Exercise 6.1

Data Source:

This project uses **Looker E-Commerce Dataset**, publicly available on Kaggle. The dataset is based on a modern retail e-commerce store which has multiple interconnected tables that reflect real-world business operations. For this analysis, the data files which include Orders, Products, Users, Inventory items, Order Items, Events and distribution centers were used to explore customer behavior, product performance, and operational efficiency.

Data Limitation:

- Majority of the date fields have timestamps, different locations have different time-zones and therefore affecting time-based analysis.
- The dataset is artificially generated, thus not representing real customer behaviors or trends.
- Some Cities, Brand names, product names are missing from the data set.

Data Ethics:

- **Privacy:** This data set doesn't include PII (personally identifiable information) because the data is synthetically generated.
- **Accountability:** The findings that come out of this data set does not represent real world e commerce behavior or consumer actions.
- **Equality:** Since the generated data is synthetic, it could have unintentional bias toward some products or some geographical locations.
- **Transparency:** Everything from the generation/cleaning/insight generation on this data set is documented to assure no misrepresentation or manipulation was made.

Analysis Criteria:

- Exploratory analysis through visualizations (scatterplots, correlation heatmaps, pair plots, and categorical plots)

- Geospatial analysis using a shapefile
- Regression analysis
- Cluster analysis
- Time-series analysis
- Analysis narrative and final results (presented in your dashboard)

Business Questions:

- What are the top selling products by order and revenue?
- How has revenue trended over time?
- What are the purchase behaviors based on Age, Gender, Location?
- What are the most successful traffic source to purchase?
- What is the average delivery and processing time by region?
- What is the ratio of completed/Canceled/Returned orders?

Data Cleaning

1. Duplicate checks, consistency checks, missing value checks were done in all of the dataframes.
    a. Missing Values: orders
        i. returned_at: *112,696 rows*
        ii. shipped_at: *43765 rows*
        iii. delivered_at: *81342 rows*
    b. Missing values: products
        i. Name: *2 rows*
        ii. Brand: *24 rows*

        Created a map to pick the most frequent brand used according to name that is not null.
        Filled missing brand using the map

        Updated missing values = 22

c. Missing Values: Users
    i. city: *958 rows*

    Created a data dictionary based on available postal codes of missing city names. Then applied the data dictionary into the city column.

    Updated missing values: 0

d. Missing Values: inventory_items
    i. Sold_at: *308,946 Rows*
    ii. product_name: *29 rows*
    iii. Product_brand: *401 rows*

    Created a brand map to pick the most frequent product_brand used according to product_name that is not null.
    Filled missing product_name using the brand_map_brand

    Updated missing values: 365

e. Missing Values: Order items
    i. shipped_at: *63,478 rows*
    ii. delivered_at: *117,918 rows*
    iii. returned_at: *163,527 rows*

f. Missing Values: Events
    i. user_id: *1,125,671*
    ii. city: *23,080*

    Some rows were updated based on postal code that had existing city. Others were updated manually based on data dictionary created.

    Updated mssing values: *815*

g. Missing Values: Distribution centers - *0*

2. Columns dropped:
    a. Events: sequence_number, session_id, ip_address *(Not necessary for the analysis)*

b. Inventory_items: Product_sku (*Not necessary for the analysis)*

c. Users: email, street_address, latitude, longitude (*Not necessary for the analysis)*

d. Products: sku (*Not necessary for the analysis)*

3. Column names edited: *None*
4. Duplicates: *None*
5. Dataframes combined: Orders + Products + Users + Order Items + Distribution Centers
6. Column types changed:
   a. delivered_at to dateframe
   b. order_created_at to dateframe
   c. shipped_at to dateframe
7. Columns Derived:
   a.

Data Profile

Orders: *125,226 rows*

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   order_id     125226 non-null  int64
 1   user_id      125226 non-null  int64
 2   status       125226 non-null  object
 3   gender       125226 non-null  object
 4   created_at   125226 non-null  object
 5   returned_at  12530 non-null   object
 6   shipped_at   81461 non-null   object
 7   delivered_at 43884 non-null   object
 8   num_of_item  125226 non-null  int64
```

Products: *29,120 rows*

```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   29120 non-null  int64
 1   cost                 29120 non-null  float64
 2   category             29120 non-null  object
 3   name                 29118 non-null  object
 4   brand                29096 non-null  object
 5   retail_price         29120 non-null  float64
 6   department           29120 non-null  object
 7   sku                  29120 non-null  object
 8   distribution_center_id  29120 non-null  int64
```

Users: *100,000 rows*

```
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   id             100000 non-null  int64
 1   first_name     100000 non-null  object
 2   last_name      100000 non-null  object
 3   email          100000 non-null  object
 4   age            100000 non-null  int64
 5   gender         100000 non-null  object
 6   state          100000 non-null  object
 7   street_address  100000 non-null  object
 8   postal_code    100000 non-null  object
 9   city           100000 non-null  object
 10  country        100000 non-null  object
 11  latitude       100000 non-null  float64
 12  longitude      100000 non-null  float64
 13  traffic_source  100000 non-null  object
 14  created_at     100000 non-null  object
```

Inventory items: *490,705 rows*

```
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   id                      490705 non-null  int64
 1   product_id              490705 non-null  int64
 2   created_at              490705 non-null  object
 3   sold_at                 181759 non-null  object
 4   cost                    490705 non-null  float64
 5   product_category        490705 non-null  object
 6   product_name            490676 non-null  object
 7   product_brand           490304 non-null  object
 8   product_retail_price    490705 non-null  float64
 9   product_department      490705 non-null  object
 10  product_sku             490705 non-null  object
 11  product_distribution_center_id  490705 non-null  int64
```

Order items: *181,759 rows*

```
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   id              181759 non-null  int64
 1   order_id        181759 non-null  int64
 2   user_id         181759 non-null  int64
 3   product_id      181759 non-null  int64
 4   inventory_item_id  181759 non-null  int64
 5   status          181759 non-null  object
 6   created_at      181759 non-null  object
 7   shipped_at      118281 non-null  object
 8   delivered_at    63841 non-null   object
 9   returned_at     18232 non-null   object
 10  sale_price      181759 non-null  float64
```

Events: *2,431,963 rows*

```
 #   Column        Dtype
---  ------        -----
```

```
 0   id              int64
 1   user_id         float64
 2   sequence_number  int64
 3   session_id      object
 4   created_at      object
 5   ip_address      object
 6   city            object
 7   state           object
 8   postal_code     object
 9   browser         object
 10  traffic_source  object
 11  uri             object
 12  event_type      object
```

Distribution Centers: *10 rows*

```
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         10 non-null     int64
 1   name       10 non-null     object
 2   latitude   10 non-null     float64
 3   longitude  10 non-null     float64
```