Employee Case Study

September 9, 2023

[1]:
```
import numpy as np
from sklearn.linear_model import LinearRegression
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split # Sklearn package's␣
↪randomized data splitting function
from scipy.stats import pearsonr
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
emp_data = pd.read_excel("I:\\Internship\\Employee Data - 4th July\\Case Study␣
↪1.xlsx",sheet_name='Filtered', header=0)
emp_data.head()
```

[2]:
```
   Respondent ID What is your gender? What is your race or ethnicity?   Age  \
0              1                    M  White (Not Hispanic or Latino)  57.0
1              2                    F              Hispanic or Latino  35.0
2              3                    M              Hispanic or Latino  59.0
3              4                    F       Black or African American   NaN
4              5                    M  White (Not Hispanic or Latino)  34.0

   What is your yearly CTA salary?  CTA Tenure (Months)   \
0                        102544.00                 66.0
1                         85386.08                179.0
2                         84439.68                111.0
3                         84439.68                  NaN
4                         77316.59                 70.0

               At which location do you work?  \
0                              S/S Heavy Mtce
1                          Chicago Ave Garage
2                          North Park  Garage
3                          Forest Glen Garage
4  567 W Lake Street - Main Location for CTA
```

```
                 What is your Position? Are you a manager or above?  \
0                      Painter (Various)                          NO
1                    Mobile Bus Mechanic                          NO
2                           Bus Mechanic                          NO
3                           Bus Mechanic                          NO
4  Project Specialist II - Communications                          NO

   Fit Item 1  …  Stay Factor: Pay/Salary  \
0           4  …               Pay/Salary
1           3  …               Pay/Salary
2           4  …               Pay/Salary
3           4  …               Pay/Salary
4           4  …               Pay/Salary

   Stay Factor: Coworker Relationships  Stay Factor: Grievance Handling  \
0              Coworker Relationships                              NaN
1                                 NaN                              NaN
2                                 NaN                              NaN
3                                 NaN               Grievance Handling
4                                 NaN                              NaN

   Stay Factor: Job Satisfaction  Stay Factor: Challenging Work  \
0                            NaN               Challenging Work
1              Job Satisfaction               Challenging Work
2                            NaN               Challenging Work
3                            NaN                            NaN
4                            NaN                            NaN

   Stay Factor: Rewards & Recognition  Stay Factor: Safety  \
0                                 NaN                  NaN
1                                 NaN                  NaN
2              Rewards & Recognition               Safety
3                                 NaN               Safety
4              Rewards & Recognition                  NaN

  Stay Factor: Workload  # of Safety Incidents  # of Absent Days/Tardiness
0                   NaN                      0                           1
1                   NaN                      0                           1
2              Workload                      0                           1
3                   NaN                      0                           1
4                   NaN                      0                           1

[5 rows x 71 columns]
```

[3]:
```
emp_data.describe()
```

[3]:

	Respondent ID	Age	What is your yearly CTA salary?	\
count	1498.00000	1455.000000	1498.000000	
mean	749.75968	47.317526	89856.098465	
std	432.91285	9.876587	20146.816820	
min	1.00000	22.000000	32780.800000	
25%	375.25000	40.000000	80419.040000	
50%	749.50000	49.000000	80419.040000	
75%	1124.75000	55.000000	99919.040000	
max	1499.00000	69.000000	367790.340000	

	CTA Tenure (Months)	Fit Item 1	Fit Item 2	Fit Item 3	\
count	1454.000000	1498.000000	1498.000000	1498.000000	
mean	127.186382	2.995995	3.008011	3.078772	
std	124.533128	1.332492	1.362712	1.352082	
min	4.000000	1.000000	1.000000	1.000000	
25%	15.000000	2.000000	2.000000	2.000000	
50%	73.000000	3.000000	3.000000	3.000000	
75%	242.000000	4.000000	4.000000	4.000000	
max	438.000000	5.000000	5.000000	5.000000	

	HiPo Item 1	HiPo Item 2	HiPo Item 3	…	\
count	1454.000000	1454.000000	1498.000000	…	
mean	3.143054	3.347318	1.896529	…	
std	1.483621	1.264586	0.840270	…	
min	1.000000	1.000000	1.000000	…	
25%	2.000000	3.000000	1.000000	…	
50%	3.000000	3.000000	2.000000	…	
75%	4.000000	4.000000	3.000000	…	
max	33.000000	5.000000	5.000000	…	

	Satisfaction Rank: Management	Satisfaction Rank: Organizational Fit	\
count	1451.000000	1451.000000	
mean	3.323225	3.367333	
std	1.741067	1.757979	
min	1.000000	1.000000	
25%	2.000000	2.000000	
50%	3.000000	3.000000	
75%	4.000000	5.000000	
max	9.000000	9.000000	

	Satisfaction Rank: Career Opportunity	\
count	1451.000000	
mean	3.275672	
std	1.783397	
min	1.000000	
25%	2.000000	
50%	3.000000	

```
75%                                    4.000000
max                                    9.000000

       Satisfaction Rank: Work Environment  \
count                             1451.000000
mean                                 3.248105
std                                  1.707727
min                                  1.000000
25%                                  2.000000
50%                                  3.000000
75%                                  4.000000
max                                  9.000000

       Satisfaction Rank: Clear Job Expectations  \
count                                   1451.000000
mean                                       3.440386
std                                        1.838964
min                                        1.000000
25%                                        2.000000
50%                                        3.000000
75%                                        5.000000
max                                        9.000000

       Satisfaction Rank: Other (specify below)  \
count                                  1451.000000
mean                                      3.691247
std                                       2.257505
min                                       1.000000
25%                                       2.000000
50%                                       3.000000
75%                                       5.000000
max                                       9.000000

       How likely is it that you would recommend the Chicago Transit Authority
to a friend or colleague?  \
count                                           1263.000000
mean                                               7.759303
std                                                2.692484
min                                                0.000000
25%                                                7.000000
50%                                                9.000000
75%                                               10.000000
max                                               10.000000

       Stay Intention: I plan on working here for another (in years):  \
count                                           1498.000000
mean                                               7.114820
```

```
std                                                  5.055751
min                                                  0.000000
25%                                                  3.000000
50%                                                  7.000000
75%                                                  9.000000
max                                                 23.000000

       # of Safety Incidents  # of Absent Days/Tardiness
count            1498.000000                 1498.000000
mean                0.022029                    0.257677
std                 0.186865                    0.695696
min                 0.000000                    0.000000
25%                 0.000000                    0.000000
50%                 0.000000                    0.000000
75%                 0.000000                    0.000000
max                 3.000000                   14.000000

[8 rows x 54 columns]
```

[4]:
```
emp_data.isna().sum()
```

[4]:
```
Respondent ID                            0
What is your gender?                    27
What is your race or ethnicity?          0
Age                                     43
What is your yearly CTA salary?          0
                                      …
Stay Factor: Rewards & Recognition    1299
Stay Factor: Safety                   1161
Stay Factor: Workload                 1249
# of Safety Incidents                    0
# of Absent Days/Tardiness               0
Length: 71, dtype: int64
```

[5]:
```
emp_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1498 entries, 0 to 1497
Data columns (total 71 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Respondent ID
1498 non-null   int64
 1   What is your gender?
1471 non-null   object
 2   What is your race or ethnicity?
```

```
1498 non-null   object
 3   Age
1455 non-null   float64
 4   What is your yearly CTA salary?
1498 non-null   float64
 5   CTA Tenure (Months)
1454 non-null   float64
 6   At which location do you work?
1498 non-null   object
 7   What is your Position?
1498 non-null   object
 8   Are you a manager or above?
1498 non-null   object
 9   Fit Item 1
1498 non-null   int64
 10  Fit Item 2
1498 non-null   int64
 11  Fit Item 3
1498 non-null   int64
 12  HiPo Item 1
1454 non-null   float64
 13  HiPo Item 2
1454 non-null   float64
 14  HiPo Item 3
1498 non-null   int64
 15  HiPo Item 4
1498 non-null   int64
 16  Satisfaction Item 1
1454 non-null   float64
 17  Satisfaction Item 2
1498 non-null   object
 18  Engagement Item 1
1498 non-null   int64
 19  Engagement Item 2
1498 non-null   int64
 20  Engagement Item 3
1498 non-null   int64
 21  Engagement Item 4
1498 non-null   int64
 22  Engagement Item 5
1498 non-null   int64
 23  Engagement Item 6
1498 non-null   int64
 24  Motivation Item 1
1489 non-null   float64
 25  Motivation Item 2
1489 non-null   float64
 26  Motivation Item 3
```

```
1489 non-null   float64
 27  Motivation Item 4
1489 non-null   float64
 28  Performance Item 1
1489 non-null   float64
 29  Performance Item 2
1489 non-null   float64
 30  Leadership Item 1
1481 non-null   float64
 31  Leadership Item 2
1481 non-null   float64
 32  Leadership Item 3
1481 non-null   float64
 33  Support Item 1
1481 non-null   float64
 34  Support Item 2
1481 non-null   float64
 35  Commitment Item 1
1481 non-null   float64
 36  Commitment Item 2
1481 non-null   float64
 37  Commitment Item 3
1481 non-null   float64
 38  Commitment Item 4
1481 non-null   float64
 39  Commitment Item 5
1481 non-null   float64
 40  Diversity Item 1
1476 non-null   float64
 41  Diversity Item 2
1476 non-null   float64
 42  Diversity Item 3 (Reverse Coded)
1476 non-null   float64
 43  Trust Item 1
1476 non-null   float64
 44  Trust Item 2
1476 non-null   float64
 45  Coworker Item 1
1476 non-null   float64
 46  Coworker Item 2
1476 non-null   float64
 47  Satisfaction Rank: Communication
1451 non-null   float64
 48  Satisfaction Rank: Compensation
1451 non-null   float64
 49  Satisfaction Rank: Coworkers/Peers
1451 non-null   float64
 50  Satisfaction Rank:  Management
```

```
1451 non-null   float64
 51  Satisfaction Rank: Organizational Fit
1451 non-null   float64
 52  Satisfaction Rank: Career Opportunity
1451 non-null   float64
 53  Satisfaction Rank: Work Environment
1451 non-null   float64
 54  Satisfaction Rank: Clear Job Expectations
1451 non-null   float64
 55  Satisfaction Rank: Other (specify below)
1451 non-null   float64
 56  How likely is it that you would recommend the Chicago Transit Authority to
a friend or colleague?  1263 non-null   float64
 57  Stay Intention: I am actively seeking another job in a different
company/organization.             1498 non-null   object
 58  Stay Intention: I plan on working here for another (in years):
1498 non-null   int64
 59  Stay Factor: Quality of Management
347 non-null    object
 60  Stay Factor: Career Development
524 non-null    object
 61  Stay Factor: Pay/Salary
958 non-null    object
 62  Stay Factor: Coworker Relationships
536 non-null    object
 63  Stay Factor: Grievance Handling
90 non-null     object
 64  Stay Factor: Job Satisfaction
630 non-null    object
 65  Stay Factor: Challenging Work
397 non-null    object
 66  Stay Factor: Rewards & Recognition
199 non-null    object
 67  Stay Factor: Safety
337 non-null    object
 68  Stay Factor: Workload
249 non-null    object
 69  # of Safety Incidents
1498 non-null   int64
 70  # of Absent Days/Tardiness
1498 non-null   int64
dtypes: float64(39), int64(15), object(17)
memory usage: 831.0+ KB
```

[6]:
```
emp_data.columns.ravel()
emp_data = emp_data.replace(' ', np.NaN)
#emp_data = emp_data.mask( emp_data == ' ')
```

```
#emp_data.loc[:,10]
#emp_data.dropna(inplace=True)
```

What is the average CTA tenure (Column F) in months?

[7]:
```
cta_tenure_mean = emp_data['CTA Tenure (Months) '].mean()
print("Average value of CTA Tenure (Months) column is ",cta_tenure_mean)
```

```
Average value of CTA Tenure (Months) column is  127.18638239339752
```

[8]:
```
emp_data_subset = emp_data[emp_data['What is your Position?'].str.
 ↪startswith('Manager')]
emp_data_subset
```

[8]:
```
      Respondent ID What is your gender? What is your race or ethnicity?  \
12               13                    M  White (Not Hispanic or Latino)
15               16                    M  White (Not Hispanic or Latino)
22               23                    F       Black or African American
26               27                    M       Black or African American
30               31                    M       Black or African American
…              …                   …                             …
1375           1377                    M       Black or African American
1408           1410                    M       Black or African American
1484           1486                  NaN       Black or African American
1485           1487                    M       Black or African American
1494           1496                    M              Hispanic or Latino

       Age  What is your yearly CTA salary?  CTA Tenure (Months)   \
12    53.0                        102721.52                   7.0
15    68.0                        102721.52                 165.0
22    53.0                         97454.38                 315.0
26    55.0                        102721.52                 243.0
30    44.0                        102721.52                  89.0
…     …                              …                    …
1375  40.0                        102721.52                 236.0
1408  34.0                        102721.52                  19.0
1484  33.0                        102721.52                  82.0
1485   NaN                        107989.37                   NaN
1494  52.0                        102721.52                   8.0

                  At which location do you work?  \
12                      Howard Terminal (Paulina)
15    567 W Lake Street - Main Location for CTA
22    567 W Lake Street - Main Location for CTA
26                            103rd Street Garage
30                            74th Street  Garage
…                                           …
1375                           311 West Institute
```

```
1408                          311 West Institute
1484                         103rd Street Garage
1485                          Chicago Ave Garage
1494                          311 West Institute

                              What is your Position?  \
12                   Manager, Transportation - Rail
15                     Manager, Facilities Security
22                 Manager, Planning Administration
26                           Manager, Bus Operations
30                           Manager, Bus Operations
…                                                 …
1375               Manager, Rail Station Management
1408               Manager, Rail Station Management
1484                    Manager, Maintenance - Bus
1485                    Manager, Maintenance - Bus
1494  Manager, Administration - Rail Station Management

     Are you a manager or above?  Fit Item 1  …  Stay Factor: Pay/Salary  \
12                           YES           3  …                      NaN
15                           YES           2  …                      NaN
22                           YES           5  …                      NaN
26                           YES           4  …                      NaN
30                           YES           1  …               Pay/Salary
…                              …         …   …                        …
1375                         YES           2  …                      NaN
1408                         YES           5  …               Pay/Salary
1484                         YES           3  …               Pay/Salary
1485                         YES           5  …                      NaN
1494                         YES           5  …                      NaN

      Stay Factor: Coworker Relationships  Stay Factor: Grievance Handling  \
12                                    NaN                              NaN
15                 Coworker Relationships               Grievance Handling
22                 Coworker Relationships                              NaN
26                                    NaN                              NaN
30                 Coworker Relationships                              NaN
…                                       …                                …
1375                                  NaN                              NaN
1408               Coworker Relationships                              NaN
1484               Coworker Relationships                              NaN
1485                                  NaN                              NaN
1494                                  NaN                              NaN

      Stay Factor: Job Satisfaction  Stay Factor: Challenging Work  \
12                              NaN                            NaN
15                 Job Satisfaction               Challenging Work
```

```
22                 Job Satisfaction                               NaN
26                              NaN              Challenging Work
30                              NaN                               NaN
…                                 …                                 …
1375                            NaN                               NaN
1408                            NaN                               NaN
1484               Job Satisfaction                               NaN
1485                            NaN              Challenging Work
1494               Job Satisfaction                               NaN

      Stay Factor: Rewards & Recognition  Stay Factor: Safety  \
12                                   NaN                  NaN
15                                   NaN               Safety
22                                   NaN               Safety
26                                   NaN                  NaN
30                                   NaN                  NaN
…                                    …                  …
1375                                 NaN                  NaN
1408                                 NaN                  NaN
1484               Rewards & Recognition                  NaN
1485                                 NaN                  NaN
1494                                 NaN                  NaN

      Stay Factor: Workload  # of Safety Incidents  # of Absent Days/Tardiness
12                      NaN                      0                           1
15                 Workload                      1                           1
22                      NaN                      0                           1
26                      NaN                      0                           1
30                      NaN                      0                           1
…                     …                    …                          …
1375                    NaN                      0                           0
1408                    NaN                      0                           0
1484               Workload                      0                           2
1485                    NaN                      0                           2
1494                    NaN                      0                           1

[61 rows x 71 columns]
```

What is the mode of CTA tenure (Column F) for only managers in months?

[9]:
```
cta_tenure_mgr_mode = emp_data_subset['CTA Tenure (Months) '].mode()
print("Mode of CTA Tenure (Months) for Manager Category is␣
 ↪",cta_tenure_mgr_mode.values)
```

```
Mode of CTA Tenure (Months) for Manager Category is  [10.]
```

What is the median salary (Column E) of all respondents?

```
[10]: salary_median = emp_data['What is your yearly CTA salary?'].median()
print("Median of all employees salary is ",salary_median)
```

```
Median of all employees salary is  80419.04
```

What is the standard deviation of stay intention (Column BG) for all respondents?

```
[11]: stddev_stayintention = emp_data['Stay Intention: I plan on working here for␣
 ↪another (in years):'].std()
print("The number of years an employee decides to stay in the organisation␣
 ↪range within ",stddev_stayintention)
```

```
The number of years an employee decides to stay in the organisation range within
5.0557513696165355
```

What is the range of Age (Column D) in years?

```
[12]: print("Age of employees ranges between ",emp_data['Age'].min(), "and␣
 ↪",emp_data['Age'].max())
```

```
Age of employees ranges between  22.0 and  69.0
```

What is the Net Promoter Score (Column BE) for all survey respondents? (See Part 2 of Instructions)

In this case study, the promoters will be the employees who opt to stay in the organisation and detractors are those who plan to leave the organisation.

- promoters - score equal to or greater than 9
- passives - score equal to 7 or 8
- detractors - score less than or equal to 6

```
[13]: #Since Respondent ID is the unique key, retrieving count using that column.
total_emp = emp_data['Respondent ID'].count()
emp_data.rename(columns = {
'How likely is it that you would recommend the Chicago Transit Authority to a␣
 ↪friend or colleague?' : 'org_recommendation'}
                , inplace = True)

empdataPromo = emp_data.loc[emp_data['org_recommendation'] >= 9]
promoters = empdataPromo['Respondent ID'].count()

empdataPassive = emp_data.loc[emp_data['org_recommendation'].isin([7,8])]
passives = empdataPassive['Respondent ID'].count()

empdatadetra = emp_data.loc[emp_data['org_recommendation'] <= 6]
detractors = empdatadetra['Respondent ID'].count()

nps = ((promoters - detractors)/total_emp) * 100
print("The Net Promoter Score is ",int(nps))
```

```
The Net Promoter Score is  22
```

Run a bivariate correlation analysis of survey items (Columns J:AU) and tenure (Column F). What three variables have the greatest correlation (absolute magnitude) with tenure? Only consider statistically significant correlations using a .05 alpha probability threshold. Report the names of the variables, the r values, and p values.

[14]:
```
emp_data.rename(columns = {
'Engagement Item 1' : 'engagement_item_1',
'Engagement Item 2' : 'engagement_item_2',
'Engagement Item 3' : 'engagement_item_3',
'Engagement Item 4' : 'engagement_item_4',
'Engagement Item 5' : 'engagement_item_5',
'Fit Item 1' : 'fit_item_1',
'Fit Item 2' : 'fit_item_2',
'Fit Item 3' : 'fit_item_3',
'# of Absent Days/Tardiness' : 'tardiness'}, inplace = True)
emp_data.rename(columns = {
'Commitment Item 1' : 'commitment_item_1',
'Commitment Item 2' : 'commitment_item_2',
'Commitment Item 3' : 'commitment_item_3',
'Commitment Item 4' : 'commitment_item_4',
'Commitment Item 5' : 'commitment_item_5'}, inplace = True)
emp_data.rename(columns = {'What is your yearly CTA salary?':'cta_salary',
                            'Engagement Item 1':'engagement_item_1',
                           'Stay Intention: I plan on working here for another␣
 ↪(in years):' : 'stay_intention'}, inplace = True)

corr_cols = ['CTA Tenure (Months) ','fit_item_1', 'fit_item_2',
       'fit_item_3', 'HiPo Item 1', 'HiPo Item 2', 'HiPo Item 3',
       'HiPo Item 4', 'Satisfaction Item 1 ', 'Satisfaction Item 2',
       'engagement_item_1', 'engagement_item_2', 'engagement_item_3',
       'engagement_item_4', 'engagement_item_5', 'Engagement Item 6',
       'Motivation Item 1', 'Motivation Item 2', 'Motivation Item 3',
       'Motivation Item 4', 'Performance Item 1', 'Performance Item 2',
       'Leadership Item 1', 'Leadership Item 2', 'Leadership Item 3',
       'Support Item 1', 'Support Item 2', 'commitment_item_1',
       'commitment_item_2', 'commitment_item_3', 'commitment_item_4',
       'commitment_item_5', 'Diversity Item 1', 'Diversity Item 2',
       'Diversity Item 3 (Reverse Coded)', 'Trust Item 1', 'Trust Item 2',
       'Coworker Item 1', 'Coworker Item 2']
```

[15]:
```
iter_cols = ['fit_item_1', 'fit_item_2',
       'fit_item_3', 'HiPo Item 1', 'HiPo Item 2', 'HiPo Item 3',
       'HiPo Item 4', 'Satisfaction Item 1 ', 'Satisfaction Item 2',
       'engagement_item_1', 'engagement_item_2', 'engagement_item_3',
       'engagement_item_4', 'engagement_item_5', 'Engagement Item 6',
       'Motivation Item 1', 'Motivation Item 2', 'Motivation Item 3',
       'Motivation Item 4', 'Performance Item 1', 'Performance Item 2',
```

```python
        'Leadership Item 1', 'Leadership Item 2', 'Leadership Item 3',
        'Support Item 1', 'Support Item 2', 'commitment_item_1',
        'commitment_item_2', 'commitment_item_3', 'commitment_item_4',
        'commitment_item_5', 'Diversity Item 1', 'Diversity Item 2',
        'Diversity Item 3 (Reverse Coded)', 'Trust Item 1', 'Trust Item 2',
        'Coworker Item 1', 'Coworker Item 2']
pair_column = ""
lcorr_vals = []
corr_df = pd.DataFrame(emp_data['CTA Tenure (Months) '])
for pair_column in iter_cols:
    corr_df = emp_data[['CTA Tenure (Months) ',pair_column]]
    corr_df.dropna(inplace=True)

    r_value,p_value = pearsonr(x=corr_df.iloc[:,0],y=corr_df.iloc[:,1])

    final_result = (pair_column,r_value,p_value)
    lcorr_vals.append(final_result)
    #corr_df.drop(columns='coorelation_column',axis=1)
lcorr_vals
```

[15]:
```
[('fit_item_1', -0.004810819555188566, 0.8545733471053413),
 ('fit_item_2', 0.02573827413280609, 0.3267142717706238),
 ('fit_item_3', -0.004713836688322202, 0.8574733755740244),
 ('HiPo Item 1', -0.6038591129549808, 3.38034147980469e-145),
 ('HiPo Item 2', 0.49763399270962616, 8.096499271065116e-92),
 ('HiPo Item 3', 0.0186386279878786, 0.47760087471159307),
 ('HiPo Item 4', -0.017979010992573444, 0.49332514731826893),
 ('Satisfaction Item 1 ', 0.720491385587094, 4.261936467570659e-233),
 ('Satisfaction Item 2', 0.008345821522639369, 0.7505913262621816),
 ('engagement_item_1', 0.01926239855822254, 0.46298671425081156),
 ('engagement_item_2', 0.026192087548919843, 0.3182541581930929),
 ('engagement_item_3', 0.010370087706281235, 0.6927720362936609),
 ('engagement_item_4', 0.01432682402658285, 0.585162627385992),
 ('engagement_item_5', -0.03613437247414592, 0.1684760192107389),
 ('Engagement Item 6', 0.011362591131266954, 0.6650784791976768),
 ('Motivation Item 1', -0.027042169180077957, 0.30429972437609504),
 ('Motivation Item 2', -0.004180798438448685, 0.8738355078350756),
 ('Motivation Item 3', 0.007334380622254307, 0.7805783508188276),
 ('Motivation Item 4', 0.009472264560702222, 0.7190212833383927),
 ('Performance Item 1', 0.015522754848669199, 0.5554630272645065),
 ('Performance Item 2', -0.0054346718045072155, 0.8364692793246219),
 ('Leadership Item 1', -0.005036947542931124, 0.848702216778762),
 ('Leadership Item 2', 0.0058057010532763164, 0.8259555865650631),
 ('Leadership Item 3', -0.014470350083268375, 0.5836299446447538),
 ('Support Item 1', 0.03233798915036808, 0.22053185213771895),
 ('Support Item 2', 0.016546444141077452, 0.5308322482768332),
 ('commitment_item_1', 0.03189622819002545, 0.22690476283675493),
```

```
 ('commitment_item_2', 0.02853066285341971, 0.2797802985320398),
 ('commitment_item_3', -0.0037989550298951433, 0.8855911530165483),
 ('commitment_item_4', -0.03180566383952124, 0.2282273029839716),
 ('commitment_item_5', 0.02114557471662422, 0.4231467802350828),
 ('Diversity Item 1', -0.014018701951172212, 0.5959450327704577),
 ('Diversity Item 2', -0.018573993607492303, 0.4823266888622963),
 ('Diversity Item 3 (Reverse Coded)',
  0.0030654269790177154,
  0.9076995346236424),
 ('Trust Item 1', -0.0340372369953979, 0.19784070631705414),
 ('Trust Item 2', -0.012141424575296671, 0.6460695055885636),
 ('Coworker Item 1', -0.03851936801465056, 0.14500087152275065),
 ('Coworker Item 2', -0.00808618489173956, 0.7597262424993099)]
```

[16]:
```python
filtered_lcorr_vals = [(a, b, c) for (a,b,c) in lcorr_vals if c<=0.05]
filtered_lcorr_vals
```

[16]:
```
[('HiPo Item 1', -0.6038591129549808, 3.38034147980469e-145),
 ('HiPo Item 2', 0.49763399270962616, 8.096499271065116e-92),
 ('Satisfaction Item 1 ', 0.720491385587094, 4.261936467570659e-233)]
```

With the above table, we could come to a conclusion that the top 3 columns with strong correlation with "CTA Tenure (Months)" is

1. Satisfaction Item 1
2. HiPo Item 1
3. HiPo Item 2

Employees those who feel highly positive about the workplace (Satisfaction Item 1) and feel that they are growing professionally while contributing to organisation's growth (HiPo Item 1 & HiPo Item 2) tend to associate longer (CTA Tenure) with the organisation

Using a t-test analysis is there a statistically significant difference (using a .05 alpha threshold) in satisfaction item 1 (Column Q) between managers and non-managers?

#In order to perform the t-test analysis against "satisfaction item 1" between managers and non managers, we will make use of the data in "Are you a manager or above?" column.

[17]:
```python
#First split the employees as manager and non-manager categories
mgr_ds=emp_data[emp_data['What is your Position?'].str.
 ↪startswith('Manager')]['Satisfaction Item 1 ']
mgr_ds.count()

non_mgr_ds=emp_data[~emp_data['What is your Position?'].str.
 ↪startswith('Manager')]['Satisfaction Item 1 ']
non_mgr_ds.count()

print("manager :-", mgr_ds.count(), "non-manager :-", non_mgr_ds.count())
```

manager :- 58 non-manager :- 1396

Since the number of values in each dataset differs, we would go with ttest_ind method.

[18]:
```
from scipy.stats import ttest_ind

ttest_ind(mgr_ds,non_mgr_ds,equal_var=False, nan_policy='omit')
```

[18]:
```
TtestResult(statistic=0.6595369257621557, pvalue=0.5120176812629944,
df=61.39136252566917)
```

Since the p-value is greater than 0.05, we can conclude that impact of role on Satisfactory item 1 is not statistically significant.

Run an OLS regression analysis with Salary (Column E) and Engagement item 1 (Column S) as your predictors and stay intention (Column BG) as your predicted outcome. Report the magnitude (i.e., unstandardized betas) and p value of both predictors.

'What is your yearly CTA salary?','Engagement Item 1','Stay Intention: I plan on working here for another (in years):'

[19]:
```
#independent variables - What is your yearly CTA salary?, Engagement Item 1
#Dependent variable - Stay Intention: I plan on working here for another (in␣
↪years):
print(emp_data['cta_salary'].isna().sum(),
emp_data['engagement_item_1'].isna().sum(),
emp_data['stay_intention'].isna().sum())
```

```
0 0 0
```

[20]:
```
#split data for train and test
x_cols = ['cta_salary','engagement_item_1']
y_cols = ['stay_intention']
x = emp_data[x_cols]
y = emp_data[y_cols]
```

[21]:
```
X_train, X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.30,␣
↪random_state=2)
```

[22]:
```
emp_data['stay_intention'].describe()
```

[22]:
```
count    1498.000000
mean        7.114820
std         5.055751
min         0.000000
25%         3.000000
50%         7.000000
75%         9.000000
max        23.000000
Name: stay_intention, dtype: float64
```

With the above data, we can could see that the mean is more or less same as median. So that the data distribution in the dependent value is perfectly symmetrical

[23]:
```
cor_col =  ['cta_salary','engagement_item_1','stay_intention']
emp_data[cor_col].corr(method='pearson',numeric_only=True)
```

[23]:

	cta_salary	engagement_item_1	stay_intention
cta_salary	1.000000	0.009631	0.443546
engagement_item_1	0.009631	1.000000	0.018384
stay_intention	0.443546	0.018384	1.000000

We see that salary has a correlation with stay_intention but engagement_item_1 does not.

[24]:
```
import statsmodels.api as sm
# not scaling or standardizing because we require unstandardized beta values
#X_train['cta_salary'] = preprocessing.scale(X_train.cta_salary.values)
model = sm.OLS.from_formula(formula='stay_intention ~ cta_salary +␣
 ↪engagement_item_1', data=emp_data)
result = model.fit()
```

[25]:
```
result.summary()
```

[25]:

Dep. Variable:	stay_intention	**R-squared:**	0.197
Model:	OLS	**Adj. R-squared:**	0.196
Method:	Least Squares	**F-statistic:**	183.3
Date:	Sat, 09 Sep 2023	**Prob (F-statistic):**	6.35e-72
Time:	16:38:09	**Log-Likelihood:**	-4388.4
No. Observations:	1498	**AIC:**	8783.
Df Residuals:	1495	**BIC:**	8799.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-3.1513	0.690	-4.569	0.000	-4.504	-1.798
cta_salary	0.0001	5.82e-06	19.131	0.000	9.99e-05	0.000
engagement_item_1	0.0682	0.112	0.609	0.543	-0.151	0.288

Omnibus:	108.092	**Durbin-Watson:**	1.977
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	520.938
Skew:	0.078	**Prob(JB):**	7.58e-114
Kurtosis:	5.885	**Cond. No.**	5.45e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.45e+05. This might indicate that there are strong multicollinearity or other numerical problems.

The OLS results are similar to correlation valies.

The p-value [0.00] of cta_salary specifies that it there exist a relation between the Salary and Stay Intention

The p-value [0.543] of engagement_item_1 clearly shows that it is statistically insignificant.

Run a negative binomial regression analysis with all Engagement items (column S:W) and Fit items (column J:L) as your predictors and Tardiness (Column BT) as your predicted outcome. Report the item names, odd ratios, and p value of only the statistically significant predictors (use a 0.05 alpha probability threshold).

[26]:
```
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

[27]:
```
x_cols_nb = ['engagement_item_1','engagement_item_2','engagement_item_3',
            ␣
 ↪'engagement_item_4','engagement_item_5','fit_item_1','fit_item_2','fit_item_3']
y_cols_nb = ['tardiness']
x = emp_data[x_cols_nb]
y = emp_data[y_cols_nb]
```

[28]:
```
from patsy.highlevel import dmatrices
nb_formula = 'tardiness ~ engagement_item_1  + engagement_item_2 +␣
 ↪engagement_item_3 + engagement_item_4 + engagement_item_5 + fit_item_1 +␣
 ↪fit_item_2 + fit_item_3'
```

[29]:
```
#Setup the X and y matrices for the training and testing data sets
Y_train_nb, X_train_nb = dmatrices(formula_like=nb_formula, data=emp_data,␣
 ↪return_type='dataframe')
Y_test_nb, X_test_nb = dmatrices(formula_like=nb_formula, data=emp_data,␣
 ↪return_type='dataframe')
```

[30]:
```
nb2_training_results = sm.GLM(Y_train_nb, X_train_nb,family=sm.families.
 ↪NegativeBinomial()).fit()
```

```
C:\ProgramData\anaconda3\lib\site-
packages\statsmodels\genmod\families\family.py:1367: ValueWarning: Negative
binomial dispersion parameter alpha not set. Using default value alpha=1.0.
  warnings.warn("Negative binomial dispersion parameter alpha not "
```

[31]:
```
print(nb2_training_results.summary(alpha=0.05))
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:             tardiness   No. Observations:                 1498
Model:                           GLM   Df Residuals:                     1489
Model Family:       NegativeBinomial   Df Model:                            8
Link Function:                   Log   Scale:                          1.0000
Method:                         IRLS   Log-Likelihood:                -945.99
Date:               Sat, 09 Sep 2023   Deviance:                       982.28
Time:                       16:38:09   Pearson chi2:                 2.16e+03
No. Iterations:                    6   Pseudo R-squ. (CS):            0.01244
Covariance Type:           nonrobust
```

```
==============================================================================
=====
                          coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-----
Intercept              -1.2008      0.395     -3.039      0.002      -1.975
-0.426
engagement_item_1       0.0770      0.062      1.232      0.218      -0.045
0.199
engagement_item_2       0.1178      0.063      1.868      0.062      -0.006
0.241
engagement_item_3       0.0194      0.061      0.318      0.750      -0.100
0.139
engagement_item_4      -0.1433      0.062     -2.297      0.022      -0.266
-0.021
engagement_item_5      -0.1361      0.053     -2.551      0.011      -0.241
-0.032
fit_item_1             -0.0570      0.044     -1.302      0.193      -0.143
0.029
fit_item_2              0.0188      0.043      0.439      0.661      -0.065
0.103
fit_item_3              0.0490      0.043      1.129      0.259      -0.036
0.134
==============================================================================
=====
```

[32]:
```
odds_ratio = np.exp((nb2_training_results.params))
odds_ratio
```

[32]:
```
Intercept            0.300938
engagement_item_1    1.080036
engagement_item_2    1.124976
engagement_item_3    1.019571
engagement_item_4    0.866518
engagement_item_5    0.872723
fit_item_1           0.944591
fit_item_2           1.018979
fit_item_3           1.050199
dtype: float64
```

Based on the p-value(0.05 alpha probability threshold), the statistically significant predictors are engagement_item_2, engagement_item_ 4 and engagement_item_5

Run an Exploratory Factor Analysis (EFA) on Commitment (columns AJ:AN). Run your analysis with one (1) fixed factor (i.e., do not use eigenvalues to determine the number of factors) using principal axis extraction and varimax rotation. Which items would you consider dropping (i.e., dimension reduction) from an aggregated measure of commitment? Report the item names

[33]:
```python
from factor_analyzer import FactorAnalyzer
from sklearn.preprocessing import StandardScaler
```

[34]:
```python
#emp_data['commitment_item_1'].fillna(emp_data['commitment_item_1'].
 ↪mode()[0],inplace=True)
#emp_data['commitment_item_2'].fillna(emp_data['commitment_item_2'].
 ↪mode()[0],inplace=True)
#emp_data['commitment_item_3'].fillna(emp_data['commitment_item_3'].
 ↪mode()[0],inplace=True)
#emp_data['commitment_item_4'].fillna(emp_data['commitment_item_4'].
 ↪mode()[0],inplace=True)
#emp_data['commitment_item_5'].fillna(emp_data['commitment_item_5'].
 ↪mode()[0],inplace=True)

efa_cols =␣
 ↪['commitment_item_1','commitment_item_2','commitment_item_3','commitment_item_4','commitmen
efa_df = pd.DataFrame()
efa_df = emp_data[efa_cols]
efa_df = efa_df.dropna()
scaler = StandardScaler()
efa_df=pd.DataFrame(scaler.fit_transform(efa_df), columns=efa_df.columns)

efa_df.shape
```

[34]: (1481, 5)

Barlett's Test of Sphericity and Kaiser-Meyer-Olkin Test

[35]:
```python
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
chi_square_value,p_value=calculate_bartlett_sphericity(efa_df)
print("chi-Square Value :" ,chi_square_value,"p-value :", p_value)
alpha = 0.05
if p_value < alpha:
    print("Good to go with Factor Analysis")
else:
    print("Factor analysis not recommended")
```

```
chi-Square Value : 98.45389237661077 p-value : 1.1107195390672575e-16
Good to go with Factor Analysis
```

[36]:
```python
#Kaiser-Meyer-Olkin (KMO) Test
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(efa_df)
print("KMO score :",kmo_model)
if kmo_model > 0.6:
    print("Data suitable for Factor analysis")
```

```
else:
    print("Factor analysis not recommended")
```

```
KMO score : 0.5951983233324004
Factor analysis not recommended
```

the KMO test (Kaiser-Meyer-Olkin) should test whether it is appropriate to use the manifest variables for factor analysis. The test involves the computation of the proportion of variance among the manifest variables. The KMO values range between 0-1 and a proportion under 0.6 would suggest that the dataset is inappropriate for factor analysis.

As KMO score is below 0.6, then factor analysis is not recomended

[37]:
```
# We actually wanted to implement 'Varimax' rotation, but since the we opted␣
↪for only one factor, no rotation will be performed.
# so not mentioning input for rotation parameter
fa = FactorAnalyzer(n_factors=1,rotation='varimax',method='principal')
fa.fit(efa_df)
```

[37]:
```
FactorAnalyzer(method='principal', n_factors=1, rotation='varimax',
               rotation_kwargs={})
```

[38]:
```
#get loadings values
pd.DataFrame(fa.loadings_,columns=['Factor1'],index=efa_df.columns)
```

[38]:

	Factor1
commitment_item_1	-0.463277
commitment_item_2	-0.408670
commitment_item_3	-0.567387
commitment_item_4	-0.557155
commitment_item_5	-0.568985

Looking at the factor loadings values, Commitment Item 1 and 2 have weak relationship with the factor. Thus, I will drop Commitment Item 1 and Commitment Item 2 from the aggregated measure of commitment.