أكاديمية طويق
TUWAIQ ACADEMY

معسكر علم البيانات و تعلم الآلة

13 -11 - 2022

أكاديمية طويق
TUWAIQ ACADEMY

نبذة عن المدرب

أكاديمية طويق
TUWAIQ ACADEMY

محتوى المعسكر

اليوم	الأسبوع الأول Getting Started	الأسبوع الثاني Data Analysis and Visualization	الأسبوع الثالث Machine Learning	الأسبوع الرابع EDA & FE in Action	الأسبوع الخامس Modeling Interpretation in Action	الأسبوع السادس Final Project
الأحد	Intro to DS	NumPy	Intro to ML	DS Knowledge Catalog	Models Families: Distance & Time Series	Final Project
الإثنين	Git & Github	Pandas	Supervised ML	**EDA1: Univariate & Multivariate Analysis**	Models Evaluation: Regression & Classification	Final Project
الثلاثاء	Python Review	Matplotlib	Supervised ML	EDA2: Association Analysis & Hypothesis Construction	Optimization Techniques	Final Project
الأربعاء	Python Review	Seaborn	Unsupervised ML	Features Engineering: Scaling, Merging & Discretization	NLP and Text Mining Basics	Final Project
الخميس	Python Review	Plotly	Unsupervised ML	Models Families: Continuous & Categorical	Neural Networks Basics	Presentation

**ملاحظة: قد تتغير المواضيع أو أوقات طرحها بناء على تقدم الطلاب.

مرحلة استكشاف البيانات EDA؟

ما هي مرحلة تحليل البيانات الاستكشافي؟

هي مرحلة يتم فيها التعرف على البيانات، وصفها، وحالتها، ومن ثم تنظيفها، وتُوجه الطريق للوصول إلى بيانات يُعتمد عليها في اتخاذ وتوجيه القرارات

الهدف من مرحلة تحليل البيانات الاستكشافي

لأننا نحاول في هذه المرحلة فهم البيانات والوصول إلى نسخة منقحة يعتمد عليها، إلا أن الطريق للوصول لهذه النسخة يُحدده الهدف من العمل على هذ المنتج.

تخدم هذه المرحلة الأهداف التالية:

1. الحصول على مرئيات عن البيانات ومشاركتها
2. فهم الهيكل الأساسي للبيانات
3. استخراج الخواص المهمة

الهدف من مرحلة تحليل البيانات الاستكشافي

لأننا نحاول في هذه المرحلة فهم البيانات والوصول إلى نسخة منقحة يعتمد عليها، إلا أن الطريق للوصول لهذه النسخة يُحدده الهدف من العمل على هذ المنتج.

تخدم هذه المرحلة الأهداف التالية:

4. الكشف عن القيم والحالات الشاذة
5. اختبار الافتراضات التي تم وضعها في مرحلة فهم المشكلة
6. توجيه الوصول إلى النتيجة النهائية

خطوات تحليل البيانات

1. تحديد المتطلبات من البيانات

تجيب هذه الخطوة على الأسئلة التالية:

- ما هو حجم البيانات؟
- هل تخدم أنواع البيانات الحالية النتيجة التي نهدف الوصول لها؟
- تحديد مصادر البيانات وكيف نشأت هذه الحقائق

2. تجميع وتنظيم البيانات للتحليل

تجيب هذه الخطوة على الأسئلة التالية:

- ما هي البيانات الأخرى التي أستطيع إضافتها لتعزز الهدف من العمل؟
- من أين سنقوم بجمعها وماهي التقنيات التي نحتاجها؟
- من سيقوم بجمعها؟

3. فحص صحة البيانات

تجيب هذه الخطوة على الأسئلة التالية:

- هل هناك حقائق فارغة؟ كيف سأتعامل معها؟
- هل هناك حقائق شاذة أو مغلوطة؟ كيف سأتعامل معها؟
- هل هناك بيانات مكررة؟

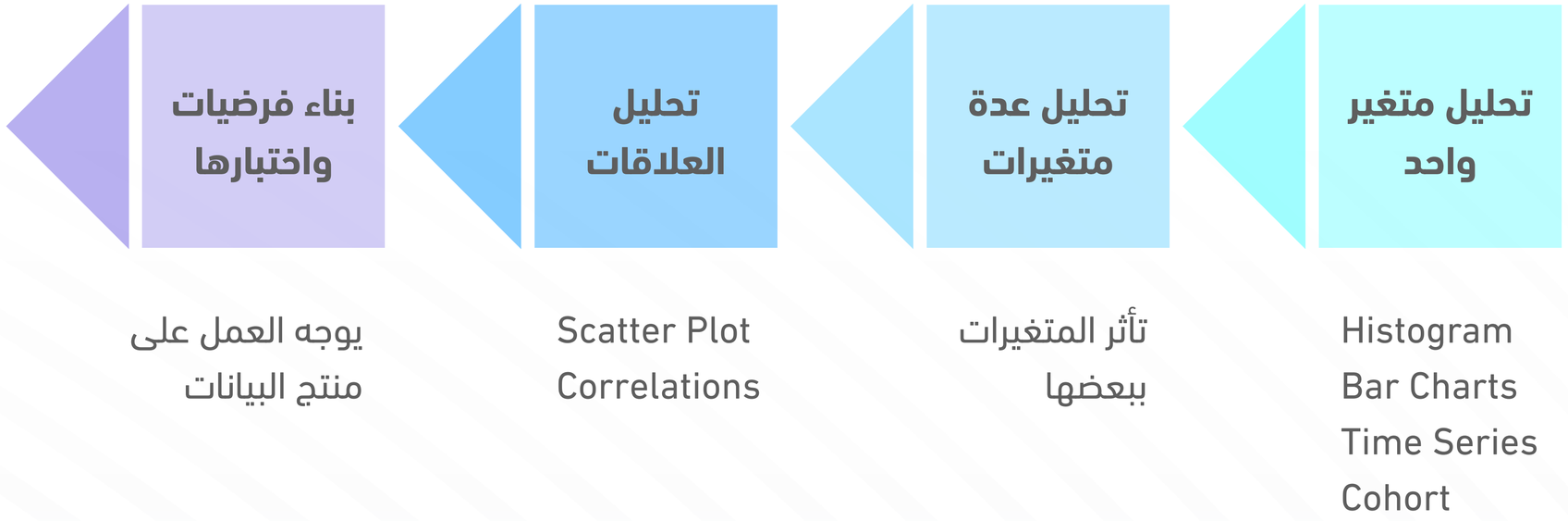
4. تحليل العلاقات في البيانات

تجيب هذه الخطوة على الأسئلة التالية:

- ما مدى تأثير كل عامود أو خاصية على النتيجة النهائية؟
- ماهي الأعمدة التي ستستمر للمرحلة التالية؟

أنواع تحليل البيانات

هناك عدة أنواع للتحليل الوصفي للبيانات



تحليل متغير واحد

تحليل متغير واحد - أرقام

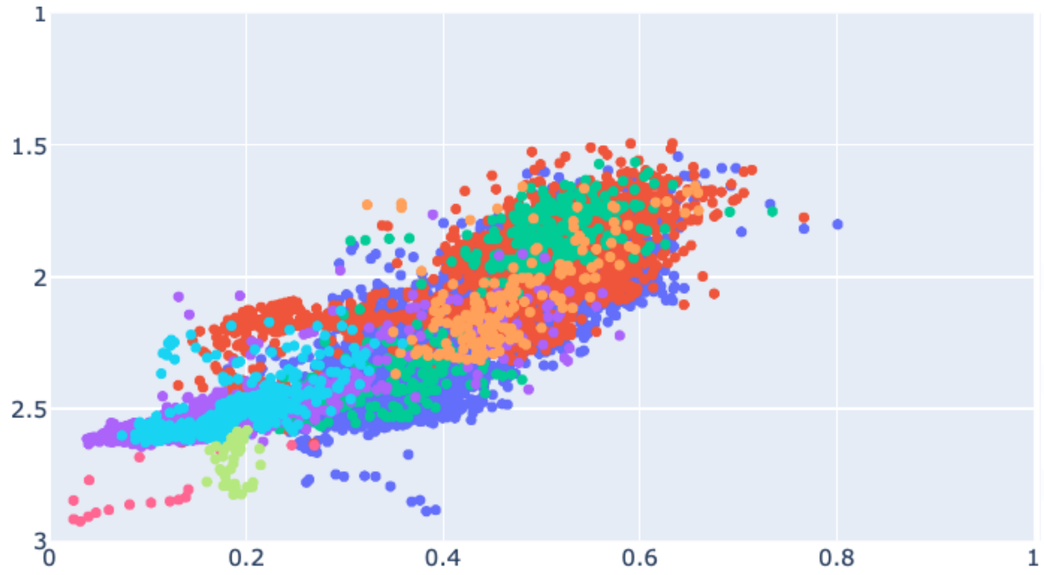
تجيب هذه الخطوة على الأسئلة التالية:

- ما هو معدل القيم؟ والشذوذ فيها؟
- ماهو توزيع كل متغير رقمي
- ما مدى اكتفاءها بالبيانات المتوفرة فيه
- هل يمكن معالجة القيم الفارغة له بإحدى الطرق؟

تحليل متغير واحد - أرقام

التمثيلات المنطبقة على تحليل المتغير الواحد:

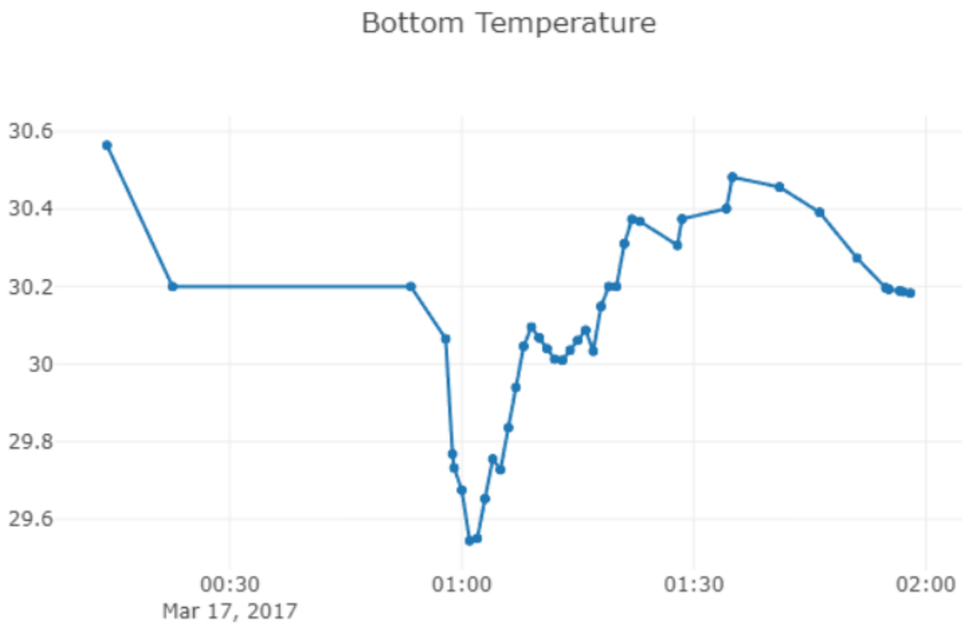
تحليل الانتشار باستخدام
Scatter Plot



تحليل متغير واحد - أرقام

التمثيلات المنطبقة على تحليل المتغير الواحد:

تحليل Trend باستخدام
التمثيل الخطي Line Plot



تحليل متغير واحد - تصنيفات

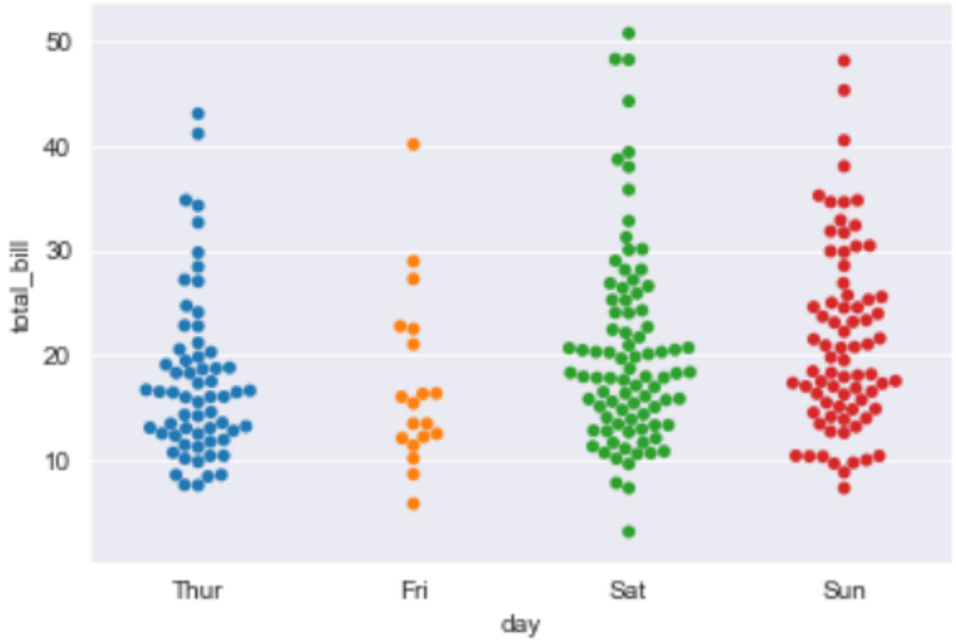
تجيب هذه الخطوة على الأسئلة التالية:

- ماهو توزيع حجم البيانات على التصنيفات
- ما هو حجم التصنيفات لدي، هل هو متزن؟ شامل؟
- هل هناك فراغات استطيع ملؤها بالاستعانة بتصنيف دارج أو تصنيف جديد؟

تحليل متغير واحد - تصنيفات

التمثيلات المنطبقة على تحليل المتغير الواحد:

تحليل الانتشار باستخدام
Swarm Plot



تحليل متغيرين أو أكثر

تحليل متغيرين أو أكثر

تجيب هذه الخطوة على الأسئلة التالية:

- ما علاقة المتغيرات ببعضها؟
- اكتشاف التجمعات (Clusters)

تحليل متغير واحد - تصنيفات

التمثيلات المنطبقة على تحليل علاقة متغيرين أو أكثر:

تحليل المجموعات المترابطة

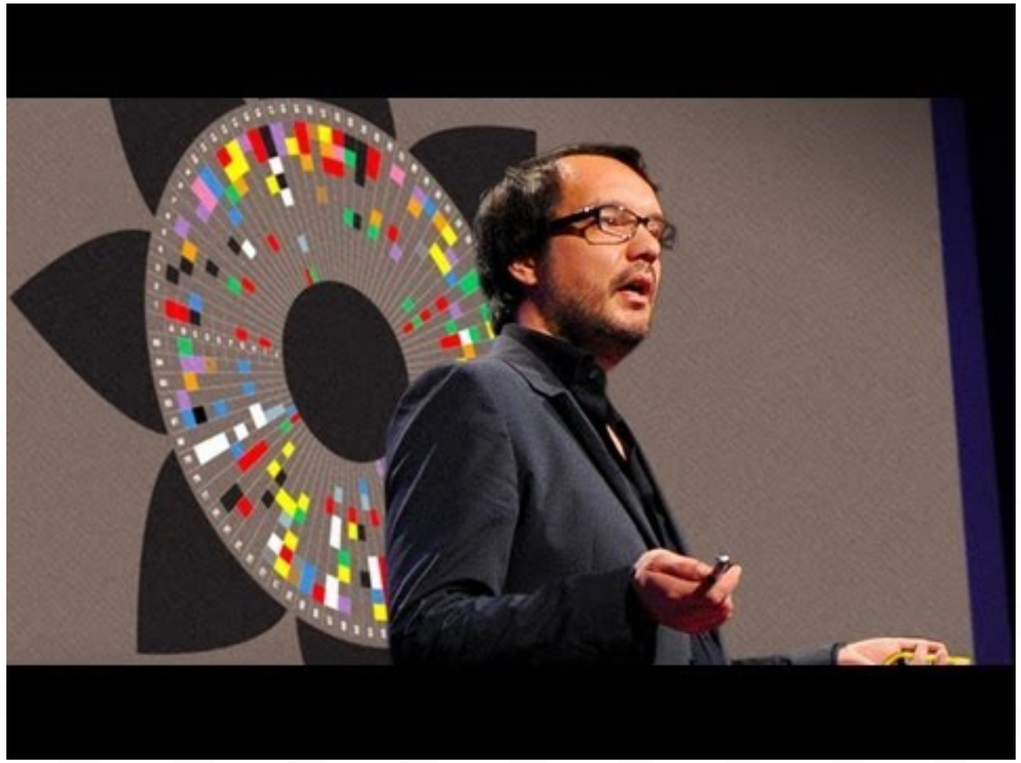
Cohort Analysis

APPS LAUNCHED ↓ % ACTIVE USERS AFTER LAUNCH →

COHORT	USERS	DAY 0	DAY 1	DAY 2	DAY 3	DAY 4	DAY 5	DAY 6	DAY 7	DAY 8	DAY 9	DAY 10
Jan 25	1,098	100%	33.9%	23.5%	18.7%	15.9%	16.3%	14.2%	14.5%	Retention over user lifetime		12.1%
Jan 26	1,358	100%	31.1%	18.6%	14.3%	16.0%	14.9%	13.2%	12.9%			
Jan 27	1,257	100%	27.2%	19.6%	14.5%	12.9%	13.4%	13.0%	10.8%	11.4%		
Jan 28	1,587	100%	26.6%	17.9%	14.6%	14.8%	14.9%	13.7%	11.9%			
Jan 29	1,758	100%	26.2%	20.4%	16.9%	14.3%	12.7%	12.5%				
Jan 30	1,624	100%	26.4%	18.1%	13.7%	15.4%	11.8%					
Jan 31	1,541	100%	23.9%	19.6%	15.0%	14.8%						
Feb 01	868	100%	24.7%	16.9%	15.8%							
Feb 02	1,143	Retention over product lifetime		18.5%								
Feb 03	1,253											
All Users	**12,487**	**100%**	**27.0%**	**19.2%**	**15.4%**	**14.9%**	**14.0%**	**13.3%**	**12.5%**	**13.1%**	**12.2%**	**12.1%**

خارج النص

فيديو: تحليل البيانات الاستكشافي



https://youtu.be/5Zg-C8AAIGg

فيديو: تحليل البيانات الاستكشافي

https://www.youtube.com/watch?v=Sm5xF-UYgdg

أمثلة للتفكير

مثال حالة استخدام: توقع حجم الصادرات للشهر التالي

فريق التخطيط للإيرادات لشعبة الجمارك في زاتكا يهدف إلى الوصول إلى قيم تقديرية لحجم الصادرات

سنمضي بالمثال التالي: بيانات الصادرات عن السعودية

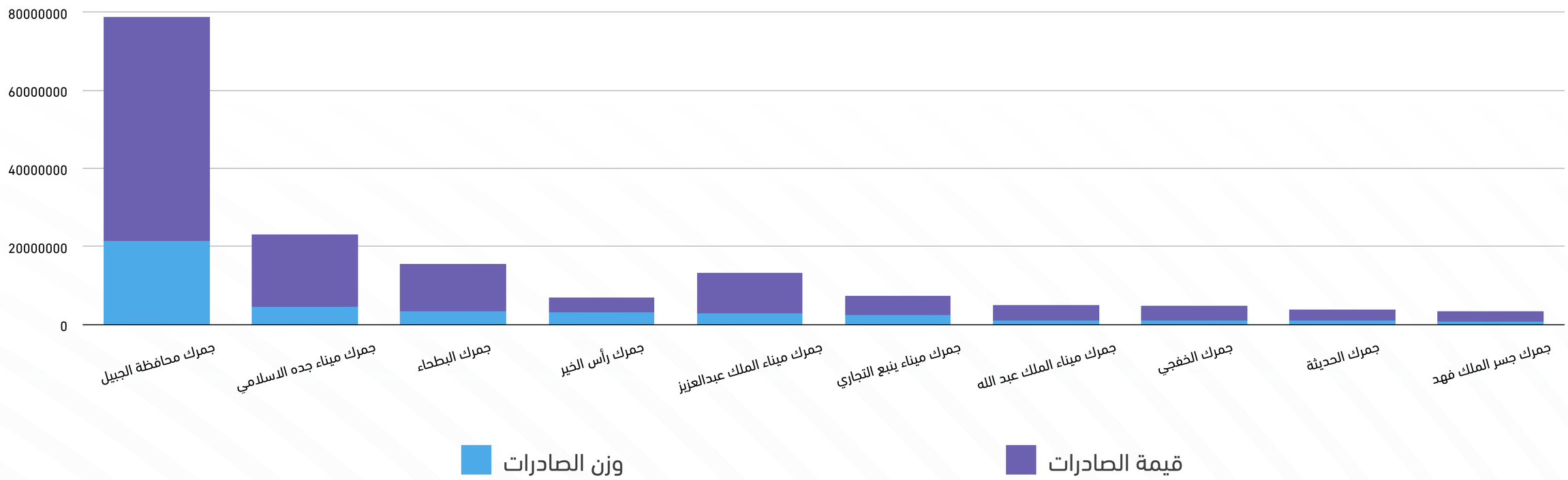
المصدر: البيانات المفتوحة ببعض التصرف لZATCA

ID	الجمرك	وزن الصادر	قيمة الصادر
12345	جمرك ميناء الملك عبدالعزيز	500	2,017.95
56789	جمرك مطارالملك عبدالعزيزالدولي	120	97,640.53

مبدئيا، هذه البيانات المبدئية التي تمت إتاحتها لفريق التحليل،
ماذا نلاحظ بهذه الأرقام؟

سنمضي بالمثال التالي: بيانات الصادرات عن السعودية

المصدر: البيانات المفتوحة ببعض التصرف لZATCA



مثال حالة استخدام: رصد الواردات المشتبه بها

فريق التدقيق اللاحق في زاتكا يهدف إلى الوصول إلى **قوائم بالشحنات المشتبه بها**

سنمضي بالمثال التالي: بيانات الواردات للسعودية

المصدر: البيانات المفتوحة ببعض التصرف لZATCA

ID	الجمرك	وزن الوارد	قيمة الوارد	نوع الوارد
12345	جمرك ميناء الملك عبدالعزيز	500	2,017.95	أغذية
56789	جمرك مطارالملك عبدالعزيزالدولي	120	97,640.53	بطاريات

ما الخطوات المحتملة للوصول إلى المطلوب من قسم التدقيق؟

أكاديمية طويق
TUWAIQ ACADEMY

استفساراتكم؟ 🤔