RMIT UNIVERSITY

Course code	ISYS3447
Course name	Intro to Business Analytics
Lecturer	Dr. Timothy Mcbush Hiele
Student name	Tran Van Phuoc - s3825778
Assignment	Assignment 2 Visualisations and Predictive Analytics Report
Word Count	1978 – Excluding Tables, Figures, In-text Citations & Reference

EXECUTIVE SUMMARY

This report aims to study how career advancement pace is perceived between genders and age groups among IntelliAuto's workforce, thereby supporting recommendations for the company to address the potential issues. Moreover, the paper also studies the key determinants of employees' employed years by conducting multiple linear regression and then using those determinants as predictors for the employee's retirement. The paper comes up with two main findings, which are the occurrence of gender and age disparity in career advancement among employees, and the key determinants of employees' employed years, which are their worked years, engagement rate, and the number of past promotions.

Table of Contents

I. INTRODUCTION

IntelliAuto is a manufacturer in the automobile industry that employs up to 5000 individuals. The firm has set a goal to upgrade its manufacturing process to align with Industry 4.0 within the next five years. Implementing new technologies is anticipated to improve production quality, increase efficiency, and the firm's competitiveness, leading to enhanced outcomes with fewer inputs. This report has been written by a data analyst at Datos Lab Corporation to study IntelliAuto's workforce situation through a survey of a random 1000-employee sample. The report discusses employees' perception of job advancement by gender and age groups; identify the key determinants of employee's employed years; and provides predictions about the retirement scenario based on those determinants.

II. DATA PRE-PROCESSING

1. Data Validation

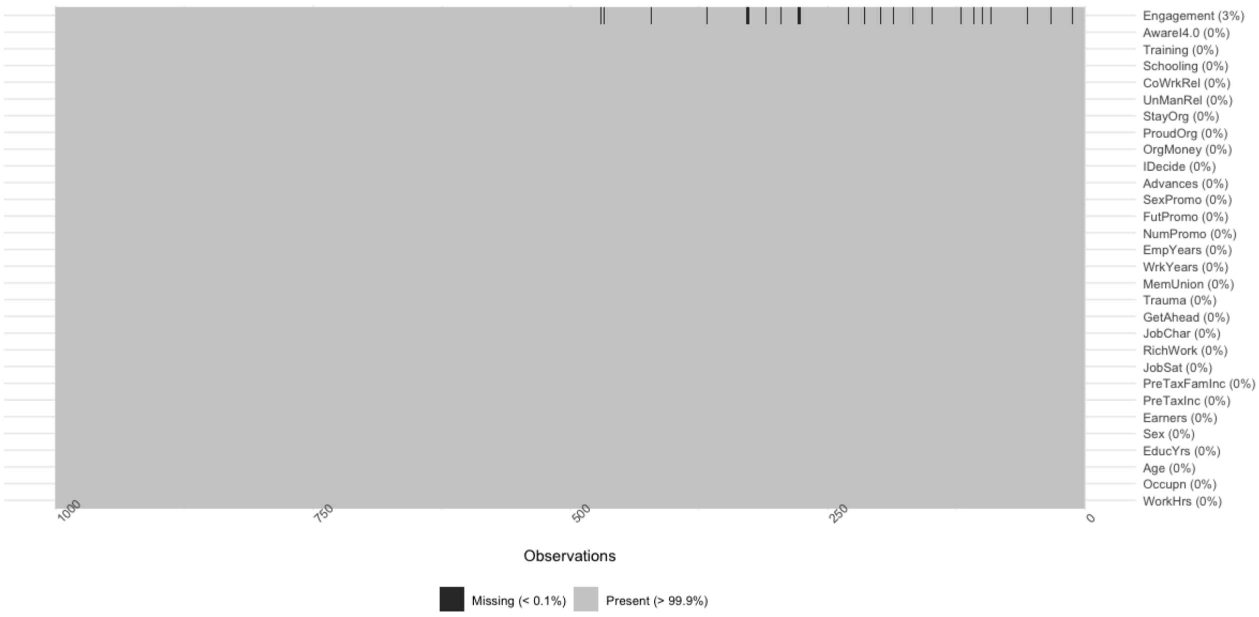


Figure 1: Missing value plot of Intelli Auto's surveyed dataset.

For this study, a dataset of IntelliAuto employees' surveyed answers, consisting of 1000 observations, is imported to RStudio; however, the Engagement column experiences up to 27 missing values (3% of total observations – Figure 1)[1]. Since data missingness could lead to a diminished statistical power of a study and produce biased estimates, missing data treatment is necessary (Kang 2013). Nevertheless, because of the large sample size and

[1] Missing value plot is conducted from miss_vis function.

the insignificance of missing value quantity, omitting rows containing missing values is the best method thanks to its simplicity. Consequently, 27 observations with missing values are excluded from the dataset.

2. Outliers Detection & Treatment

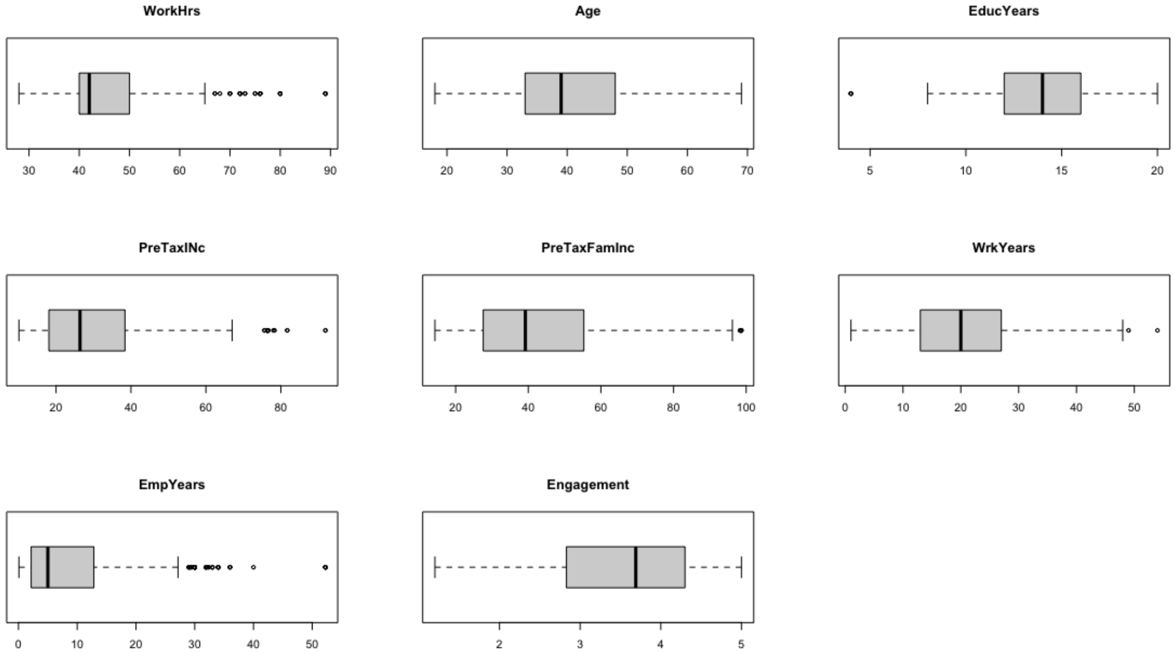


Figure 2: Box plots illustrate outliers of numeric variables[2]

According to Figure 2, most of the numerical variables contain extreme values (except Age and Engagement variables). Since linear regression is extremely sensitive to outliers, causing the model to be less reliable, outlier treatment is necessary, such as outlier removal. Nevertheless, removing outliers could cause more harm than good in this case, as the upper-bound outliers of EmpYears are the ones who are most likely to retire soon, owing to their old ages, and these outliers can provide additional information about the relationship between dependent and independent variables. Thus, removing upper-bound outliers of EmpYears variable could drastically reduce the model's accuracy as the model would not account for the characteristics of observations who are about to retire. Consequently, outliers are retained in the dataset to ensure the model's accuracy for near-retirement predictions.

[2] Despite being categorized by RStudio as numeric variables, Trauma, NumPromo, and Earners variables are specifically treated as categorical variables, owing to their small number of distinct values and high frequency (mode) characteristics.

III. METHODOLOGY

1. Multiple Linear Regression (MLR)

MLR is a statistical method applied to examine whether there are relationships between one dependent variable and multiple explanatory variables (Bell 2018). In the context of predicting an employee's years of employment, the dependent variable ($\hat{y}$) would be the years of employment (EmpYears), and the predictors (x_n) would be quantitative factors (such as the employee's age and income), and qualitative factors (such as employee's job satisfaction and their relationship with colleagues). Consequently, MLR's formula is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_n x_n$$

Where:

$\hat{y}$ is the response variable.

x_n is the predictor.

$\hat{\beta}_n$ is the predictor's slope of coefficient.

To conduct a complete MLR analysis, the following steps are taken. Firstly, bivariate analysis is conducted to identify the potential candidates that, together with the dependent variable, satisfy the linearity assumption. Secondly, the regression model is estimated by using the lm function of RStudio, and then regression results are provided, including the regression equation, coefficients, and statistical significance of the predictors. Thirdly, residual plots are analyzed to conduct residual diagnostics. Fourthly, the model's explanatory and predictive power are assessed by interpreting the model's adjusted R-squared and calculating the prediction errors, respectively. Fourthly, the interpretation of parameter's coefficients in the context of the study is provided if the predictors are statistically significant with the dependent variable. Lastly, predictions are produced from the regression equation based on the firm's retirement scenario.

2. Bivariate Analysis

Bivariate analysis refers to the analysis of determining whether there is a relationship between two variables. In the context of multiple regression, bivariate analysis can help examine the relationship between each predictor variable and the dependent variable. This helps identify the most potential predictors that can explain the greatest amount of variation in the response variable, thereby increasing model's goodness-of-fit (adjusted R-squared).

Moreover, bivariate analysis can also detect highly correlated pairs of predictors, thereby avoiding the occurrence of multicollinearity, which could lead to incorrect p-value and signs of coefficients.

Since there are two types of variables: numeric and categorical variables, different methods of bivariate analysis are applied. For numeric pairs of variables, scatterplots and correlation matrix are computed to determine the magnitude of the relationship between two continuous variables. Nevertheless, if one in two variables is categorical, grouped boxplots are preferred, as the illustration of one category having a higher median value than another category can indicate that there is a relationship between the categorical and numeric variables.

IV. FINDINGS

1. Job Advancement perceived among staff.

a. Perceived by gender.

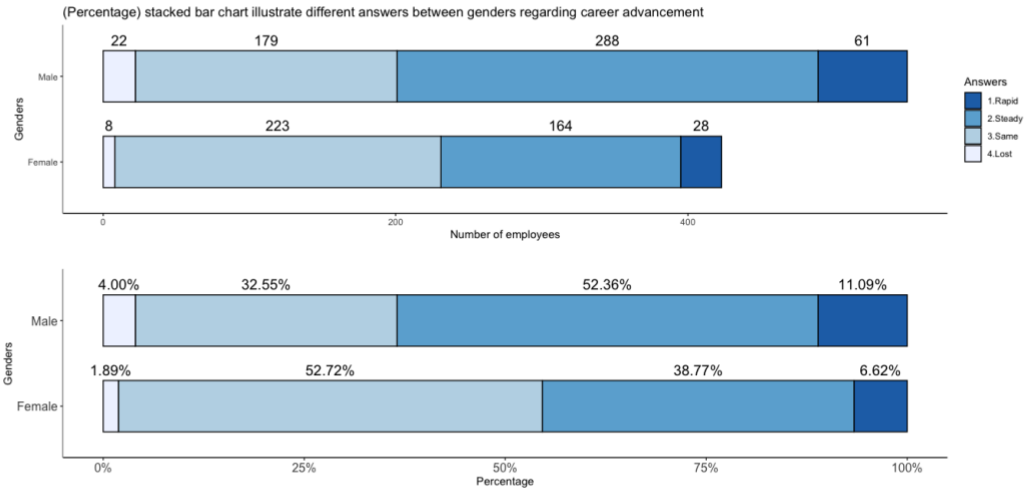


Figure 3: Likert plot illustrates different answers between the two genders regarding their career advancement pace.

According to Figure 3, regarding the male group, with 11.09% of "Rapid" and 52.36% of "Steady" answers, there are more male employees who find their career advancement progress smoothly and rapidly than the ones who do not(32.55% and 4% of "Same" and "Lost", respectively). Contrastively, there are more female employees who experience blockages in their career progress. Therefore, male employees are experiencing a faster career advancement pace than their counterparts in terms of both quantities and proportion.

b. Perceived by age groups.

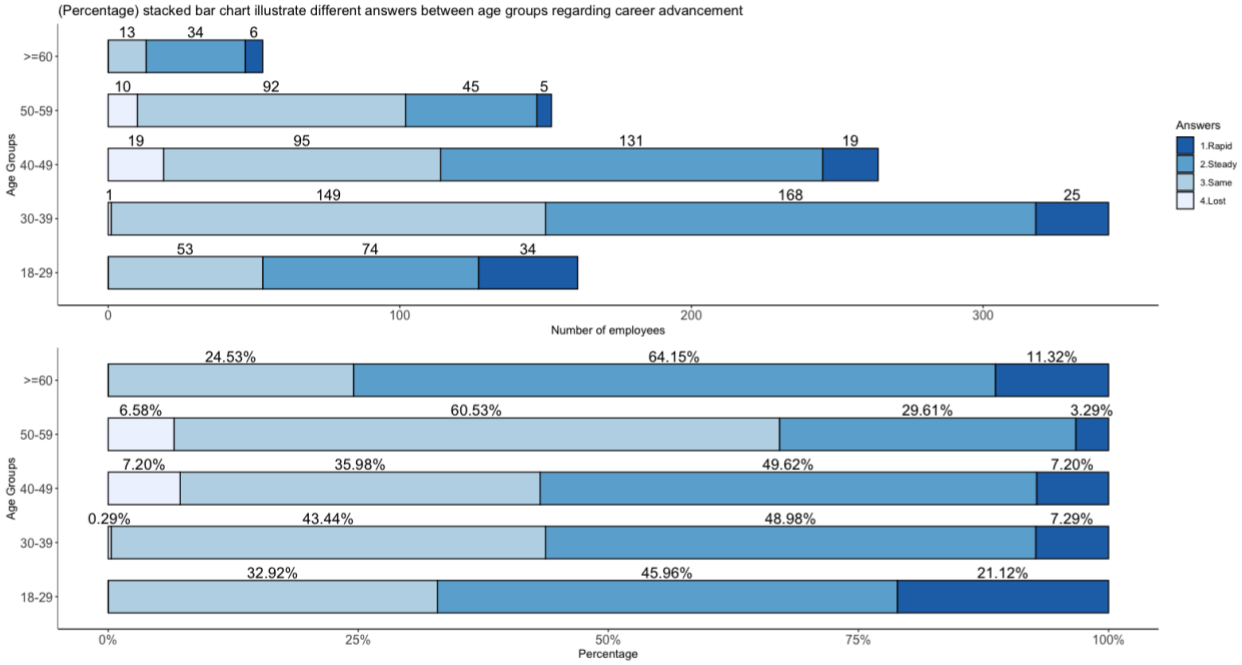


Figure 4: Likert plot illustrates different answers between age groups regarding their career advancement pace.

Overall, the older the employee is, the more difficult they find to advance their career, as the number of answers regarding “Rapid” decline per older age group[3]. Moreover, age groups between 30 and 59 years of age have the highest number of employees who find it difficult to advance their careers, resulting in being lost or stuck in their current position.

[3] the proportions of “Same” and “Lost” answers increase per older age group.

2. MLR Modelling

a. Bivariate analysis for predictor selection.

i. Numeric predictors

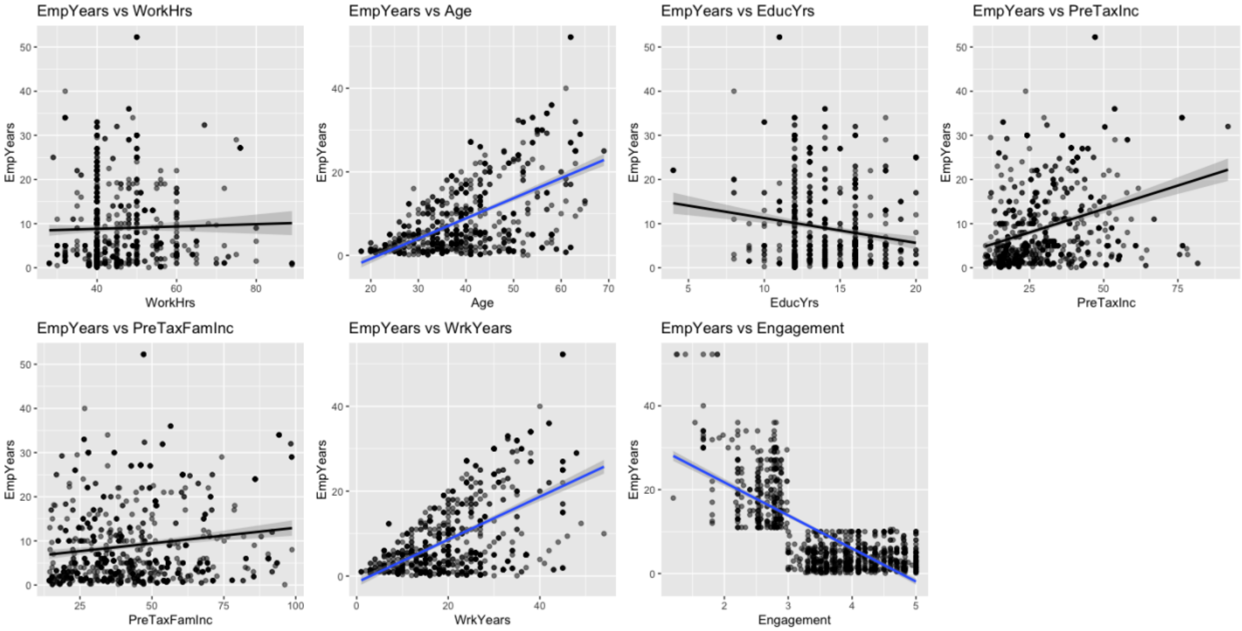


Figure 5: Scatterplots of EmpYears versus numeric variables of IntelliAuto Employee survey dataset.

According to Figure 5, there are only three variables that have noticeable linear patterns with EmpYears, which are Age, WrkYears and Engagement. Specifically, WrkYears and Age have a positive correlation with EmpYears, as their values increase along with the value of EmpYears. On the other hand, Engagement correlates negatively with EmpYears as its value increases along with EmpYears' decrease. Consequently, out of seven numeric variables: Age, WrkYears, and Engagement are the most potential numeric candidates for the regression model.

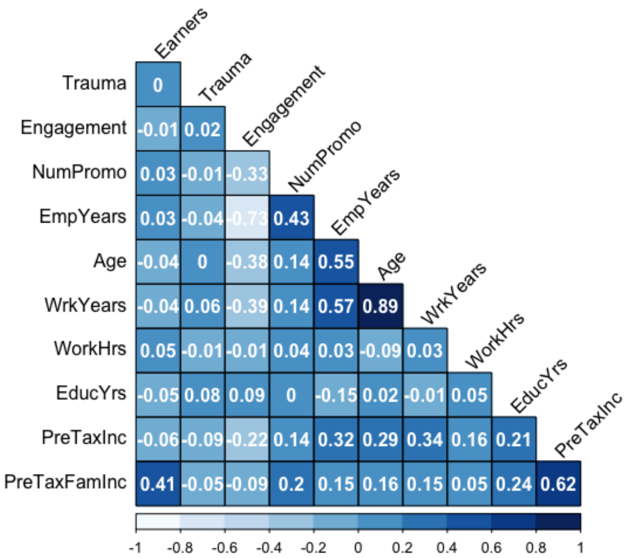


Figure 6: Correlation Matrix of numerical variables of Intelli Auto's Employee Survey dataset.

Figure 6 informs that the three candidates all have noticeable correlations with the response variable. However, the high correlation between predictors Age and WrkYears (0.89) suggests the occurrence of multicollinearity; therefore, one in two predictors should be removed from the model. In this case, the predictor Age is disqualified due to its lower correlation coefficient with EmpYears compared to WrkYears (0.55 vs 0.57).

ii. Categorical predictors

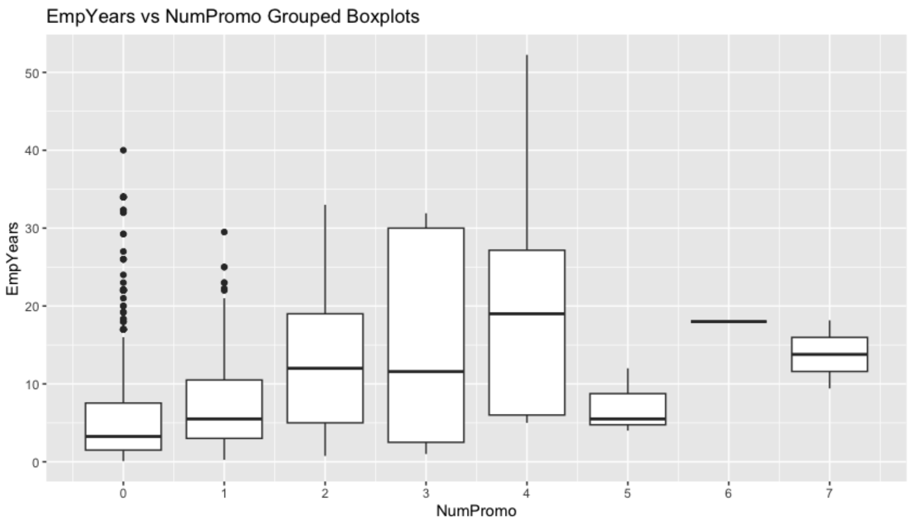


Figure 7: Grouped boxplots illustrate the relationship between EmpYears and NumPromo variables.

Despite having up to 22 categorical variables, there is only one predictor (NumPromo) that really signifies its linear relationship with the dependent variable by having its medians increased over the categories (Figure 7). Contrastively, the 21 remaining predictors have their grouped boxplot overlap with each other, which suggests that the medians of the numerical variable for each category of the categorical variable are similar, thereby not having any linear relationship with the response variable (Appendix 1).

iii. Final MLR model

$$\widehat{EmpYears} = \hat{\beta}_0 + \hat{\beta}_1\text{WrkYears} + \hat{\beta}_2\text{Engagement} + \widehat{\beta}_3\text{NumPromo1} + \widehat{\beta}_4\text{NumPromo2} + \widehat{\beta}_3\text{NumPromo3} + \widehat{\beta}_3\text{NumPromo4} + \widehat{\beta}_3\text{NumPromo5} + \widehat{\beta}_3\text{NumPromo6} + \widehat{\beta}_3\text{NumPromo7}$$

b. Regression Result

Dependent Variable: EmpYears				
Parameter	**Coefficient**	**Standard Error**	**t-Statistic**	**p-value**
(Intercept)	22.19	1.06	20.94	1.50E-80
WrkYears	0.30	0.02	16.94	1.47E-56
Engagement	-5.73	0.22	-25.86	2.06E-112
NumPromo1	0.86	0.46	1.88	0.06
NumPromo2	1.17	0.55	2.12	0.03
NumPromo3	4.06	0.67	6.09	1.65E-09
NumPromo4	8.24	0.69	11.89	1.67E-30
NumPromo5	0.94	3.08	0.31	0.76
NumPromo6	5.35	5.32	1.01	0.31
NumPromo7	-0.49	3.77	-0.13	0.90
Regression Statistics				
Multiple R-squared	0.6845			
Adjusted R-squared	0.6816			
Residual Standard Error	5.309 on 963 degrees of freedom			
F-statistic	232.2 on 9 and 963 DF			

Table 1: MLR's result.

Prediction Model Equation:

$$\widehat{EmpYears} = 22.19 + 0.3\text{WrkYears} – 5.73\text{Engagement} + 0.86\text{NumPromo1} + 1.17\text{NumPromo2} + 4.06\text{NumPromo3} + 8.24\text{NumPromo4} + 0.94\text{NumPromo5} + 5.35\text{NumPromo6} – 0.49\text{NumPromo7}$$

c. Residual Diagnostics

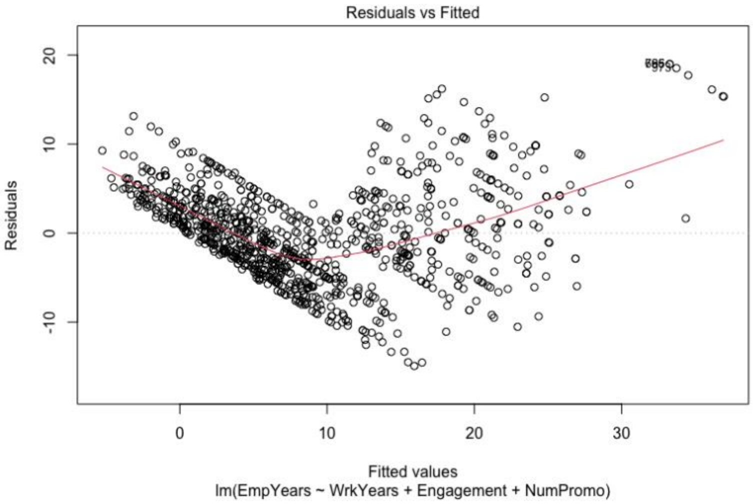


Figure 8: Fitted vs Residual plots of MLR model.

The residual plots inform that the MLR model suffers from heteroskedasticity[4], as the residuals are not evenly distributed across the fitted values: they start to spread out away from the regression curve from the fitted value of 10, implying the predictions of larger values are less reliable than lower values.

d. Model Accuracy

Goodness-of-Fit informs how well the observed data is explained by the regression model (Wooldridge et al. 2016). The regression model's adjusted R-squared is 0.6816, implying that 68.16% of total observed variations in employees' employed years are explained by the variations of 3 explanatory variables and the remaining 31.84% variations are explained by the variations of unobserved determinants. Hence, the model is a decent fit for the data.

[4] Heteroskedasticity is phenomenon where the variability of residuals of a regression model is not the same for all values of the predictor variable, affecting the accuracy of predictions (Rigobón 2003)

e. Coefficient Interpretation

Hypothesis testing (Significance level = 5%):

$$\begin{cases} H_0: \beta_n = 0 \ (The\ predictor\ has\ no\ effect\ on\ \text{EmpYears}) \\ H_a: \beta_n \neq 0 \ (The\ predictor\ has\ effect\ on\ \text{EmpYears}) \end{cases}$$

Independent variables	Coefficient	P-value vs Significance Level	Hypothesis testing outcome	Interpretation
WrkYears	0.30	1.47E-56 < 0.05	Reject H_0	Employee's current employed years increase by 0.2757 year for every 1-year increase in employee's full-time employed year, ceteris paribus.
Engagement	-5.73	2.06E-112 < 0.05	Reject H_0	On average, when employees' engagement rate increase by 1 unit, their employment years will decrease by 5.73 years, ceteris paribus.
NumPromo1	0.86	0.06 > 0.05	Fail to reject H_0	No Relationship
NumPromo2	1.17	0.03 < 0.05	Reject H_0	On average, employees who have 2 promotions have worked 1.17 years longer for current employer compared to employees who do not have any promotion, ceteris paribus.
NumPromo3	4.06	1.65E-09 < 0.05	Reject H_0	On average, employees who have 3 promotions have worked 4.06 years longer for current employer compared to employees who do not have any promotion, ceteris paribus.
NumPromo4	8.24	1.67E-30 < 0.05	Reject H_0	On average, employees who have 4 promotions have worked 8.24 years longer for current employer compared to employees who do not have any promotion, ceteris paribus.
NumPromo5	0.94	0.76 > 0.05	Fail to reject H_0	No Relationship
NumPromo6	5.35	0.31 > 0.05	Fail to reject H_0	No relationship
NumPromo7	-0.49	0.90 > 0.05	Fail to reject H_0	No relationship

Table 2: Interpretation of statistically significant predictors' coefficients.

f. Retirement scenario prediction

Assume that the legal retirement age is 65 and the employee age group of above 55 has the highest probability to retire in the next five years. Thus, the characteristics of this age group is surveyed so that the medians of the group are served as predictors' values for IntelliAuto's retirement scenario:

>= 55 Age group	Median
WrkYears	38
NumPromo	2
Engagement	2.78

Table 3: Median of three independent variables of employee age group who is 55-year-old or older.

Next, the values of the scenario are fitted into the regression equation to estimate the EmpYears's median of employee group that is likely to retire in the next five years:

$$\widehat{EmpYears} = 22.19 + 0.3\text{WrkYears} – 5.73\text{Engagement} + 1.17\text{NumPromo2}$$

$$\Leftrightarrow \widehat{EmpYears} = 22.19 + 0.3 \times 38 – 5.73 \times 2.78 + 1.17 \times 1$$

$$\Leftrightarrow \widehat{EmpYears} = 18.966$$

Consequently, the model predicts that the employee group who are most likely to retire in the next 5 years have the median of employed years of 18.966 ≈ 19. Thus, Intelli Auto must count the number of employees who have employed years of 19 years or more to prepare for retirement plan and their transition to digital manufacturing.

Number of employees with more than 19 employed years with Intelli Auto
159

Table 4: Number of employees with more than 19 employed years with IntelliAuto.

V. DISCUSSION

1. Gender and age disparity in career advancement

As there are a higher proportion of female employees experiencing stuck and lost at their current role compared to their men colleagues, women at Intelli Auto are facing gender disparity in career advancement. The difference between the two genders' answers might result from the following reason. Women at Intelli Auto receives fewer promotional opportunity compared to men (Appendix 2), as they are primarily assigned to less challenging roles (Admin Support) compared to men's dominated roles (Managerial, Professional, Technical) (Appendix 3), thereby resulting in less chance for women to have exposure to professional training to have access to high-level responsibility and tasks that are prerequisites for career advancement (Lyness & Thompson 1997; De Pater et al. 2010).

Furthermore, older employees also feel stagnant and uncertain in their career advancement, which might result from their limited knowledge and obsolete skills that are essential for utilizing modern technologies (Alhlouh & Kiss 2022). In fact, the older employees' age groups are less aware of Industry 4.0 compared to younger age groups (Appendix 4), implying that they find it complex and unfamiliar to work with new technologies, which are essential skills for their career progression. Consequently, the outdated skillset of older employee groups might result in challenges for Intelli Auto's transition plan to Industry 4.0.

2. Determinants of employees' employed years (retirement)

By using MLR analysis, IntelliAuto can predict the median employed years of the employee group that has the highest probability to retire in the next 5 years: the predicted group consists of 152 employees whose median employed years is 18.966, almost 19 years (Table 5). That group

Nevertheless, despite having a decent adjusted R-squared of 66.18%, the predictions might not be reliable as the model suffers from heteroskedasticity.

Moreover, the company can also identify the significant determinants of employees' employed years, thereby monitoring those determinants to increase employees' employed years (retention) and reduce the probability of old employees retiring early. For instance, the MLR informs that the longer the employees worked years, the lower their engagement rate; therefore, to increase their retention rate, IntelliAuto should improve their employee engagement toward job first.

VI. RECOMMENDATION

Firstly, to solve the gender disparity in the career advancement of IntelliAuto - a company that belongs to a men-dominated industry - the firm should introduce new policies that focus on the development of female employees. Specifically, the company should introduce functional and management trainee programs for women that provide training about how to perform managerial tasks; and position rotation every six months so that female employees can get their hands on different roles, thereby improving their knowledge, technical and leadership skills set that are not only benefit their career progression in the automobile industry but also improving the quality of IntelliAuto's workforce.

Secondly, regarding the solution for age disparity in job advancement, the company should also introduce workshops and training sessions for older employees to advance their technological skills, such as training sessions on how to communicate with artificial intelligence (A.I) or interact with robotic machinery, thereby enhancing their working performance. Moreover, they should hire industry expertise as career development consultants to provide the necessary navigation, roadmap and advice for employees who are lost in their mid-career development.

Lastly, regarding IntelliAuto's plans for retirement and digital manufacturing transition, the company should prepare enough robotic machinery or A.I tool that can handle the workload of 159 employees who are likely to retire in the next five years; therefore, the company will not face workforce shortages when employees retire unexpectedly. Another recommendation is that IntelliAuto should consider methods of improving employee engagement towards the job, as the engagement rate is identified as a statistically significant determinant of employees' employed years, thereby lowering the probability of early retirement among employees (improving retention rate).

VII. CONCLUSION

To sum up, by conducting a bivariate analysis of a random 1000 IntelliAuto's employees, the report discovers that there are gender and age disparity in job advancement among the workforces. Moreover, with MLR analysis, the paper also identifies key determinants of employees' employed years, which help to predict the number of employees who are likely to retire before the legal age of 65. With the aforementioned recommendations, IntelliAuto should successfully address those issues.

REFERENCE

Alhloul A and Kiss E (2022) 'Industry 4.0 as a Challenge for the Skills and Competencies of the Labor Force: A Bibliometric Review and a Survey', *Sci*, 4(3):34.

Bell N (2018) *Introduction to statistics*, Tritech Digital Media, Los Angeles, CA.

De Pater IE, Van Vianen AE and Bechtoldt MN (2010) 'Gender differences in job challenge: A matter of task allocation', *Gender, Work & Organization*, 17(4):433-453.

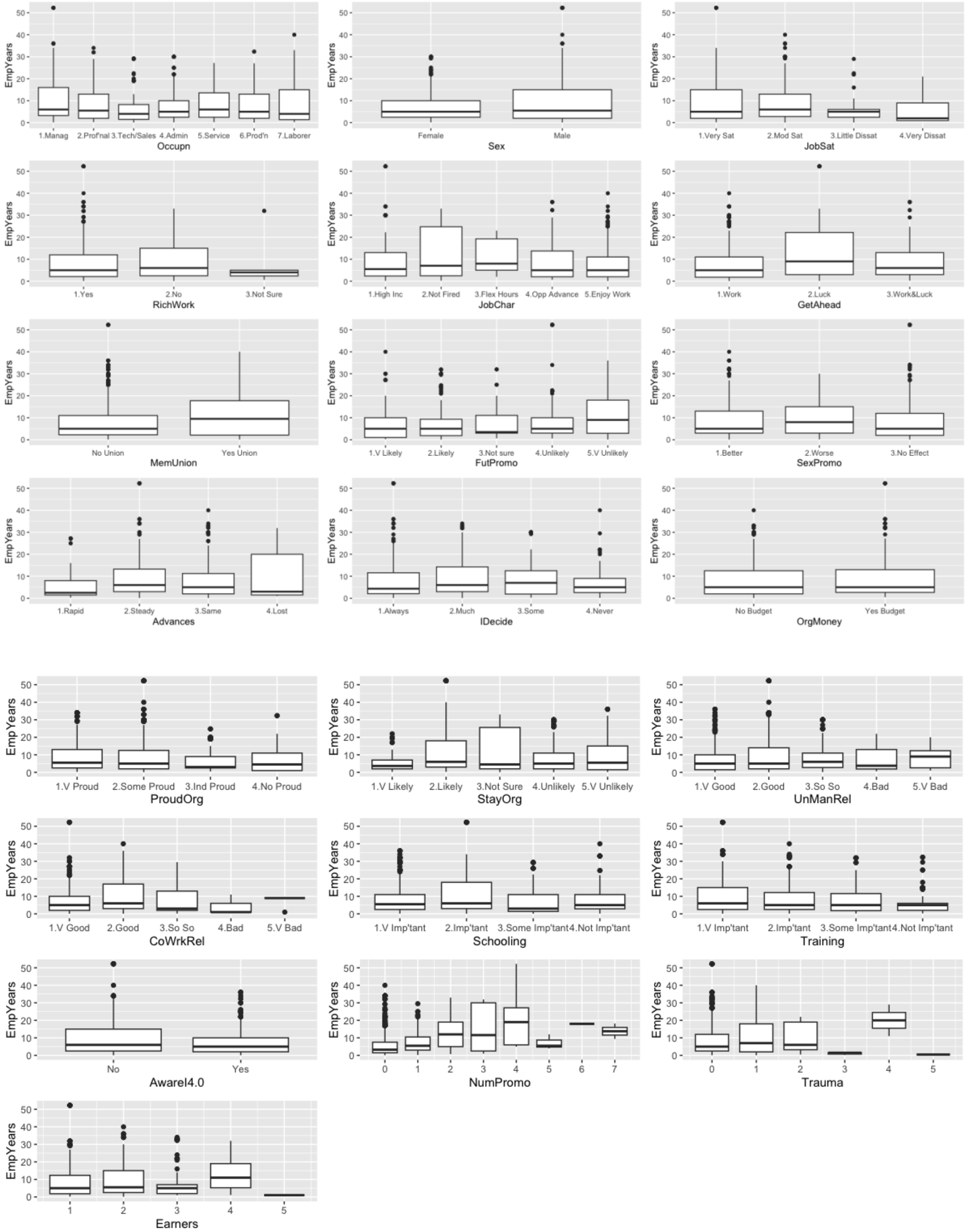
Kang H (2013) 'The prevention and handling of the missing data', *Korean journal of anesthesiology*, 64(5):402-406.

Lyness KS and Thompson DE (1997) 'Above the glass ceiling? A comparison of matched samples of female and male executives', *Journal of applied psychology*, 82(3):359.

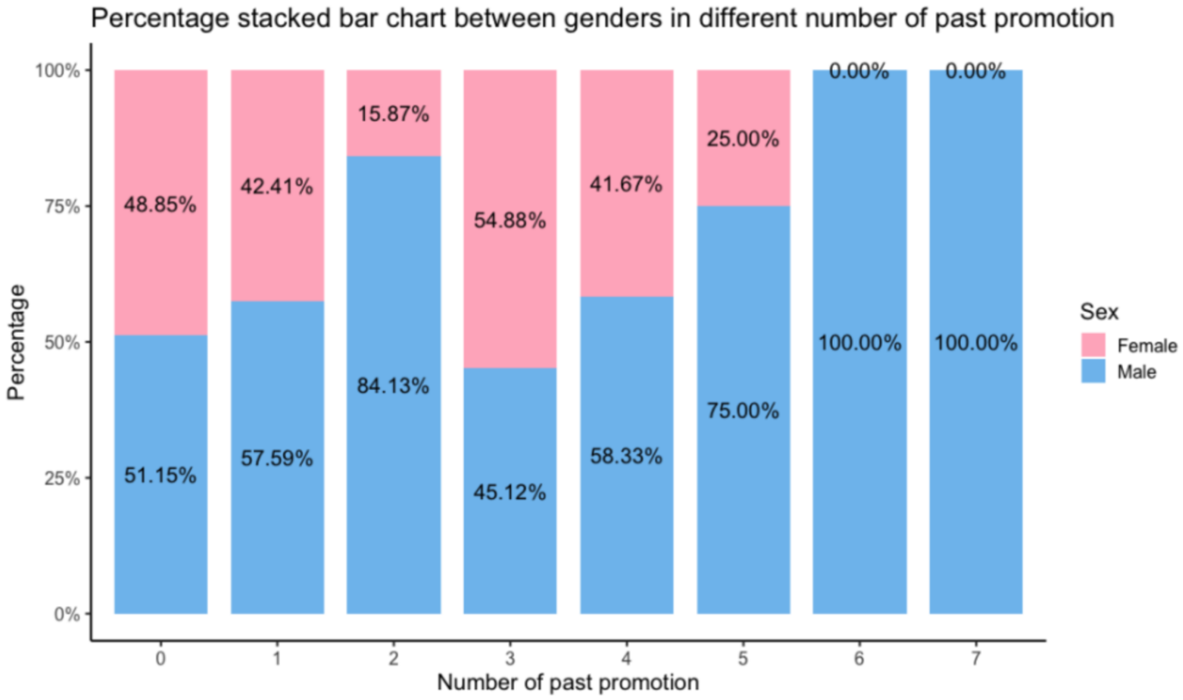
Rigobón, R 2003, ‘Identification through heteroskedasticity’, *Review of Economics and Statistics*, vol. 85, no. 4, pp. 777-792.

Wooldridge JM, Mokhtarul W and Jenny L (2016) *Introductory Econometrics: Asia-Pacific Edition. Cengage Australia*, Cengage Australia, Melbourne.

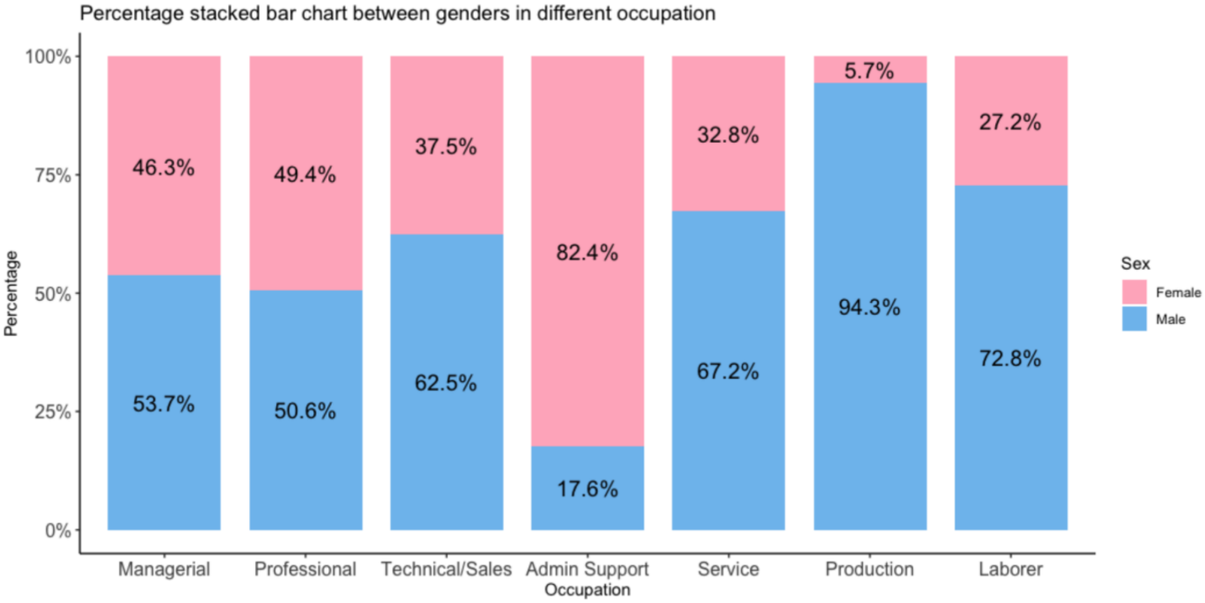
APPENDICE



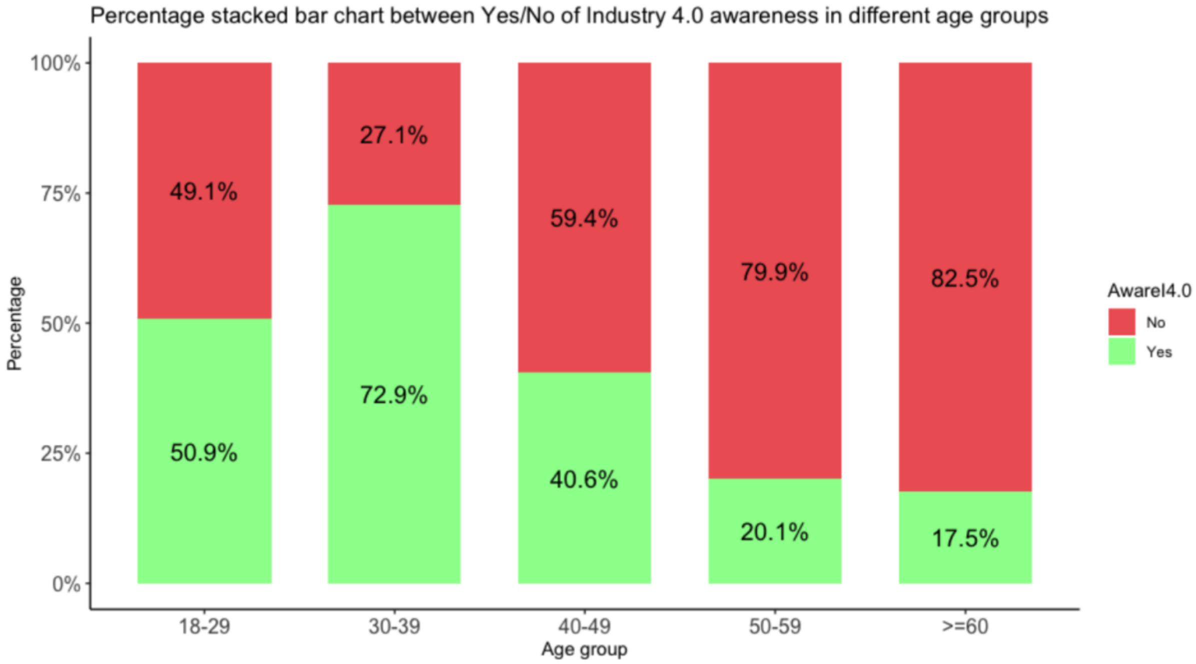
Appendix 1: Grouped boxplots illustrates the distribution of EmpYears among categorical variables.



Appendix 2: Percentage stacked bar chart between genders in different number of past promotions.



Appendix 3: Percentage stacked bar chart between genders in different occupation.



Appendix 4: Percentage stacked bar chart between Yes/No answers of Industry 4.0 awareness question among different age groups.