Raw data to Clean Data

```
import pandas as pd
 In [2]:
 In [6]:
          emp = pd.read_excel(r'/Users/babarhussain/Documents/Rawdata.xlsx')
          emp
 Out[6]:
              Name
                           Domain
                                             Location
                                       Age
                                                          Salary
                                                                    Exp
                      Datascience#$
                                                                     2+
               Mike
                                   34 years
                                              Mumbai
                                                         5^00#0
             Teddy^
                            Testing
                                      45' yr Bangalore
                                                       10%%000
                                                                     <3
          2
             Uma#r
                     Dataanalyst^^#
                                       NaN
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
                        Ana^^lytics
          3
                                             Hyderbad
                                                         2000^0
               Jane
                                       NaN
                                                                    NaN
          4
             Uttam*
                          Statistics
                                      67-vr
                                                 NaN
                                                         30000- 5+ year
          5
                Kim
                              NLP
                                       55vr
                                                 Delhi
                                                        6000^$0
                                                                    10 +
 In [8]:
          emp.shape
 Out[8]:
          (6, 6)
In [10]:
          emp.columns
Out[10]:
          Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='obj
          ect')
In [12]:
         emp.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
         #
              Column
                        Non-Null Count
                                          Dtype
                        6 non-null
         0
              Name
                                          object
         1
              Domain
                        6 non-null
                                          object
                        4 non-null
         2
              Age
                                          object
         3
              Location 4 non-null
                                          object
         4
                        6 non-null
              Salary
                                          object
         5
                         5 non-null
                                          object
              Exp
        dtypes: object(6)
        memory usage: 420.0+ bytes
In [14]: emp.isnull()
```

about:srcdoc Page 1 of 16

Out[14]:		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 [18]: emp.isnull().any().any()

Out[18]: True

In [20]: emp.describe()

Out[20]:

	Name	Domain	Age	Location	Salary	Exp
count	6	6	4	4	6	5
unique	6	6	4	4	6	5
top	Mike	Datascience#\$	34 years	Mumbai	5^00#0	2+
freq	1	1	1	1	1	1

cleaning Data

Out [23]: 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^*lytics NaN Hyderbad 2000*0 NaN 4 Uttam* Statistics 67-yr NaN 30000- 5+ year	In [23]:	е	етр												
1 Teddy^ Testing 45' yr Bangalore 10%%000 <3	Out[23]:		Name	Domain	Age	Location	Salary	Exp							
 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 		0	Mike	Datascience#\$	34 years	Mumbai	5^00#0	2+							
3 Jane Ana^^lytics NaN Hyderbad 2000^0 NaN 4 Uttam* Statistics 67-yr NaN 30000- 5+ year		1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3							
4 Uttam* Statistics 67-yr NaN 30000- 5+ year		2	Uma#r	Dataanalyst^^#	NaN	NaN	1\$5%000	4> yrs							
,		3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN							
F Vim NLD FEVE Dollai 60000¢0 10.		4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year							
5 KIIII NLP SSYI DEIIII 6000 \$0 10+		5	Kim	NLP	55yr	Delhi	6000^\$0	10+							

In [25]: emp['Name']

about:srcdoc Page 2 of 16

```
Out[25]: 0
                 Mike
          1
               Teddy^
          2
                Uma#r
          3
                 Jane
          4
               Uttam*
          5
                  Kim
          Name: Name, dtype: object
         emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
In [27]:
          emp['Name']
Out[27]:
          0
                Mike
               Teddy
          1
          2
                Umar
          3
                Jane
          4
               Uttam
          5
                 Kim
          Name: Name, dtype: object
         emp['Domain'] = emp['Domain'].str.replace(r'\W','', regex=True)
          emp['Domain']
Out[29]:
          0
               Datascience
          1
                   Testing
          2
               Dataanalyst
                 Analytics
          3
          4
                Statistics
                       NLP
          Name: Domain, dtype: object
         emp['Age']
In [31]:
Out[31]:
          0
               34 years
                 45' yr
          1
          2
                    NaN
          3
                    NaN
          4
                  67-yr
                   55yr
          Name: Age, dtype: object
         emp['Age'] = emp['Age'].replace(r'^0-9.','', regex=True)
In [33]:
          emp['Age']
               34 years
Out[33]:
          0
          1
                 45' yr
          2
                    NaN
          3
                    NaN
          4
                  67-yr
          5
                   55yr
          Name: Age, dtype: object
         emp['Age'] = emp['Age'].replace(r'[^0-9.]','',regex=True)
In [35]:
          emp['Age']
```

about:srcdoc Page 3 of 16

```
Out[35]:
                 34
          0
          1
                 45
          2
               NaN
          3
                NaN
          4
                 67
          5
                 55
          Name: Age, dtype: object
In [37]:
          emp
Out[37]:
                                      Location
             Name
                        Domain
                                Age
                                                   Salary
                                                             Exp
          0
              Mike
                    Datascience
                                  34
                                       Mumbai
                                                  5^00#0
                                                               2+
            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
            Uttam
                       Statistics
                                  67
                                          NaN
                                                  30000-
                                                          5+ year
               Kim
                           NLP
                                  55
                                                 6000^$0
                                                              10+
          5
                                          Delhi
In [39]:
          emp['Salary']
Out[39]:
          0
                 5^00#0
          1
                10%%000
          2
                1$5%000
          3
                 2000^0
          4
                 30000-
          5
                6000^$0
          Name: Salary, dtype: object
In [43]:
          emp['Salary'] = emp['Salary'].replace(r'[^0-9.]','',regex=True)
          emp['Salary']
Out[43]:
          0
                 5000
          1
                10000
          2
                15000
          3
                20000
          4
                30000
          5
                60000
          Name: Salary, dtype: object
In [45]:
          emp['Exp']
Out[45]:
          0
                     2+
          1
                     <3
          2
                 4> yrs
          3
                    NaN
          4
                5+ year
          5
                    10+
          Name: Exp, dtype: object
```

about:srcdoc Page 4 of 16

```
In [47]:
          emp['Exp'] = emp['Exp'].replace(r'[^0-9.]','',regex=True)
          emp['Exp']
Out[47]:
                   2
                   3
           1
           2
                   4
           3
                NaN
                   5
           5
                 10
           Name: Exp, dtype: object
In [49]:
          emp
Out [49]:
              Name
                         Domain
                                        Location Salary
                                  Age
                                                          Exp
                                                            2
           0
               Mike
                     Datascience
                                   34
                                         Mumbai
                                                   5000
             Teddy
                         Testing
                                   45
                                       Bangalore
                                                  10000
                                                            3
           2
              Umar
                     Dataanalyst NaN
                                            NaN
                                                  15000
                                                            4
           3
                                       Hyderbad
                                                 20000
               Jane
                        Analytics
                                 NaN
                                                         NaN
                        Statistics
                                                  30000
                                                            5
           4
             Uttam
                                   67
                                            NaN
           5
                Kim
                            NLP
                                            Delhi
                                                  60000
                                                           10
                                   55
In [51]:
          clean_data = emp.copy()
          clean_data
Out [51]:
              Name
                                        Location Salary
                         Domain
                                  Age
                                                          Exp
          0
               Mike
                     Datascience
                                   34
                                         Mumbai
                                                   5000
                                                            2
             Teddy
                                   45
                                       Bangalore
                                                  10000
                                                            3
                         Testing
           2
                     Dataanalyst NaN
                                                  15000
                                                            4
              Umar
                                            NaN
           3
               Jane
                        Analytics
                                 NaN
                                       Hyderbad
                                                  20000
                        Statistics
           4
             Uttam
                                   67
                                            NaN
                                                  30000
                                                            5
                            NLP
           5
                Kim
                                   55
                                            Delhi 60000
                                                           10
```

EDA Technique Apply

find Missing Values

```
In [55]: clean_data
```

about:srcdoc Page 5 of 16

Out[55]:		Name	Domain	Age	Location	Salary	Ехр
	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 [57]: clean_data.isnull().any().any()

Out[57]: True

In [59]: clean_data.isnull()

Out[59]: Name Domain Age Location Salary Exp 0 False True True False 3 False False True False False True 4 True False False False False False

False

False

In [61]: clean_data.isna()

False

Out[61]: Name Domain Age Location Salary Exp False False False False False False 1 False False False **False** False False 2 False **False** True False False True 3 False False False True False True **False** False False True False False 5 False False False **False** False False

In [63]: clean_data.isnull().sum()

False

about:srcdoc Page 6 of 16

False False

```
Out[63]:
                       0
          Name
          Domain
                       0
          Age
                       2
                       2
          Location
          Salary
                       0
                       1
          Exp
          dtype: int64
          clean_data['Age']
In [65]:
Out[65]:
          0
                 34
          1
                 45
          2
                NaN
          3
                NaN
          4
                 67
                 55
          5
          Name: Age, dtype: object
In [67]:
          import numpy as np
          clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_
In [71]:
          clean_data['Age']
Out[71]:
          0
                   34
          1
                   45
          2
                50.25
          3
                50.25
          4
                   67
          5
                   55
          Name: Age, dtype: object
In [73]:
          clean_data
Out[73]:
                        Domain
             Name
                                  Age
                                        Location Salary
                                                         Exp
          0
              Mike
                    Datascience
                                   34
                                         Mumbai
                                                  5000
                                                           2
             Teddy
                        Testing
                                   45
                                       Bangalore
                                                 10000
                                                           3
          2
              Umar
                    Dataanalyst
                                50.25
                                            NaN
                                                 15000
                                                           4
          3
              Jane
                       Analytics 50.25
                                       Hyderbad
                                                 20000 NaN
             Uttam
                       Statistics
                                   67
                                            NaN
                                                 30000
                                                           5
          5
               Kim
                           NLP
                                   55
                                           Delhi
                                                 60000
                                                          10
In [75]:
          emp['Exp'].isnull()
```

about:srcdoc Page 7 of 16

```
Out[75]:
          0
                False
                False
          1
          2
                False
          3
                 True
          4
                False
          5
                False
          Name: Exp, dtype: bool
          clean_data['Exp'] = emp['Exp'].fillna(np.mean(pd.to_numeric(clean_data['E
In [83]:
          clean_data['Exp']
Out[83]:
                  2
                  3
          1
          2
                  4
          3
                4.8
          4
                  5
          5
                 10
          Name: Exp, dtype: object
In [85]:
          clean_data
Out[85]:
                        Domain
                                        Location
             Name
                                  Age
                                                  Salary Exp
          0
               Mike
                    Datascience
                                    34
                                         Mumbai
                                                   5000
                                                            2
             Teddy
                         Testing
                                    45
                                       Bangalore
                                                  10000
          2
              Umar
                     Dataanalyst 50.25
                                             NaN
                                                  15000
                                                            4
          3
               Jane
                       Analytics 50.25
                                        Hyderbad
                                                  20000
                                                          4.8
                                                            5
             Uttam
                       Statistics
                                    67
                                                  30000
                                            NaN
                           NLP
          5
               Kim
                                    55
                                            Delhi
                                                  60000
                                                           10
In [87]:
          emp
Out[87]:
             Name
                        Domain
                                       Location Salary Exp
                                 Age
                    Datascience
                                                           2
          0
               Mike
                                  34
                                        Mumbai
                                                  5000
             Teddy
                         Testing
                                  45
                                      Bangalore
                                                 10000
                                                           3
                                                           4
          2
              Umar
                     Dataanalyst
                                NaN
                                           NaN
                                                 15000
                                      Hyderbad 20000
          3
               Jane
                       Analytics NaN
                                                         4.8
          4
                                                30000
                                                           5
             Uttam
                       Statistics
                                  67
                                           NaN
          5
               Kim
                           NLP
                                  55
                                           Delhi 60000
                                                          10
          clean_data['Location'] = clean_data['Location'].fillna(clean_data['Locati
In [89]:
```

about:srcdoc Page 8 of 16

clean data['Location']

```
Out[89]:
          0
                   Mumbai
                Bangalore
          1
          2
                Bangalore
          3
                 Hyderbad
          4
                Bangalore
          5
                    Delhi
          Name: Location, dtype: object
In [91]:
          clean_data
Out [91]:
             Name
                        Domain
                                 Age
                                       Location Salary Exp
          0
              Mike
                    Datascience
                                   34
                                        Mumbai
                                                  5000
                                                           2
             Teddy
                                   45
                                      Bangalore
                                                 10000
                                                           3
                        Testing
          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
                           NLP
                                                          10
               Kim
                                   55
                                           Delhi
                                                 60000
In [93]:
          clean_data.isnull().any().any()
Out[93]:
          False
In [95]:
          clean_data.isnull().sum()
Out[95]:
          Name
                       0
                       0
          Domain
          Age
                       0
          Location
                       0
          Salary
                       0
          Exp
          dtype: int64
In [99]: 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
                         6 non-null
                                          object
              Domain
         2
              Age
                         6 non-null
                                          object
          3
              Location
                        6 non-null
                                          object
          4
              Salary
                         6 non-null
                                          object
          5
                         6 non-null
              Exp
                                          object
        dtypes: object(6)
        memory usage: 420.0+ bytes
```

about:srcdoc Page 9 of 16

```
In [103...
          clean_data[['Age','Salary','Exp']] = clean_data[['Age','Salary','Exp']].a
          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
                        6 non-null
                                         int64
              Age
          3
              Location 6 non-null
                                         object
         4
              Salary
                        6 non-null
                                         int64
          5
                        6 non-null
                                         int64
              Exp
        dtypes: int64(3), object(3)
        memory usage: 420.0+ bytes
In [107... clean_data[['Name','Domain','Location']] = clean_data[['Name','Domain','L
          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
                        6 non-null
                                         int64
              Age
          3
              Location 6 non-null
                                         category
              Salary
          4
                        6 non-null
                                         int64
          5
              Exp
                        6 non-null
                                         int64
         dtypes: category(3), int64(3)
        memory usage: 938.0 bytes
In [109... clean_data
Out [109...
             Name
                       Domain Age
                                     Location Salary Exp
              Mike
                   Datascience
                                 34
                                                5000
                                                        2
          0
                                      Mumbai
          1 Teddy
                        Testing
                                 45
                                    Bangalore
                                               10000
                                                        3
          2
             Umar
                    Dataanalyst
                                 50
                                    Bangalore
                                               15000
                                                        4
          3
                      Analytics
                                 50
                                     Hyderbad 20000
                                                        4
              Jane
                                     Bangalore 30000
                                                        5
          4 Uttam
                      Statistics
                                 67
          5
                          NLP
               Kim
                                 55
                                         Delhi 60000
                                                       10
          clean_data.to_csv('clean_data.csv', index=False)
In [119...
In [121...
          import os
          os.getcwd()
```

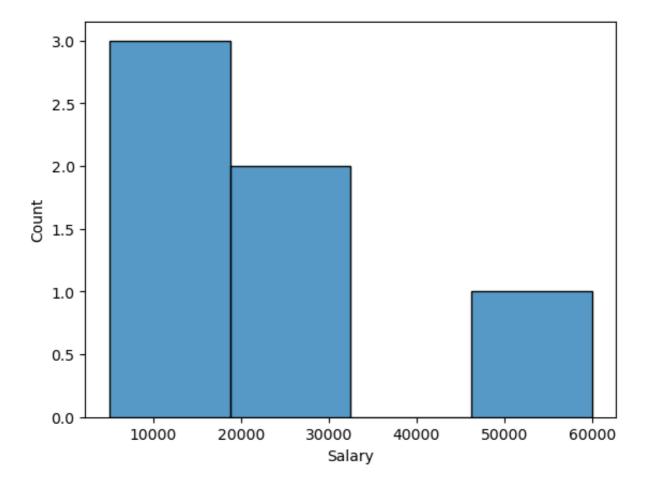
about:srcdoc Page 10 of 16

Out[121... '/Users/babarhussain/EDA Example'

```
In [123...
          clean_data
Out [123...
                        Domain Age
                                       Location Salary Exp
             Name
                                                          2
               Mike
                    Datascience
                                  34
                                        Mumbai
                                                  5000
           1 Teddy
                         Testing
                                  45
                                      Bangalore 10000
                                                          3
              Umar
           2
                     Dataanalyst
                                  50
                                      Bangalore
                                                15000
                                                          4
           3
                                      Hyderbad 20000
               Jane
                       Analytics
                                  50
           4 Uttam
                                      Bangalore 30000
                       Statistics
                                  67
                                                          5
           5
               Kim
                           NLP
                                  55
                                          Delhi 60000
                                                         10
In [125...
          import matplotlib.pyplot as plt
          import seaborn as ss
In [127...
          import warnings
          warnings.filterwarnings('ignore')
          clean_data['Salary']
In [129...
Out [129...
          0
                 5000
           1
                10000
           2
                15000
           3
                20000
           4
                30000
           5
                60000
           Name: Salary, dtype: int64
```

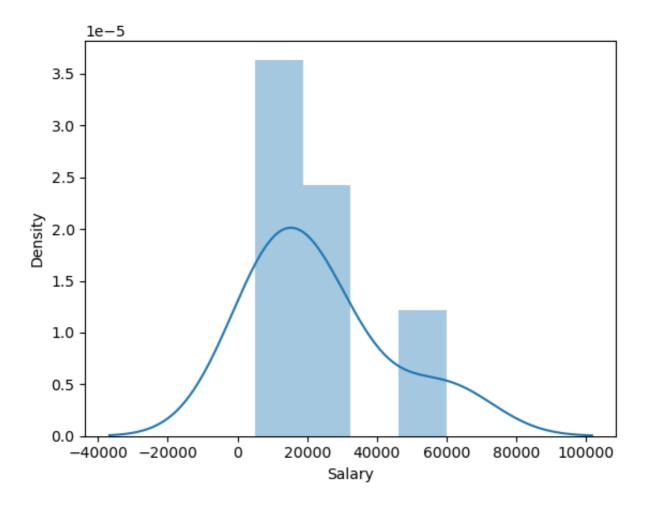
In [131... vs = ss.histplot(clean_data['Salary']) #Outlier Treatment

about:srcdoc Page 11 of 16



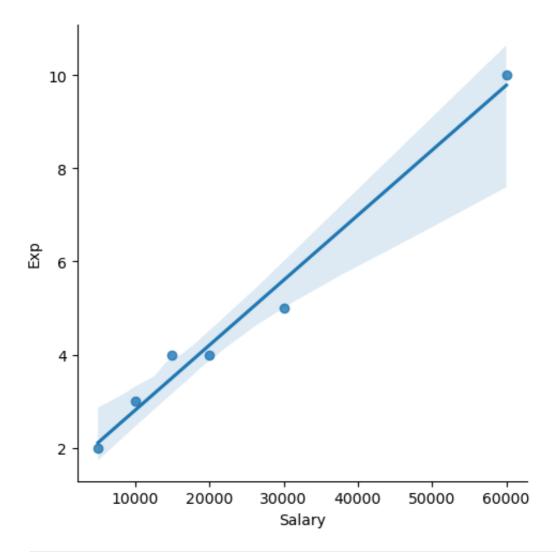
In [135... vs1 = ss.distplot(clean_data['Salary']) # univariate Analysis

about:srcdoc Page 12 of 16



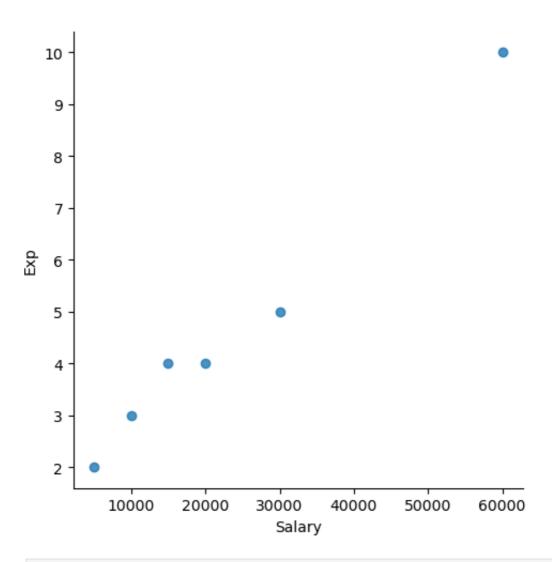
In [137... vs3 = ss.lmplot(data=clean_data, x='Salary', y='Exp') # Bivariate Analysi

about:srcdoc Page 13 of 16



In [141... vs3 = ss.lmplot(data=clean_data, x='Salary', y='Exp', fit_reg =False) # B

about:srcdoc Page 14 of 16



```
In [143... # Variable Identification
In [145... x_vr = clean_data[['Name','Domain','Age','Location','Exp']] # here all ar x_vr
```

Out[145		Name	Domain	Age	Location	Ехр
	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

```
In [147... y_vr = clean_data[['Salary']] # here Slary is dependent variable
y_vr
```

about:srcdoc Page 15 of 16

Out[147		Salary
	0	5000
	1	10000
	2	15000
	3	20000
	4	30000
	5	60000

55 60000

10

5

In [151... imputation = pd.get_dummies(clean_data, dtype=int) #dummy variables (also
imputation

Out[151		Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Un
	0	34	5000	2	0	0	1	0	
	1	45	10000	3	0	0	0	1	
	2	50	15000	4	0	0	0	0	
	3	50	20000	4	1	0	0	0	
	4	67	30000	5	0	0	0	0	

In [155... imputation = pd.get_dummies(clean_data, columns=['Name'], dtype=int,drop_
imputation

0

Out[155		Domain	Age	Location	Salary	Ехр	Name_Jane	Name_Kim	Name_Mike
	0	Datascience	34	Mumbai	5000	2	0	0	1
	1	Testing	45	Bangalore	10000	3	0	0	0
	2	Dataanalyst	50	Bangalore	15000	4	0	0	0
	3	Analytics	50	Hyderbad	20000	4	1	0	0
	4	Statistics	67	Bangalore	30000	5	0	0	0
	5	NLP	55	Delhi	60000	10	0	1	0

In []:

In []:

about:srcdoc Page 16 of 16