Hochschule Karlsruhe
University of
Applied Sciences

HKA

DSCB420 - Projektbericht

SoSe 2024

67950, 79849, 88692, 71254

Executive Summary

In der Gastronomie können externe Faktoren wie das Wetter einen erheblichen Einfluss auf die Besucherzahlen und damit auf den Umsatz haben. Ein Restaurantbesitzer möchte die Zusammenhänge zwischen den Wetterbedingungen und den täglichen Umsätzen seines Betriebes besser verstehen und nutzen, um betriebliche Entscheidungen zu optimieren. Durch die Analyse historischer Umsatzdaten in Kombination mit Wetterdaten können Muster und Trends erkannt werden, die eine fundierte Vorhersage zukünftiger Umsätze ermöglichen.

Mit unserer App wollen wir unseren Nutzern Empfehlungen und einen Überblick über ihre Gastronomieumsätze im Zusammenhang mit aktuellen Wetterdaten geben, indem wir Wetter- und Umsatzdaten nutzen. Durch detaillierte Berichte im Dashboard können wir die Umsätze der einzelnen Geschäfte mit den Wetterdaten vergleichen. Durch das Trainieren eines Modells können wir unseren Kunden auch eine Prognose geben, in welchem Zeitraum sie ihre Ressourcen weiter ausbauen können, um ihren Umsatz zu optimieren. Letztendlich wollen wir ein eigenes Produkt anbieten, das die Planung und Entscheidungsfindung erleichtert.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Anwendungsfall

Das Hauptziel dieses Anwendungsfalls ist die Entwicklung eines Modells, das den Umsatz eines Gastronomiegewerbes auf der Grundlage von Wetterdaten prognostiziert. Darüber hinaus soll die Anwendung dem Restaurantbesitzer Empfehlungen zu wetterbezogenen Themen im Zusammenhang mit seinem Umsatz geben. Dadurch soll der Restaurantbesitzer in die Lage versetzt werden, Ressourcen wie Personal und Vorräte effizient zu planen und gezielte Marketingstrategien zu entwickeln. Langfristig soll die Lösung helfen, Kosten zu senken und die Kundenzufriedenheit zu erhöhen. Außerdem wollen wir dem Nutzer einen detaillierten Bericht ausgeben, was zusätzlich die Planung des Geschäfts aushelfen soll.

2. Aufbereitung der Daten

Als Datenquellen standen die Gastronomieumsätze und Wetterdaten auf Tagesbasis zur Verfügung. Diese waren unvollständig, d.h. es fehlten Daten, es wurden Werte verwendet, die zu Fehlern bei der Analyse führen, oder es wurden unnötige Spalten verwendet, die keinen Analysezweck haben. Ziel der Aufbereitung ist es, die Datenquellen für die weitere Analyse zu filtern und zu transformieren.
Die typische Vorgehensweise beim Filtern von Datenquellen besteht darin, zunächst zu untersuchen, welche Spalten vorhanden sind und welche Fehler korrigiert werden müssen. Die Filterung und Transformation erfolgt über die Pandas-Bibliothek. Dabei wird aus Gründen der Architektur und der Historisierung jede Änderung an der Datei neu gespeichert und mit Datum/Uhrzeit versehen.
Jede Datenquelle hat ihr eigenes Jupyter-Notebook, in dem während des gesamten Filter- und Transformationsprozesses alle Dateien der Datenquelle in einer Schleife durchlaufen werden und so zumindest prototypisch auf Knopfdruck ausführbar sind.

2.1 Gastronomieumsaetze_flat.csv

Es liegen Daten vor, die die Umsätze in verschiedenen Gastronomiebetrieben von 1994 bis Anfang 2024 erfassen. Die Filterung besteht im Wesentlichen aus der Umwandlung von Spaltennamen und Werten sowie der Eliminierung mehrerer Dubletten. Das Ergebnis der Transformation ist eine Faktentabelle für Gastronomieumsätze.

2.2 Wetterdaten

Aus den zur Verfügung gestellten Wetterdaten wurden die Textdateien zu CSV-Dateien geändert und mit Pandas eingelesen. Die Wetterdaten sind nach Wetterstationen sortiert. Dabei besteht eine Wetterstation immer aus Metadaten und einer Datei, die die untersuchten Werte ausgibt. Das bedeutet, dass nur die Dateien “produkt_klima_tag_*” bereinigt und transformiert werden, da die Metadaten nur zur Behebung von Qualitätsproblemen geliefert wurden. Innerhalb der Dateien wurden für die Bereinigung hauptsächlich mit der Numpy Bibliothek fehlerhafte Werte zu NaN geändert, damit diese nicht bei Berechnungen zu Fehlwerten führen. Für die Transformation auf eine Monatsbasis wurde mit Pandas nach Datum gruppiert. Am Ende findet eine Harmonisierung statt, die alle Wetterstationen zu einer einzigen Faktentabelle zusammenfügt.

2.3 Geographiedaten

Zusätzlich wollten wir noch die Geodaten der Stationen auswerten, um zu schauen, ob die Temperaturdaten überhaupt repräsentativ sind. In den Metadaten befanden sich zusätzlich die Stationshöhe sowie die geografische Breite und Länge der einzelnen Stationen. Anhand dessen konnten wir mit Folium diese auf einer Deutschlandkarte visualisieren.

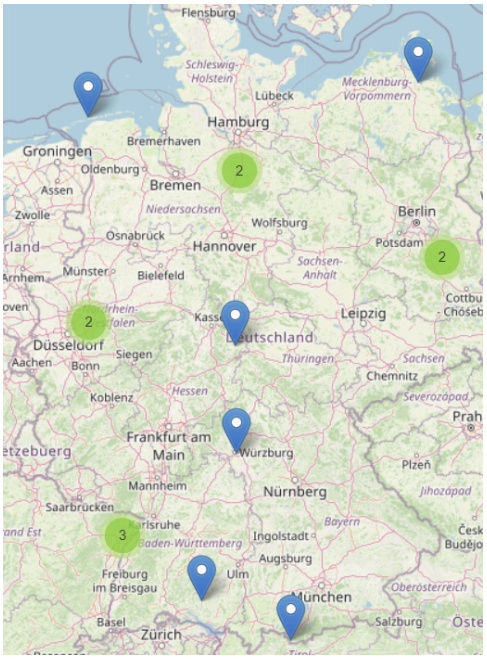

Abbildung 1: Karte der Wetterstationen

Im Allgemeinen ist die Station breit gestreut und in Deutschland werden die meisten Teile größtenteils abgedeckt. Jedoch gabs ausnahmen, wie die Stationen Bad Bergzabern und Pirmasens, die zwar ziemlich nah aneinander liegen, wir uns jedoch entschieden haben, diese wegen dem Höhenunterschied drin zu lassen. Zusätzlich haben wir noch ein mehrdimensionales Streudiagramm anhand der Daten im Zusammenhang mit der Durchschnittstemperatur erstellt. Hier wurde insbesondere auch die Höhe miteinberechnet.

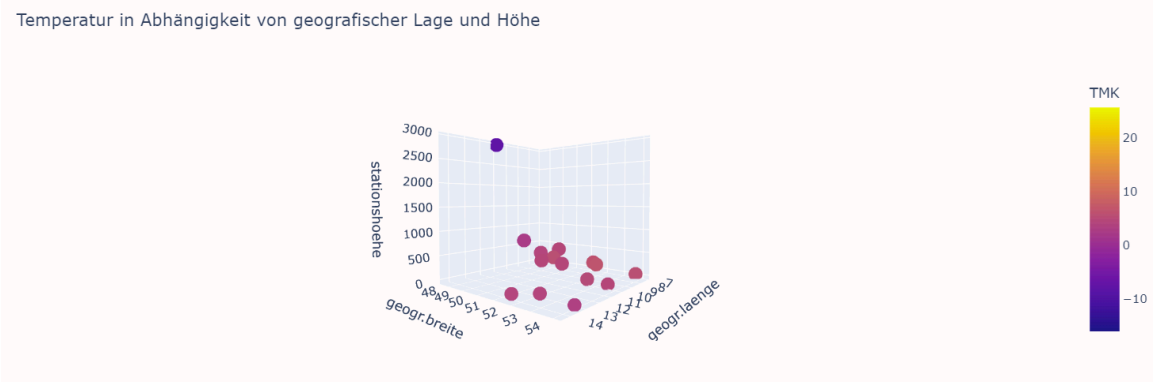

Abbildung 2: Ortsstreudiagramm der Wetterstation mit Temperatur

Die Zugspitze war hierbei ein sehr deutlicher Ausreißer. Letztendlich ließen wir sie zwar drin, da es durchaus Gastgewerbe in Deutschland gibt, die etwas höher gelegen sind und der deutsche Teil der Alpen auch nicht unbeachtet sein sollte. Die Erkenntnis hat uns trotzdem

dazu bewegt, für spätere Modellierung den Median statt den Durchschnittswert zu verwenden, um die Gewichtung des Ausreißers zu verringern.

Man konnte ansonsten auch eine Korrelation zwischen der Lage der Wetterstationen und der Temperatur herstellen. Dafür haben wir eine Korrelationsmatrix erstellt, die den Zusammenhang der Werte untereinander vergleicht.

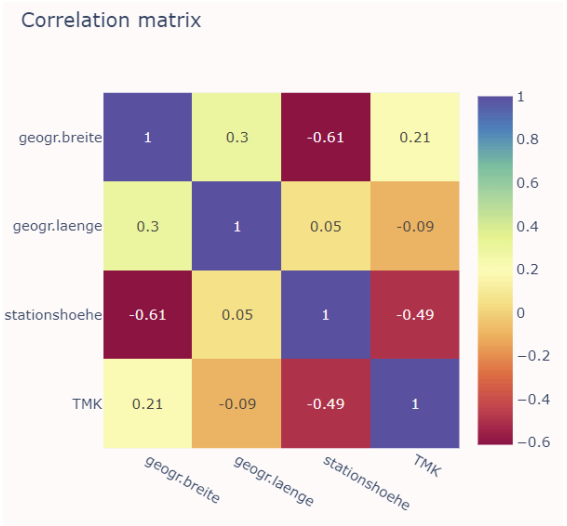

Abbildung 3: Korrelation der Wetterdaten mit den Temperaturwerten

Wie schon durch das mehrdimensionale Streudiagramm angedeutet, ist die Temperatur tendenziell niedriger, umso höher die Station liegt. Auf der anderen Seite spielt auch die geographische Breite eine Rolle. Je weiter nördlich die Wetterstation, desto höher sind tendenziell die Temperaturen. Wie bei der Korrelation zwischen Breite und Höhe ersichtlich wird, kann auch dies daran liegen, dass südliche Gebiete eher höher gelegen sind.

Beantwortung vorgegebener analytischer Fragen (Teil 1)

Zunächst wurden erst sowohl die Umsatzdaten der Campingplätze als auch die Medianwetterdaten als DataFrame importiert und anhand der monatlichen Datumswerte in eine Tabelle zusammengefügt.

Relation zwischen Temperatur und Campingplatz Umsätzen

Für die Temperatur wurde hier TMK genommen, welches das Tagesmittel der Lufttemperatur in zwei Metern Höhe darstellt.
Um schon mal vorab einen Zusammenhang graphisch zu prüfen, wurde ein Streudiagramm zwischen den Temperaturen und den Umsätzen erstellt. Dabei wurde eine Linie nach der Methode der kleinsten Quadrate benutzt, in welcher erkenntlich wird, dass eine deutliche lineare Beziehung zwischen beiden Variablen besteht.

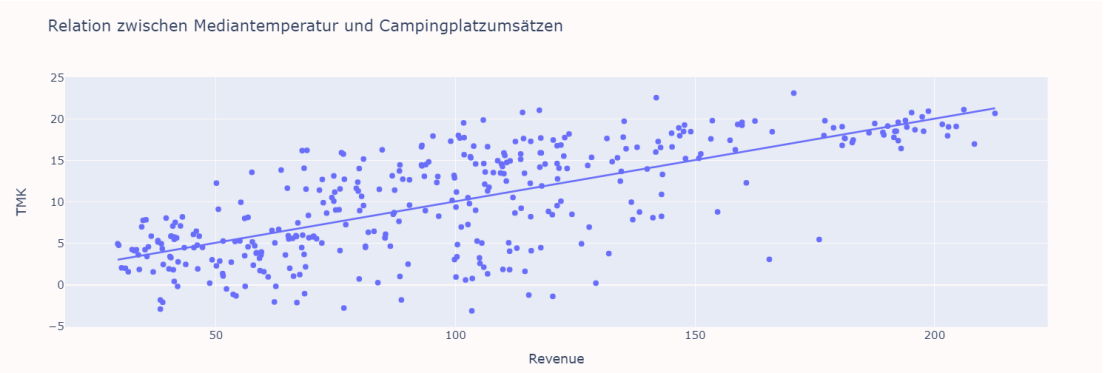

Abbildung 4: OLS-Trendlinie zwischen Campingumsätzen und Temperatur

Als nächstes wurde, da beide Variablen Verhältnisskaliert sind, der Person-Korrelationskoeffizient berechnet.

Mit r = 0.7 besteht eine hohe bis sehr hohe Korrelation
Mit p = 1.56^{-54} ist die Korrelation Hoch Signifikant, da p < 0.1

Mit dem niedrigen p-Wert besteht also eine sehr geringe Wahrscheinlichkeit, dass der Zusammenhang der Werte zufällig enstanden ist.

Somit existiert eine sehr klare und hochsignifikante Korrelation zwischen den Mediantemperaturen der vorgegebenen Wetterstationen und der Umsatzdaten der Campingplätze in Deutschland.

Relation mit zeitlicher Verzögerung

Mittels Autokorrelation wurde zusätzlich geprüft, wie hoch die Zusammenhänge von Umsatz und Temperatur sind, wenn man die Monate im Zeitraum von einem Jahr versetzt betrachtet. Hier wurde bei k=1, also nach einem Lag von einem Monat mit p_1= 0.69, der stärkste Zusammenhang festgestellt. Da dieser Wert allerdings kleiner als die Echtzeitkorrelation von p_0 = 0.7 ist, deutet das darauf hin, dass Menschen tendenziell spontan und möglicherweise Wettervorhersagen basiert, Campingplätze buchen. So könnten Campingplatzbetreiber mit der Information ihre Preise der Temperatur entsprechend anpassen. Auf der anderen Seite zeigt die Autokorrelation auch, dass es einen sehr starken saisonalen Effekt zu geben scheint. Die halbjährigen Lags sind nämlich am niedrigsten, während man nach ca. acht Monaten Verzögerung wieder einen Anstieg der Korrelation betrachten kann.

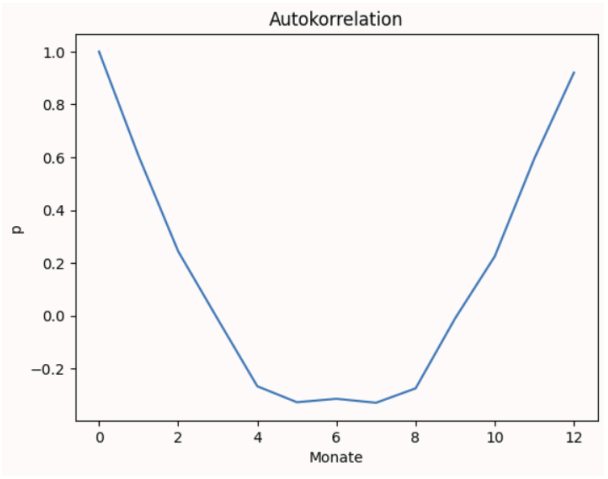

Abbildung 5: Autokorrelation zwischen Umsätzen und Temperatur

Relation zu allen Wettermesswerten

Im nächsten Schritt wurden alle Spalten der Wetterdaten genommen und auf eine Korrelation geprüft. Dafür bot sich eine Korrelationsmatrix an, mit welcher man vorher farbig einen ungefähren Eindruck bekommt, welche Werte wie stark zusammenhängen

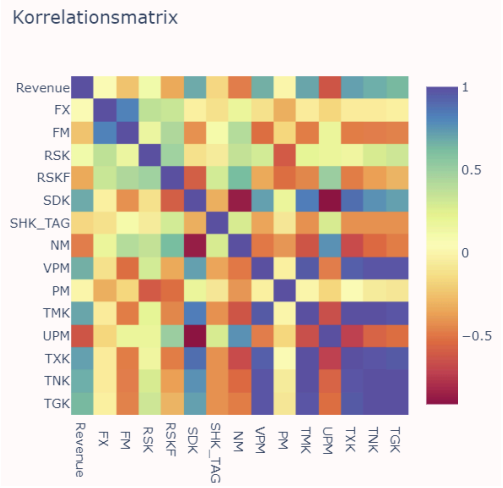

Abbildung 6: Korrelationsmatrix der Wetterdaten mit Umsatzdaten

Man erkennt eine diagonale Linie mit den Werten von 1.0, da sich hier die Werte mit sich selbst korrelieren. Die Temperaturwerte (TMK, TXK und TGK) gehen sowohl mit sich selbst einher als auch relativ stark mit den Umsätzen. Ebenso scheinen die tägliche Sonnenscheindauer (SDK) und der mittlere Dampfdruck (VPM) einen positiven Bezug zu Campingplatzumsätzen zu haben. Tendenziell negativ korrelieren dagegen der Tagesmittel des Bedeckungsgrades als auch die relative Luftfeuchtigkeit. Durch den positiven Bezug der Sonnenscheindauer und dem negativen der Bedeckung lässt sich feststellen, dass neben der Temperatur auch das Wetter einen Bedeutung für den Umsatz darstellt. Allerdings liegt dies auch daran, dass Wetter und Temperatur in engem Zusammenhang stehen, was sich auch aus der Korrelationsmatrix entnehmen lässt. Allgemein sehen die Korrelation wie folgt aus:

Wetterdaten	Kürzel	Wert
Tagesmaximum Lufttemperatur 2m	TXK	0.72
Tagesmittel Lufttemperatur 2m	TMK	0.70
tägliche Sonnenscheindauer	SDK	0.68
Tagesminimum Lufttemperatur 2m	TNK	0.67
Tagesmittel des Dampfdrucks	VPM	0.66
Minimum Lufttemperatur 5cm	TGK	0.64
tägliche Niederschlagshöhe	RSK	0.12
Tagesmaximum Windspitze	FX	0.04
Tagesmittel des Luftdrucks	PM	0.00
Tageswert Schneehöhe	SHK_Tag	-0.16
Tagesmittel Windgeschwindigkeit	FM	-0.25
Niederschlagsform	RSKF	-0.34
Tagesmittel des Bedeckungsgrades	NM	-0.47
Tagesmittel der relativen Feuchte	UPM	-0.62

Tabelle 1: Korrelationswerte zwischen Wetterdaten und Umsätzen

Trend- und Saisonbereinigung

Die Trendbereinigung ist hier Sinnvoll, da ein Vergleich zwischen Umsätzen schwierig ist, die weit auseinander liegen, da externe Einflussfaktoren wie die Inflation zu einer grundlegenden Verzerrung der Aussagekraft über die Bewertung dieser führen. Da man wie man in der unteren Grafik in Blau sehen kann, es tendenziell um einen linearen Trend handelt, wurde hier die lineare Trendbereinigung mit der Bibliothek Signal angewendet.

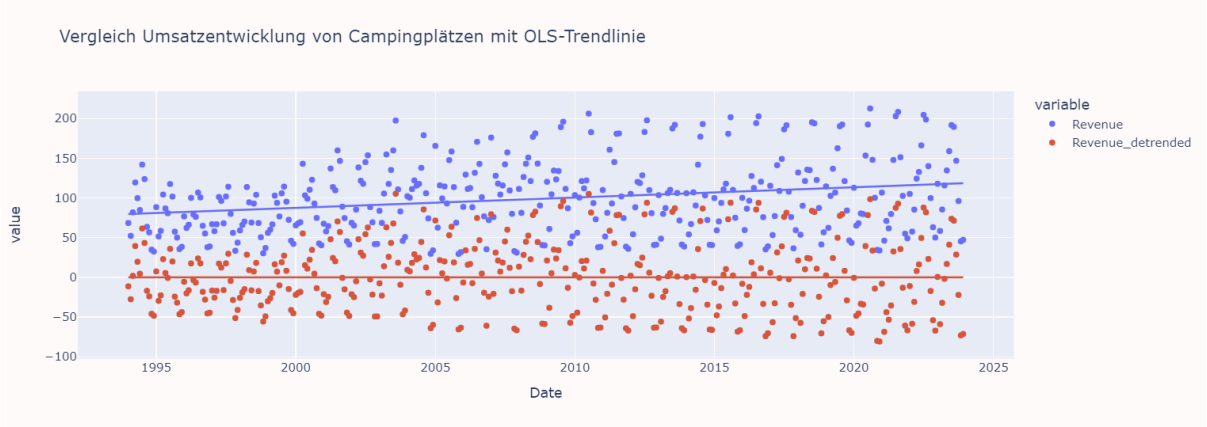

Abbildung 7: Trendvergleich zwischen originalen und Trendbereinigten Umsätzen

Die lineare Trendkomponente wurde, wie man bei den roten Werten sehen kann, entfernt, wodurch das Bestimmtheitsmaß $R^2 = 0$ ist. Somit lässt sich keine lineare Abhängigkeit zwischen Zeit und Umsätzen mehr aufweisen.

Die starke saisonale Komponente hat sich schon in der Aufgabe 2b.) in der Autokorrelation bemerkbar gemacht. Für die Saisonbereinigung wurde hier die STL-Methode (Seasonal-Trend decomposition using Loess) verwendet. Nach der Berechnung kam man zu folgendem Residuum:

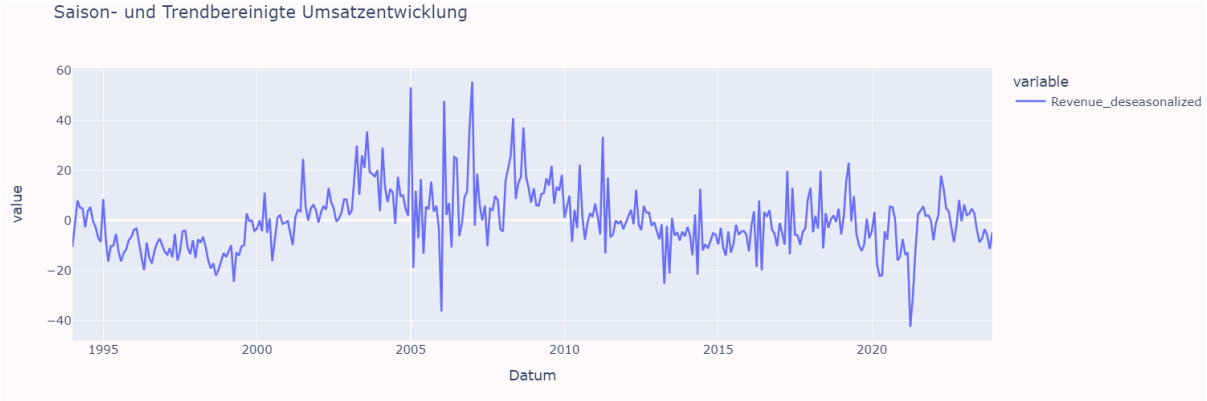

Abbildung 8: Saison- und Trendbereinigte Umsatzentwicklung

Es lässt hier weder einen Trend noch eine wirkliche Saison erkennen. Allerdings fallen trotzdem eine Ausschläge auf. Besonders zwischen 2005 und 2007 sind diese besonders hoch. Dies kann zum Beispiel mit bestimmten Events wie der Weltmeisterschaft 2006 in Deutschland zusammenhängen. Auf der anderen Seite ist ein großes Tief zwischen 2020 und 2021 erkennbar, welches untereanderem mit der Corona-Pandemie zusammenhängt.

Die unbereinigten Umsätze konnte mit einem linearen Regressionmodell zu 65% ($R^2 = 0.65$) erklärt werden, während die Umsätze, welche Saison- und Trendbereinigt wurden nur noch zu ungefähr 19% ($R^2 = 0.1898$) erklärt werden konnten. Somit sind Campingplätze stark Saison- und Trendabhängig.

Datenquelle: DAX

a) Unsere aktuellen Datengrundlagen bestehen aus den Umsatzdaten der einzelnen Gastronomiebetriebe sowie den Temperaturdaten über die jeweiligen Zeiträume.

b) Wir erweitern die Daten um eine weitere Datenquelle, den DAX. Von diesem Datensatz nutzen wir das Datum, den Open- und den Close-Wert des DAX. Die DAX-Daten umfassen den Zeitraum von 1987-12-30 bis 2024-06-03. Um die Vergleichbarkeit mit den Umsatz- und Temperaturdaten zu gewährleisten, werden die DAX-Daten auf Monat und Jahr aggregiert. Die Datenbereinigung erfolgte nach denselben Methoden wie bei den Temperatur- und Umsatzdaten.

   Der Aggregationsprozess für die DAX-Daten umfasste:

   - Konvertierung des Datums in ein DateTime-Objekt.
   - Extraktion von Jahr und Monat und Hinzufügen dieser als neue Spalte.

- Gruppierung der Daten nach Jahr und Monat.
- Aggregation der Open- und Close-Werte, wobei der erste Open-Wert und der letzte Close-Wert jedes Monats verwendet wurden.

c) Durch die Anwendung eines linearen Regressionsmodells haben wir festgestellt, dass die Open- und Close-Werte des DAX einen geringen oder sogar negativen Einfluss auf den Umsatz haben können. Die Temperatur (TMK) hat den größten Einfluss auf den Umsatz, gefolgt vom Open-Wert des DAX und zuletzt vom Close-Wert des DAX.

Die Regressionsgleichungen für verschiedene Gastronomiebetriebe lauten wie folgt:

GastroID	***Regressionsgleichung***
WZ08-55	Y = 101.3003 + 1.9843 * TMK + 0.0030 * Open + -0.0057 * Close
WZ08-551	Y = 102.4165 + 1.7872 * TMK + 0.0030 * Open + -0.0056 * Close
WZ08-552	Y = 92.1624 + 4.5558 * TMK + 0.0050 * Open + -0.0095 * Close
WZ08-553	Y = 32.8658 + 4.4996 * TMK + 0.0060 * Open + -0.0042 * Close
WZ08-559	Y = 115.9893 + 2.5709 * TMK + -0.0001 * Open + -0.0040 * Close
WZ08-56	Y = 151.3831 + 1.1297 * TMK + -0.0017 * Open + -0.0043 * Close
WZ08-561	Y = 155.4274 + 1.4237 * TMK + -0.0017 * Open + -0.0049 * Close
WZ08-562	Y = 95.1109 + 0.3009 * TMK + 0.0006 * Open + -0.0011 * Close
WZ08-563	Y = 229.6171 + 0.8970 * TMK + -0.0059 * Open + -0.0073 * Close
WZ08-561-01	Y = 165.2756 + 1.3537 * TMK + -0.0023 * Open + -0.0052 * Close
WZ08-55-01	Y = 131.5513 + 1.4559 * TMK + 0.0001 * Open + -0.0048 * Close

Tabelle 2: Regressionsgleichung der Gastgewerbe

Durch diese Analyse können wir besser verstehen, wie verschiedene Faktoren, insbesondere die Temperatur und die DAX-Werte, den Umsatz der Gastronomiebetriebe beeinflussen. Dies ermöglicht eine gezieltere Planung und Optimierung der Geschäftsstrategien.

Datenquelle: Campingplätze

Hilfreich für bessere Analysen sind die Standorte der einzelnen Gastgewerbe. Hierfür ließe sich beispielsweise durch den Vergleich mit örtlichen Wetterstationen eine bessere Vorhersage modellieren. Als Prototyp haben wir uns zunächst auf Campingplätze beschränkt. Eine Möglichkeit bietet die kostenfreie OSMnx-API der OpenStreetMap, mit der gezielt Campingplätze in Deutschland gesucht werden können. Jedoch entschieden wir uns, die Place-API von Google Maps zu benutzen, da diese mehr Campingplätze beeinhaltet und besser skalierbar ist, indem mehr Informationen, wie Öffnungszeiten oder Rezensionen abgefragt werden können. Wohlgemerkt können mehr Daten hohe Kosten verursachen, weshalb wir es bei den Geodaten belassen haben.

Da alle Campingplätze innerhalb eines Rechtecks von den äußersten Punkten in Deutschland abgefragt wurden, haben wir ein Geometrieshape von Deutschland genommen und alle ausländischen Campingplätze aus dem Datenset herausgenommen. Durch ein anderes Shape konnten wir zusätzlich noch Informationen darüber gewinnen, in welchem Bundesland, Regierungsbezirk und Kreis die Plätze liegen.

Nutzung der Daten für Stakeholder und Konsumenten

Gastronomiebesitzer

Gastronomiebesitzer können ihre Bestell- und Lagerstrategien optimieren, indem sie den zukünftigen Umsatz basierend auf Wetter- und DAX-Daten prognostizieren. Wenn das Modell einen Anstieg des Umsatzes vorhersagt, bedeutet dies, dass in den kommenden Tagen mehr Kunden im Restaurant erwartet werden. Dies erfordert eine Erhöhung der Lebensmittelbestände und eventuell auch der Personalkapazität, um die gesteigerte Nachfrage zu bedienen. Durch diese proaktive Planung können Engpässe vermieden und die Kundenzufriedenheit gesteigert werden.

Campingplatzbetreiber

Campingplatzbetreiber können die Auslastung ihres Campingplatzes für die kommenden Tage und Wochen besser vorhersagen, indem sie die Umsatzdaten und externen Faktoren wie Wetterbedingungen analysieren. Höhere prognostizierte Umsätze im Regressionsmodell deuten auf eine höhere Auslastung hin. Dies ermöglicht eine effiziente Planung und Ressourcenzuweisung, einschließlich Personalmanagement, Instandhaltung und Bereitstellung von Zusatzdiensten. Zudem können Campingplatzbetreiber Marketingaktionen gezielt einsetzen, um in Zeiten niedrigerer Auslastung zusätzliche Buchungen zu generieren.

Aktionäre

Aktionäre können fundierte Investitionsentscheidungen treffen, indem sie die Prognosen des linearen Regressionsmodells nutzen. Dieses Modell liefert eine Einschätzung des zukünftigen Umsatzes eines Unternehmens, basierend auf historischen Daten und externen Einflussfaktoren wie dem DAX und Wetterbedingungen. Durch die Analyse dieser Vorhersagen können Aktionäre besser beurteilen, ob ein Unternehmen in naher Zukunft finanzielle Zuwächse erwarten kann. Dies bietet ihnen einen strategischen Vorteil bei der Portfolioverwaltung und hilft, Risiken zu minimieren.

Idee und Umsetzung der Architektur

Die Implementierung einer Architektur, die einen prototypischen Datenfluss abbildet, umfasst verschiedene Schlüsselkomponenten und Techniken, die sowohl lokal als auch in der Cloud eingesetzt werden. In unserem Projekt werden Techniken wie ETL-Prozesse (Extraktion, Transformation, Laden) und maschinelles Lernen lokal entwickelt und über das Repository mit den Gruppenmitgliedern geteilt. Techniken wie Batch Processing und Stream Processing werden, wie in der Vorlesung und den Big Data Labs vorgestellt, über die Google Cloud Platform realisiert. Dabei haben alle Gruppenmitglieder Zugriff auf das Cloud Projekt, so dass alle beim Aufbau der Architektur mitwirken können.

Lokale Entwicklung des Data Lake

Um alle Daten in ihrer ursprünglichen Form speichern zu können, werden sie direkt aus ihrem Quellsystem in den Data Lake gespeichert, der in unserem Projekt lokal oder im Repository abgelegt wird. Über das Repository kann jeder auf die Dateien zugreifen, um sie weiter zu bearbeiten. Zusätzlich werden alle geänderten Dateien in einem Unterverzeichnis “Historisierung” mit dem aktuellen Bearbeitungszeitpunkt gespeichert. Dadurch ist gewährleistet, dass die Originaldateien unverändert bleiben und auch innerhalb der Historisierung bei einer erneuten Änderung der Datei ein neuer Zeitstempel erzeugt wird und somit keine Daten verloren gehen.

Lokale Entwicklung der ETL-Prozesse

Ebenfalls lokal/im Repository befinden sich unsere ETL-Prozesse. Das Repository enthält alle Notebooks, die auf den Data Lake referenzieren und die Datenquellen bereinigen und transformieren. Dabei ist die Bereinigung der Daten von syntaktischen (formalen) und semantischen (inhaltlichen) Fehlern in unserer Architektur in der 3. Somit werden unsere Datenquellen manuell auf Fehler untersucht und manuell korrigiert. Schließlich wird über eine Harmonisierung eine Faktentabelle erstellt, die später über Batch Processing mit Spark mit dem Star-Schema mehrere Dimensionstabellen für BigQuery und Visualisierung erzeugt.

Umsetzung des Data Warehouse Konzepts

Um die Daten für analytische Zwecke nutzen zu können, müssen sie angepasst und zusammengeführt werden (Harmonisierung). Die aufbereiteten Daten werden anschließend in einem Data Warehouse gespeichert, um Analysen und Dashboards zu erstellen. Als Data Warehouse-Lösung wurde in diesem Fall die Anwendung BigQuery auf der Google Cloud Platform verwendet. Um unsere Daten jedoch in BigQuery hochladen zu können, müssen wir zunächst einen Umweg gehen. Über das lokale Verzeichnis “Temp_DWH” werden die aktuell erstellten Faktentabellen in einem Bucket im Google Cloud Storage abgelegt. Dabei werden neben der Faktentabelle bereits vorgefertigte Dimensionstabellen, die mittels Batch Processing durch Spark erstellt wurden, in unsere Datenbank in BigQuery geladen.

Erstellung des Dashboards

Über das Data Warehouse werden die Daten direkt in ein Dashboard geladen, wo sie visualisiert und für den Endnutzer aufbereitet werden. Für die Erstellung des Dashboards wurde in diesem Fall Google Looker Studio verwendet. Der Zugriff auf die im Warehouse liegenden Daten wurde mit Hilfe des Google Connectors bereitgestellt. Dieser kann auf verschiedene Daten innerhalb der Analytics Architektur zugreifen und ermöglicht eine stets aktuelle Datenbasis für das Dashboard. Es steht eine Sicht auf die gesammelten Wetter- und Umsatzdaten zur Verfügung.

Batch Processing

Folgende Schritte wurden auf die Faktentabelle durchgeführt, um auf die gewünschte Ergebnisse zu kommen:

1. Erstellung des Dataproc-Clusters:
    - Ein Dataproc-Cluster wird erstellt, indem man die Google Cloud Console oder gcloud-Befehle verwendet. Hierbei werden die Anzahl der Nodes, die Maschinenkonfiguration und die Spark-Komponenten spezifiziert.
2. Aufteilen der Faktentabelle in Dimensionstabellen:
    - Die Faktentabelle wird von Google Cloud Storage über einem Bucket in den Cluster geladen.
    - Mit Spark SQL werden SQL-Abfragen geschrieben, die die Faktentabelle in verschiedene Dimensionstabellen aufteilen.
    - Diese Dimensionstabellen werden aus der Faktentabelle extrahiert, indem spezifische Spalten ausgewählt und entsprechende SQL-Abfragen formuliert werden. In diesem Fall werden Dimensionstabellen für unterschiedliche Wetterthemen erstellt, wie z.B. Temperatur, Niederschläge, Sonnenschein.
3. Zwischenspeichern der Dimensionstabellen:
    - Die erstellten Dimensionstabellen werden in das Parquet-Format konvertiert und im Google Cloud Storage zwischengespeichert.
    - Der Spark DataFrameWriter wird verwendet, um die Dimensionstabellen als Parquet-Dateien in den Bucket zu speichern.
4. Durchführung weiterer Analysen:

    - Die gespeicherten Dimensionstabellen können nun für weitere Analysen genutzt werden. Spark SQL kann verwendet werden, um komplexe Abfragen und Analysen auf diesen Tabellen durchzuführen.
    - Mögliche Analysen können das Berechnen von Aggregaten, das Erstellen von Reports oder das Trainieren von Machine Learning Modellen auf den Dimensionstabellen umfassen.
    - Diese Analysen werden dann in die BigQuery reingeworfen.

Stream Processing

Die Implementierung einer Pub/Sub-Verwendung mithilfe des Regressionsmodells soll Stakeholdern ermöglichen, die Daten effizient zu nutzen und fundierte Entscheidungen zu treffen. Hier ist der Prozess im Detail beschrieben:

Prozessbeschreibung

1. Tägliche Datenbeschaffung:
    - Temperaturdaten: Die aktuellen Temperaturdaten werden täglich von einer SenseBox in Luxemburg Mersch über die OpenSenseMap bezogen, dies kann erweitert werden auf die lokalen Temperaturen der Betroffenen.
    - DAX-Daten: Da der Abruf von DAX Open- und Close-Daten kostenpflichtig ist, generiert eine Cloud Function stattdessen zufällige Werte für den DAX Open- und Close-Kurs, um den natürlichen Verlauf des DAX zu simulieren.
    - Wertebereich der Simulation zwischen 15000 - 20000.
2. Modellausführung:
    - Die gesammelten Temperaturdaten und die simulierten DAX-Daten werden in eine Cloud Function eingespeist.
    - Das Regressionsmodell wird in der Cloud Function geladen und auf die neuen Daten angewendet.
    - Der Output des Modells besteht aus den vorhergesagten Umsätzen für verschiedene Gastronomiebetriebe.
3. Pub/Sub Benachrichtigungssystem:
    - Das Ergebnis der Modellvorhersage wird an einen Pub/Sub-Dienst weitergeleitet.
    - Stakeholder (Aktionäre, Gastronomiebesitzer, Campingplatzbetreiber) haben individuelle Schwellenwerte für Benachrichtigungen definiert.
        - Aktionäre: Geben eine Umsatzschwelle an, bei deren Überschreitung sie benachrichtigt werden, um Entscheidungen über ihr Aktiendepot zu treffen.
        - Gastronomiebesitzer: Geben eine Umsatzschwelle an, bei deren Überschreitung sie benachrichtigt werden, um mehr Vorräte zu kaufen.
        - Campingplatzbetreiber: Können durch die Umsatzprognosen der Gastronomiebetriebe besser planen, ob der Campingplatz in den folgenden Tagen ausgelastet sein wird.
4. Benachrichtigungen:

- Basierend auf den definierten Schwellenwerten sendet der Pub/Sub-Dienst Benachrichtigungen an die entsprechenden Stakeholder.
- Diese Benachrichtigungen können per E-Mail, SMS oder Push-Benachrichtigung erfolgen, je nach den Präferenzen der Nutzer.

5. Cloud Function:
   - Link zur Cloud Function: https://us-central1-dscb420-ad.cloudfunctions.net/test_model
6. Probleme:
   - Da es Probleme mit der Pub/Sub Implementierung gibt wurde nur die Cloud Function implementiert und kann um das Pub/Sub erweitert werden, wenn Google Cloud es zulässt.

Datenfluss

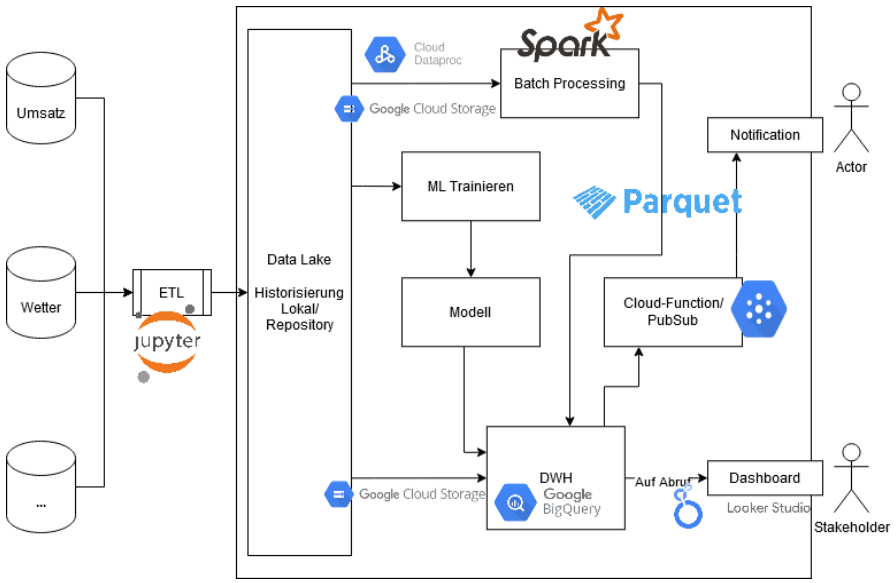

Abbildung 10: Datenfluss und angewandte Anwendungen

Als Datengrundlage dienen Umsatzdaten des Gastgewerbes von Januar 1994 bis Dezember 2023 auf Monatsebene und Wetterdaten von Wetterstationen, wobei der Betrachtungszeitraum je nach Messstation variiert. Betrachtet werden hauptsächlich Daten, die dem Zeitraum der Umsatzdaten entsprechen. Diese Datenquellen werden in einem Data Lake gespeichert und anschließend einer ETL-Verarbeitung unterzogen. Um eine Big-Data-Historisierung aufzubauen, wird jede Änderung nicht in der Originaldatei überschrieben, sondern als neue Datei im Data Lake gespeichert. Dabei werden die zuletzt geänderten Daten in das Data Warehouse geladen. Dadurch wird die Aktualität der Daten im Warehouse gewährleistet. Die Daten im Data Lake werden unter anderem für den ersten Trainingslauf des Machine-Learning-Modells verwendet. Ebenfalls im Data Lake befinden sich die Faktentabellen. Mit der Faktentabellen besitzen wir ein Galaxy-Schema und können, wenn die Tabelle im Bucket gespeichert ist, mit einem Dataproc Cluster über Spark die Tabelle in mehrere Dimensionstabellen aufteilen. Die Ergebnisse können nach dem Batch Processing im Data Format Parquet gespeichert werden, welches dann in das Data Warehouse “BigQuery” eingefügt werden kann. Bei der Nutzung von Pub/Sub bauen wir ein

Benachrichtigungssystem auf für Entwickler und Nutzer, welches auf bestimmte Ereignisse reagieren soll.
Zum Schluss wird die BigQuery-Datenbank mit Looker Studio gelesen, um die Berichte für unsere Endbenutzer zu visualisieren.

Anhang

Dokumentation:

- “Dabi 2-Architecture.drawio” Diagramm vom Datenfluss und wie diese Strukturiert ist
- “DABI2_Dashboard.pdf” Beispiel Dashboard was sich unsere Kunden vorstellen können

Analytic_Results:

- Dieser Ordner beinhaltet die Notebooks für den Aufgabenteil 2 und Machine Learning
- Aufgabe2.ipynb
- “LineareRegression_Dax_GastroUmsatz.ipynb”

BatchProcessing_GCP:

- Der Ordner beinhaltet die im Dataproc verwendeten PySpark Skripte
- “Dim_Weather.ipynb” lesen von Faktentabellen mit Spark und Umwandlung in Dimenionstabellen
- “Analytics_Rain.ipynb” vertiefte Analyse/Aufteilung mit Spark SQL; Untersuchung nach Niederschlagswerten.
- “Analytics_Sunshine.ipynb” Untersuchung nach Sonnenschein und Bedeckungsgrad
- “Analytics_Temperature.ipynb” Untersuchung von Temperaturdaten
- “Analytics_Wind.ipynb” Untersuchung von Winddaten

ETL:

- Der Ordner beinhaltet die für das Data Warehouse einzuführende ETL-Prozesse zur Filterung und Harmonisierung von Faktentabellen.