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

.

Group 9
Andrew Manz
Scott Kurtz
Clemente Rodriguez
sfasds`Zachary Scholefield

.

Final Project- CMPS 3390
Spring 2022
California State University, Bakersfield

Introduction

This project was the ultimate test of all of our skills we have learned in our collective years of programming. Our program incorporated Pandas to help analyze the accident data from the extremely large csv file. We learned how to clean datasets based on certain conditions, and we learned a lot about error handling. Getting familiar with Pandas with Python to compile data in various forms was an interesting, challenging, and sometimes frustrating task. It definitely seems that manipulating data in this format is much faster than sifting through a regular database with MYSQL. As a team, we were able to efficiently put together a project we can all be proud of.

Approach

Our group initially assigned two team members to work on Python & Pandas and the other two to work on Ruby & DARU. The main goal at this time was to figure out the syntax to answer the ten original questions. The Ruby team struggled and fought to accomplish what they could; however Daru is not well maintained and it was an added stress to all our lives. All-the-while the Python team were able to make much more progress due to the use of a current and maintained tool like Pandas; they were able answer a majority of the questions before the Ruby requirement was lifted. Once we dropped the Ruby portion, the Ruby team assisted in finishing the questions and started developing the menu, formatting into functions, and error-handling.

The team members who worked on Python started off by first familiarizing themselves with the pandas library and getting the CSV read. Afterwards the missing information and empty slots were removed from the CSV in order to clean it up. From there, the queries were created in order to answer the ten prompts. One of the team members developed the new list of choices while implementing error handling. Another developer worked on the searching algorithms. Everyone worked together to finish the queries on the searching algorithms and cleaning the code as well as implementing the timer. The main focus was to communicate and let the other team members know what we were working on throughout the duration of the project. We all have strong communications skills and if help was needed, we all jumped in and lent a hand.

Structure

The program is structured in a monolithic format, using methods to implement most functionality. The menu is implemented using a series of if statements, with an error catch for inputs that do not match any of the options presented. Our code starts with a brief introduction letting the user know that they will need to load and process the data first using options 1 and 2 respectively. After that the user can choose 3, 4,5,6, or 7. Choice 3 prints the results of the ten original questions for the assignment. To incorporate this function, we used a function that just called all 10 of the prompts, 1 for each question.

The fourth selection allows the user to select a city, state and zip code to find out the number of accidents in a particular location. Along with getting the data from the dataframe, we used functions like .title( )to capitalize the first letter of each word for input with multiple words(e.g. "City of Industry"). Option 5 allows the user to search by year, month, and day. We originally developed both option 4 and 5 to have one function calling each individual option but after re-reading the instructions, we realized we needed to change our implementation. We then formatted the output to print a generic timeframe if one of the selections was 'NA'. Otherwise, it would print in a typical date format (e.g. mm/dd/yyy). We may have been able to use the separate function concept but we realized that trying to keep track of the elapsed time while jumping through functions would be a little clunky. The sixth option has the user make a selection through a temperature range and a visibility range. Thankfully, we realized that the one function would suffice for this before trying to separate the functions only to bring them back together again. These values were then reformatted for easier readability and gave the units for clarity.

We finished our implementation by cleaning up the code and implementing the timing system. The timing system was needed to keep track of each of the individual searches. Additionally, we used a main variable that added all of the individual times together to come up with a total time. This total time is displayed after the user types '7' to exit the application. Once the functionality was on point, we then made sure all error handling was sound with if statements. Inside the menu, we ensured the user couldn't process the data without it being loaded first, run the prompts with the answers until loading and processing were performed already and so on. We did this by checking to see if the load time was zero; if so, it would throw an error and reload the menu to try again. We continued this logic throughout the menu options.

Answers

1. In what month were there more accidents reported?

- The month with the most accidents is: 12

2. What is the state that had the most accidents in 2020?

- The state with the most accidents in the year 2020 is: ['CA']

3. What is the state that had the most accidents of severity 2 in 2021?

- The state with the most accidents of severity 2 in 2021 is: ['CA']

4. What severity is the most common in Virginia?

- The most common severity in Virginia is: [2]

5. What are the 5 cities that had the most accidents in 2019 in California?

- The 5 cities in California that had the most accidents in 2019 are:

Los Angeles	449
Sacramento	178
San Diego	178
San Jose	134

Oakland	121

6. What was the average humidity and average temperature of all accidents of severity 4 that occurred in 2021?

- For accidents with severity 4 that occured in 2021:

  Average Humidity: 66.95750708215297

  Average Temperature: 57.804559109383284

7. What are the 3 most common weather conditions (weather_conditions) when accidents occurred?

- The 3 most common weather conditions are:

Fair	237116
Mostly Cloudy	78361
Cloudy	73463

8. What was the maximum visibility of all accidents of severity 2 that occurred in the state of New Hampshire?

- The maximum visibility of all accidents of severity 2 that occurred in the state of New Hampshire: 130.0

9. How many accidents of each severity were recorded in Bakersfield?

- Accidents in Bakersfield with Severity 1: 0

  Accidents in Bakersfield with Severity 2: 2250

  Accidents in Bakersfield with Severity 3: 26

  Accidents in Bakersfield with Severity 4: 20

10. What was the longest accident (in hours) recorded in Florida in the Spring (March, April, and May) of 2020?

- 95.99 Hours.

Conclusion

For people who enjoy mathematics and data analysis, this was an interesting assignment. It appears that using Pandas is a great way to find out answers to questions regarding large data sets. Our team learned the importance of planning a project out before implementing the design concept. This was especially evident with the searching choices 4 and 5. A lot of time was wasted not thinking through the functions thoroughly before starting to type. We were able to overcome so much this semester, and we are extremely proud of what we have accomplished and look forward to perhaps using these skills in our future careers.