In [ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import sklearn as sklearn
df = pd.read_csv("autodata2.csv")
df.head(5)
```

Out[ ]:

	Unnamed: 0	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels
0	0	3	122	alfa-romero	std	two	convertible	rwd
1	1	3	122	alfa-romero	std	two	convertible	rwd
2	2	1	122	alfa-romero	std	two	hatchback	rwd
3	3	2	164	audi	std	four	sedan	fwd
4	4	2	164	audi	std	four	sedan	4wd

5 rows × 30 columns

In [ ]:
```
bool_series = pd.isnull(df["price"])
print("missing values in price attribute are:\n",df[bool_series])
```

```
missing values in price attribute are:
    Unnamed: 0  symboling  normalized-losses         make aspiration  \
0           0          3                122  alfa-romero        std   \
2           2          1                122  alfa-romero        std

  num-of-doors   body-style drive-wheels engine-location  wheel-base  ...  \
0          two  convertible          rwd           front        88.6  ...   \
2          two    hatchback          rwd           front        94.5  ...

   compression-ratio  horsepower  peak-rpm  city-mpg highway-mpg price  \
0                9.0       111.0       NaN        21          27   NaN   \
2                9.0       154.0       NaN        19          26   NaN

   city-L/100km horsepower-binned  diesel  gas
0     11.190476               Low       0    1
2     12.368421            Medium       0    1

[2 rows x 30 columns]
```

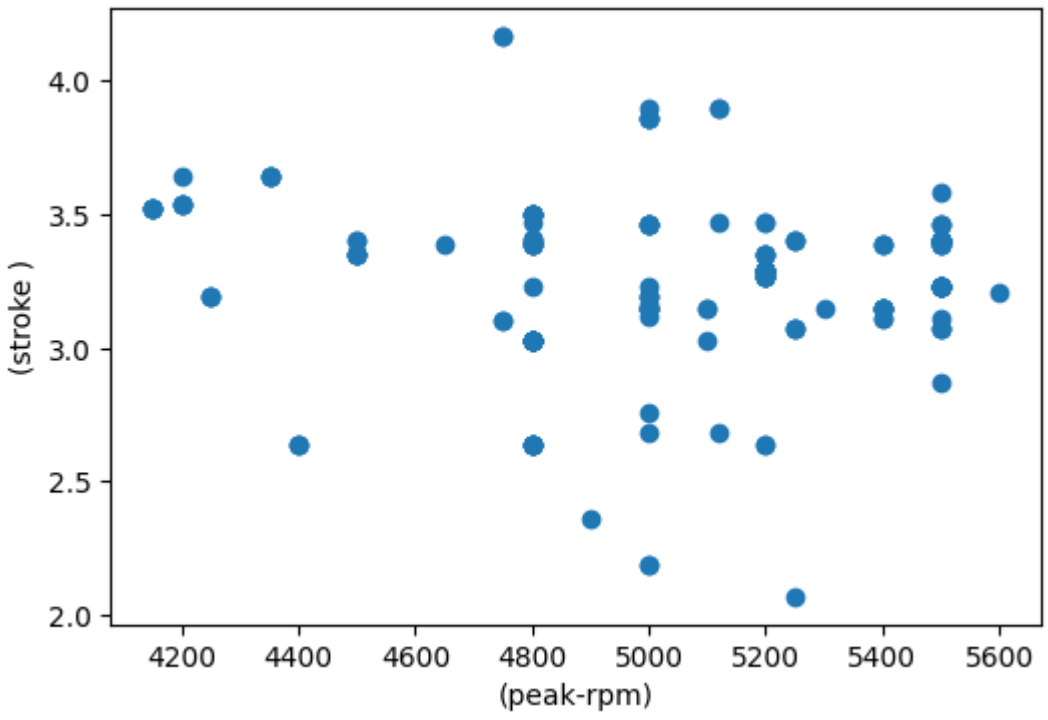
In [ ]:
```
print("missing values replaced by -9999:\n",df["price"].replace(to_replace = np.
```

```
missing values replaced by -9999:
 0     -99999.0
1      16500.0
2     -99999.0
3      13950.0
4      17450.0
         ...
196    16845.0
197    19045.0
198    21485.0
199    22470.0
200    22625.0
Name: price, Length: 201, dtype: float64
```

In [ ]:
```
avg_rpm = df["peak-rpm"].astype("int").mean(axis = 0)
print("missing values replaced by average",df["peak-rpm"].replace(np.NAN, value=
```

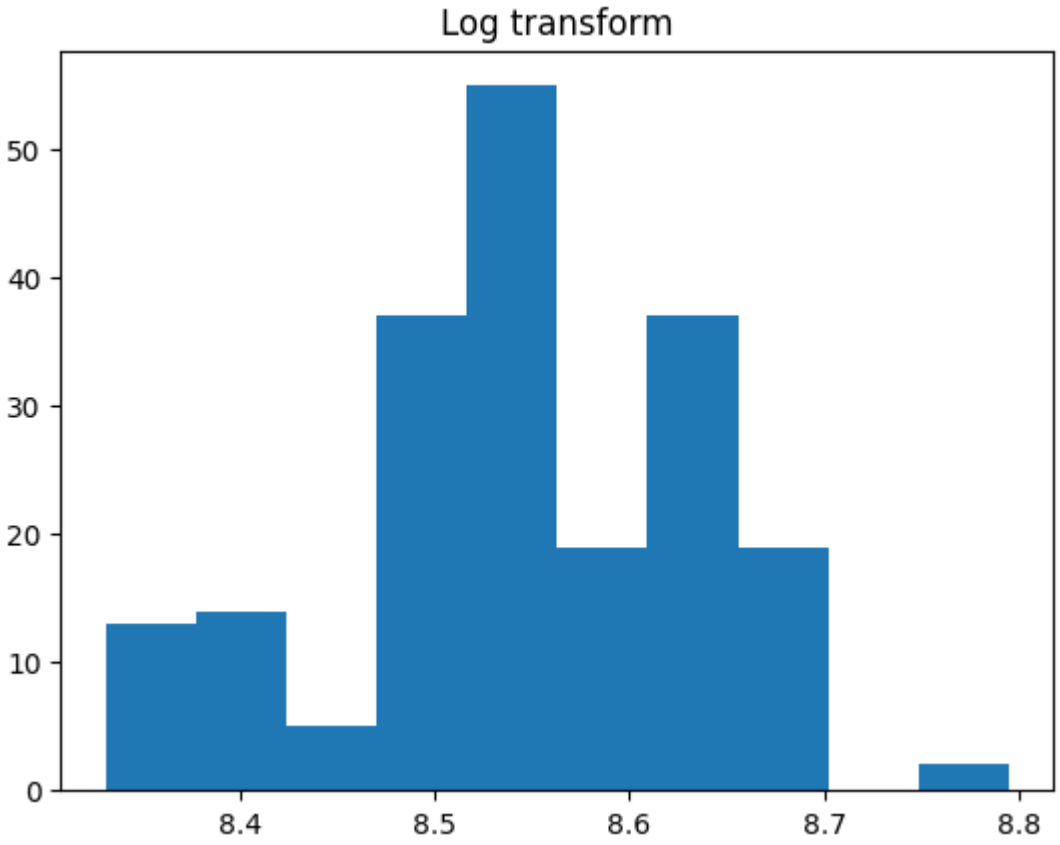
```
missing values replaced by average 0      5118.781726
1      5000.000000
2      5118.781726
3      5500.000000
4      5500.000000
          ...
196    5400.000000
197    5300.000000
198    5500.000000
199    4800.000000
200    5400.000000
Name: peak-rpm, Length: 201, dtype: float64
```

In [ ]:
```
fig, ax = plt.subplots(figsize=(6, 4))
ax.scatter(no_outliers['peak-rpm'],no_outliers['stroke'])
ax.set_xlabel('(peak-rpm)')
ax.set_ylabel('(stroke )')
plt.show()
```



In [ ]:
```
outlier_indices=np.where((df['peak-rpm'] >5750) & (df['stroke']<10.5))
no_outliers = df.drop(outlier_indices[0])
```

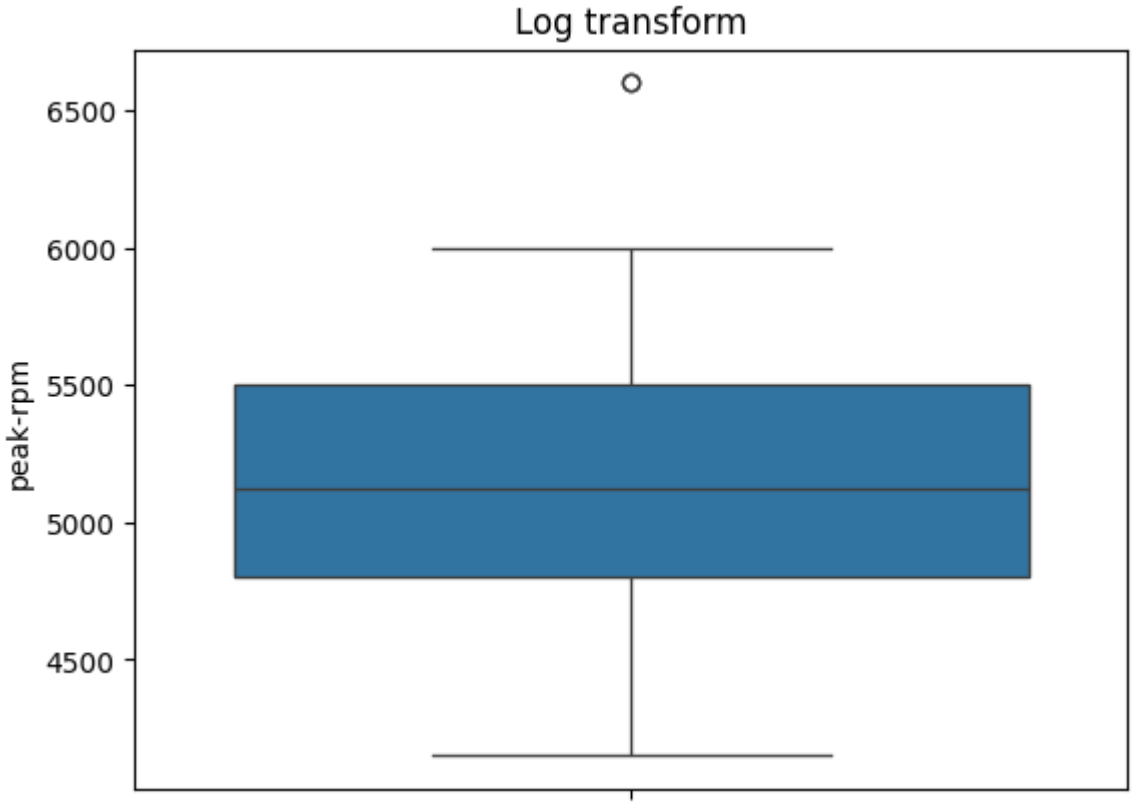
```
plt.hist(np.log(df['peak-rpm']));
plt.title('Log transform');
plt.show()
```



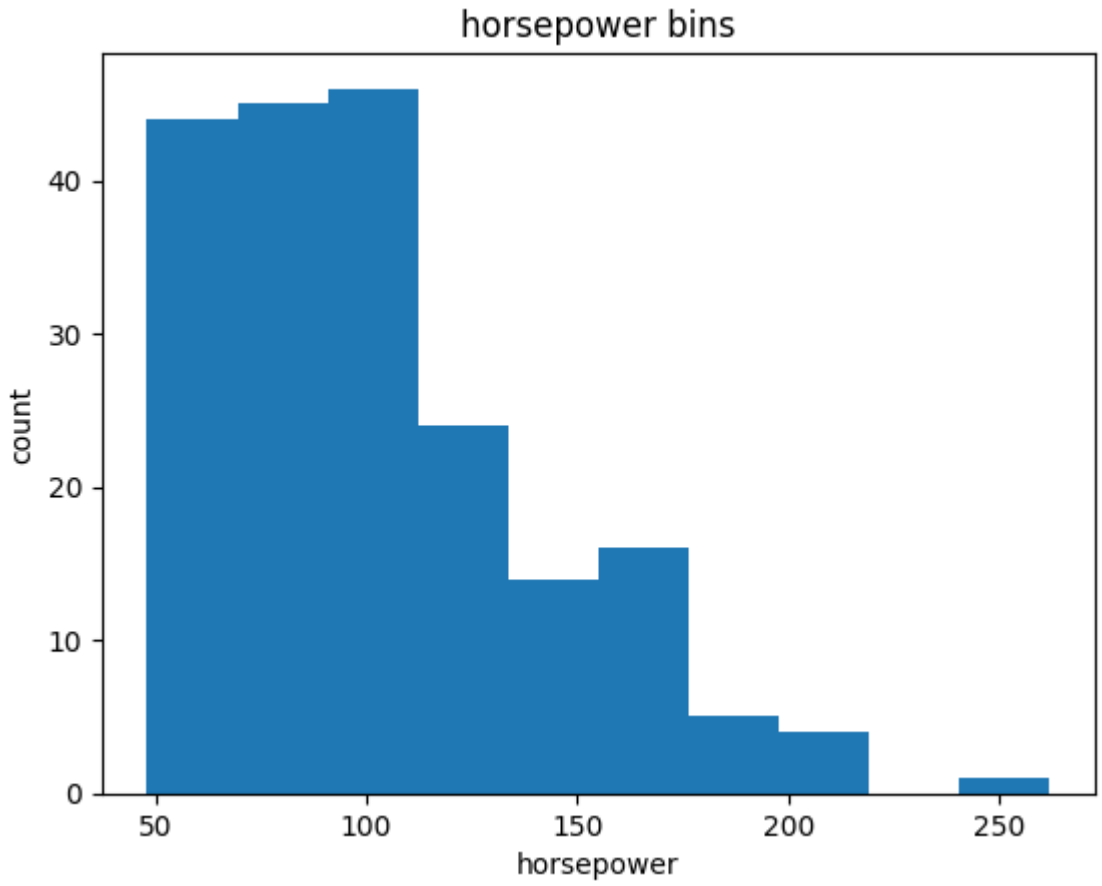
In [ ]:

```
import seaborn as sns
outlier_indices=np.where((df['peak-rpm'] >5750) & (df['stroke']<10.5))
no_outliers = df.drop(outlier_indices[0])

sns.boxplot((df['peak-rpm']));
plt.title('Log transform');
plt.show()
```



In [ ]:
```
df["horsepower"]=df["horsepower"].astype(float, copy=True)
plt.hist(df["horsepower"])
plt.xlabel("horsepower")
plt.ylabel("count")
plt.title("horsepower bins")
plt.show()
bins = np.linspace(min(df["horsepower"]), max(df["horsepower"]), 4)
group_names = ['Low', 'Medium', 'High']
df['horsepower-binned'] = pd.cut(df['horsepower'], bins, labels=group_names, inc
print("Binning\n",df[['horsepower','horsepower-binned']].head(20))
```




```
Binning
     horsepower horsepower-binned
0         111.0               Low
1         111.0               Low
2         154.0            Medium
3         102.0               Low
4         115.0               Low
5         110.0               Low
6         110.0               Low
7         110.0               Low
8         140.0            Medium
9         101.0               Low
10        101.0               Low
11        121.0            Medium
12        121.0            Medium
13        121.0            Medium
14        182.0            Medium
15        182.0            Medium
16        182.0            Medium
17         48.0               Low
18         70.0               Low
19         70.0               Low
```

In [ ]: