File Edit Selection ··· vscodeproject

emp.ipynb

0Testing > emp.ipynb > ...

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables ··· base (Python 3.12.4)

[1]
```python
import pandas as pd
```
Python

[2]
```python
emp = pd.read_excel(r"C:\Users\mohap\Downloads\Rawdata.xlsx")
```
Python

[3]
```python
emp
```
Python

	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

0 0 Connect Cell 1 of 69 Go Live

Sports headline With FIFA annou... Search ENG IN 22:28 27-03-2025

File Edit Selection ··· vscodeproject

emp.ipynb

0Testing > emp.ipynb > ...

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables ··· base (Python 3.12.4)

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^^lytics	NaN	Hyderbad	2000^0	NaN
4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^$0	10+

[4]

```
emp.columns
```

Python

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

```
emp.head()
```

[5]

Python

0 0 Connect Cell 1 of 69 Go Live

Sports headline With FIFA annou... Search ENG IN 22:29 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

OTesting > emp.ipynb > ...

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

```
5       Kim
Name: Name, dtype: object
```

```python
emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
emp['Name']
```

[20] Python

```
0     Mike
1    Teddy
2     Umar
3     Jane
4    Uttam
5      Kim
Name: Name, dtype: object
```

```python
emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
emp['Domain']
```

[21] Python

```
0    Datascience
1        Testing
2    Dataanalyst
3      Analytics
4     Statistics
5            NLP
Name: Domain, dtype: object
```

```python
emp['Age'] = emp['Age'].str.replace(r'\W','',regex=True)
emp['Age']
```

[22] Python

```
0    34years
1      45yr
```

0 0 Connect Cell 1 of 60 Go Live

Watchlist Ideas Search ENG IN 22:30 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

0Testing > emp.ipynb > ...

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

```
Name: Age, dtype: object
```

```python
emp['Age'] = emp['Age'].str.extract(r'(\d+)')
emp['Age']
```

[23] Python

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

```python
emp['Location'] = emp['Location'].str.replace(r'\W','',regex=True)
emp['Location']
```

[24] Python

```
0      Mumbai
1   Bangalore
2         NaN
3     Hyderbad
4         NaN
5       Delhi
Name: Location, dtype: object
```

```python
emp['Salary'] = emp['Salary'].str.replace(r'\W','',regex=True)
emp['Salary']
```

[25] Python

```
0     5000
1    10000
2    15000
```

0 0 Connect Cell 1 of 60 Go Live

Watchlist Ideas Search ENG IN 22:30 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

0Testing > emp.ipynb > ...

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

```python
emp['Exp'] = emp['Exp'].str.extract(r'(\d+)')
emp['Exp']
```

[27] Python

```
0      2
1      3
2      4
3    NaN
4      5
5     10
Name: Exp, dtype: object
```

```python
emp
```

[28] Python

	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

we have raw data we use regex to clean the data and removed all noise characted from the dataset

We can also work in same things in sql query as well

```python
clean_data = emp.copy()
```

0 0 Connect Cell 1 of 60 Go Live

Air: Moderate
Now

Search

ENG IN 22:31 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

0Testing > emp.ipynb > ...

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

[31]
```python
clean_data['Age']
```

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

[32]
```python
import numpy as np
```

Generate + Code + Markdown

[33]
```python
clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'])))
```

[34]
```python
clean_data['Age']
```

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

[35]
```python
clean_data['Exp']
```

0 0 Connect Cell 1 of 60 Go Live

Air: Moderate
Now

Search

ENG IN 22:32 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

OTesting > emp.ipynb > ...

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

```
Name: Exp, dtype: object
```

```python
clean_data['Location'].isnull().sum()
clean_data['Location']
```

[37] Python

```
0      Mumbai
1   Bangalore
2         NaN
3    Hyderbad
4         NaN
5       Delhi
Name: Location, dtype: object
```

```python
clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode()[0])
```

[38] Python

```python
clean_data['Location']
```

[39] Python

```
0      Mumbai
1   Bangalore
2   Bangalore
3    Hyderbad
4   Bangalore
5       Delhi
Name: Location, dtype: object
```

```python
clean_data
```

[40] Python

0 0 Connect Spaces: 4 Ln 30 of 60 Go Live

Air: Moderate
Now

Search

ENG IN 22:33 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

0Testing > emp.ipynb > ...

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

EDA Technique

```python
import matplotlib.pyplot as plt
import seaborn as sns
```

[51] Python

```python
import warnings
warnings.filterwarnings('ignore')
```

[52] Python

Generate + Code + Markdown

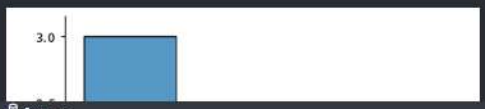
```python
clean_data['Salary']
```

[53] Python

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

```python
vis1 = sns.displot(clean_data['Salary'])
```

[54] Python



0 0 Connect Spaces: 4 Cell 30 of 60 Go Live

29°C Partly cloudy Search ENG IN 22:34 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

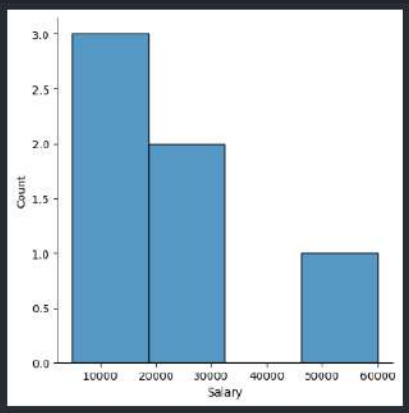
0Testing > emp.ipynb > _

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

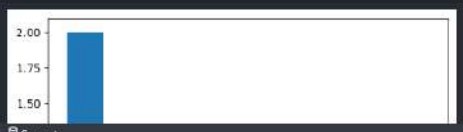
```
vis1 = sns.displot(clean_data['Salary'])
```

[54] Python




```
vis1 = plt.hist(clean_data['Salary'])
```

[55] Python



Connect Spaces: 4 Cell 30 of 69 Go Live

29°C Partly cloudy Search ENG IN 22:34 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

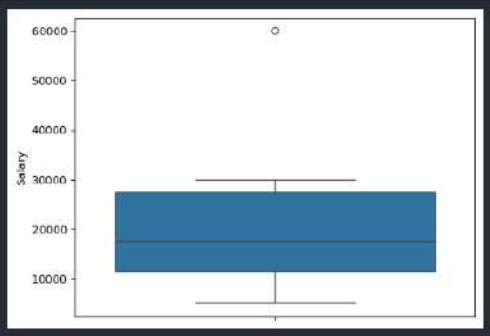
OTesting > emp.ipynb > _

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

```
vis2 = sns.boxplot(clean_data['Salary'])
```

[56] Python




```
vis3= sns.lmplot(data=clean_data, x='Exp', y='Salary')
```

[57] Python

Connect Spaces: 4 Cell 30 of 69 Go Live

29°C Partly cloudy Search ENG IN 22:35 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

0Testing > emp.ipynb > _

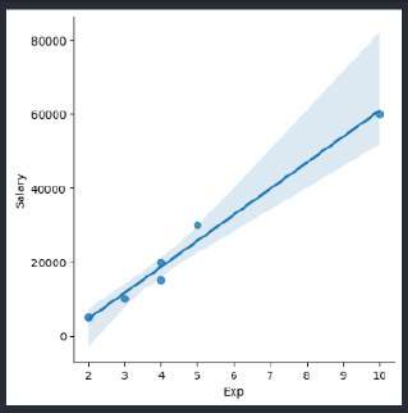
Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

[57]

```
vis3= sns.lmplot(data=clean_data, x='Exp', y='Salary')
```

Python



[58]

```
vis4 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = False)
```

Python

Connect

Spaces: 4 Cell 10 of 69 Go Live

29°C Partly cloudy

Search

ENG IN

22:36 27-03-2025

File Edit Selection View Go ...

vscodeproject

emp.ipynb

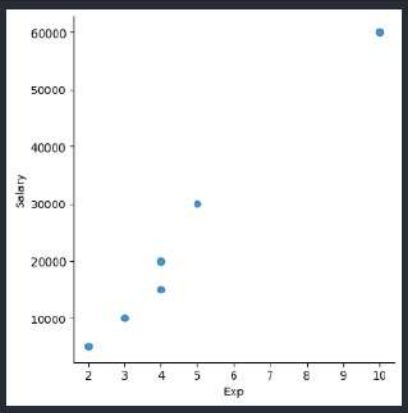
OTesting > emp.ipynb > _

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

```
vis4 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = False)
```

Python



Generate + Code + Markdown

Add Code Cell

```
clean_data[:]
```

Python

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

Connect

Spaces: 4 Cell 30 of 69 Go Live

29°C Partly cloudy

Search

ENG IN

22:36 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

0Testing > emp.ipynb > _

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

```
clean_data[:]
```

Python

	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

```
clean_data[0:6:2]
```

Python

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
2	Umar	Dataanalyst	50	Bangalore	15000	4
4	Uttam	Statistics	67	Bangalore	30000	5

```
clean_data[::-1]
```

Python

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

```
clean_data.columns
```

Python

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

Connect Spaces: 4 Cell 30 of 69 Go Live

29°C Partly cloudy Search ENG IN 22:37 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

0Testing > emp.ipynb > _

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

```
clean_data.columns
```

Python

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

```
X_iv = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
```

Python

```
X_iv
```

Python

	Name	Domain	Age	Location	Exp
0	Mike	Datascience	34	Mumbai	2
1	Teddy	Testing	45	Bangalore	3
2	Umar	Datasanalyst	50	Bangalore	4
3	Jane	Analytics	50	Hyderbad	4
4	Uttam	Statistics	67	Bangalore	5
5	Kim	NLP	55	Delhi	10

```
y_dv = clean_data['Salary']
y_dv
```

Python

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

```
emp
```

Python

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

Connect Spaces: 4 Cell 30 of 69 Go Live

29°C Partly cloudy Search ENG IN 22:38 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

0Testing > emp.ipynb > _

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

5	Kim	NLP	55	Delhi	60000	10

```
X_iv
```

Python

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

```
y_dv
```

Python

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

```
clean_data
```

Python

	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

Connect Spaces: 4 Cell 30 of 69 Go Live

29°C Partly cloudy Search ENG IN 22:38 27-03-2025

File Edit Selection View Go ... vscodeproject

emp.ipynb

0Testing > emp.ipynb > _

Generate + Code + Markdown | Run All Restart Clear All Outputs | Jupyter Variables Outline ...

base (Python 3.12.4)

[72]
```
imputation = pd.get_dummies(clean_data)
```
Python

[73]
```
imputation
```
Python

	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Uttam	Domain_Analytics	Domain_Dataanalyst	Domain_Datascience	Domain_NLP	Domain_Statistics	Domain_Testing	Location_Bangalore	Location_Delhi	Location_
0	34	5000	2	False	False	True	False	False	False	False	False	True	False	False	False	False	False	
1	45	10000	3	False	False	False	True	False	False	False	False	False	False	False	True	True	False	
2	50	15000	4	False	False	False	False	True	False	False	True	False	False	False	False	True	False	
3	50	20000	4	True	False	False	False	False	False	True	False	False	False	False	False	False	False	
4	67	30000	5	False	False	False	False	False	True	False	False	False	False	True	False	True	False	
5	55	60000	10	False	True	False	False	False	False	False	False	False	True	False	False	False	True	

raw data with lot of regex, missing, uncleandata

regex, clean

fill missing numerical & cateigroica

clean_dataset (data cleaning)

outlier treatement, univarient, bivarient, co-relation

split the data into x_i.v & y_dv

impute cateogrica data to numerical

Connect Spaces: 4 Cell 10 of 69 Go Live

29°C Partly cloudy Search ENG IN 22:39 27-03-2025