**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені ТАРАСА ШЕВЧЕНКА**



**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра прикладних інформаційних систем**

**Звіт до лабораторної роботи №2**

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

*Студентки 4 курсу*

*Групи ПП-41*

*спеціальності 122 «Комп'ютерні науки»*

*ОП«Прикладне програмування»*

*Поліщук Каріни Юріївни*

Київ 2023

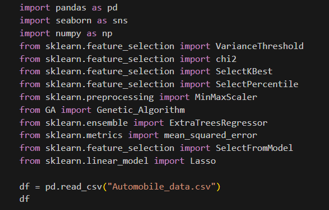
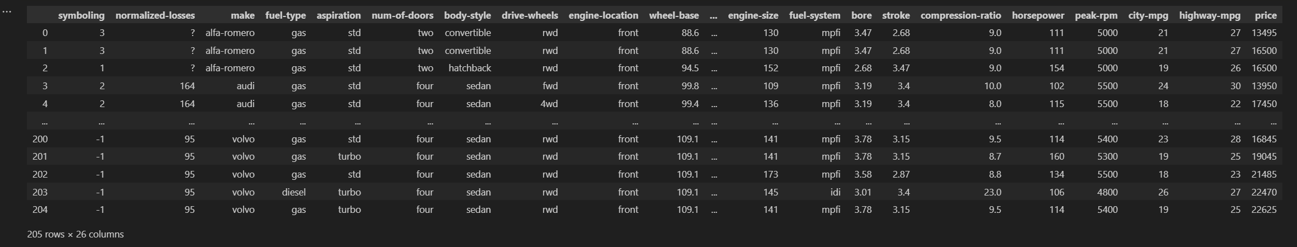
**Тема роботи:** Методи аналізу та вибору значущих ознак (Features’ Selection Procedures)

**Завдання для лабораторної роботи**

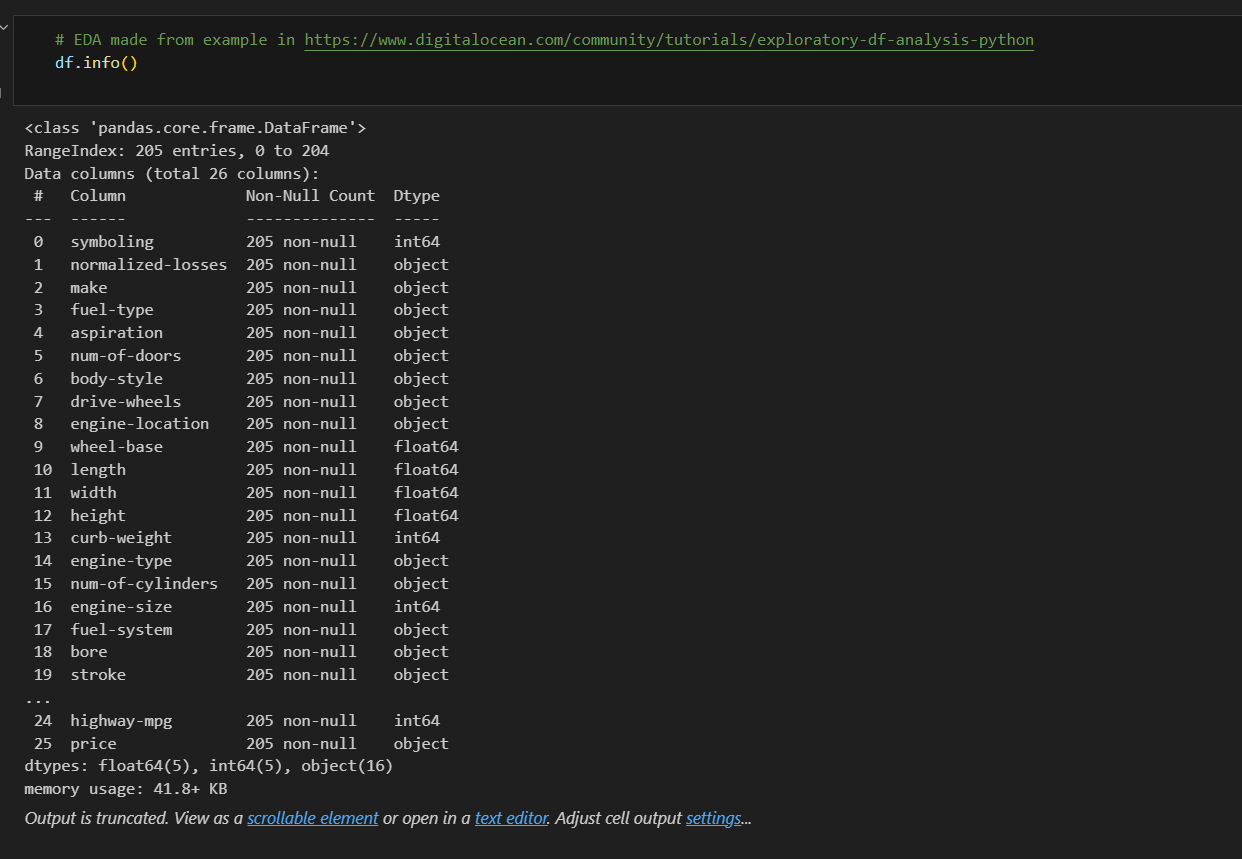
1. Ознайомитись з наданим прикладом використання різних методів відбору значущих ознак (папка Example).
2. Завантажити файли з даними у папку проекту з посилання:
3. https://drive.google.com/file/d/1su22-W8JrRZzm0mea5v8x46YmLh083qp/view?usp=sharing
4. Очистити дані та обробити відсутні дані.
5. Зробити 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

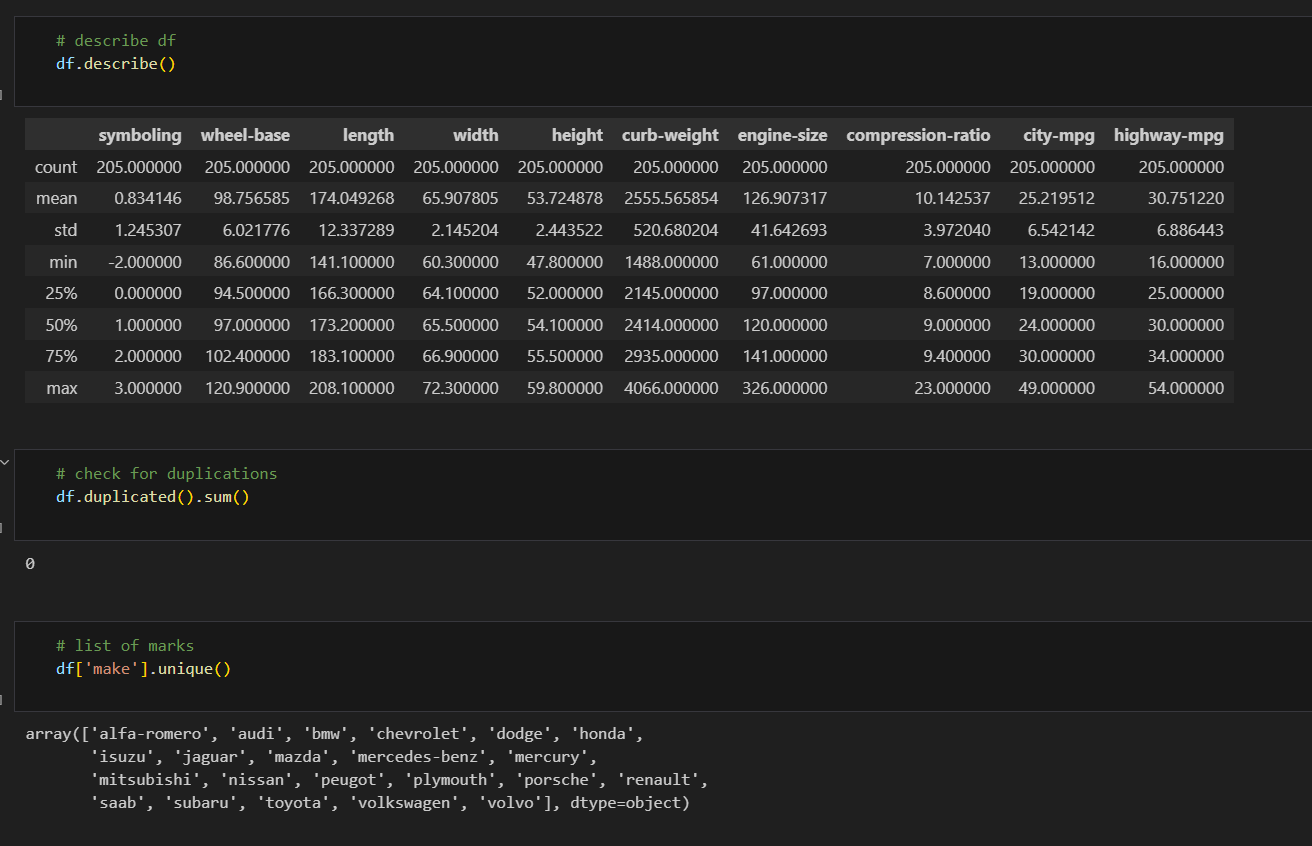
**Хід роботи:**

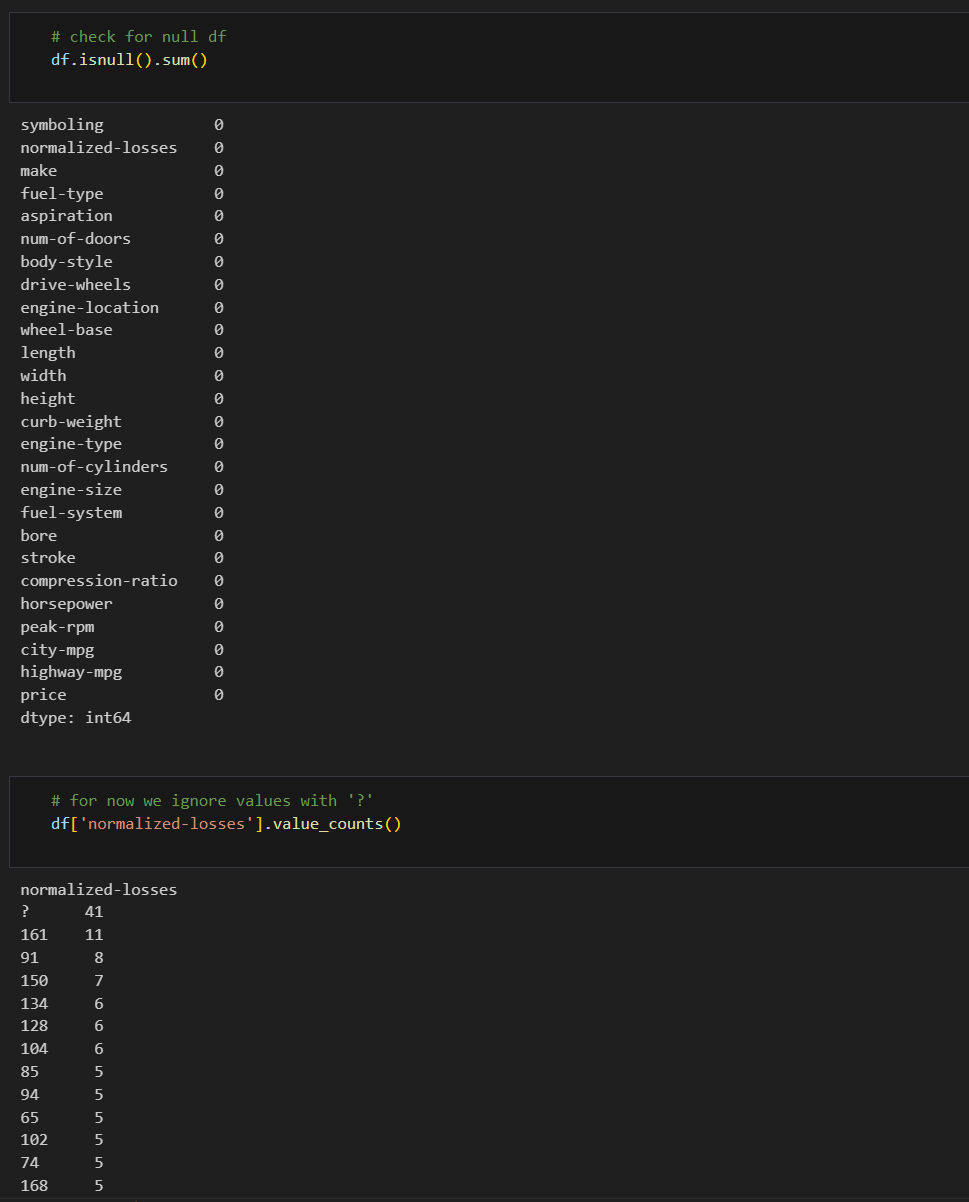
1. Проведемо підготовку та налаштування інструментів. Здійснемо передачу даних.

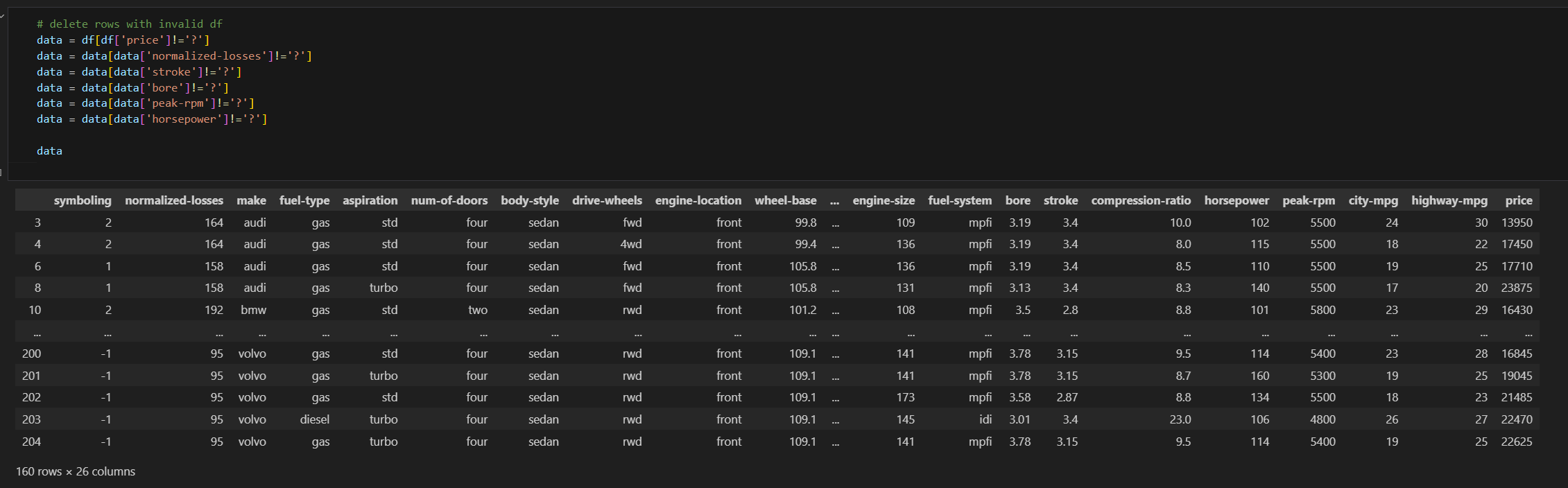
2. Зробимо реалізацію EDA



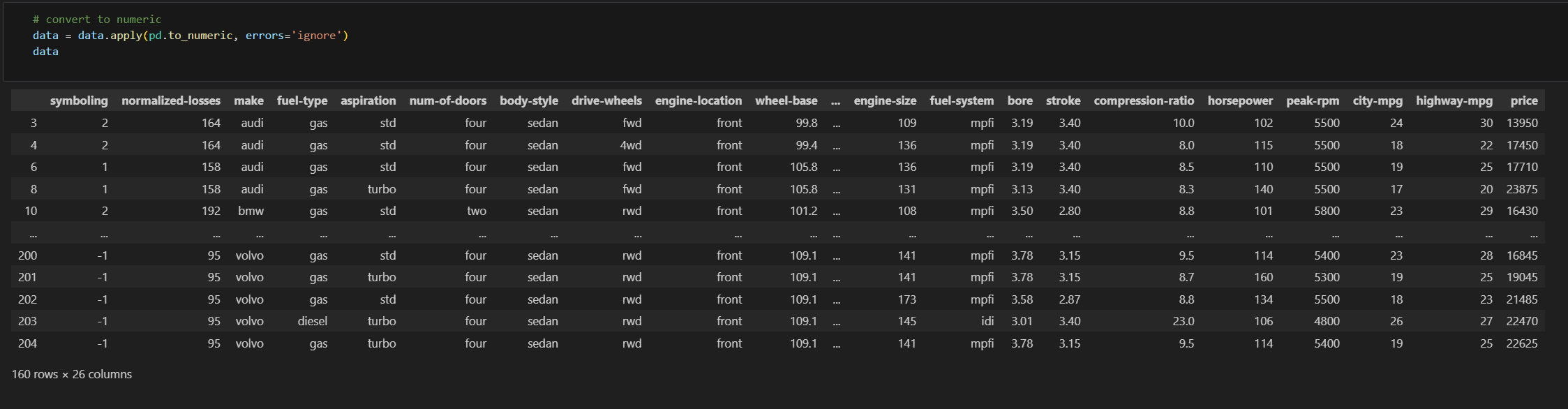




4.Проведемо очистку DataFrame. Очистка датафрейму - це процес видалення непотрібних або некоректних даних з таблиці, щоб забезпечити точність та послідовність даних.

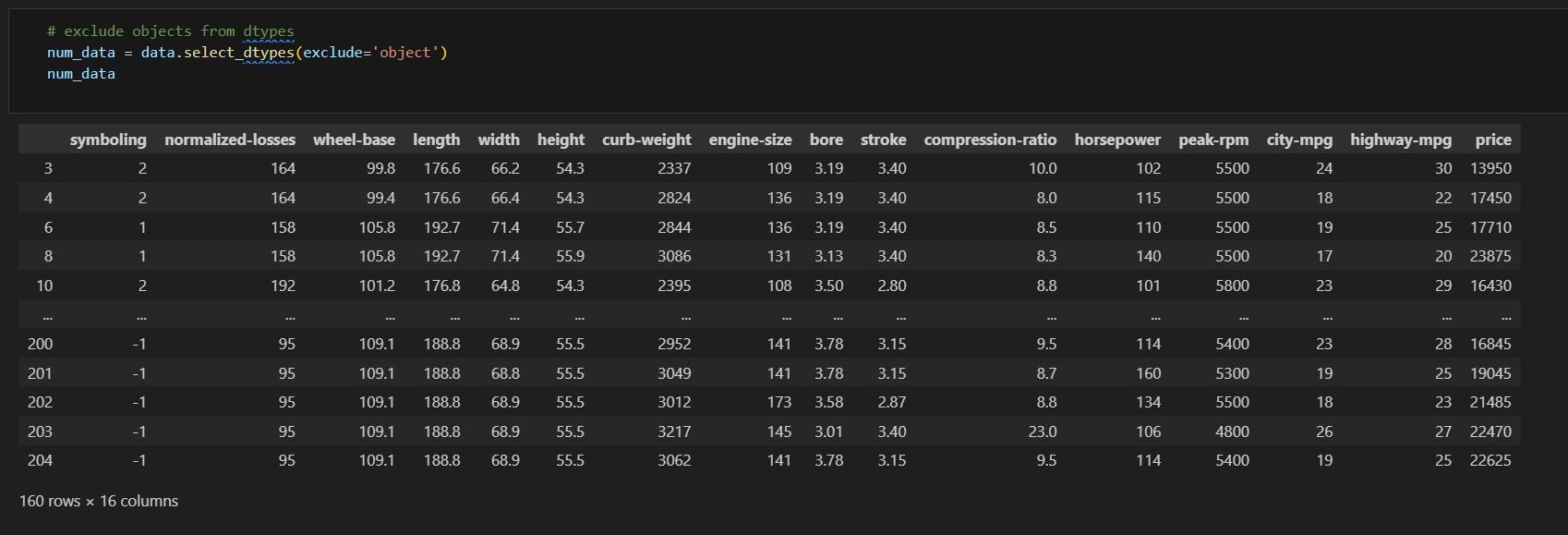


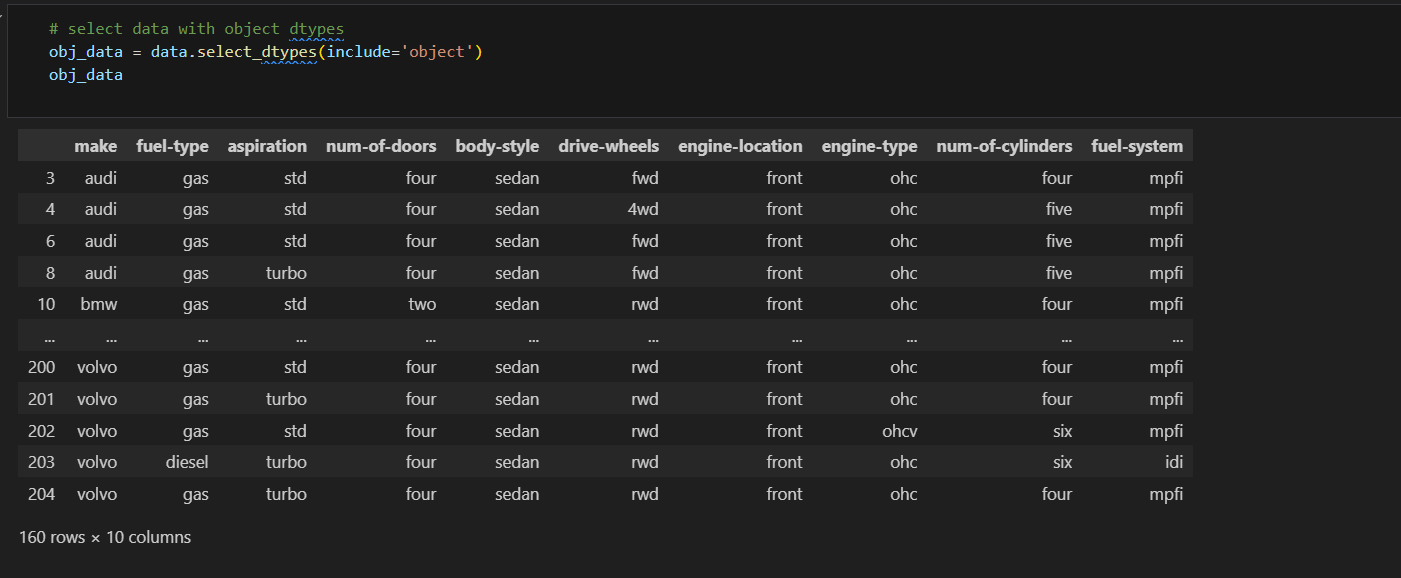
5.Проведемо конвертацію стовпців у числовий тип даних. Використаємо метод `apply()` з бібліотеки pandas. Цей метод застосовує задану функцію до кожного елемента у датафреймі. після виконання цього рядка коду, змінна `data` буде містити оновлений датафрейм, де значення всередині кожної комірки, якщо це можливо, будуть перетворені на числовий тип даних. В той же час, якщо конвертація неможлива, значення залишаться без змін, так як параметр `errors='ignore'` вказує програмі ігнорувати помилки конвертації.



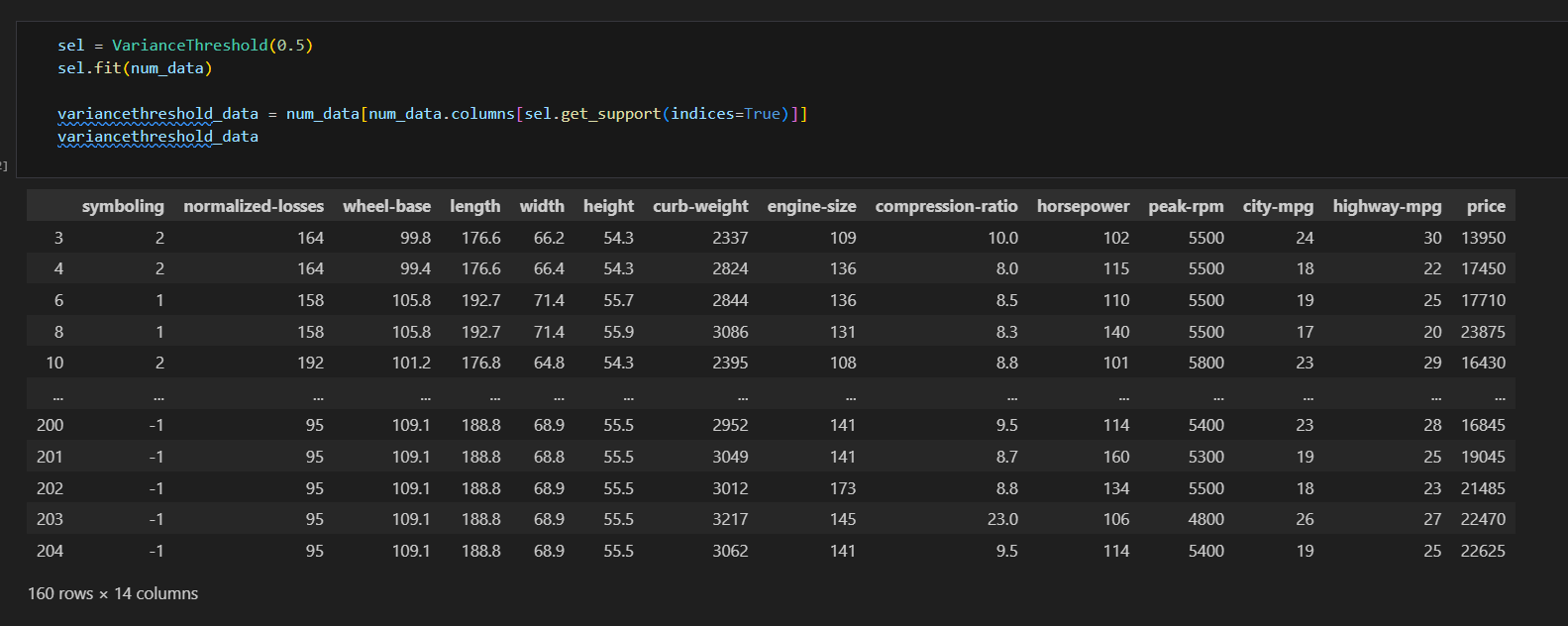
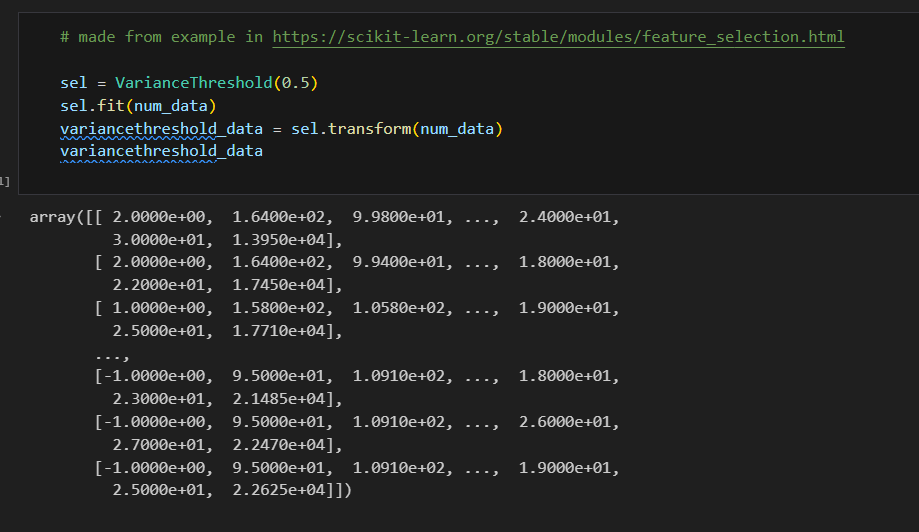
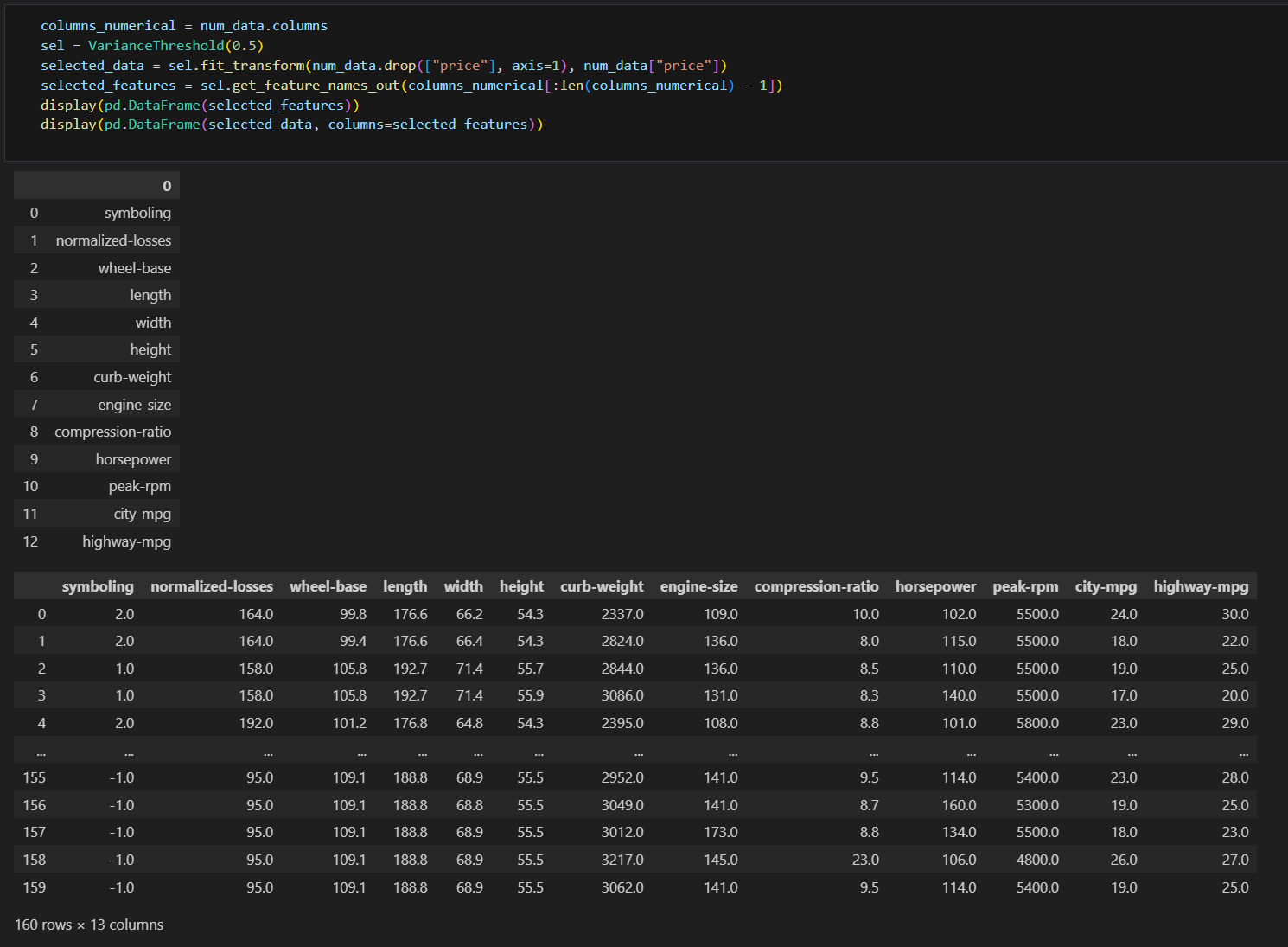
6.Зробимо виділення даних - категоріальні та числові. Метод `select\_dtypes()` дозволяє вибирати колонки датафрейму за типом даних.

В даному випадку, через аргумент `exclude` задається тип даних, який має бути виключений з вибору, тобто всі колонки з типом даних "object" будуть виключені.   
Отже, в результаті виконання цього рядка коду, змінна `num\_data` міститиме новий датафрейм, який містить лише ті колонки оригінального датафрейму `data`, які мають числовий тип даних.

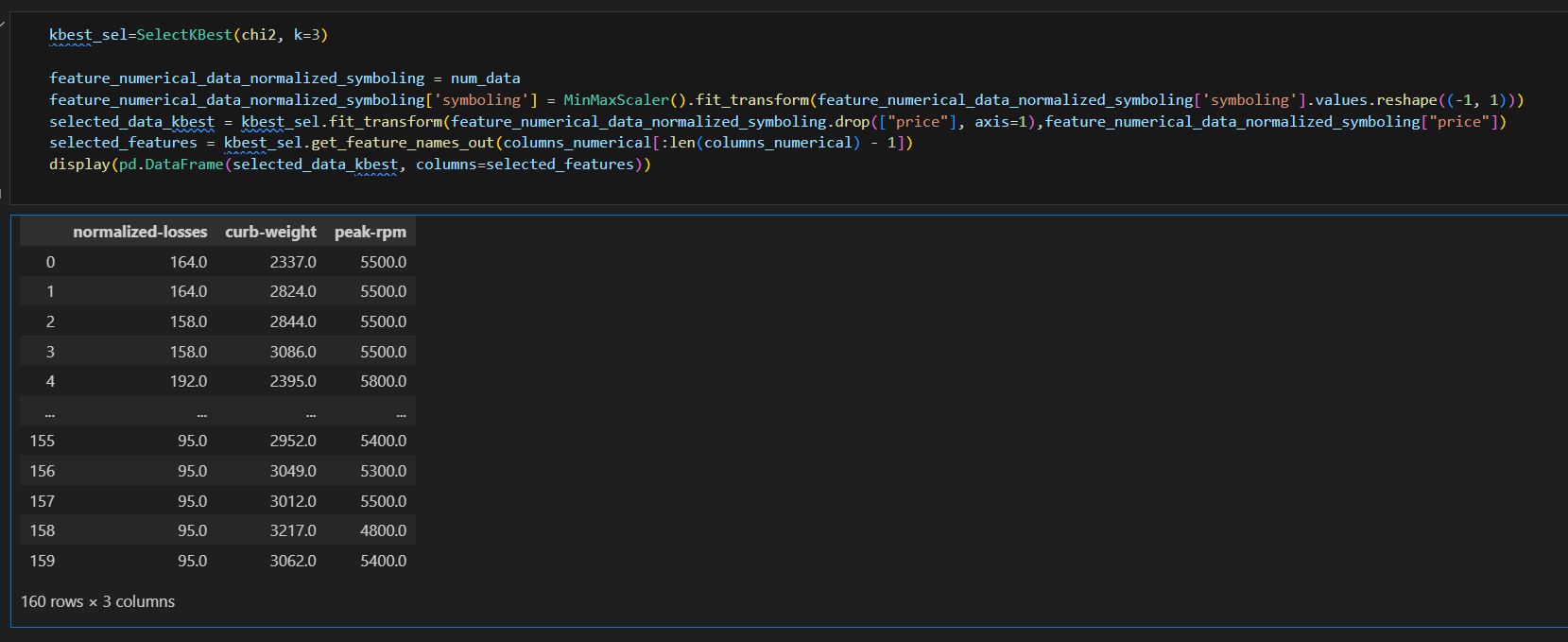


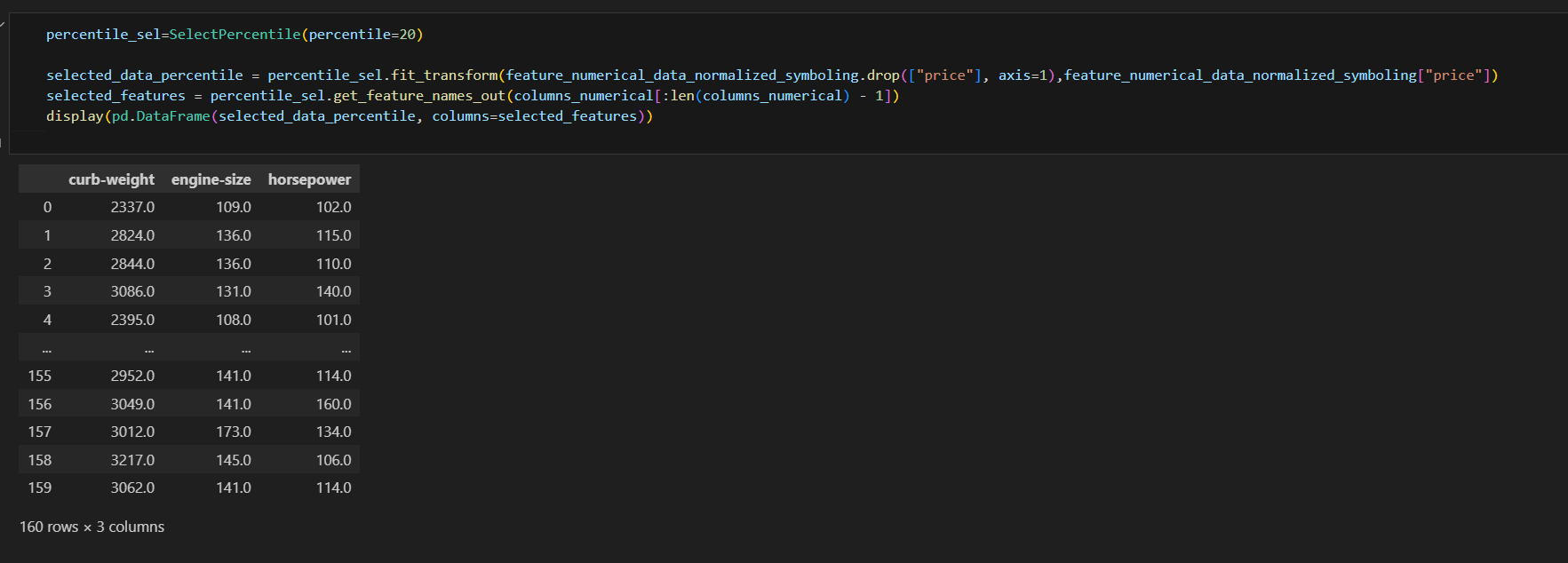


7. Застосуємо методи для оцінки значущості факторів. Було відкинуто з дисперсією менше 0.5.

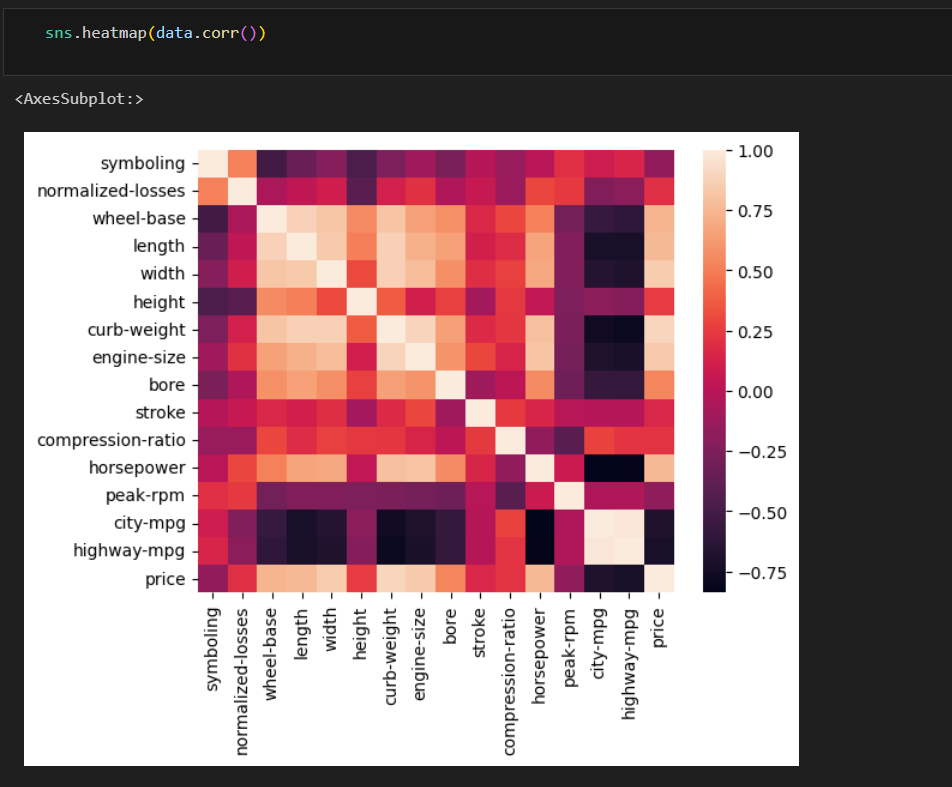
6. Зробимо Вибір k найвпливовіших факторів.Прибиремо всі від’ємні числа. Найважливіші фактори: normalized-losses, curb-weight, peak-rpm.





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

7.Зроблено візуалізацію корреляції значень даних для візуального пошуку факторів, що найбільш впливають на ціну.



**Висновки:**

У цій лабораторній я ознайомилася 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.