

## Data Science Gruppenaufgabe

Name	Matrikelnummer
Darius Bonk	22213311
Vipul Durgade	22213303
Matial Domche	22213315
Kiran Krishnakumar	22213304

Hausaufgabe im Studienfach Data Science

bei

Dr.-Ing. J.-H. Wieken, Fachhochschule Westküste

vorgelegt am 04.12.2022

## Inhaltsverzeichnis

Abbildungsverzeichnis .....	II
Tabellenverzeichnis.....	III
1. Aufgabe 1 .....	1
2. Aufgabe 2 .....	4
3. Aufgabe 3 .....	6
4. Aufgabe 4 .....	8
Literaturverzeichnis .....	i

## Abbildungsverzeichnis

**Es konnten keine Einträge für ein Abbildungsverzeichnis gefunden werden.**

## Tabellenverzeichnis

**Es konnten keine Einträge für ein Abbildungsverzeichnis gefunden werden.**

## 1. Aufgabe 1

- a) Geben Sie die vier grundlegenden Data Science-Skills an und erläutern Sie diese kurz.

1- Maschine Learning

Hier geht es darum, Modelle zu entwickeln und einzusetzen, um produktive KI-Lösungen zu implementieren und Modelle und Vorhersagen in für das Unternehmen nützlichen Begriffen zu erklären.

2- Informatik

Zur Informatik gehören Datenbank, Datenkonvertierung, technische Infrastruktur und effizienter und wartbarer Code.

3- Anwendungsgebiet

Der Anwendungsbereich umfasst das Verständnis des eigentlichen Ziels sowie das Verständnis des Bewertungsprozesses. Die Datenerhebung oder die Identifizierung möglicher Probleme im Prozess. Wichtig ist vor allem die Zielgruppengerechte Aufbereitung der Daten und Ergebnisse.

4- Datenhandlung

Die Manipulation von Daten umfasst die Erstellung von Statistiken, die Visualisierung von Daten und den Schutz von Daten, unabhängig von der verwendeten Technik.

b) Geben Sie die drei Bestandteile von SQL an und beschreiben Sie deren Anwendungsbereich.

- SQL-DDL (Data Definition Language)

Dient der Erstellung, Änderung und Löschung von Datenbankstrukturen.

- SQL-DML (Data Manipulation Language)

Dient der Abfrage, dem Einfügen, Ändern und Löschen Daten in gegebenen Strukturen

- SQL-DCL (Data Control Language)

Dient der Pflege der Datenbankinfrastruktur, beispielsweise der Zugriffsberechtigungen oder der Speicherverwaltung

- b) Erläutern Sie die Ziele, Vorteile und Nachteile der Normalisierung

Unter Normalisierung versteht man den Prozess der Reduzierung der Datenredundanz in einer Tabelle und der Verbesserung der Datenintegrität.

Normalisierung ist also eine Methode zur Organisation von Daten in einer Datenbank. Bei der Normalisierung werden die Spalten und Tabellen in der Datenbank organisiert, um sicherzustellen, dass ihre Abhängigkeiten mithilfe von Datenbankeinschränkungen korrekt implementiert werden. Normalisierung ist der Prozess der Organisation von Daten in einer geeigneten Weise. Sie wird verwendet, um die Duplizierung verschiedener Beziehungen in der Datenbank zu minimieren. Sie wird auch zur Fehlerbehebung bei Ausnahmen wie Einfügungen, Löschungen und Aktualisierungen in der Tabelle verwendet. Es hilft bei der Aufteilung einer großen Tabelle in mehrere kleine normalisierte Tabellen. Relationale Verknüpfungen und Links werden verwendet, um Redundanz zu reduzieren. Normalisierung, auch bekannt als Datenbanknormalisierung oder Datennormalisierung, ist ein wichtiger Teil des

relationalen Datenbankdesigns, da sie dazu beiträgt, die Geschwindigkeit, Genauigkeit und Effizienz der Datenbank zu verbessern.

Vorteile der Normalisierung	Nachteile der Normalisierung
Niedriger Speicherplatzbedarf	Viele Tabellen
Anomalien werden vermieden	Viele Join-Operationen nötig
Änderungen einfach	Einbußen bei der Performanz
Sicherung der Datenqualität	

## 2. Aufgabe 2

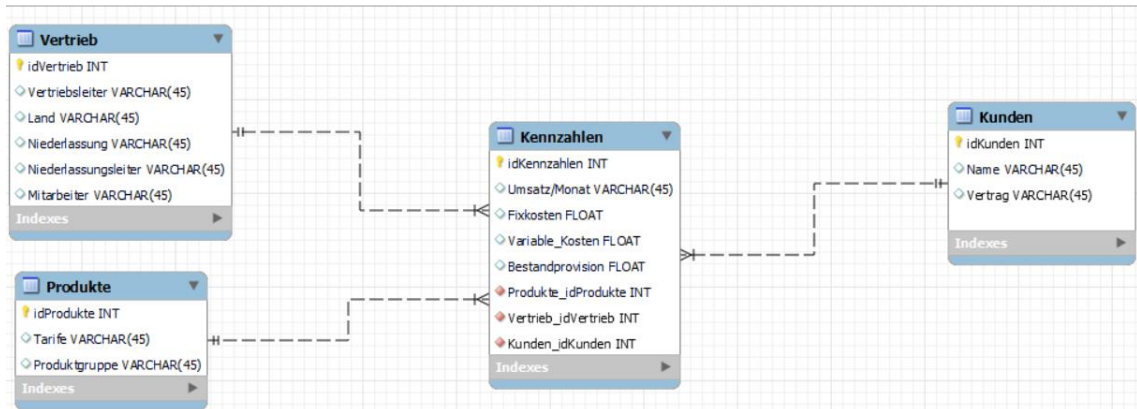
Sie stehen vor folgender Anforderung: Ein Versicherungsunternehmen möchte den Erfolg seiner Außendienststruktur ermitteln.

Dafür erhalten Sie folgende Beschreibung:

„Der Vertrieb hat einen Vertriebsleiter und ist darunter nach den Ländern Österreich, Deutschland und Schweiz aufgeteilt. Diese haben jeweils einen eigenen Landesvertriebsleiter. In jedem Land erfolgt dann eine Aufteilung nach Regionen und innerhalb der Regionen nach Niederlassungen mit den Niederlassungsleitern, die ihrerseits die Agenturen mit deren einzelnen Agenturmitarbeitern koordinieren. Unsere Produkte sind die Tarife, die zu den Produktgruppen Leben-, Kranken und Sachversicherungen gehören. Die Kunden haben wir wie den Vertrieb regional gegliedert. Ein Kunde schließt einen Vertrag ab, der mehrere Tarife beinhalten kann. Der Tarif bestimmt den Umsatz pro Monat. Der Umsatz wird außerdem für die Bestandsprovision dem Mitarbeiter zugeordnet, der den Vertrag mit dem Kunden abgeschlossen hat und gilt monatlich jeden Monat solange der Vertrag läuft. Außerdem werden unsere Kosten für die Verwaltung des Tarifs und eventuelle Zahlungen bei Schäden monatlich dem Tarif des Kunden, dem Produkttarif zugeordnet. Außerdem brauchen wir die Kosten, getrennt nach Fixkosten und variablen Kosten, die in den einzelnen Ebenen der Vertriebsstruktur monatlich entstehen, unabhängig von den Verträgen und Tarifen, sowie die Differenz zwischen Umsatz und diesen Kosten.“



a) Entwerfen Sie ein Starschema als ER-Diagramm für diese Situation.



b) Erstellen Sie SQL-Befehle, um die Tabellen für die Kennzahlen und zumindest eine Dimensionstabelle des Starschemas zu erzeugen.

### 3. Aufgabe 3

Erstellen Sie einen SQL-Befehl mit dem Schema Nordmarkt in MySQL basierend auf den Tabellen des Starschemas (ohne die Tabelle nordmarkt selbst), der die Summe der Umsätze und den Durchschnitt der Rabatte pro Monat, pro deutschem Bundesland und pro Kategorie ermittelt. Das Ergebnis soll nach Monaten aufsteigend und innerhalb der Monate nach Umsatzsumme absteigend sortiert sein.

Es soll sichergestellt sein auch wenn Daten hinzukommen, dass nur Daten aus dem Jahr 2022 angezeigt werden. Alle Kennzahlen sollen mit zwei Nachkommastellen und dem Währungssymbol € ausgegeben werden.

```

1 • SELECT Monat, CONCAT('€ ', CAST(CAST(sum(Umsatz) AS DECIMAL(18,2)) AS CHAR(55))) AS "Umsatzsumme"
2   , CONCAT('€ ', CAST(CAST(avg(Umsatz) AS DECIMAL(18,2)) AS CHAR(55))) AS "Umsatzavg",
3   Bundesland, Kategorie
4 FROM
5   (nordmarkt.kennzahlen INNER JOIN nordmarkt.bestelldatum ON nordmarkt.kennzahlen.bestelldatum_ID = nordmarkt.bestelldatum.bestelldatum_ID)
6   INNER JOIN nordmarkt.kunde ON nordmarkt.kennzahlen.kunde_ID = nordmarkt.kunde.kunde_ID
7   INNER JOIN nordmarkt.produkt ON nordmarkt.kennzahlen.produkt_ID = nordmarkt.produkt.produkt_ID
8   WHERE (nordmarkt.bestelldatum.Jahr = 2022)
9   GROUP BY nordmarkt.bestelldatum.Monat, nordmarkt.kunde.Bundesland, nordmarkt.produkt.kategorie
10  ORDER BY nordmarkt.bestelldatum.Monat, length(Umsatzsumme) DESC, Umsatzsumme DESC;

```

100% 81:10

Result Grid Filter Rows: Search Export:

	Monat	Umsatzsumme	Umsatzavg	Bundesland	Kategorie
1	1	€ 124363.75	€ 358.40	Schleswig-Holstein	Obst und Gemüse
1	1	€ 49350.66	€ 135.95	Schleswig-Holstein	Fleisch
1	1	€ 46998.35	€ 114.63	Schleswig-Holstein	Getränke
1	1	€ 21014.34	€ 179.61	Hamburg	Getränke
1	1	€ 15960.80	€ 121.84	Hamburg	Fleisch
1	1	€ 9036.14	€ 86.06	Hamburg	Obst und Gemüse
1	1	€ 7904.38	€ 42.22	Schleswig-Holstein	Fisch
1	1	€ 6902.09	€ 57.04	Schleswig-Holstein	Süßwaren
1	1	€ 4475.67	€ 248.65	Mecklenburg-Vorpommern	Fleisch
1	1	€ 3731.09	€ 95.67	Niedersachsen	Getränke
1	1	€ 2854.60	€ 109.79	Niedersachsen	Fleisch

Erweitern Sie auf Basis der Tabelle nordmarkt die Kennzahlentabelle um den Einkaufspreis\_netto und füllen Sie die Spalte mit den zugehörigen Daten. Geben Sie die benötigten SQL-Befehle an.

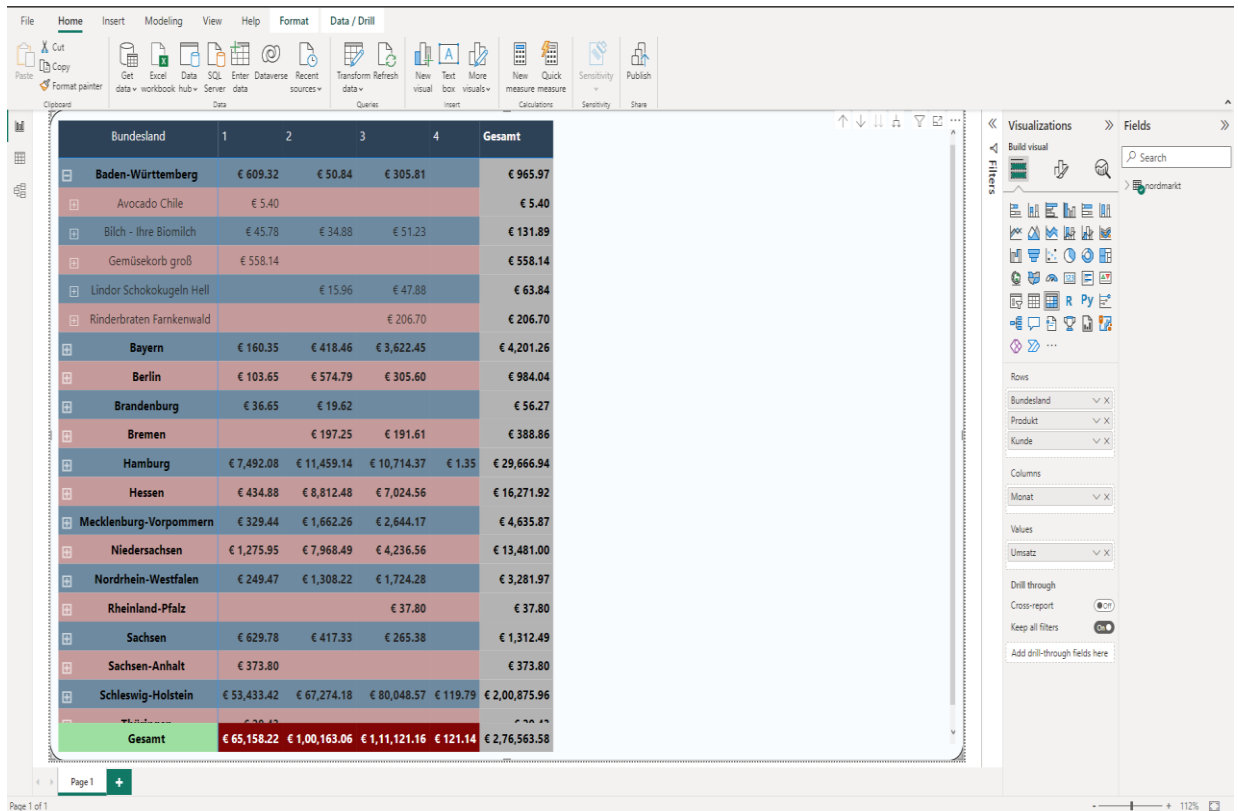
```
1 • ALTER TABLE kennzahlen
2 ADD Einkaufspreis_netto double(22,2);
3 • Update kennzahlen
4 Set kennzahlen.Einkaufspreis_netto = (select nordmarkt.Einkaufspreis_netto
5 from nordmarkt where kennzahlen.kennzahl_ID = nordmarkt.kennzahl_ID)
```

Der Einkaufspreis\_netto von Bilch-Ihre Biomilch wird zum 1.2 von 1,09 € auf 1,29 € erhöht.  
Beschreiben Sie die notwendigen Änderungen in der Datenbank (Spalten, Fremdschlüssel, Zeilen) damit sowohl Analysen für die Umsätze mit diesem Produkt

- a. mit dem ursprünglichen Preis für den gesamten Zeitraum
- b. Mit dem neuen Preis für den gesamten Zeitraum
- c. Mit dem jeweils zu dem Zeitpunkt gültigen Preis möglich sind.

## 4. Aufgabe 4

- a) Erstellen Sie auf der Basis von Nordmarkt eine tabellarische Auswertung der Umsätze in Deutschland nach Bundesland und Monat. Richten Sie die in der Datenbank vorhandenen drei Dimensionen Kunde, Produkt und Bestelldatum als Dimension ein.

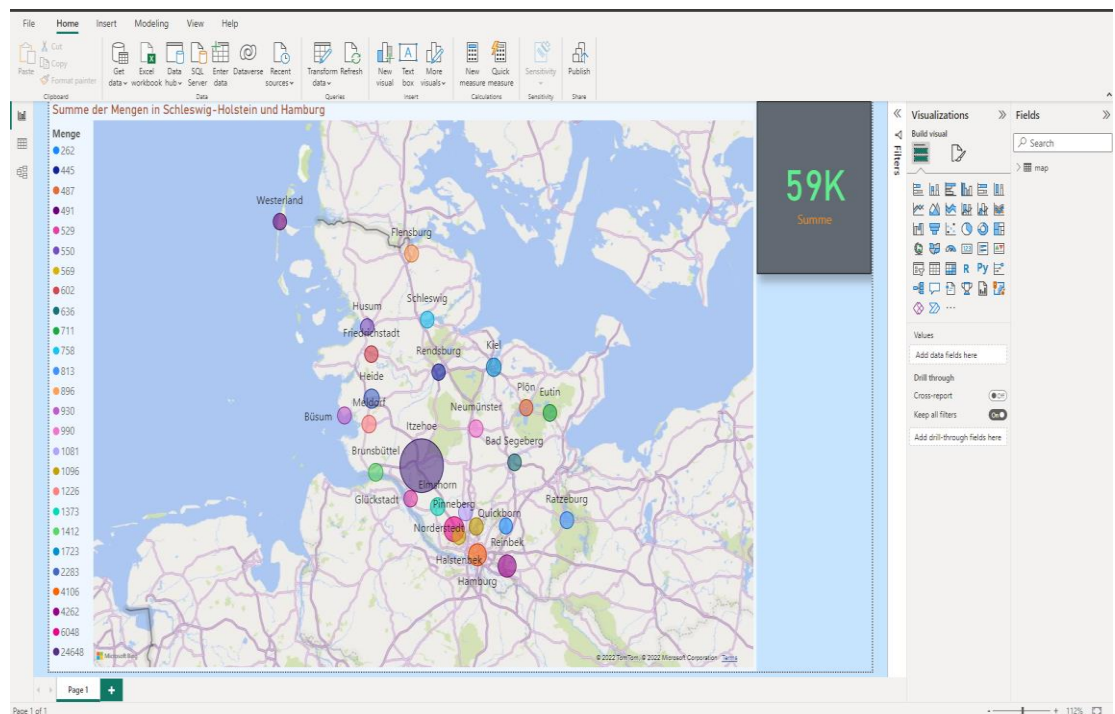


The screenshot shows the Microsoft Power BI Desktop interface. The main view displays a data table with the following structure:

Bundesland	1	2	3	4	Gesamt
Baden-Württemberg	€ 609.32	€ 50.84	€ 305.81		€ 965.97
Avocado Chile	€ 5.40				€ 5.40
Bilch - Ihre Biomilch	€ 45.78	€ 34.88	€ 51.23		€ 131.89
Gemüsekorb groß	€ 558.14				€ 558.14
Lindor Schokokugeln Hell		€ 15.96	€ 47.88		€ 63.84
Rinderbraten Farnkenwald			€ 206.70		€ 206.70
Bayern	€ 160.35	€ 418.46	€ 3,622.45		€ 4,201.26
Berlin	€ 103.65	€ 574.79	€ 305.60		€ 984.04
Brandenburg	€ 36.65	€ 19.62			€ 56.27
Bremen		€ 197.25	€ 191.61		€ 388.86
Hamburg	€ 7,492.08	€ 11,459.14	€ 10,714.37	€ 1.35	€ 29,666.94
Hessen	€ 434.88	€ 8,812.48	€ 7,024.56		€ 16,271.92
Mecklenburg-Vorpommern	€ 329.44	€ 1,662.26	€ 2,644.17		€ 4,635.87
Niedersachsen	€ 1,275.95	€ 7,968.49	€ 4,236.56		€ 13,481.00
Nordrhein-Westfalen	€ 249.47	€ 1,308.22	€ 1,724.28		€ 3,281.97
Rheinland-Pfalz			€ 37.80		€ 37.80
Sachsen	€ 629.78	€ 417.33	€ 265.38		€ 1,312.49
Sachsen-Anhalt	€ 373.80				€ 373.80
Schleswig-Holstein	€ 53,433.42	€ 67,274.18	€ 80,048.57	€ 119.79	€ 200,875.96
<b>Gesamt</b>	<b>€ 65,158.22</b>	<b>€ 1,00,163.06</b>	<b>€ 1,11,121.16</b>	<b>€ 121.14</b>	<b>€ 2,76,563.58</b>

The interface includes a ribbon with tabs like File, Home, Insert, Modeling, View, Help, Format, and Data / Drill. The right-hand pane shows the 'Visualizations' and 'Fields' sections, with 'Bundestland', 'Produkt', and 'Kunde' listed as dimensions and 'Monat' as a column. The 'Values' section shows 'Umsatz' (Sales) as the measure.

Erstellen Sie eine zweite Auswertung, die die Summe der Mengen in Form einer Landkarte wiedergibt. Berücksichtigt werden soll nur Schleswig-Holstein und Hamburg. Farblich sollen die Mengen erkennbar sein und die Werte als Beschriftung dienen.



Sie können wahlweise Power BI oder Tableau verwenden. Speichern Sie das Ergebnis als Power BI oder Tableau-Arbeitsmappe und als Screenshot des Gesamtbildschirms.

## Literaturverzeichnis

**Im aktuellen Dokument sind keine Quellen vorhanden.**