Utilizing Python to Analyze Data from US Accidents Report 2016 - 2021

Andrew Manz

Clemente Rodriguez

Scott Kurtz

Zachary Scholefield

Cal State, Bakersfield
Programming Languages
CMPS 3500
Prof. Walter Morales
Spring 2022

Introduction

Tasks:

- Read a CSV file
- Clean up the CSV file
- Create queries that answer the 10 assigned prompts
- These queries should be implemented in Python
- Create a menu that allows for user input
- Implement error handling for mistakes that a user could make

Approach

Initial Approach:

- Split teams between Ruby and Python
- Implement Daru and Pandas respectively

Approach post Ruby:

- Reformatted prompts into functions
- Enable search capability
- Created menu in if / else block
- Error handling
- Input validation

Structure

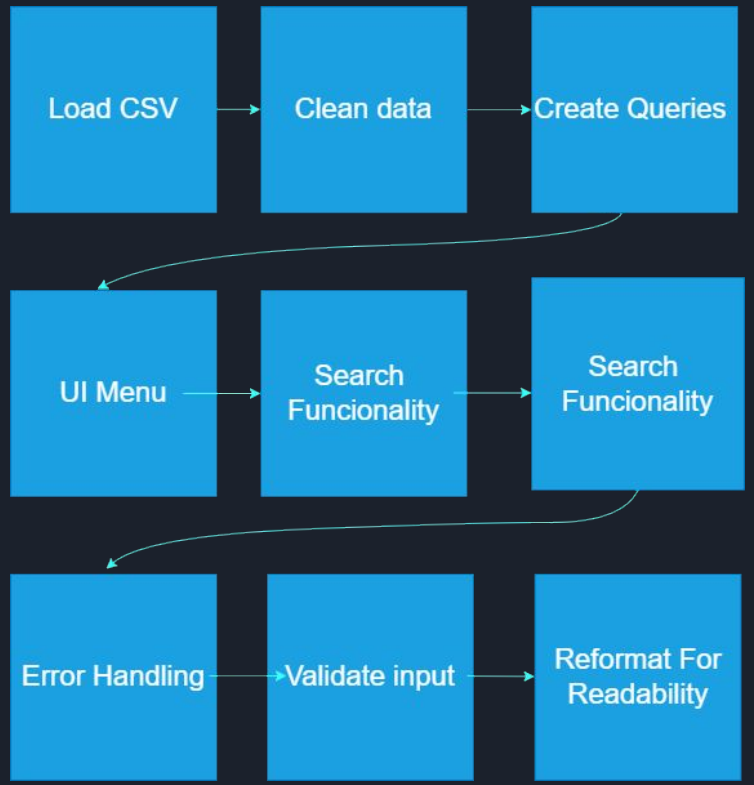
Modularity via functions

Use of methods: value_counts, type, to_list, length, if/conditionals

Try / except blocks

UI Menu

Input validation



Implementation

```
Welcome to a data processing application to find records of
accidents that occurred in the U.S. from 2016 to 2021.

You will need to (1) load the information first and then
(2) Process the data before you can search.
************************************************************************

************************************************************************
*                              MENU                                    *
************************************************************************
1: Load  the data
2: Process the data
3: Print the answers to the questions
4: Search accidents by place (city, state, zip)
5: Search accidents by time(year, month, day)
6: Search accidents by conditions (temperature range and visibility range)
7: Quit
```

Loading Data:

```
Loading input data set:
*******************************************
[ 2022-05-12 20:31:27.896124 ] Starting Script
[ 2022-05-12 20:31:31.716794 ] Loading US_Accidents_data.csv
[ 2022-05-12 20:31:31.716794 ] Total Columns Read: 22
[ 2022-05-12 20:31:31.717294 ] Total Rows Read: 711335
Time to load is: 3.8217 seconds
```

Processing Data:

```
Processing and cleaning input data set:
*****************************************************
[ 2022-05-12 20:32:27.263877 ] Performing Data Cleanup
[ 2022-05-12 20:32:30.789991 ] Total Rows Read after cleaning is: 596147
Time to process is: 3.5261 seconds
```

Error handling:

```
!!! PLEASE LOAD AND PROCESS DATA BEFORE LOADING ANSWERS TO PROMPTS!!!


************************************************************************
*                              MENU                                    *
************************************************************************
```

Implementation

Beginning of prompt list:

```
Prompt 1:
[ 2022-05-12 20:36:04.171889 ] In what month were there more accidents reported?
[ 2022-05-12 20:36:04.175390 ] December

Prompt 2
[ 2022-05-12 20:36:04.175890 ] What is the state that had the most accidents in 2020?
[ 2022-05-12 20:36:04.233400 ] ['CA']
```

Temp and visibility search:

```
Input the lowest temperature of the range in ℉: 25
Input the highest temperature of the range  in ℉: 30
Input the lowest visibility of the range (0-10) in miles: .1
Input the farthest visibility of the range (0-10) in miles: 6

The number of accidents with temperature between 25.0 ℉ and  30.0 ℉
and visibility between 0.1 mi  and  6.0 mi is:
 5771
```

Time search:

```
You will be given the opportunity to search by month, day, and year.
If you only want to limit your search by one or two factors,
Type: NA, when given those options.
************************************************************

Please type the year between 2016 and 2021: 2016
Please type the month as integer (Jan is 1,Feb is 2, etc.): 12
Please type the day: 5
The number of accidents on 12/5/2016 is:
 183
```

Location search:

```
You will be given the opportunity to search by city, state, or zip code.
************************************************************

Please type the name of the city you would like to search.
LaS vEgAs

The number of accidents in Las Vegas was:
 345

Please type the name of the state you would like to search.
Format: CA, NV, WA, etc.
Nv
The number of accidents in NV was:
 1376

Please type the zip code you would like to search.
Format: 12345
89074
The number of accidents in 89074 was:
 7

Time for location search :  0.077
```

Conclusions

Pandas is the best

We learned how to manage a large-scale project

We took raw data and were able to extrapolate meaningful information

Data analysis is a very powerful tool when working with large datasets

Teamwork makes the dream work