```
import pandas as pd
pd. version
'2.2.2'
pip install --upgrade openpyxl
Requirement already satisfied: openpyxl in c:\users\bhavesh govind
bhork\anaconda3\lib\site-packages (3.1.5)
Requirement already satisfied: et-xmlfile in c:\users\bhavesh govind
bhork\anaconda3\lib\site-packages (from openpyxl) (1.1.0)
Note: you may need to restart the kernel to use updated packages.
emp = pd.read excel(r'C:\Users\Bhavesh Govind Bhork\Downloads\
Rawdata.XLSX')
emp
     Name
                   Domain
                                 Age
                                       Location
                                                  Salary
                                                               Exp
0
     Mike
            Datascience#$
                                         Mumbai
                                                  5^00#0
                                                                2+
                           34 years
                             45' yr
1
  Teddv^
                                                                <3
                  Testing
                                      Bangalore
                                                 10%%000
2
   Uma#r
           Dataanalyst^^#
                                                 1$5%000
                                 NaN
                                            NaN
                                                            4> yrs
              Ana^^lytics
3
                                       Hvderbad
                                                  2000^0
     Jane
                                 NaN
                                                               NaN
4
  Uttam*
               Statistics
                               67-yr
                                            NaN
                                                  30000 -
                                                           5+ year
5
                      NLP
                                55yr
                                                 6000^$0
      Kim
                                          Delhi
                                                               10+
id(emp)
1948209461248
emp.columns
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'],
dtype='object')
emp.shape
(6, 6)
emp
                   Domain
     Name
                                       Location
                                                  Salary
                                                               Exp
                                 Age
0
     Mike
            Datascience#$
                            34 years
                                         Mumbai
                                                  5^00#0
                                                                2+
  Teddy^
1
                  Testing
                              45' yr
                                      Bangalore
                                                 10%%000
                                                                <3
2
    Uma#r
           Dataanalyst^^#
                                 NaN
                                                 1$5%000
                                                            4> yrs
                                            NaN
              Ana^^lytics
3
                                                  2000^0
     Jane
                                 NaN
                                       Hyderbad
                                                               NaN
4
               Statistics
  Uttam*
                               67-yr
                                            NaN
                                                  30000-
                                                           5+ year
5
      Kim
                      NLP
                                55yr
                                          Delhi
                                                 6000^$0
                                                               10+
emp.head()
```

```
Domain
                                        Location
     Name
                                                    Salary
                                                                 Exp
                                  Age
0
     Mike
            Datascience#$
                             34 years
                                          Mumbai
                                                    5^00#0
                                                                  2+
1
   Teddy^
                   Testing
                               45' yr
                                       Bangalore
                                                   10%%000
                                                                  <3
2
    Uma#r
           Dataanalyst^^#
                                  NaN
                                                   1$5%000
                                              NaN
                                                              4> yrs
               Ana^^lytics
3
     Jane
                                  NaN
                                        Hyderbad
                                                    2000^0
                                                                 NaN
4
  Uttam*
                Statistics
                                67-yr
                                              NaN
                                                    30000-
                                                             5+ year
emp.tail()
                                      Location
     Name
                    Domain
                                Age
                                                  Salary
                                                               Exp
                            45' yr
1
   Teddv^
                   Testina
                                     Bangalore
                                                 10%%000
                                                                <3
2
           Dataanalyst^^#
    Uma#r
                                NaN
                                           NaN
                                                 1$5%000
                                                            4> yrs
               Ana^^lytics
3
                                NaN
                                      Hyderbad
                                                  2000^0
                                                               NaN
     Jane
4
                Statistics
                                           NaN
                                                  30000 -
   Uttam*
                              67-yr
                                                           5+ year
5
                                         Delhi
      Kim
                       NLP
                               55yr
                                                 6000^$0
                                                               10 +
emp.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #
     Column
                Non-Null Count
                                 Dtype
- - -
 0
     Name
                6 non-null
                                 object
 1
                6 non-null
     Domain
                                 object
 2
                4 non-null
                                 object
     Age
 3
     Location
                4 non-null
                                 object
 4
                6 non-null
     Salary
                                 object
 5
     Exp
                5 non-null
                                 object
dtypes: object(6)
memory usage: 420.0+ bytes
emp
     Name
                    Domain
                                  Age
                                        Location
                                                    Salary
                                                                 Exp
                                                    5^00#0
0
     Mike
            Datascience#$
                             34 years
                                          Mumbai
                                                                  2+
1
   Teddy^
                               45' yr
                                                                  <3
                   Testing
                                       Bangalore
                                                   10%%000
2
    Uma#r
           Dataanalyst^^#
                                  NaN
                                              NaN
                                                   1$5%000
                                                              4> vrs
3
               Ana^^lytics
     Jane
                                  NaN
                                        Hyderbad
                                                    2000^0
                                                                 NaN
4
   Uttam*
                Statistics
                                67-yr
                                              NaN
                                                    30000-
                                                             5+ year
5
      Kim
                       NLP
                                           Delhi
                                                   6000^$0
                                                                 10 +
                                 55yr
emp.isnull()
                                     Salary
    Name
          Domain
                     Age
                          Location
                                                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
```

```
emp.isna()
   Name Domain Age Location Salary Exp
                                       False
  False False False
                         False
                                False
1 False
         False False
                         False
                                False False
2 False False True
                                False
                         True
                                       False
3 False
          False True
                         False
                                False True
4 False
          False False
                         True
                                False False
5 False False False
                         False
                                False False
emp.isnull().sum()
          0
Name
Domain
          0
          2
Age
          2
Location
Salary
Exp
dtype: int64
emp['Name']
0
      Mike
1
    Teddv^
2
     Uma#r
3
      Jane
4
    Uttam*
5
       Kim
Name: Name, dtype: object
```

DATA CLEAING OR DATA CLEANSING

```
emp['Name']
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)
emp['Name'] #cleaned Name atribute
0
      Mike
1
     Teddy
2
      Umar
3
      Jane
4
     Uttam
```

```
Kim
Name: Name, dtype: object
emp['Domain']
0
      Datascience#$
1
            Testing
2
     Dataanalyst^^#
3
        Ana^^lytics
4
         Statistics
5
Name: Domain, dtype: object
emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
#formula for cleaning
emp['Domain'] #cleaned domain
0
     Datascience
1
         Testing
2
     Dataanalyst
3
       Analytics
4
      Statistics
5
             NLP
Name: Domain, dtype: object
emp
    Name
               Domain
                             Age
                                   Location
                                               Salary
                                                           Exp
0
    Mike
          Datascience
                       34 years
                                     Mumbai
                                               5^00#0
                                                            2+
                                  Bangalore
1
  Teddy
                          45' yr
                                                             <3
              Testing
                                              10%%000
2
                                                        4> yrs
   Umar
          Dataanalyst
                             NaN
                                         NaN
                                              1$5%000
3
            Analytics
                             NaN
                                   Hyderbad
                                               2000^0
                                                           NaN
    Jane
4
  Uttam
           Statistics
                           67-yr
                                         NaN
                                               30000-
                                                       5+ year
5
     Kim
                  NLP
                                      Delhi 6000^$0
                            55yr
                                                            10 +
emp['Age']
0
     34 years
       45' yr
1
2
          NaN
3
          NaN
4
        67-yr
5
         55yr
Name: Age, dtype: object
emp['Age'] = emp['Age'].str.replace(r'\W','',regex=True) #by this
irregular text removed but not words we need to remove text totally
emp['Age']
```

```
0
     34years
1
        45yr
2
         NaN
3
         NaN
4
        67yr
5
        55yr
Name: Age, dtype: object
emp['Age']=emp['Age'].str.extract(r'(\d+)') #by this code we can
remove the text totally from our numerical data
emp['Age'] #sucesfully removed text data
0
      34
1
      45
2
     NaN
3
     NaN
4
      67
5
      55
Name: Age, dtype: object
emp['Location'] = emp['Location'].str.replace(r'\W','',regex=True)
emp['Location']
0
        Mumbai
1
     Bangalore
2
           NaN
3
      Hyderbad
4
           NaN
5
         Delhi
Name: Location, dtype: object
emp
    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
            Analytics
                              Hyderbad
                                         2000^0
                                                     NaN
    Jane
                       NaN
4
  Uttam
           Statistics
                         67
                                   NaN
                                         30000 -
                                                 5+ year
                         55
     Kim
                  NLP
                                 Delhi
                                        6000^$0
                                                     10 +
emp['Salary'] = emp['Salary'].str.replace(r'\W','',regex=True)
emp['Salary']
0
      5000
1
     10000
2
     15000
3
     20000
4
     30000
```

```
60000
Name: Salary, dtype: object
emp.head()
    Name
               Domain
                       Age
                              Location Salary
                                                    Exp
    Mike
                        34
                                         5000
                                                     2+
0
          Datascience
                                Mumbai
                        45
                                        10000
1
  Teddy
              Testing
                             Bangalore
                                                     <3
2
    Umar
          Dataanalyst
                       NaN
                                   NaN
                                        15000
                                                 4> yrs
3
                              Hyderbad
    Jane
            Analytics
                       NaN
                                        20000
                                                   NaN
  Uttam
           Statistics
                        67
                                   NaN
                                       30000
                                               5+ year
emp['Exp']=emp['Exp'].str.extract(r'(\d+)') #by this code we can
remove the text totally from our numerical data
emp['Exp']
       2
0
1
       3
2
       4
3
     NaN
4
       5
5
      10
Name: Exp, dtype: object
emp
    Name
               Domain Age
                              Location Salary
                                               Exp
0
    Mike
          Datascience
                        34
                                Mumbai
                                         5000
                                                 2
                                                 3
1
  Teddy
              Testing
                        45
                             Bangalore
                                        10000
2
                                        15000
                                                 4
    Umar
          Dataanalyst
                       NaN
                                   NaN
3
    Jane
            Analytics
                       NaN
                              Hyderbad
                                        20000
                                               NaN
4
                                   NaN
                                        30000
                                                 5
  Uttam
           Statistics
                        67
5
     Kim
                  NLP
                        55
                                 Delhi
                                        60000
                                                 10
clean data = emp.copy()
clean_data
    Name
               Domain
                       Age
                              Location Salary
                                               Exp
    Mike
                                         5000
0
        Datascience
                        34
                                Mumbai
                                                 2
                                                 3
  Teddv
                        45 Bangalore
                                        10000
1
              Testing
2
   Umar
          Dataanalyst
                                   NaN
                                       15000
                                                 4
                       NaN
3
    Jane
            Analytics
                       NaN
                              Hyderbad
                                        20000
                                               NaN
  Uttam
4
           Statistics
                        67
                                   NaN
                                        30000
                                                 5
5
     Kim
                  NLP
                        55
                                 Delhi 60000
                                                 10
```

LETS APPLY EDA TECHNIQUE:-

```
clean_data
```

```
Name
                              Location Salary
               Domain
                       Age
                                                Exp
0
    Mike
          Datascience
                                Mumbai
                                         5000
                         34
                                                  3
1
  Teddy
              Testing
                        45
                             Bangalore
                                        10000
2
                                                  4
   Umar
          Dataanalyst
                                   NaN
                                        15000
                        NaN
3
    Jane
            Analytics
                        NaN
                              Hyderbad
                                        20000
                                                NaN
4
  Uttam
           Statistics
                         67
                                   NaN
                                        30000
                                                  5
5
     Kim
                         55
                                 Delhi 60000
                  NLP
                                                 10
clean data.isnull().sum()
Name
            0
            0
Domain
            2
Age
Location
            2
Salary
            0
Exp
            1
dtype: int64
clean_data['Age']
0
      34
1
      45
2
     NaN
3
     NaN
4
      67
5
      55
Name: Age, dtype: object
import numpy as np
clean data['Age']=
clean data['Age'].fillna(np.mean(pd.to numeric(clean data['Age'])))
clean data['Age']
        34
0
1
        45
2
     50.25
3
     50.25
4
        67
5
        55
Name: Age, dtype: object
clean data['Exp']=clean data['Exp'].fillna(np.mean(pd.to numeric(clean
data['Exp'])))
clean data['Exp']
       2
0
1
       3
2
       4
3
     4.8
```

```
4
       5
5
      10
Name: Exp, dtype: object
clean data
                          Age
                                Location Salary
    Name
               Domain
                                                  Exp
0
    Mike
          Datascience
                           34
                                  Mumbai
                                            5000
                                                    2
                                                    3
1
  Teddy
              Testing
                           45
                               Bangalore
                                          10000
2
                                                    4
   Umar
          Dataanalyst 50.25
                                     NaN
                                          15000
3
            Analytics 50.25
    Jane
                                Hyderbad
                                          20000
                                                  4.8
                                                    5
4
  Uttam
           Statistics
                           67
                                     NaN
                                          30000
5
                           55
     Kim
                  NLP
                                   Delhi
                                          60000
                                                   10
clean data['Location']
0
        Mumbai
1
     Bangalore
2
           NaN
3
      Hyderbad
4
           NaN
5
         Delhi
Name: Location, dtype: object
clean data['Location'] =
clean_data['Location'].fillna(clean_data['Location'].mode()[0])
clean data['Location']
0
        Mumbai
1
     Bangalore
2
     Bangalore
3
      Hyderbad
4
     Bangalore
5
         Delhi
Name: Location, dtype: object
clean data
                                Location Salary
    Name
               Domain
                          Age
                                                  Exp
0
    Mike
          Datascience
                           34
                                  Mumbai
                                            5000
                                                    2
                           45
                                          10000
                                                    3
1
  Teddy
              Testing
                               Bangalore
2
                               Bangalore
    Umar
          Dataanalyst 50.25
                                          15000
                                                    4
3
            Analytics 50.25
                                                  4.8
    Jane
                               Hyderbad
                                          20000
                                                    5
4
  Uttam
           Statistics
                           67
                               Bangalore
                                          30000
5
     Kim
                  NLP
                           55
                                   Delhi
                                          60000
                                                   10
clean data.isnull()
    Name
                          Location
          Domain
                     Age
                                    Salary
                                               Exp
                                     False
   False
           False
                  False
                             False
                                             False
   False
           False False
                             False
                                     False
                                             False
```

```
False
           False False
                                           False
                            False
                                    False
3
   False
           False False
                            False
                                    False
                                           False
4
   False
           False False
                            False
                                    False
                                           False
5 False
           False False
                            False
                                    False False
emp.info
                                               Domain Age
<bound method DataFrame.info of</pre>
                                    Name
                                                              Location
Salary
        Exp
                                        5000
    Mike Datascience
                        34
                               Mumbai
                                                 2
1
   Teddv
              Testina
                        45
                            Bangalore
                                       10000
                                                 3
2
   Umar Dataanalyst
                       NaN
                                  NaN
                                       15000
                                                 4
3
          Analytics
                             Hyderbad
                                       20000
    Jane
                       NaN
                                              NaN
4
           Statistics
                                                 5
   Uttam
                        67
                                  NaN
                                      30000
5
                        55
                                Delhi
     Kim
                  NLP
                                      60000
                                                10>
                           #system bydefault capture it as a object
clean data.info()
but we have to convert it as integer
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
               Non-Null Count
     Column
                               Dtype
0
               6 non-null
                               object
     Name
1
     Domain
               6 non-null
                               object
 2
               6 non-null
                               object
     Age
 3
     Location 6 non-null
                               object
 4
     Salary
               6 non-null
                               object
 5
     Exp
               6 non-null
                               object
dtypes: object(6)
memory usage: 420.0+ bytes
clean data['Age'] = clean data['Age'].astype(int)
                  #hence we sucessfully converted object into
clean data.info()
integer
<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
                               obiect
 1
               6 non-null
     Domain
                               object
 2
               6 non-null
                               int32
     Age
 3
     Location 6 non-null
                               object
4
    Salary
               6 non-null
                               object
 5
     Exp
               6 non-null
                               object
dtypes: int32(1), object(5)
memory usage: 396.0+ bytes
```

```
clean data['Salary'] = clean data['Salary'].astype(int)
clean data.info()
                  #Here we converted salary into int type
<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
               6 non-null
 1
     Domain
                               obiect
2
     Age
               6 non-null
                               int32
 3
     Location 6 non-null
                               object
 4
               6 non-null
                               int32
     Salary
 5
     Exp
               6 non-null
                               object
dtypes: int32(2), object(4)
memory usage: 372.0+ bytes
clean data
    Name
               Domain Age
                             Location Salary
                                               Exp
0
    Mike Datascience
                        34
                               Mumbai
                                         5000
                                                  2
                                                  3
1
  Teddv
              Testina
                        45
                            Bandalore
                                        10000
2
   Umar
         Dataanalyst
                        50
                            Bangalore
                                        15000
                                                  4
3
            Analytics
                             Hyderbad
    Jane
                        50
                                        20000
                                               4.8
4
           Statistics
                                                  5
  Uttam
                            Bangalore
                                        30000
                        67
5
     Kim
                  NLP
                        55
                                Delhi
                                        60000
                                                10
clean data['Salary'] = clean data['Salary'].astype(int)
clean data['Exp']
                   = clean data['Exp'].astype(int)
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
               6 non-null
                               object
     Name
1
               6 non-null
     Domain
                               obiect
 2
     Age
               6 non-null
                               int32
 3
                               object
    Location 6 non-null
4
     Salary
               6 non-null
                               int32
 5
     Exp
               6 non-null
                               int32
dtypes: int32(3), object(3)
memory usage: 348.0+ bytes
clean data['Name'] = clean data['Name'].astype('category')
clean_data['Location'] = clean_data['Location'].astype('category')
clean data['Domain'] = clean data['Domain'].astype('category')
```

```
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
     Age
                               int32
 3
    Location 6 non-null
                               category
4
               6 non-null
     Salary
                               int32
 5
               6 non-null
                               int32
     Exp
dtypes: category(3), int32(3)
memory usage: 866.0 bytes
clean data
                             Location Salary
    Name
               Domain Age
                                               Exp
0
    Mike Datascience
                        34
                               Mumbai
                                         5000
                                                 2
                                                 3
1
  Teddy
              Testing
                        45
                            Bangalore
                                        10000
2
                            Bangalore
                                                 4
   Umar Dataanalyst
                        50
                                        15000
3
          Analytics
                        50
                             Hyderbad
                                        20000
                                                 4
    Jane
                                                 5
4
  Uttam
           Statistics
                        67
                            Bangalore
                                        30000
5
     Kim
                  NLP
                        55
                                Delhi
                                        60000
                                                10
clean data.to csv('clean data.csv')
PermissionError
                                          Traceback (most recent call
last)
Cell In[172], line 1
----> 1 clean data.to csv('clean data.csv')
File c:\Users\Bhavesh Govind Bhork\anaconda3\Lib\site-packages\pandas\
util\ decorators.py:333, in
deprecate nonkeyword arguments.<locals>.decorate.<locals>.wrapper(*arg
s, **kwargs)
    327 if len(args) > num allow args:
    328
           warnings.warn(
    329
msg.format(arguments= format argument list(allow args)),
                FutureWarning,
    330
    331
                stacklevel=find_stack_level(),
    332
--> 333 return func(*args, **kwargs)
File c:\Users\Bhavesh Govind Bhork\anaconda3\Lib\site-packages\pandas\
core\generic.py:3967, in NDFrame.to csv(self, path or buf, sep,
na rep, float format, columns, header, index, index label, mode,
```

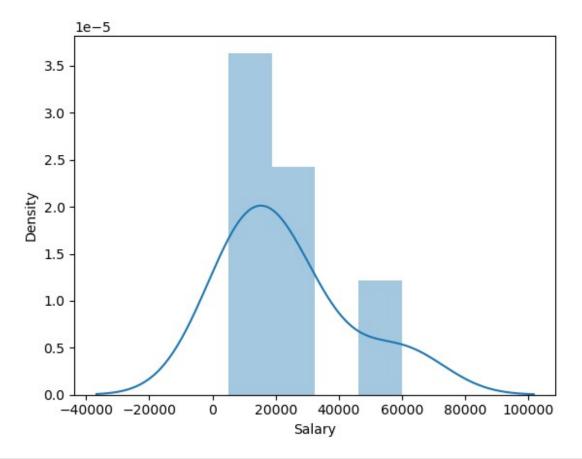
```
encoding, compression, quoting, quotechar, lineterminator, chunksize,
date format, doublequote, escapechar, decimal, errors,
storage options)
   3956 df = self if isinstance(self, ABCDataFrame) else
self.to frame()
   3958 formatter = DataFrameFormatter(
   3959
            frame=df,
   3960
            header=header,
           3964 decimal=decimal,
   (\ldots)
   3965 )
-> 3967 return DataFrameRenderer(formatter).to csv(
   3968
            path or buf,
            lineterminator=lineterminator,
   3969
   3970
            sep=sep,
   3971
            encoding=encoding,
   3972
            errors=errors,
   3973
            compression=compression,
   3974
            quoting=quoting,
   3975
            columns=columns,
   3976
            index label=index label,
            mode=mode,
   3977
   3978
            chunksize=chunksize,
            quotechar=quotechar,
   3979
   3980
            date format=date format,
            doublequote=doublequote,
   3981
   3982
            escapechar=escapechar,
   3983
            storage options=storage options,
   3984 )
File c:\Users\Bhavesh Govind Bhork\anaconda3\Lib\site-packages\pandas\
io\formats\format.py:1014, in DataFrameRenderer.to csv(self,
path or buf, encoding, sep, columns, index label, mode, compression,
quoting, quotechar, lineterminator, chunksize, date format,
doublequote, escapechar, errors, storage options)
            created buffer = False
    995 csv formatter = CSVFormatter(
    996
            path or buf=path or buf,
    997
            lineterminator=lineterminator,
   (\ldots)
           1012 formatter=self.fmt,
   1013 )
-> 1014 csv formatter.save()
   1016 if created buffer:
            assert isinstance(path or buf, StringIO)
File c:\Users\Bhavesh Govind Bhork\anaconda3\Lib\site-packages\pandas\
io\formats\csvs.py:251, in CSVFormatter.save(self)
    247 """
    248 Create the writer & save.
    249 """
```

```
250 # apply compression and byte/text conversion
--> 251 with get handle(
    252
            self.filepath or buffer,
    253
            self.mode.
    254
            encoding=self.encoding,
    255
            errors=self.errors,
    256
            compression=self.compression,
    257
            storage options=self.storage options,
    258 ) as handles:
    259
            # Note: self.encoding is irrelevant here
    260
            self.writer = csvlib.writer(
    261
                handles.handle,
                lineterminator=self.lineterminator,
    262
   (\ldots)
            267
                        quotechar=self.quotechar,
    268
            self. save()
    270
File c:\Users\Bhavesh Govind Bhork\anaconda3\Lib\site-packages\pandas\
io\common.py:873, in get handle(path or buf, mode, encoding,
compression, memory map, is text, errors, storage options)
    868 elif isinstance(handle, str):
    869
            # Check whether the filename is to be opened in binary
mode.
    870
            # Binary mode does not support 'encoding' and 'newline'.
            if ioargs.encoding and "b" not in ioargs.mode:
    871
    872
                # Encoding
--> 873
                handle = open(
    874
                    handle,
    875
                    ioargs.mode,
    876
                    encoding=ioargs.encoding,
    877
                    errors=errors,
    878
                    newline="",
    879
    880
            else:
    881
                # Binary mode
    882
                handle = open(handle, ioargs.mode)
PermissionError: [Errno 13] Permission denied: 'clean data.csv'
import os
os.getcwd()
'c:\\Users\\Bhavesh Govind Bhork\\Downloads'
```

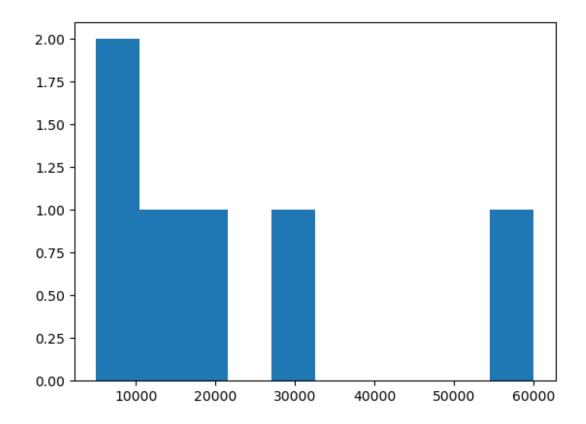
^ we cleaned the RAW DATA into CLEAN DATA and also imported that data into our system

```
import warnings
warnings.filterwarnings('ignore')
clean_data['Salary']
```

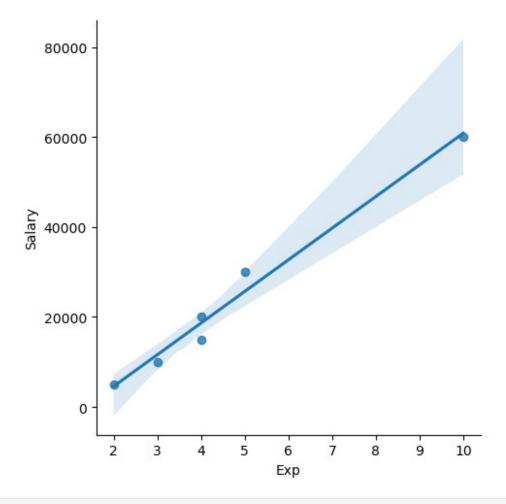
```
0    5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: int32
import seaborn as sns #we have to import library first to visualize the data
visl = sns.distplot(clean_data['Salary'])
```



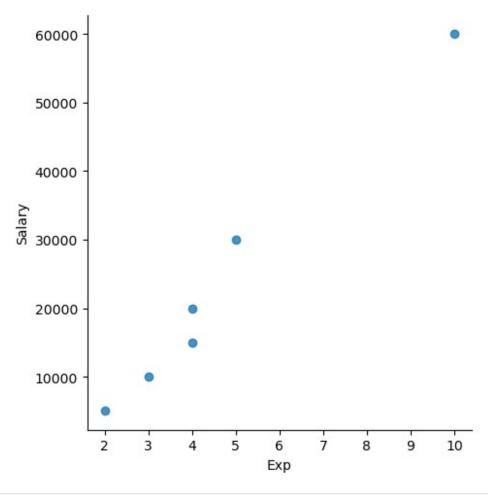
```
import matplotlib.pyplot as plt
vis2 = plt.hist(clean_data['Salary'])
```



vis4 = sns.lmplot(data=clean_data, x='Exp', y='Salary') #linear model
plot



vis5 = sns.lmplot(data=clean_data, x='Exp', y='Salary', fit_reg=False)



```
clean_data[:]
    Name
                Domain
                        Age
                               Location
                                          Salary
                                                  Exp
    Mike
          Datascience
                         34
                                 Mumbai
                                            5000
                                                     2
                                                    3
1
  Teddy
               Testing
                         45
                              Bangalore
                                           10000
2
                                                    4
    Umar
          Dataanalyst
                         50
                              Bangalore
                                           15000
3
    Jane
            Analytics
                         50
                               Hyderbad
                                           20000
                                                    4
4
                                                    5
  Uttam
           Statistics
                         67
                              Bangalore
                                           30000
5
     Kim
                   NLP
                         55
                                  Delhi
                                           60000
                                                   10
x_4 = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
x_4 #these all are independent vaariables
    Name
                        Age
                               Location
                Domain
                                         Exp
    Mike
0
          Datascience
                         34
                                 Mumbai
                                            2
1
   Teddy
              Testing
                              Bangalore
                                            3
                         45
2
    Umar
          Dataanalyst
                              Bangalore
                                            4
                         50
3
                               Hyderbad
                                            4
    Jane
            Analytics
                         50
4
   Uttam
           Statistics
                         67
                              Bangalore
                                            5
5
     Kim
                   NLP
                         55
                                  Delhi
                                           10
```

```
y dv = clean data[['Salary']]
y dv #this is dependent variable
   Salary
     5000
0
1
    10000
2
    15000
3
    20000
4
    30000
5
    60000
emp
                                                  Exp
    Name
                Domain
                         Age
                                Location Salary
    Mike
           Datascience
                          34
                                  Mumbai
                                            5000
                                                     2
0
                                                     3
1
                                           10000
   Teddy
               Testing
                          45
                               Bangalore
2
                                                     4
    Umar
           Dataanalyst
                                           15000
                         NaN
                                     NaN
3
                                Hyderbad
    Jane
             Analytics
                         NaN
                                           20000
                                                  NaN
4
  Uttam
            Statistics
                          67
                                     NaN
                                           30000
                                                     5
5
     Kim
                    NLP
                          55
                                   Delhi
                                           60000
                                                    10
clean_data
    Name
                         Age
                                Location
                                           Salary
                Domain
                                                    Exp
0
    Mike
           Datascience
                          34
                                  Mumbai
                                             5000
                                                      2
                                                      3
1
   Teddy
                          45
                               Bangalore
                                            10000
               Testing
2
                               Bangalore
                                                      4
    Umar
           Dataanalyst
                          50
                                            15000
3
                                                      4
    Jane
             Analytics
                          50
                                Hyderbad
                                            20000
                                                      5
4
   Uttam
            Statistics
                          67
                               Bangalore
                                            30000
5
     Kim
                   NLP
                          55
                                   Delhi
                                            60000
                                                     10
x 4
                                           Exp
    Name
                Domain
                         Age
                                Location
    Mike
                                  Mumbai
0
           Datascience
                          34
                                             2
1
   Teddy
               Testing
                          45
                               Bangalore
                                             3
2
                                             4
    Umar
           Dataanalyst
                          50
                               Bangalore
3
    Jane
             Analytics
                          50
                                Hyderbad
                                             4
4
   Uttam
            Statistics
                          67
                               Bangalore
                                             5
5
     Kim
                    NLP
                          55
                                   Delhi
                                            10
y_dv
   Salary
0
     5000
1
    10000
2
    15000
3
    20000
4
    30000
5
    60000
```

cl	ean_da	ta						
0 1 2 3 4 5	Name Mike Teddy Umar Jane Uttam Kim	Datas T Dataa Ana Stat	Domain cience esting nalyst lytics istics NLP	34 Mu 45 Banga 50 Banga 50 Hyde 67 Banga	mbai ! lore 1! lore 1! rbad 20 lore 3	lary Exp 5000 2 0000 3 5000 4 0000 4 0000 5		
im	putati	on = pd	.get_du	ımmies(clean	_data, d	type= <mark>int</mark>)		
im	putati	on						
Na	Age me_Uma	Salary r \	Exp N	lame_Jane N	ame_Kim	Name_Mike	Name_Teddy	
0	3 4	5000	2	0	0	1	0	
1	45	10000	3	0	0	0	1	
0	50	15000	4	0	0	0	0	
1 3	50	20000	4	1	0	0	0	
0 4	67	30000	5	0	0	0	0	
0 5	55	60000	10	0	1	0	0	
0				·	_			
Dο		Uttam atascie		_Analytics	Domain_D	ataanalyst		
0	a±11_b	0	nee (0		0		
1		0		0		0		
0 2		0		0		1		
0 3 0		0		1		0		
0 4		1		0		Θ		
0		0		0		0		
5 0		U		U		U		
				Statistics	Domain_	Testing		
Lo 0	cation	_Bangal 0	ore \	0		0		0
1		0		0		1		1
				•		_		

2	0	0	0	1
3	0	0	0	0
4	0	1	0	1
5	1	0	0	Θ
_				_
0 1 2 3 4 5	- 0 0 0 0 0 1	ntion_Hyderbad Loc 0 0 0 1 0	ation_Mumbai 1 0 0 0 0	
imputat	ion.columns			
	'Name_Teddy', 'N 'Domain_Dataanal 'Domain_Statisti	'Exp', 'Name_Jane lame_Umar', 'Name_U yst', 'Domain_Data .cs', 'Domain_Testi , 'Location_Hyderb	ttam', 'Domain_An science', 'Domain ng', 'Location_Ba	alytics', _NLP', ngalore',
len(imp	utation.columns)			
19				

in this project we have covered:-

- . raw data with lot of regx, missing, uncleaned regex, clean
- . filled missing numerical and category
- . clean data set
- . outlier treatment, univarient, bivarient, corelation
- . split the data into x_4 and y_dv
- . impute category data to numerical
- . EDA done