Assignment 2

Semester 2 2023

PAPER NAME: Data Analysis

PAPER CODE: COMP517

Student ID	Student Names
23194064	**Ha Viet Ly**
23194073	**Thuan Nguyen**
23194080	**Soyun Choi**

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 One: Exploring Data and Testing Hypotheses: Uncovering Insights from Dataset 4
Task 1: Data Preparation and Exploration 4
a, Data Preparation 4
b, Data Exploration 6
c, Data Distribution 8
d, Multivariate analysis of data 10
Task 2: Assumptions, and Hypothesis Formulation 10
a, Objective 10
b, Assumptions 10
c, Hypotheses 11
Task 3: Statistical Technique: Hypothesis Testing 11
a, Statistical method 11
b, Hypothesis Testing 11
Task 4: Discussion and Conclusion 13
a, Discussion 13
b, Conclusion 13
Part Two: Regression Analysis 14
Task 1: Identify Potential Predictor Variables 14
Task 2: Assumptions for Regression Analysis 15
a, Assumptions 15
b, Relevance to Analysis 15
c, Multicollinearity Testing 15
d, Linearity Testing 16
Task 3: Regression Analysis 17
Task 4: Assumptions of Linear Regression 18
a, Assumptions 18
b, Assumption of Normality of Residual 18
c, Assumption of Homoscedasticity and Independence of residuals 19
Task 5: Discussion and Conclusion 20
a, Discussion 20
b, Conclusion 20
c, Limitations and Further Research: 20

Table of Figures

Figure 1: Duplicates of Dataset 4
Figure 2: Box Plots for Numerical Data 4
Figure 3: Dataset's outliers 4
Figure 4: Scatter Plots for Columns with Outliers 5
Figure 5: Dataset's Missing values 5
Figure 6: Information of the Dataset 6
Figure 7: First few rows of the Dataset 6
Figure 8: Summary Statistics of the Dataset 7
Figure 9: Distribution for Numerical Columns 8
Figure 10: Distribution for Categorical Columns 9
Figure 11: Distribution of Experience Category and Performance Rating by Department 10
Figure 12: One-way ANOVA Results 11
Figure 13: Tukey's Post Hoc Test 12
Figure 14: Correlation Heatmap of All Numerical Data 14
Figure 15: Correlation Heatmap of Independent Variables 15
Figure 16: Scatter Plots of TrainingHours and Salary columns 16
Figure 17: OLS Regression Results 17
Figure 18: Q-Q Plot of Residuals 18
Figure 19: Anderson-Darling Normality Test 19
Figure 20: Homoscedasticity Plot 19

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

Task 1: Data Preparation and Exploration

a, Data Preparation

Firstly, the dataset is checked for any duplicates, outliers and missing values that are not handled.

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

---Number of duplicates: 0

---Percentage of duplicates: 0.0 %
```

Figure 1: Duplicates of Dataset

As the given figure above, no presence of duplicate rows is detected. Therefore, no action needs to be taken to address the duplication of rows in the dataset.

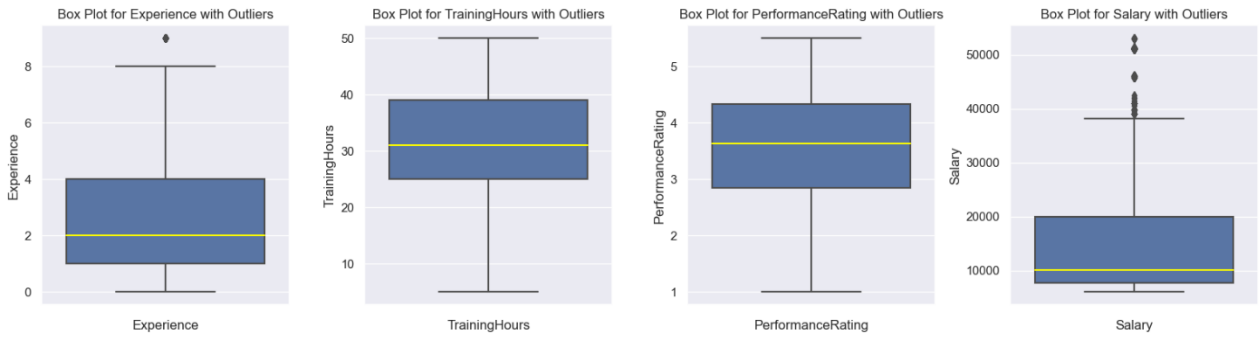

Figure 2: Box Plots for Numerical Data

```
---Sum of Outliers:
 Experience            61
TrainingHours           0
PerformanceRating       0
Salary                152
dtype: int64

---Columns with Outliers:
['Experience', 'Salary']
```

Figure 3: Dataset's outliers

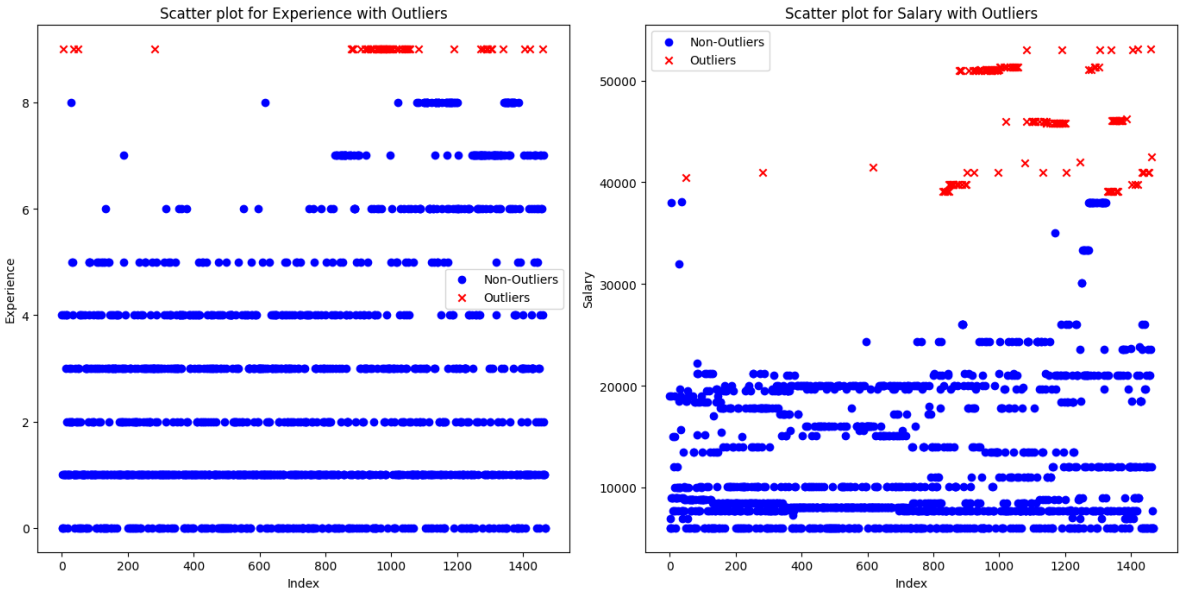

Figure 4: Scatter Plots for Columns with Outliers

Besides, Interquartile Range (IQR) is used to detect extreme values without making any assumptions about the underlying data distribution. According to the generated results, outliers can be seen in the 'Experience' and 'Salary' columns. These outliers should be kept as they provide valuable insights and accuracy of the dataset:

- Experience: all the 61 extreme values in this column are '9' indicating nine years of experience. However, 9 is the appropriate number of years that one employee can work in the company, and nine years of experience should be acknowledged.
- Salary: 152 outliers of this column lie between 40000NZD and above 50000NZD, indicating a much higher salary than the average salary. But there are many factors contributing to one's salary so keeping all these values could represent the real-life circumstances.

Moreover, no missing values are detected among all the 7 columns. Every row in the dataset contains a valid and complete value.

In conclusion, the dataset is cleaned and ready to be analysed with appropriate outliers, no duplicates, and no missing values.

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

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

Figure 5: Dataset's Missing values

b, Data Exploration

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

---Dataset's Shape:
(1468, 7)
```

Figure 6: Information of the Dataset

Dataset consists of 7 columns:

- 5 numerical columns: EmployeeID, Experience, TrainingHours, PerformanceRating, Salary
- 2 categorical columns: Department, Gender

Dataset has a shape of 1468 rows by 7 columns. In this context, the dataset stored 1468 employee records including 7 attributes for each of them.

```
---Dataset's First few rows:
   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 7: First few rows of the Dataset

With the dataset's first few rows, some information is gained as:

- EmployeeID is a 4-digit number that started from 1001, and the dataset is presented in the ascending order of EmployeeID.
- The names of departments are shown either in full, such as Marketing and Sales, or in abbreviated, such as IT or HR.
- Gender of the first 5 employees is either male or female.

- Experience, TrainingHours, and Salary of an employee is represented as an integer with the unit of the columns are year and hour, respectively.
- PerformanceRating of an employee is a number with 2 decimal places, ranging from 1.00 to 5.50 for the first 5 employees.

---Summary Statistics:

	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

Figure 8: Summary Statistics of the Dataset

Summary Statistics:

- There is data for 1468 employees.
- Experience ranges from 0 to 9 years, which are divided into four levels: Entry-level, Junior, Mid-level, Senior. Those employees have approximately 3 years of experience on average.
- Training hours range from 5 to 50 hours.
- Performance ratings span from 1 to 5.5. The average rating is around 3.5.
- Salaries range from 6000 to 53100. The average salary of all departments is approximately 16107.

c, Data Distribution

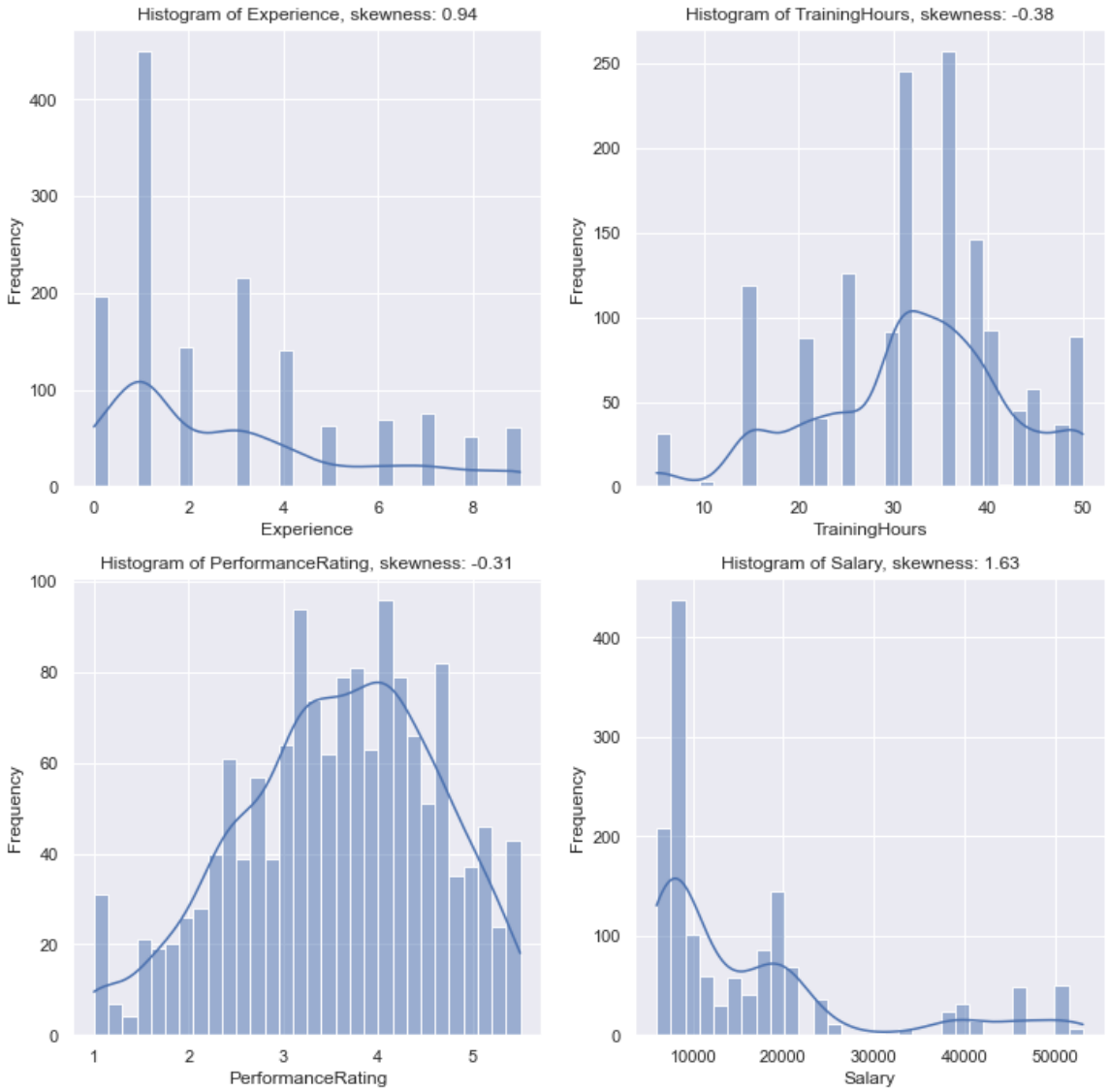

Figure 9: Distribution for Numerical Columns

From the figure above,

- The distribution of performance ratings seems to be slightly left-skewed (skewness: -0.31) with many employees receiving ratings between 3 and 5.
- The distribution of training hours appears to be slightly left-skewed (skewness: -0.38) and the shape of the distribution might be random.
- The distribution of experience is right-skewed (skewness:0.94) and relatively uniform, indicating that there are employees with caried years of experience in the dataset.
- The distribution of Salary is significantly right-skewed (skewness: 1.63), which is indicating that most of the employees have low salary.

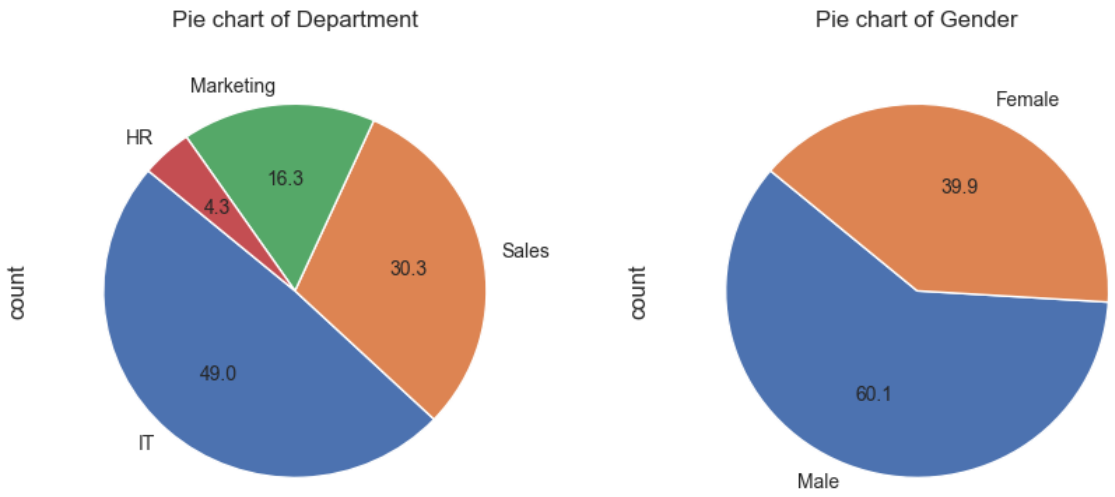

Figure 10: Distribution for Categorical Columns

From the figure above,
Distribution of different departments can be interpreted as:

- The majority of employees work for the IT department, accounting for nearly half of the whole company at 49%.
- Following this, the Sales and Marketing departments accounted for 30.3% and 16.3% respectively.
- In contrast, the HR department has the fewest employees, representing a mere 4.3% of the workforce.

Distribution of different genders can be interpreted as:

- More than half of the company's employees are Male, constituting 60.1% of the total employees.
- At 39.9%, female employees make up the smaller segment of the workforce.

d, Multivariate analysis of data

Figure 11: Distribution of Experience Category and Performance Rating by Department

The boxplot displays the relationships between employees' years of experience on their performance ratings within different departments, segmented by years of experience. The boxplot provides a visual representation of median ratings, interquartile range, and potential outliers of performance rating in each department for distinct experience levels.

In conclusion, from the initial exploration, such as histogram of performance ratings, experience, and boxplot above, it is observed that there are variations in performance ratings across departments and experience levels and necessarily to be investigated.

Task 2: Assumptions, and Hypothesis Formulation

a, Objective

This report aims to investigate potential variations in employee performance ratings across different departments. If such variations exist, identify which departments have ratings that are noticeably higher or lower.

b, Assumptions

- The dataset represents an accurate representation of the organisation’s workforce.
- The performance ratings are reliable and consistently measured across departments.

c, Hypotheses

- Null hypothesis (H0):
  μIT = μMarketing = μSales = μHR (all departments have the same mean of performance ratings.)
  There is no statistically significant difference between the mean performance rating across these departments.
- Alternative hypothesis (Ha):
  At least one department has a means of performance ratings that is different from the others.
  There is a statistically significant difference among the mean performance rating across these departments.

Task 3: Statistical Technique: Hypothesis Testing

a, Statistical method

For this analysis, the appropriate statistical technique is ANOVA (Analysis of Variance).
It is suitable for this because:

- Independent variable (Department) is categorical with multiple levels.
- Dependent variable (Performance Rating) is continuous.
- To test if the means of Performance Rating are different across various departments.

b, Hypothesis Testing

```
Unique Department Counts:
['IT' 'Marketing' 'Sales' 'HR']

----------------------------------------------------------------------------------------------------

One-way ANOVA Results:
P-value: 0.0000

The null hypothesis is rejected in favour of the alternative hypothesis
There is statistically significant different among the mean performance rating across these department

----------------------------------------------------------------------------------------------------

One-way ANOVA Results:
F-statistic: 61.45
Critical F-value: 2.61

The test statistic is in the tail of the F-distribution, and the null hypothesis is rejected
There is statistically significant different among the mean performance rating across these department
```

Figure 12: One-way ANOVA Results

From the result we get, the p-value is very small (0.0000), and the F-statistic = 61.45 is much higher than the Critical F-value = 2.61. Therefore, the Null Hypothesis is rejected and there is statistically significant difference among the mean performance rating across these departments, which means that at least one department has a mean performance rating that is different from the others.

For determining which specific pairs of departments exhibit significant differences in performance ratings, Tukey's post-hoc test is the chosen method.

The results from Tukey's post-hoc test provide pair-wise comparisons between the departments. Its benefit is to identify which specific pairs of departments have significantly different mean performance ratings.

```
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 13: Tukey's Post Hoc Test

As can be seen from the result above, there are several department pairs that have significant differences in performance ratings as indicated by the "reject" columns (True indicates a significant difference).

To further understand these differences, the positive mean values of "meandiff" columns indicates first department (group1) has a higher average performance rating meanwhile the negative mean value of the second (group2) has a higher average rating.

Additionally, to determine where those significant differences lie when there are more than two groups, the mean differences (meandiff) and the p-values adjusted for multiple comparisons (p_adj) need to be considered:

- HR and IT: The mean difference between the HR and IT's performance ratings is 0.3715, which is higher than the p-value (p-adj) that is 0.0217, indicating that there is statistically significant difference between these two groups (reject = True).
- HR and Marketing: The mean difference between the HR and Marketing's performance ratings is 1.027 which is higher than the p-value (p-adj) that is 0.0, indicating that there is statistically significant difference between these two groups (reject = True).
- HR and Sales: The mean difference between the HR and Sales' performance ratings is 1.0256, which is higher than the p-value (p-adj) that is 0.0, indicating that there is statistically significant difference between these two groups (reject = True).
- IT and Marketing: The mean difference between the IT and Marketing's performance ratings is 0.6555, which is higher than the p-value (p-adj) that is 0.0, indicating that there is statistically significant difference between these two groups (reject = True).
- IT and Sales: The mean difference between the IT and Sales' performance ratings is 0.6541, which is higher than the p-value (p-adj) that is 0.0, indicating that there is statistically significant difference between these two groups (reject = True).
- Marketing and Sales: The mean difference between the Marketing and Sales' performance ratings is -0.0014, which is smaller than the p-value (p-adj) is 1.0, indicating that there is no statistically significant difference between these two groups (reject = False).

Task 4: Discussion and Conclusion

a, Discussion

- Variations in Performance Ratings: The ANOVA test revealed significant variations in performance ratings across different departments. The post-hoc Tukey test further identified specific department pairs that exhibited these significant differences.

- Implications: The differences in performance ratings across departments could arise from several factors:
    - Departmental work culture and dynamics.
    - Variability in leadership and management styles.
    - Differences in training programs and resources available to employees in different departments.
    - Variations in job roles, responsibilities, and performance assessment criteria.

- Actionable Insights:
    - For departments with lower average performance ratings, leadership may need to investigate the root causes and implement interventions such as training programs, mentorship initiatives, or changes in performance assessment criteria.
    - Departments with notably higher ratings can be studied as potential models. Best practices from these departments could be shared across the organization to drive overall improvement.

b, Conclusion

The analysis revealed significant differences in employee performance ratings across departments. Some departments exhibited notably higher or lower ratings than others. Organizations should delve deeper into department-specific factors that may influence these ratings and implement strategies to ensure consistent and fair performance evaluation across the board. The data-driven insights obtained can guide interventions aimed at fostering a culture of continuous improvement and consistent performance assessment.

Part Two:
Regression Analysis

Task 1: Identify Potential Predictor Variables.

Before building the regression analysis, it is necessary to identify potential predictor (independent) variables that might be correlated with the employee's performance rating (dependent variable).

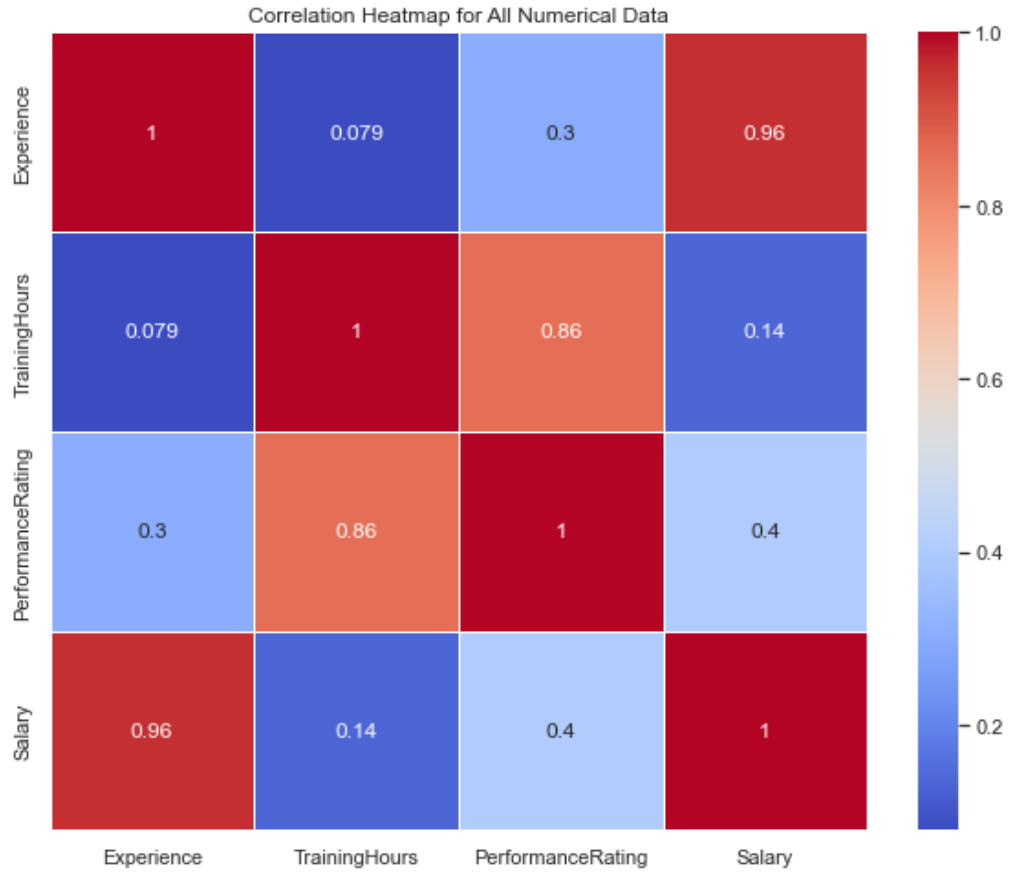

Figure 14: Correlation Heatmap of All Numerical Data

Based on the correlation heatmap and the correlation values,
Potential Predictor Variables:

- Training Hours: there is a strong positive correlation (r=0.86) between the number of training hours and performance rating. The reason for choosing training hours is that employees with more training hours might be better with the skills and knowledge required for their roles, leading to better performance.
- Salary: There is a moderate positive correlation (r=0.40) between salary and performance rating. The reason salary can be potential independent variable is that roles with greater responsibility or complexity might come with higher remuneration and different performance expectation.

- Experience: it is easily observed that experienced employees might have better domain knowledge, expertise, and flexibility of handling different situations, leading to higher performance ratings.

Task 2: Assumptions for Regression Analysis

a, Assumptions

The assumptions for multiple linear regression are:

- Linearity: There should be a linear connection between the independent and dependent variables, meaning that changes of independent variables are proportional to changes in dependent variable.
- Independent: the residuals (errors) should be independent.
- The variance of the residuals is constant (Homoscedasticity) across all levels of the independent variables.
- Normality of Residuals: The observed data comes from a normal distribution.
- No Multicollinearity: There should be no linear connection between the independent and dependent variables. (above 0.7) as it would be challenging to separate their effects on each other.

b, Relevance to Analysis

These assumptions above ensure that the regression results are unbiased, valid, efficient, and reliable. Violations can lead to incorrect conclusions or inefficient estimations.

c, Multicollinearity Testing

Before proceeding to regression analysis, there are two assumptions that need to be tested: The Linearity assumption and multicollinearity among predictor variables are crucial to be checked.

Figure 15: Correlation Heatmap of Independent Variables

According to the above correlation heatmap (Figure 15) between independent variables:

- TrainingHours vs Salary and TrainingHours vs Experience: These two pairs have low correlation values (less than 0.7), suggesting that multicollinearity is not a concern for them.
- Multilinearity is seen as some pairs of independent variables have correlation coefficients (r) greater than or equal to 0.7, which in this analysis is Salary and Experience: 0.96

Multicollinearity can inflate the variance of the regression coefficients. Therefore, one variable between Salary and Experience needs to be removed to eliminate multilinearity due to problems such as unstable coefficient estimates and reduced interpretability of the model. Also, for highly correlated pairs, one variable could be used to determine the other one.

In conclusion, Salary is more highly correlated with target variable (PerformanceRating) than Experience (0.4 > 0.3), strongly believe that keeping Salary and removing Experience might be better for further analysis.

After finishing the multicollinearity check, there are now only Salary and TrainingHours as independent variables.

d, Linearity Testing

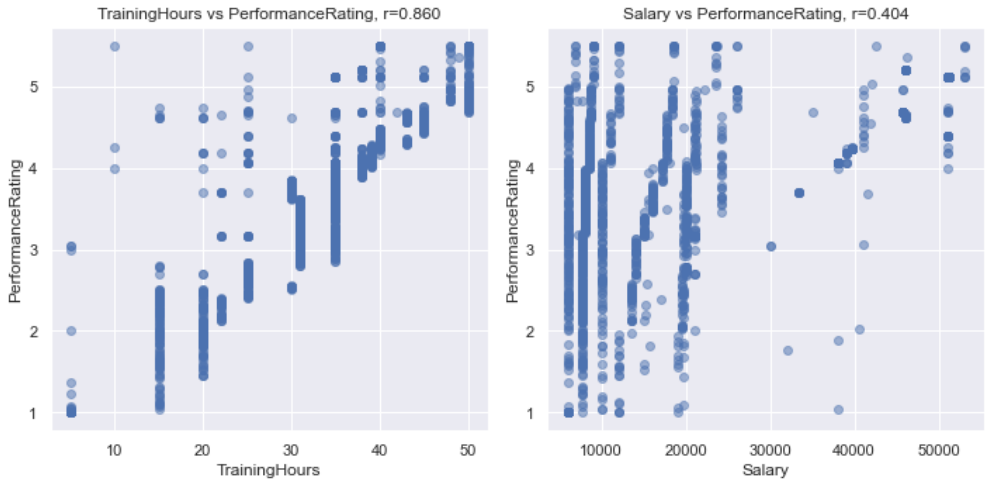

Figure 16: Scatter Plots of TrainingHours and Salary columns

According to the above scatterplots:

- There seems to be a strong clear positive linear relationship between the number of training hours and performance rating (r = 0.860).
- The relationship of Salary and Performance Rating is moderate positive (r = 0.404) but seems to be weaker compared to training hours and performance rating.

In conclusion, both two variables checked above both have positive linear correlation with dependent variable (Performance Rating). Therefore, the selected predictor variables seem to reasonably satisfy the linearity condition.

Task 3: Regression Analysis

After checking for Linearity assumption and Multicollinearity, the identified predictor variables (Salary and TrainingHours) can be used to present the Regression Analysis.

OLS Regression Results

Dep. Variable:	PerformanceRating	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.823
Method:	Least Squares	F-statistic:	3412.
Date:	Wed, 18 Oct 2023	Prob (F-statistic):	0.00
Time:	12:30:26	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 17: OLS Regression Results

Multiple Linear Regression Model is used. This model can be interpreted as:

- R-squared: 0.823, indicating a relatively good amount 82.3 % of the variability in the dependent variable can be explained by the independent variables in this model:
- F-statistic: 3412. is high, indicating at least one independent variable significantly contributes to explaining the dependent variable.
- Prop (F-statistic): 0.00 is low, indicating strong evidence against the null hypothesis that all coefficients are zero.
- The intercept (const) is 0.4306, indicating when TrainingHours and Salary is 0, predicted PerformanceRating is 0.4306
- P-values for Coefficients: The p-values associated with the coefficients are extremely low (zero):
    - The intercept, TrainingHours, and Salary coefficients are statistically significant.
    - In this context, TrainingHours, and Salary significantly contributes to predicting Performance Rating.

According to Figure 17, there is one noticeable note suggesting that the condition number (6.97e+04) is large, and some issues may be associated with it. This note can be explained by the presence of outliers in the 'Salary' column. These outliers are rationalised in Part One Task 1, and will be seen more clearly in Q-Q Plot of Residuals in Part Two Task 4.

In conclusion, given all other independent variables are kept constant, for every 1 unit increase in TrainingHours, the PerformanceRating increases by around 0.08 units, whereas for every 1 unit increase in Salary, the PerformanceRating increases by 2.52e-05 units.

Task 4: Assumptions of Linear Regression

a, Assumptions

After performing the regression, this part is to check the remaining assumptions to validate the reliability of the regression analysis and the accuracy of its results. As mentioned as Task 2, the assumptions now left:

- Independence of residuals
- Homoscedasticity (constant variance of residuals)
- Normality of residuals

b, Assumption of Normality of Residual

Check for assumptions using diagnostic plots and tests:

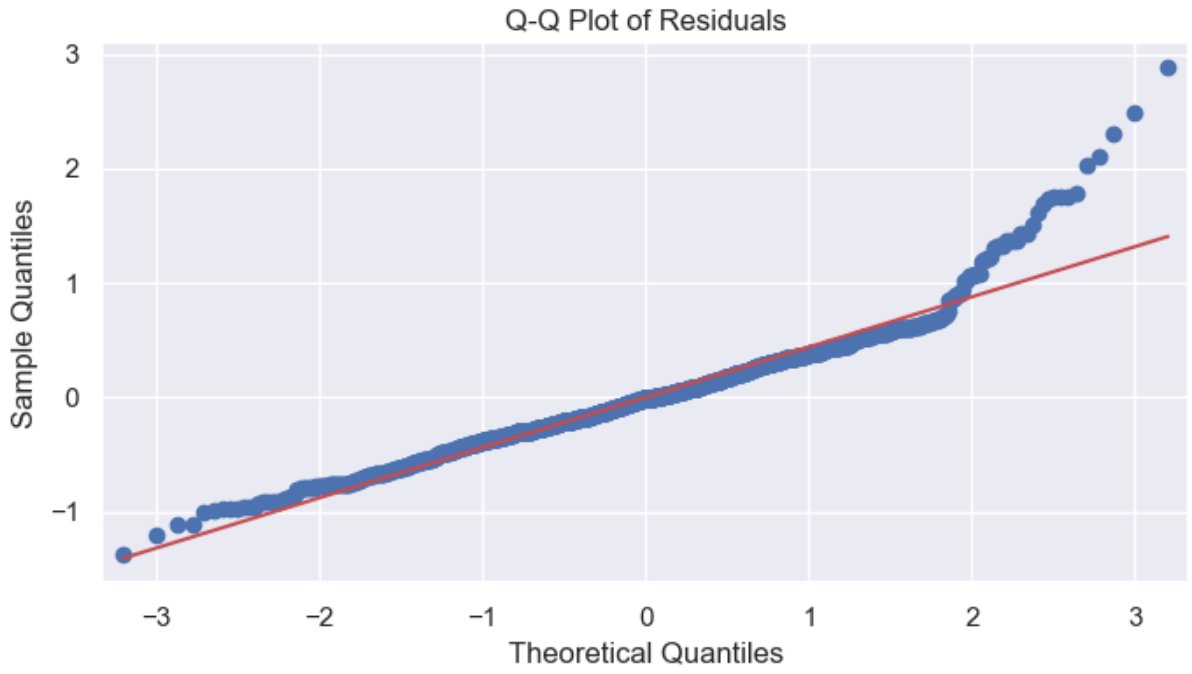

Figure 18: Q-Q Plot of Residuals

The plot above checks the normality of residuals. The points should approximately lie on the reference line. As can be observed above, most of the points closely follow a straight line except for some points as the end, it suggests that the residuals are approximately normally distributed but still having some potential outliers or heavy-tailed distributions. Due to those outliers, it is the reason that might make the condition number in Part Two Task 3 large.

```
Anderson-Darling Statistic: 7.917607488034719
Critical Values: [0.574 0.654 0.785 0.916 1.089]
Significance Levels: [15.  10.   5.   2.5  1. ]
```

Figure 19: Anderson-Darling Normality Test

Anderson-Darling Statistic: 7.92 is greater than Critical Value of (0.785) at Significance Levels of 5; reject the null hypothesis. The observed data is likely to come from a normal distribution.

As a result, due to all the reasons mentioned above, assumption of normality of residual is met.

c, Assumption of Homoscedasticity and Independence of residuals

Next, the plot shown below, which is Residuals vs Fitted values (Predicted Values), helps to check Independence of residuals and Homoscedasticity. The residuals should be scattered randomly around the horizontal axis and there should be no discernible pattern or trend in the plot.

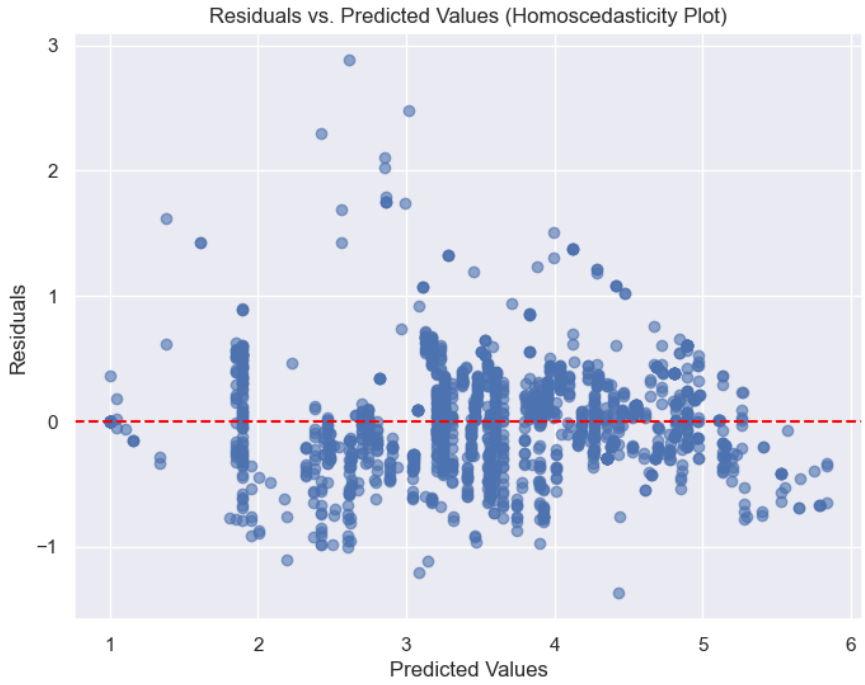

Figure 20: Homoscedasticity Plot

According to the above plot, the points seem to be a random scatter around the red line with little deviation. These points also not classic cone or fan shape: not Heteroscedasticity. It is suggested that the assumptions are reasonably met.

As a result, all the assumptions of the linear regression model seem to be reasonably satisfied.

Task 5: Discussion and Conclusion

a, Discussion

From the Regression model that is made in Task 3 Part 2 (figure 17):

- TrainingHours and Performance Rating: There is a positive relationship between the number of training hours and performance rating. This finding aligns with the intuitive understanding that employees who undergo more training are likely better equipped to perform their tasks, leading to higher ratings.
- Salary and Performance Rating: There is also a positive relationship between salary and performance rating, although the effect size is smaller. This could be because higher salaries are associated with roles of higher responsibility or complexity. The performance expectations and assessment criteria might be different for these roles.
- Model Fit: The model explains approximately 82.3% of the variance in performance ratings, which indicates a good fit. However, as mentioned in Task 1 Part 1, the presence of outliers in Salary (as indicated by the high condition number), which is kept because those outliers might reflect by a lot of factors in reality. Therefore, it is acceptable, and the results can be interpreted with caution.

b, Conclusion

The regression analysis revealed that both training hours and salary are significant predictors of employee performance ratings. While training has a more substantial effect, salary also plays a role in determining performance ratings. Organizations can leverage these insights to make informed decisions about training programs and compensation strategies to optimize employee performance.

c, Limitations and Further Research:

- Multicollinearity: The high correlation between salary and experience was a limitation in this analysis. Future analyses could use techniques like ridge or lasso regression to handle multicollinearity.
- Other Factors: The dataset might not capture all factors influencing performance ratings. Factors like job satisfaction, team dynamics, or leadership style could also play a role.
- Causality: This analysis is correlational, and causality cannot be inferred. Experimental or longitudinal designs would be needed to determine causal relationships.