DATA SCIENCE LAB – BIVARIATE ANALYSIS

1.SUV DATASET

User ID	Gender	Age	Estimated	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1
15714658	Male	48	41000	1
15599081	Female	45	22000	1
15705113	Male	46	23000	1
15631159	Male	47	20000	1

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import math
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

```
data = pd.read_csv("suv_data.csv")
data.head(5)
```

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

```
data.shape
```

```
(400, 5)
```

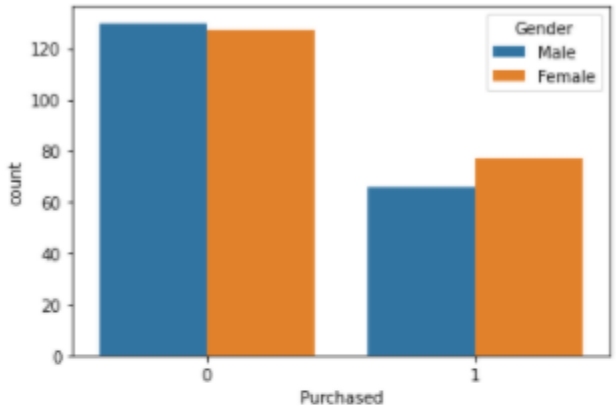
```
data.isnull().sum()
```

```
User ID            0
Gender             0
Age                0
EstimatedSalary    0
Purchased          0
dtype: int64
```

Plot with required axes

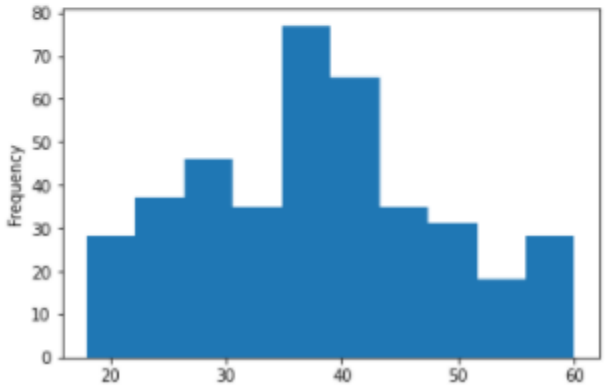
```
sns.countplot(x="Purchased", hue = "Gender", data=data)
```

```
<AxesSubplot:xlabel='Purchased', ylabel='count'>
```




```
data["Age"].plot.hist()
```

```
<AxesSubplot:ylabel='Frequency'>
```



User-id column is dropped

```
data.drop("User ID", axis=1 ,inplace=True)
data
```

	Gender	Age	EstimatedSalary	Purchased
0	Male	19	19000	0
1	Male	35	20000	0
2	Female	26	43000	0
3	Female	27	57000	0
4	Male	19	76000	0
...	...	...	...	...
395	Female	46	41000	1
396	Male	51	23000	1
397	Female	50	20000	1
398	Male	36	33000	0
399	Female	49	36000	1

400 rows × 4 columns

Getting the dummies of gender attribute as True and False/ 0 and 1

```
sex = pd.get_dummies(data["Gender"], drop_first=True)
sex
```

	Male
0	1
1	1
2	0
3	0
4	1
...	...
395	0
396	1
397	0
398	1
399	0

400 rows × 1 columns

Merge sex with dataset

```
data =pd.concat([data,sex], axis=1)
data
```

	Gender	Age	EstimatedSalary	Purchased	Male
0	Male	19	19000	0	1
1	Male	35	20000	0	1
2	Female	26	43000	0	0
3	Female	27	57000	0	0
4	Male	19	76000	0	1
...	...	...	...	...	...
395	Female	46	41000	1	0
396	Male	51	23000	1	1
397	Female	50	20000	1	0
398	Male	36	33000	0	1
399	Female	49	36000	1	0

400 rows × 5 columns

Drop original gender attribute after replacing with dummies

```
data.drop("Gender",axis=1 ,inplace=True)
data
```

	Age	EstimatedSalary	Purchased	Male
0	19	19000	0	1
1	35	20000	0	1
2	26	43000	0	0
3	27	57000	0	0
4	19	76000	0	1
...	...	...	...	...
395	46	41000	1	0
396	51	23000	1	1
397	50	20000	1	0
398	36	33000	0	1
399	49	36000	1	0

400 rows × 4 columns

Test and train on Purchase column

```
x=data.drop("Purchased", axis=1)
y=data["Purchased"]
```

```
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2, random_state=42)
```

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

```
model = LogisticRegression()
model.fit(x_train, y_train)
```

```
LogisticRegression()
```

```
from sklearn.metrics import accuracy_score
predic = model.predict(x_test)
accuracy_score(y_test, predic)
```

```
0.8875
```

2. Job Change of Data Scientists

enrollee_	city	city_deve	gender	relevent_	enrolled_	education	major_dis	experienc	company_	company_	last_new_	training_hours
32403	city_41	0.827	Male	Has releve	Full time (	Graduate	STEM	9	<10		1	21
9858	city_103	0.92	Female	Has releve	no_enroll	Graduate	STEM	5		Pvt Ltd	1	98
31806	city_21	0.624	Male	No releve	no_enroll	High School		<1		Pvt Ltd	never	15
27385	city_13	0.827	Male	Has releve	no_enroll	Masters	STEM	11	Oct-49	Pvt Ltd	1	39
27724	city_103	0.92	Male	Has releve	no_enroll	Graduate	STEM	>20	10000+	Pvt Ltd	>4	72
217	city_23	0.899	Male	No releve	Part time	Masters	STEM	10			2	12
21465	city_21	0.624		Has releve	no_enroll	Graduate	STEM	<1	100-500	Pvt Ltd	1	11
27302	city_160	0.92	Female	Has releve	no_enroll	Graduate	STEM	>20			>4	81
12994	city_173	0.878	Male	Has releve	no_enroll	Graduate	STEM	14			4	2
16287	city_21	0.624	Male	Has releve	Full time (	Graduate		3	50-99	Funded St	1	4
10856	city_103	0.92	Male	Has releve	no_enroll	Masters	Other	>20			>4	196
9272	city_90	0.698	Male	Has releve	no_enroll	Graduate	STEM	20	Oct-49	Pvt Ltd	2	51
14249	city_46	0.762	Male	Has releve	no_enroll	Graduate	STEM	8	100-500	Other	never	48
24372	city_98	0.949		Has releve	no_enroll	Masters	STEM	4	100-500	Pvt Ltd	1	134
14070	city_103	0.92		No releve	no_enroll	Graduate	STEM	5			never	10
24914	city_21	0.624		Has releve	Full time (	Graduate	STEM	13	1000-4999	Pvt Ltd	1	125
7865	city_21	0.624	Male	Has releve	no_enroll	Masters	STEM	4	100-500	Pvt Ltd	1	4
7463	city_13	0.827	Male	Has releve	no_enroll	Masters	Business [	2	50-99	Pvt Ltd	1	31
21514	city_21	0.624		Has releve	no_enroll	Graduate	STEM	6		Pvt Ltd	4	23
29033	city_21	0.624	Male	No releve	Full time course			2			never	110
15359	city_103	0.92		No releve	Full time (	Graduate	STEM	2			never	74
16001	city_103	0.92		Has releve	no_enroll	Graduate	STEM	7	10000+		1	44
25202	city_21	0.624	Male	Has releve	no_enroll	Graduate	STEM	6	1000-4999	Pvt Ltd	3	33
5058	city_103	0.92	Male	No releve	Full time (	Graduate	STEM	1			1	81
23570	city_118	0.722	Male	Has releve	no_enroll	Graduate	STEM	19	100-500	Pvt Ltd	>4	19
19139	city_103	0.92	Female	Has releve	Part time	Graduate	STEM	15		Public Sec	>4	48

Import dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import math
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

```
data = pd.read_csv("aug_test.csv")
data.head(5)
```

	enrollee_id	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size
0	32403	city_41	0.827	Male	Has relevent experience	Full time course	Graduate	STEM	9	<10
1	9858	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM	5	NaN
2	31806	city_21	0.624	Male	No relevent experience	no_enrollment	High School	NaN	<1	NaN
3	27385	city_13	0.827	Male	Has relevent experience	no_enrollment	Masters	STEM	11	10/49
4	27724	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	10000+

```
data.shape
```

```
(2129, 13)
```

```
data.isnull().sum()
```

```
enrollee_id                0
city                       0
city_development_index     0
```

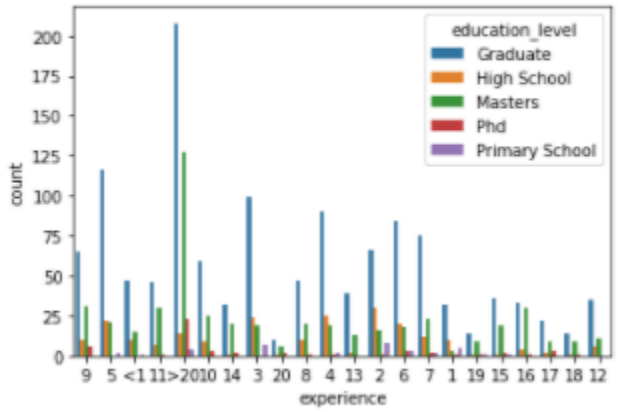
```
data.isnull().sum()
```

```
enrollee_id                0
city                       0
city_development_index     0
gender                   508
relevent_experience        0
enrolled_university       31
education_level           52
major_discipline         312
experience                 5
company_size             622
company_type             634
last_new_job              40
training_hours             0
dtype: int64
```

Plot over experience and education level

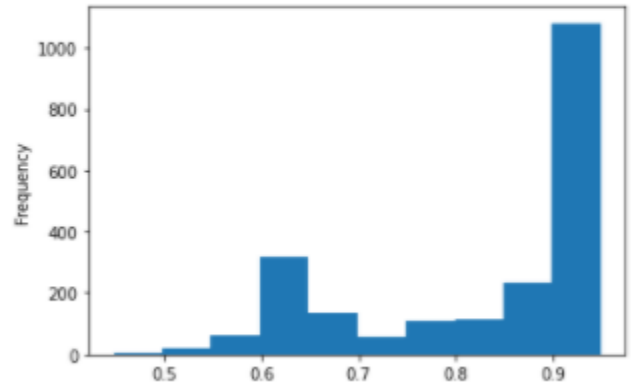
```
sns.countplot(x="experience", hue = "education_level", data=data)
```

```
<AxesSubplot:xlabel='experience', ylabel='count'>
```




```
data["city_development_index"].plot.hist()
```

```
<AxesSubplot:ylabel='Frequency'>
```



Drop enrollee_id attribute

```
data.drop("enrollee_id", axis=1 ,inplace=True)
data
```

	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size	co
0	city_41	0.827	Male	Has relevent experience	Full time course	Graduate	STEM	9	<10	
1	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM	5	NaN	
2	city_21	0.624	Male	No relevent experience	no_enrollment	High School	NaN	<1	NaN	
3	city_13	0.827	Male	Has relevent experience	no_enrollment	Masters	STEM	11	10/49	
4	city_103	0.920	Male	Has relevent experience	no_enrollment	Graduate	STEM	>20	10000+	
...	...	...	...	...	...	...	...	...	...	
2124	city_103	0.920	Male	No relevent experience	no_enrollment	Graduate	Humanities	16	NaN	
2125	city_136	0.897	Male	Has relevent experience	no_enrollment	Masters	STEM	18	NaN	
2126	city_100	0.887	Male	No relevent experience	no_enrollment	Primary School	NaN	3	NaN	
2127	city_102	0.804	Male	Has relevent experience	Full time course	High School	NaN	7	100-500	
2128	city_102	0.804	Male	Has relevent experience	no_enrollment	Masters	STEM	15	10000+	

2129 rows × 12 columns

Dummying gender same as in suv data

```
sex = pd.get_dummies(data["gender"], drop_first=True)
sex
```

	Male	Other
0	1	0
1	0	0
2	1	0
3	1	0
4	1	0
...	...	...
2124	1	0
2125	1	0
2126	1	0
2127	1	0
2128	1	0

2129 rows × 2 columns

```
data =pd.concat([data,sex], axis=1)
data
```

	city	city_development_index	gender	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size	company
0	city_41	0.827	Male	Has relevent experience	Full time course	Graduate	STEM	9	<10	
1	city_103	0.920	Female	Has relevent experience	no_enrollment	Graduate	STEM	5	NaN	F
2	city_21	0.624	Male	No relevent experience	no_enrollment	High School	NaN	<1	NaN	F
3	city_13	0.827	Male	Has relevent experience	no_enrollment	Masters	STEM	11	10/49	F

```
data.drop("gender",axis=1 ,inplace=True)
data
```

	city	city_development_index	relevent_experience	enrolled_university	education_level	major_discipline	experience	company_size	company_type
0	city_41	0.827	Has relevent experience	Full time course	Graduate	STEM	9	<10	NaN
1	city_103	0.920	Has relevent experience	no_enrollment	Graduate	STEM	5	NaN	Pvt Ltd
2	city_21	0.624	No relevent experience	no_enrollment	High School	NaN	<1	NaN	Pvt Ltd
3	city_13	0.827	Has relevent experience	no_enrollment	Masters	STEM	11	10/49	Pvt Ltd
4	city_103	0.920	Has relevent experience	no_enrollment	Graduate	STEM	>20	10000+	Pvt Ltd
...	...	...	...	...	...	...	...	...	...
2124	city_103	0.920	No relevent experience	no_enrollment	Graduate	Humanities	16	NaN	Public Sector
2125	city_136	0.897	Has relevent experience	no_enrollment	Masters	STEM	18	NaN	NaN
2126	city_100	0.887	No relevent experience	no_enrollment	Primary School	NaN	3	NaN	Pvt Ltd
2127	city_102	0.804	Has relevent experience	Full time course	High School	NaN	7	100-500	Public Sector
2128	city_102	0.804	Has relevent experience	no_enrollment	Masters	STEM	15	10000+	Pvt Ltd

2129 rows × 13 columns

Drop education_level attribute as it consists of string values

```
data.drop("education_level", axis=1,inplace=True)
data
```

	city_development_index	training_hours	Male	Other
0	0.827	21	1	0
1	0.920	98	0	0
2	0.624	15	1	0
3	0.827	39	1	0
4	0.920	72	1	0
...	...	...	...	...
2124	0.920	15	1	0
2125	0.897	30	1	0
2126	0.887	18	1	0
2127	0.804	84	1	0
2128	0.804	11	1	0

2129 rows × 4 columns

Training and testing over Training Hours

```
x=data.drop("training_hours", axis=1)
y=data["training_hours"]
```

```
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2, random_state=42)
```

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

```
model = LogisticRegression()
model.fit(x_train, y_train)
```

```
LogisticRegression()
```

```
from sklearn.metrics import accuracy_score
predic = model.predict(x_test)
accuracy_score(y_test, predic)
```

```
0.014084507042253521
```