****

**ANL 252**

**PYTHON FOR DATA ANALYTICS**

**Group-Based Assignment**

**January 2023 Presentation**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI Number** |
| Chia Zhi Sheng Keath | H2270045 |
| Teo Meng Hee Tiffany | N2082088 |
| Xiong XiaoFeng | Y2110266 |
| Chong Xin Hui Sandy | W2110447 |

**Tutorial Group: T05 / Group 07**

**Instructor’s Name: MUNISH KUMAR**

**Submission Date: 19 Feb 2023**

**Declaration Page**

We, members of group 7, do hereby declare that we each contributed to this assignment and that we collectively agree to a shared grade.

|  |  |  |
| --- | --- | --- |
| Name | Contribution | Signature |
| Chia Zhi Sheng Keath  H2270045 | I did question 1 (a) and (b) – 4 charts, codes and corresponding table, editing descriptions | *keath* |
| Teo Meng Hee Tiffany  N2082088 | I did question 1 (a) and (b) – 4 charts, codes and corresponding table, writing descriptions | *Tiffany* |
| Chong Xin Hui Sandy  W2110447 | I did question 1 (a) | Sandy |
| Xiong XiaoFeng  Y2110266 | I did question 1 (c) | *Xiong* |

**Question 1**

**(a)**

**Chart 1**

To start off, we wanted to analyse how the monthly pay is distributed across the 1250 employees in various departments, job role and job grade. We can analyse it using histogram, which is a chart that visualises continuous variable, that is “SalaryPay”. Using excel, the salary distribution histogram chart is as shown Figure 1.1.1.

**Figure 1.1.1**

***Histogram on Company’s monthly pay distribution***

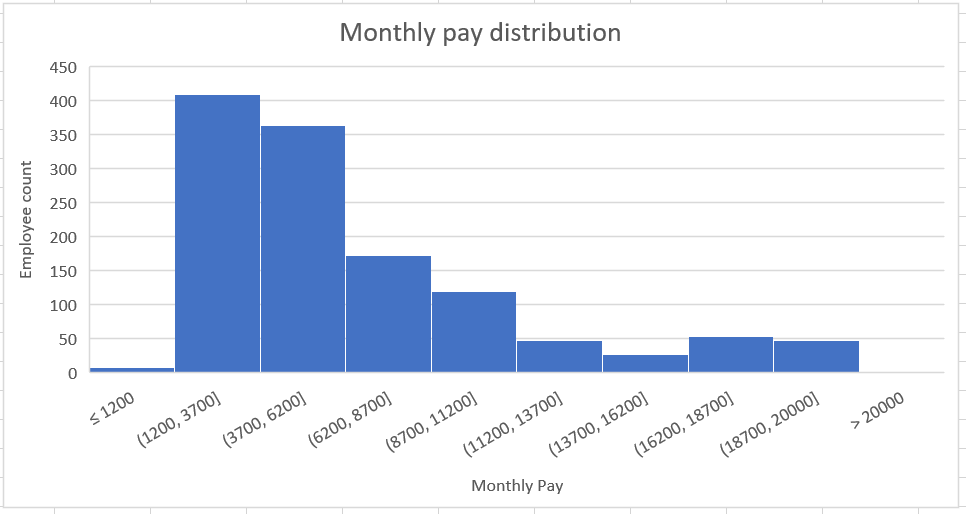
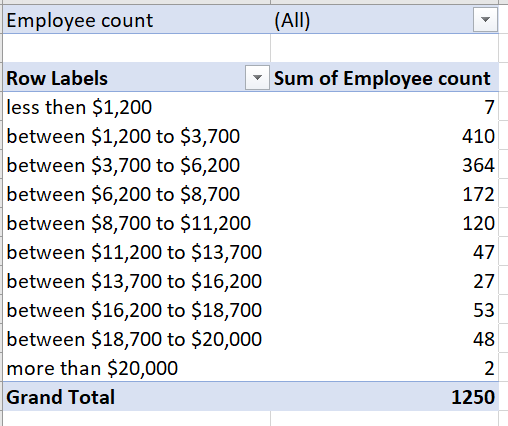


Chart 1 shows a right-skewed histogram distribution, where the peak of the curve is on the graph’s left side. This suggests that most of the employees’ monthly salary data falls on the lower side of the salary distribution range. Specifically, between $1,200 and $6,200. This insight gained could be interesting for the company to evaluate its compensation competitiveness and attractiveness towards existing employees, fresh graduates, or future employees. For example, employees may leave the company due to low monthly pay, or people may not join the company because of unattractive pay.

**Figure 1.1.2**

***Chart 1’s corresponding table***



The corresponding table’s content is simply the column of employees’ monthly pay. Additionally, since the automatic excel function displayed too narrow bins, which will cause noise that distracts from proper insights, we adjusted the bins width, overflow, and underflow bin. Specifically, the salary bin width was adjusted to $2,500, 10 bins, overflow bin of $20,000 and underflow bin of $1,200.

**Chart 2**

Next, we wanted to identify some characteristics of employees that left the company. For this, we will assess employees that left the company and their job characteristics of job grade. Furthermore, we will analyse them in their respective units. A stacked bar chart will allow us to create bars that can represent different proportions adding to the total, such that comparisons can be made between different proportions and the whole of the various bar categories. In this case, compare leavers in the 3 business units and their respective 5 job grades. Each business unit is represented by a bar with height representing the number of employees that have left the company. Each bar is then stacked with different colours representing employees’ job grades. Chart output from excel is as shown below.

**Figure 1.2.1**

***Stacked bar chart on employees who left and their job grade.***

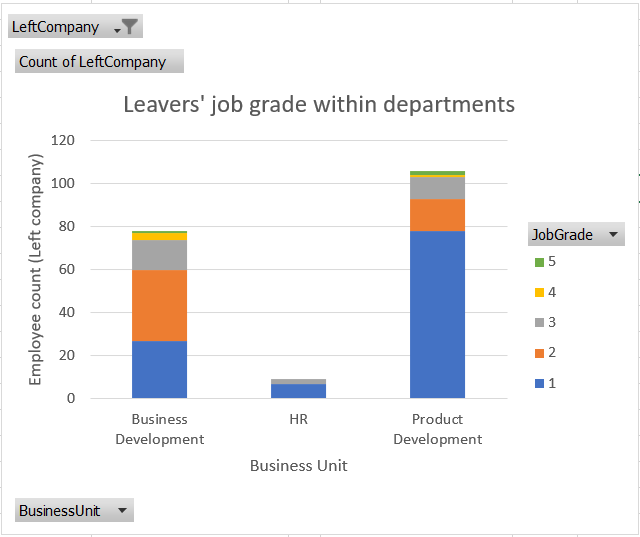
****

Figure 1.2.1 shows that the tallest bar is the product development department, suggesting that it has the most number of leavers, followed by business development, then HR. But the interesting observation is that the blue area, representing job grade level 1, has a significantly large area for product development and the HR department. For business development, although blue shade is not the greatest area, it is still the second largest area. The orange area, representing job grade level 2 is greater. Nevertheless, grade 2 still implies a lower-level job grade compared to 3 and above. The insights gained from this observation is that most of the leavers are in the lower job grade level of 1 and 2, irregardless of their business units. Following chart 1’s analysis, we could deduce that perhaps employees in the lower-level roles are leaving because salary is not competitive enough.

**Figure 1.2.2**

***Chart 2’s corresponding table***

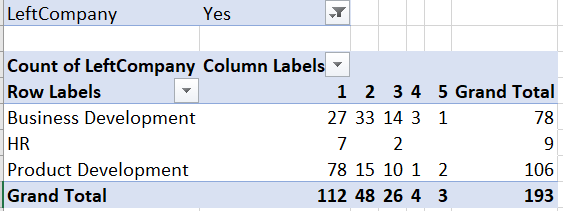


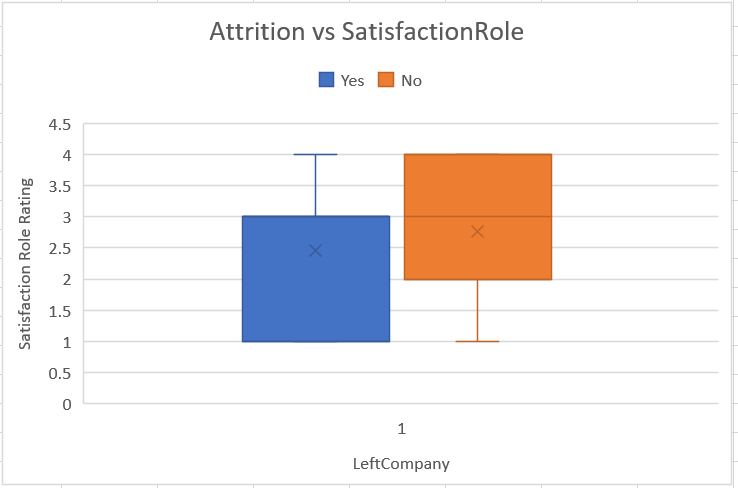
Figure 1.2.2 shows us the filtered data of total 193 employees who have left, by job grade and departments. From here, it shows that there is a certain level of correlativeness for employees who left between their grade and department. “JobGrade” attribute is pulled to a column with 5 different job grade values. The “BusinessUnit” attribute is under rows.

**Chart 3**

For Chart 3, we wanted to analyse “CompanyLeft” and “SatisfactionRole” to see whether employees’ satisfaction in their job roles have an impact on their decision to leave the company. Potential insights can be gained to establish effective retention strategies if there is a relationship. We used a box plot to have a clear summary of various quantitative attributes like upper and lower quartile, median and more. We created 2 box plots, 1 representing employees who left the company and the other representing employees still in the company.

**Figure 1.3.1**

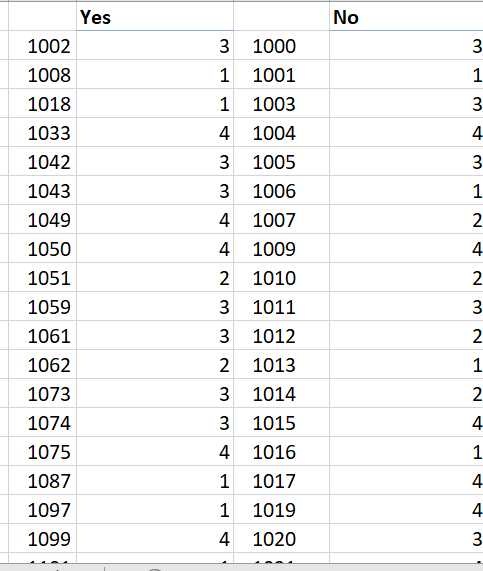
***Box plot on “CompanyLeft” and “SatisfactionRole”***



Firstly, looking at the blue box plot which are employees that left, its 25th percentile is a satisfaction rating of 1 while both its 50th and 75th percentile is rating 3. This suggests that the majority of their role satisfaction is 3 and below. Secondly, for the orange box plot representing employees that are still with the company, the 25th, 50th and 75th percentiles are satisfaction rating 2, 3 and 4, respectively. So, very few of them have very lower role satisfaction of below 2. Hence, generally, employees that left have lower satisfaction in their job role. This could be a useful insight if the company wants to improve its attrition.

**Figure 1.3.2**

***Chart 3’s corresponding table***



The corresponding table for chart 3 includes “EmpID” and employees’ respective input on their satisfaction rating for their job role. Employees are filtered and split into ‘Yes’ and ‘No’ columns for the “LeftCompany” variable. The list of employees that are ‘Yes’ for “LeftCompany” is 193, while ‘No’ is 1057. Thus, the corresponding table extends further down than what is shown.

**Chart 4**

Finally, we wanted to investigate whether investment in training and development for employees is effective on their performance. Since we want to find the correlation between the 2 numeric attributes, “TrainingAttended” and “Performance rating”, we constructed a scatter plot.

**Figure 1.4.1**

***Scatter plot on “TrainingAttended” and “Performance rating”***

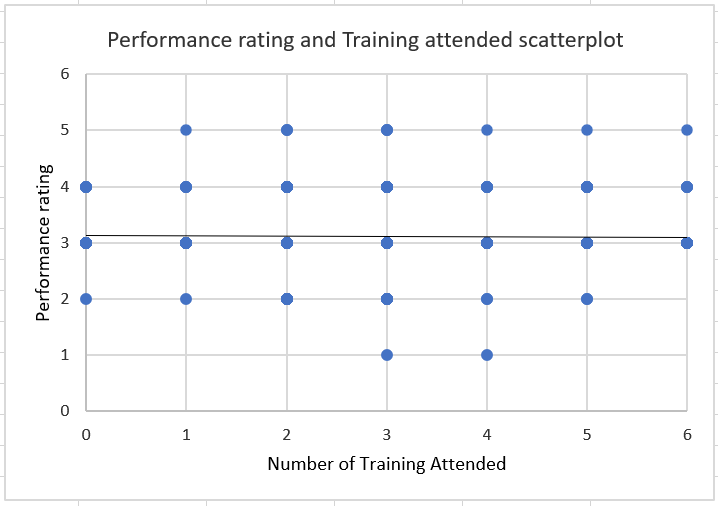
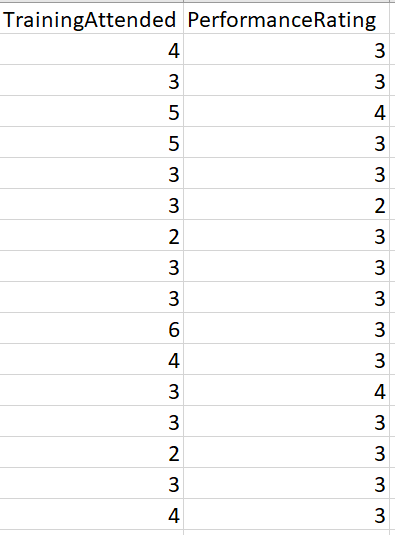


Figure 1.4.1 shows that the number of trainings attended is between 0 and 6. For performance rating, employees can be rated between 1 to 5, of which our group sees rating ‘3’, ‘4’, and ‘5’ to suggest good performance. The scatterplots points align closely to the grid instead of scattered, following the fixed scale-point in both axises’ variables and whole number inputs. The trendline is horizontal, suggesting that there is no relationship between the 2 variables. This is rather interesting as we initially thought that training typically impacts greater performance, but it seems that training does not impact performance in this case. Insights the company can gain from this is that employees’ performance will be as given, regardless of whether they had previously attended training. So, certain employees that have poor performance ratings, that is 1 or 2, are not due to the lack of training. Instead, it could be due to other factors.

**Figure 1.4.2**

***Chart 4’s corresponding table***



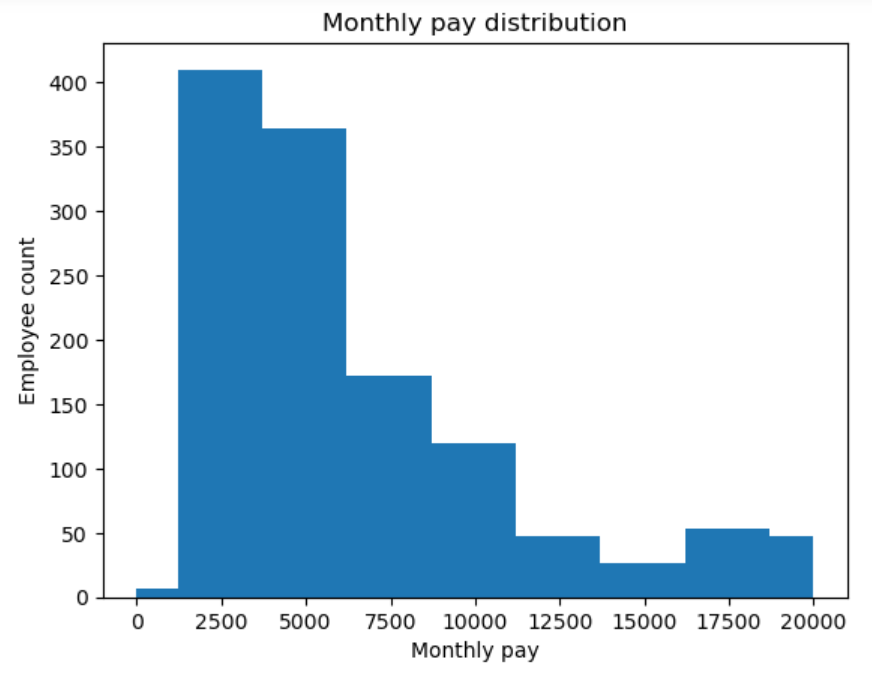
The corresponding table for chart 4 is essentially just “TrainingAttended” and “PerformanceRating” columns.

**(b)**

**Chart 1: Histogram on monthly pay distribution**

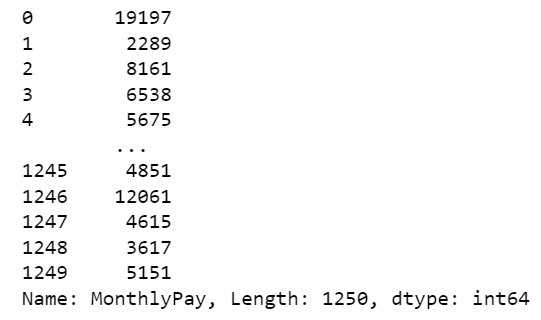
**Figure 2.1.1**

***Python’s chart 1 output***



**Figure 2.1.2**

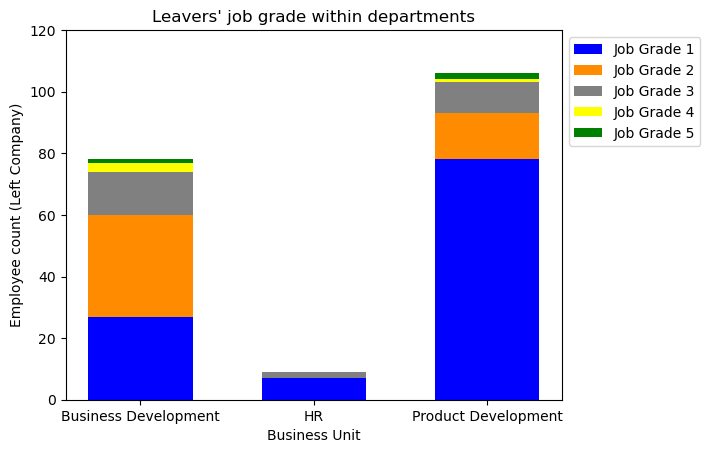
***Python’s chart 1 corresponding table***



**Chart 2: Stacked bar**

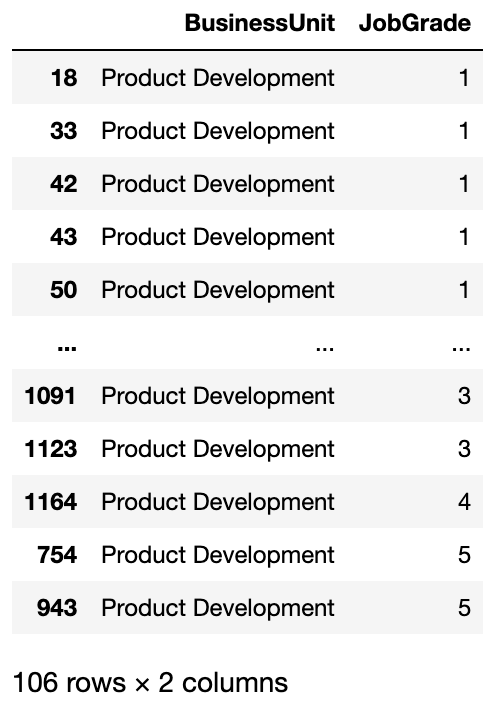
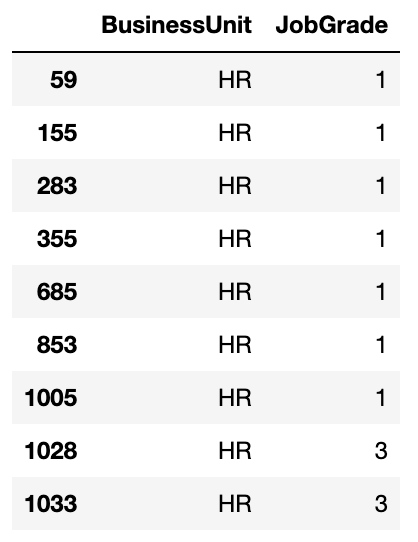
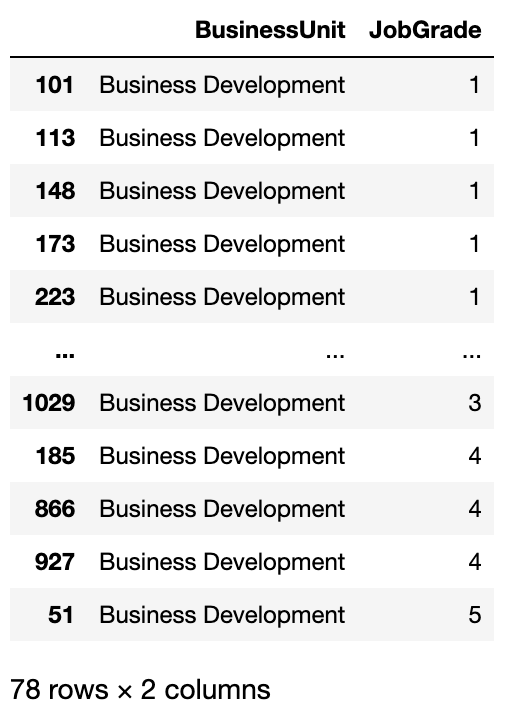
**Figure 2.2.1**

***Python’s chart 2 output***



**Figure 2.2.2**

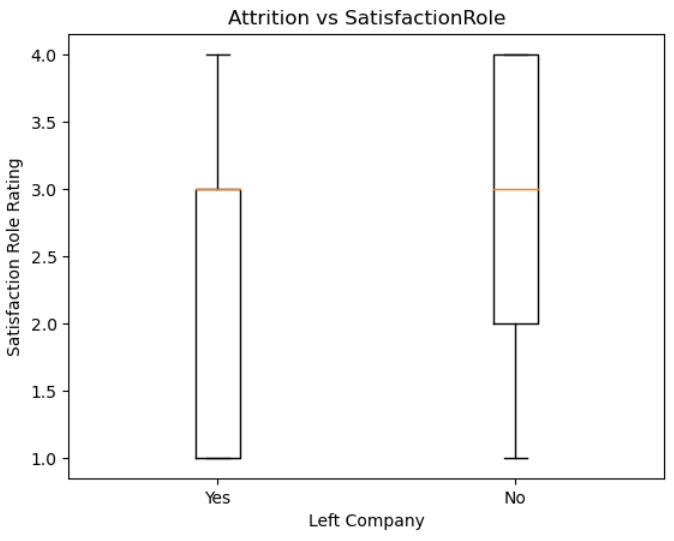
***Python’s chart 2 corresponding table***



**Chart 3**

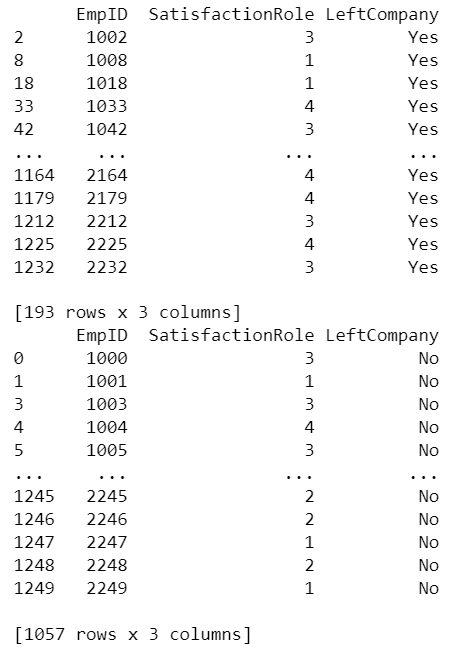
**Figure 2.3.1**

***Python’s chart 3 output***



**Figure 2.3.2**

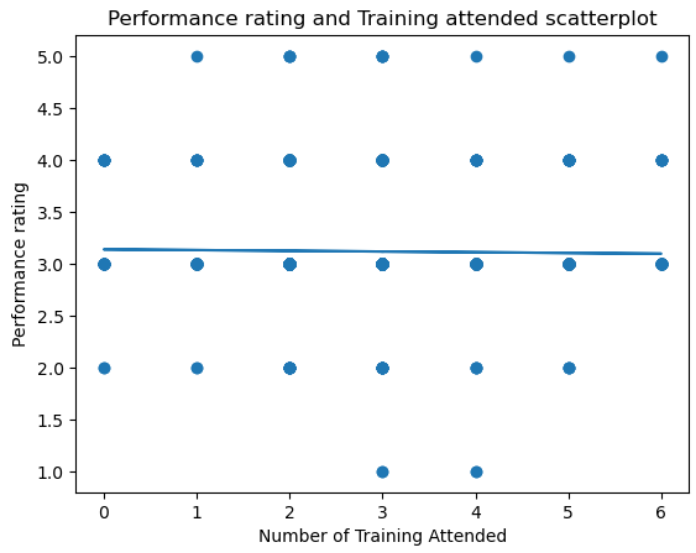
***Python’s chart 3 corresponding table***



**Chart 4**

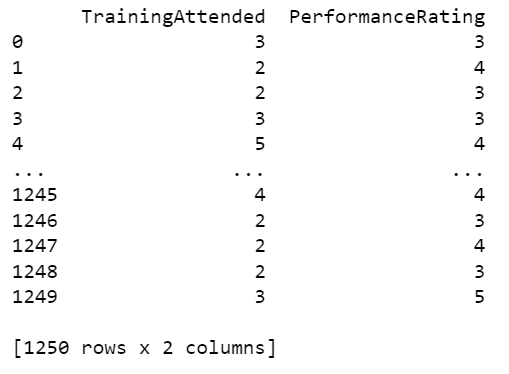
**Figure 2.4.1**

***Python’s chart 4 output***



**Figure 2.4.2**

***Python’s chart 4 corresponding table***



**(c)**

**Output for Option 1 — Current staff strength:**

There are 1057 people in the company.

Business Devlopment: 299 people;Product Development: 716 people;HR: 42 people

Male: 623 people;Female: 434 people

**Output for Option 2 — Current staff annual remuneration:**

The current total annual staff remuneration expenditure is $87,182,508

The current median monthly salary is $5,238

**Output for Option 3 — Eligible long service award recipients:**

EmpID BusinessUnit Education FieldOfStudy Gender Age \

295 1295 Business Development 4 Business Tech Male 59

466 1466 Product Development 3 Business Tech Male 55

1242 2242 Business Development 4 Marketing Male 57

896 1896 Business Development 4 Marketing Male 59

ProximityToHome Travel JobGrade JobRole ... TrainingAttended \

295 10 Rare 4 Account Manager ... 1

466 1 Rare 5 Section Manager ... 2

1242 26 Rare 5 Section Manager ... 3

896 3 Rare 4 Section Manager ... 3

WorkLifeBalance WithCompany CurrentRole LastPromoted CurrentManager \

295 2 37 10 1 8

466 3 36 10 4 13

1242 3 36 6 2 13

896 4 35 6 1 16

SatisfactionCompany SatisfactionRole SatisfactionCoworkers \

295 4 3 3

466 4 1 3

1242 3 4 3

896 3 1 1

LeftCompany

295 No

466 No

1242 No

896 No

[4 rows x 26 columns]

Employees with 35 or more years of service:

EmpID: 1295 Job role: Account Manager Age: 59 With company: 37 years

EmpID: 1466 Job role: Section Manager Age: 55 With company: 36 years

EmpID: 2242 Job role: Section Manager Age: 57 With company: 36 years

EmpID: 1896 Job role: Section Manager Age: 59 With company: 35 years

**Output for Option 4 — Exit:**

Would you like to continue? (Indicate yes or no)

When input ‘yes’, the program will continue

When input ‘no’, the program will be paused.

**Appendixes**

**Appendix 1: Codes for creating chart 1, histogram, in Python**

# import pandas package into the program

import pandas as pd

# use pandas function to import and read the dataset saved as .csv from external source

EE\_data = pd.read\_csv("GBA.csv")

# select monthly pay column by its variable name in the dataset

monthly\_pay = EE\_data["MonthlyPay"]

print(monthly\_pay)

# import matplotlib visualisation package into the program

import matplotlib.pyplot as plt

# plot histogram with arguments of monthly pay as x-axis values, histtype, bin ranges

plt.hist(monthly\_pay, histtype = 'bar', bins=[0,1200,3700,6200,8700,11200,13700,16200,18700,20000])

# apply plot options to add chart title, and axes labels

plt.title("Monthly pay distribution")

plt.xlabel("Monthly pay")

plt.ylabel("Employee count")

**Appendix 2: Codes for creating chart 2, stacked bar chart, in Python**

##import pandas package, numpy and matplotlib

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

employee= pd.read\_csv("GBA.csv")

##double checking if the columns are correct

employee

##filter job grade

left= employee.query('LeftCompany=="Yes"')

jobgrade1=left.query('JobGrade==1')

jobgrade2=left.query('JobGrade==2')

jobgrade3=left.query('JobGrade==3')

jobgrade4=left.query('JobGrade==4')

jobgrade5=left.query('JobGrade==5')

##TABLE OUTPUT

leftbd=left.query('BusinessUnit=="Business Development"')

print(leftbd.sort\_values(by=["BusinessUnit", "JobGrade"])[["BusinessUnit", "JobGrade"]])

lefthr=left.query('BusinessUnit=="HR"')

print(lefthr.sort\_values(by=["BusinessUnit", "JobGrade"])[["BusinessUnit", "JobGrade"]])

leftpd=left.query('BusinessUnit=="Product Development"')

print(leftpd.sort\_values(by=["BusinessUnit", "JobGrade"])[["BusinessUnit", "JobGrade"]])

##BUSINESS DEVELOPMENT

bd1=int(len(jobgrade1.query('BusinessUnit=="Business Development"')))

bd2=int(len(jobgrade2.query('BusinessUnit=="Business Development"')))

bd3=int(len(jobgrade3.query('BusinessUnit=="Business Development"')))

bd4=int(len(jobgrade4.query('BusinessUnit=="Business Development"')))

bd5=int(len(jobgrade5.query('BusinessUnit=="Business Development"')))

##HR

hr1=int(len(jobgrade1.query('BusinessUnit=="HR"')))

hr2=int(len(jobgrade2.query('BusinessUnit=="HR"')))

hr3=int(len(jobgrade3.query('BusinessUnit=="HR"')))

hr4=int(len(jobgrade4.query('BusinessUnit=="HR"')))

hr5=int(len(jobgrade5.query('BusinessUnit=="HR"')))

##PRODUCT DEVELOPMENT

pd1=int(len(jobgrade1.query('BusinessUnit=="Product Development"')))

pd2=int(len(jobgrade2.query('BusinessUnit=="Product Development"')))

pd3=int(len(jobgrade3.query('BusinessUnit=="Product Development"')))

pd4=int(len(jobgrade4.query('BusinessUnit=="Product Development"')))

pd5=int(len(jobgrade5.query('BusinessUnit=="Product Development"')))

##plotting stacked bar

x=["Business Development","HR","Product Development"]

one=[bd1,hr1,pd1]

two=[bd2,hr2,pd2]

three=[bd3,hr3,pd3]

four=[bd4,hr4,pd4]

five=[bd5,hr5,pd5]

## for data to be stacked

b\_three=list(np.add(one,two))

b\_four=list(np.add(b\_three,three))

b\_five=list(np.add(b\_four,four))

plt.bar(x,one,label="Job Grade 1",width=0.6,color="blue")

plt.bar(x,two,bottom=one,label="Job Grade 2",width=0.6,color="darkorange")

plt.bar(x,three,bottom=b\_three,label="Job Grade 3",width=0.6, color="grey")

plt.bar(x,four,bottom=b\_four,label="Job Grade 4",width=0.6, color ="yellow")

plt.bar(x,five,bottom=b\_five,label="Job Grade 5",width=0.6,color="green")

plt.ylim(0,120)

plt.ylabel("Employee count (Left Company)")

plt.xlabel("Business Unit")

plt.title("Leavers' job grade within departments")

plt.legend(bbox\_to\_anchor=(1, 1), loc='upper left')

plt.show()

**Appendix 3: Codes for creating chart 3, box plots, in Python**

##import pandas package, numpy and matplotlib

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

employee= pd.read\_csv('GBA.csv')

##double checking if the columns are correct

employee.columns

##filtering out those who left the company and those who did not

left= employee.query('LeftCompany=="Yes"')

stay= employee.query('LeftCompany=="No"')

#print chart 3's corresponding table

print(left.drop(columns = ["BusinessUnit", "FieldOfStudy", "Gender", "Age", "ProximityToHome", "Travel", "JobGrade",

"JobRole", "MaritalStatus", "MonthlyPay", "PreviousEmployers", "SalaryIncrement",

"PerformanceRating", "WorkingExperience", "TrainingAttended", "WorkLifeBalance", "WithCompany",

"CurrentRole", "LastPromoted", "CurrentManager", "SatisfactionCompany", "SatisfactionCoworkers",

"Education"]))

print(stay.drop(columns = ["BusinessUnit", "FieldOfStudy", "Gender", "Age", "ProximityToHome", "Travel", "JobGrade",

"JobRole", "MaritalStatus", "MonthlyPay", "PreviousEmployers", "SalaryIncrement",

"PerformanceRating", "WorkingExperience", "TrainingAttended", "WorkLifeBalance", "WithCompany",

"CurrentRole", "LastPromoted", "CurrentManager", "SatisfactionCompany", "SatisfactionCoworkers",

"Education"]))

plt.boxplot(left.SatisfactionRole, positions=[0])

plt.boxplot(stay.SatisfactionRole, positions=[1])

## cleaning up axes

plt.title("Attrition vs SatisfactionRole")

plt.xticks([0,1], ['Yes', 'No'])

plt.xlabel("Left Company")

plt.ylabel("Satisfaction Role Rating")

plt.show()

**Appendix 4: Codes for creating chart 4, scatter plot, in Python**

# import pandas package into the program

import pandas as pd

# use pandas function to import and read the dataset saved as .csv from external source

EE\_data = pd.read\_csv("GBA.csv")

# select TrainingAttended and PerformanceRating column for chart 4's dataframe

chart4\_df = EE\_data[["TrainingAttended", "PerformanceRating"]]

print(chart4\_df)

# import matplotlib visualisation package into the program

import matplotlib.pyplot as plt

# plot scatterplot with x and y values

x = EE\_data["TrainingAttended"]

y = EE\_data["PerformanceRating"]

plt.scatter(x, y)

# import numpy package into the program

import numpy as np

# add trendline in scatterplot

z = np.polyfit(x, y, 1)

p = np.poly1d(z)

plt.plot(x, p(x))

# apply plot options to add chart title, and axes labels

plt.title("Performance rating and Training attended scatterplot")

plt.xlabel("Number of Training Attended")

plt.ylabel("Performance rating")

**Appendix 5: Codes for Q1 (c)**

# User interactive input to check the employee information

# It will keep work in loop until user expresses not to continue

# Import pandas library

import pandas as pd

# Import data set

EmployeeInformation = pd.read\_csv("GBA.csv")

# Display data set

print(EmployeeInformation)

# Initial variable list

EmployeeProfile = ['1.Current staff strength','2.Current staff annual renumeration','3.Eligible long service award recipients','4.Exit']

# Tell user to input the correct option

print(f"This is the list of employee profile {EmployeeProfile}. Please enter any one of 1, 2, 3 and 4")

# Loop function

continue\_query = 'yes'

while continue\_query == 'yes':

# Create a user input option for the employee information

user\_input = int(input("Enter the option you want to look for: "))

# Provide information of the current total staff strength, and a breakdown of staff strength by business unit and gender

if user\_input == 1:

print("There are 1057 people in the company.\nBusiness Devlopment: 299 people;Product Development: 716 people;HR: 42 people\nMale: 623 people;Female: 434 people")

# Provide information of the current total annual staff remuneration expenditure, and the current median monthly salary

elif user\_input == 2:

print("The current total annual staff remuneration expenditure is $87,182,508\nThe current median monthly salary is $5,238 ")

# Output information of the employees (EmpID), job role and age, ordered by descending age, with 35 or more years of service

elif user\_input == 3:

print(EmployeeInformation.iloc[[295,466,1242,896],:])

print("Employees with 35 or more years of service:\nEmpID: 1295 Job role: Account Manager Age: 59 With company: 37 years\nEmpID: 1466 Job role: Section Manager Age: 55 With company: 36 years\nEmpID: 2242 Job role: Section Manager Age: 57 With company: 36 years\nEmpID: 1896 Job role: Section Manager Age: 59 With company: 35 years")

# Asking user whether want to exit the program

else:

continue\_query = str(input("Would you like to continue? (Indicate yes or no)"))