Von Daten zu Personas: Eine strategische Kundensegmentierung für Travel Tide

Eine datengestützte Kundensegmentierung für Travel Tide

Mastery Projekt | Masterschool

[Interaktive Analyse und Implementierung (Google Colab)]

von Sinem Ara-Yücel | September 2025

Management Summary

Ziel dieses Projekts war es, eine datengestützte Grundlage für personalisiertes Marketing bei Travel Tide zu schaffen, indem der Kundenstamm in handhabbare Gruppen segmentiert wird. Mittels unüberwachten maschinellem Lernen wurden erfolgreich fünf distinkte Kunden-Personas identifiziert, die von selten buchenden „Effizienten Städtereisenden“ bis hin zu hochprofitablen „Premium-Familienplanern“ reichen.

Jede Persona weist einzigartige Verhaltensweisen und Bedürfnisse auf, was einen maßgeschneiderten Marketingansatz für das neue Prämienprogramm ermöglicht. Die zentrale Handlungsempfehlung ist, diese Erkenntnisse für gezielte E-Mail-Kampagnen zu nutzen, indem für jedes Segment die jeweils passendsten Vorteile beworben werden. Diese personalisierte Strategie zielt darauf ab, das Kundenengagement, die Loyalität und letztlich den Lifetime Value unserer Kunden signifikant zu steigern.

Business Case & Projektziele

Business Case

Der Online-Reisemarkt ist hart umkämpft. Die Gewinnung von Neukunden ist kostspielig, weshalb die Steigerung der Kundenbindung und die Maximierung des Lifetime Values bestehender Kunden von strategischer Bedeutung für Travel Tide sind.

Um ein effektives, personalisiertes Prämienprogramm zu entwickeln, das die Kundenloyalität nachhaltig erhöht, müssen wir zunächst ein tiefes, datengestütztes Verständnis für unsere Kunden entwickeln. Die zentrale geschäftliche Herausforderung besteht darin, die Frage zu beantworten: "Welche unterschiedlichen Kundentypen nutzen unsere Plattform und was charakterisiert sie?" Dieses Projekt wird die notwendigen Erkenntnisse liefern, um diese Frage zu beantworten und damit die Grundlage für gezielte, personalisierte Marketingmaßnahmen zu schaffen.

Projektziel: Entwicklung einer datengestützten Kundensegmentierung als strategische Grundlage für personalisiertes Marketing

Um unsere Kunden effektiver und persönlicher ansprechen zu können, segmentieren wir in diesem Projekt unseren gesamten Kundenstamm in klar definierte Gruppen. Ziel ist es, von einer undifferenzierten Massenansprache wegzukommen und stattdessen maßgeschneiderte Angebote zu schaffen, die exakt auf die Bedürfnisse der jeweiligen Kundengruppe zugeschnitten sind.

Als konkretes Ergebnis wird ein praktischer Leitfaden entwickelt, der 3 bis 5 zentrale Kunden-Personas vorstellt. Für jede dieser Personas werden die folgenden drei Kernfragen beantwortet:

- Wer ist der Kunde? (Ein prägnanter Name und eine griffige Kurzbeschreibung, z.B. "Der effiziente Städtereisende").
- Wie verhält er sich? (Die wichtigsten Merkmale zu Demografie, typischem Buchungsverhalten, Budget und bevorzugten Reisearten).
- Wie relevant ist die Gruppe? (Ihr prozentualer Anteil an unserer gesamten Kundenbasis zur Einschätzung des Marktpotenzials).

Diese Ausarbeitung liefert dem Marketing-Team eine sofort einsetzbare Grundlage, um die Kommunikation, insbesondere für das neue Prämienprogramm, zielgerichtet zu gestalten und die Konversionsraten signifikant zu erhöhen.

Datengrundlage: Schema, Struktur und initiale Qualitätsanalyse

Datenquelle und Aufbereitung

Die Analyse basiert auf internen Produktionsdaten von Travel Tide aus den letzten Jahren. Für die Segmentierung wurde dieser Datensatz gezielt auf eine relevante Kundengruppe gefiltert.

Berücksichtigt wurden ausschließlich aktive Nutzer mit **mehr als 7 Seitenklicks**, deren Interaktionen **nach dem 4. Januar 2023** stattfanden, um eine Analyse des aktuellen Kundenverhaltens zu gewährleisten. Die Datengrundlage bilden die Tabellen: users, sessions, flights und hotels.

Zeilenanzahl je Datensatz in den Rohrtabellen:

- User: 1.020.926
- hotels: 1.918.617
- flights: 1.901.038
- sessions: 5.408.063

Überprüfung der Nullwerte, um diese zu berücksichtigen.

- Users: keine NULL-Werte
- hotels: keine NULL-Werte
- flights:
    - return_time = 88.734
- sessions:
    - tripd_id = 3.072.218
    - flight_discount_amount = 4.522.267
    - hotel_discount_amount = 4.716.683

Datensatz

Nachfolgend eine Beschreibung der einzelnen Tabellen und ihrer Spalten:

Benutzer : demografische Informationen zum Benutzer

- user_id : eindeutige Benutzer-ID (Schlüssel, int)
- Geburtsdatum : Geburtsdatum des Benutzers (Datum/Uhrzeit)
- Geschlecht : Geschlecht des Benutzers (nominal)
- verheiratet : Familienstand des Benutzers (binär)
- has_children : ob der Benutzer Kinder hat oder nicht (binär)
- home_country : Wohnsitzland des Benutzers (nominal)
- home_city : Wohnort des Benutzers (nominal)
- home_airport : bevorzugter Heimatflughafen des Benutzers (nominal)
- home_airport_lat : geografische Nord-Süd-Position des Heimatflughafens (dezimal)
- home_airport_lon : geografische Ost-West-Position des Heimatflughafens (dezimal)
- sign_up_date : Datum der TravelTide-Kontoerstellung (Datum/Uhrzeit)

Sitzungen : Informationen zu einzelnen Browsersitzungen (Hinweis: Es werden nur Sitzungen mit mindestens 2 Klicks berücksichtigt)

- session_id : eindeutige Browser-Sitzungs-ID (Schlüssel, Zeichenfolge)
- user_id : die Benutzer-ID (Fremdschlüssel, int)
- trip_id : ID, die Flug- und Hotelbuchungen zugeordnet ist (Fremdschlüssel, Zeichenfolge)
- session_start : Zeitpunkt des Beginns der Browsersitzung (Zeitstempel)
- session_end : Zeitpunkt des Endes der Browsersitzung (Zeitstempel)
- flight_discount : ob ein Flugrabatt angeboten wurde oder nicht (binär)

- hotel_discount : ob ein Hotelrabatt angeboten wurde oder nicht (binär)
- flight_discount_amount : Prozentsatz vom Basistarif (Dezimalzahl)
- hotel_discount_amount : Prozentsatz vom Basisnachtpreis (Dezimal)
- flight_booked : ob der Flug gebucht wurde oder nicht (binär)
- hotel_booked : ob das Hotel gebucht wurde oder nicht (binär)
- page_clicks : Anzahl der Seitenklicks während der Browsersitzung (int)
- Stornierung : ob der Zweck der Sitzung darin bestand, eine Reise zu stornieren (binär)

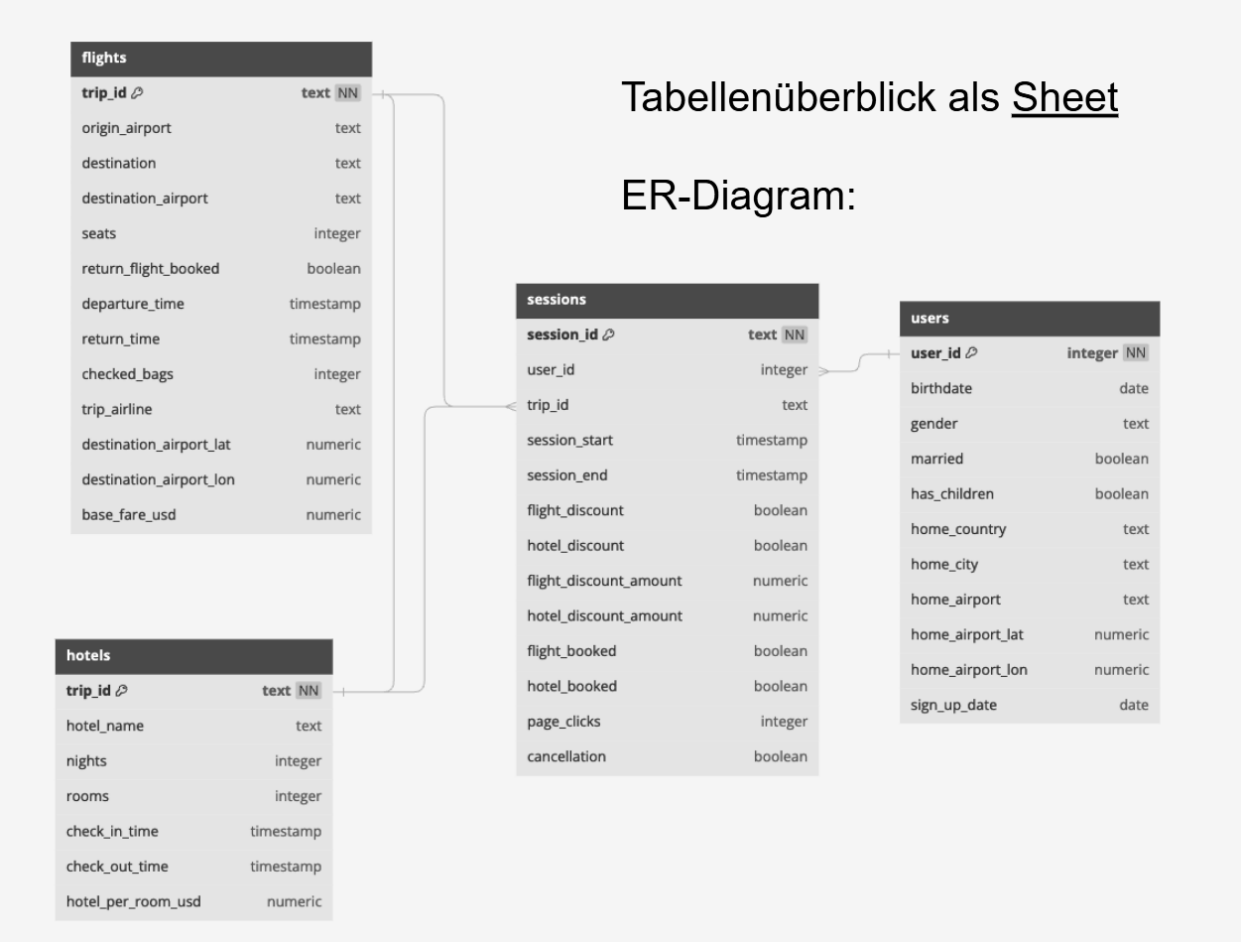
Flüge : Informationen zu gekauften Flügen

- trip_id : eindeutige Reise-ID (Schlüssel, Zeichenfolge)
- origin_airport : Heimatflughafen des Benutzers (nominal)
- Zielort : Zielort (nominal)
- destination_airport : Flughafen in der Zielstadt (nominal)
- Sitzplätze : Anzahl der gebuchten Sitzplätze (int)
- return_flight_booked : ob ein Rückflug gebucht wurde oder nicht (binär)
- departure_time : Abflugzeit vom Abflughafen (Zeitstempel)
- return_time : Zeitpunkt der Rückkehr zum Ausgangsflughafen (Zeitstempel)
- checked_bags : Anzahl der aufgegebenen Gepäckstücke (int)
- trip_airline : Fluggesellschaft, die den Benutzer vom Abflug- zum Zielort bringt (nominell)
- destination_airport_lat : geografische Nord-Süd-Position des Zielflughafens (dezimal)
- destination_airport_lon : geografische Ost-West-Position des Zielflughafens (dezimal)
- base_fare_usd : Preis des Flugpreises vor Rabatt (Dezimalzahl)

Hotels : Informationen zu gekauften Hotelaufenthalten

- trip_id : eindeutige Reise-ID (Schlüssel, Zeichenfolge)
- hotel_name : Hotelmarkenname (nominal)
- Nächte: Anzahl der im Hotel verbrachten Nächte (int)
- Zimmer : Anzahl der beim Hotel gebuchten Zimmer (int)
- check_in_time : Beginn des Hotelaufenthalts des Benutzers (Zeitstempel)
- check_out_time : Uhrzeit, zu der der Hotelaufenthalt des Benutzers endet (Zeitstempel)
- hotel_per_room_usd : Preis für den Hotelaufenthalt pro Zimmer und Nacht vor Rabatt (Dezimal)

Datenbankschema

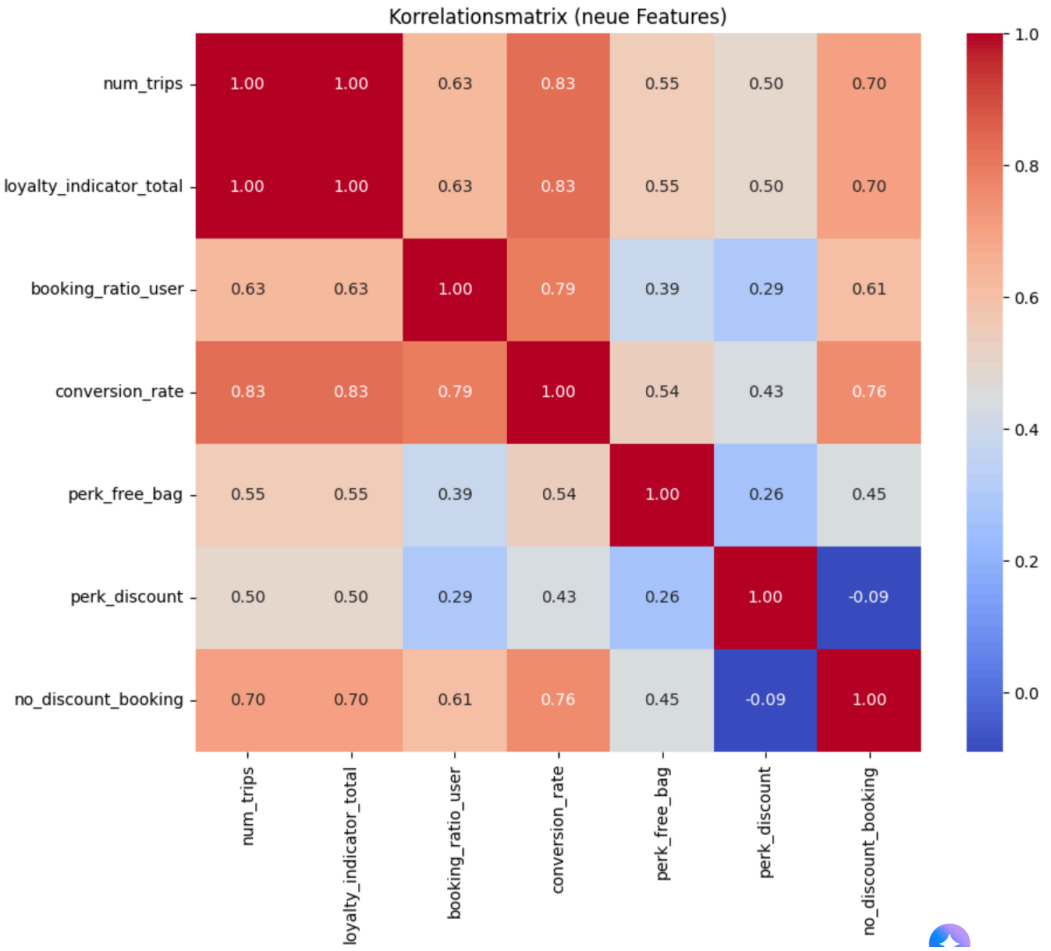


Vorgehensweise und Methodik

Angewandte Methoden

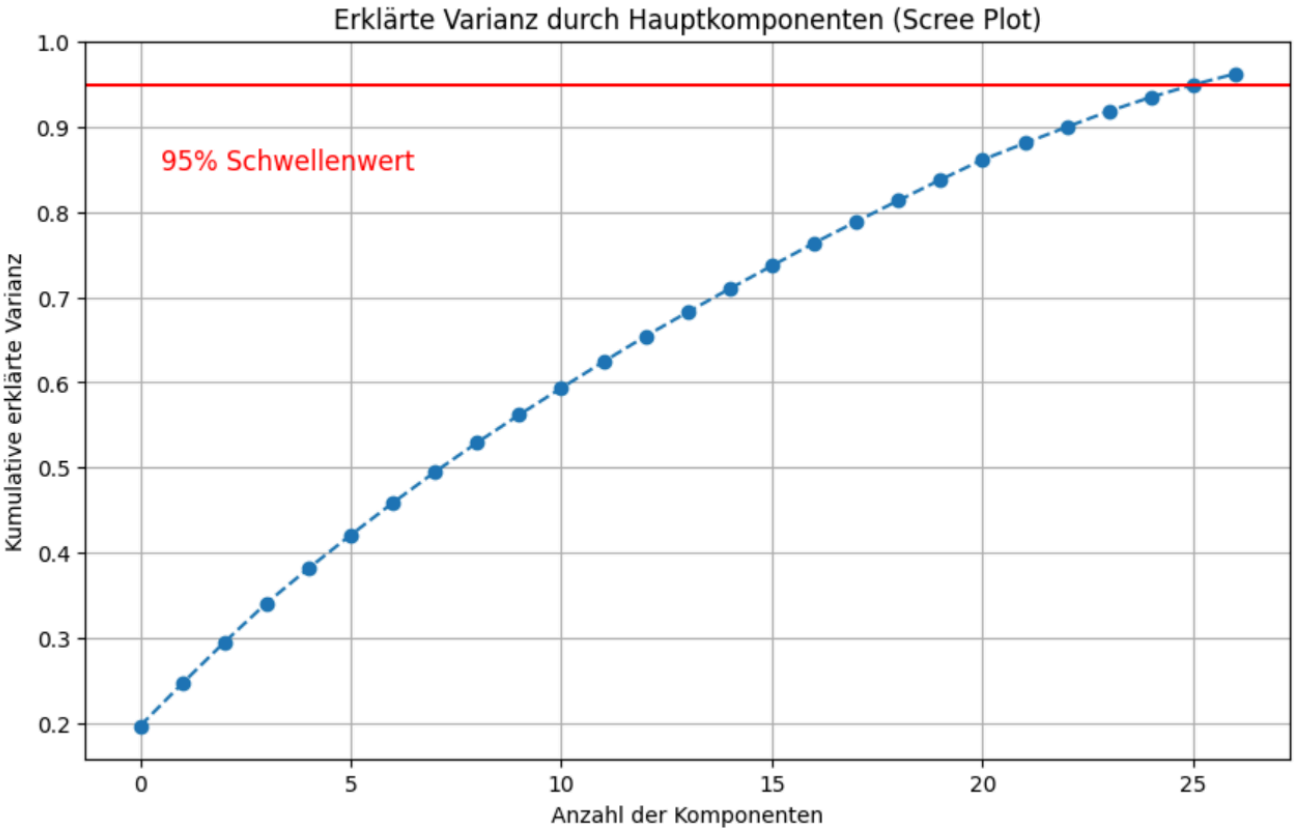
Um verborgene Muster und natürliche Gruppierungen im Nutzerverhalten ohne vordefinierte Annahmen zu entdecken, wird als primäre Methode das unüberwachte maschinelle Lernen eingesetzt.

- **Explorative Datenanalyse (EDA):** Das Projekt begann mit einer umfassenden EDA, um Datenverteilungen zu visualisieren, Korrelationen zu identifizieren und ein grundlegendes Verständnis der Rohdatensätze zu gewinnen.
- **Feature Engineering & Skalierung:** Bestehende Daten werden zu aussagekräftigen Merkmalen kombiniert. Da Clustering-Algorithmen auf Distanzmessungen basieren, werden alle numerischen Features auf eine einheitliche Skala gebracht (Standardisierung), um Verzerrungen zu vermeiden



- **Hashing Encoding:** Textbasierte kategoriale Merkmale mit vielen einzigartigen Werten (z. B. Heimatstadt) wurden in eine feste Anzahl numerischer Merkmale umgewandelt. Dieser Prozess macht wertvolle, nicht-numerische Informationen für den Clustering-Algorithmus nutzbar, ohne eine unüberschaubare Anzahl neuer Spalten zu erzeugen

- **Hauptkomponentenanalyse (PCA):** Um den hochdimensionalen Datensatz für das Clustering zu optimieren, wurde die PCA verwendet, um die Anzahl der Merkmale zu reduzieren und gleichzeitig ein Maximum an Informationen (Varianz) zu erhalten. Dieser Schritt verdichtet komplexe Daten zu ihren wesentlichen Komponenten, was zu einem stabileren und effizienteren Clustering führt.



- K-Means-Clustering: Im Kern wird der K-Means-Algorithmus verwendet, um die Nutzerdaten in eine bestimmte Anzahl von Clustern (Segmenten) zu unterteilen. Die Mitglieder eines Clusters ähneln sich in ihrem Verhalten und ihren Merkmalen stark, während sie sich von den Mitgliedern anderer Cluster deutlich unterscheiden

Vorgehensweise (Projektschritte)

Das Projekt folgt einem strukturierten Prozess:

1. **Datenbeschaffung & -aufbereitung:** Die relevanten Daten aus den users, sessions, flights und hotels Tabellen werden mittels der entwickelten SQL-Abfrage extrahiert und zu einem einzigen Analyse-Datensatz zusammengeführt.

2. **Explorative Datenanalyse (EDA)**: Untersuchung der Daten auf Muster, Anomalien und Zusammenhänge, um Hypothesen für die Segmentierung zu bilden.

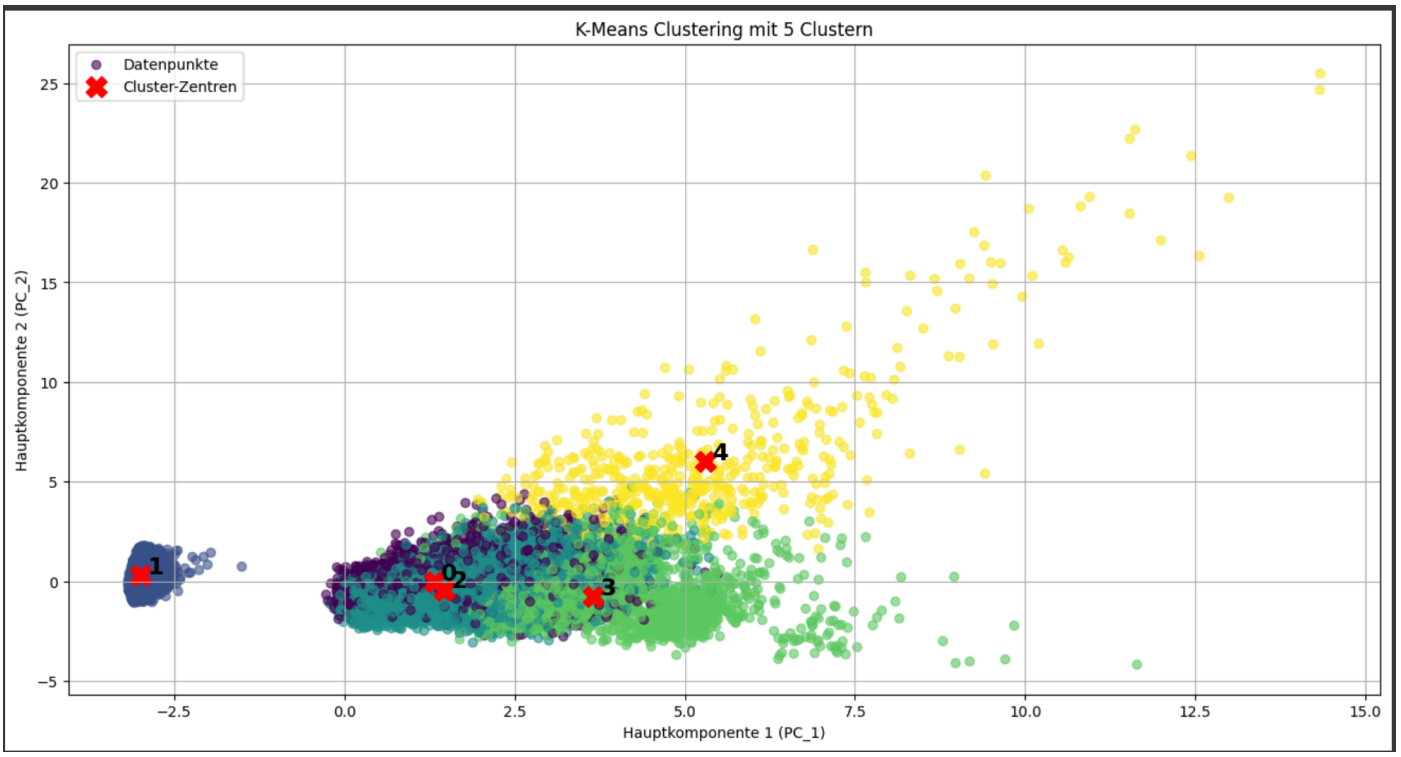
3. **Feature Engineering & Skalierung**: Aufbereitung der Daten für das Modelltraining, inklusive der Normalisierung der numerischen Features.

4. **Modelltraining**: Anwendung des K-Means-Algorithmus. Mittels der "Ellenbogenmethode" wird die optimale Anzahl von Clustern für die Segmentierung ermittelt.

5. **Segmentanalyse & Interpretation:** Detaillierte Analyse der resultierenden Cluster. Jedes Segment wird auf seine charakteristischen Merkmale (z.B. Preisbewusstsein, Reisedauer) untersucht, um greifbare "Personas" zu definieren.

6. **Ergebnispräsentation & Handlungsempfehlungen:** Die finalen Kundensegmente werden visualisiert und beschrieben. Darauf aufbauend werden konkrete, datengestützte Empfehlungen für das Design des personalisierten Prämienprogramms formuliert.

Analyseergebnisse: Die 5 Travel Tide-Personas



Cluster 0 = Der/Die effiziente Städtereisende

Diese Kunden sind typischerweise zwischen 30 und 50 Jahre alt und reisen meist ohne Kinder. Sie wissen genau, was sie wollen: einen kurzen Tapetenwechsel, meistens über ein Wochenende.

Sie buchen selten, aber wenn, dann sehr gezielt und entschlossen. Sie planen eine bestimmte Kurzreise, buchen sie und sind dann wieder weg. Loyalität zu einer Plattform ist ihnen nicht so wichtig wie ein guter Deal für ihren spezifischen Trip. Sie reisen mit leichtem Gepäck (Handgepäck).

Top 3 Beweise aus den Daten

- Nur eine Reise: Der num_trips-Wert liegt bei ca. 1.
- Kurzer Aufenthalt: Die nights liegen im Schnitt bei 2-3 Nächten.
- Leichtes Gepäck: Der checked_bags-Wert ist extrem niedrig, sie reisen also fast nur mit Handgepäck.

Marketing-Empfehlung: Die attraktivsten Vorteile für diese Gruppe

- Top 👍: "Keine Stornogebühren" und "Exklusive Rabatte".
- Flop 👎: "Kostenfreies Aufgabegepäck" ist für sie irrelevant.

Cluster 1 = Die Schaufensterbummler

Diese Nutzer bilden die mit Abstand größte Gruppe. Sie sind auf der Plattform aktiv und klicken sich um, haben aber noch nie eine Buchung getätigt. Sie zeigen Interesse, aber zögern vor dem finalen Schritt. Hier liegt das größte ungenutzte Potenzial.

Top 3 Beweise aus den Daten:

- Keine Reisen: Der num_trips-Wert ist exakt 0.
- Interesse vorhanden: Der page_clicks-Wert ist über 0 (im Schnitt 18), sie sind also auf der Seite aktiv.
- Keine Konversion: Die conversion_rate ist 0, sie schließen nie ab.

Marketing-Tipp: Ziel ist es, die Hürde für die erste Buchung zu senken.

- Top 👍: "Exklusive Rabatte" (z.B. ein Erstanmelder-Bonus) und "Keine Stornogebühren", um das Risiko zu minimieren.

Cluster 2 = Die effizienten Städtereisenden (Gruppe B)

Diese Gruppe ist in ihrem Verhalten fast identisch mit Cluster 0. Es handelt sich ebenfalls um zielgerichtete Einmal-Bucher, die unkomplizierte und kurze Trips mit leichtem Gepäck bevorzugen. Die separate Gruppierung deutet auf feine Unterschiede hin (z.B. bevorzugte Reiseziele oder Airlines), die in diesen Daten nicht sichtbar sind.

Top 3 Beweise aus den Daten:

- Nur eine Reise: Der num_trips-Wert liegt bei ca. 1.
- Kurzer Aufenthalt: Die nights liegen im Schnitt bei 2-3 Nächten.
- Leichtes Gepäck: Der checked_bags-Wert ist extrem niedrig.

Marketing-Tipp: Die Strategie ist identisch zu Cluster 0.

- Top 👍: "Keine Stornogebühren" und "Exklusive Rabatte".

Cluster 3 = Die aufstrebenden Stammkunden

Diese Kunden haben bereits mehrfach gebucht und zeigen erste Anzeichen von Loyalität. Sie interagieren deutlich mehr mit der Webseite (page_clicks ist hoch), was darauf hindeutet, dass sie die Plattform aktiv für ihre Reiseplanung nutzen. Ziel ist es, diese Kunden fest an die Marke zu binden.

Top 3 Beweise aus den Daten:

- Mehrfach-Bucher: Der num_trips-Wert liegt im Schnitt bei über 2.
- Beginnende Loyalität: Der loyalty_indicator_total ist ebenfalls über 2.
- Hohes Engagement: Der page_clicks-Wert (im Schnitt 96) ist deutlich höher als bei den Einmal-Buchern.

Marketing-Tipp: Belohne ihre bisherige Treue und gib ihnen einen Anreiz für die nächste Buchung.

- Top 👍: "1 Hotelübernachtung gratis bei Flugbuchung" ist ein starker Anreiz für die nächste Reise. Auch "Exklusive Rabatte" wirken als Belohnung.
    - Ein einfacher Vorteil wie "Kostenloses Essen" ist nett, aber nicht der entscheidende Hebel, um sie zu halten.

Cluster 4 = Die Premium-Familienplaner

Dies ist eine kleine, aber extrem wertvolle Gruppe. Sie buchen zwar nicht oft, aber wenn, dann geben sie sehr viel Geld für lange Reisen mit der ganzen Familie aus. Das sind die Kunden, die den großen Jahresurlaub planen und dabei auf Qualität und Komfort achten.

Top 3 Beweise aus den Daten:

- Extrem hohe Ausgaben: Der total_trip_cost-Wert ist mit über 6.200 € mit Abstand am höchsten.
- Lange Reisen & viel Gepäck: Die nights (ca. 7) und checked_bags (ca. 2) sind die höchsten aller Cluster.
- Oft mit Familie: Der has_children-Wert ist mit ca. 50% am höchsten.

Marketing-Tipp: Behandle diese VIPs mit hochwertigen Angeboten, die ihren Bedürfnissen entsprechen.

- Top 👍: Alle Vorteile sind relevant! Besonders wirksam sind "Kostenfreies Aufgabegepäck" und "1 Hotelübernachtung gratis", da sie bei langen, teuren Reisen einen hohen monetären Wert darstellen.

Zusammenfassende Handlungsempfehlungen

Die datengestützte Analyse unseres Kundenstamms hat fünf klar voneinander abgrenzbare Kundensegmente aufgedeckt. Jedes dieser Segmente repräsentiert eine einzigartige Persona mit spezifischen Verhaltensmustern und Bedürfnissen, die eine gezielte Marketing Ansprache erfordert.

Im Folgenden werden die Ergebnisse der Clusteranalyse detailliert vorgestellt. Für jede der fünf identifizierten Personas wird eine Beschreibung ihres typischen Verhaltens geliefert und daraus die jeweils wirksamste Handlungsempfehlung abgeleitet, um sie erfolgreich für das Travel Tide-Prämienprogramm zu gewinnen.

Fazit und Ausblick

Dieses Projekt hat erfolgreich Rohdaten in ein klares, umsetzbares Segmentierungsmodell überführt, das die strategische Grundlage für personalisiertes Marketing bei Travel Tide bildet. Der nächste logische Schritt ist die Implementierung gezielter A/B-Test-Kampagnen auf Basis der abgeleiteten Persona-Empfehlungen, um die direkten Auswirkungen auf Konversionsraten und Kundenloyalität zu messen und eine kontinuierliche Optimierung unserer Marketingmaßnahmen zu ermöglichen.

Anhang

Explorative Datenanalyse (EDA)

weiblich: 453.654

männlich: 558.986

other: 8.286

155.593 sind Familien (Kinder & Verheiratet) = 15,24 %

319.637 unverheiratet und haben Kinder. = 31%

394.817 sind verheiratet und haben keine Kinder = 38,67%

Altersgruppenverteilung der User:

Junger Erwachsener (18-29)	180.134
Erwachsener (30-49)	563.730
Senior (50+)	277.062

774.134 (14,31%) User haben mindestens 1x sowohl Flug als auch Hotel gebucht. *(unique)*

1.665.150 User haben insgesamt Flug & Hotel gebucht. *(Wiederkehrende Kund:innen enthalten)*

326.558 haben nur Flüge gebucht. *(Wiederkehrende Kund:innen enthalten)*

344.137 haben nur Hotel gebucht. *(Wiederkehrende Kund:innen enthalten)*

Top 10 User: Ø- & Gesamt-Aufenthaltsdauer

User_id	Total Aufenthaltsdauer	Durchschnittliche Aufenthaltsdauer
40110	06:35:58	00:35:59
82350	06:26:07	00:29:42
29390	06:22:46	00:42:31
141579	06:21:24	00:54:29
1107	06:13:09	00:37:18
187226	05:57:30	00:51:04
1550	05:11:24	00:31:08
15795	04:58:39	00:37:19
102822	04:57:27	00:29:44
47787	04:51:04	00:24:15

Die 10 beliebtesten Hotel's mit dem meisten Übernachtungen in Summe:

Hotelname	Übernachtungen (Nights)
Radisson - new york	58527
Rosewood - new york	58204
Starwood - new york	57967
NH Hotel - new york	57881
Best Western - new york	57867
Wyndham - new york	57691
Conrad - new york	57681
InterContinental - new york	57677
Banyan Tree - new york	57441
Marriott - new york	57422

Das beliebteste Reiseziel ist New-York (USA)

Die unbeliebtesten Reiseziele mit den wenigstens Übernachtungen in Summe:

Hotelname	Übernachtungen (Nights)
Banyan Tree - dalian	5
Hilton - quito	6
Radisson - qingdao	7
Wyndham - tianjin	9
Crowne Plaza - lagos	9