Mark Mersnik
ID: 010455185

EE257 Homework 1

1) a) A flexible method performs better when n is large, and p is small. Due to the large number of samples and low number of p will prevent the flexible method from overfitting the data.

   b) An inflexible method would be better when there a large number of predictors p and a small number of observations n. The small number of samples with a large number of predictors would cause a flexible method to overfit the data and underperform with new test data.

   c) A flexible method would fit a lot better to non-linear data due to its complex fitting while inflexible would fit better to linear.

   d) An inflexible method would fit better since it won't take into account the errors as much as a flexible method would.

2) a) The scenario is a regression problem with interest being in inference. This is due to the fact that we want to understand the factors that affect CEO salary rather than predict what the salary may be. The n is our 500 firms and the p is the 3 factors, profit, number of employees, and CEO salary.

   b) This scenario is a classification problem that we are trying to predict. This is due to the fact that the prediction we are trying to make is a binary result, either success or failure. The n is the 20 products and the p is the 13 factors, price charged for the product, marketing budget, competition price, and the 10 other factors.

   c) This scenario is a regression problem that we are trying to predict. This is due to the fact that we are trying to find a percentage change in the US dollar without needing to understand why. The n is for the 52 weeks that we recorded data in 2012 and the p is 3 for the % change in the US market, % change in the British market, and the % change in the German market.

3) Dataset information derived by hand:

3) a) revenue

$Q1 = 1240$

$Q3 = 1654$ — No Outliers

$IQR = 1654 - 1240 = 414$

$K = 1.5$ $\quad K * IQR = 621$

Range: $[1240 - 621 \quad 1654 + 621] = [619 \quad 2275]$

Computer science doctorates:

$Q1 = 861$

$Q3 = 1611$ — No Outliers

$IQR = 1611 - 861 = 750$

$1.5 * 750 = 1125$

Range: $[861 - 1125 \quad 1611 + 1125] = [-264 \quad 2736]$

b) Revenue:

Mean: $14415 / 10 = \boxed{1441.5}$

Median: $(1307 + 1435)/2 = \boxed{1371}$

Standard Dev:

$$\sigma_1 = \sqrt{\frac{\Sigma(x_i - 1441.5)^2}{10}} = \sqrt{\frac{506086.5}{10}} = \boxed{224.9637}$$

Computer Science doctorates:

Mean: $11951 / 10 = \boxed{1195.1}$

Median: $(948 + 1129)/2 = \boxed{1038.5}$

$$\sigma_2 = \sqrt{\frac{\Sigma(x_i - 1195.1)^2}{10}} = \sqrt{\frac{1369330.9}{10}} = \boxed{370.0447}$$

c) $$\rho = \frac{\Sigma(x_i - 1441.5)(y_i - 1195.1)}{\sqrt{\Sigma(x_i - 1441.5)^2}\sqrt{\Sigma(y_i - 1195.1)^2}}$$

$$\rho = \frac{820033.5}{(711.3976)(1170.1841)} = \boxed{0.98507}$$

d) Yes, according to the data given there is a high correlation between the revenue arcades bring in to the amount of computer science doctorates given.

4)

In [2]:
```
import pandas as pd
powerdata = pd.read_csv("servo.csv", names=["motor", "screw", "pgain", "vga
powerdata.head()
```

Out[2]:

	motor	screw	pgain	vgain	class
0	E	E	5	4	0.281251
1	B	D	6	5	0.506252
2	D	D	4	3	0.356251
3	B	A	3	2	5.500033
4	D	B	6	5	0.356251

In [3]:
```
powerdata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 5 columns):
motor    167 non-null object
screw    167 non-null object
pgain    167 non-null int64
vgain    167 non-null int64
class    167 non-null float64
dtypes: float64(1), int64(2), object(2)
memory usage: 6.6+ KB
```

In [4]:
```
powerdata["motor"]
```

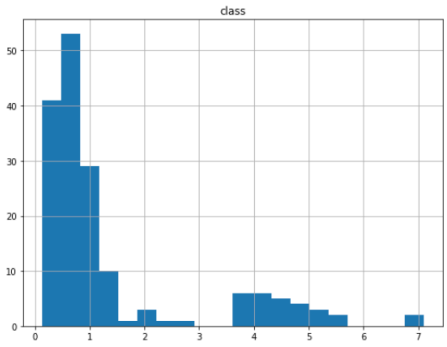
Out[4]:
```
0      E
1      B
2      D
3      B
4      D
      ..
162    B
163    B
164    C
165    A
166    A
Name: motor, Length: 167, dtype: object
```

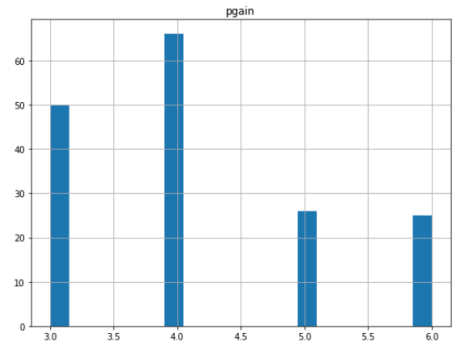
In [5]:
```
powerdata.describe()
```

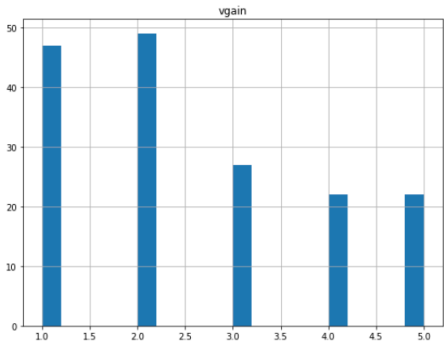
Out[5]:

	pgain	vgain	class
count	167.000000	167.000000	167.000000
mean	4.155689	2.538922	1.389708
std	1.017770	1.369850	1.559635
min	3.000000	1.000000	0.131250
25%	3.000000	1.000000	0.503126
50%	4.000000	2.000000	0.731254
75%	5.000000	4.000000	1.259369
max	6.000000	5.000000	7.100108

In [11]:
```
%matplotlib inline
import matplotlib.pyplot as plt
powerdata.hist(bins = 20, figsize = (20, 15))
plt.show()
```

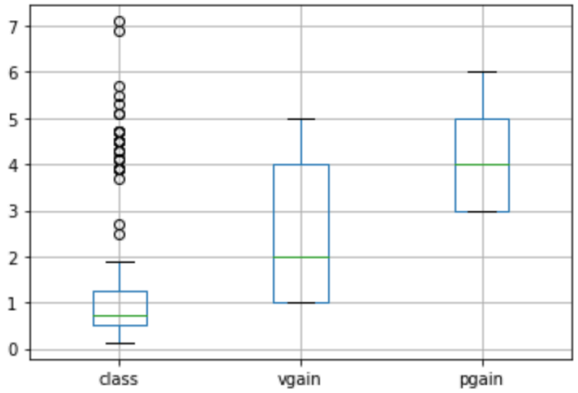






```
In [12]: powerdata.boxplot(column=['class', 'vgain', 'pgain'])
```

```
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x1183991d0>
```




```
In [14]: corr_matrix = powerdata.corr()
         corr_matrix ["class"].sort_values(ascending = False)
```

```
Out[14]: class    1.000000
         vgain   -0.364383
         pgain   -0.598129
         Name: class, dtype: float64
```

```
In [ ]:
```

5) The three videos that we were asked to watch covered the advancements of AI. Watson and Sophia were examples of our attempts at creating a robot that has general intelligence. This is an intelligence that is much like a human and will one day become undistinguishable. Eventually this will form into superintelligence that surpass that of any human that has ever lived. As for the last video, “Human Need Not Apply”, it covered the use of AI in automation. Both of these subjects bring up a lot of philosophical and ethical questions. When looking at the philosophical, we consider Watson and Sophia who will soon have an intelligence that appears to be exactly like that of a human. However, does this mean that they are conscious and feel true emotions? This is a very important question that we need to consider and closely follow. As for the ethical questions, does having a robot replace humans in such a wide variety of fields become viable in the long run? If robots can do everything humans can do, better and for less money, what are we left to do? Will we be happy without purpose or responsibility? All of these questions are important to consider. The drastic impact of AI is inevitable. Whether it will be beneficial or not, in the long run, is unknown so I think we should keep a close eye on it, while it progresses. Lastly, if it were up to me, I believe that AI should be used to automate most jobs, if it does advance us a lot quicker, but that creative jobs should be left to humans, as our way to express ourselves in a healthy manner.