In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

path='Visadataset.csv'
visa_df=pd.read_csv(path)
cat=visa_df.select_dtypes(include='object').columns
num=visa_df.select_dtypes(exclude='object').columns
```

In [2]:
```
# we have two columns
# continent and another columns is case_status
visa_df['continent'].value_counts()
```

Out[2]:
```
continent
Asia             16861
Europe            3732
North America     3292
South America      852
Africa             551
Oceania            192
Name: count, dtype: int64
```

In [3]:
```
visa_df['case_status'].value_counts()
```

Out[3]:
```
case_status
Certified    17018
Denied        8462
Name: count, dtype: int64
```

In [4]:
```
# how many asia students got certified
# how many asia student got denied

# idea
# con1=visa_df['continent']=='Asia'
# con2=shortlist certified
# con=con1&con2
# extract the data len
```

In [5]:
```
con1=visa_df['continent']=='Asia'
con2=visa_df['case_status']=='Certified'
con=con1&con2
len(visa_df[con])
```

Out[5]:
```
11012
```

In [6]:
```
continent_labels=visa_df['continent'].unique()
certifid_list=[]
denied_list=[]
for i in continent_labels:
    con1=visa_df['continent']==i
    con2=visa_df['case_status']=='Certified'
    con3=visa_df['case_status']=='Denied'
    certi_con=con1&con2
    deni_con=con1&con3
    certifid_list.append(len(visa_df[certi_con]))
    denied_list.append(len(visa_df[deni_con]))
```

```
certifid_list,denied_list
pd.DataFrame(zip(certifid_list,denied_list),
             index=continent_labels,
             columns=['Certified','Denied'])
```

Out[6]:

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

In [7]:
```
con=visa_df['case_status']=='Certified'
visa_df[con].groupby('continent').size()
```

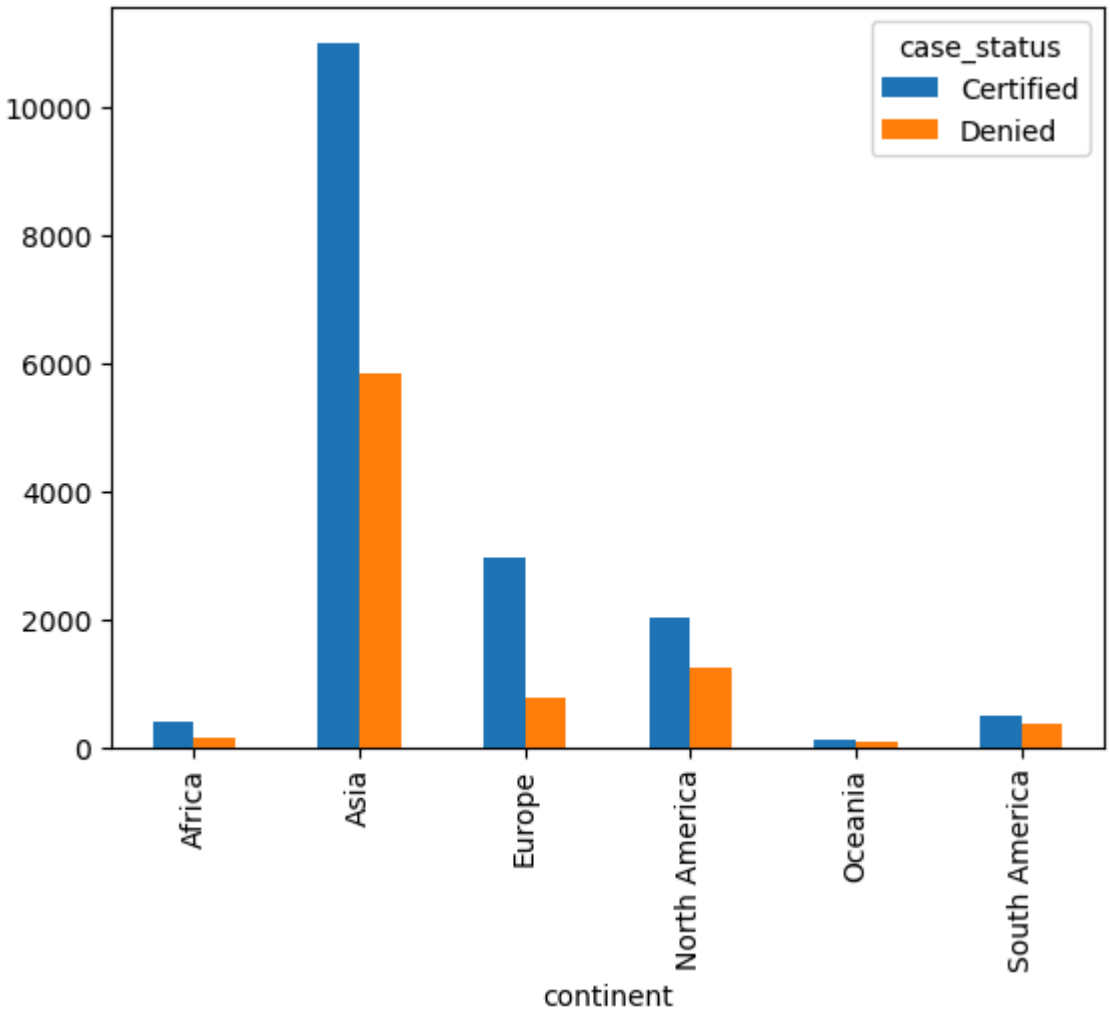
Out[7]:
```
continent
Africa             397
Asia             11012
Europe            2957
North America     2037
Oceania            122
South America      493
dtype: int64
```

In [8]:
```
col1=visa_df['continent']
col2=visa_df['case_status']
r1=pd.crosstab(col1,col2)
```

In [9]:
```
col1=visa_df['continent']
col2=visa_df['case_status']
r2=pd.crosstab(col2,col1)
```

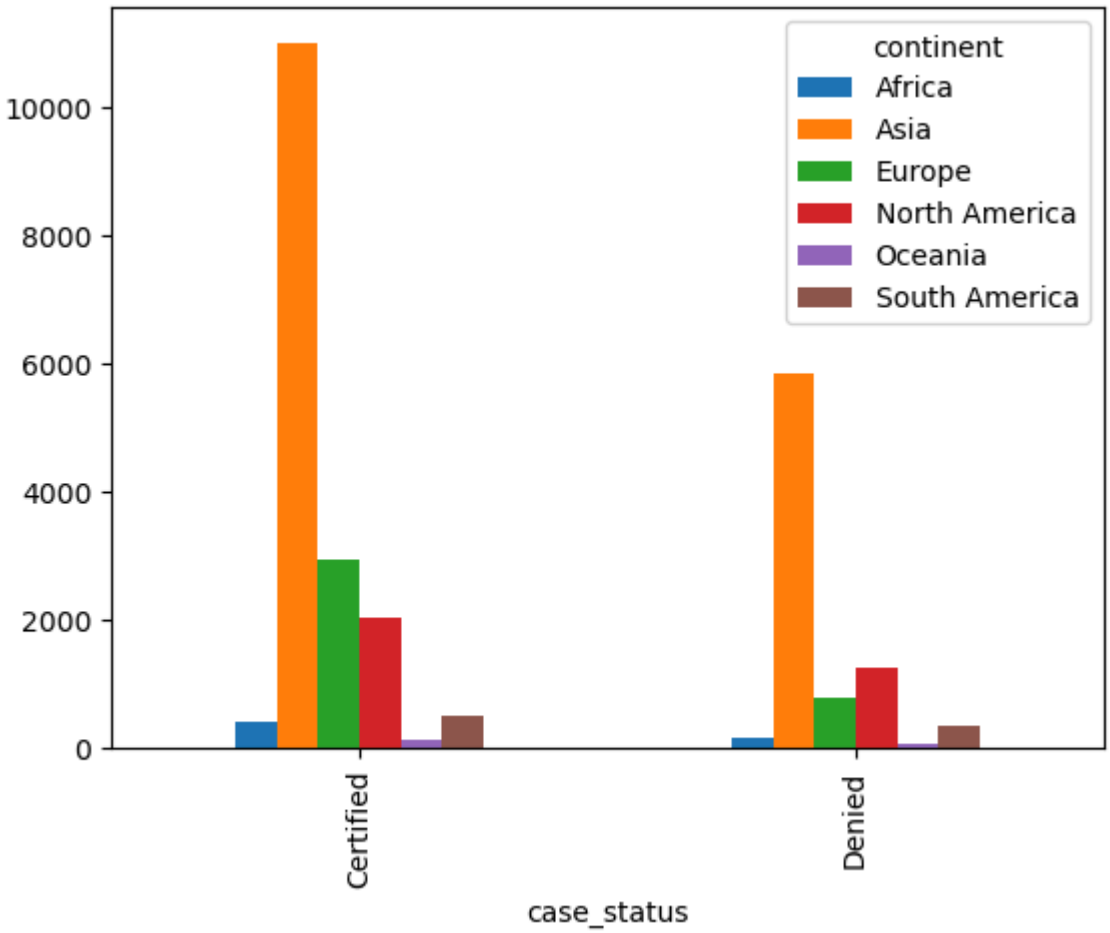
In [10]:
```
r1.plot(kind='bar')
```

Out[10]:
```
<Axes: xlabel='continent'>
```



In [11]:
```
r2.plot(kind='bar')
```

Out[11]:
```
<Axes: xlabel='case_status'>
```



In [12]:
```
# continent
# education of employee
# case_status

# there are 11k applicants certified from Asia
#   Now i want to know these 11k based on education
```

In [13]:
```
# select col1
# select col2
# select col3
# cols=[col2,col3]
# pd.crosstab(index,columns)
# pd.crosstab(col1,cols)
```

In [14]:
```
col1=visa_df['continent']
col2=visa_df['case_status']
col3=visa_df['education_of_employee']
cols=[col1,col2]
pd.crosstab(col3,cols)
```

Out[14]:

continent	Africa		Asia		Europe		Nort
case_status	Certified	Denied	Certified	Denied	Certified	Denied	Certifie
education_of_employee							
Bachelor's	81	62	4407	2761	1040	259	64
Doctorate	43	11	780	143	788	58	20
High School	23	43	676	1614	162	328	21
Master's	250	38	5149	1331	967	130	97

In [15]:
```
col1=visa_df['continent']
col2=visa_df['case_status']
col3=visa_df['education_of_employee']
cols=[col3,col2]
pd.crosstab(col1,cols)
```

Out[15]:

education_of_employee	Bachelor's		Doctorate		High School		
case_status	Certified	Denied	Certified	Denied	Certified	Denied	Certifie
continent							
Africa	81	62	43	11	23	43	25
Asia	4407	2761	780	143	676	1614	514
Europe	1040	259	788	58	162	328	96
North America	641	584	207	51	210	191	97
Oceania	38	28	19	3	19	17	4
South America	160	173	75	14	74	63	18

In [16]:
```
col1=visa_df['continent']
col2=visa_df['case_status']
col3=visa_df['education_of_employee']
cols=[col1,col3]
pd.crosstab(col2,cols)
```

Out[16]:

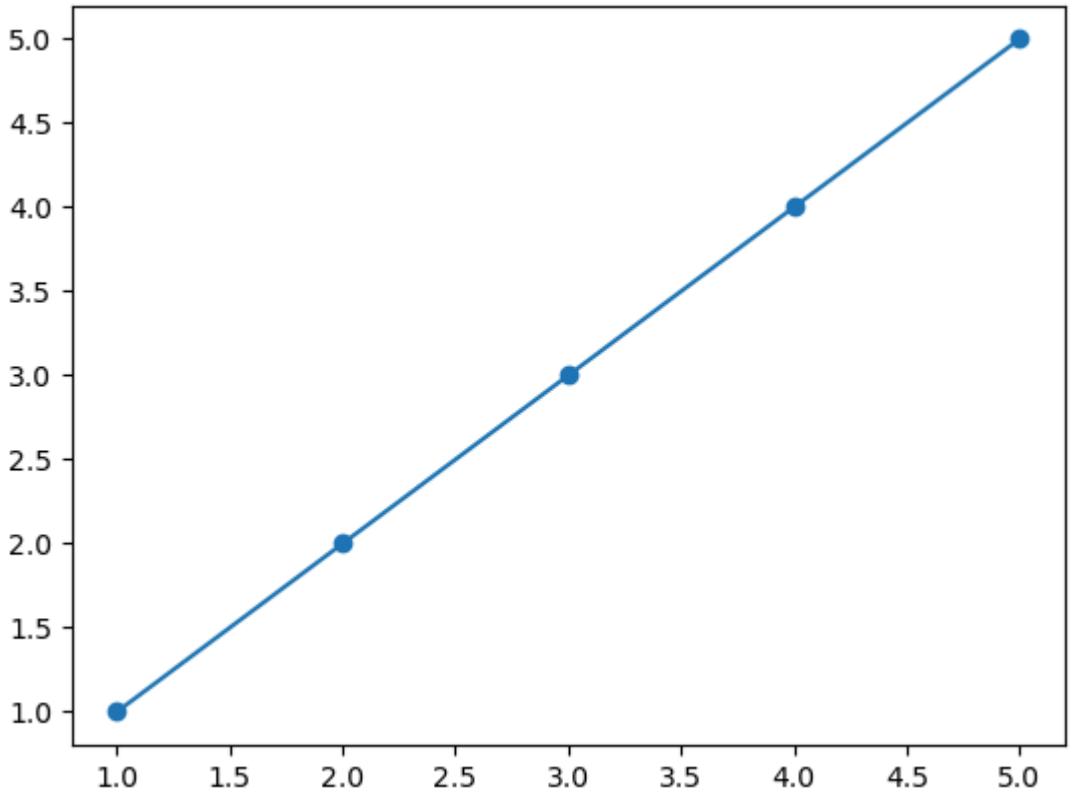
continent	Africa					
education_of_employee	Bachelor's	Doctorate	High School	Master's	Bachelor's	Doctorate
case_status						
Certified	81	43	23	250	4407	780
Denied	62	11	43	38	2761	143

2 rows × 24 columns

Numerical-Numerical

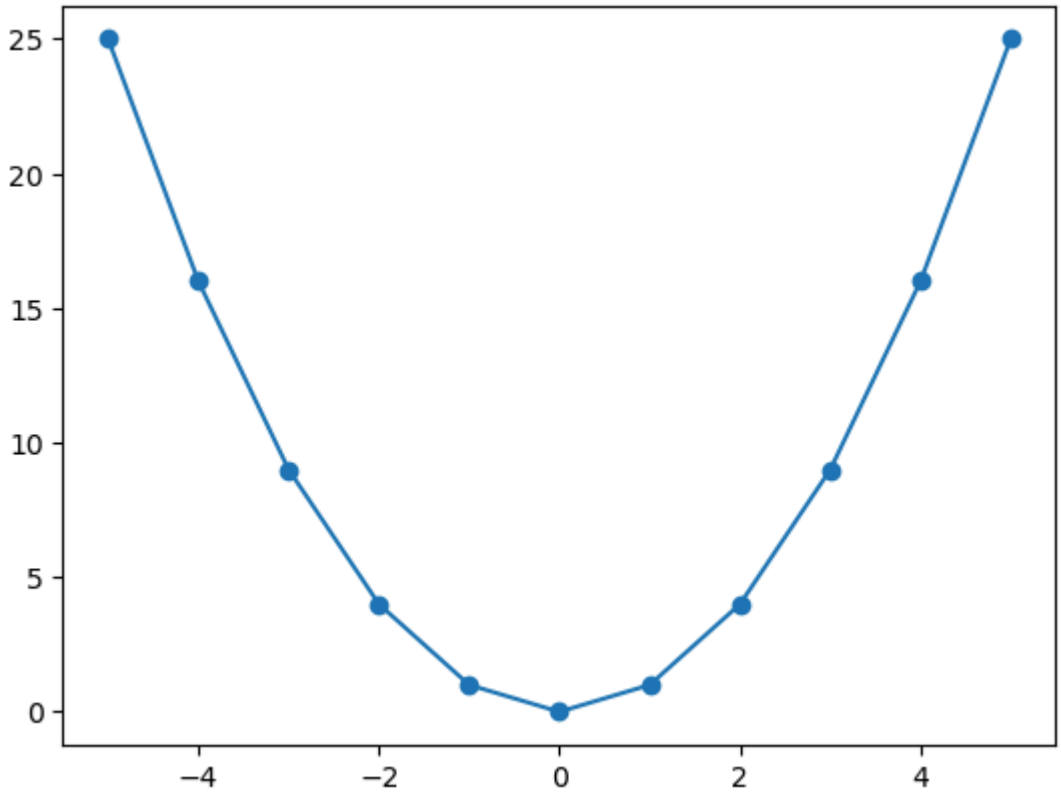
In [37]:
```
x=[1,2,3,4,5]
y=[1,2,3,4,5]
plt.scatter(x,y)
plt.plot(x,y)
```

Out[37]:
```
[<matplotlib.lines.Line2D at 0x2292f9769c0>]
```



In [47]:
```
#x=[i for i in range(1,10)]
x=list(range(-5,6))
y=[i**2 for i in x]
plt.scatter(x,y)
plt.plot(x,y)
```

Out[47]:
```
[<matplotlib.lines.Line2D at 0x229306dd160>]
```

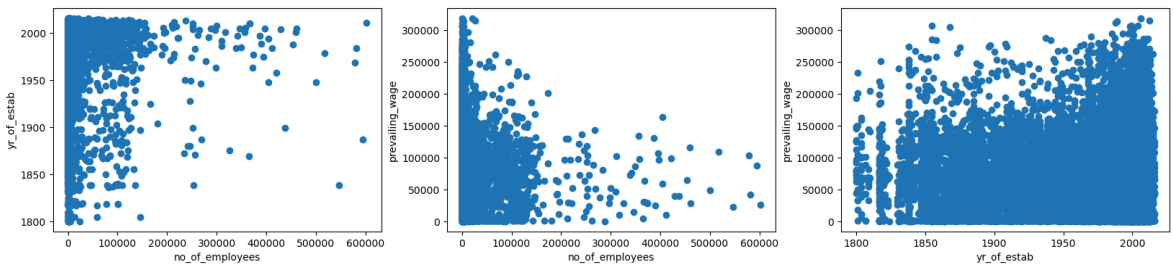


In [49]:
```
num
```

Out[49]:
```
Index(['no_of_employees', 'yr_of_estab', 'prevailing_wage'], dtype='object')
```

In [61]:
```
col1=visa_df['no_of_employees']
col2=visa_df['yr_of_estab']
col3=visa_df['prevailing_wage']
plt.figure(figsize=(20,4))
plt.subplot(1,3,1).scatter(col1,col2)
plt.xlabel('no_of_employees')
plt.ylabel('yr_of_estab')
#=========================================
plt.subplot(1,3,2).scatter(col1,col3)
plt.xlabel('no_of_employees')
plt.ylabel('prevailing_wage')
#=============================================
plt.subplot(1,3,3).scatter(col2,col3)
plt.xlabel('yr_of_estab')
plt.ylabel('prevailing_wage')
```

Out[61]:
```
Text(0, 0.5, 'prevailing_wage')
```



Correlation

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

In [67]:
```
visa_df.corr(numeric_only=True)
```

Out[67]:

	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

Heat map

In [74]:
```
import seaborn as sns
corr=visa_df.corr(numeric_only=True)
sns.heatmap(corr,annot=True)
```

Out[74]: <Axes: >

In [ ]: