Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science 2

Sommersemester 2021

Übungsblatt 4

Aufgabe 1 * (Verkaufszahlen)

Die Firma Rossmann hat einen Datensatz zu ihren tagesbasierten Umsätzen von Januar 2013 bis Juli 2015 veröffentlicht. Der Datensatz findet sich unter anderem auf der Kaggle Plattform als Beispiel zur Umsatzvorhersage.
Sie finden den Datensatz unter:

`Kurse/DataScience2/data/rossmann-train.csv`

In dieser Aufgabe geht es darum, dass Sie sich nochmal etwas mit der Zeitreihen-Darstellung von Pandas auseinandersetzen.

1. Lesen Sie den Datensatz in einen DataFrame ein und schauen Sie sich die verfügbaren Spalten an. Erzeugen Sie daraus einen DateFrame, der die mittleren Umsätze pro Wochentag enthält und plotten Sie diesen durchschnittlichen Verlauf.

2. Erstellen Sie eine Series mit der Spalte `Sales` und dem Datum als Index. Der Index sollte natürlich im `datetime64` Format sein (Sie erinnern sich ja sicherlich noch an die Funktion `pd.to_datetime(..)`).

3. Benutzen Sie `resample` um den durchschnittlichen monatlichen Umsatz zu berechnen und erstellen Sie einen Plot dazu.

4. Erzeugen Sie ein Trend-Modell indem Sie eien lineare Regression für den durchschnittlichen monatlichen Umsatz berechnen. Erstellen Sie einen Plot mit monatlichen Umsätze und der Vorhersage ihrer linearen Regression.

   Hinweis: Wenn Sie für zwei Series-Objekte die `.plot()` Funktion in einer Codezelle ausführen, werden die beiden Series-Objekte im gleichen Plot erzeugt.

5. Subtrahieren Sie ihren Monats-Trend von der ursprünglichen monatlichen Durchschnittsfolge und plotten Sie das Ergebnis. Was fällt Ihnen auf?
   Wann sind die monatlichen Peak-Zeiten bezogen auf den Umsatz?

6. Erstellen Sie ein lineares Regressionsmodell für den durchschnittlichen Wochenverlauf der Verkäufe bei Rossmann. Nutzen Sie dazu auch Funktionen vom Grad 2 oder höher. Plotten Sie den Wochenverlauf und ihr Regressionsmodell.

7. Schreiben Sie eine Schleife um über die Wochen zu iterieren und berechnen Sie den durchschnittlichen Trainingsfehler ihres Wochenregressionsmodells auf den ersten `k` Wochen des Datensatzes.

   (`k` können Sie hier beliebig wählen.)

Die Teilaufgaben 6. und 7. sind als Zusatzaufgaben gedacht.