

Project 2: Analyze Survey Data

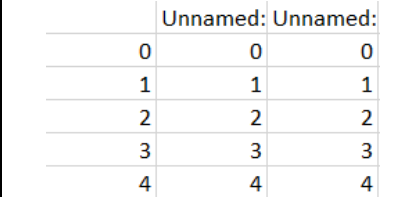
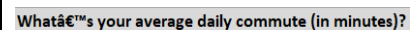
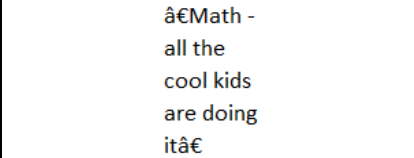

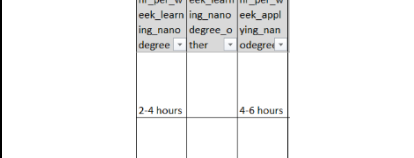
UDACITY: DATA FOUNDATION NANODEGREE
EMILY LAU

Summary

In this project, I will analyze a real dataset about current Udacity students across a number of programs.

1. Data Cleaning

Below shown the list of actions taken to clean the data before data exploration and analysis.

Screenshot	Action	Reason
	First column renamed as "index".	First column is empty and it looks like it is the numbering for each row.
	Column B & C removed	Column B & C also known as Unnamed: 0 is the same as the first column.
	Change all column names to be informative instead of as question. Remove all weird characters, change all to lowercase and replace spaces with underscores.	Columns are harder to read and this allows the data to be easier to read in other means such as SQL, etc.
	Removed weird characters using find and replace function: - 400 "œ" - 452 "• "	Data contain weird and unreadable characters.
	Created new column to calculate the current age based on birthdate.	Age would be easier to perform analysis than birthdate.
	Replace values with the average.	Unable to conduct analysis on values such as "2-4 hours"

2. Data Exploration

2.1. Youngest and oldest participant

The youngest participants are between the age of 0 to 1 and oldest is 118 which does not seem to be right. After comparing the age against the birthdate field, it seems that those with empty birthdate will result with the age of 118. As such, all the age with the value of 0, 1 and 118 are removed from this dataset.

2.2. Average number of hours of sleep

Based on the dataset the lowest is 1 and the highest is 9,141,984 hours of sleep in a day. There are also participants with 45, 65 and 85 hours of sleep in a day, which

exceed the number of hours in a day. The values that exceed the number of hours in a day could be a typo and are removed from this dataset.

2.3. Average number of hours commuting in a day

One day is equal to 24 hours, which is 1,440 minutes. Based on the dataset, the lowest is 0 and highest is 600 which is still reasonable as it did not exceed the number of hours in a day.

2.4. Average number of hours spent sitting in a day exceeds the number of hours in a day.

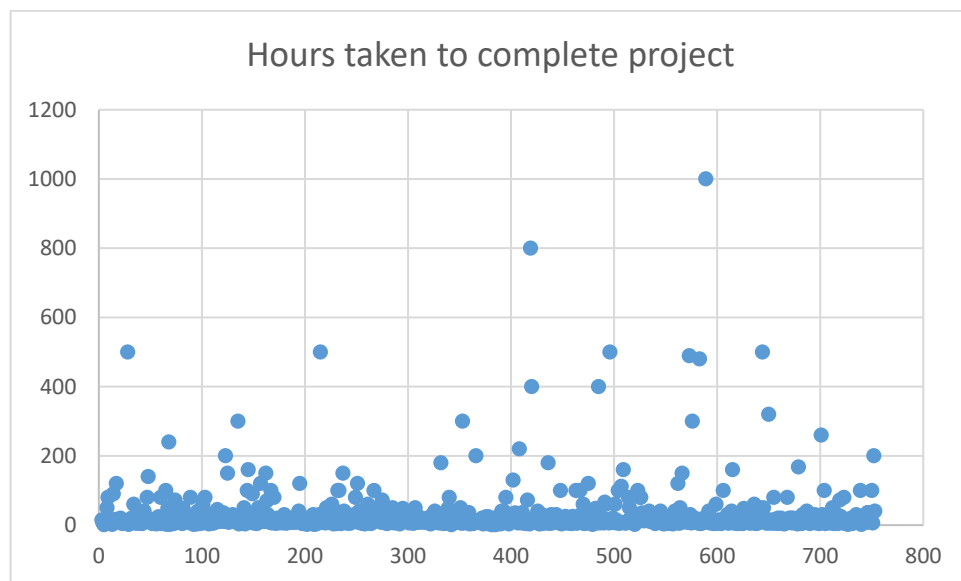
One day is equal to 24 hours. Based on the dataset, the lowest is 1 hour and highest is 800 hour, which is impossible as it exceed the number of hours in a day and they will be removed from the dataset.

2.5. Years of experience

The lowest is 0 and the highest is 40 which seems reasonable.

2.6. Average hours worked on a project exceeds the number of hours in a day.

The lowest average spent is 1 hour and the highest is 1,000 hour. Based on the scatter chart shown below, there are a few outliers who spent more than 800 hours.



2.7. The Maximum number of books read or listened to in a year exceeds the number of days in a year.

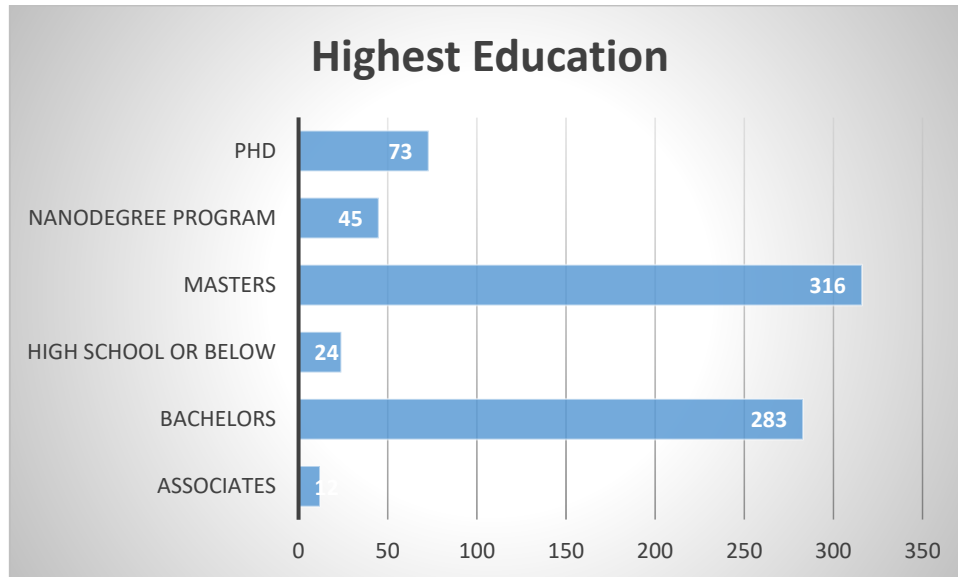
The lowest number of books read or listened to in a year is 0 and highest is 600. Even though it exceeds the number of books read exceed the number of days in a year but it have been proven in an article that there are people who did read more than 365 books in a year which means that this is not impossible.

Article: <http://latimesblogs.latimes.com/jacketcopy/2009/01/how-to-read-462.html>

3. Data Exploration

3.1. What was the most frequent “Highest Degree Earned”?

Masters is the most frequent highest degree earned followed by bachelors and Phd.

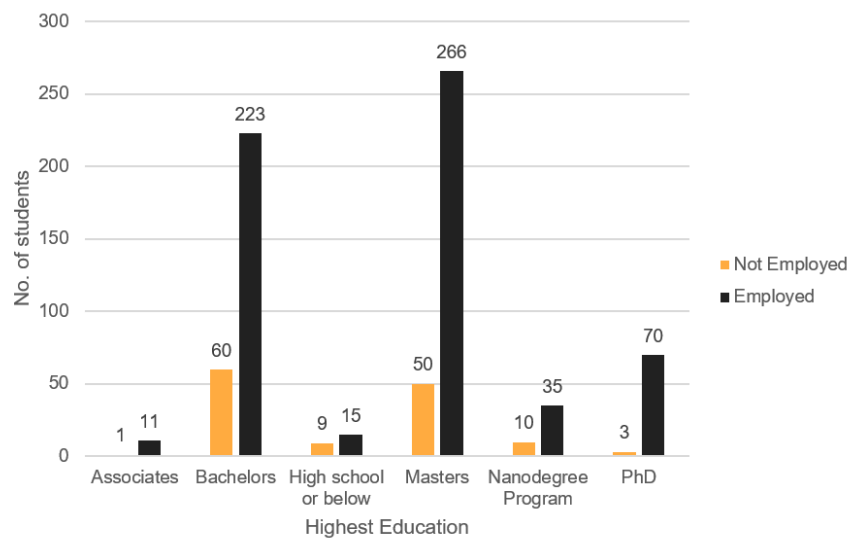


3.2. What is the most common Nanodegree Program students of the survey are taking?

Based on the tabulation shown below, the top three most popular Nanodegree Program is Deep Learning Foundation, Machine Learning Engineer followed by Data Analyst.

Intro to Program	Business Analyst	Data Analyst	Machine Learning Engineer	Artificial Intelligence	Deep Learning Foundations	Self-Driving Car Engineer	Robotics	None	Other.6
		Data Analyst							
23	19	157	235	111	291	15	8	46	43

3.3. What proportion of students who completed the survey are employed?



Highest Education	Not Employed	Employed
Associates	1	11
Bachelors	60	223
High school or below	9	15
Masters	50	266
Nanodegree Program	10	35
PhD	3	70
	133	620

Based on the chart and table, 620 (80%) of Udacity graduates are employed. Most of the graduates are equipped with Masters (43%) and Bachelor (36%). It is also shown that PhD graduates have fewer unemployment.

3.4. Of students who have a Master's or PhD, what percent responded as not employed?

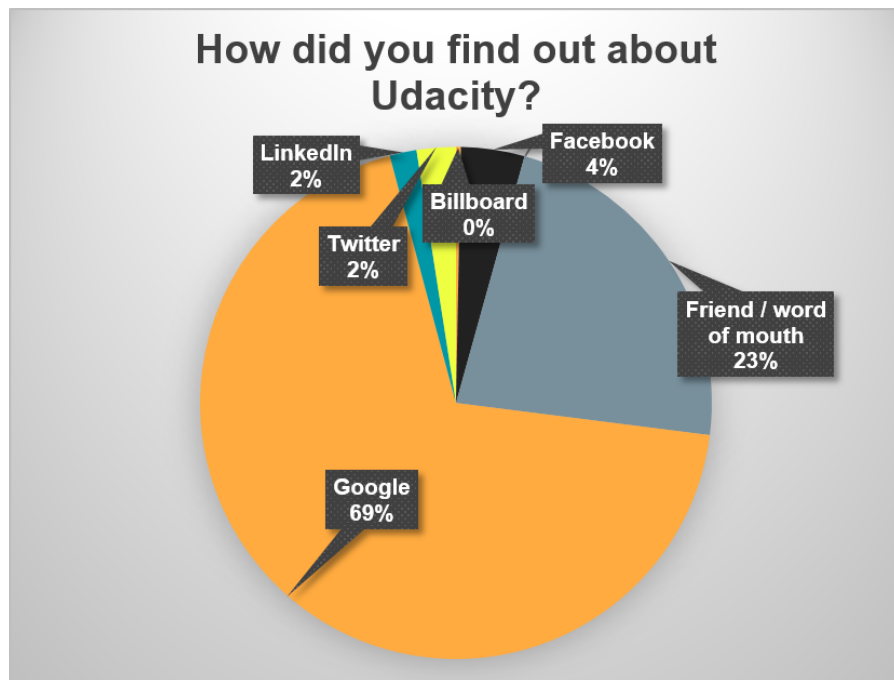
Based on the table shown below 53 out of 389 (14%) students responded as not employed.

Highest Education	Not Employed	Employed	Grand Total
Masters	50	266	316
PhD	3	70	73
Grand Total	53	336	389

3.5. What is the daily commute time (in minutes) such that 50% of students must travel further?

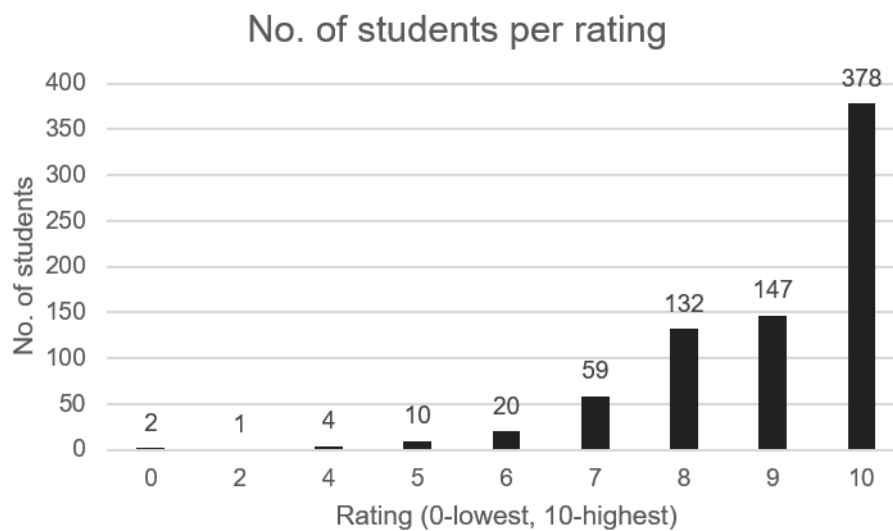
The median commute time per day is 35 mins.

3.6. What is the most common way people find out about Udacity?



Based on the graph shown on the left, 446 students found out about Udacity through Google, which is also the top source. Followed by through friends or word of mouth with 147 students. The rest of the sources such as Twitter, LinkedIn, Facebook and Billboard are less than 100 students.

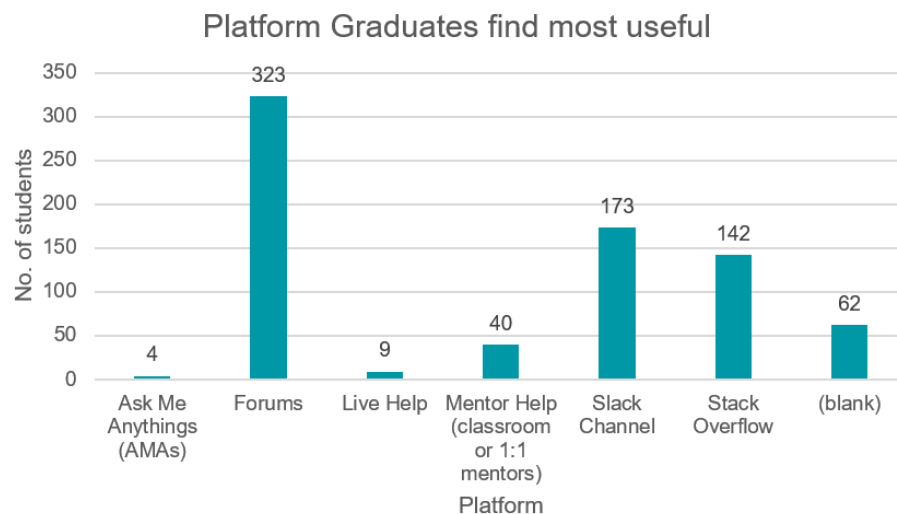
3.7. Do Udacity graduates like the Nanodegree program?



Min	0
Max	10
Mean	8.98
Median	10
Mode	10
Standard Deviation	1.36

The chart shows a skewed-left distribution, which means the mean is less than median, this also shows that most graduates like the Nanodegree program. The min is 0 and max is 10 which tells us how widely spread out the extreme observation of the distribution are. The mean of the distribution is 8.98 and 10 for both the median and mode, which means most students, will give a rate a minimum rating of 8. A standard deviation of 1.36 tells us that most of the numbers are very close to the average. However, this is just survey respondents and is not from the entire Udacity Student population thus it does not tell us what everyone thinks.

3.8. What are the platforms graduate find most useful?



Based on the chart, the platform where graduates find most useful is Forums (42%). Followed by Slack Channel (23%) and Stack Overflow (19%). It also show that the lease useful ones are AMAs and Live Help.