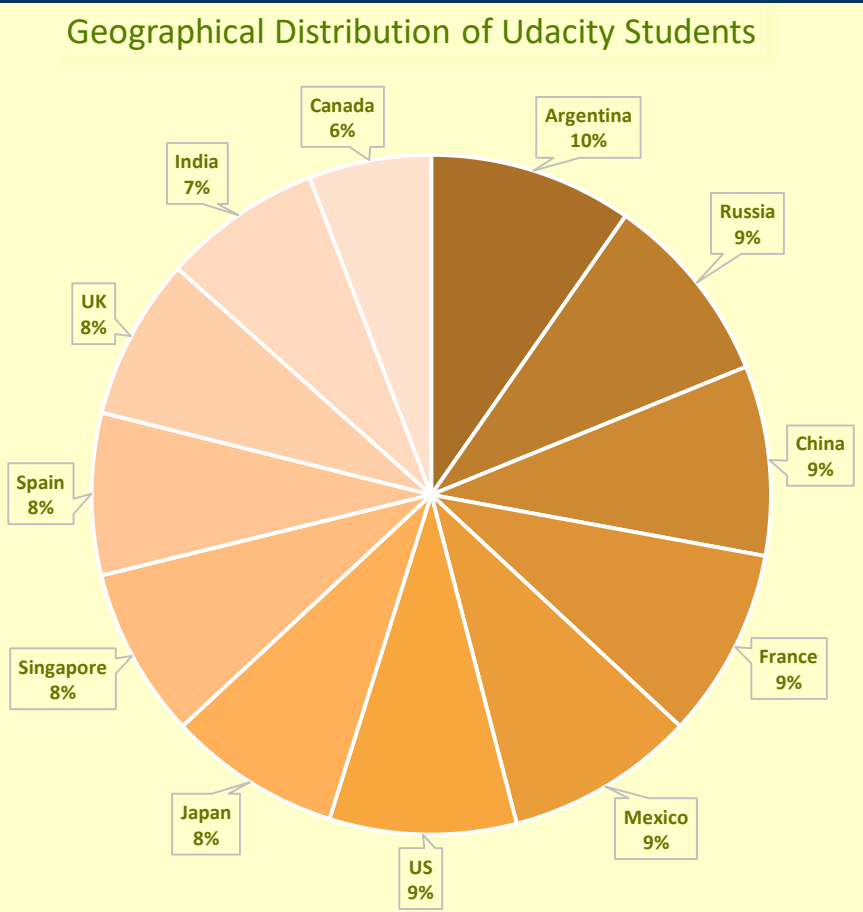
Marketing Analytics Nanodegree Program

Project 2
Analyze Survey Data

By Audrey Tan
10/04/2019

What countries did the surveyed students come from?



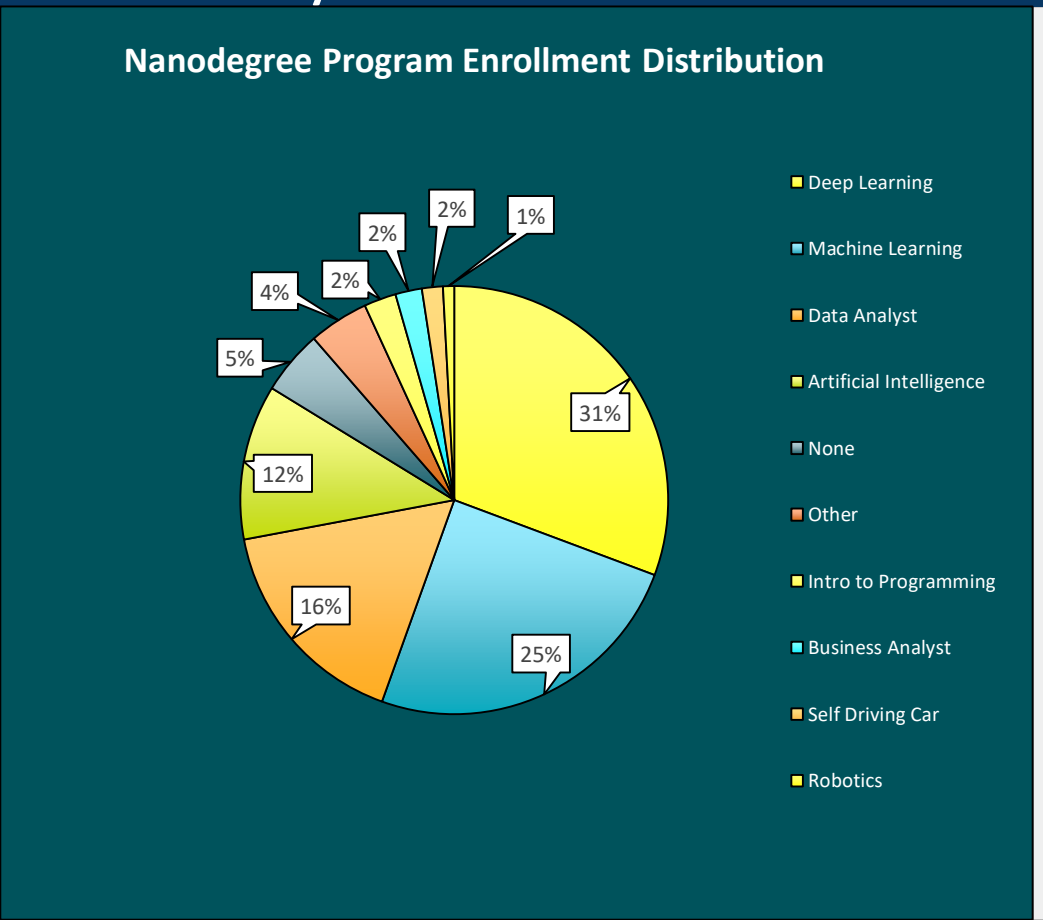
The surveyed students in the sample dataset came from around the globe from North and South Americas, to Europe and Asia.

The sample dataset shows student geographical locations are quite even distributed ranging between 6% to 10% of total population.

The distribution ranking is as follows with Argentina ranks the highest at 10% and Canada the lowest at 6%:

1. Argentina – 10%
2. Russia, China, France, Mexico, US - 9%
3. Japan, Singapore, Spain and UK - 8%
4. India – 7%
5. Canada - 6%

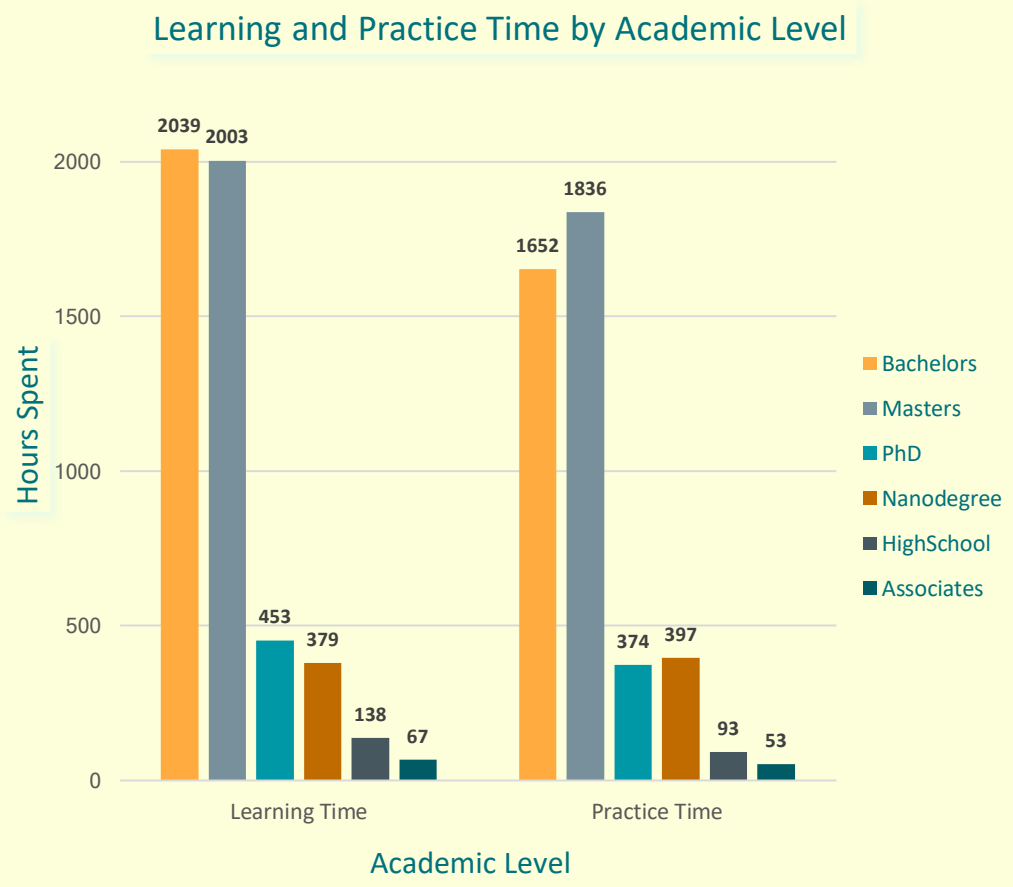
What are the most enrolled Nanodegree Programs among the surveyed students?



Among the surveyed students, the most enrolled Nanodegree Programs are in the following descending order of popularity:

1. Deep Learning
2. Machine Learning
3. Data Analyst
4. Artificial Intelligence
5. None
6. Other
7. Intro to Programming
8. Business Analyst
9. Self Driving Car
10. Robotics

How much time did the surveyed students from different academic levels spend on learning and practicing?

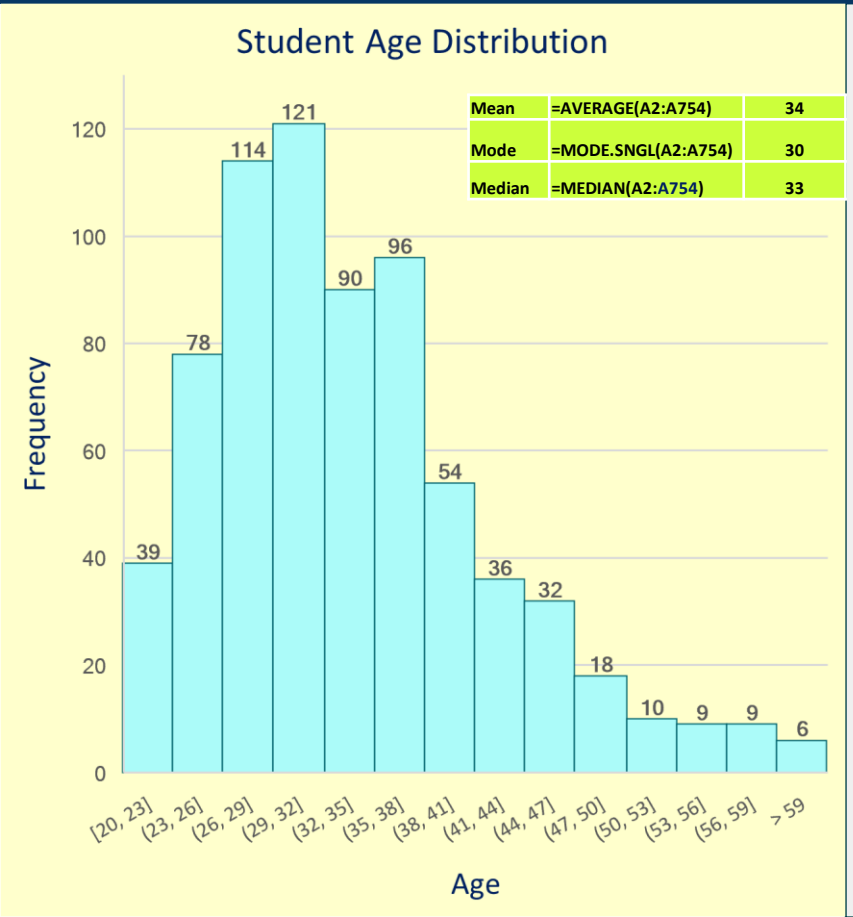


The clustered bar chart shows the patterns of hours spent among 6 academic levels in learning and practice time.

In learning time, Bachelors and Masters top the chart with 2039 and 2003 hours respectively, followed by PhD, Nanodegree, High School and Associates with 453, 379, 138 and 67 hours respectively, although the number of hours spent by the latter 4 groups are significantly lower than that of the former 2 groups.

Interestingly, the trends for the 6 groups in practice time mirror that in learning time, with Masters and Bachelors toping the chart and spending significantly more hours than that by PhD, Nanodegree, High School and Associates.

What is the age distribution among the surveyed students?

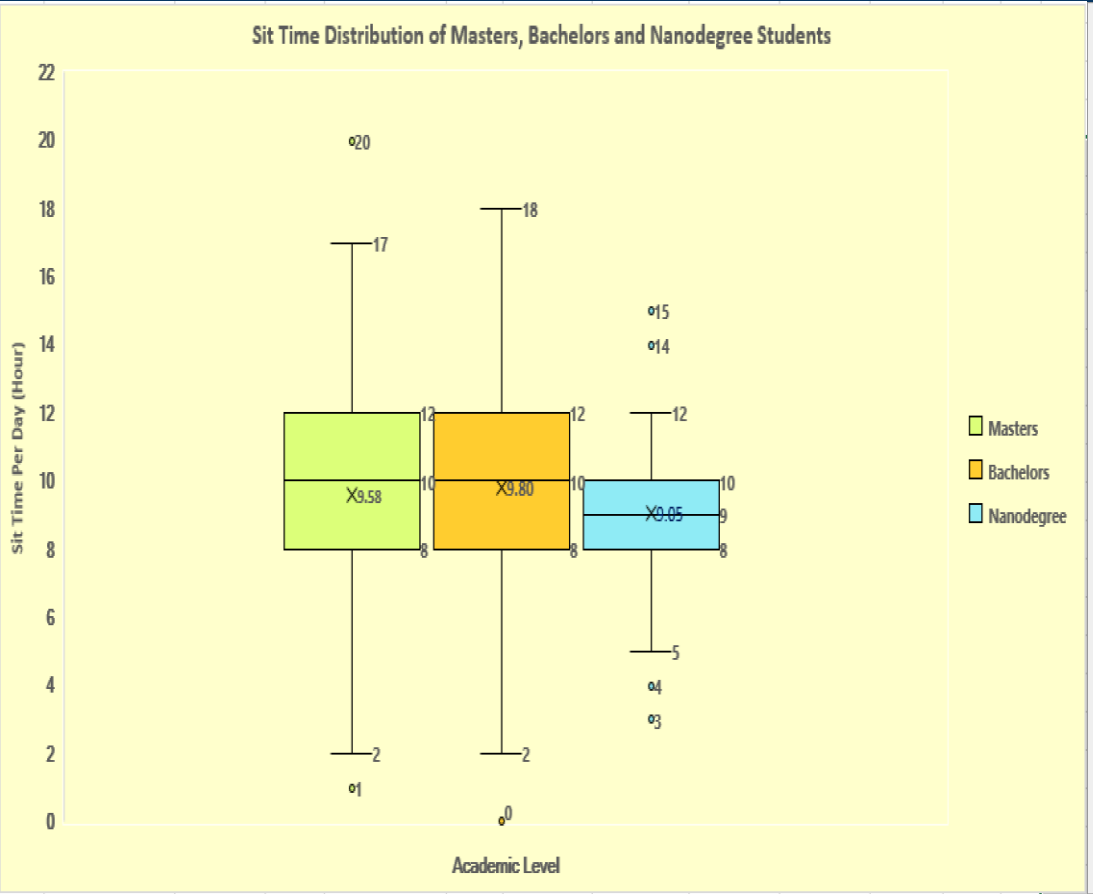


The histogram, sans the outliers (i.e. without DOB or misstated DOB), shows a right skewed distribution with a "tail" stretching toward the right.

The usual metrics Mean (34), Mode (30) and Median (33) used in measures of center are no longer identical in value as in a normal distribution. The mean is greater than the median and no longer a better measure of center for this distribution. The standard deviation is likely not an appropriate measure of variability (or spread) for this distribution as well.

The mode in this dataset is 30, it is the most frequently occurring age and falling between 26 and 32. It tells us on average the students in the dataset were in their early 30s.

Is there any difference in sit time pattern among surveyed students who hold Masters, Bachelors and Nanodegree?



The box plots for students of Masters, Bachelors and Nanodegree look symmetrical, the means (9.58, 9.80, 9.05) for all 3 are quite close in values.

Their interquartile range IQR (4,4,2), standard deviations (2.93, 3.08, 2.58) and ranges (15, 16, 7 excluding outliers) are pretty close in values too.

All these suggest they have similar sit time pattern and distribution, especially so for students of Masters and Bachelors.

Extreme outliers can cause drastic shift in the values of range and standard deviation, but won't affect IQR. Thus, we can use IQR to identify outliers and also as a less sensitive measure of the spread of a data set.

Summary Statistics, Measures of Center and Spread for Sit Time Distribution of Masters, Bachelor and Nanodegree Students

Summary Statistics	Master	Bachelor	Nanodegree
Minimum	2	2	5
Q3	12	12	10
Q2	10	10	9
Q1	8	8	8
Maximum	17	18	12
Mean	9.58	9.80	9.05
Median	10	10	9
Std Deviation	2.93	3.08	2.56
Range	15	16	7
IQR	4	4	2

Appendix I: StdDev, Range, IQR, Mean, Median calculation for Sit Time Distribution of Masters, Bachelor and Nanodegree students

Masters	Std Dev	=STDEV.S(B1:B313)	**2.93**
	Range	=LARGE(B1:B313,2)-SMALL(B1:B313,2)	**15**
	IQR	=QUARTILE.EXC(B1:B313,3)-QUARTILE.EXC(B1:B313,1)	**4**
	Mean	=AVERAGE(B1:B313)	**9.58**
	Median	=MEDIAN(B1:B313)	**10**
Bachelors	Std Dev	=STDEV.S(D1:D313)	**3.08**
	Range	=LARGE(D1:D313,2)-SMALL(D1:D313,2)	**16**
	IQR	=QUARTILE.EXC(D1:D313,3)-QUARTILE.EXC(D1:D313,1)	**4**
	Mean	=AVERAGE(D1:D313)	**9.80**
	Median	=MEDIAN(D1:D313)	**10**
Nanodegree	Std Dev	=STDEV.S(F1:F313)	**2.56**
	Range	=LARGE(F1:F313,4)-SMALL(F1:F313,3)	**7**
	IQR	=QUARTILE.EXC(F1:F313,3)-QUARTILE.EXC(F1:F313,1)	**2**
	Mean	=AVERAGE(F1:F313)	**9.05**
	Median	=MEDIAN(F1:F313)	**9**

Appendix II: Notes on cleaning the Survey Dataset

The following steps were applied to clean up the sample dataset:

- Age – students without DOB and/or misstated DOB are excluded
- Avg. sleep time per night - misstated hours like 45,65,85,9141984 and no stated hours are excluded
- Avg. sitting time per day (hours) – incorrect hours like 50,56,60,88,200,540,610,720,800 are excluded
- No of books read per year – set blank value to 0
- Job Industry – set blank value to “Unspecified”
- Employer – clean up blank values
- Learning hours – clean up non numeric and invalid hour values
- Practicing hours – clean up non numeric and invalid hour values
- Likelihood to recommend – replace 0 value as 1