

Wintersemester 2022/23

Inhalt

1	Anwendungsfall und Datensatz	2
2	Organisation und einzureichende Dokumente	4
3	Aufgabenstellung	4
3.1	Überblick	4
3.2	Vorab gesammelte, konkrete Fragen der Mitarbeiter	5
3.3	Aufgaben zur Erstellung eines prototypischen „Business Dashboards“	5
4	Anforderungen an die Einreichung	6
4.1	Projektbericht	6
4.2	Jupyter-Notebooks	6
4.3	PowerBI-Dashboard Prototyp	7
5	Anforderungen an die Präsentation	7

1 Anwendungsfall und Datensatz

In dieser Übung sollen Ihre Kompetenzen in der Datenanalyse ausgeweitet werden. Auf Basis von Visualisierungen und statistischen Analysen eines betriebswirtschaftlichen Anwendungsfalls sollen wieder Schlüsse aus den Daten gezogen und Hypothesen über den Betrachtungsgegenstand aufgestellt werden. Die Perspektive der rein „explorativen Analyse“, die bspw. in einem Jupyter Notebook umgesetzt werden kann, wird dieses Mal erweitert um die operative Sichtweise auf „Business Analytics“, d.h. die organisatorische Einbettung und Verwendung Ihrer Analysen.

Ihre Gruppe stellt ein Team einer Data Science-Beratungsgesellschaft dar und Sie beraten einen **Online Lebensmittel-Lieferanten** aus Kalifornien.

Die vorzunehmenden Analysen enthalten grob folgende Aspekte: (1) **Vorverarbeitung** und Bereinigung der Daten, (2) Identifikation von Zusammenhängen sowie deren Prüfung unter Verwendung **statistischer Methoden** in Kombination mit (3) der **Visualisierung** Ihrer Erkenntnisse. Die Visualisierung ist in diesem Übungsblatt allerdings zweigeteilt zu verstehen: Zum einen wird sie zur „explorativen Analyse“ inklusive der Unterstützung Ihrer statistischen Analysen (in einem Jupyter Notebook) verstanden. Zum anderen soll eine prototypische Umsetzung eines visuellen „Business Dashboards“ erreicht werden. Das Tool soll sich für die operative Einbettung in die Business Intelligence-Architektur des Unternehmens eignen. In diesem Fall soll PowerBI verwendet werden.

Die Nutzer des Lebensmittel-Lieferdienstes können online bestellen und bekommen die gewünschten Produkte nach Hause geliefert. Der Datensatz enthält als **Basisstruktur die Nutzerbestellungen**, die eindeutig über das Attribut „order_id“ identifizierbar sind. Eine Bestellung kann aus einer oder mehreren Bestellzeilen bestehen, d.h. es ist über die „product_id“ sichtbar, welche Produkte der Nutzer (identifizierbar über „user_id“) bestellt hat. Eine Bestellung erstreckt sich somit ggf. über mehrere Zeilen, bspw. order_id 4 enthielt 3 Produkte (40078, 46175, 9065) und wurde von user_id 193635 getätigt.

Der Datensatz enthält einige Details zu den einzelnen Bestellungen:

- Die Reihenfolge, in der die Produkte in den Warenkorb gelegt wurden ist in „add_to_cart_order“ hinterlegt.
- Ob das Produkt schon einmal von einem Nutzer bestellt wurde ist unter „reordered“ erkennbar. *Achtung: wir sehen nur einen Ausschnitt der Daten, d.h. es kann Nutzer geben, deren erste Bestellung in unserem Ausschnitt dennoch „reordered“ Produkte enthält.*
- In der Spalte „days_since_prior_order“ wird angezeigt, wie viele Tage die vorherige Bestellung des Nutzers her ist. Der Wert kann „n/a“ sein, wenn es sich um die erste Bestellung eines Nutzers (im zugrundeliegenden Zeitraum) handelt. *Achtung: Hier könnte es aufgrund des Datenausschnitts ggf. Inkonsistenzen geben.*
- Die Zeitinformationen sind relativ begrenzt, da der Kunde Ihnen nicht die vollständigen Informationen zukommen lassen wollte. Sie haben aber den Wochentag („order_dow“) sowie die Stunde („order_hour_of_day“) der Bestellungen erhalten.
- Die bestellten Produkte sind nicht nur über die „product_id“ erkennbar, sondern auch über ihren „product_name“. Zusätzlich enthält der Datensatz Informationen zu dem „Gang“, in dem die Produkte in einem Supermarkt stehen („aisle_id“ bzw. „aisle“) sowie eine Zuordnung der Produkte zu Kategorien („department_id“ bzw. „department“).
- Auch die räumlichen Informationen über die Bestellungen sind relativ begrenzt: Sie erhalten lediglich Auskunft darüber, in welchem County der Nutzer bestellt hat („county“). *Achtung: Die County-Zuordnung entspricht nicht den realen Daten, d.h. es lohnt sich nicht Wettereinflüsse, Feiertage (ist ohnehin nicht möglich) oder weitere Faktoren mit County-Bezug aus externen Datenquellen abzurufen und zu analysieren. Dennoch lohnt sich eine Betrachtung der Counties im Hinblick auf die vorzunehmenden Analysen und Visualisierungen.*
- In der Spalte „tip“ ist erkennbar, ob der Kunde dem Lieferanten (Fahrer) Trinkgeld gegeben hat.

Ihr Consulting-Projekt wurde von **zwei Anspruchsgruppen** des Lieferdienstes initiiert: Die Personalabteilung und die Disposition zeigen unterschiedliche Wünsche, um aus den Daten spezifische Erkenntnisse zu erlangen. Diese sind in den folgenden zwei Absätzen grob dargestellt. Die konkrete Aufgabenstellung finden Sie in Kapitel 3.

Das Unternehmen ist dabei, die Arbeitsbedingungen seiner Fahrer neu zu gestalten. Insbesondere haben viele Fahrer den Eindruck, dass die Chance auf Trinkgeld bei unterschiedlichen Aufträgen sehr unterschiedlich hoch sei, und dies als ungerecht empfunden worden. Die **Personalabteilung** will dem auf den Grund gehen und zunächst anhand von objektiven Daten prüfen, ob sich tatsächlich systematische Muster bei der Trinkgeldvergabe nachweisen lassen. Falls dies der Fall ist, sollen die erkannten Muster entweder bei der Entlohnung der Fahrer durch das Unternehmen, oder bei der Zuordnung der Fahrer zu den Aufträgen berücksichtigt werden – letzteres z. B. indem alle Fahrer gleich oft vielversprechende und weniger aussichtsreiche Fahrten zugewiesen bekommen. Ihr Auftrag bei der Datenanalyse besteht also darin, nachvollziehbare Muster zu suchen und dabei zu ermitteln, wovon die Wahrscheinlichkeit abhängt, ob bei einer Bestellung Trinkgeld gegeben wird oder nicht.

Die **Disposition** erwartet von Ihrem Team Erkenntnisse über die regionalen Unterschiede im Kaufverhalten der Kunden. Im Fokus sollten dabei die räumlichen Unterschiede des Absatzes bspw. in Bezug auf die Produktkategorien, die „Gänge“ oder umsatzstarke Produkte stehen. Die Disposition erhofft sich dadurch eine bessere Versorgung der einzelnen Counties. *Folgende Info zum Geschäftsmodell ist für die Aufbereitung entscheidend: Der Online-Lieferdienst betreibt pro County ein Distributionszentrum, von dem aus die Bestellungen ausgeliefert werden. Die Disponenten, die für die Belieferung der Zentren verantwortlich sind, wünschen sich Informationen darüber, welche Produkte (bzw. Kategorien oder „Gänge“) in welchem County stärker, schwächer oder eher durchschnittlich nachgefragt werden. Auf dieser Basis soll die Belieferung der Distributionszentren optimiert werden. Achtung: Ein Prognosemodell ist nicht zu erstellen. Es reicht zunächst eine statische Auswertung des in den Daten sichtbaren Zustands, d.h. „Descriptive Analytics“ wie in Kapitel 3.2 beschrieben.*

Vor Auftragserteilung wurden an Sie einige spezielle Fragen von Mitarbeitern der Personalabteilung und der Disposition zusammengetragen, die in verschiedenen Zusammenhängen kürzlich aufgetreten sind (siehe Abschnitt 3.2). Diese sollen von Ihnen im Rahmen der oben formulierten, „breiteren“ Zielen beantwortet werden.

Wenden Sie für dieses Übungsblatt wieder Ihre über das bisherige Studium hinweg und in der Veranstaltung DSCB310 erlangten Kompetenzen an, um für beide Anspruchsgruppen die erhofften Erkenntnisse aus den Daten möglichst effektiv zu kommunizieren bzw. eine operative Einbettung Ihrer Analysen in den Geschäftsprozess mittels des „Business Dashboards“ vorzubereiten.

Die Erkenntnisse der Projektarbeit sind in einem Projektbericht festzuhalten und in einem (simulierten) „Pitch“ beim Kunden zu präsentieren. Weitere Hinweise zu der Erwartungshaltung der Kunden erhalten Sie in den folgenden Kapiteln.

2 Organisation und einzureichende Dokumente

Die Übung ist durch Projektgruppen bestehend aus 3 - 4 Studierenden umzusetzen; im Regelfall sind dies die gleichen Gruppen wie bei dem ersten Projekt. Die Gruppen müssen in ILIAS registriert sein (vgl. Hinweise in den Vorlesungen).

- **Start Übung 2:** Montag, 5.12.2022
- **Abgabe Projektbericht (Detailhinweise s. u.): Montag, 16.01.2023, 8 Uhr**
Über ILIAS muss der Projektbericht von der Gruppe als PDF-Datei bis zur Deadline eingereicht werden (Abschnitt „Übungsaufgaben“ → „Baustein Übung“ – dort können Sie die Datei mit dem Projektbericht je Gruppe hochladen)
- **Präsentation (Detailhinweise s.u.): Freitag 20.01.2023** (pro Gruppe 15 Minuten Präsentation, ca. 10 Minuten Diskussion). Achtung: die Präsentationen sind am **20.01.2023 um 8 Uhr** in ILIAS als PDF einzureichen. Es handelt sich um einen Pflichttermin, bei dem bei allen Präsentationen sämtliche Teilnehmer anwesend sein sollen.

3 Aufgabenstellung

3.1 Überblick

Die Aufgabenstellung ist wie in Kapitel 1 dargestellt zweigeteilt:

- (1) Eine Analyse des Datensatzes sowohl in Hinblick auf die Personal- als auch auf die Dispositionsfragestellung wird erwartet. Die Stakeholder sollen in Ihrem **Projektbericht**, der durch ein **Jupyter-Notebook** gestützt werden soll, Antworten auf die unten dargestellten konkreten Fragen (Kapitel 3.2) sowie weitere Informationen, die für ihre oben formulierten allgemeinen Ziele relevant sind, erhalten. Der Projektbericht soll Ihre Kernerkenntnisse aus den Analysen enthalten und jeweils auf die entsprechenden Passagen im Jupyter-Notebook referenzieren. Alle Analysen sollen im Text leicht verständlich kommuniziert werden, aber auch im Detail im Code nachvollzogen werden können. Bauen Sie den Projektbericht und Ihr Notebook entsprechend systematisch auf. Sie dürfen auch zwei oder mehr Notebooks einreichen, wenn es bspw. Sinn ergibt, die finalen Ergebnisse von einzelnen Detailanalysen, die nicht vielversprechend waren, zu trennen.
- (2) Zusätzlich sollen Sie ein prototypisches „**Business Dashboard für die Disposition**“ mit **PowerBI** umsetzen. Hauptanwender dieses Dashboard-Prototyps soll die Disposition sein, Sie dürfen aber optional zusätzlich für die Personalabteilung auch ein Dashboard kreieren. Weitere Hinweise hierzu finden Sie in Kapitel 3.3.

Wichtig für alle Schritte: Relevant sind die in den Vorlesungen/Übungen vorgestellten Konzepte (statistische Methoden, Vorverarbeitungskonzepte, Visualisierungen, Aspekte zu PowerBI) **inklusive der Vorlesungen 1 Woche vor dem Abgabetermin** sowie alle Inhalte aus den vorherigen Semestern.

3.2 Vorab gesammelte, konkrete Fragen der Mitarbeiter

Die Mitarbeiter der **Personalabteilung** haben folgende konkrete Aufgaben bzw. Fragen für Ihren Beratungsauftrag formuliert, die im Rahmen Ihrer Gesamtanalyse explizit zu beantworten sind:

Frage P.1: Prüfen Sie, ob ein Zusammenhang zwischen dem Trinkgeldgeben und dem Vorhandensein eines der folgenden Produkte in einer Bestellung existiert: 24852, 24964, 2120

Frage P.2: Lassen sich *regionale* Unterschiede im Trinkgeldverhalten erkennen?

Frage P.3: Welche Attribute einer Bestellung wirken sich auf das Trinkgeldverhalten aus?

Frage P.4: Spielt die Vergangenheit eines Users eine Rolle in Hinblick auf die Trinkgeldwahrscheinlichkeit, oder kommt es nur auf Inhalt und Parameter der aktuellen Bestellung an?

Die Mitarbeiter der **Disposition** haben ebenfalls einige konkrete Fragen formuliert:

Frage D.1: Prüfen Sie, ob es beim Produkt 13176 auffällige Muster beim Wiederbestellverhalten gibt.

Frage D.2: Spielen die Artikel aus dem Department „produce“ in allen Counties eine gleich große Rolle?

Frage D.3: Welche Counties sind sich ähnlich in Hinblick auf die jeweiligen „Top 10“-Produkte?

Empfehlung für den Projektbericht: Beantworten Sie die Fragen und belegen Sie Ihre Aussagen mit geeigneten statistischen Analysen. Nehmen Sie in den Projektbericht die Kernerkenntnisse dieser Analysen auf, begründen Sie Ihre Auswahl der angewendeten Methoden und referenzieren Sie auf die entsprechenden Bereiche im Notebook. Ergänzen Sie die statistischen Analysen, sofern möglich und Ihrer Meinung nach geeignet, durch passende Visualisierungen. Diese sollten sich möglichst gut zur Kommunikation Ihrer Erkenntnisse eignen, so dass die in Analytics „ungeschulten Stakeholder“ die Ergebnisse leicht nachvollziehen können.

3.3 Aufgaben zur Erstellung eines prototypischen „Business Dashboards“

Die **Disposition** erhofft sich – wie in Kapitel 1 dargestellt – Erkenntnisse über das Kaufverhalten der Kunden insbesondere im Hinblick auf räumliche Unterschiede des Absatzes.

Nutzen Sie ein Jupyter-Notebook, um **räumliche Unterschiede** im Hinblick auf Produktkategorien („departments“), Gänge („aisles“) oder auch umsatzstarke Produkte zu ermitteln. Hierzu sind ggf. Vorverarbeitungsschritte bzw. Aggregationen nötig. Auch Visualisierungen im Notebook können helfen, Ihre Erkenntnisse zu untermauern und eine nachvollziehbare „Data Story“ aufzubauen. Ziel ist wie oben erläutert eine bessere Information der Disponenten. Ihre Analysen des aktuellen Stands in den Daten sollen der Disposition erlauben, die Distributionszentren der Counties im Hinblick auf „besondere“ Produkte/Gänge/Kategorien besser zu versorgen. Es handelt sich um eine statische Analyse, d.h. Sie müssen mögliche Veränderungen in den Daten und zeitliche Aspekte nicht berücksichtigen. Insbesondere ist auch kein Prognosemodell gefordert. Sie dürfen natürlich weitere Methoden (auch Prognosen) nutzen, um Erkenntnisse über die Daten zu gewinnen; Diese sollten aber nicht zu prominent in den Fokus der Analysen gebracht werden.

Erarbeiten Sie aus Ihren Erkenntnissen **Handlungsempfehlungen** für die Disposition, anhand derer eine bessere Versorgung der einzelnen Counties erreicht werden kann. Sie dürfen wo nötig Annahmen treffen, wenn bspw. für Sie relevante Informationen nicht angegeben sind. Stellen Sie diese bitte transparent dar.

Die Python-basierte explorative Analyse soll die Grundlage eines „**Business Dashboards**“ für die Disposition sein. In dem Dashboard sollen auf leicht nachvollziehbare Weise Ihre Analyseergebnisse sichtbar werden, so dass Dispositionsmitarbeiter entsprechend Ihrer erarbeiteten

Handlungsempfehlungen entscheiden könnten. Der Prototyp ist somit eine Vorstufe zur Einbettung der „Descriptive Analytics“ in den operativen Geschäftsbetrieb. *Wichtig: Wir gehen von einem stationären Zustand der Daten aus (s.o.), d.h. Sie müssen sich keine Gedanken zu Änderungen in den Daten machen, die sich in folgenden Perioden ergeben würden (dies dürfen Sie natürlich, wird aber nicht gefordert). Visualisieren Sie den aktuellen Stand der Daten.*

Setzen Sie das Business Dashboard **prototypisch mit PowerBI** um. Die Hauptanwender kommen wie gesagt aus der Disposition. Sie dürfen zusätzlich auch ein Dashboard für die Personalabteilung kreieren, dies ist jedoch nicht gefordert. *Wichtig: Da in der Vorlesung bis zu dem Zeitpunkt der Abgabe nur begrenzte PowerBI-Kenntnisse vorliegen werden, dürfen Sie für noch nicht behandelte Visualisierungen auch Python-Visualisierungen vorschlagen, um Ihre Dashboard-Ideen umzusetzen. Erstellen Sie – falls Sie von der Möglichkeit Gebrauch machen – ein „Mockup“, d.h. bspw. einen Screenshot Ihres PowerBI-Dashboards ergänzt um Screenshots der Python-Visualisierungen. Die Verwendung von PowerBI ist allerdings verpflichtend und die behandelten Möglichkeiten in PowerBI sollten „ausgereizt“ werden.*

4 Anforderungen an die Einreichung

Einzureichende Dokumente:

- **Projektbericht (PDF)**
- **Jupyter Notebook(s)**
- **PowerBI-Dashboard (pbix-Datei)**

4.1 Projektbericht

- Erstellen Sie eine Titelseite Bezeichnung „DSCB310 – Abgabe Übungsblatt 2 (WS2022/23)“ sowie die Matrikelnummern der Gruppenteilnehmer (keine Namen!) und die Gruppennummer
- **8-11 Seiten (A4) inkl. Visualisierungen**, die Titelseite zählt nicht dazu, verlagern Sie Details in den Anhang bzw. in das Jupyter Notebook, beschränken Sie sich auf die Kernaussagen
 - o Schriftart Arial, Schriftgrad 11, Zeilenabstand „mehrfach 1,2“
 - o Seitenränder 2,5cm
 - o Abstand zwischen Absätzen 3 Punkte (Einstellung unter „Absatz“ in Word)
 - o Abstand nach Überschriften 6 Punkte (es sollten Kapitel / Überschriften zur Strukturierung verwendet werden)
- Erstellen Sie ein Inhalts- und Abbildungs- sowie Tabellenverzeichnis und ggf. ein Abkürzungsverzeichnis.
- Achten Sie auf Systematik im Argumentationsaufbau, präzise Formulierungen, ein durchgängiges Abbildungsdesign mit hoher Abbildungsqualität, sorgfältige Prüfungen der formellen Aspekte, eine zielgruppengerechte Aufbereitung der Inhalte, etc.
- Reichen Sie den Projektbericht als PDF-Datei ein

4.2 Jupyter-Notebooks

Erstellen Sie zusätzlich ein bzw. mehrere sauber strukturierte Jupyter-Notebooks. Achten Sie auf die Hinweise in Kapitel 3.

Wichtig: Sie dürfen die Jupyter-Visualisierungen oder -Tabellen auch in einem anderen Tool wie Excel/Visio/PowerPoint oder ähnlichem nachbearbeiten, um Aspekte hervorzuheben (sofern die Methode der Hervorhebung in der Veranstaltung nicht behandelt wurde). Bitte kennzeichnen Sie entsprechende Abbildungen mit einer Fußnote an der Abbildungsunterschrift und dem Hinweis in der Fußnote („Erstellt in Jupyter-Notebook und nachbearbeitet mit <Tool-Name>“).

4.3 PowerBI-Dashboard Prototyp

Erstellen Sie einen Prototyp der die in Kapitel 3.3 beschriebenen Erwartungen des Kunden erfüllt. Begründen Sie die Auswahl der abgebildeten Elemente im Dashboard und erläutern Sie, wie das Dashboard der Disposition helfen kann, Ihre Handlungsempfehlungen umzusetzen. Die Handlungsempfehlungen müssen natürlich auch genannt und kurz erläutert werden.

5 Anforderungen an die Präsentation

Alle Gruppen müssen die Ergebnisse dieser Übung präsentieren.

Formelle Anforderungen an die Präsentation

- Zeitlicher Rahmen der Präsentation: 15 Minuten Präsentation, ca. 10 Minuten Diskussion
 - o Genaues Timing ist ein Bewertungskriterium!
- Sie dürfen selbst entscheiden wer präsentiert (auch wie viele Studierende), jedoch muss im Laufe des Semesters jeder mindestens einmal präsentiert haben.

Wichtige Hinweise zu den Bewertungskriterien: Achten Sie insbesondere auf

- eine saubere Struktur und Systematik in der Präsentation (unterstützt bspw. durch eine Agenda),
- übersichtliche und leicht nachvollziehbare Folien (es gibt keine Formatvorlage, hier können Sie selbst kreativ sein),
- eine zielgruppengerechte Aufbereitung der Inhalte,
- eine Darstellungsform, die auf eine Präsentation zugeschnitten ist (nicht 1:1 Kopie aus dem Projektbericht!) und klares Foliendesign,
- lesbare Abbildungen, Tabellen, etc.,
- genaues Timing (üben Sie die Präsentation inkl. Übergabe an Kommilitonen etc.),
- einen flüssigen Vortragsstil ohne ablesen (üben!), sowie
- Nutzung „interaktiver“ Inhalte – sofern möglich und angebracht. Beispiele:
 - o Vereinzelte Ausführung und Visualisierung in Jupyter-Notebook
 - o Animationen können bei der systematischen Herleitung komplexer Inhalte helfen (aber nicht übertreiben!)
 - o Live-Vorstellung des „Business Dashboards“