Анализ временных рядов

Порядок

В общем:

- наблюдения (data points) не упорядочены
- чтобы предсказать новое наблюдение, может потребоваться учесть все наблюдения

У **временных рядов**:

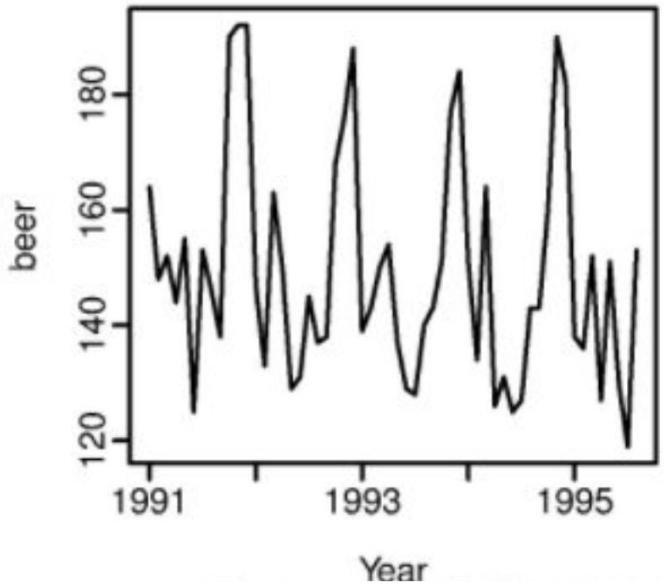
- наблюдения упорядочены хронологически
- для предсказания нового наблюдения: более новые (недавние) наблюдения более релевантны, чем более старые наблюдения

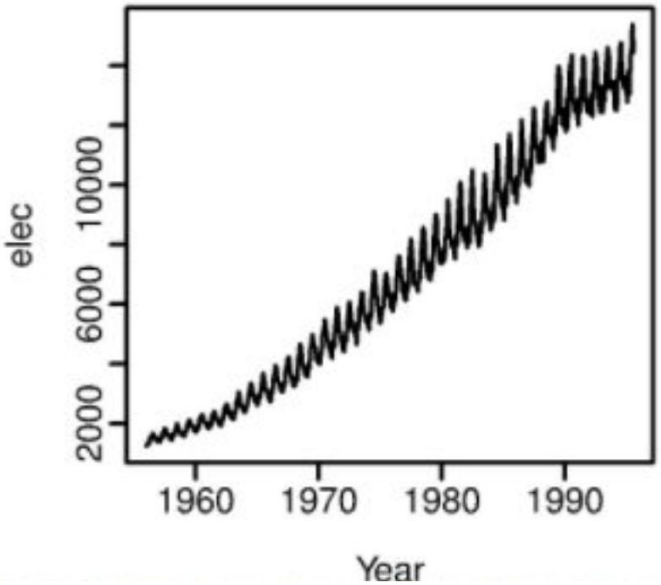
Закономерности во временных рядах

Зачастую временной ряд может быть разложен на следующие элементы:

- **Тренд/тенденция**: долговременное увеличение/уменьшение некоторых значений в данных
- **Сезонность**: циклическая закономерность в данных (например: по дням недели, по кварталам года)
- **Шум**: недетерминированный элемент в данных

Сезонность или Тренд





Resampling

Resampling (передискретизация) изменяет **частоту** наблюдений во временном ряду.

Upsampling: (повышающийся дискретизация) увеличение частоты наблюдений, например, переход от дней к часам.

Downsampling: (понижающая дискретизация) уменьшение частоты наблюдений, например, от дней к неделям.

`pandas.DataFrame.resample`

Resampling

Resampling используется, если существующие наблюдения были собраны с неправильной частотой:

- они или слишком редкие или слишком частые

Upsampling: (повышение частоты) обычно требует большей аккуратности, так как мы по сути *гадаем*, каким может быть недостающее промежуточное наблюдение, используя известные соседние наблюдения.

Например: имеются наблюдения, выполненные с частотой один раз в час, а необходимы такие же данные, но собранные с частотой один раз в 30 минут. Недостающие промежуточные данные можно сконструировать, взяв в качестве значения среднее арифметическое двух соседних наблюдений, одного слева и одного справа (*линейная интерполяция*).

`pandas.DataFrame.resample`

Resampling

Resampling используется, если существующие наблюдения были собраны с неправильной частотой:

- они или слишком редкие или слишком частые

Downsampling: (понижение частоты) обычно надежнее, чем upsampling, потому что оно агрегирует (объединяет) данные, тем самым уменьшая их гранулярность (степень детализации).

Например, имеются наблюдения, выполненные с частотой один раз в час, а необходимы наблюдения, выполненные с частотой один раз в день. Мы можем вычислить среднее арифметическое почасовых наблюдений и использовать его в качестве дневного значения.

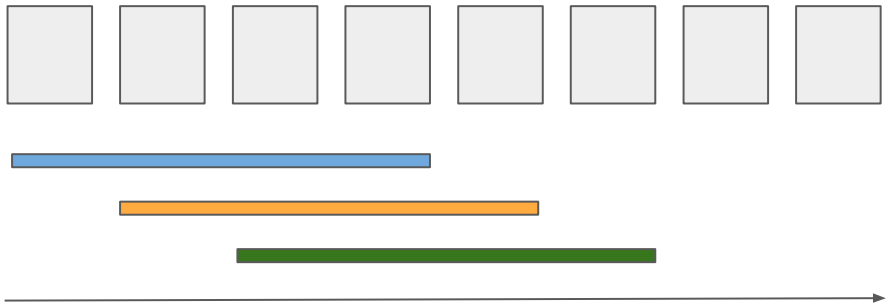
`pandas.DataFrame.resample`

Скользящее окно

Отличие временных рядов: **не все данные одинаковы!**

“Скользящее окно” – некоторая функция применяется к фрагментам данных (slice) фиксированного размера. “Окно” движется по данным.

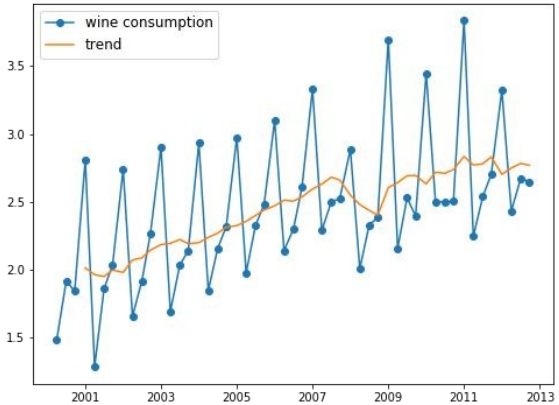
`pandas.DataFrame.rolling`



Окно размера 4

Скользящее среднее

- Лучше выявляет **тренд** в данных.
- Создает новый временной ряд, в котором отдельные наблюдения есть результат усреднения наблюдений в рамках некоторого окна фиксированного размера.
- Окно перемещается вдоль временного ряда.
- Аналогично можно вычислить скользящее стандартное отклонение, скользящую медиану и т.д.

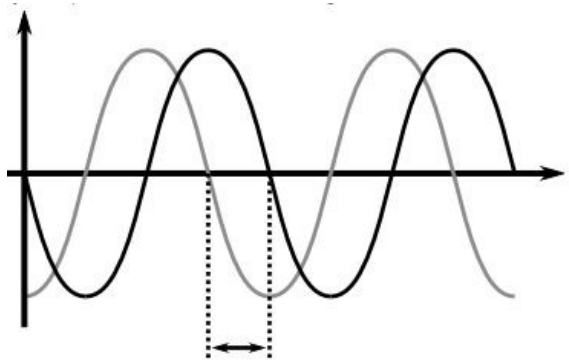


Сдвиг временного ряда

Иногда необходимо сдвинуть весь временной ряд целиком. Цели:

- избавиться от некоторого известного запаздывания (latency)
- легче выявить причинно-следственные отношения
- создать lag features

Index	Value(t)	Value(t-1)
t	50	40



`pandas.DataFrame.shift`

Вычисление разностей (Differencing)

Это один из способов преобразования временного ряда.

Новое наблюдение получается вычитанием предыдущего наблюдения из текущего. Например, если задан сдвиг (lag) в 1 единицу времени:

```
difference(t) = observation(t) - observation(t-1)

  где: observation(t)   – наблюдение в момент t
       observation(t-1) – наблюдение в предыдущий момент t-1
```

Использование разностей позволяет удалить из данных **тренд** (lag=1) и **сезонность** (lag=m) и тем самым выделить **шум**.

```
pandas.DataFrame.diff
```

Автокорреляция

Автокорреляция это степень сходства временного ряда с самим собой, взятым с некоторым временным сдвигом.

Например, если взять
наблюдения в диапазоне [1:10],
наблюдения в диапазоне [5:15]
и сравнить их между собой. Насколько похожи значения этих двух выборок?

Позволяет выявить **сезонность** во временном ряду.

```
pandas.DataFrame.autocorr
```

Нужно запомнить

- Мы часто рассматриваем временной ряд как состоящий из **тренда**, **сезонности** и некоторого **шума**.

- В зависимости от решаемой задачи и целей, бывает полезно произвести передискретизацию (**resampling**) данных. Upsampling требует большей аккуратности, чем downsampling.

- **Не все наблюдения созданы равными!** Мы можем выбросить некоторые прошлые наблюдения используя прием “**скользящее окно**”.

- **Вычисление разностей** полезно, когда надо удалить из данных тренд и сезонность.

Практика

time_series_data_analysis.ipynb