**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**імені ТАРАСА ШЕВЧЕНКА**

**Факультет інформаційних технологій**

**Кафедра прикладних інформаційних систем**

****

**Звіт**

до виконання лабораторної роботи № 2

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

на тему:

«Методи аналізу та вибору значущих ознак (Features’ Selection Procedures)»

**Виконано:**

студ. групи ПП-42, підгрупа 1

Дикий Максим

**Київ – 2022**

**Завдання для лабораторної роботи**

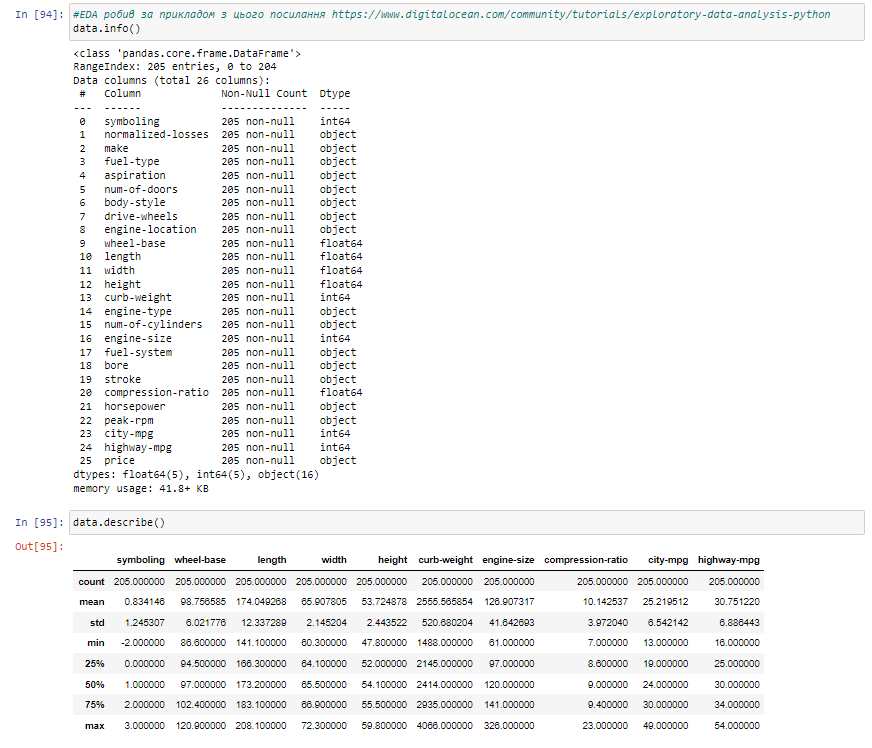
* Ознайомитись з наданим прикладом використання різних методів відбору значущих ознак (папка Example).
* Завантажити файли з даними у папку проекту з посилання:

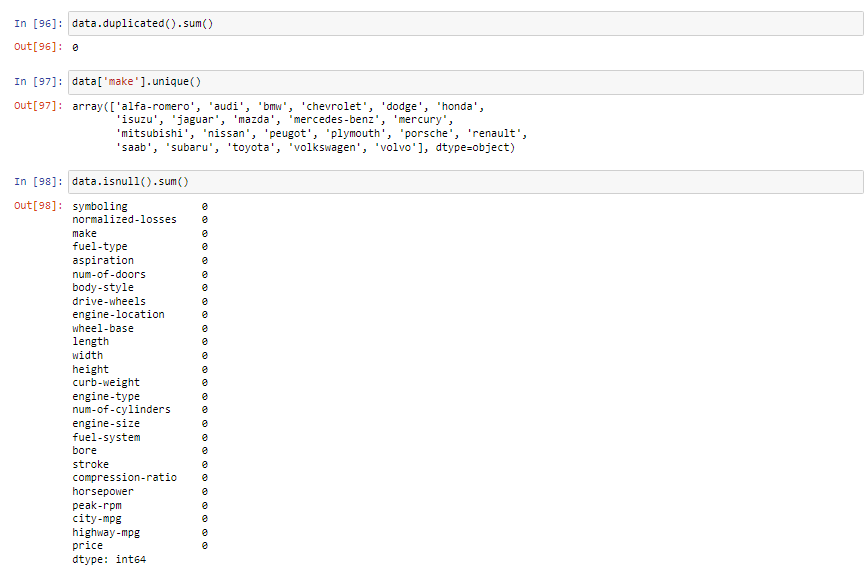
https://drive.google.com/file/d/1su22-W8JrRZzm0mea5v8x46YmLh083qp/view?usp=sharing

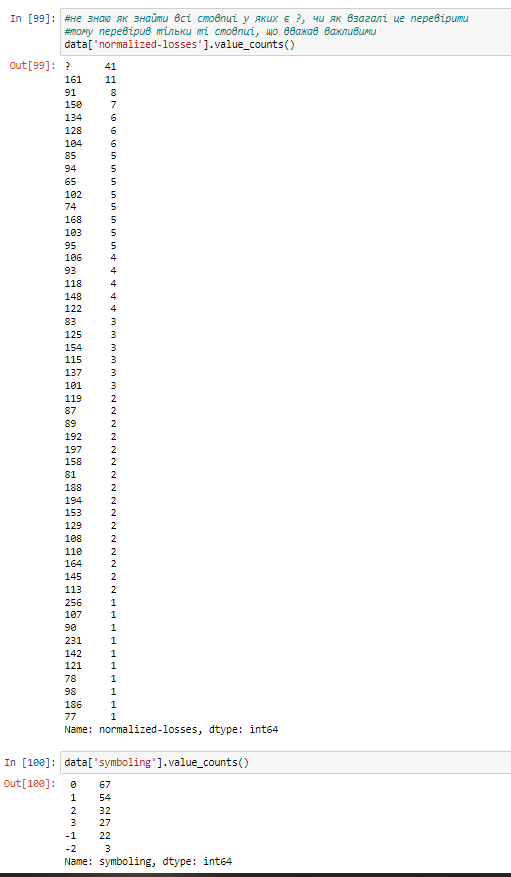
* Очистити дані та обробити відсутні дані.
* Зробити EDA по ознаках.
* Проаналізуйте надані дані, використовуючи методи з прикладу та документації, та зберіть результати аналізу у результуючий ранжируваний датафрейм, в якому лівим індексом будуть ознаки, а колонки – результати однофакторного аналізу ознак. Подумайте над системою ранжування такою, яка б враховувала наявність багатьох факторів ранжування (припустимо, що всі вони мають однакову вагу на прийняття вами рішення).
* Проаналізуйте ознаки на взаємозалежність, та побудуйте відповідні heatmap засобами seaborn по кожному з використаних методів дослідження.
* Зберіть висновки у звіт (графіки, висновки текстом у окремому файлі), який потребує належного оформлення, структури тощо.
* Завантажити звіт та файл ipynb з виконаними завданнями на git в окрему папку з відповідною назвою лабораторної роботи

**Хід роботи:**

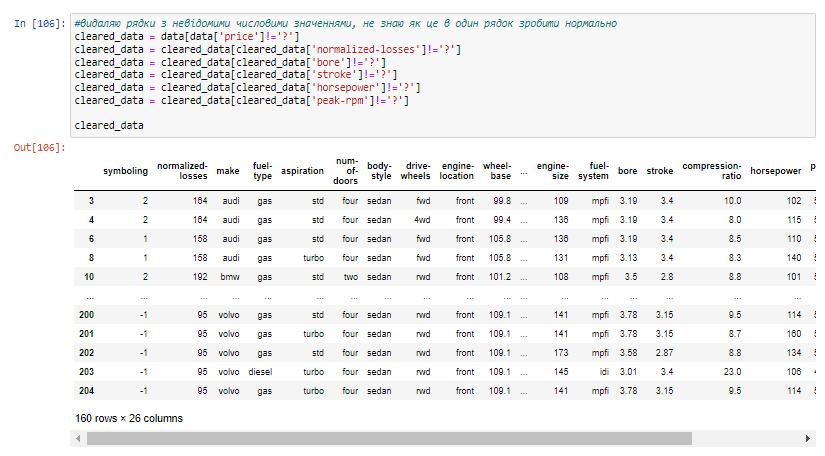
Провів EDA



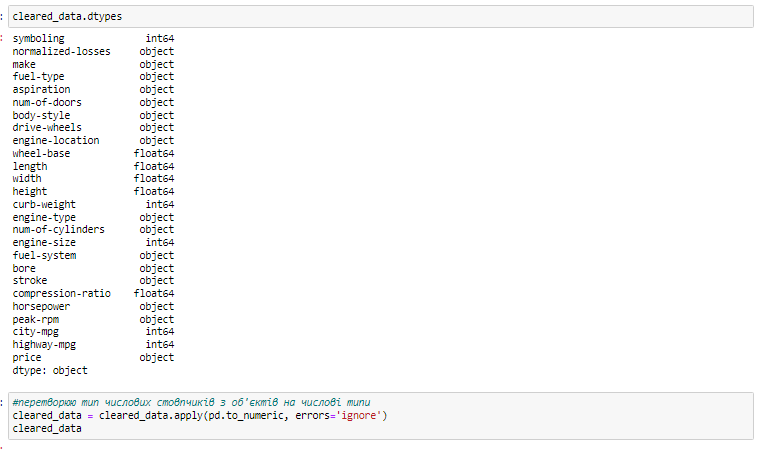




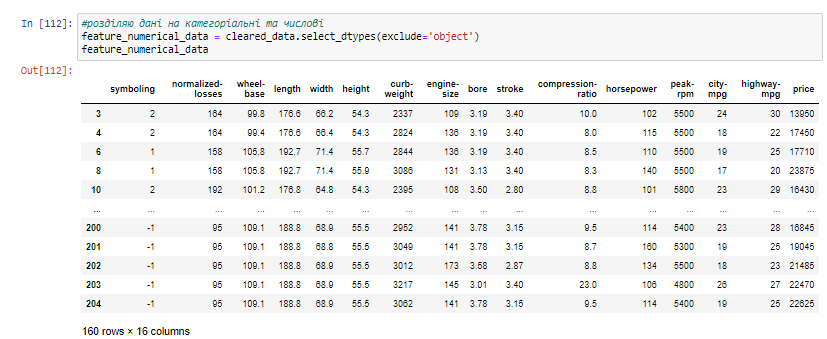
Видалив порожні дані

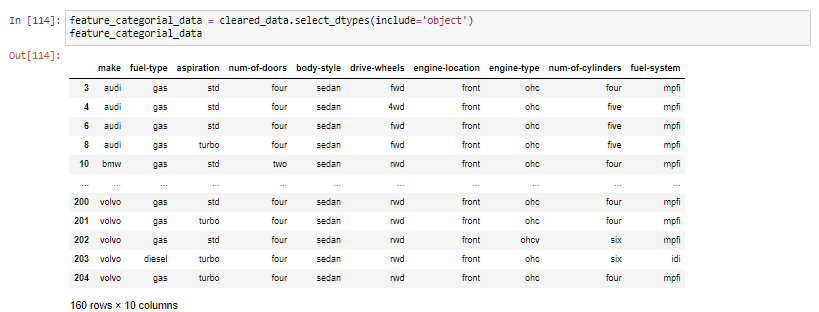


Привів числові стовпці до числового типу даних

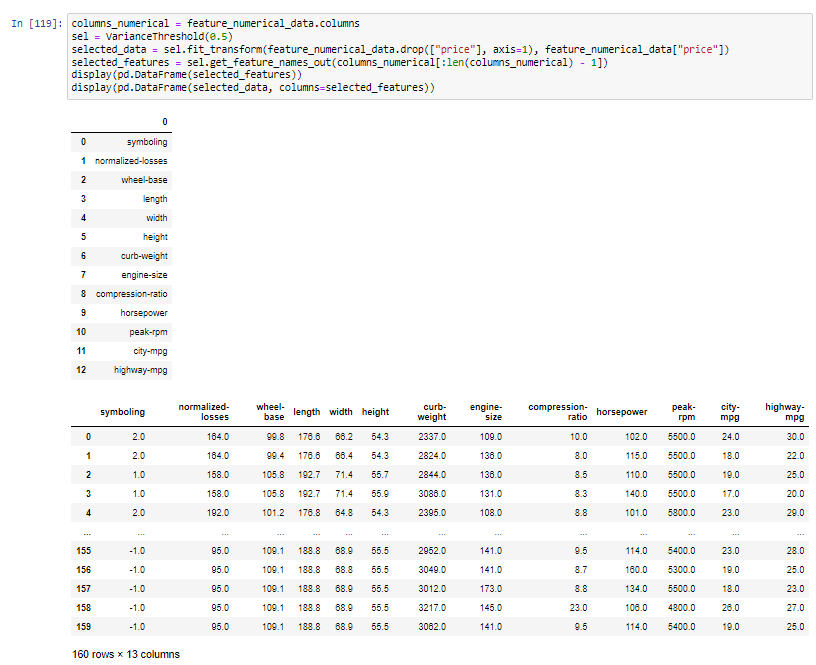


Далі розділив дані на категоріальні та числові

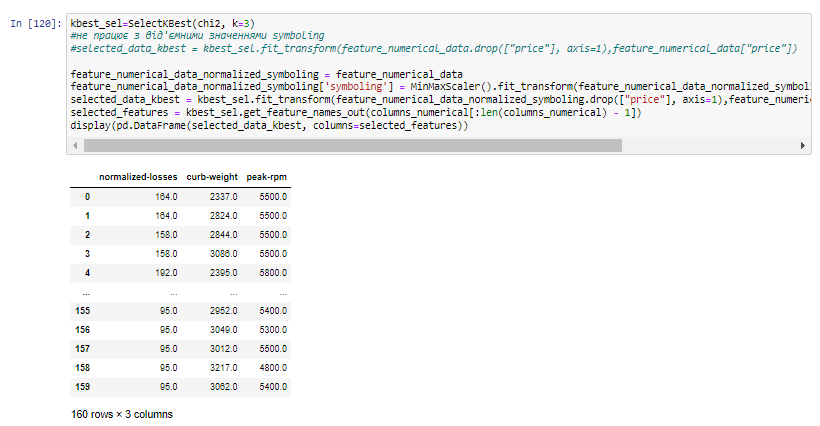




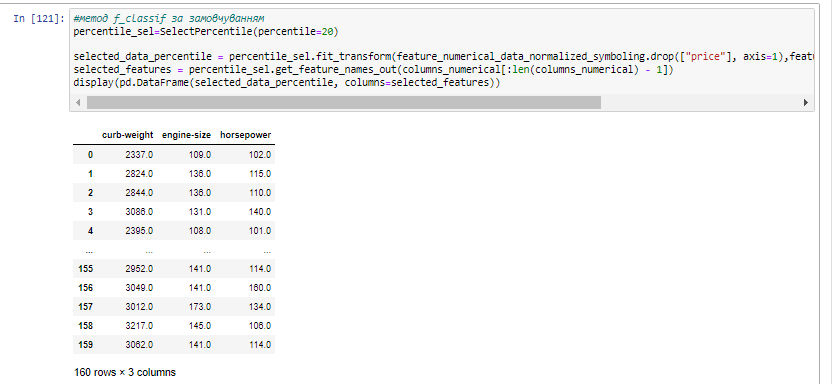
Далі почав використовувати методи для оцінки значущості факторів. Спочатку відкинув стовпці з дисперсією менше 0.5. Цей метод відсіяв стовпчики bore та stroke



Наступний метод обирає k найвпливовіших факторів, тут використовував chi square для обирання факторів. Додатково необхідно було прибрати всі від’ємні числа. Найважливішими факторами є normalized-losses, curb-weight, peak-rpm



Останній метод залишає вказаний відсоток найважливіших факторів, залишив за замовчуванням дисперсійний аналіз. Цей метод вважає найважливішими такі стовпчики - curb-weight, engine-size, horsepower



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

На останок візуалізував корреляцію значень даних, це також дозволяє візуально знайти фактори, що найбільш впливають на ціну

