Summarizing the Relationship between Two Features

Associated Variables

When variables are associated, information about the value of one variable provides information about the value of the other variable. For example, average temperature might be associated with ice cream sales because people tend to buy more ice cream in summer months, when the temperature is hotter. This doesn't necessarily mean that the higher temperature **causes** more people to buy ice cream, but it does mean that we can predict ice cream sales more accurately if we know what the temperature is outside.

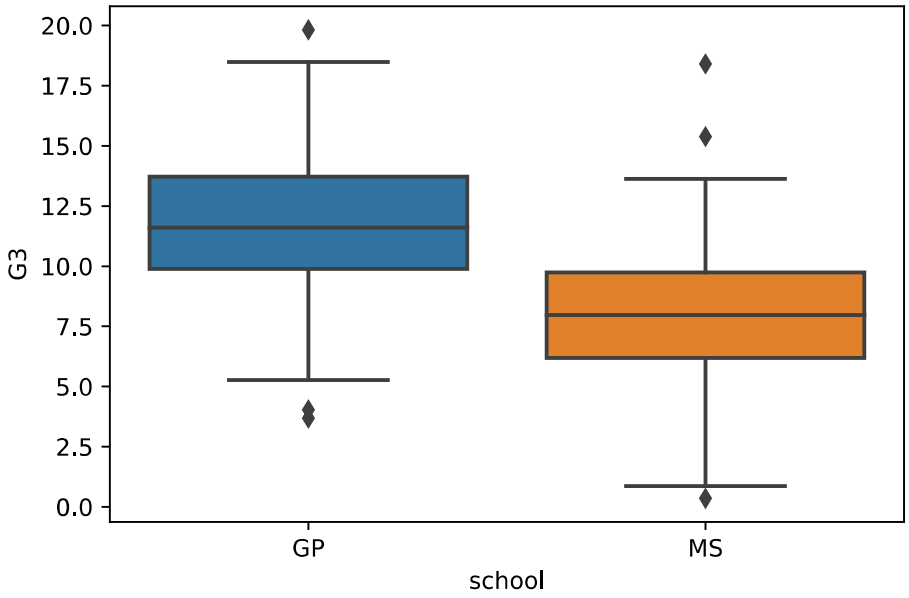
Mean and Median Differences

Mean and median differences are summary statistics that can be used to assess an association between a quantitative variable and a categorical variable. For example, if we want to evaluate whether there is an association between whether or not someone receives a drug and their heart rate, we might calculate the difference in mean heart rate for people who took the drug compared to the mean heart rate for people who did not take the drug. If we find that people who took the drug have an average heart rate that's 10 beats per minute lower, that would provide evidence of an association.

```
import numpy as np

hr_drug = data.hr[data.group == 'drug']
hr_no_drug = data.hr[data.group == 'no drug']

# to calculate a mean difference:
mean_diff = np.mean(hr_drug) - np.mean(hr_no_drug)

# to calculate a median difference:
median_diff = np.median(hr_drug) - np.median(hr_no_drug)
```

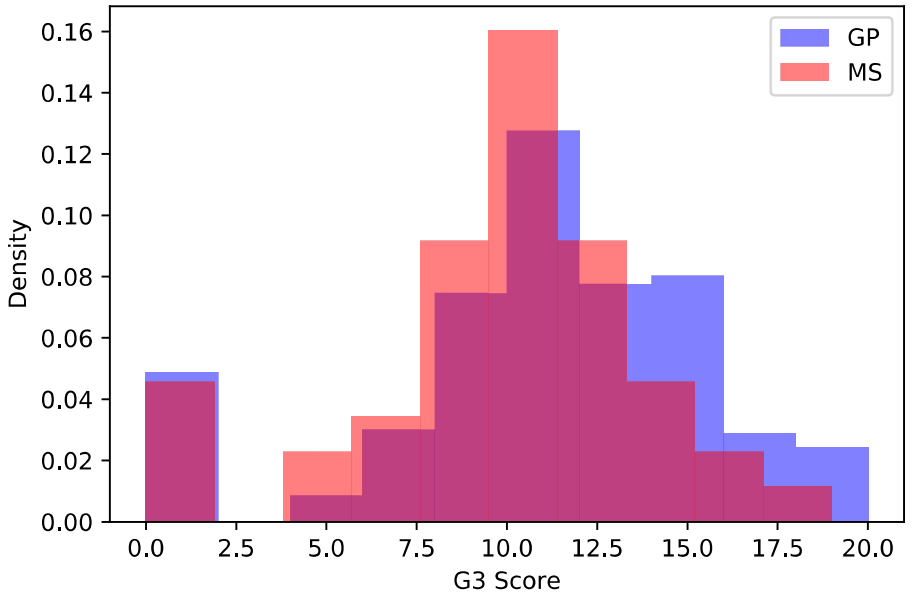
Side-by-Side Box Plots

Side-by-side box plots can be used along with mean and median differences to assess whether a quantitative variable and a categorical variable are associated. More overlap in the box plots indicates less association while less overlap in the box plots indicates a stronger association. For example, this image shows a side-by-side box plot of math scores at two different schools. Students seem to be scoring higher at one school than the other, suggesting that there might be an association between what school a student attends and their math score.



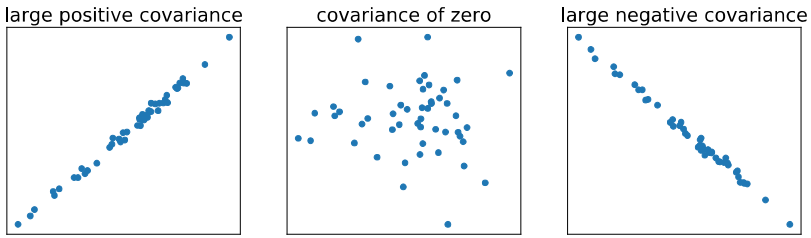
Overlaid Histograms

Overlaid histograms can be used along with mean and median differences to assess an association between a quantitative variable and a categorical variable. After normalizing the histograms, more overlap indicates less association and less overlap indicates a stronger association. The example image shows math scores for students at two different schools. We see that scores tend to be higher for students at the GP school, but there is a lot of overlap in these distributions — suggesting that the association is not very strong.



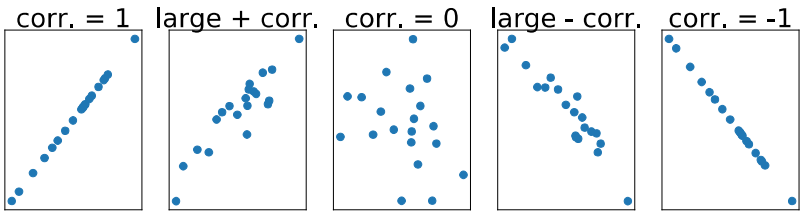
Covariance

Covariance ranges from negative infinity to positive infinity and is used to measure the strength of a linear association between two quantitative variables. A large negative covariance indicates a strong negative linear association where large values of one variable are associated with small values of the other. A large positive covariance indicates a strong positive linear association where large values of one variable are associated with large values of the other. A covariance of 0 indicates there is no linear relationship.



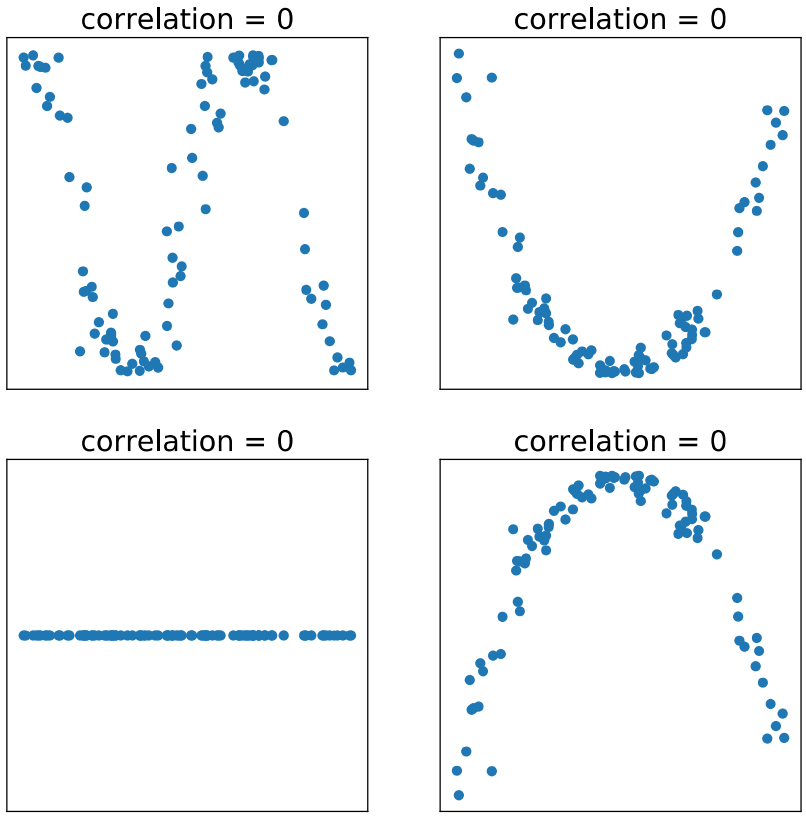
Correlation

Correlation ranges from negative one to positive one and is used to measure the strength of a linear association between two quantitative variables. A correlation closer to negative one indicates a strong negative linear where large values of one variable are associated with small values of the other. A correlation closer to positive one indicates high positive linearity where large values of one variable are associated with large values of the other. A correlation of 0 indicates there is no linear relationship. The figure shows pairs of variables with correlations ranging from negative one to one.



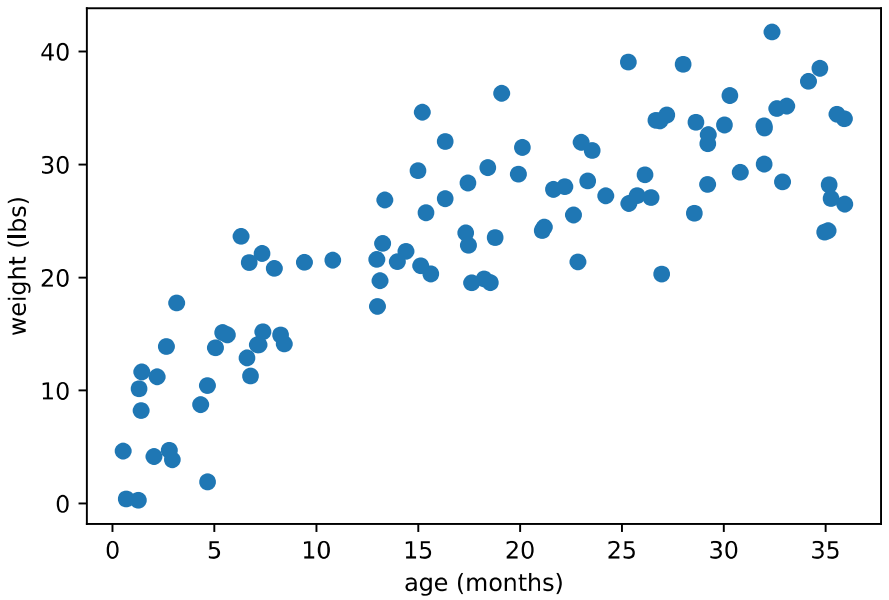
Non-linear Associations

Correlation and covariance measure the strength of a linear relationship, so a correlation or covariance of 0 means the variables are not **linearly** associated (or linearly associated with a slope of zero, which can be visualized as a horizontal line). However, other non-linear relationships between the variables may still exist. The example plots all show pairs of variables with a correlation close to zero.



Scatter Plots

A scatter plot can be used to visually inspect whether there is an association between two quantitative variables. If there is a pattern in the plot, the variables are associated; if there is no pattern, the variables are not associated. For example, this plot shows a pair of associated variables; children who are older tend to weigh more.



Contingency Tables

To determine whether two categorical variables are associated, it is helpful to look at a contingency table. For example, suppose that we ask 85 people whether they prefer chips or candy and whether they prefer pizza or cake. We could assemble their responses into a contingency table, which might look like:

	chips	candy
pizza	20	10
cake	5	50

This table indicates that 20 people said they prefer chips and pizza.

```
# to create a contingency table in python:
import pandas as pd
pd.crosstab(variable1, variable2)
```

Contingency Table Proportions

When determining whether two variables are associated, it can be helpful to look at a contingency table of proportions. Contingency tables are often given in frequencies and can be converted to proportions by dividing each frequency by the total number of observations. The provided code sample shows two equivalent ways of creating a contingency table of proportions in Python.

```
import pandas as pd

# calculate "by hand"
Xtab_freq = pd.crosstab(data.var1, data.var2)
Xtab_prop = Xtab_freq/len(data)

# calculate with pd.crosstab
Xtab_prop = pd.crosstab(data.var1, data.var2, normalize =
True)
```

Marginal Proportions

A marginal proportion in a contingency table is the proportion of observations in a single category of one variable. If given a contingency table of proportions, the marginal proportion can be calculated by taking the row and column sums. If given a contingency table of frequencies, the marginal proportion can be calculated by dividing the row or column sum by the total number of observations.

```
Consider the following contingency table showing food
preferences for two binary questions:
        chips   candy
pizza      20      10
cake       15       5

Then the marginal for chips is: proportion preferring chips
= (20+15)/(20+15+10+5)
```

Chi-Square Statistic

The Chi-Square statistic is used to summarize an association between two categorical variables. The Chi-Square statistic ranges from zero to infinity. The more associated two variables are, the larger the Chi-Square statistic will be.

```
#python implementation
from scipy.stats import chi2_contingency
frequency_table = pd.crosstab(df.variable_1, df.variable_2)
chi2, pval, dof, expected =
chi2_contingency(frequency_table)
```