In [7]:
```python
import pandas as pd
df = pd.read_csv("E:/tips.csv")
```

In [8]:
```python
df.head()
```

Out[8]:

	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

In [9]:
```python
df.tail()
```

Out[9]:

	total_bill	tip	sex	smoker	day	time	size
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

In [10]:
```python
df
```

Out[10]:

	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
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [11]:
```python
df.shape
```

Out[11]:
```
(244, 7)
```

In [14]:
```python
df.columns
```

Out[14]:
```
Index(['total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'size'], dtype='object')
```

In [22]:
```python
df["sex"].value_counts()
```

Out[22]:
```
Male      157
Female     87
Name: sex, dtype: int64
```

In [24]:
```python
df["tip"].value_counts()
```

Out[24]:
```
2.00    33
3.00    23
4.00    12
5.00    10
2.50    10
        ..
4.34     1
1.56     1
5.20     1
2.60     1
1.75     1
Name: tip, Length: 123, dtype: int64
```

In [28]:
```python
df["total_bill"].value_counts()
```

Out[28]:
```
13.42    3
13.81    2
15.98    2
17.92    2
10.07    2
        ..
24.71    1
21.16    1
28.97    1
22.49    1
18.78    1
Name: total_bill, Length: 229, dtype: int64
```

In [5]:
```python
import pandas as pd
df = pd.read_csv("E:/tips.csv")
```

In [6]:
```python
df
```

Out[6]:

	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
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [8]:
```python
df["size"].value_counts()
```

Out[8]:
```
2    156
3     38
4     37
5      5
1      4
6      4
Name: size, dtype: int64
```

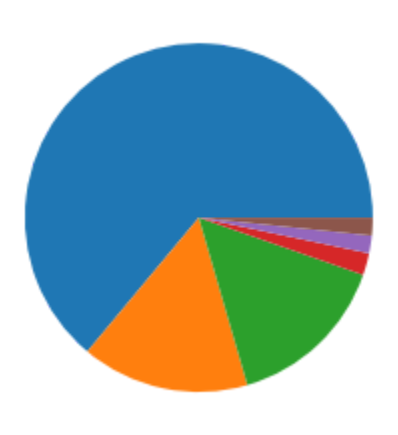
In [9]:
```python
df['size']
```

Out[9]:
```
0      2
1      3
2      3
3      2
4      4
      ..
239    3
240    2
241    2
242    2
243    2
Name: size, Length: 244, dtype: int64
```

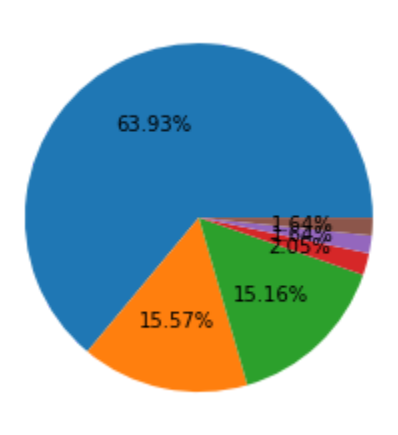
In [16]:
```python
import matplotlib.pyplot as plt
plt
```

Out[16]:
```
<module 'matplotlib.pyplot' from 'E:\\anaconda\\lib\\site-packages\\matplotlib\\pyplot.py'>
```

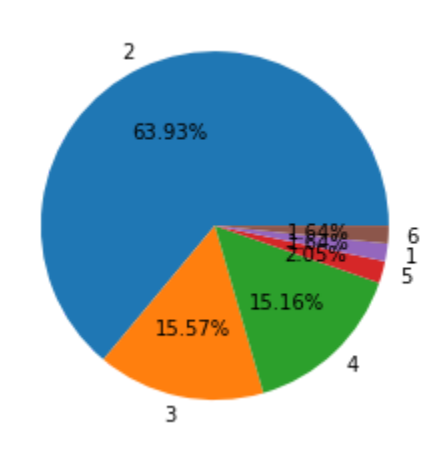
In [51]:
```python
import matplotlib.pyplot as plt
z=plt.pie(df['size'].value_counts())
```



In [50]:
```python
import matplotlib.pyplot as plt
z=plt.pie(df['size'].value_counts(),autopct="%0.2f%%")
```



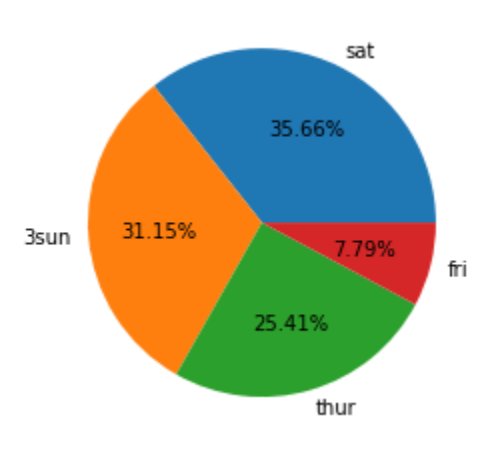
In [46]:
```python
import matplotlib.pyplot as plt
z=plt.pie(df['size'].value_counts(),labels=['2','3','4','5','1','6'],autopct="%0.2f%%")
```



In [35]:
```python
df['day'].value_counts()
```

Out[35]:
```
Sat     87
Sun     76
Thur    62
Fri     19
Name: day, dtype: int64
```

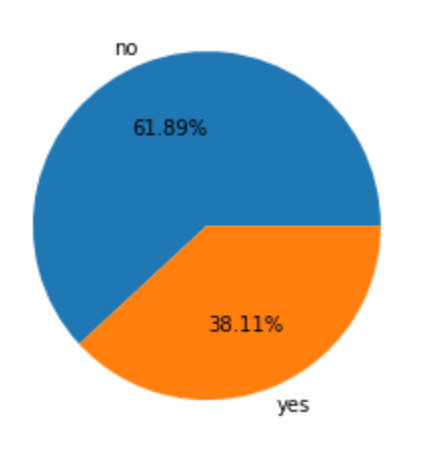
In [47]:
```python
import matplotlib.pyplot as plt
z=plt.pie(df['day'].value_counts(),labels=['sat','3sun','thur','fri'],autopct="%0.2f%%")
```



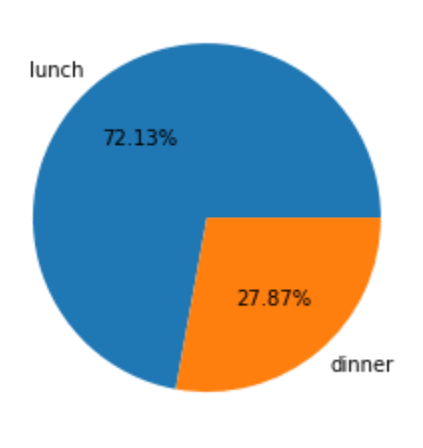
In [39]:
```python
df['smoker'].value_counts()
```

Out[39]:
```
No     151
Yes     93
Name: smoker, dtype: int64
```

In [48]:
```python
import matplotlib.pyplot as plt
z=plt.pie(df['smoker'].value_counts(),labels=['no','yes'],autopct="%0.2f%%")
```



In [49]:
```python
import matplotlib.pyplot as plt
z=plt.pie(df['time'].value_counts(),labels=['lunch','dinner',],autopct="%0.2f%%")
```



In [43]:
```python
df['total_bill'].max()
```

Out[43]:
```
50.81
```

In [45]:
```python
df['total_bill'].min()
```

Out[45]:
```
3.07
```

In [53]:
```python
df['tip'].value_counts()
```

Out[53]:
```
2.00    33
3.00    23
4.00    12
5.00    10
2.50    10
        ..
4.34     1
1.56     1
5.20     1
2.60     1
1.75     1
Name: tip, Length: 123, dtype: int64
```

In [54]:
```python
df['tip'].max()
```

Out[54]:
```
10.0
```

In [55]:
```python
df['tip'].min()
```

Out[55]:
```
1.0
```

In [5]:
```python
import pandas as pd
df = pd.read_csv("E:/tips.csv")
```

In [6]:
```python
df
```

Out[6]:

	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
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [7]:
```python
df['day'].unique()
```

Out[7]:
```
array(['Sun', 'Sat', 'Thur', 'Fri'], dtype=object)
```

In [9]:
```python
df['day'].value_counts()
```

Out[9]:
```
Sat     87
Sun     76
Thur    62
Fri     19
Name: day, dtype: int64
```

In [10]:
```python
df['time'].unique()
```

Out[10]:
```
array(['Dinner', 'Lunch'], dtype=object)
```

In [12]:
```python
df['time'].value_counts()
```

Out[12]:
```
Dinner    176
Lunch      68
Name: time, dtype: int64
```

In [13]:
```python
#which day more collection
g=df.groupby(df['day'])
```

In [18]:
```python
g.max()
```

Out[18]:

	total_bill	tip	sex	smoker	time	size
day						
Fri	40.17	4.73	Male	Yes	Lunch	4
Sat	50.81	10.00	Male	Yes	Dinner	5
Sun	48.17	6.50	Male	Yes	Dinner	6
Thur	43.11	6.70	Male	Yes	Lunch	6

In [19]:
```python
#which day less collection
g.min()
```

Out[19]:

	total_bill	tip	sex	smoker	time	size
day						
Fri	5.75	1.00	Female	No	Dinner	1
Sat	3.07	1.00	Female	No	Dinner	1
Sun	7.25	1.01	Female	No	Dinner	2
Thur	7.51	1.25	Female	No	Dinner	1

In [20]:
```python
g.max().sum()
```

Out[20]:
```
total_bill                    182.26
tip                            27.93
sex                 MaleMaleMaleMale
smoker                  YesYesYesYes
time           LunchDinnerDinnerLunch
size                              21
dtype: object
```

In [21]:
```python
df['size'].max()
```

Out[21]:
```
6
```

In [22]:
```python
df.loc[df['size']==6]
```

Out[22]:

	total_bill	tip	sex	smoker	day	time	size
125	29.80	4.2	Female	No	Thur	Lunch	6
141	34.30	6.7	Male	No	Thur	Lunch	6
143	27.05	5.0	Female	No	Thur	Lunch	6
156	48.17	5.0	Male	No	Sun	Dinner	6

In [1]:
```python
import pandas as pd
df = pd.read_csv("E:/tips.csv")
```

In [2]:
```python
df
```

Out[2]:

	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
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [3]:
```python
df['total_bill'].max()
```

Out[3]:
```
50.81
```

In [5]:
```python
df.loc[df['total_bill']==50.81]
```

Out[5]:

	total_bill	tip	sex	smoker	day	time	size
170	50.81	10.0	Male	Yes	Sat	Dinner	3

In [7]:
```python
df['tip'].max()
```

Out[7]:
```
10.0
```

In [8]:
```python
df.loc[df['tip']==10.0]
```

Out[8]:

	total_bill	tip	sex	smoker	day	time	size
170	50.81	10.0	Male	Yes	Sat	Dinner	3

In [10]:
```python
# only female data
df.loc[df['sex']=='Female']
```

Out[10]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
11	35.26	5.00	Female	No	Sun	Dinner	4
14	14.83	3.02	Female	No	Sun	Dinner	2
16	10.33	1.67	Female	No	Sun	Dinner	3
...	...	...	...	...	...	...	...
226	10.09	2.00	Female	Yes	Fri	Lunch	2
229	22.12	2.88	Female	Yes	Sat	Dinner	2
238	35.83	4.67	Female	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

87 rows × 7 columns

In [12]:
```python
# only male data
df.loc[df['sex']=='Male']
```

Out[12]:

	total_bill	tip	sex	smoker	day	time	size
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
5	25.29	4.71	Male	No	Sun	Dinner	4
6	8.77	2.00	Male	No	Sun	Dinner	2
...	...	...	...	...	...	...	...
236	12.60	1.00	Male	Yes	Sat	Dinner	2
237	32.83	1.17	Male	Yes	Sat	Dinner	2
239	29.03	5.92	Male	No	Sat	Dinner	3
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2

157 rows × 7 columns

In [21]:
```python
# machine learning
# convert the string values to numeric
from sklearn.preprocessing import LabelEncoder
le1=LabelEncoder()
le2=LabelEncoder()
le3=LabelEncoder()
le4=LabelEncoder()
df['sex']=le1.fit_transform(df['sex'])
df['smoker']=le2.fit_transform(df['smoker'])
df['day']=le3.fit_transform(df['day'])
df['time']=le4.fit_transform(df['time'])
```

In [22]:
```python
df
```

Out[22]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	0	0	2	0	2
1	10.34	1.66	1	0	2	0	3
2	21.01	3.50	1	0	2	0	3
3	23.68	3.31	1	0	2	0	2
4	24.59	3.61	0	0	2	0	4
...	...	...	...	...	...	...	...
239	29.03	5.92	1	0	1	0	3
240	27.18	2.00	0	1	1	0	2
241	22.67	2.00	1	1	1	0	2
242	17.82	1.75	1	0	1	0	2
243	18.78	3.00	0	0	3	0	2

244 rows × 7 columns

In [31]:
```python
# separate input and output
X=df.drop(columns='tip')
Y=df['tip']
```

In [33]:
```python
# split the data for training and testing
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2)
```

In [35]:
```python
X_train
```

Out[35]:

	total_bill	sex	smoker	day	time	size
104	20.92	0	0	1	0	2
82	10.07	0	0	3	1	1
235	10.07	1	0	1	0	2
141	34.30	1	0	3	1	6
174	16.82	1	1	2	0	2
...	...	...	...	...	...	...
119	24.08	0	0	3	1	4
31	18.35	1	0	1	0	4
208	24.27	1	1	1	0	2
216	28.15	1	1	1	0	5
3	23.68	1	0	2	0	2

195 rows × 6 columns

In [36]:
```python
Y_train
```

Out[36]:
```
104    4.08
82     1.83
235    1.25
141    6.70
174    4.00
       ...
119    2.92
31     2.50
208    2.03
216    3.00
3      3.31
Name: tip, Length: 195, dtype: float64
```

In [42]:
```python
# create a ML model
from sklearn.neighbors import KNeighborsRegressor
K=KNeighborsRegressor(n_neighbors=5)
# train the model
K.fit(X_train,Y_train)
```

Out[42]:
```
KNeighborsRegressor()
```

In [43]:
```python
# test the model
Y_pred=K.predict(X_test)
```

In [44]:
```python
Y_pred
```

Out[44]:
```
array([3.46 , 2.302, 3.29 , 2.582, 5.05 , 2.644, 3.616, 4.168, 4.516,
       2.11 , 1.572, 3.732, 2.244, 1.924, 2.038, 2.112, 2.138, 2.918,
       1.872, 3.026, 4.256, 2.55 , 4.142, 4.132, 1.584, 2.802, 3.   ,
       1.77 , 3.146, 3.042, 2.448, 2.778, 2.406, 2.548, 2.038, 2.632,
       2.506, 3.234, 3.124, 2.344, 1.64 , 1.792, 2.002, 5.446, 1.584,
       3.304, 3.086, 3.45 , 4.45 ])
```

In [46]:
```python
Y_test.values
```

Out[46]:
```
array([3.5 , 3.07, 3.5 , 3.  , 5.17, 2.64, 3.48, 1.17, 4.73, 1.68, 1.  ,
       3.75, 2.5 , 1.5 , 5.15, 2.  , 1.44, 3.08, 1.47, 3.27, 5.  , 1.64,
       5.07, 3.09, 2.  , 3.  , 3.21, 1.67, 2.71, 6.5 , 3.71, 4.3 , 3.48,
       2.47, 2.52, 2.24, 2.02, 3.48, 3.5 , 2.  , 1.58, 1.5 , 1.98, 2.5 ,
       1.48, 6.5 , 3.5 , 1.5 , 3.  ])
```

In [52]:
```python
# find mean squared error
from sklearn.metrics import mean_squared_error
mse=mean_squared_error(Y_test,Y_pred)
mse
```

Out[52]:
```
1.3855133061224494
```

In [53]:
```python
# r 2 score
from sklearn.metrics import r2_score
rse=r2_score(Y_test,Y_pred)
rse
```

Out[53]:
```
0.2003851265010247
```

In [ ]: