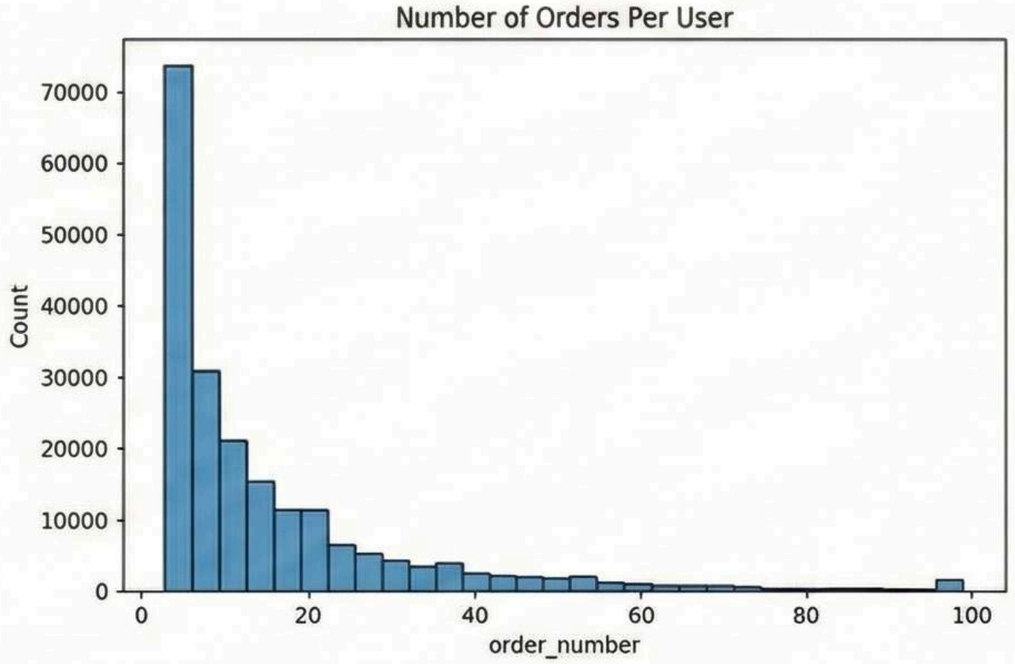
تحليل عدد الطلبات لكل مستخدم

توزيع تكرار الشراء بين العملاء



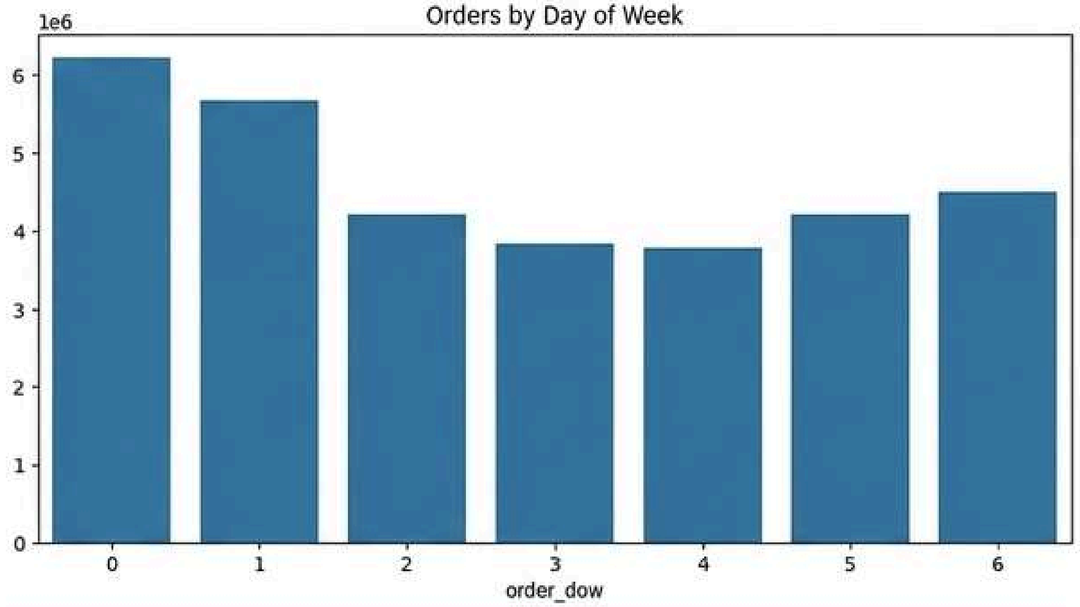
🔍 ماذا تخبرنا البيانات؟

يوضح الرسم البياني "هستوجرام" توزيع عدد الطلبات التي قام بها كل مستخدم. نلاحظ الأغلبية العظمى من المستخدمين قاموا بعدد 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 الأوفياء 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

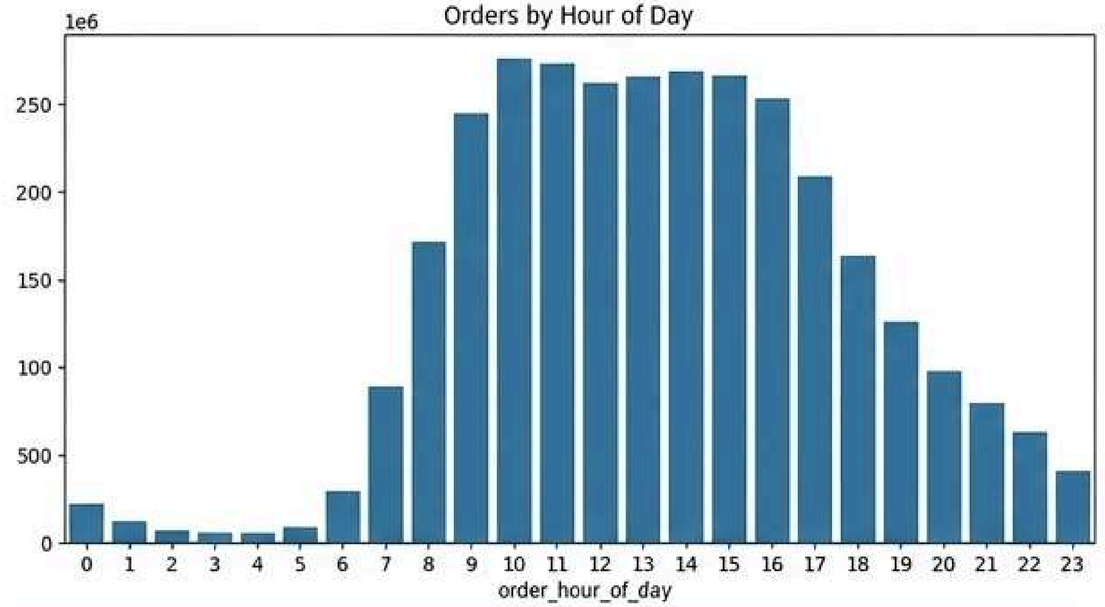
تحليل الطلبات اليومي (Orders by Day of Week)



🔍 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).

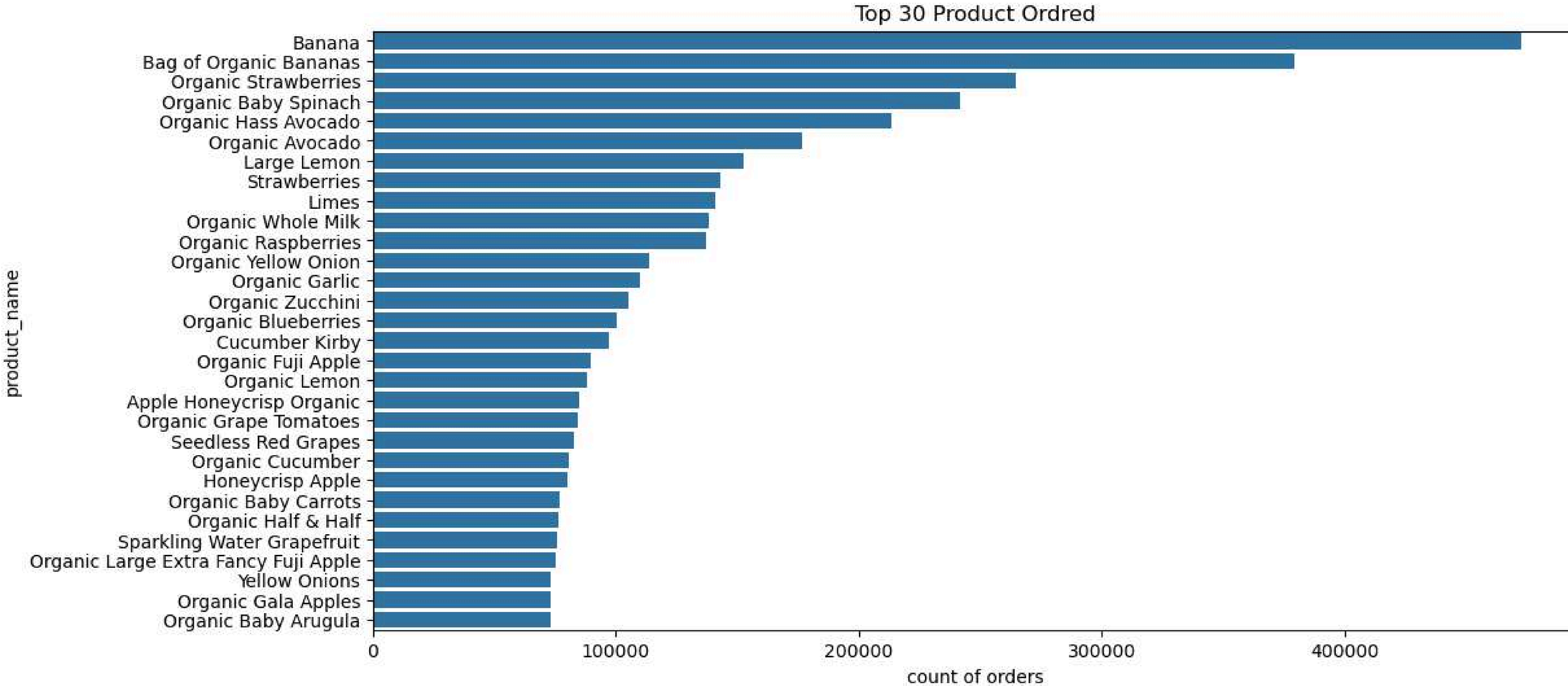
تحليل الطلبات الساعي (Orders by Hour of Day)

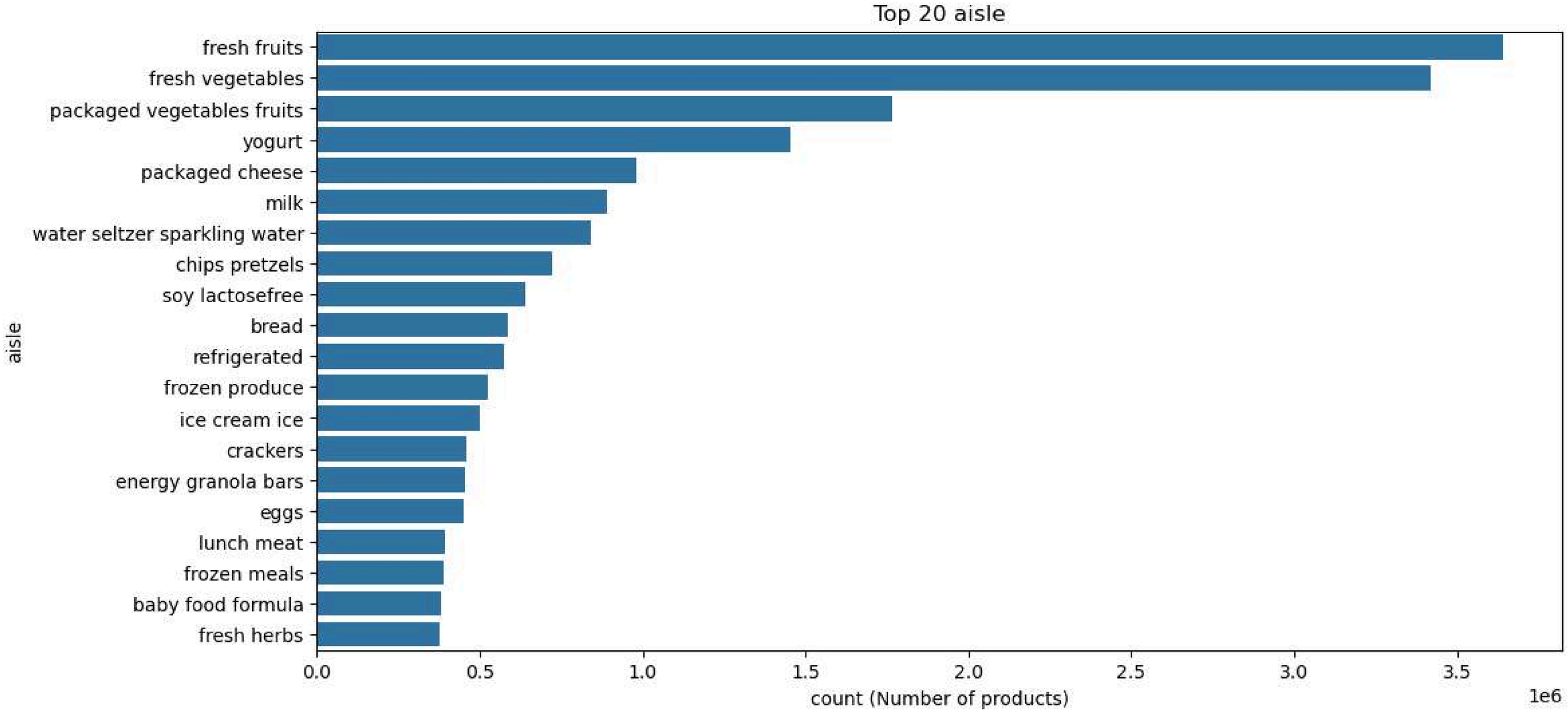


🔍 ماذا تخبرنا البيانات؟

تبين البيانات أن ذروة نشاط الشراء تكون خلال ساعات النهار، وتحديداً بين الساعة 9 و4 عصراً (الساعات 9-16)، مع أعلى معدلات طلب حوالي الساعة 10 صباحاً و 2 ظهراً.

ينخفض النشاط بشكل كبير في ساعات الصباح الباكر (من منتصف الليل حتى 6 صباحاً).

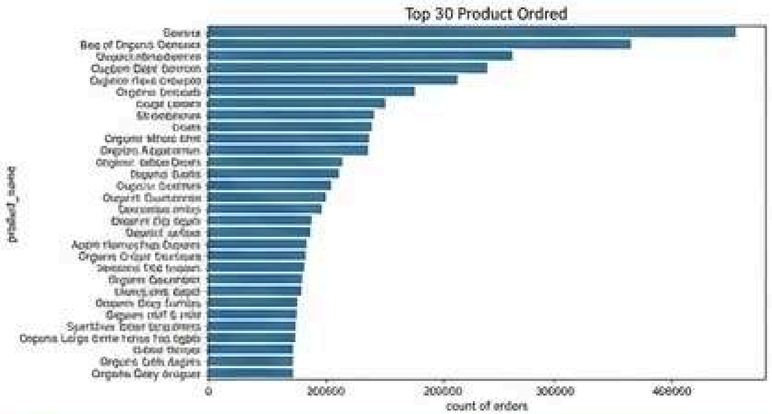

Top 30 Product Ordred
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Organic Hass Avocado
Organic Avocado
Large Lemon
Strawberries
Limes
Organic Whole Milk
Organic Raspberries
Organic Yellow Onion
Organic Garlic
Organic Zucchini
Organic Blueberries
Cucumber Kirby
Organic Fuji Apple
Organic Lemon
Apple Honeycrisp Organic
Organic Grape Tomatoes
Seedless Red Grapes
Organic Cucumber
Honeycrisp Apple
Organic Baby Carrots
Organic Half & Half
Sparkling Water Grapefruit
Organic Large Extra Fancy Fuji Apple
Yellow Onions
Organic Gala Apples
Organic Baby Arugula
product_name
0
100000
200000
300000
400000
count of orders


Top 20 aisle
fresh fruits
fresh vegetables
packaged vegetables fruits
yogurt
packaged cheese
milk
water seltzer sparkling water
chips pretzels
soy lactosefree
bread
refrigerated
frozen produce
ice cream ice
crackers
energy granola bars
eggs
lunch meat
frozen meals
baby food formula
fresh herbs
aisle
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e6
count (Number of products)

تحليل المنتجات والممرات الأكثر طلباً

نظرة على تفضيلات العملاء وتوزيع الطلبات حسب الفئات والممرات

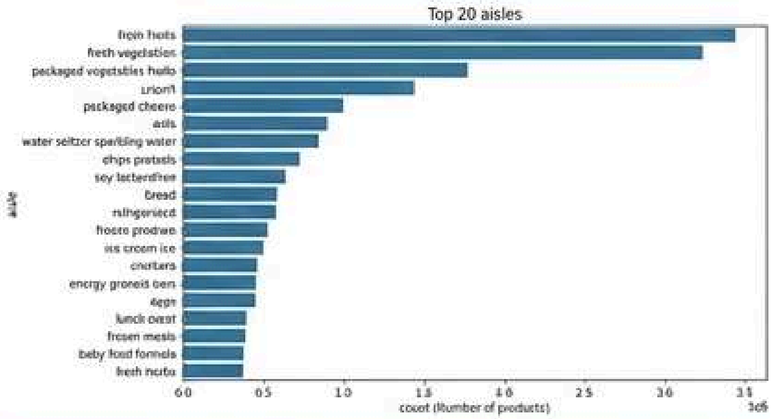
أكثر 30 منتجاً طلباً (Top 30 Products)



🔍 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.

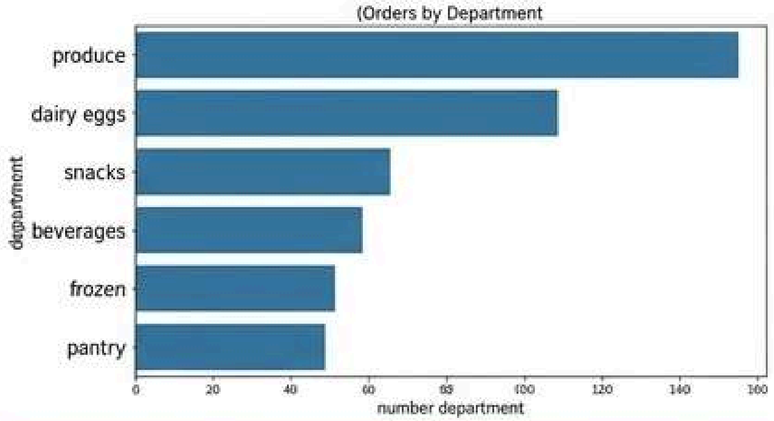
أكثر 20 ممراً نشاطاً (Top 20 Aisles)



🔍 ماذا تخبرنا البيانات؟

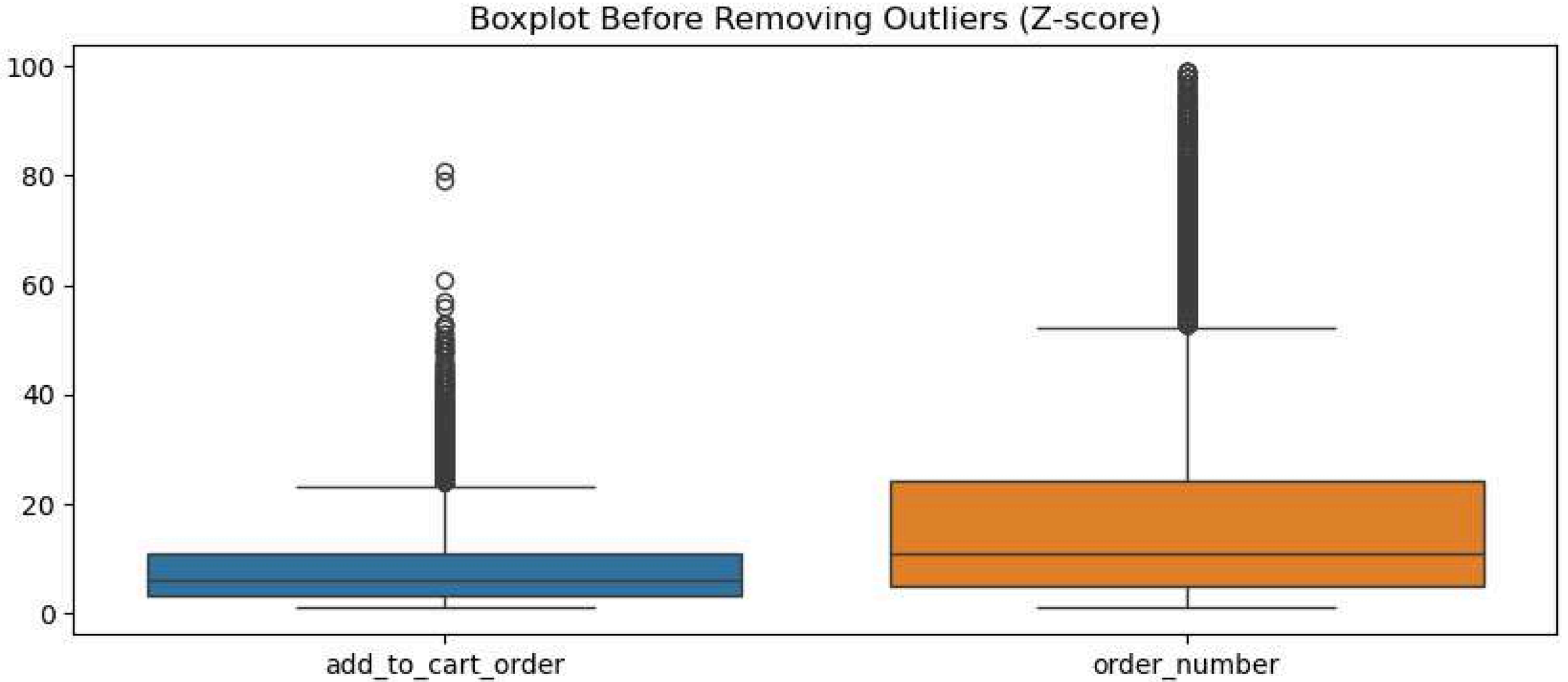
تتصدر ممرات الفواكه والخضروات الطازجة القائمة بفارق كبير، تليها الخضروات والفواكه المعبأة المعبأة، ومنتجات الألبان (الزبادي، الجبن، الحليب). هذا يؤكد الأهمية القصوى للمنتجات الطازجة والأساسية في سلة مشتريات العملاء.

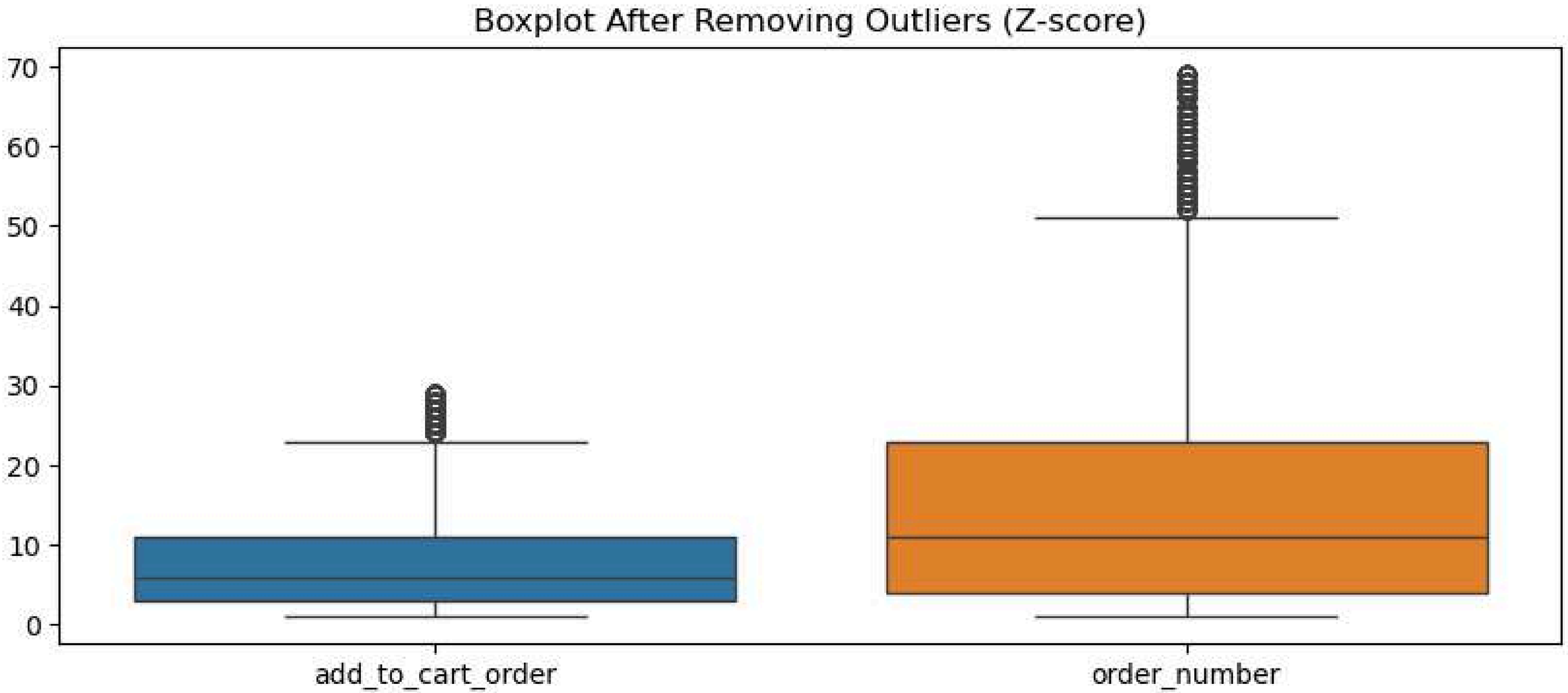
توزيع الطلبات حسب القسم (افتراضي (Orders by Department)



🔍 ماذا تخبرنا البيانات؟

يتوافق توزيع الطلبات حسب القسم مع تحليل المنتجات والممرات، 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.

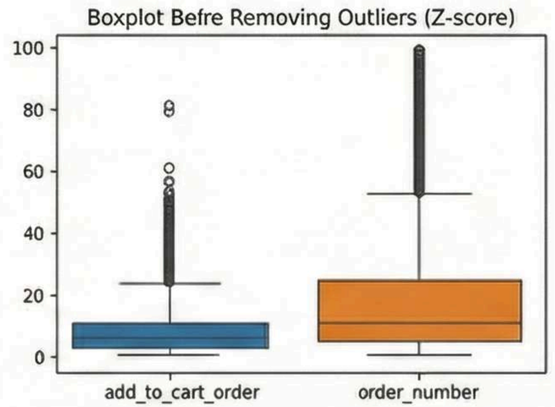

Boxplot Before Removing Outliers (Z-score)
100
80
60
40
20
0
add_to_cart_order
order_number


Boxplot After Removing Outliers (Z-score)
70
60
50
40
30
20
10
0
add_to_cart_order
order_number

معالجة القيم المتطرفة (Outlier Treatment)

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

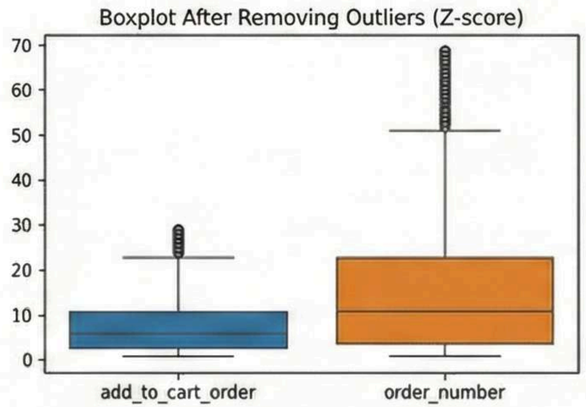
قبل إزالة القيم المتطرفة (Before Removing Outliers - Z-oore)



🔍 ماذا تخبرنا البيانات؟

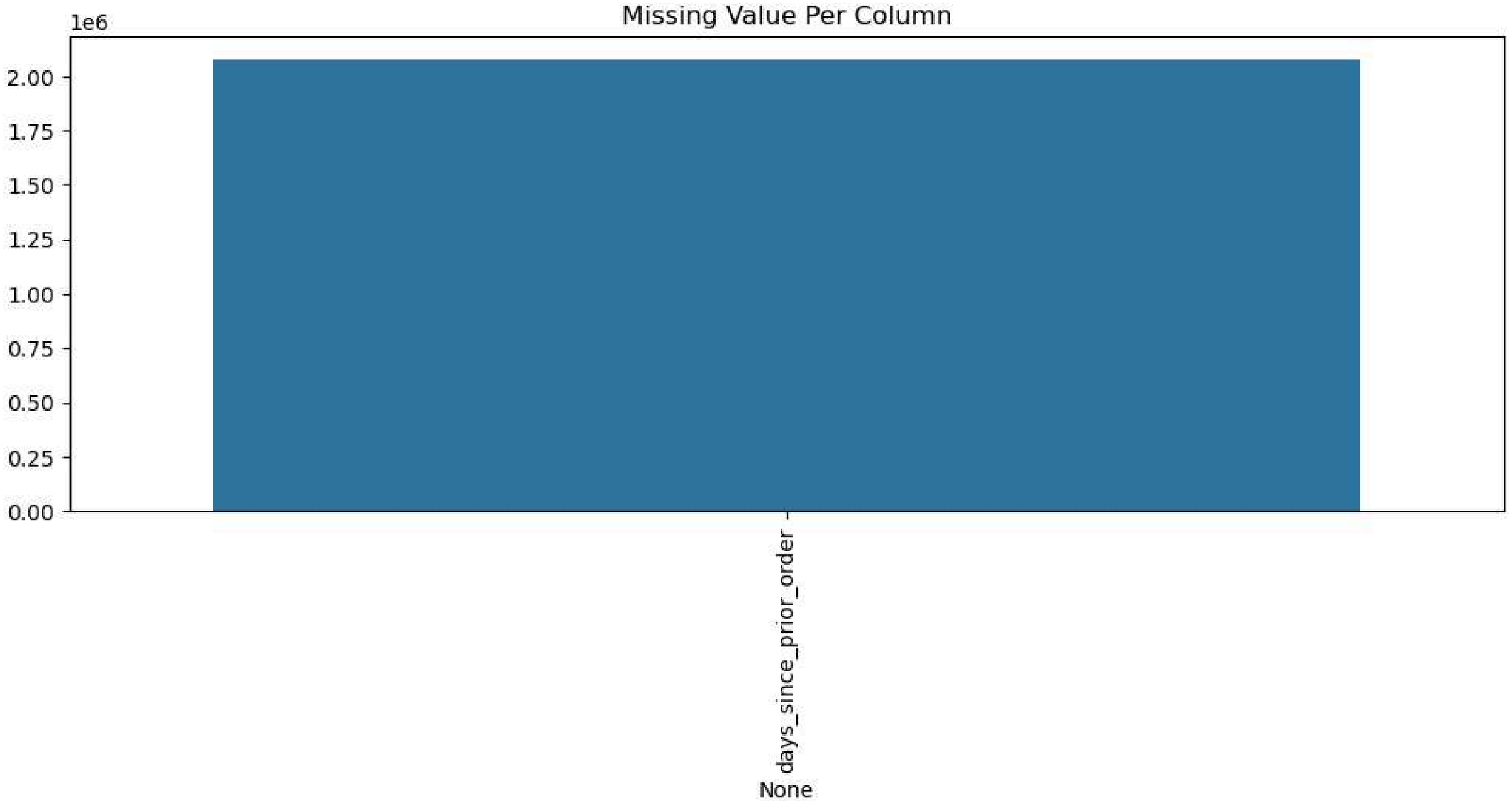
يوضح المخطط الصندوقي قبل المعالجة وجود عدد كبير من القيم المتطرفة (النقاط الفردية خارج "الشوارب"). في `add_to_cart_order`، تصل القيم إلى 80، بينما في `order_number`، تصل إلى 100، مما يشير إلى وجود ملاحظات شاذة قد تؤثر على التحليل الإحصائي.

بعد إزالة القيم المتطرفة (After Removing Outliers - Z-score)



🔍 ماذا تخبرنا البيانات؟

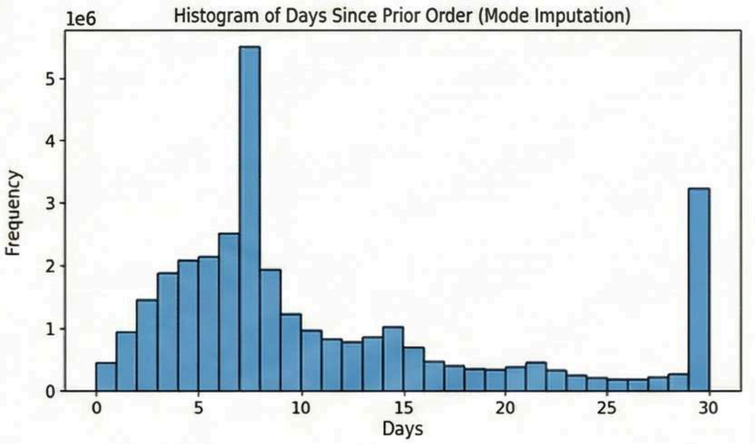
بعد تطبيق طريقة Z-core، تم تقليص عدد القيم المتطرفة بشكل كبير. أصبحت القيم القصوى لـ `add_to_cart_order` عند 30 ولـ `order_number` عند 70 تقريباً. هذا يؤدي إلى توزيع بيانات أكثر نظافة وموثوقية لبناء النماذج.


Missing Value Per Column
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
days_since_prior_order
None

مقارنة طرق معالجة القيم المفقودة للفترة بين الطلبات

تحليل تأثير الطرق المختلفة على توزيع البيانات واختيار الطريقة الأنسب منطقياً

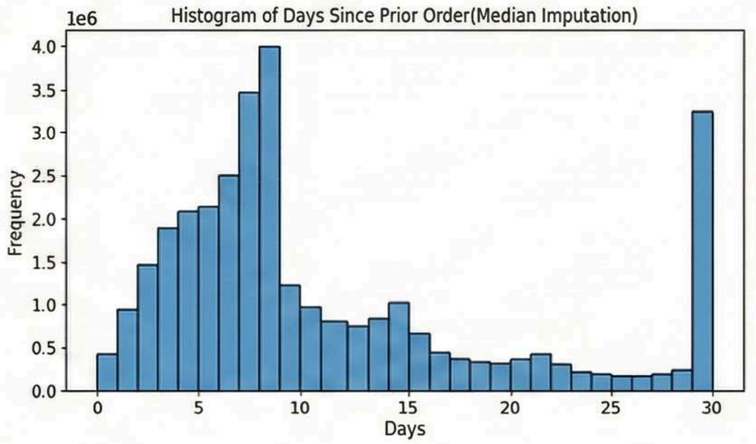
طريقة بالمنوال (Mode Imputation)



🔍 ماذا تخبرنا البيانات؟

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

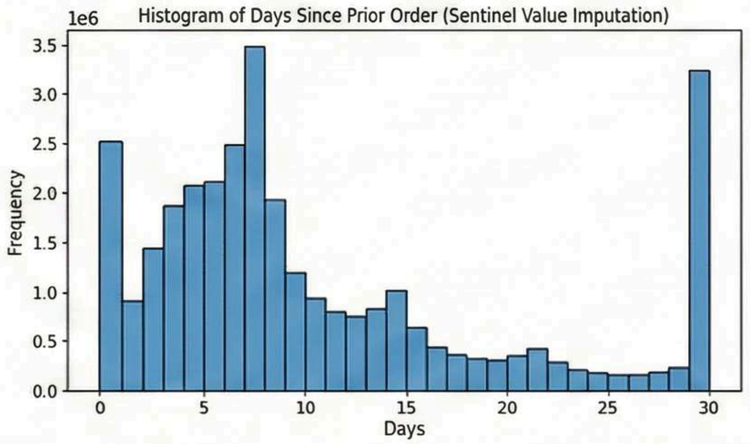
طريقة بالوسيط (Median Imputation)



🔍 ماذا تخبرنا البيانات؟

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

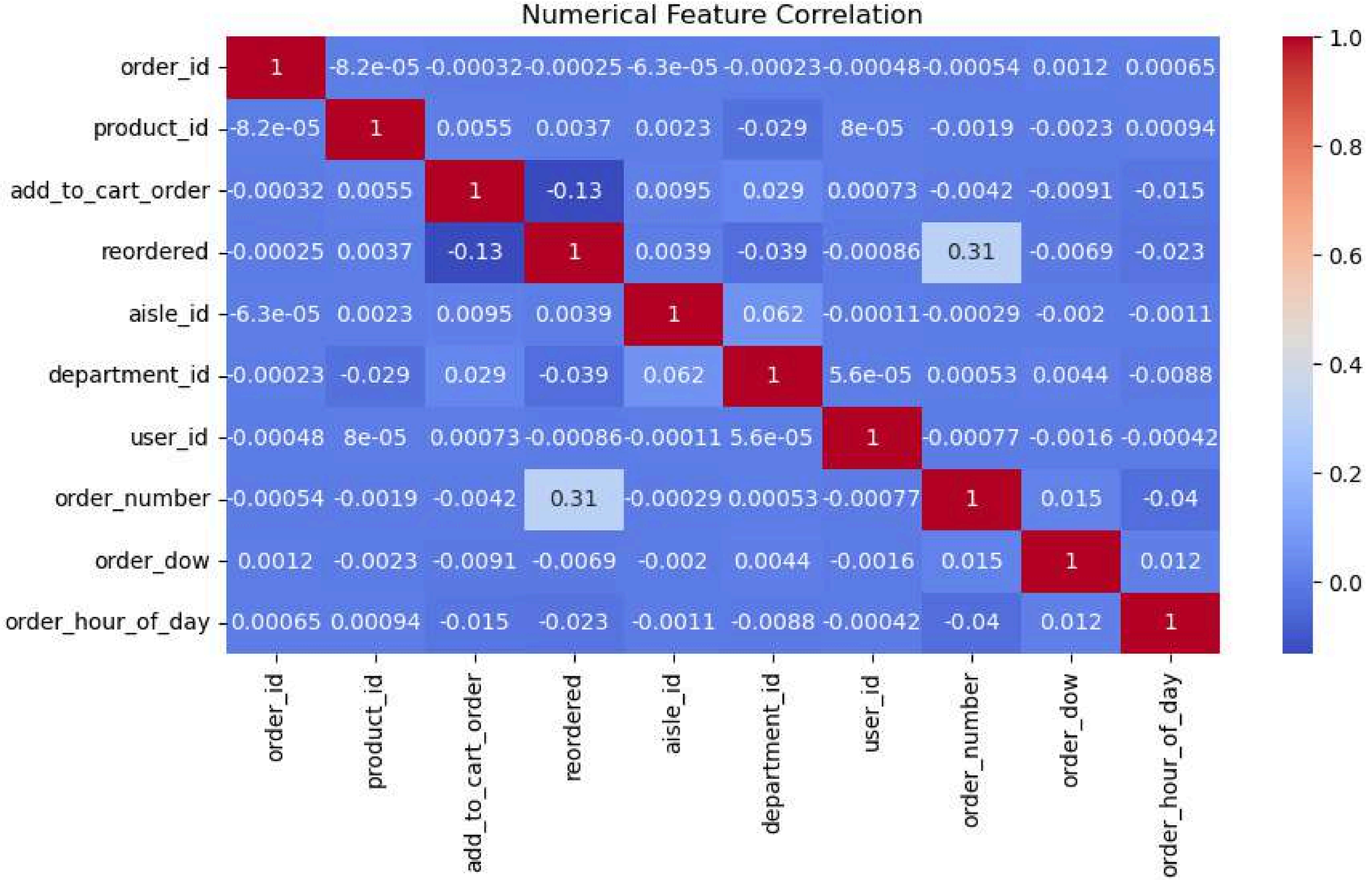
ملء بالقيمة الثابتة 0 (Sentinel Value - 0) ⭐ الخيار الأنسب



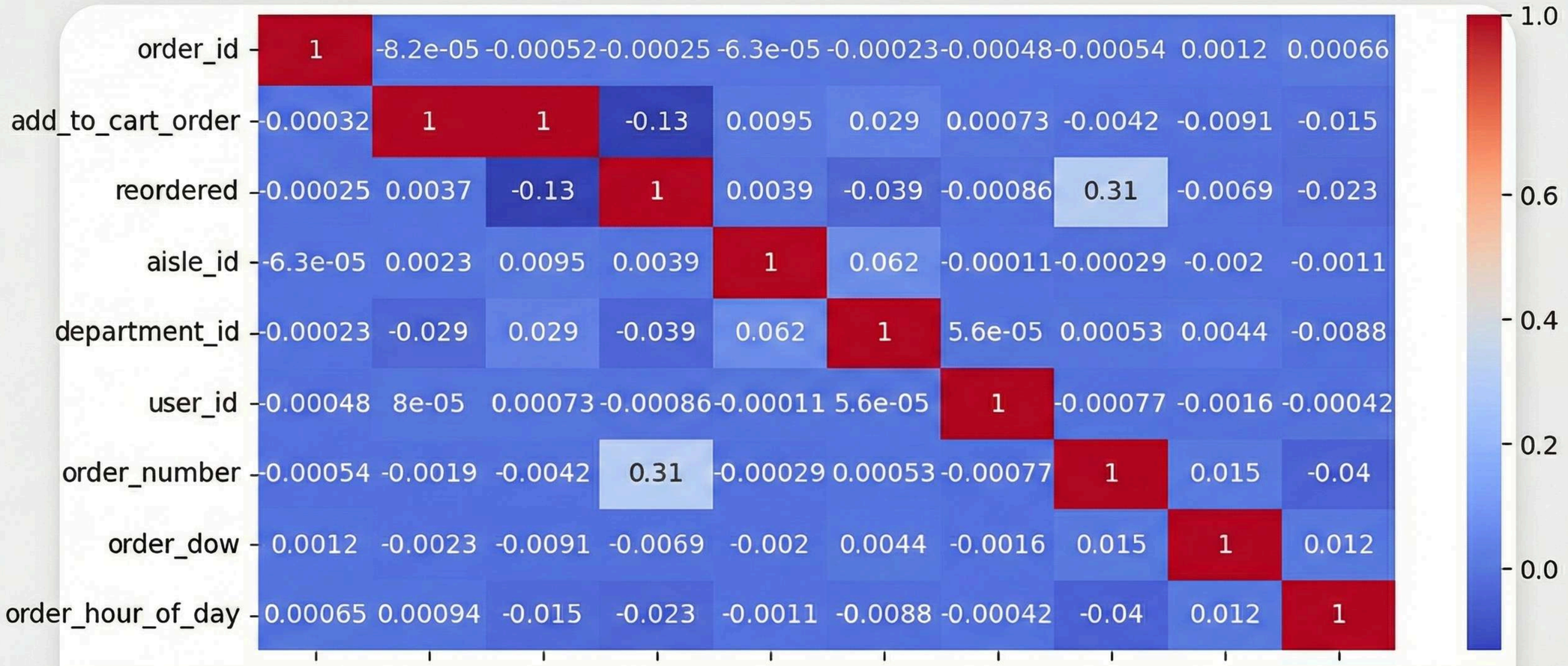
🔍 ماذا تخبرنا البيانات؟

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).

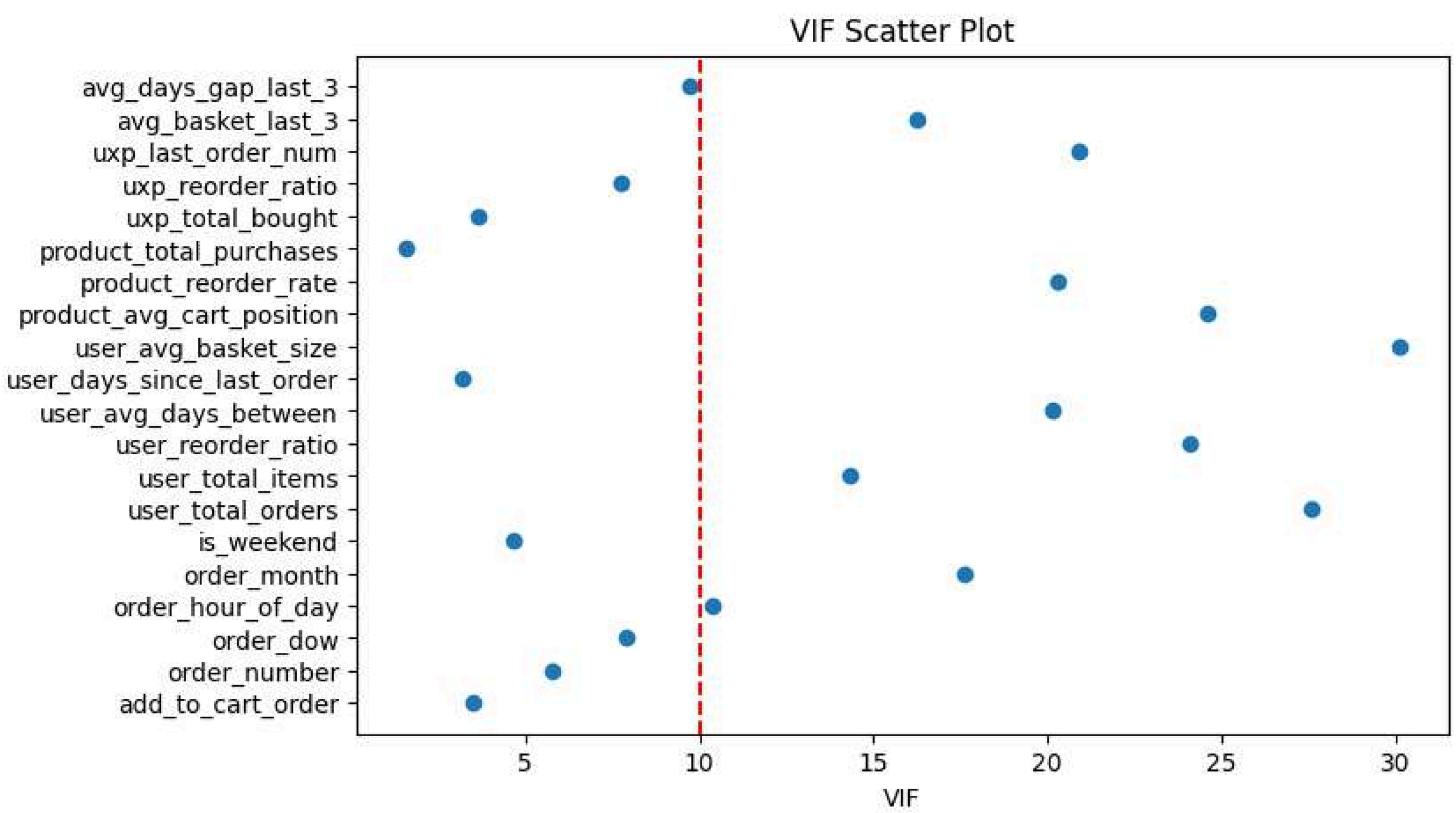

Numerical Feature Correlation
order_id 1 -8.2e-05 -0.00032 -0.00025 -6.3e-05 -0.00023 -0.00048 -0.00054 0.0012 0.00065
product_id -8.2e-05 1 0.0055 0.0037 0.0023 -0.029 8e-05 -0.0019 -0.0023 0.00094
add_to_cart_order -0.00032 0.0055 1 -0.13 0.0095 0.029 0.00073 -0.0042 -0.0091 -0.015
reordered -0.00025 0.0037 -0.13 1 0.0039 -0.039 -0.00086 0.31 -0.0069 -0.023
aisle_id -6.3e-05 0.0023 0.0095 0.0039 1 0.062 -0.00011 -0.00029 -0.002 -0.0011
department_id -0.00023 -0.029 0.029 -0.039 0.062 1 5.6e-05 0.00053 0.0044 -0.0088
user_id -0.00048 8e-05 0.00073 -0.00086 -0.00011 5.6e-05 1 -0.00077 -0.0016 -0.00042
order_number -0.00054 -0.0019 -0.0042 0.31 -0.00029 0.00053 -0.00077 1 0.015 -0.04
order_dow 0.0012 -0.0023 -0.0091 -0.0069 -0.002 0.0044 -0.0016 0.015 1 0.012
order_hour_of_day 0.00065 0.00094 -0.015 -0.023 -0.0011 -0.0088 -0.00042 -0.04 0.012 1
order_id
product_id
add_to_cart_order
reordered
aisle_id
department_id
user_id
order_number
order_dow
order_hour_of_day
1.0
0.8
0.6
0.4
0.2
0.0

تحليل نسبة إعادة الطلب وعدم إعادته



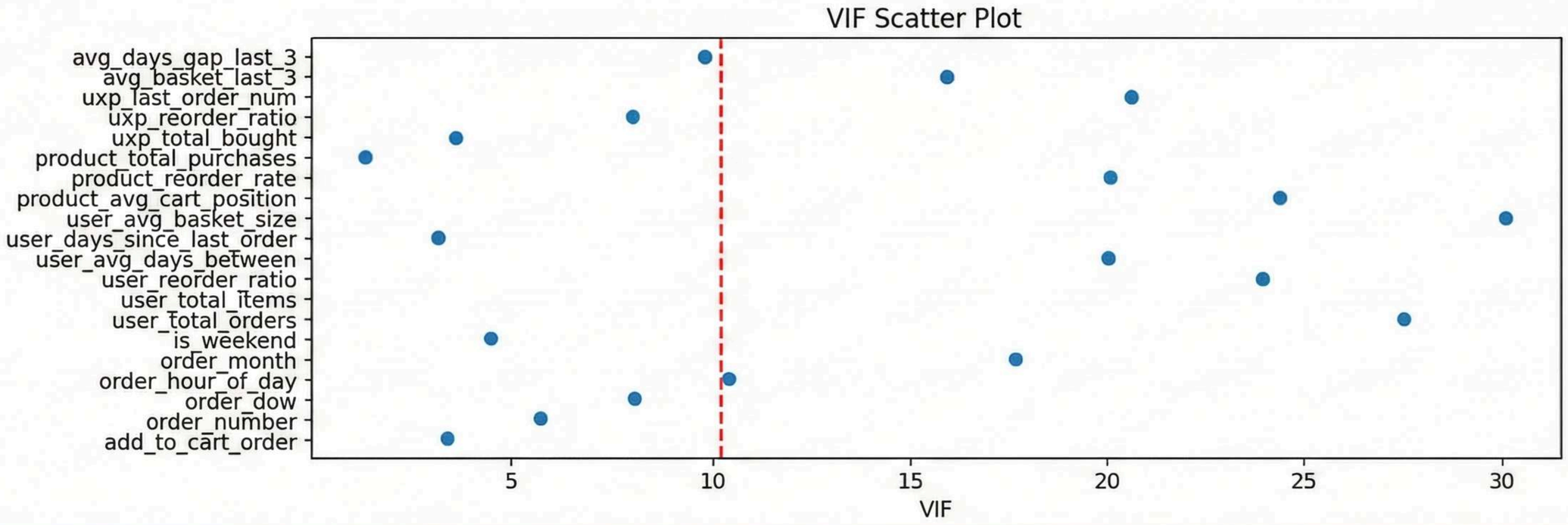
🔍 ماذا تخبرنا البيانات؟

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
avg_days_gap_last_3
avg_basket_last_3
uxp_last_order_num
uxp_reorder_ratio
uxp_total_bought
product_total_purchases
product_reorder_rate
product_avg_cart_position
user_avg_basket_size
user_days_since_last_order
user_avg_days_between
user_reorder_ratio
user_total_items
user_total_orders
is_weekend
order_month
order_hour_of_day
order_dow
order_number
add_to_cart_order
5
10
15
20
25
30
VIF

تحليل التعددية الخطية (VIF)

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



🔍 ماذا تخبرنا البيانات؟

يوضح الرسم البياني قيم عامل تضخم التباين (VIF) لكل متغير مستقل. النقاط التي تقع على يمين الخط الأحمر المتقطع (VIF > 10) تشير إلى (VIF > 10) تشير إلى وجود تعددية خطية عالية قد تؤثر سلباً على نموذج الانحدار، مثل `product_reorder_r` و `user_avg_basket_size` `uxp_reorder_ratio` بينما المتغيرات على الا اليسار (VIF < 10) تعتبر ذات ارتباط مقبول ولا تمثل مشكلة كبيرة، مثل `uxp` و `order_dow` `is_weekend` و `product_total_purchases`.