Time Series

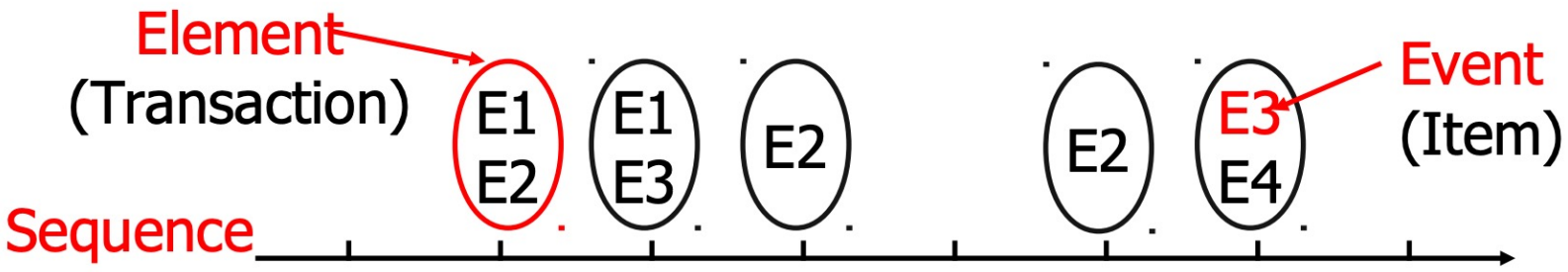
Exercise 3

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Sequential Pattern Mining

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Sequence Database	Sequence	Element (Transaction)	Event (Item)
Customer Data	Purchase history of a given customer	A set of items bought by a customer at time t	Books, dairy products, CDs, etc
Web Server Logs	Browsing activity of a particular Web visitor	A collection of files viewed by a Web visitor after a single mouse click	Home page, index page, contact info, etc
Chord Progressions	Chords played in a song	Individual notes hit at a time	Notes (C, C#, D, ...)



Sequences

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- A sequence is an ordered list of elements (transactions)

$$s = < e_1\ e_2\ e_3\ \ldots >$$

  - Each element contains a collection of items (events)

$$e_i = \{i_1, i_2, \ldots, i_k\}$$

  - Each element is attributed to a specific time

- Length of a sequence |s| is given by the number of elements of the sequence.

- A k-sequence is a sequence that contains k events (items).

Subsequences

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- A sequence $\langle a_1\, a_2 \ldots a_n \rangle$ is contained in another sequence $\langle b_1\, b_2 \ldots b_m \rangle$ ($m \geq n$) if there exist integers $i_1 < i_2 < \ldots < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, …, $a_n \subseteq b_{i_n}$

Data sequence <b>	Subsequence <a>	Contain?
< {2,4} {3,5,6} {8} >	< {2} {3,5} >	Yes
< {1,2} {3,4} >	< {1} {2} >	No
< {2,4} {2,4} {2,5} >	< {2} {4} >	Yes

- The *support* of a subsequence w is defined as the fraction of data sequences that contain w
- A *sequential pattern* is a frequent subsequence (i.e., a subsequence whose support is ≥ *minsup*)

Generalized Sequential Pattern Algorithm (GSP)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Step 1:
  - Make the first pass over the sequence database D to yield all the 1-element frequent subsequences
- Step 2: Repeat until no new frequent subsequences are found
  1. Candidate Generation:
     - Merge pairs of frequent subsequences found in the (k-1)*th* pass to generate candidate sequences that contain k items
  2. Candidate Pruning:
     - Prune candidate k-sequences that contain infrequent (k-1)-subsequences (Apriori principle)
  3. Support Counting:
     - Make a new pass over the sequence database D to find the support for these candidate sequences
  4. Candidate Elimination:
     - Eliminate candidate k-sequences whose actual support is less than *minsup*

Applying GSP

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data sequence <b>
< {2,4} {3,5,6} {8} >
< {1,2} {3,4} >
< {2,4} {2,4} {2,5} >

- **minsup: 50%**
- Frequent 1-sequences?
- Frequent 2-sequences?
- Frequent 3-sequences?

1-sequences	2-sequences	3-sequences
{1}	{2}, {3}	{2,4}, {5}
{2}	{2,4}	
{3}	{2}, {4}	
{4}	{2}, {5}	
{5}	{4}, {5}	
{6}		
{8}		

Task: Study Histories

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Student	Semester 1	Semester 2	Semester 3	Semester 4
1235894	CS101, CS103, CS104, MA101	CS105, MA102, MA103	MA104	MA201,CS106
1237843	CS101, MA101, MA102	CS103, MA103	CS105	CS107
1238843	CS106	MA101, MA102	CS101, CS102, CS103	CS104, MA201
1240834	MA101, MA102	CS101, CS102, CS103	CS104, CS105	MA201, CS106
1243984	CS101, CS102, CS103	MA101	MA104, CS107	MA201, CS106
1245543	MA101, CS101, CS102, CS103	CS107	CS106	MA102
1247509	CS101, CS103, MA101	MA103	CS105	CS106
1256832	CS101, MA101	MA102, MA103	CS103, CS104	MA201
1256934	CS101, CS102, CS103	CS104, MA101	MA102, CS105	MA201, CS106
1257905	MA101	CS104, MA104	CS102	CS101, CS106, CS103

TODO

Find all sequences of courses with minsup = 0.75

Full list of courses:

CS101
CS102
CS103
CS104
CS105
CS106
CS107
MA101
MA102
MA103
MA104
MA201

Task: Study Histories

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

1-sequences	
CS101	10
CS102	6
CS103	10
CS104	6
CS105	5
CS106	8
CS107	3
MA101	10
MA102	7
MA103	4
MA104	3
MA201	6

2-sequences			
{CS101, CS103}	8	{CS103, CS106}	1
{CS101}, {CS103}	2	{CS103}, {CS106}	6
{CS103}, {CS101}	0	{CS106}, {CS103}	1
{CS101, CS106}	1	{CS103, MA101}	3
{CS101}, {CS106}	6	{CS103}, {MA101}	2
{CS106}, {CS101}	1	{MA101}, {CS103}	5
{CS101, MA101}	5	{CS106, MA101}	0
{CS101}, {MA101}	2	{CS106}, {MA101}	1
{MA101}, {CS101}	3	{MA101}, {CS106}	7

No 3-sequences possible as we have only one frequent 2-sequence.

Time Series Prediction
Component Models

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

A **time series** can consist of four components:

- Long - term trend (T_t)
- Cyclical effect (C_t)
- Seasonal effect (S_t)
- Random variation (R_t)

this is what we want to find

we need to eliminate those

Additive Model:

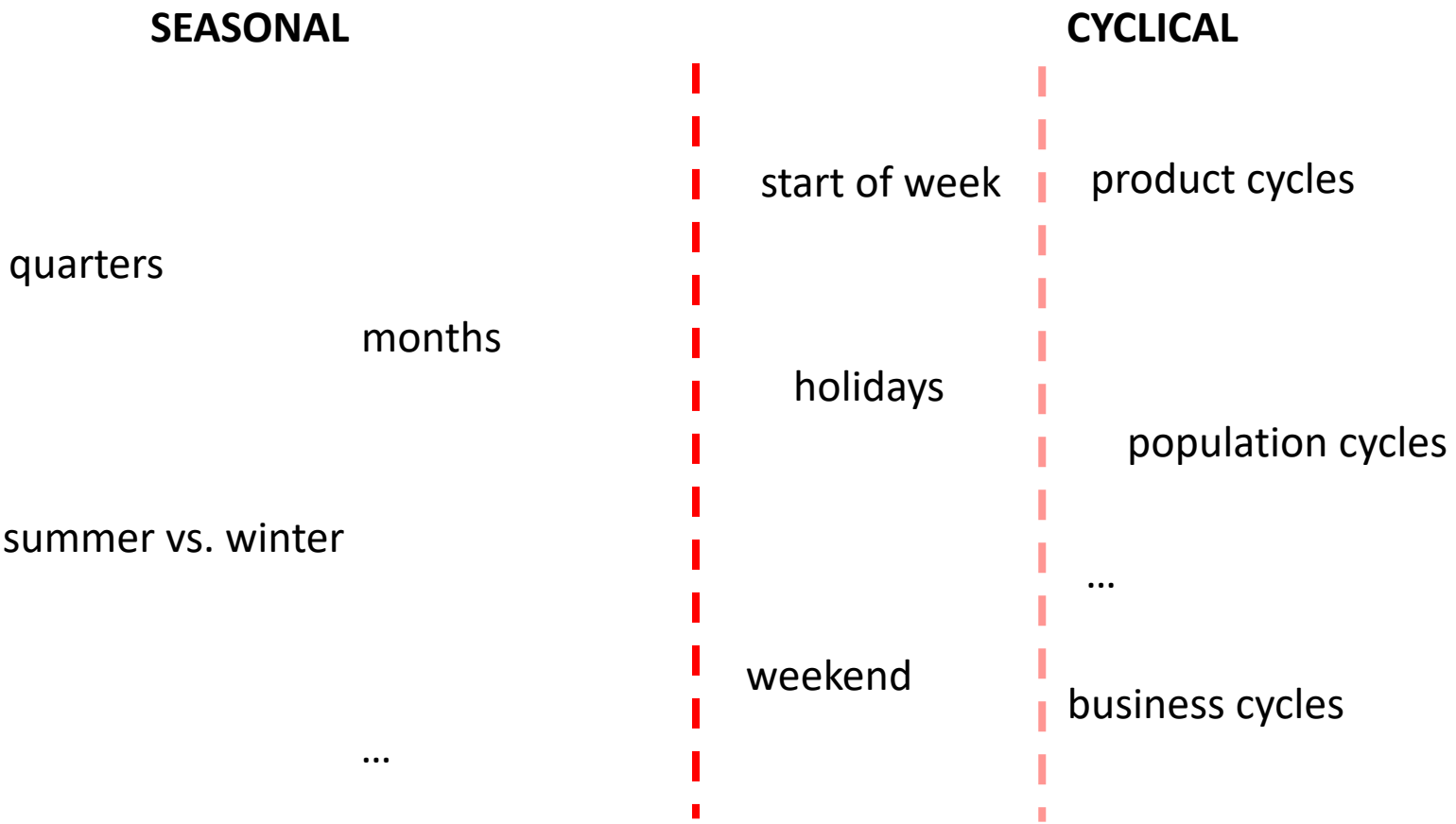
- Series = $T_t + C_t + S_t + R_t$

Multiplicative Model:

- Series = $T_t \times C_t \times S_t \times R_t$

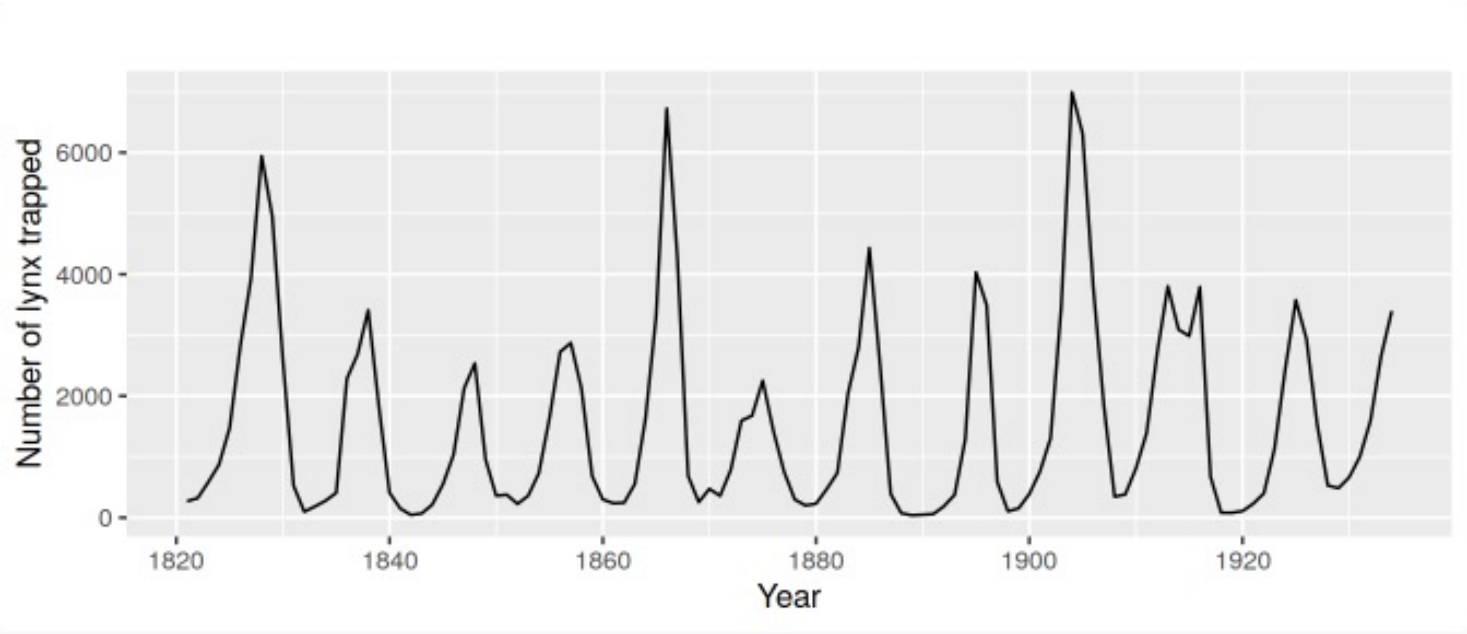
Time Series Prediction

Seasonal vs. Cyclical Effects



Time Series Prediction

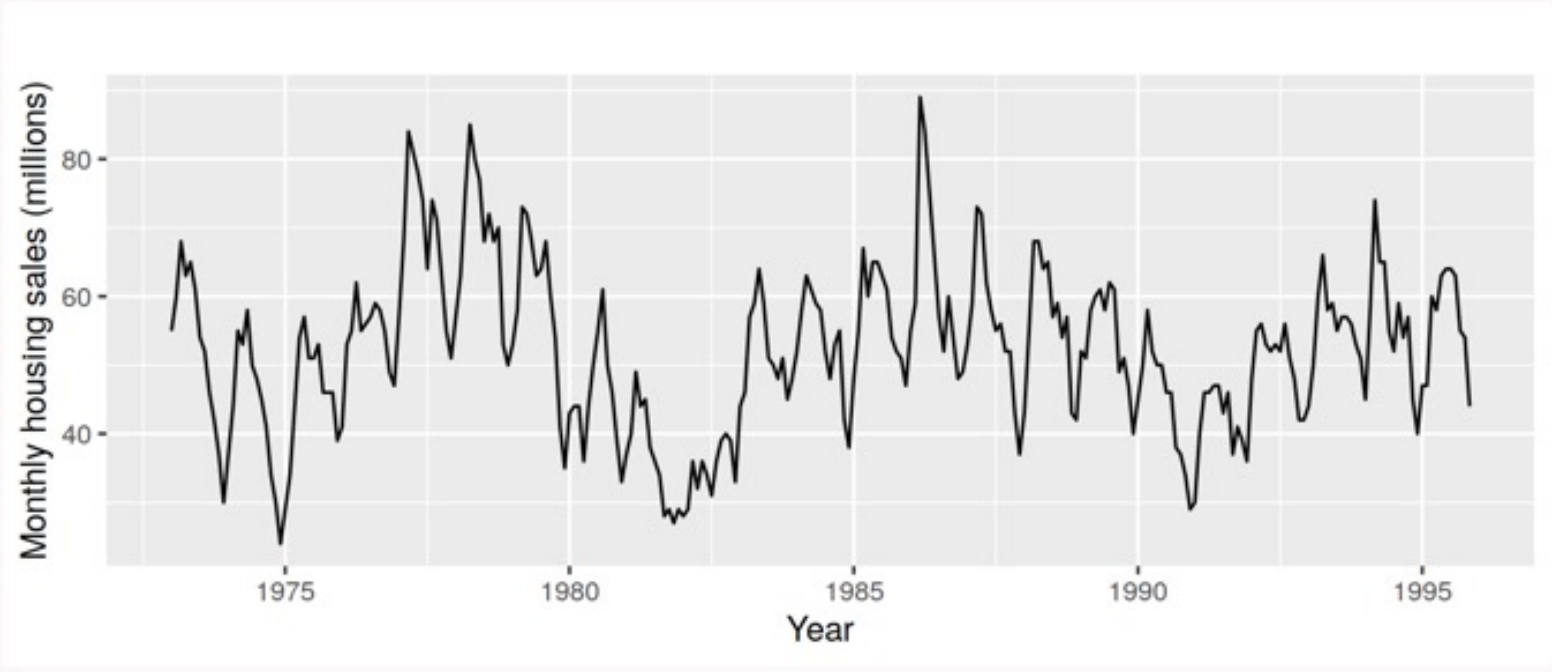
Seasonal vs. Cyclical Effects



CYCLICAL

Time Series Prediction

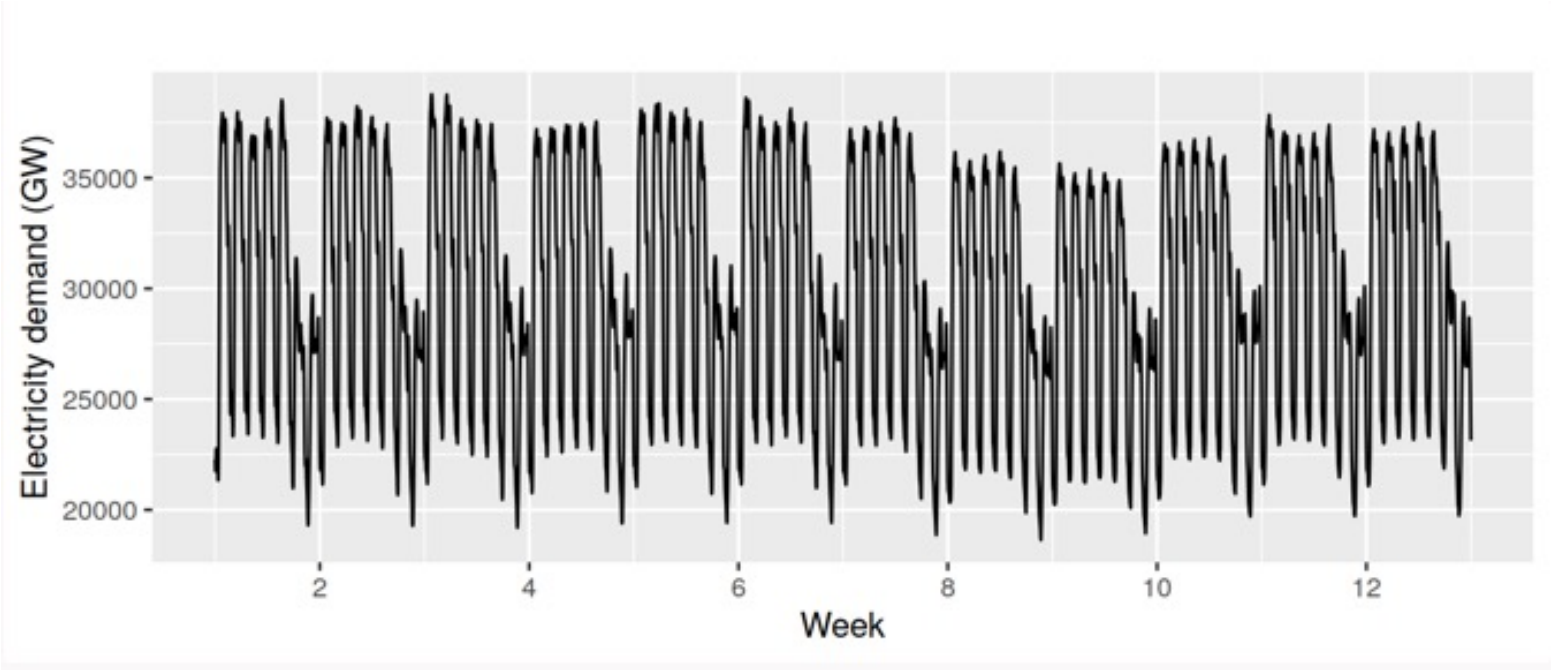
Seasonal vs. Cyclical Effects



SEASONAL & CYCLICAL

Time Series Prediction

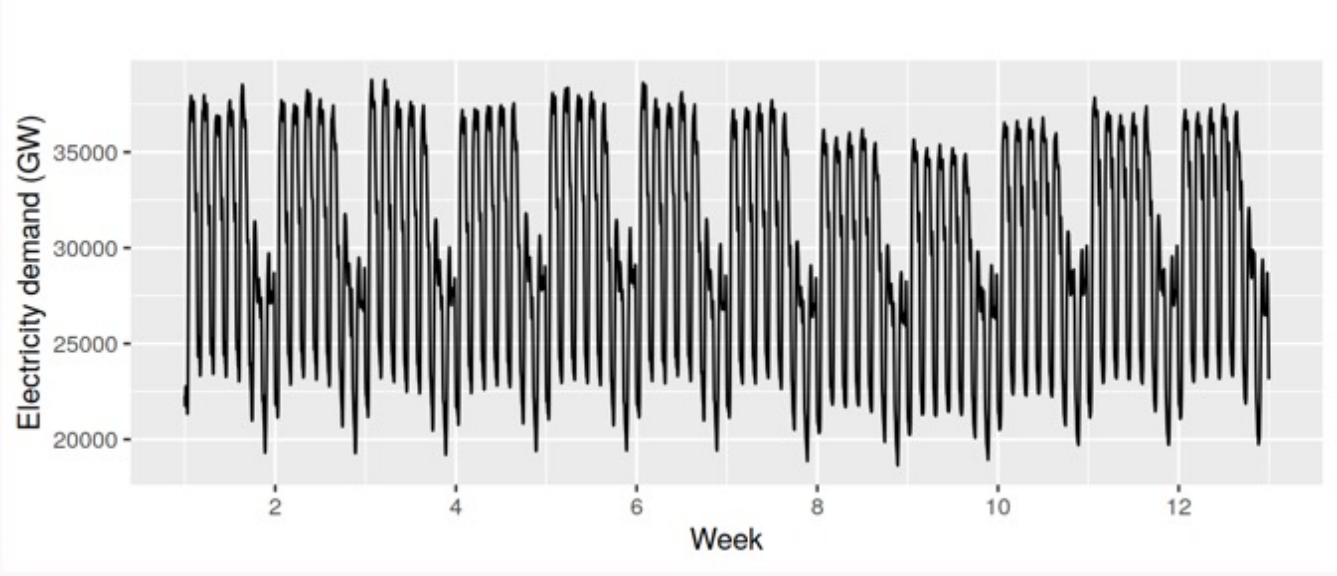
Seasonal vs. Cyclical Effects



CYCLICAL
(on the long term also SEASONAL)

Time Series Prediction

Smoothing



From the given dataset, you get one datapoint per day which is the peak electricity demand. You want to create a predictor for the peak electricity demand at a given day *t*.

TASK:

(1) If using simple exponential smoothing, will the peak demand for a Saturday likely be overpredicted or underpredicted? Why?

(2) Would the Holt-Winters method solve this problem?

If yes, which cycle length should be chosen?

Time Series Prediction
Smoothing

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

TASK:
(1) If using simple exponential smoothing, will the peak demand for a Saturday likely be overpredicted or underpredicted? Why?
Simple exponential smoothing puts much weight on very recent observations, so it would likely overpredict the demand for the Saturday.

(2) Would the Holt-Winters method solve this problem?
If yes, which cycle length should be chosen?
As it is a cyclic time series, Holt-Winters method is a perfect fit. It explicitly introduces a term for cyclic predictions.
A cycle length of 7 would make sense.