Student ID Number: 0783063
18029208
22180242

Assignment 2

Semester 2 2023

PAPER NAME: Data Analysis

PAPER CODE: COMP517

Student ID	Student Names
0783063	Billie-Jean Laing
22180242	Juchang Kim
18029208	William Bank Sukjaem

Due Date: Midnight Friday 20th Oct 2023
TOTAL MARKS: 100

INSTRUCTIONS:

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

Table of Contents

Part 1: Exploring Data and Testing Hypotheses 3
Data Preparation and Exploration: 3
Multivariate analysis 5
Assumption and Hypothesis Formulation 7
Statistical Technique: Hypothesis Testing 8
Discussion 10
Conclusion 11
Part 2: Regression Analysis 13
Identify Potential Predictor Variables 13
Assumptions for Regression Analysis 13
Assumption Tests 14
Regression Analysis 20
Assumptions of Linear Regression 22
Discussion and Conclusion 24
Appendices 25
Figure 1.0 – Summary Statistics of Numeric Data 25
Table 1.0 – Number of Outliers with a Z-Score threshold of 3 25
Table 1.1 – Summary Statistics of KiwiLearn Employees 26
Table 1.2 – Distribution Skew of Numeric Data 26
Figure 1.1 – Distribution of Numeric Columns 26
Figure 1.2 – Number of employees in each department 27
Figure 1.3 – Relationship between Years of Experience and Performance Ratings by Department 27
Figure 1.4 – Probability of Distribution of Performance Rating 28
Figure 1.5 – One way ANOVA results based on P-value 28
Figure 1.6 – One-way ANOVA results based on F-statistic 29
Figure 1.7 – Tukey’s HSD Post Hoc test results 29
Table 1.8 – Mean value of Performance rating per Department 29
Figure 2.0 – Correlation Heatmap of all variables 30
Figure 2.1 – Correlation Heatmap of Independent Variables 31
Figure 2.2 – Scatter Plots for each independent variable against the dependent variable 31
Figure 2.3 – VIF Results to test for multicollinearity 31
Figure 2.4 – OLS Regression Results (Training Hours, Salary) 32
Figure 2.5 – OLS Regression Results (Training Hours, Experience) and VIF results 32
Figure 2.6 – OLS Regression Results (Training Hours, Experience) used for Regression Analysis 33
Figure 2.7 – QQ Plot of Residuals for visual assessment 34
Figure 2.8 – Anderson-Darling Statistic output results 35
Figure 2.9 – Homoscedasticity Plot showing Residuals vs Predicted Values 35

Part 1: Exploring Data and Testing Hypotheses

Data Preparation and Exploration:

The dataset encompasses 1468 data points, each representing categorical and numerical information about employees at KiwiLearn. The information relates to Gender, Department, Experience, Training Hours, Salary and Performance Rating. Through the cleansing process we can confirm that the data has no missing values, no duplicates and no negative values. The data has outliers. There are two methods to uncover outliers. One method is through box plots while the other method is through the z score. Figure 1.0 below shows some box plots visually displaying the summary statistics including the outliers for all numeric columns. Through the box plots we can determine that there is at least one outlier in the Experience column and at a minimum more than ten outliers in the Salary column.

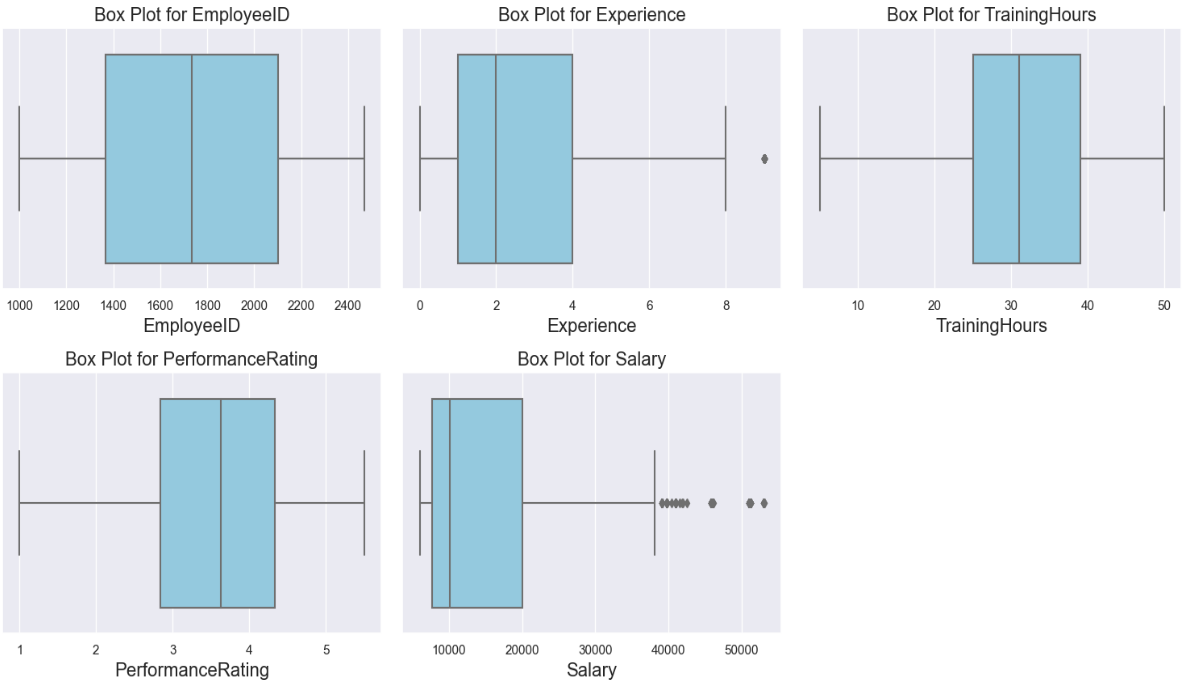
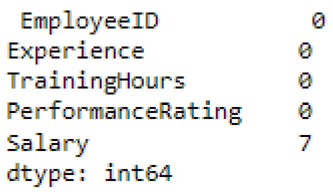


Figure 1.0 – Summary Statistics of Numeric Data

The number of outliers are different when using the Z-Score. At a threshold of 3 (which indicates extreme outliers), seven outliers were identified in the data. This is shown in Table 1.0 below.

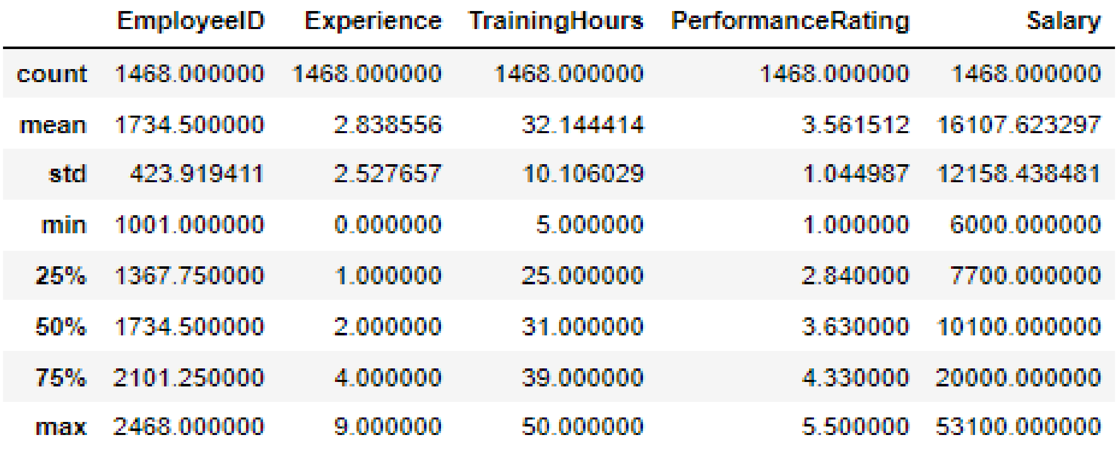


```
 EmployeeID          0
Experience          0
TrainingHours       0
PerformanceRating   0
Salary              7
dtype: int64
```

Table 1.0 – Number of Outliers with a Z-Score threshold of 3

Table 1.0 indicates no outliers in the Experience column unlike the Box Plot outliers and only 7 outliers present in the Salary column unlike the high amount above 10 present in the Box Plot for Experience. For the purpose of this report, we have decided to base our outlier assessment on the outliers derived from the Z-Score. In this case, the 7 outliers present in the Salary column will not impact the integrity of the data as there is an interesting relationship between Salary, Experience and Performance Rating that needs to be explored. Keeping the outliers would be a true representation and reflection of the population of employees at KiwiLearn. This report will proceed with the outliers present in the dataset.

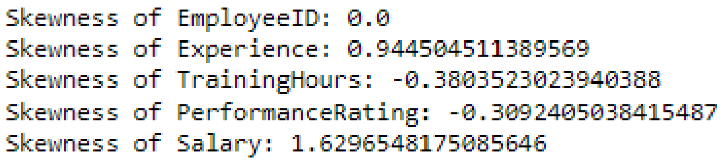
Table 1.1 shows a summary of the numerical statistics of the employee details. Regarding employee experience at KiwiLearn, the average tenure is approximately 2.84 years, placing most employees in the Junior category, as this falls within the 5-year experience threshold. Employee performance is evaluated on a scale of 1 to 5.5, where 1 represents poor performance and 5.5 signifies exceptional performance. With an average employee tenure in the Junior category, the mean performance rating of 3.56 is relatively strong.



	EmployeeID	Experience	TrainingHours	PerformanceRating	Salary
count	1468.000000	1468.000000	1468.000000	1468.000000	1468.000000
mean	1734.500000	2.838556	32.144414	3.561512	16107.623297
std	423.919411	2.527657	10.106029	1.044987	12158.438481
min	1001.000000	0.000000	5.000000	1.000000	6000.000000
25%	1367.750000	1.000000	25.000000	2.840000	7700.000000
50%	1734.500000	2.000000	31.000000	3.630000	10100.000000
75%	2101.250000	4.000000	39.000000	4.330000	20000.000000
max	2468.000000	9.000000	50.000000	5.500000	53100.000000

Table 1.1 – Summary Statistics of KiwiLearn Employees

Table 1.2 and Figure 1.1 illustrates the Skew and Distribution of each Numeric Column in the dataset. The Employee ID column indicates a list of all the employee numbers. This column is neither positive or negative skewed. The Experience column is positively skewed indicating that most of the employees have under 4 to 5 years of experience. The Training Hours column indicates a negative skew which means a majority of the employees have received at least 30 hours of training within the last year. The Performance Rating is also negatively skewed meaning a majority of the employee ratings fall with 3 and 5.5. The Salary column is positively skewed meaning the salary of most employees fall under $2,500.00 per month.



```
Skewness of EmployeeID: 0.0
Skewness of Experience: 0.944504511389569
Skewness of TrainingHours: -0.3803523023940388
Skewness of PerformanceRating: -0.3092405038415487
Skewness of Salary: 1.6296548175085646
```

Table 1.2 – Distribution Skew of Numeric Data

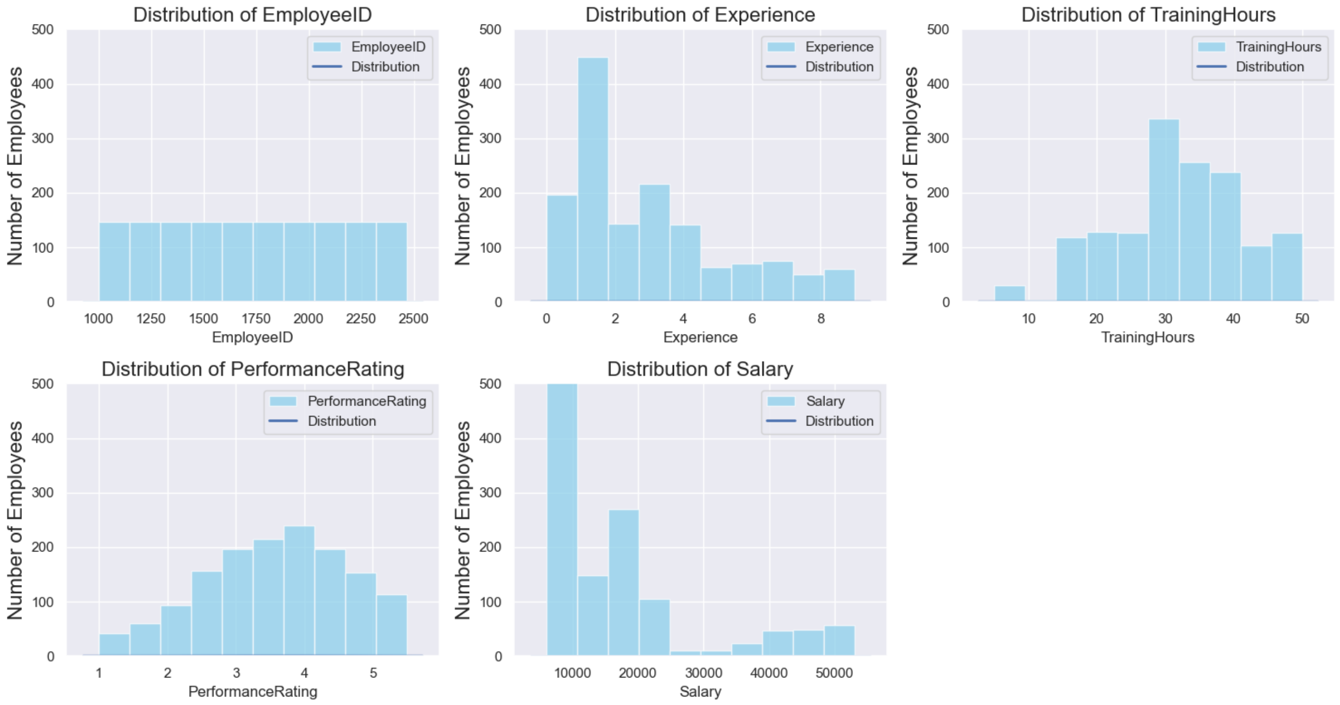


Figure 1.1 – Distribution of Numeric Columns

Multivariate analysis

Figure 1.2 shows the distribution of employees per department. This will help provide context when analysing the experience level by performance rating for each department.

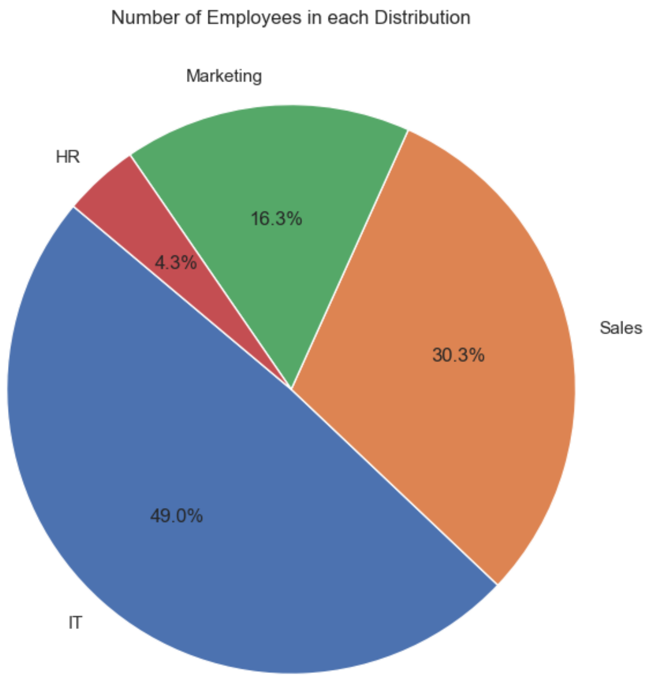


Figure 1.2 – Number of employees in each department

The IT department holds the largest number of employees followed by Sales, Marketing and HR.

Figure 1.3 categorises employees by department and illustrates the experience level by performance rating. The departments that appear to achieve a mean rating of above 3.5 across all of the experience levels are Sales and Marketing. At Entry, Junior and Senior level, the Sales and Marketing departments have a performance rating mean of around 4 however at Mid-Level the Performance Rating increases to around 4 and higher. IT and HR have a Performance Rating level of around 3 at Entry and Junior level however things change at Mid and Senior level. The IT department has an increased performance rating at the Mid level experience range with their employee ratings sitting between 4.6 and 5.4. HR sees a drop in the mean performance rating to 2 at Mid level but both IT and HR begin to stabilize and rise slightly to around 3.5 - 3.6 at Senior level. When it comes to comparing the performance of Sales and Marketing at Entry and Junior level with IT and HR at Entry and Junior level it is important to recognise the difference in size of department. The performance of each department is related to the performance and work ethic of the manager. Smaller departments with excellent performing managers with good work ethic might have more tightly-knit teams, enabling better coordination and potentially higher average performance ratings due to closer supervision and more personalized attention from managers. Smaller departments with bad performing managers with under par work ethic would perform the opposite. This could explain why Sales and Marketing perform better than IT and HR perform poorly.

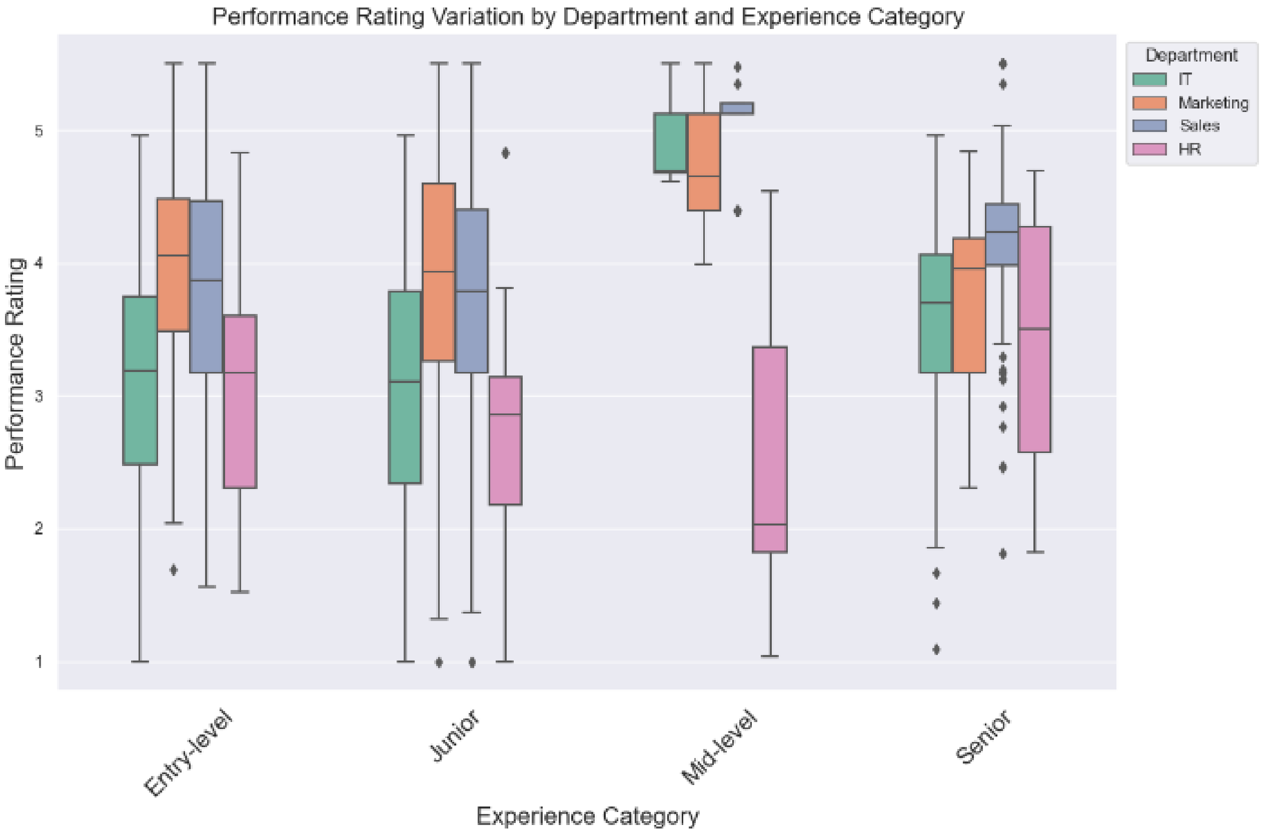


Figure 1.3 – Relationship between Years of Experience and Performance Ratings by Department

Assumption and Hypothesis Formulation

The purpose of this analysis is to investigate the potential variations in employee performance ratings across the departments of an organisation called KiwiLearn. KiwiLearn is enthusiastic about examining its data to gain insights for future decision-making. This analysis will help them identify the departments that exhibit notably higher or lower ratings and the possible reasons and relationships around this. Figure 1.4 shows the current probability of the distribution of employee performance ratings.

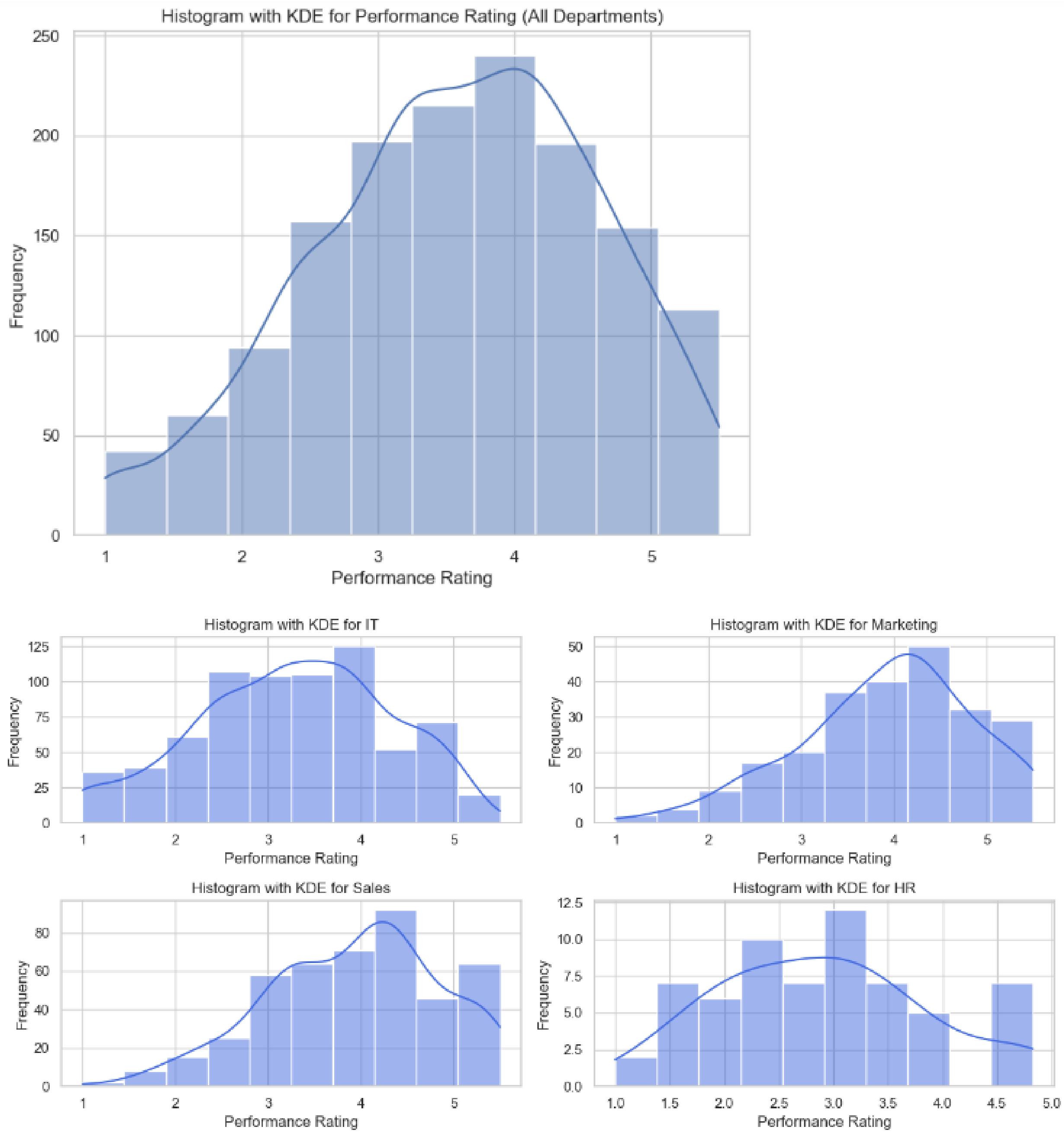
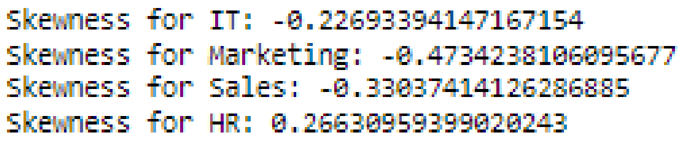


Figure 1.4 – Probability of Distribution of Performance Rating

A calculation was done to indicate the Skewness of the Performance Rating histogram for all departments. The skew amounted to -0.31. This indicates a negative skew in the performance rating across all departments. The skews of the departments individually are displayed below in Table 1.3. IT, Marketing and Sales show a negative skew while HR shows a positive skew. Based on the skews shown for each department in Figure 1.4 and Table 1.3 this could be the first indication that the means of each department are different, validating the need to perform hypothesis testing.



```
Skewness for IT: -0.22693394147167154
Skewness for Marketing: -0.473423810609567
Skewness for Sales: -0.33037414126286885
Skewness for HR: 0.26630959399020243
```

Table 1.3 – Skew of Each Department's Performance Rating

The following assumptions will be made:

- The dataset is representative of the entire employee population at KiwiLearn.
- The performance ratings follow a normal distribution within each department.
- The years of experience are accurately recorded and categorization into experience levels is appropriate.

Further analysis and calculations will be based on these assumptions.

Using the following Null and Alternative Hypothesis we will determine whether the skews of each department shown above indicate a statistical difference of the mean value of employee performance ratings across departments.

Null Hypothesis (H_0): There is no significant difference in the average performance ratings amongst different departments.

H_0 : μ_1 IT = μ_2 Sales = μ_3 Marketing = μ_4 HR

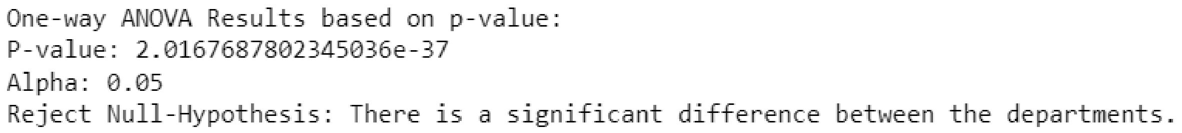
Alternative Hypothesis (H_a): There is a significant difference in the average performance ratings amongst different departments.

H_a : μp ≠ μq (p,q: [IT, Sales, Marketing, HR] & p ≠ q})

Statistical Technique: Hypothesis Testing

We will gather evidence to support or to reject the null hypothesis through Analysis of Variance (ANOVA). We are comparing 4 departments therefore using the one-way ANOVA method is suitable as it allows for comparisons among three or more groups. We will be able to identify whether there are significant statistical differences in the mean between each department.

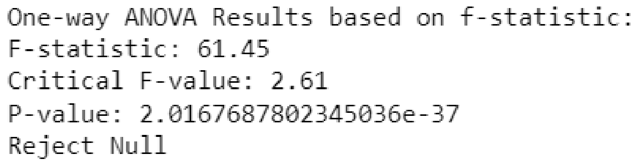
Figure 1.5 below shows the ANOVA results based on the P Value. If the p-value is less than the significance level (here αlpha = 0.05) then reject the null hypothesis (H0). In this case, the p-value is 2.0168e-37, which is extremely small and significantly less than alpha 0.05. Therefore, we reject the null hypothesis because there is a significant difference between at least two department means.



```
One-way ANOVA Results based on p-value:
P-value: 2.0167687802345036e-37
Alpha: 0.05
Reject Null-Hypothesis: There is a significant difference between the departments.
```

Figure 1.5 – One way ANOVA results based on P-value

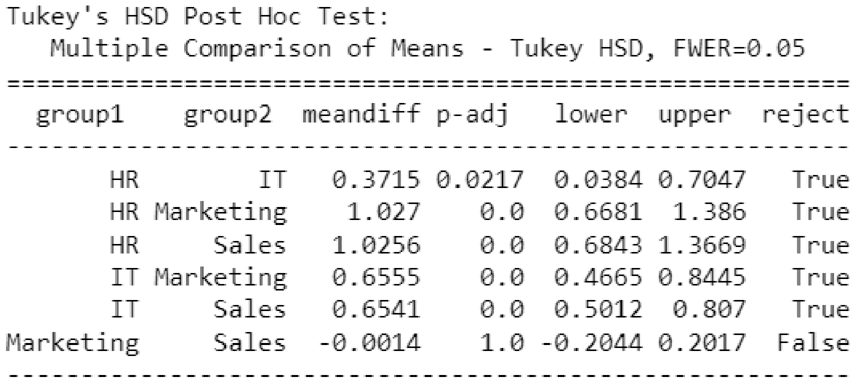
Figure 1.6 below shows the ANOVA results based on the F Statistic. If f-statistic is greater than the critical F-value, it means that the F-statistic is in the tail of the F-distribution, and you should reject the null hypothesis. In this case the f-statistic is 61.45 which is greater than the critical f-value of 2.61. Therefore, we reject the null hypothesis because there is a significant difference in performance ratings among different departments.



```
One-way ANOVA Results based on f-statistic:
F-statistic: 61.45
Critical F-value: 2.61
P-value: 2.0167687802345036e-37
Reject Null
```

Figure 1.6 – One-way ANOVA results based on F-statistic

As the P-value and F-statistic indicate a significant difference across departments a Tukey's HSD Post-Hoc Test needs to be conducted to identify specific departments with significant differences. Figure 1.7 displays the results from Tukey's HSD post hoc test.



```
Tukey's 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 1.7 – Tukey's HSD Post Hoc test results

Comparing HR and IT, the mean difference is 0.3715, and the p-value (p-adj) is 0.0217, indicating that there is a statistically significant difference between these HR and IT.

Comparing HR and Marketing, the mean difference is 1.027, and the p-value (p-adj) is 0.0, indicating that there is a statistically significant difference between these HR and Marketing.

Comparing HR and Sales, the mean difference is 1.0256, and the p-value (p-adj) is 0.0, indicating that there is a statistically significant difference between HR and Sales.

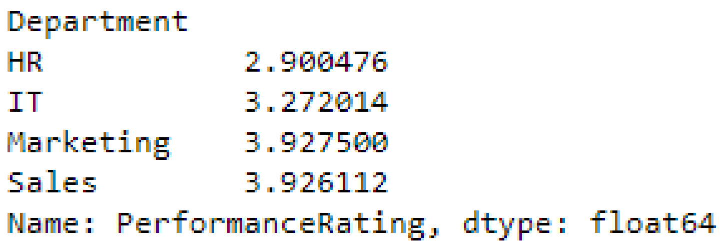
Comparing IT and Marketing, the mean difference is 0.6555, and the p-value (p-adj) is 0.0, indicating that there is a statistically significant difference between IT and Marketing.

Comparing IT and Sales, the mean difference is 0.6541, and the p-value (p-adj) is 0.0, indicating that there is a statistically significant difference between IT and Sales.

Comparing Marketing and Sales, the mean difference is -0.0014, and the p-value (p-adj) is 1.0, indicating that there is no statistically significant difference between Marketing and Sales.

The Tukey's HSD test in Figure 1.7 shows significant differences in all department combinations except the combination of Sales and Marketing.

To support the evidence found in Tukey's HSD test the mean values of each department have been calculated and shown in Table 1.8. It is apparent that Sales and Marketing have very similar average performance ratings, Marketing 3.927 (3 decimals) while Sales 3.926(3 decimals) as confirmed in Tukey's test. It appears the departments with the largest difference in performance rating are HR and Marketing, HR with an average of 2.900 (3 decimals) and Marketing's average of 3.927 (3 decimals). As per the mean performance scores below, the remaining departments are different to each other and Tukey's HSD test has confirmed that the differences are statistically significant. The department with the highest performance rating is Marketing, followed by Sales, then IT and finally HR.



```
Department
HR           2.900476
IT           3.272014
Marketing    3.927500
Sales        3.926112
Name: PerformanceRating, dtype: float64
```

Table 1.8 – Mean value of Performance rating per Department

Discussion

Based on the results of the above analysis, it is evident that there are significant differences in performance ratings among the different departments. The implications that could arise include a change in the culture of KiwiLearn as an organisation, employee morale and motivation could be impacted and management may need to assess their strategic focus depending on if they had plans to innovate the business in areas with lower performance ratings.

In terms of the culture at KiwiLearn the differing performance ratings could cause a divide in departments. There's a chance that higher rating departments such as Sales and Marketing could foster a success mindset culture with competitiveness and innovation whilst lower performing departments such as HR and IT may continue with their complacency and (perceived) underachievement. In the event of a divide in the departments, there is a further impact to KiwiLearn in the form of reduced collaboration and communication amongst the departments. Collaboration and Communication across departments is essential in the overall successful performance of an organisation. To bring more context to these performance ratings, it would be interesting to investigate and analyse the determinants of the performance rating to ensure it is based fairly across each department.

Like organizational culture, a decline in employee morale and motivation can generate similar repercussions within KiwiLearn. Diminished morale and motivation might lead to lower job satisfaction and increased turnover rates, as employees may feel disheartened and undervalued.

Considering the amplification of negativity versus positivity in today's environment, KiwiLearn would undoubtedly prioritise avoiding the long-term persistence of decreased employee morale and motivation.

Regarding KiwiLearn's strategic focus, management should evaluate and adjust their strategic priorities when introducing innovative practices to departments. Their innovative practices should match their targeted departments after considering the morale/motivation and culture. Implementing the correct strategic focus to the correct department/employees could yield significant improvements towards their plans ensuring that innovation is strategically aligned with the goal of enhancing company performance and growth.

Possible reasons for the differences in ratings among departments could stem from different management styles and differing work cultures among the different departments. There is a chance that some departments could have more distinct and challenging roles and responsibilities that require a unique skill set. Moreover, the allocation of resources and support may differ, contributing to discrepancies in performance ratings. For example the IT might possess stronger technical skills causing their performance rating to be higher whereas Sales would prioritise reaching sales and revenue targets.

KiwiLearn has the opportunity to understand the specific factors that contribute to the variations in performance ratings among departments. They could conduct employee surveys or interviews to discover deeper insights into the specific drivers of performance within each department. Doing so will help them identify best practices to implement across underperforming departments to improve performance. This will help identify best practices from high-performing departments that can be implemented across lower departments to improve overall performance.

Conclusion

The analysis has revealed that we can reject the Null Hypothesis of "There is no significant difference in the average performance ratings amongst different departments". In turn it has been uncovered that the Alternative Hypothesis of "There is a significant difference in the average performance ratings amongst different departments" is true. The evidence to Reject the Null Hypothesis came from the ANOVA method as this method allowed us to compare all 4 departments. The Anova results based on the P-Value conclude that the p-value is 2.0168e-37, which is extremely small and significantly less than alpha 0.05 meaning we can reject the null hypothesis. The Anova results based on the F Statistic conclude that the f-statistic is 61.45 which is greater than the critical f-value of 2.61, therefore, we can reject the null hypothesis.

After uncovering that we can reject the null hypothesis based on the ANOVA Results, Tukey's HSD Post Hoc Test solidified the findings by providing us with the exact calculation of mean difference and specifically for which departments we can reject the null hypothesis. Tukey's results revealed the departments HR and IT, HR and Marketing, HR and Sales, IT and Marketing, and IT and Sales have significantly different average performance values that range from a mean difference value of 0.3715 at the least to 1.027 at the most therefore confirming that the null hypothesis can be rejected for these departments. The only department combination that had no significant performance difference was Sales and Marketing and in this single case, we can fail to reject the null hypothesis. The departments with the largest difference in performance rating are HR and Marketing, HR with an average performance rating of 2.900 (3 decimals) and Marketing's average performance rating of 3.927 (3 decimals). The Departments with the smallest difference in performance rating are Sales and Marketing, Sales with an average of 3.926 (3 decimals) and Marketing's average of 3.927 (3 decimals). The highest performing department is Marketing, followed by Sales, then IT and finally HR.

It is recommended that KiwiLearn conduct additional research within their organisation to understand the specific factors that contribute to the variations in performance ratings among departments. If they fail to do so, they could hinder their success for future innovative plans as there may be little communication and collaboration across the organisation. This may also cause implications in the form of shifts in organisational culture and employee morale and motivation potentially resulting in decreased staff retention.

Part 2: Regression Analysis

Identify Potential Predictor Variables

The selection of independent variables to compare against a dependent variable are essential in regression analysis. For optimal use, it is advisable that the independent variables are numerical and have a potential relationship with the dependent variable. In this analysis, our dependent variable is employee performance rating. We have opted to exclude the employee ID column from the analysis as it does not contribute any supplementary insights to the evaluation of employee performance ratings.

This report will focus on the following three variables as Potential Predictor Variables:

- Years of Experience
- Training Hours
- Salary

“Years of Experience” has been selected as a variable due to the strong connection between experience and performance. Typically, employees with greater experience tend to demonstrate improved performance, indicating the presence of valuable skills acquired over time.

"Training Hours" has been selected as a variable due to the potential link between training and improved performance. A higher number of training hours is believed to contribute to better performance outcomes, as continual training efforts can effectively enhance employees' skill sets and overall performance levels.

"Salary" has been selected as a variable due to the potential association between higher salaries and superior performance. A higher salary often signifies better performance, possibly indicating a relationship between compensation and employees' performance levels. Additionally, higher salaries are commonly associated with heightened performance expectations, which can motivate employees to deliver superior work outcomes.

Assumptions for Regression Analysis

This Regression Analysis will be based on the following assumptions:

- Linearity: There is a linear relationship between the predictor variables and the dependent variable.
- Homoscedasticity: The variance of the residuals must be constant across predicted variables.
- Normality of Residuals: Residuals follow a pattern of normal distribution.
- No Perfect Multicollinearity: Predictor variables are not perfectly correlated amongst other predictor variables

The linearity assumption is fundamental as it establishes the foundation of the regression model, asserting that there is a linear relationship between the predictor variables (Years of Experience, Training Hours, Salary) and the dependent variable (Employee Performance Rating). This assumption is critical because linear regression predicts the mean of the dependent variable as a linear combination of the independent variables. When this linearity assumption is violated, the model might misconstrue both the intensity and direction of the relationships among the variables - Years of Experience, Training Hours, Salary, and Employee Performance Rating, potentially obscuring the true nature of their associations.

Residuals should be consistent across all levels of predictors. This is the concept of Homoscedasticity. Deviations from homoscedasticity can compromise the reliability of the model's inferences. Heteroscedasticity (the opposite of Homoscedasticity), as indicated by a non-uniform pattern or a "horn" shape in the homoscedasticity plot, can contribute to inefficient coefficient estimates, thereby impacting the precision of the model. Overall, the absence of homoscedasticity can lead to biased standard errors, misinterpretations of the model's fit, and unreliable predictions.

In the context of Normality of Residuals, when the residuals do not follow a normal distribution, there's a risk of introducing bias into the coefficient estimates. This can subsequently lead to flawed interpretations of the relationship between Years of Experience, Training Hours and Salary (independent variables) and Employee Performance Rating (dependent variable). In addition, the assumption of normality is crucial for ensuring the validity of various statistical inferences, such as hypothesis tests (like t-tests and F-tests for coefficients). Any departure from normality in the residuals may result in erroneous conclusions from these tests.

Regarding Multicollinearity, violating the assumption of no Perfect Multicollinearity can render it challenging for any regression model to yield reliable solutions. This situation may lead to an elevation in the number of standard errors, contributing to unstable model suggestions and imprecise coefficient estimates. Consequently, it becomes increasingly difficult to accurately interpret the influence of each independent variable (Years of Experience, Training Hours, Salary) on the dependent variable (Employee Performance Rating), potentially leading to misinformed conclusions.

Assumption Tests

The assumptions of No Perfect Multicollinearity and Linearity should be tested before performing the regression.

Multicollinearity

Multicollinearity will need to be checked among the Years of Experience, Training Hours and Salary (predictor variables) using a correlation heatmap. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. In the context of multiple linear regression. Typically, a common threshold to identify multicollinearity is when the correlation coefficient (often denoted as "r") between two variables exceeds 0.7 (or any other chosen threshold).

Figure 2.0 shows the heatmap of correlation coefficient values between each variable. As we are looking for the presence of multicollinearity we will focus on the variables with correlation coefficients of 0.7 and above specifically between the Performance Rating (dependent variable) and each of the independent variables. There is a correlation coefficient of 0.86 between Performance Rating and Training Hours. This is exactly what we are looking for in terms of multicollinearity. This coefficient value indicates that Training Hours (independent variable) has a strong relationship with the Performance Rating (dependent variable) therefore TrainingHours should be identified as the independent variable for this regression analysis. The correlation coefficient values of Performance Rating against Experience and Performance Rating against Salary are all low between 0.1 and 0.4.

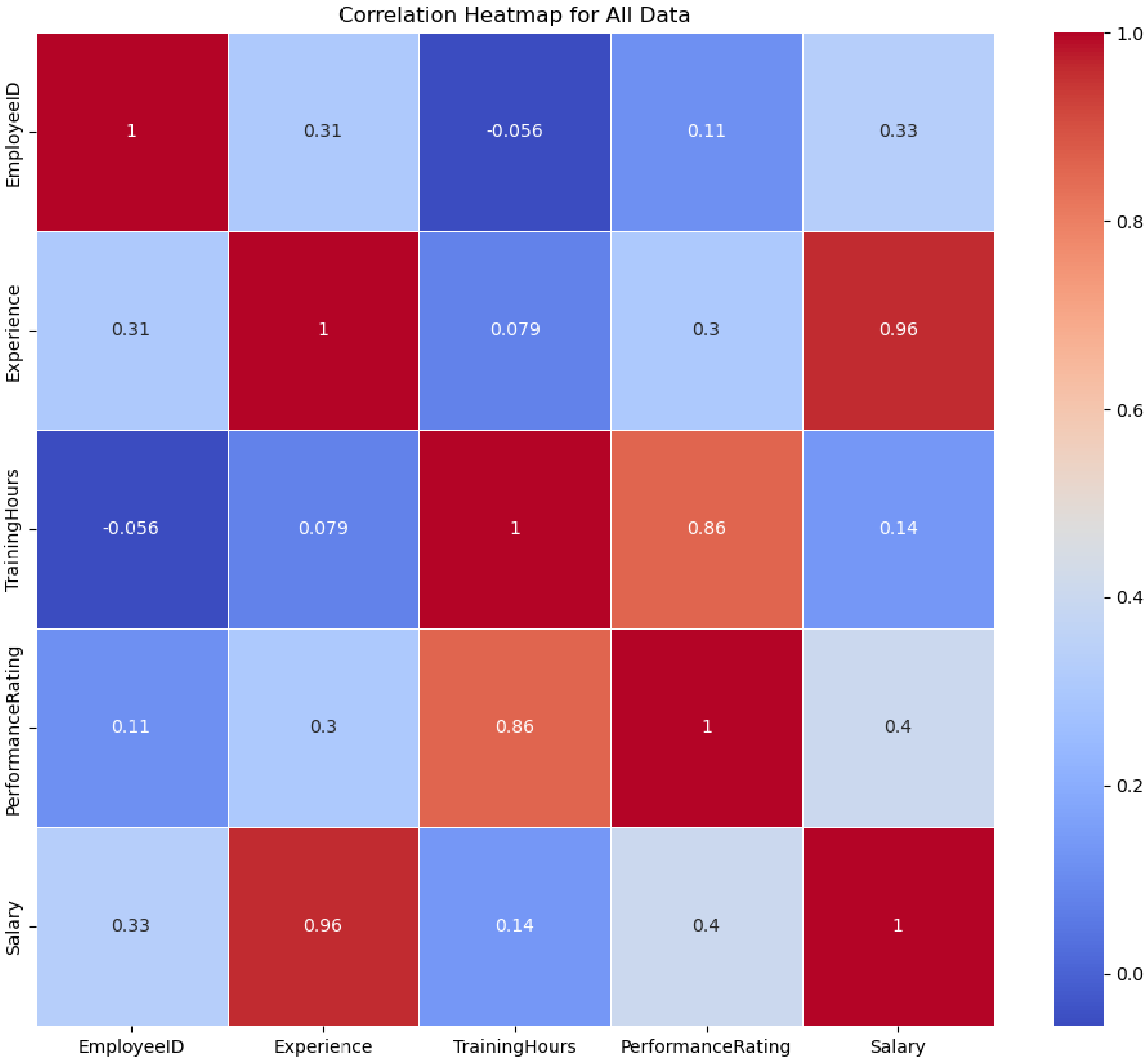


Figure 2.0 – Correlation Heatmap of all variables

Figure 2.1 shows the heatmap of correlation coefficient values between the independent variables. Our attention is drawn to the variables with correlation coefficients equal to or greater than 0.7. There is a correlation coefficient of 0.96 between independent variables Experience and Salary with a significant impact on the target variable (Performance Rating). To mitigate the issues arising from multicollinearity, it is advisable to remove one of these highly correlated independent variables (Experience or Salary). This is crucial because multicollinearity can lead to problems such as unstable coefficient estimates and reduced interpretability of the model.

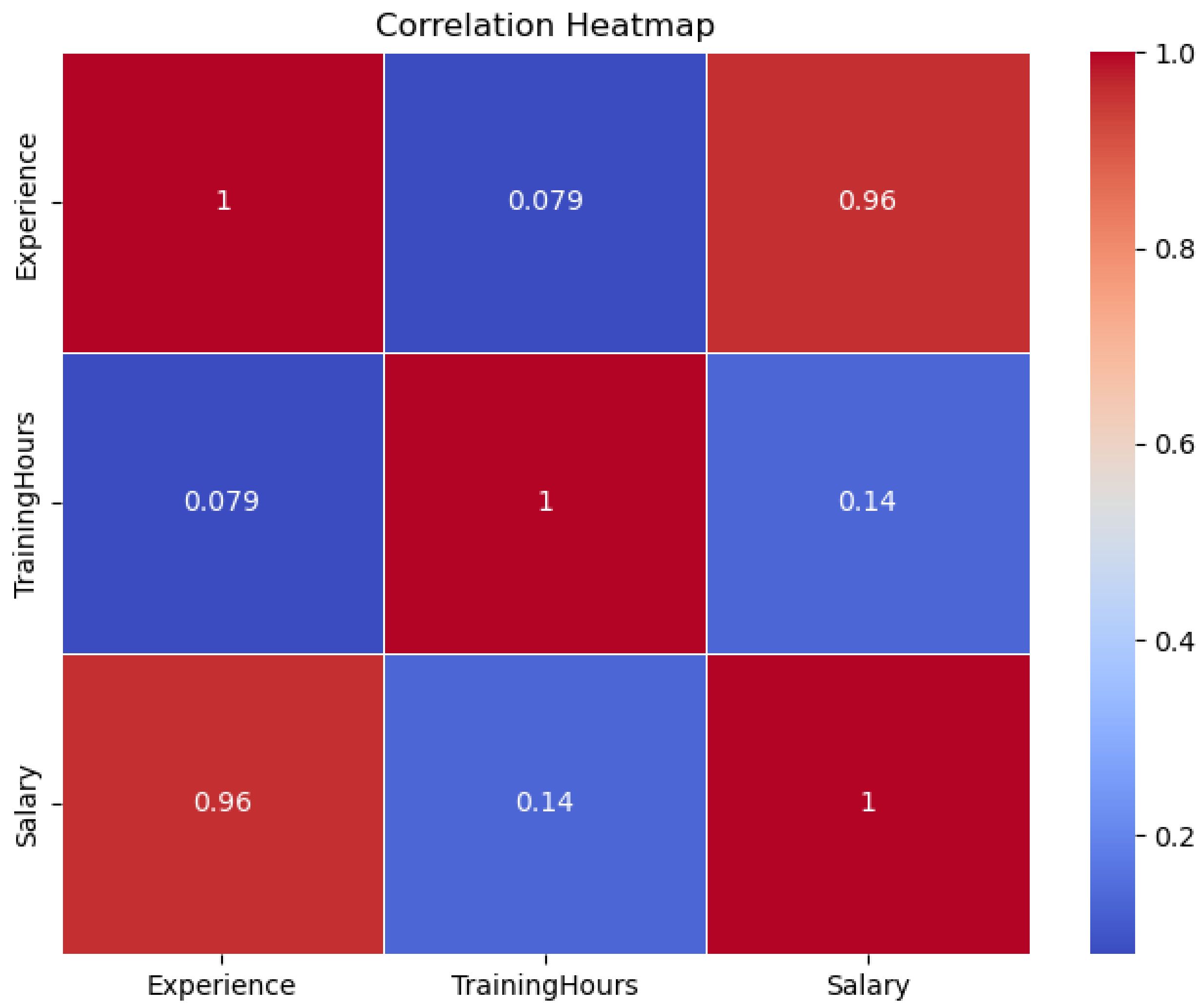


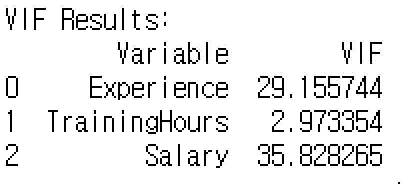
Figure 2.1 – Correlation Heatmap of Independent Variables

As shown in the correlation heatmap of all data on the previous page (Figure 2.0), when looking at the variables and their correlation coefficient values, specifically Experience with Performance Rating = 0.3, Salary with Performance Rating = 0.4, the correlation coefficient value of Salary with Performance Rating is higher. As a result the independent variable (Experience) should be removed.

VIF is a diagnostic tool used to detect multicollinearity in a multiple regression analysis. It helps understand how the variance of the estimated regression coefficients are being inflated due to the presence of correlation among the predictor variables. High VIF values should be addressed to ensure the reliability of the regression analysis.

- VIF = 1: No correlation between the predictor and the other variables.
- VIF > 1 but < 5: Moderate correlation. This is generally acceptable in many cases.
- VIF > 5: High correlation. This is a sign of problematic multicollinearity.

If there are variables with high VIF, it is advisable to consider removing one or more of these variables.



```
VIF Results:
        Variable        VIF
0     Experience  29.155744
1  TrainingHours   2.973354
2         Salary  35.828265
```

Figure 2.3 – VIF Results to test for multicollinearity

The VIF values show each independent variable and how much they are related to each other. In this case, the value for Salary 35.828 and Experience 29.156 have high VIF values. As the VIF value is above the threshold of 5 this is an indication of high correlation and therefore a sign of multicollinearity. In theory, one or all of the independent variables that display signs of multicollinearity should be removed. If both variables are removed only Training Hours will remain as the last remaining independent variable. In this case, this would be an example of Simple Linear Regression and not a Multiple Regression model. Because of this, the Salary variable will be removed which reflects the highest VIF value among the independent variables.

To summarise, through the Correlation Heatmap we were able to identify TrainingHours and Salary as the independent Variables. Through Linearity via the scatterplots we were able to identify Training Hours and Salary as independent variables. Through the VIF test, we were able to identify Training Hours and Experience as independent variables.

Linearity

To check the Linearity assumption in the multiple linear regression model, scatterplots of the independent variables (X) against the dependent variable (y) will be created. These plots will be visually inspected for linear relationships. A linear relationship typically appears as a roughly straight pattern or trend in the data points, where the points seem to follow a specific direction or form a recognizable shape, like a line or a curve. If the relationship is linear, the data points tend to cluster around a line, indicating a proportional increase or decrease in the Employee Performance Rating variable corresponding to changes in either Years of Experience, Training Hours and Salary. However, if the scatterplot shows a scattered, irregular pattern without a discernible trend, this may suggest the absence of a linear relationship.

Figure 2.2 below shows scatter plots which portray the linearity between the independent variable and the target variable. These will show us whether the data points are gathered in a straight pattern or trend and whether the data points follow a specific direction or form a recognizable shape, line or curve. By looking at the scatter plots below we can determine the plot with the most easily recognisable pattern / trend is the Training Hours vs Performance Rating plot. This plot shows signs of having an overall positive linear relationship although the line is not perfectly formed. This matches the values shown in the correlation heatmap where Training and Performance Rating had a correlation coefficient of 0.86. The closer the value is to one, the easier it is to visualise the trend in the scatter plot. The correlation coefficient for Salary and Performance Rating is 0.4 while the correlation coefficient for Experience and Performance Rating is 0.3. making it easier to identify a positive linear relationship in Salary vs Performance Rating when compared to Experience and Performance Rating. Overall all three plots display to some degree a positive linear relationship with Training Hours and Performance Rating displaying the strongest linear relationship.

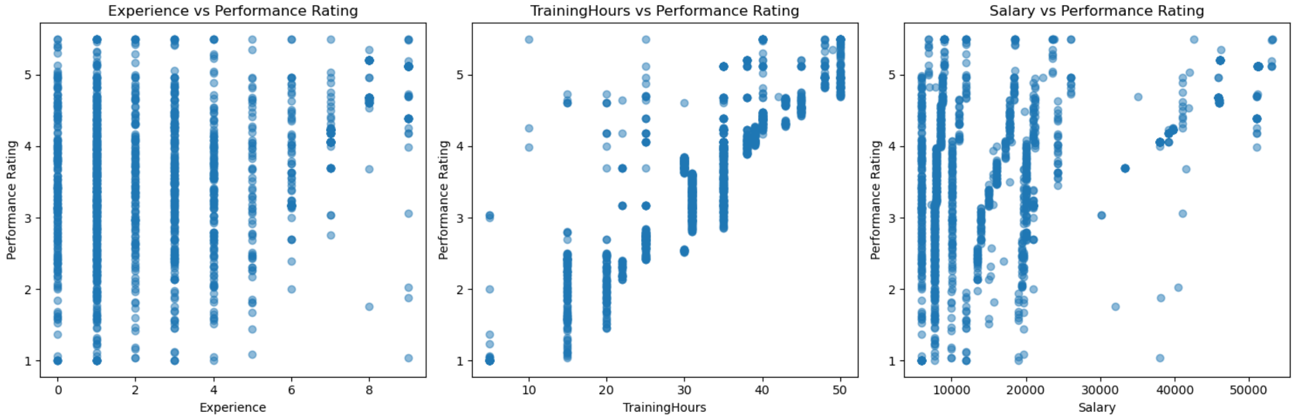
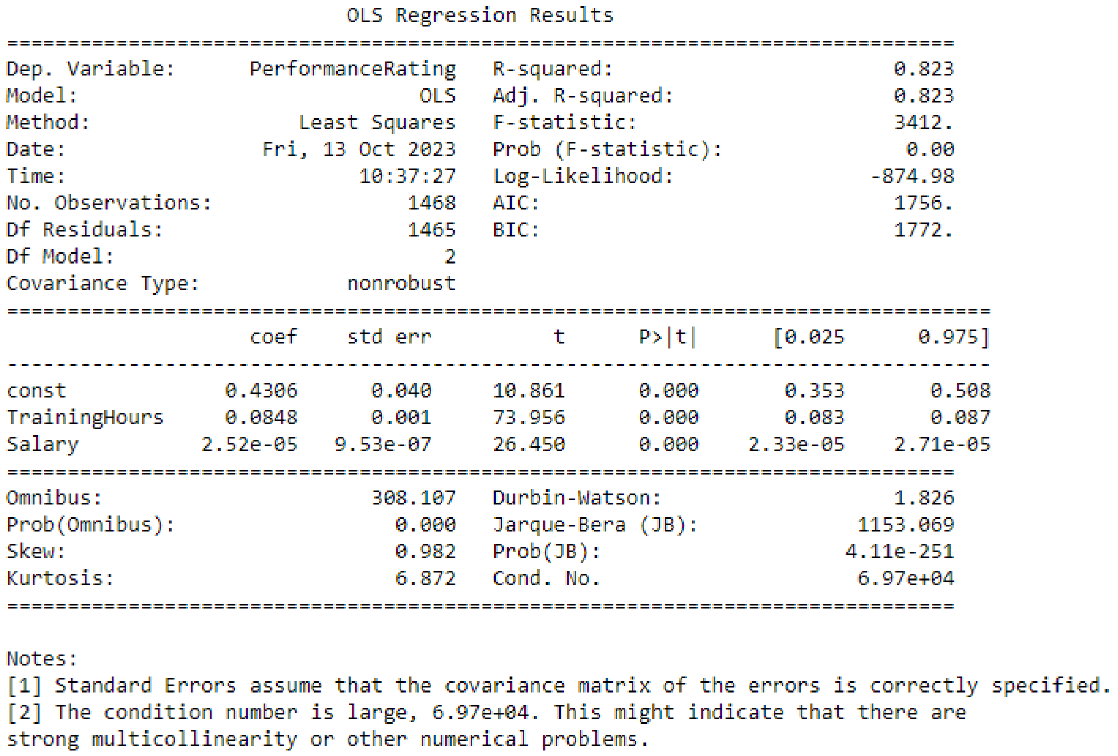


Figure 2.2 – Scatter Plots for each independent variable against the dependent variable

OLS Regression Results

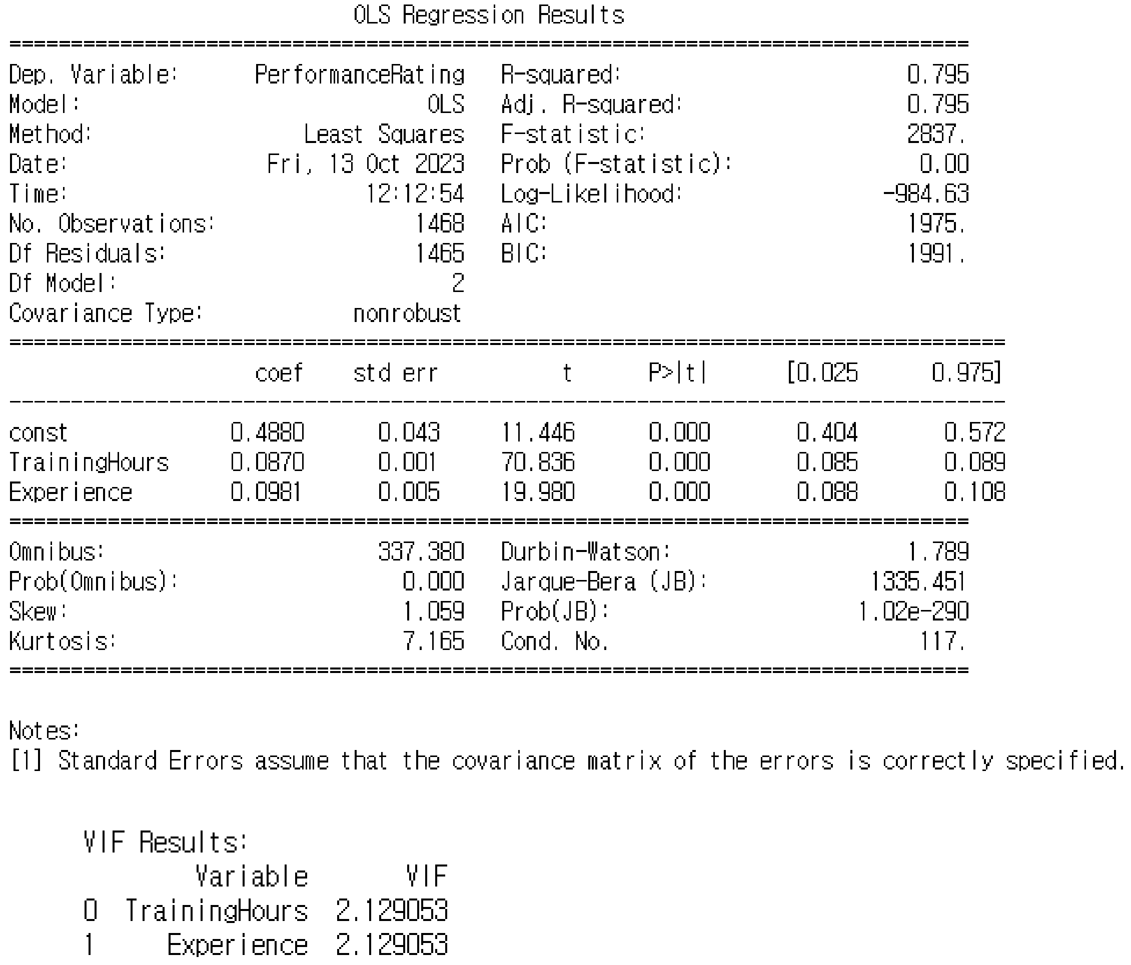
There are two different types of independent variables identified. By using the OLS regression models, we will identify which is more suitable for regression analysis. The OLS models are displayed in the following figure 2.4 and figure 2.5.



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:      PerformanceRating   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.823
Method:                 Least Squares   F-statistic:                     3412.
Date:                Fri, 13 Oct 2023   Prob (F-statistic):               0.00
Time:                        10:37:27   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 2.4 – OLS Regression Results (Training Hours, Salary)

Figure 2.4 shows the large condition number of 6.97e+04 is evidence indicating strong multicollinearity. Based on the evidence presented, the variables Training Hours and Salary are not suitable for multiple regression analysis. The t-statistic value of approximately 10.861 is also good to compare with other OLS model results to identify the independent variables. The OLS model with the higher t-statistic is commonly selected and this is the approach we will take.



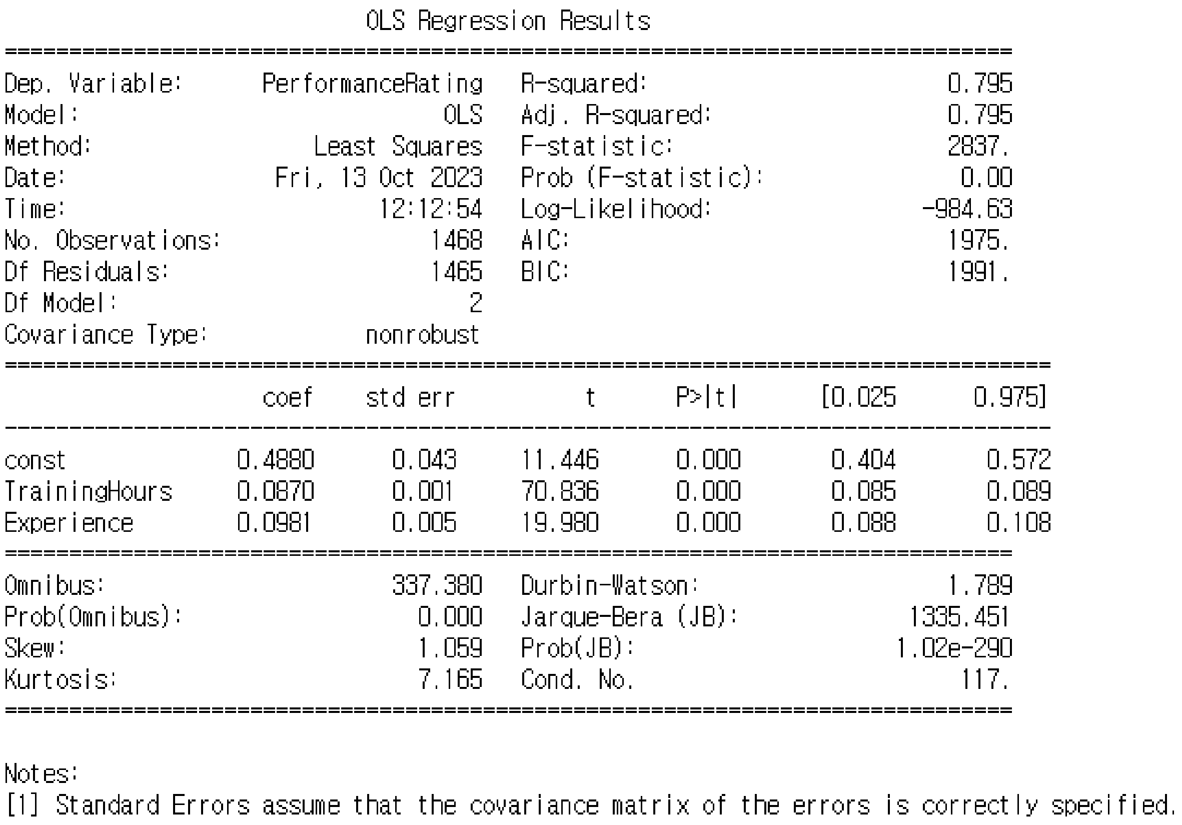
```
                            OLS Regression Results
==============================================================================
Dep. Variable:      PerformanceRating   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.795
Method:                 Least Squares   F-statistic:                     2837.
Date:                Fri, 13 Oct 2023   Prob (F-statistic):               0.00
Time:                        12:12:54   Log-Likelihood:                -984.63
No. Observations:                1468   AIC:                             1975.
Df Residuals:                    1465   BIC:                             1991.
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             0.4880      0.043     11.446      0.000       0.404       0.572
TrainingHours     0.0870      0.001     70.836      0.000       0.085       0.089
Experience        0.0981      0.005     19.980      0.000       0.088       0.108
==============================================================================
Omnibus:                      337.380   Durbin-Watson:                   1.789
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1335.451
Skew:                           1.059   Prob(JB):                    1.02e-290
Kurtosis:                       7.165   Cond. No.                         117.
==============================================================================

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

```
VIF Results:
        Variable       VIF
0  TrainingHours  2.129053
1     Experience  2.129053
```

Figure 2.5 – OLS Regression Results (Training Hours, Experience) and VIF results

Figure 2.5 OLS regression results does not show that this model might indicate a strong multicollinearity through the low condition number of 1.02e-290. The VIF values in each independent variable are lower than before and these values are within reason to be accepted. Based on the results shown and identified in figure 2.5 and figure 2.4, the identified independent variables are TrainingHours and Experience which will be used for multiple regression analysis. Additionally to compare with the previous OLS model with the current t-statistic value of 11.446, the t-statistic in figure 2.5 is slightly higher than the t-statistic value of 10.861 of figure 2.4; therefore, we will pick OLS model in figure 2.5 because it has higher t-statistic which matches the approach we forementioned.

Regression Analysis



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      PerformanceRating   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.795
Method:                 Least Squares   F-statistic:                     2837.
Date:                Fri, 13 Oct 2023   Prob (F-statistic):               0.00
Time:                        12:12:54   Log-Likelihood:                -984.63
No. Observations:                1468   AIC:                             1975.
Df Residuals:                    1465   BIC:                             1991.
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             0.4880      0.043     11.446      0.000       0.404       0.572
TrainingHours     0.0870      0.001     70.836      0.000       0.085       0.089
Experience        0.0981      0.005     19.980      0.000       0.088       0.108
==============================================================================
Omnibus:                      337.380   Durbin-Watson:                   1.789
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1335.451
Skew:                           1.059   Prob(JB):                    1.02e-290
Kurtosis:                       7.165   Cond. No.                         117.
==============================================================================

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

Figure 2.6 – OLS Regression Results (Training Hours, Experience) used for Regression Analysis

The model shown in the regression analysis displays an R-squared of 0.795 which approximately represents 79.5% of the variation in the Performance Rating which can be attributed to the independent variables. The high f-statistic of 2837.0 and the low p-value of 0.00 shows that the overall model is statistically significant which evidently indicates that at least training hours and experience has a substantial correlation to performance rating. The coefficients of training hours and experience, are also both highly significant with a P>|t| (P-values) of 0.000. This means for every unit of training hours, there is an estimated increase of approximately 0.4822 units in performance rating, and for every unit of experience. These are the predictions that can be made based of the relevant coefficients. These results indicate a strong relationship between the chosen independent and dependent variables further enhancing the reliability of the regression model.

Assumptions of Linear Regression

From the analysis, we assume a linear relationship between the predictor variables, Training Hours and Experience, and Performance Rating, being the dependent variable. We also assume that the residuals follow a pattern of normal distribution in regard to the normality of residuals.
The results of the remaining assumptions, Normality of Residuals and Homoscedasticity will be discussed next as they are relevant to the linear regression.

Normality of Residuals

Regarding the check for Normality of Residuals, there are two methods that can check the Normality. The first method is through a Q-Q plot and the second method is through the Anderson-Darling Test. The Q-Q plot is a graphical tool used in statistics that will be able to assess whether the KiwiLearn Dataset follows a pattern of Normal Distribution. It is a visual way to compare the quantiles of the data with the quantiles of theoretical Normal Distribution. If the data points closely follow a straight line it suggests that the residuals are approximately normally distributed. Any deviations from this line could indicate non-normality.

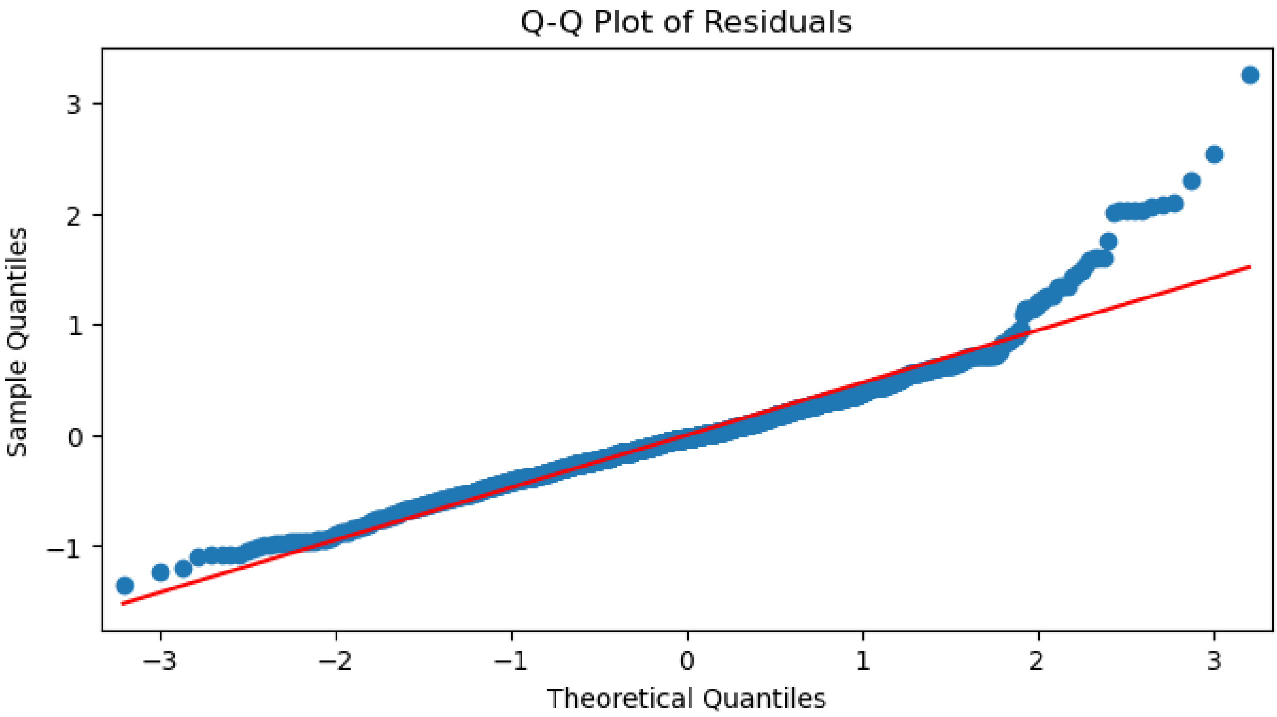
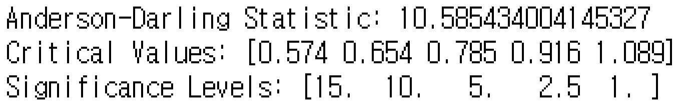


Figure 2.7 – Q-Q Plot of Residuals for visual assessment

Figure 2.7 is a visualisation of the Q-Q plot. The red line represents the quantile residuals of normal distribution and the blue dots represent the quantile residual data points from the KiwiLearn dataset. During the Theoretical Quantiles between around point -3.5 and -2 it is clear that the KiwiLearn quantile residuals strayed away from the red Normal Distribution line. From points -2 to approximately 2, the KiwiLearn data points remained in proportion with the Normal Distribution line however from point 2 onwards, the KiwiLearn data significantly deviated from the path of normal distribution. Because of the deviation of the KiwiLearn data points from the normal distribution line present between around points -3.5 to -2 and point 2 onwards, the Q-Q plot concludes that the residuals are not normally distributed therefore indicating some non-normality present in the KiwiLearn Dataset.

The Anderson - Darling Test (AD Test) will be conducted now. In the AD Test, the null hypothesis states that the KiwiLearn dataset follows a specified distribution which is typically Normal Distribution. The alternative hypothesis is that the KiwiLearn dataset does not follow a specified distribution. In this test, the p-value represents the probability of obtaining a test statistic as extreme (or more extreme) as the one observed under the null hypothesis that data follows a normal distribution. It gives more weight to the tails of the distribution, making it more sensitive to deviations in the tails. The test statistic is compared to critical values to determine whether there is enough evidence to reject the null hypothesis (the assumption of normality). In this hypothesis testing, a small p-value indicates that the observed data is unlikely to come from a normal distribution.



```
Anderson-Darling Statistic: 10.585434004145327
Critical Values: [0.574 0.654 0.785 0.916 1.089]
Significance Levels: [15.  10.   5.   2.5  1. ]
```

Figure 2.8 – Anderson-Darling Statistic output results

Figure 2.8 shows the statistical output results of the Anderson-Darling test. To interpret the results, if the A-D Statistic is greater than the critical value at a specific significance level then the null hypothesis can be rejected. If the A-D Statistic is smaller than the critical value at a specific significance level then we fail to reject the null hypothesis. In this case, the critical value is 0.785 and the significant level is 5. The result of the AD test is 10.59 (2 decimals) which is greater than the Critical Value of 0.785 at a Significance Level of 5, therefore there is strong enough evidence to reject the null hypothesis that states the KiwiLearn dataset follows normal distribution.

It appears that the Q-Q plot and the Anderson-Darling Statistic Test both conclude that there is enough evidence to Reject the Null hypothesis that the KiwiLearn dataset follows a pattern of normal distribution. As both tests reach the same conclusion we can trust the accuracy and base the rest of the analysis upon these findings.

As per the findings of the Q-Q plot and the AD test, when the residuals do not follow a normal distribution, there's a risk of introducing bias into the coefficient estimates. This can subsequently lead to flawed interpretations of the relationship between Years of Experience, Training Hours and Salary (independent variables) and Employee Performance Rating (dependent variable). Due to the deviation from the normality of the residuals, we must be aware that any inferences drawn from the data may be fraught with inaccuracies.

As the assumption of Normality of Residuals was not met there are at least two strategies that can be implemented to address the violation. The first strategy is to do a thorough data cleanse that involves a deep analysis of the outliers. As shown in the Q-Q plot there are outliers present points around -3.5 to -2 and point 2 onwards. Once the data is cleansed it is possible that the residual data points will fit perfectly upon the red line. If the data still does not display normal distribution, the second option after a data cleanse would be to perform a Box Cox transformation or logarithmic transformation on the data. These transformations can help achieve normal distribution.

Homoscedasticity

Homoscedasticity is also known as constant variance. It implies that the variance of the residuals should be constant across all levels of the independent variables. To check for homoscedasticity, we can create a plot of the residuals versus the predicted values (fitted values) from our regression model. A pattern in the plot that fans out or narrows down as the predicted values change can indicate heteroscedasticity (non-constant variance). Opposite to the Q-Q plot, we want the data points to be situated randomly and not following the red dashed line. Data that has no pattern with reference to the red dashed line (y = 0) is an indication of homoscedasticity (constant variance).

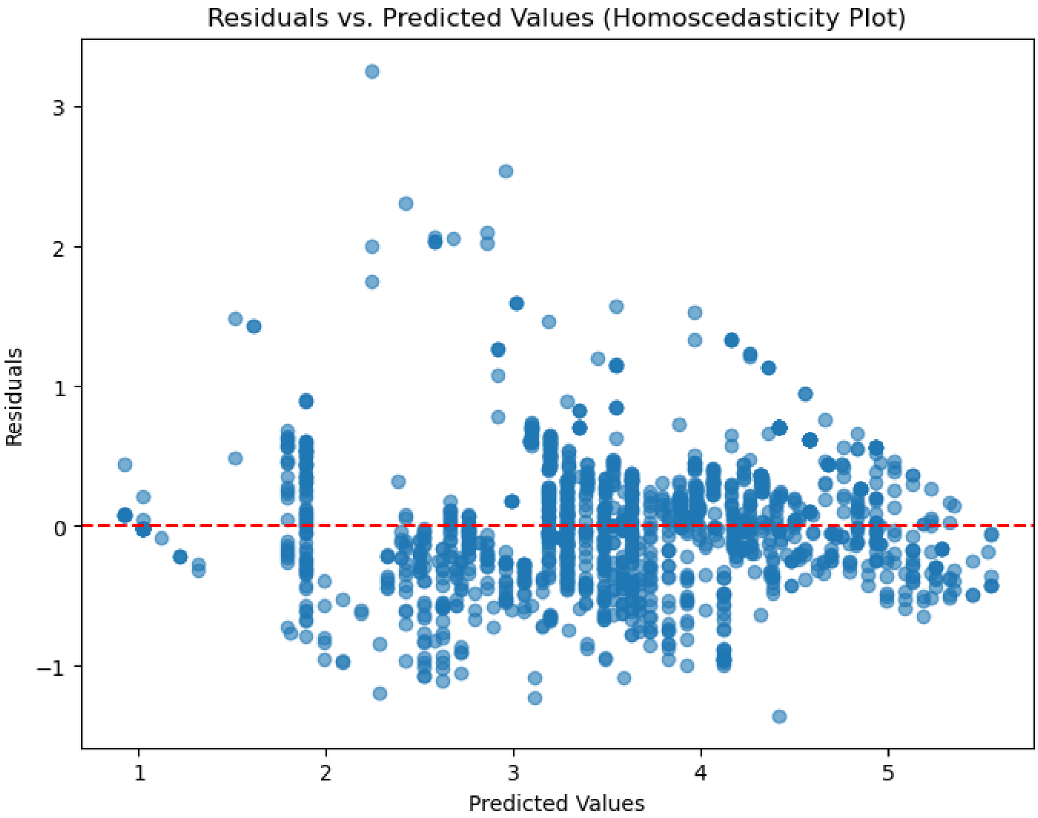


Figure 2.9 – Homoscedasticity Plot showing Residuals vs Predicted Values

Figure 2.9 shows the Homoscedasticity Plot of Residuals vs Predicted Value. The x-axis represents the predicted values (fitted value) from our regression model while the y axis represents the residuals which are the differences between the observed and predicted values of the dependent variable. The red dashed line (y = 0) is used as a reference to help identify any patterns in the residuals. As previously stated we are looking out to see whether the data points are randomly scattered around the red dashed line or whether the data points sit in close proximity to the red dashed line. KiwiLearn data points appear to be randomly scattered through-out the plot with no real connection to the red dashed line. There is no symmetry or distinct shape or pattern of the data points and the red dashed line, which indicates the data exhibits homoscedasticity. Through this homoscedasticity plot we can confidently say that the variance of the residuals is consistent across all levels of the predictors. This consistency in the spread of the KiwiLearn's data points around the mean value indicates that the model's assumption of Homoscedasticity is being met, enhancing the reliability of the regression analysis ensuring no biased standard errors or misinterpretations of the model's fit.

Discussion and Conclusion

Through the regression analysis, we were able to identify some interesting insights about the relationships between Performance Ratings and Training Hours, Salary and Experience. In general, a significant relationship is observed between all independent variables and the dependent variable, despite the presence of multicollinearity issues. The correlation heatmap, scatter plots, and VIF analysis highlight the substantial impact of Training Hours on the performance rating. The regression analysis highlights the strong dependence of performance ratings on both Training Hours and Experience. Specifically the Correlation Heatmap drew attention to Training Hours and Salary as independent variables, while, the analysis of scatter plots confirmed Training Hours and Salary as influential independent variables. Subsequently, the VIF test revealed the significance of Training Hours and Experience as independent variables.

The insights we have gained reveal patterns amongst our identified variables:

- Years of Experience: The positive relationship between Years of Experience and Performance Rating highlights the significance of accumulated experience in enhancing employee performance. This relationship underscores the value of skills developed over an extended period, contributing to improved job performance.
- Training Hours: The notably strong positive association between Training Hours and Performance Rating emphasizes the critical role of training in augmenting employee performance. This finding supports the notion that continuous training efforts can lead to enhanced skill sets and superior job performance, thus highlighting the importance of investing in employee development programs.
- Salary: The positive linear relationship between Salary and Performance Rating suggests that higher compensation levels might be indicative of superior job performance. Additionally, the association between higher salaries and increased performance expectations implies that competitive remuneration could serve as a motivating factor, encouraging employees to deliver exemplary results.

In summary, despite the issues of multicollinearity and real world complexities, our analysis reveals important relationships between training hours, experience, salary and how an employee performs based on performance rating.

Limitations and Improvements

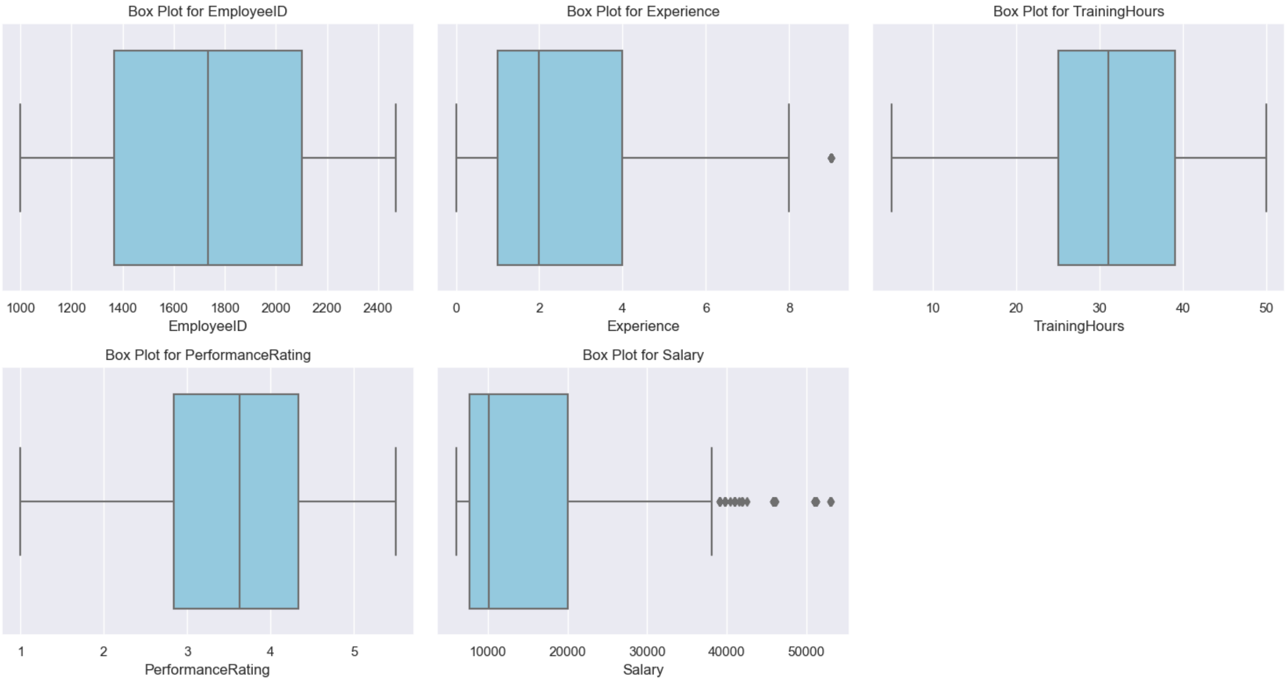
The analysis reveals certain limitations, notably the presence of multicollinearity issues, particularly between the variables of Experience and Salary. To successfully perform the regression analysis, four assumptions were made. These assumptions were: No Perfect Multicollinearity, Linear Regression, Normality of Residuals and Homoscedasticity. The analysis confirms that the assumptions of Normality in Residuals was not met, linear regression and homoscedasticity were met, and no perfect multicollinearity was partially met. The issue of multicollinearity can make the interpretation of the individual coefficients and overall model difficult.

To address the issue, we would have to refine the data aiming for the residual data points that will ideally fit perfectly around the red line. If the data still does not display normal distribution, another option after a data cleanse would be to perform a Box Cox transformation or logarithmic transformation on the data. These transformations can help achieve normal distribution.

There are some opportunities to develop and build a deep analysis to get insight for further research. By switching the independent variables and dependent variable, various results could arise providing additional information to describe each variable's interaction and effects. The analysis could be expanded to take into account external factors such as market condition, hiring season and industry trend. This would result in a deeper analysis to find out more accurate information about interaction between performance rating with other variables.Additionally, collaborating with the human resources team to gather comprehensive information on employee-related factors can provide valuable insights for analyzing performance ratings and related metrics.

Appendices

Figure 1.0 – Summary Statistics of Numeric Data



Box Plot for EmployeeID

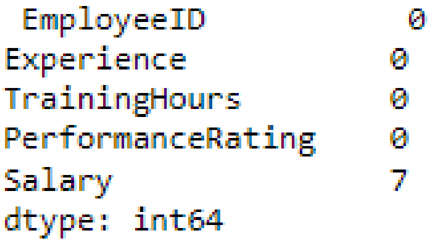
Box Plot for Experience

Box Plot for TrainingHours

Box Plot for PerformanceRating

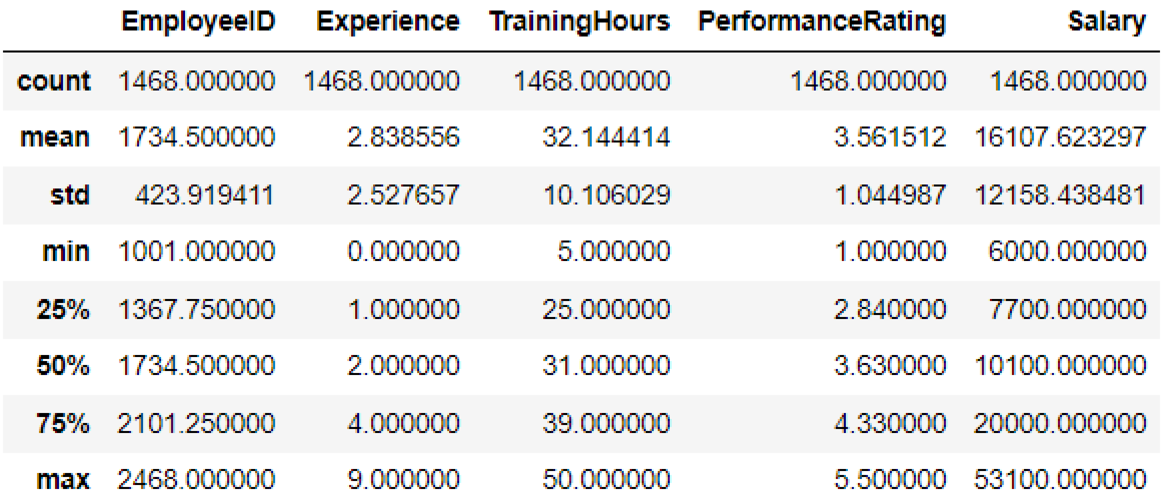
Box Plot for Salary

Table 1.0 – Number of Outliers with a Z-Score threshold of 3



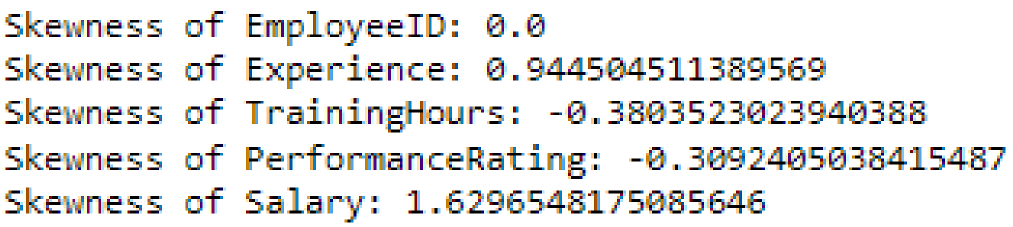
```
 EmployeeID          0
Experience          0
TrainingHours       0
PerformanceRating   0
Salary              7
dtype: int64
```

Table 1.1 – Summary Statistics of KiwiLearn Employees



	EmployeeID	Experience	TrainingHours	PerformanceRating	Salary
count	1468.000000	1468.000000	1468.000000	1468.000000	1468.000000
mean	1734.500000	2.838556	32.144414	3.561512	16107.623297
std	423.919411	2.527657	10.106029	1.044987	12158.438481
min	1001.000000	0.000000	5.000000	1.000000	6000.000000
25%	1367.750000	1.000000	25.000000	2.840000	7700.000000
50%	1734.500000	2.000000	31.000000	3.630000	10100.000000
75%	2101.250000	4.000000	39.000000	4.330000	20000.000000
max	2468.000000	9.000000	50.000000	5.500000	53100.000000

Table 1.2 – Distribution Skew of Numeric Data



```
Skewness of EmployeeID: 0.0
Skewness of Experience: 0.944504511389569
Skewness of TrainingHours: -0.3803523023940388
Skewness of PerformanceRating: -0.3092405038415487
Skewness of Salary: 1.6296548175085646
```

Figure 1.1 – Distribution of Numeric Columns

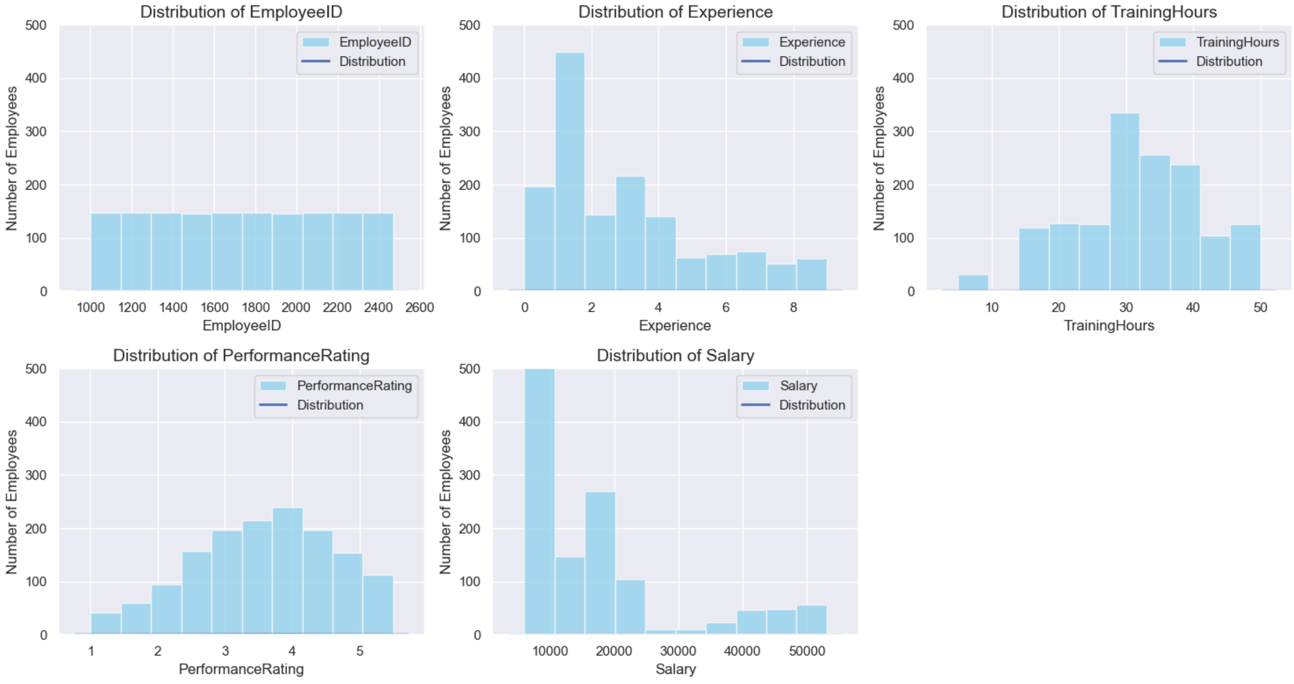


Figure 1.2 – Number of employees in each department

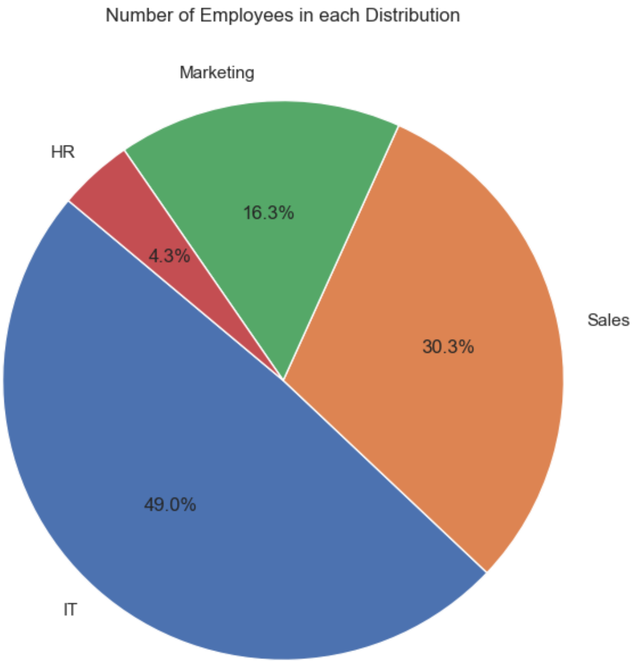


Figure 1.3 – Relationship between Years of Experience and Performance Ratings by Department

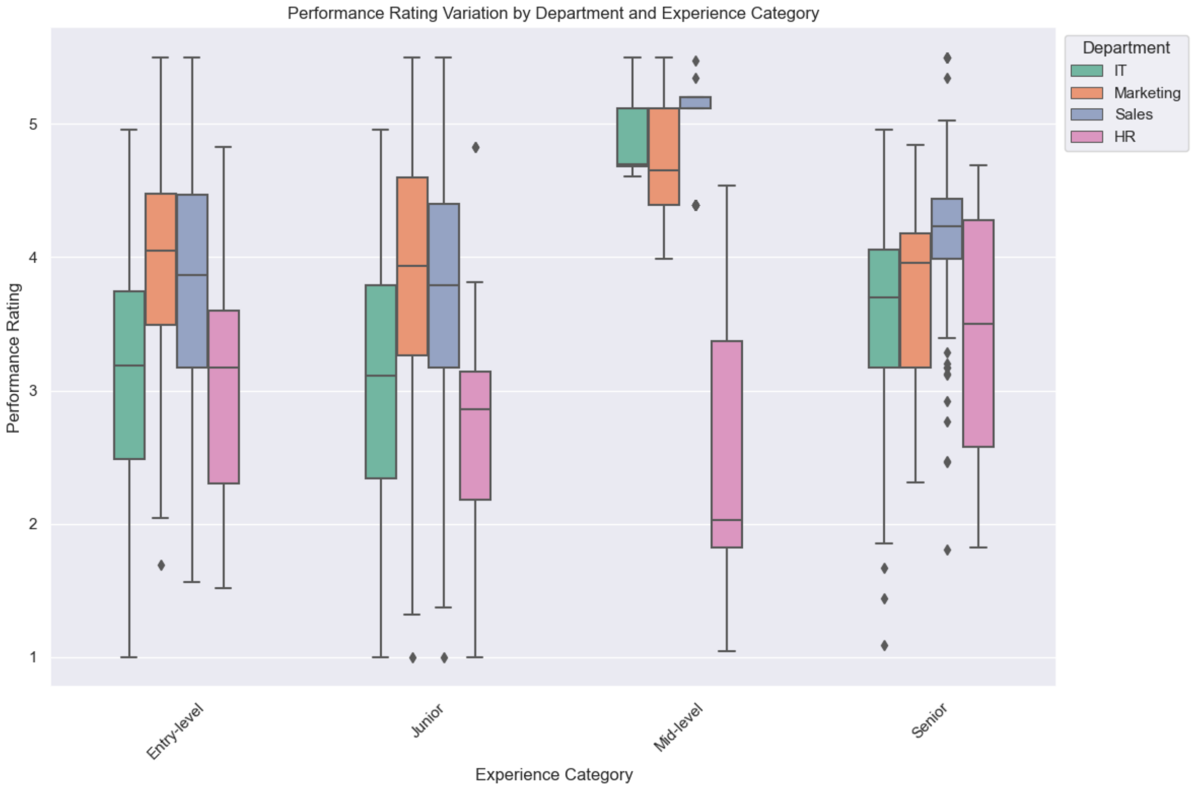


Figure 1.4 – Probability of Distribution of Performance Rating

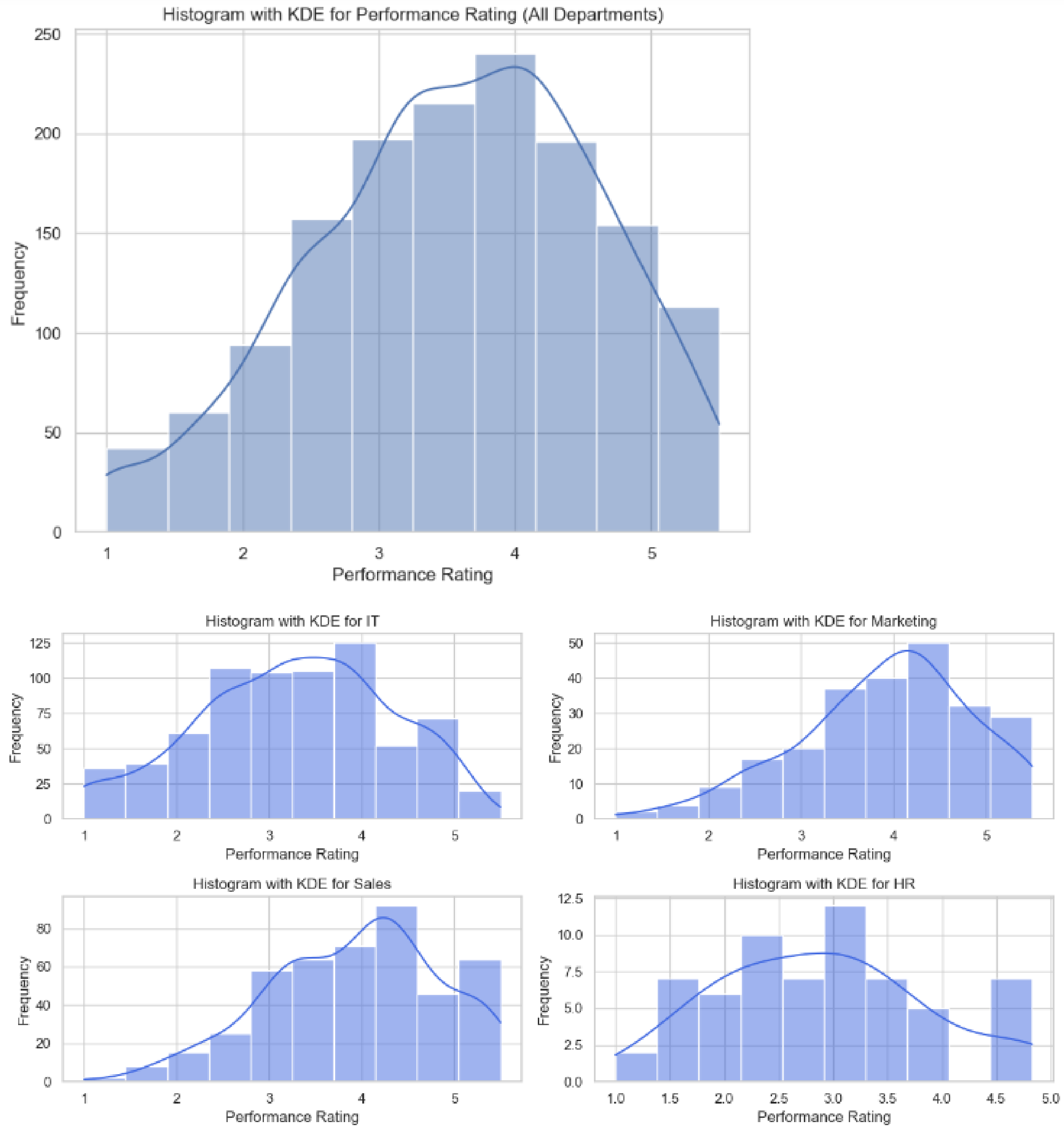
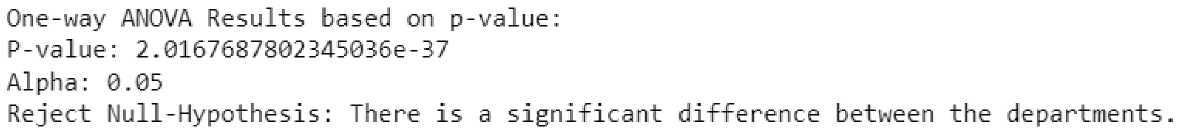
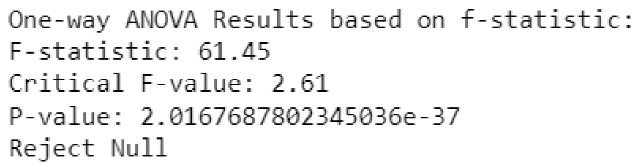


Figure 1.5 – One way ANOVA results based on P-value



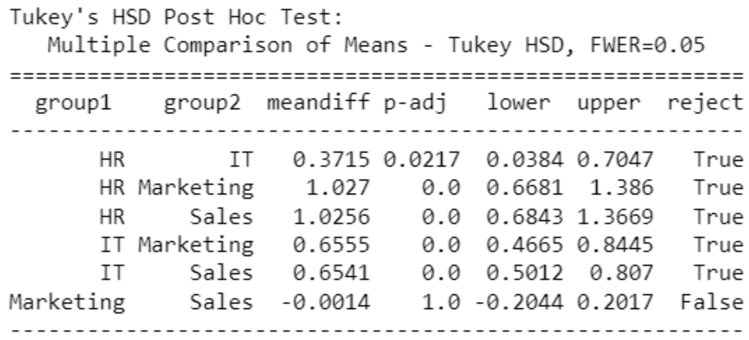
```
One-way ANOVA Results based on p-value:
P-value: 2.0167687802345036e-37
Alpha: 0.05
Reject Null-Hypothesis: There is a significant difference between the departments.
```

Figure 1.6 – One-way ANOVA results based on F-statistic



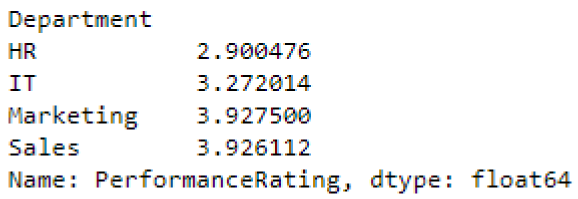
```
One-way ANOVA Results based on f-statistic:
F-statistic: 61.45
Critical F-value: 2.61
P-value: 2.0167687802345036e-37
Reject Null
```

Figure 1.7 – Tukey’s HSD Post Hoc test results



```
Tukey's 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
---------------------------------------------------------
```

Table 1.8 – Mean value of Performance rating per Department



```
Department
HR           2.900476
IT           3.272014
Marketing    3.927500
Sales        3.926112
Name: PerformanceRating, dtype: float64
```

Figure 2.0 – Correlation Heatmap of all variables

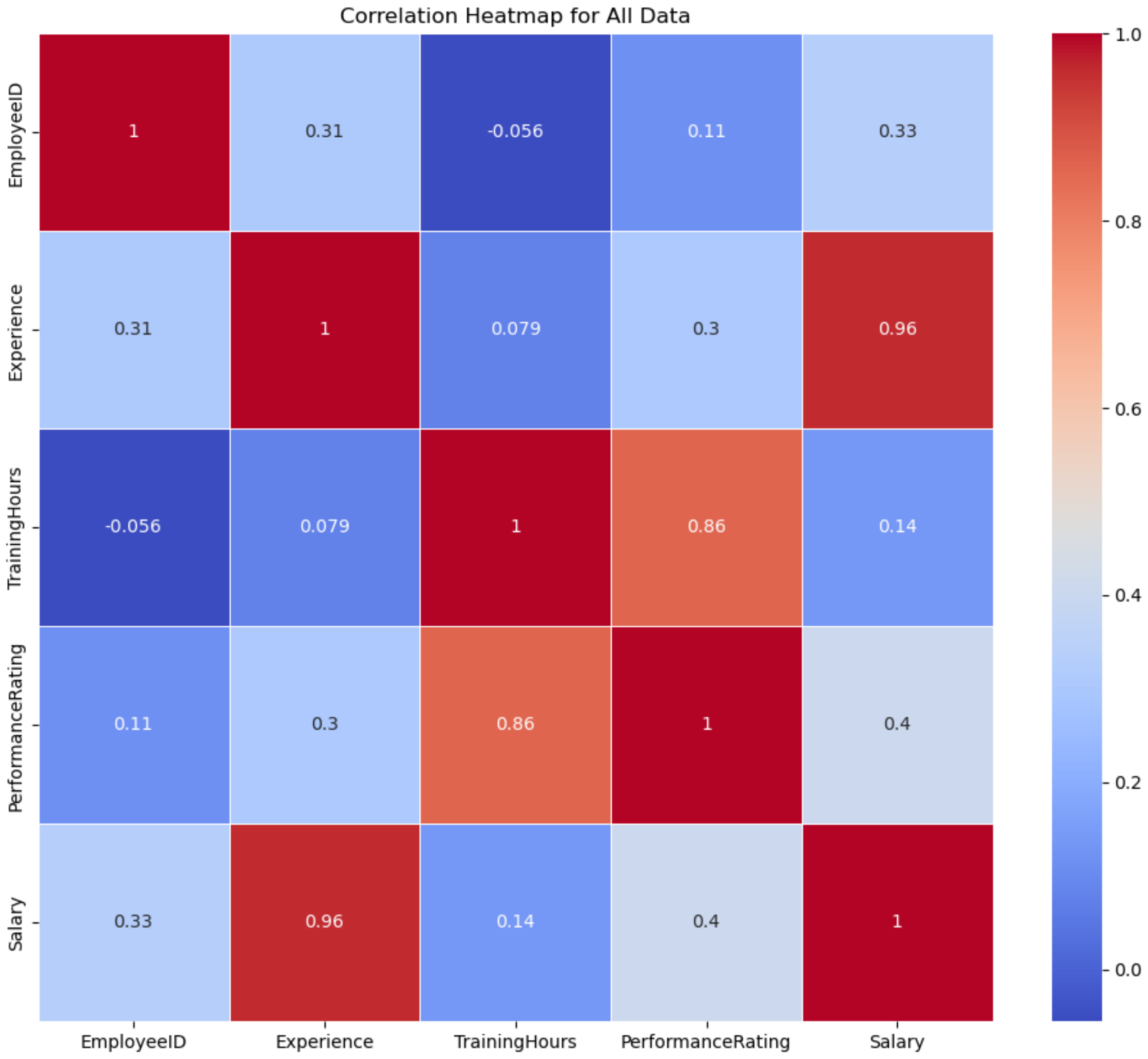


Figure 2.1 – Correlation Heatmap of Independent Variables

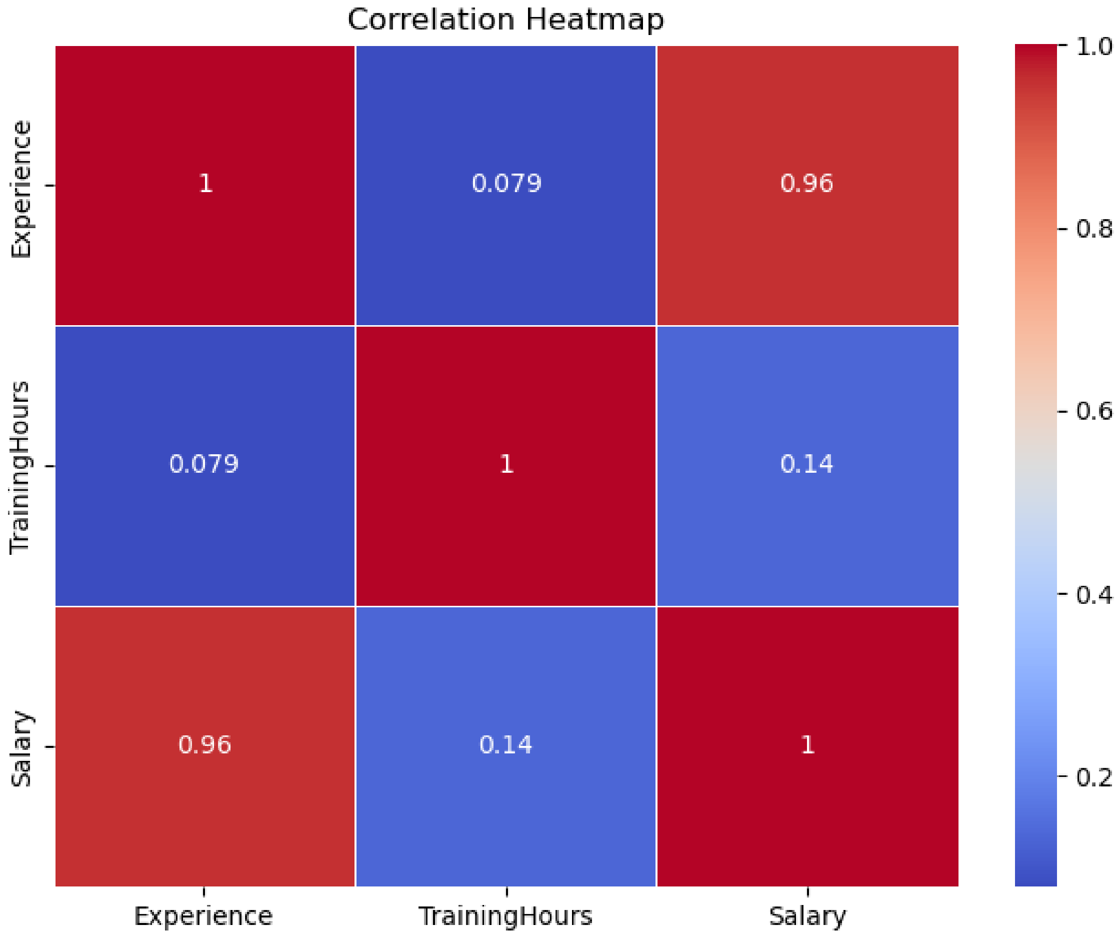


Figure 2.2 – Scatter Plots for each independent variable against the dependent variable

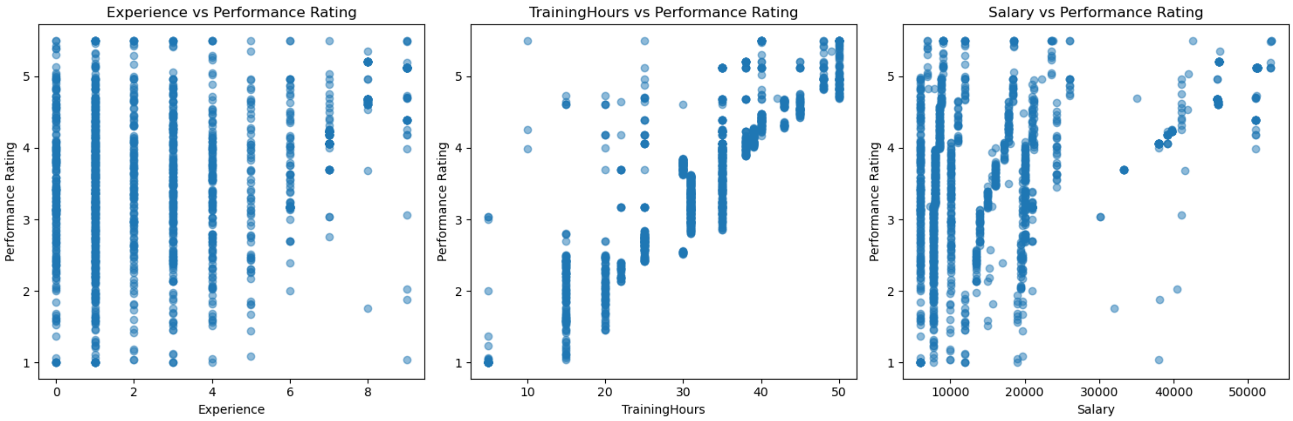
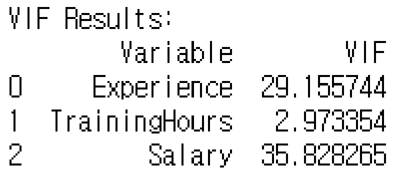
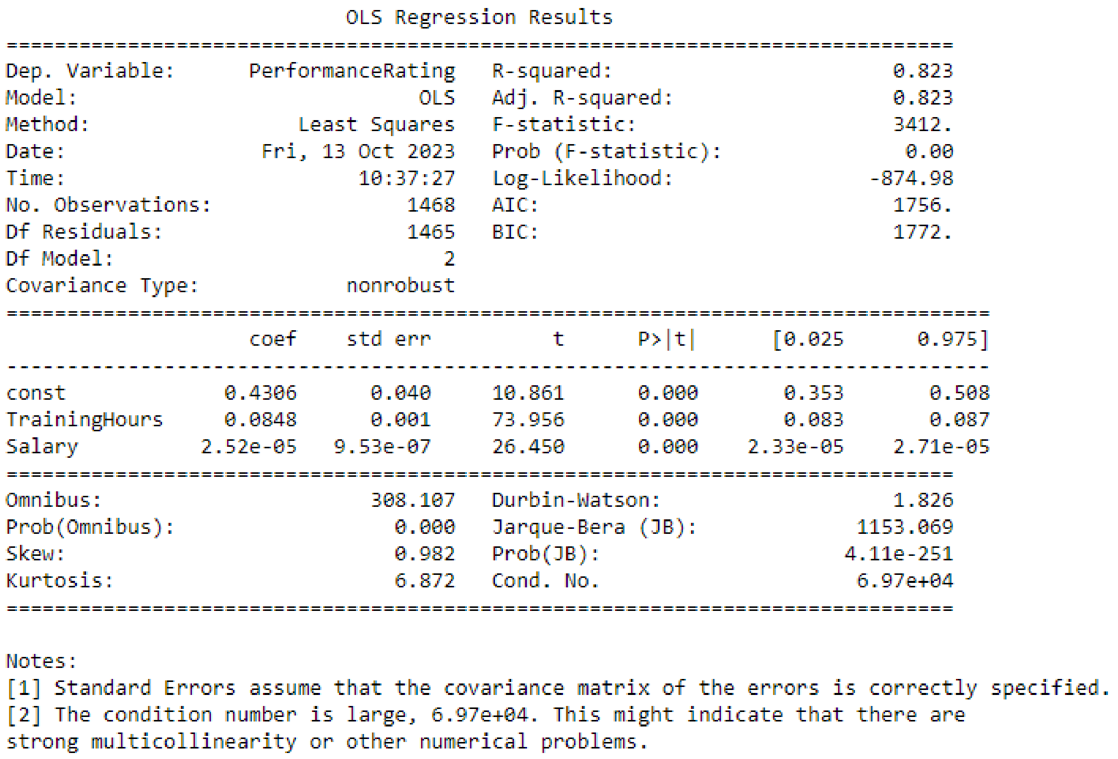


Figure 2.3 – VIF Results to test for multicollinearity



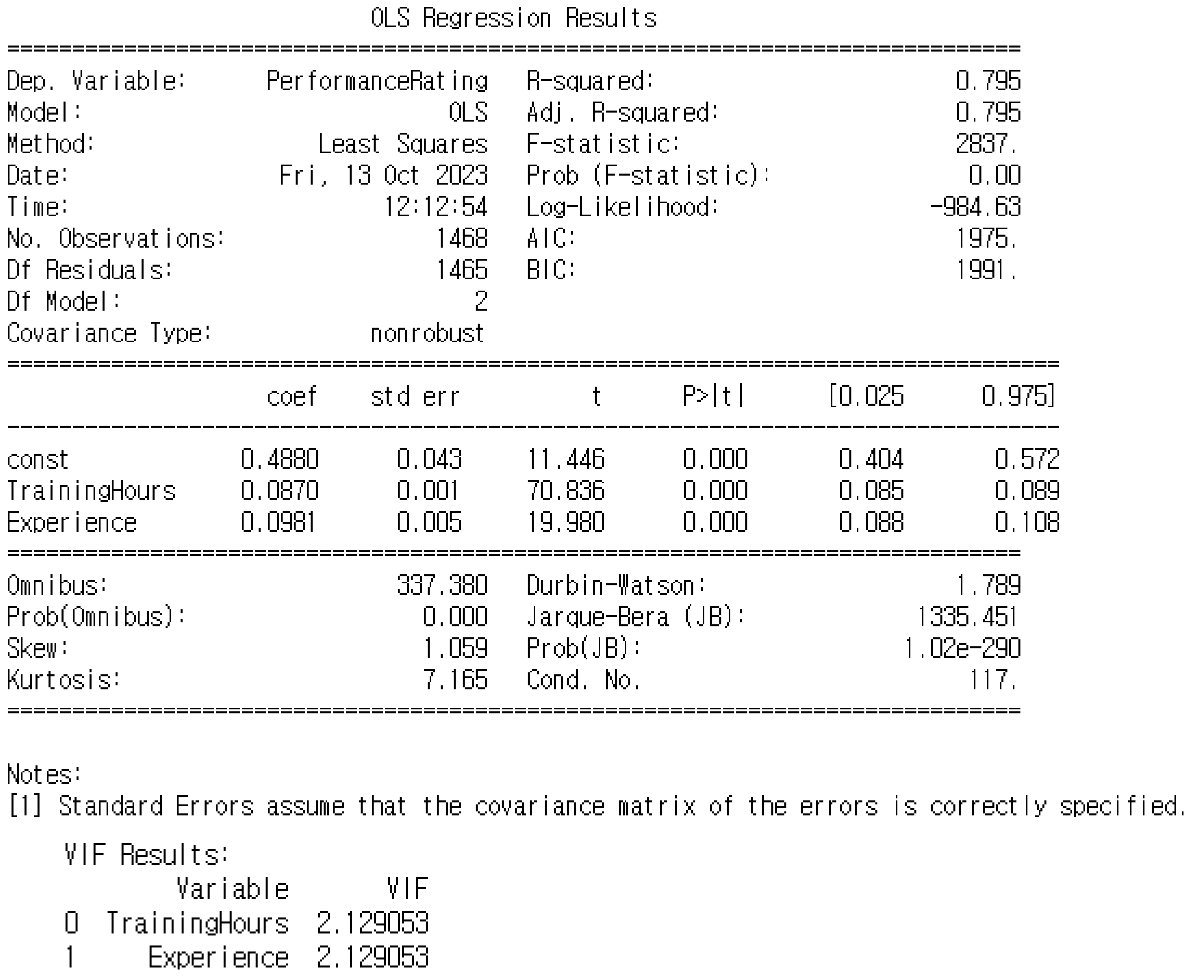
```
VIF Results:
        Variable        VIF
0     Experience  29.155744
1  TrainingHours   2.973354
2         Salary  35.828265
```

Figure 2.4 – OLS Regression Results (Training Hours, Salary)



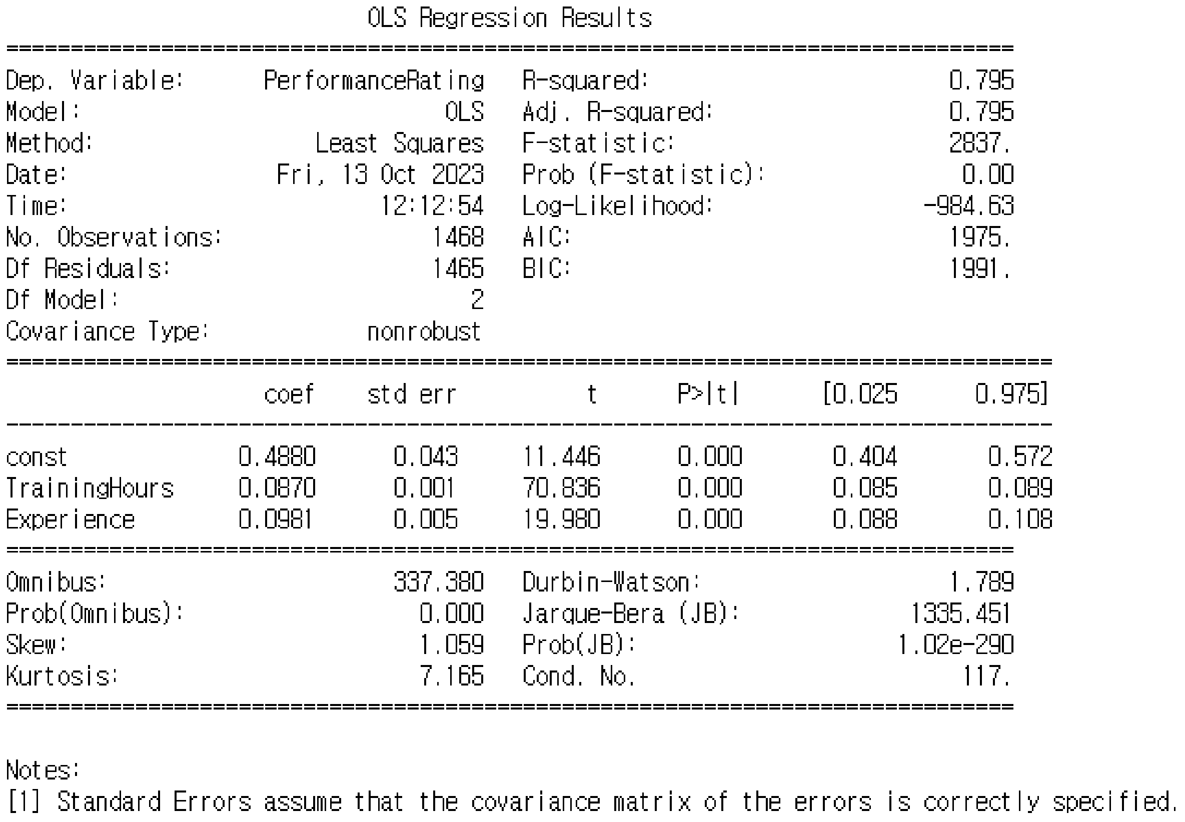
```
                            OLS Regression Results
==============================================================================
Dep. Variable:      PerformanceRating   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.823
Method:                 Least Squares   F-statistic:                     3412.
Date:                Fri, 13 Oct 2023   Prob (F-statistic):               0.00
Time:                        10:37:27   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 2.5 – OLS Regression Results (Training Hours, Experience) and VIF results



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      PerformanceRating   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.795
Method:                 Least Squares   F-statistic:                     2837.
Date:                Fri, 13 Oct 2023   Prob (F-statistic):               0.00
Time:                        12:12:54   Log-Likelihood:                -984.63
No. Observations:                1468   AIC:                             1975.
Df Residuals:                    1465   BIC:                             1991.
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             0.4880      0.043     11.446      0.000       0.404       0.572
TrainingHours     0.0870      0.001     70.836      0.000       0.085       0.089
Experience        0.0981      0.005     19.980      0.000       0.088       0.108
==============================================================================
Omnibus:                      337.380   Durbin-Watson:                   1.789
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1335.451
Skew:                           1.059   Prob(JB):                    1.02e-290
Kurtosis:                       7.165   Cond. No.                         117.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
    VIF Results:
            Variable       VIF
    0  TrainingHours  2.129053
    1     Experience  2.129053
```

Figure 2.6 – OLS Regression Results (Training Hours, Experience) used for Regression Analysis



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      PerformanceRating   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.795
Method:                 Least Squares   F-statistic:                     2837.
Date:                Fri, 13 Oct 2023   Prob (F-statistic):               0.00
Time:                        12:12:54   Log-Likelihood:                -984.63
No. Observations:                1468   AIC:                             1975.
Df Residuals:                    1465   BIC:                             1991.
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             0.4880      0.043     11.446      0.000       0.404       0.572
TrainingHours     0.0870      0.001     70.836      0.000       0.085       0.089
Experience        0.0981      0.005     19.980      0.000       0.088       0.108
==============================================================================
Omnibus:                      337.380   Durbin-Watson:                   1.789
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1335.451
Skew:                           1.059   Prob(JB):                    1.02e-290
Kurtosis:                       7.165   Cond. No.                         117.
==============================================================================

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

Figure 2.7 – Q-Q Plot of Residuals for visual assessment

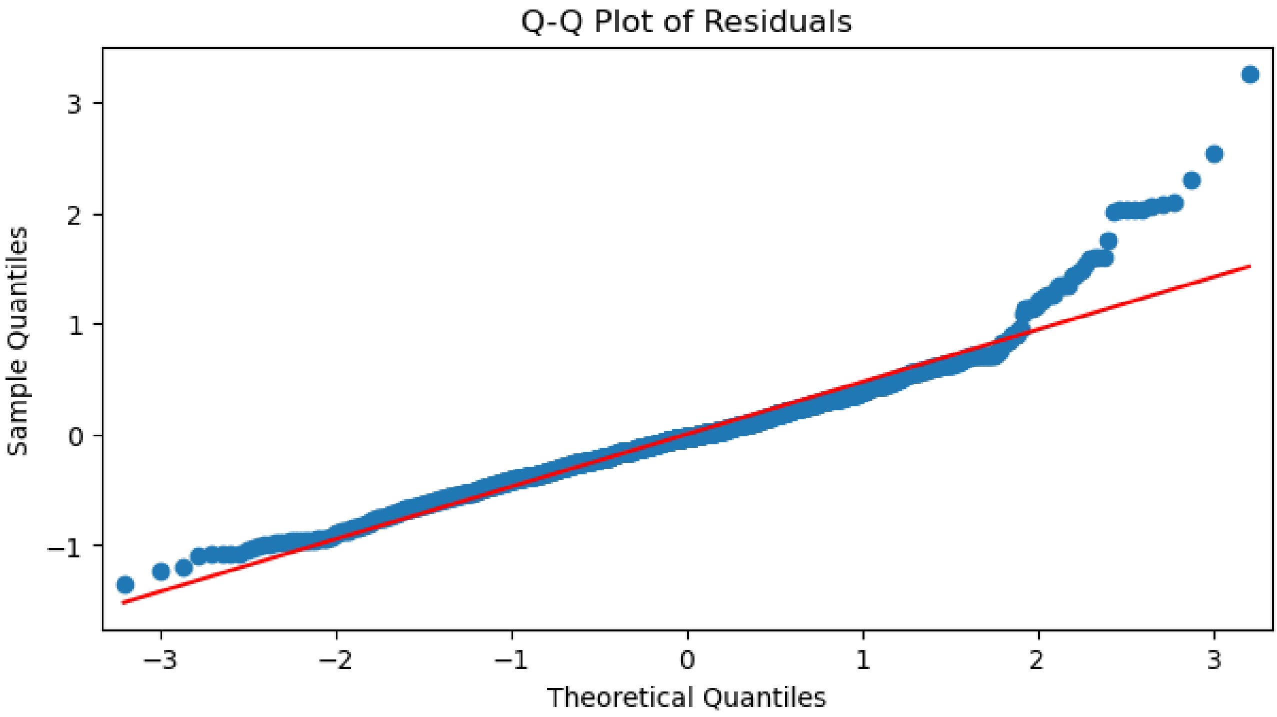
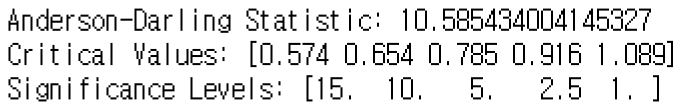
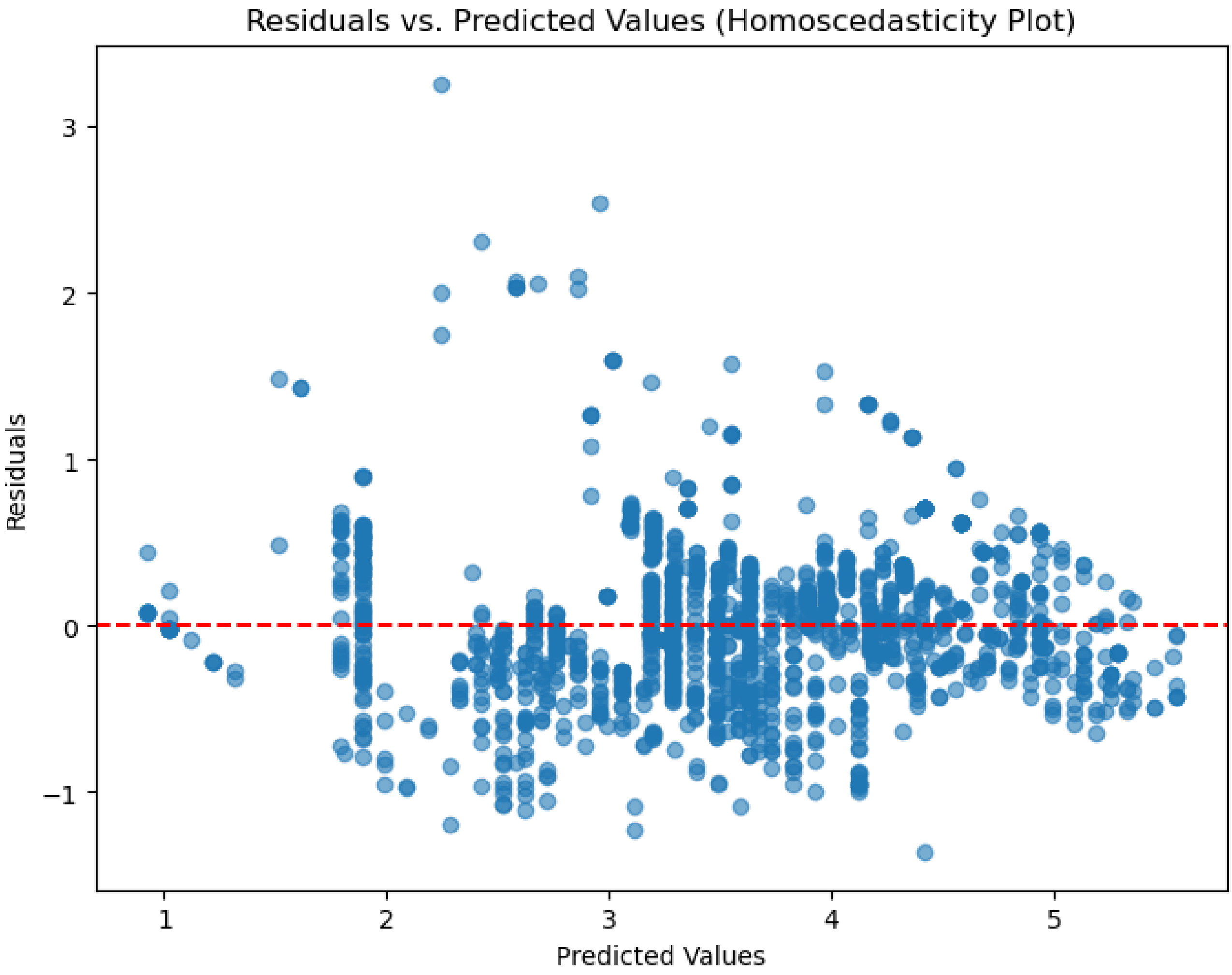


Figure 2.8 – Anderson-Darling Statistic output results



```
Anderson-Darling Statistic: 10.585434004145327
Critical Values: [0.574 0.654 0.785 0.916 1.089]
Significance Levels: [15.  10.   5.   2.5  1. ]
```

Figure 2.9 – Homoscedasticity Plot showing Residuals vs Predicted Values


Residuals vs. Predicted Values (Homoscedasticity Plot)
Residuals
Predicted Values