CS 46 W - Project Specification - CSV Processing Categorization

Version 3

1. Project Introduction & Overview

1.1 A Quick Reminder: Don't Panic

There's lots of detailed information in this document. Please don't panic. Ask me clarification questions regarding anything that is confusing. I strongly suggest that you take the approach of doing a little work on the project every week. If you wait until the last week, then you are guaranteed to panic.

Image credit: https://theventureguide.com/wp-content/uploads/2016/07/Elephant-1-bite-at-a-time-750x875.jpg

KEEP CALM AND EAT THAT ELEPHANT ONE BITE AT A TIME

CS 46 W - Project Specification - CSV Processing Categorization

1.2 Backstory

Your sibling Bob Smith heard that you were taking a programming course in Python (you might have mentioned it on Twitter or something). So Bob emails you asking for help on a budgeting problem. Here is a snippet of that email:

Ok so as you know I don't trust Intuit and their Quicken finance software. A few years back I was burned by their software in that I would download statements (as qif or qex or whatever the damn format was at the time) and use Quicken to categorize it. Then one day I did my usual monthly Costco credit card statement download and Quicken didn't work. It said something like "You need to use latest version of Quicken to do your budget". I was so pissed that I swore to never use Quicken or any other proprietary budgeting software (and don't get me started about [Intuit and simplified tax forms](#)).

Anyway, I love downloading those simple credit card CSVs of my monthly bill but categorizing them is becoming a bear. Can you use your new found programming knowledge to make the budget categorization process easier on me?

1.3 Overview

The following section define the specification of the project as follow:

- "Bob's current (manual) process of budget categorization" section: describes how Bob does his budgeting categorization now. This section is the "customer's use case". It should help provide context and put you in the customer's "shoes".
- "Summary of project requirements": describes the solution to Bob's process in the form of your project. It also describes what you need to hand in.
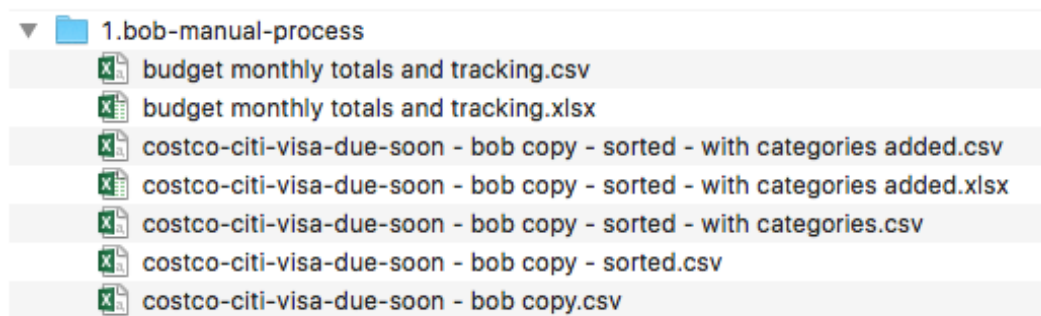
Note: **you can find all files and folders mentioned in this document** at the following box.com address: https://app.box.com/s/up19papd1ec60vfqcp4yl7ss92q99xcd

CS 46 W - Project Specification - CSV Processing Categorization

2. Bob's current (manual) process of budget categorization

Here is how Bob does it now. Your goal will be to automate this process so Bob's life is less miserable when the Costco credit card bill arrives and it is time to do the monthly categorization of the Costco credit card (for use in his budgeting process).

Note: You will find the files mentioned in this section within the "1.bob-manual-process" folder at the box.com url that was mentioned in the overview step (above).

▼ 📁 1.bob-manual-process
  📄 budget monthly totals and tracking.csv
  📄 budget monthly totals and tracking.xlsx
  📄 costco-citi-visa-due-soon - bob copy - sorted - with categories added.csv
  📄 costco-citi-visa-due-soon - bob copy - sorted - with categories added.xlsx
  📄 costco-citi-visa-due-soon - bob copy - sorted - with categories.csv
  📄 costco-citi-visa-due-soon - bob copy - sorted.csv
  📄 costco-citi-visa-due-soon - bob copy.csv

CS 46 W - Project Specification - CSV Processing Categorization

2.1 Go to citi.com website and download latest statement transactions as csv file

CS 46 W - Project Specification - CSV Processing Categorization
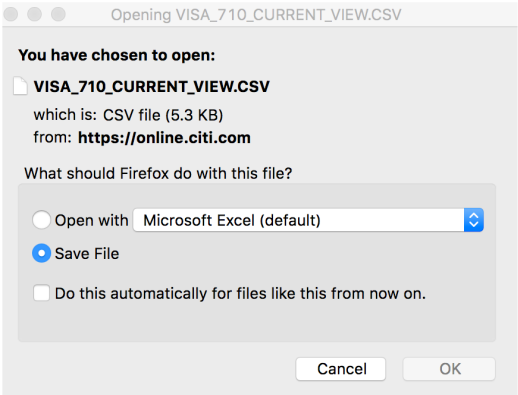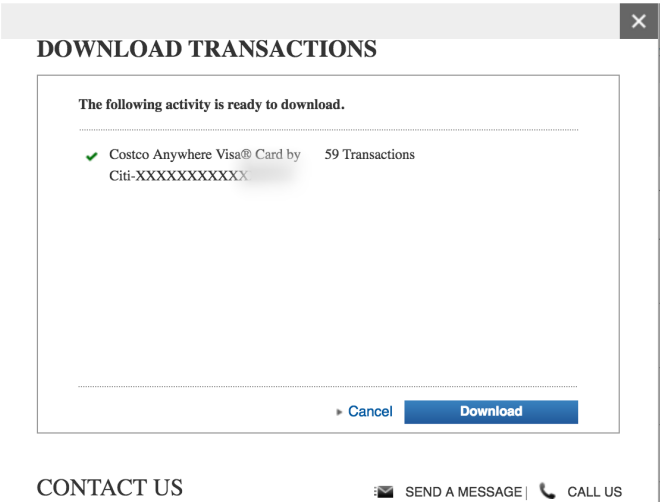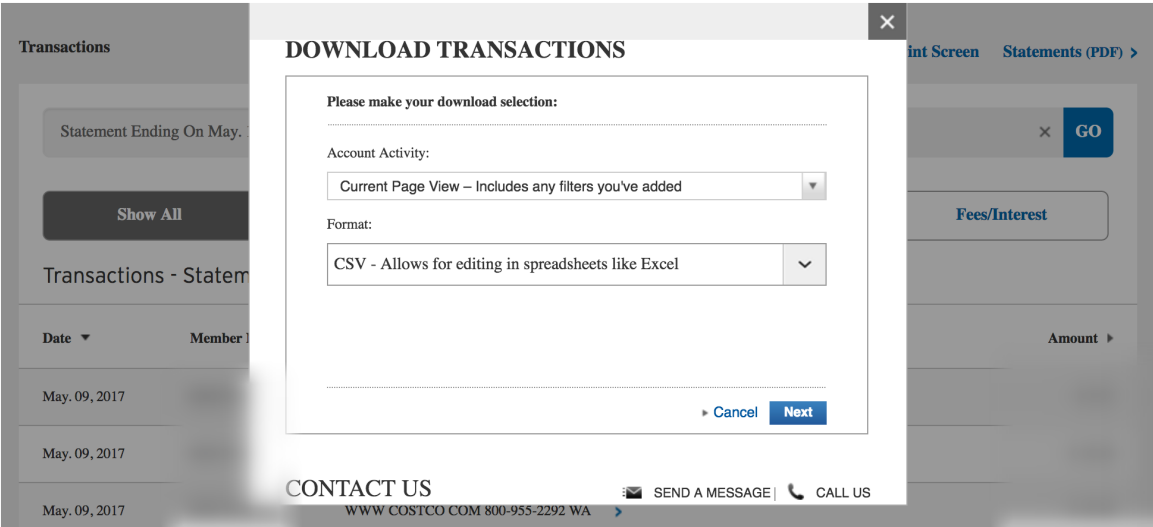
CS 46 W - Project Specification - CSV Processing Categorization

2.2 Make a copy of csv file and opens it up in excel

See: costco-citi-visa-due-soon - bob copy.csv

CS 46 W - Project Specification - CSV Processing Categorization

2.3 Sort based on business name

See: costco-citi-visa-due-soon - bob copy - sorted.csv

Sorting based on C column in Excel.
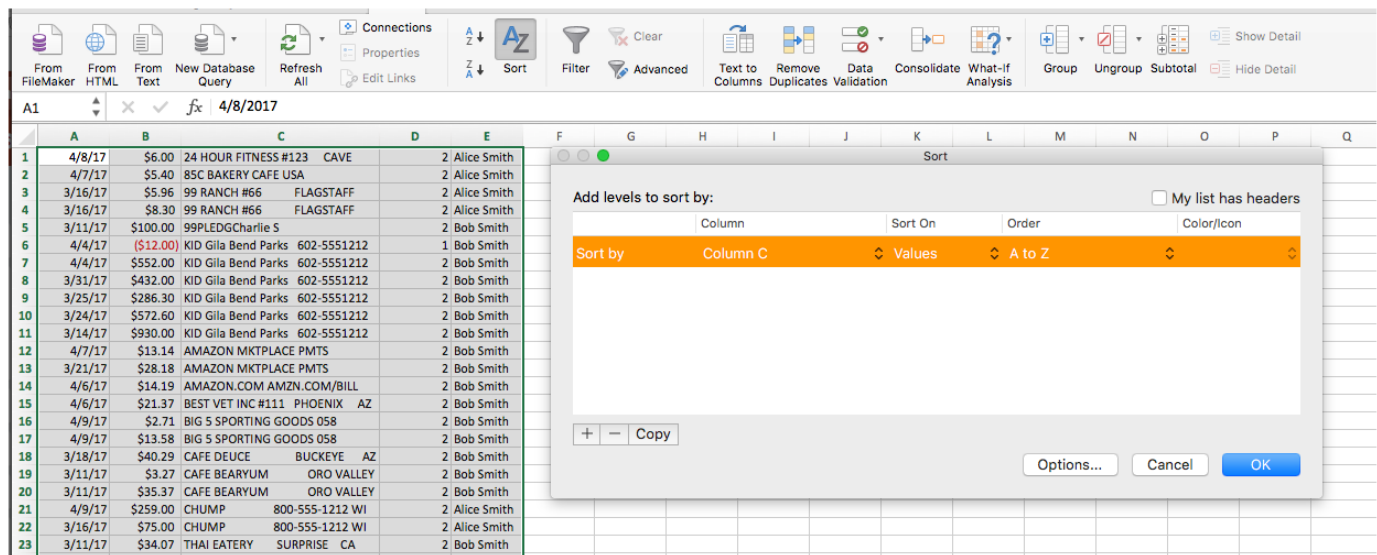
CS 46 W - Project Specification - CSV Processing Categorization

2.4 Insert category column and other column headings

See: costco-citi-visa-due-soon - bob copy - sorted - with categories.csv

Since this is Bob and Alice's "shared" card (they both use if for shared expenses), he ignores the "who" column and the "not sure" column.

Bob then inserts a category column at the end as well as some additional columns

date	amount	vendor	not sure	who	category
4/8/17	$6.00	24 HOUR FITNESS #123		2 Alice Smith	
4/7/17	$5.40	85C BAKERY CAFE USA		2 Alice Smith	
3/16/17	$5.96	99 RANCH #66		2 Alice Smith	
3/16/17	$8.30	99 RANCH #66		2 Alice Smith	
3/11/17	$100.00	99PLEDGCharlie S		2 Bob Smith	

2.5 Put a category name for each row

See: costco-citi-visa-due-soon - bob copy - sorted - with categories.csv

Note that these categories refer to flexible expenses. So fixed expenses like mortgage, car insurance, home insurance would be part of a different spreadsheet and are not part of this. We're only focusing on credit card transactions.

Possible categories for this shared (between Alice & Bob) credit card are:

- entertainment: optional activities like eating out, movies, etc..
- food: non-optional food purchases
- household: items related to home (like Ikea furniture items, home depot items, etc...)
- pet: items related to pet such as food, treats, toys
- pet-medical: items related to pet health
- kid: items related specifically to kid (doesn't include common things like food)
- kid-medical
- travel: items related to family trip (hotels, gas, etc...)

- utilities: gas, electric
- gift: gifts from family (like for birthday parties that child goes to, relatives, etc...)
- charity: items for charity (not accounting for tax deductibility or anything like that)
- find-detail: this category would typically be for Amazon purchases where bob would need to go into his amazon account and find the item and then categorize it in the spreadsheet later
- ignore: items to ignore in categorization (like payment of last bill)
- fixed: a fixed expense like insurance
  - fixed-insurance

Note: If there's a vendor that your program cannot identify, then it should put a category name of WARNING-NO-CATEGORY.
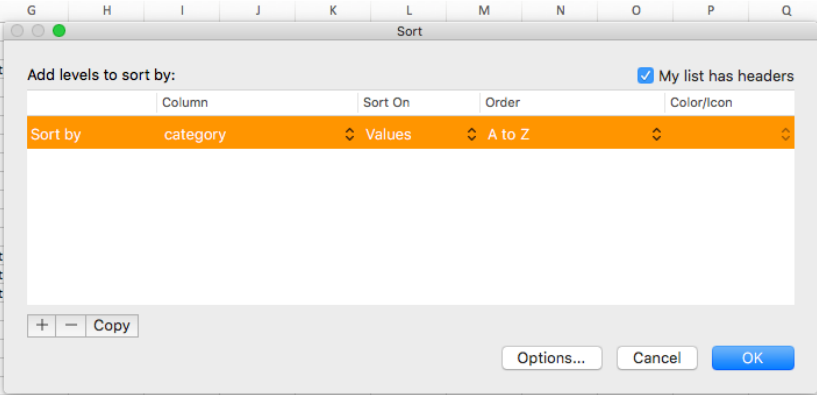
	A	B	C	D	E	F	G
1	date	amount	vendor	not sure	who	category	
2	4/8/17	$6.00	24 HOUR		2 Alice Smith	kid	
3	4/7/17	$5.40	85C BAKERY		2 Alice Smith	entertainment	
4	3/16/17	$5.96	99 RANCH		2 Alice Smith	food	
5	3/16/17	$8.30	99 RANCH		2 Alice Smith	food	
6	3/11/17	$100.00	99PLEDGCh		2 Bob Smith	charity	
7	4/7/17	$13.14	AMAZON		2 Bob Smith	find-detail	
8	3/21/17	$28.18	AMAZON		2 Bob Smith	find-detail	
9	4/6/17	$14.19	AMAZON.C		2 Bob Smith	find-detail	
10	4/6/17	$21.37	BEST VET		2 Bob Smith	pet-medical	
11	4/9/17	$2.71	BIG 5		2 Bob Smith	household	
12	4/9/17	$13.58	BIG 5		2 Bob Smith	household	
13	3/11/17	$3.27	CAFE		2 Bob Smith	entertainment	
14	3/11/17	$35.37	CAFE		2 Bob Smith	entertainment	
15	3/18/17	$40.29	CAFE DEUCE		2 Bob Smith	entertainment	

CS 46 W - Project Specification - CSV Processing Categorization

2.6 Sort by category and put a "total for category" column

see: costco-citi-visa-due-soon - bob copy - sorted - with categories added.csv

Sort by category and put a new column named "total category"

CS 46 W - Project Specification - CSV Processing Categorization

2.7 Add totals for each category and put it in "total for category column"

- See:
    - costco-citi-visa-due-soon - bob copy - sorted - with categories added.csv - how it looks on csv
    - costco-citi-visa-due-soon - bob copy - sorted - with categories added.xlsx - here totals are preserved (in terms of sum done via Excel formula)

Note: For this project you will **NOT** use or generate xlsx files. Any xlsx files in the box.com location are just there to show data in a formatted way (since csv files lose all Excel formatting).

CS 46 W - Project Specification - CSV Processing Categorization

	G28			$f x$	=SUM(B3:B28)		

	A	B	C	D	E	F	G	H
1	date	amount	vendor	not sure	who	category	total for category	
2	3/11/17	$100.00	99PLEDGCh	2	Bob Smith	charity	$100.00	
3	4/7/17	$5.40	85C BAKERY	2	Alice Smith	entertainment		
4	3/11/17	$3.27	CAFE	2	Bob Smith	entertainment		
5	3/11/17	$35.37	CAFE	2	Bob Smith	entertainment		
6	3/18/17	$40.29	CAFE DEUCE	2	Bob Smith	entertainment		
7	3/16/17	$18.87	DIM SUM	2	Bob Smith	entertainment		
8	4/10/17	$19.27	DREAMSCA	2	Bob Smith	entertainment		
9	3/12/17	$26.97	GHIRARDELL	2	Bob Smith	entertainment		
10	3/19/17	$18.00	GROUPON	2	Alice Smith	entertainment		
11	3/30/17	$4.89	IN N OUT	2	Alice Smith	entertainment		
12	3/23/17	$9.73	IN-N-OUT	2	Bob Smith	entertainment		
13	3/11/17	$2.66	IN-N-OUT	2	Bob Smith	entertainment		
14	3/11/17	$7.23	IN-N-OUT	2	Bob Smith	entertainment		
15	3/16/17	$10.00	KEE WAH	2	Alice Smith	entertainment		
16	4/2/17	$29.93	KOJA	2	Bob Smith	entertainment		
17	3/24/17	$9.99	Netflix.com	2	Bob Smith	entertainment		
18	3/11/17	$4.49	NOTHING	2	Bob Smith	entertainment		
19	4/8/17	$31.33	OLIVE	2	Bob Smith	entertainment		
20	4/2/17	$35.96	PHO	2	Bob Smith	entertainment		
21	4/9/17	$2.18	REDBOX	2	Bob Smith	entertainment		
22	3/19/17	$61.17	SOSO SUSHI	2	Bob Smith	entertainment		
23	4/1/17	$6.40	SQ	2	Bob Smith	entertainment		
24	3/19/17	$5.46	SQ	2	Bob Smith	entertainment		
25	4/9/17	$10.25	STARBUCKS	2	Bob Smith	entertainment		
26	4/1/17	$10.70	STARBUCKS	2	Bob Smith	entertainment		
27	3/11/17	$34.07	THAI	2	Bob Smith	entertainment		
28	4/9/17	$3.36	WENDYS #3	2	Alice Smith	entertainment	$447.24	
29	4/7/17	$13.14	AMAZON	2	Bob Smith	find-detail		
30	3/21/17	$28.18	AMAZON	2	Bob Smith	find-detail		

CS 46 W - Project Specification - CSV Processing Categorization

2.8 Transfer the total for each calculated category to this month's budget sheet for the credit card

- see:
    - budget monthly totals and tracking.csv: csv version
    - budget monthly totals and tracking.xlsx: xlsx version preservice formatting (like centering and such)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Which Card?	Month	entertainment	food	household	pet	pet-medical	kid	kid-medical	travel	utilities	gift	charity	find-detail	ignore	fixed-insurance
2																
3																
4	Costco Visa	"March 2017"	$447.24	$727.38	$137.96	$8.95	$21.37	$3,404.32			$59.99	$5.99	$100	$55.51	($1,234.78)	$305
5																
6																

3. Summary of project requirements

3.1 The Goal

The goal of this project is to automate Bob's manual process of budget categorization of the Costco credit card. So basically your program should accept "costco-citi-visa-due-soon - bob copy.csv" as input and ultimately generate "budget monthly totals and tracking.csv" as output. The "Details and suggestions..." section (below) will go over the approach.

Note: the file that your program accepts does not have to be exactly "costco-citi-visa-due-soon - bob copy.csv". Since this name has spaces - your file can be named without any spaces in it (the steps below will give you guidance as to the filenames you should use if filename with spaces are a problem).

CS 46 W - Project Specification - CSV Processing Categorization

3.2 Project Due Date, How To Submit, and Grading Process

Due Date:

- While I would prefer to get the project earlier rather than later. The final deadline for the project is as follows:
  - For online class: Two weeks past the last week of class on (at 9:00 pm Pacific). For example, if the the course runs from 6/26/17 to 8/18/17 - then the project would be due 2 weeks after 8/18/17 - so the project deadline would be 9/1/17 at 9:00 pm pacific.
  - **Anything handed after the above deadlines will be considered late and points will be taken away for this.**

How to Submit:

Please send me a **zip file (not .zipx or any other extension)** via messaging in Canvas email system or send directly to my email address. Contents of the zip file is described in the "What you need to hand-in" step (below).

Grading Process:

- There will be 2 versions of your program (sort of like 2 phases - details below)
- version-1 of your program will be %80 of your grade (this is the easier problem)
- version-2 of your program will be %20 of your grade
- **You need to complete version-1 before proceeding with version-2**

3.3 What you need to hand-in

Please put all of the following files in a folder that has your name on it (for example - if I was submitting the project my folder name would be EliLev) and then compress/zip the folder. Send me this zip file through Canvas or via email.

Your FirstnameLastname folder should have the following folders and files:

CS 46 W - Project Specification - CSV Processing Categorization

1. you should have 2 folders named (without quotes of course):
   - "version-1"
   - "version-2"

2. your code files:
   - You'll have two sets of code files in each of the above folders
   - All python files (it can be just 1 file or more than 1 file). I should be able to take this code and run it on the Bob csv file
     - If the csv filename has to to be a certain name - indicate this in a README.txt file
   - A file named "how_my_code_works.txt" where you explain exactly how your Python code works. You can do this in 2 ways:
     - Explain your code in essay format pointing out the functions and how the program works
     - Copy your code into this file and comment it heavily. Note that the code file that I use in testing your project should be free of comments
     - If you used code from the book - just mention which chapter (no extra detail needed)

3. README.txt: This file will explain how to run your program (it should be 1 or 2 sentences).
4. challenges.txt: a pure text file that describes all the problems that you encountered (no .docx, .rtf, or any other formats)
   - Describe what challenges you encountered?
   - If you had problems completing the project – record the issues that you had in detail.

5. code-from-web.txt: This text file is required **only if you used code from the web in your program**. Please make sure to:
   - Provide the URL of the source of the code
   - Paste the code in this document and explain what each line does (put a comment next to each line and write explanation)

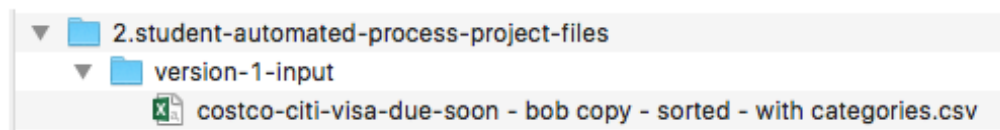## 3.4 What is in version 1 of your program? (%80 of your grade)

Version 1 of your program will use the input file(s) from the box folder named "version-1-input" that is in the "2.student-automated-process-project-files" folder.

CS 46 W - Project Specification - CSV Processing Categorization

- Your program will be given the "costco-citi-visa-due-soon - bob copy - sorted - with categories.csv" and it needs to:
  - Find totals for each category (remember that every row will have a category)
    - Note 1: please don't hard code the addition of the category based rows; the rows of "costco-citi-visa-due-soon - bob copy - sorted - with categories.csv" will be randomly sorted (in terms of row locations).
    - Note 2: you can create a file/files that your program will use as a place where it will store the mapping of vendor names and categories

  - Create a file named "output.version-1.budget-monthly-totals-and-tracking.csv" that will have the content shown in "budget monthly totals and tracking.csv" from above (in "Bob's current (manual) process of budget categorization" section)

Note 1: if the input file name is a problem because of spaces in the name - feel free to rename the "costco-citi-visa-due-soon - bob copy - sorted - with categories.csv" to "input.version-1.csv"

Note 2: In most programming (and in the case of this project) there is no "perfect program with the perfect answer". Focus on doing the best that you can with the stated problem and the knowledge that you have and iterate as you have more information in term of the problem statement and your knowledge.

```
▼  📁 2.student-automated-process-project-files
    ▼  📁 version-1-input
          📄 costco-citi-visa-due-soon - bob copy - sorted - with categories.csv
```

## 3.5 What is in version 2 of your program? (%20 of your grade)

Note: **You must complete version 1 of your program before proceeding with version 2.**

Version 2 of your program will use the input file(s) from the box folder named "version-2-input" that is in the "2.student-automated-process-project-files" folder.

CS 46 W - Project Specification - CSV Processing Categorization

- Your program will be given the "costco-citi-visa-due-soon - bob copy - sorted - with sparse categories.csv" and it needs to:
    ◦ Create a mapping of categories and vendors (since not all vendors will be categorized)
    ◦ Use the above mapping to put categories on rows that don't have a vendor (for a row that cannot be identified with a category, then your program should label the row with a category of "WARNING-NO-CATEGORY")
    ◦ Create a csv file (named output.version-2.intermediate-matched-vendors-sorted-categories.csv) that will have a row with vendor name and a category
    ◦ Find totals for each category that is found. If you used the "WARNING-NO-CATEGORY" category - include a total for it.
    ◦ Create a file named "output.version-2.budget-monthly-totals-and-tracking.csv" that will have the content shown in "budget monthly totals and tracking.csv" from above (in "Bob's current (manual) process of budget categorization" section)

Note 1: if the input file name is a problem because of spaces in the name - feel free to rename the "costco-citi-visa-due-soon - bob copy - sorted - with sparse categories.csv**" to "input.version-2.csv"**

```
▼ 📁 2.student-automated-process-project-files
    ▶ 📁 version-1-input
    ▼ 📁 version-2-input
            📗 costco-citi-visa-due-soon - bob copy - sorted - with sparse categories.csv
```
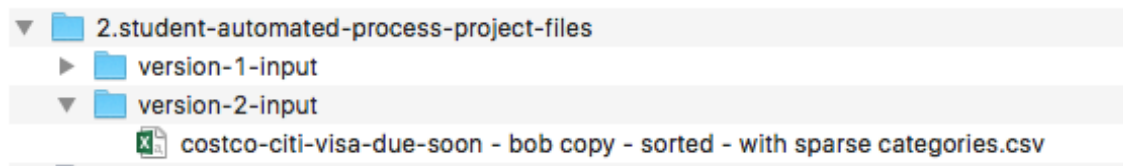
3.6 Resources that you may use

The following are the resources that you can use in creating your project. If you want to use something that's not listed here **please contact me**.

1. (most preferable) You may use any information and any code in our course's book ("Automate the Boring Stuff with Python" - 3rd edition). Please make sure to cite which chapters/code you used in your project's documentation (you don't have to explain this code - just cite it).
2. You may use https://www.draw.io/ for flowchart creation for the project (or any other program that generates a .png or .jpg output.
3. You may use content that you find on the web with the following stipulations:

CS 46 W - Project Specification - CSV Processing Categorization

- You must cite the source of any code or information that you get (i.e. URL for site). If the information is not publicly available through the web then you cannot use that source of information
- For any code that you use outside of the book's code, you need to provide a text file with the code that you copied and a line-by-line explanation of what the code does (see the "What you need to hand-in" step for details)

3.7 Extra credit

Before attempting any extra credit: please^100 do the basics of what I ask in the "What you need to hand-in" step from above for version 1 and version 2. Extra credit will augment what you hand in for version 1 and version 2 of your program, it wont substitute for it.

Possible extra credit:

- Flowchart for how your version 1 program works
- Flowchart for how your version 2 program works

4. A last reminder: Don't Panic!

If you've made it to this last section and you're hyper-ventilating I want you to consider a couple of things:

1. Take a [breath](#).
2. I know that this class is an addition to your many other responsibilities (work at home, work at work, life in general). My goal is to give you the best educational experience and you need to figure out how to best work on this project while dealing with life in general. Some ideas:

CS 46 W - Project Specification - CSV Processing Categorization

- Take "bites" of the project in small pieces. For example, rather than deal with the whole version-1 input csv file create a new file with a couple of rows. Work on getting this file to process correctly.
- Outside of the chapters that we cover in class, the other chapters in the book provide the information that you need. For those chapters, you can skim and quickly experiment. With the 'basics' background from the main class content - you should be able to do this.

3. If you're still panicking send me an email answering with the following statements filled out (we can brainstorm some things):
   - I'm panicked/worried because of _____
   - My current ideas on how I can deal with the above worry is _____

image credit: http://img.picturequotes.com/2/9/8712/dont-panic-quote-1.jpg