CSUB

Class Project

Group 7

Hunberto Pascual

Kenneth Wood

Nathan Wardinsky

CMPS: 3500

Professor Morales

* Introduction

ClassProjectGroup7.py is a python implementation of basic data analysis, allowing users to load data from a CSV file into a data frame. Users then can explore, describe, and analyze said data. Exploring data allows for printing columns and their name, dropping any column, counting distinct values, searching for any value, and sorting columns. Describe data does mathematical calculations on a specified column, which includes the mean, median, mode, etc. Analysis answers questions such as how many airlines are in the dataset. This project taught us how to remotely work cooperatively with others and deal the unexpected, like losing a team member. Since this was a three-member team we had to put in quite a bit more work than we initially expected, juggling this project along with our classes and assignments was the biggest struggle.

* How the group approached the project

Our first step as a group was to connect on discord and set up a GitHub repository so that we could work cooperatively. The first goal was to write code that would read CSV files and load its data into a data frame. After that, we implemented the first couple of tasks so everyone could see how data frames work and have a base to build off. We pushed our code onto a GitHub repository and scheduled meetings via discord to plan and check our progress. Since we were working remotely off a GitHub repository, all coding was done on our personal computers, and the final version of our code was copied onto Odin.

Once the foundation was set, we split up the tasks; one person focused on creating the general structure for our program along with a menu interface, while the others worked on the describing data section. Once the general structure was built along with a working menu interface, the rest was smooth sailing; everyone added in their own methods to the appropriate Class. After the completion of Exploring Data and near completion of Describing Data, we shifted our focus on the analysis questions, implementing error handling and run-time measuring. Lastly, in the final week before the due date, we wrote this Report, discussed our presentation, and tied up any loose ends and bugs.

* Explain how data loading, data cleaning, and search capabilities were implemented

Data loading from CSV files was handled using data frames and lists to isolate specific columns when convenient. The loadFile() function prompts the user for the name of a file to load and, if successfully found loads it into a global data frame accessible to all methods. An error message is displayed if the file is not found, and the user is asked to try again. For data cleaning, we removed the index column and duplicate rows were dropped. In the case of the analysis questions, we had to filter out some duplicate values by creating a list of all unique values and filtering duplicates.

Search capabilities were implemented using data frames and lists; for example, the searchColumn() method, which finds a value in a column as specified by the user, first gets the column and value from the user, stores them as capitalized strings. Then using that input locates the matching column in the data frame and stores its data as a list of capitalized strings. Doing so allows our methods to search for any number or string without worrying about data type or capitalization. Then it’s a simple for loop iterating through the list to find the specified value.

* Explain the structure of your code, how Error Handling was implemented, and how variable naming and comments were managed in your code to improve its readability

One of our main priorities was getting the general structure of our program done as early as possible so that we wouldn’t have to do any major revisions later. Our program follows an Object-Oriented approach; we have three main classes, ExploreData, DescribeData, and Analysis which house all corresponding methods. Each Class also has its sub-menu, exporeDataMenu(), describeDataMenu(), and analysisMenu(), each of which displays a menu and handles all looping for its respective Class. Outside of the classes, we also have two other general functions, mainMenu() and loadData(); since these two could be implemented as singular functions, we did not create a class for them.

MainMenu() is the main control of the program; it allows the user to select data Loading, ExploreData, DescribeData, and Analysis and then calls the corresponding class submenu. The submenu then prints all available operations, including the option of returning to the mainMenu, and calls the necessary method as requested by the user. Error handling was implemented using try and except statements and general clever coding. Errors that could be foreseen were avoided using if and else loops to handle user input so that regardless of what the user enters, no error would occur. For less apparent errors, we purposely set out to cause them and implemented the corresponding except statements; some methods required their own error handling so the looping of our program would work properly and print helpful error messages in the correct location.

As a group, we aimed to make our program easy to read by naming everything as descriptively as possible and including comments where necessary. For example, col\_to\_describe is the name of the column that will be sent to the DescribeData methods. The analysis portion of our program contains quite a bit of comments due to the complexity of some of the questions. Our goal was to write reliable software and I think we accomplished that wonderfully.

* Answers to questions

1. How many airlines are included in the data set? Print the first 5 in alphabetical order.

17 Airlines in the data set.

Five first airlines in alphabetical order:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Alaska Airlines Inc.

Allegiant Air

American Airlines Inc.

American Eagle Airlines Inc.

Atlantic Southeast Airlines

1. How many departing airports are included in the data set? Print the last 5 in alphabetical order.

96 departing airports in the data set.

Last five departing airports in alphabetical order:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Tucson International

Tulsa International

Washington Dulles International

Will Rogers World

William P Hobby

1. What airline has the oldest plane? Print the five airlines with the oldest planes recorded

Delta Air Lines Inc. was the airline with oldest plane at: 32 years

Top five airlines with oldest planes:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Delta Air Lines Inc.

United Air Lines Inc.

American Airlines Inc.

Southwest Airlines Co.

SkyWest Airlines Inc.

1. What is the airport that averaged the greatest number of passengers recorded in 2019? Print the 5 airport that averaged the greatest number of passengers in 2019.

Atlanta Municipal airport had the highest average passengers at 4365661 passengers

Top five airports with highest averaged passengers:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Atlanta Municipal

Chicago O'Hare International

Dallas Fort Worth Regional

Los Angeles International

Stapleton International

1. What is the airline that averaged the greatest number of employees (Flight attendants and ground service) in 2019? Print the 5 airlines that averaged the greatest number of employees in 2019.

American Eagle Airlines Inc. was the airline with the most average employees at 0.0003484

Top five airlines with highest averaged employees:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

American Eagle Airlines Inc.

United Air Lines Inc.

JetBlue Airways

Delta Air Lines Inc.

Hawaiian Airlines Inc.

1. What was the month of the year in 2019 with most delays overall? And how many delays were recorded in that month?

June had the most delays in 2019 with a total count of: 7219 delays

1. What was the day in 2019 with most delays overall? And how many delays were recorded in that day?

Thursday had the most delays in 2019 with a total count of: 10429 delays

1. What airline carrier experience the most delays in January, July, and December

Southwest Airlines Co. had the most delays in January with a total count of 949

Southwest Airlines Co. had the most delays in July with a total count of 1345

Southwest Airlines Co. had the most delays in December with a total count of 1575

1. What was the average plane age of all planes with delays operated by American Airlines inc.

The average plane age of all planes with delays operated by American Airlines is: 11.098

1. How many planes were delayed for more than 15 minutes during days with heavy snow > 15 inches

0 planes were delayed for more than 15mins due to heavy snow of 15in or greater.

1. What are the 5 Airports that had the most delays in 2019?

Top 5 airports with the most delays in 2019:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Chicago O'Hare International with 3913 delays

Dallas Fort Worth Regional with 3567 delays

Atlanta Municipal with 3527 delays

Stapleton International with 2901 delays

Douglas Municipal with 2344 delays

* Conclusion

For a team of 3 people our project operates quite well, progress was steady throughout the duration of our project and thanks to the early structural creation of our program and everyone’s good work ethic we avoided any major hurdles. We adopted a primarily object-oriented approach to our python implementation of file data analysis, utilizing classes, methods, and functions to load a CSV file is into a data frame which can then be modified, explored, described, and analyzed by users.