Badjessa B. Bahoumda

CSE 12

April 2, 2016

Analysis of the cancer dataset

The cancer dataset used in my project is an abstract of all the instances of cancer cases recorded by the Center for Disease Control and Prevention (CDC). This data is called the United States Cancer Statistics and it represents the amount of cancer cases by area, by sex, by age group and also contains information about the mortality and the causal site. This data is a compilation of cancer cases and analysis of the data set aims to help understand it better.

The questions that are answered in my project are as follow:

* What is/are the most common site(s) of cancer between the brain, the female breast and the corpus?
* Is there a correlation between the average age and the count of cancer instances?
* Is cancer more common in female or male or both?
* Which has a higher concentration of site between Delaware, New Jersey, North Carolina?

The first question give insight into two of most encountered cancer in the world and the third one that may not be necessarily common knowledge. It allows us to understand which is more likely to be found and thus gives an insight into the reason why the corpus cancer might not be known by our society while brain and breast cancer are widely recognized.

The second question was simply meant to see whether cancer was caused at random or was most likely to appear in children than adult. This also is important when monitoring individuals for different cases, it allows us to identify which age group is most likely to be considered at risk.

The third question gives insight into whether cancer would most likely found in either female or males. This also allows close monitoring of either sex if it was to say that one was most likely to have cancer than the other.

In regard to the fourth question, by figuring out whether there was a higher concentration in one area, we are able to look at specific aspects of that area which could lead to figuring out external factors in the causation of cancer.

There some limitations to understanding the dataset in depth. The first one I encountered was the age grouping. This made it harder to write a code allowing to configure a correlation or relationship between sites and age. There is also a lack of detail, each list in the data set gives some general information about the disease and where it is contracted but does not really gives any significant detail thus does not allow for further research.

There is a Visual illustration in the project and it is showing a correlation between the average age and the counts of cancer mortality.

Program Development:

The code starts by importing the cancer dataset. Next all the variables are defined and in order to answer the first question, we list the site and for each site we record the number of Brain site, the number of Female breast site and the corpus site. Then by comparing all these numbers together, we find the most common cancer site. We then attempt to answer the third question by recording the sex for each site and list of site earlier. Similarly, to the first question we compare the values again and print the highest value. In order to answer the fourth question, we create variable that have for values the list of cancer instances by each of the three area. Then for each individual area we record the number of sites, and to finally answer we compare the values and print the highest value. In order to answer the second question, we use a visual representation, a scatter plot to be precise. We make a list containing and all the averages and all the counts and then the code computed create a scatter plot of the average age by the count.

These conclusions affect the population in the areas recorded, the cancer researchers at the CDC. This data affects the people in the areas where the cases of cancer of recorded because they could be part of the data, it could be somebody in their families or a close friend. It also allows to tell whether a region should be considered at risk. It allows the population to make decision about how to live their lives if they were in such a region. Furthermore, it also affects the researchers at the CDC because analysis of this data is a step toward finding out causes of the different type of cancers and their concentrations in different regions. This allows for further research regarding cures and also protection measures.