Data Professional Salary Prediction

Said KHALID

June 26, 2024

Contents

1 Introduction

This project explores predicting salaries of data professionals using advanced machine learning techniques. It aims to understand the most influential factors affecting salaries in this field using regression models.

[49]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.linear_model import Lasso,Ridge
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
from sklearn.metrics import r2_score
from sklearn.preprocessing import  StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import  mean_absolute_error
from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.tree import DecisionTreeClassifier,plot_tree
from sklearn.ensemble import  RandomForestClassifier
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.cluster import KMeans,AgglomerativeClustering
from sklearn.metrics import silhouette_score
import scipy.cluster.hierarchy as sch
from sklearn.preprocessing import  LabelEncoder
from sklearn.decomposition import  PCA
from sklearn.ensemble import RandomForestRegressor
from sklearn.feature_selection import VarianceThreshold
from statsmodels.tsa.seasonal import seasonal_decompose
from dateutil.parser import  parse
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.statespace.sarimax import SARIMAX
import statsmodels.api as sm
from imblearn.over_sampling import  SMOTE
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
```

2 Load the Dataset

Initially, we must load the dataset and examine its structure

[3]:
```
df=pd.read_csv('Salary Prediction of Data Professions.csv')
df
```

[3]:

	FIRST NAME	LAST NAME	SEX	DOJ	CURRENT DATE	DESIGNATION
0	TOMASA	ARMEN	F	5-18-2014	01-07-2016	Analyst
1	ANNIE	NaN	F	NaN	01-07-2016	Associate
2	OLIVE	ANCY	F	7-28-2014	01-07-2016	Analyst
3	CHERRY	AQUILAR	F	04-03-2013	01-07-2016	Analyst
4	LEON	ABOULAHOUD	M	11-20-2014	01-07-2016	Analyst
...	...	...	..	...	...	...
2634	KATHERINE	ALSDON	F	6-28-2011	01-07-2016	Senior Manager
2635	LOUISE	ALTARAS	F	1-14-2014	01-07-2016	Analyst
2636	RENEE	ALVINO	F	1-23-2014	01-07-2016	Analyst
2637	TERI	ANASTASIO	F	3-17-2014	01-07-2016	Analyst
2638	GREGORY	ABARCA	M	9-18-2014	01-07-2016	Analyst

	AGE	SALARY	UNIT	LEAVES USED	LEAVES REMAINING	RATINGS
0	21.0	44570	Finance	24.0	6.0	2.0
1	NaN	89207	Web	NaN	13.0	NaN
2	21.0	40955	Finance	23.0	7.0	3.0
3	22.0	45550	IT	22.0	8.0	3.0
4	NaN	43161	Operations	27.0	3.0	NaN
...	...	...	...	...	...	...
2634	36.0	185977	Management	15.0	15.0	5.0
2635	23.0	45758	IT	17.0	13.0	2.0
2636	21.0	47315	Web	29.0	1.0	5.0
2637	24.0	45172	Web	23.0	7.0	3.0
2638	24.0	49176	Marketing	17.0	13.0	2.0

	PAST EXP
0	0
1	7
2	0
3	0
4	3
...	...
2634	10
2635	0
2636	0
2637	1
2638	2

[2639 rows x 13 columns]

[4]:
```
df.head(5)
```

[4]:

	FIRST NAME	LAST NAME	SEX	DOJ	CURRENT DATE	DESIGNATION	AGE
0	TOMASA	ARMEN	F	5-18-2014	01-07-2016	Analyst	21.0
1	ANNIE	NaN	F	NaN	01-07-2016	Associate	NaN
2	OLIVE	ANCY	F	7-28-2014	01-07-2016	Analyst	21.0
3	CHERRY	AQUILAR	F	04-03-2013	01-07-2016	Analyst	22.0
4	LEON	ABOULAHOUD	M	11-20-2014	01-07-2016	Analyst	NaN

	SALARY	UNIT	LEAVES USED	LEAVES REMAINING	RATINGS	PAST EXP
0	44570	Finance	24.0	6.0	2.0	0
1	89207	Web	NaN	13.0	NaN	7
2	40955	Finance	23.0	7.0	3.0	0
3	45550	IT	22.0	8.0	3.0	0
4	43161	Operations	27.0	3.0	NaN	3

[5]:
```
df.tail(5)
```

[5]:

	FIRST NAME	LAST NAME	SEX	DOJ	CURRENT DATE	DESIGNATION	AGE
2634	KATHERINE	ALSDON	F	6-28-2011	01-07-2016	Senior Manager	36.0
2635	LOUISE	ALTARAS	F	1-14-2014	01-07-2016	Analyst	23.0
2636	RENEE	ALVINO	F	1-23-2014	01-07-2016	Analyst	21.0
2637	TERI	ANASTASIO	F	3-17-2014	01-07-2016	Analyst	24.0
2638	GREGORY	ABARCA	M	9-18-2014	01-07-2016	Analyst	24.0

	SALARY	UNIT	LEAVES USED	LEAVES REMAINING	RATINGS	PAST EXP
2634	185977	Management	15.0	15.0	5.0	10
2635	45758	IT	17.0	13.0	2.0	0
2636	47315	Web	29.0	1.0	5.0	0
2637	45172	Web	23.0	7.0	3.0	1
2638	49176	Marketing	17.0	13.0	2.0	2

[6]:
```
df.sample(5)
```

[6]:

	FIRST NAME	LAST NAME	SEX	DOJ	CURRENT DATE	DESIGNATION	AGE
1232	LONNIE	ACERRA	M	12-18-2014	01-07-2016	Analyst	21.0
478	REBECCA	ALPAUGH	F	02-02-2014	01-07-2016	Analyst	22.0
2530	VILMA	APRIGLIANO	F	1-18-2014	01-07-2016	Analyst	23.0
197	CARIDAD	ARMWOOD	F	06-11-2014	01-07-2016	Analyst	23.0
2555	JEWELL	ANGELI	F	3-27-2012	01-07-2016	Manager	32.0

	SALARY	UNIT	LEAVES USED	LEAVES REMAINING	RATINGS	PAST EXP
1232	48626	Marketing	27.0	3.0	2.0	0
478	49744	Web	28.0	2.0	2.0	0
2530	46984	Web	24.0	6.0	4.0	0
197	45022	Finance	23.0	7.0	3.0	0
2555	112363	Finance	18.0	12.0	3.0	7

[7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2639 entries, 0 to 2638
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   FIRST NAME        2639 non-null   object
 1   LAST NAME         2637 non-null   object
 2   SEX               2639 non-null   object
 3   DOJ               2638 non-null   object
 4   CURRENT DATE      2639 non-null   object
 5   DESIGNATION       2639 non-null   object
 6   AGE               2636 non-null   float64
 7   SALARY            2639 non-null   int64
 8   UNIT              2639 non-null   object
 9   LEAVES USED       2636 non-null   float64
 10  LEAVES REMAINING  2637 non-null   float64
 11  RATINGS           2637 non-null   float64
 12  PAST EXP          2639 non-null   int64
dtypes: float64(4), int64(2), object(7)
memory usage: 268.2+ KB
```

[8]:
```
df.columns=df.columns.str.lower()
df
```

[8]:

	first name	last name	sex	doj	current date	designation
0	TOMASA	ARMEN	F	5-18-2014	01-07-2016	Analyst
1	ANNIE	NaN	F	NaN	01-07-2016	Associate
2	OLIVE	ANCY	F	7-28-2014	01-07-2016	Analyst
3	CHERRY	AQUILAR	F	04-03-2013	01-07-2016	Analyst
4	LEON	ABOULAHOUD	M	11-20-2014	01-07-2016	Analyst
...	...	...	..	...	...	...
2634	KATHERINE	ALSDON	F	6-28-2011	01-07-2016	Senior Manager
2635	LOUISE	ALTARAS	F	1-14-2014	01-07-2016	Analyst
2636	RENEE	ALVINO	F	1-23-2014	01-07-2016	Analyst
2637	TERI	ANASTASIO	F	3-17-2014	01-07-2016	Analyst
2638	GREGORY	ABARCA	M	9-18-2014	01-07-2016	Analyst

	age	salary	unit	leaves used	leaves remaining	ratings
0	21.0	44570	Finance	24.0	6.0	2.0
1	NaN	89207	Web	NaN	13.0	NaN
2	21.0	40955	Finance	23.0	7.0	3.0
3	22.0	45550	IT	22.0	8.0	3.0
4	NaN	43161	Operations	27.0	3.0	NaN
...	...	...	...	...	...	...
2634	36.0	185977	Management	15.0	15.0	5.0
2635	23.0	45758	IT	17.0	13.0	2.0
2636	21.0	47315	Web	29.0	1.0	5.0

```
2637  24.0   45172         Web          23.0               7.0      3.0
2638  24.0   49176   Marketing          17.0              13.0      2.0

      past exp
0            0
1            7
2            0
3            0
4            3
...        ...
2634        10
2635         0
2636         0
2637         1
2638         2

[2639 rows x 13 columns]
```

3 Summary Statistics

We utilize summary statistics to gain an overview of the numerical features

[9]:
```
df.describe()
```

[9]:
```
                age         salary  leaves used  leaves remaining      ratings  \
count  2636.000000    2639.000000  2636.000000       2637.000000  2637.000000
mean     24.756449   58136.678287    22.501517          7.503223     3.486159
std       3.908228   36876.956944     4.604469          4.603193     1.114933
min      21.000000   40001.000000    15.000000          0.000000     2.000000
25%      22.000000   43418.000000    19.000000          4.000000     2.000000
50%      24.000000   46781.000000    22.000000          8.000000     3.000000
75%      25.000000   51401.500000    26.000000         11.000000     4.000000
max      45.000000  388112.000000    30.000000         15.000000     5.000000

          past exp
count  2639.000000
mean      1.566881
std       2.728416
min       0.000000
25%       0.000000
50%       1.000000
75%       2.000000
max      23.000000
```

[10]:
```
columns_to_select = ['age', 'salary', 'leaves used', 'leaves remaining',␣
 ↪'ratings', 'past exp']
df1=df[columns_to_select]
```

```
df1.corr()
```

[10]:

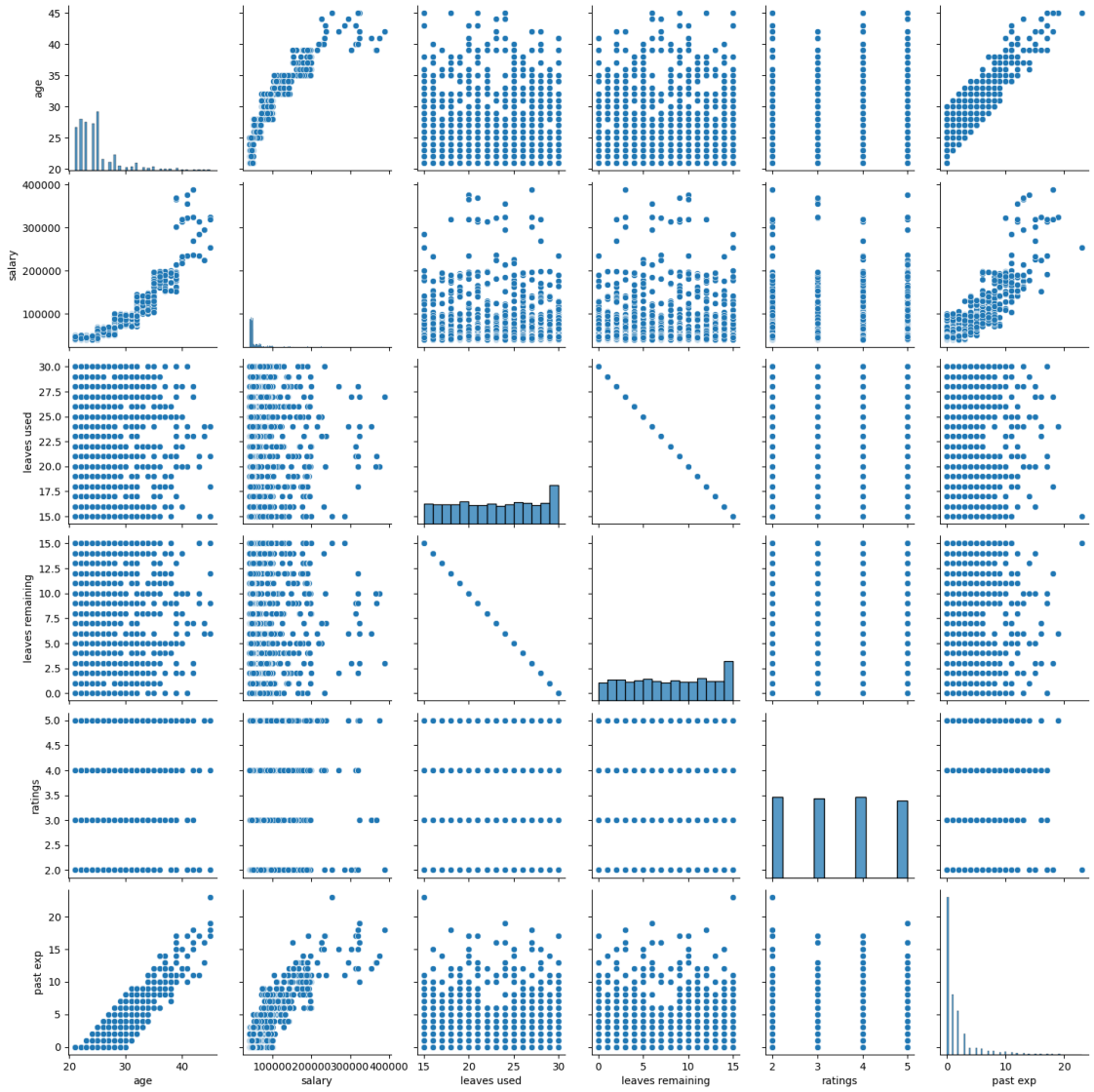
	age	salary	leaves used	leaves remaining	ratings	past exp
age	1.000000	0.872213	0.007825	-0.006515	0.036801	0.903926
salary	0.872213	1.000000	0.006498	-0.005422	0.020248	0.854046
leaves used	0.007825	0.006498	1.000000	-1.000000	0.002200	0.008601
leaves remaining	-0.006515	-0.005422	-1.000000	1.000000	-0.003415	-0.006728
ratings	0.036801	0.020248	0.002200	-0.003415	1.000000	0.040123
past exp	0.903926	0.854046	0.008601	-0.006728	0.040123	1.000000

4 Data Visualization

We will employ various plots to visualize the distribution of salaries and investigate relationships between SALARY and other features.

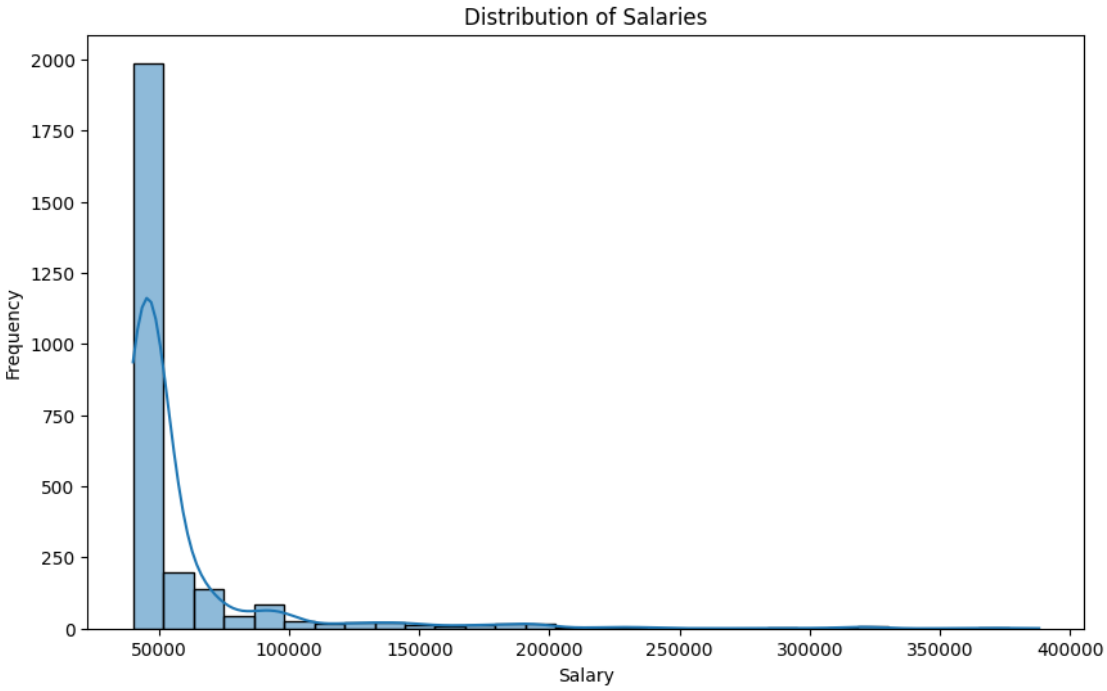
[11]:
```
sns.pairplot(df)
```

[11]:
```
<seaborn.axisgrid.PairGrid at 0x1a0c1a3a290>
```



4.1 Distribution of Salaries

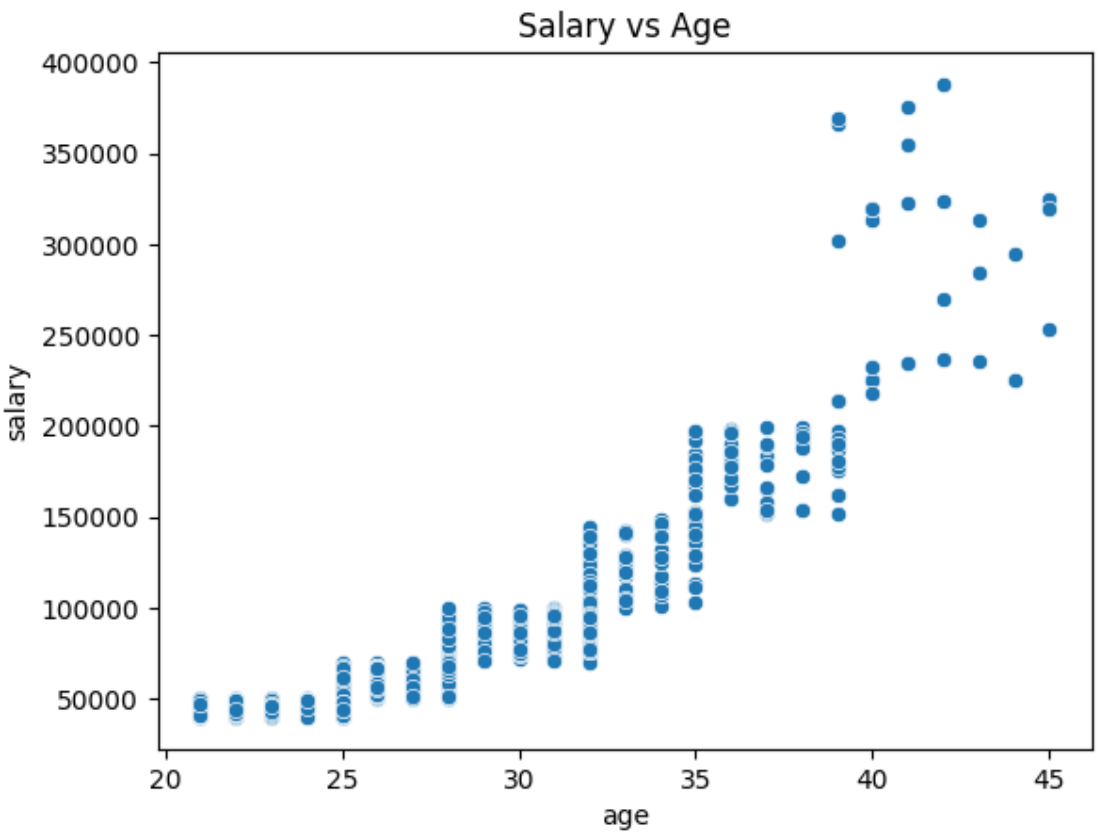
[12]:
```
plt.figure(figsize=(10,6))
sns.histplot(df['salary'],bins=30,kde=True)
plt.title('Distribution of Salaries')
plt.xlabel('Salary')
plt.ylabel('Frequency')
plt.show()
```



4.2 Relationshep between salary and age

[13]:
```
sns.scatterplot(x='age',y='salary',data=df)
plt.title('Salary vs Age')
```

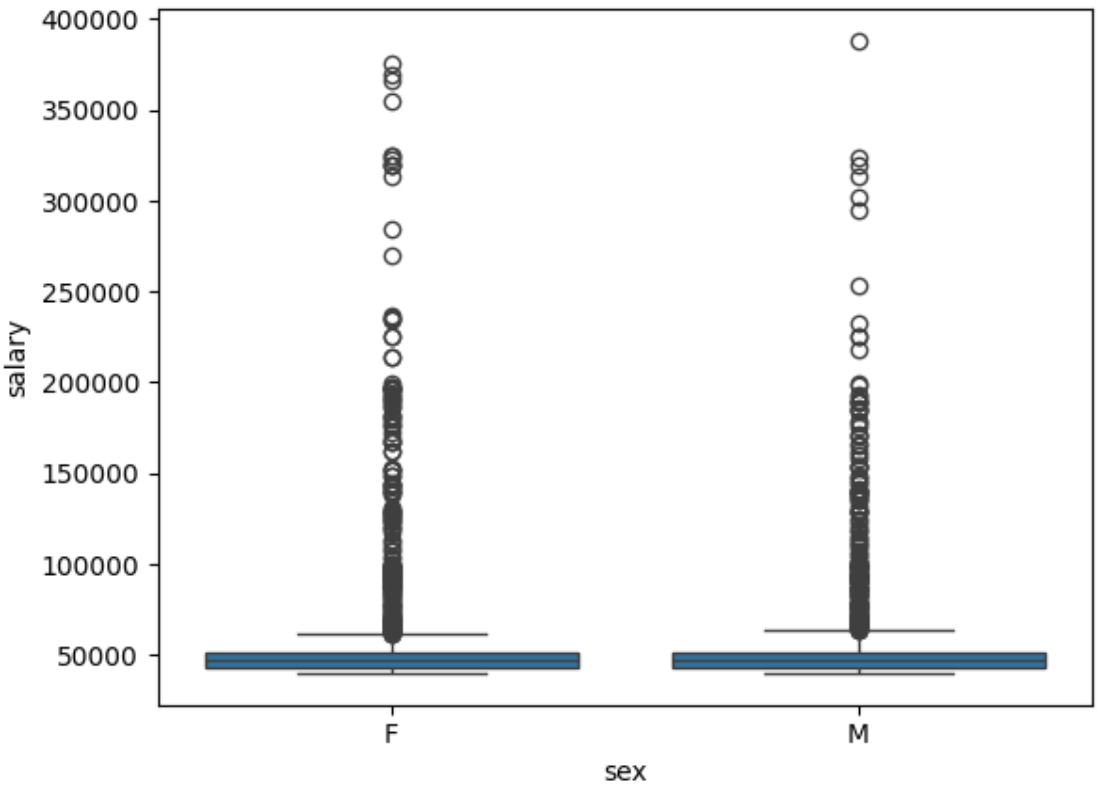
[13]:
```
Text(0.5, 1.0, 'Salary vs Age')
```



4.3 The salary by gender

[14]:
```
sns.boxplot(x="sex",y='salary',data=df)
```

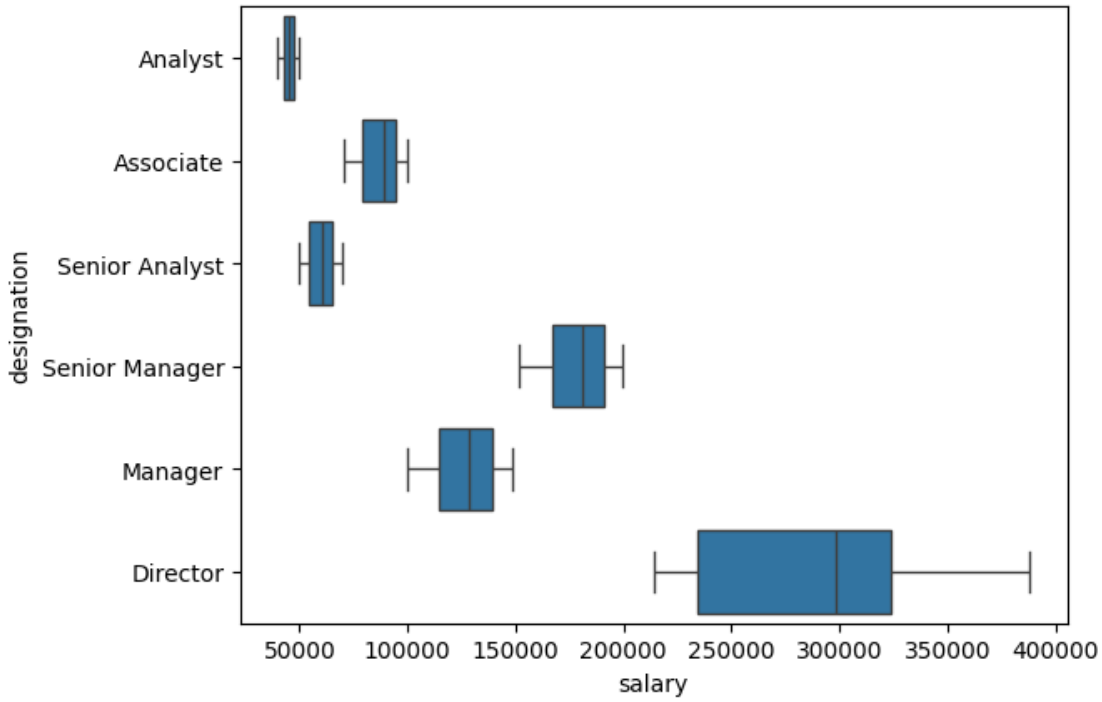
[14]:
```
<Axes: xlabel='sex', ylabel='salary'>
```



4.4 The salary by designation

[16]:
```
sns.boxplot(x='salary',y='designation',data=df)
```

[16]:
```
<Axes: xlabel='salary', ylabel='designation'>
```



5 Correlation Matrix

[17]:
```
df1
```

[17]:

	age	salary	leaves used	leaves remaining	ratings	past exp
0	21.0	44570	24.0	6.0	2.0	0
1	NaN	89207	NaN	13.0	NaN	7
2	21.0	40955	23.0	7.0	3.0	0
3	22.0	45550	22.0	8.0	3.0	0
4	NaN	43161	27.0	3.0	NaN	3
...	...	...	...	...	...	...
2634	36.0	185977	15.0	15.0	5.0	10
2635	23.0	45758	17.0	13.0	2.0	0
2636	21.0	47315	29.0	1.0	5.0	0
2637	24.0	45172	23.0	7.0	3.0	1
2638	24.0	49176	17.0	13.0	2.0	2

[2639 rows x 6 columns]

[18]:
```
df1.corr()
```

[18]:

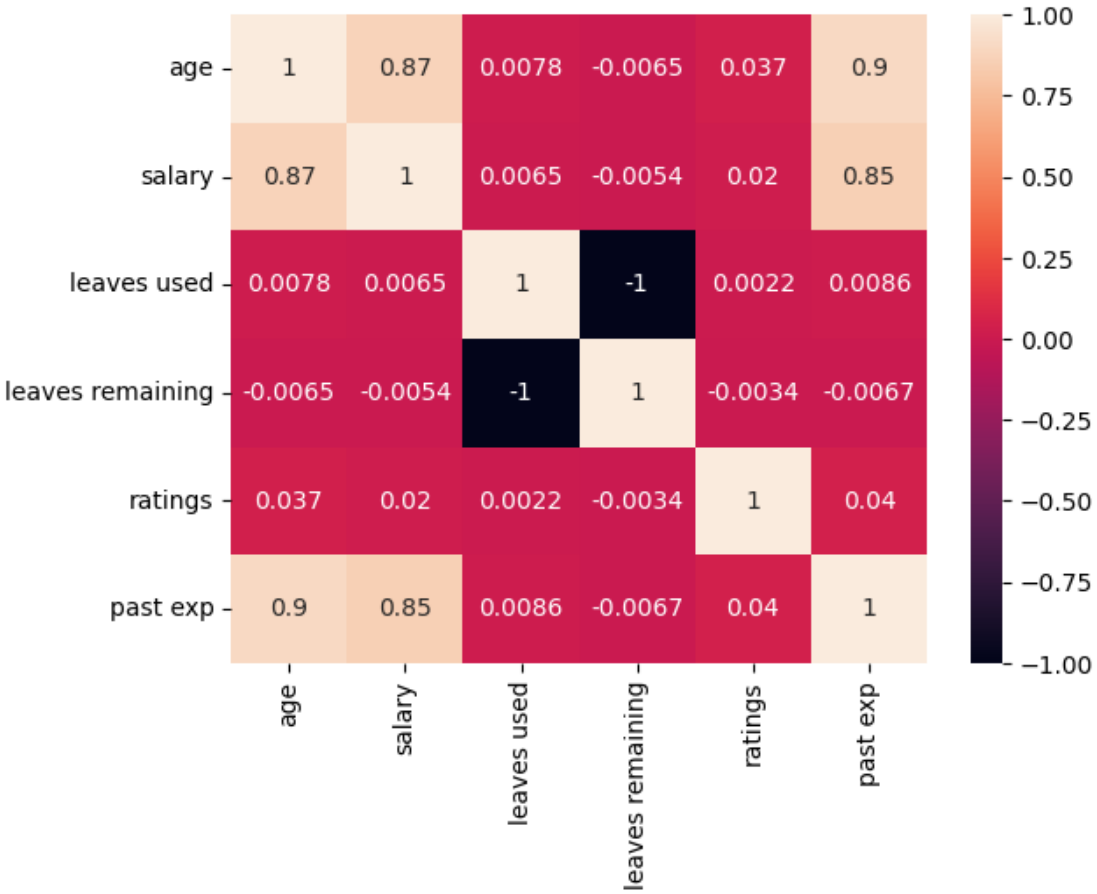
	age	salary	leaves used	leaves remaining	ratings	\
age	1.000000	0.872213	0.007825	-0.006515	0.036801	

```
salary            0.872213  1.000000     0.006498         -0.005422  0.020248
leaves used       0.007825  0.006498     1.000000         -1.000000  0.002200
leaves remaining -0.006515 -0.005422    -1.000000          1.000000 -0.003415
ratings           0.036801  0.020248     0.002200         -0.003415  1.000000
past exp          0.903926  0.854046     0.008601         -0.006728  0.040123

                  past exp
age               0.903926
salary            0.854046
leaves used       0.008601
leaves remaining -0.006728
ratings           0.040123
past exp          1.000000
```

[19]:
```
sns.heatmap(df1.corr(),annot=True)
```

[19]: <Axes: >



6 Data preprocessing

Prepare the data for model training. This involves handling missing values, encoding categorical variables, and scaling or normalizing features as necessary.

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

[20]:
```
first name          0
last name           2
sex                 0
doj                 1
current date        0
designation         0
age                 3
salary              0
unit                0
leaves used         3
leaves remaining    2
ratings             2
past exp            0
dtype: int64
```

[21]:
```
df.head()
```

[21]:

	first name	last name	sex	doj	current date	designation	age	\
0	TOMASA	ARMEN	F	5-18-2014	01-07-2016	Analyst	21.0	
1	ANNIE	NaN	F	NaN	01-07-2016	Associate	NaN	
2	OLIVE	ANCY	F	7-28-2014	01-07-2016	Analyst	21.0	
3	CHERRY	AQUILAR	F	04-03-2013	01-07-2016	Analyst	22.0	
4	LEON	ABOULAHOUD	M	11-20-2014	01-07-2016	Analyst	NaN	

	salary	unit	leaves used	leaves remaining	ratings	past exp
0	44570	Finance	24.0	6.0	2.0	0
1	89207	Web	NaN	13.0	NaN	7
2	40955	Finance	23.0	7.0	3.0	0
3	45550	IT	22.0	8.0	3.0	0
4	43161	Operations	27.0	3.0	NaN	3

[22]:
```
le=LabelEncoder()
for col in df.select_dtypes(include='object'):
    df[col]=le.fit_transform(df[col])
df.head()
```

[22]:

	first name	last name	sex	doj	current date	designation	age	salary	\
0	2208	2436	0	751	0	0	21.0	44570	
1	127	2475	0	967	0	1	NaN	89207	
2	1770	1671	0	865	0	0	21.0	40955	
3	392	2137	0	109	0	0	22.0	45550	

```
4        1377        161    1  494             0            0   NaN   43161

   unit  leaves used  leaves remaining  ratings  past exp
0     0         24.0               6.0      2.0         0
1     5          NaN              13.0      NaN         7
2     0         23.0               7.0      3.0         0
3     1         22.0               8.0      3.0         0
4     4         27.0               3.0      NaN         3
```

7 Machine Learning Model Development

7.1 Linear Regression

[23]:
```
df.fillna(df.mean(),inplace=True)
```

[24]:
```
x=df.drop('salary',axis=1)
y=df['salary']
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=42)
x_train,x_test,y_train,y_test
```

[24]:
```
(      first name  last name  sex  doj  current date  designation   age  unit  \
 2395         627       1637    0  304             0            0  22.0     2
 440         2130       1821    0  101             0            0  24.0     2
 508         1867        615    1  399             0            4  27.0     1
 76          1254       1553    0  214             0            0  24.0     5
 522          968       2172    0  711             0            0  24.0     4
 ...          ...        ...  ...  ...           ...          ...   ...   ...
 1638          56        910    1  495             0            0  25.0     3
 1095         285       1762    0  380             0            5  36.0     0
 1130        1233       1585    0  668             0            0  23.0     5
 1294        1871        131    1  880             0            0  23.0     5
 860          787       1996    0  886             0            0  23.0     3

       leaves used  leaves remaining  ratings  past exp
 2395         18.0              12.0      2.0         0
 440          26.0               4.0      4.0         0
 508          26.0               4.0      5.0         4
 76           29.0               1.0      3.0         1
 522          15.0              15.0      4.0         1
 ...           ...               ...      ...       ...
 1638         25.0               5.0      4.0         2
 1095         18.0              12.0      2.0         8
 1130         20.0              10.0      4.0         0
 1294         18.0              12.0      5.0         1
 860          25.0               5.0      4.0         1

 [2111 rows x 12 columns],
```

```
      first name  last name  sex  doj  current date  designation   age  unit  \
1322        1533       2466    0  566             0            5  36.0     1
1185        1595       2139    0  705             0            0  24.0     1
2572         673       2036    0  244             0            0  24.0     5
1709         323       1417    0  406             0            0  24.0     5
809         1141       2053    0  240             0            1  28.0     5
...          ...        ...  ...  ...           ...          ...   ...   ...
812         1670       1507    0  522             0            4  28.0     5
544         2184       2314    0  189             0            4  28.0     2
1278         971       1108    1  488             0            0  25.0     0
1199         716        868    1   60             0            0  25.0     2
2257        2275       1592    0  400             0            0  22.0     2

      leaves used  leaves remaining  ratings  past exp
1322         27.0               3.0      3.0         6
1185         23.0               7.0      3.0         1
2572         21.0               9.0      4.0         1
1709         27.0               3.0      5.0         1
809          29.0               1.0      2.0         0
...           ...               ...      ...       ...
812          15.0              15.0      4.0         2
544          15.0              15.0      2.0         2
1278         15.0              15.0      4.0         2
1199         29.0               1.0      3.0         2
2257         19.0              11.0      5.0         0

[528 rows x 12 columns],
2395     40031
440      45482
508      51930
76       43638
522      48476
         ...
1638     46071
1095    181470
1130     49449
1294     49873
860      45636
Name: salary, Length: 2111, dtype: int64,
1322    179845
1185     48441
2572     40707
1709     43144
809      84967
         ...
812      54166
544      55693
```

```
1278     42014
1199     45188
2257     43064
Name: salary, Length: 528, dtype: int64)
```

[25]:
```
model=LinearRegression()
model.fit(x_train,y_train)
```

[25]:
```
LinearRegression()
```

[28]:
```
y_pred=model.predict(x_test)
print(y_pred)
mse=mean_squared_error(y_test,y_pred)
print("Mean squared error:", mse)
```

```
[136022.72737355  51571.65550578  51733.51461151  52386.51639039
  68473.71543761  42162.60521125  80754.65322171  41536.08258218
 186117.71944406  31470.7402991   79338.84399651  67259.82283521
  66374.32632472  30464.92046199  68557.24955969  61188.80527572
  66590.71975954  30969.76774241  66112.22300529  31439.99404433
  41852.7097288   46089.41141328  47440.40340117  50326.37251095
  36084.10553438  41515.17772144  62082.23831498  60644.49586398
  55153.54794554  70922.36548195  59633.88298643  30338.30658761
  43173.68886651  41670.70482527  35375.93026778 127055.75872186
  56333.42922916  41830.52240761  57076.94345504  52918.24443049
  51247.82487012  42187.69748236  51279.05122012  55939.97766745
  35068.66043252  51880.08795026  36488.64595322  62804.49532303
  31261.37322452  62647.75989335 111678.53411642  42552.28458946
  32354.28979343  37661.96121358  56692.00343534  32971.87365601
  51987.41228738  40859.34152306  52187.12385072  31523.1140597
  50147.90629116  63830.71526064  51981.96221074  56461.21000363
  52784.58690739  66060.04872795 160830.0988623   62876.74518194
  30283.32390481  60189.57597279  36218.35841608 111264.01662809
  35275.07531312  36552.5260828   31068.97220067  59715.55348193
 102323.73766353  75693.71751743  49846.68912893  47100.93722304
  30767.20221494 199959.18907297  56904.89810923  85344.63881524
  30296.46112954  36742.09625285  79629.55103692 100750.42022107
  44483.97523503 117617.0738149   41072.30947851  42078.54442991
  56615.39824877  57456.14854778  62631.96640278  34398.15412432
  81086.48481686 113695.04009071  62766.64876444  32555.76327285
  56218.97155365  38194.72568454  46663.5236408   93546.64604546
  36547.62686694  53177.73777329  29671.49851276  52765.74193683
  55980.67050992  30067.77684704  86791.77848674  66731.80704762
  51930.91848653  40750.04908942 116610.30624934  50917.51229419
  46669.60047431  60925.71470673  35985.31754887  36721.46033519
  52872.52777436  36661.93949258  60442.94000958  75914.24263133
  48040.4359975   78646.43876348  45293.02820812 169691.9745912
  42530.57328809  52223.45031854 191388.12617227  29251.6218725
```

30946.27583628 30975.0710635   56810.47338909 45473.85751933
35263.65685778 62197.51754707 47991.07342683 60213.12736794
55117.7272866   32391.84408977 32579.29419022 62079.3128174
71354.40425934 87255.44771623 31068.97220067 50721.78233652
67926.56498979 42920.72192906 57034.11814108 42639.50997737
41682.63292975 52199.52073106 80170.53343737 31436.2723545
42020.12536082 32213.48085462 56927.89398358 31378.59770367
35069.7996125 146638.07171929 172109.84943997 117882.84941234
46452.4881162   87041.98817719 61252.75140286 52623.37042453
158641.56068579 61810.90441932 56463.99042496 61092.53401368
30858.86711095 37037.07339633 121778.01652508 36841.53432106
41307.08899857 117674.01099982 57493.66043315 75146.56102277
55556.85844384 56175.6999249   30681.85004   65555.19580166
57080.24343753 29288.13830236 103566.0354521   39800.86846606
89147.80545755 214315.784759   170383.03540199 30404.43830109
29955.82821896 35469.97004153 31053.74841849 33323.77833639
36701.23946362 32897.53374841 31423.9197479   50947.05099964
61503.15871864 154554.42365867 60466.4664804   45274.39734485
37765.88853278 74818.39672068 31266.24668404 36520.20049046
46259.42743237 245155.87593082 46864.53955259 36490.12386749
49919.44092068 61165.36327011 35399.83044951 36627.86009651
43095.77083373 79922.14387447 59612.13416267 31098.53080556
199233.54678375 31863.87989484 52332.44020608 42766.39781969
143904.07649721 62830.05741601 169935.24546171 30404.43830109
38198.60253425 32888.72890795 72383.36503692 61190.5904129
120801.60354284 37120.93829685 46931.51066702 54815.12254756
61822.75107599 36159.0115183   58799.19883537 131642.10494654
45481.01322816 116745.98769906 120545.51260049 60486.2407842
33111.04179099 31840.88129712 31819.33754227 50718.51390152
36936.92786195 35822.99903819 41766.9932395   53304.31690083
67140.99625411 31004.81287208 61485.11599548 30045.44313673
61478.56457503 48033.22829783 65954.10109136 42214.12455161
35305.03142811 30946.31181953 42736.54662089 35447.26960654
89997.56344614 52187.34203576 29829.10900228 45523.61760664
66086.27517066 43797.34607253 61407.88512885 46293.58441566
30473.9710288   30782.40280961 60193.14662566 41333.90844513
60732.30246909 51368.24237597 31563.85130674 47232.62627577
65853.66481005 50583.92350038 66591.947218   37792.00807017
32296.8032755   51032.38082055 30987.05085679 46439.65689561
66738.24238514 43737.06740581 43215.07729916 62819.52508941
31872.25093688 66897.82517642 30503.73983781 37768.7777871
35972.17688125 116778.10192049 66601.12687893 63217.91151299
35036.96969162 56195.92971394 51965.01663681 66078.79434744
42502.2859305   48345.51034515 45839.30044002 75188.88513998
30508.19836944 81470.74100186 30881.40022999 31219.51705641
101071.33601467 102208.97504969 52325.61680899 50975.14304922
60239.16214491 31329.13783091 47353.34084774 41884.09445711
30687.99507259 36670.70475413 38738.86617288 37024.54308392

```
 49231.22855886  37255.79396642  61052.75056677  56134.20480741
 50535.6866154   37583.1235111   32891.42401922 116303.15034919
 56759.89643539  51933.86810373  36656.02371179  52574.14942173
 44609.54471132  62500.13619909  60994.4603806  127672.08768168
 56974.87540512  72968.67819292  67574.88830182  51795.28053256
 51710.50854822  51967.49385839  42437.90119021  56233.29440414
 41325.1692052   41475.77633394  31083.24428602 224646.1513929
 62072.20610491  47211.47410095  29533.73530965  43177.37329829
 66610.91235951  47592.82595092  65915.05388919  61768.41857837
 45348.6141533   65836.95152127  30776.92611949  65128.96937726
 46916.34519761  80862.48828544  43475.09253988  41875.71304668
 55864.27607349  41649.59374503 166269.74651109  29345.77469177
 31964.82976596  56177.21627529  40620.95431049  30945.50209163
 36796.42348744  34405.54630912  51906.03901642  38224.34966578
 65404.98098917  37449.22496381  46999.36492626  36352.41395428
 47823.64980485  57997.05628943  98275.77138015  43091.76699223
 32374.80194226  62072.46781166  42530.57328809  52081.07218191
 36465.11831409  49699.38596967  53309.65929195 145390.31430283
 66281.93224771  41547.47387243  37584.58060029  46507.70610722
 31072.89969992  42197.61173762  78998.67657678  38474.49705131
 31963.13409582  47365.69467402  57240.14122967 108828.10995895
 29867.2157829   46588.69534656  45775.78803024  46840.37918494
 46884.62665278  60304.8441929   75049.87502094  57333.01120674
 60103.87344638  36267.51667884  56929.22773133  31824.44374367
199959.18907297  56968.56511225 184829.16923548  47466.57627338
 61521.22629034  30033.67543754  36995.33409478  66856.51979043
 61560.35309929  64516.794795    52453.91331387  50450.3809705
 90060.2920836   63439.43019511  64966.21964436  35228.19316602
116244.79811226  57606.66126602  94791.47036509  38448.69570018
145136.57586495  65322.82802687  61283.704267   202970.4318977
 89731.79464782  61348.22095617  62238.70970238  35584.51599161
 56284.25129302  41987.77447138  43140.32603025 187237.6241122
 32234.79587836  36398.27609047  47838.26718714  45821.98514814
 44824.06830566  62715.88055179  35946.08026481  30605.06413409
 36015.98667742  43808.18144348  37141.44975917  36953.44513448
 86359.07871904  48133.07652537  31808.05880512  35328.40399858
 53306.96463933  38121.29819805  31712.07399313  56822.82818442
 52350.13058436  40843.28292159 107550.83861299  37715.68074179
 56100.57697391  53010.01426237  62704.82825628  61861.86822955
 65954.10109136  72402.17583552  46992.12863581  42087.14107761
 62618.58905687  76800.31383659  30702.57092685  55434.14344005
 61663.53324219  62591.11883957  60694.88008106  81881.50491712
 36906.6736029   91738.47467069 123954.08909628  30875.48212845
 45977.43295053  35932.96357362  50563.73095262  50947.40249277
 42614.5817637   47715.59719783  61480.72249837  60446.83365462
 33323.95065377  30548.77163406  31048.42763737  60662.07109911
 50372.28996106  31142.47899164  34915.51491842  30811.86911684
 51002.34005935  32339.21778098  35591.58942897  43504.11228197
```

```
  41463.67047239  65535.01979459  71345.89576864  41602.76134544
  55309.34645625  61085.92328638  46743.8646617   77436.31493681
  76622.75954561  60294.74003575  60814.76147739  36837.82485957]
Mean squared error: 399539912.04154205
```

[29]:
```
mae=mean_absolute_error(y_test,y_pred)
print("Mean absolute error:", mae)
```

```
Mean absolute error: 12531.69107970105
```

[31]:
```
r2=r2_score(y_test,y_pred)
print('R-squared score :',r2)
```

```
R-squared score : 0.7593911334181406
```

[32]:
```
model.score(x_test,y_test)
```

[32]:
```
0.7593911334181406
```

[33]:
```
model.score(x_train,y_train)
```

[33]:
```
0.7930278833905532
```

7.2 Decision Trees

[34]:
```
dt=DecisionTreeRegressor()
dt.fit(x_train,y_train)
```

[34]:
```
DecisionTreeRegressor()
```

[35]:
```
y_pred1=dt.predict(x_test)
y_pred1
```

[35]:
```
array([186356.,  44649.,  49190.,  40329.,  73397.,  46966.,  62169.,
        49434., 320148.,  40414.,  61867.,  49866.,  50084.,  48018.,
        68295.,  43714.,  45172.,  40524.,  43563.,  43383.,  41920.,
        41516.,  41353.,  65596.,  43733.,  44305.,  41045.,  54573.,
        65212.,  52690.,  43328.,  42148.,  44214.,  47807.,  40318.,
       132054.,  48602.,  48730.,  43070.,  43433.,  44856.,  42455.,
        45370.,  43295.,  48587.,  59115.,  42160.,  49826.,  41227.,
        49644.,  76143.,  45527.,  41378.,  45062.,  43310.,  41925.,
        69716.,  42213.,  47764.,  48223.,  64913.,  47903.,  40246.,
        45217.,  46779.,  42214., 197246.,  43031.,  49453.,  44145.,
        40800., 128247.,  41250.,  42950.,  45275.,  40634.,  71816.,
        51369.,  60827.,  45174.,  40363., 323196.,  45340.,  51565.,
        41124.,  43308.,  53275.,  93659.,  46752.,  88029.,  40718.,
        42848.,  45389.,  47257.,  44856.,  45464.,  52324.,  87324.,
        42686.,  47001.,  40113.,  40363.,  47691.,  76887.,  44323.,
        41777.,  45379.,  45748.,  43391.,  48539.,  69432.,  49896.,
```

```
 57000.,  46261.,  91602.,  41386.,  46822.,  43751.,  47146.,
 45573.,  44108.,  46026.,  47736.,  52678.,  43740.,  88281.,
 40339., 164773.,  48587.,  41500., 323196.,  42473.,  47990.,
 45379.,  57000.,  46217.,  41284.,  44657.,  48689.,  40786.,
 45748.,  43007.,  44718.,  42962.,  54749.,  98323.,  45275.,
 46791.,  61575.,  42848.,  43137.,  46703.,  41951.,  40240.,
 50687.,  41966.,  41445.,  44513.,  43063.,  48343.,  44806.,
140401., 161839.,  91316.,  40438.,  72359.,  41781.,  48817.,
162639.,  48682.,  45597.,  60778.,  47104.,  44948.,  83066.,
 43676.,  48789.,  98323.,  45776.,  69643.,  47534.,  48817.,
 42760.,  62169.,  40329.,  46741.,  82083.,  40085.,  60522.,
301872., 154106.,  47561.,  46741.,  48587.,  44145.,  41930.,
 41262.,  41008.,  48015.,  44664.,  47843., 135858.,  54048.,
 44824.,  43905.,  60707.,  40720.,  44146.,  49190., 324783.,
 41120.,  41920.,  50130.,  65324.,  41861.,  41164.,  46247.,
 82993.,  52704.,  43051., 180778.,  48492.,  45298.,  45735.,
162639.,  41718., 153769.,  47561.,  44469.,  43036.,  51566.,
 65324.,  91602.,  42345.,  45721.,  40180.,  44220.,  48185.,
 51055., 132054.,  41225., 100952.,  72626.,  47257.,  46813.,
 44094.,  49070.,  43605.,  42041.,  49095.,  43620.,  40329.,
 42047.,  41726.,  56977.,  48555.,  66984.,  47571.,  43938.,
 49052.,  48667.,  47157.,  43132.,  40363.,  66338.,  47069.,
 48959.,  48305.,  45249.,  42178.,  54987.,  48287.,  48781.,
 47759.,  40998.,  41899.,  42666.,  48082.,  45229.,  41394.,
 45674.,  41777.,  48736.,  48581.,  48497.,  43185.,  43051.,
 41657.,  98716.,  40085.,  45896.,  45674.,  48611.,  42056.,
 48587.,  44406.,  41904., 143778.,  42735.,  47517.,  45289.,
 51747.,  42455.,  48098.,  43638.,  48751.,  41308.,  51240.,
 49660.,  72328.,  43733.,  40251.,  88029.,  97400.,  47422.,
 41803.,  64193.,  46326.,  45229.,  45380.,  49736.,  41593.,
 40354.,  46164.,  42686.,  44767.,  66244.,  48101.,  44212.,
 46157.,  40354.,  76143.,  48164.,  50813.,  49581.,  48689.,
 44799.,  49321.,  40786., 129305.,  51053.,  85391.,  63008.,
 67458.,  56181.,  45776.,  45525.,  48969.,  45329.,  40277.,
 44094., 225339.,  40414.,  42455.,  43414.,  43173.,  40270.,
 41125.,  43328.,  47573.,  44799.,  41166.,  43783.,  49559.,
 47573.,  99416.,  46813.,  46306.,  68505.,  46844., 189435.,
 40646.,  48179.,  47877.,  47216.,  48801.,  40541.,  45781.,
 40485.,  40816.,  48662.,  45172.,  41668.,  43063.,  41962.,
 60465.,  90201.,  48272.,  40332.,  46227.,  48587.,  48180.,
 48682.,  43125.,  40565., 186356.,  51566.,  42375.,  40702.,
 49052.,  46662.,  44160.,  86413.,  46612.,  41196.,  47936.,
 46779.,  72886.,  45062.,  44145.,  45340.,  42672.,  40357.,
 41989.,  69329.,  43308.,  41777.,  49624.,  45597.,  43726.,
323196.,  51565., 213987.,  45352.,  42826.,  49271.,  40277.,
 40572.,  64060.,  47764.,  45606.,  41308.,  50687.,  61956.,
 43272.,  42840., 117354.,  42560.,  93529.,  45100., 132054.,
```

```
       58848.,  42280., 161799.,  56772.,  54475.,  43228.,  45550.,
       47257.,  45450.,  47967., 253284.,  46862.,  42148.,  45460.,
       43295.,  46367.,  40572.,  47431.,  44964.,  40275.,  44146.,
       40818.,  43352.,  60183.,  49253.,  40955.,  40001.,  46349.,
       44521.,  49901.,  45748.,  45125.,  48823.,  90201.,  49976.,
       59481.,  60125.,  48696.,  46064.,  43938.,  97400.,  45621.,
       45704.,  45278.,  96378.,  41516.,  46791.,  59877.,  44856.,
       61046.,  56518.,  48689.,  78789., 140762.,  43168.,  47774.,
       47216.,  40113.,  43714.,  48817.,  42737.,  44856.,  49075.,
       42848.,  49752.,  48410.,  44557.,  40113.,  44697.,  48443.,
       41455.,  47477.,  49032.,  45062.,  45538.,  45370.,  47201.,
       50813.,  45680.,  57923.,  43474.,  45896.,  58617.,  55693.,
       49499.,  40113.,  44005.])
```

[36]:
```
mse1=mean_absolute_error(y_test,y_pred1)
mse1
```

[36]:
```
5150.545454545455
```

[37]:
```
r2_1=r2_score(y_test,y_pred1)
r2_1
```

[37]:
```
0.9301813151795492
```

7.3 Random Forest

[39]:
```
rf=RandomForestRegressor()
rf.fit(x_train,y_train)
```

[39]:
```
RandomForestRegressor()
```

[40]:
```
y_pred2=rf.predict(x_test)
y_pred2
```

[40]:
```
array([185386.13,  46250.24,  46285.84,  43762.69,  85103.56,  46136.85,
        58877.19,  45819.83, 259717.07,  43627.79,  61187.96,  44598.2 ,
        53128.8 ,  44356.78,  64367.81,  45931.8 ,  45235.47,  44938.04,
        43332.34,  45627.35,  43496.69,  44137.84,  43963.41,  58533.11,
        43889.34,  45628.77,  45871.14,  62365.7 ,  60075.87,  59203.46,
        43868.74,  45593.05,  44192.29,  44755.92,  45621.71, 122186.27,
        45383.04,  46065.26,  45810.49,  45435.6 ,  45215.08,  42655.72,
        46671.78,  44105.61,  43929.78,  59706.26,  44958.79,  44666.39,
        45419.95,  45914.07,  83385.98,  46036.67,  44701.82,  45572.47,
        46595.64,  44647.59,  61776.63,  42881.51,  45931.59,  44545.48,
        62920.94,  45369.8 ,  45706.28,  44766.27,  45369.58,  43503.72,
       188243.42,  45261.72,  46298.66,  44353.53,  46067.41, 127036.57,
        44030.55,  44942.69,  45419.54,  44011.08,  88724.5 ,  58855.37,
        57441.92,  45369.6 ,  41563.53, 282936.11,  44500.5 ,  57936.4 ,
```

```
 44162.04,  44816.74,  55329.57,  86544.87,  45247.82,  86950.23,
 42993.02,  43708.74,  45834.96,  44238.  ,  44920.65,  45294.15,
 54070.11,  87517.18,  46865.69,  45419.72,  43768.08,  44639.97,
 46467.86,  85048.39,  45058.95,  43727.72,  45459.16,  44969.91,
 44873.19,  44302.79,  61389.85,  45555.1 ,  60974.21,  45079.07,
 85546.44,  43503.84,  44855.63,  44380.78,  45503.77,  46133.08,
 45008.69,  45730.82,  44411.97,  62589.91,  44374.24,  86134.48,
 43942.61, 177862.39,  45526.18,  45824.68, 281804.4 ,  44099.76,
 46002.84,  44531.53,  57857.01,  45386.07,  45259.93,  45177.07,
 45954.02,  44117.87,  45116.69,  44219.69,  45504.68,  45232.9 ,
 60622.64,  86298.63,  45419.54,  46140.31,  60785.41,  44261.66,
 44620.86,  46059.62,  45611.55,  44812.12,  58423.74,  44926.52,
 44716.62,  45862.87,  45341.65,  45208.37,  45061.12, 127700.08,
168049.38,  85165.59,  44973.76,  85214.6 ,  45305.43,  45148.94,
184630.52,  44860.36,  44826.95,  59611.82,  44120.09,  44899.98,
 82858.64,  43645.28,  44927.2 ,  91384.61,  44603.06,  64307.91,
 45041.91,  45606.62,  45397.98,  57330.69,  44512.25,  44692.6 ,
 83916.54,  45385.9 ,  60402.88, 295434.31, 169818.45,  45685.62,
 44766.49,  47266.22,  43628.24,  44269.81,  44828.38,  43828.39,
 45453.58,  44784.33,  44873.34, 131434.65,  57106.53,  44569.29,
 44833.5 ,  61073.17,  44409.43,  45293.92,  45103.69, 309408.77,
 46285.77,  43669.3 ,  58581.71,  61710.29,  43935.08,  42666.25,
 44937.14,  83834.49,  56496.98,  44454.69, 182545.61,  46254.65,
 45492.09,  45313.65, 176700.97,  46505.61, 178814.36,  45685.62,
 45062.22,  43656.74,  60529.77,  63053.82,  87076.49,  45486.45,
 46499.89,  44620.  ,  44136.01,  44537.84,  57474.84, 123088.17,
 45443.35, 122800.48,  77378.99,  44286.06,  44043.83,  45762.69,
 45555.56,  45971.98,  45141.35,  45376.  ,  45101.23,  42390.32,
 45208.02,  44383.06,  59612.31,  45020.13,  59775.88,  43830.51,
 44969.49,  45105.24,  46190.84,  45956.06,  45769.74,  43495.54,
 63858.79,  45013.71,  45392.35,  43687.57,  44595.53,  45724.27,
 55628.41,  44377.94,  45314.39,  44960.8 ,  44399.78,  44246.55,
 44999.77,  44761.61,  45077.18,  44957.53,  44338.37,  44340.4 ,
 45737.97,  46567.26,  44506.92,  46460.91,  45068.11,  45339.44,
 82019.19,  44426.62,  45863.39,  44508.9 ,  45214.28,  44744.85,
 46140.19,  43887.82,  43834.88, 140083.08,  44167.75,  45037.42,
 44079.87,  56522.92,  46068.25,  46185.74,  46132.11,  45110.46,
 44786.08,  54432.5 ,  45215.14,  86870.24,  44980.41,  44692.04,
 83983.48,  86514.53,  44346.55,  45326.5 ,  59794.32,  45071.78,
 45514.33,  45278.88,  46214.74,  43979.32,  45191.46,  44874.12,
 45845.13,  46593.9 ,  59038.12,  44811.28,  44418.09,  45120.55,
 44712.95,  83251.89,  45408.23,  54311.43,  45030.87,  45223.65,
 44527.28,  45039.27,  42165.99, 125796.56,  55430.48,  89035.06,
 59654.72,  63945.45,  57674.85,  45550.56,  45448.45,  45894.11,
 43831.28,  44165.18,  45715.8 , 251658.37,  44300.  ,  45258.05,
 43965.23,  44142.1 ,  41948.81,  44804.64,  44435.29,  47248.46,
 44821.64,  44460.66,  45015.99,  45876.63,  45271.56,  85656.64,
```

```
 44736.15,  44608.93,  63567.78,  44156.38, 180385.39,  43992.  ,
 45921.74,  45418.95,  46580.13,  45956.12,  46510.17,  44416.01,
 44963.16,  45937.13,  45153.25,  45313.75,  45390.57,  44913.61,
 44933.48,  58795.44,  85574.17,  47415.45,  44548.94,  46008.27,
 45526.18,  45421.2 ,  45072.15,  43857.61,  43941.85, 180225.54,
 61189.38,  45534.11,  45267.82,  45648.69,  43515.8 ,  44496.24,
 85689.56,  45162.04,  44944.54,  46026.16,  44866.46,  90582.38,
 44336.05,  44376.78,  45023.57,  45152.53,  44343.14,  45618.18,
 56587.94,  45621.23,  43503.83,  45511.24,  45047.14,  44635.51,
282936.11,  57449.72, 231359.18,  45518.68,  45726.94,  45953.97,
 44751.51,  45135.47,  60375.55,  44644.09,  45402.14,  45303.15,
 53635.34,  62213.19,  45251.04,  44029.16, 121117.08,  44549.67,
 89819.13,  45617.64, 122500.75,  59023.6 ,  44824.9 , 182975.06,
 59252.43,  60458.37,  45583.02,  44316.65,  44289.49,  43917.21,
 44825.08, 253214.37,  45338.27,  44720.56,  44574.49,  44885.03,
 45465.45,  44119.12,  45190.76,  45248.26,  44981.92,  45041.28,
 44488.98,  44996.45,  61207.3 ,  46452.87,  43723.45,  43487.99,
 46295.7 ,  45021.19,  44023.53,  46207.07,  46184.13,  43991.34,
 85883.76,  46569.97,  60800.78,  61819.02,  44520.24,  46118.64,
 44969.49,  92080.11,  45026.09,  46218.44,  45088.44,  88753.73,
 44170.11,  46297.4 ,  60697.97,  43852.64,  61126.62,  59454.61,
 45475.38,  90669.91, 130792.19,  44265.66,  44877.81,  46043.44,
 43800.81,  45620.94,  45384.86,  44808.45,  43715.71,  46482.56,
 43646.66,  44838.6 ,  44472.59,  45930.25,  45620.84,  45722.8 ,
 44828.17,  45445.11,  46044.79,  44483.29,  45061.  ,  44844.96,
 44559.55,  45272.  ,  56728.99,  45844.02,  58692.15,  44534.49,
 45495.  ,  58206.81,  57079.73,  45434.82,  45303.49,  44143.75])
```

[41]:
```
mse2=mean_absolute_error(y_test,y_pred2)
mse2
```

[41]: 4317.633579545454

[42]:
```
r2_2=r2_score(y_test,y_pred2)
r2_2
```

[42]: 0.7593911334181406

7.4 Gradient Boosting

[57]:
```
from sklearn.ensemble import GradientBoostingRegressor
gb=GradientBoostingRegressor()
gb
```

[57]: GradientBoostingRegressor()

[58]:
```
gb.fit(x_train,y_train)
y_pred3=gb.predict(x_test)
y_pred3
```

[58]:
```
array([186728.59405888,  45070.84294477,  45198.82691064,  44341.84407785,
        81371.41323456,  44993.81104778,  60012.55950194,  45292.22543005,
       248282.98082785,  46592.29470295,  61274.77811785,  45274.74071182,
        59735.6478639 ,  44849.63653917,  59562.40381207,  45161.56262571,
        45448.3841448 ,  45194.83557534,  45522.0078771 ,  45264.25926915,
        44988.09440933,  44778.21524534,  46882.0249771 ,  58839.86833143,
        44286.40330087,  45198.82691064,  45389.35348674,  59522.80679308,
        57887.95649934,  62591.41274145,  44642.91314146,  45139.87559998,
        44823.72634765,  45119.02063677,  45640.15781882, 125635.32727244,
        44615.7056458 ,  45640.15781882,  45432.884061  ,  45006.098115  ,
        44537.2394815 ,  45078.99462454,  45286.24496378,  45221.14434919,
        45017.14267203,  58688.08561964,  45018.65590407,  45274.74071182,
        44790.28125379,  44587.97695142,  86642.23965434,  45450.27752269,
        45314.25621641,  45213.7572048 ,  44864.10177876,  45199.51443938,
        59204.92608822,  45309.57735237,  45081.32438744,  45085.08571018,
        59483.59333209,  44978.04936435,  44593.11955432,  45164.24146418,
        45136.28284166,  45413.06957987, 186931.85101111,  45006.098115  ,
        45278.50203456,  44819.0474836 ,  45108.88380439, 125762.96925135,
        45124.0482688 ,  44786.5358787 ,  44738.92565803,  45108.88380439,
        87646.79300841,  62649.97458091,  58100.49904975,  45116.37823239,
        44781.2420095 , 264536.96121917,  44992.77327337,  59883.19844901,
        45432.884061  ,  45055.87586898,  59289.09144244,  88894.72870221,
        45274.74071182,  90167.73528486,  45057.51810314,  44893.35971153,
        45078.99462454,  44577.47324456,  44618.63139951,  45096.81779595,
        62163.98100779,  90230.96688258,  45342.48116009,  44501.69199067,
        45442.7622679 ,  44706.14473918,  45662.66650804,  87594.44965773,
        45400.98807522,  45162.59161552,  45033.80201001,  45274.74071182,
        44987.8286236 ,  45260.32944427,  63013.99841966,  45083.07521514,
        59045.27933053,  45339.48554159,  90030.84177141,  44919.26359311,
        44826.53169483,  45146.06921721,  45448.029432  ,  45116.37823239,
        45430.69625946,  45662.66650804,  44978.04936435,  60168.12390419,
        45134.08208087,  80152.93278318,  45129.40321683, 171199.64414164,
        45052.36301712,  45413.06957987, 271827.00401645,  45105.03438906,
        45650.51025939,  45018.65590407,  58348.51390726,  45377.1417301 ,
        45535.89748447,  44883.00727097,  45157.30448105,  45178.48422959,
        45274.74071182,  45274.74071182,  45209.71232804,  44997.10725823,
        60654.44617652,  81780.817943  ,  44738.92565803,  45472.39956559,
        60041.25320889,  44923.08379384,  45288.85447512,  45491.67976648,
        44995.20890615,  44701.46587513,  62368.96231575,  44546.35044181,
        45187.28274522,  44623.92399413,  45349.83798215,  45104.17282435,
        44665.36206579, 133943.42598743, 182617.30259966,  86053.57138903,
        45129.40321683,  80534.69544279,  45221.14434919,  45288.85447512,
       176385.92577107,  45511.26102972,  45511.97406502,  58644.45092484,
```

```
 44691.60634861,  44827.76257445,  85794.1490609 ,  45198.82691064,
 45081.32438744,  88401.9103605 ,  45380.6990696 ,  59388.48715456,
 45337.80229604,  45288.85447512,  44765.5357463 ,  59400.48182928,
 44642.91314146,  45108.88380439,  82527.9497899 ,  45211.5091141 ,
 62986.54252619, 277891.9809019 , 168738.90039137,  45206.83025006,
 45104.20494035,  45409.16796786,  45052.36301712,  45081.32438744,
 45146.06921721,  45211.5091141 ,  45487.00090244,  45346.34227149,
 45260.20227068, 138765.94656518,  59570.96937986,  45337.7564775 ,
 44890.53602468,  62106.7225662 ,  45146.06921721,  46165.64961153,
 45005.41058626, 322934.67857443,  45042.79419411,  44988.09440933,
 59108.51092825,  60479.88519398,  44799.35694899,  45413.06957987,
 44940.66575649,  85820.74034348,  58705.43443186,  45432.884061  ,
209704.65011714,  45116.37823239,  45272.41094892,  44756.36427675,
181250.84723925,  45323.40682739, 180595.88818939,  45206.83025006,
 45020.34088041,  44903.40147157,  60566.0173976 ,  60146.06445062,
 90084.73851361,  44889.65889076,  45413.06957987,  45598.91325014,
 45140.4622746 ,  44992.77327337,  59045.27933053, 131604.71688385,
 44584.52646629, 112661.03466036,  82680.55045995,  44512.7284148 ,
 45134.08208087,  45522.0078771 ,  44902.68955341,  45400.98807522,
 45508.4810449 ,  44508.30643429,  45140.53689773,  44642.91314146,
 45081.32438744,  45339.48554159,  59952.58116597,  45314.25621641,
 59803.55832567,  45153.76647928,  44665.36206579,  45219.65325549,
 45199.51443938,  45080.39793431,  45131.10214142,  44638.23427741,
 59668.64404532,  45187.22737216,  43061.87010304,  45270.06184778,
 43981.86477561,  45199.51443938,  59905.12912671,  44976.40579038,
 44978.04936435,  44601.43789814,  44395.78900595,  45134.08208087,
 45276.25394386,  44770.88956894,  44768.65134398,  45274.74071182,
 44454.33340543,  38500.59858714,  45413.06957987,  45613.74900946,
 44706.14473918,  44992.77327337,  45270.06184778,  45474.81230571,
 77291.70467395,  45211.5091141 ,  45116.37823239,  44642.91314146,
 45018.6152385 ,  45270.92207704,  45104.20494035,  44923.08379384,
 45070.85048315, 118939.26990888,  44653.9501601 ,  44992.77327337,
 44799.35694899,  59328.17456284,  45352.08607284,  45662.66650804,
 45491.67976648,  45134.08208087,  45017.14267203,  59323.49569879,
 45016.92173935,  90283.46144419,  44645.09151361,  45328.59155272,
 89406.14230851,  87147.37682743,  45234.73478448,  45020.34088041,
 59426.14334308,  45270.06184778,  44005.17085609,  44681.13800431,
 52839.56574666,  45291.90025362,  45274.74071182,  45339.48554159,
 45081.32438744,  45231.80341264,  59195.07083018,  45274.74071182,
 45188.79309856,  45334.15367477,  45081.32438744,  83483.00538347,
 45613.74900946,  57322.10784493,  45144.10414949,  45399.11193341,
 45120.83989203,  45413.06957987,  44881.35661547, 124411.93853901,
 58590.74311371,  81295.40637794,  59372.69523026,  59685.46549213,
 59103.83206421,  45511.26102972,  45535.89748447,  45209.71232804,
 44521.29486857,  44590.71856238,  45129.40321683, 241692.62152025,
 45616.28862399,  45116.37823239,  44978.04936435,  45450.27752269,
 45096.81779595,  45339.48554159,  44770.88956894,  45219.03271228,
```

```
 45120.83989203,  44785.96167076,  45630.00863453,  45185.5847218 ,
 45270.92207704,  81629.27812048,  45134.08208087,  45599.43491032,
 58657.75872978,  45317.46747187, 186187.05740217,  44903.40147157,
 45067.8705437 ,  44919.06023298,  45274.74071182,  45600.64231423,
 45254.96782042,  44917.54700093,  44980.39622545,  44771.26505051,
 45034.30221383,  45294.48332501,  45349.83798215,  45630.00863453,
 45211.5091141 ,  59424.21860326,  88072.70877756,  45126.24386513,
 44953.91107431,  45046.88058214,  45052.36301712,  45460.67121203,
 38391.80331968,  44275.28164377,  44059.90116364, 178737.26150852,
 60526.961925  ,  45135.59531292,  45081.32438744,  45452.58508446,
 44819.0474836 ,  45129.40321683,  79931.00582887,  45274.74071182,
 45015.91429312,  45211.5091141 ,  45322.90866508,  87562.2001049 ,
 45096.81779595,  45177.20893547,  45432.884061  ,  45556.42459625,
 44813.58357716,  45274.74071182,  59516.91202317,  45055.87586898,
 44703.21898547,  45337.84330743,  45879.60556271,  45199.19172264,
264536.96121917,  58514.26619818, 226751.07601634,  45339.48554159,
 45194.83557534,  45491.67976648,  44653.9501601 ,  45081.32438744,
 59701.7840139 ,  45081.32438744,  45349.47656151,  45225.82321324,
 62453.28130314,  59411.19280236,  44589.20533033,  44797.87835778,
123751.58328966,  45015.91429312,  88819.16273868,  45491.67976648,
124533.17099992,  59340.64368942,  45071.03858079, 215112.17014106,
 63150.30405379,  59402.72991997,  44775.83115375,  45276.25394386,
 44508.04955075,  45272.41094892,  45134.08208087, 246556.12711311,
 45123.99577444,  45507.50709766,  45187.22737216,  45072.9611158 ,
 45211.5091141 ,  45351.02113554,  44987.8286236 ,  43727.38091989,
 44638.23427741,  45213.7572048 ,  45408.26667664,  44490.27928422,
 62515.94675394,  45337.84330743,  45033.32655237,  44589.20533033,
 45535.89748447,  45430.69625946,  45199.51443938,  45450.27752269,
 45370.60239877,  44940.66575649,  89689.66379624,  45281.39598133,
 59679.34322092,  59108.51092825,  45272.41094892,  45129.40321683,
 44665.36206579,  81575.16800001,  45262.37713685,  45209.1793512 ,
 45006.098115  ,  88071.82749272,  45070.85048315,  45314.25621641,
 58753.1463284 ,  44308.27566629,  59786.65292482,  62503.82590645,
 44627.57740846,  88514.33898123, 113109.84805257,  44525.97373261,
 45337.7564775 ,  45274.74071182,  44940.66575649,  45161.56262571,
 45158.66974846,  45268.92720676,  44534.117874  ,  45479.83877986,
 44642.91314146,  44879.74808852,  45076.6455234 ,  45274.74071182,
 45274.74071182,  45387.04592497,  44703.21898547,  45270.06184778,
 44992.77327337,  45129.99050301,  45092.13893191,  44709.01496964,
 45171.50318675,  44790.5404571 ,  59285.79051113,  45601.68300102,
 58706.64483408,  45081.32438744,  45209.07834075,  62567.05750417,
 62227.21260551,  45213.7572048 ,  45351.02113554,  44992.77327337])
```

[59]:
```
mse3=mean_absolute_error(y_test,y_pred3)
print('Mean Absolute Error:',mse3)
```

```
Mean Absolute Error: 4661.074643221528
```

[60]:
```
r2_3=r2_score(y_test,y_pred3)
print('R-squared:',r2_3)
```

```
R-squared: 0.9313643404629065
```

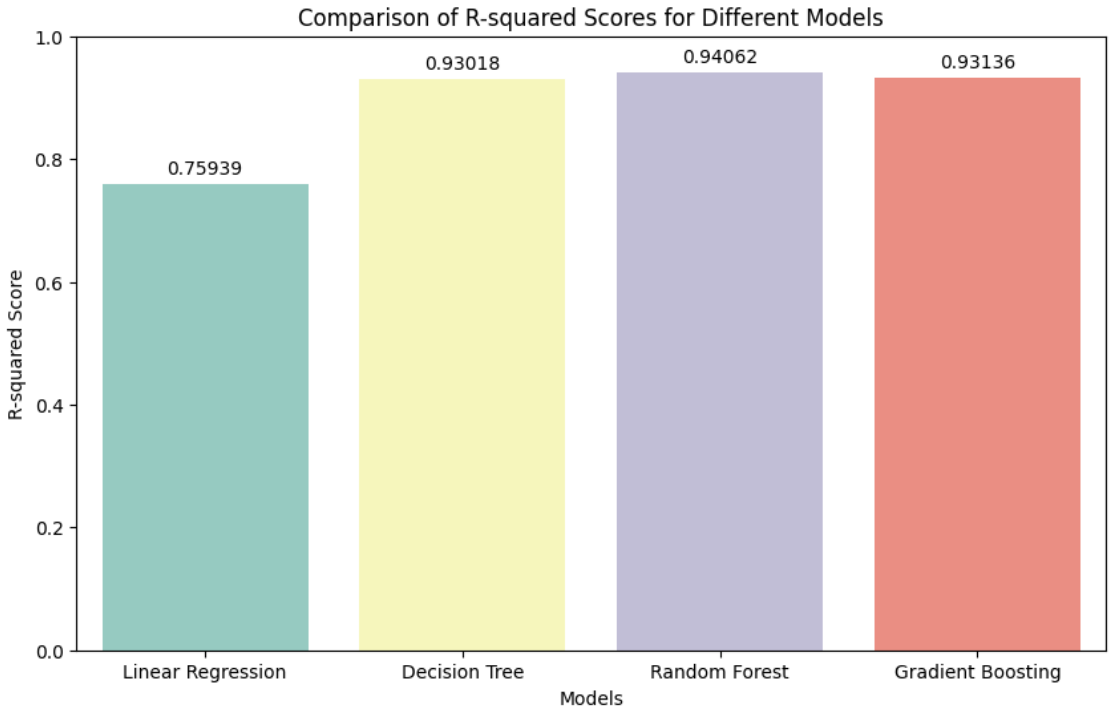
7.5 Comparison of R-squared Scores for different Models

[68]:
```
r2=0.75939113
r2_1=0.93018131
r2_2=0.94062267
r2_3=0.93136434
models=['Linear Regression','Decision Tree','Random Forest','Gradient Boosting']
r2_scores=[r2,r2_1,r2_2,r2_3]
plt.figure(figsize=(10,6))
ax=sns.barplot(x=models,y=r2_scores,palette='Set3')
for m,score in enumerate(r2_scores):
    ax.text(m , score +0.01 ,f'{score:.5f}', ha='center' , va='bottom')
plt.xlabel('Models')
plt.ylabel('R-squared Score')
plt.title('Comparison of R-squared Scores for Different Models')
plt.ylim(0,1)
plt.show()
```

```
C:\Users\KHALID\AppData\Local\Temp\ipykernel_21836\1556638125.py:8:
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same
effect.

  ax=sns.barplot(x=models,y=r2_scores,palette='Set3')
```


Comparison of R-squared Scores for Different Models
R-squared Score
1.0
0.8
0.6
0.4
0.2
0.0
0.75939
0.93018
0.94062
0.93136
Linear Regression
Decision Tree
Random Forest
Gradient Boosting
Models

[ ]:

[ ]:

8 Conclusion

This project focused on predicting salaries of data professionals has provided valuable insights into the realm of data science and machine learning. Through an in-depth analysis of a comprehensive dataset including attributes such as experience, education, job title, and geographic location, we developed and evaluated multiple regression models.

The results highlight the effectiveness of various models:

- Random Forest (RF) achieved an impressive R^2 of 0.9406, showcasing its robust predictive power in capturing the complexities of salary prediction.
- Gradient Boosting (GB) closely followed with an R^2 of 0.9314, demonstrating strong performance in refining predictions through iterative learning.
- Linear Regression (RL), while showing moderate performance with an R^2 of 0.7594, provided a baseline understanding of salary prediction based on linear relationships.

These findings underscore the importance of model selection and optimization in accurately predicting salaries within the data profession. Furthermore, the project's insights into influential factors such as experience, education, and job role provide actionable information for individuals and organizations aiming to make informed decisions regarding compensation strategies.

Project Outcome

The project successfully:

- Developed and evaluated predictive models for salary prediction using machine learning techniques.
- Identified Random Forest as the most effective model for predicting data professional salaries, achieving an R^2 of 0.9406.
- Demonstrated the applicability of Gradient Boosting for refining predictions with an R^2 of 0.9314.

In conclusion, this project not only enhanced our proficiency in data analysis and predictive modeling but also illustrated the significant impact of data-driven approaches in guiding strategic decisions within the dynamic landscape of data professions.

9 References

Dataset : Kaggle - Salary Prediction of Data Professions