

# Содержание.

1. Что такое временные ряды
2. Классификация временных рядов
3. Основные понятия для анализа временных рядов
4. Базовые преобразования временных рядов
5. Необходимые статистические критерии
6. STL - разложение

# Что есть временной ряд?

По критерию размерности ряды могут быть одномерными и многомерными. Одномерные ряды представляют зависимость одной метрики (параметра) от времени, т.е. следующая конечная последовательность чисел:

$$ts = (t_i, v_i), t_i < t_j, \forall i < j,$$

где  $t$  - время,  $v$  - отслеживаемая метрика.

$$ts = (t_i, v1_i, v2_i, \dots, vn_i), t_i < t_j, \forall i < j$$

где  $t$  - время,  $v1..vn$  - отслеживаемые метрики.

# Что временным рядом не является

Во временных рядах значение признака измеряется через постоянные интервалы времени.

Если интервалы разные хотя бы для двух разных  $t$ , мы имеем дело со стохастическим процессом.

# Какие бывают временные ряды?

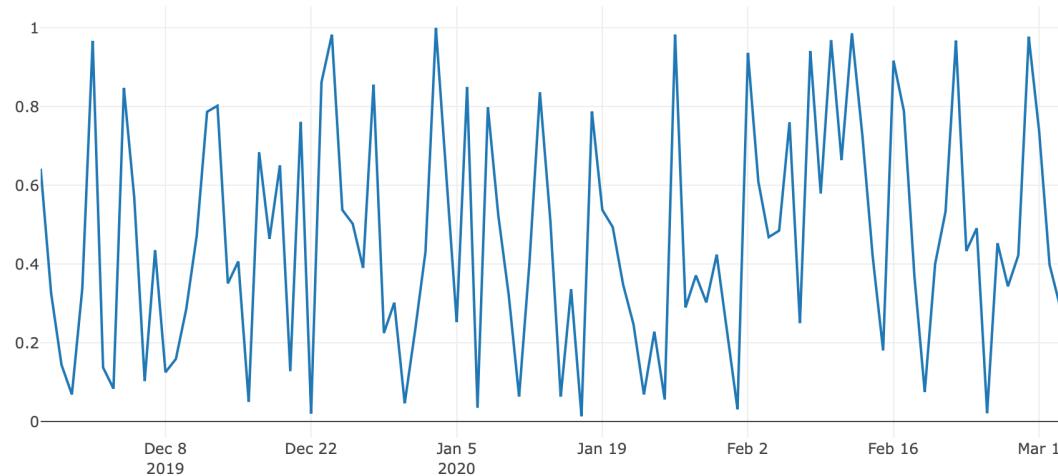
1. Стационарные
2. Интегрированные
3. Сезонные
4. Циклические
5. Детерминированные

Определять тип временного ряда важно для построения автоматической системы предсказания.

# Стационарный ряд

Самым простым типом временного ряда является так называемый стационарный временнной ряд. Стационарный временнной ряд это такой случайный процесс, дисперсия и матожидание которого не зависят от времени, а между соседними значениями нет значимой корреляции.

Ряд  $y_1, \dots, y_T$  **стационарен**, если  $\forall s$  распределение  $y_t, \dots, y_{t+s}$  не зависит от  $t$ , т. е. его свойства не зависят от времени.



# Интегрированный ряд

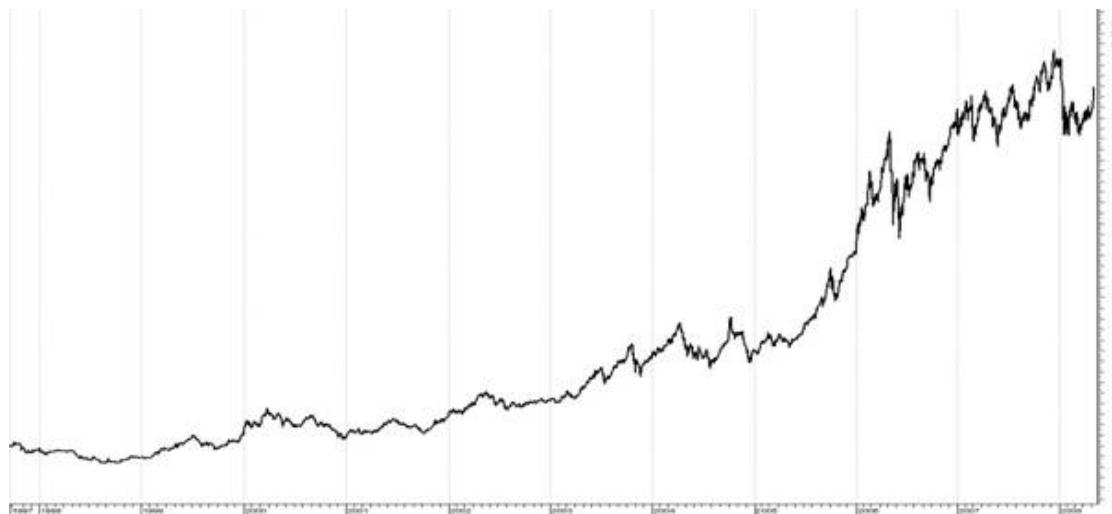
Следующий простой тип временных рядов называется интегрированным рядом. Интегрированный ряд это такой ряд, при дифференцировании которого получается стационарный ряд. Самым известным примером такого ряда является модель случайного блуждания.

$$y_{t+1} = y_t + e_t, e_t -$$

случайная компонента.

$$y(t) = at + b + e_t, e_t -$$

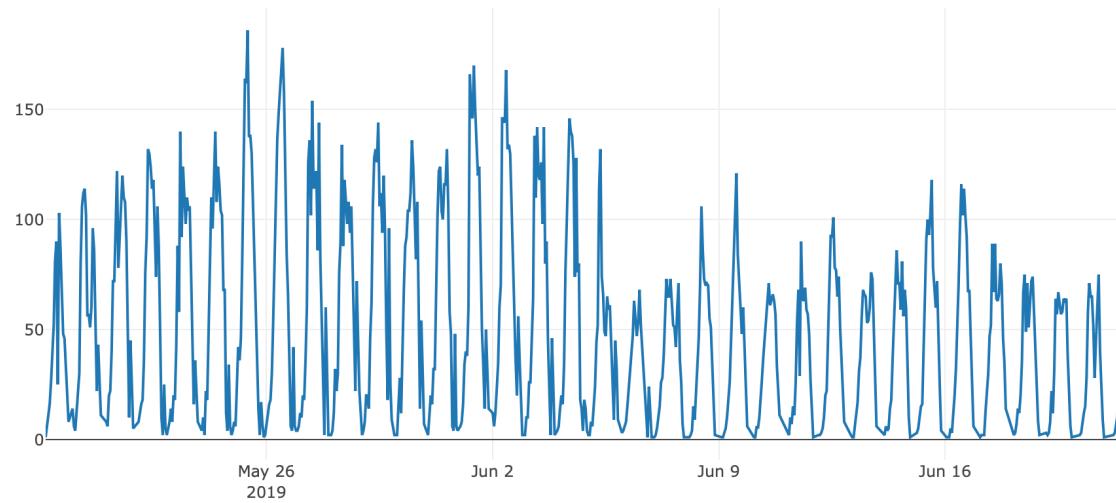
случайная компонента.



# Сезонный ряд

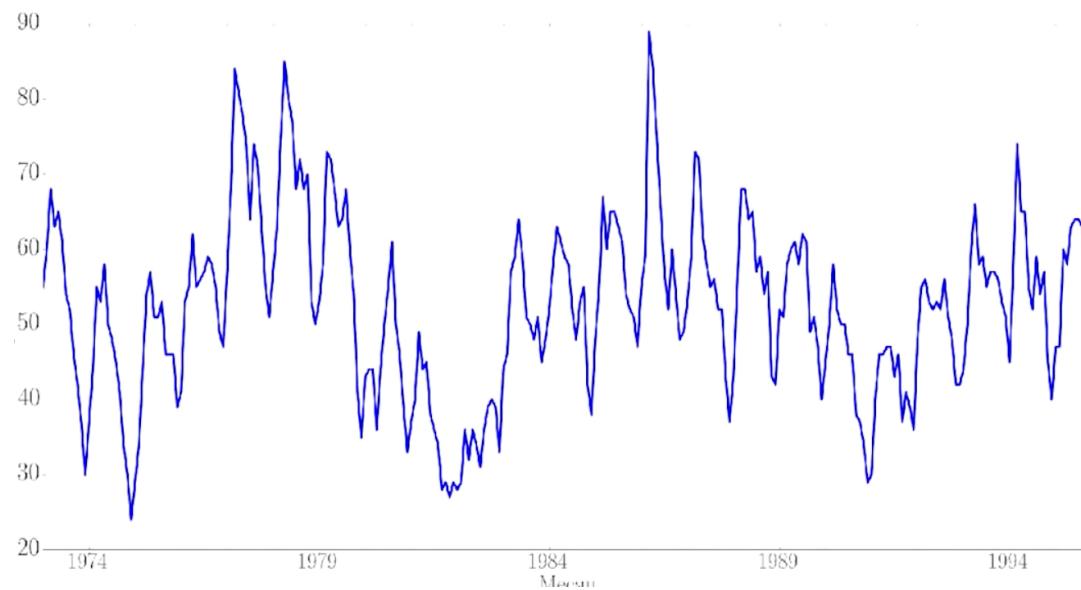
Теперь, если ряд не является ни стационарным, ни интегрированным, это значит, что данный ряд может быть классифицирован в зависимости от наличия или отсутствия у него сезонности - периодичного повторения паттерна через равные промежутки времени.

$$y_t = y_{t-k}, \forall t$$



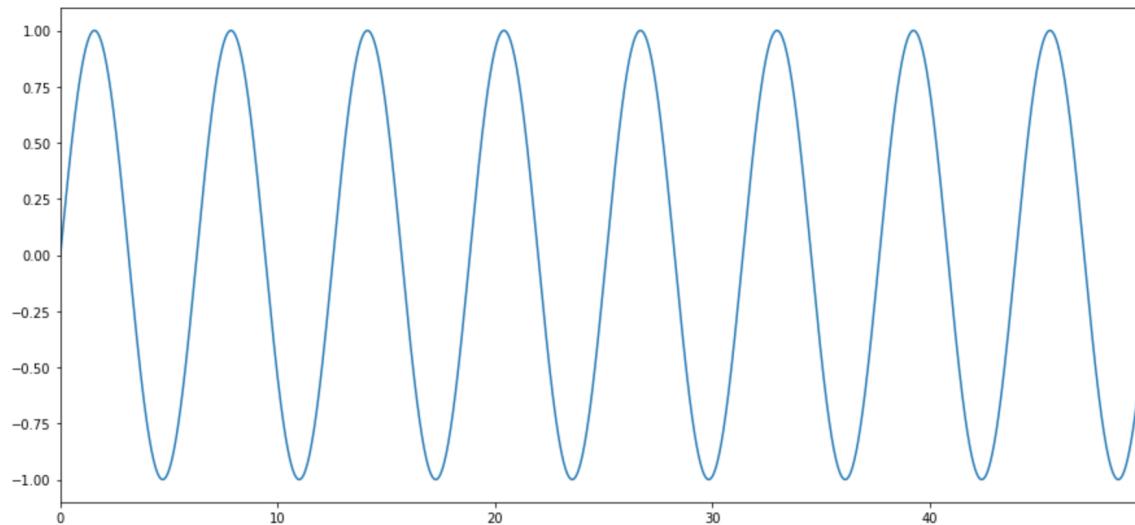
# Циклический ряд

Циклическость отличается от периодичности тем, что период циклическости в данном случае не является постоянной величиной.



# Детерминированный ряд

Детерминированный ряд - любой ряд, не содержащий случайной компоненты, который точно описывается некой аналитической функцией от времени.



# Компоненты ряда

Итак, каждый временной ряд может содержать в себе следующие компоненты.

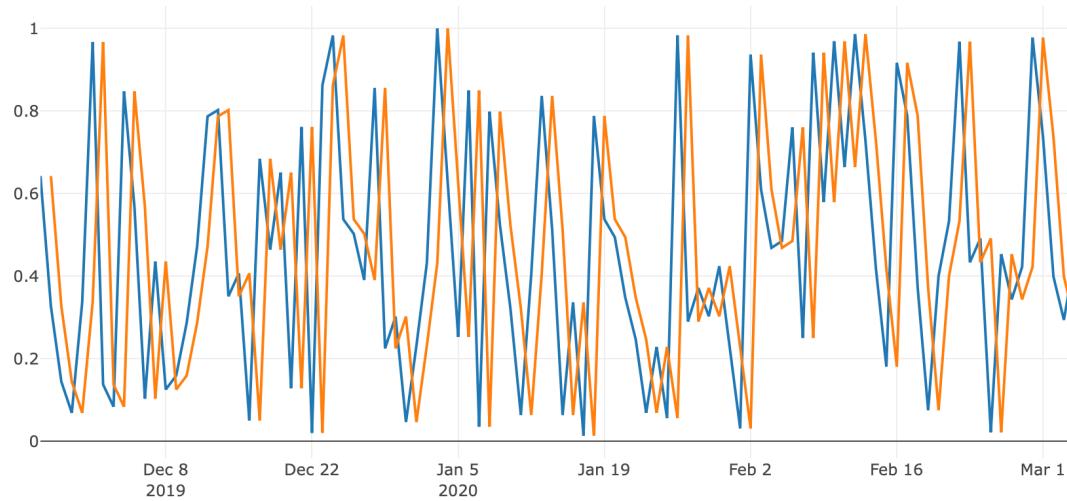
- Линейный тренд
- Сезонность
- Цикличность
- Ошибка

# Базовые понятия для анализа временных рядов

- Лаги временного ряда
- Период сезонности
- Автокорреляция
- Гетероскедастичность
- Стационарность

# Лаги временного ряда

Лагом  $k$  для точки  $t$  временного ряда  $Y$  называется значение данного ряда в точке  $t-k$ .



# Период сезонности

Ряд  $Y_t$  имеет период сезонности  $s$ , если  $\forall t \quad Y_{t-s} = Y_t$

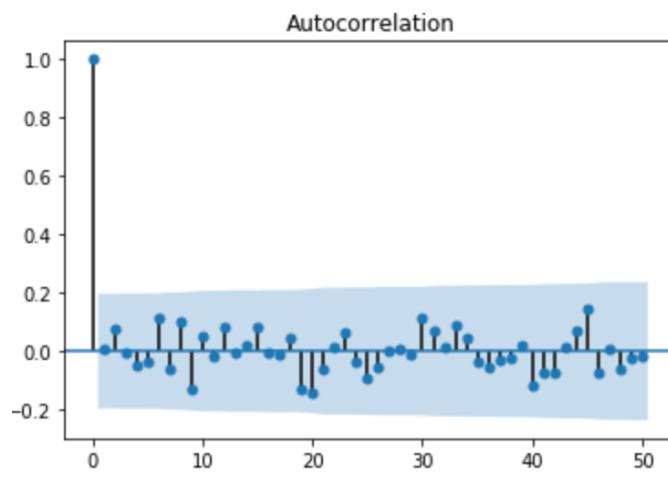
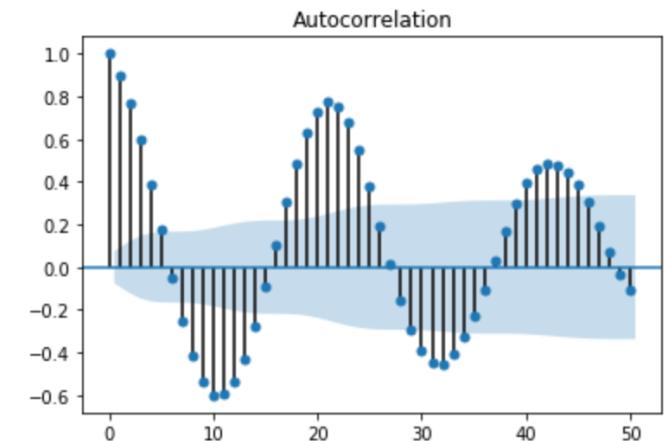
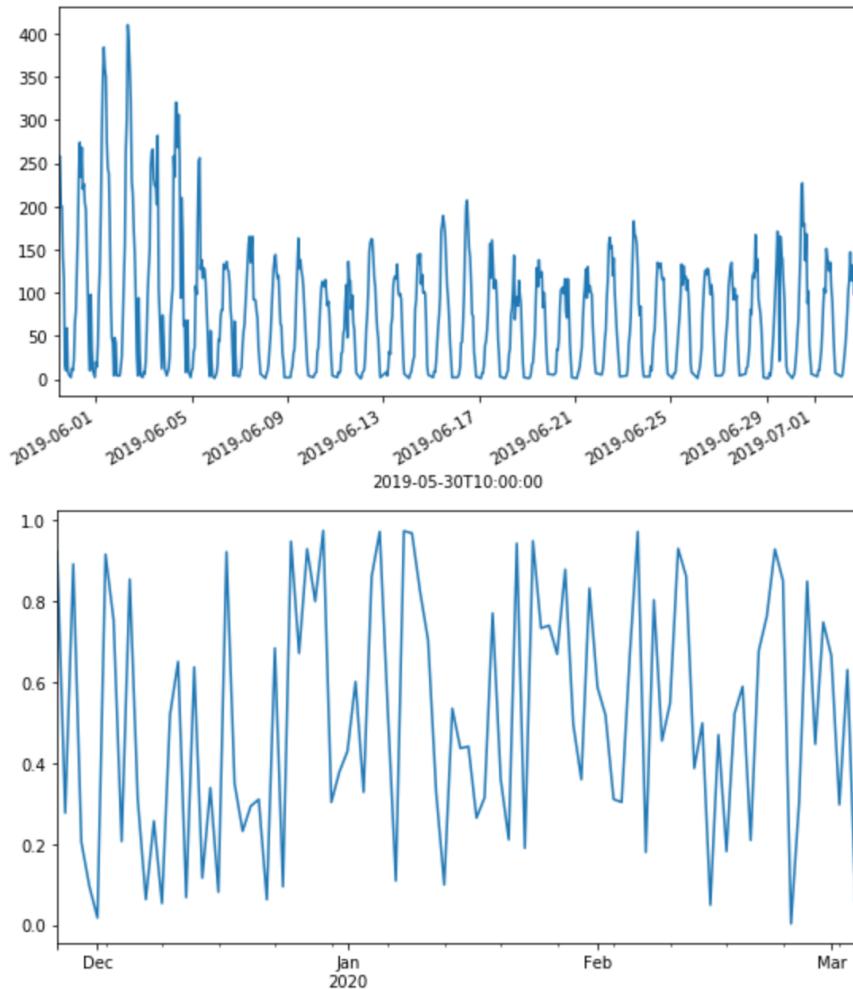
# Автокорреляция

Автокорреляция временного ряда - это классическая корреляция Пирсона, взятая относительно самого себя, сдвинутого на некий лаг  $k$ .

$$r_{XY} = \frac{\text{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}.$$

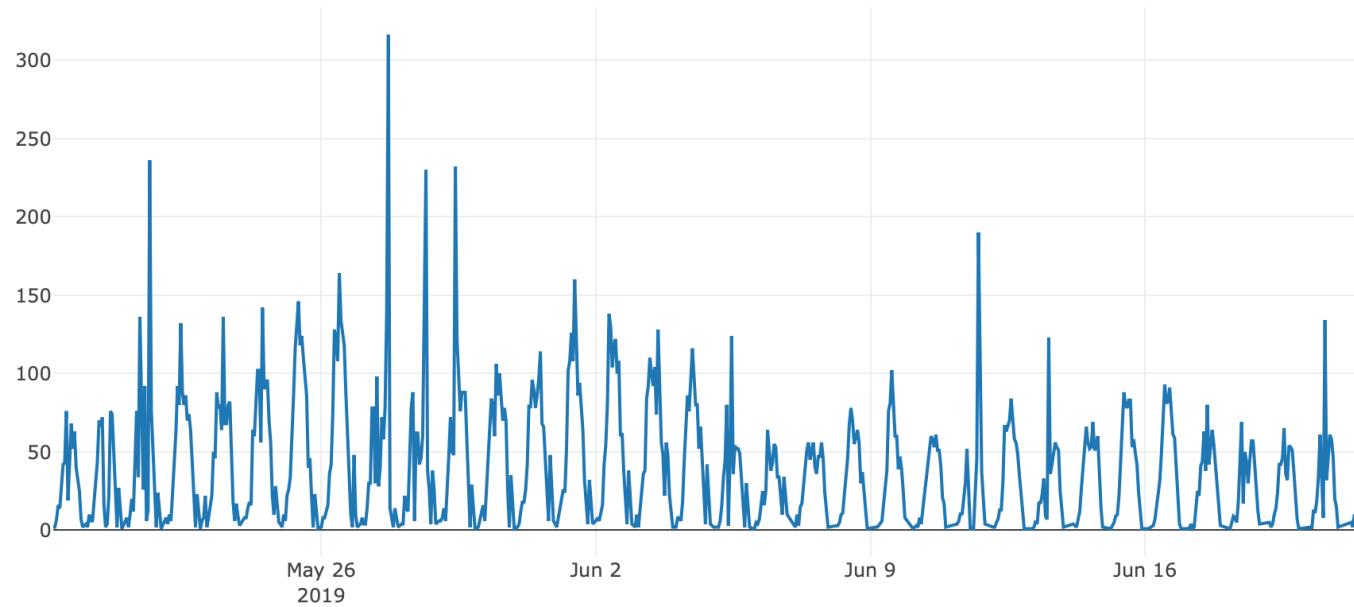
Здесь  $\mathbf{X}$  - исходный временной ряд,  $\mathbf{Y}$  - сдвинутый относительно самого себя.

# Коррелограммы



# Гетероскедастичность

Гетероскедастичность - непостоянство дисперсии.

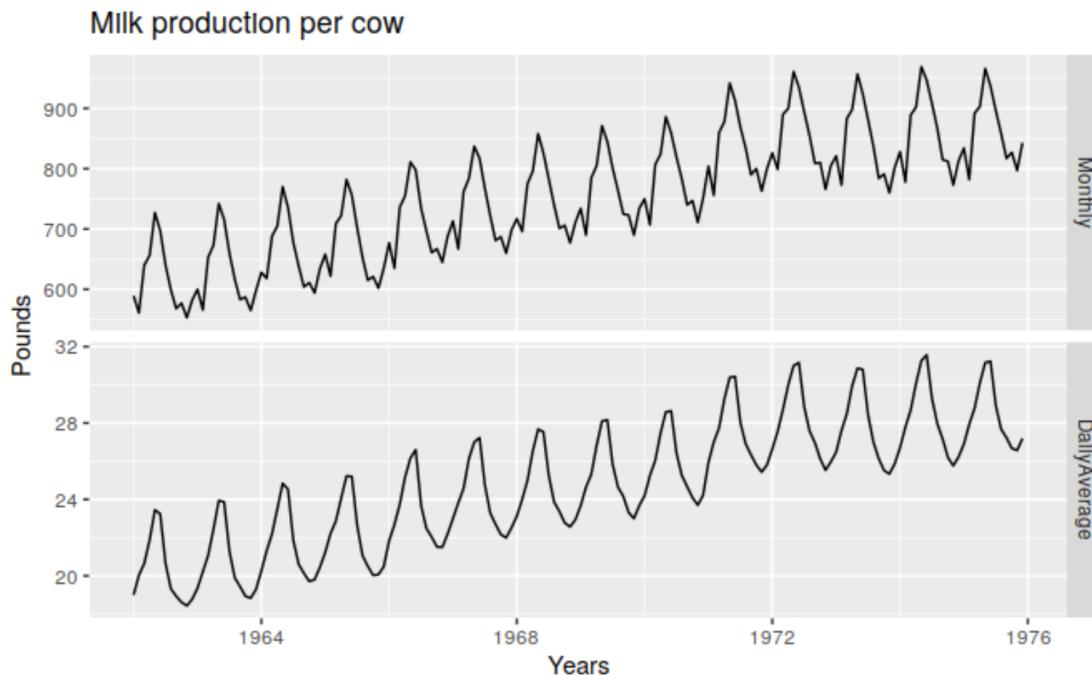


# Базовые преобразования временных рядов

- Временные преобразования
- Стабилизация дисперсии - логарифмирование, преобразование Бокса-Кокса.
- Дифференцирование ряда

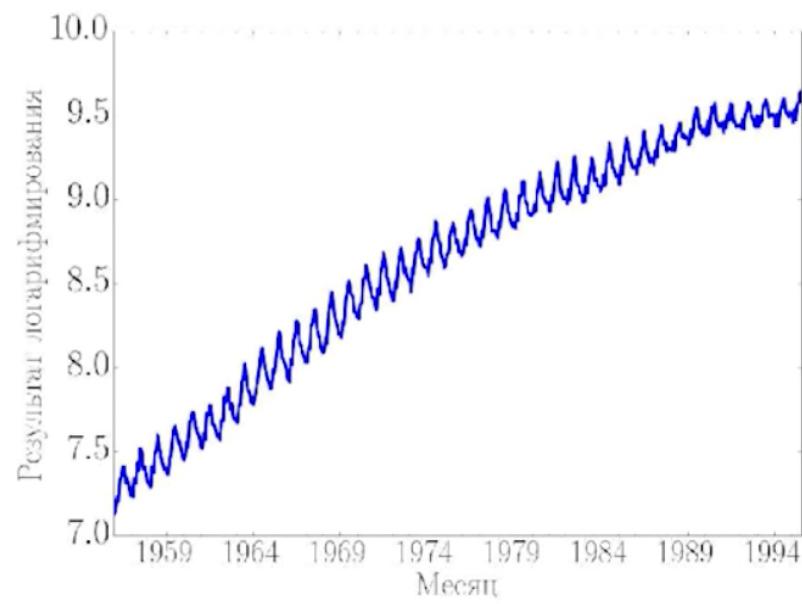
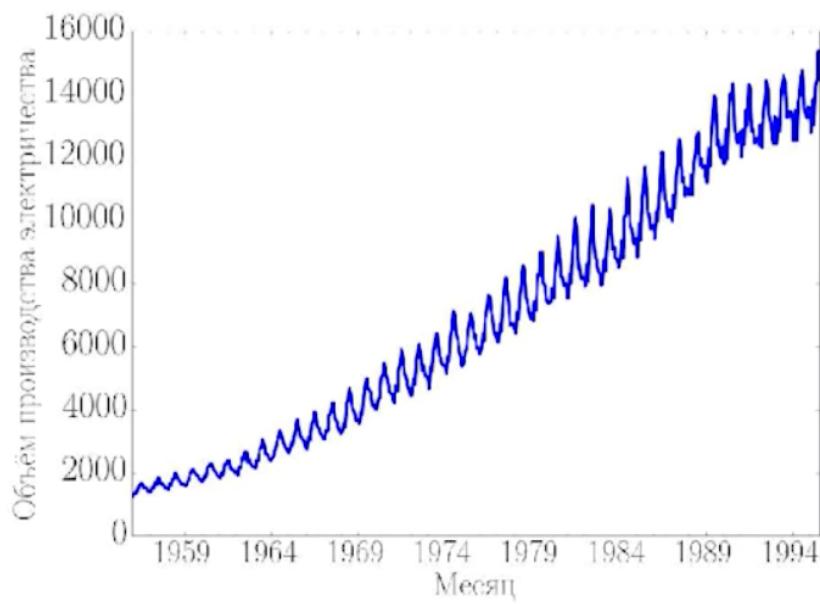
# Временные преобразования ряда

Зачастую очень помогает изменить “гранулярность” ряда, т.е. временные промежутки значений. Или взять определенный период ряда. На классическом примере ниже, переход от месячных измерений удоев поголовья скота к средним дневным, значительно упростил структуру временного ряда.



# Стабилизация дисперсии.

При непостоянной дисперсии первое, что стоит сделать - взять логарифм от значений ряда.



# Стабилизация дисперсии.

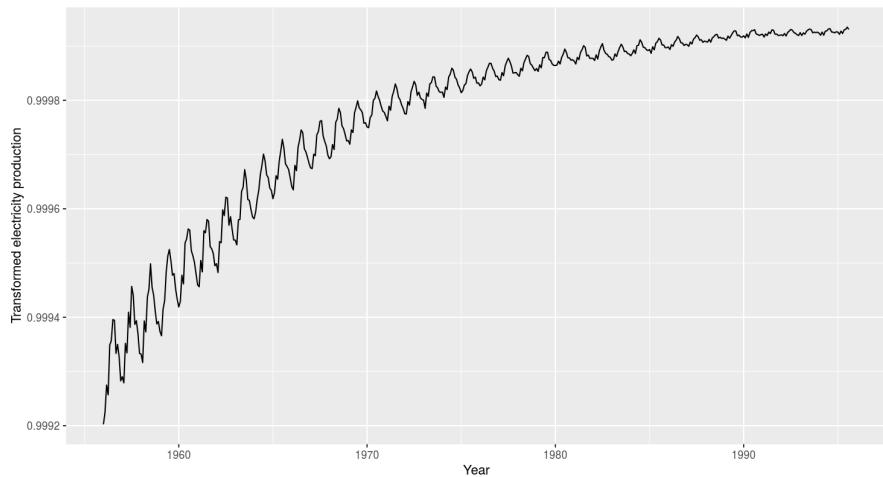
Развитие идеи логарифмирования - преобразование Бокса-Кокса.

$$y_t = \begin{cases} \exp(w_t) & \text{if } \lambda = 0; \\ (\lambda w_t + 1)^{1/\lambda} & \text{otherwise.} \end{cases}$$

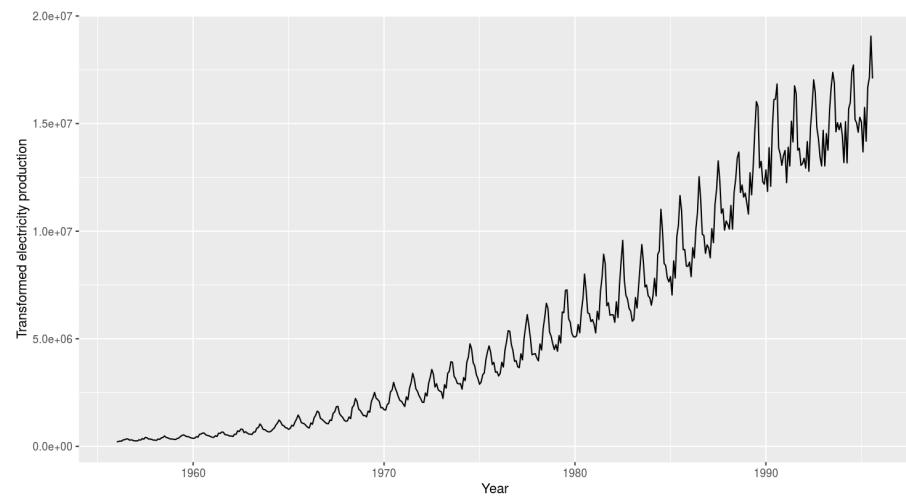
# Стабилизация дисперсии.

Развитие идеи логарифмирования - преобразование Бокса-Кокса.

$$\lambda = -1$$

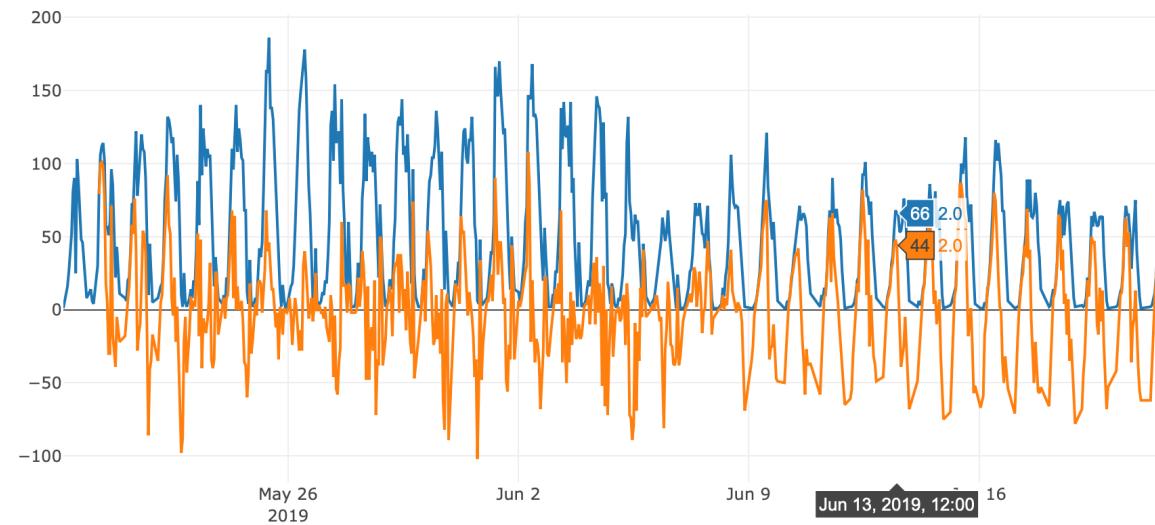


$$\lambda = 2$$



# Дифференцирование ряда

Дифференцирование порядка  $k$  - взятие разности  $Y_t^{diff} = Y_t - Y_{t-1}$



# Построение периодограммы временного ряда

Метод, основанный на разложении временного ряда при помощи преобразования Фурье. Используется для автоматического нахождения периода сезонности и может быть в данном контексте использован как альтернатива графику автокорреляций.

Преобразование Фурье предполагает разложение ряда на  $N$  гармонических компонент с частотами  $f_k$

$$X(f_{k/N}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn}$$

Теперь, мы можем построить периодограмму, т.е. зависимость плотности вероятности от частоты. Наше предположение будет заключаться в том, что частоты с наибольшей плотностью вероятности являются нашим периодом сезонности.

$$\mathcal{P}(f_{k/N}) = \|X(f_{k/N})\|^2 \quad k = 0, 1 \dots \lceil \frac{N-1}{2} \rceil \quad L_k = 1/f_k.$$

# Критерий Дики-Фуллера - на стационарность.

- временной ряд:  $y^T = y_1, \dots, y_T$ ;
- нулевая гипотеза:  $H_0$ : ряд нестационарен;
- альтернатива:  $H_1$ : ряд стационарен;
- статистика: неважно;
- нулевое распределение: табличное.

# Критерий Ланга-Бокса - на автокорреляцию.

$$\tilde{Q} = n(n+2) \sum_{k=1}^m \frac{\hat{\rho}_k^2}{n-k},$$

где  $n$  — число наблюдений,  $\hat{\rho}_k$  — автокорреляция  $k$ -го порядка, и  $m$  — число проверяемых лагов.

$H_0$ : данные являются случайными (то есть представляют собой [белый шум](#)).

$H_a$ : данные не являются случайными.

# STL разложение

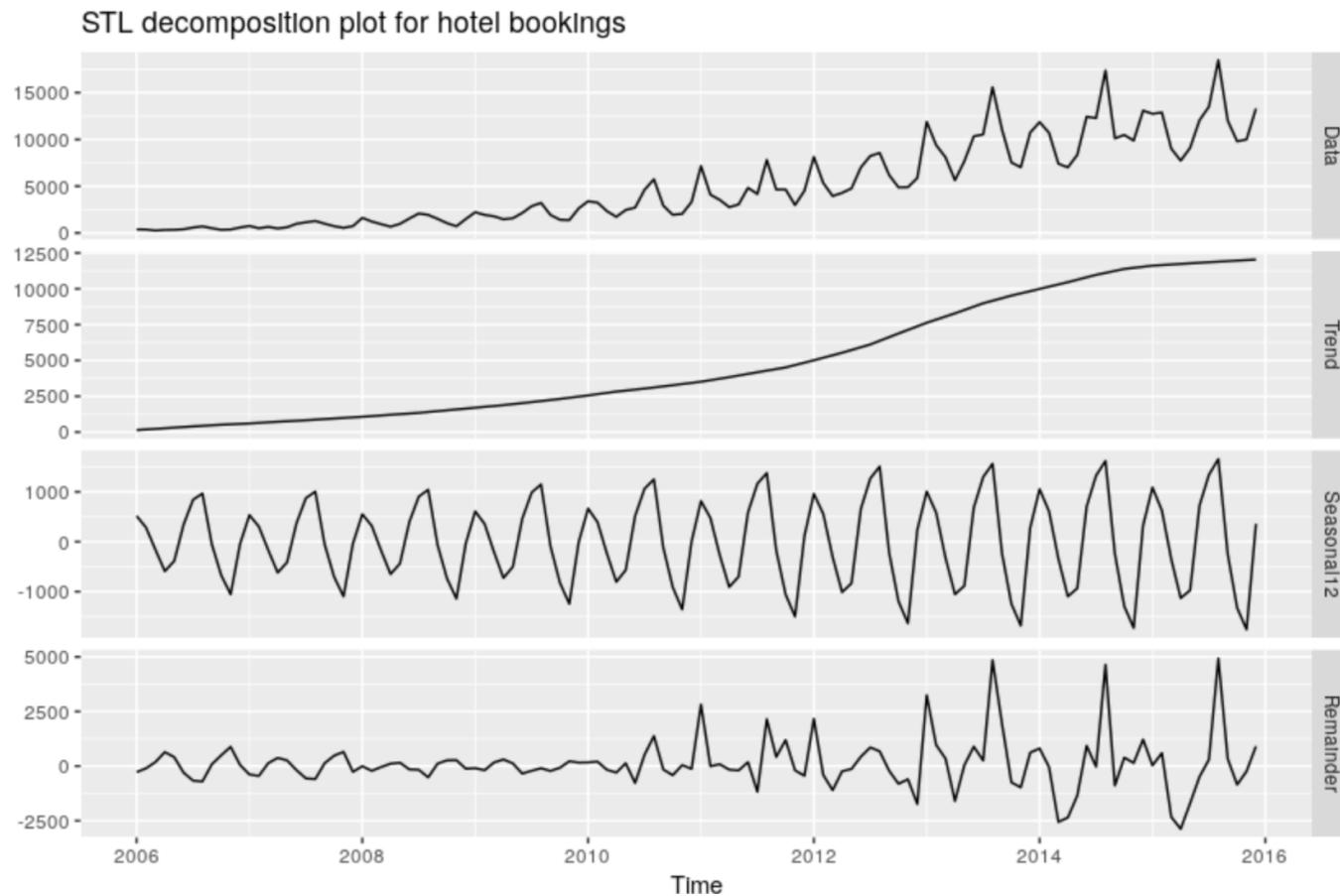
- Как мы уже выяснили, временные ряды состоят из нескольких компонент: сезонности, цикличности, тренда и случайного шума
- Одним из наиболее популярных методов разложения временного ряда на данные компоненты является STL - разложение
- Разложение используется для случаев, когда нам необходимо отделить сезонные эффекты от трендов
- Также данное разложение представляет собой один из наиболее простых и эффективных методов поиска аномалий

# Алгоритм работы

\* Примечание: для простоты реализации на семинаре, здесь приводится “наивная” реализация stl разложения - для рядов с около линейным трендом и использованием скользящей медианы в качестве фильтра. В реальном разложении используется достаточно сложный алгоритм из двух последовательных циклов сглаживания ряда при помощи метода локальных регрессий loess (подробнее можно найти по ссылке <http://www.nniiem.ru/file/news/2016/stl-statistical-model.pdf>)

1. Аппроксимировать линейный тренд при помощи метода МНК.
2. Вычесть его из временного ряда  $Y_t - T_t = S_t + R_t$ .
3. Определить период сезонности  $k$  временного ряда одним из удобных методов: коррелограмма, периодограмма.
4. Сгладить оставшийся ряд “сезонность+шум” с периодом  $k/2$ .  
(конкретное окно сглаживания подбирается эмпирически)
5. Вычесть получившуюся сезонную компоненту  
$$Y_{s+r} - S_s = S_t + R_t - S_t = R_t.$$
6. Остаток  $R_t$  - шум.
7. Проверить остаток на стационарность статистикой Дики-Фуллера.

# Пример STL разложения



# Ограничения STL разложения

- Применим только для рядов так называемого аддитивного типа, т.е. для тех, которые описываются следующей формулой:

$$Y_t = T_t + S_t + R_t$$

- Не поддерживает ряды с мультипликативной сезонностью
- Для рядов длиной не менее двух периодов сезонности