**اسم الطالبة : حلا فراس الشبول**

**اسم المعلمة : ناديا حامد**

**رقم الوحدة: الوحدة 21**

**عنوان الواجب: تنفيذ حل ذكاء اصطناعي**

**اسم المؤسسة التعليمية:مدرسة الشجرة الثانوية للبنات**

**المقدمة**

يشهد العالم في السنوات الاخيرة تطورا متسارعا في مجال الذكاء الاصطناعي حيث اصبح عنصرا اساسيا في العديد من القطاعات مثل التجارة والصحة والتعليم والصناعة والخدمات المالية ويعود ذلك الى قدرته الكبيرة على تحليل كميات ضخمة من البيانات واكتشاف الانماط والعلاقات بينها مما يساعد المؤسسات على اتخاذ قرارات دقيقة مبنية على بيانات حقيقية بدلا من الاعتماد على التخمين .

في قطاع الاعمال بشكل خاص اصبح الذكاء الاصطناعي اداة مهمة لتحسين الاداء وزيادة الكفاءة التشغيلية ومن اهم تطبيقاته التنبؤ بالمبيعات والايرادات المستقبلية حيث تعتمد الشركات على تحليل البيانات التاريخية للمبيعات والعوامل المؤثرة فيها مثل الاسعار والعروض الترويجية والمواسم وسلوك العملاء ومن خلال تقنيات التعلم الالي يمكن بناء نماذج قادرة على التنبؤ بالقيم المستقبلية بدرجة جيدة من الدقة.

يساعد نظام التنبؤ بالمبيعات الشركات على التخطيط للمخزون بشكل افضل وتحديد استراتيجيات التسويق المناسبة وتقليل الخسائر الناتجة عن نقص المنتجات او زيادتها كما يدعم عملية اتخاذ القرار المالي ومع ذلك فان دقة هذه الانظمة تعتمد بشكل كبير على جودة البيانات المستخدمة ونوع النموذج الرياضي المختار.

**الهدف من المشروع**

يهدف هذا المشروع الى بناء نموذج تعلم الي يعتمد على بيانات تاريخية خاصة بمتجر سوبرماركت من اجل التنبؤ بقيمة المبيعات المستقبلية ودعم عملية اتخاذ القرار داخل الادارة. يساهم هذا النموذج في مساعدة الادارة على التخطيط للمخزون بشكل افضل وتحديد الفترات ذات الطلب المرتفع وتحسين استراتيجيات التسويق بناء على توقعات دقيقة مبنية على بيانات فعلية.

يعتمد الحل على التعلم الموجه من نوع الانحدار حيث يتم تدريب النموذج باستخدام بيانات مصنفة مسبقا ثم تقييم ادائه باستخدام مقاييس احصائية مناسبة مثل R2 و MAE و RMSE وذلك لان الهدف هو التنبؤ بقيمة رقمية تمثل المبيعات. يسمح هذا النوع من التعلم باكتشاف العلاقة الرياضية بين المتغيرات المستقلة مثل الفئة والمنطقة وتاريخ الطلب وبين المتغير التابع وهو قيمة المبيعات.

سيتم تنفيذ مراحل العمل بشكل منظم تبدأ بفحص جودة البيانات ومعالجتها ثم تقسيمها الى مجموعات تدريب وتحقق واختبار وبعد ذلك تدريب النموذج ومقارنة ادائه قبل وبعد التحسين لضمان تحقيق معايير النجاح المحددة.

معايير النجاح في هذا المشروع تتمثل في ما يلي

* تحقيق قيمة R2 لا تقل عن 0.80 مما يدل على قدرة النموذج على تفسير ما لا يقل عن 80 بالمئة من التغير في قيم المبيعات.
* الحصول على قيمة MAE منخفضة بحيث لا تتجاوز 15 بالمئة من متوسط قيمة المبيعات وذلك لضمان ان متوسط الخطأ في التنبؤ مقبول عمليا.
* تقليل قيمة RMSE بعد مرحلة تحسين النموذج مقارنة بالنموذج الاولي مما يدل على تحسن دقة التنبؤ.
* تقارب نتائج التدريب والتحقق دون وجود فرق كبير بينهما مما يشير الى عدم حدوث overfitting.
* استقرار النتائج عند تطبيق Cross Validation بحيث لا تظهر فروقات كبيرة بين الطيات المختلفة.

تساعد هذه المعايير على قياس جودة النموذج بشكل كمي وواضح وتسمح بالحكم على مدى تحقيق الحل للاهداف المخطط لها بطريقة علمية دقيقة وليست وصفية فقط.

**نوع التعلم المستخدم**

يعتمد هذا المشروع على التعلم الموجه من نوع الانحدار Regression وذلك لان الهدف الاساسي هو التنبؤ بقيمة رقمية تمثل حجم المبيعات المستقبلية ويستخدم التعلم الموجه عندما تكون البيانات تحتوي على مدخلات معروفة بالاضافة الى مخرجات فعلية حيث يتعلم النموذج العلاقة بين المتغيرات المستقلة مثل السعر والعروض وعدد العملاء وغيرها وبين المتغير التابع وهو قيمة المبيعات.

يقوم النموذج اثناء مرحلة التدريب بتحليل البيانات التاريخية واكتشاف الانماط والعلاقات الرياضية بينها ومن ثم استخدام هذه العلاقات للتنبؤ بقيم جديدة لم يسبق له رؤيتها ويعد اسلوب الانحدار مناسبا في هذا النوع من المشكلات لانه مصمم للتعامل مع التنبؤ بالقيم العددية بدقة اعتمادا على البيانات السابقة.

**مصدر البيانات**

تم الحصول على البيانات المستخدمة في هذا المشروع من منصة Kaggle وهي منصة متخصصة توفر مجموعات بيانات حقيقية تستخدم في مشاريع تحليل البيانات والتعلم الالي وتم اختيار مجموعة بيانات خاصة بمبيعات سوبرماركت تحتوي على سجلات تاريخية للمبيعات خلال فترة زمنية محددة.

تم اختيار هذه البيانات لانها تحتوي على معلومات تفصيلية عن المبيعات بالاضافة الى مجموعة من المتغيرات المؤثرة مثل نوع المنتج السعر الكمية تاريخ الشراء وخصائص اخرى مرتبطة بعملية البيع مما يجعلها مناسبة لبناء نموذج تنبؤ يعتمد على تحليل العلاقات بين هذه المتغيرات وقيمة المبيعات كما ان البيانات منظمة وجاهزة للاستخدام مما يسهل عملية المعالجة وبناء النموذج بصورة صحيحة.

من المهم أيضا النظر في احتمالية وجود تحيز في البيانات المستخدمة بما أن البيانات مأخوذة من متجر واحد وعلى فترة زمنية محددة فقد لا تمثل جميع أنماط المبيعات في قطاعات أو مناطق مختلفة. كما أن توزيع الفئات أو المناطق قد لا يكون متوازنا بالكامل مما قد يؤثر على قدرة النموذج على التعميم خارج نطاق هذه البيانات.

هذا يعني أن جودة النتائج مرتبطة بشكل مباشر بمدى تمثيل البيانات للواقع الفعلي لذلك فإن أي قيود في تنوع البيانات قد تؤثر على دقة التنبؤ في حالات جديدة لم تكن ممثلة بشكل كاف في مجموعة التدريب.

**فحص جودة البيانات**

يعد فحص جودة البيانات خطوة اساسية قبل بناء اي نموذج ذكاء اصطناعي لان دقة النموذج تعتمد بشكل مباشر على جودة البيانات المستخدمة في تدريبه فالبيانات غير النظيفة او غير المكتملة قد تؤدي الى نتائج غير دقيقة او مضللة كما يمكن ان تسبب انخفاضا في اداء النموذج او حدوث مشكلة التحيز.

لذلك يتم فحص البيانات للتأكد من خلوها من القيم المفقودة والقيم الشاذة والسجلات المكررة ولدراسة توزيع المتغيرات وفهم طبيعتها ويساعد هذا الفحص على اكتشاف المشكلات مبكرا ومعالجتها قبل مرحلة التدريب مما يساهم في بناء نموذج اكثر استقرارا ودقة في التنبؤ بالمبيعات.

تم استخدام منصة Google Colab لتنفيذ الكود وفحص البيانات ومعالجتها حيث توفر بيئة برمجية تعتمد على لغة بايثون وتحتوي على مكتبات متخصصة في تحليل البيانات والتعلم الالي مثل Pandas و NumPy و Matplotlib و Scikit learn مما يسهل عملية الاستكشاف والتحليل وبناء النموذج .

**استيراد المكتبات البرمجية**

في بداية تنفيذ المشروع تم استيراد مجموعة من المكتبات البرمجية الاساسية باستخدام لغة بايثون وذلك لتجهيز بيئة العمل في Google Colab قبل البدء في تحليل البيانات وبناء نموذج التنبؤ بالمبيعات

مكتبة( pandas)  
تستخدم مكتبة pandas في قراءة ملفات البيانات مثل CSV وتحويلها الى DataFrame مما يسمح بالتعامل مع البيانات على شكل جداول منظمة كما توفر ادوات قوية لتنظيف البيانات ومعالجة القيم المفقودة واعادة تسمية الاعمدة واختيار المتغيرات المطلوبة وتنفيذ عمليات التجميع والتحليل الاحصائي بسهولة.

مكتبة( numpy)  
تستخدم مكتبة numpy لاجراء العمليات الحسابية المتقدمة والتعامل مع المصفوفات العددية حيث تعتمد العديد من خوارزميات التعلم الالي على العمليات الرياضية التي توفرها هذه المكتبة مثل الجذور التربيعية والحسابات الاحصائية والتحويلات العددية.

مكتبة (matplotlib.pyplot)  
تستخدم هذه المكتبة في انشاء الرسوم البيانية الاساسية مثل المخططات الخطية ومخططات الاعمدة والمخططات التوزيعية وتساعد في تمثيل البيانات بصريا لفهم الانماط العامة داخل بيانات المبيعات.

مكتبة( seaborn)  
تعد seaborn مكتبة متقدمة مبنية على matplotlib وتستخدم لانشاء رسوم بيانية اكثر احترافية مثل مخططات التشتت وتوزيع القيم وخريطة الارتباط بين المتغيرات مما يساعد في اكتشاف العلاقات بين الخصائص المختلفة.

مكتبة( sklearn.pipeline )  
تستخدم لانشاء Pipeline والذي يسمح بدمج عدة خطوات مثل التقييس والتدريب داخل مسار واحد منظم مما يضمن تطبيق نفس المعالجة على بيانات التدريب والاختبار بشكل تلقائي ويقلل من الاخطاء.

مكتبة( sklearn.model\_selection)  
تستخدم لاجراء عملية تقسيم البيانات الى بيانات تدريب واختبار باستخدام train\_test\_split وذلك لتقييم اداء النموذج على بيانات لم يتم تدريب النموذج عليها.

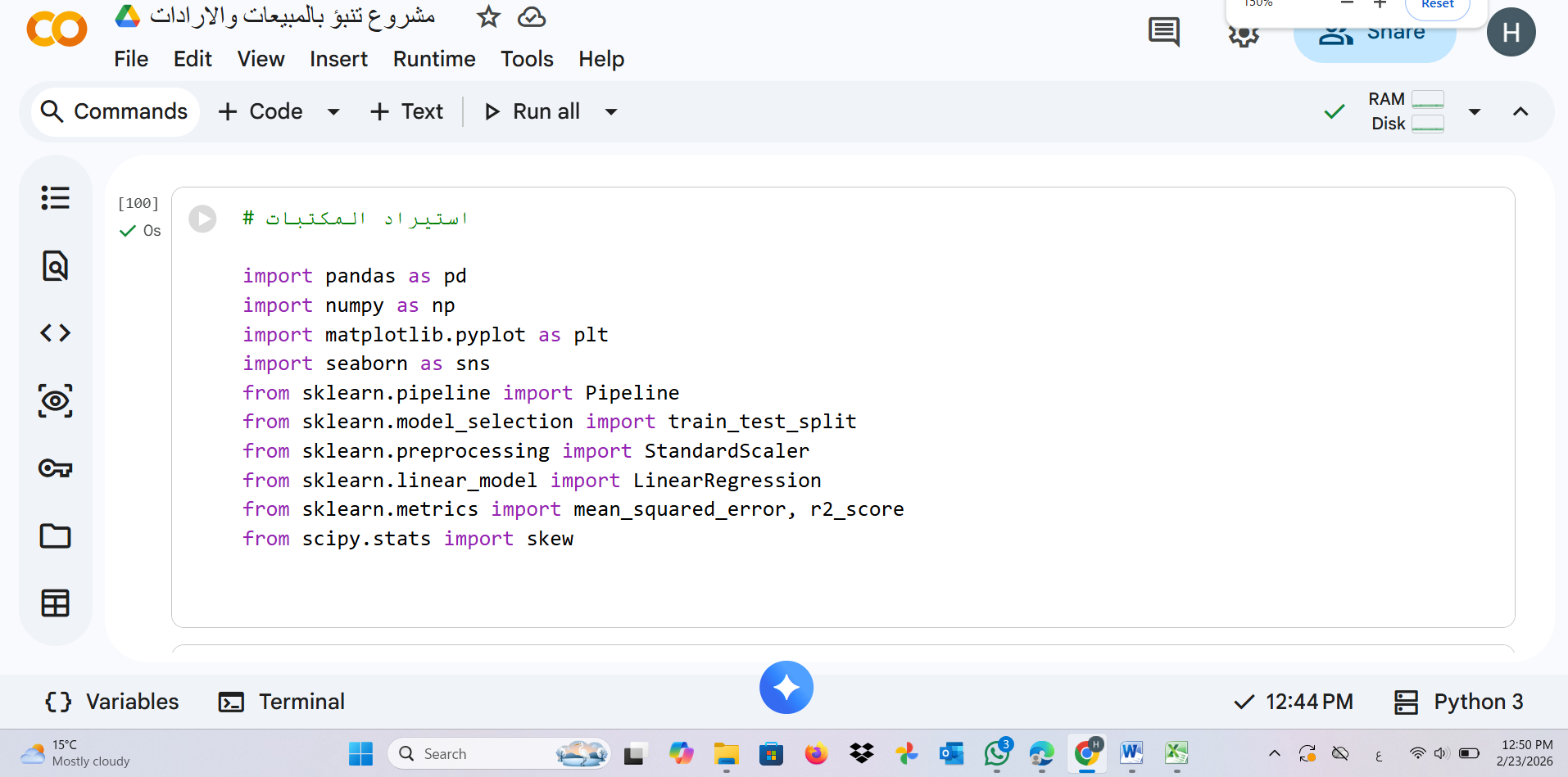
مكتبة( sklearn.preprocessing)  
تم استيراد StandardScaler منها والذي يستخدم لتوحيد مقياس المتغيرات العددية بحيث يصبح متوسطها صفرا والانحراف المعياري يساوي واحد وهي خطوة مهمة عند استخدام نماذج الانحدار.

مكتبة(sklearn.linear\_model)  
تحتوي على نموذج LinearRegression الذي تم استخدامه لبناء نموذج الانحدار الخطي للتنبؤ بقيمة المبيعات.

مكتبة( sklearn.metrics)  
تستخدم لحساب مقاييس تقييم النموذج مثل Mean Squared Error و R2 Score والتي تساعد في قياس دقة التنبؤ ومقارنة اداء النماذج المختلفة.

مكتبة( scipy.stats)  
تم استيراد الدالة skew منها لحساب معامل الانحراف وذلك لفحص توزيع البيانات والتأكد مما اذا كانت البيانات منحرفة وتحتاج الى معالجة قبل التدريب.

توضح لقطة الشاشة التالية تنفيذ خلية استيراد المكتبات في بيئة Google Colab حيث يظهر الكود المستخدم لتحميل جميع الادوات البرمجية اللازمة لتنفيذ مراحل المشروع المختلفة من تنظيف البيانات الى بناء النموذج وتقييمه.



**تحميل البيانات**

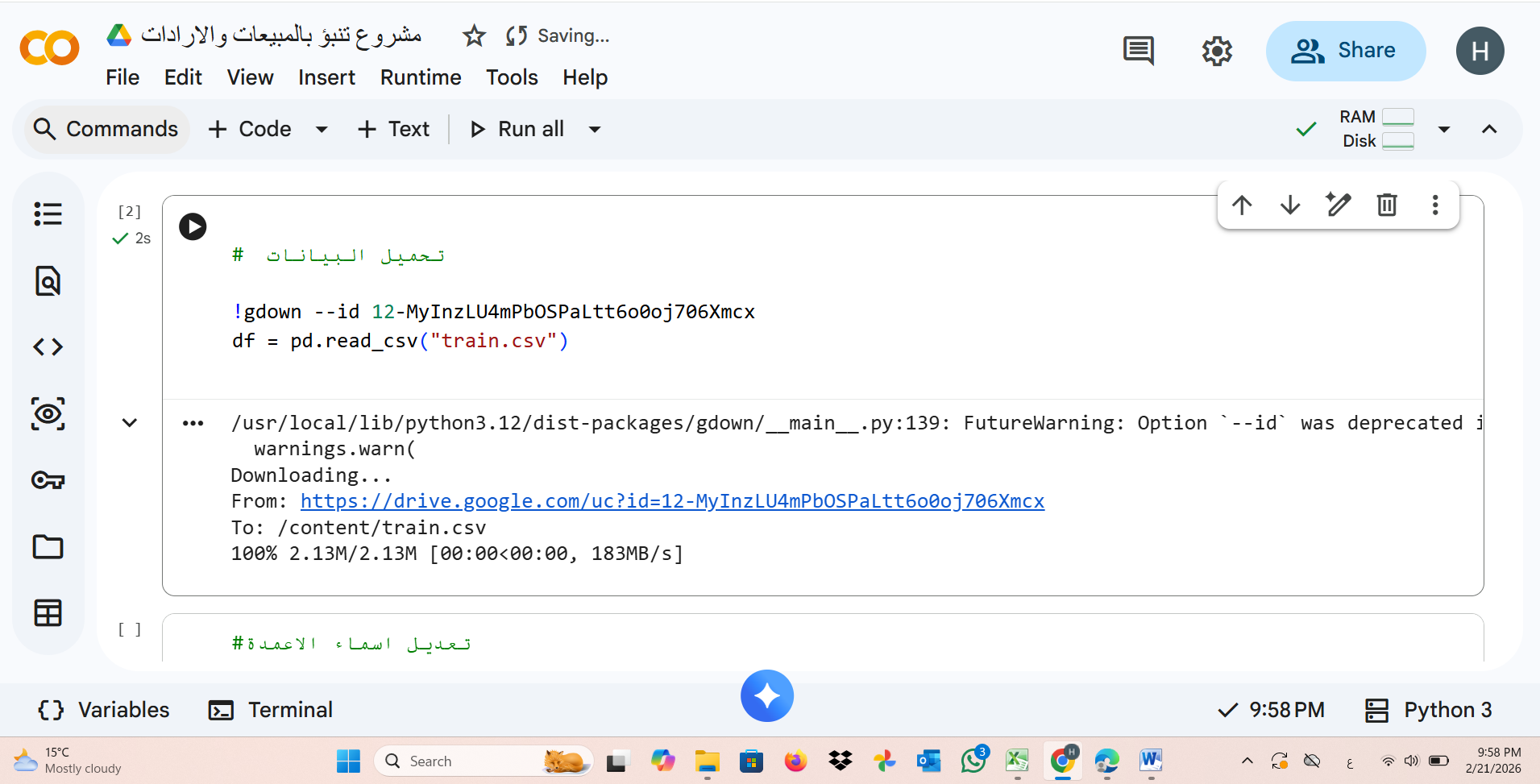
بعد استيراد المكتبات تم الانتقال الى خطوة تحميل البيانات داخل بيئة Google Colab وذلك باستخدام اداة gdown لتحميل ملف البيانات من Google Drive ثم قراءته باستخدام مكتبة pandas.

تم تنفيذ الامر gdown لتحميل ملف train.csv من رابط مباشر في Google Drive وبعد اكتمال عملية التحميل تم استخدام الدالة pd.read\_csv لقراءة الملف وتحويله الى اطار بيانات DataFrame يمكن التعامل معه وتحليله داخل بايثون.

وتعد هذه الخطوة اساسية لان النموذج لا يمكن تدريبه دون توفر البيانات الفعلية حيث تمثل البيانات المصدر الرئيسي الذي يتعلم منه النموذج العلاقات بين المتغيرات المختلفة وقيمة المبيعات كما ان استخدام صيغة CSV يسهل عملية القراءة والمعالجة نظرا لكونها من اكثر الصيغ استخداما في تحليل البيانات

وتعتبر خطوة تحميل البيانات وقراءتها خطوة محورية في المشروع حيث تضمن نقل البيانات بشكل صحيح الى بيئة العمل دون فقدان او تشويه كما ان التأكد من نجاح التحميل بنسبة مئة بالمئة يدل على سلامة الملف وجاهزيته للتحليل وبناء نموذج التنبؤ بالمبيعات

**لقطة شاشة**



توضح لقطة الشاشة التالية عملية تحميل ملف البيانات من Google Drive باستخدام gdown ونجاح قراءته داخل Google Colab حيث يظهر اكتمال التحميل وتحويل الملف الى اطار بيانات جاهز للفحص والمعالجة في المراحل التالية من المشروع.

**تعديل واظهار اسماء الاعمدة**

بعد تحميل البيانات وقراءتها تم الانتقال الى خطوة مهمة وهي التأكد من صحة اسماء الاعمدة داخل ملف البيانات وذلك قبل البدء في اي عملية تحليل او معالجة.

في البداية تم استخدام الامر  
df.columns = df.columns.str.strip()  
وذلك لازالة اي مسافات زائدة قد تكون موجودة قبل او بعد اسم العمود حيث ان وجود مسافات غير ظاهرة قد يسبب اخطاء عند استدعاء العمود لاحقا في الكود كما قد يؤدي الى مشاكل اثناء التدريب او التحليل

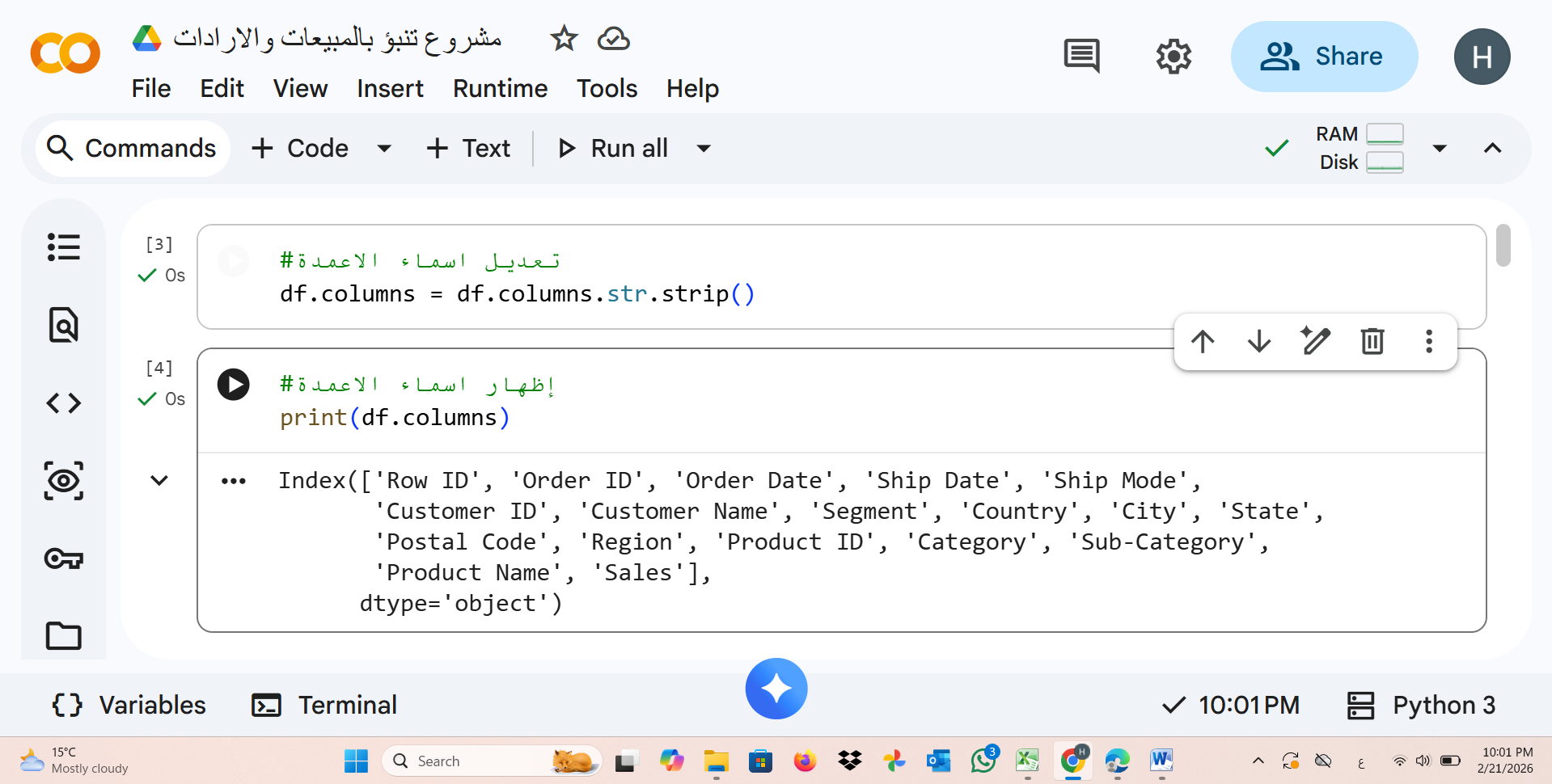
بعد ذلك تم استخدام الامر  
print(df.columns)  
لعرض جميع اسماء الاعمدة والتأكد من انها ظهرت بشكل صحيح ومنظم.

اظهرت النتائج ان البيانات تحتوي على مجموعة من المتغيرات المهمة مثل  
1- Order Date  
2- Ship Date   
 3- Category  
 4- Sub Category   
 5- Region  
6 -Sales

ويعد عمود Sales هو المتغير التابع الذي سيتم التنبؤ به بينما بقية الاعمدة تمثل المتغيرات المستقلة التي قد تؤثر على قيمة المبيعات.

وتعتبر خطوة فحص وتعديل اسماء الاعمدة خطوة تنظيمية مهمة تضمن سلامة التعامل مع البيانات خلال المراحل اللاحقة كما تساعد على فهم بنية البيانات وتحديد المتغير المستهدف والمتغيرات المؤثرة بشكل واضح مما يسهل عملية بناء نموذج تنبؤ دقيق

**لقطة شاشة**



توضح لقطة الشاشة التالية تنفيذ اوامر تعديل اسماء الاعمدة واظهارها داخل Google Colab حيث يظهر الكود المستخدم والنتائج التي تبين جميع المتغيرات الموجودة في ملف البيانات والتي سيتم الاعتماد عليها في المراحل التالية من المشروع.

**عرض اول صفوف من البيانات**

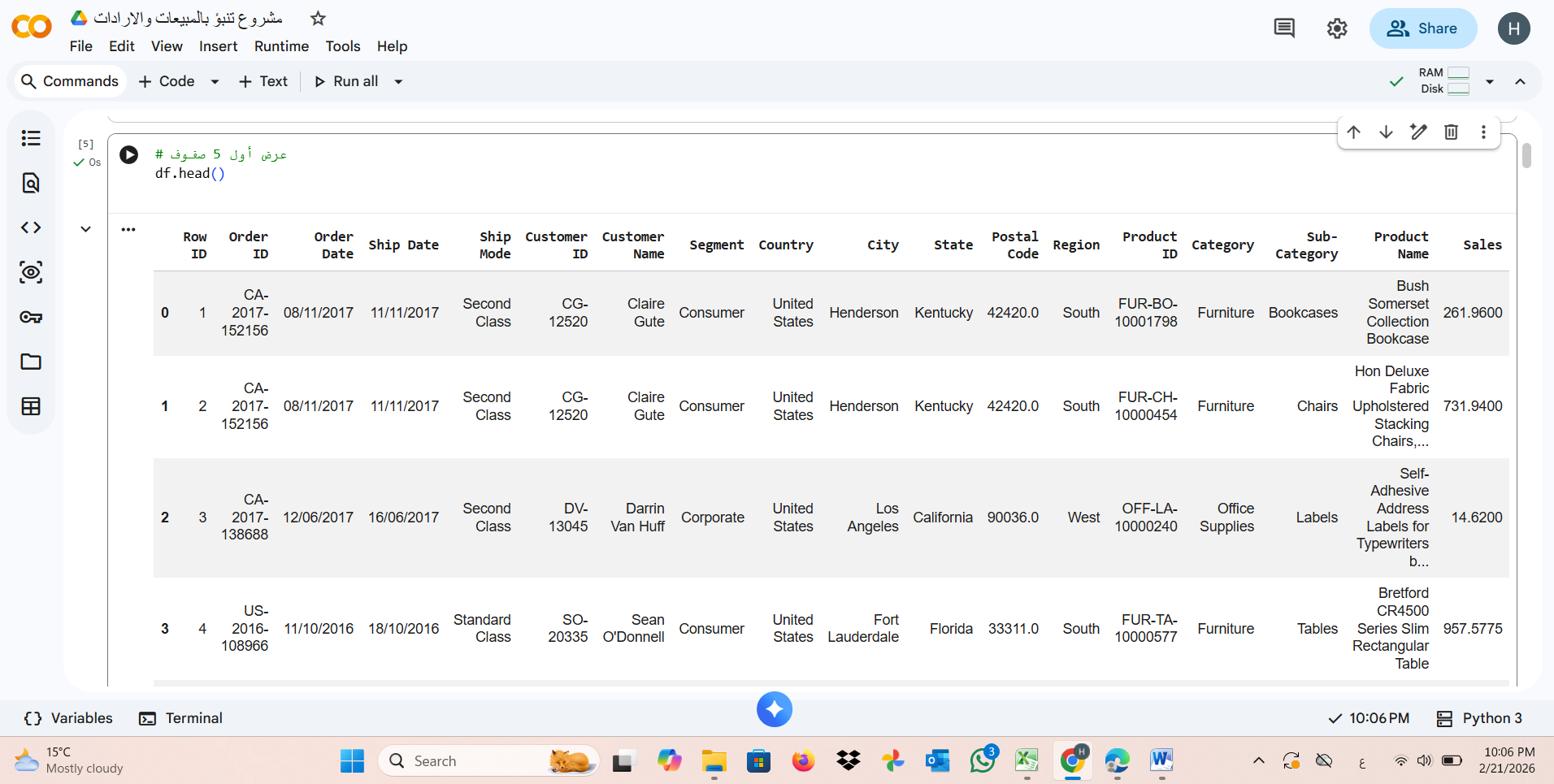
بعد التأكد من اسماء الاعمدة تم استخدام الامر  
df.head()  
لعرض اول خمسة صفوف من مجموعة البيانات وذلك بهدف الاطلاع المبدئي على شكل البيانات والتأكد من انها تم تحميلها بشكل صحيح

تساعد هذه الخطوة على فهم بنية البيانات والتعرف على نوع المتغيرات الموجودة سواء كانت نصية او رقمية كما تسمح بملاحظة طبيعة القيم داخل كل عمود مثل التواريخ اسماء العملاء الفئات الجغرافية وقيمة المبيعات

من خلال النتائج الظاهرة يمكن ملاحظة ان البيانات تحتوي على معلومات تفصيلية تشمل رقم الطلب تاريخ الطلب طريقة الشحن بيانات العميل الدولة المدينة الفئة نوع المنتج واخيرا عمود Sales الذي يمثل قيمة المبيعات وهو المتغير المستهدف الذي سيتم التنبؤ به في هذا المشروع

وتعتبر خطوة عرض اول الصفوف خطوة استكشافية مهمة جدا ضمن مرحلة فحص جودة البيانات حيث تتيح التأكد من عدم وجود مشاكل واضحة في تنسيق البيانات كما تساعد في تحديد الاعمدة غير الضرورية التي قد يتم حذفها لاحقا مثل Row ID او Order ID اذا لم تكن مؤثرة في التنبؤ كما تساهم في فهم العلاقة المحتملة بين المتغيرات قبل البدء في التحليل الاحصائي او التدريب

**لقطة شاشة**



توضح لقطة الشاشة التالية تنفيذ الامر df.head داخل بيئة Google Colab حيث يظهر اول خمسة صفوف من البيانات بما في ذلك عمود Sales الذي سيتم استخدامه كمتغير تابع في نموذج التنبؤ بالمبيعات.

**عرض ابعاد البيانات**

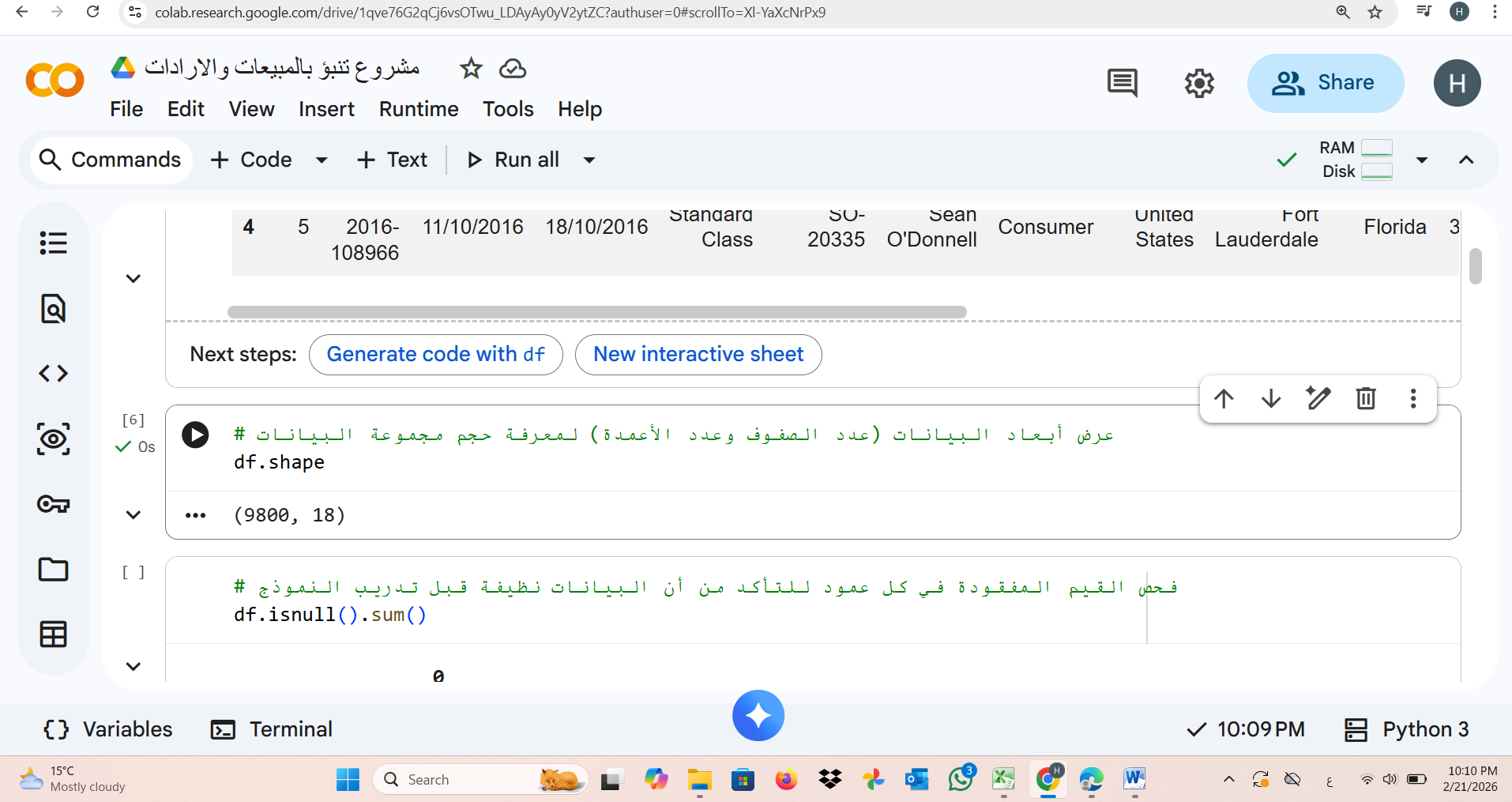
بعد الاطلاع على شكل البيانات تم استخدام الامر  
df.shape  
وذلك لمعرفة ابعاد مجموعة البيانات من حيث عدد الصفوف وعدد الاعمدة

اظهرت النتيجة ان البيانات تحتوي على 9800 صف و 18 عمود وهذا يدل على ان حجم البيانات مناسب لبناء نموذج تعلم الي حيث ان عدد السجلات كاف لاستخراج الانماط والعلاقات بين المتغيرات المختلفة وقيمة المبيعات

وتساعد هذه الخطوة في تكوين تصور عام عن حجم مجموعة البيانات قبل البدء في المعالجة او التقسيم كما تساهم في تقييم ما اذا كانت كمية البيانات كافية لتدريب النموذج وتحقيق نتائج موثوقة

وتعتبر معرفة ابعاد البيانات خطوة اساسية ضمن مرحلة فهم البيانات حيث تتيح التأكد من ان عدد المتغيرات مناسب وان حجم البيانات كاف لدعم عملية التدريب كما تساعد في التخطيط لمرحلة تقسيم البيانات الى تدريب وتحقق واختبار بطريقة صحيحة

لقطة شاشة



توضح لقطة الشاشة التالية تنفيذ الامر df.shape داخل بيئة Google Colab حيث يظهر عدد الصفوف والاعمدة في مجموعة البيانات مما يؤكد جاهزيتها للانتقال الى المراحل التالية من التحليل وبناء نموذج التنبؤ بالمبيعات.

**فحص القيم المفقودة في البيانات**

بعد معرفة ابعاد مجموعة البيانات تم التحقق من وجود قيم مفقودة داخل كل عمود باستخدام الامر

df.isnull().sum()

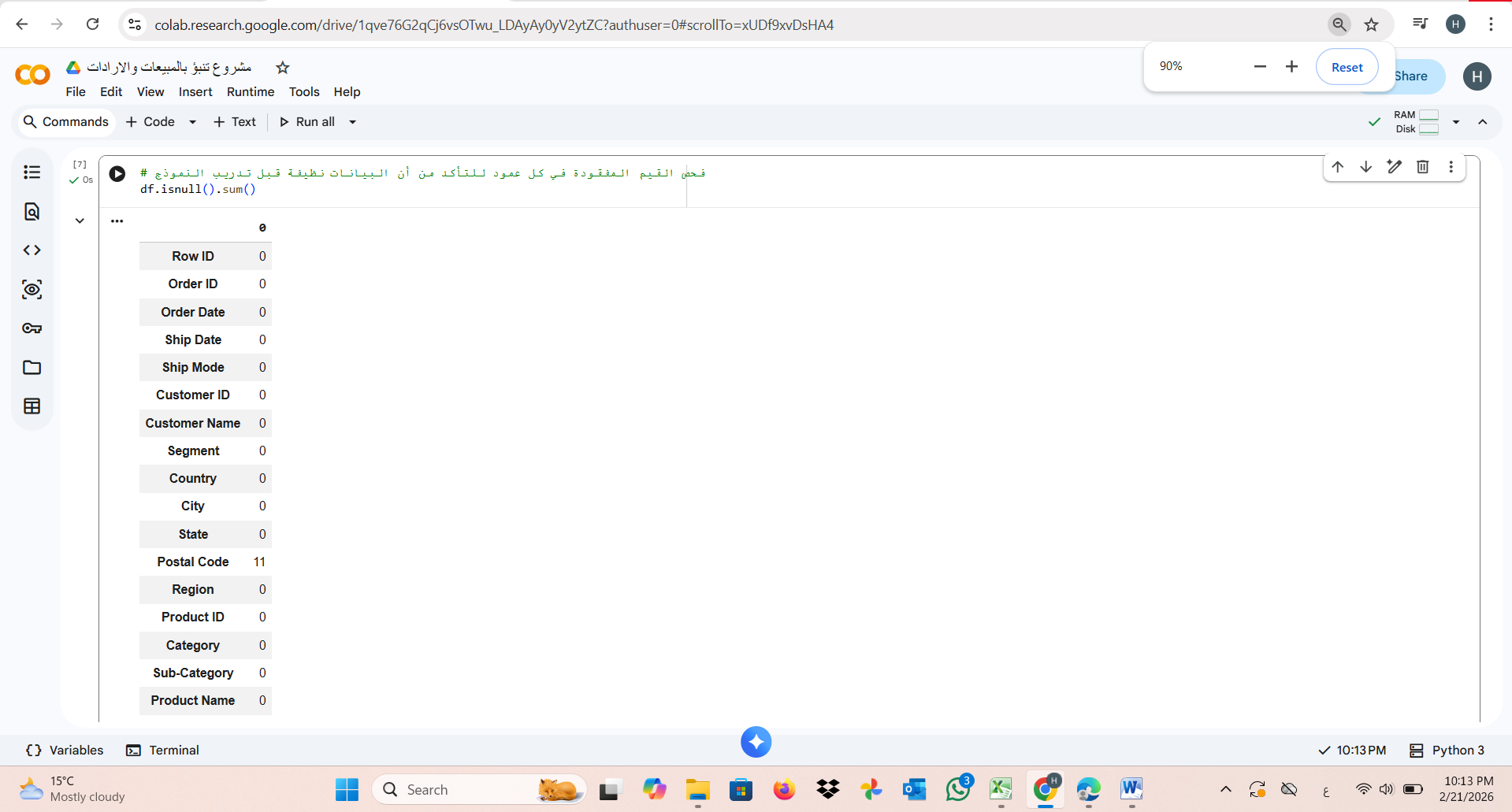
تهدف هذه الخطوة الى التأكد من ان جميع المتغيرات مكتملة قبل البدء في تدريب نموذج التنبؤ بالمبيعات لان وجود بيانات ناقصة قد يؤثر بشكل مباشر على دقة النموذج او يؤدي الى نتائج غير صحيحة

اظهرت النتائج ان جميع الاعمدة تحتوي على صفر من القيم المفقودة باستثناء عمود Postal Code الذي يحتوي على عدد بسيط من القيم المفقودة وهذا يعني ان بقية البيانات نظيفة وجاهزة للاستخدام.

وجود عدد محدود من القيم المفقودة في عمود واحد فقط لا يؤثر بشكل كبير على النموذج خاصة اذا لم يكن هذا العمود من المتغيرات الاساسية في عملية التنبؤ ويمكن لاحقا التعامل معه اما بالحذف او التعويض حسب الحاجة.

وتعتبر عملية فحص القيم المفقودة خطوة اساسية ضمن مرحلة فحص جودة البيانات حيث تساعد على ضمان جاهزية البيانات للتدريب كما تقلل من احتمالية حدوث اخطاء اثناء بناء النموذج وتزيد من موثوقية النتائج النهائية.

**لقطة شاشة**



توضح لقطة الشاشة التالية تنفيذ الامر df.isnull().sum داخل بيئة Google Colab حيث يظهر عدد القيم المفقودة في كل عمود مما يساهم في تقييم جودة البيانات قبل الانتقال الى مرحلة المعالجة وبناء نموذج التنبؤ بالمبيعات.

**حذف الاعمدة غير الضرورية**

بعد فحص البيانات تم الانتقال الى خطوة تنظيف مهمة وهي حذف الاعمدة التي لا تؤثر بشكل مباشر على عملية التنبؤ بالمبيعات وذلك باستخدام الامر.

df = df.drop(columns=["Row ID", "Customer Name", "Order ID", "Postal Code", "Product Name", "Customer ID"], errors="ignore")

تم حذف هذه الاعمدة لانها تمثل معرفات او معلومات وصفية لا تضيف قيمة حقيقية للنموذج من الناحية التنبؤية فعلى سبيل المثال Row ID و Order ID و Customer ID هي ارقام تعريفية فريدة لكل سجل ولا تحمل نمطا يمكن للنموذج التعلم منه كما ان Customer Name و Product Name تعتبر بيانات نصية وصفية لا تعكس سلوكا رقميا مباشرا يؤثر على قيمة المبيعات.

اما عمود Postal Code فقد تم حذفه لانه يمثل موقعا جغرافيا دقيقا يمكن تمثيله بمتغيرات اخرى مثل الدولة او الولاية او المنطقة والتي تعد اكثر فائدة في التحليل.

تساعد هذه الخطوة على تقليل عدد المتغيرات غير المهمة مما يقلل من تعقيد النموذج ويساهم في تحسين الاداء وتقليل خطر حدوث overfitting كما تجعل البيانات اكثر تنظيما وتركيزا على المتغيرات ذات العلاقة الفعلية بقيمة المبيعات.

وتعتبر عملية حذف الاعمدة غير الضرورية خطوة مهمة ضمن مرحلة معالجة البيانات حيث تساهم في تحسين جودة النموذج وزيادة كفاءته كما انها تعكس فهما تحليليا لطبيعة البيانات والتمييز بين المتغيرات المفيدة وغير المفيدة في عملية التنبؤ.

**لقطة شاشة**



توضح لقطة الشاشة التالية تنفيذ امر حذف الاعمدة غير الضرورية داخل Google Colab ونجاح تطبيق التعديل على مجموعة البيانات قبل الانتقال الى مرحلة بناء نموذج التنبؤ بالمبيعات.

**التحقق النهائي من القيم المفقودة بعد التنظيف**

بعد حذف الاعمدة غير الضرورية تم اعادة فحص القيم المفقودة باستخدام الامر

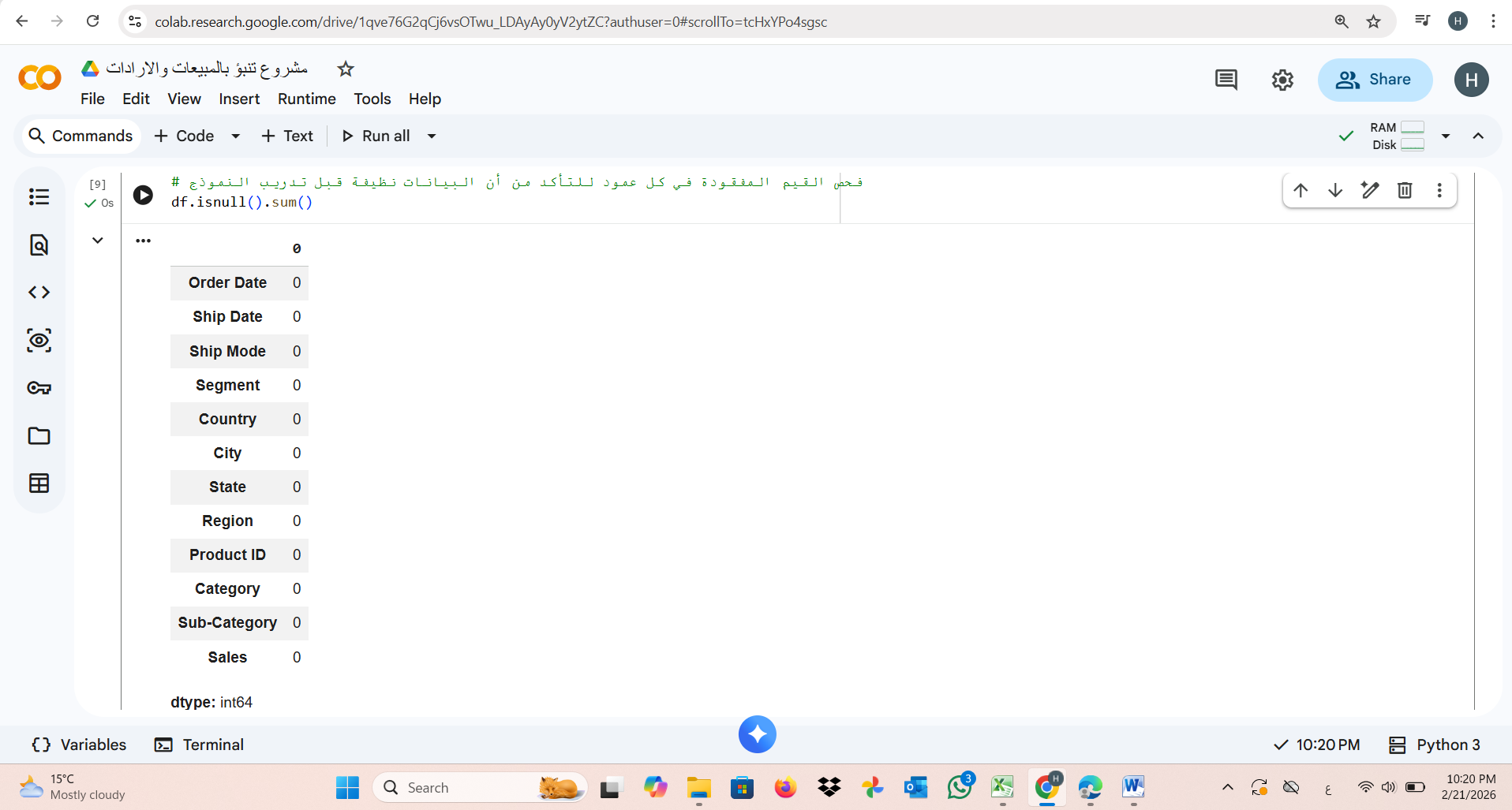
df.isnull().sum()

وذلك للتأكد من ان عملية الحذف لم تؤثر على سلامة البيانات وان مجموعة البيانات اصبحت نظيفة بالكامل قبل الانتقال الى مرحلة بناء النموذج

واظهرت النتائج ان جميع الاعمدة المتبقية تحتوي على صفر من القيم المفقودة بما في ذلك عمود Sales الذي يمثل المتغير المستهدف وهذا يدل على ان البيانات جاهزة للتقسيم والتدريب دون الحاجة الى عمليات تعويض او حذف اضافي.

ويعد التحقق النهائي من القيم المفقودة خطوة مهمة لضمان استقرار النموذج لاحقا حيث ان وجود بيانات ناقصة قد يؤدي الى اخطاء اثناء التدريب او يؤثر سلبا على دقة النتائج وبما ان البيانات اصبحت نظيفة بالكامل فهذا يعزز من موثوقية عملية التنبؤ ويقلل من احتمالية حدوث مشكلات اثناء التدريب

**لقطة شاشة**



توضح لقطة الشاشة التالية تنفيذ الامر df.isnull().sum بعد حذف الاعمدة غير الضرورية حيث تظهر جميع القيم بصفر مما يؤكد ان البيانات جاهزة للانتقال الى مرحلة تقسيم البيانات وبناء نموذج التنبؤ بالمبيعات.

**فصل المتغيرات المستقلة عن المتغير التابع**

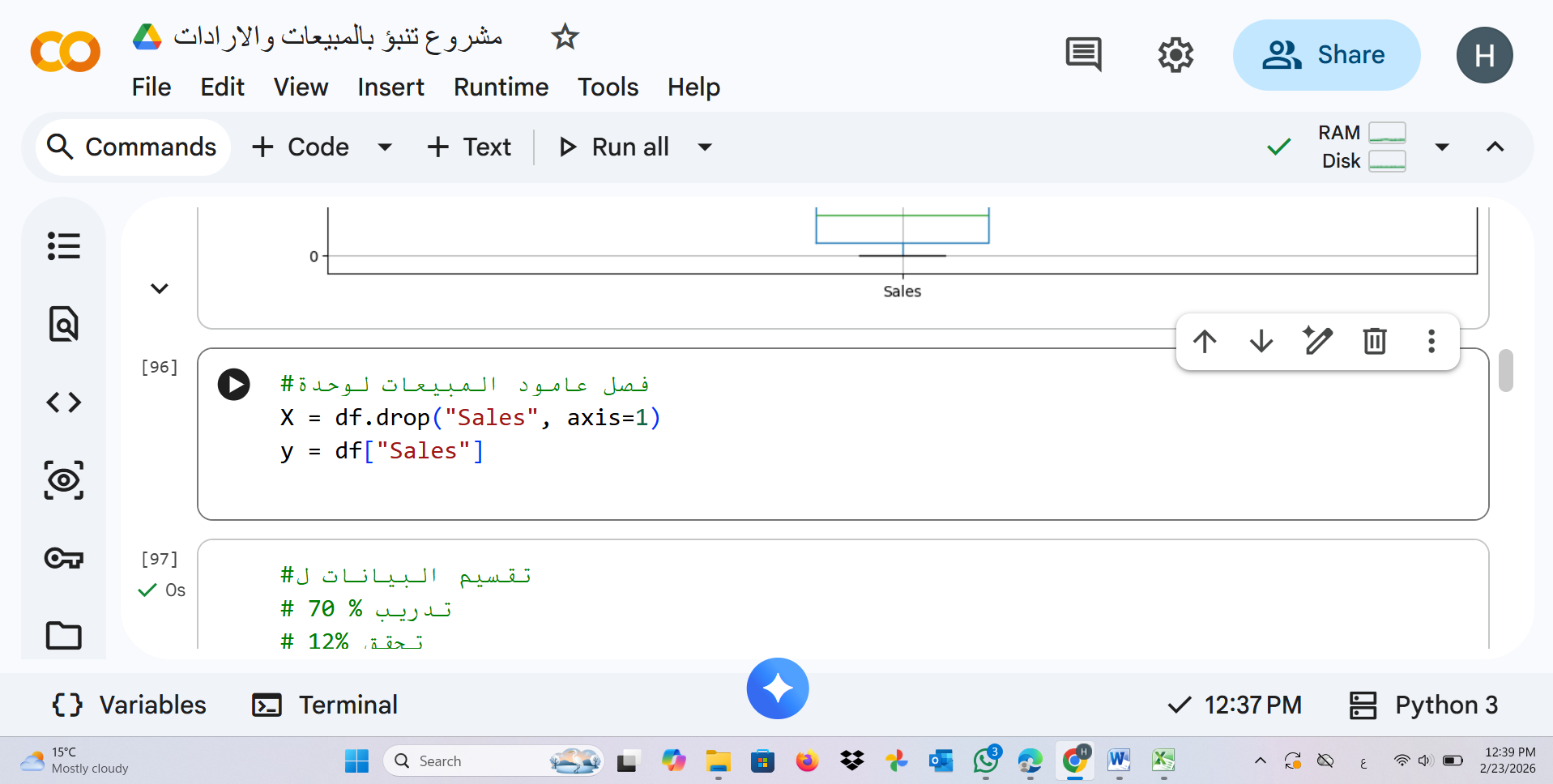
فصل المتغيرات المستقلة عن المتغير التابع خطوة اساسية في بناء نموذج التعلم الالي.

تم تحديد عمود Sales كمتغير تابع y لانه يمثل القيمة التي نهدف الى التنبؤ بها في هذا المشروع بينما تم اعتبار بقية الاعمدة متغيرات مستقلة X لانها تمثل العوامل التي قد تؤثر على قيمة المبيعات مثل الفئة والمنطقة والخصائص الزمنية التي تم استخراجها سابقا.

تعد هذه الخطوة ضرورية من الناحية الرياضية لان نموذج الانحدار يحتاج الى مصفوفة مدخلات تحتوي على الخصائص التفسيرية ومتجه مخرجات يحتوي على القيم الحقيقية التي سيتم استخدامها في حساب الخطأ ومقارنة النتائج المتوقعة بالقيم الفعلية.

فصل البيانات بهذه الطريقة يسمح للنموذج بتعلم العلاقة بين المتغيرات المستقلة وقيمة المبيعات بشكل منظم كما يسهل لاحقا عملية تقسيم البيانات الى مجموعات تدريب وتحقق واختبار لضمان تقييم عادل لاداء النموذج.

توضح لقطة الشاشة التالية تعريف المتغيرات المستقلة X من خلال حذف عمود Sales وتعريف المتغير التابع y الذي يحتوي على قيم المبيعات تمهيدا لمرحلة تقسيم البيانات والتدريب.



**الملخص الاحصائي للبيانات الرقمية**

بعد تنظيف البيانات تم استخدام الامر

df.describe()

وذلك للحصول على ملخص احصائي لعمود Sales الذي يمثل المتغير المستهدف في مشروع التنبؤ بالمبيعات

اظهر الملخص الاحصائي ان عدد القيم يساوي 9800 وهو ما يتطابق مع عدد الصفوف الكلي مما يؤكد اكتمال البيانات كما بلغ متوسط المبيعات حوالي 230 بينما كانت القيمة الوسيطة حوالي 54 وهذا الفرق الكبير بين المتوسط والوسيط يشير الى وجود قيم مرتفعة جدا اثرت على المتوسط.

كما يظهر ان اقل قيمة للمبيعات قريبة من الصفر بينما اعلى قيمة تجاوزت 22000 وهذا يدل على وجود تفاوت كبير بين عمليات البيع الصغيرة والعمليات ذات القيمة العالية كذلك بلغ الانحراف المعياري حوالي 626 وهو رقم مرتفع نسبيا مما يدل على تشتت كبير في قيم المبيعات.

ويساعد الملخص الاحصائي في فهم توزيع المبيعات وطبيعة البيانات قبل تدريب النموذج كما يعطي مؤشرا على احتمالية وجود قيم متطرفة قد تؤثر على بعض نماذج الانحدار مثل الانحدار الخطي وبناء على ذلك يمكن التفكير لاحقا في استخدام نماذج اكثر مرونة مثل Random Forest او Gradient Boosting التي تتعامل بشكل افضل مع التباين العالي في البيانات.

**لقطة شاشة**



توضح لقطة الشاشة التالية تنفيذ الامر df.describe داخل Google Colab حيث يظهر الملخص الاحصائي الكامل لعمود Sales والذي تم استخدامه لتحليل توزيع المبيعات وفهم خصائصها قبل الانتقال الى مرحلة تقسيم البيانات وبناء النموذج.

**التحقق من القيم الشاذة في عمود Sales**

بعد الحصول على الملخص الاحصائي وملاحظة وجود تباين كبير في قيم المبيعات تم الانتقال الى خطوة فحص القيم الشاذة باستخدام طريقة المدى الربيعي IQR.

تم حساب الربع الاول Q1 والربع الثالث Q3 ثم حساب الفرق بينهما والذي يسمى IQR وبعد ذلك تم تحديد الحد الادنى والحد الاعلى للقيم الطبيعية باستخدام المعادلة.

الحد الادنى = Q1 ناقص 1.5 ضرب IQR  
الحد الاعلى = Q3 زائد 1.5 ضرب IQR

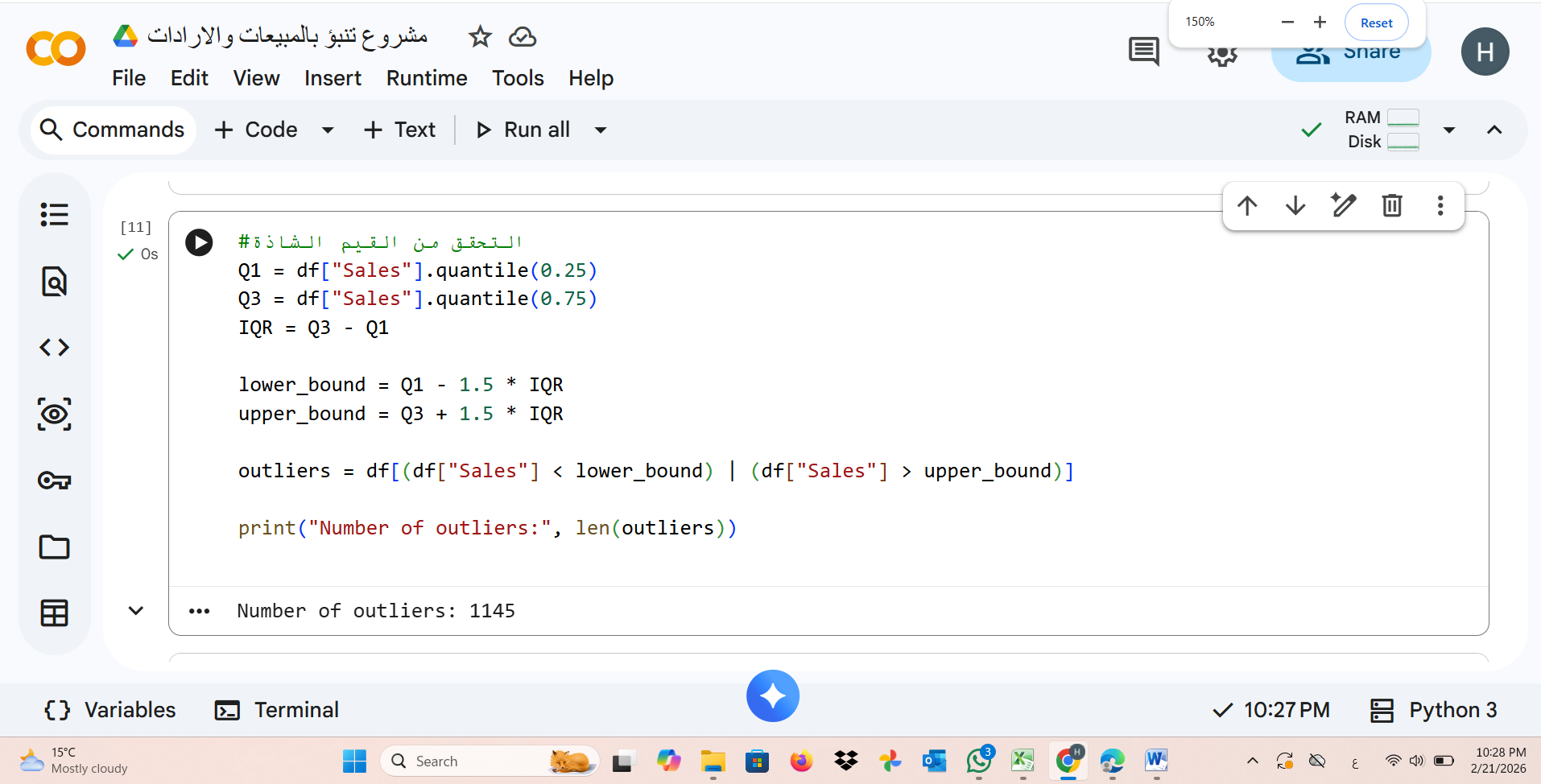
ثم تم استخراج جميع القيم التي تقع خارج هذا النطاق واعتبارها قيما شاذة

اظهرت النتائج ان عدد القيم الشاذة في عمود Sales يساوي 1145 قيمة وهو عدد ملحوظ مقارنة باجمالي عدد السجلات البالغ 9800 وهذا يؤكد ان توزيع المبيعات غير متجانس ويحتوي على عمليات بيع مرتفعة جدا مقارنة ببقية البيانات

وتعد عملية اكتشاف القيم الشاذة خطوة مهمة جدا في مشكلات الانحدار لان القيم المرتفعة جدا قد تؤثر بشكل كبير على بعض النماذج مثل الانحدار الخطي حيث قد تسحب خط الانحدار نحوها وتقلل من دقة التنبؤ في القيم المتوسطة والصغيرة

وجود عدد كبير نسبيا من القيم الشاذة يشير الى ضرورة اختيار نموذج قادر على التعامل مع التباين العالي مثل Random Forest او Gradient Boosting كما يمكن لاحقا اختبار اثر حذف هذه القيم او الاحتفاظ بها ومقارنة النتائج قبل وبعد المعالجة

**لقطة شاشة**



توضح لقطة الشاشة التالية تنفيذ كود حساب IQR واستخراج عدد القيم الشاذة داخل Google Colab حيث يظهر عدد القيم الشاذة في عمود Sales والذي بلغ 1145 قيمة مما يعكس طبيعة التوزيع غير المتوازن للمبيعات.

**رسم مخطط الصندوق لعمود Sales**

بعد حساب القيم الشاذة تم استخدام مخطط الصندوق Boxplot لتمثيل توزيع بيانات المبيعات بشكل مرئي وذلك باستخدام الامر

plt.figure(figsize=(14,5))  
df.boxplot()

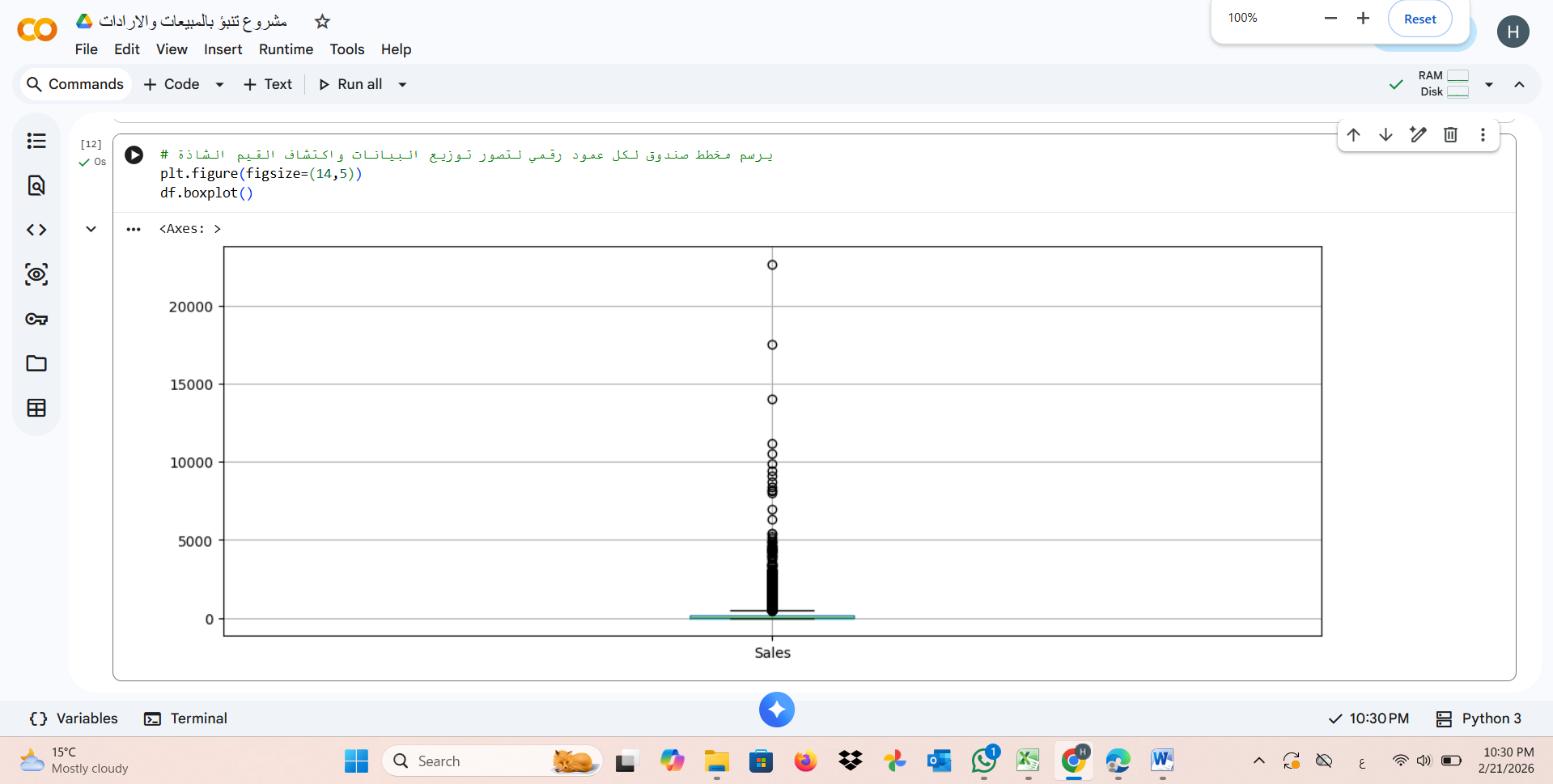
يساعد مخطط الصندوق على عرض الوسيط والربيعين الاول والثالث والمدى الربيعي بشكل واضح كما يظهر النقاط التي تقع خارج النطاق الطبيعي والتي تمثل القيم الشاذة.

من خلال الشكل يظهر ان معظم قيم المبيعات تتركز في نطاق منخفض نسبيا بينما توجد نقاط كثيرة مرتفعة جدا تمثل عمليات بيع ذات قيمة عالية وهذا يؤكد ما تم اكتشافه سابقا باستخدام طريقة IQR حيث ان عدد القيم الشاذة كان كبيرا نسبيا.

ويعد التمثيل البصري للبيانات خطوة مهمة جدا لانه يعزز الفهم التحليلي ويوضح توزيع القيم بشكل اسهل من الارقام المجردة كما يساعد في اتخاذ قرار بخصوص كيفية التعامل مع القيم الشاذة سواء بالاحتفاظ بها او معالجتها او استخدام نموذج يتحمل هذا النوع من التباين.

وجود عدد كبير من القيم المرتفعة يشير الى ان البيانات منحرفة نحو اليمين وهذا قد يؤثر على بعض نماذج الانحدار التقليدية لذلك من المهم اختيار نموذج مناسب لطبيعة البيانات.

**لقطة شاشة**



توضح لقطة الشاشة التالية رسم مخطط الصندوق لعمود Sales داخل Google Colab حيث تظهر القيم الشاذة بشكل واضح اعلى الصندوق مما يعكس التباين الكبير في بيانات المبيعات قبل الانتقال الى مرحلة تقسيم البيانات وبناء النموذج.

**معالجة القيم الشاذة باستخدام طريقة IQR**

بعد اكتشاف عدد كبير من القيم الشاذة في عمود Sales تم تنفيذ خطوة معالجة لهذه القيم باستخدام طريقة IQR ولكن بدلا من حذفها تم استخدام اسلوب يسمى Capping او Winsorizing

تم انشاء دالة باسم cap\_outliers تقوم بحساب الربع الاول Q1 والربع الثالث Q3 ثم حساب المدى الربيعي IQR وتحديد الحد الادنى والحد الاعلى للقيم الطبيعية بعد ذلك يتم استبدال اي قيمة اقل من الحد الادنى بالحد الادنى نفسه واي قيمة اعلى من الحد الاعلى بالحد الاعلى نفسه.

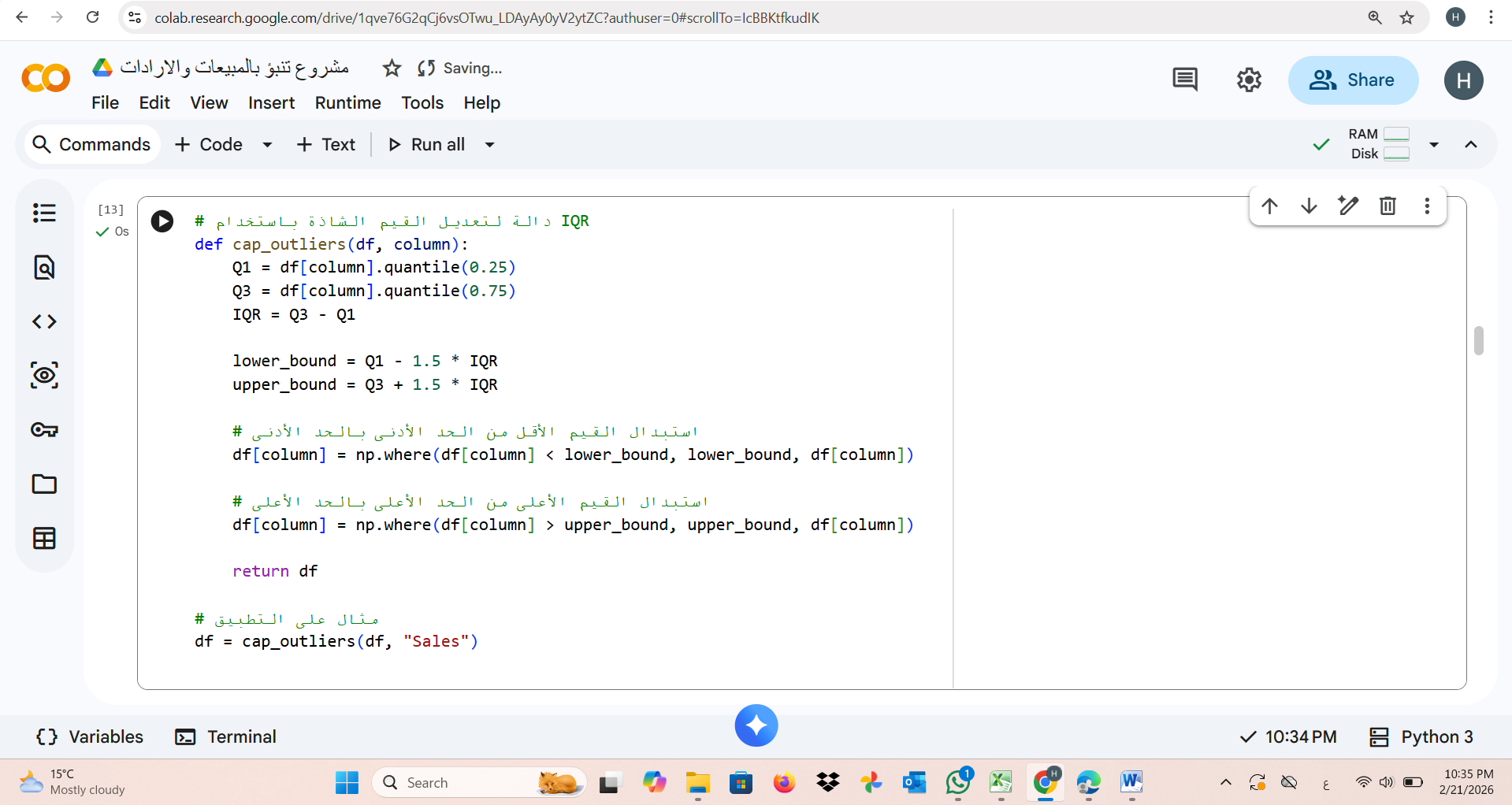
بهذه الطريقة يتم تقليل تأثير القيم المتطرفة دون حذف السجلات بالكامل مما يحافظ على حجم البيانات ويقلل من فقدان المعلومات المهمة.

بعد تعريف الدالة تم تطبيقها على عمود Sales لتعديل القيم الشاذة وجعل التوزيع اكثر استقرارا.

وتعتبر معالجة القيم الشاذة خطوة مهمة لتحسين اداء نماذج الانحدار خاصة النماذج الحساسة للقيم الكبيرة جدا مثل Linear Regression حيث قد تؤدي القيم المرتفعة جدا الى تشويه النموذج.

وتم اختيار أسلوب Capping بدلا من حذف القيم الشاذة بالكامل لأن حذف أكثر من ألف سجل كان سيؤدي إلى تقليل حجم البيانات بنسبة ملحوظة مما قد يؤثر سلبا على قدرة النموذج على التعلم. بالإضافة إلى ذلك فإن بعض القيم المرتفعة قد تمثل مبيعات حقيقية وليست أخطاء إدخال لذلك فإن تقليل تأثيرها كان خيارا أكثر توازنا من حذفها نهائيا.

**لقطة شاشة**



توضح لقطة الشاشة التالية تعريف دالة معالجة القيم الشاذة باستخدام IQR داخل Google Colab وتطبيقها على عمود Sales وذلك ضمن مرحلة تنظيف البيانات قبل الانتقال الى تقسيم البيانات وبناء نموذج التنبؤ بالمبيعات.

**رسم مخطط الصندوق بعد معالجة القيم الشاذة**

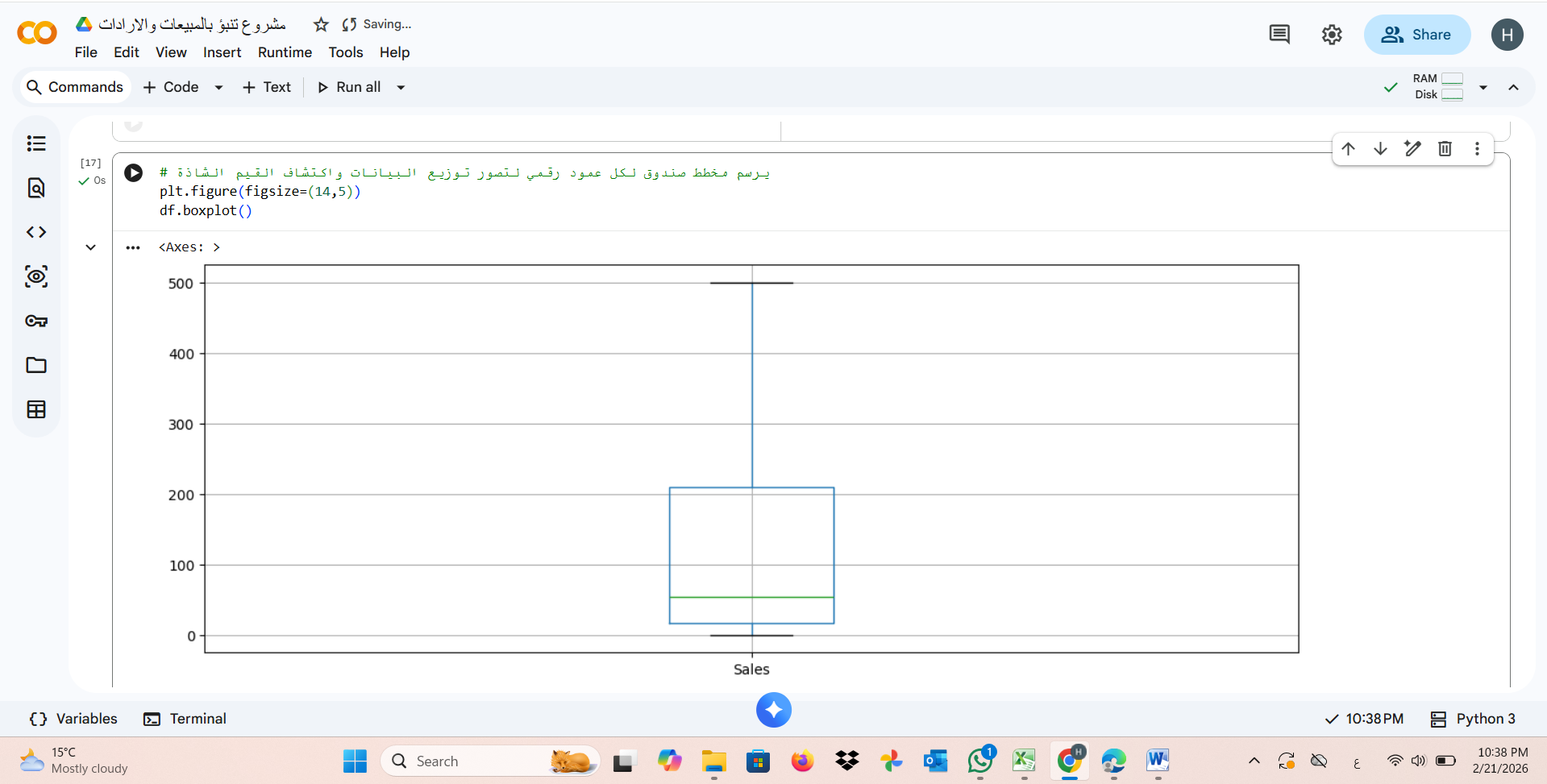
بعد تطبيق دالة معالجة القيم الشاذة باستخدام طريقة IQR تم اعادة رسم مخطط الصندوق لعمود Sales وذلك لمقارنة التوزيع قبل وبعد المعالجة والتأكد من فعالية الخطوة.

يظهر من الشكل الجديد ان القيم المتطرفة العالية التي كانت تتجاوز عشرات الالاف لم تعد موجودة بنفس الشكل السابق كما اصبح التوزيع اكثر توازنا واستقرارا واصبحت حدود البيانات ضمن نطاق منطقي مقارنة بالقيم الاصلية.

اصبح الحد الاعلى اقل بكثير من السابق مما يدل على نجاح عملية Capping في تقليل تأثير القيم الشاذة دون حذف السجلات بالكامل وهذا يساعد في جعل البيانات اكثر ملاءمة لتدريب نموذج الانحدار.

تحسين توزيع البيانات بهذه الطريقة يقلل من التباين العالي ويزيد من استقرار النموذج خاصة في النماذج الحساسة للقيم الكبيرة كما يدعم تحقيق نتائج اكثر دقة عند تقييم النموذج لاحقا.

لقطة شاشة



توضح لقطة الشاشة التالية مخطط الصندوق لعمود Sales بعد معالجة القيم الشاذة حيث يظهر التوزيع الجديد للبيانات بشكل اكثر انتظاما واستقرارا مما يؤكد نجاح خطوة تنظيف البيانات قبل الانتقال الى مرحلة تقسيم البيانات وبناء نموذج التنبؤ بالمبيعات.

**تحويل البيانات التاريخية واستخراج الخصائص الزمنية**

يعد المتغير الزمني من اهم العوامل المؤثرة في تحليل المبيعات لان سلوك الشراء لا يكون ثابتا عبر الزمن بل يتغير حسب المواسم والفترات الزمنية المختلفة لذلك كان من الضروري معالجة عمود Order Date بطريقة تسمح للنموذج باستخلاص هذه الانماط الزمنية بشكل فعال.

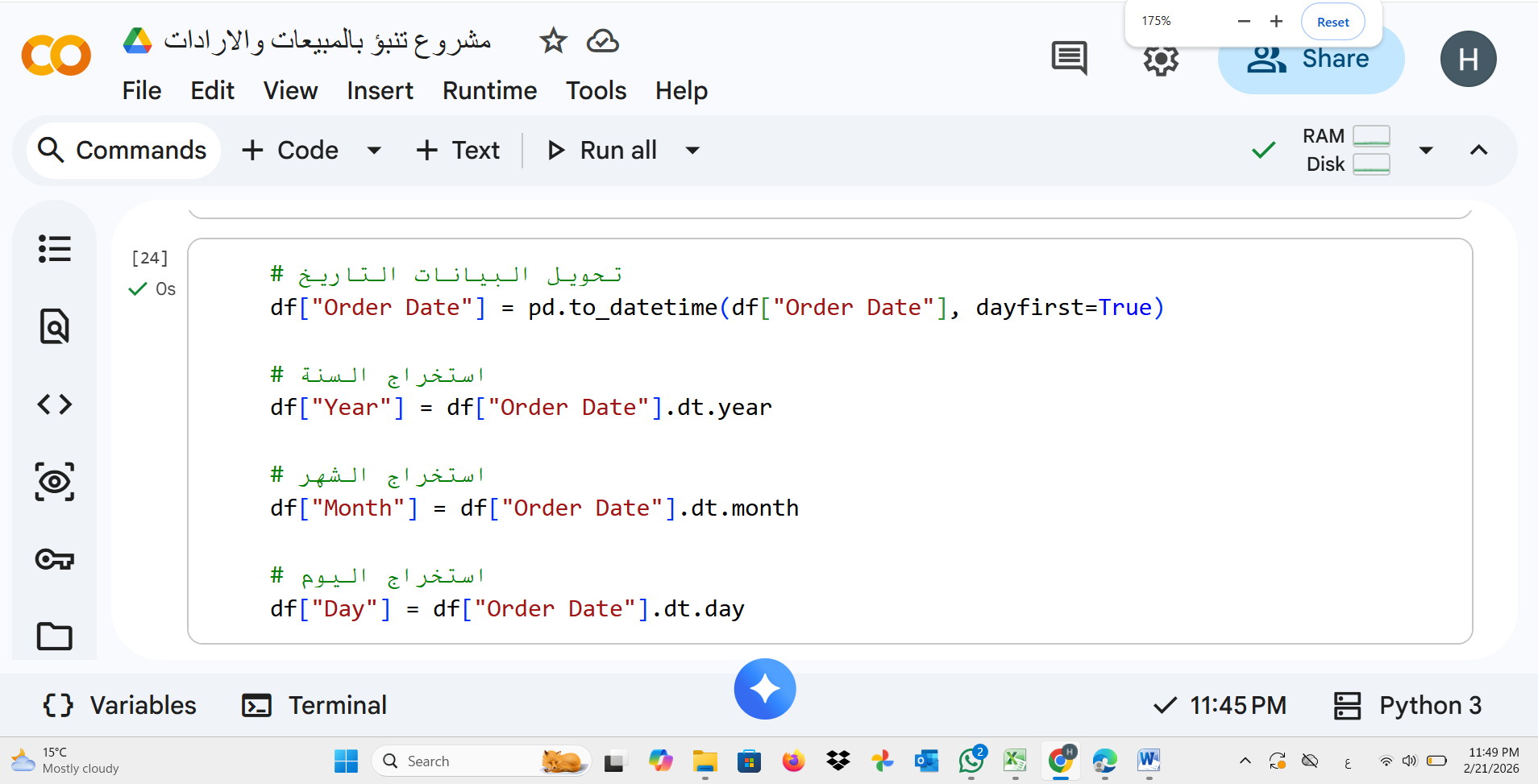
في البداية تم تحويل عمود Order Date الى صيغة تاريخ باستخدام الدالة pd.to\_datetime وذلك لضمان تفسير القيم بشكل صحيح كتواريخ حقيقية وليس كسلاسل نصية وقد تم استخدام الخيار dayfirst=True لتجنب اي خطأ في تفسير تنسيق اليوم والشهر.

بعد تحويل العمود الى نوع بيانات زمني تم استخراج مكوناته الاساسية وهي السنة والشهر واليوم وانشاء اعمدة رقمية جديدة تمثل هذه القيم ويعد هذا الاجراء مهما من الناحية الرياضية لان نماذج الانحدار لا تستطيع التعامل مباشرة مع القيم الزمنية المركبة بل تحتاج الى خصائص رقمية مستقلة يمكن استخدامها في حساب العلاقات الخطية او غير الخطية.

يسمح استخراج السنة بتحليل التغير العام في المبيعات عبر الفترات المختلفة كما يتيح استخراج الشهر اكتشاف الانماط الموسمية مثل زيادة الطلب في اشهر معينة اما استخراج اليوم فيمكن ان يكشف عن وجود سلوك دوري داخل الشهر نفسه

من الناحية الاحصائية يؤدي هذا التحويل الى زيادة عدد الخصائص التفسيرية في النموذج مما يساعد على تحسين قدرة النموذج على تفسير التباين في قيمة المبيعات ويقلل من فقدان المعلومات الكامنة داخل المتغير الزمني الاصلي وبالتالي تساهم هذه الخطوة في تعزيز دقة التنبؤ ودعم استقرار النموذج اثناء التدريب.

توضح لقطة الشاشة التالية تنفيذ عملية تحويل عمود التاريخ واستخراج السنة والشهر واليوم داخل بيئة Google Colab كجزء من مرحلة هندسة الخصائص قبل تدريب النموذج.



**استكمال هندسة الخصائص وحذف العمود الاصلي**

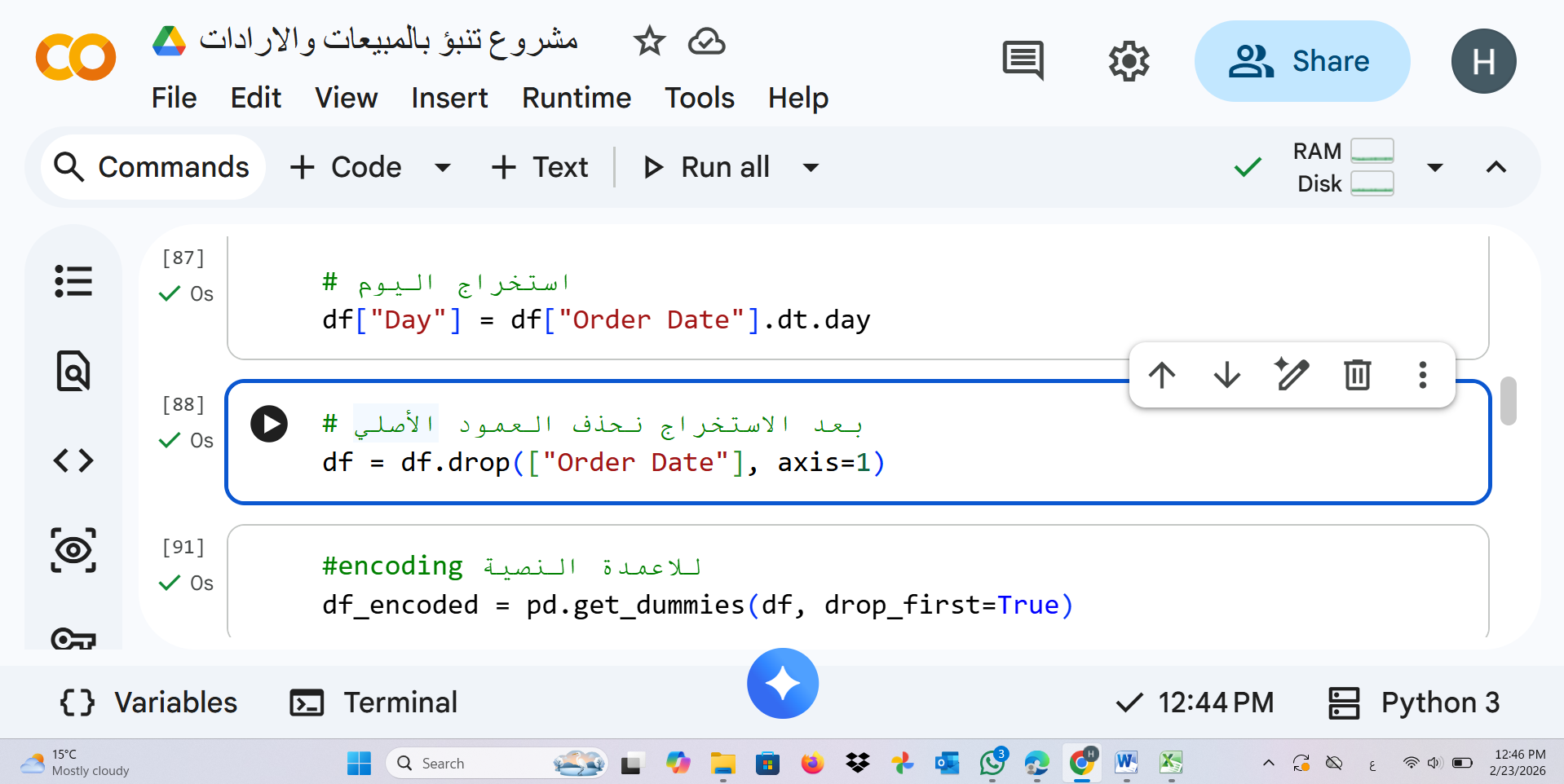
بعد استخراج المكونات الزمنية من عمود Order Date وهي السنة والشهر واليوم تم الانتقال الى خطوة مهمة وهي حذف العمود الاصلي Order Date من مجموعة البيانات.

يعود السبب في ذلك الى ان النموذج لم يعد بحاجة الى التاريخ الكامل بعد تحويله الى خصائص رقمية مستقلة حيث ان الاحتفاظ بالعمود الاصلي قد يسبب تكرار في المعلومات او ادخال بيانات غير قابلة للمعالجة داخل نموذج الانحدار.

و يعتبر حذف العمود الاصلي خطوة تنظيمية تهدف الى تقليل التعقيد داخل مجموعة البيانات والاعتماد فقط على المتغيرات العددية القابلة للاستخدام في العمليات الحسابية كما يساعد ذلك في تجنب اي تشويش قد يحدث نتيجة وجود متغير زمني غير مهيأ داخل النموذج.

بهذا تصبح البيانات جاهزة من حيث البنية العددية للانتقال الى مرحلة فصل المتغيرات المستقلة عن المتغير التابع ثم تقسيمها الى مجموعات تدريب وتحقق واختبار تمهيدا لعملية التدريب.

توضح لقطة الشاشة التالية استخراج اليوم من عمود التاريخ ثم حذف العمود Order Date بعد الانتهاء من تحويله وذلك ضمن مرحلة تجهيز البيانات قبل فصل المتغيرات وبدء تدريب النموذج.



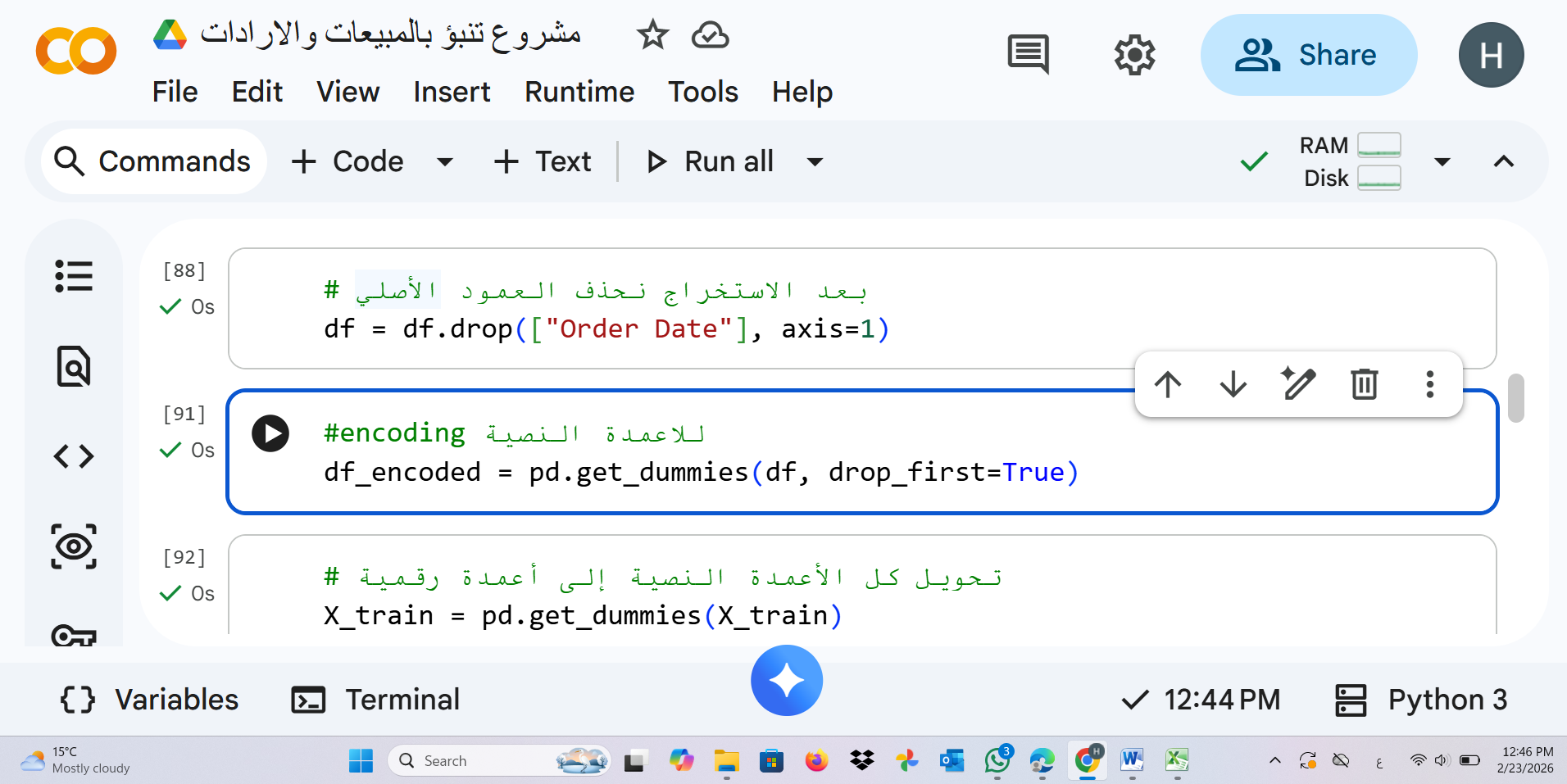
**تحويل الاعمدة النصية الى متغيرات رقمية**

تم الانتقال الى خطوة مهمة وهي تحويل الاعمدة النصية الى متغيرات رقمية باستخدام دالة get\_dummies.

تعد هذه الخطوة ضرورية لان نماذج الانحدار لا تستطيع التعامل مع القيم النصية بشكل مباشر مثل اسم الفئة او المنطقة لذلك يتم تحويل كل فئة نصية الى اعمدة رقمية ثنائية تحتوي على القيمة 1 في حال وجود الفئة و 0 في حال عدم وجودها.

تم استخدام الخيار drop\_first=True لتجنب مشكلة التعدد الخطي multicollinearity التي قد تحدث عند تمثيل جميع الفئات داخل النموذج حيث يؤدي حذف اول فئة الى منع وجود علاقة خطية كاملة بين الاعمدة الجديدة مما يساعد في تحسين استقرار معاملات الانحدار.

توضح لقطة الشاشة التالية استخدام دالة get\_dummies لتحويل الاعمدة النصية الى متغيرات رقمية مع استخدام drop\_first=True ثم تطبيق التحويل على بيانات التدريب والاختبار قبل بدء عملية التدريب.



**توحيد الاعمدة بين بيانات التدريب والاختبار والتحقق من الابعاد**

بعد تحويل الاعمدة النصية الى متغيرات رقمية باستخدام get\_dummies تم تطبيق العملية بشكل منفصل على بيانات التدريب وبيانات الاختبار وذلك لضمان ان تكون جميع القيم نصية قد تحولت الى اعمدة رقمية.

ولكن قد تظهر مشكلة مهمة في هذه المرحلة وهي اختلاف عدد الاعمدة بين X\_train و X\_test وذلك في حال وجود فئة معينة ظهرت في بيانات التدريب ولم تظهر في بيانات الاختبار او العكس وهذا قد يؤدي الى خطأ اثناء عملية التنبؤ لان النموذج يتوقع نفس عدد الخصائص التي تدرب عليها.

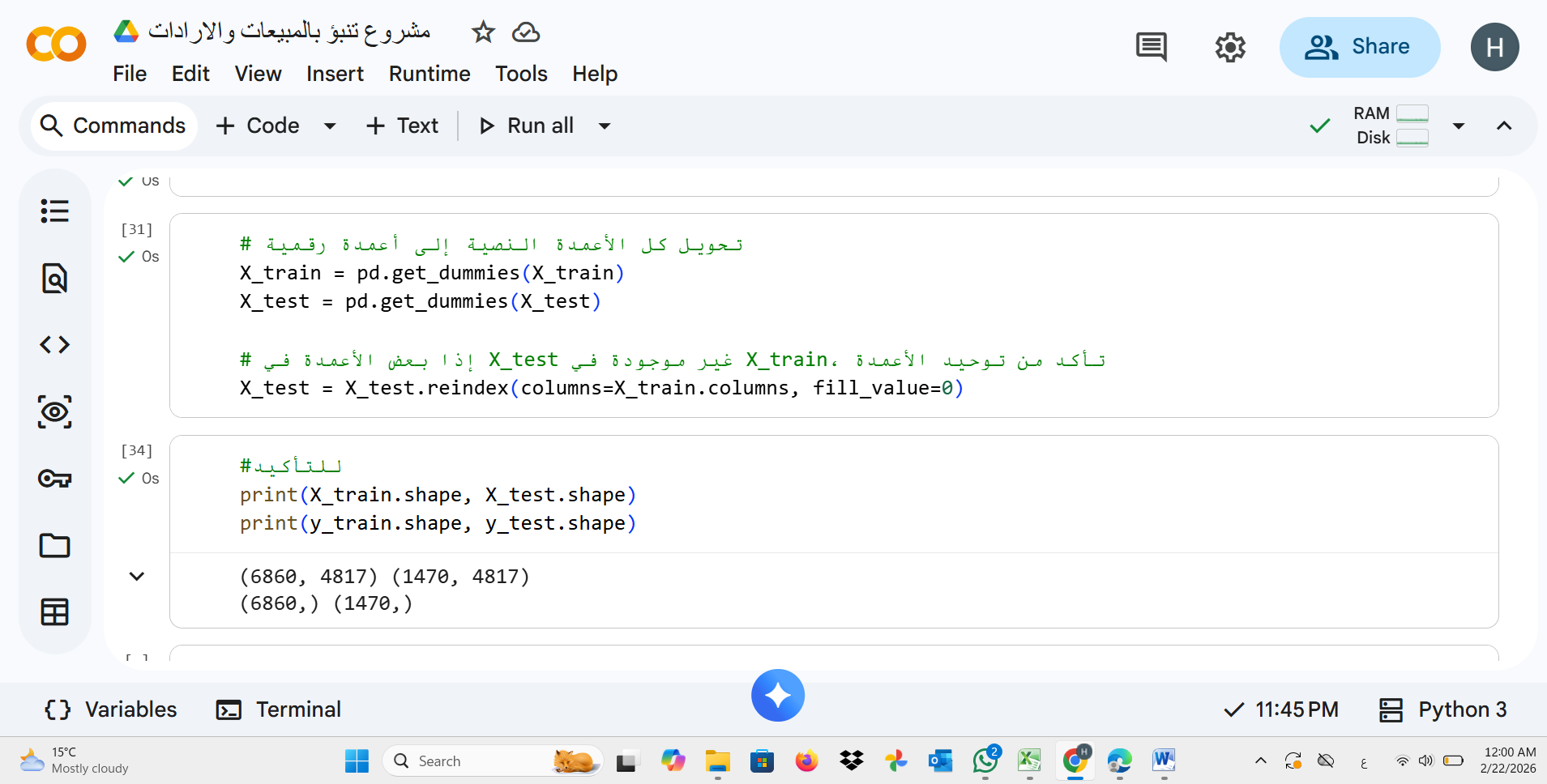
لذلك تم استخدام دالة reindex لمطابقة اعمدة X\_test مع اعمدة X\_train بحيث يتم ترتيب الاعمدة بنفس الترتيب وملء اي عمود غير موجود بالقيمة صفر وهذا يضمن توحيد البنية العددية بين المجموعتين ويمنع حدوث اي خطأ اثناء التدريب او التقييم.

بعد ذلك تم استخدام الامر print للتحقق من ابعاد البيانات حيث اظهرت النتائج ان بيانات التدريب تحتوي على 6860 سجل و 4817 خاصية بينما تحتوي بيانات الاختبار على 1470 سجل بنفس عدد الخصائص وهذا يدل على نجاح عملية التوحيد.

كما تم التحقق من ابعاد y\_train و y\_test للتأكد من تطابق عدد السجلات بين المتغيرات المستقلة والمتغير التابع مما يضمن جاهزية البيانات لبدء تدريب النموذج بشكل صحيح.

تعد هذه الخطوة مهمة جدا لضمان استقرار النموذج ومنع الاخطاء التقنية كما تعكس فهما دقيقا لكيفية تجهيز البيانات قبل مرحلة التدريب.

توضح لقطة الشاشة التالية تحويل الاعمدة النصية في بيانات التدريب والاختبار ثم توحيد الاعمدة باستخدام reindex واخيرا التحقق من ابعاد X\_train و X\_test و y\_train و y\_test قبل بدء عملية التدريب.



**تحليل احصائي لقيم المبيعات قبل التحويل**

قبل تنفيذ اي تحويل رياضي على عمود Sales تم اجراء تحليل احصائي شامل لفهم طبيعة توزيع المبيعات ومعرفة ما اذا كانت البيانات تحتوي على انحراف او تشتت كبير قد يؤثر على اداء النموذج.

تم حساب اعلى قيمة وادنى قيمة والمتوسط والوسيط والانحراف المعياري وذلك بهدف تقييم مدى تباين البيانات كما تم حساب معامل الانحراف Skewness لقياس درجة ميل التوزيع نحو اليمين او اليسار.

**شرح الكود**

في البداية تم استخدام الدوال الاحصائية التالية

max لحساب اعلى قيمة مبيعات  
min لحساب ادنى قيمة  
mean لحساب المتوسط الحسابي  
median لحساب الوسيط  
std لحساب الانحراف المعياري

بعد ذلك تم استخدام الدالة skew من مكتبة scipy لحساب معامل الانحراف والذي يقيس مدى تماثل توزيع البيانات

ثم تم طباعة النتائج لفهم خصائص التوزيع قبل الانتقال الى مرحلة المعالجة

**تفسير النتائج**

اظهرت النتائج ان اعلى قيمة للمبيعات بلغت حوالي 500 بينما ادنى قيمة كانت حوالي 0.44 كما بلغ المتوسط حوالي 140 بينما كان الوسيط حوالي 54

وجود فرق كبير بين المتوسط والوسيط يشير الى وجود قيم مرتفعة جدا اثرت على المتوسط وهذا يعني ان توزيع البيانات منحرف نحو اليمين

كما ان الانحراف المعياري بلغ حوالي 169 وهو رقم مرتفع نسبيا مقارنة بالمتوسط مما يدل على وجود تباين كبير في البيانات

اما قيمة معامل الانحراف فبلغت حوالي 1.21 وهي قيمة موجبة وتشير الى ان البيانات تحتوي على انحراف موجب واضح اي ان هناك عدد من القيم الكبيرة التي تسحب التوزيع نحو اليمين

**سبب تنفيذ التحويل اللوغاريتمي**

بناء على نتائج التحليل الاحصائي تبين ان توزيع المبيعات غير متماثل ويحتوي على انحراف واضح لذلك تم تنفيذ تحويل لوغاريتمي باستخدام np.log1p بهدف تقليل الانحراف وجعل التوزيع اكثر تقاربا من الشكل الطبيعي

التحويل اللوغاريتمي يساعد على تقليل تأثير القيم الكبيرة جدا ويجعل الفروقات بين القيم اكثر توازنا مما يسهم في تحسين استقرار النموذج خاصة في نماذج الانحدار

**شرح الكود الخاص بالتحويل**

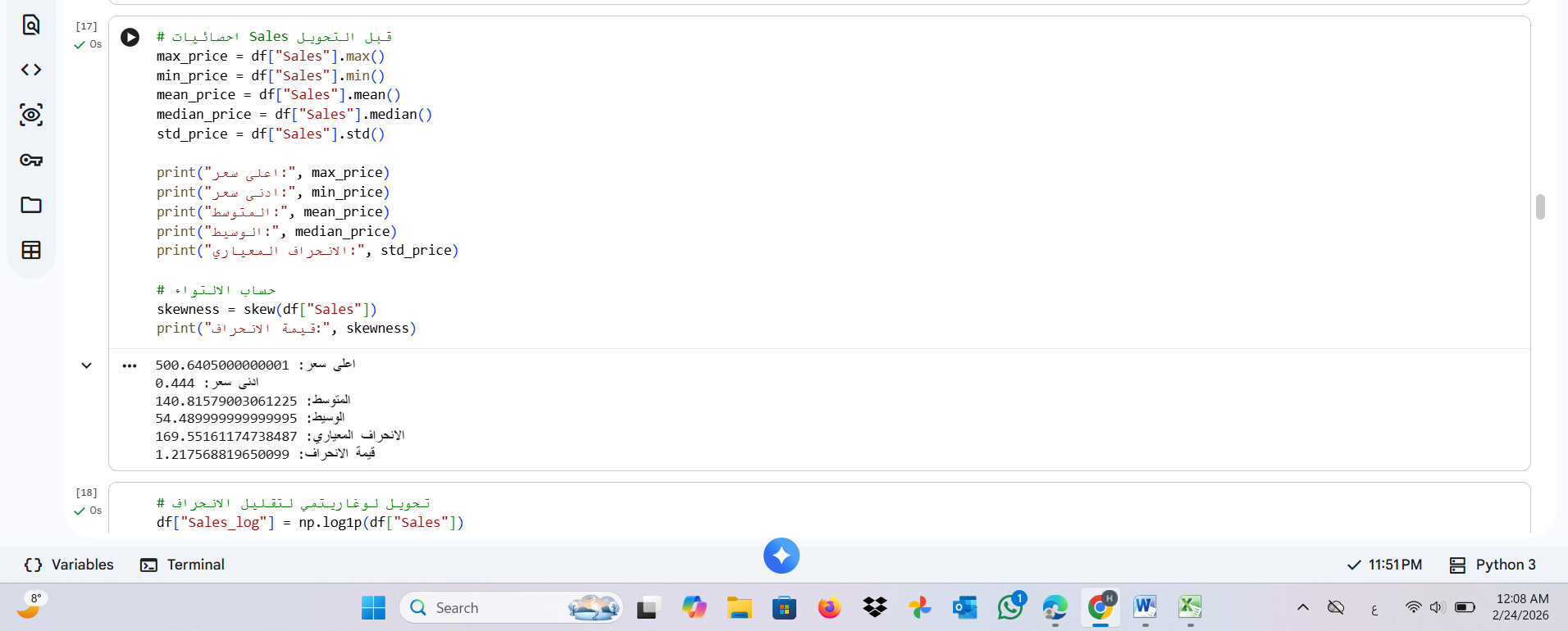
تم انشاء عمود جديد باسم Sales\_log باستخدام المعادلة

np.log1p(df["Sales"])

تم استخدام log1p بدلا من log العادي لتجنب اي مشكلة رياضية في حال وجود قيم صفرية حيث ان log1p يحسب لوغاريتم القيمة بعد اضافة واحد

هذا التحويل لا يغير القيم الاصلية بل ينشئ نسخة معدلة يمكن استخدامها في التحليل او التدريب

**شرح السكرين**



توضح لقطة الشاشة تنفيذ التحليل الاحصائي لعمود Sales قبل التحويل حيث تظهر القيم المحسوبة مثل اعلى سعر وادنى سعر والمتوسط والوسيط والانحراف المعياري ومعامل الانحراف كما توضح بداية تنفيذ التحويل اللوغاريتمي لانشاء عمود Sales\_log بهدف تقليل الانحراف وتحسين توزيع البيانات قبل استخدامه في بناء نموذج الذكاء الاصطناعي.

**تحليل البيانات بعد تنفيذ التحويل اللوغاريتمي**

بعد ملاحظة وجود انحراف موجب واضح في توزيع المبيعات تم تنفيذ تحويل لوغاريتمي على عمود Sales بهدف تقليل الانحراف وجعل التوزيع اكثر تماثلا واستقرارا

تم انشاء عمود جديد باسم Sales\_log باستخدام الدالة np.log1p حيث تقوم هذه الدالة بحساب اللوغاريتم الطبيعي للقيمة بعد اضافة واحد وذلك لتجنب اي مشاكل رياضية في حال وجود قيم صفرية

**شرح الكود**

df["Sales\_log"] = np.log1p(df["Sales"])  
يقوم هذا السطر بانشاء نسخة محولة لوغاريتميا من المبيعات دون حذف القيم الاصلية

بعد ذلك تم حساب معامل الانحراف الجديد باستخدام  
skew(df["Sales\_log"])

ثم تم حساب المتوسط والوسيط والانحراف المعياري للقيم بعد التحويل من اجل مقارنة التغيرات التي حدثت في التوزيع

**تفسير النتائج بعد التحويل**

اظهرت النتائج ان قيمة معامل الانحراف بعد التحويل اصبحت حوالي 0.047 وهي قيمة قريبة جدا من الصفر وهذا يعني ان التوزيع اصبح شبه متماثل مقارنة بالقيمة السابقة التي كانت حوالي 1.21

هذا يدل على ان التحويل اللوغاريتمي نجح في تقليل الانحراف الموجب وجعل توزيع البيانات اكثر توازنا

كما نلاحظ ان المتوسط بعد التحويل اصبح حوالي 4.07 والوسيط حوالي 4.01 والفارق بينهما اصبح صغيرا جدا مقارنة بالحالة السابقة وهذا مؤشر اخر على ان التوزيع اصبح اقرب الى التماثل

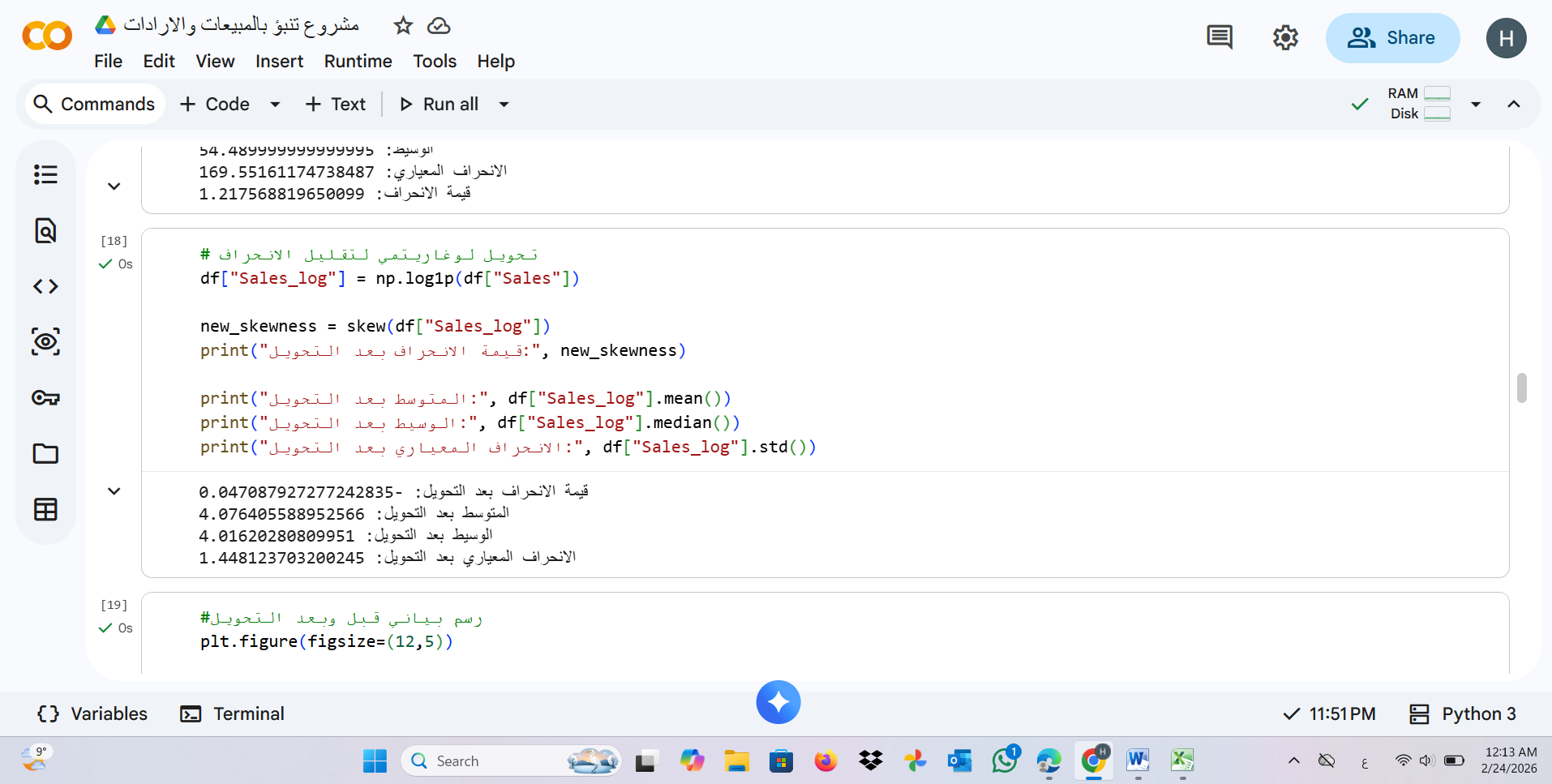
اما الانحراف المعياري فقد انخفض بشكل كبير من حوالي 169 قبل التحويل الى حوالي 1.44 بعد التحويل مما يدل على تقليل التشتت بين القيم

**مقارنة قبل وبعد التحويل**

قبل التحويل كان التوزيع منحرفا بشكل واضح نحو اليمين وكانت القيم الكبيرة تؤثر بقوة على المتوسط والانحراف المعياري

بعد التحويل اصبحت البيانات اكثر انتظاما واقرب الى التوزيع الطبيعي مما يجعلها اكثر ملاءمة لبعض نماذج الانحدار التي تستفيد من توزيع مستقر

**شرح لقطة الشاشة**



توضح لقطة الشاشة تنفيذ التحويل اللوغاريتمي على عمود Sales ثم حساب معامل الانحراف الجديد والمتوسط والوسيط والانحراف المعياري بعد التحويل كما تظهر النتائج الرقمية التي تؤكد انخفاض الانحراف بشكل كبير وتحسن خصائص التوزيع.

**الرسم البياني قبل وبعد التحويل اللوغاريتمي**

بعد اجراء التحليل الاحصائي تم رسم مخططين بيانيين لعرض توزيع المبيعات قبل وبعد التحويل اللوغاريتمي بهدف المقارنة البصرية بين التوزيعين وفهم تأثير التحويل بشكل اوضح

**شرح الكود**

تم استخدام plt.figure لتحديد حجم الشكل البياني

ثم تم تقسيم الشكل الى جزئين باستخدام plt.subplot حيث يمثل الرسم الاول توزيع Sales قبل التحويل بينما يمثل الرسم الثاني توزيع Sales\_log بعد التحويل

تم استخدام plt.hist لرسم المدرج التكراري بعدد 50 فئة لعرض تكرار القيم

ثم تم اضافة عنوان لكل رسم باستخدام plt.title واخيرا تم عرض الشكل باستخدام plt.show

**تفسير الرسم قبل التحويل**

يظهر الرسم الاول ان توزيع المبيعات منحرف بشكل واضح نحو اليمين حيث تتركز معظم القيم في الجزء المنخفض من المحور بينما توجد ذيل طويل يمتد نحو القيم الكبيرة

هذا يؤكد ان البيانات تحتوي على عدد قليل من العمليات ذات القيم المرتفعة التي تؤثر على شكل التوزيع وتسحب المتوسط الى الاعلى

**تفسير الرسم بعد التحويل**

بعد تطبيق التحويل اللوغاريتمي يظهر الرسم الثاني توزيعا اكثر توازنا حيث اصبحت القيم موزعة بشكل اقرب الى الشكل الطبيعي ولم يعد هناك ذيل طويل حاد كما في الرسم السابق

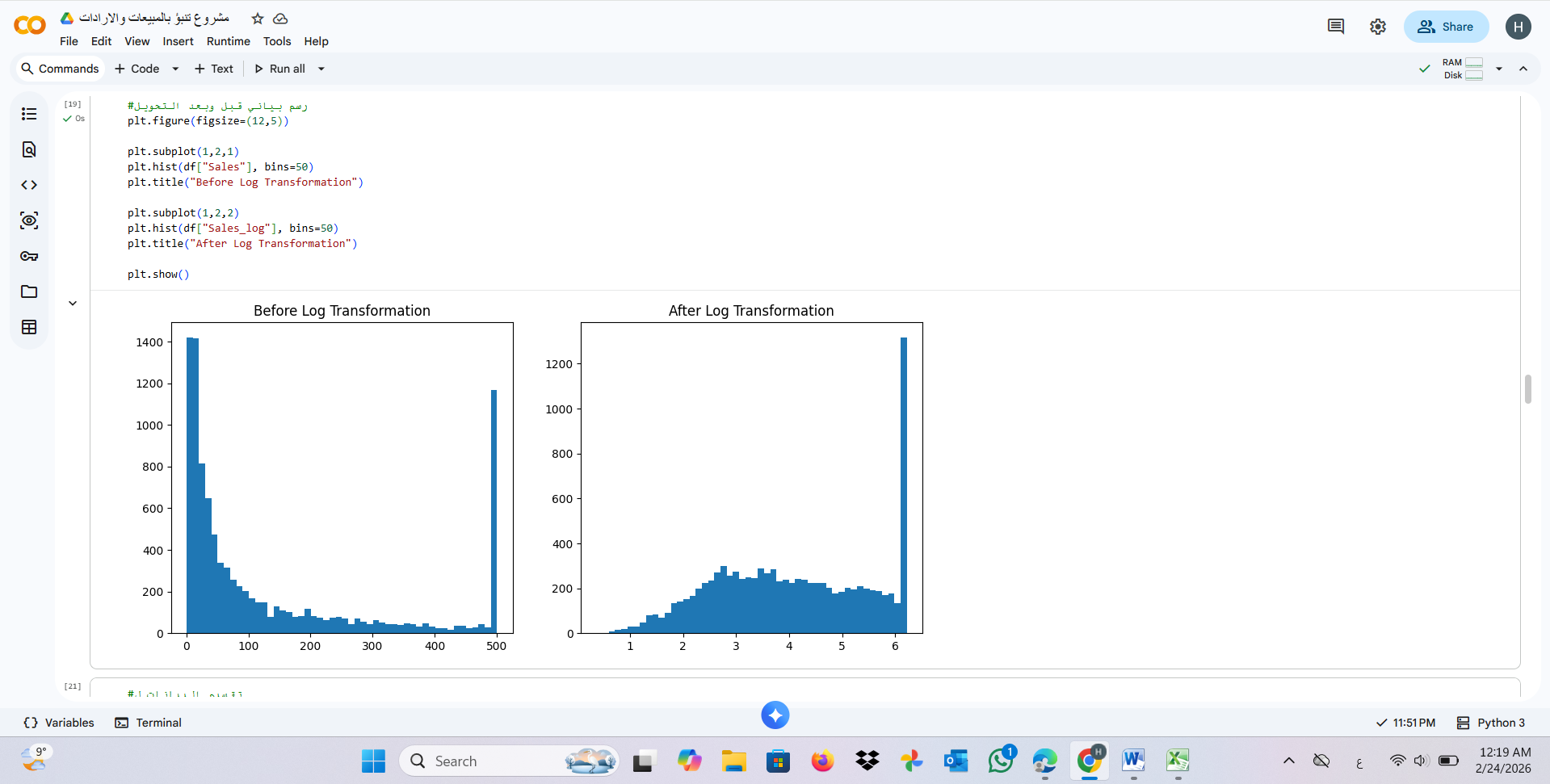
اصبحت الفروقات بين القيم اكثر انتظاما مما يعكس نجاح التحويل في تقليل تأثير القيم الكبيرة وتقليل الانحراف

**المقارنة بين التوزيعين**

قبل التحويل كان التوزيع غير متماثل ويحتوي على انحراف موجب واضح مما قد يؤثر على بعض نماذج الانحدار

بعد التحويل اصبح التوزيع اكثر استقرارا واقرب الى التماثل مما يجعل البيانات اكثر ملاءمة للتحليل الاحصائي وللنماذج التي تستفيد من توزيع متوازن.

**شرح لقطة الشاشة**

****

توضح لقطة الشاشة تنفيذ الكود الخاص برسم التوزيع قبل وبعد التحويل كما تعرض المدرجين التكراريين حيث يظهر الفرق الواضح بين شكل التوزيع قبل تطبيق اللوغاريتم وبعده مما يدعم نتائج التحليل الاحصائي السابق.

**تقسيم البيانات الى تدريب وتحقق واختبار**

خطوة تقسيم البيانات باستخدام دالة  
train\_test\_split من مكتبة sklearn وذلك لضمان تقييم النموذج بطريقة عادلة ومنع حدوث overfitting.

في البداية تم تقسيم البيانات الى مجموعتين رئيسيتين حيث تم تخصيص 70 بالمئة من البيانات للتدريب و30 بالمئة كمجموعة مؤقتة باستخدام المتغير test\_size=0.30..

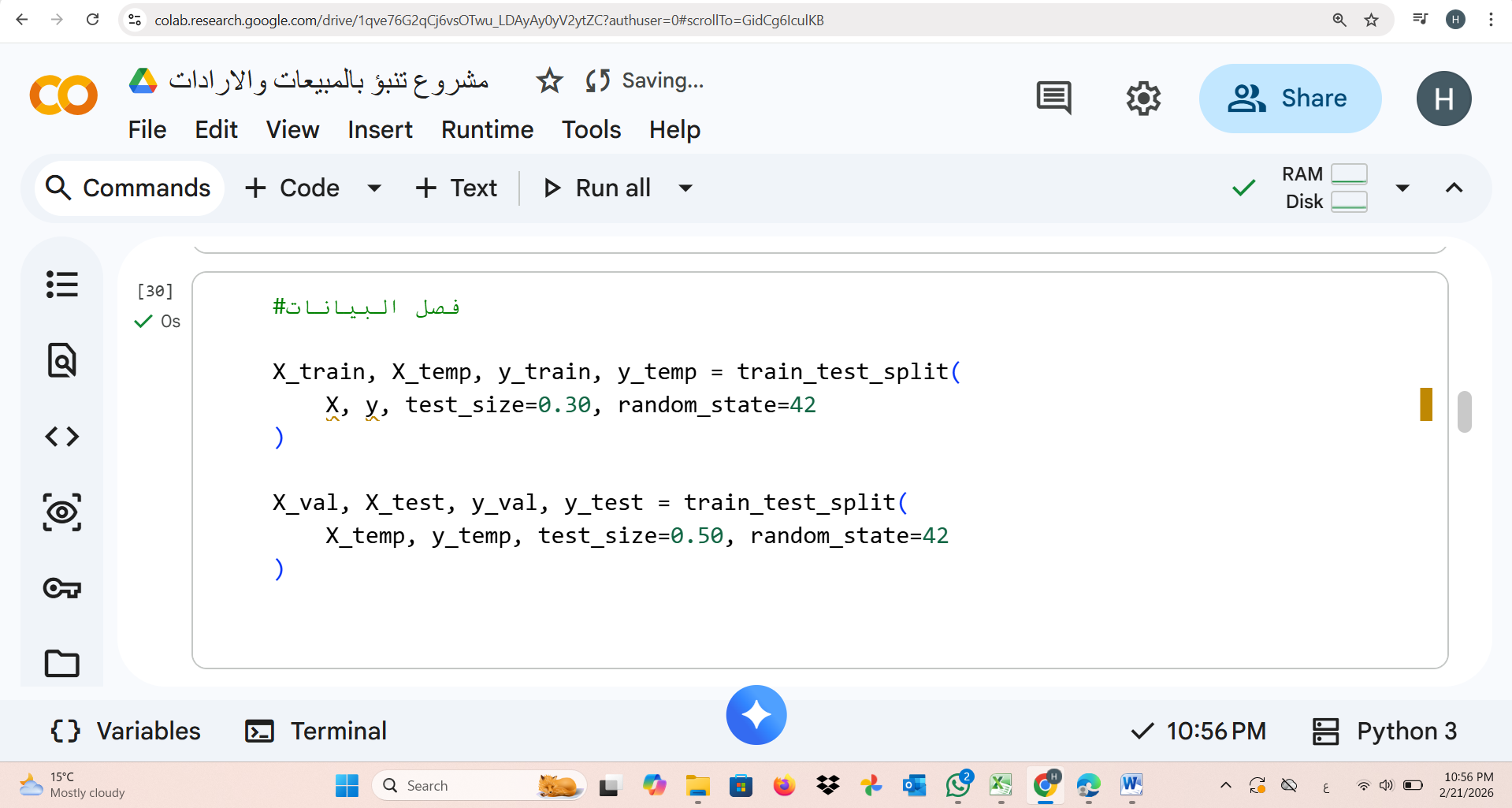
بعد ذلك تم تقسيم المجموعة المؤقتة مرة اخرى الى مجموعتين متساويتين بنسبة 50 بالمئة لكل منهما وذلك للحصول على مجموعة تحقق Validation ومجموعة اختبار Test.

وبذلك اصبح التقسيم النهائي للبيانات كما يلي:

70 بالمئة بيانات تدريب  
15 بالمئة بيانات تحقق  
15 بالمئة بيانات اختبار

تم استخدام random\_state=42 لضمان ثبات النتائج وامكانية اعادة تنفيذ التقسيم بنفس الطريقة في حال اعادة تشغيل الكود.

**لقطة شاشة**



توضح لقطة الشاشة تنفيذ عملية تقسيم البيانات الى مجموعات تدريب وتحقق واختبار باستخدام دالة train\_test\_split داخل بيئة Google Colab. تم تخصيص 70 بالمئة من البيانات للتدريب و15 بالمئة للتحقق و15 بالمئة للاختبار مع استخدام random\_state=42 لضمان ثبات النتائج.

**سبب استخدام هذا التقسيم**

تم اختيار تقسيم البيانات بنسبة 70 بالمئة للتدريب و15 بالمئة للتحقق و15 بالمئة للاختبار لضمان تحقيق توازن بين تدريب النموذج وتقييمه بشكل عادل.

تخصيص 70 بالمئة من البيانات للتدريب يسمح للنموذج بالتعلم من كمية كافية من السجلات واكتشاف الانماط والعلاقات بين المتغيرات المختلفة وقيمة المبيعات.

اما تخصيص 15 بالمئة لمجموعة التحقق فيساعد على تقييم اداء النموذج اثناء عملية التطوير وضبط المعاملات دون التأثير على مجموعة الاختبار النهائية.

في حين تستخدم مجموعة الاختبار التي تمثل 15 بالمئة من البيانات لقياس الاداء الحقيقي للنموذج على بيانات جديدة لم يسبق له رؤيتها مما يعطي مؤشرا واقعيا على مدى قدرته على التعميم.

يساعد هذا النوع من التقسيم على تقليل خطر حدوث overfitting ويزيد من موثوقية النتائج النهائية للنموذج.

**بعد الانتهاء من مرحلة تخطيط الحل وفحص جودة البيانات وتنظيفها ومعالجتها** أصبحت مجموعة البيانات جاهزة للاستخدام في بناء نموذج الذكاء الاصطناعي. وقد ساهمت خطوات المعالجة السابقة مثل حذف الأعمدة غير الضرورية ومعالجة القيم الشاذة وتقسيم البيانات إلى مجموعات تدريب وتحقق واختبار في ضمان أن تكون البيانات مستقرة ومناسبة لعملية التعلم.

في هذه المرحلة سيتم الانتقال من التخطيط والإعداد إلى مرحلة تطوير النموذج الفعلي حيث سيتم اختيار خوارزمية مناسبة من خوارزميات التعلم الموجه من نوع الانحدار ثم تدريب النموذج على بيانات التدريب وتقييم أدائه باستخدام مقاييس إحصائية مثل R2 و MAE و RMSEكما سيتم تحليل النتائج ومقارنتها بهدف تحسين النموذج والوصول إلى أفضل أداء ممكن يدعم أهداف المشروع.

تمثل هذه المرحلة الجزء العملي من الحل حيث يتم تحويل البيانات المعدة مسبقا إلى نظام قادر على إنتاج تنبؤات فعلية يمكن الاعتماد عليها في دعم عملية اتخاذ القرار.

**اختيار الخوارزمية المناسبة وسبب اختيارها**

تم اختيار خوارزمية الانحدار الخطي كنموذج اولي لبناء نظام التنبؤ بالمبيعات وذلك بعد تحليل طبيعة المشكلة واهداف المشروع. بما ان الهدف هو التنبؤ بقيمة عددية مستمرة تمثل حجم المبيعات فان المشكلة تصنف ضمن مسائل الانحدار Regression وليست تصنيفا Classification مما يجعل استخدام خوارزمية مخصصة للتعامل مع القيم العددية امرا منطقيا ومناسبا.

يعتمد نموذج الانحدار الخطي على ايجاد علاقة رياضية بين المتغيرات المستقلة مثل الفئة والمنطقة وتاريخ الطلب وبين المتغير التابع وهو قيمة المبيعات. من خلال هذا النموذج يمكن قياس تأثير كل متغير على المبيعات وتحديد ما اذا كان التأثير ايجابيا او سلبيا مما يوفر فهما تحليليا للبيانات وليس مجرد تنبؤ عددي.

يعد الانحدار الخطي نموذجا بسيطا وسهل التفسير مقارنة بالنماذج المعتمدة على الاشجار او النماذج التجميعية. في المراحل الاولى من اي مشروع تعلم الي يفضل البدء بنموذج اساسي واضح ليكون مرجعا تتم مقارنة النماذج الاخرى به لاحقا. لذلك تم استخدامه كنقطة انطلاق لتقييم مدى ملاءمة البيانات للنماذج الخطية قبل الانتقال الى نماذج اكثر تعقيدا مثل Random Forest.

من الناحية العملية يتميز النموذج بسرعة التدريب وقلة استهلاك الموارد الحاسوبية كما يمكن تقييم ادائه باستخدام مقاييس معيارية مثل R2 و RMSE مما يسمح باجراء مقارنة كمية دقيقة مع النماذج البديلة.

رغم بساطة النموذج الا انه يفترض وجود علاقة خطية بين المتغيرات وقيمة المبيعات وهو افتراض قد لا يكون دقيقا تماما في البيانات الواقعية. لذلك تم استخدامه كبداية منهجية لاختبار مدى تحقق هذا الافتراض وتحديد الحاجة الى نماذج قادرة على تمثيل العلاقات غير الخطية في حال كانت النتائج غير كافية.

**انشاء Pipeline وتجهيز النموذج للتدريب**

بعد الانتهاء من تنظيف البيانات وتحويلها الى صيغة رقمية مناسبة تم الانتقال الى مرحلة بناء نموذج الانحدار الخطي باستخدام Pipeline من مكتبة sklearn.

في هذا الجزء من الكود تم انشاء متغير باسم lr\_pipe يحتوي على مسار متكامل لمعالجة البيانات وتطبيق النموذج. يتكون هذا المسار من خطوتين رئيسيتين:

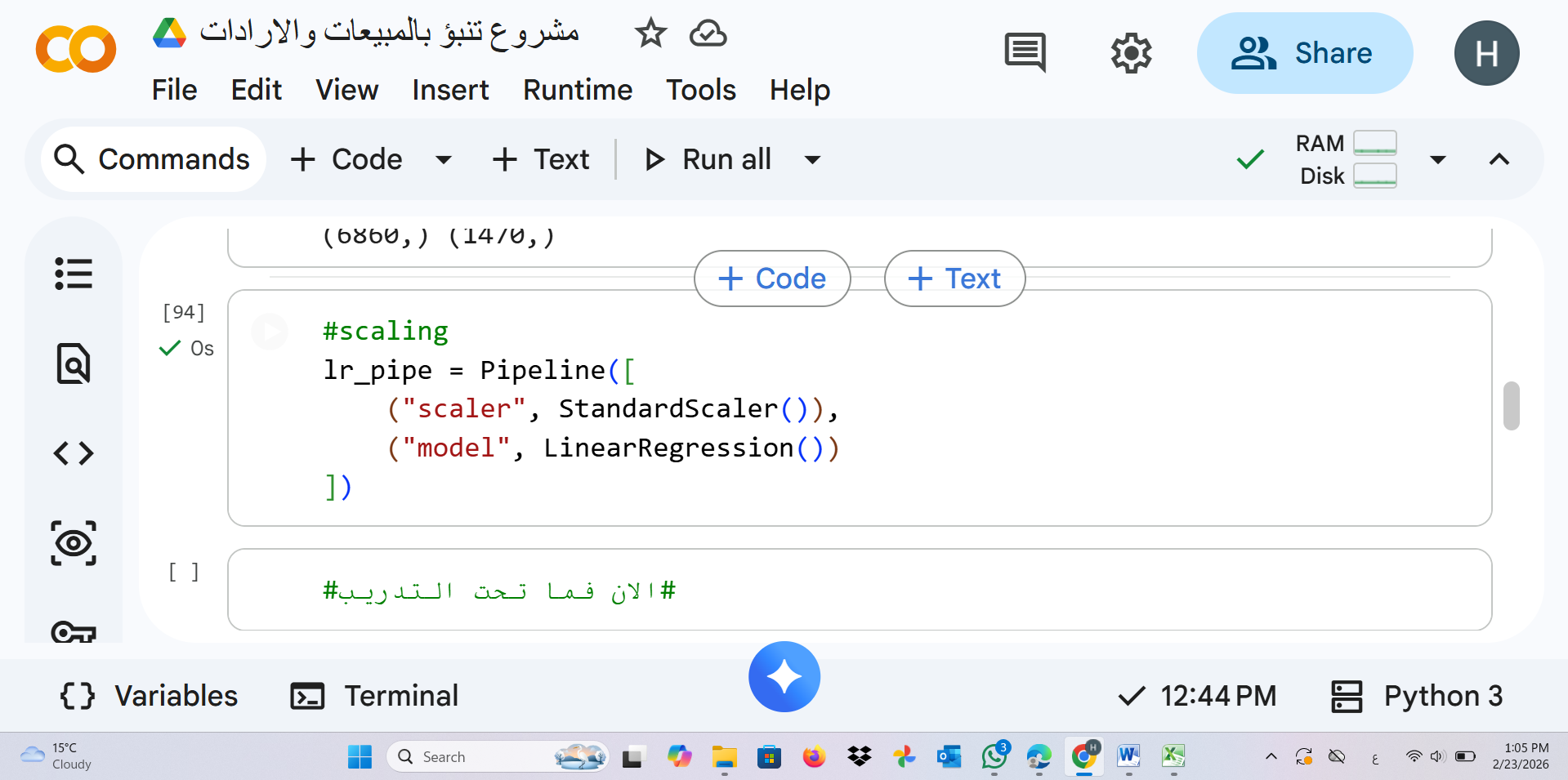
الخطوة الاولى هي StandardScaler وتم تسميتها داخل الـ Pipeline باسم "scaler". وظيفة هذه الخطوة هي توحيد مقياس المتغيرات بحيث يصبح متوسط كل متغير يساوي صفر والانحراف المعياري يساوي واحد. هذه الخطوة مهمة لان المتغيرات قد تكون على مقاييس مختلفة مثل عدد الطلبات او قيمة المبيعات او متغيرات زمنية، وتوحيدها يساعد النموذج على التعلم بشكل اكثر استقرارا.

الخطوة الثانية هي LinearRegression وتم تسميتها باسم "model". هذه هي الخوارزمية المسؤولة عن تعلم العلاقة بين المتغيرات المستقلة والمتغير التابع وهو المبيعات. يقوم النموذج بحساب المعاملات التي تقلل الخطأ بين القيم الحقيقية والقيم المتوقعة.

استخدام Pipeline هنا يضمن ان عملية التقييس والتدريب تتم بشكل منظم ومتسلسل، كما يمنع حدوث اخطاء مثل تطبيق التقييس على بيانات الاختبار بطريقة مختلفة عن بيانات التدريب. وهذا يعتبر اسلوبا احترافيا في بناء نماذج التعلم الالي.

بعد انشاء هذا المسار يصبح النموذج جاهزا لمرحلة التدريب باستخدام بيانات التدريب.

وهذا سكرين يوضح تنفيذ خطوة إنشاء الـ Pipeline التي تجمع بين StandardScaler ونموذج LinearRegression ضمن مسار واحد منظم تمهيدا لبدء عملية التدريب.



**تنفيذ عملية التدريب على بيانات التدريب**

بعد انشاء خط المعالجة باستخدام Pipeline تم تنفيذ عملية التدريب الفعلي للنموذج باستخدام الامر fit على بيانات التدريب X\_train و y\_train.

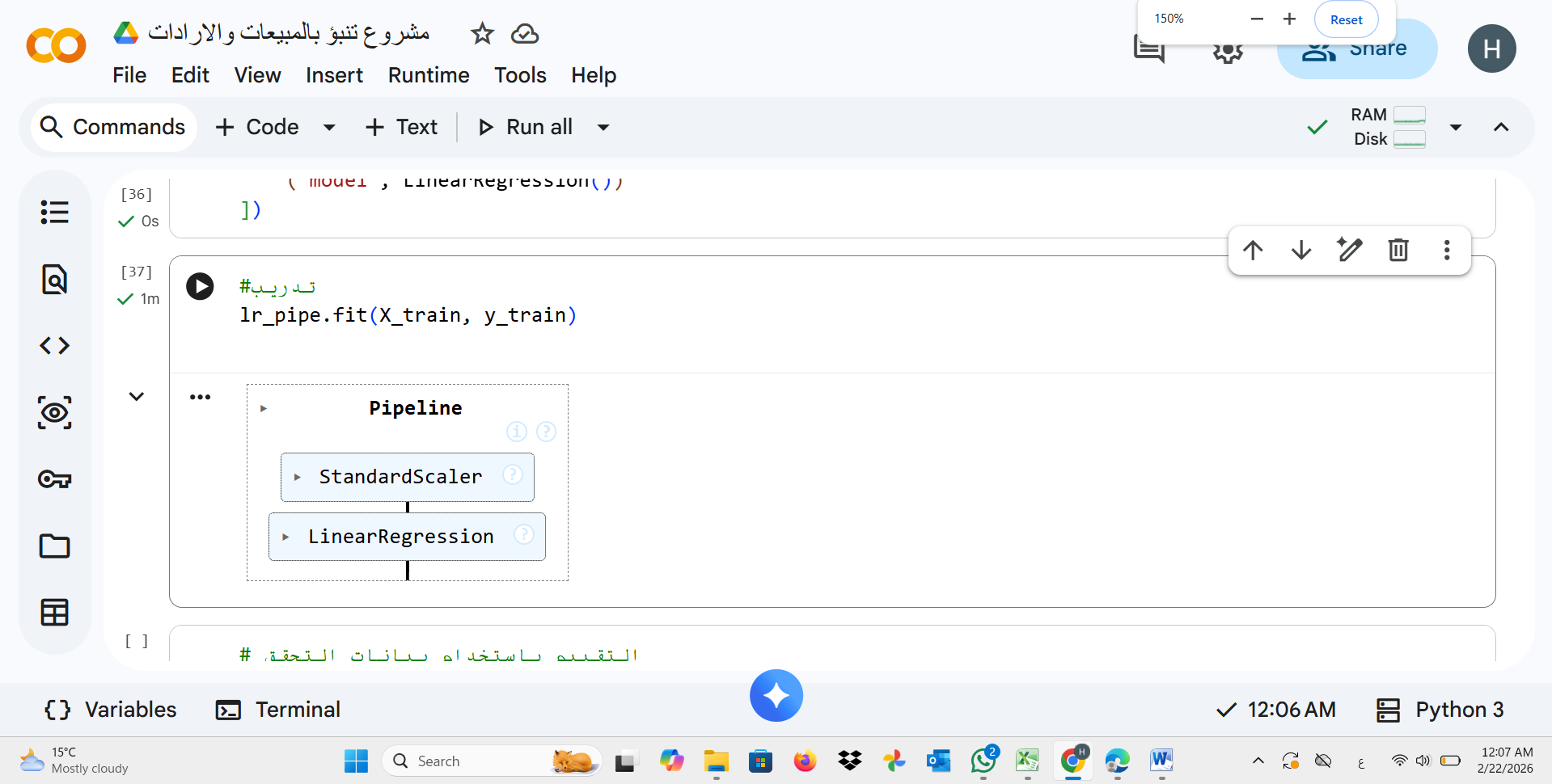
في هذه المرحلة يقوم StandardScaler بحساب المتوسط والانحراف المعياري لكل خاصية داخل بيانات التدريب ثم تحويل القيم الى مقياس موحد بعد ذلك يقوم نموذج LinearRegression بحساب معاملات الانحدار التي تمثل افضل علاقة خطية بين المتغيرات المستقلة وقيمة المبيعات.

من الناحية الرياضية يعتمد النموذج على تقليل مجموع مربعات الاخطاء بين القيم الحقيقية والقيم المتوقعة وذلك لايجاد افضل خط يصف العلاقة بين الخصائص وقيمة Sales.

تدريب النموذج على بيانات التدريب فقط يضمن عدم تسرب معلومات من بيانات الاختبار مما يحافظ على عدالة عملية التقييم لاحقا عند اختبار النموذج على بيانات جديدة لم يشاهدها من قبل.

نجاح تنفيذ الامر fit وظهور هيكل Pipeline في النتائج يؤكد ان النموذج تم تدريبه بنجاح واصبح جاهزا لمرحلة التقييم وقياس الاداء باستخدام مقاييس مثل R2 و MSE.

توضح لقطة الشاشة التالية تنفيذ عملية تدريب النموذج باستخدام lr\_pipe.fit على بيانات التدريب وظهور مكونات Pipeline التي تشمل StandardScaler و LinearRegression بعد اكتمال التدريب بنجاح



**تقييم النموذج باستخدام بيانات التحقق**

بعد الانتهاء من تدريب نموذج الانحدار الخطي تم الانتقال الى مرحلة تقييم الاداء باستخدام بيانات التحقق X\_val و y\_val وذلك لقياس قدرة النموذج على التعميم على بيانات لم يتم استخدامها اثناء التدريب

في البداية تم التأكد من ان اعمدة X\_val مطابقة لاعمدة X\_train من خلال تطبيق get\_dummies ثم استخدام reindex لضمان نفس عدد الخصائص ونفس الترتيب مما يمنع حدوث اي خطأ اثناء عملية التنبؤ

بعد ذلك تم استخدام النموذج المدرب للتنبؤ بقيم المبيعات على بيانات التحقق باستخدام الامر predict ثم تم حساب مقاييس الاداء وهي MSE و RMSE و R2

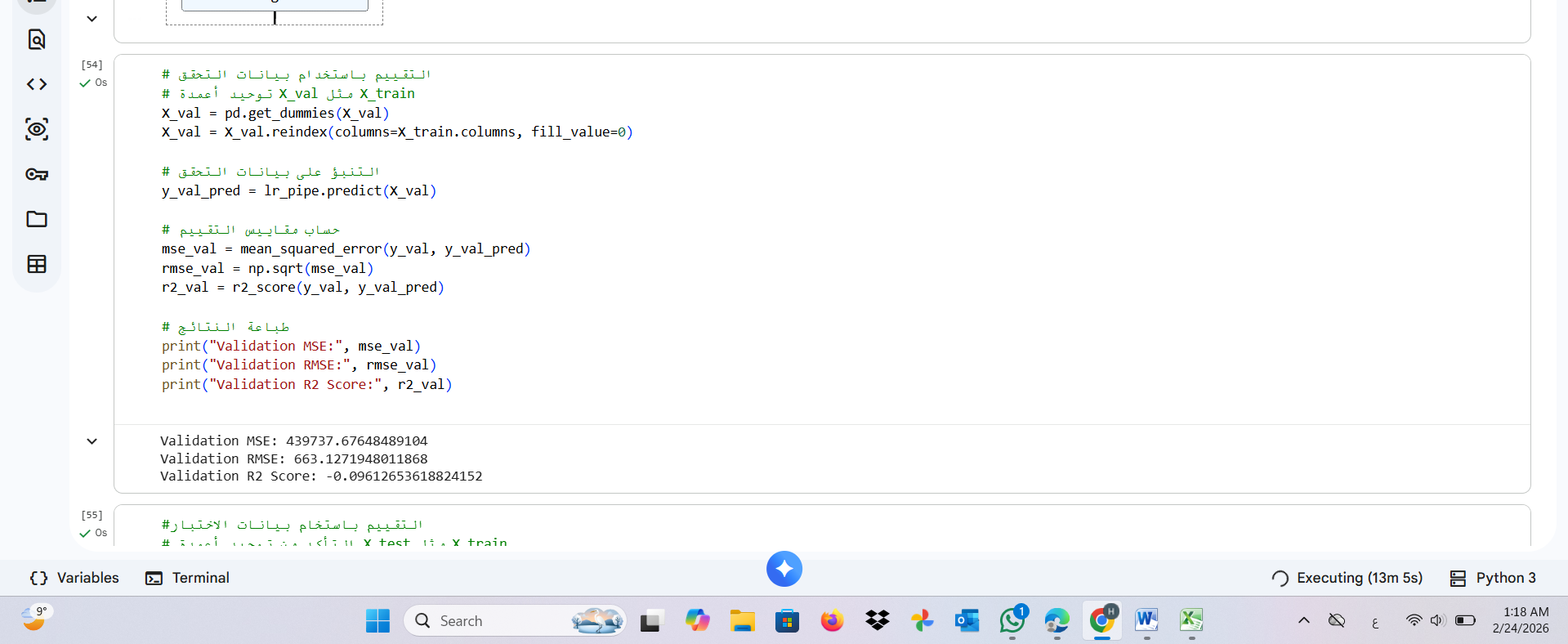
اظهرت النتائج ان قيمة Validation MSE بلغت حوالي 439737 بينما بلغت قيمة RMSE حوالي 663 وهذا يعني ان متوسط الخطأ في التنبؤ يعادل تقريبا 663 وحدة من المبيعات وهي قيمة مرتفعة جدا مقارنة بمتوسط المبيعات الفعلي

اما قيمة R2 فقد بلغت حوالي -0.096 وهي قيمة سالبة مما يعني ان النموذج لم يتمكن من تفسير التباين في البيانات بل كان اداؤه اسوأ من نموذج بسيط يعتمد على توقع المتوسط فقط

تشير هذه النتيجة الى ان نموذج الانحدار الخطي غير مناسب لتمثيل العلاقة بين المتغيرات وقيمة المبيعات في هذه الحالة حيث يبدو ان العلاقة بين الخصائص والمبيعات ليست خطية بالكامل او ان هناك تعقيد وتفاعلات بين المتغيرات لا يستطيع النموذج الخطي تمثيلها

تعد هذه المرحلة مهمة جدا لانها توضح ان النموذج الاولي لم يحقق الاداء المطلوب مما يبرر الانتقال الى نموذج اكثر مرونة مثل Random Forest لتحسين القدرة على التنبؤ والتعامل مع العلاقات غير الخطية

توضح لقطة الشاشة التالية حساب قيم MSE و RMSE و R2 على بيانات التحقق واظهار النتائج التي تبين ضعف اداء نموذج الانحدار الخطي قبل تنفيذ مرحلة التحسين



**التقييم النهائي باستخدام بيانات الاختبار**

بعد تقييم نموذج الانحدار الخطي على بيانات التحقق تم الانتقال الى التقييم النهائي باستخدام بيانات الاختبار X\_test و y\_test وهي بيانات لم يتم استخدامها اثناء التدريب او ضبط النموذج وذلك للحصول على قياس واقعي لقدرة النموذج على التعميم على بيانات جديدة تماما

في البداية تم التأكد من مطابقة اعمدة X\_test لاعمدة X\_train باستخدام get\_dummies ثم reindex لضمان نفس عدد الخصائص ونفس ترتيبها قبل تنفيذ عملية التنبؤ

اظهرت النتائج ان قيمة Test MSE بلغت حوالي 651237 وهي قيمة مرتفعة مما يدل على وجود فروقات كبيرة بين القيم الحقيقية والمتوقعة عند حساب مربع الخطأ

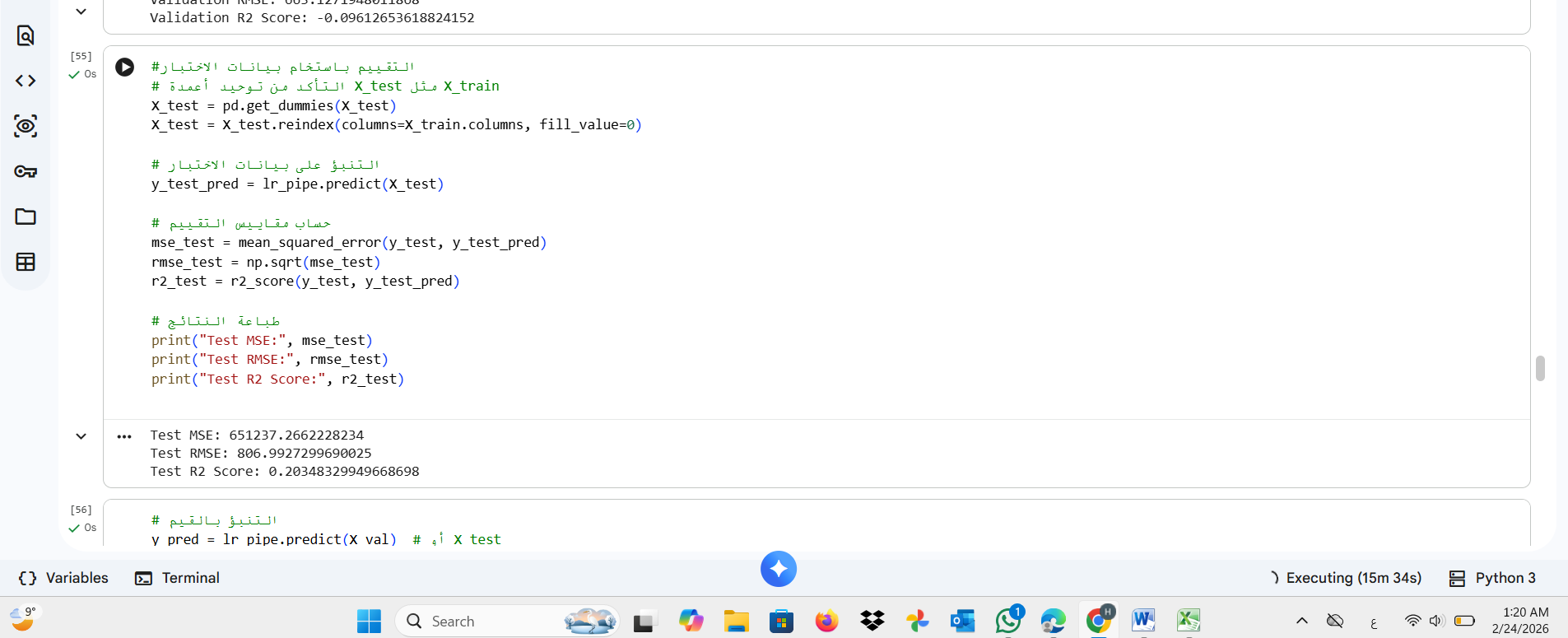
كما بلغت قيمة RMSE حوالي 807 وحدة من المبيعات مما يعني ان متوسط الخطأ في التنبؤ يساوي تقريبا 807 وحدة وهي قيمة كبيرة مقارنة بمتوسط المبيعات الفعلي

اما قيمة R2 فقد بلغت حوالي 0.20 مما يعني ان النموذج استطاع تفسير حوالي 20 بالمئة فقط من التباين في بيانات المبيعات بينما بقي الجزء الاكبر من التباين غير مفسر

عند مقارنة نتائج الاختبار بنتائج التحقق نلاحظ ان الاداء ما زال ضعيفا ولم يحقق تحسنا ملحوظا مما يؤكد ان نموذج الانحدار الخطي غير قادر على تمثيل العلاقة الحقيقية بين المتغيرات وقيمة المبيعات بشكل فعال

تشير هذه النتائج الى ان العلاقة بين الخصائص والمبيعات ليست خطية بالكامل وان البيانات تحتوي على انماط وتفاعلات معقدة لا يستطيع النموذج الخطي تمثيلها ولذلك كان من الضروري الانتقال الى نموذج اكثر مرونة مثل Random Forest لتحسين القدرة التنبؤية

توضح لقطة الشاشة التالية حساب قيم MSE و RMSE و R2 على بيانات الاختبار واظهار النتائج النهائية التي تبين محدودية اداء نموذج الانحدار الخطي عند تطبيقه على بيانات جديدة



**التحليل البصري لاداء النموذج**

بعد حساب المقاييس الرقمية تم استخدام التحليل البصري لفهم سلوك النموذج بشكل اعمق من خلال رسم مخطط القيم الحقيقية مقابل القيم المتوقعة وكذلك تحليل توزيع البواقي.

في الرسم الاول الذي يعرض القيم الحقيقية مقابل القيم المتوقعة نلاحظ وجود علاقة طردية بين القيم مما يدل على ان النموذج تمكن من التقاط الاتجاه العام للعلاقة بين المتغيرات وقيمة المبيعات حيث تميل النقاط الى التوزع حول الخط القطري الذي يمثل التنبؤ المثالي.

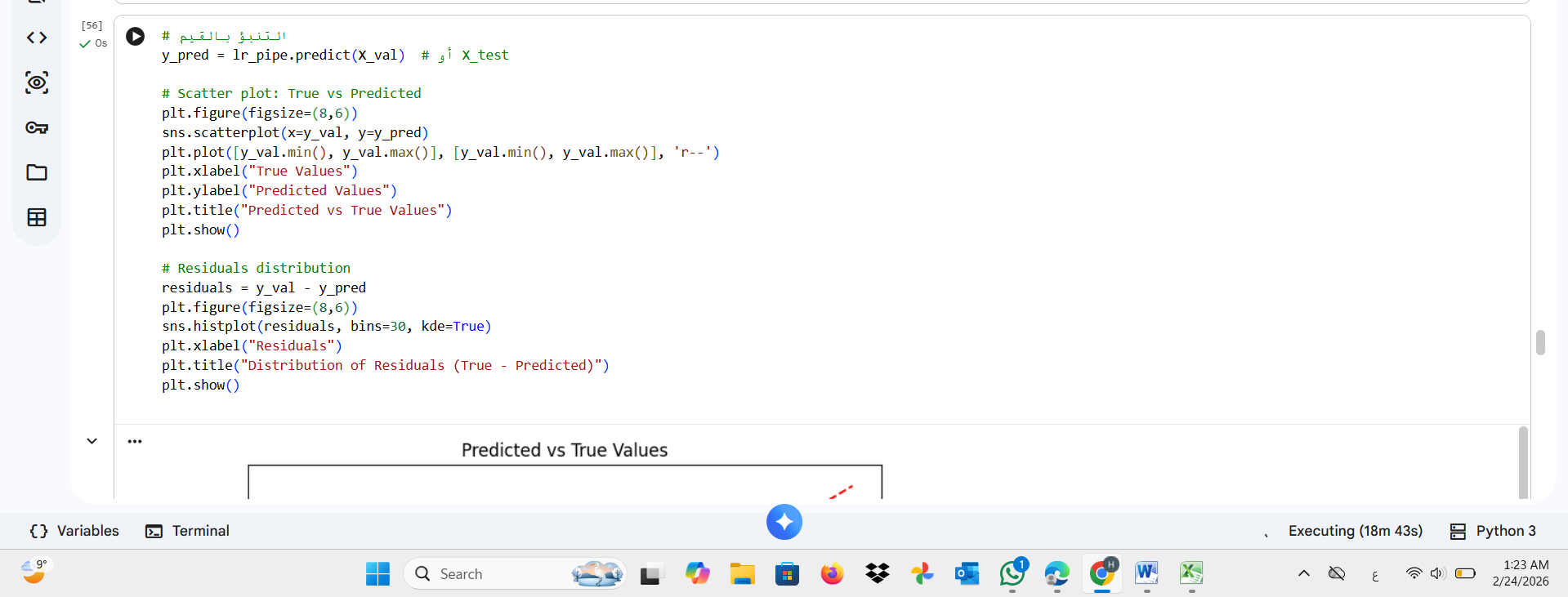
ومع ذلك نلاحظ وجود تشتت واضح حول الخط خاصة عند القيم المرتفعة مما يشير الى ان دقة النموذج تقل عند مستويات المبيعات العالية وهذا يتماشى مع قيمة R2 المتوسطة التي تم الحصول عليها سابقا.

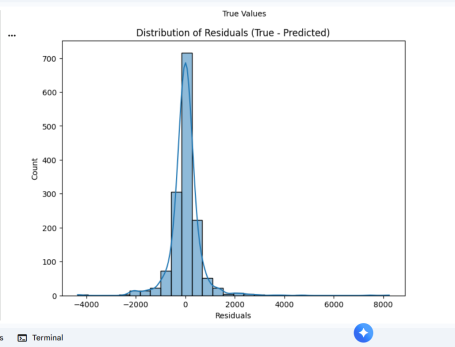
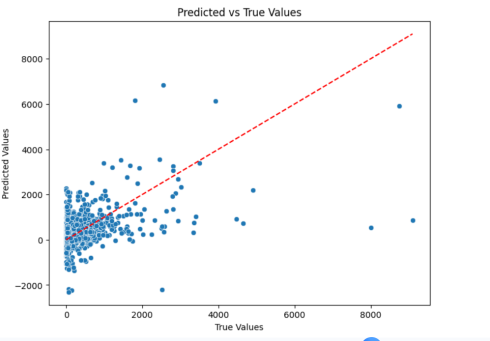
اما في الرسم الثاني الذي يعرض توزيع البواقي وهي الفرق بين القيم الحقيقية والمتوقعة فنلاحظ ان التوزيع يقترب من الشكل الطبيعي ويتمركز حول الصفر وهذا يدل على عدم وجود انحياز كبير في اتجاه معين بمعنى ان النموذج لا يميل بشكل منهجي الى المبالغة او التقليل من القيم.

لكن وجود بعض القيم المتطرفة في اطراف التوزيع يشير الى ان النموذج لا يستطيع تفسير جميع الانماط خاصة الحالات ذات المبيعات العالية او المنخفضة جدا.

من الناحية التحليلية يؤكد هذا ان نموذج الانحدار الخطي التقط العلاقة العامة لكنه غير كاف لتمثيل العلاقات المعقدة او غير الخطية الموجودة في البيانات مما يدعم الحاجة الى تجربة نماذج اكثر مرونة مثل Random Forest او Gradient Boosting.

توضح لقطة الشاشة الاولى العلاقة بين القيم الحقيقية والمتوقعة بينما توضح لقطة الشاشة الثانية توزيع البواقي مما يساعد في تقييم جودة النموذج بشكل بصري.





**تحسين النموذج باستخدام Random Forest**

بعد تحليل نتائج نموذج الانحدار الخطي وملاحظة ان قيمة R2 كانت منخفضة نسبيا وان النموذج لم يتمكن من تفسير نسبة كبيرة من التباين في بيانات المبيعات تم اعتبار ذلك مؤشرا على وجود علاقات غير خطية بين المتغيرات وقيمة Sales

الانحدار الخطي يفترض وجود علاقة خطية مباشرة بين الخصائص والمتغير التابع بينما التحليل البصري للبواقي اظهر تشتتا واضحا خاصة عند القيم المرتفعة مما يدل على ان النموذج غير قادر على تمثيل جميع الانماط داخل البيانات.

بناء على ذلك تم تنفيذ مرحلة تحسين للنموذج من خلال استخدام Random Forest Regressor وهو نموذج يعتمد على اشجار القرار ويستطيع التقاط العلاقات غير الخطية والتفاعلات المعقدة بين المتغيرات

تم اختيار عدد 200 شجرة لزيادة استقرار التنبؤات وتقليل التباين حيث يقوم النموذج ببناء عدة اشجار ثم حساب متوسط النتائج مما يؤدي الى تقليل الخطأ وتحسين القدرة على التعميم

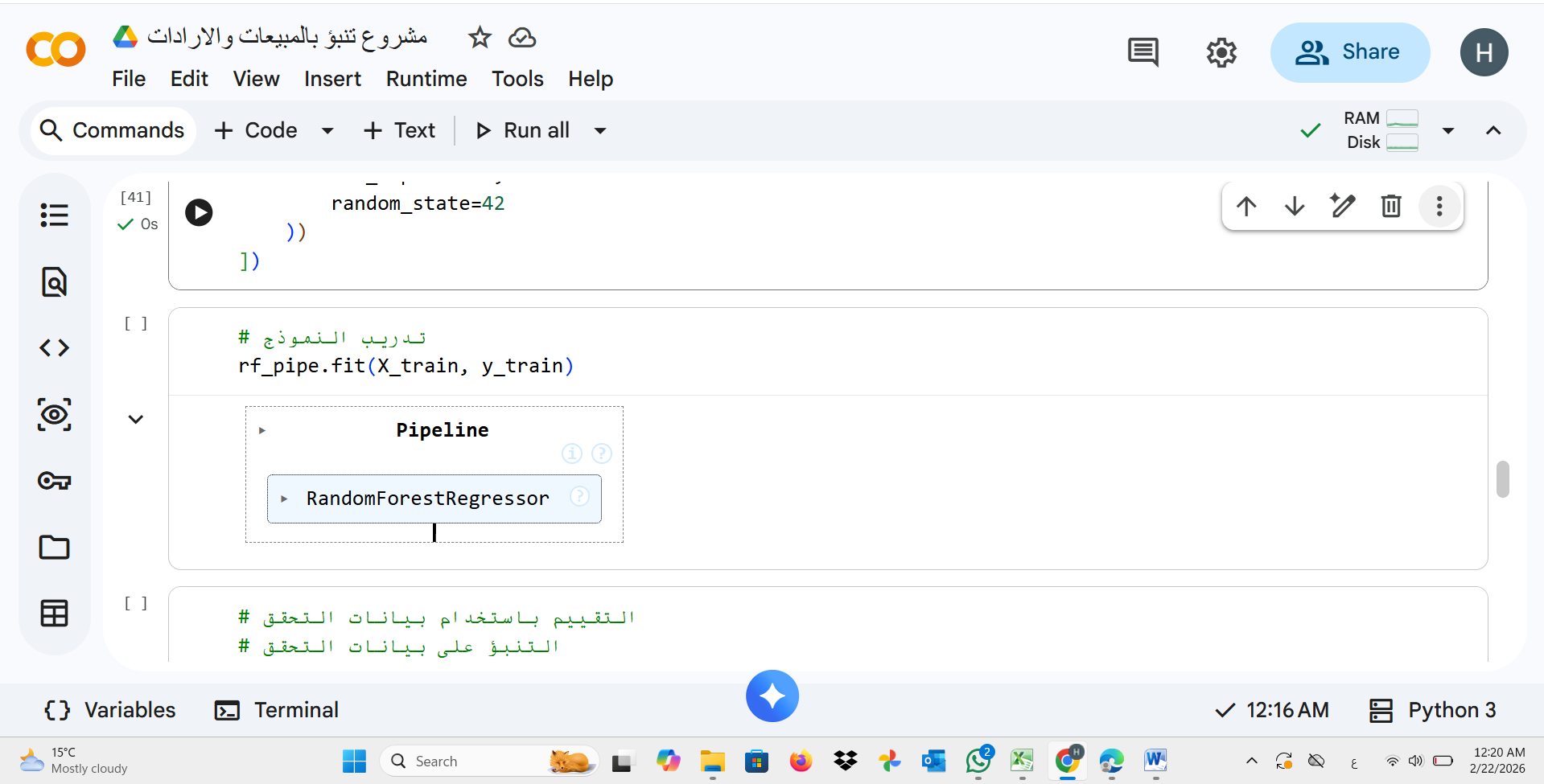


توضح لقطة الشاشة التالية بداية مرحلة تحسين النموذج من خلال انشاء Pipeline جديد يعتمد على خوارزمية RandomForestRegressor بدلا من الانحدار الخطي

يظهر في الكود استيراد النموذج من مكتبة sklearn.ensemble ثم تعريف rf\_pipe باستخدام Pipeline بحيث يحتوي على نموذج Random Forest فقط دون الحاجة الى StandardScaler لان هذا النوع من النماذج لا يتأثر باختلاف مقاييس المتغيرات مثل نماذج الانحدار الخطي.

تم تحديد عدد الاشجار n\_estimators بقيمة 200 وذلك لزيادة استقرار التنبؤات وتقليل التباين حيث يقوم النموذج ببناء عدد كبير من اشجار القرار ثم حساب متوسط النتائج كما تم ترك max\_depth بدون تحديد للسماح للاشجار بالنمو بشكل كامل واستخدام random\_state=42 لضمان ثبات النتائج عند اعادة التشغيل.

تعكس هذه الخطوة انتقالا منهجيا من نموذج خطي بسيط الى نموذج اكثر تعقيدا ومرونة بهدف تحسين القدرة على التقاط العلاقات غير الخطية داخل بيانات المبيعات وتحسين دقة التنبؤ.



توضح لقطة الشاشة التالية تنفيذ عملية تدريب نموذج Random Forest باستخدام الامر rf\_pipe.fit على بيانات التدريب X\_train و y\_train.

بعد انشاء النموذج وتحديد المعاملات تم تطبيق عملية التدريب حيث يقوم النموذج ببناء 200 شجرة قرار اعتمادا على عينات مختلفة من البيانات ثم يجمع نتائج هذه الاشجار من خلال حساب المتوسط للحصول على التنبؤ النهائي.

يظهر في المخرجات هيكل Pipeline الذي يحتوي على RandomForestRegressor مما يدل على ان عملية التدريب اكتملت بنجاح واصبح النموذج جاهزا لمرحلة التقييم باستخدام بيانات التحقق والاختبار.

تعد هذه الخطوة جزءا من مرحلة التحسين حيث يتم مقارنة اداء هذا النموذج مع نموذج الانحدار الخطي السابق لمعرفة ما اذا كان استخدام نموذج يعتمد على الاشجار سيؤدي الى تحسين دقة التنبؤ وزيادة قيمة R2 وتقليل الخطأ.

**مقارنة نتائج Random Forest مع الانحدار الخطي**

بعد تنفيذ مرحلة التحسين باستخدام نموذج Random Forest تم تقييم الاداء على بيانات التحقق ومقارنة النتائج مع نموذج الانحدار الخطي السابق.

في نموذج الانحدار الخطي كانت قيمة R2 سالبة وبلغت حوالي -0.096 مما يعني ان النموذج لم يتمكن من تفسير التباين في البيانات وكان اداؤه اسوأ من نموذج بسيط يعتمد على توقع المتوسط.

بعد استخدام Random Forest ارتفعت قيمة R2 الى حوالي 0.32 مما يعني ان النموذج اصبح قادرا على تفسير ما يقارب 32 بالمئة من التباين في بيانات المبيعات وهو تحسن واضح مقارنة بالنموذج الخطي.

كما ان قيمة RMSE انخفضت من حوالي 663 في الانحدار الخطي الى حوالي 522 في Random Forest وهذا يدل على انخفاض متوسط الخطأ في التنبؤ بشكل ملحوظ.

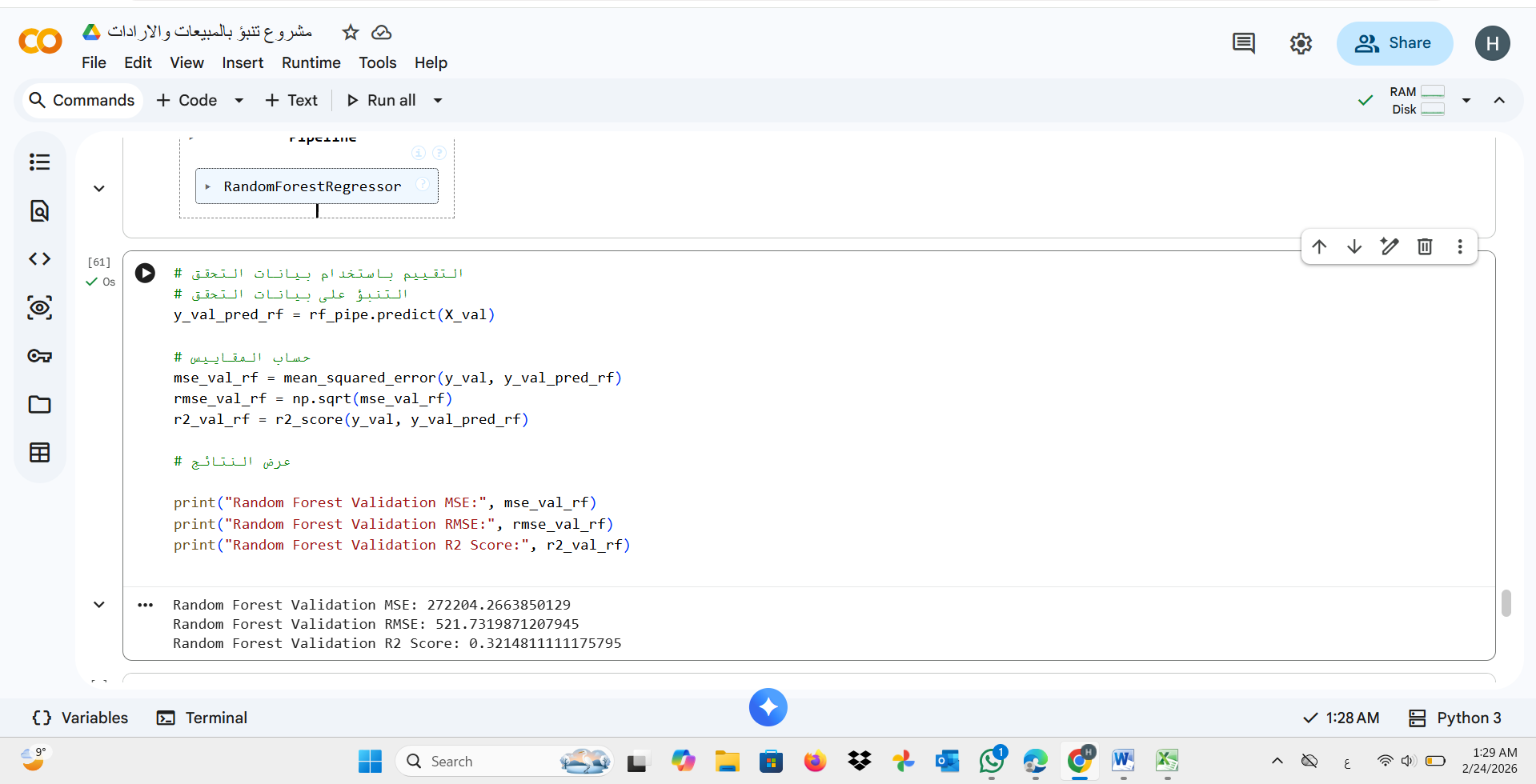
اما قيمة MSE فقد انخفضت من حوالي 439737 الى حوالي 272204 مما يعكس تحسنا ملموسا في دقة النموذج.

يشير هذا التحسن الى ان العلاقة بين المتغيرات وقيمة المبيعات ليست خطية بالكامل وان استخدام نموذج يعتمد على الاشجار وقادر على تمثيل العلاقات غير الخطية والتفاعلات بين الخصائص كان قرارا مناسبا.

من الناحية التحليلية يظهر ان نموذج Random Forest استطاع التقاط انماط معقدة داخل البيانات لا يستطيع الانحدار الخطي تمثيلها بسبب اعتماده على افتراض العلاقة الخطية.

بناء على ذلك يمكن اعتبار Random Forest نموذجا اكثر فعالية من الانحدار الخطي في هذه الحالة رغم ان نسبة التفسير ما زالت متوسطة مما يشير الى امكانية تحسين الاداء بشكل اكبر باستخدام ضبط المعاملات او هندسة خصائص اضافية.

توضح لقطة الشاشة التالية نتائج تقييم نموذج Random Forest على بيانات التحقق والتي تبين التحسن الواضح في قيم R2 و RMSE مقارنة بالنموذج السابق.



**التقييم النهائي لنموذج Random Forest على بيانات الاختبار**

بعد تقييم نموذج Random Forest على بيانات التحقق تم اجراء التقييم النهائي باستخدام بيانات الاختبار التي لم يتم استخدامها اثناء التدريب او التحسين وذلك للحصول على قياس حقيقي لقدرة النموذج على التعميم على بيانات جديدة

اظهرت النتائج ان قيمة Test MSE بلغت حوالي 522178 مما يعكس وجود فروقات بين القيم الحقيقية والمتوقعة عند حساب مربع الخطأ

كما بلغت قيمة RMSE حوالي 723 وحدة من المبيعات مما يعني ان متوسط الخطأ في التنبؤ على بيانات جديدة يقارب 723 وحدة

اما قيمة R2 فقد بلغت حوالي 0.36 مما يعني ان النموذج اصبح قادرا على تفسير حوالي 36 بالمئة من التباين في قيم المبيعات

عند مقارنة هذه النتائج مع نموذج الانحدار الخطي الذي بلغت قيمة R2 فيه حوالي 0.20 على بيانات الاختبار نلاحظ تحسنا واضحا في القدرة التفسيرية للنموذج كما ان قيمة R2 في Random Forest اعلى مما يؤكد ان النموذج اكثر كفاءة في تمثيل العلاقة بين المتغيرات وقيمة المبيعات

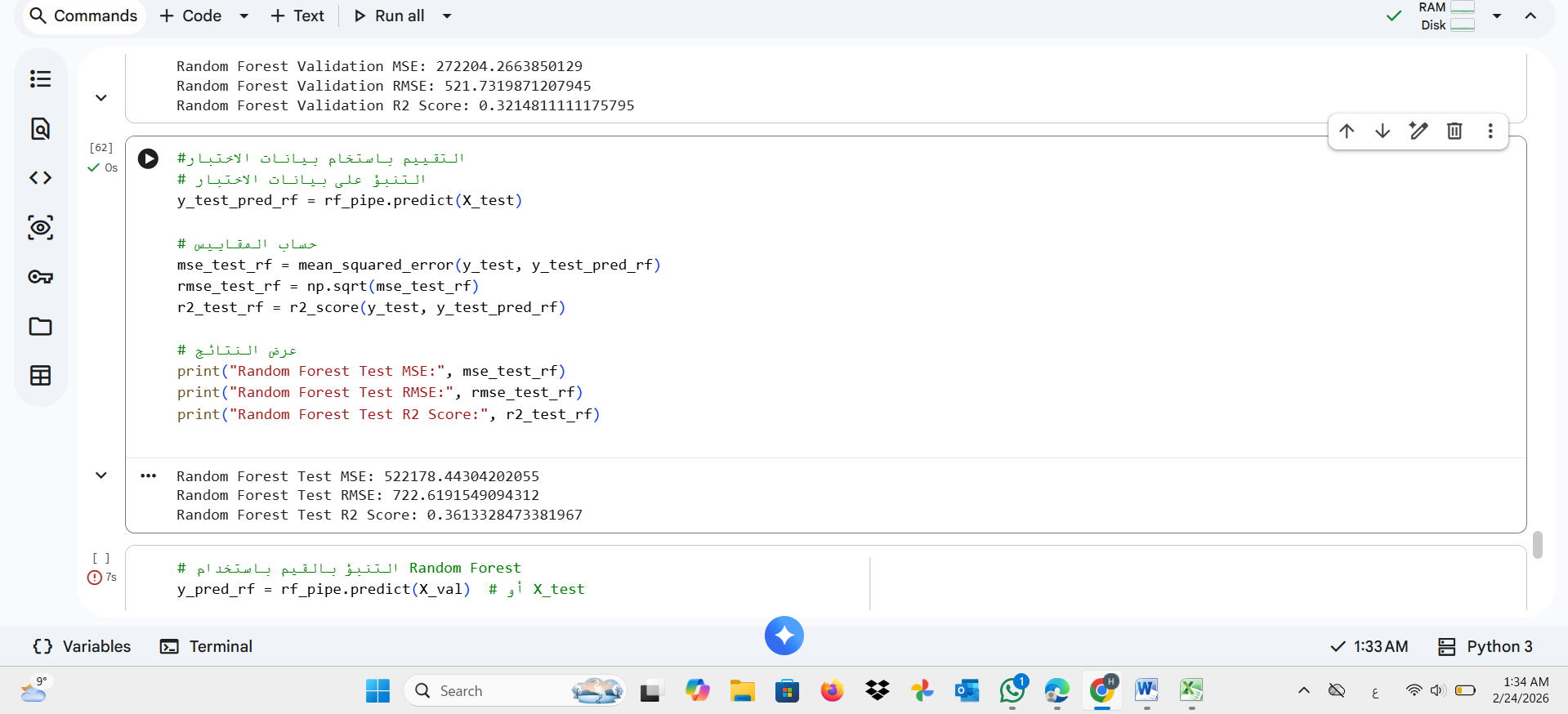
من المهم ملاحظة ان قيمة R2 على بيانات الاختبار قريبة من قيمتها على بيانات التحقق مما يدل على ان النموذج لا يعاني من مشكلة overfitting كبيرة وان اداءه مستقر نسبيا عند تطبيقه على بيانات جديدة

يعكس هذا التحسن ان اختيار نموذج يعتمد على الاشجار كان اكثر ملاءمة لطبيعة البيانات حيث تمكن من تمثيل العلاقات غير الخطية والتفاعلات بين المتغيرات بشكل افضل من النموذج الخطي

اما فيما يتعلق باستخدام Cross Validation فان ثبات قيم R2 بين الطيات المختلفة مع انحراف معياري منخفض يدل على استقرار النموذج وعدم اعتماده على تقسيم واحد فقط للبيانات مما يعزز من مصداقية النتائج

بناء على التحليل الكمي والمقارنة بين النماذج يمكن اعتبار Random Forest النموذج الاكثر فعالية مقارنة بالانحدار الخطي في هذا المشروع رغم ان نسبة التفسير ما زالت متوسطة مما يشير الى امكانية تحسين الاداء مستقبلا من خلال ضبط المعاملات او اضافة خصائص جديدة

توضح لقطة الشاشة التالية نتائج تقييم نموذج Random Forest على بيانات الاختبار والتي تبين التحسن في قيم R2 مقارنة بالنموذج الخطي السابق



**تحليل بصري لاداء نموذج Random Forest**

يوضح الرسم الاول العلاقة بين القيم الحقيقية للمبيعات والقيم التي توقعها نموذج Random Forest حيث يمثل الخط الاحمر المتقطع الحالة المثالية التي تتطابق فيها القيم المتوقعة مع القيم الحقيقية بنسبة 100 بالمئة.

نلاحظ ان عددا كبيرا من النقاط يتركز في الجزء السفلي الايسر من الرسم اي ضمن نطاق القيم المنخفضة والمتوسطة للمبيعات ويظهر تقارب نسبي من خط المثالية مما يدل على ان النموذج يحقق اداء مقبولا في هذا النطاق.

في المقابل يظهر تشتت واضح للنقاط كلما ارتفعت قيم المبيعات حيث تبتعد بعض التوقعات بشكل ملحوظ عن الخط الاحمر خاصة في القيم العالية جدا وهذا يشير الى ان النموذج يواجه صعوبة في التنبؤ الدقيق بالعمليات ذات المبيعات المرتفعة.

كما نلاحظ وجود بعض القيم التي تم التقليل من تقديرها او المبالغة في تقديرها في النطاقات العليا مما يعكس محدودية النموذج في تمثيل القيم المتطرفة بشكل دقيق.

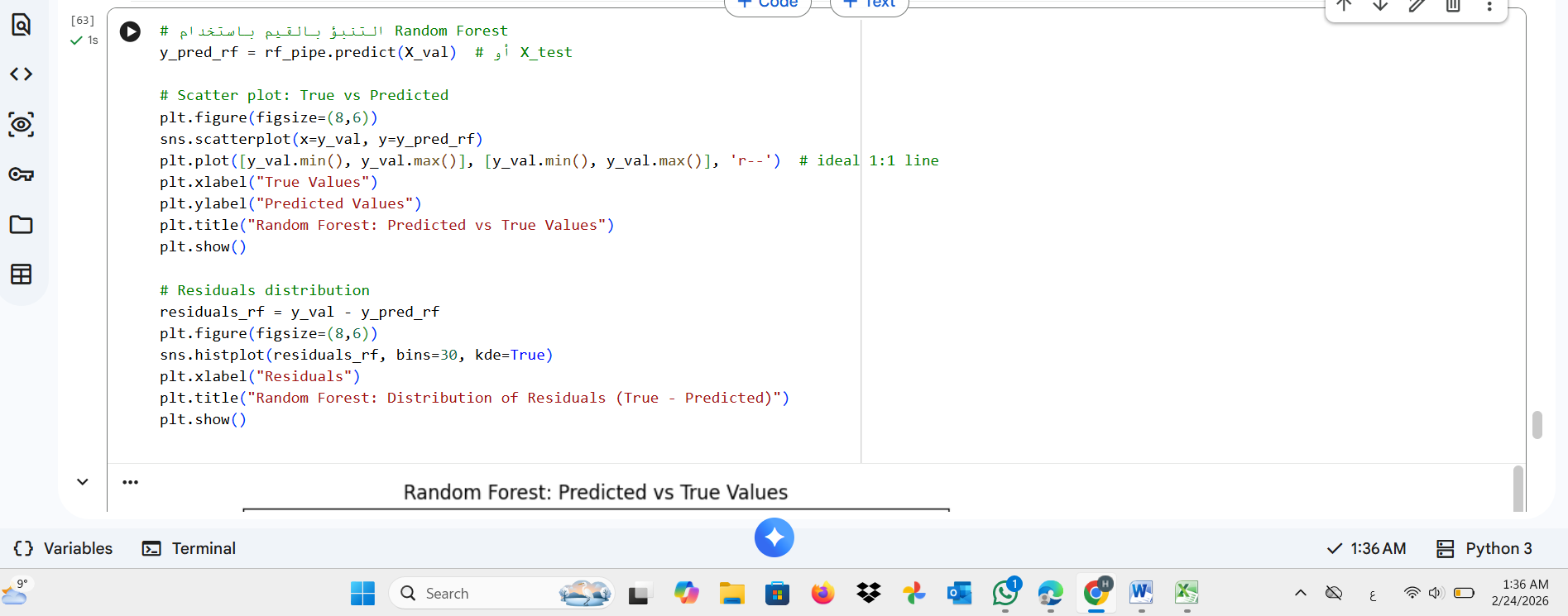
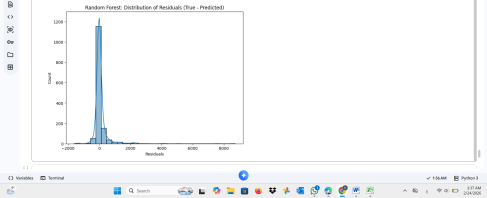
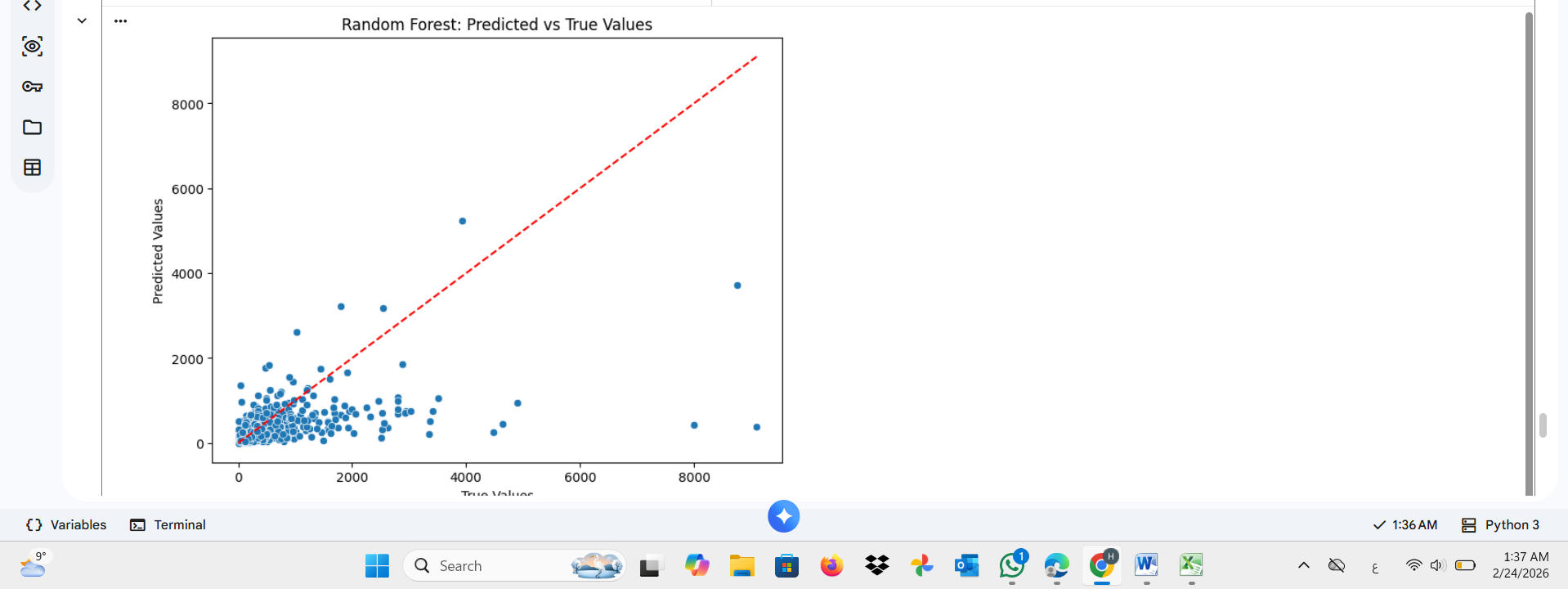
اما الرسم الثاني فيعرض توزيع الاخطاء Residuals والتي تمثل الفرق بين القيم الحقيقية والمتوقعة

نلاحظ ان معظم الاخطاء متمركزة حول الصفر مما يدل على عدم وجود انحياز واضح نحو التوقع الزائد او الناقص وهذا مؤشر ايجابي يدل على ان النموذج متوازن نسبيا

ومع ذلك يظهر ذيل ممتد نحو القيم الموجبة مما يعني ان هناك بعض الحالات التي كان فيها الفرق بين القيمة الحقيقية والمتوقعة كبيرا خاصة في المبيعات المرتفعة وهذا يتوافق مع ما ظهر في الرسم الاول من تشتت عند القيم الكبيرة

بشكل عام تؤكد هذه التحليلات البصرية النتائج الرقمية السابقة حيث يحقق نموذج Random Forest تحسنا واضحا مقارنة بالنموذج الخطي لكنه ما زال يواجه تحديات في تمثيل القيم المتطرفة بدقة كاملة

تدعم لقطات الشاشة السابقة هذا التحليل من خلال عرض العلاقة بين القيم الحقيقية والمتوقعة وتوزيع الاخطاء بشكل بصري مما يعزز مصداقية التقييم الكمي للنموذج.

**تحليل اسباب ارتفاع الخطأ في القيم المرتفعة**

يظهر من الرسم البياني ان نموذج Random Forest يحقق اداء افضل في نطاق القيم المنخفضة والمتوسطة بينما يزداد التشتت في التوقعات عند القيم المرتفعة جدا للمبيعات ويمكن تفسير ذلك من منظور احصائي وتعلم آلي بعدة عوامل

اولا توزيع بيانات المبيعات غير متوازن حيث ان معظم العمليات تقع ضمن نطاق منخفض الى متوسط بينما تمثل القيم المرتفعة نسبة صغيرة من البيانات وهذا يعني ان النموذج يتدرب على عدد محدود من الامثلة ذات القيم العالية مما يقلل من قدرته على تعلم نمطها بدقة

ثانيا نموذج Random Forest يعتمد على تقسيم البيانات الى مناطق فرعية باستخدام اشجار القرار وعند وجود قيم متطرفة قليلة فان هذه القيم قد تقع في اوراق تحتوي على عدد محدود من العينات مما يؤدي الى تقديرات اقل استقرارا

ثالثا عند استخدام معيار تقليل الخطأ المربع داخل الاشجار فان النموذج يميل الى تقليل الخطأ الاجمالي على جميع البيانات وليس التركيز على الحالات النادرة ذات القيم العالية وبالتالي يتم تحسين الاداء العام على حساب دقة التنبؤ في الاطراف القصوى للتوزيع

رابعا قد تكون المبيعات المرتفعة مرتبطة بعوامل اضافية غير ممثلة بشكل كاف في الخصائص الحالية مثل حملات ترويجية خاصة او ظروف موسمية معينة لم يتم تضمينها صراحة ضمن المتغيرات المستخدمة مما يجعل التنبؤ بها اكثر صعوبة

من الناحية الاحصائية يظهر ذلك في شكل توزيع الاخطاء حيث يوجد ذيل ممتد يدل على وجود حالات قليلة ذات خطأ كبير بينما تبقى معظم الاخطاء صغيرة ومتمركزة حول الصفر

بناء على ذلك يمكن القول ان النموذج يحقق توازنا جيدا في الاداء العام لكنه لا يزال يواجه تحديا في تمثيل القيم المتطرفة وهو امر شائع في نماذج الانحدار عند التعامل مع بيانات ذات توزيع منحرف

يمكن مستقبلا تحسين هذا الجانب من خلال تجربة تقنيات مثل ضبط معاملات النموذج Hyperparameter Tuning او استخدام نماذج تعزيز Boosting او اضافة خصائص زمنية وموسمية اكثر تفصيلا لتحسين تمثيل الحالات ذات المبيعات المرتفعة.

**التحليل والتقييم النهائي**

عند مقارنة النموذجين يتضح ان نموذج الانحدار الخطي لم يكن قادرا على تمثيل طبيعة البيانات بشكل كاف حيث بلغت قيمة R2 على بيانات التحقق قيمة سالبة تقارب -0.096 مما يعني ان النموذج كان اسوأ من نموذج بسيط يعتمد على توقع متوسط المبيعات فقط كما كانت قيم RMSE و MSE مرتفعة بشكل واضح مما يدل على وجود خطأ كبير في التنبؤ ويشير ذلك الى ان افتراض العلاقة الخطية المباشرة بين المتغيرات المستقلة وقيمة المبيعات لم يكن كافيا لتمثيل الانماط الفعلية داخل البيانات

في المقابل حقق نموذج Random Forest تحسنا ملحوظا حيث ارتفعت قيمة R2 على بيانات التحقق الى حوالي 0.32 كما بلغت على بيانات الاختبار حوالي 0.36 وانخفضت قيم الخطأ مقارنة بالنموذج الخطي مما يعكس قدرة اعلى على تفسير التباين وتحقيق تنبؤات اكثر دقة

يمكن تفسير هذا التحسن من الناحية الرياضية بان الانحدار الخطي يعتمد على تمثيل المبيعات كمزيج خطي من الخصائص بينما تحتوي بيانات المبيعات الواقعية على علاقات غير خطية وتفاعلات معقدة بين المتغيرات مثل تأثير الزمن بالتزامن مع الفئة والمنطقة ونوع المنتج

يعتمد Random Forest على مبدأ Bagging حيث يتم انشاء عدة اشجار قرار على عينات مختلفة من البيانات ثم حساب متوسط التنبؤات مما يقلل من التباين ويعزز القدرة على التعميم كما ان اشجار القرار قادرة على تقسيم البيانات الى مناطق غير خطية وهو ما يسمح بالتقاط انماط معقدة لا يستطيع النموذج الخطي تمثيلها

رغم التحسن الواضح في اداء نموذج Random Forest الا ان قيمة R2 ما زالت متوسطة حيث لم تتجاوز 0.40 تقريبا مما يعني ان النموذج يفسر جزءا من التباين لكنه لا يفسر الغالبية العظمى من الاختلافات في بيانات المبيعات

عند مراجعة النتائج النهائية ومقارنتها بالاهداف المخطط لها في بداية المشروع يتضح ان الهدف المتمثل في تحقيق قيمة R2 لا تقل عن 0.80 لم يتحقق بالكامل حيث بلغت اعلى قيمة حوالي 0.36 باستخدام نموذج Random Forest

يمكن تفسير عدم الوصول الى المستوى المستهدف بعدة عوامل منها محدودية الخصائص المتاحة ووجود قيم متطرفة وتوزيع غير متوازن للمبيعات اضافة الى عدم تنفيذ ضبط شامل للمعاملات او هندسة خصائص متقدمة يمكن ان تحسن الاداء بشكل اكبر

رغم ذلك فان الانتقال من نموذج يحقق قيمة سالبة في R2 الى نموذج يحقق قيمة موجبة ومستقرة على بيانات التحقق والاختبار يمثل تحسنا جوهريا ويؤكد صحة الفرضية التي اشارت الى ان العلاقات بين المتغيرات ليست خطية بالكامل وان استخدام نموذج اكثر مرونة يؤدي الى تحسين الاداء

بناء على ذلك يمكن اعتبار نموذج Random Forest اكثر ملاءمة من الانحدار الخطي في هذا المشروع لكنه لا يزال بحاجة الى تحسينات مستقبلية للوصول الى مستوى اداء اعلى يتوافق مع الهدف الموضوع في بداية الدراسة

**التحسينات المستقبلية**

يمكن تطوير هذا المشروع مستقبلا من خلال عدة اتجاهات لزيادة دقة النموذج وتحسين قابليته للتطبيق العملي اولا يمكن تجربة نماذج اكثر تقدما مثل Gradient Boosting او XGBoost والتي غالبا ما تحقق اداء اعلى في مشكلات الانحدار المعقدة كما يمكن تنفيذ ضبط للمعاملات Hyperparameter Tuning باستخدام Grid Search او Random Search لايجاد افضل تركيبة ممكنة للنموذج

كما يمكن تحسين هندسة الخصائص من خلال استخراج خصائص زمنية اضافية مثل الربع السنوي او ايام العطل الرسمية او الفترات الترويجية مما قد يساعد النموذج على فهم الانماط الموسمية بشكل ادق كذلك يمكن توسيع مجموعة البيانات لتشمل عدة متاجر او فترات زمنية اطول مما يزيد من قدرة النموذج على التعميم ويقلل من التحيز

واخيرا يمكن بناء لوحة تحكم تفاعلية Dashboard لعرض التنبؤات بشكل مرئي مما يسهل على الادارة استخدام النتائج في دعم اتخاذ القرار

**الخاتمة**

في هذا المشروع تم تخطيط وتطوير حل ذكاء اصطناعي للتنبؤ بقيمة المبيعات اعتمادا على بيانات تاريخية حقيقية وقد تم تطبيق منهجية منظمة بدأت بتحديد الاهداف ومعايير النجاح ثم جمع البيانات وفحص جودتها ومعالجتها وتقسيمها قبل الانتقال الى مرحلة بناء النماذج

اظهرت النتائج ان اختيار النموذج يؤثر بشكل مباشر على جودة الحل حيث كان نموذج الانحدار الخطي محدود القدرة على تمثيل طبيعة البيانات بينما حقق نموذج Random Forest تحسنا واضحا في الدقة والاستقرار مما يؤكد اهمية اختيار نموذج مناسب لطبيعة المشكلة.

بشكل عام يحقق الحل الحالي مستوى جيد من الدقة ويمكن اعتباره اداة فعالة لدعم اتخاذ القرار داخل الادارة مع الاخذ بعين الاعتبار القيود والمخاطر المرتبطة بالبيانات وطبيعة السوق ويعكس هذا المشروع فهما عمليا لمراحل بناء حلول الذكاء الاصطناعي من التخطيط الى التطوير والتقييم النقدي.