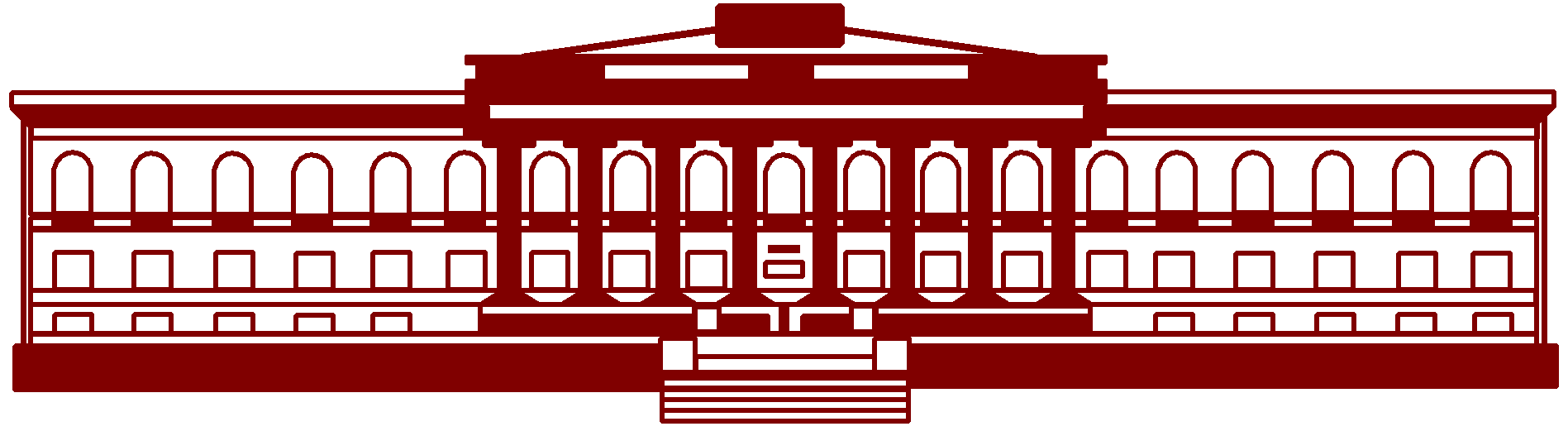
1. **КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**
2. **імені ТАРАСА ШЕВЧЕНКА**
4. ****
5. **ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**
6. **Кафедра прикладних інформаційних систем**
7. **Звіт до лабораторної роботи №3**
8. **з курсу**
9. **«Data Science та Big Data»**
10. *студентки 4 курсу*
11. *групи ПП-42*
12. *спеціальності 122 «Комп'ютерні науки»*
13. *ОП «Прикладне програмування»*
14. *Романович І.В*
15. *Викладач:*
16. *Шолохов О.В*
17. *Білий Р.О*
18. **Київ – 2023**

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

**Мета:** метою лабораторної роботи є отримання практичних навичок аналізу та вибору значущих ознак для моделі за допомогою кореляційного аналізу, таблиць сопряжіння, аналізу багатомірні залежності та дихотомії, дисперсійного аналіз – ANOVA, критерій Хі-квадрат тощо.

**Контекст:**

Ви – data analyst у компанії, яка торгує підтриманими автомобілями по всій Америці (викупає у власника, та перепродає). Ваше керівництво надало вам завдання проаналізувати наявні дані та виявити серед них фактори (ознаки), які впливають на ціну, а також структуру взаємозалежності факторів, та оформити результати дослідження у звіт.

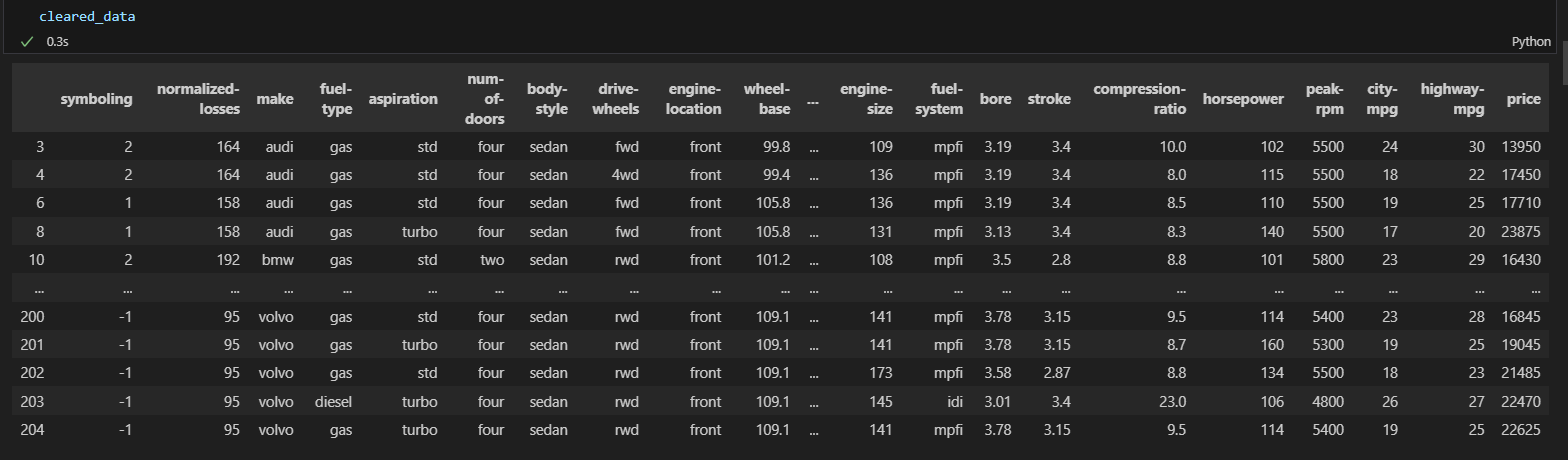
Наданий вам набір даних складається з даних з автомобільного щорічника Ward's Automotive Yearbook за 1985 рік.

**Завдання:**

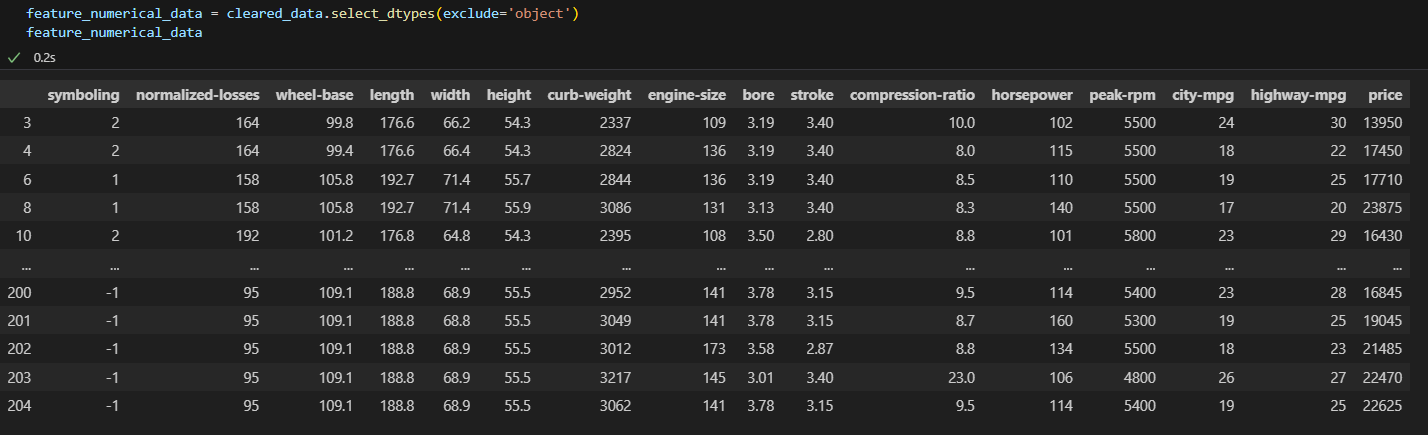
1. Завантажити файли з даними у папку проекту з посилання:
2. Очистити дані та обробити відсутні дані.
3. Зробити EDA по ознаках.
4. Проаналізуйте надані дані, використовуючи методи з прикладу та документації, та зберіть результати аналізу у результуючий ранжируваний датафрейм, в якому лівим індексом 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).
5. Проаналізуйте ознаки на взаємозалежність, та побудуйте відповідні heatmap засобами seaborn по кожному з використаних методів дослідження.
6. Зберіть висновки у звіт (графіки, висновки текстом у окремому файлі), який потребує належного оформлення, структури тощо.

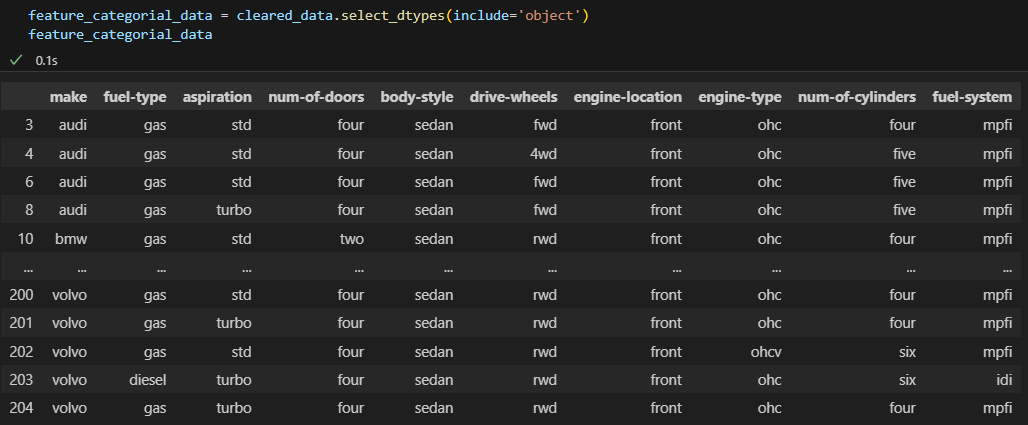
**Хід роботи:**

Очистимо дані

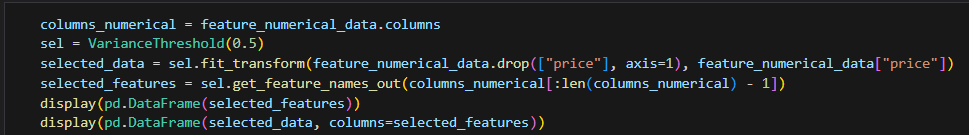


Далі розділяємо дані на категоріальні та числові

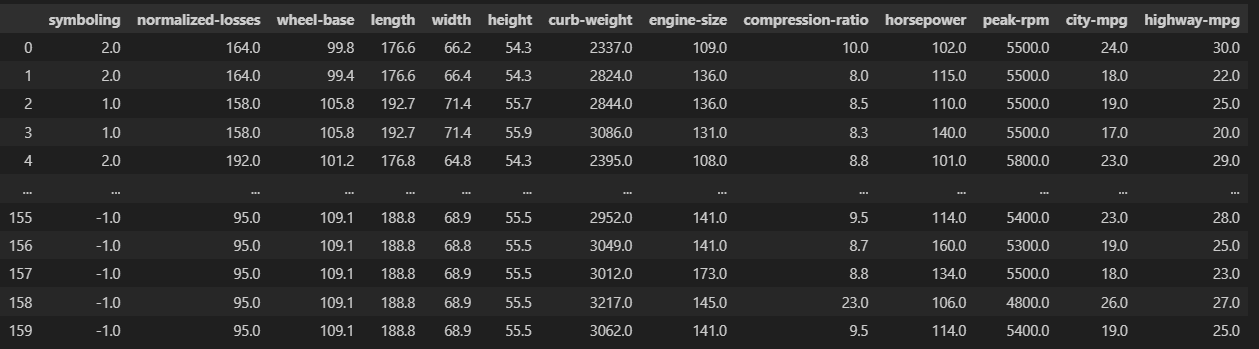




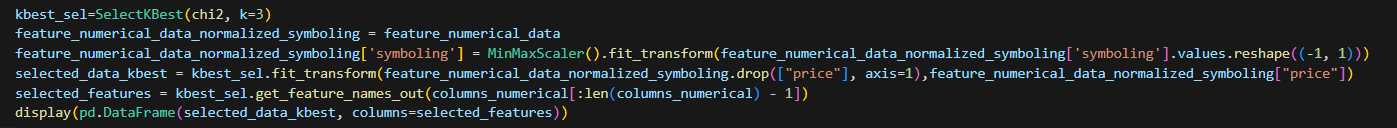
Далі використовуємо методи для оцінки значущості факторів. Спочатку відкидаємо стовпці з дисперсією менше 0.5. Цей метод відсіяв стовпчики bore та stroke.

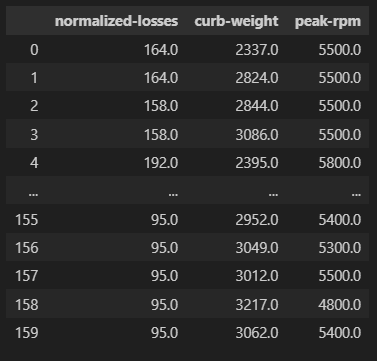




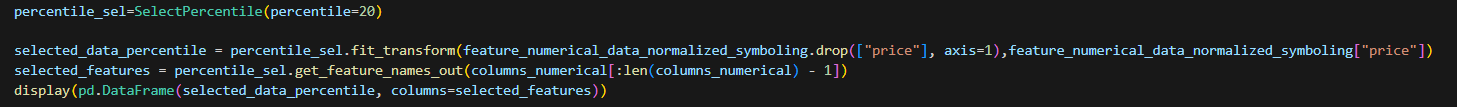


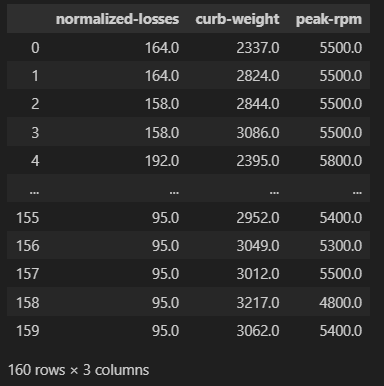
Наступний метод обирає k найвпливовіших факторів, тут використовував chi square для обирання факторів. Додатково необхідно було прибрати всі від’ємні числа. Найважливішими факторами є normalized-losses, curb-weight, peak-rpm





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

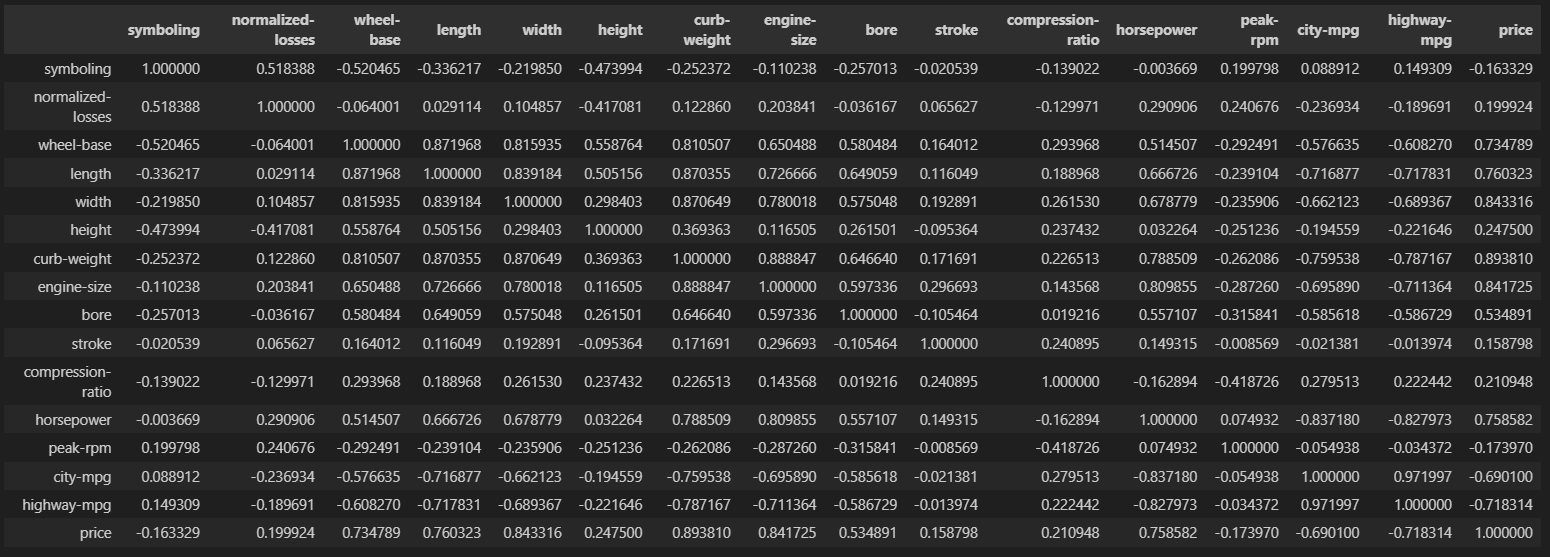




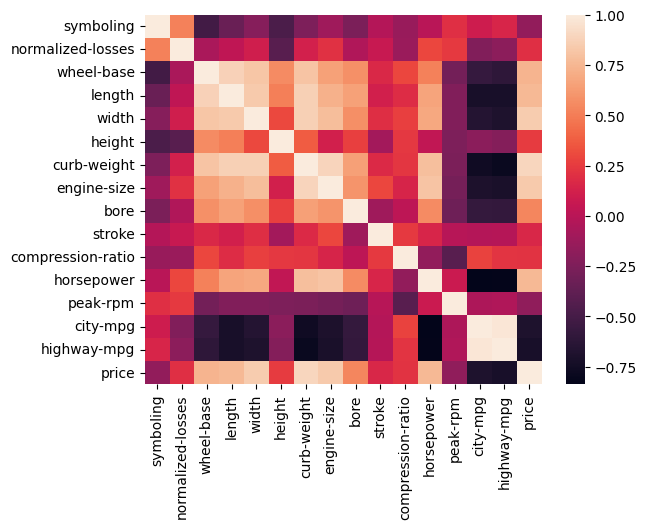
Отже, за результатами цих трьох методів можна сказати що найбільш на ціну впливають такі показники - normalized-losses, curb-weight, peak-rpm, engine-size, horsepower. Важливо зазначити, що найвпливовішим є показник **curb-weight.**

Наостанок візуалізуємо кореляцію значень даних, це також дозволяє візуально знайти фактори, що найбільш впливають на ціну.









**Висновки:**

Після проведення аналізу ринку у сфері вживаних автомобілів в США було зроблено важливі висновки, які сприятимуть більш ефективному розробленню стратегій з продажу та маркетингу.

Одні із ключових факторів, що визначають ціну автомобіля - їхні розміри, об'єм двигуна та потужність. Великі авто з потужними двигунами, як правило, коштують дорожче. Крім того, престиж бренду та стиль кузова також впливають на цінову категорію, що означає, що автомобілі відомих виробників та авто з елегантним дизайном частіше мають вищі ціни.