Badjessa B. Bahoumda

CSE-12

Project 2 Write Up.

***Abstract:***

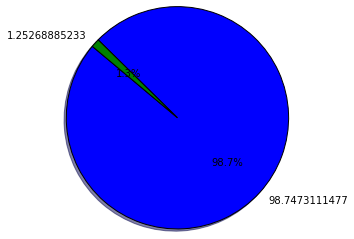
The data used for my second project was the college scoreboard dataset. This College scorecard dataset is an abstract of information regarding a sample of college and Universities in the United States. College Scorecards make it easier for students to search for a college that is a good fit for them. They can use the College Scorecard to find out more about a college's affordability and value so they can make more informed decisions about which college to attend.

***Questions/Conclusions(Visualizations):***

I used three questions accompanied by visualizations when analyzing my dataset. The question and the results from my query as follow:

1. List the Historically Black College and Universities, their price and show the percentage they make up of the schools in the dataset.
   * Visualization: Pie Chart

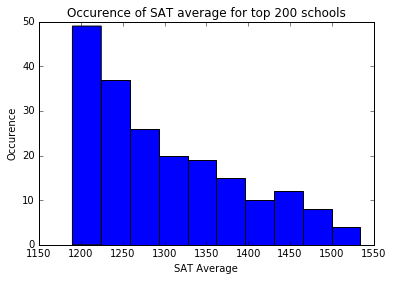
According to the query, there were 99 schools in the College Scoreboard that were Historically Black colleges and Universities (HBCU). The Price range varied from $1100 to over $20000. However, it was interesting to find that although 99 may seem like a big number they only made up about 1.3% of the total colleges and Universities in the dataset. The visualization looked like this:



The green represents the HBCU’s in the dataset and the blue represent the rest of the schools and as stated earlier HBCU’s only make up 1.3% of the dataset. This result was not something surprising but it was something that I felt was important to look at when looking at colleges because it was something that I considered crucial when I was looking at various schools.

1. What are the top 200 schools with the highest SAT Averages? out of the list, which SAT Averages occurred more often and which did not?
   * Visualization: Histogram

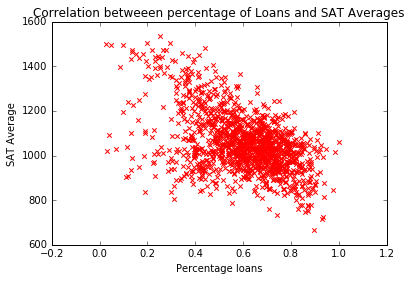
The top 200 SAT scores Averages ranged from 1190 to 1534. Some of the schools with the high SAT Averages were Princeton University, Brown University, Harvard University and the California Institute of Technology. Lower Averages schools were Belmont University, University of California, University of Colorado Boulder. When looking at the occurrences of the SAT Averages, we could see that the low SAT scores occurred more often than the High SAT averages. The histogram looked like this:



The result was a bit surprising to me because I assumed that most top schools would have really high SAT Averages and the histogram proved that was not true. It was significant because it shows that SAT scores are not the only that are considered of an applicant and is a significant boost for students who feel not qualified to apply to some schools.

1. Is there a correlation between the SAT Averages and the percentage loans out of a sample of 1400 schools?
   * Visualization: Scatter Plot

The third Correlation asked to show whether or not there was a correlation between the SAT Averages and the percentage loans in the schools in the dataset. In order to show correlation, we have to create a scatter plot. Our plot visualization had this form:



The correlation that we recognized was that the higher the SAT score the lower the percentage. This could be translated into the fact that schools offer bigger financial aid to student with high SAT Scores.

***Program Development:***

I started my code by importing both sqlite3 and matplotlib in python. Sqlite3 will be used to analyze the dataset and matplotlib is essential to create visualizations in our code. Next I created connections between sqliet3 and the College Scoreboard dataset that I will be using for my project. I then created a cursor to be able to get result and then started answering the questions regarding the dataset. For the first question, I executed the query which allowed me to pull the name and the net price of all schools that we HBCU, however it was important to create two queries for this task because some of the schools were private and somewhere public. The first query added the where clause in which the schools are public or (=1), the second one was for schools that were private whether for profit or nonprofit (>1). The next query counted the number of schools that were HBCU’s and the one afterwards counted all of the schools in the dataset. This is important in order to calculate the proportions and create a pie chart showing the percentage of HBCU’s in the entire dataset. We then integrate both result into a list and then calculate the proportion of each to the total sum of each. Next we create pie chart from the matplotlib imported using the plt.pie function.

To answer the second question, I created a query which returns the institution names and the SAT Averages that are of the highest value in the dataset. I then created a list of all the SAT Averages, it is important to note that I sorted the list so that the value was listed from the smallest value to the biggest. This is of importance because a histogram needs a sequence in the values. Next, using the plt.hist function I plot a histogram of the SAT Averages and I can see how often an SAT Average occurs in the dataset.

Next, to answer the third question, I made a query which returned the names of colleges, their percentage loans and their SAT Averages and set the limit of the values given at 1400. I then created a list of all the percentage loans and another one containing the SAT Averages. Finally, I created the scatter plot using plt.scatter plot.

***Limitations:***

There was not much information in the dataset regarding ACT averages. The dataset also did not have any information regarding about the type of degrees offered in the schools or anything regarding the type of sports or activities. I felt these types of information play a role in the decision to attend a school.

***Social Impact:***

The conclusions drawn from my analysis affects students and schools. They affect students because it gives them a clue in their search for a school that would be a good fit for them. The analysis allows the student to be able to filter the school based on the type of schools they want, which SAT scores they have and also on potential financial aid that could be given to them. The conclusions affect the schools because it allows them to look at their ranking compared to other schools, see the average performance of incoming freshmen at their school and an overview of the financial state of student attending their schools.

***Reflections:***

The first problem, I encountered was making a query with data that was in Unicode, but after emailing my professor, I was able to resolve it. The second issue, I encountered had to do with not receiving null as a result of the query, I found it hard to resolve but through a method of trial-and-error, I was able to solve it. I also was not able to list the Universities like I originally wanted to since the console did not show all of the data, but I was able to fix that by showing the visualization?

***Acknowledgements:***

Professor Eric Fouh.