lpt6uwazy

February 24, 2025

[1]:
```
import pandas as pd
import numpy as np
```

[3]:
```
df = pd.read_csv('salary.csv')
```

[5]:
```
df
```

[5]:
```
    Age  Gender Education Level                 Job Title  \
0    43   Other             PhD              Data Analyst
1    23  Female             PhD           Biotechnologist
2    25  Female     High School        Research Scientist
3    32   Other        Master's        Research Scientist
4    41    Male      Bachelor's              Data Analyst
..  …     …              …                          …
95   28   Other     High School            Lab Technician
96   30  Female             PhD        Research Scientist
97   45  Female        Master's  Quality Control Specialist
98   31    Male        Master's        Research Scientist
99   54    Male             PhD              Data Analyst

    Years of Experience  Salary
0                     2   78913
1                     9  110403
2                     7   39666
3                     3   91913
4                    26   40868
..                  …      …
95                   19  125371
96                   24  149805
97                    9  146587
98                   11   49128
99                    2   68487

[100 rows x 6 columns]
```

[13]:
```
#mean
mean_values = df.mean(numeric_only=True)
```

```
print(mean_values)
```

```
Age                          39.75
Years of Experience          14.97
Salary                    93616.05
dtype: float64
```

[15]:
```
df.loc[:,'Age'].mean()
```

[15]:
```
39.75
```

[21]:
```
#Median
median_value = df.median(numeric_only=True)
print(median_value)
```

```
Age                           40.5
Years of Experience           15.0
Salary                    100757.0
dtype: float64
```

[23]:
```
df.loc[:,'Age'].median()
```

[23]:
```
40.5
```

[29]:
```
#mode
mode_value = df.mode(numeric_only=True).iloc[0]
print(mode_value)
```

```
Age                          43.0
Years of Experience           9.0
Salary                    31887.0
Name: 0, dtype: float64
```

[33]:
```
df.loc[:,'Age'].mode()
```

[33]:
```
0    43
Name: Age, dtype: int64
```

[35]:
```
#Minimum
df.min()
```

[35]:
```
Age                                 22
Gender                          Female
Education Level             Bachelor's
Job Title              Biotechnologist
Years of Experience                  0
Salary                           31887
dtype: object
```

[37]:
```
df.loc[:,'Age'].min(skipna = False)
```

[37]:
```
22
```

[39]:
```
#maximum
df.max()
```

[39]:
```
Age                                     60
Gender                               Other
Education Level                        PhD
Job Title              Research Scientist
Years of Experience                     30
Salary                              149963
dtype: object
```

[41]:
```
df.loc[:,'Age'].max(skipna = False)
```

[41]:
```
60
```

[45]:
```
# Standard Deviation
std_values = df.std(numeric_only=True)
print(std_values)
```

```
Age                       11.428632
Years of Experience        8.678843
Salary                 35796.187433
dtype: float64
```

[47]:
```
df.loc[:,'Age'].std()
```

[47]:
```
11.428632305032671
```

[51]:
```
#Categorical Variable: Genre
#Quantitative Variable : Age
df.groupby(['Gender'])['Age'].mean()
```

[51]:
```
Gender
Female    36.296296
Male      41.941176
Other     40.230769
Name: Age, dtype: float64
```

[53]:
```
from sklearn import preprocessing
enc = preprocessing.OneHotEncoder()
enc_df = pd.DataFrame(enc.fit_transform(df[['Gender']]).toarray())
enc_df
```

[53]:
```
      0    1    2
0   0.0  0.0  1.0
1   1.0  0.0  0.0
2   1.0  0.0  0.0
3   0.0  0.0  1.0
4   0.0  1.0  0.0
..  ...  ...  ...
95  0.0  0.0  1.0
96  1.0  0.0  0.0
97  1.0  0.0  0.0
98  0.0  1.0  0.0
99  0.0  1.0  0.0

[100 rows x 3 columns]
```

[55]:
```
df_u = df.rename(columns={'Salary)': 'Income'}, inplace=False) # Fix␣
 ↪the␣parenthesis
print(df_u.groupby('Gender')['Salary'].mean()) # Fix the grouping and indexing
```

```
Gender
Female    98657.666667
Male      91067.970588
Other     92347.102564
Name: Salary, dtype: float64
```

[57]:
```
df_encode = df_u.join(enc_df)
print(df_encode) # Use the correct variable name
```

```
    Age  Gender Education Level                     Job Title  \
0    43   Other             PhD                  Data Analyst
1    23  Female             PhD              Biotechnologist
2    25  Female     High School            Research Scientist
3    32   Other        Master's            Research Scientist
4    41    Male      Bachelor's                  Data Analyst
..  ...     ...             ...                           ...
95   28   Other     High School                Lab Technician
96   30  Female             PhD            Research Scientist
97   45  Female        Master's   Quality Control Specialist
98   31    Male        Master's            Research Scientist
99   54    Male             PhD                  Data Analyst

    Years of Experience  Salary    0    1    2
0                     2   78913  0.0  0.0  1.0
1                     9  110403  1.0  0.0  0.0
2                     7   39666  1.0  0.0  0.0
3                     3   91913  0.0  0.0  1.0
4                    26   40868  0.0  1.0  0.0
..                  ...     ...  ...  ...  ...
```

```
95                     19  125371  0.0  0.0  1.0
96                     24  149805  1.0  0.0  0.0
97                      9  146587  1.0  0.0  0.0
98                     11   49128  0.0  1.0  0.0
99                      2   68487  0.0  1.0  0.0

[100 rows x 9 columns]
```

[61]:

```
import pandas as pd
# Calculate skewness for numerical columns
skewness = df_encode.select_dtypes(include=['number']).skew()
print("Skewness of numerical columns:")
print(skewness)
```

```
Skewness of numerical columns:
Age                       0.084161
Years of Experience       0.101069
Salary                   -0.190685
0                         1.051977
1                         0.685851
2                         0.457949
dtype: float64
```

[63]:

```
import numpy as np
from scipy import stats
```

[65]:

```
z = np.abs(stats.zscore(df['Salary']))
```

[67]:

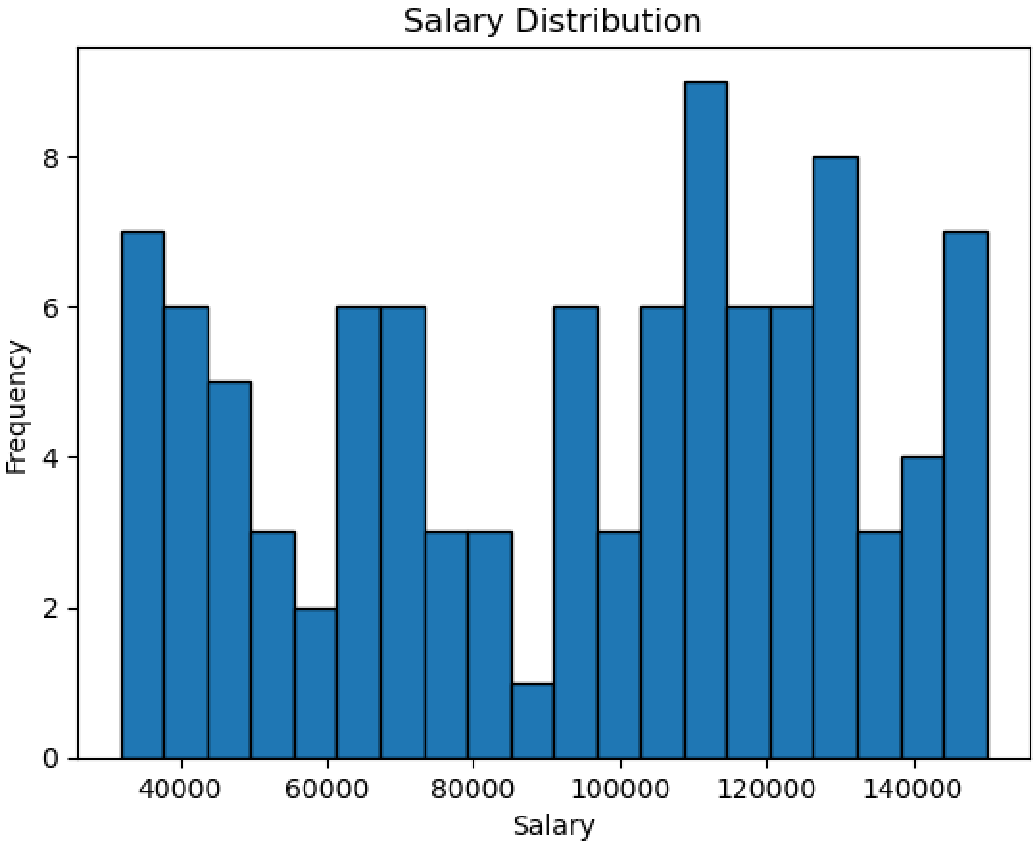
```
z
```

[67]:

```
0      0.412813
1      0.471322
2      1.514738
3      0.047816
4      1.480990
         …
95     0.891573
96     1.577599
97     1.487248
98     1.249076
99     0.705540
Name: Salary, Length: 100, dtype: float64
```

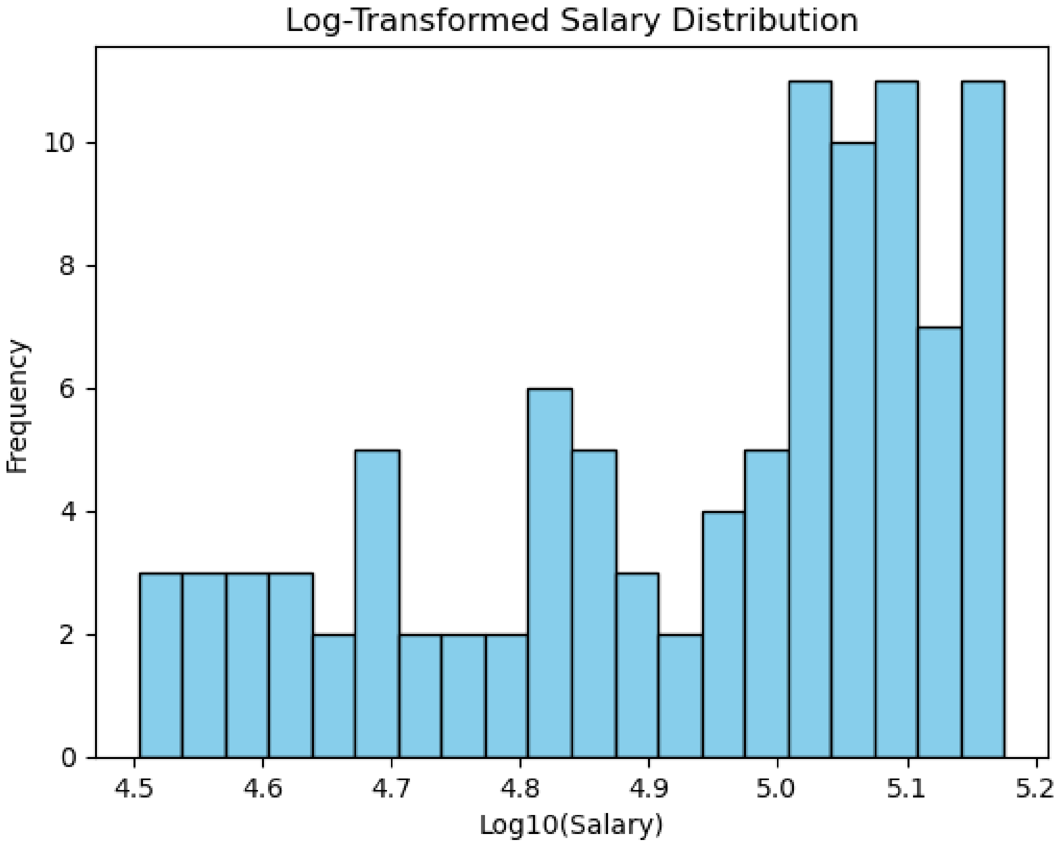
[71]:

```
import pandas as pd
new_df = df.copy()
new_df['Salary'].plot(kind='hist', bins=20, edgecolor='black')
plt.xlabel('Salary')
plt.ylabel('Frequency')
plt.title('Salary Distribution')
plt.show()
```



[77]:

```
# Added the missing closing parenthesis
df['log_math'].plot(kind='hist', bins=20, edgecolor='black', color='skyblue')
plt.xlabel('Log10(Salary)')
plt.ylabel('Frequency')
plt.title('Log-Transformed Salary Distribution')
plt.show()
```


Log-Transformed Salary Distribution
Frequency
0
2
4
6
8
10
4.5
4.6
4.7
4.8
4.9
5.0
5.1
5.2
Log10(Salary)

```
[ ]:
```