Projektbericht
DSCB310
Abgabe Übungsblatt 2
WS2021/22
Matrikelnummer der Gruppe 3:
67610
77573
67641
77587
77399

Fakultät für Informatik und Wirtschaftsinformatik
Data Science

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Dieser Projektbericht dient der Dokumentation der Analyse ,,Online Lebensmittel-Lieferanten aus Kalifornien“ an der Hochschule Karlsruhe im Bachelorstudiengang Data Science im Wintersemester 2021/2022. Der fortschreitende Trend im Onlineshoppingsegment führt zur Etablierung des Online-Lebensmittelhandel. Die Tendenzen zeigen klar auf die Nachhaltigkeit.

Die Zielvorstellung dieses Projektes ist eine zielgruppengerechte Analyse des Datensatzes. Die daraus resultierenden Ergebnisse ermöglichen es, der Marketing-Abteilung beziehungsweise der Dispositionen des Lieferdienstes, tiefere Einsichten in das Geschäftsfeld zu gewähren, sodass Rückschlüsse für die betrieblichen Handlungsbereiche gezogen werden können. Das Ziel ist es demnach, auf Grundlage der Erkenntnisse aus diversen Analysen, eine Unterstützung im Durchdringen der betrieblichen Handlungsbereiche zu sein und als Stütze für künftige Handlungen des Unternehmens zu dienen.

Hinweise: Die Nummerierungen im Bericht stimmen nicht mit dem des Notebooks überein.

Die referenzierenden Fußnoten weisen auf die Nummerierungen im Notebook hin.

Die Handlungsempfehlungen der *Marketingabteilung* werden fortlaufend durch Sternchen* und die Empfehlungen der °Disposition° mittels Kreis° am Empfehlungsanfang und Ende gekennzeichnet.

2 Übersicht der Daten und Datenvorverarbeitung

Die ‚shoppingCarts.parquet‘ Datei gibt Aufschluss über alle Bestellungen über einen gewissen Zeitraum. Jede Zeile entspricht dabei einem Produkt aus einer Bestellung.

Es wurden circa 6.1 Millionen Produkte in 600.000 Bestellungen von 37.000 Kunden bestellt, die jeweils durch 14 Spalten beschrieben werden.[1] Die Bestellungen wurden in 58 counties getätigt, welche alle zu Kalifornien, des westlichen Bundesstaates der USA, gehören. Trotz dem fehlenden Datum lässt sich die Zeitspanne, durch die ‚unique‘ Wertausgabe von 31, auf einen Monat vermuten.

Durch die Überprüfung der NaN-Werte[2], Duplikate[3], Datentypen der einzelnen Spalten[4] und Optimierung des Speichers[5] wurden die bereits konsistenten Daten verfeinert. Es wurden keine gravierenden Datenqualitätsfehler erkannt. Kleine Datenqualitätsprobleme wie die Abteilung ‚missing‘ wurden nicht berücksichtigt.

3 Datenanalysen

3.1 Produktübersicht

Generell wurde im gegebenen Zeitraum genau 45.616 verschiedenen Produkte bestellt. Diese lassen sich in 21 Klassen kategorisieren und sind in 134 Gängen aufzufinden[6]. Die zehn Bestseller[7], das bedeutet die Handelsartikel, deren Absatzvolumen überdurchschnittlich hoch

[1] Kapitel 1.2.1
[2] Kapitel 1.2.2
[3] Kapitel 1.2.3
[4] Kapitel 1.3.1
[5] Kapitel 1.3.2
[6] Kapitel 2.1.1
[7] Kapitel 2.1.2

ist, des Online Lebensmittel-Lieferanten sind ausschließlich in der Produktkategorie ‚produce' aufzufinden und sind im Folgenden tabellarisch aufgeführt.

	product_name	total orders	department	aisle
1	Banana	90369	produce	fresh fruits
2	Bag of Organic Bananas	70552	produce	fresh fruits
3	Organic Strawberries	50847	produce	fresh fruits
4	Organic Baby Spinach	45860	produce	packaged vegetables fruits
5	Organic Hass Avocado	40208	produce	fresh fruits
6	Organic Avocado	33119	produce	fresh fruits
7	Large Lemon	28725	produce	fresh fruits
8	Strawberries	27719	produce	fresh fruits
9	Limes	27022	produce	fresh fruits
10	Organic Raspberries	26106	produce	packaged vegetables fruits

Tabelle 1: Die zehn Bestsellerprodukte

Es handelt sich ausschließlich um Obst oder Gemüse. Sie befinden sich infolgedessen nur in dem Gang der frischen Früchte oder im Gang des verpackten Gemüses und Obst. Das Lieblingsprodukt mit 90396 absoluten Bestellungen sind die ‚einfachen' Bananen. Hervorstechend ist die erhöhte Nachfrage der Kunden allein nach Bio-Produkten in den Bestsellern. Sechs der zehn sind naturbelassen.

Die zehn beliebtesten Kategorien sind im folgenden Balkendiagramm grafisch dargestellt[8]:

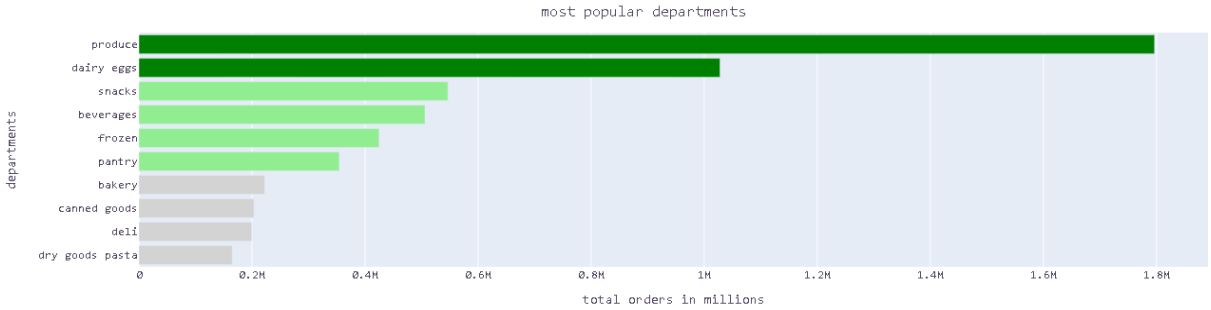


Abbildung 1: Die zehn Bestsellerkategorien

Je kräftiger die grüne Farbe, desto größer die absoluten Bestellungen. Durch die Analyse der beliebtesten zehn Produkte war zu spekulieren, dass die Kategorie ‚produce' mit knapp 1.8 Millionen bestellten Produkten circa 29 Prozent des Gesamtabsatzvolumen des Online-Lieferanten ausmacht und somit die Bestsellerkategorie belegt. Durch die Einbindung der zweitbesten Kategorie ‚dairy eggs', welche ebenfalls die eine Millionenmarke überschreitet, besitzen allein diese zwei Kategorien einen Anteil von circa 46 Prozent aller Bestellungen.

3.2 Bestellungsübersicht

Durch die Aggregierung nach den größten Bestellungen beziehungsweise den Bestellungen mit den meisten Artikeln ergibt sich die Rangliste[9], aus der ersichtlich ist, dass die Topbestellung insgesamt 145 Artikel enthält. Nahezu alle Topbestellungen beinhalten knapp 100 Artikel. Auffällig sind die sich wiederholenden ‚user_id's' in den Top zehn Bestellungen. Daraus kann abgeleitet werden, dass es sich um wiederholte Großbestellungen eines ‚customer' handelt, die somit zu den kaufstärkeren Kunden gehören sollten. Dies wird allerdings im Verlauf dieser Dokumentation genauer erläutert.

Mit Hilfe des Säulendiagramms sind klare Tendenzen der Verteilung von der Anzahl der Artikel pro Bestellung zu erkennen.

[8] Kapitel 2.1.3

[9] Kapitel 2.2.1

Abbildung 2: Die Häufigkeitsverteilung der Produktanzahl pro Bestellung

Es handelt sich um eine Häufigkeitsverteilung, die näherungsweise normalverteilt ist, da sie eine rechtsschiefe Glockenfunktion abbildet. Über die Artikelmenge kann keine genauere Auskunft gegeben werden, da aus dem vorliegenden Datensatz nicht ersichtlich wird, welche Menge der einzelnen Artikel bestellt wird[10]. Die meisten Bestellungen beinhalten fünf Artikel und insgesamt beinhalten 62% aller Bestellungen eine Artikelanzahl von 1 bis 10. *Infolge kann eine Schwelle für den kostenlosen Versand von zehn Produkten im Onlineshop gesetzt werden. Demnach steigt der Reiz die Versandkosten zu sparen und die kleineren Bestellungen werden tendenziell größer. Alternativ kann für einen jährlichen Beitrag von 39,99 Euro für Premiumuser der Versand kostenfrei sein, wodurch die Häufigkeitsverteilung ebenfalls auf der x-Achse nach rechts verschoben werden sollte und somit zu einem höheren Absatz führen könnte.*

3.3 Übersicht der Regionen

In Kalifornien gibt es große regionale Nachfrageabweichungen beim Online Lebensmittel-Lieferanten. Es existiert eine Differenz der absoluten Bestellungen des besten und schlechtesten ‚county' von einer halben Million.

Das Kartendiagramm stellt mittels Farbverlauf die absoluten Verkäufe der Produkte je ‚county' dar. Je dunkler das mintgrün, desto höher die Beliebtheit des Produktes. Der Durchschnitt der gesamten Artikel pro ‚county'[11] liegt bei 105.961.

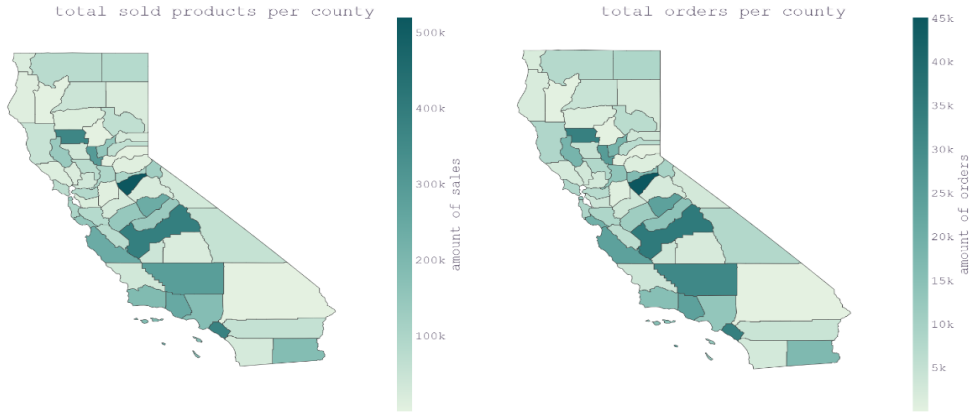


Abbildung 3: Absolute Anzahl an bestellten Artikeln und Bestellungen je Region

[10] Kapitel 2.2.2

[11] Kapitel 2.3.1.6

Im Vergleich dazu wurden auf der folgenden Karte die absoluten Bestellungen je ‚county' abgebildet. Hier liegt der Durchschnitt eines ‚counties' bei 10.430 Bestellungen[12].

In beiden Abbildungen sind grundlegend sehr ähnliche Muster zwischen Bestellungen und Artikeln zu erkennen. Die Top zehn ‚counties' weisen eine sehr ähnliche Rangfolge der Bestellungen und Artikel auf[13]. Die Topregion Calaveras hat über eine halbe Million Bestellungen.

Abbildung 4: Top 10 der kaufstärksten Regionen

Das Schlusslicht besteht aus Butte, Placer, San Bernardino und Trinity. Die absoluten Bestellungen beinhalten in der Summe weniger als 1000 Artikel je Region über den gesamten Zeitraum des Datensatzes. In Butte wurden sogar nur vier Artikel verkauft. Die Gesamtbestellungen sind mit drei, 33, 79 und 109 marginal[14]. *In diesen Regionen muss der Fokus gezielt auf intensives Marketing auf den sozialen Netzwerken gelegt werden. Dadurch wird die richtige Zielgruppe über dasselbe Medium erreicht, wie bestellt wird.*

°Eine individuelle Umstrukturierung der Lagerzentren sortiert nach der Anzahl an Bestellungen je Kategorie bspw. Beginn von ‚produce' mit der ‚aisle_id' von eins etc. wird empfohlen. Beliebte Produkte an den Ganganfang setzen, um Lagerwege zu sparen wie z.B. Bananen und Bioerdbeeren im Gang ‚fresh fruits' ist ebenfalls sinnvoll. Eine Fusionierung der nachfrageschwächsten Region Tehama mit Butte, Placer mit El Dorado und Riverside mit San Bernardino wird auch empfohlen, um die Distributionslager völlig auszuschöpfen ohne die Puffer zu schmälern, da Lieferengpässe weiterhin problemlos bewältigt werden müssen. (Annahme: Je kleiner die ‚aisle_id', desto näher am Lieferein- und Ausgang)°

3.4 Kundenverhalten

Im Folgenden wird das Verhalten der Kunden genauer analysiert. Dabei wurden die besten Kunden über den gegebenen Zeitraum durch verschiedene Kriterien ermittelt. Untypisch ist die maximale Anzahl an Bestellungen von exakt 100 pro Kunde im vorliegenden Datensatz. Es gibt über 156 Kunden, die genau 100 Bestellungen getätigt haben[15]. Daraus lassen sich zwei mögliche Begründungen ableiten. Erstens könnten die User einen Premiumaccount erhalten, die untypischerweise eine neue ‚user_id' erhalten, wodurch bessere Konditionen für die folgenden Bestellungen gewährt werden. Alternativ nutzen die User das Schlupfloch des Gratisversand der ersten 100 Bestellungen und erstellen sich anschließend einen neuen Account, bei dem die ersten 100 Bestellungen erneut versandkostenfrei sind. Die fünf schwächsten Kunden gaben lediglich drei Bestellungen in Auftrag. Aus den spezifischen

[12] Kapitel 2.3.1.7
[13] Kapitel 2.3.1.2, 2.3.1.3
[14] Kapitel 2.3.1.4, 2.3.1.5
[15] Kapitel 3.1.1

Bestellungen geht nicht hervor, weshalb die Kunden nicht wiederbestellt haben. Denn sie enthielten hauptsächlich beliebte Produkte[16]. *Mittels einer zeitlich begrenzten Rabattierung von 15 Prozent per Mail auf die nächste Bestellung für unregelmäßig aktive Kunden, d.h. einen Durchschnitt des Wertes ‚days_since_prior_order' von über 30, innerhalb der nächsten 24 Stunden kann die bisher geschwächte Kundenbindung gestärkt werden. Sofern dies scheitert, kann ein zweiminütiger Fragebogen per Mail zugesandt werden, indem unverbindlich mögliche Gründe der Unregelmäßigkeit im Kauf abgefragt werden, der bei erfolgreicher Teilnahme mit einem fünf Euro Rabattcode belohnt wird.*

Darüber hinaus wurden die Kunden mit den besten Bestellungen zusammengeführt, indem das Verhältnis der Anzahl der bestellten Produkte und die Anzahl der Bestellungen gebildet wurde. Die Top 10 Users[17] kaufen durchschnittlich 50 Produkte pro Bestellung. Außerdem liegt das Verhältnis bei den Kunden mit den meisten absoluten Produktbestellungen bei 30 Artikel pro Bestellung.

Um das Verhalten des Wiederverkaufes tiefgründiger zu verstehen, wurden die Produkte, welche in mindestens 40 Bestellungen waren, in folgender Tabelle[18] ausgegeben, die von nahezu keinem User wiederbestellt wurden.

product_id	product_name	total_orders	total_reorders	bought_in_counties	department_id	aisle_id	reorder_percentage
23369	Tamarind Paste	57	0.0	30	13	72	0.000000
22087	Whole Celery Seed	40	0.0	19	13	104	0.000000
6904	Toasted Sesame Seeds	40	0.0	23	13	104	0.000000
34592	Organic Allspice	114	1.0	34	13	104	0.877193
49284	Fennel Seed	152	2.0	39	13	104	1.315789
29658	Organic Ground White Pepper	74	1.0	24	13	104	1.351351
30928	Organic Gluten Free Yellow Cornmeal	73	1.0	26	13	17	1.369863
1658	Double Superfine Mustard Powder, Original English	69	1.0	25	13	104	1.449275
41348	Shortening, All-Vegetable	64	1.0	29	13	17	1.562500
44626	Ground Allspice	101	2.0	30	13	104	1.980198

Tabelle 2: Schwächsten zehn Wiederbestellraten aller Produkte mit mind. 40 Bestellungen

Von den schlechtesten 50 Produkten, aggregiert nach der Wiederverkaufsrate, sind 42 vom ‚department' pantry (‚department_id': 13) und 29 in dem Gang ‚spices seasoning' (‚aisle_id': 104)[19]. Diese Produkte sind zwar unbeliebt, werden aber auf Grund des Vorratskaufes grundsätzlich unregelmäßiger gekauft. Dennoch sollte die Qualität dieser Artikel überprüft werden.

3.5 Zeitanalyse

In diesem Absatz werden die Zeiten der Buchungen genauer untersucht. Hierbei werden nachstehend die einzelnen Tage des Datensatzes durch jeweils eine Säule im Diagramm visualisiert.

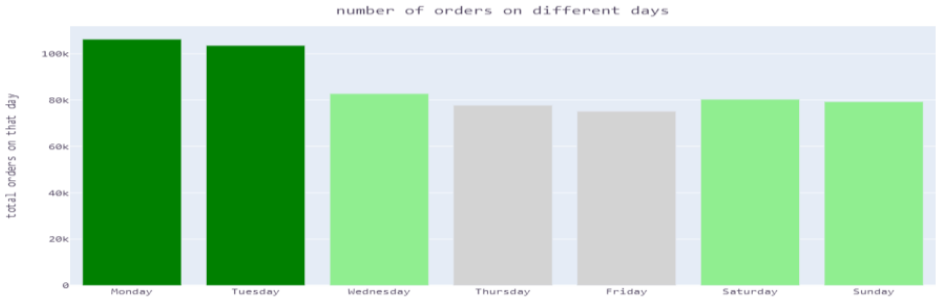


Abbildung 5: Bestellungen über alle Wochentage

Es besteht ein klarer Trend der Bestellungen über alle Wochentage. Am Montag und Dienstag zeichnen sich die größten Nachfragen mit

[16] Kapitel 3.3
[17] Kapitel 3.2
[18] Kapitel 3.4
[19] Kapitel 3.4.1

jeweils über 100.000 Bestellungen ab. In der Mitte der Woche flacht die Kurve auf das Minimum am Freitag von etwa 75.000 Bestellungen ab. Sodass die Nachfrage am Wochenende auf wieder bis zu 80.000 Bestellungen steigt.

Auch über alle Stunden der Tage sind klare Tendenzen zu erkennen. Die nachfragestärksten Stunden liegen von neun bis 17 Uhr bei einem Bestellschnitt von über 45.000 pro Stunde. In den Abendstunden sinken die Bestellungen auf unter 10.000 ab und am frühen Morgen liegen sie sogar bei unter einem Tausend.

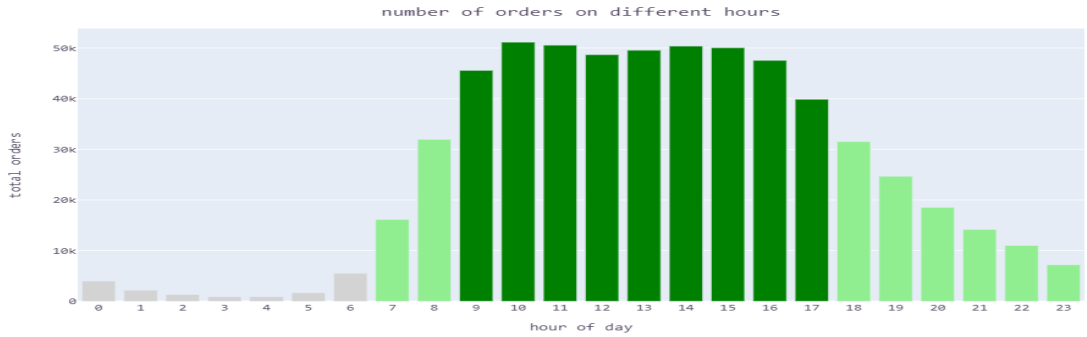


Abbildung 6: Bestellungen über alle Stunden

°Durch diese Auswertung empfiehlt sich, sofern möglich, den Prozess der Beschaffung der Ware auf die Uhrzeiten von 19 Uhr bis 7 Uhr zu beschränken, damit die Beschaffungen und Auslieferungen nicht miteinander kollidieren, optimal an den Wochentagen Donnerstag und Freitag. Vorausgesetzt, die Bestellungen werden überwiegend am selben Tag zum Versand fertiggestellt.°

3.6 Warenkorbanalyse

Genau 90 Prozent der Verkäufe fallen auf genau 19,7 Prozent der Produkte zurück. Das bedeutet, dass genau 8986 Produkte 5531092 Verkäufe ausmachen[20]. Außerdem machen 783 Produkte genau 50 Prozent aller ‚orders' aus[21]. Das wiederum deutet auf eine klare Ungleichverteilung der Verkäufe aller Produkte hin. Diesbezüglich ist ein Säulendiagramm mit allen Produkten je Kategorie. Alle Kategorien besitzen einen wesentlich geringeren Anteil an Bestseller Produkten im Vergleich zu den Unpopulären.

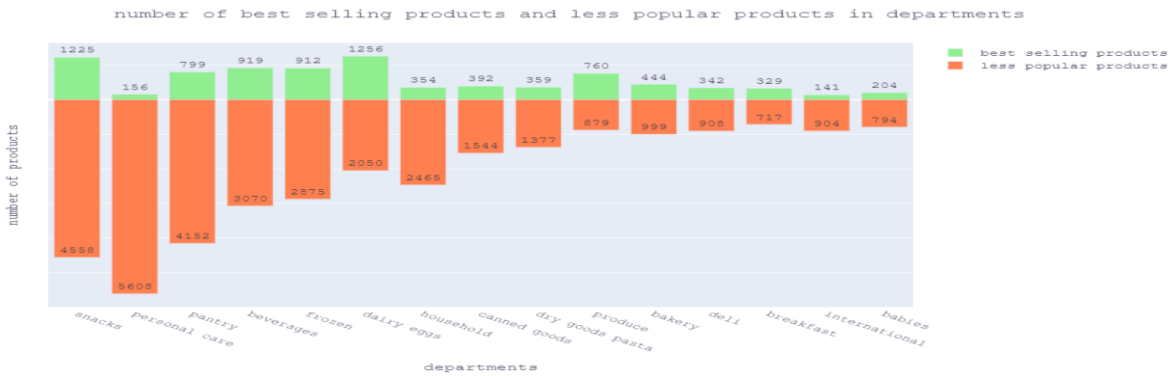


Abbildung 7: Absoluter Anteil der Verteilung aller Produkte je Kategorie

[20] Kapitel 4.2.1.1

[21] Kapitel 4.2.1.2

Anknüpfend dazu wird in Abbildung elf der relative Anteil der meistverkauften Produkte über alle Verkäufe je Kategorie dargestellt. Dies bestärkt die ungleiche Nachfrage innerhalb der Kategorien. Die Aufteilung des Balkens zeigt das Verhältnis der Bestsellerprodukten von 90 Prozent zu den unbeliebten Produkten von zehn Prozent.

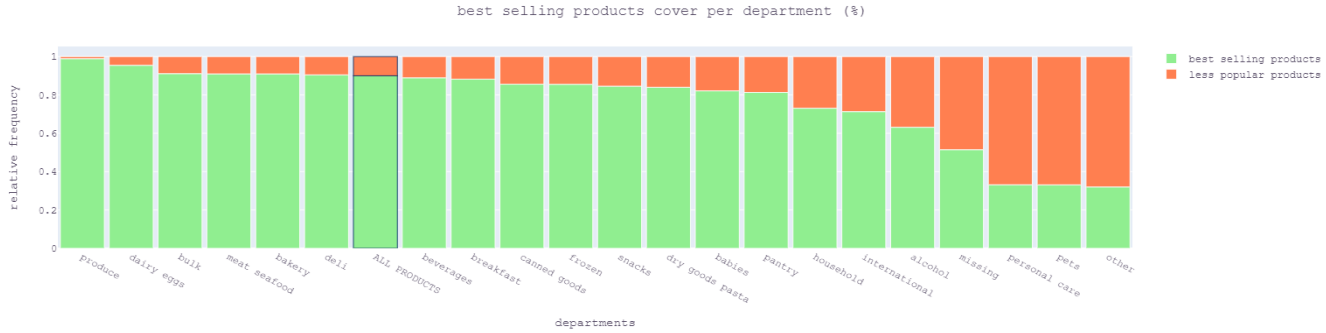


Abbildung 8: Relativer Anteil der meistverkauften Produkte über alle Verkäufe je Kategorie

Womöglich besitzt der Online Lebensmittel-Lieferant eine zu große Produktvielfalt. °Auf Grund der Tatsache, dass knapp zwei Prozent der Produkte die Hälfte aller Verkäufe ausmachen, kann ein mögliche Minderung der Produktvielfalt sinnvoll sein, um hohe Lagerkosten zu sparen.° Das einzige Manko hierbei sind die fehlenden Informationen über den Absatz je Produkt. Im weiteren Verlauf werden weiter Warenkorbanalysen durchgeführt.

4 Konkrete Fragen der Mitarbeiter

4.F1 Frage 1

Prüfen Sie, ob es belastbare Unterschiede im Wiederbestellverhalten zwischen den Produkten mit den product_ids 6217, 14778 und 23579 gibt.

Unter Zuhilfenahme des Säulendiagramms wurden die drei Produkte ‚Peanut Butter Chocolate Chip Fruit & Nut Food Bar, Organic Chocolate Chip Chewy Granola Bars und Pumpkin Spice Protein Bar' untereinander durch die Anzahl der Bestellungen und Wiederbestellungen verglichen. Alle drei Produkte gehören der Produktkategorie ‚snacks' an.

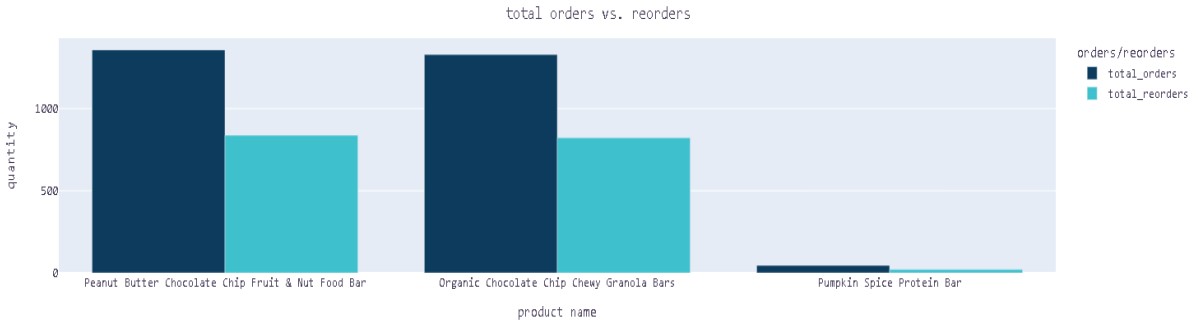


Abbildung 9: Absolute Bestellungen und Wiederbestellungen der drei Riegel

Die ersten beiden Produkte weisen keine belastbaren Unterschiede bei der absoluten Anzahl an Bestellungen auf, die bei 1329 und 1357 liegen, sowie bei der Wiederbestellrate, die jeweils bei circa 62 Prozent liegen. Durch die Unbeliebtheit, das heißt eine geringe Anzahl an Bestellungen des Artikels ‚Pumpkin Spice Protein Bar' von 46, ist ein Vergleich mit den ersten zwei Riegeln nicht aussagekräftig und somit nicht belastbar differenzierbar[22]. Die zwei

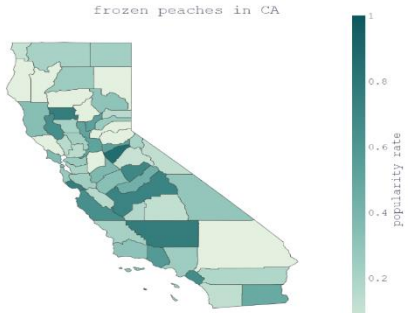
[22] Kapitel 2.2.F1.1+F1.2

beliebteren Riegel wurden in über 40 ‚counties' bestellt, wohingegen das nachfrageschwächste Produkt nur in 14 ‚counties' geordert wurde[23]. Dennoch ist die Wiederbestellrate dieses Artikels über 15 Prozent geringer. Durch die Spezifizierung auf Riegel und Proteinriegel innerhalb der Produktkategorie wurde durch die nahezu gleiche Wiederkaufraute bestätigt[24], dass es sich beim dritten Riegel um ein weniger nachgefragtes Produkt handelt und somit als ein möglicher Kandidat eines Auslaufmodels deklariert werden kann. *Sofern der Proteinriegel neu im Segment gelistet ist und bisher nicht viel Aufsehen erregen konnte, kann zur Verbreitung ein Bundle aus ‚Pumpkin Spice Protein Bar' und den zwei beliebtesten Proteinriegel ‚Brownie Crunch High Protein Bar' und ‚High Protein Bar Chunky Peanut Butter' erstellt werden, bei dem die Käufer in Summe 20 Prozent sparen können. Aufgrund dessen kommt der Riegel in mehr ‚counties' in Umlauf und bekommt die Chance, von den Konsumenten getestet zu werden.*

4.F2 Frage 2

Schwankt die Popularität von Produkt 9390 zwischen den Regionen?

Grundlegend sind klare Schwankungen der Popularität des Produkts ‚frozen peaches' zu erkennen. Das Kartendiagramm veranschaulicht die Popularität mittels des grünen Farbverlaufes. Je dunkler das grün, desto höher die Popularität des Produktes. Sie bildet sich aus der gewichteten Addition von 70 Prozent des normalisierten relativen Anteiles aller User im jeweiligen ‚county', die das Produkt erworben haben, und 30 Prozent der normalisierten durchschnittlichen Bestellmenge der User vom Produkt des ‚counties'.



Die höchste Popularität befindet sich in den bereits ermittelten nachfragestärkeren Regionen wie Glen, Kern und Calaveras. Hingegen liegt die Beliebtheit der nachfrageschwächsten ‚counties' wie Butte, Placer und San Bernardino bei null.

Abbildung 10: Popularität des Produktes 'frozen peaches' in Kalifornien

4.F3 Frage 3

Sind die Produkte 9390, 2713, 21883 und 16753 in den gleichen Regionen populär und unpopulär, oder unterscheiden sich die Muster?

Die Popularität der Produkte ‚frozen peaches', ‚fresh frozen chopped spinach', ‚fresh frozen chopped spinach' und ‚pinot grigio wine' werden nachstehend durch Mapplots dargestellt. Prinzipiell sind Muster zwischen den Produkten auf regionaler Ebene ersichtlich.

Verstärkte räumliche Übereinstimmungen der vier Karten ähneln den allgemein nachfragestärkeren Regionen. Die ersten beiden Artikel befinden sich in den Kategorien ‚frozen'. Die Produkte drei und vier befinden sich im ‚department' und ‚alcohol'. Folgerichtig wäre eine gute Korrelation zwischen den Alkoholprodukten. Allerdings liegt der Korrelationskoeffizient bei nur 0,36[25].

[23] Kapitel 2.2.F1.3

[24] Kapitel 2.2.F1.4

[25] Kapitel 2.3.F3.3

Unüblich sind die gegensätzlichen Popularitäten der Regionen zwischen den alkoholischen Artikeln. Die Gleichheit der Muster von Karte eins und zwei sind, wie erwartet, nahe zu identisch. Außerdem besteht eine gute Musterübereinstimmung zwischen dem ‚irish whiskey' und den beiden ‚frozen' Artikel.

Alles in allem besitzen alle vier Karten ein gewissen Grad an regionalem Muster hinsichtlich der Popularität. Lediglich die Alkohole weisen kaum Übereinstimmung in ihrer Beliebtheit auf.

Abbildung 11: Popularität der Produkte 9390, 2713, 21883 und 16753

4.F4 Frage 4

Welche Counties sind sich ähnlich im Hinblick auf den Produktmix?

Nach Ermittlung des individuellen Produktmix aller Regionen entstehen Cluster, die Gemeinsamkeiten im Kaufverhalten aufweisen. Jedes ‚county' wird einer der sechs Gruppen zugeordnet und nachfolgend durch ein Kartendiagramm dargestellt.

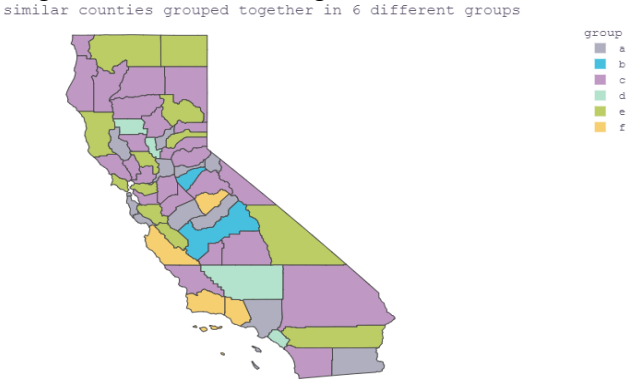


Abbildung 12: Gruppiert vergleichbare Produktmixe

Alle sechs Gruppen untereinander besitzen einen ähnlichen Produktmix. Unter Zuhilfenahme der Tabellen erkennt man die Größe der einzelnen Gruppen. Auffällig hierbei sind die Muster der einzelnen Gruppierungen. Denn die generell meisten Bestellungen pro Region hat die Gruppe b mit Calaveras und Fresno, gefolgt von d mit Glenn, Kern, Orange, Sutter und f mit Mariposa, Monterey, Santa Barbara und Ventura. Daraus kann geschlossen werden, dass der Produktmix und die Kaufkraft der ‚counties' stark miteinander zusammenhängen.

group		group		group		group		group		group	
county		county		county		county		county		county	
Yuba	a	Alpine	a	Sonoma	c	Napa	c	Kern	d	San Benito	e
Santa Cruz	a	San Francisco	a	Stanislaus	c	Butte	c	Orange	d	Contra Costa	e
Los Angeles	a	Fresno	b	Tehama	c	Colusa	c	Santa Clara	e	Riverside	e
Sacramento	a	Calaveras	b	Trinity	c	Del Norte	c	Yolo	e	Plumas	e
Imperial	a	San Luis Obispo	c	Tulare	c	Mono	c	Marin	e	Nevada	e
Madera	a	San Diego	c	Tuolumne	c	El Dorado	c	Inyo	e	Mariposa	f
Lake	a	San Bernardino	c	Shasta	c	Humboldt	c	Siskiyou	e	Monterey	f
San Mateo	a	San Joaquin	c	Placer	c	Lassen	c	Mendocino	e	Ventura	f
Merced	a	Sierra	c	Alameda	c	Sutter	d	Modoc	e	Santa Barbara	f
Amador	a	Solano	c	Kings	c	Glenn	d				

Tabelle 3: Zuweisung der 'counties' mittels Produktmixähnlichkeit zu einer Gruppe

Des Weiteren besteht auf Produkt- und Kategorie Ebene generell eine sehr hohe Korrelation zwischen nahezu allen ‚counties' mit Ausnahme der bestellschwachen ‚counties' wie San Bernardino[26].

4.F5 Frage 5

Welche der TOP 50-Produkte sind sich ähnlich im Hinblick auf die regionale Verteilung?

Die Top 50 Produkte wurden mittels einer Clusteranalyse in vier Gruppen kategorisiert. Die erstellten Cluster ähneln sich untereinander im Hinblick auf die regionale Verteilung[27]. Es handelt sich bei den Produkten nahezu ausschließlich um Obst und Gemüse. Deshalb befinden sich in den Gruppen sehr ähnliche Artikel. Allerdings heben sich die einzigen zwei Getränke ‚sparkling water grapefruit' und ‚spring water' der Top 50 in eine eigene Gruppe heraus. Alles in allem sind die regionalen Verteilungen aller Top 50-Produkte sehr ähnlich.

4.F6 Frage 6

Fallen Ihnen weitere Zusammenhänge zwischen Produkten, Kategorien, oder Gängen auf, die Sie der Marketing-Leitung zur Optimierung der Website-Platzierungen empfehlen würden?

Grundlegend wird die Häufigkeitsverteilung der Rangfolge von Produkten absteigend, die zum Warenkorb hinzugefügt werden, von den Abteilungen 'produce' und ‚dairy eggs' dominiert[28]. Interessant hierbei ist die erhöhte Wahrscheinlichkeit der Abteilungen ‚snacks', ‚frozen' und ‚panty', dass diese Produktgruppen tendenziell später zum Warenkorb hinzugefügt werden.

Ein Vergleich der Kategorien ‚beverages' und ‚alcohol' gibt Aufschluss über die zeitliche Abhängigkeit der täglichen Nachfrageänderungen. In folgendem Barplot werden die normalisierten relativen Anteile der beiden Kategorien über die ganze Woche abgebildet, da von ‚beverages' generell wesentlich mehr Produkte bestellt werden.

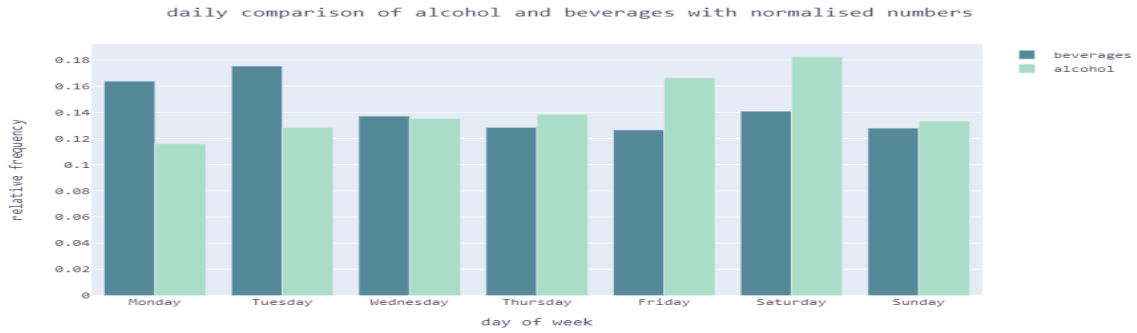


Abbildung 13: Tägliche Vergleiche der normalisierten relativen Werte zweier Kategorien

Demungeachtet ändert sich die Nachfrage der beiden Produktgruppen. Getränke werden zu Beginn der Woche häufiger bestellt. Die Nachfrage nach dem Alkohol steigt logischerweise Richtung Wochenende. Diese Behauptung wurde durch einen Chi-Quadrat -Test überprüft und bestätigt[29]. Das bedeutet es besteht ein hochsignifikanter Unterschied in der zeitlichen Verteilung von Produkten aus den Kategorien ‚beverages' und ‚alcohol'. Zu empfehlen ist eine Priorisierung vom Alkohol auf der Startseite der Webseite ab Donnerstag bis einschließlich

[26] Kapitel 2.4.F4.5 + Kapitel 2.4.F4.6
[27] Kapitel 2.4.F5.4 + F5.5+ F5.6
[28] Kapitel 4.3.F6.1
[29] Kapitel 4.3.F6.2

Sonntag. Zu Beginn der Woche kann der Fokus der Startseite auf nachfragestärkere Kategorien gelegt werden.

Der Trend der biologisch beziehungsweise ökologischen Lebensmittel wird durch das folgende Säulendiagramm untermauert. Hierbei wurden alle Regionen auf biologische und nicht biologische Produkte separiert, um diese miteinander zu vergleichen.

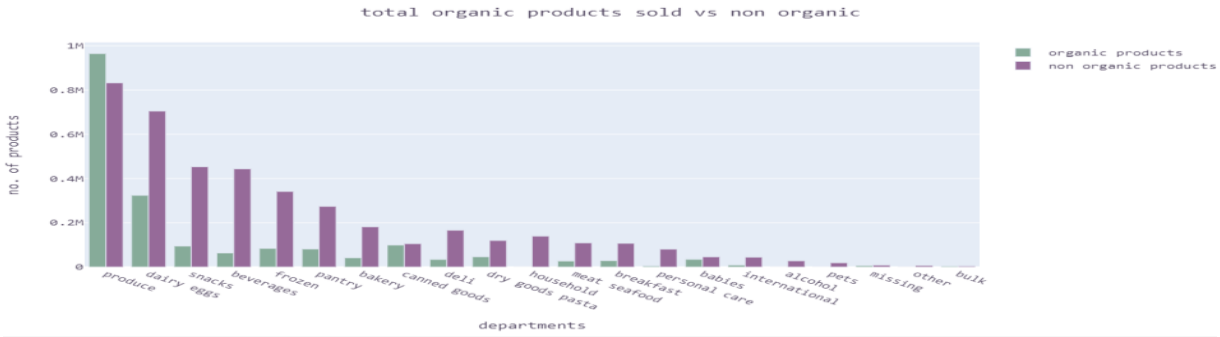


Abbildung 14: Absoluten Verkäufe von ‚organic' und ‚non organic' Produkten

In der nachfragestärksten Kategorie werden bereits mehr ‚organic' Produkte bestellt. Knapp 32 Prozent aller Verkäufe beinhalten ‚organic' Artikel. Obwohl nur circa zehn Prozent der Produkte im Sortiment ‚organic' sind[30]. Um eine zielgruppengerechte Werbung zu schalten, empfiehlt sich daher einen Slogan zu wählen, welcher für die Nachhaltigkeit plädiert. Ein Beispiel dafür wäre ‚Vorsprung durch Nachhaltigkeit'.

Die besten ‚aisle' Kombinationen, das bedeutet die am häufigsten zusammenbestellten ‚aisle' Paare über alle Bestellungen, sind ‚fresh fruits' und ‚fresh vegetables' mit über 1.26 Million[31]. Die Top zehn der besten Produktkombination, sind wie erwartet die allgemeinen Bestseller des Lieferanten, setzen sich ausschließlich aus Obst und Gemüse zusammen[32]. Die beliebtesten Produkte sollten auf der Webseite schneller zu erreichen sein, damit der Komfort der Kunden durch die geringe Suchzeit maximiert wird. Außerdem ist das Hinzufügen einer individuell angepassten Produktempfehlung durch einen Algorithmus sinnvoll, welcher Produkte empfiehlt, die andere Kunden ebenfalls zusammen gekauft haben wie zum Beispiel ‚Kunden, die ‚milk' gekauft haben, kauften obendrein ‚Vitamin D' (Platz 21 der besten Produktpaare).

5 Resümee und Ausblick

Durch die Ausarbeitung dieses reellen Datensatzes konnten alle wesentlichen und bereits erlernten Fähigkeiten im Kontext von Data Analysis und Business Intelligence angewandt werden. Um eine breitere Datenbasis, d.h. tiefergehende Datenanalysen zu gestalten, ist ein größerer Datensatz erforderlich, welcher zusätzliche Spalten mit mehr Informationen wie Menge, Preis, Zeit, Lagerverfügbarkeit etc. anbietet. Obendrein wären Kundendaten nützlich, sodass die Zielgruppe analysiert werden kann.

Das Ziel hierbei ist tiefgründigere Handlungsentscheidungen ableiten zu können.

[30] Kapitel 4.3.F6.3

[31] Kapitel 4.3.F6.4.1

[32] Kapitel 4.3.F6.4.2