

praksh sir ds project/ x Data cleaning PANDAS - Jupyter x +

localhost:8888/notebooks/praksh%20sir%20ds%20project/Data%20cleaning%20PANDAS.ipynb

jupyter Data cleaning PANDAS Last Checkpoint: 41 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [1]: `import pandas as pd`

In [2]: `Data = pd.read_excel(r'C:\Users\hp\OneDrive\Documents\Desktop\Rawdata.xlsx')`
Data

Out[2]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience#\$	34 years	Mumbal	5^00#0	2+
1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
2	Uma#r	Dataanalyst^^#	NaN	NaN	1\$5%000	4> yrs
3	Jane	Ana^^lytics	NaN	Hyderabad	2000^0	NaN
4	Uttam^	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^\$0	10+

In [3]: `Data.shape` *## dimension of the dataframe*

Out[3]: (6, 6)

```
In [4]: len(Data)
```

```
Out[4]: 6
```

```
In [5]: len(Data.columns)
```

```
Out[5]: 6
```

```
In [8]: Data["Domain"]
```

```
Out[8]: 0    Datascience#$  
        1      Testing  
        2  Dataanalyst^^#  
        3    Ana^^lytics  
        4    Statistics  
        5         NLP  
        Name: Domain, dtype: object
```

```
In [9]: Data.isnull()
```

```
Out[9]:
```

	Name	Domain	Age	Location	Salary	Exp
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	True	True	False	False
3	False	False	True	False	False	True
4	False	False	False	True	False	False
5	False	False	False	False	False	False

```
In [96]: Data.isnull().sum()
```

```
Out[96]: Name      0
Domain    0
Age       2
Location  2
Salary    0
Exp       1
dtype: int64
```

```
In [7]: Data["Name"]
```

```
Out[7]: 0    Mike
        1    Teddy^
        2    Uma#r
        3    Jane
        4    Uttam*
        5    Kim
        Name: Name, dtype: object
```

```
In [10]: Data["Age"]
```

```
Out[10]: 0    34 years
        1    45' yr
        2    NaN
        3    NaN
        4    67-yr
        5    55yr
        Name: Age, dtype: object
```

```
In [11]: Data["Location"]
```

```
Out[11]: 0    Mumbai
        1    Bangalore
        2    NaN
        3    Hyderabad
        4    NaN
        5    Delhi
```

```
In [11]: Data["Location"]
```

```
Out[11]: 0    Mumbai  
1    Bangalore  
2         NaN  
3    Hyderabad  
4         NaN  
5         Delhi  
Name: Location, dtype: object
```

```
In [12]: Data["Salary"]
```

```
Out[12]: 0    5^00#0  
1    10%%000  
2    1$5%000  
3    2000^0  
4    30000-  
5    6000^$0  
Name: Salary, dtype: object
```

```
In [13]: Data["Exp"]
```

```
Out[13]: 0    2+  
1    <3  
2    4> yrs  
3    NaN  
4    5+ year  
5    10+  
..    -    ..    . . .
```

```
In [15]: Data[["Name","Domain","Age"]]
```

Out[15]:

	Name	Domain	Age
0	Mike	Datascience#\$	34 years
1	Teddy^	Testing	45' yr
2	Uma#r	Dataanalyst^^#	NaN
3	Jane	Ana^^lytics	NaN
4	Uttam*	Statistics	67-yr
5	Kim	NLP	55yr

```
In [16]: Data[["Name","Domain","Age","Location","Salary","Exp"]]
```

Out[16]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience#\$	34 years	Mumbal	5^00#0	2+
1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
2	Uma#r	Dataanalyst^_^#	NaN	NaN	1\$5%000	4> yrs
3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^\$0	10+

Data cleansing

In [17]: Data

Out[17]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience#\$	34 years	Mumbai	5^00#0	2+
1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
2	Uma#r	Dataanalyst^^#	NaN	NaN	1\$5%000	4> yrs
3	Jane	Ana^alytics	NaN	Hyderbad	2000^0	NaN
4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000*\$0	10+

In [18]: Data["Name"] = Data["Name"].str.replace(r"\W", "")

C:\Users\hp\AppData\Local\Temp\ipykernel_14840\2624503681.py:1: FutureWarning: The default value of regex will change from True to False in a future version.
Data["Name"] = Data["Name"].str.replace(r"\W", "")

praksh sir ds project/ x Data cleaning PANDAS - Jupyter x +

localhost:8888/notebooks/praksh%20sir%20ds%20project/Data%20cleaning%20PANDAS.ipynb

jupyter Data cleaning PANDAS Last Checkpoint: 44 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [19]: Data["Name"]

Out[19]:

0	Mike
1	Teddy
2	Umar
3	Jane
4	Uttam
5	Kim

Name: Name, dtype: object

In [20]: Data["Domain"] = Data["Domain"].str.replace(r'\W', '')

C:\Users\hp\AppData\Local\Temp\ipykernel_14840\804684565.py:1: FutureWarning: The default value of regex will change from True to False in a future version.
Data["Domain"] = Data["Domain"].str.replace(r'\W', '')

In [21]: Data["Domain"]

Out[21]:

0	Datascience
1	Testing
2	Dataanalyst
3	Analytics
4	Statistics
5	NLP

Name: Domain, dtype: object

localhost:8888/notebooks/praksh%20sir%20ds%20project/Data%20cleaning%20PANDAS.ipynb

jupyter

Data cleaning PANDAS

Last Checkpoint: 44 minutes ago (autosaved)

Python 3 (ipykernel)

Logout

File

Edit

View

Insert

Cell

Kernel

Widgets

Help

Trusted

+

Run

Code

In [22]:

▶

Data["Age"] = Data["Age"].str.extract('(\d+)')

In [23]:

▶

Data["Age"]

Out[23]:

0

34

1

45

2

NaN

3

NaN

4

67

5

55

Name: Age, dtype: object

In [24]:

▶

Data

Out[24]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5^00#0	2+
1	Teddy	Testing	45	Bangalore	10%%000	<3
2	Umar	Dataanalyst	NaN	NaN	1\$5%000	4> yrs
3	Jane	Analytics	NaN	Hyderbad	2000^0	NaN
4	Uttam	Statistics	67	NaN	30000-	5+ year
5	Kim	NLP	55	Delhi	6000^\$0	10+

praksh sir ds project/ x Data cleaning PANDAS - Jupyter: x +

localhost:8888/notebooks/praksh%20sir%20ds%20project/Data%20cleaning%20PANDAS.ipynb

jupyter Data cleaning PANDAS Last Checkpoint: 44 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [25]: Data["Location"] = Data["Location"].str.replace(r'\W', "")

C:\Users\hp\AppData\Local\Temp\ipykernel_14840\1052224569.py:1: FutureWarning: The default value of regex will change from True to False in a future version.

Data["Location"] = Data["Location"].str.replace(r'\W', "")

In [26]: Data["Location"]

Out[26]:

0	Mumbai
1	Bangalore
2	NaN
3	Hyderabad
4	NaN
5	Delhi

Name: Location, dtype: object

In [27]: Data["Salary"] = Data["Salary"].str.replace(r'\W', '')

C:\Users\hp\AppData\Local\Temp\ipykernel_14840\1649229267.py:1: FutureWarning: The default value of regex will change from True to False in a future version.

Data["Salary"] = Data["Salary"].str.replace(r'\W', '')

praksh sir ds project/ x Data cleaning PANDAS - Jupyter x +

localhost:8888/notebooks/praksh%20sir%20ds%20project/Data%20cleaning%20PANDAS.ipynb

jupyter Data cleaning PANDAS Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [28]: Data["Salary"]

Out[28]:

0	5000
1	10000
2	15000
3	20000
4	30000
5	60000

Name: Salary, dtype: object

In [29]: Data["Exp"] = Data["Exp"].str.extract('(\d+)')

In [30]: Data["Exp"]

Out[30]:

0	2
1	3
2	4
3	NaN
4	5
5	10

Name: Exp, dtype: object

In [31]: Data

Out[31]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	NaN	NaN	15000	4
3	Jane	Analytics	NaN	Hyderbad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [32]: Data

Out[32]:

Name	Domain	Age	Location	Salary	Exp
------	--------	-----	----------	--------	-----

In [34]: `clean_data`

Out[34]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	NaN	NaN	15000	4
3	Jane	Analytics	NaN	Hyderbad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

Missing value treatement

```
In [35]: import numpy as np
```

```
In [36]: clean_data
```

Out[36]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	NaN	NaN	15000	4
3	Jane	Analytics	NaN	Hyderbad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [37]: `clean_data.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype
---  -
 0   Name        6 non-null      object
 1   Domain      6 non-null      object
 2   Age         4 non-null      object
 3   Location    4 non-null      object
 4   Salary      6 non-null      object
 5   Exp         5 non-null      object
dtypes: object(6)
memory usage: 416.0+ bytes
```

In [38]: `clean_data.head(1)`

Out[38]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbal	5000	2


```
In [39]: clean_data["Age"]
```

```
Out[39]: 0    34  
1    45  
2    NaN  
3    NaN  
4    67  
5    55  
Name: Age, dtype: object
```

```
In [40]: clean_data["Age"] = clean_data["Age"].fillna(np.mean(np.mean(pd.to_numeric(clean_data["Age"]))))
```

```
In [41]: clean_data["Age"]
```

```
Out[41]: 0    34  
1    45  
2    50.25  
3    50.25  
4    67  
5    55  
Name: Age, dtype: object
```

praksh sir ds project/

Data cleaning PANDAS - Jupyter

+

localhost:8888/notebooks/praksh%20sir%20ds%20project/Data%20cleaning%20PANDAS.ipynb

jupyter Data cleaning PANDAS Last Checkpoint: an hour ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

In [42]:

Data

Out[42]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	NaN	NaN	15000	4
3	Jane	Analytics	NaN	Hyderabad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [43]:

clean_data

Out[43]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	NaN	15000	4
3	Jane	Analytics	50.25	Hyderabad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [44]: clean_data["Exp"] = clean_data["Exp"].fillna(np.mean(np.mean(pd.to_numeric(clean_data["Exp"]))))
```

```
In [45]: clean_data
```

Out[45]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	NaN	15000	4
3	Jane	Analytics	50.25	Hyderbad	20000	4.8
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [47]: clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode()[0])
```

```
In [48]: clean_data
```

Out[48]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	Bangalore	15000	4
3	Jane	Analytics	50.25	Hyderbad	20000	4.8
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [50]: clean_data["Age"] = clean_data["Age"].astype(int)
```

```
In [51]: clean_data["Salary"] = clean_data["Salary"].astype(int)
```

```
In [52]: clean_data["Exp"] = clean_data["Exp"].astype(int)
```

```
In [53]: clean_data
```

Out[53]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [55]: clean_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 6 entries, 0 to 5  
Data columns (total 6 columns):  
#   Column      Non-Null Count  Dtype  
---  ---  
0   Name         6 non-null     object  
1   Domain       6 non-null     object  
2   Age          6 non-null     int32  
3   Location     6 non-null     object  
4   Salary       6 non-null     int32  
5   Exp          6 non-null     int32  
dtypes: int32(3), object(3)  
memory usage: 344.0+ bytes
```

📄 + 🔍 📄 ⬆️ ⬆️ ▶️ Run 🛑 ↺️ ⏩ Code ▾ 🗨

```
In [56]: clean_data["Name"] = clean_data["Name"].astype("category")
```

```
In [57]: clean_data["Domain"] = clean_data["Domain"].astype("category")
```

```
In [58]: clean_data["Location"] = clean_data["Location"].astype("category")
```

```
In [59]: clean_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Name        6 non-null     category
1   Domain      6 non-null     category
2   Age         6 non-null     int32
3   Location    6 non-null     category
4   Salary      6 non-null     int32
5   Exp         6 non-null     int32
dtypes: category(3), int32(3)
memory usage: 862.0 bytes
```

```
In [60]: clean_data
```

```
Out[60]:
```

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [61]: clean_data.to_csv("clean_data.csv")
```

```
In [62]: import os  
os.getcwd()
```

```
Out[62]: 'E:\\python 2023\\praksh sir ds project'
```

```
In [63]: clean_data.columns
```

```
Out[63]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```



```
In [64]: ► import matplotlib.pyplot as plt # visualisation
import seaborn as sns # Advance visualization
```

```
In [65]: ► import warnings
warnings.filterwarnings("ignore")
```

```
In [66]: ► clean_data
```

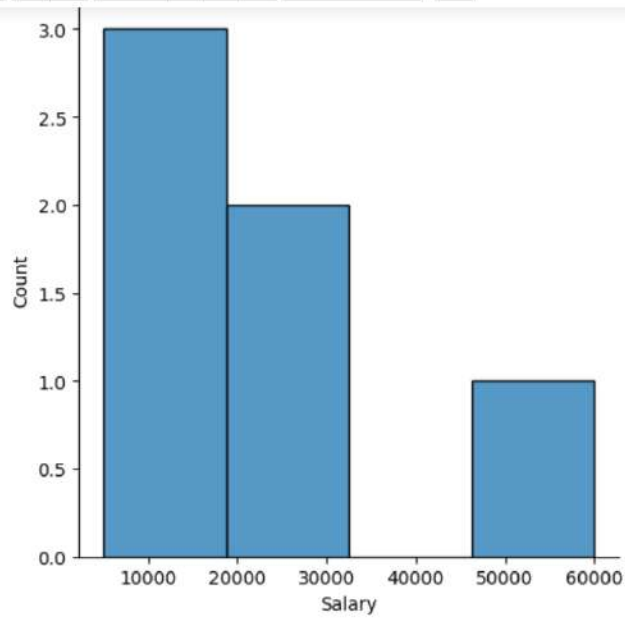
Out[66]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [67]: clean_data["Salary"]
```

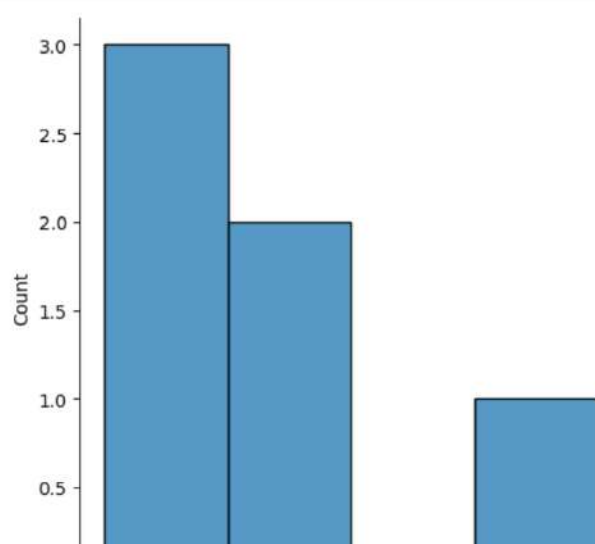
```
Out[67]: 0    5000  
1   10000  
2   15000  
3   20000  
4   30000  
5   60000  
Name: Salary, dtype: int32
```

```
In [68]: vis1 = sns.displot(clean_data['Salary'])
```



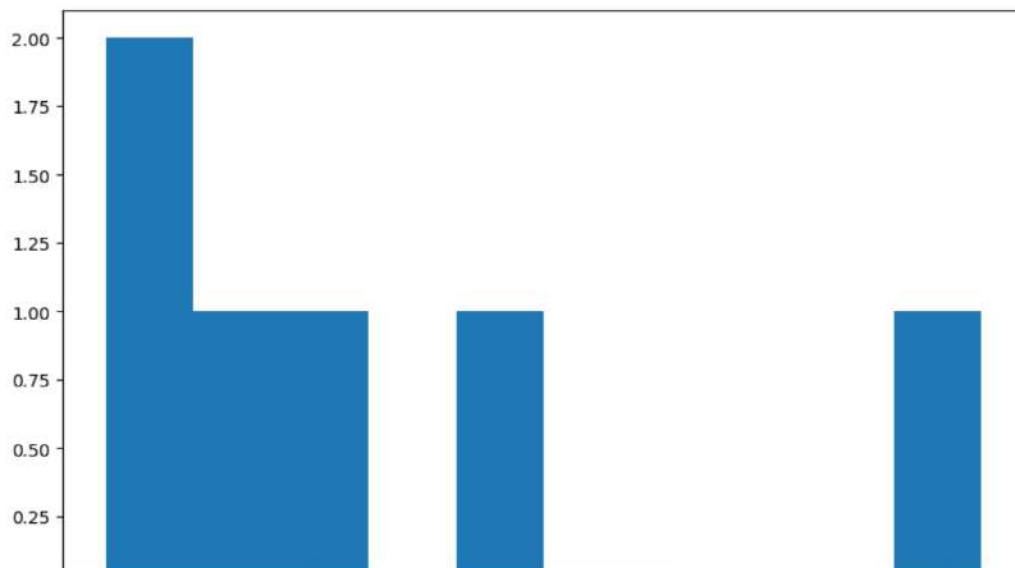
```
In [69]: plt.rcParams["figure.figsize"] = 10,6
```

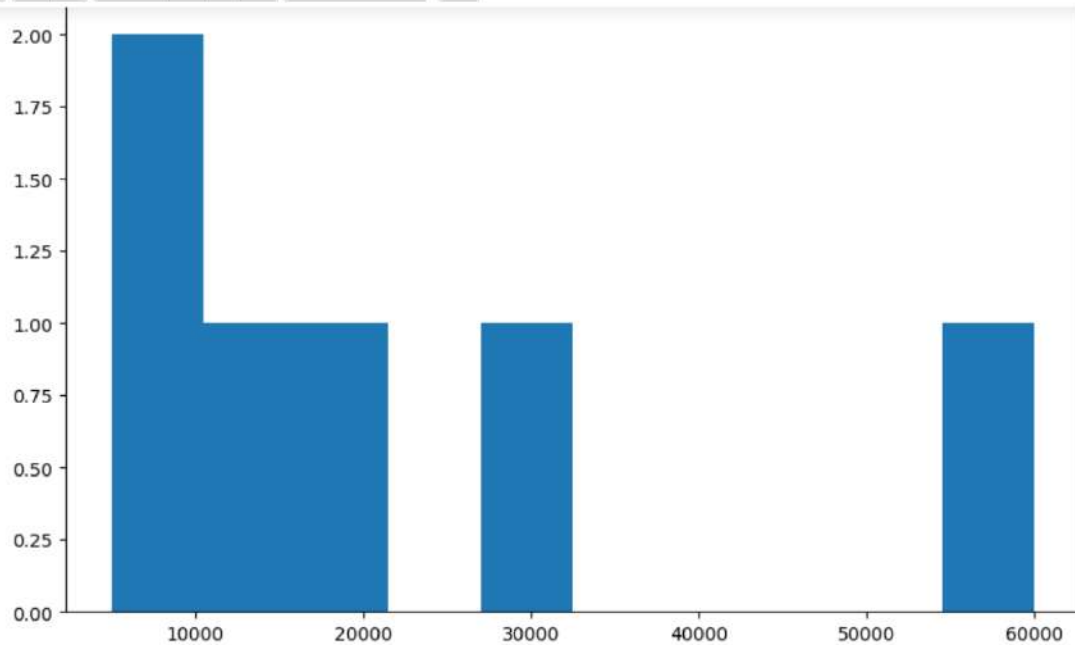
```
In [70]: vis1 = sns.displot(clean_data["Salary"])
```



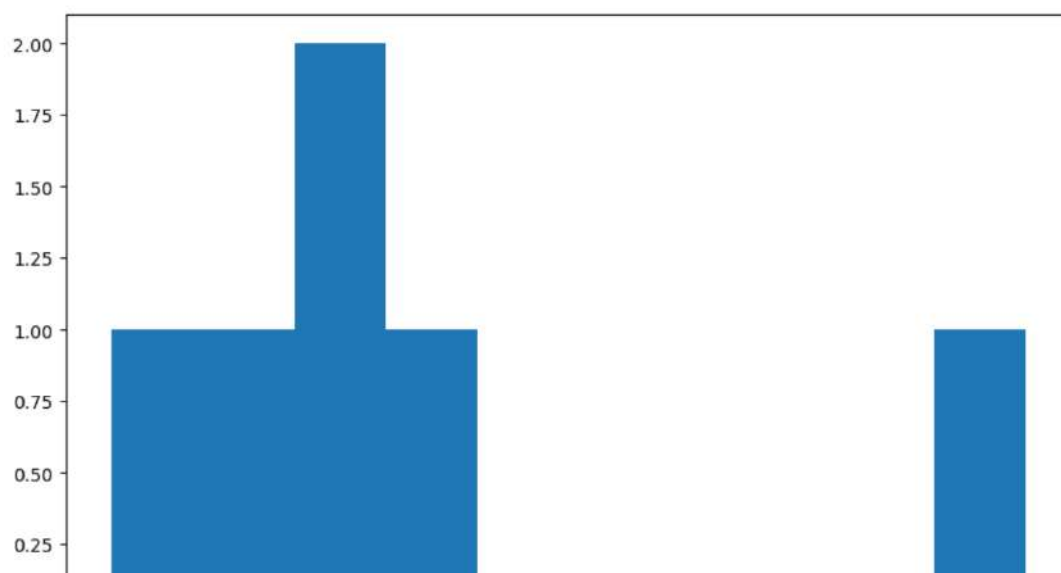


```
In [71]: vis2 = plt.hist(clean_data["Salary"])
```

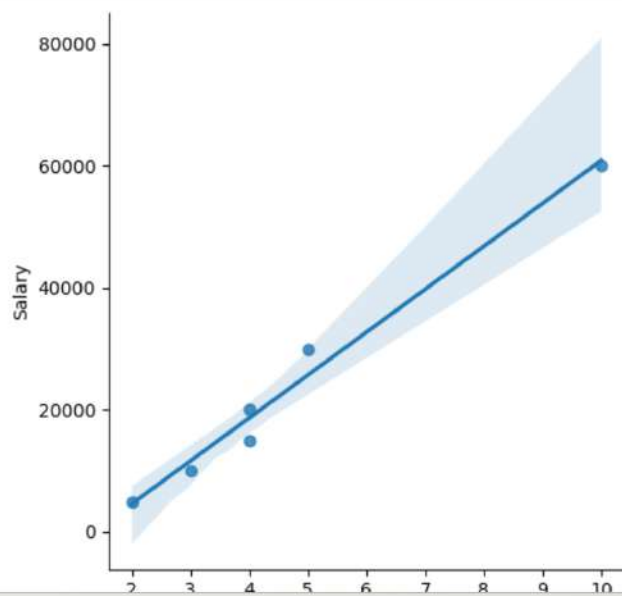




```
In [72]: vis3 = plt.hist(clean_data["Exp"])
```

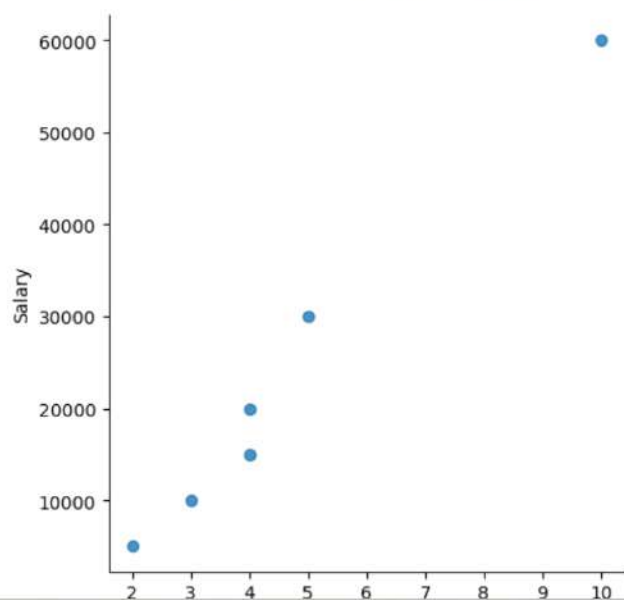


```
In [73]: vis4 = sns.lmplot(data=clean_data, x = "Exp", y="Salary")
```





```
In [74]: vis5 = sns.lmplot(data=clean_data, x = "Exp", y="Salary", fit_reg = False)
```



In [77]: `clean_data[:]`

Out[77]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [78]: `clean_data[:2]`

Out[78]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3

In [79]: `clean_data[2:]`

Out[79]:

	Name	Domain	Age	Location	Salary	Exp
2	Umar	Dataanalyst	50	Bangalore	15000	4

```
In [83]: ► x_iv = clean_data.drop(["Salary"],axis=1)
```

```
In [84]: ► clean_data
```

Out[84]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [86]: x_iv.columns
```

```
Out[86]: Index(['Name', 'Domain', 'Age', 'Location', 'Exp'], dtype='object')
```

```
In [87]: clean_data
```

```
Out[87]:
```

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [88]: ► y_dv =clean_data.drop(["Name", "Domain", "Age", "Location","Exp"],axis=1)
```

```
In [89]: ► y_dv
```

Out[89]:

Salary	
0	5000
1	10000
2	15000
3	20000
4	30000
5	60000

File Edit View Insert Cell Help Run Debug Code

In [92]: y_dv

Out[92]:

	Salary
0	5000
1	10000
2	15000
3	20000
4	30000
5	60000

```
In [94]: imputation = pd.get_dummies(clean_data)
```

```
In [95]: imputation
```

Out[95]:

	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Uttam	Domain_Analytics	Domain_Dataanalyst	Domain_Datasc
0	34	5000	2	0	0	1	0	0	0	0	0	
1	45	10000	3	0	0	0	1	0	0	0	0	
2	50	15000	4	0	0	0	0	1	0	0	1	
3	50	20000	4	1	0	0	0	0	0	1	0	
4	67	30000	5	0	0	0	0	0	1	0	0	
5	55	60000	10	0	1	0	0	0	0	0	0	