Міністерство освіти і науки України

Київський національний університет імені Тараса Шевченка

Факультет інформаційних технологій

Кафедра прикладних інформаційних систем

Звіт

з лабораторної роботи №3

з дисципліни “Data Science та Big Data”

Виконала:

Костюк Д.О.

група ПП-42

Перевірив:

Київ 2023

**Лабораторна робота №3**

## Тема роботи: Методи аналізу та вибору значущих ознак

## Мета роботи: 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.

## Завдання

• 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/1su22W8JrRZzm0mea5v8x46YmLh083qp/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 врахувала 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.

**Хід виконання**

Висновки:

Найбільше на ціну впливає розмір двигуна, майже усі методи визначили його як надважливий.

Окрім цього методи вибору значущих ознак визначали важливою характеристикою вагу корпусу.

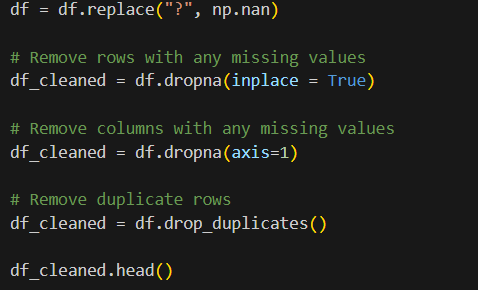
Розміри авто, тобто довжина, висота та ширина, найбільше корелюють між собою та з вагою, колісною базою і розміром двигуна

Символіка, марка, кількість циліндрів, паливна система, тип двигуна, сильно впливають на ціну автомобіля, причому найбільший вплив має виробник.

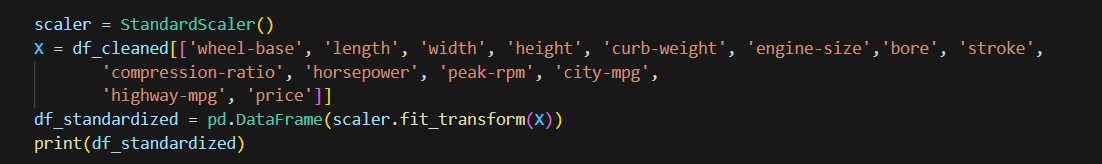
Оптимізації ціни можна досягти за рахунок покращення функціональності, оптимізації кількості циліндрів, раціонального використання ресурсів двигуна.

Хід виконання лабораторної

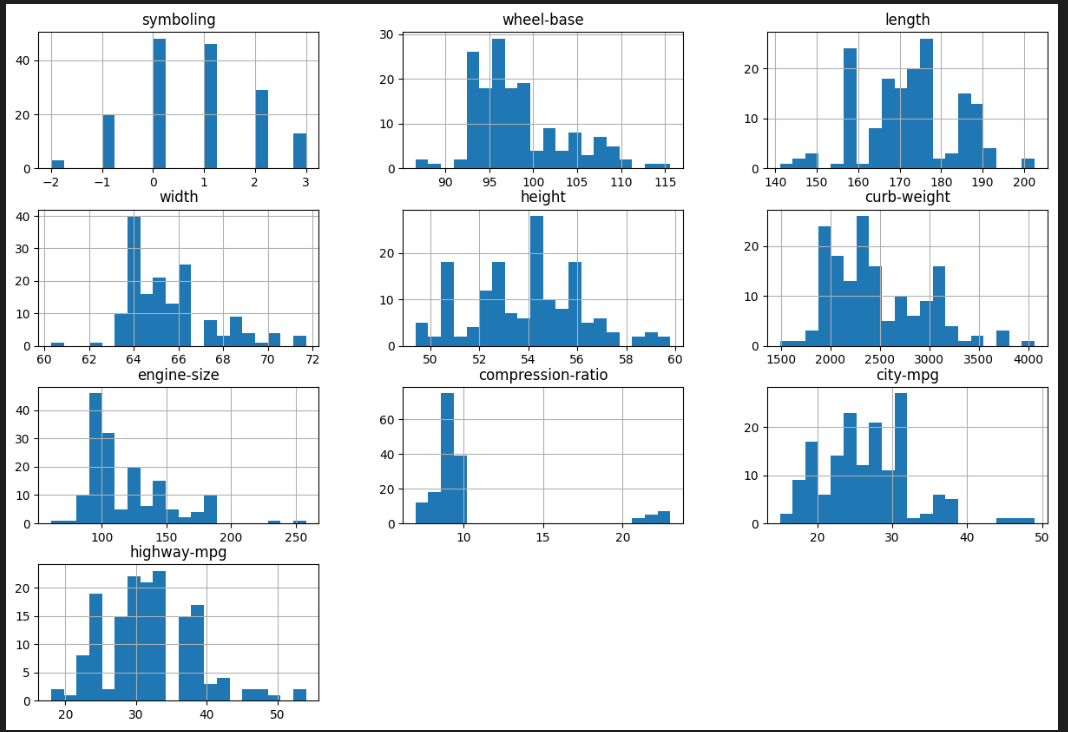
Очистила дані, які не містять значень:



Нормалізувала дані датасету:

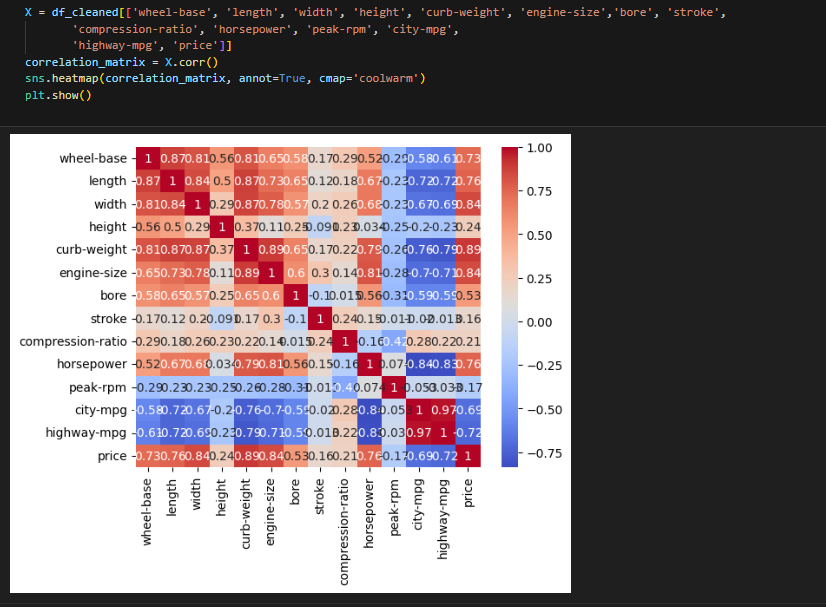


Гістограма з очищеними та нормалізованими даними:

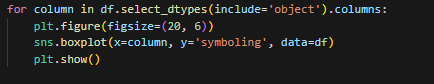


Створення графіків для певних значень із датасету для EDA

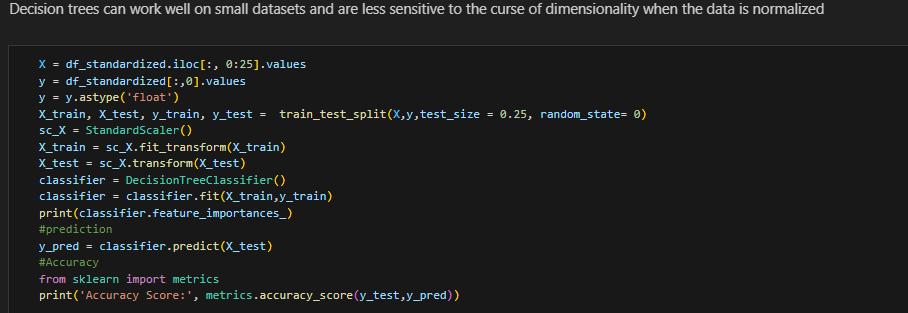
Матриця кореляції для значень датасету:



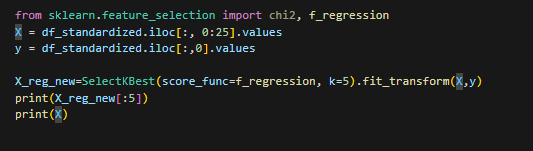
Створення графіків залежності числових значень від категоричних:



Дерево прийняття рішень:



K-best:



## Висновок: у ході виконання даної лабораторної роботи я отримала практичні навички 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.