Data Analyzer

Ahmed Arif Khan
1324501
Computer Science
2XB3 L01

Abstract:

There is a lot of open data online but it gets hard to manage all that information. This application will solve this problem by taking the information and giving you powerful tools to visualize it. It will also find correlations with the inputted data and variables such as age, gender, time, and if it finds a strong correlation then it displays it to the user and then the user can analyze why that is happening. When it displays the information to the user it will be in a graphical format to help users see how each piece of information relates to one another. To test our application we are going to use sample data that we can mathematically check if the correlation exists or not. We will need to check every piece of our application with sample data and also give it normal data and manually check if it correct or not.

Objectives and scope of the project:

My friends and me were trying to use the open data available to make a program that will help find good places to buy houses but using the open data was overwhelming and hard to manage. It required a lot of time and effort just to find the useful information. There is a vast amount of open data publicly available on the Internet for everyone to use except the problem is that it is inconvenient to find and work with. First you need to find the information that you will be using from many different sources. Gather it all up and fix the small changes for example standardizing the date format, or fixing the formats for all the locations. After collecting a large amount of information it gets very complex to help manage the data because you have around thousands of pieces of information and it gets hard to see how they relate to each other. This application will fix this problem by gathering all the information in one place and giving the user powerful visual tools to help them see the information in a more manageable way that will help the users of this application find solutions to real world problems. Before we find solutions or make any difference we need to know what problems people are facing. With this we can accurately find which group of people are suffering with which problems. For example: Lets say the program looks at all the motor vehicle accident rate with respect to gender, age, and so on. Then it concludes to saying that there is a strong correlation between the age group of 17-20 and motor vehicle accident. Then users can go and figure out why this is happening and find solutions. In every day life people usually come up with a hypothesis of certain problems existing and then test for them. That is a valid strategy to help check if your hypothesis was correct but it leads to confirmation bias. Our program will fix this by giving the users sets of graphs that already have a strong correlation so that people can already see the problem they might have not notices before. This is a very important problem to solve in today's day and age because technology gives us a vast amount of information, which we cannot use effectively. It becomes hard to manage. With more people will be aware of the problems that happen in our world. Solutions can only occur if people are aware of the problems around them. There are going to be times where a similar solution can apply to many different situations and if the average user could see how other solutions are made they could apply that into that situation. All in all the most effective way to make the world better is to be aware of

the problems in it. This software will allow us to live in a world where people are aware.

Input/output:

The input will be large statistical information of Canada such as crime rate, income, and health. This information will be retrieved from Canada's open data websites. When the user goes into our application there will be two main features, explore and analyze. In analyze the user can pick any number of data sets such as hand gun violence and graph it with respect to time, age, gender, geographical location and so on. The other feature will be called explore. In this the computer will chose any topic. It will find the distribution of the topic by graphing it with respect to all the groups such as gender, age, and wealth. It will then find the correlation between that specific information and group and when the correlation becomes to high it displays it to the user in the respective graphical format. The program could graph the information in any desired chart for example pie graph, bar graph, line graph and so on. This will allow the computer to do the mindless information management so the user can go and analyze it and figure out why these trends exist and help prevent them for the future.

Algorithmic challenges of the project:

When we have to visualize the graph it will need to be able to change from the different types of charts with ease. To do this we will we will need to store the data in a BST to help us be able to switch to different visual formats.
Then it will need to graph the data and find its correlation with respect to many different variables. For the program it would need to be able to solve this problem in a very fast manor because whatever the algorithm complexity is, it will need to recursively do it a large number of times.
This program will be taking information from many different databases and combine them. When you combine different pieces of information together you arise at problems where the program does not know things like a latitude and longitude coordinate and the location name wont be equivalent. So we are going to need to make a searching algorithm that converts all of the GPS coordinates to the standard location format.

Project timetable:

There is a grant chart at the end of this that shows the 5 dates in their respective order. This is just elaborating on what the chart means by each row has.

General Layout: Make a more defined input and output diagram. Make a sample diagram of how the view is going to look on paper. Make a general layout of the amount of classes and how they are going to interact with each other, make the test files for all the different parts of the program, build sample data to test with.
Build The View: Make the view, meaning the code that displays the menu the different graphs. In the current phase of development it will just display sample data.

Build The Model: This will hold the logic of the program. This will be making the part that graphs the data in many different ways and finds the correlation between two variables until it finds two variables that have a strong correlation.
Make the controller: Connects everything in the program together. It just serves as a link between the user input, the model, and the view.
Finish Up: Make the presentation, combine all the stuff together, and now thoroughly test every aspect of the program.

• **References**:

- So far I did not use any references to the Internet. I am going to be using a lot of data from the Internet but I don't know which ones yet.