

**ANL252**

**Python for Data Analytics**

# **Tutor-Marked Assignment**

**July 2022 Presentation**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Chan Zhi Hao** | **N2070913** |

**Tutorial Group: ­­­­­­­­­­­­ T 9**

**Instructor’s Name: Dr. Kumar Munish**

**Submission Date: 07/08/2022**

Q1a.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **Average of Salary** | **Column Labels** |  |  |
| **Row Labels** | **F** | **M** | **Grand Total** |
| Admin | 80159.5 | 73523 | 77315.28571 |
| C-Level | 250000 | 0 | 250000 |
| Engineering | 99044.6 | 92689.4 | 95867 |
| IT | 87905.38889 | 101534.087 | 95550.7561 |
| Manufacturing | 59655.22857 | 59015.34375 | 59412.90533 |
| Sales | 75900.27273 | 66952.90909 | 71426.59091 |
| **Grand Total** | **67686.52778** | **71063.81132** | **69118.496** |
|  |  |  |  |

Table 1.1: Average salary for individual department

Figure 1.1: Line chart of average salary for the individual department

Figure 1.1 shows a simple line chart with markers of the salary for the individual department. There are 105 female employees in manufacturing and only 1 female employee at C-level. Based on the average, the only C-level female employee earns 250000 and the manufacturing each female employee earns 59655. There are 64 male employees under manufacturing and 3 males’ employee in the admin. Based on the average, IT male employee earn the most of 101534 and the manufacturing each male employee earn 59015.

There are a few factors that will affect the employee's salary and they are years of experience, education, gender, and performance reviews. NUS research shows in Singapore median pay of women and men working full-time reveals that women earn 16.3 per cent less than men in 2018. For the marital, citizenship and minority will not affect the employee's salary. These three factors also will not affect the employee's overall performance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Row Labels** | **PerformanceScore (Meet and Exceed)** | **Survey (4 and above)** | **Satisfaction (4 and above)** | **Absence (10 days and below)** |
| **Admin** | **2** | **10** | **8** | **7** |
| **Engineering** | **3** | **13.61** | **13** | **8** |
| **IT** | **15** | **67.02** | **67** | **101** |
| **Manufacturing** | **37** | **165.23** | **166** | **187** |
| **Sales** | **4** | **18.32** | **19** | **26** |
| **Grand Total** | **61** | **274.18** | **273** | **329** |

Table 1.2: Individual department overall performance

|  |  |  |
| --- | --- | --- |
| Unit (Department) | Percentage Performance score (X-axis) | Average Point  (Y-axis) |
| Admin | 28.6 | 3.6 |
| Engineering | 30 | 3.5 |
| IT | 36.6 | 5.7 |
| Manufacturing | 21.9 | 3.1 |
| Sales | 18.2 | 2.9 |

Table 1.3: Scatter plot data

Figure 1.2: Scatter Plot for the individual department overall score

Figure 1.2 shows a stacked column of the individual department's overall performance. It showed manufacturing has the higher and admin have the lower overall performance. There are 7 admin staff, 10 engineering staff, 41 IT staff, 169 manufacturing staff, 22 sales staff and 1 C-level staff altogether. For the best performance, I’m looking at 3 criteria. First, satisfaction must scale 4 and above. Second, engagement survey results must be 4 and above. Last, the absence must be 10 days and below. After looking at all the criteria, the overall best performance is the IT department which has a 36.6%. The percentage performance score is calculated by (15/41) X 100% = 36.6% and average point is 235.02/41= 5.7. 15 out of 41 employees have met the 3 criteria listed above. The lowest score is sales department which has a (4/22) X 100% = 18.2% and average point is 63.32/22 = 2.9.

Summaries, the only C-level female employee earns the highest salary, but we can’t take this into consideration because the data will only survey by 1 person in that department. We will take the second highest salary which will be IT males and the average salary is $101534. Then the lower salary will be the manufacturing male and the average salary is $59015. In this line chart, I am looking at the average salary for individual departments.

I have used the 3 criteria (satisfaction, engagement survey and absence) to calculate the average point of the individual department. IT department has the highest average point of 5.7 and the highest percentage performance score of 36.6%. The sales department has the lowest average point of 2.9 and the lowest percentage performance score of 18.2%. In the scatter plot. I have used the percentage performance score as an x-axis and the average point as a y-axis to plot the chart. In this scatter plot, I am looking at the overall score for the individual departments.

Q1b.

|  |  |
| --- | --- |
| In  [1]: | import matplotlib.pyplot as plt |
| In  [2]: | Female\_scores = [80160,250000,99045,87905,59655,75900]  Male\_scores = [73523,0,92689,101534,59015,66952]  Department = ["Admin","C-Level","Engineering","IT","Manufacturing","Sales"] |
| In  [3]: | plt.plot(Department, Female\_scores, color = "green", marker = "o", markerfacecolor = "black",markeredgecolor="black")  plt.plot(Department, Male\_scores, color = "blue", marker = "o", markerfacecolor = "black",markeredgecolor="black")  plt.xlabel("Department")  plt.ylabel("Average salary")  plt.xticks(Department,labels = Department)  plt.yticks([50000,100000,150000,200000,250000,300000],["50000","100000","150000","200000","250000","300000"])  plt.grid(axis="y")  plt.legend(["F", "M"])  plt.title("Average salary for the individual department")  plt.xticks(Department, labels = Department)  plt.show() |
| Out  [3]: | **Screenshot** |
| In  [4]: | x=28.6  y=3.6  x1=30  y1=3.5  x2=36.6  y2=5.7  x3=21.9  y3=3.1  x4=18.2  y4=2.9 |
| In[5]: | plt.scatter(x,y, s=100, c='orange',marker='s',edgecolor='orange', linewidth=1,alpha=0.75)  plt.scatter(x1,y1, s=100, c='grey',marker='^',edgecolor='grey', linewidth=1,alpha=0.75)  plt.scatter(x2,y2, s=100, c='yellow',marker='x',edgecolor='yellow', linewidth=1,alpha=0.75)  plt.scatter(x3,y3, s=100, c='black',marker='2',edgecolor='black', linewidth=1,alpha=0.75)  plt.scatter(x4,y4, s=100, c='blue',marker='o',edgecolor='blue', linewidth=1,alpha=0.75)  plt.xlabel("Percentage performance score")  plt.ylabel("Average point")  plt.grid()  plt.xticks([0,5,10,15,20,25,30,35,40],["0","5","10","15","20","25","30","35","40"])  plt.yticks([1,2,3,4,5,6],["1","2","3","4","5","6"])  plt.legend(["Admin","Engineering","IT","Manufacturing","Sales"],loc='upper left')  plt.title("Overall score")  plt.show() |
| Out  [5]: | **Screenshot** |

Comment:

Import matplotlib package helps us plot the bar-chart, histogram, and scatter plot. I have used 2 charts (a line chart and a scatter plot) to plot the data. Then enter the data to plot.

1. Function for line chart plot

plt. plot (x , y , color , linestyle , linewidth , marker , markerfacecolor , markeredgecolor , markersize)

2. Function for scatter plot

plt. scatter (x , y , color = None , marker = None , linewidths = None , edgecolors = None)

Q1c.

|  |  |
| --- | --- |
| In [1]: | import pandas as pd  import matplotlib.pyplot as plt  import numpy as np  from datetime import datetime |
| In [2]: | data=pd.read\_csv("TMA\_Data.csv") |
| In [3]: | data.head() |
| Out[3]: | **Screenshot** |
| In [4]: | Join\_Date = data['JoinDate']=pd.to\_datetime(data['JoinDate'])  Left\_Date = data['LeftDate']=pd.to\_datetime(data['LeftDate'])  print(Join\_Date,Left\_Date) |
| Out[4]: | **Screenshot** |
| In [5]: | Left\_Date2 = Left\_Date.fillna("2022-05-01")  print(Left\_Date2) |
| Out[5]: | **Screenshot** |
| In [6]: | Days = (Left\_Date2 - Join\_Date).dt.days  print(Days) |
| Out[6]: | **Screenshot** |
| In [7]: | Years = Days / 365 |
| In [8]: | max\_number = max(Years)  print("The maximum length of service is",round(max\_number,1),"years") |
| Out[8]: | The maximum length of service is 16.3 years |
| In [9]: | data\_min = data["Days"].min(axis=0)  print("The minimum length of service is",round(data\_min,1),"years") |
| Out[9]: | The minimum length of service is 0.1 years |
| In[10]: | mean\_number = sum(Years)/ len(Years)  print("The average length of service is",round(mean\_number,1),"years") |
| Out[10]: | The average length of service is 6.8 years |

Comment:

Import the library and the file. To check the correct file, I have the key in data.head() to see the file. I am using pandas.to\_datetime function to convert the argument to datetime. Then use the pandas fillna function to fill in the missing data. Since the result wants to be in years, divide 365 because 1 year has 365 days. Lastly, enter the statement.

Q1d.

|  |  |
| --- | --- |
| In [1]: | import matplotlib.pyplot as plt  import numpy as np  from datetime import datetime  import pandas as pd |
| In [2]: | data1=pd.read\_csv("TMA\_Data.csv") |
| In [3]: | newcolumn1 = data1.sort\_values(by="LeftDate")  pd.set\_option("display.max\_column",None)  pd.set\_option("display.max\_row",None)  newcolumn1.head(251)  extract = newcolumn1[['Staff','LeftDate']].reset\_index(drop= True)  extract.head(251)  y = extract.values[0:85,0]  z = extract.values[0:85,1] |
| In [4]: | left = {}  for i in range(len(z)):  left[y[i]] = z[i]  current = extract.values[85:250, 0] |
| In [5]: | today = datetime.today()  d1 = today.strftime("%d/%m/%Y") |
| In [6]: | name=str(input("Please enter the staff name "))  answer = name  def gameplay():  answer = input(name)  if answer in left:  print("The employee has left the organisation on {}".format(left[answer]))  elif answer in current:  print("This employee still working as of", d1)  else:  print("No such person in the organisation")  while True:  answer1 = input("Do you want to make another query within the organisation? (Y/N) ").upper()  if answer1 == "Y":  gameplay()  elif answer1 == "N":  print("Your search ends here \nThank you and have a nice day!")  break  else:  print(" Please enter Y or N ") |
| Out[6]: | **Screenshot** |

Comment:

Import library and file. Then sort the value of the left date and use pd.set\_option to set the value of the specified option. pandas.DataFrame.reset\_index to reset the index or a level of it. 85 people have left the organisation. Then we can know left and current employees in the organisation. Then using the if-else loop and while loop is for decision making. First, will ask the user to enter the employee that the user wants to search. If the employee has left or is current, it will display the statement message. Hence, if the user enters the wrong employee’s name it will show “No such person in the organisation.” While loop is the user enters the correct employee then it will execute the statement.

Usually, I will print() each step so as to know whether I am on the right path before proceeding to the next step.