AUT

Student ID Number: 20121076, 23216497

Assignment 2

Semester 2 2024

PAPER NAME: Data Analysis

PAPER CODE: COMP517

Student ID	Student Names
20121076	Humam Al-Taiff
23216497	Kristian Ortega

Due Date: Midnight Friday 18th Oct 2024
TOTAL MARKS: 100

INSTRUCTIONS:

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

Table of Contents

List of Figures

Part One

Features of the Dataset

```
(1468, 7)
```

Figure 1.1.1: Result of data frame shape command in python

Shape: The number of data points (rows) of the dataset is 1468, and there are 7 variables (columns).

```
   EmployeeID Department  Gender  Experience  TrainingHours  \
0        1001         IT    Male           4              5
1        1002  Marketing  Female           0             50
2        1003      Sales    Male           0              5
3        1004         HR    Male           1              5
4        1005         HR  Female           9              5

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

Figure 1.1.2: Result of data frame head command in python

Head: The features recorded (column labels) Employee ID, Department, Gender, Experience, Training Hours Performance Rating.

```
EmployeeID             int64
Department            object
Gender                object
Experience             int64
TrainingHours          int64
PerformanceRating    float64
Salary                 int64
dtype: object
```

Figure 1.1.3: Data types of Kiwilearn dataset

Data Types: Employee ID, Experience, Training hours and Salary are integer data types, meaning they are only represented by whole numbers. Department and Gender are object data types meaning they are represented by strings. Performance rating is a float data type represented by a number with a decimal. Observing the values in the dataset we can see the float is rounded to the 2nd decimal point.

Cleaning the Dataset

Handling Duplicate Data: Checking for duplicate data, we can see that there are no duplicated rows of data in our dataset.

```
Empty DataFrame
Columns: [EmployeeID, Department, Gender, Experience, TrainingHours, PerformanceRating, Salary]
Index: []
```

Figure 1.2.1: Result of printing duplicate rows, showing a there is an empty data frame

Handling Missing Values: Checking missing values in the dataset, we find that there are none.

```
EmployeeID           0
Department           0
Gender               0
Experience           0
TrainingHours        0
PerformanceRating    0
Salary               0
dtype: int64
```

Figure 1.2.2: Result of printing rows with missing values, showing a there is an empty data frame

Handling Outliers: Checking outliers, we find that there are 7 rows outliers and that they are due to a significantly larger salary than other employees. In discovering these outliers, we decided that we would not transform their values as they represent significant data points to the analysis

```
z_scores=zscore(df['TrainingHours'])
outliers=(np.abs(z_scores)>3)
print(df[outliers])

z_scores=zscore(df['PerformanceRating'])
outliers=(np.abs(z_scores)>3)
print(df[outliers])

z_scores=zscore(df['Salary'])
outliers=(np.abs(z_scores)>3)
print(df[outliers])
```

```
Empty DataFrame
Columns: [EmployeeID, Department, Gender, Experience, TrainingHours, PerformanceRating, Salary]
Index: []
Empty DataFrame
Columns: [EmployeeID, Department, Gender, Experience, TrainingHours, PerformanceRating, Salary]
Index: []
      EmployeeID Department  Gender  Experience  TrainingHours  \
1082        2083         IT  Female           9             35
1189        2190         IT    Male           9             25
1306        2307      Sales    Male           9             35
1338        2339      Sales  Female           9             48
1404        2405      Sales    Male           9             48
1421        2422  Marketing  Female           9             48
1460        2461         IT    Male           9             10

      PerformanceRating  Salary
1082               5.12   53010
1189               5.12   53010
1306               5.12   53010
1338               5.19   53010
1404               5.48   53020
1421               5.50   53100
1460               5.50   53100
```

Figure 1.2.3: Outliers calculated and displayed, showing the outlier data point are sourced from the salary column

Exploring the Clean Dataset

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

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

Figure 1.3.1: Summary statistics of clean dataset

Evaluating the summary statistics of numerical columns, we find that the mean salary is $16,107.62, the mean experience is 2.84 hours, the mean training hours are 32.14 hours, and the mean performance rating is 3.56.

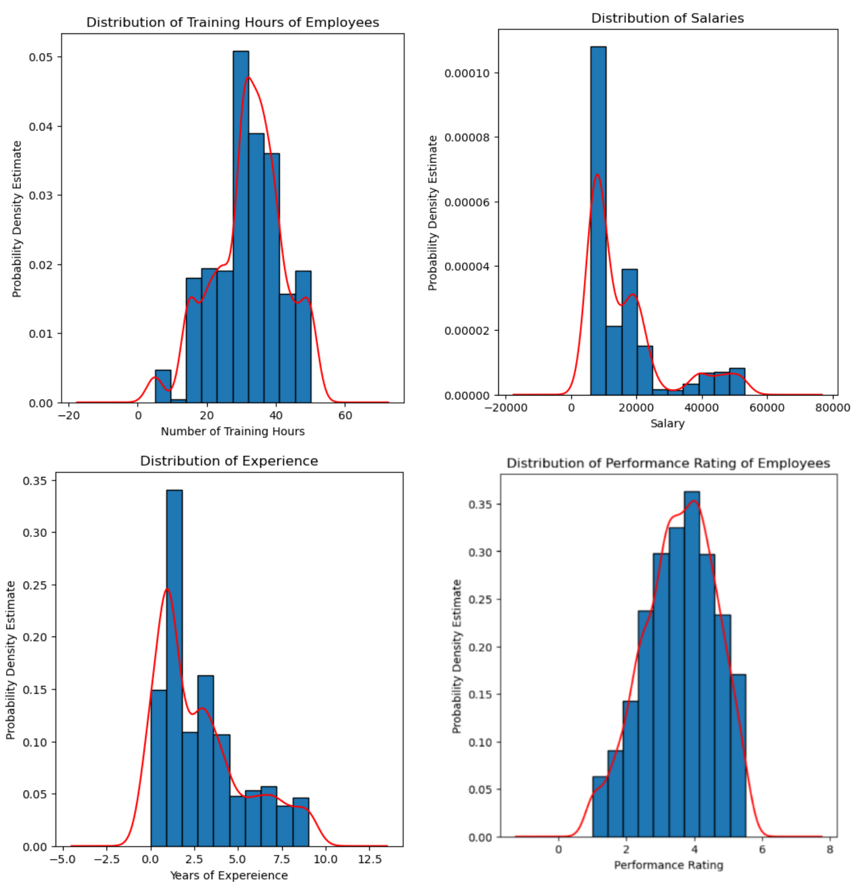

Figure 1.3.2: Histograms with KDE of numerical columns in Kiwilearn Dataset

The distribution of numerical columns we find that Training hours follow a narrow normal distribution, Salaries and Experience follow a positively skewed distribution, Performance rating follows a normal distribution with a slight negative skew.

```
print(df.groupby('Department').size())
print(df.groupby('Department')['PerformanceRating'].mean())
print(df.groupby('Department')['PerformanceRating'].var())
```

```
Department
HR            63
IT           720
Marketing    240
Sales        445
dtype: int64
Department
HR           2.900476
IT           3.272014
Marketing    3.927500
Sales        3.926112
Name: PerformanceRating, dtype: float64
Department
HR           0.957463
IT           1.073170
Marketing    0.873183
Sales        0.862889
Name: PerformanceRating, dtype: float64
```

Figure 1.3.3: Result of code showing the number of data points grouped by department and their mean and variance in performance rating respectively

Looking at the department categorical column we can see that there are vastly different numbers of data points for departments. The average performance rating for different departments is similar (within 1 performance rating point). The variance of performance rating between departments is similar, suggesting that despite the difference in sampling the departments data points are equally spread.

```
Gender
Female    585
Male      883
```

Figure 1.3.4: Number of Female and Male employees at Kiwilearn

Investigating features of the gender variable; we find that there are significantly more male employees than female employees sampled in this dataset.

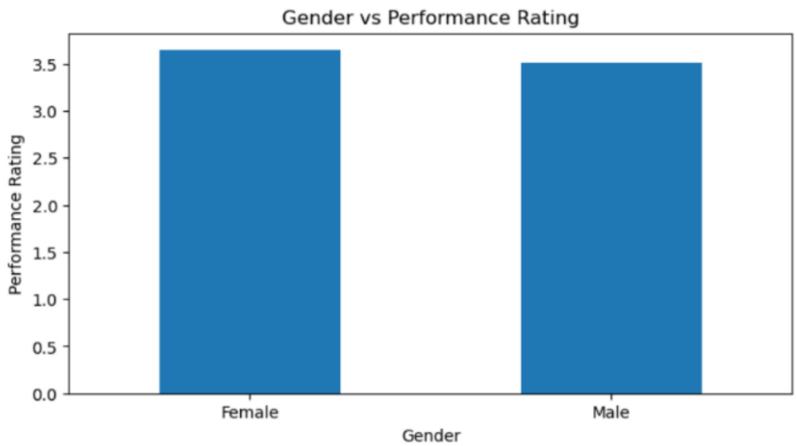

Figure 1.3.5: Bar chart comparing the Performance rating of employees across genders

Evaluating the relationship between gender and performance rating we find that there isn't a significant difference.

Multivariate Analysis

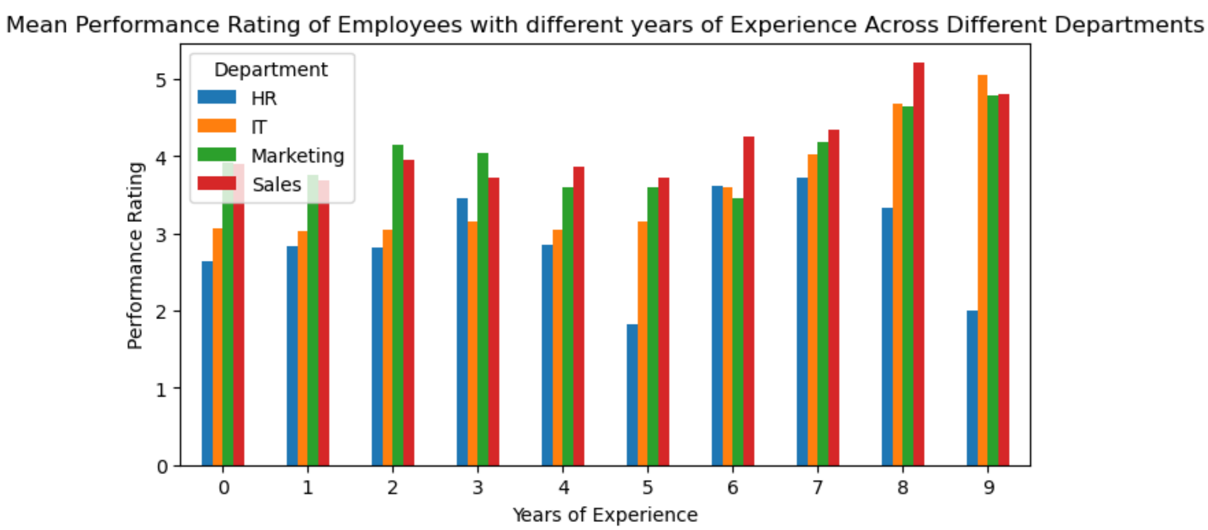

Figure 1.4.1: Grouped Bar chart comparing the mean Performance rating of employees across different departments and years of experience in

From the graph above, you can observe quite a few things and interesting data trends. At a glance, you can tell that employee performance had a general upward trend as the years of experience increased regardless of department.

In regards for the HR department, you can observe that HR employee performance has little to no correlation to years of experience, with HR employee performance, even dipping at 5 and 9 years of experience. The HR department also has the lowest employee performance for greater years of experience among the other departments.

In regards for the IT department, you can observe that it keeps a relatively consistent trend, with employee performance generally increasing with years of experience. The IT department also seems to have more average or modest data compared to the departments

In regards for the Marketing department, you can observe that employee performance starts off strong in the Marketing department, having the highest numbers of all the departments at 0 – 1 years of experience, however employee performance fell between years 2 – 6, then steadily rising again at 7 years of experience.

In regards for the Sales department, you can observe that it generally has the most consistent employee performance to years of experience ratio; with the employee performance having higher ratings from years 6 onwards with a spike at 8 years of experience.

When comparing the departments in terms of years, you can make these observations.

0 – 3 years of experience:

- The HR department has the lowest employee performance across all the departments.
- The IT department's employee performance is quite modest in comparison to the marketing and sales department, but is not the department with the lowest employee performance.
- The marketing department has the highest performance across all the departments.
- The sales department is a close second having the second highest employee performance

4 – 6 years of experience:

- The HR department dips between the 4th and 6th years of experience, with a dip in employee performance at the 5th year of experience then rising back up on the 6th year.
- The IT department has steady employee performance growth between 4 – 6 years of experience, however it does not make any rapid improvements compared to the other three departments.
- The marketing department goes through a dip in employee performance, with it gradiually lowering until 6 years of experience.
- The Sales department has a slow decrease in employee performance between 4 – 5 years of experience then rising quite significantly at the 6th year of experience, outperforming the other three departments.

7 – 9 years of experience

- The HR department's employee performance starts to dip and lowers during the 7 – 9 years of experience, with a significant decrease at the 9th year of experience.

- The IT department employee performance rises rapidly between these three years, even outperforming the other three departments at 9 years of experience.
- The marketing department continues to keep a relatively high employee performance average in comparison to the other departments, though it does start to fall behind the IT and sales departments from 8 years of experience onwards.
- The sales department continues to perform strong, with employee performance peaking at the 8^{th} year of experience.

Objectives

The objective of this case study is to gain insight into the various departments within KiwiLearn and investigate any potential variation in employee performance rating across these departments.

Assumptions

Before reviewing the data and analyzing it; we have made a few assumptions about the data and its collection:

- There are no major external factors (e.g. policy changes or mass lay off only affecting one department)
- There is no bias in the data collection (e.g. the data was not influenced by anyone who favors a specific department)
- Each department has consistent management styles. (So employee performance variation is not due to a strong or weak management style)
- Data collection has no self-selection bias. (We assume that the data collected was not influenced by any personal bias, exaggeration, or downplay of performance.)
- Employee Years of Experience reflect the employee's relevant work history. This is to assume that those with 0-2 years of experience are relatively new to the industry itself rather than KiwiLearn.

Hypotheses

Null Hypothesis: There is no significate variance in employee performance across all departments in KiwiLearn; and any variance could be attributed to random chance and not due to any difference between departments.

Alternative Hypothesis: If there is variation of performance rating across employees then it is linked to or caused by the department that they are working.

Statistical Method

We used ANOVA analysis for this investigation due to ANOVA having numerous pros and benefits. These benefits include:

- Being able to compare multiple groups simultaneously
  - ANOVA allows comparison between more than two groups at once, meaning we could compare all departments at the same time
- Detecting difference across multiple groups
  - ANOVA analysis can detect whether there are any significant variations between groups which could be used to help identify any relating factors
- Help Identify sources of variance
  - ANOVA analysis divides the overall variance into different components which can allow a user to identify the source of any variance in the data.
- ANOVA analysis can be used for further analysis
  - ANOVA can be used as a preliminary step for further statistical analysis.

The purpose of the ANOVA analysis was to find if there was any statistically significant variation between the averages of the HR, IT, Marketing, and Sales departments. We applied this to our analysis by comparing the variances of the four departments in KiwiLearn simultaneously without the need to conduct multiple pair-limited comparisons which could have increased the risk of data and/or comparison errors.

Results

```
One-way ANOVA Results:    Degree of Freedom within the data sets:3
F-statistic: 61.45        Degree of Freedom between the data sets:1464
P-value: 0.0000           Critical F-Value: 8.528336500713385
```

Figure 1.9.1: Results of ANOVA analysis

The ANOVA results we received show some interesting data and information based on inferencing this data.

From our ANOVA results we received an F-statistic value of 61.45; which indicates the variance between the means of departments is significant as it is larger than our calculated Critical F-Value.

From our ANOVA results, we also received a p-value of 0.0000. This p-value is quite smaller than the normal p-value threshold of 0.05. This means we can reject the null hypothesis, indicating there is a statistically significant difference in employee performance based on years of experience between departments in KiwiLearn.

Tukey's Post-Hoc Test

Tukey's post-hoc test is used to assess any significant difference between pairs of group means. This test is generally used to follow up a one-way ANOVA once the F-test shows a significant difference between groups, or in this case, departments. We're using this test to detect the source of any significant difference in employee performance based on specific departments.

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 1.10.1: Results of Tukey's Post Hoc Test

From our Tukey's post-hoc test we got some significant and interesting values such as:

- HR department vs IT department. The average difference is 0.3715 with a p-value of 0.0217; this shows a significant difference in performance between these two departments with HR employees having slightly better performance.
- IT department vs Marketing department. The average difference is 0.6555 with a p-value of 0.0; this shows that the IT employees outperform the employees from the Sales department.
- HR department vs Sales department. The average difference is 1.0256 with a p-value of 0.0; this shows that the employees of the HR department outperform the employees' Sales department.

These results have some interesting implications that suggest the HR and IT departments have a higher performance rating than the Marketing and Sales departments. This suggests that certain factors within these departments may be affecting employee performance.

Discussion

There could be multiple reasons why there is varying employee performance across the departments of KiwiLearn. Some examples of possible reasons could be:

- Different management styles across departments.
    - Different styles of management and leadership within each department could influence employee performance in their respective department. For example, a poor leader could lead to less and stellar performance while a capable leader could lead to higher performance.
- Total employees.
    - Some departments may have more or less employee's than another department, and thus the average of employee performance rating… hmmm, yeah no I think average accounts of different sizes in data huh…
- Nature of the departments work.
    - Since some departments have different tasks, some may have more or less employee performance rating due to factors out of the companies jurisdiction, such as the Sales or Marketing departments being reliant on marketplace factors.

Based on our analysis findings, we can direct Kiwilearn to improve on these things:

- Management and leadership team review. This could reveal any potential poor leadership or poor management practices that negatively affect the department.
- Hold performance rating-centered seminars for departments with lower performance ratings. Departments with higher performance ratings could possibly lead these seminars to get each department up to code.
- KiwiLearn could further examine and investigate the potential reasons for the differences between each department. They could examine quantifiable variables such as IQ, psychoanalytical test scores, or department support (software, financial). To determine what is the source of the disparity.

Conclusion

In conclusion, based on our findings and investigations in the data of KiwiLearn, we can recognize that there is a relationship between performance rating and departments, however, this relation does not necessarily imply causation. The observed relationship between employee performance and their department has some differences, however, further investigation into the relationship between these performance and department via a heatmap, show that these variables are weakly related and in fact using training hours would give a better indication of employee performance rating.

Part Two

Objectives

To identify significant relationships between performance rating and other variables and create a Multiple Linear Regression Model using appropriately related variables.

Selecting Independent Variables

In selecting predictor variables for our multiple linear regression model, we focused exclusively on numerical columns. To account for any correlation between departments and performance rating we transformed Departments to 1,2,3,4 using an ordinal scale (1 is least performance rating, 4 is most). We identified four key variables that are likely to influence performance ratings: training hours, department, experience, and salary. Training hours are expected to enhance performance ratings, as guided learning can lead to improved employee skills. Experience is another critical factor; employees with greater experience tend to feel more competent and confident in their roles, which may translate to higher performance. Additionally, salary could incentivize employees to work harder, thus positively affecting their performance ratings. Department may impact the performance rating of employees as there may be many differences between employees of departments like culture, facilities and job objectives that could raise or reduce performance rating.

We excluded the gender variable, as it does not have a direct numerical relationship with performance ratings. However, it's worth noting that transforming this variable into dummy variables could be a consideration in future analyses. Selecting appropriate predictor variables is essential for ensuring the accuracy and validity of our regression model.

Assumptions

- **Must be linear relationship between the dependent and independent variables**
  - The relationship between the dependent and independent variables must be linear as otherwise our regression model will not fit the predicted trends.
    A non-linear relationship between dependent and independent variables suggests that the two variables are not significantly linked and that using a linear regression model is not accurate at predicting their relationship.
- **Normally distributed error (Residual)**
  - Normally distributed residuals are important in indicating the independence from each datapoint and support the assumption of homoscedasticity. Residuals that are distributed normally indicate that our model captures sources of variance and errors are random and not because of inappropriate modelling decisions.
- **No Multicollinearity**

  - Predictor variables should not be related to each other. Multi collinearity is important to avoid confounding factors in our model, enhancing its accuracy.
- **Homoscedasticity**
  - Homoscedasticity is important in a linear regression analysis as it indicates that the model is well-defined, meaning that the dependent variable is adequately defined by the predictor variable. If there is too much variance in the residuals then it would indicate that the independent variables are not well defined and thus, are not relevant to the analysis.
- **The variance of the residuals must be constant across predicted variables**
  - The variance of the residuals should be constant to ensure that each residual data point does not affect each other (essentially homoscedasticity). This would allow for accurate modelling, with limited bias.

Testing Multi-Collinearity and Linearity

Before doing the multiple linear regression, we should test for multi-collinearity between independent variables using a correlation index heat map.

Also, we should test for linearity between our independent variables and performance rating (our dependent variable.

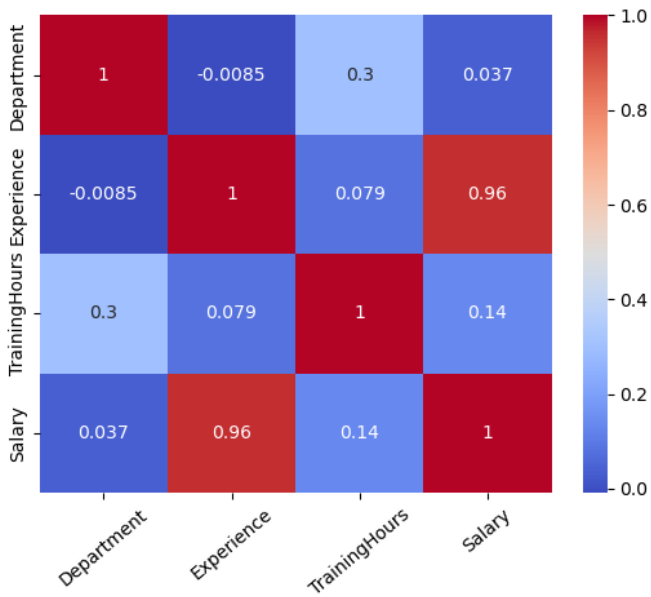

Figure 2.3.1: Heat map of potential predictor variables

In our heat map we can see there is a strong co-linearity between the Salary and Experience independent variables. Thus, for our analysis we should only consider using one of these variables and not both.

Figure 2.3.2: Scatter Plot of potential predictor variables

In our scatter plots we can see that the variables with the strongest linear relationships are Training hours and Experience. The 'Departments' category does not have an apparent linear relationship. Training hours is the closest to a linear relationship with the dependent variable while Experience is more scattered. Therefore, our independent variables for multiple linear regression analysis should be TrainingHours and Experience.

Multiple Linear Regression Analysis

```
Regression 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:                Wed, 16 Oct 2024   Prob (F-statistic):               0.00
Time:                        12:39:50   Log-Likelihood:                -984.63
No. Observations:                1468   AIC:                             1975.
Df Residuals:                    1465   BIC:                             1991.
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             0.4880      0.043     11.446      0.000       0.404       0.572
Experience        0.0981      0.005     19.980      0.000       0.088       0.108
TrainingHours     0.0870      0.001     70.836      0.000       0.085       0.089
==============================================================================
Omnibus:                      337.380   Durbin-Watson:                   1.789
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1335.451
Skew:                           1.059   Prob(JB):                    1.02e-290
Kurtosis:                       7.165   Cond. No.                         117.
==============================================================================

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

Figure 2.4.1: Results of Multiple linear OLS regression

R^2 value: Our value is 0.795 which indicates a high linear relatability to the dependent variable. i.e., 79.5% of the variance in the dependent variable can be explained by the independent variables, training hours and experience (our predictor variables are closely related to our dependant variable).

F-statistic: Our F-statistic in our OLS results is remarkably high. This indicates that our independent variables are significantly tied to our dependent variable; performance rating.

Coef: Our results for our coef suggest that training hours and experience affect the growth of performance rating at the same rate albeit experience affects performance rating slightly more represented by a higher coef.

Std err: The OLS results indicate that we have a small standard error for our variables suggesting that we have precise estimates for our variables.

t-value: A large t-value insinuates significant relationships. Here we can see that the training hours may present as a more related variable to performance rating than experience.

P>|t| Value: Values less than our assumed alpha level 0.05 indicate that our results are significant and support rejecting our null hypothesis. In this case we can see the results of our independent variables mean that the relationships are significant.

Skew: Our skew value, of 1.059 indicates that there is a positive skew to our residuals.

Kurtosis: Our kurtosis indicates potential outliers, and a deviation of residuals from normality.

Omnibus Test: Our low Omnibus p-value (less than our significance value) indicates that the null hypothesis can be rejected, suggesting that the residuals are not normally distributed.

Durbin-Watson Statistic: Our Durbin-Watson statistic is slightly less than 2, which suggests our residuals are slightly correlated.

Jarque-Bera Test: Our probability of our Jarque-Bera Test indicates that the residuals do not follow a normal distribution (as it is smaller than our significance value).

Significance

Our Regression analysis results indicate a strong correlation between our independent variables, Training Hours and Experience, and our dependent variable, performance rating. As represented by our R^2 value. Similarly, our analysis demonstrates that our results are certainly significant, meaning we can reject our Null Hypothesis and confidently state that Training Hours and Experience significantly affect the performance rating of employees at Kiwilearn. When we look at the degree of the effect, we can see that Experience will influence the performance rating at a slightly higher rate than training hours. This is illustrated by examining the co-efficient of the variables in our OLS test. Also, as they are both positive co-efficients, we can see that they have a positive linear relationship with performance rating. Meaning as training hours and experience increases the employee's performance rating increases. Our standard error being low supports the validity of our analysis, as it means there was little room for error in the analysis.

However, the results of our analysis indicate that we do not meet some of our assumptions, invalidating our findings. Considering this, we should analyze our results further using different tests to verify their quality.

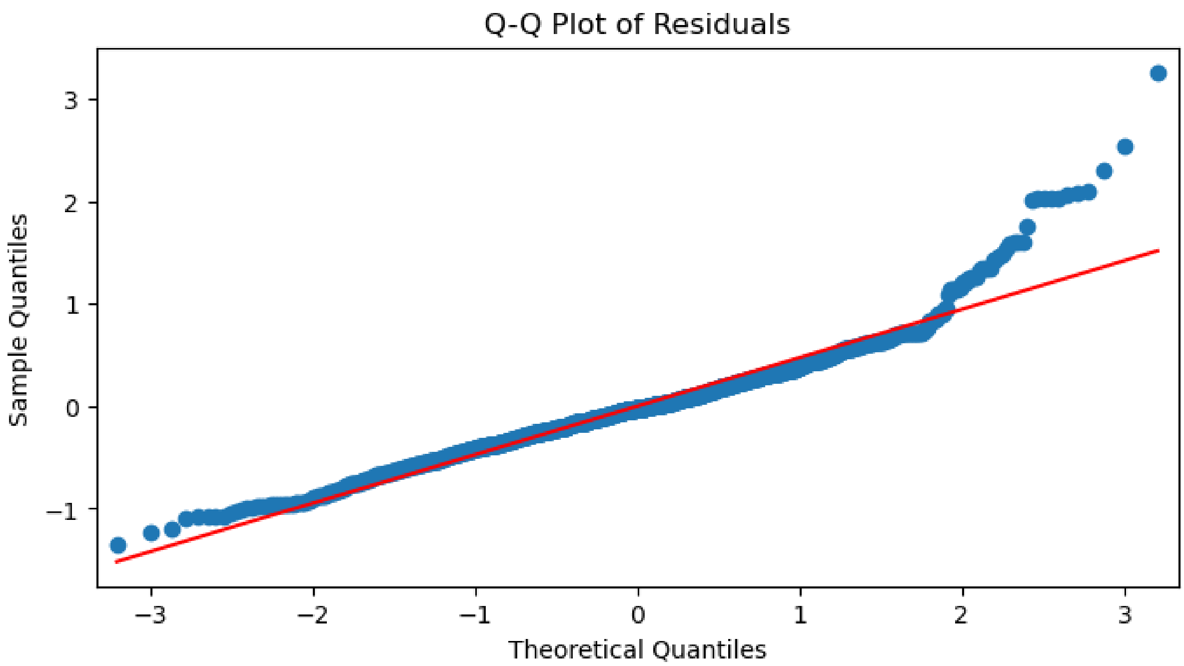

Figure 2.5.1: Q-Q plot of residuals of our OLS Multiple Linear Regression model

<u>Evaluating our Assumptions</u>

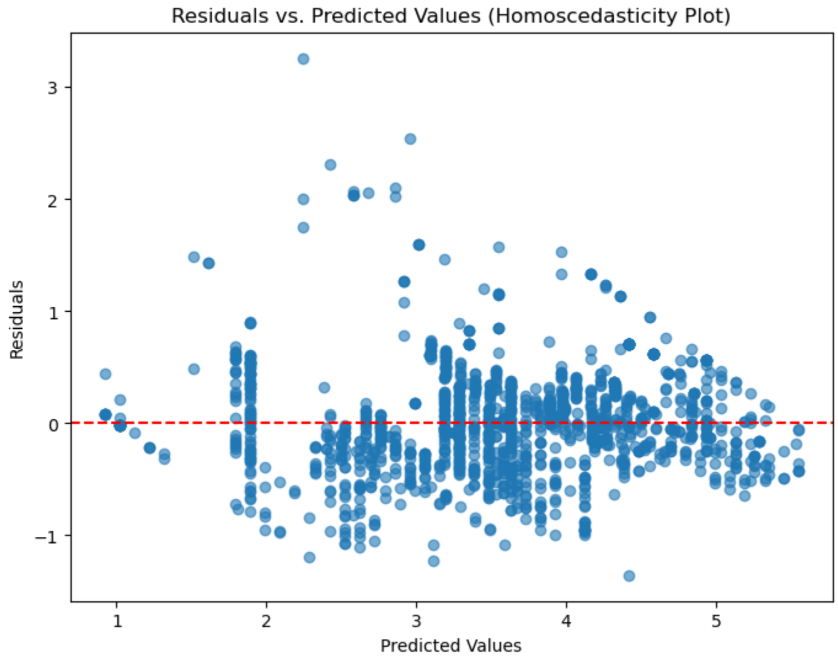

Figure 2.6.1: Residuals vs Predicted values plot

Looking at the homoscedasticity plot, the residuals appear random, indicating that they are distributed without systematic patterns. However, the magnitude of residuals at the 2-3 range of predicted values is significantly larger than at other levels, suggesting a violation of the assumption of constant variance of residuals.

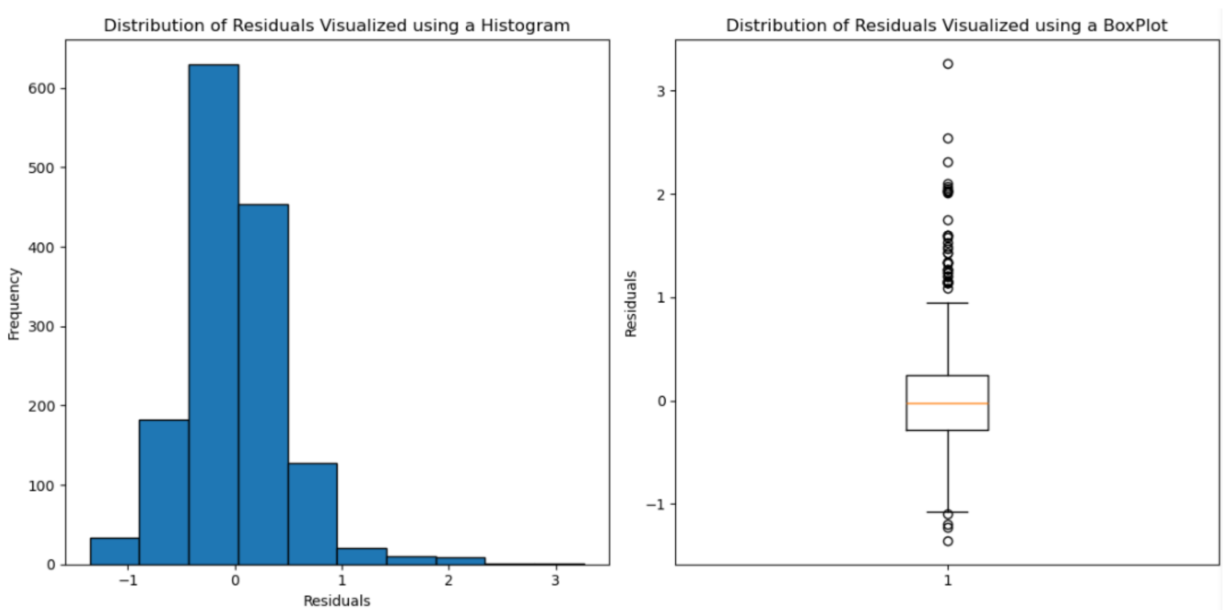

Figure 2.6.2: Histogram and Boxplot of residuals, demonstrating their distribution

This trend is more clearly illustrated in the histogram and boxplot of the residuals, where we observe a positive skew. This indicates potential heteroscedasticity, where the variance of the residuals varies across levels of the independent variables, suggesting that other factors may be influencing the results of our regression.

```
Anderson-Darling Statistic: 10.5854340041451
Critical Values: [0.574 0.654 0.785 0.916 1.089]
Significance Levels: [15.  10.   5.   2.5  1. ]
```

Figure 2.6.3: Anderson-Darling Statistic, critical values at different significant levels

Additionally, the residuals do not appear to be normally distributed. The Anderson-Darling statistic being larger than the critical value supports this conclusion, indicating that the residuals do not follow a normal distribution, which can affect the validity of hypothesis tests and confidence intervals derived from the model.

The QQ plot further illustrates these issues, as deviations from the regression line at the extremes suggest that the residuals deviate from normality, reinforcing concerns about heteroscedasticity. Overall, the presence of heteroscedasticity and non-normally distributed residuals raises significant concerns about the reliability and accuracy of our regression model, suggesting that predictions may be less reliable, and that further investigation or model adjustments may be necessary.

Overall, the presence of heteroscedasticity and non-normally distributed residuals raises concerns about the reliability and accuracy of our regression model, suggesting that predictions may be less reliable, and that further investigation or model adjustments is necessary.

Improvements to our Assumptions

To improve upon meeting our assumptions we should:

- **Look to deconstruct generalized variables –** This is because omitted variables may influence the distribution of variables without us knowing or being able to account for the influence
- **Use different modelling methods –** Using a variety of modelling methods like Weighted Least squares may aid in establishing residuals with a more constant variance and better distribution.
- **Address Outliers –** Outliers may disproportionately influence our data set and while we did not end up using the salary column in our analysis, the data points of those outliers were kept and may have influenced the distribution of our residuals giving it the positive skew we observe.

- **Consider transforming Variables -** We could apply transformations to our dependent variable or independent variables to help achieve normality in residuals. E.g. transforming performance rating into a logarithmic or exponential scale or doing this for training hours.
- **Request for an increased sample size**: a larger sample size can help in achieving normality due and assist in creating better reliability and validity to our results.

Conclusion

In conclusion, our results demonstrate a significant correlation between the independent variables (training hours and experience) and the dependent variable (performance rating), validating part of our objective by identifying key factors that impact performance ratings. However, the findings are limited by issues related to the assumptions required for a Multiple Linear Regression Model. Specifically, concerns about heteroscedasticity and potential under-sampling within our population undermine the model's accuracy. Consequently, we are unable to achieve our secondary objective of developing a reliable model for predicting performance ratings. Future research should address these limitations to enhance the robustness of the analysis.

Limitations

There are several limitations in this study that may introduce error or bias. First, the lack of transparency regarding data collection methods prevents us from addressing potential biases that could influence our findings. Additionally, the use of generalized variables may exacerbate issues with homoscedasticity. Without clarity on how performance ratings are calculated, our understanding of their relationship with other variables is limited.

Several potential biases remain unaddressed, including sample bias, availability bias, reporting bias, and omitted variable bias. Sample bias arises from the significant differences in employee samples across departments. Availability bias may occur if data collection relied solely on a single database, neglecting other relevant variables. Reporting bias could stem from data gathered through employee surveys. Furthermore, the presence of omitted variable bias is suggested by the observed heteroscedasticity, indicating that important variables may be missing or need further exploration. Finally, without insight into the calculation of performance ratings, we cannot identify other variables that may be associated with them.

Future Research

Future research should focus on several key areas to enhance the veracity of our findings. First, it would be beneficial to deconstruct the training hours variable to better understand its impact on performance ratings. Additionally, gaining insight into the calculation of performance ratings is crucial, as this will help identify other variables that may influence them. Researchers could also explore using Weighted Least Squares (WLS) or alternative methods to create a more accurate predictive model. Lastly, transforming the training hours variable, for example by applying a logarithmic scale, may help in addressing issues related to non-linearity and variance.

Appendix

KiwiLearn Code

http://localhost:8888/lab/tree/COMP517_Assignment_2_Code.ipynb