البرمجيات والذكاء الاصطناعي

بايثون للأبحاث

حوّل بيانات خامًا فوضوية إلى نتيجة واضحة يمكن الدفاع عنها — المهارات المحددة في Python التي تفصل بين تحليل بحثي وسكربت بسيط.

المستوى: مبتدئ إلى متوسط **مدة التنفيذ:** عطلتا نهاية أسبوع **التكلفة التقريبية:** $0

قبل البدء: يفترض هذا المخطط أنك تستطيع بالفعل كتابة Python الأساسية (المتغيرات، الحلقات، الدوال) — راجع SA-001 إذا لم يكن ذلك مريحًا بعد. هذا المخطط يتناول استخدام Python على بيانات حقيقية، لا تعلّم اللغة من الصفر.

01 الهدف

كتابة Python تعمل ليست نفس مهارة كتابة Python تنتج نتيجة بحثية يمكن لشخص آخر أن يثق بها. للكود البحثي أولويات مختلفة عن سكربت تكتبه لنفسك: كل خطوة يجب أن تكون قابلة لإعادة الإنتاج، وكل تحويل للبيانات يجب أن يكون ظاهرًا لا يدويًا، وكل رسم بياني يجب أن يصمد أمام سؤال "كيف بالضبط حصلت على هذا الرقم؟". بنهاية هذا المخطط، ستكون قادرًا على أخذ مجموعة بيانات خامة وفوضوية وتحويلها إلى نتيجة منظّفة ومحلَّلة ومرسومة بيانيًا — مع أثر واضح يوضّح بالضبط كيف وصلت إليها.

02 المتطلبات الأساسية

- راحة مع Python الأساسية: المتغيرات، الحلقات، الدوال، وتشغيل سكربت أو دفتر ملاحظات
- مجموعة بيانات للتدرب عليها — بيانات عامة من Kaggle أو بيانات ميدانية/مخبرية خاصة بك تعمل بنفس الكفاءة
- لا حاجة لخلفية إحصائية للبدء، رغم أنها تساعد عند وصولك إلى القسم ٠٦

03 الأدوات الموصى بها

الأداة	الغرض	ملاحظات
Jupyter Notebook / JupyterLab	بيئة تفاعلية لاستكشاف البيانات خطوة بخطوة	مجاني — ثبّته عبر `pip install jupyterlab`، أو استخدم Google Colab دون تثبيت
pandas	تحميل البيانات الجدولية وتنظيفها وإعادة تشكيلها	مجاني — المكتبة الأساسية لمعظم العمل على البيانات البحثية
NumPy	العمليات العددية وحسابات المصفوفات	مجاني — pandas مبنية فوقها
Matplotlib / Seaborn	الرسم البياني والتصور	مجاني — Seaborn يعطي رسومًا افتراضية أنظف بكود أقل
SciPy	الاختبارات الإحصائية ودوال الحوسبة العلمية	مجاني — ضرورية بمجرد تجاوز الإحصاء الوصفي
Git	التحكم بإصدارات كود التحليل (وليس بياناتك الخام)	مجاني — انظر ملاحظة القسم ٠٤ حول ما ينتمي إلى التحكم بالإصدارات

04 الأدوات المطلوبة

- Python 3.10 أو أحدث مع pip، أو حساب Google Colab (دون حاجة للتثبيت)
- محرر أكواد أو واجهة Jupyter

قابلية إعادة الإنتاج وزلة البيانات

لا تُعدّل أبدًا ملف بيانات خام يدويًا في Excel أو محرر نصوص — إذا احتجت إلى إصلاح أو تصفية شيء، افعل ذلك بالكود، بحيث يكون التحويل الدقيق ظاهرًا وقابلًا للتكرار. أبقِ ملفك الخام الأصلي دون تغيير واكتب نسختك المنظّفة إلى ملف جديد. هذه العادة الوحيدة هي ما يسمح لشخص آخر — بما في ذلك أنت، بعد ستة أشهر — بالتحقق بالضبط من كيفية إنتاج رقم ما، وهي الفرق بين نتيجة وادّعاء.

05 سير عمل البيانات البحثية

يتبع التحليل البحثي في Python خط أنابيب ثابتًا، بغض النظر عن المجال. إبقاء هذه المراحل منفصلة — بدلًا من التنظيف والتحليل والرسم البياني جميعها في كتلة واحدة متشابكة — هو ما يجعل التحليل سهل التحقق وسهل الإعادة عند وصول بيانات جديدة.

شكل ١ — المراحل الخمس لتحليل بحثي قابل لإعادة الإنتاج في PYTHON

المرحلة	المهام النموذجية	المكتبة الأساسية
التنظيف	معالجة القيم المفقودة، إصلاح الأنواع، إزالة التكرارات	pandas
التحليل	الإحصاء الوصفي، التجميع، اختبارات الفرضيات	NumPy / SciPy
التصور البياني	رسوم التوزيع، خطوط الاتجاه، المقارنات	Matplotlib / Seaborn
التقرير	خلايا markdown تشرح كل قرار ونتيجة	Jupyter Notebook

06 سير العمل خطوة بخطوة

1. **حمّل البيانات الخام دون تعديلها.** اقرأها مباشرة في pandas عبر `pd.read_csv()` — يبقى الملف الأصلي دون تغيير على القرص.
2. **افحص قبل أن تنظّف.** شغّل `.info()` و`.describe()` و`.head()` لفهم ما تتعامل معه فعليًا قبل تغيير أي شيء.
3. **عالج البيانات المفقودة أو المشوّهة بوضوح.** قرّر عمدًا إن كنت ستحذف القيم المفقودة أو تملؤها أو تُشير إليها — واكتب السبب في خلية markdown، لا في ذهنك فقط.
4. **احسب الإحصاء الوصفي أولًا.** المتوسط، الوسيط، الانحراف المعياري، والعدد حسب المجموعة — قبل اللجوء لأي اختبار متقدم، تأكد من فهمك للشكل الأساسي للبيانات.
5. **ارسم بيانيًا قبل أن تستنتج.** غالبًا ما يكشف مدرّج تكراري أو رسم تشتت قيمًا شاذة أو أنماطًا تخفيها الإحصائيات الموجزة وحدها.
6. **شغّل الاختبار الإحصائي المناسب، إذا تطلّب سؤالك ذلك.** اختبار t أو اختبار مربع كاي في SciPy يتطلب سطرًا واحدًا من الكود — الجزء الأصعب هو اختيار الاختبار الصحيح لبياناتك وسؤالك.
7. **وثّق كل خطوة في خلايا markdown أثناء العمل.** اشرح ما تفعله كل خلية كود ولماذا، بلغة واضحة — عامل الدفتر كسجل مختبري، لا مجرد سكربت.
8. **احفظ البيانات المنظّفة بشكل منفصل عن البيانات الخام.** صدّرها إلى ملف جديد مثل `data_cleaned.csv`، دون الكتابة فوق الأصل أبدًا.

07 شرح الكود

يستعرض هذا مرحلتي التنظيف والتحليل على مجموعة بيانات صغيرة بأسلوب بحثي — لاحظ أن كل خطوة صريحة وظاهرة، بدلًا من سطر واحد كثيف يقوم بخمسة أشياء دفعة واحدة.

```
import pandas as pd
import numpy as np
from scipy import stats

# 1. Load raw data — never modify this file directly
raw = pd.read_csv("survey_responses_raw.csv")

# 2. Inspect before touching anything
print(raw.info())
print(raw.describe())

# 3. Handle missing values explicitly, with a documented decision
#    Decision: drop rows missing the outcome variable; fill missing age with the median
cleaned = raw.dropna(subset=["response_score"]).copy()
cleaned["age"] = cleaned["age"].fillna(cleaned["age"].median())

# 4. Descriptive statistics by group
summary = cleaned.groupby("group")["response_score"].agg(["mean", "std", "count"])
print(summary)

# 5. A statistical test comparing two groups
group_a = cleaned[cleaned["group"] == "A"]["response_score"]
group_b = cleaned[cleaned["group"] == "B"]["response_score"]
t_stat, p_value = stats.ttest_ind(group_a, group_b, equal_var=False)
print(f"t = {t_stat:.3f}, p = {p_value:.4f}")

# 6. Save cleaned data separately — raw file stays untouched
cleaned.to_csv("survey_responses_cleaned.csv", index=False)
```

ملاحظة: `equal_var=False` يشغّل اختبار Welch لـ t، الذي لا يفترض تساوي التباين بين المجموعتين — خيار افتراضي أكثر أمانًا من اختبار t القياسي ما لم تتحقق من ذلك الافتراض.

08 الأخطاء الشائعة

الخطأ	لماذا يضر	الحل
تعديل ملف البيانات الخام يدويًا	يجعل التحليل مستحيل التحقق أو الإعادة عند وصول بيانات جديدة	حوّل البيانات فقط عبر الكود، واكتب دائمًا إلى ملف جديد

الخطأ	لماذا يضر	الحل
تشغيل خلايا الدفتر بترتيب غير متسلسل	ينتج نتائج تعتمد بصمت على تاريخ تنفيذ مخفي	استخدم بانتظام "إعادة تشغيل النواة وتشغيل الكل" للتأكد من عمل الدفتر من البداية للنهاية
حذف القيم المفقودة دون التحقق من سبب فقدانها	قد يحيّز النتائج بصمت إذا لم يكن الفقدان عشوائيًا	تحقق مما إذا كانت القيم المفقودة تتركز حول مجموعة أو شرط معين قبل حذفها
اختيار اختبار إحصائي قبل فهم شكل البيانات	العديد من الاختبارات تفترض شكل توزيع (كالتوزيع الطبيعي) قد لا تملكه بياناتك	ارسم بيانيًا أولًا — مدرّج تكراري يتطلب سطرًا واحدًا وقد يجنّبك اختبارًا خاطئًا
رسوم بيانية دون تسميات محاور أو وحدات	رسم بياني لا يستطيع المراجع تفسيره عديم الفائدة عمليًا، مهما كان التحليل الأساسي صحيحًا	سمِّ كل محور، وكل وحدة، وكل مفتاح رسم قبل اعتبار الرسم البياني مكتملًا

09 الخطوة التالية

من تحليل نظيف إلى مساهمة بحثية

سير العمل في هذا المخطط هو الأساس الدقيق الذي يُبنى عليه **SA-003 (أساسيات تعلّم الآلة)** — النموذج لا يكون أفضل من البيانات المنظفة والمفهومة التي تغذيه. إذا تضمّن مشروعك بيانات ميدانية أو استبيانية أصلية، اجمع هذا المخطط مع سلسلة البحث والكتابة العلمية (خاصة RS-003، البحث الحاسوبي وعلم البيانات) لمعرفة كيفية كتابة منهجيتك رسميًا، أو مع **ES-001 (تصميم أول مشروع بحثي بيئي)** إذا كانت بياناتك من عمل ميداني.