Mini Project 1: Data Preparation and Warehousing

Due date: Tuesday, May 10, 8:00pm

Lab Assignment

Assignment Preparation

This is a **pair programming** project. Team up with another student during the **Tuesday, April 26** class.

Overview

During the first four weeks of DATA 301 we concentrated on learning how Python works with data frames, and how Python's pandas package can be used to obtain/ingest, prepare, and analyze data. We spent some time specifically discussing the notions of data warehousing and data cubes, and the kinds of data analysis one could do with their help.

In this project you will prepare and analyze a dataset of e-commerce sales of an on-line store.

Dataset. The dataset is a Kaggle.com dataset called Superstore, available here:

https://www.kaggle.com/datasets/vivek468/superstore-dataset-final

The CSV file with the data is available on the course web page as well.

The dataset presents a list of sales made by an on-line store over the course of several years. The data dictionary (the text below is copied verbatim from the Kaggle dataset description) is as follows:

- Row ID: Unique ID for each row.

- Order ID: Unique Order ID for each Customer.
- Order Date: Order Date of the product.
- Ship Date: Shipping Date of the Product.
- Ship Mode: Shipping Mode specified by the Customer.
- Customer ID: Unique ID to identify each Customer.
- Customer Name: Name of the Customer.
- Segment: The segment where the Customer belongs.
- Country: Country of residence of the Customer.
- City: City of residence of of the Customer.
- State: State of residence of the Customer.
- Postal Code: Postal Code of every Customer.
- Region: Region where the Customer belong.
- Product ID: Unique ID of the Product.
- Category: Category of the product ordered.
- Sub-Category: Sub-Category of the product ordered.
- Product Name: Name of the Product
- Sales: Sales of the Product.
- Quantity: Quantity of the Product.
- Discount: Discount provided.
- Profit: Profit/Loss incurred.

Goals. The ultimate goal of your project is to analyze the sales and the profit of the store in a variety of ways, to give the store management some insight on how the store is doing, which products are selling, what customer segments are bringing in the most profit and so on. To achieve these goals you also need to properly prepare the data.

Data Preparation

Your goal is to build a data warehouse that aggregates the following information:

1. Number of sales
2. Number of sold items

3. Total revenue from the sales
4. Profit/loss from the sales

by the following dimensions:

1. Month and Year of sale (based on the date when the order is received)
2. Region
3. State
4. Shipping Mode
5. Consumer Segment
6. Category and sub-category of the product

Some analytical questions below may require use of additional data, but the main analyses will be performed with the data warehouse that has the information above.

To accomplish this goal, you need to transform some of your existing variables into new ones. Specifically:

1. From the "Order Date" column extract the day of month, month, and year of the date of the order.
2. From the "Ship Date" column extract the day of month, month, and year of the date of the order.
3. Compute, for each order the number of days between the day the order was placed and the day it was shipped. Note: orders can be placed in one month, and shipped in another - you need to figure out how to properly compute the number of days between the order and shipping.
4. Using the "Sales" attribute and the "Profit" attribute, compute, for each order, the percent of profit/loss taken on that order.

Add all newly computed variables to the data frame you are maintaining. Then, with the enhanced data frame, perform a series of transormations to construct the data warehose data frame described above.

Analysis

Your goal is to answer the following questions.

Question 1. Analyze the sales by category by year. Are the sales in each category improving year-after-year? What are the categories of products with the best improvement in sales? What are the categories of products that lagged?

Question 2. Do the same analysis for subcategories, concentrating specifically on the subcategories of the categories that perform the best/the worst.

Question 3. Analyze the categories of the products for profitability by year. Compare sales numbers and revenue to profit for each category/year. Are there categories with large revenue and lower than expected profit? How about the other way around: are there cateogories that generate excellent profits despite smaller sales/revenue.

Question 4. Analyze sales vs profit by year and month. Are there specific months in each year where sales spike and/or profits increase? What are overall trends?

Question 5. What product categories are most popular in different regions? Are they the same or different? Is this popularity stable over time, or does it change?

Question 6. How are purchases different across different segments of customers? What categories and subcategories of products are most popular with each segment? What categories or subcategories generate most (least) profit for each segment.

Question 7. Analyze relationship between the time between order and shipment by shipping mode. Is this time the same or different for different shipping modes? Is the time between order and shipment remaining stable for each shipment mode over time (year-by-year), or is it changing? If it is changing - in what direction?

Question 8. Which segments of customers favor different types of shipping modes? Note, because of possible disbalance in the totlal number of shipments using each of the modes, you may want to analyze no just the raw numbers of orders, but also the percentages (i.e. what percentage of all first class orders is shipped to each segment, and so on).

Question 9. Visualize the relationships between the profit and the quantity of products sold in each category in each month. If using scatterplots, use size of the dots and their color to incorporate information from other variables available to you (number of orders, volume of sales, category of the product, year, etc...). Explain what you see in your visualizations (note, you can have multiple visualizations here - as many as makes sense for you to create).

Question 10. Visualize the relationships between the revenue and the profit for different categories/sub-categories of the products for different segments of the customers. If using scatterplots, use the size of the dots and their color to incorporate other useful for analysis variables.

Question 11. Create heatmaps (if you have not done so to answer earlier questions) showing the distributions of orders

- by customer segment and shipment type
- by customer segment and category of product
- by category of product and region
- by region and and month of sale for each year individually

Question 12. Analyze the distribution of sales by region. Have sales to specific regions changed significantly over time (year-to-year) as compared to other regions? Are there any states that account for a larger than usual number of sales/revenue/profit?

Question 13. Analyze the distriution of sales by region/state and product category and customer segment. Describe any interesting things you see.

For each question, apply appropriate slice-and-dice and rollup operations to construct the data frame containing precisely the data you need to study it. Conduct analysis via appropriate text-based, and graphics-based visualizations. Provide explanations for what you are doing, and the results you are observing.

Collaboration

While this lab is a pair-programming lab, and ”pair-programming” assumes ”both people work on the same task together”, I understand that you may need to do individual work due to different schedules outside of class. I am ok with work being split and one person performing the specific analytical tasks required for a given question. But I **expect** and **strongly recommend** that you discuss **what analysis to perform** for each question **together**.

Submission

Create one Jupyter notebook that has all code and explanatory text related to the tasks in this mini-project. Place all your work there in order of the questions. Your notebook can assume that the data file is located in the same directory as the notebook itself. Name the notebook Project01-Name1-Name2.ipynb where Name1 and Name2 are the last names of the students submitting the notebook.

Use `handin` to submit as follows:

```
$ handin dekhtyar 301-project01 <file>
```

Good Luck!