MINI PROJECT 3:

EXPLORATORY DATA ANALYSIS

OBJECTIVE

The goal of this Mini Project 3 is to conduct exploratory data analysis (EDA). Additionally you may want to incorporate basic modeling techniques to further your analysis.

BACKGROUND

Now that you have cleaned your data you will conduct exploratory data analysis to gain further understanding on the underlying dynamics inherent in your data. You will query your data to extract subsets of the overall dataset to conduct more specific analysis on key variables. Additionally you will aggregate your data using different criteria in order to compare and contrast different partitions of your data. You should also look to see the relationship between different variables.

You should liberally use the Excel tools you have learned including querying functions, aggregation functions including conditional aggregations (e.g. AverageIF), pivot tables, etc.

A useful and important tool to incorporate in your data analysis is to use visualizations of your data or aggregations of the data. Visualizing your data allows you to see patterns that may not be apparent when looking at detailed information. Remember the adage: “a picture is worth a thousand words”.

Use Excel’s data visualization tools including bar charts, histograms, pie charts, etc.

FOUR STEPS OF EXPLORATORY DATA ANALYSIS

1. Selecting columns of interest and target feature(s):
   - Which columns in your data sets will help you answer the questions posed by your problem statement?
   - Which columns represent the key pieces of information you want to examine (i.e. your target variables)?
   - How many numerical, textual, datetime etc. columns are in your dataset?
   - Pick out any similar columns among your disparate data sets for potential linking later on the EDA process

2. Explore Individual columns for preliminary insights:
   - How many null values are present in your data (what percentage)?
   - Plot one-dimensional distributions of numerical columns (ex. histograms) and observe the overall shape of the data (i.e. normal distribution, skewed, multimodal, discontinuous)
   - Calculate subgroup size of text/categorical data (i.e. with pivot tables)
   - Explore any date/datetime columns for basic trends. How long is the period of time covered by the dataset? Do any seasonality trends immediately become apparent?

3. Plot two-dimensional distributions of your variables of interest against your target variable(s):
   - Across different values of your independent variable, how does the dependent variable change?
   - Which interactions of variables provide the most interesting insights?
   - What trends do you see in the data? Do they support or contradict the hypothesis of your problem statement?

4. Analyze any correlations between your independent and dependent variables:
   - Understand and resolve surprising correlations between these variables, and use this information to validate your initial hypothesis. (You can do this using a scatter plot graph)

MORE IS BETTER!

There are many ways to do this. Expect to create a lot more material than you end up using. As you explore the data, you will run many tests that don't yield anything useful, and you will create many charts that don't end up helping you answer your questions. This is good! Keep track of what you did. What didn't work can help you tell your story, as much as explaining what did work!

Additional Excel EDA Resources:

- Data Analysis using Excel
- Exploratory data analysis
- The Ultimate Guide to EDA with Excel
- Dave on Data Youtube Channel

NEW DATASET:

For MP3 you will work with a new dataset [Additional store data] related to Specialty Foods' business. The dataset for this mini-project contains sales information related to purchases at Special Foods' stores. See the Data Dictionary below for the information contained in this new dataset.

DATA DICTIONARY:

- **Invoice id:** Computer generated sales slip invoice identification number
- **Branch:** Branch of supercenter (3 branches are available identified by A, B and C).
- **City:** Location of supercenters
- **Customer type:** Type of customers, recorded by Members for customers using member card and Normal for without member card.
- **Gender:** Gender type of customer
- **Product line:** General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel
- **Unit price:** Price of each product in $
- **Quantity:** Number of products purchased by customer
- **Tax:** 5% tax fee for customer buying
- **Total:** Total price including tax
- **Date:** Date of purchase (Record available from January 2019 to March 2019)
- **Time:** Purchase time (10am to 9pm) shown in 24 hour format in the data
- **Payment:** Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet)

- **COGS:** Cost of goods sold
- **Gross margin percentage:** Gross margin percentage
- **Gross income:** Gross income.
- **Rating:** Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

DIRECTIONS

PART I: INTRODUCTION & BACKGROUND

CASE STUDY: SPECIALTY FOODS INC.

You will continue to work on the Specialty Foods case for the exploratory data analysis step of a typical data analytics project. **Please refer to mini project 1 for a detailed description of the company, the previous data provided and other related information including your analytic tasks and objectives.**

INTRODUCTION:

Recall that given your data analytic skills, your manager has asked you to help the marketing team by gathering insights into the type of customers the company has and the products they buy. You are also asked to review past campaigns and suggest improvements for future marketing campaigns. In addition to gaining a better understanding of the business your analysis should result in specific recommendations on how the company can improve business results.

ANALYSIS:

For this mini project each team should work on exploratory data analysis. Please review the information provided at the beginning of this mini project that covers the various steps in the EDA process.

This mini project focuses on Exploratory Data Analysis & Data Visualization which follows on from mini project 1 & 2 that covered understanding the Business Problem and Data Curation, respectively. Mini Project 4 will cover advanced software tools specifically SQL and Tableau.

DATASETS:

1. Original datasets for Mini Projects 1 and 2 [sales, marketing, and customer]
2. New dataset for Mini Project 3 [Additional store data]

PART II: ANALYSIS / DATA ANALYTICS

DELIVERABLE:

Place the answers to the questions below in a new spreadsheet in the customer file and call it: Answers. If additional sheets feel free to add them and name the sheets appropriately to identify what questions you are answering.

QUESTIONS:

After doing a great job with your original data [sales, marketing, and customer], your marketing team sends you new data in a CSV format to analyze in order to advance the business. This time the data is for the multiple stores they have.

You should explore each variable further with aggregation functions, pivot tables and further statistical analysis. What new observations do you have now that you have performed EDA that you did not have before? See detailed questions below to help guide you in your Exploratory Data Analysis.

In addition to brainstorming with your team members, you should discuss with your TAs and learn from other fellows to explore other techniques for exploratory data analysis.

For your analysis your team should begin by answering the questions below. However these questions are aimed only to get you started and practice your skills. You should consider further analysis and determine what additional questions will help in both understanding the business and making recommendations to improve the business's results.

1. It is important to always check any data you receive for errors, mistakes, etc. So before starting to work with this new Dataset, we recommend that you check out the data set and clean whatever needs to be cleaned. Once, the data has been cleaned, you will need to write a small paragraph that explains your process, edits, and changes.
2. Place your cleaned data In the "Store Data Clean" sheet:
   - a) Create a new column named "Net Income". In this column calculate the Net Income by deducting the taxes from total (Total - Tax).
   - b) Find which location has the highest Net Income.
3. Using a Pivot Table:
   - a) Calculate the sum of Payment methods for each Customer Type with each Gender.
   - b) Based on your calculations, create a chart with all the metrics and appropriate labels.
   - c) Write a small paragraph describing what you see in the chart and what you can conclude.
4. Using VLookup or Index+Match, add the Income Column from the original data provided (for MP1 & MP2) onto the Clean Store Data sheet.
5. Calculate the Mean, Mode and Median from the COGS column:
   - a) Create a chart representing your COGS.
   - b) How would you describe the distribution of this variable? Is it skewed, symmetric, or multimodal?
6. Calculate the sum of the Total:
   - a) From 1/1/2019 to 1/31/2019
   - b) From 2/1/2019 to 2/28/2019
   - c) From 3/1/2019 to 1/30/2019
   - d) Compare the 3 results. Which month had the highest Total? Make a hypothesis on why that month had the highest Total.

BONUS QUESTION (OPTIONAL):

For the Optional section please read this material on statistical analysis / modeling and feel free to reach out to your TAs with any questions.

Below are some guides on statistical methods and how to model your data to gain insights. Note that this material is beyond the scope of this program and is provided to challenge you to 1) research new data analytic techniques 2) conduct independent, self-learning which is a valuable aspect in your data science journey as there are many techniques & methods available based on the data and specific use case, i.e. business problem.

As this is optional there is no specific deliverable. Rather write a short paragraph on what you learned and found interesting, particularly highlight any statistical analysis that could help in analyzing Specialty Foods.

STATISTICAL METHODS (OPTIONAL)

Here are some resources:

- The Beginner's Guide to Statistical Analysis | 5 Steps & Examples
- 5 Statistical Analysis Methods That Take Data to the Next Level
- 7 Types of Statistical Analysis Techniques (And Process Steps)

MODELING (OPTIONAL)

To learn more about modeling you can check out the following:

- Data Science Modelling: 8 Easy Steps
- Top 10 Data Science Algorithms You Must Know About
- 6 Predictive Models Every Beginner Data Scientist should Master

PART III: TEAM WORK & INDIVIDUAL PARTICIPATION

An important part of this program requires each fellow to work in a team to complete the four Mini Projects. Together these mini projects cover key aspects of an overall data analytic project similar to those you will face when employed as an analyst.

Similarly many times as an analyst you will work in teams. As part of this program each fellow needs to balance personal and work commitments with the team work required in this program. It is your team's responsibility to determine how best to communicate with each other to complete the Mini Projects and when you will meet outside of the regularly scheduled Tuesday and/or Saturday sessions, if necessary.

DELIVERABLE:

In addition to answering the questions in Part II above, each team member should create a separate word document explaining what role they played in completing the Mini Project. In the document you will need to include the following:

- **Team name:** List the name that your team agreed upon.
- **Responsibilities:** List what your responsibilities were throughout this project. In addition, if a teammate did not participate please list their names and indicate that they did not contribute.
- **Learning Outcomes:** Each team member will be expected to submit their own individual word document. In addition to listing your responsibilities please explain what you learned or got out of this Mini Project.

You can access the Written Answer Template here.