Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science 1

Wintersemester 2020/2021

Aufgabenstellung zur Hausarbeit (Probelauf)

Die Prüfungsleistung zum Modul *Data Science 1* findet als Hausarbeit statt. Die Aufgabenstellung zur Hausarbeit finden Sie in diesem Dokument.

Für die Bearbeitung der Aufgabenstellung und die Erstellung Ihrer Hausarbeit steht wieder der Jupyter-Notebook Server zu Verfügung. Die Abgabe der Hausarbeit erfolgt dann als PDF-Export Ihres Jupyter-Notebooks. Das PDF Ihres Notebooks laden Sie als Lösung in der zugehörigen Aufgabe im Moodle Kurs hoch.
Im Verzeichnis

`Hausarbeit-Test`

in Ihrem Notebook-Account finden Sie ein `data/` Verzeichnis und ein leeres Notebook, dass Sie für die Bearbeitung dieser Hausarbeit nutzen können.

Andere Formen der Abgabe sind nicht vorgehen.

Für die Bearbeitung der Aufgabenstellung haben Sie ab Themenausgabe eine Woche Zeit. Der exakte Zeitraum wird in der zugehörigen Aufgabe im Moodle Kurs vermerkt.

Als Materialien können Sie sämtliche Unterlagen aus der Vorlesung und den Übungen mit benutzen, im Internet recherchieren oder weitere Bücher/Kurse mit verwenden. Geben Sie bitte bei Verwendung von umfangreicherem Programm-Code aus dem Netz (mehr als 3-4 Zeilen) die Quelle kurz mit an.

Aufgabenstellung

Wir betrachten im Folgenden einen Lieferdienst für Frischeprodukte. Ziel des Lieferdienstes sind möglichst kurze Wege bei der Auslieferung von Waren. Dazu verfügt der Anbieter über eine Reihe von Warenverteilzentren, in denen frische Ware gekühlt gelagert wird.
Orte werden im Folgenden jeweils über ein 2-dimensionales Koordinatensystem festgelegt, d.h. jedem Warenverteilzentrum ist ein `x` und ein `y` Wert zugeordnet.

Aufgabe 1 (Python Basics)

Das Python-Modul `lieferdienst` stellt die Funktionen `kunden()` und `verteilzentren()` zur Verfügung. Beide Funktionen liefern eine Liste von 3-Tupeln zurück, die jeweils eine eindeutige ID, die `x`- und die `y`-Koordinate enthalten, d.h. jedes Tupel hat die Form

```
(id, x, y)
```

Beispiel:

```
import lieferdienst

lieferdienst.kunden()
# [ (1, 0.5, 0.2), (2, 0.2, 0.3), (3, 0.4, 0.2),...]

lieferdienst.verteilzentren()
# [("DC1", 0.25, 0.25), ("DC2", 0.25, 0.5),...]
```

1. Schreiben Sie eine Funktion `entfernung(kunde,wvz)`, die den Abstand zwischen einem Kunden und einem Warenverteilzentrum berechnet. Als Abstandsfunktion soll der euklidische Abstand benutzt werden.

2. Schreiben Sie eine Funktion `naechstesVZ(kunde)`, die für einen Kunden das nächstgelegene Verteilzentrum bestimmt.

3. Berechnen Sie aus der Kundenliste eine neue Liste, die für jeden Kunden eine weitere Komponente bekommt, die das jeweils nächstgelegene Verteilzentrum enthält!

4. Schreiben Sie eine Funktion `kundenverteilung()`, die für die Verteilzentren jeweils die Anzahl der zugeordneten Kunden berechnet. Die Funktion soll eine Liste mit Tupeln der folgenden Form liefern:

   ```
   (verteilzentrumID, anzahlKunden)
   ```

Aufgabe 2 (Pandas und Statistiken)

Der Lieferdienst führt natürlich über die bestellten Waren pro Kunde genau Buch. Kunden bestellen Waren für jeweils eine Woche im voraus, z.B. 3 Portionen Äpfel für Dienstag, Donnerstag und Sonntag, 1 Portion am Samstag.
Die folgende Tabelle zeigt je Kunde die bestellten Artikel pro Tag. Der Wert ist jeweils als *Anzahl Portionen* gedacht.

Kunde	Woche	Kategorie	Produkt	Mo	Di	Mi	Do	Fr	Sa	So
K7	5/2021	Obst	Äpfel	3	0	0	3	0	1	3
K7	5/2021	Fertiggerichte	Nudeln	1	2	1	3	3	3	2
K4	5/2021	Fertiggerichte	Bratkartoffeln	0	3	1	2	2	1	3
K4	5/2021	Gemüse	Paprika	0	0	0	3	1	0	3
K4	5/2021	Getränke	Bier	3	2	3	1	1	1	2

Sie finden im Verzeichnis `data/` die Datei `bestellungen.csv`, die die Daten von einer Reihe von Kunden enthält.

1. Für wieviele verschiedene Kalenderwochen sind in der Datei Bestellungen enthalten?
   Wieviele Kunden enthalten die Daten?

2. Berechnen Sie eine Spalte `Gesamt`, die für jede Zeile die Gesamtanzahl an Portionen für ein Produkt pro Kunde enthält.

3. Die Kategorien `Gemüse` und `Obst` enthalten die kritischen Frischeprodukte, für die auf jeden Fall genügend Lagerkapazität im lokalen Warenverteilzentrum vorhanden sein muss.

   Berechnen Sie die Portionenanzahlen für diese Frischeprodukte pro Kalenderwoche. Dafür bietet sich z.B. die `groupby` Funktion von `DataFrame` an.

   Welche Kalenderwoche hat die meisten/wenigsten Frischeportionen? Wie sehr schwankt dieser Wert? (Standardabweichung?)

4. Berechnen Sie den Anteil der Frischeportionen an der Gesamtzahl bestellter Portionen über alle Wochen.

Aufgabe 3 (Modell-Training)

An dieser Stelle wird es in der finalen Hausarbeit noch eine Aufgabe zur Vorhersage/Modellierung von Daten mit dem *SciKit-Learn* Modul stehen. Krankheitsbedingt fällt diese Aufgabe für den Test der Hausarbeit allerdings weg.