In [ ]:
```
DM Assignment 3 Data Preprocessing
```

In [2]:
```
import pandas as pd
from numpy import random
import matplotlib.pyplot as plotting
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

In [4]:
```
pd.set_option('display.max_columns', None)
pd.set_option('max_colwidth', None)
df = pd.read_csv("dat.data",header=None, delim_whitespace=True)
df = df.rename(columns={0: 'mpg', 1: 'cylinders', 2: 'displacement', 3: 'horsepo
                        4: 'weight', 5: 'acceleration', 6: 'model year', 7: 'ori
                        8: 'car name'})
```

In [ ]:
```
Question 1:-
    Cars and attributes in dataset
```

In [6]:
```
print("\nUnique cars ->",len(df['car name'].unique()))
print("Attributes ->",len(df.columns))
```

```
Unique cars -> 305
Attributes -> 9
```

In [ ]:
```
Question 2:-
    How many distinct car companies are represented in the data set? What is the
    Do some internet search that can tell you about the history and popularity o
```

In [7]:
```
cars=df['car name']
mpg=df['mpg']
distinctCarCompanies=[]
for i in cars:
    i=i.split(" ")
    if i[0] not in distinctCarCompanies:
        distinctCarCompanies.append(i[0])

print("\nDist car Companies ->",len(distinctCarCompanies))
print("Best mpg car ->",df.loc[df['mpg'] == mpg.max(),'car name'].iloc[0],
      "mpg -> ",mpg.max())
```

```
Dist car Companies -> 37
Best mpg car -> mazda glc mpg ->  46.6
```

In [9]:
```
eight_cylinders_cars=df.loc[df['cylinders'] == 8, 'car name']
eight_cylinders_car_companies=[]
for i in eight_cylinders_cars:
    i=i.split(" ")
    eight_cylinders_car_companies.append(i[0])

counter = 0
```

```
frequent_car_company = eight_cylinders_car_companies[0]
for i in eight_cylinders_car_companies:
    curr_frequency = eight_cylinders_car_companies.count(i)
    if(curr_frequency> counter):
       counter = curr_frequency
       frequent_car_company = i

print("Company with most 8-cylinder cars ->",frequent_car_company)
three_cylinders=df.loc[df['cylinders'] == 3, 'car name'].tolist()
print("Cars with three cylinders ->",', '.join(three_cylinders))
```

```
Company with most 8-cylinder cars -> ford
Cars with three cylinders -> mazda rx2 coupe, maxda rx3, mazda rx-4, mazda rx-7
gs
```

In [ ]:

```
Question 3:-
    What is the range, mean, and standard deviation of each attribute? Pay atten
    missing values.
```

In [10]:

```
df['horsepower'] = df['horsepower'].apply(lambda x: float(x.replace('?','NaN')))
print(df.describe(percentiles=[0.5])[1:])
```

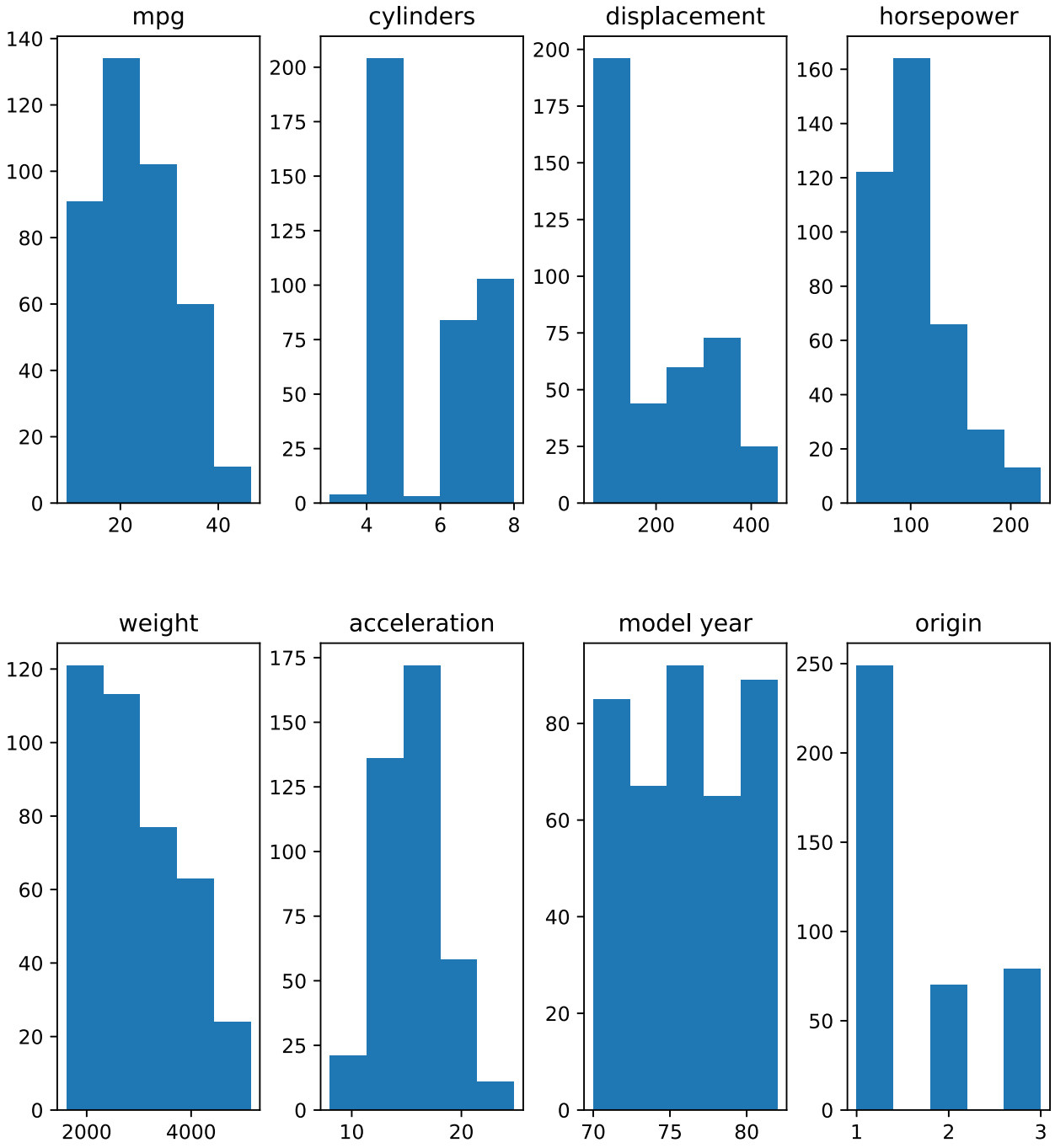
```
              mpg  cylinders  displacement  horsepower       weight  \
mean    23.514573   5.454774    193.425879  104.469388  2970.424623
std      7.815984   1.701004    104.269838   38.491160   846.841774
min      9.000000   3.000000     68.000000   46.000000  1613.000000
50%     23.000000   4.000000    148.500000   93.500000  2803.500000
max     46.600000   8.000000    455.000000  230.000000  5140.000000

       acceleration  model year    origin
mean      15.568090   76.010050  1.572864
std        2.757689    3.697627  0.802055
min        8.000000   70.000000  1.000000
50%       15.500000   76.000000  1.000000
max       24.800000   82.000000  3.000000
```

In [ ]:

```
Question 4:-
    Plot histograms for each attribute. Pay attention to the appropriate choice
    Write 2-3 sentences summarizing some interesting aspects of the data by look
```

In [13]:

```
df.hist(bins=5,grid=False,layout=[2,4],figsize=[9,10])
plotting.show()
print()
```

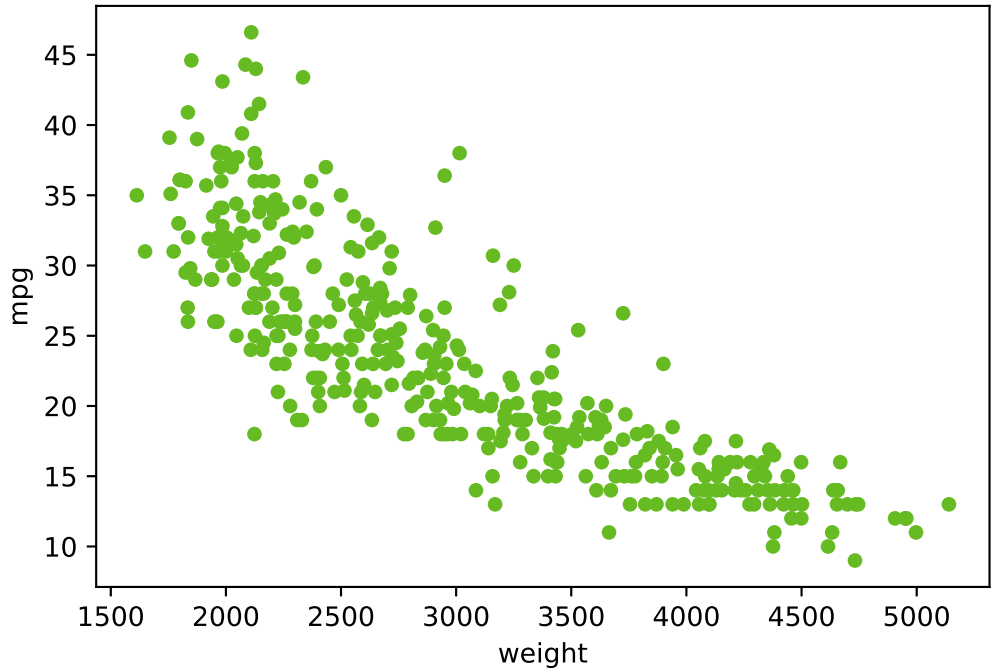


In [ ]:

```
Question 5:-
    Plot a scatterplot of weight vs. MPG attributes. What do you conclude about
    between the attributes? What is the correlation coefficient between the 2 at
```

In [14]:

```
df.plot.scatter(x='weight', y='mpg', c='#6B2')
plotting.show()
print("\nThe correlation coefficient between weight and mpg is",
      df['weight'].corr(df['mpg']))
```




```
The correlation coefficient between weight and mpg is -0.8317409332443352
```

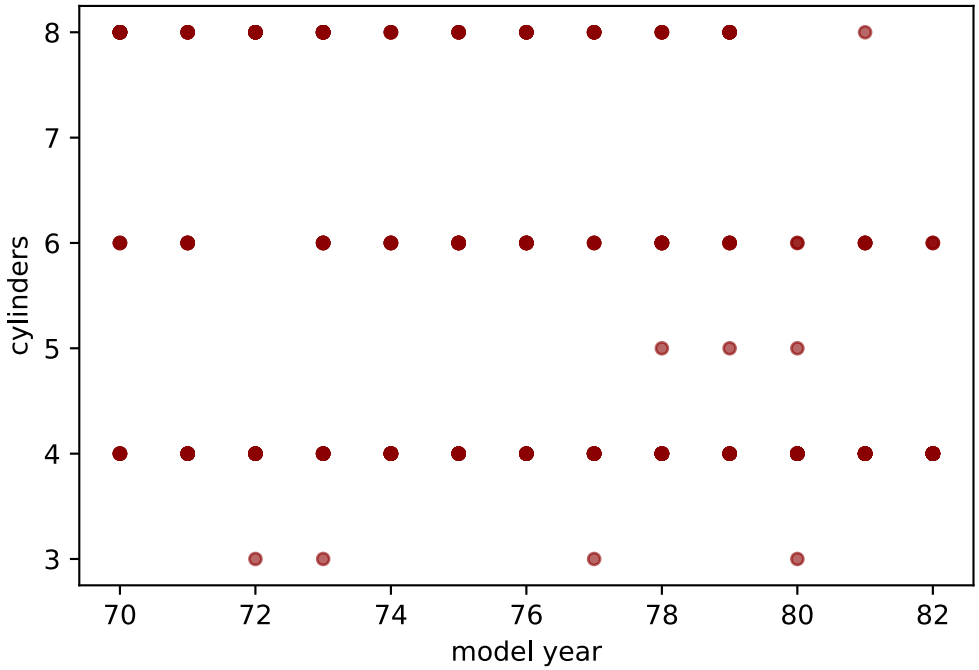
In [ ]:

```
Question 6:-
    Plot a scatterplot of year vs. cylinders attributes. Add a small random nois
    the scatterplot look nicer. What can you conclude? Do some internet search a
    car industry during 70's that might explain the results
```

In [16]:

```
print("\nBEFORE ADDING RANDOM NOISE")
print("The scatterplot of model year vs cylinders is given below:")
df.plot.scatter(x='model year', y='cylinders', c='DarkRed',alpha=0.6)
plotting.show()
```

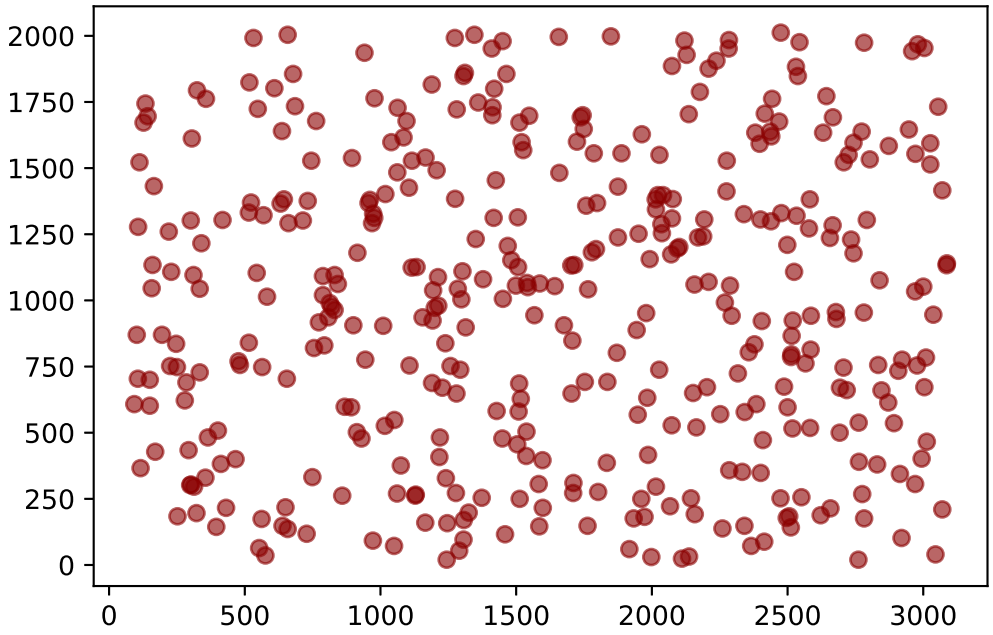
```
BEFORE ADDING RANDOM NOISE
The scatterplot of model year vs cylinders is given below:
```



In [17]:
```
cylinders=list(df['cylinders'])
modelyear=list(df['model year'])

for i in range(len(cylinders)):
    cylinders[i]+=random.randint(0,1000)*2+10
    modelyear[i]+=random.randint(0,1000)*3+15
```

In [18]:
```
print("\nAFTER ADDING RANDOM NOISE")
print("The scatterplot of model year vs cylinders is given below:")
plotting.scatter(modelyear,cylinders, c='DarkRed',alpha=0.6)
plotting.show()
```

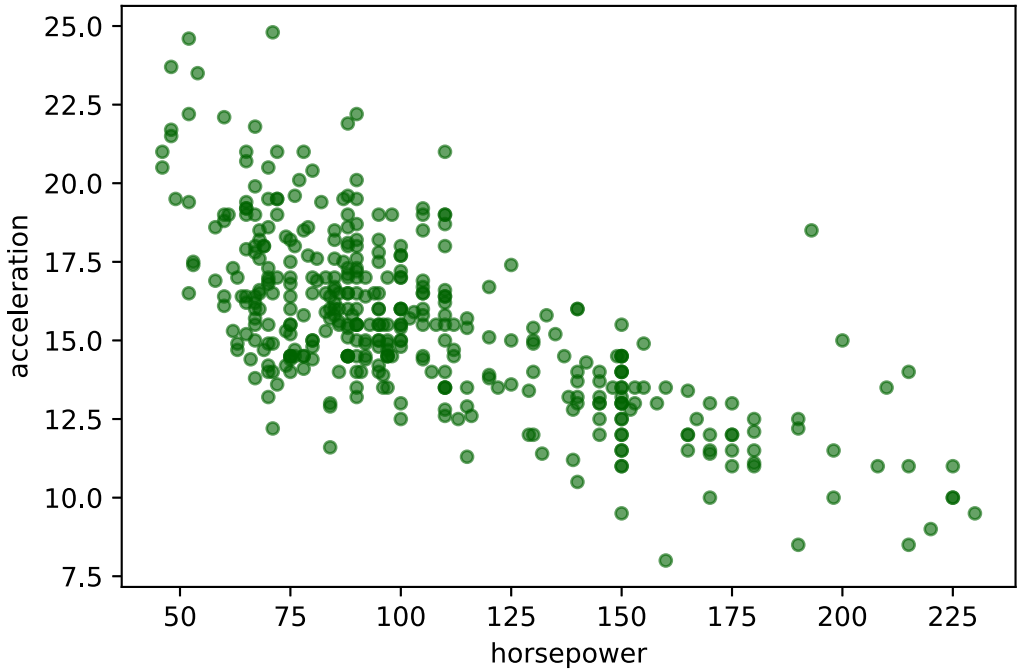
```
AFTER ADDING RANDOM NOISE
The scatterplot of model year vs cylinders is given below:
```



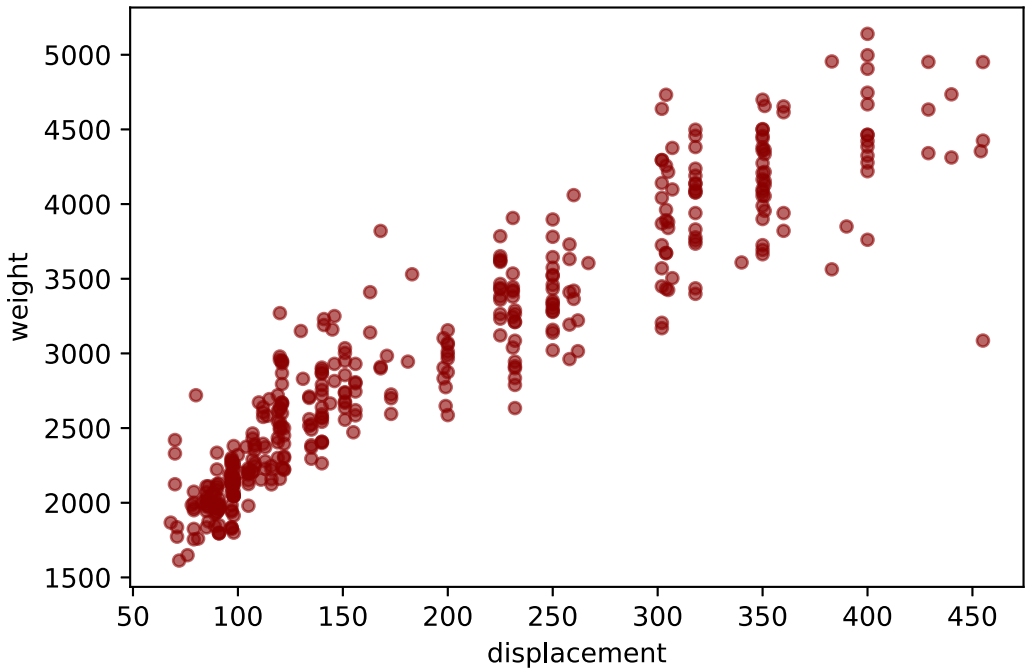
In [ ]:
```
Question 7:-
    Show 2 more scatterplots that are interesting do you. Discuss what you see.
```

In [19]:
```
print("The scatterplot of horsepower vs acceleration is given below:")
df.plot.scatter(x='horsepower', y='acceleration', c='DarkGreen',alpha=0.6)
plotting.show()
print()

print("The scatterplot of displacement vs weight is given below:")
df.plot.scatter(x='displacement', y='weight', c='DarkRed',alpha=0.6)
plotting.show()
```

```
The scatterplot of horsepower vs acceleration is given below:
```



The scatterplot of displacement vs weight is given below:



In [ ]:

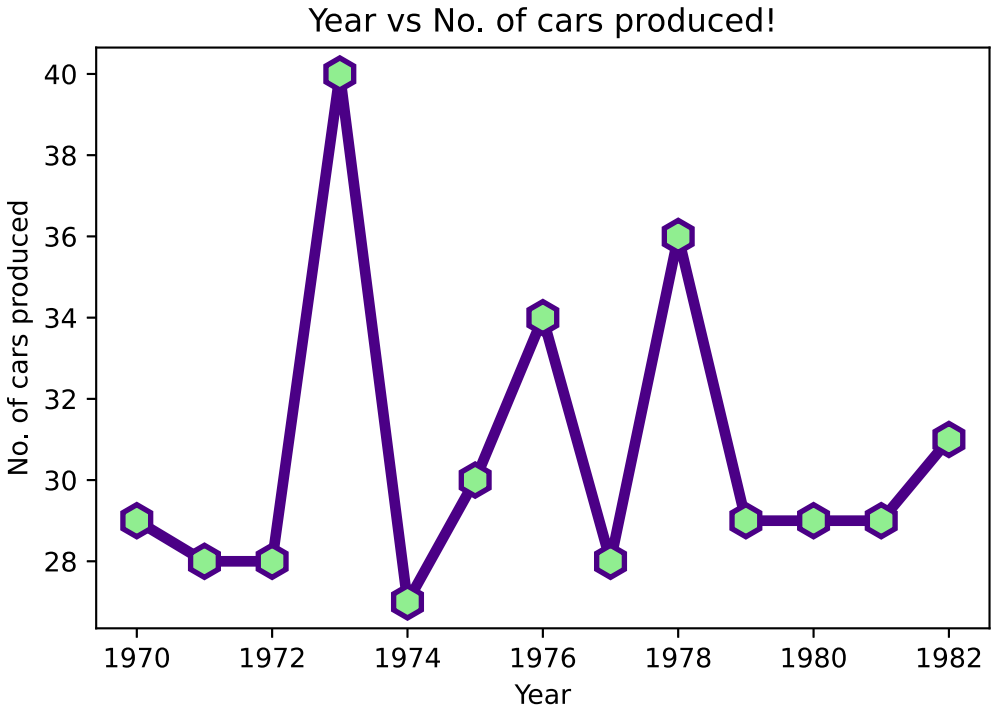
```
Question 8:-
    Plot a time series for all the companies that show how many new cars they in
    each year. Do you see some interesting trends?
```

In [20]:

```
dictValues = {}
for i in df['model year']:
    if i not in dictValues.keys():
        newCars = df.loc[df['model year'] == i, 'car name']
        i+=1900
        dictValues[i]=newCars.shape[0]

x=list(dictValues.keys())
y=list(dictValues.values())
```

```
print("The time series plot between years and number of cars produced is shown b
plotting.plot(x, y, color='#4b0086', linewidth=4, marker='h', markerfacecolor='l
              markeredgewidth=2, markersize=12)
plotting.xlabel('Year')
plotting.ylabel('No. of cars produced')
plotting.title('Year vs No. of cars produced!')
plotting.show()
```

The time series plot between years and number of cars produced is shown below:



In [ ]:

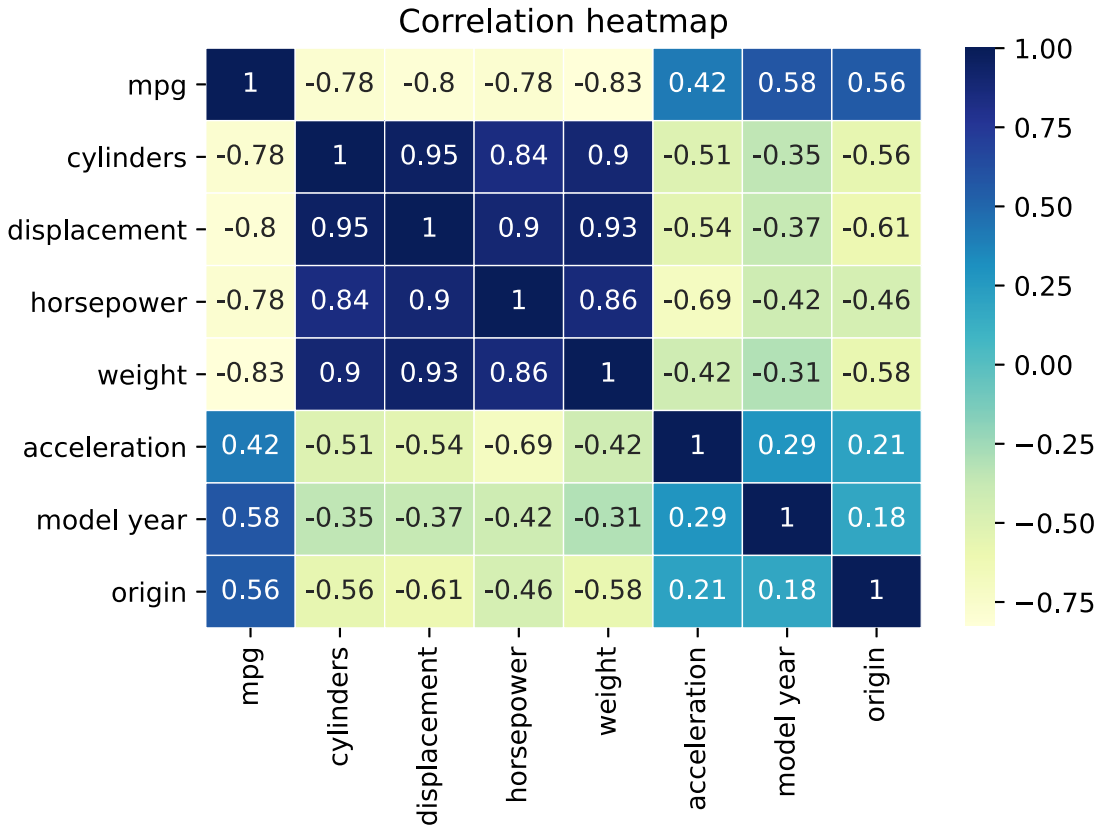
```
Question 9:-
    Calculate the pairwise correlation, and draw the heatmap with Matplotlib. Do
    interesting correlation?
```

In [21]:

```
correlation = df.corr()
heatmap = sns.heatmap(correlation, cbar=True, annot=True, cmap="YlGnBu", linewid
heatmap.set_title("Correlation heatmap")
```

Out[21]:

```
Text(0.5, 1.0, 'Correlation heatmap')
```


Correlation heatmap
mpg
cylinders
displacement
horsepower
weight
acceleration
model year
origin
1 -0.78 -0.8 -0.78 -0.83 0.42 0.58 0.56
-0.78 1 0.95 0.84 0.9 -0.51 -0.35 -0.56
-0.8 0.95 1 0.9 0.93 -0.54 -0.37 -0.61
-0.78 0.84 0.9 1 0.86 -0.69 -0.42 -0.46
-0.83 0.9 0.93 0.86 1 -0.42 -0.31 -0.58
0.42 -0.51 -0.54 -0.69 -0.42 1 0.29 0.21
0.58 -0.35 -0.37 -0.42 -0.31 0.29 1 0.18
0.56 -0.56 -0.61 -0.46 -0.58 0.21 0.18 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75