Employee Churn HR Analysis

Evan Moore

October 17, 2016

Exploratory Analysis

My first step is importing the data and taking a look at the summary to get a sense of the variables' values across all employees.

```
library(RCurl)
x <- getURL("https://raw.githubusercontent.com/tommyjee/ugrid/master/Proj_3/employee_retention_data.csv
employees <- read.csv(text = x)
library(dplyr)
library(lubridate)
library(ggplot2)
employees <- tbl_df(employees)
summary(employees)
```

```
##   employee_id       company_id                   dept        seniority
##  Min.   :    36   Min.   : 1.000   customer_service:9180   Min.   : 1.00
##  1st Qu.:250134   1st Qu.: 1.000   data_science    :3190   1st Qu.: 7.00
##  Median :500793   Median : 2.000   design          :1380   Median :14.00
##  Mean   :501604   Mean   : 3.427   engineer        :4613   Mean   :14.13
##  3rd Qu.:753137   3rd Qu.: 5.000   marketing       :3167   3rd Qu.:21.00
##  Max.   :999969   Max.   :12.000   sales           :3172   Max.   :99.00
##
##      salary            join_date          quit_date
##  Min.   : 17000   2012-01-03:  105   2015-05-08:  111
##  1st Qu.: 79000   2011-08-29:  104   2015-11-27:  102
##  Median :123000   2014-05-12:  104   2015-03-06:  100
##  Mean   :138183   2012-03-26:  102   2015-04-10:   99
##  3rd Qu.:187000   2014-03-24:  102   2015-08-21:   99
##  Max.   :408000   2013-12-09:  100   (Other)   :12999
##                   (Other)   :24085   NA's      :11192
```

Nothing seems to stick out, besides the confusing max value of 99 for seniority . Let's investigate that further.

```
sort(unique(employees$seniority), decreasing = TRUE)
```

```
##  [1] 99 98 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10  9
## [24]  8  7  6  5  4  3  2  1
```

```
subset(employees, seniority > 30)
```

```
## # A tibble: 2 x 7
##   employee_id company_id      dept seniority salary  join_date  quit_date
##         <dbl>      <int>    <fctr>     <int>  <dbl>     <fctr>     <fctr>
## 1       97289         10  engineer        98 266000 2011-12-13 2015-01-09
## 2      604052          1 marketing        99 185000 2011-07-26 2013-12-06
```

Only a couple odd values in the dataset; let's throw them out since we have so much other data to work with.

```
employees <- subset(employees, seniority < 30)
```

Now that we have a more accurate dataset, we can split it into chunks of those who stayed and those who quit, and check those out individually to see if anything notable comes up.

```
stay <- subset(employees, is.na(employees$quit_date))
quit <- subset(employees, !is.na(employees$quit_date))
summary(quit)
```

```
##   employee_id       company_id                   dept        seniority
##  Min.   :    36   Min.   : 1.000   customer_service:5094   Min.   : 1.00
##  1st Qu.:252838   1st Qu.: 1.000   data_science    :1682   1st Qu.: 7.00
##  Median :501208   Median : 2.000   design          : 778   Median :14.00
##  Mean   :502062   Mean   : 3.459   engineer        :2361   Mean   :14.12
##  3rd Qu.:755098   3rd Qu.: 5.000   marketing       :1782   3rd Qu.:21.00
##  Max.   :999969   Max.   :12.000   sales           :1811   Max.   :29.00
##
##      salary            join_date          quit_date
##  Min.   : 17000   2011-08-29:   98   2015-05-08:  111
##  1st Qu.: 81000   2011-10-10:   93   2015-11-27:  102
##  Median :122000   2011-05-23:   92   2015-03-06:  100
##  Mean   :135639   2011-06-06:   92   2015-04-10:   99
##  3rd Qu.:180000   2012-01-03:   91   2015-08-21:   99
##  Max.   :379000   2011-04-11:   90   2015-09-11:   99
##                   (Other)   :12952   (Other)   :12898
```

```
summary(stay)
```

```
##   employee_id       company_id                   dept        seniority
##  Min.   :   107   Min.   : 1.000   customer_service:4086   Min.   : 1.00
##  1st Qu.:247897   1st Qu.: 1.000   data_science    :1508   1st Qu.: 7.00
##  Median :499790   Median : 2.000   design          : 602   Median :14.00
##  Mean   :501079   Mean   : 3.388   engineer        :2251   Mean   :14.12
##  3rd Qu.:750487   3rd Qu.: 5.000   marketing       :1384   3rd Qu.:21.00
##  Max.   :999840   Max.   :12.000   sales           :1361   Max.   :29.00
##
##      salary            join_date          quit_date
##  Min.   : 17000   2015-07-06:   93   2011-10-13:    0
##  1st Qu.: 76000   2015-06-08:   92   2011-10-14:    0
##  Median :123000   2015-03-16:   91   2011-10-21:    0
##  Mean   :141238   2015-03-09:   90   2011-10-28:    0
##  3rd Qu.:195000   2015-01-20:   89   2011-11-11:    0
##  Max.   :408000   2015-02-09:   89   (Other)   :    0
##                   (Other)   :10648   NA's      :11192
```

We can see that those who stay tend to have higher average salaries by about 6 thousand dollars than those who quit, but otherwise everything seems to be normal between the two sets, including having identical seniority values. Less salary for those who quit could indicate that employees may be motivated by a desire for greater pay as a reason to quit. Many more customer service employees are leaving than any other department, but this is explained by the much higher number of employees in that category represented in the dataset. To get a better sense of which departments are suffering the most from churn, we can by dividing the count of each in `quit` by the total count for the `employees` dataset to figure out what percentage of each job ended up leaving.

```
library(plyr)
by_dept <- count(quit, 'dept')[,2] / count(employees, 'dept')[,2]
names(by_dept) <- c("Customer Service","Data Science","Design","Engineer","Marketing","Sales")
by_dept
```

```
## Customer Service     Data Science           Design         Engineer
##        0.5549020        0.5272727        0.5637681        0.5119254
##        Marketing            Sales
##        0.5628553        0.5709332
```

```
sum(by_dept) / 6 #average between all departments
```

```
## [1] 0.5486095
```

So, data scientists and engineers have the lowest turnover at 52-53%, while the rest of the departments have between 55 and 57%. Considering the specialized nature of the work data scientists and engineers do, it makes sense that they would have more permanent jobs and benefits that may lead to less quitting. On the other hand, customer service jobs often come and go, not mandating any advanced experience and being popular jobs for young people who may have school or other responsibilites that could make them quit or switch jobs. As for the other departments, being somewhere in between the specialty of customer service and data science/engineering, it's not immediately clear what would make them have higher-than-average turnover rates.

Salary seems to be a reasonable variable to go off of (being the only clear difference between the general `stay` and `quit` datasets), so I tried some analysis of the salaries in specific departments between those who stayed and those who quit to get a better sense of the relationship. First, for the department with the greatest amount of quitting, sales:

```
sales_quit <- filter(quit, dept=='sales')
sales_stay <- filter(stay, dept=='sales')
sumSales <- sum(sales_quit$salary / nrow(sales_quit)) - sum(sales_stay$salary / nrow(sales_stay))
sumSales
```

```
## [1] -638.6808
```

There's only a few hundred dollars difference - salary is probably not affecting whether people working in sales quit or not. For data science, on the other hand...

```
data_quit <- filter(quit, dept=='data_science')
data_stay <- filter(stay, dept=='data_science')
sumData <- sum(data_quit$salary / nrow(data_quit)) - sum(data_stay$salary / nrow(data_stay))
sumData
```

```
## [1] -16124.55
```

A more notable difference of 16 thousand dollars could potentially point to a reason why a data scientist may leave and look for a higher-paying job. Engineers display a similar, but slightly smaller, gap, perhaps explaining why they quit less than data scientists.

```
engin_quit <- filter(quit, dept=='engineer')
engin_stay <- filter(stay, dept=='engineer')
sumEngin <- sum(engin_quit$salary / nrow(engin_quit)) - sum(engin_stay$salary / nrow(engin_stay))
sumEngin
```

```
## [1] -13642.39
```

The difference in salaries from all departments:

```
##            Sales     Data Science      Engineering Customer Service
##        -638.6808      -16124.5541      -13642.3913        4665.7560
##        Marketing           Design
##        2432.7532        5561.3081
```

Interestingly, the rest of the departments have a positive difference, meaning those who quit actually made more. While this contradicts the conclusions made from the `quit`/`stay` summaries, it is in fact reasonable to consider the inflated values for data scientists and engineers in a different category as the rest of the jobs

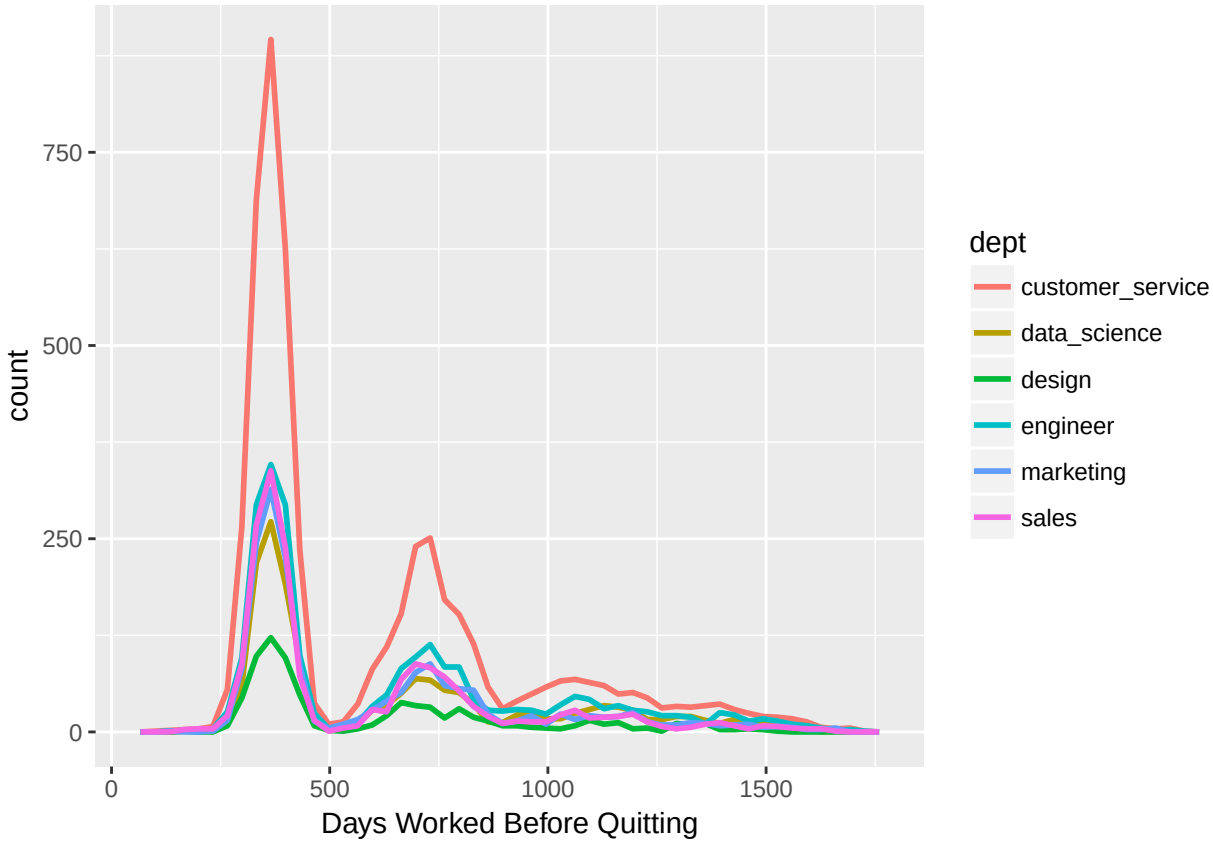
due to their higher prestige and salaries. This indicates that the 6 thousand dollar difference between these datasets was almost solely caused by the much higher salaries from data scientists and engineers, further confirming my suspicion that these jobs should be treated separately than the others ones. Perhaps this implies that those who quit were getting payed more, and therefore had more responsibility or potential as deemed by the company, making them more attractive targets at other jobs should they decide they want a new work experience. This still doesn't explain why they would quit in the first place, though. Let's take a look at average differences in seniority among each department between quit and stay to see if there are any connections to be untangled there.

```
##            Sales     Data Science      Engineering Customer Service
##      -0.03147879      -1.46133266      -1.07300158       0.82045821
##        Marketing           Design
##       0.34801370       0.84590781
```

As the summaries suggest, there is no major difference between the seniorities of those who quit versus those who stayed. We see that in half the cases, quitting indicates a lower seniority, while in the other half it is the opposite. Like the salary summaries, data science and engineering are in a league of their own, while sales is barely positive and the rest are negative values, indicating that outside of the high-tier jobs those with more seniority coming into the job may be more likely to quit. For data science and engineering, more previous experience indicates that they are more likely to stay, perhaps reflecting the terminal nature of these jobs. Customer service and design both have values of seniority close to a year greater for those who quit, which implies that hiring employees with less seniority for these jobs may lead to less churn.

Finally, we can look at the length of time that each worker stayed at their job in order to gain more insight about which departments may have the quickest vs. longest churn rates overall. The lubridate package makes these types of calculations easy.

```
ggplot(quit, aes(x = (as.numeric(work_days) / 86400))) + geom_freqpoly(bins= 50, aes(col = dept), size =
```



Judging by this plot, employee churn spikes heavily after about on year of work, then has a smaller spike after about two years, before roughly evening out for any number of days after. Both large chunks are normally distributed, showing that the numbers start ramping up after about 250 days, peak at one year and then slowly ramp down until about 500 days before beginning the same process again, albeit on a smaller scale. This make the data heavily skewed right, as empoyees who have been working for multiple years in a row will be less likely to quit than someone who works for a year and realizes they are not fit for the job. The departments' sizes are reflected on the y-axis - customer service has the greatest number of employees by a large margin, and therefore has a much higher quit amount than the other jobs. Design is on the other end of the spectrum, featuring noticably lower values. The rest of the departments are roughly equivalent, with data scientists having slightly lower amount. Let's look at each department individually based on average number of days worked to get some more insight.

```
##            Sales     Data Science      Engineering Customer Service
##         580.8099         650.8284         649.1881         602.2899
##        Marketing           Design
##         600.7279         602.8709
```

Data scientists and engineers stay about a month and a half longer than the norm, with the other three jobs closely related and sales lagging slightly behind. Interesting how the relationships between the variables in this dataset exhibit similar traits through a variety of different situations.

Conclusions

Overall, my results were inconclusive - while I was able to determine some new relationships between variables in the `quit` subset, I was not able to clearly determine what is making employees quit. While for data science and engineering, higher salaries offered by other jobs point to a potential reason for quitting, this relationship does not hold for the other departments. The greatest amount of employees will leave during their first year, so the company should focus more on making the transition into working smoother to keep people from quitting so early. Similar tactics should be applied in the second year to alleviate the smaller churn spike that comes after. High numbers of customer service jobs quitting should not be the focus of the company since these jobs often come and go due to the nature of the employees: instead, the higher than average number of sales, design, and marketing jobs quitting should be looked into further to see what may be causing that. For customer service and design in particular, hiring employees with less experience coming in should lead to less churn, at the cost of potentially higher training expenses if required. Sales could also be looked because employees spend on average the least amount of time working in that department before quitting.

If we were to include more variables, an important one for predicting churn would be whether the employee's partner has changed jobs, as that would provide a clear reason to quit, even if there was nothing otherwise making the employee do so. And while somewhat harder to work with, sentiment analysis could be applied to any sort of feedback employees have given about their jobs and predictors could be utilized as a result. Getting information on why the employees left would also be immensely valuable, whether they were fired for incompetence/not being a good fit, left of their own accord for a new offer, or quit for unrelated reasons potentially related to the workplace enviroment. Seeing the spread of reasons why (even if not entirely accurate) would give a better sense of what may be driving churn in a company.