Anwendung explorativer Datenanalyse zur Identifikation von Optimierungspotenzialen hinsichtlich des Umsatzes am Beispiel eines Online-Publishers mit Affiliate-Ansatz

Tobias Zwingmann

30.4.2019

Inhaltsverzeichnis

1. Einleitung **3**

1.1 Thema des Projektes 3

1.2 Fragestellung 3

1.3 Zielsetzung 4

2. Einordnung in den Kontext Data Science **4**

3. Dokumentation des Lösungsweges **5**

3.1 Business Understanding 5

3.1.1 Affiliate Marketing als Geschäftsmodell 5

3.1.3 Hintergründe zum konkreten Fallbeispiel 12

3.2 Data Understanding 14

3.3 Data Preparation 16

3.4 Explorative Datenanalyse 22

3.5 Analyse der Klickraten 25

3.5.1 Analyse der CTR-Verteilung 25

3.5.2 Regressionsmodellierung zur Vorhersage der Klicks in Abhängigkeit von Seitenaufrufen 31

3.5.3 Analyse der RPC-Verteilung 35

3.5.4 Regressionsmodellierung zur Vorhersage der Klicks in Abhängigkeit von Seitenaufrufen 39

3.6 Deployment: Aufbau des Scoring-Widgets 42

4. Zusammenfassung der Ergebnisse **45**

5. Kritische Würdigung **47**

1. Einleitung

1.1 Thema des Projektes

Im vorliegenden Projekt wird für ein Unternehmen im Bereich Online-Marketing ein Informationsportal untersucht, welches in erster Linie durch die Ausspielung programmatischer Werbeanzeigen sowie durch im Content platzierte Affiliate-Links Erlöse erzielt.

Während vom Seitenbetreiber Daten bereits in Echtzeit genutzt werden, um beispielsweise die Auslieferung und Positionierung von Werbeanzeigen automatisch zu optimieren, bedarf es auf strategischer Ebene anderer Ansätze, um die richtigen Entscheidungen für die Entwicklung des Geschäftsmodells zu treffen.

Im Projekt wünscht sich die Unternehmensleitung eine Einschätzung darüber, wie einzelne Inhalte und Themen auf dem Informationsportal zukünftig besser positioniert werden könnten, um die Gesamteinnahmen der Seite dauerhaft zu erhöhen. Dabei geht es explizit nicht um die Optimierung einzelner Klickpreise per se, sondern vorgelagert um die Generierung strategischer Handlungsempfehlungen, die entsprechende Aktionen für bestimmte Themenbereiche der Seite vorschlägt.

Diese Analysen sollen überwiegend durch quantitative Informationen aus verschiedenen Datenquellen gedeckt werden. Das Management erwartet dabei objektive und validierbare Informationen aus entsprechenden Nutzungsdaten der Webseite selbst und damit verbundener Systeme zur Unterstützung der strategischen Entscheidungsfindung.

1.2 Fragestellung

Zur Erstellung dieser strategischen Handlungsempfehlungen im Kontext des Online-Marketings (hier speziell: Affiliate-Marketing) müssen folgende zwei zentrale Probleme geklärt werden:

Das erste Problem bezieht sich auf die Definition eines Sets von Erfolgskennzahlen (KPIs). Als konzeptionelles Rahmengerüst dienen hierbei die branchenüblichen Referenzwerte Klickrate (CTR) und Umsatz-Pro-Klick (RPC). Diese beiden Kennzahlen müssen zum einen hinreichend konkretisiert (z. B. Definition des Zeitbezugs) und zum anderen in eine sinnvolle Verbindung miteinander gebracht werden. Die für die Berechnung der Kennzahlen notwendigen Daten liegen in unterschiedlichen Systemen vor (hier: Google Analytics für CTR und Amazon Partnernet für RPC). Zur Auflösung dieses Problems werden einerseits fest definierte Regelwerke (Business Rules) und geeignete Primärschlüssel zur Verknüpfung unterschiedlicher Datenquellen benötigt.

Die zweite Problemdimension bezieht sich auf die qualitative Beurteilung der einzelnen Kennzahlen, d.h. konkret: Ab wann ist ein bestimmter Wert für eine gegebene KPI gut oder schlecht?

Unterschiedliche Unterseiten haben unterschiedliche Besucher und Klickraten sowie Umsätze unterliegen zum Teil starken Schwankungen im Zeitverlauf (z. B. Abhängigkeit von Wochentagen). Daher reichen Trivialmodelle wie zum Beispiel statistische Mittelwerte zur Einschätzung der Güte einer KPI-Kennzahl nicht aus. Als Lösung bieten sich statistische Verfahren an, die beispielsweise die schiefe Verteilung der Input-Daten entsprechend berücksichtigen und valide Einschätzungen erlauben, ob ein zu einer bestimmten Zeit und auf einer bestimmten Seite erreichter CTR-Wert akzeptabel ist, oder nicht.

Zur Kommunikation der Ergebnisse wird eine instruktive Visualisierung angestrebt, die die Handlungsempfehlungen klar aufzeigt. Diese Visualisierung soll zudem programmatisch reproduzierbar sein, sodass die gewonnenen Einsichten auf Basis neuer Daten schnell erneut ausgewertet bzw. erneut Entscheidungen getroffen werden können.

1.3 Zielsetzung

Als zentrales Deliverable soll eine interaktive Visualisierung erstellt werden, die eine Einordnung von einzelnen Unterseiten des Informationsportals mit geeigneten KPIs in eine strategische Handlungsfeldmatrix vornimmt und entsprechende Handlungsempfehlungen vorschlägt, wie der im Beispielfall generierte Umsatz weiter gesteigert werden kann.

2. Einordnung in den Kontext Data Science

Die Lösung der gegebenen Problemstellung erfordert die effiziente Verknüpfung spezifischen Domänenwissens, statistischer Kenntnisse sowie der programmiertechnischen Fähigkeit zur Handhabung der unterschiedlichen Datenformate und Quellsysteme in Verbindung mit einer entsprechenden Datenmodellierung zur Ableitung zielführender Erkenntnisse. Der Schwerpunkt der Arbeit liegt auf der Datenanalyse und hier insbesondere auf dem Teilbereich der Visualisierung, d. h. die instruktive Darstellung komplexer Sachverhalte in programmatischer Form.

Für dieses Ziel fließt ein Großteil der Arbeit in die entsprechende Aufbereitung der Input-Daten zur korrekten Berechnung der KPIs (CTR und RPC) sowie der anschließenden Modellentwicklung zur Bewertung der KPI-Performance (Score-Wert), die dann abschließend visualisiert werden können.

3. Dokumentation des Lösungsweges

3.1 Business Understanding

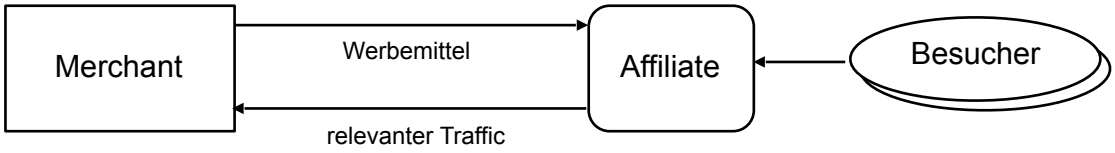
3.1.1 Affiliate Marketing als Geschäftsmodell

Unter „Affiliate-Marketing“ wird allgemein ein auf Online-Kooperationen basierendes Marketing- und Vertriebskonzept im Bereich des E-Business verstanden (Vgl. Kollmann, T. (2011), S. 299.). Affiliate-Marketing ist als Spezialbereich des Online-Marketings zu betrachten (Vgl. Jurišová, V. (2013), S. 5.).

Eine Affiliate-Partnerschaft besteht typischerweise aus einer Dreiecksbeziehung zwischen einem Online-Händler (sogenannter Merchant bzw. Programmbetreiber), einem Webseitenbetreiber (sogenannter Affiliate bzw. Publisher) und dessen Webseitenbesuchern (Vgl. Büttgen, M. (2003), S. 323.).

Merchant und Affiliate gehen dabei eine Kooperation ein: Der Webseitenbetreiber (Affiliate) bewirbt Produkte des Merchants auf seiner Website und wirkt damit als Absatzmittler des Merchants. Im Gegenzug erhält der Webseitenbetreiber eine erfolgsabhängige Vergütung. Das Verhältnis von Affiliate, Merchant und Webseitenbesuchern ist schematisch in Abbildung 2 dargestellt. Affiliate Marketing kann damit als Übertragung des seit Jahrzehnten existierenden Prinzips der Vertriebs- oder Netzwerkpartner in die Online-Welt gesehen werden. (Vgl. Lammenett, E. (2015), S. 45.)

Abbildung 2: Beziehungen im Affiliate-Marketing

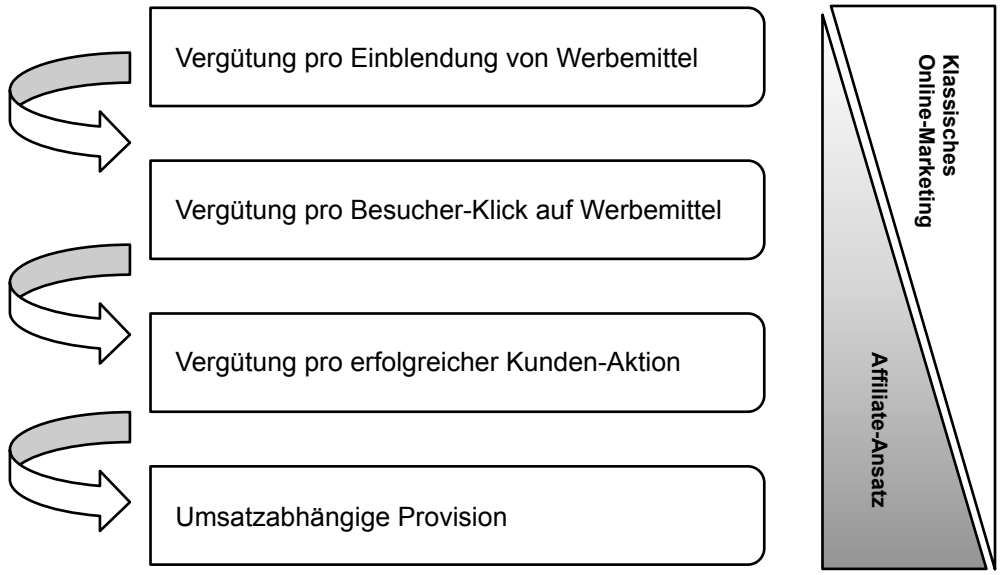


In der Regel bauen Merchants über Affiliate-Programme (Partnerprogramme) sehr viele Kooperationen zu unterschiedlichen Publishern auf. Das Ziel dabei ist es, den Zugang zu kontext- oder zielgruppenrelevanten Content- bzw. Community-Webseiten (Nischenwebseiten) zu verbessern. Zwar weisen diese Webseiten in der Regel im Vergleich zu reichweitenstarken Plattformen eine geringere Besucherfrequenz auf, bieten dafür aber eine erhöhte Themenrelevanz und damit bessere Chancen für einen Geschäftsabschluss (Vgl. Lücke, F., Webering, J. (2003), S. 13 sowie Tollert, D. (2009), S. 12.). Merchants stellen für die Bewerbung ihrer Produkte unterschiedliche Werbemittel bereit. Prinzipiell handelt es sich dabei um die gleichen Werbeformen, die auch im klassischen Online-Marketing genutzt werden, d. h. vor allem Grafikbanner und Text-Links.

Der essenzielle Unterschied zwischen Affiliate-Marketing und der herkömmlichen Bannerwerbung ist, dass die Partner beim Affiliate-Marketing grundsätzlich erfolgsabhängig vergütet werden, d. h. zum Beispiel bei Erreichung eines Kaufabschlusses (Pay per Sale) oder der Registrierung eines Neukunden (Pay per Lead). Kaufabschlüsse werden in der Regel innerhalb einer bestimmten Frist angerechnet (Vgl. Tollert, D. (2009), S. 17f.).

In der klassischen Bannerwerbung werden hingegen häufig die reine Einblendung eines Werbemittels (Pay per View) oder der Klick auf ein Werbemittel (Pay per Click) vergütet. Darüber hinaus existieren natürlich auch Mischformen, sodass in einigen Fällen auch Affiliate-Programme mit beispielsweise einer Pay-per-Click-Abrechnung angeboten werden. Diese Formate kommen in der Gesamtbetrachtung allerdings äußerst selten vor. Eine schematische Darstellung der Abgrenzung zwischen klassischem Online-Marketing und Affiliate-Marketing ist in Abbildung 3 zu sehen.

Abbildung 3: Abgrenzung des klassischen Online-Marketings und des Affiliate-Ansatzes anhand der Vergütungsmodelle im E-Business3.1.2 Definition relevanter Kennzahlen



Zur Messung der Leistung einer Website werden unterschiedliche Kennzahlen erhoben. Im Folgenden sollen die zwei wichtigsten Kennzahlensysteme für die beiden häufigsten Vergütungsmodelle Pay per Click (PPC) und Pay per Sale (PPS) erläutert und einander gegenübergestellt werden.

Im klassischen PPC-Ansatz bilden die sogenannte „Click-Through-Rate“ (CTR) und der „Cost-per-Click“ (CPC) die beiden maßgeblichen Kennzahlen zur Bestimmung der Leistungsfähigkeit eines Online-Projektes. Zur weiteren Erläuterung sollen nachfolgend einige Notationen und Definitionen eingeführt werden.

Die Indexvariable s bezeichnet im Kontext der vorliegenden Arbeit eine jeweils fest definierte Seiten-URL, inklusive der dieser Seite zugeordneten Werbemittel. Die Indexvariable t referenziert eine bestimmte Periode, in der Leistungen (z. B. Klicks oder Umsätze) beobachtet wurden. Die Perioden weisen untereinander jeweils die gleiche Länge auf und überschneiden sich zeitlich nicht. Eine Periode ist beispielsweise ein genaues Tagesdatum. Die Anzahl aller Perioden wird mit der Variable p gekennzeichnet und als Beobachtungszeitraum definiert. Die Anzahl aller Seiten-URLs wird mit der Variable n notiert, die zugleich als Seitenumfang dient.

Die Anzahl der Aufrufe einer bestimmten Seiten-URL in einer festgelegten Periode t wird als v_{st} definiert. Entsprechend lässt sich die Menge aller Aufrufe je Seiten-URL in allen beobachteten Perioden als Matrix mit $n \times p$ Elementen von v_{st} darstellen (Formel 1).

Matrizen werden nachfolgend mit fett gedruckten Großbuchstaben symbolisiert.

Formel 1: Definition der Matrix $\mathbf{V}$

$$\mathbf{V} := (v_{st})_{s=1,\ldots,n;t=1,\ldots,p} \in \mathbb{R}^{n\times p}$$

Vektoren sollen nachfolgend als nicht-fettgedruckte Großbuchstaben dargestellt werden. Die Matrix $\mathbf{V}$ kann durch Vektorisierung auch als Vektor V ausgedrückt werden. Die Vektorisierung von $\mathbf{V}$ ist in Formel 2 festgelegt.

Formel 2: Definition der Vektorisierung der Matrix $\mathbf{V}$

$$V := vect(\mathbf{V}) := (v_{11}, v_{12}, \ldots, v_{1p}, \ldots, v_{n1}, \ldots, v_{np}) \in \mathbb{R}^{np}$$

Die in einer Periode t erzielten Klicks auf ein oder mehrere Werbemittel einer bestimmten Seiten-URL s sollen mit c_{st} bezeichnet werden. Auch hier wird die Gesamtheit aller Klicks pro Webseite und Periode innerhalb eines Beobachtungszeitraums als Matrix mit $n \times p$ Elementen definiert. Die Definition der Matrix $\mathbf{C}$ ergibt sich analog zur Matrix $\mathbf{V}$ in Formel 3.

Formel 3: Definition der Matrix $\mathbf{C}$

$$\mathbf{C} := (c_{st})_{s=1,\ldots,n;t=1,\ldots,p} \in \mathbb{R}^{n\times p}$$

Analog zu $\mathbf{V}$ und V kann auch hier eine Vektorisierung der Matrix vorgenommen werden. Die Vektorisierung von $\mathbf{C}$ ist in Formel 4 dargestellt.

Formel 4: Definition der Vektorisierung von $\mathbf{C}$

$$C := vect(\mathbf{C}) := (c_{11}, c_{12}, \ldots, c_{1p}, \ldots, c_{n1}, \ldots, c_{np}) \in \mathbb{R}^{np}$$

Aus Sicht eines Webseitenbetreibers misst die Klickrate CTR allgemein, wie oft ein Werbemittel im Verhältnis zu dessen Einblendungen angeklickt wird. Der Klick auf ein bezahltes Werbemittel kann in diesem Kontext auch als Konversion interpretiert werden. Die Kennzahl CTR lässt sich sowohl auf einzelne Werbemittel als auch auf eine ganze Seiten-URL anwenden. Im vorliegenden Fallbeispiel soll die Interpretation der Seiten-CTR verwendet werden, welche die Leistung aller Werbemittel auf einer Seiten-URL misst. Hintergrund ist, dass nicht die Leistungssteigerung einzelner Werbebanner im Vordergrund steht, sondern die Identifikation von seitenübergreifenden Optimierungspotenzialen.

Entsprechend dazu ist die Kennzahl der Seiten-CTR in Formel 5 notiert für $s = 1, \ldots, n$ und $t = 1, \ldots, p$.

Formel 5: Definition der Kennzahl CTR

$$ctr_{st} := \frac{c_{st}}{v_{st}}$$

Die Matrix $\mathbf{CTR}$ soll alle Einzelwerte der n Seiten-CTR in den p Perioden beinhalten und ist in Formel 6 definiert.

Formel 6: Definition der Matrix $\mathbf{CTR}$

$$\mathbf{CTR} := (ctr_{st})_{s=1,\ldots,n;t=1,\ldots,p} \in \mathbb{R}^{n \times p}$$

Der Vektor CTR stellt auch hier die Vektorisierung von $\mathbf{CTR}$ dar und ist in Formel 7 dargelegt.

Formel 7: Definition der Vektorisierung der $\mathbf{CTR}$

$$CTR := vect(\mathbf{CTR}) := (ctr_{11}, ctr_{12}, \ldots, ctr_{1p}, \ldots, ctr_{n1}, \ldots, ctr_{np}) := (ctr_{st}) \in \mathbb{R}^{np}$$

Für die Fälle, in denen auf die durchschnittliche CTR verwiesen wird, soll das arithmetische Mittel aller Elemente des Vektors CTR gemeint sein. Die durchschnittliche Seiten-CTR ist als $\overline{CTR}$ gekennzeichnet und in Formel 8 erfasst.

Formel 8: Definition der Durchschnitts-CTR

$$\overline{CTR} := \frac{(ctr_{11} + ctr_{12} + \ldots + ctr_{1p} + \ldots + ctr_{n1} + \ldots + ctr_{np})}{np}$$

Während die Kennzahl CTR die Akzeptanz eines Werbemittels durch Webseitenbesu-cher beschreibt, stellt die Kennzahl CPC (Cost per Click) grundsätzlich einen Indikator für den Preis dar, den ein Werbekunde für einen Klick auf dessen Anzeigen zu zahlen bereit ist. Der tatsächlich bezahlte Preis pro Klick kann dabei je nach Werbemittel, Nutzergruppen, Zeit, Platzierung oder weiteren Variablen stark variieren. Die Preise werden in der Regel automatisiert durch ein WerbeNetzwerk im Gebotsverfahren berechnet.

Die Kennzahl CPC definiert sich allgemein aus dem Verhältnis von Klicks auf Werbemittel, die auf einer bestimmten Seiten-URL s platziert sind, zu den Erlösen e, die daraus in einer bestimmten Periode t resultieren. Sie kann analog zur Kennzahl ctr_{st} für alle $s = 1, \ldots, n$ und $t = 1, \ldots, p$ gebildet werden und definiert sich gemäß Formel 9.

Formel 9: Definition der Kennzahl CPC

$$cpc_{st} \coloneqq \frac{e_{st}}{c_{st}}$$

Die einzelnen Elemente von e_{st} lassen sich analog zur Bildung von V und C als Vektor $E \coloneqq (e_{11}, e_{12}, \ldots, e_{1p}, \ldots, e_{n1}, \ldots, e_{np}) \in \mathbb{R}^{np}$ formulieren.

Aus Sicht eines Webseitenanbieters ist die Kennzahl CPC nur im Kontext mehrerer Beobachtungen (d. h. mehrerer Klicks) sinnvoll, da die Einzelwerte je nach Bietersituation stark schwanken können. Analog zur Matrix $\mathbf{CTR}$ kann auch hier eine Matrix $\mathbf{CPC}$ mittels aller Elemente von cpc_{st} für $s = 1, \ldots, n$ und $t = 1, \ldots, p$ gebildet werden. Folglich ergibt sich die Vektorisierung dieser Matrix ebenfalls analog zu Formel 10.

Formel 10: Definition der Vektorisierung der Matrix $\mathbf{CPC}$

$$CPC \coloneqq vect(\mathbf{CPC}) \coloneqq (cpc_{11}, cpc_{12}, \ldots, cpc_{1p}, \ldots, cpc_{n1}, \ldots, cpc_{np}) \coloneqq (cpc_{st}) \in \mathbb{R}^{np}$$

Zur Bestimmung der Leistungsfähigkeit einer Webseite werden in der Regel CPC-Mittelwerte über mehrere Perioden gebildet. Das als $\overline{CPC}$ definierte arithmetische Mittel der Elemente von CPC bestimmt sich analog zu $\overline{CTR}$ und ist in Formel 11 dargestellt.

Formel 11: Definition der Durchschnitts-CPC

$$\overline{CPC}: = \frac{(cpc_{11} + cpc_{12} + \ldots + cpc_{1p} + \ldots + cpc_{n1} + \ldots + cpc_{np})}{np}$$

Der Gesamtumsatz einer Website mit PPC-Ansatz entspricht letztlich den kumulierten Erlösen aus allen Werbemittel-Klicks. Unter Verwendung der zuvor genannten Kennzahlen lässt sich jedoch der Gesamtumsatz auch als Summe für $k = np$ Seitenaufrufe innerhalb eines Beobachtungszeitraums, multipliziert mit der hierin erzielten durchschnittlichen CTR und den durchschnittlichen CPC über alle Seiten und über alle Werbemittel, wie in Formel 12 dargestellt, berechnen:

Formel 12: Berechnung des Umsatzes im PPC-Ansatz

$$U \coloneqq \left(\sum_{(i=1)}^{k} V_i \right) \overline{CTR} \times \overline{CPC})$$

Unter der Annahme, dass die seitenweite durchschnittliche CTR und die durchschnittlichen CPC konstant bleiben, führt ein Anstieg der Seitenaufrufe zu einem höheren Gesamtumsatz. Gleichfalls

erhöht sich der Umsatz auch dann, wenn bei konstanten Seitenaufrufen die durchschnittliche CTR oder die durchschnittlichen CPC oder beide Kennzahlen gleichzeitig gesteigert werden.

Die Systematik der Umsatzermittlung bei Webseiten mit PPS-Ansatz ist hierzu fast identisch. Der wesentliche Unterschied liegt lediglich in der Substitution der Kennzahl CPC durch die Kenngröße Revenue-per-Click (RPC). RPC gibt die Höhe der vermittelten Netto-Umsätze im Verhältnis zu den Klicks auf ein oder mehrere Werbemittel an.

Die Höhe der Netto-Umsätze, die in einer bestimmten Periode t durch Klicks auf die auf einer Seite s platzierten Werbemittel generiert wurden, sollen im Folgenden als r_{st} bezeichnet werden. Somit ergibt sich auch hier analog zu c_{st} eine $n \times p$ Matrix, welche die Netto-Umsätze über mehrere (n) Seiten-URLs in einer fest definierten Anzahl (p) von Perioden enthält. Diese Matrix soll mit $\mathbf{R}$ gekennzeichnet werden und ist in Formel 13 definiert.

Formel 13: Definition der Matrix $\mathbf{R}$

$$\mathbf{R} := (r_{st})_{s=1,\ldots,n;t=1,\ldots,p} \in \mathbb{R}^{n\times p}$$

Die Vektorisierung dieser Matrix erfolgt analog zu den vorherigen Definitionen und soll als Vektor R bezeichnet werden.

Formel 14: Definition der Vektorisierung von $\mathbf{R}$

$$R := vect(\mathbf{R}) := (r_{11}, r_{12}, \ldots, r_{1p}, \ldots, r_{n1}, \ldots, r_{np}) := (r_{st}) \in \mathbb{R}^{np}$$

Die Kennzahl RPC ergibt sich entsprechend der Berechnung der CPC als Quotient der durch die Werbemittel einer Seite s vermittelten Netto-Umsätze und der Anzahl der Klicks auf die dieser Seite zugeordneten Werbemittel in einer bestimmten Periode t. Somit gilt wie auch für den Umsatz pro Klick (RPC) die Definition in Formel 15 für $s = 1, \ldots, n$ und $t = 1, \ldots, p$.

Formel 15: Definition der Kennzahl RPC

$$rpc_{st} := \frac{r_{st}}{c_{st}}$$

Die Matrix RPC ergibt sich entsprechend der Formel 16.

Formel 16: Definition der Matrix $\mathbf{RPC}$

$$\mathbf{RPC} := (rpc_{st})_{s=1,\ldots,n;t=1,\ldots,p} \in \mathbb{R}^{n\times p}$$

Auch hier soll die Vektorisierung der Matrix definiert werden, sodass alle Elemente von $\mathbf{RPC}$ auch in Form eines Vektors RPC dargestellt werden können. Die Vektorisierung ist in Formel 17 formuliert.

Formel 10: Definition der Vektorisierung der Matrix $\mathbf{RPC}$

$$RPC \coloneqq vect(\mathbf{RPC}) \coloneqq (rpc_{11}, rpc_{12}, \ldots, rpc_{1p}, \ldots, rpc_{n1}, \ldots, rpc_{np}) \coloneqq (rpc_{st}) \in \mathbb{R}^{np}$$

Der durchschnittliche RPC definiert sich ebenfalls als das arithmetische Mittel aller Elemente des Vektors RPC. Die Berechnung der Kennzahl RPC ist in Formel 18 dargestellt.

Formel 18: Definition des Durchschnitts-RPC

$$\overline{RPC}: = \frac{(rpc_{11} + rpc_{12} + \ldots + rpc_{1p} + \ldots + rpc_{n1} + \ldots + rpc_{np})}{np}$$

Der Gesamterlös eines Affiliates ergibt sich aus der Verrechnung des vermittelten Umsatzes mit der jeweils individuell durch den Merchant festgelegten Provisionsrate. (Je nach Merchant und Provisionsmodell werden einzelne Produkte oder Produktkategorien unterschiedlich vergütet. Da es hier eine Vielzahl von Ausprägungen gibt und die Aussagen dazu praktisch nicht zu verallgemeinern sind, soll die Provisionsrate an dieser Stelle nicht weiter berücksichtigt werden. Sie lässt sich jedoch bei Bedarf einfach als Multiplikationsfaktor der RPC-Kennzahl ergänzen.)

Die RPC-Kennzahl kann isoliert (d. h. ohne Provisionsbetrachtung) als inhaltliches Äquivalent zur CPC-Kennzahl verstanden werden, da sie ebenfalls Auskunft über den Wert eines Werbemittels gibt. Lässt man die Provisionsberechnung außen vor, kann der durch einen Affiliate vermittelte Gesamtumsatz auch wie in Formel 19 folgt notiert werden:

Formel 19: Berechnung des Umsatzes im PPS-Ansatz

$$U \coloneqq \left(\sum_{(i=1)}^{k} V_i \right) \overline{CTR} \times \overline{RPC})$$

Im Gegensatz zur CPC- kann die RPC-Kennzahl grundsätzlich auch Null betragen. Dieser Fall tritt dann ein, wenn zwar Klicks auf ein Werbemittel stattfanden, aber keine Einkäufe beim Merchant getätigt wurden. Im PPC-Ansatz würde ein Werbemittel genau dann automatisch nicht mehr eingeblendet werden, wenn kein Werbebudget dafür bereitsteht oder kein Angebot platziert wurde. Im PPS-Ansatz ist also zusätzlich zur ersten Konversion (Klick auf ein Werbemittel) eine Konversion auf zweiter Ebene (Einkauf beim Merchant) notwendig, um Umsätze zu generieren.

Zusammenfassend lässt sich feststellen, dass die Kennzahlen CTR und CPC (bzw. im PPS-Ansatz RPC) sehr gut zur Analyse von Umsatzoptimierungen eingesetzt werden könnten, da sie unabhängig voneinander optimierbar sind und dabei ebenso unterschiedliche Handlungsfelder bedienen. Die Kennzahlen sind zudem relativ robust, leicht interpretierbar und gut auf unterschiedliche Sachverhalte übertragbar – unabhängig vom gewählten Erlösmodell. Robuste Verfahren zur Umsatzoptimierung lassen sich daher bei PPC-Modellen auf der Optimierung der CTR- und CPC-

Kennzahl sowie bei PPS-Modellen auf einer Optimierung der CTR- und RPC-Kennzahl basierend, aufbauen.

3.1.3 Hintergründe zum konkreten Fallbeispiel

Das im Rahmen dieser Arbeit vorgestellte Fallbeispiel ist eine Online-Projekt, dessen Geschäftsmodell hauptsächlich auf einem Affiliate-Ansatz beruht. Der Verfasser der vorliegenden Arbeit ist gleichzeitig Betreiber der Webseite und verfügt damit über uneingeschränkten Zugriff auf die Leistungsdaten des Projektes.

Das hier vorgestellte Projekt trägt den Titel „HDMI Guru“ und wird im Folgenden kurz als „Webseite“ bezeichnet. Die Webseite wurde im Februar 2012 unter der Domain https://www.welches-hdmi-kabel.de registriert und dient seither als Plattform für die Veröffentlichung von deutschsprachigen Inhalten zu Themen der Unterhaltungselektronik. Im August 2016 umfasste der Webauftritt insgesamt 50 Artikel, die in unregelmäßigen Abständen veröffentlicht wurden. Inhaltlich fokussiert sich das Angebot auf Blogbeiträge und Ratgeber-Artikel zum Themenkomplex der HDMI-Schnittstelle, dem aktuellen de-facto Standard zur Verkabelung im Heimkinobereich. Jährlich greifen rund 800.000 Besucher auf das Angebot zu.

Aufgrund der hohen thematischen Spezialisierung und der relativ überschaubaren Reichweite kann diese Seite auch als typischer Vertreter eines Nischenwebseiten-Konzeptes angesehen werden.

Das Geschäftsmodell der Webseite basiert im Wesentlichen auf einem Affiliate-Ansatz. Rund 90 % der Umsätze werden über das deutschsprachige Partnerprogramm von Amazon.de generiert. Hierüber werden an Amazon vermittelte Einkäufe in den Ländern Deutschland, Österreich und der Schweiz abgedeckt.

Die übrigen rund 10 % der Einnahmen generieren sich durch Pay-Per-Click Werbeeinblendungen, die vom Google-Adsense-Netzwerk bereitgestellt werden.

Im Jahr 2018 wurden so mithilfe des Amazon Partnerprogramms rund 26.000 Produkte im Gesamtwert von rund 430.000 Euro vermittelt.

Aufgrund der hohen Relevanz des Amazon-Partnerprogramms für den Geschäftserfolg der vorgestellten Affiliate-Webseite soll dieses Programm nachfolgend in den wesentlichen Grundzügen erläutert werden:

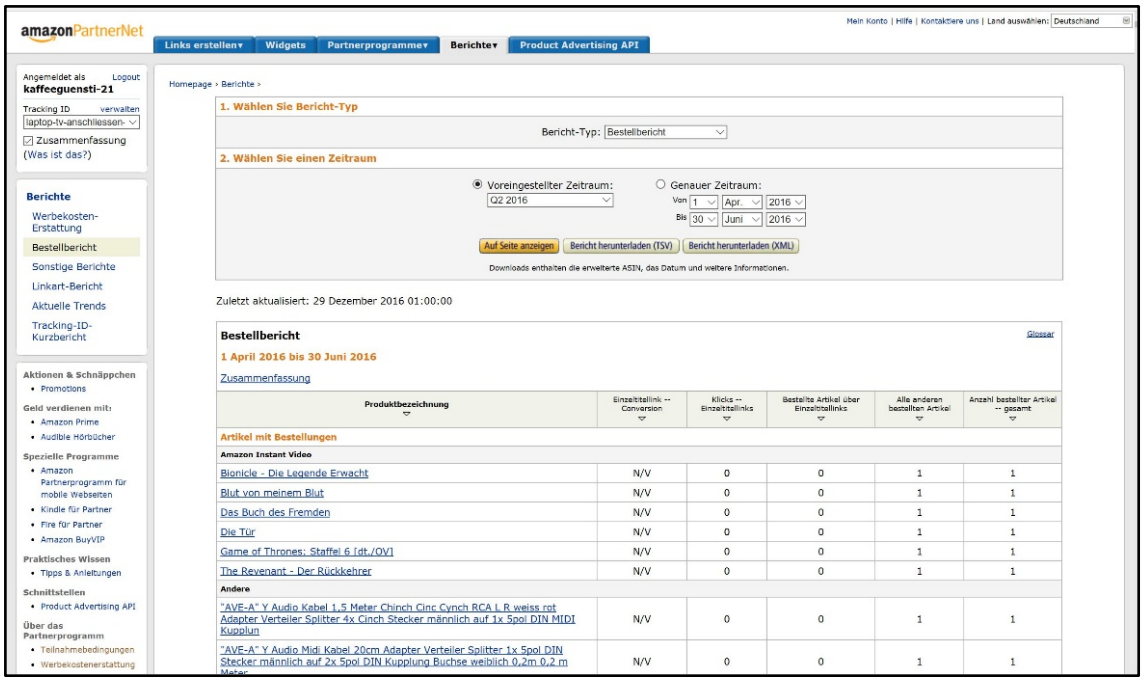
Das Amazon Partnerprogramm bietet Betreibern von Webseiten eine Möglichkeit, nachvollziehbare Links zum Produktsortiment von Amazon.de zu generieren und diese dann auf der Betreiber-Webseite einzubinden. Kauft ein Besucher bei Amazon ein, nachdem er über einen solchen Partner-Link auf das Angebot von Amazon.de gelangt ist, so erhält der Affiliate eine Provision auf den Gesamtwert des vermittelten Warenkorbs.

Dabei werden alle Käufe, die innerhalb von 24 Stunden nach Klick des Affiliate-Links getätigt werden, vergütet. Die Nachverfolgung geschieht über das Setzen eines sogenannten Tracking-Cookies. Die Identifizierung eines Affiliate-Links erfolgt über das Hinzufügen eines URL-Parameters in Form einer sogenannten „Tracking-ID“.

Ein Publisher kann mehrere Tracking-IDs zur Identifizierung unterschiedlicher Webseiten einsetzen. Die Provisionssätze unterscheiden sich je nach Produktkategorie und betragen zwischen 1 % bis maximal 10 % des Nettobestellwerts. Vergütet werden nur sogenannte „qualifizierte Bestellungen“, das umfasst im Wesentlichen Bestellungen, die nicht durch den Affiliate selbst getätigt wurden und die nicht durch den Käufer retourniert wurden.

Die Teilnahme am Amazon-Partnerprogramm ist für Webseitenbetreiber kostenlos. Statistiken zur Leistungsfähigkeit des Programms können durch den Affiliate über das Analyse- und Abrechnungstool „Amazon PartnerNet“ erstellt werden. Abbildung 4 zeigt exemplarisch ein Bildschirmfoto des Amazon PartnerNets.

Abbildung 4: Screenshot Amazon PartnerNet



Die Besucher des Fallbeispiels werden hauptsächlich über Suchmaschinenmarketing akquiriert: Rund 85% der Webseitenbesucher erreichen das Angebot über die organische Suche von Google.de. Rund 12% der Besucher gelangen per Direktzugriff auf die Seite. Weitere 3% gelangen über Verweise von anderen Webseiten oder sozialen Netzwerken auf das Angebot.

Das Grundkonzept der Webseite verfolgt dabei die Idee, dass Internetnutzer ein bestimmtes Thema im Bereich der HDMI-Thematik über Google.de recherchieren.

Ziel ist es nun, zu dieser bestimmten Suchanfrage eine möglichst gute Platzierung in den organischen Suchergebnissen von Google.de zu erreichen und den Besucher zielgerichtet auf die eigene Seite zu lenken. Dort sollte der Besucher dann nicht nur die nachgefragten Informationen, sondern außerdem Verweise bzw. Empfehlungen zu Produkten vorfinden, die für die Problemlösung hilfreich sind.

Dies sind in der Regel Verweise auf Kabel, Adapter oder weitere Elektronikzubehörartikel. Die entsprechenden Produkt-Links wurden zuvor manuell über das Amazon-Partnerprogramm generiert und in die einzelnen Ratgeber-Artikel eingefügt.

Zur Messung der Webseiten-Metriken wird das kostenlose und zugleich weit verbreitete Tool „Google Analytics“ eingesetzt. Über Google Analytics werden unter anderem Zugriffe auf einzelne Beiträge (Artikel) über die jeweiligen (im Zeitverlauf gleichbleibenden) Artikel-URLs ausgewertet.

Das Tracking der Affiliate-Performance wird über sogenannte Tracking-IDs realisiert. Dabei wird für jede Unterseite (d. h. für jeden Beitrag) eine einzigartige Tracking-ID erstellt, die an die ausgehenden Produkt-Links zu Amazon übergeben wird. Über diese Tracking-ID lässt sich im Amazon Partnerprogramm sehen, welche Unterseite wie viele Klicks und wie viele Produkte erzielt hat.

Die im Amazon PartnerNet hinterlegten Tracking-IDs stellen somit einen eindeutigen inhaltlichen Bezug zu den einzelnen Seiten-URLs in Google Analytics her und können als Schlüssel angesehen werden, der Daten aus Google Analytics und Amazon Partnernet verknüpft.

Als Untersuchungszeitraum wird der zum Zeitpunkt des Beginns dieser Arbeit aktuellste Zwölfmonatszeitraum vorgegeben. Dies entspricht dem 18. Dezember 2017 bis 16. Dezember 2018.

3.2 Data Understanding

Die zur Analyse des Fallbeispiels benötigten Daten werden von zwei unterschiedlichen Quellen bereitgestellt: Besucherzahlen werden über das Tool Google Analytics erhoben, Daten zu Klicks und vermittelten Umsätzen über das Amazon PartnerNet.

Die Daten aus Google Analytics werden in Form von kommagetrennten Text-Dateien (CSV) angeliefert. Google Analytics generiert pro Seiten-URL eine CSV-Datei für einen individuell einstellbaren Zeitraum. Im Fallbeispiel wurden somit 13 unterschiedliche CSV-Dateien für jeweils den gleichen Zeitraum zwischen 18.12.2017 und 16.12.2018 generiert. Der Aufbau der CSV-Dateien ist identisch. Jede CSV-Datei enthält einen sogenannten Header-Bereich, der aggregierte Informationen und Metadaten zur aktuellen Datei anzeigt, und den eigentlichen Datenbereich, der zwei Spalten (Datum und Seitenaufrufe) zum ausgewählten Zeitraum umfasst. Ein Auszug aus der Export-CSV-Datei für die Seiten-URL http://www.welches-hdmi-kabel.de/hdmi-kabel-unterschiede/ ist in Abbildung 5 dargestellt.

Abbildung 5: Auszug einer Export-Datei (CSV) aus Google Analytics

```
# ----------------------------------------
# Welches HDMI Kabel?
# Aufschlüsselung nach Content
# 20150901-20160831
# ----------------------------------------

Seite,Seitenaufrufe,Einzelne Seitenaufrufe,Durchschn. Besuchszeit auf Seite,Absprungrate,% Ausstiege
/unterschiede/,110.688,100.220,00:06:41,"76,49 %","78,39 %"
,110.688,100.220,00:06:41,"76,49 %","78,39 %"

Index: Tag,Seitenaufrufe
01.09.15,136
02.09.15,159
03.09.15,161
04.09.15,177
05.09.15,160
06.09.15,160
07.09.15,159
```

Um Informationen zu Klicks und Umsätzen zu erhalten, müssen zwei verschiedene Berichte aus dem Amazon PartnerNet generiert werden. Der sogenannte Tracking-ID-Kurzbericht enthält Informationen zur Summe der Klicks, die jeweils über die einzelnen Tracking-IDs des Partners im Berichtszeitraum generiert wurden. Die Tracking-ID-Kurzberichte werden ebenfalls als CSV-Dateien angeliefert und enthalten eine jeweils wöchentliche Zusammenfassung der Klicks über alle Seiten-URLs. Im Fallbeispiel ergibt dies 52 Textdateien für die definierten zwölf Berichtsmonate. Die Textdateien sind identisch aufgebaut und erhalten neben den Angaben zu Klicks auch Informationen zu den in der jeweiligen Woche für die einzelnen Tracking-IDs generierten Bestellungen. Abbildung 6 zeigt einen Auszug für den Tracking-ID-Kurzbericht des Zeitraums 01.01.2018 bis 07.01.2018.

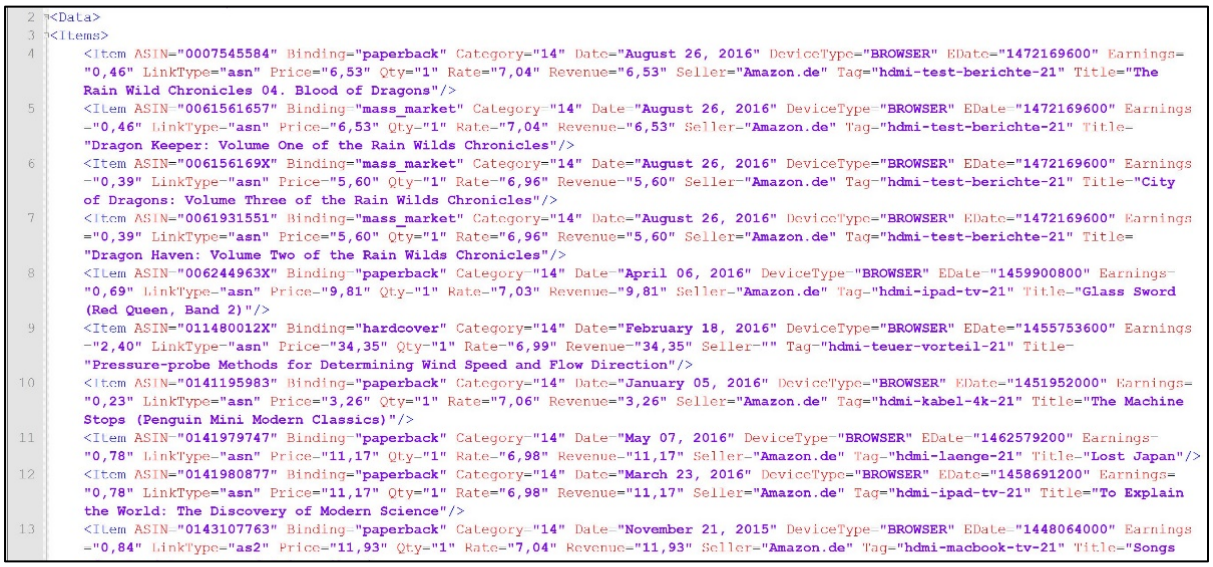
Abbildung 6: Auszug einer Export-Datei (TXT) eines Tracking-ID-Kurzberichts aus Amazon PartnerNet

```
Tracking-ID-Kurzbericht für kaffeequensti-21 1 September 2015 bis 6 September 2015
Tracking-ID Klicks  Anzahl der Bestellungen Anzahl ausgelieferter Artikel   Anzahl der Besucher Umsatz durch
Produkte (EUR )  Werbekostenerstattung (EUR )
3d-hdmi-kabel-21    5   0   0   5   0,00    0,00
airplay-adapter-21  7   0   0   5   0,00    0,00
beliebte-filme-21   21  1   0   10  0,00    0,00
beliebte-hdmi-kabel-21  1   0   0   1   0,00    0,00
dvi-hdmi-adapter-21 3   0   0   3   0,00    0,00
galaxy-an-tv-21 185 16  16  154 223,86  6,76
game-capture-21 40  12  4   28  242,19  11,68
handball-baden-21   1   0   0   1   0,00    0,00
hdmi-1-21   34  11  2   31  21,70   0,65
hdmi-arc-21 75  15  17  71  669,18  26,00
hdmi-cec-21 2   0   0   2   0,00    0,00
hdmi-da-wandler-21  1   0   0   1   0,00    0,00
hdmi-dvi-21 1   0   0   1   0,00    0,00
hdmi-dvi-adapter-21 4   0   0   4   0,00    0,00
```

Die benötigten Informationen zu den vermittelten Umsätzen werden nicht in die Tracking-ID-Kurzberichte exportiert. Hierzu muss ein sogenannter „Werbekosten-Erstattungsbericht“ erstellt werden. Der Werbekosten-Erstattungsbericht wird in Form einer XML-Datei angeliefert, die Informationen zu allen im Berichtszeitraum vermittelten Produkten enthält, inklusive der daraus resultierenden Umsätze und Provisionen (Werbekostenerstattung). Auch sind Informationen zum jeweiligen Bestelldatum und zur jeweiligen Tracking-ID, über die der Umsatz generiert wurde, integriert. Der Werbekosten-Erstattungsbericht für den Berichtszeitraum ist damit der umfangreichste angelieferte Datensatz bezüglich des Fallbeispiels, er umfasst rund 23.800 Zeilen Text mit jeweils

einer eigenen XML-Struktur, die grob in 15 Dimensionen übersetzt werden können. Ein Auszug des Werbekosten-Erstattungsberichts ist in Abbildung 7 dargestellt.

Abbildung 7: Auszug einer Export-Datei (XML) eines Werbekosten-Erstattungsberichts aus Amazon PartnerNet



3.3 Data Preparation

Zur Datenvorbereitung sowie für alle weiteren Datenanalysen wird die frei erhältliche Software R in der Version 3.5.2 in Verbindung mit der Entwicklungsumgebung RStudio verwendet.

Die angelieferten Rohdaten aus Google Analytics und dem Amazon PartnerNet sind unterschiedlich formatiert und stehen zunächst in keinem unmittelbaren Bezug zueinander. Zunächst gilt es hierbei das Entitäts-Identifikations-Problem aufzulösen. Die Zuordnung der beiden Datensets aus dem Amazon PartnerNet ist hierbei noch die leichtere Aufgabe: Zwar liegen beide Datensets in unterschiedlichen Granularitätsstufen vor - Klicks wöchentlich und Umsätze täglich - aber beide lassen sich über das Merkmal “Tracking-ID” einer jeweils einheitlichen Entität zuordnen. Eine Tracking-ID ist ein URL-Parameter, der über einen eingehenden Link mitgegeben wird und über den sich sowohl Klicks als auch Umsätze eindeutig zuordnen lassen. Allerdings sind diese Tracking-IDs in Google Analytics vollständig unbekannt. Die Zählung einzelner Seitenaufrufe erfolgt auf Ebene einzelner Seiten-URLs. Die Aufgabe besteht darin, einzelne Tracking-IDs einzelnen Seiten-URLs zuzuordnen. Um dies valide zu tun müssen zwei Annahmen erfüllt sein:

1. Alle ausgehenden Affiliate-URLs einer Seiten-URL müssen mit dem identischen Tracking-ID Parameter ausgestattet sein.
2. Jede Tracking-ID darf genau einer Seiten-URL zugeordnet sein.

Beide Annahmen sind für das vorliegende Projekt grundsätzlich erfüllt und so lässt sich das Mapping von Seiten-URLs zu Tracking-IDs mithilfe eines Web-Scrapings durchführen. Hierzu wird die Seite des Gesamtprojektes mittels einer entsprechenden Sitemap initialisiert, die alle Unterseiten (Seiten-URLs) der Webseite enthalte. Ein Scraping-Algorithmus arbeitet alle in der Sitemap gelisteten URLs ab, und extrahiert den Tracking-ID-Parameter aus ausgehenden URLs. Dieser wird anschließend in eine Mapping-Tabelle geschrieben, die die Zuordnung zwischen Seiten-URLs und Tracking-IDs herstellt. Mithilfe dieser Mapping-Tabelle können nun Seitenaufrufe für eine bestimmte Seiten-URL aus Google Analytics den Klicks und Umsätzen aus Amazon PartnerNet zugeordnet werden. Der Code für das Web-Scraping und Mapping lautet wie folgt:

```
# 0. Preprocessing - Aufbau der Datensets ----
# Match Website-URLs with Amazon Affiliate Tracking IDs
# Scrape Website and associate URLs with Tracking ID indicated as "tag=..."
in outgoing links.

if (file.exists("data/html_urls.rds")) {

  html_urls <- read_rds("data/html_urls.rds") # (If exists, use this file
and do not execute web scraping again)

} else {

 # Read a sitemap to fetch all URLs a website has for scraping

 sitemap <- read_xml("https://www.welches-hdmi-kabel.de/sitemap.xml")

 # This sitemap is nested so we have to get all sitemaps first
 sitemap_urls <- sitemap %>%
   xml_children() %>%
   xml_text() %>%
   str_replace_all(".xml.*", ".xml")

 sitemap_urls <- sitemap_urls[-1]

 html_urls = tibble("value" = character())

 n = length(sitemap_urls)

 # Iterate through each sitemap and extract the URLs starting with
https:...
 for (i in 1:n ){
   html_urls <- rbind(html_urls,
   sitemap_urls[i] %>%
   read_xml() %>%
   xml_children() %>%
   xml_children() %>%
   xml_text() %>%
   as_tibble() %>%
   filter(str_detect(value, 'https:'))
 )
   }
 write_rds(html_urls, "data/html_urls.rds")
}
```

```r
# Scrape the URLs from the sitemap for Amazon Affiliate Tag. If found,
write into table and link to the resp. URL. Count the number of occurences
of outgoing links per Website.
if (file.exists("data/matched_tags_to_url.rds")) {

  matched_tags_to_url <- read_rds("data/matched_tags_to_url.rds") # Again
don't execute unless file does not exist

} else {

 matched_tags_to_url = tibble("Tag" = character(),
                              "n" = integer(),
                              "URL" = character())

 n = nrow(html_urls)

 j = 1L

 for (i in 1:n) {

   url <-  html_urls[[1]][i]

   parsed_url_links <-
     url %>%
     read_html() %>%
     html_nodes("body") %>%
     html_nodes("a") %>%
     html_attr("href") %>%
     as.tibble()

    if(nrow(parsed_url_links %>% filter(str_detect(value, "tag="))) > 0)
{ #Code will only be executed, if outgoing link including the string "Tag="
is found on the website:
      matched_tags_to_url[j,] <-
       parsed_url_links %>%
       filter(str_detect(value, "tag=")) %>%
       mutate(Tag = str_extract(value, "tag=.*-21")) %>%
       mutate(Tag = str_remove(Tag, "tag=")) %>%
       group_by(Tag) %>%
       count %>%
       mutate(URL = url)
      j <- j+1L
    }

 }
 rm(parsed_url_links, url, j, i, n)

 # In rare cases where an URL contains a parameter which is not tracked by
Amazon EU, but for example Amazon US, NAs will be generated. Filter out
NA's for further analysis.
 matched_tags_to_url <- matched_tags_to_url %>%
   filter(!is.na(Tag))

  write_rds(matched_tags_to_url, "data/matched_tags_to_url.rds")

}
```

Nach Auflösung des Entitäts-Identifikations-Problems müssen noch diverse Schritte zur Datenbereinigung durchgeführt werden. Dies beinhaltet diverse Formatierungen, wie beispielsweise einheitliche Datumsformate, die Aggregation auf eine einheitliche Granularitätsstufe (Kalenderwochen) und die Zusammenführung aller drei Quellen in ein einheitliches Datenset. Einige Datenexporte enthalten Informationen zu unerwünschten Tracking-IDs, beispielsweise Tracking-IDs, die für Testprojekte oder durch externe Partner genutzt wurden. Datensätze mit Informationen zu diesen Tracking-IDs werden aus dem Gesamtdatensatz gelöscht. Des Weiteren sollen nur Umsätze analysiert werden, die nicht retourniert wurden. Entsprechend wurden alle Umsätze, die ein Rücksende-Kennzeichen haben, aus dem Werbekosten-Erstattungs-Datensatz entfernt.

Der vollständige Code für die Datenaufbereitung lautet wie folgt:

```
# 1. Umsätze (Orders) von Amazon----
fee_orders_xml <- read_xml("data/Amazon Orders 18.12.2017 -
16.12.2018/1549029224178-Fee-Orders-7a68a8a6-7342-4f73-a750-3e26a56c5147-
XML.xml")

fee_orders_xml <-
  fee_orders_xml %>%
  xml_children() %>%
  xml_children() %>%
  xml_attrs()

fee_orders_tbl <-
  fee_orders_xml %>%
  as.data.frame() %>%
  t %>%
  as_tibble(row.names=F)

rm(fee_orders_xml)

# Aggregiere Bestellungen auf Tagesebene
fee_orders_daily_tbl <-
  fee_orders_tbl %>%
  mutate(Datum = parse_date_time(Datum, "Ymd HMS")) %>%
  mutate(Datum = as.Date(Datum)) %>%
  mutate(Preis = as.numeric(str_replace(Preis, ",", ".")),
         Menge = as.integer(Menge)) %>%
  mutate(Umsatz = Preis * Menge) %>%
  group_by(Tag, Datum) %>%
  summarise(Umsatz = sum(Umsatz)) %>%
  mutate(CW = get_cw_and_year_from_date(Datum))

fees_weekly_tbl <- fee_orders_daily_tbl %>%
  group_by(Tag, CW) %>%
  summarise(Umsatz = sum(Umsatz))

# 2. Klicks auf Amazon Partnerlinks ----
file.names <- list.files("data/Amazon Klicks 18.12.2017 - 16.12.2018/")
path.names <- paste0("data/Amazon Klicks 18.12.2017 - 16.12.2018/",
file.names)
```

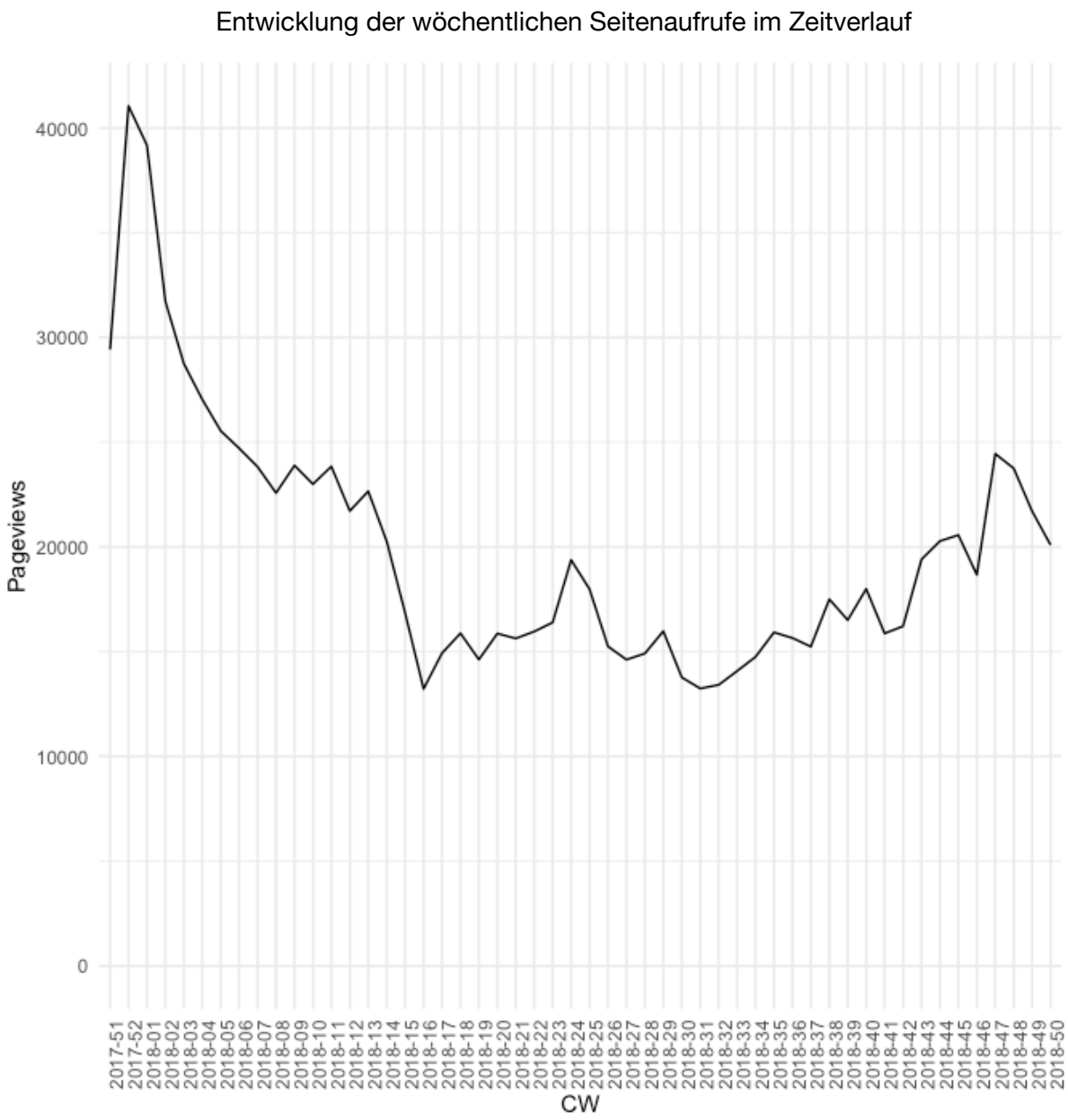
```r
clicks_tbl <- tibble(
  "Tracking-ID" = character(),
  "Klicks" = numeric(),
  "Artikel bestellt" = numeric(),
  "Artikel geliefert" = numeric(),
  "Umsatz (€)" = numeric(),
  "Ad Gebühren (€)"  = numeric()
  )

clicks_tbl <- map_df(path.names, bind_clicks_csv_to_tbl) #Diese Tabelle
enthält noch keine Angaben zum Datum

clicks_date_tbl <- map_df(path.names, bind_click_dates_csv_to_tbl) #Diese
Tabelle enthält nur Datum und Dateiname

clicks_weekly_tbl <-
  clicks_tbl %>%
  left_join(clicks_date_tbl) %>%
  separate(Range, c("Start", "End"), " to ") %>%
  select(-File) %>%
  mutate(Start = mdy(Start),
         End = mdy(End),
         CW = get_cw_and_year_from_date(Start),
         CW_check = get_cw_and_year_from_date(End),
         DaysInWeek = End - Start + 1
         )%>%
  arrange(CW) %>%
  rename(Tag = `Tracking-ID`)

# Sanity Checks: Die einzelnen CSV-Files wurden manuell aus Amazon
Partnernet generiert.Hier wird geprüft, ob Start- und Enddatum des
jeweiligen CSV-Files in der gleichen Kalenderwoche sind und ob
# jede Kalenderwoche aus 7 Einzeltagen besteht.

if (all(clicks_weekly_tbl$CW == clicks_weekly_tbl$CW_check) &
all(clicks_weekly_tbl$DaysInWeek == 7)) {
  clicks_weekly_tbl <-
    clicks_weekly_tbl %>%
    group_by(Tag, CW) %>%
    summarise(Klicks = sum(Klicks))
} else {
  print("Prüfe Rohdaten im Ordner Amazon Klicks")
}

# 3. Pageviews from Google Analytics ----
file.names <- list.files("data/Google Analytics 01.12.2017 - 31.12.2018/")
path.names <- paste0("data/Google Analytics 01.12.2017 - 31.12.2018/",
file.names)

pageviews_tbl <- tibble(
  "Page path level 1" = character(),
  "Date" = character(),
  "Pageviews" = character(),
  "Unique Pageviews" = character(),
  "Avg. Time on Page" = character(),
  "Bounce Rate" = character(),
```

```
  "% Exit" = character()
  )

pageviews_tbl <- map_df(path.names, bind_pageview_csv_to_tbl)
#Parsing Fehler verursacht durch CSV-Format, welches eigentlich 2 Berichte
in einer Datei enthält

#Clean up data frame
pageviews_tbl <- pageviews_tbl %>%
  filter(!is.na(`Page path level 1`)) %>%
  filter(!is.na(Date)) %>%
  mutate(URL = str_remove(`Page path level 1`, "/$")) %>% #Entferne Slashes
am Ende
  filter(str_detect(URL, "^/")) %>%
  mutate(Pageviews = str_remove(Pageviews, ","),
         Pageviews = as.numeric(Pageviews)) %>%
  group_by(URL, Date) %>%
  summarise(Pageviews = sum(Pageviews)) %>%
  ungroup() %>%
  mutate(URL = paste0("https://www.welches-hdmi-kabel.de",URL)) %>%
  mutate(CW = get_cw_and_year_from_date(ymd(Date)))

pageviews_weekly_tbl <-
  pageviews_tbl %>%
  group_by(URL, CW) %>%
  summarize(Pageviews = sum(Pageviews))

# 4. Join Views, Klicks and Revenues into one dataframe ----

# Kleinster Granularitätsfaktor: Kalenderwoche Tag und URL
date_range <- fees_weekly_tbl$CW %>% unique() %>% sort

weekly_raw_tbl <-
  expand(matched_tags_to_url, Tag, date_range ) %>%  #Basis Dataframe mit
jeder Kombination aus Tag und CW in Daterange
  rename(CW = date_range) %>%
  left_join(matched_tags_to_url) %>%
  mutate(URL = str_remove(URL, "/$")) %>%
  left_join(pageviews_weekly_tbl, by=c("URL", "CW")) %>%
  left_join(clicks_weekly_tbl, by=c("CW", "Tag")) %>%
  left_join(fees_weekly_tbl, by=c("CW", "Tag")) %>%
  mutate(Pageviews = replace_na(Pageviews, 0),
         Klicks = replace_na(Klicks, 0),
         Umsatz = replace_na(Umsatz, 0)) %>%
  mutate(Umsatz = replace_na(Umsatz, 0)) %>%
  group_by(Tag, CW) %>%
  summarise(Pageviews = sum(Pageviews),
            Klicks = sum(Klicks),
            Umsatz = sum(Umsatz)) %>%
  ungroup() %>%
  mutate(CTR = ifelse(Pageviews == 0, 0, round(Klicks / Pageviews,4))) %>%
  mutate(RPC = ifelse(Klicks == 0, 0, round(Umsatz / Klicks, 4)))

write_rds(weekly_raw_tbl, "data/weekly_raw_tbl.rds")
```

3.4 Explorative Datenanalyse

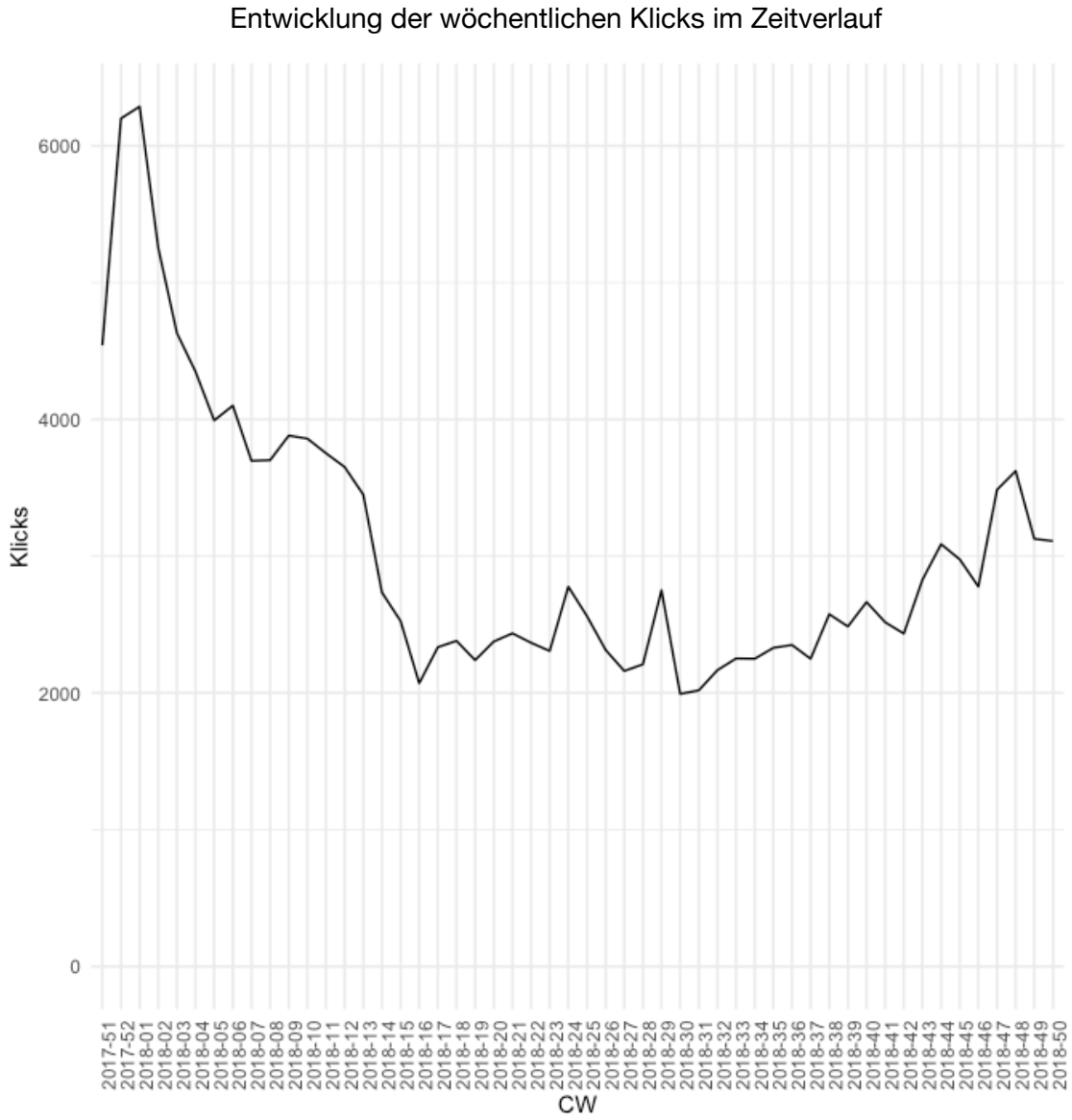
Um ein allgemeines Verständnis für die Beschaffenheit und Verteilung der Daten zu bekommen, soll die allgemeine Struktur- und Häufigkeitsverteilung der Merkmale Seitenaufrufe, Klicks und Umsätze untersucht werden. Die Untersuchungen dienen dazu, Besonderheiten in der Datenverteilung zu identifizieren und diese vor weiteren Analysen ggf. zu bearbeiten.

Zunächst wird die Verteilung der kumulierten wöchentlichen Seitenaufrufe (V), Klicks (C) und Umsätze (R) grafisch untersucht:



Hier ist zu erkennen, dass die Seitenaufrufe im Verlauf des Jahres 2018 zunächst abnahmen, ausgehend von einem Höhepunkt in der KW 52 des Jahres 2017. Ab circa Mitte 2018 stabilisierten sich die Seitenaufrufe wieder und nahmen gegen Ende des Jahres wieder zu. Die hohen

Seitenaufrufe in den Wintermonaten könnten ein Indikator für das in diesem Zeitraum anlaufende Weihnachtsgeschäft sein. Die Weihnachtsfeiertage liegen in der Regel in KW51 - KW52.

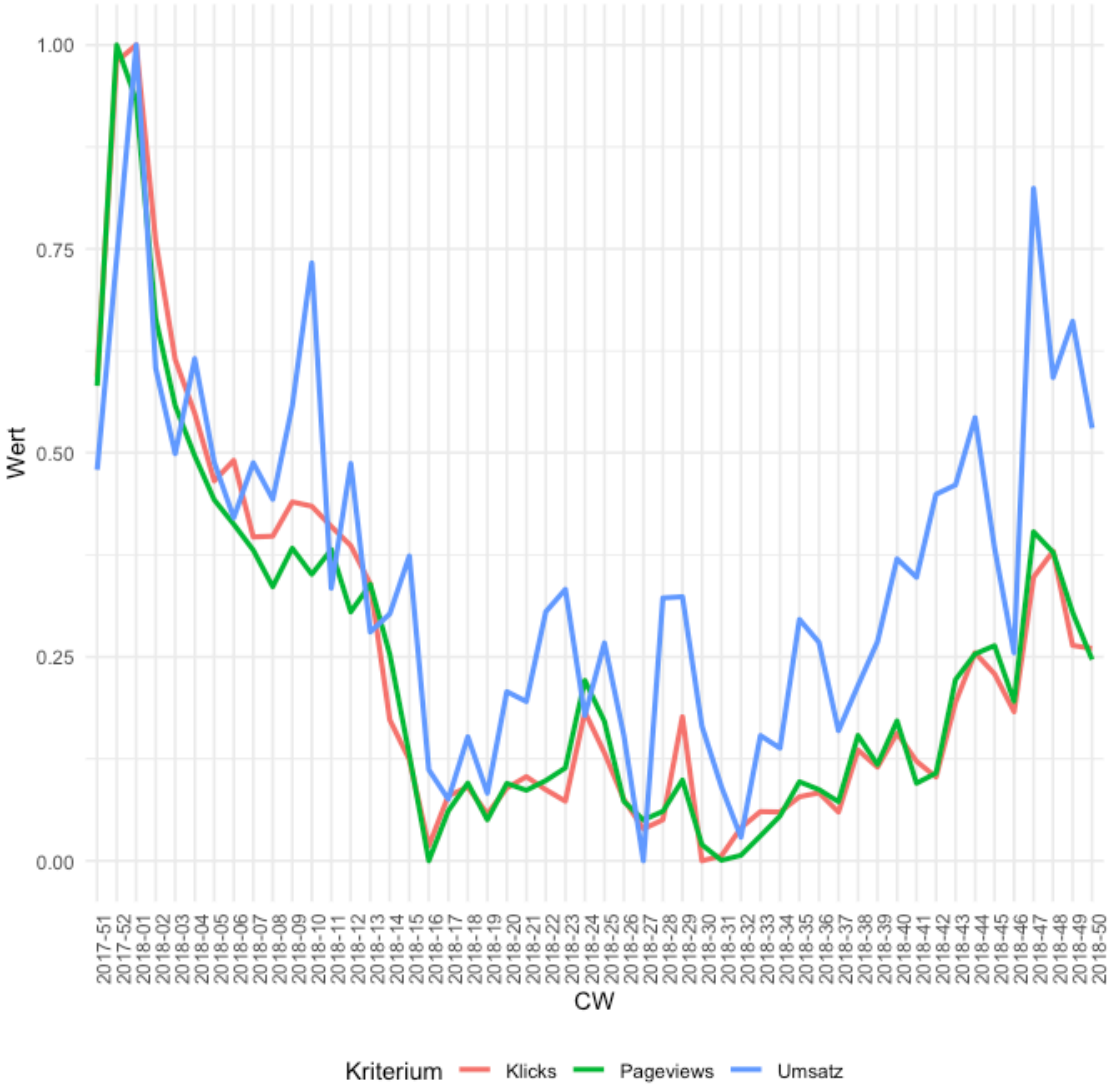


Die Entwicklung der kumulierten Klicks je Kalenderwoche ähnelt der Entwicklung der Seitenaufrufe sehr stark. Hier wird ebenfalls das Maximum in KW 52 erreicht. Die Darstellung ist insofern bezüglich der Abbildung der Seitenaufrufe plausibel, als dass ohne Seitenaufrufe keine Klicks generiert werden können.

Die wöchentliche Umsatzentwicklung im Zeitverlauf zeigt einen wesentlichen Unterschied im Vergleich zu den beiden vorherigen Darstellungen. Zum wird das Maximum nicht in KW 52, sondern in KW 1 erreicht, zum anderen gibt es einen zweiten Ausreißer nach oben, der in KW 47 stattfindet.

Zur genaueren Untersuchung werden Seitenaufrufe, Klicks und Umsätze in einem Chart dargestellt:

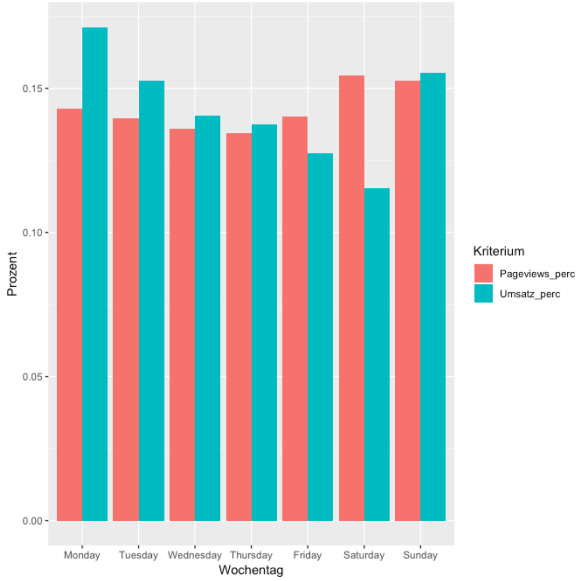
Entwicklung der wöchentlichen Seitenaufrufe, Klicks und Umsätze im Zeitverlauf



Aus dieser Betrachtung ergeben sich zwei zentrale Erkenntnisse: Einerseits korrelieren alle drei Werte miteinander. Seitenaufrufe und Klicks korrelieren sehr stark, der Umsatz im Vergleich dazu weniger. Die Umsatzkurze zeigt immer wieder Ausreißer nach oben, die in den Seitenaufrufen und Klicks nicht auftauchen.

Der Grund hierfür liegt vor allem in der Berechnungssystematik der Umsätze durch den Merchant Amazon. Als Umsätze werden grundsätzlich nur Waren gezählt, die tatsächlich an den Besteller verschickt wurden. An Sonntagen findet generell (mit der Ausnahme von digitalen Gütern) kein Versand statt. Die Abweichung zwischen Bestell- (Klick) und Versanddatum (Umsatz) kann somit durchaus ein bis zwei Tage umfassen und damit in unterschiedliche Kalenderwochen fallen, was in der folgenden Abbildung deutlich wird.

Analyse der Umsätze und Seitenaufrufe für Wochentage

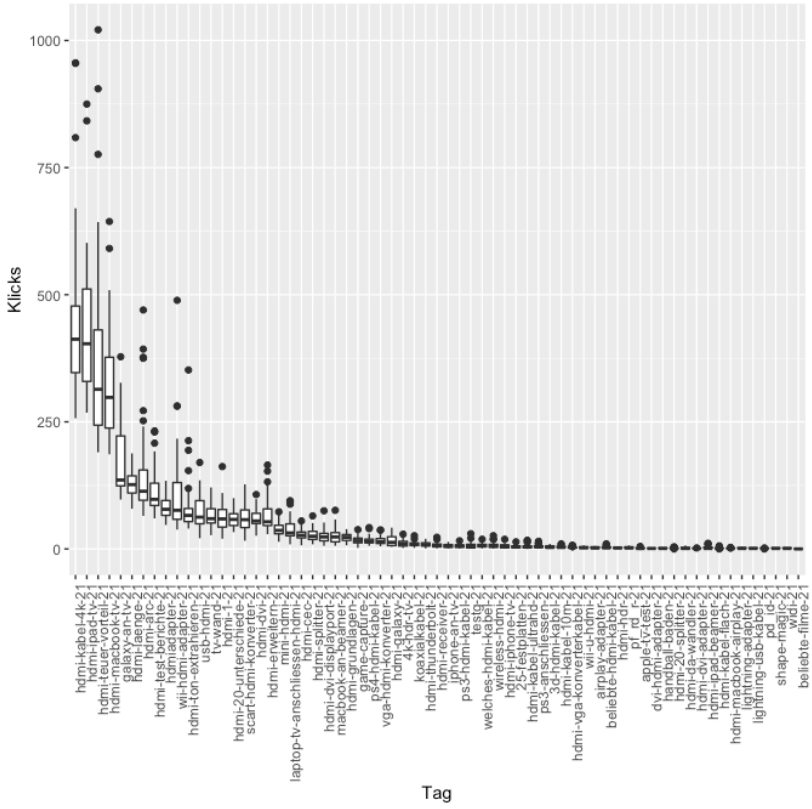


Die Abbildung zeigt, dass Umsätze am stärksten Montags generiert werden, wohingegen Seitenaufrufe vermehrt am Wochenende stattfinden. Der hohe Umsatz an Montagen ist jedoch vor allem auch auf die Bestellungen an Wochenende zurückzuführen, die Montags versandt werden.

3.5 Analyse der Klickraten

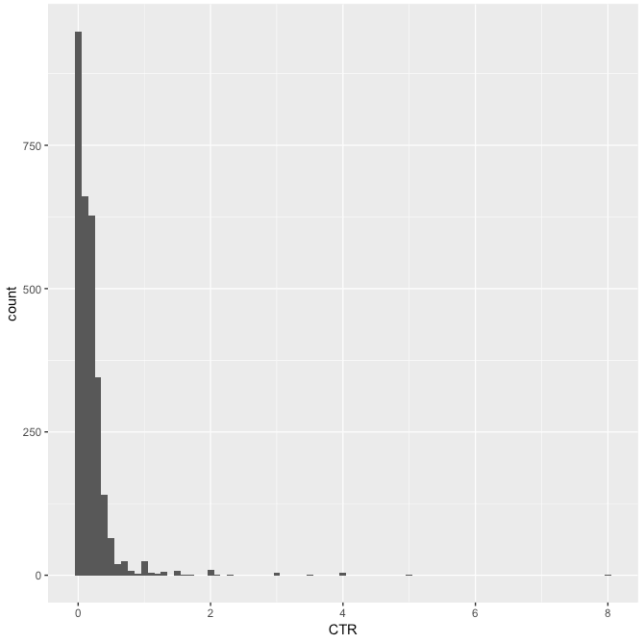
3.5.1 Analyse der CTR-Verteilung

Zur Analyse der ersten Konversionsebene soll das Klickverhalten in Bezug auf die Seitenaufrufe untersucht werden (Click-Through-Rate). Hierzu ist es zunächst hilfreich, ein Verständnis davon zu entwickeln wie sich Klicks nicht nur im Zeitverlauf, sondern auch über einzelne Seiten-URLs verteilen. Dabei hilft folgender Boxplot:

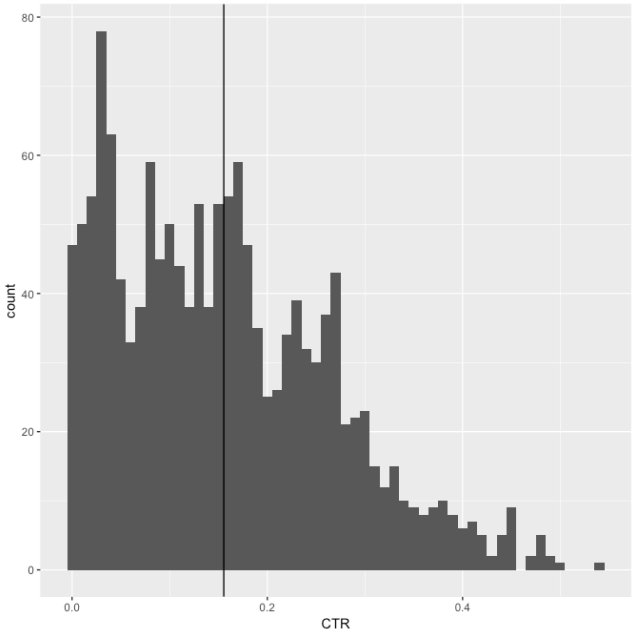


Durch diese Grafik wird deutlich, dass ein paar wenige Seiten-URLs die meisten Klicks absolut generieren. Dementsprechend weist die Verteilung einen Tail auf, der viele URLs umfasst, aber in Summe wenig Klicks generiert. Diese Erkenntnis ist für die spätere Benchmark-Modellierung relevant und muss entsprechend berücksichtigt werden.

Um weitere Aussagen über die Performance der CTR treffen zu können, muss zunächst analysiert werden, inwiefern überhaupt eine Varianz innerhalb der Daten vorliegt.

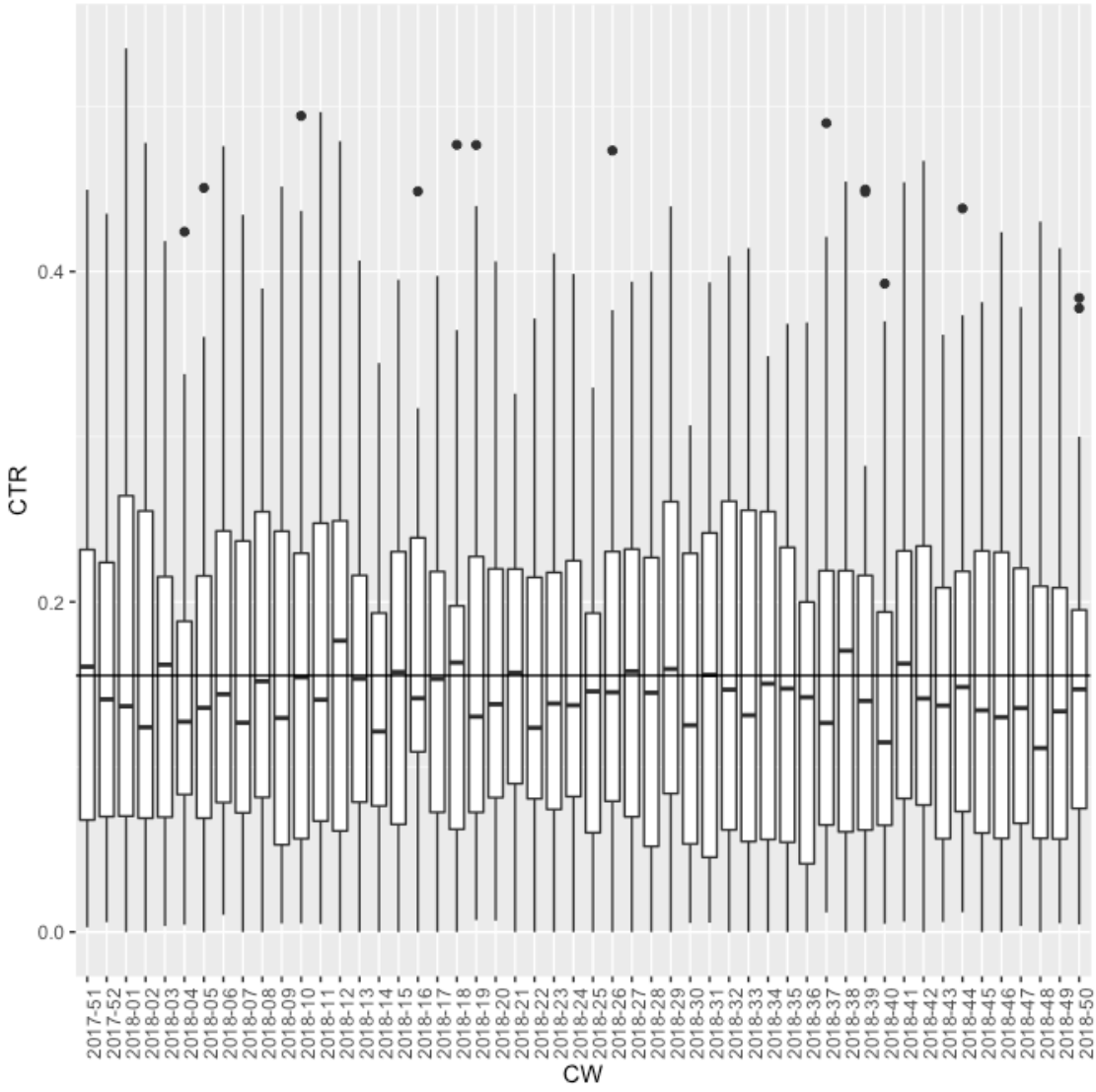


Betrachtet man diese ungefilterte Darstellung werden vor allem zwei Dinge deutlich: Erstens gibt es sehr viele Beobachtungen, bei denen die CTR 0 ist. Andererseits gibt einige extreme Ausreißer nach oben mit bis zu 8 Klicks pro Seitenaufruf. Beide Extremfälle sind für die Ableitung von Benchmarks wenig hilfreich. Daher wird die Verteilung unter der Bedingung analysiert, dass eine Seiten-URL innerhalb einer Woche mindestens 10 Aufrufe pro Tag, das heißt 70 Aufrufe pro Woche erreicht hat. Unterhalb dieser Grenze sind keine verlässlichen Aussagen zum Klickverhalten möglich.



Berücksichtigt man diese Bedingung, ergibt sich ein etwas aussagekräftigeres Bild. Zum einen existiert eine relativ starke Varianz, das heißt die CTR streut relativ breit um den Mittelwert 0,16. Offensichtlich ist der Mittelwert zudem kein optimales Lagemaß, da die Daten nicht normal verteilt sind.

Schwankt die CTR in Abhängigkeit von der Zeit? Diese Frage beantwortet folgende Abbildung:

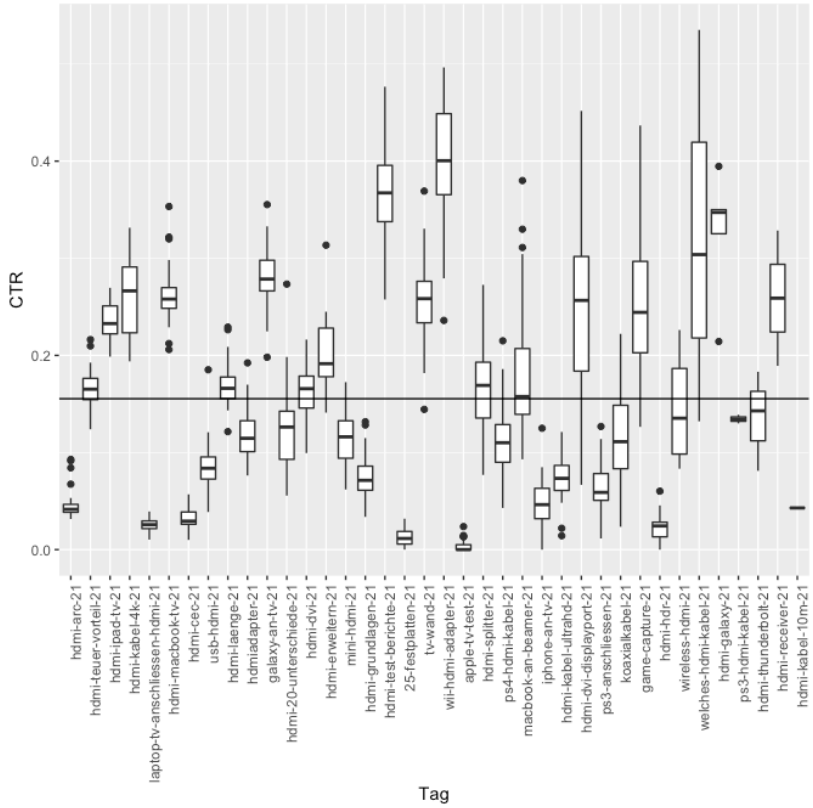


Das Ergebnis: Der CTR-Mittelwert pro Kalenderwoche ist über den Zeitverlauf gesehen relativ konstant. Hier scheint kein belastbares Muster vorzuliegen.

Schwankt die CTR nach Seiten-URL, performen also manche Seiten besser als andere?

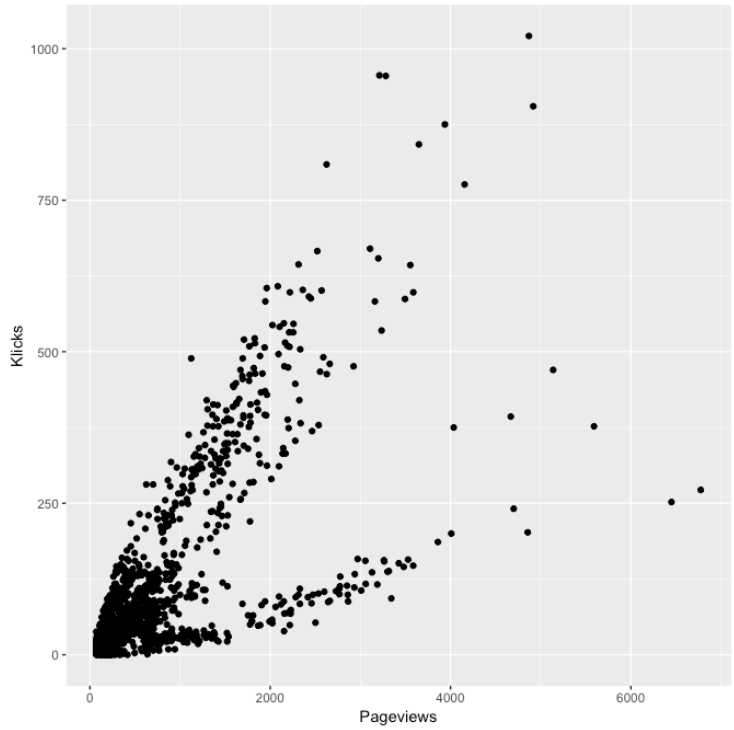
Die Erkenntnis: Ja, die CTR variiert sehr stark nach Seiten-URL. Der geringste CTR-Median pro Seiten-URL liegt bei 0,01 und der höchste bei 0,4. Die Sortierung der Seiten-URLs erfolgt hier

anhand der jeweiligen absoluten Seitenaufrufe in absteigender Reihenfolge. Zwei der Top-5-URLs weisen sehr schwache CTR-Werte auf.



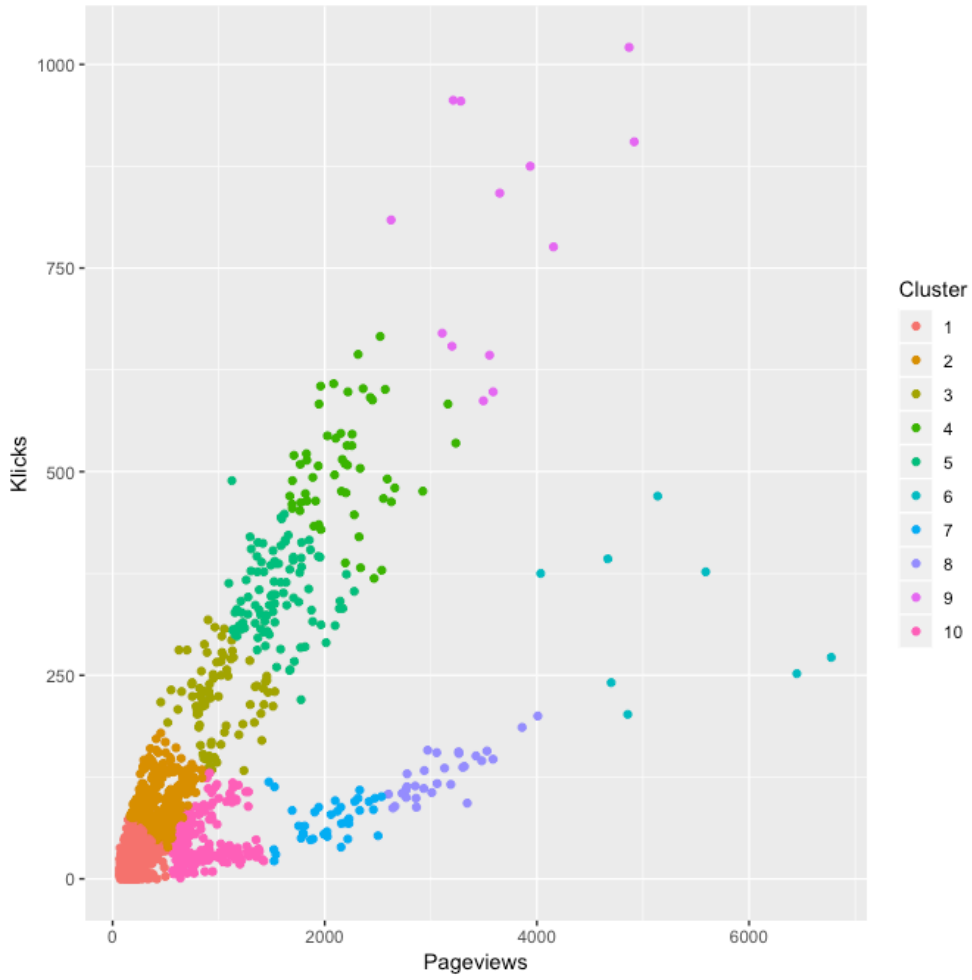
Abschließen soll betrachtet werden, inwiefern die CTR im Vergleich zu den absoluten Seitenaufrufen schwankt. Bleibt die CTR auch bei steigenden Seitenaufrufen konstant? Zur Verdeutlichung wird der direkte Zusammenhang zwischen Klicks und Seitenaufrufen visualisiert.

Verhältnis von Seitenaufrufen zu Klicks, absolut"



Bliebe die CTR jeweils konstant, müsste die Verteilung durch eine einfache Regressionsgerade modelliert werden können. Die Erkenntnis: Grundsätzlich existiert tatsächlich ein linearer Zusammenhang zwischen den Variablen, je mehr Pageviews, desto mehr Klicks. Allerdings scheinen hier mindestens 2 Gruppen von Daten zu existieren, einige mit besserer und einige mit schlechterer Performance.

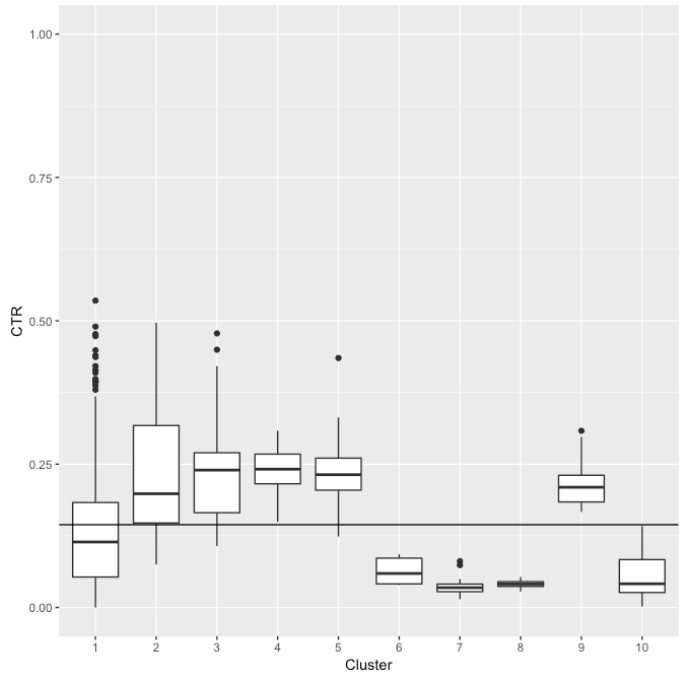
Cluster-Analyse: Auflösung der zwei Sub-Gruppen



Das Ziel der Analyse ist es, einen geeigneten Benchmark für die Kennzahl CTR abzuleiten. Aus diesem Grund soll die Datengruppe mit schlechterer Performance gezielt ausgeschlossen werden. Um diese Daten mit schlechter Performance zu markieren, soll eine Clusteranalyse durchgeführt werden.

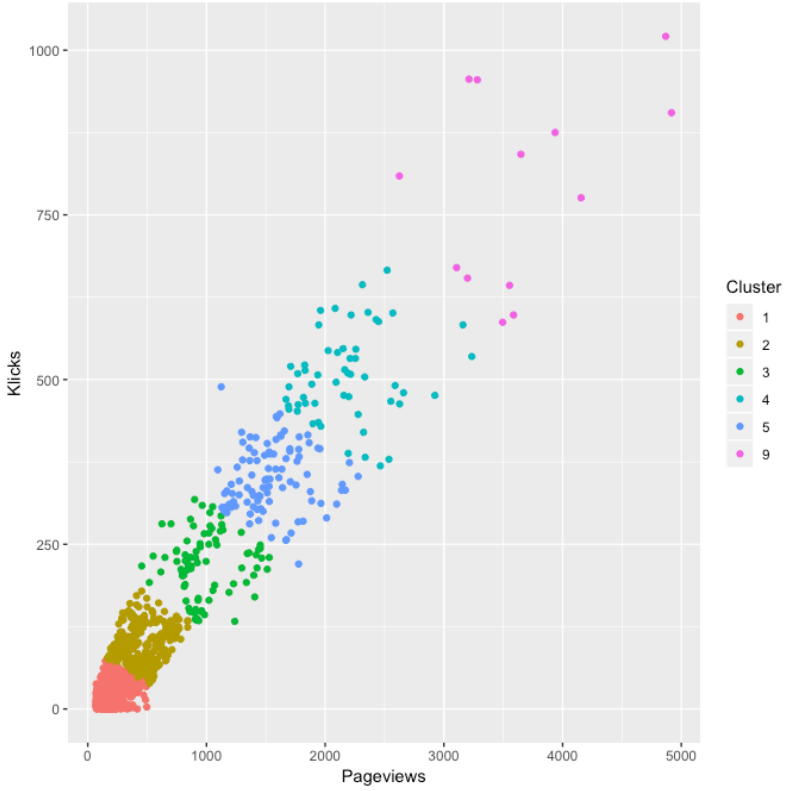
Als Clusteralgorithmus wird k-Means verwendet. Als Zielclusterannahme wird k = 10 definiert, was sich nach einigen Versuchen als guter Kompromiss zwischen Clusteranzahl und Clusterähnlichkeit herausstellte. Im Folgenden wird die durchschnittliche CTR pro Cluster berechnet und die Cluster mit den jeweils geringsten CTR-Medianen als Low-Performer markiert.

CTR Verteilung über die Cluster

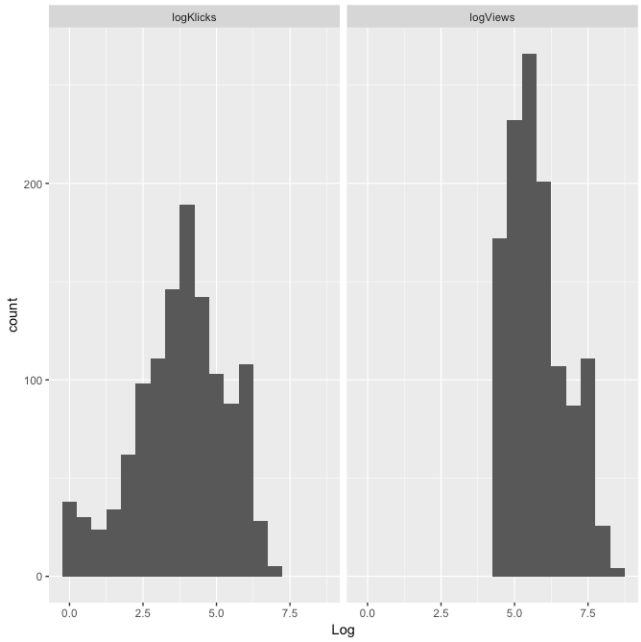


Es sollen nur solche Cluster in die Benchmarkmodellierung einfließen, deren Cluster-Median mindestens die Hälfte des gesamten Medians beträgt. Dies betrifft die Cluster 1, 2, 3, 4, 5 und 9. Beschränkt man die Verteilung von Seitenaufrufen zu Klicks nun auf die verbliebenen Cluster, ergibt sich folgendes Bild:

Verhältnis von Seitenaufrufen zu Klicks, absolut, nach Bereinigung



Diese Verteilung lässt sich nun gut mittels einer Regression modellieren, wobei Seitenaufrufe (Pageviews) als Input- und Klicks als Output-Variable gelten. Da die Variablen Klicks und Pageviews wie oben gezeigt nicht normalverteilt sind, muss diese Annahme zunächst hergestellt werden. In dem vorliegenden Fall geschieht dies über eine Log-Transformation Beide Variablen nähern sich nun deutlich besser einer Normalverteilung an. Die abgeschnitten wirkende Verteilung der logViews resultiert aus der Vorbedingung, dass in der Modellierung nur Seiten-URLs mit mindestens 70 Aufrufen pro Woche betrachtet wurden.



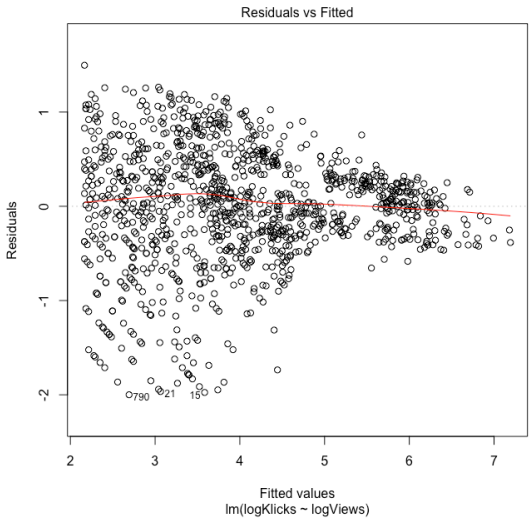
3.5.2 Regressionsmodellierung zur Vorhersage der Klicks in Abhängigkeit von Seitenaufrufen

```
#Temporäres Modell zur Ausreißererkennung in den Reisduen
clicks_lm_1 <- lm(logKlicks ~ logViews, data = ctr_modelling_df)
summary(clicks_lm_1)

##
## Call:
## lm(formula = logKlicks ~ logViews, data = ctr_modelling_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1014 -0.2802  0.1277  0.5632  1.8317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.6869     0.1595  -23.11   <2e-16 ***
## logViews      1.2904     0.0272   47.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9146 on 1204 degrees of freedom
## Multiple R-squared:  0.6515, Adjusted R-squared:  0.6513
## F-statistic:  2251 on 1 and 1204 DF,  p-value: < 2.2e-16
```

Das erste Modell trifft die Daten mit einem R^2 Wert von 0.65 noch nicht besonders gut. Eine Ausreißeranalyse auf den Residuen soll helfen, dass Modell weiter zu verbessern. Als Algorithmus wird ein Hampel-Test verwendet.

```
#Ausreißer markieren
ctr_modelling_outliers <- hampel.proc(clicks_lm_1$residuals)

#Neues Modell ohne Ausreißer
clicks_lm_2 <- lm(logKlicks ~ logViews, data=ctr_modelling_df[-
ctr_modelling_outliers,])
print(summary(clicks_lm_2))

##
## Call:
## lm(formula = logKlicks ~ logViews, data = ctr_modelling_df[-
ctr_modelling_outliers,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99961 -0.32240  0.03037  0.38468  1.49655
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.92191    0.10789  -27.08   <2e-16 ***
## logViews     1.18993    0.01827   65.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6054 on 1126 degrees of freedom
## Multiple R-squared:  0.7902, Adjusted R-squared:  0.7901
## F-statistic:  4242 on 1 and 1126 DF,  p-value: < 2.2e-16
```

Das neue Modell nach Ausreißeranalyse trifft die Daten mit einem R^2 von 0.79 noch etwas besser. Die Analyse der Residuen soll zeigen, ob diese auch normalverteilt und das Modell damit akzeptabel ist:



Das Ergebnis: Das Modell trifft bei besonders kleinen prognostizierten Klicks die Daten noch nicht besonders gut. Dies dürfte darin begründet sein, dass mit der initialen Annahme von 70 Seitenaufrufen, eine Menge Zugriffe mit sehr wenig Traffic herausgefiltert wurde, aber gleichzeitig immer noch sehr viele Beobachtungen auftauchen, die wenig bis gar keine Klicks zeigen.

In den Performance-Cluster beträgt der Median-CTR, dass im Durchschnitt pro Seitenaufruf mindestens 0.17 Klicks generiert werden sollten, bedeutet das für 70 Seitenaufrufe pro Woche mindestens 12 Klicks absolut. Das Modellierungs-Datenset soll daher auch auf diese Anzahl von Minimum-Klicks beschränkt werden:

```
#Log-Transformation der Predictors zur besseren Modellierung:
ctr_modelling_df <-
  ctr_df %>%
  filter(Cluster %in% clusters_perform & Klicks >=
round(median_ctr_performance * 70,0) ) %>%
  mutate(logKlicks = log1p(Klicks),
         logViews = log1p(Pageviews))

p <- ctr_modelling_df %>%
  select(logKlicks, logViews) %>%
  gather("Variable", "Log") %>%
  ggplot(aes(x=Log)) +
  geom_histogram(binwidth = .3) +
  facet_grid(.~Variable)

print(p)

#Temporäres Modell zur Ausreißererkennung in den Reisduen
clicks_lm_3 <- lm(logKlicks ~ logViews, data = ctr_modelling_df)

#Ausreißer markieren
ctr_modelling_outliers <- hampel.proc(clicks_lm_3$residuals)

#Neues Modell ohne Ausreißer
clicks_lm_4 <- lm(logKlicks ~ logViews, data=ctr_modelling_df[-
ctr_modelling_outliers,])
print(summary(clicks_lm_4))

##
## Call:
## lm(formula = logKlicks ~ logViews, data = ctr_modelling_df[-
ctr_modelling_outliers,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32355 -0.28588  0.00706  0.31653  1.05863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.75152    0.09218  -19.00   <2e-16 ***
## logViews     1.01866    0.01518   67.11   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4463 on 964 degrees of freedom
## Multiple R-squared:  0.8237, Adjusted R-squared:  0.8235
## F-statistic:  4504 on 1 and 964 DF,  p-value: < 2.2e-16
```

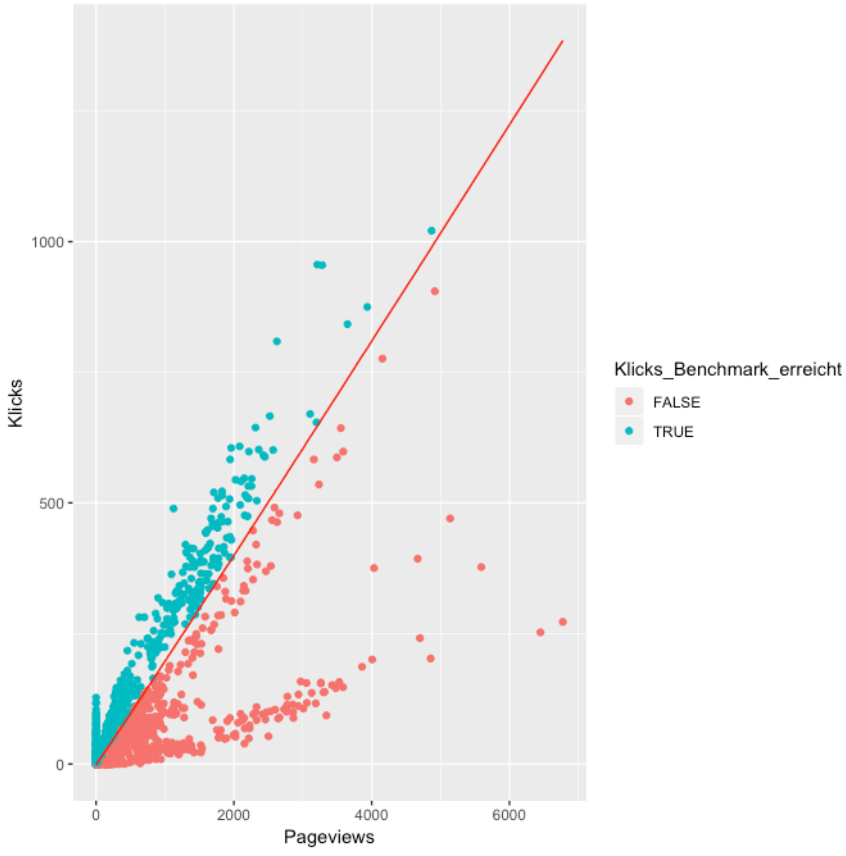
```
plot(clicks_lm_4, which=1)
```

Das Modell ClicksLM_4 verbessert sich durch diesen zusätzlichen Constraint auf einen R^2 von 0.82 und weist gleichzeitig eine besser normalverteile Darstellung der Residuen auf. Der hohe Wert des Shapiro-Wilk-Test auf Normalverteilung der Residuen bestätigt diese Annahme:

```
shapiro.test(clicks_lm_4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  clicks_lm_4$residuals
## W = 0.99253, p-value = 8.721e-05
```

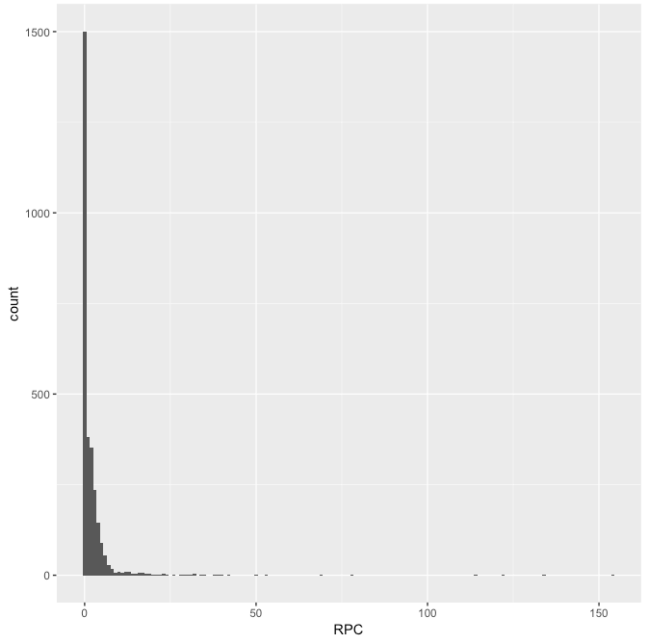
Das Modell Clicks_LM_4 lässt sich nun dafür nutzen, um die Benchmark-Klicks im gesamten Datensatz, das heißt auch für alle Seitenaufrufe unter 70 zu verwenden. Des weiteren wird eine zusätzliche Spalte angefügt, die angibt ob der Benchmark jeweils erreicht wurde oder nicht. Diese Benchmark-Erreichung lässt sich auf grafisch im Gesamtdatensatz visualisieren, wobei die Linie die Regressionsgerade (=CTR-Benchmark) ist und die Farbe die Erreichung des Benchmarks markiert:

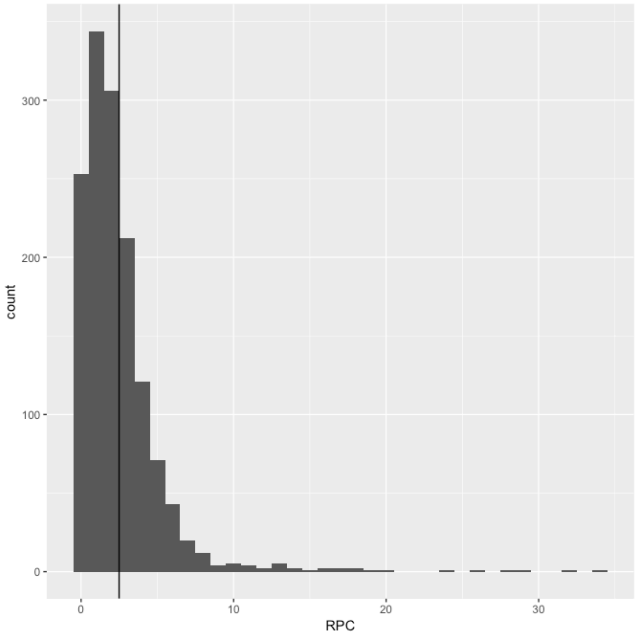


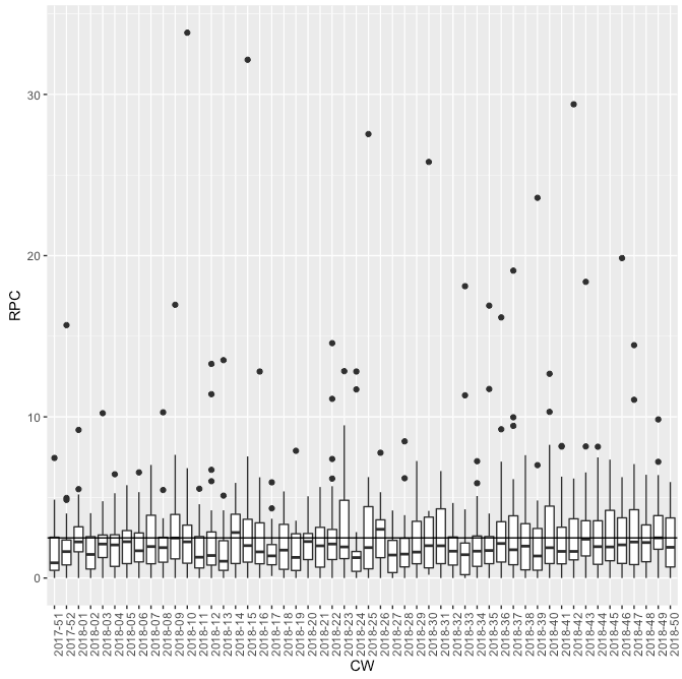
3.5.3 Analyse der RPC-Verteilung

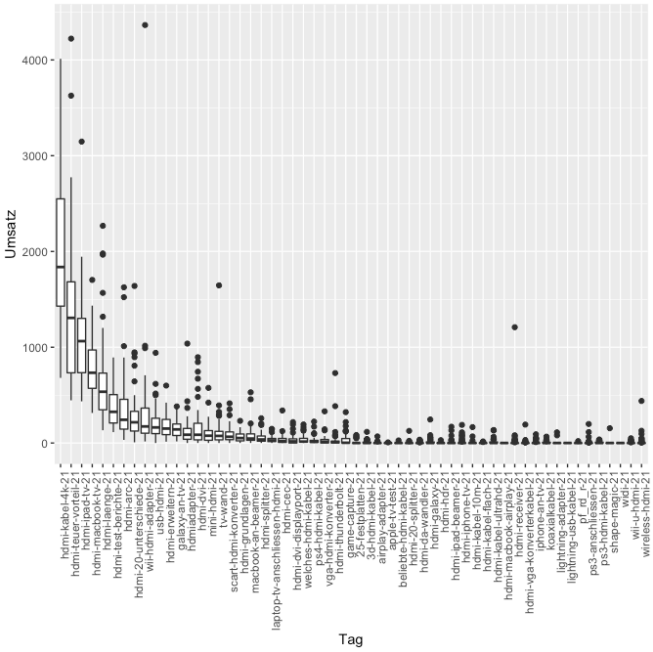
In der zweiten Konversionsebene soll die Kenngröße Umsatz in Bezug auf das Klickverhalten analysiert werden, um so nähere Erkenntnisse über die Benchmarkgröße RPC (Revenue-Per-Click) zu gewinnen.

Da die Abfolge der Analyse im Wesentlichen identisch zur CTR-Betrachtung ist, werden die Auswertungen etwas kompakter zusammengefasst. Auch für die Kenngröße RPC ist es zunächst hilfreich, die Verteilung der Umsätze über den gesamten Datensatz zu betrachten. Die folgenden Abbildungen zeigen die Verteilung der Kennzahl RPC allgemein, im Zeitverlauf und über einzelne Seiten-URLs.



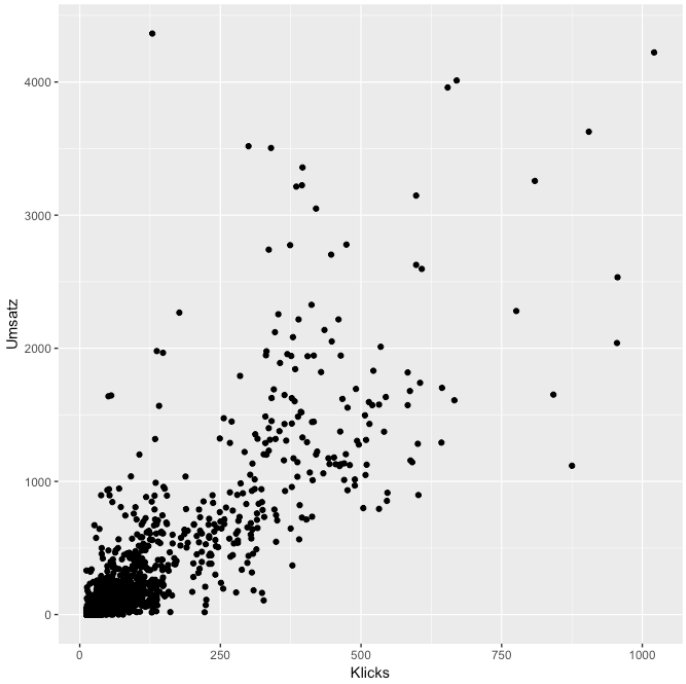






Auch hier lässt sich erkennen, dass nur eine wenige Anzahl von Seiten-URLs die meisten Umsätze generiert. Entsprechendes schlägt sich auch in der Betrachtung der Verteilung der RPC im gesamten Datenset nieder. Auch hier soll wieder ein Mindestwert definiert werden, in diesem Fall analog zur vorherigen Betrachtung die Baseline von 12 Klicks pro Woche. Ähnlich wie bereits bei der CTR-Analyse wird deutlich, dass die Verteilung der Umsätze stärker über die einzelnen Seiten-URLs schwankt als über die Zeit, was bedeutet, dass die Seiten-URL ein größerer Einflussfaktor auf die Performance des RPC-Benchmarks zu sein scheint. Gleichzeitig gibt es auch mehr Ausreißer, sowohl über Seiten-URLs als auch über die Zeitdimension verteilt.

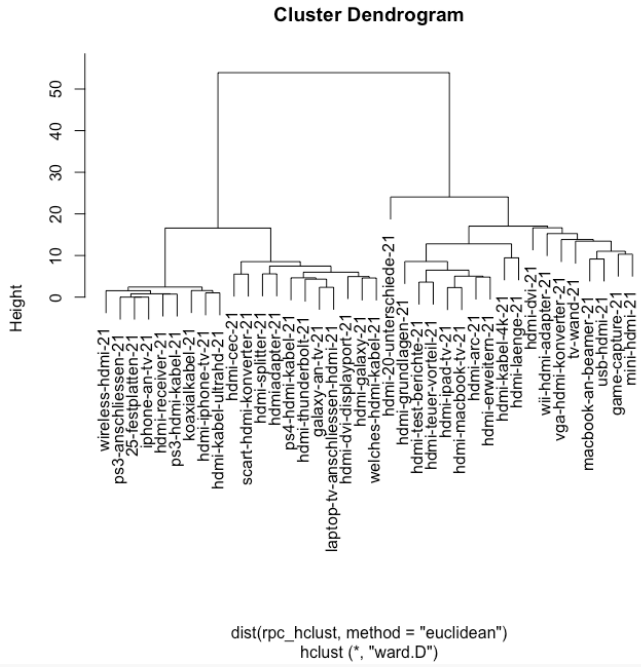
Verhältnis von Klicks zu Umsätzen, absolut



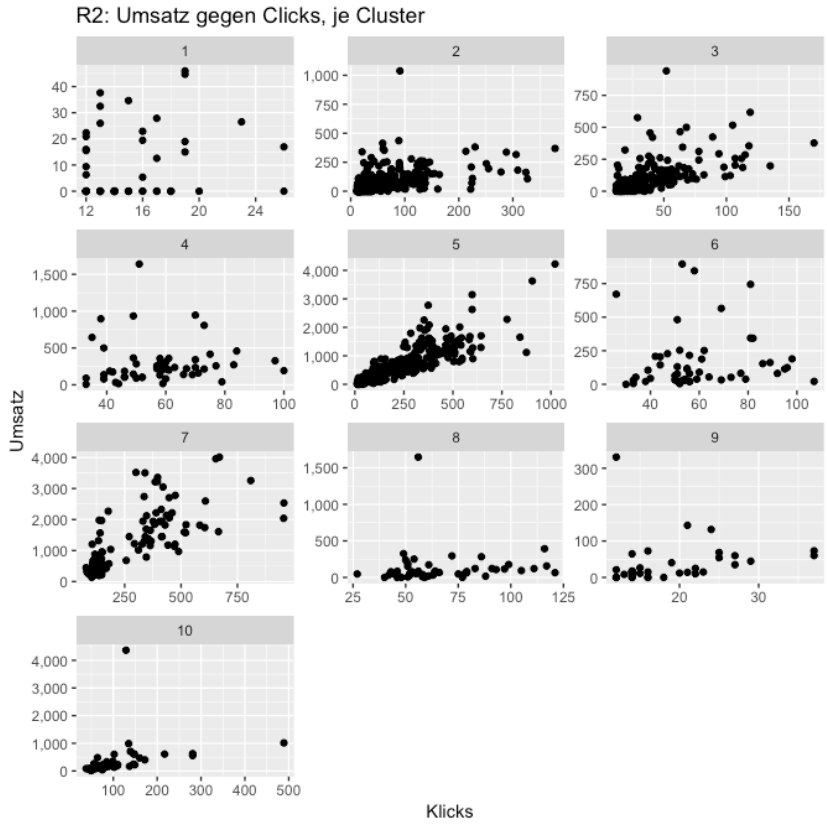
Die Erkenntnis: Aufgrund der vielen Ausreißer, ist der lineare Zusammenhang hier nicht so offensichtlich wie bei der Analyse der Seitenaufrufe und Klicks. Zur Ableitung der Benchmarks ist also die entsprechende Bereinigung der Daten entscheidend. Auch hier sollen zunächst mithilfe einer Clusteranalyse gut performende und schlecht performende Punkte getrennt werden. An dieser Stelle wird ein hierarchischer Clusteralgorithmus verwendet, um die Strukturen innerhalb der Daten besser erkenntlich zu machen.

```
rpc_hclust <- weekly_raw_tbl %>%
  filter(Klicks >= weekly_clicks_min) %>%
  select(Tag, RPC, Klicks, Umsatz, CW) %>%
  mutate(RPC = scale(RPC, center = TRUE, scale = TRUE) ) %>%
  select(-Klicks, -Umsatz) %>%
  spread(Tag, RPC) %>%
  select(-CW)

#Replace missing with minimal value
```

```
rpc_hclust <-
  rpc_hclust %>%
  mutate_all(funs(replace(., is.na(.), min(rpc_hclust, na.rm = T)))) %>%
  t()

rpc_hclust <- hclust(dist(rpc_hclust, method = "euclidean"),
method="ward.D")
plot(rpc_hclust)
```

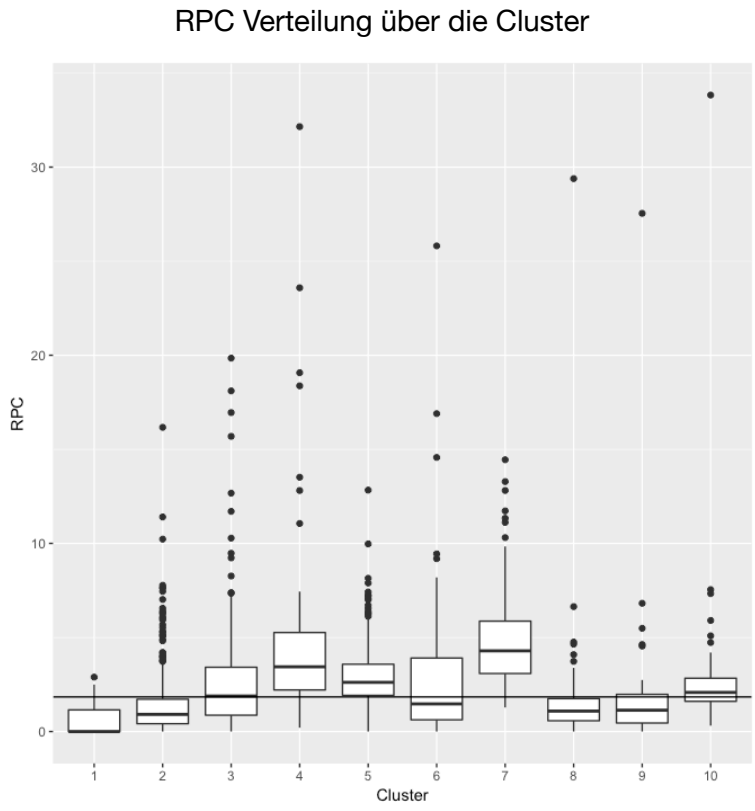



```
rpc_hclust_assignment <- cutree(rpc_hclust, h=12) %>%
  as.list() %>%
  as_tibble() %>%
  t()

rpc_hclust_assignment = data.frame("Tag" = rownames(rpc_hclust_assignment),
                                   "Cluster" = rpc_hclust_assignment[,1])

rpc_df <- weekly_raw_tbl %>%
  select(Tag, CW, Umsatz, Klicks, RPC) %>%
  filter(Klicks >= weekly_clicks_min) %>%
  left_join(rpc_hclust_assignment, by="Tag") %>%
  mutate(Cluster = as.factor(Cluster))

#Clicks vs. Views in each cluster
 rpc_df %>%
  ggplot(aes(Klicks, Umsatz)) +
  geom_point() +
  scale_y_continuous(labels = comma) +
  ggtitle("R2: Umsatz gegen Clicks, je Cluster") +
  ylab("Umsatz") +
  xlab("Klicks") +
  facet_wrap(~Cluster, ncol=3, scales = "free")
```




```
rpc_df %>%
  ggplot(aes(x=Klicks, y=Umsatz, color=Cluster)) +
  geom_point()
```

Nun werden die RPC-Mediane der einzelnen Cluster analysiert und High-Performer markiert. Nach Einschränkung auf die High Performance Cluster, die hier durch einen Median von mindesten 70% des Gesamt-Medians definiert werden, bleiben folgende Cluster übrig: Cluster 3, 4, 5, 7 und 10.

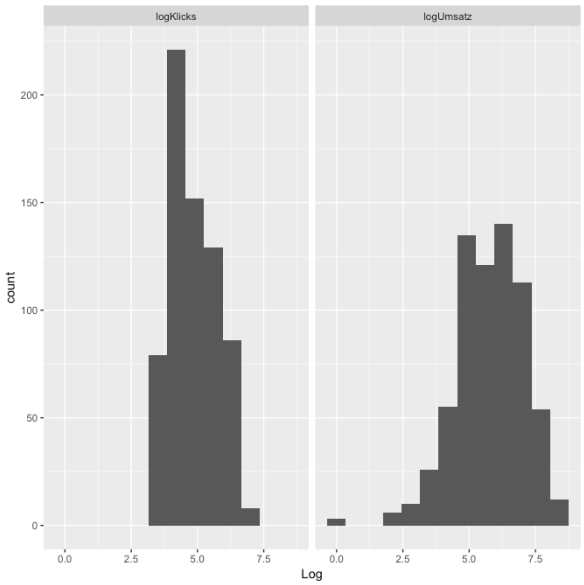


Verhältnis von Klicks zu Umsatz, absolut, vor (links) nach (rechts) Bereinigung

3.5.4 Regressionsmodellierung zur Vorhersage der Klicks in Abhängigkeit von Seitenaufrufen

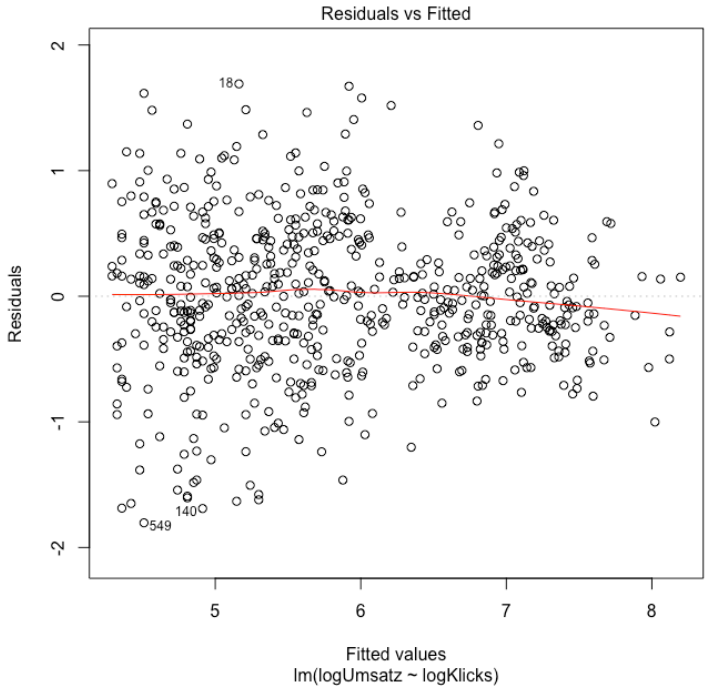
Der Aufbau des entsprechenden Regressionsmodells erfolgt analog zur CTR-Analyse per Skript:

```
#Log-Transformation der Predictors zur besseren Modellierung:
rpc_modelling_df <-
  rpc_df %>%
  filter(Cluster %in% clusters_perform_rpc & Klicks >=
round(median_rpc_performance * weekly_clicks_min,1) ) %>%
  mutate(logKlicks = log1p(Klicks),
         logUmsatz = log1p(Umsatz))

p <- rpc_modelling_df %>%
  select(logKlicks, logUmsatz) %>%
  gather("Variable", "Log") %>%
  ggplot(aes(x=Log)) +
  geom_histogram(binwidth = .7) +
  facet_grid(.~Variable)

print(p)
```




```
#Temporäres Modell zur Ausreißererkennung in den Residuen
revenue_lm_1 <- lm(logUmsatz ~ logKlicks, data = rpc_modelling_df)

#Ausreißer markieren
rpc_modelling_outliers <- hampel.proc(revenue_lm_1$residuals)

#Neues Modell ohne Ausreißer
revenue_lm_2 <- lm(logUmsatz ~ logKlicks, data=rpc_modelling_df[-
rpc_modelling_outliers,])

print(summary(revenue_lm_2))

##
## Call:
## lm(formula = logUmsatz ~ logKlicks, data = rpc_modelling_df[-
rpc_modelling_outliers,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80362 -0.36114 -0.00501  0.40160  1.69053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3142     0.1343    2.34   0.0196 *
## logKlicks     1.1375     0.0272   41.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5931 on 648 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.7292
## F-statistic:  1749 on 1 and 648 DF,  p-value: < 2.2e-16

plot(revenue_lm_2, which=1)
```

Das Modell Revenue_LM_2 erreicht einen R^2 von 0.73 und weist gleichzeitig eine akzeptable Residuenverteilung auf.



Der Shapiro-Wilk-Test auf Normalverteilung unterstreicht die Normalverteilungsannahme ebenfalls:

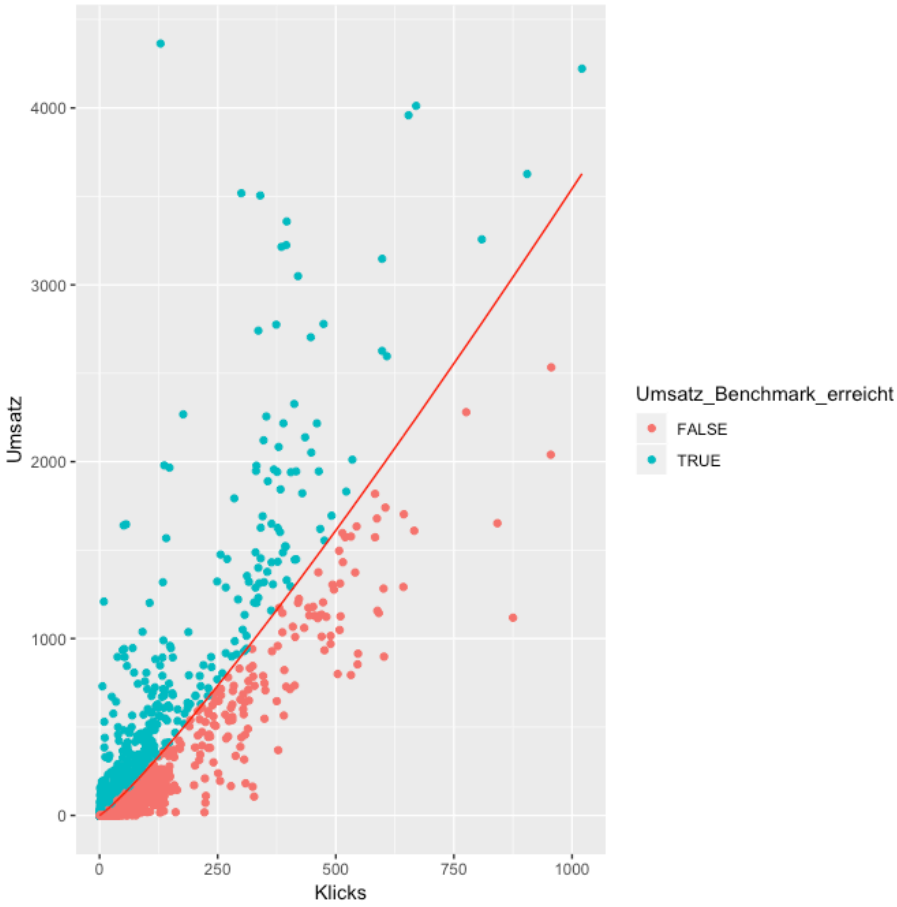
```
shapiro.test(revenue_lm_2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  revenue_lm_2$residuals
## W = 0.99345, p-value = 0.006239
```

Das Modell Revenue_LM_2 wird nun dafür genutzt, um die Benchmark-Umsätze im gesamten Datensatz zu berechnen:

```
## # A tibble: 10 x 5
##    Tag                CW      Umsatz Umsatz_Benchmark
Umsatz_Benchmark_erre…
##    <chr>              <chr>    <dbl>            <dbl> <lgl>
##  1 hdmi-kabel-4k-21   2017-51 1610.             2232 FALSE
##  2 hdmi-ipad-tv-21    2017-51 1283.             1986 FALSE
##  3 hdmi-teuer-vorte… 2017-51 2627.             1975 TRUE
##  4 hdmi-macbook-tv-… 2017-51 1129.             1481 FALSE
##  5 wii-hdmi-adapter… 2017-51  557.              838 FALSE
##  6 galaxy-an-tv-21    2017-51   71.2            648 FALSE
##  7 hdmi-arc-21        2017-51  282.             576 FALSE
##  8 hdmi-test-berich… 2017-51  518.              502 TRUE
##  9 hdmi-erweitern-21 2017-51  333.              334 FALSE
## 10 hdmi-laenge-21     2017-51  608.              334 TRUE
```

Diese Benchmark-Erreichung wird wie folgt grafisch visualisiert:



3.6 Deployment: Aufbau des Scoring-Widgets

Mit den nun neu gewonnen Informationen über Definition und Erreichung eines Benchmarks für einen bestimmten Datenpunkt in einer Kalenderwoche lassen sich als Seitenbetreiber gewinnbringende Erkenntnisse ableiten.

Als Seitenbetreiber mit Affiliate-Ansatz stehen im Prinzip drei grobe Strategien zur Verfügung, um die Umsätze einer Seite zu erhöhen:

1. Mehr Traffic auf die Seite bringen, bspw. durch Werbemaßnahmen und so die Anzahl der Seitenaufrufe steigern
2. Die CTR erhöhen, indem beispielsweise mehr Affiliate-Links auf einer Seiten-URL platziert werden oder die Platzierung prominenter erfolgt
3. Die RPC erhöhen, indem die verlinkten Produkte besser den Kundeninteressen angepasst werden.

Die Informationen zur Benchmarkerreichung teilen einem Seitenbetreiber mit, welche Strategie er auf welcher Seiten-URL anwenden muss. Der Scoring-Datensatz sieht so aus, dass über alle Seiten-URL der Mittelwert der Benchmarkerreichung gebildet wird. Dieser Mittelwert wird damit zum Benchmark-Score. Hat eine Seiten-URL immer den Benchmark erreicht oder geschlagen, beträgt der Score 100%, wurde der Benchmark nie erreicht beträgt die Quote 0%. Die folgende Tabelle zeigt die Seiten-URLs mit den jeweils höchsten Scores für die RPC-Benchmarkerreichung:

```
## # A tibble: 10 x 4
##    Tag                     Score_CTR Score_RPC Pageviews_Gesamt
##    <chr>                       <dbl>     <dbl>            <dbl>
##  1 pf_rd_r-21                  1         0.92                37
##  2 hdmi-laenge-21              0.0962    0.827            39728
##  3 hdmi-kabel-4k-21            0.962     0.75             90696
##  4 shape-magic-21              1         0.730               77
##  5 hdmi-kabel-flach-21         0.979     0.729              173
##  6 lightning-usb-kabel-21      1         0.727               15
##  7 hdmi-macbook-airplay-21     0.481     0.692              596
##  8 hdmi-20-unterschiede-21     0.0577    0.673            26011
##  9 hdmi-teuer-vorteil-21       0.0385    0.654           113371
## 10 apple-tv-test-21            0         0.635             9895
```

```
## # A tibble: 10 x 4
##    Tag                              Score_CTR Score_RPC Pageviews_Gesamt
##    <chr>                                <dbl>     <dbl>            <dbl>
##  1 galaxy-an-tv-21                      1         0                31888
##  2 beliebte-hdmi-kabel-21               1         0.0690              51
##  3 koaxialkabel-21                      0.192     0.0962            4251
##  4 laptop-tv-anschliessen-hdmi-21       0         0.0962           76784
##  5 hdmiadapter-21                       0         0.115            37218
##  6 ps3-anschliessen-21                  0.0192    0.135             3938
##  7 25-festplatten-21                    0         0.173            15722
##  8 hdmi-galaxy-21                       1         0.173             1926
##  9 hdmi-cec-21                          0         0.192            45107
## 10 hdmi-dvi-displayport-21              0.808     0.192             5309
```

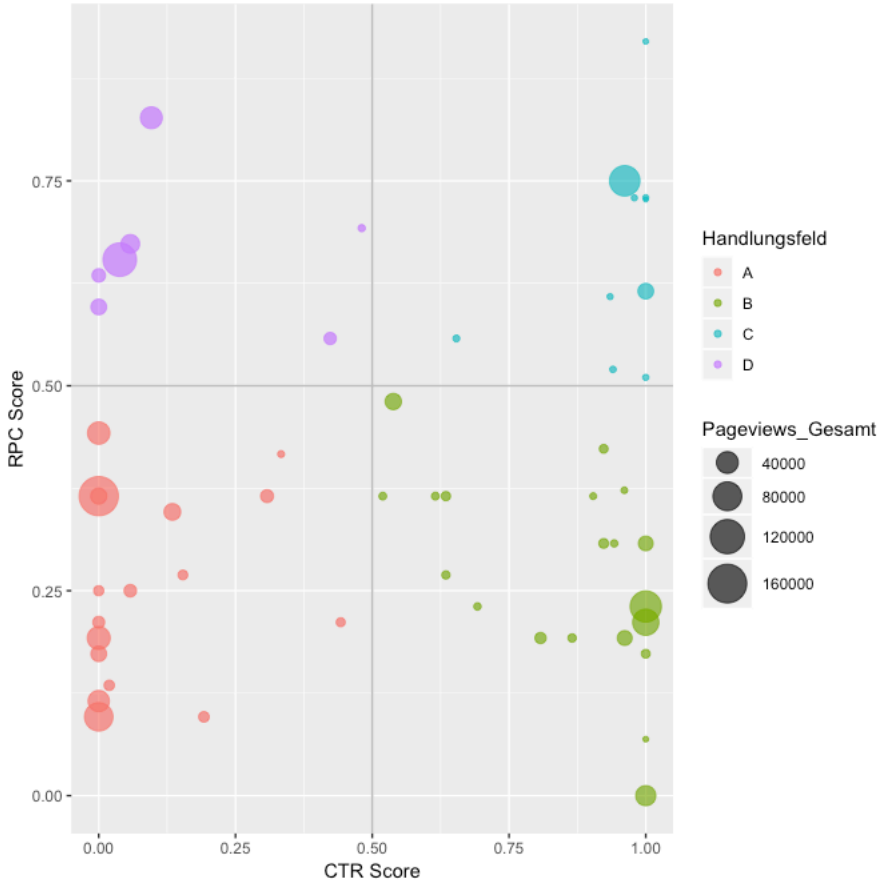
Wie eingangs erwähnt lassen sich diese Berechnungen nutzen, um daraus strategische Handlungsempfehlungen abzuleiten. Diese entstehen, indem bestimmte Score-Erreichungen bestimmten Handlungsfeldern zugewiesen werden:

```
scores_df <-scores_df %>%
  mutate(Handlungsfeld = case_when(
    Score_CTR > 0.5 & Score_RPC <=0.5 ~ "B",
    Score_CTR > 0.5 & Score_RPC > 0.5 ~ "C",
    Score_CTR <= 0.5 & Score_RPC > 0.5 ~ "D",
    TRUE ~ "A") # Geringer CTR Score und geringer RPC Score
    )
```

Zur besseren Veranschaulichung lassen sich diese Scores in einer Handlungsfeldmatrix darstellen:

```
scores_df %>%
  ggplot(aes(x=Score_CTR, y=Score_RPC, color=Handlungsfeld,
size=Pageviews_Gesamt)) +
  geom_point(alpha = .7) +
  scale_size_continuous(range = c(1,10)) +
  xlab("CTR Score") +
  ylab("RPC Score") +
  geom_hline(yintercept = .5, color="grey") +
  geom_vline(xintercept = .5, color=„grey")
```

Scoring Widget (Handlungsfeldmatrix)



Die einzelnen Handlungsfelder lassen sich nun wie folgt übersetzen:

Handlungsfeld A (roter Bereich) beschreibt Seiten-URLs, die sowohl hinsichtlich der Klickraten als auch hinsichtlich der vermittelten Umsätze eine eher schwache Performance zeigen. Hier sollte ergründet werden, ob diese schwache Performance strukturell bedingt ist (bspw. durch die Verlinkung der falschen Produkte oder eine schlechte Platzierung von Werbelinks) oder ob dies eher auf mangelnde Datenqualität (z.B. begründet durch zu wenig Seitenaufrufe) zurückzuführen ist. Des Weiteren sollte analysiert werden, ob in diesem Feld Seiten-URLs mit hohem Besuchervolumen anzutreffen sind. Falls „ja“, sollten hier schnellstmöglich geeignete Maßnahmen eingeleitet werden, um zuallererst eine Verbesserung der Klickrate und in einem zweiten Schritt die Beobachtung der resultierenden Umsatzkennzahlen zu adressieren.

Handlungsfeld B (grüner Bereich) beschreibt ein strategisch besonders umsatzrelevantes Handlungsfeld. Hier werden Seiten-URLs gruppiert, die bezüglich der individuellen Klickrate bereits eine relativ gute Performance aufweisen. Allerdings bietet die Konversion auf zweiter Ebene noch erhebliches Verbesserungspotenzial in Form von vermittelten Umsätzen. Es bietet sich hier an, mit der Verlinkung von ggf. unterschiedlichen Produkten zu experimentieren und beispielsweise im

Rahmen weiterer Tests (z.B. A/B-Tests) zu untersuchen, ob sich die vermittelten Umsätze pro Klick durch die Verlinkung unterschiedlicher Produkte verbessern lässt.

In Handlungsfeld C (türkisfarbener Bereich) wurden Seiten-URLs gruppiert, die bereits überdurchschnittliche Performance hinsichtlich der Klickraten (CTR) und Klickpreise (RPC) aufweisen. Seiten-URLs in diesem Feld sollten unbedingt prioritär im Fokus bleiben und es sollten außerdem Anstrengungen unternommen werden, die absoluten Seitenaufrufe für diese Seiten-URLs zu erhöhen, bspw. durch geeignete Maßnahmen absichern.

Datenpunkte in Handlungsfeld D (lilafarbener Bereich) stellen Seiten-URLs dar, deren Performance hinsichtlich der zweiten Konversionsebene überdurchschnittliche Werte annimmt, deren Klickraten jedoch extrem niedrig sind. Der Fokus dieser Seiten sollte darauf liegen, zunächst die Klickraten zu erhöhen, beispielsweise durch die wirksamere oder häufigere Platzierung von Werbelinks bzw. Werbemitteln auf den einzelnen Seiten. Des Weiteren sollte beobachtet werden, wie sich die RPCs bei zunehmenden Klicks verhalten und in welches Handlungsfeld sich die URLs dadurch ggf. entwickeln.

Ferner lässt sich durch die Benchmarkmodellierung auch ein Umsatzpotenzial errechnen unter der Annahme, dass der Benchmark für CTR und RPC immer mindestens erreicht wäre. Diese Information zur Umsatzprognose kann als weiterer Kontext im Diagramm dargestellt werden, beispielsweise interaktiv bei Markierung einzelner Punkte.

Im konkreten Fallbeispiel wurde diese Visualisierung als interaktives HTML-Widget bereit, welches neben der jeweiligen Seiten-URL auch das dazu gehörige Umsatzpotenzial anzeigt, sobald mit der Maus über den jeweiligen Punkt gefahren wird.

4. Zusammenfassung der Ergebnisse

In der vorliegenden Arbeit wurden statistische Methoden der explorativen Datenanalyse auf ein konkretes Fallbeispiel angewendet, um daraus Empfehlungen abzuleiten, wie der im Beispielfall generierte Umsatz gesteigert werden kann.

Die Analysen haben gezeigt, dass bereits mit einfachen Analysemethoden und leicht interpretierbaren Modellen wesentliche Optimierungspotenziale hinsichtlich der Umsatzsteigerung identifiziert werden konnten.

Im Fokus der Untersuchungen stand dabei die strategische Entscheidungsebene, die maßgeblich durch menschliches Mitwirken geprägt wird. Der Kern der Handlungsempfehlung zielte somit insbesondere auf Entscheidungen ab, die in strategische Bereiche wie der Geschäftsfeldentwicklung oder das interne Benchmarking fallen.

Mithilfe unterschiedlicher Visualisierungen konnten sowohl die Kernprobleme des Fallbeispiels (ungleichmäßige Verteilung des Umsatzes über Zeit und Webseiten) aber auch die wesentlichen

Potenziale (strategische Handlungsfelder) instruktiv dargestellt werden. Des Weiteren konnten durch die Nutzung einfacher Clusterverfahren natürliche Strukturen innerhalb der Daten, bezogen auf die Klick- und Umsatzleistung, aufgedeckt werden, welche maßgeblich zur Ableitung zielführender Strategien beigetragen haben. Zudem unterstützte die Clusteranalyse dabei, lineare Wirkungszusammenhänge zwischen den Variablen Seitenaufrufe und Klicks sowie Klicks und Umsätze präziser darzustellen: So wurden einfache Regressionsmodelle auf Basis einzelner Cluster entwickelt, um die wesentlichen Wirkungszusammenhänge zwischen den Variablen Seitenaufrufen und Klicks bzw. Klicks und Umsätze über alle Seiten-URLs hinweg besser modellieren zu können. Gleichzeitig wurde bei der Modellierung darauf geachtet, dass die jeweiligen CTR bzw. RPC-Mediane in den zu modellierenden Clustern höher waren, als in den nicht modellierten Clustern.

Dies hatte zur Folge, dass die Regressions-Prognosen gleichzeitig als Ziel-Werte für alle (d.h. auch die nicht modellierten Cluster) übernommen werden konnten. So ließen sich individuelle Scorings in Bezug auf die Erreichung oder Nicht-Erreichung des Zielwertes auf Basis jedes einzelnen Datenpunktes berechnen. Im Gegensatz zu CTR- und RPC-Analysen, die beispielsweise nur auf der Betrachtung von Durchschnittswerten einzelner Werbemittel oder Seiten-URLs basieren, liefern diese Ergebnisse deutlich genauere Benchmarks.

Die Validierung der Regressionsmodelle wurde mit Hilfe von Standardverfahren wie zum Beispiel der Residuenanalyse durchgeführt.

Auf Basis der jeweils pro Seiten-URL erreichten CTR- und RPC-Scorings konnten insgesamt vier strategische Handlungsfelder identifiziert werden, die jeweils unterschiedliche Maßnahmen für die ihnen zugeordneten Seiten-URLs implizieren.

Als Nebeneffekt haben die Untersuchungen außerdem gezeigt, dass die meisten Umsätze jeweils zu Wochenbeginn und Wochenende generiert werden. Um Umsatzausfälle zu reduzieren, sollten Wartungsarbeiten daher tendenziell eher in der Wochenmitte als am Ende einer Woche stattfinden.

Auf einer übergeordneten Ebene hat das Fallbeispiel indirekt demonstriert, dass im Bereich des Online-Marketings selbst kleinste Projekte bereits relativ viel Datenaufkommen und -komplexität verursachen. Doch benötigt man immer direkt Machine-Learning-Methoden um aus diesen Daten belastbare Erkenntnisse für die Zukunft zu gewinnen? Die Antwort ist ganz klar: Nein.

Denn selbst einfachste Basis-Werkzeuge der explorativen Datenanalyse bieten immer noch ein hoch wirksames Rüstzeug: Clusteranalysen und Regressionsmodelle unterstützen dabei, die groben Strukturen und Zusammenhänge innerhalb der Daten besser zu verstehen und einzuordnen. Einfache Visualisierungen wie Histogramme oder Boxplots geben nach wie vor eine gute, erste Orientierung über die Verteilung von Daten, unabhängig von deren Umfang. Diese Werkzeuge sollten intensiv genutzt werden, auch – und insbesondere – in besonders datenintensiven und vermeintlich hoffnungslos komplexen Datenströmen, wie jene aus der Welt der Internetökonomie.

5. Kritische Würdigung

Aus dem vorliegenden Beobachtungszeitraum von zwölf Monaten lassen sich keine fundierten Aussagen zu Trends und saisonalen Effekten in Bezug auf die Seitenaufrufe und damit auch in Bezug auf Klicks und Umsätze treffen. Da die Vermutung naheliegt, dass saisonale Effekte einmal jährlich auftreten, insbesondere geprägt durch das Weihnachtsgeschäft, sollten idealerweise mindestens zwei weitere Vergleichszeiträume vorliegen, um diese Hypothese zu testen. Die Aussagen zu den Seitenaufrufen können nur dann exakt formuliert werden, wenn weitere Zeitreihen vorliegen.

Des Weiteren sollten die Betrachtungsebene idealerweise nicht nur auf Kalenderwochen- sondern auf mindestens Tagesebene erfolgen, um die Genauigkeit der Modelle weiter zu verbessern und ggf. weitere Einflussgrößen in die Modelle hinzuzunehmen (bspw. Wochentage). Da die Datensets für Umsätze und Seitenaufrufe bereits täglich vorliegen, müssten hier nur noch tagesaktuelle Werte für Klicks generiert werden, um diesen Schritt zu gehen.

Die im Rahmen dieser Arbeit durchgeführten Modellierungen, dazu zählen insbesondere die Clusteranalysen sowie die Regressionen, wurden vor dem Hintergrund einer explorativen Datenanalyse durchgeführt. Dies bedeutet, dass die Feinjustierung dieser Modelle noch zu leisten ist. Im Beispiel der Regressionsanalysen wurden die allgemeinen Anwendungsvoraussetzungen, darunter insbesondere die Annahme der Linearität, untersucht und anschließend Regressionskoeffizienten zur Modellierung vorgeschlagen. Hier sollte auch noch die Robustheit des Modells mittels Übertragung auf neue Test-Daten überprüft werden, um beispielsweise eine Überanpassung des Models auszuschließen. Im Falle der Clusteranalyse sollten die Gruppenunterschiede weiterführend mithilfe von Teststatistiken validiert werden. Insgesamt sind die Empfehlungen der explorativen Cluster- und Regressionsanalyse als Ausgangspunkt für weitere Detailanalysen zu verstehen und nicht als finales Endprodukt. Die Modelle müssten auf neuen Daten regelmäßig neu validiert werden, um zu sehen, ob die Anwendbarkeit nach wie vor gegeben ist.