أساسيات تحليل البيانات: الدليل النظري المتكامل

تحليل البيانات هو الفن والعلم الذي يهدف إلى تحويل البيانات الخام إلى رؤى قابلة للتنفيذ (Actionable Insights) تساعد في اتخاذ القرارات. هذا الدليل يستعرض المفاهيم الأساسية التي يحتاجها كل محلل بيانات للبدء في هذا المجال.

---

1. ما هو إطار البيانات (DataFrame)؟

عندما تتعامل مع البيانات باستخدام لغة مثل Python (وتحديداً مكتبة `pandas`)، فإن الهيكل الأساسي الذي ستعمل عليه يُسمى **DataFrame** * **الوصف:** يمكنك تخيل الـ DataFrame كجدول يشبه جداول Excel، يتكون من صفوف (Rows) وأعمدة (Columns). * **الأعمدة (Features/ Variables):** كل عمود يمثل "متغيراً" أو سمة معينة من البيانات (مثل: التاريخ، المدينة، التكلفة، أو اسم المنتج). ويجب أن يحتوي العمود الواحد على نفس نوع البيانات (مثلاً أرقام فقط، أو نصوص فقط). * **الصفوف (Observations/Records):** كل صف يمثل "سجلاً" أو عملية واحدة كاملة (مثلاً: عملية بيع واحدة تمت في تاريخ معين وبواسطة مندوب محدد).

الدوال الأساسية لاستكشاف إطار البيانات:

- `df.head(n)`: عرض أول `n` صفوف لأخذ نظرة سريعة على شكل البيانات (غالباً نستخدم `head(5)` أو `head(10)`).
- `df.info()`: للحصول على ملخص تقني يشمل عدد الصفوف، أسماء الأعمدة، أنواع البيانات (`dtype`)، وما إذا كانت هناك قيم مفقودة (Non-Null Count).
- `df.describe()`: يعطيك ملخصاً إحصائياً للأعمدة الرقمية (المتوسط، الانحراف المعياري، الربعيات، والحدود الدنيا والقصوى).

---

2. تنظيف البيانات (Data Cleaning)

لا يوجد مشروع تحليل بيانات دون الحاجة إلى التنظيف. البيانات الخام غالباً ما تكون غير دقيقة، ناقصة، أو غير متسقة.

- **القيم المفقودة (Missing Values):** هي الخلايا الفارغة في الجدول (تظهر كـ `NaN`). يمكن التعامل معها إما بحذف الصفوف/الأعمدة إذا كان نقصها لا يؤثر، أو باستبدالها بقيم مثل "المتوسط الحسابي" (Mean) أو القيمة الأكثر تكراراً (Mode).
- **القيم الشاذة أو المتطرفة (Outliers):** قيم تبتعد بشكل غير طبيعي عن باقي البيانات (مثلاً إدخال تكلفة بـ 50,000,000 لمنتج تكلفته الطبيعية 3,000). يتم اكتشافها عادةً باستخدام المدى الربيعي (IQR) ومعالجتها لتجنب تشويه التحليل.
- **التوحيد والإملاء (Standardization):** من الأخطاء الشائعة كتابة نفس الاسم بطرق مختلفة (مثل `Mjeeed` و `Mjeed`، أو `Lulu` و `Lulu Hyper`). نستخدم دوال مثل `replace()` لتوحيدها، حيث أن النظام يعتبرها مدخلات مختلفة تماماً إن لم تتوحد.
- **تنسيق التواريخ:** تحويل السلاسل النصية أو الأرقام التي تمثل التواريخ إلى صيغة تاريخ قياسية (`datetime`) لتسهيل التحليل الزمني (كالتحليل حسب الشهر أو السنة).

---

3. هندسة الميزات (Feature Engineering)

تعني إنشاء أعمدة جديدة بناءً على البيانات الموجودة بهدف استخراج رؤى أعمق لا يمكن استنتاجها من البيانات الخام مباشرة. * **الحسابات الرياضية المباشرة:** مثل طرح عمود التكلفة `Cost` من عمود المبيعات `Sale` لإنشاء عمود الربح `Profit`. ($ ext{Profit} = ext{Sale} - ext{Cost}$)
*__التصنيف الشرطي:__ بناء متغيرات جديدة لتجميع البيانات (مثلاً تصنيف المناديب إلى "ذكور" و"إناث" بناءً على الاسم، أو تصنيف الأعمار إلى "شباب"، "بالغين"، "كبار").

---

4. استخراج الرؤى (Extracting Insights) والتجميع (Aggregation)

لفهم "الصورة الكبرى" (The Big Picture)، نحتاج إلى تلخيص ملايين السجلات في أرقام مفيدة.

- **دالة `groupby()`:** من أهم دوال التحليل. تقوم بتقسيم البيانات إلى مجموعات بناءً على عمود معين (مثل المدينة `City` أو المنتج `Prod`)، ثم نُطبق دالة حسابية عليها (كالمجموع `sum` أو المتوسط `mean`).
  - مثال: للإجابة على "أي مدينة حققت أعلى مبيعات إجمالية؟"، نقوم بتجميع البيانات بالمدينة وحساب مجموع المبيعات وترتيبها.
- **الدمج المنطقي (Logical Filtering):** القدرة على عزل جزء محدد من البيانات باستخدام الشروط المنطقية (`&` لـ "و"، و `|` لـ "أو"). مثل: "أريد فقط المبيعات التي تمت في الرياض **و** كانت من فئة الألعاب".

---

5. التحليل الإحصائي الاستدلالي (Inferential Statistics)

المحلل المحترف لا يكتفي بالنظر إلى المتوسطات فقط، بل يسأل: "هل هذا الفارق حقيقي ومؤثر، أم أنه مجرد صدفة؟"

- **اختبار t للعينات المستقلة (Independent T-Test):** يُستخدم لمقارنة متوسطات مجموعتين مستقلتين (مثل أرباح الذكور مقابل أرباح الإناث) لمعرفة ما إذا كان الفرق بينهما ذو "دلالة إحصائية" (Statistically Significant).
- **قيمة p (p-value):** هي المؤشر الأساسي في هذا الاختبار.
  - إذا كانت $p \le 0.05$: نرفض فرضية العدم (Null Hypothesis)، ونقول إن الفرق حقيقي وجوهري.
  - إذا كانت $p > 0.05$: نقبل فرضية العدم، مما يعني أن الفرق الظاهري هو مجرد صدفة أو تذبذب طبيعي في البيانات، ولا يمكن الاعتماد عليه كقاعدة.

---

6. التمثيل البصري (Data Visualization)

"الصورة تساوي ألف كلمة". تحويل الجداول المعقدة إلى رسوم بيانية يساعد على اكتشاف الأنماط وإيصال الفكرة بسرعة للإدارة.

- **الرسم الشريطي (Bar Chart):** ممتاز لمقارنة الكميات بين الفئات المختلفة (مثلاً مقارنة المبيعات بين المدن).
- **الرسم الشريطي المجمع (Grouped Bar Chart):** يسمح بإضافة بُعد إضافي، مثل عرض مبيعات المدن، ومقسمة داخلياً حسب نوع المنتج.
- **الرسم الدائري (Pie/Donut Chart):** لبيان الحصص والنسب المئوية (مثل نسبة مساهمة كل منتج في الأرباح الكلية).
- **الرسوم التفاعلية (Interactive Charts):** أدوات مثل Plotly تتيح إنشاء لوحات تحكم (Dashboards) تسمح للمستخدم بالفلترة (عبر القوائم المنسدلة)، والتكبير، ومشاهدة التفاصيل الدقيقة عند تحريك الفأرة، وهي أداة احترافية جداً في عروض الأعمال.