assignment 14 (own)

March 8, 2022

1 Bivariate analysis regarding correlation on 1 combination of 2 columns of categorical data on Mushrooms dataset

[27]:
```
import pandas as pd
```

[28]:
```
import seaborn as sns
```

[29]:
```
from scipy.stats import chi2_contingency
```

As my previous dataset had mostly numerical data, I am using this dataset for categorical excersises. The data is categorical data about mushrooms. Each mushrooms has data on whether it is poisinous or not and information about the mushrooms properties. https://www.kaggle.com/hatterasdunton/mushroom-classification-updated-dataset?select=mushroomsupdated.csv

[34]:
```
df = pd.read_csv('mushrooms.csv', sep=',')
sns.set_style("dark")
```

[35]:
```
df.head()
```

[35]:
```
       class cap-shape cap-surface cap-color     bruises     odor  \
0  Poisonous    Convex      Smooth     Brown     Bruises  Pungent
1     Edible    Convex      Smooth    Yellow     Bruises   Almond
2     Edible      Bell      Smooth     White     Bruises    Anise
3  Poisonous    Convex       Scaly     White     Bruises  Pungent
4     Edible    Convex      Smooth     Green  No Bruises     None

  gill-attachment gill-spacing gill-size gill-color  ...  \
0             Free        Close    Narrow      Black  ...
1             Free        Close     Broad      Black  ...
2             Free        Close     Broad      Brown  ...
3             Free        Close    Narrow      Brown  ...
4             Free      Crowded     Broad      Black  ...

  stalk-surface-below-ring stalk-color-above-ring stalk-color-below-ring  \
0                   Smooth                  White                  White
1                   Smooth                  White                  White
```

```
2                   Smooth                   White                  White
3                   Smooth                   White                  White
4                   Smooth                   White                  White

  veil-type veil-color ring-number   ring-type spore-print-color population  \
0   Partial      White         One     Pendant             Black  Scattered
1   Partial      White         One     Pendant             Brown   Numerous
2   Partial      White         One     Pendant             Brown   Numerous
3   Partial      White         One     Pendant             Black  Scattered
4   Partial      White         One  Evanescent             Brown   Abundant

   habitat
0    Urban
1  Grasses
2  Meadows
3    Urban
4  Grasses

[5 rows x 23 columns]
```

I would like to see if color and poininousness are correlated. I know this is the case with animals, but I don't know about mushrooms

[36]:

```
table = df.groupby(['cap-color','class']).size().unstack('class', fill_value=0)
table
```

[36]:

```
class      Edible  Poisonous
cap-color
Brown        1264       1020
Buff           48        120
Cinnamon       32         12
Green        1048        808
Pink           56         88
Purple         16          0
Red           624        876
White         720        320
Yellow        400        672
```

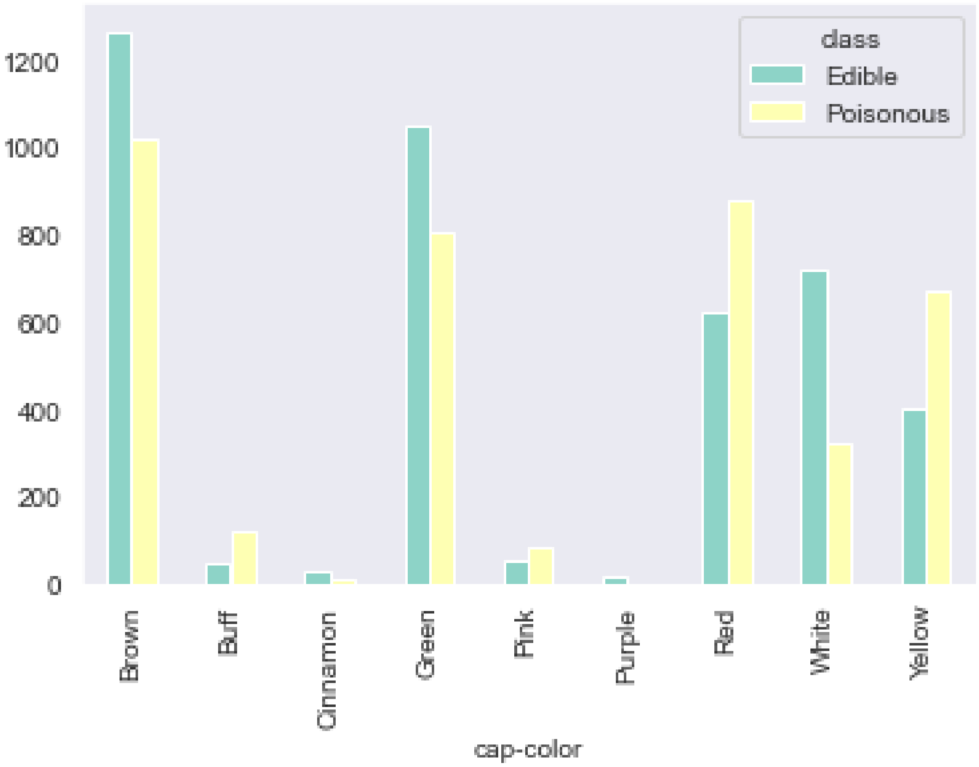
There are some significant differences in ratios: Brown doesn't say anything, white means it is probably okay and all purple mushrooms (in this dataset at least) are poisionous.

[39]:

```
table.plot(kind='bar')
```

[39]:

```
<AxesSubplot:xlabel='cap-color'>
```



The bar plot also shows significantly different ratios

```
[38]: chi2_contingency(table)
```

```
[38]: (375.346859678969,
       3.495286115362265e-76,
       8,
       array([[1183.04677499, 1100.95322501],
              [  87.01920236,   80.98079764],
              [  22.79074348,   21.20925652],
              [ 961.35499754,  894.64500246],
              [  74.58788774,   69.41211226],
              [   8.28754308,    7.71245692],
              [ 776.95716396,  723.04283604],
              [ 538.69030034,  501.30969966],
              [ 555.26538651,  516.73461349]]))
```

The chance of the two variables being correlated is very high.