How to Make and Interpret a Simple Plot

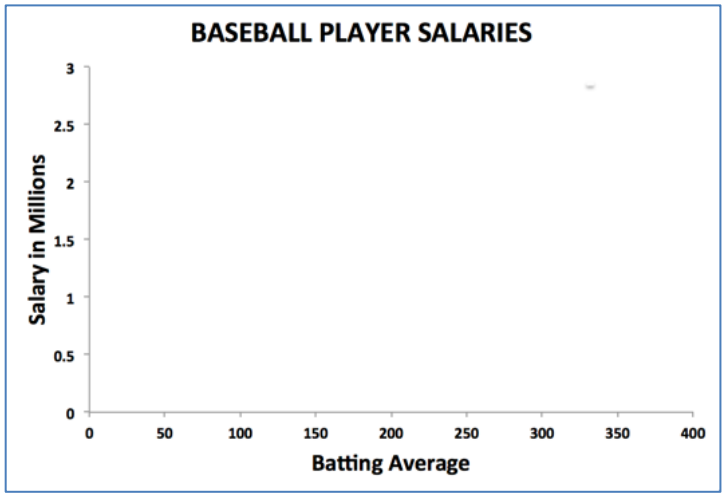
Some of the students in ASTR 101 have limited mathematics or science backgrounds, with the result that they are sometimes not sure about how to make plots that demonstrate the existence of *a relationship* between various measured or quoted values, or how to interpret such a relationship. This short guide will help you with that tool. (Many of you will not need this advice, of course.)

Here is a simple example. Let's consider a hypothetical set of four professional baseball players of varied skills and levels of accomplishment. Suppose we find the following information about their batting averages and their pay (see the table just below; the numbers are all *made up,* and only the first two of the players are or recently were actually in the big leagues):

Player	Career batting average	Baseball salary (in millions!)
Jose Bautista	320	2.85
Albert Puholz	290	1.76
Joe Smith	170	0.80
Dave Hanes	050	0.02

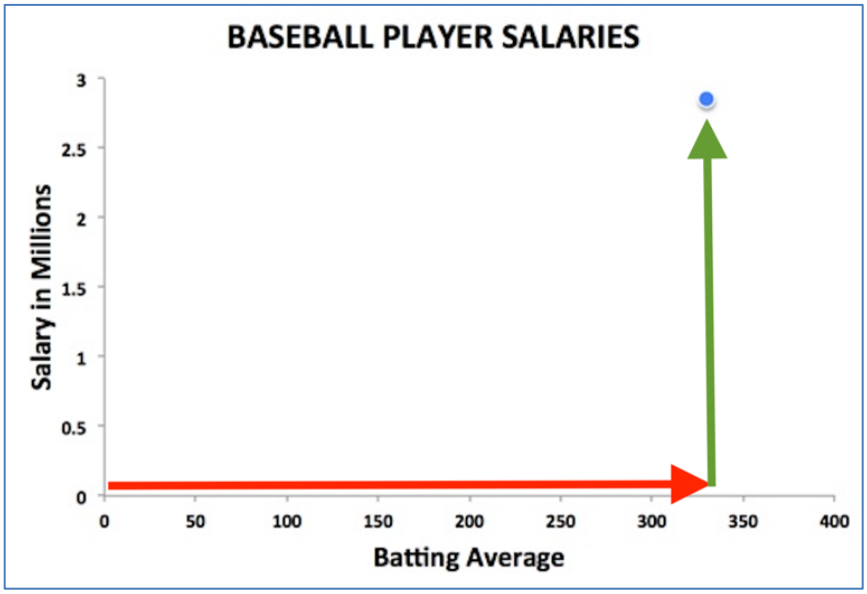
What you notice right away is that there is *a very strong correlation:* the player with the outstandingly best batting average has the highest salary – deservedly so! – while the least skillful player (Hanes!) has by far the lowest baseball salary, a meagre $20,000. But suppose you were signing up a fifth player, one who is of middling skill, with a batting average of (say) 225. You might like to know what reasonable salary to offer that player, to see that he is not way out of line with the others. The easiest way to figure this out is *first to plot the data – that is, see how one set of numbers depends on the other*. This will tell you , in a nicely visual way, how strongly a player's salary generally depends on his batting average.

In preparing to make the plot, it is useful first to look at the *range* of each set of numbers. The table shows that the batting averages lie between 0 and 400, while the salaries (in millions of dollars) range from nearly 0 to almost 3. This means that *scales* of the two sides of the plot (graph) will need to be very different, like so:

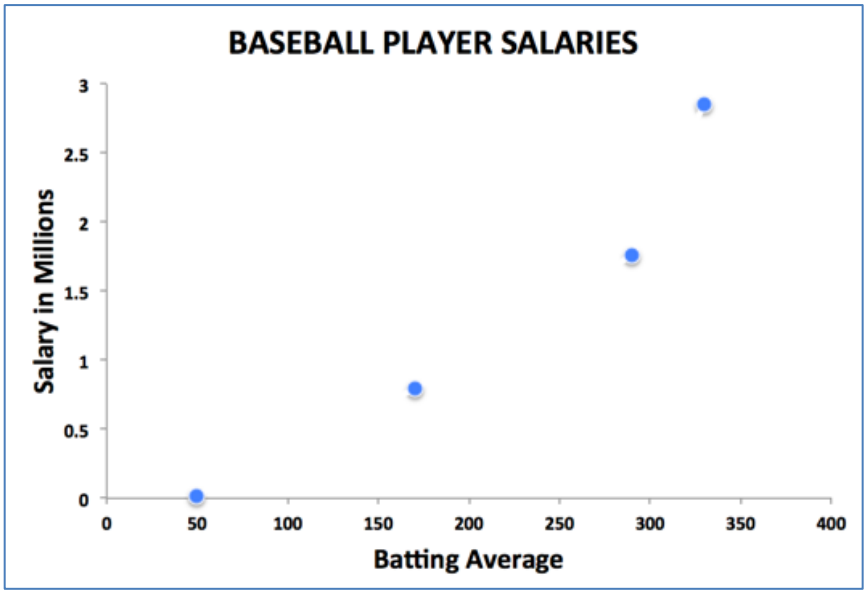


Note the important fact that in making this plot, we make sure that the numbers are *uniformly spread out* along each axis. This is very important.

Now to *plot* the data. Start with Jose Bautista, who has a batting average of 320 and a salary of 2.85 million. Move to the *right* (the red arrow, in the figure below) until you reach a value of ~320 along the horizontal axis; then move *up* (the green arrow) until you are opposite a salary of 2.85 (million). Make a mark there, as shown. The location of this symbol represents the correspondence between Jose's performance and his pay.



In the same way, you now have to add a plotted point for each of the other players, – the rest of the information you have. Here is what you will get:



There are a few things to note:

- Right away, you notice a strong *correlation*: points that lie farther to the right (high batting averages) also lie higher up (bigger salaries). This tells us at *a glance* that more productive players earn higher salaries, which is hardly surprising, and gives us a quick sense of how strong that dependence is. The important thing is that the plotted points do not lie at *random locations.*

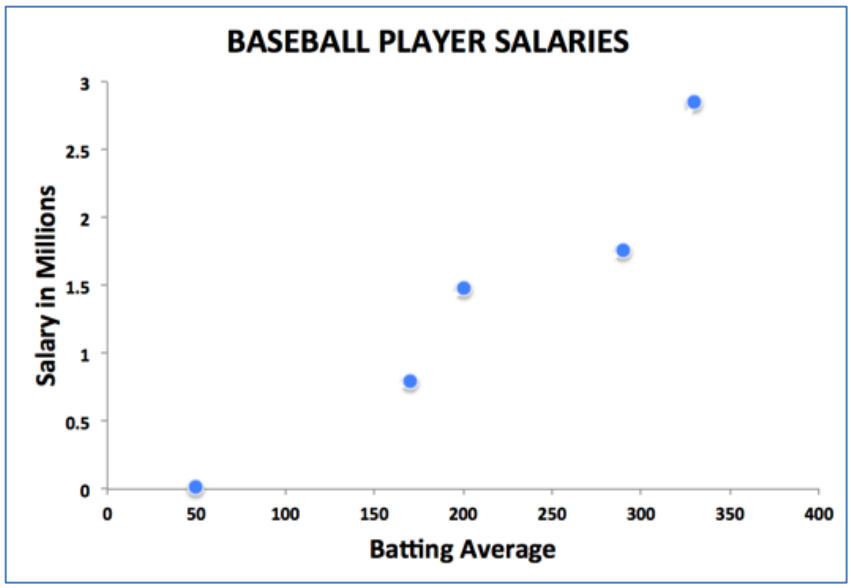
- *It could have been otherwise!* Suppose that a player's salary was chosen completely at random (say, by drawing a card out of a deck, or if a superstitious manager decided to base a player's salary on his date of birth, regardless of his skills). In that case, even a weak player might be very highly paid, and we would not expect to see a strong correlation. The plotted points could be all over the place, in what is known as a *scatter diagram.*

- Look again at the plot at the bottom of the previous page. You can see that the observed relationship is *not linear* – that is, you couldn't draw a single straight line that passes through or very near all the plotted points. In many circumstances, there is no obvious reason that the relationship *should* take that simple form – especially here, where I have simply made up the numbers, completely out of my head! But in scientific studies we sometimes do see a linear dependence, which tells us that the fundamental relationship between the plotted quantities is particularly simple.

- Here's a rather silly example of that: suppose that you cut pieces of completely random length off a long uniform piece of lumber. In that case, you'd expect the weight of each piece to depend on its length: if piece A is *twice the length* of piece B, it will *weigh twice as much*, and so on. If you make measurements of length and weight of the pieces of wood, and then draw up a plot, you will necessarily see a nicely *linear* relationship.

- Looking back at our plot of baseball player salaries, however, you will note that even though the relationship is not linear, it is fairly *smooth* – not up-and-down all over the place. (Please remember that I made up these numbers, so in this case, it is by my design.) In real life, however, we don't always expect to see a smooth correspondence because there may be other factors at play that may lead to a considerable *scatter* in the relationship.

To understand this last point, imagine now adding a new shortstop to your baseball team – a real superstar. He is admittedly a mediocre hitter, batting only 200, but he is an outstanding defensive fielder, so he is *absolutely essential* for the success of your team. For that reason, he will be earning 1.48 million dollars. When we add him to the team, and his information to the plot, we see this:



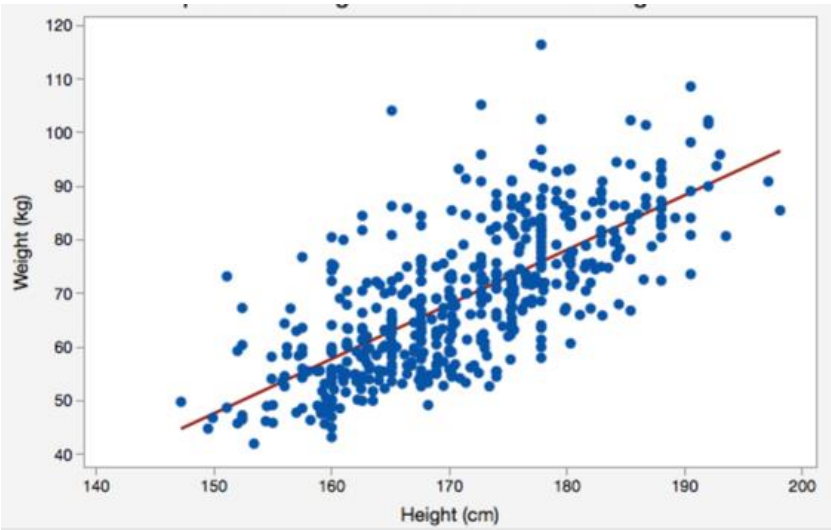
The correlation is still strong, but it is now a bit more 'jumpy', reflecting the fact that there are other considerations (not just batting average) that go into determining salary.

In this case, the scatter is *real* – that is, we know the salaries and batting averages precisely, and can trust the correctness of the plot. On the other hand, the observed scatter in a plot may simply reflect *errors in the original measurements* -- in other words, *how good the data are!* Sloppy measurements, or misreported batting averages, are going to give a messy plot, with the points sprinkled here and there.

Here's an example: reconsider the example I gave above, that of cutting pieces of wood of random lengths off of a long plank. You might take pains to measure the lengths very carefully, using a tape measure; but suppose you 'weigh' them only *very approximately*. For instance, you might bounce each piece up and down in your hand and merely 'guesstimate' how much it weighs. The plotted data points will consequently be fairly uncertain, and as a result the nice smooth relationship you might have expected will be 'smeared out' in your plot.

To reinforce that point, let's look at some *real data* for a moment – in this case, a plot that shows the relationship between the heights and weights of a sample of adults. (See the next page.) I cut and pasted this diagram from the web, and don't know exactly where it comes from. Let's imagine for the moment that it represents the information gathered from the incoming first-year class at Queen's University.

As you can see, in this case there is a *very strong correlation*: it tells you, not surprisingly, that *tall people are generally heavier than short people.* This is simply *because they are bigger overall.* (The red line shows the general trend, and it was presumably determined by some statistical analysis of the dataset.) But there is quite a lot of scatter. Why?



It could be that there is actually a *very tight correspondence,* and if we had made super-careful measurements, maybe the data points would have fallen *exactly along the red line*. That fundamental relationship might have been blurred, however, if we had the bad luck of hiring a really careless assistant, someone whose measurements were very sloppy. In that case, the scatter of points would be largely attributable to random errors, in one or both of the measured attributes (height and weight).

As you know from real life, *that's not the case here*. (The data are probably quite precise and trustworthy.) Other factors, however, have an important influence: most importantly, a person's build or muscularity, so that two people of the same height may differ a lot in weight. Notice, for instance, the point right at the very top of the graph: it represents a person of moderate height (180 cm) but quite heavy weight (117 kg or so) – in other words, someone with a very stocky stature. It would probably be wrong to suspect that the point shows up there because of shoddy measurements or bad record-keeping by the investigator!

Such graphs (plots) can be used in various interpretive ways. For instance, if you learn that your cousin is 175 cm in height, you might reasonably conclude that she *probably* weighs about 70 kg (as inferred from the red line – the 'general relationship' – in the figure above). On the other hand, the intrinsic scatter in the distribution means that you would not be surprised if her weight was somewhat more or less.

Now let's look back at the plot for our imaginary baseball team. Suppose we decide to sign up a new player of average skills (say, with a 225 batting average). As a starting point, a simple glance at the plot suggests that it might be reasonable to pay him something in the range of 1.25 million dollars. (Of course, his agent will try to persuade you that he has special skills that justify a higher salary.). That is an example of how a schematic depiction of the relationship can guide your understanding and decision-making.

Cautionary Notes when Making a Plot

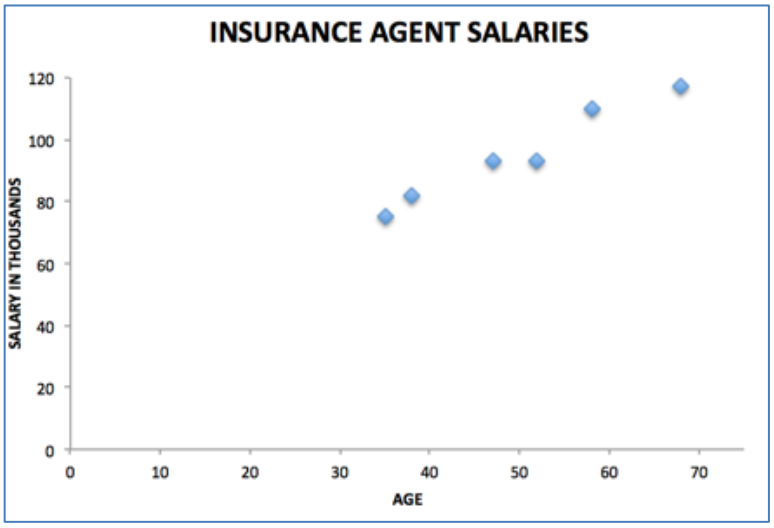
In Virtual Lab 2, when studying the orbits of the Galilean moons of Jupiter, you are asked to make a couple of plots. You have seen above how this is done, but there are a couple of special things to note. The second of these can be a real pitfall, if you are not careful.

Point #1: Choosing Appropriate Scales

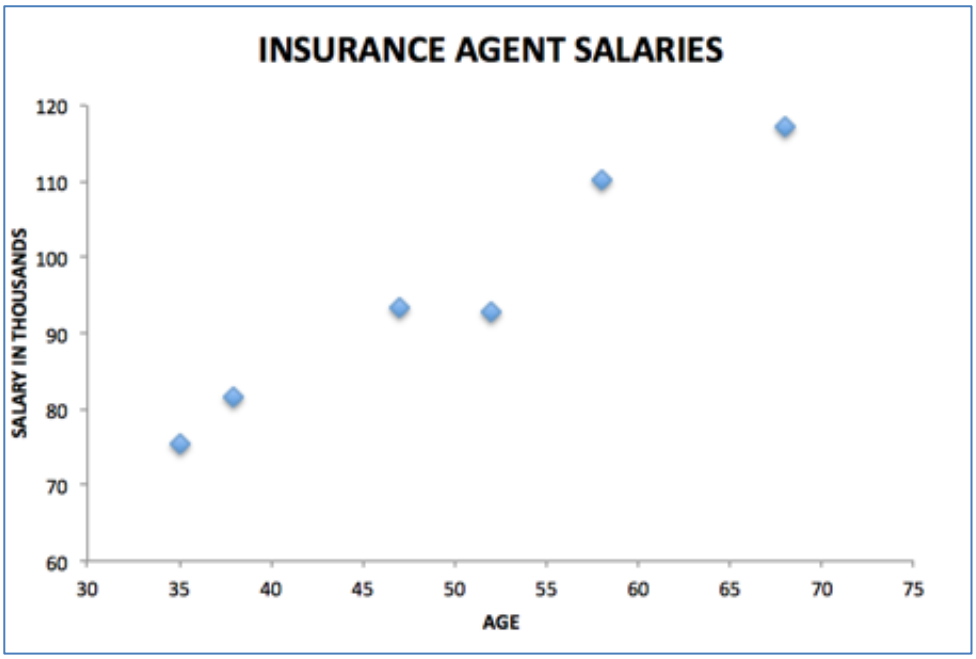
The data you are going to plot may not always start near zero! Suppose, for instance, you want to explore how the salary of a typical independent insurance agent depends on her or his age. The very youngest agent you encounter might be 35 years old, say, and you might get values like those in the table below. Once again, *I am making these up.* I have no idea what an actual insurance agent earns, or if age is a dominant factor (although I expect that experience plays an important role)

Agent	Age	Salary (in thousands)
Smith	35	75.3
Jones	38	81.7
Brown	47	93.3
Doe	52	92.9
Miller	58	110.1
Green	68	117.3

Given these data, it might seem logical to make a plot that shows ages up to 75, and salaries up to 120K, both starting from zero. If you do so, it will look like this:



Although this is okay, the data points you have plotted all lie towards the upper right of the diagram– there is a lot of 'wasted space' in the graph. It would be *cosmetically* better to adopt different scales, using one that runs from 30 to 75 for the agent's age and another from 60 to 120 (thousands) for the agent's salary. The plot you get will now look like this:



You can see that it uses 'the whole page' much more effectively than before. (For a 'real world' example of this point, look at the plot I showed a couple of pages ago – the weight versus height of adults – and consider the scales that were adopted in making that plot.)

By the way, you may note that in the particular example of the insurance agents, the plot above seems to reflect a roughly *linear* relationship: a straight line passing through the set of points from lower left to upper right would be a reasonable representation. But this is by accident! (Remember that I made up the numbers.)

Point #2: How to Use Excel

We sometimes advise you to "use an Excel spreadsheet to make a plot." That's fine, ***if you know how!*** Unfortunately, it is very easy to get it wrong, for a simple reason.

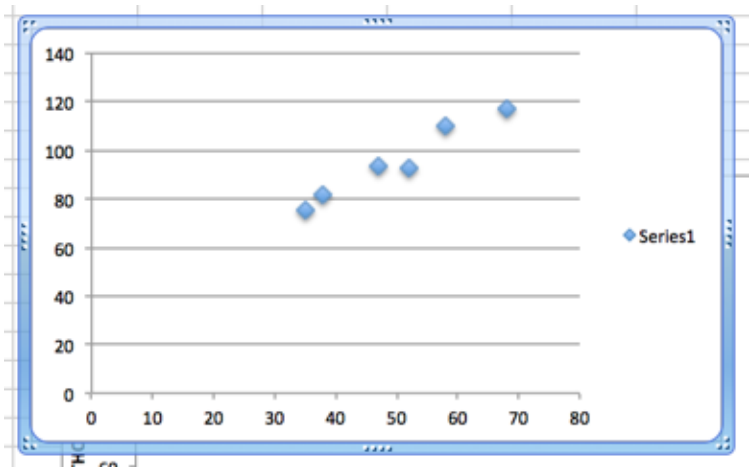
Here is why. Let's begin by entering the data for our hypothetical insurance agents (from the table two pages back) into an Excel spreadsheet, as shown here:

	A	B	C
1		Age	Salary (thousands)
2	Smith	35	75.3
3	Jones	38	81.7
4	Brown	47	93.3
5	Doe	52	92.9
6	Miller	58	110.1
7	Green	68	117.3
8			

We want to understand how the entries in column C (the salaries) are related to those in column B (the ages). To demonstrate that, select the cells from B2 though C7, as shown here:

	A	B	C
1		Age	Salary (thousands)
2	Smith	35	75.3
3	Jones	38	81.7
4	Brown	47	93.3
5	Doe	52	92.9
6	Miller	58	110.1
7	Green	68	117.3
8			

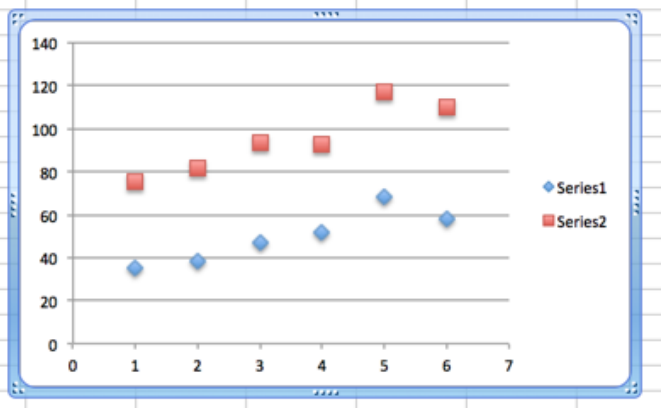
and then press "Chart", followed by "Scatter" and "Marked scatter" in the menu bar across the top of the Excel window. You will get the figure shown on the next page:



and can then format it to suit your tastes. ***Well done!*** This is just what we want. (You can even adjust the range of the scales as well, to 'fill up' the page in the way we saw earlier.). ***But beware!*** If you mistakenly select *all the columns* in the rows that have data, like this:

	A	B	C
1		Age	Salary
2	Smith	35	75.3
3	Jones	38	81.7
4	Brown	47	93.3
5	Doe	52	92.9
6	Green	68	117.3
7	Miller	58	110.1
8			

and then hit "Chart" as before, you get something quite unexpected, looking like so:



You have indeed plotted the *data, but not quite in the way intended.* What does the graph show? Well, each one of the numbers across the bottom (1 to 6) represents one of the six rows of data (that is, number 1 is for agent Smith, number 2 is Jones, etc), while the coloured points above each number shows the actual data: the blue symbols represent their ages; the red symbols represent their salaries. (Look directly above number 3, for example, to learn that agent Brown is in her 40s and earns about $95K yearly.) Unfortunately, this sort of representation does not allow you to see *very clearly and directly* how salaries depend on ages.

It is true, in this case, that you can get a *sense* of the relationship. Note, for instance, that agent 5 (Mr Green) has both an advanced age (his blue point is higher than the other blue ones) and also a relatively good salary (his red point is likewise higher than the other red points). Meanwhile, agent 1 (Ms Smith) is relatively low in both respects. But the actual relationship does not jump out at you in the very clear way that you will see if you actually plot *one set of numbers directly against the others,* as we did on the previous page.

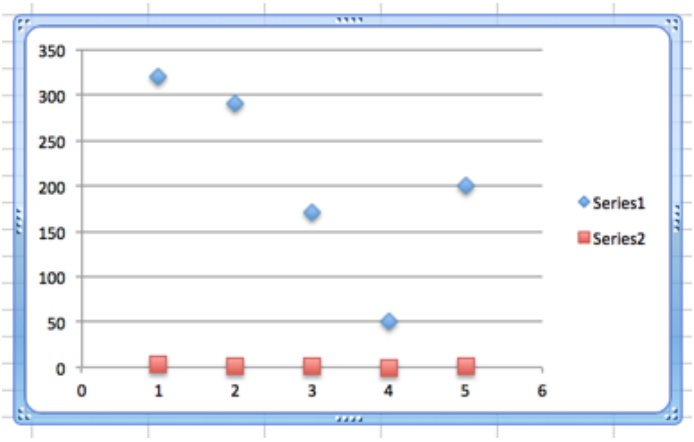
As a matter of fact, you were actually *a bit lucky* in this case because the ages of the agents (numbers in the range 35-70) are not very far removed from their salaries (numbers ranging between 75 and 117, in thousands of dollars). This means that the numerical scale shown on the left of the plot just above applies to both, and with a bit of effort you can interpret what's going on. To see a worst-case scenario, however, let's go back to our baseball team (including our superstar shortstop). We now have

	A	B	C
1		Average	Salary
2	Bautista	320	2.85
3	Puholz	290	1.76
4	Smith	170	0.8
5	Hanes	50	0.02
6	Shortstop	200	1.48

and if we select the data the 'wrong' way, including the column of names like so:

	A	B	C
1		Average	Salary
2	Bautista	320	2.85
3	Puholz	290	1.76
4	Smith	170	0.8
5	Hanes	50	0.02
6	Shortstop	200	1.48
7			

then our “chart” command yields this figure:



Once again, the numbers across the bottom correspond to the five players. The blue symbols reflect their various batting averages, and the red dots are their salaries in millions. (Hanes, with the abysmal batting average and pathetic salary, is number 4.) The problem here is that the scale on the left goes from 0 to 350, as required by the need to accommodate the range of batting averages, but the salaries (in *millions of dollars*) are all very small, with values like 1 or 2. For that reason, the red symbols all lie at the very bottom of the figure, and *we can’t even tell which salary is the largest and which is smallest.* There is no hope of determining the relationship between the actual performance and the pay of a player from such a plot.

Summary: if you want to assess, interpret and use any relationship that may exist, ***you must be sure to plot one set of data directly against the other.*** This is ***essential.*** In this document, we have explained exactly how to do so, either by hand or using a tool like Excel.

Feel free to contact us if there are any aspects on which you need clarification.