Aufgabenstellung 3. Teilleistung Gruppenarbeit / GitHub

Bitte bearbeiten Sie diese Aufgabe in einer **Gruppe von 4-6 Personen**. Eine Liste mit ihren Gruppenmitgliedern finden Sie auf Moodle.

Bitte nutzen Sie für die Bearbeitung der im folgenden geschilderten Aufgabe GitHub und Git und die in Sitzung 7 erlernten Inhalte. Die Bewertung dieser Teilleistung wird an der Anwendung der in Sitzung 7 erlernten Inhalte (sowohl des Vortrags als auch der Übungsaufgabe) festgemacht. Ihre Skripte müssen **durchlaufen ohne Fehler und die Aufgabestellung richtig umsetzen**, allerdings wird darüber hinaus Ihr Programmieren nicht streng bewertet (die Dokumentation Ihrer Funktionen und Skripte allerdings schon!). Wichtig für die Bewertung ist, dass Sie als **ganze Gruppe** unter Beteiligung **aller Mitglieder** die Aufgabe über ein **GitHub Repository** bearbeiten und auf diesem erkennbar ist, dass Sie über einen **längeren Zeitraum gemeinsam** an der Aufgabe gearbeitet haben und damit die **in Sitzung 7 kennen gelernten Workflows** zur Verwendung von GitHub eingesetzt haben.

Abzugeben ist über Moodle für Ihre Gruppe **ein Link zu Ihrem GitHub Repository**. Stellen Sie sicher, dass ich dieses einsehen kann. Anhand von diesem wird die Bewertung stattfinden. **Abgabefrist für den Link: 28.02.2023, 23.59h**.

Über Ihr GitHub Repository unter Verwendung von Git und in gleichbeteiligter Gruppenarbeit, bearbeiten Sie bitte die folgenden Aufgaben:

1. **Zwei** Gruppenmitglieder erstellen ein R-Skript (verfügbar auf Ihrem Repository), in welchem Sie einen Datensatz simulieren. Dieser soll Beobachtungen von 100 Personen auf 5 Variablen beinhalten, sowie eine ID-Spalte. Die Variablen sollen sein:
    - (1) *Alter*, simuliert aus einer Normalverteilung mit Erwartungswert 25 und einer Standardabweichung von 2,
    - (2) *Studienfach*, zufällig gezogen für alle Personen aus einer Auswahl von „Statistik“, „Data Science“, „Mathe“ und „Informatik“, wobei die Fächer „Statistik“ und „Data Science“ mit gleicher Wahrscheinlichkeit studiert werden sollen, „Informatik“ mit einer etwas geringeren Wahrscheinlichkeit und „Mathe“ mit der geringsten Wahrscheinlichkeit,
    - (3*) Interesse an Mathematik*, wobei hier ein von Ihnen frei zu wählender Zusammenhang mit Studienfach bestehen darf (aber nicht muss); hier können Sie überlegen, welchen Zusammenhang Sie für plausibel halten und entsprechend simulieren; die Variable soll nur Werte $\in \{1,2,3,4,5,6,7\}$ annehmen können (wie Sie dafür sorgen ist Ihnen frei überlassen), wobei 1 = sehr geringes Interesse, und 7 = sehr hohes Interesse abbildet,
    - (4) *Interesse an Programmieren*, wobei hier ein von Ihnen frei zu wählender Zusammenhang mit Studienfach bestehen darf (aber nicht muss); hier können Sie überlegen, welchen Zusammenhang Sie für plausibel halten und entsprechend simulieren, die Variable soll nur Werte $\in \{1,2,3,4,5,6,7\}$ annehmen können (wie Sie dafür sorgen ist Ihnen frei überlassen), wobei 1 = sehr geringes Interesse, und 7 = sehr hohes Interesse abbildet,
    - (5) *Mathe-LK (ja/nein),* eine dichotome Variable, die kodiert, ob jemand in der Schule Mathe-LK hatte oder nicht; hierbei darf (muss aber nicht) ein

Zusammenhang Ihrer Wahl mit Studienfach, Interesse an Mathematik und Interesse an Programmieren bestehen.
    - Diejenigen Gruppenmitglieder, die für (1.) zuständig sind, sollen **die anderen nicht über die simulierten Zusammenhänge informieren.** D.h., die anderen Mitglieder sollen sich dieses in (1.) erstellte Skript nicht angucken bevor Aufgabe (4.) nicht erledigt ist.
2. Der Datensatz soll als csv-Datei in Ihrem GitHub-Repository gespeichert werden.
3. **Gemeinsam als ganze Gruppe** erstellen Sie bitte zwei weitere R-Skripte. Im 4. Schritt sollen diejenigen Gruppenmitglieder, die nicht an (1.) gearbeitet haben, den Datensatz analysieren (Deskription und Visualisierung). Hierzu sollen nun Funktionen erstellt werden, die dabei genutzt werden.
    - Funktionen-R-Skript 1 soll (mindestens) folgende Funktionen enthalten:
        - (a) Eine Funktion, die verschiedene geeignete deskriptive Statistiken für metrische Variablen berechnet und ausgibt
        - (b) Eine Funktion, die verschiedene geeignete deskriptive Statistiken für kategoriale Variablen berechnet und ausgibt
        - (c) Eine Funktion, die geeignete deskriptive bivariate Statistiken für den Zusammenhang zwischen zwei kategorialen Variablen berechnet ausgibt
        - (d) Eine Funktion, die geeignete deskriptive bivariate Statistiken für den Zusammengang zwischen einer metrischen und einer dichotomen Variablen berechnet und ausgibt
        - (e) Eine Funktion, die eine mindestens ordinal skalierte Variable quantilbasiert kategorisiert (z.B. in „niedrig“, „mittel“, „hoch“)
        - (f) Eine Funktion, die eine geeignete Visualisierung von drei oder vier kategorialen Variablen erstellt
        - Freiwillig: weitere zur Deskription und Visualisierung geeignete Funktionen
    - Funkionen-R-Skript 2 soll Helfer-Funktionen enthalten, die nicht selbst zur Deskription und Visualisierung der Daten verwendet werden, sondern die nur in Funktionen-Skript 1 Anwendung finden (→ interne Funktionen). Funktionen-R-Skript 2 muss mindestens eine Funktion enthalten.
    - Denken Sie auch an eine gute Dokumentation aller Funktionen. Nutzen Sie Ihr GitHub Repository um darüber zu diskutieren, welche Funktionen sinnvoll und notwendig sind.
4. **Diejenigen Gruppenmitglieder, die nicht an (1.) gearbeitet haben**, sollen dann mit Hilfe der in (3.) in Funktionen-R-Skript 1 erstellten Funktionen den Datensatz aus (2.) analysieren (Deskription und Visualisierung). Hierzu soll ein viertes Skript in Ihrem Repository erstellt werden. Hierbei sollte Ihr R-Skript im Minimum jede der Funktionen (a) bis (f) aus Funktionen-R-Skript 1 einmal anwenden.
5. Diskutieren Sie über Ihr GitHub Repository **als ganze Gruppe** Ihre Ergebnisse. Die Gruppenmitglieder, die an (1.) gearbeitet haben, können dann hier Aufschluss über die wahren zugrunde gelegten Zusammenhänge geben.