MITHESH R, J002

Loading Packages

In [27]:

```
import pandas as pd
import numpy as np
from statistics import mean
import matplotlib.pyplot as plt
```

Statistics Package Exercise 1: Exploring Variables in a Dataset View dataset

In [28]:

```
dep = pd.read_csv('/Users/home/Downloads/OneDrive_1_07-07-2020/depression.csv'
)
dep.head()
```

Out[28]:

```
   Unnamed: 0  Hospt  Treat  Outcome        Time  AcuteT  Age  Gender
0           1      1      0        1   36.143002     211   33       1
1           2      1      1        0  105.142998     176   49       1
2           3      1      1        0   74.570999     191   50       1
3           4      1      0        1   49.714001     206   29       2
4           5      1      0        0   14.429000      63   29       1
```

Question 1: What are the categorical variables in this dataset? From the above image, we can see that ‘Hospt’ , ‘Treat’ , ‘Outcome’ and ‘Gender’ are categorical variables, i.e have outcomes in levels.

1. Hostp because the numbers represent codes, which are used to identify individual hospitals and place them into categories.
2. Treat because the treatment received by the patients is in the form of categories (Lithium, Imipramine, or Placebo)
3. Outcome since recurrence is in the form of two categories (Recurrence or No Recurrence)
4. Gender because the numbers represent two distinct categories: Female and Male.

Question 2: What are the quantitative variables in this dataset? From the above image, we can see that ‘Time’ , ‘AcuteT’ and ‘Age’ are quantitative variables, i.e have outcomes in datatype int or numeric (continous)

1. Time since it can take on multiple numerical values, which have arithmetic meaning (i.e., it makes sense to add, subtract, multiply, divide, or compare the magnitude of such values)
2. Age since it can take on multiple numerical values, which represent a characteristic of the patient
3. AcuteT because it can take on multiple numerical values to represent a characteristic of the patient.

Statistics Package Exercise 2 : Tallying Data and Creating Pie Charts

Loading Dataset for question 3: friends.csv

In [29]:

```
friends = pd.read_csv('/Users/home/Downloads/OneDrive_1_07-07-2020/friends.csv')
friends.head()
```

Out[29]:

	Unnamed: 0	Friends
0	1	No difference
1	2	No difference
2	3	No difference
3	4	No difference
4	5	No difference

In [30]:

```
friends.Friends.value_counts()
```

Out[30]:

```
No difference    602
Opposite sex     434
Same sex         164
Name: Friends, dtype: int64
```

In [31]:

```
friends.Friends.value_counts()/len(df1)
```

Out[31]:

```
No difference     0.501667
Opposite sex      0.361667
Same sex          0.136667
Name: Friends, dtype: float64
```

In percentage

In [32]:

```
friends.Friends.value_counts()*100/len(df1)
```

Out[32]:

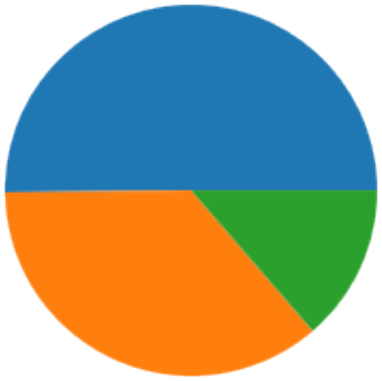
```
No difference     50.166667
Opposite sex      36.166667
Same sex          13.666667
Name: Friends, dtype: float64
```

In [33]:

```
plt.pie(friends.Friends.value_counts())
```

Out[33]:

```
([<matplotlib.patches.Wedge at 0x7f9359b57390>,
  <matplotlib.patches.Wedge at 0x7f9359ac52d0>,
  <matplotlib.patches.Wedge at 0x7f9359b57d10>],
 [Text(-0.005759554721866945, 1.0999849215009294, ''),
  Text(-0.45266576442884915, -1.0025436178611113, ''),
  Text(1.0001597573937753, -0.45790878970601206, '')])
```

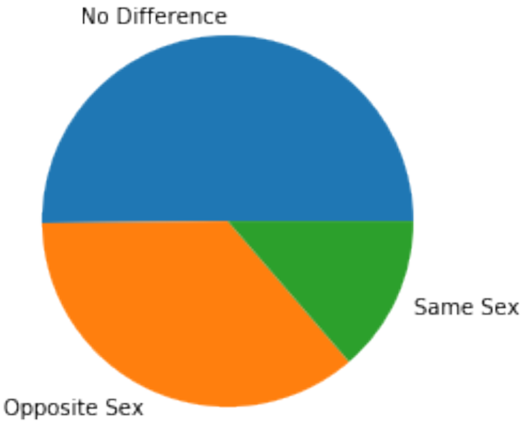


In [34]:

```
legend = ['No Difference', 'Opposite Sex', 'Same Sex']
plt.pie(friends.Friends.value_counts(), labels = legend)
```

Out[34]:

```
([<matplotlib.patches.Wedge at 0x7f9359bfda90>,
  <matplotlib.patches.Wedge at 0x7f9359bfde90>,
  <matplotlib.patches.Wedge at 0x7f9359c0a490>],
 [Text(-0.005759554721866945, 1.0999849215009294, 'No Difference')
,
  Text(-0.45266576442884915, -1.0025436178611113, 'Opposite Sex'),
  Text(1.0001597573937753, -0.45790878970601206, 'Same Sex')])
```



Question 3: Describe the distribution of the variable “friends” in dataset - Survey that asked 1,200 U.S. college students about their body perception

The students are NOT divided equally among the three categories. About 50% of the students find it as easy to make friends with the opposite sex as with the same sex. Among the remaining 50% of the students, the majority (36.2%) find it easier to make friends with people of the opposite sex, and the remainder (13.7%) find it easier to make friends with people of their own sex.

Statistics Package Exercise 3: Creating and Describing Histograms Loading Dataset for question 4 to 6: actor_age.csv

In [35]:

```
actors = pd.read_csv('/Users/home/Downloads/OneDrive_1_07-07-2020/actor_age.cs
v')
actors.head()
```

Out[35]:

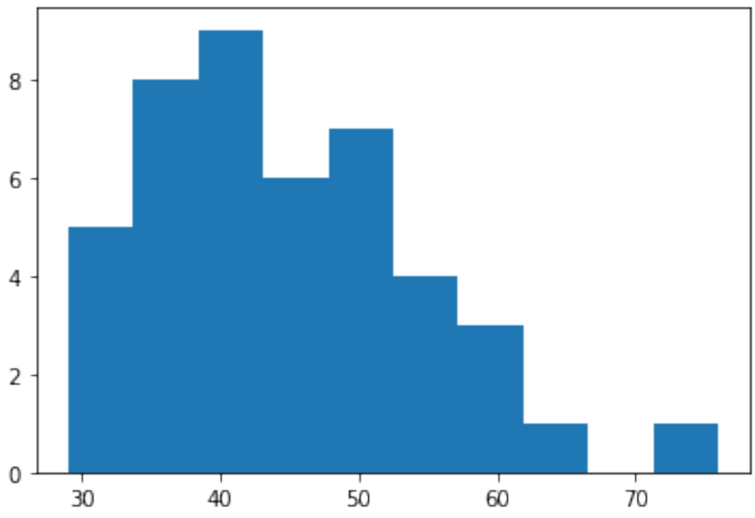
	Unnamed: 0	Age
0	1	43
1	2	42
2	3	48
3	4	49
4	5	56

In [36]:

```
plt.hist(actors.Age)
```

Out[36]:

```
(array([5., 8., 9., 6., 7., 4., 3., 1., 0., 1.]),
 array([29. , 33.7, 38.4, 43.1, 47.8, 52.5, 57.2, 61.9, 66.6, 71.3
, 76. ]),
 <a list of 10 Patch objects>)
```

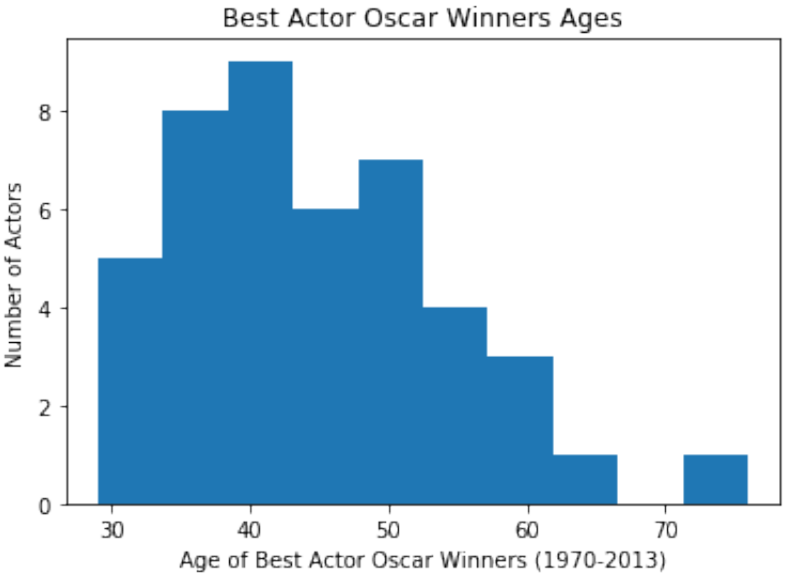


In [37]:

```
plt.hist(actors.Age)
plt.gca().set(xlabel='Age of Best Actor Oscar Winners (1970-2013)', ylabel='Nu
mber of Actors', title='Best Actor Oscar Winners Ages')
```

Out[37]:

```
[Text(0, 0.5, 'Number of Actors'),
 Text(0.5, 0, 'Age of Best Actor Oscar Winners (1970-2013)'),
 Text(0.5, 1.0, 'Best Actor Oscar Winners Ages')]
```



Question 4: Describe the distribution of the ages of the Best Actor Oscar winners. Be sure to address shape, center, spread and outliers (Dataset - Best Actor Oscar winners (1970-2013))

1. Shape: the distribution is skewed right. This means that most actors receive the best acting Oscar at a relatively younger age (before age 48), and fewer at an older age.
2. Center: The distribution seems to be centered at around 42-43. This means that about half the actors are 42 or younger when they receive the Oscar, and about half are older.
3. Spread: The age distribution ranges from about 30 to about 75. The entire dataset is covered, then, by a range of 45 years. It should be noted, though, that there is one high outlier at around age 75, and the rest of the data ranges only from 30 to 60.
4. Outliers: As mentioned above, there is one high outlier at around age 75.

In [38]:

```
actors.Age.describe()
```

Out[38]:

```
count    44.000000
mean     44.977273
std       9.749153
min      29.000000
25%      38.000000
50%      43.500000
75%      50.250000
max      76.000000
Name: Age, dtype: float64
```

Question 5: Getting information from the output: A. How many observations are in this data set? B. What is the mean age of the actors who won the Oscar? C. What is the five-number summary of the distribution? (Dataset - Best Actor Oscar winners (1970-2013))

A. There are n = 44 observations in the data set (representing the age of the Best Actor Oscar winners of the 44 years from 1970 through 2013). B. Mean = 44.98 C. The five-number summary is: min = 29, Q1 = 38, M = 43.5, Q3 = 50.5, Max = 76

Question 6: Get information from the five-number summary: A. Half of the actors won the Oscar before what age? B. What is the range covered by all the actors' ages? C. What is the range covered by the middle 50% of the ages? (Dataset - Best Actor Oscar winners (1970-2013))

A. Half the actors won the Oscar before age 43.5 (the median). B. The range covered by all the ages is: Range = Max - min = 76 - 29 = 47. C. The range covered by the middle 50% of the ages is: IQR = Q3 - Q1 = 50.5 - 38 = 12.5

Statistics Package Exercise 4: Creating Side-by-Side Boxplots

Loading Dataset: grad_data.csv

In [39]:

```
grad = pd.read_csv('/Users/home/Downloads/OneDrive_1_07-07-2020/grad_data.csv'
)
grad.head()
```

Out[39]:

	Unnamed: 0	College.A	College.B	College.C	College.D	College.E	College.F
0	1	57.6	70.1	54.5	80.1	71.3	68.8
1	2	43.2	69.6	55.6	77.3	62.6	61.0
2	3	49.6	67.3	56.9	74.7	54.5	57.7
3	4	51.4	76.7	71.0	79.2	57.5	66.4
4	5	69.9	69.4	73.3	84.6	55.0	75.2

In [40]:

```
grad.describe()
```

Out[40]:

	Unnamed: 0	College.A	College.B	College.C	College.D	College.E	College.F
count	8.00000	8.000000	8.000000	8.000000	8.000000	8.000000	8.000000
mean	4.50000	60.962500	71.287500	65.175000	79.112500	60.775000	72.775000
std	2.44949	11.955027	3.061483	8.279018	3.853547	5.892307	11.116108
min	1.00000	43.200000	67.300000	54.500000	74.100000	54.500000	57.700000
25%	2.75000	50.950000	69.550000	56.575000	76.650000	56.875000	65.050000
50%	4.50000	63.750000	70.150000	67.650000	79.000000	59.150000	72.000000
75%	6.25000	70.500000	73.050000	71.575000	81.100000	63.700000	81.275000
max	8.00000	73.800000	76.700000	74.800000	84.600000	71.300000	87.400000

In [42]:

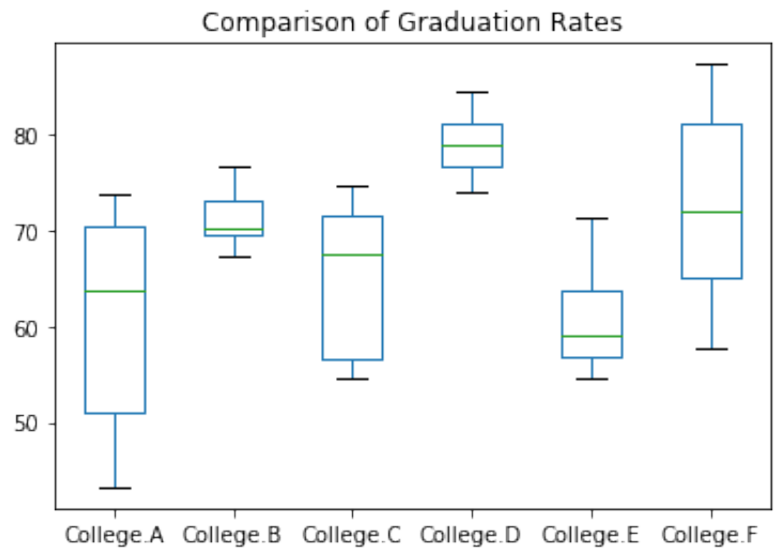
```
grad1 = grad.drop('Unnamed: 0', axis=1)
```

In [43]:

```
grad1.plot(kind='box', title='Comparison of Graduation Rates')
```

Out[43]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9359f89990>
```



In [44]:

```
ratings = pd.read_csv('/Users/home/Downloads/OneDrive_1_07-07-2020/ratings.csv
')
ratings.head()
```

Out[44]:

	Unnamed: 0	Class.I	Class.II	Class.III
0	1	1	1	1
1	2	1	1	2
2	3	5	1	3
3	4	5	1	4
4	5	5	1	5

In [45]:

```
rating = ratings.drop('Unnamed: 0', axis=1)
```

In [46]:

```
rating.std(axis=0)
```

Out[46]:

```
Class.I       1.568929
Class.II      4.000000
Class.III     2.631174
dtype: float64
```

Question 7: What are the standard deviations of the three rating distributions? Was your intuition correct? (Dataset - 27 students in the class were asked to rate the instructor on a number scale of 1 to 9)

Here are the three standard deviations: Class I: 1.569 Class II: 4.000 Class III: 2.631 Note that through this example, we also learn that the number of distinct values represented in a histogram does not necessarily indicate greater variability.

Question 8: Assume that the average rating in each of the three classes is 5 (which should be visually reasonably clear from the histograms), and recall the interpretation of the SD as a “typical” or “average” distance between the data points and their mean. Judging from the table and the histograms, which class would have the largest standard deviation, and which one would have the smallest standard deviation? Explain your reasoning (Dataset - 27 students in the class were asked to rate the instructor on a number scale of 1 to 9)

In class I, almost all the ratings are 5, which is also the mean. The average distance between the observations and the mean, then, would be very small. In class II most of the observations are far from the mean (at 1 or 9). The average distance between the observations and the mean in this case would be larger. Class III is the case where some of the observations are close to the mean, and some are far, so the average distance between the observations and the mean would be somewhere in between class I and II. Ranking of Std. Dev from smallest to largest - Class I < Class III < Class II

End of assignment. Return to top. By Mithesh R, J002. ML