Problem Statement ¶

A real estate agent want help to predict the house price for regions in USA.He gave us the dataset to work on to use linear regression model.Create a model that helps him to estimate of what the house would sell for

Import libraries

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
# To import dataset
df=pd.read_csv('Instagram csv')
df
```

Out[3]:

	Impressions	From Home	From Hashtags	From Explore	From Other	Saves	Comments	Shares	Likes	Profile Visits	F
0	3920	2586	1028	619	56	98	9	5	162	35	
1	5394	2727	1838	1174	78	194	7	14	224	48	
2	4021	2085	1188	0	533	41	11	1	131	62	
3	4528	2700	621	932	73	172	10	7	213	23	
4	2518	1704	255	279	37	96	5	4	123	8	
...	...	...	...	...	...	...	...	...	...	...	
114	13700	5185	3041	5352	77	573	2	38	373	73	
115	5731	1923	1368	2266	65	135	4	1	148	20	
116	4139	1133	1538	1367	33	36	0	1	92	34	
117	32695	11815	3147	17414	170	1095	2	75	549	148	

	Impressions	From Home	From Hashtags	From Explore	From Other	Saves	Comments	Shares	Likes	Profile Visits	F
118	36919	13473	4176	16444	2547	653	5	26	443	611	

119 rows × 13 columns

In [4]:
```
# To display top 10 rows
df.head(10)
```

Out[4]:

	Impressions	From Home	From Hashtags	From Explore	From Other	Saves	Comments	Shares	Likes	Profile Visits	Fol
0	3920	2586	1028	619	56	98	9	5	162	35	
1	5394	2727	1838	1174	78	194	7	14	224	48	
2	4021	2085	1188	0	533	41	11	1	131	62	
3	4528	2700	621	932	73	172	10	7	213	23	
4	2518	1704	255	279	37	96	5	4	123	8	
5	3884	2046	1214	329	43	74	7	10	144	9	
6	2621	1543	599	333	25	22	5	1	76	26	
7	3541	2071	628	500	60	135	4	9	124	12	
8	3749	2384	857	248	49	155	6	8	159	36	
9	4115	2609	1104	178	46	122	6	3	191	31	

Data Cleaning and Pre-Processing

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119 entries, 0 to 118
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   Impressions     119 non-null    int64 
 1   From Home       119 non-null    int64 
 2   From Hashtags   119 non-null    int64 
 3   From Explore    119 non-null    int64 
 4   From Other      119 non-null    int64 
 5   Saves           119 non-null    int64 
 6   Comments        119 non-null    int64 
 7   Shares          119 non-null    int64 
 8   Likes           119 non-null    int64 
 9   Profile Visits  119 non-null    int64 
 10  Follows         119 non-null    int64 
 11  Caption         119 non-null    object
 12  Hashtags        119 non-null    object
dtypes: int64(11), object(2)
memory usage: 12.2+ KB
```

In [6]:
```
df.describe()
```

Out[6]:

	Impressions	From Home	From Hashtags	From Explore	From Other	Saves	Comn
count	119.000000	119.000000	119.000000	119.000000	119.000000	119.000000	119.0C
mean	5703.991597	2475.789916	1887.512605	1078.100840	171.092437	153.310924	6.6E
std	4843.780105	1489.386348	1884.361443	2613.026132	289.431031	156.317731	3.54
min	1941.000000	1133.000000	116.000000	0.000000	9.000000	22.000000	0.0C
25%	3467.000000	1945.000000	726.000000	157.500000	38.000000	65.000000	4.0C
50%	4289.000000	2207.000000	1278.000000	326.000000	74.000000	109.000000	6.0C
75%	6138.000000	2602.500000	2363.500000	689.500000	196.000000	169.000000	8.0C
max	36919.000000	13473.000000	11817.000000	17414.000000	2547.000000	1095.000000	19.0C

In [7]:
```
df.columns
```

Out[7]:
```
Index(['Impressions', 'From Home', 'From Hashtags', 'From Explore',
       'From Other', 'Saves', 'Comments', 'Shares', 'Likes', 'Profile Visit
s',
       'Follows', 'Caption', 'Hashtags'],
      dtype='object')
```

In [8]:

```
a = df.dropna(axis='columns')
a.columns
```

Out[8]:

```
Index(['Impressions', 'From Home', 'From Hashtags', 'From Explore',
       'From Other', 'Saves', 'Comments', 'Shares', 'Likes', 'Profile Visit
s',
       'Follows', 'Caption', 'Hashtags'],
      dtype='object')
```

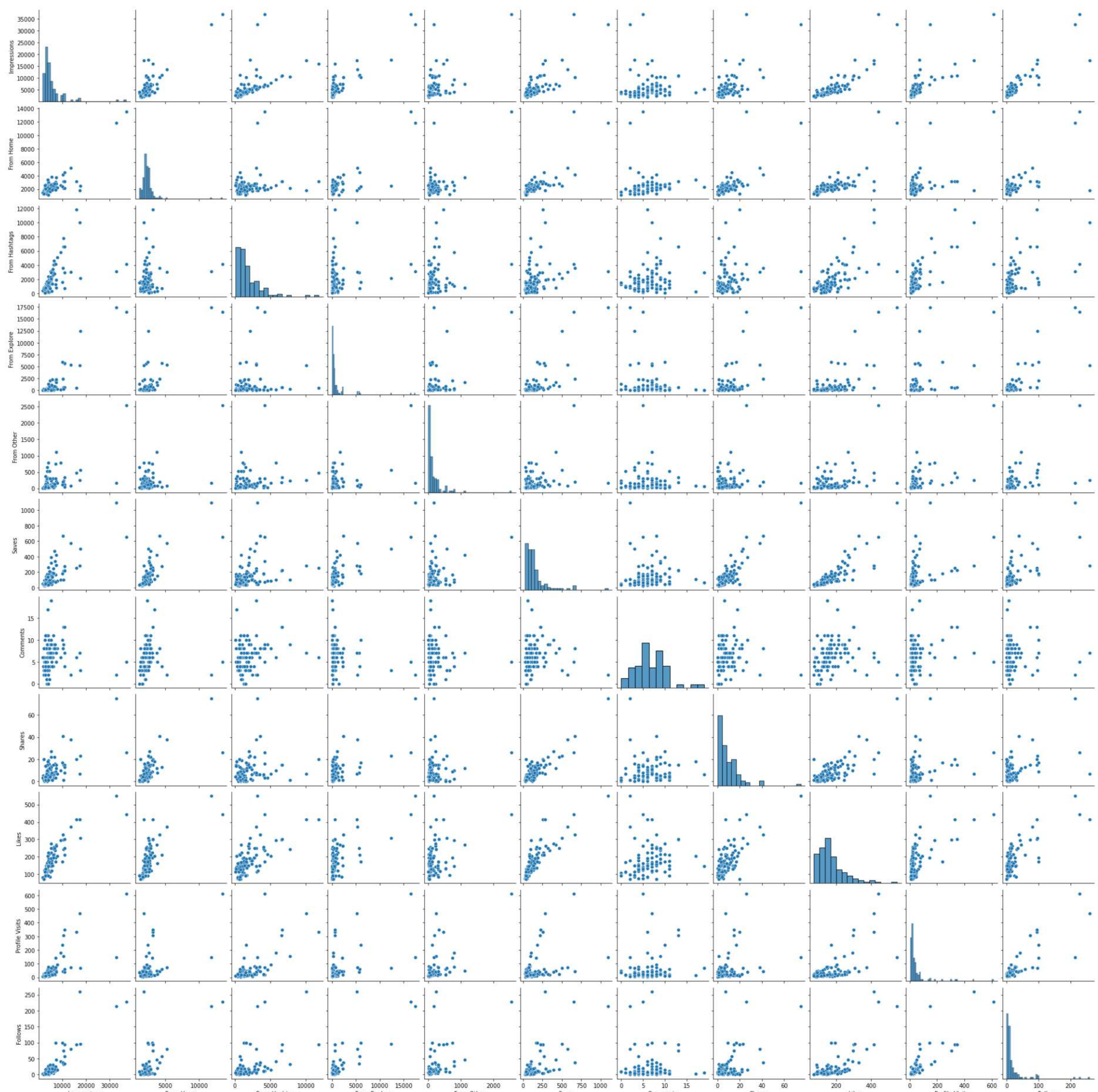
EDA and Visualization

In [9]:

```
sns.pairplot(a)
```

Out[9]:

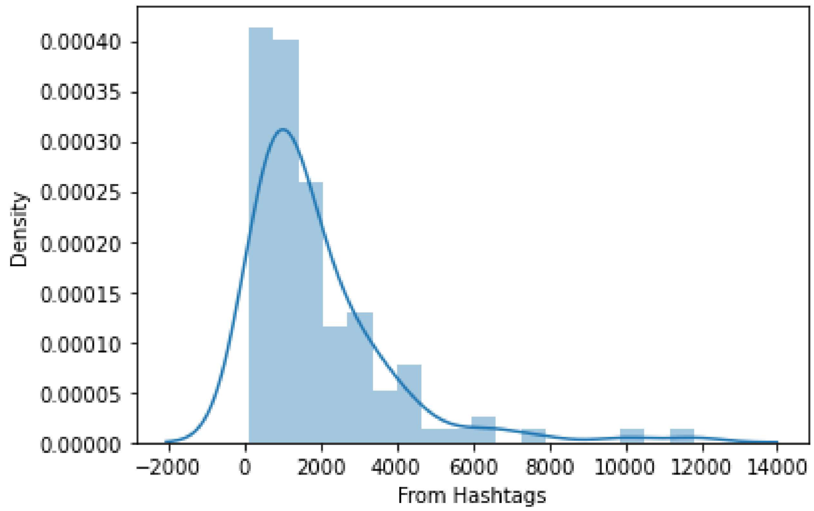
```
<seaborn.axisgrid.PairGrid at 0x1bccde1b550>
```



In [10]:
```
sns.distplot(a['From Hashtags'])
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: Fut
ureWarning: `distplot` is a deprecated function and will be removed in a futu
re version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function for hi
stograms).
  warnings.warn(msg, FutureWarning)
```

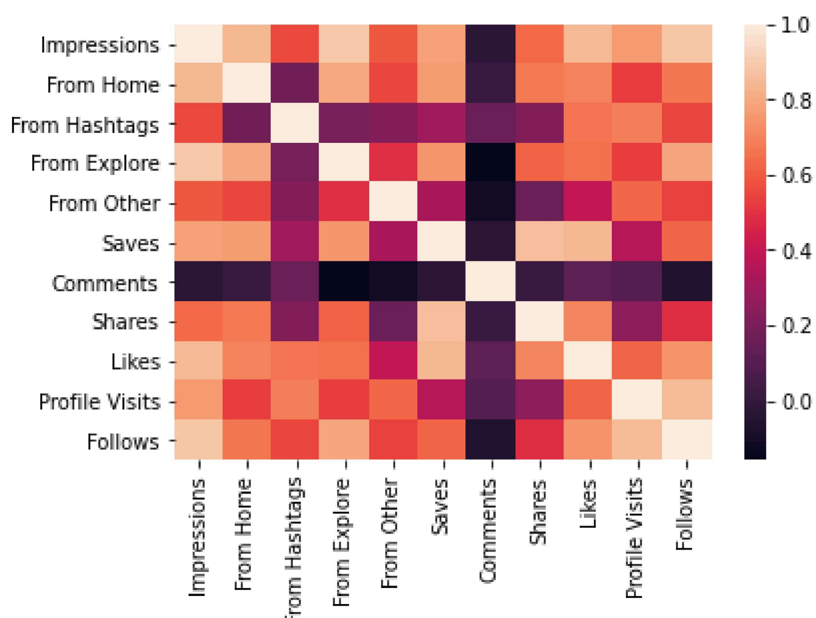
Out[10]: <AxesSubplot:xlabel='From Hashtags', ylabel='Density'>



In [11]:
```
a1=a[['Impressions', 'From Home', 'From Hashtags', 'From Explore',
       'From Other', 'Saves', 'Comments', 'Shares', 'Likes', 'Profile Visits',
       'Follows']]
```

In [12]:
```
sns.heatmap(a1.corr())
```

Out[12]: <AxesSubplot:>



To Train the Model - Model Building

We are going to train Linear Regression model;We need to split out data into two variables x and y where x is independent variable (input) and y is dependent on x(output). We could ignore address column as it is not required for our model.

In [13]:
```
x=a1[['Impressions', 'From Home', 'From Explore',
       'From Other', 'Saves', 'Comments', 'Shares', 'Likes', 'Profile Visits',
       'Follows']]
y=a1['From Hashtags']
```

To split my dataset into training and test data

In [14]:
```
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.3)
```

In [15]:
```
from sklearn.linear_model import LinearRegression

lr=LinearRegression()
lr.fit(x_train,y_train)
```

Out[15]:
```
LinearRegression()
```

In [16]:
```
print(lr.intercept_)
```

```
-77.7277810377102
```

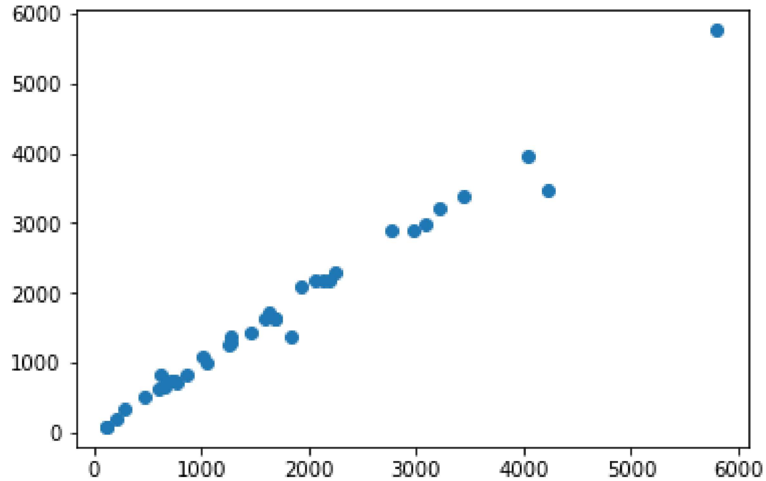
In [17]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

Out[17]:

	Co-efficient
Impressions	0.962427
From Home	-0.989016
From Explore	-0.967234
From Other	-1.072870
Saves	0.141394
Comments	0.358981
Shares	0.235983
Likes	0.533632
Profile Visits	0.833593
Follows	-0.816640

In [18]:
```
prediction=lr.predict(x_test)
plt.scatter(y_test,prediction)
```

Out[18]: <matplotlib.collections.PathCollection at 0x1bcd4ec4850>



In [19]:
```
print(lr.score(x_test,y_test))
```

0.9834497604361685

In [20]:
```
from sklearn.linear_model import ElasticNet
en = ElasticNet()
en.fit(x_train,y_train)
```

Out[20]: ElasticNet()

In [22]:
```
print(en.coef_)
```

```
[ 0.96247457 -0.98880091 -0.96745331 -1.07315863  0.14249777  0.29113032
  0.20527044  0.53334414  0.83086286 -0.80897327]
```

In [23]:
```
print(en.intercept_)
```

-77.66928279920944

In [24]:
```

print(en.predict(x_test))
```

```
[2302.96451709  828.03065634 3465.00024205 1630.93369163 1253.87178545
 2189.62491115   87.51290318 2897.43851731 3951.92022851  749.33686708
  839.74491859 3375.48440175  665.96485931  354.54019749  637.19830644
 1630.93369163 2988.90677117 5761.82402806 1090.35593276 2097.49452655
 1378.26128332   87.51290318  990.57133071 2901.57306491 2174.35509193
  724.07474235  504.12778871  196.9834181  1648.80956723 1381.77894624
 2183.43072187 1284.02538199 1710.70764902  735.18545716 3213.11707032
 1428.71293659]
```

In [25]:
```
print(en.score(x_test,y_test))
```

```
0.9834550038936513
```

In [26]:
```
from sklearn import metrics
```

In [27]:
```

print("Mean Absolytre Error:",metrics.mean_absolute_error(y_test,prediction))
```

```
Mean Absolytre Error: 89.92502129480482
```

In [28]:
```
print("Mean Squared Error:",metrics.mean_squared_error(y_test,prediction))
```

```
Mean Squared Error: 27006.849321501835
```

In [29]:
```
print("Root Mean Squared Error:",metrics.mean_squared_error(y_test,prediction)
```

```
Root Mean Squared Error: 27006.849321501835
```