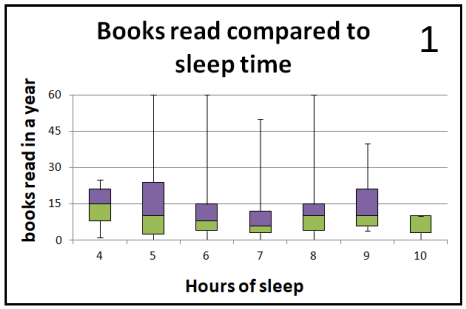
Project 2

Analyze Survey Data

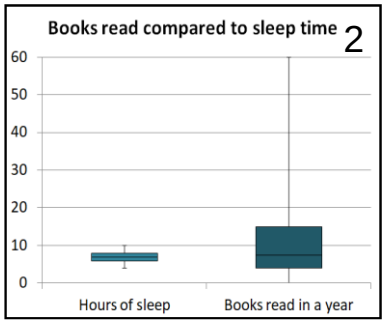
M. Isabel Gonzalez Rosas

Is there a relationship between hours of sleep and quantity of books read in a year? Does it vary regarding location?



Here we can see the relationship between the quantity of books read in a year and hours of sleep from a sample of 734 Udacity students after cleaning some outliers and missing information from the original sample.

Both mode and median are 7 hours of sleep a day and mean is almost 7 too (6.94) showing an almost centered/symmetric distribution in plot 1. This is so since most frequent hours of sleep(mode) is 7 and standard deviation is just 0.95. Plot 2 shows that values for hours of sleep are not too disperse from the mean, 7 hours, which is also the mode and, as the range is also small, going from 4 to 10 hours of sleep (6 hours range), there is not much room to disperse to.

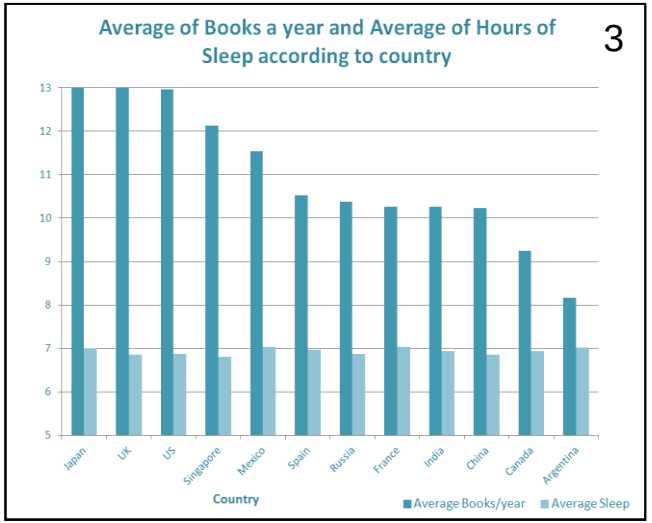


This is not the case for quantity of books read in a year where, although mean and mode are near 10, median is 7.50 and the values go from 0 to 60 books in a year. This causes a standard deviation of almost 11 books (10.90).

	Sleep_a_day	Books_a_year
Mean	6.94	10.98
Median	7.00	7.50
Mode	7.00	10.00
Standard Deviation	0.95	10.90
Variance	0.91	118.84
Range	6.00	60.00
Minimum	4.00	0.00
Maximum	10.00	60.00
Sum	5091.00	8060.00
First quartile	6.00	4.00
Third quartile	8.00	15.00
Sample	734.00	734.00

By looking at plot 1, ***its seems there is no relationship between hours of sleep and books read a year.*** All medians are 15 books or less, disregarding the quantity of hours slept. Plot is neither left nor right skewed but mostly centered. This can be due to a lack of strength for reading due to poor sleep in case of 4 hour-sleep, or lack of time to read due to excessive sleeping, in the 9 and 10 hour-sleep group. But students sleeping five to eight hour have similar averages for books read in a year.

However, although average hours of sleep is rounded to 7 in every country on the sample, plot 3 shows that ***the average of books read a year does varies according to location***, going from an average of 13.14 books in Japan to 8.17 in Argentina.

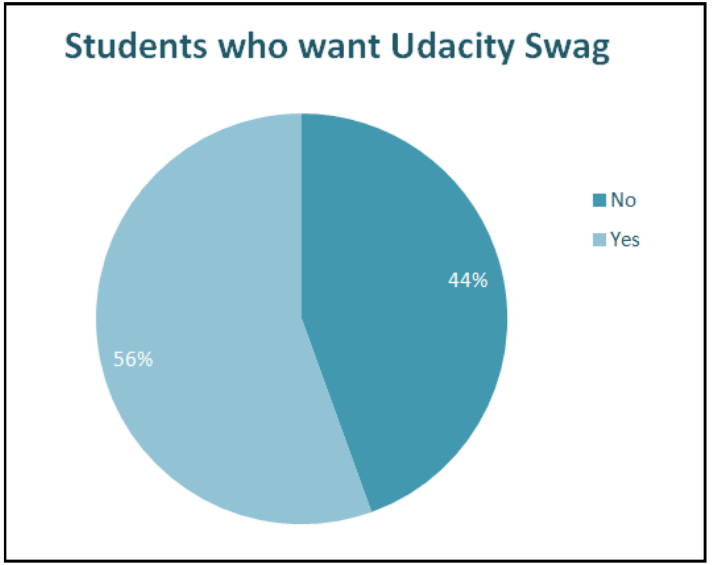


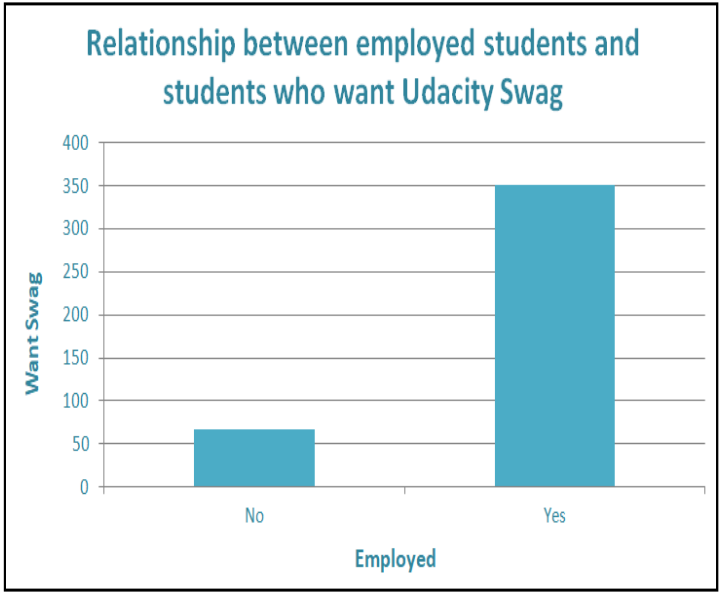
What students from the sample are more interested in acquiring Udacity swag, employed or unemployed ones?

From this plots it can be seen that 56% of the Udacity sample of 753 students (left after cleaning students with missing employment information), want Udacity swag. That's 418 students.

Among those 418, it seems the majority of them (almost 84%) is employed.

So it can be said that within the sample students, ***employed students are more interested in acquiring Udacity swag***.





Employed	SUM Udacity Swag
No	67
Yes	351
Total	**418**

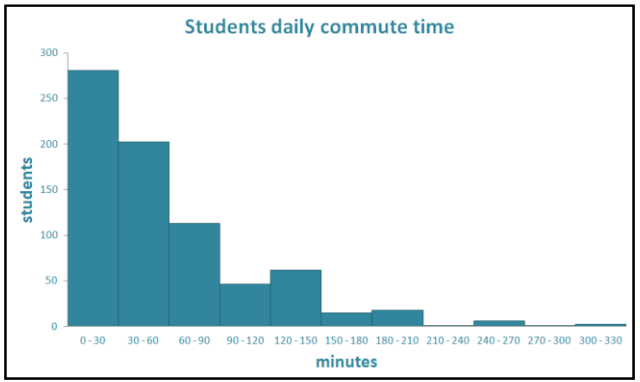
Want Swag	Sum
No	335
Yes	418
Total	**753**

Is there any difference regarding commute time and country where the student lives?

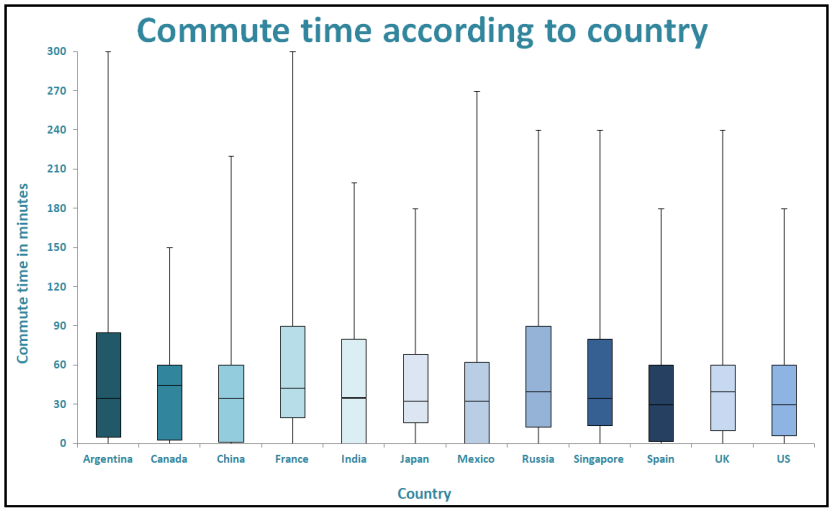
After cleaning some outliers with more than 300 minutes of daily commute time and some missing information from the students sample, the histogram shows that most students commute time varies from 0 minutes to an hour and a half (90 minutes) within a range of 300 minutes. The plot shows a right skewed distribution, having a mean of 49 minutes, greater than the 35 minutes median and 0 minutes for the mode.

After dividing the data according to the country of precedence, it can be seen, in the box plot, that this is also true for all of the countries from the sample since none third quartile exceeds 90 minutes commute time and median for all of them ranges just from 30 (Argentina, Spain, US) to 45 minutes (Canada).

It can be affirmed that ***there is no difference in commute time regarding country where the student lives*.**



Commute time	
Mean	49
Median	35
Mode	0
Std. Dev.	51
Q1	8
Q3	60
Range	300
Min	0
Max	300
Count	747



Box Plot on Ms Office 2007: https://www.youtube.com/watch?v=ZFbPnwKwVWk

	Argentina	Canada	China	France	India	Japan	Mexico	Russia	Singapore	Spain	UK	US
Mean	50.99	48.34	43.13	59.29	48.93	45.71	47.40	54.99	54.05	41.02	49.07	41.56
Median	30.00	45.00	35.00	42.50	35.00	32.50	32.50	40.00	35.00	30.00	40.00	30.00
Mode	0.00	60.00	0.00	30.00	0.00	0.00	0.00	120.00	0.00	0.00	0.00	0.00
Std. Dev.	57.33	45.67	47.15	57.03	50.88	41.14	51.38	53.41	56.90	44.08	56.97	42.11
Q1	5.00	2.75	1.00	20.00	0.00	16.25	0.00	13.00	13.75	2.00	10.00	6.25
Q2	85.00	60.00	60.00	90.00	80.00	68.75	62.50	90.00	80.00	60.00	60.00	60.00
Range	300.00	150.00	220.00	300.00	200.00	180.00	270.00	240.00	240.00	180.00	240.00	180.00
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max	300.00	150.00	220.00	300.00	200.00	180.00	270.00	240.00	240.00	180.00	240.00	180.00

Histogram on Ms Office: https://www.youtube.com/watch?v=uLmKJH7zE-Y

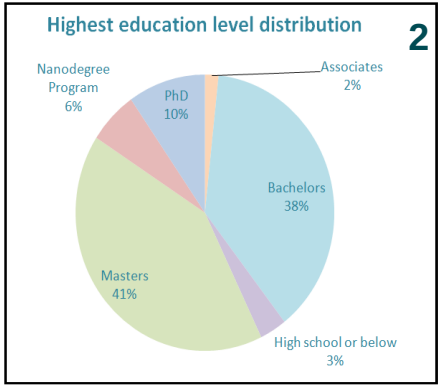
Is there a relationship regarding age and students whose highest education level is a nanodegree?

As per graphs 1 and 2 it can be seen that the majority of students' from the sample higher level of education is Bachelor or Master. There is a 6% of students whose higher education level is a nanodegree.

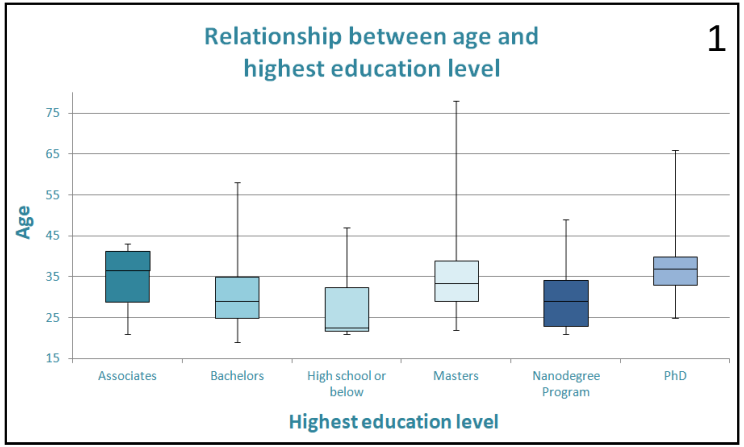
Among those 44 students most of them are on their twenties and early thirties as it can be seen on plot 3, having a mode of 22 and median of 29.

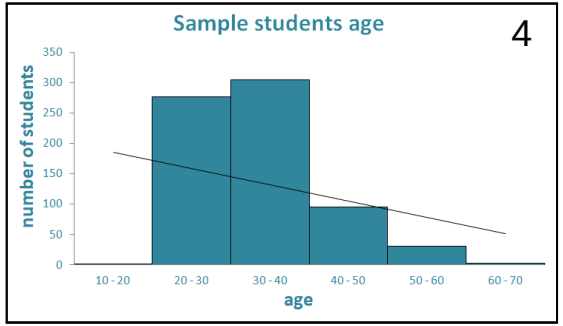
Although range varies greatly between nd students (28) and all students from sample (59), it seems the standard deviation is rounded to 8 in both cases and the average age (mean) is not so different, having 30 and 33 years.

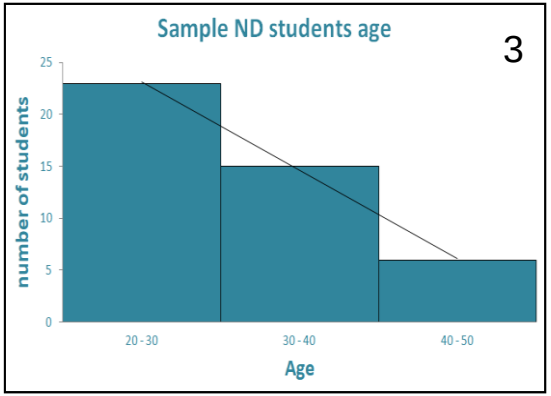
It can be seen regarding them that ***there is a tendency that shows that quantity of students whose highest education level is a nanodegree decreases while age increases**** (maybe because students tend to finish other studies such as Bachelor or Master while growing older or maybe because online training is not so popular on elders) as it shows plot 3.



Age of	ND	All
Mean	29.98	33.23
Median	29.00	32.00
Mode	22.00	29.00
Std. Dev.	7.54	8.37
Q1	23.00	27.00
Q3	34.25	37.00
Range	28.00	59.00
Minimium	21.00	19.00
Maximum	49.00	78.00
Count	44.00	712.00







* This is also true for all other students from the sample as it can be seen in plot 4.