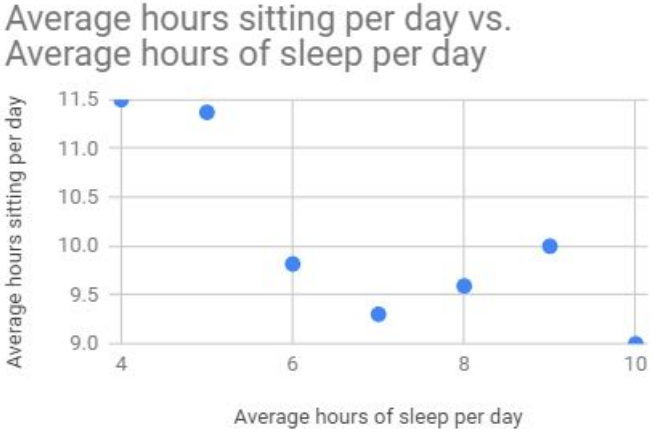
Project: Analyze Survey Data

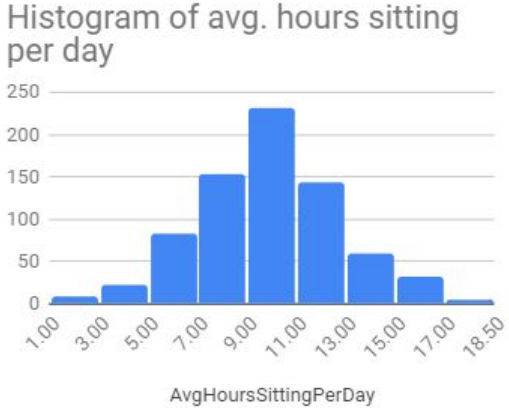
Hanna Kondrashova

Udacity Data Foundations

2018

Do respondents who sleep more hours per day spend more time on sitting also?

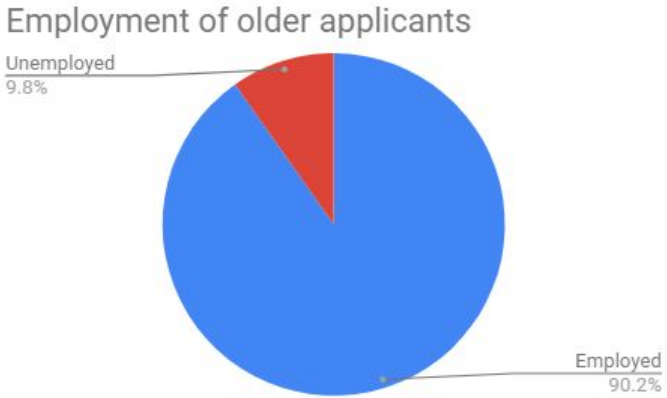


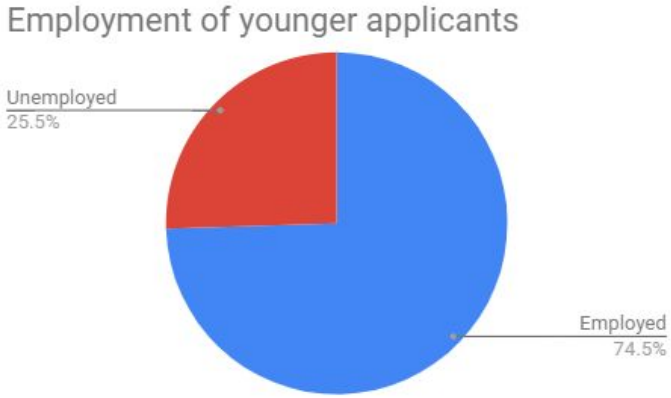


Mean	9.63
Stdev	2.95
Median	10
Mode	10.00

After removing outliers, 736 paired samples left. For each value of 'average hours of sleep per day' (Q1), correspondent values of 'hours sitting per day' (Q2) were grouped and averaged (each group contained >= 4 values). The result, approximation of E(Q2|Q1), shows that Q2 doesn't demonstrate any remarkable functional dependency on Q1. Besides that, Q2 seems to be ruled by marginal distribution (see histogram) having mean very near to median and mode (see table and Note 1), and standard deviation (2.95 hours) which explains scatter (2.5 hours) on the first figure (see Note 2).

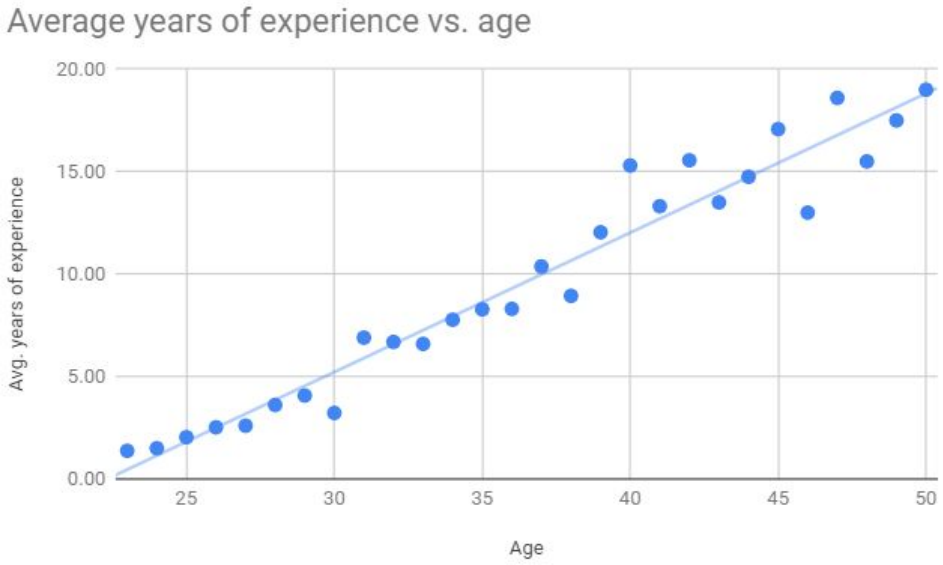
Do older respondents have less unemployment ratio than younger ones?





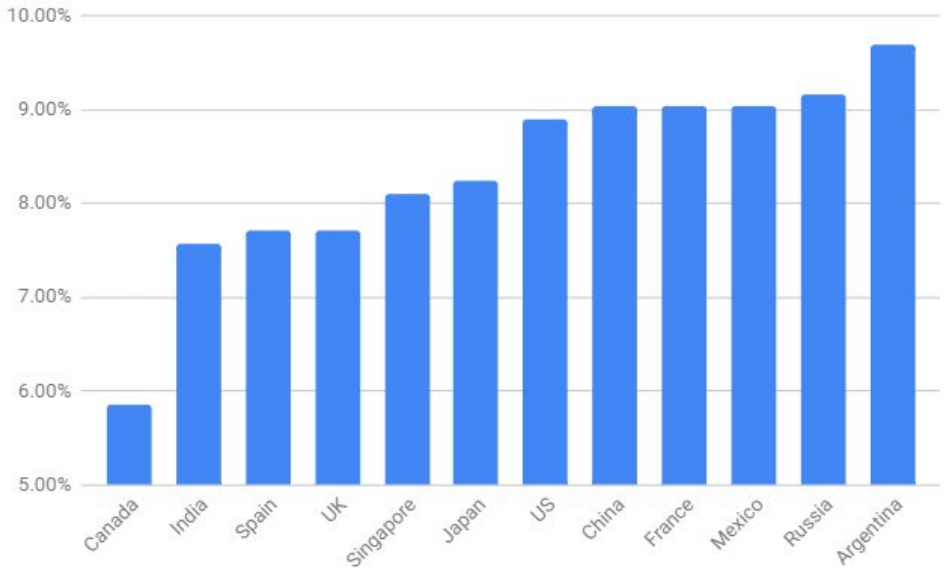
The median value of age after removing outliers (age of 1, 2 years and blanks) is 32 years, so the younger respondents are considered to be below this age, and older above and including it. The pie diagrams clearly depict affirmative answer (see Note 3): 9.8 percent of unemployed for older and 25.5 for younger.

Does experience (in years) of the respondents grow with age?



Removing outliers and inconsistent data (years of experience more than age, blanks and values of age of 1, 2 years, etc.) for calculating E(Q1 | Q2), the 'average age of experience' (Q1) conditional on 'age of respondents' (Q2), we have 543 pairs to analyze. Averaging for >=4 samples per each age (so that it is consistent) gives a clear affirmative answer on the given question (see Note 4) representing a linear increasing dependency shown on the picture (see Note 5).

Are there any prevailing countries/cities where the respondents come from?



The answer is negative: the diagram shows the ratio per each country, so that 12 countries are distributed between 5.94% and 9.69%, which is near 100/12 = 8.33%. Besides that, the countries except minimal and maximal values lay very near to each other (difference within 1.60%), so they represent distribution near to uniform, which means that a priory the respondent can be from any of those countries with equal chance.

Notes

1. The marginal distribution of Q2 seems to be normal, but one cannot state that unless doing a chi-square goodness of fit test (or similar). More insight shows that Q1 is also normal (with its own mean and stdev), so mutual PDF(Q1, Q2) seems to be factorizable in two normal PDFs, which means Q1 and Q2 are independant and will have zero correlation. This actually can be verified by chi-square test of independence. Those tests were not in course scope, so they are omitted.

2. According to the properties of normally distributed variables, the sample average of N i.i.d. normal variables has stdev of each of them divided by √N. We had the smallest group of 4, so the 2 stdev scatter will be 2*2.95/√4 = 2.95 hours, which corresponds to observed 2.5 hours on the drawing (it's smaller because actually there are more groups with larger N prevailing). One should notice, that stdev is actually biased here, so for the whole population it is calculated applying Bessel's correction (which is applied for the value in the table).

Notes

3. It must be taken in account, that this sentence is true only for represented data, not the whole population. Also higher age doesn't guarantee better chance of finding a job, so it's not causation. The median value used to classify the respondents is shifted to younger age, so for the older respondents less data is available, which means less accurate averaging.

4. Here there's definitely a strong correlation present, which in fact doesn't mean causation (you aren't guaranteed to have more years of experience just being older). Also not all the population is investigated here, and for higher age there's less data for averaging, which makes it less accurate. Besides that, there are other factors influencing on the years of experience, such as the graduation age. Also there should be *relevant* or *non-relevant* experience taken in account, which is not possible with the given data.

5. Strictly, there could be a real linear regression line found to describe this dependency. For that, the age shouldn't be rounded down to avoid having multiple values of Q2 with respect to single Q1, which makes it more consistent with the experiment and easier to handle, but it's harder to visualize.