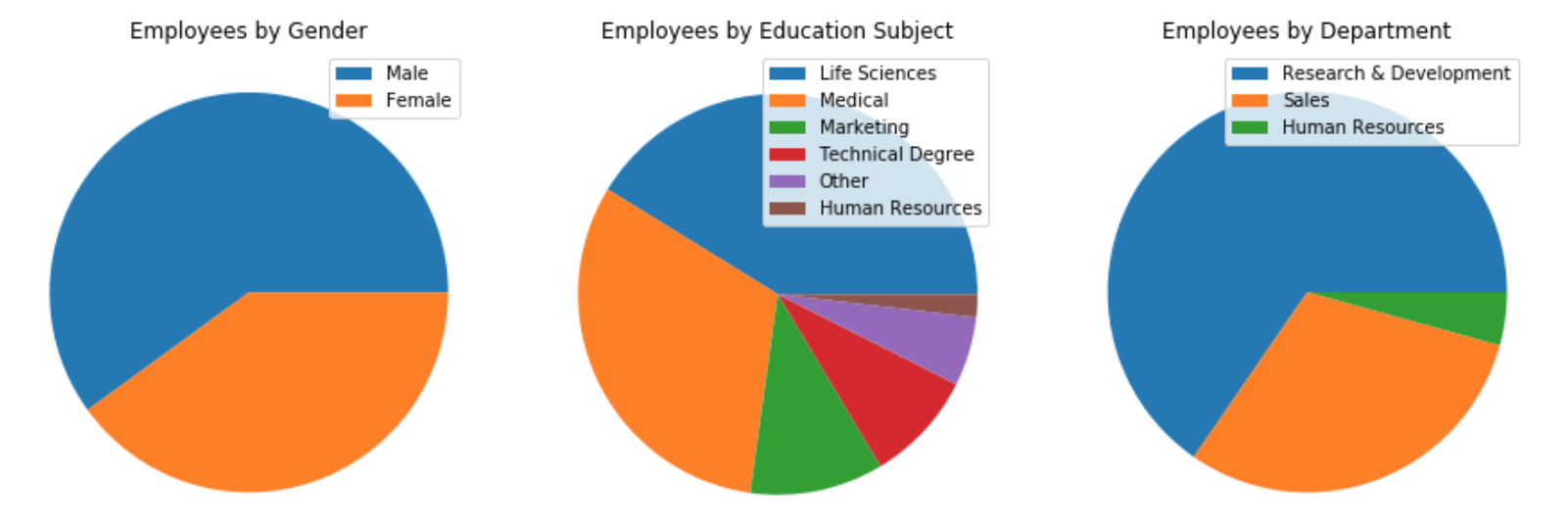
# An analysis of employee data from IBM

The database covers a wide variety of measures of IBM’s employees. Unfortunately, the context of the database was lacking, as it only includes a number of employees who left the company at some point; thus it’s impossible to tell whether this attrition rate was over the course of a month, a year, or a decade.

*However*, even though the lack of context makes any conclusions about IBM’s employee turnover impossible, there are a number of interesting things that can still be gleaned from the data.



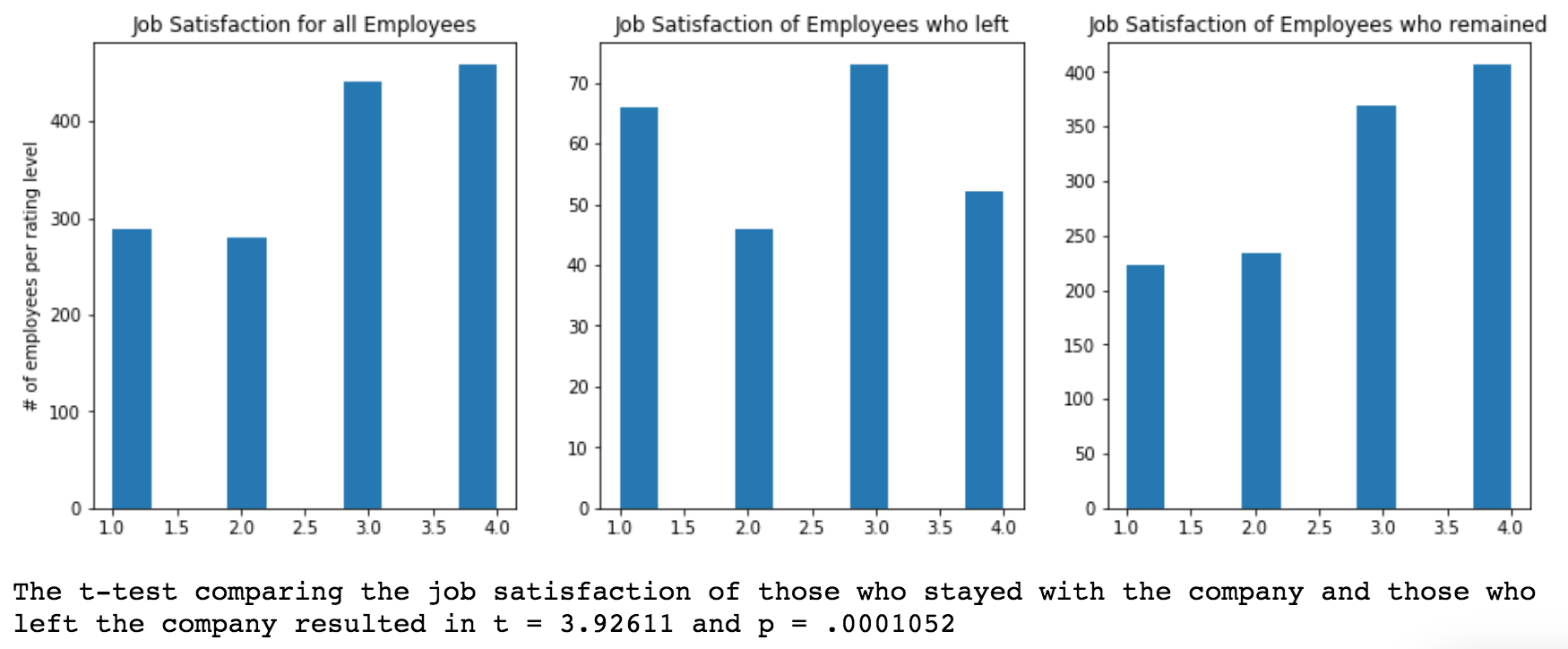
This is a basic overview of some of the notable proportions within the employee population. A slim majority of employees are male, almost three quarters got their degrees in either the life sciences or medical school, and the R&D department is huge compared to the Sales and HR departments.

## Overarching question: What could have contributed to attrition?

When looking at a population without any kind of context or knowledge of past trends, there are quite a few reasons why people could be leaving. Unhappiness with their job, low pay, too long between promotions, an excessively long commute, and many other factors can contribute to a person’s leaving their employer. How much of an influence they have is much harder to pin down, but comparing the differences between the population who chose to leave and those who remained with the company should at least point us in a useful direction.

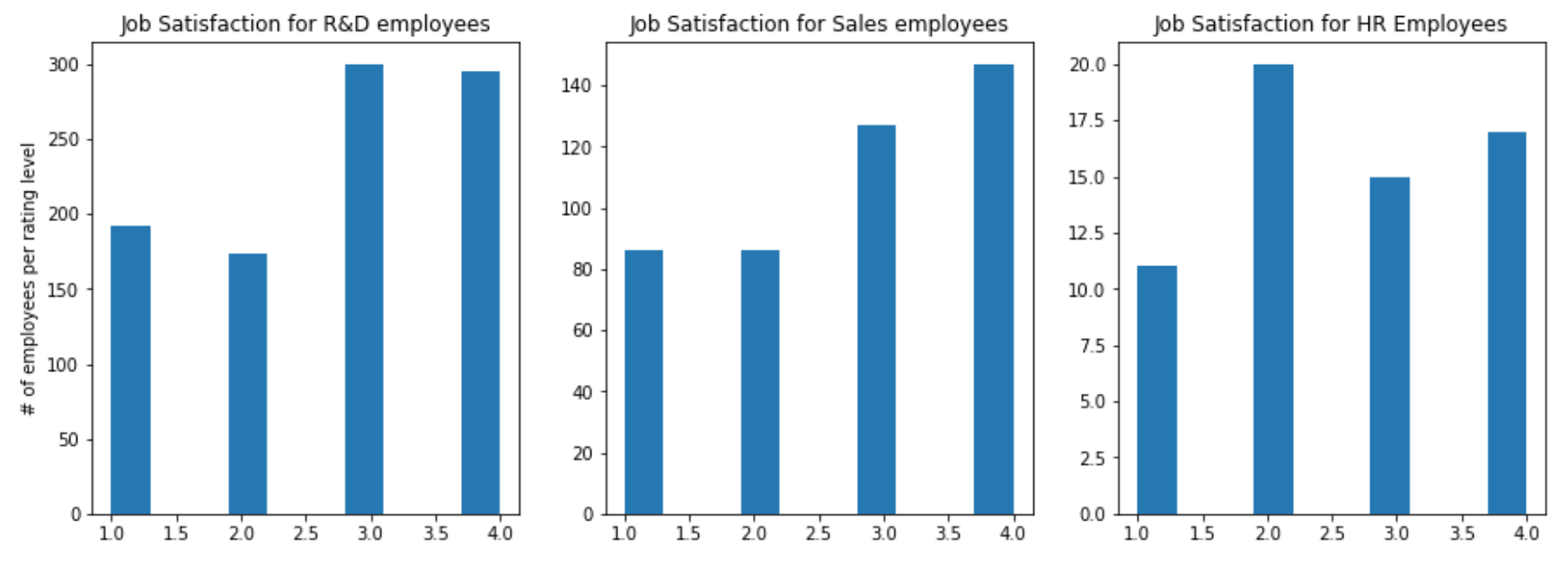
### Question 1: How much of an impact did job satisfaction have on attrition?

Job satisfaction is the first factor I chose to look at, given its likely correlation with someone’s decision to seek greener pastures. Comparing the self-reported ratings of those who chose to leave with those who stayed at IBM, one finds a noticeable (and completely unsurprising) uptick in the proportion of employees who left and rated their job satisfaction as only ‘1’ (the lowest possible rating) alongside a similar drop in the proportion who rate their job satisfaction as ‘4’ (the highest possible rating). A t-test confirms that the difference in the two populations is distinct and not simply an effect of random noise.

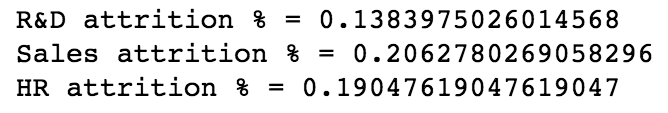


### Sidebar: job satisfaction between departments

Following up on job satisfaction as a subject, I was curious as to whether there was a discernable difference between the three departments, and it turns out that while R&D and Sales have job satisfaction numbers similar to the employee population as a whole, HR has a significantly higher proportion of employees who have less-than-stellar job satisfaction ratings:



However, despite this, HR had a slightly *lower* rate of attrition than Sales:



Pinning down the reasons for this discrepancy between departments might be possible with the data at hand, but is beyond the scope of this paper.

### Question 2: How much did commute distance impact decisions to leave?

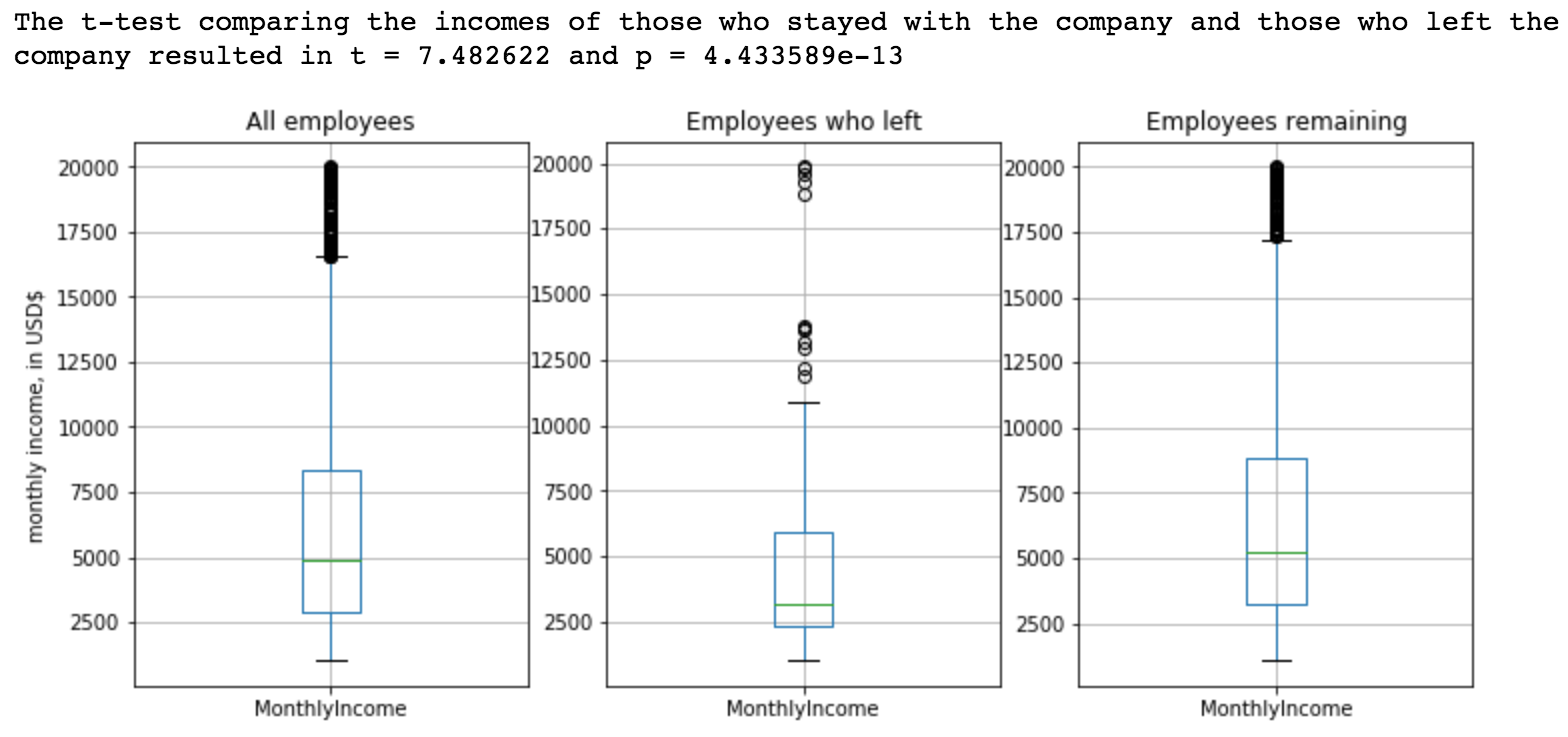
The distance that people commute is a known factor in how likely they are to stay with a company. How much did the population who left differ in this measure from those who stayed?



Somewhat, but not immensely. The average commute distance was indeed higher for those who left, but not immensely so. Further, the wide variance in both populations mean that the p-value for this comparison is too high to place great confidence in this measure as one of the deciding factors in whether people leave or stay.

### Question 3: How much did lower income correlate with decisions to leave?

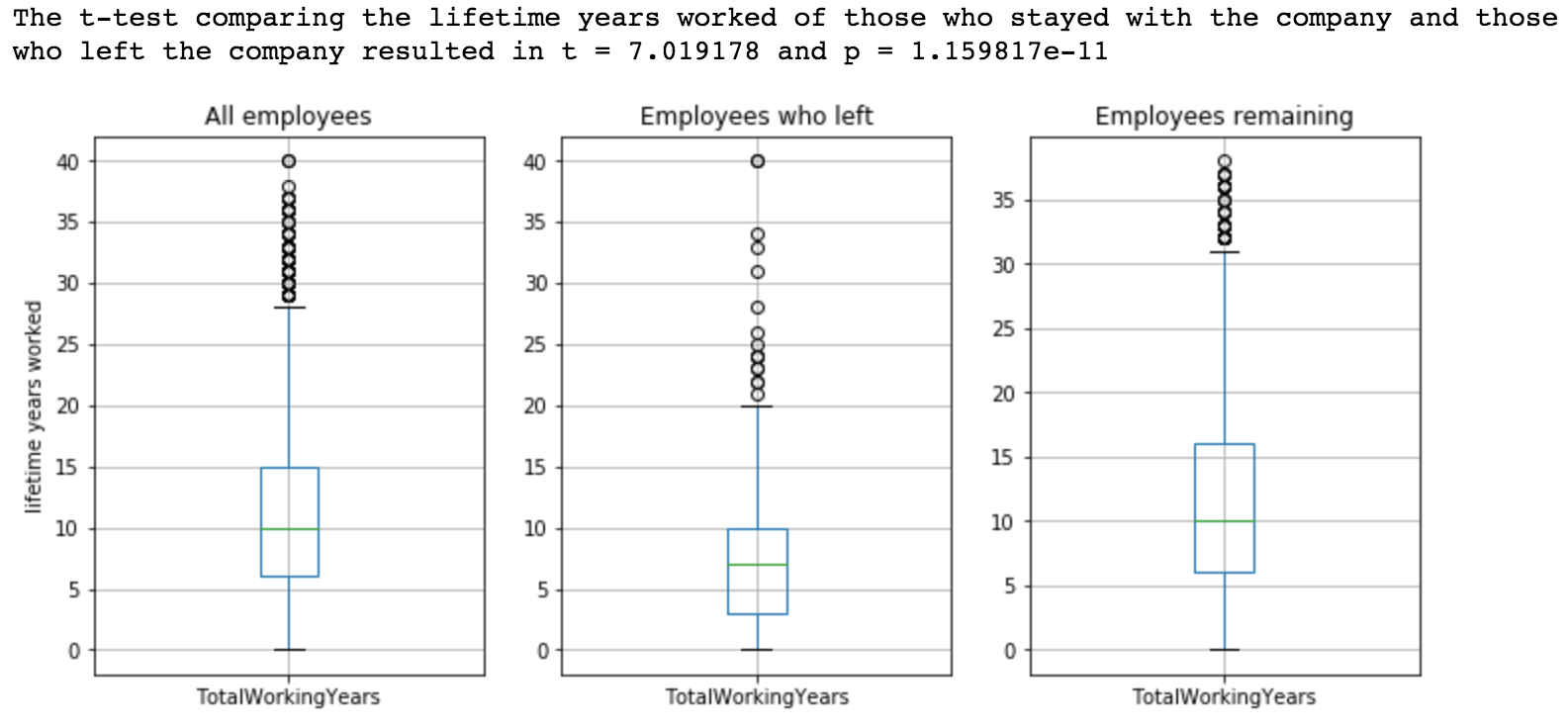
Income is one of the classic reasons why people leave one workplace for another. How did the two sub-populations compare here?



While there is a large amount of variance and many extreme outliers in both populations, *on average* the people who left had noticeably lower incomes than those who stayed. The t-test further underlines the difference between the two populations. This isn’t exactly a surprising conclusion, and it’s complicated by other factors that I’ll get into below, but it’s a worthwhile point to make in this discussion.

### Career length

One question that a lot of HR experts are asking these days is whether age has an impact on how often someone switches between jobs. The reasons for the (possible) shift in behavior patterns are immaterial for this paper, but the fact remains that it’s a point of curiosity for many, so it’s worth looking into:



The data displayed here show that there is a distinct difference in length of career (again, on average) between those who stayed at IBM and those who left, with a t-test almost as conclusive as the one for income.

This ties in with the relatively lower wages between those who left and those who stayed: the people who have worked more years tend to have higher salaries, and so if there’s a difference in length of careers then that almost certainly has an impact on the disparity of average income between the two populations. We need to be very careful in not drawing direct causal relationships between these factors and peoples’ decisions to leave, as there are often multiple factors that impact those decisions, and not all of them are going to be direct. Strong correlation is as much as we are likely to get, with this data set.

### Years worked at IBM

“Company loyalty” is a known phenomena; people get used to the way things are done at a particular company, and become reluctant to leave. How much of an effect did the length of time employees had been working at IBM have on their odds of leaving?



The comparison and t-test aren’t as conclusive here as in the two prior tests, but it is still plenty good enough to show a strong correlation between length of time spent working for IBM and the likelihood of someone staying with the company.

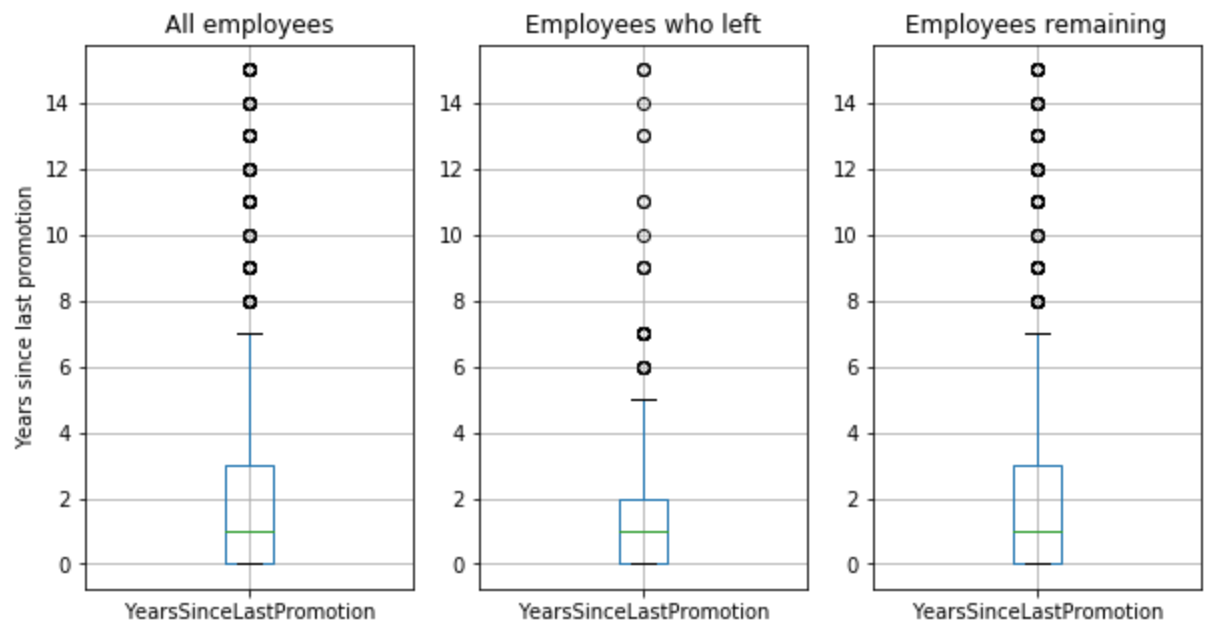
(This, unfortunately, gets into hazy territory given the lack of timeline context for the database – because we don’t know how much time the database covers or how much total turnover there is every year, we can’t pin down exactly how important this correlation is. With a better idea of how many new employees were entering or leaving the company per time period, we would be able to draw better conclusions about the importance that length of time working there had in people’s decisions to leave or stay.)

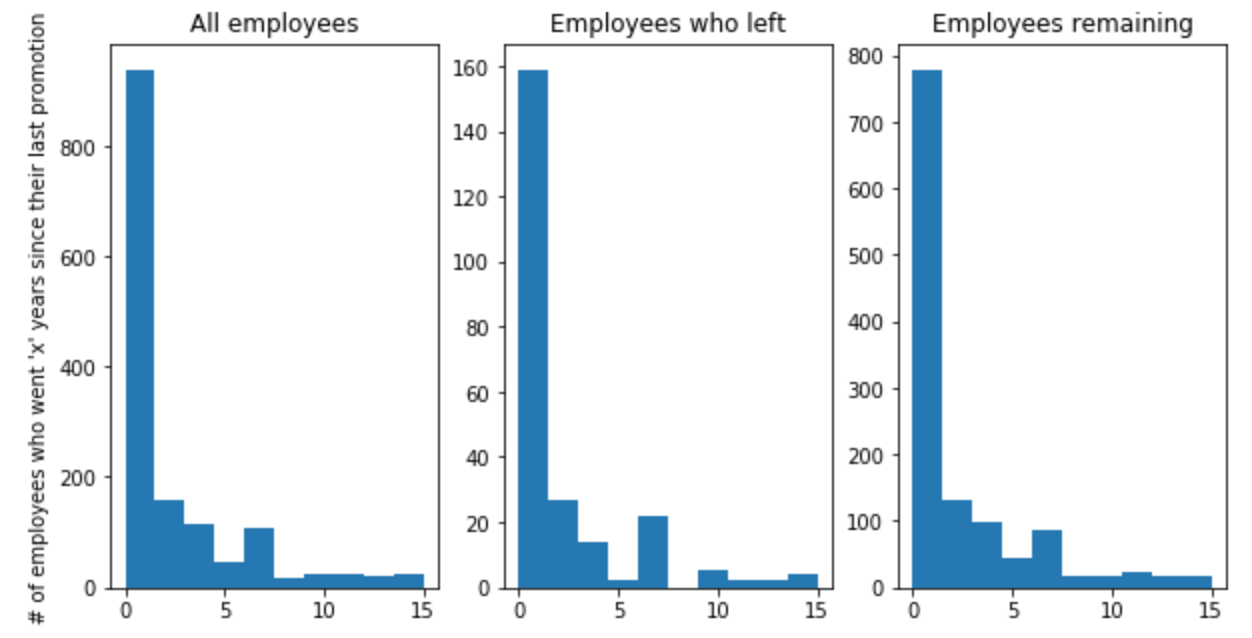
### Years since last promotion

Possibility for advancement is another “obvious” reason why people choose to stay or leave a company, and the database has a metric that mostly works to measure that, by counting the years since an employee’s last promotion. The issue one runs into with this measure, though, is that it runs afoul of the issue that people who leave because of lack of advancement tend not to stay beyond a certain point. While that point differs for each individual, the odds are that there are very few people who quit their job after ten years without promotion, and many more who would quit much earlier. This means that our baseline assumption when using this measure as a proxy for why employees left the company – that those who left did so because they went to long without a promotion – is fundamentally flawed, because those who are likely to leave at all for this reason are likely to do so *early*, skewing the results towards the lower end of the scale.

And this is exactly what the data show: those who left tended to do so within 5 years of their latest promotion. While some stuck around for up to 15 years, they are significant outliers to the “attrition” population, and the p-value for this comparison is so high that the test is effectively impossible to draw any conclusions from.

Side note: There’s also an oddly uniform hump around the 7-year mark; no matter which population you’re looking at, people tended to go get promoted when they hit the 7 year mark, and nobody went more than 15 years without one. Both of these points are rather odd, and I wish there were more context to be found here to answer this particular puzzle. (The answer’s likely in the HR or managerial handbook for the company, which sadly wasn’t included with the database.)





The t-test comparing the years since their last promotion of those who stayed with the company and those who left resulted in t = 1.287927 and p = 0.198651.

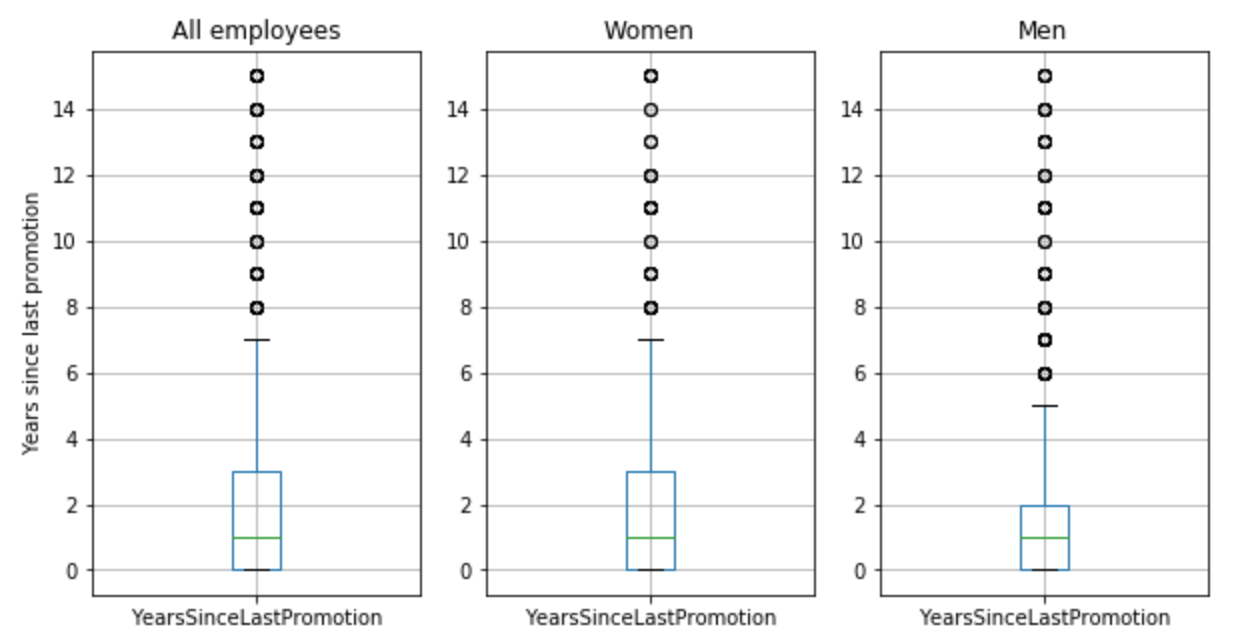
## Conclusion: too little evidence

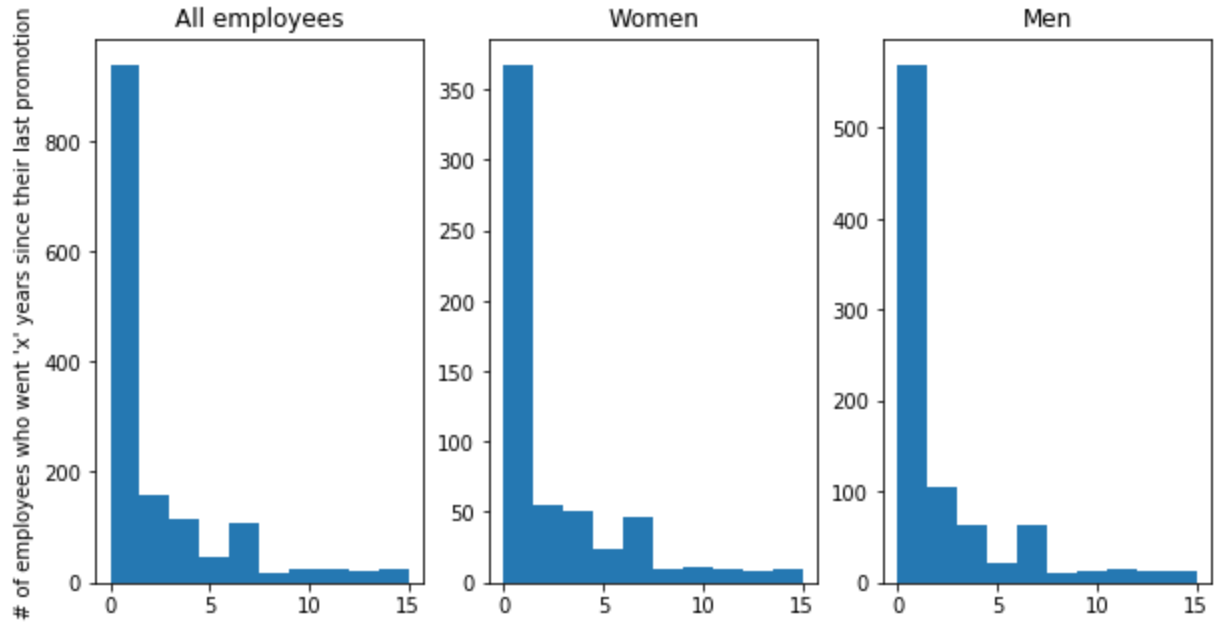
There’s too little causal data here for someone of my caliber in statistics to be able to draw firm conclusions about *why* the majority of people who left the company did so. While there are certainly some strong correlations, none are definitive, in my eyes, and the utter lack of context surrounding the database makes my very reluctant to render judgment on any of them. That said, the strong correlation with job satisfaction, income, and time spent at the company does lend itself to the stereotypical analysis of employee loyalty. Whether that’s an *accurate* analysis in this case, I couldn’t say without more data.

\*\*\* what did you learn?

### PS / Sidebar: Gender’s impact on promotions

A random test I ran to see if there were anything else interesting in the database was to compare the difference in years since last promotion for men versus for women. The results were inconclusive, at best, with a p value of over 0.3, but still suggestive from the box chart that there *might* be some degree of bias there.





The t-test comparing the years since their last promotion of men versus women resulted in t = 1.0287927 and p = 0.303721

# box plots

plt.figure(figsize=(10, 5))

plt.subplot(1, 3, 1)

employees\_sincepromo.boxplot()

plt.title('All employees')

plt.ylabel('Years since last promotion')

plt.subplot(1, 3, 2)

attrition\_sincepromo.boxplot()

plt.title('Employees who left')

plt.subplot(1, 3, 3)

remain\_sincepromo.boxplot()

plt.title('Employees remaining')

plt.show()

# histograms

plt.figure(figsize=(10, 5))

plt.subplot(1, 3, 1)

plt.hist(employees['YearsSinceLastPromotion'])

plt.title('All employees')

plt.ylabel('# of employees who went \'x\' years since their last promotion')

plt.subplot(1, 3, 2)

plt.hist(attrition['YearsSinceLastPromotion'])

plt.title('Employees who left')

plt.subplot(1, 3, 3)

plt.hist(remain['YearsSinceLastPromotion'])

plt.title('Employees remaining')

plt.show()