Fachhochschule Nordwestschweiz
Hochschule für Informatik

GDV - Grundlagen der Datenvisualisierung

Datenvisualisierung von Immobilienverkäufen (2001–2022)

Abbildung 1: Hartford, Connecticut, USA Downtown Skyline.[1]

Sandra Senn
BSc Data Science Studentin

Windisch, Frühlingssemester 2025

[1] https://portal.ct.gov/choosect/

Inhaltsverzeichnis

Einführung

Datenvisualisierung ist ein zentrales Werkzeug, um komplexe Zusammenhänge in grossen Datensätzen sichtbar und verständlich zu machen. Gerade in einer Zeit, in der immer mehr Daten gesammelt und ausgewertet werden, ist es entscheidend, Informationen nicht nur korrekt, sondern auch übersichtlich und zielgruppengerecht zu präsentieren. Visualisierungen helfen dabei, Muster zu erkennen, Ausreisser zu identifizieren und Entwicklungen über die Zeit zu verfolgen[2]. Sie sind unverzichtbar, um datenbasierte Entscheidungen zu unterstützen, Forschungsergebnisse zu vermitteln oder neue Hypothesen zu entwickeln.

Dieses Dokument verfolgt das Ziel, die Grundlagen der Datenvisualisierung anhand eines praxisnahen Beispiels systematisch zu erläutern. Es richtet sich an alle, die verstehen möchten, wie aus Rohdaten durch gezielte Aufbereitung und bewusstes Design aussagekräftige Grafiken entstehen. Im Mittelpunkt steht dabei nicht nur die technische Umsetzung, sondern auch das Verständnis für die theoretischen Hintergründe, die Auswahl geeigneter Diagrammtypen, die Berücksichtigung menschlicher Wahrnehmung und die kritische Reflexion der eigenen Visualisierungen. Das Dokument führt einen Schritt für Schritt von der Datenaufbereitung über die Gestaltung bis hin zur Evaluation der Visualisierungen und bietet damit einen umfassenden Leitfaden für eigene Projekte im Bereich Data Science oder Business Analytics.

Als Datengrundlage dient ein umfangreicher Immobilienverkaufsdatensatz aus Connecticut (USA), der Transaktionen aus den Jahren 2001 bis 2022 umfasst. Der Datensatz enthält detaillierte Informationen zu jeder Immobilientransaktion. Neben dem Verkaufsdatum und dem Verkaufspreis werden unter anderem Stadtname, Adresse, Immobilien- und Nutzungstyp, Schätzwert, Sales Ratio sowie verschiedene Bemerkungen der Gutachter und der Behörden erfasst. Zusätzlich sind geografische Koordinaten vorhanden, sodass auch räumliche Analysen möglich sind. Die Daten bieten vielfältige Analysemöglichkeiten. Von Preisentwicklung über die Identifikation von Hotspots bis hin zur Untersuchung von Zusammenhängen zwischen Lage, Immobilienart und Verkaufserfolg ist alles möglich.[3]

Die folgenden Kapitel zeigen, wie diese Rohdaten bereinigt, transformiert und für verschiedene Visualisierungstypen aufbereitet werden. Dabei wird sowohl auf die Besonderheiten quantitativer (z.B. Verkaufspreis, Fläche) als auch kategorialer Daten (z.B. Stadt, Immobilientyp) eingegangen. Ziel ist es, die wichtigsten Prinzipien der Datenvisualisierung praxisnah zu vermitteln und die Leser:innen in die Lage zu versetzen, eigene, fundierte Visualisierungen zu erstellen und kritisch zu bewerten.

[2] Cleveland, W. S. (1984). *Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, 79(387), 531–554. https://doi.org/10.1080/01621459.1984.10478080.*

[3] Saeed, O. M. (15. Juni 2025). Kaggle. Von tate of Connecticut. (2023). Real Estate Sales 2001–2022 [Data set].: https://www.kaggle.com/datasets/omniamahmoudsaeed/real-estate-sales-2001-2022

LE 1 – Grundlagen der Visualisierung, Diagrammtypen[4][5]

Bedeutung von Visualisierung

„Ein Bild sagt mehr als tausend Worte“. Diese Aussage trifft besonders auf die Datenvisualisierung zu. Gerade bei grossen, komplexen Datensätzen wie die Immobilienverkäufen von 2001 bis 2022 wird deutlich, wie essenziell Visualisierungen sind um Muster, Trends und Ausreisser sichtbar zu machen. Während tabellarische Daten oft unübersichtlich und schwer zu interpretieren sind, ermöglichen Visualisierungen eine sofortige, intuitive Erfassung der wichtigsten Informationen. Sie helfen, verborgene Strukturen zu erkennen, Hypothesen zu entwickeln und datenbasierte Entscheidungen zu treffen.

Warum Visualisierung?

- Sie macht quantitative und qualitative Informationen verständlich.
- Sie unterstützt die Exploration grosser Datenmengen.
- Sie ermöglicht es, Zusammenhänge, Ausreisser und Trends schnell zu erfassen.
- Sie ist entscheidend für die Kommunikation von Ergebnissen an verschiedene Zielgruppen, von Analysten bis zur breiten Öffentlichkeit.

[6]Beispiel: Dinosaurier-Daten[7]

Das berühmte Beispiel der „Dinosaurier-Daten“ (Anscombe’s Quartet, Datasaurus Dozen) zeigt, dass identische statistische Kennzahlen völlig unterschiedliche Muster verbergen können, die erst durch Visualisierung sichtbar werden. Ein Datensatz kann als Mittelwert und Streuung identisch erscheinen, aber als Grafik ein Dinosaurier, eine Gerade oder eine zufällige Wolke sein.

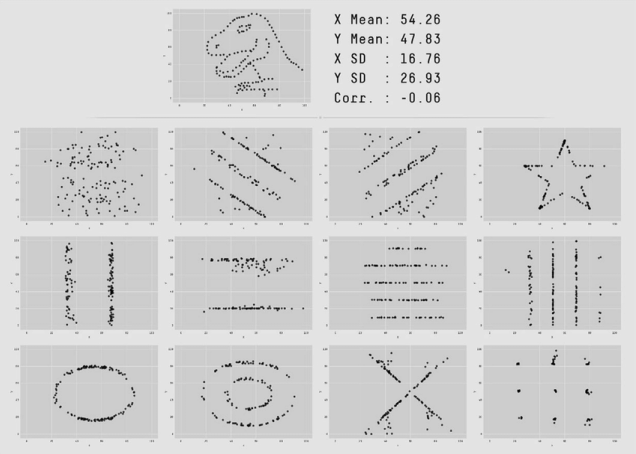

Abbildung 2: Anscombe’s Quartet, Datasaurus Dozen

[4] Wickham, H. (2010). A layered grammar of graphics. Journal of Computational and Graphical Statistics, 19(1), 3–28. https://doi.org/10.1198/jcgs.2009.07098.

[5] Wilkinson, L. ((2005)). The Grammar of Graphics (2nd ed.). Springer.

[6] https://www.researchgate.net/figure/The-Datasaurus-Dozen-Matejka-Fitzmaurice-2017-an-extension-of-Anscombes-quartet_fig2_325941519

[7] https://www.youtube.com/watch?v=DbJyPELmhJc

Geeignete Visualisierungen für Immobilienverkäufe

Die Auswahl des richtigen Diagrammtyps hängt immer von der Fragestellung und den Eigenschaften der Daten ab.

Im Kontext von Immobilienverkäufen sind insbesondere folgende Visualisierungstypen relevant:

Diagrammtyp	Zweck / Fragestellung	Beispiel für Immobilienverkäufe
Liniendiagramm	Entwicklung über die Zeit, Trends	Durchschnittlicher Verkaufspreis pro Jahr
Balkendiagramm	Vergleich von Mengen oder Mittelwerten zwischen Gruppen	Anzahl Verkäufe pro Stadtteil
Boxplot	Verteilung und Ausreisser innerhalb von Gruppen	Preisspanne in verschiedenen Stadtteilen
Heatmap	Darstellung von Dichte oder Intensität über zwei Dimensionen (z.B. Zeit und Ort)	Preisentwicklung nach Stadtteil und Jahr
Karte (Map)	Geografische Muster, regionale Unterschiede	Verteilung der Verkäufe auf einer Stadtkarte
Histogramm	Verteilung einer einzelnen numerischen Variable	Häufigkeit der Verkaufspreise
Streudiagramm	Beziehung zwischen zwei numerischen Variablen	Fläche vs. Verkaufspreis

Abbildung 3: Top 7 Visualisierungen für den Immobilien Datensatz

Liniendiagramm: Zeitreihenanalyse der Verkaufspreise

Das Liniendiagramm „Durchschnittlicher Verkaufspreis pro Jahr" zeigt die Entwicklung der Immobilienpreise von 2001 bis 2022 und ist die ideale Wahl für die Darstellung von Zeitverläufen.

Vorteile des Liniendiagramms für Immobiliendaten

Trendvisualisierung: Langfristige Preisentwicklungen werden sofort sichtbar, etwa der kontinuierliche Anstieg seit 2001

Krisenerkennnung: Einbrüche oder Stagnationen (z.B. um 2008 während der Finanzkrise) sind klar erkennbar

Kontinuitätswahrnehmung: Die Linie vermittelt den Eindruck einer kontinuierlichen Entwicklung über die Zeit

Vergleichbarkeit: Mehrere Zeitreihen (z.B. verschiedene Städte) können einfach überlagert werden

Präzision: Genaue Werte für einzelne Jahre können abgelesen werden

Nachteile und Limitationen

Interpolation: Die Linien suggerieren Werte zwischen den Messpunkten, die möglicherweise nicht existieren

Überlagerung: Bei vielen Zeitreihen kann das Diagramm unübersichtlich werden

Skalierungseffekte: Die Wahl der y-Achsen-Skalierung kann Trends übertreiben oder verschleiern

Ausreisser: Extreme Werte einzelner Jahre können die Gesamtdarstellung verzerren

Spezifische Anwendung im Immobilien-Datensatz
Das Liniendiagramm eignet sich besonders gut für die Immobiliendaten, da es Marktzyklen, Krisenauswirkungen und langfristige Preissteigerungen deutlich macht. Die kontinuierliche Zeitachse spiegelt die Realität des Immobilienmarktes wider, in dem Preise graduell über die Zeit ändern.

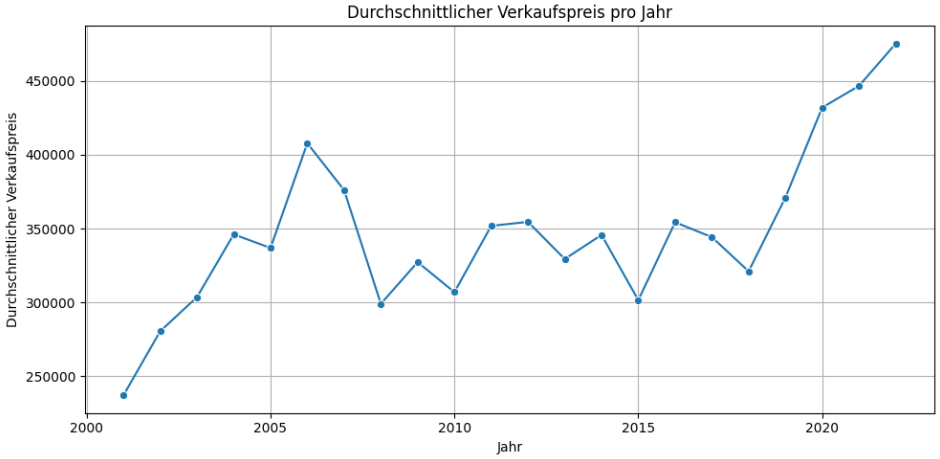

Abbildung 4: Durchschnittlicher Verkaufspreis der Immobilien zwischen 2001 und 2022

Balkendiagramm: Vergleich der Verkaufszahlen zwischen Städten

Das Balkendiagramm „Anzahl Verkäufe in den Top 10 Städten" visualisiert die Verteilung der Transaktionen zwischen verschiedenen Städten und ermöglicht direkte Vergleiche.

Vorteile des Balkendiagramms für Immobiliendaten

Direkter Vergleich:	Unterschiede zwischen den Städten sind auf einen Blick erkennbar
Ranking:	Die natürliche Sortierung zeigt sofort die aktivsten Märkte
Präzision:	Balkenlängen ermöglichen eine genaue Wahrnehmung von Mengenunterschieden
Klarheit:	Jede Stadt ist eindeutig als separate Kategorie dargestellt
Skalierbarkeit:	Weitere Städte können einfach hinzugefügt werden

Nachteile und Limitationen

Statischer Vergleich:	Keine zeitliche Entwicklung sichtbar
Kategoriale Beschränkung:	Nur diskrete Kategorien (Städte) darstellbar
Platzverbrauch:	Bei vielen Kategorien wird das Diagramm schnell unübersichtlich
Fehlender Kontext:	Grösse oder Bevölkerung der Städte wird nicht berücksichtigt
Auswahlbias:	Fokus nur auf Top-10-Städte kann kleinere, aber interessante Märkte ausblenden

Spezifische Anwendung im Immobilien-Datensatz
Das Balkendiagramm ist ideal für die Identifikation der wichtigsten Immobilienmärkte im Datensatz. Es zeigt, welche Städte die höchste Marktaktivität haben und ermöglicht eine Fokussierung der weiteren Analyse auf diese Hotspots. Die klare Hierarchie hilft bei der Priorisierung von Geschäftsentscheidungen oder Investitionsstrategien.

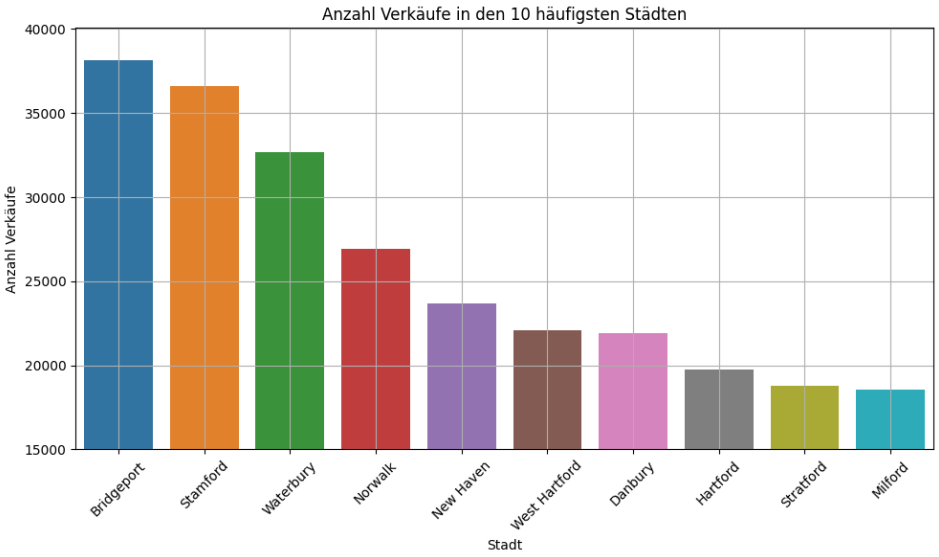

Abbildung 5: die 10 Städte in denen am meisten Immobilien verkauft werden

Kommunikative und explorative Funktion von Visualisierungen

Kommunikation
Beide Diagrammtypen erfüllen unterschiedliche kommunikative Funktionen. Das Liniendiagramm erzählt die Geschichte der Marktentwicklung über zwei Jahrzehnte, während das Balkendiagramm die geografische Verteilung der Marktaktivität vermittelt.

Exploration
In der Analysephase helfen beide Visualisierungen, verschiedene Aspekte zu erkunden: Das Liniendiagramm deckt zeitliche Muster und Anomalien auf, das Balkendiagramm identifiziert räumliche Schwerpunkte und Marktkonzentrationen.
Die Kombination beider Diagrammtypen ermöglicht eine umfassende Analyse sowohl der zeitlichen als auch der räumlichen Dimensionen des Immobilienmarktes und bildet die Grundlage für weiterführende, detailliertere Untersuchungen.

LE 2 – Visuelle Wahrnehmung

Warum visuelle Wahrnehmung zentral für Datenvisualisierung ist

Datenvisualisierung lebt davon, dass Menschen Informationen schnell und intuitiv erfassen können. Unsere Wahrnehmung ist darauf spezialisiert Muster, Unterschiede und Zusammenhänge visuell zu erkennen. Das geschieht viel schneller, als es mit Zahlenkolonnen möglich wäre. Gerade bei umfangreichen Datensätzen wie Immobilienverkäufen aus verschiedenen Städten zwischen 2001 und 2022 ist es entscheidend, Visualisierungen so zu gestalten, dass sie die Stärken der menschlichen Wahrnehmung nutzen und ihre Schwächen berücksichtigen

Wahrnehmungsprinzipien: Wie visuelle Variablen unsere Interpretation steuern[8][9]

Die menschliche Wahrnehmung verarbeitet visuelle Informationen anhand bestimmter grundlegender Kanäle (nach Bertin und Cleveland & McGill):

- **Farbe (Hue, Saturation, Value):** Farben helfen, Gruppen oder Kategorien (z.B. verschiedene Städte) zu unterscheiden. Kräftige, kontrastreiche Farben lenken die Aufmerksamkeit auf wichtige Elemente, während dezente Farben Hintergrundinformationen transportieren.
- **Grösse:** Grössere Symbole oder Balken wirken wichtiger oder zahlenmässig bedeutender. In einem Balkendiagramm über Verkaufszahlen pro Stadt wird die Wahrnehmung von Mengenunterschieden durch die Länge der Balken unterstützt.
- **Position:** Die Position eines Elements im Diagramm ist einer der stärksten Kanäle. Zeitverläufe werden meist auf der x-Achse abgebildet, Werte auf der y-Achse. So können Trends und Ausreisser schnell erkannt werden.
- **Form:** Unterschiedliche Marker-Formen (z.B. Kreise, Dreiecke) können verschiedene Ereignisse oder Kategorien kennzeichnen.
- **Helligkeit und Sättigung:** Helle oder gesättigte Farben stechen hervor, während blasse Farben in den Hintergrund treten. Dies ist besonders nützlich, um kritische Ereignisse (wie Marktkrisen) hervorzuheben.

Gestaltgesetze: Wie wir Muster erkennen

Die Gestaltpsychologie beschreibt, wie Menschen visuelle Informationen strukturieren. Diese Prinzipien helfen uns, Zusammenhänge und Gruppen auf einen Blick zu erkennen.

Die wichtigsten Gestaltgesetze für die Datenvisualisierung sind:

- **Proximität (Nähe):** Elemente, die nah beieinander liegen, werden als zusammengehörig wahrgenommen. In einem Scatterplot werden z.B. Datenpunkte, die räumlich gruppiert sind, als Cluster interpretiert.
- **Ähnlichkeit:** Elemente mit gleicher Farbe, Form oder Grösse werden als Gruppe wahrgenommen. In einem Boxplot mit unterschiedlichen Farben für verschiedene Städte erkennt man sofort, welche Box zu welcher Stadt gehört.
- **Verschlossenheit (Closure):** Das Gehirn ergänzt fehlende Informationen, um vollständige Formen zu sehen. Linien, die fast einen Kreis bilden, werden als Kreis wahrgenommen, auch wenn ein Teil fehlt.
- **Kontinuität:** Linien oder Muster, die in eine Richtung verlaufen, werden als zusammengehörig interpretiert. Ein Trend im Liniendiagramm wird als fortlaufende Entwicklung verstanden.
- **Figur-Grund-Prinzip:** Das Auge unterscheidet automatisch zwischen Vordergrund (z.B. hervorgehobene Ereignisse) und Hintergrund (z.B. allgemeiner Trend).

Kritische Reflexion: Wahrnehmungsfallen und Designentscheidungen[10]

Nicht jede visuelle Entscheidung ist automatisch effektiv. Es gibt Wahrnehmungsfallen, die zu Fehlinterpretationen führen können:

[8] Bertin, J. (1983). Semiology of Graphics: Diagrams, Networks, Maps. University of Wisconsin Press.

[9] Cleveland, W. S. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, 79(387), 531–554. https://doi.org/10.1080/01621459.1984.10478080.

[10] Bertin, J. (1983). Semiology of Graphics: Diagrams, Networks, Maps. University of Wisconsin Press.

- **Zu viele Farben:** Wenn zu viele Städte gleichzeitig in unterschiedlichen Farben dargestellt werden, kann das zu Verwirrung führen. Es empfiehlt sich, maximal 8 bis 10 Farben zu verwenden und weitere Städte gegebenenfalls zu gruppieren.
- **Falsche Skalen:** Verzerrte Achsen oder unpassende Skalierungen können Trends überbetonen oder verschleiern.
- **Cherry Picking (selektive Datenauswahl):**Ein weiteres Risiko in der Visualisierung ist das sogenannte Cherry Picking. Dabei werden gezielt nur bestimmte Zeiträume, Kategorien oder Datenpunkte ausgewählt, um eine gewünschte Aussage zu unterstreichen, während andere relevante Informationen ausgeblendet werden.
- **Überlappende Elemente:** In dicht besetzten Diagrammen (z.B. Scatterplots mit vielen Städten) sollten Transparenz oder Jitter verwendet werden, um Überlagerungen zu vermeiden.

Best Practice
Die Auswahl der visuellen Variablen und die Anwendung der Gestaltgesetze sollten immer auf die Zielgruppe und den Nutzungskontext abgestimmt werden. Für Analysten sind detaillierte, interaktive Visualisierungen sinnvoll, während für die breite Öffentlichkeit klare, reduzierte Darstellungen besser geeignet sind.
Dabei gilt: Eine Grafik sollte immer möglichst sauber und einfach interpretierbar sein. Zusätzliche Elemente wie Gridlines, Farben, Formen oder weiterführende Informationen sollten nur dann hinzugefügt werden, wenn sie einen echten Mehrwert für das Verständnis oder die Aussage der Grafik bieten. Alles, was nicht zur Klarheit beiträgt, sollte vermieden werden, um die Informationsaufnahme nicht unnötig zu erschweren.

Fazit

Die Gestaltung von Visualisierungen für Immobilienverkäufe aus verschiedenen Städten zwischen 2001 und 2022 profitiert massgeblich von der gezielten Anwendung von Wahrnehmungsprinzipien und Gestaltgesetzen. Durch den bewussten Einsatz von Farbe, Grösse, Position und Form sowie durch die Hervorhebung kritischer Ereignisse wird sichergestellt, dass die wichtigsten Informationen schnell, intuitiv und fehlerfrei erfasst werden können. Die Erkenntnisse aus der Wahrnehmungspsychologie sind damit ein zentrales Fundament für jede wirkungsvolle Datenvisualisierung.

LE 3 – Entwurfsprinzipien vs. Daten

Datenaufbereitung: Vom Rohdatensatz zur Visualisierung

Die Grundlage jeder aussagekräftigen Visualisierung ist eine sorgfältige Datenaufbereitung. Der Datensatz „Real Estate Sales 2001–2022“ umfasst Transaktionen aus verschiedenen Städten und enthält Angaben wie Verkaufsdatum, Verkaufspreis, Stadtname und weitere Merkmale der Immobilien. Rohdaten sind jedoch selten sofort nutzbar.

Für eine saubere, verständliche Visualisierung sind mehrere Schritte notwendig:

Datenbereinigung

Zunächst müssen fehlerhafte, unvollständige oder inkonsistente Einträge entfernt werden. Häufige Fehler sind Zeilen mit fehlenden Verkaufspreisen oder ungültigen Datumsangaben. Auch Tippfehler bei Städtenamen sollten vereinheitlicht werden, um Mehrfachzählungen zu vermeiden. Solche Anpassungen mussten bei diesem Datensatz nicht vorgenommen werden, da die Kaggle-Daten bereits sauber aufgearbeitet sind. Bei einigen weniger wichtigen Informationen wie „Non Use Code“, „Assessor

Remarks“ oder auch „OPM remarks“ fehlen mehr als 70% der Daten. Hier kann man prüfen, ob die Daten aufgrund zu weniger Messwerte gar nicht berücksichtigt werden.

Aggregation

Um Trends und Muster sichtbar zu machen, wurden die Daten auf verschiedene Ebenen aggregiert. Für Zeitreihen wurden die durchschnittlichen Verkaufspreise pro Jahr und Stadt berechnet. Für Vergleichsdiagramme (z.B. Boxplots) wurden die Verkaufspreise nach Städten gruppiert.

Kategorisierung und Datentypen[11]

Ein zentrales Element der Datenaufbereitung ist die Unterscheidung und korrekte Behandlung von unterschiedlichen Datentypen. Grundsätzlich unterscheidet man zwischen quantitativen (numerischen) und kategorialen (qualitativen) Daten:

- **Quantitative Daten** sind messbare, numerische Werte wie Verkaufspreis oder Wohnfläche. Sie ermöglichen mathematische Berechnungen (z.B. Durchschnitt, Summe) und werden typischerweise als kontinuierliche Variablen behandelt. Für diese Art von Daten sind Visualisierungen wie Liniendiagramme, Histogramme oder Boxplots geeignet, da sie Verteilungen, Trends und Ausreisser sichtbar machen.
- **Kategoriale Daten** sind Merkmale, die in Gruppen oder Kategorien eingeteilt werden, wie beispielsweise der Name der Stadt oder der Immobilientyp. Sie dienen der Gruppierung und dem Vergleich, nicht der Berechnung. Für kategoriale Daten bieten sich Balkendiagramme oder gruppierte Boxplots an, da sie den Vergleich zwischen den einzelnen Kategorien ermöglichen.

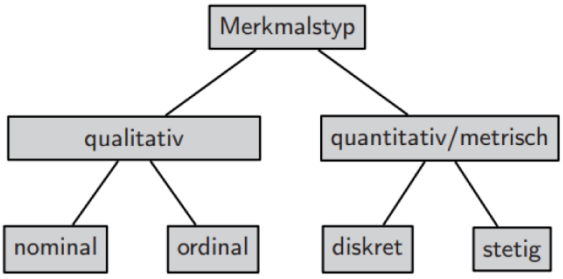

Abbildung 6: Kategorisierung von Datentypen (E. Cramer & U. Kamps, Grundlagen der Wahrscheinlichkeitsrechnung und Statistik, Springer Verlag, 2017

Im vorliegenden Immobilien-Datensatz ist zum Beispiel „Stadt“ eine kategoriale Variable, während „Verkaufspreis“ eine quantitative Variable ist. Um die Übersichtlichkeit zu wahren, wurden nur die Städte mit den meisten Transaktionen für detaillierte Vergleiche ausgewählt (z.B. die Top 10 nach Verkaufsanzahl). Seltener vertretene Städte wurden ggf. in einer Kategorie „Andere“ zusammengefasst.

Umgang mit Ausreissern

Gerade im Immobilienmarkt gibt es extreme Ausreisser (z.B. Luxusimmobilien). Diese sollen identifiziert und je nach Visualisierung entweder entfernt, geclippt oder explizit hervorgehoben. Das wird gemacht, um die Hauptverteilung nicht zu verzerren und dennoch Transparenz über die Datenstruktur zu wahren.

Transformation und Typisierung

Häufig müssen numerische Variablen wie „Verkaufspreis“ auf ein einheitliches Zahlenformat gebracht werden. Auch Datumsangaben müssen oft in Jahreszahlen umgewandelt werden, um Zeitreihen zu ermöglichen. Das mussten wir hier nicht machen, da die Daten Qualität hoch ist. Die klare

[11] Wilkinson, L. ((2005)). The Grammar of Graphics (2nd ed.). Springer

Unterscheidung zwischen kontinuierlichen Variablen (Preis, Fläche) und kategorialen Variablen (Stadt) ist essenziell für die Wahl des richtigen Diagrammtyps.

Design-Entscheidungen: Layout, Farbe, Typografie und visuelle Hierarchie

Die Gestaltung der Visualisierung folgt den Prinzipien der Klarheit, Lesbarkeit und Zielgruppenorientierung. Jede Designentscheidung wurde bewusst im Hinblick auf die Datenstruktur und die Bedürfnisse der Nutzer getroffen.

Layout

Die Anordnung der Diagramme orientiert sich an der logischen Reihenfolge der Analyse: Zunächst werden übergeordnete Trends (z.B. Zeitreihen der Preise) gezeigt, danach Vergleiche zwischen den Städten (z.B. Boxplots, Balkendiagramme). Achsen sind klar beschriftet, und die Diagrammgrössen wurden so gewählt, dass auch bei vielen Städten alle Beschriftungen lesbar bleiben.

Farbe

Farben dienen dazu, verschiedene Städte klar zu unterscheiden. Dabei wurden kontrastreiche, farbenblinde-freundliche Paletten verwendet, um die Zugänglichkeit für alle Nutzer zu gewährleisten. Kritische Ereignisse (z.B. Finanzkrise 2008, Pandemie 2020) wurden durch auffällige, warme Farben (z.B. Rot) und spezielle Linienstile (gestrichelt) hervorgehoben.

Typografie

Grosse, klare Schriftarten wurden für Achsen, Titel und Legenden gewählt, um die Lesbarkeit auch auf kleinen Bildschirmen oder in Präsentationen sicherzustellen. Wichtige Beschriftungen (z.B. Jahreszahlen bei Krisen) wurden fett und in auffälligen Farben dargestellt.

Visuelle Hierarchie

Wichtige Informationen (z.B. Medianwerte, kritische Jahre) wurden durch Grösse, Farbe oder Markierungen hervorgehoben. Weniger wichtige Elemente (z.B. Hintergrundraster) wurden dezent gehalten, um den Fokus nicht zu stören. Die Reihenfolge der Städte in Diagrammen wurde nach Relevanz (z.B. Verkaufsanzahl) sortiert.

Daten- und Design-Reflexion: Wie beeinflusst das Design die Datenwahrnehmung?[12]

Die Verbindung zwischen Datenaufbereitung und Design ist zentral für die Aussagekraft jeder Visualisierung. Designentscheidungen müssen stets die Datenstruktur, -qualität und -aussage reflektieren.

Umgang mit Ausreissern

Durch das Clipping extremer Verkaufspreise in Boxplots werden die Hauptverteilungen klarer sichtbar, ohne dass Ausreisser das Diagramm verzerren. Gleichzeitig können Ausreisser in separaten Visualisierungen dargestellt oder mit speziellen Markern hervorgehoben werden, um Transparenz zu schaffen.

[12] Wickham, H. (2010). A layered grammar of graphics. Journal of Computational and Graphical Statistics, 19(1), 3–28. https://doi.org/10.1198/jcgs.2009.07098.

Kategorisierung der Städte

Die Entscheidung, nur die häufigsten Städte einzeln zu zeigen, beruht auf dem Prinzip der Übersichtlichkeit. Eine zu grosse Zahl an Kategorien würde das Diagramm überladen und die Vergleichbarkeit erschweren. Die Gruppierung seltener Städte in „Andere“ folgt Empfehlungen zur Reduktion kognitiver Last.

Wahl des Diagrammtyps

Die Unterscheidung zwischen kontinuierlichen und kategorialen Daten beeinflusst die Visualisierung massgeblich. Zeitreihen (kontinuierlich) werden als Liniendiagramm, Städtevergleiche (kategorial) als Boxplot oder Balkendiagramm dargestellt. Diese Zuordnung folgt den Empfehlungen aus „Guide to Data Types and How to Graph Them“ und „Data: Continuous vs. Categorical“.

Transparenz und Nachvollziehbarkeit

Jede Transformation und Aggregation der Daten wird dokumentiert und – wo sinnvoll – im Diagramm oder der Legende erläutert. So wird sichergestellt, dass die Visualisierung nicht nur optisch ansprechend, sondern auch nachvollziehbar und vertrauenswürdig ist.

Fazit

Die Analyse des Datensatzes „Real Estate Sales 2001–2022“ zeigt, wie eng Datenaufbereitung und Designentscheidungen in der Datenvisualisierung miteinander verknüpft sind. Durch sorgfältige Bereinigung, Aggregation und Kategorisierung der Daten sowie durch gezielte Designentscheidungen bezüglich Layout, Farbe, Typografie und visueller Hierarchie werden komplexe Informationen verständlich und nutzbar gemacht. Die Reflexion über diese Zusammenhänge ist zentral, um Visualisierungen zu schaffen, die nicht nur schön, sondern auch korrekt, transparent und zielgruppenorientiert sind.

LE 4 – Grammatik der Grafikwerkzeuge

Das Konzept der Grammar of Graphics

Die Grammar of Graphics ist ein universelles Grundgerüst für die systematische Erstellung von Datenvisualisierungen und bildet die theoretische Grundlage vieler moderner Visualisierungstools wie ggplot2 oder auch Python-Bibliotheken wie seaborn. Das Konzept wurde von Leland Wilkinson entwickelt und von Hadley Wickham weiterentwickelt. Es beschreibt Visualisierungen als Zusammenspiel modularer Komponenten, die flexibel kombiniert werden können, um aus Daten aussagekräftige Grafiken zu erzeugen.

Das beigefügte Schaubild („Major Components of the Grammar of Graphics“) visualisiert diese Komponenten als eine Pyramide, die von den Rohdaten bis zum finalen Koordinatensystem reicht.
Jede Stufe fügt eine weitere Ebene der Abstraktion und Kontrolle hinzu.

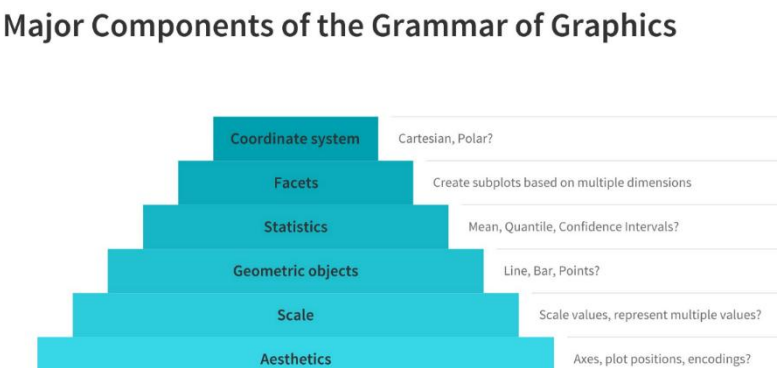

Abbildung 7: Abbildung aus dem Artikel "A Comprehensive Guide to the Grammar of Graphics for Effective Visualization of Multi-dimensional…"

Die sieben Hauptkomponenten der Grammar of Graphics[13][14]

1. **Data**
   Am Anfang jeder Visualisierung steht der zugrundeliegende Datensatz. Die Auswahl und Qualität der Daten sind entscheidend: Sie bestimmen, welche Fragen beantwortet werden können und welche Visualisierungstypen sinnvoll sind. Es ist wichtig, die Daten vorab zu bereinigen, auf Vollständigkeit und Konsistenz zu prüfen und sie gegebenenfalls zu transformieren. Beispielsweise bilden bei der Analyse von Immobilienverkäufen die Transaktionsdaten aus verschiedenen Städten und Jahren die Basis. Fehlerhafte oder lückenhafte Daten können zu irreführenden Darstellungen führen, weshalb eine sorgfältige Datenvorbereitung unerlässlich ist.

2. .**Aesthetics**
   Aesthetics beschreibt, wie Variablen aus den Daten auf visuelle Eigenschaften wie Position (x/y-Achse), Farbe, Grösse oder Form abgebildet werden. Die Wahl der Mappings sollte stets die Lesbarkeit und Interpretierbarkeit unterstützen. So empfiehlt es sich, zentrale Variablen wie Zeit auf die x-Achse und den Preis auf die y-Achse zu legen, während Kategorien wie Städte durch Farben unterschieden werden können. Zu viele gleichzeitige Mappings können die Grafik überfrachten und sollten vermieden werden. Die bewusste Auswahl der Ästhetik-Kanäle ist entscheidend für die Aussagekraft der Visualisierung.

3. **Scale**
   Skalen legen fest, wie Werte auf Achsen oder in Farbverläufen abgebildet werden – etwa linear, logarithmisch, diskret oder kontinuierlich. Die Wahl der Skala beeinflusst massgeblich, wie Trends und Unterschiede wahrgenommen werden. Falsch gewählte Skalen können Entwicklungen verzerren oder verschleiern. So sollte beispielsweise bei Preisentwicklungen eine lineare Skala verwendet werden, um echte Veränderungen sichtbar zu machen. Farbskalen müssen so gewählt werden, dass sie auch für Menschen mit Farbsehschwäche unterscheidbar sind. Eine bewusste Skalierung unterstützt die Vergleichbarkeit und Verständlichkeit der Grafik.
4. **Geometric objects (Geoms)**
   Geometrische Objekte sind die grafischen Grundelemente, mit denen Datenpunkte dargestellt werden – etwa Linien für Zeitreihen, Balken für Häufigkeiten oder Punkte für Streudiagramme. Die Wahl des passenden Geoms richtet sich nach dem Datentyp und der Fragestellung: Linien eignen sich für Entwicklungen über die Zeit, Balken für den Vergleich von Gruppen, Punkte für die Darstellung von Zusammenhängen zwischen zwei Variablen. Eine klare und konsistente Auswahl der Geoms erleichtert die Interpretation und verhindert Missverständnisse.

5. **Statistics**
   Statistische Transformationen wie Aggregationen, Mittelwerte, Quantile oder Konfidenzintervalle helfen dabei, Muster im Datensatz sichtbar zu machen, die auf Rohdatenebene nicht direkt erkennbar wären. Sie werden eingesetzt, wenn Daten zu unübersichtlich sind oder Zusammenfassungen benötigt werden. Beispielsweise kann der durchschnittliche Verkaufspreis pro Jahr und Stadt berechnet und visualisiert werden, um Trends zu erkennen. Es ist wichtig, die gewählten statistischen Methoden transparent zu machen und deren Aussagekraft kritisch zu reflektieren.

[13] Wilkinson, L. ((2005)). The Grammar of Graphics (2nd ed.). Springer

[14] Wickham, H. (2010). A layered grammar of graphics. Journal of Computational and Graphical Statistics, 19(1), 3–28. https://doi.org/10.1198/jcgs.2009.07098.

6. **Facets**
   Facetten ermöglichen die Aufteilung einer Grafik in mehrere Teilgrafiken, die verschiedene Gruppen, Kategorien oder Zeiträume separat darstellen. So lassen sich beispielsweise für jede Stadt eigene Zeitreihenplots erzeugen, um regionale Unterschiede zu analysieren. Facetten sind besonders hilfreich, wenn komplexe Zusammenhänge übersichtlich dargestellt werden sollen. Dabei sollte die Anzahl der Facetten begrenzt bleiben, um die Übersichtlichkeit zu wahren, und die Achsenskalierung sollte vergleichbar sein.

7. **Coordinate system**
   Das Koordinatensystem bestimmt, wie die Daten im Raum angeordnet werden – meist kartesisch (klassische x/y-Achsen), manchmal aber auch polar (z.B. für Kreisdiagramme) oder geografisch (für Karten). Die Wahl des Koordinatensystems sollte immer zur Fragestellung passen: Zeitreihen werden typischerweise im kartesischen System dargestellt, während geografische Daten auf Karten abgebildet werden. Ein passendes Koordinatensystem unterstützt die Orientierung und verhindert Verzerrungen in der Wahrnehmung.

Fazit

Die Grammar of Graphics ist mehr als nur ein technisches Konzept. Sie ist eine Denkweise, die das Erstellen von Visualisierungen strukturiert und nachvollziehbar macht. Durch die Kombination der sieben Komponenten lassen sich aus Daten klare, analytisch präzise und für die Zielgruppe verständliche Geschichten erzählen.
Ein bewusster Umgang mit jeder Komponente ermöglicht es, Visualisierungen gezielt zu steuern, Missverständnisse zu vermeiden und die Aussagekraft zu maximieren. Die modulare Struktur fördert die Kreativität und Flexibilität: Jede Komponente kann einzeln angepasst werden, ohne die gesamte Visualisierung neu zu entwerfen. Das erleichtert es, Visualisierungen iterativ zu verbessern und auf neue Fragestellungen anzupassen.
Zudem unterstützt die Grammar of Graphics die Kommunikation zwischen Datenanalysten, Designern und Entscheidungsträgern, indem sie eine gemeinsame Sprache für die Beschreibung und Bewertung von Visualisierungen bereitstellt. Sie schafft so die Grundlage für datengetriebene Entscheidungen und fördert eine kritische Reflexion über die Aussagekraft und Grenzen von Datenvisualisierungen.
Die Grammar of Graphics ist das Fundament für moderne, effektive Datenvisualisierungen. Sie macht Visualisierungen nachvollziehbar, flexibel und aussagekräftig

LE 5 – Evaluation

Warum Visualisierungen evaluieren?[15]

Visualisierungen sind ein zentrales Werkzeug, um komplexe Daten verständlich und Muster sichtbar zu machen und somit Entscheidungen zu unterstützen. Doch selbst eine formal korrekte Grafik kann ihr Ziel verfehlen, wenn sie von der Zielgruppe nicht richtig verstanden wird oder zu Fehlinterpretationen führt. Die Evaluation von Visualisierungen ist daher essenziell, um sicherzustellen, dass die intendierte Botschaft tatsächlich ankommt und die Nutzer:innen effizient und fehlerfrei mit den Informationen arbeiten können. Evaluation hilft, Schwächen im Design zu erkennen, Missverständnisse zu minimieren und die Zugänglichkeit für verschiedene Nutzergruppen (z.B. Expert:innen vs. Laien) zu gewährleisten.

Qualitative vs. quantitative Methoden[16]

Für die Evaluation von Visualisierungen gibt es sowohl qualitative als auch quantitative Ansätze. Qualitative Methoden wie Interviews, Fokusgruppen oder Think-Aloud-Protokolle liefern tiefgehende Einblicke in die Wahrnehmung, die Verständlichkeit und die Interpretationsstrategien der Nutzer:innen. Sie eignen sich besonders, um subjektive Eindrücke, Schwierigkeiten oder Verbesserungsvorschläge zu erfassen. Quantitative Methoden hingegen messen objektive Kriterien wie Bearbeitungszeit, Fehlerquoten oder Präferenzwerte in standardisierten Tests. Sie ermöglichen es, verschiedene Visualisierungsdesigns systematisch zu vergleichen und statistisch auszuwerten, etwa durch A/B-Tests oder kontrollierte Experimente.

Nutzerstudie: Planung und Durchführung

Eine Nutzerstudie zur Evaluation einer Visualisierung sollte folgende Aspekte berücksichtigen:

Zielgruppe
Die Zielgruppe der Nutzerstudie sind Leserinnen und Leser der Tageszeitung Hartford Courant. Diese Gruppe umfasst sowohl Personen ohne Vorkenntnisse in Datenvisualisierung als auch solche mit grundlegender Erfahrung im Umgang mit Diagrammen.

Anzahl der Teilnehmenden
Für eine aussagekräftige Nutzerstudie werden in der Literatur in der Regel mindestens 10 bis 12 Personen pro Gruppe empfohlen, um belastbare und generalisierbare Ergebnisse zu erhalten2. In dieser Mini-Evaluation wurden je vier Personen pro Gruppe befragt – einmal Laien und einmal Expertinnen und Experten aus dem Data Science Studium. Diese geringe Anzahl reicht für eine fundierte Aussage nicht aus, ist jedoch für den Rahmen dieser Arbeit und eine erste Einschätzung der Visualisierung angemessen.

Mini-Evaluation: Liniendiagramm „Durchschnittlicher Verkaufspreis pro Jahr“[17]

Für das oben gezeigte Liniendiagramm wurde eine Mini-Evaluation mit zwei Nutzergruppen durchgeführt: Laien ohne Vorkenntnisse in Datenvisualisierung und Expert:innen (Kommilitonen im Data Science Studium) mit Erfahrung im Umgang mit Diagrammen.

[15] Cleveland, W. S. (1984). *Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, 79(387), 531–554. https://doi.org/10.1080/01621459.1984.10478080.*

[16] Rohrer, C. (2008). User Research Methods: From Qualitative to Quantitative. Nielsen Norman Group.

[17] Rohrer, C. (2008). User Research Methods: From Qualitative to Quantitative. Nielsen Norman Group.

Aufgaben

1. Das Jahr mit dem höchsten durchschnittlichen Verkaufspreis identifizieren.
2. Einen Zeitraum mit deutlichem Preisrückgang benennen.
3. Den allgemeinen Trend der Preisentwicklung beschreiben.

Unabhängige und abhängige Variablen

In der Nutzerstudie wurden verschiedene unabhängige und abhängige Variablen betrachtet, um die Verständlichkeit und Effektivität der Visualisierung zu evaluieren. Zu den unabhängigen Variablen zählen der verwendete Visualisierungstyp, in diesem Fall ein Liniendiagramm, das gewählte Farbschema sowie die Markierung einzelner Datenpunkte, etwa durch Kreise, um die Zuordnung zu bestimmten Jahren zu erleichtern. Diese Gestaltungsmerkmale beeinflussen, wie die Teilnehmenden die dargestellten Informationen wahrnehmen und interpretieren.

Als abhängige Variablen wurden mehrere Aspekte gemessen: Zum einen die Bearbeitungszeit, die die Teilnehmenden für jede Aufgabe benötigten, um beispielsweise das Jahr mit dem höchsten durchschnittlichen Verkaufspreis zu identifizieren. Zum anderen wurde die Fehlerquote bei der Lösung der Aufgaben erfasst, also wie häufig falsche Antworten gegeben wurden. Darüber hinaus spielte die Zufriedenheit der Nutzerinnen und Nutzer eine Rolle, ebenso wie die subjektive Verständlichkeit der Visualisierung, die durch gezielte Nachfragen und das erhaltene Feedback erhoben wurde. Diese abhängigen Variablen geben Aufschluss darüber, wie gut die Visualisierung von der Zielgruppe verstanden und genutzt werden kann.

Studiendurchführung

Die Studie wurde vor Ort als Interview durchgeführt, um die Verständlichkeit und Nutzerfreundlichkeit der Visualisierung im direkten Dialog mit den Teilnehmenden zu evaluieren. Dabei wurde darauf geachtet, dass die erhobenen Daten anonym und ausschliesslich für den Zweck der Evaluation erfasst wurden. Dies gewährleistet den Schutz der Persönlichkeitsrechte und entspricht den Anforderungen an den Datenschutz.
Alle Teilnehmenden wurden vor Beginn ausführlich über den Zweck sowie den Umfang der Studie informiert. Sie gaben ihr mündliches Einverständnis zur Teilnahme, was einen wichtigen ethischen Aspekt im Forschungsprozess darstellt und die informierte Einwilligung sicherstellt. Die Durchführung der Interviews ermöglichte es, gezielt Rückfragen zu stellen und auf individuelle Verständnisschwierigkeiten einzugehen.
Die Auswertung der Studie erfolgte qualitativ, indem die Antworten und das geäusserte Feedback der Teilnehmenden systematisch analysiert wurden. Auf diese Weise konnten sowohl positive Aspekte als auch Verbesserungspotenziale der Visualisierung identifiziert und anschliessend in die Weiterentwicklung des Designs integriert werden

Ergebnisse

- Beide Gruppen konnten die Hauptaufgaben grundsätzlich lösen. Die Markierung der Datenpunkte mit Kreisen (marker="o") wurde von den meisten als hilfreich empfunden, um einzelne Jahre schnell zuzuordnen.
- Einige Laien hatten zunächst Schwierigkeiten, die y-Achse korrekt zu interpretieren, insbesondere bei der Einschätzung der absoluten Preisunterschiede zwischen den Jahren. Erst nach gezieltem Nachfragen wurde die Skala bewusst wahrgenommen.
- Die Expert:innen wiesen darauf hin, dass die Grafik insgesamt klar strukturiert ist, empfahlen jedoch, auffällige Ereignisse (wie starke Preissprünge) zusätzlich zu annotieren, um die Interpretation weiter zu erleichtern.

- Beide Gruppen äusserten, dass die Reduktion auf eine Linie ohne zusätzliche Farben oder Formen die Lesbarkeit verbessert hat. Die Lesbarkeit der Achsenbeschriftungen wurde als gut bewertet.

Fazit

Die Mini-Evaluation zeigt, dass das Liniendiagramm für die Zielgruppe grundsätzlich verständlich ist und die wichtigsten Informationen effizient vermittelt. Dennoch kann die Verständlichkeit durch gezielte Annotationen und eine noch klarere Hervorhebung von Ausreissern weiter verbessert werden. Die Evaluation bestätigt die Relevanz von Nutzerfeedback im Designprozess und unterstreicht, dass auch scheinbar einfache Visualisierungen von einer strukturierten Überprüfung profitieren.

Die Evaluation von Visualisierungen ist ein unverzichtbarer Schritt, um sicherzustellen, dass Daten nicht nur korrekt, sondern auch verständlich und zielgruppengerecht vermittelt werden. Die Kombination aus qualitativen und quantitativen Methoden sowie die Integration von Nutzerfeedback führen zu besseren, effektiveren Visualisierungen.

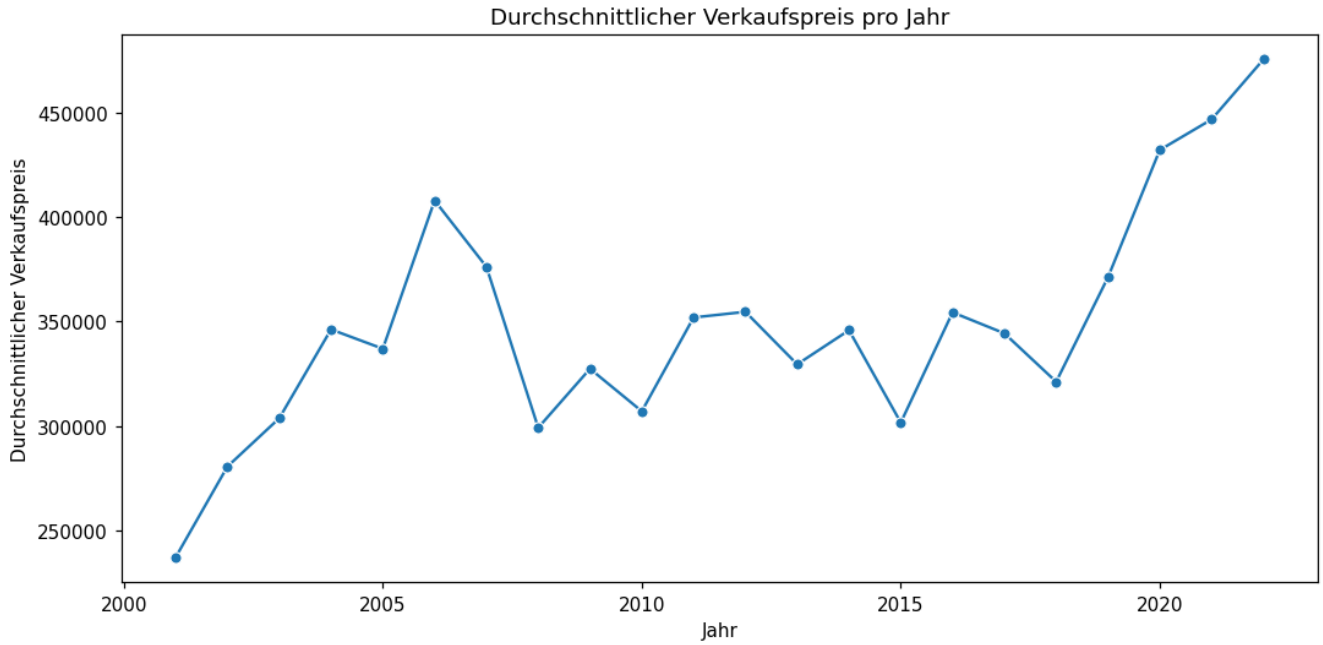

Abbildung 8: Zu evaluierende Grafik

Abbildungsverzeichnis

Literaturverzeichnis

Bertin, J. (1983). *Semiology of Graphics: Diagrams, Networks, Maps. University of Wisconsin Press.*

Cleveland, W. S. (1984). *Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, 79(387), 531–554. https://doi.org/10.1080/01621459.1984.10478080.*

Rohrer, C. (2008). *User Research Methods: From Qualitative to Quantitative. Nielsen Norman Group.*

Saeed, O. M. (15. Juni 2025). *Kaggle*. Von tate of Connecticut. (2023). Real Estate Sales 2001–2022 [Data set].: https://www.kaggle.com/datasets/omniamahmoudsaeed/real-estate-sales-2001-2022 abgerufen

Wickham, H. (2010). *A layered grammar of graphics. Journal of Computational and Graphical Statistics, 19(1), 3–28. https://doi.org/10.1198/jcgs.2009.07098.*

Wilkinson, L. ((2005)). *The Grammar of Graphics (2nd ed.). Springer.*