Data Science in Business:

Introduction to Machine Learning in Python

Dr. Peter Molnar

Sarah Zeis

Carly Wieting

Course Overview

Class 1: Introduction to Machine Learning and Set-Up Python

Class 2: Data Exploration

Class 3: Machine Learning Models (Decision Tree and KNN)

Class 4: Analyze Celebrity Tweets

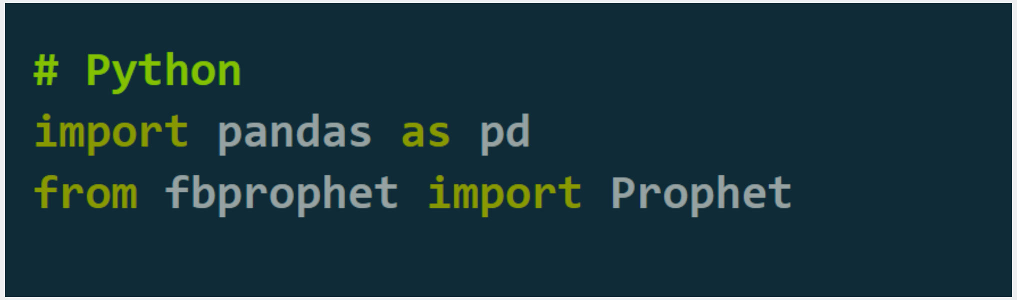
Class 5: Forecasting with Facebook Prophet

What is forecasting?

Forecasting (AKA time series analysis) is a complex phenomenon that uses historical data inputs that are predictive of future trends. This can improve the business’ strategy, planning, budget, etc…

Facebook Prophet

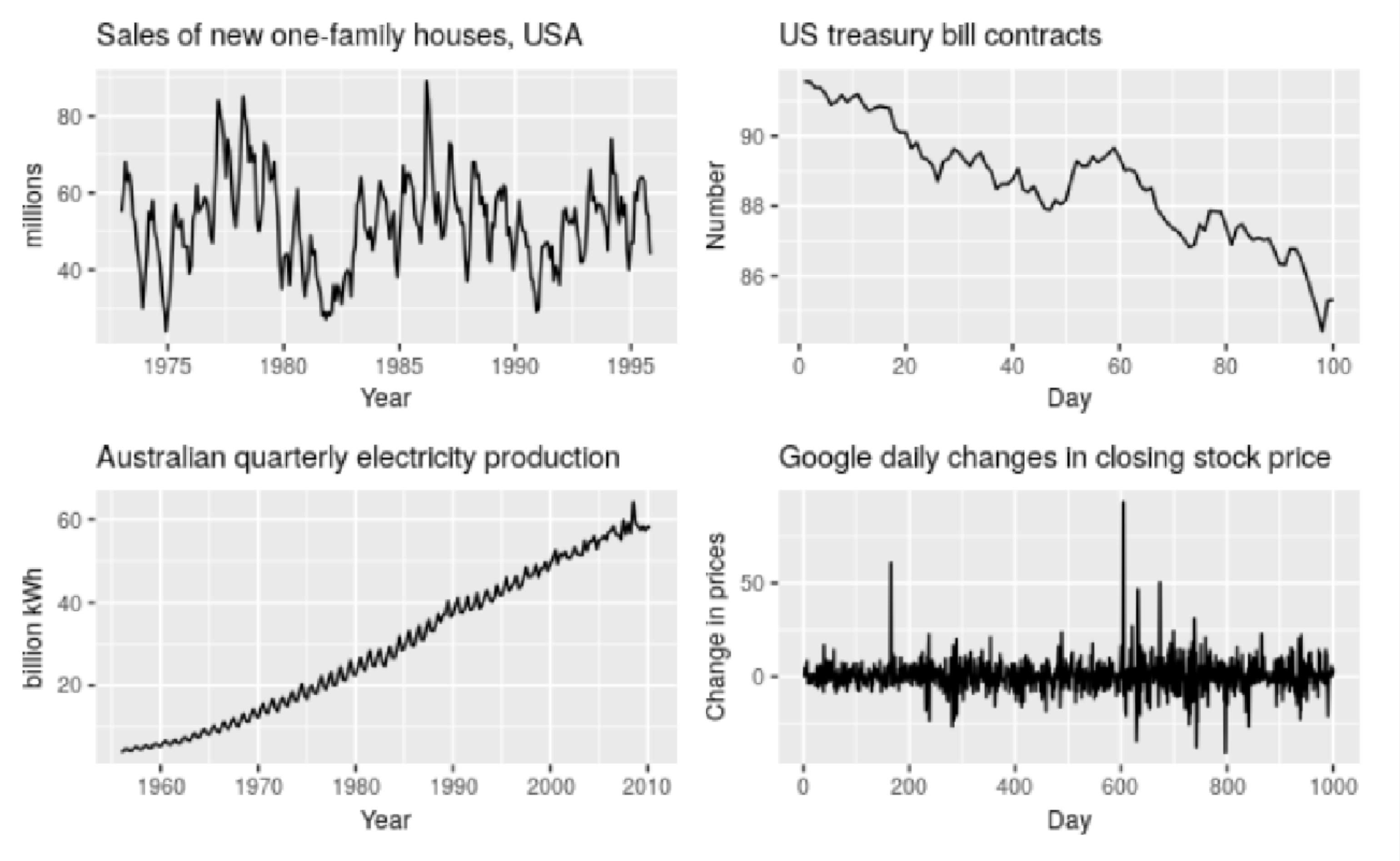
- Invented by Facebook’s data science team
- Automates some features and can be easily tuned by people of a variety of backgrounds



```
# Python
import pandas as pd
from fbprophet import Prophet
```

Uses of forecasting

- **Sales/Revenue**
- **Demand**
- **Stock prices**
- **Next day of customer purchase**
- Earthquakes
- Weather forecasting



Linear Regression vs. Forecasting

$y = m*x + b + error$	$x_t = m*x_{t-1} + b + error$
x is some feature used to predict y	x is past values, y is future values
Example is using age to predict income	Example is using past sales to predict future sales

Outline of Forecasting

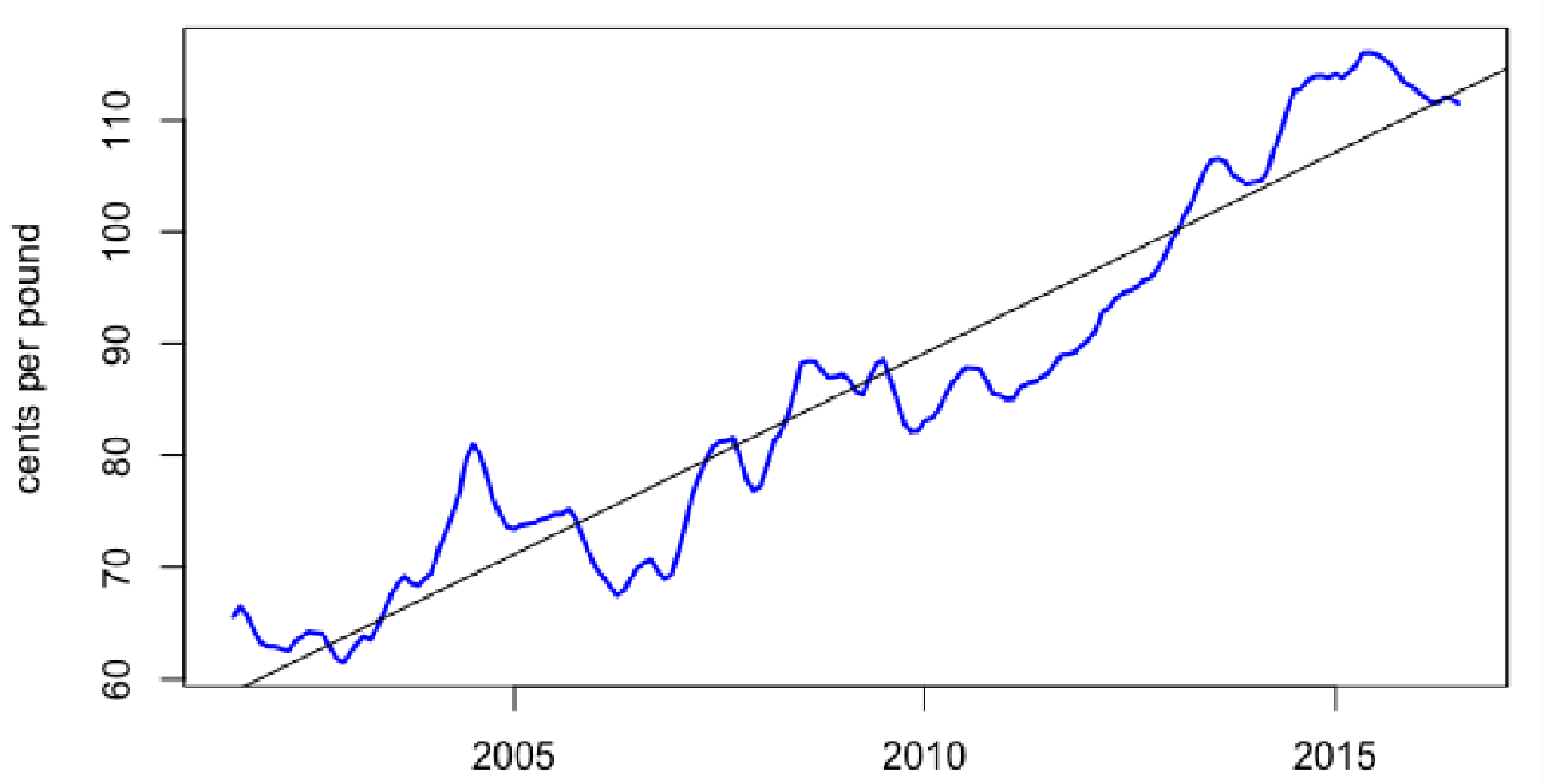
1. Visualize/plot the data
2. Make the data stationary
3. Transformation
4. Estimation
5. Forecasting and Prediction
6. Model Diagnostics

Visualizing the data

After plotting the data, you can see there is an upwards trend

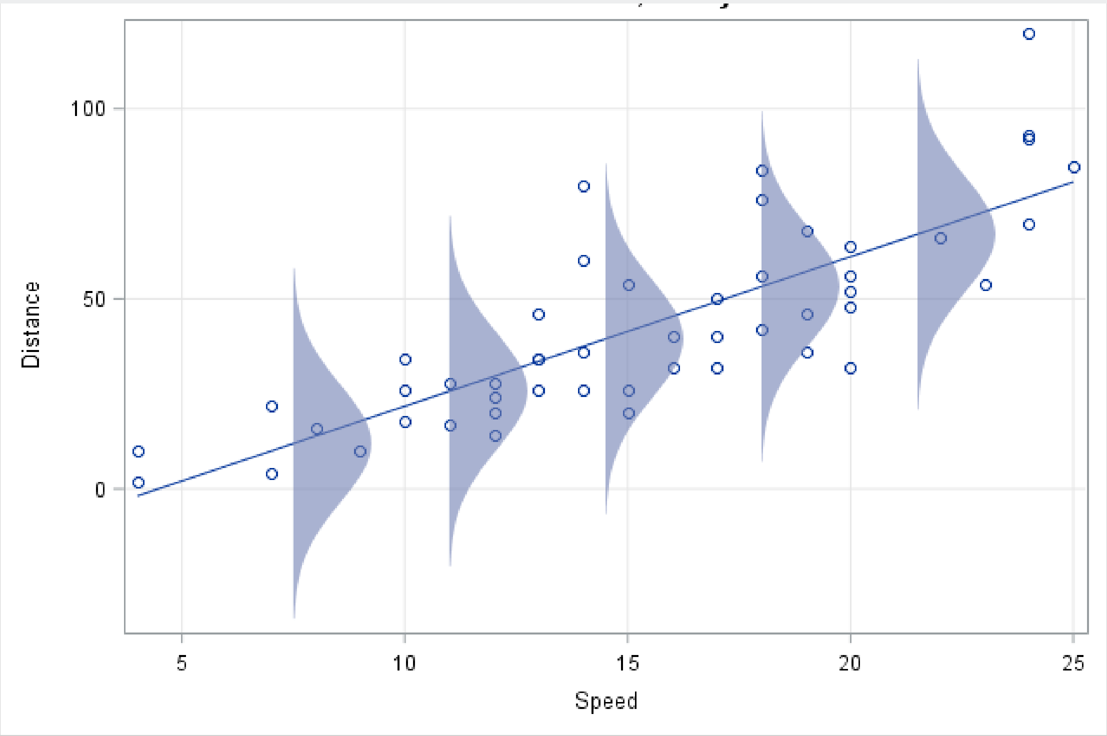
This means the data is not stationary

Take a difference of the data to remove the difference



Assumptions of the Model

- The errors are completely random (no correlations)
- The error is normally distributed



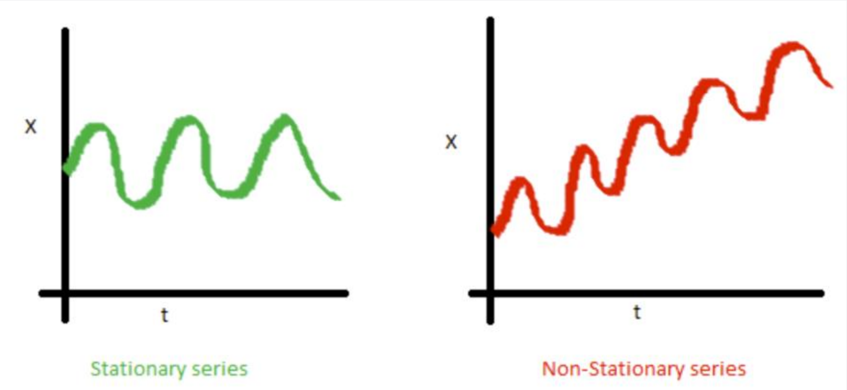
Stationary Forecasting

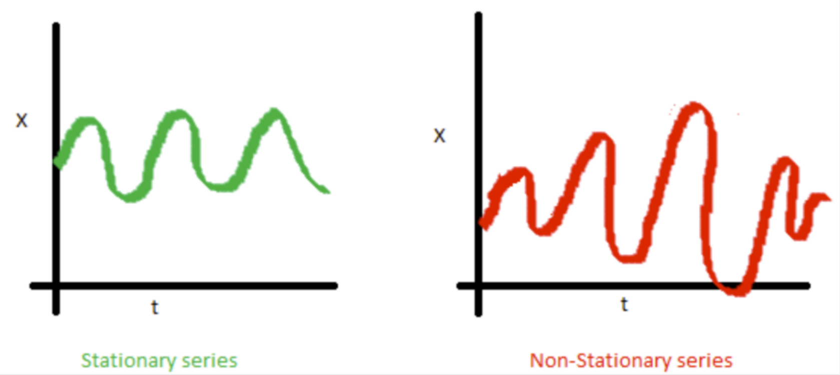
You cannot do time series analysis without stationary data. Fortunately, this is only a few lines of code in Python

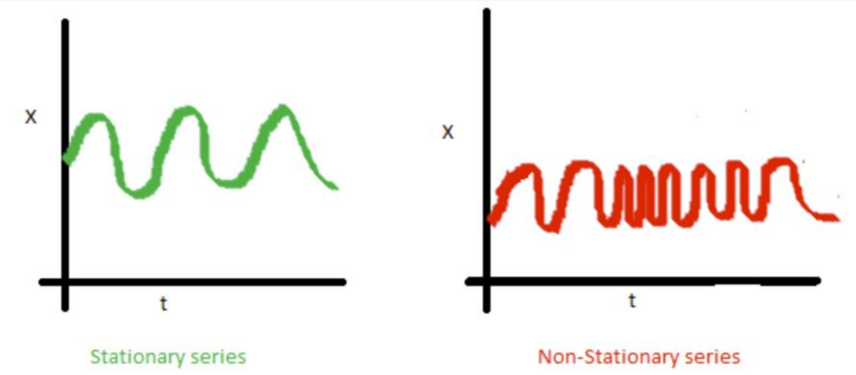
In the first graph, we can make the time series flat as opposed to having an upwards or downwards trend (expected value(x) = 0).

In the second graph, we want to make sure the barriers of the time series data are in a consistent range throughout (variance(x) = 0).

In the third graph, we want to make sure there waves are spread out evenly (covariance(x_t, x_{t+s}) = 0).





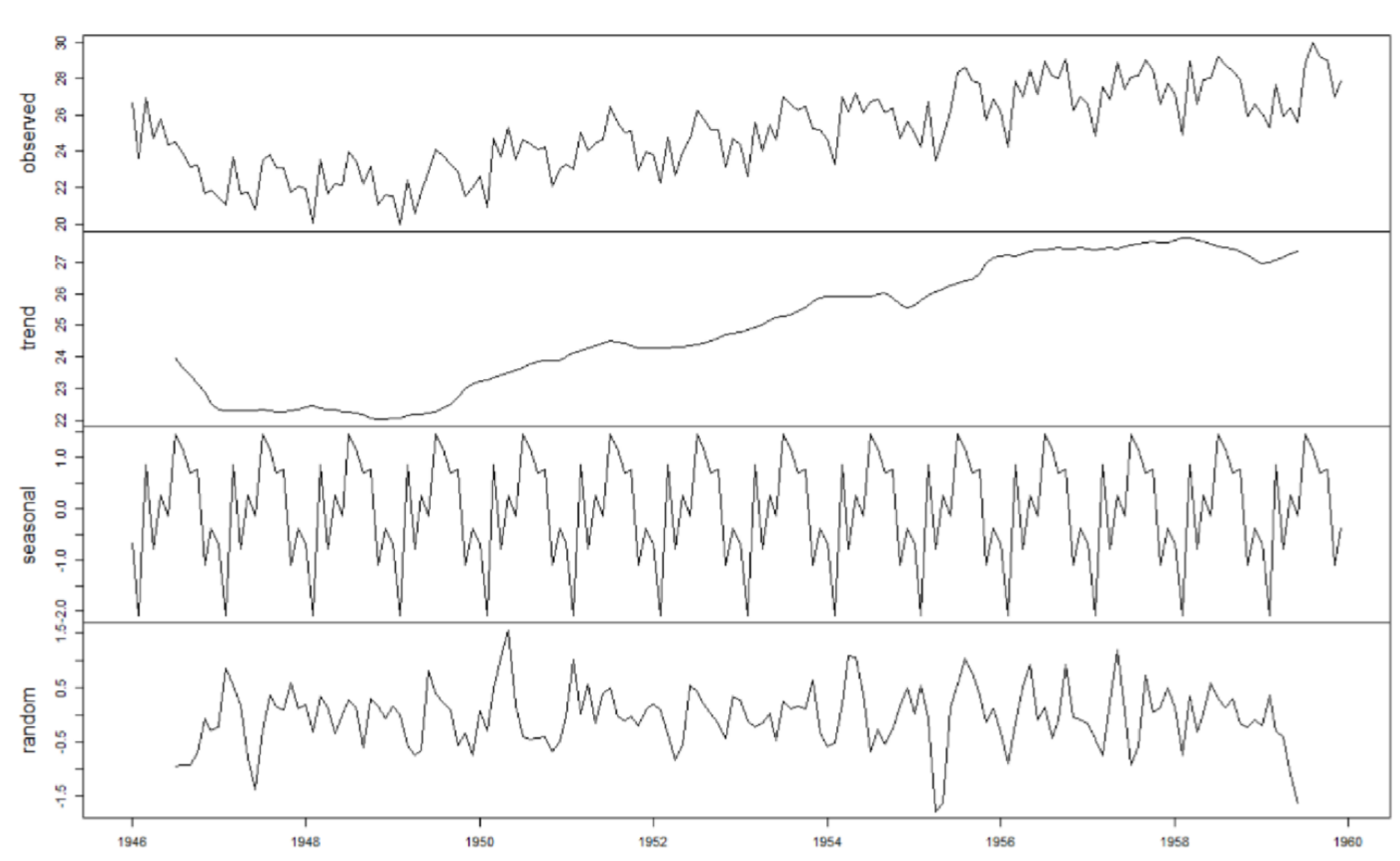


Decomposition of Data

The first graph shows the raw data

You can decompose the data into three components:

1. Trend
2. Seasonality
3. Residuals

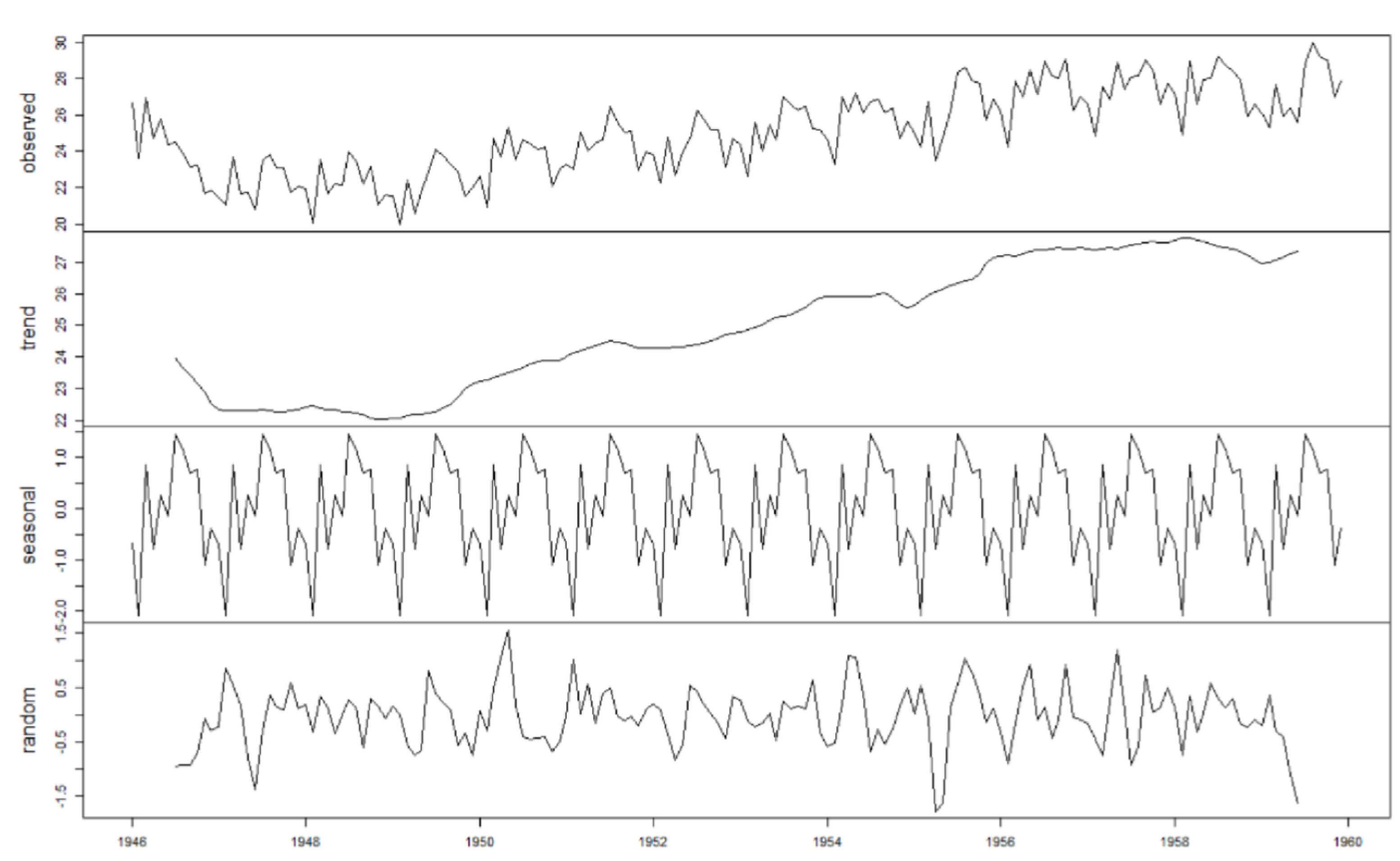


Transformation of the Data

The trend can be removed by taking the difference.

This simply subtracts previous days values from the current days

This results in a combination of the seasonality and random components



GeorgiaState University | J. MACK ROBINSON COLLEGE OF BUSINESS

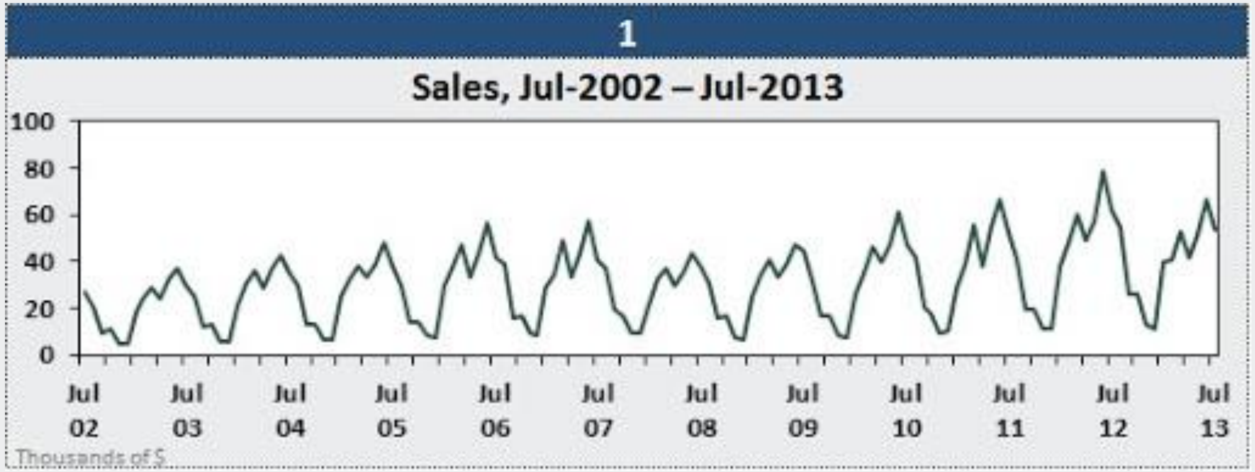
Robinson

What is seasonality?

Seasonality is the variations that occur at regular time intervals, such as hourly, daily, monthly, yearly

It usually relates to the business cycle. For example, you would expect the most sales in December around the holidays/Black Friday.

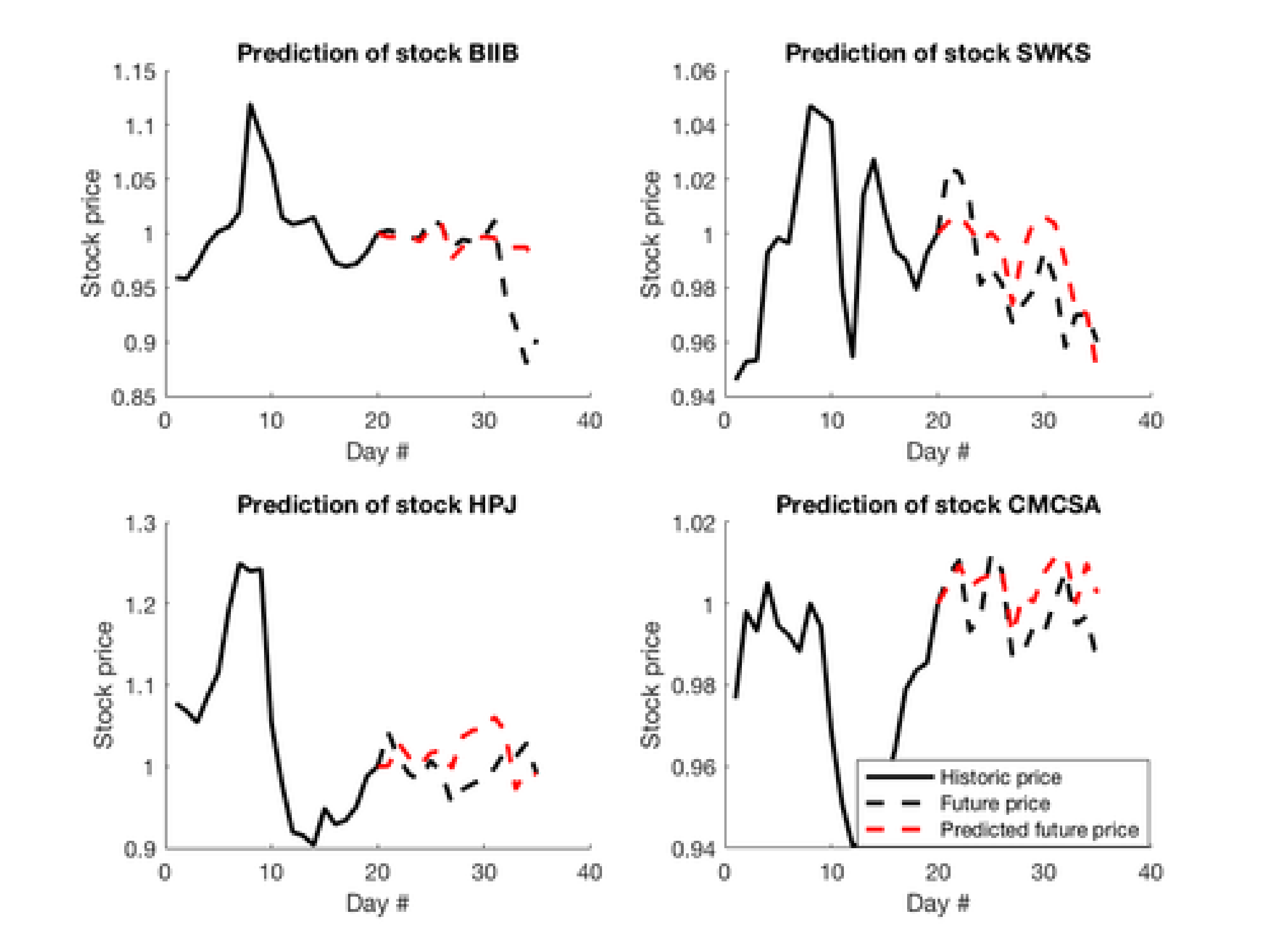
What are some other seasonalities in business?



Forecasting and Prediction

You can take your dataset and create the test and train set by splitting. Then do the prediction on the test set and see how accurate the model is.

Normally, you predict one day ahead, then get the rest of the future values recursively.

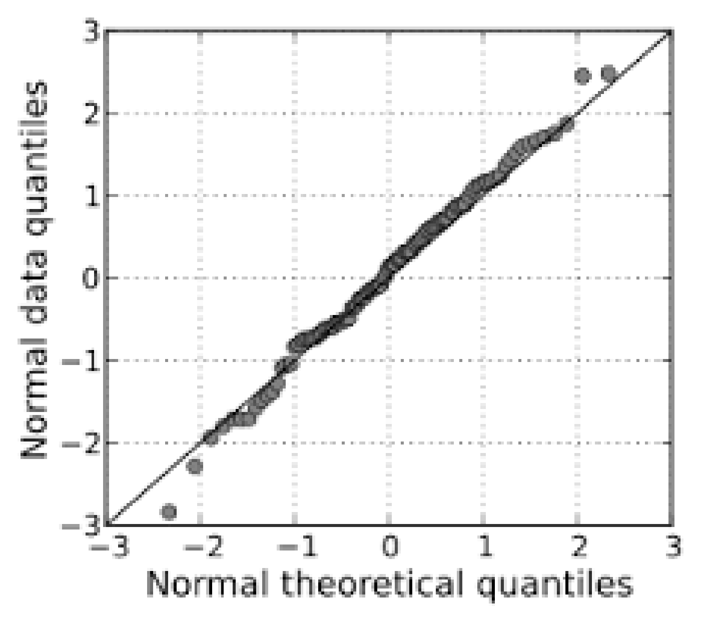


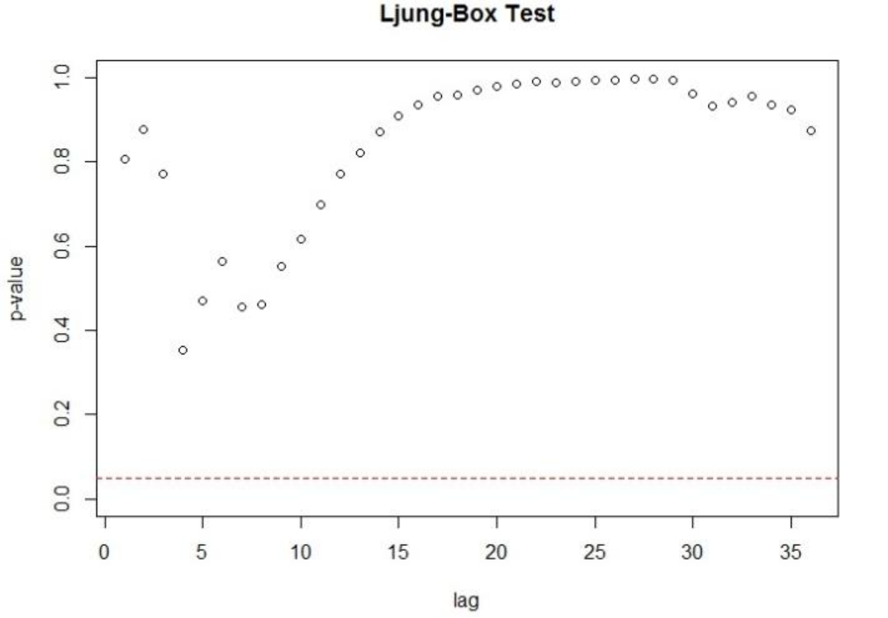
Model Diagnostics

Normal Q-Q plot and Ljung-Box Statistics

Normal Q-Q plot makes sure the residuals are normally distributed. Just make sure all of the points are all in a line and don't deviate significantly.

Ljung-Box Statistics checks if there are correlations between each of the points. Just make sure all of the points are above the dotted line and ignore any patterns of the points.





Additional References

Facebook Prophet Documentation:

https://facebook.github.io/prophet/docs/quick_start.html#python-api

Facebook Prophet Github:

https://github.com/facebook/prophet

Introduction to ARIMA models:

https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

It’s Python time!