Datenvisualisierung mit ggplot2

Sebastian Kuhn

4 9 2021

# Datenvisualisierung mit ggplot2 - Videocasts

## Inhalt

* [Datenvisualisierung mit ggplot2 - Teil 1: Grundlagen](#basics)
* [Die “Grammar of Graphics”](#grammar)
* [Ein erster ggplot-Befehl](#Befehl)
* [Literatur, Blogs und schnelle Hilfe](#Hilfe)
* [Die Datenbasis](#data)

## Datenvisualisierung mit ggplot2 - Teil 1: Grundlagen

Herzlich Willkommen zum ersten Teil einer kleinen Video-Tutorial-Reihe zum Thema “Datenvisualisierung mit ggplot2”. In diesem ersten Teil der Reihe werden wir uns die grundlegende Funktionsweise des Pakets ggplot anschauen und dafür erstmal auf einer etwas theoretischeren Ebene stehen bleiben - aber keine Angst: die Praxisbeispiele kommen noch früh genug in den nächsten Videos! Vorher glaube ich allerdings, dass es erstmal ganz sinnvoll ist, sich einen Überblick über die Grundprinzipien und die Ideen zu verschaffen, die mit ggplot verbunden sind - und genau das soll das Thema dieses Videos sein.

Wenn wir damit fertig sind, werde ich Ihnen in den kommenden Videos anhand einiger Beispiele zeigen, wie man Diagramme in ggplot Schritt für Schritt aufbaut. Wir starten dabei mit einem Streudiagramm und werden uns später weitere Darstellungsformen und auch komplexere Grafiken anschauen. Am Ende werden Sie - zumindest hoffe ich das - die Grundprinzipien einer effektiven Datenvisualisierung verstehen, und auch wissen, wie man mit ggplot eine ziemlich breite Palette an Diagrammen erstellen kann.

Vorab noch ein kurzer Disclaimer: Ich gehe in dieser Einführung davon aus, dass Sie die benötigte Software - also vor allem R und R-Studio - schon installiert haben und dass Sie auch schon Grundkenntnisse in R haben. Zudem ist es mir wichtig, hier zu sagen, dass ich auch mit einer Videoreihe keinen umfassenden Überblick über alle Funktionen von ggplot geben kann - vielmehr müssen wir hier zwangsläufig an der Oberfläche bleiben und auch die Beispiele, die ich Ihnen zeigen werde sind keinesfalls als “Kochrezepte” oder als strikte Regelwerke zu verstehen. Wie fast immer bei der Arbeit mit R gibt es auch bei ggplot häufig mehr als einen Weg zum Ziel - und gerade bei der Visualisierung von Daten stehen Ihnen zahllose Möglichkeiten zur Verfügung die zum Teil auch mit Geschmack zu tun haben. Das bedeutet allerdings auch, dass ggplot2 natürlich sehr viel mehr kann als Ihnen im Rahmen dieser Reihe hier zeigen kann.

Das Ziel dieser Tutorialreihe besteht im Grunde darin, dass Sie einen schnellen Eindruck von der Funktionalität von ggplot bekommen und schnell selbst damit loslegen können, Plots damit zu erstellen. Sie sollten am Ende der Reihe dazu in der Lage sein, einen in ggplot2 geschriebenen Code im Prinzip zu verstehen, wenn Sie ihn sehen und Sie sollten einen Blick dafür bekommen, wie die “Grammatik” von ggplot funktioniert, wie die verschiedenen Ebenen, Formen und Daten zusammengesetzt werden und wie Sie Abbildungen in ggplot selbst Schritt für Schritt aufbauen können.

Noch ein Hinweis an der Stelle: Haben Sie keinen zu großen Respekt vor kompliziert aussehenden Codes - seien Sie sich versichert, dass die wenigsten Menschen, die praktisch täglich mit ggplot arbeiten sämtliche Teile des Codes auswendig und fehlerfrei herunterschreiben können. Natürlich wird bei der Erstellung von schicken Grafiken mit ggplot andauernd nebenher gegoogelt und auf Cheat-Sheets gespickelt - und das ist auch okay so. Ganz im Gegenteil: Im Netz findet sich eine sehr aktive Community, die zahlreiche Tipps und Tricks zu ggplot teilt und auch häufig die Codes und Daten für unglaublich tolle Darstellungen zur Verfügung stellt. Ich werde Ihnen da am Ende des Tutorials noch zwei, drei Hinweise geben. Nutzen Sie diese Möglichkeiten und seien Sie kreativ - das ist letztlich auch Sinn und Zweck der Vernstaltung von ggplot2.

### Die “Grammar of Graphics”

So, jetzt kommen wir aber endlich zum Punkt und schauen uns an, was es mit ggplot2 auf sich hat. ggplot gehört zu der tidyverse-Familie, die eine Reihe sehr nützlicher Pakete aus dem Umfeld von R-Studio-Entwicklern beinhaltet und das wichtige ist - diese sind und werden auch gut gepflegt. Ursprünglich wurde ggplot von Hadley Wickham im Jahr 2005 entwickelt. Wickham hat sich dabei von einem Konzept namens “Grammar of Graphics” inspirieren lassen, das Ende der 1990er Jahre von Leland Wilkinson entwickelt wurde. Das zeigt sich im Übrigen auch an dem Namen des Pakets - die beiden “g” in ggplot eben für “grammar” und “grpahics” stehen.

Ich kann und will an der Stelle nicht zu sehr ins Detail gehen, aber ich glaube es macht durchaus Sinn, sich den Grundgedanken der “Grammar of Graphics” einmal anzuschauen. Die Idee von Wilkinson besteht darin, eine ‘Grammatik’ für Grafiken zu entwickeln - das sagt ja auch schon der Name. Vielleicht fragen Sie sich jetzt, was hat denn “Grammatik” mit Grafik zu tun?. Wenn Sie sich die Sprache vorstellen, dann spielt da die Grammatik eine ganz wichtige Rolle bei der systematischen Beschreibung von Sprache und Grammatik stellt im Prinzip ein Regelwerk zur Verwendung der Sprache zur Verfügung. Beherrscht man dieses Regelwerk, dann ist auf dieser Basis eine unzählige Anzahl an (mehr oder weniger) sinnvollen sprachlichen Aussagen möglich. Analog dazu soll die “Grammar of Graphics” eine Art Sammlung von Konzepten und Regeln darstellen, aus denen wir im Prinzip alle denkbaren und sinnvollen Grafiken und Datenvisualisierungen erstellen können. In seinem - übrigens sehr lesenswerten - Aufsatz zu den Grundlagen von ggplot schreibt Wickham selbst, dass diese Grammatik die Grundlage zur Erstellung von graphischen “Gedichten” darstellt. Da wird schon deutlich, dass neben diesem systematischen Aspekt der Grammatik auch ein künstlerischer Aspekt eine Rolle spielt, aber bevor wir uns diesem widmen können müssen wir uns erstmal die Bestandteile der Grammatik genauer anschauen um dazu in der Lage zu sein, systematisch Grafiken mit ggplot2 erstellen zu können.

Schauen wir uns nun die Bestandteile der Grammatik einmal näher an. In der Pyramide links sehen sie die wichtigsten Elemente der “Grammar of Graphics” und das Besondere daran ist, dass Sie mit Hilfe dieser Komponenten so ziemlich jede erdenkliche statistische Grafik beschreiben - und mit ggplot eben auch erstellen - können.

#### Data

Ganz unten finden Sie da die Daten - und klar, ohne Daten gäbe es nichts, was Sie irgendwie visualisieren könnten. Daten helfen Ihnen Daten aber auch nicht viel - vielmehr müssen Sie irgendwie festlegen, welche Variablen aus Ihren Daten in welcher Form abgebildet werden sollen. Und hiermit ist dann auch das “Aesthetics”-Element der “Grammar of Graphics” angesprochen.

#### Aesthetics

Mit ästhetischen-Funktionen verlinken Sie letztlich in ggplot2 Ihre Variable mit bestimmten ästhetischen Darstellungen in der Grafik - beispielsweise legen Sie hier also fest, welche Variablen auf der x- oder y-Achse dargestellt werden sollen oder ob Ihre Variablen mit bestimmten Farben in der Darstellung verbunden werden. Weil es hier um die logische Verbindungen zwischen den Daten und den sichtbaren Elementen in Ihrer Darstellung geht, wird an der Stelle häufig vom sogenannten “aesthetic mapping” gesprochen.

#### Scale

Sobald dann klar ist, welche Daten für eine Visualisierung verwendet werden sollen, muss auch die Skalierung festgelegt werden. Wenn wir hier von Skalen oder Skalierung sprechen meinen wir beispielsweise die Wertebereiche der x- und y-Achsen. Es macht zum Beispiel keinen Sinn, eine Zeitachse von 1800 bis 2020 darzustellen, wenn Sie nur Daten aus den Jahren 2010 bis 2020 verwenden. Das ist klar. Mit der Skalierung sind aber auch Farb- oder Größenskalen und zahlreiche andere Eigenschaften und Parameter von ggplot-Objekten gemeint. Das wird vermutlich klarer, wenn wir uns die Beispiele in den nächsten Videos anschauen.

#### Geometric Objects

Sobald feststeht, welche Daten in einem Plot dargestellt werden sollen muss wir natürlich auch festlegen, in welcher Form die Daten visualisiert werden - ob Sie zum Beispiel Punktwolke visualisiert haben möchten, Boxplots oder Balkendiagramme - und wenn wir von derlei Visualisierungen sprechen (im ggplot2-Kontext), dann ist hier die Rede von “geometrischen Objekten”. Für praktisch alle gängigen Formen von Datenvisualisierungen existieren in ggplot mittlerweile eigene geometrische Funktionen, zum Beispiel erstellt die Funktion geom\_point() Streudiagramme, geom\_bar() erstellt Balkenplots und geom\_boxplot() erstellt Boxplots.

#### Statistics

Als nächstes Element der “Grammar of Graphics” kommen wir zu statistischen Transformationen oder zu den “Statistics”. Was ist darunter zu verstehen? Nicht immer wollen Sie den exakten Wert Ihrer Variablen in einer Abbildung darstellen, vielmehr sind Sie vielleicht an Mittelwerten, Summen oder Auszählungen interessiert oder wollen z.B. den Logarithmus von Werten verwenden. Daher wird in der “Grammar of Graphics” und in ggplot2 die Möglichkeit einer statistischen Transformation vor der eigentlichen Darstellung offen gelassen. Viele Darstellungen wie z.B. Boxplots oder Histogramme setzen solche statistischen Transformationen ja im Prinzip von vorneherein voraus - bei anderen sind sie dagegen eher optional. Wenn Sie keine statistischen Transformation vornehmen, dann ist im Kontext von ggplot2 immer mal wieder die Rede von der “transformation identity”. Dann werden die Daten nicht verändert, sondern direkt aus dem Datensatz an die jeweilige Ästhetik weitergegeben.

#### Facets

Das nächste Element in der Pyramide sind die “Facets” - also auf Deutsch so etwas wie Facettenspezifikationen. Konkret ist hier gemeint, dass die ausgewählten Daten vor der eigentlichen Visualisierung noch in verschiedene Teil-Datensätze aufgesplittet werden können. Sie könnten z.B. Ihre Daten nach Ländern oder Jahren separieren und auch in separaten Darstellungen plotten. In diesem Fall verwenden wir dann eine Facettenspezifikation, die zum Beispiel für jedes Land erst einen eigenen Teildatensatz erstellt und erst dann die eigentliche Darstellung.

#### Coordinate System

Zuletzt ist für die Erstellung einer Grafik immer die Festlegung eines bestimmten Koordinatensystems notwendig, das auch zu den Daten und den geometrischen Objekten passt. Am häufigsten wird hier das sogenannte “kartesische” Koordinatensystem verwendet - für viele Darstellungen - wie zum Beispiel für Streudiagramme - passt das meistens, für andere Darstellungen - wie für Kuchendiagramme - bieten sich aber andere Systeme wie zum Beispiel das polare Koordinatensystem an.

#### Bahnhof

So, vielleicht verstehen Sie jetzt nur noch Bahnhof fragen sich, warum das alles so kompliziert sein muss - und ich gebe zu, dass das alles erstmal ziemlich viel und ziemloich trocken erscheint, ich glaube aber, dass Sie sehr davon profiteren werden, wenn sie zumindest ein Gefühl für die Grundidee von ggplot haben, bevor Sie selbst richtig loslegen.

Und zur Beruhigung: ggplot wird Ihnen viel Arbeit abnehmen, da die Entwickler sich hier viele Gedanken über intelligente Voreinstellungen und default-Werte gemacht haben und Sie letztlich einfache Grafiken schon durch die Angabe der Daten, der “Aesthetics” und der geometrischen Objekte erstellen können ohne dass Sie all die anderen Elemente, die wir gerade angesprochen haben, händisch festlegen oder verändern müssten. Wenn Sie keine weiteren Angaben - also außer der Daten, der “Aesthetics” und der geometrischen Objekte - mehr machen, dann verwendet ggplot grundsätzlich eine Reihe von sehr sorgfältig ausgewählten default-Werte, die in aller Regel zu sehr vernünftigen und vorzeigbaren Ergebnissen führen. Wenn Sie natürlich eine Darstellung für eine Publikation erstellen wollen oder tiefere Einsichten in Ihre Daten haben wollen, dass werden Sie in der Regel noch einiges anpassen müssen - für einen schnellen Einblick in die Daten sind die Voreinstellungen aber - in den meisten Fällen - völlig ausreichend.

### Ein erster ggplot-Befehl

So, jetzt wollen wir uns aber mal anschauen, wie diese Grundsätzte der “Grammar of Graphics” in ggplot umgesetzt sind. Dafür schauen wir uns jetzt einfach mal einen einfachen Befehl in ggplot schematisch an. Hier auf der Folie finden Sie diesen Befehl.

Was hier passiert ist im Grunde das Folgende: Mit der Funktion ggplot() teilen Sie R im Prinzip einfach mit, dass Sie jetzt gerne ggplot verwenden würden. Diese Funktion erwartet als Argument in aller Regel immer die Angabe des Datensatzes (und zwar des Datensatzes in dem sich die Variablen befinden, die eben dargestellt werden sollen). Sie übergeben diesen Datensatz im Argument data =, wobei diese Angabe im Prinzip optional ist. Die Funktion ggplot() erwartet grundsätzlich als erstes einen Datensatz, daher können Sie die Angabe von “data =” im Prinzip auch einfach weglassen. In dem vorliegenden Fall übergeben Sie der Funktion ggplot() also einen fiktiven Datensatz namens “df” und legen als nächstes in der Aesthetics-Funktion (also a-e-s) fest, welche Variablen Sie nun konkret verwenden wollen und stellen das sogenannte “Mapping” zwischen den Daten und Ihrer Abbildung her. In diesem Fall wollen Sie also die fiktive Variable x auf der x-Achse darstellen und die fiktive Variable y auf der y-Achse. Damit haben Sie die Daten und die Ästhetik des Plots festgelegt. Nun muss ggplot nur noch wissen, welche Darstellungsform verwendet werden soll. Daher fügen Sie nun mit einem + eine der zahlreichen geom\_-Funktionen hinzu - in diesem Fall sollen Punkte geplottet werden - es soll also ein Streudiagramm erstellt werden.

Manchmal macht es Sinn, sich einen ggplot-Befehl einmal ausgesprochen vorzustellen - und das würde in der Variante wie folgt klingen: “Hey ggplot, nimm doch bitte den Datensatz gf, stelle mir die Werte der Variable x auf der x-Achse dar und die Werte der Variable y auf der y-Achse. Und zeichne dann bitte für jede Wertkombination einen Punkt.”

Und das war’s dann auch schon - alle anderen Entscheidungen - beispielsweise über die Art des Koordniatensystems oder die Skalierung - würde ggplot in diesem Fall selbst treffen (da Sie es nicht getan haben) - und in aller Regel würde das dann auch zu einem vorziegbaren Ergebnis führen.

### Literatur, Blogs und schnelle Hilfe

So, und weil das alles in der Theorie nicht so richtig viel Spaß macht, freue ich mich schon auf die kommenden Videos in denen ich Ihnen zeigen werde, wie ggplot in der Praxis funktioniert. Vorher möchte ich Ihnen zum Ende des Videos aber noch ein paar Tipps zur Literatur und Hilfestellungen rund um ggpplot an die Hand geben. Ich habe ja schon am Anfang des Videos gesagt, dass Sie keinen allzu großen Respekt vor kompliziert aussehenden Codes haben sollten, sondern sich mit Offenheit und mit Kreativität auf ggplot einlassen sollten - und aus meiner Erfahrung heraus ist es hierzu sehr hilfreich, sich einfach mal im Netz anzuschauen, was andere Menschen so alles mit ggplot machen und dann auch gegebenenfalls besonders gelungene oder schöne Grafiken einfach einmal selbst zu replizieren. Glücklicherweise gibt es dazu mittlerweile im Netz einen Haufen toller Blogs und ich habe Ihnen hier nur einmal eine ganz kleine Auswahl mitgebracht.

Sehr ans Herz legen will ich Ihnen beispielsweise die [R-graph-gallery](https://www.r-graph-gallery.com/), die eine Vielzahl an möglichen ggplots beinhaltet und auch gleich den Code dazu bereitstellt - und auch beispielsweise den Blog von [Cédric Scherer](https://www.cedricscherer.com/), der wirklich sehr schöne und informative Grafiken erstellt - wie zum Beispiel auch die, die Sie hier auf der Folie sehen können - und der auch zahlreiche Tutorials zu ggplot2 auf seinen Seiten bereit hält - da lohnt sich ein Blick auf jeden Fall.

Dann finden Sie hier auf der Folie auch zwei Literaturempfehlungen - die beiden Bücher von Kiran Healy und von Claus Wilke halte ich wirklich für uneingeschränkt empfehlenswert und auch hier ist das schöne, dass beide Autoren ihren Code auf ihren Homepages bzw. auf den github-Seiten auch frei zur Verfügung stellen. Das heißt, Sie können im Grunde alle Abbildungen, die Sie in den Büchern finden auch mit mehr oder weniger großem Aufwand selbst auch nachbauen.

Zuletzt will ich Sie noch auf die umfangreiche Dokumentation des Pakets auf den [tidyverse](https://ggplot2.tidyverse.org/)-Seiten hinweisen - hier finden Sie in der Regel schnell Informationen zu allen Funktionen von ggplot2 - und übrigens auch einen sehr hilfreichen Spickzettel mit den wichtigsten Befehlen und Funktionen, den Sie einfach ausdrucken und auf Ihren Schreibtisch legen können. Wenn auch das nicht mehr hilft, dann lege ich Ihnen zum Schluss noch die Plattform [stackoverflow](https://stackoverflow.com/) ans Herz - hier finden Sie insbesondere dann Rat wenn Sie ganz besonders knifflige Fragen oder Probleme haben - da lohnt sich dann ein Blick wenn Sie nicht mehr weiter kommen.

Und nun wünsche ich Ihnen schonmal viel Spaß beim stöbern und freue mich auf die kommenden Videos!