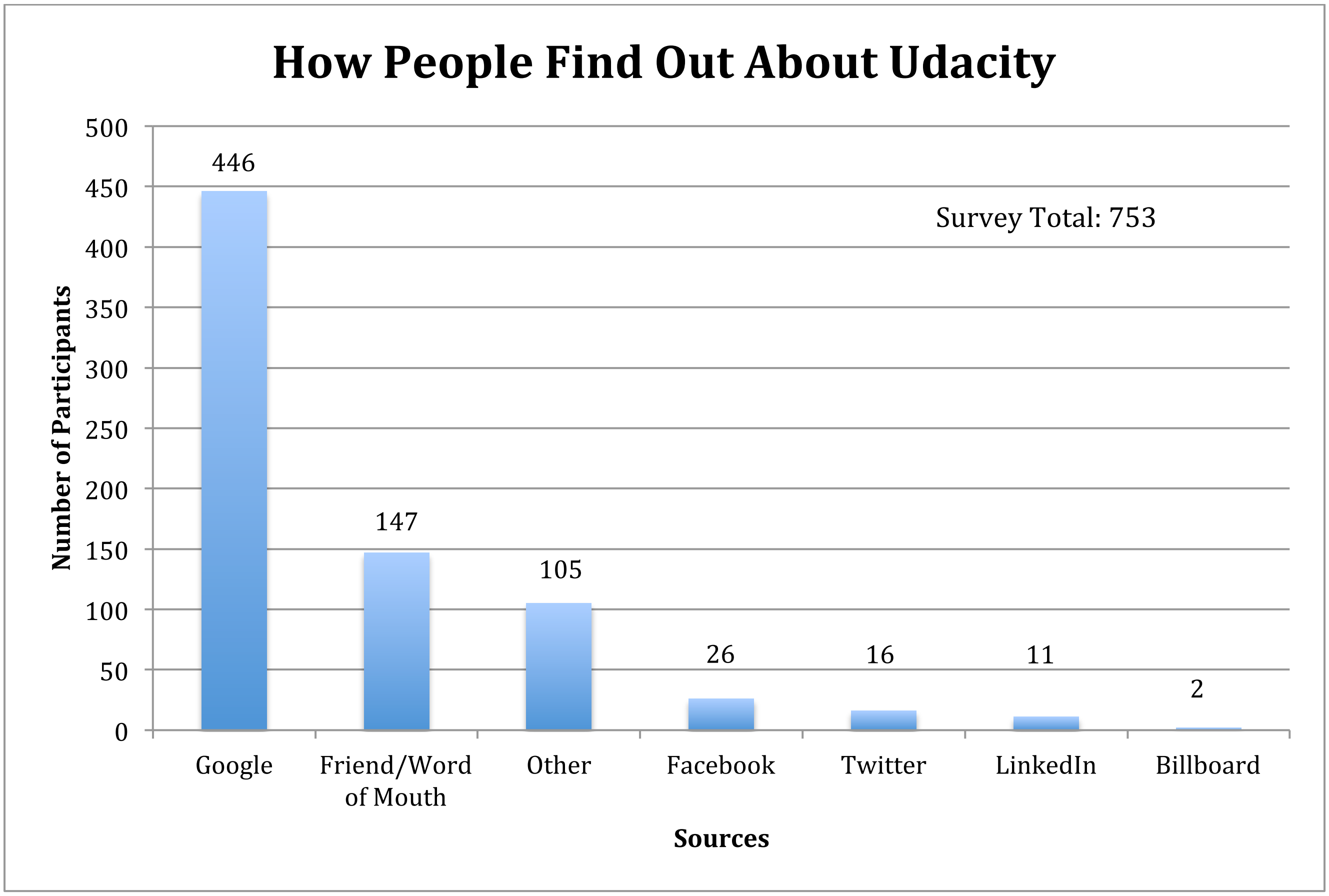
Chi Wing (Winnie) Ng

******One thing to note about this analysis project: The data provided for this project are from Survey respondents and not from the entirety Udacity Student Population**

Question 1: How do people know about Udacity?

Figure 1: A Visualization of Question 1



<u>**Conclusion:**</u> From this chart, we can see that students find out about Udacity through Google searches. From a statistical point of view, they make up about 60% of the overall survey. It should be stated that "Other" as a response has many written responses that could have been under Friend/ Word of Mouth such as "School Reviews", "AI class", "Sebastian keynote @ IBM World of Watson". Regardless, it would not have affected the result that Google is the highest search source of where participants find Udacity.

Question 2: How many hours per week do students spend on learning materials? Is there a difference in hours between employed and unemployed?

	Min	Lower Quartile	Upper Quartile	Max
Unemployed	1	5	10	40
Employed	1	4	6	80

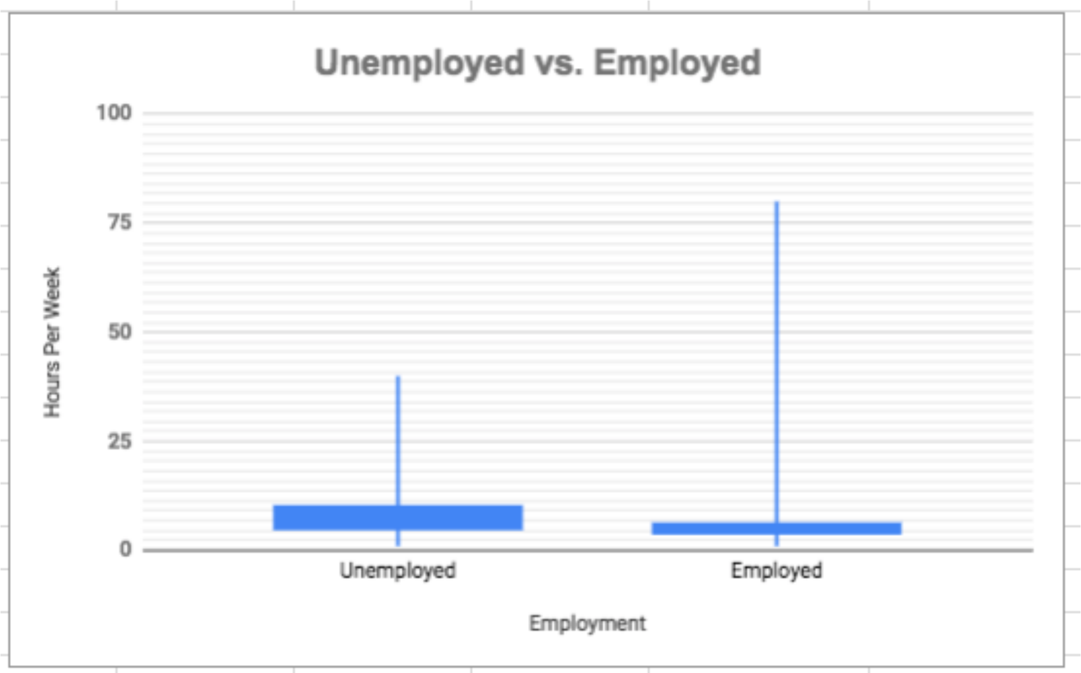


Figure 2a: Boxplot that shows a visual results between unemployed and employed

For this problem, I thought it would be better to show a box plot because there are participants who chose “Other” as a response and gave their own number that wasn’t within the multiple choices. Also the number of participants who were unemployed is fewer than the numbers employed.

Other things to note when analyzing the data:

- I omitted data with answers “10+”, “More than 10”, and any answer than either doesn’t give a number or a range as they are not specific quantitative responses and can affect the data. The answer is too vague to give an accurate representation. We could put in “10 hours” for that kind of response but the quantitative range is too high (1-80) and will not give an accurate representation of the data.
- If the response was a range, I opt for the minimum of the range (For example, if it was a response, “20-30 hours”, I used the 20 hours in the range. I chose to utilize the range response this way because we can assume that the responders did at least 20 hours)

As a result, instead of 753 participants, there are a total of 693 participants in this data.

Conclusion: From the boxplot in figure 2a, we can see that those employed spends longer time on the learning materials than those unemployed. For those unemployed, the response was maximum 40 hours while employed was 80 hours, twice the amount of time.

There is a large spread of data given that the range is 79. With a large outlier, the range is not a very good estimate of how the data behaves and is even more irrelevant in this case given that it only takes two points into considerations.

However, this result has other factors that were not taken into account for: the different Nanodegree programs, whether or not students had prior background before applying for the program.

Question 3: What is the average number of books read per year? How does it look in a normal distribution?

Table 3.1

Mean	13.48335553
Median	8
Mode	10
St. Dev	28.87447643

From Table 3.1, the average number of books read per year is about 13 books. The median number of books read per year is 8, which means half of the survey reads less than or equal to 8 books per year and the other half of the survey read more than or equal to 8 books per year. With a mode equal to 10, this means that most people read about 10 books per year. Seeing that the median and mode are less than the mean, it seems more than half of the responders read less than the average number of books per year.

Another thing to note is that it seems than the mean is higher than the median, which can be said that the curve distribution is skewed to the right, as proven in the figure below:

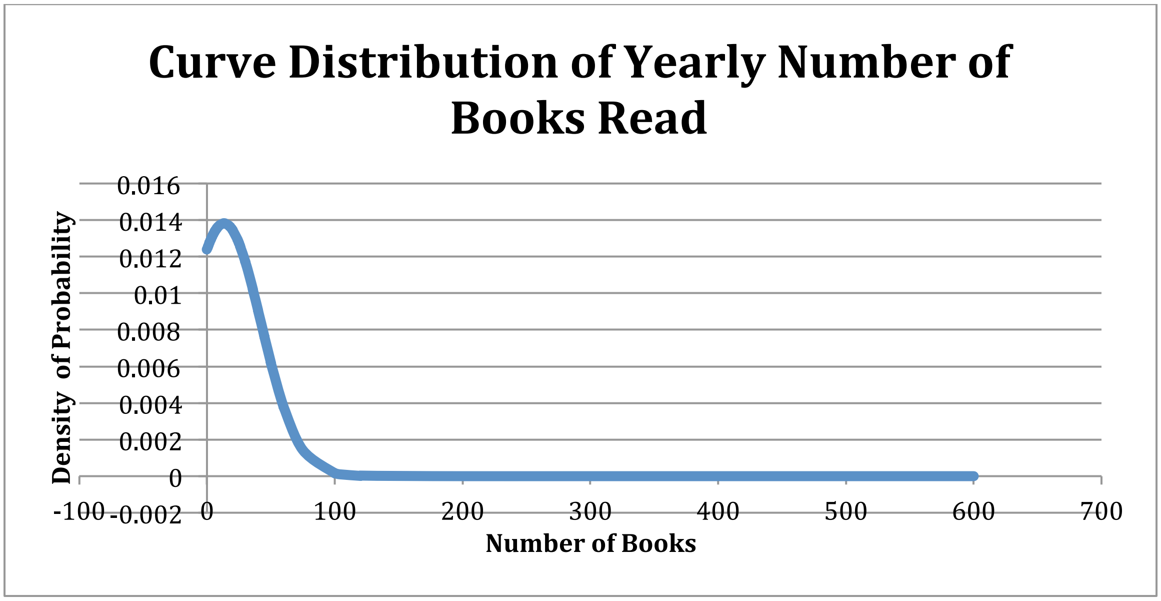


Figure 3a: Curve Distribution of Yearly Number of Books Read by Udacity Students

Conclusion: From Figure 3a, we see that it almost resembles a normal distribution. This could be due to the outlier that is stretching the data. And with a standard deviation of about 28.9 and a mean of 13.4, this is true because the larger the standard deviation, the wider the spread of the range of values from the data points. In other words, the range of the data points is far from the mean.

Question 4: Is there any relationship between the number of hours of sleep and daily commute of Udacity students?

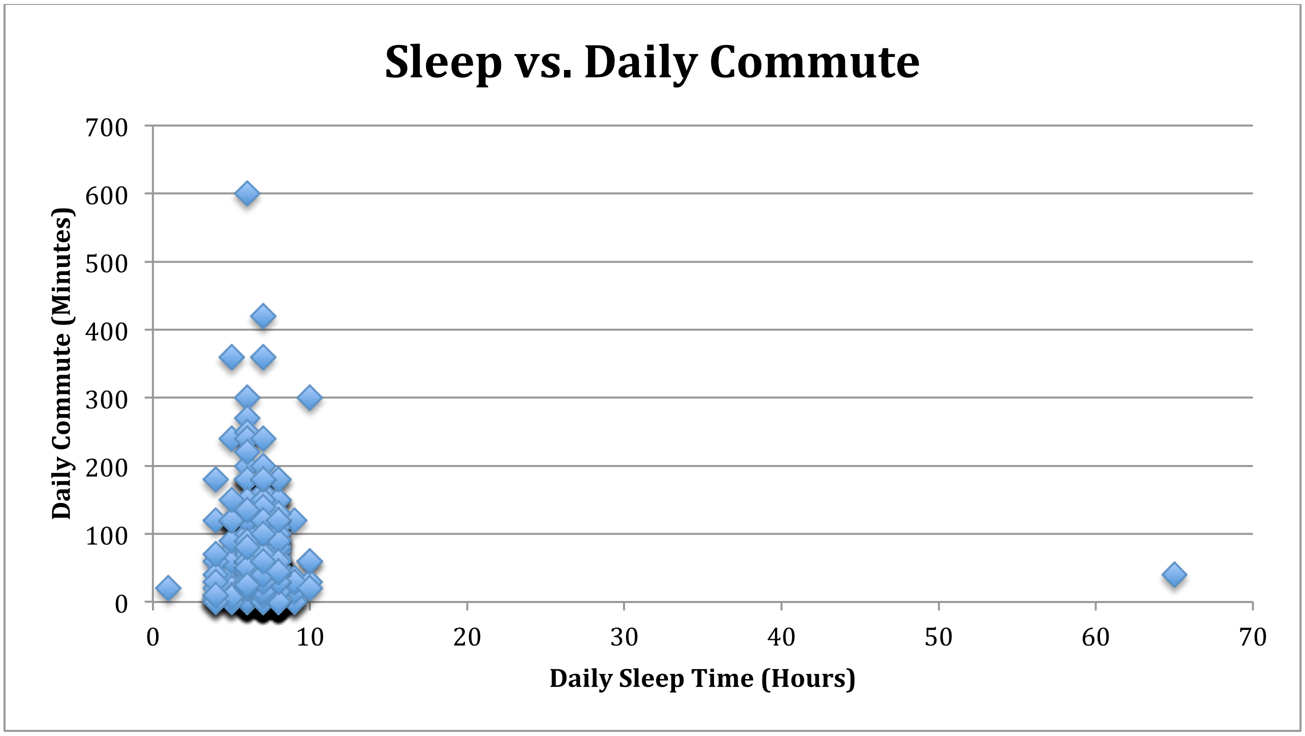


Figure 4a: Scatterplot of the relationship between sleep time and daily commute

For this question, I thought the best way to look at the relationship would be to use a scatter plot. It seems that there are two very large outliers in this graph that does not give a very good visual representation.

Another thing to note is the unit; Unit for Daily Commute are in minutes while sleep time are in hours. However, I don't think it will affect our results much. Even if we were to change sleep time from hours to minutes, it would be relative and would produce the same result.

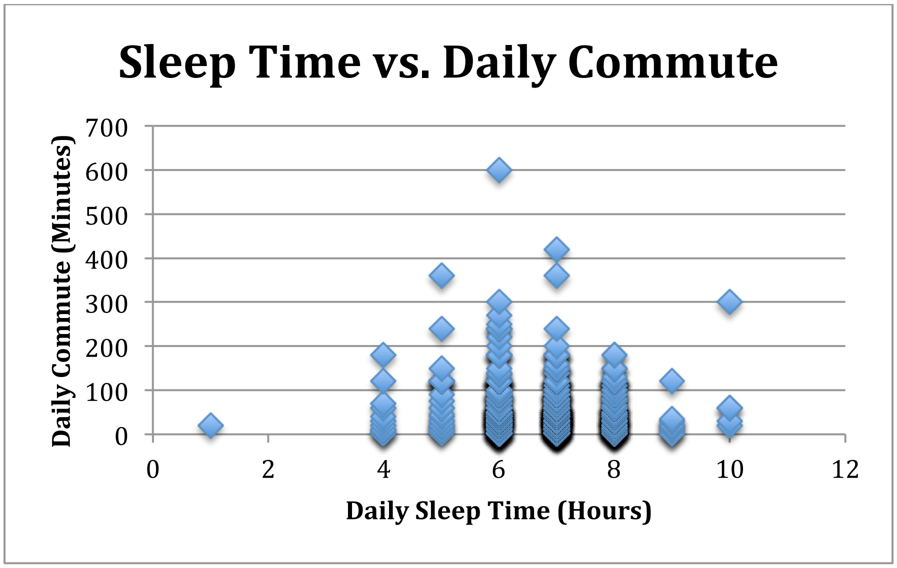


Figure 4b: Scatterplot of the relationship between sleep time and daily commute after removing the outlier.

Conclusion: After removing the outlier, it’s interesting to see that the visual resembles almost like a bell curve. Regardless, it seems there is no clear relationship between the number of sleep hours and daily commute. There could be factors that affect the results that were not accounted for: where the person lived, commuting location (the survey question never stated if the daily commute was for work), and whether the term “Daily Commute” meant the total amount of commute in the participant’s day.