

**ANL252 (Online)**

**Python for Data Analytics**

# **Tutor-Marked Assignment**

**July 2022 Presentation**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Mako Wang Jun** | **Y2172686** |

**Tutorial Group: ­­­­­­­­­­­­ T 09**

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

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

**Question 1**

**(a)**

|  |  |
| --- | --- |
| **Satisfaction Score** | **Frequency** |
| 1 | 1 |
| 2 | 4 |
| 3 | 84 |
| 4 | 80 |
| 5 | 81 |

|  |  |
| --- | --- |
| **Satisfaction** | |
| Mean | 3.944 |
| Median | 4 |
| Range | 4 |
| Maximum | 5 |

In the organization, 250 employees have their satisfaction score generated on a scale of 1 to 5. The bar chart represents the satisfaction score of the 250 employees.

We observe that the distribution on the satisfaction scores is skewing to the right which shows that majority of employees have satisfaction scores of 3 to 5. There are 5 data which are below 3. These seemed to be on the lower side.

The mean and median scores are 3.944 and 4 respectively. In this case, mean would be a better measurement of satisfaction score. The mean score represents the average satisfaction score of an employee and 3.944 is on the higher side of the scale.

As shown on the summarized table, we have a total of 5 employees with scores of 1 to 2, 2% of the total employees which is in an acceptable range.

|  |  |
| --- | --- |
| **Absence** | |
| Mean | 10.312 |
| Median | 10 |
| Mode | 4 |
| Standard Deviation | 5.858321922 |
| Skewness | 0.037526433 |
| Range | 19 |
| Minimum | 1 |
| Maximum | 20 |
| Sum | 2578 |
| Count | 250 |

|  |  |
| --- | --- |
| **Absence Rate** | **Frequency** |
| 0-3 | 39 |
| 4-6 | 43 |
| 7-9 | 36 |
| 10-12 | 27 |
| 13-15 | 44 |
| 16-18 | 34 |
| 19-21 | 27 |

In the organization, we have kept records of 250 employees and their amount of absence in the previous year. The histogram shows the absence rate of these employees. The absence rate ranges from 1 to 20.

We can observe that the distribution of employee’s absence rate is close to evenly spread. This shows that the data are evenly distributed and there are not outliers.

The mean and median absence rate are 10.312 and 10 respectively. In this case, mean would be a better measurement of absence rate.

The standard deviation is 5.858. The variation is not acceptable in this case since the coefficient of variation is 56.81%. As the CV is greater than 30%, there is a cause of concern for employees that are taking too much absence days.

**(b)**

# creation of DataFrame

import pandas as pd

import numpy as np

#Create a Dictionary of series

d = {'Satisfaction Score':pd.Series(['1','2','3','4','5']),

'Frequency':pd.Series([1,4,84,80,81])}

#Create a DataFrame

df = pd.DataFrame(d)

print(df)

import pandas as pd

df = pd.read\_csv (r'TMA\_Data.csv')

# To create summary statistics

mean1 = df['Satisfaction'].mean()

sum1 = df['Satisfaction'].sum()

max1 = df['Satisfaction'].max()

min1 = df['Satisfaction'].min()

count1 = df['Satisfaction'].count()

median1 = df['Satisfaction'].median()

std1 = df['Satisfaction'].std()

var1 = df['Satisfaction'].var()

print ('Mean: ' + str(mean1))

print ('Sum: ' + str(sum1))

print ('Max: ' + str(max1))

print ('Min: ' + str(min1))

print ('Count: ' + str(count1))

print ('Median: ' + str(median1))

print ('Std: ' + str(std1))

print ('Var: ' + str(var1))

# importing the required module

import matplotlib.pyplot as plt

# x-coordinates of left sides of bars

left = [1, 2, 3, 4, 5]

# set limits

plt.ylim(top=90)

# heights of bars

height = [1, 4, 84, 80, 81]

# labels for bars

tick\_label = ['1', '2', '3', '4', '5']

# plotting a bar chart

plt.bar(left, height, tick\_label = tick\_label,

width = 0.6, color = ['blue'])

# naming the x-axis

plt.xlabel('Satisfaction Score')

# naming the y-axis

plt.ylabel('Frequency')

# plot title

plt.title('Satisfaction Scores of 250 Employees')

# function to show the plot

plt.show

# creation of DataFrame

import pandas as pd

import numpy as np

#Create a Dictionary of series

d = {'Absence Rate':pd.Series(['0-3','4-6','7-9','10-12','13-15','16-18','19-21']),

'Frequency':pd.Series([39,43,36,27,44,34,27])}

#Create a DataFrame

df = pd.DataFrame(d)

print(df)

import pandas as pd

df = pd.read\_csv (r'TMA\_Data.csv')

# To create summary statistics

mean1 = df['Absence'].mean()

sum1 = df['Absence'].sum()

max1 = df['Absence'].max()

min1 = df['Absence'].min()

count1 = df['Absence'].count()

median1 = df['Absence'].median()

std1 = df['Absence'].std()

var1 = df['Absence'].var()

print ('Mean: ' + str(mean1))

print ('Sum: ' + str(sum1))

print ('Max: ' + str(max1))

print ('Min: ' + str(min1))

print ('Count: ' + str(count1))

print ('Median: ' + str(median1))

print ('Std: ' + str(std1))

print ('Var: ' + str(var1))

# importing the required module

import matplotlib.pyplot as plt

import numpy as np

# frequencies

absence = [15,20,14,20,2,9,1,17,19,5,1,14,4,4,12,12,14,16,13,15,20,9,12,4,13,4,16,16,6,14,6,9,4,1,9,20,8,13,11,6,11,18,10,10,7,13,10,2,19,20,10,14,9,2,3,2,3,10,4,1,4,13,9,14,7,2,3,8,5,6,15,14,16,3,20,16,7,9,3,18,11,8,3,9,19,3,16,1,18,13,7,16,6,4,2,19,16,16,3,1,19,4,9,13,17,16,19,17,8,11,13,7,15,2,10,15,19,1,12,3,8,1,15,9,7,19,17,2,11,20,10,4,11,11,15,15,7,2,13,11,2,19,16,16,13,12,14,4,1,20,8,6,2,5,7,20,14,9,1,15,13,7,16,16,6,11,19,8,16,5,4,6,17,19,2,18,17,7,20,15,16,18,19,18,14,8,14,20,5,15,3,20,16,14,12,15,4,4,6,13,7,4,6,19,5,11,11,19,4,10,7,6,7,1,9,14,3,6,14,3,16,8,5,15,17,14,20,5,2,17,7,11,3,4,4,13,7,12,6,4,18,2,6,2,5,15,14,4,13,16,]

# set limits

plt.ylim(top=50)

# setting the ranges and no. of intervals

range = (0, 21)

bins = 7

# plotting a histogram

plt.hist(absence, bins, range, color = 'blue',

histtype = 'bar', rwidth = 0.8)

# x-axis label

plt.xlabel('Number of Absence')

# frequency label

plt.ylabel('Frequency')

# plot title

plt.title('Absence rate of 250 Employees')

# function to show the plot

plt.show()

**(c)**

# importing required module

import pandas as pd

df = pd.read\_csv("TMA\_Data.csv")

#replace all NA cells in LeftDate column with 1st May 2022

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

df.head()

#convert from string to datetime format

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

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

df.head()

#create new column showing difference in datetime

df['LengthOfService'] = ((df['LeftDate'] - df['JoinDate']).dt.days/365).round(1)

df.head()

# Calculate min, max and average length of service

MaxLengthOfService = df['LengthOfService'].max()

print('Maximum Length Of Service: '+str(MaxLengthOfService))

MinLengthOfService = df['LengthOfService'].min()

print("Minimum Length Of Service: "+str(MinLengthOfService))

AvgLengthOfService = df['LengthOfService'].mean()

print("Average Length Of Service: "+str(AvgLengthOfService))

**(d)**

# importing required module

import pandas as pd

import numpy as np

import csv

filepath = "TMA\_Data.csv"

with open(filepath) as csvfile:

reader = csv.DictReader(csvfile)

Staff = input("Please input full name of staff name:")

for row in reader:

if Staff == row['Staff']:

print(f"This person was/is a staff of the organization")

break

else:

print('This staff does not exist ')

terminate\_qns = str(input("Do you want to go again (Y/N):"))

if terminate\_qns == "Y":

Staff = input("Please input full name of staff name:")

for row in reader:

if Staff == row['Staff']:

print(f"This person was/is a staff of the organization")

break

else:

print('This staff does not exist ')

elif terminate\_qns == "N":

print("Goodbye")

else:

print("Error in input")