Hochschule Offenburg
offenburg.university

Entwicklung eines Maschine Learning Modells ohne die Verwendung von Hotel spezifischen Vergangenheitsdaten

William Mendat

MASTERARBEIT

zur Erlangung des akademischen Grades Master of Science (M.Sc.)

Studiengang Informatik Master

Fakultät Elektrotechnik, Medizintechnik und Informatik
Hochschule für Technik, Wirtschaft und Medien Offenburg

30.03.2024

Durchgeführt bei der Firma happyhotel

Betreuer
Prof. Dr.-Ing. Janis Keuper, Hochschule Offenburg
Prof. Dr. rer. nat. Klaus Dorer, Hochschule Offenburg

Mendat, William:
Entwicklung eines Maschine Learning Modells ohne die Verwendung von Hotel spezifischen Vergangenheitsdaten / William Mendat. –
MASTERARBEIT, Offenburg: Hochschule für Technik, Wirtschaft und Medien Offenburg, 2024. 33 Seiten.

Mendat, William:
Development of a machine learning model without using hotel-specific historical data / William Mendat. –
MASTER THESIS, Offenburg: Offenburg University, 2024. 33 pages.

Vorwort

Die Entwicklung eines Maschine Learning Modells ohne die Verwendung von Vergangenheitsdaten ist ein umfangreiches und sehr interessantes Thema. Es ist ein sehr weitreichendes Thema, an dem sehr lange geforscht werden könnte und an sich auch eine Wissenschaft für sich ist. Nicht alle Facetten und Möglichkeiten, dieses doch recht Komplexe Thema zu bewältigen, werden in der folgenden Arbeit dargestellt aber es wird einblick gegeben, wie an dieses Thema herangegangen werden kann.

Ich möchte mich an dieser Stelle bei allen Personen bedanken, die mich während dieser Bachelor-Thesis unterstützt haben. Ein besonderer Dank geht dabei an Prof. Dr.-Ing. Janis Keuper für die Betreuung während der Arbeit und die Möglichkeit, dieses Thema überhaupt bearbeiten zu können. Des Weiteren möchte ich mich bei Kai Schmidt und Marius Müller bedanken, die mich während der Bearbeitung tatkräftig Unterstütz haben. Als letztes möchte die happyhotel an sich danken, dass sie mir erlaubt haben bei Ihnen in der Firma diese Thesis zu schreiben.

Eidesstattliche Erklärung

Hiermit versichere ich eidesstattlich, dass die vorliegende Bachelor-Thesis von mir selbstständig und ohne unerlaubte fremde Hilfe angefertigt worden ist, insbesondere, dass ich alle Stellen, die wörtlich oder annähernd wörtlich oder dem Gedanken nach aus Veröffentlichungen, unveröffentlichten Unterlagen und Gesprächen entnommen worden sind, als solche an den entsprechenden Stellen innerhalb der Arbeit durch Zitate kenntlich gemacht habe, wobei in den Zitaten jeweils der Umfang der entnommenen Originalzitate kenntlich gemacht wurde. Ich bin mir bewusst, dass eine falsche Versicherung rechtliche Folgen haben wird.

Offenburg, 30.03.2024

William Mendat

Sperrvermerk

Die vorliegende Abschlussarbeit beinhaltet vertrauliche Informationen und interne Daten des Unternehmens happyhotel. Sie darf aus diesem Grund nur zu Prüfzwecken verwendet und ohne ausdrückliche Genehmigung durch die happyhotelweder Dritten zugänglich gemacht, noch ganz oder in Auszügen veröffentlicht werden. Die Sperrfrist endet 5 Jahre Jahre nach dem Einreichen der Arbeit bei der Hochschule Offenburg. Unbeschadet hiervon bleibt die Weitergabe der Arbeit und Einsicht in die Arbeit an die mit der Prüfung befassten Mitarbeiter der Hochschule und Prüfer möglich, die ihrerseits zur Geheimhaltung verpflichtet sind, sowie die Verwendung der Arbeit in eventuellen prüfungsrechtlichen Rechtsschutzverfahren nach Maßgabe der geltenden verwaltungsprozessualen Regeln.

Zusammenfassung

Entwicklung eines Maschine Learning Modells ohne die Verwendung von Hotel spezifischen Vergangenheitsdaten

Test Abstarct

Abstract

Development of a machine learning model without using hotel-specific historical data

Test Abstract

Inhaltsverzeichnis

1. Einleitung

In einer Welt, die sich mit rasanter Geschwindigkeit digitalisiert, suchen die Menschen stehts nach Wegen, um die Komplexität des modernen Lebens zu bewältigen. Diese Digitalisierung hat eine stetig wachsende Sehnsucht nach der Vorhersage zukünftiger Ereignisse hervorgebracht - sei es in der Wirtschaft, der Gesundheitsbranche oder auch im Bereich des Dienstleistungssektors wie dem Hotelgewerbe. Es ist ein Streben nach Präzision, ein Bestreben, aus Daten und Mustern eine art Kristallkugel zu formen, um die Zukunft vorhersagen zu können.

Albert Einstein hat einst mit einem Buchtitel von Ihm gesagt:

> If you want to know the future, look at the past. [1]

Dieser Gedanke illustriert die gängige Annahme, dass die Vergangenheit Hinweise auf die Zukunft liefern kann. Es ist interessant anzumerken, dass dieses Zitat auch als Titel eines Buches von Einstein dient, welches seine philosophischen Ansichten zur Zeit, Raum und Vorhersage behandelt.

Doch was passiert, wenn diese Vergangenheitsdaten nicht verfügbar sind oder nicht genutzt werden können? In Branchen wie der Hotelindustrie, die oft noch auf traditionelle, statische Preisstrategien zurückgreifen, stellt sich die Frage, wie eine effektive Vorhersage ohne spezifische historische Daten möglich ist.

Es wird immer deutlicher, dass ein dynamischerer Ansatz im Hotelwesen erforderlich ist, um die Umsatzoptimierung durch Revenue Management zu steigern. Dies erfordert die Anpassung von Preismodellen an sich ändernde Nachfrage und andere Einflussfaktoren. Eine mögliche Lösung liegt in der Verlagerung der traditionellen Rolle des Revenue Managements auf Modelle, die auf breiteren Datenquellen und fortgeschrittenen Methoden des maschinellen Lernens basieren.

Die Suche nach einem solchen Modell, das ohne die spezifischen Vergangenheitsdaten eines bestimmten Hotels auskommt, bildet das Herzstück dieser Forschungsarbeit. Der Fokus liegt darauf, alternative Datenquellen zu erkunden und innovative Ansätze zu entwickeln, um Prognosen und Entscheidungsgrundlagen für das Revenue Management in der Hotellerie zu schaffen. Ziel ist es, dass diese nicht ausschließlich auf vergangenen Daten eines spezifischen Hotels basieren, sondern auf einer Vielzahl von allgemeinen, zugänglichen Informationen und fortschrittlichen Analysemethoden beruhen. Es geht darum, einen Weg zu finden, wie Hotels, selbst ohne ihre spezifischen vergangenen Daten, zukünftige Entscheidungen im Bereich des Revenue Managements treffen können, um ihre Leistung zu optimieren und ihre Wettbewerbsfähigkeit zu stärken.

1.1. happyhotel

Die vorliegende Masterthesis fängt mit einem umfassenden Überblick über die Firma happyhotel an. In diesem ersten Kapitel wird eingehend auf die fundamentale Idee und das herausragende Produkt von happyhotel eingegangen, welches einen signifikanten Beitrag zur Weiterentwicklung der Hotelbranche leistet.

Im Anschluss wird der Fokus auf das gegenwärtige Vorgehen des Unternehmens, der Dynamischen Preisgenerierung, gelegt. Diese fortschrittliche Methode, die auf einer Künstlichen Intelligenz und umfassender Datenanalyse basiert, ermöglicht es happyhotel, in Echtzeit auf auf den momentanen Markt zu reagieren und optimale Preise für Hotels zu generieren.

Abschließend wird in diesem einführenden Kapitel die zugrunde liegende Problematik hervorgehoben, die den Ausgangspunkt dieser Arbeit bildet. Dabei wird der Fokus auf eine Herausforderung gerichtet, die mit der Dynamischen Preisgenerierung einhergeht. Diese Problematik dient als Basis für die nachfolgende Analyse und Forschung, die darauf abzielt, innovative Lösungsansätze und Optimierungen im Rahmen der Preisstrategie von happyhotel zu entwickeln.

1.1.1. Das Unternehmen

Die Firma happyhotel wurde im Jahr 2019 von den drei Gründern Sebastian Kuhnhardt, Marius Müller und Rafael Weißmüller gegründet. Sie wollten wie der Name schon vermuten lässt Hotels glücklicher machen. Angefangen hat es mit der Erkenntnis von Sebastian, dass sich viele Hoteliers nicht mit der Dynamischen Preisgestaltung beschäftigen. Meist vertrauen diese Hoteliers einfach auf ihr Bauchgefühl, welcher Preis zur momentanen Situation passen könnte oder passen ihre Preise gar nicht an.

Somit stellten sie sich die Fragen:

- Wie können die Preise für die Übernachtung in einem Zimmer besser vorausgesagt werden?
- Wonach sollten sich die Preise richten und wie kann man sie bestimmen?

Eine Lösung musste her um mehr Dynamik in die Preisgestaltung zu bringen. Sie erschufen die Software happyhotel ein Revenue Management System. Mit happyhotel kann der Hotelier sein gesamtes Hotel analysieren und Ihm werden Preisvorschläge für seine Zimmerkategorien generiert um mehr Umsatz zu erzeugen.

Abbildung 1.1: happyhotel

1.1.2. Dynamische Preisgenerierung

Wie im vorherigen Kapitel erwähnt fokussiert sich happyhotel auf die Dynamische Preisgenerierung. Um diese Preise zu generieren braucht es vor allem zwei Sachen:

- Die Daten des Hotels wie zum Beispiel Buchungen
- Ein Vorgehen um aus den gesammelten Daten Preise zu generieren

Die Daten bekommen sie aus den verschiedensten Quellen. Sogennante Property Management Systeme kurz PMS sind Systeme um ein Hotel zu verwalten. In diesem Property Management Systeme können Hotels zum Beispiel ihre Zimmer verwalten oder aber auch Buchungen anlegen und pflegen. Mit den Herstellern dieser

Property Management Systeme arbeitet happyhotel zusammen um an die Daten des Hotels zu gelangen.
Da die Daten vorhanden sind, braucht es ein Vorgehen um aus den Daten einen Preis zu generieren. Dazu ist happyhotel auf die folgenden zwei Ideen gekommen:

- Buchungskurvenmodell
- Kombination aus RevPAR und Buchungskurvenmodell

Das Buchungskurvenmodell war die erste Idee von happyhotel. Bei dem Buchungskurvenmodell wird sich die Vergangenheit angeschaut um das zukünftige Buchungsverhalten vorherzusagen. Ziel dabei ist es die Auslastung für einen Tag vorherzusagen um anhand dessen einen akkuraten Preis zu bestimmen.

Da bei dem Buchungskurvenmodell sich einige schwächen aufgezeigt haben, wurde ein neues Modell erschaffen um die Schwächen entgegen zu wirken. Es sollte nun der RevPAR wert vorhergesagt werden und mit dem Buchungskurvenmodell angepasst werden. RevPAR steht dabei für Revenue per Available Room. Auch bei diesem Modell wird sich die Vergangenheit des Hotels angeschaut um den Zukünftigen Umsatz pro verfügbarem Zimmer vorherzusagen und basierend darauf den endgültigen Preis zu ermitteln.

1.1.3. Momentane Problematik

Wie es Albert Einstein schon sagte: Soll die Zukunft vorhersagen gesagt werden, so sollte sich die Vergangenheit angeschaut werden. Auf diesem Grundprinzip ist happyhotel auch vorgegangen, sie schauen sich bei beiden Ansätzen die Vergangenheit an um Vorhersagen über die Zukunft zu tätigen.

Doch was passiert, wenn die Daten der Vergangenheit nicht vorhanden sind? Dies kann zum Beispiel passieren wenn ein Hotel die Software nutzen möchte, welches erst in der Zukunft eröffnet. Auch dieses Hotel soll mit adäquaten Preisvorschlägen gefüttert werden. Die soeben beschriebene Situation ist ein generelles Problem im Maschine Learning Bereich. Es kommt nicht allzu selten vor, dass keine Vergangenheitsdaten aus den verschiedensten Gründen vorliegen. Dieser Problematik soll in dieser Arbeit auf dem Grund gegangen werden.

Ziel dieser Arbeit ist es, ein Modell zu entwickeln welches Preisempfehlungen für ein Hotel liefert, für das es bisher noch keine Vergangenheitsdaten gibt. Dieses Modell soll dann für folgende zwei Scenarien genutzt werden können:

- Neue happyhotel Kunden ohne Daten
- Nachfrageeinschätzung für bestimmte Märkte

1.2. Vorgehensweise

Vor dem Eintauchen in die Lösung eines Problems ist es entscheidend, einen klaren Weg dorthin festzulegen. Antoine de Saint-Exupéry hat mit den Worten *Ein Ziel ohne Plan ist nur ein Wunsch* treffend darauf hingewiesen, dass ein bloßes Ziel ohne einen durchdachten Plan lediglich eine vage Vorstellung bleibt. Das Verständnis und die Festlegung einer angemessenen Vorgehensweise sind daher der Schlüssel, um ein Problem effektiv anzugehen. Aufgrund dessen wird in dem folgenden Abschnitt dieser Arbeit auf die Vorgehensweise eingegangen. Zudem werden innerhalb dieser Sektion die Benchmark-Hotels ermittelt an welchen getestet werden kann, ob die Vorgehensweise ein Erfolg war.

1.2.1. Vorgehen in der Datenwissenschaft

Ein Projekt welches in der Datenwissenschaft (engl. Data Science) angesiedelt ist, beginnt in der Regel mit einem geschäftlichen Problem, so wie es auch in dieser Thesis der Fall ist. Sobald das Problem klar definiert ist, sollten ein oder mehrere Konzepte ausgearbeitet werden, wie das Problem gelöst werden kann. Diese Konzepte sollen als Leitfaden dienen um das angestrebte Ziel zu erreichen. In der Regel ist es ratsam mehr als ein Konzept auszuarbeiten, da so ausweichmöglichkeiten festgelegt werden können, sollte ein Konzept nicht funktionieren. So werden auch in dieser Arbeit, in dem Kapitel *Konzepte*, für das vorliegende Problem Konzepte ausgearbeitet und Evaluiert.

Mit der klaren Definition des Problem und der darauffolgenden Konzeptionieren, kann der Datenwissenschaftler mit Hilfe des *OSEMN*-Vorgangs die Problematik angehen [2].

Der *OSEMN*-Vorgang besteht aus den folgenden Elementen:

- Obtian data (Erhalten von Daten)
- Scrub data (Daten reinigen)
- Explore data (Untersuchen von Daten)
- Model data (Modelldaten)
- Interpret results (Interpretieren von Ergebnissen)

Obtian data

Die wichtigste Ressource des 21. Jahrhundert besteht in den Daten, die zur Verfügung stehen. Dies erläuterte Klaus Schwab, der Gründer des Weltwirtschaftsforums, mit seinen Worten: *Die wertvollste Ressource des 21. Jahrhunderts sind nicht mehr Öl, sondern Daten*. Aufgrund dessen besteht der erste Schritt, nach der Evaluierung der Konzepte, in der Beschaffung der Daten. Es muss zunächst ein Überblick geschaffen werden. Zum Überblick gehören Informationen wie:

- Welche Daten bereits vorhanden sind.
- Welche Daten eventuell noch intern neu erworben werden müssen.
- Welche Daten aus dem Internet gezogen werden können

Sobald die Daten beschaffen worden sind, kann mit dem nächsten Schritt fortgefahren werden.

Scrub data

Der nächste Schritt besteht darin, die Daten zu bereinigen. Zum bereinigen der Daten gehört der Vorgang mit dem die Daten in ein standardisiertes Format gebracht werden. Dazu gehört der Umgang mit fehlenden Daten, sowie die Korrektur von Fehlern und das Entfernen von sogenannten *outlier* [2]. Outlier sind Daten, welche im Verhältnis zu der gesamten Datenmenge aus der reihe Tanzen.

Explore data

Die Datenuntersuchung oder auch Datenanalyse dient dazu, um mit den Daten vertraut zu werden und ein besseres Verständnis für die Daten zu Gewinnen. Dies ist ein sehr Wichtiger Schritt bei einem *Data Science* Projekt, da nur dann ein gutes Ergebnis erzielt werden kann, wenn die Daten, mit den gearbeitet werden kann, verstanden sind. Das Verständnis über die Daten trägt zu dem auch maßgeblich dazu bei, den richtigen Ansatz für die Modellierung zu finden.

Model

Nach dem Erforschen der Daten, kann ein *Maschine Learning* Modell eingesetzt werden. Es existieren viele verschieden Modelle, die meist einen für einen speziellen Fall implementiert worden sind. Die Auswahl des Modells hängt ganz davon ab, welche Art von Problem gelöst werden soll und welche Daten vorhanden sind, um dieses Problem zu lösen.

Interpret results

Der letzte Schritt besteht darin, die gesammelten Ergebnisse zu Interpretieren. Dazu gehörend ist die Entscheidung, ob die erzielten Ergebnisse gut oder schlecht sind. Meistens werden zur Hilfe der Entscheidung ob die Ergebnisse gut oder schlecht sind, Diagramme, Grafiken und Tabellen erstellt. Es soll hier zudem auch Entschieden werden, ob die Ergebnisse verwendet werden sollten oder ob nicht weiter geforscht werden muss. So entsteht ein Kreislauf, der mit dem erstellen weiteren Konzepte startet und mit dem Interpretieren der Ergebnisse endet.

1.2.2. Benchmark Hotels

Noch vor der Konzeptionierung zur eigentlichen Lösung der Problematik, sollten schon vorhandene Hotels in unserer Datenbank als *Benchmark-Hotels* ausgesucht werden. Die Idee dahinter ist es, Hotels auszusuchen, bei denen viele Daten vorhanden sind und so zu tun als wären gar keine Daten von diesen Hotels vorhanden. Mithilfe dieser Hotels sollen denn die ausgearbeiteten Konzepte von Kapitel *Konzepte* validiert werden. Ein Konzept wird dann als gut empfunden, wenn es in der

Lage ist, gute Preisvorschläge für die Benchmark-Hotels zu erzeugen.

Da nicht jedes Hotel, welches in der Datenbank vorhanden ist in frage kommt, wurden einige Kriterien aufgestellt, die ein Hotel erfüllen muss, um als Benchmark-Hotel gelten zu können. Diese Kriterien sehen wie folgt aus:

- Haben Daten von mehr als zwei Jahren
- Haben einen Benutzer mit der Role *Revenue Manager*
- Haben oft Preise geändert
- Haben oft vorgeschlagene Preise von happyhotel nicht angenommen

Der hintergrund warum die letzten drei Kriterien dazu kamen, liegt darin, dass diese Hotels mit den Preisvorschlägen von happyhotel nicht zufrieden sind und vermutlich auch manuell *Revenue Managment* betreiben. Die Hoffnung besteht darin, dass das Konzept nicht nur für Hotels gute Preisvorschläge liefert, sondern, dass das auch verwendet werden kann, als Alternative zu den vorhandenen zwei Modellen, für Hotels die noch manuell Dynamische Preise gestalten.

Es wurde zur Analyse der Hotels eine *json*-Datei erstellt. Diese *json*-Datei besteht aus einer Liste von einzelnen Objekten. Jedes Objekt innerhalb der Liste repräsentiert ein Hotel in der Datenbank. Ein Beispiel für diese *json*-Datei könnte wie folgt aussehen:

```
[
 {
     "company_id": "11111111111111",
     "not_accepted_recs": 10,
     "price_change_average": 5
 }
]
```

Listing 1.1: Beispielhafte json-Datei

Das Feld *not_accepted_recs* beschreibt dabei wie viele Preisvorschläge das Hotel nicht angenommen hat beziehungsweise ignoriert oder abgelehnt hat und das Feld *price_change_average* ist die durchschnittliche Preisänderung pro Tag. Anhand von diesen Informationen konnten zwei Hotels als Benchmark-Hotels ausgesucht werden.

2. Konzepte

Wie bereits in dem vorangegangenen Kapitel *Vorgehensweise* hervorgehoben wurde, bildet die Entwicklung und Ausarbeitung von Konzepten einen essenziellen Eckpfeiler bei der Bewältigung und Lösungsfindung für komplexe Probleme. Im nachfolgenden Abschnitt wird eine ausführliche Vorstellung und Evaluation der erarbeiteten Konzepte präsentiert. Dabei liegt der Fokus darauf zu ergründen, welche Konzepte das Potenzial besitzen, weiterverfolgt zu werden, um die spezifische Herausforderung zu meistern. Besondere Aufmerksamkeit gilt hierbei der grundlegenden Idee jedes Konzeptes sowie der detaillierten Darlegung, wie jedes einzelne Konzept dazu beitragen kann, dynamische Preisgestaltung für ein Hotel auch ohne vorhandene Daten zu generieren

Ziel dieses Kapitels ist es einen umfassenden Überblick über die Konzepte zu geben, ihre Relevanz für die Forschung zu betonen und den Weg für die darauffolgenden Analysen und Schlussfolgerungen zu ebnen.

2.1. Hotel Daten von vielen Hotels

Das Konzept *Hotel Daten von vielen Hotels* verfolgt die grundlegende Idee, sämtliche bis dato gesammelten Hoteldaten zu konsolidieren und ein umfassendes, übergeordnetes Modell des maschinellen Lernens zu entwickeln.

Die bereits vorhandenen Daten aus zahlreichen Hotels bildet die Basis für den Aufbau eines solchen Modells. Dieses Vorhaben sieht vor, das bereits existierende RevPAR-Modell zu modifizieren und zu erweitern. Zunächst wird angestrebt, alle Hotels in eine vergleichbare Form zu bringen. Eine mögliche Herangehensweise hierbei ist die Definition bestimmter Hotelmerkmale und ihre Auflistung als Vektor,

um eine Vergleichbarkeit zu ermöglichen.

Im nächsten Schritt ist eine Anpassung des RevPAR-Modells erforderlich, da dieses normalerweise auf Buchungsdaten basiert. Für Hotels ohne historische Buchungsdaten ist es offensichtlich nicht möglich, diese als Features zu verwenden, da sie schlichtweg nicht verfügbar sind. Stattdessen sollen die charakteristischen Merkmale jedes Hotels dem jeweiligen RevPAR zugeordnet werden.

Sobald das RevPAR-Modell entsprechend umstrukturiert ist, können sämtliche Hotels in der Datenbank als Datensätze dem Modell zugeführt werden. Falls ein Hotel ohne vergangene Buchungsdaten auftaucht, können basierend auf seinen charakteristischen Eigenschaften Prognosen über den zu erwartenden RevPAR getroffen werden. In diesem Szenario muss das Hotel lediglich, wie alle anderen Hotels auch, eine Zuordnung zwischen dem RevPAR und dem tatsächlichen Preis festlegen.

Die Aufmachung dieses Konzeptes soll im folgenden Schaubild nochmal Bildlich verdeutlicht werden:

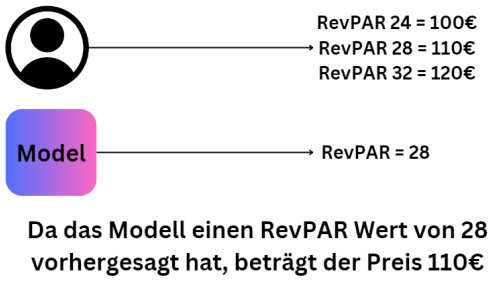

Abbildung 2.1: RevPAR-Modell Vorgehen

Dieser Ansatz zielt darauf ab, eine umfassende Verwendung der vorhandenen Daten zu ermöglichen und somit auch für Hotels ohne historische Buchungsdaten eine Prognose des RevPAR auf der Grundlage ihrer individuellen Eigenschaften zu ermöglichen.

2.2. Mitbewerber Modell

Die grundlegende Idee des Konzeptes: Mitbewerber Modell ist es, die Daten von der Konkurrenz zu benutzen um daraufhin Preisvorschläge zu generieren.

Durch dritt Anbieter wie *HQ-Revenue* können Konkurrenzdaten genutzt werden um ein Modell aufzubauen. HQ-Revenue ist ein Anbieter, welcher Internetseiten wie Bookings.com oder trivago *scraped* um an Hotelpreise oder andere Daten zu kommen. Dabei gibt es viele verschiedene Vorgehensweisen um Preise für ein Hotel ohne Vergangenheitsdaten zu entwickeln. Ein Primitiver Ansatz dabei wäre es, wenn alle Preise von den Konkurrenten genommen werden und damit der Durchschnitt ermittelt wird. Dies hat natürlich nichts mit Maschine Learning oder geschweige denn Data Science zu tun aber es wäre ein Ansatz der Verfolgt werden könnte.

Dieser Ansatz birgt jedoch eine Problematik: Nicht jeder Kunde von happyhotel hat auch automatisch Konkurrenzdaten zur Verfügung, diese müssen noch dazu gebucht werden. Deswegen wird folgender Ansatz verfolgt.

So wie im vorherigen Konzept *Hotel Daten von vielen Hotels* werden auch hier die Hotels in eine Vergleichbare Form gebracht. Das Ziel ist es dann ein oder mehrere Hotels zu finden, die vermeintlich ähnlich sind. Sobald ein oder mehrere ähnliche Hotels gefunden worden sind, können die Konkurrenzdaten von den ähnlichen Hotels genutzt werden.

Als Zielvariable des Modells werden dann die Preise des ähnlichsten Hotels verwenden. Jedoch sollen hierbei die Preise von den Konkurrenten und von dem ähnlichsten Hotel nicht einfach so benutzt werden, sondern lediglich das Verhältnis. Die Preise sollen anhand von dem Durchschnittlichen Preis in Verhältnis gebracht werden und dieses Verhältnis soll vorhergesagt werden.

Das Hotel ohne Vergangenheitsdaten muss in diesem Fall dann einen Durchschnittlichen Preise angeben, anhand dessen mit dem vorhergesagten Verhältnis der tatsächliche Preis abgeleitet werden kann. Dies soll im folgenden Schaubild noch einmal dargestellt werden:

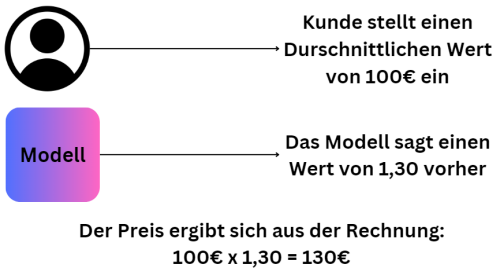

Abbildung 2.2: Mitbewerber Modell

2.3. Ähnliche Hotels

Dieses Konzept der *Ähnlichen Hotels* verschmilzt in gewisser Weise die Ideen der beiden Konzepte *Mitbewerber Modell* und *Hotel Daten von vielen Hotels*. Dieser Ansatz zielt darauf ab, ähnliche Hotels zu identifizieren und basierend auf den Daten dieser Hotels ein Modell zu entwickeln.

Im Gegensatz zum Konzept *Hotel Daten von vielen Hotels* besteht bei diesem Ansatz die Möglichkeit, konkrete Buchungsdaten der jeweiligen Hotels zu verwenden. Die primäre Herausforderung liegt jedoch darin, die ähnlichsten Hotels zu identifizieren. Nachdem diese ähnlichen Hotels ausfindig gemacht wurden, kann das bereits vorhandene Modell *Kombination aus RevPAR und Buchungskurve* ohne jegliche Anpassungen genutzt werden.

Dieses Modell wird dann, ähnlich wie bei anderen Hotels, mit den Buchungsdaten der identifizierten ähnlichen Hotels gefüttert, um Preise zu generieren. In diesem Szenario muss der Kunde lediglich eine Zuordnung zwischen dem RevPAR-Wert und dem konkreten Preis des Hotels festlegen.

Dieser Ansatz kombiniert die Vorteile beider vorherigen Konzepte, indem er sowohl auf die Ähnlichkeits-findung zwischen Hotels als auch auf die Nutzung spezifischer Buchungsdaten abzielt. Durch die Verwendung vorhandener Modelle ohne umfangreiche Modifikationen können so gezielt Preisvorhersagen für ähnliche Hotels generiert werden.

2.4. Synthetische Daten erstellen

Das Konzept der *Erstellung synthetischer Daten* markiert einen innovativen Ansatz innerhalb der Konzepte und weicht von den bisherigen Strategien ab. Dieser Ansatz verfolgt die Idee, ein Modell mit sämtlichen verfügbaren Daten zu trainieren und darauf aufbauend synthetische zukünftige Daten zu generieren. Ziel ist es, eine Art Simulation zu erstellen, die den Buchungsverlauf eines Hotels nachbildet.

Mittels dieser Simulation wird angestrebt, Vorhersagen darüber zu treffen, wie viele Buchungen für bestimmte Zimmerkategorien an bestimmten Tagen eingehen werden. Dadurch soll die Möglichkeit geschaffen werden, einen dynamischen Preis entsprechend dem erwarteten Buchungsverlauf zu gestalten.

Die Grundidee hinter dieser Vorgehensweise liegt in der Schaffung eines virtuellen Modells, das basierend auf vergangenen Daten und Mustern potenzielle zukünftige Buchungen simuliert. Hierbei sollen verschiedene Szenarien durchgespielt werden, um die wahrscheinlichsten Buchungstrends abzuschätzen und somit einen fundierten Ansatz für die dynamische Preisgestaltung zu generieren.

2.5. Evaluation

Nachdem nun die ausgearbeiteten Konzepte vorgestellt wurden, gilt es diese zu Bewerten um festzulegen mit welchen Konzepten fortgefahren werden soll. Jedes der vorgestellten ist auf seine Art valide und hat auch Berechtigung verfolgt zu werden. Um deshalb entscheiden zu können, welches Konzept überhaupt nachgegangen werden soll oder in welcher Reihenfolge die Konzepte ausprobiert werden sollen, werden die Konzepte nach den folgenden Kriterien bewertet:

- Aufwand
- Erfolgswahrscheinlichkeit
- Impact

Aufwand und Erfolgswahrscheinlichkeit sind selbsterklärend. Der Impact bezieht sich darauf, in wie fern happyhotel im generellen von dem Konzept profitieren könnte und ob das Konzept nicht auch für schon vorhandene Kunden eingesetzt werden könnte.

Jedes Konzept kann bei jedem Kriterium eine Zahl zwischen 1 bis 5 erzielen, wobei 5 das Beste und 1 das Schlechteste in dem jeweiligen Kriterium bedeutet. Die Ergebnisse der Evaluation sind wie folgt in der Tabelle dargestellt:

Konzepte	**Aufwand**	**Erfolgsw.**	**Impact**	**Result**
Daten von vielen Hotels	4	4	4	12
Mitbewerber	4	3	3	10
Ähnliche Hotels	4	5	4	13
Synthetischen Daten	1	3	5	9

Tabelle 2.1.: Evaluierung der Konzepte

Nach der Evaluierung wurde bestimmt, dass das Konzept *Ähnliche Hotels* das größte Potenzial hat und soll dementsprechend auch verfolgt werden. Je nach Zeit und Ergebnisse werden die Konzepte *Mitbewerber Modell* und *Hotel Daten von vielen Hotels* auch verfolgt und evaluiert werden. Die Erstellung einer Simulation durch Synthetische Daten ist auch ein sehr interessantes Konzept, würde aber im rahmen dieser Thesis zu weit gehen.

3. Ähnliche Hotels

Die Evaluierung der verschiedenen Konzepte, wie sie im vorangegangenen Kapitel *Konzepte* ausführlich dargelegt wurden, führte zu einer kritischen Entscheidung bezüglich des anfänglichen Fokus für die strategische Ausrichtung. Nach einer umfassenden Analyse und Bewertung wurde klar, dass das Konzept der *Ähnliche Hotels* sich als überzeugende und vielversprechende Lösung für das vorliegende Problem erweist.

Dieses Kapitel konzentriert sich daher detailliert auf das Konzept der *Ähnliche Hotels*. Es zielt darauf ab, die spezifische Methodik und Herangehensweise dieses Konzepts zu beleuchten. Die kommenden Abschnitte bieten eine eingehende Analyse, die durch umfassende Datenanalysen und Experimente unterstützt wird. Dadurch soll ein tiefgreifendes Verständnis für die erfolgreiche Umsetzung des Konzepts der *Ähnliche Hotels* vermittelt werden.

Wie in der Sektion *Ähnliche Hotels* beschrieben, liegt der Fokus zunächst darauf, ähnliche Hotels zu identifizieren, die als Grundlage für die Anpassung und Anwendung des bereits bestehenden RevPAR-Modells dienen sollen.

3.1. Finden von ähnlichen Hotels

Das finden von ähnlichen Hotels ist die Grundlage dieses Konzeptes. Um dies zu erreichen soll das vorgehen, welches in der Sektion *Ähnliche Hotels* beschrieben wurde, verfolgt werden. Angefangen mit der Datenbeschaffung, muss sich zunächst ein überblick darüber gemacht werden, welche Daten schon vorhanden sind und welche Daten gegeben falls noch besorgt werden müssen, bevor dann mit der Datenanalyse und der Modellierung weiter gemacht werden kann.

3.1.1. Datenbeschaffung

Durch die verschiedenen Property Management Systeme, sind für die einzelnen Hotels, schon Stammdaten wie die Zimmer oder Zimmerkategorien vorhanden. Zudem wird der jeweilige Kunde beim Einrichten von happyhotel auch schon nach verschiedenen Daten befragt. Unter den Daten, die bei einem Kunden abgefragt werden, gehören Informationen wie zum Beispiel die Adresse des jeweiligen Hotels.

Werden nun alle schon vorhanden Informationen zusammengetragen, die bisher in der Datenbank zur Verfügung stehen, entsteht dabei das folgende Dataframe:

company_id	median_min	median_max	areatype_count	area_count	region	city
60dca381ae221ce505d2263a	10950.0	12550.0	6	12	Baden-Württemberg	Ohlsbach
61baffd61e564422dba0903f	17900.0	24900.0	5	16	Schleswig-Holstein	Schleswig
60ce02508af8912a494d2b29	30400.0	37400.0	16	68	Rheinland-Pfalz	Rhodt unter Rietburg
614c8db23707c0c5c3bd91e6	6000.0	8600.0	15	49	Nordrhein-Westfalen	Eschweiler
600e7ff88c1fe4620c48db08	12400.0	18500.0	6	27	Baden-Württemberg	Reutlingen

Abbildung 3.1: Alle schon vorhanden Features

In Abbildung 3.1 sind die Daten zu sehen, die bislang zur Verfügung stehen, wobei sie die *median_min* und *median_max* werte auf den Median aller Zimmerkategorie-Preise bezieht.

Dadurch das *region* und *city* manuell vom Kunden eingetragene Werte sind, ergab sich eine gewisse Skepsis ob alle Werte norm-konform eingetragen wurden. Aufgrund dieser Skepsis sollte eine kleine Datenanalyse getätigt werden.

Bei der Datenanalyse wurde jeweils nach der Region und nach der Stadt gruppiert um zu prüfen ob die so Benutzt werden können. Im folgenden sind die Werte jeweils für die Region und für die Stadt zu sehen:

```
Index(['Baden-Württemberg', 'Bavaria', 'Bayern', 'Berlin', 'Brandenburg',
       'Free and Hanseatic City of Hamburg', 'Hansestadt Hamburg', 'Hessen',
       'Kanton Zürich', 'Land Berlin', 'Land Salzburg',
       'Mecklenburg Vorpommern', 'Niedersachsen', 'Nordrhein-Westfalen',
       'Rheinland-Pfalz', 'Saarland', 'Sachsen', 'Sachsen-Anhalt',
       'Schleswig-Holstein', 'Wien'],
      dtype='object', name='region')
```

Abbildung 3.2: Alle vorhanden Regionen

```
Index([' München', 'Aachen', 'Allenbach', 'Appenweier', 'Aschheim', 'Bad Ems',
       'Bad Kreuznach', 'Bad Schlema', 'Baesweiler', 'Bayreuth', 'Bendorf',
       'Berlin', 'Berlin - Karlshorst', 'Borkum', 'Cuxhaven', 'Dachau',
       'Dortmund', 'Dülmen', 'Eisenberg (Pfalz)', 'Erlangen', 'Eschweiler',
       'Freiburg im Br.', 'Freiburg im Breisgau', 'Freising',
       'Gerlinden/Maisach', 'Haltern am See', 'Hamburg', 'Harrislee',
       'Isernhagen', 'Karlsruhe', 'Kirchberg an der Jagst', 'Koblenz', 'Köln',
       'Königswinter', 'Lautenbach', 'Lutherstadt Eisleben', 'Mannheim',
       'Monschau', 'München', 'Nürnberg', 'Ohlsbach', 'Olching', 'Parsberg',
       'Parsberg-Hörmannsdorf', 'Plauen', 'Putbus - Lauterbach', 'Remscheid',
       'Reutlingen', 'Rhodt unter Rietburg', 'Rosenheim', 'Rüdesheim am Rhein',
       'Saarlouis', 'Salzburg', 'Sauerlach', 'Schermbeck', 'Scheßlitz',
       'Schleswig', 'Schwabach', 'Schönwald im Schwarzwald',
       'Seehausen am Staffelsee', 'Sonsbeck', 'St. Peter-Ording', 'Stuttgart',
       'Templin', 'Todtnau', 'Todtnau-Muggenbrunn', 'Trassem', 'Trier', 'Ulm',
       'Uster/Zürich', 'Warstein', 'Weilheim im Oberbayern', 'Wemding',
       'Wernau', 'Westerstede', 'Wien', 'Wunstorf'],
      dtype='object', name='city')
```

Abbildung 3.3: Alle vorhanden Städte

Es ist eindeutig zu erkennen, dass es sowohl bei der Region als auch bei der Stadt zu einer Inkonsistenz kommt. So wurde die wurde Region *Berlin* sowohl als *Berlin*, als auch mit *Land Berlin* angegeben. Auch bei der Stadt ist die Inkonsistenz deutlich zu sehen, so wurde *München* mit Leerzeichen vorne dran angegeben oder die Stadt Freiburg einmal mit der Abkürzung *Br.* und einmal ausgeschrieben *Breisgau* angegeben. Es ergab sich also, dass diese Werte so wie sie sind, nicht verwendet werden können.

Beschaffung von Region, City und Stadtgröße

Durch eine schon im Vorfeld getätigte Arbeit, existiert zu jedem Hotel in unserer Datenbank, die Koordinaten, repräsentiert durch die zwei Werte Long- und Latitude. Mithilfe von diesen zwei Werten sollte ein Skript geschrieben werden um die Region, die Stadt und Größe der Stadt zu beschaffen.

Für das Skript wurde *Nominatim* API verwendet um die einzelnen Werte zu beschaffen. Im folgenden wird das Skript präsentiert:

```
1     from geopy.geocoders import Nominatim
2
3  def get_region_city_size(latitude, longitude):
4     geolocator = Nominatim(user_agent="city_size_app")
5     location = geolocator.reverse((latitude, longitude), language='de')
6
7     address = location.raw['address']
8
9     if 'city' in address:
10      region = address["city"]
11      if 'state' in address:
```

```
12        region = address["state"]
13      return {
14            "city": address["city"],
15            "region": region,
16            "size": "Großstadt"
17        }
18    elif 'town' in address:
19        return {
20            "city": address["town"],
21            "region": address["state"],
22            "size": "Kleinstadt"
23        }
24    elif 'village' in address:
25      region = None
26      if 'county' in address:
27        region = address["county"]
28      if 'state' in address:
29        region = address["state"]
30      return {
31            "city": address["village"],
32            "region": region,
33            "size": "Kleinstadt"
34        }
35    else:
36        return dict()
```

Listing 3.1: Einfaches Recommendation System für Film vorschläge

Die Funktion *get_region_city_size* nimmt als Parameter die Long- und Latitude Werte und erzeugt dadurch ein *Dictionary* mit den Werten *region*, *city* und *size*.

Beschaffung von der Hotelart

Einer der wichtigsten Eigenschaften, die ein Hotel vorweisen kann, ist die Hotelart von dem jeweiligen Hotel. Die Hotelart hängt maßgeblich mit der zur grundlegenden Preisgestaltung ab. Dies ist einfach zu erklären, da Hotels existieren, die eher auf Wellness ausgelegt sind und somit prinzipiell teurer sind als einfache Urlaubshotels. Somit ist die Art eines Hotels essentiell um ähnliche Hotels zu finden. So soll auch für das Modell die Hotelart vorhanden sein. Dieses Feature muss jedoch erst beschafft werden, da diese Information nicht in der Datenbank hinterlegt ist.

Leider ist der Versuch, die Hotelart auf einem automatisiertem Weg zu bekommen, gescheitert und es blieb nichts anderes übrig als die Hotelart eines jedem Hotels manuell herauszufinden.

Nachdem die Hotelart beschaffen wurde, sieht der Feature-Datensatz wie folgt aus:

company_id	median_min	median_max	areatype_count	area_count	region	city	art
6284f18969588aff1c3ace83	8900.0	19900.0	8	147	Berlin	Berlin	Stadthotel
615f129a52ab7509853d352c	7400.0	23400.0	3	27	Baden-Württemberg	Karlsruhe	Stadthotel
6155bdf03707c0492904de86	8000.0	31000.0	4	19	Bayern	Erlangen	Wellnesshotel
5fc4e11c3c09ba0c169c93b3	9000.0	30000.0	2	188	Berlin	Berlin	Businesshotel
627232f4ac34477a98aba6af	4900.0	12900.0	9	61	Rheinland-Pfalz	Trier	Stadthotel

Abbildung 3.4: Alle schon vorhanden Features 2

3.1.2. Datenvorverarbeitung

Nach der Datenbeschaffung erfolgt die Datenvorverarbeitung, die darauf abzielt, die Qualität und Integrität der Daten sicherzustellen. Ein zentraler Schwerpunkt liegt dabei auf der Identifikation und Eliminierung fehlerhafter oder ungültiger Datensätze. Häufig wird innerhalb der Datensätze nach Nullwerten gesucht, um diese entweder durch valide Daten zu ersetzen oder in einigen Fällen gänzlich zu entfernen. Im vorliegenden Fall ist es Wichtig, dass sämtliche als verwendbar gekennzeichneten Hotels in die Analyse einbezogen werden. Die Datensätze sollen nicht einfach verworfen werden; vielmehr erfolgt eine gezielte Substitution von Nullwerten durch valide Daten, um eine konsistente und zuverlässige Grundlage für die weiterführende Analyse zu gewährleisten [3].

Im folgenden ist die Anzahl aller Nullwerte innerhalb des Datensatzes zu sehen:

```
median_min        0
median_max        0
areatype_count    0
area_count        0
region            0
city              0
art               0
dtype: int64
```

Abbildung 3.5: Summe aller Nullwerte im Datensatz

In Abbildung 3.5 ist zu sehen, dass innerhalb des es Datensatzes keine Nullwerte gibt und somit auch keine Datenvorverarbeitung getätigt werden muss.

3.1.3. Allgemeine Datenanalyse

Die Datenanalyse ist ein, wenn nicht sogar der Wichtigste schritt im Data Science Bereich [3]. Dabei sollen die Daten ergründet und verstanden werden. Dieser Schritt soll nicht nur ein Verständnis für die vorliegenden Daten schaffen, sondern auch *outlier* erfassen. Meist wird auch versucht innerhalb der Datenanalyse zusammenhänge zu der Zielvariable zu finden. Dies setzt jedoch voraus, dass eine Zielvariable vorhanden ist. In dem vorliegenden Fall existiert keine Zielvariable, da nicht bekannt ist welche Hotels mit welchen Hotels ähnlich sind. Aufgrund dessen, dass keine Zielvariable vorhanden ist, wird die folgende Datenanalyse lediglich dazu genutzt um die Daten besser zu verstehen. Zudem soll festgestellt werden, welche Daten wie als Features verwendet werden können.

Für die Datenanalyse soll die Python-Bibliothek *Seaborn* benutzt werden. Die *Seaborn*-Bibliothek stellt ein leistungsstarkes Werkzeug dar, das speziell für die Erstellung von statistischen Grafiken in Python entwickelt wurde [4]. Das Hauptziel besteht darin, die verschiedenen Verteilungen der Features auf anschauliche Weise darzustellen, um einen umfassenden Überblick über die zugrunde liegenden Daten zu ermöglichen. Durch die Nutzung der Funktionalitäten von *Seaborn* wird eine effiziente und ästhetisch ansprechende Visualisierung erreicht, die es ermöglicht, Muster, Ausreißer oder Trends in den Daten leichter zu identifizieren.

Region Features

Zuallererst soll ein grober Überblick über die Städte der Hotels innerhalb der Datenbank erstellt werden. Dazu wird innerhalb des Datensatzes nach der Stadt um die Anzahl der Hotels einer Stadt zu ermitteln.

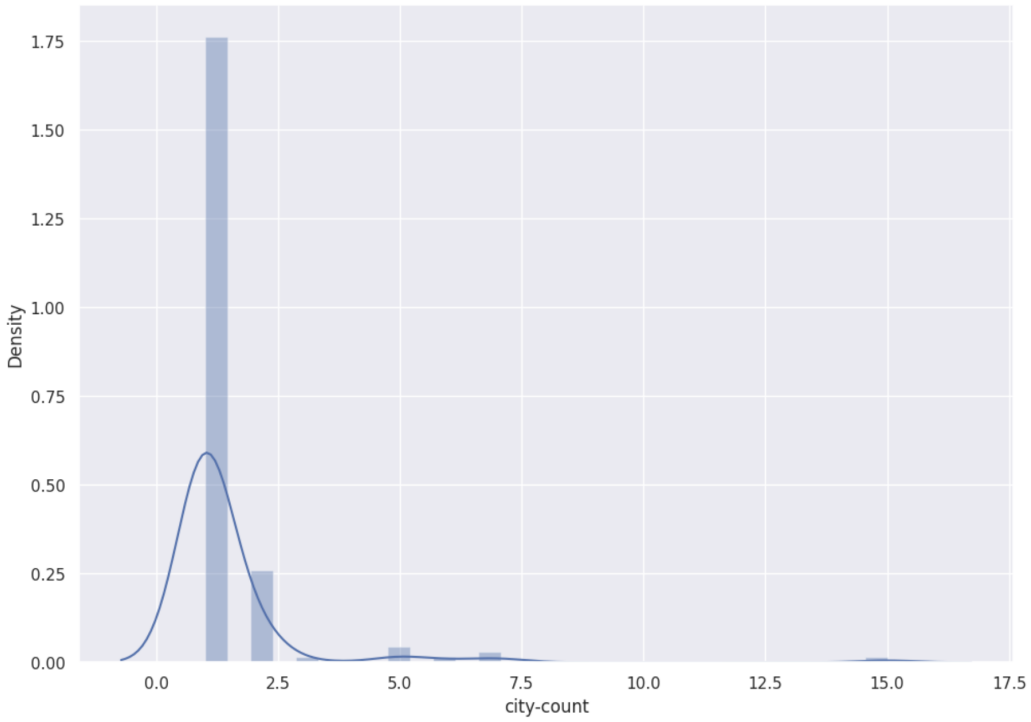

Abbildung 3.6: Verteilung der Städte

Die Analyse der Abbildung 3.6 offenbart, dass die überwiegende Mehrheit der in der Datenbank erfassten Städte lediglich über ein einziges Hotel verfügt, welches happyhotel in Anspruch nimmt. Zugleich verdeutlicht die Abbildung 3.6, dass es lediglich einen geringen Anteil von Städten gibt, in denen fünf oder mehr Hotels registriert sind. Dies ist etwas problematisch, da das Feature *City* zu unausgewogen ist und so nicht mit in das Modell gegeben werden kann. Es muss also in einer anderen Form verwendet werden.

Momentan werden, wie in der Abbildung 3.4 gezeigt, die Stadt und die Region als zwei separate Features aufgelistet. Die Idee ist es nun die zwei Features zu verschmelzen und die Hotels in Regionen aufzuteilen um mehr Informationen zu erhalten. Anstatt also Region und Stadt separat zu haben soll es ein Feature Region geben, welches wie folgt aufgebaut ist:

- Befindet sich innerhalb einer Stadt Fünf oder mehr Hotels, so wird die Stadt ohne jegliche Modifikation als Region genommen.
- Befindet sich innerhalb einer Stadt weniger als Fünf Hotels, so setzt sich die Region aus der ursprünglichen Region, also dem Bundesland und der Größe der Stadt zusammen nach dem Schema: Region-Größe

Für die Idee muss zunächst ermittelt werden, in welchen Städte sich Fünf oder mehr Hotels befinden. Auch diese Information kann wie folgt Visualisiert werden:

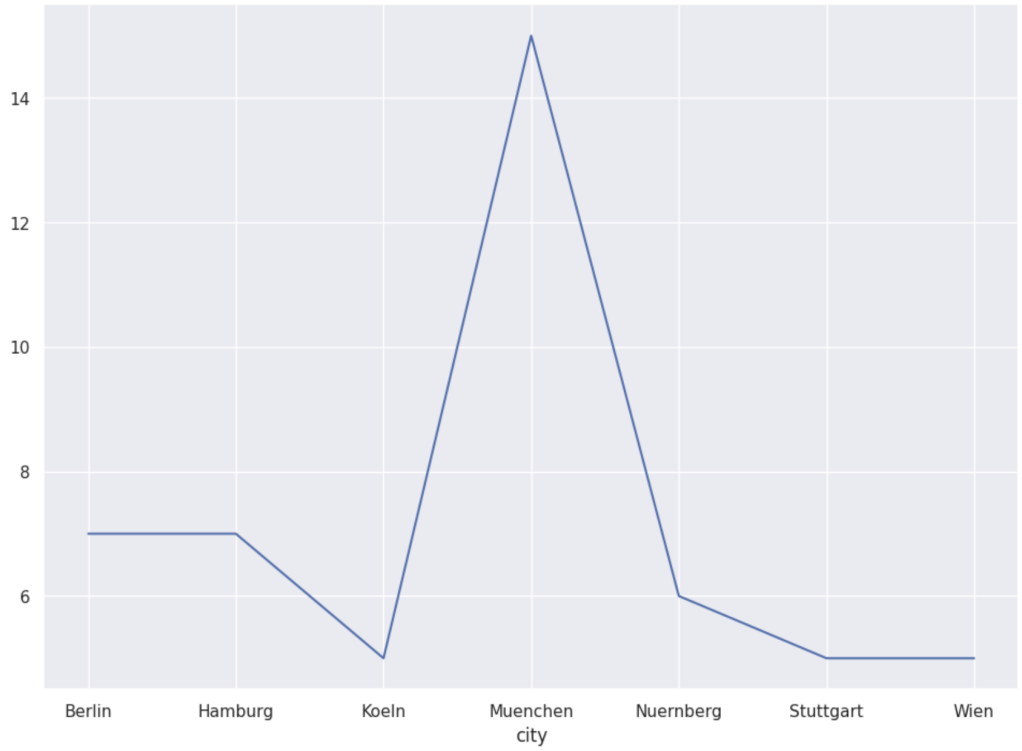

Abbildung 3.7: Städte mit Fünf oder mehr Hotels

Ganz klar zu erkennen ist, dass die Sieben Städte Berlin, Hamburg, Köln, München, Nürnberg, Stuttgart, und Wien die Städte sind, in denen Fünf oder mehr Hotels vorhanden sind. Die aufgelisteten Städte können dementsprechend so übernommen werden und für alle anderen wird die Regel von oben angewandt.
Nach der Umformulierung von Region und Stadt sieht der Datensatz wie folgt aus:

company_id	median_min	median_max	areatype_count	area_count	art	region2
6247262ad7da93fb2b9ac50d	10500.0	29500.0	7	18	Stadthotel	Wien
5f7db0bb34ee036332daffea	12000.0	60000.0	5	38	Businesshotel	Berlin
5faac8b2448f3913904ffaed	9500.0	35000.0	6	205	Businesshotel	Muenchen
637353f228d9f119bb9e6d2b	12900.0	52900.0	12	75	Stadthotel	Nuernberg
6492f9f68b5ac216293bee08	7500.0	8300.0	5	9	Familienhotel	Bayern-Kleinstadt
65005a8d765491b555190edb	9150.0	17150.0	9	36	Ferienhotel	Baden-Wuerttemberg-Grossstadt
63791654c755fe850934264c	12900.0	30000.0	8	26	Familienhotel	Niedersachsen-Kleinstadt
641039e8b6b33774ec0335d5	11500.0	30000.0	10	189	Apartementhotel	Berlin
61420e574f15836e1a37a173	7900.0	29000.0	3	28	Businesshotel	Nuernberg
6221d0f96b58f774c9890657	8900.0	14900.0	2	42	Apartementhotel	Bayern-Kleinstadt

Abbildung 3.8: Datensatz nach der Umformulierung

Auch hier soll wieder die Verteilung visualisiert werden

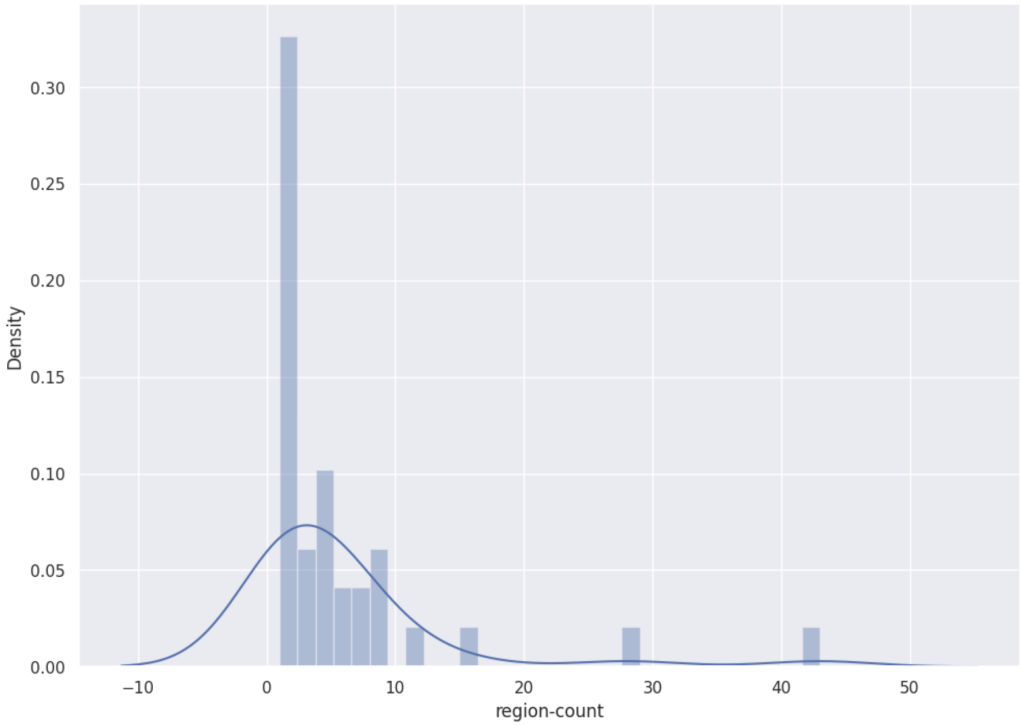

Abbildung 3.9: Verteilung des neuen Features *region2*

Es zeigt sich, dass noch immer ein großer Anteil des Datensatzes niedrigeren Bereich sich befindet, jedoch konnte mit der Modifikation ein bisschen mehr Varianz in die Daten gebracht werden.

Preis Features

Als nächstes sollen die Preis Features, namentlich betitelt mit *median_min* und *median_max*, erkundet und visualisiert werden. Hierzu soll wie auch bei der Region zunächst die Verteilung betrachtet werden:

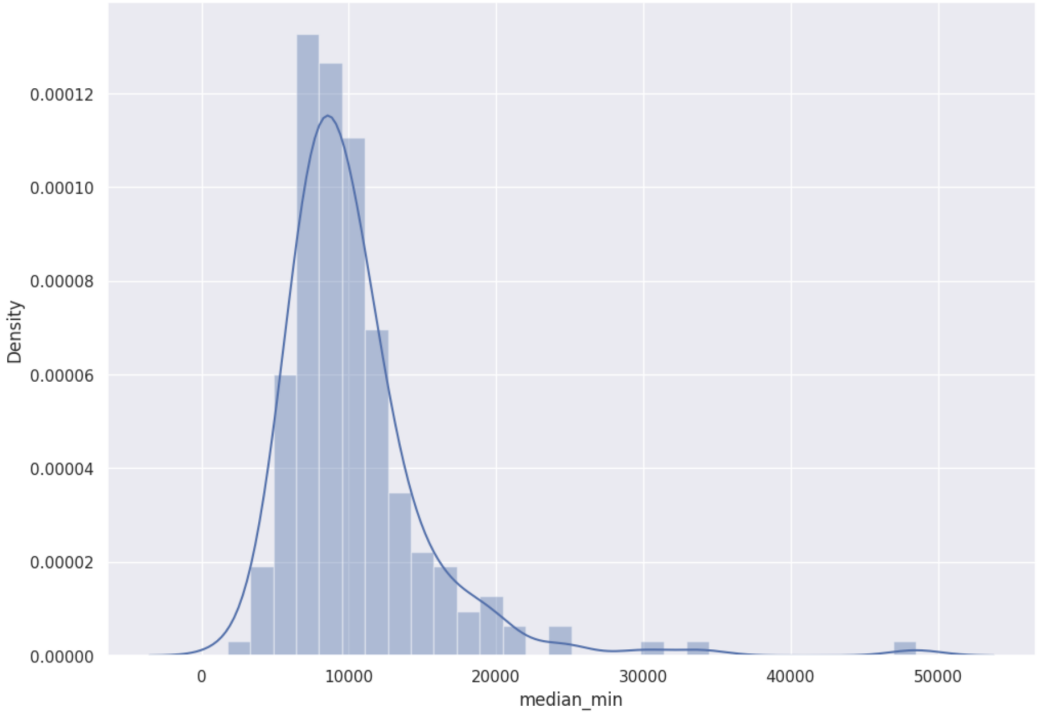

Abbildung 3.10: Verteilung von dem Preis Features *median_min*

Abbildung 3.10 zeigt, dass das Feature *median_min* eine Normalverteilung mit ein paar outlier aufweist. Eine Normalverteilung sagt aus, dass die Verteilung mehr Daten um den Mittelwert herum aufweist. Die Datenverteilung nimmt ab, wenn sich vom Zentrum entfernt wird. Die resultierende Kurve ist symmetrisch zum Mittelwert und bildet eine glockenförmige Verteilung [5].

Eine weitere interessante Information die noch aus dem Feature *median_min* gelesen werden kann, ist der Durchschnittliche Wert Region. Dazu soll der Datensatz nach der Region gruppiert werden und den Durchschnittlichen Wert ermittelt werden.

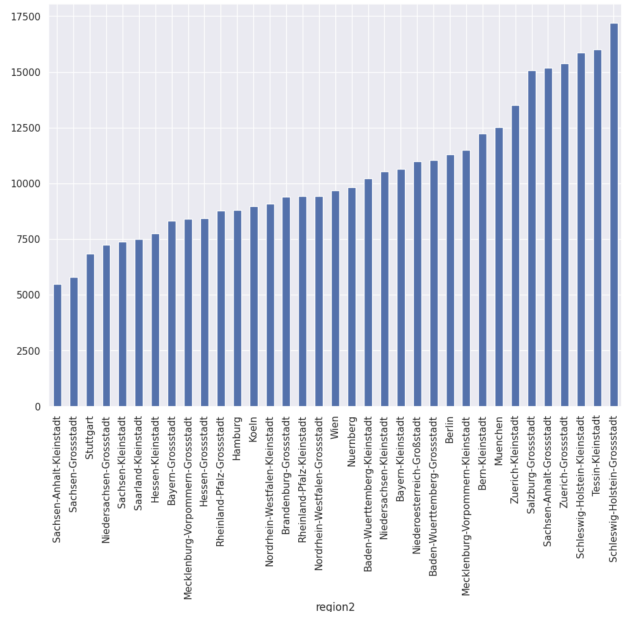

Abbildung 3.11: Durchschnittlicher minimal Preis pro Region

Die Erkenntnis die aus der Abbildung 3.11 genommen werden kann, ist die, dass es einen deutlichen unterschied macht, in welcher Region das Hotel liegt wenn auf den Minimalen Median Preis des Hotel geschaut wird. Das gleich kann auch mit dem Maximalen Median Preis eines Hotels gemacht werden. Auch hier soll sich zunächst die Verteilung visualisiert werden:

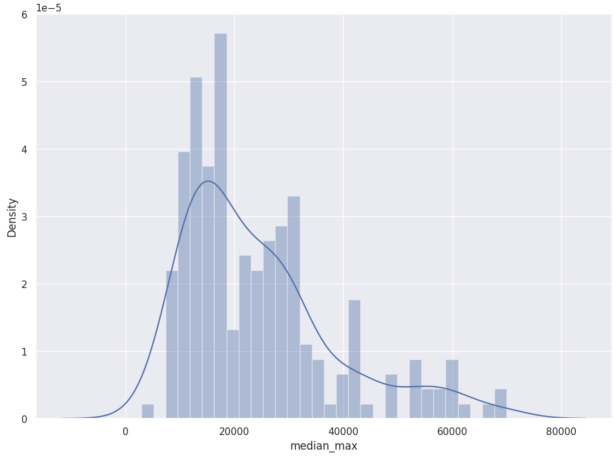

Abbildung 3.12: Verteilung von dem Preis Features *median_max*

Anders als bei dem Minimalen Median Preis, kann bei dem Maximalen Median Preis keine Normalverteilung erkannt werden. Hier wirken die Werte recht verstreut.

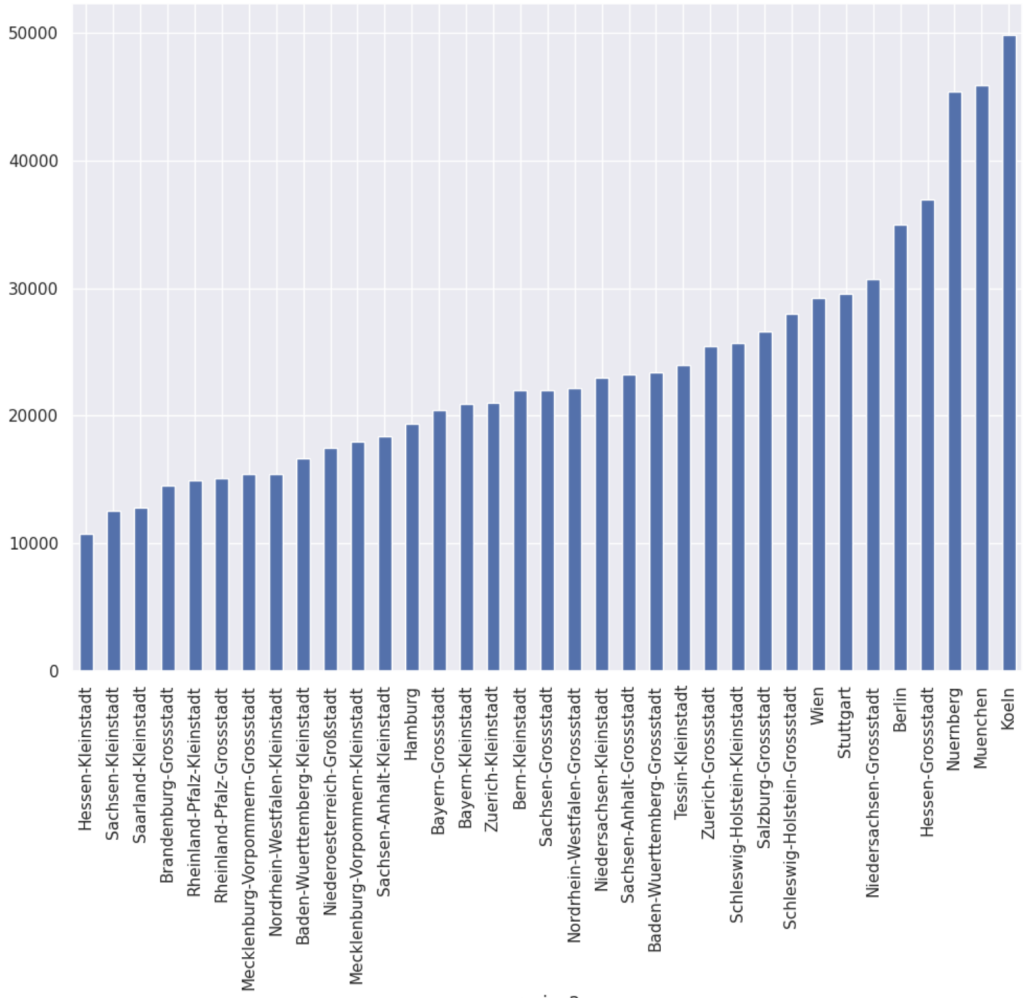

Abbildung 3.13: Durchschnittlicher maximal Preis pro Region

Abbildung 3.13 zeigt, dass es auch deutliche Unterschiede bei den einzelnen Regionen gibt. Eine weitere interessante Erkenntnis ist die, dass es bei dem Maximalen Median Preis eher so ist, dass die Großstädte wie Köln und München eher zu einem höheren Maximalen Preis tendieren.

Hotelart Feature

Die Hotelart bildet einen essenziellen Bestandteil, um umfassende Einblicke in die Charakteristiken eines Hotels zu gewinnen. Sie liefert nicht nur Informationen über den Zweck und die Ausrichtung der Unterkunft, sondern ermöglicht auch eine präzise Identifikation der Zielgruppe, die das Hotel anspricht [6]. Die Zielgruppe eines Hotels ist zudem eine wertvolle Information wenn es darum geht Preise für das

Hotel zu gestalten, zumindest ist so die Annahme. Auch in diesem Fall kann wieder nach der Hotelart gruppiert werden und jeweils der Minimale und Maximale durchschnittliche Preis angezeigt werden.

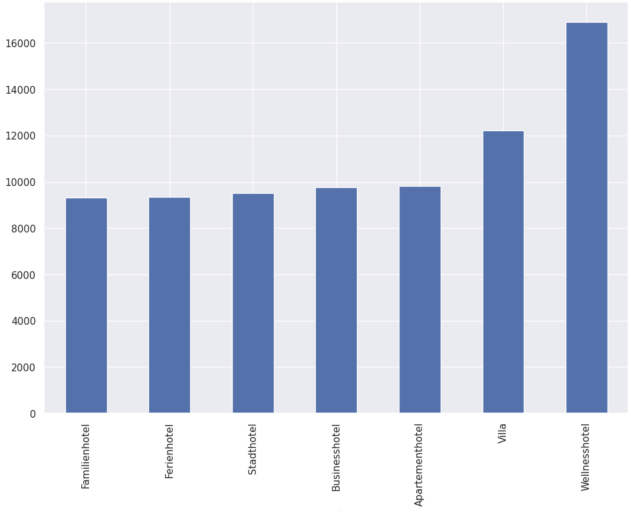

Abbildung 3.14: Durchschnittlicher minimal Preis pro Hotelart

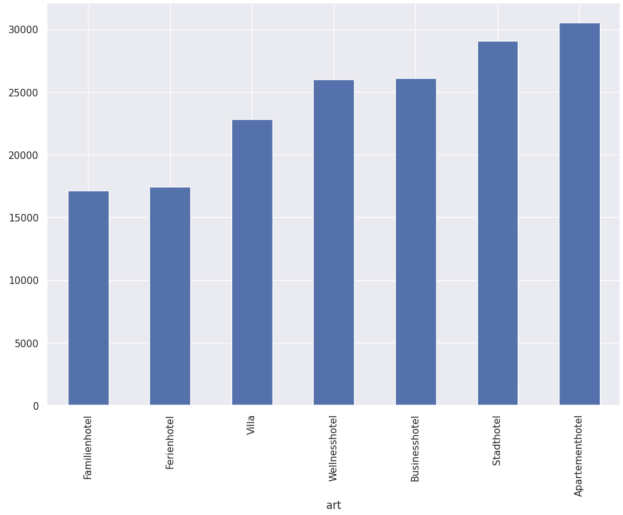

Abbildung 3.15: Durchschnittlicher maximal Preis pro Hotelart

Anhand von den zwei Abbildungen 3.14 und 3.15 zeigt sich, dass tatsächlich einen unterschied bei den Preisen auf die Hotelart bezogen gibt. Zudem zeigt sich, dass die zwei Hotelarten *Ferienhotel* und *Familienhotel* in beiden Fällen die gleiche Information wiedergibt und somit auch zu *Ferienhotel* zusammengefasst werden kann. Zudem könnten anhand von *median_min* noch weitere Hotelarten zusammengefasst werden, jedoch wenn beide Informationen zusammen betrachtet werden, so bleibt es lediglich bei *Ferienhotel* und *Familienhotel*.

Zimmer Features

Die letzten zwei Features innerhalb des Datensatzes sind die Features *area_count* und *areatype_count* welche die Größe des jeweiligen Hotels repräsentieren. Die Frage die sich hier also stellt ist, ob das Preisverhältnis in irgendeiner Art mit der Größe des Hotels zusammenhängen kann. Hierzu soll zunächst auch erstmal die Verteilung der Zimmeranzahl angeschaut werden:

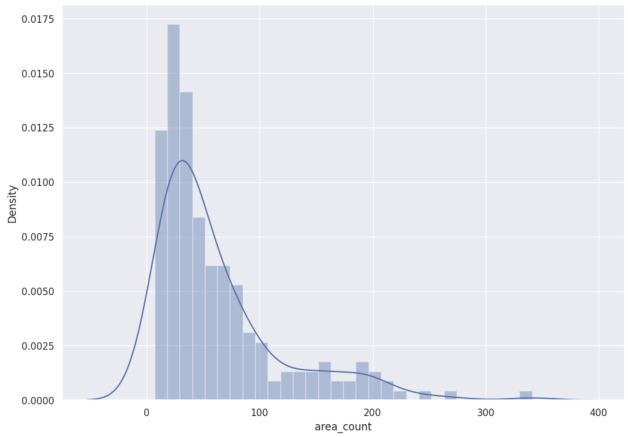

Abbildung 3.16: Verteilung nach Zimmeranzahl

Die Verteilung zeigt, dass auch recht ausgeprägt ist und nur im Ansatz einer Normalverteilung gleicht. Des Weiteren soll nach der Anzahl der Zimmer gruppiert werden um zu überprüfen wie oft jeder Anzahl vorkommt:

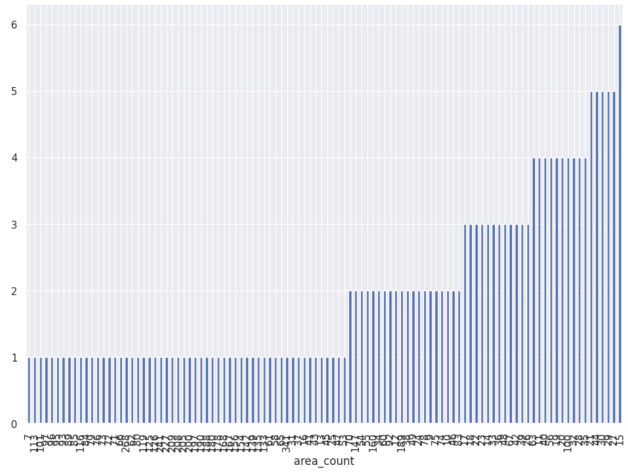

Abbildung 3.17: Häufigkeit der Zimmeranzahl im Datensatz

Abbildung 3.17 hat gezeigt, dass das Feature *area_count* zu ausgeprägt ist. Aufgrund dessen, dass das Feature zu ausgeprägt ist und keine Idee vorhanden ist, wie

dieses Feature umformuliert werden könnte, wurde beschlossen *area_count* und *areatype_count* aus dem Datensatz zu entfernen.

Der finale Datensatz welcher für das Modell benutzt werden soll, sieht wie folgt aus:

company_id	median_min	median_max	art	region2
6247262ad7da93fb2b9ac50d	10500.0	29500.0	Stadthotel	Wien
5f7db0bb34ee036332daffea	12000.0	60000.0	Businesshotel	Berlin
5faac8b2448f3913904ffaed	9500.0	35000.0	Businesshotel	Muenchen
637353f228d9f119bb9e6d2b	12900.0	52900.0	Stadthotel	Nuernberg
6492f9f68b5ac216293bee08	7500.0	8300.0	Ferienhotel	Bayern-Kleinstadt
65005a8d765491b555190edb	9150.0	17150.0	Ferienhotel	Baden-Wuerttemberg-Grossstadt
63791654c755fe850934264c	12900.0	30000.0	Ferienhotel	Niedersachsen-Kleinstadt
641039e8b6b33774ec0335d5	11500.0	30000.0	Apartementhotel	Berlin
61420e574f15836e1a37a173	7900.0	29000.0	Businesshotel	Nuernberg
6221d0f96b58f774c9890657	8900.0	14900.0	Apartementhotel	Bayern-Kleinstadt

Abbildung 3.18: Finaler Datensatz für das Modell

3.1.4. Evaluation der Ähnlichen Hotels

Der vorliegende Datensatz ist nun verfügbar, und grundsätzlich kann die Modellierung fortgesetzt werden. Allerdings stellt sich die Frage, ob die identifizierten Hotels tatsächlich ähnlich zum ursprünglichen Hotel sind. Es ist von entscheidender Bedeutung, nachzuweisen, dass die ausgewählten Hotels auf irgendeine Weise miteinander vergleichbar sind. Aus diesem Grund wird im nachfolgenden Abschnitt ein Mechanismus entwickelt, um die Ähnlichkeit der Hotels zu überprüfen und zu gewährleisten.

Angesichts des angestrebten Ziels, nämlich der dynamischen Generierung von Preisen, wurde zunächst in Erwägung gezogen, die Preise der einzelnen Hotels zu vergleichen. Zu diesem Zweck wurde initial ein *Dataframe* erstellt, das sämtliche gültigen Hotels sowie ihre Preisinformationen für einen bestimmten Zeitraum umfasst.

	5eff10592f24eb0df2ef4825	5ea83ceb60b7a90e412a8047	612dd09e6abd9a3651042cb3	614c8b5e3707c081fcbd48e8	6006f4878c1fe4571a439326
0	6900.0	7150.0	99900.0	4800.0	13050.0
1	7000.0	7350.0	99900.0	4600.0	11950.0
2	7000.0	7550.0	99900.0	4700.0	11950.0
3	7000.0	7850.0	99900.0	4600.0	11950.0
4	7000.0	7500.0	99900.0	4700.0	11950.0
5	7000.0	7350.0	99900.0	4700.0	11950.0
6	7000.0	7350.0	99900.0	4700.0	11950.0
7	7000.0	7150.0	99900.0	4600.0	11950.0
8	7000.0	7350.0	7400.0	4600.0	11950.0
9	7000.0	7550.0	7900.0	4600.0	13750.0

Abbildung 3.19: Preise von allen Hotels für das Jahr 2022

Abbildung 3.19 zeigt einen exemplarischen Auszug aus dem DataFrame. Zudem wurde anhand diesem DataFrame noch die dazugehörige Korrelationsmatrix erstellt. Die Korrelationsmatrix ist dafür da um zusammenhänge zwischen den Vektoren zu finden. Dabei beschreibt ein Wert nahe 1 einen hohen positiven Zusammenhang der zwei Vektoren und ein Wert nahe -1 einen hohen negativen Zusammenhang der zwei Vektoren. Ein Korrelationswert gegen 0 beschreibt beschreibt keinerlei Zusammenhang der Vektoren [7]. Die Vermutung ist es, dass wenn die Preise von 2 Hotels korrelieren und die Preise sich überschneiden, so werden dass ähnliche Hotels sein.

Diese Vermutung soll dementsprechend mit einigen Hotels getestet werden:

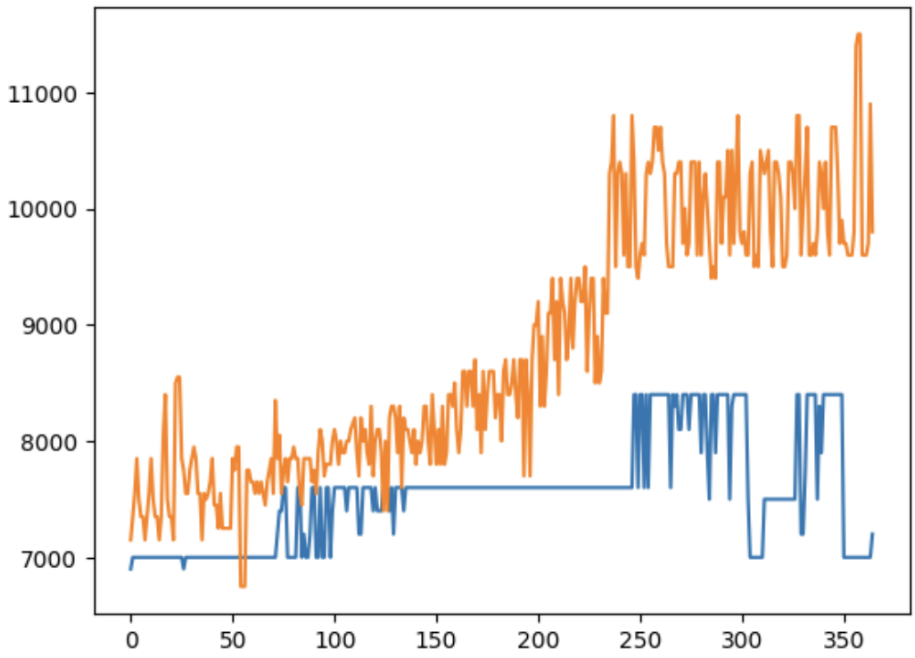

Abbildung 3.20: Visualisierung der Preise zweier Hotels

	5eff10592f24eb0df2ef4825	5ea83ceb60b7a90e412a8047
5eff10592f24eb0df2ef4825	1.000000	0.600991
5ea83ceb60b7a90e412a8047	0.600991	1.000000

Abbildung 3.21: Korrelationswerte der zwei Hotels

Abbildung 3.20 illustriert den Preisverlauf von zwei Hotels im Jahr 2022, während Abbildung 3.21 die Korrelation zwischen diesen beiden Hotels zeigt. Der festgestellte Korrelationswert von 0,6 erweist sich als bemerkenswert, insbesondere vor dem Hintergrund, dass die Preise in Abbildung 3.20 beträchtlich voneinander abweichen. Infolgedessen wurde die Überlegung angestellt, die Preise zu skalieren und daraufhin miteinander zu vergleichen. Entscheidend für ähnliche Hotels ist lediglich die Tendenz wie sich die Preise verhalten.

Werden die Preise nun Skaliert ändert sich an der Korrelation nichts und die Skalierten Preise sehen nun wie folgt aus:

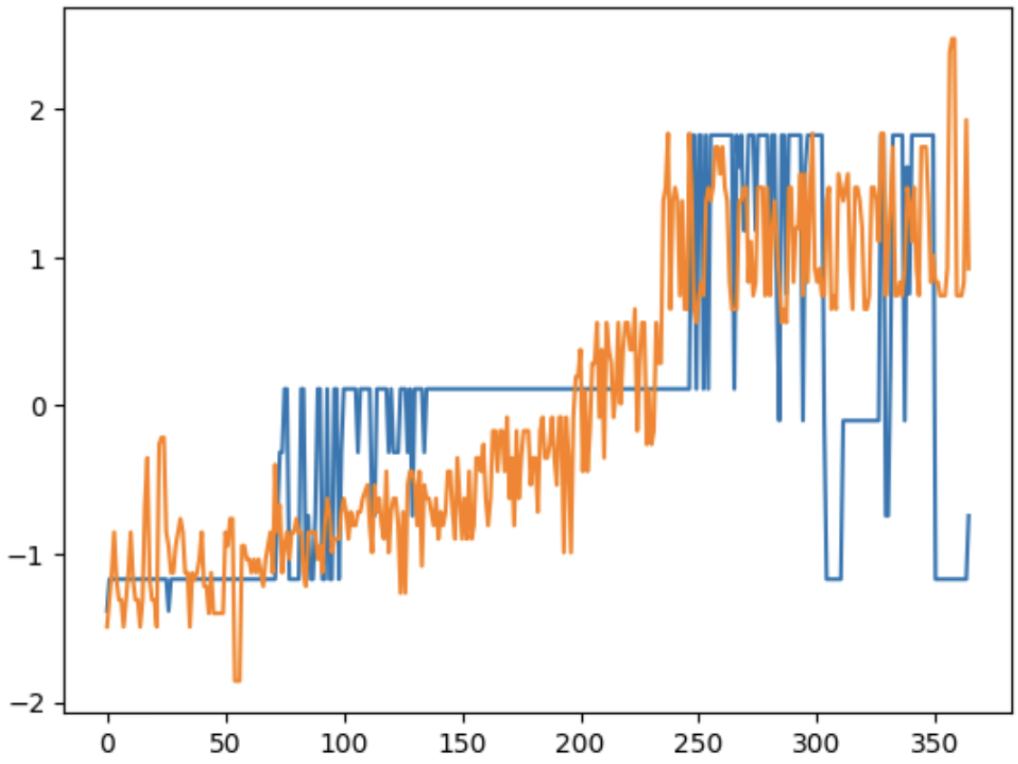

Abbildung 3.22: Visualisierung der Skalierten Preise zweier Hotels

Eine zusätzliche Betrachtung ergab die Frage, ob ein ähnliches Muster nicht auch durch die Verwendung des RevPAR erzielt werden könnte. Die Verwendung des

RevPAR-Werts erscheint in diesem Kontext sinnvoller als die ausschließliche Berücksichtigung der Zimmerpreise, da das nachfolgende Modell letztendlich darauf abzielt, den RevPAR-Wert vorherzusagen.

Im folgenden werden die gleichen zwei Hotels mit dem RevPAR-Wert verglichen:

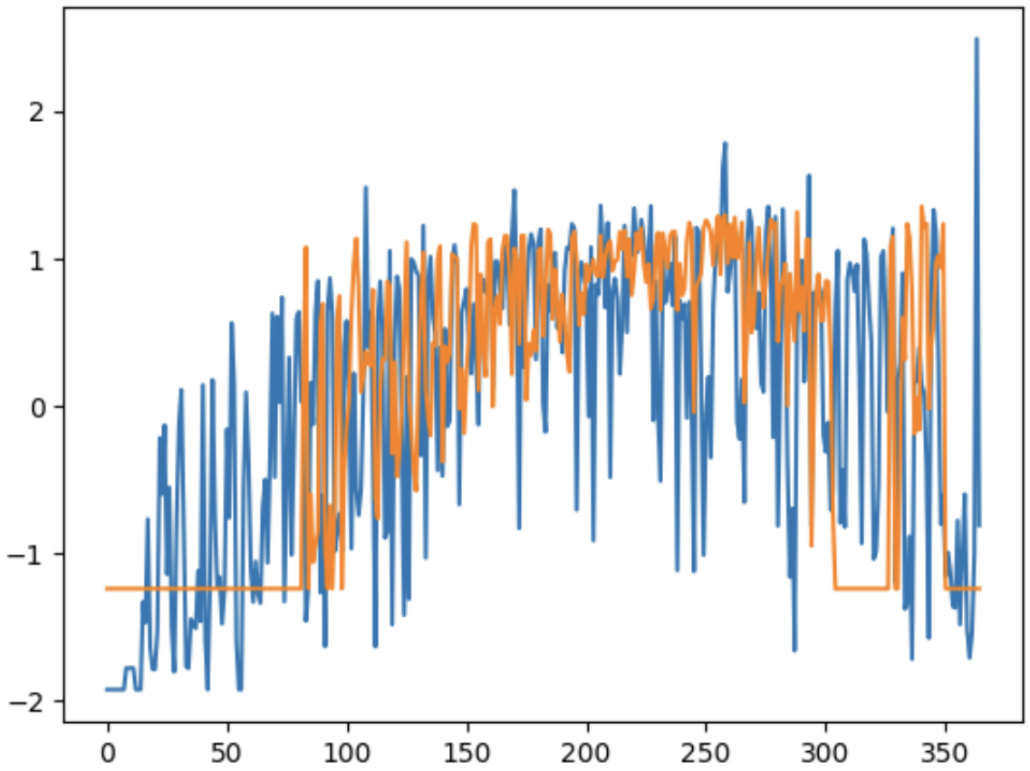

Abbildung 3.23: Visualisierung der Skalierten RevPAR-Werte zweier Hotels

	5eff10592f24eb0df2ef4825	5ea83ceb60b7a90e412a8047
5eff10592f24eb0df2ef4825	1.00000	0.26945
5ea83ceb60b7a90e412a8047	0.26945	1.00000

Abbildung 3.24: Korrelationswerte der RevPAR-Werte zweier Hotels

Abbildung 3.24 offenbarte einen abweichenden Korrelationswert im Vergleich zu demjenigen, der bei der Betrachtung der Preisentwicklung ermittelt wurde. Daraufhin wurde die Entscheidung getroffen, dass der Korrelationswert der RevPAR-Werte als aussagekräftiger betrachtet wird als derjenige der reinen Preisentwicklung. Infolgedessen wurde festgelegt, dass dieser Wert als Evaluation für die Ähnlichkeit zwischen den Hotels herangezogen wird.

Dieser Korrelationswert der RevPAR-Werte kann lediglich zur Evaluation der Modelle verwendet werden, da dieser Korrelationswert für ein Hotel nicht vorhanden ist.

Tabellenverzeichnis

Abbildungsverzeichnis

Listings

Literatur

[1] M. KARATAŞ und A. Einstein, *If You Want to Know the Future, Look at the Past.* Amazon Digital Services LLC - KDP Print US, 2017, ISBN: 9781976758331. Adresse: https://books.google.de/books?id=u_14twEACAAJ.

[2] Amazon Web Services, Inc., *Was ist Datenwissenschaft? – Datenwissenschaft erklärt – AWS*, 15.11.2023. Adresse: https://aws.amazon.com/de/what-is/data-science/.

[3] A. Agarwal, „Linear Regression on Boston Housing Dataset - Towards Data Science“, *Towards Data Science*, 5.10.2018. Adresse: https://towardsdatascience.com/linear-regression-on-boston-housing-dataset-f409b7e4a155.

[4] Melanie, *Seaborn: Alles über das Python-Tool zur Datenvisualisierung - Weiterbildung Data Science | DataScientest.com*, 2023. Adresse: https://datascientest.com/de/seaborn-alles-ueber-das-python-tool-zur-datenvisualisierung.

[5] Shrishty, „What Is Normal Distribution & Standard Deviation in Statistics“, *Simplilearn*, 5.08.2021. Adresse: https://www.simplilearn.com/tutorials/statistics-tutorial/what-is-normal-distribution.

[6] S. User, *Die wichtigsten Hotelarten im Überblick*, 20.01.2024. Adresse: https://www.toursol.at/de/willkommen/toursol-blog/99-hotelarten.

[7] D. S. Team, „Was ist eine Korrelationsmatrix?“, *Data Science*, 3.05.2020. Adresse: https://datascience.eu/de/mathematik-statistikatik/was-ist-eine-korrelationsmatrix/.

A. Anhang

Test