

DSAI Probe-Schularbeit (Mock Exam Worksheet): Big Data & Stochastik

Name: _____ Datum: _____ Punkte: ____ / 50 Note: ____ (Notenschlüssel: ab 44=1, 38=2, 30=3, 25=4)

Teil A: Big Data - Definition, 3 V's, Principles & Technology (18 Punkte)

A1. Definition & The 3 V's (6 Punkte): Nenne die „3 V's“ von Big Data und gib zu jedem V mindestens zwei konkrete Beispiele aus den Unterlagen an.

Nr.	Name des „V“ (English / Deutsch)	Mindestens 2 konkrete Beispiele / Merkmale
1. V _____		
2. V _____		
3. V _____		

A2. Velocity - Die 3 Datenzustände (3 Punkte): Erkläre den Unterschied zwischen Data in Use, Data in Motion und Data at Rest.

- Data in Use:
- Data in Motion:
- Data at Rest:

A3. Big Data Principles (6 Punkte): Erkläre die 4 Prinzipien „More“, „Messy“, „Correlation“ und „Datafication“ (inkl. der Beispiele Flugbuchung & Automotor).

- „More“ & „Messy“:
- „Correlation“ (inkl. Beispiel):
- „Datafication“ (inkl. Beispiel):

A4. Big Data Technology (3 Punkte): Nenne die 4 Technologien aus den Unterlagen und wofür „NoSQL“ steht.

1) _____ 2) NoSQL = „_____“ 3) _____ 4) _____

Teil B: Stochastik, Metriken & Daten-Beschaffenheit (16 Punkte)

B1. Aufbau der Stochastik & Population vs. Sample (4 Punkte): In welche 2 Hauptgebiete teilt sich die Stochastik, und welche 2 Teilgebiete hat die Statistik? Wie lautet die Größenbeziehung zwischen Sample S und Population P?

- Stochastik → 1) _____ 2) Statistik (a: _____, b: _____)
- Beziehung Sample S zu Population P: $S_1 \neq S_2 \dots \neq S_n$, $\sum S_i =$ _____, Größenvergleich: _____ $\ll$ _____

B2. Statistische Metriken (6 Punkte): Schreibe die Formeln für Mittelwert (Mean), Varianz (Variance) und Standardabweichung (Std. Deviation) auf. Wofür eignet sich besonders der Median?

- Mittelwert $\bar{x} =$ _____ • Varianz $s^2 =$ _____ • Std.-Abw. $s =$ _____
- Vorteil / Anwendung des Medians:

B3. Daten-Beschaffenheit & Data Science Tabellen (6 Punkte): Unterscheide Qualitative vs. Quantitative Daten (mit Beispielen). Wie macht man qualitative Daten berechenbar? Was steht in Data-Science-Tabellen in Zeilen (rows) bzw. Spalten (columns)?

- Qualitativ vs. Quantitativ + Umwandlung:
- 2D Data Table Convention: Rows (Zeilen) = _____, Columns (Spalten) = _____

Teil C: Praxis- & Rechenaufgaben - Histogramm (16 Punkte)

C1. Kurzaufgabe aus dem Handout (4 Punkte): Ein Datensatz hat das Minimum $\min = 12$ und das Maximum $\max = 65$. Es sollen $\#groups = 6$ Gruppen gebildet werden. Berechne die Range und die Bin (class) width!

• Step 1 (Range) = _____ • Step 2 (Bin width) = _____

C2. Hauptaufgabe Histogramm (12 Punkte): Gegeben ist eine Stichprobe von $n = 27$ Messwerten mit $\min = 152$ und $\max = 191$ (siehe Heft-Übung).

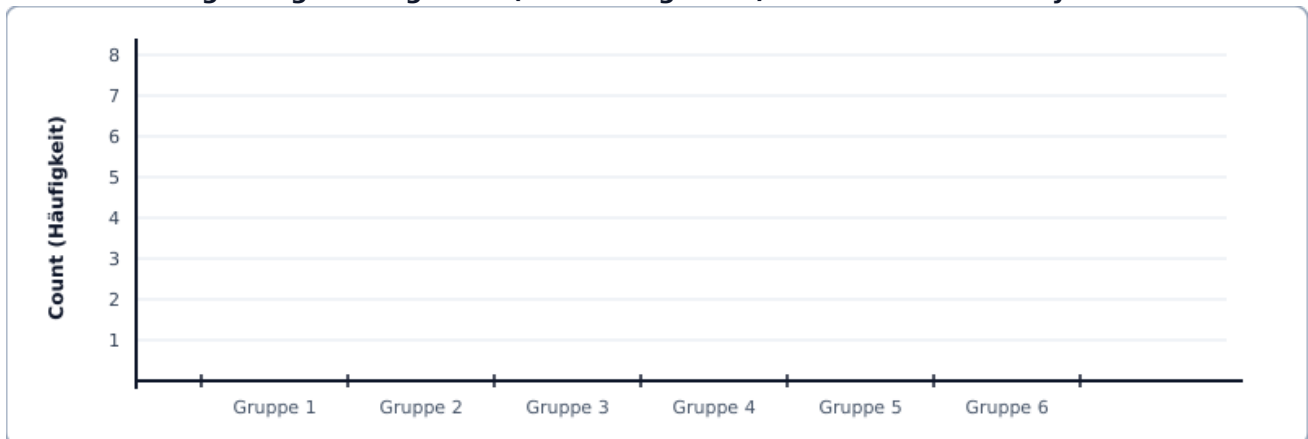
a) Berechne die Spannweite (Range): $Range = \max - \min =$ _____

b) Berechne die Gruppenanzahl mit Sturges' Rule $[1 + 3,322 \cdot \log_{10}(27)]$ und die Klassenbreite (Bin Width):

c) Vervollständige die Intervall-Tabelle (Start bei 152, je 7 Werte pro Intervall) für die gegebenen Counts (1, 6, 5, 8, 4, 3):

Gruppe (Group)	Intervall (von - bis)	Count (Häufigkeit)
1	152 - ____	1
2	____ - ____	6
3	____ - ____	5
4	____ - ____	8
5	____ - ____	4
6	____ - ____	3

d) Zeichne das zugehörige Histogramm (Balkendiagramm) in das Koordinatensystem:



Musterlösung & Bewertungsschlüssel (Answer Key - 50 Punkte)

Notenschlüssel: 44-50 Pkt. = Sehr gut (1) | 38-43 Pkt. = Gut (2) | 30-37 Pkt. = Befriedigend (3) | 25-29 Pkt. = Genügend (4) | <25 Pkt. = Nicht genügend (5)

Lösung A1 - The 3 V's of Big Data (6 Punkte):

- 1. Volume (Datenmenge):** Sensor data, Quality testing data, E-mails from customers and suppliers, Evaluation data on portals (web shops, web pages), Content of social networks.
- 2. Velocity (Geschwindigkeit):** Velocity of creation of data, Velocity of managing data (in real-time), Data in Use / Data in Motion / Data at Rest.
- 3. Variety (Datenvielfalt):** Heterogeneous data sources & formats (text, videos, images), Structured and non-structured data, Internal and external data.

Lösung A2 - Die 3 Datenzustände bei Velocity (3 Punkte):

- **Data in Use:** Active data stored in databases that are under constant changes (aktive Daten unter ständiger Veränderung).
- **Data in Motion:** Data stored temporarily in a computer that will soon be read or updated (temporäre Daten im Stream/RAM).
- **Data at Rest:** Inactive data stored in databases (data warehouse ...) or archives (ruhende Archiv-/Warehouse-Daten).

Lösung A3 - Die 4 Big Data Principles (6 Punkte):

- **“More”:** Not only collecting samples, but collecting complete information (alle Daten statt nur Stichproben).
- **“Messy”:** Accept that data quality will be low; the large amount of data helps to get correct results.
- **“Correlation”:** Analyze WHAT, but not WHY. Beispiel: Nicht das Prinzip des Yield Managements von Airlines analysieren, sondern einfach die beste Zeit für die Flugbuchung finden (best time for booking).
- **“Datafication”:** Everything is information — collect it and use it. Beispiel: Vibrationen eines Automotors sammeln und nutzen, um vorherzusagen, wann der Motor ausfällt (forecast that engine breaks down).

Lösung A4 - Big Data Technology (3 Punkte):

- 1) Usage of hundreds or thousands of parallel running computers | 2) **NoSQL = “Not Only SQL”** | 3) **MapReduce (Hadoop map)** - parallel and distributed algorithm | 4) **In-memory analytics**.

Lösung B1 - Stochastik & Population vs. Sample (4 Punkte):

- **Stochastik** gliedert sich in **1) Wahrscheinlichkeitstheorie** und **2) Statistik (a: Deskriptive Statistik, b: Inferenz-Statistik / Schließende Statistik)**.
- **Population P vs. Sample S:** $S_1 \neq S_2 \dots \neq S_n$, $\sum S_i = P$, und $S_size \ll P_size$ (Stichprobe ist viel kleiner als die Population).

Lösung B2 - Statistische Metriken (6 Punkte):

- **Mittelwert (Mean):** $\bar{x} = (1/n) \cdot \sum x_i$ (Populationssymbol: μ)
- **Varianz (Variance / Streuung):** $s^2 = (1/n) \cdot \sum (x_i - \bar{x})^2$ (Populationssymbol: σ^2)
- **Standardabweichung (Std. Deviation):** $s = \sqrt{s^2}$ (hat wieder dieselbe Einheit wie die Daten x_i)
- **Median:** Zentralwert der sortierten Daten; ideal **für die Ausreißer-Erkennung (Outlier Detection)**, da er robust gegen Extremwerte ist.

Lösung B3 - Daten-Beschaffenheit & Data Science Tabellen (6 Punkte):

- **Qualitative Daten:** Kategorien/Labels (z. B. Haarfarbe: blond, schwarz; Auto; Bewertung: gut, mittel, schlecht) → man kann nicht direkt damit rechnen → Umwandlung in Zahlen durch **Encoding**!
- **Quantitative Daten:** „Zahlen“ → man kann direkt damit rechnen.
- **2D-Tabellen-Konvention:** **Rows (Zeilen) = Instances, Columns (Spalten) = Features**.

Lösung C1 & C2 - Histogramm-Aufgaben (16 Punkte):

- **C1 (Handout):** Step 1: Range = 65 - 12 = **53**. Step 2: Range / Groups = 53 / 6 = 8,83 → aufgerundet **Bin Width = 9**.
- **C2 (Heft-Übung, n = 27):** a) Range = 191 - 152 = **39**. b) Sturges' Rule: $1 + 3,322 \cdot \log_{10}(27) = 5,75 \approx$ **6 Gruppen**; Klassenbreite = $\lceil 39 / 6 \rceil =$ **7 Werte pro Intervall**.
- **c) Intervalle:** Gruppe 1: **152-158** (Count 1) | Gruppe 2: **159-165** (Count 6) | Gruppe 3: **166-172** (Count 5) | Gruppe 4: **173-179** (Count 8) | Gruppe 5: **180-186** (Count 4) | Gruppe 6: **187-193** (Count 3). Summe = 27.

