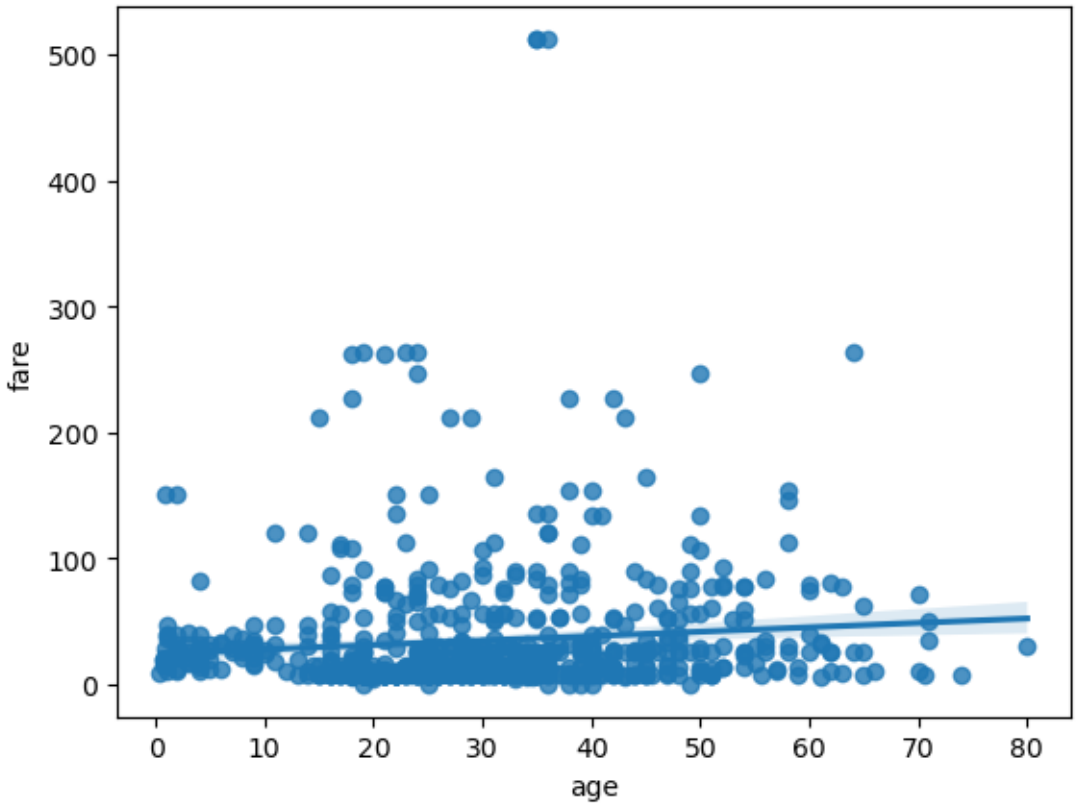
Lecture22

April 23, 2024

[5]:
```
import seaborn as sns
df=sns.load_dataset('titanic')
```

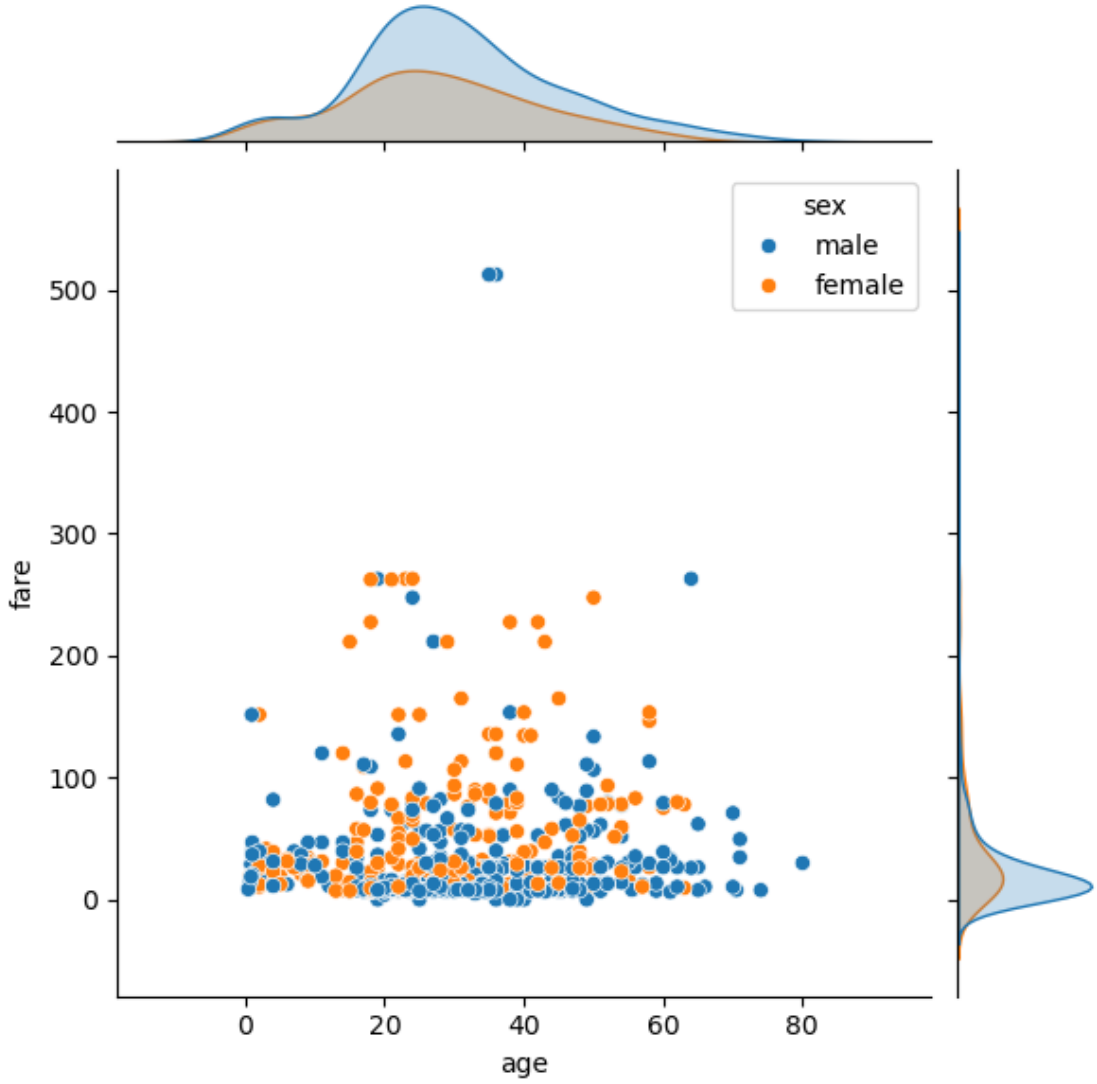
[4]:
```
# displot(data, x, hue)
# sns.countp[lot(data, x, hue)
# sns.kdeplot(data, x, hue)
# sns.stripplot(data, x, y, hue)
# sns.swarmplot(data, x,y, hue)
# sns.barplot(data, x,y, hue)
# sns.pointplot(data, x,y, hue)
```

[7]:
```
# regplot (cont. vs cont.)
sns.regplot(data=df, x='age', y='fare');
```



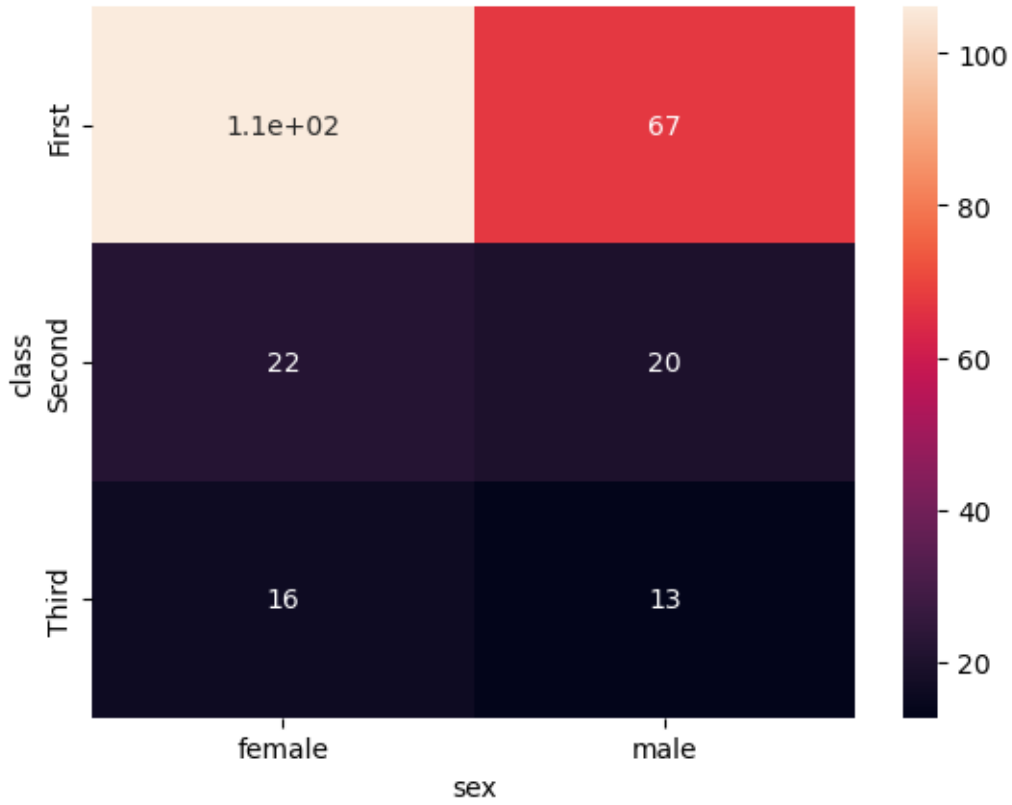
[14]:
```
# jointplot
# sns.jointplot(data=df, x='age', y='fare', kind='reg')
sns.jointplot(data=df, x='age', y='fare', hue='sex')
```

[14]: <seaborn.axisgrid.JointGrid at 0x7fb69e30ffd0>



[22]:
```
# heatmap
pt=df.pivot_table(index='class', columns='sex', values='fare', observed=True)
sns.heatmap(pt, annot=True)
```

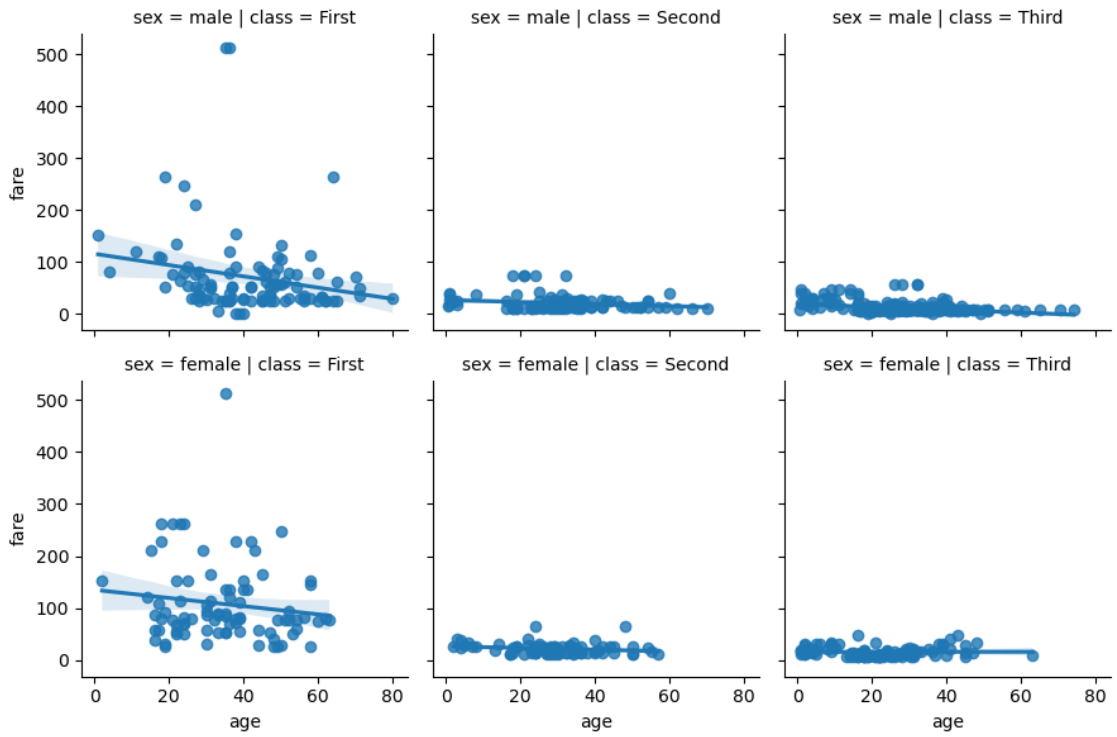
[22]: <Axes: xlabel='sex', ylabel='class'>



[25]:
```
import matplotlib.pyplot as plt
```

[28]:
```
# FacetGrid
# sns.FacetGrid(data=df, col='class', row='sex').map(plt.scatter, 'age', 'fare')
sns.FacetGrid(data=df, col='class', row='sex').map(sns.regplot, 'age', 'fare')
```

[28]: <seaborn.axisgrid.FacetGrid at 0x7fb68f1c7750>



[33]:
```
df.head()
```

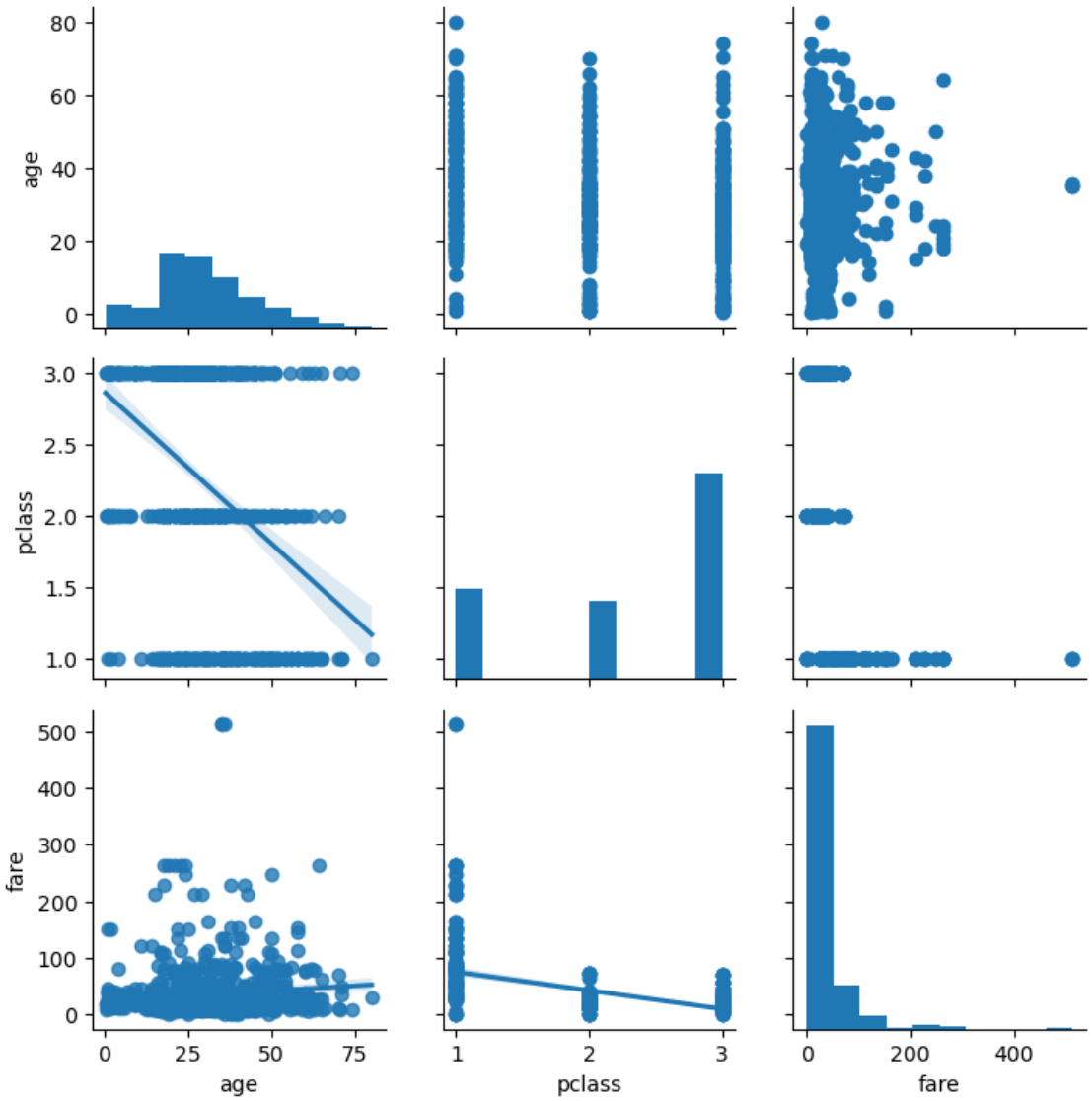
[33]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	\
0	0	3	male	22.0	1	0	7.2500	S	Third	
1	1	1	female	38.0	1	0	71.2833	C	First	
2	1	3	female	26.0	0	0	7.9250	S	Third	
3	1	1	female	35.0	1	0	53.1000	S	First	
4	0	3	male	35.0	0	0	8.0500	S	Third	

	who	adult_male	deck	embark_town	alive	alone
0	man	True	NaN	Southampton	no	False
1	woman	False	C	Cherbourg	yes	False
2	woman	False	NaN	Southampton	yes	True
3	woman	False	C	Southampton	yes	False
4	man	True	NaN	Southampton	no	True

[35]:
```
# PairGrid
# sns.PairGrid(data=df).map(plt.scatter)
sns.PairGrid(data=df[['age','pclass','fare']]).map_diag(plt.hist).map_upper(plt.
↪scatter).map_lower(sns.regplot)
```

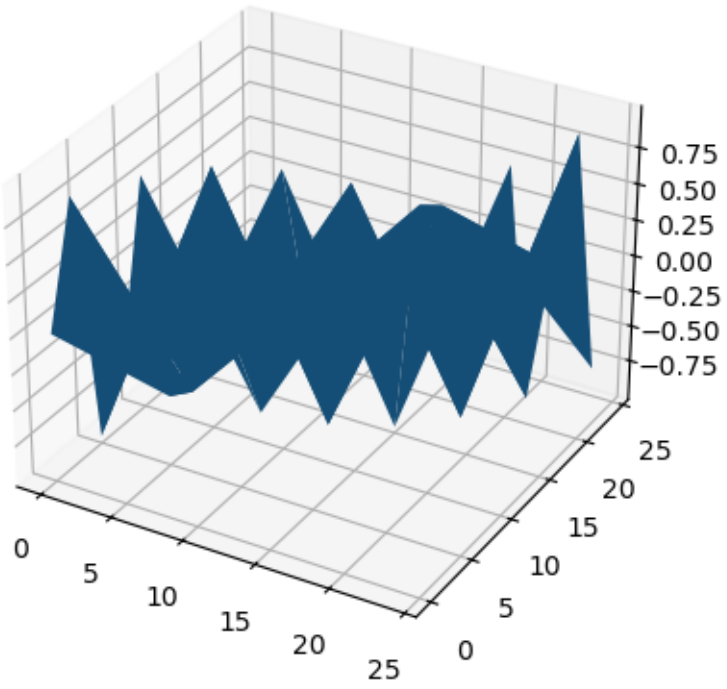
[35]:
```
<seaborn.axisgrid.PairGrid at 0x7fb682d98c10>
```



[40]:
```
# 3d plot
ax=plt.axes(projection ='3d')
# ax.plot3D(x,y,z)
# x=np.arange(25)
# y=np.arange(5,30)
# z=x+y
# ax.plot3D(x,y,z)

import numpy as np
x=np.arange(25).reshape(5,5)
y=np.arange(25).reshape(5,5)
z=np.sin(x+y)
ax.plot_surface(x,y,z)
```

[40]: <mpl_toolkits.mplot3d.art3d.Poly3DCollection at 0x7fb6826fc090>



[41]:
```
# Summary
# One variable
    # Cat
    # sns.countplot(data, x)
    # Cont
    # plt.hist(x)
    # plt. pie(x)
    # plt.boxplot(x)
    # sns.displot(data, x)
    # sns.kdeplot(data, x)
# Two variables
    # cat vs cat
    # sns.barplot(data,x,y)
    # sns.pointplot(data, x, y)
    # sns.heatmap(pivoit_table)

    # Cont vs Cont
    # plt.bar(x,y)
    # plt.barh(x,y)
    # plt.scatter(x,y)
    # plt.plot(x,y)
```

```
    # sns.regplot(data,x,y)
    # sns.lmplot(data, x,y)
    # sns.jointplot(data, x,y)
    # sns.PairGrid(data, x,y)

    # Cont vs Cat
    # sns.swarmplot(data, x, y)
    # sns.stripplot(data, x, y)
    # sns.FacetGrid()
    # hue
```

1 Data Pre-processing

[43]:
```
# pip install scikit-learn
```

[53]:
```
import sklearn.preprocessing as skp
import pandas as pd
df=pd.read_csv('datafiles/tips.csv')
df.head()
```

[53]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

1.0.1 Normalization –> [0,1]

[60]:
```
df['total_bill'].values.reshape(-1,1).shape
```

[60]:
```
(244, 1)
```

[62]:
```
df['total_bill_norm']=skp.MinMaxScaler().fit_transform(df['total_bill'].values.
↪reshape(-1,1))
```

[63]:
```
df.head()
```

[63]:

	total_bill	tip	sex	smoker	day	time	size	total_bill_norm
0	16.99	1.01	Female	No	Sun	Dinner	2	0.291579
1	10.34	1.66	Male	No	Sun	Dinner	3	0.152283
2	21.01	3.50	Male	No	Sun	Dinner	3	0.375786
3	23.68	3.31	Male	No	Sun	Dinner	2	0.431713
4	24.59	3.61	Female	No	Sun	Dinner	4	0.450775

1.0.2 Standarization –> mean=0, std=1

[65]:
```
df['total_bill_std']=skp.StandardScaler().fit_transform(df['total_bill'].values.
↪reshape(-1,1))
```

1.0.3 Exclude Outliers

[66]:
```
# df[df.apply(lambda x:np.abs(x-x.mean())/x.std() <3).all(axis=1)]
```

1.0.4 Label Encoding (categorical –> numerical labels)

[69]:
```
# skp.LabelEncoder().fit_transform(df['day'])
```

1.0.5 Label Binarization

[73]:
```
labels=skp.LabelBinarizer().fit_transform(df['day'])
tmpdf=pd.DataFrame(labels, columns=sorted(df.day.unique()))
pd.concat([df,tmpdf], axis=1)
```

[73]:

	total_bill	tip	sex	smoker	day	time	size	total_bill_norm	\
0	16.99	1.01	Female	No	Sun	Dinner	2	0.291579	
1	10.34	1.66	Male	No	Sun	Dinner	3	0.152283	
2	21.01	3.50	Male	No	Sun	Dinner	3	0.375786	
3	23.68	3.31	Male	No	Sun	Dinner	2	0.431713	
4	24.59	3.61	Female	No	Sun	Dinner	4	0.450775	
..	…	…	…	…	…	…	…	…	
239	29.03	5.92	Male	No	Sat	Dinner	3	0.543779	
240	27.18	2.00	Female	Yes	Sat	Dinner	2	0.505027	
241	22.67	2.00	Male	Yes	Sat	Dinner	2	0.410557	
242	17.82	1.75	Male	No	Sat	Dinner	2	0.308965	
243	18.78	3.00	Female	No	Thur	Dinner	2	0.329074	

	total_bill_std	Fri	Sat	Sun	Thur
0	-0.314711	0	0	1	0
1	-1.063235	0	0	1	0
2	0.137780	0	0	1	0
3	0.438315	0	0	1	0
4	0.540745	0	0	1	0
..	…	…	…	…	…
239	1.040511	0	1	0	0
240	0.832275	0	1	0	0
241	0.324630	0	1	0	0
242	-0.221287	0	1	0	0
243	-0.113229	0	0	0	1

[244 rows x 13 columns]

1.0.6 Data Binning (Continues –> categorical)

[78]:
```
df['tip_category']=skp.KBinsDiscretizer(n_bins=3, encode='ordinal').
↪fit_transform(df['tip'].values.reshape(-1,1))
```

[81]:
```
df['tip_category'].replace(0, 'small' , inplace=True)
df['tip_category'].replace(1, 'average' , inplace=True)
df['tip_category'].replace(2, 'large' , inplace=True)
```

[82]:
```
df
```

[82]:

	total_bill	tip	sex	smoker	day	time	size	total_bill_norm	total_bill_std	tip_category
0	16.99	1.01	Female	No	Sun	Dinner	2	0.291579	-0.314711	small
1	10.34	1.66	Male	No	Sun	Dinner	3	0.152283	-1.063235	small
2	21.01	3.50	Male	No	Sun	Dinner	3	0.375786	0.137780	large
3	23.68	3.31	Male	No	Sun	Dinner	2	0.431713	0.438315	large
4	24.59	3.61	Female	No	Sun	Dinner	4	0.450775	0.540745	large
..	…	…	…	…	…	…	…	…	…	…
239	29.03	5.92	Male	No	Sat	Dinner	3	0.543779	1.040511	large
240	27.18	2.00	Female	Yes	Sat	Dinner	2	0.505027	0.832275	small
241	22.67	2.00	Male	Yes	Sat	Dinner	2	0.410557	0.324630	small
242	17.82	1.75	Male	No	Sat	Dinner	2	0.308965	-0.221287	small
243	18.78	3.00	Female	No	Thur	Dinner	2	0.329074	-0.113229	average

[244 rows x 10 columns]

[ ]:

[ ]: