Beschreibung der Daten

Der Instacart Datensatz besteht aus über 3.2 Millionen Bestellungen. Er ist aufgeteilt in die Datentabellen:

- orders: Information über die Bestellungen.
- opp: Information über die in den Bestellungen enthaltenen Produkte.
- tips: Information, ob ein tip gegeben wurde oder nicht.
- products: Information zu den bestellbaren Produkten. Es sind fast 50.000 Produkte vorhanden, welche über 4 Spalten beschrieben werden.
- aisles: Information zu den Inseln, in welchen die Produkte angeboten werden. Im Datensatz sind 134 Supermarktgänge gelistet, welche durch eine schriftliche Beschreibung genau aussagen, was in diesem Gang zu finden ist.
- departments: Information zu den Abteilungen, in welchen die Produkte angeboten werden. Es gibt 21 Abteilungen, in welche jedes Produkt eingeordnet wird. Der Datensatz besteht aus zwei Spalten.

Nachfolgend werden die Spalten des Datensatz orders beschrieben:

- order_id: Numerisches Feature, welches für jede Bestellung eine einzigartige ID enthält. Insgesamt gibt es 3.214.874 Millionen distinkte Bestellungen.
- user_id: Numerisches Feature, welches für jeden distinkten Nutzer eine einzigartige ID enthält. Die Bestellungen wurden von 206209 distinkten Nutzern getätigt.
- tip: Binäres Feature mit der Information, ob ein Tip gebeben wurde. 0 bedeutet, dass kein Tip gegeben wurde, 1 bedeutet, dass ein Tip gegeben wurde. Bei circa 1.9 Millionen Bestellungen wurde Trinkgeld gegeben, bei 1.4 Millionen nicht.
- eval_set: Kategorisches Feature mit der Information, zu welchem Set die Beobachtung gehört. Verfügbare Werte sind: train und test. Train kann für das Training benutzt werden. Für Test-Beobachtungen soll eine Vorhersage generiert werden. Das Feature ist daher im Trainingsset stetig und enthält keine weitere Information.
- order_number: Numerisches Feature mit der Information, die wievielte Bestellung es eines Nutzers ist. Jede Person hat mindestens 3-mal und maximal 99-mal bestellt. Durchschnittlich hat jede Person ca. 17 Bestellungen aufgegeben.
- order_dow: Numerisches Feature mit der Information über den Wochentag der Bestellung. Die Information ist als Zahl zwischen 0 und 6 encodiert. Es wird angenommen, dass die Tage 0 und 1 mit mehr Bestellungen die Tage Samstag und Sonntag sind. Die weiteren Tage sind in der üblichen Reihenfolge. Die meisten Bestellungen erfolgen am Wochenende.
- order_hour_of_day: Numerisches Feature mit der Information über die Uhrzeit der Bestellung. Die wenigsten Bestellungen gehen rund um 3 Uhr nachts ein, die meisten um 10 Uhr vormittags.
- days_since_prior_order: Numerisches Feature mit der Information über die Anzahl der vergangenen Tage seit der letzten Bestellung. Es vergehen zwischen keinem und mehr als 30 Tagen, bis die Personen erneut bestellen. Durchschnittlich wird alle 10,7 Tage erneut bestellt.
- aisles_X: 134 Numerische Feature, die beschreiben, welchem Gang die bestellten Produkte zugeordnet sind. 0 beschreibt, dass eine Bestellung keinen Artikel aus dem

jeweiligen Gang beinhaltet. 1 beschreibt, dass mindestens ein Artikel aus dem Gang enthalten ist.

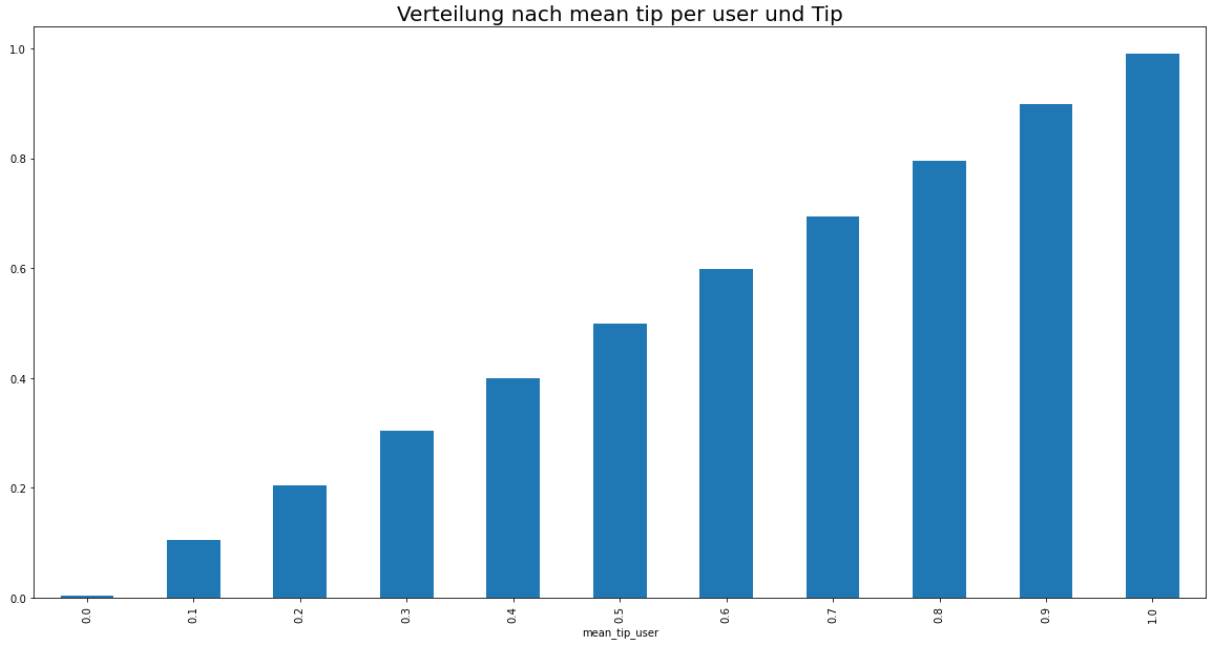
- departments_X: 21 Numerische Feature, die beschreiben, welcher Abteilung die bestellten Produkte zugeordnet sind. 0 beschreibt, dass eine Bestellung keinen Artikel aus der Abteilung beinhaltet. 1 beschreibt, dass mindestens ein Artikel aus der Abteilung enthalten ist.

Im ersten Schritt wurden orders und tips verbunden. Über eine Aggregation wurden zudem die Information zu den Aisles und Departments der in einer Bestellung enthaltenen Produkte integriert.
Die Daten pro Bestellung sind vollständig, bis auf die Spalte “Wochentag der Bestellung” (order_dow), bei der Werte mit NaN angegeben sind, welche später behandelt werden. Alle Daten sind ganzzahlige Werte (int), außer days_since_prior_order, welches ein Kommazahlenwert (float) ist. Eine weitere Ausnahme ist das Objekt eval_set.
Für die Optmierung der Datentypen wurden die Speicherbereiche für die int- und float-Werte jeweils minimal gewählt.

Daten verstehen

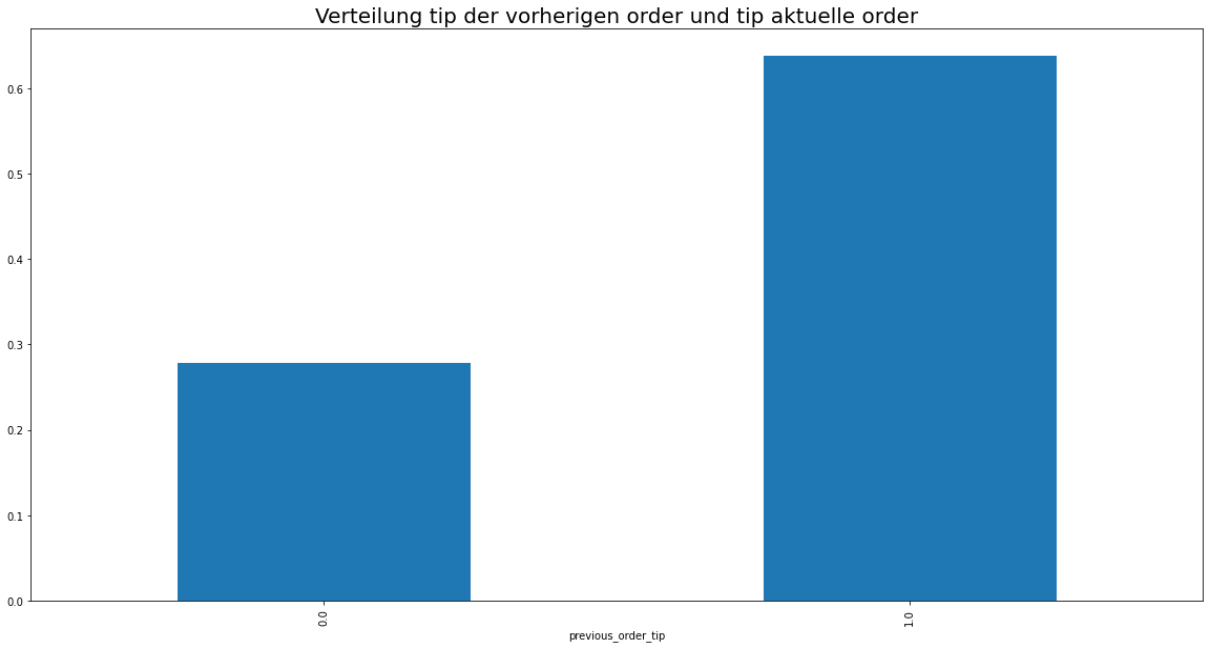
Mean tip per user

Es ist erkennbar, dass Tip stark von dem bisherigen durchschnittlichen Trinkgeld des Users abhängt. Die Wahrscheinlichkeit, ob ein User Trinkgeld gibt, entspricht für jeden Wert nahezu dem bisherigen Durchschnitt des Trinkgeldes. Es wird daher eine hohe Korrelation vermutet.



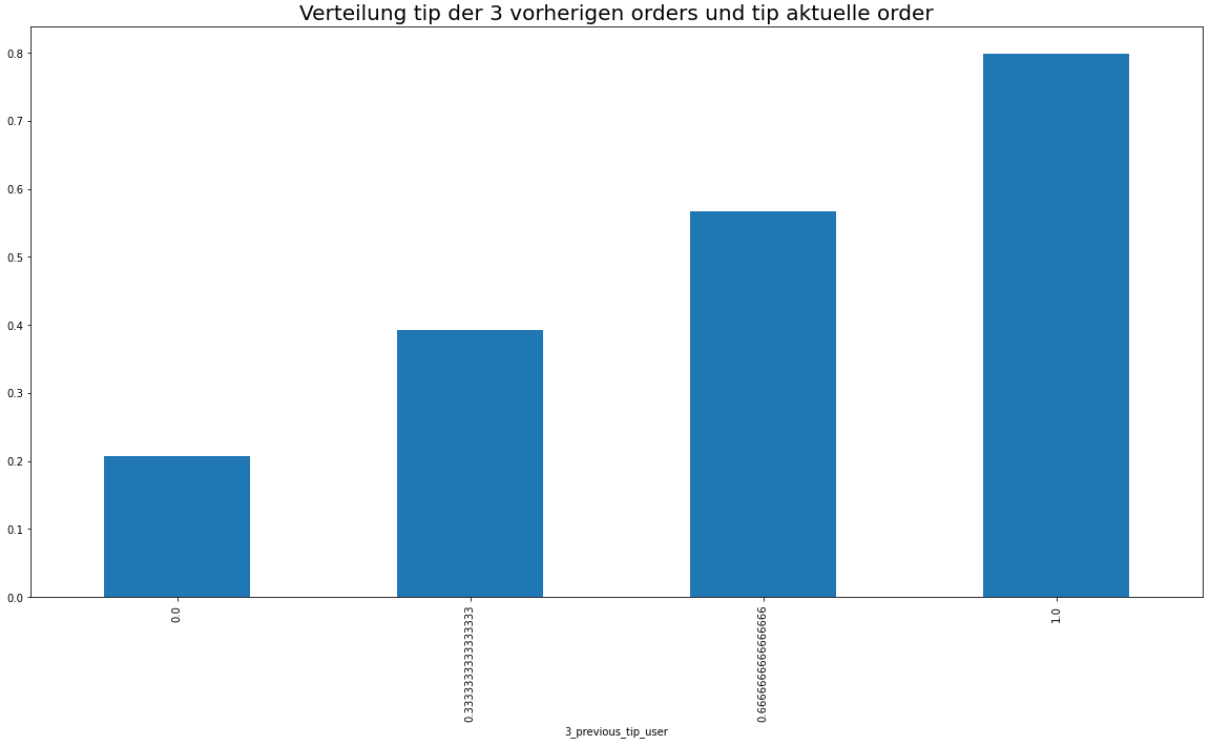
Tip Bestellung zuvor

Bei dem Vergleich der Tips der Bestellung zuvor mit den Tips der aktuellen Bestellung wird ebenfalls ein großer Zusammenhang offensichtlich. Bei über 60% der Bestellungen bei denen ein Trinkgeld gegeben wurde, wurde bei der aktuellen wieder ein Trinkgeld gegeben. Bei den Bestellungen, bei denen kein Trinkgeld gegeben wurde, was das nur bei circa 28 % der Bestellungen der Fall.



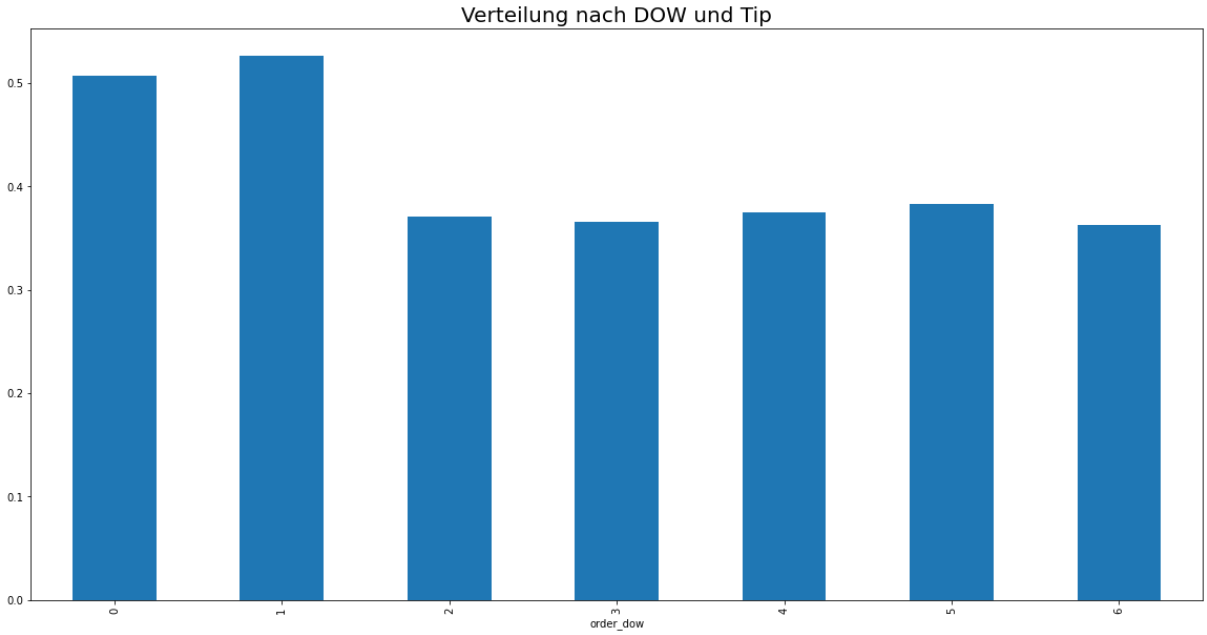
Tip letzten 3 Bestellungen

Ein ähnliches Bild ergibt sich bei der Betrachtung der letzten drei Bestellungen. Je häufiger bei den vergangenen drei Bestellungen ein Trinkgeld gegeben wurde, desto häufiger wurde auch in der aktuellen Bestellung ein Trinkgeld gegeben. Ein starker Zusammenhang wird daher offensichtlich, wobei er jedoch nicht so stark erscheint, wie in den ersten beiden Diagrammen, da die Wahrscheinlichkeitsunterschiede nicht ganz so groß sind.



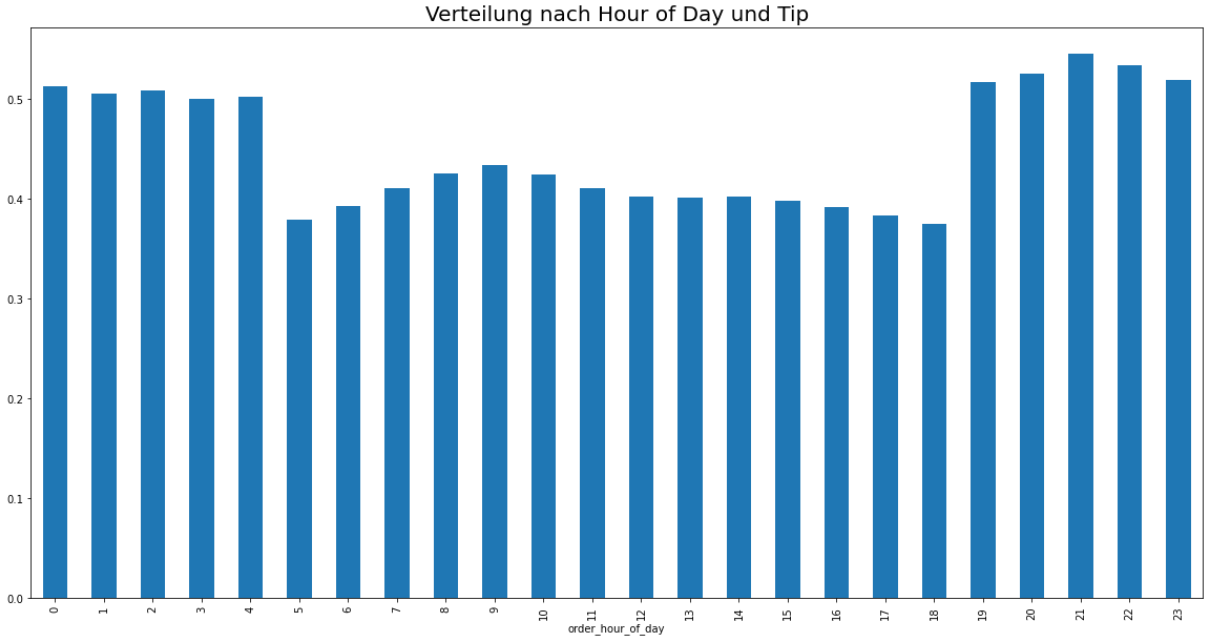
Wochentag

Bei der Betrachtung der relativen Häufigkeit von Tip nach Wochentag ist ein deutlicher Unterschied zwischen dem Wochenende und Werktagen erkennbar. Am Sonntag ist die Wahrscheinlichkeit eines Trinkgeldes am höchsten. Zwischen den Wochentagen sind keine relevanten Unterschiede erkennbar.



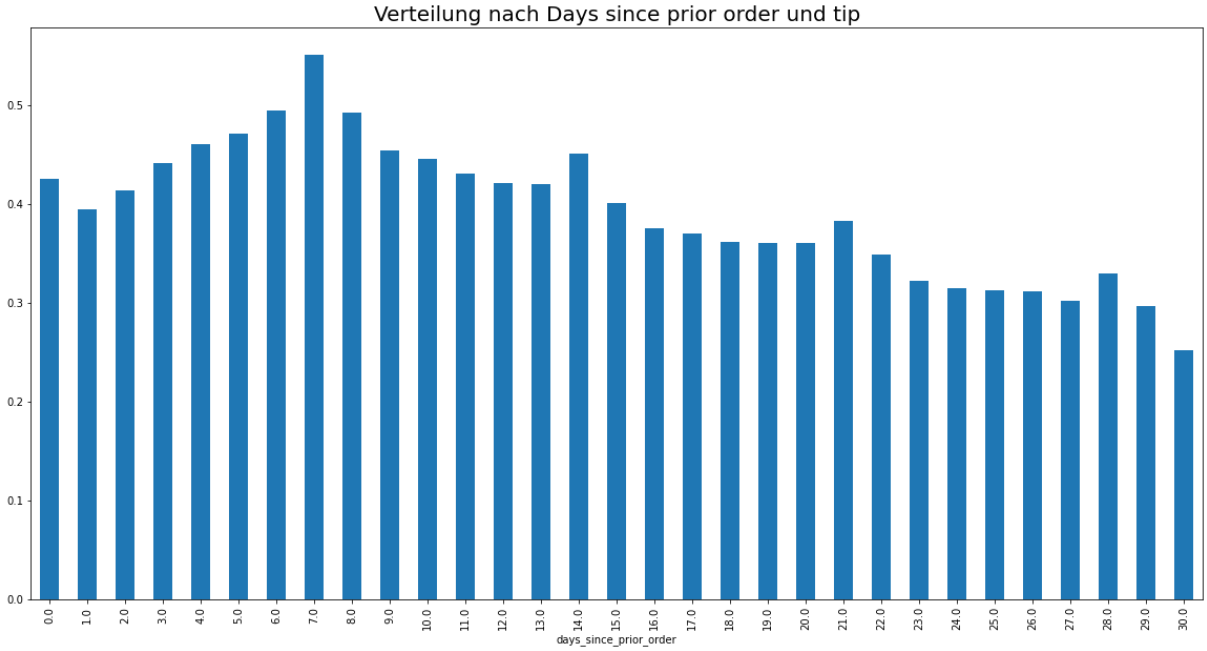
Uhrzeit

Bei der Betrachtung der relativen Wahrscheinlichkeit nach der Uhrzeit lässt sich erkennen, dass die Wahrscheinlichkeit für einen Tip abends deutlich höher ist als tagsüber. Zwischen 4 und 5 Uhr ist ein abrupter Rückgang der Wahrscheinlichkeit für ein Trinkgeld zu erkennen, während sie zwischen 18 und 19 Uhr wieder sprunghaft steigt. Zusätzlich steigt die Wahrscheinlichkeit für ein Trinkgeld morgens von 5 Uhr bis 9 Uhr leicht an, um dann bis 18 Uhr wieder leicht zu sinken.



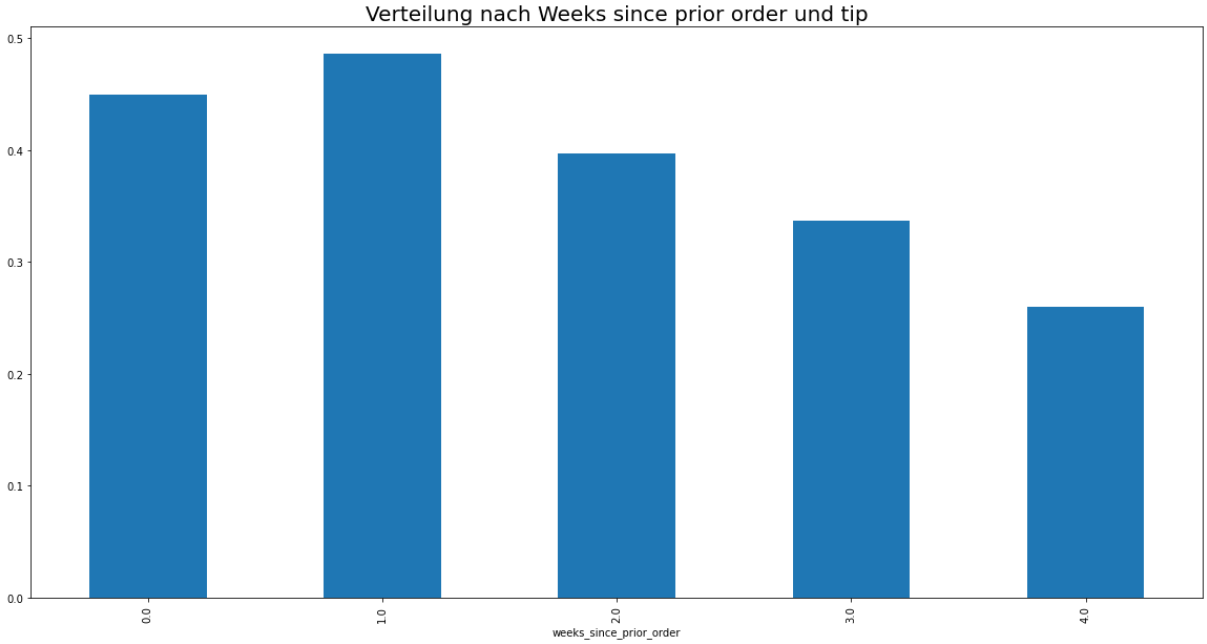
Days since prior Order

Anschließend haben wir die Tip-Wahrscheinlichkeit mit dem Abstand zur letzten Bestellung verglichen. Die Wahrscheinlichkeit für einen Tip steigt dabei innerhalb der ersten 7 Tage an, bevor sie anschließend wieder sinkt. Ausnahmen sind immer die Zahlen der 7er-Reihe, bei denen immer eine Steigerung der Wahrscheinlichkeit zum Vortag erkennbar ist. Das lässt darauf schließen, dass bei Bestellungen im wöchentlichen oder mehrwöchigen Rhythmus die Wahrscheinlichkeit für ein Trinkgeld höher ist.



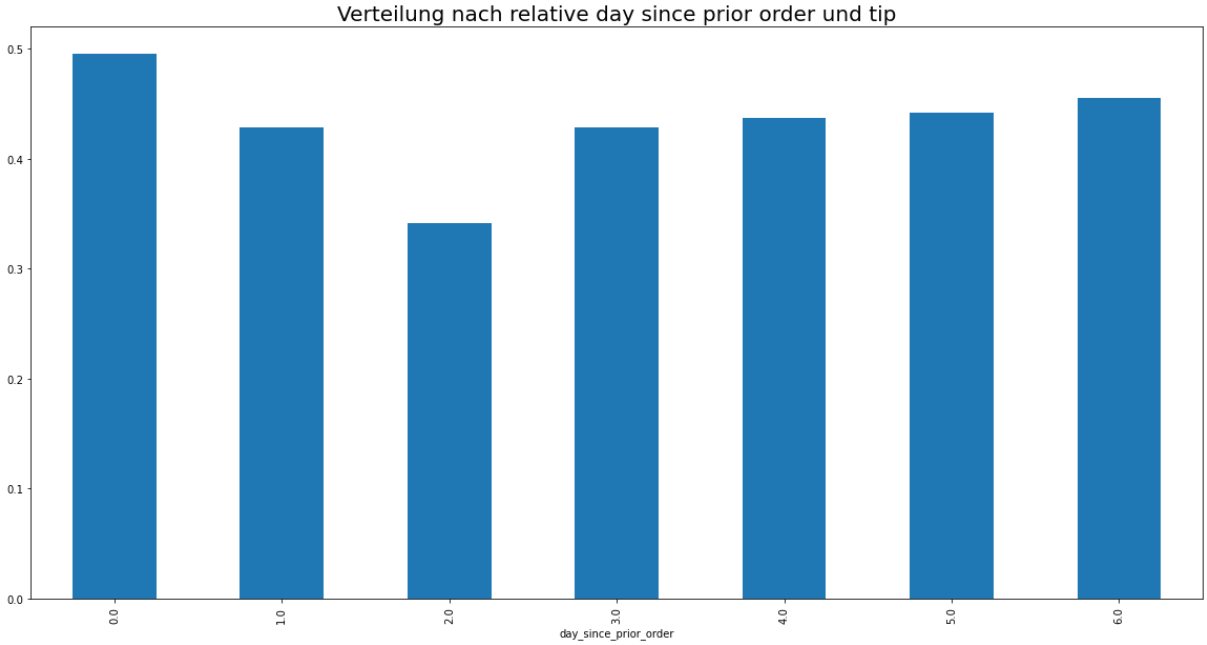
Weeks since prior order

Zur besseren Übersichtlichkeit haben wir anschließend wir uns die Tip-Wahrscheinlichkeit wochenweise angeschaut. Hierbei wurden unsere Erkenntnisse ais dem vorherigen Diagramm bestätigt. Bei Abständen zur letzten Bestellung von bis zu einer Woche steigt die Wahrscheinlichkeit, während sie bei größeren Abständen konstant zurückgeht.



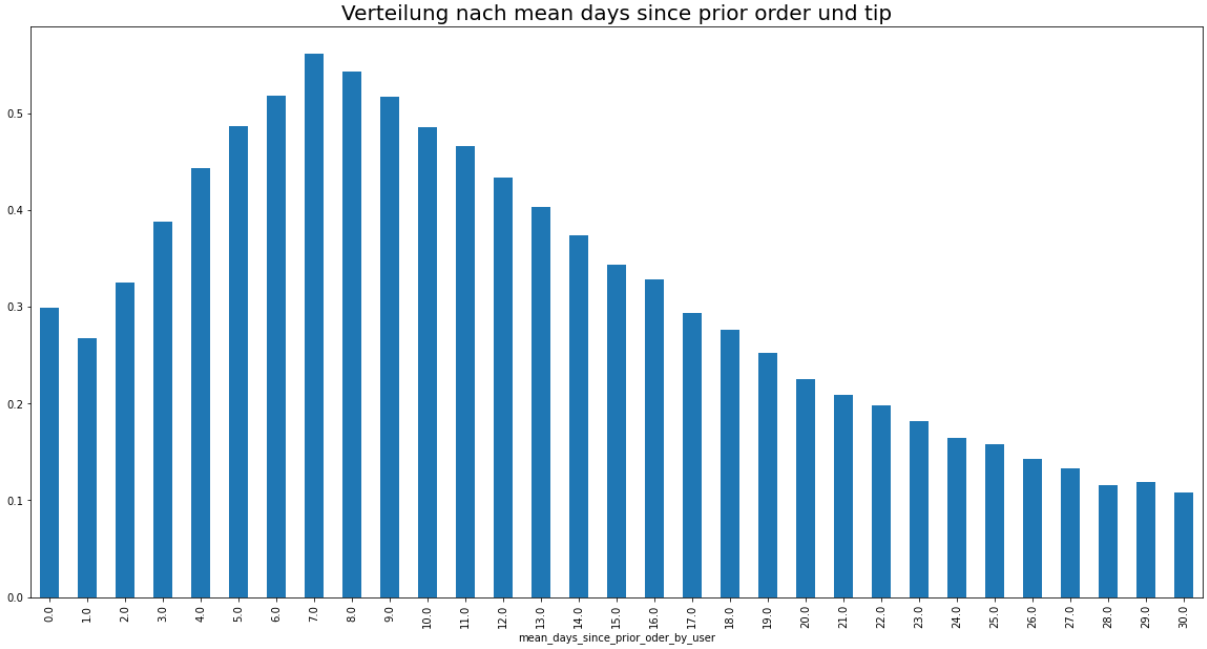
Relative day since prior order

Zusätzlich haben wir die Auffälligkeit untersucht, dass bei wöchentlichen Rhythmen die Tip-Wahrscheinlichkeit höher ist. Dazu haben wir den relativen Abstand zur letzten Bestellung betrachtet, also wie viel Wochentage später die Bestellung aufgegeben wurde. Dabei wurde unsere Erkenntnis bestätigt, dass die Tip-Wahrscheinlichkeit bei wöchentlichen Rhythmen höher ist.



Mean days since prior order

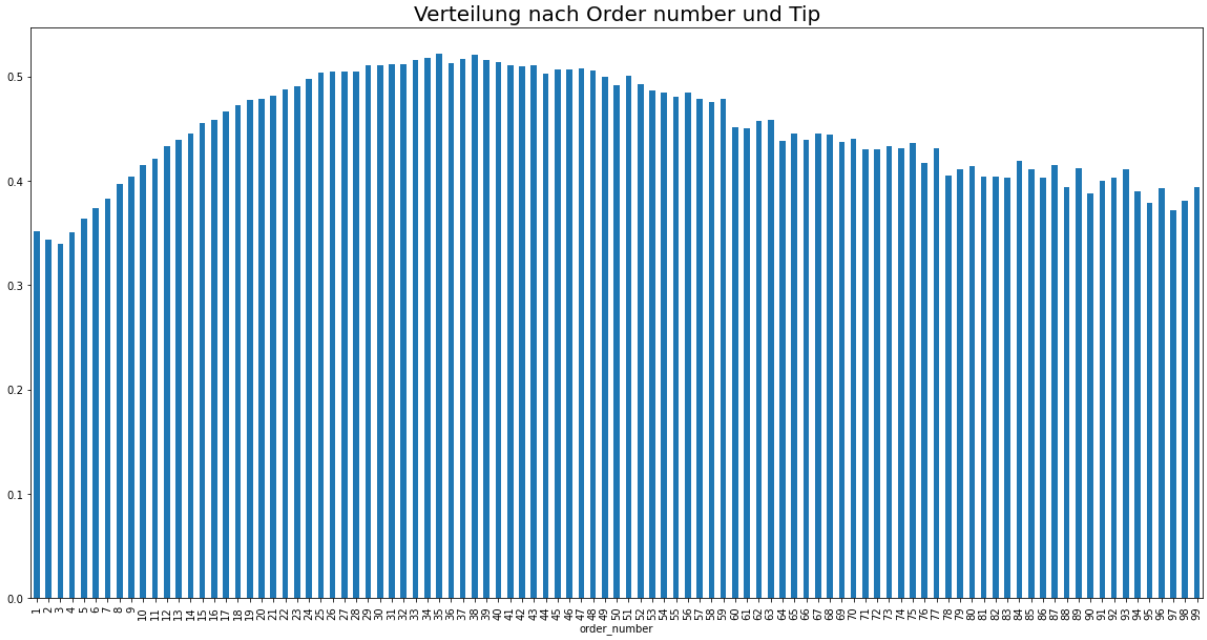
Bei der Betrachtung der Tip-Wahrscheinlichkeit nach zeitlichem Abstand pro User ist erkennbar, dass User mit einem sehr kurzen durchschnittlichen Bestellabstand eher selten ein Trinkgeld geben. Bis zu einem durchschnittlichen Bestellabstand von 7 Tagen steigt anschließend die Tip-Wahrscheinlichkeit, während danach stetig sinkt.



Order number

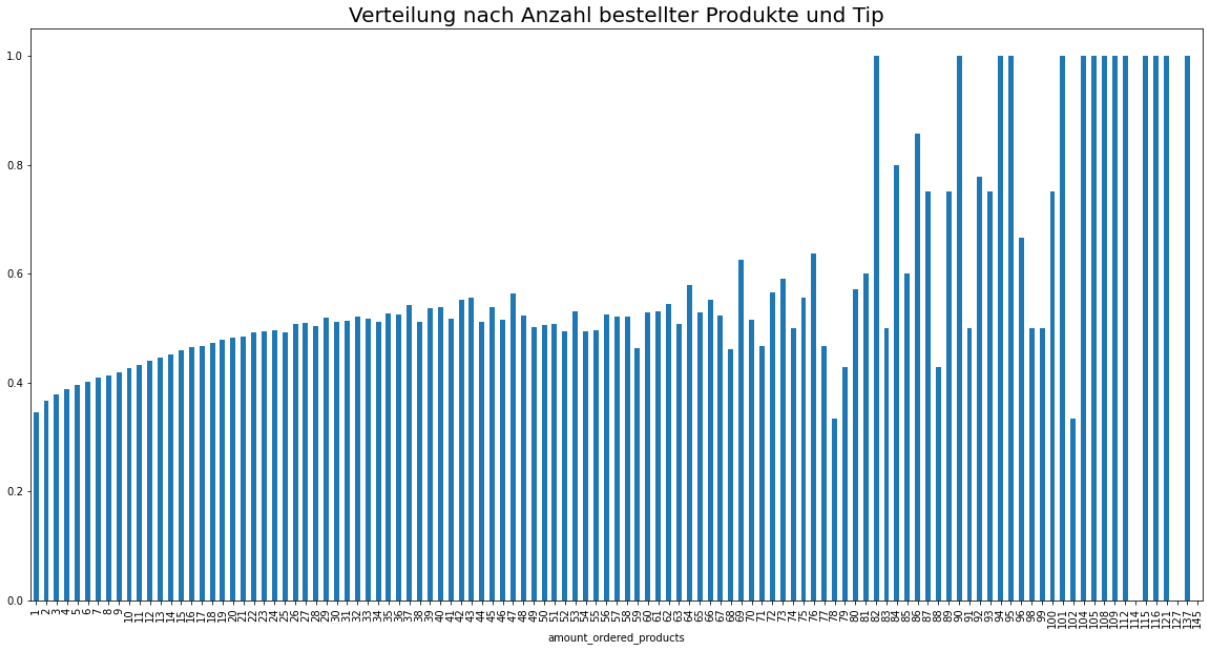
Bei dem Vergleich der Bestellnummer zur Tip-Wahrscheinlichkeit ist bei Bestellung zwei und drei ein kleiner Rückgang im Vergleich zur ersten Bestellung erkennbar. Anschließend steigt

die Tip-Wahrscheinlichkeit zur Bestellnummer linear bis zu einer höchsten Tip-Wahrscheinlichkeit bei Bestellnummer 35. Anschließend fällt die Tip-Wahrscheinlichkeit wieder. Dabei sind einige Ausreißer erkennbar, was aber vermutlich an einer geringen Anzahl an Usern liegt die überhaupt über 70-mal bestellen. Dadurch haben einzelne Zufälligkeiten einen größeren Einfluss auf die Werte.



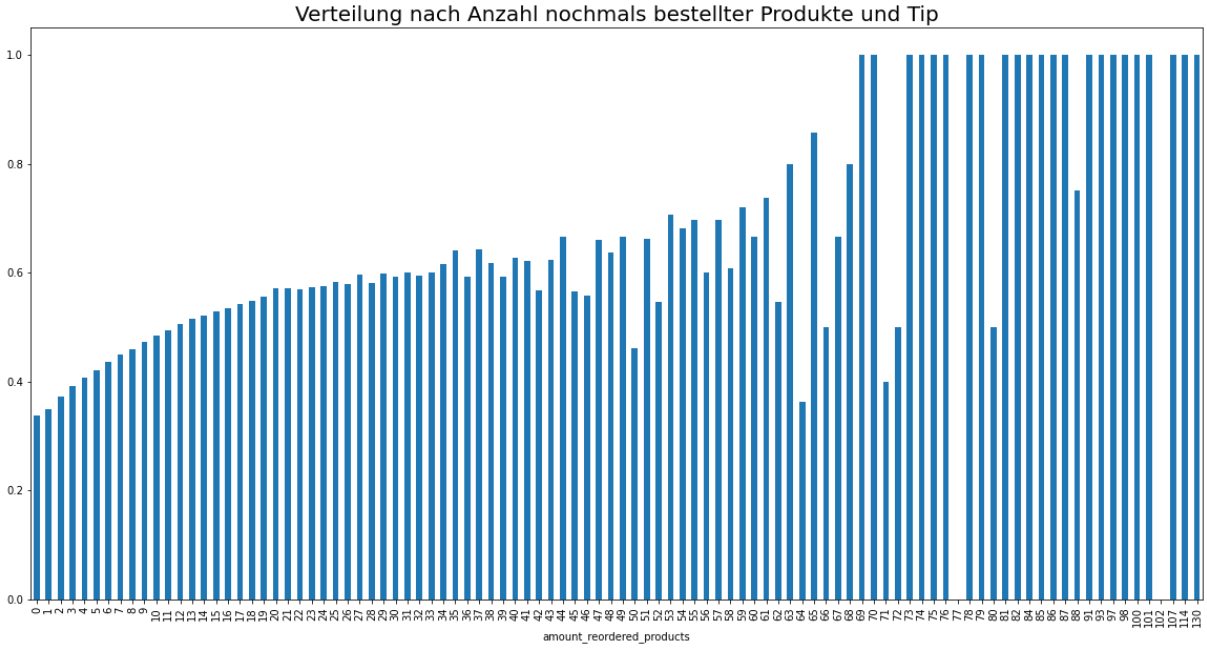
Anzahl bestellter Produkte

Bei der Betrachtung der Anzahl der bestellten Produkte zur Tip-Wahrscheinlichkeit ist am Anfang ein linearer Zusammenhang erkennbar. Bis zu einer Warenkorbgröße von circa 45 steigt die Tip-Wahrscheinlichkeit, während sie danach leichtfällt. Ab einer Warenkorbgröße von 70 sind starke Schwankungen erkennbar und einige Werte mit einer Tip-Wahrscheinlichkeit von 0% oder 100%. Das liegt vermutlich, daran dass es kaum Bestellungen mit so einer großen Anzahl an Produkten gibt und somit schon einzelne Bestellungen zu den Werten führen. Daher sind die Werte ab einer Warenkorbgröße von 70 nicht signifikant.



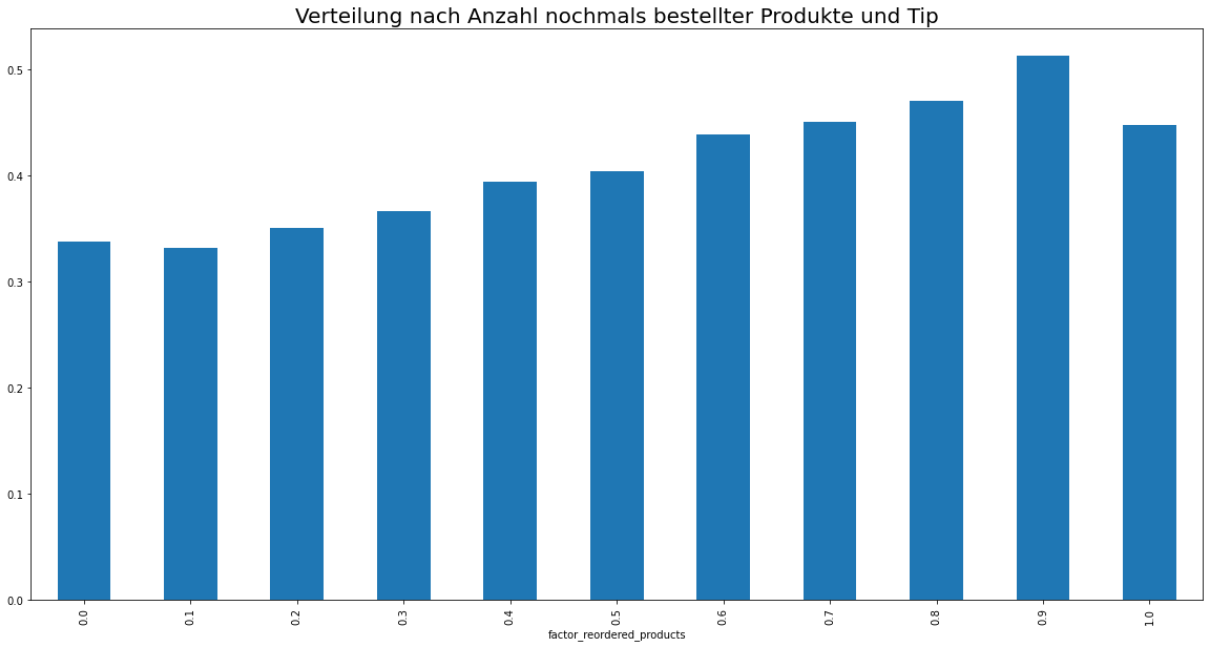
Erneut bestellte Produkte

Beim Vergleich der Anzahl an erneut bestellten Produkten zur Tip-Wahrscheinlichkeit ist am Anfang eine deutliche Steigerung erkennbar. Es lässt sich daraus schließen, dass je höher die Anzahl der erneut bestellten Produkte, desto höher ist die Tip-Wahrscheinlichkeit. Ab einem Wert von 45 erneut bestellten Produkten nimmt die Schwankung der Tip-Wahrscheinlichkeit wieder deutlich zu. Wie im Diagramm vorher sind diese Werte daher wohl nicht signifikant aufgrund von zu wenigen Bestellungen mit so vielen erneut bestellten Produkten.



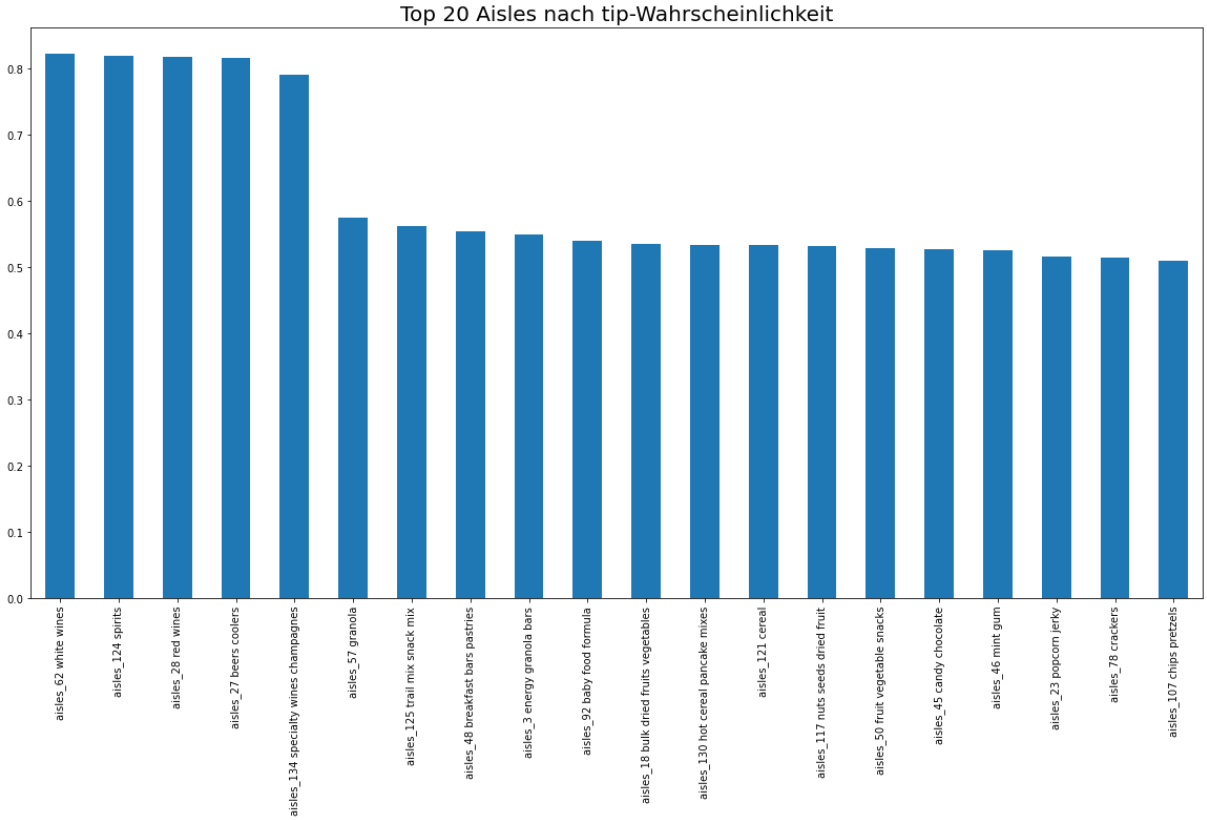
Anteil erneut bestellter Produkte

Bei der Betrachtung des Anteils erneut bestellter Produkte an allen Produkten einer Bestellung kann eine konstante Steigerung der Tip-Wahrscheinlichkeit festgestellt werden. Auffällig hoch ist die Tip Wahrscheinlichkeit bei einem Anteil von 90 % der erneut bestellten Produkte. Dagegen sinkt die Tip-Wahrscheinlichkeit bei einem 100-prozentigen Anteil wieder deutlich.



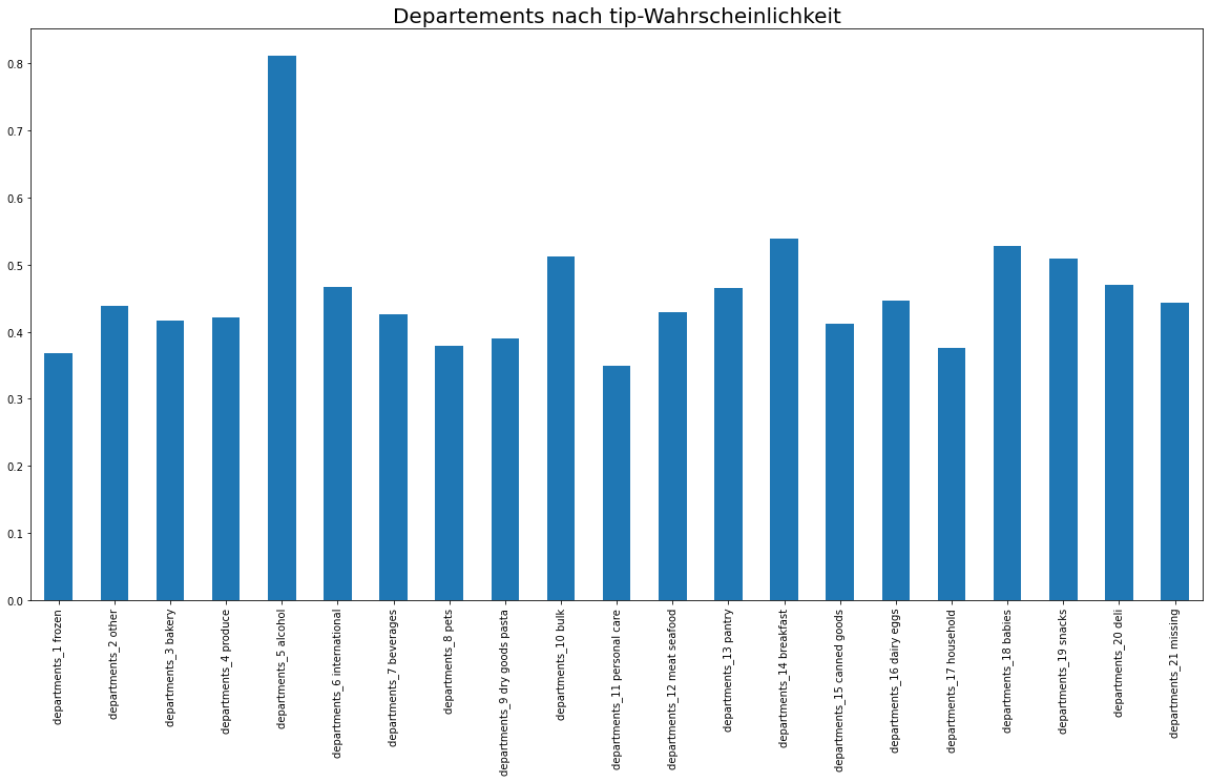
Aisles

Bei Betrachtung der der Tip-Wahrscheinlichkeit je Aisle wird eine klare Tendenz deutlich. Die Tip-Wahrscheinlichkeit der Aisles “ white wines”, “Spirits”, “red wines” und “speciality wines champagnes” sind mit über 80 % signifikant höher als die Tip-Wahrscheinlichkeit der anderen Aisles. Da es sich bei den vier genannten Aisles, ausschließlich um alkoholische Getränke handelt, wird vermutet, dass bei Bestellung von Alkohol die Tip-Wahrscheinlichkeit deutlich höher ist.



Departments

Diese Vermutung wird beim Blick auf die Tip-Wahrscheinlichkeit je Department bestätigt. Die Wahrscheinlichkeit für einen Tip, falls ein Produkt aus dem Department „alcohol“ in einer Bestellung enthalten war, ist signifikant höher als bei den anderen Departments. Ansonsten sind zwar Unterschiede zwischen den Departments erkennbar, es lassen sich jedoch keine relevanten Erkenntnisse daraus gewinnen.



Fazit

Abschließend wird festgehalten wovon es abhängig, ob ein Trinkgeld gegeben wird:

- Dem durchschnittlichen Trinkgeld des Users
- Dem durchschnittlichen Trinkgeld des Users in den letzten drei Bestellungen
- Ob bei der letzten Bestellung des Users Trinkgeld gegeben wurde
- Ob am Wochenende oder werktags bestellt wurde
- Ob tagsüber oder nachts bestellt wurde
- Dem Abstand zur letzten Bestellung
- Ob in wöchentlichen Rhythmen bestellt wird
- Den durchschnittlichen Bestellabständen des Users
- Der Bestellnummer
- Der Anzahl der bestellten Produkte
- Dem Anteil erneut bestellter Produkte
- Ob Alkohol mitbestellt wurde