

**ANL252**

**Python for Data Analytics**

# **Tutor-Marked Assignment**

**July 2022 Presentation**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Shun Pei Ling Felicia** | **H1882537** |

**Tutorial Group: ­­­­­­­­­­­­ T 09**

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

**Submission Date: 14/08/2022**

**Question 1**

Table

Description automatically generated

*Figure 1*

*Figure 2*

The histogram in Figure 2 illustrates the frequency distribution of the salaries of the 250 staffs from an organization. The values for salary in the histogram are categorized to decrease the number of bins and the height of each bar represents the number of staffs that belong to each bin.

Based on the above visual representation, we can infer that the distribution is not symmetrical. In other words, it is skewed towards the left and hence, the distribution is said to be positively skewed. In this case, the mean would be larger than the median or the mode. From the above graph, we notice that majority of the staffs earn a salary between $50,000 and $62,499. It comprises 92 staffs as observed from the pivot table in Figure 1, accounting for the highest proportion of 37% of the total staff population.

On the other hand, we have noted that there are outliers and extreme values. With reference to Figure 1, there are 2 staffs earning a salary range between $175,000 and $187,499 and this forms the outlier. There is a high earner earning a salary falling in the range of $237,500 to $250,000 and this forms the extreme value. The presence of these high earners is the source of the skewed distribution as the extreme salary value that are much greater than the mean outweighs the fewer number of salary value which are much lower than the mean.

Table

Description automatically generated

*Figure 3*

*Figure 4*

The above bar chart illustrates the demographics of staffs by its functional units. We have observed that the Manufacturing unit is the largest unit in the organization in terms of its size. It comprises a total of 169 staffs, accounting for 68% of the total staff population. This provides an insight to the nature of the business the organization is engaging. With a proportion accounting for more than half of the population, it may suggest that the organization may be a manufacturing company. In contrast, the smallest unit belongs to the C-Level unit. With this, we deduce that the organization may be a small firm as it is led by only 1 corporate level executive staff. This can be further supported by a small population of 250 staffs. In addition, this staff may be source of the skewed distribution of salaries observed in Figure 2, earning the highest salary ranging between $237,500 to $250,000, considering of the fact that she is the owner of the organization.

According to the chart, we can infer that the Manufacturing unit is predominantly female. Out of the 169 staffs, 105 or 62% of them are female. This is an interesting observation considering that the manufacturing industry is often perceived as labour-intensive and dominated by males. A possible rationale is that it is a woman-owned or -led company, as indicated by the blue colour bar under C-Level unit, which refers to female gender. Such company would be more likely to implement human resource employment policies and practices that are gender-responsive such as maternity leave and better job security for women. These features may have appealed to the female employees or its female employee retention, resulting in an overall higher female proportion as compared to males. Another possible rationale could be that the organization is a beauty manufacturer, producing cosmetic products. Considering that females are consumers of beauty products themselves, such industry may appeal more to them as they are able to bring value to the company based on their usage experience.

On the other hand, the IT unit is the second largest unit, comprising of 41 staffs or 16% of the total staff population. This is of no surprise as the IT unit plays a vital role in the manufacturing industry, ensuring that the equipment used in the manufacturing process function according to programs, as well as innovate new systems to automate manual procedures. Unlike the Manufacturing unit, the IT unit is predominantly male. It consists of 23 staffs, accounting for 56% of the IT unit. This could be due to the stereotypes and perception of the nature of STEM careers.

The third largest unit belongs to the Sales unit. It comprises of 22 staffs, accounting for 9% of the total population. In terms of gender ratio, there is an even proportion of males and females. Another unit that has an even gender proportion belongs to the Engineering unit.

1. Histogram

# import relevant modules with pd as an alias for pandas and plt as an alias for matplotlib.pyplot

import pandas as pd

import matplotlib.pyplot as plt

# load csv file "TMA\_Data.csv" into a dataframe name employee\_data

employee\_data = pd.read\_csv("/Users/feliciashun/Desktop/SUSS/Y5S1/ANL252/TMA\_Data.csv")

# display data rows

# employee\_data.head()

# capture dataframe

df = pd.DataFrame(employee\_data, columns = ["Salary"])

# cast "Price" column of the dataframe into a list

salary\_list = list(employee\_data["Salary"])

# salary\_min and salary\_max are the minimum and maximum value respectively of the salary list

salary\_min = min(salary\_list)

salary\_max = max(salary\_list)

# print salary\_min and salary\_max to see its values

print(salary\_min)

print(salary\_max)

# create bins and labels. Start from 37,500 as the lowest value to maximum value of 250,000.

salary\_bins = [37500,50000,62500,75000,87500,100000,112500,125000,137500,150000,175000,187500,200000,212500,225000,237500,250000]

salary\_labels = ['37500-49999','50000-62499','62500-74999','75000-87499','87500-99999','100000-112499','112500-124999','125000-137499','137500-149999','150000-174999','175000-187499','187500-199999','200000-212499','212500-224999','225000-237499','237500-250000']

df['salaryrange'] = pd.cut(df.Salary, bins = salary\_bins, labels = salary\_labels)

# create pivot table

pivot\_table = pd.pivot\_table(df, index ='salaryrange',values=['Salary'], aggfunc='count')

pivot\_table

# figure assigns an area for the chart. Figsize represents the figure size in inches.

plt.figure(figsize = (20,10))

# histogram with 20 bins, range from 0 to salary\_max, bins are centered between the bin edges, vertical orientation, relative width of 80%

plt.hist(salary\_list, bins = 20, range = (0, salary\_max), color = "orange", align = "mid", orientation = "vertical", rwidth = 0.8)

plt.title("Distribution of salaries", fontsize = 30, weight = "bold")

plt.xticks(ticks = range(37500, 262500, 12500), labels = range(37500, 262500, 12500))

plt.yticks(ticks = range(0, 110, 10), labels = range(0, 110, 10))

plt.xlabel("Salary ($)", fontsize = 18)

plt.ylabel("Number of staffs", fontsize = 18)

plt.show()

Bar Chart

# import relevant modules with pd as an alias for pandas and plt as an alias for matplotlib.pyplot

import pandas as pd

import matplotlib.pyplot as plt

# read csv file "TMA\_Data.csv" into a dataframe name employee\_data

employee\_data = pd.read\_csv("/Users/feliciashun/Desktop/SUSS/Y5S1/ANL252/TMA\_Data.csv")

# capture dataframe

df = pd.DataFrame(employee\_data, columns = ["Gender", "Unit"])

# create frequency columns based on gender

df['Frequency'] = df.groupby(["Gender"])["Gender"].transform("count")

# create pivot table. Null values replaced with 0 using fill\_value.

df\_pivot = df.pivot\_table(index = ["Unit"], values = ["Frequency"], columns = ["Gender"], aggfunc = "count", fill\_value = 0)

df\_pivot

# plot bar chart using the dataframe

bar\_chart = df\_pivot.plot.bar(stacked = True, color = ["red", "blue"], figsize = (10,10), rot = 1)

plt.title("Gender porportion of staffs by unit", fontsize = 20, weight = "bold")

plt.yticks(ticks = range(0, 190, 20), labels = range(0, 190, 20))

plt.xlabel("Unit", fontsize = 18)

plt.ylabel("Number of staffs", fontsize = 18)

1. # import relevant modules with pd as an alias for pandas and np as an alias for numpy

import pandas as pd

import numpy as np

# read csv file "TMA\_Data.csv" into a dataframe name employee\_data

employee\_data = pd.read\_csv("/Users/feliciashun/Desktop/SUSS/Y5S1/ANL252/TMA\_Data.csv")

# capture dataframe

df = pd.DataFrame(employee\_data)

#fill missing values NaN under LeftDate column with "1/5/2022"

df["LeftDate"].fillna("5/1/2022", inplace = True)

# format JoinDate date format

df["JoinDate"] = pd.to\_datetime(df.JoinDate)

# format LeftDate date format

df["LeftDate"] = pd.to\_datetime(df.LeftDate)

# calculate length of service in years

df['LengthOfService'] = df['LeftDate'] - df['JoinDate']

df['LengthOfService'] = df['LengthOfService']/np.timedelta64(1,'Y')

print(df)

# calculate minimum length of service

min\_srv = df['LengthOfService'].min()

min\_years = round(min\_srv,1)

# calculate maximum length of service

max\_srv = df['LengthOfService'].max()

max\_years = round(max\_srv,1)

# calculate average length of service

avg\_srv = df['LengthOfService'].mean()

avg\_years = round(avg\_srv,1)

print(f"The minimum, maximum and average length of service of the staffs are {min\_years}, {max\_years}, {avg\_years} years respectively.")

Based on the output of the code written, the minimum, maximum and average length of service of the staffs are 0.1, 16.3 and 6.8 years respectively.

1. # import relevant modules with pd as an alias for pandas

import pandas as pd

# read csv file "TMA\_Data.csv" into a dataframe name employee\_data

employee\_data = pd.read\_csv("/Users/feliciashun/Desktop/SUSS/Y5S1/ANL252/TMA\_Data.csv")

# capture dataframe

df = pd.DataFrame(employee\_data, columns = ["Staff"])

# print user input using while loop to allow iterations unless user chooses otherwise

print("--------------------------------------")

print("Welcome to our organization's records.")

valid\_input = False

while valid\_input == False:

try:

enquire\_input = str(input("Would you like to enquire if a person is a staff of our organization? (Y/N):"))

if enquire\_input == "Y":

input\_name = str(input("Please enter the full name that is registered with our organization:"))

if input\_name in df.values:

print(f"The name, {input\_name}, can be found in our organization's records.")

else:

print(f"The name, {input\_name}, cannot be found in our organization's records.")

elif enquire\_input == "N":

valid\_input = True

print("The enquiry shall end.")

except ValueError:

print("Wrong out")