Dummy variables to handle outliers and special events

Outliers

Contents

REVIEW METHODS OF HANDLING OUTLIERS

INTRODUCE DUMMY VARIABLES

Methods for handling outliers

PRE-PROCESS THE DATA

IMPUTE THE OUTLIERS

MODELLING

MODEL THE OUTLIER USING FEATURES

Pre-processing outliers

- Treat outlier as missing data and impute the value.

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	**70**
2020-02-15	30
2020-02-16	25

Pre-processing outliers

- Treat outlier as missing data and impute the value.

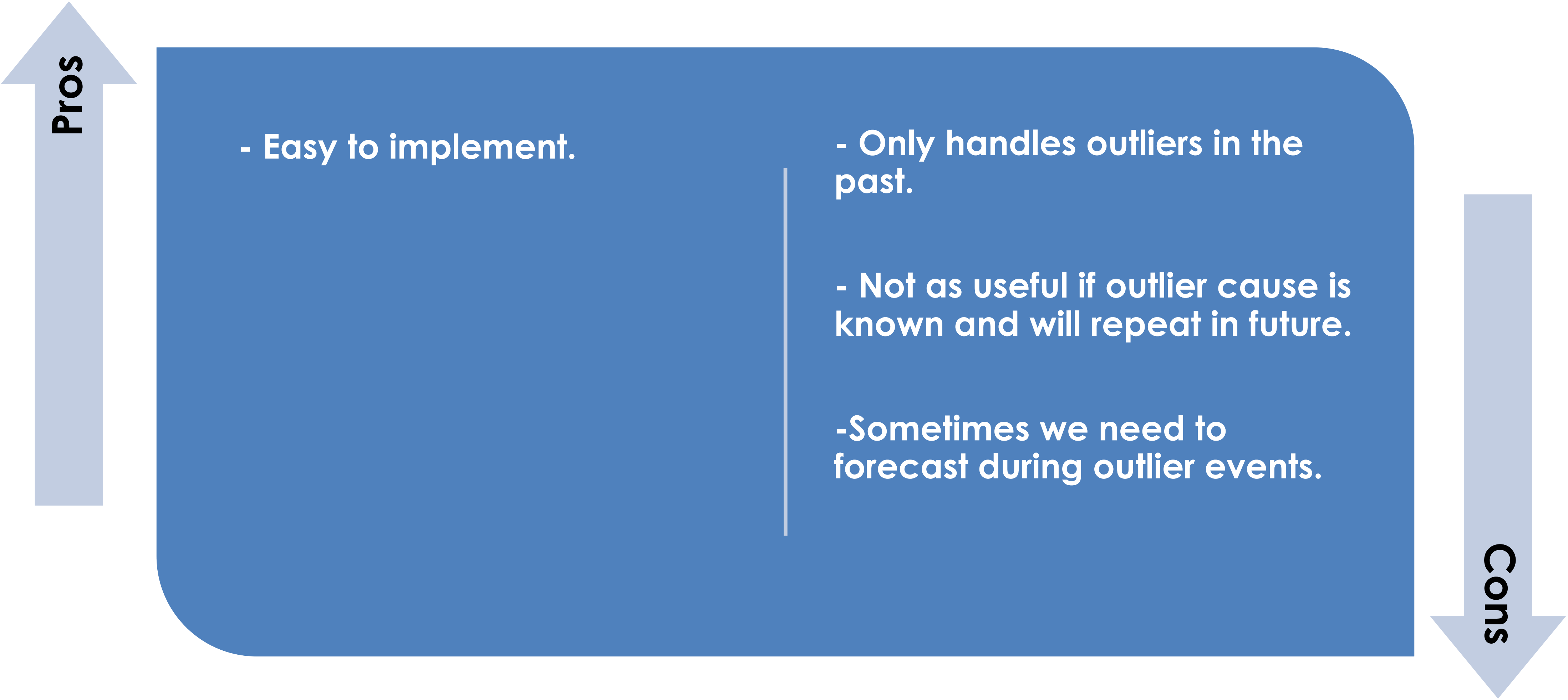
Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	**NaN**
2020-02-15	30
2020-02-16	25

Pre-processing outliers

- Treat outlier as missing data and impute the value.

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	**30**
2020-02-15	30
2020-02-16	25

Pre-processing outliers



- Easy to implement.

- Only handles outliers in the past.

- Not as useful if outlier cause is known and will repeat in future.

-Sometimes we need to forecast during outlier events.

Modelling outliers & special events

- It is important to understand the cause of the outlier (e.g., recording errors vs external event).
- Examples: Public holidays, marketing events such as Black Friday & Boxing Day sales, etc.
- As we know the future dates of these events it can be helpful to model the outlier.
- This can be done using a **dummy variable**.

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	**70**
2020-02-15	30
2020-02-16	25

Dummy variables

- A dummy variable (aka indicator variable) is a variable that only takes two values: 1 or 0.
- It is used to indicate the presence or absence of an effect.

Date	Sales
2020-02-12	23
2020-02-13	30
2020-02-14	70
2020-02-15	30
2020-02-16	25

Dummy variables

- A dummy variable (aka indicator variable) is a variable that only takes two values: 1 or 0.

- It is used to indicate the presence or absence of an effect.

Date	Sales	Is_valentines
2020-02-12	23	0
2020-02-13	30	0
2020-02-14	70	**1**
2020-02-15	30	0
2020-02-16	25	0

Dummy variables

- A dummy variable (aka indicator variable) is a variable that only takes two values: 1 or 0.

- It is used to indicate the presence or absence of an effect.

Date	Sales	Is_valentines	Is_public_hol
2020-02-12	23	0	0
2020-02-13	30	0	0
2020-02-14	70	**1**	0
2020-02-15	30	0	0
2020-02-16	25	0	0

Dummy variables

- A dummy variable (aka indicator variable) is a variable that only takes two values: 1 or 0.
- It is used to indicate the presence or absence of an effect.

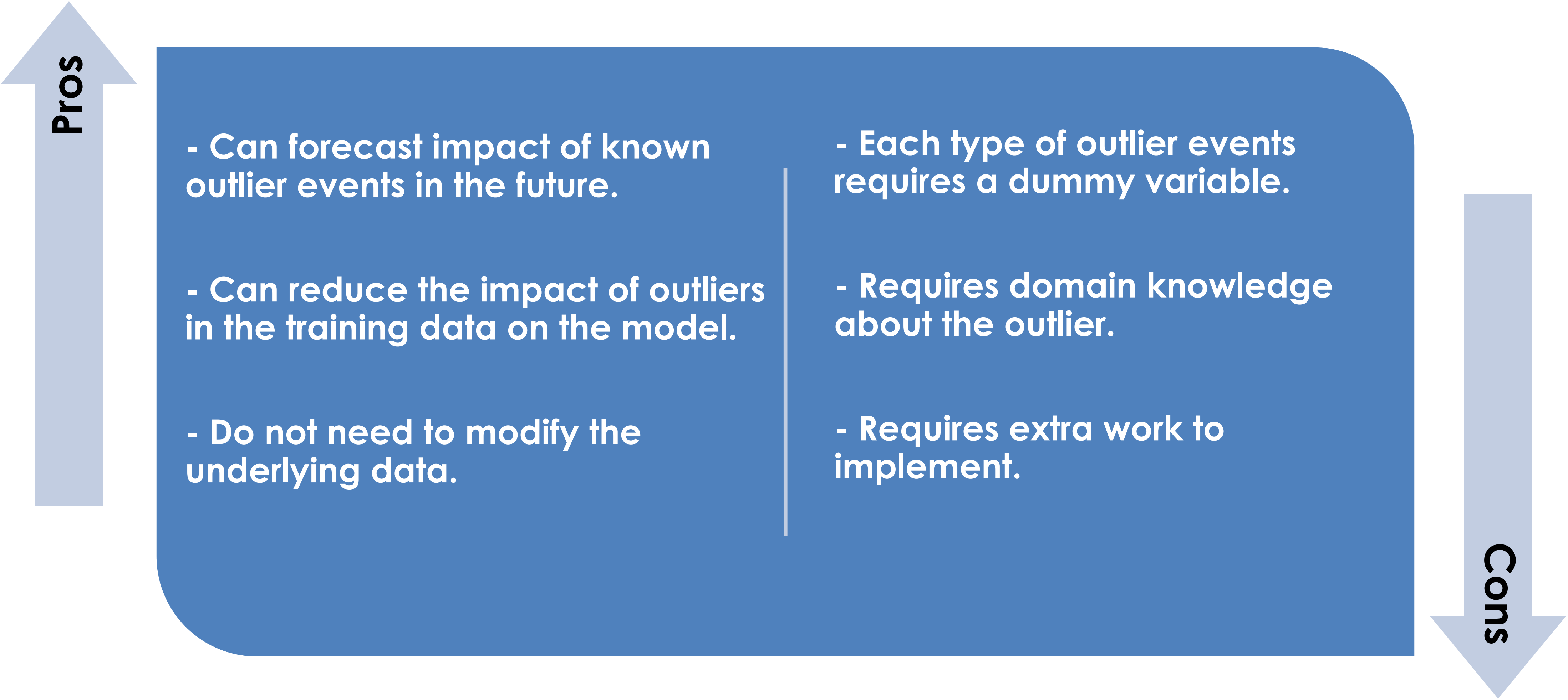
Date	Sales	Is_valentines	Is_public_hol
2020-02-12	23	0	0
2020-02-13	30	0	0
2020-02-14	70	1	0
2020-02-15	30	0	0
2020-02-16	25	0	0
…	…	…	…
2020-02-22	5	0	**1**

Dummy variables

- A dummy variable (aka indicator variable) is a variable that only takes two values: 1 or 0.
- It is used to indicate the presence or absence of an effect.

Date	Sales	Is_valentines	Is_public_hol	Is_promo_period
2020-02-12	23	0	0	**1**
2020-02-13	30	0	0	**1**
2020-02-14	70	1	0	**1**
2020-02-15	30	0	0	0
2020-02-16	25	0	0	0
…	…	…	…	…
2020-02-22	5	0	1	0

Dummy variables



Pros

- Can forecast impact of known outlier events in the future.

- Can reduce the impact of outliers in the training data on the model.

- Do not need to modify the underlying data.

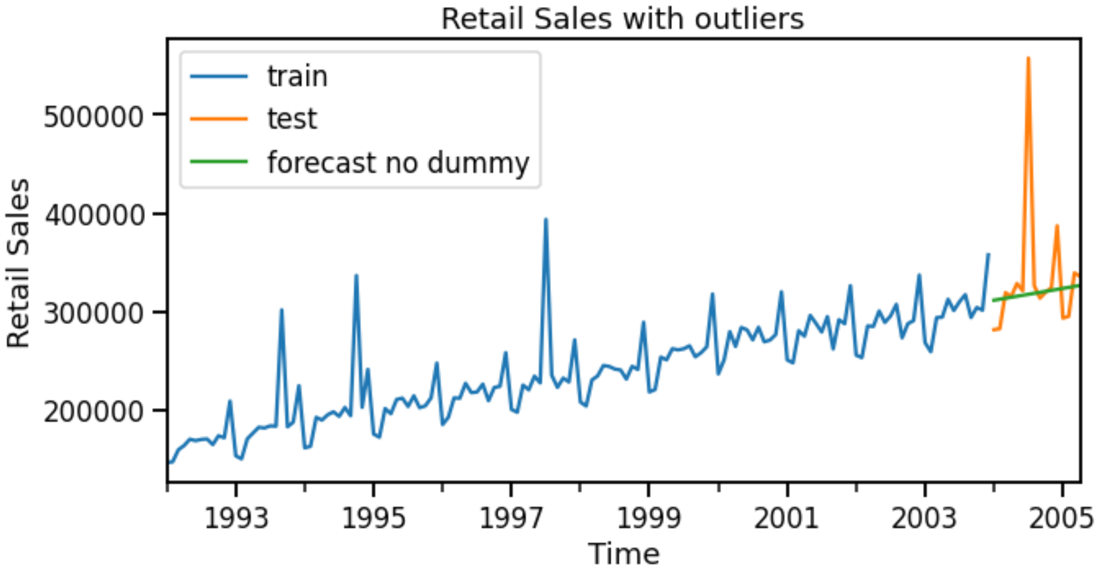
Cons

- Each type of outlier events requires a dummy variable.

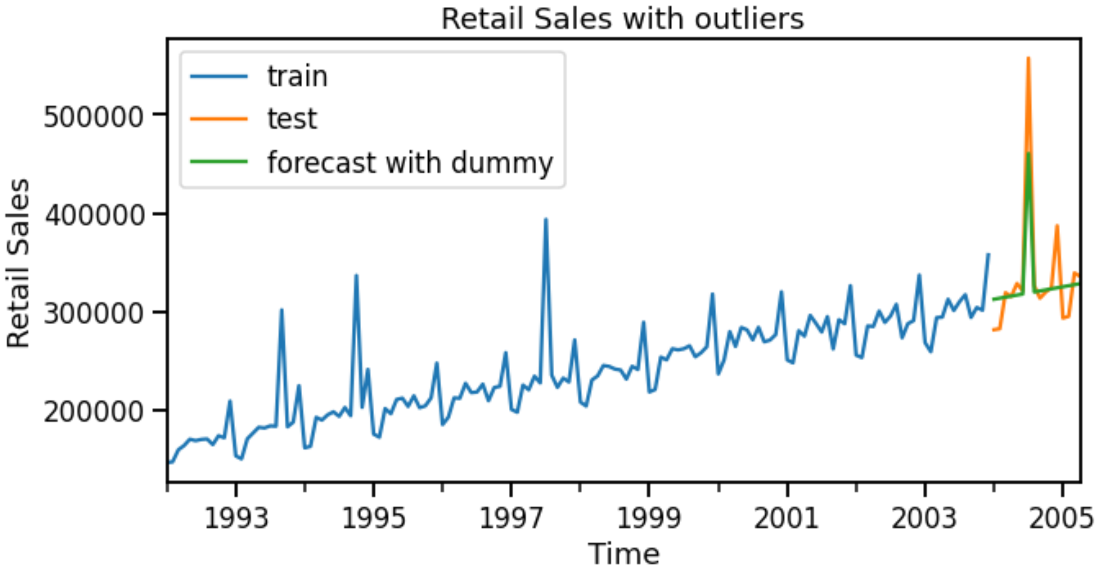
- Requires domain knowledge about the outlier.

- Requires extra work to implement.

Example: Modelling retail sales



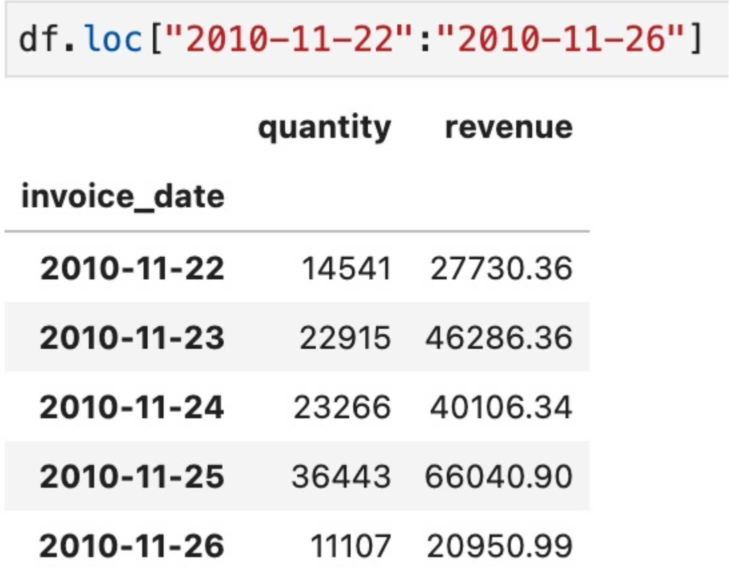
Model trained on just a trend feature.



Model trained on trend & is_outlier feature.

Implementation

- We want to convert a set of known dates into a binary variable.

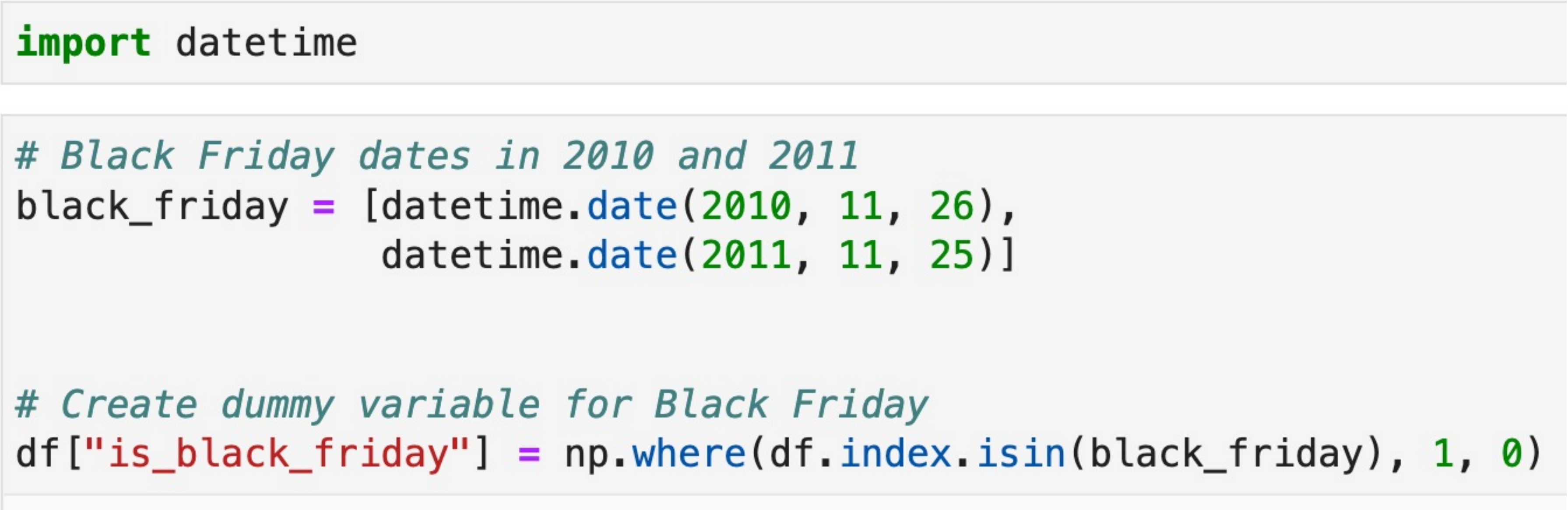


```
df.loc["2010-11-22":"2010-11-26"]
```

	quantity	revenue
invoice_date		
2010-11-22	14541	27730.36
2010-11-23	22915	46286.36
2010-11-24	23266	40106.34
2010-11-25	36443	66040.90
2010-11-26	11107	20950.99

Implementation

- We want to convert a set of known dates into a binary variable.



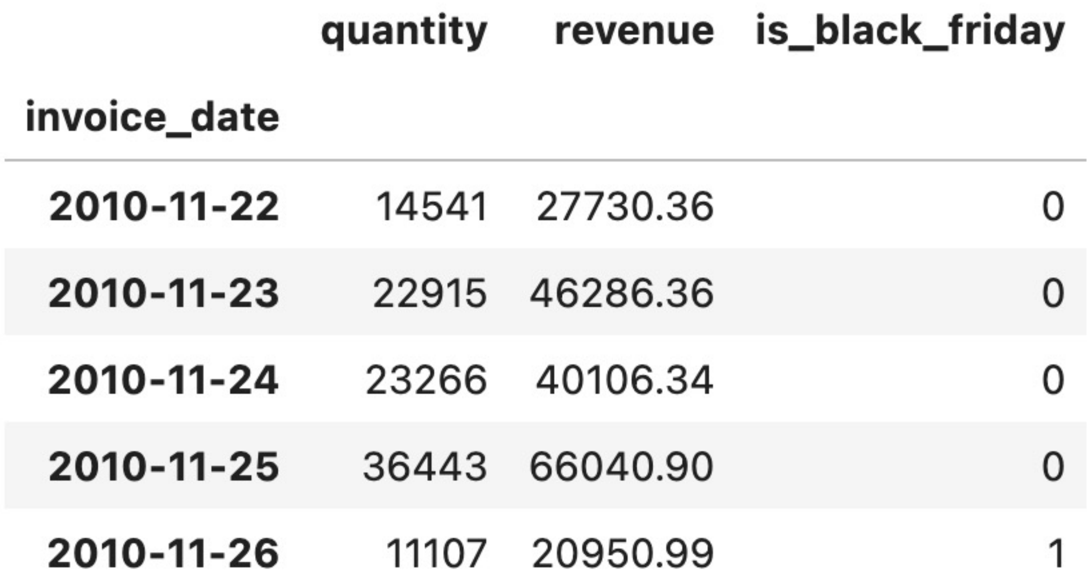
```
import datetime
```

```
# Black Friday dates in 2010 and 2011
black_friday = [datetime.date(2010, 11, 26),
                datetime.date(2011, 11, 25)]


# Create dummy variable for Black Friday
df["is_black_friday"] = np.where(df.index.isin(black_friday), 1, 0)
```

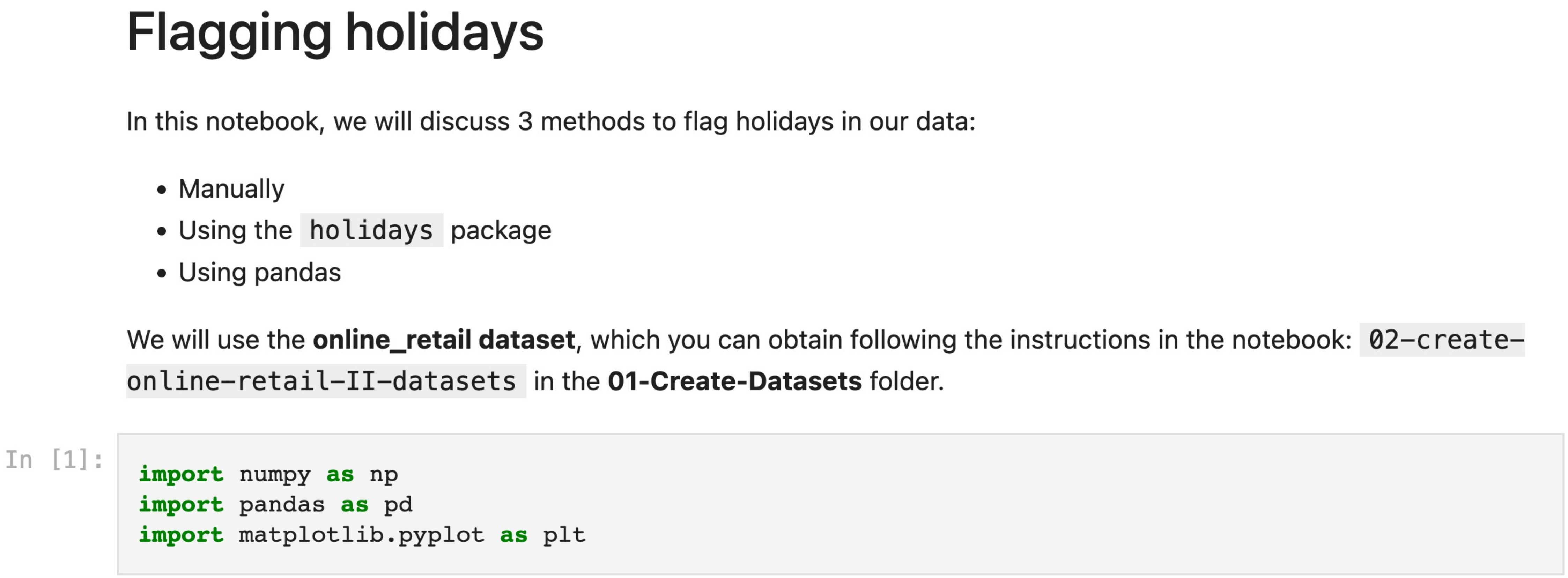
Implementation

- We want to convert a set of known dates into a binary variable.



	quantity	revenue	is_black_friday
invoice_date			
2010-11-22	14541	27730.36	0
2010-11-23	22915	46286.36	0
2010-11-24	23266	40106.34	0
2010-11-25	36443	66040.90	0
2010-11-26	11107	20950.99	1

See “highlighting-holidays” in Time Features section



Flagging holidays

In this notebook, we will discuss 3 methods to flag holidays in our data:

- Manually
- Using the `holidays` package
- Using pandas

We will use the **online_retail dataset**, which you can obtain following the instructions in the notebook: `02-create-online-retail-II-datasets` in the **01-Create-Datasets** folder.

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Summary

If the outlier cause is known we can use this to improve our forecasts.

Dummy variables can be used to model the impact of outliers.

Dummy variables can remove the impact of an outlier on a model.