3.3 Relationships Between Quantitative Variables

May 9, 2019

1 3.3 Relationships between Two Quantitative Variables

In this chapter, we discuss ways to summarize and visualize relationships between *quantitative* variables. To illustrate the concepts, we use the Ames housing data set.

In [1]:
```
%matplotlib inline
import pandas as pd

housing_df = pd.read_csv("https://raw.githubusercontent.com/dlsun/data-science-book/mas
housing_df.head()
```

Out[1]:

	Order	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	\
0	1	526301100	20	RL	141.0	31770	Pave	
1	2	526350040	20	RH	80.0	11622	Pave	
2	3	526351010	20	RL	81.0	14267	Pave	
3	4	526353030	20	RL	93.0	11160	Pave	
4	5	527105010	60	RL	74.0	13830	Pave	

	Alley	Lot Shape	Land Contour	...	Pool Area	Pool QC	Fence	\
0	NaN	IR1	Lvl	...	0	NaN	NaN	
1	NaN	Reg	Lvl	...	0	NaN	MnPrv	
2	NaN	IR1	Lvl	...	0	NaN	NaN	
3	NaN	Reg	Lvl	...	0	NaN	NaN	
4	NaN	IR1	Lvl	...	0	NaN	MnPrv	

	Misc Feature	Misc Val	Mo Sold	Yr Sold	Sale Type	Sale Condition	SalePrice
0	NaN	0	5	2010	WD	Normal	215000
1	NaN	0	6	2010	WD	Normal	105000
2	Gar2	12500	6	2010	WD	Normal	172000
3	NaN	0	4	2010	WD	Normal	244000
4	NaN	0	3	2010	WD	Normal	189900

[5 rows x 82 columns]

1.1 Visualization

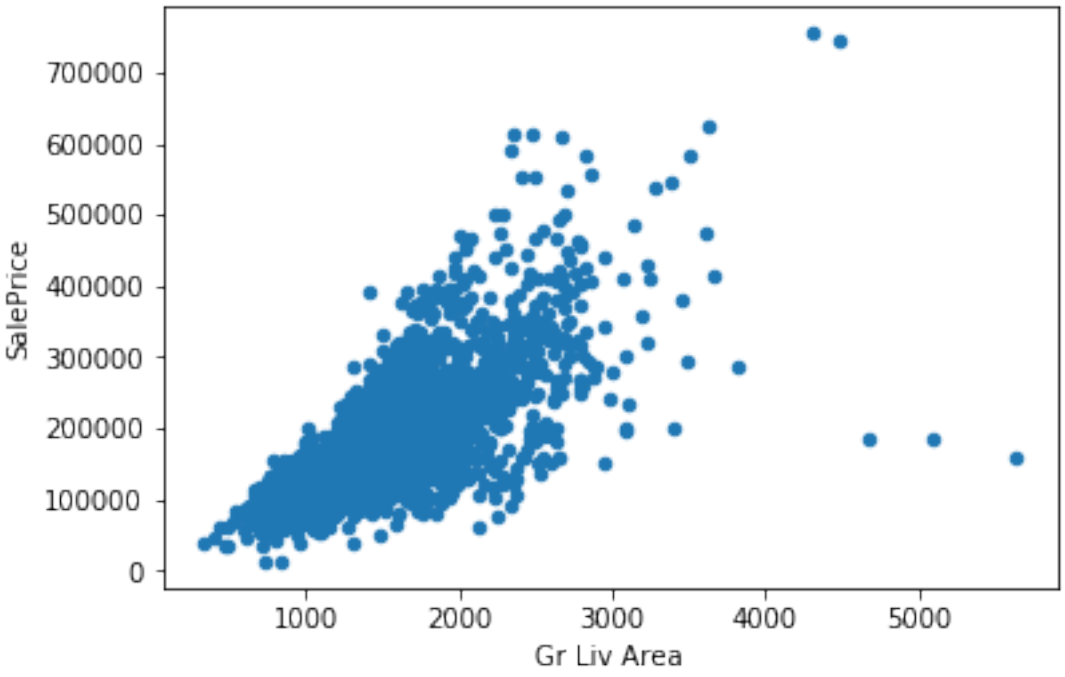
Let's start by visualizing the relationship between the square footage (of the dwelling) and the sale price, both of which are quantitative variables. To do this, we can make a **scatterplot**. In

a scatterplot, each observation is represented by a point. The (x, y) coordinates of each point represent the values of two variables for that observation.

To make a scatterplot in `pandas`, we use the `.plot.scatter()` method of `DataFrame`. Since there are multiple columns in the `DataFrame`, we have to specify which variable is x and which variable is y.

```
In [2]: housing_df.plot.scatter(x="Gr Liv Area", y="SalePrice")
```

```
Out[2]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff6343812e8>
```



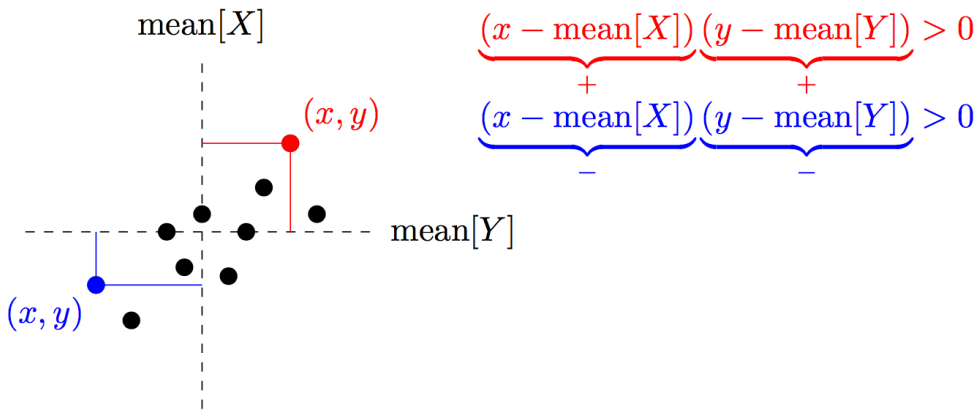
We see that square footage (of the dwelling) and the sale price have a positive relationship. That is, the greater the living area, the higher the sale price.

1.2 Summary Statistics

To summarize the relationship between two quantitative variables X and Y, we can report the *covariance* between them, defined as

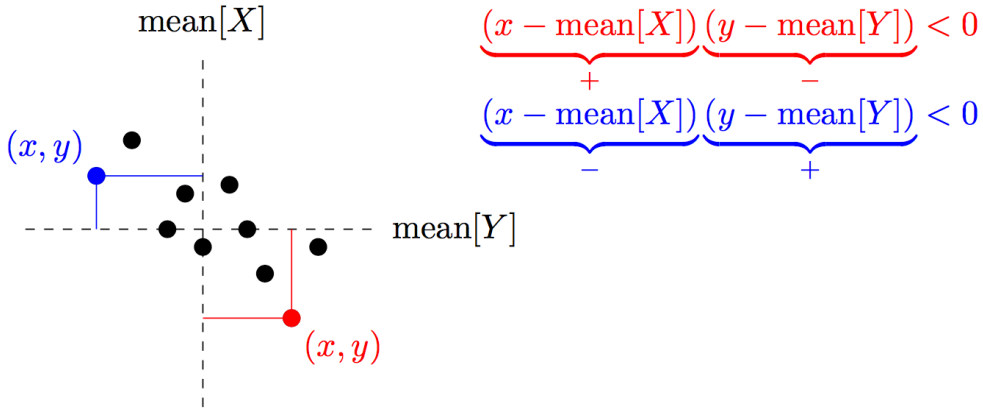
$$\text{Cov}[X, Y] = \frac{1}{n-1} \sum_x \sum_y (x - \text{mean}[X])(y - \text{mean}[Y])$$

The sign of the covariance will match the direction of the relationship between the two variables. The figures below illustrate why. If two variables are positively related, then the scatterplot might look as follows, with most points in the upper-right and lower-left quadrants (when you divide up the plane into four quadrants based on the means of X and Y).

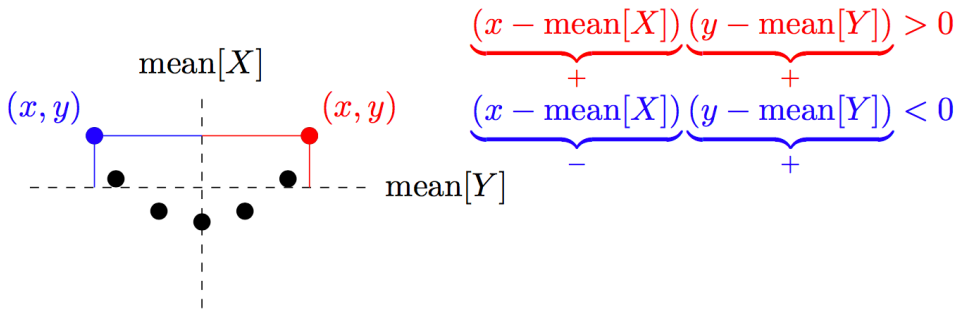


Each point on this scatterplot contributes to the sum that makes up the covariance. Any point in the upper-right quadrant (where x and y are both greater than their respective means) has a positive contribution, since the product of two positive numbers is positive. A point in the lower-left quadrant (where x and y are both less than their respective means) also has a positive contribution, since the product of two negative numbers is also positive. Therefore, on the whole, the covariance will be positive for two variables with a positive relationship.

We can also consider two variables with a negative relationship. A scatterplot of two negatively-related variables might look as follows, with most points in the upper-left and lower-right quadrants. Points in both of these quadrants will have a negative contribution towards the covariance, since the product of a positive and a negative number is negative.



What does it mean for the covariance to be *zero*? It does not necessarily mean that there is *no* relationship at all between the two variables; it just means that the two variables do not move in a consistent direction. For example, the two variables below have *zero* covariance because the negative contributions from the upper-left and lower-right quadrants perfectly cancel out the positive contributions from the upper-right and lower-left quadrants. However, it would be inaccurate to say that X and Y have *no* relationship; they have a strong relationship, but it just is not consistently in one direction.



To calculate the covariance between two quantitative variables, we use the `.cov()` method in `pandas`. This method is attached to one `Series` and takes another `Series` of the same length as input. It returns the covariance between the two `Series`.

```
In [3]: housing_df["Gr Liv Area"].cov(housing_df["SalePrice"])
```

```
Out[3]: 28542199.568276513
```

The covariance between the two variables is positive, as should be apparent from the scatterplot above. Larger houses sell for higher prices.

One criticism of the covariance is that the value itself is difficult to interpret, and covariances are not comparable across different variables. As we did with the χ^2 distance in the previous section, we can normalize the covariance. This *normalized covariance* is called the **correlation** and is symbolized r:

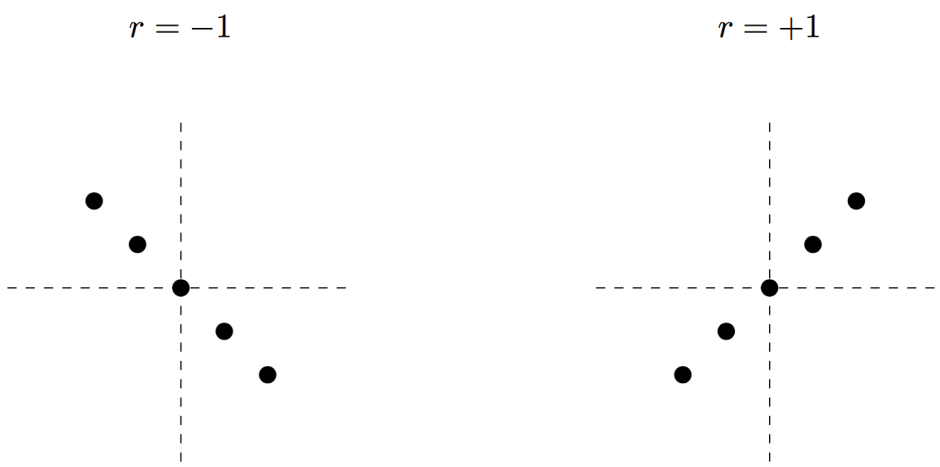
$$r = \frac{\text{Cov}[X, Y]}{\text{SD}[X]\text{SD}[Y]}$$

The correlation has all of the important properties of covariance:

- A positive correlation indicates a positive relationship between the variables. As one increases, so does the other.
- A negative correlation indicates a negative relationship between the variables. As one increases, the other tends to decrease.
- A zero correlation means that the two variables do not move in a consistent direction, but does not necessarily mean that they have *no* relationship.

But the correlation is also guaranteed to be between -1 and 1, so it can be compared across data sets.

What does a maximal correlation of ± 1 mean? It means that the data fall perfectly along a line.



Correlation is calculated in `pandas` in much the same way that covariance is, using the `.corr()` method:

```
In [4]: housing_df["Gr Liv Area"].corr(housing_df["SalePrice"])
```

```
Out[4]: 0.7067799209766279
```

Like the covariance, the correlation r is positive, but it is a number between -1 and $+1$. $r = +1$ would mean that all of the points on the scatterplot fell perfectly along a line (with positive slope). Although the points in the scatterplot do not all fall perfectly on a line, they do seem to hover around an underlying line. This explains why the covariance is close to, but not equal to, 1.

2 Exercises

Exercise 1. What is the correlation between any variable and itself? Check your answer with any (quantitative) variable from the Ames housing data set.

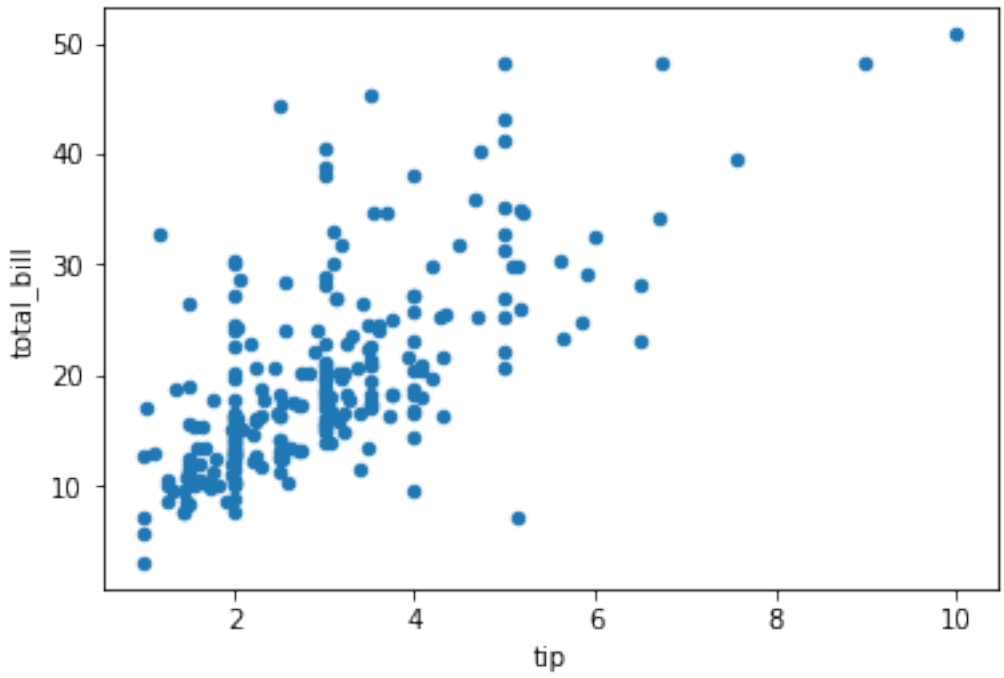
```
In [10]: (housing_df["SalePrice"].corr(housing_df["SalePrice"]),
          housing_df["SalePrice"].cov(housing_df["SalePrice"]), housing_df.SalePrice.var())
```

```
Out[10]: (1.0, 6381883615.6884375, 6381883615.6884365)
```

Exercises 2-3 deal with the Tips data set (`https://raw.githubusercontent.com/dlsun/data-science-book/`
Exercise 2. Make a scatterplot showing the relationship between the tip and the total bill.

```
In [6]: tips = pd.read_csv("https://raw.githubusercontent.com/dlsun/data-science-book/master/da
        tips.plot.scatter("tip", "total_bill")
```

```
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x7ff6322aeda0>
```



Exercise 3. Which pair of variables in this data set have the highest correlation with each other?

```
In [14]: tips.head()
         (tips["total_bill"].corr(tips["tip"]), tips["total_bill"].corr(tips["size"]), tips["t

         tips.corr()
```

Out[14]:

	total_bill	tip	size
total_bill	1.000000	0.675734	0.598315
tip	0.675734	1.000000	0.489299
size	0.598315	0.489299	1.000000

Exercise 4. To build your intuition about correlation, play this correlation guessing game. There is even a two-player mode that allows you to play against a friend in the class.