# Block 1: Multiple-Choice Questions (Extended)

## Foundations & Empirical Fitting

Data Science and Strategy for Business

# Question 1

**Welche drei Bereiche bilden die Schnittmenge “Data Science”?**

A. Marketing, Vertrieb, Finanzen
B. Statistik, Informatik, Domänenwissen
C. Machine Learning, Deep Learning, AI
D. Datenbanken, Cloud, APIs

# Question 1

**Welche drei Bereiche bilden die Schnittmenge “Data Science”?**

A. Marketing, Vertrieb, Finanzen
B. Statistik, Informatik, Domänenwissen
C. Machine Learning, Deep Learning, AI
D. Datenbanken, Cloud, APIs

**Answer: B**
Data Science kombiniert statistische Methoden, informatische Werkzeuge und fachliches Domänenwissen.

# Question 2

**Welche Phase des CRISP-DM nimmt typischerweise die meiste Zeit in Anspruch?**

A. Business Understanding
B. Data Preparation
C. Modeling
D. Deployment

# Question 2

**Welche Phase des CRISP-DM nimmt typischerweise die meiste Zeit in Anspruch?**

A. Business Understanding
B. Data Preparation
C. Modeling
D. Deployment

**Answer: B**
Data Preparation (inkl. Data Understanding) beansprucht oft 60-80% der gesamten Projektzeit.

# Question 3

**Ist CRISP-DM ein linearer oder zyklischer Prozess?**

A. Streng linear: jede Phase wird genau einmal durchlaufen
B. Zyklisch: iteratives Vor- und Zurückspringen ist normal
C. Hybrid: nur die ersten 3 Phasen sind iterativ
D. Linear mit optionalen Rücksprüngen

# Question 3

**Ist CRISP-DM ein linearer oder zyklischer Prozess?**

A. Streng linear: jede Phase wird genau einmal durchlaufen
B. Zyklisch: iteratives Vor- und Zurückspringen ist normal
C. Hybrid: nur die ersten 3 Phasen sind iterativ
D. Linear mit optionalen Rücksprüngen

**Answer: B**
CRISP-DM ist zyklisch und iterativ. Insbesondere zwischen Modeling und Data Preparation wird häufig hin- und hergesprungen.

# Question 4

**Warum ist Domänenwissen in Data-Science-Projekten entscheidend?**

A. Es ermöglicht die Interpretation der Ergebnisse und verhindert unsinnige Modelle
B. Es ersetzt statistische Methoden
C. Es ist nur für die Datensammlung wichtig
D. Domänenwissen ist weniger wichtig als technische Skills

# Question 4

**Warum ist Domänenwissen in Data-Science-Projekten entscheidend?**

A. Es ermöglicht die Interpretation der Ergebnisse und verhindert unsinnige Modelle
B. Es ersetzt statistische Methoden
C. Es ist nur für die Datensammlung wichtig
D. Domänenwissen ist weniger wichtig als technische Skills

**Answer: A**
Domänenwissen hilft bei der Interpretation von Mustern, bei der Feature-Auswahl und verhindert statistisch korrekte, aber sachlich unsinnige Modelle.

# Question 5

**Was ist ein häufiges Missverständnis über AI/Machine Learning?**

A. AI kann jedes Problem lösen, ohne Daten oder Domänenwissen
B. AI-Modelle benötigen qualitativ hochwertige Daten
C. AI-Projekte folgen strukturierten Prozessen wie CRISP-DM
D. AI-Ergebnisse müssen validiert werden

# Question 5

**Was ist ein häufiges Missverständnis über AI/Machine Learning?**

A. AI kann jedes Problem lösen, ohne Daten oder Domänenwissen
B. AI-Modelle benötigen qualitativ hochwertige Daten
C. AI-Projekte folgen strukturierten Prozessen wie CRISP-DM
D. AI-Ergebnisse müssen validiert werden

**Answer: A**
"AI is not magic" – AI benötigt gute Daten, klare Ziele und Domänenwissen. Ohne diese Grundlagen liefert auch das beste Modell keine nützlichen Ergebnisse.

# Question 6

**Was bestimmt primär den Wert von Daten für ein Unternehmen?**

A. Die Menge der gesammelten Daten
B. Die Aktualität und Relevanz für strategische Entscheidungen
C. Die Anzahl der Datenbanktabellen
D. Die verwendete Speichertechnologie

# Question 6

**Was bestimmt primär den Wert von Daten für ein Unternehmen?**

A. Die Menge der gesammelten Daten
B. Die Aktualität und Relevanz für strategische Entscheidungen
C. Die Anzahl der Datenbanktabellen
D. Die verwendete Speichertechnologie

**Answer: B**
Daten sind wertvoll, wenn sie aktuell, relevant und nutzbar für Entscheidungen sind. Große Datenmengen ohne Relevanz bringen keinen strategischen Vorteil.

# Question 7

**Wie verteilt sich die Zeit in typischen Data-Science-Projekten?**

A. 80% Modellierung, 20% Datenaufbereitung
B. 60-80% Datenaufbereitung, 20-40% Modellierung und Deployment
C. 50% Business Understanding, 50% Modeling
D. 90% Deployment, 10% Data Preparation

# Question 7

**Wie verteilt sich die Zeit in typischen Data-Science-Projekten?**

A. 80% Modellierung, 20% Datenaufbereitung
B. 60-80% Datenaufbereitung, 20-40% Modellierung und Deployment
C. 50% Business Understanding, 50% Modeling
D. 90% Deployment, 10% Data Preparation

**Answer: B**
Die meiste Zeit geht für Data Understanding und Data Preparation drauf. Modeling ist oft der kleinste Teil des Projekts.

# Question 8

**Warum sollten Sie relative statt absolute Pfade in R verwenden?**

A. Relative Pfade sind schneller
B. Relative Pfade ermöglichen Reproduzierbarkeit und Zusammenarbeit
C. Absolute Pfade werden von R nicht unterstützt
D. Es gibt keinen Unterschied

# Question 8

**Warum sollten Sie relative statt absolute Pfade in R verwenden?**

A. Relative Pfade sind schneller
B. Relative Pfade ermöglichen Reproduzierbarkeit und Zusammenarbeit
C. Absolute Pfade werden von R nicht unterstützt
D. Es gibt keinen Unterschied

**Answer: B**
Relative Pfade funktionieren auf jedem Computer und ermöglichen so Zusammenarbeit und Reproduzierbarkeit. Absolute Pfade wie "C:/Users/Anna/..." funktionieren nur auf einem Rechner.

# Question 9

**Welche dplyr-Funktion erstellt neue Spalten in einem Datensatz?**

A. filter()
B. select()
C. mutate()
D. arrange()

# Question 9

**Welche dplyr-Funktion erstellt neue Spalten in einem Datensatz?**

A. filter()
B. select()
C. mutate()
D. arrange()

**Answer: C**
`mutate()` erstellt neue Spalten oder modifiziert bestehende. `filter()` wählt Zeilen, `select()` wählt Spalten, `arrange()` sortiert.

# Question 10

**Wozu dient der Pipe-Operator %¿% in tidyverse?**

A. Zur Verkettung von Operationen und besserer Lesbarkeit
B. Zum Kommentieren von Code
C. Zum Laden von Paketen
D. Zur Fehlerbehandlung

# Question 10

**Wozu dient der Pipe-Operator %¿% in tidyverse?**

A. Zur Verkettung von Operationen und besserer Lesbarkeit

B. Zum Kommentieren von Code

C. Zum Laden von Paketen

D. Zur Fehlerbehandlung

**Answer: A**

Der Pipe-Operator leitet das Ergebnis einer Operation als Eingabe an die nächste weiter. Dadurch wird Code lesbarer: `daten %>% filter() %>% select()`.

# Question 11

**Was ist das Hauptziel der explorativen Datenanalyse (EDA)?**

A. Das finale Modell zu trainieren
B. Muster, Ausreißer und Datenqualität zu verstehen
C. Hypothesentests durchzuführen
D. Den Datensatz zu verkleinern

# Question 11

**Was ist das Hauptziel der explorativen Datenanalyse (EDA)?**

A. Das finale Modell zu trainieren
B. Muster, Ausreißer und Datenqualität zu verstehen
C. Hypothesentests durchzuführen
D. Den Datensatz zu verkleinern

**Answer: B**
EDA dient dem Verständnis der Datenstruktur, der Identifikation von Mustern, Ausreißern und Qualitätsproblemen – vor der eigentlichen Modellierung.

# Question 12

**In der Regression Umsatz $= 500 + 25 \times$ Werbung bedeutet der Koeffizient 25:**

A. Der Umsatz beträgt 25 CHF
B. Pro zusätzlichem CHF Werbung steigt der Umsatz um 25 CHF
C. 25% des Umsatzes kommen von Werbung
D. Die Korrelation beträgt 0.25

# Question 12

**In der Regression Umsatz = 500 + 25×Werbung bedeutet der Koeffizient 25:**

A. Der Umsatz beträgt 25 CHF
B. Pro zusätzlichem CHF Werbung steigt der Umsatz um 25 CHF
C. 25% des Umsatzes kommen von Werbung
D. Die Korrelation beträgt 0.25

**Answer: B**
Der Koeffizient gibt die Änderung der abhängigen Variable (Umsatz) bei einer Einheit Erhöhung der unabhängigen Variable (Werbung) an.

# Question 13

**Ein $R^2$ von 0.64 bedeutet:**

A. 64% der Datenpunkte liegen auf der Regressionslinie
B. 64% der Varianz in Y wird durch das Modell erklärt
C. Die Korrelation beträgt 0.64
D. Das Modell hat 64 Variablen

# Question 13

**Ein $R^2$ von 0.64 bedeutet:**

A. 64% der Datenpunkte liegen auf der Regressionslinie
B. 64% der Varianz in Y wird durch das Modell erklärt
C. Die Korrelation beträgt 0.64
D. Das Modell hat 64 Variablen

**Answer: B**
$R^2$ misst den Anteil der Gesamtvarianz in Y, der durch die Prädiktoren erklärt wird. (Hinweis: Die Korrelation wäre $r = \sqrt{0.64} = 0.8$)

# Question 14

**Was zeigt $R^2$ NICHT?**

A. Den Anteil erklärter Varianz
B. Ob das Modell kausal ist
C. Die Güte der Anpassung
D. Die Stärke des linearen Zusammenhangs

# Question 14

**Was zeigt $R^2$ NICHT?**

A. Den Anteil erklärter Varianz
B. Ob das Modell kausal ist
C. Die Güte der Anpassung
D. Die Stärke des linearen Zusammenhangs

**Answer: B**
$R^2$ misst nur die Anpassungsgüte, nicht Kausalität. Ein hohes $R^2$ kann auch bei reinen Korrelationen oder Scheinzusammenhängen auftreten.

# Question 15

**Was minimiert die OLS-Methode (Ordinary Least Squares)?**

A. Die Summe der Residuen
B. Die Summe der quadrierten Residuen
C. Die Anzahl der Ausreißer
D. Die Anzahl der Prädiktoren

# Question 15

**Was minimiert die OLS-Methode (Ordinary Least Squares)?**

A. Die Summe der Residuen
B. Die Summe der quadrierten Residuen
C. Die Anzahl der Ausreißer
D. Die Anzahl der Prädiktoren

**Answer: B**
OLS minimiert die Summe der quadrierten Abweichungen (Residuen) zwischen beobachteten und vorhergesagten Werten: $\min \sum (y_i - \hat{y}_i)^2$.

# Question 16

**Was bedeutet “ceteris paribus” bei der Interpretation von Regressionskoeffizienten?**

A. Alle Variablen sind gleich wichtig
B. Der Effekt gilt nur für bestimmte Fälle
C. Alle anderen Variablen werden konstant gehalten
D. Die Koeffizienten sind standardisiert

# Question 16

**Was bedeutet “ceteris paribus” bei der Interpretation von Regressionskoeffizienten?**

A. Alle Variablen sind gleich wichtig
B. Der Effekt gilt nur für bestimmte Fälle
C. Alle anderen Variablen werden konstant gehalten
D. Die Koeffizienten sind standardisiert

**Answer: C**
Ceteris paribus (“alles andere gleich”) bedeutet, dass der Koeffizient den Effekt einer Variable zeigt, während alle anderen Variablen konstant gehalten werden.

# Question 17

**Ein Interaktionsterm Preis$\times$Marke in der Regression bedeutet:**

A. Preis und Marke sind korreliert
B. Der Effekt des Preises hängt von der Marke ab
C. Preis und Marke haben den gleichen Effekt
D. Einer der beiden Effekte ist nicht signifikant

# Question 17

**Ein Interaktionsterm Preis$\times$Marke in der Regression bedeutet:**

A. Preis und Marke sind korreliert
B. Der Effekt des Preises hängt von der Marke ab
C. Preis und Marke haben den gleichen Effekt
D. Einer der beiden Effekte ist nicht signifikant

**Answer: B**
Ein Interaktionsterm modelliert, dass der Effekt einer Variable (Preis) unterschiedlich ist, je nach Ausprägung einer anderen Variable (Marke).

# Question 18

**Was ist der Unterschied zwischen $R^2$ und adjusted $R^2$?**

A. Adjusted $R^2$ bestraft zusätzliche Prädiktoren
B. Adjusted $R^2$ ist immer höher als $R^2$
C. Es gibt keinen Unterschied
D. Adjusted $R^2$ ignoriert Ausreißer

# Question 18

**Was ist der Unterschied zwischen $R^2$ und adjusted $R^2$?**

A. Adjusted $R^2$ bestraft zusätzliche Prädiktoren
B. Adjusted $R^2$ ist immer höher als $R^2$
C. Es gibt keinen Unterschied
D. Adjusted $R^2$ ignoriert Ausreißer

**Answer: A**
Adjusted $R^2$ korrigiert für die Anzahl der Prädiktoren und sinkt, wenn ein hinzugefügter Prädiktor die Erklärungskraft nicht verbessert. $R^2$ steigt immer.

# Question 19

**Overfitting tritt auf, wenn:**

A. Das Modell zu wenige Variablen hat
B. Das Modell die Trainingsdaten zu gut lernt und schlecht generalisiert
C. Die Daten zu wenig Varianz haben
D. Die Koeffizienten zu klein sind

# Question 19

**Overfitting tritt auf, wenn:**

A. Das Modell zu wenige Variablen hat
B. Das Modell die Trainingsdaten zu gut lernt und schlecht generalisiert
C. Die Daten zu wenig Varianz haben
D. Die Koeffizienten zu klein sind

**Answer: B**
Overfitting bedeutet, dass das Modell das Rauschen in den Trainingsdaten mitlernt und dadurch auf neuen Daten schlecht performt.

# Question 20

**Warum ist die Out-of-Sample-Performance wichtiger als die In-Sample-Performance?**

A. Out-of-Sample ist einfacher zu berechnen
B. Out-of-Sample zeigt die Generalisierungsfähigkeit auf ungesehene Daten
C. In-Sample-Metriken sind unzuverlässig
D. Es gibt keinen Unterschied

# Question 20

**Warum ist die Out-of-Sample-Performance wichtiger als die In-Sample-Performance?**

A. Out-of-Sample ist einfacher zu berechnen
B. Out-of-Sample zeigt die Generalisierungsfähigkeit auf ungesehene Daten
C. In-Sample-Metriken sind unzuverlässig
D. Es gibt keinen Unterschied

**Answer: B**
In-Sample-Performance ist immer optimistisch, da das Modell die Daten "kennt". Out-of-Sample zeigt, wie gut das Modell auf neue, ungesehene Daten generalisiert.