DEUTSCHER MARKETING VERBAND

COMPETENCE CIRCLE **#CCTIM**

MIT DEN RICHTIGEN DATEN DIE BESSEREN ENTSCHEIDUNGEN TREFFEN

Editorial

Crowdsource Reviews: Kundenbewertungen verlieren an Aussagekraft und behindern Innovationen.

Der Competence Circle Technologie, Innovation & Management #ccTIM ist ein Kompetenzzentrum für AI, Blockchain, Data Analytics, digitale Geschäftsmodelle und weitere technologische Trends. Denn die fortschreitende Globalisierung und Digitalisierung stellt Unternehmen vor neue Herausforderungen.

Wie schaffen wir "Added Value" für die Gesellschaft und für unsere Kunden mit neuen Technologien?

In diesem Whitepaper dreht sich alles um die richtigen Daten, um damit die besseren Entscheidungen treffen zu können. Sie erfahren, warum immer mehr Unternehmen uns nach dem Einkauf direkt zu einer Kundenbewertung auffordern und wenn man Produktbewertungen analysiert, warum sie an Aussagekraft verlieren und sogar Innovationen behindern können.

Sie bekommen einen Einblick „hinter die Kulissen" von Studien - zu oft glauben wir den provokanten Headlines und den Studien. Doch wie werden die Daten eigentlich richtig gesammelt und aufbereitetet und wie kann ich mit einem geschulten Blick analysieren, ob die Studie tatsächlich korrekt erstellt wurde? Zum Schluss bekommen Sie anhand von Checklisten noch Tipps, wie Sie selber eine aussagekräftige Studie erstellen.

Prof. Dr. Urs E. Gattiker
Präsident Marketing Club Lago
Leiter Competence Circle
„Technologie, Innovation & Management" #ccTIM

Abstract

Beruflich und privat segeln wir durch ein kaum noch überschaubares Datenmeer. Unsere Schreibtische werden überschwemmt mit Kundenbewertungen für Produkte und Dienstleistungen sowie Nutzerdaten von Facebook, Instagram, Linkedin oder Twitter sowie Website-Analytics von den vielen Tools, die wir im Marketing einsetzen. Zusätzlich bieten Unternehmen ihren Kunden vermehrt Studien zum kostenlosen Download an. Ranglisten, wie sie in Blogs publiziert werden, reproduzieren Studiendaten im Überfluss.
Wie können Unternehmen strategisch von Big Data und Small Data profitieren?

Gemäß Schätzungen weigern sich ein Drittel der Online-Käufer in den USA, Produkte zu kaufen, die keine positive Bewertungen haben. Gleichzeitig weist das Bundeskartellamt als unabhängige Wettbewerbsbehörde, deren Aufgabe der Schutz des Wettbewerbs in Deutschland ist darauf hin (2020-10), dass gefälschte und manipulierte Nutzerbewertungen beim Online-Shopping auf dem Vormarsch sind.

Die Datenflut steigt und wir wissen nicht mehr, welchen Daten wir tatsächlich vertrauen können, um daraus die richtigen Entscheidungen für uns abzuleiten. Wir müssen Studien, Ranglis-

Inhalt

ten, Nutzerbewertungen sowie Daten von Facebook oder Google kritisch hinterfragen. In diesem Whitepaper führen wir daher zwei substantielle Erkenntnisse aus:

Auf der einen Seite dürfen wir Studiendaten über Korrelationen keineswegs als Beweise für ein kausales Verhältnis zwischen Ursache und Wirkung lesen. Zusammenhänge bedeuten nicht, dass das eine zwingend aus dem anderen folgt.
Auf der anderen Seite verlieren insbesondere Kundenbewertungen an Aussagekraft und behindern Innovation – und das nicht nur, weil sie häufig gefälscht oder manipuliert wurden. Viele Fehler liegen schon in den Bewertungssystemen selbst.

Für ein besseres Verständnis bieten wir Ihnen relevante Informationen, Checklisten, Tipps und Tricks mit vielen aktuellen Beispielen aus der Praxis. Wir illustrieren, wie Leser Studienresultate, Daten und Fakten aus der Werbung und Marktforschung aber auch Crowdsource Reviews und ihre zugrunde liegenden Bewertungssysteme besser verstehen – und somit einschätzen können, inwiefern sie ihnen vertrauen können oder eher weniger.

Die Leithemen

Wie können Unternehmen strategisch von Big Data und Small Data profitieren?

1 **Mithilfe von Daten Licht ins Dunkel bringen:**
Eine Entscheidung aus dem Bauch heraus ist meist mit einem großen Risiko verbunden. Im Marketing sind wir auf zuverlässige Daten und Insights angewiesen, die uns zu besseren Entscheidungen verhelfen.
Doch welchen Daten können wir vertrauen? Von Studien über Ranglisten bis hin zu Crowdsource Reviews müssen wir unseren kritischen Blick trainieren, um mithilfe von den richtigen Daten Licht ins Dunkel bringen zu können.

2 **Studien richtig verstehen: Korrelation ist nicht gleich Kausalität!**
Der häufigste Fehler in den Medien ist es wohl, eine Korrelation als einen Beweis für einen kausalen Zusammenhang zwischen Ursache und Wirkung darzustellen. Übrigens macht auch George Clooney Daten nicht zuverlässiger und vertrauenswürdiger. Ein Faktencheck hilft hier.

3 **Datenerhebung im eigenen Unternehmen:**
Wenn wir die Meinungen von Kunden abbilden wollen gilt es, eine möglichst repräsentative Gruppe von Teilnehmern einer Umfrage zu sichern. Die Variablen Sprache, Alter, Geschlecht, Einkommen usw. sind fast immer wichtige Messpunkte, denn sie können Aussagen stark beeinflussen. Je mehr Faktoren Einfluss auf das Ergebnis haben, desto größer ist die benötigte Datenmenge und damit auch die Stichprobe, um daraus eine verlässliche Aussage treffen zu können.

4 **Crowdsource Reviews und Bewertungssysteme durchschauen:**
Damit Kundenbewertungen uns einen wahren Mehrwert liefern, müssten sie einer Analyse wie der kritischen Betrachtung des Lektors einer wissenschaftlichen Studie standhalten. Doch viele Bewertungen sind oft gefälscht und auch die Bewertungssysteme selbst entsprechen oft nicht dem wissenschaftlichen Standard. Inwiefern helfen uns Kundenbewertungen so noch weiter, wenn sie an Aussagekraft und Nützlichkeit verlieren und Innovationen behindern?

Schlagworte

#Analytics, #Big Data, #Digital Marketing, #Content Marketing, #Crowdsource Reviews, #Kundenbewertungen, #Marketing-Studien, #Reliabilität, #Validität, #Studie, #Trust, #Airbnb, #Alibaba, #Amazon, #Apple, #Deliveroo, #drkpi, #eBay, #Motel One, #Uber, #Zalando

Zitiervorschlag

Gattiker, Urs E.; Sinistra, Patrizia; Babuzki, Johanna & Temmen, Taina (Januar 2021).
Mit den richtigen Daten die besseren Entscheidungen treffen.
Whitepaper. Duesseldorf: Deutscher Marketing Verband e.V. (DMV). https://test.drkpi.ch/download/38/

Datenqualität bewerten und verbessern

Marketing baut auf der Suche nach Daten und ihrer kritischen Analyse auf, um Erkenntnisse zu gewinnen und daraus bessere Entscheidungen treffen zu können. Marketers brauchen daher Zugang zu aussagekräftigen Daten und Insights. Detaillierte Informationen über Kunden und ihr Kaufverhalten, die Antworten auf strategische Fragen liefern, sind Beispiele dafür. Von Interesse ist dabei etwa, was Kundenrezensionen bewirken können.

Vielleicht ist das ein Grund dafür, warum Ranglisten über Produkte und Dienstleistungen, Konzertsäle oder andere Tourismusdestinationen Hochkonjunktur haben. Auch Politiker nutzen internationale Ranglisten wie das Weltbank-Ranking zum Geschäftsklima als Argumentationshilfen (World Bank, 2019-05). Solche Ranglisten, die auf unbestimmten Faktoren basieren, werden mittlerweile in inflationären Maß produziert. Jeder will sich einen Platz auf dem Treppchen sichern. Ein Rangplatz im Mittelfeld ist dabei schon Grund genug für für einen symbolischen Sprung von der Klippe.

Doch welche Kriterien werden genutzt, um zum Beispiel entweder Basel oder Luzern mit "dem besten Konzertsaal der Schweiz" ausstatten zu können? Das können wir nur nachvollziehen, wenn sinnvolle Auswahlkriterien klar definiert und publiziert sind. Beispielsweise, wie messen wir die unvergleichliche Akustik welche z.B. den Münchner Gasteig zum besten Konzertsaal der Welt macht? Laut Mauró (2015-02) braucht es für ein optimales Klangergebnis verschiedene Faktoren, die zusammenspielen müssen. Doch welche Faktoren er ausgewählt hat und wie er dies bewertete, um seine Rangliste der besten 10 Konzertsäle der Welt zu erstellen, bleibt dabei ein Rätsel.

Diese wenigen Beispiele zeigen bereits: Daten müssen unabhängig ihrer Quelle erst einmal auf ihre Gültigkeit (Validität) und Zuverlässigkeit (Reliabilität) überprüft werden. Nicht alles lässt sich so einfach messen, wie es oftmals dargestellt wird. Korrelationen lassen sich leichter aufzeigen als eindeutige, kausale Verhältnisse – und sie sorgen für die besseren Schlagzeilen in den Medien. Außerdem lassen Daten – im Kontext betrachtet – oft sehr unterschiedliche Rückschlüsse zu.

Leitfragen für einen Schnelltest

Tagtäglich finden wir in den Medien über provokante Thesen abgeleitete fragwürdige Aussagen bzw. verfälschte Ergebnisse aus beliebigen Studien. Oft sind diese aus zweiter Hand. Wer die zugrundeliegenden Studien dann nicht selbst liest, ist nur so schlau wie der Journalist als Nutzer der Studienresultate. Um Daten vertrauen zu können, müssen wir die zugrundeliegenden Studien daher selbst lesen und einige Punkte überprüfen.

Gesunder Menschenverstand und ein wenig Zeit minimieren die Risiken und sparen uns oft später Geld und Ärger, wenn die falschen Entscheidungen im Marketing getroffen werden, die im schlimmsten Fall zu gewaltigen Flops führen können. Wir bieten Ihnen daher eine Checkliste als Vorlage zur Beurteilung einer beliebigen Studie. Damit stellen wir sicher, dass wir eine Studie wirklich verstehen und mögliche Schwachstellen erkennen (siehe Seite 04).

In einer Gruppe von Fachfremden können bereits 59 Prozent allein auf Basis der Beschreibung von Hypothese und Methodik einer Studie korrekt angeben, ob eine Studie Sinn macht und repliziert werden kann (Hoogeveen, Sarafoglou & Wagenmakers, 2020-09). Es lohnt sich also auch für Nicht-Experten, hier 30 Minuten Zeit zu investieren.

Wird eine Studie – wie beispielsweise die des Bundeskartellamts (2020-10-06) zu gefälschten und manipulierten Nutzerbewertungen beim Online-Kauf publiziert – lohnt es sich, auch den angehängten dreiteiligen Fragebogen genauer unter die Lupe zu nehmen. Unternehmen mussten hier Aussagen wie beispielsweise

- **"Besser bewertete Produkte erscheinen in der Default-Einstellung weiter oben."**
- **"Häufiger bewertete Produkte erscheinen in der Default-Einstellung weiter oben."**

mittels einer Skala von eins (sehr stark) bis fünf (sehr schwach) bewerten.

Um die Frage zu beantworten, inwiefern Unternehmen Kundenbewertungen beeinflussen, hat das Bundeskartellamt also eher generelle Tendenzen gemäß der Selbsteinschätzung der Unternehmen abgefragt. Wie viele manipulierte Bewertungen für einige Unternehmen publiziert werden, wurde dabei jedoch nicht untersucht.

Allerdings ist fraglich, ob die Frage nach der genauen Anzahl gefälschter Rezensionen ein zuverlässiges Ergebnis liefern kann. Der Grund ist die Response-Bias durch die soziale Erwünschtheit. Das bedeutet, dass Studienteilnehmer oftmals das äußern, was von der Gesellschaft eher akzeptiert wird.

Wir können also nicht sicher sein, ob die endgültige schwarze Zahl auf vollkommen ehrlichen Antworten basieren würde. Wie viele Teilnehmer würden eine Antwort auf diese Frage gänzlich verweigern (Non-Response)? Damit hatte das Bundeskartellamt übrigens auch zu kämpfen (siehe Abschlussbericht des Bundeskartellamts, 2020-11, S.10-11).

Checkliste für die Evaluation von beliebigen Studienberichten

	Leitfragen	Beispiele von relevanten Faktoren
	Validität: Ist die Studie sinnvoll und gültig?	
1.	**Auswahl des Untersuchungsgegenstandes:** Gibt es Neigungen oder Voreingenommenheiten der Forschenden? Ist die Studie repräsentativ?	Wer ist verantwortlich für die Studie, wer finanziert sie und aus welchem Interesse (falls bekannt)? Wer führt sie durch? Sind die Parteien unabhängig, beispielsweise bei einer Medikamentenstudie? Ist die Stichprobe groß genug? Wird eine repräsentative Auswahl an Versuchspersonen getroffen? Wer sind die Befragten (Kunden, Studenten, Mitarbeiter, zufällig ausgesuchte Teilnehmer anhand von Mobil- und Festnetz-Anschlüssen, etc.)? Inwiefern sind beispielsweise Erkenntnisse aus einer Studie an amerikanischen Studenten auf Kunden in der D-A-CH-Region übertragbar?
2.	**Studienverlauf:** Wird die Studie einmal, mehrmals, über einen kurzen Zeitraum oder langfristig durchgeführt? Hat die Studie mehrere Phasen?	Gibt es zum Beispiel mehrere Gruppen, um einen Vorher-Nachher-Vergleich zu machen? Wird der Verlauf der Studie durch externe Faktoren, wie bestimmte Ereignisse oder durch einen Lerneffekt der Versuchspersonen beeinflusst? Wurden Langzeitfolgen überprüft?
3.	**Datenerhebung:** Wie wurden die Daten gesammelt (Online-Fragebogen, im Labor, Telefoninterview, Online-Simulationen, per App auf Mobiltelefon, z.B. GPS, etc.)?	Wurden strukturierte Interview-Fragen vorbereitet? Werden jedem Teilnehmer dieselben Fragen gestellt oder gibt es Abweichungen, z.B. andere Formulierungen der Versuchsleiter in persönlichen Gesprächen? Sind die Fragebögen einsehbar bzw. angehängt? Hat die Bewertungsskala mit beispielsweise fünf oder sieben Optionen einen neutralen Mittelpunkt oder nicht?
4.	**Eindeutigkeit:** Gibt es einen klaren, kausalen Zusammenhang zwischen Ursache und Wirkung oder kann lediglich eine Korrelation aufgezeigt werden?	Sind Definitionen klar und eindeutig? Sind die Fragen klar und nicht mehrdeutig formuliert? Wird wirklich genau gemessen, was Ziel der Messung ist, oder werden noch andere Faktoren unbewusst bzw. gezwungenermaßen mitgemessen? Gibt es mögliche Störfaktoren?
	Reliabilität: Ist die Studie zuverlässig und reproduzierbar?	
5.	**Ort und Zeitpunkt der Studie:** Messungen, die zu verschiedenen Zeitpunkten oder an unterschiedlichen Orten durchgeführt werden, produzieren häufig unterschiedliche Ergebnisse.	Wann und wo wird die Studie durchgeführt? Wirkt sich beispielsweise die Jahreszeit oder der Ort auf die Studie aus (Winter in Großstadt oder Sommer am Strand)? Gibt es besondere Ereignisse, die die Studienergebnisse beeinflussen könnten (z.B. Wahl eines neuen Präsidenten, Marktveränderungen, etc.)?
6.	**Soziale Erwünschtheit:** Antworten Teilnehmer der Studie ehrlich beziehungsweise sind sie fähig, wahrheitsgemäß zu antworten?	Gibt es beispielsweise unangenehme Fragen? (Im persönlichen Gespräch neigen Probanden eher dazu, sich in ein besseres Licht zu rücken als in anonymen Fragebögen.) Wissen Teilnehmer die Antworten oder glauben sie vielleicht nur eine Antwort sei wahrheitsgemäß, obwohl sie nicht der Realität entspricht?
7.	**Methodik:** Nutzt die Studie eine gut fundierte, objektive Methodik, die im Studienbericht genauer beschrieben ist? Sind die Erklärungen zur Methodik nachvollziehbar und verständlich?	Wie wurden Interviews geführt? Sind diese aufgezeichnet worden und wie wurden diese analysiert? Sind Fragebögen standardisiert, oder besteht die Möglichkeit, dass Fragen auf unterschiedliche Art und Weise gestellt werden (z.B. im Interview)? Haben die Durchführenden die Ergebnisse beeinflusst? Weist die Studie bereits auf eigene Schwachstellen hin, die in Kauf genommen wurden (z.B. aus Kostengründen)?

8.	**Wiederholbarkeit:** Können wir die Resultate mithilfe von eigenen, kleineren Experimenten kritisch überprüfen, um sie für die Optimierung unseres Produktes und/oder der Services zu nutzen?	Große Mengen von Daten bieten nicht automatisch einen Mehrwert. Sie können nicht einfach auf jede beliebige Personengruppe übertragen werden und können im schlimmsten Fall zu falschen Rückschlüssen führen. Auch in kleineren Mengen erzeugen Daten einen Wert – zum Beispiel mit einem A/B-Test auf der eigenen Website. Damit können wir insbesondere überprüfen, ob die Ergebnisse der Studie auf unsere Zielgruppe übertragbar sind.
9.	**Objektive Interpretation:** Lassen die genutzten (und genau bestimmten) Indikatoren sowie Kennzahlen die von den Autoren gemachten Rückschlüsse zu?	Wurden die Daten korrekt interpretiert oder wird eine bestimmte Perspektive nicht berücksichtigt? Eine Korrelation beweist keine Kausalität. Klare Verhältnisse zwischen Ursache und Wirkung sind sehr schwer zu beweisen, bei derartigen Behauptungen ist daher immer Vorsicht geboten.

Zeitaufwand Checkliste: 30–60 Minuten.

Notiz: Aktuell ist die Beständigkeit (Volatilität) der Daten besonders wichtig: Wie lange sind die Daten noch aktuell? Ist eine Studie, die "vor Corona" durchgeführt wurde, auch auf die Zeit während des Lockdowns und "nach Corona" übertragbar? Wurde die Studie durch den Lockdown beeinflusst beziehungsweise wurden Studienergebnisse signifikant von den Auswirkungen des Lockdowns beeinträchtigt?

Korrelation ist nicht gleich Kausalität

Der häufigste Fehler von Lesern einer Studie ist es, eine Korrelation als Kausalität zu erfassen. Eindeutige Kausal-Beweise geben auch bessere Schlagzeilen ab. Korrelationen haben weniger Wirkung, da sie nur vage Zusammenhänge aufzeigen. Daher kommt es vor, dass Headlines in Medien oft Studienresultate auf eine möglichst prägnante Aussage reduzieren, die leider manchmal auch falsch sein kann.

Ein Beispiel für die Schlagzeile einer Studie ist "Wer viel streamt, geht häufiger ins Kino." (siehe Gattiker, 2020-05). Doch exzessives Video-Streaming führt nicht grundsätzlich zu mehr Kinobesuchen, genauso wenig wie Kälte automatisch den Schneefall zur Folge hat. Diese Dinge werden von weiteren Faktoren beeinflusst, wie etwa das allgemeine Interesse an Filmen. Die Studie wies auf einen Zusammenhang hin, der nicht Ursache und Wirkung beschreibt, sondern eben nur eine Korrelation.

Ein weiteres großes Problem ist, dass wir häufig getäuscht werden – und zwar vom Aussehen der grafischen Aufmachung einer Studie oder einfach von einer Top-Schlagzeile. Das Aussehen schlägt dabei den eigentlichen Inhalt und die Qualität der Studie. Auch der Überbringer der Botschaft, in dem Fall ein Werbebotschafter, ist oft wichtiger als die Qualität des Produktes oder der Daten. Insbesondere wenn es sich dabei um prominente Persönlichkeiten wie Lady Gaga oder Georg Clooney handelt.

Marketing-Studien: Aussehen schlägt Inhalt

In den meisten Fällen sind Medienfachleute an einer Studie mit weniger als 1.000 Teilnehmern wenig interessiert. Ob die Studie gut ist oder nicht, können Journalisten nicht immer beurteilen, wenn ihnen das Know-How fehlt. Das Resultat: Die Wahl fällt auf eine Studie, die eine Top-Schlagzeile hergibt. Und ehe man sich versieht, wird aus einer Korrelation fälschlicherweise eine Kausalität.

In der Praxis bedeutet das zum Beispiel auch, dass ein Datensatz mit 1.000 Teilnehmern in der Schweiz interessanter ist als ein Datensatz mit 300 Teilnehmern aus Deutschland – auch wenn letzterem die fachlich bessere Studie zugrunde liegt. Wird die Studie mit den 1.000 Schweizer Teilnehmern nun auch noch auf die deutsche Bevölkerung übertragen, sollten bei allen die Alarmglocken läuten.

Es ist auch kein Geheimnis, dass Studien ein beliebtes Marketing-Instrument sowohl im B2B- als auch im B2C-Bereich geworden sind. Marketing-Studien können insbesondere das Interesse der informierten Kunden wecken, die sich nicht mehr so leicht von herkömmlicher TV- oder Print-Werbung täuschen lassen und das Internet zur Recherche nutzen.

Unternehmen wie Adobe, IBM, Microsoft oder Namics geben gerne ihre aufwendigen Studien in Auftrag und bieten das finale, grafisch aufbereitete Dokument schließlich kostenfrei zum Download an. Doch die Daten können nicht immer differenziert genug analysiert werden. Wenn die möglichen Effekte von Land, Sprache, Alter der Teilnehmer, Industriezweig und so weiter jedoch nicht berücksichtigt werden, sind Gültigkeit (Validität) und Zuverlässigkeit (Reliabilität) nicht immer garantiert.Und das Design gewinnt mehr an Priorität als der eigentliche Inhalt.

Daten im Kontext betrachtet

Der Big-Mac-Index vergleicht Daten im Kontext: Die Kaufkraft verschiedener Währungen wird mithilfe der Preise für einen Big Mac verglichen. Dabei zeigt sich, dass der Schweizer Franken im Jahr 2020 dem US-Dollar gegenüber mit gut 20 Prozent überbewertet ist. Beim Euro ist es eine gut 16-prozentige Unterbewertung. Der Index zeigt auf einfache Art und Weise auf, wie Wechselkurse die Kaufkraftparitäten nicht unbedingt mit einbeziehen.
(https://www.economist.com/news/2020/07/15/the-big-mac-index).

Aufgrund von Steuerbetrug und Schwachstellen im System entgehen den EU-Mitgliedstaaten Einnahmen aus der Mehrwertsteuer. Die Differenz zwischen den erwarteten und den tatsächlichen Mehrwertsteuereinnahmen wird als Mehrwertsteuerlücke bezeichnet. In Euro angegeben, weist Italien 2018 die größte Mehrwertsteuerlücke auf, gefolgt vom Vereinigten Königreich und Deutschland. Griechenland ist erst auf Platz sechs zu finden. Das entnehmen wir einer Rangliste der Financial Times (2020-09, S.1).

Doch diese absoluten Zahlen sind irreführend. Wenn wir die Daten in einem anderen Kontext betrachten – wie viel Prozent der zu erwartenden Mehrwertsteuereinnahmen des jeweiligen Landes noch fehlen – sieht die Rangliste ganz anders aus (und deutet auf ein anderes Problem hin): In Deutschland macht die Mehrwertsteuerlücke unter 10 Prozent der Gesamtsumme aus. In Griechenland hingegen ist sie bei rund 30 Prozent der gesamten Mehrwertsteuereinnahmen.

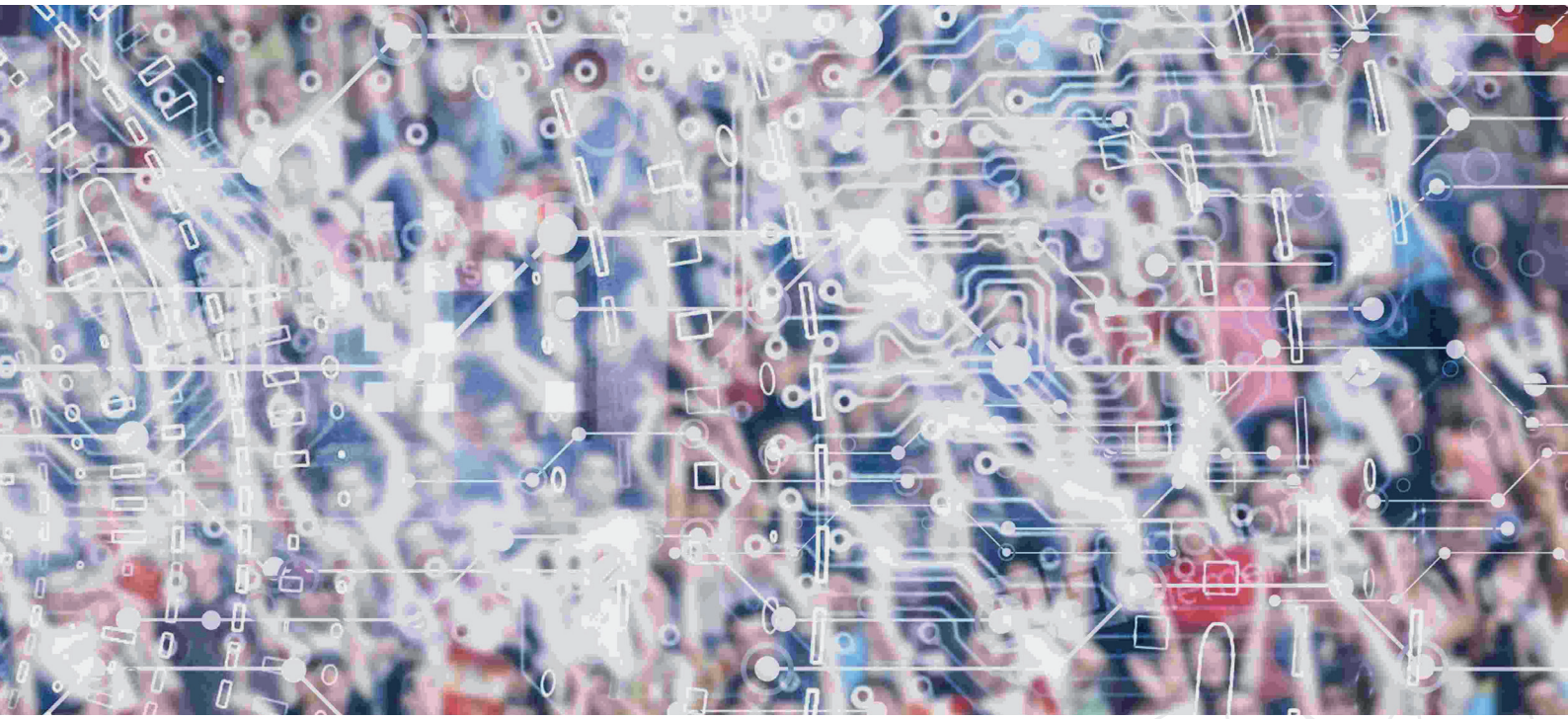
In der EU hat Griechenland vielleicht ein größeres Problem in Sachen Mehrwertsteuer als Deutschland, doch ist das Land auch Vorbild: Zum Beispiel hat Griechenland den Asylantrag für gut 162 Flüchtlinge pro 100.000 Einwohner schon beim Erstgesuch bewilligt. Für die Schweiz waren es 126, für Deutschland 85, für Malta 82 und für Schweden 59 Flüchtlinge (pro 100.000 Einwohner) (EASO, 2020).

Verglichen mit verschiedensten Ressourcen der Länder (z.B. flächenmäßige Landesgröße in Quadratmetern statt Einwohnerzahl) könnten wir willkürlich immer neue Bilder zeichnen, neue Rangplätze verteilen und mal das eine, mal das andere Land weiter oben oder unten platzieren.

Auch bei Unternehmensentscheidungen können wir sehr gut unterscheiden zwischen internen und externen Kontexten, die es zu berücksichtigen gilt. Der externe Kontext ist zum Beispiel der Vergleich mit der Konkurrenz auf dem Markt.

Ein einfaches Beispiel: Aufgrund des Lockdowns macht ein Restaurantbetreiber – verglichen mit anderen Restaurants in der Region – vielleicht nur einen geringen Verlust. Das sagt noch nichts darüber aus, ob es nicht schon dieser Verlust ist, der ihn in den Ruin treiben kann, weil zum Beispiel die Rücklagen nicht reichen, um auch die zweite Phase zu überbrücken.

Kürzlich wies ein Kunde von drkpi® darauf hin, dass das Leser-Engagement auf der unternehmenseigenen Community Page rückläufig sei. Dank des drkpi® PageTracker



konnte jedoch ein Langzeitvergleich für das Unternehmen des Kunden mit drei Konkurrenten errechnet werden. Und plötzlich sah das Ganze anders aus. Was das Leser-Engagement betraf, war die Community-Seite im Vergleich zur Konkurrenz extrem gut und in Deutschland in den Top 10 der Community Pages in diesem Industriezweig. Die Konsequenz: Das Investment in die Community Pages wurde nicht heruntergefahren, denn wie sich zeigte, sind sie essentiell für die Qualität und Performance der eigenen Webseite und für das Content Marketing.

Daten sind immer dann besonders interessant, wenn diese auch im Kontext betrachtet werden. "Wie viele Corona-Neuinfektionen gibt es pro 100.000 Einwohner?" Die Antwort kann – ganz klar – für eine dicht besiedelte Großstadt wie Berlin etwas anderes als für den Schwarzwald bedeuten.

Die Liste an Beispielen lässt sich über alle Bereiche hinweg bis ins Unendliche fortführen, denn es lassen sich beliebig viele neue Kontexte formulieren. Die Frage ist natürlich, wann eine Betrachtung eines bestimmten Kontexts sinnvoll ist und – in Marketing-Angelegenheiten – wann sie für uns sinnvoll ist. Eindeutig ist, dass die Betrachtung von Daten im Kontext helfen kann, Situationen besser einzuschätzen.

Natürlich begegnen wir in der Öffentlichkeit und in den Medien auch häufig professionell wirkenden Statistiken und Behauptungen, die Daten eher in das eine als in das andere Licht rücken. Ein bestimmter Kontext wird vorsätzlich oder mangels Wissen nicht beleuchtet. Hier hilft oft eine Recherche, um sich ein neutrales Bild zu machen oder zumindest mehrere Perspektiven kennenzulernen.

Sind die Daten vertrauenswürdig?

Daten sind nicht einfach gegeben. Egal wie vertrauenswürdig eine Grafik oder eine Studie wirkt, wir müssen uns immer die kritische Frage stellen: „Woher kommen diese Daten eigentlich?", damit wir Auswertungen vertrauen und in eine Entscheidung einfließen lassen können: Wer hat die Daten wie gesammelt – wann und wo? Wie wurden die Daten gesammelt und analysiert? Wie objektiv und zuverlässig ist die Interpretation der Ergebnisse?

Letzten Endes kamen die Ergebnisse von irgendwo her: Jemand hat gezählt oder gemessen und vielleicht noch gerechnet – idealerweise systematisch und mit Sorgfalt. Diese Bemühungen erfordern Investment in Form von Budget und Zeit sowie Fachwissen. Sie sind weder selbstverständlich noch garantiert.

Auf der anderen Seite ist eine gute Studie daran zu erkennen, dass im Studienbericht auch auf die in Kauf genommenen Schwächen, beispielsweise auf Probleme einer bestimmten Messtechnik oder auf Abweichungen selbst, hingewiesen wird.

Wir alle wollen aussagekräftige Daten und Insights. Wir benötigen detaillierte Informationen über unsere Kunden in jedem Bereich. Nur mit diesen können wir nachweislich aufzeigen, ob unsere Marketing-Aktivitäten erfolgreich sind.

Die Sammlung und Zusammenführung aller verteilt vorliegender Marketing-Daten wird daher oft als kritischer Schritt angepriesen. Nur dank dieser Zusammenführung und Vereinheitlichung kann eine umfassende Datenanalyse stattfinden. Doch was nützt die Zusammenführung von Daten, wenn diese nicht vertrauenswürdig sind?

Daten sammeln ist aber noch lange nicht alles. Dass es einen Unterschied macht, wie Daten erhoben werden, illustriert eine Studie über den Fleischwaren-Kauf. Die Studie mit 18.000 Verkäufen in 18 Edeka Supermärkten zeigt, dass Konsumenten oft immer noch Billigfleisch bevorzugen. Wenn Billigfleisch pro Kilo nur 50 Cent billiger ist als Fleisch aus besseren Haltungsbedingungen, greifen zwei Drittel der Käufer zum günstigeren Produkt (Enneking, Kleine-Kalmer, Dauermann & Voigt, 2019-01).

Dieses geringe Kaufinteresse stand dabei im Widerspruch zu den Ergebnissen der parallel durchgeführten Befragung mit fast 700 Konsumenten. In den kurzen Interviews gaben deutlich mehr Konsumierende an, Tierwohl-Produkte zu bevorzugen.

Fazit: Konsumenten geben im persönlichen Gespräch vielleicht an, dass sie sich eine artgerechte Haltung von Hühnern, Rindern oder Schweinen wünschen. Doch der eigene Geldbeutel ist beim unbeobachteten Zahlen an der Kasse dann doch wichtiger – wir haben es wieder mit dem Problem der sozialen Erwünschtheit zu tun.

Es ist schwierig im Voraus zu prüfen, welchen Aufpreis der Kunde tolerieren würde, wenn es dabei um soziale Akzeptanz geht. Nach Steve Jobs sagen Nutzer scheinbar immer "Ja" zu möglichen neuen Features bei technischen Produkten. Ob sie diese dann wirklich nutzen würden und willens sind, dafür zu zahlen, steht auf einem anderen Blatt.

Zum Träumen sagt kaum jemand "Nein" – bis wir den Preis sehen. Wer bares Geld zahlen soll, überlegt es sich dann doch auch schnell wieder anders. Dieses Problem begegnet uns in der Marktforschung besonders häufig.

Zuletzt ist auch bei Statistik-Dienstleistern, wie zum Beispiel Statista, ein zweiter Blick auf die Herkunft der Daten empfehlenswert. Erst die Überprüfung, woher die Daten eigentlich kommen, zeigt uns, wie verlässlich die Daten wirklich sind. Dabei ist es vielleicht auch fragwürdig, warum die Referenzen für eine Statistik wie dem Big-Mac-Index bei Statista andere sind, als beim Economist, dem Erfinder dieses Indexes. Diese Zeitung errechnet und publiziert diesen Vergleich gemäß Kaufkraftparität bereits seit 1986 (siehe: https://de.statista.com/statistik/daten/studie/199335/umfrage/big-mac-index--weltweiter-preis-fuer-einen-big-mac/).

Checkliste für die Datenerhebung

Mit den richtigen Daten bessere Entscheidungen zu treffen, heißt auch zu experimentieren. Zum Beispiel ist es für einen Onlineshop-Betreiber interessant zu erfahren, wie die Anzahl nicht abgeschlossener Käufe (Warenkorb-Abbrüche) und somit nicht gemachter Umsätze auf der eigenen Website am besten verringert werden kann.

Im Rahmen einer großen Studie sandte Alibaba den Besuchern seines Online-Shops einen Rabattgutschein, und zwar in dem Moment, als sie ihre gefüllten virtuellen Einkaufswagen stehen ließen und den Einkauf somit abgebrochen haben. Der Gutschein war 24 Stunden gültig (Zhang, Dai, Duong et.al., 2018-12).

Auch eine unkomplizierte Umfrage auf der eigenen Website ist ein gutes Mittel, um möglichst einfach und schnell eine große Menge von Daten zu sammeln. Es muss also nicht immer ein Marktforschungsinstitut mit einer Budget-verzehrenden Studie beauftragt werden. Doch wie die Tabelle unten zeigt, gibt es im Vorfeld einiges zu berücksichtigen, um sich Zeit, Budget und sogar Ärger zu sparen und, um die richtigen Daten zu bekommen.

Zuerst gilt es festzulegen, bei welchen strategischen Fragen, welche Daten Licht ins Dunkel bringen können. Was muss gemessen werden und wie muss es gemessen werden, um konkrete und praktikable Antworten auf diese zu erhalten? Dann können wir bewerten, ob wir die notwendigen Ressourcen haben, um die Datenerhebung und -analyse selbst in die Hand zu nehmen, oder ob externe Spezialisten als Dienstleister in Anspruch genommen werden sollten.

Checkliste zur Datenerhebung und -analyse im eigenen Unternehmen

Fragen und Schritte bei der Vorbereitung	Beispiele
Zielsetzung: Formulieren Sie ein bis drei sinnvolle strategische Fragen, die es zu beantworten gilt. Die Fragen sollen Antworten liefern können, die an der praktischen Umsetzung orientiert sind.	Zu allgemein: „Wie können wir unsere Verkaufszahlen im Online-Shop erhöhen?“ (Unmengen von Variablen können hier in Betracht gezogen werden.) → Präziser (und leichter zu überprüfen): „Wie können wir die Anzahl nicht abgeschlossener Käufe im Online-Shop verringern?“ – siehe Alibaba-Studie.
Notiz: Je präziser die Fragestellung ist, desto weniger Variablen sind im nächsten Schritt meist nötig.	
Definition der Variablen oder KPIs: Welche Daten können diese Fragen möglichst präzise beantworten? Wie werden sie gemessen? Welche Daten und Methoden brauchen wir für sinnvolle Berechnungen? Welche Daten können ohne größeren Aufwand zusätzlich gesammelt oder berechnet werden? In welchem Kontext müssen wir die Daten betrachten?	Bei Alibaba z.B.: Anzahl der Online-Shop-Besucher, Anzahl der Online-Käufe (z.B. mit und ohne genutzte Gutscheine), Summe der gekauften Produkte, Einnahmen insgesamt, Anzahl der in den Warenkorb gelegten Produkte, die nicht gekauft wurden, etc. ... → Weitere interessante Vergleichswerte im Detail sind z.B. die Dauer des Besuchs der einzelnen Produktseiten oder die Gesamtzahl bei mehreren Besuchen derselben Seite.
Notiz: Oberste Priorität ist dabei sicherzustellen, dass die Daten GDPR- oder DSGVO-konform gesammelt und verarbeitet werden (siehe Gattiker, Temmen & Sinistra, 2018-04).	
Recherche: Welche Daten hat das Unternehmen bereits, z.B. aus vergangenen Experimenten? Was sagen vorhandene Studien? Gibt es sekundäre Daten (Bundesamt für Statistik etc.) und reichen uns diese vielleicht sogar für eine Entscheidung aus? Gibt es bestimmte Methoden und Ideen, mit denen wir selbst experimentieren können? Mit welchen Mitteln können die fehlenden Daten für unser Unternehmen erhoben werden?	Die Studie zu Alibaba kann beispielsweise für die eigenen Zwecke in einem kleinen Rahmen reproduziert werden – mit Änderungen oder weiteren Variablen, wo nötig. → Daten für eine Umfrage können wir mit einem Online-Tool erfassen, das wir jedoch gut beherrschen sollten. → Manche Daten – beispielsweise SEO-Daten zur Performance der Website – bekommen wir sehr schnell z.B. über Google Analytics oder drkpi® PageTracker.
Notiz: Die intensive Recherche hilft bei der Entscheidung, ob wir die Daten selbst erheben können oder ob ein externer Spezialist oder eine professionelle Studie nötig ist.	

Methode festlegen: Wie können wir Variablen messen und genau die Daten bekommen, die wir benötigen? Online-Umfrage oder via App, im Kaufhaus oder telefonisch, etc.? Welche Methode eignet sich am besten?	Fragenkataloge für Umfragen erstellen: Wir können Expertise einkaufen (beispielsweise an einer Universität) oder selbst einen Fragenkatalog entwickeln. Werden allen nötigen Variablen erfasst und sinnvoll gemessen?
Achtung: Formulierungen sind wichtig. Lassen sie alle möglichen Antworten auf eine Frage zu oder nur bestimmte? Formulierungen können die Antworten der Teilnehmer stark beeinflussen!	
Wer sind die Studienteilnehmer: Kunden oder auch neue Zielgruppen? Rekrutierung der Teilnehmer nach Wohnort, Herkunft, Sprache, Geschlecht, Ausbildung, Einkommen und anderen kritischen Faktoren (siehe Krebsstudie Beispiel 1, Seite 15). Je mehr dieser Variablen einen Effekt auf die Ergebnisse haben können, desto höher muss die Teilnehmerzahl sein, um eine repräsentative Stichprobe zu erhalten. Braucht es eine Kontrollinstanz?	Wollen wir beispielsweise auch genügend Frauen aus Deutschland auf einer Online-Shopping-Plattform für Männermode befragen, werden 1.000 deutsche Teilnehmer hier nicht reichen, wenn 80% davon Männer sind und aus jedem Bundesland eine repräsentative Anzahl Frauen jeder Altersklasse benötigt wird. Hier gilt es, Wahrscheinlichkeiten zu berücksichtigen (siehe Beispiel auf Seite 18).
Zeitpunkt und Dauer: Wann wird die Studie durchgeführt? Wird ein Experiment einmalig oder mehrmals oder langfristig (mit oder ohne Kontrollgruppe) durchgeführt? Wirken sich verschiedene Jahreszeiten oder sogar Tageszeiten auf das Ergebnis aus?	Das Beispiel Alibaba zeigt: Kurzfristig werden dank der Gutscheine mehr Produkte verkauft. Langfristig verändert sich das Kaufverhalten signifikant: Online-Shop-Besucher legen z.B. mehr Produkte in den Warenkorb, um Gutscheine per Mail zu erhalten. Eine solche Strategie kann auf lange Sicht dadurch als weniger profitabel entlarvt werden.
Notiz: Das Kaufverhalten in der Vorweihnachtszeit kann für bestimmte Produkte anders sein als im Sommer und umgekehrt. Macht es Sinn, Daten über 12 Monate zu sammeln oder nur über einen bestimmten Zeitraum auszuwählen?	
Methode/Datenanalyse: Quantitative (das heißt, parametrische oder nichtparametrische Tests) oder qualitative Analyse/Studie (beispielsweise Ethnographie).	Qualitative Datenerhebung kann sehr hilfreich sein, ist aber meist sehr zeitintensiv. Quantitative Tests lassen sich leichter durchführen.
Reporting: Bleiben die Resultate nur intern zugänglich oder auch auf der Website für Journalisten, Medien und Kunden verfügbar?	Die Alibaba-Studie wurde veröffentlicht und leistet einen wertvollen Beitrag in der Marketing-Community.
Notiz: Publizieren Sie Resultate, sollte die Berichterstattung klar und transparent sein. Auch ein Whitepaper kann sinnvoll sein. Die perfekte Studie gibt es nicht. Weisen Sie daher auch auf mögliche Schwachstellen der eigenen Studie hin, wenn Sie diese veröffentlichen.	
Umsetzung der Resultate: Inwiefern und wie genau können die Daten genutzt werden? Was soll im Marketing/Verkauf mithilfe dieser Daten optimiert werden?	Die Alibaba-Studie zeigt, dass Gutscheine die Verkaufszahlen kurzfristig steigern können, langfristig nimmt die Kaufbereitschaft für Produkte ohne Rabatt jedoch ab aufgrund der strategischen Schnäppchen-Jagd der lernenden Konsumentenn („Wenn ich Produkte im Warenkorb zurücklasse, bekomme ich einen Gutschein“).
Notiz: Meist impliziert dies schon die Fragestellung, die Sie zu Beginn formuliert haben (Kosten senken, mehr Sales generieren etc.). Sie sollten also bereits wissen, wie sich mögliche Antworten auf die formulierte Frage umsetzen lassen!	

Notiz: Die Vorschläge können als grobe Richtlinien bei der Vorbereitung der Datenerhebung im eigenen Unternehmen genutzt werden. Für die spezifischen Fragestellungen, Umfragen und andere Experimente müssen die wichtigen Details natürlich im Einzelfall entschieden werden.

Es kann selbstverständlich immer passieren, dass wir nach der Analyse und Interpretation der gesammelten Daten keine eindeutigen Antworten auf die formulierten Fragen erhalten. Das kann daran liegen, dass die zu Beginn formulierten Fragen nicht präzise genug waren. Ein anderer Fehler ist die Selektion der Studienteilnehmer oder aber der Variablen, die gemessen wurden. Fehler in der Analyse sind ebenfalls möglich. Es kann auch passieren, dass wir nicht die Antwort bekommen, die wir hören wollten.

Deshalb ist es wichtig, alles so detailliert und sorgfältig wie möglich im Voraus zu planen und wenn möglich, doch einen Spezialisten zurate zu ziehen, damit böse Überraschungen, Frust und Misserfolge vermieden werden können.

Praxisbeispiele zur Datenerhebung und Studientauglichkeit folgen am Ende des Whitepapers (Seite 16-18)

Wie sind professionelle Studien aufgebaut?
Zwei Beispiele: STATISTIK AUSTRIA (positiv) und Promarca Brand of the Year 2020 (negativ)

Was muss bei der Nutzung von Crowd-Plattformen zur Datensammlung berücksichtigt werden?
Ein Beispiel anhand von Amazon Mechanical Turk

Wie werden Stichproben richtig zusammengesetzt?

Wie kann eine möglichst hohe Datenqualität gesichert werden?

Besser Entscheidungen dank Experimenten treffen

Das Experiment des Online-Händlers Alibaba zeigt, wie hilfreich eigene Experimente im Marketing sein können, um beispielsweise den Verkauf zu optimieren. Ähnliche Testungen aber auch Umfragen kann ein Unternehmen relativ leicht selbst umsetzen, ggf. auch mit der Hilfe von Spezialisten professionell gestalten.

Bei Alibaba geht es nicht nur darum, ob ein bestimmter Rabatt funktioniert, sondern auch um das wie und das warum. Diese Daten im eigenen Unternehmen zu erheben und zu sammeln, kann für strategische Entscheidungen extrem wertvoll sein, denn wir müssen nicht auf die Daten der Anderen vertrauen und diese auf uns übertragen, was immer mit Unsicherheiten verbunden ist. Bei einer eigenen Erhebung ist allerdings darauf zu achten, dass die Daten gemäß DSGVO sicher gespeichert werden.

Eine andere Art von Experiment zeigte Steve Jobs:
Steve Jobs wird nachgesagt, dass er ein „Stickler“ für Details war, was die Benutzerfreundlichkeit der Apple-Produkte betraf. Aus diesem Grund wollte er immer selbst sehen, wie Nutzer ein neues iPhone bedienten. Er war davon überzeugt, dass er damit genau beobachten könne, wie benutzerfreundlich oder auch kompliziert die Geräte tatsächlich waren. Eine solche Beobachtung kann auch schon bei einer nicht-repräsentativen Stichprobe von nur fünf Kunden hilfreiche Erkenntnisse bringen. Im Falle eines neuen Produktes kann ein “Test auf Herz und Nieren” eines neuen Finanz-Produktes vom Management das Risiko für einen Flop-Launch minimieren. Das verhindert einen Imageverlust und höhere Kosten (Gattiker, 2020-10).

Datenerhebung durch Crowdsourcing?

Nach einem Aufenthalt im Hotel, einem Online-Einkauf oder einem Flug mit der Lieblings-Airline passiert es fast immer: Wir erhalten die Aufforderung, eine Bewertung abzugeben.

Crowdsource Reviews scheinen auf den ersten Blick ein intelligentes System zu sein: Das Unternehmen spart – im Vergleich zur aufwendigen Studie – viel Zeit und Geld, indem die Arbeit direkt an Kunden ausgelagert wird.
Die statistischen Berechnungen sind leicht zu bewältigen, programmierbar und können automatisiert im Hintergrund ablaufen.

Gleichzeitig können innerhalb kurzer Zeit Massen animiert werden, Daten abzugeben, die zwar subjektiv, im Idealfall aber auch unabhängig sind und dem Unternehmen so einen großen Mehrwert liefern können. Je mehr Bewertungen abgegeben werden, desto weniger problematisch ist auch die Subjektivität. So die Theorie.
Doch stellen wir den Blick auf die verschiedenen Bewertungssysteme einmal scharf, kommen schnell Zweifel auf. Wie werden diese Daten genau gesammelt und wie vertrauenswürdig und nützlich sind sie? Wie aussagekräftig sind Kundenbewertungen und wie gut sind die Zensuren auf den verschiedenen Plattformen wirklich?

Bei den klassischen Crowdsource-Bewertungssystemen mit fünf Sternen, die wir in den meisten Online-Portalen finden, gibt die Anzahl der Sterne den Durchschnitt aller Kundenbewertungen per Sterne-Vergabe an. Wir könnten nun erwarten, dass die schlechteste Bewertung (1,0 Sterne) auch die schlechtesten Produkte auf der Plattform markiert. Die beste Bewertung (5,0 Sterne) sollte ungefähr gleichermaßen selten zu finden, denn nur so sind wir in der Lage, die Angebote innerhalb eines Portals untereinander vergleichen zu können. Der Mittelwert von 3,0 Sternen sollte durchschnittliche Produkte beschreiben und dementsprechend am häufigsten zu finden sein. Es handelt sich in der Theorie um eine Normalverteilung.

Die Beispiele in den nächsten Paragraphen zeigen jedoch auf, dass die Verteilungen im Vergleich zu dieser erwarteten Normalverteilung stark verzerrt sind, sodass die Daten an Aussagekraft verlieren.
Ein Beispiel im Voraus (siehe Zervas, Proserpio & Byers, 2015-01): Wenn ein Online-Portal angibt, dass die dort gelisteten Hotels im Durchschnitt mit 4,5 Sternen von fünf Sternen bewertet werden, entspricht das nicht der zu erwartenden Normalverteilung mit einem Mittelwert von 3,0. Die Nutzerbewertungen zeichnen eine verzerrte Verteilung (skewed distribution) und wir können gar nicht mehr einschätzen, welches Hotel nun wirklich ein sehr gutes oder ein sehr schlechtes Hotel ist, denn anscheinend sind alle top. Doch gibt es wirklich so gut wie gar keine schlechten Hotels mehr? Wohl kaum.

Eine negative Ein- oder Zwei-Sterne-Bewertung in einer App beeinflusst das Markenimage. Gemäß einer Umfrage von Apptentive (nicht datiert) trifft dies auf rund 55 Prozent ihrer Studienteilnehmer zu. 71 Prozent der Befragten gaben dabei an, dass eine Vier- oder Fünf-Sterne-Bewertung der App einer bekannten Marke im App-Store ihre Sicht auf die Marke als Ganzes positiv beeinflusst.

Appentive weist auch darauf hin, dass eine Verbesserung der durchschnittlichen Bewertung von zwei auf drei Sterne die Conversion-Rate im App-Store um 306 Prozent steigert. Klettert die App von drei auf vier Sterne, kann das die Conversion nochmals um bis zu 92 Prozent steigern.

Kundenbewertungen verbessern die Chancen für einen erfolgreichen Verkaufsabschluss. Doch wenn die Bewertungen mit Tricks gezielt manipuliert werden, sind Aussagekraft und Nützlichkeit dieser Daten für Konsumenten abgeschwächt. Sie könnten langfristig sogar mehr schaden als helfen.

Kundenbewertungen bei Airbnb und TripAdvisor

In einer Analyse von über 600.000 Airbnb-Liegenschaften stellten Forscher (Luca & Zervas, 2015-05) fest, dass fast 95 Prozent der Angebote entweder 4,5 oder fünf Sterne erhielten. Fünf Sterne sind dabei die höchstmögliche Bewertung. Sie stellten diese Airbnb-Bewertungen denjenigen von TripAdvisor mit etwa 500.000 gelisteten Hotels gegenüber. Hier liegt der Durchschnitt bei weitaus niedrigeren 3,8 Sternen.

Eine wichtige Unterscheidung im Bewertungssystem der beiden Portale: Im Gegensatz zu TripAdvisor erlaubt es Airbnb nicht nur den Gästen, Unterkünfte zu bewerten, sondern auch umgekehrt: Gastgeber können die Gäste ebenfalls bewerten. Bewerten sich Gäste und Gastgeber gegenseitig, könnte dies dazu führen, dass alle Nutzer tendenziell positivere Bewertungen abgeben – um im Gegenzug selbst eine möglichst positive Bewertung zu erhalten (oder zumindest keine negative Bewertung zurückzubekommen).

Somit scheinen TripAdvisor-Nutzer ehrlichere Bewertungenen abzugeben als Airbnb-Nutzer. Normalerweise könnte man annehmen, dass der Mittelwert der eingereichten Evaluation bei drei von fünf Punkten liegt – nur dann hat das Bewertungssystem einen Sinn und kann dazu dienen, die Hotels in einem Portal miteinander zu vergleichen. Doch das Ranking erscheint nach oben hin verzerrt zu sein.

Dies wiederum schwächt auch hier den Wert der Bewertungen für zukünftige Gäste ab, da anscheinend fast alle Airbnb-Angebote einen Top-Service bieten.
Wie viele Bewertungen dabei zusätzlich teilweise „manipuliert“ wurden, um gegenüber den wirklich guten Unterkünften konkurrenzfähig zu bleiben, muss ebenfalls hinterfragt werden.

Uber-Evaluationen auf Basis von Reziprozität

Uber lässt Fahrer ebenfalls auch die Fahrgäste bewerten. Das heißt, dass nicht nur Kunden Uber-Fahrer bewerten, sondern auch umgekehrt. Somit kann Uber auf Basis der Bewertungen nicht nur den Fahrern kündigen, sondern auch den Kunden "Hausverbot" erteilen. Dies führt zu der unaufrichtigen gegenseitigen Zusicherung hoher Bewertungen – derselbe Fehler, dieselbe Methode, wie im Bewertungssystem von Airbnb, die zu Verfälschungen führen kann. Wenn das Rating eines Fahrers unter den Wert von 4,6 von maximal fünf Punkten fällt (Moore, 2020-01), kann das Unternehmen den Uber-Fahrer "deaktivieren". Da Uber-Fahrer keine Auftragnehmer, sondern Angestellte sind, ist das im Prinzip die Kündigung.

Ein Bewertungssystem auf Basis der Reziprozität liefert weder Kunden noch Anbietern einen besonderen Mehrwert. Es führt im obigen Beispiel im schlimmsten Fall zu Diskrimierung und Entlassungen. Wir möchten keine schlechten Bewertungen abgeben, wenn der Angestellte aufgrund dieser seinen Arbeitsplatz verlieren kann. Fazit: Wir sollten uns auf Bewertungen in einem solchen System schlicht und einfach nicht allein verlassen.

Amazon-Bewertungen gegen Vergütung

1996 ging es bei Amazon nur darum, ein gekauftes Buch zu bewerten. Heute verkaufen Amazon und unzählige andere Unternehmen auf dem Amazon-Marketplace Produkte jeglicher Art, die ebenfalls bewertet werden können. Mit illegalen Tricks bauen einige Billig-Händler ihre Vormachtstellung beim weltgrößten Online-Portal aus.

Beliebt sind zum Beispiel "Produkttester-Gruppen" für Amazon-Produkte auf Facebook. Wer in einer solchen Gruppe ist, kann die zu testenden Produkte bei Amazon bestellen, hinterlässt eine positive Bewertung in deutscher Sprache und erhält im Gegenzug dann den Kaufpreis zurückerstattet. Oben drauf gibt es oft noch einen Bonus (Kaufmann, 2019-09).

Die Verlockung für Händler, zweifelhafte Methoden wie diese zu nutzen, ist offensichtlich. Wer täglich mehrere Verkäufe auf Amazon verzeichnen kann, die positiv bewertet werden, wird mit einem hohen Rangplatz in der Amazon-Suche belohnt. Das kann den Umsatz des Händlers schnell um mehrere 100 Prozent steigern.

Es wundert nicht, dass die Methode weiterhin genutzt wird, wie verschiedene Recherchen zeigen (Hitchins, 2019-04) – obwohl Amazon „das Erstellen, Ändern oder Posten von Inhalten im Austausch gegen eine Vergütung jeglicher Art (einschließlich kostenloser oder verbilligter Produkte) oder im Namen anderer" in den Community-Richtlinien ausdrücklich verbietet.

Apple und Google: Die Inflation der App-Bewertungen ist schlecht für Innovationen

Gaming-Apps werben um eine Bewertung vom Nutzer schon kurz nachdem dieser eine hohe Punktzahl im Game erreicht hat. Auch Banking-Apps scheinen Nutzer um Bewertungen zu bitten, wenn sie wissen, dass "Zahltag" ist. Und Sport-Apps geben den Anstoß zu bewerten, wenn das Team des Fans, der die App nutzt, gerade gewonnen hat.

App-Entwickler manipulieren uns mit Belohnungsreizen und nutzen unsere Stimmung gezielt, um ihr Ranking im App-Store zu verbessern. Die durchschnittliche Zahl der Bewertungen für Apps von Apple stieg von 19.000 im Jahr 2017 auf über 100.000 im Jahr 2019 an (McGee, 2020-09).

Der Auslöser für diese Inflation war ein scheinbar harmloses Update von Apple zur Stärkung des Verbraucher-Engagement im September 2017 – "In-App-Prompts". Die Nutzer mussten nicht mehr proaktiv in den App-Store gehen, um eine App zu bewerten – ein System, das oft nur frustrierte Nutzer anzog. Sie können Bewertungen nun auch bequem in der App selbst abgeben.

Im Gegensatz dazu kletterten die Bewertungen im Play Store von Google, der in diesem Zeitraum keine In-App-Bewertungen nutzte, von 33.000 auf nur 43.000 Bewertungen pro App. Erst 2020 wurden In-App-Reviews auch hier implementiert.

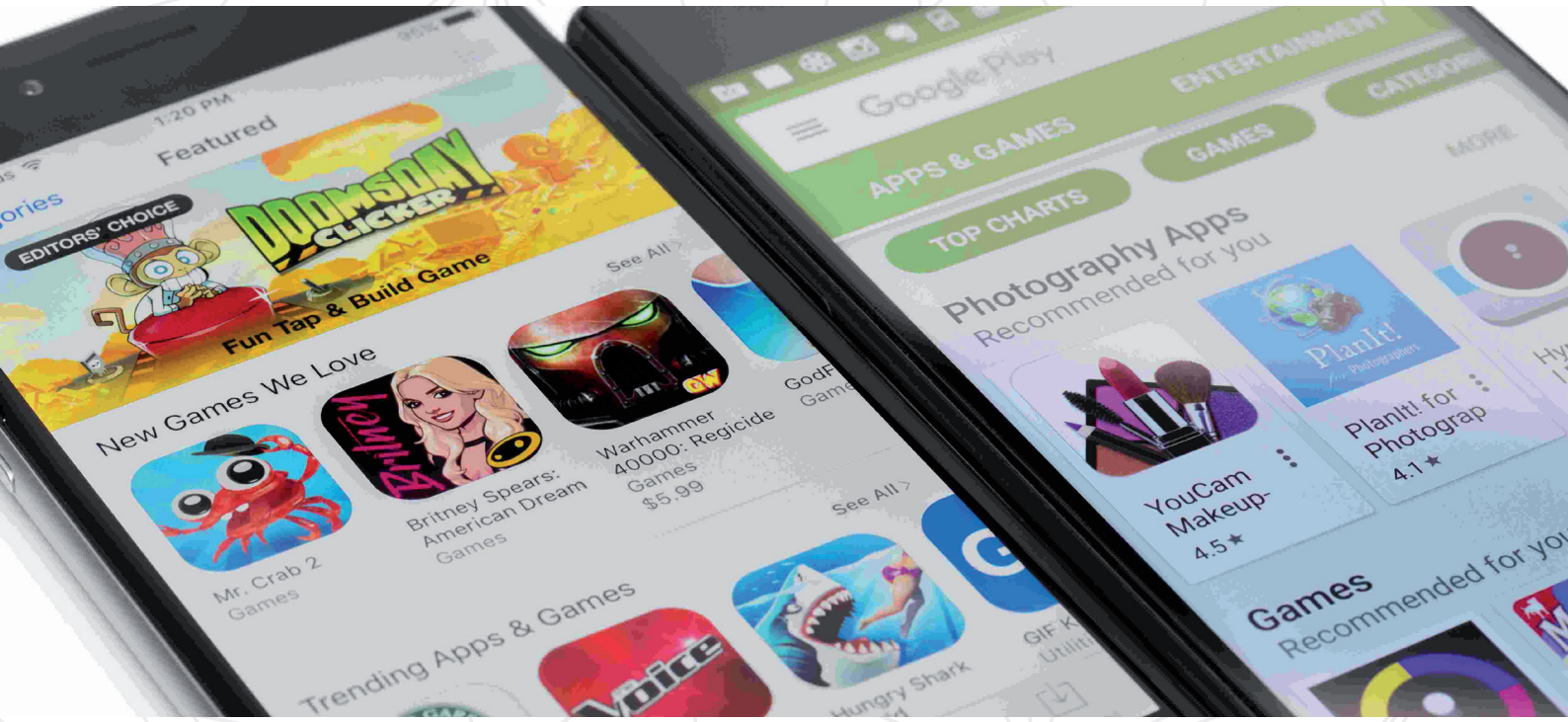
Das neue System ermöglicht es Entwicklern, Schlupflöcher auszunutzen und die Verbraucher dazu zu bringen, ihre Bewertungen aufzublähen. Beispielsweise können Entwickler einen Verbraucher vorbereiten, bevor er die App bewertet. Sie fragen den App-Nutzer sicher nicht nach einer Bewertung, wenn dieser gerade ein Spiel verloren oder eine schlechte Nachricht bekommen hat. Erst wenn der Nutzer nach der richtigen Vorbereitung mit positiven Reizen eine In-App-Bewertung vorgenommen hat, initiiert die App die "offizielle" Bewertung, die dann auch den App-Store erreicht.

Das Resultat ist, dass Entwickler von mittelmäßigen Apps dennoch eine Top-Bewertung erhalten können. Bessere Bewertungen erlauben es den Entwicklern dann auch noch sich auf ihren Lorbeeren auszuruhen, anstatt die Apps zu verbessern (McGee, 2020-09).

Übrigens zeigen die Daten auch, dass App-Bewertungen mit angehängten schriftlichen Ausführungen seit September 2017 keinerlei Inflation erfahren haben (Lin, 2020-08). Die Idee, dass mehr Sterne die bessere Qualität von Apps für iPhone oder iPad widerspiegeln, kann also nicht bestätigt werden. Es ist eher auf die langen schriftlichen Erklärungen zu achten.

Lin (2020-08) zitiert auch Entwickler, die das Tool dafür loben, dass es ihnen geholfen hat: „Nur eine Woche, nachdem wir In-App-Reviews implementiert haben, haben wir die höchste Bewertung aller Zeiten erreicht."

Dies ist ein klares Bekenntnis: Entwickler können einen Anstieg der Bewertungen durch In-App-Reviews erwarten – unabhängig davon, ob sie ihre App tatsächlich verbessert haben. Der Ansporn für kleine Verbesserungen und große Innovationen durch die realen Bedürfnisse und Wünsche der User scheint damit immer geringer zu werden.



Inflationäre Crowdsource-Bewertungen: Schlecht für Verbraucher?

Wenn kaum jemand auf einer Fünf-Punkte-Skala unter 3,5 rangiert, ist eine Bewertung von 3,5 schon ziemlich schlecht. Eine durchschnittliche Bewertung die "wirklich einen Unterschied macht" und nur etwas darunter liegt, kann für ein Mietobjekt jedoch verheerend sein oder den Fahrer direkt auf die Straße setzen. Wir wollen niemanden verletzen und erst recht nicht Grund dafür sein, dass jemand seine Lebensgrundlage verliert – aufgrund einer schlechten Bewertung.

Nicht nur gefälschte und manipulierte Bewertungen sind das große Problem. Die Beispiele oben zeigen, dass die Glaubwürdigkeit von Massenrezensionen durch die Funktionsweisen und Schwachstellen der Bewertungssysteme auf den verschiedenen Plattformen fast völlig aufgehoben werden kann.

Wie die Beispiele oben zeigen, wird die Aussagekraft der Bewertungen fast überall stark abgeschwächt. Das Prinzip der gegenseitigen Bewertungen ist für die Verbraucher, die sich informieren wollen, frustrierend, denn gegenseitige Bewertungen fallen vielfach positiver aus. Die Angst vor Diskriminierung schränkt die Nutzbarkeit noch weiter ein. Verbesserungen und Innovationen bleiben dabei auf der Strecke.

Zu guter Letzt muss hinterfragt werden, inwiefern die Kunden- oder Nutzerbewertungen repräsentativ sind in Bezug auf Variablen wie Einkommen, Alter, Beruf und Geschlecht. Legitime Bewertungen beim Online- oder Offline-Kauf können verzerrt sein, wenn bestimmte Personengruppen häufiger Bewertungen hinterlassen als andere. Vielbeschäftigte Menschen haben weniger Zeit Produktbewertungen zu schreiben und sind deshalb eher unterrepräsentiert.

Beispielsweise ist die Personengruppe, die ihre Meinung über Restaurants in den USA besonders aktiv austauscht, weitestgehend weiß und relativ wohlhabend. Film-, Fernseh- und Spielkritiken werden hingegen von jungen Männern dominiert (Hill, 2019-09). Ob hier die Zielgruppe des einzelnen Restaurant-Betreibers oder des Filmemachers immer angemessen repräsentiert wird, muss von Fall zu Fall überprüft werden.

Trotz der beschriebenen Probleme sind Kundenbewertungen für Hotels, Beratungsunternehmen oder gar Werkzeugmaschinenhersteller für uns interessant – speziell Bewertungen mit angehängten schriftlichen Ausführungen (Lin, 2020-08) oder mit selbstgedrehten Videos. Diese geben dem Unternehmen und uns als potenzielle Kunden Informationen, die beispielsweise auf Probleme in einem der Hotels einer Kette oder auf Missstände in Sachen Gastfreundlichkeit hinweisen. Likes und Sterne helfen vergleichsweise wenig. Eine genaue Beschreibung des Problems bringt uns deutlich weiter.

Das eigene Engagement, das heißt die Reaktion auf eine Bewertung als Unternehmen, ist hier der Schlüssel zum Erfolg. Bekommen wir ein ausführliches positives Feedback – zum Beispiel zu einer Newsletter-Mail privat oder öffentlich auf einer Website – ist ein Dankeschön für Unternehmen hier ein Pflichtprogramm.

Bei einer sehr schlechten Bewertung ist die Entschuldigung das Mindeste und die teilweise Wiedergutmachung des Fehlers das Ziel – bei einem Hotel zum Beispiel mit einem Gutschein für den nächsten Aufenthalt. Das hilft auch den zukünftigen Kunden, die sehen, dass Bewertungen beantwortet werden und Kunden souveräne Hilfestellungen bei Problemen oder kulante Ersatzleistungen bekommen.

Fazit: Daten im Digital Marketing von Studien bis zu Crowdsource Reviews

Man nennt dies die Zeit von Big Data. Doch von der Datenflut aus Berichten, Ranglisten, Statistiken und Studien (von oftmals fraglicher Qualität) werden wir geradezu erschlagen und die vielen ganzen und halben Sterne „rasen uns um die Köpfe". Einen Überblick zu erlangen, ist schwierig.

Eine Korrelation zwischen Video-Streaming und der Anzahl an Kinobesuchen bedeutet nicht, dass ein kausaler Zusammenhang besteht. Gute Kundenbewertungen für ein Unternehmen bedeuten nicht, dass diese Bewertungen den Tatsachen entsprechen. Positive Bewertungen sind wichtig, um konkurrenzfähig zu bleiben – ob gefälscht oder echt. Sie sichern einem Verkäufer auf Amazon einen guten Rangplatz in den Suchergebnissen der Plattformen.

Wenn wir Daten nicht im Kontext betrachten und von allen Seiten genau beleuchten, tappen wir eben weiterhin im Dunkeln (zumindest teilweise). Auch unsere Entscheidungen auf Basis dieser Daten – sowohl privat als auch im Beruf – können sich „als mindestens als sehr riskant entpuppen". Das A und O sind deshalb gut durchdachte, klare Definitionen, Regeln und Transparenz bei der Bewertung von Daten aus Studien und anderen Quellen. Aber auch bei der Entscheidungsfindung im Unternehmen sind diese essentiell.

Dieses Whitepaper hat aufgezeigt, mit welchen einfachen Mitteln die Zuverlässigkeit von Daten überprüft werden kann. Mithilfe von unserer Checkliste (siehe Seite 04/05) kann eine Studie sehr leicht auf Validität, Reliabilität und Objektivität überprüft werden. Wichtig ist: Man muss kein Wissenschaftler sein, um unplausible Datenerhebungen oder Statistiken und falsche Schlussfolgerungen zu durchschauen.

Auch bei Ranglisten, wie sie in Magazinen, Blogs und Whitepapers häufig veröffentlicht werden, lohnt es sich, anhand der Checkliste zu überprüfen, welche Kriterien wie gemessen wurden, um das Ranking zu erstellen. Ranglisten können nach diesem oder jenem Kriterium beliebig neu geordnet und in einem anderen Licht präsentiert werden. Können (reproduzierte) Daten in einem Whitepaper einer kritischen Prüfung wie dieser standhalten, handelt es sich wohl um einen sehr guten, zuverlässigen Report. Solche Reports oder Studienberichte gibt es, wenn auch weniger als wir uns dies wünschen.

Anders sieht es bei Kundenbewertungen aus: Sie sind so gut wie nie repräsentativ, denn vielbeschäftigte Menschen hinterlassen weniger Bewertungen, weil sie ganz einfach weniger Zeit dafür haben. Interessanter sind für uns als Verbraucher eher die längeren Texte der Bewertenden, die ein Produkt anhand positiver und negativer Aspekte auseinandernehmen. Natürlich sind die einzelnen Reviews sehr subjektiv, doch gerade das kann uns bei persönlichen Kaufentscheidungen auch helfen.

Sind die Daten richtig und vertrauenswürdig? Das müssen wir uns eben jedes Mal aufs Neue fragen – bevor wir dem Charme von George Clooney oder einer Influencerin erliegen und ihren Worten zu einem Produkt oder einer Studie trauen. Ein seriöser Faktencheck aus eigener Hand ist so gut wie immer nötig.

Takeaways: 10 wichtige Lektionen, um Daten besser nutzen zu können

1. Daten sind wichtig, doch sie müssen immer kritisch hinterfragt werden.
2. Das Aussehen einer Studie darf nicht über die Datenqualität hinwegtäuschen.
3. Studien und andere Datenquellen müssen immer einer Prüfung auf Validität, Reliabilität und Objektivität unterzogen werden.
4. Eine der wichtigsten Fragen ist: Woher kommen diese Daten eigentlich?
5. Klare Definitionen und transparente Regeln sind essentiell.
6. Soziale Erwünschtheit kann die Datenqualität stark beeinträchtigen.
7. Daten müssen im richtigen Kontext betrachtet werden.
8. Korrelation ist nicht gleich Kausalität.
9. Dank eigener Experimente und ausgewerteten Daten in der internen Datenbank des Unternehmens können bessere Entscheidungen getroffen werden.
10. Die besten Erkenntnisse ergeben sich aus der Kombination von Statistik und persönlicher Erfahrung.

Beispiel 1: STATISTIK AUSTRIA – höchst professionelle Studien

Die Informationen zur Methodik einer Studie sind essentiell. Die untere Erklärung ist ein gutes Beispiel, wie man mit wenig Text dem Leser einer Studie sehr genau erklären kann, was gemacht wurde und welche Methoden dabei zum Einsatz gekommen sind. Ein vorbildliches Beispiel ist das STATISTIK AUSTRIA Bundesanstalt Statistik Österreich, welches wichtige Informationen zur Methodik gleich in der Pressemitteilung liefert.

Methodische Informationen, Definitionen: Statistik Austria führte im Auftrag des Bundesministeriums für Gesundheit sowie der Bundesgesundheitsagentur von Oktober 2013 bis Juni 2015 eine auf der Europäischen Gesundheitsbefragung basierende österreichweite Erhebung zum Thema Gesundheit durch. Insgesamt wurden 15.771 zufällig ausgewählte Personen im Rahmen eines telefonischen Interviews und eines schriftlichen Fragebogens befragt. Die Ergebnisse sind repräsentativ für die Bevölkerung in Privathaushalten ab 15 Jahren (hochgerechnet 7,2 Mio. Personen).

*Themen der Befragung waren zum einen der **Gesundheitszustand** der Bevölkerung, also das Auftreten bestimmter Krankheiten und Gesundheitsprobleme, Schmerzen, funktionaler Beeinträchtigungen sowie das Ausmaß bzw. der Bedarf an Unterstützung bei Aktivitäten des täglichen Lebens. Ein zweiter Themenbereich betraf das **Gesundheitsverhalten**. Hier wurden Daten zu Risikofaktoren (Rauchen, Alkohol, Adipositas) sowie zu Ernährung, körperlicher Aktivität und Gesundheitsvorsorge erhoben. Ein dritter Aspekt betraf die **Inanspruchnahme von Gesundheitsleistungen**. Erstmals wurden in der Gesundheitsbefragung Informationen zum Gesundheitszustand er im Haushalt lebenden Kinder erhoben.*
Die Interviews wurden telefonisch und computerunterstützt durchgeführt (CATI-Computer Assisted Telephone Interviewing). Die Erhebung über die Lebensqualität und körperliche Aktivität erfolgte mittels eines selbst auszufüllenden Papierfragebogens im Anschluss an das Telefoninterview. Die österreichweite Ausschöpfung lag bei 40,7 Prozent. Die Teilnahme an der Befragung war generell freiwillig. Zu beachten ist ein möglicherweise sozial erwünschtes Antwortverhalten besonders bei einzelnen, möglicherweise sensiblen oder für die Zielperson unangenehmen Fragen, wie zum Alkoholkonsum oder zum Rauchen.

Es wurde auch sichergestellt, dass die Anzahl der "zufällig ausgewählten" Personen in schwach besiedelten Gebieten Österreichs im Verhältnis zu Wien genügend hoch war. Daher wurden in Wien über 4.000 Menschen befragt und in den Randregionen 450. Nur damit konnte sichergestellt werden, dass das Ergebnis auch für die schwach besiedelten Gebiete repräsentativ ist. Nur das erlaubte eine Unterteilung der 450 Antworten eines schwach besiedelten Gebietes zum Beispiel nach Geschlecht und Altersgruppen. Dabei waren die Untergruppen dann laut Statistik Austria groß genug für statistische Analysen und Vergleiche.
Das ist eines der vielen, guten Beispiele von STATISTIK AUSTRIA, wie man Definitionen und Methoden dem Publikum sachgerecht und verständlich erklärt. Einsicht unter: http://www.statistik.at/web_de/presse/105561.html

Ein **weiteres Beispiel von STATISTIK AUSTRIA vom 2020-12-11** gibt es hier: 4,7% der österreichischen Bevölkerung hatten Mitte/Ende Oktober Antikörper gegen SARS-CoV-2 (siehe: http://www.statistik.at/web_de/presse/124959.html).
Wieviel Arbeit hinter der Studie steckt und wie viel Organisation inklusive der Kosten, beschreibt die Sektion zur Stichprobenselektion hier: Bei einer Stichprobengröße von 7.823 in Privathaushalten wohnhaften Personen ab 16 Jahren beantworteten rund 2.711 Personen bis Ende Oktober einen Fragebogen. Davon vereinbarten 2.504 Personen einen Termin für eine umfangreiche Coronavirus-Testung (Nasen-Rachen-Abstrich, Antikörperschnelltest und Blutabnahme) bei einer der 53 Teststationen des Roten Kreuzes. Zwischen 12. und 14. November wurden die Testungen österreichweit durchgeführt.

Beispiel 2: Promarca Brand of the Year 2020 – Hintergrundinformationen unbefriedigend

Das Negativ-Beispiel unten ist wohl eines, das keine repräsentativen Daten anbietet. Die Studie ist sicherlich schwer zu replizieren (Reliabilität) und mittels ein oder zwei Fragen hat man das komplexe Konstrukt von Vertrauen in eine Marke wohl kaum gut eruiert.

Die öffentlich zugänglichen Informationen sagen sehr wenig über die Methodologie der Studie aus. In der Medienmitteilung heißt es unter anderem zur Methodik:
Promarca Brand of the Year 2020 *– Die vertrauenswürdigste Promarca-Marke aus dem Havas Brand Predictor [...] Zum sechsten Mal zeichnet Promarca die vertrauenswürdigste Marke aus den Reihen ihrer Mitglieder aus. Der Award «Promarca Brand of the Year 2020» geht an Ricola. Der Kräuterbonbonhersteller konnte sich zum zweiten Mal gegenüber seinen Konkurrenten durchsetzen [...]*
Die Brand Predictor Studie von Havas Schweiz und management tools research ist ein international etabliertes und repräsentatives Markenbewertungsverfahren und ermittelt seit 2012 effizient die Dynamik sowie das erarbeitete Vertrauen von Marken und macht zuverlässige Vorhersagen zu ihrer Zukunftsrelevanz. Die Abbildung der wahrgenommenen Dynamik und des verdienten Vertrauens stützt sich jeweils auf eine zentrale Schlüsselfrage. So führt die Fragestellung, ob eine Marke bezüglich ihrer Beliebtheit in der Bevölkerung an Boden gewinnt, an Boden verliert oder stagniert, rechnerisch zu einem Wert, der als Indikator der Markendynamik gilt. Die Evaluation der Vertrauenswürdigkeit basiert hingegen auf der Fragestellung, zu welchem Grad die Marke bisher ihre Leistungsversprechen erfüllen konnte.

Interessant ist, dass nur eine einzige Frage, die dann als Indikator der Markendynamik dienen soll, gestellt wird. Eine weitere Frage dient der Evaluation, zu welchem Grad die Marke bisher ihre Leistungsversprechen erfüllen konnte. Wie der Wert, der "als Indikator der Markendynamik" berechnet wird, ist nicht klar.

Dass der diesjährige Gewinner in einem Jahr von Platz 11 auf 1 vorrücken konnte, überrascht. Das Ranking scheint sehr großen Schwankungen zu unterliegen. Die Studie mag etabliert sein, wie die Medienmitteilung feststellt.

Aber ob es sich um ein repräsentatives Markenbewertungsverfahren handelt, kann mit diesen wenigen Angaben nicht überprüft werden. Deshalb sind die Resultate einer solchen Studie mit größerer Vorsicht zu genießen. Replizierbarkeit und Validität (Wird hier tatsächlich gemessen, was gemessen werden soll?) dieser Statistik darf mit Recht hinterfragt werden. Journalisten nutzen dennoch einfach die Pressemitteilung ohne die Daten zu hinterfragen (Stefano, 2020-06).
Ist ja schließlich "ein international etabliertes und repräsentatives Markenbewertungsverfahren."

Datensammlung mit Crowd-Plattformen – Beispiel 3: Amazon's Mechanical Turk (MTurk)

Die Durchführung psychologischer und Verbraucherorientierter Forschung über Online-Plattformen und Panels wie Amazon's Mechanical Turk (MTurk) wird sehr häufig genutzt. Nichtsdestotrotz kann die Arbeit mit diesem praktischen Pool von Teilnehmern eine einzigartige Reihe von Herausforderungen mit sich bringen. In Übereinstimmung mit früheren Ergebnissen zeigt eine Untersuchung von Ophir, Sisso, Asterhan, Tikochinski und Reichart (2020), dass bis zu 11 Prozent der MTurk-Teilnehmer, verglichen mit nur drei bis sieben Prozent der Gesamtbevölkerung in den USA, die diagnostischen Kriterien für eine schwere Depression erfüllen können.

Es ist natürlich interessant, die Gründe für die erhöhten Prävalenzraten von Depressionen zu kennen. Kenntnis dieser Gründe ist jedoch auch Voraussetzung für eine genaue Interpretation von Forschungsergebnissen, die mit Hilfe der MTurk-Teilnehmer generiert wurden.

In einer kürzlich durchgeführten Studie untersuchten Kim und Duhachek (2020) anhand einer MTurk-Stichprobe, wie sich Personen auf Informationen verlassen, die von nichtmenschlichen Agenten wie künstlicher Intelligenz und Robotern bereitgestellt werden. Insbesondere untersuchte die Studie, wie sich Überzeugungsversuche durch nichtmenschliche Agenten von Überzeugungsversuchen durch menschliche Agenten unterscheiden könnten.

Wie sich die Prävalenz von Depressionen bei den MTurk-Teilnehmern (siehe oben) auf diese Ergebnisse auswirken könnte, wird nicht untersucht. Auch der Fakt, dass die Teilnehmer auf der MTurk Plattform deutlich jünger wie auch mehr körperlich inaktiver sind und über eine schlechtere Schlafqualität berichten als die national repräsentative Stichprobe für die USA in der NHANES Studie bedarf einer genaueren Analyse (Ophir, Sisso, Asterhan, Tikochinski & Reichart, 2020).

Entscheidend ist, wie gut diese Datensets oder Panels die Bevölkerung repräsentieren oder reflektieren. Beispielsweise zeigen Polls vor Wahlen, dass die Dinge nicht immer so ablaufen wie die Pollster vorher eruierten. Zwei Tage vor den Wahlen (Oktober 15, 2020) in New Zealand, hat die Newshub Reid Research-Umfrage ergeben, dass Jacinda Ardern's Labour Partei 45,8 Prozent der Stimmen erhalten würde, was einem Rückgang von 4,3 Prozent entspricht. Die zweite Partei National, würde gemäß dieser Umfrage 31,1 Prozent erhalten, ein Plus von 1,5 Prozent (siehe https://www.nzherald.co.nz/nz/election-2020-labour-down-and-nz-first-up-in-final-poll/RUAOVF45J256TSZTXT34TE6NDA/).

Die tatsächlichen Ergebnisse waren denn ganz andere. Die Pollster lagen wieder einmal völlig daneben. Am Schluss erhielt die Labour Party 49,1 Prozent der Stimmen, verglichen mit 26,8 Prozent (Minus von 2,9 Prozent gegenüber den letzten Wahlen) für ihren größten Herausforderer, die konservative National Party. Die Labour-Partei gewann 64 der 120 Sitze im Einkammerparlament des Landes. Das ist die höchste Zahl von Parlamentariern seit Einführung des Verhältniswahlrechts in Neuseeland im Jahr 1996. Dies erlaubt der Labour Partei ohne Koalitionspartner zu regieren (https://www.bbc.com/news/world-asia-54519628).

Online-Panels aber auch Polls haben ihre Charakteristiken (beispielsweise Alter der Teilnehmer, Ausbildung, und Gesundheit), welche die Genauigkeit und Repräsentativität dieser Daten beeinflussen. Das Resultat sind dann Prognosen oder Studiendaten, die gewisse Probleme aufweisen. Das heißt, ungenau sind oder nicht genau der Zielgruppe entsprechen. Übrigens: Nicht jeder Pollster ist so transparent wie Colmar-Brunto hier: https://www.colmarbrunton.co.nz/what-we-do/1-news-poll/ Das erhöht vielleicht nicht die Genauigkeit der Resultate, aber sicherlich das Verständnis beim Leser und bei Journalisten, wie ungenau diese sein können.

PS: Wenn bis 11 Prozent der MTurk-Teilnehmer, verglichen mit nur drei bis sieben Prozent der Gesamtbevölkerung in den USA, die diagnostischen Kriterien für eine schwere Depression erfüllen können, wie wirkt sich dies auf deren Antworten aus? Anders ausgedrückt, haben andere Freelancer Crowdsourcing Plattformen, auf denen ich bestimmte Arbeiten wie ein Logo-Design vergeben kann, auch Faktoren (z.B. Gewicht, Alter, usw.) welche diese von der Bevölkerung unterscheiden?
Und wenn ja, wie wirken sich diese auf die Arbeit aus?

Stichproben-Selektion und Datenqualität
Wie setzen wir unsere Stichprobe zusammen?

Nehmen wir an, wir wollen wissen, wie gut unser Produkt bei Kunden ankommt. Vielleicht müssen wir aber auch herausfinden, inwiefern wir die Markenstärke oder die Nutzung von Hashtags, wie zum Beispiel #JustDoIt, #KitKatBreak, #drkpiPageTracker, aus Kampagnen verbessern konnten.

Egal, was das Ziel ist, wir brauchen vertrauenswürdige Daten zu einem Preis (zum Beispiel Budget, Zeit und Inanspruchnahme limitierter Ressourcen wie Personal), den wir uns leisten können.

Im Folgenden ein **Beispiel** für eine Datenerhebung.
Hier brauchen wir ein Optimum oder einen Wert für Geld (Value for Money):

- **Geschlecht (vier Gruppen wie Frau, Mann, LGBT, andere Gruppe etc.),**
- **Sprach- und/oder geographische Region (vier oder fünf Gruppen in der Schweiz: Deutsch, Französisch, Italienisch, andere Sprachen wie beispielsweise Englisch, Rätoromanisch etc.),**
- **Arbeit: Stunden pro Woche,**
- **Ausbildung: keine abgeschlossene Ausbildung/Lehre, abgeschlossene Ausbildung/Lehre, Abitur/Matura, Fachhochschul-/Universitäts-Abschluss (Bachelor, Master) oder**
- **Anzahl der Schuljahre (neun Jahre Grundschule, drei oder vier Jahre duale Ausbildung, etc.)**

Für jede Gruppe wollen wir vielleicht am Ende gut 50 Teilnehmer. Doch für eine repräsentative Stichprobe ist es notwendig, dass ein Bewohner von Hamburg dieselbe Chance hat selektioniert zu werden wie ein Bewohner aus Oberbayern.

Dies bedeutet zum Beispiel in der Schweiz, dass circa 6,5 Prozent der Teilnehmer die italienische Sprache als Erstsprache haben. Die Folge ist, dass wir fast zehn Mal mehr deutschsprachige Schweizer in einer Gruppe haben. Geschlecht (vier Gruppen) und Sprache (vier Gruppen) ergibt 16 Gruppen. Wenn wir in jeder dieser Gruppen mindestens 50 Italienisch sprechende Teilnehmer haben wollen, brauchen wir alleine für die Variable Geschlecht 200 Teilnehmer mit Italienisch als primärer Sprache/Muttersprache. Das heißt, 50 mit Italienisch als Hauptsprache bedingt die Auswahl von gut 550 deutschsprachigen Teilnehmern x 4 = gut 1.850 Teilnehmer die primär Deutsch sprechen (siehe auch: https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/sprachen-religionen/sprachen.html).

Es überrascht deshalb nicht, dass eine repräsentative Studie in der Schweiz schnell über 10.000 Teilnehmer haben kann, um der zufälligen Selektion der Teilnehmer - beispielsweise gemäß Bevölkerungsanteil nach Landessprache – gerecht zu werden. Dies sprengt oft den Rahmen der Möglichkeiten für das Marketing in einem Unternehmen. Dies ist speziell dann der Fall, wenn mehrere Länder in die Studie einbezogen werden müssen.

Stichproben-Selektion und Datenqualität
Wie können wir eine optimale Datenqualität sichern?

Datenqualität ist die Basis zum Erfolg. Wenn ein Unternehmen Daten von seinen Kunden sammelt, muss die Stichprobe die weite Bandbreite der Kunden sehr gut repräsentieren. Das heißt, gemäß Geschlecht, Altersgruppe, Land, Einkommensgruppe etc. müssen die gesammelten Daten diejenigen der wirklichen Kunden genau widerspiegeln.

Man kann Datenqualität aber auch etwas formaler beschreiben. Grundsätzlich spielen sicherlich vier Faktoren eine wichtige Rolle, ob unsere Big-Data-Sammlung von guter Qualität oder eben nicht ist:

- **Richtigkeit (Veracity) der Daten**
  Sind die Daten für das Problem oder die gestellten Fragen angemessen?

- **Zuverlässigkeit und Replizierbarkeit (Reliability) der Daten**
  Inwiefern kommen zum Beispiel zwei Personen zur gleichen Beurteilung eines Produktes oder einer Dienstleistung?

- **Gültigkeit (Validität) der Daten**
  Sind die Daten valide und messen diese, was sie messen sollen - mit hoher Qualität?

- **Beständigkeit (Volatilität) der Daten**
  Wie lange sind die Daten noch aktuell? Beispielsweise ist bei einer Voraussage einer Grippe-Epidemie, eines Verkehrsstaus oder eines Vulkanausbruches die Beständigkeit relativ tief.

Wenn wir Daten sammeln oder uns die Resultate einer Studie anschauen, lohnt es sich sicherlich auf diese drei Dinge zu achten:

1. **Kulturelle und regulatorische Unterschiede:** Daten wurden zum Beispiel in den USA gesammelt und jemand hat Rückschlüsse daraus über Kundenverhalten in der D-A-CH-Region gemacht. Macht das tatsächlich Sinn?

2. **Methodik:** Zusammen gemischt – das heißt, die einen Daten sind aus Land A und die anderen aus Land B mit anderer Altersgruppe. Ist das angemessen oder sind Kunden aus Österreich wirklich anders als diejenigen aus dem Tessin?

3. **Kontroll-Daten.** Das heißt, es wurden verschiedenen Gruppe in den Test einbezogen. Beispiel ist die erste Gruppe erhält das neue Medikament, die zweite Gruppe die Placebo-Pille ohne Wirkung, die dritte Gruppe erhält kein Medikament und die vierte Gruppe bekommt ein Konkurrenzprodukt. Dies hilft genau zu eruieren, inwiefern das neue Produkt die Krankheitssymptome bekämpft.

Ganz grundsätzlich sollten wir uns Gedanken machen, ob Daten aus Land A auch relevant für Land B sind.

Auch die **Unternehmensgröße** kann bei Umfragen eine Rolle spielen. Das heißt, die Antwort einer Führungskraft im Marketing in einem DAX.-Unternehmen wird sich vielleicht von derjenigen eines KMU (kleine oder mittlere Unternehmen) stark unterscheiden.

Laut der Definition der Europäischen Union und dem statistischen Bundesamt (Deutschland) beschäftigen KMU bis 249 Vollzeitbeschäftigte und haben bis 50 Mio. Euro Umsatz. In der Schweiz sind fast 90 Prozent aller Unternehmen Mikrobetriebe mit bis zu 9 Vollzeitbeschäftigten und 8,5 Prozent kleine Unternehmen mit bis zu 49 Vollzeitbeschäftigten (siehe auch https://drkpi.com/ransomware/).

Wenn wir zum Beispiel eine Umfrage machen wollen, um mehr über die Trends in der Rekrutierung von Personal herauszufinden, muss uns klar sein, welche Personen aus welchen Unternehmen wir befragen wollen. Beispielsweise sind unsere Kunden eher KMU- oder DAX- Unternehmen. Je nach Fall sollten wir uns auch bemühen, dass dies in unserer Stichprobe reflektiert wird.

KMU sind vielleicht weniger bereit einen externen Dienstleister in ihre Rekrutierungsarbeit mit einzubeziehen, als vielleicht ein DAX-Unternehmen.Es gibt aber auch Beispiele aus der Online-Werbung aus den USA, die aufhorchen lassen, denn diese zeigen starke Unterschiede im Verhalten von Konzernen mit KMU auf. Beispielsweise hat die Kampagne “Black Lives Matter” dazu geführt, dass Großbetriebe wie Unilever oder The North Face wieder einmal für eine gewisse Zeit Werbung auf Facebook gestoppt haben - ganz im Gegensatz zu KMU (Murphy, 2020-06-27/28).

Referenzliste

Apptentive (Publikation nicht datiert, Datenerhebung 2018). Mobile app ratings and reviews: Where to start and how to win. Aufgerufen 12.09.2020 auf: https://explore.apptentive.com/c/Ratings-and-Reviews_PDF_V1?x=5DpafR&#zoom=page-fit

Bundeskartellamt (2020-10). Gefälschte und manipulierte Nutzerbewertungen beim Online-Kauf. Sektoruntersuchung Nutzerbewertungen – Abschlussbericht. Aufgerufen am 15.10.2020 auf: https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Sektoruntersuchungen/Sektoruntersuchung_Nutzerbewertungen_Bericht.pdf?__blob=publicationFile&v=3
Fragebogen als Download: https://www.bundeskartellamt.de/Nutzerbewertungen_Produkt.html?nn=13060878

DIN ISO 20488:2019-03 (2019-03). Online-Kundenbewertungen – Grundsätze und Anforderungen für die Erhebung, Moderation und Veröffentlichung (ISO 20488:2018). Aufgerufen 10.10.2020 auf: https://beuth.de/de/norm/din-iso-20488/300183899

EASO (2020). Asylum Report 2020. Annual report on the situation of asylum in the European Union. European Asylum Support Office. DOI: 10.2847/531878. Aufgerufen am 12. Dez. 2020 auf https://easo.europa.eu/sites/default/files/EASO-Asylum-Report-2020.pdf

Enneking Ulrich, Kleine-Kalmer, Ruth, Dauermann, Angelika & Voigt, Rieke (2019-01). Kaufbereitschaft bei verpackten Schweinefleischprodukten im Lebensmitteleinzelhandel – Realexperiment und Kassenzonen-Befragung. Hochschule Osnabrück. Aufgerufen am 29.06.2020 auf: https://www.hs-osnabrueck.de/fileadmin/HSOS/Homepages/Personalhomepages/Personalhomepages-AuL/Enneking/Tierwohlstudie-HS-Osnabrueck_Teil-Realdaten_17-Jan-2019.pdf

Gattiker, Urs E. (2020-10-26). 1CSX Credit Suisse Digital Bank: Top oder Flop. [Blog Eintrag - drkpi]. Aufgerufen am 10 Oktober 2020 auf: https://test.drkpi.ch/csx-credit-suisse-digital-banking-usability-test/

Gattiker, Urs E. (2020-05-18). Datengesteuertes Marketing: Datenqualität und NZZ Journalismus. [Blog Eintrag - drkpi]. Aufgerufen am 10 Oktober 2020 auf: https://drkpi.com/en/datengesteuertes-marketing-show-me-the-numbers-1/

Gattiker, Urs E.; Temmen, Taina: Sinistra, Patrizia (2018-04, 2. rev. Auflage). EU-Datenschutzgrundverordnung (DSGVO): Was ist Sache für Marketing Manager, Geschäftsleitung und Vorstand? Whitepaper. Düsseldorf: Deutscher Marketing Verband e.V. (DMV). Aufgerufen am 2019-08-31 auf: https://MCLago.com/download/13/

Harford, Tim (2020-09-20). Statistics, lies and the virus: Tim Harford's five lessons from a pandemic. Financial Times, online. Aufgerufen am 2. Dezember, 2020 auf: https://www.ft.com/content/92f64ea9-3378-4ffe-9fff-318ed8e3245e

Hill, Andrew (2019-09-09). Rating systems return to haunt the gig economy. Financial Times, Work & Careers, S. 12. Aufgerufen am 1.06.2020 auf: https://www.ft.com/content/eb8b7c0e-ce71-11e9-99a4-b5ded7a7fe3f

Hitchins, Nathalie (2019-04-19). Revealed: Amazon struggling to stem the tide of fake reviews. Which? (UK Consumer Association). Aufgerufen am 7.07.2020 auf: https://press.which.co.uk/whichpressreleases/revealed-amazon-struggling-to-stem-the-tide-of-fake-reviews/

Hoogeveen, Suzanne, Sarafoglou, Alexandra & Wagenmakers, Eric-Jan (2020-09). Laypoepole can predict which social-science studies will be replicated successfully. Advances in Methods and Practices in Psychological Science, 3(3), pp. 267–285. DOI: https://doi.org/10.1177/2515245920919667

Kaufmann, Moritz (2019-09-22). Chinesen unterwandern Amazon. Mit illegalen Tricks bauen Billig-Händler ihre Vermachtstellung beim weltgrößten Online Portal aus. NZZ am Sonntag, S. 26. Aufgerufen am 4.06.2020 auf: https://nzzas.nzz.ch/wirtschaft/amazon-wird-von-den-china-haendlern-unterwandert-ld.1510374?reduced=true

Lee, Dave & Murphy, Hannah (2020-08-13). Technology. Facebook groups trading fake Amazon reviews still in business. Financial Times, Companies & Markets, S. 6. Aufgerufen am 14.08.2020 auf: https://www.ft.com/content/d4af6504-924e-4f94-b82e-0f02671faa12

Lin, Scott (2020-08-05). Leverage the In-App review API for your Google Play reviews. [Blog Eintrag - Android Developers]. Aufgerufen am 1.09.2020 auf: https://android-developers.googleblog.com/2020/08/in-app-review-api.html

Luca, Michael & Zervas, Georgios (2015-05-01). Fake it til you make it: Reputation, competition, and Yelp review fraud. Harvard Business School NOM Unit Working Paper No. 14-006. DOI: http://dx.doi.org/10.2139/ssrn.2293164

Mauró, Helmut (2015-02-09). Die besten Konzertsäle der Welt: Hier drin spielt die Musik. Süddeutsche (online). Aufgerufen am 12. Dez. 2020 auf https://www.sueddeutsche.de/kultur/die-besten-konzertsaele-der-welt-hier-drin-spielt-die-musik-1.2339789

McGee, Patrick (2020-09-08). Why app inflation is bad for consumers. FT Big Read. The Mobile Economy. Financial Times, S. 17. Aufgerufen am 11.09.2020 auf: https://www.ft.com/content/217290b2-6ae5-47f5-b1ac-89c6ccebab41

Moore, Elaine (2020-01-19). Why can't we resist online reviews. Financial Times, FT Weekend, Life & Arts, S. 17. Aufgerufen am 10.05.2020 auf: https://www.ft.com/content/1b2bbd7a-365f-11ea-a6d3-9a26f8c3cba4

Murphy, Hanna (2020-06-26). Unilever pulls Facebook and Twitter ads on hate speech concerns. Financial Times. Aufgerufen am 24. Dez. 2020 auf https://www.ft.com/content/5e9624fa-d121-44a9-ba84-d456385e50ab

Ophir, Sisso, Asterhan, Tikochinski und Reichart (2020), Kim, Tae Woo & Duhachek, Adam (2020-04). Artificial Intelligence and Persuasion: A Construal-Level account. Psychological Science, 31(4), S. 363–380
Aufgerufen am 12. Dez. 2020 auf https://doi.org/10.1177/0956797620904985

Ophir, Aakov, Sisso, Itay, Asterhan, Christa S. C., Tikochinski, Refael & Reichart, Roi (2020-01). The Turker blues: Hidden factors behind increased depression rates among Amazon's Mechanical Turkers. Clinical Psychological Science, 8(1), S. 65-83. Aufgerufen am 12. Dez. 2020 auf https://doi.org/10.1177/2167702619865973 auch auf https://www.researchgate.net/publication/336012041_The_Turker_Blues_Hidden_Factors_behind_Increased_Depression_Rates_in_Amazon%27s_Mechanical

Pressemitteilung: Promarca Brand of the Year 2020 Die vertrauenswürdigste Promarca-Marke aus dem Havas Brand Predictor (2020-06-04). Aufgerufen am 6 Juni 2020 auf: https://documentcloud.adobe.com/link/track?uri=urn:aaid:scds:US:57507c8e-d0a1-4b7b-bfdf-bfc64f6e508d#pageNum=1

Stefano, Roberto (2020-06-04). Nachhaltiges Vertrauen. Markenartikel. Handelszeitung, Special Marken, S. 25. Aufgerufen am 6 Juni 2020 auf: https://www.promarca.ch/wp-content/uploads/2020/06/Markenspecial-1.pdf

World Bank (2019-05-24). Doing Business 2020. Washington DC: World Bank. Aufgerufen am 12.Dezember 2020 auf https://www.doingbusiness.org/en/reports/global-reports/doing-business-2020

Zervas, Georgios, Proserpio, Davide & Byers, John (2015-01-28). A first look at online reputation on Airbnb, where every stay is above average. SSRN (online). DOI: http://dx.doi.org/10.2139/ssrn.2554500

Zhang, Dennis, Dai, Hengchen, Dong, Lingxiu, et.al. (2018-12). How do price promotions affect customer behaviour on retailing platforms? Evidence from a large randomised experiment on Alibaba. Management Science, 27(12), 2343-2345. Aufgerufen 7.09.2020 auf: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3029707

Autor und Leiter des CoCi:
Professor Urs E. Gattiker Ph.D.
drkpi® CyTRAP Labs GmbH
Präsident MCLago

Autorin:
Patrizia Sinistra
Vault Security Systems AG
Vorstand MCLago

Autorin:
Johanna Babuzki
CONRADY GRUPPE
Verwaltungs GmbH

Autorin und Leiterin des CoCi:
Taina Temmen
Vorstand DMV
COO & Co-Founder
EDITIVE

Kontakt DMV-Geschäftsstelle
Telefon: 0211.864 06-0
competence@marketingverband.de

Kontakt Competence Circle
Taina Temmen
temmen@marketingverband.de

Urs E. Gattiker
gattiker@marketingverband.de

Competence Circle

Die neun Competence Circle bilden eine inhaltliche Themen- und Kompetenz-Plattform für den DMV und sorgen mit ihrer Expertise u.a. durch die Erstellung der Whitepapers für einen Know-how Transfer auf allen Ebenen des Deutschen Marketing Verbands. Die einzelnen Gruppen stehen für folgende neun Themen:

1 **Bewegtbild**
2 **Customer Excellence**
3 **Data Driven Marketing & Decision Support Pricing**
4 **Employer Branding**
5 **Markenmanagement**
6 **Marketingplanung und -optimierung**
7 **Pricing & Market Strategy**
8 **Sponsoring**
9 **Technologie, Innovation & Management #ccTIM**

Whitepaper #ccTIM

Gattiker, Urs E. & Temmen, Taina (2020-01). Blockchain-Technologie: Wie es Die Lieferkette und das Marketing verändert. Duesseldorf: Deutscher Marketing Verband e.V. (DMV). https://MCLago.com/download/41/

Gattiker, Urs E.; Temmen, Taina; Sinistra, Patrizia (2019-01). Künstliche Intelligenz: Roboter Lisa räumt die Küche auf und jobbt als Wirtschaftsprüfer. Whitepaper. Duesseldorf: Deutscher Marketing Verband e.V. (DMV). https://MCLago.com/download/30/

Gattiker, Urs E., Temmen, Taina, Sinistra, Patrizia (2018-04, 2. rev. Auflage). EU-Datenschutzgrundverordnung (DSGVO): Was ist Sache für Marketing Manager, Geschäftsleitung und Vorstand? Whitepaper. Düsseldorf: Deutscher Marketing Verband e.V. (DMV). https://MCLago.com/download/13/

40+ Whitepaper der Competence Circle des Deutschen Marketing Verband:
https://www.marketingverband.de/marketingkompetenz/competence-circles/

Impressum

Herausgeber
Deutscher Marketing Verband e.V. (DMV)
Sternstrasse 58, D-40479 Düsseldorf
Fon +49 (0) 211.864 06-0
info@marketingverband.de
marketingverband.de

Bildrechte: Adobe Stock

ISSN (Print) 2512-5842
ISSN (Online) 2512-5656