

3.0 Methodology

3.1 Data Source and Tool Used

The dataset used in this study for evaluating employee promotions was obtained from Kaggle (Zaman, 2021). The dataset comprises various employee-related information (Figure 2). The dataset contains a total of 54,808 observations. To ensure a representative sample for the study, a simple random sampling with control sorting was used. Using a specified seed (123), 3,500 observations were randomly sampled for further analysis (Figure 2). The software, SAS Studio, is used in this study to perform data pre-processing, data transformation and feature engineering techniques.

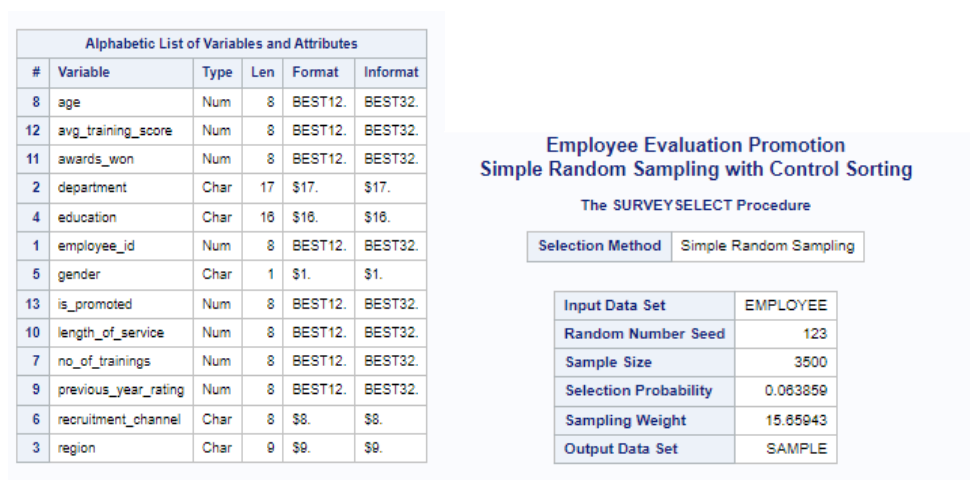


Figure 2. The metadata and simple random sampling of the dataset with control sorting.

3.2 Data Pre-processing

A dataset may contain noisy, missing and inconsistent data. Hence, data preparation is important to improve data quality and results. The technique involves data cleaning, handling of missing values and outliers.

3.2.1 Data Cleaning

First, cleaning of data is applied to address duplicates and inconsistent values. Duplicates should be removed to avoid bias. In the dataset, each employee has a unique employee ID, which is used to check for duplications. The analysis revealed no duplicate employee IDs (Figure 3). Additionally, the value "HR" under the department attribute is renamed to "Human Resources," and the value "R&D" is renamed to "Research & Development" (Figure 4).

Total rows: 0 Total columns: 2

	employee_id	duplicate_count

Figure 3. The results of duplicates in the dataset.

department	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Analytics	352	10.06	352	10.06
Finance	179	5.11	531	15.17
HR	156	4.46	687	19.63
Legal	58	1.66	745	21.29
Operations	750	21.43	1495	42.71
Procurement	411	11.74	1906	54.46
R&D	68	1.94	1974	56.40
Sales & Marketing	1058	30.23	3032	86.63
Technology	488	13.37	3500	100.00

department	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Analytics	352	10.06	352	10.06
Finance	179	5.11	531	15.17
Human Resources	156	4.46	687	19.63
Legal	58	1.66	745	21.29
Operations	750	21.43	1495	42.71
Procurement	411	11.74	1906	54.46
Research & Develo	68	1.94	1974	56.40
Sales & Marketing	1058	30.23	3032	86.63
Technology	488	13.37	3500	100.00

Figure 4. Rename the value under department attribute.

3.2.2 Missing Values

There are 271 missing values under the "previous year rating" attribute, 154 missing values under the "average training score" attribute, and 139 missing values in the "education level" attribute (Figure 5). Missing values can occur due to human errors, privacy issues, or interruptions in data collection, and they can have an impact on the performance of prediction models (Rençberoğlu, 2019). Imputation techniques are used to handle missing values to remedy this. While numerical missing values are replaced with the mean or median, categorical missing values are substituted with the mode, which is the attribute's most prevalent value. In this dataset, the missing values under "previous year rating" and "average training score" are imputed with mean, while missing values under "education level" is imputed with the mode, specifically "Bachelor's". Imputing missing values with the mode helps preserve the overall distribution and ensures that the imputed values represent the majority class. Imputing missing values with the mean maintains the overall average and reduces the impact on the distribution of the numerical variable. After the imputation, no missing values are detected (Figure 6).

The Missing Value of Numerical Data in Employee Promotion Evaluation Dataset

The MEANS Procedure

Variable	N Miss
employee_id	0
no_of_trainings	0
age	0
previous_year_rating	271
length_of_service	0
awards_won	0
avg_training_score	154
is_promoted	0

total_count	department_missing_count	education_missing_count	gender_missing_count	recruitment_missing_count	region_missing_count
3500	0	139	0	0	0

Figure 5. The total number of variables with missing values.

Missing Data Frequencies

The FREQ Procedure

department	Frequency	Percent
Non-missing	3500	100.00

region	Frequency	Percent
Non-missing	3500	100.00

education	Frequency	Percent
Non-missing	3500	100.00

gender	Frequency	Percent
Non-missing	3500	100.00

recruitment_channel	Frequency	Percent
Non-missing	3500	100.00

employee_id	Frequency	Percent
Non-missing	3500	100.00

no_of_trainings	Frequency	Percent
Non-missing	3500	100.00

age	Frequency	Percent
Non-missing	3500	100.00

previous_year_rating	Frequency	Percent
Non-missing	3500	100.00

length_of_service	Frequency	Percent
Non-missing	3500	100.00

awards_won	Frequency	Percent
Non-missing	3500	100.00

avg_training_score	Frequency	Percent
Non-missing	3500	100.00

is_promoted	Frequency	Percent
Non-missing	3500	100.00

Figure 6. The number of missing values among the variables after imputation.

3.2.3 Outliers

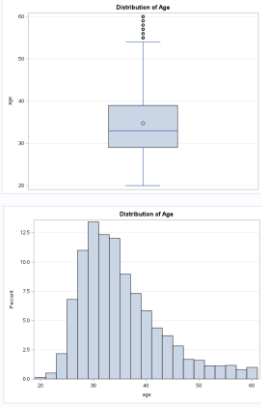
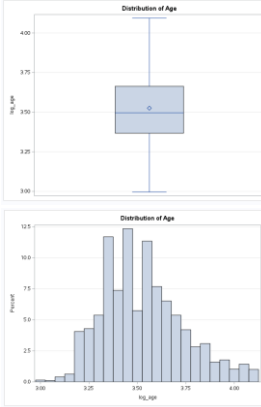
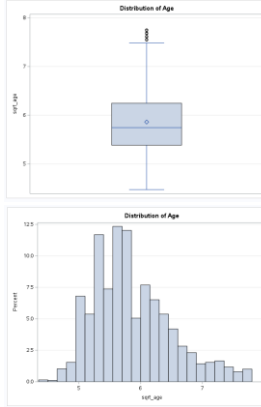
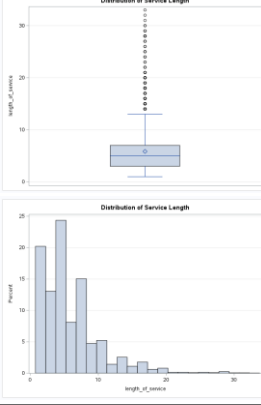
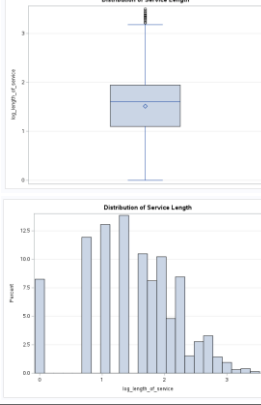
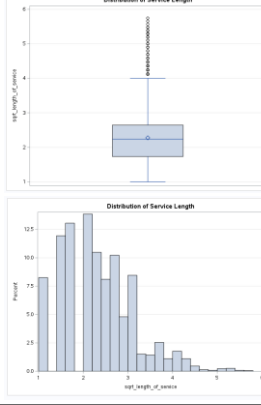
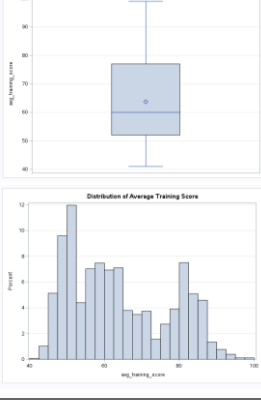
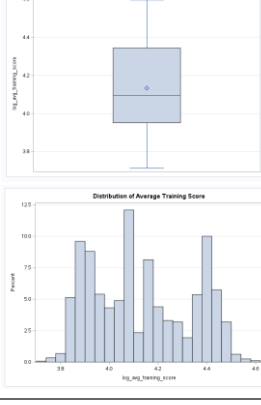
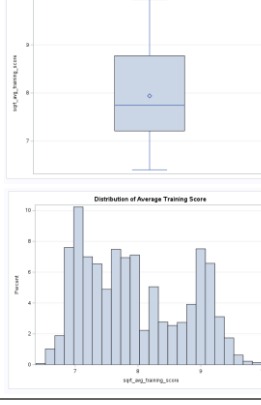
The handling of outliers can involve various approaches, including removal, imputation, and transformation. In this study, the removal technique is rejected due to the potential loss of valuable information, biased results, distorted distribution, and reduced statistical power. Similarly, the imputation approach is also rejected as it can introduce data distortion, bias, loss of information, and influence on statistical models, violating assumptions and impacting the interpretation and validity of the model.

Instead, the transformation technique is applied as it mitigates the impact of extreme values by normalizing the data, improving linearity, reducing leverage, and stabilizing variances. Logarithmic and square root transformations are commonly used techniques to modify the distribution of variables in data analysis. In this study, both transformations are applied to the age, service length, and average training score attributes for comparison (Table 1). The results indicate that the logarithmic transformation is preferred.

The logarithmic transformation helps reduce the magnitude of extreme values, making the distribution more symmetric. It also enhances linearity, improving the interpretability of

the models. By applying the logarithmic transformation, the range of values is compressed, resulting in more consistent variance across the data range and better visualization without the influence of extreme values. This transformation addresses issues related to highly skewed data distributions and unequal variances, effectively treating heteroscedasticity issues and improving the overall data quality.

Table 1. Comparison of original, logarithmic transformation, square root transformation of the attributes age, service length and average training score.

Attributes	Original	Logarithmic Transformation	Square Root Transformation
Age			
Service Length			
Average Training Score			
Observation	Logarithmic transformation is more efficient. The distribution of variable is more symmetric and bring the data closer to a normal distribution.		

3.3 Exploratory Data Analysis

This phase involves summarizing, visualizing, and gaining an in-depth understanding of the traits of a dataset (Halibas et al., 2019). Figure 7 presents a pie chart depicting the distribution of promotion status among the employees in the dataset. It shows that out of the 3,500 observations, 3,208 (91.66%) employees were not promoted, while 292 (8.34%) employees were promoted. This distribution indicates a high bias towards the non-promoted class, highlighting the imbalanced nature of the dataset in terms of promotion outcomes.

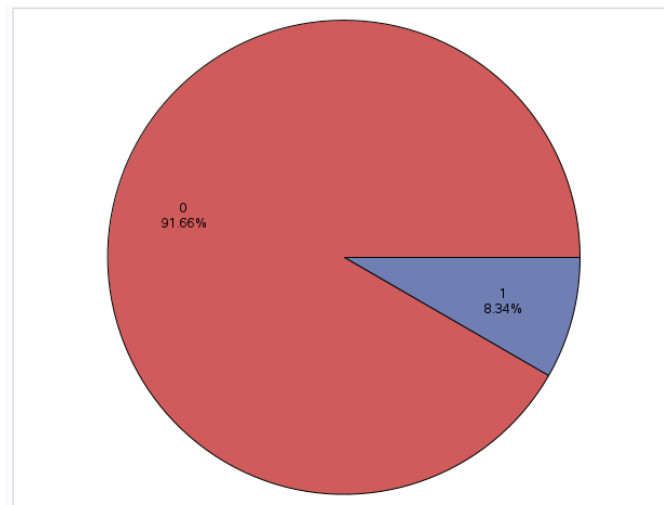


Figure 7. The pie chart of promoted and unpromoted employees.

3.3.1 Pearson's Correlation

The correlation coefficients and p-values in Figure 8 provide valuable insights into the relationships between variables in the dataset, as determined by Pearson correlation analysis. The coefficient threshold of a strong correlation is supported by 0.75 value, a moderate one falls between 0.45 - 0.75, and a weak correlation is below 0.45 (Halibas et al., 2019). A p-value below 0.0001 signifies statistical significance, indicating that the observed correlations are unlikely to have occurred by chance alone.

Based on the analysis in Figure 8, several observations can be made. First, moderate positive correlation ($r = 0.63440$, $p < 0.0001$) between the logarithm of age and the logarithm of service length. This suggests that as individuals' age increases, their length of service tends to increase as well. Furthermore, a negative relationship is discovered from training numbers with the logarithm of age ($r = -0.08280$, $p < 0.0001$). It indicates as individuals' age hikes up, the number of trainings they undergo tends to decrease. Additionally, the logarithm of average

training score exhibits a negative correlation with the logarithm of age ($r = -0.07291$, $p < 0.0001$) and a positive correlation linked to the previous year rating ($r = 0.07557$, $p < 0.0001$). These findings suggest that as individuals' age increases, their average training scores tend to decrease, while higher previous year ratings are associated with higher average training scores. There are no strong correlation observed based on the threshold 0.75 (Halibas et al., 2019); hence no attribute is removed. Employee ID was removed as it has no contribution to the target outcome.

Figure 9 displays the relationship between the features and the target outcome (promotion status). Positive correlations are observed between employee ID, previous year rating, awards won, and average training score with the target variable, while negative correlations are observed between the number of training sessions, age, and length of service with the promotion status (Figure 9).

The correlation between employee ID and promotion status is small and lacks statistical significance. This suggests that employee ID alone may not have a substantial impact on promotion prospects. A statistically significant negative correlation is found between the number of training sessions and promotion status, indicating that a higher number of training sessions may be associated with a lower likelihood of promotion. The correlation between \log_age and promotion status is negative, but it is not statistically significant. This suggests that age may not have a significant impact on an employee's promotion prospects. In contrast, previous year rating exhibits a strong positive correlation with promotion status, indicating that employees with higher ratings are more likely to be promoted. Both awards won and $\log_avg_training_score$ demonstrate strong positive correlations with promotion status, indicating that employees who have won awards or achieved higher average training scores are more likely to be promoted. Conversely, $\log_length_of_service$ shows a significant negative correlation with promotion status, suggesting that employees with longer service lengths may have a lower likelihood of being promoted.

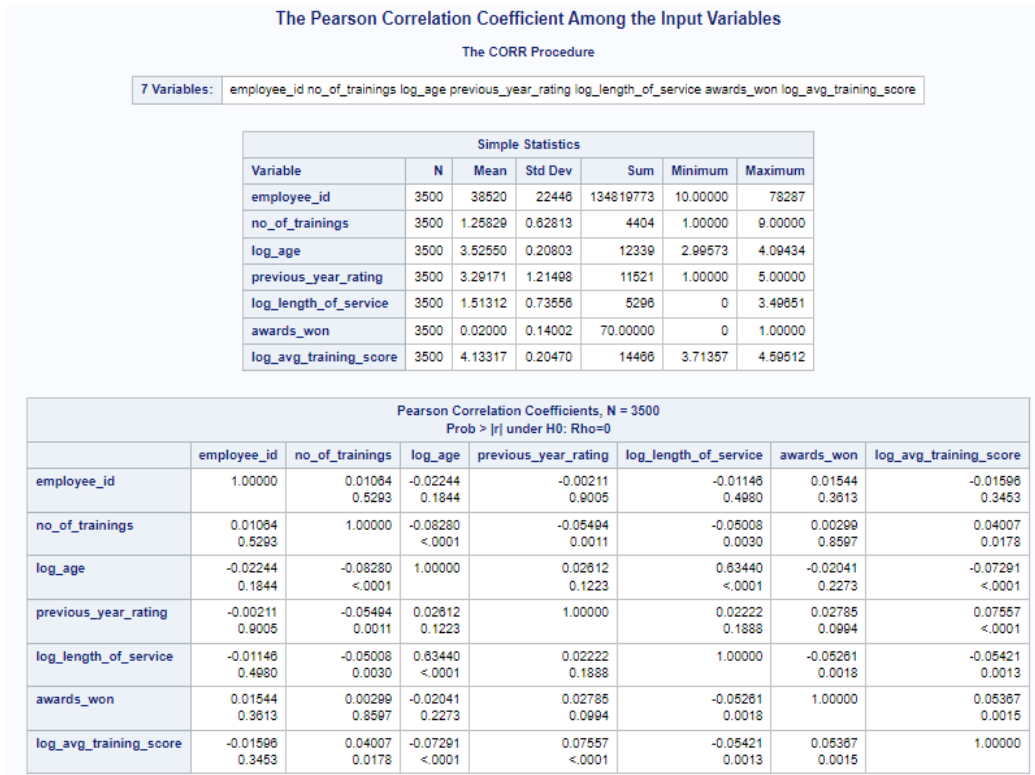


Figure 8. The Pearson's Correlation Coefficient among the input features.

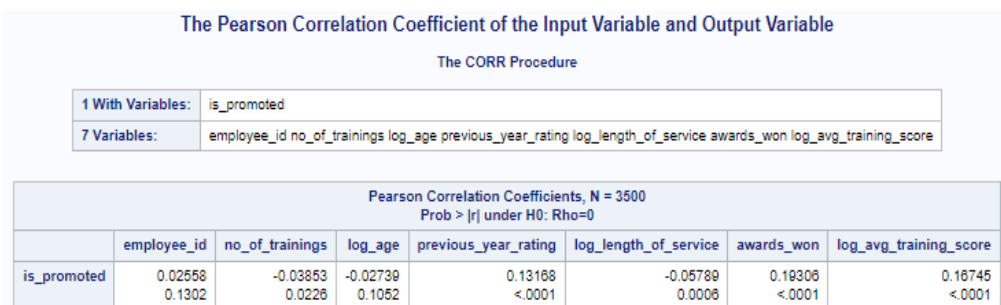


Figure 9. The Pearson's Correlation Coefficient of the input features and output feature.

3.3.2 Bivariate Analysis

Figure 10 to Figure 20 shows the frequency table and bar plot of the input variables by the promotion status. In Figure 10, department Sales & Marketing has the highest promoted employees (26.37%) followed by operations department (20.21%) and technology department (19.89%).

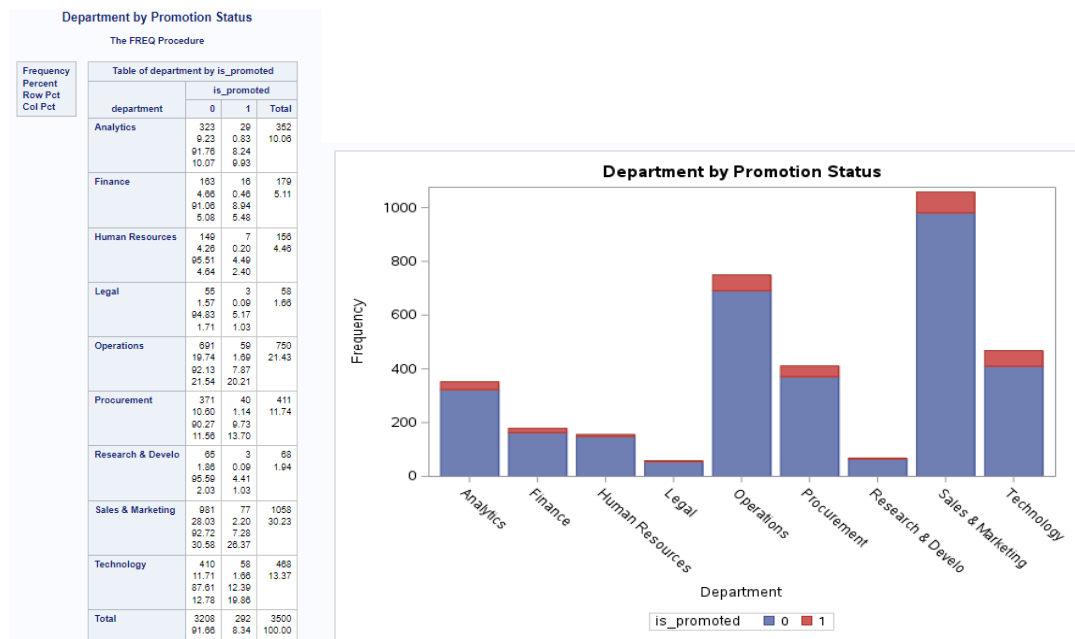


Figure 10. The frequency table and bar plot of the department by promotion status.

In Figure 11, most of the employee worked at region 2. It has the highest promotion rates (25.34%) among the regions. The lowest promotion rate is from region 6, region 9 and region 18 with 0% promotion rate.

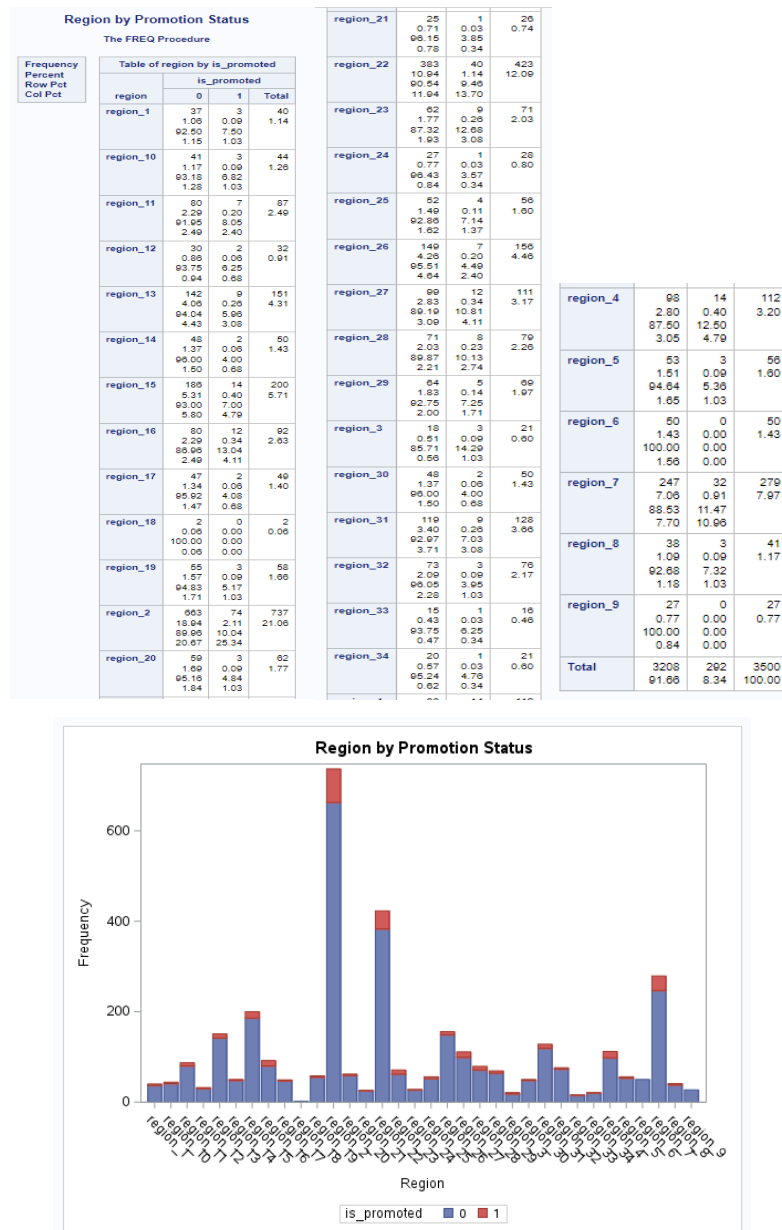


Figure 11. The frequency table and bar plot of the region by promotion status.

In Figure 12, most of the promoted employees have Bachelor's degree education level (65.41%), followed by Master's degree and above (32.53%) and below secondary (2.05%).

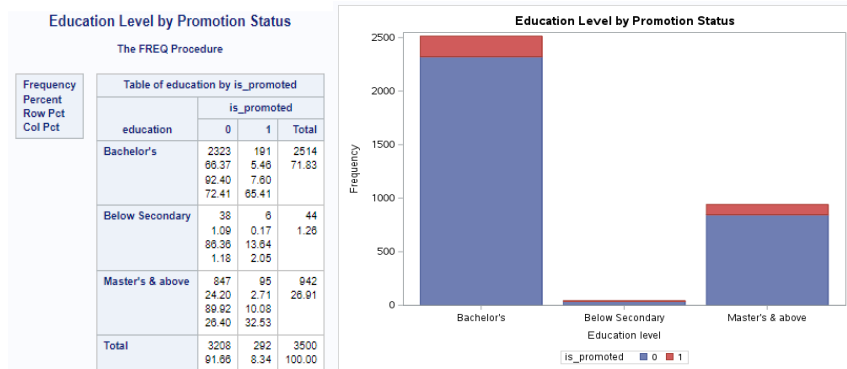


Figure 12. The frequency table and bar plot of the education level by promotion status.

In Figure 13, male (70.21%) has more promoted employees than female (29.79%).

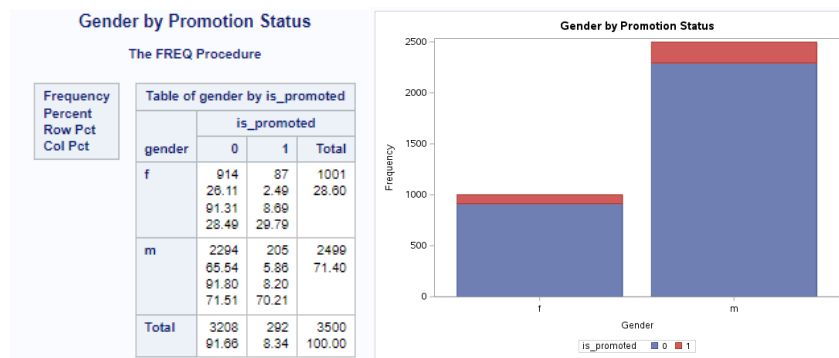


Figure 13. The frequency table and bar plot of the gender by promotion status.

In Figure 14, most of the promoted employees recruited from other channel (56.16%), followed by sourcing channel (40.07%) and referral (3.77%).

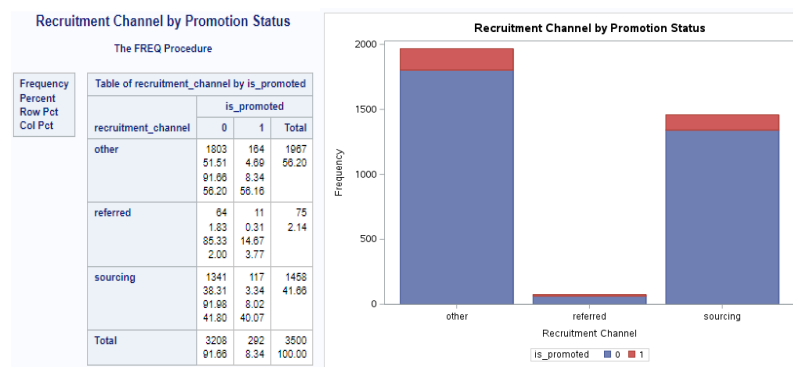


Figure 14. The frequency table and bar plot of the recruitment channel by promotion status.

In Figure 15, the number of employees and promotion status trends decreased when the training number is increased.

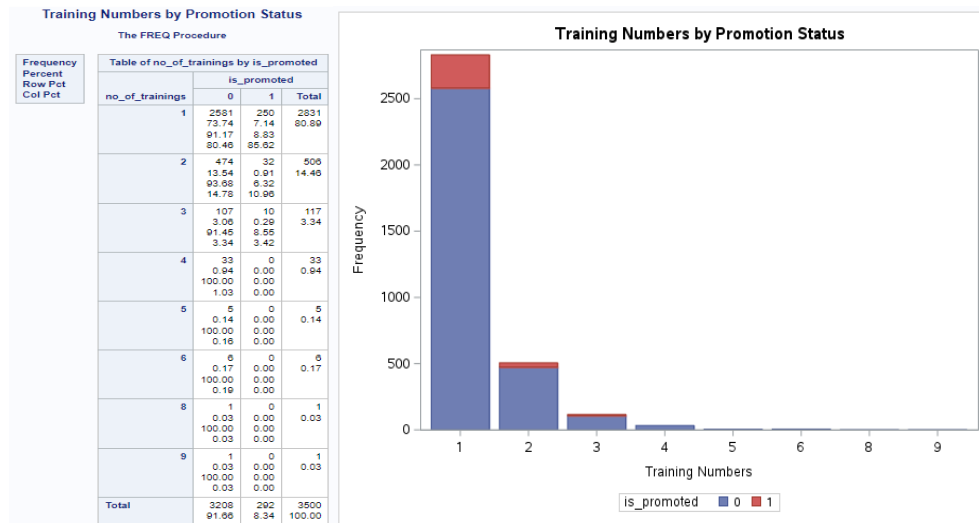


Figure 15. The frequency table and bar plot of the training numbers by promotion status.

In Figure 16, previous year rating of 5 has the highest promotion rate up to 38.01% followed by previous year rating of 3 (37.67%) and previous year rating of 4 (16.78%).

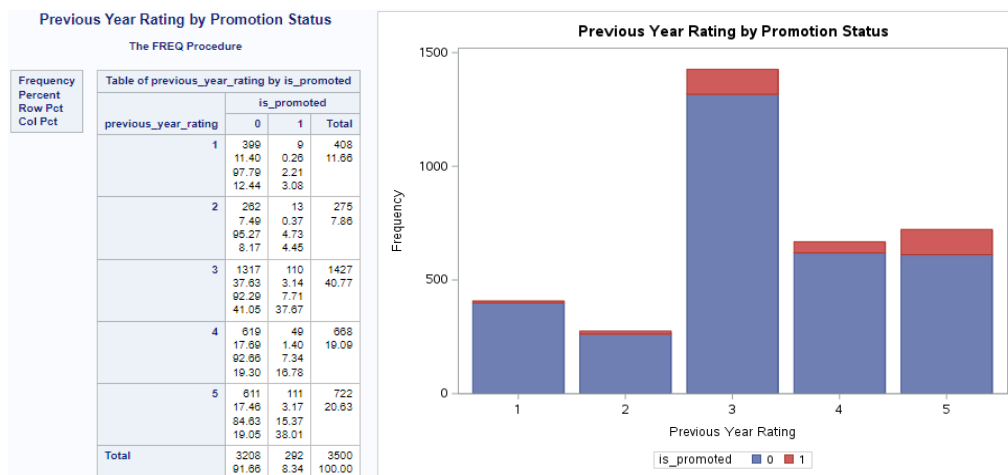


Figure 16. The frequency table and bar plot of the previous year ratings by promotion status.

In Figure 17, the range and mean of log age in promoted employees is lower than unpromoted employees.

Log Age by Promotion Status					3.36725853					3.6888794541					3.9318256327				
The FREQ Procedure																			
Frequency Percent Row Pct Col Pct	Table of log_age by is_promoted				3.4011973817				3.7135720667				3.9512437186						
	is_promoted																		
	log_age	0	1	Total															
	2.9957322736	5	0	5	190	22	212	62	9	101	16	5	21						
		0.14	0.00	0.14	5.43	0.03	5.46	2.63	0.09	2.90	0.46	0.14	0.60						
		100.00	0.00	100.00	89.92	10.38	100.00	91.09	8.91	100.00	76.19	23.81	100.00						
		0.16	0.00	0.16	5.92	7.63	13.55	2.87	3.08	5.95	0.50	1.71	4.46						
	3.0445224377	4	0	4	238	20	258	94	3	97	19	0	19						
		0.11	0.00	0.11	6.80	0.67	7.37	2.40	0.09	2.49	0.54	0.00	0.54						
		100.00	0.00	100.00	92.25	7.75	100.00	96.55	3.45	100.00	100.00	0.00	100.00						
		0.16	0.00	0.16	6.42	6.65	13.07	2.92	1.03	3.95	0.59	0.00	0.59						
	3.0910424534	12	2	14	188	20	208	69	7	76	21	1	22						
		0.34	0.08	0.40	5.37	0.67	6.04	1.69	0.20	1.89	0.50	0.03	0.53						
		85.71	14.29	100.00	90.38	9.62	100.00	90.55	3.45	94.00	95.45	4.55	100.00						
		0.37	0.08	0.43	5.86	6.65	12.51	1.94	2.40	4.34	0.65	0.34	1.00						
	3.1354942159	17	5	22	208	16	224	77	4	81	16	1	17						
		0.49	0.14	0.63	5.94	0.46	6.40	2.20	0.11	2.31	0.46	0.03	0.49						
		77.27	22.73	100.00	92.85	7.14	100.00	95.09	4.94	100.00	94.12	5.88	100.00						
		0.53	1.71	2.24	6.48	5.48	11.96	2.40	1.37	3.77	0.50	0.34	0.84						
	3.1780503903	47	7	54	188	13	201	42	7	49	16	3	19						
		1.34	0.20	1.54	5.37	0.37	5.74	1.20	0.20	1.40	0.46	0.09	0.54						
		87.04	12.96	100.00	93.53	6.47	100.00	85.71	14.29	100.00	84.21	15.79	100.00						
		1.47	2.40	3.87	5.86	4.45	10.31	1.31	2.40	3.71	0.50	1.03	1.53						
	3.2188758249	82	6	88	205	15	220	47	3	50	20	2	22						
		2.34	0.17	2.51	5.95	0.43	6.38	1.34	0.09	1.43	0.57	0.08	0.65						
		93.18	8.82	100.00	93.18	6.82	100.00	94.00	6.00	100.00	90.91	9.09	100.00						
		2.59	2.05	4.64	6.39	5.14	11.53	1.47	1.03	2.50	0.62	0.65	1.27						
	3.258096538	142	8	150	120	11	131	31	1	32	14	1	15						
		4.09	0.23	4.29	4.59	0.37	4.96	1.28	0.14	1.40	0.40	0.03	0.43						
		64.97	5.32	70.29	93.18	7.34	100.00	89.60	10.20	100.00	93.33	6.67	100.00						
		4.43	2.74	7.17	5.11	4.45	9.56	1.37	1.71	3.08	0.44	0.34	0.78						
	3.295836866	187	21	208	120	11	131	31	1	32	12	1	13						
		4.77	0.90	5.37	3.80	0.31	4.11	0.71	0.03	0.74	0.34	0.03	0.37						
		88.33	11.17	100.00	91.97	8.03	100.00	90.98	3.13	94.10	93.33	6.67	100.00						
		5.21	7.19	12.40	3.93	3.77	7.70	0.97	0.34	1.31	0.44	0.34	0.78						
	3.322045102	175	22	197	123	9	132	25	2	27	15	1	16						
		5.00	0.63	5.63	3.51	0.28	3.77	0.71	0.06	0.77	0.43	0.03	0.46						
		88.33	11.17	100.00	93.18	6.82	100.00	92.59	7.41	100.00	93.75	6.25	100.00						
		5.48	7.53	13.01	3.43	4.79	8.22	0.78	0.88	1.66	0.47	0.34	1.01						
	3.36725583	100	55	155	110	14	124	23	0	23	19	0	19						
		0.65	0.36	1.01	0.72	0.11	0.83	0.15	0.00	0.15	0.12	0.00	0.12						
		100.00	36.13	136.13	88.42	10.58	99.00	100.00	0.00	100.00	100.00	0.00	100.00						
		0.65	0.36	1.01	0.61	0.07	0.68	1.00	0.34	1.34	61.88	8.34	100.00						
	3.36725583	100	55	155	93	11	104	32	1	33	3208	262	3500						
		0.65	0.36	1.01	0.61	0.07	0.68	0.19	0.03	0.22	61.88	8.34	100.00						
		100.00	36.13	136.13	0.61	0.07	0.68	1.00	0.34	1.34	61.88	8.34	100.00						

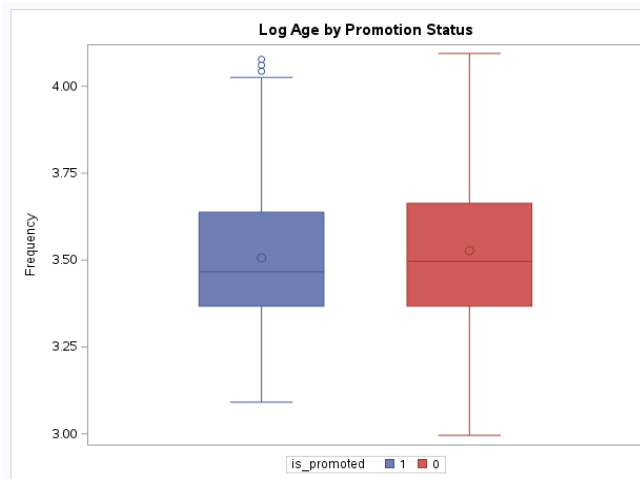


Figure 17. The frequency table and box plot of the log age by promotion status.

In Figure 18, the range and mean of log service length in promoted employees is lower than unpromoted employees.

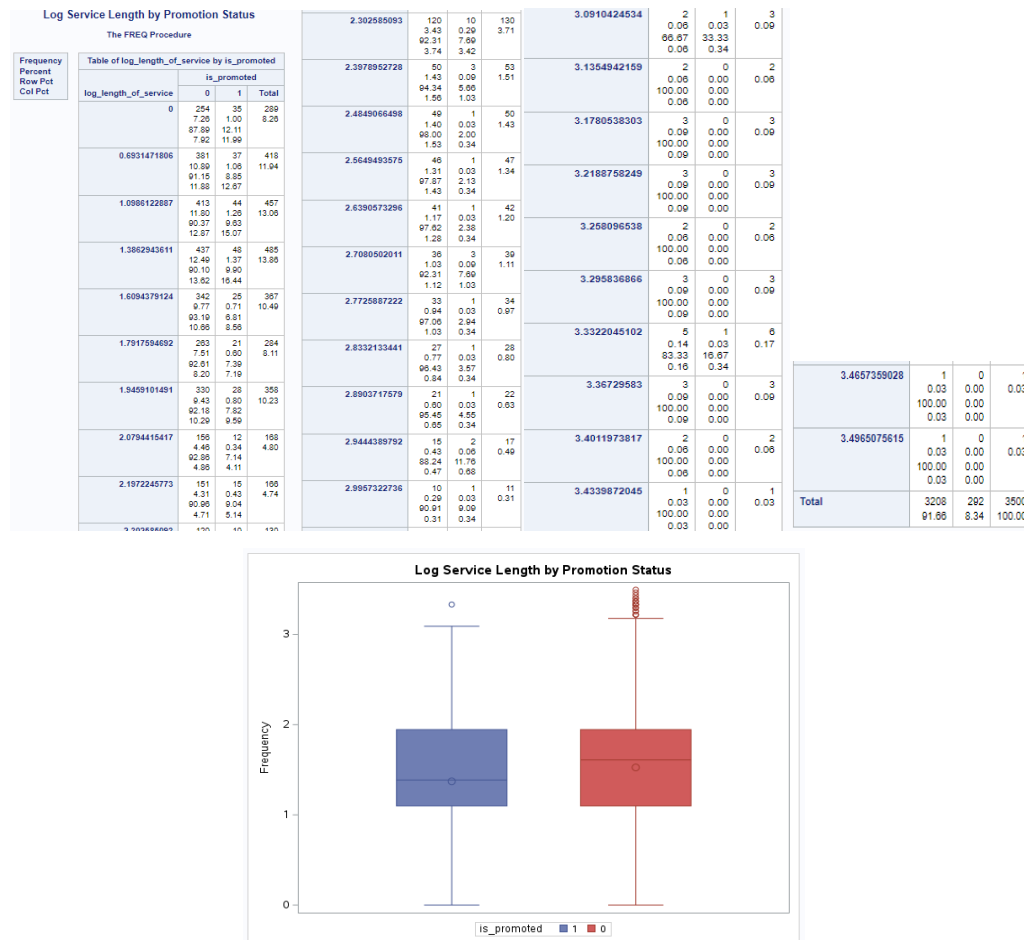


Figure 18. The frequency table and box plot of the log service length by promotion status.

In Figure 19, the unawarded employees has 89.04% promotion rate while awarded employees has 10.95%.

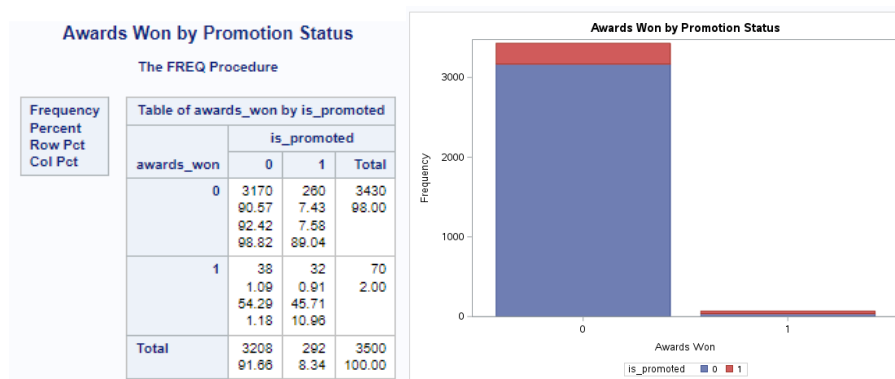


Figure 19. The frequency table and bar plot of the awards won by promotion status.

In Figure 20, the mean and interquartile range of log average training score in promoted employees is higher than the unpromoted employees.

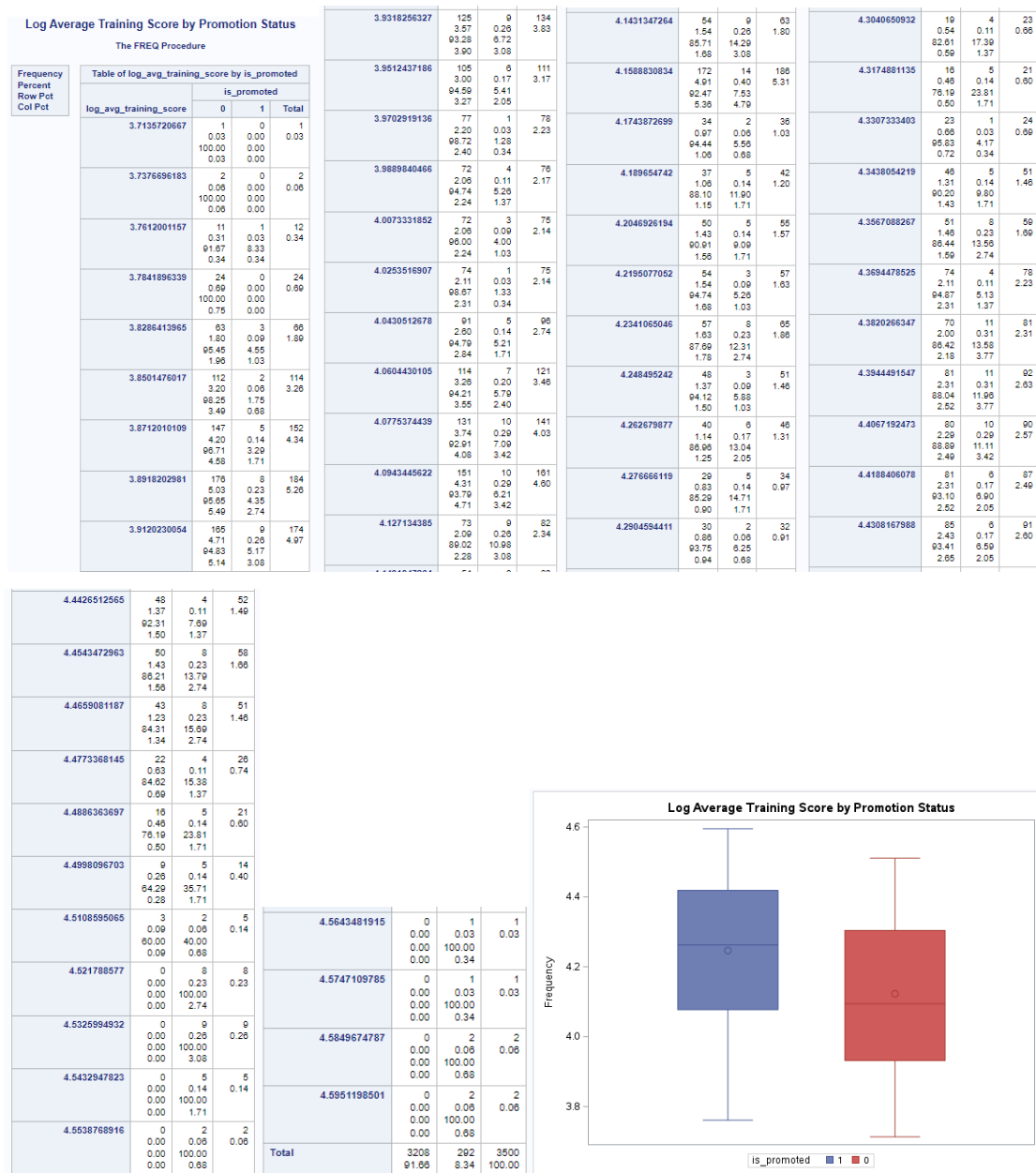


Figure 20. The frequency table and box plot of the log average training score by promotion status.

Figure 21 shows the frequency table and bar plot of awards won by education. The awards won has only 2% while no awards won has 98%. For the employee with awards won, there is only one of them from below secondary education level, followed by 12 from master's and above and 57 from bachelor's degree.

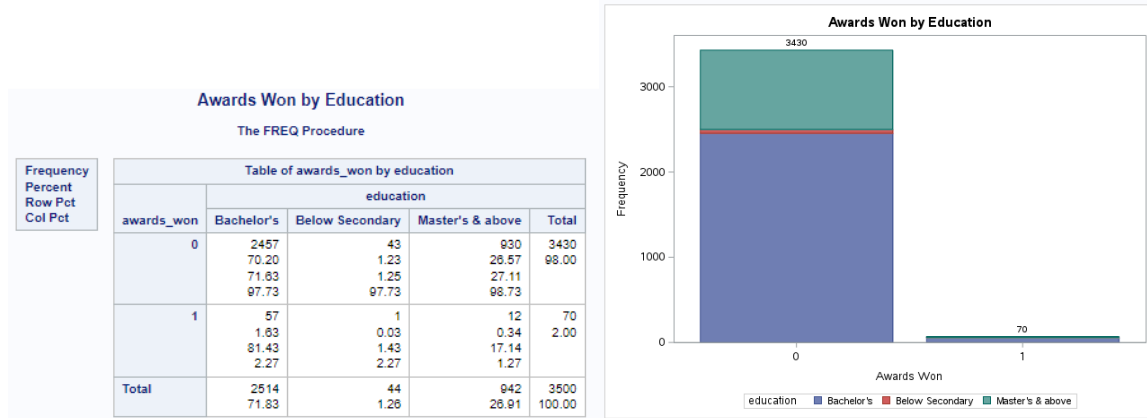


Figure 21. The frequency table and bar plot of the awards won by education.

Figure 22 shows the box plot of the log average training score by education. The result showed that the highest mean under “master’s and above” followed by “bachelor’s degree” and “below secondary”. Upper range of the average training score by below secondary is the lowest.

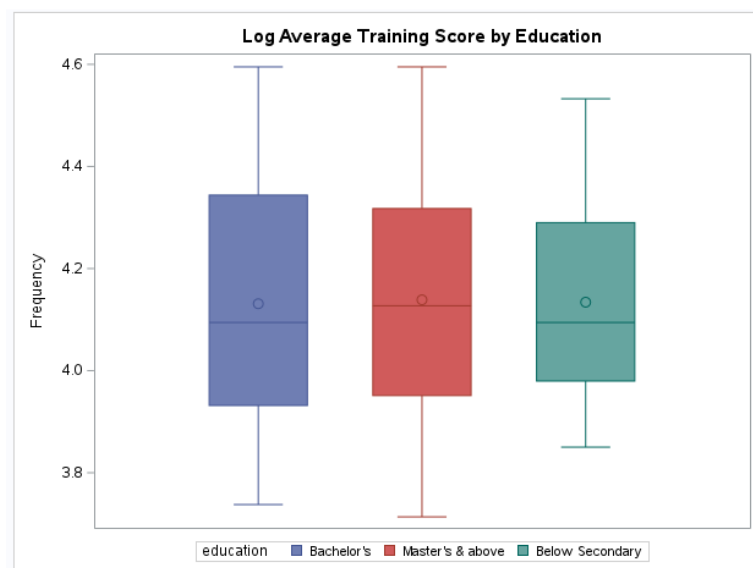


Figure 22. The box plot of the log average training score by education.

Figure 23 shows the scatterplot of “age” and “length of service” by “previous year rating”. The results show a linear relationship of “age” and “length of service” by “previous year rating”.

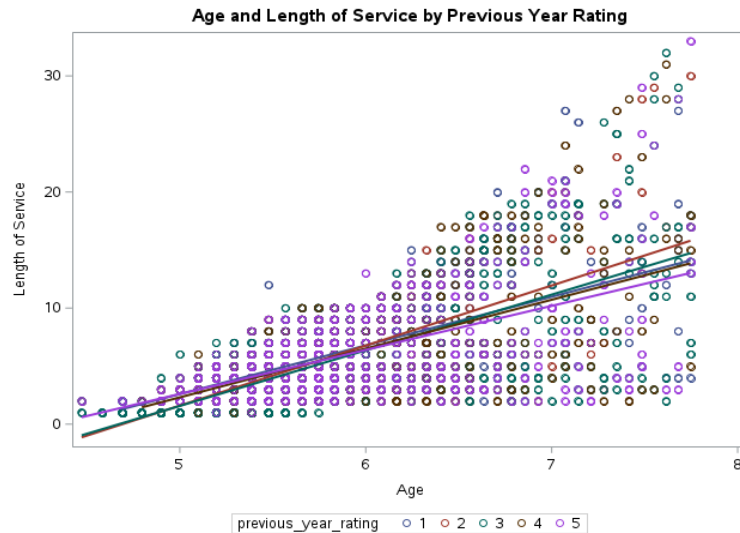


Figure 23. The scatter plot of age and length of service by previous year rating.

Figure 24 shows the box plot of previous year rating and service length of promoted status. The distribution of previous year rating and length of service by promoted employees are spread.

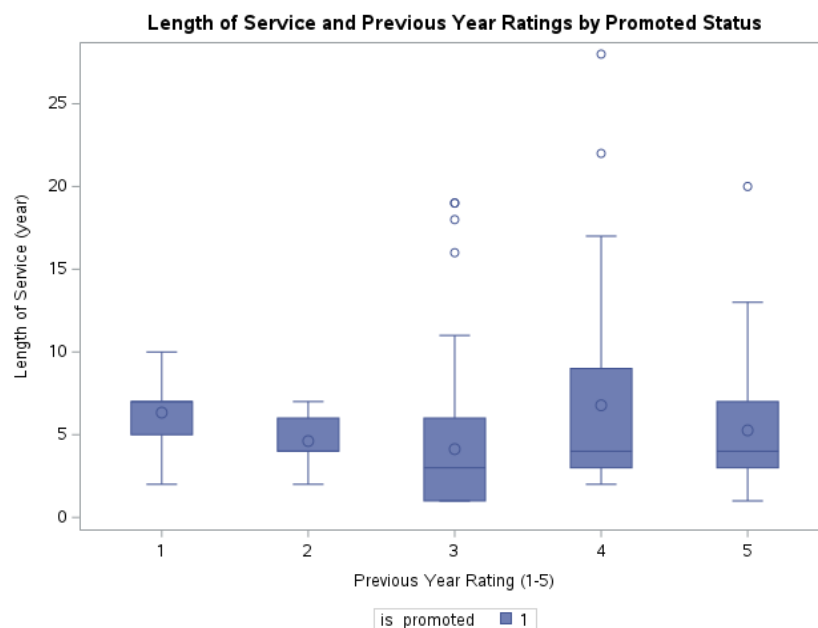


Figure 24. The box plot of previous year rating and service length by promoted status.

3.4 Feature Engineering

3.4.1 Feature Transformation

Data transformation involves the method to normalise and aggregate the data with the purpose to enhance data mining accuracy and efficiency. Data encoding is applied to the data to transform the nominal data into numerical data or binomial data. Label encoding is applied on education and gender as they have inherent order or hierarchy. One-hot encoding is applied on department, region and recruitment channels as there are multiple values with no inherent order or hierarchy. Variables such as employee ID and promotion status are not encoded as they are identifiers or target variables. The department, region, education and recruitment channel and gender are transformed by applying one-hot or label encoding techniques (Figure 25).

Obs	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won	avg_training_score	is_promoted	log_age	log_length_of_service	log_avg_training_score	education_encoding	gender_encoding
1	Operations	region_13	Bachelor's	m	other	1	35	4	6	0	57	1	3.55535	1.79116	4.04305	1	0
2	Technology	region_7	Master's & above	f	other	1	39	4	4	0	78	1	3.66356	1.38629	4.35671	2	1
3	Human Resources	region_28	Bachelor's	m	other	1	28	2	2	0	52	1	3.33220	0.69315	3.95124	1	0
4	Operations	region_15	Master's & above	m	sourcing	2	44	5	11	0	92	1	3.76419	2.39790	4.52179	2	0
5	Human Resources	region_22	Bachelor's	m	sourcing	1	28	3	1	0	51	1	3.33220	0.00000	3.99183	1	0

department_analytics	department_finance	department_hr	department_legal	department_operations	department_procurement	department_rnd	department_sales	department_tech	recruitment_other	recruitment_referred	recruitment_sourcing	region_1	region_2	region_3	region_4	region_5	region_6	region_7	region_8	region_9	region_10
0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

region_11	region_12	region_13	region_14	region_15	region_16	region_17	region_18	region_19	region_20	region_21	region_22	region_23	region_24	region_25	region_26	region_27	region_28	region_29	region_30	region_31	region_32	region_33	region_34
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Figure 25. The first 5 observations of the dataset after data encoding.

3.4.2 Feature Creation

Feature creation is applied to numerical attributes to transform and group the continuous values into categories. This process is known as discretization or binning technique. Binning is applied to increase the robustness of the model and avoid overfitting (Rençberoğlu, 2019). Based on the dataset, the age, length of service and average training score are discretized into specific bins. Binning is applied on original data instead of logarithm data. This is due to the interpretability issues as binning on original data is easier to understand and interpret. It is more straightforward for group comparisons with reliability.

The variable "age" has been categorised into four bins, denoted as "BIN_age" (Figure 26). This binning approach allows for a simplified representation of the age variable, providing insights into the distribution of employees across different age groups. It reveals that the largest proportion of employees falls within the 30-39 age range, while the younger and older age groups have relatively fewer individuals.

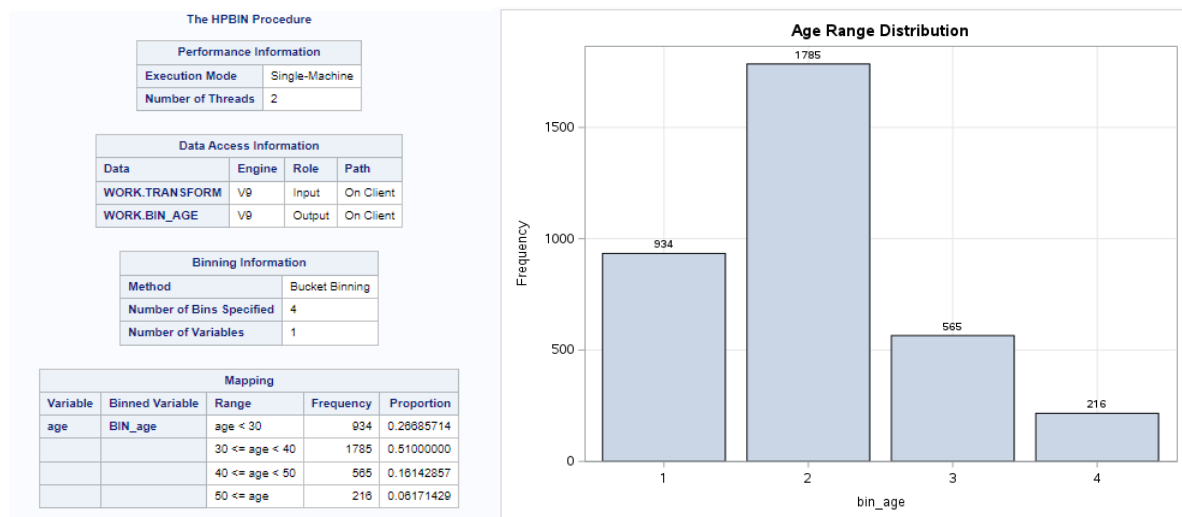


Figure 26. The binning and bar plot of the age range distribution.

The variable "length_of_service" was divided into four bins resulting in the creation of the binned variable "BIN_length_of_service" (Figure 27). It highlights the distribution of employees across different ranges of service length, with the majority having shorter service lengths and a decreasing proportion as the service length increases.

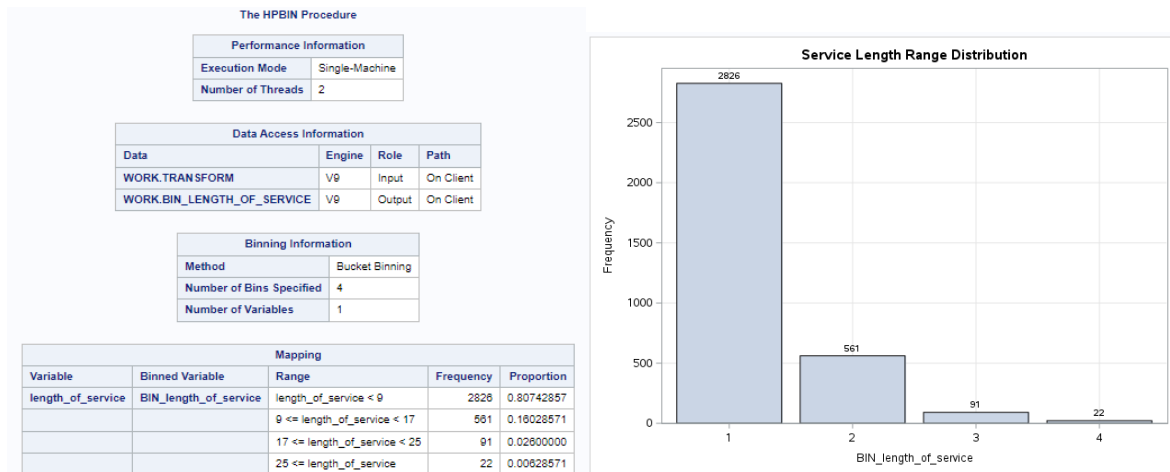


Figure 27. The binning and bar plot of service length range distribution.

The variable "avg_training_score" has been divided into four bins, labelled as "BIN_avg_training_score" (Figure 28). It reveals that the largest proportion of employees has scores below 55.5 with 1203 employees. Relatively fewer employees have scores between 70 and 84.5, and the smallest group has scores 84.5 and above.

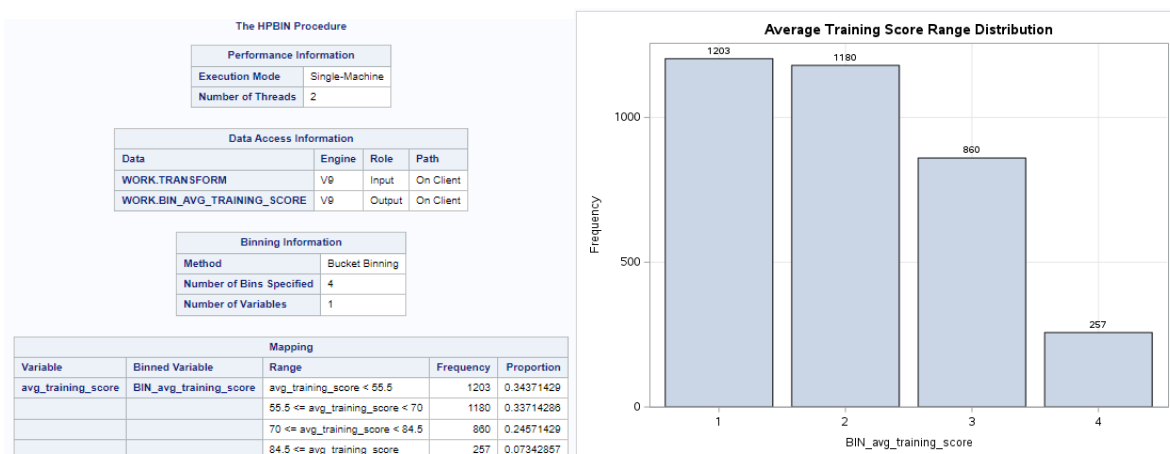


Figure 28. The binning and bar plot of average training score range distribution.