PR2-DataWrangling

February 6, 2025

[25]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

[2]:
```
df = pd.read_csv('dataset.csv')
```

[4]:
```
df.head()
```

[4]:

	Roll_No.	Name	FE Score	SE Score	Placement	Department
0	1	Sukesh	8.0	8.0	No	IT
1	2	Sukesh	9.0	6.0	No	IT
2	3	Haresh	7.0	9.0	Yes	CS
3	4	Sukesh	8.0	7.0	Yes	IT
4	5	Haresh	10.0	0.0	No	AI&DS

[10]:
```
df.isna().sum()
```

[10]:
```
Roll_No.       0
Name           0
FE Score      10
SE Score       9
Placement      0
Department     0
dtype: int64
```

[19]:
```
df['FE Score'] = df["FE Score"].fillna(df["FE Score"].mean())
```

[13]:
```
df['FE Score'].mean()
```

[13]:
```
7.842931937172775
```

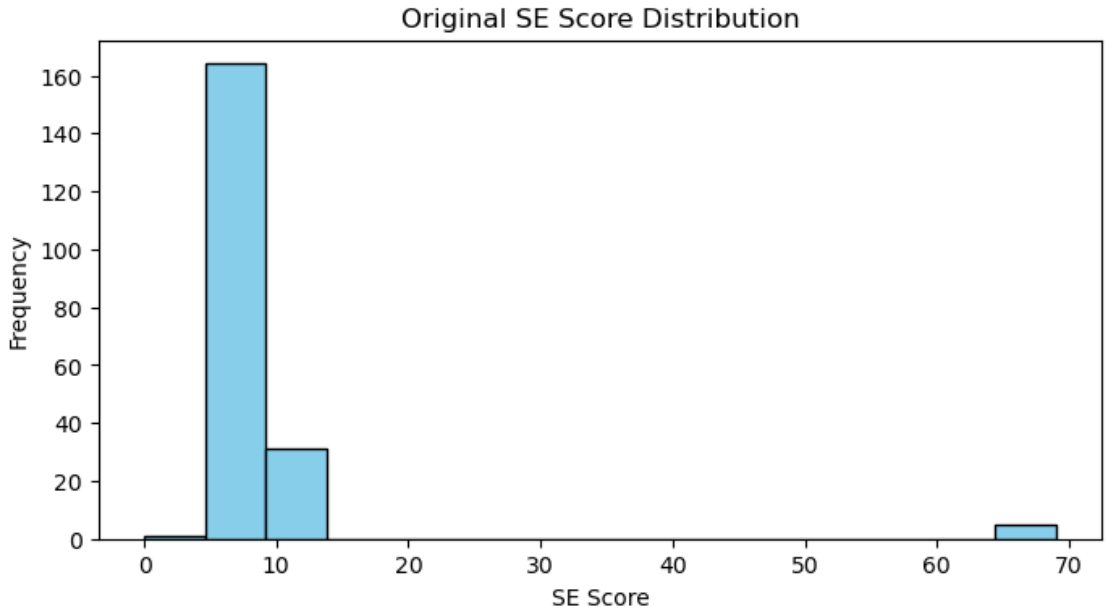
[17]:
```
df['SE Score']=df['SE Score'].fillna(df['SE Score'].median())
```

[20]:
```
df.isna().sum()
```

[20]:
```
Roll_No.      0
Name          0
```

```
FE Score      0
SE Score      0
Placement     0
Department    0
dtype: int64
```

[26]:
```
plt.figure(figsize=(8, 4))
plt.hist(df['SE Score'].dropna(), bins=15, color='skyblue', edgecolor='black')
plt.title("Original SE Score Distribution")
plt.xlabel("SE Score")
plt.ylabel("Frequency")
plt.show()
```



[28]:
```
df['Log_SE_Score'] = df['SE Score'].apply(lambda x: np.log1p(x) if pd.notna(x)␣
↪else x)
```

[29]:
```
plt.figure(figsize=(8, 4))
plt.hist(df['Log_SE_Score'].dropna(), bins=15, color='salmon',␣
↪edgecolor='black')
plt.title("Log Transformed SE Score Distribution")
plt.xlabel("Log(SE Score)")
plt.ylabel("Frequency")
plt.show()
```


Log Transformed SE Score Distribution
Frequency
Log(SE Score)
0
20
40
60
80
100
0
1
2
3
4

[ ]: