Assignment 2

Group 97

Semester 2, 2024

PAPER NAME: Data Analysis

PAPER CODE: COMP517

Student ID	Student Names
14885318	Lucas Hashemi
23218672	Trinity Thompson

DUE DATE: Midnight Monday 21st Oct 2024

TOTAL MARKS: 100

INSTRUCTIONS:

1. **The following actions** may be deemed to constitute a breach of the General Academic **Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
3. **Attach your code for all the datasets in the appendix section**.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Background

As a leading education provider based in New Zealand focused on providing tertiary students with high-quality learning resources - KiwiLearn is devoted to the continuous development of its employees and subsequently the improvement of its services within the educational sector.

Expansion into online services in the form of a flexible online platform has allowed KiwiLearn to meet the changing needs of students and institutions nationally. KiwiLearn's success is driven by the collaborative efforts of four key departments – Sales, Marketing, IT, and HR. In addition to this, employees are given opportunities to upskill through dedicated training hours to effectively stay on top of the latest trends in education and technology.

Now, KiwiLearn is looking to examine their employee performance data to gain insights that will help them make informed, data-driven decisions moving forward, further improving their success.

1.2 Purpose of the Analysis

The purpose of this analysis is to examine employee data provided by KiwiLearn to gain valuable insights surrounding their employees' performance that will benefit future decision-making. Minor exploratory data analysis of the dataset will be undertaken to record initial observations and ensure data cleanliness before progressing to a hypothesis-driven analysis utilizing methods such as one-way ANOVA testing to understand differences across different department sectors within KiwiLearn. Finally, regression analysis will be conducted to discern any other key factors or relationships that impact employees' performance.

1.3 Dataset Overview

KiwiLearn's "Employee_Performance" dataset will be the focus of the analysis within this report. The data is comprised of important employee information which will be explored to observe any performance differences across departments, as well as discover underlying relationships within the dataset. As an internally conducted analysis, it is assumed that the data provided by KiwiLearn has truthfully and accurately been collected for effective analysis to take place.

Included within the dataset is information pertaining to each employee, including an 'Employee ID' identifier, designation of 'Gender' and current 'Department' of employment. Furthermore, attributes related to employees' work performance and history is incorporated – such as 'Years of Experience' on a scale of 0-9 and 'Training Hours' the employee has undertaken within the past year. The monthly 'Salary' of each employee is also included, alongside a 'Performance Rating' from 1 (Poor performance) to 5.5 (Exceptional performance).

Of particular interest within this report is the 'Performance Rating' of employees and subsequent related factors of both numerical and categorical nature.

2. Exploring Data and Testing Hypotheses

2.1 Data Preparation and Exploration

2.1.1 First Look at the Data (First Five Entries):

To derive the format and types of information or values contained within the 'Employee_Performance' dataset, an initial glimpse of the data has been provided.

Table 1: Summary of Employee Performance Dataset – First Five Entries

```
-----First Few Rows of Employee Performance Dataset:
   EmployeeID Department  Gender  Experience  TrainingHours  \
0        1001         IT    Male           4              5
1        1002  Marketing  Female           0             50
2        1003      Sales    Male           0              5
3        1004         HR    Male           1              5
4        1005         HR  Female           9              5

   PerformanceRating  Salary
0               1.00   19000
1               5.50    6900
2               1.00    6000
3               1.00    6000
4               1.04   38000
```

Details of the first five employees at KiwiLearn can be seen in Table 1, identified by unique Employee IDs from 1001 to 1005. Six key attributes are visible across the visible entries: 'Department' (IT, Marketing, Sales, HR), 'Gender', years of 'Experience' (ranging from 0 to 9 years) and 'Training Hours' are recorded - with one Marketing employee completing 50 hours. 'Performance Rating' can also be seen, with most employees attaining a value of 1.00, except for one employee with a rating of 5.50. Monthly 'Salary' is also stated, with a range from 6,000 to 38,000 within this sample. It is evident that variations in experience, training, performance, and salaries are present among the employees.

2.1.2 Dataset Overview and Structure

A succinct overview of the data frame visible in Table 2 showcases the shape, types of data and number of entries in the dataset amongst other characteristics. This concise summary provides valuable insight prior to analysis and aids in identifying any relevant data gaps that may require further attention.

Table 2: Basic Information of Dataset

```
----DataFrame Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1468 entries, 0 to 1467
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   EmployeeID         1468 non-null   int64
 1   Department         1468 non-null   object
 2   Gender             1468 non-null   object
 3   Experience         1468 non-null   int64
 4   TrainingHours      1468 non-null   int64
 5   PerformanceRating  1468 non-null   float64
 6   Salary             1468 non-null   int64
dtypes: float64(1), int64(4), object(2)
memory usage: 80.4+ KB
None
```

Comprised of 1,468 entries (rows) indexed from 0 to 1,467 and featuring seven attributes (columns) indexed from 0 to 6, the 'Employee_Performance' dataset possesses a shape of (1468, 7). As seen in Table 2, the dataset includes four integer columns ('EmployeeID', 'Experience', 'TrainingHours', and 'Salary'), one float column

('PerformanceRating'), and two object columns ('Department' and 'Gender'). This indicates a diverse mix of numerical and categorical data types. Notably, the dataset appears complete, with 1,468 non-null entries for all seven columns.

2.1.3 Identifying and Handling Missing Data

Despite the initial overview suggesting no null-values are present within the dataset, it is best to ensure accuracy by specifically checking for missing values. Missing data may obscure analysis results by introducing bias or hiding key relationships due to information loss. To ensure accurate results and inferences can be derived from this analysis, missing variables must be dealt with if present.

Table 3: Summary of Missing Values in the Dataset

```
-----Number of Missing Values:
EmployeeID           0
Department           0
Gender               0
Experience           0
TrainingHours        0
PerformanceRating    0
Salary               0
dtype: int64

-----Percentage of Missing Values:
EmployeeID           0.0
Department           0.0
Gender               0.0
Experience           0.0
TrainingHours        0.0
PerformanceRating    0.0
Salary               0.0
dtype: float64
```

Upon review, there appear to be no missing values in the dataset – evidenced by Table 3. Subsequently, no adjustments are needed as all the information pertaining to each employee is fully available within the dataset.

2.1.4 Identifying and Handling Duplicates

Table 4: Summary of Duplicate Rows in the Dataset

```
-----Number of Duplicate Rows: 0
```

Examination of the dataset indicated that no duplicate entries were present, hence no entries will be adjusted or removed. All data entries of employees are unique with no repeated records, ensuring an accurate analysis can be undertaken without repeated information skewing results.

2.1.5 Identifying and Handling Outliers

Visual Method – Box Plot and Histogram:

Visualisation methods including box plots and histograms assist in understanding the spread of data and identifying potential outliers contained within the dataset. Whilst box plots showcase beneficial metrics such as the interquartile range (IQR) and median, they distinctly outline potential outliers in the form of dots outside the main data range (whiskers). Histograms, on the other hand, allow for a quick glimpse into the distribution of data – particularly in relation to symmetry and skewness.

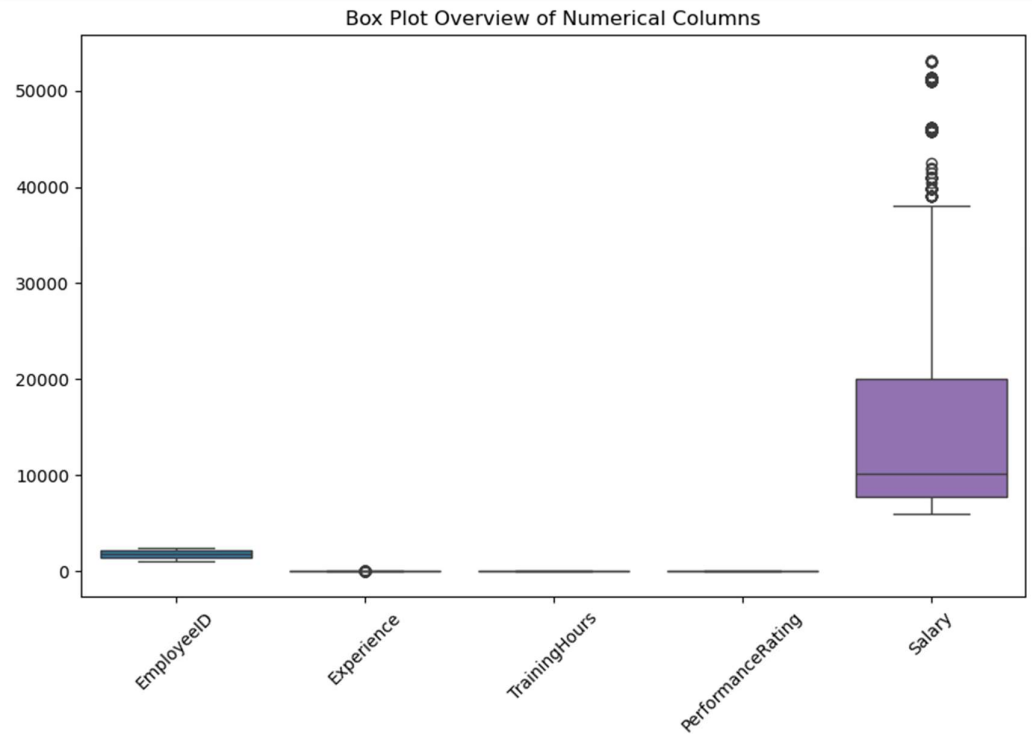

Figure 1: Box Plot overview of numerical columns for outlier detection.

Potential outliers can be discerned upon inspection of Figure 1. Two numerical columns appear to have outliers - 'Experience' and 'Salary'. Examination of the following figures will allow for potential outliers to be observed more closely.

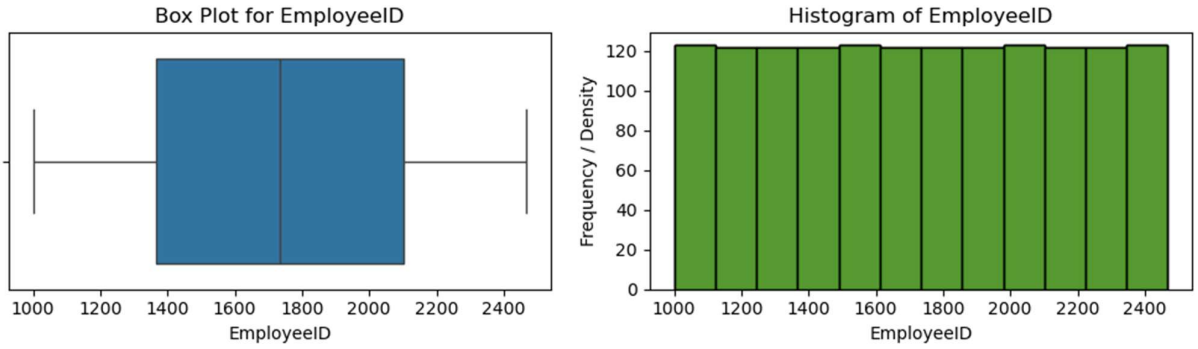

Figure 2: Box Plot and Histogram of 'EmployeeID' for outlier analysis.

'EmployeeID' has been included for the initial analysis despite being of an index nature in order to ensure no abnormal values were included. As there appear to be none visible in Figure 2, 'EmployeeID' will not be considered for any future outlier visualisations.

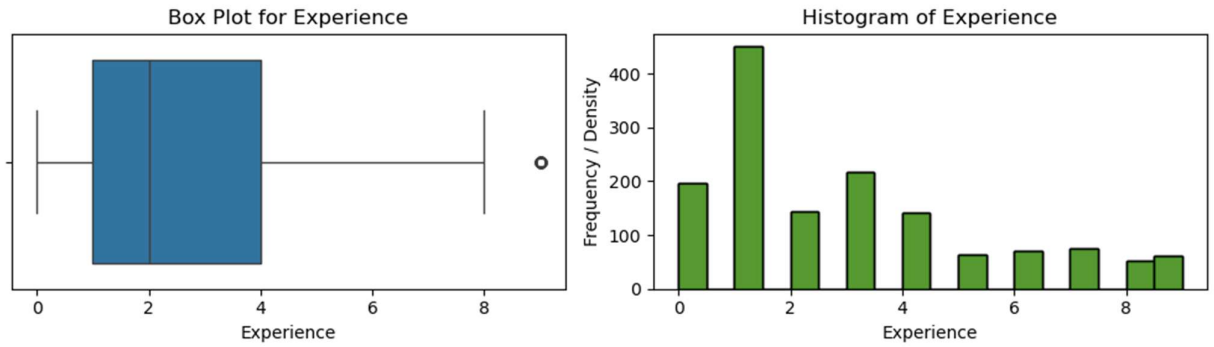

Figure 3: Box Plot and Histogram of 'Experience' for outlier analysis.

A few outliers can be observed within Figure 3. The placement of the outlier suggests that there are only a few senior employees with greater than 8 years' experience, whilst most staff are of junior level. The skewness present in the histogram attests to this.

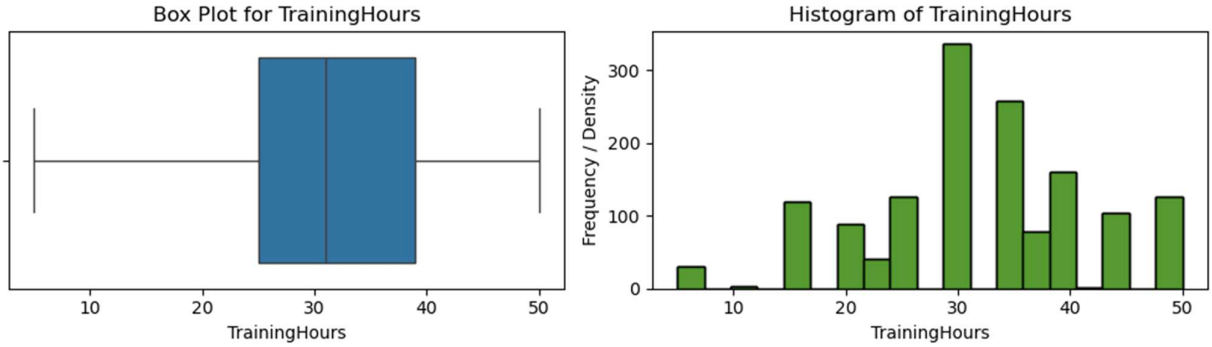

Figure 4: Box Plot and Histogram of 'Training Hours' for outlier analysis.

There are no outliers visible within Figure 4, suggesting that training hours amongst employees are relatively normal in distribution. This is reflected in the histogram, which appears largely symmetrical.

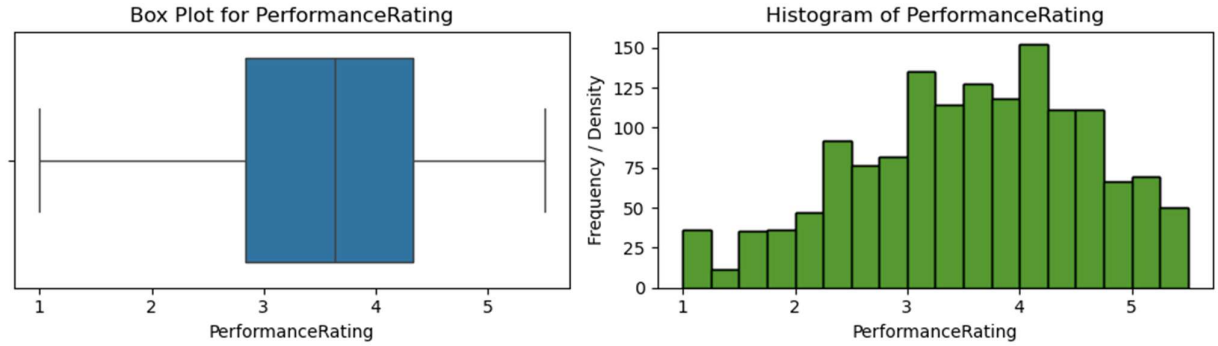

Figure 5: Box Plot and Histogram of 'Performance Rating' for outlier analysis.

According to Figure 5, the distribution of ‘Performance Rating’ among KiwiLearn employees contains no abnormal values.

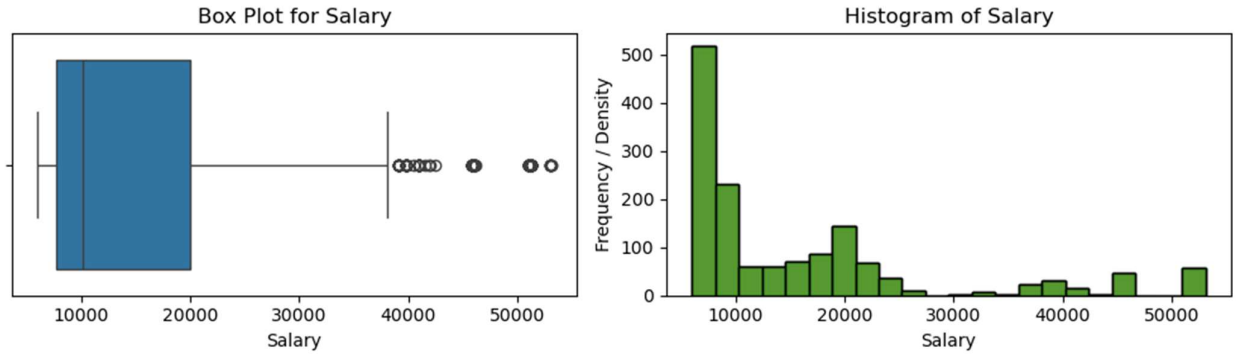

Figure 6: Box Plot and Histogram of 'Salary' for outlier analysis.

Several outliers are visible in Figure 6, suggesting disparity in the 'Salary' of KiwiLearn employees. The large number of outliers present in the box plot, alongside the clear skew in the histogram suggest that further investigation should be undertaken.

Statistical Method – Skew:

Table 5: Skewness for Numerical Columns

```
----Skewness of Numerical Columns:

EmployeeID           0.00
Experience           0.95
TrainingHours       -0.38
PerformanceRating   -0.31
Salary               1.63
dtype: float64
```

Skewness values of numerical columns were calculated to understand the distribution of data – particularly asymmetry – and observe potential columns containing outliers. Skewness also provides insight into an appropriate method for outlier detection. The calculated skewness values for 'Experience', 'Training Hours', 'Salary', and 'Performance Rating' are displayed in Table 5.

'Training Hours' and 'Performance Rating' have skewness values of –0.38 and –0.31 respectively, indicating a slight negative skew. This suggests that both 'Training Hours' and 'Performance' have a few lower values pulling the distribution to the left, but overall, a relatively symmetrical distribution can be observed. Contextualising, it can be determined that most employees have a similar number of training hours and receive comparable performance ratings.

On the other hand, 'Experience' is positively skewed with a value of 0.95 - suggesting many employees have few years of experience, whilst a small number have numerous. Subsequently, the distribution is stretched to the right.

A highly positively skewed distribution can be surveyed in 'Salary'. The skewness value of 1.63 indicates that in general, employees earn lower salaries, however this is disproportionally skewed by the few employees with incredibly high salaries, which pulls the distribution to the right. This may be indicative of the hierarchical pay discrepancies between those in senior and entry-level roles, which reflects the industry standard and could be determined reasonable.

Statistical Method – IQR:

Table 6: Summary of Outliers - IQR

```
Summary of Outliers in the Dataset (IQR Method):
              Column  Number of Outliers  Percentage of Outliers
0         EmployeeID                   0                    0.00
1         Experience                  61                    4.16
2      TrainingHours                   0                    0.00
3  PerformanceRating                   0                    0.00
4             Salary                 152                   10.35
5              Total                 213                    2.90
```

Considering the skewness values calculated in Table 5, the IQR method has been selected as the primary statistical outlier detection method due to its tolerance of skewed data. The summary of outliers identified using this method are outlined in Table 6. 61 outliers were detected in the 'Experience' column, amounting to approximately 4.16% of the total entries within the column. This reflects the notion observed in Figure 3 that most employees have relatively low experience, whilst a few have numerous. 'Salary' contains 152 outliers, constituting 10.35% of the columns' entries. Whilst a noticeable percentage, this aligns with previous observations in Figure 6 that a small number of employees earn significantly larger salaries than the majority – which is a common occurrence in the workplace.

No outliers were detected in 'Training Hours' and 'Performance Rating', suggesting the disparity between employees' training hours and performance ratings are not abnormal.

Statistical Method – Z-Score:

Table 7: Summary of Outliers - Z-score

```
Summary of Outliers in the Dataset (Z-score Method):
              Column  Number of Outliers  Percentage of Outliers
0         EmployeeID                   0                    0.00
1         Experience                   0                    0.00
2      TrainingHours                   0                    0.00
3  PerformanceRating                   0                    0.00
4             Salary                   7                    0.48
5              Total                   7                    0.10
```

For confirmation and exploratory purposes, the Z-Score method was also employed in consideration of the more symmetrical skewness values observed in Table 5. As seen in Table 7, a different result can be observed after employing this method. Here, no outliers are observed in 'Experience', 'Training Hours', or 'Performance Ratings'.

Whilst this confirms that there are indeed no outliers within the more normally distributed 'Training Hours' and 'Performance Salary', using the Z-Score method on skewed data (such as 'Experience' and 'Salary' may result in less outliers being detected. As the Z-Score method relies on the mean and standard deviation, which become skewed by extreme values, not all outliers may be accurately identified. The seven outliers detected in 'Salary' (0.48% of the column) and the zero outliers detected in 'Experience' reflect this. The true number of outliers may lie somewhere between these two methods but indicate there is only a small amount of potentially abnormal data requiring investigation.

The insights derived from these visual and statistical outlier analyses showcase that much of the dataset is consistent, with only 'Salary' and 'Experience' requiring further investigation. These attributes will be investigated visually before determining an appropriate handling method.

Outlier Visualisation – Scatter Plot (IQR)

To visualise outliers against their non-outlier counterparts, scatterplots have been produced using the IQR method. Outliers appear in both 'Experience' and 'Salary', clustering at the higher end of the value range.

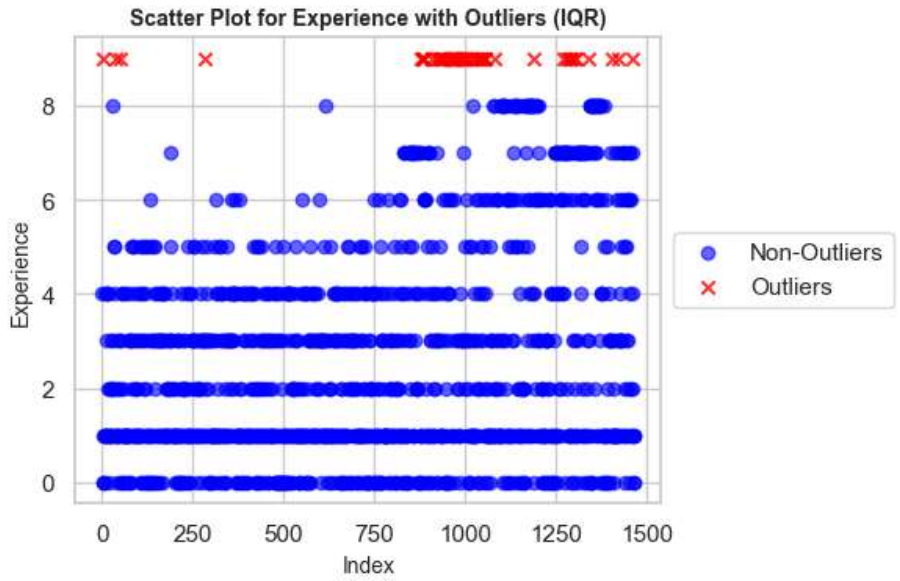

Figure 7: Scatter plot of Outliers vs Non-Outliers for 'Experience' (IQR)

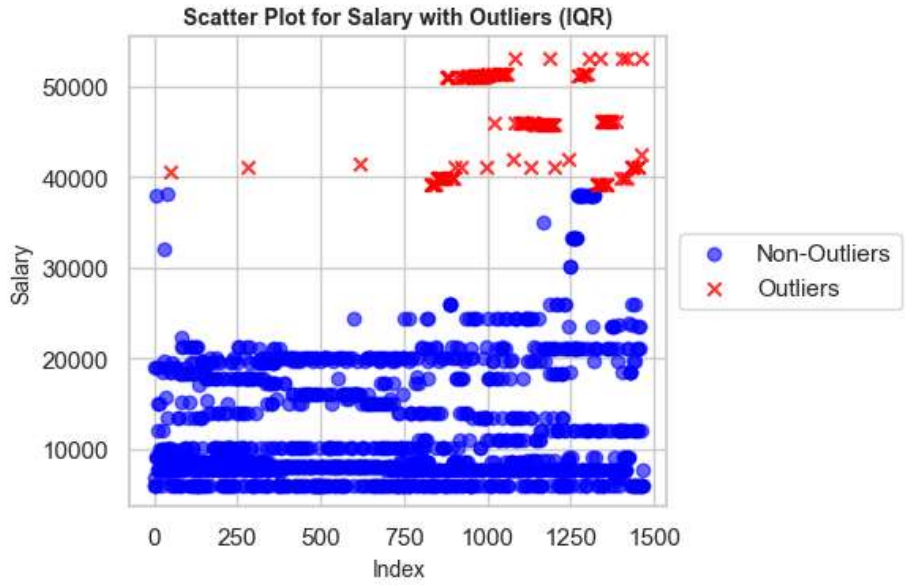

Figure 8: Scatter plot of Outliers vs Non-Outliers for 'Salary' (IQR)

Handling Outliers:

Collating observations from distribution visuals in Figures 2-8 alongside statistical data depicted in Tables 6-7, a conclusion can be drawn regarding the validity of potential outliers present in 'Experience' and 'Salary'. Upon review, alongside further investigation into the specific rows containing outliers – it was determined that potential outliers should not be removed, as they may be representative of key correlations between 'Salary' and 'Experience'.

In the case of 'Experience', the outliers initially detected are the employees who possess 9 years of experience. The small number of employees with larger numbers of experience compared to the copious with small numbers is indicative of trends within companies – as there are less employees the further up the hierarchical chain. Thus, these data points were deemed valid.

In the case of 'Salary', although higher salaries were flagged as outliers, these points reflect the distribution of pay across a multitude of employment levels, and there is not enough evidence to remove this data without introducing bias. Furthermore, every potential outlier outlined were employees with seven years of experience or greater. It is reasonable that senior employees command higher compensation, hence they should remain in the dataset.

2.1.6 Summary Statistics

To understand the central tendency, spread, and distribution of numerical data within the dataset, and consequently gain insight into the dataset's characteristics, patterns or anomalies - summary statistics were calculated. Metrics including the mean, standard deviation, minimum, quartiles, and maximum values for each attribute are determined.

Table 8: Summary Statistics of the Dataset

```
----Summary Statistics of the Dataset:

        EmployeeID   Experience  TrainingHours  PerformanceRating  \
count  1468.000000  1468.000000    1468.000000        1468.000000
mean   1734.500000     2.838556      32.144414           3.561512
std     423.919411     2.527657      10.106029           1.044987
min    1001.000000     0.000000       5.000000           1.000000
25%    1367.750000     1.000000      25.000000           2.840000
50%    1734.500000     2.000000      31.000000           3.630000
75%    2101.250000     4.000000      39.000000           4.330000
max    2468.000000     9.000000      50.000000           5.500000

             Salary
count   1468.000000
mean   16107.623297
std    12158.438481
min     6000.000000
25%     7700.000000
50%    10100.000000
75%    20000.000000
max    53100.000000
```

Summary statistics of the 'Employee_Performance' dataset's numerical columns are displayed in Table 9. On average, employees have approximately 2.8 years of experience, with a standard deviation of 2.53 years (2dp). This showcases employees' large range of experience levels from entry-level to senior roles. A minimum of 0 years and maximum of 9 years can be seen. Training hours average at 32.14 (2dp), spanning from 5 to 50 hours – a relatively distinct spread. The standard deviation of 10.11 hours (2dp) suggests some variability in training received. Performance ratings are comparable across employees with a mean of 3.56 (2dp) and standard deviation of 1.04 (2dp) - the lowest amongst the variables displayed. A minimum of 0 and a maximum of 5.5 is observed. Salaries vary widely, with an average of $16,107 a month. 'Salary' possesses the largest standard deviation of 12,158.44 (2dp) demonstrating significant income differences between employees. The lowest salary recorded is $6,000, whereas the highest is $53,100 - suggesting the presence of highly compensated workers within the organisation. 25% of

workers earn less than $7700 whereas the top 25% of workers earn greater than $20000 – which may be reflective of different hierarchical roles and salary brackets.

As the potential outliers were deemed valid, or representative of key differences amongst employees, the summary statistics have not been altered by pre-processing changes as the data was retained. Observing the results in Table 9, a quick snapshot of the variation in employee experience, salary, and performance within the company has been obtained.

2.1.7 Data Distribution and Visualization

Visualize Distribution: Numerical Variables – Box Plots and Histogram with KDE:

Visualisation methods can be used alongside summary statistics to gain key insight into data distribution. Numerical variables displayed in Figures 9-12 can be compared with previous findings to further understand the dataset.

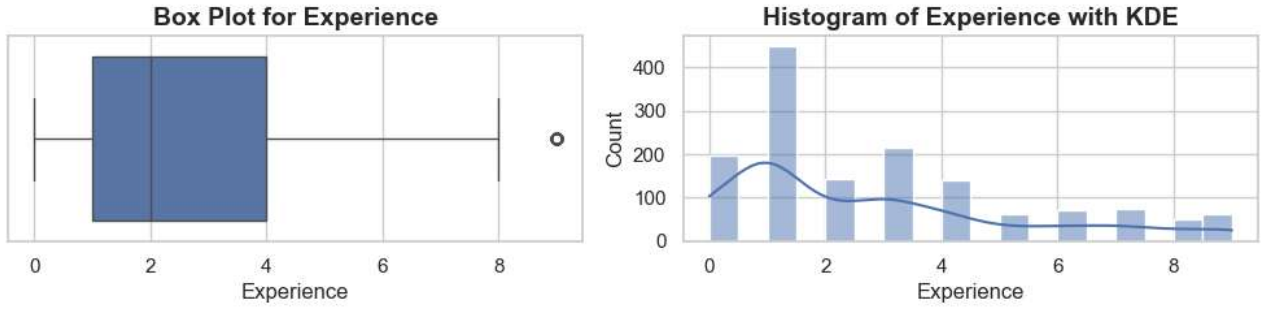

Figure 9: Box Plot and Histogram with KDE for 'Experience'

'Experience' exhibits a positive skew, with a longer right tail as seen in Figure 9 – suggesting many employees at KiwiLearn have relatively low experience (approximately 0-2 years), indicating a higher number of entry-level staff. The median lies at two years, further attesting to this. Employees with higher levels of experience are fewer, dropping off significantly at five years of experience. This reflects the skewness value of 0.95 (2dp). calculated previously.

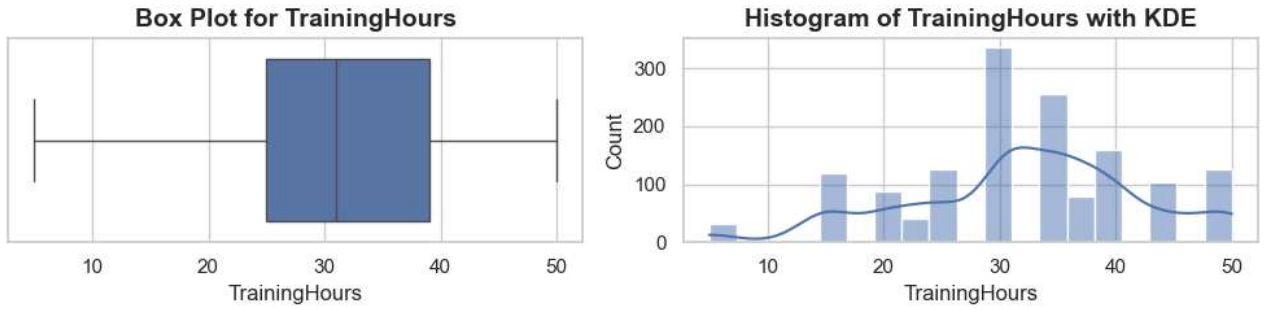

Figure 10: Box Plot and Histogram with KDE for 'Training Hours'

The distribution of 'Training Hours' is slightly positively skewed, with a longer tail visible on the left. Figure 10 showcases that 50% of employees have logged between 25 and 39 training hours over the past year, with a median of approximately 32. This suggests KiwiLearn provides equal training opportunities for employees at all levels and values continuous improvement.

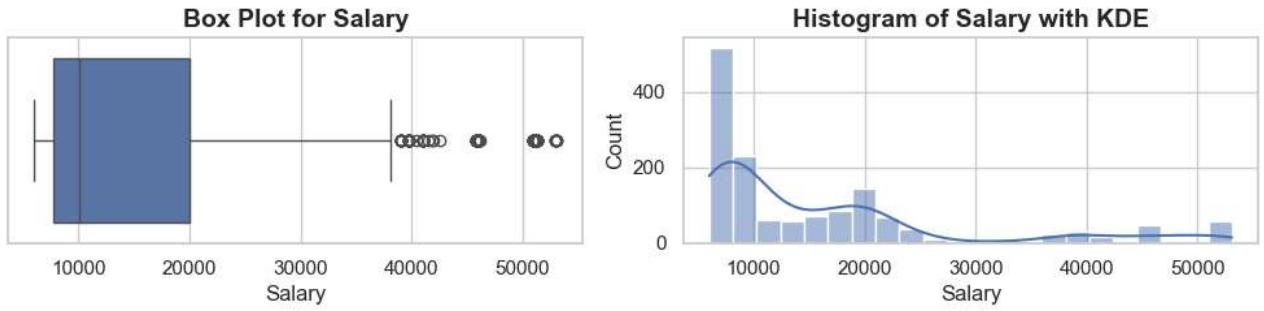

Figure 11: Box Plot and Histogram with KDE for 'Salary'

A highly positively skewed distribution can be observed in Figure 11 of 'Salary' aligning with the calculated skewness value of 1.63 (2dp). Varying from $6,000 to over $50,000 around a median of approximately $10,000. A small number of high earners can be observed on the right, pulling the salary higher. This is likely to occur with senior roles present in the dataset. As most employees are entry or junior level, a lower salary range is more common with 75% of employees earning less than $20000 and 50% less than $10000.

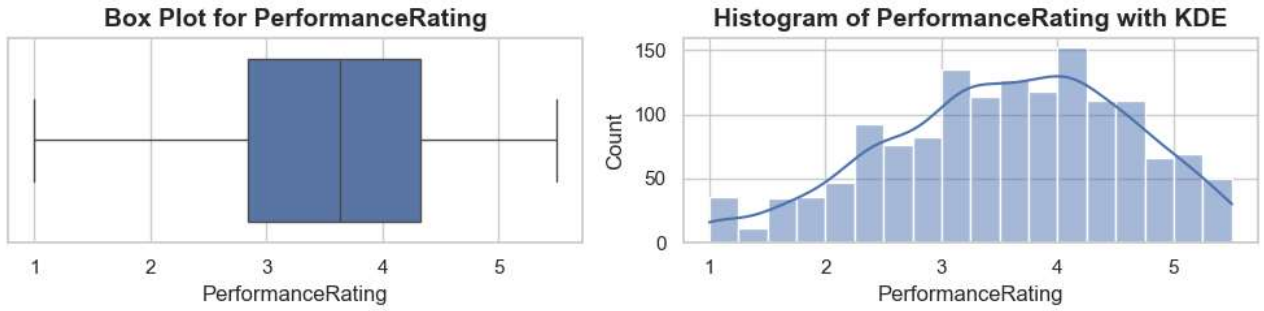

Figure 12: Box Plot and Histogram with KDE for 'Performance Rating'

'Performance Rating' displayed in Figure 12 demonstrates a more normal distribution, but still possesses a slight negative skew. Ratings range from 1 to 5.5, with a median at approximately 3.5. 75% of employees receive ratings higher than approximately 2.8, suggesting more employees obtain positive ratings than negative – a net positive for KiwiLearn.

Visualize Distribution: Categorical Variables – Pie Chart and Bar Plot:

Prior to analysis, the distribution of categorical variables must also be acknowledged. Considering this reports investigation of employee performance across departments, it is important to note size disparities if present. Utilizing visualisation methods such as pie charts and bar plots, the distribution of 'Department' and 'Gender' can be observed. Relevant insights into the representation of staff across various sectors and the employment practices of KiwiLearn may be revealed through such analysis.

Figure 13: Distribution of 'Department' at KiwiLearn using Pie Chart and Bar Plot as visualisation methods.

Analysing Figure 13, the proportion of departments at KiwiLearn appears uneven, with a difference of 44.7% between the largest and smallest departments. 'IT' holds the most employees at 49% - 18.7% greater than the next largest, 'Sales'. Contextualised, these proportions seem reasonable. With services hosted on an online platform, KiwiLearn would require many employees to oversee its technological components, and any successful company requires a competent and diverse sales staff to drive profit. Marketing is next largest – comprising 16.3%. The smallest department, 'HR', constitutes only 4.3% of employees in the dataset. Whilst the spread of employees across departments may be attributed to the focus of KiwiLearn's services, it is important to recognise these differences in employee numbers when analysing performance across departments.

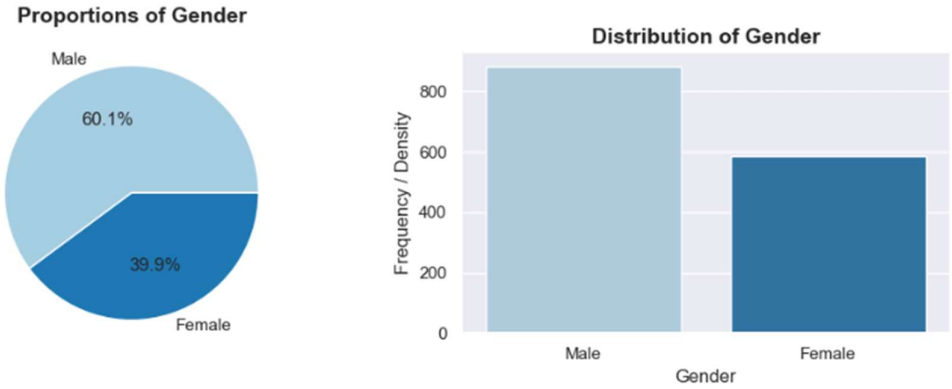

Figure 14: Distribution of 'Gender' at KiwiLearn using Pie Chart and Bar Plot as visualisation methods.

Compared to 'Department', the distribution of 'Gender' within the dataset displayed in Figure 14 is relatively balanced. A greater number of males are employed, comprising 60.1% compared to females at 39.9% - a difference of 20.2%. This 60-40 split is somewhat equal, suggesting relatively equal gender employment opportunities at KiwiLearn.

2.1.8 Multivariate Analysis: Visualising Experience vs. Performance Ratings Across Departments

Multivariate analysis will be performed to determine any variations in employee performance based on experience across different departments. Utilising separate scatter plots and box plots, insights into the relationship between performance and experience can be assessed, alongside any potential discrepancy of evaluations across sectors of the workplace.

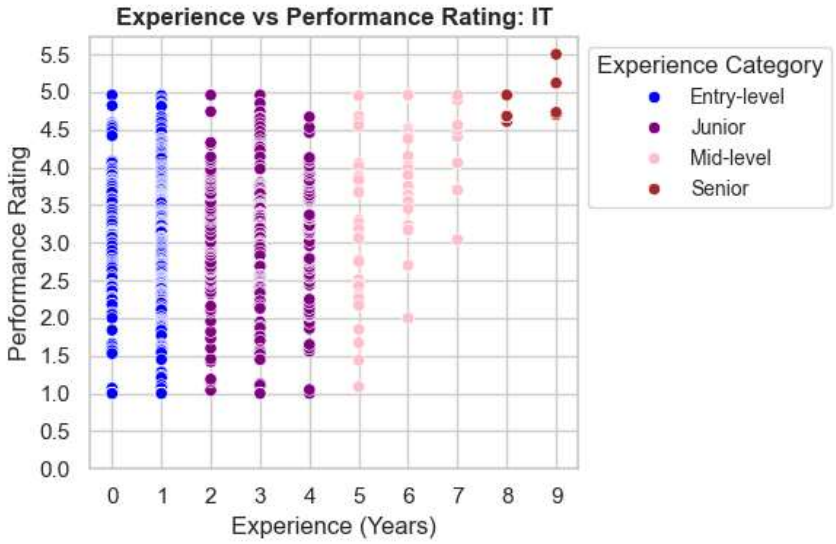

Figure 15: Scatter plot - Experience vs Performance Rating in the IT Department

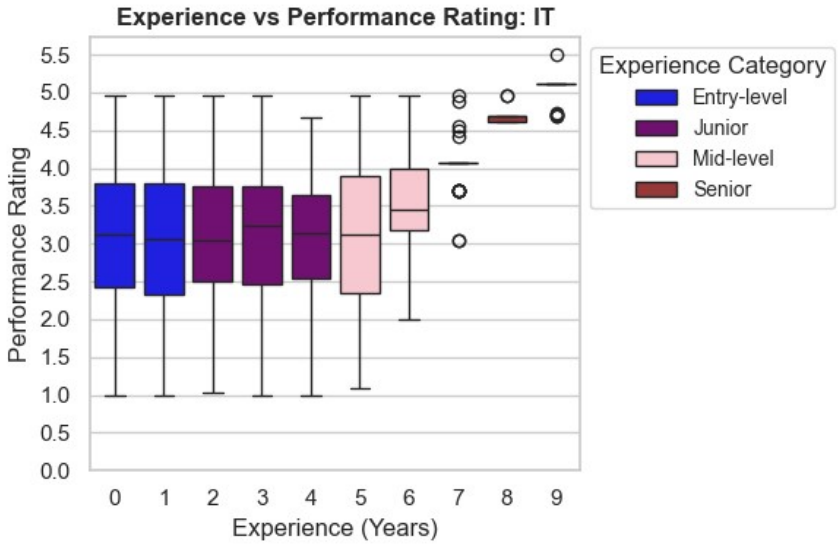

Figure 16: Box plot - Experience vs Performance Rating in the IT Department

Observing Figures 15 and 16 of the IT department, large variability in performance ratings can be seen in employees with 0 to 5 years of experience. This is particularly prevalent among entry-level and junior employees which possess the greatest range amongst experience categories with a minimum of 1 and maximum of 5.5. Mid-level employees with 5 years of experience show similar variability. However, employees with 6 or more years of experience begin to exhibit consistently higher performance ratings, as shown by the increase in median values. Senior employees with 8 years of experience demonstrate a notable rise in median performance ratings, exceeding 4.0, with a further increase to above 5.0 at 9 years, consistently achieving higher and more stable performance ratings than that of other levels. As the only experience category exhibiting a result of 5.5, and with minimum values never falling below 3, this suggests that increased experience correlates with higher, stable performance levels within the IT department.

Figure 17: Scatter plot - Experience vs Performance Rating in the Marketing Department

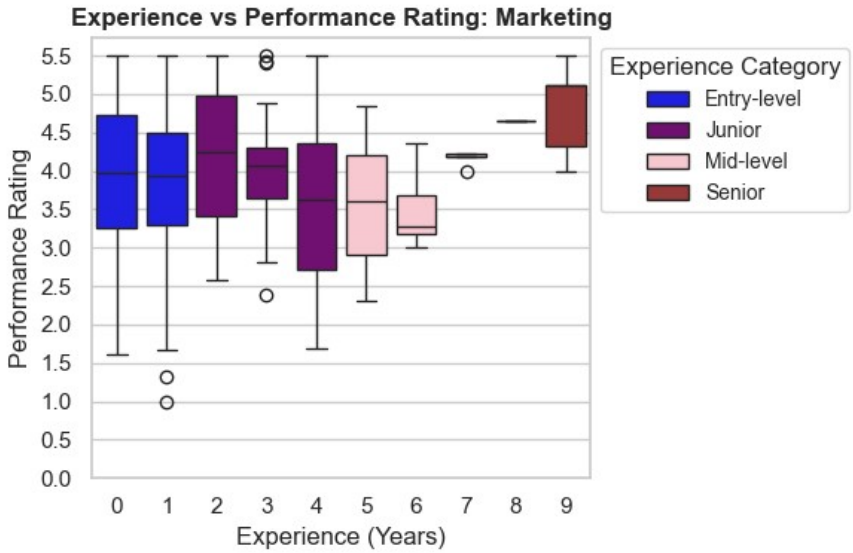

Figure 18: Box plot - Experience vs Performance Rating in the Marketing Department

A somewhat similar trend can be observed in the Marketing department, as shown in Figures 17 and 18. Entry-level employees exhibit a broad range of performance ratings, from 1 to 5.5 - the highest variability amongst the experience levels. Junior employees follow a comparable pattern, as indicated by the median values. Notably, employees at multiple experience levels have received ratings of 5.5. Employees at the mid-level demonstrate a smaller range, with performance ratings between approximately 2.3 and 4.9. Within the Marketing department, it is interesting to note that employees with 6 years of experience possessed the lowest median rating at just above 3. Employees with 6-7 years of experience appear to perform well more consistently, with a tighter range of ratings all falling above a minimum of 3. Finally, Senior employees (8+ years of experience) maintain constantly high-performance rating between 4 and 5.5 - a range of 1.5. The trends depicted suggest that with increased experience, employees in the Marketing department achieve higher performance ratings.

Figure 19: Scatter plot - Experience vs Performance Rating in the Sales Department

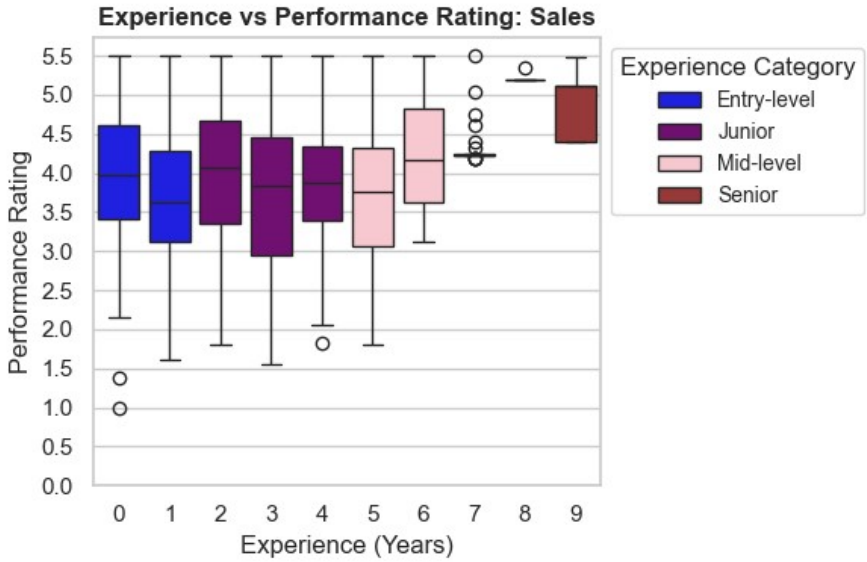

Figure 20: Box plot - Experience vs Performance Rating in the Sales Department

Comparable patterns can also be observed in the Sales department displayed in Figures 19 and 20. Highest variability within employees in the Sales department is also amongst entry-level and junior employees. Variability begins to narrow in employees with 4-5 years of experience. Those with 6-7 years' experience cluster between 3 to 5.5 then between 4 and 5.5 with a median of approximately 4.2. Senior employees (8+ years) remain the strongest performing employees with a range from approximately 4.4 to 5.5. Once again, it is notable that employees at all experience levels have received an exceptional rating (5.5), however it is only mid-level employees and below who obtain ratings lower on the scale – solidifying the notion that there is relative variability amongst employees until senior level where higher ratings are consistently achieved.

Figure 21: Scatter plot - Experience vs Performance Rating in the HR Department

Figure 22: Box plot - Experience vs Performance Rating in the HR Department

Unlike the previous departments, HR (depicted in Figures 21 and 22) does not exhibit a clear trend of increasing performance with higher experience. No pattern is discernible, with medians of senior employees falling below that of their mid-level counterparts in some cases. Employees with 5 years of experience appear to have the lowest scores on average at approximately 1.8. Overall, the performance ratings are scattered, suggesting experience may not be a significant factor in determining performance within the HR department.

Upon review of Figures 15-22, it appears that, generally, there is a correlation between experience and performance ratings within departments at KiwiLearn. This is most evident in the IT, Sales and Marketing where mid-level to senior-level employees show increasingly higher performance ratings on average. However, HR does not exhibit the same sentimentality in trend, with scattered points and variability across all experience levels. This suggests a low correlation and less predictability between experience and performance rating within the department. Recognising this, it can be inferred that an employee's years of experience do contribute to their overall performance ratings across most departments.

2.2 Assumptions and Hypothesis Formulation

2.2.1 Analysis Objective

The objective of this analysis is to investigate potential variations in employee performance ratings across different departments at KiwiLearn - HR, IT, Marketing, and Sales. Utilising hypothesis testing methods such as one-way ANOVA (Analysis of Variance) test, departments with notably higher or lower performance ratings will be identified, thereby revealing how department affiliation influences employee performance. Thus, KiwiLearn will gain insight into how departments within the organisation assess and evaluate performance rating.

Research Question: Is there a statistically significant difference in 'Performance Rating' across different departments at KiwiLearn?

2.2.2 Assumptions of Analysis

Before conducting the analysis, several key assumptions have been made:

- **Normality**: The data (performance ratings) within each department are assumed to be normally distributed. This assumption is necessary for performing an ANOVA test, with violations potentially producing unreliable results (predicted values, estimation of variance or affecting f-distribution critical values).
- **Homogeneity of Variances**: The variance in performance scores is assumed to be equal across departments. Violating this assumption could result in incorrect rejection of the null hypothesis during one-way ANOVA.
- **Independence**: The data (performance ratings) collected from one department does not influence other departments – samples are independent. This assumption is essential to ascertain valid statistical inferences and avoid biased estimates.

Visualize Normality - Histogram with KDE:

To check for normality and obtain a clearer picture of how ratings differ between various departments, histogram plots with Kernal Density Estimates (KDE) will be employed.

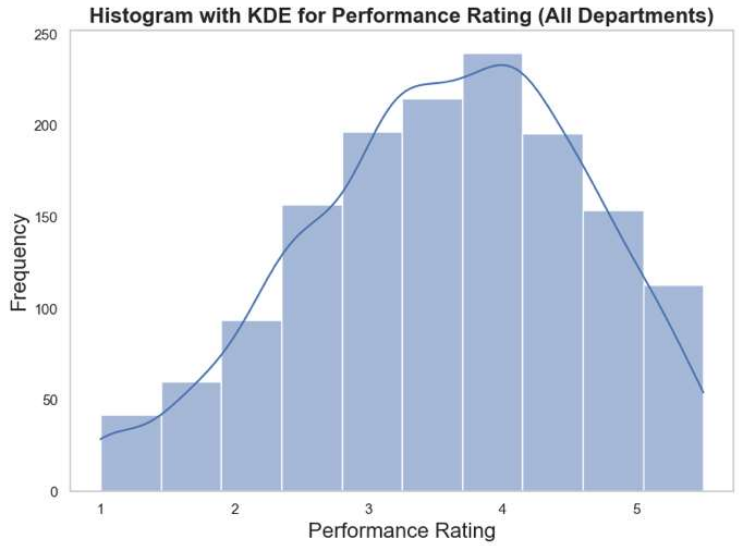

Figure 23: Histogram with KDE for Performance Ratings (All Departments)

A relatively normal distribution of performance ratings can be observed across all test streams, with a slight negative skew present. Generally, ANOVA testing is robust to minor deviations from normality, especially when using large sample sizes such as the 'Employee_Performance' dataset. As such the condition of normality has been met.

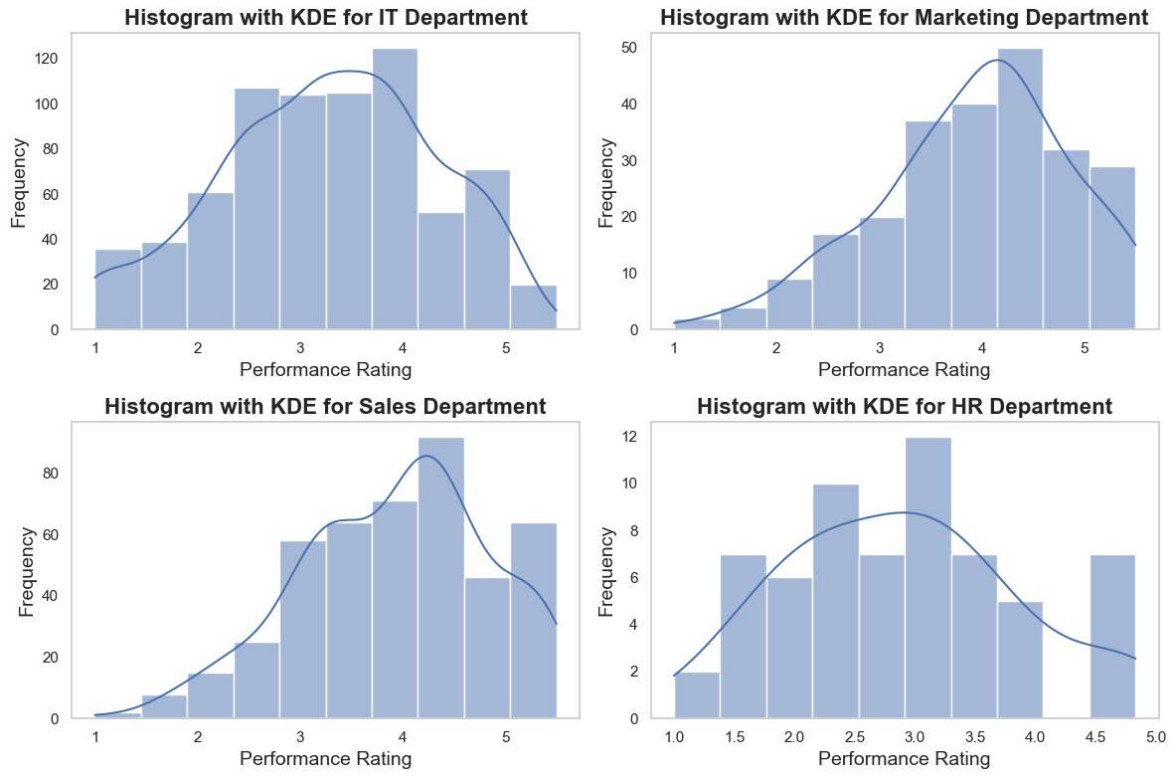

Figure 24: Histogram with KDE for Performance Ratings (Individual Departments)

Within the individual streams, The IT department appears most normally distributed, clustered around the KDE peak of approximately 3.5. Similarly, HR exhibits a normal distribution with a KDE peak of approximately 3. Both the Sales and Marketing department appear quite negatively skewed, with longer left tails and KDE peaks at approximately 4.2 and 4.1 respectively. This indicates that Sales and Marketing may be more generous with their ratings compared to other departments.

2.2.3 Hypothesis Statements (Null and Alternative Hypothesis)

Null Hypothesis (H_0): All group means are equal. There are no significant differences in performance ratings between the different departments.

- **H_0:** $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (Where μ_i is the mean performance rating for department *i)*

Alternative Hypothesis (H_a): At least one group mean is different from the others. There is a significant difference in performance ratings among the different departments.

- **H_a:** $\exists$ *i, j* such that $\mu_i \neq u_j$ (Where μ_i is the mean performance rating for department *i* and u_j is the mean performance rating for a different department, *j)*

2.3 Statistical Technique: Hypothesis Testing

2.3.1 Explanation of Statistical Method

The statistical method of one-way ANOVA (Analysis of Variance) testing will be conducted to compare the employee performance ratings across KiwiLearn's four departments: IT, HR, Sales, and Marketing. The parametric test is often used to examine and juxtapose a metric of two or more groups to determine statistical differences and is therefore appropriate to compare the mean performance ratings across different departments. Should statistically significant differences in mean be found across departments during ANOVA, a Tukey's HSD post-hoc test will be performed to identify which specific departments differ significantly from each other.

Using these statistical methods, differences in performance rating across departments can effectively and definitively be revealed.

2.3.2 Results of Hypothesis Testing

Table 9: Results of one-way ANOVA

Dataset statistical output	
	Sample Size
Department IT:	720 observations
Department Marketing:	240 observations
Department Sales:	445 observations
Department HR:	63 observations
One-way ANOVA Results:	
f-statistic:	61.45
Critical f-value:	2.61
p-value:	0.0000 (4dp)

The number of employees within each department can be observed within Table 11 – a useful metric when comparing performance ratings and aiding in contextualisation. The results from the one-way ANOVA test are also displayed. The F-statistic (61.45) and p-value (0.00, 2dp) aid in identifying statistically significant performance ratings between departments.

P-Value / Significance Level Analysis

Significance level: $\alpha = 0.05$

Conclusion: $0.000 < 0.05 \therefore$ Reject the null hypothesis.

The p-value of 0.000 (4dp) is unequivocally smaller than the alpha value of 0.05 displayed in Table 11. At the 5% significance level, it can be concluded that among at least some departments, performance ratings are significantly different.

F-Value / F-Statistic Analysis

Critical f-Value: 2.61

Conclusion: $61.45 > 2.61 \therefore$ Reject the null hypothesis.

Table 11 displays that the F-statistic (61.45) is greater than the critical F-value (2.61). Hence at the 5% significance level it can be concluded that there is a statistically significant difference in performance ratings between at least some of the departments.

2.3.3 Tukey's Post-hoc Test Results

Tukey's Honestly Significant Different post hoc test will be utilised to determine which specific department pairs have statistically significant different in performance ratings. Whilst ANOVA results revealed that overall, a significant difference can be observed among the groups, this statistical method accurately identifies the specific pairs with major differences and is of particular use when more than three groups are involved. Table 12 outlines findings produced by the test.

Table 10: Results of Tukey's HSD Post Hoc Test

```
Tukey's HSD Post Hoc Test Results:
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1    group2  meandiff p-adj   lower  upper  reject
-----------------------------------------------------
       HR        IT   0.3715 0.0217  0.0384 0.7047   True
       HR Marketing    1.027    0.0  0.6681  1.386   True
       HR     Sales   1.0256    0.0  0.6843 1.3669   True
       IT Marketing   0.6555    0.0  0.4665 0.8445   True
       IT     Sales   0.6541    0.0  0.5012  0.807   True
Marketing     Sales  -0.0014    1.0 -0.2044 0.2017  False
-----------------------------------------------------
```

Comparing the HR and IT departments, the mean difference is 0.3715, with a p-value (p-adj) of 0.0217, indicating a statistically significant difference between the two departments. Reject the null hypothesis.

A mean difference of 1.027 is present between HR and Marketing. The p-value (p-adj) of 0.0 indicatesa statistically significant difference between the departments. Reject the null hypothesis.

HR and Sales have a mean difference of 1.0256, with a p-value (p-adj) of 0.0, indicating there is a statistically significant difference between these departments. Reject the null hypothesis.

Comparing the IT and Marketing departments, the mean difference is 0.6555, with a p-value (p-adj) of 0.0, indicating a statistically significant difference is present between the two. Reject the null hypothesis.

IT and Sales possess a mean difference of 0.6541, with a p-value (p-adj) of 0.0. A statistically significant difference can be observed. Reject the null hypothesis.

A mean difference of -0.0014 is seen between Marketing and Sales. However, the p-value (p-adj) of 1.0 stipulates there is no statistically significant difference between these departments. Do not reject the null hypothesis.

Evaluating the above interpretations derived from Table 12, statistically significant differences can be observed between the following departments:

- HR and IT.
- HR and Marketing.
- HR and Sales.
- IT and Marketing.
- IT and Sales.

These results indicate that employees' performance ratings within the HR department are significantly different from all other departments. Similarly, IT has noticeably different ratings compared to Marketing and Sales, whilst the latter have no major differences in performance ratings. These insights are beneficial to KiwiLearn, allowing them to investigate any potentially critical or overly generous departments and take specific action to improve performance in certain departments.

2.4 Discussion and Conclusion

2.4.1 Conclusion – Interpretation of Results and Summary of Analysis

Through the exploratory analysis performed on KiwiLearn's "Employee_Performance" dataset, the potential variations in performance ratings across departments were investigated.

Utilising visualisations including scatterplots and box plots, the relationship between 'Experience' and 'Performance Rating' was explored. Figures 15 to 20 indicated the presence of a relationship between experience and performance rating, with entry and junior-level employees showing a wide range of performance variability compared to their senior counterparts who received consistently high ratings. The trend of increasing consistency and higher performance as experience increased was visible in the IT, Sales and Marketing departments. Interestingly, the HR department did not exhibit the same trend, with performance ratings scattered with no discernible pattern across all experience levels, as seen in Figures 21-22.

These findings suggest that whilst years of experience do correlate or partially influence employee performance, employees at any experience level can still obtain remarkably high-performance ratings – so a correlation relationship may be more appropriate than a causal one. Results provide valuable insight for KiwiLearn to focus their efforts, showcasing that entry and junior-level employees could benefit from more targeted training and support to obtain the same consistency as their senior counterparts. Management could look at what's working for high-performing juniors and create tailored training plans to improve this experience levels' overall performance in efforts- reducing variability.

Observing the histogram with KDE plots of individual departments displayed in Figures 23-24, it appears that Marketing and Sales have higher ratings on average than the other departments with a mean of approximately 4. This could indicate that managers within these departments are more generous with their ratings than others. However, whilst recognising the skew of these graphs favours higher performance ratings, the mean of other departments indicated by the KDE peak at approximately 3 shows that the difference between department averages may not be overly significant.

To determine potential variance in performance ratings across departments, One-way ANOVA testing was conducted. Results revealed a statistically significant difference between at least some of the departments. A p-value of 0.000 (4dp) falling below an alpha value of 0.5, and an f-statistic of 61.45 being greater than 2.61, elicited a rejection of the null hypothesis at the 5% significance level (Table 11). Tukey's Post Hoc test (Table 12) disclosed that both HR and IT had statistically significant differences in performance rating from all other departments.

This type of performance metric appears to work fairly well for IT, Sales, and Marketing. However, it may not be suitable for HR, which stands out with lower ratings. Management may need to investigate whether this is due to the nature of HR's work, how performance is measured or whether departmental issues not shown in the dataset could be affecting performance. Adjusting the metrics and providing more resources may aid in remedying this.

The noticeable difference in employee numbers across departments must also be considered. IT has 720 employees, Sales has 445, Marketing has 240, but HR only has 63. This difference could affect the results and may suggest HR is understaffed, with individuals handling a heavier workload, which could be impacting their performance.

Overall, this analysis provides a clear picture of how different departments at KiwiLearn assess and evaluate performance. Understanding these differences can guide better decisions to improve employee development and performance across the organization

3. Regression Analysis

To uncover underlying relationships between the independent parameters associated with employee performance, this exploration will utilize the statistical technique of regression analysis. Subsequently, insights into how employee performance ratings correlate with and are impacted by different variables in the dataset will be revealed. By employing this method, we aim to produce a regression model that effectively showcases how independent variables within the dataset influence employee performance ratings.

3.1 Identify Potential Predictor Variables

Ahead of performing this analysis, potential variables that may correlate with employee performance ratings must be identified. Upon inspection of the independent variables present within the dataset, it is likely that 'Experience' and 'Training Hours' would impact the performance rating of an employee. By implementing a correlation matrix, these suppositions can be reviewed, and a conclusion can be drawn.

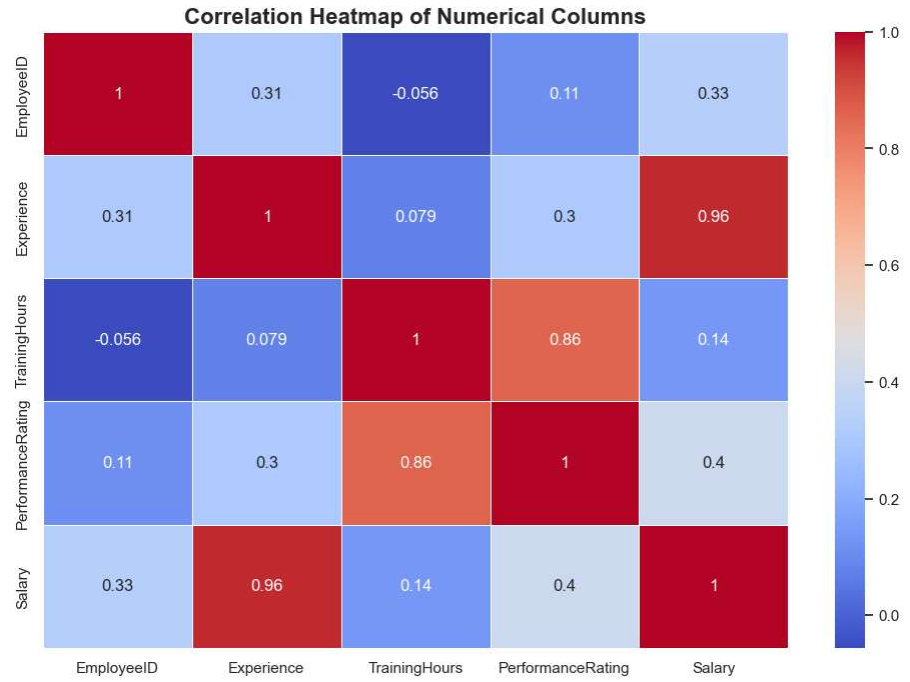

Figure 25: Correlation Heatmap for Numerical Columns

Upon examination, the correlation heatmap of numerical columns within the dataset (Figure 25) indicates that the independent variables with a potential correlation or impact on 'Performance Rating' are 'Experience', 'Training Hours' and 'Salary'.

'Training Hours' appears to have a strong positive relationship with 'Performance Rating' – suggested by the correlation coefficient of 0.86. This assumption can be deemed rational, as it is plausible employees with a higher number of training hours within the past year are likely to receive higher performance ratings.

With a correlation coefficient of 0.4, 'Salary' has a moderately positive relationship with 'Performance Rating'. This correlation value implies a relationship between the two variables is present. However, it may hint that employees that exhibit high performance ratings receive a high salary, rather than contrariwise

Although not as impactful as training hours within the past year, Experience' is realistically expected to impact the 'Performance Rating' an employee receives. This assumption is attested to by the correlation coefficient of 0.3 possessed – a somewhat moderate positive relationship.

'Employee ID' has the lowest correlation coefficient with Performance Rating, at a low value of 0.11. It is used more as an indexing or identification variable, and thus will not be considered as a potential predictor variable.

3.2 Assumptions for Regression Analysis

There are several assumptions that must be met to carry out successful regression analysis on a dataset and produce valid results.

Firstly, there must be no multicollinearity present between independent variables. Indicated by a correlation coefficient of 0.7 or greater, this multicollinearity can lead to unstable coefficient estimates and reduce interpretability - as individual effects become harder to discern. A linear relationship between the dependent and independent variables must also be present. This implies that changes in independent variables result in proportional changes in the dependent variable – in this case, performance ratings of employees. Violating this assumption can lead to incorrect conclusions or statistics such as the f-test and t-test results, which rely on linearity.

Normal distribution of residuals is another assumption that must be met for effective linear regression analysis. Estimates of standard error for coefficients may be impacted if this condition is not met, as well as p-values, which assume a normal distribution.

Finally, homoscedasticity should be exhibited, meaning the variability of residuals should remain constant. Constant variance indicates a suitable model fit, whereas heteroscedasticity suggests the model may not accurately capture the relationship between the independent and dependent variables. Violating homoscedasticity will affect the efficiency and validity of the statistical inferences derived from the model (e.g. t-tests, f-tests etc.)

Prior to creating the linear regression model, it is best practice for the preliminary assumptions of multicollinearity and linearity to be checked. As these assumptions significantly impact the validity and interpretability of the regression results, it is crucial these conditions are satisfied before advancing the analysis.

3.2.1 Multicollinearity

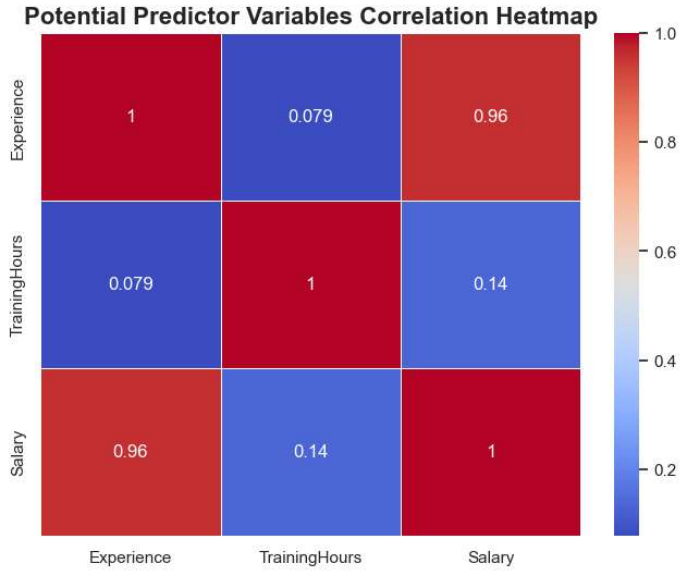

Figure 26: Correlation Matrix Heatmap of Independent Variables.

The heatmap depicted in Figure 26 showcases the correlation coefficients of independent variables within the dataset and highlights the multicollinearity present. Multicollinearity is determined to be existent when correlation coefficients of equal to or greater than 0.7 are visible between independent variables. In Figure 26, 'Salary' and 'Experience' have a correlation coefficient of 0.96 - a near perfect positive relationship – indicating high multicollinearity.

The strong relationship between these variables will influence the dependent variable, 'Performance Rating' and obscure the inferences drawn from the linear regression analysis should it not be confronted. To address this, an independent variable from the pair will be removed to limit the influence of multicollinearity on the statistical inferences made using the linear regression model. The independent variable that is least linear will be removed from the pair after checking for linearity between 'Performance Rating' and each independent variable.

3.2.2 Linear Relationships

To ensure accurate and reliable multiple linear regression analysis, the independent and dependent variables involved must possess a linear relationship. This allows for valid statistical inferences to be drawn from the regression model. Potential predictor variables that do not exhibit a linear relationship with the dependent variable will not be included in the regression analysis moving forward.

Figure 27: Scatterplots of independent variable 'Experience' against the dependent variable, 'Performance Rating'.

Figure 28: Scatterplots of independent variable 'TrainingHours' against the dependent variable, 'Performance Rating'.

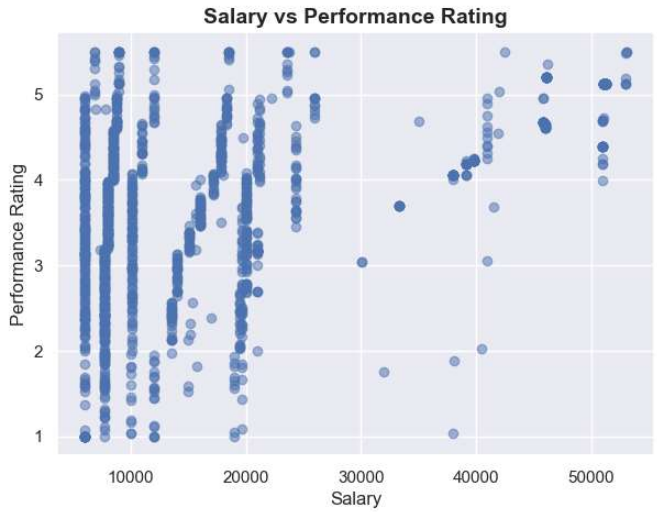

Figure 29: Scatterplot of independent variable 'Salary' against the dependent variable, 'Performance Rating'.

The clearest linear relationship can be observed between 'Training Hours' and Performance Rating'. A strong pattern is visible with points clustered in a straight line, exhibiting a positive linear relationship. Moving forward, 'Training Hours' will be included within the regression analysis.

Similarly, it appears that 'Experience' also has a linear relationship when plotted against 'Performance Rating'. There is a clear trend that 'Performance Rating' increases as 'Experience' does, despite not possessing as strong of a relationship compared to 'Training Hours'. Thus, it will take the place of the second predictor variable to form multiple linear regression analysis.

'Salary' does not possess a linear relationship with 'Performance Rating'. There is no apparent linear pattern visible within the graph, with points clustered together on the left-hand side, and others sporadically plotted on the right. This indicates a non-linear relationship; thus 'Salary' will not be included in the analysis moving forward.

3.3 Regression Analysis

3.3.1 Multiple Linear Regression Model

Using the selected independent variables 'Training Hours' and 'Experience', alongside the addition of a constant variable, multiple linear regression has been executed. The following regression results summary has been produced using the least squares method and provides insight into the overall fit and significance of the model, alongside individual variables' effect on employee performance ratings.

Table 11: Multiple Linear Regression Model Summary

OLS Regression Results

Dep. Variable:	PerformanceRating	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	2837.
Date:	Mon, 07 Oct 2024	Prob (F-statistic):	0.00
Time:	01:15:53	Log-Likelihood:	-984.63
No. Observations:	1468	AIC:	1975.
Df Residuals:	1465	BIC:	1991.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4880	0.043	11.446	0.000	0.404	0.572
Experience	0.0981	0.005	19.980	0.000	0.088	0.108
TrainingHours	0.0870	0.001	70.836	0.000	0.085	0.089

Omnibus:	337.380	Durbin-Watson:	1.789
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1335.451
Skew:	1.059	Prob(JB):	1.02e-290
Kurtosis:	7.165	Cond. No.	117.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

According to the R-Squared value of 0.795 displayed in Table 13, it appears that the overall model fit is suitable. 79.5% of the variability of the independent variable 'Performance Rating' can be attributed to the independent variables 'Training Hours' and 'Experience'. This is consistent with the adjusted R-Squared value which remains the same, solidifying the significance of the independent variable's role in the variation of employee performance ratings.

Provided in the table is an F-statistic of 2837 for the overall model significance. The associated p-value of 0.00 insinuates that the model is statistically significant, and at least one of the independent variables - 'Training Hours' or 'Experience' has a significant impact on employee performance ratings.

Insight is also provided into the individual predictor variables' significance in the form of coefficients, t-statistics and associated p-values. In the case of 'Experience' a coefficient of 0.098 (3dp) is present. The t-statistic of 19.980 and corresponding p-value 0.00 indicate that the coefficient is highly significant. Similarly, 'Training Hours' has a coefficient of 0.087 (3dp), a t-statistic of 70.836 and an associated p-value of 0.000 - suggesting that the coefficient is also greatly significant.

The absolute t-statistic of 'Training Hours' is larger than that of 'Experience', indicating that the independent variable 'Training Hours' is largely more significant than 'Experience'. Whilst this suggests 'Training Hours' is more influential on 'Performance Rating', both independent variables appear to be noteworthy.

The constant or intercept value can also be derived from the table. When both independent variables 'Training Hours' and 'Experience' are zero, 'Performance Rating' is expected to be at approximately 0.49 (2dp).

3.4 Assumptions of Linear Regression

After producing a regression model summary, it is imperative that the final assumptions of regression analysis are checked to ensure the results produced are valid, or whether further investigation is needed to ascertain the analysis is accurate. This includes the assumption of residual normality, and homoscedasticity.

3.4.1 Normality of Residuals

To determine if the residuals of regression analysis follow a specific distribution, a Q-Q or quantile-quantile plot can be used. This graphical tool enables the analyst to compare the quantiles of both the sample and theoretical distribution – in this case a normal distribution. If the points of the plot follow the line closely, it indicates that the residuals are of a normal nature.

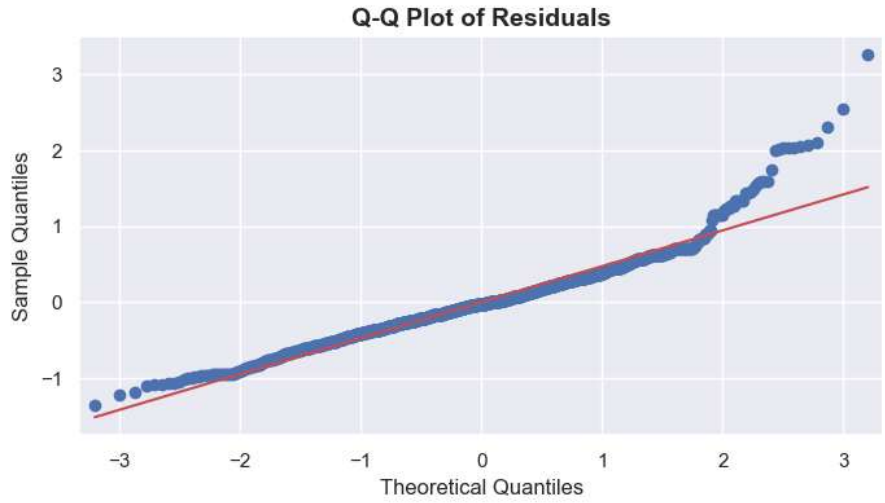

Figure 30: Q-Q plot of regression residuals

A conclusion can be drawn as to whether the residuals of the multiple linear regression are normally distributed by viewing the placement of points within the Q-Q plot displayed in Figure 30. Examining the plot, whilst a significant majority of the points closely follow the reference line, there are multiple points that begin to deviate at the tails, which may indicate non-normality of the distribution. A heavy-tailed distribution appears to be present, insinuating that the tails of the sample are heavier than the theoretical distribution. This is indicated by the portion of points on the left and particularly the right-tail which lie above the reference line.

Recognising the few strong indicators of non-normality within the residuals, it would be advantageous to conduct another normality test and corroborate results to ensure the assumption is satisfied or solidify its violation. In contrast to the graphical tool of a Q-Q plot, the Anderson-Darling test, a statistical method, provides insight into the normality of residuals based on a critical value. Should the statistic fall below the critical value, the null hypothesis remains – in this case a normal distribution – whereas if the statistic is higher than the selected critical value, the null hypothesis is rejected.

Table 12: Anderson-Darling Normality Test Results

Normality Test Results:					
Anderson-Darling Statistic:	10.59 (2dp)				
Critical Values:	[0.574	0.654	0.785	0.916	1.089]
Significance Levels:	[15	10	5	2.5	1]

The Anderson-Darling statistic of 10.59 (2dp) derived from Table 14 is greater than the critical value 0.785, indicating at the chosen significance level of 5, the null hypothesis is rejected, and the residuals showcase a non-normal distribution. At all other significance levels, the statistic produced is still drastically greater than the critical values – solidifying a non-normal nature of residuals. This testifies to the observation of the Q-Q plot in Figure 30.

Acknowledging the graphical and statistical methods used to determine normality of residuals, and the subsequent violation derived from these tests, further consideration must be taken to assess the legitimacy of the linear regression model results. A lack of residual normality can influence statistical tests such as t-tests and f-tests, and lead to incorrect p-values, inflated or deflated standard errors and confidence intervals. In turn this also affects any predictions drawn, especially in the tails of the distribution.

Action must be taken to ensure accurate and reliable inferences. Whilst they will not be conducted, mitigation strategies could include data transformation to stabilise variance and make the distribution more symmetric, such as the box-cox, square root or log transformation. Alternatively, methods such as robust regression techniques including least absolute deviations (LAD) which are less sensitive to non-normality, or quantile regression which estimates the conditional median instead of the mean (and is therefore less susceptible to non-normal data) could be considered to address non-normality of residuals. These methods may improve the reliability of the analysis.

3.4.2 Homoscedasticity

The final assumption is that of homoscedasticity, or constant variance of residuals. Similarly to other assumptions of regression analysis, it must be satisfied to ensure validity and is an indicator of model suitability. The presence of a pattern in the plot (distinctly fanning out or narrowing) indicates heteroscedasticity, whereas a lack of any identifiable pattern results in homoscedasticity.

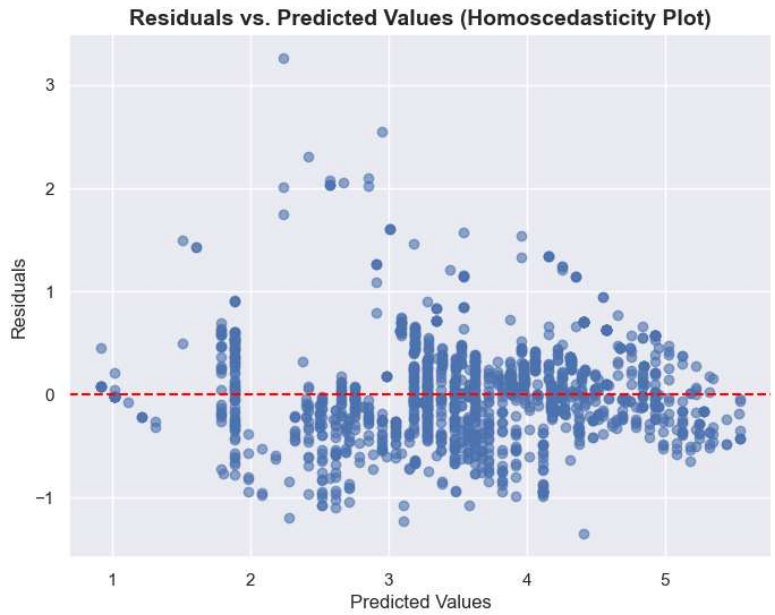

Figure 31: Homoscedasticity plot of residuals vs predicted values.

There is no apparent pattern seen within the homoscedasticity plot (Figure 26). Plot points are spread erratically with no particularly clear fanning or narrowing visible. Points are spread closely and distantly from the reference line at all levels of the independent variables 'Training Hours' and 'Experience'. This lack of pattern indicates the assumption of homoscedasticity is satisfied.

Homoscedasticity is a crucial assumption of regression analysis that impacts the validity of statistical inferences derived from the model, and whilst the presence of it is a positive sign, caution should still be exercised when interpreting results due to the absence of non-normality of residuals.

3.5 Discussion and Conclusion

3.5.1 Conclusion – Interpretation of Results and Summary of Analysis

By conducting regression analysis on KiwiLearn's "Employee_Performance" dataset, underlying relationships related to employee's performance ratings have been revealed and explored.

Prior to the analysis, potential predictor variables were drawn from the employee-related variables by checking for multicollinearity and observing linearity. The correlation matrix in Figure 25 indicated 'Experience', 'Training Hours' and 'Salary' were potential predictor variables. Analysing Figure 26, it was discovered that 'Experience' and 'Salary' were highly correlated and thus would skew inferences drawn from the regression model. To mitigate this, linearity was checked in Figures 27-29, and 'Salary' was subsequently removed as a predictor variable.

Utilizing multiple linear regression, insights into the resulting potential influencing or correlating factors 'Training Hours' and 'Experience' have been gained. The relationship between both predictor variables and the independent variable 'Performance Rating' are statistically significant, indicated by the zero p-values presented in Table 13. Observing the t-statistics present, 'Training Hours' seems to be the most significant predictor variable with a higher value of 70.836. The predicted intercept value of 'Performance Rating' when the independent variables are zero was also produced. Overall, the fit of the regression model was deemed suitable and statistically significant with an R-Squared and adjusted R-Squared value of 0.795, an F-statistic of 2837 as well as an associated p-value of 0.00 (Table 13).

3.5.2 Limitations and Potential Bias

There were, however, limitations to the analysis conducted. Most crucial assumptions of linear analysis were satisfied; however, the normality of residuals was not. Whilst non-normality of residuals does not necessarily invalidate the regression model completely, it is important to consider the potential effects and examine any possible causes for a more complete analysis. As discussed previously, non-normality may impact validity of hypothesis tests and confidence intervals – as many regression tests assume normal distribution. Transforming the data or performing more robust regression techniques may be beneficial.

Other data limitations or potential biases may be present in the data. The analysis was performed under the assumption that no selection bias was present, and that the full employee staff, or a representative sample were included. Should this not be the case, the results concluded may not be representative of the broader employee population at KiwiLearn. Similarly, it was assumed that employee performance ratings were determined fairly rather than subjectively. If this were not true, there may be a bias present in the data influenced by personal reasons or other external factors – such as a specific department being less generous with their ratings. The data does not indicate whether values contained are an average over time (e.g. A year) or taken only at one point. Should the data be indicative of only one point, it may not capture the dynamic nature of performance ratings and the influence of other variables over time.

Finally, it must be noted that whilst regression analysis identifies associations/correlations between 'Training Hours', 'Experience' and 'Performance Rating', this does not necessarily imply causation.

3.5.3 Further Research and Analysis Improvements

Within the current analysis, further improvements would include mitigating the effects of the non-normal residuals using aforementioned techniques. For further research and improvements in overall analysis, it would be beneficial to conduct a longitudinal study of employee's performance at KiwiLearn. Collecting the data over an extended period will allow for improved understanding of causation by recording changes in performance and the effects of independent variables over time. This also accommodates for industry trends and mitigates other external factors/anomalies. It would also be advantageous to include other objective performance metrics such as sales numbers, bugs/errors fixed, or students reached alongside the currently more subjective performance ratings to reduce any potential bias and observe how different types of performance are measured.