Beispielfragen für die Klausur Business Analytics

Die Aufgaben basieren zur besseren Nachvollziehbarkeit hier auch auf Beispielen aus unseren Vorlesungsunterlagen.

Aufgabenteil 1 – Grundlagen und Begriffe

a. In welchem Zusammenhang stehen *Variablen*, *Prädiktoren* und die *Spalten eines Datensatzes*?

b. Unterscheiden Sie die Nutzung von Daten zur *Deskription* und zur *Prädiktion*.

c. Wie können wir allgemein die *Qualität von Vorhersagen* auf Basis der Modelle aus der Veranstaltung beurteilen? Nennen Sie auch konkrete Beispiele.

d. Was verstehen wir unter der *Dimensionalität* von Daten?

Aufgabenteil 2 – Szenarien, Ansätze und Methoden

Nennen Sie für die gebenenen Szenarien geeignete Verfahren aus unserer Veranstaltung und begründen Sie Ihre Auswahl (Arten von Prädiktoren und Zielvariablen, ggf. Eigenschaften der Daten und Verfahren).

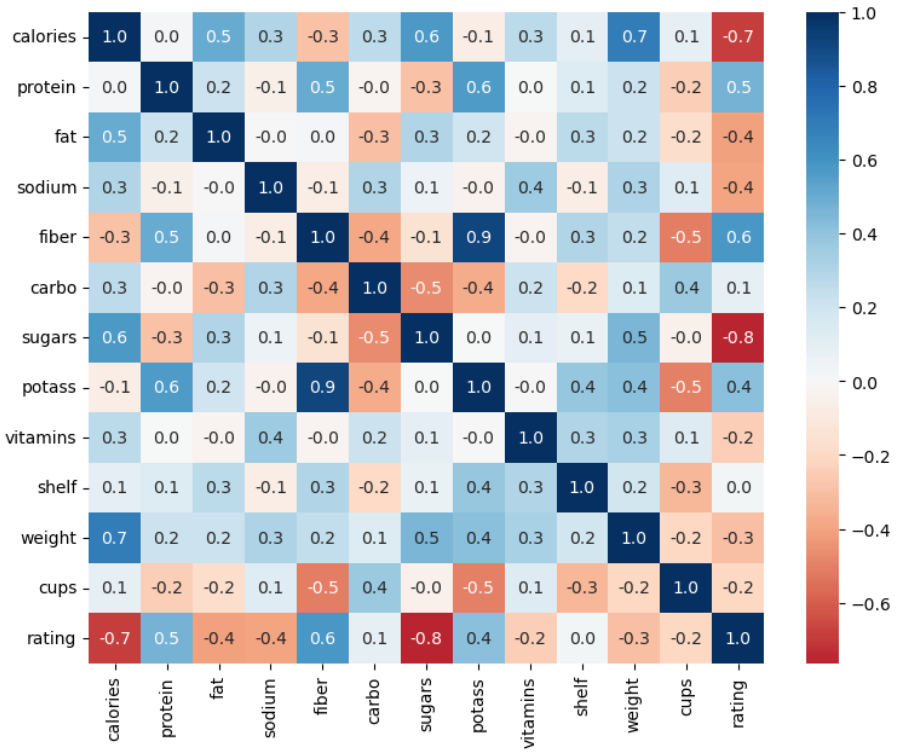
a. Sie bekommen einen großen Datensatz mit sehr vielen potentiellen Prädiktoren. Die Daten stammen aus einem Anwendungsbereich, der Ihnen unbekannt ist.

b. In einem Fertigungsprozess stehen drei unterschiedliche Filtersysteme zur Verfügung. Für eine Reihe von Produkten soll der Einfluss der Filter auf den Schadstoffausstoß untersucht werden.

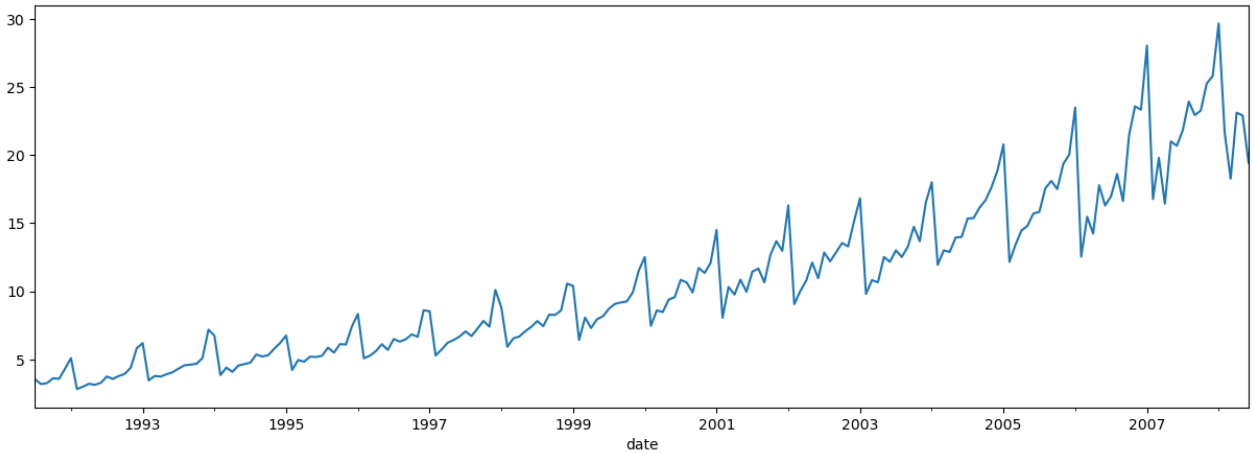
c. Aus den Verlaufsdaten der Wasserstände aus einem Küstenabschnitt soll die Gefahr von Hochwasser prognostiziert werden.

Aufgabenteil 3 – Vorbereitung und initiale Datenbetrachtung

a. Betrachten sie die folgende Korrelationsmatrix. Welche Aussagen können Sie über die Variablen als potentielle Prädiktoren treffen? Wie können Sie die Matrix für ein Vorhersagemodell nutzen?



b. Betrachten Sie die folgende Zeitreihe. Was verstehen wir unter Komponenten von Zeitreihen? Welche können Sie hier erkennen und welche Eigenschaften haben sie hier?



c. Welche Arten von Saisonalitäten gibt es in Zeitreihen und wie unterscheiden sie sich?

d. Betrachten sie folgenden Datensatz:

	age_08_04	km	fuel_type	hp	met_color	automatic	cc	doors	quarterly_tax	weight
0	23	46986	Diesel	90	Y	0	2000	3	210	1165
1	23	72937	Diesel	90	Y	0	2000	3	210	1165
2	24	41711	Diesel	90	Y	0	2000	3	210	1165
3	26	48000	Diesel	90	N	0	2000	3	210	1165
4	30	38500	Diesel	90	N	0	2000	3	210	1170
...	...	...	...	...	...	...	...	...	...	...
995	68	42750	Petrol	110	Y	0	1600	3	69	1050
996	67	42102	Petrol	110	Y	0	1600	5	85	1075
997	63	41586	Petrol	110	Y	0	1600	5	19	1114
998	64	41200	Petrol	110	N	0	1600	5	85	1070
999	57	40214	Petrol	86	N	0	1300	3	69	1025

Wenn Sie alle enthaltenen Variablen als Prädiktoren für eine multiple lineare Regression nutzen möchten, warum und wie müssten Sie die Daten vorbereiten? Erläutern Sie die Funktionsweise der passenden Funktion aus Pandas.

Aufgabenteil 4 – Verfahren

a. Die erschöpfende Suche einer *Variablenauswahl für die lineare Regression* liefert folgendes Ergebnis:

	n	r2adj	AIC
0	1	0.767901	10689.712094
1	2	0.801160	10597.910645
2	3	0.829659	10506.084235
3	4	0.846357	10445.174820
4	5	0.849044	10435.578836
5	6	0.853172	10419.932278
6	7	0.853860	10418.104025
7	8	0.854297	10417.290103
8	9	0.854172	10418.789079
9	10	0.854036	10420.330800
10	11	0.853796	10422.298278

Wie viele Variablen würde das Modell Ihrer Wahl haben? Erläutern Sie die Kriterien bzw. was sagen die Kennzahlen hier aus (Formeln müssen Sie nicht angeben)?

b. Wir nutzen eine *logistische Regression* zur Vorhersage, ob eine Kundin ein Darlehen akzeptiert. Prädiktoren sind Alter, Berufserfahrung und Einkommen. Das resultierende Modell liefert

```
intercept  -11.520720605525696
               Age  Experience    Income
coeff   0.212317   -0.213463  0.037524
```

Geben Sie den Logit an. Wie wirken sich jeweils Änderungen der Werte der Prädiktoren aus, wenn alle anderen unverändert bleiben?

c. Eine *zweifaktorielle Varianzanalyse* der Absatzzahlen von Schokolade mit verschiedenen Platzierungen und Verpackungen liefert folgende Ergebnisse:

	Quelle	Quadrate (erklärt)	Fehler (unerklärt)	Gesamtabweichung	Partielles Eta-Quadrat
0	Platzierung	1944.200000	238.0	2182.200000	0.890936
1	Verpackung	240.833333	238.0	478.833333	0.502959
2	Interaktion	48.466667	238.0	286.466667	0.169188

Erläutern Sie das Vorgehen der zweifaktoriellen Varianzanalyse im Allgemeinen und interpretieren Sie die hier gegebenen Kennzahlen.