e) Yes, this data does support the researchers' conjecture that there is information in a person's voice to help identify the taller person because each data statistic points towards this. Both the mean and median are 62% which could not be considered a coincidence. Furthermore, both the 1^{st} and 3^{rd} quartiles are above 50% as well (57% and 67.5% respectively). All of this shows that on average, most people can tell whether somebody is taller than another person more than half of the time.

Question assigned to the following page: 2.4

d) **1st Quartile:**

49, 53, 55, 56, 56, 56, 58, 58, 58, 59, 61, 61

The 1st quartile is between 56 and 58, therefore, it is 57.

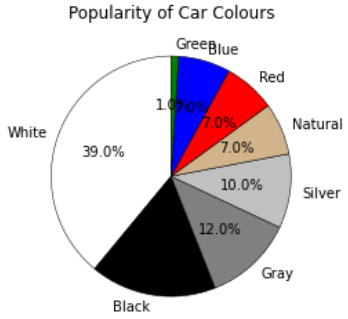
3rd Quartile:

63, 65, 66, 67, 67, 67, 68, 68, 69, 69, 70, 70

The 3rd quartile is between 67 and 68, therefore, it is 67.5.

Question assigned to the following page: 4

In [76]:
```
import numpy
import pandas
import matplotlib.pyplot as plt
import seaborn as sns
from numpy import percentile

car_colours_csv = pandas.read_csv('ex01-25carcolor.csv')

car_colors = car_colours_csv['Color'].tolist()

colors_lowercase = []

for color in car_colors:
    colors_lowercase.append(color.lower())

colors_lowercase[4] = 'tan'

popular_values = car_colours_csv['Popular'].tolist()

car_pie = plt.pie(popular_values, labels = car_colors, colors = colors_lowercase, startangle = 90, autopct ='%1.1f%%',
                  wedgeprops = {"edgecolor" : "black",'linewidth': 0.5,  'antialiased': True})
car_pie = plt.title('Popularity of Car Colours')
```



Question assigned to the following page: 1.4

c) The smallest annual return was below -40% and the largest annual return was between 50% and 60%

d) About 27.47% (25/91) of all years had a return below 0%.

Question assigned to the following page: 7.3

c) No, two subjects with the same weight loss would not have the same benefits from a weight-reduction program. If one was already in good shape or underweight, losing a significant amount of weight would not be good but it may be good for somebody who is overweight. Weight loss is proportional! It depends on your initial weight!

I do see an advantage in considering excess weight loss. First of all, losing excess weight may not be good. One can become underweight from this. Furthermore, the excess weight lost may be muscle as well which is not a good thing. One should aim to lose fat, not muscle. As for a reduction in BMI, I do not see an advantage in this variable as the BMI has been known to be an unreliable source for tracking one's health/weight. It is not accurate as it fails to take into account a number of different factors.

Question assigned to the following page: 2.3

c) **Mean:**

(49 + 53 + 55 + 56 + 56 + 56 + 58 + 58 + 58 + 59 + 61 + 61 +63 + 65 + 66 + 67 + 67 + 67 + 68 + 68 + 69 + 69 + 70 + 70) / 24 = 62.041666

The mean is 62.041666

Median:

Median = 49, 53, 55, 56, 56, 56, 58, 58, 58, 59, 61, 61, 63, 65, 66, 67, 67, 67, 68, 68, 69, 69, 70, 70

The median is between 61 and 63, therefore, the median is 62.

Question assigned to the following page: 3

3. **Mean:**

(31 + 35 + 33 + 31 + 29 + 34 + 29 + 30 + 29 + 28) / 10 = 30.9

The mean is 30.9

Standard Deviation:

Mean = 30.9 (from above)

Deviation for each data point:

31 - 30.9 = 0.1

35 - 30.9 = 4.1

33 – 30.9 = 2.1

31 – 30.9 = 0.1

29 – 30.9 = -1.9

34 – 30.9 = 3.1

29 – 30.9 = -1.9

30 – 30.9 = -0.9

29 – 30.9 = -1.9

28 – 30.9 = -2.9

Deviation Squared for each data point:

$0.1^2 = 0.01$

$4.1^2 = 16.81$

$2.1^2 = 4.41$

$0.1^2 = 0.01$

$(-1.9)^2 = 3.61$

$3.1^2 = 9.61$

$(-1.9)^2 = 3.61$

$(-0.9)^2 = 0.81$

$(-1.9)^2 = 3.61$

$(-2.9)^2 = 8.41$

Sum = 0.01 + 16.81 + 4.41 + 0.01 + 3.61 + 9.61 + 3.61 + 0.81 + 3.61 + 8.41 = 50.9

Sum/(# of data points – 1):

50.9/9 = 5.655555

Final Calculation:

$\sqrt{5.655555} = 2.37814$

The standard deviation is 2.37814

4.

Question assigned to the following page: 1.3

c) The smallest annual return was below -40% and the largest annual return was between 50% and 60%

d) About 27.47% (25/91) of all years had a return below 0%.

Question assigned to the following page: 5.3

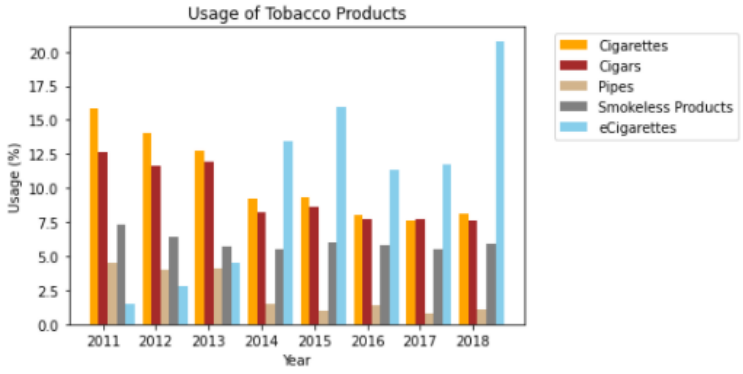
c) Tobacco usage of teens of any product stayed the same between 2011 and 2018, however, the products they used did change. As seen in the second bar graph comparing each individual product, all of the products saw a decline in its usage or stayed relatively the same with no increase. E-cigarettes on the other hand did see a significant increase.

Question assigned to the following page: 6.3

c) These values show that fish is incredibly healthy for humans because the omega-3 to omega-6 ratio is greater than 1, and therefore, it has more omega-3 than omega-6, which is what health experts recommend.

Question assigned to the following page: 5.2

In [78]:
```
import numpy
import pandas
import matplotlib.pyplot as plt
import seaborn as sns
from numpy import percentile

tobacco_use = pandas.read_csv('tobacco_data.csv')

cigarettes = tobacco_use['Cigarettes']
cigaretteList = cigarettes.tolist()

cigars = tobacco_use['Cigars']
cigarList = cigars.tolist()

pipes = tobacco_use['Pipes']
pipeList = pipes.tolist()

smokeless = tobacco_use['Smokeless_tobacco']
smokelessList = smokeless.tolist()

ecigarettes = tobacco_use['Ecigarettes']
ecigList = ecigarettes.tolist()

years = tobacco_use['year']
year_list = years.tolist()

x = numpy.arange(len(year_list))

bar_width = 0.17
fig, data=plt.subplots()
cigaretteBar = data.bar(x-bar_width,cigaretteList, bar_width,label = 'Cigarettes', color='orange')
cigarBar = data.bar(x, cigarList, bar_width,label='Cigars', color='brown')
pipeBar = data.bar(x+bar_width, pipeList, bar_width,label = 'Pipes', color = 'tan')
smokelessBar = data.bar(x + 2*bar_width, smokelessList, bar_width, label = 'Smokeless Products', color='gray')
ecigBar = data.bar(x+3*bar_width, ecigList, bar_width, label = 'eCigarettes', color='skyblue')
plt.xlabel('Year')
plt.ylabel('Usage (%)')
plt.title('Usage of Tobacco Products')
plt.xticks(x, year_list)
plt.legend(bbox_to_anchor=(1.05, 1.0), loc='upper left')
plt.show()
```



Question assigned to the following page: 2.2

b) The shape of the dataset is bimodal (according to the split stem plot) because there appear to be 2 modes (high 50s and high 60s). The center of the dataset seems to be in the low 60s range, specifically, 61. As for variability, there are no extreme outliers; each data value seems to be in the normal range. Furthermore, the 1^{st} and 3^{rd} quartiles seem to be normal as well, as the 1^{st} quartile seems to be in the high 50s section and the 3^{rd} quartile seems to be in the high 60s section.

Question assigned to the following page: 7.2

In [73]:

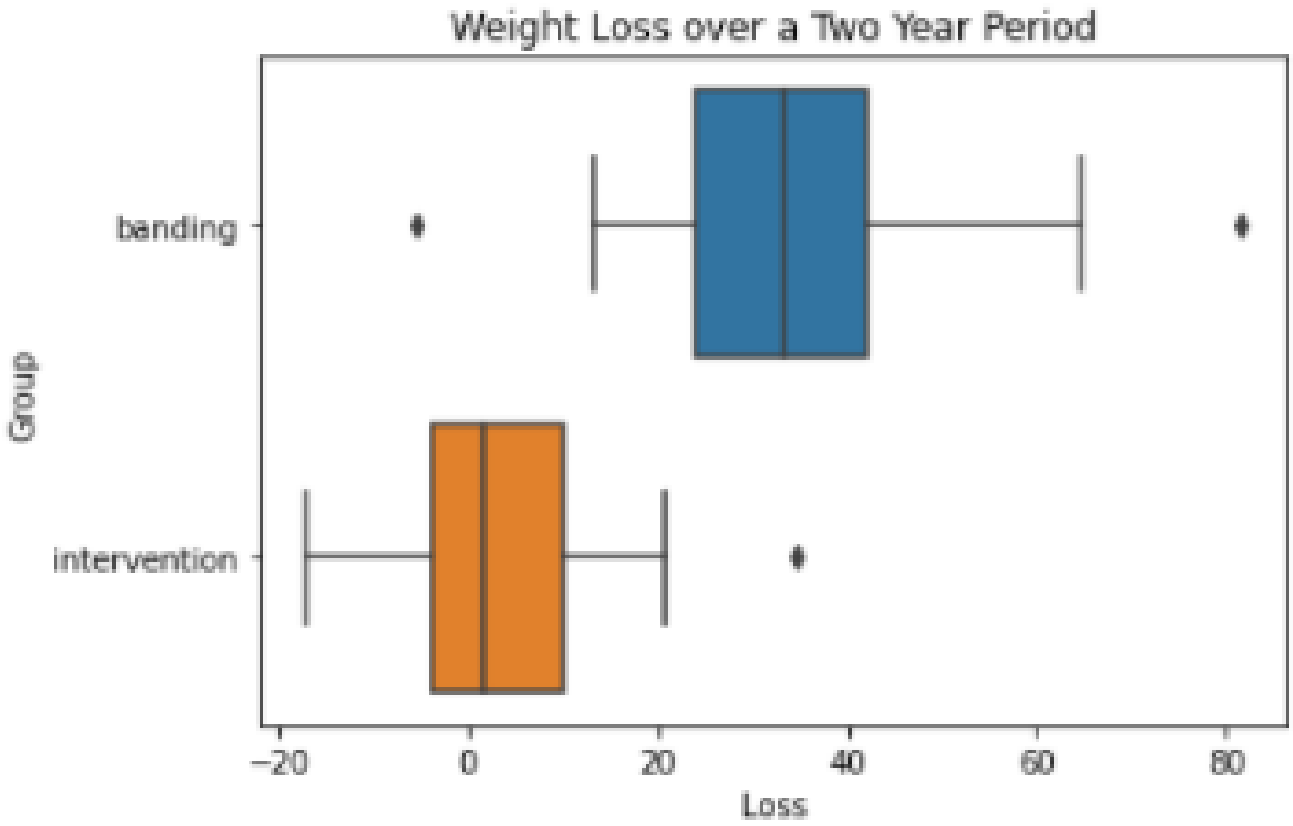
```
import numpy
import pandas
import matplotlib.pyplot as plt
import seaborn as sns
from numpy import percentile

gastricDF = pandas.read_csv('ex02-42gastric.csv')
gastricBP = sns.boxplot(x = gastricDF['Loss'], y=gastricDF['Group'])
plt.title('Weight Loss over a Two Year Period')

bandingValues = gastricDF[gastricDF['Group']=='banding']

quartilesBanding = percentile(bandingValues['Loss'] ,[25,50,75])

iqrBanding = (quartilesBanding[2]-quartilesBanding[0])

maximumBanding = (quartilesBanding[2] + 1.5 * iqrBanding)


minimumBanding = (quartilesBanding[0] - 1.5 * iqrBanding)


outliersBanding = []
for num in bandingValues['Loss']:
    if num > float(maximumBanding):
        outliersBanding.append(num)


interventionValues = gastricDF[gastricDF['Group']=='intervention']

quartilesIntervention = percentile(interventionValues['Loss'],[25,50,70])

iqrIntervention = (quartilesIntervention[2]-quartilesIntervention[0])

maximumIntervention = (quartilesIntervention[2]+1.5*iqrIntervention)

minimumIntervention = (quartilesIntervention[0]-1.5 * iqrIntervention)

outliersIntervention = []
for num in interventionValues['Loss']:
    if num > float(maximumIntervention):
        outliersIntervention.append(num)
```

Question assigned to the following page: 7.2

```
outliersIntervention = []
for num in interventionValues['Loss']:
    if num > float(maximumIntervention):
        outliersIntervention.append(num)

# Calculations directly BELOW were done via Jupyter notebook through the .describe() method, and therefore,
# may differ from hand calculations
print('Numerical Summary for Banding:')
print(bandingValues.describe())

print('\nNumerical Summary for Intervention:')
print(interventionValues.describe())

print()

# Calculations directly BELOW were done via Jupyter notebook through the hand calculation method, and therefore,
# may differ from Python methods
print('Five-Number Summary for Banding (with high outliers)')
print('Min:', minimumBanding)
print('Q1:', quartilesBanding[0])
print('Median:', quartilesBanding[1])
print('Q3:', quartilesBanding[2])
print('Max:', maximumBanding)
print('High Outliers:', outliersBanding)

print('\nFive-Number Summary for Intervention (with high outliers)')
print('Min:', minimumIntervention)
print('Q1:', quartilesIntervention[0])
print('Median:', quartilesIntervention[1])
print('Q3:', quartilesIntervention[2])
print('Max:', maximumIntervention)
print('High Outliers:', outliersIntervention)
```

Question assigned to the following page: 7.2

```
Numerical Summary for Banding:
             Loss
count  24.000000
mean   34.866667
std    18.123002
min    -5.400000
25%    24.100000
50%    33.350000
75%    42.025000
max    81.400000

Numerical Summary for Intervention:
             Loss
count  18.000000
mean    3.011111
std    13.217541
min   -17.000000
25%    -4.000000
50%     1.700000
75%    10.200000
max    34.600000

Five-Number Summary for Banding (with high outliers)
Min: -2.7875000000000005
Q1: 24.1
Median: 33.349999999999994
Q3: 42.025000000000006
Max: 68.91250000000001
High Outliers: [81.4]

Five-Number Summary for Intervention (with high outliers)
Min: -19.0
Q1: -4.0
Median: 1.7
Q3: 6.0
Max: 21.0
High Outliers: [34.6]
```


Weight Loss over a Two Year Period
bending
intervention
Group
−20
0
20
40
60
80
Loss

Question assigned to the following page: 7.2

Gastric banding was certainly a better method for weight loss when compared to lifestyle intervention if the only factor considered was total weight lost (costs, safety, etc. not included).

Question assigned to the following page: 1.2

b) The approximate center of the distribution is the 10% mark. There would be roughly 40 years with lower returns than this range and roughly 51 years with higher returns than this range. These numbers were estimated after analyzing and getting the sum of the years visually.

c) The smallest annual return was below -40% and the largest annual return was between 50% and 60%

Question assigned to the following page: 6.2

b) The shape of this distribution is some variation of a right-skewed graph, but not totally. Only 7 foods have more omega-3 than omega-6. Because of this, most modern food oils are not good for our health because most of the time, they contain more omega-6 than omega-3.

Question assigned to the following page: 1.1

1.

a) The overall shape of the distribution of monthly returns seems to be a bell shape that is slight left skewed.

Question assigned to the following page: 2.1

2.

a)

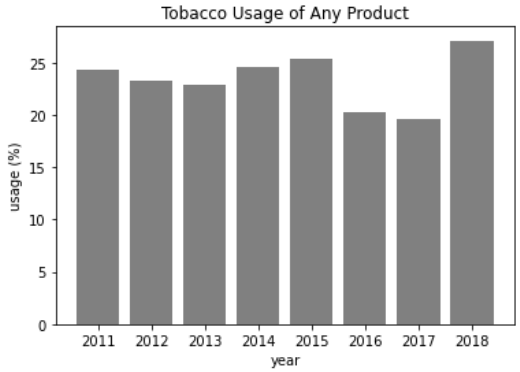
Stem	Leaves
4	9
5	3 5 6 6 6 8 8 8 9
6	1 1 3 5 6 7 7 7 8 8 9 9
7	0 0

Stem	Leaves
4	
4	9
5	3
5	5 6 6 6 8 8 8 9
6	1 1 3
6	5 6 7 7 7 8 8 9 9
7	0 0
7	

I prefer the second plot (the one with split stems) because it gives more information quicker. I can see more easily that more numbers are in a higher range in the 50s and 60s section than in the lower range of the same stem.

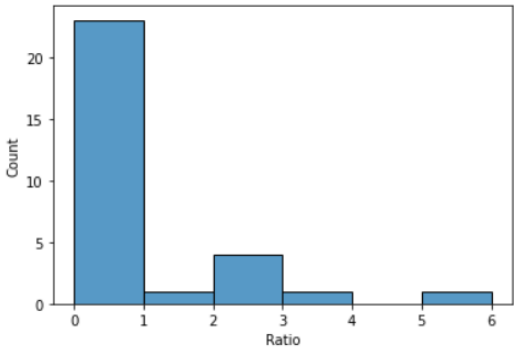
Question assigned to the following page: 5.1

In [77]:
```
import numpy
import pandas
import matplotlib.pyplot as plt
import seaborn as sns
from numpy import percentile

tobacco_use = pandas.read_csv('tobacco_data.csv')
anyTobaccoProduct = tobacco_use['Any_tobacco']

years = tobacco_use['year']

anyTPvalues = anyTobaccoProduct.tolist()

year_list = years.tolist()

tobacco_bar = plt.bar(year_list, anyTPvalues, color = 'gray')
tobacco_bar = plt.xlabel('year')
tobacco_bar = plt.ylabel('usage (%)')
tobacco_bar = plt.title('Tobacco Usage of Any Product')

# There is no pattern of change in the usage of any tobacco product. There is no growth or decline. It is basically a flatline.
```



Question assigned to the following page: 6.1

In [68]:
```
import numpy
import pandas
import matplotlib.pyplot as plt
import seaborn as sns
from numpy import percentile

oilsData = pandas.read_csv('ex01-34foodoils.csv')
oils = oilsData['Oil']
ratios = oilsData['Ratio']

sns.histplot(data=oilsData, x="Ratio", binwidth=1)
plt.show()
```



Question assigned to the following page: 7.1

7.

a) In the context of this study, a negative value meant that the subject gained weight during the two-year period instead of losing it as would be expected.