MINI PROJECT 2:

DATA CURATION / WRANGLING

OBJECTIVE

The goal of this Mini Project 2 is to practice the various steps in the data curation process.

BACKGROUND

DATA CURATION

Data Curation, also referred to as data cleaning, involves three steps: Data Sourcing, Data Profiling, and Data Wrangling. Below are more detailed explanations of each step in the data curation process followed by instructions on what to turn in for the Data Curation deliverable.

STEP 1: DATA SOURCING

You have already been given the primary dataset consisting of the three files. It is sometimes necessary or helpful to supplement your data with additional datasets to assist in your analysis. For the purposes of this Mini Project this is not required but you should be aware and think about what additional, external data sources could complement your analysis.

STEP 2: DATASET PROFILING

Data profiling refers to the process of "examining, analyzing, reviewing and summarizing data sets to gain insight into the quality of data". The purpose of this step is to make sure the data is suitable for analysis.

There are three main types of data profiling:

1. Structure discover
   - Checking data types and format for consistency
   - Mathematical checks where necessary (e.g. using sum, min, max)
   - Exploring distinct values in each column

2. Content discovery
   - Looking for errors in individual rows?
   - Identifying zero/blank/null values

3. Relationship discovery
   - Check if data is related to other cells

Typical output for this step in the data curation process would be identifying and documenting your findings from the above. There are advanced profiling techniques that we encourage fellows to try. Please see **panoply** for more information.

STEP 3: DATA WRANGLING

With your datasets sourced, you now need to wrangle them. This entails cleaning your datasets and setting them up to be compatible with one another. Keeping track of this is important. You may need to demonstrate how you transformed the raw dataset into what you used in your analysis. You might need to go back and see how the variables from the different datasets align.

The exact steps differ from project to project but here are some examples of data wrangling:

- Merging multiple data sources
- Identifying missing data and how to handle it
- Deleting unnecessary data
- Identifying outliers
- Checking and fixing data types

The main steps of data wrangling are:

1. Data structuring
2. Data cleaning
3. Data enriching
4. Data validating
5. Data publishing (this step is not necessary)

To learn more about data wrangling, we recommend the following:

- **Data Wrangling: What It Is & Why It's Important (hbs.edu)**
- **Data Wrangling (Wikipedia)**

WHY ARE WE DOING ALL THIS?

When looking for good data, a lot of what we do leads nowhere. This can be extremely frustrating because it can feel like we wasted our time. However none of this time was wasted! Remember the devil is in the details. Gaining a better understanding of your data allows you to identify your key variables and organize your data well for analysis. Subsequently you will have a sense for what types of visualizations to create!

Typical output from this step is detailed documentation of the wrangling steps you used on your data. In this section these may include:

1. Listing all the cleaned datasets you are using.
   - Document how each one was wrangled
2. Note down the size of the dataset in terms of rows, columns and file size.
3. List the original source(s) of the cleaned data set.
4. Creating a data table schema of the cleaned dataset.
   - This will outline each field in the dataset, value type, and a brief description of each column.

DIRECTIONS

PART I: INTRODUCTION & BACKGROUND

CASE STUDY: SPECIALTY FOODS INC.

You will continue to work on the Specialty Foods case for the data curation step of a typical data analytics project. **Please refer to Mini Project 1 for a detailed description of the company, the data and other related information including your analytic tasks and objectives.** For your convenience we have reposted some key information below.

INTRODUCTION:

Specialty Foods Inc. is a food retailer focusing on the higher end of the market. You are a new member of the marketing team that was hired based on your data analytic skills. The company is interested in improving business results through more data-driven analysis and decision making. Traditionally the marketing department has launched campaigns to increase sales using qualitative analysis that has focused on previous experience and an understanding of the market.

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.

BUSINESS OVERVIEW:

Specialty Foods sells products within five main categories: wine, meats, fruits, seafood, and sweets. Each of the aforementioned categories are further divided into standard and premium products. The company has three sales channels and items can be purchased through physical, in-store locations, catalog sales, and through the company's website. The marketing department periodically uses different campaigns to increase sales.

ANALYSIS:

For this Mini Project each team should work on curating the data. Please review the information provided at the beginning of this Mini Project that covers the various steps in the data curation process.

This Mini Project focuses on Data Curation which builds upon Mini Project 1 that covered understanding the Business Problem. Subsequent Mini Projects will cover additional topics in the analytic process and include Exploratory Data Analysis & Modeling and Data Visualization & Storytelling including the advanced tools of SQL and Tableau.

DATASETS:

In order to conduct your analysis you are given data on the company's **customers**, **sales**, and previous **marketing** campaigns. The data you have available to you contains both socio-demographic and company specific information. There are 3 separate data sources each containing information on customers, sales and marketing campaigns.

Please refer to Mini Project 1 for a detailed description of the company, the customers, sales, and marketing data, and other related infomation including your analytics tasks and objectives.

PART II: ANALYSIS / DATA ANALYTICS

DELIVERABLE:

Place the answers to the questions below in a new spreadsheet in the customer file and call it: **Answers2**. If you need additional sheets feel free to add them and name the sheets appropriately to identify what questions you are answering.

QUESTIONS:

Remember you should get your data in good shape to conduct your analysis. Your team should answer the questions below regarding your data. However you should not limit your work to only these questions. As a data analytics professional you should ask additional questions about the quality of your data and seek out other techniques you can use to produce a clean dataset to conduct your analysis in addressing your business problem.

In addition to brainstorming with your team members, you should discuss with your TAs and learn from other fellows to explore other techniques in the data curation step.

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. What common column / variable can you find in your tables in order to link the separate data sources? Once you have identified this link, how would you join the tables? You should create a new file that contains all the information from the three separate tables and place it in one spreadsheet. Hint: use VLookup()
2. Are there other ways to connect your tables? What other functions have been covered in lectures that enable you to do this? Can you find other ways to link your data? Hint: use Google, YouTube, Microsoft documentation to search for other ways to link the data.
3. When you clean your data sometimes you may want to combine similar information found in separate columns into one column. Do you have similar categories of data that could be combined in order to conduct your analysis? You should combine some of your columns that contain similar information into a new column. Hint: look at some of the social-demographic columns, for example information related to marital status contained in separate columns can be combined into one new column.
4. As a data analyst, you should get an idea of the type of information you have in each column and what the values represent such as monetary values, item counts or categorical variables. What are the data types for the information you have?
5. Statistics provide useful summary information of data. You should perform basic summary statistics on the most "important" columns (give this thought). Hint: what are the extreme values (min, max)? Which values occur the most (mode)? How often do values occur (frequency)? What is the range of values? Within the range of values is the data concentrated more toward the lower end of the range or the higher, or is it evenly distributed, etc? What is the distribution of values?
6. Are there simple Excel charts that can assist you in visually analyzing the data from the previous questions? Please provide examples of these charts for some of the "important" variables.

7. Part of data curation involves cleaning your data and removing values that may hinder your analysis as they are incorrect inputs or are so different from the rest of the data that they do not provide valuable information. You may want to remove some of the rows of data that are irrelevant but be careful as some values that may appear to be irrelevant but may actually represent valuable information that needs to be considered in your analysis. You should identify which values of each of the columns, if any, do not make sense and then determine if you should remove the row. Hint: look for blanks and missing values, values that are outside of the typical range (eg negative or extreme values) or do not make sense given what the data represents (eg decimals).

BONUS QUESTION (OPTIONAL):

For this optional question, create an additional new spreadsheet in the customer file and name it '**Optional**'.

1. Once you have curated your data are there any issues you foresee that may cause problems in your data analysis for the business problem you are addressing?
2. Now that you have understood your data better, can you think of additional analytical tools that will help you to perform your analysis?
3. Having begun the process of working on a data analytic project, can you see how data-driven decisions may improve on qualitative analysis management may have previously used to improve business results? If so, what are some examples?

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 should create a word document explaining what role each member played in completing the Mini Project. In the document you will need to include the following:

- Team name: Choose something everyone can agree on that represents the group members.
- Responsibilities: Each member should take responsibility for various aspects of project-based work. In this section, write the name of the fellow and what was their contribution. If any fellow did not participate please list those individuals.
- Individual Submission: Each team member should separately provide a brief explanation in the document explaining what they individually learned or got out of this Mini Project. Please include your name.