Date-15-12-23

In [3]:
```
# import the packages
# read the data

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
file_path="C:\\Users\\kurre\\OneDrive\\Documents\\Naresh IT\\datafiles\\Vis
visa_df=pd.read_csv(file_path)
visa_df
```

Out[4]:

	case_id	continent	education_of_employee	has_job_experience	requires_job_traini
0	EZYV01	Asia	High School	N	
1	EZYV02	Asia	Master's	Y	
2	EZYV03	Asia	Bachelor's	N	
3	EZYV04	Asia	Bachelor's	N	
4	EZYV05	Africa	Master's	Y	
...	...	...	...	...	
25475	EZYV25476	Asia	Bachelor's	Y	
25476	EZYV25477	Asia	High School	Y	
25477	EZYV25478	Asia	Master's	Y	
25478	EZYV25479	Asia	Master's	Y	
25479	EZYV25480	Asia	Bachelor's	Y	

25480 rows × 12 columns

We draw two categorical columns analysis

In [5]:
```
# Continent colums value counts

visa_df['continent'].value_counts()
```

Out[5]:
```
continent
Asia             16861
Europe            3732
North America     3292
South America      852
Africa             551
Oceania            192
Name: count, dtype: int64
```

In [6]:
```
visa_df['case_status'].value_counts()
```

Out[6]:
```
case_status
Certified    17018
Denied        8462
Name: count, dtype: int64
```

In [ ]:
```
#Q) out of all Asian applicants how many got Visa
#   Out of all Europe applicants how many got Visa
```

In [7]:
```
con1=visa_df['continent']=='Asia'
con2=visa_df['case_status']=='Certified'
con=con1&con2
len(visa_df[con])
```

Out[7]:
```
11012
```

In [8]:
```
visa_df['continent'].unique()
visa_df['continent'].value_counts().keys()
```

Out[8]:
```
Index(['Asia', 'Europe', 'North America', 'South America', 'Africa',
       'Oceania'],
      dtype='object', name='continent')
```

In [9]:
```
# Generalised
lables=visa_df['continent'].unique()
certified_count=[]
denied_count=[]
for i in lables:
    con1=visa_df['continent']==i
    con2=visa_df['case_status']=='Certified'
    con3=visa_df['case_status']=='Denied'
    certified_count.append(len(visa_df[con1&con2]))
    denied_count.append(len(visa_df[con1&con3]))


pd.DataFrame(zip(lables,certified_count,denied_count),
             columns=['continent','certified','denied'])
```

Out[9]:

	continent	certified	denied
0	Asia	11012	5849
1	Africa	397	154
2	North America	2037	1255
3	Europe	2957	775
4	South America	493	359
5	Oceania	122	70

In [10]:

```
pd.DataFrame(zip(lables,certified_count,denied_count),
                          columns=['continent','certified','denied']).set_index('contine
```

Out[10]:

	certified	denied
continent		
Asia	11012	5849
Africa	397	154
North America	2037	1255
Europe	2957	775
South America	493	359
Oceania	122	70

$pd.crosstab$

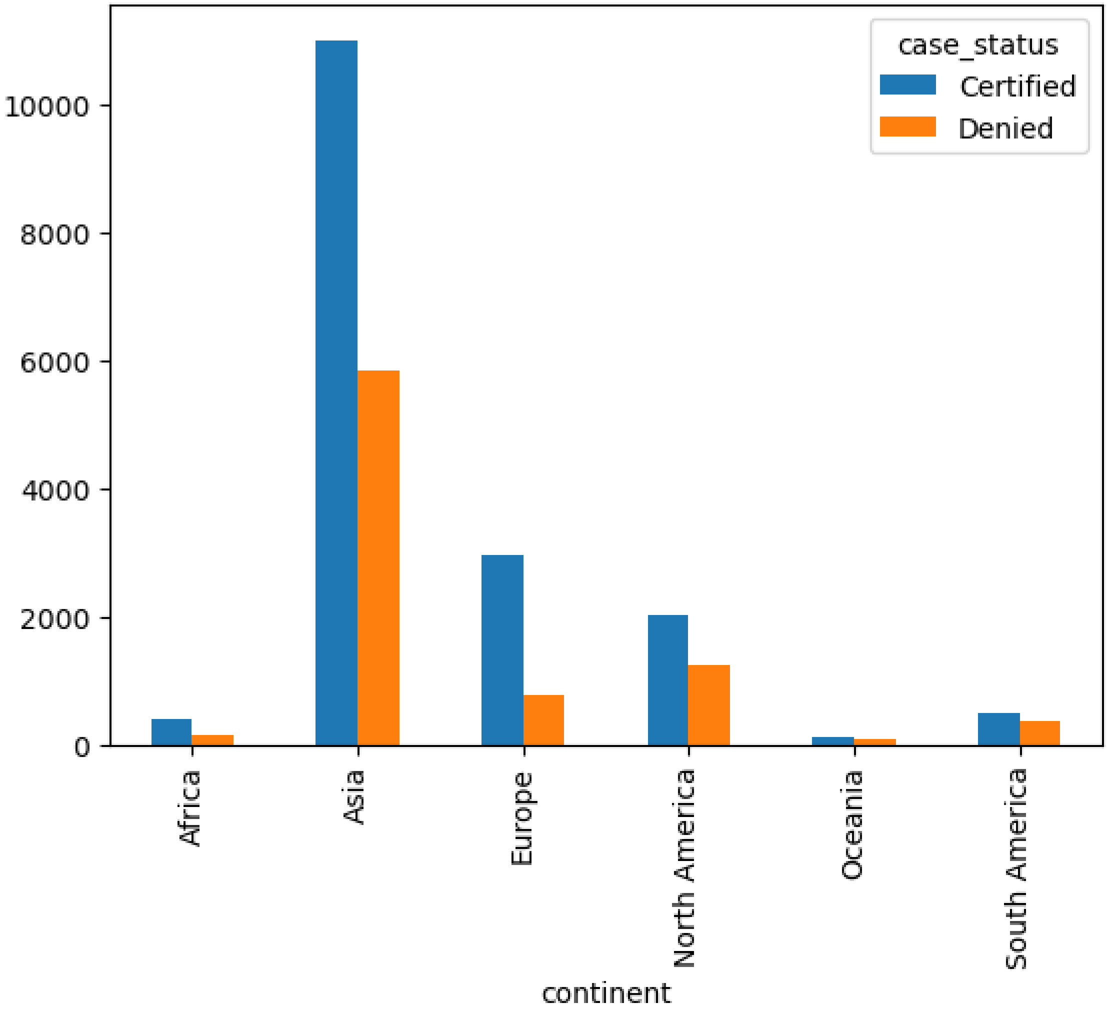
In [11]:

```
col1=visa_df['continent']
col2=visa_df['case_status']

result1=pd.crosstab(col1,col2)
result1
```

Out[11]:

case_status	Certified	Denied
continent		
Africa	397	154
Asia	11012	5849
Europe	2957	775
North America	2037	1255
Oceania	122	70
South America	493	359

```
In [13]: result1.plot(kind='bar')
         plt.show()
```



We repeated multiple columns

In [14]:

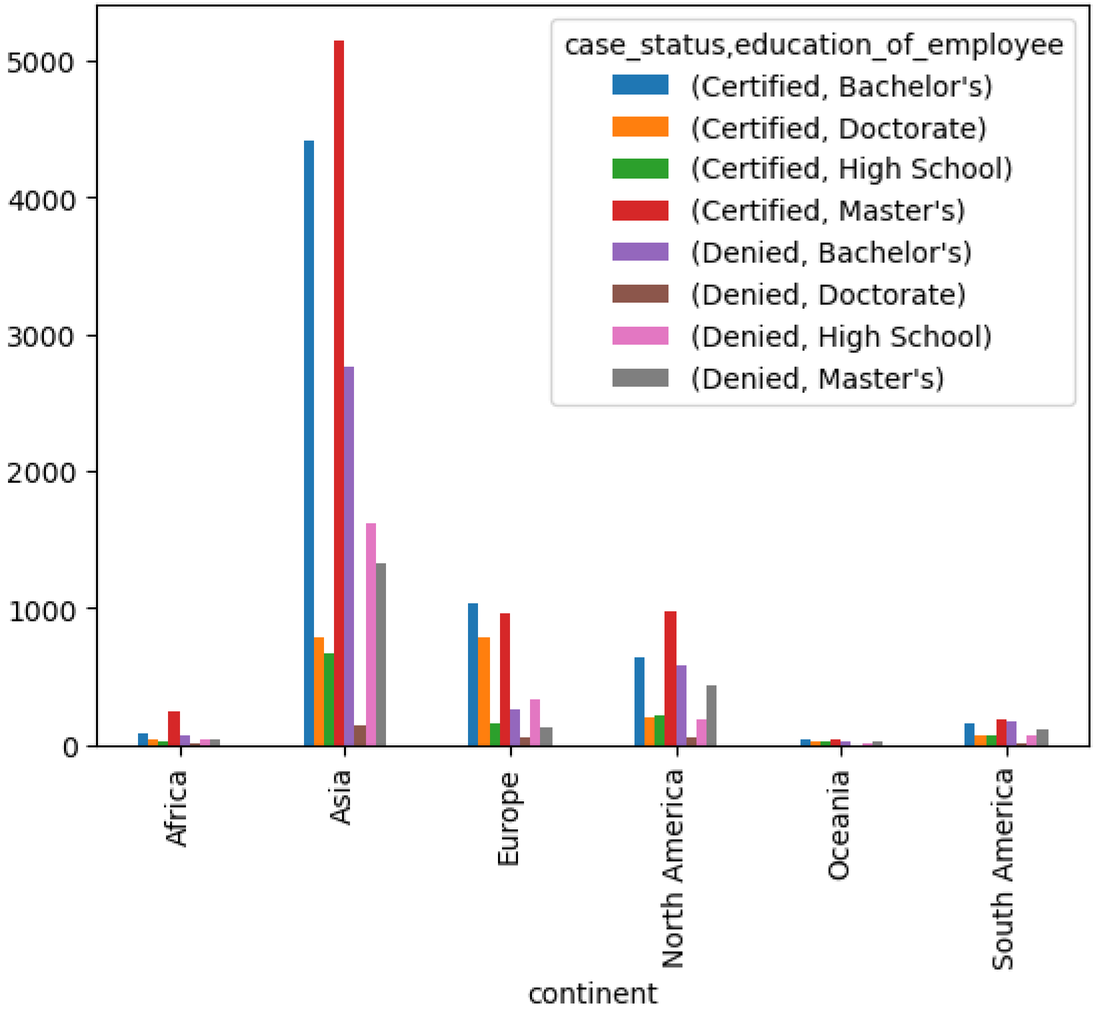
```
#Continent
#Education
#Case status

col1=visa_df['continent']
col2=visa_df['case_status']
col3=visa_df['education_of_employee']
col=[col2,col3] # values
result2=pd.crosstab(col1,col)
result2
```

Out[14]:

case_status	Certified						
education_of_employee	Bachelor's	Doctorate	High School	Master's	Bachelor's	Doctorate	Hig Scho
continent							
Africa	81	43	23	250	62	11	4
Asia	4407	780	676	5149	2761	143	16
Europe	1040	788	162	967	259	58	3
North America	641	207	210	979	584	51	1
Oceania	38	19	19	46	28	3	
South America	160	75	74	184	173	14	6

In [15]:
```
result2.plot(kind='bar')
```

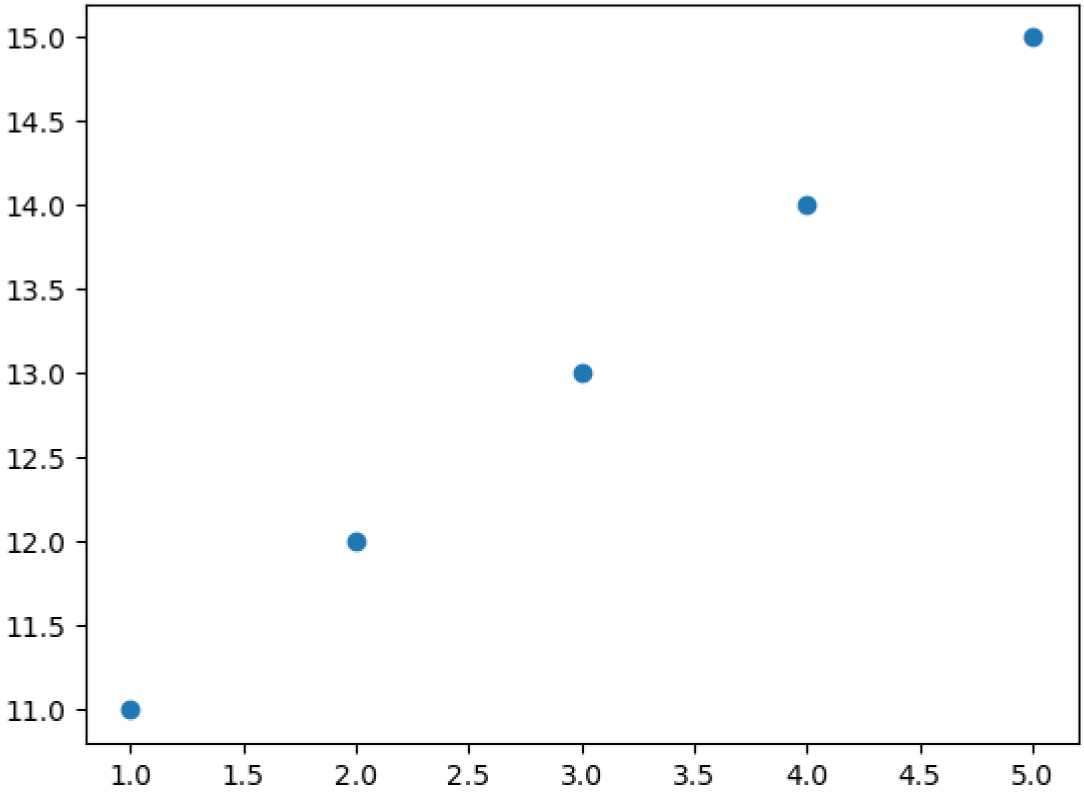
Out[15]:
```
<Axes: xlabel='continent'>
```



We draw two numerical columns analysis

Numerical vs Numerical

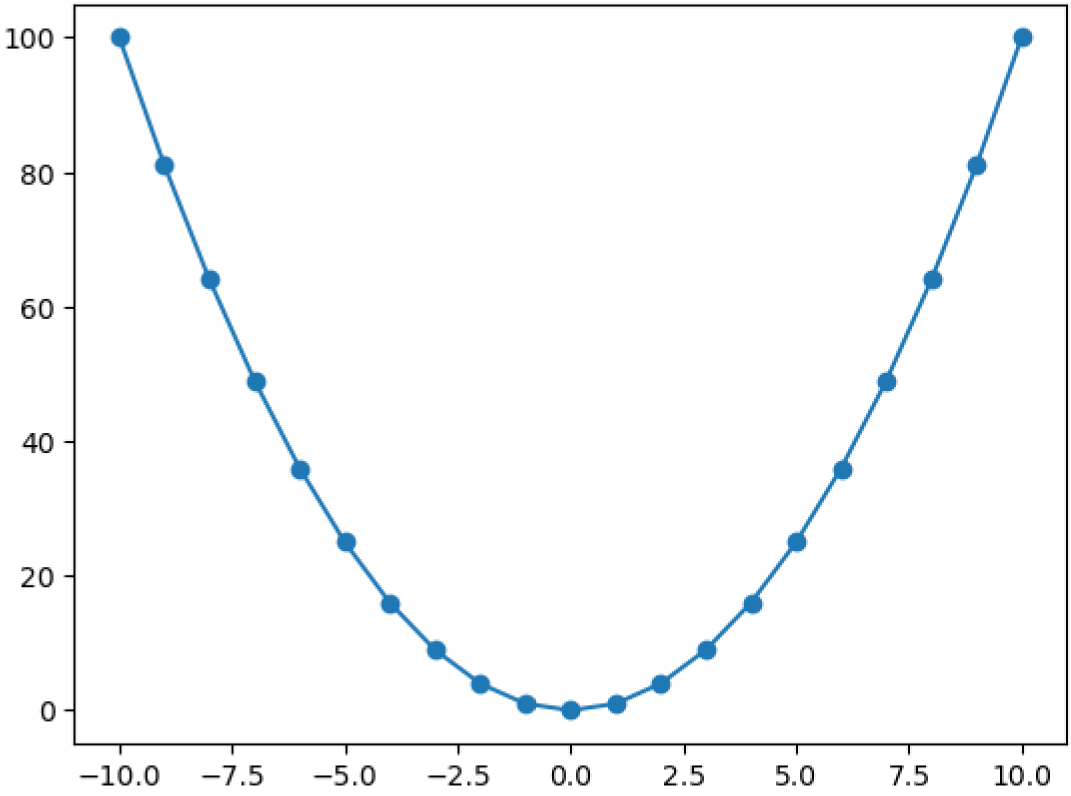
```
In [16]: x=[1,2,3,4,5]
         y=[11,12,13,14,15]

         #(1,11),(2,12),(3,13),(4,14),(5,15)
         plt.scatter(x,y)

Out[16]: <matplotlib.collections.PathCollection at 0x22afb580c90>
```



```
In [17]: x=[i for i in range(-10,11)]
         y=[i*i for i in x]
         x

Out[17]: [-10, -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 1
         0]

In [18]: y

Out[18]: [100, 81, 64, 49, 36, 25, 16, 9, 4, 1, 0, 1, 4, 9, 16, 25, 36, 49, 64, 81,
         100]
```

In [19]:
```
plt.scatter(x,y)
plt.plot(x,y)
```

Out[19]:
```
[<matplotlib.lines.Line2D at 0x22afb909f50>]
```



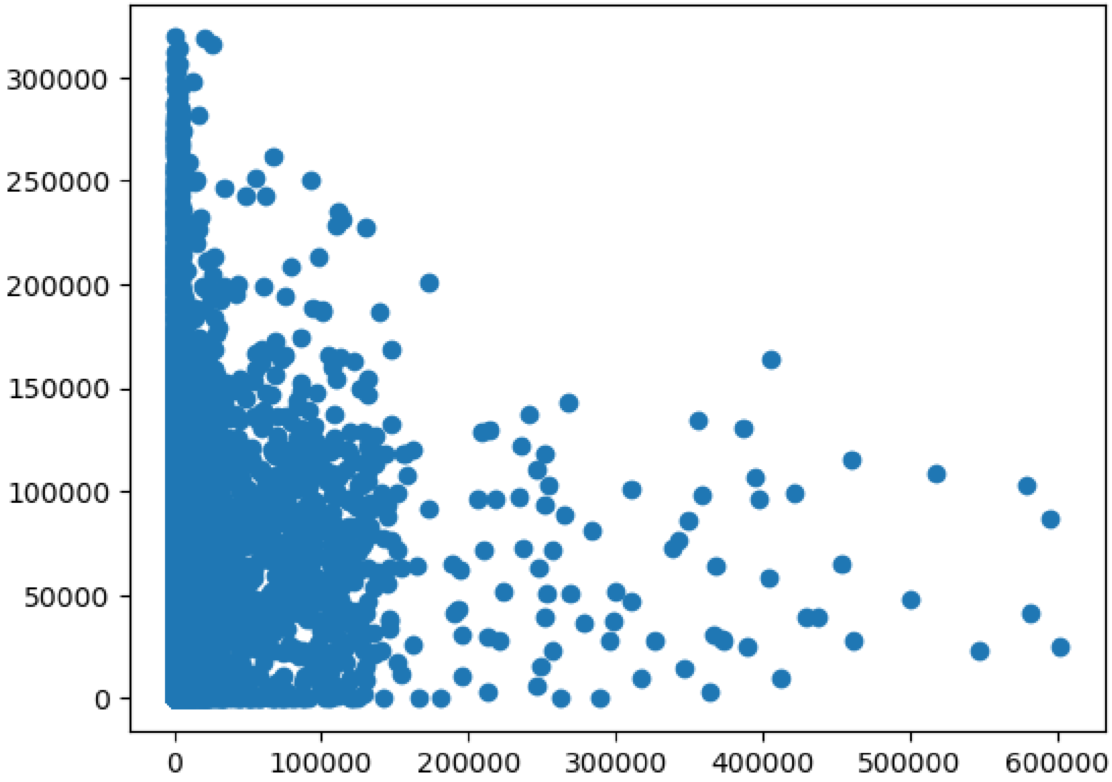
- Scatter plots for only numerical analysis
- Scatter plots provides an idea , both variables are related or not related
- Postivie relation
    - Increase in the curve
- Negative relation
    - Decrease in the curve
- No realtion
    - Neither increase nor Decrease

In [20]:
```
dtypes=dict(visa_df.dtypes)
num=[i for i in dtypes if dtypes[i]!='O']
num
```

Out[20]:
```
['no_of_employees', 'yr_of_estab', 'prevailing_wage']
```

In [21]:
```
col1=visa_df['no_of_employees']
col2=visa_df['prevailing_wage']
plt.scatter(col1,col2)
```

Out[21]:
```
<matplotlib.collections.PathCollection at 0x22afb901050>
```



In [22]:
```
#Covariance-matrix

#How many numerical variables are there : 3

#              no_employee   yr     wage

#no_employee    var          cov    cov

#yr             cov          var    cov

#age            cov          cov     var
```

$correlation-coeffiecinet$

- Denoted with r
- r range from -1 to 1
- postive relation range = (0,1]
- negative relation range = [-1,0)
- no relation = 0

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

$Corr()$

In [23]:
```
visa_df.corr(numeric_only=True)  # applicable for you need to see numeric_o

# in the data frame we have both cat and numerical column
# correlation applicable for only numerical column
# Explicitly mention numeric= True

# If people has pandas old version
# they dont have numeric_only argument
# for them  visa_df.corr() works
```

Out[23]:

	no_of_employees	yr_of_estab	prevailing_wage
no_of_employees	1.000000	-0.017770	-0.009523
yr_of_estab	-0.017770	1.000000	0.012342
prevailing_wage	-0.009523	0.012342	1.000000

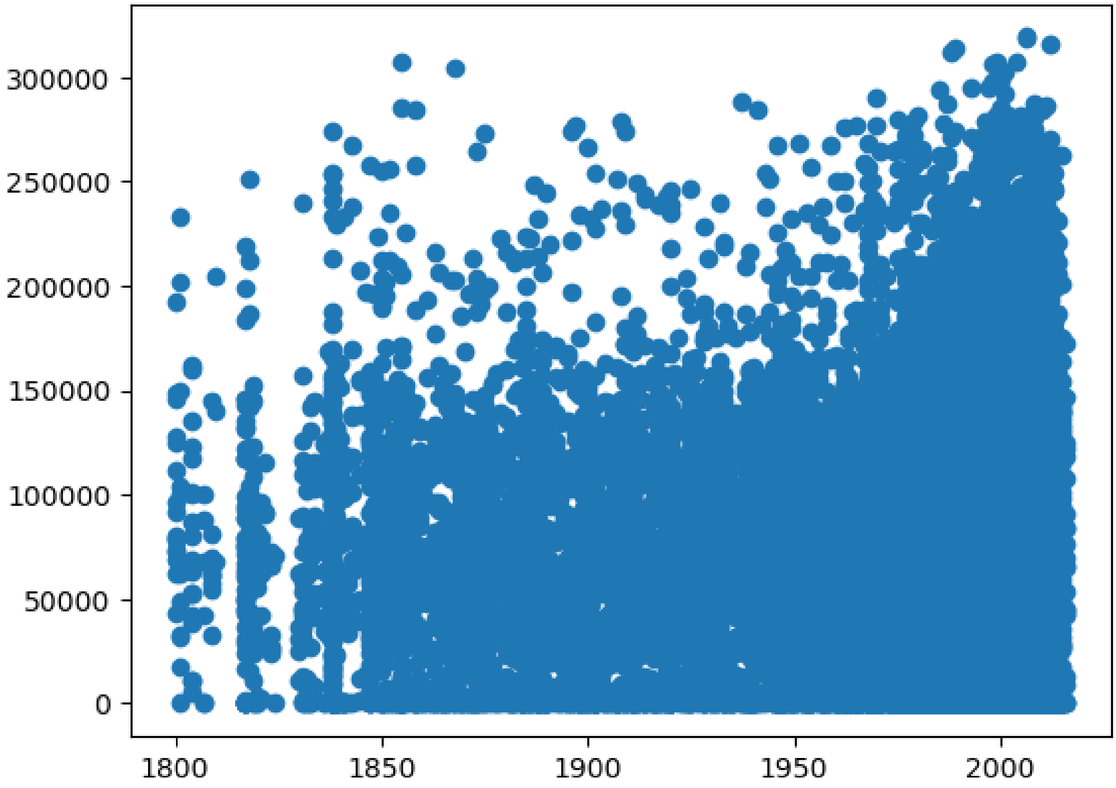
In [25]:
```
pd.__version__  # double underscore
```

Out[25]:
```
'2.0.3'
```

In [ ]:
```
#pip unisntall pandas

#pip install pandas==2.0.3
```

In [26]:
```
plt.scatter(visa_df['yr_of_estab'],visa_df['prevailing_wage'])
```

Out[26]:
```
<matplotlib.collections.PathCollection at 0x22afbcdd690>
```



- EDA session-1
  - Read the data
  - Create the data frame using list
  - Create the data frame using dictionary
  - How to save the dataframe
  - How to add new column
  - How to drop new column
- EDA session-2:
  - shape/size
  - columns/dtypes
  - head/tail
  - take/loc/iloc
  - isnull/len
- EDA session-3 Categorical data analysis
  - How to read a column
  - unique/nunique
  - value counts
  - we created a frequncy table by our own skill
  - bar chart
  - pie chart
- EDA session -4 Numerical data analysis
  - How to read a column
  - statistical measurements
  - mean/median/count/max/min/std/25/50/75
  - describe function
  - using numpy we draw measurements

        - Histogram
        - we checked the empiricle rule
    - EDA session-5 Outlier analysis
        - We draw box plot
        - we implemented how to find outlier
        - we remove the outliers
        - we imputed with median
        - np.where
    - EDA session-6: Bi variate and multivariate analysis
        - we draw two cat columns analysis
        - we implemented by our own skill
        - pd.crosstab
        - draw the plots
        - we repeated multiple columns
        - for two numerical columns plt.scatter
        - correlation data.corr
        - matrix
        - heatmap

In [ ]:

In [ ]:

Date-18-12-2023

In [ ]:

```
# read the packages
# read the data
```

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
file_path="C:\\Users\\kurre\\OneDrive\\Documents\\Naresh IT\\datafiles\\Vis
visa_df=pd.read_csv(file_path)
visa_df
```

Out[2]:

	case_id	continent	education_of_employee	has_job_experience	requires_job_traini
0	EZYV01	Asia	High School	N	
1	EZYV02	Asia	Master's	Y	
2	EZYV03	Asia	Bachelor's	N	
3	EZYV04	Asia	Bachelor's	N	
4	EZYV05	Africa	Master's	Y	
...	...	...	...	...	
25475	EZYV25476	Asia	Bachelor's	Y	
25476	EZYV25477	Asia	High School	Y	
25477	EZYV25478	Asia	Master's	Y	
25478	EZYV25479	Asia	Master's	Y	
25479	EZYV25480	Asia	Bachelor's	Y	

25480 rows × 12 columns

In [5]:
```
# corr function
visa_df.corr(numeric_only=True)
```

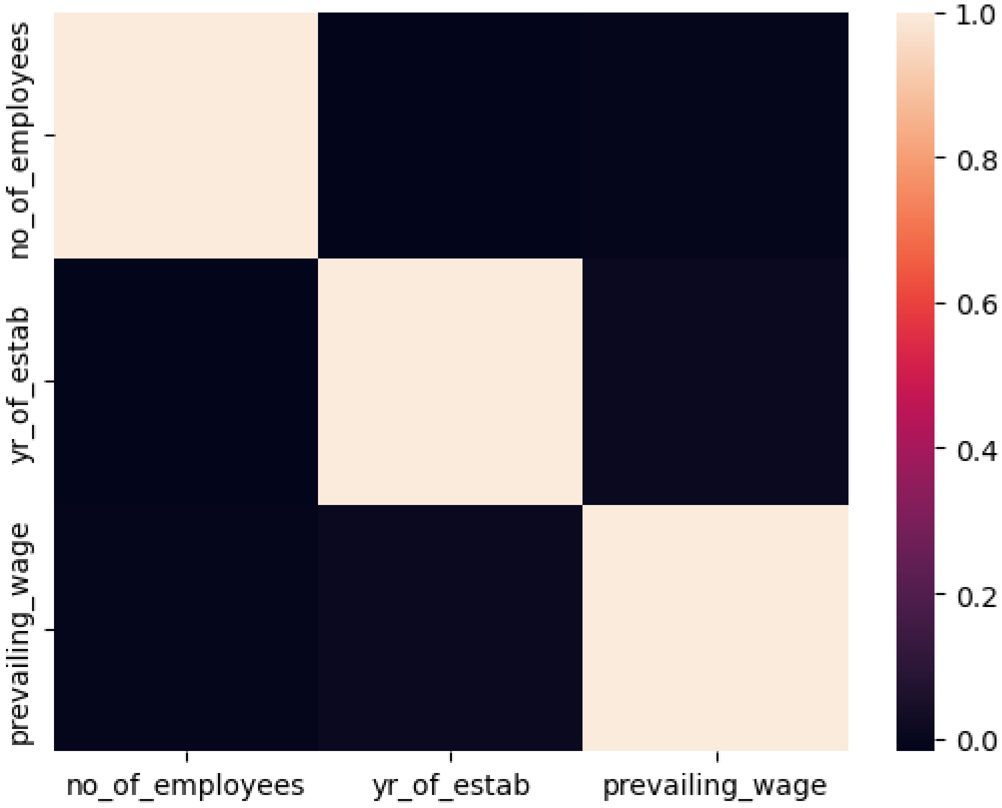
Out[5]:

	no_of_employees	yr_of_estab	prevailing_wage
no_of_employees	1.000000	-0.017770	-0.009523
yr_of_estab	-0.017770	1.000000	0.012342
prevailing_wage	-0.009523	0.012342	1.000000

In [ ]:
```
# matrix
# showing values in a matrix
# showing values in a picture: Heatmap
```

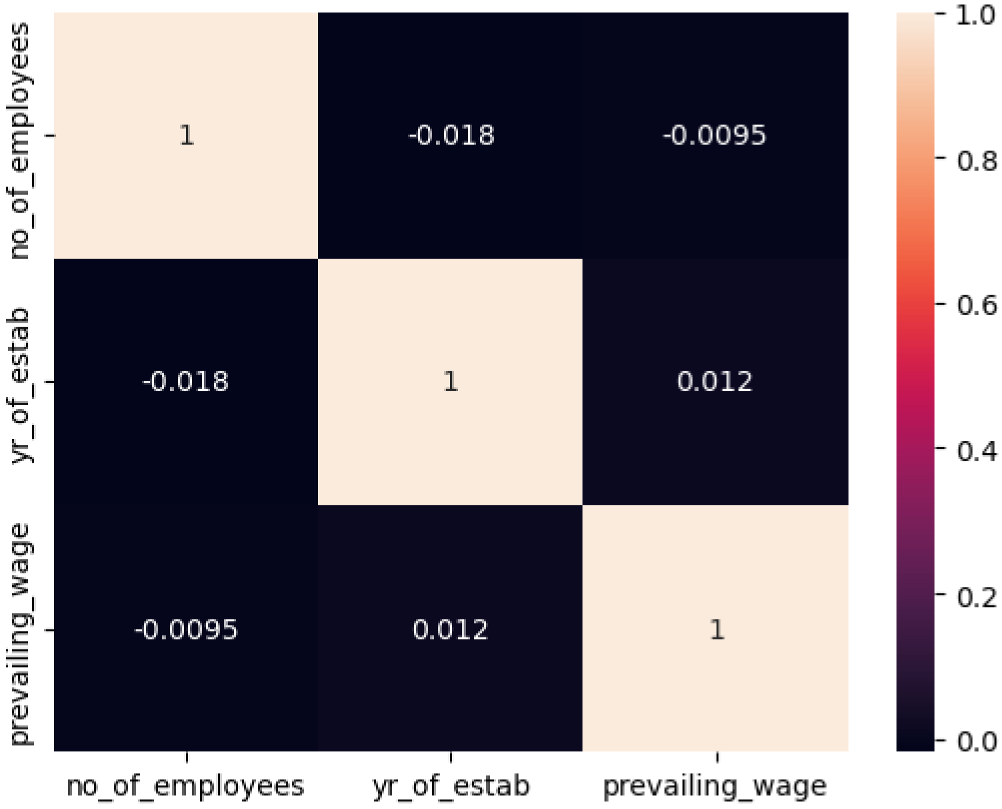
In [8]:
```
corr_data=visa_df.corr(numeric_only=True)
sns.heatmap(corr_data)
```

Out[8]:
```
<Axes: >
```



In [9]:
```
corr_data=visa_df.corr(numeric_only=True)
sns.heatmap(corr_data,annot=True)        # for see the value use 'annot'
```

Out[9]:
```
<Axes: >
```



In [10]:

```
# wine quality dataset

file_path1="C:\\Users\\kurre\\OneDrive\\Documents\\Naresh IT\\datafiles\\wi
wine_df=pd.read_csv(file_path1)
wine_df
```

Out[10]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
3193	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75
3194	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3195	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71
3196	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3197	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66

3198 rows × 12 columns

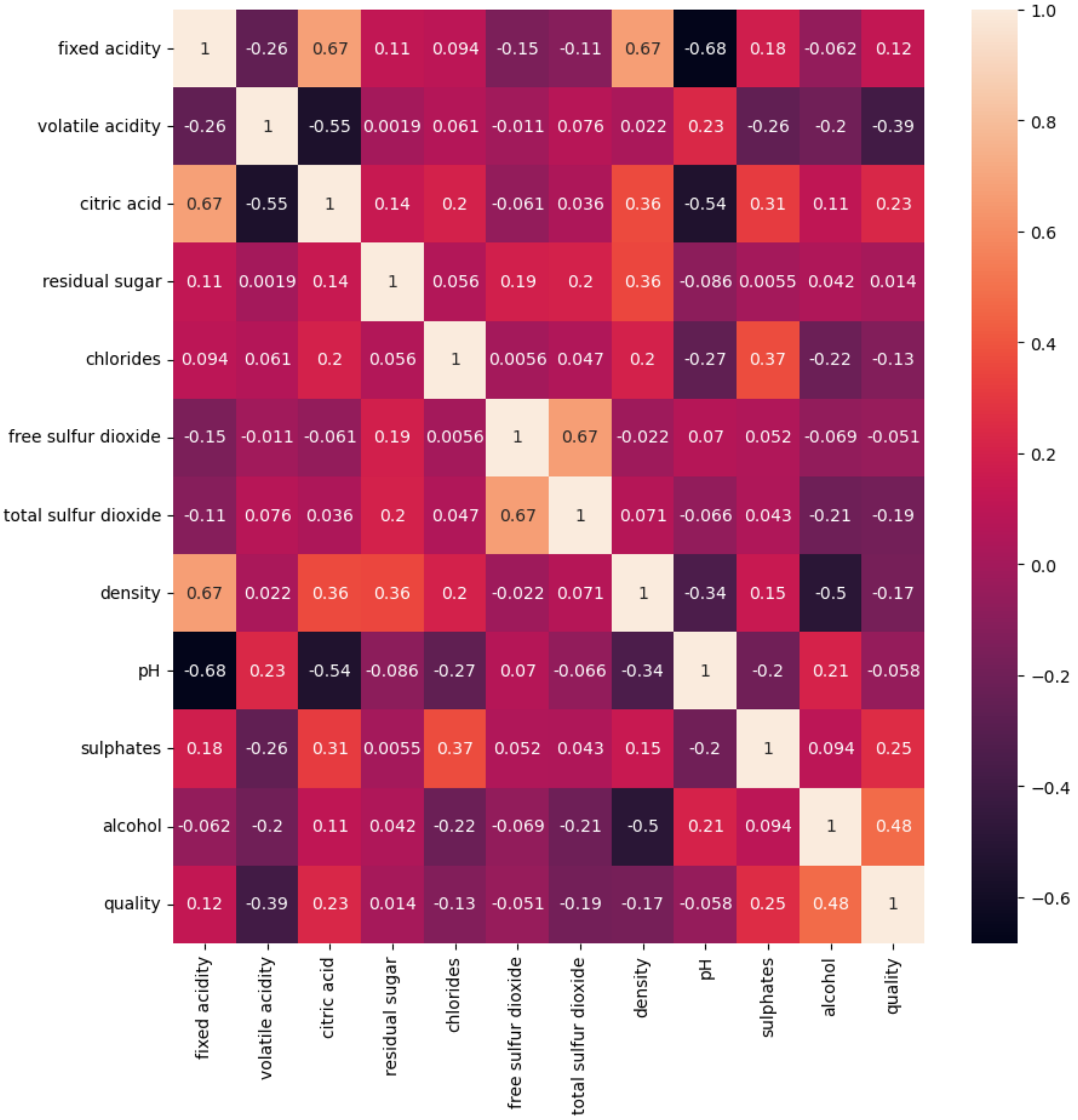
In [12]:
```
wine_df.corr()
```

Out[12]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	densi
fixed acidity	1.000000	-0.256131	0.671703	0.114777	0.093705	-0.153794	-0.113181	0.6680
volatile acidity	-0.256131	1.000000	-0.552496	0.001918	0.061298	-0.010504	0.076470	0.0220
citric acid	0.671703	-0.552496	1.000000	0.143577	0.203823	-0.060978	0.035533	0.3649
residual sugar	0.114777	0.001918	0.143577	1.000000	0.055610	0.187049	0.203028	0.3552
chlorides	0.093705	0.061298	0.203823	0.055610	1.000000	0.005562	0.047400	0.2006
free sulfur dioxide	-0.153794	-0.010504	-0.060978	0.187049	0.005562	1.000000	0.667666	-0.0219
total sulfur dioxide	-0.113181	0.076470	0.035533	0.203028	0.047400	0.667666	1.000000	0.0712
density	0.668047	0.022026	0.364947	0.355283	0.200632	-0.021946	0.071269	1.0000
pH	-0.682978	0.234937	-0.541904	-0.085652	-0.265026	0.070377	-0.066495	-0.3416
sulphates	0.183006	-0.260987	0.312770	0.005527	0.371260	0.051658	0.042947	0.1485
alcohol	-0.061668	-0.202288	0.109903	0.042075	-0.221141	-0.069408	-0.205654	-0.4961
quality	0.124052	-0.390558	0.226373	0.013732	-0.128907	-0.050656	-0.185100	-0.1749

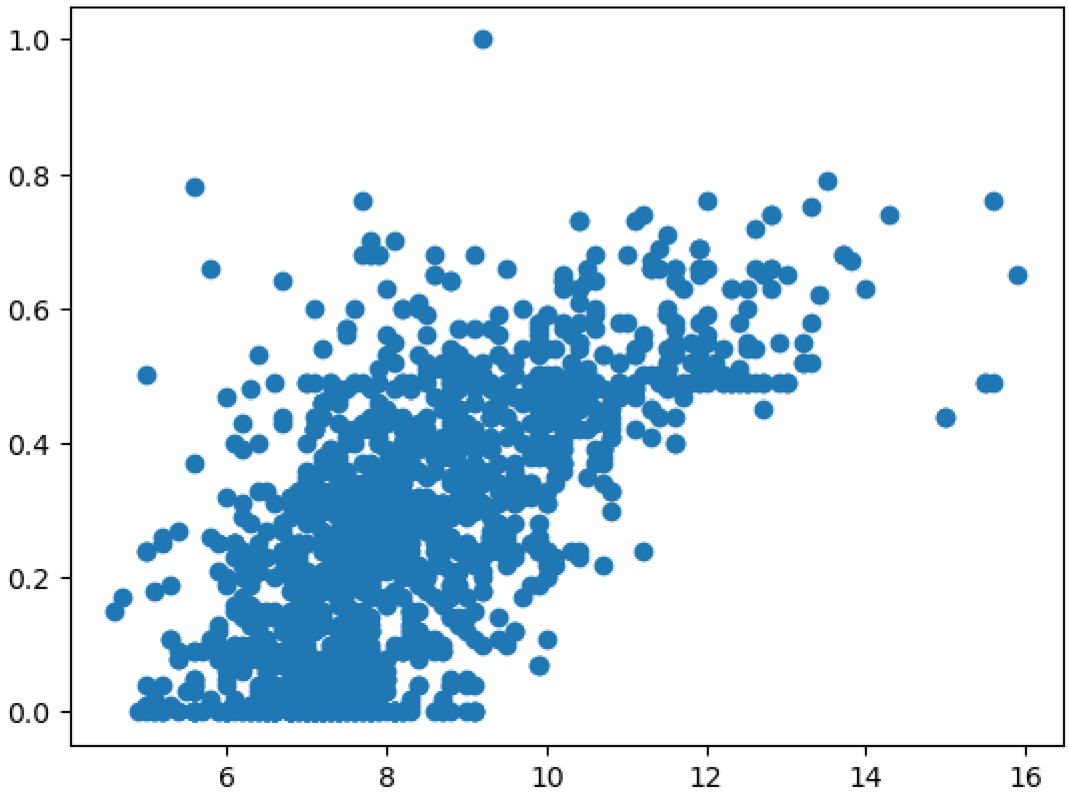
```
In [13]: plt.figure(figsize=(10,10))
         sns.heatmap(wine_df.corr(),annot=True)
```

```
Out[13]: <Axes: >
```



```
In [14]: plt.scatter(wine_df['fixed acidity'],wine_df['citric acid'])
```

```
Out[14]: <matplotlib.collections.PathCollection at 0x1d7f39fcd90>
```



```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```