Charter School Budget Analysis EDA

October 5, 2023

1 Budget Analysis EDA

Exploratory data analysis of budget data of a small charter school district.

[4]:
```
# importing packages

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import datetime as dt
```

[122]:
```
# importing data
df = pd.read_csv(r"/Users/scipio/Downloads/Charter_School_Budget_Data.csv")
df.head()
```

[122]:

	Name	Position	Department	Campus	Finance_Budget_Salary	Actual_Budget_Salary	Budgeted	Employment_Start_Date
0	Filled	Safety Specialist	Custodial	HO	47000.0	79577.0	Budgeted	NaN
1	Filled	Safety Specialist	Custodial	HO	47000.0	79577.0	Budgeted	NaN
2	Filled	Safety Specialist	Custodial	HO	47000.0	79577.0	Budgeted	NaN
3	VACANT	Chief Advancement Officer	School Admin	HO	172500.0	0.0	Budgeted	NaN
4	ELIMINATED	School Business Administrator (Part-Time)	Business Office	HO	51741.0	0.0	Budgeted	NaN

[6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 192 entries, 0 to 191
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Name                    192 non-null    object
 1   Position                192 non-null    object
 2   Department              192 non-null    object
 3   Campus                  192 non-null    object
 4   Finance_Budget_Salary   192 non-null    float64
 5   Actual_Budget_Salary    192 non-null    float64
 6   Budgeted                192 non-null    object
 7   Employement_Start_Date  152 non-null    object
dtypes: float64(2), object(6)
memory usage: 12.1+ KB
```

[7]:
```
#Shape
df.shape
```

[7]:
```
(192, 8)
```

There is 192 rows and 8 columns in the dataset.

[8]:
```
#Null values
df.isnull().sum()
```

[8]:
```
Name                     0
Position                 0
Department               0
Campus                   0
Finance_Budget_Salary    0
Actual_Budget_Salary     0
Budgeted                 0
Employement_Start_Date   40
dtype: int64
```

There are 40 null values in the Employment Start Date column.

[9]:
```
# Vacant positions
vacant_filter = ['VACANT','SUB - VACANT']

df[df['Name'].isin(vacant_filter)].groupby('Name')['Name'].count()
```

[9]:
```
Name
SUB - VACANT     1
VACANT          26
Name: Name, dtype: int64
```

There are 27 vacancies.

[10]:
```
df[df['Name']== 'ELIMINATED'].groupby('Name')['Name'].count()
```

[10]:
```
Name
ELIMINATED    6
Name: Name, dtype: int64
```

6 positions were eliminated.

[11]:
```
# Filled Positions

filter_values = ['VACANT','ELIMINATED','SUB-VACANT']

filled_positions = df[~df['Name'].isin(filter_values)]

filled_positions.shape[0]
```

[11]:
```
160
```

There are 160 filled positions

[12]:
```
df['Department'].nunique()
```

[12]:
```
14
```

There are 14 departments in the dataset:

1. Custodial
2. School Admin
3. Business Office
4. School Admin - Clerical
5. Improvement of Instruction Services
6. High School (HS)
7. Kindergarten to Eight Grade (K-8)
8. Information Technology (IT)
9. Student Success Team (SST)
10. Transportation
11. Clerical Business
12. Other Prof. Staff -CST
13. Other - Instructional
14. Salaries (Support) - Nurse

[13]:
```
df['Campus'].nunique()
```

[13]:
```
5
```

There are 5 Campuses in the dataset:

1. Primary School (PS)
2. Intermediate School (IS)
3. Middle School (MS)
4. High School (HS)

5. Home Office (HO)

[14]:
```
df['Budgeted'].nunique()
```

[14]:
```
2
```

There are two budget categories,

1. Budgeted: The staff member's salary was accounted for in the budget.
2. Unbudgeted: The staff member's salary was not accounted for in the budget.

[15]:
```
# Unbudgeted Hires Count
df[df['Budgeted'] == 'Unbudgeted'].groupby('Budgeted')['Name'].count()
```

[15]:
```
Budgeted
Unbudgeted    11
Name: Name, dtype: int64
```

There were 11 unbudgeted hires.

[16]:
```
#Unbudgeted Salaries
unbudgeted = df[df['Budgeted'] == 'Unbudgeted']

round(unbudgeted['Actual_Budget_Salary'].sum(),2)
```

[16]:
```
878822.56
```

The unbudgeted hires salaries totaled $878,822.56

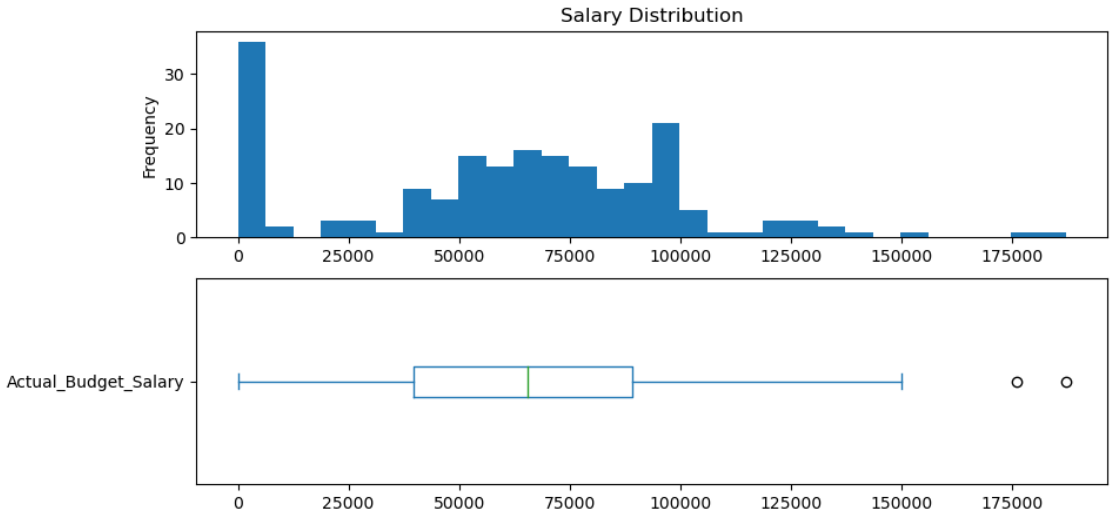
[17]:
```
round(df['Actual_Budget_Salary'].describe(),2)
```

[17]:
```
count       192.00
mean      60931.50
std       39072.17
min           0.00
25%       39799.44
50%       65385.46
75%       89251.73
max      187250.00
Name: Actual_Budget_Salary, dtype: float64
```

[120]:
```
# Creating a subplot
fig, axs = plt.subplots(nrows = 2, ncols = 1, figsize = (10,5))

df['Actual_Budget_Salary'].plot(kind = 'hist', bins = 30, title = 'Salary␣
↪Distribution', ax = axs[0])
df['Actual_Budget_Salary'].plot(kind = 'box', vert = False, ax = axs[1])
```

[120]:
```
<Axes: >
```



There is a multimodal right skew in the distribution of the salary data as indicated in the histogram. The Boxplot also indiactes that there are some salary values that are outliers in the dataset.

[124]:
```
#Calculating employment length
employment_length =  df[df['Employment_Start_Date'].notnull()]

employment_length['Employment_Start_Date'] = pd.
 ↪to_datetime(employment_length['Employment_Start_Date'])
```

```
/var/folders/3k/bzmghyyj1j51lkx1mc36njjw0000gn/T/ipykernel_63217/964332287.py:4:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  employment_length['Employment_Start_Date'] =
pd.to_datetime(employment_length['Employment_Start_Date'])
```

[105]:
```
#Creating datetime object
date1 = dt.datetime(2023,9,22)

#inserting datetime object into dataframe
employment_length['Current'] = date1

# Creating a column named Employment Length
employment_length['Employment_Length_Years'] = employment_length['Current'].dt.
 ↪year - employment_length['Employement_Start_Date'].dt.year
```

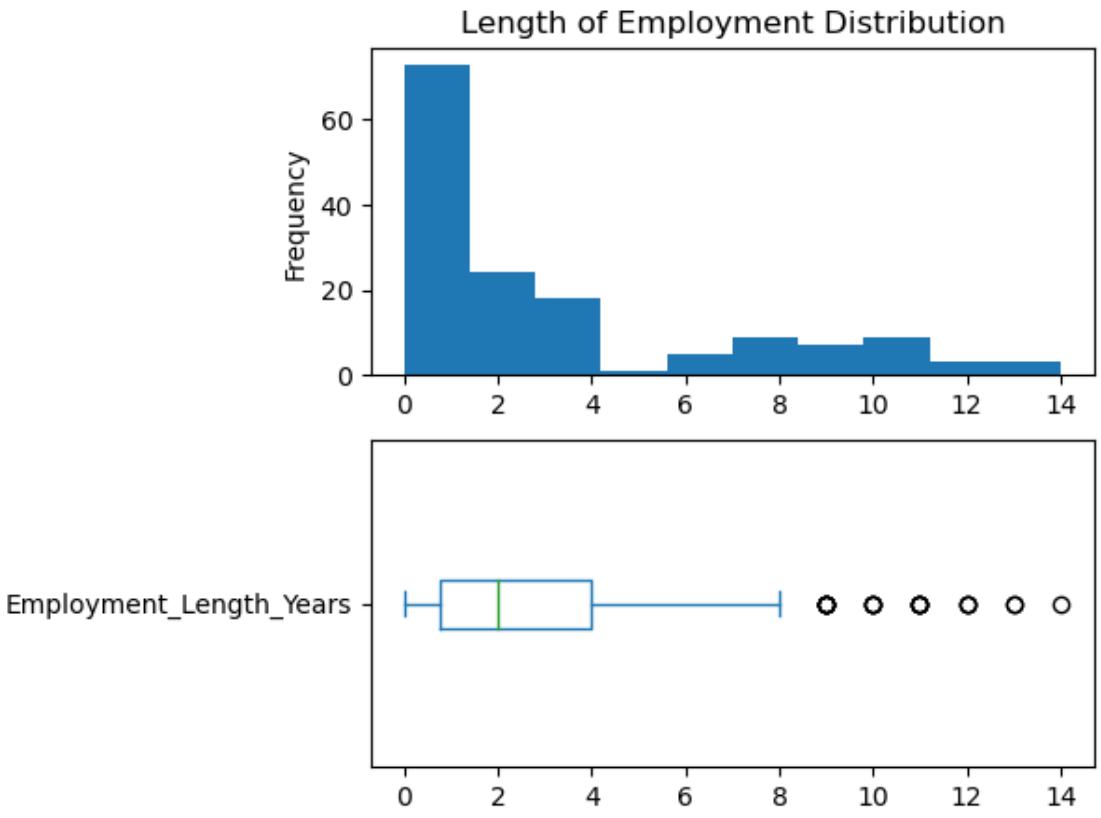
[109]:
```
round(employment_length['Employment_Length_Years'].describe(),2)
```

[109]:
```
count    152.00
mean       3.16
std        3.70
min        0.00
25%        0.75
50%        2.00
75%        4.00
max       14.00
Name: Employment_Length_Years, dtype: float64
```

[117]:
```
# Creating subplots
fig,axs = plt.subplots(nrows = 2, ncols = 1, figsize = (5,5))

employment_length['Employment_Length_Years'].plot(kind = 'hist', title =␣
 ↪'Length of Employment Distribution', ax = axs[0])
employment_length['Employment_Length_Years'].plot(kind = 'box', vert = False,␣
 ↪ax = axs[1])
```

[117]:
```
<Axes: >
```



There is a right skew in the distribution of length of employment as indicated in the histogram.

This is indicative in the difference between the mean and median in which employees that have been employed with the organization have inflated the mean. Additionally, employees that have been employed at the organization for more than 8 years are outliers as indicated by the boxplot above.