NAME : RETHINAGIRI G

ROLL NO: 225229130

COURSE TITLE : DATA AND VISUAL ANALYTICS LAB

Lab7. Data Visualization in Seaborn

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

1. Visualizing Statistical Relationships

Import train_upvote_mini.csv file

In [2]:
```
df=pd.read_csv("C:\\Users\\user\\Downloads\\train_upvote_mini.csv")
df.head()
```

Out[2]:

	ID	Tag	Reputation	Answers	Username	Views	Upvotes
0	52664	a	3942.0	2.0	155623	7855.0	42.0
1	327662	a	26046.0	12.0	21781	55801.0	1175.0
2	468453	c	1358.0	4.0	56177	8067.0	60.0
3	96996	a	264.0	3.0	168793	27064.0	9.0
4	131465	c	4271.0	4.0	112223	13986.0	83.0

What is its size?

In [3]:
```
df.shape
```

Out[3]:
```
(15440, 7)
```

Show the types of each feature

In [4]:
```
df.dtypes
```

Out[4]:
```
ID              int64
Tag            object
Reputation    float64
Answers       float64
Username        int64
Views         float64
Upvotes       float64
dtype: object
```

How many unique "tag" available?

In [5]:
```
df['Tag'].unique()
```

Out[5]:
```
array(['a', 'c', 'r', 'j', 'p', 's', 'h', 'o', 'i', 'x'], dtype=object)
```

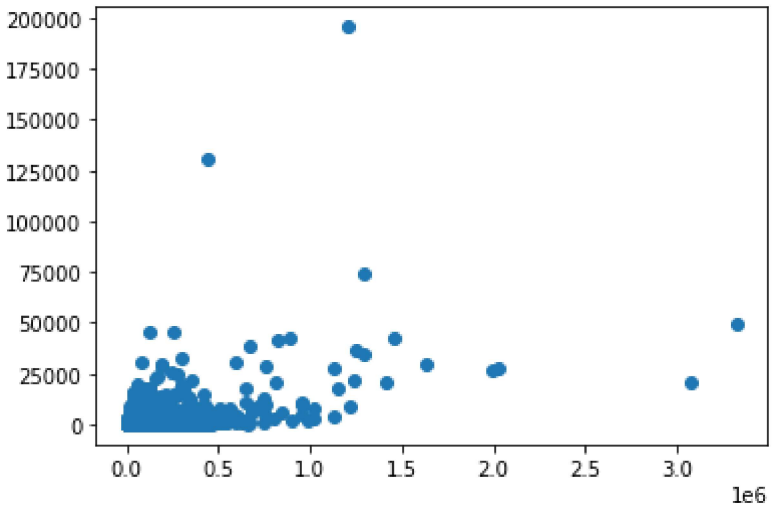
Visualize with Scatterplot

Does no. of views correlate no of upvotes?.

Show scatterplot (inherited from matplotlib) and relplot between "views" and "upvotes"

In [6]:
```
plt.scatter(x=df['Views'], y=df['Upvotes'])
```

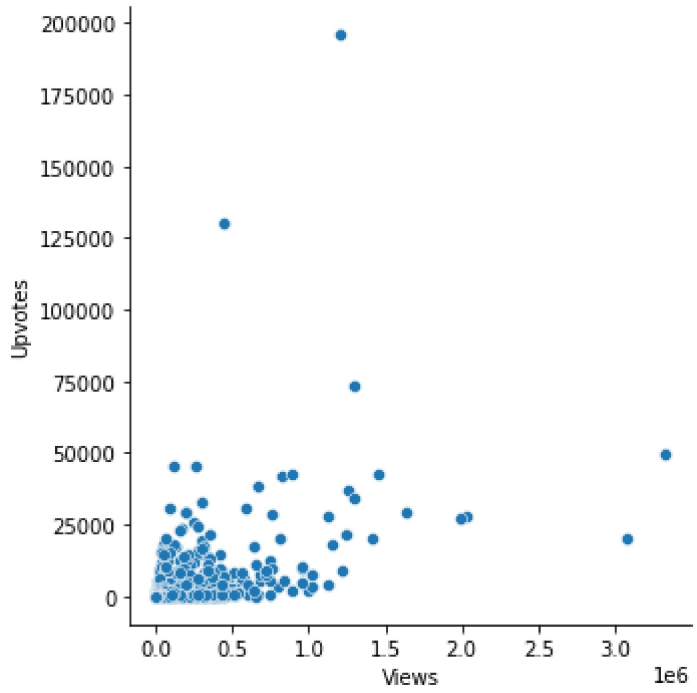
Out[6]:
```
<matplotlib.collections.PathCollection at 0x1a07cd91400>
```



Plot replot between "Views" and "Upvotes"

In [7]:
```
sns.relplot(data=df, x="Views", y="Upvotes")
```

Out[7]:
```
<seaborn.axisgrid.FacetGrid at 0x1a07ce12ca0>
```

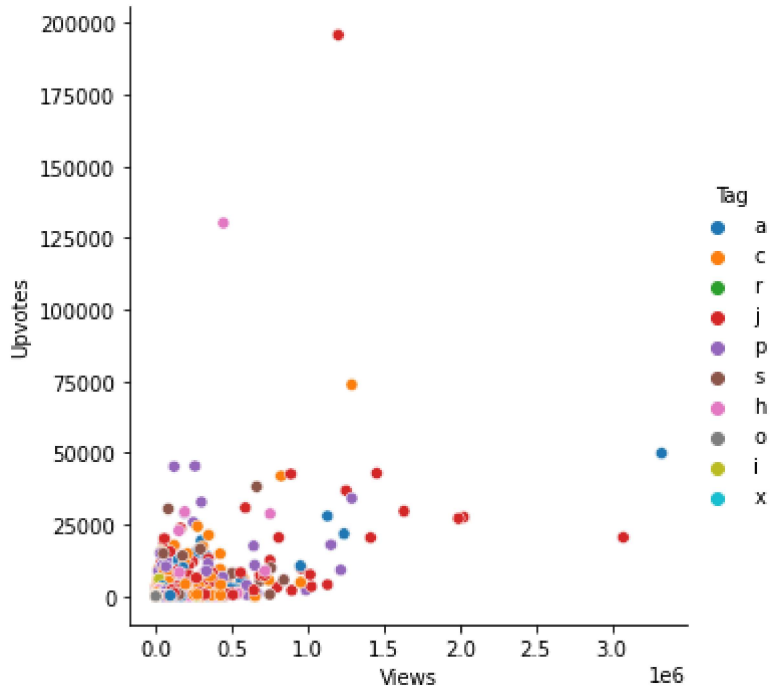


Next, we want to see the tag associated with data.

Plot relplot between "Views" and "Upvotes" with hue as "Tag"

In [8]:
```
sns.relplot(data=df, x="Views", y="Upvotes",hue='Tag')
```

Out[8]:
```
<seaborn.axisgrid.FacetGrid at 0x1a07ce67e80>
```

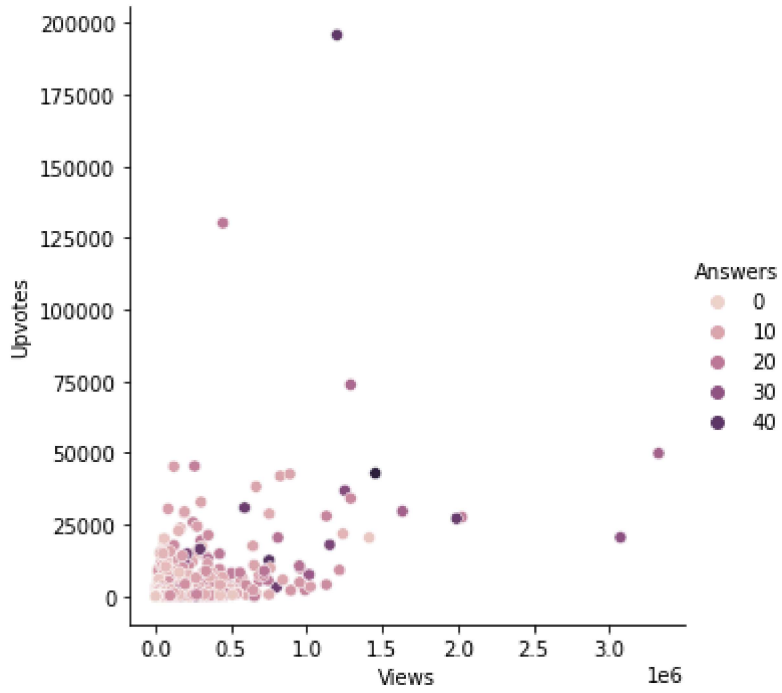


Hue Plot

Plot relplot between "Views" and "Upvotes" with hue as "Answers"

In [9]:
```
sns.relplot(data=df, x="Views", y="Upvotes",hue="Answers")
```

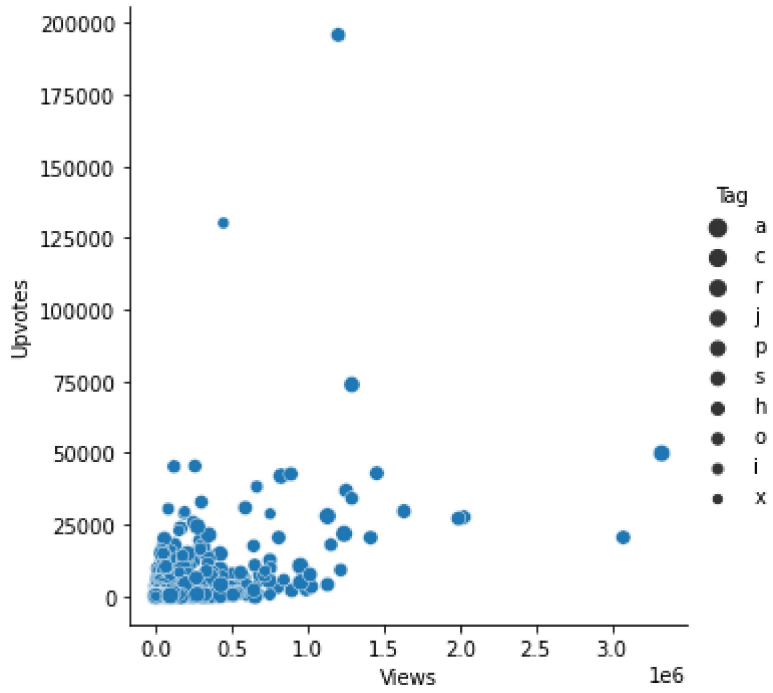
Out[9]:
```
<seaborn.axisgrid.FacetGrid at 0x1a07cf04a00>
```



Plot relplot between "Views" and "Upvotes" with size as "Tag"

In [10]:
```
sns.relplot(data=df, x="Views", y="Upvotes",size="Tag")
```

Out[10]:
```
<seaborn.axisgrid.FacetGrid at 0x1a07cf0aee0>
```

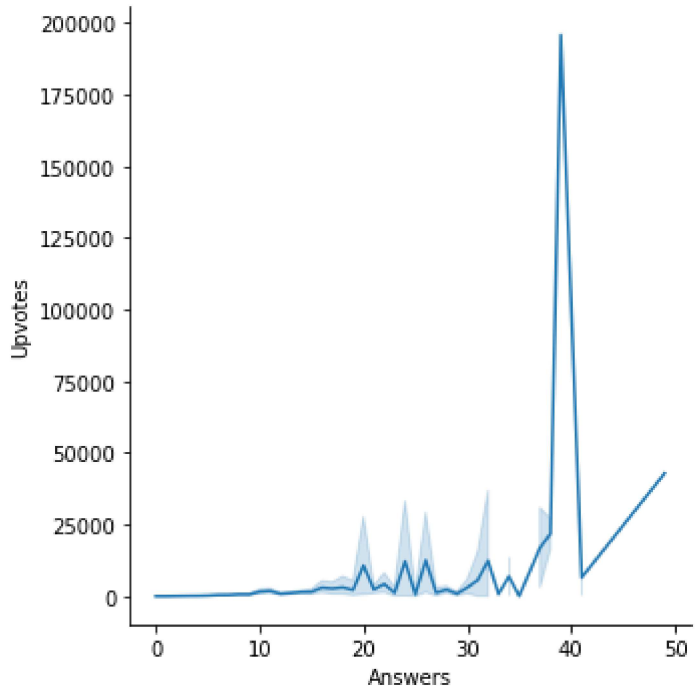


Does no of times question answered impact the no. of upvotes?

Plot line chart using relplot between "Answers" and "Upvotes"

In [11]:
```
sns.relplot(data=df, x="Answers", y="Upvotes",kind="line")
```

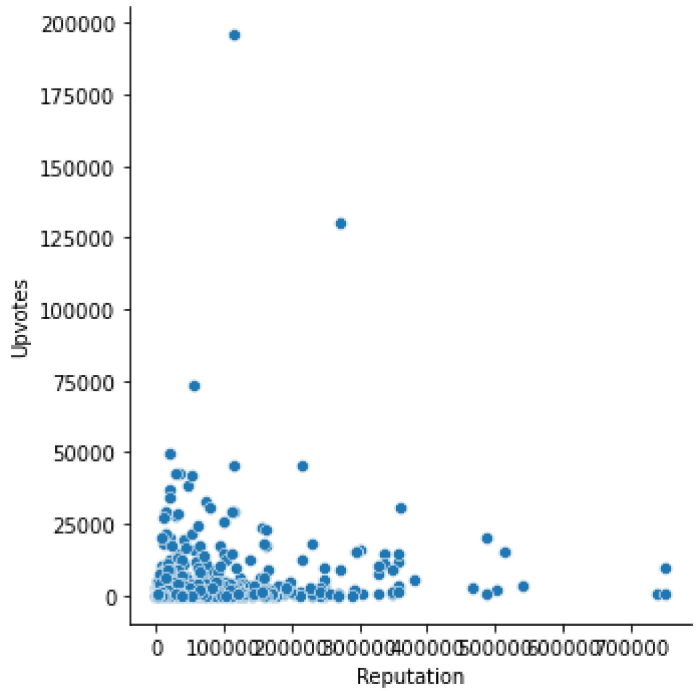
Out[11]:
```
<seaborn.axisgrid.FacetGrid at 0x1a07e0845e0>
```



Does Reputation score of question author impact no of upvotes?. Draw replot.

In [12]:
```
sns.relplot(data=df, x="Reputation", y="Upvotes")
```

Out[12]:
```
<seaborn.axisgrid.FacetGrid at 0x1a07cee6d60>
```



2. Visualizating Categorial Data

Various Categorial Plots in Seaborn

Categorical scatterplots:

- stripplot() (with kind="strip"; the default)
- swarmplot() (with kind="swarm")

Categorical distribution plots:

- boxplot() (with kind="box")
- violinplot() (with kind="violin")

- boxenplot() (with kind="boxen")

Categorical estimate plots:

- pointplot() (with kind="point")
- barplot() (with kind="bar")
- countplot() (with kind="count")

Dataset - HR analytics description

Jitter Plot

In [13]:
```
df1=pd.read_csv("C:/Users/user/Downloads/train_hr_mini.csv")
df1.head()
```

Out[13]:

	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_se
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	
1	65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	
3	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	
4	48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	

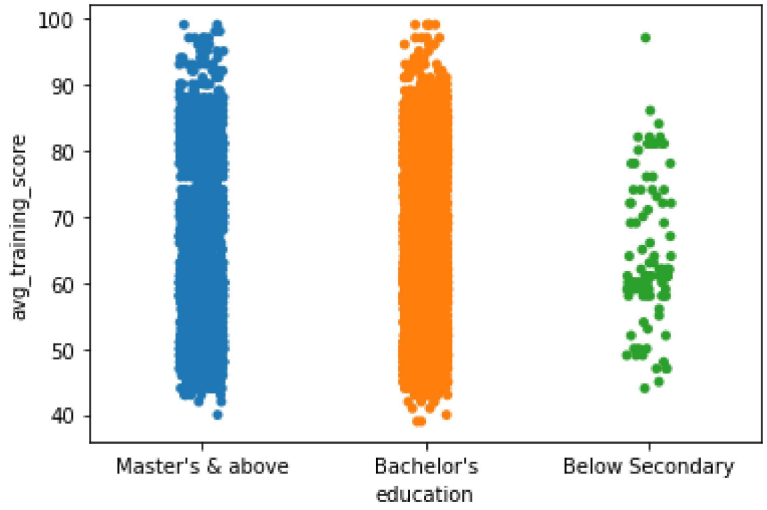
In [14]:
```
df1.shape
```

Out[14]:
```
(6397, 14)
```

Show Jitter plot between education and avg_training_score

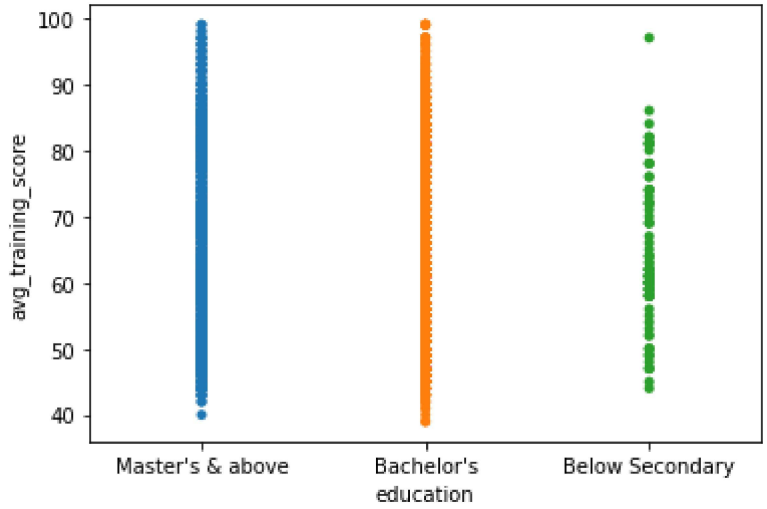
In [15]:
```
sns.stripplot(data=df1, x="education", y="avg_training_score", jitter=True)
```

Out[15]: <AxesSubplot:xlabel='education', ylabel='avg_training_score'>



In [16]:
```
sns.stripplot(data=df1, x="education", y="avg_training_score", jitter=False)
```

Out[16]: <AxesSubplot:xlabel='education', ylabel='avg_training_score'>



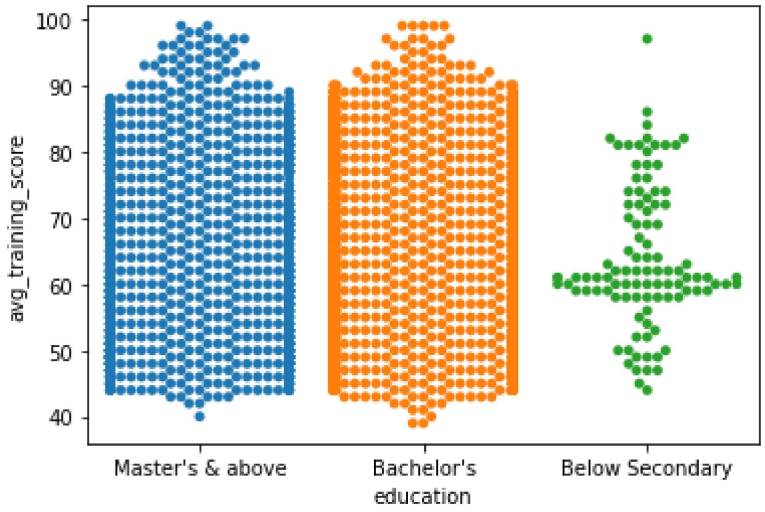
Sworm Plot

Plot Swarm plot between education category and avg_training_score

In [17]:
```
sns.swarmplot(data=df1, x="education", y="avg_training_score")
```

```
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 74.2% of the points cann
ot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 88.1% of the points cann
ot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

Out[17]:
```
<AxesSubplot:xlabel='education', ylabel='avg_training_score'>
```



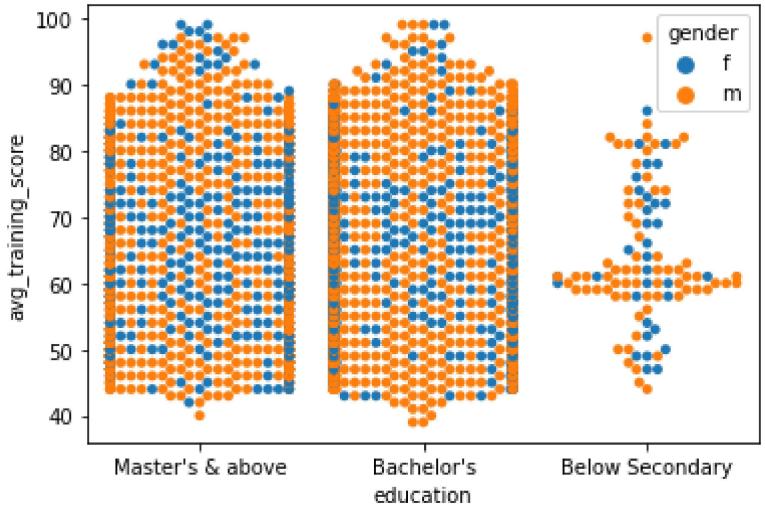
Hue Plot

Show Hue Plot to see the gender distribution in the plot of education category and avg_training_score.Here, hue is "gender".

In [18]:
```
sns.swarmplot(data=df1, x="education", y="avg_training_score",hue="gender")
```

```
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 74.2% of the points cann
ot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 88.1% of the points cann
ot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

Out[18]: <AxesSubplot:xlabel='education', ylabel='avg_training_score'>

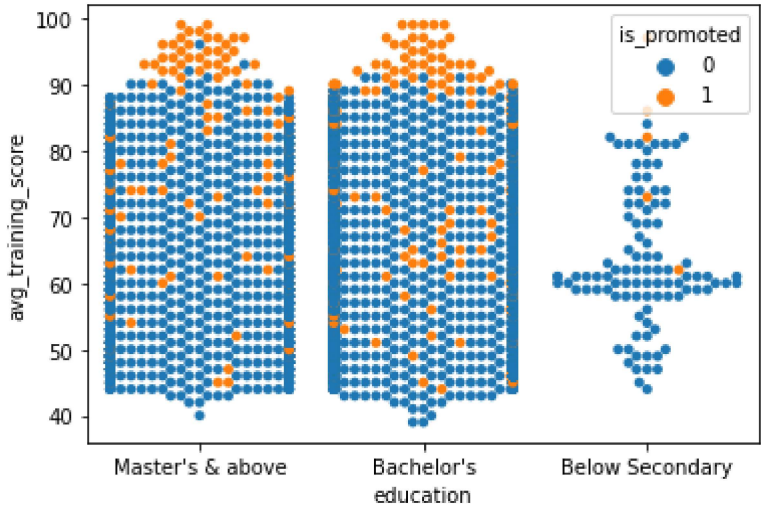


Who are all promoted considering education and avg training score?. Draw swarm plot with hue as "is_promoted"

In [19]:
```
sns.swarmplot(data=df1, x="education", y="avg_training_score",hue="is_promoted")
```

```
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 74.2% of the points cann
ot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 88.1% of the points cann
ot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

Out[19]: <AxesSubplot:xlabel='education', ylabel='avg_training_score'>

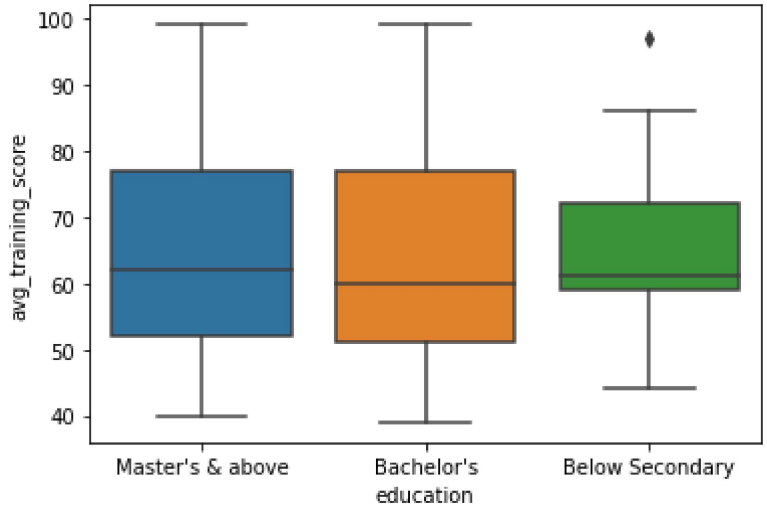


Box Plot

Draw box plot between education and avg_training_score

In [20]:
```
sns.boxplot(data=df1, x="education", y="avg_training_score")
```

Out[20]:
```
<AxesSubplot:xlabel='education', ylabel='avg_training_score'>
```

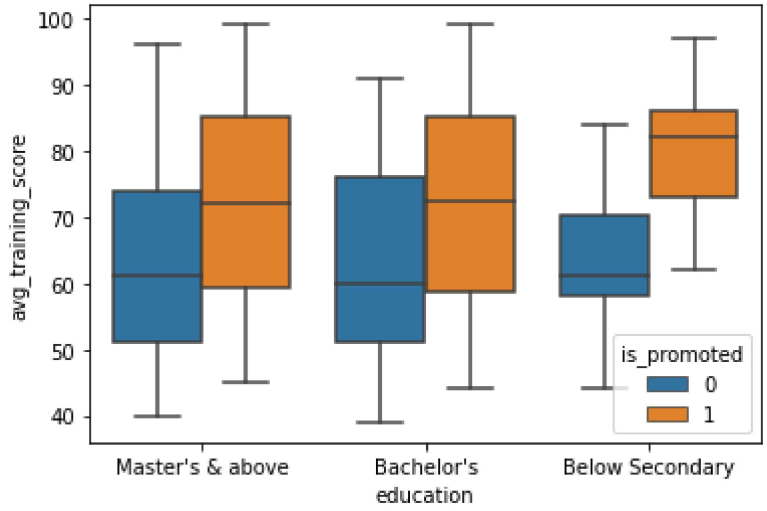


Box Plot with Hue Dimension

Who are promoted and not promoted considering education and avg_training_score?. Draw Box Plot.

In [21]:
```
sns.boxplot(data=df1, x="education", y="avg_training_score",hue="is_promoted")
```

Out[21]:
```
<AxesSubplot:xlabel='education', ylabel='avg_training_score'>
```

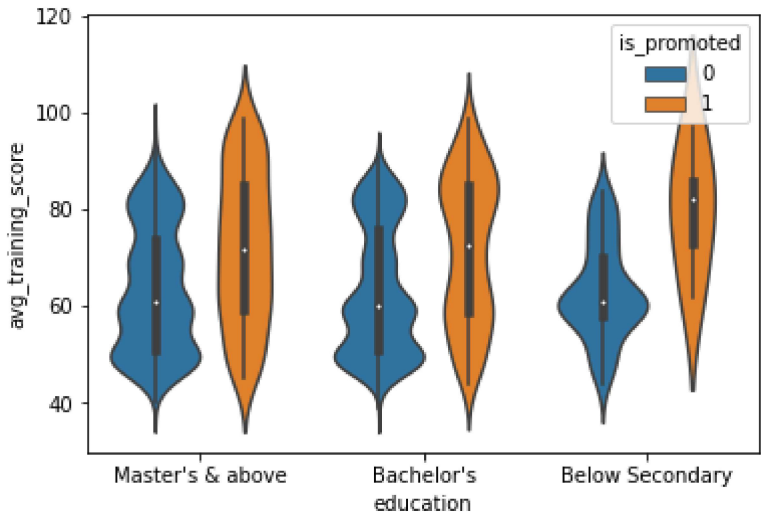


Violin Plot

Show violin plot between education categories and avg training score with hue as "is_promoted" target variable

In [22]:
```
sns.violinplot(data=df1, x="education", y="avg_training_score",hue="is_promoted")
```

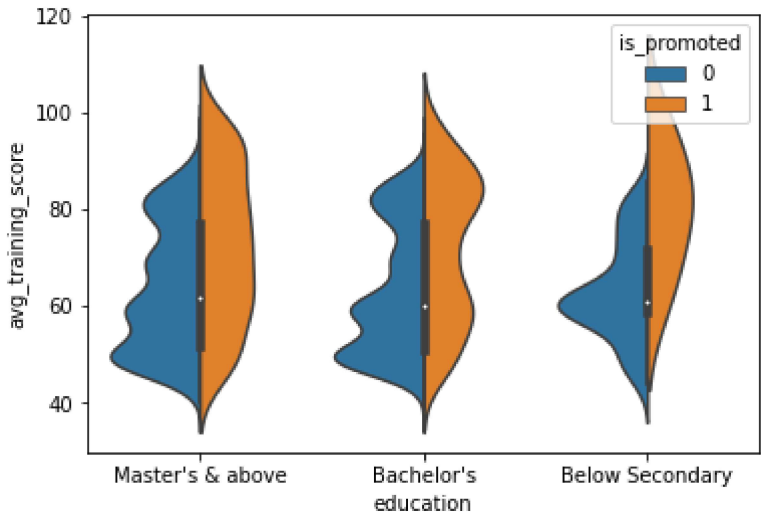
Out[22]:
```
<AxesSubplot:xlabel='education', ylabel='avg_training_score'>
```



Draw Violin plot with only 2 hue levels, use split attribute

In [23]:
```
sns.violinplot(data=df1, x="education", y="avg_training_score", hue ='is_promoted', split=3)
```

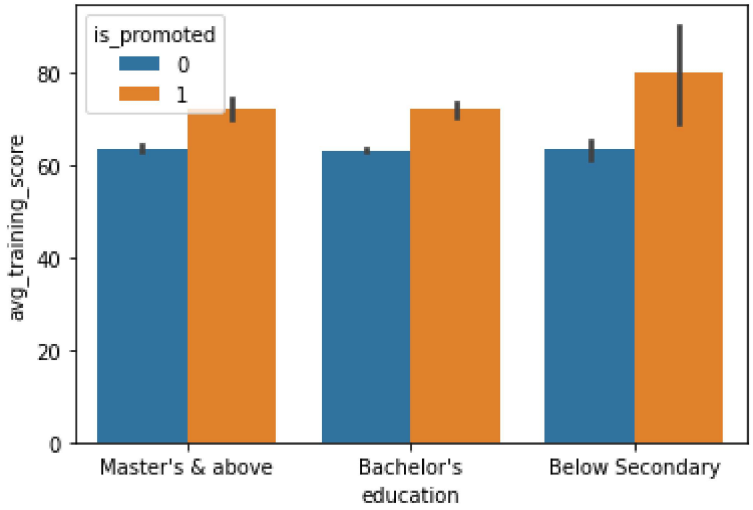
Out[23]:
```
<AxesSubplot:xlabel='education', ylabel='avg_training_score'>
```



Using catplot(), draw a Bar Chart between "education" and "avg_training_score", with hue as "is_promoted"

In [24]:
```
sns.barplot(data=df1, x="education", y="avg_training_score", hue ='is_promoted')
```

Out[24]:
```
<AxesSubplot:xlabel='education', ylabel='avg_training_score'>
```

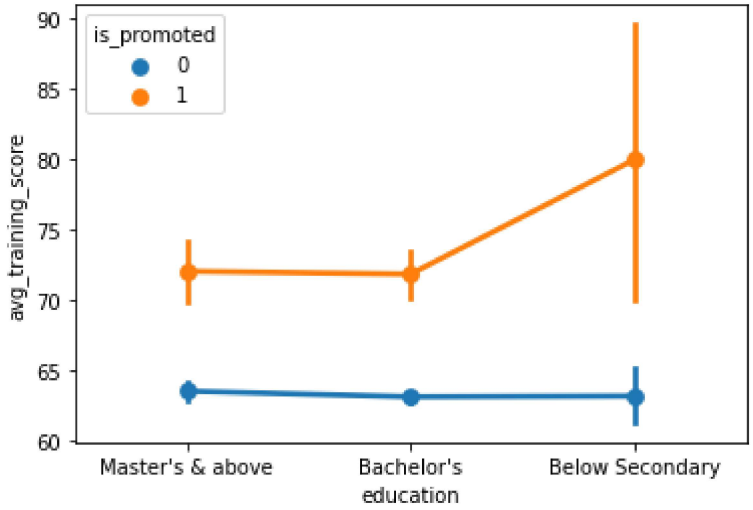


Point Plot

Show point plot between education and avg training score with hue promotion category

In [25]:
```
sns.pointplot(data=df1, x="education", y="avg_training_score", hue ='is_promoted')
```

Out[25]:
```
<AxesSubplot:xlabel='education', ylabel='avg_training_score'>
```



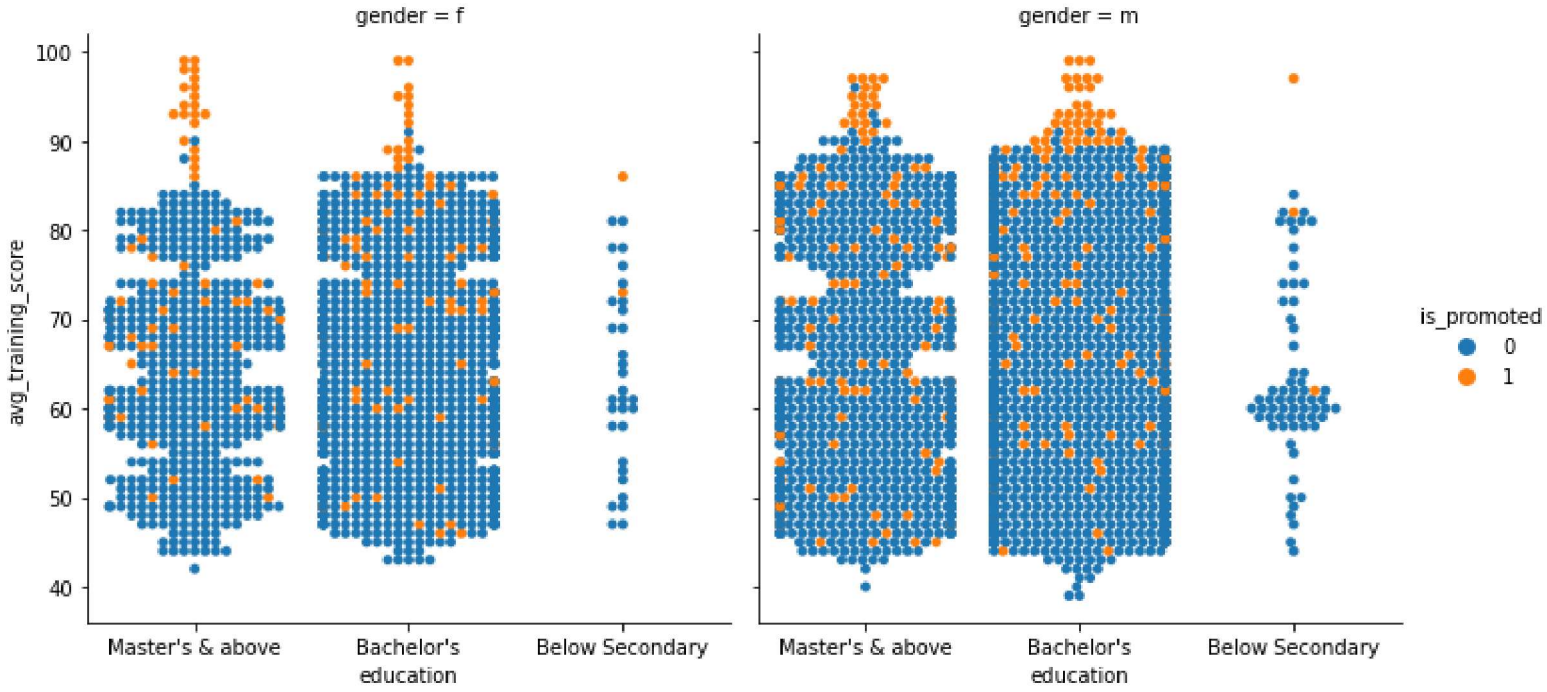
Multiple Dimension in Seaborn

Draw swarm plot for education, avg training score, hue as is_promoted for male and female category

In [26]:
```
sns.catplot(data=df1, x="education", y="avg_training_score", hue ='is_promoted', kind='swarm',col='gender')
```

```
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 6.4% of the points canno
t be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 40.9% of the points cann
ot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 38.9% of the points cann
ot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\user\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 73.2% of the points cann
ot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

Out[26]:
```
<seaborn.axisgrid.FacetGrid at 0x1a07f861820>
```



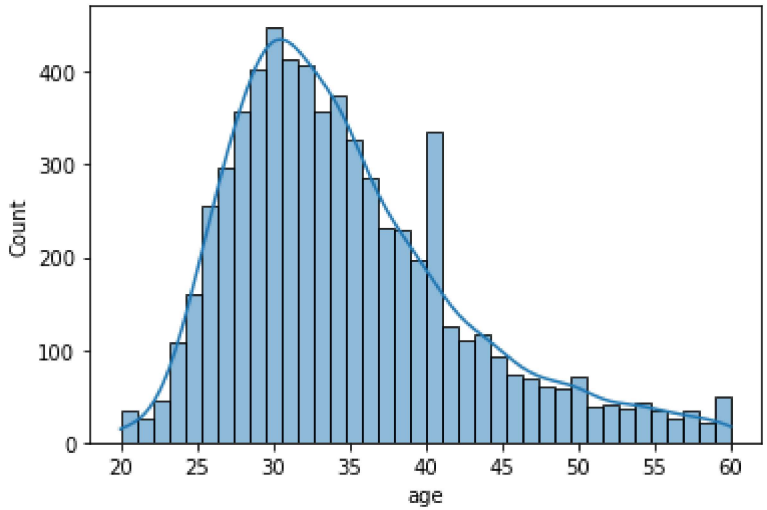
3. Visualizing the Distribution of Data

Plot Univariate Distributions

Plot Histogram with kernel density estimate value for age attribute

In [27]:
```
sns.histplot(data=df1,x='age',kde=True)
```

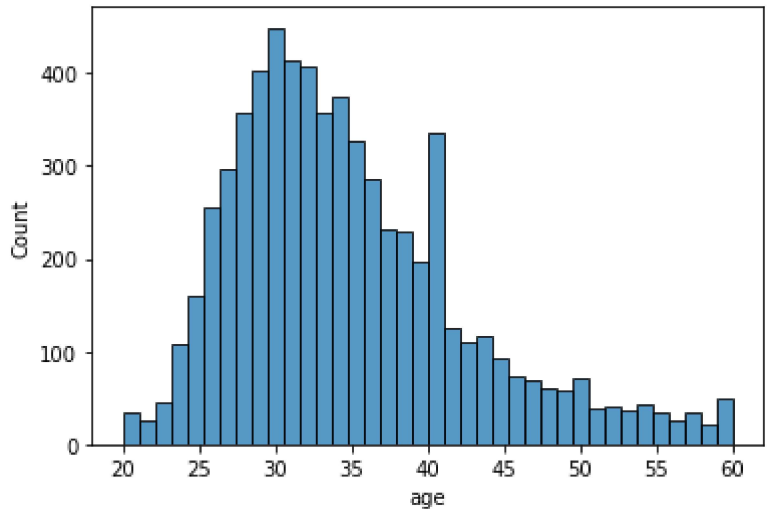
Out[27]:
```
<AxesSubplot:xlabel='age', ylabel='Count'>
```



Show only Histogram for age variable, without KDE

In [28]:
```
sns.histplot(data=df1,x='age',kde=False)
```

Out[28]:
```
<AxesSubplot:xlabel='age', ylabel='Count'>
```



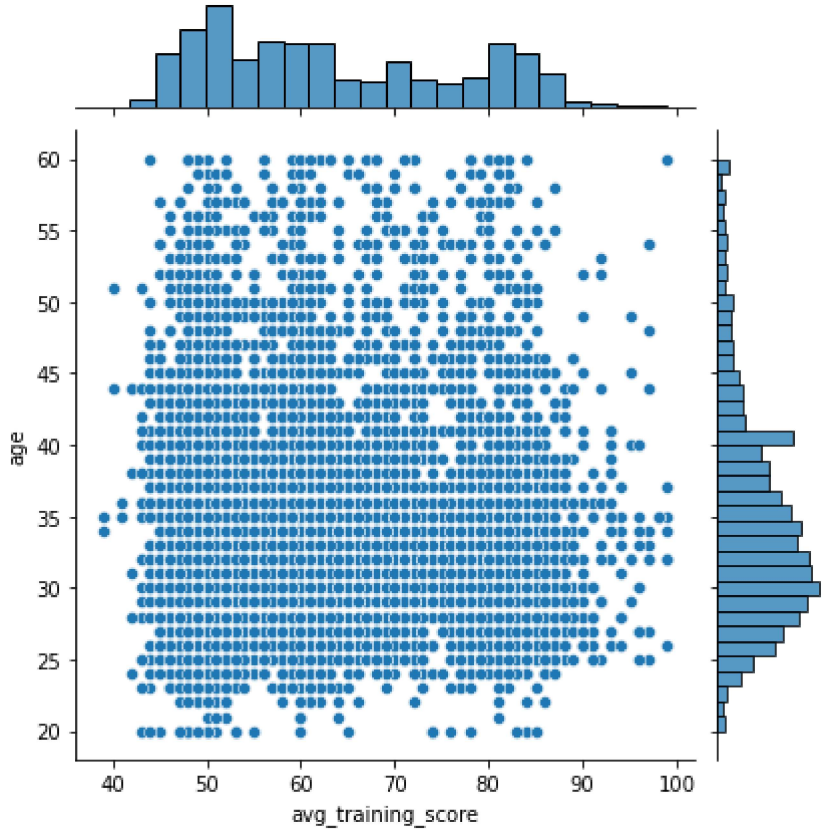
Plot Bivariate Distributions

Joint Plot

Draw a joint plot between avg_training_score and age

In [29]:
```
sns.jointplot(x='avg_training_score',y='age', data=df1 )
```

Out[29]:
```
<seaborn.axisgrid.JointGrid at 0x1a001032a00>
```

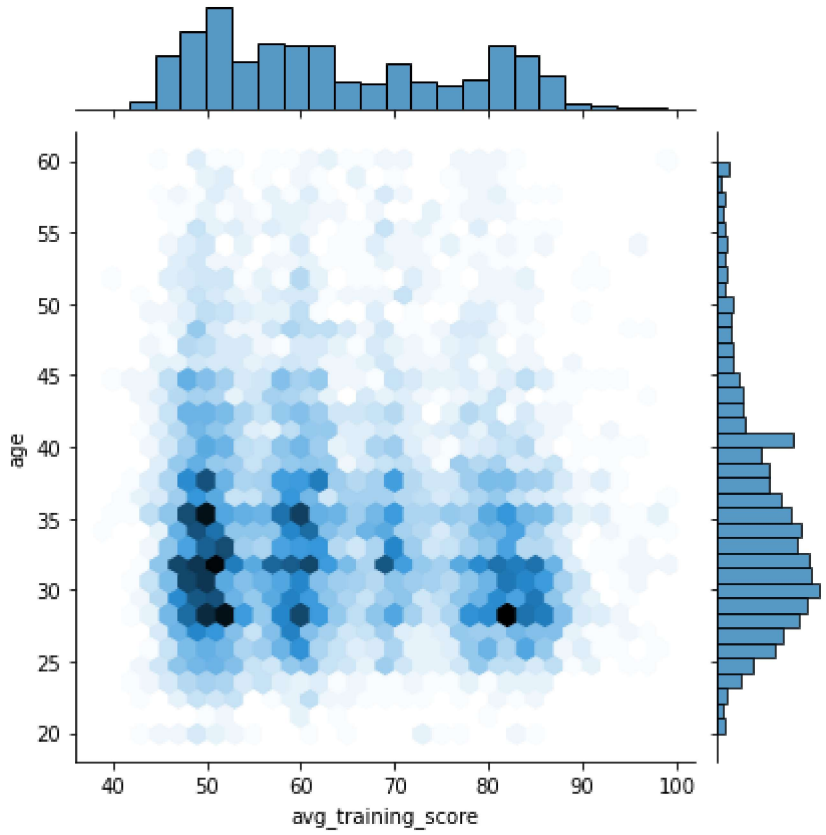


Hex Plot

Draw a hexplot for depicting the relationship between avg training score and age

In [30]:
```
sns.jointplot(x='avg_training_score',y='age', data=df1,kind='hex' )
```

Out[30]:
```
<seaborn.axisgrid.JointGrid at 0x1a0010a5520>
```

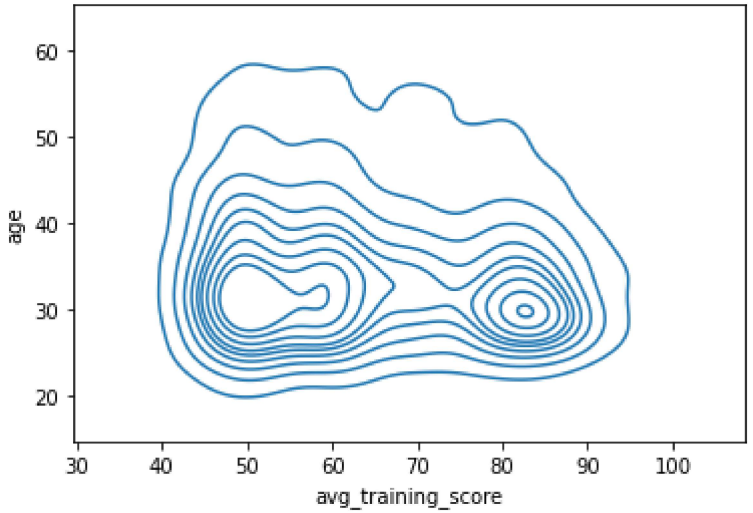


KDE Plot

Show KDE Plot to visualize age vs avg training score

In [31]:
```
sns.kdeplot(x='avg_training_score',y='age', data=df1 )
```

Out[31]:
```
<AxesSubplot:xlabel='avg_training_score', ylabel='age'>
```

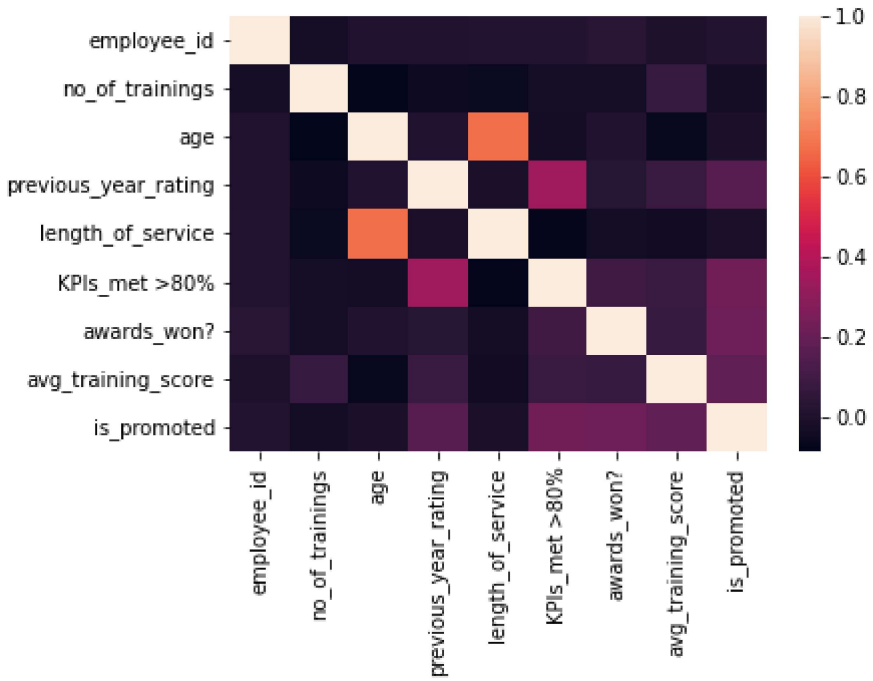


Heat Map

Draw heatmap for the dataset

In [32]:
```
sns.heatmap(data=df1.corr())
```

Out[32]: <AxesSubplot:>



Can you answer these questions about the previous heatmap?

- What's the strongest and what's the weakest correlated pair (except the main diagonal)?

The strongest correlated pair will have the highest correlation coefficient value, which will appear as a bright red cell in the heatmap. The weakest correlated pair will have the lowest correlation coefficient value, which will appear as a dark blue cell in the heatmap.

- What are the three variables most correlated with the target variable, is_promoted ?

Previuos_year_rating KPIs met>80% Awards_Won?

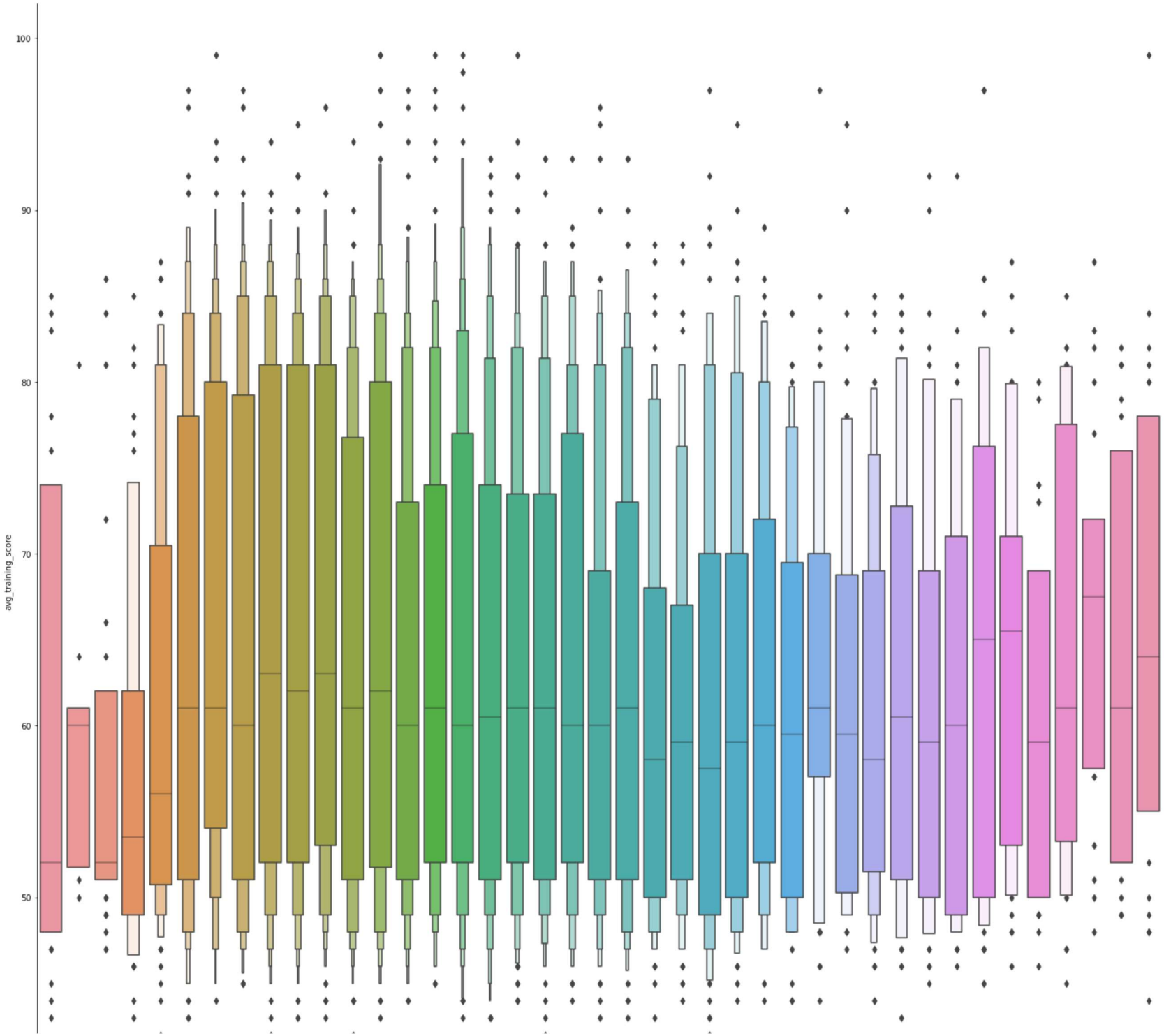
Boxen Plot

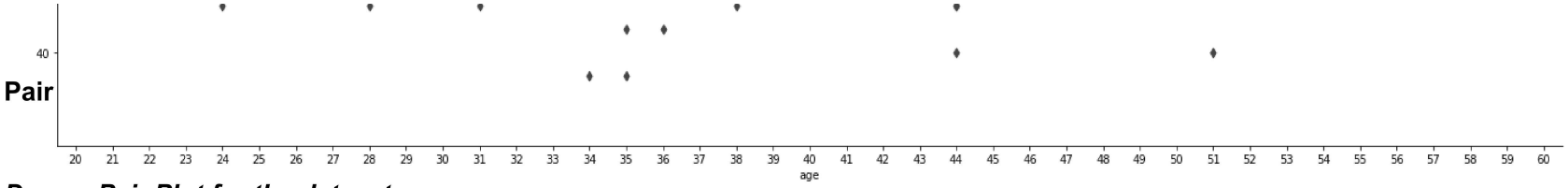
Draw Boxen Plot between "age" and "avg_training_score, with hue "is_promoted"

Adjust height and aspect values to make chart pretty

In [34]:
```
sns.catplot(x='age',y='avg_training_score', data=df1, kind='boxen', height=20, aspect=1 )
```

Out[34]:
```
<seaborn.axisgrid.FacetGrid at 0x1a0014736d0>
```

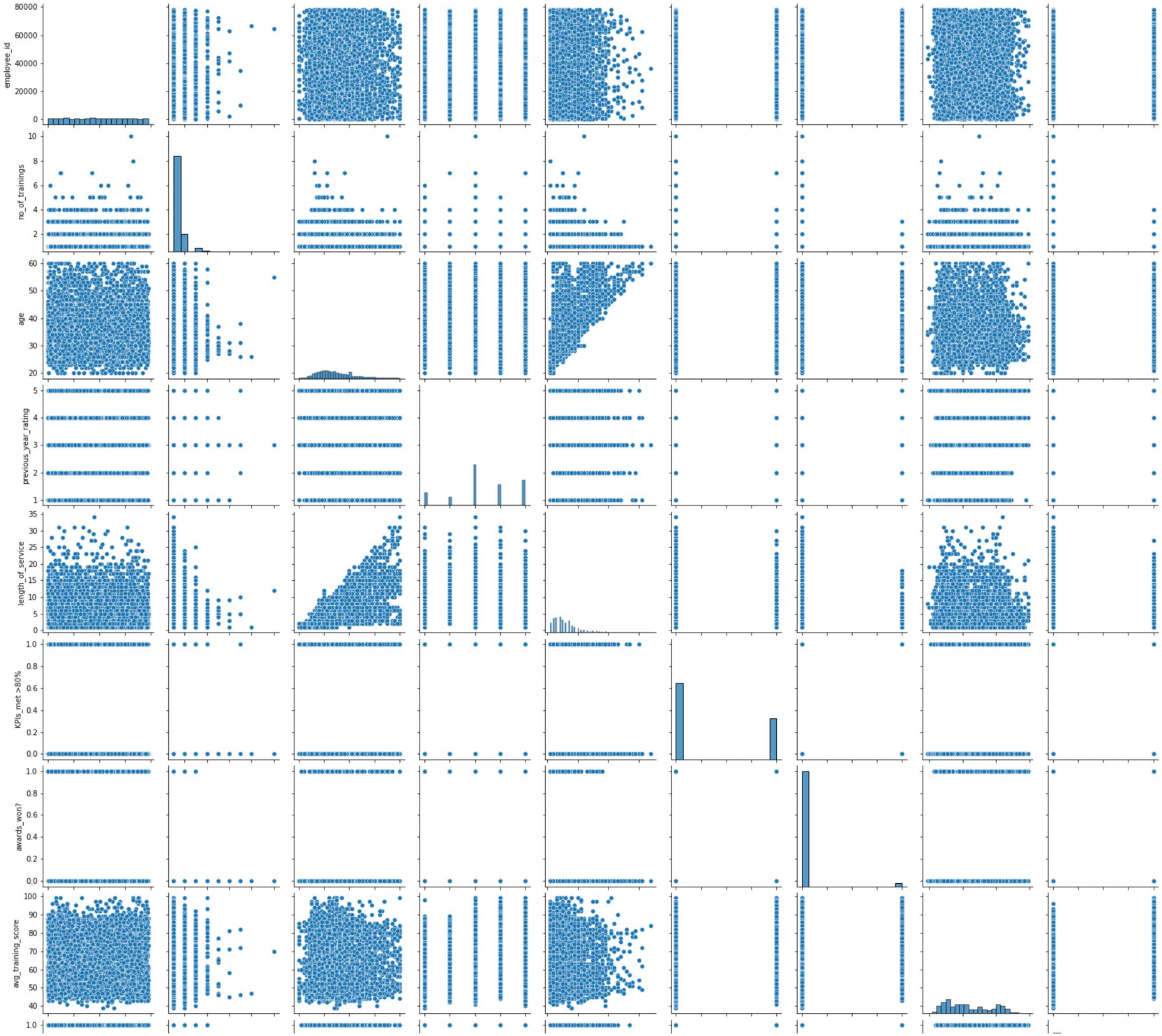

avg_training_score
100
90
80
70
60
50

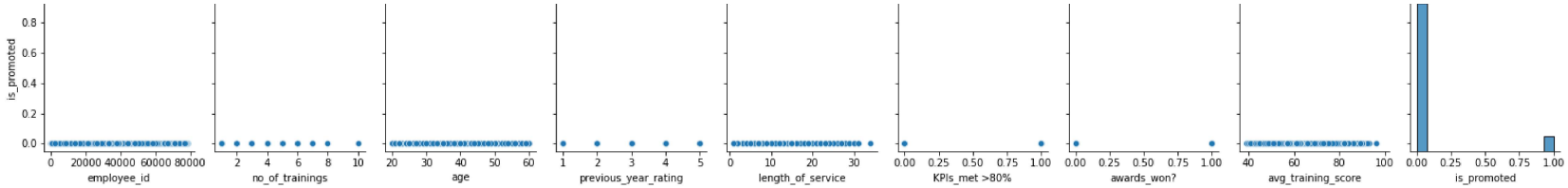


Draw a Pair Plot for the dataset

In [35]:
```
sns.pairplot(data=df1)
```

Out[35]:
```
<seaborn.axisgrid.PairGrid at 0x1a001f48dc0>
```


employee_id
no_of_trainings
age
previous_year_rating
length_of_service
KPIs_met >80%
awards_won?
avg_training_score


is_promoted
0.0
0.2
0.4
0.6
0.8
0
20000
40000
60000
80000
employee_id
2
4
6
8
10
no_of_trainings
20
30
40
50
60
age
1
2
3
4
5
previous_year_rating
0
10
20
30
length_of_service
0.00
0.25
0.50
0.75
1.00
KPIs_met >80%
0.00
0.25
0.50
0.75
1.00
awards_won?
40
60
80
100
avg_training_score
0.00
0.25
0.50
0.75
1.00
is_promoted