Analyzing Udacity Nanodegree Survey Data

Udacity DFND

Project 2: Analyze survey results

Submitted by:
Manal Almehmadi

Analyzing Udacity Nanodegree Survey Data

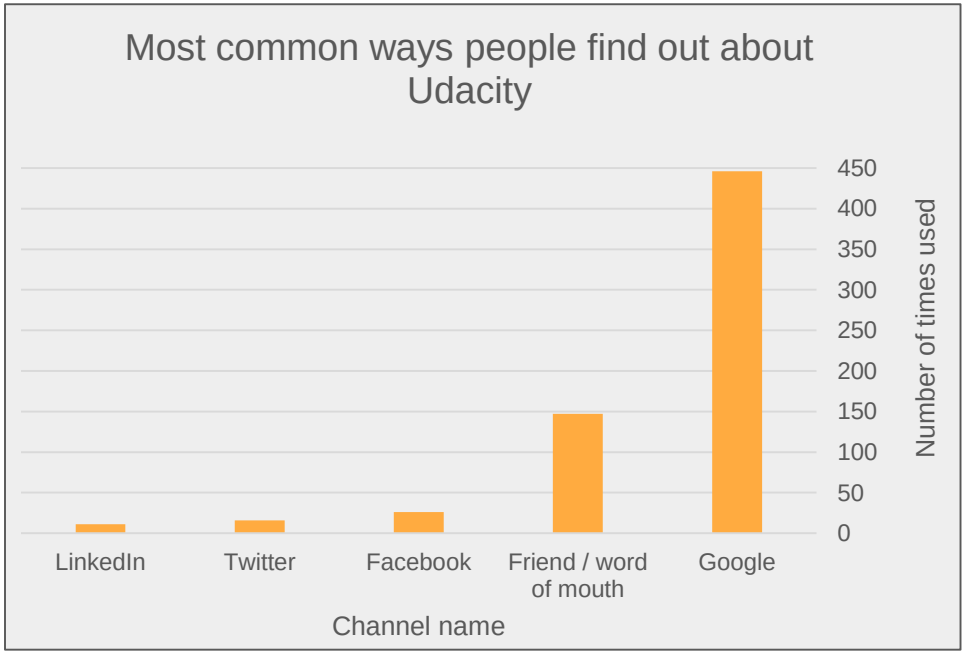
In this project, I will analyze a dataset about Udacity students across a number of nanodegree programs using Microsoft Excel. The dataset is a little messy so, I need to clean the data first.

Cleaning the dataset

What I did to clean the dataset are:

- Combine multiple columns that answer one question into a single column and remove the original columns like “ways to find out about Udacity” field.
- Change some column name to be short, and still informative.
- Compute age from “birth date” filed and remove it.
- Remove outlier values in multiple fields like “age_years” filed, “hours of sleep per night” filed, “Hours spend in study per week” filed, “hours of spend sitting per day” because some of them contains hours over the day hours.

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

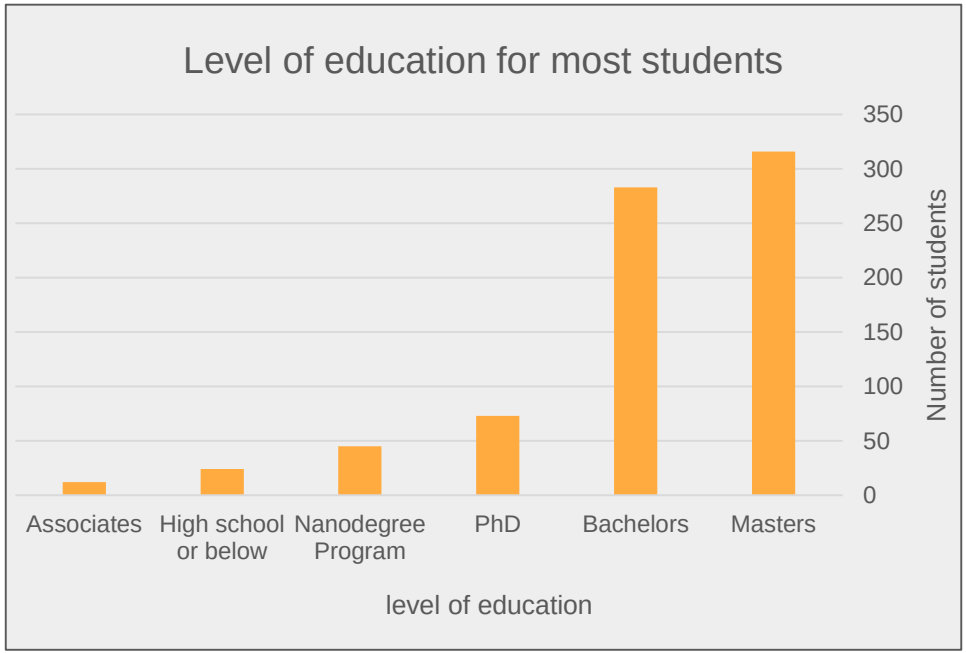


As shown in the bar chart, which represents the top 5 channel students knew the Udacity through it. So Udacity can use these channel in ads to attract more students and becomes more popular.

We can analyze from the chart the top 5 channels (consecutive) are Google, from friends, Facebook, Twitter, LinkedIn. Google has the highest number of people with a percentage of 59%.

channel	Google	Friend. Word of month	Facebook	Twitter	LinkedIn
%	59 %	19 %	3 %	2 %	1 %

What is the highest level of education for most students?

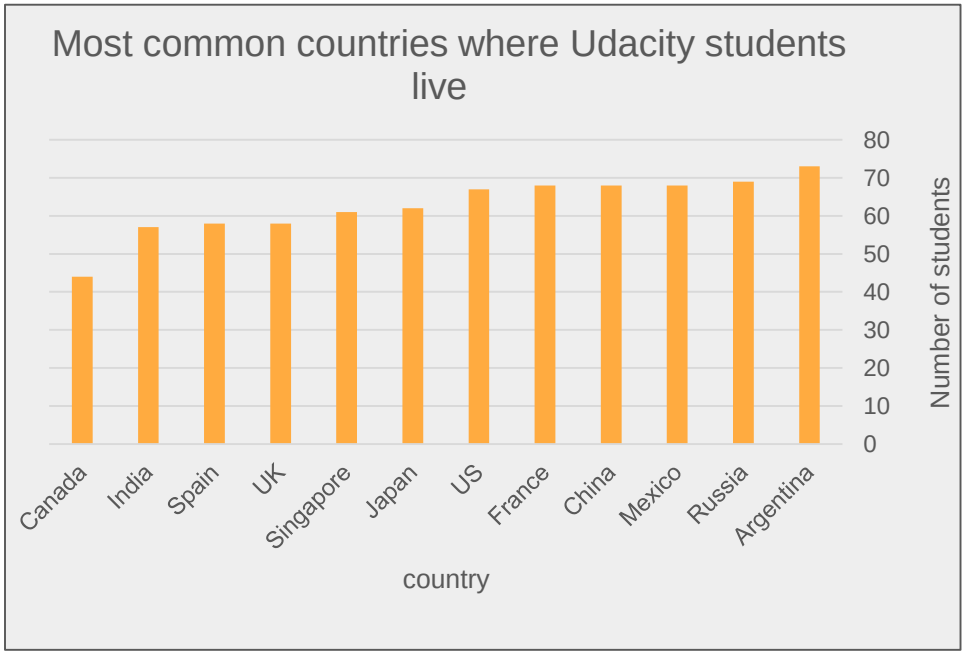


It is good for Udacity to know the education level of its students. So it can focus on what is the most appropriate nanodegree program it should offer and the interests of its students.

Based on data from the survey respondents and after being represented at the bar chart, it is clear that most of the students of Udacity have **Master** degree with 316 students.

Education	PhD	Masters	Bachelors	Nanodegree program	High school or below	Associates
Num. students	73	316	283	45	24	12

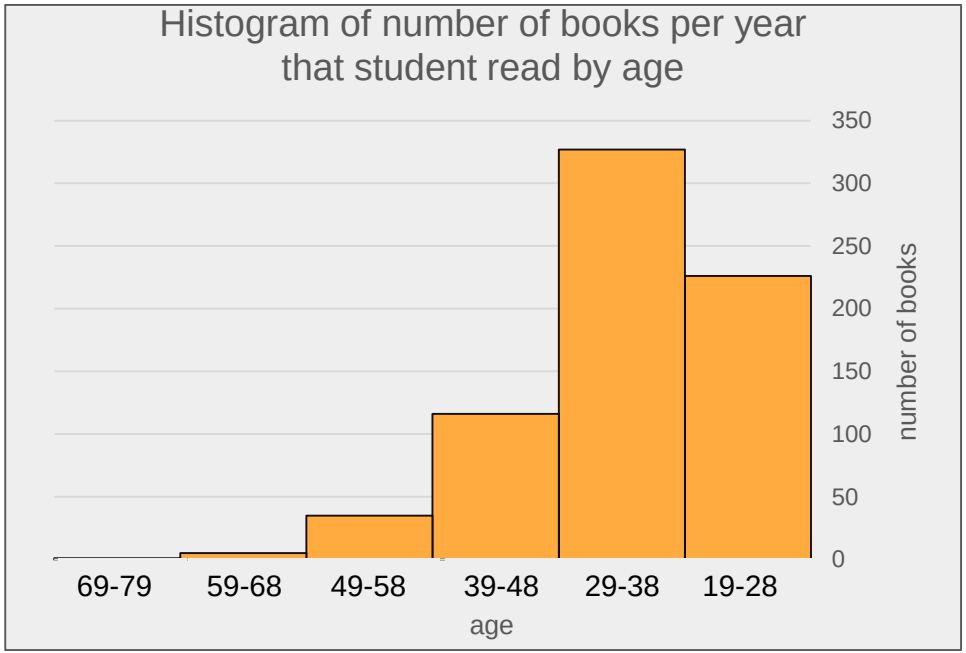
What are the most common countries/cities where students live?



To answer this question, I have created the bar chart with countries name on X-axis and count of students live in it on Y-axis. It shows the top 10 countries the Udacity students live in, the bins show slight difference in number of students. however Argentina has the most students live in with 73 students.

Argentina	Russia	Mexico	China	France	US	Japan	Singapore	UK	Spain	India	Canada
73	69	68	68	68	67	62	61	58	58	57	44

What is the age group most of the students?



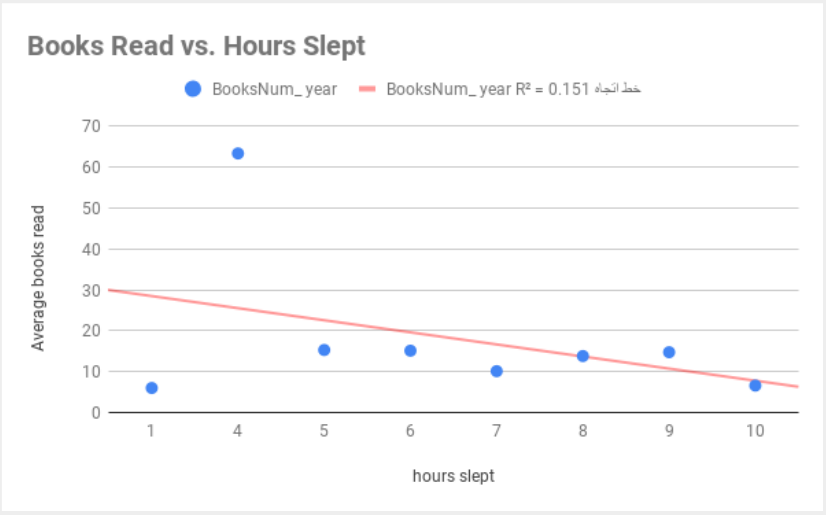
As shown in the histogram, the most students' age in range 29-38.

A look at the chart of age group for the most of the students reveal that the data is Symmetric distribution. So **(Mean = Median = Mode)** but here, **median = mean=32**, the **mode is 29** means most students' age is 29. The **high SD** (I calculated SD after removing the outliers because it affects on the standard deviation) shows the ages of Udacity students are very spread and the Udacity has nanodegree programs appropriate to different ages. Also, **range (60)** which is difference between min and max, represent data is spread from the mean.

continue

Median	32
Mean	32
Mode	29
Min	19
Max	79
Range	60
Stander deviation	8.37

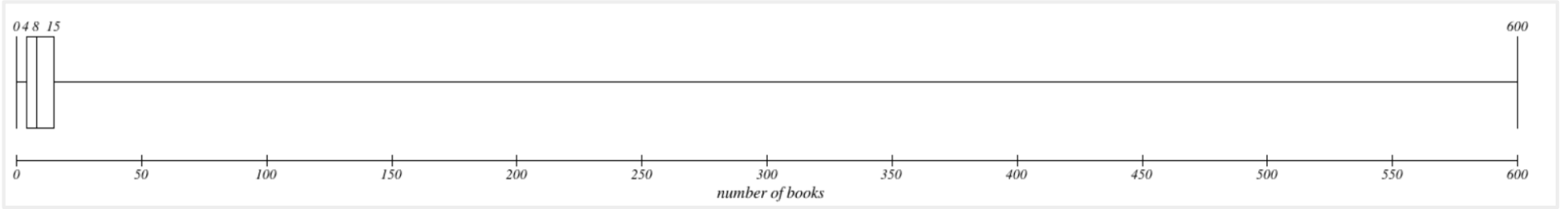
Do Udacity students if they get fewer hours of sleep, will read books more?



I created these charts on Google spreadsheet because the my Microsoft Excel 2013 take several steps to create box plot. This is a link for spreadsheet

https://docs.google.com/spreadsheets/d/19FO64fJnkqA47jT7pQ1e1Gc_PgDV3OIoeKAOMLgE2k8/edit?usp=sharing

Box plot of books read per year



Do Udacity students if they get fewer hours of sleep, will read books more?

The scatter plot shows a negative correlation with hours slept to books read and very weak (R= 0.15) , mean if the hours of sleep increase, the number of books read per year will decrease. Which are support my question's assumption. The **mode is 10** means the most Udacity students read 10 books per year. The standard deviation is very high for number of books read at approximately 28.9 that mean the data is spread from the mean (13.5)

	min	quartile 1	quartile 2	quartile 3	max	mean	mode	standard deviation
Number of books	0	4	8	15	600	13.5	10	28.87

This data is from survey respondents and is not from the entire Udacity students population.

Thank you