تعرف على فاطمة

وهي محللة بيانات متحمسة، شرعت في رحلة في
باستخدام (*EDA*) تحليل البيانات الاستكشافي
مجموعة بيانات معقدة تهدف إلى توجيه
استراتيجيات شركتها.

نظرًا لفهمها لـEDA كعنصر حاسم في علوم البيانات المتأصلة في العلوم والتكنولوجيا والهندسة والرياضيات(STEM)، استخدمت التقنيات الإحصائية وتكنولوجيا الحوسبة والنهج الهندسي لحل المشكلات بشكل منهجي، ومزج النظرية العلمية مع التطبيق العملي لتحقيق الأهداف المتنوعة لـEDA.

وبينما تعمقت فاطمة في فهم الأهداف متعددة الأوجه لتحليل البيانات الاستكشافي EDA

كان كل هدف بمثابة ضوء إرشادي في استكشافها:

1.

اكتشاف الحالات الشاذة

تضمنت مهمتها الأولية تحديد القيم المتطرفة أو الأنماط غير العادية في البيانات، والتي غالبًا ما تكشف عن رؤى أو مجالات رئيسية تحتاج إلى مزيد من الفحص.

2.

اختبار الفرضيات في تحليل البيانات

الاستكشافي سمح لفاطمة باختبار افتراضاتها حول البيانات، وهو خطوة حيوية لتأكيد أو نفي نظرياتها وفهم خصائص مجموعة البيانات.

3.

التحقيق في البيانات

كرست فاطمة الكثير من الوقت للتحقيق العميق في البيانات، وفحص المتغيرات وعلاقاتها المتبادلة، كما يفعل المحقق الذي يجمع الأدلة معًا لتكوين صورة أوسع.

4.

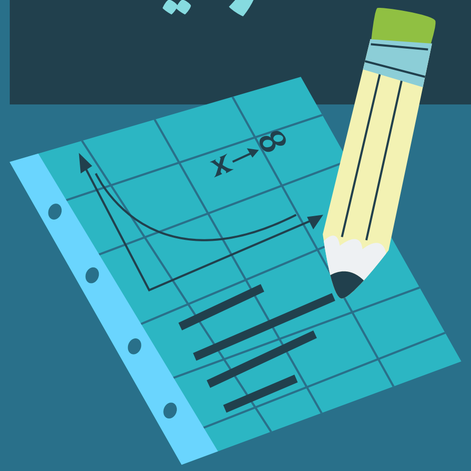
اكتشاف الأنماط

وجدت فاطمة الإثارة في الكشف عن أنماط مثل الاتجاهات أو الارتباطات أو التجمعات، والتي قدمت رؤى قيمة وغالباً ما وجهت المزيد من التحليل.

بالنسبة لفاطمة، عندما شرعت في رحلة تحليل البيانات، كانت الإحصائيات الوصفية

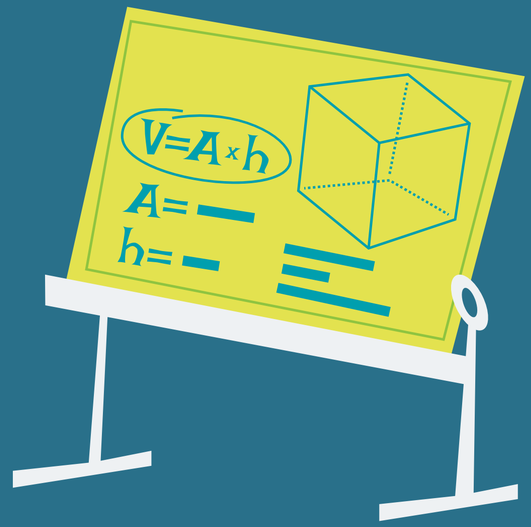
هي الخطوة الأولى في فهم مجموعة البيانات المعقدة التي أمامها.

ومن خلال تطبيق هذه الإحصائيات، حصلت على نظرة عامة أولية لبياناتها

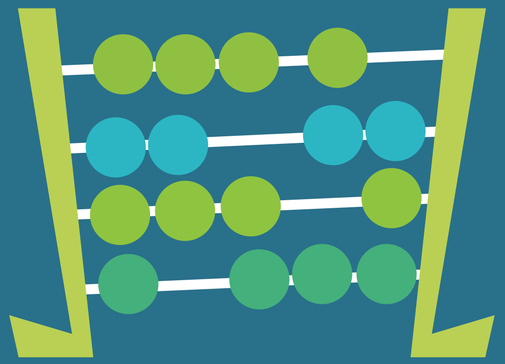


لقد حسبت المتوسط والوسيط للعثور على الاتجاه المركزي، مما منحها فهمًا سريعًا للقيم المتوسطة في مجموعة البيانات الخاصة بها.

كشف النطاق والانحراف المعياري عن مدى انتشار نقاط بياناتها، مما يشير إلى مستوى الاتساق في البيانات.

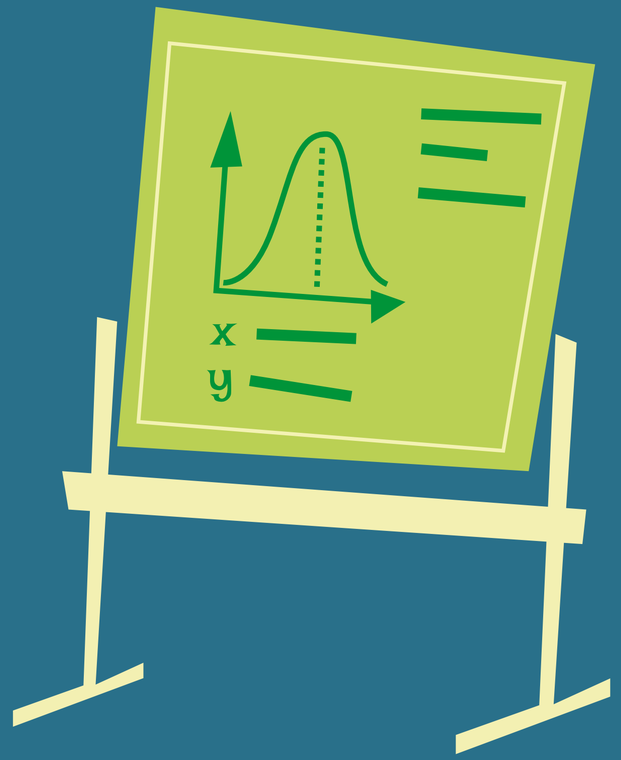


من خلال فحص الانحراف يمكن لفاطمة أن تستنتج ما إذا كانت بياناتها تحمل أي انحراف أو ارتفاع غير عادي.



ساعدتها توزيعات الترددات في تصور البيانات، مما يسهل اكتشاف الأنماط والشذوذات.

ومن خلال تحليل الأرباع والنسب المئوية، تمكنت من تحديد القيم المتطرفة وفهم توزيع البيانات عبر مستويات مختلفة.

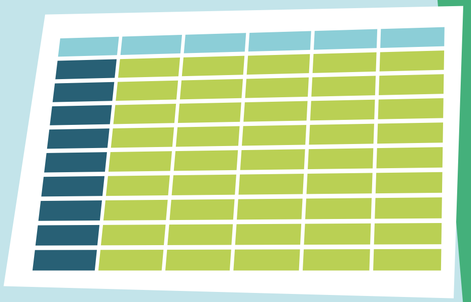


أجرت فاطمة تشخيصًا للبيانات، لتحديد أي مشكلات في مجموعة البيانات اللتي قد تؤثر على دقة التحليل أو موثوقيته.

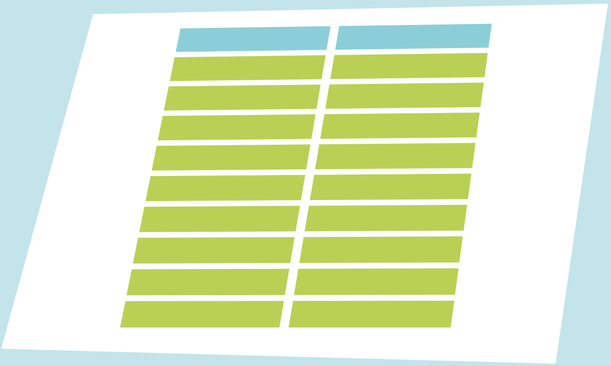
ما هي الخطوات الأساسية للتشخيص؟

بدأت فاطمة بتحديد نوع البيانات لكل عمود في مجموعة البيانات الخاصة بها، مما ساعد في فهم ما إذا كانت البيانات هي عددية، أو تصنيفية، أو نصية، أو تاريخية/زمنية

ثم قامت بتصنيف أبعاد مجموعة البياناتٍ عن طريق عد الصفوف والأعمدة، مما قدم فهمًا واضحًا لحجمها ونطاقها لتخطيط تحليلها



تحديد نوع البيانات في العمود: بعد ذلك، تحدد فاطمة أنواع البيانات الدقيقة لعمودين محددين، مثل عدد صحيح، أو عدد عشري، أو سلسلة نصية، أو القيمة المنطقية

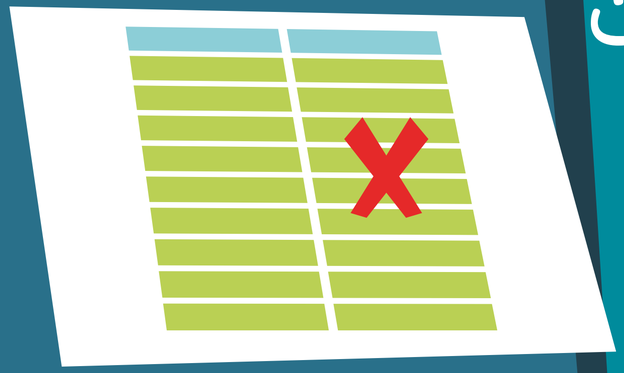


تتحقق فاطمة من وجود قيم فارغة أو مفقودة في كل عمود، وهي خطوة حاسمة لتقييم اكتمال البيانات وسلامتها

تقييم الخصائص الإحصائية: في النهاية، قامت فاطمة بتقييم مقاييس إحصائية رئيسية مثل المتوسط والوسيط والانحراف المعياري لكل عمود لفهم توزيع البيانات والاتجاهات المركزية، مما يعد المرحلة الأولى لتحليل أعمق

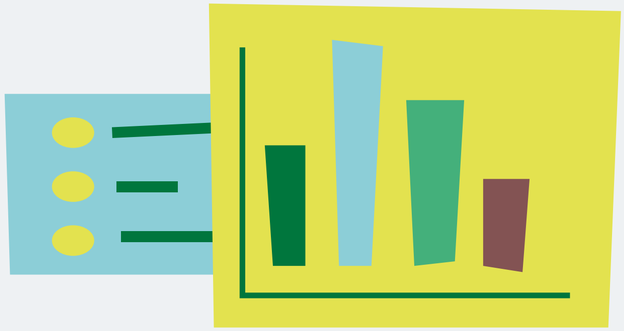
تتقدم فاطمة إلى تنظيف البيانات بعد تشخيص البيانات، بهدف تحسين جودة مجموعة البيانات الخاصة بها.

تبدأ بالتعامل مع القيم الخالية، وتقرر ما إذا كانت تريد إزالتها أو إدراجها. ثم تقوم بعد ذلك بتقييم مدى ملاءمة كل عمود، وإزالة الأعمدة غير المساهمة.

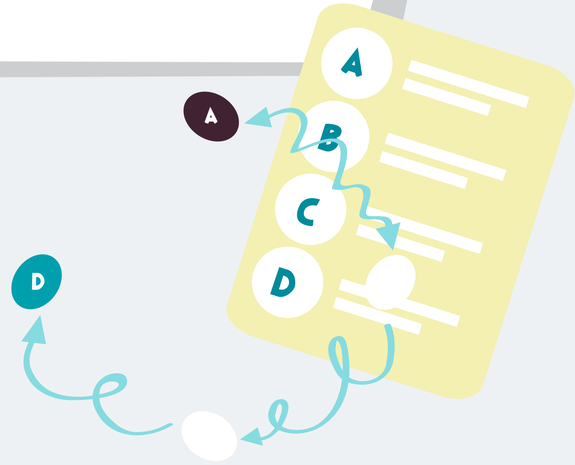


وبعد ذلك، تقوم بتحديد السجلات المكررة وإزالتها

إذا واجهت فاطمة تحديات مع بنية أنواع البيانات في مجموعة البيانات الخاصة بها، وتحديداً التفريق بين الأنواع العددية والتصنيفية، فيجب عليها تصحيح هذه التصنيفات.



تلعب البيانات العددية دوراً رئيسيًا في العمليات الحسابية وتحسين استخدام الذاكرة في إطارات البيانات(*data frames*), وهو أمر مهم للمكتبات المستخدمة في الحسابات الإحصائية وتعلم الآلة.



لكن نطاق قيمتها يمكن أن يؤثر على أداء النموذج، لذلك يتم استخدام التوحيد *Normalization* أو التقليص *Scaling* لتوحيد النطاقات، وتحقيق توازن في تأثير الميزات على التحليل وبناء النموذج.

يعمل التوحيد على ضبط القيم ضمن نطاق [0، 1]، مما يفيد الخوارزميات الحساسة للقياس مثل *K-Nearest Neighbors*.

$$X_{Normalized} = \frac{X_{intial}}{max\ (X)}$$

أما التقييس(*Standardization*)، من ناحية أخرى، فيتم إلى متوسط يساوي صفر وانحراف معياري يساوي واحد، وهو مثالي للطرق مثل آلات المتجهات الداعمة (*Support Vector Machines*) التي تفترض توزيعًا طبيعيًا للبيانات.

$$X_{Standardized} = \frac{X_{intial} - \max(X)}{std(X)}$$

$$X_{Normalized} = \frac{X_{intial}}{max\ (X)}$$

الاختيار بينهما يعتمد على الخوارزمية الخاصة وخصائص البيانات، حيث يكون التوحيد حساسًا للقيم الطرفية، في حين أن التقييس أقل حساسية لها

$$X_{Standardized} = \frac{X_{intial} - max\ (X)}{std\ (X)}$$

تعكس مجموعة فاطمة من البيانات المتحولة، التي أصبحت الآن سردًا يتجاوز الأرقام والفئات، رحلتها الماهرة في تحليل البيانات الاستكشافي *EDA*, تسلط هذه العملية الضوء على قوة النهج المنهجي والثاقب في الكشف عن قصص البيانات المخفية.