Student ID Number: 23215381,21151144

Assignment 2

Semester 2 2024

PAPER NAME: Data Analysis

PAPER CODE: COMP517

Student ID	Student Names
23215381	Ali Abbaspour Mojdehi
21151144	Mirasha Fernando

Due Date: Midnight Friday 18th 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. **Submit your code separately**.

Table of Contents

List of Figures

INTRODUCTION

This report is based on 'EmployeePerformance.csv', a dataset containing information about employees at KiwiLearn. KiwiLearn is a leading educational provider in New Zealand. A Brief description of the variables in the dataset is given below.

- **Employee ID**: A unique identifier assigned to each employee.
- **Department**: Names of the four departments at KiwiLearn: IT, Marketing, Sales and HR.
- **Gender**: Male or Female.
- **Experience**: Contains the total number of years of professional experience each employee has on a range of 0-9 years.
- **Salary**: The monthly income earned by each employee.
- **Performance Rating**: A measure of the performance of each employee within the company, on a scale of 1-5.5.
- **Training Hours**: The number of hours each employee has spent training within the past year.

The purpose of this analysis is to show what kind of factors affect the performance ratings of employees in different departments. Factors such as salary, training hours, experience, gender and department will be examined to determine how much of an effect they have on the performance level of employees.

We will be using the below libraries for this analysis.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm
from statsmodels.stats.multicomp import MultiComparison
import statsmodels.api as sm
from scipy.stats import skew, kurtosis
```

Figure 1 – List of libraries used for the analysis

Further explanation on the list of libraries shown in figure 1 is given below.

- The '**NumPy**' library is used for basic mathematical calculations.
- The '**Pandas**' library is used for data manipulation and analysis. Used for data frame cleaning, filtering, aggregating etc.
- '**Seaborn**' and 'matplotlib.pyplot' is mainly used for data visualization. Informative and attractive plots can be created using this.

- '**Scipy'** is used for scientific calculations. Mostly for statistical tests.
- '**statsmodel'** is used for statistical modelling. Good for hypothesis testing, time series analysis, etc.
- '**statsmodels.stats.multicomp.MultiComparison'** is used for multiple comparison specifically for the post-hoc tests.
- '**scipy.stats.skew'** and '**scipy.stats.kurtosis'** is used to measure skewness and kurtosis of the distribution of the variables in the dataset.

PART ONE - Exploring Data and Testing Hypotheses

Data Preparation and Exploration

Figure 2 shows an overview of the dataset. It has 7 variables, including 3 categorical variables and 4 numerical variables.

	EmployeeID	Department	Gender	Experience	TrainingHours	PerformanceRating	Salary
0	1001	IT	Male	4	5	1.00	19000
1	1002	Marketing	Female	0	50	5.50	6900
2	1003	Sales	Male	0	5	1.00	6000
3	1004	HR	Male	1	5	1.00	6000
4	1005	HR	Female	9	5	1.04	38000
5	1006	IT	Female	4	5	1.05	19000
6	1007	Marketing	Male	1	50	5.50	9000
7	1008	Marketing	Female	1	50	5.50	9000
8	1009	Sales	Male	0	5	1.37	6000
9	1010	HR	Male	1	15	1.47	7700

Figure 2 - Overview of the dataset

For analysis purposes, we will convert the department and gender variables into numerical data. This is done by assigning the four departments with numerical values. HR as 1, IT as 2, Marketing as 3 and Sales as 4. Same with gender, 1 and 2 for male and female respectively. Figure 3 shows the summary statistics of the updated dataset.

	Mean	Median	Standard Deviation	Variance
EmployeeID	1734.500000	1734.50	423.919411	1.797077e+05
Experience	2.838556	2.00	2.527657	6.389051e+00
TrainingHours	32.144414	31.00	10.106029	1.021318e+02
PerformanceRating	3.561512	3.63	1.044987	1.091997e+00
Salary	16107.623297	10100.00	12158.438481	1.478276e+08
dprtmnt	2.726839	2.00	0.944060	8.912492e-01
gend	1.398501	1.00	0.489756	2.398614e-01

	Minimum	25th Percentile	50th Percentile (Median)
EmployeeID	1001.0	1367.75	1734.50
Experience	0.0	1.00	2.00
TrainingHours	5.0	25.00	31.00
PerformanceRating	1.0	2.84	3.63
Salary	6000.0	7700.00	10100.00
dprtmnt	1.0	2.00	2.00
gend	1.0	1.00	1.00

	75th Percentile	Maximum	Skewness	Kurtosis
EmployeeID	2101.25	2468.0	0.000000	-1.200001
Experience	4.00	9.0	0.944505	-0.105225
TrainingHours	39.00	50.0	-0.380352	-0.077315
PerformanceRating	4.33	5.5	-0.309241	-0.461060
Salary	20000.00	53100.0	1.629655	1.764965
dprtmnt	4.00	4.0	0.261255	-1.341147
gend	2.00	2.0	0.414627	-1.828084

Figure 3 - summary statistics of the variables

If we look at **experience**, the mean (2.8) is higher than the median (2.0), indicating a right skew in the data, this is confirmed by the skew reading of 0.94. With the minimum reading being 0 and max being 9.0, the standard deviation of 2.5 is relatively high, the positive skew of 0.94 also leads us to the assumption that there are more employees with low experience and there is high variation in their experience levels.

Training Hours has a high standard deviation but low skew value indicating the distribution is relatively normal. The mean is slightly higher than the median indicating a slight skew, which means more employees have training hours above the mean. Minimum training hours is 5 hours.

In **performance rating**, mean and median are close in value. The standard deviation and skew value are low indicating a distribution that is close to normal. Half of the employees have a performance rating below 3.63.

Salary has a wide range from 6000 to 53100, it also has a high standard deviation, indicating significant variability in salaries across the employees. The distribution is positively skewed, with a long tail on the right, indicating a few employees earn significantly more than the average.

```
Missing values of the dataset:
EmployeeID           0
Department           0
Gender               0
Experience           0
TrainingHours        0
PerformanceRating    0
Salary               0
dtype: int64
```

Figure 4 - missing values

```
Number of duplicate rows: 0
```

Figure 5 - duplicates rows

Figures 4 and 5, shows us that there are no missing values or duplicated data. If we test for the outliers, we can see outliers only in the salary variable (Figure 6).

```
Number of outliers in each variable:
EmployeeID           0
Experience           0
TrainingHours        0
PerformanceRating    0
Salary               7
dprtmnt              0
gend                 0
dtype: int64
```

Figure 6 - number of outliers

Further consideration about the salary variable, the outliers in it are important to our analysis. This information could be related to/affecting our dependant variable. We cannot remove this data. Figures 7 and 8 will provide context to these outliers.

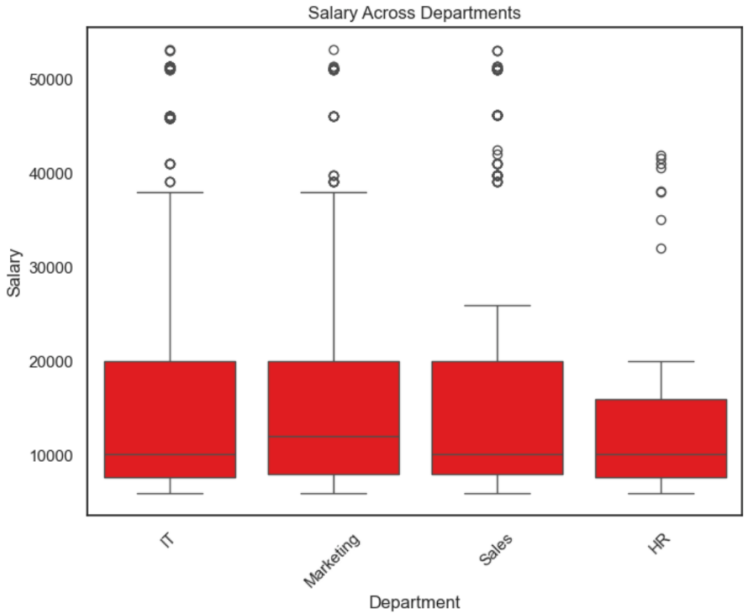

Figure 7 - boxplots showing salary within each department

	EmployeeID	Department	Gender	Experience	TrainingHours	\
1082	2083	IT	Female	9	35	
1189	2190	IT	Male	9	25	
1306	2307	Sales	Male	9	35	
1338	2339	Sales	Female	9	48	
1404	2405	Sales	Male	9	48	
1421	2422	Marketing	Female	9	48	
1460	2461	IT	Male	9	10	

	PerformanceRating	Salary	dprtmnt	gend
1082	5.12	53010	2	2
1189	5.12	53010	2	1
1306	5.12	53010	4	1
1338	5.19	53010	4	2
1404	5.48	53020	4	1
1421	5.50	53100	3	2
1460	5.50	53100	2	1

Figure 8 – Tabular view of outliers with all information.

Figure 7 shows the boxplots of salary within each department, and we can see that it contains several outliers. With closer examination of this data, we see that all these outliers are of employees with 9+ years of experience and exceptional performance ratings (Figure 8). Since our exploration is focused on this area of the data, it is crucial to include these outliers. Keeping this data will give us a more accurate understanding of how salary and performance are linked. These outliers represent important points about the workplace structure, and they may be useful for us as extreme salary holders show exceptional performance ratings. Removing this data will inhibit us from clearly understanding the effect of salary on the performance levels of the employees within each department.

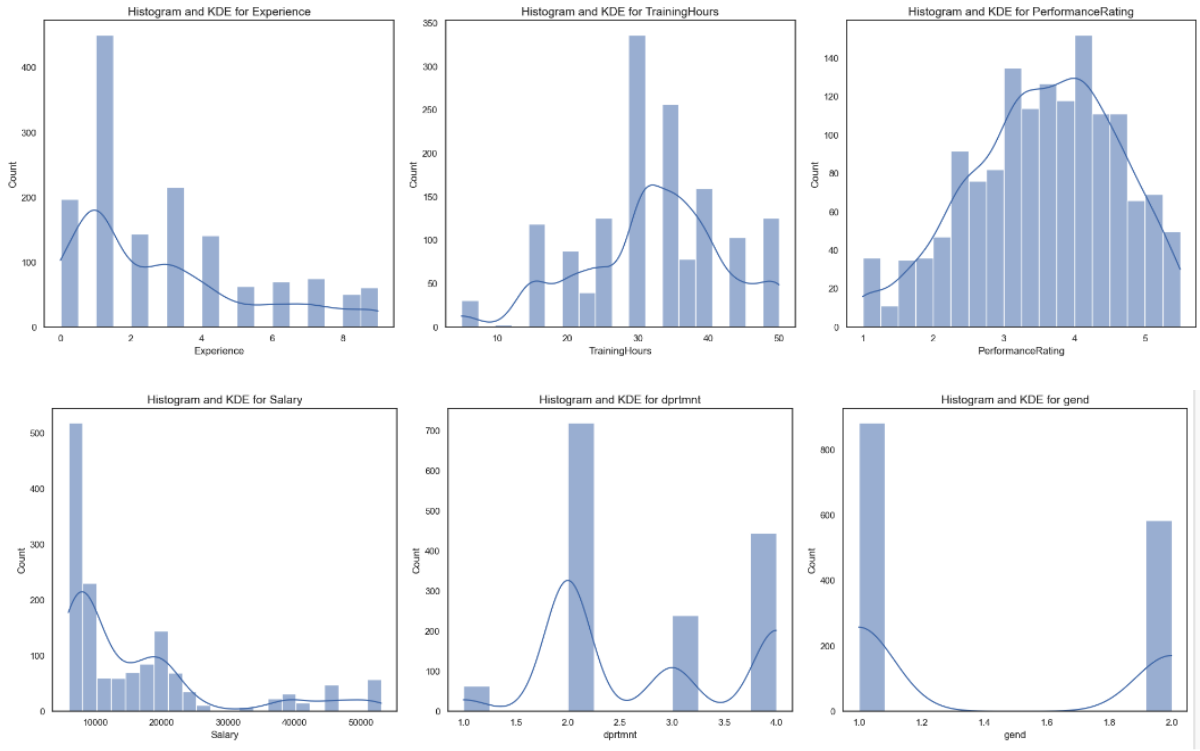

Figure 9 Histograms of all relevant columns with KDE

Figure 9 with the histograms give us a clear visualization of the distribution of each variable in our dataset. Starting from the top left we have a visualization of the relative size of each department, IT is clearly the largest, HR is the smallest.

The graph showing gender tells us there is more men in the company than women.

The salary plot gives a good indication of the pay discrepancy within the company, generally it looks like employees are paid a similar monthly salary, however we have already seen some outliers.

The experience figure on the bottom left tells us there are more employees with less experience currently working at the company.

The TrainingHours graph indicates majority of employees receive above average training, indicating a positive company culture regarding professional development.

The performance rating graph displays an almost textbook bell-shaped curve indicating a close to normal distribution of employee performance ratings.

To provide a more vivid visual representation of key statistical values within each variable we made a series of boxplots (Figure 10), this allowed us to develop a better fundamental understanding of our data.

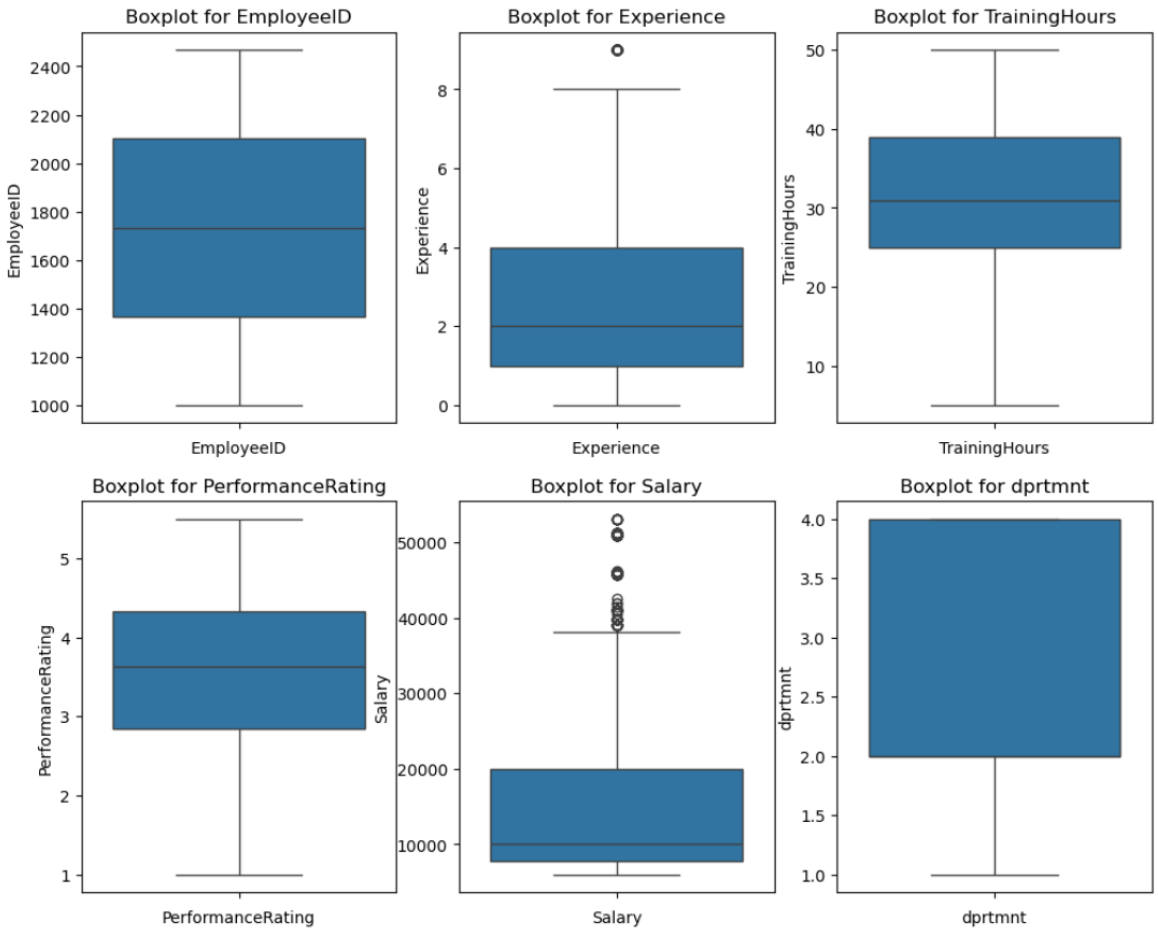

Figure 10 boxplots for all variables

Figure 10 allows us to quickly observe and compare key statistics relevant to our analysis. At a glance we can derive median values, IQR, range and any potential outliers for each variable.

Experience

- Median value is around 2 years of experience across the whole company, suggesting that many employees are early in their career.
- The middle 50% of employees (IQR) have between 1 and 4 years of experience.
- There are outliers shown but as the range only spans from 1-9 and a majority of employees have low experience, this is possibly just a representation of the lack of senior employees.

TrainingHours

- The median training hour is about 30 Hours.

- Most employees receive between 25 and 40 hours of training a year (IQR).
- There is a wide range within this variable with the maximum being 50 hours and minimum being close to 0.

PerformanceRating

- The median performance rating is around 3.5-4, suggesting a general high level of performance in the company.
- Most employees are rated between 3 and 4.5 in terms of performance (IQR)
- The scale spans from 1(low) – 5.5 (high).

Salary

- The median monthly income for an employee at KiwiLearn is around $10,000
- Most employees receive an income between $8,000 and $20,000 (IQR)
- The salary variable has a very wide range indicating significant disparity between the two extremes in terms of pay. We also notice the most outliers in this variable. The higher salaries are potentially skewing the data in this column and the average salary is likely much lower.

Following this examination of the relevant variables present in the dataset we thought it was appropriate to develop a representation of the makeup of the business. An easy-to-understand visualization in this case is a pie chart.

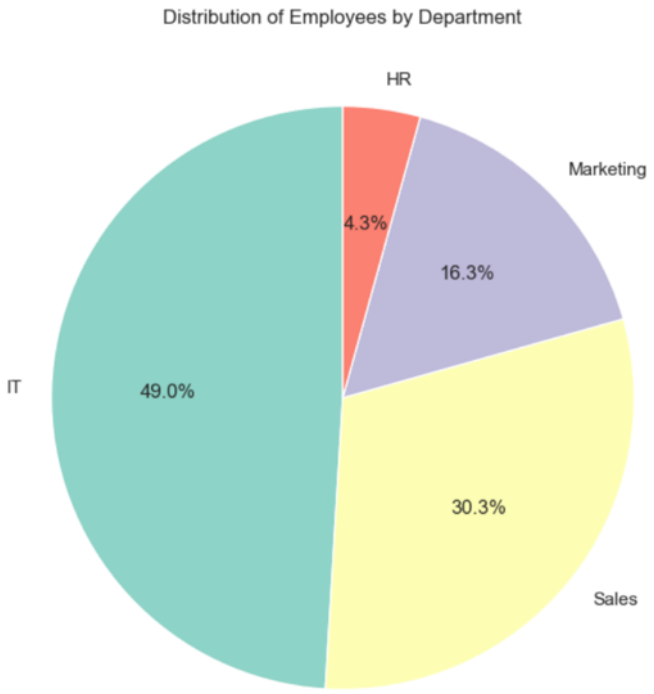

Figure 11 pie plot showing distribution of employees by department

Figure 11 is a representation of the distribution of Employees within the firm, each colour represents a department, the size of each segment corresponds to the percentage of the employees in that department. We can use this as a visual representation of how the organization allocates its workforce, this will help us in understanding the structure and focus of the organization.

Immediately we can make a few key observations.

IT makes up almost half of the entire company, indicating a technology driven approach to the development of its resources. IT clearly plays a significant role in the company's operations.

Sales is the second largest department with 30.0% of total employees coming from sales. This suggests sales is also a large part of the business model, potentially for customer acquisition, revenue generation and client interactions.

Marketing is the third largest department with 16.3% of total employees. Marketing makes up a smaller portion than sales (it is still significant), potentially indicating an active sales focused approach as opposed to a more passive advertisement approach to distribute its product.

HR is the smallest department with 4.3% of total employees. HR is likely focused more of administrative tasks like recruitment and employee relations, HR departments are often the smallest in any organization however they play a significant role in daily operations.

The previous explorations of the data have provided us with the necessary context to go forth with our analysis. From this point onwards we will be looking at the interplay between variables, trying to find patterns in the data and examining correlations found between our variables. Our focus will be determining the causes of the variations within employee performance rating across different departments. We will use the following multivariate analyses to further our understanding of the data.

Multivariate Analysis

To strengthen our understanding of the relationships between variables in the EmployeePerformance dataset we performed an in-depth multivariate analysis. This presented us with a wide array of visualizations, allowing us to closely examine multiple variables simultaneously and draw insights as to how they may be affecting one another.

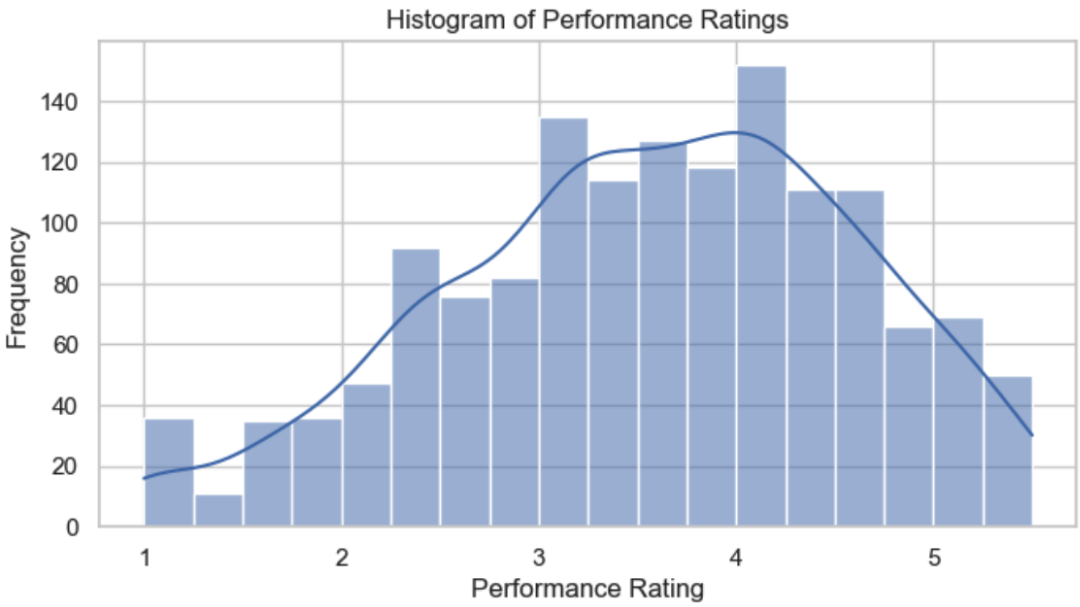

Figure 12 Histogram (with KDE) showing the distribution of performance rating scores

Initially we wanted to evaluate the overall spread of the Performance Rating variable as this was to be a key for our analysis. We can see from figure 12 that the distribution is roughly bell shaped although a slight positive skew is observed. This indicates that while most employees have a performance rating of 3-4, a small group of employees have an above average rating score of above 4. The peak being around 3-4 suggests this is the average rating of the employees. This visualization brought forth the question, which departments have the highest performance rating on average?

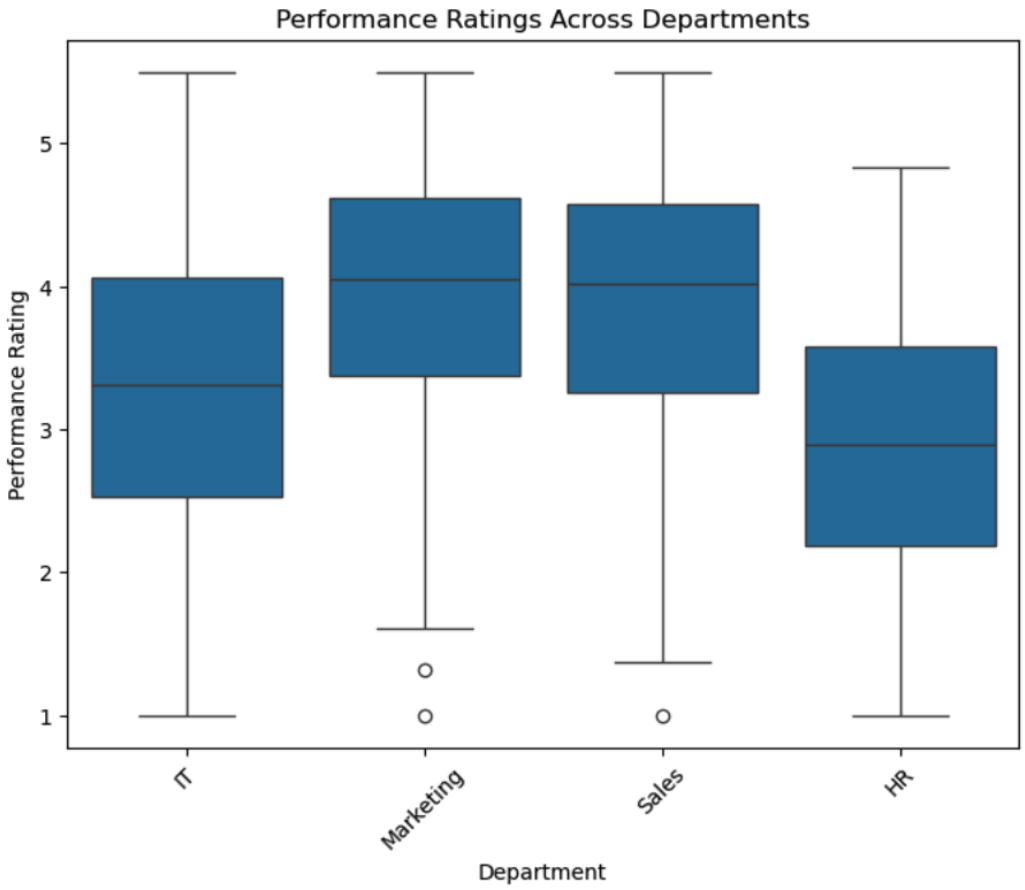

Figure 13 Boxplots showing Performance rating in each Department

Figure 13 is a series of boxplots showing performance rating scores for each Department. We can see the median (represented by the line through each of the boxes) in Marketing and Sales are much higher than that of IT and HR. Furthermore, the interquartile range (represented by the size of the respective boxes) shown in Marketing and Sales are noticeably smaller than the remaining two departments. Another observation from this plot is that the whiskers (representing range of data) for Marketing, Sales and IT all extend to the upper limit (5.5) of the scale, while HR has a significantly shorter range. These observations tell us several things about the departments at KiwiLearn.

- HR seems to be the lowest performing department, consisting of a low median and wider variability in the performance ratings of its employees.
- IT department also has a low median and high degree of variability in scores; however, it has a much wider range than HR and a much higher median rating. We know from previous analyses that IT is the largest out of all departments, so the range and IQR being larger are more understandable.
- Sales and Marketing look very similar on this visualization, with the highest median scores and more consistent ratings (shorter boxes). This suggests the employees in these departments perform more uniformly with fewer extreme ratings.

This got us questioning why certain departments perform better than others. The organization must ask itself, what are the factors contributing to employee performance?

A tool we can use as analysts to easily find and measure correlations between different variables, is a simple correlation heatmap which is used to visualize the correlation coefficient (R).

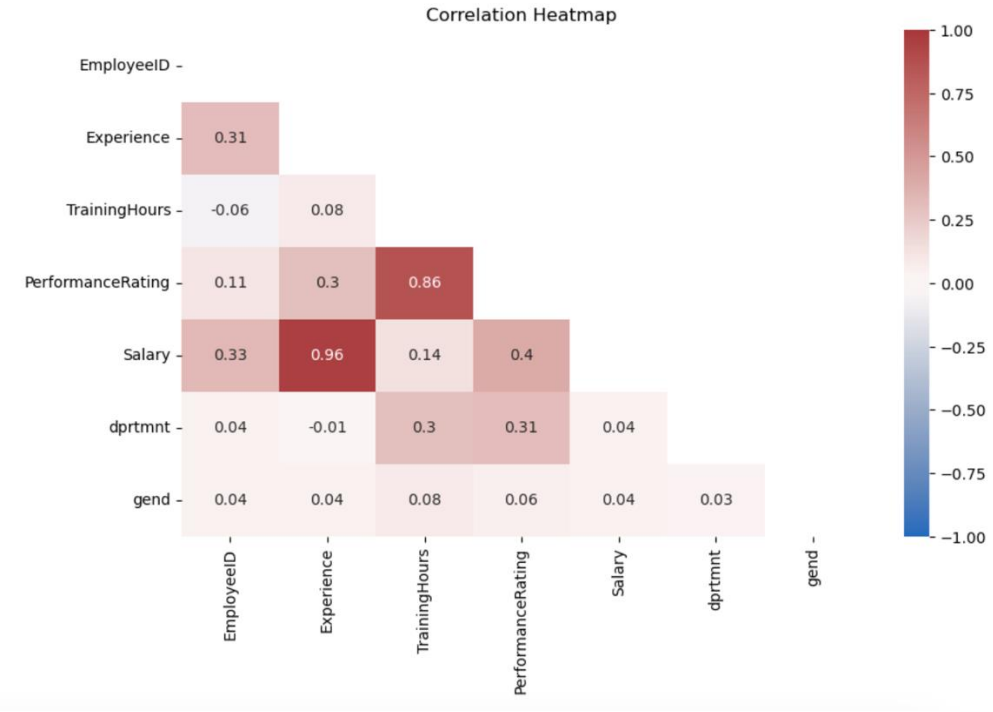

Figure 14 Correlation Heatmap showing r coefficient between all variables

While the heatmap in figure 14 can help us understand *all* variables better, our analysis is focused on the **PerformanceRating** variable, and so we shall refrain from straying too far from this recognised dependant variable. Instead, we will be concerned only with the independent variables that show a strong relation with the dependant.

Key findings from Correlation Matrix

PerformanceRating & TrainingHours – R = 0.86

- The relationship shown here between these two variables is difficult to overlook. The positive coefficient indicates that an increase in either one results in an increase in the other. From this we can conclude that employees who receive more training, have a higher performance rating. Therefore, training must directly improve employee performance. Conversely, employees with less training tend to have poorer performance rating scores, an interesting investigation point. **(Which departments receive more training?)**

PerformanceRating & Salary – R = 0.4

- This is a logical assumption to draw, stating that employees who perform better are rewarded with a higher salary. Although this relationship is not as strong as the one between salary and experience, it is significant enough to be mentioned and studied further in this analysis. **(Which departments earn more money?)**

Salary & Experience – R = 0.96

- The most significant relationship shown on this correlation matrix is the observed relationship between these two variables. Employees with greater industry experience are rewarded, tending to have higher salaries. Although this may seem like a positive for the company, it could be negatively effecting output as the relationship between experience and performance rating is not as significant as it could be. Are employees being rewarded for the wrong things? **Maybe the relationship between performance rate and salary should be stronger to motivate employees to perform to a higher standard.**

PerformanceRating & Experience – R = 0.3

- As previously mentioned, this relationship is notable but somewhat underwhelming considering the implication of the correlation between experience and salary. This also raises the issue of multicollinearity between salary and experience as they are so highly correlated with one another, we cannot effectively use both of these variables in our final analysis. **(Which departments have higher experience on average?)**

PerformanceRating & Department – R = 0.31

- The correlation between PerformanceRating and Department is notable however department does not seem to be the most significant determinant of employee performance. We know from previous analysis there is clearly a relationship between these variables, however, the correlation matrix implies that it is not one of major significance.

How does training differ between departments

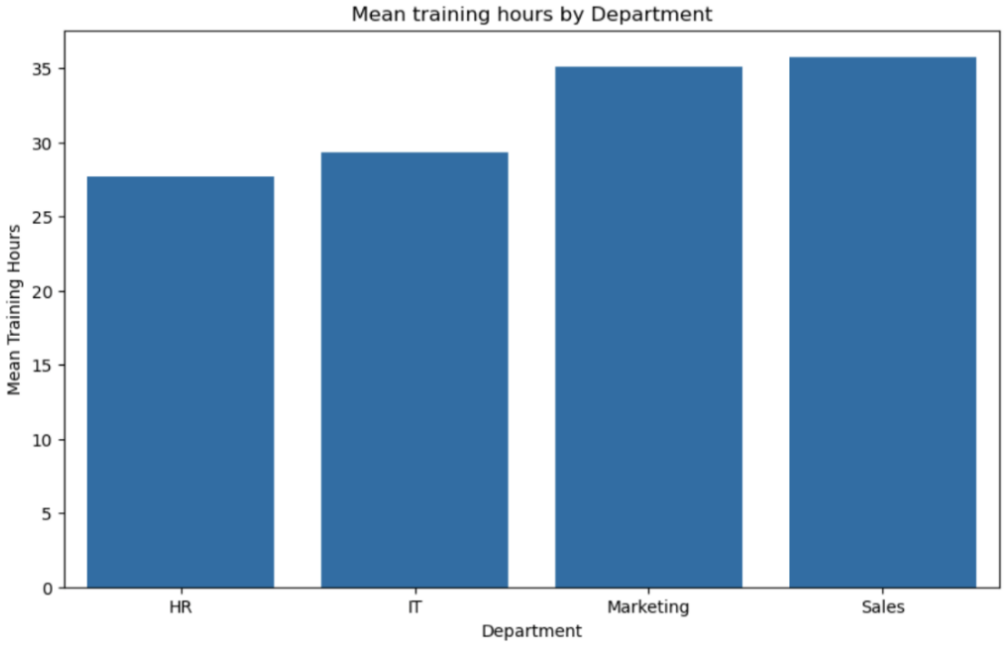

Figure 15 Average TrainingHours for each Department

The bar plot in figure 15 with mean TrainingHours for each department provides us with additional context after having analysed the correlation matrix. Drawing from insights gained through previous visualizations we can make a few generalisations regarding our departments and their performance rating.

Firstly, it is possible that Sales and Marketing have stronger internal management and a greater focus on output when compared to other departments. We can see this by looking at the provided figure and figure 13 showing performance rating, while keeping in mind the relative size of these departments (large but not largest, figure 11).

Secondly, we see a striking resemblance when comparing the difference in mean training hours between the departments shown here and the median values of performance rating. Below is a visualization showing mean PerformanceRating for each Department in the same format.

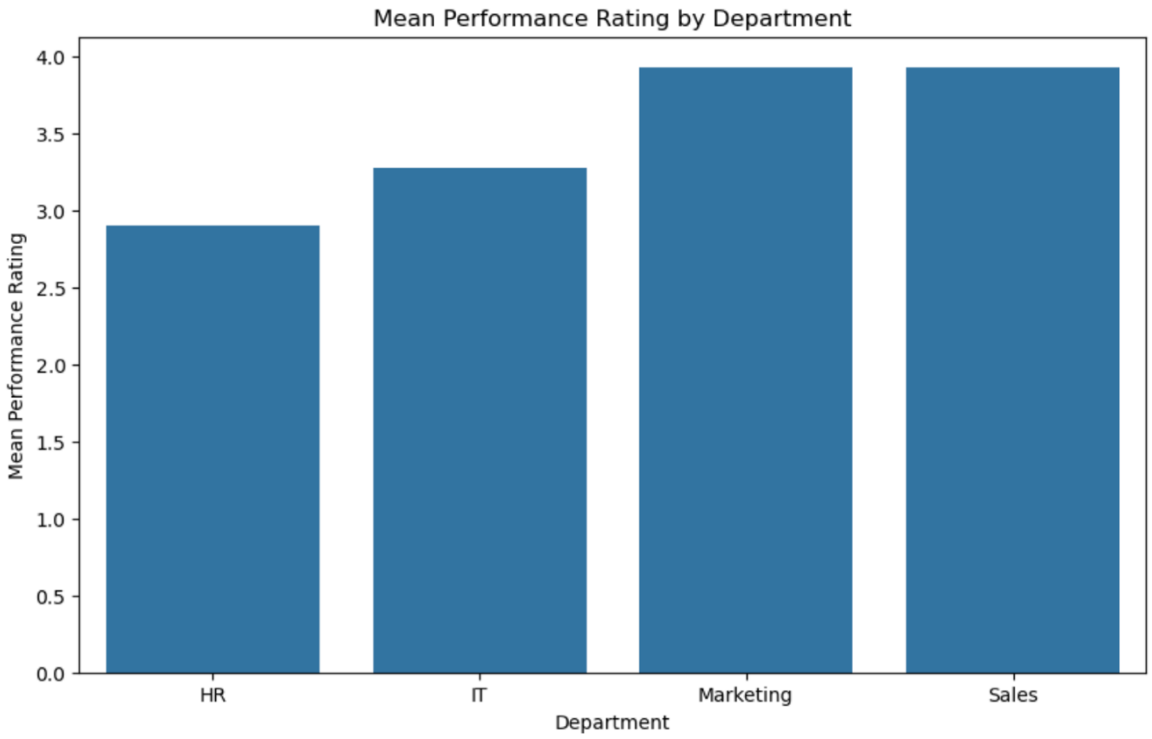

Figure 16 Mean PerformanceRating in each Department

The advantage of displaying this graph is our ability to look between figure 15 and 16 to notice the similarities in the ratio of difference between each department. It is obvious to us that the two graphs are strikingly similar, this observation aligns with the correlation value shared between these two variables (TrainingHours and PerformanceRating, r value = 0.86, figure 14). This solidifies our assertion that TrainingHours is key in determining PerformanceRating among the employees.

Salary Analysis

Another significant variable shown in previous visualizations has been Salary. It is important for us to further understand this variable's spread, effect and response to other attributes of our dataset. Firstly, we will examine the salary ranges within each department through the use of boxplots.

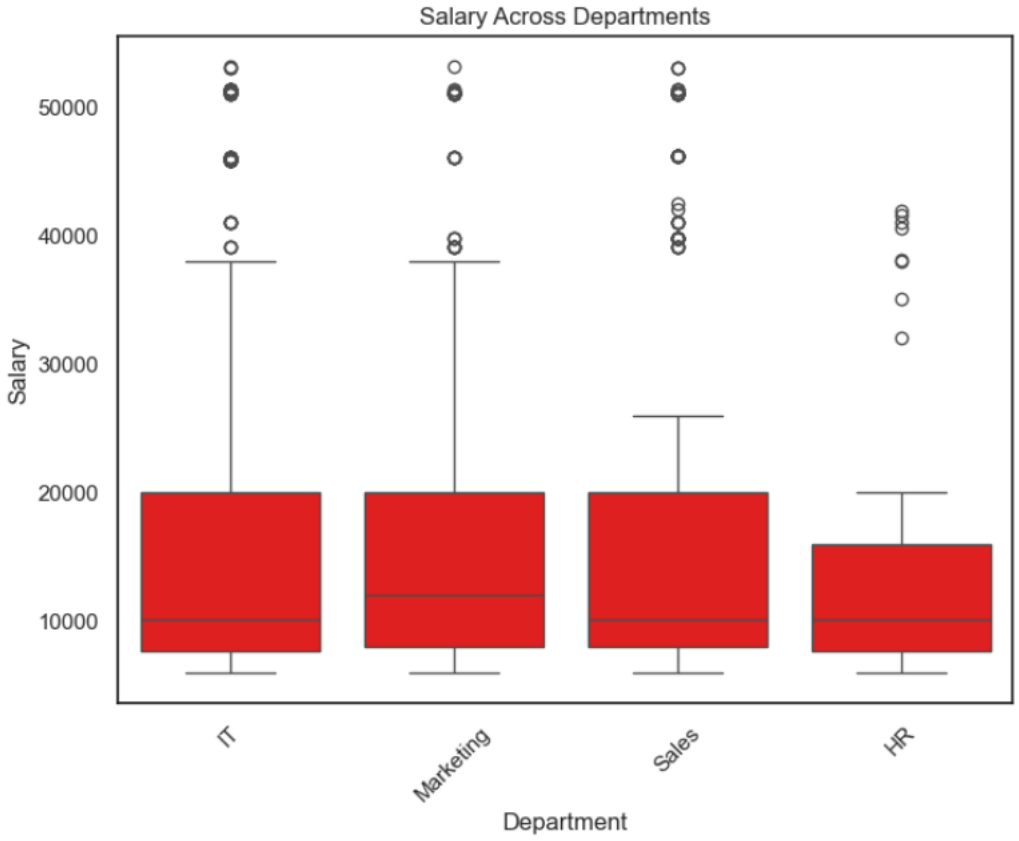

Figure 17 Salary through the lens of Department

Breaking down Salary as shown in figure 17 provides us with important additional context, notably median salary is relatively consistent throughout every department (around $10,000). Only marketing has a noticeable advantage in terms of salary within the group, marketing employees generally earn more than employees from other departments (closer to $15,000). In general, we can observe that HR employees tend to earn less while maintaining a more consistent and tight spread of salaries while IT, Marketing and Sales have larger salary ranges and a higher number of outliers that earn significantly more than the average. Sales in particular shows the greatest amount of variability with the most outliers, suggesting that salary structures may be tied to performance metrics (commission).

Another correlation shown to us in figure 14 was that of Salary and Experience (0.96).

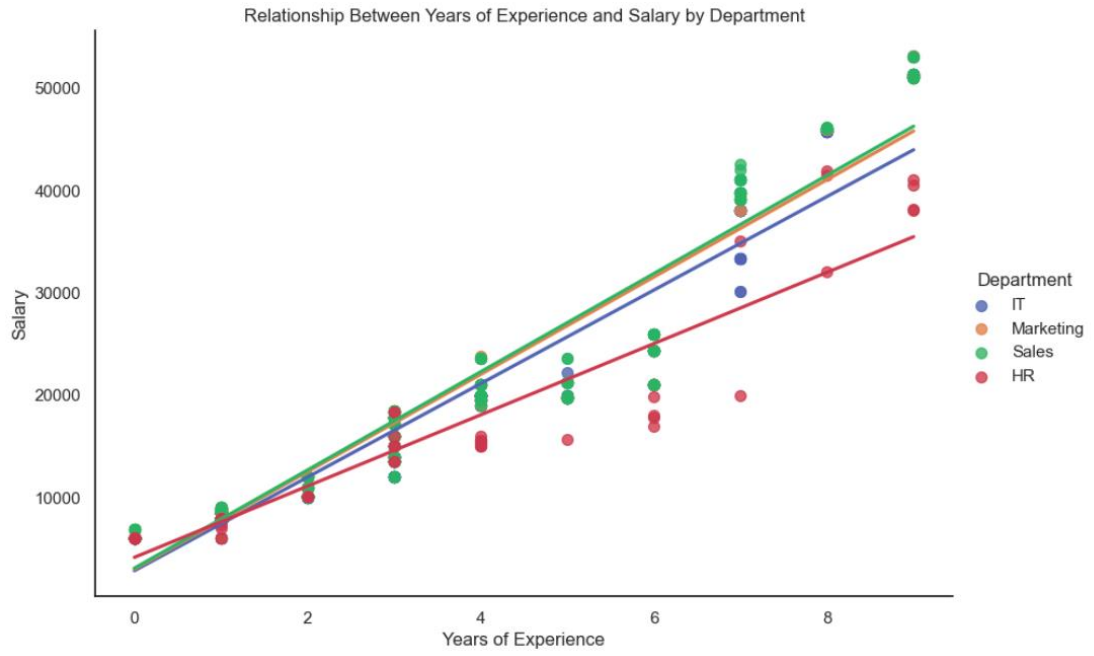

Figure 18 Salary based on Experience in each department

The high correlation between these two variables results in an almost linear graph (figure 18), where every increase in a unit of experience appears to result in an increase in salary. This is typical in most work environments as you are expected to have more knowledge as your experience increases, therefore your performance is also expected to increase. Looking at figure 18 we can observe that sales, marketing and IT have similar slopes indicating that experience influences salary similarly across these departments, employees working in these departments receive consistent pay increases as their experience increases. HR has a gentler slope suggesting the salary cap for employees in this department is lower. The linear nature of these trend lines makes it easy to predict expected salary based on years of experience and department, this suggests employees have specific pay bands for different experience levels. (Potential automatic pay increases).

The flatter slope seen in HR could be a result in underinvestment in this department. Could an increase in pay result in a better performance rating overall? Should salary be more of a response to performance than experience? **Do employees with more experience tend to perform better?**

Analysis of experience vs performance rating in departments

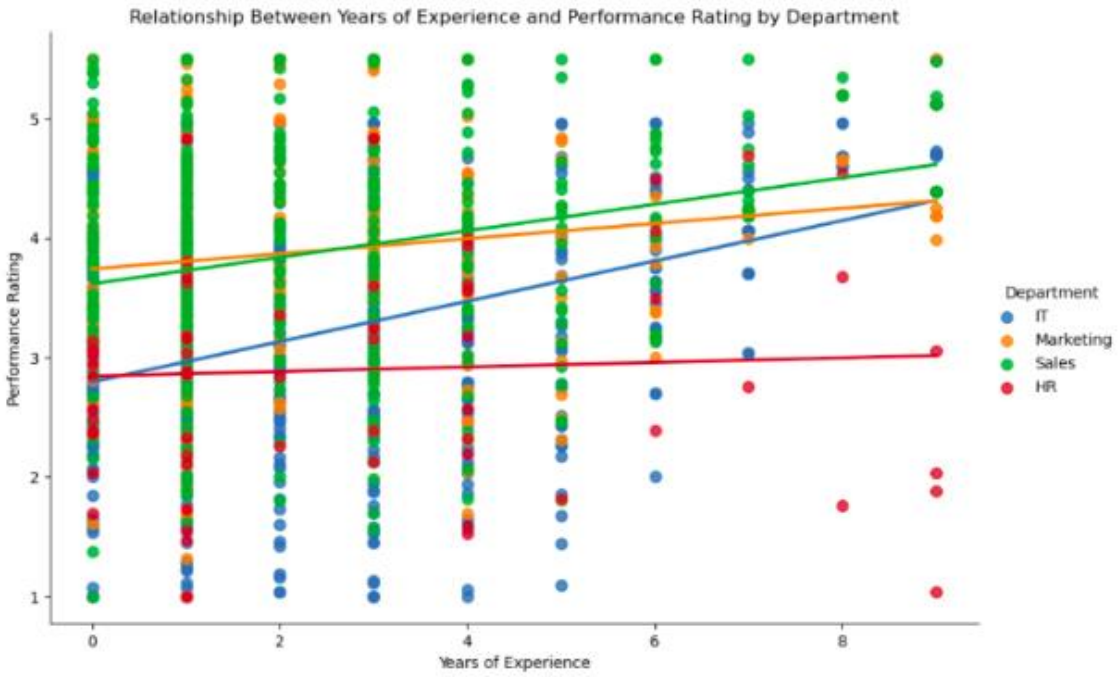

Figure 19 years of experience and performance rating for each department

This scatterplot on figure 19 shows the effect experience has on performance ratings. There are some key insights to be gained when comparing the two plots (18 & 19). First, we can see sales has a steeper slope compared to marketing while in the previous graph the slopes were very similar. This tells us that experience has a strong impact on both performance *and* salary in sales, indicating that employees in sales may have a pay structure tied to performance. It is common practice for salespeople to earn bonuses and commissions based on good performance, therefore, more experience in sales results in a significant increases in both variables.

IT can be seen on this graph with the steepest slope, telling us that as an IT professional's experience increases, their performance significantly improves. However, we can see from the last graph that experience does not seem to have as much of an effect on salary for IT. The relationship between experience and performance for IT is impressive, if compensation were more closely tied to performance for IT employees, maybe we would see an increasing trend in IT performance similar to that found in sales.

The most troublesome relation shown here is that within HR, displaying an almost flat line. Experience very slightly effects performance rating. This is concerning for us as HR seems to have very limited growth as a department, potentially leading to issues with

employee motivation and retention. **An interesting statistic would be to look at employee turnover for each department.**

Insights gained from multivariate analysis

The multivariate analysis of the EmployeePerformance dataset provides valuable insights into the key factors affecting employee performance and compensation. The analysis highlights that **training hours play a crucial role in improving performance**, with a strong positive correlation (R = 0.86). Departments like Sales and Marketing benefit from consistent performance ratings and higher salaries, suggesting effective management practices. However, HR stands out as a department with low performance ratings, limited salary growth, and minimal improvement with experience, indicating potential issues with employee motivation and retention.

The analysis also reveals a **disconnect between salary and performance in some departments**, such as IT, where performance improves significantly with experience, but salary growth does not match the trend. Furthermore, multicollinearity between salary and experience raises questions about whether the company is **rewarding tenure over performance**, and whether performance-based pay structures could be more effective in motivating employees across departments.

Future analyses could explore turnover rates and workload distribution to further understand the dynamics driving performance across the different departments.

By aligning **compensation structures more closely with performance**, especially in IT and HR, the company can promote a more motivated and high-performing workforce.

Assumptions and Hypothesis Formulation

Objective definition

The formulated objective of this analysis is to investigate potential variations in employee performance rating across departments. If variations exist, the objective will be to identify the departments that exhibit significantly higher or lower ratings.

Assumptions

Before conducting the analysis, there are few assumptions to be made. They are stated as below.

1) We are under the assumption that the data has been collected in a fair and controlled manner, the data is **independent** (one employee's information does not affect another).

2) The data follows a **normal distribution** allowing us to make inferences about the results of each test we conduct with confidence and clarity.

3) There are **linear relations** between variables allowing us to draw insights (increases in an independent variable can result in changes in the dependant variable).

4) The data does not contain **multicollinearity**, the independent variables do not show a high degree of correlation as this can affect the results of the analysis.

5) The data shows **homogeneity** of variance or an equal variance among the residuals. Uneven variances in samples can cause biased or skewed results.

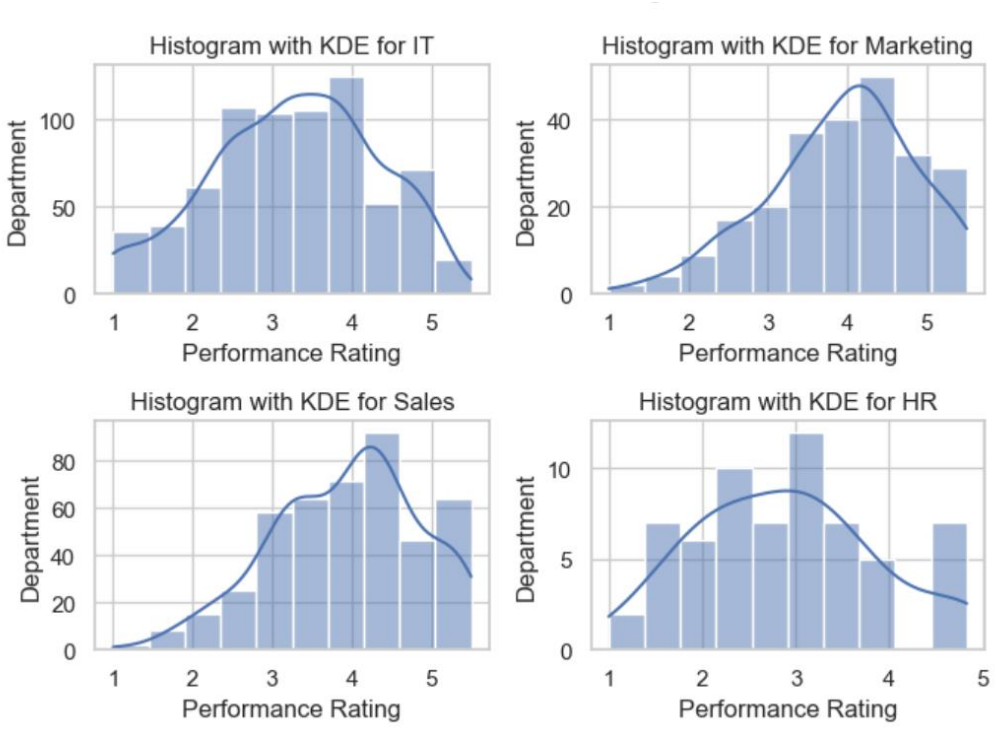

Figure 20 - histograms with KDE showing performance rates for each department

Figure 20 shows us the histograms with Kernel Density Estimates (KDE) for the four departments. Performance rating for IT department appears fairly normal with a bell-shaped KDE centred around the performance level of 4. The ratings are mostly around 2.5 and 5 with the peak at 4. Overall, department IT seems to have more employees with higher ratings and only a few receiving lower ratings.

Distribution of the marketing department is slightly skewed to the left with most employees with 4 and 5 performance ratings. Though there are few with lower ratings, the marketing department consists of employees performing at a high level.

Likewise, sales department is also left skewed with a majority of employees showing higher performance ratings. Most ratings are between 3 and 4.5 with plenty achieving 5.

HR department shows a less smooth distribution. Ratings range from 1 to 5 but less consistent that other departments. Diverse set of ratings with employees spread across different levels.

Hypothesis Formulation

After stating the assumptions, we are to create two hypotheses that align with our analysis objective. They are the null hypothesis and the alternative hypothesis. The null hypothesis of a test states no effect or no relationship between variables and the alternative hypothesis states an effect or relationship within the variables.

Null Hypothesis (H0): There is no significant difference in performance rating between different departments (all group means are equal).

Alternative Hypothesis (HA): There is a significant difference in performance rating between different departments. (the alternative hypothesis or the research hypothesis contradicts the null hypothesis, suggesting that there is at least one group mean that is statistically different from the rest).

Statistical Technique: Hypothesis Testing

Explanation of statistical method

Using the assumptions previously mentioned we used one-way ANOVA to test our hypotheses. This is a statistical method used to compare the means between and among three or more groups and see if they are statistically significant. Using ANOVA, we can decide if the differences observed are genuine differences or due to random chance.

The ANOVA test generates an F-statistic which is calculated by dividing two mean squares. This value is a measure of how much the group means differ relative to the variability within each group.

ANOVA will give you a P-value and F-statistic as a result,

The P-Value is a probability measure telling us how likely it is to observe the differences in the data if the null hypothesis was actually true. If the p-value is lower than the chosen critical value (0.05 in this case), it tells you that the result of the analysis is unlikely to have occurred by chance and so you may reject the null hypothesis. If the P-value is higher than the chosen critical value/threshold it tells you that the result of the analysis is likely to have occurred by chance and so you cannot confidently reject the null hypothesis.

The F-statistic is a ratio of the variance between the groups to the variance within the groups. Simply put, F = variance between groups / variance within groups. This is a helpful measure for us as we can use this output to measure the significance of the variance. A low amount of variation between groups will result in a number close to 1, whereas a high amount of variance will result in a higher number. The higher the number, the more significant you can say your result is.

A limitation of ANOVA is that it does not tell you which group mean is statistically different compared to the rest.

Hypothesis testing

We will now conduct the ANOVA test for our dataset.

```
One way ANOVA results
F-Statistic: 61.45
Critical F-Value: 2.61
P-Value: 0.00
We reject the null hypothesis, there is a significant difference in performance rating across the different departments
```

Figure 21 - test results of the ANOVA test

Figure 21 gives us the test results of ANOVA. As you can see, the F-statistic is 61.45, this is greater than the critical F-value of 2.61, which means that the variation between the groups is considerably larger than the variation within the groups. The p-value is 0.00, which is less than the significance level (0.05). This means that there is strong evidence against the null hypothesis.

We reject the null hypothesis, confirming there is at least one group mean that is statistically different to the rest.

Post HOC test

Tukey's Honestly Significance Difference Test (HSD) is used after ANOVA analysis to find out which specific groups have a significant variance in means. We use Tukey's HSD test in this case because the ANOVA test has confirmed our rejection of the null hypothesis, but we still do not know which groups differ. Tukey's HSD tests every possible pair of groups to see which difference in mean is statistically significant.

Tukey's HSD **is applicable** to our data because the result of our ANOVA tells us that there is a significant difference in means in the groups we tested, however it tells us nothing about which specific group means are different. Conducting Tukey's test will show us the specific group pairs where the null hypothesis can confidently be rejected. Tukey's test results in several important statistical measures about each pair of groups.

Meandiff

- This is the calculated difference in the mean values of the group pairs (group 2 – group 1). A positive meandiff value suggests that the second group in the pair has a higher mean value than the first group in the pair. A negative mean diff value suggests that the first group in the pair has a higher mean than the second. A meandiff value close to 0 suggests that the means are extremely close and may not be significantly different in value, whereas a meandiff value further away from 0 suggests there is a more significant difference in the mean values of the groups.

p-adj

- This is the p-value that has been adjusted for multiple comparisons, reducing the chance of false positives. The p-value was discussed for the ANOVA test and the same explanation applies here only to the specific group pairs. A p-value less

than the threshold amount suggests that the variation seen is unlikely to have been observed by chance, giving us evidence to reject the null hypothesis, whereas a p-value higher than the threshold amount suggests that the result is likely to have occurred by chance in nature so we cannot confidently reject the null hypothesis.

Confidence intervals - lower and upper

- The test has a 95% confidence interval for the difference between the means. The interval shown (lower - upper) shows the range of values that is likely to contain the true mean difference between two groups.
- If the confidence interval does not include 0 it means that the result is statistically significant (**reject** the null hypothesis)
- If the confidence interval does contain 0, it suggests the result is not statistically significant. (**do not reject** the null hypothesis)

Reject

- Gives a true or false result. True if we are to reject the null hypothesis, false if we cannot reject the null hypothesis.

Tukey's HSD results

```
Tukeys HSD post hoc test:
   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
-----------------------------------------------------
```

Figure 22 - results of the Tukey's HSD post hoc test

As you can see from figure 22, the Post HOC test shows that only Marketing and Sales do not have a significant difference in mean performance rating score. Every other department pair is shown to have a significant difference in mean performance rating. Below is a detailed explanation of the results.

- **HR vs IT –** The mean difference is 0.3715 with a p-value if 0.0217. The p-value is less than 0.05 and the confidence interval does not include zero (0.0384 –

0.7047), therefore, we can conclude that there is a statistically significant difference between the mean performance ratings of HR and IT departments. Since the mean difference is positive, we can confirm that IT department has higher performance ratings compared to HR.

- **HR vs marketing –** The mean difference is 1.027 with a p-value of 0.0. This is a very strong statistical difference, as the p-value is zero and the mean difference is comparatively high. The confidence interval (0.6681 – 1.386) doesn't include zero which means that the mean difference between the HR and marketing is statistically significant. The positive mean difference confirms that marketing has higher performance ratings than HR.
- **HR vs sales –** The mean difference is 1.0256 with a p-value if 0.0 and the confidence interval is from 0.6843 to 1.3669. The positive mean difference tells us that Sales has higher performance ratings compared to HR. The p-value and the confidence interval strongly confirm that the mean difference is statistically significant.
- **IT vs marketing –** The p-value is 0.0 and the confidence interval is 0.4665 to 0.8445 which does not include zero, this proves that the mean difference between the two departments is statistically significant. The mean difference is 0.6555 which is not as high as when compared to previous pairs, but this shows that marketing has higher performance ratings compared to IT.
- **IT vs sales –** The p-value is 0.0 and the confidence interval is 0.5012 – 0.807 with no overlap with zero. The mean difference shows that sales has significantly higher performance ratings compared to IT and the mean difference between these two departments is statistically significant.
- **Marketing vs sales –** The mean difference between these two departments is -**0.0014** and the p-value is 1.0. The confidence interval is -0.2044 to 0.2017 and this includes zero which means that this is not statistically significant. We can confirm that there is no significant difference between the mean performance ratings in marketing and sales departments.

Discussion and Conclusion

The analysis conducted on the KiwiLearn's employee performance ratings across different departments has given important insights into how performance varies within the organization. The result of the ANOVA tells us that there is a significant difference between the mean performance levels in each department and the post HOC test tells us that there are noticeable differences between HR and IT, HR and marketing, HR and sales, IT and marketing, IT and sales.

- The HR department has significantly lower performance ratings compared to all other departments (IT, Marketing and Sales).
- Marketing and Sales stand out as having the highest mean performance rating with no significant difference between these two departments
- IT maintains a middle of the road position in all analysis, with a better rating than HR but not as good of a rating as sales or marketing

Potential reasons for differences

Leadership & management structure

- An aspect we have talked about through our analysis is that of leadership within the departments. Leadership is crucial in a business and a poor management strategy can result in low output and sub optimal performance. Looking at the results, it can be said that sales and marketing may have stronger management systems in place. A more robust management system means employees are more likely to perform to a higher standard, this is shown in the lack of variability in the performance ratings of these two departments, especially when considering their size.
- Alternatively, we can look to HR for a potential example of poor management and leadership. HR is the smallest department and should therefore be the easiest to manage. The results of our testing show that HR is consistently underperforming in comparison to the other larger departments. The fact that this department has made no effort to change their situation, tells us that there could be a potential management challenge.

Differences in role complexity

- Sales and marketing may have easy to define and measure work goals, resulting in clearer pathways set towards higher achievement.
- HR and IT however consist of more complex problem-solving tasks and workloads with more difficult to measure goal practices. This not only makes creating an easy to navigate path towards achievement more difficult, but it also means the performance metrics used may not be applicable to these departments as their role is completely different.
- i.e. if a performance metric for sales is 'customer satisfaction', this same metric cannot be used for IT or HR.

Incentive structure

- Sales associates may be getting paid in commission or receive bonuses based on monthly sales. This would incentivise sales employees to perform better as they would see an almost immediate financial gain from high performance.

- A lack in performance linked rewards in departments such as IT or HR could result in lower motivation and consequently, lower performance ratings.

Actionable insights

The results from this analysis tell us that Sales and Marketing are performing well compared to the other departments. KiwiLearn should use these high performing departments as benchmarks when setting standards for the other departments. An investigation should be conducted into all aspects of each department to find out what Sales and Marketing are doing differently to be achieving at such a high rate. Specifically, we would like to offer a number of actionable insights for KiwiLearn to conduct in order to help increase the performance of lagging departments.

Leadership development

- HR is underperforming while Sales and Marketing are performing well. We suggest the implementation of mentorship programs where leaders from high performing departments (sales and marketing) run workshops or tutorials for the leaders in other departments where they can share successful management strategies. This would help the leaders of all departments develop stronger management skills and enable them to set a higher standard for their respective departments. An improvement in leadership will create clearer expectations, improve employee engagement and create a more performance-oriented culture within the workplace.

Align compensation with performance metrics

- The introduction of performance-based incentives (rewards for high performing employees in each department) would result in employees being more motivated to set and achieve goals based on specific and measurable KPIs. A quarterly review of performance would mean employees have a specific timeframe after which they can expect to have their performance analysed by the organization. Knowing that they can influence their bonuses or salaries based on how they perform each quarter would result in employees putting more effort into their daily work.

Increase in training for all departments

- Our prior analysis showed that Training Hours have a significant effect on performance rating. An increase in investment in the yearly training for HR and IT would mean that employees have more opportunity to grow within the company and see a yearly development in their skills and career. Department specific training budgets would ensure that all departments are receiving the same

opportunity for professional development, boosting confidence and enhancing performance across the whole company.

Conclusion

The results from our one-way ANOVA analysis confidently conveyed a significant difference in mean performance ratings across the different departments.

The F-statistic value was 61.45, far more than the calculated critical F-value of 2.61. The P-value was 0.00, lower than the threshold of 0.05 (5%). These outputs gave us reason to conduct an HSD test on our departments to find out which specific departments have a significant difference in mean performance rating.

The result of our HSD stated that all departments are significantly different in mean performance rating when compared as pairs, other than Sales and Marketing. This result coincides with the visualizations shown earlier in the report. Sales and marketing have been shown to share a very similar mean performance rating (figures 13 & 16).

Examining the meandiff column in the HSD test (figure 21) we are drawn to the pairwise comparisons between HR & Marketing, and HR & Sales. The mean difference for each of these comparisons is 1.027 and 1.0256 respectively, affirming the idea that these departments are on opposite ends of the performance rating scale. HR has consistently underperformed in our analysis, while sales and marketing have dominated the comparisons. IT is the largest department and yet performed much better than the small HR department (HR & IT meandiff = 0.3715, IT has a higher mean performance rating). Comparing IT to one of the stronger departments (marketing) we see a mean difference of 0.6555, this although significant, is almost half the difference between HR and marketing. Proving again, that HR as a department has exceptionally low ratings.

Part Two: Regression Analysis

Identify Potential Predictor Variables

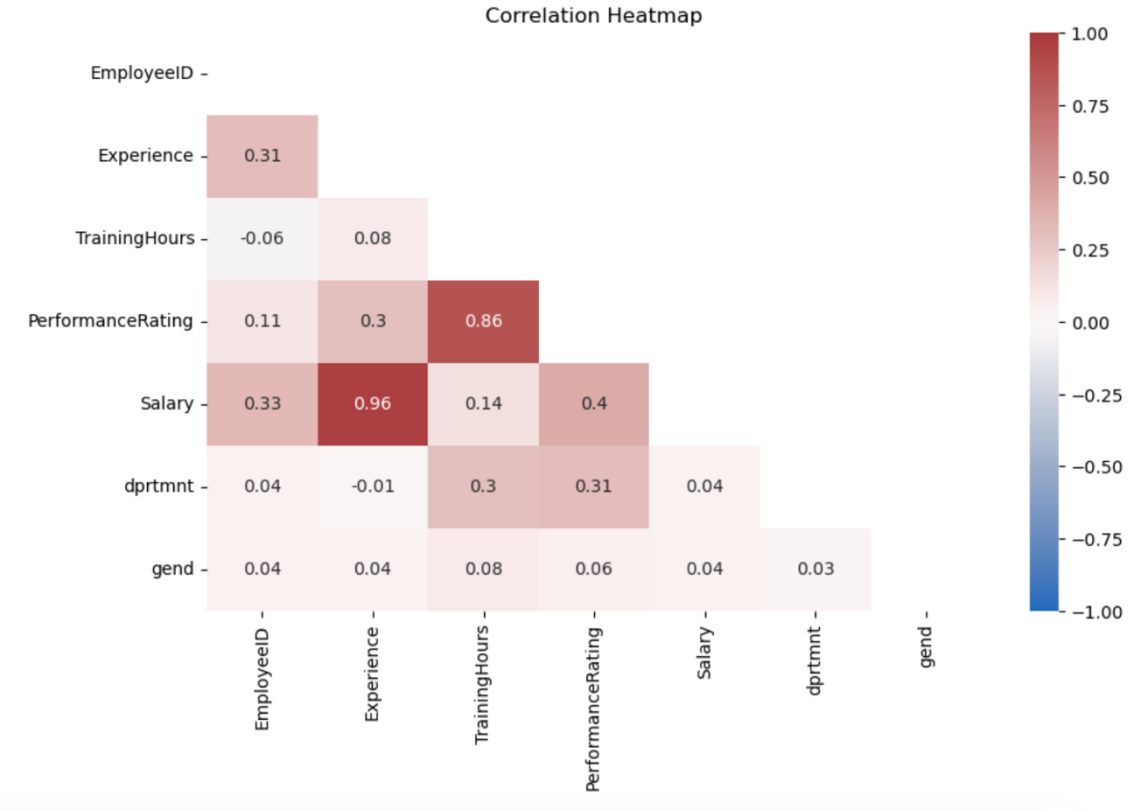

Figure 23 re-analysis of the Correlation Heatmap

Re-analysis of the original correlation matrix is required to choose independent variables to use in the forthcoming regression analysis. This, combined with our rich understanding of the dataset will help us to identify variables that are meaningfully correlated with the dependant variable. A variable showing high correlation with performance rating indicates to us that it would be a good predictor. Our prior analysis and correlation matrix provide us with a few promising candidates as we move towards completing a regression analysis. These variables are listed below with their corresponding correlation coefficient with PerformanceRating:

1. **Years of experience – 0.3**

Employees with more experience might perform better as they are more skilled and familiar with their roles. The correlation is only moderate suggesting that experience alone is not enough to predict performance rates.

2. **Training hours – 0.86**

The high correlation suggests that training plays a significant role in increasing employee performance. Employees with more training are likely to have better skills, knowledge and motivation to excel in their careers.

3. **Salary – 0.4**

The salary variable can potentially effect employee performance in two ways, higher salaries could motivate employees to perform better, or high performers may be rewarded with more pay. Salary only has a moderate correlation shown; however, it makes an interesting variable for exploration.

4. **Department – 0.31**

Different departments may have different expectations or challenges within them, influencing employee performance in varying ways.

We chose these as predictor variables because they showed the highest degree of correlation with PerformanceRating and seemed to make logical sense when considering the real-world implications of the data.

Before we can fit the regression model we must make a couple of assumptions about our data.

Assumptions for Regression Analysis and the relevance to our analysis

Before conducting a regression analysis, it is important for us to make and verify assumptions about our data to ensure the results of the analysis are valid and reliable enough to interpret. The assumptions we make are

1) **Multicollinearity**

- The predictor variables should not display a high degree of correlation with one another. This is to isolate the individual effect of each independent variable on the dependant variable. If multicollinearity is found within the predictor variables and not dealt with before applying the regression model, it may result in:
    1. **Unstable coefficients** - small changes in the data causing large changes in the coefficients.
    2. **Inflated standard errors** – if p-values increase it may lead to us wrongly concluding that a predictor is not significant when it actually is.

We can check for multicollinearity using a correlation heatmap.

2) **Linear Relationships**

- The relationship between the dependant variable and the independent variable should be linear. A non-linear relationship may result in:
    1. **Underfitting** – If a linear model is used for non-linear data, the true pattern of the data will be missed, leading to underfitting. Underfitting in simple terms is when the model is too simple to explain the relationships in the data.
    2. **Biased coefficients** – The linear model will try to fit a straight line through data that might instead follow a curve or other pattern. This would result in incorrect predictions and insights.

We can check for linearity using scatterplots with regression lines.

3) Independence of observations

- Observations are expected to have been collected using statistically valid sampling methods ensuring there are no hidden relationships among variables. Dependant data can result in:
    1. **Incorrect conclusions** – We may find significance where there is none or miss significance where it exists.

4) Normality

- We assume the data follows a normal distribution. This ensures that coefficients are unbiased, and p-values and confidence intervals are reliable. If normality is violated It may result in:
    1. **Inaccurate statistical tests** – p-values and confidence intervals will be misleading, leading to us making incorrect conclusions about the significance of predictors.

We can check for normality using Q-Q plots and AD tests.

5) Homoscedasticity

- The variance of the residuals should remain constant across all independent variables i.e., the spread of the residuals should be the same for all predicted variables. If homoscedasticity is violated it may result in:
    1. **Inefficient and biased coefficients** – when heteroscedasticity is present the variance of the regression coefficient estimates but the model doesn't show this, simply put the model might make a term statistically significant when in reality it is not. Therefore, the results of the analysis become unreliable.

We can test for homoscedasticity by creating a fitted value vs residual plot.

Assumption Testing

Out of the five assumptions mentioned above two are to be tested before performing regression and they are **linearity** and **multicollinearity**.

The reason why we should test linearity before performing regression is that it makes sure that our model is appropriate and reliable. The coefficients generated from the model are reliable and the model provides a good fit for the data, which means that the predicted values will be close to the observed values. If we fail to test this before regression, then it will produce predictions that are far from the actual data points. Overall, testing for linearity will reduce the risk of having misleading results. It also gives us an idea of whether we should do a transformation or use a non-linear model.

The reason why we should test for multicollinearity before regression is because it helps to ensure the accuracy, consistency and the accountability of the model. Even small changes in the data can cause huge changes in the estimated coefficients if multicollinearity exists. It also inflates the standard errors of the coefficients, which means even a genuine relationship between the dependent and the independent variable will be shown as statistically insignificant. Specifically, it's impossible to interpret the model because the model cannot depict the unique effect of one variable while controlling other variables.

Multicollinearity check

To effectively check for multicollinearity, we must create a separate data frame containing only our chosen predictor variables and a correlation heatmap for them. Variables displaying high correlation must be removed or transformed.

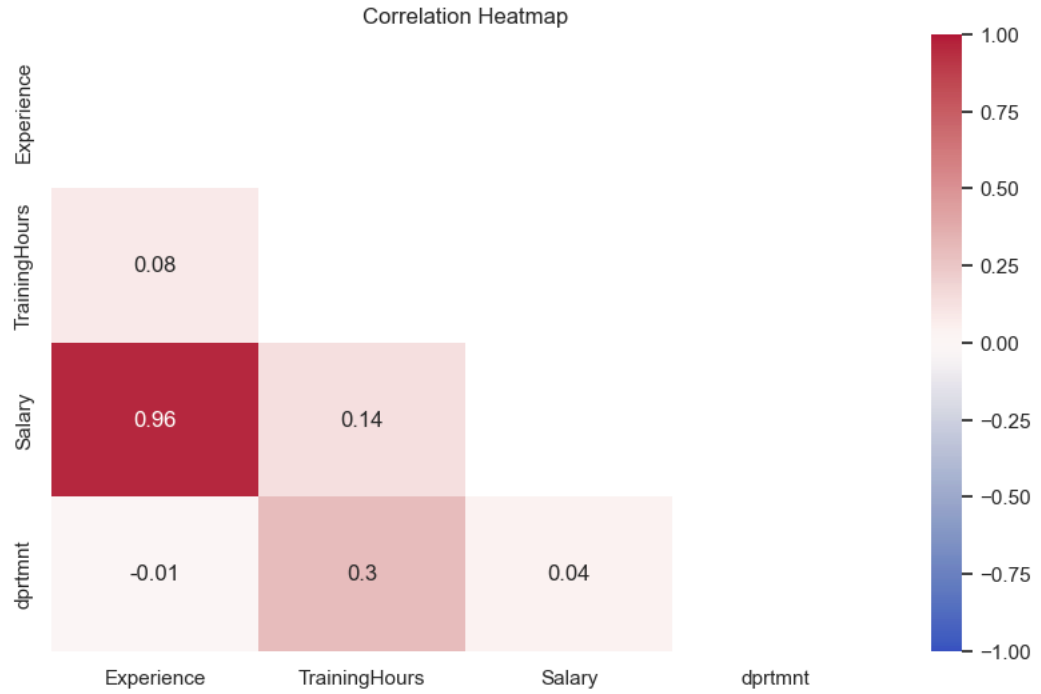

Figure 24 Correlation Heatmap for Predictor Variables

Generally, we would only consider variables with correlation coefficients greater than 0.7 as potential multicollinear variables. Looking at figure 24, shows that only one pair of variables is showing a degree of correlation above the threshold (0.7), Experience and Salary. We know we must properly deal with this pair before applying the linear regression model. Before deciding on further actions, we will check for linearity among the predictor variables.

Linearity check

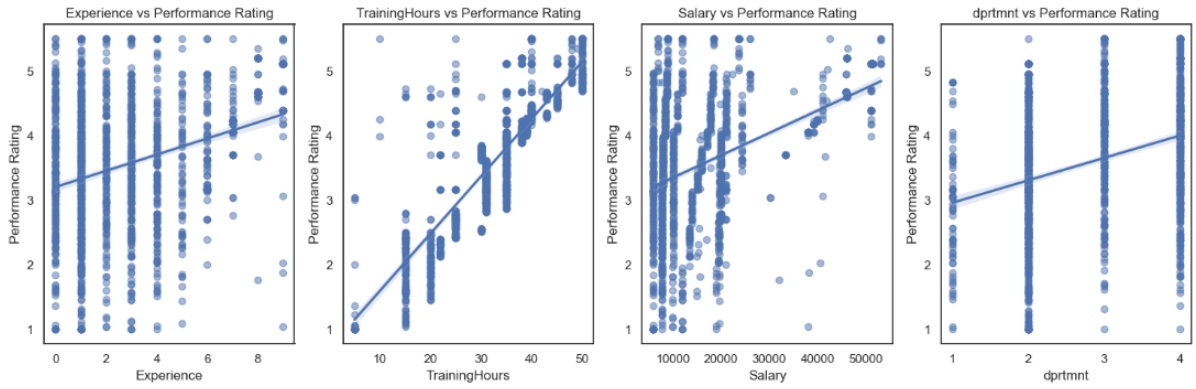

Figure 25 Scatter plots showing linearity between predictor variables and Performance Rate

Inspecting the scatterplots on figure 25 with regression lines showing our chosen predictor variables relationship with the dependant variable can help us determine

which specific variables should be included in the regression analysis. We will combine results from both the previous correlation heatmap and these linearity plots to decide on which variables to remove from the final analysis.

Experience

- Displays a positive trend, however the slope is relatively flat, suggesting performance rating does increase with experience but not significantly. This indicates experience may not be a strong predictor.
- Previous analysis has shown a high degree of correlation between experience and salary, the weak relationship between shown experience and performance suggests the removal of experience as a predictor.

TrainingHours

- Displays a strong positive linear relationship, as TrainingHours increase, performance ratings increase. We see low variability in the data points against the trend line indicating an obvious pattern. This is a very good predictor for performance and should definitely be used in the final model.
- Training hours had no significant correlation with any other independent variable.

Salary

- Displays a significant positive trend, however the data points are more scattered compared to training hours. Does not display a perfectly linear relationship, suggesting salary alone cannot fully predict performance. However, the effect is undeniable and should be considered for the final analysis.
- The correlation between salary and experience means we cannot use both of these variables; salary has a stronger relationship with performance than experience does.

Department

- Displays the weakest positive trend, not a good predictor variable for our regression analysis. We know already that sales and marketing have better performance than HR and IT, we want to use other variables to understand why.

Conclusions

We have chosen two predictor variables based on the preceding tests. Our final predictor variables are **TrainingHours** and **Salary**. These variables show the most significant correlation with performance and have the strongest linear relationship with

our performance variable, making them the most appropriate to use in the regression model.

Regression Analysis

Figure 26 gives us the results from our OLS regression model.

OLS Regression Results

Dep. Variable:	PerformanceRating	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.823
Method:	Least Squares	F-statistic:	3412.
Date:	Tue, 15 Oct 2024	Prob (F-statistic):	0.00
Time:	12:32:09	Log-Likelihood:	-874.98
No. Observations:	1468	AIC:	1756.
Df Residuals:	1465	BIC:	1772.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4306	0.040	10.861	0.000	0.353	0.508
TrainingHours	0.0848	0.001	73.956	0.000	0.083	0.087
Salary	2.52e-05	9.53e-07	26.450	0.000	2.33e-05	2.71e-05

Omnibus:	308.107	Durbin-Watson:	1.826
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1153.069
Skew:	0.982	Prob(JB):	4.11e-251
Kurtosis:	6.872	Cond. No.	6.97e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.97e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 26 OLS regression results table

R-Squared – 0.823

The regression model explains 82.3% of the variance in performance rating, indicating a strong fit. This means that the independent variables have a significant impact on performance.

Adj. R-Squared – 0.823

The adjusted R-squared value accounts for the number of predictors in the model, a big difference in R-Squared and Adj R-Squared would indicate overfitting.

F-Statistic - 3412

Tests the overall significance of the model, a high F-Stat in conjunction with a low p-value (0.00 in this case), suggests the model is statistically significant.

Const coefficient – 0.4306

The predicted baseline performance rating when both TrainingHours and salary are 0.

TrainingHours coefficient – 0.0848

For each additional hour of training, performance rating increases by 0.0848 points, meaning that with an additional 12 hours of training, employees can increase their performance rating by 1 point! With the p-value of 0.00 the effect is said to be statistically significant.

Salary coefficient – $2.52 * 10^{-5}$

For every dollar increase in salary, the performance rate increases by 0.0000252 points telling us that the effect of salary on performance in positive but small. The p-value of 0.00 suggests that salary is a statistically significant predictor of performance.

Omnibus - 308.107

The omnibus test of normality evaluates whether residuals from the regression model are normally distributed through the use of skewness and kurtosis evaluation. If the p-value for the omnibus test is less than 0.05 it is an indication that the residuals deviate from normality. Our results show a high omnibus reading with a p-value of 0.00, indicating the residuals may not be normally distributed. This is a clear violation of the assumption of normality for the OLS regression model, resulting in unreliability of the statistical inferences.

Jarque – Berra (JB) – 1153.069

Another test of normality focussing on skew and kurtosis. Again, the p-value for this test is less than the threshold (0.05), strongly indicating that the residuals do not follow a normal distribution.

Condition Number – $6.97 * 10^{4}$

The condition number measures the sensitivity of the regression models estimates to small changes in the data, a large condition number can suggest that the independent variables are highly correlated or not well scaled. Our condition number is extremely high suggesting a severe issue in the model. Because we know the independent variables do not show a high degree of multicollinearity in this case, we can assume the high condition number is a result of the scaling issue. Salary is measured in thousands while training hours in tens, leading to an ill conditioned matrix.

Conclusion:

While the results of the OLS model tell us the independent variables are statistically significant in predicting performance rating, the issues raised by the results of the omnibus and JB tests suggest that we cannot trust the inferences made due to non-normality of the residuals. Furthermore, the condition number suggests our

independent variables are poorly scaled, making the regression output unreliable. We must test the remaining assumptions of the regression analysis to confirm whether our residuals are normally distributed or not.

Assumptions of Linear Regression

Normality check

To test for normality in the residuals we use a quantile-quantile (Q-Q) plot. This is a graphical tool used to assess whether a set of data follows a normal distribution.

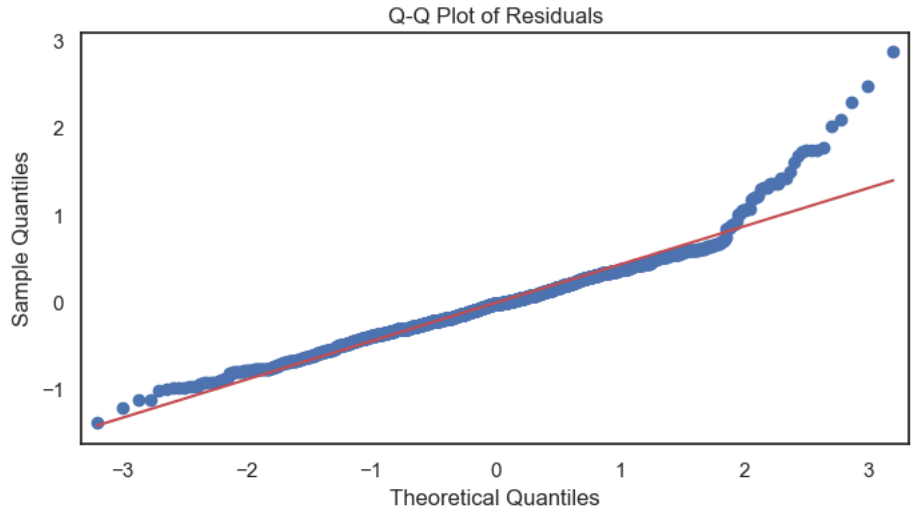

Figure 27 Q-Q plot and AD test results

The red line in the graph (figure 27) represents normal distribution, we can see the tails show significant deviation representing excess kurtosis. The middle section of the plot shows promising results however the tails indicate more extreme residuals than expected in a normal distribution.

Another test we ran to confirm the non-normality of distribution is the Anderson Darling (AD) test. The AD test compares the empirical cumulative distribution function of the sample data to the theoretical cumulative distribution function of a normal distribution. If the sample data matches a normal distribution closely, the test statistic will be smaller. The AD test returns a number of critical values at different significant levels, if the test statistic exceeds the critical value at your chosen significance level, you must reject the null hypothesis (H0: the data is normally distributed). In our case the AD test statistic is 7.9, this is much higher than the critical value of 0.785 at a significance level

of 0.05. We reject the null hypothesis and can confirm that our data is not normally distributed.

Homoscedasticity check

Homoscedasticity implies that the variance of the residuals should be constant across all levels of the independent variables. In order to accurately check for homoscedasticity, we must create a plot of the residuals against the predicted values from our regression model. If the plot results in a pattern that fans out or narrows as predicted values change, it can be an indication of heteroscedasticity or non-constant variance. There should be no discernible pattern in the following plot.

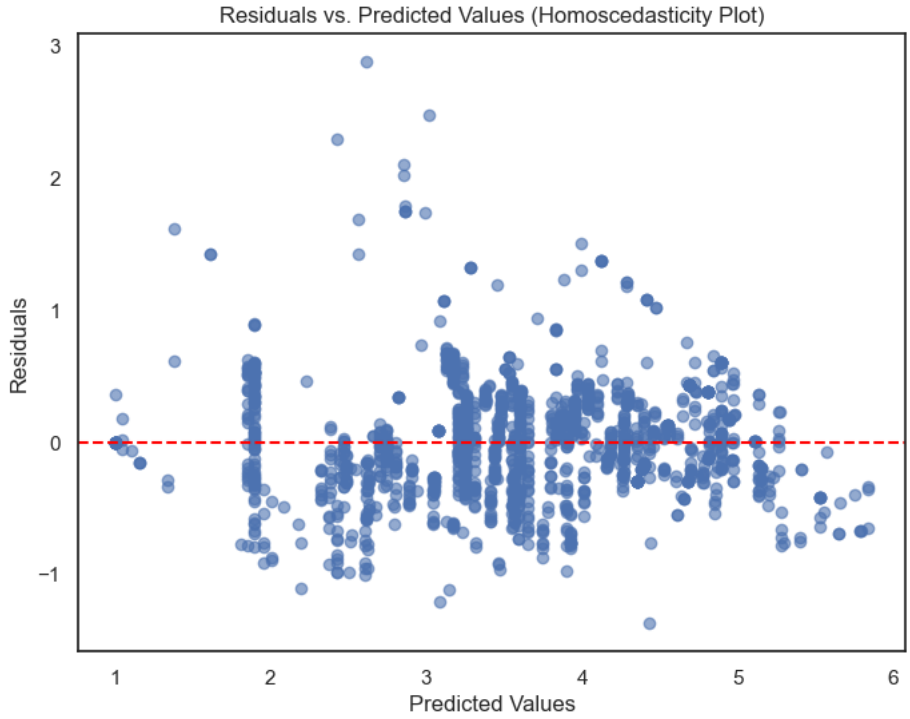

Figure 28 plot of residuals vs predicted values (Homoscedasticity test)

The plot on figure 28 consists of a visible pattern, the residuals appear to become narrower as the predicted values increase - a mild fan shape – which is a sign of heteroscedasticity.

Strategies to address violations

Our data is showing signs of both non-normality and heteroscedasticity, these are violations of the assumptions made before conducting the OLS regression analysis meaning the results of said analysis are invalid and unreliable. There are many ways to address these violations that would make our results more reliable.

Transform the variables

Applying a transformation to variables can help stabilize variance – reducing heteroscedasticity – and normalize the residuals. This method would address both violations from our data. There are many possible transformations however the one most suited to our case would be a **Box-Cox transformation**. These types of transformations are especially effective to **stabilize variance**. The Box-Cox transformation finds the best transformation parameter to make the data as close to a **normal distribution** as possible, meaning all possible transformation methods are considered in a Box-Cox transformation. Using this method, we can greatly improve the accuracy and reliability of our OLS model.

Use a more robust regression analysis

Using another regression analysis method such a **quantile regression** is common practice when conditions of linear regression are not met. Quantile regression estimates are more robust against outliers as different measures of central tendency can be used to comprehensively analyse the relationship between variables. Instead of focusing on mean, we can use median or IQR. Quantile regression has been successfully used in scenarios with complex interactions between different factors, we can see this in our own data with the difference in scale found between salary and training hours. Quantile regression can easily handle heteroscedasticity. Furthermore, this regression analysis method shows **how** predictors influence the distribution, while OLS method only shows the **average** effect. Opening the analysis up to other methods could provide the business with better results.

Discussion and Conclusion

Objective and key findings

In this report we explored the key factors influencing **employee performance** at KiwiLearn and identified significant variations across the different departments. Through our analysis we found that **TrainingHours** had the most substantial impact on performance, we saw departments with more training tended to perform better overall suggesting that **increasing training could boost performance** especially in underperforming departments such as HR and IT. Sales and Marketing showed the highest performance ratings with no significant difference between them. HR and IT showed lower mean and median performance with HR performing the worst overall. We also noticed salary's influence on performance, although this was less noticeable than training hours. The strong correlation between **salary and experience** raises concerns

about whether the company is rewarding tenure over performance, highlighting the need for **performance-based pay structures.**

Limitations

Despite valuable insights gained, the analysis faced many challenges. Importantly the data violated some of the assumptions for linear regression. **Multicollinearity** between experience and salary resulted in us not being able to use both variables in the final regression model, meaning we may have omitted valuable insights from our analysis. **Non-normality** of the residuals and **heteroscedasticity** indicate that the results from the regression model should be interpreted with plenty of caution. **Lack of adequate data** meant that we could not get the full picture of the factors influencing performance. Future research should include variables such as: workload, leadership quality, turnover rates, employee satisfaction and engagement metrics to better understand the variations in performance.

Suggestions for future areas of research

Although the analysis exposed areas of improvement within the organization, further research would be beneficial for the business to enhance the reliability and depth of the insights gained. Some ways to improve the analysis include:

Incorporation of additional variables

- Training hours, salary and experience have been shown to affect performance in employees, however we can be certain there are many more factors at play. Including other variables can provide a more holistic picture of the inner workings of the organization. Data should be collected on workload, job satisfaction, leadership quality and employee motivation levels for example, to create a better model on the spread of performance rating.
- An interesting study would be into the annual turnover rates within the organization, which departments are consistently providing value for their employees increasing employee retention. This could assist in the identification of management issues through the company.

Time series analysis

- Performing a timeseries analysis of performance ratings through different periods of the year could provide interesting insights regarding seasonal effects on workload and performance. A cross-sectional snapshot of the business may miss important dynamics and trends. Longitudinal data captured over time would allow us to see high and low periods and trends in each department.

Conduct department specific investigations

- Each department has different tasks, goals and success markers, which could affect performance in different ways. Conducting department specific analysis would account for the nuances found within each business group. Specifically, a focus on leadership styles, role complexity and incentive/bonus pay structure.
- Conduct anonymous surveys to gain personal data from employees without fear of repercussions.

Apply different regression models

- The data used violated some of the assumptions necessary for linear regression, further analysis should apply a different regression model such as quantile regression which is more robust against these statistical violations.
- Further analysis could also consider transforming some of the variables in order to better fit the assumptions necessary for different regression models.

Our analysis uncovered the variations in employee performance across different departments, we have discussed the factors at play and come to a series of actionable insights for the organization. By implementing mentorship programs to increase the standard of leadership and increasing funding for training in low performing departments along with more closely aligning compensation with performance company wide, KiwiLearn can foster a high performing culture across all department groups.