الاسم: ديما محمد منير القلفه

**تقرير تحليل بيانات السيارات**

تم استخدام مجموعة بيانات السيارات المتاحة على موقع Kaggle

[raw.githubusercontent.com/alexeygrigorev/mlbookcamp-code/master/chapter-02-car-price/data.csv](https://raw.githubusercontent.com/alexeygrigorev/mlbookcamp-code/master/chapter-02-car-price/data.csv)

**المشاكل التي واجهتها**

1. **أسماء الأعمدة غير متناسقة:** كانت أسماء الأعمدة تحتوي على مسافات وحروف كبيرة، مما يجعل التعامل معها صعبًا في الكود.
2. **بيانات مفقودة:** بعض الأعمدة تحتوي على قيم مفقودة (NaN)، خاصة في أعمدة مثل "عدد الأبواب" و"فئة السوق".
3. **تنسيق البيانات**: بعض البيانات النصية كانت غير متسقة، مثل أنواع الوقود وأنواع ناقل الحركة، حيث كانت تحتوي على مسافات ورموز غير ضرورية.
4. **توزيع الأسعار غير متوازن**: كانت هناك بعض القيم المتطرفة في عمود السعر (MSRP) التي قد تؤثر على التحليل.

**كيفية تنظيف البيانات**

1. **توحيد أسماء الأعمدة:**

* تم تحويل جميع أسماء الأعمدة إلى أحرف صغيرة.
* تم استبدال المسافات بشرطات سفلية (\_) لتحسين قابلية القراءة والتعامل في الكود.
* مثال: Engine Fuel Type أصبح engine\_fuel\_type.

1. **معالجة القيم المفقودة:**

* تم تحديد الأعمدة التي تحتوي على قيم مفقودة.
* حذف الصفوف ذات القيم المفقودة و تخزين الداتا التي تم حذف القيم المفقودة في داتا جانبية في حال أردنا استبدال القيم المفقودة بقيم أخرى مستقبلاً.

1. **تنظيف البيانات النصية:**

* تم تحويل جميع البيانات النصية إلى أحرف صغيرة.
* تم استبدال المسافات والرموز غير الضرورية ب underscore لتحسين التناسق.

1. **تحليل توزيع الأسعار:**

* تم إنشاء رسم بياني لتوزيع الأسعار (MSRP) لتحديد القيم المتطرفة وفهم توزيع البيانات.

**ملاحظات إضافية**

* البيانات تحتوي على معلومات غنية ومتنوعة عن السيارات، مثل سنة الصنع، قوة المحرك، عدد الأسطوانات، واستهلاك الوقود.
* يمكن استخدام هذه البيانات لبناء نماذج تنبؤية لتقدير أسعار السيارات بناءً على مواصفاتها.
* هناك بعض الأعمدة مثل "فئة السوق" تحتوي على قوائم متعددة القيم مفصولة بفواصل، والتي يمكن تقسيمها لتحليل أكثر تفصيلاً.
* الخطوات التالية يمكن أن تشمل تحليلات أكثر تعمقًا، مثل دراسة العلاقة بين قوة المحرك واستهلاك الوقود، أو تأثير سنة الصنع على السعر.