2. Взаимосвязь данных

Диаграмма рассеяния

График, где значения соединяются линиями, хорош, если иллюстрирует непрерывную связь. Если же в нашем распоряжении данные, которые никак не связаны друг с другом, гораздо лучше обозначить их точками. Это возможно на особом типе графиков — диаграмме рассеяния (scatter plot):

```
data.plot(x='column_x', y='column_y', kind='scatter')
```

Корреляция

Очевидный недостаток диаграммы рассеяния в том, что местами может оказаться огромное количество точек, слившихся в единую массу. В облаке значений не разглядеть области более высокой плотности. Есть два способа сделать график нагляднее:

Сделать точки полупрозрачными, задав параметр alpha и подобрать его оптимальное значение

Построить график, поделённый на шестиугольные области

График делят на ячейки; пересчитывают точки в каждой ячейке. Затем ячейки заливают

цветом: чем больше точек — тем цвет гуще. Такой график называется hexbin -

графиком, разделённым на шестиугольные области. Число ячеек по горизонтальной

оси задают параметром gridsize, аналогом bins для hist() .

```
data.plot(x='column_x', y='column_y', kind='hexbin',
gridsize=our_gridsize, sharex=False, grid=True)
```

Смысл этого графика, как и у гистограммы — отображение частотности. Но на гистограмме показана только одна величина, а здесь две. Повышенная частота определённых сочетаний указывает на закономерность. Часто цель анализа данных в том и состоит, чтобы показать связь двух величин.

Взаимозависимость двух величин называется взаимная корреляция. График позволяет утверждать, что две величины явно взаимосвязаны, или взаимно коррелируют. В том случае, если существует прямая зависимость величин (чем больше одна, тем больше другая), то говорят, что корреляция положительная. В том случае, если существует обратная зависимость величин (чем больше одна, тем меньше другая), то говорят, что корреляция отрицательная.

В предположении, что целевой показатель связан с другими показателями линейно, взаимосвязь оценивается с помощью коэффициента корреляции Пирсона.

Он помогает определить, как сильно меняется одна величина при изменении другой; и принимает значения от - 1 до 1. Если с ростом первой величины, растёт вторая, то коэффициент корреляции Пирсона — положительный. Если при изменении одной величины другая остаётся прежней, то коэффициент равен 0. Если рост одной величины связан с уменьшением другой, коэффициент отрицательный. Чем ближе коэффициент корреляции Пирсона к крайним значениям: 1 или 1, тем сильнее взаимозависимость. Если значение близко к нулю, значит связь слабая, либо отсутствует вовсе. Бывает, что коэффициент нулевой не оттого, что связи между значениями нет, а потому что у неё более сложный, не линейный характер. Потому-то коэффициент корреляции такую связь не берёт.

Коэффициент Пирсона находят методом corr() . Метод применяют к столбцу с первой

величиной, а столбец со второй передают в параметре. Какая первая, а какая — неважно:

```
print(data['column_1'].corr(data['column_2']))
print(data['column_2'].corr(data['column_1']))
```

Совместное распределение для множества величин

Когда в задаче нужно найти попарные взаимосвязи величин, это можно сделать с помощью попарных диаграмм рассеяния. В Pandas такую задачу решают не data.plot() , а специальным методом: pd.plotting.scatter_matrix(data): pd.plotting.scatter_matrix(data)

Помимо попарных диаграмм рассеяния, можно получить попарный коэффициент корреляции для всех величин. Это можно сделать с помощью матрицы корреляции:

```
data.corr()
```

Необходимо для выданного датасета
1) построить диаграмму рассеяния
2) оценить совместное распределение для множества величин
3) реализовать самостоятельно алгоритм линейной регрессии
4) оценить ее точность на основе f-меры, сравнить с программной реализацией модели на Phyton.