**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені ТАРАСА ШЕВЧЕНКА**



**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра прикладних інформаційних систем**

**Звіт до лабораторної роботи №3**

**з курсу «Data Science та Big Data»**

*Студента 4 курсу*

*Групи ПП-41*

*спеціальності 122*

*«Комп'ютерні науки»*

*ОП«Прикладне програмування»*

*Гуцала Миколи Олександровича*

Київ 2023

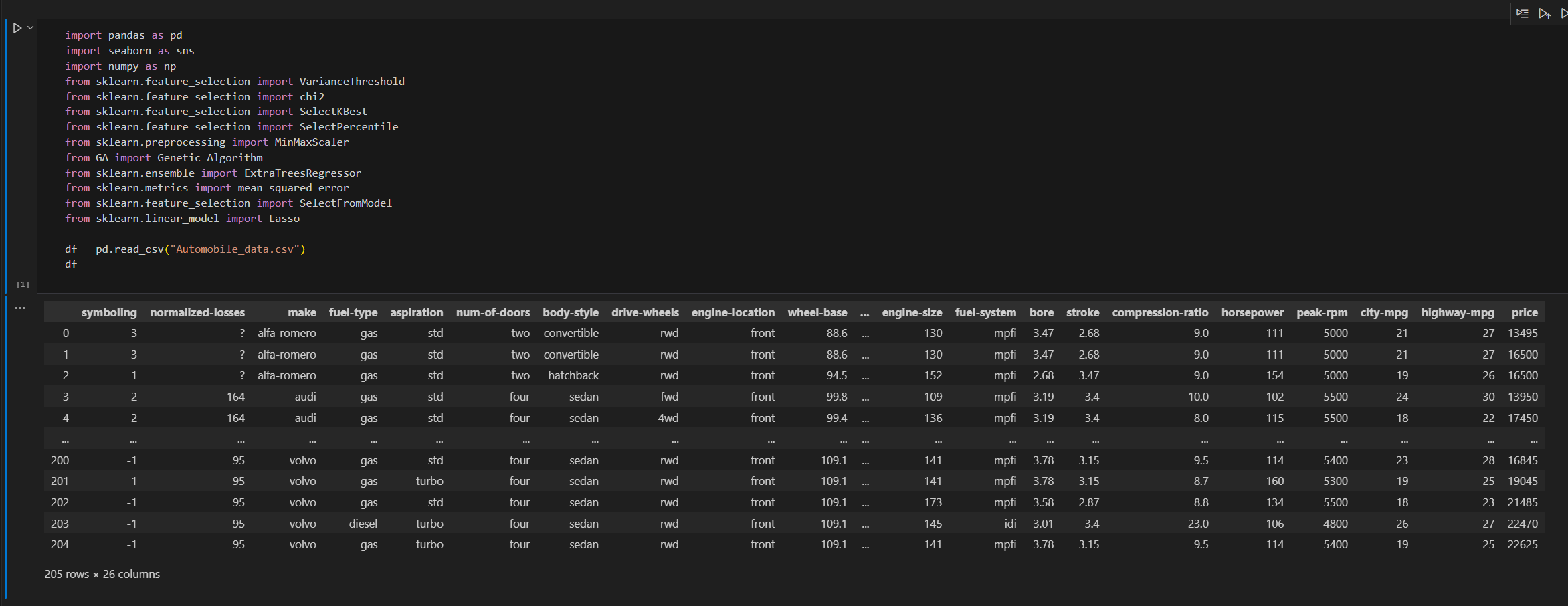
**Тема роботи:** Методи аналізу та вибору значущих ознак (Features’ Selection Procedures)

**Завдання для лабораторної роботи**

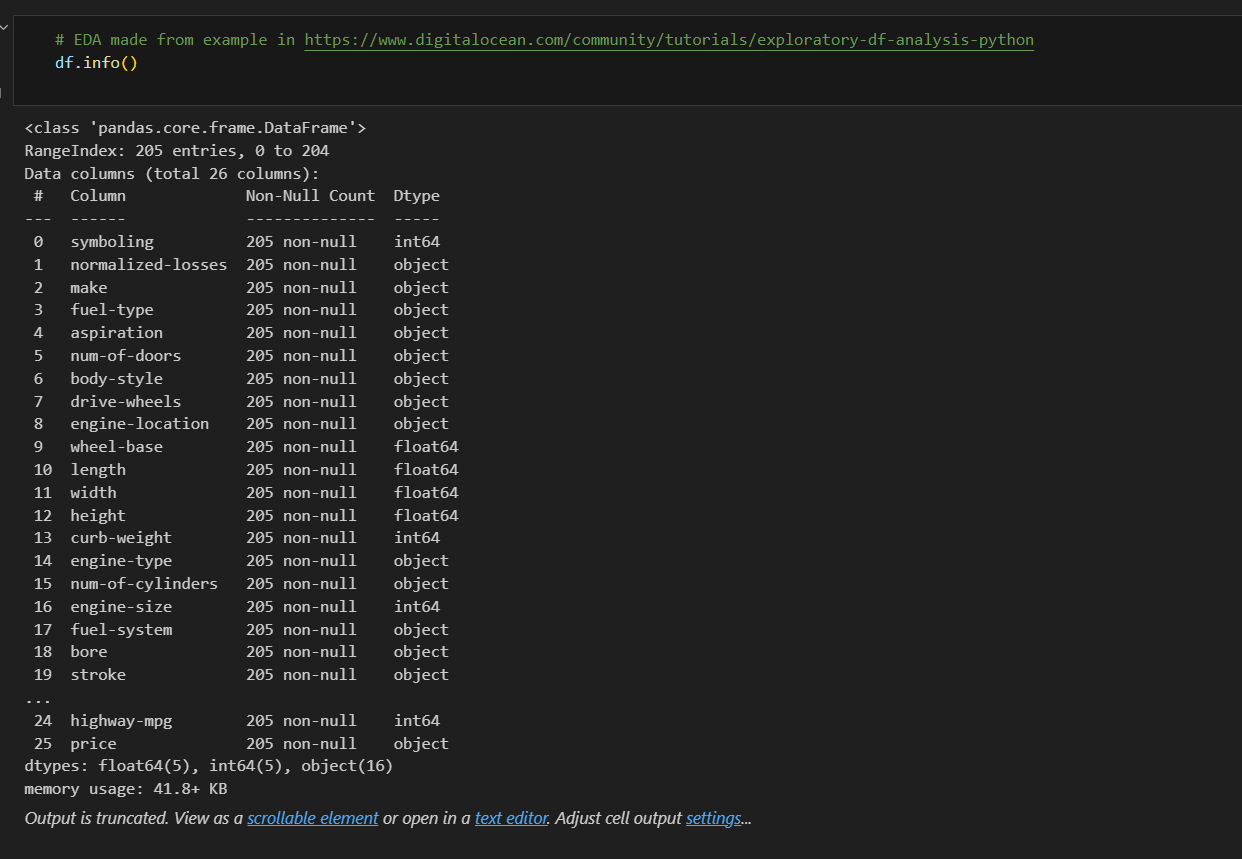
1. Ознайомитись з наданим прикладом використання різних методів відбору значущих ознак (папка Example).
2. Завантажити файли з даними у папку проекту з посилання:
3. https://drive.google.com/file/d/1su22-W8JrRZzm0mea5v8x46YmLh083qp/view?usp=sharing
4. Очистити дані та обробити відсутні дані.
5. Зробити EDA по ознаках.
6. Проаналізуйте надані дані, використовуючи методи з прикладу та документації, та зберіть результати аналізу у результуючий ранжируваний датафрейм, в якому лівим індексом будуть ознаки, а колонки – результати однофакторного аналізу ознак. Подумайте над системою ранжування такою, яка б враховувала наявність багатьох факторів ранжування (припустимо, що всі вони мають однакову вагу на прийняття вами рішення).
7. Проаналізуйте ознаки на взаємозалежність, та побудуйте відповідні heatmap засобами seaborn по кожному з використаних методів дослідження.
8. Зберіть висновки у звіт (графіки, висновки текстом у окремому файлі), який потребує належного оформлення, структури тощо.
9. Завантажити звіт та файл ipynb з виконаними завданнями на git в окрему папку з відповідною назвою лабораторної роботи

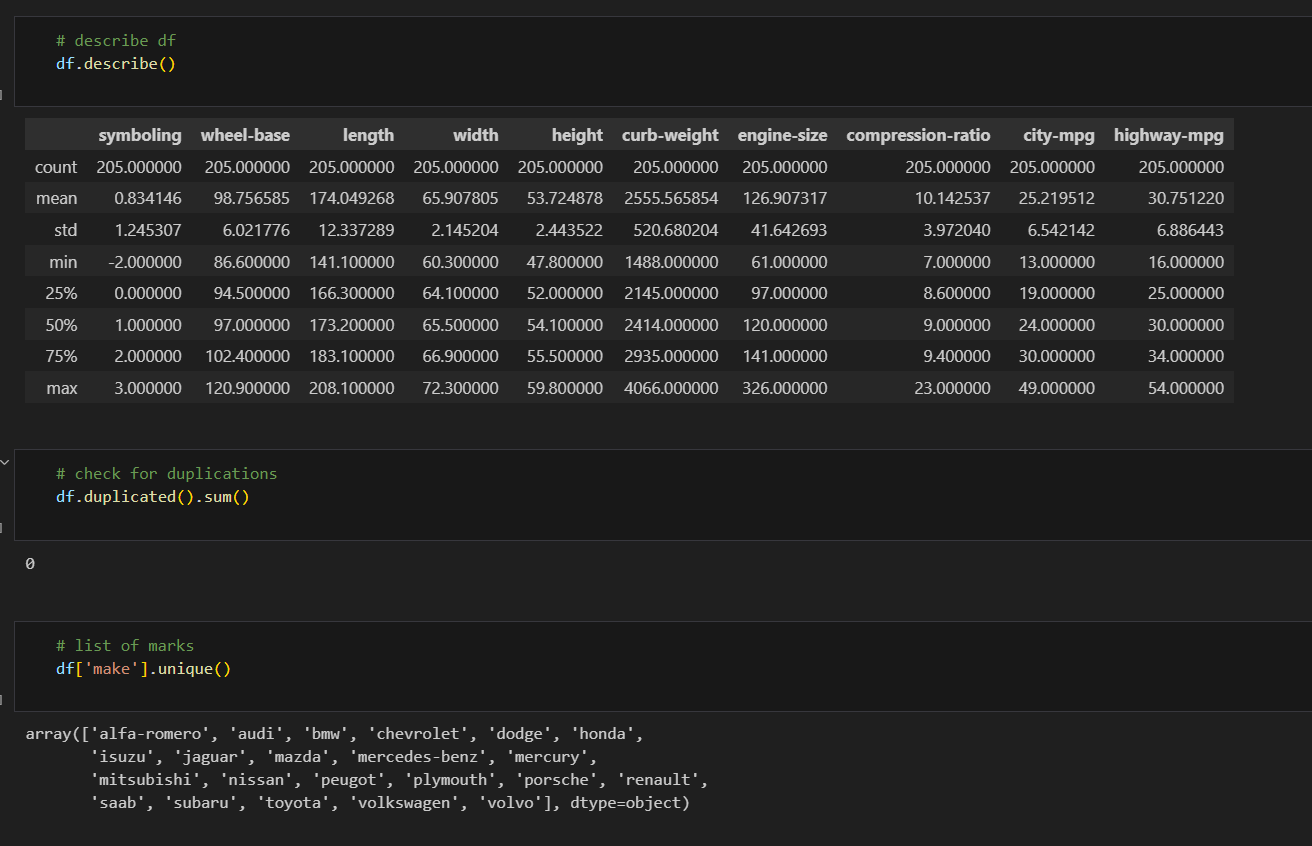
**Хід роботи:**

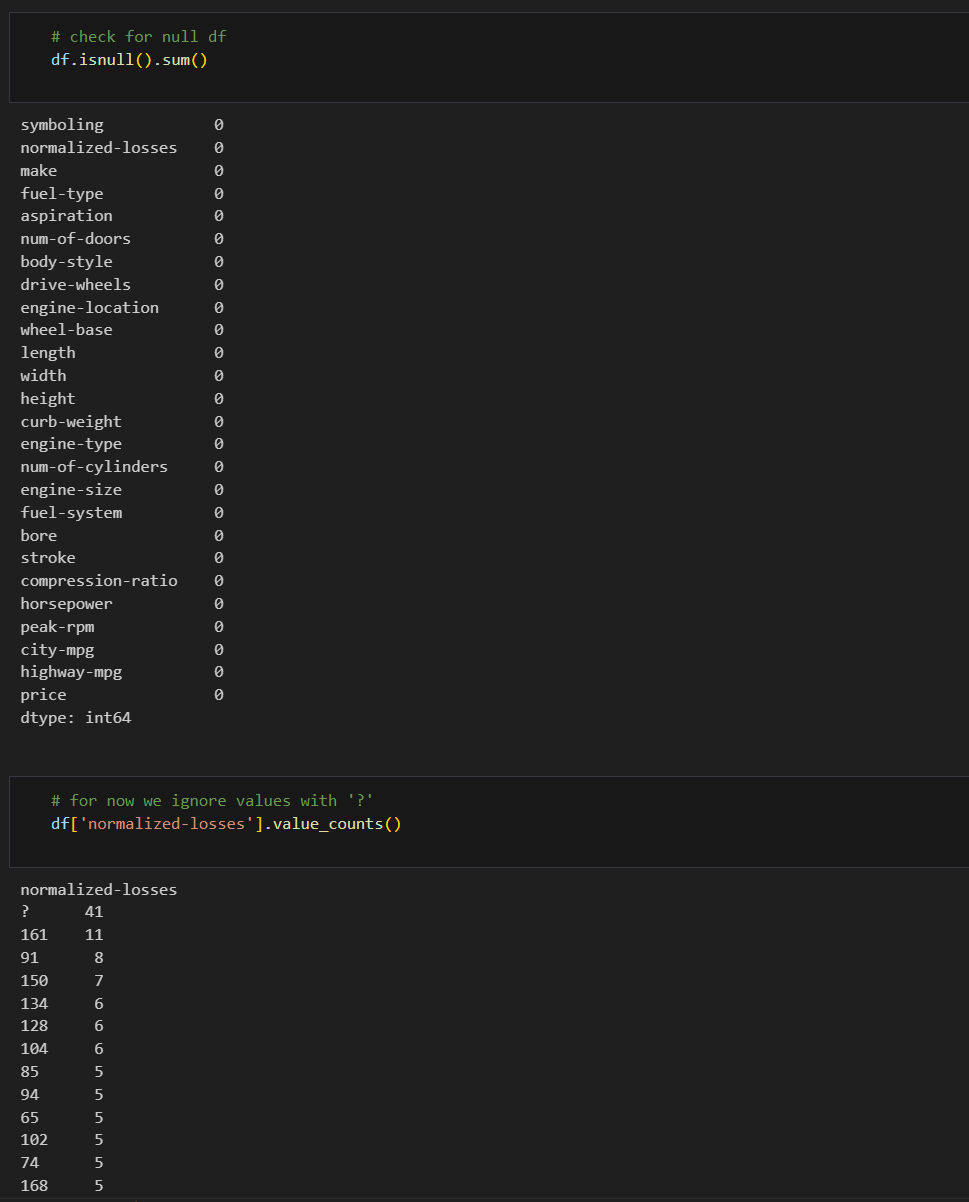
Інсталювання інструментів та завантаження даних



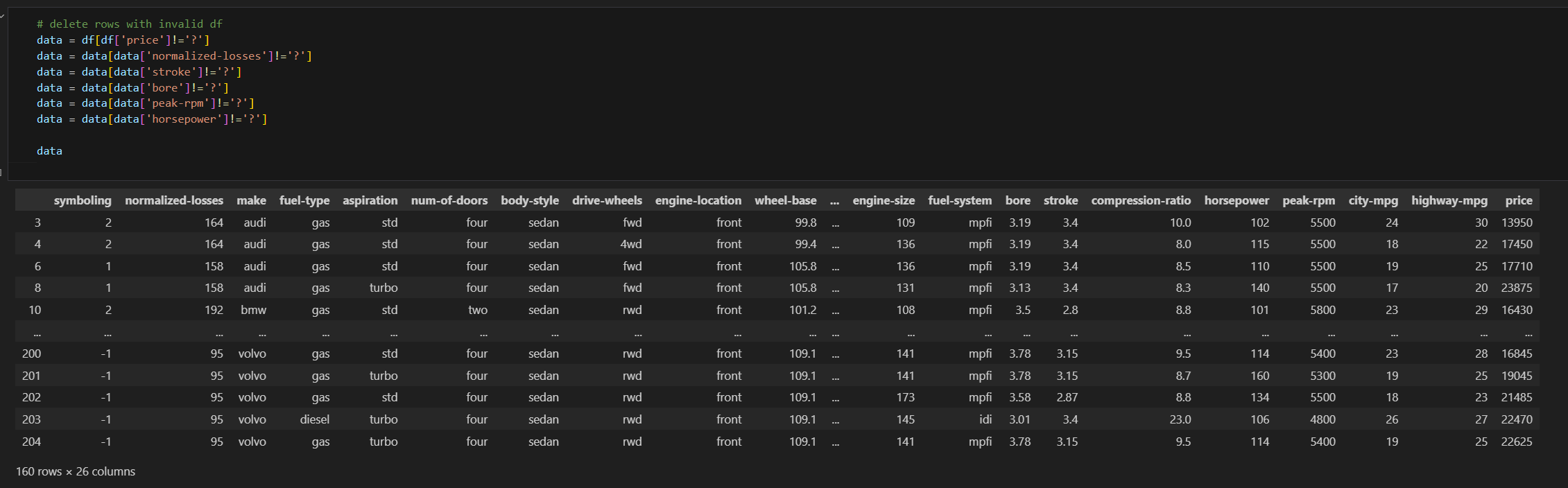
Підготовка та реалізація EDA



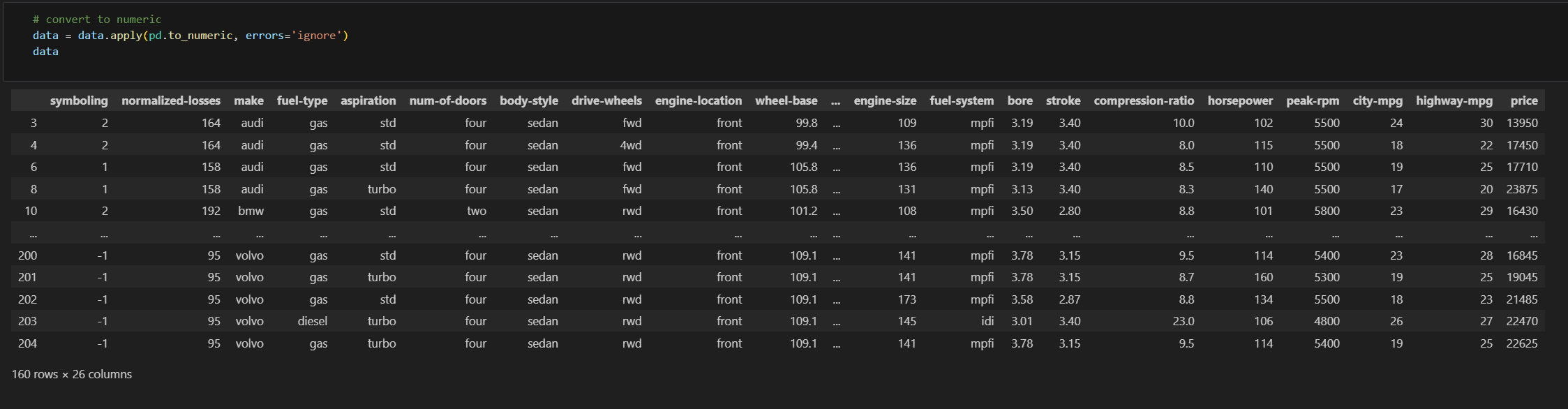




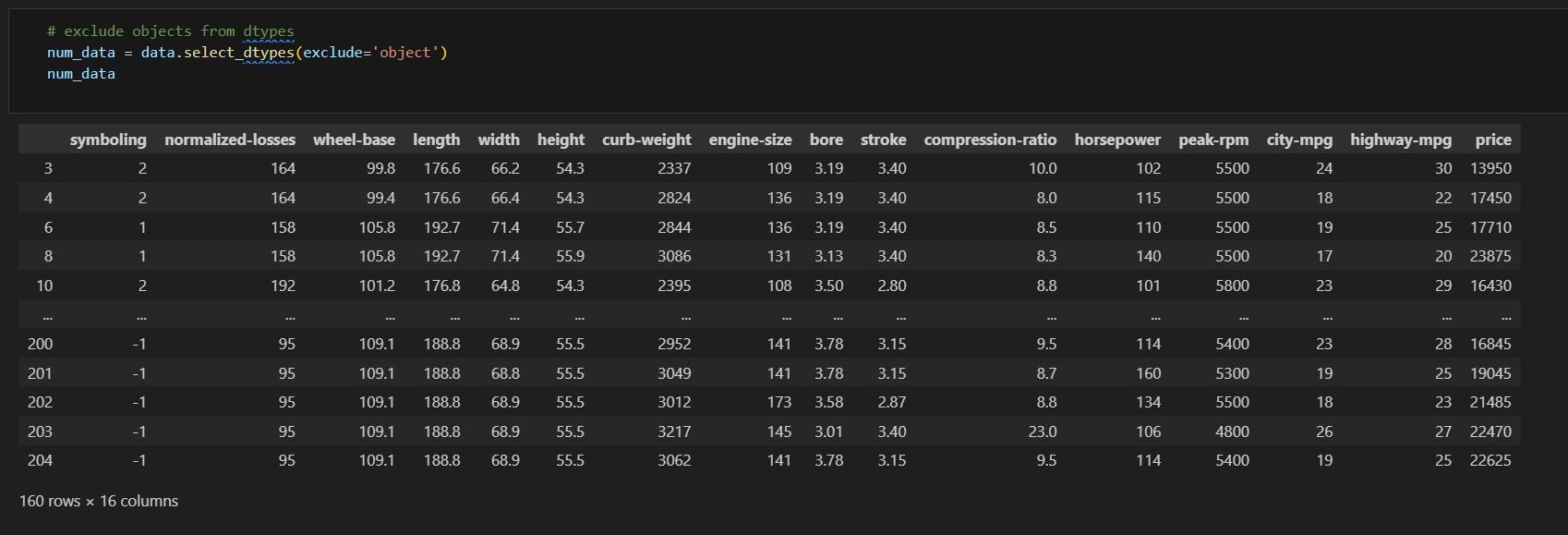
Далі було проведено очистку датафрейму

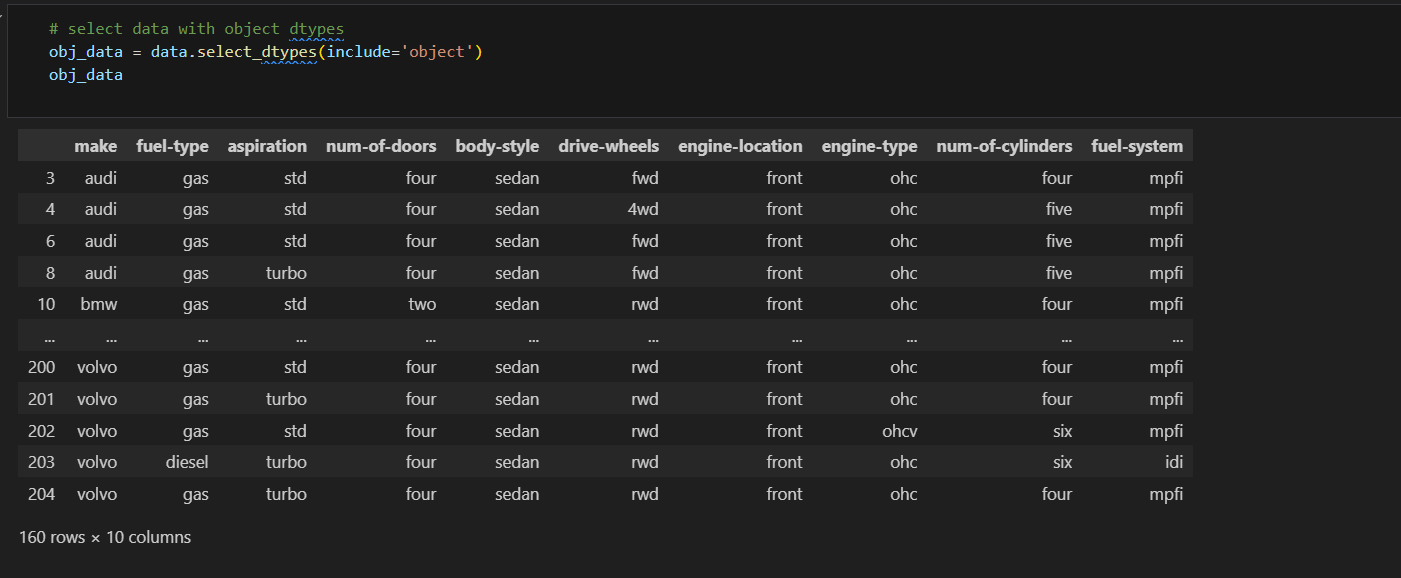


Конвертація стовпців у числовий тип даних

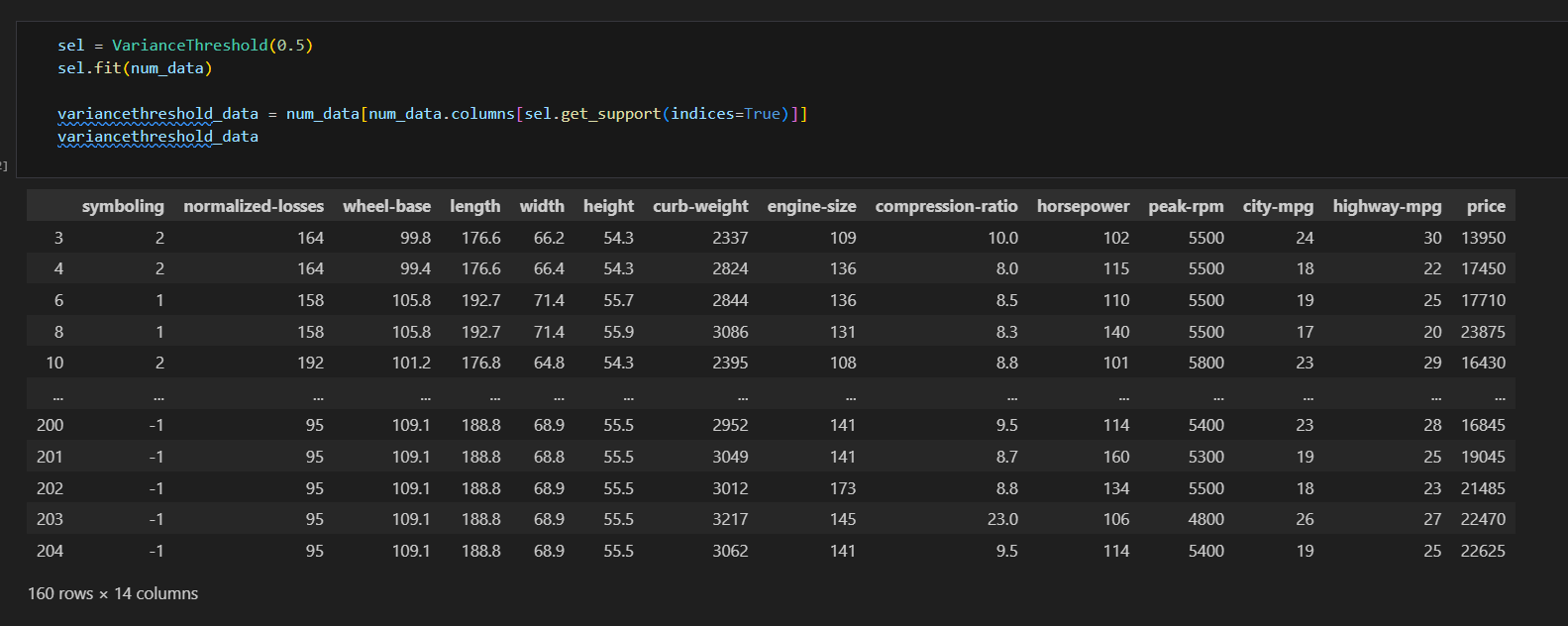
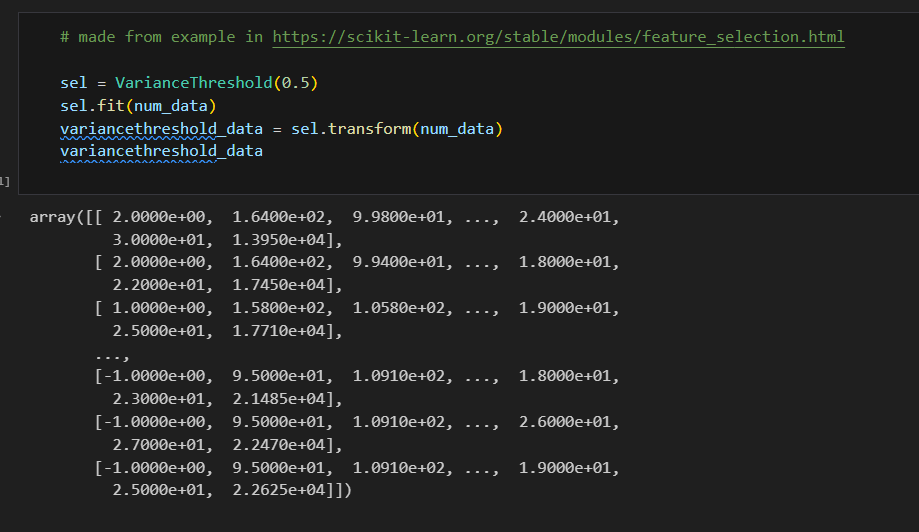
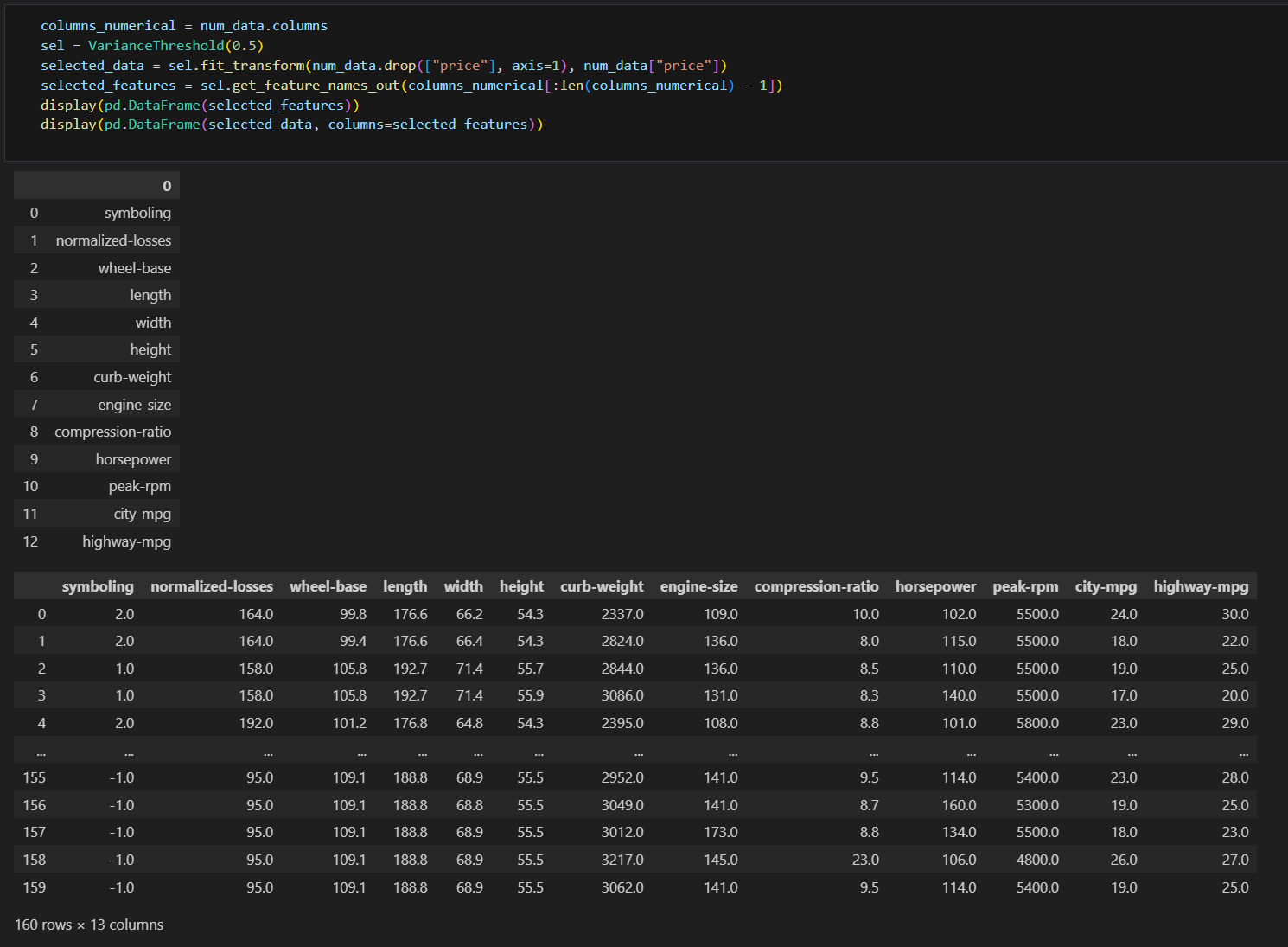


Виділення даних - категоріальні та числові

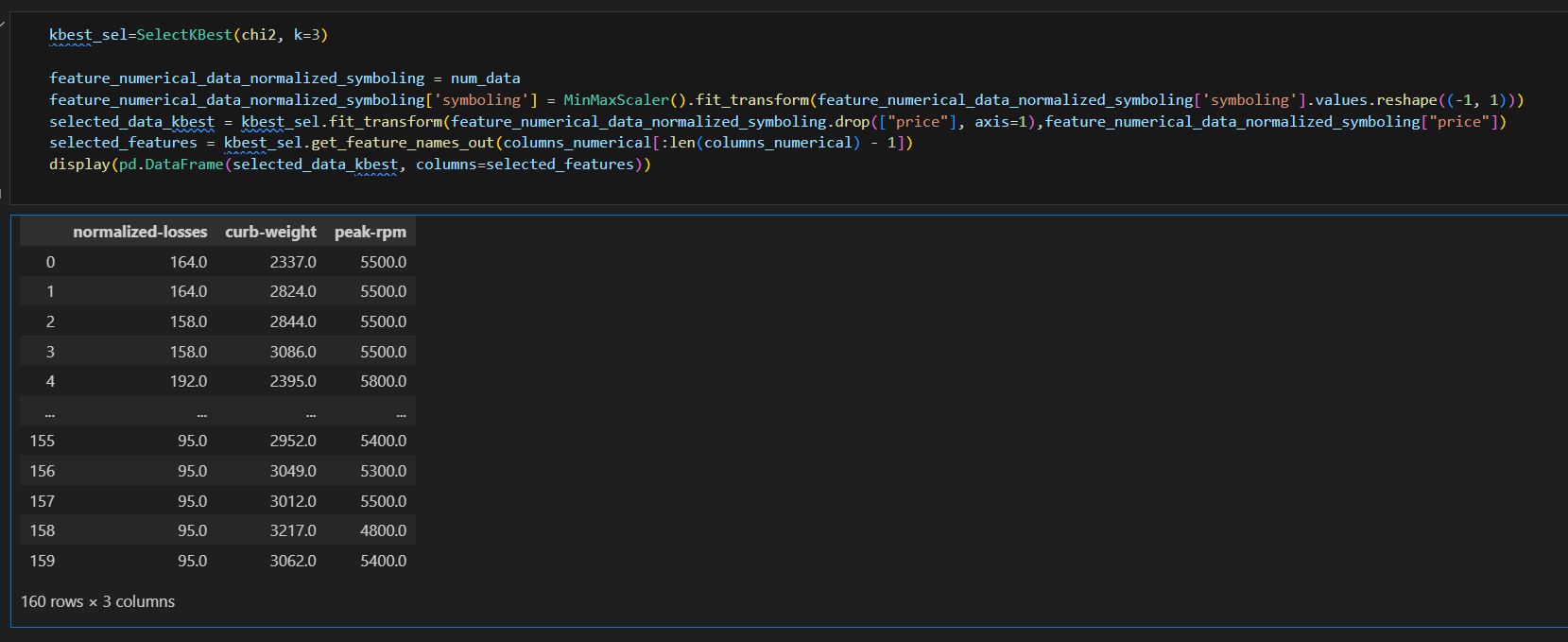




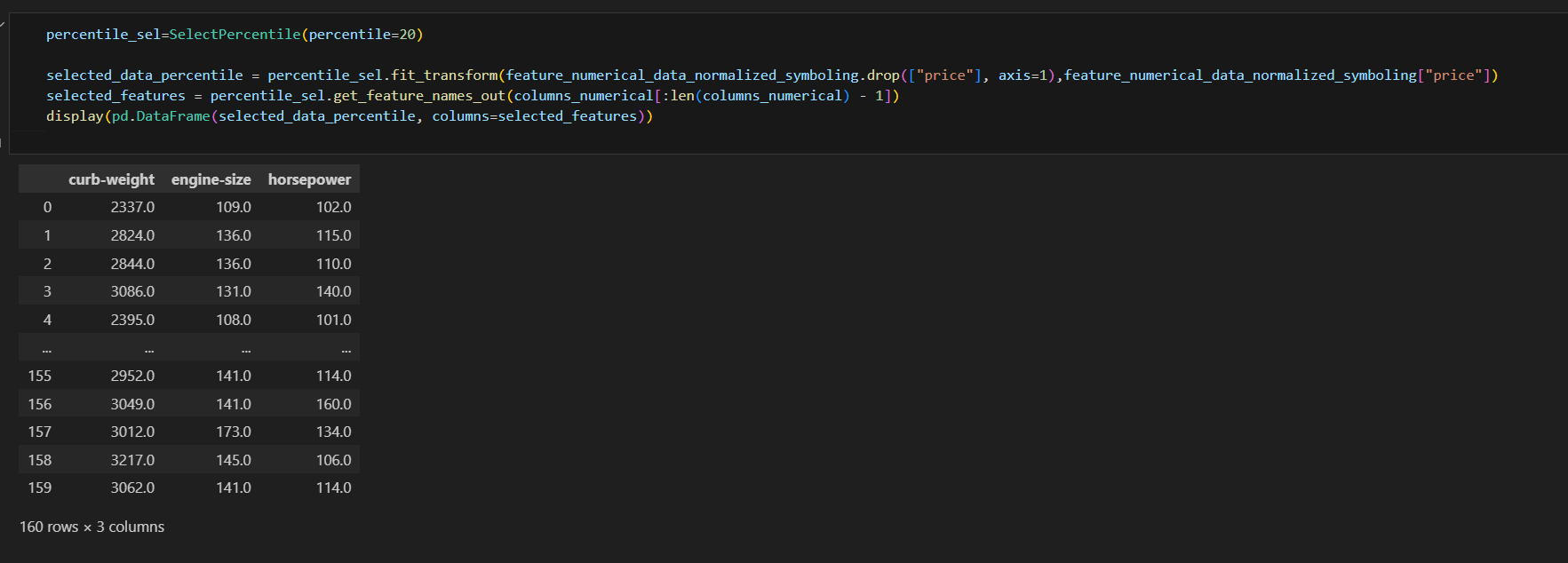
Використання методів для оцінки значущості факторів. Було відкинуто з дисперсією менше 0.5.

Вибір k найвпливовіших факторів, тут використано chi2 для обирання факторів. Також необхідно було прибрати всі від’ємні числа. Найважливіші фактори: normalized-losses, curb-weight, peak-rpm.

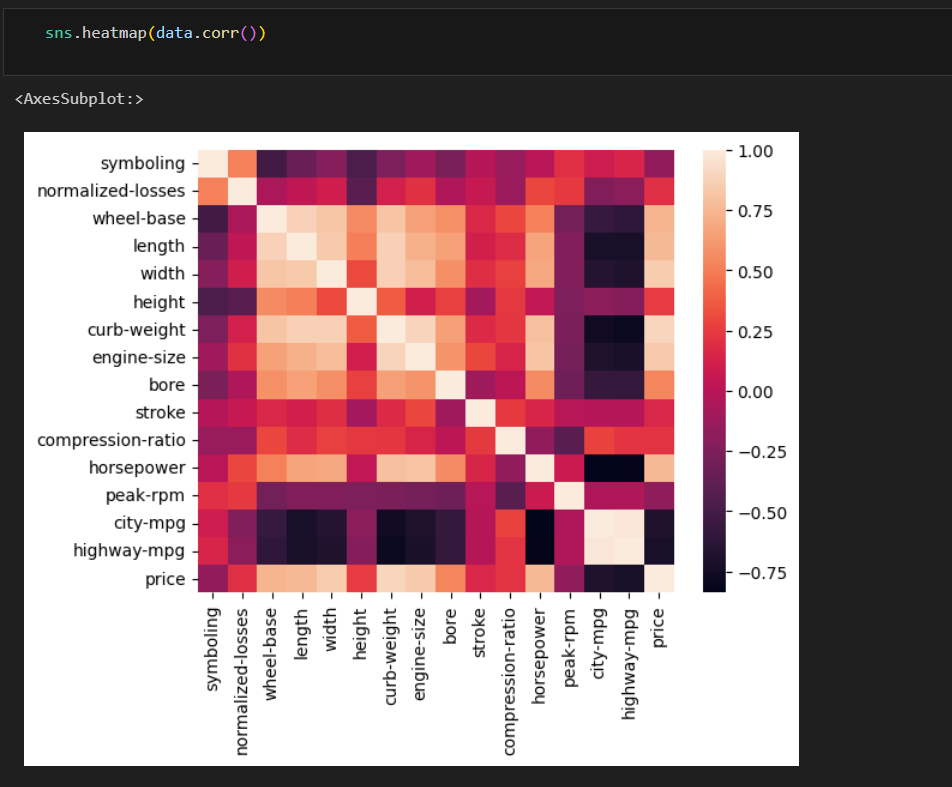


Наступний метод повертає вказаний відсоток найважливіших факторів. Цей метод вважає найважливішими такі стовпчики - curb-weight, engine-size, horsepower



Отже, за результатами цих трьох методів можна зробити висновки, що найбільш на ціну впливають такі показники - normalized-losses, curb-weight, peak-rpm, engine-size, horsepower. Важливо зазначити, що найвпливовішим є показникcurb-weight.

Також візуалізовано корреляцію значень даних, для візуального пошуку факторів, що найбільш впливають на ціну



**Висновки:**

У цій лабораторній я ознайомився з наданим прикладом використання різних методів відбору значущих ознак. Було очищено дані та оброблено відсутні дані. Також зроблено EDA по ознаках, проаналізовано надані дані, використовуючи методи з прикладу та документації, та зібрано результати аналізу у результуючий ранжируваний датафрейм, в якому лівим індексом є ознаки, а колонки – результати однофакторного аналізу ознак.   
 Проаналізовано ознаки на взаємозалежність, та побудовано відповідні heatmap засобами seaborn.