Course Code: ANL 252 Python for Data Analytics

# Assignment: TMA01

# Name: Shawn Ng Zhen Xiang

# SUSS PI No: Q2172585

# Submission Date: 13 August 2022

**Question 1a)**



Based off a 10 bin histogram, we can see that majority of the company falls between the 45k-65k range in salary. Which is about 153/250. Also 94% of the employees have a salaray that is below 106k which is the third bin.

When we compare Survey results(X axis) to the Salary (Y Axis) we can also determine that the spread of the survey is similar across the different pay scales at a lower range, of < 100k.

However, above that amount the higher paid employees tend to sit towards a higher survey result compared the lower paid employees.

**Question 1b)**

# import libraroes

import pandas as pd

import math

import numpy as np

# we use plt as an alias for matplotlib.pyplot

import matplotlib.pyplot as plt

# read the csv file

tma\_data = pd.read\_csv("tma\_data.csv")

print(tma\_data.head())

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

salary = list(tma\_data["Salary"])

# rmin and rmax are the minimum and maximum value of the price list

rmin = min(salary)

rmax = max(salary)

# figure allocates an area for the chart. Figsize defines the size of the figure in inches. dpi = dots per inch

plt.figure(figsize=(10,6), dpi=100)

# histogram with 10 bins, range from rmin to rmax, bins are centered between the bin edges, vertical orientation, relative width 0f 80%

plt.hist(salary, bins=10, range=(rmin, rmax), align="mid", orientation="vertical", rwidth=.8)

plt.title("Histogram of Salary")

plt.xlabel("Salary")

plt.ylabel("Frequency")

plt.style.use('ggplot')

plt.show()

#matplotlib.pyplot.scatter(x, y, s=None, c=None, marker=None, cmap=None, norm=None, vmin=None, vmax=None, alpha=None, linewidths=None, \*, edgecolors=None, plotnonfinite=False, data=None, \*\*kwargs)[source]

#plt.scatter(x, y, s=area, c=colors, alpha=0.5)

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

x = list(tma\_data["Survey"])

y = list(tma\_data["Salary"])

gender = list(tma\_data["Gender"])

# plotting scatter graph of salary and Survey results

plt.figure(figsize=(10,6), dpi=100)

plt.scatter (x, y, alpha=1)

plt.xlabel("Survey")

plt.ylabel("Salary")

plt.show()

**Question 1c)**

tma\_data["Length\_Of\_Service"] = np.nan

from pandas import Timestamp

from statistics import mean

#adding new column

tma\_data["Length\_Of\_Service"] = np.nan

#changing date type

tma\_data['LeftDate']=tma\_data['LeftDate'].fillna(Timestamp('20220501'))

tma\_data[['LeftDate','JoinDate']]= tma\_data[['LeftDate','JoinDate']].apply(pd.to\_datetime)

#Calculations

tma\_data['Length\_Of\_Service'] = tma\_data['LeftDate'] - tma\_data['JoinDate']

tma\_data['Length\_Of\_Service'] = tma\_data['Length\_Of\_Service'].dt.days.astype('int16')

yearsmaxworked = round(int(max(tma\_data["Length\_Of\_Service"]))/365, 1)

yearsminworked = round(int(min(tma\_data["Length\_Of\_Service"]))/365, 1)

avg = round(mean (tma\_data['Length\_Of\_Service'])/365,1)

print (f"The maximum years worked is {yearsmaxworked}")

print (f"The minimum years worked is {yearsminworked}")

print (f"The average number of years worked is {avg}")

**Question 1d)**

import pandas as pd

import math

import numpy as np

tma\_data = pd.read\_csv("tma\_data.csv")

staff = list(tma\_data["Staff"])

while True:

user\_input = str(input("Who do you want to look for in the staff database"))

if user\_input in (staff):

yesno\_input= input(f"{user\_input} is in the database. Would you like to look for someone else? Yes/No")

if yesno\_input == 'Yes':

continue

elif yesno\_input == 'No':

break

else:

print(f"Error, please try again")

else:

yesno\_input= input(f"{user\_input} is not found. Would you like to look for someone else? Yes/No")

if yesno\_input == 'Yes':

continue

elif yesno\_input == 'No':

break

else:

print(f"Error, please try again")