Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science 1

Wintersemester 2023 / 2024

Hausarbeit

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 muss bis spätestens 23:59 Uhr am 29.2.2024 in der zugehörigen Aufgabe im Moodle Kurs hochgeladen werden.
Andere Formen der Abgabe sind nicht vorgehen.

Die Hausarbeit kann in Gruppenarbeit von bis zu drei Personen bearbeitet werden. In diesem Falle genügt *eine* Abgabe.

In jedem Fall sind in der Hausarbeit am Anfang des Notebooks die Namen und Matrikelnummern aller daran beteiligten Personen zu vermerken (gilt auch für Einzelabgaben).

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.

Die Verwendung von ChatGPT oder ähnlichen Hilfsmitteln ist nicht gestattet. Die Prüfungsordnung sieht für Verdachtsfälle die Möglichkeit mündlicher Nachprüfungen vor.

Aufgabe 1 (Python Basics)

Der Mode-Markt ist sowohl online als auch in lokalen Stores ein Markt mit einem hohen Preisdruck und viel Optimierung. Im Bereich Data Science geht es hierbei um Marketing-Analysen, Produktentwicklung, Kundensegmente, dynamische Preisgestaltung und vieles mehr.
In dieser Aufgaben betrachten wir die Produkte und Einkäufe einer Mode-Kette. Dazu gibt es auf dem DataScience Server das Modul **fashion**, über das einige Daten und Funktionen zur Verfügung gestellt werden.
Mit der Funktion **bestellungen()** wird eine Liste zurückgeliefert, in der jedes Element ein Tupel mit den folgenden Werten ist:

```
(bestellNr, datum, kundenNr, alter, [ artikel1, artikel2, ...  ]  )
```

Hier ist ein kleines Beispiel, wie die Daten zu benutzen sind:

```
import fashion

bestellungen = fashion.bestellungen()

bestellung_nr50 = bestellungen[50]
#  (1579197, '2019-01-22', 8, 32, [67994, 44888, 40797,
                                   54725, 16930])
```

Wie in dem Python Code zu sehen ist, hat beispielsweise die Bestellung am Index 50 der Liste die folgenden Werte:

```
(1579197, '2019-01-22', 8, 32, [67994, 44888, 40797, 54725, 16930])
```

Das heisst, die Bestellung mit der Nummer 1579197 wurde am 22.1.2019 aufgegeben und zwar von Kunde Nummer 8, der aktuell 32 Jahre alt ist. In dieser Bestellung wurden die Artikel 67994, 44888, 40797, 54725 und 16930 bestellt.

Für eine derartige Liste sollen Sie die folgenden Aufgaben lösen:

1. Schreiben Sie eine Funktion **kunden(liste)**, die als Parameter die obige Liste bekommt und die Menge der Kundennummern zurückgibt, für die es Bestellungen in der Liste gibt. Dabei soll jede Kundennummer nur einmal in der Ergebnisliste vorkommen.

2. Schreiben Sie eine Funktion **bestellungen_von(liste, kunde)**, die für die gegebene Liste und eine vorgegebene Kundennummer eine Liste aller Bestellungen dieses Kunden zurückgibt.

3. Geben Sie eine Funktion **anzahl_bestellungen(liste)** an, die eine Liste von Bestellungen in der obigen Form bekommt, und eine Liste mit Tupeln der Art als Ergebnis zurückliefert:

   ```
   [ (kundenNr, anzahlBestellungen), ...]
   ```

   D.h. im Ergebnis steht jeweils ein Tupel aus einer Kundenummer und der Anzahl von Bestellungen für diesen Kunden.

4. Eine wichtige Kennzahl für z.B. das Marketing ist die Länge der Kundenbindung, die sogenannte *customer lifetime*.

   Schreiben Sie eine Funktion `lifetime(orders, kundenNr)`, die für eine gegebene Kundennummer die Anzahl der Tage zwischen der ersten und der letzten Bestellung berechnet.

   Das Modul `fashion` enthält eine Funktion `tage_zwischen(datum1, datum2)`, mit der die Differenz zweier Daten aus der obigen Bestell-Liste berechnet werden kann.

5. Der Datensatz enthält bestellungen zwischen dem 20.9.2018 und dem 22.9.2020. Eine häufige Definition von verlorenen Kunden erfolgt über die Anzahl der Tage, die der letzte Einkauf eines Kunden zurückliegt. Schreiben Sie eine Funktion `verlorene_kunden(orders)`, die die Kundennummern aller Kunden zurückgibt, die seit mehr als 60 Tagen nichts mehr gekauft haben.

   Gehen Sie dabei vom 22.9.2020 als dem aktuellen Datum der Ermittlung aus – sonst wären ja alle Kunden *verloren*.

6. Schreiben Sie eine Funktion `kauf_frequenz(orders, kundenNr)`, die die durchschnittliche Anzahl der Tage zwischen zwei Einkäufen des Kunden mit der vorgegebenen `kundenNr` berechnet.

Aufgabe 2 (Pandas und Statistiken)

In dieser Aufgabe sollen ein paar der Daten aus dem Fashion-Markt weiter analysiert werden. Sämtliche Dateien zu dieser Aufgaben finden Sie im Verzeichnis

`Kurse/DataScience1/data/fashion`

auf dem Notebook Server. Als Grundlage dient eine Tabelle mit Bestelldaten, die Sie in der Datei `bestellungen.csv` finden:

ID	Datum	KundenNr	ArtikelNr	ProduktNr	Preis	Kanal
1299843	2018-12-28	1559	57268	716258	0.02540678	2
1299844	2018-12-28	1582	49048	693764	0.030491525	1
1299889	2018-12-28	4622	33207	637673	0.02201695	2
1299831	2018-12-28	14	34866	643985	0.02159322	2
1299831	2018-12-28	14	51780	699755	0.036	2
1299845	2018-12-28	1736	43338	676387	0.033881355	1
1299832	2018-12-28	23	47211	688262	0.05083051	2
1299833	2018-12-28	39	39836	662857	0.036	1

Table 1: Die Datei `bestellungen.csv`

Die Spalten *ID*, *Datum* und *KundenNr* enthalten die Bestell-ID, das Bestelldatum und die Kundennummer des Bestellers. Die Spalte *ArtikelNr* ist der eindeutige Artikel, die *ProduktNr* bezeichnet das Produkt. Ein Produkt kann z.B. in unterschiedlichen Farben oder Größen angeboten werden. Die Varianten haben alle die gleiche *ProduktNr* aber unterschiedliche *Artikelnummern*.
Der Preis ist ein bereits skalierter Preiswert. Die *Kanal*-Spalte enthält die Information über den Weg, über den der Artikel gekauft wurde – dies ist für die Hausarbeit jedoch nicht relevant.

Produkt-Daten

Zu dem Mode-Anbieter gibt es auch eine Reihe von CSV-Dateien, die zusätzliche Informationen zu Produkten und Kunden enthalten. Die Datei `produkte.csv` enthält zu jeder Artikelnummer die zugehörigen Produktinformationen:

ID	Produktname	Hauptkategorie	Kategorie	Unterkategorie
134	Bobby elastic waist belt	Ladieswear	Womens Big accessories	Accessories
135	FIFTY SHADES moulded halternec	Ladieswear	Womens Swimwear, beachwear	Swimwear
136	FIFTY SHADES tie brief	Ladieswear	Womens Swimwear, beachwear	Swimwear
137	Robin 3pk solid	Menswear	Men Underwear	Under-, Nightwear
138	4p Claw	Ladieswear	Womens Small accessories	Accessories
139	4p Claw	Ladieswear	Womens Small accessories	Accessories

Table 2: Die Tabellenstruktur in der Datei `produkte.csv`.

Die Spalten sind eigentlich selbsterklärend: die ID ist die Artikel-ID, die auch in der Bestellungstabelle vorkommt. Die Bezeichnung des Artikels steht in der Spalte *Produktname*, die Eingruppierung innerhalb des Produktkatalogs erfolgt über die Spalten *Hauptkategorie*, *Kategorie* und *Unterkategorie*.

Kunden-Daten

Ein paar wenige Details gibt es auch über die Kunden. Die Datei `kunden.csv` enthält die folgende Tabellenstruktur:

ID	Alter	Aktiv	ClubStatus	NewsletterFrequency
1371946	32.0	0.0	ACTIVE	NONE
1371947	24.0	0.0	ACTIVE	NONE
1371948	55.0	1.0	ACTIVE	Regularly
1371949	34.0	1.0	ACTIVE	Regularly

Table 3: Die Struktur der Tabelle `kunden.csv`.

Die Kundendaten enthalten das aktuelle Alter des Kunden, den Club-Status und die Häufigkeit, die der Kunde den Newsletter zugeschickt bekommt.

Sie sollen sich im Folgenden mit diesen Daten beschäftigen. Stellen Sie sich vor, dies wären Daten ihres Unternehmens, d.h. Sie sind an Fragen interessiert wie z.B.

- Was sind unsere umsatzstärksten Wochentage? Monate?
- Wie hat sich der Umsatz über die letzten Monate entwickelt? Trend?
- Welche Altersgruppe kauft am häufigsten bei uns ein?

Hintergrund dieser Aufgabe ist es, dass Sie sich mit einem unbekannten Datensatz vertraut machen und mit Hilfe von Pandas untersuchen, welche Informationen aus den Daten herausgesucht werden können.

Die Aufgaben:

1. Zunächst sollen ein paar generelle Informationen berechnet werden:
   - Welche Spalten/Datentypen haben die Datensätze?
   - Wieviele Artikel/Kunden/Bestellungen gibt es in dem Datensatz?
   - Über welchen Zeitraum sind die Bestellungen aufgezeichnet worden?
   - Gibt es fehlende Werte? Gibt es Duplikate in den Daten? Welche Spalten enthalten fehlende Werte und wie viele?
2. Wieviele Produkte werden im Schnitt pro Bestellung gekauft? Wieviel Artikel enthält die Bestellung mit den meisten gekauften Produkten?
3. Erstellen Sie ein Diagramm, das die Anzahl der Bestellungen pro Woche darstellt. Ist ein Umsatztrend erkennbar? Vergleichen Sie die Anzahl der Bestellungen von Jan/Feb/März/April 2020 jeweils mit dem entsprechenden Vorjahresmonat.

   Hinweis: Hier müssen Sie mit Datumsangaben umgehen. Dazu hatten wir in den Übungen z.B. die Funktion `pd.to_datetime` kennengelernt.

4. Teilen Sie die Kunden in Altersgruppen ein. Nutzen Sie dafür die Gruppierungen 10-20, 20-30, 30-40, usw. (alternativ können Sie auch eine eigene Altersgruppierung wählen).

   Wie ist die Altersstruktur der Kunden? Erstellen Sie einen Histogram-Plot dazu!

   Hinweis: Schauen Sie sich dazu z.B. die Pandas Funktion `cut` an (siehe Pandas Dokumentation).

5. Unterscheidet sich die Altersstruktur bei Club- bzw. Nicht-Club Mitgliedern?

6. Gruppieren Sie die Bestellungen nach Altersgruppen - Welche Altersgruppe erzeugt die meisten Bestellungen? Welche Altersgruppe bestellt die meisten Artikel?

7. Aus welcher Hauptkategorie werden die meisten Produkte angeboten?

   Aus welcher Hauptkategorie werden die meisten Produkte verkauft?

Hinweis zur Bearbeitung

Die Vorlesung hat einige der Grundlagen zu Pandas vermittelt. Natürlich ist Pandas deutlich umfangreicher, als man es innerhalb einer 1-semestrigen Vorlesung vermitteln kann. Für die Lösung einiger dieser Aufgaben ist es daher erforderlich, sich weiter mit Pandas zu beschäftigen. Dazu gehört u.a. die Verbindung mehrerer Tabellen (JOIN), was Sie beispielsweise aus dem Bereich der Datenbanken in Wirtschaftsinformatik 2 bereits kennen sollten.
Die Dokumentation für den JOIN findet sich z.B. unter

`https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.join.html`

Es sei hier noch angemerkt, dass es hilfreich ist, wenn der *index* des DataFrames, den man an einen bestehenden DataFrame heften möchte für den JOIN relevant ist. So sollte z.B. der DataFrame für die Kunden als Index am besten die Kundennummer enthalten, bevor dieser an die Bestellungen ge-joint wird.

Es geht bei der Bearbeitung dieser Aufgaben nicht nur um die reine Programmierung in Python. Ziel ist es, die Daten entlang der Teilaufgaben zu analysieren und die Ergebnisse in einem gewissen Rahmen zu interpretieren.
Dazu gehört zu jeder Teilaufgabe, dass Sie kurz skizzieren, wie Sie vorgehen wollen, welche Teil-DataFrames sie ggf. berechnen wollen und was Sie am Ergebnis ggf. kritisch betrachten (z.B. Datenqualität, etc.). Auch dafür haben Sie in Data Science und Kursen wie Wirtschaftsstatistik Methoden und Werkzeuge kennengelernt.