**Assignment 2**

**Semester 2 2024**

## PAPER NAME: Data Analysis

**PAPER CODE:** COMP517

|  |  |
| --- | --- |
| **Student ID** | **Student Names** |
|  |  |
|  |  |
|  |  |

**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

1. **Please email** [**DCT.EXAM@AUT.AC.NZ**](mailto:DCT.EXAM@AUT.AC.NZ) **if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
2. **Attach your code for all the datasets in the appendix section**.

**Contents**

[**PAPER NAME: Data Analysis 1**](#_Toc179977300)

[**Part One: Exploring Data and Testing Hypotheses: Uncovering Insights from Dataset 3**](#_Toc179977301)

[**Task 1: Data Preparation and Exploration 3**](#_Toc179977302)

[**Task 2: Assumptions and Hypothesis Formulation 3**](#_Toc179977303)

[**Task 3: Statistical Technique: Hypothesis Testing 4**](#_Toc179977304)

[**Task 4: Discussion and Conclusion 5**](#_Toc179977305)

[**Part Two: Regression Analysis 6**](#_Toc179977306)

[**Task 1: Identify Potential Predictor Variables 6**](#_Toc179977307)

[**Task 2: Assumptions for Regression Analysis 7**](#_Toc179977308)

[**Task 4: Assumptions of Linear Regression 10**](#_Toc179977309)

[**Task 5: Discussion and Conclusion 12**](#_Toc179977310)

**List of Figures**

[**Figure 1: Tukey’s Test: Mean Performance Ratings by Department 5**](#_Toc179977319)

[**Figure 2 : Residuals vs. Fitted Values (Linearity Check) 8**](#_Toc179977320)

[**Figure 3 : Q-Q Plot (Normality of Residuals) 8**](#_Toc179977321)

[**Figure 4: Residuals vs. Fitted Values (Homoscedasticity Check) 10**](#_Toc179977322)

[**Figure 5:Q-Q Plot of Residuals (Normality Check) 11**](#_Toc179977323)

# Part One: Exploring Data and Testing Hypotheses: Uncovering Insights from Dataset

## Task 1: Data Preparation and Exploration

 **Initial Data Exploration:**

* The dataset has 1,468 entries with 7 columns: There are EmployeeID, Department, Gender, Experience, Training Hours, Performance Rating and Salary has been considered under this analysis..
* There are no missing values in our dataset, the data types used in the dataset are integers, floats and objects.
* Employees have 0-9 years of experience (average: 2.8 years), performance ratings range from 1.0 to 5.5 (mean: 3.56), and salaries vary from $6,000 to $53,100 (mean: $16,107).

 **Distribution of Performance Ratings:**

* Most ratings are between 2.8 and 4.3, with fewer extremes.

 **Distribution of Experience:**

* Most employees have less than 5 years of experience, with a sharp decline in higher experience levels.

 **Multivariate Analysis:**

* Scatter plot shows no clear trend between experience and performance rating; IT department shows higher mid-level to senior experience concentration.
* No strong correlations between experience and performance ratings across departments.

 **Boxplot of Performance Ratings by Department:**

* Marketing and Sales have wider performance rating ranges, while IT has higher medians and HR has lower medians, indicating departmental performance variations.

## Task 2: Assumptions and Hypothesis Formulation

**Objective of the Analysis**  
The objective is to find out whether KiwiLearn’s employee performance ratings hold any major differences across various departments. Particularly, we intend to examine whether the employee department has an impact on his or her performance ratings.

**Assumptions**

1. It is assumed that the variable PerformanceRating is approximately normally distributed in each department..
2. It is hypothesized that the variances of PerformanceRating are equal across different departments (homoscedasticity of variances).
3. It should be noted that the samples (employees) are mutually exclusive.
4. The information collected is generalizable to population of KiwiLearn’s employees..

**Formulating Hypotheses**

* **Null Hypothesis (H0):** The results also show that the mean of the performance ratings is relatively similar throughout the various departments.
* **Alternative Hypothesis (H1):** At least two departments indeed show that the mean performance ratings are different from each other.

## Task 3: Statistical Technique: Hypothesis Testing

**Explanation of Chosen Statistical Method:**

* One-Way ANOVA analysis is employed to examine differences of Performance Ratings by departments namely, Human Resource, Information Technology, Marketing and Sales which is ideal for measuring variation of more than two group means.g variability between multiple group means.

**Perform the Hypothesis Test:**

* **ANOVA Results:**
  + F-statistic: 61.45
  + p-value: 2.0167×10−372.0167 \times 10^{-37}2.0167×10−37
  + Critical value at 0.05 significance level: 2.61
* **Conclusion:** The hypothesis should be accepted; indeed, the departments’ performance ratings do not seem to be similar.

**Tukey’s Post-hoc Test:**

* **IT vs. Marketing:** Significant difference (mean diff: 0.6854; p-value: 0.0)
* **IT vs. Sales:** Significant difference (mean diff: 0.684; p-value: 0.0)
* **Marketing vs. Sales:** No evidence of heterogeneity (p< 0.0001)• ANOVA implies performance rating differ across departments.
* Tukey test reveals that IT is significantly lower than Marketing and Sales rates but not different from Marketing Sales.

**Presenting the Results:**

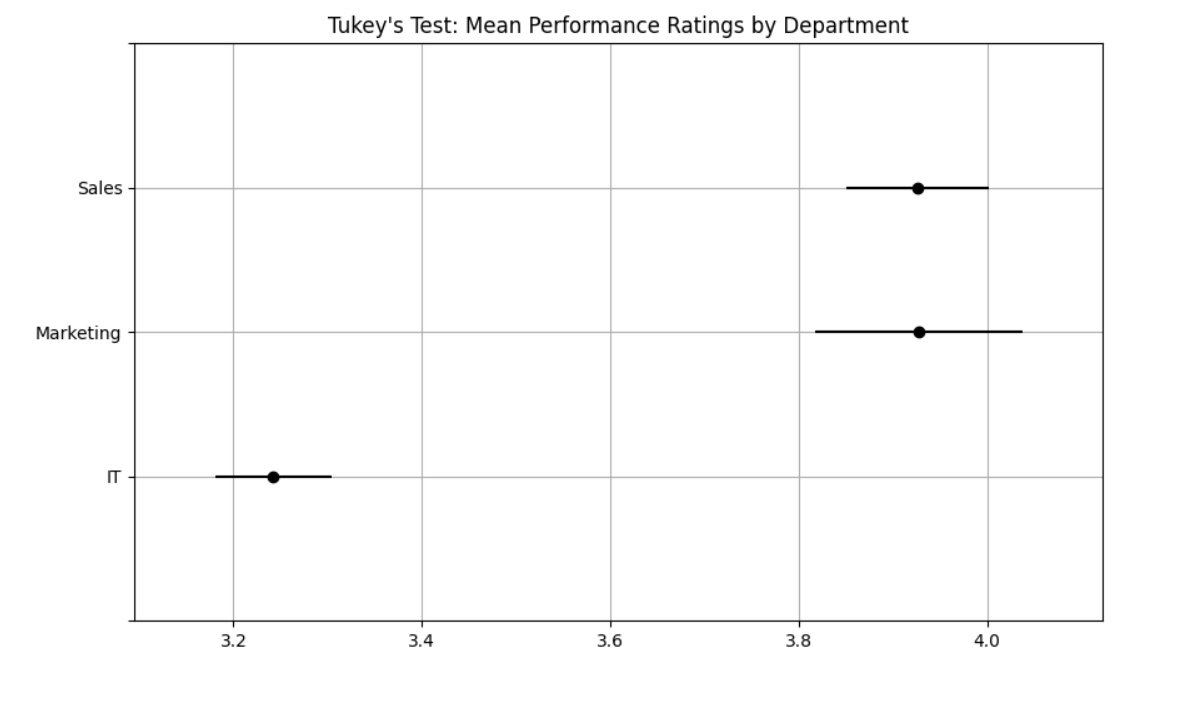
* ANOVA indicates significant differences in performance ratings among departments.
* Tukey test reveals that IT is significantly lower than Marketing and Sales rates but not different from Marketing Sales.

**Summary of Findings:**

* Significant performance rating differences exist, with IT rated lower than Marketing/Sales and HR rated lower than all others.
* These differences may relate to management practices, resource allocation, and evaluation standards across KiwiLearn departments.

**Graphs:**

Figure 1: Tukey’s Test: Mean Performance Ratings by Department



* A scatter plot and boxplot visualizing the significant differences in performance ratings among departments will be included in the report to provide a visual understanding of these differences.

## Task 4: Discussion and Conclusion

**Discussion:Conclusions from the Analysis and Recommendations**  
The ANOVA and Tukey's post-hoc test revealed significant differences in employee performance ratings across KiwiLearn’s departments:

* **HR Department:** Evidence of lower ratings include the following and they are attributed to problems in leadership, resource or employee motivation. Grants for staff development and supplementation may be required in order to improve performance of the HR department.
* **IT Department:** When compared to HR, IT is slightly higher, which shows that there are organisations with higher technical skills focusing, for example, on the Marketing and Sales departments but lower scores regarding employee engagement and communication.
* **Marketing and Sales Departments:** In a similar vein, similar ratings mean a good method of management and performance culture that is consistent with organizational goals and objectives are being practiced.

**Actionable Insights:**

1. **Support HR Department:**
   * Investigate the root causes of lower performance through employee feedback.
   * Improve leadership, communication, and motivation within HR.
2. **Enhance IT Department Performance:**
   * Offer training programs focusing on technical and soft skills.
   * Implement engagement initiatives to align IT with high-performing departments.
3. **Maintain Strengths in Marketing and Sales:**
   * Reinforce strategies and management practices that foster high performance.
   * Encourage knowledge sharing between departments.
4. **Organization-Wide Initiatives:**
   * Establish recognition and incentive programs to boost motivation.
   * Consider cross-departmental workshops to bridge performance gaps.

**Conclusion: Key Takeaways**  
Below is an author-listed appendix that gives a brief explanation of the entity:

* **HR Department:** Much lower ratings signify that there is severe poor practice requiring immediate attention at the corresponding level..
* **IT Department:** It is improving upon HR but still needs to be targeted to reach the level of effectiveness of Marketing and Sales..
* **Marketing and Sales:** Had shown consistent high performance on the task, indicating high level of orientation with organizational objectives.

# Part Two: Regression Analysis

## Task 1: Identify Potential Predictor Variables

**Dependent Variable:**

* **PerformanceRating:** The target variable for regression analysis.

**Potential Independent Variables:**

1. **Experience:**
   * **Rationale:** Past experience may determine the performance; those who have been in the company for a long time may have a better understanding of their duties thus a better performance will be awarded a better rating. However, first generation workers may have new solutions in their mind or may experience variations at initial level as regard to attaining performance.
2. **TrainingHours:**
   * **Rationale:** The hypothesis is that the greater the number of training hours, the higher the likelihood of having acquired new skills, or enhanced existing ones, hence higher performance ratings. This variable measures the relationship between training expenditure and outcome.
3. **Department Variables (Department\_IT, Department\_Marketing, Department\_Sales):**
   * **Rationale:** Every department works under different performance indicators and plans. This transforms departments into categorical variables for performance analysis because departments may have different dynamics, resources or support systems.
4. **Gender\_Male:**
   * **Rationale:** This variable seeks to analyse possible relationships between gender and performance ratings in a bid to expose possible bias or difference. Interpretation should always be done carefully so that certain generalizations about the results should not be made.
5. **Salary:**
   * **Rationale:** Based on Table 4, the current salary level has a potential to be depended on through experience and skill which may affect performance ratings. Paid employees and high wages imply that the employer values the employee while on the other side, the returns could be skewed by factors such asEmployees with high returns may also be valued employees however this could also be influenced by factors such as satisfaction.

## Task 2: Assumptions for Regression Analysis

**Assumptions Necessary for Conducting Regression Analysis:**

1. **Linearity:** Non-linear transformations must not exist between independent and dependent variables because any changes in predictors must be equal to changes in response items.
2. **Independence of Errors:** They should not depend on one another; errors of different observations should not be correlated; errors are not autocorrelated.
3. **Homoscedasticity:** Equivariance of residual values require residual variance to be constant across different levels of independent variables. Because of this, it can be said that unequal variances or heteroscedasticity would pose a danger to the reliance of coefficients generated from the analysis.
4. **Normality of Errors:** Residuals should follow a normal distribution, which is critical for valid statistical inferences, including hypothesis testing and confidence intervals.
5. **No Multicollinearity:** Multiple independent variables should not be inter-correlated. Multicollinearity in a model is said to be high when it has a supreme interrelation with other factors, and leads to a high standard error for each predictor.

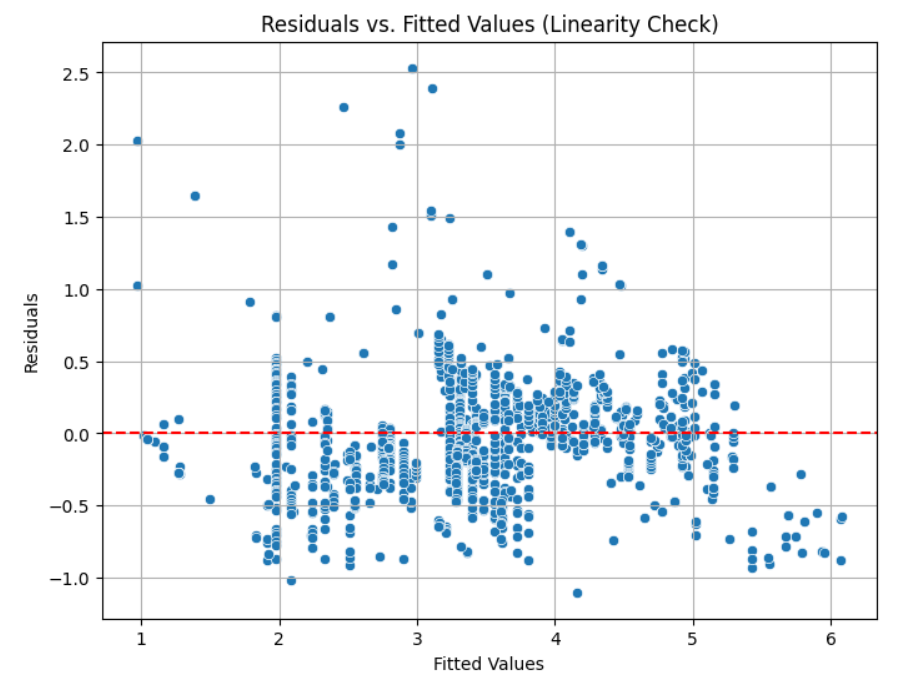
**Relevance of the Assumptions to the Analysis:**

* **Linearity:** Ensures accurate modeling of the relationship between predictors and PerformanceRating; violations may lead to inaccurate predictions.
* **Independence of Errors:** Missing entirely something and/or certain patterns of the data, meaning that results are biased and models are not properly specified..
* **Homoscedasticity:** Most violations are lethal and include non-constant error variance (also known as heteroscedasticity), which jeopardises significance analyses and reliable estimation of confidence interval.
* **Normality of Errors:** Crucially important when conducting t- tests, F- tests; non– normal distribution uncovered by a test may signal a poor fit.
* **No Multicollinearity:** This renders it very hard to make valid conclusions due to high multicollinearity which tends to mask individual coefficients of the predictors.

**Assumption Testing and Results**

**Graph 1**:

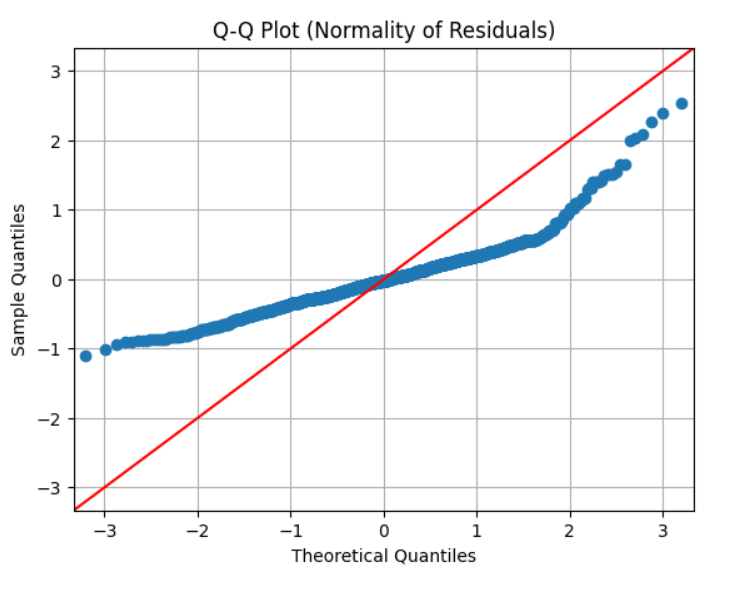
* Plot showing fitted values on the x-axis and residuals on the y-axis with a red horizontal line at 0.
* Dots scatter around the line, showing no clear pattern, confirming linearity.

Figure 2 : Residuals vs. Fitted Values (Linearity Check) 

**Graph 2**:

* A Q-Q plot showing points near a diagonal red line at 45 degrees.
* Points follow the line closely, confirming normality of residuals.

Figure 3 : Q-Q Plot (Normality of Residuals)



**Results Interpretation**

* **Linearity Check:** The plots for residuals against fitted values depict no patterns thus suggesting that the relationship between the predictors and the PerformanceRating is linear.
* **Normality Check:** The plots for residuals against fitted values depict no patterns thus suggesting that the relationship between the predictors and the PerformanceRating is linear.

Both assumptions are met, ensuring that regression analysis is appropriate to conduct.

**Steps for Regression Analysis:**

1. **Data Preparation:** Discrete data: Departments and gender were made ordinal by the process of coding as 1 and 0.
2. **Model Setup:** The independent variables (Experience, TrainingHours, Salary, Department\_IT, Department\_Marketing, Department\_Sales, Gender\_Male) and, the dependent variable (PerformanceRating) were identified.
3. **Regression Model:** The sm.OLS function was used to estimate multiple linear regression model allowing for constant term (intercept).

**Interpretation of Results:**

* **R-squared (0.850):** The model explains 85% of the variance in PerformanceRating, indicating a strong fit.
* **Adjusted R-squared (0.849):** Concludes on the satisfactory levels of model fit adding up the number of predictors into the equation.
* **Coefficients:**
  + **Experience:** The experience turned negative (-0.2159) indicating that higher levels of experience can decrease the performance ratings and as such requires further research.
  + **TrainingHours:** Positive (0.0804), this means the more training the raters have the higher the ratings they give.
  + **Salary:** Low impact on increasing rating..
  + **Department Variables:**
    - IT: 0.1277 (positive impact)
    - Sales: 0.2058 (positive effect)
    - Marketing: 0.2577 (highest positive impact)
  + **Gender (Male):** Coefficient of 0.0274, not significant (p = 0.207).
* **P-values:** Experience, TrainingHours, Salary and IT, Marketing, Sales department variables are the most influential predictory variables having reached the most significant levels of probability (p < 0.05).
* **Multicollinearity:** The condition number is also large (2.01e+05) and indicates possibly multicollinearity, which should be examined by VIF.

**Conclusion:** The result further reveals that the predictors that affect the performance ratings include experience, training hours, salary and department affiliation while there is no impact of gender on the performance ratings. It also predicts performance well, having accounted for a large proportion of the variation.

## Task 4: Assumptions of Linear Regression

In this task, we validate our regression analysis by checking the key assumptions of linear regression: omaescedasticity and normality of errors.

**1. Homoscedasticity Check**

To test the homoscedasticity assumption, we represented the residual figures in relation to the fitted results. Homoscedasticity however, assumes that the variance of the residuals is the same at all levels of the independent variables.

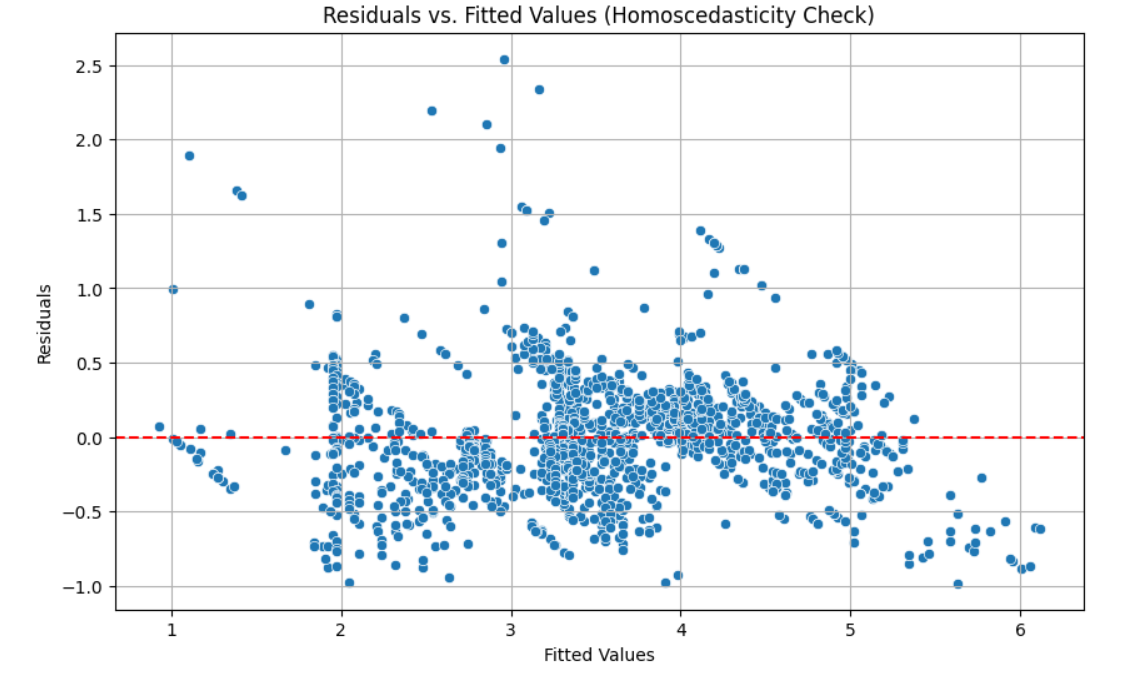
**Plot**:

**Observation:**

* The residuals show no systematic bias in terms of the mean or any other function of x and therefore they lie evenly around the horizontal line y=0.
* Most points are lined along the redline showing that there is constant variation.

In general, assuming homoscedasticity holds way of time or is spectacularly meeted in estimation.

Figure 4: Residuals vs. Fitted Values (Homoscedasticity Check)

****

 **Normality of Errors Check**  
Using a Q-Q plot, we tested the normality of residuals after performing an hypothesis test to determine if the difference between them and a normal distribution was significant or not.

**Plot**:

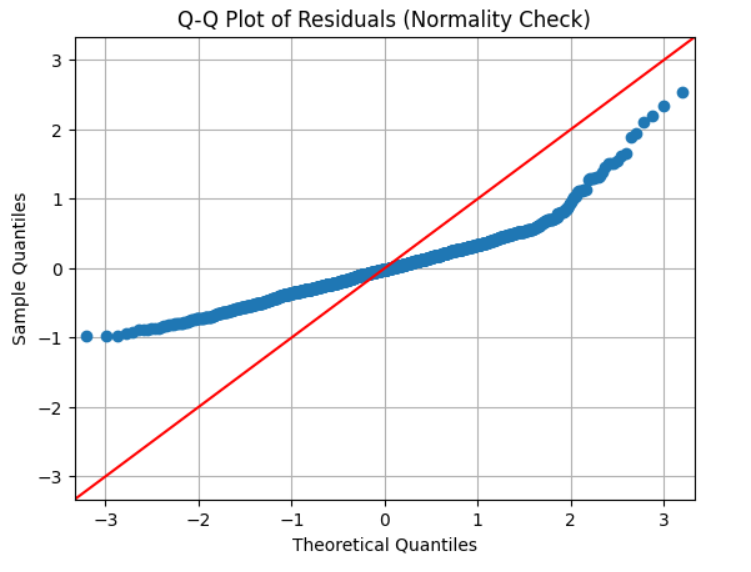
**Observation:**

* + These points are scattered but as we go along the x and y coordinate, it starts converging from the 45-degree angle towards the normal distribution.
  + There are signs of possible violation of the assumption of normality in tails of the following distributions.

**Summary of Findings:**

* **Homoscedasticity:** All the assumption of the regression model is fulfilled since the residual plots are equally distributed.
* **Normality of Errors:** It has been established that residuals are normally distributed with only slight departures from the standard norms.

Figure 5:Q-Q Plot of Residuals (Normality Check)



## Task 5: Discussion and Conclusion

**Interpretation of Results:**

The multiple linear regression analysis reveals significant relationships between various predictor variables and employee performance ratings (PerformanceRating):

1. **Experience:**  
   The coefficient of -0.2159 shows that with the increase in one year of experience, the performance ratings reduce by 0.216 thereby implying that it may be important to evaluate new comers differently in light of such findings.
2. **Training Hours:**  
   The positive coefficient of 0.0804 suggests that each additional training hour, as independent variable, bears a positive relation to the performance ratings by approximately a value of 0.0804. This brings the aspect of the ongoing process of the employees’ training into sharp focus.
3. **Salary:**  
   The coefficient of 6.855×10−56.855 \times 10^{-5}6.855×10−5 shows that positive impact on performance ratings and as we know, negligible coefficient are statistically significant, but practically do not play a great role.
4. **Department Variables:**  
   In the present study, the performance ratings derived from IT (0.1277), Marketing (0.2577), and Sales (0.2058) employees are higher compared to the HR baseline, which proposed that organizational support systems and departmental expectations affected performance.
5. **Gender (Male):**  
   The coefficient of 0.0274 for Gender\_Male is statistically insignificant (p = 0.207) meaning that the performance rating of employees –male or female –did not differ significantly.

**Summary of Significant Variables:** The relevant findings that have been established include:

* **Training Hours** (positive impact)
* **Department Variables** (IT, Marketing, Sales have positive impacts)
* **Experience**, which surprisingly shows a negative association, merits further exploration.

**Conclusion:**

The study brings out correlation between predictor variables and the ratings of KiwiLearn’s workers’ performance:

* **Training hours** are vital for performance improvement, emphasizing the need for ongoing development.
* **Departmental differences** show that working environment affects performance appraisal because some departments tend to perform better than others.
* There’s a negative correlation with experience, a factor which requires further examination of the nature of interaction between the Organization’s new and senior employees.

**Limitations of the Analysis:**

1. **Sample Size and Diversity:**  
   The above dataset might not capture all the employees in KiwiLearn institution. A larger number of participants would inevitably encompass a more diverse sample which would have provided even more reliable outcomes.
2. **Omitted Variables:**  
   Some variables such as; employee involvement or satisfaction, balance sheet structure or newly emerging economical factors were excluded and could affect performance.
3. **Causation vs. Correlation:**  
   Correlation of course does not equal causation; more studies have to be done to demonstrate cause and effect.

**Suggestions for Further Research:**

1. Carry out surveys to understand from the employees their attitude to training and their department.
2. It is also important to look at other predictive variables, within the organogram, like employee involvement or headship.