**Module 2**

**About Assignment**

You have been hired as a **Power BI Developer** for **X Company**. Your manager is struggling to get the data in the right format. The manager is asking to look into the data:

The manager is finding various issues as follows:

**Data Understanding Issue:** Your manager is finding it difficult to understand the empty and null percentages in the dataset. He is also worried about the descriptive statistics (distribution) for all the top 1000 rows. You have been asked to show it in the power query.

**Imputation of missing values:** The data architect just informed your manager about a possible missing value with the column Region in the excel provided. He insisted on replacing all the missing values in Region with West. Now you are being asked to perform the data imputation.

**Date Format issue:** The data has been fetched from the European region’s server due to which the data format is of European standards **dd/mm/yyyy**. As the data is generated and relevant for the US time zone. Your manager is asking you to convert the date to US/Canadian format i.e **mm/dd/yyyy**

Q1) You've been given retail data, but it's spread across five different files this time. Your ultimate objective is to create a high-level dashboard with various cuts and insights. To begin, you must include all five tables in your data model and verify that the relationships between them are adequately specified for future usage. You must also provide some fields in the data to allow the various displays, such as Net Units, Weekday versus Weekend performance, and so on.

Dataset: Sales folder - all 5 files

1 - Load all the **5 files**, one by one into the data model - Make sure the header contains the field names in all the files

2 - Drop records from table **‘PinCode-Geo’** where ‘Zone’ is missing. Drop records from **‘Mod3\_Raw\_CityTier\_v0 1’** where ‘CityTier’ is missing.

3 - For the common columns between tables, make sure the relationship is present. For the table ‘Mod3\_Raw\_CityTier\_v0 1’,

make sure the ‘City’ field has a relationship with ‘City’ from ‘PinCode-Geo’ table

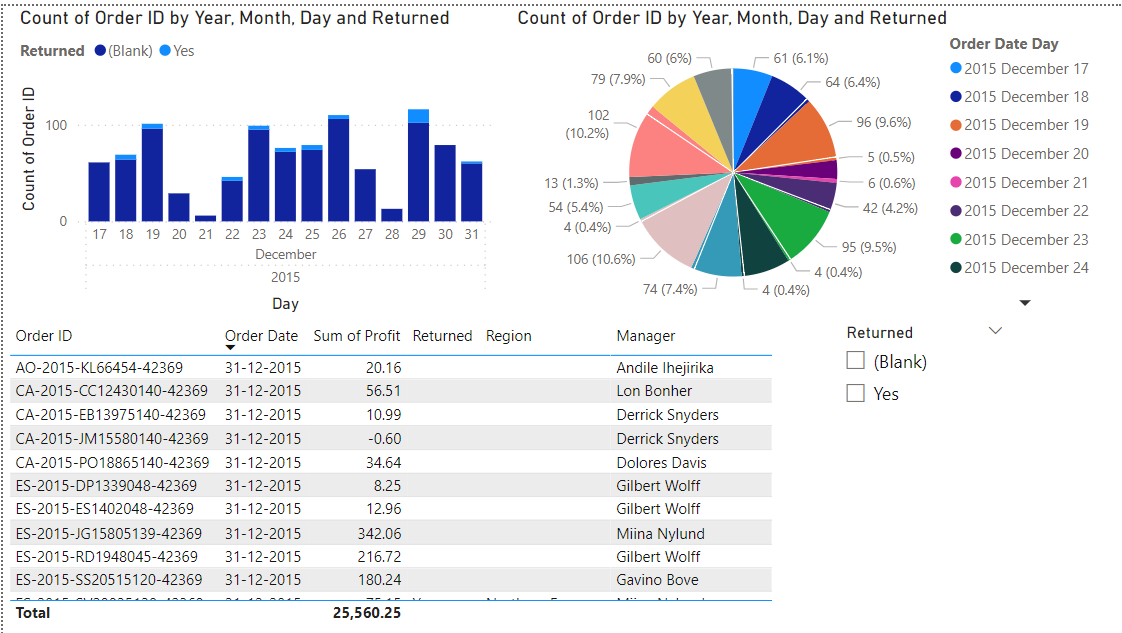
**Answers:**

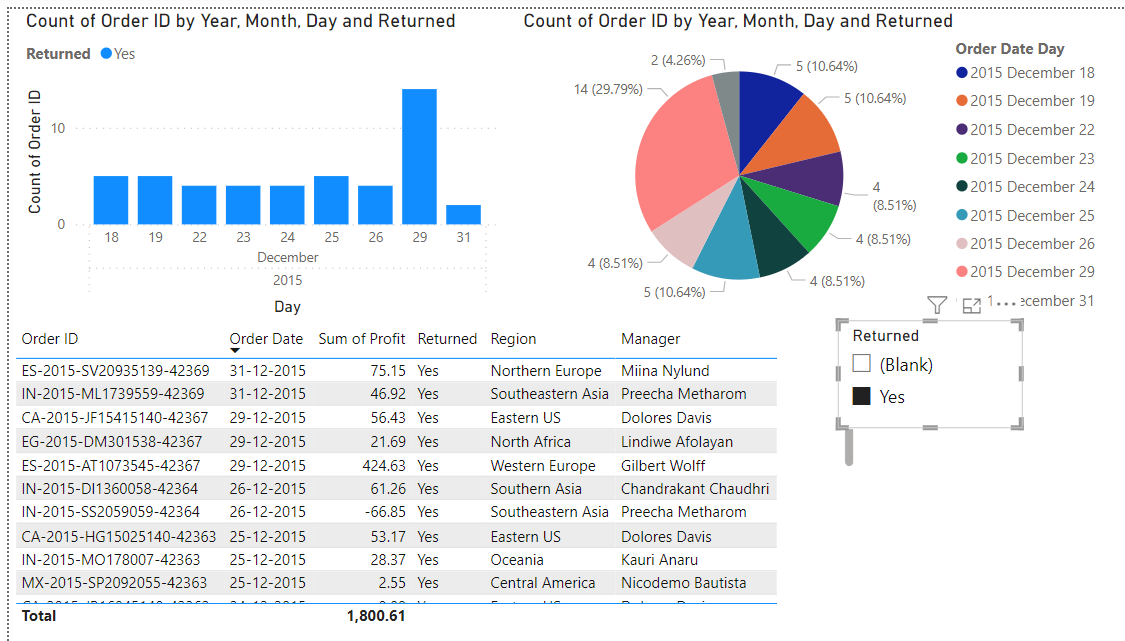
The dataset Mod 3\_global\_superstore\_2016\_v0.1 is a separate dataset and has relationship only with the Orders, Returns, People.

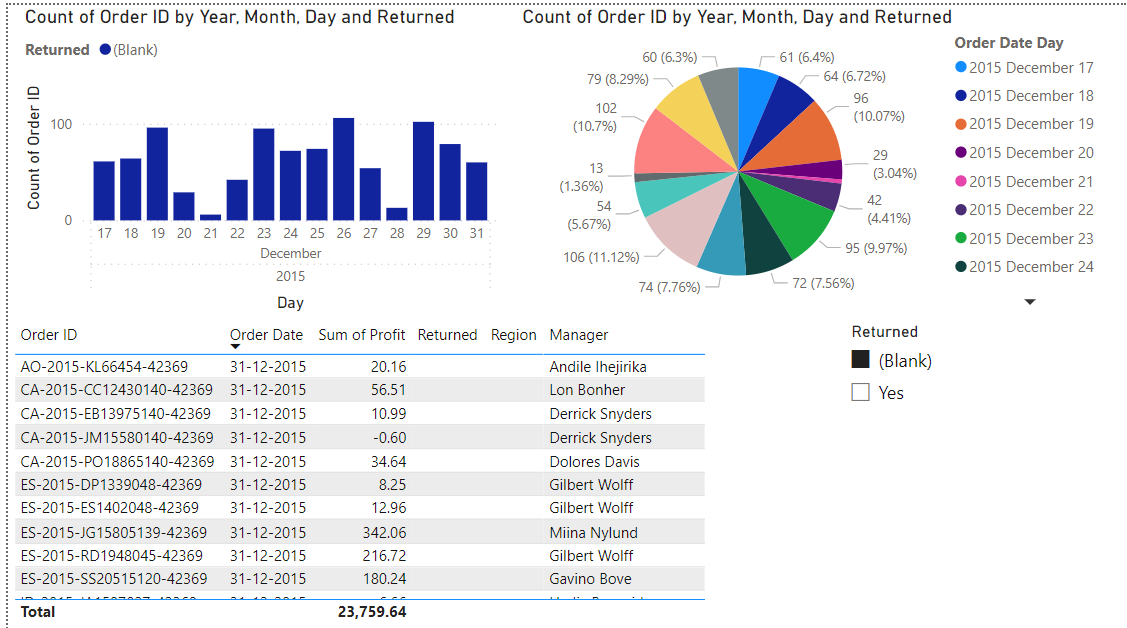
Here People data is representing the manager data. The relationships were created between Orders & Returns by linking the order ID and Orders & People(Manager) by linking the Region.

The plots are drafted as follows for the top 1000 rows filtered with newest order dates in mm/dd/yyyy format:

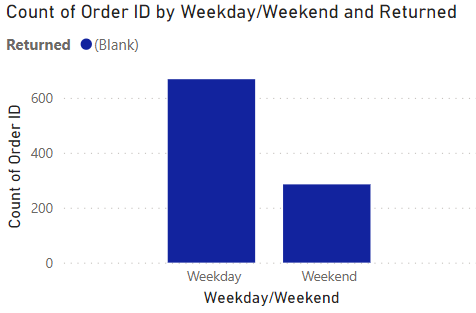
1. Clustered column representing the total number of orders not returned and total number of orders returned year wise.
2. Table representing the Order IDs, profit, Returned, Region and Manager.
3. Slicer for selecting the returns or all the orders.



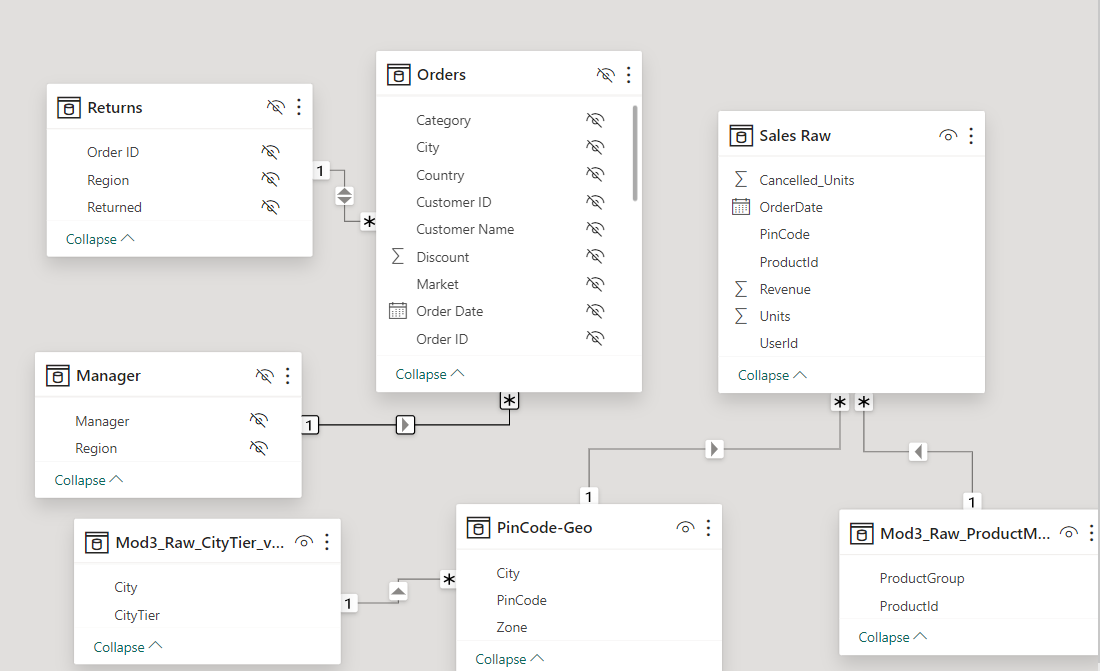




Weekday/Weekend = IF(WEEKDAY('Orders'[Order Date],2)>5, "Weekend", "Weekday")

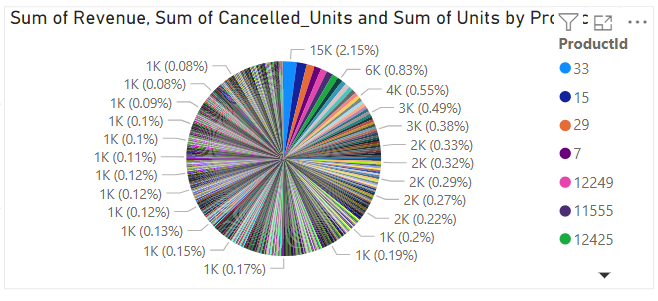


The datasets Mod3\_Raw\_CityTier\_v0.1, Mod3\_Raw\_PinCodeGeo\_v0.1, Mod3\_Raw\_Sales\_v0.1 and Mod3\_Raw\_ProductMap\_v0.1 have relationships among them and they have been established as follows



The plots are drafted as follows for the top 1000 rows filtered with newest order dates in mm/dd/yyyy format:

1. Table representing the Order date, sum of revenue, zone, city tier, sum of units and cancelled units. Since Product group has only one value in 1000 rows, this column has not been used.
2. Product ID is also not used due to less insights on it. Only one insight of product 33 performing well in terms of revenue was found



1. Pie chart and Clustered column representing the sum of revenue for each city tier and zone date wise with additional information like sum of units and sum of cancelled units.
2. Slicer for selecting the revenue based and zone and tier wise.

