Time series analysis

Lecture 1: Introducion

Alaa Bakhti

Your expectations?

Class: 15 hours

Syllabus

- Introduction
  - Patterns & characteristics
  - Stationarity
  - Autocorrelation (ACF, PACF)
- Forecasting
- Anomaly detection

Grading:

- Project on Anomaly detection or

  Forecasting: 80%
- Class participation: 20%

Anomaly detection Project

Subject: anomaly detection on a time series dataset.

Dataset: you need to propose a labeled dataset that needs to be validated before you start working on it

Requirements:

- 2 methods needs to be implemented: 1 based on heuristics and 1 on modeling
- Univariate or Multivariate (30%)
- Analysis of the cases where the anomalies were not detected (FN) and the wrong anomalies (FP) (60%)
- Comparison between predicted and correct anomalies
- Apply the models for all the data (sensors) and not only one
- Configuration of which model to use for which sensor

Deliverables: Github repository respecting the DSP course guidelines (requirements.txt, data folder, readme.md, etc) with a well documented notebook(s) (10%)

❌ labels should not be used for the prediction

Forecasting project

Subject: Forecasting (no Stock price prediction)

Dataset: you need to propose a dataset that needs to be validated before you start working on it

Requirements:

- 2 methods needs to be implemented: 1 statistical (ARIMA, etc) and another
- Univariate or Multivariate (30%)
-

Deliverables: Github repository respecting the DSP course guidelines (requirements.txt, data folder, readme.md, etc) with a well documented notebook(s) (10%)

Defense

- **Duration:** 35 min - 25 min presentation 10 min Q&A
- The defense will be composed of 3 parts:
- Presentation of the followings (15 min)
    - Use case and dataset
    - Different models used and how they work
    - Results of the different models, their advantages & limitations when applied in the use case
    - References of the different resources used for the project (research papers, blogs, etc)
- Demonstration of the different models (10 min)
- Q&A (10 min)

Introduction

Time series

Definition

- Series of data points indexed in time order: for sensor data, a single observation is composed of the sensor measured value and the time it was measured in
- Have a natural temporal ordering
- Examples: Sensor data, sale data, stock market

Real world problems

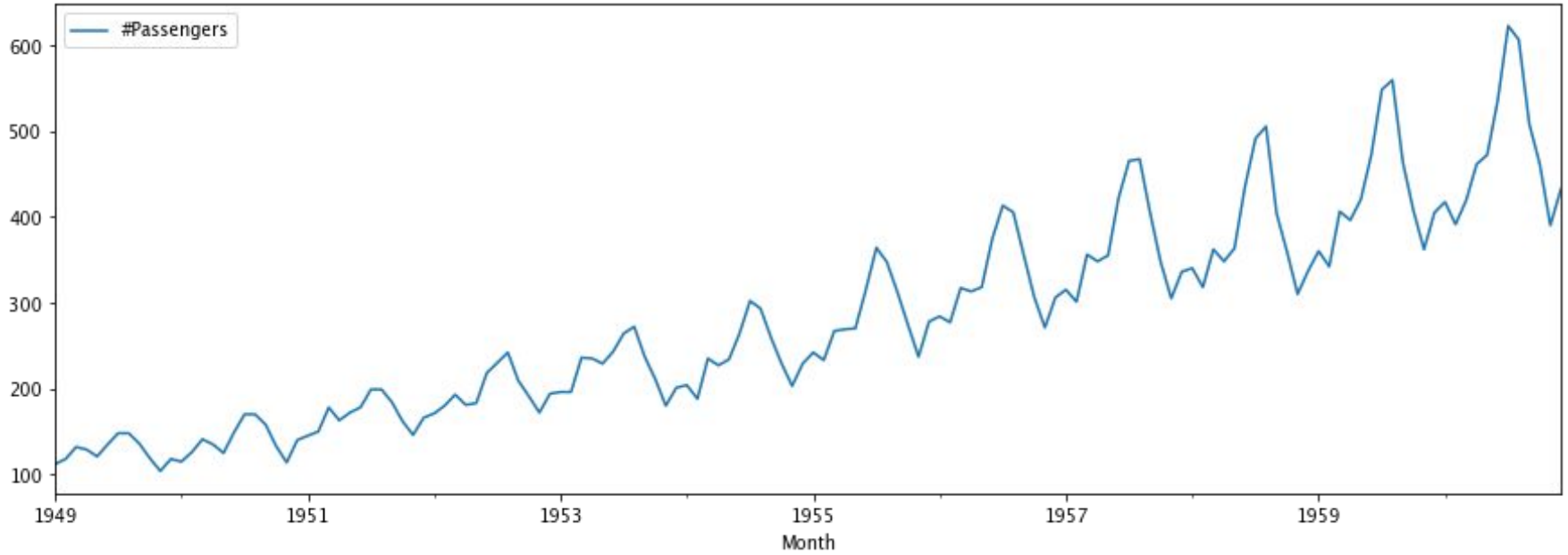
- Time series annotation
  - Outlier / anomaly detection: assign a label to each data point (anomaly or not)
  - Change point detection (segmentation)
- Forecasting: predict the future values of the time series based on previously observed values

Patterns and characteristics

Trend

*A trend exists when there is a **long-term** increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend as “changing direction,” when it might go from an increasing trend to a decreasing trend - Rob J Hyndman*

- General tendency for the time series to go up or down over time
- For a trend to emerge, enough data needs to be available (2 weeks worth of data is not sufficient)



Number of air passengers per month

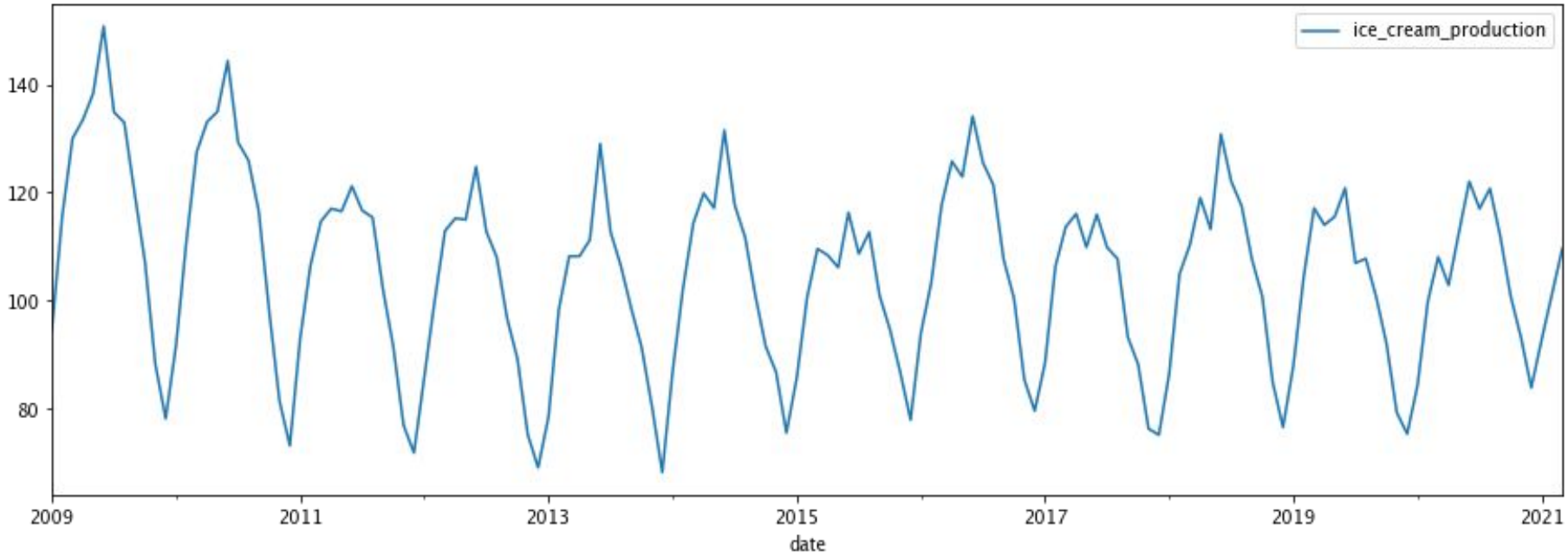
Seasonality

A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. ***Seasonality is always of a fixed and known frequency*** *- Rob J Hyndman*

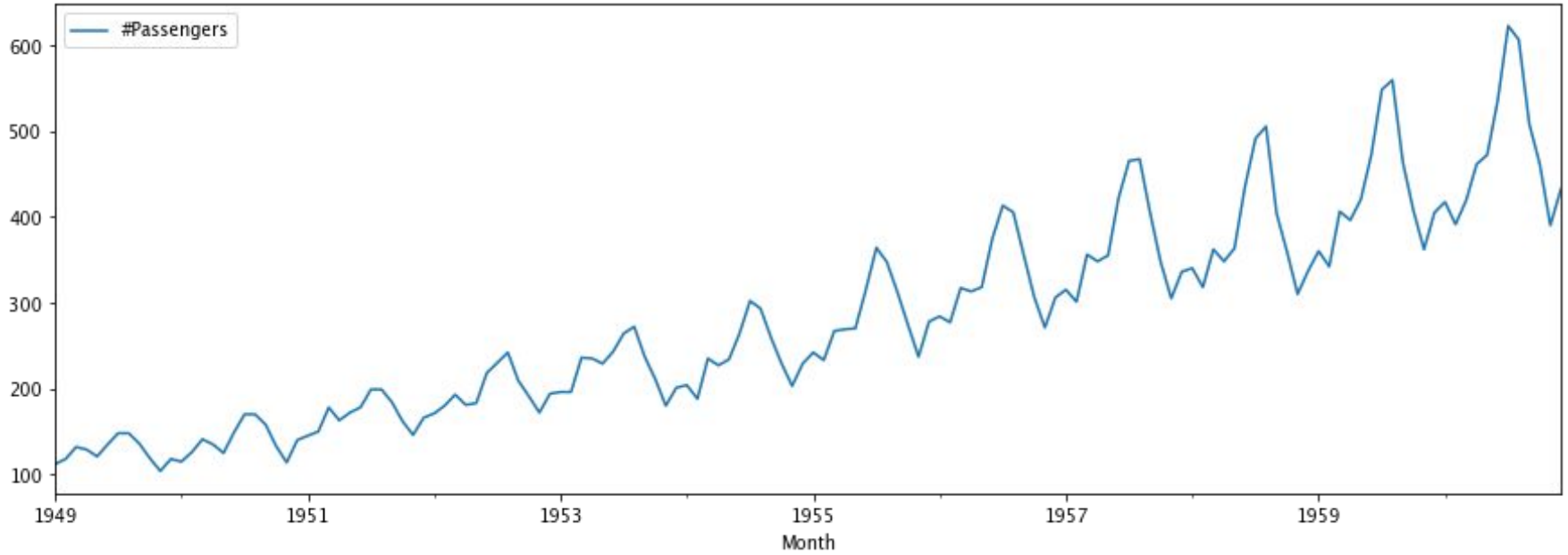
- **Regular** pattern in the time series related to the calendar
- Season = week, nb days, nb hours, etc
- A seasonal time series has peaks and troughs at predictable times

Examples:

- Electricity consumption is high during the day and low during the night
- Ice cream consumption is high during the summer and low the rest of the year



Ice cream production in the US



Number of air passengers per month

Cyclicity

*A cycle occurs when the data exhibit rises and falls that are **not of a fixed frequency**. These fluctuations are usually due to economic conditions, and are often related to the "business cycle." - Rob J Hyndman*

- Irregular peaks and troughs at an unpredictable times
- They can not be predicted according to time like seasonal time series

Examples

- Stock prices

Stationarity

A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary, the trend and seasonality will affect the value of the time series at different times. - Rob J Hyndman

Conditions: the distribution of data doesn't change over time:

- No trend: constant mean and variance over time
- No seasonality
- Autocorrelation

In general, a stationary time series will have no predictable patterns in the long-term

Stationarity

Why do we care?

- Stationarity is an important characteristic for time series modeling
- If a process is not stationary, it becomes difficult to model (estimate the different parameters)
- If parameters vary overtime, too many parameters should be estimated

How to check if a time series is stationary?

- Visually by plotting the time series
- With statistical tests like the Augmented Dickey–Fuller test (ADF)

Augmented Dickey–Fuller test

- Test for trend non-stationarity
- Null-hypothesis is time series is non-stationary

It is recommended to always plot the time series to determine if it is stationary or not

How to make a non-stationary time series stationary?

- Differentiating
  - Removes the trend from the time series
  - **1st order:** subtract the previous value from each value of the time series (X2_new = X2 - X1)
  - **2nd order** (difference 2 times), 3rd order, etc
  - **Seasonal difference**: If the time series is seasonal with season = 3, you can test the 3rd order difference.
- Other transformations
  - Log (eliminate exponential growth, stabilises the variance)
  - Exponential
  - Square root
- Sequence of transformations: Log then difference, etc

Autocorrelation in time series

Lag (Backshift) operator

In time series analysis, the lag operator (L) or backshift operator (B) operates on an element of a time series to produce the previous element - Wikipedia

- Notation
  - y(t) = the time series at t
  - y(t-1) = the time series at t - 1
  - L = the lag operator
- L y(t) = y(t - 1): the time series *1* period ago
- L^2 y(t) = y(t - 2): the time series *2* time periods ago
- L^3 y(t) = y(t - 3): the time series *3* time periods ago
- ...
- Lk y(t) = y(t - k): the time series *k* time periods ago

Autocorrelation function (ACF)

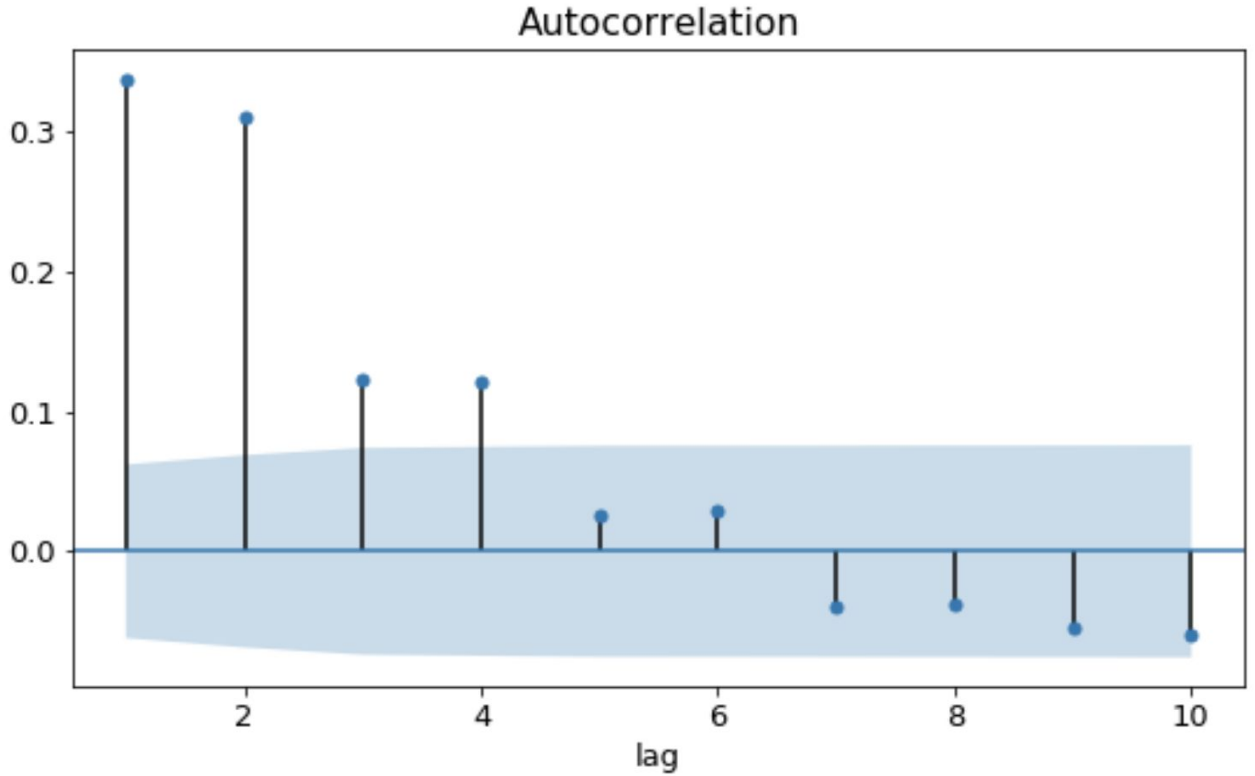
The time series value at t may depend on its past values (t-1, t-2, t-3, etc)

Autocorrelation

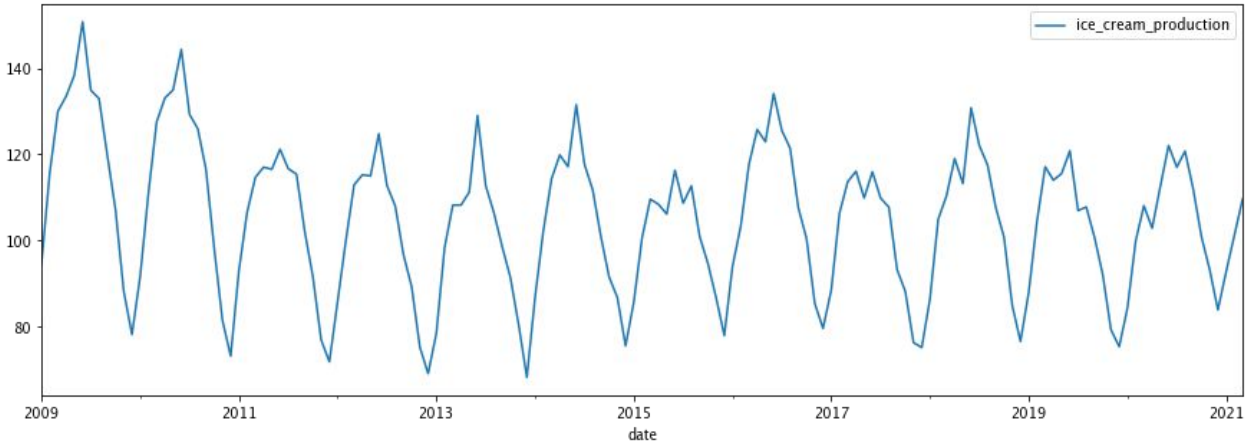
- The correlation of a time series with a lagged copy of itself (lag 1) *corr(y(t), y(t-1))*
- Describes how well a value of the series is related with its past values

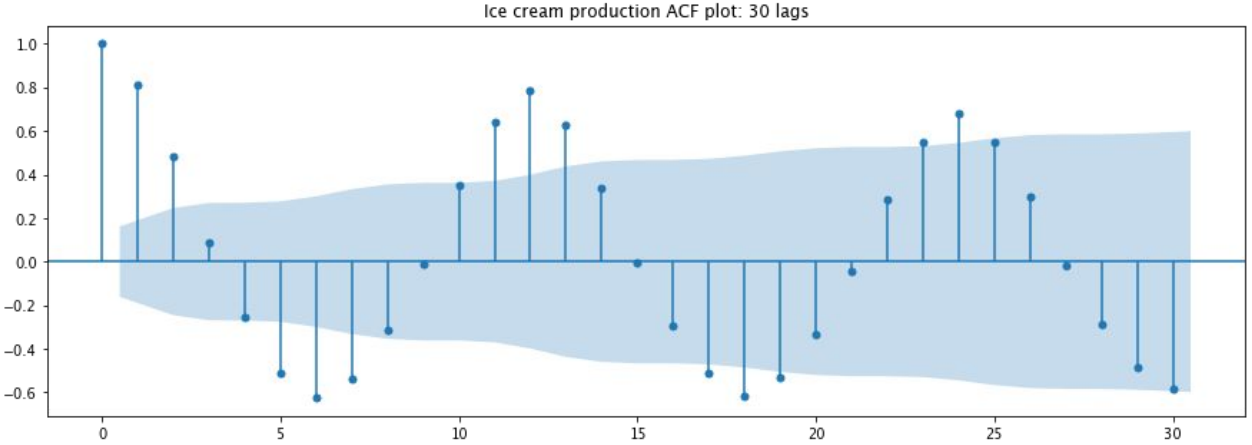
Autocorrelation function (ACF)

- Measures the autocorrelation of the time series for different lags (lag 1, lag 2, etc): *corr(y(t), y(t-k))*
- ACF plot: the blue regions represents the error band, any value inside it is not statistically significant and can be considered as 0 autocorrelation

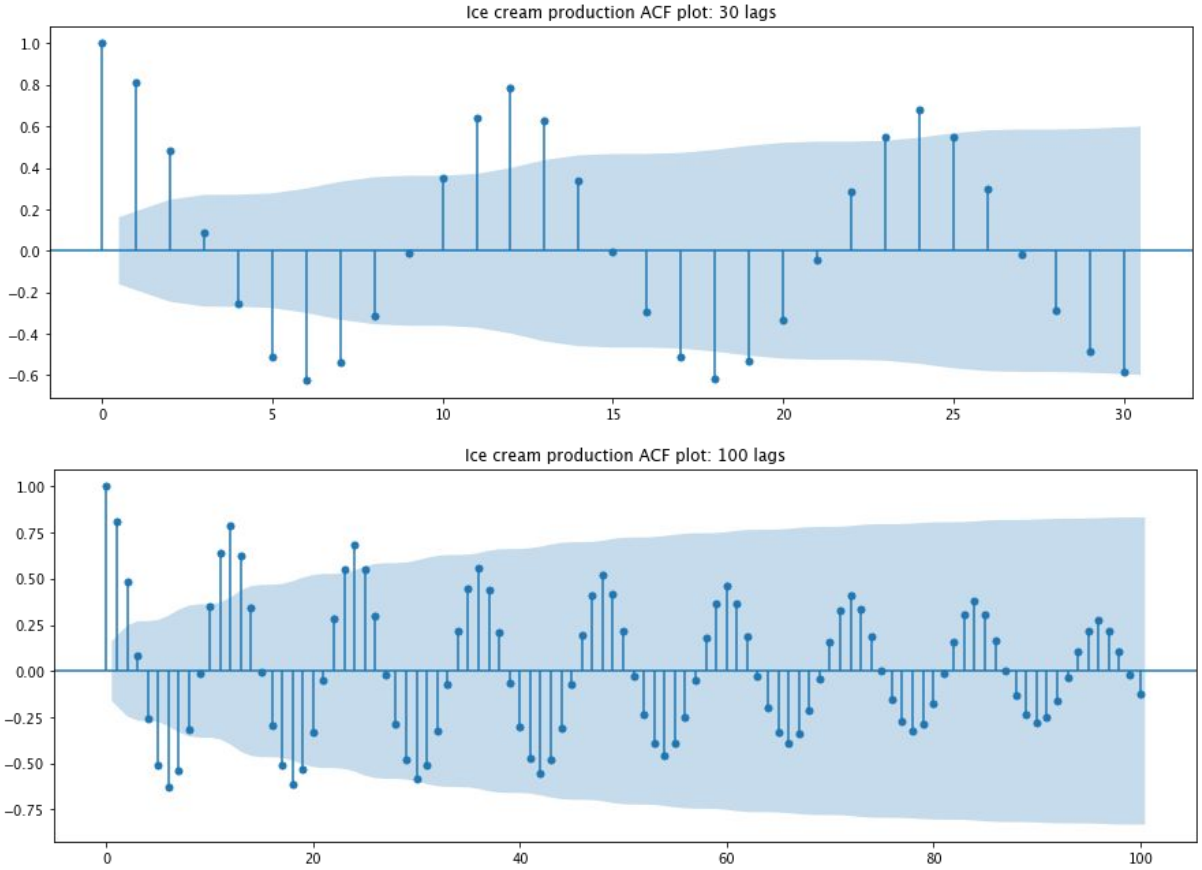


ACF plot of a time series

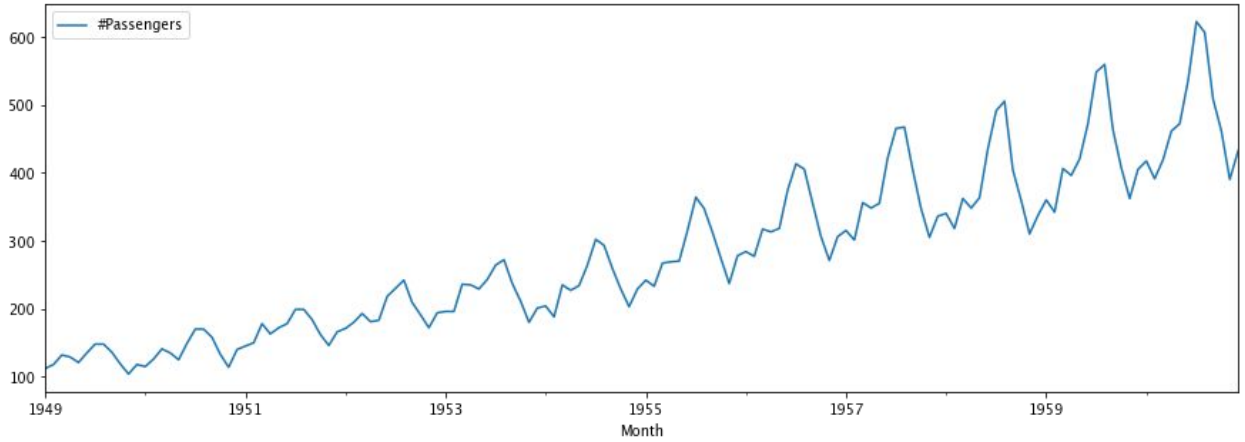


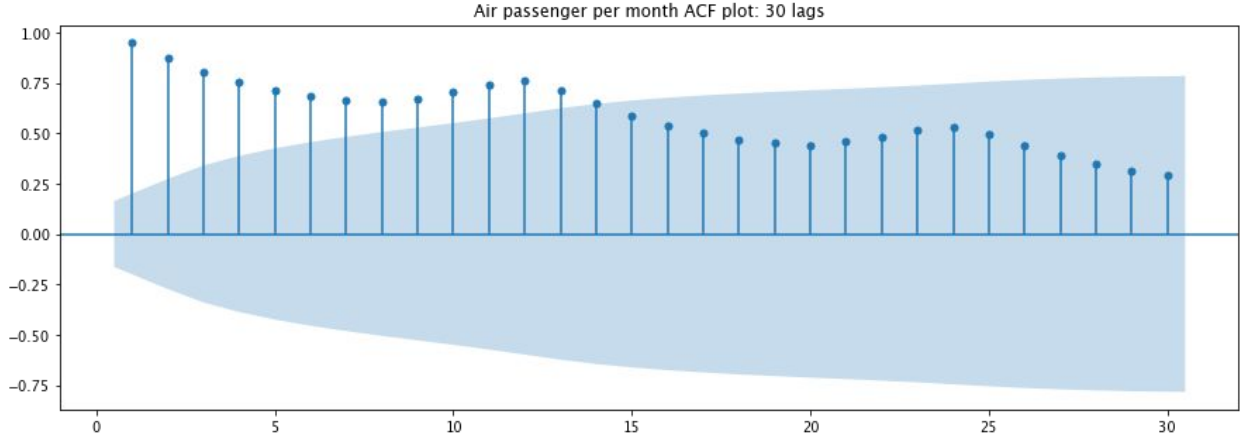


ACF plot of a time series with seasonality

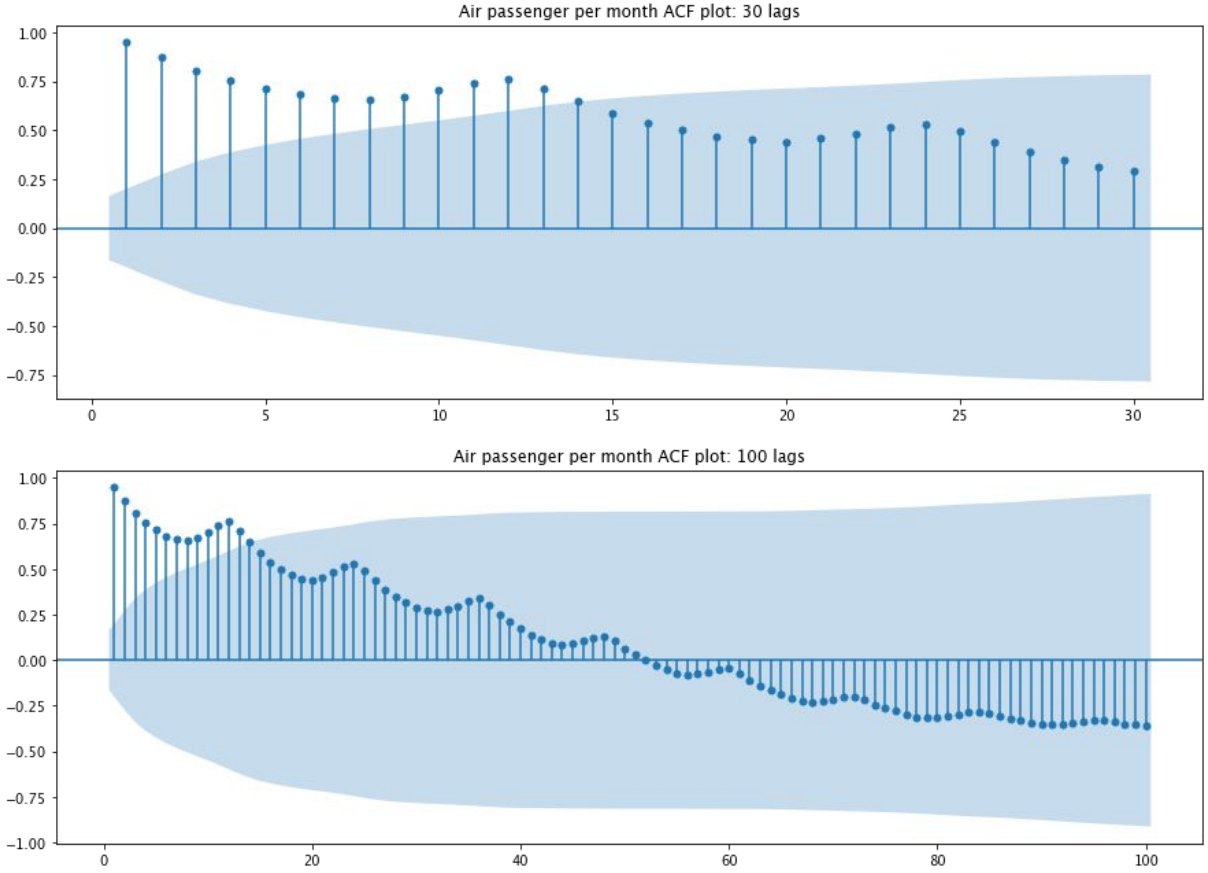


ACF plot of a time series with seasonality





ACF plot of a time series with trend and seasonality

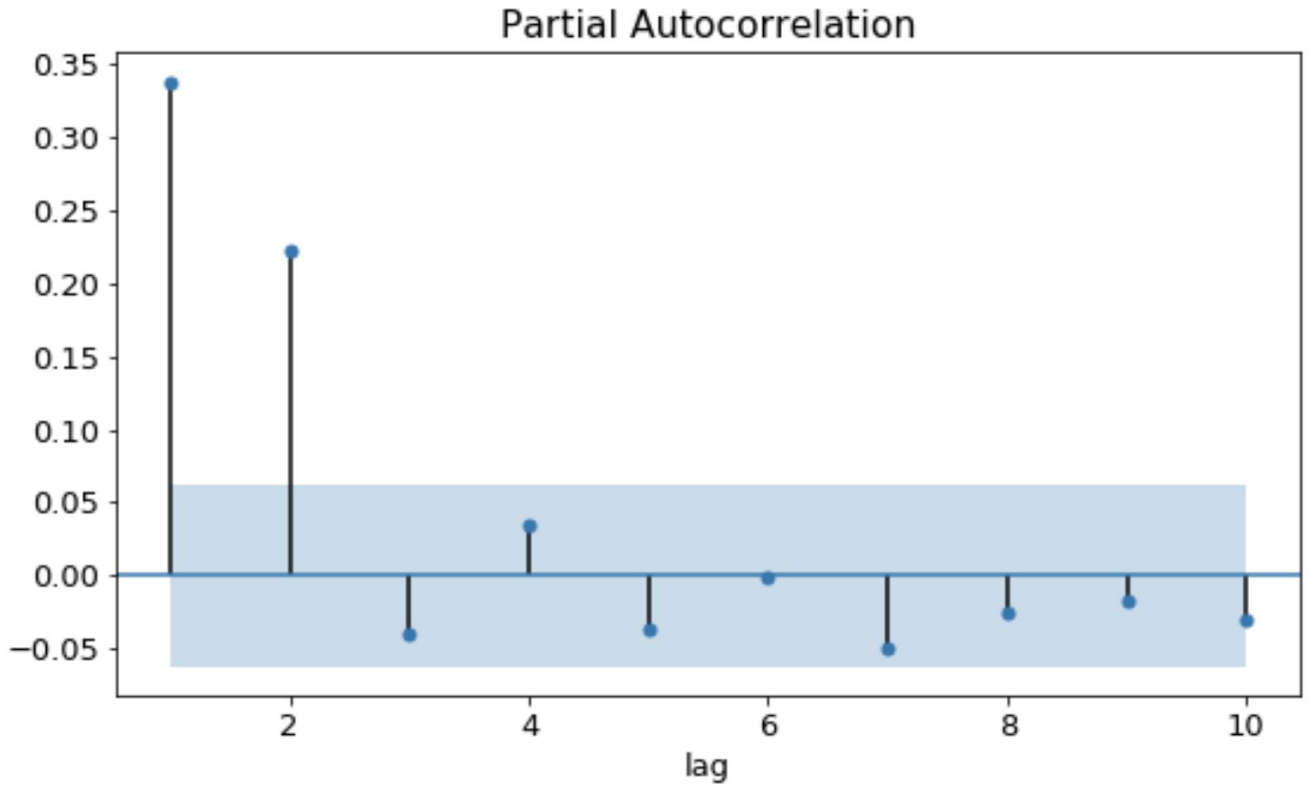


ACF plot of a time series with trend and seasonality

Partial autocorrelation function (PACF)

If y(t) and y(t-1) are correlated, then y(t-1) and y(t-2) must also be correlated. However, then y(t) and y(t-2) might be correlated, simply because they are both connected to y(t-1), rather than because of any new information contained in y(t-2) that could be used in forecasting y(t). To overcome this problem, we can use partial autocorrelations. These measure the relationship between y(t) and y(t-k) after removing the effects of lags 1, 2, 3, ..., k-1. - Rob J Hyndman

- Example: the PACF of 4 is the direct correlation of the value 4 time periods ago on the current value of the time series without considering the intermediate time periods (1, 2 and 3)

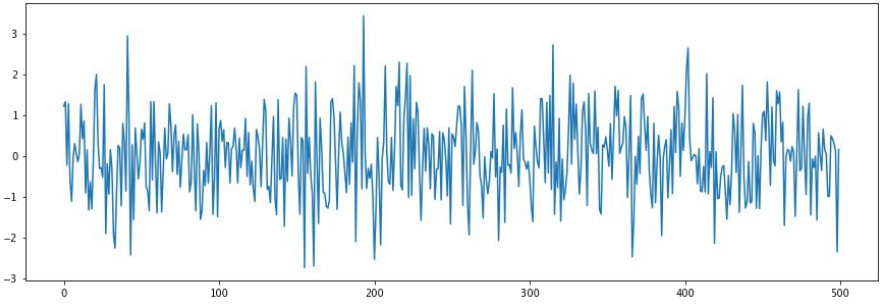


PACF plot of a time series

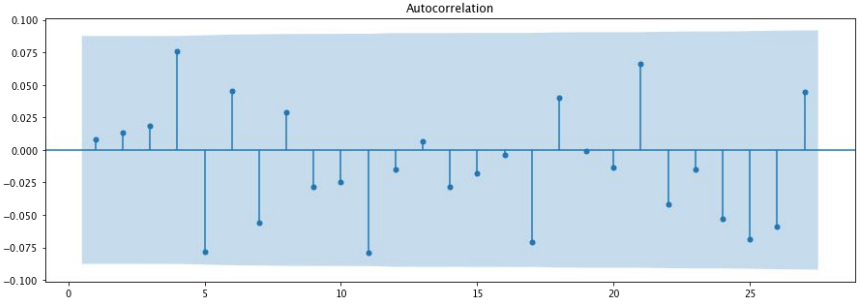
White noise

A non predictable time series that is composed only of random observations

- No trend, no seasonality, no cyclicity
- Zero mean and constant variance
- “zero” autocorrelation at all lags: the ACF plot is composed of only insignificant spikes: 95% of the series autocorrelations are not statistical significant and can be considered as zero



White noise time series



ACF plot of a white noise time series

Practical work

Practical work

For this practical work, we will be using the Air passenger dataset from kaggle

- Setup time series
  - Load the data
  - Convert the column types (datetime, etc)
  - Set the date column as a datetime index
  - Plot the time series
- Stationarity
  - Use the augmented Dickey-Fuller test to check if the time series is stationary or not. You can use the statsmodels library
  - If it is not, apply some transformation on the time series to make it stationary (differing, log, square root, etc)
- ACF
  - Plot the ACF function for the different time series (raw, 1st order difference, etc)
  - Plot the PACF function of the transformed time series

Useful links

- Forecasting: Principles and Practice - Rob J Hyndman and George Athanasopoulos (2nd ed)
- Packages
  - Time series toolbox (Facebook) https://facebookresearch.github.io/Kats/
  - Forecasting package (Linkedin) https://linkedin.github.io/greykite/