In [ ]:
```
"""
Lian Gan A16998869
Yihui Zhang A16631173
Rui Yan A15875396
"""
```

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os
```

In [ ]:
```
df_raw = os.path.join('mtcars.csv')
df = pd.read_csv(df_raw)
c_df = df.loc[:, df.columns!='model']
```

Q1 sample mean

In [ ]:
```
c_df.mean()
```

Out[ ]:
```
mpg      20.090625
cyl       6.187500
disp    230.721875
hp      146.687500
drat      3.596563
wt        3.217250
qsec     17.848750
vs        0.437500
am        0.406250
gear      3.687500
carb      2.812500
dtype: float64
```

Q1 variance

In [ ]:
```
c_df.var()
```

Out[ ]:
```
mpg        36.324103
cyl         3.189516
disp    15360.799829
hp       4700.866935
drat        0.285881
wt          0.957379
qsec        3.193166
vs          0.254032
am          0.248992
gear        0.544355
carb        2.608871
dtype: float64
```

Q2 sample variance-covariance matrix

In [ ]:
```
c_df.cov()
```

Out[ ]:

	mpg	cyl	disp	hp	drat	wt	qsec
mpg	36.324103	-9.172379	-633.097208	-320.732056	2.195064	-5.116685	4.509149
cyl	-9.172379	3.189516	199.660282	101.931452	-0.668367	1.367371	-1.886855
disp	-633.097208	199.660282	15360.799829	6721.158669	-47.064019	107.684204	-96.05168
hp	-320.732056	101.931452	6721.158669	4700.866935	-16.451109	44.192661	-86.77008
drat	2.195064	-0.668367	-47.064019	-16.451109	0.285881	-0.372721	0.08714
wt	-5.116685	1.367371	107.684204	44.192661	-0.372721	0.957379	-0.30548
qsec	4.509149	-1.886855	-96.051681	-86.770081	0.087141	-0.305482	3.19316
vs	2.017137	-0.729839	-44.377621	-24.987903	0.118649	-0.273661	0.67056
am	1.803931	-0.465726	-36.564012	-8.320565	0.190151	-0.338105	-0.20496
gear	2.135685	-0.649194	-50.802621	-6.358871	0.275988	-0.421081	-0.28040
carb	-5.363105	1.520161	79.068750	83.036290	-0.078407	0.675790	-1.89411

Q2 sample correlation matrix

In [ ]:
```
c_df.corr()
```

Out[ ]:

	mpg	cyl	disp	hp	drat	wt	qsec	vs
mpg	1.000000	-0.852162	-0.847551	-0.776168	0.681172	-0.867659	0.418684	0.664039
cyl	-0.852162	1.000000	0.902033	0.832447	-0.699938	0.782496	-0.591242	-0.810812
disp	-0.847551	0.902033	1.000000	0.790949	-0.710214	0.887980	-0.433698	-0.710416
hp	-0.776168	0.832447	0.790949	1.000000	-0.448759	0.658748	-0.708223	-0.723097
drat	0.681172	-0.699938	-0.710214	-0.448759	1.000000	-0.712441	0.091205	0.440278
wt	-0.867659	0.782496	0.887980	0.658748	-0.712441	1.000000	-0.174716	-0.554916
qsec	0.418684	-0.591242	-0.433698	-0.708223	0.091205	-0.174716	1.000000	0.744535
vs	0.664039	-0.810812	-0.710416	-0.723097	0.440278	-0.554916	0.744535	1.000000
am	0.599832	-0.522607	-0.591227	-0.243204	0.712711	-0.692495	-0.229861	0.168345
gear	0.480285	-0.492687	-0.555569	-0.125704	0.699610	-0.583287	-0.212682	0.206023
carb	-0.550925	0.526988	0.394977	0.749812	-0.090790	0.427606	-0.656249	-0.569607

The first thing that I find through the results from the correlation and covariance matrices is that Correlation of X and Y is inversely related with the covariance matrix of X and Y. Data in both matrices are having the same signs. When the data is positive in covariance matrix, the data in correlation matrix will be much smaller; when the data is negative in covariance matrix, the corresponding data in the correlation data will be larger since the equation is Correlation(X,Y) = cov(X,Y) / sqrt(var(X)var(y)). Take the relationship between the number of cylinders and mpg in the given dataset as the example, for the cyl-mpg data in the

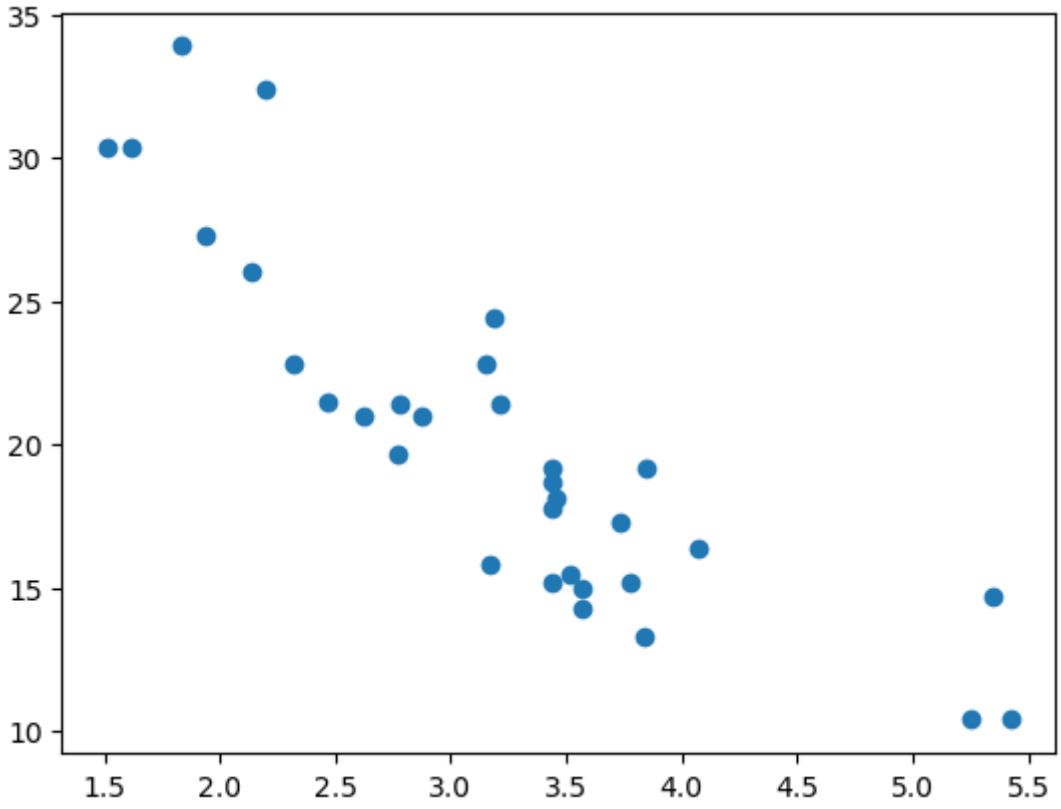
covariance matrix, the number is -9.172379, and the corresponding number in the correlation matrix is -0.85. The correlation matrix also tells the relationship between two variables. If the correlation number is approaching -1, it tells the relation is strongly negative; if the correlation number is approaching 1, it tells the relation is strongly positive; if the correlation number is approaching 0, it tells there is not an obvious relation between the two variables. Take the -0.85 from mpg-cyl as an example, it shows there is a negative relation between the number of cylinders and mpg, which means as the number of cylinders increases, the mpg decreases.

Q3 scatter plot between wt (Weight) and mpg (Miles per gallon)

y axis is weight, x axis is mpg

In [ ]:
```
plt.scatter(x=df['wt'],y=df['mpg'])
```

Out[ ]:
```
<matplotlib.collections.PathCollection at 0x7fd885f60f70>
```



Q4

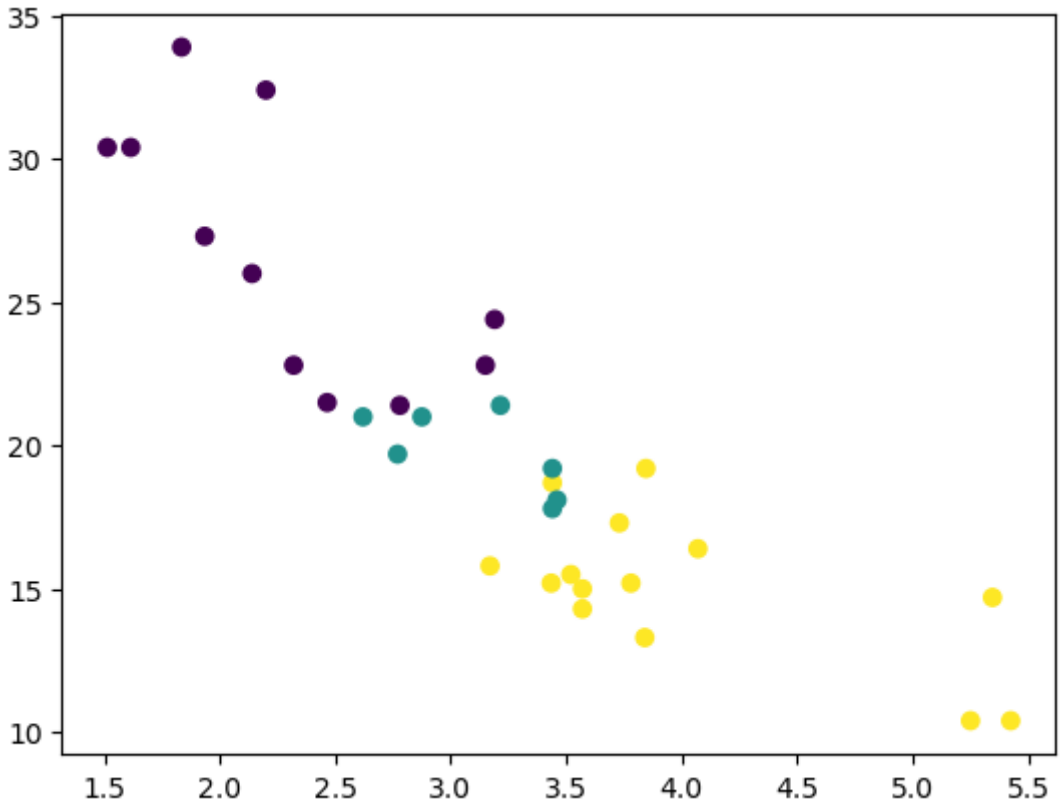
a scatter plot to show the relationship between wt (Weight), mpg (Miles per gallon) and cyl (Number of cylinders).

y axis is weight, x axis is mpg

lighter color(yellow) means higher value in cyl, darker color(purple) means lower value in cyl

In [ ]:
```
plt.scatter(x=df['wt'],y=df['mpg'],c=df['cyl'])
```

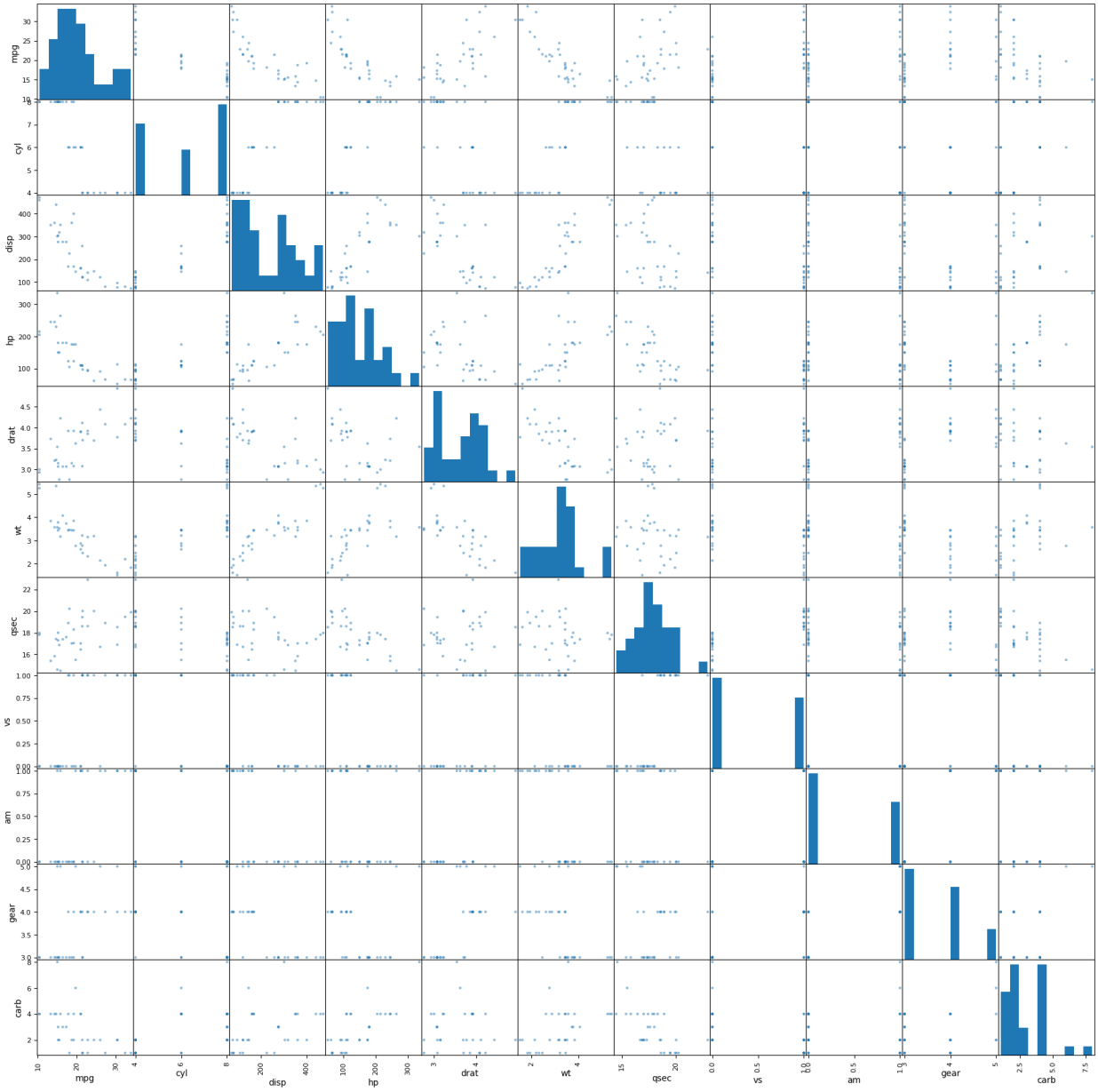
Out[ ]:
```
<matplotlib.collections.PathCollection at 0x7fd886c04940>
```



Q5 pairwise scatter plot for all variables

For self-pair, the plot is replaced by histagram, because data in scatter plot are on y=x which doesn't helpful

In [ ]:
```
mtx = pd.plotting.scatter_matrix(df,figsize=(22,22))
```



Q6

I partially agree with the engineer's suggestion that the relationship between wt and mpg is subject to the number of cylinders. By looking at the model we drew in Q4, as the numbers of cylinders and weight increase, the mpg decreases. However, by looking at the model in Q5, there is no strong and obvious relationship between the number of cylinders and the weight. Therefore, the relationship between wt and mpg is not solely subject to the number of cylinders.

Contribution: Everyone in this group contributes equally.