Context:

Recently, Stack Overflow had some interesting findings about the salaries of programmers who used spaces and programmers who used tabs. Namely, about which particular IDE setting corresponded with higher paychecks. This information was collected from their 2017 Annual Developer Survey. This raised a few questions about how much using spaces vs using tabs actually affected someone's paycheck, and whether or not it was an effect that was strong enough to prompt a change in behavior for aspiring or current developers. Our cohort decided to investigate these findings separately.

One member of the cohort decided to take the data as prepared and perform additional analyses on it to find out what other insights might be gleaned from further investigation. The other went the route of looking at the raw data and preparing it in a way that didn't artificially remove variables before analysis.

The Raw Data Processing Route:

Actions Taken:

The original dataset had 51392 observations of 154 variables. The schema of the 2017 survey indicated that not all questions were asked to all participants, and this was also taken into consideration when processing the data.

First the data was restricted to those who responded who identified as professional developers and as employed at least part time (as a contractor, freelancer, full-time worker, or self-employed) . It was then further restricted to those who answered the question of tabs vs spaces, those who answered the question about salary, and variables for questions not posed to current professional developers were removed. Questions about respondents' use of Stack Overflow were also removed.

Next, the remaining data was separated out into indicator variables in order to fit a linear model to it to explore the relationships between the variables.

Due to low response rates, an indicator variable was added to the following variables to indicate non-response, and then NAs were replaced with average values for the variable: Career Satisfaction, Job Satisfaction,

The resulting data frame of 12426 observations was fitted to a cross validated linear model. The magnitude of the coefficients were examined and then passed into a correlation matrix.

Results:

Country of origin was by far the strongest predictor of salary. Followed somewhat distantly by an unwillingness to answer the question about job satisfaction. After that, the largest predictors were either having coded for a job for 20 years, or – negatively predictive – having only coded for a job for 1-2 years. Feeling like one is greatly underpaid was also predictive of Salary, followed by Career Satisfaction.

Whether a developer used spaces vs tabs was 60th in terms of predictive power when it came to variable coefficients and their magnitudes.

Results from the correlation matrix indicated that working in the US and having programmed for 20+ years are very highly correlated with making a higher salary. Having coded for a job for 20+ years and working for a publicly traded corporation were also highly correlated with salary. Following that was not having attended university and being white.

Finally, while solidly correlated with salary but not highly predictive, came using spaces, working for a company of 10+k employees, prioritizing getting things done in hiring new employees, career satisfaction, and valuing retirement benefits.

Comparative analysis by country and year

Actions Taken:

Different methods were used to investigate how the relationship between salary and spacing choice varied by country. The 2017 dataset, this time restricted to the predictors used by Robinson, was compared with the 2015 survey results. The 2017 data included respondents' choice of spaces or tabs, salary, nationality, years of experience, coding language, educational background, and company size. The 2015 results included respondents' choice of spaces or tabs (the only year in which this question was asked, other than 2017), salary, nationality, age, gender, years of experience, and educational background. The data was restricted to responses from professional developers for the purposes of comparison. For this survey, the question about tabs and spaces included a "Huh?" option, whose respondents were omitted from the analysis. For both years' datasets, respondents who said either "tabs" or "it depends" were lumped into a single group to simplify

analysis, since the effect of spacing choice upon salary was approximately similar for both those groups in Robinson's analysis.

The first type of analysis utilized the Akaike Information Criterion to investigate whether spacing choice was a useful variable to include in a model for predicting salary. First, the AIC was computed for the data from both 2015 and 2017 for each country. Then the AIC was recomputed once for each predictor variable with that variable being omitted. The differences between the values for the original and modified datasets yielded the information gained by including each variable in the analysis.

Unregularized linear regressions were also used to provide a direct measure of statistical significance for each country.

Results:

The comparative analysis revealed that the unexplained impact of spacing choice upon salary varied significantly by country. For 2017, the differential AIC from spacing choice was high for the UK, US, and India, being ranked 2nd, 3rd, and 2nd respectively out of more than sixty variables, with the absolute changes in AIC being 20.5, 40.1 and 7.1. But the importance of spacing choice is not as great as it sounds since most of those variables were binary indicators for specific educational levels and coding languages. The differential AIC was also low or negligible for all the other countries, ranging between -2 and 1. Years of experience was consistently ranked as the most important predictor for salaries; the second-most important in the U.S. was PHP specialization.

The 2015 data revealed a slightly different pattern: spacing choice was the second-most important variable (out of ten) for the US and UK, but insignificant for all others, including India.

The unregularized regressions revealed that spacing choice was a statistically significant predictor ($p<0.05$) of salary in the 2017 data for the US, the UK, India, and Poland, but less significant or insignificant for Germany, France, Canada, and Australia. For 2015, spacing choice was a statistically significant predictor of salary for the US and the UK, but less significant or insignificant for other countries.

Conclusions:

Spacing choice versus other factors

Spaces vs Tabs has a weaker predictive effect than the standard items one might expect to affect salary. It's no replacement, for example, for decades' worth of experience coding. It was also rather telling that the next highest predictor was a disinclination to report job satisfaction, and that feeling greatly underpaid had a larger effect on salary than whether one used tabs or spaces.

Choice of spacing method is a robust predictor of salary in the U.S. and U.K., but the effect is weaker, unreliable or nonexistent in other countries. This suggests that the explanation for the effect is sociocultural or may be related to nation-specific industry practices. However, it does not imply that it is a wholly Anglosphere phenomenon, since Australia and Canada were also included in the analysis.

The fact that there was a strong effect for developers in India in 2017 but not in 2015 is puzzling. The number of Indian respondents was relatively high in both years.

The survey collection methodology introduces a handful of biases (such as voluntary response and non-response biases) into the equation. This can be seen readily by looking at the fact that not responding to the Job Satisfaction or Career Satisfaction questions had much higher coefficient magnitudes assigned to them than tabs vs spaces. Which might lead one to posit that attitude about one's work would also have a strong effect on one's compensation for it.

There is some sort of measurable effect that seems to be captured by the spaces vs tabs question, however its actual real-life effect is somewhat questionable. So, while it's an interesting variable to look at, perhaps the first advice one might give a new developer, after having looked at this data, would not be "use spaces instead of tabs, and you'll go far."