Analysis of Factors Affecting Resignations of University Employees

An exploratory study was conducted to identify factors influencing voluntary resignations at a large research university over the past twenty years. A linear regression model was fit using techniques including cross-validation and a power transformation of the response variable. The university's HR department was concerned with an observed higher turnover among their youngest employees (i.e. Millennials) and among female employees. Using analysis of variance and predicted marginal means, we find that sex is not a significant predictor for length of employment. Furthermore, although an employee's generation and age at resignation are significant factors, Millennials do not exhibit a practically significant different length of employment compared to other generational groups. This finding disrupts stereotyped representations of generational factors in the workforce and suggests that younger employees resigning sooner can be better explained as a feature of their age rather than their generational group.

Introduction

Our client is a member of the Human Resources department at a large public research university. They observed higher numbers of voluntary resignations among young university employees in recent years compared to all the employees on average. Also, female employees appear to be resigning more frequently compared to their male counterparts. Hence, we ask whether an employee's length of employment before resignation is significantly affected by age, generation, gender, or job-specific factors such job type, employee rank (analogous to an employee's seniority), and union membership.

Methodology

The client provided a dataset with employee resignations at the university occurring between 1996 and 2015. The dataset starts with over 7000 observations, where each observation represents the resignation of an employee. We use a multiple linear regression model to explain YEARS_OF_SERVICE. The variables in the final model are defined in Table 1.

Final model variables	Definition
YEARS_OF_SERVICE (response)	Length (years) of most recent continuous service term
SEX	Employee's self-identified gender
SERVICE_DATE	Starting date of most recent continuous service term
CLUSTER	Job position's category
BIRTH_DATE	Year of birth (number of days from 1900)
GEN_AGE	Generation and the age bracket at resignation
UNION_RANK	Whether the job is unionised and the job rank

Table 1. Final model variables and their brief definitions.

Starting with the initial dataset, we performed multiple rounds of data-cleaning. This included the removal of obviously erroneous observations and redundant variables, as well as declaring variable type (i.e. continuous or categorical). We also removed observations where YEARS_OF_SERVICE is less than four months based on client feedback, leaving 6721 observations.

We resolved additional problems including variables with multicollinearity, imbalance, and high granularity. For example, the two variables SERVICE_DATE (the start date of an employee's most recent period of continuous service) and JOB_ENTRY_DATE (the date when an employee began their current job) are identical except for those employees that switch jobs within the university. Hence, we eliminated one based on the mean-squared prediction error averaged over five runs of five-fold cross-validation. We chose cross-validation over other goodness-of-fit measures such as AIC because the greater generalisability of results from cross-validation renders it a more appealing technique for the client (Hastie et al. 2009).

Resignation age	Traditionalists	Baby Boomers	Gen. X	Gen. Y
24 & Under	🟨	🟨	168	831
25 to 34	🟨	32	1123	1972
35 to 44	🟨	341	1123	85
45 to 54	5	531	211	🟥
55 to 59	28	155	🟥	🟥
60 & Over	27	89	🟥	🟥

Table 2. Blue: possible values of GEN_AGE covered by dataset with sample size for each level shown; red: impossible values at present; yellow: possible values not covered by dataset's time span.

Additionally, we created the new variable GEN_AGE to facilitate the comparison of employees from different generations within resignation age ranges and vice versa. The GEN_AGE variable is ostensibly the interaction of an employee's generation and age at resignation, but with impossible and out-of-bound combinations of the two factors dropped (Table 2). By combining some age brackets we also addressed the imbalance of the original age bracket variable (Appendix Fig. 6 and Table 4).

To reduce granularity of the given UNION_CODE variable, we replaced it by a binary variable that indicates whether an employee is unionised or not. However, since the client formed the levels of RANK using union information, the variables are correlated and so combining them to form UNION_RANK retains the information in both while improving the model (Appendix Fig. 7).

The final model describes the relationship between YEARS_OF_SERVICE and the six explanatory variables in Table 1. We diagnosed the validity of the statistical inference arising from the model (Appendix Fig. 3). Independence of observations is justified as employees can rationally be modelled as resigning for independent reasons. We verified the normality of the error term via a Q-Q plot. We observed high leverage points in the residual plot but concluded they are not of concern because their removal does not change the fit of the data. The homoscedasticity assumption is satisfied only following a power transformation of the response variable, where the power was found via a box-cox transformation (Appendix Fig. 2).

As a result we have built a model with an adjusted R^2 of 0.63 and has diagnostic plots that indicate that the model satisfies the necessary assumptions for valid statistical inference.

Results

We began with analysis-of-variance of our final model via type-II sum-of-squares (which is suitable for imbalanced data) to identify significant variables.

	Sum-of-squares	Df	F value	Pr(>F)
SEX	0.06	1	1.01	0.32
SERVICE_DATE	528.16	1	8505.20	<0.001
CLUSTER	7.06	6	18.94	<0.001
BIRTH_DATE	139.62	1	2248.44	<0.001
GEN_AGE	335.76	14	386.20	<0.001
UNION_RANK	12.91	7	29.71	<0.001
Residuals	415.44	6690		

Table 3. All variables except SEX *are significant; note that* SERVICE_DATE*'s estimated coefficient has a magnitude of < 0.001 and so is practically insignificant.*

We visualised the results of the estimated model using a technique called *predicted marginal means (PMM)* (Appendix p. 7). This method predicts the years of service of a "typical" employee given common factor levels by averaging over all other quantitative and categorical variables in the model.

The gender of an employee is not a significant predictor of the years of service of those employees who voluntarily resign, as seen in Fig. 1(a). This can be explained by the higher total proportion of female employees at the university.

Turning to the GEN_AGE variable, within the two age brackets 25 to 34 and 35 to 44 (Fig.'s 1(b) and 1(c)) there is a significant difference in length of employment between a typical employee in Generation X compared to those in Generation Y (Millennials) and Baby Boomers. Interestingly, the direction of the difference between the 25 to 34 and 35 to 44 generational cohorts flips. After consulting with the client, we learned that these findings may reflect the effect of economic variables not considered in the model. Generation X employees entered the job market during economically challenging circumstances and needed to switch jobs often early in their careers. Now, they tend to hold onto hard-won career positions within the university.

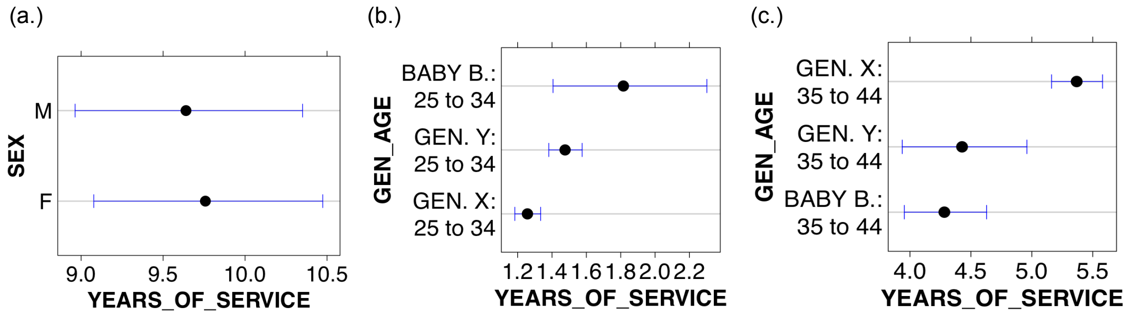

Figure 1. PMM for factor combinations. The center dot is the estimated marginal mean and the bar is the corresponding 95% confidence interval. (a.) PMM years of service between female and male employees. (b.) PMM years of service between generations when age at resignation is 25-34. (c.) PMM years of service between generations when age at resignation is 35-44.

We found no significant differences among YEARS_OF_SERVICE for employees in age brackets greater than 35 to 44, suggesting that generational factors have little influence on the decision of employees older than 45 to voluntary resign (Appendix Fig. 8). Furthermore, despite the existence of significant differences between the generations' YEARS_OF_SERVICE within brackets 25 to 34 and 35 to 44, this is relatively small compared to the magnitude of the difference in YEARS_OF_SERVICE *between* age brackets themselves. That is, the predicted span in YEARS_OF_SERVICE between the age brackets is much greater for older age brackets compared to the predicted span in in younger age brackets.

This is interesting in light of common narratives about Millennials who are thought to switch jobs more quickly and often compared to other generational groups. The model's predictions contradict such narratives: the difference in magnitude for the YEARS_OF_SERVICE on the longitudinal axis between Fig.'s 1(b) and 1(c) (1.2 to 2.2 YEARS_OF_SERVICE versus 4.0 to 5.5) suggests that regardless of generation, younger employees are predicted to resign more quickly.

Finally, for the significant predictor UNION_RANK, employees with union and rank codes that correspond to higher-level management positions had a greater YEARS_OF_SERVICE than those in other ranks (Appendix Fig. 9).

Conclusion

We find that differences in GEN_AGE and UNION_RANK lead to practically significant differences in YEARS_OF_SERVICE for employees who voluntarily resigned. Indeed, an employee's age at resignation has a greater effect on length of employment than their generation, which means higher turnover can be expected among younger people in general regardless of their generational bracket.

Future work on this model can bring in earlier years in the dataset to study young resignations at different historical times. One of the biggest limitations of this study was the 20-year time span of the data, 1996-2015. That is, given data from 1966-1996, we could draw more general conclusions about young employees by including the Traditionalists and Baby Boomers at the 24 & Under age brackets.

Additionally, we believe it would be productive to apply a dimensionality reduction algorithm like t-SNE (van der Maaten & Hinton 2008) on the dataset to visualize the high-dimensional structure of the employee resignations in 2-D or 3-D in order to reveal patterns among the resignations that exist naturally in the higher-dimensional space.

Finally, a separate, second dataset is available that contains "snapshot" information recording the employment status of all employees each year. This opens up the possibility of a survival analysis, where we model the expected time until an employee's resignation. One of the main limitations of this study is the conditional nature of the response variable, and doing a survival analysis on the snapshot dataset could overcome this limitation.

Appendix

The response variable

Note that the response variable is an employee's length of employment is conditional on the fact that the employee voluntarily resigned and had worked for at least four months. The original skewed distribution is visible in the left histogram of Fig. 2, and we applied a power transformation to the response variable in order to improve the model diagnostics. The right histogram in Fig. 2 shows the improved shape of the distribution after raising the response to the power of 3.30, as suggested by the box-cox test.

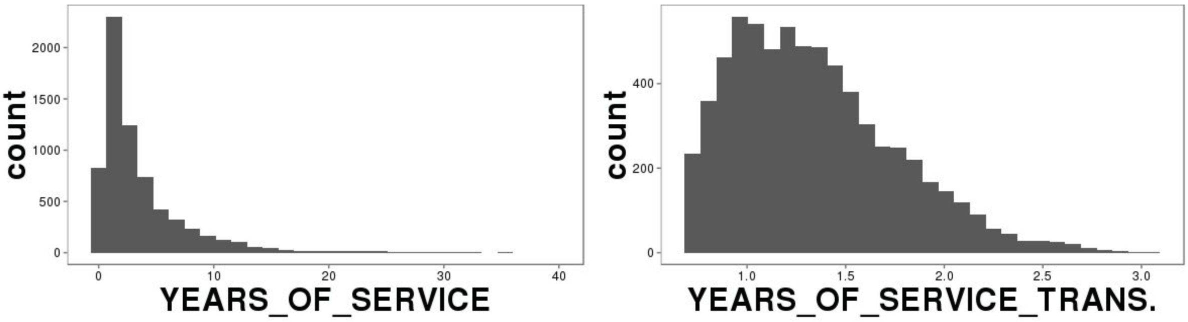

Figure 2. Histogram of `YEARS_OF_SERVICE` *before (left) and after (right) transformation.*

Model diagnostics

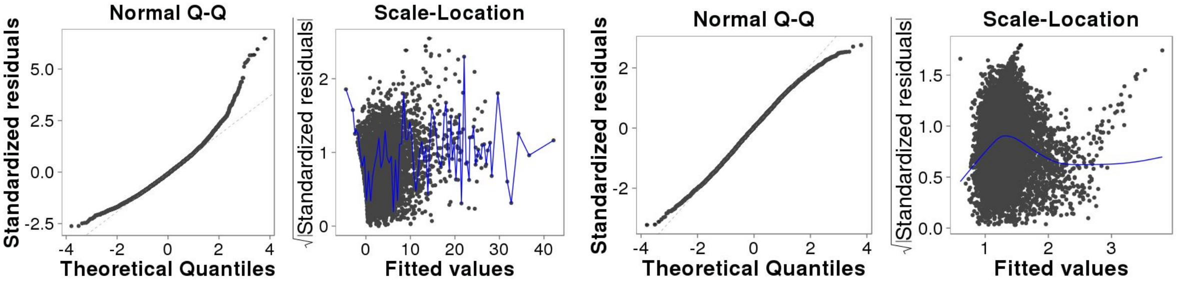

Figure 3. Left two plots are diagnostics of an early model and the right two plots are those of the final model; note the vast improvement in the Q-Q plot and the more random scatter in the fitted values versus residuals plot.

Visualising the categorical variables and variable balance problems

A number of the categorical variables presented serious balance problems. A comparison among the histograms in Fig.'s 5-7 depicts the problem clearly.

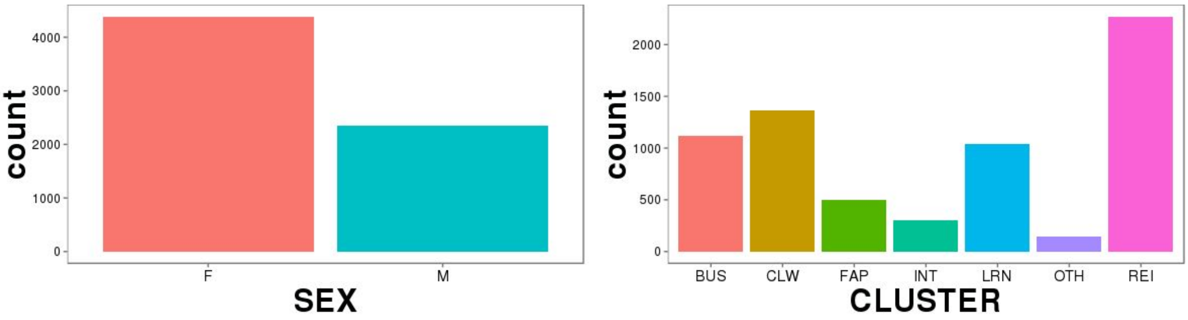

Figure 4. Distribution of `SEX` *and* `CLUSTER`*; both variables are reasonably balanced.*

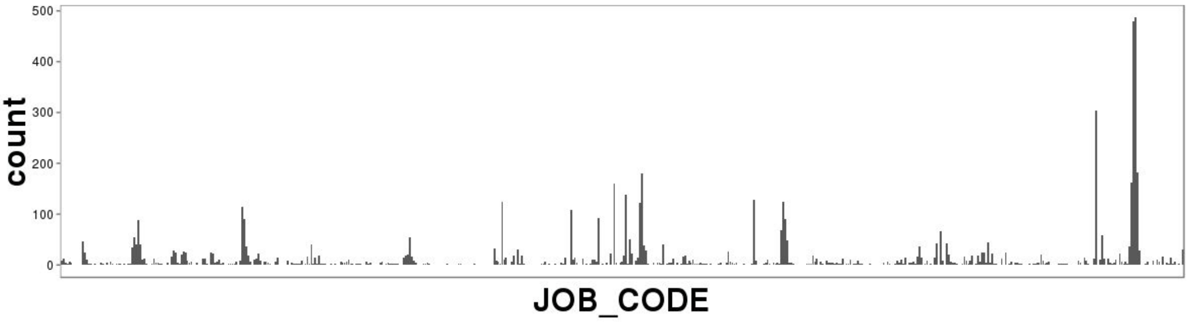

Figure 5. Distribution of JOB_CODE*; extreme balance issues since many job codes have one observation only, whereas others have hundreds. This leads to very large error ranges. This imbalance is in part caused by high granularity of the variable–a large number of possible job codes. Labels on the x-axis omitted for clarity.*

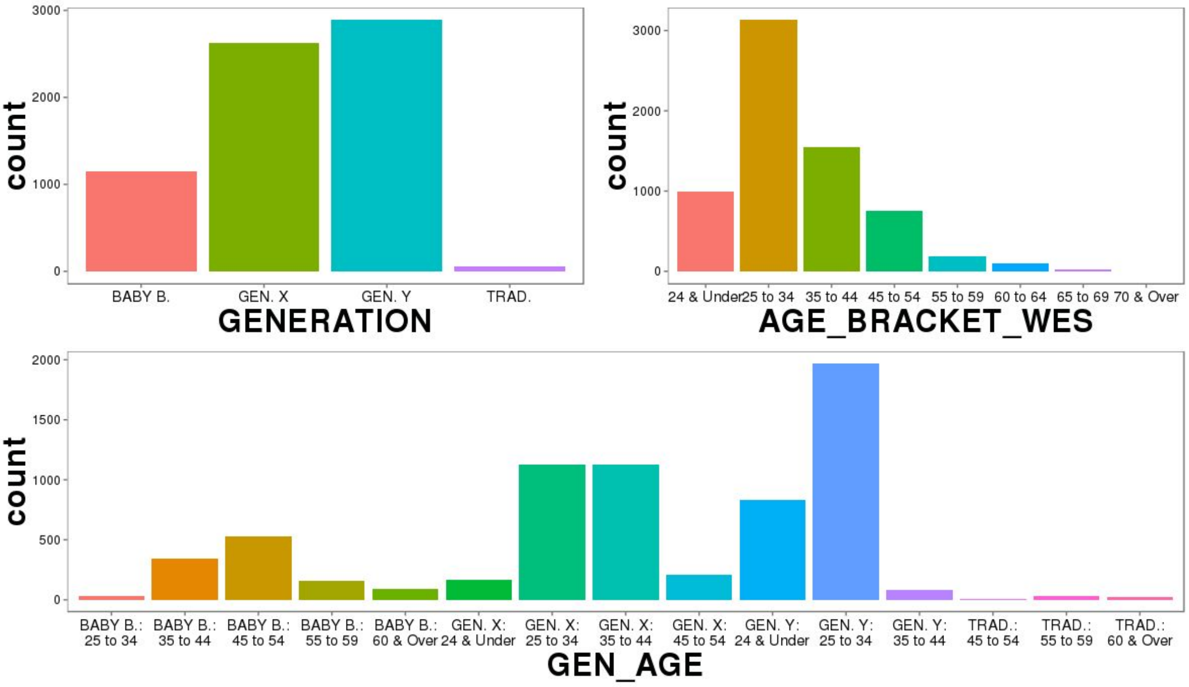

Figure 6. Top two plots are distributions of variables provided by the client, which we "merge" into the GEN_AGE *after some releveling of the* AGE_BRACKET_WES *variable.* GEN_AGE*'s distribution is given by the bottom plot; note the decreased number of age brackets in* GEN_AGE *compared to* AGE_BRACKET_WES *still leads to unavoidable low number of observations for levels involving traditionalists, which contributes to some of the wide error bars in the plots under Secondary results (p. 8).*

GEN_AGE exhibits additional problems aside from balance—the temporal limitation of the dataset's time span (1996-2015), as explored in tables 3 and 4.

GENERATION	Resignation age: lower bound	Resignation age: upper bound
Traditionalists: ≤ 1945	51	N/A
Baby Boomers: [1946, 1964]	32	69
Generation X: [1965, 1978]	18	50
Generation Y: [1979, 2000]	N/A	36

Table 4. Given that the dataset covers 1996-2015, specifying an employee's generation implies that their age at the time of voluntary resignation must fall within certain bounds.

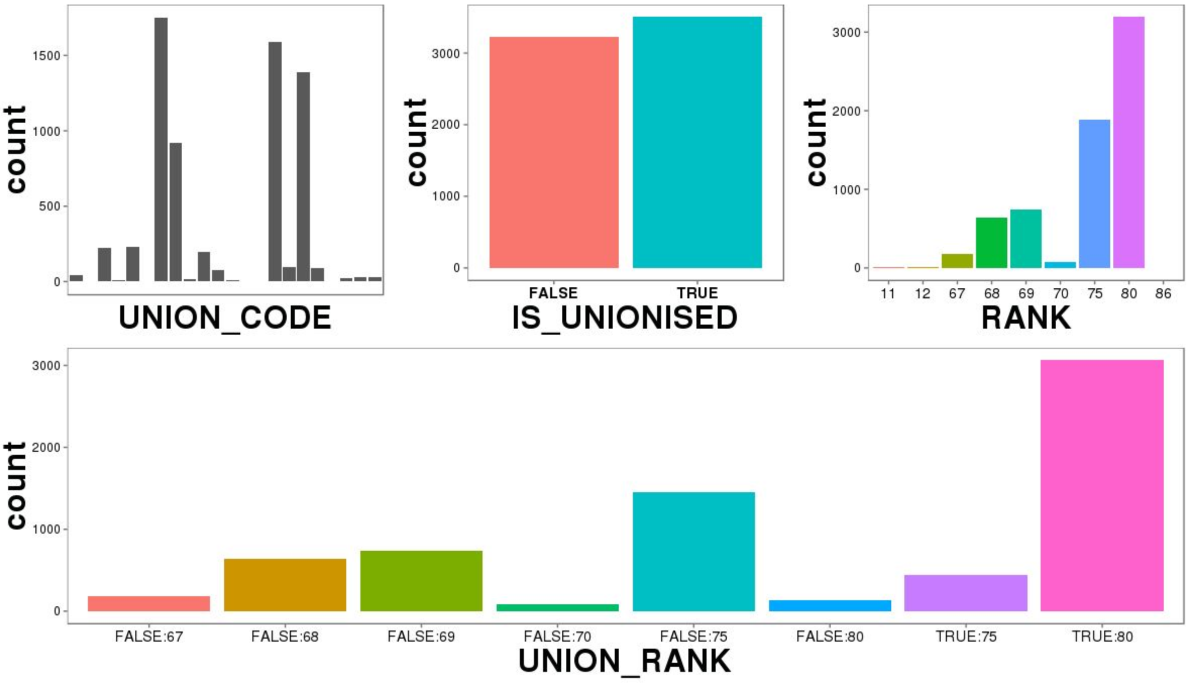

Figure 7. Distributions of `UNION_CODE` *and* `RANK`*, which are client-provided variables. Since* `UNION_CODE` *has similar granularity issues as* `JOB_CODE` *described above, though to a lesser degree, we created a binary variable* `IS_UNIONISED` *indicating union membership.* `IS_UNIONISED` *and* `RANK` *are nearly collinear, as* `RANK` *is partially derived from* `UNION_CODE`*, and so our solution is combining* `IS_UNIONISED` *with* `RANK`*, where observations with ranks 11, 12 and 86 are removed based on client information.*

Predicted marginal means and associated confidence intervals

Given a level of a categorical variable or a continuous variable of interest, the following steps construct the associated predicted marginal mean and 95% confidence interval:

1. Take the estimated linear model coefficient of the variable of interest and construct a 95% confidence interval around the coefficient using the estimated standard deviation (standard error) of the coefficient. Note that the standard error decreases as the corresponding number of observations increases.
2. Calculate the average for each of the other variables, multiply by their corresponding estimated coefficient and take the sum.
3. Translate/move the entire confidence interval and the coefficient by the sum calculated in step (2).
4. Reverse the power transformation used for the final model (raise all values obtained after step (3) to the power of 3.30).

Secondary results

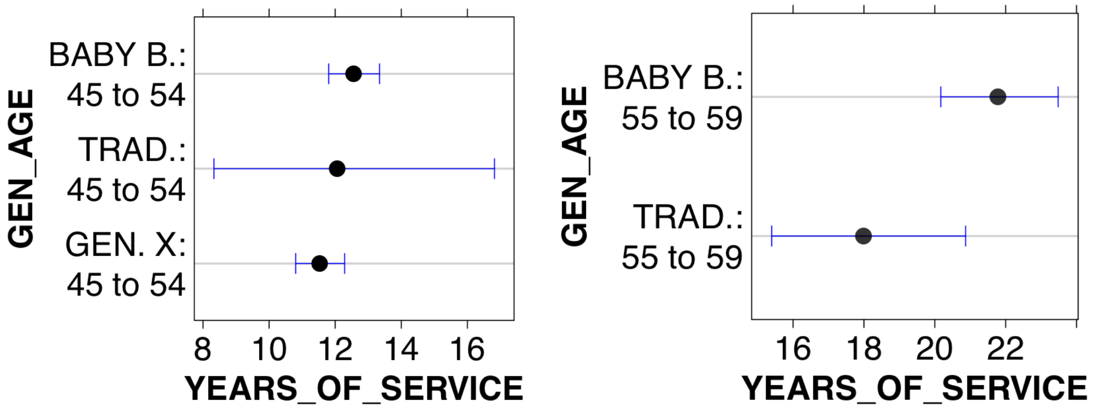

Figure 8. (Left) Among employees aged 45-54 at resignation, there was no significant difference in average YEARS_OF_SERVICE *between generations. (Right) Among employees aged 55-59 at resignation, there was no significant difference in average* YEARS_OF_SERVICE *between baby boomers and traditionalists. Again as discussed in the Results, by comparing the x-axis ranges, we find age at resignation has more practical effect on* YEARS_OF_SERVICE.

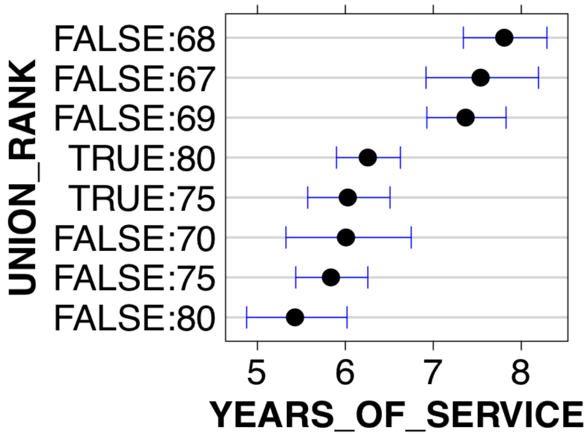

Figure 9. Union and rank membership splits into two subsets; employees in the one with ranks 67-69 (more senior/management positions) have higher YEARS_OF_SERVICE *on average than employees in the other subset containing the remaining levels.*

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science+Business Media, 2009.

Lenth, Russell. *Using lsmeans.* R Package Vignette. Date of access: 16 April 2016. https://cran.r-project.org/web/packages/lsmeans/vignettes/using-lsmeans.pdf.

L.J.P. van der Maaten and G.E. Hinton. *Visualizing High-Dimensional Data Using t-SNE*. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

Wickham, Hadley. *ggplot2 book*. Online book, 2013. http://ggplot2.org/book/.