MSCA 31000 6 – ASSIGNMENT 1
CHAPTER 1

1. **A teacher wishes to know whether the males in his/her class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared. Is this an example of descriptive or inferential statistics?**
   This is an example of *descriptive statistics.* It is not inferential as a sample is not taken from the population (the teacher's class) to understand the population. Also, the data is about the attitudes of males and females in the class which is collected by a questionnaire. The outcome would show which gender is more conservative which is a summary of the data collected thereby making it an example of descriptive statistics.

3. **If you are told that you scored in the 80th percentile, from just this information would you know exactly what that means and how it was calculated? Explain.**
   Consider the following table for this answer:

Score	Rank
10	1
13	2
14	3
18	4
20	5

   The above statement could mean any of the following, but the third point is the precise information to be inferred from this:
   - My score is the lowest score that is *greater than* 80% of the scores considered, which in the case of the above table would be the score 20.
   - My score is the lowest score that is *greater than or equal to* 80% of the scores considered, which in the case of the above table would be the score 18.
   - The *precise information inferred* would be using the below calculation:
   - Compute the Rank of the 80th percentile using the formula

$$R = \left(\frac{P}{100}\right)(N+1)$$

   Where P is the desired percentile (80th percentile) and N is the number of cases).
   So in the above case, $R = \left(\frac{80}{100}\right)(5+1) = 4.80$
   - If R was an integer, then the score at rank R would be the 80th percentile. However, since it's an integer, the below further calculations are required:
     a. Let IR be the integer part of the rank R and FR the fractional part of the rank R. The rank is then computed using the formula:

$$R = (FR)(Difference\ between\ scores\ at\ rank\ IR\ and\ IR+1) + lowest\ score\ amongst\ scores\ at\ rank\ IR\ and\ IR+1$$

     b. In the above example, the calculations are as follows:

$$IR = 4, FR = 0.80, Score\ at\ rank\ IR\ (4) = 18, Score\ at\ rank\ (IR+1)\ (4+1) = 20$$
$$R = (0.80)(20-18) + 18 = 19.60$$

   - Thus, the score at the 80th percentile is 19.60.

5. **Give an example of an independent and a dependent variable.**

Consider the experiment – can the new drug invented help Alzheimer's patients improve their memory? In this experiment, the Alzheimer's patients were randomly divided into groups to receive the new drug or a placebo (control variable). After 12 weeks, they were given memory tests. The results showed that those who were given the drug performed better at the memory test.
In this scenario,

- the independent variable is the type of treatment: new drug for Alzheimer's, placebo
- the dependent variable is the memory test: has the memory improved or not on administration of the drug.

7. **Specify the level of measurement used for the items in Question 6.**
    - Rating of the quality of a movie on a 7-point scale - *Ordinal*
    - Age - *Ratio*
    - Country you were born in - *Nominal*
    - Favorite Color - *Nominal*
    - Time to respond to a question - *Ratio*

9. **The formula for finding each student's test grade (g) from his or her raw score (s) on a test is as follows: g = 16 + 3s**
Is this a linear transformation? If a student got a raw score of 20, what is his test grade?
This is a linear transformation as the score is multiplied by a number and added to another.
If the raw score is 20, the student's test grade is:
$$g = 16 + 3(20)$$
$$g = 16 + 60$$
$$g = 76$$
The student's test grade = 76.

11. **Which of the frequency polygons has a large positive skew? Which has a large negative skew?**
Frequency polygon A has a large *positive skew*. *Frequency polygon B* has a large *negative skew*.

CHAPTER 2

1. **Name some ways to graph quantitative variables and some ways to graph qualitative variables.**
Some ways to graph Quantitative variables are:
    - Stem and Leaf Display
    - Histograms
    - Frequency Polygons
    - Box plots
    - Bar charts
    - Line graphs
    - Scatter plots
    - Dot plots

Some ways to graph Qualitative variables are:

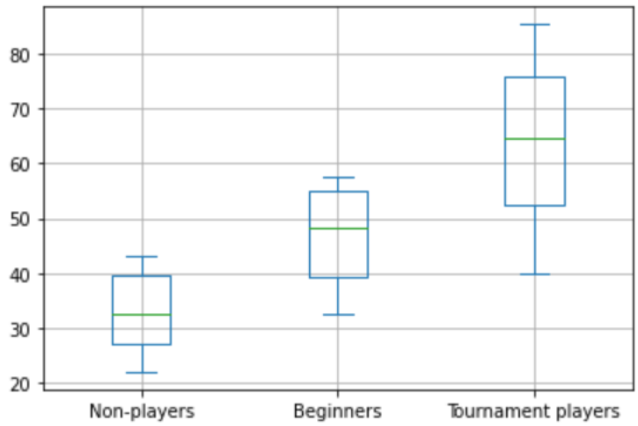
- Pie charts
- Bar charts

3. **An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the total number of pieces correctly remembered from three chess positions. Create side-by-side box plots for these three groups. What can you say about the differences between these groups from the box plots?**

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

In [2]:
```
import pandas as pd
import numpy as np
df = pd.read_csv('C:/Users/Administrator/Downloads/Stats Class/Book1.csv')
df.plot.box(grid='True')
```

Out[2]: <AxesSubplot:>



From the above graph, the following inferences can be made:

- The non-players were able to remember a maximum of 43.2 positions, beginners were able to remember a maximum of 57.7 and tournament players were able to remember a maximum of 85.3 positions. Thus, it can be concluded that tournament players are able to remember chess position much better than the other two categories most probably as they are seasoned players.
- Tournament players have a larger positive skew than the other two categories.

- All categories have a larger negative skew than their positive skew showing that the number of people between the 25th & 50th percentile were able to remember more moves than the rest of the population.
- The median of tournament players is the highest at 64.6 positions which is almost double that of the median for non-players which is at 32.6 positions.

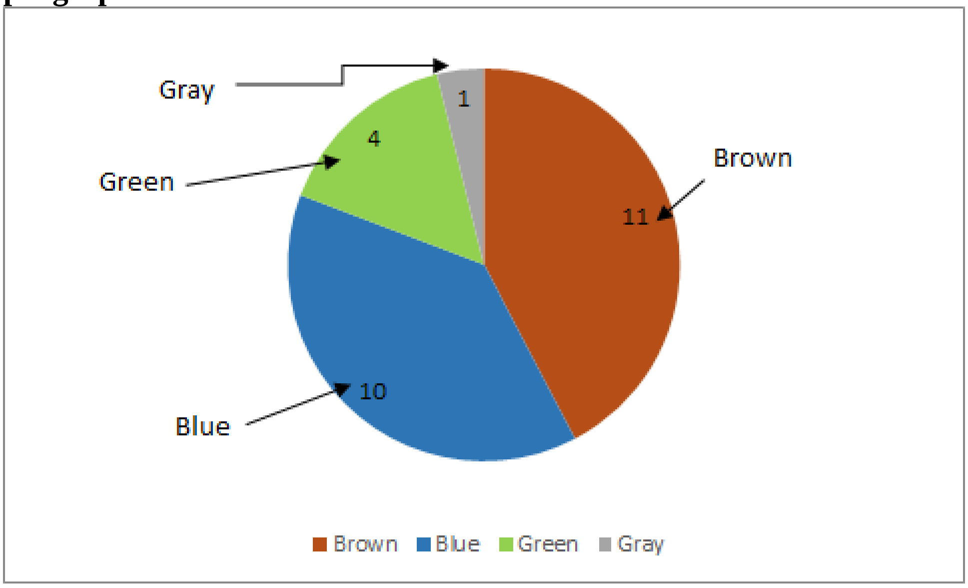
5. **In a box plot, what percent of the scores are between the lower and upper hinges?**
   In a box plot, the box stretches between the lower hinge (25th percentile) and higher hinge (75th percentile) thereby representing 50% of the scores.

7. **For the data from the 1977 Stat. and Biom. 200 class for eye color, construct:**

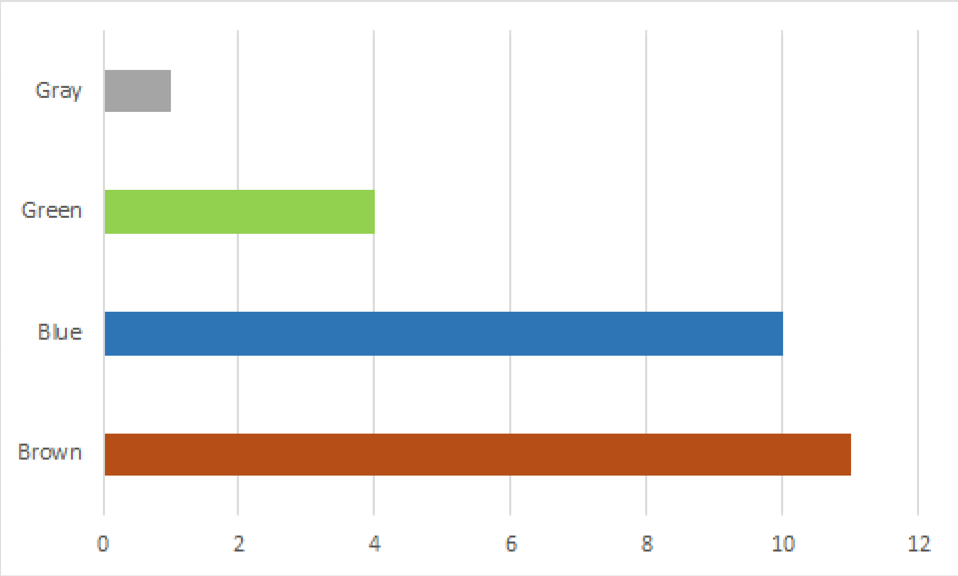
Eye Color	Number of students
Brown	11
Blue	10
Green	4
Gray	1

All graphs created using Excel

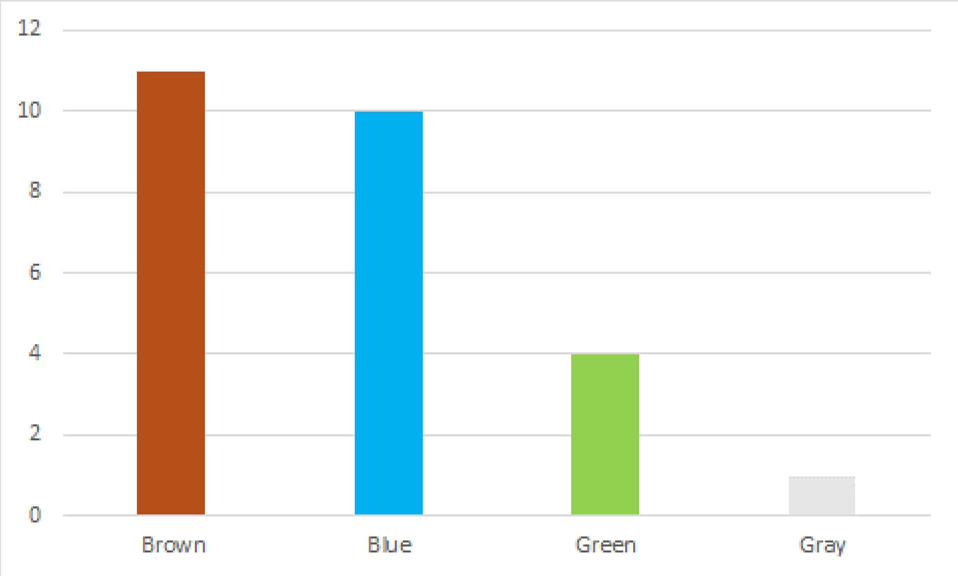
a. pie graph



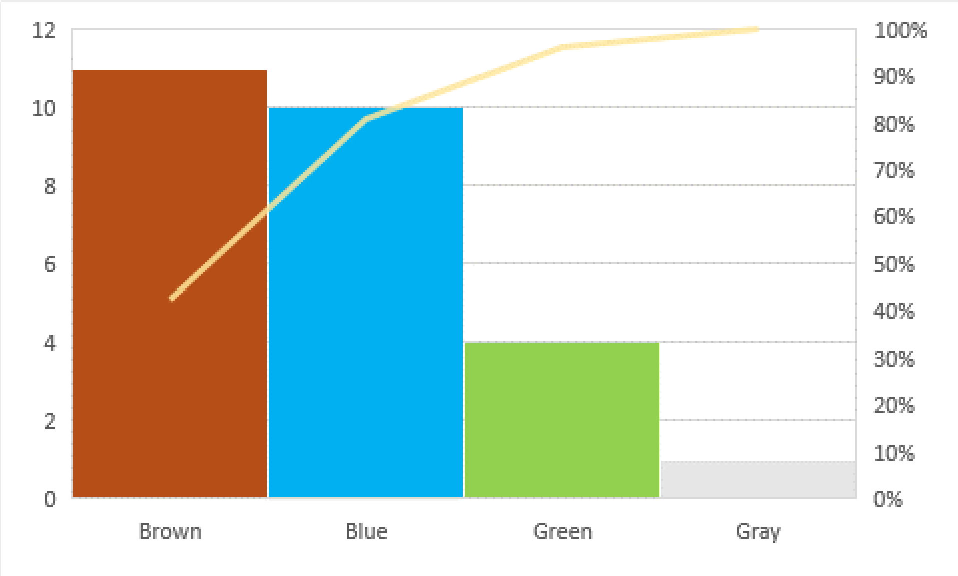
b. horizontal bar graph



c. **vertical bar graph**



d. **a frequency table with the relative frequency of each eye color**



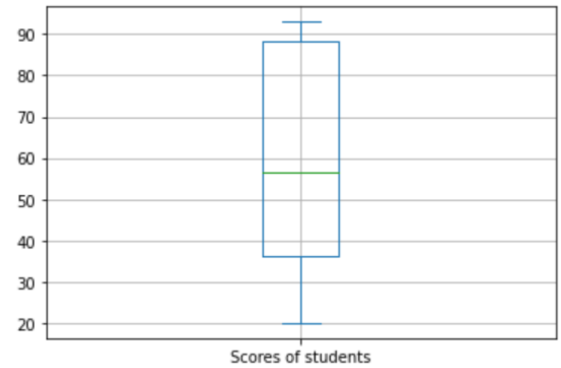
9. **Which of the box plots below has a large positive skew? Which has a large negative skew?**
Plot B has a large positive skew. Plot C has a large negative skew.

CHAPTER 3

1. **Make up a dataset of 12 numbers with a positive skew. Use a statistical program to compute the skew. Is the mean larger than the median as it usually is for distributions with a positive skew? What is the value for skew?**
Consider the scores of 12 students on a test as per the below data in the numpy array:

```
df = pd.DataFrame(np.array([20,25,25,40,50,55,58,80,88,89,93,92]), columns=['Scores of students'])
df.plot.box(grid='True')
print("Mean of data = " + str(df['Scores of students'].mean()))
print("Median of data = " + str(df['Scores of students'].median()))
if df['Scores of students'].mean() > df['Scores of students'].median():
    print("Mean is larger than median as seen for positive skews")
else:
    print("Mean is not larger than median as seen for positive skews")
print("Value of Skew = " + str(df['Scores of students'].skew()))
```

```
Mean of data = 59.583333333333336
Median of data = 56.5
Mean is larger than median as seen for positive skews
Value of Skew = -0.12201759752633712
```



3. **Make up three data sets with 5 numbers each that have:**
(a) **the same mean but different standard deviations.**

						Mean	Standard Deviation
Set A	5	5	5	5	5	5	0
Set B	3	4	5	6	7	5	1.414213562
Set C	4	4	5	6	6	5	0.894427191

(b) the same mean but different medians.

						Mean	Median
Set A	5	6	7	8	9	7	7
Set B	5	5	8	8	9	7	8
Set C	5	6	6	9	9	7	6

(c) the same median but different means.

						Mean	Median
Set A	10	20	30	40	50	30	30

Set B	10	20	30	40	40	28	30
Set C	20	20	30	50	50	34	30

5. **A sample of 30 distance scores measured in yards has a mean of 7, a variance of 16, and a standard deviation of 4. (a) You want to convert all your distances from yards to feet, so you multiply each score in the sample by 3. What are the new mean, variance, and standard deviation? (b) You then decide that you only want to look at the distance past a certain point. Thus, after multiplying the original scores by 3, you decide to subtract 4 feet from each of the scores. Now what are the new mean, variance, and standard deviation?**
   a) The new mean = 7 * 3 = 21, variance = 16 * 3*3 = 144, standard deviation = 4 *3 =12
   b) The new mean = 21 – 4 = 17, variance = 144, standard deviation = 12

7. **For the test scores in question #6, which measures of variability (range, standard deviation, variance) would be changed if the 22.1 data point had been erroneously recorded as 21.2?**
   The range would remain the same however the standard deviation and variance would change.

9. **For the numbers 1, 3, 4, 6, and 12:**
   a) Find the value (v) for which Σ(X-v)2 is minimized. = *70.8*

	Value	Squared Deviation from Median	Squared Deviation from Mean
	1	= power(abs(1-4),2) = 9	= power(abs(1-5.2),2) = 17.64
	3	= power(abs(3-4),2) = 1	= power(abs(3-5.2),2) = 4.84
	4	= power(abs(4-4),2) = 0	= power(abs(4-5.2),2) = 1.44
	6	= power(abs(6-4),2) = 4	= power(abs(6-5.2),2) = 0.64
	12	= power(abs(12-4),2) = 64	= power(abs(12-5.2),2) = 46.24
Total	26	78	70.8
Mean	5.2		
Median	4		

   b) Find the value (v) for which Σ|x-v| is minimized. = *14*

	Value	Absolute Deviation from Median	Absolute Deviation from Mean
	1	=abs(1-4) = 3	=abs(1-5.2) = 4.2
	3	=abs(3-4) = 1	=abs(3-5.2) = 2.2
	4	=abs(4-4) = 0	=abs(4-5.2) = 1.2
	6	=abs(6-4) = 2	=abs(6-5.2) = 0.8
	12	=abs(12-4) = 8	=abs(12-5.2) = 6.8
Total	26	14	15.2
Mean	5.2		
Median	4		

11. **An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess positions.**

Compare the performance of each group. Consider spread as well as central tendency.

Rank	Non-players	Beginners	Tournament players
1	22.1	32.5	40.1
2	22.3	37.1	45.6
3	26.2	39.1	51.2
4	29.6	40.5	56.4
5	31.7	45.5	58.1
6	33.5	51.3	71.1
7	38.9	52.6	74.9
8	39.7	55.7	75.9
9	43.2	55.9	80.3
10	43.2	57.7	85.3
Mean	33.04	46.79	63.89
Median	32.6	48.4	64.6
Mode	43.2	No Mode	No Mode
Trimean	32.75	53.8375	68.74375
Geometric mean	32.12256	45.96684	62.07393
Trimmed Mean	33.1375	47.2125	64.1875
Range	21.1	25.2	45.2
IQR	40.575	17.15	27.2
Variance	58.0804	73.3969	219.6269
Standard Deviation	7.62105	8.567199	14.81981

13. **True/False: The best way to describe a skewed distribution is to report the mean.**
– False, as a mean is not enough to describe a skewed distribution since it would need to be compared to the median of the data.

15. **Compare the mean, median, trimean in terms of their sensitivity to extreme scores.**
For *symmetric distributions*, the mean, median, trimean are equal except in bimodal distributions. However, when distributions are skewed, differences occur between the mean, median and trimean. When distributions have a *positive skew*, the mean is typically higher than the median, although it may not be in bimodal distributions. Typically, the trimean will fall between the median and the mean.

17. **A set of numbers is transformed by taking the log base 10 of each number. The mean of the transformed data is 1.65. What is the geometric mean of the untransformed data?**
The geometric mean of the untransformed data is the antilog of 1.65 = 44.6683592151

19. **The histogram is in balance on the fulcrum. What are the mean, median, and mode of the distribution (approximate where necessary)?**
Mean = 4.4

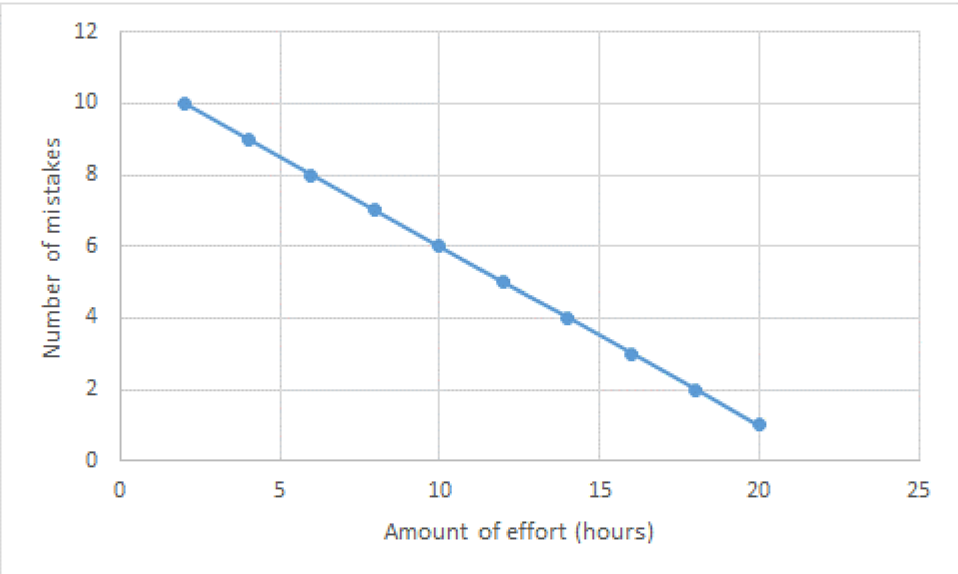
Median = ~3.3
Mode = 1

CHAPTER 4

1. **Describe the relationship between variables A and C. Think of things these variables could represent in real life.**
The variables A and C are having a linear relationship and have a negative association as the value of C decreases as A increases. An example in real life is variable A is amount of time spent on a project and variable B is number of mistakes made in a project. So as the time spent on the project increases, the number of mistakes decreases.

3. **Make up a data set with 10 numbers that has a negative correlation.**

Amount of effort (hours)	Number of mistakes
20	1
18	2
16	3
14	4
12	5
10	6
8	7
6	8
4	9
2	10



5. **Would you expect the correlation between High School GPA and College GPA to be higher when taken from your entire high school class or when taken from only the top 20 students? Why?**

The correlation would be higher when taken only from the top 20 students, as the sample is biased towards well performing students since those students who did well in High School would most likely also do well in College as toppers.

7. **For this same class, the relationship between the amount of time spent studying and the amount of time spent socializing per week was also examined. It was determined that the more hours they spent studying, the fewer hours they spent socializing. Is this a positive or negative association?** It is a *Negative association*.

9. **Students took two parts of a test, each worth 50 points. Part A has a variance of 25, and Part B has a variance of 36. The correlation between the test scores is 0.8. (a) If the teacher adds the grades of the two parts together to form a final test grade, what would the variance of the final test grades be? (b) What would the variance of Part A - Part B be?**
   The formula for these is:
   $$\boldsymbol{\sigma^2 = \sigma^2{}_A + \sigma^2{}_B + 2\rho\sigma^2{}_A\sigma^2{}_B}$$
   A) Variance of final test grades = 36 + 25 + (2)(36)(25) = 1861
   B) Variance of Part A – Part B is same as answer A.

11. **True/False: It is possible for variables to have r=0 but still have a strong association.** *False, it means they have no association between them.*

13. **True/False: After polling a certain group of people, researchers found a 0.5 correlation between the number of car accidents per year and the driver's age. This means that older people get in more accidents.** *False, the correlation between the two is no strong enough to come to this conclusion.*

15. **True/False: To examine bivariate data graphically, the best choice is two side by side histograms.** *False, a scatter plot is one of the best choices.*

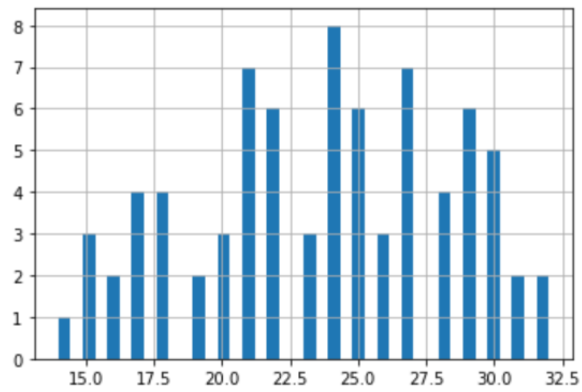
ANGRY MOODS CASE STUDY QUESTIONS

10. **Plot a histogram of the distribution of the Control-Out scores.**

In [23]:

```python
import pandas as pd
import numpy as np
df = pd.read_excel('C:/Users/Administrator/Downloads/Stats Class/angry_moods.xls')
df['Control-Out'].hist(bins = 40)
```

```
Out[23]: <AxesSubplot:>
```



11. What is the overall mean Control-Out score? What is the mean Control-Out score for the athletes? What is the mean Control-Out score for the non-athletes?

$$Overall\ Mean\ of\ Control\ out\ scores = \frac{sum\ of\ all\ scores}{Number\ of\ scores}$$

$$Overall\ Mean\ of\ Control\ out\ scores = \frac{1848}{78}$$

$$Thus, \quad \boldsymbol{Overall\ Mean\ of\ Control\ out\ scores = \ 23.69230769}$$

$$Mean\ of\ ControlOut\ score\ for\ athletes = \frac{sum\ of\ all\ controlout\ score\ for\ athletes}{Number\ of\ athletes}$$

$$Mean\ of\ Control\ out\ score\ for\ athletes = \frac{617}{25}$$

$$Thus, \quad \boldsymbol{Mean\ of\ Control\ out\ score\ for\ athletes = \ 24.68}$$

$$Mean\ of\ ControlOut\ score\ for\ non\ athletes = \frac{sum\ of\ all\ controlout\ score\ for\ non\ athletes}{Number\ of\ non\ athletes}$$

$$Mean\ of\ Control\ out\ score\ for\ non\ athletes = \frac{1231}{35}$$

$$Thus, \quad \boldsymbol{Mean\ of\ Control\ out\ score\ for\ non\ athletes = \ 23.22641509}$$

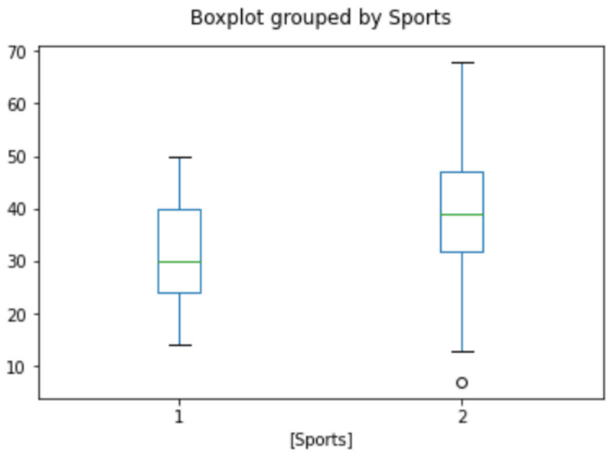
Overall mean for Control-Out score = 23.69230769
Mean of Control-Out score for athletes = 24.68
Mean of Control-Out score for non-athletes = 23.22641509

17. Plot parallel box plots of the Anger Expression Index by sports participation. Does it look like there are any outliers? Which group reported expressing more anger?

```
In [40]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         df = pd.read_excel('C:/Users/Administrator/Downloads/Stats Class/angry_moods.xls')

         df_to_plot = df[['Sports', 'Anger_Expression']]
         axes = df_to_plot.boxplot(by='Sports', grid = False)
         axes.set_title('')

Out[40]: Text(0.5, 1.0, '')
```



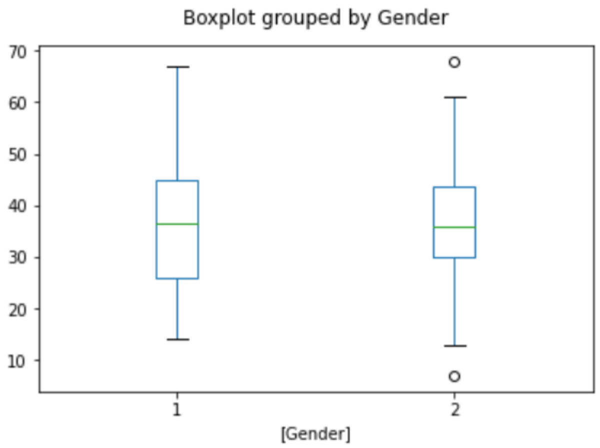
From the above box plot, it does look like Non-athletes have an outlier of 7. The non-athletes were reported expressing more anger.

18. Plot parallel box plots of the Anger Expression Index by gender.

```
In [41]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         df = pd.read_excel('C:/Users/Administrator/Downloads/Stats Class/angry_moods.xls')

         df_to_plot = df[['Gender', 'Anger_Expression']]
         axes = df_to_plot.boxplot(by='Gender', grid = False)
         axes.set_title('')

Out[41]: Text(0.5, 1.0, '')
```



20. What is the correlation between the Control-In and Control-Out scores? Is this correlation statistically significant at the 0.01 level?

The correlation = 0.719283

This correlation is not statistically significant at the 0.01 level as it's p-value is not less than level 0.01.

```
In [43]:  import pandas as pd
          import numpy as np
          from scipy.stats.stats import pearsonr
          df = pd.read_excel('C:/Users/Administrator/Downloads/Stats Class/angry_moods.xls')

          df[['Control-In', 'Control-Out']].corr(method = 'pearson')
          pearsonr(df['Control-In'], df['Control-Out'])

Out[43]:  (0.7192834133867423, 1.1899926358135796e-13)
```

Control-Out (X)	Control-In (Y)	x (deviation from mean of X)	y (deviation from mean	xy	x^2	y^2
23	20	-0.692307692	-1.961538	1.3579882	0.4792899	3.8476331
25	24	1.307692308	2.0384615	2.6656805	1.7100592	4.1553254
28	28	4.307692308	6.0384615	26.011834	18.556213	36.463018
23	23	-0.692307692	1.0384615	-0.718935	0.4792899	1.0784024
26	28	2.307692308	6.0384615	13.934911	5.3254438	36.463018
25	23	1.307692308	1.0384615	1.3579882	1.7100592	1.0784024
31	27	7.307692308	5.0384615	36.819527	53.402367	25.386095
22	31	-1.692307692	9.0384615	-15.29586	2.8639053	81.693787
22	24	-1.692307692	2.0384615	-3.449704	2.8639053	4.1553254
29	29	5.307692308	7.0384615	37.357988	28.171598	49.539941
24	25	0.307692308	3.0384615	0.9349112	0.0946746	9.2322485
24	22	0.307692308	0.0384615	0.0118343	0.0946746	0.0014793
29	30	5.307692308	8.0384615	42.66568	28.171598	64.616864
23	27	-0.692307692	5.0384615	-3.488166	0.4792899	25.386095
17	11	-6.692307692	-10.96154	73.357988	44.786982	120.15533
28	24	4.307692308	2.0384615	8.7810651	18.556213	4.1553254
27	24	3.307692308	2.0384615	6.7426036	10.940828	4.1553254
15	18	-8.692307692	-3.961538	34.434911	75.556213	15.693787
26	26	2.307692308	4.0384615	9.3195266	5.3254438	16.309172
15	14	-8.692307692	-7.961538	69.204142	75.556213	63.386095
29	15	5.307692308	-6.961538	-36.9497	28.171598	48.463018
29	23	5.307692308	1.0384615	5.5118343	28.171598	1.0784024
14	13	-9.692307692	-8.961538	86.857988	93.940828	80.309172
16	18	-7.692307692	-3.961538	30.473373	59.171598	15.693787

	17	14	-6.692307692	-7.961538	53.281065	44.786982	63.386095
	21	21	-2.692307692	-0.961538	2.5887574	7.2485207	0.9245562
	18	13	-5.692307692	-8.961538	51.011834	32.402367	80.309172
	26	24	2.307692308	2.0384615	4.704142	5.3254438	4.1553254
	17	14	-6.692307692	-7.961538	53.281065	44.786982	63.386095
	22	22	-1.692307692	0.0384615	-0.065089	2.8639053	0.0014793
	28	23	4.307692308	1.0384615	4.4733728	18.556213	1.0784024
	27	21	3.307692308	-0.961538	-3.180473	10.940828	0.9245562
	21	19	-2.692307692	-2.961538	7.9733728	7.2485207	8.7707101
	25	24	1.307692308	2.0384615	2.6656805	1.7100592	4.1553254
	21	16	-2.692307692	-5.961538	16.050296	7.2485207	35.539941
	16	18	-7.692307692	-3.961538	30.473373	59.171598	15.693787
	24	25	0.307692308	3.0384615	0.9349112	0.0946746	9.2322485
	30	26	6.307692308	4.0384615	25.473373	39.786982	16.309172
	22	16	-1.692307692	-5.961538	10.088757	2.8639053	35.539941
	21	22	-2.692307692	0.0384615	-0.10355	7.2485207	0.0014793
	22	19	-1.692307692	-2.961538	5.0118343	2.8639053	8.7707101
	27	24	3.307692308	2.0384615	6.7426036	10.940828	4.1553254
	30	30	6.307692308	8.0384615	50.704142	39.786982	64.616864
	20	17	-3.692307692	-4.961538	18.319527	13.633136	24.616864
	27	20	3.307692308	-1.961538	-6.488166	10.940828	3.8476331
	32	22	8.307692308	0.0384615	0.3195266	69.017751	0.0014793
	24	22	0.307692308	0.0384615	0.0118343	0.0946746	0.0014793
	19	19	-4.692307692	-2.961538	13.89645	22.017751	8.7707101
	25	19	1.307692308	-2.961538	-3.872781	1.7100592	8.7707101
	17	18	-6.692307692	-3.961538	26.511834	44.786982	15.693787
	22	23	-1.692307692	1.0384615	-1.757396	2.8639053	1.0784024
	15	16	-8.692307692	-5.961538	51.819527	75.556213	35.539941
	29	19	5.307692308	-2.961538	-15.71893	28.171598	8.7707101
	28	20	4.307692308	-1.961538	-8.449704	18.556213	3.8476331
	24	22	0.307692308	0.0384615	0.0118343	0.0946746	0.0014793
	32	32	8.307692308	10.038462	83.39645	69.017751	100.77071
	24	16	0.307692308	-5.961538	-1.83432	0.0946746	35.539941
	21	24	-2.692307692	2.0384615	-5.488166	7.2485207	4.1553254
	21	21	-2.692307692	-0.961538	2.5887574	7.2485207	0.9245562
	27	22	3.307692308	0.0384615	0.1272189	10.940828	0.0014793
	19	19	-4.692307692	-2.961538	13.89645	22.017751	8.7707101
	25	26	1.307692308	4.0384615	5.2810651	1.7100592	16.309172
	25	24	1.307692308	2.0384615	2.6656805	1.7100592	4.1553254
	20	19	-3.692307692	-2.961538	10.934911	13.633136	8.7707101
	31	32	7.307692308	10.038462	73.357988	53.402367	100.77071
	30	29	6.307692308	7.0384615	44.39645	39.786982	49.539941
	27	23	3.307692308	1.0384615	3.4349112	10.940828	1.0784024
	30	29	6.307692308	7.0384615	44.39645	39.786982	49.539941
	27	27	3.307692308	5.0384615	16.66568	10.940828	25.386095
	21	17	-2.692307692	-4.961538	13.357988	7.2485207	24.616864
	24	23	0.307692308	1.0384615	0.3195266	0.0946746	1.0784024
	18	17	-5.692307692	-4.961538	28.242604	32.402367	24.616864
	18	23	-5.692307692	1.0384615	-5.911243	32.402367	1.0784024
	18	15	-5.692307692	-6.961538	39.627219	32.402367	48.463018
	30	32	6.307692308	10.038462	63.319527	39.786982	100.77071
	29	24	5.307692308	2.0384615	10.819527	28.171598	4.1553254
	20	18	-3.692307692	-3.961538	14.627219	13.633136	15.693787
	24	26	0.307692308	4.0384615	1.2426036	0.0946746	16.309172
Total	1848	1713	-8.52651E-14	1.279E-13	1284.0769	1692.6154	1882.8846
Mean	23.69230769	21.96153846					

$$r = (\sum xy)/(\sqrt{\sum x^2 \sum y^2})$$

$\sum x^2 \sum y^2 =$	3186999.467
$\sqrt{\sum x^2 \sum y^2} =$	1785.216924
r =	0.719283413

21. Would you expect the correlation between the Anger-Out and Control-Out scores to be positive or negative? Compute this correlation.

I would expect the correlation to be negative as a person who would let anger out generally wouldn't have enough control to set it out. Computed as below the correlation = -0.582683

In [45]:

```python
import pandas as pd
import numpy as np
from scipy.stats.stats import pearsonr
df = pd.read_excel('C:/Users/Administrator/Downloads/Stats Class/angry_moods.xls')

df[['Anger-Out', 'Control-Out']].corr(method = 'pearson')
```

Out[45]:

	Anger-Out	Control-Out
Anger-Out	1.000000	-0.582683
Control-Out	-0.582683	1.000000

	Anger-Out (X)	Control-Out (Y)	x (deviation from mean of X)	y (deviation from mean	xy	x^2	y^2
	18	23	1.923076923	-0.692308	-1.331361	3.6982249	0.4792899
	14	25	-2.076923077	1.3076923	-2.715976	4.3136095	1.7100592
	13	28	-3.076923077	4.3076923	-13.25444	9.4674556	18.556213
	17	23	0.923076923	-0.692308	-0.639053	0.852071	0.4792899
	16	26	-0.076923077	2.3076923	-0.177515	0.0059172	5.3254438
	16	25	-0.076923077	1.3076923	-0.100592	0.0059172	1.7100592
	12	31	-4.076923077	7.3076923	-29.7929	16.621302	53.402367
	13	22	-3.076923077	-1.692308	5.2071006	9.4674556	2.8639053
	16	22	-0.076923077	-1.692308	0.1301775	0.0059172	2.8639053
	12	29	-4.076923077	5.3076923	-21.63905	16.621302	28.171598
	12	24	-4.076923077	0.3076923	-1.254438	16.621302	0.0946746
	17	24	0.923076923	0.3076923	0.2840237	0.852071	0.0946746
	18	29	1.923076923	5.3076923	10.207101	3.6982249	28.171598
	27	23	10.92307692	-0.692308	-7.56213	119.31361	0.4792899
	18	17	1.923076923	-6.692308	-12.86982	3.6982249	44.786982
	9	28	-7.076923077	4.3076923	-30.48521	50.08284	18.556213
	13	27	-3.076923077	3.3076923	-10.17751	9.4674556	10.940828
	20	15	3.923076923	-8.692308	-34.10059	15.390533	75.556213
	16	26	-0.076923077	2.3076923	-0.177515	0.0059172	5.3254438
	23	15	6.923076923	-8.692308	-60.17751	47.928994	75.556213
	12	29	-4.076923077	5.3076923	-21.63905	16.621302	28.171598
	15	29	-1.076923077	5.3076923	-5.715976	1.1597633	28.171598
	26	14	9.923076923	-9.692308	-96.17751	98.467456	93.940828
	17	16	0.923076923	-7.692308	-7.100592	0.852071	59.171598

	20	17	3.923076923	-6.692308	-26.25444	15.390533	44.786982
	9	21	-7.076923077	-2.692308	19.053254	50.08284	7.2485207
	24	18	7.923076923	-5.692308	-45.10059	62.775148	32.402367
	12	26	-4.076923077	2.3076923	-9.408284	16.621302	5.3254438
	23	17	6.923076923	-6.692308	-46.33136	47.928994	44.786982
	14	22	-2.076923077	-1.692308	3.5147929	4.3136095	2.8639053
	23	28	6.923076923	4.3076923	29.822485	47.928994	18.556213
	13	27	-3.076923077	3.3076923	-10.17751	9.4674556	10.940828
	18	21	1.923076923	-2.692308	-5.177515	3.6982249	7.2485207
	11	25	-5.076923077	1.3076923	-6.639053	25.775148	1.7100592
	15	21	-1.076923077	-2.692308	2.8994083	1.1597633	7.2485207
	20	16	3.923076923	-7.692308	-30.17751	15.390533	59.171598
	11	24	-5.076923077	0.3076923	-1.56213	25.775148	0.0946746
	16	30	-0.076923077	6.3076923	-0.485207	0.0059172	39.786982
	11	22	-5.076923077	-1.692308	8.591716	25.775148	2.8639053
	15	21	-1.076923077	-2.692308	2.8994083	1.1597633	7.2485207
	17	22	0.923076923	-1.692308	-1.56213	0.852071	2.8639053
	18	27	1.923076923	3.3076923	6.3609467	3.6982249	10.940828
	16	30	-0.076923077	6.3076923	-0.485207	0.0059172	39.786982
	14	20	-2.076923077	-3.692308	7.6686391	4.3136095	13.633136
	18	27	1.923076923	3.3076923	6.3609467	3.6982249	10.940828
	11	32	-5.076923077	8.3076923	-42.17751	25.775148	69.017751
	12	24	-4.076923077	0.3076923	-1.254438	16.621302	0.0946746
	18	19	1.923076923	-4.692308	-9.023669	3.6982249	22.017751
	21	25	4.923076923	1.3076923	6.4378698	24.236686	1.7100592
	22	17	5.923076923	-6.692308	-39.63905	35.08284	44.786982
	22	22	5.923076923	-1.692308	-10.02367	35.08284	2.8639053
	26	15	9.923076923	-8.692308	-86.25444	98.467456	75.556213
	11	29	-5.076923077	5.3076923	-26.94675	25.775148	28.171598
	17	28	0.923076923	4.3076923	3.9763314	0.852071	18.556213
	16	24	-0.076923077	0.3076923	-0.023669	0.0059172	0.0946746
	9	32	-7.076923077	8.3076923	-58.7929	50.08284	69.017751
	11	24	-5.076923077	0.3076923	-1.56213	25.775148	0.0946746
	17	21	0.923076923	-2.692308	-2.485207	0.852071	7.2485207
	13	21	-3.076923077	-2.692308	8.2840237	9.4674556	7.2485207
	18	27	1.923076923	3.3076923	6.3609467	3.6982249	10.940828
	19	19	2.923076923	-4.692308	-13.71598	8.5443787	22.017751
	15	25	-1.076923077	1.3076923	-1.408284	1.1597633	1.7100592
	15	25	-1.076923077	1.3076923	-1.408284	1.1597633	1.7100592
	18	20	1.923076923	-3.692308	-7.100592	3.6982249	13.633136
	12	31	-4.076923077	7.3076923	-29.7929	16.621302	53.402367
	15	30	-1.076923077	6.3076923	-6.792899	1.1597633	39.786982
	14	27	-2.076923077	3.3076923	-6.869822	4.3136095	10.940828
	14	30	-2.076923077	6.3076923	-13.10059	4.3136095	39.786982
	15	27	-1.076923077	3.3076923	-3.56213	1.1597633	10.940828
	21	21	4.923076923	-2.692308	-13.25444	24.236686	7.2485207
	18	24	1.923076923	0.3076923	0.591716	3.6982249	0.0946746
	17	18	0.923076923	-5.692308	-5.254438	0.852071	32.402367
	24	18	7.923076923	-5.692308	-45.10059	62.775148	32.402367
	14	18	-2.076923077	-5.692308	11.822485	4.3136095	32.402367
	10	30	-6.076923077	6.3076923	-38.33136	36.928994	39.786982
	11	29	-5.076923077	5.3076923	-26.94675	25.775148	28.171598
	15	20	-1.076923077	-3.692308	3.9763314	1.1597633	13.633136
	15	24	-1.076923077	0.3076923	-0.331361	1.1597633	0.0946746
Total	1254	1848	2.13163E-14	-8.53E-14	-887.1538	1369.5385	1692.6154
Mean	16.07692308	23.69230769					

$$r = (\sum xy)/(\sqrt{\sum x^2 \sum y^2})$$

$\sum x^2 \sum y^2 =$	2318101.87
$\sqrt{\sum x^2 \sum y^2} =$	1522.531402
r =	-0.582683447