Übung: Experten der Datentransformation

In dieser Übung werden alle Studierenden im Modul einer von 4 Gruppen zugeordnet. Jede Gruppe hat die Aufgabe, sich im Selbststudium auf eine der vier Arten der Datentransformation zu spezialisieren:

1. Auswahl von Spalten mit `select`
2. Filtern von Zeilen mit `filter`
3. Veränderung bestehender Spalten und Hinzufügen neuer Spalten mit `mutate` und `transmute`
4. Gruppierung und Zusammenfassung von Zeilen mit `group_by` und `summarize`

Gruppe A: Expert*innen für `select`

Ihr werdet euch intensiv mit dem Auswählen von Spalten mittels der `select`-Funktion aus dem `{dplyr}`-Paket befassen. Konzentriert euch dabei besonders auf die folgenden Aspekte:

- Wie verschafft man sich einen Überblick über verfügbare Spalten?
- Wie wählt man Spalten anhand ihres Namens aus?
- Wie kann man effizient nach Namensmustern suchen und auswählen?
- Wie kann man Spalten *ausschließen* (abwählen) anstatt auszuwählen?
- Ist es möglich, Spalten nach ihrem Datentyp zu selektieren?
- Wie funktioniert die Auswahl anhand der Position einer Spalte?
- Kann man Spalten anhand der Definition einer Menge mit der `c()`-Funktion auswählen?

Eure Aufgabe: Überlegt, wie ihr eure gewonnenen Erkenntnisse zum Auswählen von Spalten am besten euren Kommiliton*innen vermitteln könnt. Nutzt dazu anschauliche Codebeispiele anhand eines Datensatzes eurer Wahl. Achtet darauf, dass der gewählte Datensatz eure Beispiele unterstützt. Ihr könnt auch visuelle Hilfsmittel wie Folien oder Flipcharts verwenden. Ihr habt insgesamt 30 Minuten Zeit, um eure Ergebnisse zu präsentieren.

Gruppe B: Expert*innen für `filter`

Ihr werdet euch intensiv mit dem Filtern von Zeilen mittels der `filter`-Funktion aus dem `dplyr`-Paket auseinandersetzen. Fokussiert euch dabei insbesondere die folgenden Teilaspekte:

- Wie kann man auf Basis einfacher Vergleiche filtern?
- Auf welche Weise kann man numerische Werte filtern?
- Kann man auf Basis einer Zugehörigkeit zu einer Menge filtern?
- Wie erzeuge ich komplexe Filterbedingungen mit logischen Verknüpfungen?
- Wie finde und entferne ich fehlende Werte?
- Wie kann ich Filterbedingungen verneinen?
- Kann ich auch in unstrukturierten Texte suchen und danach Filtern?

Eure Aufgabe: Überlegt euch, wie ihr die gewonnenen Erkenntnisse zum Aus- oder Abwählen von Zeilen am einfachsten euren Kommiliton*innen erklären könnt. Nutzt dafür anschauliche Codebeispiele anhand eines Datensatzes eurer Wahl. Stellt sicher, dass der gewählte Datensatz eure Beispiele unterstützt. Auch visuelle Hilfsmittel wie Slides oder Flipcharts sind erlaubt. Ihr bekommt insgesamt 30 Minuten für die Vorstellung eurer Ergebnisse.

Gruppe C: Expert*innen für `mutate` und `transmute`

Ihr steigt tief in das Verändern bestehender und das Erzeugen neuer, berechneter Spalten mittels der `mutate`- oder `transmute`-Funktion aus dem `dplyr`-Paket ein. Schaut euch dabei insbesondere die folgenden Teilaspekte genauer an:

- Wie kann ich einfache berechnete Spalten erzeugen?
- Wie kann ich Bestandteile eines Datumswerts in eine neue Spalte extrahieren (z. B. nur das Jahr)?
- Wie kann ich bestimmen, an welcher Stelle eine neue Spalte eingefügt wird?
- Wie erreiche ich es, dass nur die neuen Spalten im Ergebnis erhalten bleiben?
- Wie erreiche ich es, dass neben der neuen auch die in der Berechnung verwendeten Spalten im Ergebnis erhalten bleiben?
- Wie kann ich den Datentyp existierender Spalten verändern?
- Wie kann ich eine Spalte umbenennen?

Eure Aufgabe: Überlegt euch, wie ihr die gewonnenen Erkenntnisse zum Verändern oder Neuhinzufügen von Spalten am einfachsten euren Kommiliton*innen erklären könnt. Nutzt dafür anschauliche Codebeispiele anhand eines Datensatzes eurer Wahl. Stellt sicher, dass der gewählte Datensatz eure Beispiele unterstützt. Auch visuelle Hilfsmittel wie Slides oder Flipcharts sind erlaubt. Ihr bekommt insgesamt 30 Minuten für die Vorstellung eurer Ergebnisse.

Gruppe D: Expert*innen für `group_by` und `summarize`

Ihr seid für die letzte Form der Datentransformation zuständig und studiert die Möglichkeiten, Daten mithilfe von Funktionen aus dem `dplyr`-Paket zu gruppieren und zusammenzufassen. Beleuchtet dabei insbesondere die folgenden Teilaspekte genauer:

- Wie kann man einfache Aggregationen, wie die Summe einer Spalte über alle Einträge oder die Anzahl Datensätze insgesamt, erstellen?
- Wie gruppiere ich meine Daten nach einem Merkmal und berechne anschließend eine aggregierte Kennzahl wie den Durchschnitt oder die Summe?
- Wie berechne ich die relative Häufikeit für jede Gruppe, zum Beispiel den prozentualen Anteil pro Geschlecht?
- Was machen die Funktionen `distinct`, `count` und `tally`?
- Wie kann ich schnell eine Übersicht über die häufigsten Werte einer nominal skalierten Variable bekommen?
- Was macht die `tabyl`-Funktion aus dem `janitor`-Paket?

Eure Aufgabe: Überlegt euch, wie ihr die gewonnenen Erkenntnisse zum Gruppieren und Zusammenfassen von Zeilen am einfachsten euren Kommiliton*innen erklären könnt. Nutzt dafür anschauliche Codebeispiele anhand eines Datensatzes eurer Wahl. Stellt sicher, dass der gewählte Datensatz eure Beispiele unterstützt. Auch visuelle Hilfsmittel wie Slides oder Flipcharts sind erlaubt. Ihr bekommt insgesamt 30 Minuten für die Vorstellung eurer Ergebnisse.

Hilfsmittel

Sucht nach eigenen Quellen, die euch am besten weiterhelfen, und teilt am Ende eurer Einführung eine kleine Linkliste mit euren Kommiliton*innen. Schaut aber unbedingt auch hier vorbei:

- Wickham, Hadley, et al. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. 2nd edition, O'Reilly Media, Inc, 2023 (Online frei verfügbar).
- Offizielle Webseite des `dplyr`-Pakets
- ChatGPT von OpenAI