

**ANL252\_Python for Data Analytics**

**Tutor-marked Assignment – July Semester 2022**

**Student PI No: E2210907**

**Full Name:** Lynn Tan Ee Leng

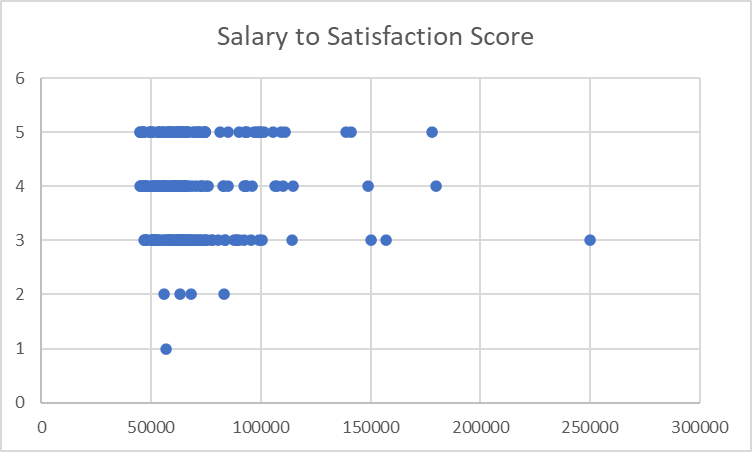
**Submission Date:** 14 August 2022

1a.

**Hypothesis 1**

Employees with higher salaries will have higher satisfaction scores.

Plotting a chart of salary to satisfaction score using the “Salary” and “Satisfaction” columns:



Conclusion: The data does not support the hypothesis. There is no obvious correlation between salary and satisfaction score. In fact, the employee with the highest salary gave an average satisfaction score of 3.

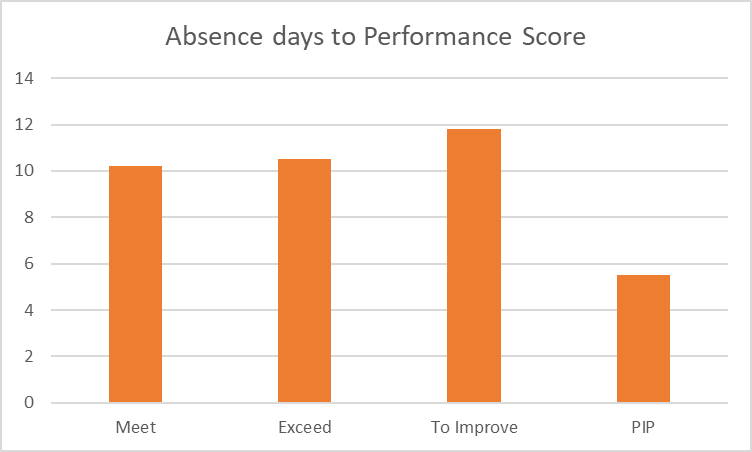
**Hypothesis 2**

Employees with fewer absence days had better performance scores.

Employees are grouped according to their performance score before calculating the mean absence days in each group, using the “AVERAGEIF” function in excel.

|  |  |  |
| --- | --- | --- |
| Meet |  | 10.199 |
| Exceed |  | 10.53333 |
| To Improve | | 11.82353 |
| PIP |  | 5.5 |

Plotting a graph of absence days to performance score:



Conclusion: Those with performance scores of “Meet” and “Exceed” have lower absence days compared to “To Improve”. This supports the hypothesis that lower absence days lead to better performance scores. The data on “PIP” is inconclusive since there is only 1 employee with such a score thus unable to come to a meaningful conclusion.

1b.

**Chart 1**

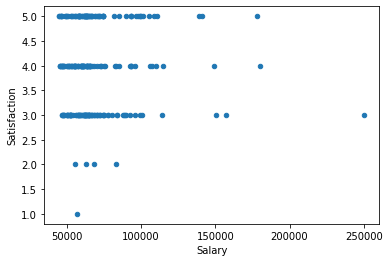
Code:

import pandas as pd

df = pd.read\_csv('TMA\_Data.csv', parse\_dates=['BirthYear', 'JoinDate', 'LeftDate'])

df.plot.scatter(x='Salary', y='Satisfaction')

Output:



**Chart 2**

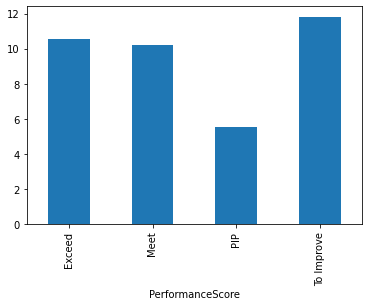
Code:

df\_groups = df.groupby('PerformanceScore')

mean\_absence\_of\_performance = df\_groups['Absence'].agg('mean')

mean\_absence\_of\_performance.plot.bar()

Output:



1c.

Code:

df.LeftDate.fillna('2022-05-01', inplace=True)

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

from datetime import timedelta

year = timedelta(days=365)

minimum\_duration = df.Duration.min()

print(f"{minimum\_duration / year :.1f}")

maximum\_duration = df.Duration.max()

print(f"{maximum\_duration / year :.1f}")

mean\_duration = df.Duration.mean()

print(f"{mean\_duration / year :.1f}")

Output:

The minimum duration of service is 0.1 years.

The maximum duration of service if 16.3 years.

The average length of service is 6.8 years.

1d.

Code:

def program():

while True:

print("Enter name to check if person is in organization or hit <enter> key to quit")

name = input("Enter: ")

if name == '':

break

elif name in df.Staff.values:

print(f"{name} is a staff in the organization")

else:

print(f"Could not find {name}")

print("Application Ended")