Capstone 2: Working with Business Data in Python

For this project, you will clean, analyze, and visualize business-related data using Python and its data analysis libraries.

Description

In this capstone project, you will demonstrate your ability to use Python to clean and analyze data. The scenario for this project is again working with data in a business environment:

- Your growing business is developing a marketing list, but the application you use to collect customer info generates that list in an awkward format, and the data contains some odd errors and missing values. You need to develop a Python program to take that list as an input, separate it into columns and rows, and clean up bad data, then output the results to a .CSV file.
- You also now have a set of 2024 transaction data to help you understand how your business is performing. You need to create an analysis of this data that reveals some information or insights about your business, including at least one chart or graph.

Basic Requirements

Your capstone project must include **all** of the following to be considered complete:

- Python program (in the form of a console application or a Jupyter Notebook) which separates customer records into columns, performs data cleanup, and exports the results to a .CSV file
- Jupyter Notebook using Python, pandas, and Matplotlib to analyze given sales data, including:
    - presentation of the data as a DataFrame, showing sample of the top 10 rows and including DataFrame description (i.e. basic statistical summary)
    - additional insights generated from the data using aggregation or statistical functions
    - at least one data plot
    - a short text explanation of each item, formatted in Markdown
- Well-organized project repository including:
    - at least one commit per day
    - brief description of project in README file, explaining the purpose of the project and describing both the major components bulleted above
    - final commit of project content by 10:00AM on presentation day

As before, at the conclusion of the capstone week, you will present the results of your project to the class. You will also participate in Q&A and sharing feedback on your classmates' presentations.

Note: You will have plenty of opportunities in this project to keep you busy and stretch your skills. Just make sure the core requirements above are met!

Project Details & Implementation

1. The capstone project includes two major components: cleanup of the customer list and analysis of the transaction data set. You can tackle these in whatever order you choose, but you will need to show a basic implementation of *both* of these components for your capstone to be considered complete. Plan your time this week accordingly!

2. Additionally, professional-level organization, documentation, and commit history are expected for your project repository in this second capstone. You must make a minimum of one commit per day, and all relevant project files must be included in your repo. Each commit should include a message that meaningfully reflects what was updated. Your repository should also include a README file that summarizes the purpose and major features of the project.
   Note: If you include anything in your capstone presentation that is not saved in your repository, it will not be counted toward your capstone grade.

3. Include appropriate code comments throughout both project components.

4. If you run into any hurdles with your project, don't wait to reach out to your instructor with questions! Ask for help if you get stuck. You are also welcome to ask your peers for their suggestions and feedback as you work through problem-solving. Just make sure, as with the first capstone, that the substance of your project reflects *your own independent work*. This is your opportunity to test your skills and show off what you can do!

- Customer List Cleanup -

5. It is recommended that you start by reviewing the data in the customer list to understand what it includes, how it is organized, and what kind of steps you will need to take to process it and return a clean result set. Some key considerations include:

   * Stray characters are appearing in the name field. The only characters that should appear in a name are letters, hyphens for hyphenated names, and periods following an initial.
   * Some of the phone numbers were incorrectly entered. A valid phone number should be in the format NNN-NNN-NNNN (numbers separated by hyphens).
   * For some records, `sms-opt-out` was left blank. Customers who opted *in* to receive marketing text messages are marked with N. Customers who opted out, either by declining to sign up for texts or canceling by replying STOP, are marked Y. You must decide how to deal with the blank values for this field. (You may want to do a little quick internet research on issues related to SMS marketing as you determine your approach.)

6. You will need to write a program that performs data cleanup on the list. Your program may be formatted in one of two ways:

7. Your cleanup program may be saved as a .py program file that can be run from a command-line interface (like Git Bash, Windows Command Prompt, or Mac Terminal). If you choose this approach, your final program should include the following elements:
    * ask the user for the file name (or full file path) for the file to be processed
    * output status messages to the console with each data cleaning step executed
    * write the cleaned data to a DataFrame and print a preview to the console
    * output a final status message confirming the save name and location of the generated .CSV file

8. Alternatively, your cleanup program may take the form of a Jupyter Notebook. If you choose this approach, your notebook should include the following elements:
    * ask the user to input the file name (or full file path) for the file to be processed
    * include separate executable code blocks for each data cleaning function performed along with appropriate labels for each step
    * write the cleaned data to a DataFrame and preview it in the notebook
    * return a final status message confirming the save name and location of the generated .CSV file

- Transaction Analysis -

9. It is recommended that you start by reviewing the data in the customer list to understand what it includes, how it is organized, and what kind of steps you will need to take to process it and return a clean result set. Some key considerations include:
    * Understand whether the data is normalized or non-normalized and how this may impact your approach to analysis. Be mindful of double-counting duplicated values!

10. Your analysis of the transaction data must take the form of a Jupyter Notebook, which *must* include the following elements:
    * conversion of the provided data set to a DataFrame, with a sampling of the top 10 rows and statistical summary using the describe attribute
    * additional insights generated from the data through the use of aggregation and/or statistical functions (aim for at least 3!)
    * at least one data plot, including basic formatting such as a title, x- and y-axis labels, and any other formatting appropriate to the data
    * a short text explanation of each item in the notebook, using Markdown cells (logical headings are also recommended)

- Capstone Project Add-ons -

11. Once you have met the core requirements for the capstone project, you may add on to your project with one or more of the below suggestions.
    * ***OPTIONAL***: Add a third component: combining the customer list data and transaction data to show information about transactions by customer. This may be in the format of either a separate .py script or new notebook. Once you have combined the data sets, create your own analysis of customer purchases.
    * ***OPTIONAL***: Perform an analysis of the customer data set to identify patterns or groupings you may notice in your customer data. You might also try adding in basic geographic information based on the area code associated with each phone number. (See public data from the NANPA, specifically https://nationalnanpa.com/enas/geoAreaCodeNumberReport.do for a list of area codes by state.)

- Capstone Presentations -

12. You will again be presenting your project results to the class. Presentations will be followed by Q&A and feedback from your fellow students and the instructor as time permits.

13. When it is your turn to present, imagine you are meeting with your business co-owners or investors who are currently working with you on this business and have a stake in its success.
    * **You will have 10 minutes for your presentation.** Be prepared to walk through the process and results of both your customer list cleanup and transaction data analysis.
    * Share your screen while presenting, and give a live demonstration of each component of your project. Specifically, you must run your cleanup program and demonstrate the output and results. You must also demo your notebook with analysis of transaction data, walking through the approach you took and the results you found.
    * Use notes or a brief set of slides to help you walk through your presentation in a well-organized manner. If you do choose to use any PowerPoint slides, keep them short and to-the-point.
    * Ideally, talk about at least one interesting or unexpected thing that you learned from **each** data set.
    * ***REMINDER***: The most important thing for your capstone presentation is showing the data – what you learned from the data, what insights you uncovered, and how you propose to use those insights to make decisions in the business.

14. **When your classmates are presenting**, take notes about anything that strikes you as particularly helpful, anything you find confusing, or any questions that you have.

* **Your participation in Q&A is expected and considered part of your capstone work.** You don't have to chime in on every project, but you should ask at least one question or share at least one piece of feedback.
* You can also use Zoom reactions during your classmates' presentations to react to things you find helpful, insightful, or clever.
* When giving feedback, focus on **constructive feedback**. The purpose of providing feedback is to help the speaker improve and refine their work. For every speaker, try to identify at least one thing you found effective and at least one thing that could be improved or clarified. Examples:

  I really liked where you...
  It was really useful to see the information you found on...
  I was confused about the part where...
  I would have liked to hear more about...

Tips for a successful capstone project

- Make sure you implement all the requirements from page 1!
- Save all your project files to your repository, and make sure you regularly commit changes using meaningful commit messages.
- Use good naming conventions throughout your project, including for big things like file names as well as little things like variables and functions inside your code.
- For your class presentation, it is *highly* recommended that you take some time to do a practice run or "dress rehearsal" of what you plan to say, and time yourself. Ten minutes can go fast!
- On presentation day, have the files you plan to share open or easily accessible, so that you're ready to go when your turn comes up.
- You are welcome to brainstorm ideas with your classmates or ask each other for feedback while you are developing your projects. Just make sure that the substance of your project reflects *your own independent work*. This is your opportunity to test your skills and show off what you can do!