Analyzing Survey Data

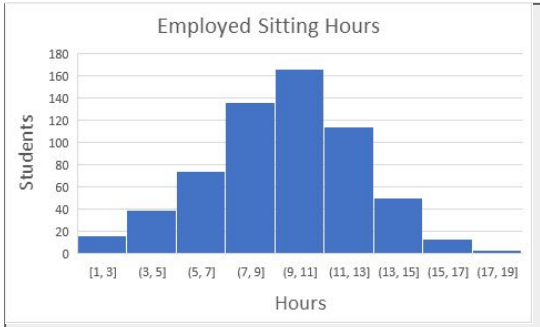
By: Hunter Tarman

Questions we will be answering:

- What is the ratio of employed to unemployed students from the survey? Who sits longer on average, employed or unemployed students?
- What is the age spread and central tendencies of the surveyed students?
- How many Nanodegrees were completed in total and which program was completed the most? The least?
- Which Nanodegree Program required the most study hours? Least study hours?
- How can we determine the difficulty of each Nanodegree? Which is most and least difficult?

*Please note: All information used in this presentation has been collected from a sample of respondents and does not reflect an analysis on the entire population of Udacity students.

Employed vs. Unemployed Daily Hours of Sitting



Total Employed Students:

611

Total Unemployed Students:

131

*Any times given over 24 hours were discarded from the data set.

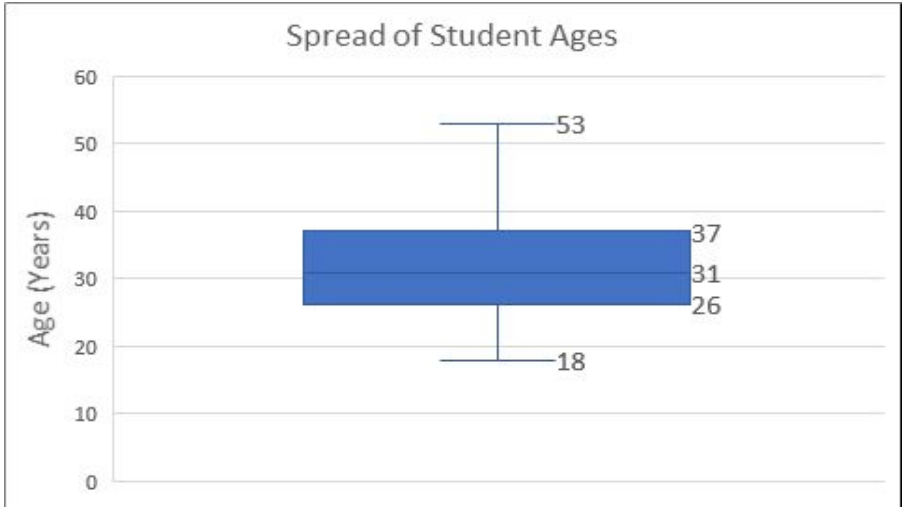
Before examining the hours spent sitting, we will glance at the ratio of employed to unemployed students, which is approximately 5:1 (gathered from the totals to the left). We can see that the participants in this survey are predominantly employed.

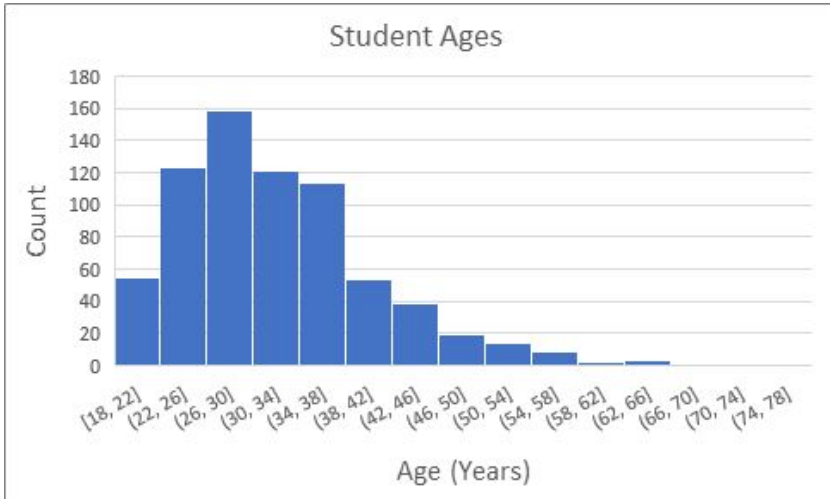
Next, we have two normal distributions of employed and unemployed students with partial skewing in the unemployed set. We can observe that the means of these groups are likely similar from these two visuals, and they are. The employed hold an approximate average of 9.6 sitting hours and the unemployed hold approximately 9.9 sitting hours. Both groups are sitting for similar lengths of the day on average, regardless of their employment status.

Student Age Spread and Central Tendencies

We have 712 age submissions from the survey. To the right, we can see the spread of that data.

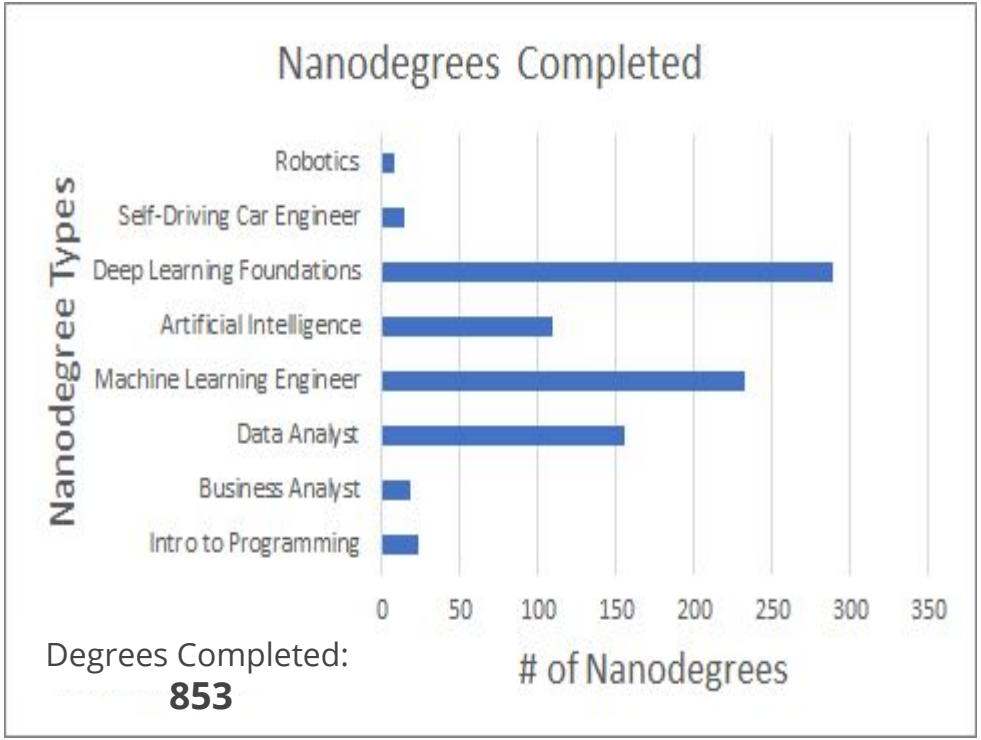
I have removed the outliers and by doing so, we are given a range of 35 years. The median sits at 31 years, within an interquartile range of 11 years. From this plot, we can suggest that a majority of students from the survey are likely between the ages of 26 and 37. The mode probably lies somewhere within these parameters as well.





To the left, we can see the the mode does indeed sit within the interquartile range at an age of 29 years. The average age for students is 33 years, just slightly greater than the median. This is seems accurate because the distribution, although normal, is skewed to the right and pulls the mean along with it. Last, the data gives us a standard deviation of 8.36 years. We can see that our earlier suggestion, being that a majority of students likely sit between the ages of 26 and 37, is roughly accurate. The standard deviation tells us that 68% of students lie between the ages of 24 and 41 years of age.

Nanodegrees Completed

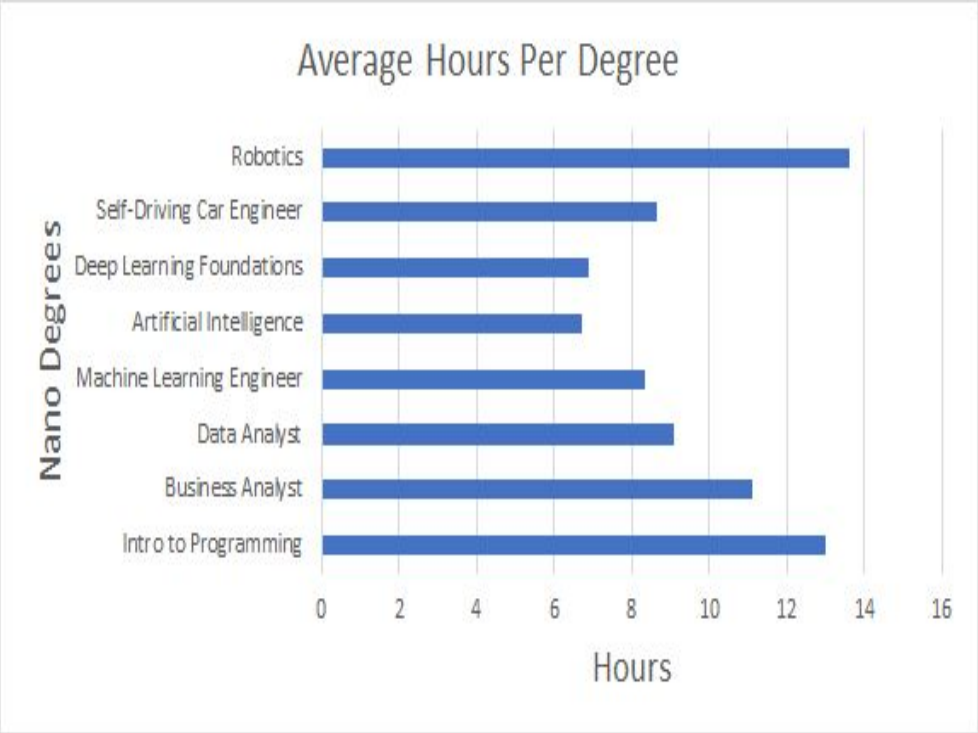


In total, there are 853 Nanodegrees completed with only 753 participants in the study. Some students must have completed several Nanodegrees, although we cannot determine the amount each individual completed with only this information.

We can see that the Deep Learning Foundations program is the most completed, with 289 submitted. The least completed Nanodegree is the Robotics with only 8 submitted. Maybe there is a correlation between completed Nanodegrees and the study time for each.

*Some individuals completed several Nanodegrees and each were counted in this data set.

Study Hours Per Nanodegree



Let's look into the recorded study hours for each Nanodegree. To the left are the daily averages of study hours spent on each of the surveyed programs.

We can see that the Robotics Nanodegree actually pulls the lead with approximately 13.6 hours studied per student on average. If we revert back to the previous slide, we know that this degree is also the least completed.

It would seem that there is a probable correlation between hours studied per Nanodegree and the completion of that Nanodegree.

*Individual Nanodegree study hours are determined by each participant's input; if no hours are given, the Nanodegree is kept but with no study hours. This may alter the accuracy of the data set. Some participants have completed multiple Nanodegrees, so the hours entered by a single participant are duplicated for each program they have completed.

Nanodegree Difficulty

Nanodegrees	Total Weekly Hours/Degree	Total Degrees Completed/ Degree Type	Weekly Hours/Degree Average	Difficulty Score
Intro to Programming	299	23	13.0	**87%**
Business Analyst	211	19	11.1	**74%**
Data Analyst	1415	156	9.1	**60%**
Machine Learning Engineer	1945	233	8.3	**56%**
Artificial Intelligence	741	110	6.7	**45%**
Deep Learning Foundations	1987	289	6.9	**46%**
Self-Driving Car Engineer	130	15	8.7	**58%**
Robotics	109	8	13.6	**91%**

We will call this correlation Nanodegree Difficulty. Let's go ahead and determine the difficulties of our Nanodegrees, set by this formula:

[(Total weekly study hrs per Nanodegree)/(Total Nanodegrees completed per Nanodegree type)/15]

The score itself is determined with values based on a range from 0 - 15, 15 valued at 100%.

If we look at Intro to Programming, there is a Difficulty Score of 87%, however this is one of the entry level programs. Increasing our variables will likely help for calculating a more accurate score, rather than only using the average time taken a week to complete a given Nanodegree.

The Robotics Nanodegree, as indicated from the previous slides, has the highest score because it holds the highest weekly average.