# Data Analysis: A Practical Introduction for Absolute Beginners

## Lab 9: Education Data

### Learning Objectives

- Find the mean/average ranking among different variables in a data set.
- Create a line graph to visualize data.

### Data Set

Mod5Lab2.csv

### What You'll Need

To complete the lab, you will need the online version of Microsoft Excel.

### Overview

Imagine that you're a data analyst at a university, and you're looking at data from six different courses from one particular instructor: This instructor is teaching intro, intermediate, and advanced versions of two classes ("Regression" and "Machine Learning"). The data show the results of post-course surveys that each student took, where they ranked different aspects of the class on a scale of 0 to 5.

In this lab, you'll help the instructor see how the students ranked each course overall, how they ranked the teacher, and how each course stacked up against the others.

### Exercise 1: Class and Teacher Rankings

1. Open the data set in Excel, which shows the results of a post-course survey from 270 different students. Here's a snapshot of the data:

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | overall.course | overall.instructor | rigor | inspired | clarity | class | level |
| 2 | 1 | 3 | 4 | 3 | 4 | 4 | Regression | Intro |
| 3 | 2 | 1 | 1 | 2 | 1 | 2 | Regression | Intro |
| 4 | 3 | 2 | 2 | 4 | 2 | 2 | Regression | Intro |
| 5 | 4 | 4 | 4 | 3 | 3 | 3 | Regression | Intro |
| 6 | 5 | 5 | 5 | 5 | 4 | 5 | Regression | Intro |
| 7 | 6 | 4 | 1 | 2 | 2 | 2 | Regression | Intro |
| 8 | 7 | 5 | 4 | 5 | 5 | 4 | Regression | Intro |
| 9 | 8 | 3 | 1 | 1 | 2 | 3 | Regression | Intro |
| 10 | 9 | 4 | 3 | 3 | 3 | 3 | Regression | Intro |
| 11 | 10 | 2 | 3 | 2 | 2 | 2 | Regression | Intro |

Each row represents one student's survey results (there should be 270 different students). Here's what each column represents:

**overall.course** = the student's ranking of the course as a whole (on a scale from 0 to 5)

**overall.instructor** = the student's ranking of the instructor/teacher during that course (on a scale from 0 to 5)

**rigor** = the student's ranking of the course difficulty/exhaustiveness (on a scale from 0 to 5)

**inspired** = the student's ranking of how inspired they felt after taking the course (on a scale from 0 to 5)

**clarity** = the student's ranking of how clear the course materials were (on a scale from 0 to 5)

**class** = the title of the course (either Regression or Machine Learning)

**level** = the level of the course (either Intro, Intermediate, or Advanced)

The untitled column A values give the student number for ID purposes. Notice that these student numbers are slightly different than the actual row numbers — so Student 1 is actually in row 2 of the spreadsheet, because all the column titles are in row 1.

2. Now you want to look at the class averages from two of these variables: overall course and overall instructor. To do this, find the average (mean) ranking for "overall.course" and for "overall.instructor" — but break it down by each of the six class types. Those six classes are: (1) Regression Intro, (2) Regression Intermediate, (3) Regression Advanced, (4) Machine Learning Intro, (5) Machine Learning Intermediate, and (6) Machine Learning Advanced.

   Thankfully, the data are already arranged by course. Set up a new table off to the side of the data (in the same spreadsheet), with columns for each course, level, average overall course score, and average overall instructor score. It should look something like this:

| I | J | K | L | M |
|---|---|---|---|---|
| | **class** | **level** | **course avg** | **instructor avg** |
| | Regression | Intro | | |
| | Regression | Intermediate | | |
| | Regression | Advanced | | |
| | Machine Learning | Intro | | |
| | Machine Learning | Intermediate | | |
| | Machine Learning | Advanced | | |

3. Click into cell L2. In that cell, use Excel's AVERAGE function to find the average/mean value of "overall.course" (column B) for all the students in the class Regression (column G) and level Intro (column H). The syntax here is **= AVERAGE(first cell:last cell)**.

   Since only the first 45 students in the list took the Regression Intro course, your first cell is B2 and your last cell is B46. You can either type those cells directly into the AVERAGE function (with a colon in between), or just type in =AVERAGE( ), click inside the parentheses, and highlight the whole range of cells between B2 and B46. Either way works.


   Hit Enter and Excel will calculate the average ranking.

| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | |
| Regression | Intermediate | | |
| Regression | Advanced | | |
| Machine Learning | Intro | | |
| Machine Learning | Intermediate | | |
| Machine Learning | Advanced | | |

   There we go. The average score from the students in the Regression Intro course was a 3 (out of 5).

4. Now find the average instructor score for that same course (Regression Intro). Use the AVERAGE function again. This time, you want the ranking each student gave for the *instructor* (column C), so your cell range runs from C2 to C46.


Hit Enter.

| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | | |
| Regression | Advanced | | |
| Machine Learning | Intro | | |
| Machine Learning | Intermediate | | |
| Machine Learning | Advanced | | |

The average ranking for the instructor in the Regression Intro course was roughly 3.07 when rounded to two decimal places.

5. Run through the same steps for the Regression Intermediate course. You'll need to scroll down a bit to find the students in this class: Regression Intermediate runs from Student 91 to Student 135. Use the AVERAGE function again with cells B92 to B136 to find the course average. (Remember, the cell numbers are 1 digit off from the student number, so Student 91 is in row 92.)


Hit Enter again.

| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | |
| Regression | Advanced | | |
| Machine Learning | Intro | | |
| Machine Learning | Intermediate | | |
| Machine Learning | Advanced | | |

The students in the Regression Intermediate class gave the course an average score of about 4.67 (out of 5). Not too shabby!

6. Find the instructor average for that same set of students. This time, use C92 to C136 as your cell range, since you're looking at the "overall.instructor" data for the Regression Intermediate class.


| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | | |
| Machine Learning | Intro | | |
| Machine Learning | Intermediate | | |
| Machine Learning | Advanced | | |

The students in this course gave the teacher an average ranking of about 4.67 (out of 5).

7. Repeat those same steps for the other four courses, paying careful attention to the cell ranges and column names. Here's what you'll use for the **Regression Advanced** course:

Course average:


| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | 2.95555556 | |
| Machine Learning | Intro | | |
| Machine Learning | Intermediate | | |
| Machine Learning | Advanced | | |

Instructor average:

| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | 2.95555556 | 3.111111111 |
| Machine Learning | Intro | | |
| Machine Learning | Intermediate | | |
| Machine Learning | Advanced | | |

8. Here's what you'll use for the **Machine Learning Intro** averages:

Course average:

$f_x$ =AVERAGE(B47:B91)

| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | 2.95555556 | 3.111111111 |
| Machine Learning | Intro | 2.47777778 | |
| Machine Learning | Intermediate | | |
| Machine Learning | Advanced | | |

Instructor average:

$f_x$ =AVERAGE(C47:C91)

| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | 2.95555556 | 3.111111111 |
| Machine Learning | Intro | 2.47777778 | 2.366666667 |
| Machine Learning | Intermediate | | |
| Machine Learning | Advanced | | |

9. Here's what you'll punch in for the **Machine Learning Intermediate** averages:

Course average:


| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | 2.95555556 | 3.111111111 |
| Machine Learning | Intro | 2.47777778 | 2.366666667 |
| Machine Learning | Intermediate | 1.5 | |
| Machine Learning | Advanced | | |

Instructor average (don't forget to switch over to column C for this range!):


| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | 2.95555556 | 3.111111111 |
| Machine Learning | Intro | 2.47777778 | 2.366666667 |
| Machine Learning | Intermediate | 1.5 | 1.711111111 |
| Machine Learning | Advanced | | |

10. And finally, here's what you'll use for the **Machine Learning Advanced** averages:

Course average:

| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | 2.95555556 | 3.111111111 |
| Machine Learning | Intro | 2.47777778 | 2.366666667 |
| Machine Learning | Intermediate | 1.5 | 1.711111111 |
| Machine Learning | Advanced | 3.44444444 | |

Instructor average:

$f_x$ =AVERAGE(C227:C271)

| J | K | L | M |
|---|---|---|---|
| **class** | **level** | **course avg** | **instructor avg** |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | 2.95555556 | 3.111111111 |
| Machine Learning | Intro | 2.47777778 | 2.366666667 |
| Machine Learning | Intermediate | 1.5 | 1.711111111 |
| Machine Learning | Advanced | 3.44444444 | 3.622222222 |

11. Now you can see how each course stacked up against the other courses. The students seemed to like the Regression Intermediate course the best, both in terms of the course content itself and the instructor.

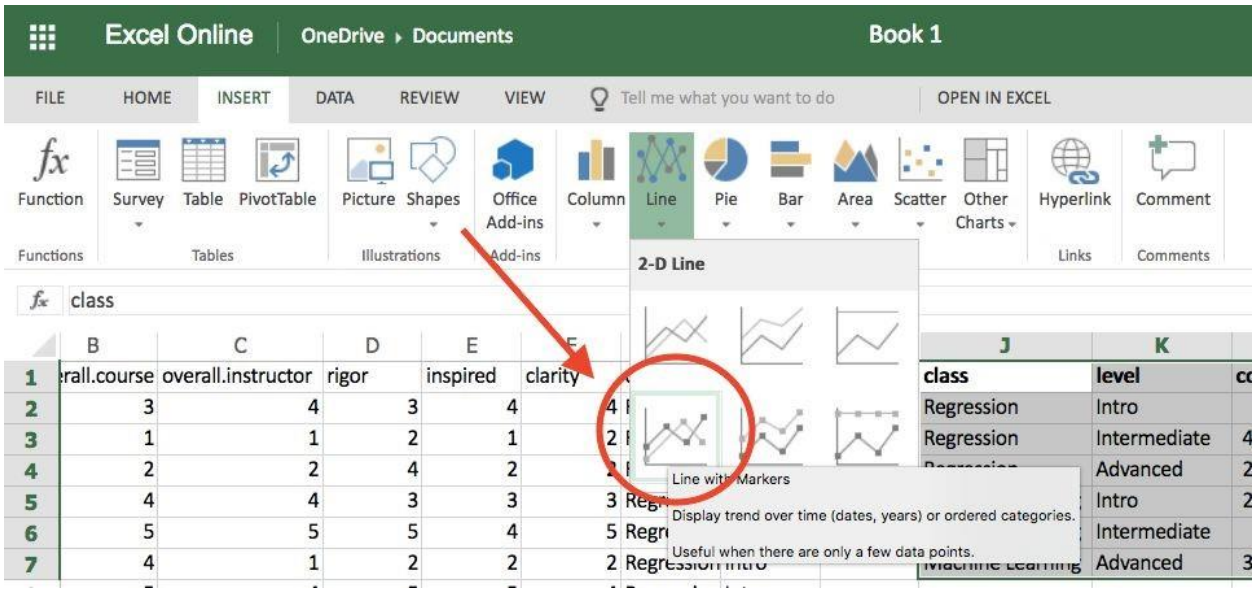
## Exercise 2: Visualizing Those Rankings

Now let's create a quick graph of all those average rankings we just found.

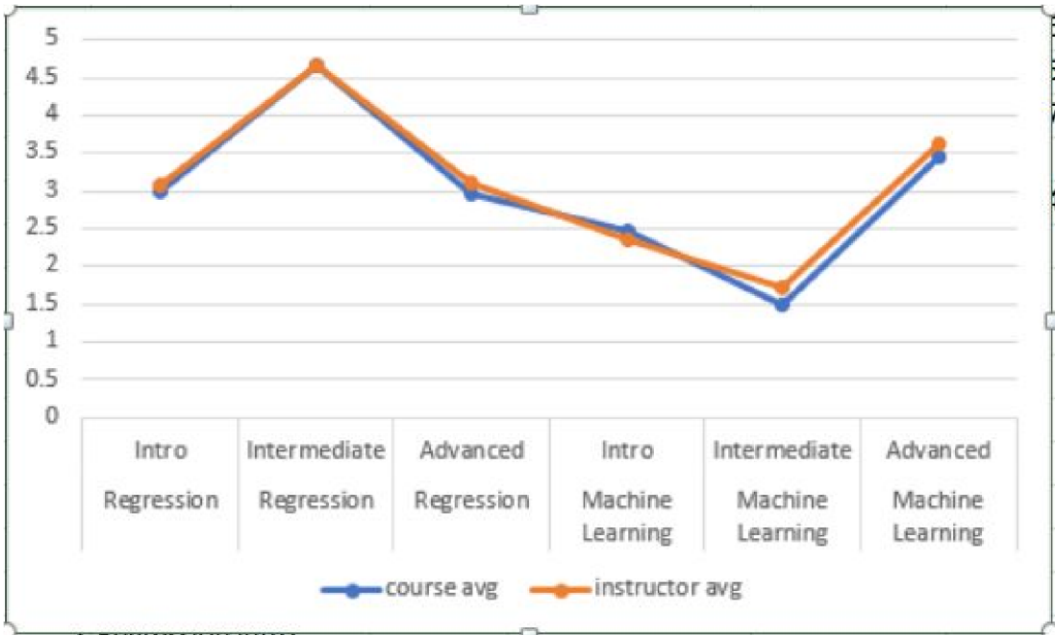
1. Highlight the entire table of new data that you created in Exercise 1.

| J | K | L | M |
|---|---|---|---|
| class | level | course avg | instructor avg |
| Regression | Intro | 3 | 3.066666667 |
| Regression | Intermediate | 4.66666667 | 4.666666667 |
| Regression | Advanced | 2.95555556 | 3.111111111 |
| Machine Learning | Intro | 2.47777778 | 2.366666667 |
| Machine Learning | Intermediate | 1.5 | 1.711111111 |
| Machine Learning | Advanced | 3.44444444 | 3.622222222 |

2. Click on the Insert tab in the ribbon. In the Charts section, click Line > Line with Markers (it's usually the icon in the lower-left corner).




3. You should now see a graph that shows both the course and instructor averages for each of the six courses. Like so:




Notice how one color (blue in our example) represents the course averages, and the other color (orange) represents the instructor averages. Now it's easier to see how these averages relate to one another.

For example, the students clearly gave the best rankings to the Regression Intermediate class. Now the instructor can use this data analysis to figure out what they did better in that course, and try to replicate it in their other courses.

It also seems likely that the instructor poured more effort into the Intermediate version of the Regression course, possibly at the expense of the Intermediate version of the Machine Learning course. Maybe the instructor was teaching both of the Intermediate courses at the same time and neglected the Machine Learning side a bit.