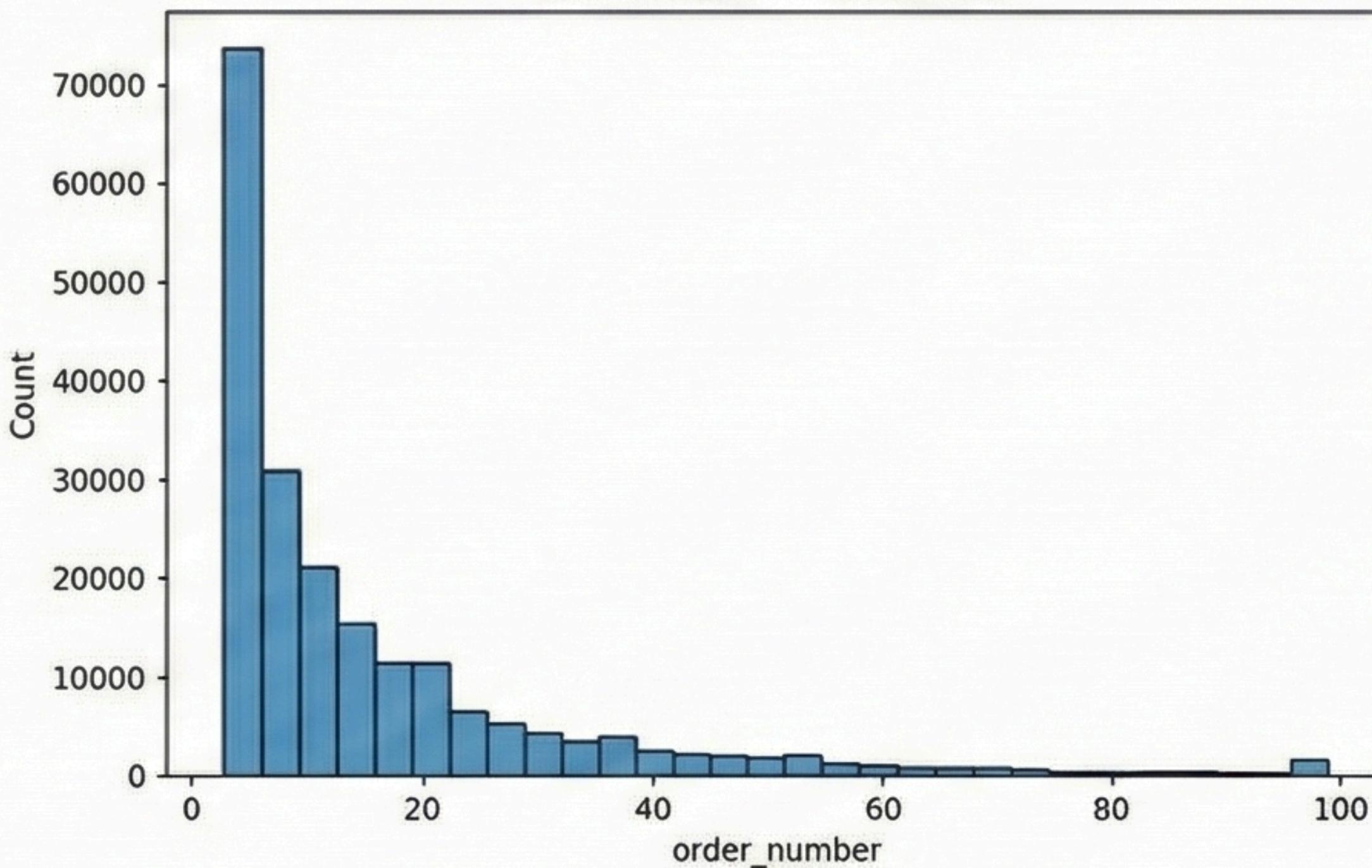


تحليل عدد الطلبات لكل مستخدم

توزيع تكرار الشراء بين العملاء

Number of Orders Per User



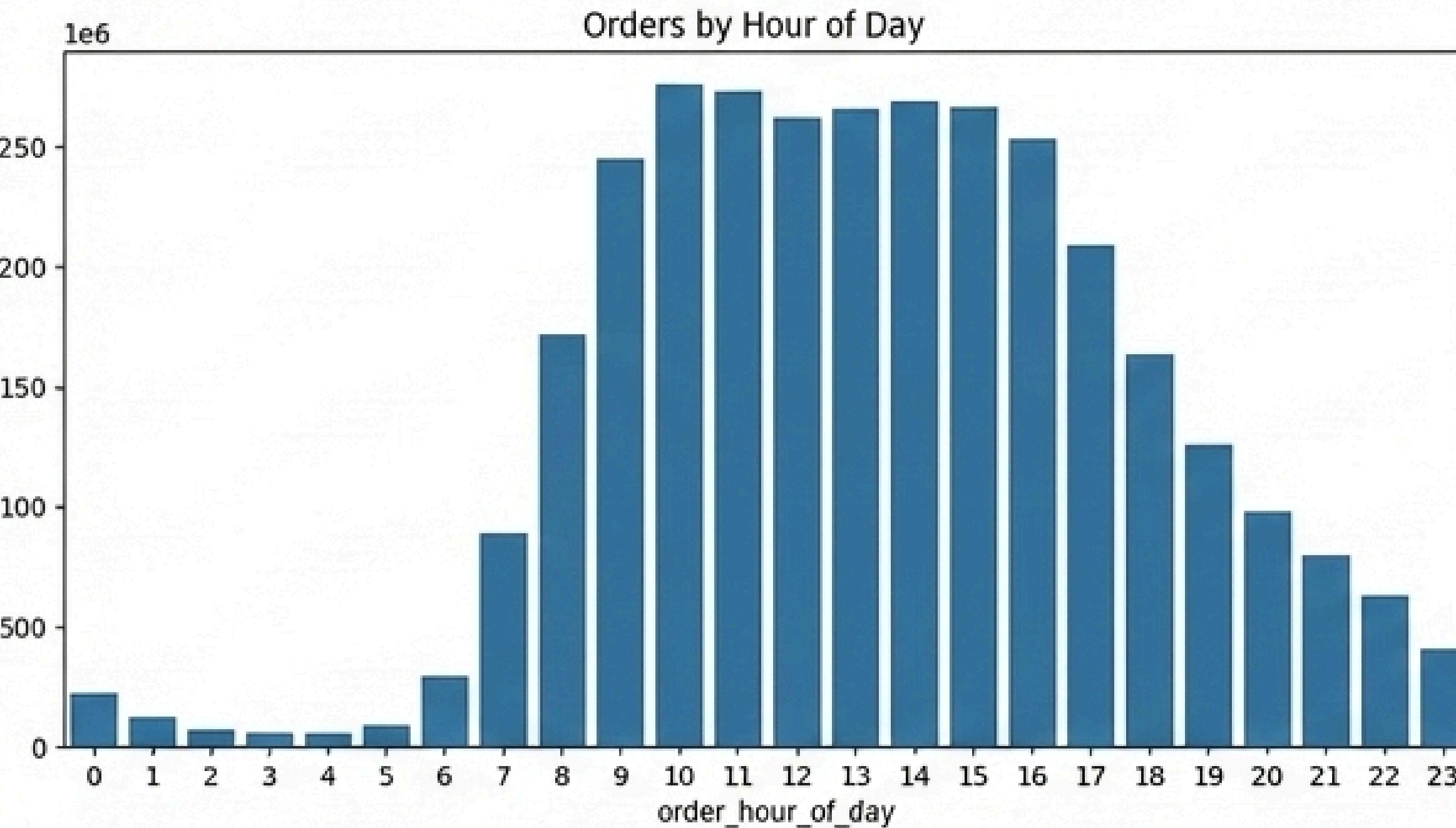
ماذا تخبرنا البيانات؟

يوضح الرسم البياني "هستوجرام" توزيع عدد الطلبات التي قام بها كل مستخدم. نلاحظ الأغلبية العظمى من المستخدمين قاموا بعمر قليل من الطلبات. القمة (أعلى عمود) تظهر عند عدد طلبات منخفض جداً (حوالي 4-5 طلبات)، حيث يتجاوز عدد هؤلاء المستخدمين 70,000 مستخدم. ينخفض عدد المستخدمين بشكل حاد وسريع كلما زاد عدد الطلبات. هناك "ذيل طويل" يمثل شريحة من العملاء الأوفيساء الذين قاموا بعدد كبير جداً من الطلبات (يصل إلى 100)، ولكن عددهم قليل مقارنة بإجمالي المستخدمين.

تحليل توقيت الطلبات (يومي وساعي)

نظرة على فترات الذروة ونمط الشراء الزمني للعملاء

تحليل الطلبات الساعي (Orders by Hour of Day)



ماذا تخبرنا البيانات؟

تبين البيانات أن ذروة نشاط الشراء تكون خلال ساعات النهار، وتتعدد بين الساعة 9 و4 عصراً (الساعات 9-16)، مع أعلى معدلات طلب حوالى الساعة 10 صباحاً و2 ظهراً. ينخفض النشاط بشكل كبير في ساعات الصباح الباكر (من منتصف الصباح حتى 6 صباحاً).

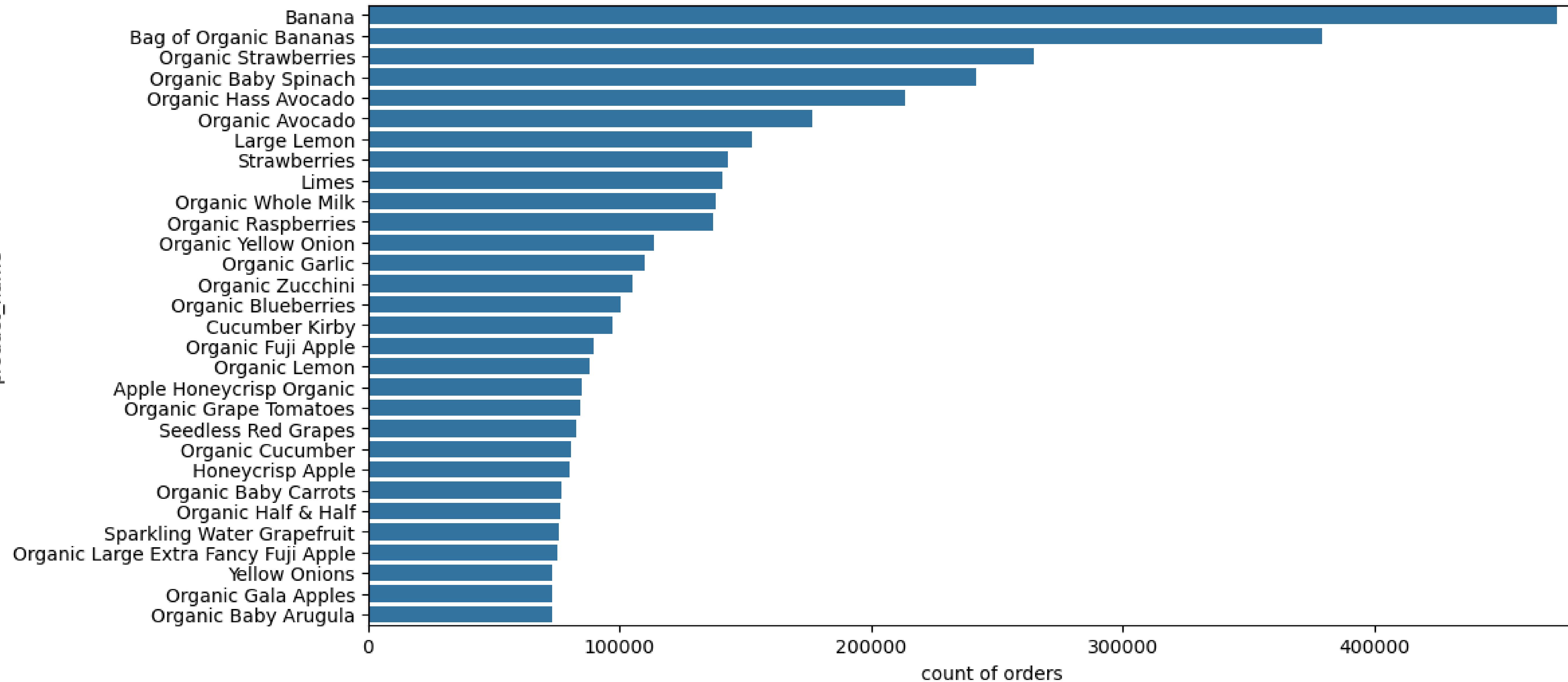
تحليل الطلبات اليومي (Orders by Day of Week)



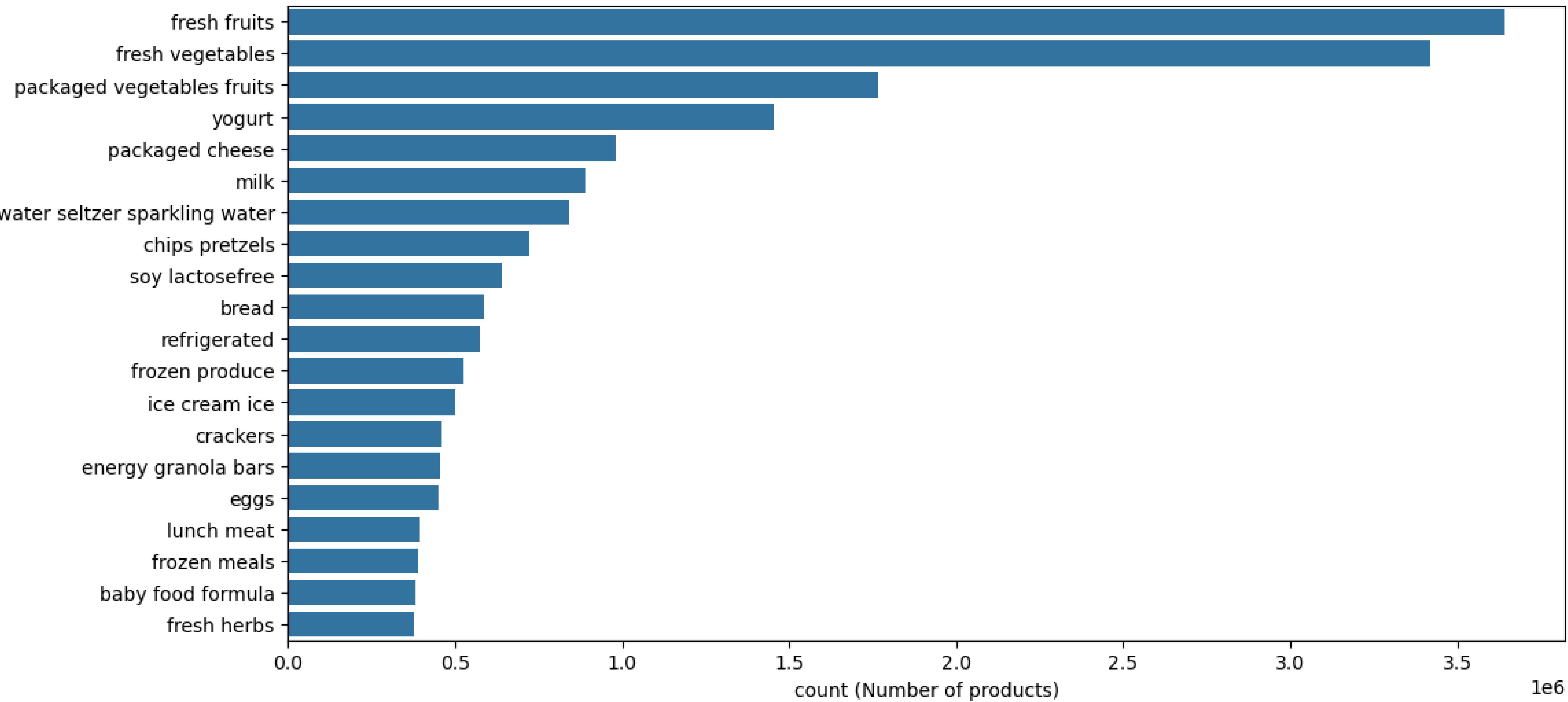
ماذا تخبرنا البيانات؟

يظهر الرسم البياني أن بداية الأسبوع (الأيام 0 و 1) هي الأنشط في عمليات عمليات الشراء ، مع انخفاض ملحوظ في منتصف الأسبوع (الأيام 3 و 4)، ثم ارتفاع طفيف في نهايته (الأيام 5 و 6).

Top 30 Product Ordred



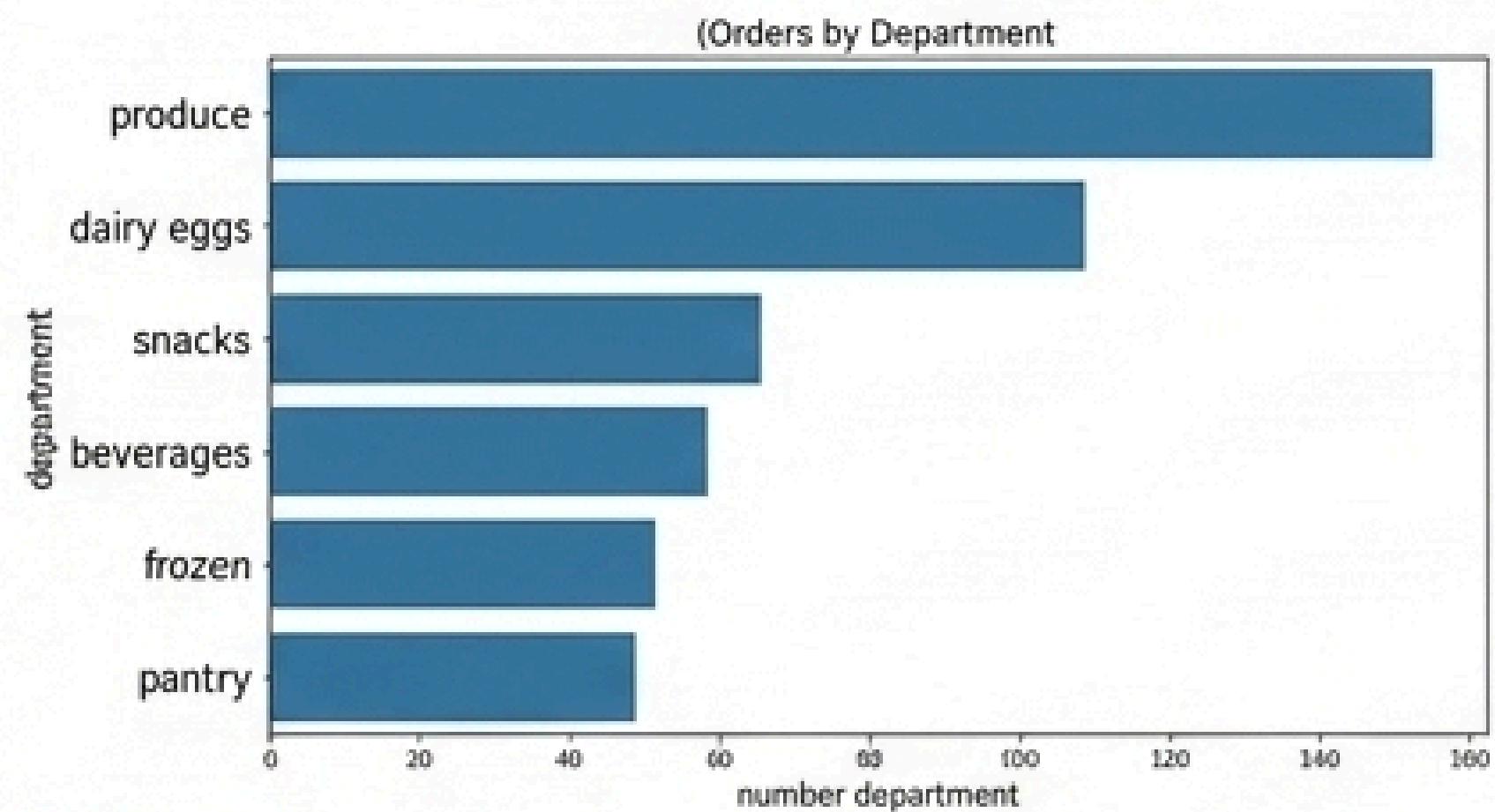
Top 20 aisle



تحليل المنتجات والممارسات الأكثر طلباً

نظرة على تفضيلات العملاء وتوزيع الطلبات حسب الفئات والممرات

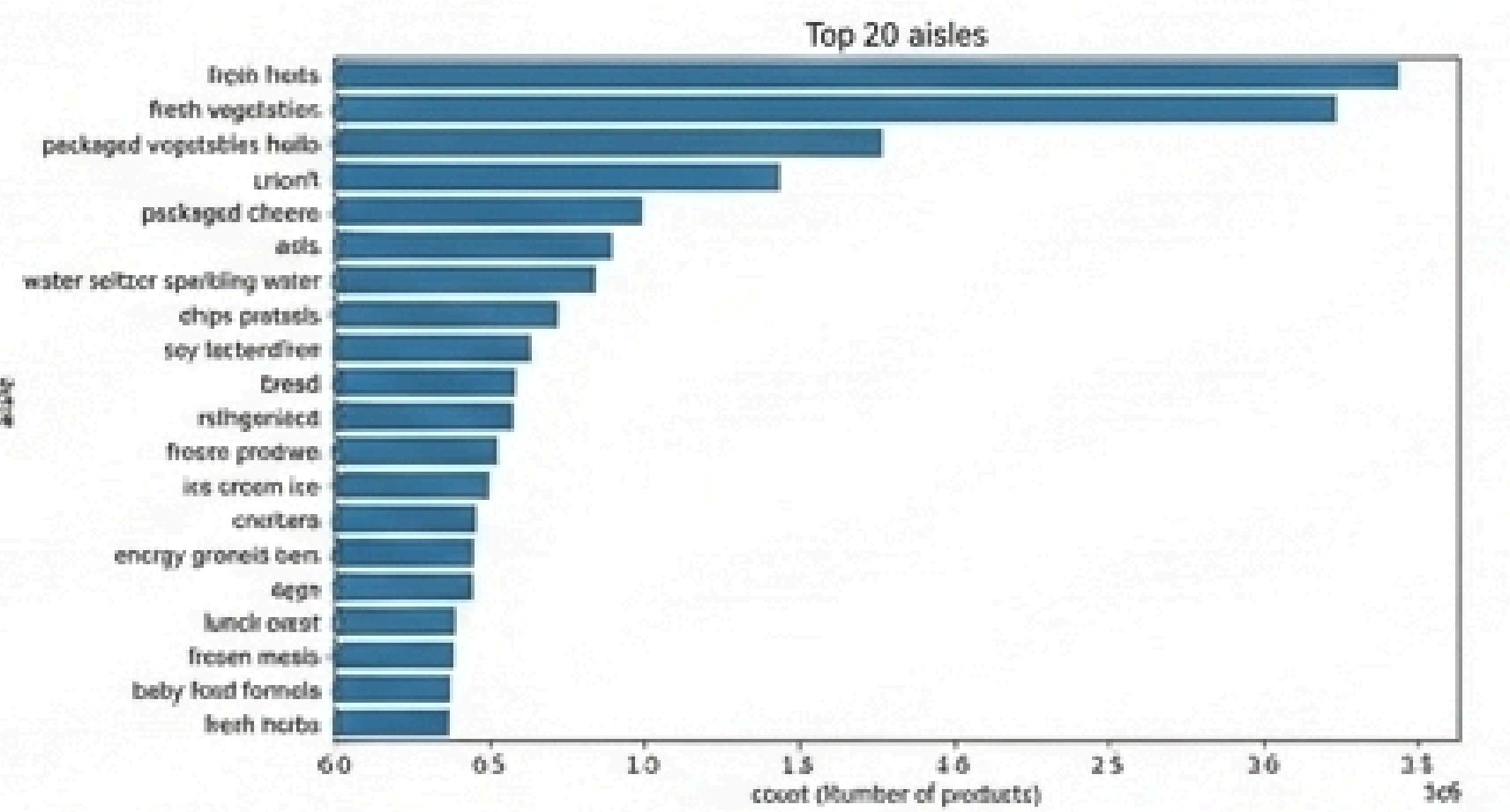
توزيع الطلبات حسب القسم (افتراضي) (Orders by Department)



ماذا تخبرنا البيانات؟

يتواافق توزيع الطلبات حسب القسم مع تحليل المنتجات والممارسات، حيث يحتل قسم 'المنتجات الطازجة' (Produce) المرتبة الأولى بلا منازع، يليه قسم 'الألبان والبيض'. هذا يعزز الاستنتاج بأن التركيز الأكبر للعملاء ينصب على المواد الغذائية الأساسية والطازجة.

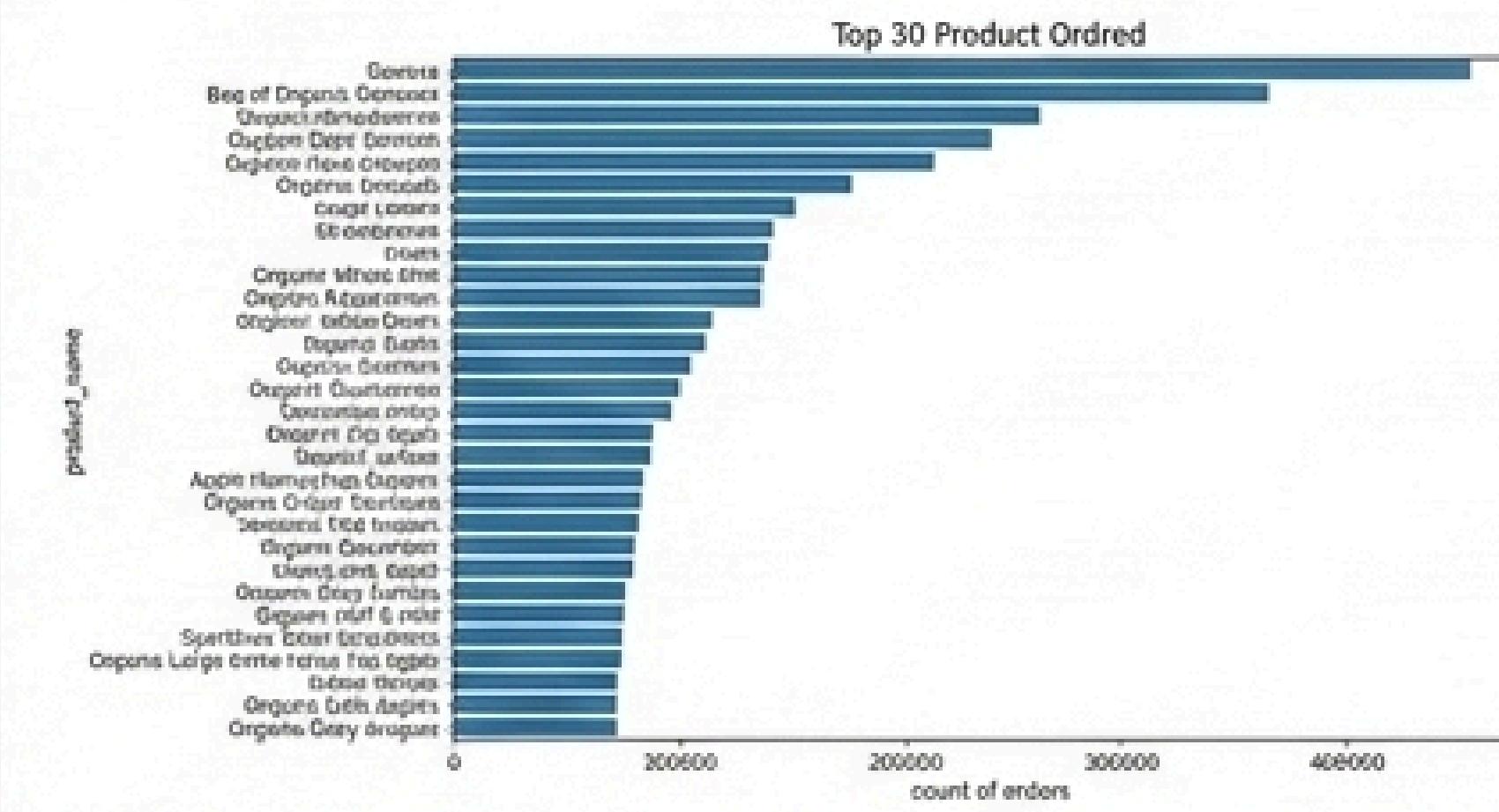
أكثـر 20 مـمراً نـشـاطـاً (Top 20 Aisles)



ماذا تخبرنا البيانات؟

تصدر ممرات الفواكه والخضروات الطازجة القائمة بفارق كبير، تليها الخضروات والفواكه المعبأة المعبأة، ومنتجات الألبان (الزيادي، الجبن، الحليب). هذا يؤكد الأهمية القصوى للمنتجات الطازجة الأساسية في سلة مشتريات العملاء.

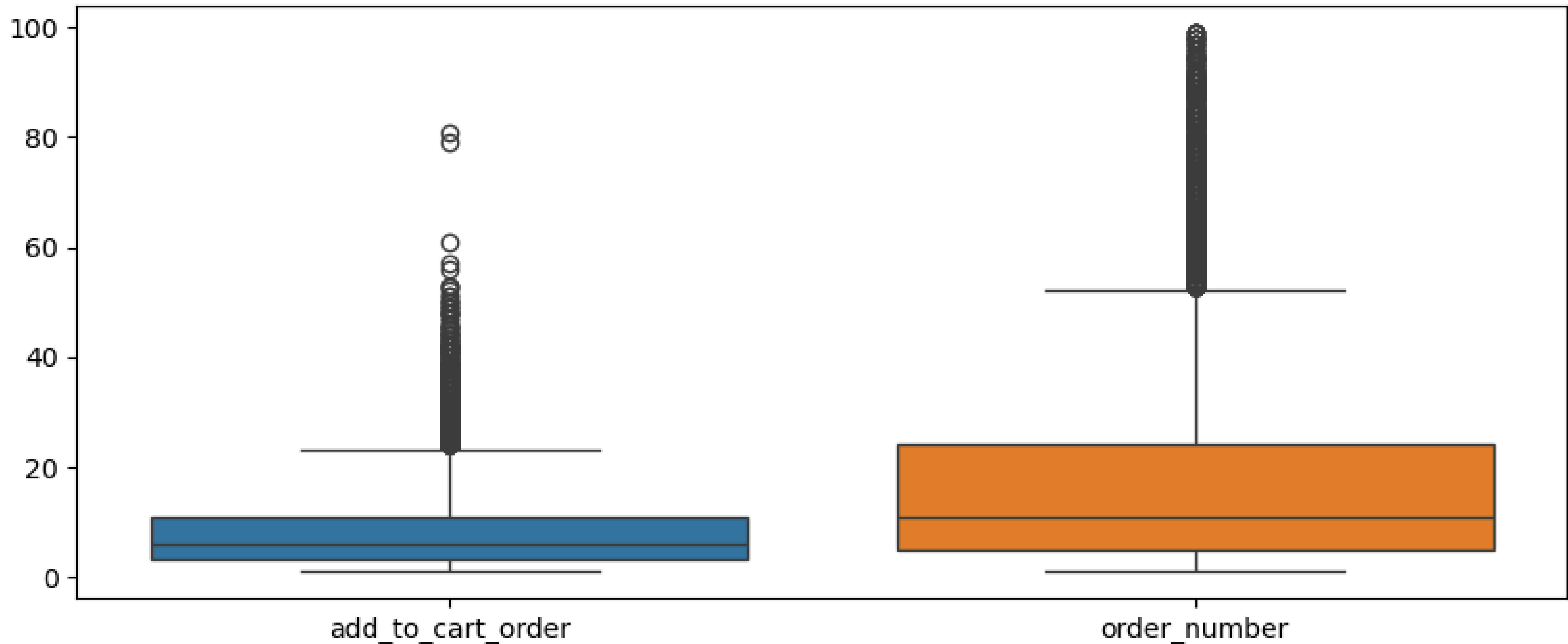
أكثـر 30 منتجـاً طلـباً (Top 30 Products)



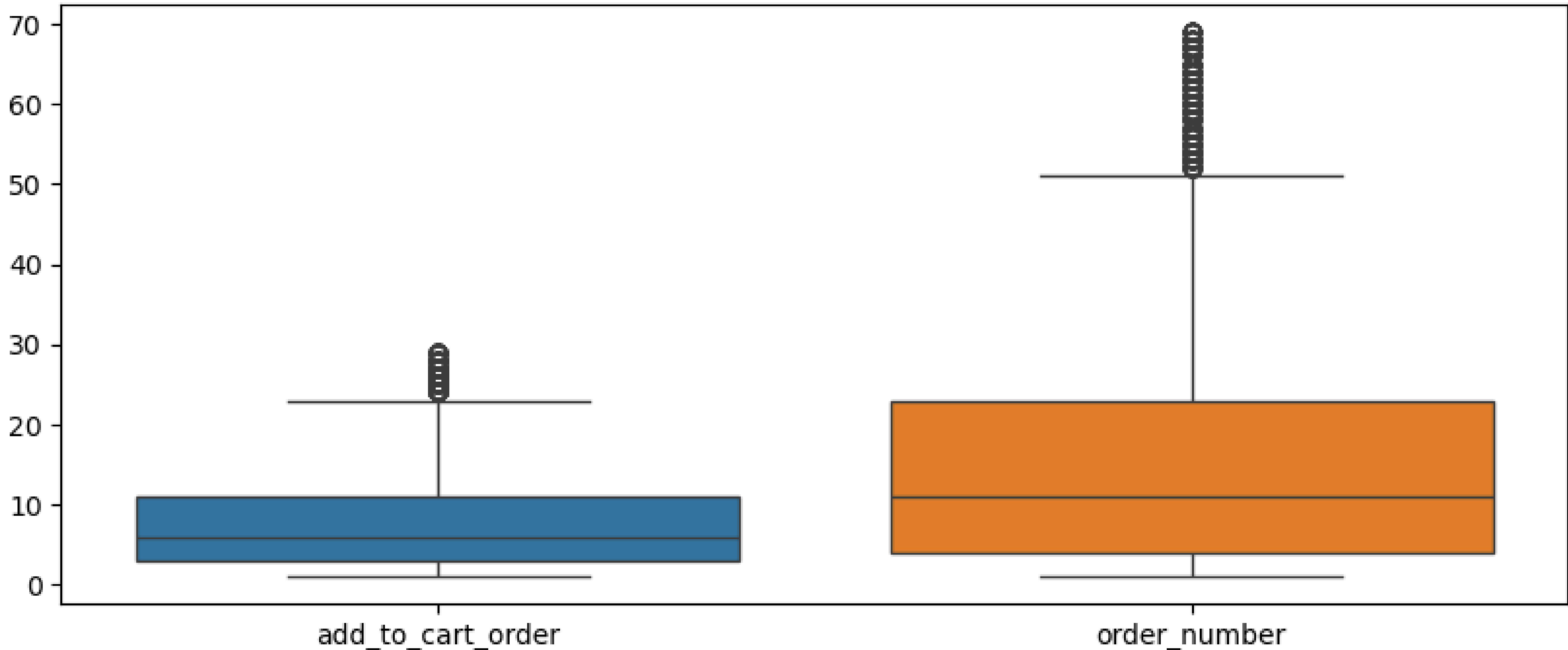
ماذا تخبرنا البيانات؟

تظهر البيانات هيمنة واضحة للمنتجات الطازجة والعضوية على رأس القائمة، وخاصة الموز بنوعيه، والفراولة، والسبانخ الصغيرة. هذا يعكس توجهاً قوياً لدى العملاء نحو الخيارات الصحية والأساسية.

Boxplot Before Removing Outliers (Z-score)



Boxplot After Removing Outliers (Z-score)



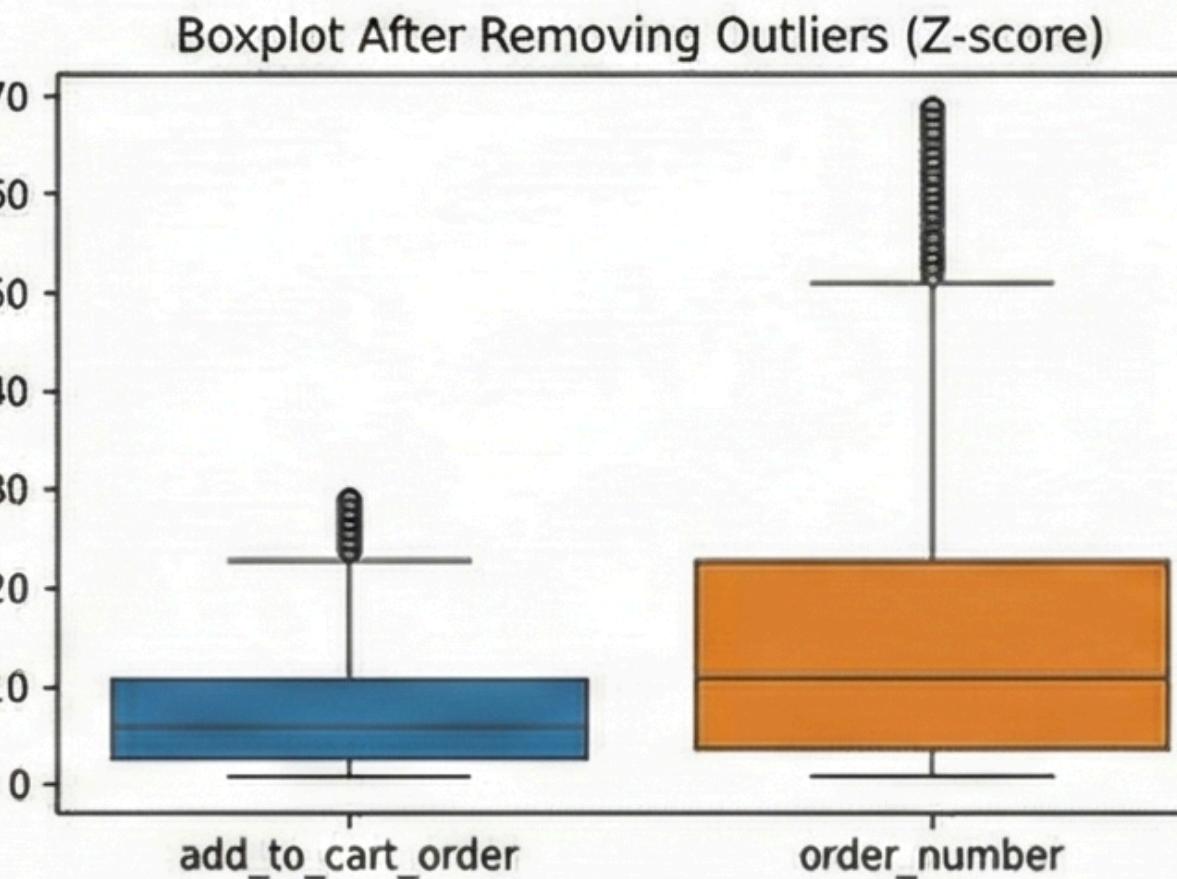
معالجة القيم المتطرفة (Outlier Treatment)

تأثير إزالة القيم الشاذة باستخدام طريقة Z-score على توزيع المتغيرات

بعد إزالة القيم المتطرفة (Z-score)

ماذا تخبرنا البيانات؟

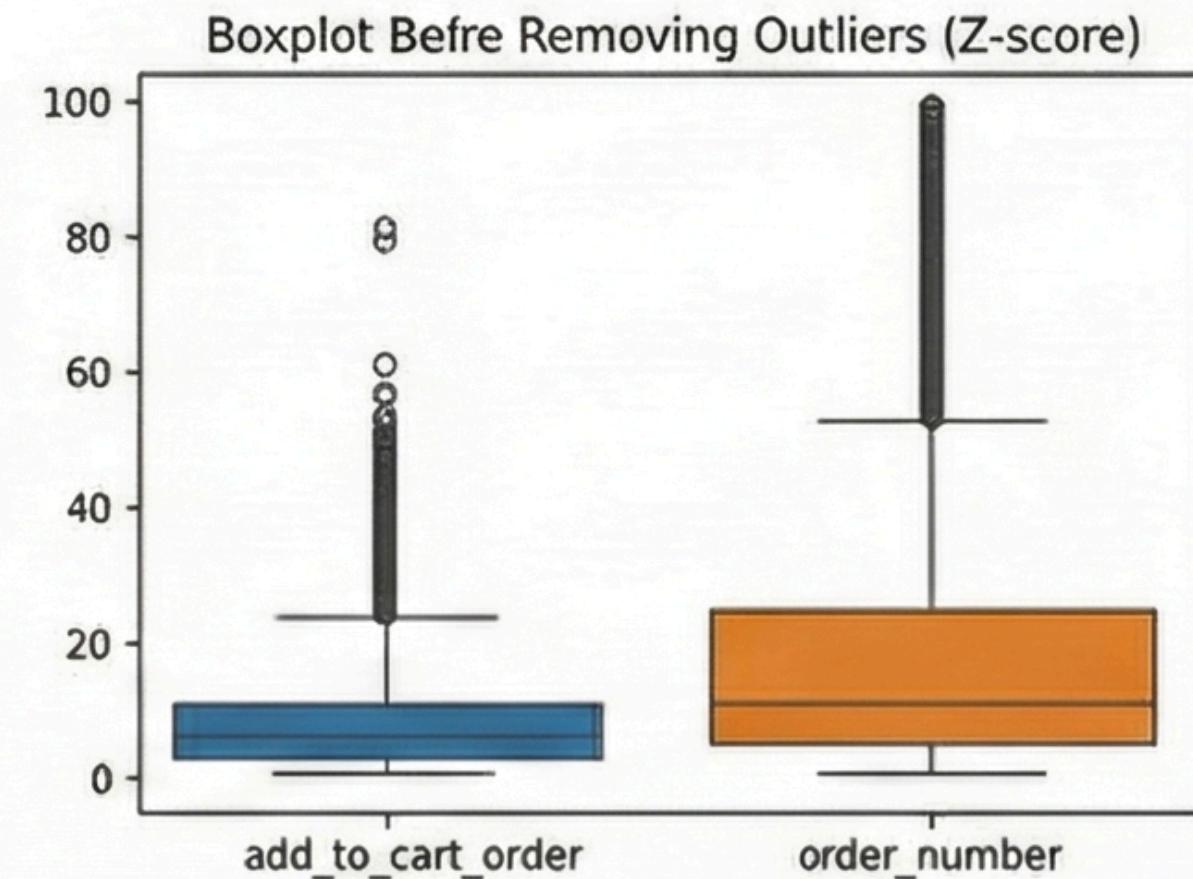
بعد تطبيق طريقة Z-score، تم تقليل عدد القيم المتطرفة بشكل كبير. أصبحت القيم القصوى لـ `add_to_cart_order` عند 30 وـ `order_number` عند 70 تقريباً. هذا يؤدي إلى توزيع بيانات أكثر نظافة وموثوقية لبناء النماذج.



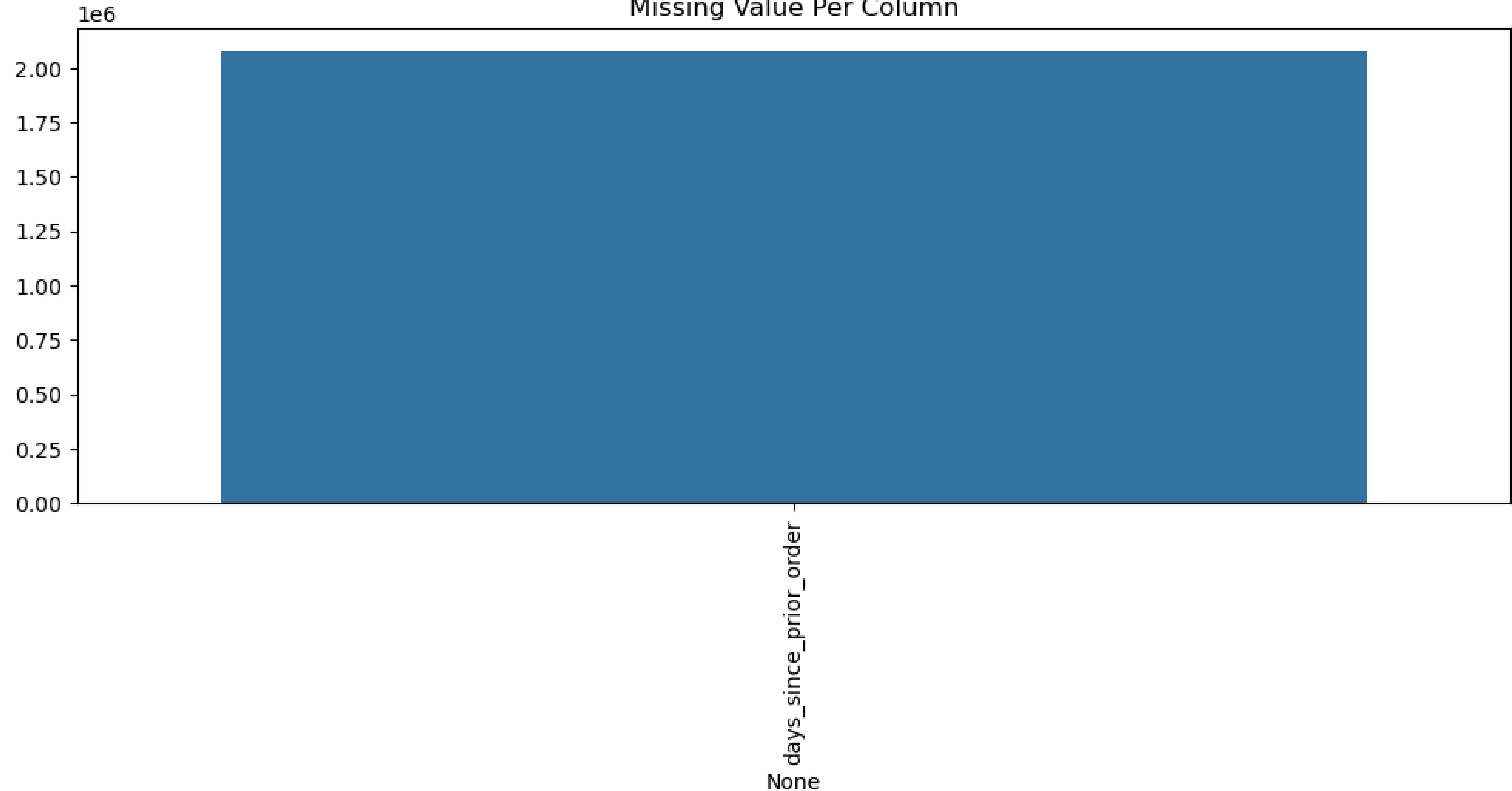
قبل إزالة القيم المتطرفة (Z-score)

ماذا تخبرنا البيانات؟

يوضح المخطط الصندوقي قبل المعالجة وجود عدد كبير من القيم المتطرفة (النقاط الفردية خارج "الشوارب"). في `add_to_cart_order`، تصل القيم إلى 80، بينما في `order_number`، تصل إلى 100، مما يشير إلى وجود ملاحظات شاذة قد تؤثر على التحليل الإحصائي.



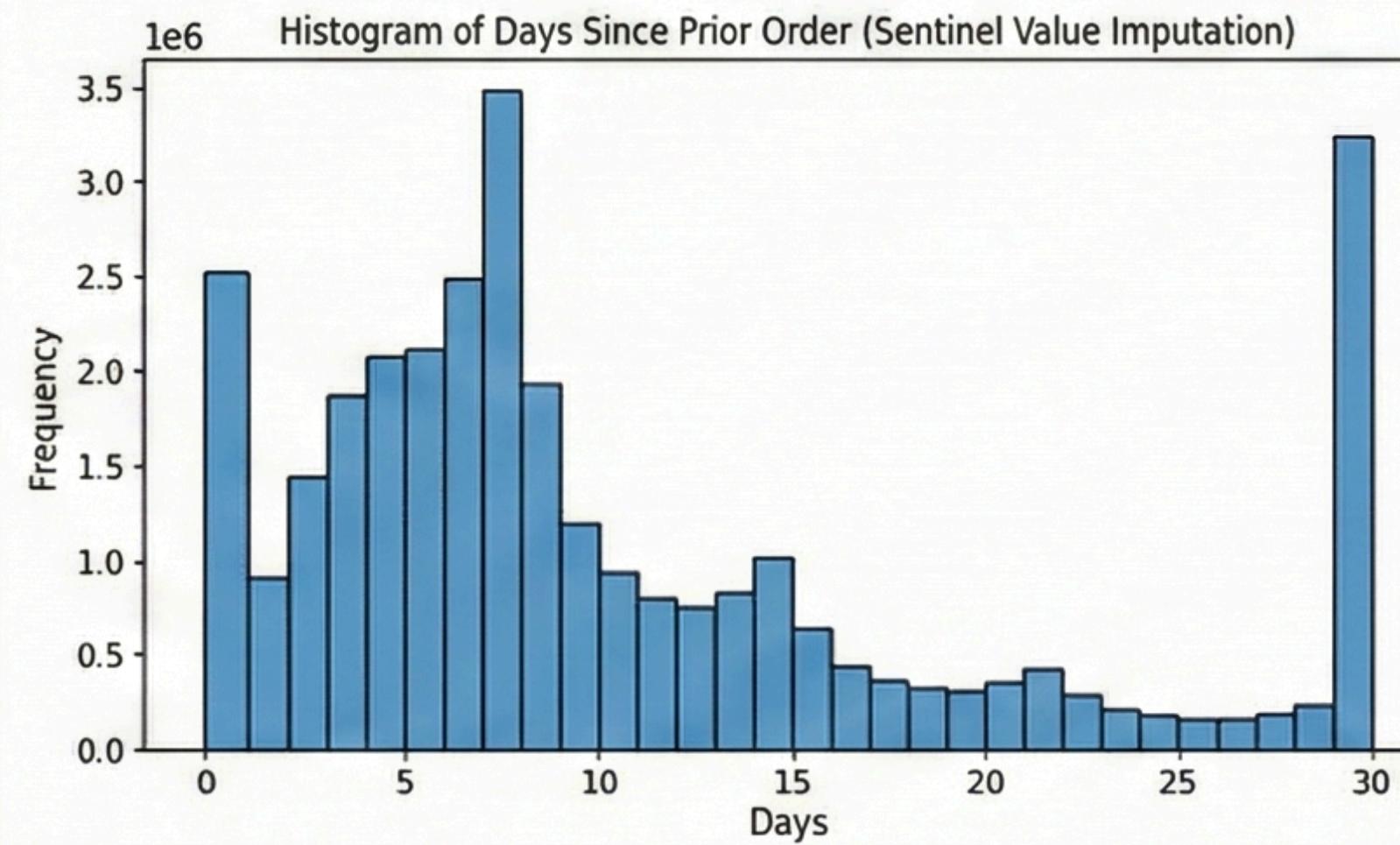
Missing Value Per Column



مقارنة طرق معالجة القيم المفقودة للفترة بين الطلبات

تحليل تأثير الطرق المختلفة على توزيع البيانات و اختيار الطريقة الأنسب منطقياً

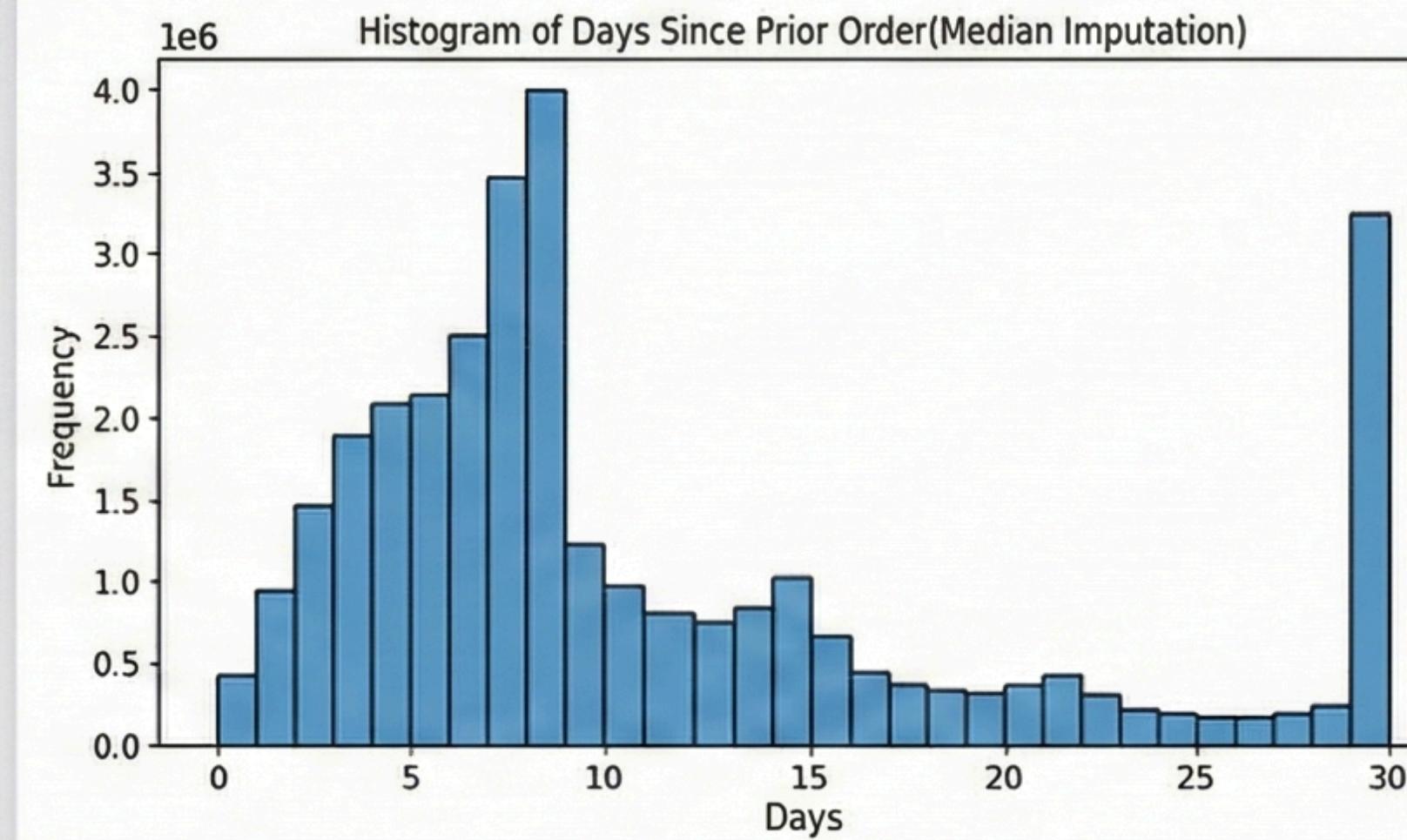
ملء بالقيمة الثابتة 0 (Sentinel Value - 0) ★ الخيار الأنسب



ماذا تخبرنا البيانات؟

اختيار القيمة الثابتة 0 لملء القيم المفقودة هو الأنسب والأكثر منطقية. هذا العمود الكبير عند الصفر يمثل بدقة 'الطلب الأول' للعميل (حيث لا توجد فترة سابقة). بالإضافة إلى ذلك، نرى الأنماط المعتادة والممثلة بشكل صحيح لإعادة الطلب الأسبوعي (7 أيام) والشهري (30 يوماً).

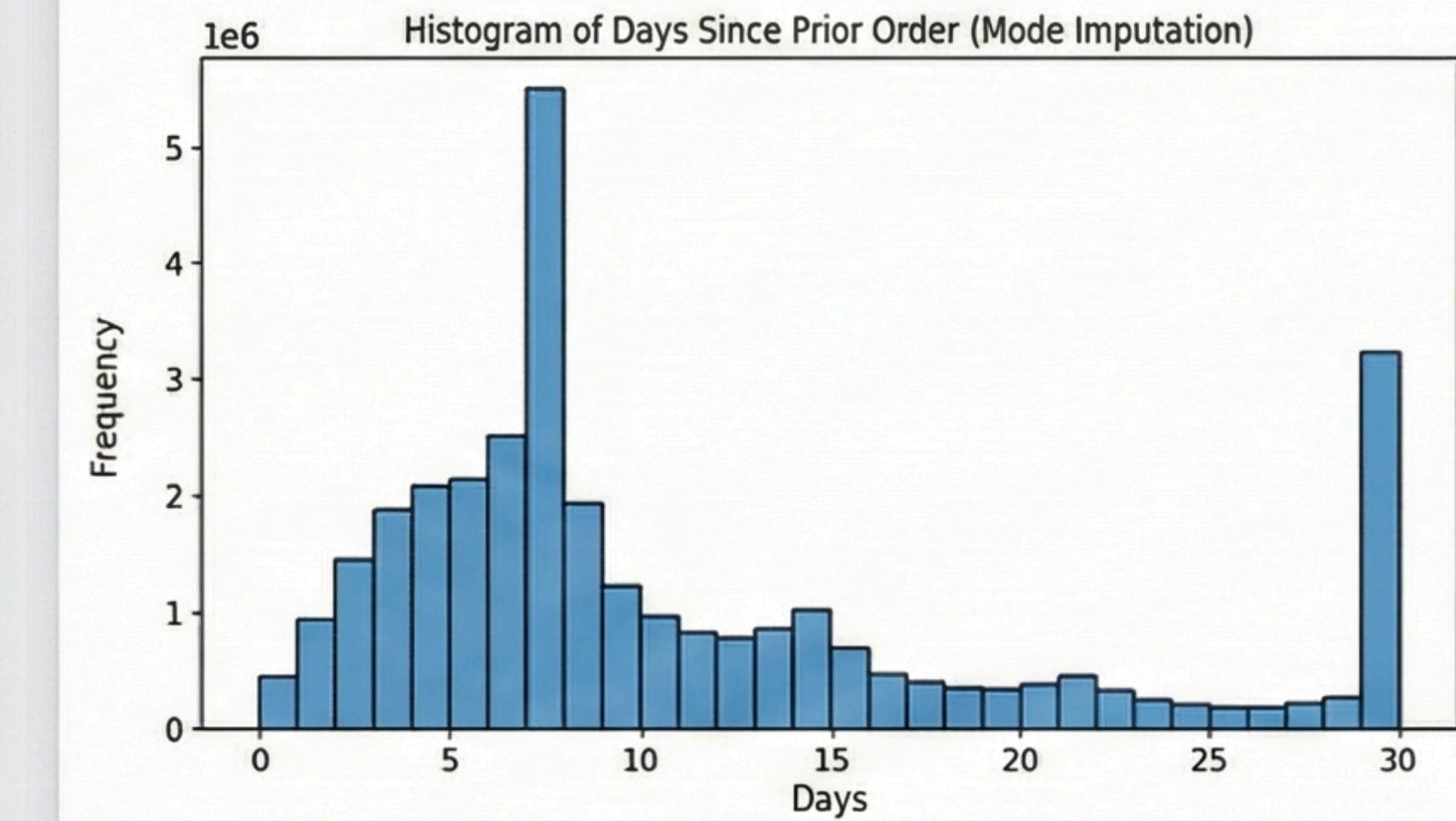
طريقة بالوسط (Median Imputation)



ماذا تخبرنا البيانات؟

استخدام القيمة المتوسطة (الوسط) يعطي توزيعاً مشابهاً للمنوال، مع قمم واضحة عند 7 و 30 يوماً، لكن القمة عند 7 أيام تبدو تبديلاً تدريجياً أقل حدة قليلاً، مما قد يكون أكثر تمثيلاً للوسط الحقيقي للبيانات مقارنة بالمنوال.

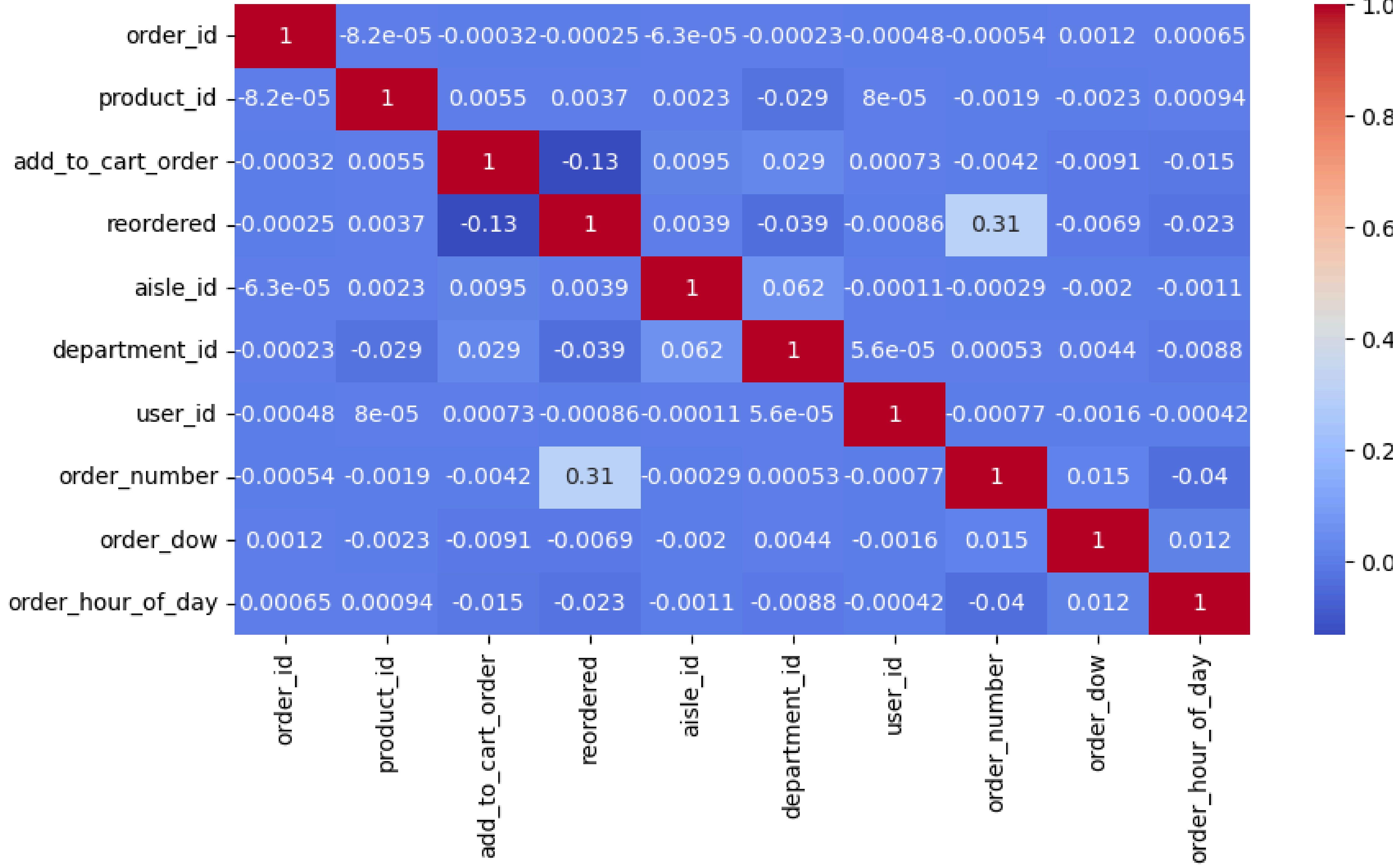
طريقة بالمنوال (Mode Imputation)



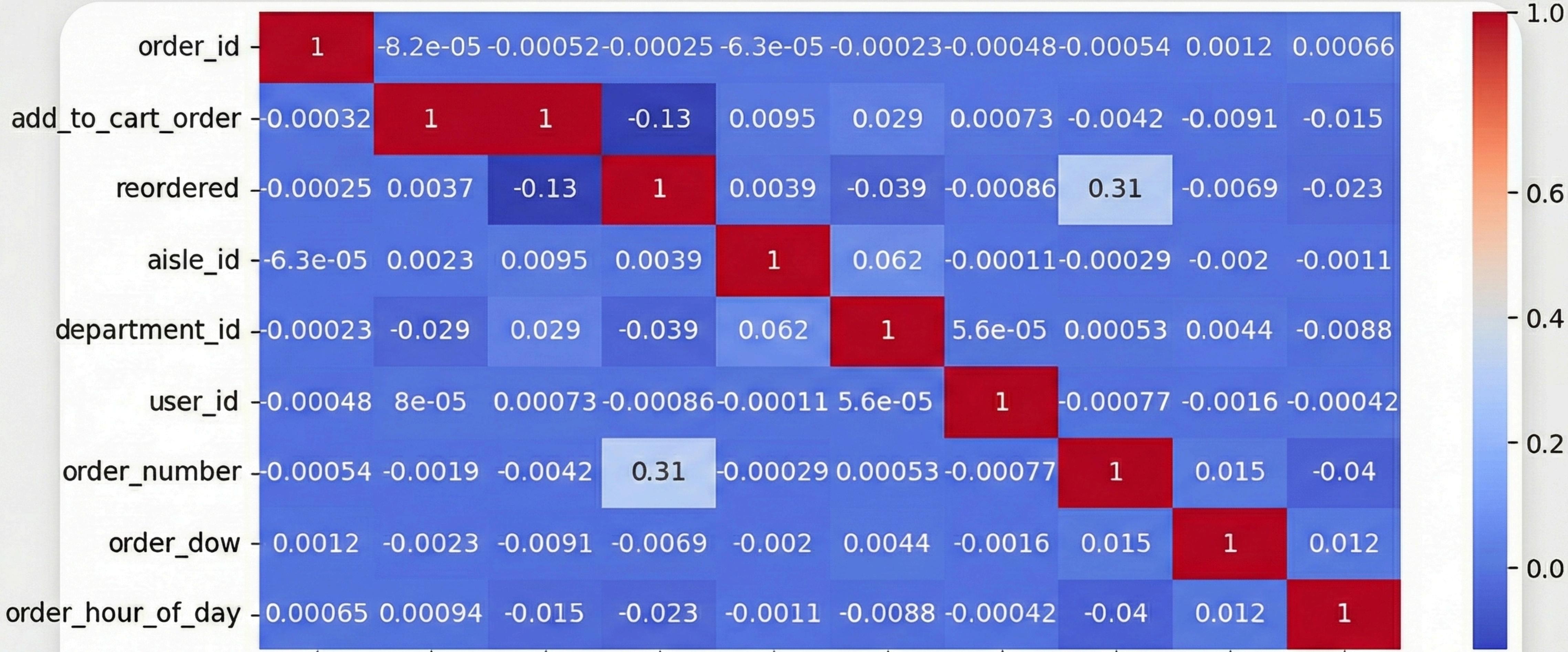
ماذا تخبرنا البيانات؟

استخدام القيمة الأكثر تكراراً (المنوال) يُظهر قمة حادة جداً وغير طبيعية عند 7 أيام، مما قد يؤدي إلى تحيز في التحليل والمبالغة في تقدير تكرار هذه الفترة الزمنية المحددة.

Numerical Feature Correlation



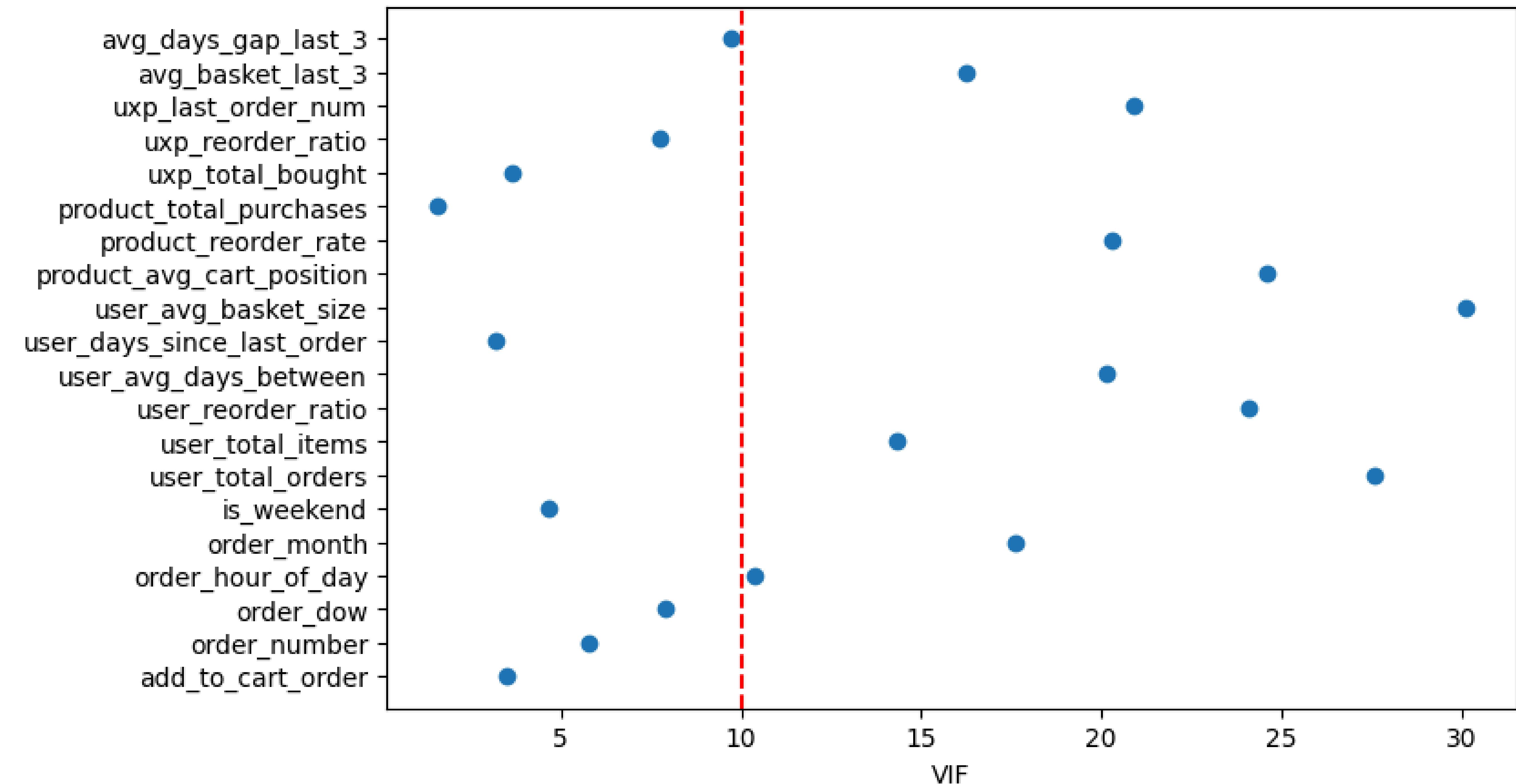
تحليل نسبة إعادة الطلب وعدم إعادته



ماذا تخبرنا البيانات؟

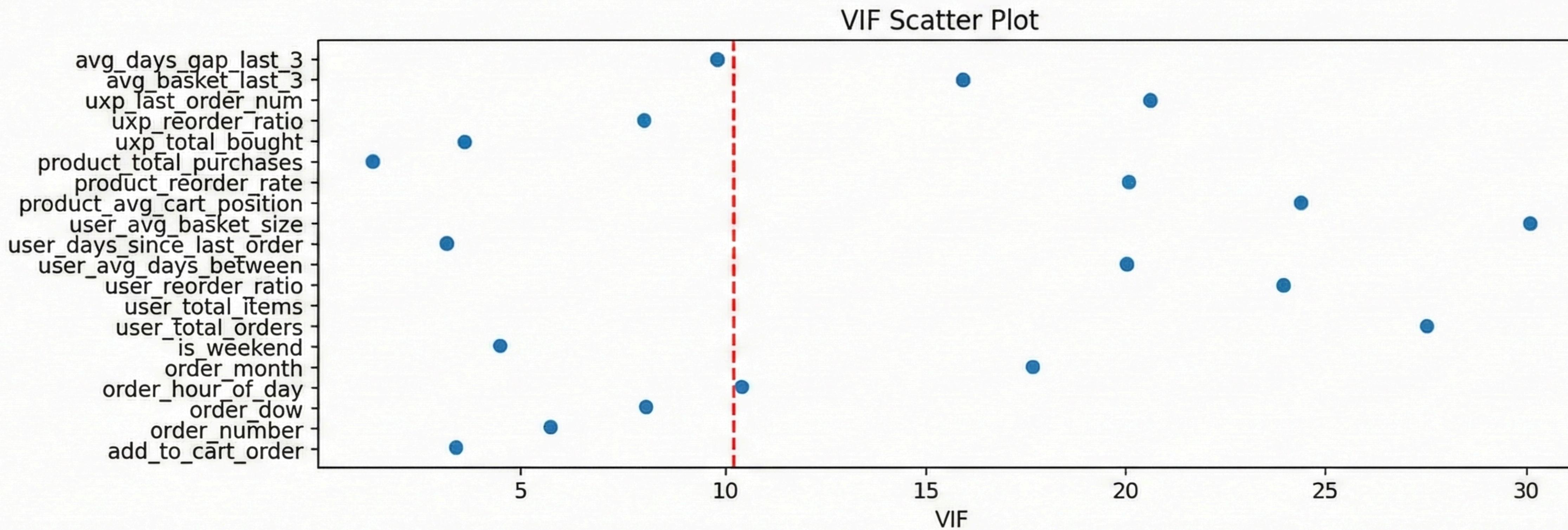
يوضح هذا الرسم البياني العلاقة القوية بين ميزة 'reordered' (إعادة الطلب) وميزة 'order_number' (رقم الطلب). تشير القيمة العالية (1) باللون الأحمر الداكن إلى وجود ارتباط تام في حالات إعادة الطلب، بينما تشير القيمة المنخفضة (0.31) باللون الأزرق الفاتح إلى نسبة عدم إعادة الطلب. باقي المتغيرات تظهر ارتباطات ضعيفة جدًا.

VIF Scatter Plot



تحليل التعددية الخطية (VIF)

تقييم ارتباط المتغيرات المستقلة للكشف عن التعددية الخطية التي قد تؤثر على دقة النماذج



ماذا تخبرنا البيانات؟

يوضح الرسم البياني قيمة عامل تضخم البيانات (VIF) لكل متغير مستقل. النقاط التي تقع على يمين الخط الأحمر المقطوع ($VIF > 10$) تشير إلى وجود تعددية خطية عالية قد تؤثر سلباً على نموذج الانحدار، مثل `user_avg_basket_size` و `product_reordered_ratio` و `order_dow` بينما المتغيرات على الـ اليسار ($VIF < 10$) تعتبر ذات ارتباط مقبول ولا تمثل مشكلة كبيرة، مثل `uxp_reordered_ratio` و `product_total_purchases` و `is_weekend`.