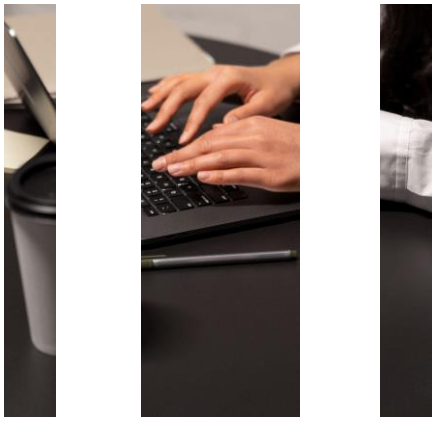
Extração, Tratamento e Carregamento de Dados

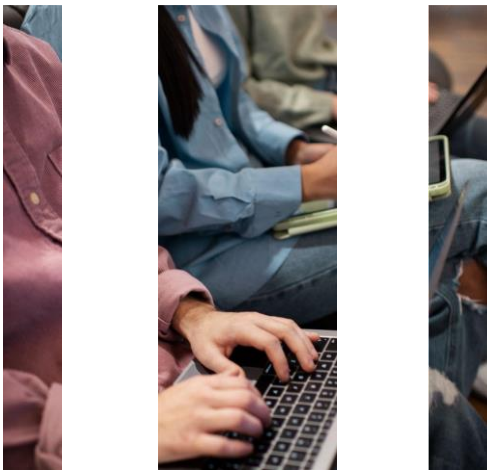


1

05

Módulo 5

Extração, Tratamento e Carregamento de Dados



Balduíno Mateus

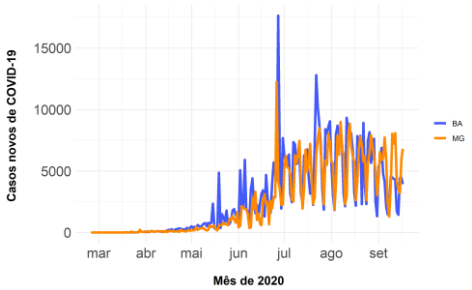
2

Séries Temporais

As séries temporais são informações registada em intervalos de tempo

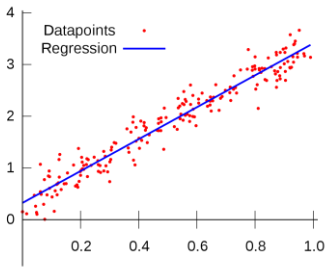
Exemplos:

- **Horas:** Registo de funcionamento de uma máquina;
- **Dias:** Previsão meteorológica;
- **Mês :** Taxa de desemprego
- **Etc..**



3

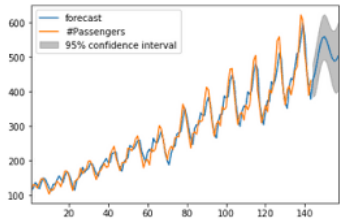
Séries Temporais



- As séries temporais são diferentes das regressões!
- **Regressão** para caso de previsão faz-se uma interpolação

4

Séries Temporais



- As séries temporais são deferentes das regressões!
- **Séries temporais** para caso de previsão faz-se uma extrapolação

5

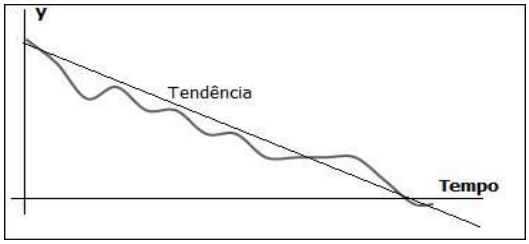
Séries Temporais

Componentes de uma série temporal:

- Componente Tendência;
- Componente Sazonal;
- Componente Cíclica.

6

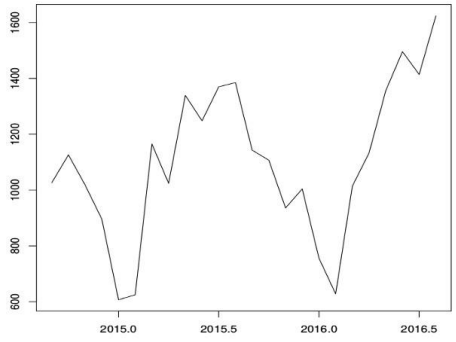
Séries Temporais



- Tendência: Movimento oculto nos dados crescente, decrescente ou estacionária

7

Séries Temporais

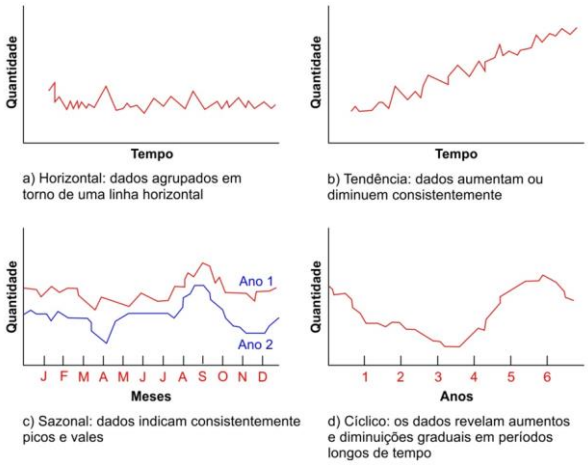


- Sazonal: Flutuações regulares dentro de um período completo de tempo (dia, semana, mês, etc.) Representam um tipo de padrão que se repete.

8

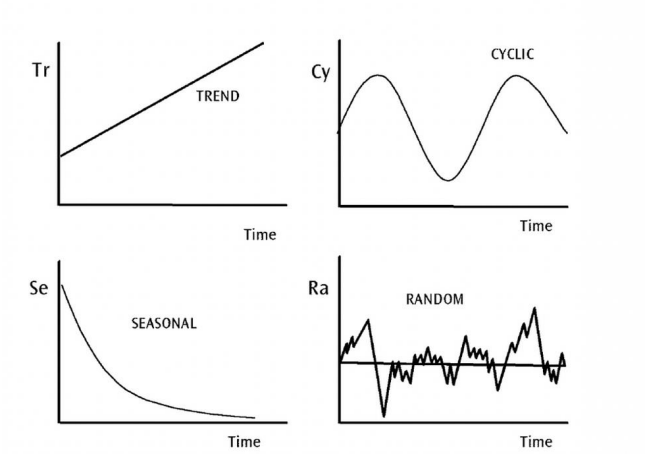
Séries Temporais

Cíclica: Flutuações de longo prazo nos dados e são similares aos fatores sazonais.



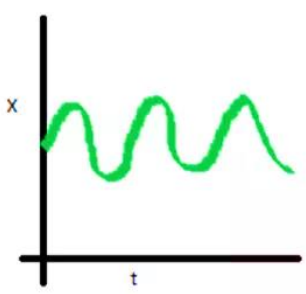
9

Séries Temporais

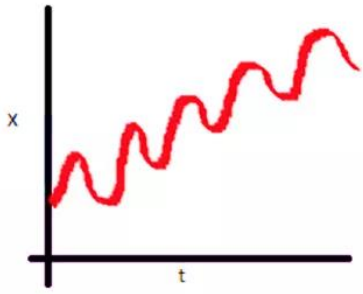


10

Séries Temporais



Stationary series



Non-Stationary series

11

Séries Temporais

Alguns problemas de séries temporais:

- Quantos produtos será vendido?
- Quantos casos de covid Portugal terá no próximo mês?
- Consumo energético de uma cidade nos próximas semanas?
- Quanto triliões de toneladas de aço será produzido no próximo ano?
- Cotação do bitcoin no próxima hora.

12

Séries Temporais

- O pandas oferece várias ferramentas incorporadas e algoritmos de dados para séries temporais. É possível trabalhar com séries temporais bem grandes de modo eficaz e manipular, agregar e fazer uma reamostragem de séries temporais irregulares e de frequência fixa facilmente.

13

Séries Temporais

14

Séries Temporais

Modelos séries temporais

Modelos tradicional

- Modelo Univariados
  - ARIMA
  - SARIMA
- Modelos Multivariados
  - Vetor Autorregressivo

Modelos de Aprendizado de máquinas

- RNN

15

Séries Temporais

ARIMA:

- AR: Autoregression- Utiliza valores numéricos históricos para prever o futuro (lag). Parâmetros (p)
- I: Integrated – Técnica para remover a tendência na série temporal e facilitar a análise (Tornar a série estacionária). Parâmetros (d)
- MA: Moving Average – Usa erros residuais a partir da média movel. Parâmetros (p)

16

Séries Temporais

- Conversão entre string e datetime
- from datetime import datetime
- data = datetime(2022,7,2) → Objetos Timestamp do pandas
- str(data) → Converter em string

'2022-07-02 00:00:00'

- data.strftime('%Y-%m-%d') → Passar para datetime

'2022-07-02'

17

Séries Temporais

- Gerar sequências de datas e intervalos de tempo de frequência fixa
- data = pd.date_range('2022-05-01', periods=24, freq='H')
- df = pd.DataFrame(data, columns=['TimeStemp'])

TimeStemp
2022-05-01 19:00:00
2022-05-01 20:00:00
2022-05-01 21:00:00
2022-05-01 22:00:00
2022-05-01 23:00:00

18

Séries Temporais

- time = pd.date_range('2022-05-01', periods=24, freq='H')
- data = pd.DataFrame({'TimeStemp':time, 'Data': np.random.rand(24)})
- data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24 entries, 0 to 23
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   TimeStemp  24 non-null     datetime64[ns]
 1   Data       24 non-null     float64
dtypes: datetime64[ns](1), float64(1)
memory usage: 512.0 bytes
```

19

Séries Temporais

	SNo	Name	Symbol	Date	High	Low	Open	Close	Volume	Marketcap
0	1	Bitcoin	BTC	2013-04-29 23:59:59	147.488007	134.000000	134.444000	144.539993	0.0	1.603769e+09
1	2	Bitcoin	BTC	2013-04-30 23:59:59	146.929993	134.050003	144.000000	139.000000	0.0	1.542813e+09
2	3	Bitcoin	BTC	2013-05-01 23:59:59	139.889999	107.720001	139.000000	116.989998	0.0	1.298955e+09
3	4	Bitcoin	BTC	2013-05-02 23:59:59	125.599998	92.281898	116.379997	105.209999	0.0	1.168517e+09
4	5	Bitcoin	BTC	2013-05-03 23:59:59	108.127998	79.099998	106.250000	97.750000	0.0	1.085995e+09

20

Séries Temporais

```
df['Date'] = pd.to_datetime(df['Date'], format='%Y-%m-%d %H:%M:%S')
df=df.set_index('Date')
del df['SNo']
```

Date	Name	Symbol	High	Low	Open	Close	Volume	Marketcap
2013-04-29 23:59:59	Bitcoin	BTC	147.488007	134.000000	134.444000	144.539993	0.0	1.603769e+09
2013-04-30 23:59:59	Bitcoin	BTC	146.929993	134.050003	144.000000	139.000000	0.0	1.542813e+09
2013-05-01 23:59:59	Bitcoin	BTC	139.889999	107.720001	139.000000	116.989998	0.0	1.298955e+09
2013-05-02 23:59:59	Bitcoin	BTC	125.599998	92.281898	116.379997	105.209999	0.0	1.168517e+09
2013-05-03 23:59:59	Bitcoin	BTC	108.127998	79.099998	106.250000	97.750000	0.0	1.085995e+09

21

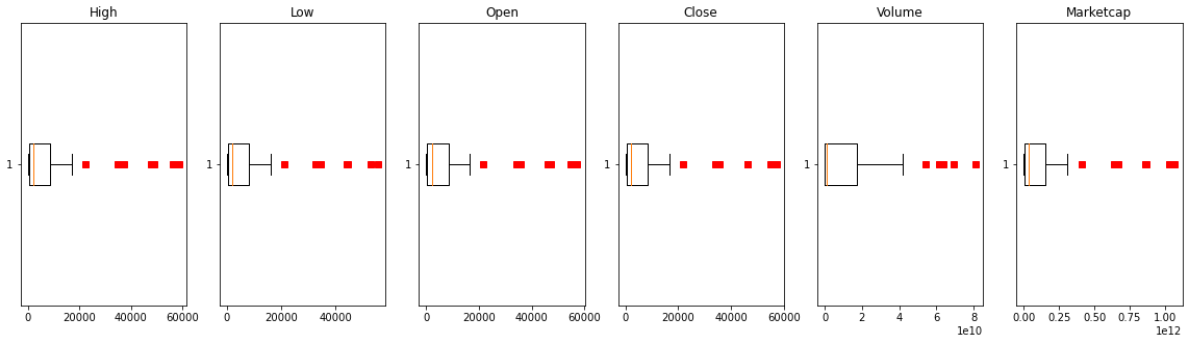
Séries Temporais

- Para que variar a taxa de amostragens basta usar a função:
- df_hora = df.resample('M').mean()

Date	High	Low	Open	Close	Volume	Marketcap
2013-04-30	147.209000	134.025002	139.222000	141.769997	0.0	1.573291e+09
2013-05-31	123.949096	114.253513	120.292097	119.992741	0.0	1.339718e+09
2013-06-30	111.300543	104.602963	108.856067	107.761407	0.0	1.216792e+09
2013-07-31	93.868936	86.719010	90.311422	90.512207	0.0	1.034233e+09
2013-08-31	116.002226	111.388452	113.041936	113.905484	0.0	1.317466e+09

22

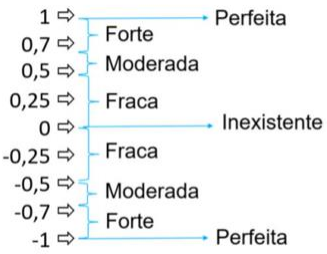
Séries Temporais

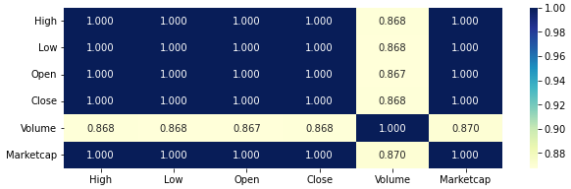


23

Séries Temporais

Séries Temporais – Correlação

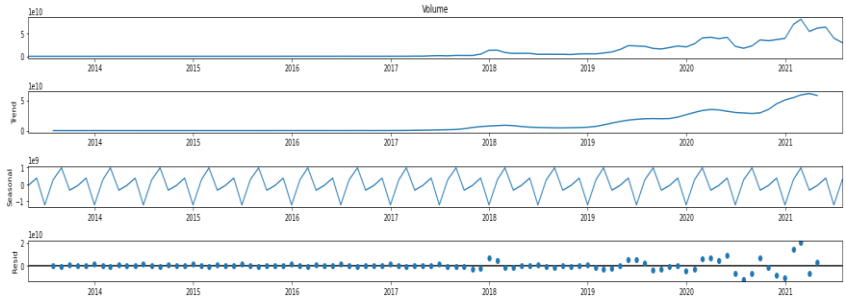




24

Séries Temporais

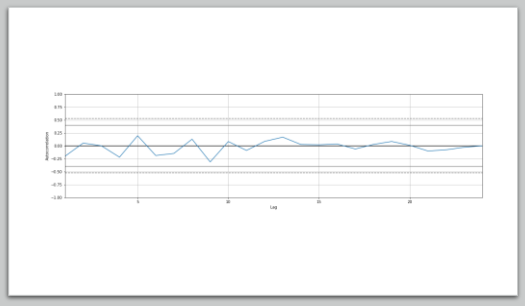
```
from statsmodels.tsa.seasonal import seasonal_decompose
plt.figure(figsize=(10,20))
plt.rcParams["figure.figsize"] = (larg,alt)
comp= seasonal_decompose(df['coluna'],model='additive', period=6)
comp.plot()
plt.show()
```

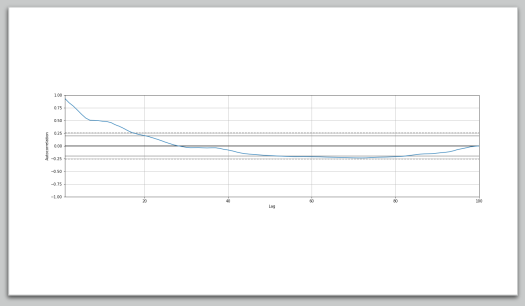


25

Séries Temporais

- O correlograma é uma ferramenta comumente usada para verificar aleatoriedade em um conjunto de dados .

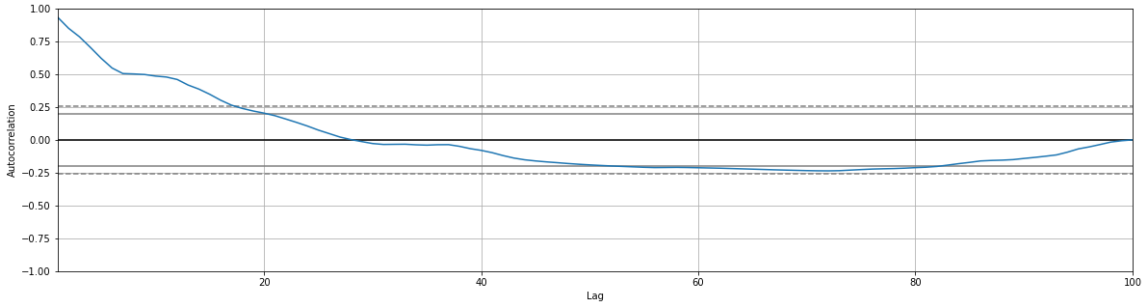




26

Séries Temporais

- Autocorrelação refere-se ao grau de correlação das mesmas variáveis entre dois intervalos de tempo sucessivos .



27

Séries Temporais

- Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

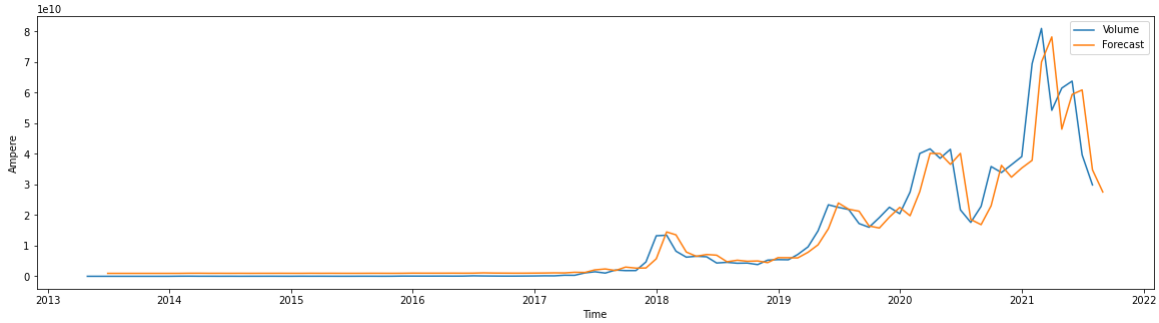
- MAPE = Mean Absolute Percentage Error

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

28

Séries Temporais

AR

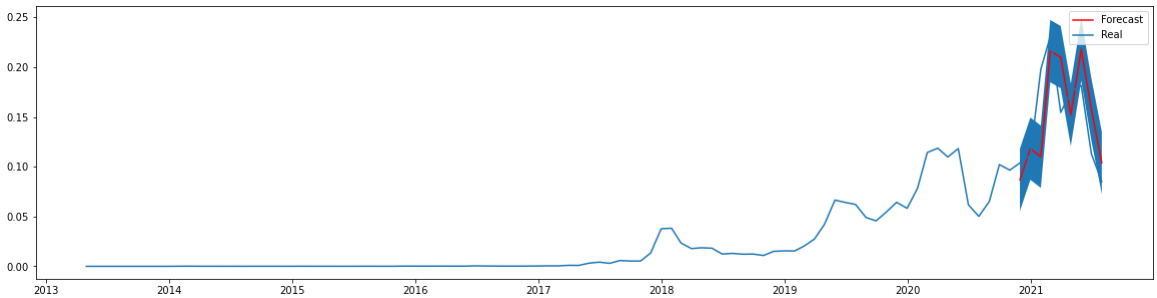


MAPE=1313923345507.071

29

Séries Temporais

ARIMA

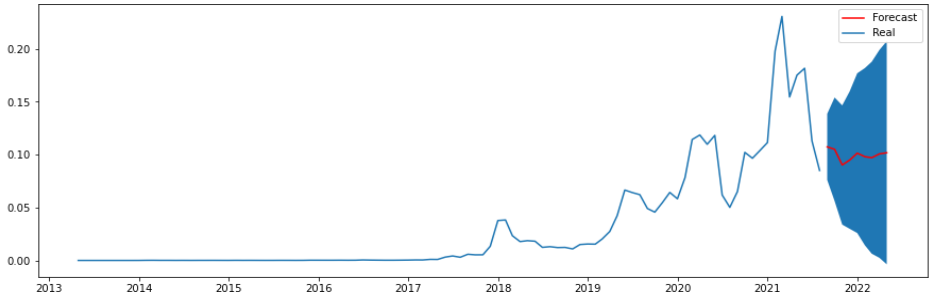


MAPE = 1182344266338.9602

30

Séries Temporais

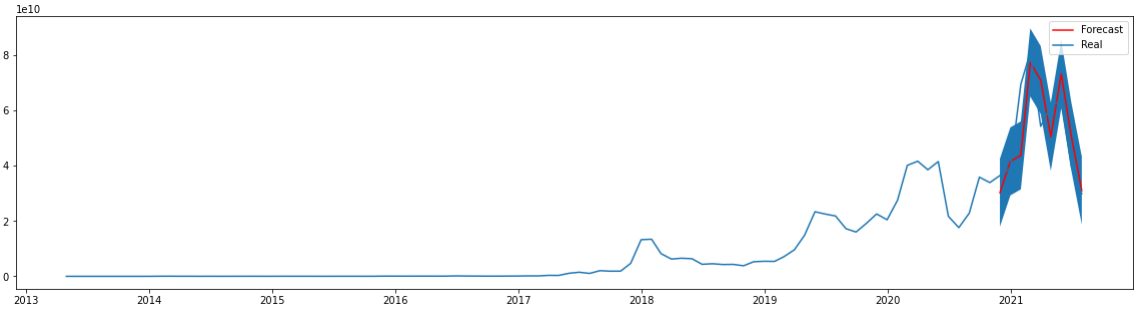
ARIMA



31

Séries Temporais

SARIMA




```
MAPE = 968336489053.5988
```

32

Exercício 22

Faça a previsão da variável do 'Marketcap' dataframe.

33

Exercício 23

- Faça importação do ficheiro 'data_covid.csv' e realize as seguintes operações:
- Elimine os valores duplicados e os Nan;
- Apresente o sumário estatístico das 5 primeiras variáveis;
- Apresente o histograma, qqplot e o boxplot das variáveis anteriores;
- Faça o estudo da correlação entre as 5 variáveis;
- Verifique se existe sazonalidade na variável 'confirmados_arscentro';
- Verifique a estacionariedade da série;
- Faça a previsão da mesma variável para 5 dias e apresenta o erro MAPE.

34