Tom Alby

WEBANALYSE: Wie aus Daten Taten folgen, Teil II

Im ersten Teil wurden die ersten beiden Schritte der Datenanalyse beleuchtet, die Definition des Geschäftsproblems und die Akquise der Daten. In diesem Teil geht es um die Analyse selbst.

Sind die Ziele bestimmt und die Daten erfasst, die für das Messen der Zielerreichung notwendig sind, so ist in einer idealen Welt gar keine Analyse notwendig. Ein Dashboard oder ein Report zeigt an, dass der eingeschlagene Weg zur Zielerreichung der richtige ist und dass das Ziel mit hoher Wahrscheinlichkeit in der avisierten Zeitspanne erreicht wird.[1] Diese ideale Welt ist wahrscheinlich eher die Ausnahme als die Regel[2], doch schon dieser Teil der Analyse birgt Fehlerpotenzial.

Deskriptive Analytik: Daten richtig beschreiben

Nachdem es im ersten Teil darum ging, ob die richtigen Daten gesammelt werden, dreht sich jetzt alles zunächst darum, dass diese erfassten Daten richtig beschrieben werden. Eine Diagnose, warum etwas von den Erwartungen abweicht, findet hier noch nicht statt. Das klingt zunächst einmal trivial und langweilig, aber überraschenderweise bekommt bei der korrekten Beschreibung von Daten keines der gängigen Analytics-Produkte eine gute Note. Ganz im Gegenteil.

Das verständliche Beschreiben von Daten ist Aufgabe der deskriptiven Statistik. In einem Report oder einem Dashboard werden Daten zusammenfassend beschrieben, sodass ein schneller Überblick und somit eine einfache „Einschätzung der Lage“ möglich sind. Genau das bieten Adobe Analytics, Google Analytics und Co: Eine einfache und mehr oder weniger attraktiv gestaltete grafische Benutzeroberfläche mit Beschreibungen der Daten. Damit ein solches Tool für möglichst viele Nutzer funktioniert, werden Standard-Darstellungen genutzt, die für die Mehrheit der Nutzer verständlich sind. Ein häufiges Beispiel ist die Verwendung des arithmetischen Mittels, auch einfach „Durchschnitt“ genannt. Der Durchschnitt ist deshalb attraktiv, weil er Informationen in einer Zahl verdichtet und der Allgemeinheit geläufig ist. Allerdings, und das ist den meisten Anwendern nicht bewusst, gibt ein Durchschnitt allein nur dann einigermaßen verlässliche Informationen über die Daten, wenn eine Normalverteilung vorliegt. Der Durchschnitt gibt dann nämlich auch den häufigsten Wert (Modus) und den Wert genau in der Mitte (Median) wieder. Liegt keine Normalverteilung vor, so kann es sein, dass der häufigste Wert woanders liegt und der Durchschnitt verzerrt und somit nicht mehr aussagekräftig ist. Oder, wie ein alter Statistiker-Witz lautet, man kann auch in einem See mit einer durchschnittlichen Wassertiefe von 20 Zentimetern ertrinken.

Ein Beispiel für eine Normalverteilung ist in Abbildung 1 zu sehen. Hier wurde ein sogenanntes Histogramm als Darstellung gewählt,

[1] Die smarte Erstellung von Reports und Dashboards wird in einem weiteren Teil behandelt.
[2] Wahrscheinlich, denn sicher kann man nicht sein, weil man nur von den Fällen erfährt, wo es nicht funktioniert ☺.

DER AUTOR

Tom Alby ist mehrfacher Buchautor und von der „brand eins“ als „Datenfreak” bezeichneter Digitalexperte. Nach Stationen wie Google und Ask.com ist er seit 2018 CDTO bei Euler Hermes.

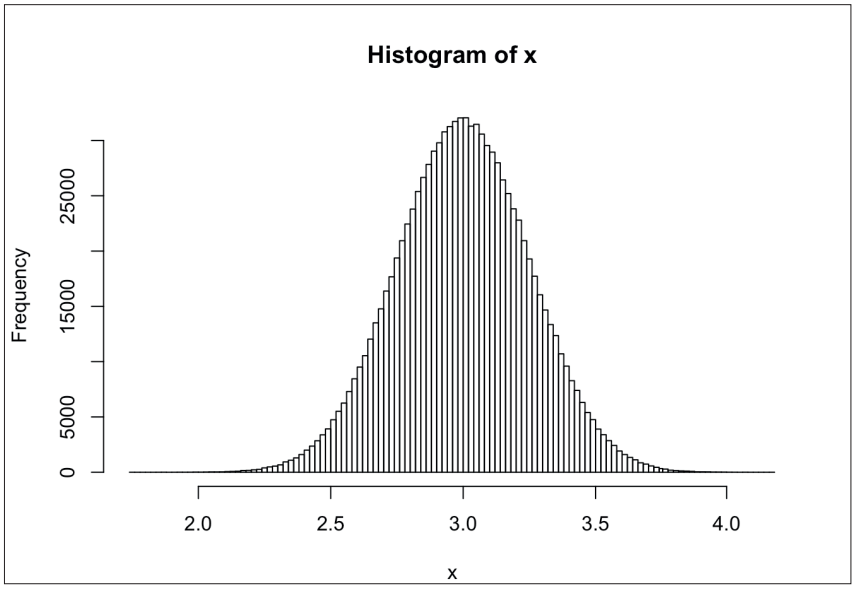

Abb. 1: Histogramm von Daten in einer Normalverteilung

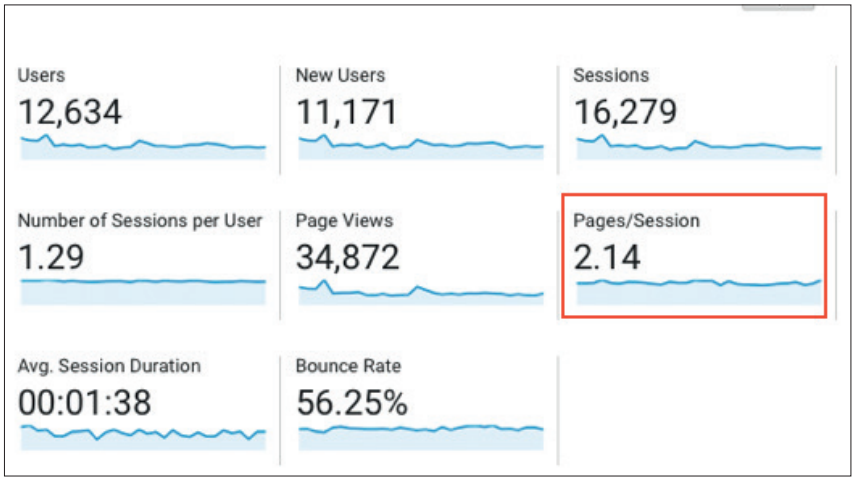

Abb. 2: Die durchschnittliche Anzahl der Seiten pro Session liegt hier über 2

da diese Art der Datenvisualisierung ideal ist für eine Häufigkeitsverteilung. Jeder Balken zeigt, wie häufig jeder Wert einer Variablen vorkommt (siehe Info-Kasten).

Dieses Problem einer fehlenden Normalverteilung und der Auswirkung auf den Durchschnitt soll an einem Beispiel aus der Webanalyse verdeutlicht werden. Die Anzahl der besuchten Seiten pro Sitzung wird häufig als Indikator dafür gesehen, wie interessant Nutzer die besuchten Seiten fanden. Abbildung 2 zeigt eine Statistik aus Google Analytics, in der ein Durchschnitt als Maßzahl verwendet wird.

Daraus lässt sich aber nicht schließen, dass die meisten Nutzer zwei Seiten sehen. Denn die Verteilung der Seiten pro Session zeigt ein ganz anderes Bild, wie in Abbildung 3 zu sehen ist.

Hier handelt es sich wieder um eine Art Histogramm, nur dass es um 90 Grad nach rechts gedreht wurde. Ganz offensichtlich existiert hier keine Normalverteilung, die Verteilung ist rechtsschief. Die meisten Nutzer sehen sich nur eine Seite an.[3] Der Durchschnitt ist deutlich verzerrt. Es sollte daher immer zunächst die Verteilung angesehen werden (die bei Google Analytics & Co meistens im Interface versteckt ist), bevor eine Zahl kommuniziert wird. Der Median, also die Mitte der Werte, wenn alle sortiert in einer Reihe aufgelistet würden, und der Modus (der häufigste Wert) liegen beide bei 1. Die Aussage „Die meisten Nutzer schauen sich eine Seite an“ hat eine ganz andere Wirkung als „Im Durchschnitt schauen sich die Nutzer etwas mehr als zwei Seiten an“ (wobei sich hier schon die Frage stellt, was „etwas mehr“ als zwei Seiten sein soll). Bei der

Page Depth	Sessions	Page Views
<1	66	0
1	10,005	10,005
2	2,465	4,930
3	1,425	4,275
4	742	2,968
5	495	2,475
6	288	1,728
7	201	1,407
8	134	1,072
9	110	990
10	96	960

Abb. 3: Die Verteilung der Anzahl der Seiten pro Sitzung, recht versteckt im Google Analytics Interface

INFO

Die Normalverteilung heißt übrigens nicht deswegen so, weil nur sie Normalität darstellt. Tatsächlich sind auch andere Verteilungen „normal“. Ihren Namen hat die Normalverteilung nur darum erhalten, weil Adolphe Quetelet durch die Messung der Brustumfänge von Tausenden Soldaten eine solche Verteilung erhielt und sie deshalb für normal hielt. Bereits Jahrzehnte zuvor erwähnte Gauß diese Verteilung, sodass sie auch manchmal Gauß-Verteilung genannt wird.

[3] Und seltsamerweise existieren Sitzungen ohne Pageviews, aber das ist ein anderes Thema.

> „Die Verteilung der Daten zu verstehen, ist die Grundlage für eine korrekte Analyse.“

Browser	Acquisition			Behaviour		
	Users	New Users	Sessions	Bounce Rate	Pages/Session	Avg. Session Duration
	21,234 % of Total: 100.00% (21,234)	18,498 % of Total: 100.20% (18,461)	28,062 % of Total: 100.00% (28,062)	13.13% Avg for View: 13.13% (0.00%)	1.23 Avg for View: 1.23 (0.00%)	00:03:25 Avg for View: 00:03:25 (0.00%)
1. Chrome	**14,324** (67.53%)	12,360 (66.82%)	19,173 (68.32%)	9.74%	1.23	00:03:44
2. Safari	**2,516** (11.86%)	2,304 (12.46%)	3,099 (11.04%)	33.20%	1.25	00:02:07
3. Edge	**1,223** (5.77%)	1,034 (5.59%)	1,862 (6.64%)	8.27%	1.21	00:03:06
4. Firefox	**1,111** (5.24%)	960 (5.19%)	1,433 (5.11%)	9.42%	1.19	00:03:16
5. Internet Explorer	**1,100** (5.19%)	988 (5.34%)	1,359 (4.84%)	8.02%	1.26	00:03:06
6. Opera Mini	**215** (1.01%)	194 (1.05%)	239 (0.85%)	74.06%	1.49	00:02:04
7. Android Webview	**211** (0.99%)	209 (1.13%)	226 (0.81%)	37.61%	1.05	00:01:18
8. Samsung Internet	**190** (0.90%)	160 (0.86%)	245 (0.87%)	9.80%	1.23	00:03:42
9. Opera	**185** (0.87%)	161 (0.87%)	264 (0.94%)	18.18%	1.20	00:03:46
10. UC Browser	**97** (0.46%)	91 (0.49%)	113 (0.40%)	39.82%	1.15	00:01:35

Abb. 4: Browser-Report in Google Analytics

ersten Aussage „fühlt“ man bereits die nächste Aktion, bei der zweiten spürt man eher ein „Na und?“.

Der Durchschnitt der Seiten pro Sitzung ist nicht der einzige Durchschnitt in Abbildung 2. Und für jeden dieser Werte ist auf den ersten Blick nicht klar, ob eine Normalverteilung vorliegt oder nicht. Durchschnittliche Sitzungsdauer? Durchschnittliche Anzahl von Sitzungen? Kaum nützlich, um daraus eine Aktion abzuleiten. Und selbst wenn eine Normalverteilung vorläge, so wäre die nächste Frage, wie groß die Streuung um den Mittelwert ist. Je breiter die Streuung, desto weniger ist der Durchschnitt geeignet, etwas über die Daten auszusagen.

Auch wenn fraglich ist, ob die Anzahl besuchter Seiten pro Sitzung überhaupt ein guter KPI ist, so wird an diesem Beispiel deutlich, dass, auch wenn schlaue Menschen Werkzeuge wie Google Analytics & Co bauen, dies einen nicht davon befreit, das eigene Gehirn anzustrengen, um die Daten besser zu verstehen. Gleichzeitig wäre die Mehrzahl der Benutzer verwirrt, wenn sie einen Median und Quartile neben dem Durchschnitt im Dashboard sähen. Es ist sicherlich keine böse Absicht der Tool-Anbieter, dass sie das in ihre Interfaces integrieren, was für die meisten Anwender einen leichten Einstieg in die Webanalyse bedeutet. Für eine fundierte Analyse reicht das aber in der Regel nicht aus.

Woran liegt‘s? Die diagnostische Analytik

Zeigen die korrekt beschriebenen Kennzahlen Abweichungen auf dem Weg zur Zielerreichung an, so ist die Frage, was die Treiber dafür sind. Aus den korrekt beschriebenen Daten soll nun ermittelt werden, was getan werden muss, um die Abweichung in Zukunft zu verhindern, oder, sollte die Abweichung positiv sein, wie man mehr davon haben kann. Gegen eine höhere Conversion-Rate hat kein Online-Shop etwas. Dadurch, dass zunächst das Hauptziel in Unterziele aufgeteilt wurde, sollten schon Anknüpfungspunkte für eine weitergehende Analyse vorhanden sein.

Aber auch wenn alle Ziele erreicht werden, lohnt es sich immer, zu fragen, ob durch Optimierungen nicht noch mehr herausgeholt werden kann. Um bei dem vorherigen Beispiel zu bleiben: Jeder Besucher, der sich nur eine Seite ansieht, wird bei einem E-Commerce-Shop zumindest in dieser Sitzung nichts kaufen. Bei einer Content-Website mit Werbung kann ein Single-Page-Visit gut sein, wenn Nutzer die Seite verlassen, weil sie auf eine Werbung geklickt haben. In beiden Fällen ist es aber notwendig, die Zahl nicht isoliert zu sehen, denn so kann keine Aktion von ihr abgeleitet werden (im Denglischen auch: „Sie ist nicht actionable.“). Wie auch im ersten Teil dieser Serie brauchen wir einen Bezug. Entweder existieren Kenntnisse, welcher Wert gut ist, oder aber es werden Segmente untereinander verglichen. Der erste Ansatz ist schwierig, da wenige Websites wirklich gut miteinander vergleichbar sind. Der zweite Ansatz dagegen ist eingeschränkt, denn wenn die Website insgesamt ein Problem hat, so würde dies nicht deutlich werden, weil ja jedes Segment unter diesem allgemeinen Problem „leiden“ würde. Dennoch ist dieser Ansatz häufig zielführend.

Segmente sind so was wie die Killerapplikation in der Webanalyse. In der Statistik ist dies seit jeher ein übliches Vorgehen: Die zu untersuchende Population wird in Teilpopulationen aufgeteilt, die sich je nach Fragestellung in ihren Merkmalsausprägungen unterscheiden. Ein Beispiel ist die Aufteilung von Wahlergebnissen in Teilpopulationen nach Bundesländern, Stadt versus Land, Geschlecht, Altersklasse oder Einkommensgruppe. Kommt zum Beispiel bei einer Analyse heraus, dass die Jungwähler eher grün wählen, so lässt sich daraus konkret ableiten, welche Schwerpunkte bedient werden müssen, um auch dieses Segment

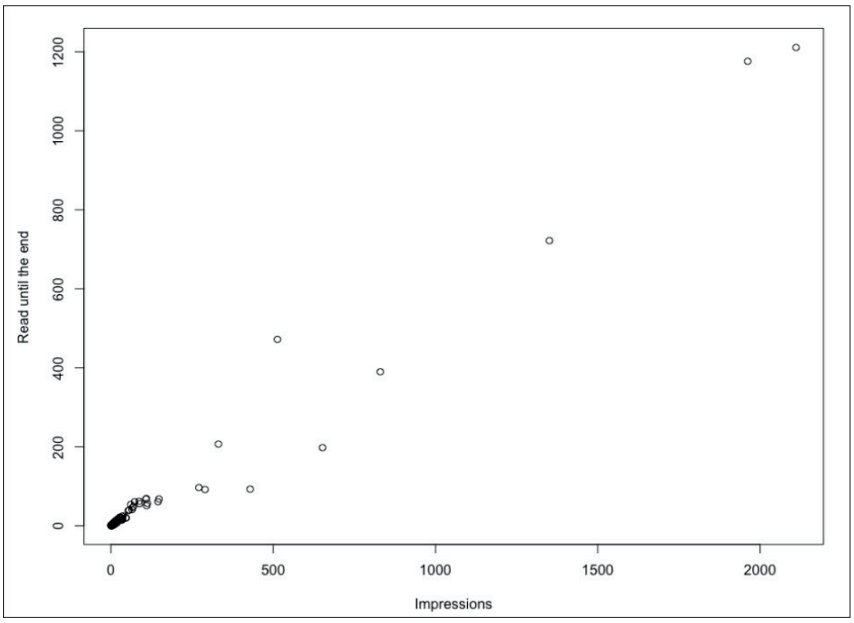

Abb. 5: Impressionen von Seiten auf der x-Achse und die Häufigkeit ihres kompletten Konsums („zu Ende gelesen") auf der y-Achse; jeder Punkt im Plot ist eine Seite

ansprechen zu können.

Nicht anders verhält es sich in der Webanalyse. So kann segmentiert werden anhand:

» Akquise-Kanal (Suche versus direkt zum Beispiel)
» Browser
» Betriebssystem
» Gerätekategorie (Mobil, Telefon, Tablet)
» Neuer versus wiederkehrender Nutzer (wobei diese Einteilung mit viel Vorsicht zu genießen ist, da nur Browser und nicht Nutzer erfasst werden)
» Vieler weiterer Teilpopulationen

Auf den ersten Blick können hier allerdings auch verwunderliche Daten zutage gefördert werden, wie in Abbildung 4 zu sehen.

Safari hat eine vielfach höhere Absprungrate („Bounce-Rate") als der am häufigsten genutzte Browser Chrome. Gleichzeitig sind die Anzahl Seiten pro Sitzung („Pages/Session") geringfügig höher. Entweder haben die Nutzer des Safari-Browsers bei der Anzahl der Seiten pro Besuch eine andere Verteilung erzeugt oder hier stimmt etwas anderes nicht. Tatsächlich handelt es sich hier um eine angepasste Absprungrate, das heißt, dass ein Absprung als solcher gezählt wird, wenn der Nutzer keine Interaktion zeigt, sie oder er also weder scrollt noch einen Link anklickt usw. Im nächsten Schritt würde zunächst einmal dieses Phänomen genauer untersucht werden. Von den „großen" Browsern sticht nur Safari hier heraus. Die „kleineren" Browser haben zwar auch hohe Absprung-Raten, werden aber auch von viel weniger Nutzern verwendet, sodass erst einmal der größte Abweicher in Bezug auf Nutzer analysiert wird.

Ein anderes Beispiel in der Analyse sind Zusammenhänge zwischen einzelnen Merkmalen. Oft möchte man wissen, ob man etwas an einem Ergebnis ändern kann, wenn eine Variable, auf die man Einfluss hat, verändert wird. Wer eine Seite mit Werbung betreibt, ist zum Beispiel daran interessiert, die Werbeeinnahmen zu erhöhen, und sucht nach Variablen, die das Ergebnis positiv beeinflussen. Allerdings bedeutet ein Zusammenhang in den Zahlen nicht immer auch, dass dieser Zusammenhang tatsächlich existiert. Daher wird auch von einem statistischen Zusammenhang gesprochen, wenn von einer Korrelation die Rede ist. Eine Korrelation ist keine Ursache-Wirkung-Beziehung.

Hohe Korrelationen sind relativ einfach in einer Visualisierung erkennbar, da die geplotteten Datenpunkte an einer imaginären Linie ausgerichtet sind wie in Abbildung 5 (hier am Beispiel eines positiven Korrelationskoeffizienten von 0,98; 1 wäre das Maximum). Man könnte das auch so ausdrücken: „Je mehr x, desto mehr y", beziehungsweise: „Je weniger x, desto weniger y", oder, bei einer negativen Korrelation: „Je weniger x, desto mehr y", beziehungsweise: „Je mehr x, desto weniger y."

In diesem Fall wird die Zahl der Impressionen einer Seite mit der Häufigkeit, dass sie zu Ende gelesen wurde, in Relation gesetzt. Offensichtlich kann hier ein Zusammenhang erwartet werden, denn je öfter eine Seite besucht wird, desto häufiger sollte sie auch die Chance haben, bis zum Schluss gelesen zu werden; zu wissen, welche Variable eine andere beeinflusst, ist allerdings nicht die Regel bei einer Korrelation. Interessant wäre nun noch zu schauen, ob auch die Länge eines Textes einen Einfluss auf die Wahrscheinlichkeit hat, dass ein Text zu Ende gelesen wird. Zu wissen, dass es einen Zusammenhang gibt, ist aber nicht unmittelbar handlungsrelevant. Um daraus eine Aktion abzuleiten, werden weitere Analyseschritte benötigt.

Von der Statistik lernen: Die explorative Datenanalyse

Wie an diesen Beispielen zu sehen ist, besteht bei einer tiefergehenden Analyse die Gefahr, dass man schnell vom Hundertsten ins Tausendste kommt, sich in dem Datenwust verliert und am Ende nicht mal mehr erinnert, welche Fragestellung eigentlich verfolgt wurde. So findet man manchmal etwas, verliert das kleine Goldstückchen dann aber während der Analyse, weil weitere Gold-Nuggets woanders vermutet werden und dafür zwischendurch so viele Schritte unternommen werden, dass eine Reproduzierbarkeit

des ersten Goldstückchens schwierig bis unmöglich ist. Auch ist es für andere Analysten, die auf den Ergebnissen aufbauen, nicht immer einfach, die Gedanken zu reproduzieren. Und selbst als Einzelkämpfer erinnert man sich nicht immer, was man vor drei Wochen herausgefunden hatte. Auch dafür existiert eine Lösung, wenngleich sie erfordert, zunächst einmal von den Interfaces, die Google & Co bieten, Abstand zu nehmen.

Im Data-Science-Bereich hat sich der Ansatz der explorativen Datenanalyse (EDA) als extrem nützlich erwiesen. Eingeführt wurde der Begriff bereits in den 1960er-Jahren durch John Tukey, der von der Statistik eine stärkere Beschäftigung mit Daten forderte. Dank der Vielfalt frei verfügbarer Programmier-Entwicklungsumgebungen sowie leistungsfähiger Rechner kann heute jeder mit einfachen Mitteln Daten systematisch explorativ analysieren. Dazu werden sogenannte Notebooks verwendet, die für verschiedene Programmiersprachen zur Verfügung stehen.

Das Besondere an den Notebooks ist, dass sowohl die Gedanken als auch der Code sowie die Ergebnisse des Codes und deren Interpretationen zusammengeführt werden. Der Autor eines Notebooks führt die Leser durch seine Gedankenwelt, zeigt den Code, der für den Ausdruck seiner Gedanken verwendet wurde, und erläutert auch seine Interpretation der Ergebnisse. Andere Analysten können sehen, ob der Code korrekt ist, und müssen sich nicht allein auf die Aussagen des Analysten verlassen. So kann auch zu einem späteren Zeitpunkt, wenn der Analyst vielleicht nicht mehr an Bord ist, alles nachvollzogen werden.

Patrick Lürwer hatte R-Notebooks bereits mit seiner Reihe „R für SEO“ in vorherigen Website-Boosting-Ausgaben vorgestellt und diese für das Erstellen eines Reports verwendet.[4] In

```r
plot(new_ga_data$n,new_ga_data$cv, xlab = "Impressions", ylab = "Read until the end")
abline(lm(new_ga_data$cv ~ new_ga_data$n, data = new_ga_data))
```

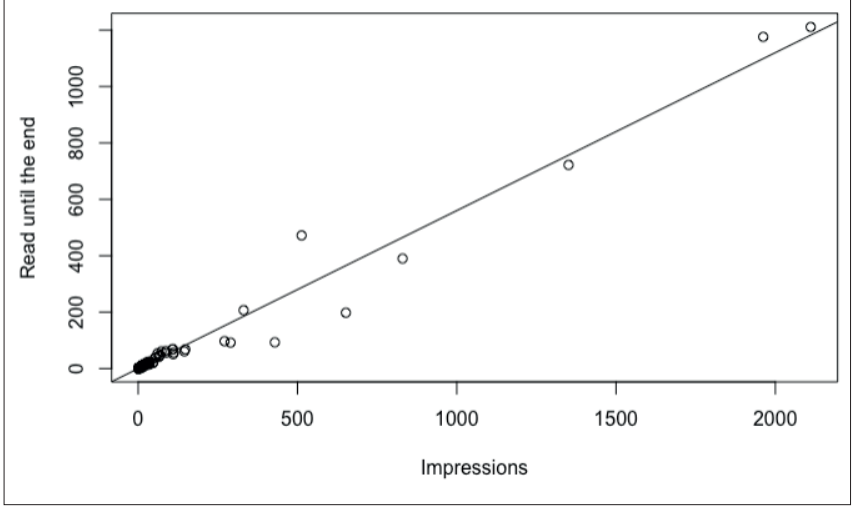

Abb. 6: Plot einer Regressionslinie in dem zuvor diskutierten Plot

der explorativen Datenanalyse werden Notebooks weniger für ein regelmäßiges Reporting, sondern, wie der Name schon sagt, zur Dokumentation der Exploration verwendet. Profi-Tipp: Gleich oben im Notebook das Ziel der Analyse definieren und dann bei jedem Abschnitt die Frage beantworten, ob die Frage beantwortet werden kann.

In den folgenden Abschnitten soll eine kurze Analyse vorgestellt werden. Wie bei Patrick wird auch hier das Package tidyverse von Hadley Wickham verwendet. Der Vorteil des tidyverse ist, dass mit wenigen Befehlen eine komplette Analyse durchgeführt werden kann und die analytischen Vorgehensweisen in einfache Programmbefehle übersetzt werden. Die folgenden Beispiele stammen aus einem Notebook, das unter *https://alby.link/websiteboostingnotebook* zusammen mit den Daten zur Verfügung gestellt wird. Die HTML-Version findet sich unter *https://alby.link/websiteboosting-beispieleda*. In beiden Versionen werden die einzelnen Schritte des Codes ausführlich erläutert.

Nachdem die Libraries und die Daten geladen und transformiert wurden, wird der zuvor in Abbildung 5 verwendete Plot der Impressionen im Verhältnis zu den zu Ende gelesenen Texten zusätzlich mit einer Regressionslinie versehen (siehe Abbildung 6). Eine solche Analyse wird üblicherweise dazu verwendet, eine Prognose zu erstellen: „Wenn x, wie viel ist dann y?“ Zuvor wurde bei der Korrelation nur der Zusammenhang festgestellt, nun wird der Wert der abhängigen Variablen beim Eintreten eines Werts der unabhängigen Variablen vorhergesagt. An der Regressionslinie ist erkennbar, welche Texte unter oder über der Regressionslinie, also dem zu erwartenden Wert, liegen.

Für die weitere Analyse wird nun ein kleiner Trick verwendet: Anstatt der Datenpunkte selber werden Residuen angesehen. Da meistens nicht alle Beobachtungen auf einer Regressionslinie liegen, werden die vertikalen Abstände der Beobachtungspunkte von dem zu erwartenden Wert gemessen. Dies sind die Residuen. Der Punkt ganz rechts oben hat einen geringeren Abstand zur Regressionslinie als der Punkt links daneben; das Residuum des ganz rechten Punkts beträgt

[4] Es existieren sog. Notebooks auch für Python, siehe zum Beispiel *https://jupyter.org/*.

```r
library(modelr)
mod <- lm(new_ga_data$cv ~ new_ga_data$n, data = new_ga_data)
tada <- new_ga_data %>%
  add_residuals(mod)
tada %>%
  filter(n > 50) %>%
  mutate(myPagePath = str_trunc(pagePath, 30)) %>%
  ggplot(aes(myPagePath, resid, group = 1)) +
  theme(axis.text.x = element_text(angle = 90)) +
  geom_ref_line(h = 0) +
  geom_line()
```

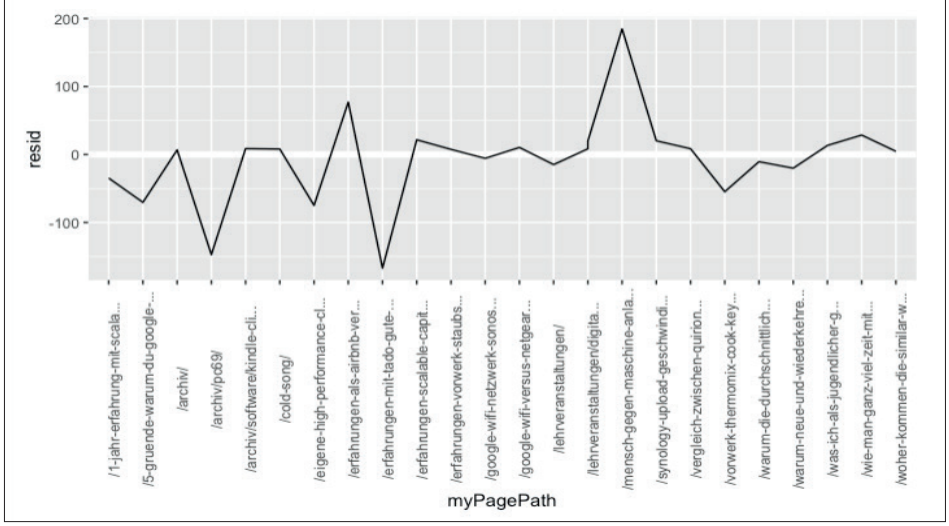

Abb. 7: Plot der Residuen

28,58, das des Punktes links daneben 77,04. Werden die Residuen geplottet, ergibt sich eine Visualisierung wie in Abbildung 7. Um in diesem Beispiel noch etwas lesen zu können, wurden nur Seiten mit mindestens 50 Aufrufen in dem Beobachtungszeitraum einbezogen.

Anders als in dem vorherigen Plot sind die Werte nun nicht mehr nach einem numerischen Wert sortiert; auf der x-Achse sind die einzelnen Seitenpfade abgebildet. Somit kann einfach abgelesen werden, welche Seiten besonders stark von dem zu erwartenden Wert abweichen.[5] Interessant ist hier, dass auch Seiten mit nur wenigen Hundert Aufrufen hohe Residuen haben können. Diese Analyse ist unmittelbar handlungsrelevant: Die Seiten mit hohen positiven Residuen haben etwas, was Seiten mit hohen negativen Residuen nicht haben. Die Texte können genauer analysiert werden, auch in Bezug darauf, mit welchen Suchbegriffen die Nutzer kommen, ob ihre Intention hier befriedigt wird. Sind Scroll-Daten vorhanden, so könnte zusätzlich nachgesehen werden, bis wohin die meisten Nutzer scrollen.

Ein Beispiel für eine längere EDA zum Thema SEO und Data Science kann unter *https://alby.link/eda* angesehen werden. Dieses Notebook wurde im Anschluss an einen kontrovers diskutierten Vortrag auf einer SEO-Konferenz zur Verfügung gestellt. Gerade im Bereich Suchmaschinenoptimierung, in dem viele Meinungen existieren, aber selten Fakten über anekdotische Evidenz hinaus geteilt werden, sind solche Ansätze zielführend.

In der nächsten Ausgabe lesen Sie dann, wie man weitere Analyse-Ansätze nutzen kann und wie dazu passende A/B-Tests aufgesetzt werden können.

[5] Normalerweise wird nun auch ein Histogramm der Residuen erstellt, das eine Normalverteilung zeigen sollte. Dieser Schritt wurde ausgelassen, dem Leser sei aber versichert, dass eine Normalverteilung der Residuen vorliegt.