para trabajar durante el módulo de SQL.
2. En el Explorador de Objetos, que está en la parte izquierda de la ventana de la aplicación, haga click derecho sobre el Server que se creó por defecto (probablemente con nombre “PostgreSQL 15”), y use la opción “Create” y luego “Database”. Se abrirá el cuadro de diálogo “Create – Database”, ahí coloque el nombre de la nueva base de datos “custshop” en el campo “Database”. No es necesario modificar ninguna otra opción. Haga click en el botón “Save”. Con esto quedará creada la base de datos completamente en blanco.
3. Descargue el archivo “custshop.tar” en su computadora. Se encuentra disponible dentro del folder de recursos de Google Drive, dentro del folder “Proyecto”. Puede encontrar el enlace directo al archivo en el Aula Virtual, en la sección “General” del bootcamp.
4. Haga click derecho sobre el nombre de la nueva base de datos y use la opción “Restore…”. En el cuadro de diálogo de la opción Restore, en el campo “Format” use la opción “Custom or tar”, luego en el campo “Filename” use el ícono de folder para ubicar el archivo custshop.tar entre sus archivos locales. La ruta completa del archivo quedará indicada en ese campo.
5. En la pestaña “Data/Objects” active las opciones “Pre-data”, “Data” y “Post-data”. No es necesario cambiar ninguna de las demás opciones. Finalmente, haga click en el botón “Restore”.
6. Una vez que el proceso haya terminado, haga un refresh en el Explorador de Objetos: click derecho sobre el nombre de la base de datos “custshop” y opción “Refresh…”. Luego verifique que se crearon las cinco tablas de la base de datos. Puede ubicarlas dentro de la estructura jerárquica: Schemas -> public -> Tables. Puede abrir una ventana de Query Tool y escribir queries para verificar que las tablas tienen datos. Por ejemplo, la tabla principal se llama customer_shopping, y debe tener 99,457 filas.

La base de datos se encuentra ahora lista para iniciar el trabajo.

Acerca de la base de datos custshop:

La base de datos custshop (Customer Shopping) contiene datos de las transacciones de compra realizadas en las tiendas de un conjunto seleccionado de los principales centros comerciales de la ciudad de Estambul en Turquía.

Puede encontrar una descripción general de la base de datos, y descripciones de algunas de las columnas en la página de Kaggle que corresponde a este dataset:

https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Tome en cuenta que la base de datos custshop está basada en el dataset de Kaggle, pero no es exactamente igual: se han hecho algunos cambios: se han modificado los datos de edad, se ha agregado columna de distrito, y además se ha organizado el dataset plano en varias tablas, con relaciones de llave primaria y llave foránea.

Parte I: Exploración de los datos en la base de datos de PostgreSQL utilizando el lenguaje SQL

Identifique las relaciones entre las tablas y, para facilitar su trabajo, cree un diagrama entidad-relación de la base de datos.

Escriba consultas (queries) en SQL que den respuesta a las siguientes preguntas:

1. Identifique la tabla principal, es decir, la que tiene la información de las transacciones. ¿Cuántas filas tiene esta tabla?
2. Suponga que las ventas se pueden calcular como el producto de la cantidad de items por el precio. ¿Cuál es el gran total de ventas en toda la tabla transaccional?
3. ¿Cuál es el rango de fechas de las transacciones, es decir la menor fecha y la mayor fecha?
4. Muestre la fecha, el número de factura, código de cliente y centro comercial de todas las transacciones del año 2021 o 2022 que pertenecen a centros comerciales en el distrito de Levent o de Besiktas, realizadas por clientes del género femenino, de entre 18 y 25 años, de productos en las categorías de Clothing, Shoes o Cosmetics, que pagaron con tarjeta de crédito o de débito y en las que el monto total de la factura fue de 10 mil liras turcas o más.
5. Muestre el total de unidades vendidas (cantidad), el conteo de transacciones (factura) y el total de ventas, agrupados por año y mes, en orden cronológico
6. Muestre el total de ventas por cada centro comercial.
7. Muestre el total de ventas, desglosado por género y por edad del cliente.
8. Muestre el precio promedio por item que corresponde a cada año-mes. El número de items facturados está indicado por la columna quantity.
9. Muestre el monto promedio por factura que corresponde a cada año-mes.
10. Muestre las ventas desglosadas por fecha, categoría de producto, centro comercial, distrito, forma de pago, edad y género del cliente.

Parte II: Preparación de datos en archivo plano para exportarlos a otras herramientas

Escriba un query para extraer los datos de las transacciones que están en la base de datos custshop, de manera que incluya las siguientes columnas (en este orden específico):

- invoice_no
- invoice_date
- shopping_mall,
- district,
- customer_id,
- gender
- age
- category
- payment_method
- quantity
- price

Exporte el resultado del query a un archivo en formato csv, con este nombre: “customer_shopping.csv”

Parte III: Creación de un modelo de datos con Excel Power Query

Carga de los datos de la tabla principal del modelo (fact table)

Utilice el archivo customer_shopping.csv como origen de datos para crear una tabla (objeto query) en el modelo de datos de Power Pivot.

Asegúrese de que las columnas de la tabla en el modelo de datos tienen tipos de datos adecuados.

Cambie el nombre del objeto query a “Customer Shopping” (sin guión bajo y con iniciales mayúsculas). Recuerde hacer un refresh de datos después del cambio de nombre, para que se aplique correctamente en el modelo de datos.

Cambie el nombre de todas las columnas de esa tabla (objeto query) para adaptarlos al mismo estándar: con espacio en lugar de guión bajo entre palabras, y con iniciales mayúsculas.

Tabla de Dimensión de Calendario

Cree una tabla de Calendario en un archivo separado de Excel, que incluya al menos las siguientes columnas:

- Date
- Year
- Month Num (1 al 12)
- Day (1 al 31)
- Month (Nombre completo del mes, en inglés o en español)
- Mon (Nombre abreviado del mes, en inglés o en español)
- Mon Year (Nombre abreviado del mes, seguido del año, en el formato “Jan-2015” o “Ene-2015”, si es en español)
- Mon Yr (Nombre abreviado del mes, seguido del año en formato de dos dígitos, por ejemplo: “Jan-15” o “Ene-15”)
- Day of Week Num (Número de día de la semana, del 1 al 7, siendo 1 el lunes)
- DOW (Abreviatura del día de la semana, en inglés o en español)
- Week of Year (Número correlativo de la semana del año, vea la función de Excel llamada WEEKNUM, en inglés; o NUM.DE.SEMANA, en español)
- Week Starting On (Fecha en que inicia la semana, iniciando los lunes)
- Quarter (Trimestre calendario: “Q1”, “Q2”, “Q3” y “Q4”)

Puede usar como apoyo o referencia el archivo Calendar Example.xlsx, que se encuentra en el folder de recursos del Proyecto en Google Drive.

Verifique que esta tabla de calendario incluye todo el intervalo de fechas que tienen los datos de la tabla principal (fact table).

Revise que todas las columnas de la tabla calendario se cargan al Modelo de Datos con el formato adecuado, tal como aparecen en el archivo de datos fuente de la tabla.

Cargue la tabla de calendario al Modelo de Datos de Power Pivot, cree la relación con la respectiva columna de fecha en la tabla fact.

IMPORTANTE: Agregue y configure las columnas que sean necesarias para que en los reportes aparezcan todos los atributos de fecha en orden cronológico.

Columna Calculada para el monto de venta (Sales)

Modifique el objeto query de la tabla principal para agregar la columna Sales, que se calcularía como el producto de la cantidad y el precio unitario. Asegúrese de que la nueva columna tenga el tipo de dato adecuado.

Creación de las medidas del modelo

Ahora crearemos las medidas (measures) base del modelo, utilizando simplemente la función de agregación de sumatoria sobre las columnas correspondientes de la tabla fact. Antes de crear las medidas, modifique el nombre de las columnas fact para poder utilizar esos nombres para los objetos medida y evitar un error por nombre duplicado. Una posibilidad para eso es agregar un guión bajo al final del nombre de la columna fact, por ejemplo, la columna “Quantity” pasaría a llamarse “Quantity_”, y la columna “Sales” pasaría a ser “Sales_”.

Ahora cree, dentro de la tabla fact, las dos medidas base: **Quantity** y **Sales**. Configure el formato de la medida Sales para que se despliegue un número fraccionario decimal (Decimal Number) pero que no muestre los centavos, y que además use la coma como separador de miles. El formato de la medida Quantity deberá ser de Número Entero, y también usando la coma como separador de miles.

Note que no vamos a crear una medida para la columna “Price” porque esa columna **es un fact no aditivo**, puesto que implícitamente es el resultado de una división (ventas dividido entre cantidad de items) por lo tanto no tiene sentido hacerle una operación de agregación, como una simple sumatoria. Cualquier reporte que trate con el concepto de Precio Promedio debe calcularse de otra manera, y no con base en la columna “Precio”.

Por último cree una medida adicional que calculo el precio promedio por item. Esta medida tendrá el nombre **Average Price**, y deberá definirse utilizando la fórmula apropiada con base en las otras medidas ya creadas.

Parte IV: Creación de Reportes Analíticos con Excel Power Pivot

1. En la primera hoja del archivo/libro de Excel donde está el Modelo de Datos, cree los gráficos dinámicos que se describen a continuación, utilizando el propio Modelo de Datos como fuente. Aplique los principios de diseño de gráficos que estudiamos en las sesiones, para lograr que estos gráficos comuniquen de forma efectiva la información contenida en los datos.

    a) Gráfico de barras verticales agrupadas (clustered columns) que muestre las ventas totales mensuales de todos los centro comerciales juntos. El gráfico debe mostrar los meses en el eje horizontal y los años como series del gráfico.
    b) Gráfico de líneas que muestre el precio promedio (Average Price) de los items que pertenecen a las categorías “Technology”, “Shoes” y “Clothing”. Estas categorías deben ser series del gráfico, es decir, una línea por cada categoría. El gráfico mostrará los meses en el eje horizontal, y deberá estar creado para filtrar un año específico, y un centro comercial específico. Puede implementar estos filtros como filtros del gráfico o como segmentadores (slicers) junto al gráfico.
    c) Gráfico que muestre, para un año, mes y centro comercial seleccionado, el total de ventas por día, desglosado por método de pago. El gráfico debe permitir comparar visualmente las ventas diarias por cada método de pago, así como las ventas diarias totales de un día con otro. Elija el tipo de gráfico que más efectivamente atienda a este requerimiento.

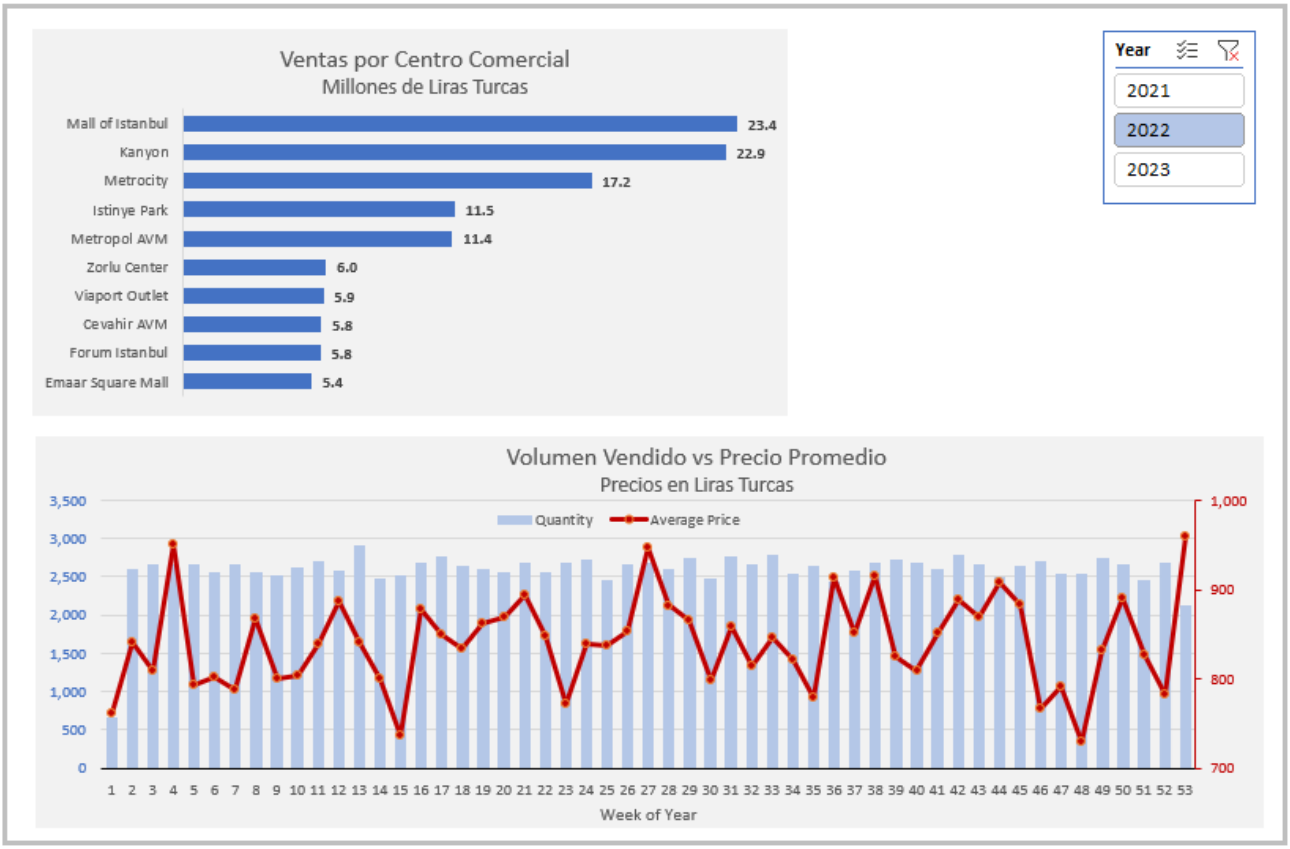
2. En la segunda hoja del archivo, cree una tabla dinámica como la que se muestra en la siguiente imagen:

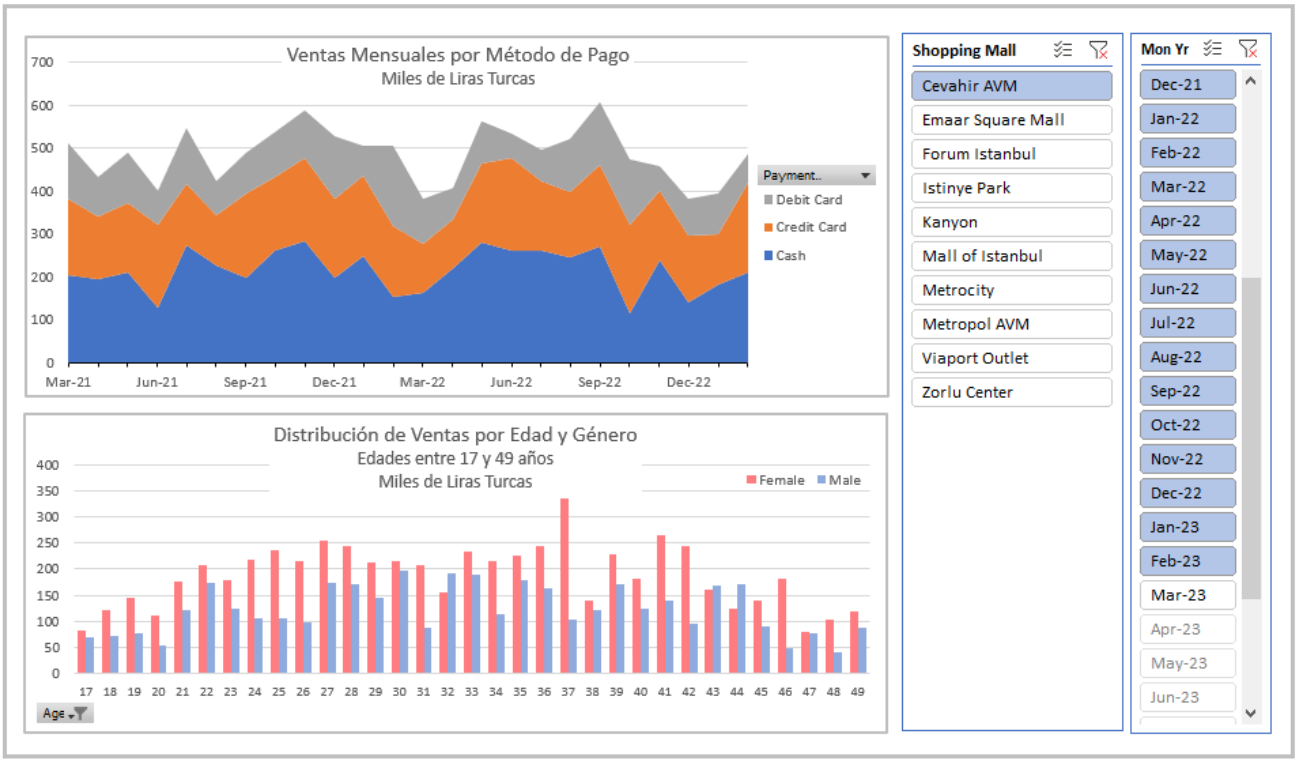
Shopping Mall	Kanyon				
Year	2023				
Sales		**Mon**			
Category	**Payment Method**	**Jan**	**Feb**	**Mar**	**Grand Total**
Clothing	Cash	340,591	314,184	89,424	744,198
	Credit Card	281,475	330,388	91,824	703,688
	Debit Card	197,753	220,259	53,114	471,126
Clothing Total		**819,819**	**864,831**	**234,362**	**1,919,012**
Technology	Cash	349,650	306,600	25,200	681,450
	Credit Card	163,800	127,050	16,800	307,650
	Debit Card	122,850	151,200	61,950	336,000
Technology Total		**636,300**	**584,850**	**103,950**	**1,325,100**
Shoes	Cash	166,847	185,453	114,032	466,332
	Credit Card	154,844	150,643	39,611	345,098
	Debit Card	97,828	78,022	2,401	178,250
Shoes Total		**419,519**	**414,117**	**156,044**	**989,680**
Cosmetics	Cash	21,428	24,071	7,847	53,346
	Credit Card	23,420	24,925	7,116	55,460
	Debit Card	11,141	8,417	1,748	21,306
Cosmetics Total		**55,989**	**57,412**	**16,711**	**130,112**
Toys	Cash	16,486	9,068	2,258	27,812
	Credit Card	11,505	5,770	1,362	18,637
	Debit Card	6,953	5,268	1,362	13,583
Toys Total		**34,944**	**20,106**	**4,982**	**60,032**
Books	Cash	4,075	2,879	515	7,469
	Credit Card	1,894	1,863	758	4,515
	Debit Card	2,318	1,470		3,788
Books Total		**8,287**	**6,212**	**1,273**	**15,771**
Food & Beverage	Cash	2,944	3,175	711	6,830
	Credit Card	2,186	2,034	622	4,843
	Debit Card	1,611	445	549	2,605
Food & Beverage Total		**6,741**	**5,654**	**1,883**	**14,278**
Souvenir	Cash	2,111	1,138	70	3,320
	Credit Card	1,654	2,182	493	4,328
	Debit Card	1,032	1,044	106	2,182
Souvenir Total		**4,798**	**4,364**	**669**	**9,830**
Grand Total		**1,986,396**	**1,957,545**	**519,874**	**4,463,815**

3. En la tercera hoja del archivo, cree los gráficos que se muestran en las siguientes imágenes.

   Procure que sus gráficos se asemejen tanto como sea posible a los que aparecen en la imagen, cuidando detalles como los atributos que se muestran en el reporte, las proporciones de las barras, los marcadores en las líneas, la escala en los ejes, las etiquetas de valores, el título de los gráficos y la presencia o ausencia de otros elementos visuales.

   Los objetos deben ser creados de tal manera que los segmentadores afecten a los gráficos que se encuentran dentro de su propio recuadro.


Ventas por Centro Comercial
Millones de Liras Turcas
Mall of Istanbul 23.4
Kanyon 22.9
Metrocity 17.2
Istinye Park 11.5
Metropol AVM 11.4
Zorlu Center 6.0
Viaport Outlet 5.9
Cevahir AVM 5.8
Forum Istanbul 5.8
Emaar Square Mall 5.4
Year
2021
2022
2023
Volumen Vendido vs Precio Promedio
Precios en Liras Turcas
Quantity
Average Price
Week of Year


Ventas Mensuales por Método de Pago
Miles de Liras Turcas
Payment..
Debit Card
Credit Card
Cash
Distribución de Ventas por Edad y Género
Edades entre 17 y 49 años
Miles de Liras Turcas
Female
Male
Age
Shopping Mall
Cevahir AVM
Emaar Square Mall
Forum Istanbul
Istinye Park
Kanyon
Mall of Istanbul
Metrocity
Metropol AVM
Viaport Outlet
Zorlu Center
Mon Yr
Dec-21
Jan-22
Feb-22
Mar-22
Apr-22
May-22
Jun-22
Jul-22
Aug-22
Sep-22
Oct-22
Nov-22
Dec-22
Jan-23
Feb-23
Mar-23
Apr-23
May-23
Jun-23

4. Finalmente, considere la siguiente situación:

   Las ventas de la cateogoría “Books” en el centro comercial “Forum Istanbul” tuvieron una leve reducción de 1.6% en el año 2022 con respecto a las ventas del año anterior. Sin embargo el equipo de ventas asegura que en 2022 en realidad el número de libros vendidos aumentó con respecto al año anterior.

   Haga un análisis precio-volumen, utilizando como base una tabla dinámica sencilla, para dar una explicación a esta cuestión, mostrando la variación en ventas, verificando si la afirmación del equipo de ventas es correcta (incremento en número de libros vendidos), y explicando por qué las ventas se habrían reducido a pesar de haberse vendido más unidades en 2022.

   Utilice una cuarta hoja en el archivo para hacer este análisis. Tome en cuenta que puede apoyarse de otras celdas, fuera de la tabla dinámica, para realizar los cálculos complementarios que soporten las conclusiones de su análisis.

Entregables de la Primera Entrega

Para esta primera entrega del proyecto, cada grupo debe enviar:

1. Un archivo de texto con extensión .sql, que contenga todos los queries de la Parte I y de la Parte II. Puede utilizar la propia herramienta de pgAdmin (Query Tool) para crear y guardar este archivo. Utilice comentarios de SQL para identificar el número de ejercicio al que corresponde cada query, por ejemplo:

```sql
--Query del Ejercicio #1
SELECT *
FROM CUSTOMER;

--Query del Ejercicio #2
SELECT *
FROM RENTAL;
```

   Utilice el siguiente formato de nombre para el archivo: Proyecto DA2 Queries Grupo X.sql, por ejemplo “Proyecto DA2 Queries Grupo 1.sql”.

2. El archivo de Excel donde está el Modelo de Datos que se creó en la Parte III y los reportes (gráficos dinámicos y tablas dinámicas) que se crearon en la Parte IV. Utilice el siguiente formato de nombre para el archivo: Proyecto DA2 Data Model Grupo X.xlsx, por ejemplo “Proyecto DA2 Data Model Grupo 1.xlsx”.

Ambos archivos deben ser enviados a la dirección de email: alexander.melgar@kodigo.org a más tardar en la fecha indicada en el Aula Virtual Moodle.