
How PayScale Uses Data Science to Predict Compensation

It All Starts With Data

Like all providers of compensation information, at PayScale our success relies upon the quality and quantity of our data. We gather it from visitors to our site. They come for many different reasons but mostly to:

- **Prepare to ask for a raise.**
- **Find out how their salary compares to others in similar job positions.**
- **Evaluate a job offer.**

During their visit they share data about themselves in exchange for whatever information they need. Call it an old-fashioned barter, or call it newfangled crowdsourcing, it allows us to get our compensation data from the most accurate, detailed source available: employees themselves. The information is gathered through a survey that takes just minutes to complete and that adapts to the individual as they tell us more about themselves.

Sifting for Gold

Like any enterprise depending upon human-generated data, we can't stop with the raw ore. It's our job to sift out the good material and process it for consumption. At PayScale, we've spent a lot of energy thinking about the best way to do this. In the most simple of terms, the following is what we do:

- **Weed Out the Outliers:** Is a data point far outside what's expected for a job? We discard data that doesn't make sense.
- **Defend Against Attempts to "Stuff the Ballot Box":** Not that we see much evidence of this, but we automatically detect and reject too much data coming from any one person.
- **Standardize the Data:** You say "computer programmer." I say "software developer." We have the technology to realize we're talking about the same job.
- **Augment the Data:** PayScale knows a lot more than just what we receive from our surveys. For example, if you tell us you're in Kuna, Idaho, we know that is in the Boise, Idaho metropolitan area. Plus, we know that the last census placed 606,376 people in the Boise area, and how being in a city of that size affects compensation.

All told, and after sorting out what we don't want and keeping just the confirmed data, PayScale has more than 35 million profiles – and counting. That is a lot of useful data.

Building a Better Mousetrap

Gathering a compensation data set that is both broad and deep is only half the battle. What makes PayScale different from its competitors, who rely on less specific, employer-submitted data, is that we use modern data mining and predictive modeling techniques to yield the highly accurate compensation predictions possible, across a huge range of circumstances.

Connecting the Dots

All data sets have gaps. Maybe you want to know how a software developer with seven years of experience programming in Java in Harlan, Kentucky should be paid. What if PayScale doesn't have many profiles that fit that description? What if there has never been a programmer like that in Harlan? How does PayScale make a prediction?

PayScale's MarketMatch™ algorithm uses advanced mathematics and statistics to answer these very questions. Using our huge database of individually entered data, we can determine how software developers are paid, in general. More importantly, we can adjust that prediction to account for having seven years of experience, for being in a small town in the upper South, and for programming in Java rather than in C++. Without an algorithm that knows how to make these adjustments, one would just have to say, "The data to price that job doesn't exist." With our sophisticated mathematical approach, we can make a prediction in which you can have confidence.

Making the Ideal Match

Traditional, employer-submitted compensation data providers allow you to specify only a few compensable factors when pricing a job position. PayScale allows you to create a much more detailed description of the job you are trying to price. Our MarketMatch™ algorithm weighs dozens of details to make very accurate pay predictions. And, it also allows us to find the people in our vast data set that are most like the one you've described according to the factors that matter most in determining compensation.

It's like dating: what matters most to one person may be inconsequential to another. At PayScale, we know that most employers pay more for employees that have more experience or who are located in large urban areas. We understand those relationships, but we also know the exceptions.

For example, family doctors often make more money in the most rural areas because of the scarcity of doctors who wish to work there. Also, their pay is often fairly steady, excluding cost-of-living increases, from the beginning of their careers until they retire. This, and much more hard-earned wisdom, is encoded in MarketMatch™. When PayScale provides anonymized comparison data so you can see examples of the data we actually use for predictions, we find the data that matters for pay, whether that's showing you data that is near your location or that has the extra skill that is really driving compensation for the position you are pricing.

Putting It All Together

Historically, compensation decisions involved some guesswork; you found the best comparison data points you could and then you made ad hoc adjustments to price the position in question. These days, however, compensation analysis should be a science, not an art. But it takes the union of deep data and clever statistical modeling to end up with the right answers. We take both the data and the statistics seriously at PayScale – it's what we do.