Übung: Daten visualisieren

Am Beispiel der Tagesschau-Daten

Prof. Dr. Nicolas Meseth

Eine gute Visualisierung macht sichtbar, was in Tabellen und Kennzahlen verborgen bleibt. Aber Visualisierungen können auch täuschen: durch abgeschnittene Achsen, ungeschickte Filterentscheidungen oder suggestive Skalierungen. In dieser Übung lernt ihr zunächst die wichtigsten Diagrammtypen für Häufigkeiten, Gruppenvergleiche, Zeitreihen und Zusammenhänge kennen. Anschließend setzt ihr euch gezielt mit typischen Fallstricken auseinander, die auch erfahrenen Analytikerinnen und Analytikern unterlaufen.

Schritt 1: Daten laden und vorbereiten

Diese Übung baut auf der vorherigen Übung zur Datentransformation auf. Stellt sicher, dass euer Projekt bereits eingerichtet ist und der Datensatz in `data/` liegt.

1. Ladet die benötigten Pakete mit `pacman::p_load()`: `tidyverse`, `ggridges` und `scales`. Lest anschließend den Datensatz ein und erstellt in einem Schritt alle abgeleiteten Variablen, die ihr für diese Übung benötigt: `year`, `month`, `weekday` (mit `lubridate::wday(..., label = TRUE)`) sowie `title_length` (Zeichenanzahl des Titels). Erstellt außerdem den reduzierten Datensatz `news_slim` und den aggregierten `news_timeline`.

Schritt 2: Häufigkeiten und Verteilungen

Die grundlegenden Visualisierungsformen für einzelne Variablen sind das Balkendiagramm für kategoriale und das Histogramm für metrische Variablen.

2. Erstellt ein Balkendiagramm, das die Häufigkeitsverteilung der Ressorts zeigt. Filtert vorab auf die sechs Hauptressorts (`"inland"`, `"ausland"`, `"wirtschaft"`, `"wissen"`, `"investigativ"`, `"faktenfinder"`). Die Balken sollen nach Häufigkeit *absteigend* sortiert sein.

Beschriftet beide Achsen sinnvoll, wählt einen aussagekräftigen Diagrammtitel und verwendet ein schlichtes Theme.

3. Erstellt ein Histogramm der Variable `word_count`. Filtert zuvor Artikel mit mehr als 3.000 Wörtern heraus, damit die Verteilung übersichtlich bleibt. Experimentiert mit verschiedenen `bins`-Werten, zum Beispiel 30, 60 und 100, um eine aussagekräftige Klasseneinteilung zu finden.

Fügt dem Histogramm mit `geom_vline()` eine vertikale Linie für den Median hinzu. Beschreibt die Verteilung. Ist sie symmetrisch, links- oder rechtsschief? Was sagt das über typische Artikellängen auf tagesschau.de aus?

4. Untersucht, welche Tags am häufigsten vergeben wurden. Filtert zunächst alle Tags heraus, die mindestens 200 Mal vorkommen, und entfernt `NA`-Werte.

Achtung: Die Tags `"FAKTENFINDER"` und `"#FAKTENFINDER"` bezeichnen dasselbe Konzept. Bereinigt das mit `str_remove_all()`, bevor ihr zählt.

Stellt die Top-Tags als sortiertes Balkendiagramm dar, absteigend nach Häufigkeit. Welche thematischen Schwerpunkte lassen sich aus der Tag-Verteilung ablesen?

Schritt 3: Gruppenvergleiche visualisieren

Nicht nur einzelne Variablen, sondern Unterschiede zwischen Gruppen sind oft am aufschlussreichsten. Boxplots, Violinplots und Ridgeline-Diagramme eignen sich dafür besonders gut. Heatmaps helfen, zwei kategoriale Dimensionen gleichzeitig zu erkunden.

5. Erstellt einen Boxplot, der die Verteilung der Wortanzahl für die sechs Hauptressorts zeigt. Filtert Artikel mit mehr als 3.000 Wörtern heraus. Sortiert die Ressorts nach ihrem Median, Tipp: `fct_reorder(ressort, word_count, .fun = median)`, sodass das Ressort mit dem niedrigsten Median oben steht. Verwendet `coord_flip()` für eine horizontale Darstellung.

Welches Ressort hat die größte Streuung? Welches die kleinste?

6. Ergänzt den Boxplot aus Aufgabe 5 um überlagerte Datenpunkte. Setzt `alpha = 0.1` und `size = 0.5`, um Überlappungen zu reduzieren. Achtet auf die Reihenfolge der `geom_`-Aufrufe. Der Boxplot sollte *über* den Punkten liegen, nicht darunter.

Ab welcher Datenmenge gerät diese Darstellung an ihre Grenzen? Was ist der Vorteil dieser kombinierten Form gegenüber einem reinen Boxplot? Schaut euch mal `geom_sina`

aus dem `ggforce`-Paket an.

7. Installiert das Paket `ggridges` und erstellt mit `geom_density_ridges()` ein Ridgeline-Diagramm der Wortanzahl-Verteilung für alle Hauptressorts. Die Filterung auf `word_count < 3000` bleibt bestehen. Füllt die Flächen mit `fill = ressort` und wählt eine ansprechende Farbpalette.

Beschreibt den Unterschied zwischen einem Boxplot und einem Ridgeline-Diagramm. Welche Informationen zeigt jede Darstellungsform, welche verbirgt sie? Wann würdet ihr welche Form bevorzugen?

8. Erstellt eine Heatmap, die zeigt, wie viele Artikel an welchem Wochentag in welchem Hauptressort erschienen. Berechnet dafür zunächst die Häufigkeiten pro Wochentag und Ressort mit `group_by()` und `summarize()`. Visualisiert das Ergebnis mit `geom_tile()` und `scale_fill_viridis_c()`.

Zeigt die Wochentage auf der x-Achse, Montag bis Sonntag, die Ressorts auf der y-Achse. Was fällt euch beim Wochenendverhalten auf? Unterscheiden sich die Ressorts darin?

Schritt 4: Zeitliche Entwicklungen visualisieren

Der Tagesschau-Datensatz erstreckt sich über fast zwei Jahrzehnte. Zeitreihenvisualisierungen helfen uns, Trends, Wachstum und saisonale Muster zu erkennen.

9. Erstellt ein Liniendiagramm, das die Gesamtanzahl der Artikel pro Jahr zeigt. Markiert die Datenpunkte zusätzlich mit `geom_point()`. Beschriftet die y-Achse sinnvoll und dreht die x-Achsenbeschriftungen, falls nötig.

Welche Jahre stechen besonders hervor? Versucht, den starken Anstieg ab 2023 inhaltlich zu erklären. Berücksichtigt dabei, dass 2006 und 2026 möglicherweise keine vollständigen Jahrgänge sind. Wie könnte man diesen Sachverhalt im Diagramm kenntlich machen?

10. Erstellt ein Liniendiagramm, das die Artikelanzahl pro Jahr *für die drei Ressorts `"inland"`, `"ausland"` und `"wirtschaft"` gleichzeitig* als drei farbige Linien zeigt. Fügt eine Legende hinzu und wählt aussagekräftige Farben mit `scale_color_manual()` oder einer vorgefertigten Palette.

Welches Ressort ist in den letzten Jahren am stärksten gewachsen? Gibt es Phasen, in denen ein Ressort vorübergehend besonders aktiv war?

11. Erstellt ein gestapeltes Flächendiagramm, `geom_area()`, der Artikelanzahl pro Jahr für die sechs Hauptressorts. Wählt `position = "stack"`, also gestapelt. Vergleicht die Darstellung mit dem Liniendiagramm aus Aufgabe 10.

Was lässt sich aus der Gesamtfläche ablesen, was im Liniendiagramm nicht sichtbar war? Welche Darstellungsform ist besser geeignet, um den Anteil einzelner Ressorts am Gesamtvolumen zu zeigen?

12. Erstellt eine Heatmap, die zeigt, wie viele Artikel pro Jahr *und Monat* erschienen. Die x-Achse soll die Monate zeigen, Januar bis Dezember, die y-Achse die Jahre, mit dem neuesten Jahr oben, Tipp: Faktor mit umgekehrter Reihenfolge. Verwendet `geom_tile()` und `scale_fill_viridis_c(option = "plasma")`.

Lassen sich saisonale Muster erkennen? Ab welchem Jahr nimmt die Datendichte deutlich zu? Was könnte das über die Entstehungsgeschichte des Datensatzes verraten?

Schritt 5: Zusammenhänge und mehrdimensionale Analysen

Scatter-Plots sind das klassische Werkzeug, um Zusammenhänge zwischen zwei metrischen Variablen zu erkunden. Wenn eine dritte Variable ins Spiel kommt, setzen wir Farbe oder Facettierung ein.

13. Gibt es einen Zusammenhang zwischen der Länge eines Titels `title_length` und der Länge des zugehörigen Artikels `word_count`? Zieht zunächst mit `slice_sample(n = 3000)` ein Zufallssample, um die Darstellung übersichtlich zu halten. Erstellt dann einen Scatter-Plot der beiden Variablen und fügt mit `geom_smooth(method = "lm")` eine lineare Trendlinie hinzu.

Beschreibt Stärke und Richtung des sichtbaren Zusammenhangs. Überprüft euren visuellen Eindruck mit `cor()`. Wie stark ist die Korrelation tatsächlich? Ist das Ergebnis überraschend, und warum oder warum nicht?

14. Erweitert den Scatter-Plot aus Aufgabe 13 um eine dritte Dimension. Färbt die Punkte nach `ressort` ein, nur die sechs Hauptressorts. Verwendet anschließend `facet_wrap(~ ressort)`, um für jedes Ressort ein eigenes Teildiagramm zu erzeugen.

Unterscheidet sich der Zusammenhang zwischen Titellänge und Artikellänge zwischen den Ressorts?

15. Untersucht den Tagesrhythmus der Redaktion. Berechnet für jede Stunde des Tages, 0 bis 23 Uhr, die Anzahl der veröffentlichten Artikel. Visualisiert das Ergebnis als Flächen- oder Balkendiagramm.

Wann ist die Redaktion am produktivsten? Wann ist die Aktivität am niedrigsten? Gibt es Muster, die auf einen typischen Redaktionsalltag hindeuten?

Schritt 6: Fallstricke bei Visualisierungen

Visualisierungen lügen selten — aber sie täuschen oft. Und meistens liegt das nicht an böser Absicht, sondern an unüberlegten Standardentscheidungen: einer Achse, die nicht bei Null beginnt, einem Filter, der stillschweigend Datenpunkte entfernt, oder einer Skalierung, die Trends dramatischer wirken lässt als sie sind. In diesem Schritt reproduziert ihr typische Fehler absichtlich und korrigiert sie anschließend.

16. Erstellt zwei Versionen desselben Balkendiagramms der Ressort-Häufigkeiten (Hauptressorts, sortiert nach Häufigkeit):

- Version A: Die y-Achse beginnt bei 15.000 (z.B. mit `coord_cartesian(ylim = c(15000, NA))`).
- Version B: Die y-Achse beginnt bei 0.

Beschreibt, wie sich der visuelle Eindruck unterscheidet. Welches Ressort wirkt in Version A wie viel Mal häufiger als ein anderes, obwohl der tatsächliche Unterschied viel kleiner ist?

Edward Tufte hat dafür das Prinzip des *Proportional Ink* formuliert: Die Menge an Tinte (oder Pixeln), die eine Datengröße repräsentiert, soll proportional zu dieser Größe sein. Erklärt in eigenen Worten, warum Version A dieses Prinzip verletzt. Wann ist es dagegen akzeptabel, die Achse nicht bei 0 beginnen zu lassen?

17. Ihr habt in Aufgabe 3 den Filter `word_count < 3.000` angewendet, bevor ihr das Histogramm gezeichnet habt. Wie stark beeinflusst dieser Filter das Bild?

Erstellt zwei Boxplots der Wortanzahl für die Hauptressorts nebeneinander: einen ohne Filter und einen mit dem Filter `word_count < 3000`. Berechnet außerdem, wie viel Prozent der Artikel pro Ressort durch den Filter entfernt werden.

Verändert der Filter Median oder IQR nennenswert? Ändert er die *Interpretation* der Ressortunterschiede? Was sollte man dokumentieren, wenn man einen solchen Filter anwendet?

18. Ein Boxplot zeigt Median, Quartile und Ausreißer, aber er verbirgt die Form der Verteilung. Stellt die Wortanzahl für das Ressort `"ausland"` auf drei Arten dar: als Histogramm, als Boxplot und als Violinplot (`geom_violin()`). Filtert jeweils auf `word_count < 3000`.

Was sieht man im Histogramm, das der Boxplot nicht zeigt? Was leistet der Violinplot im Vergleich? In welchen Situationen würdet ihr trotzdem einen reinen Boxplot wählen?

19. Zeitreihen reagieren besonders empfindlich auf die Wahl der Achsenskalierung. Nehmt die jährliche Artikelanzahl für die Jahre 2015 bis 2025 und erstellt zwei Liniendiagramme:

- Version A: y-Achse beginnt bei 0.
- Version B: y-Achse zeigt nur den Bereich der tatsächlichen Datenwerte, z.B. mit `coord_cartesian(ylim = c(2000, NA))`.

Welche Version lässt den Wachstumstrend dramatischer wirken? Ist Version B "falsch"? Wann ist es bei Liniendiagrammen vertretbar, die y-Achse nicht bei 0 zu beginnen? Wie könnte man Leser darauf hinweisen, dass die Achse ausgeschnitten ist?

20. Doppelte y-Achsen (*dual axes*) sind verlockend, wenn zwei Zeitreihen sehr unterschiedliche Skalen haben. Versucht, die jährliche Artikelanzahl für `"inland"` und `"ausland"` in einem einzigen Diagramm mit zwei y-Achsen darzustellen. In ggplot2 geht das über `sec_axis()` in `scale_y_continuous()`.

Warum ist diese Darstellung grundsätzlich problematisch? Konstruiert ein Beispiel, in dem die visuelle Aussage über den Zusammenhang beider Reihen durch die Wahl des Skalierungsfaktors umgekehrt werden kann. Erstellt dann ein *Small Multiples*-Diagramm mit `facet_wrap()` als sauberere Alternative und vergleicht beide.

Schritt 7: Reflexion

21. Erklärt in eigenen Worten den Unterschied zwischen `summarize()` und `mutate()` in Kombination mit `group_by()`. Nennt je ein konkretes Beispiel aus dieser Übung, bei dem ihr lieber `summarize()` und eines, bei dem ihr lieber `group_by() + mutate()` eingesetzt habt, und begründet eure Wahl.

22. Ihr habt in dieser Übung fast zwei Jahrzehnte Tagesschau-Berichterstattung analysiert. Welche drei Beobachtungen haben euch am meisten überrascht oder sind euch besonders aufgefallen? Versucht, jede Beobachtung auch inhaltlich, über das bloße Zahlenmuster

hinaus, zu erklären oder einzuordnen.

23. Betrachtet alle Visualisierungsformen aus dieser Übung: Histogramm, Balkendiagramm, Boxplot, Boxplot mit Punktewolke, Violinplot, Ridgeline-Plot, Heatmap, Liniendiagramm, Flächendiagramm und Scatter-Plot. Erstellt eine eigene gedankliche Klassifikation. Welche Form ist für welche Kombination aus Fragestellung (univariat, bivariat, trivariat) und Variablentyp (metrisch, kategorial, zeitlich) am besten geeignet? Gibt es Überschneidungen, in denen mehrere Formen in Frage kämen? Was bestimmt dann die Wahl?

24. Datenqualität ist ein unterschätztes, aber entscheidendes Thema. Nennt mindestens vier konkrete Hinweise auf potenzielle Qualitätsprobleme oder Einschränkungen im Tagesschau-Datensatz, die euch im Laufe dieser Übung aufgefallen sind. Denkt dabei an fehlende Werte, uneinheitliche Kodierungen, unvollständige Jahrgänge, Veränderungen der Datenquelle über die Zeit und mögliche Selektionseffekte. Wie würdet ihr diese Probleme in einer veröffentlichungswürdigen Analyse dokumentieren und behandeln?

25. In dieser Übung haben wir ausschließlich mit den strukturierten Merkmalen im Datensatz gearbeitet. Wie könnten wir die unstrukturierten Merkmale wie den Artikeltext oder die Titel in unsere Analysen einbeziehen? Was müssten wir vorher tun?