Probeaufgabe Data Scientist

Nachdem wir erfolgreich die Speisekarten von mehr als 120.000 Gastronomien sammeln konnten, ist eine der aktuellen Herausforderungen der Aufbau einer Datenpipeline, die die Datensätze bereinigt und über unsere API in die Datenbank einspielt.

Ein besonderes Augenmerk liegt hierbei auf den Preisen der Speisen und Getränke, welche durch die verschiedenen Datenquellen in der Regel nicht normalisiert sind. So können die Preise unter Umständen folgenden Formate annehmen:

5,50€
5,50 €
5.50 €
5.50 €
5.50,-
5.50
5,50
550
...

Nicht selten schleichen sich hier auch Fehler bei der Extraktion ein, sodass das Komma oder der Punkt nicht korrekt erkannt wird und der Preis statt 5,50€ 550€ beträgt.

In diesem Ordner findest du eine Sammlung von extrahierten Speisekarten (~1600 Stück). Die Aufgabe besteht nun darin, eine Funktion / Methode in Python zu schreiben, welche die Datensätze nimmt, die Preise bereinigt und normalisiert, sodass der Preis stets in Cent angeben wird (z.B. 5,50€ = 550).
Auf Basis dieses bereinigten Datensatzes soll dann die Preisverteilung von Coca Cola visualisiert werden. Dies kann als Excel, als matplot Graph oder einem anderen Medium deiner Wahl geschehen. Es geht hierbei in erster Linie darum, einen Überblick zu erhalten.

Sollten Ausreißer in den Werten oder Datensätzen auftreten, sind diese ebenfalls zu bereinigen, zu exkludieren oder zumindest zu taggen. Ausreißer könnten bspw. ein Cola Preis von 20cent oder 20€ sein, also Werte, die eindeutig falsch sind. Außerdem können manche Speisekarten unvollständig oder leer sein.