RocketScience.ai

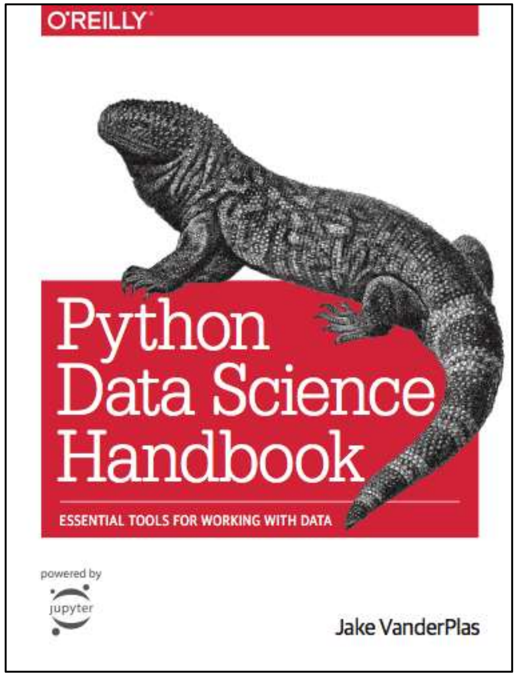
TENSOR.BY

ML-course

4. Forecasting for Time Series in Python

Kate Miniukovich (Data Scientist),
miniukovich@rocketscience.ai

Reference



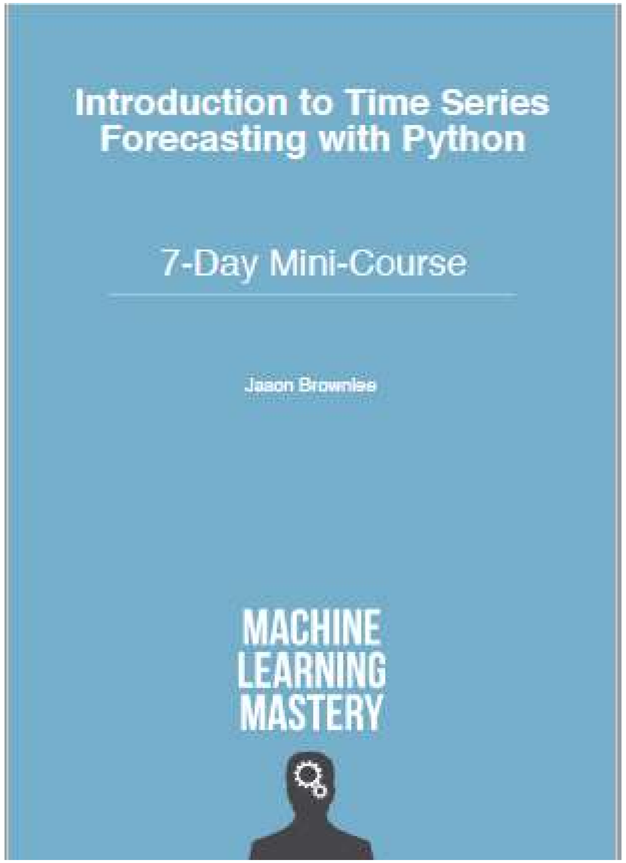
Jake VanderPlas

Python Data Science Handbook

https://jakevdp.github.io/PythonDataScienceHandbook/

03.11-Working-with-Time-Series.ipynb

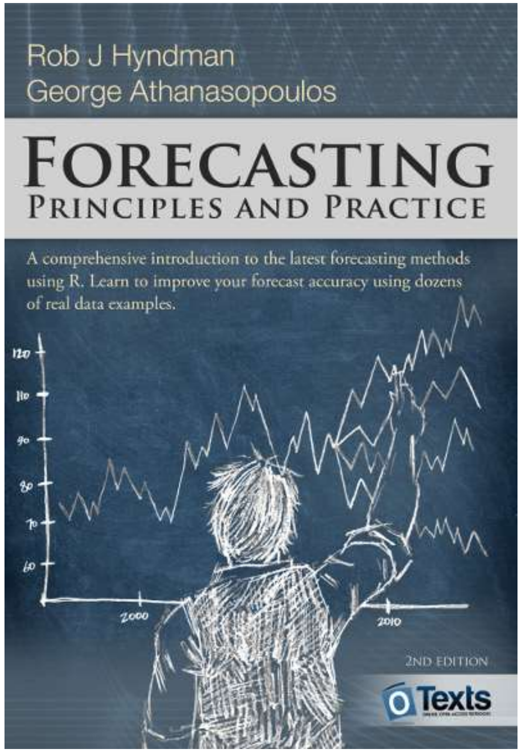
Reference



Jason Brownlee How to Create an ARIMA Model for Time Series Forecasting with Python

https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

Reference



Forecasting: Principles and Practice

Rob J Hyndman and George Athanasopoulos

Monash University, Australia

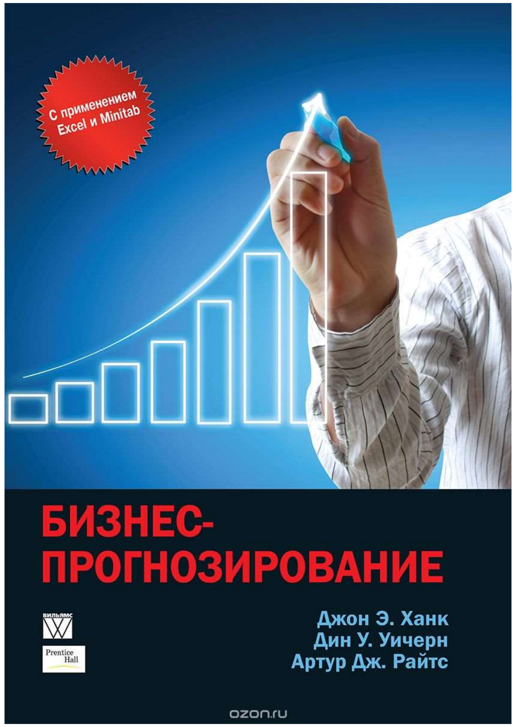
https://otexts.org/fpp2/

Reference

Бизнес-прогнозирование

7-е издание

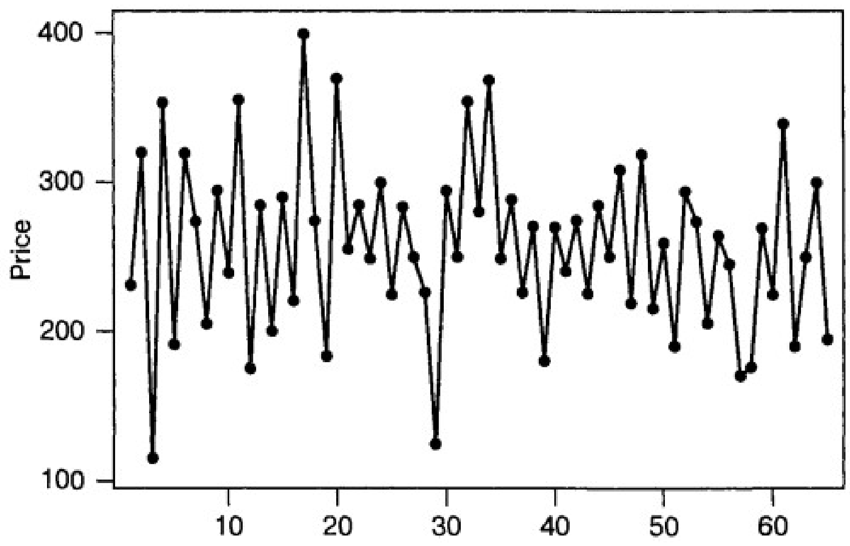
Джон Э. Ханк, Дин У. Уичерн, Артур Дж. Райтс



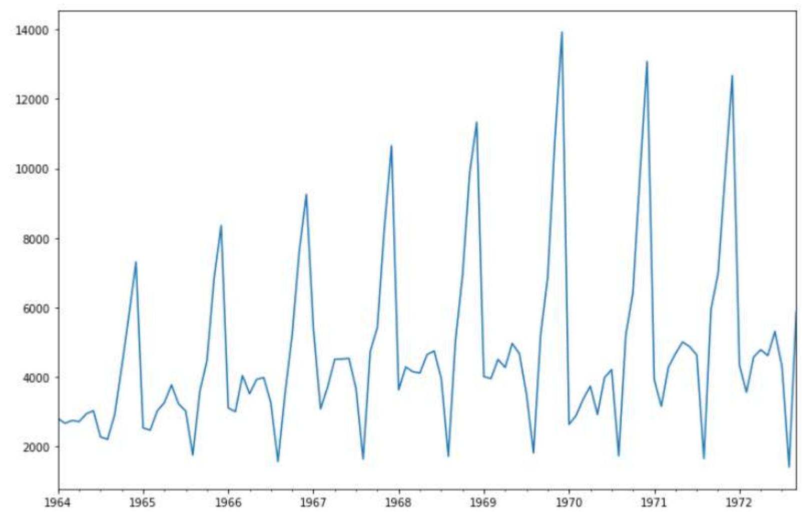
Definition

Time Series (TS) *- the value represented by a set of observations that were collected at successive intervals of time.*

Closing Prices for ISC Corporation Stock



Monthly sales of champagne Perrin Freres label from January 1964 to September 1972



Stationary and non stationary TS

Stationary TS - mean and variance don't change over time.

Non stationary TS

- **Trend**

A long-term increase or decrease in the data.

- **Seasonality**

There are periodic changes in the data, uniformly repeated from year to year.

- **Cyclicity**

There are rises and falls in the data, that do not have a fixed period.

Forecasting for TS

- **Naïve** methods *(stationary, trend, seasonality)*,
- **Box-Jenkins** methods *(stationary, trend, seasonality, cyclicity)*,
- **Other** methods.

Метод	Модель данных	Временная отдаленность	Тип модели	Минимальные требования к данным: Несезонные	Минимальные требования к данным: Сезонные
Наивный	СТ, Т, С	К	ВР	1	
Простые средние	СТ	К	ВР	30	
Скользящие средние	СТ	К	ВР	4–20	
Экспоненциальное сглаживание	СТ	К	ВР	2	
Линейное экспоненциальное сглаживание	Т	К	ВР	3	
Квадратичное экспоненциальное сглаживание	Т	К	ВР	4	
Сезонное экспоненциальное сглаживание	С	К	ВР		2×с
Адаптивная фильтрация	С	К	ВР		5×с
Простая регрессия	Т	С	К	10	
Множественная регрессия	Ц, С	С	К	10×В	
Классическое разложение	С	К	ВР		5×с
Экспоненциальные трендовые модели	Т	С, Д	ВР	10	
Подгонка S-кривой	Т	С, Д	ВР	10	
Модели Гомперца	Т	С, Д	ВР	10	
Возрастающие кривые	Т	С, Д	ВР	10	
"Перепись-II"	С	К	ВР		6×с
Модели Бокса-Дженкинса	СТ, Т, Ц, С	К	ВР	24	3×с

Модели данных: СТ — стационарные; Т — трендовые; С — сезонные; Ц — циклические.
Отдаленность прогноза во времени: К — краткий период (менее трех месяцев); С — средний период; L — большой период.
Тип модели: ВР — временной ряд; К — каузальная.
Сезонные: с — продолжительность сезонности.
Величина: В — количество величин.

Джон Э. Ханк «Бизнес-прогнозирование» (с.108)

Denotes

$Y_1, \dots, Y_t$ - *real data in time 1,…, t;*

$\bar{Y}$ - *mean of* $Y_1, \dots, Y_t$*;*

$\hat{Y}_{t+1}$ - *forecast in time t+1*

Naïve models

Simple naive model

$$\hat{Y}_{t+1} = Y_t$$

Naive model with trend

$$\hat{Y}_{t+1} = Y_t + (Y_t - Y_{t-1})$$

Naive model with quarterly seasonality

$$\hat{Y}_{t+1} = Y_{t-3}$$

Naive model with quarterly seasonality and trend

$$\hat{Y}_{t+1} = Y_{t-3} + \frac{(Y_t - Y_{t-1}) + \cdots + (Y_{t-3} - Y_{t-4})}{4}$$

Box-Jenkins methods

Autocorrelation ???

Autocorrelation

Autocorrelation - linear relationship between a value and its lag in one or more time periods.

The autocorrelation is measured using the **autocorrelation coefficient.**

Autocorrelation coefficient with a delay of k moments

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

Example

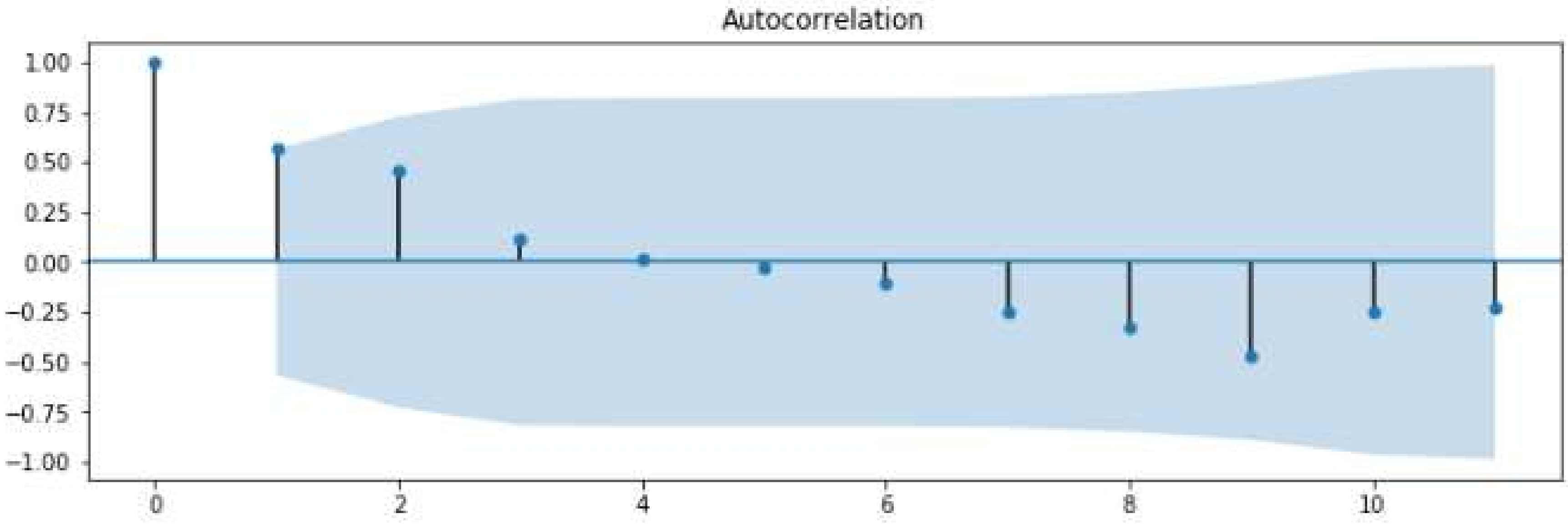
$$r_1 = \frac{\sum_{t=2}^{n}(Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

Время, t	Y_t	Y_{t-1}	$(Y_t - \bar{Y})$	$(Y_{t-1} - \bar{Y})$	$(Y_t - \bar{Y})^2$	$(Y_t - \bar{Y})(Y_{t-1} - \bar{Y})$
1	123	_	−19	_	361	_
2	130	123	−12	−19	144	228
3	125	130	−17	−12	289	204
4	138	125	−4	−17	16	68
5	145	138	3	−4	9	−12
6	142	145	0	3	0	0
7	141	142	−1	0	1	0
8	146	141	4	−1	15	−4
9	147	146	5	4	25	20
10	157	147	15	5	225	75
11	150	157	8	15	64	120
12	160	150	18	8	324	144
Сумма	1 704		0		1 474	843

$$\bar{Y} = \frac{1704}{12} = 142$$

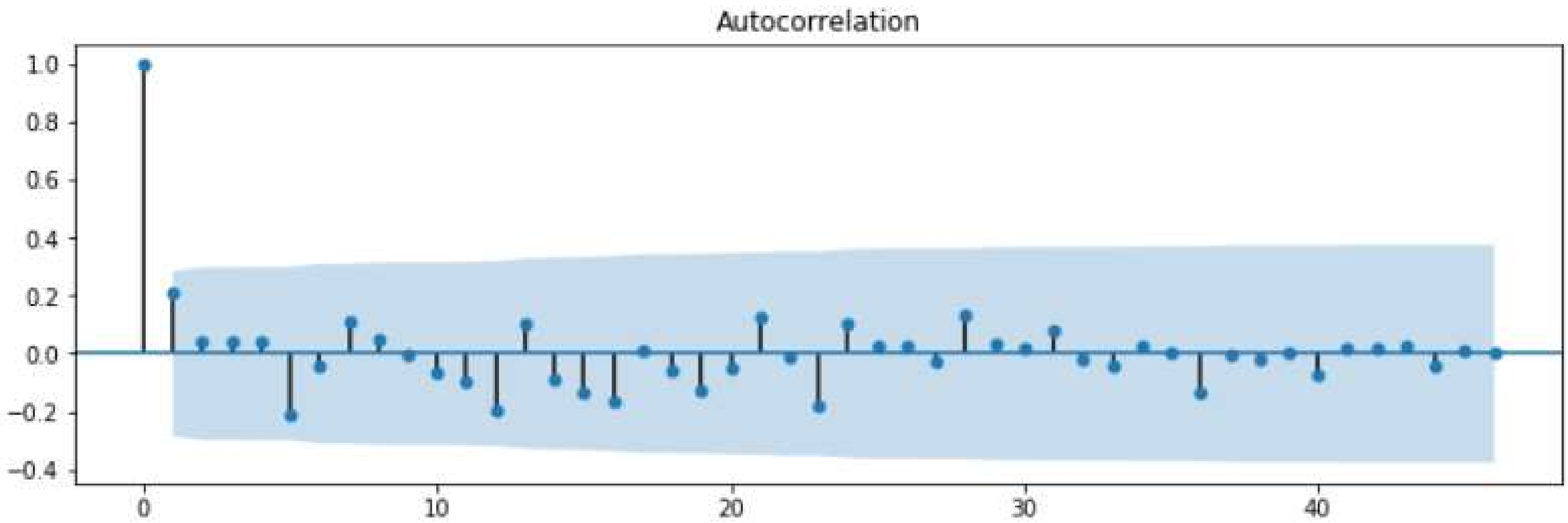
$$r_1 = \frac{843}{1474} = 0,572$$

Autocorrelation plot



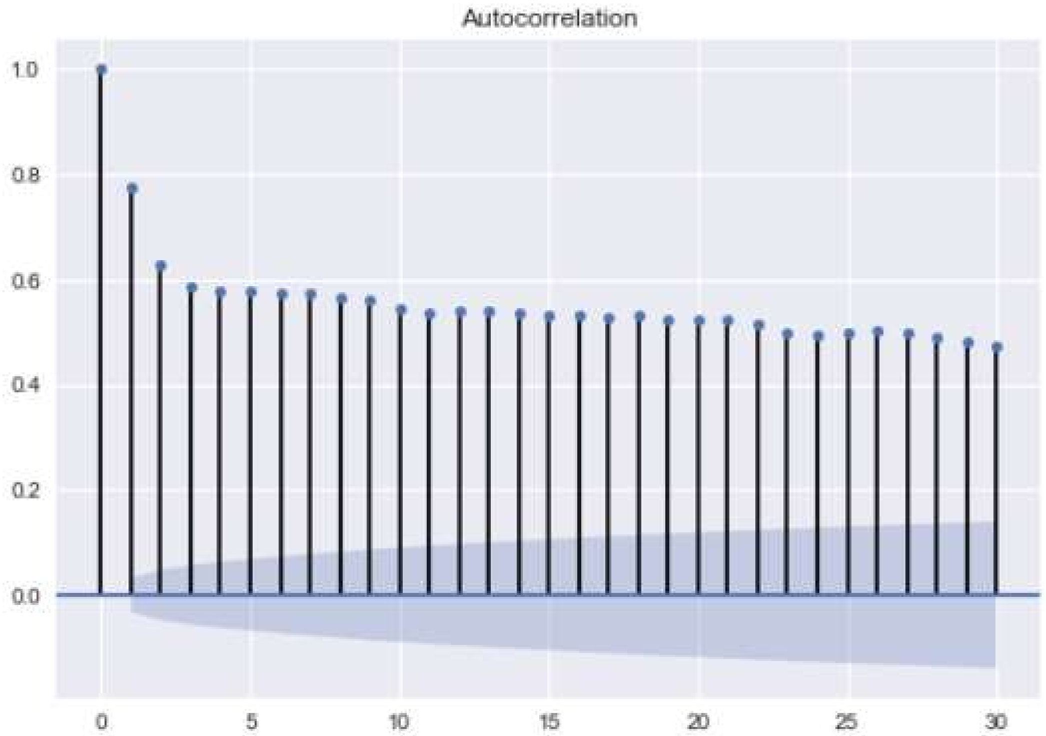
Autocorrelation plot analysis

1. If the ***autocorrelation coefficients for any lag k are close to zero***, then there is **no autocorrelation**, i.e. TS is random.



Autocorrelation plot analysis

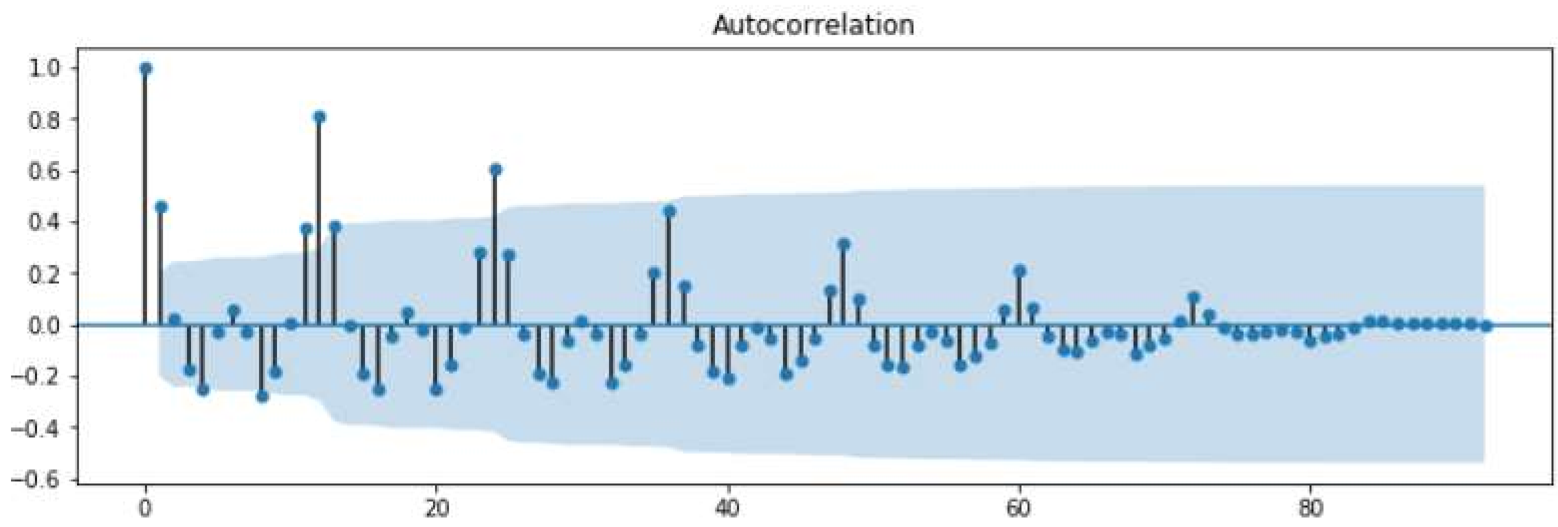
2. If the ***autocorrelation coefficients for the first few periods of delay are significantly different from zero***, and with the increase of the period gradually decrease to zero, then TS has a **trend**.



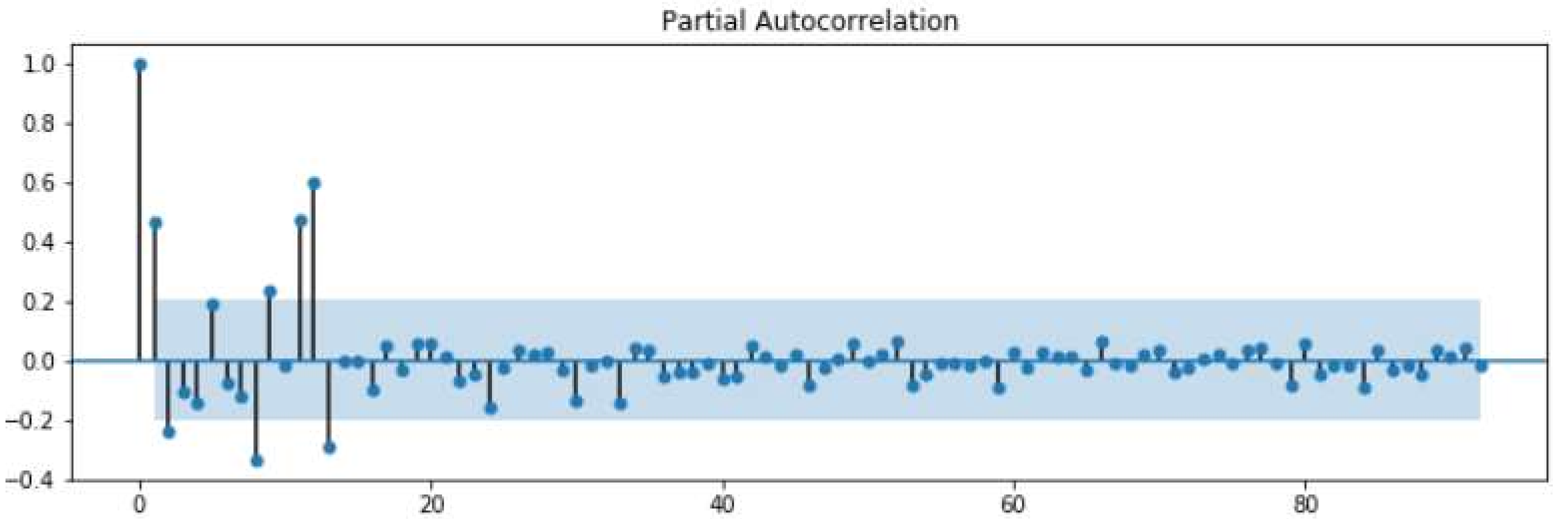
Autocorrelation plot analysis

3. If a ***significant coefficient of autocorrelation is observed for periods of lag equal to the seasonal period*** or multiples of it, then the series has **seasonality**.

The seasonal lag period is 4 for quarterly data and 12 for monthly.



Partial Autocorrelation plot



Box-Jenkins methods

For stationary TS

- AutoRegressive model of the order *p*, ***AR(p)***
- Moving Average model of the order *q*, ***MA(q)***
- Models with AutoRegression and Moving Average, ***ARMA(p, q)***

For stationary and non-stationary TS

- AutoRegressive Integrated Moving Average, ***ARIMA(p,d,q)*** - *for stationary TS d=0*

AutoRegressive model of the order *p*, AR(p)

$$\hat{Y}_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p}$$

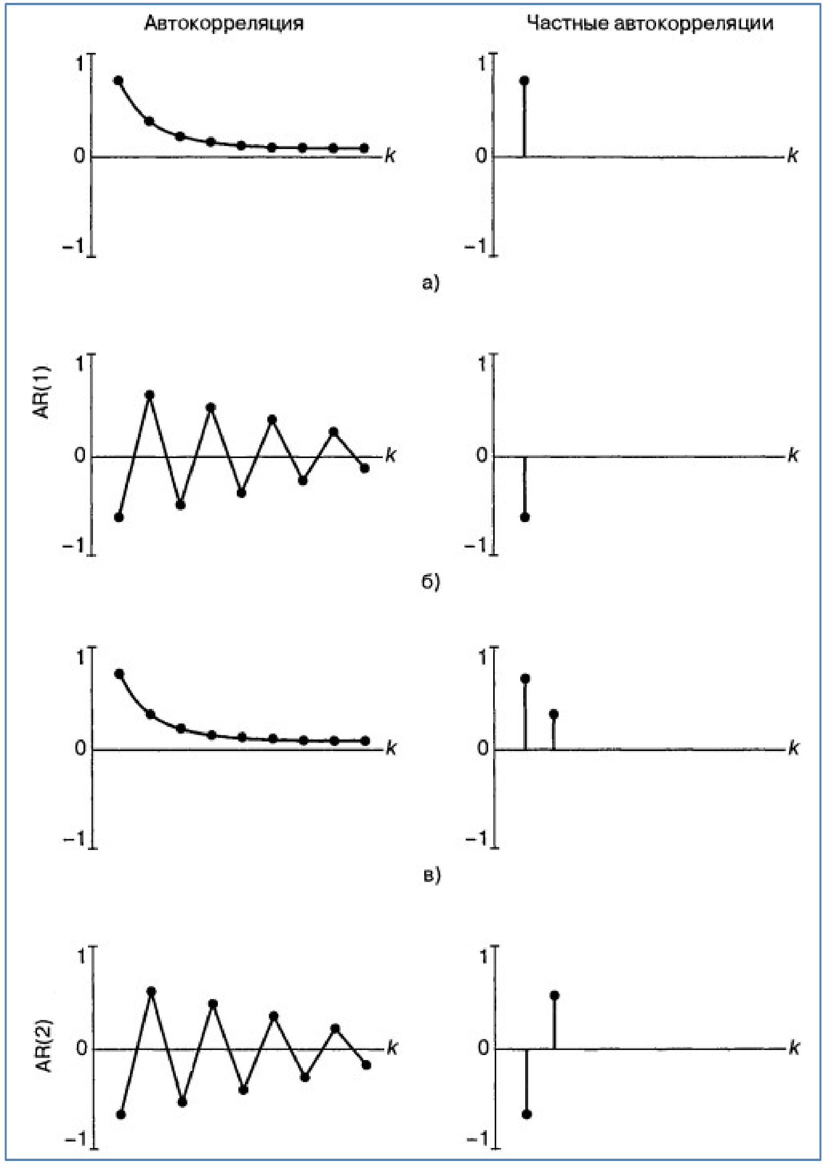
where

$\varphi_0, \varphi_1, \dots, \varphi_p$ - *estimated coefficients (not necessarily in the sum of 1 and can be either positive or negative)*

AR(1): $\hat{Y}_t = \varphi_0 + \varphi_1 Y_{t-1}$

AR(2): $\hat{Y}_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2}$

Autocorrelation and Partial Autocorrelation plots for AR(1) and AR(2)



Moving Average model of the order *q*, MA(q)

$$\hat{Y}_t = \mu - \omega_1 e_{t-1} - \omega_2 e_{t-2} - \cdots - \omega_q e_{t-q}$$

where

μ - *mean of* $Y_1, \dots, Y_t$ $(\bar{Y})$

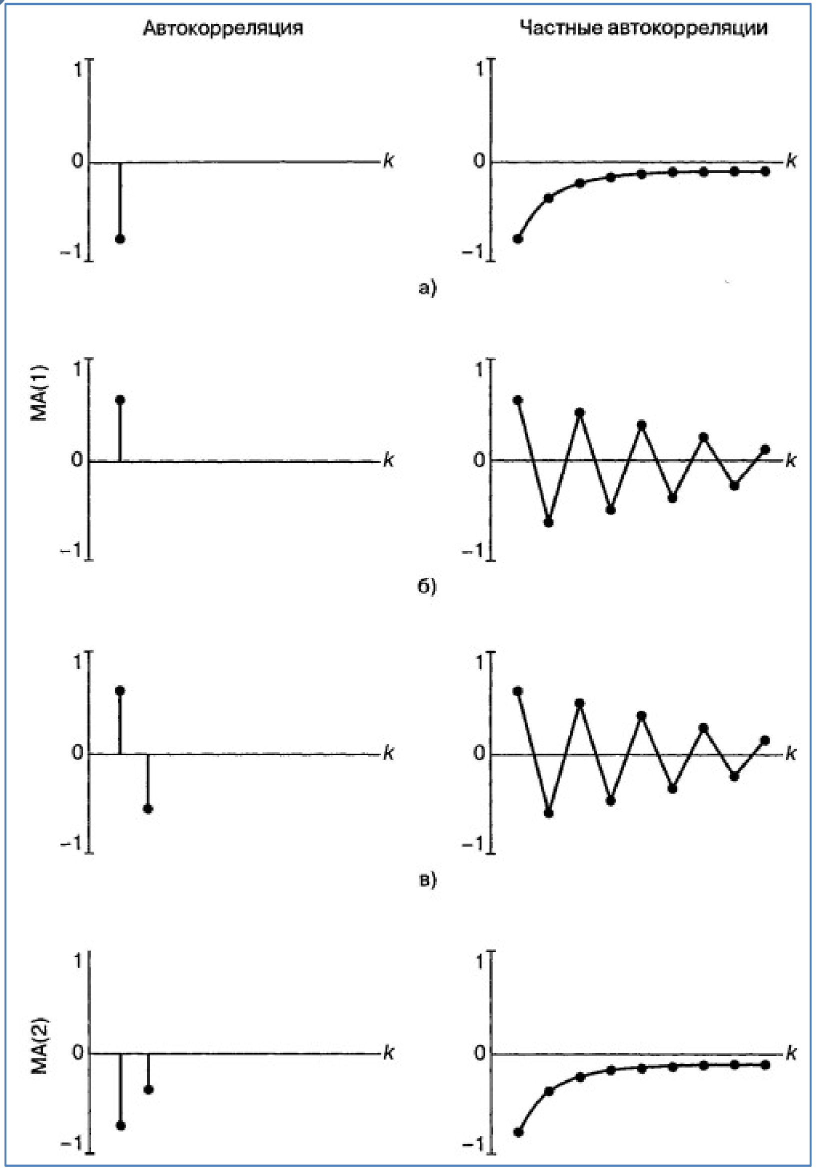
$\omega_0, \omega_1, \dots, \omega_q$ - *estimated coefficients (not necessarily in the sum of 1 and can be either positive or negative)*

$e_{t-1}, \dots, e_{t-q}$ - *errors in previous periods*

MA(1): $\hat{Y}_t = \mu - \omega_1 e_{t-1}$

MA(2): $\hat{Y}_t = \mu - \omega_1 e_{t-1} - \omega_2 e_{t-2}$

Autocorrelation and Partial Autocorrelation plots for MA(1) and MA(2)

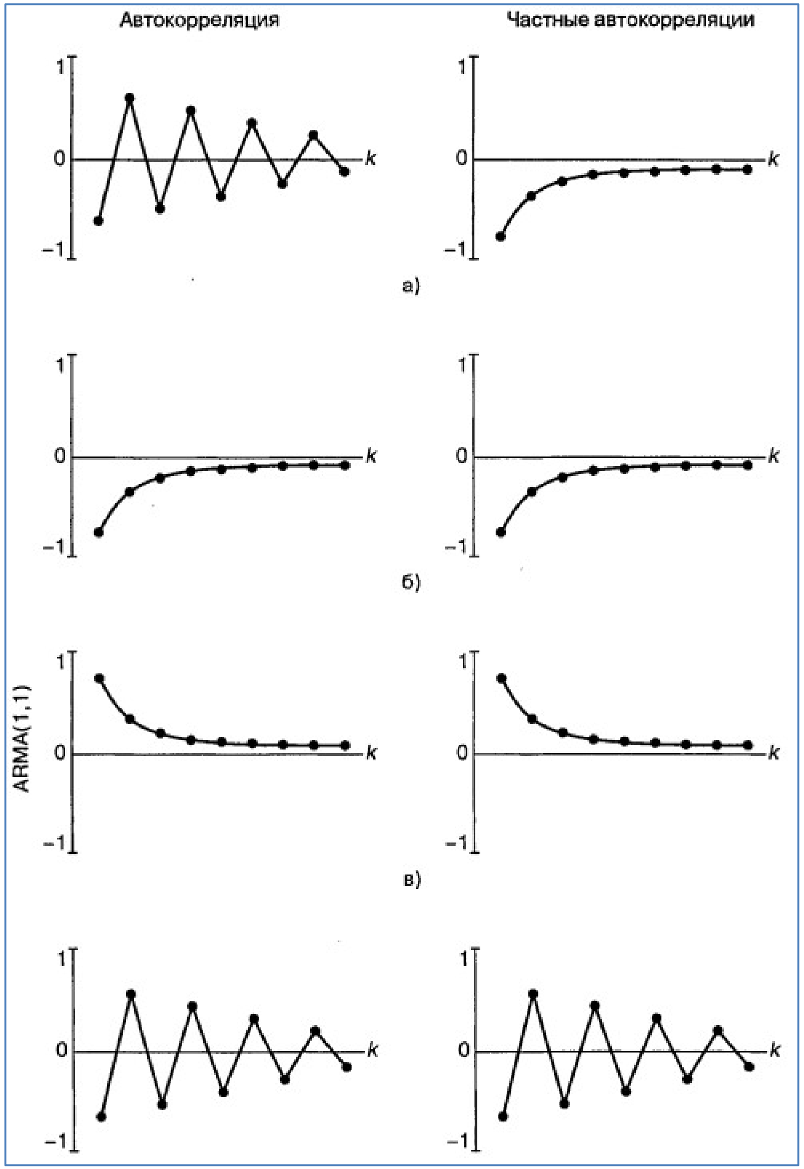


Models with AutoRegression and Moving Average, *ARMA(p, q)*

$$\hat{Y}_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \mu - \omega_1 e_{t-1} - \omega_2 e_{t-2} - \cdots - \omega_q e_{t-q}$$

ARMA(1,1): $\hat{Y}_t = \varphi_0 + \varphi_1 Y_{t-1} + \mu - \omega_1 e_{t-1}$

Autocorrelation and Partial Autocorrelation plots for ARMA(1,1)



Autocorrelation and Partial Autocorrelation plots for MA(q), AR(p), ARMA(p,q)

Model	Autocorrelation	Partial Autocorrelation
MA(q)	Terminates at step q	*Smoothly tends to zero*
AR(p)	*Smoothly tends to zero*	Terminates at step p
ARMA(p,q)	*Smoothly tends to zero*	*Smoothly tends to zero*

AutoRegressive Integrated Moving Average, *ARIMA(p,d,q)*

If TS is not stationary, it should be converted to a stationary one in order to apply *ARMA(p,q)*.

One way to convert is to replace TS itself with TS of differences.

TS of the first differences : $\Delta Y_t = Y_t - Y_{t-1}$

If TS of the first differences is not stationary, then consider TS of the second differences: $\Delta^2 Y_t = \Delta(\Delta Y_t) = Y_t - 2Y_{t-1} + Y_{t-2}$

The taking of the differences can be carried out until we obtain a stationary TS. The number of repetitions of taking the differences needed to obtain stationary TS is denoted by ***d***.

How to understand that TS is not stationary?

- TS plot demonstrates trend or seasonality or cyclicity in data.
- Autocorrelation and Partial Autocorrelation plots demonstrate the absence of a rapid disappearance of coefficients.
- Augmented Dickey–Fuller test

Forecasting errors. Models performance

$e_t = Y_t - \hat{Y}_t$ - **forecasting error in time t.**

1) **Mean Absolute Derivation** (the error is measured in the same units as TS)

$$MAD = \frac{1}{n}\sum_{t=1}^{n}|e_t|$$

2) **Mean Squared Error, Root Mean Squared Error** (the error is measured in the same units as TS, highlights large forecast errors)

$$MSE = \frac{1}{n}\sum_{t=1}^{n}{e_t}^2$$

$$RMSE = \sqrt{MSE}$$

3) **Mean Absolute Percentage Error** (the error shows how large the forecast errors are in comparison with the actual values of TS)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{Y_t}$$

4) **Mean Percentage Error** (the error determines whether the forecast is biased - constantly overvalued or undervalued)

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{e_t}{Y_t}$$

Building *ARIMA(p,d,q)*

Step 1. Determining p, d, q

Step 2. Finding model coefficients using train data, prediction on valid data, check performance (e.g. RMSE)

Step 3. Verifying model based on prediction errors analysis

If a model isn't adequate, go to step 1, otherwise model is ready to use.

?! Hyperparameters Gridsearch instead of step 1

Forecasting for TS

Example: models for Champagne dataset

Models comparison based on RMSE

Model	Valid
Simple naive	3186.501
Naive with trend	
Naive with seasonality	
Naive with trend and seasonality	
ARIMA(1,1,1)	951.260
Best grid result – ARIMA(0,0,1)	939.464

Q & A

Thank you!