**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**імені ТАРАСА ШЕВЧЕНКА**

Факультет інформаційних технологій

**Кафедра прикладних інформаційних систем**

****

**Звіт**

до виконання лабораторної роботи №3

з дисципліни «Data Science та Big Data»

на тему:

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

(*тема роботи*)

ВАРІАНТ №5

**Виконано:**

студ. групи ПП- 41, підгрупа 1

Горошко Андрія

**Перевірено:**

Білий Р.О.

**Київ – 2023**

**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/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 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.

**Виконання:**

Заімпортимо всі потрібні бібліотеки та виведемо початкому таблицю:

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

import numpy as np

from sklearn.preprocessing import LabelEncoder

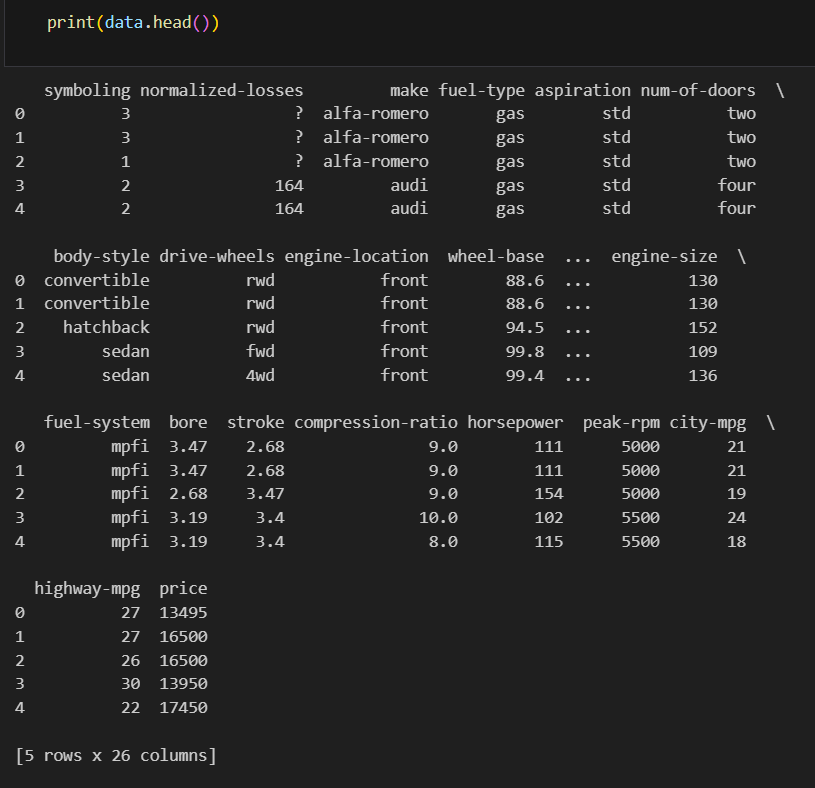
from sklearn.feature\_selection import VarianceThreshold, SelectKBest, f\_classif, mutual\_info\_classif, SelectFromModel

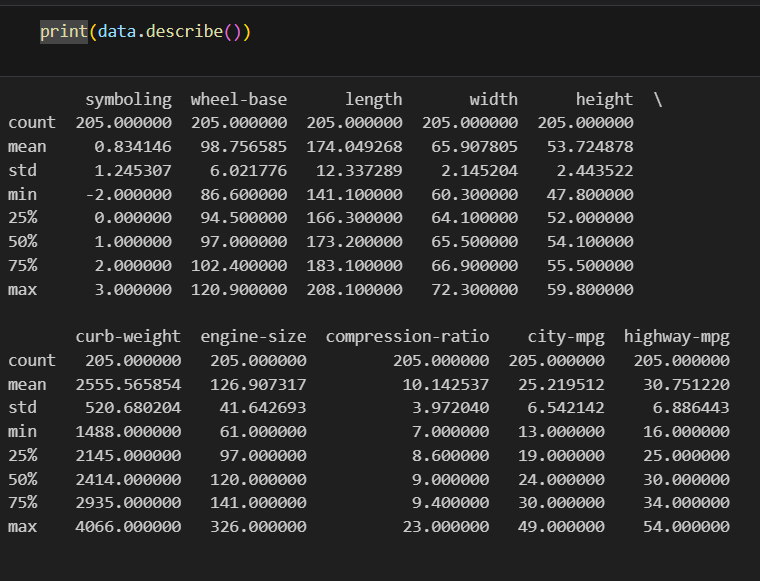
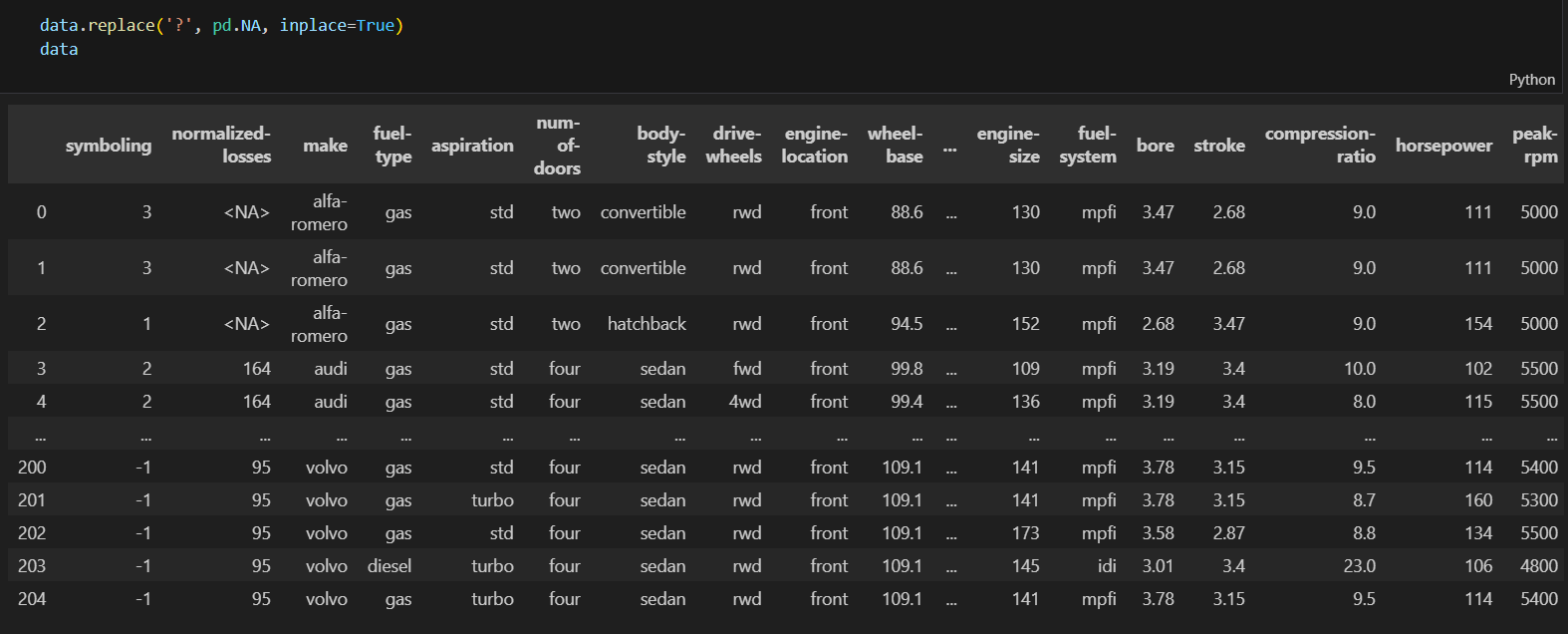
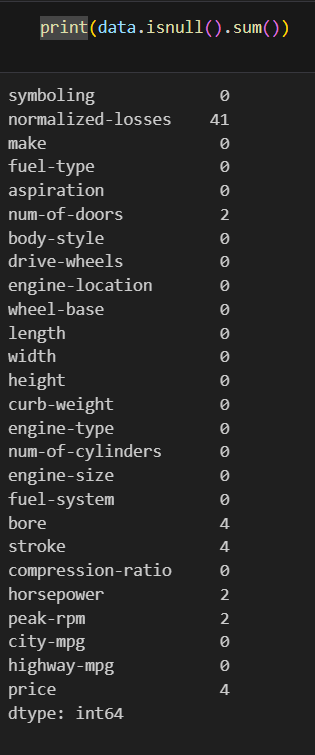
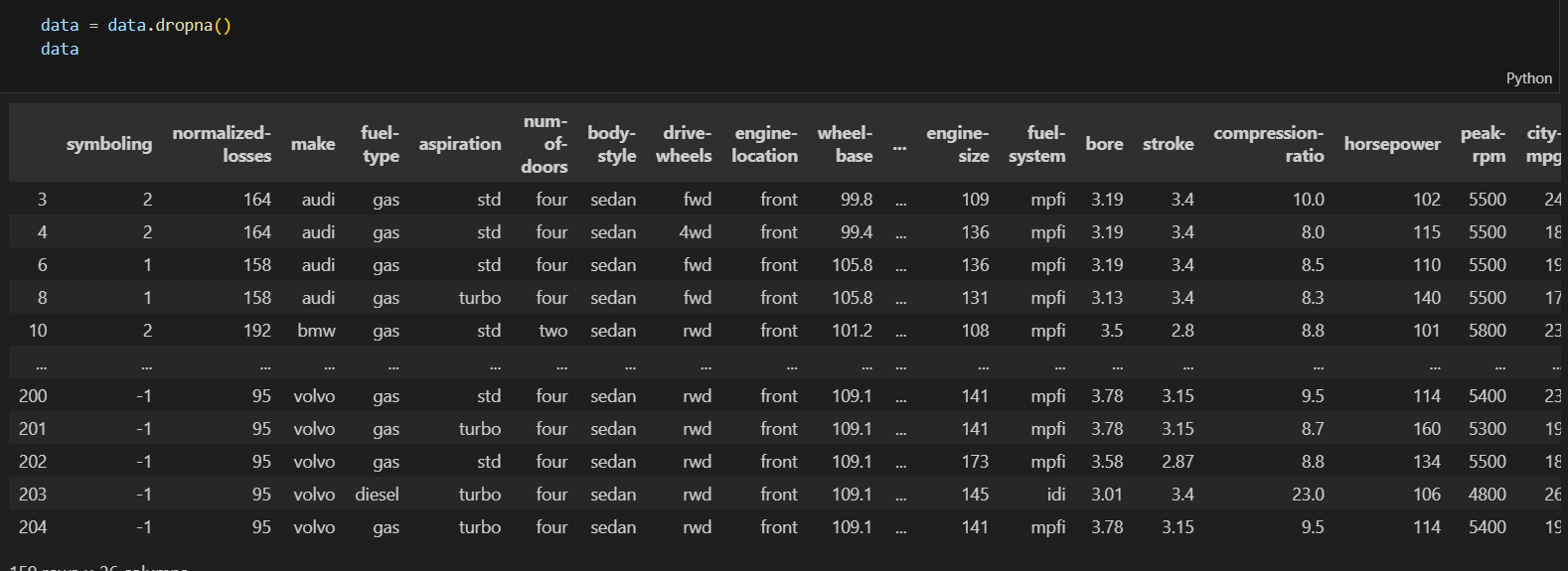
from sklearn.linear\_model import LogisticRegression, Lasso

data = pd.read\_csv(f'D:/Python/lab3/Automobile\_data.csv', sep=',')

data

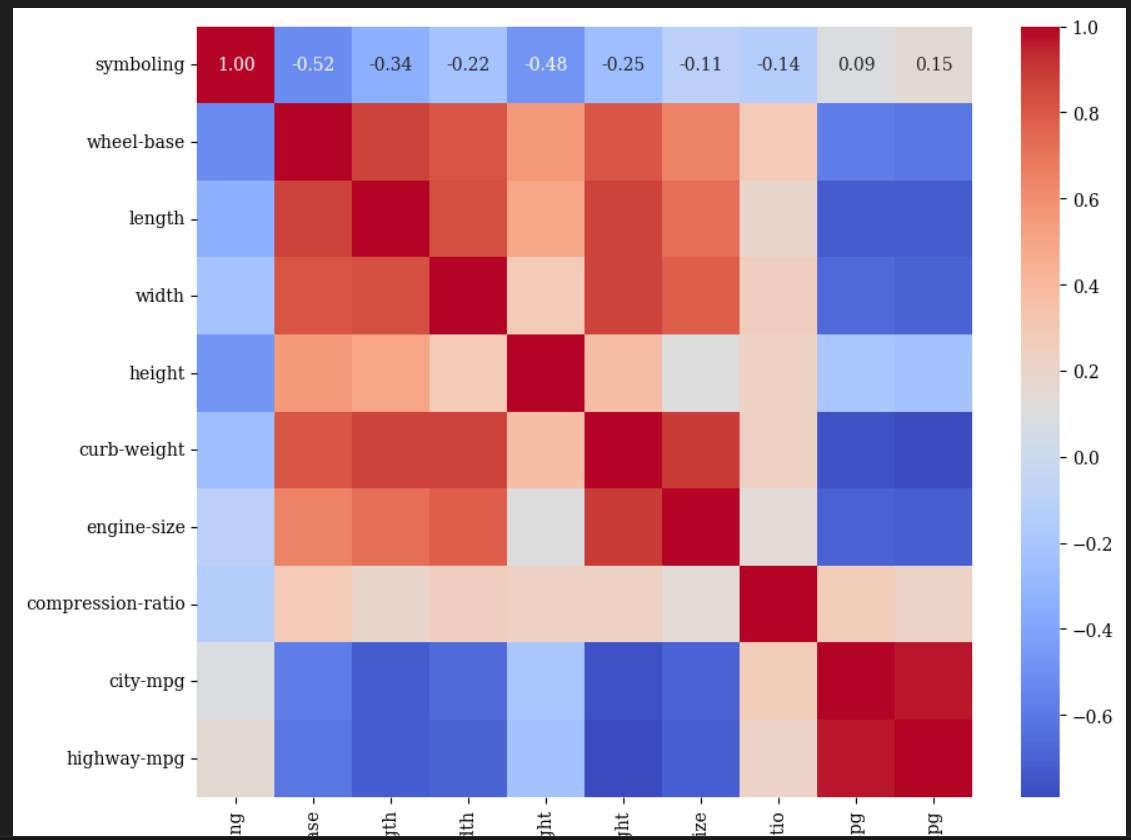


Проведемо процес нормалізації даних: 

Далі приступаємо до аналізу даних:

На першому скріншоті зображена теплова картина співвідношень даних



Далі я побудував діаграми розсіювання для всіх змінних.

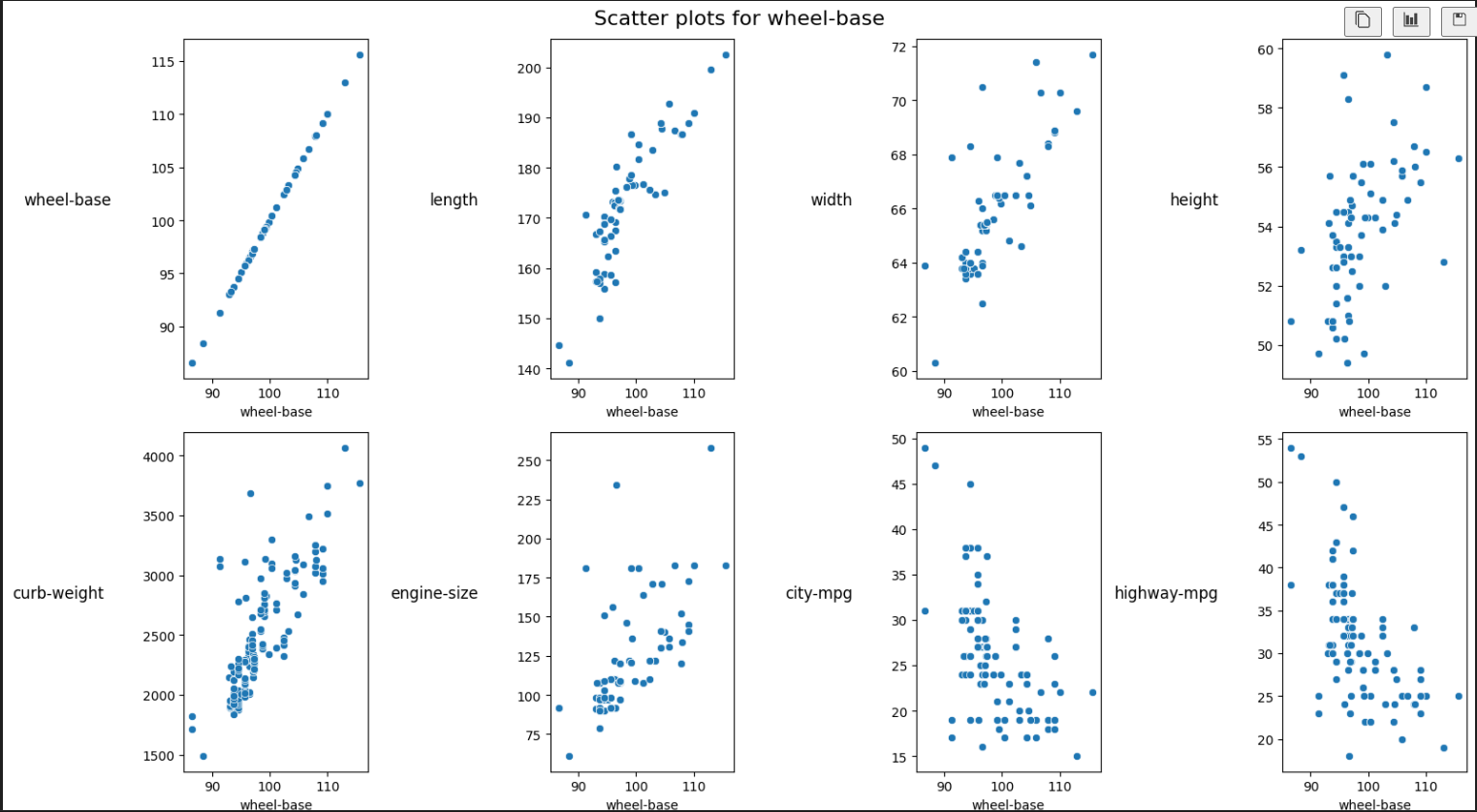
Основна мета діаграм розсіювання полягає у візуалізації зв'язку між двома числовими змінними.

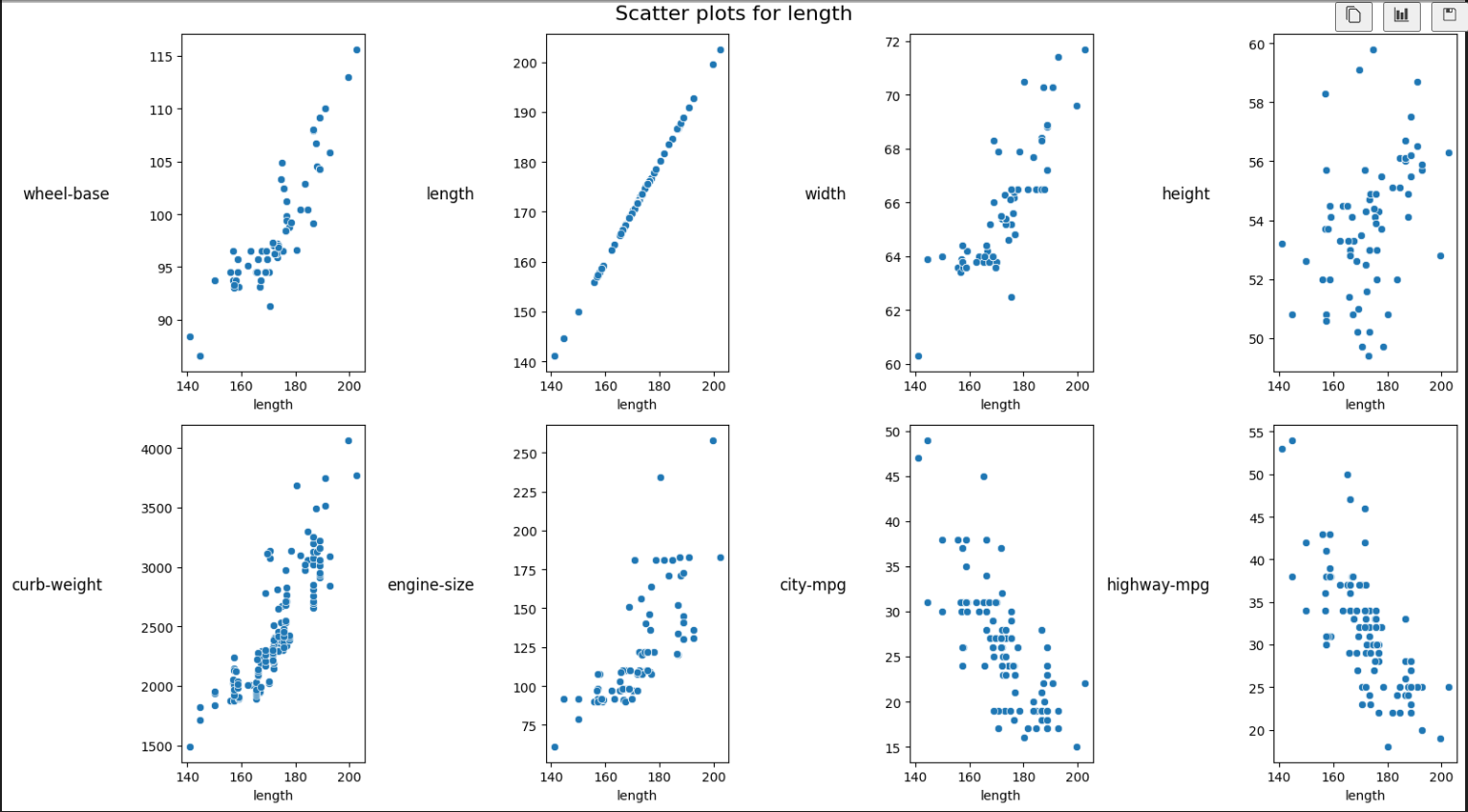
Кожен графік розсіювання відповідає конкретній парі змінних, де вісь x представляє одну з ознак, а вісь y відображає ціну.

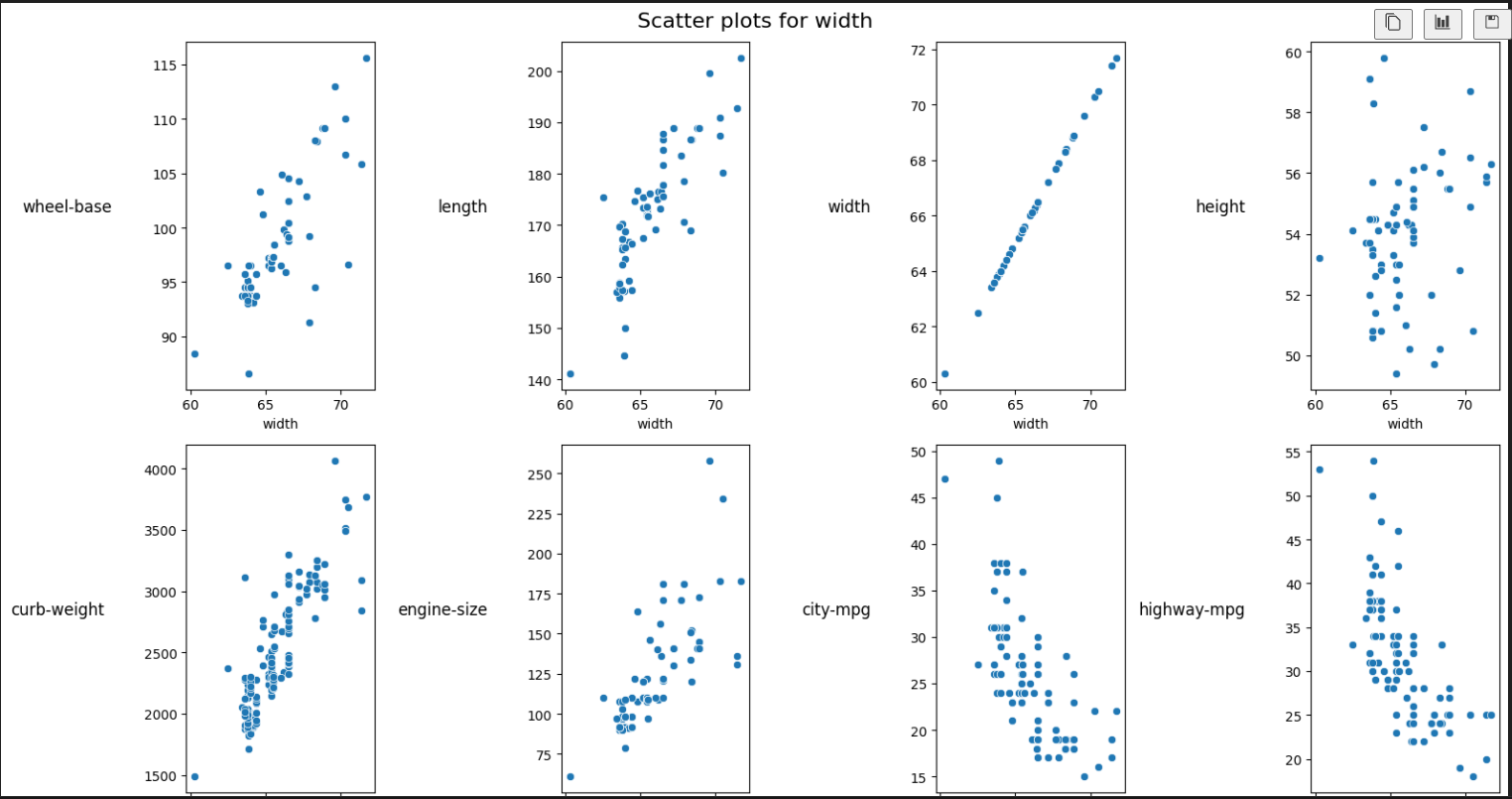
Ці графіки дозволяють спостерігати за змінами вартості автомобіля в залежності від конкретної числової характеристики,

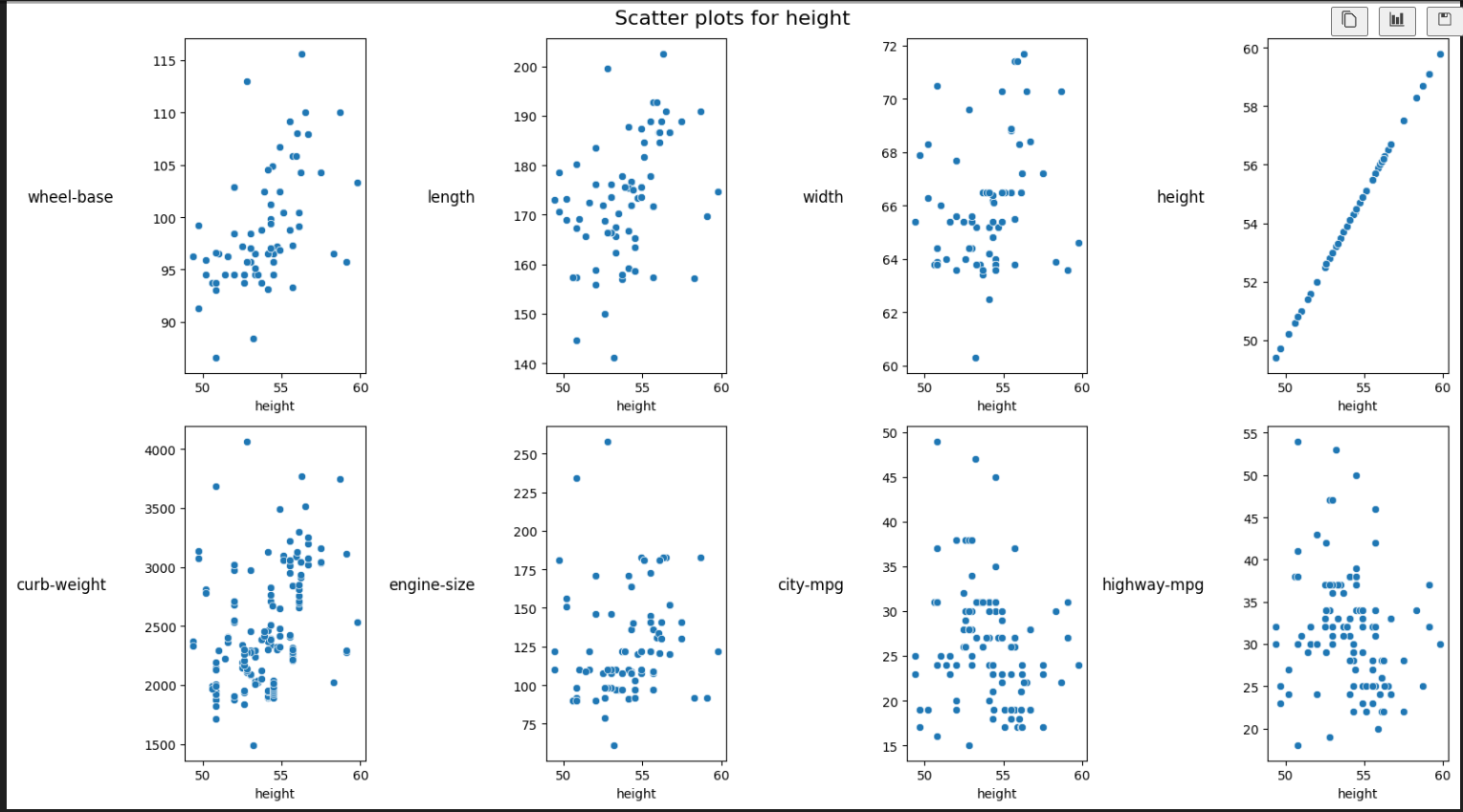
що допомагає візуалізувати можливі статистичні зв'язки між властивостями та цільовим параметром - ціною.

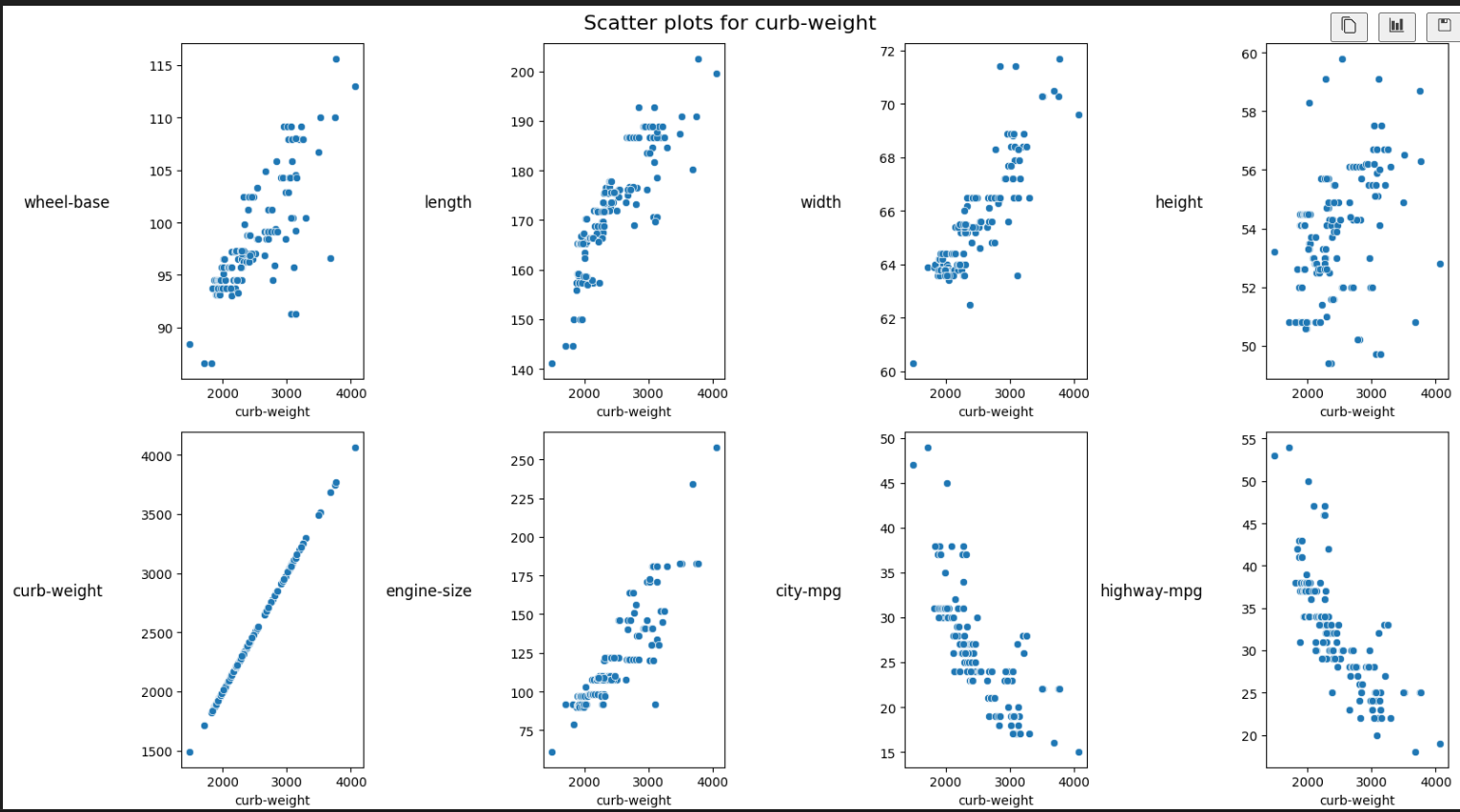
Із Графіків розсіювання для відношень ціни та числових властивостей бачимо, що не всі дані гарно корелюють між собою і хоч якось впливають один на одного, що впринципі очевидно.

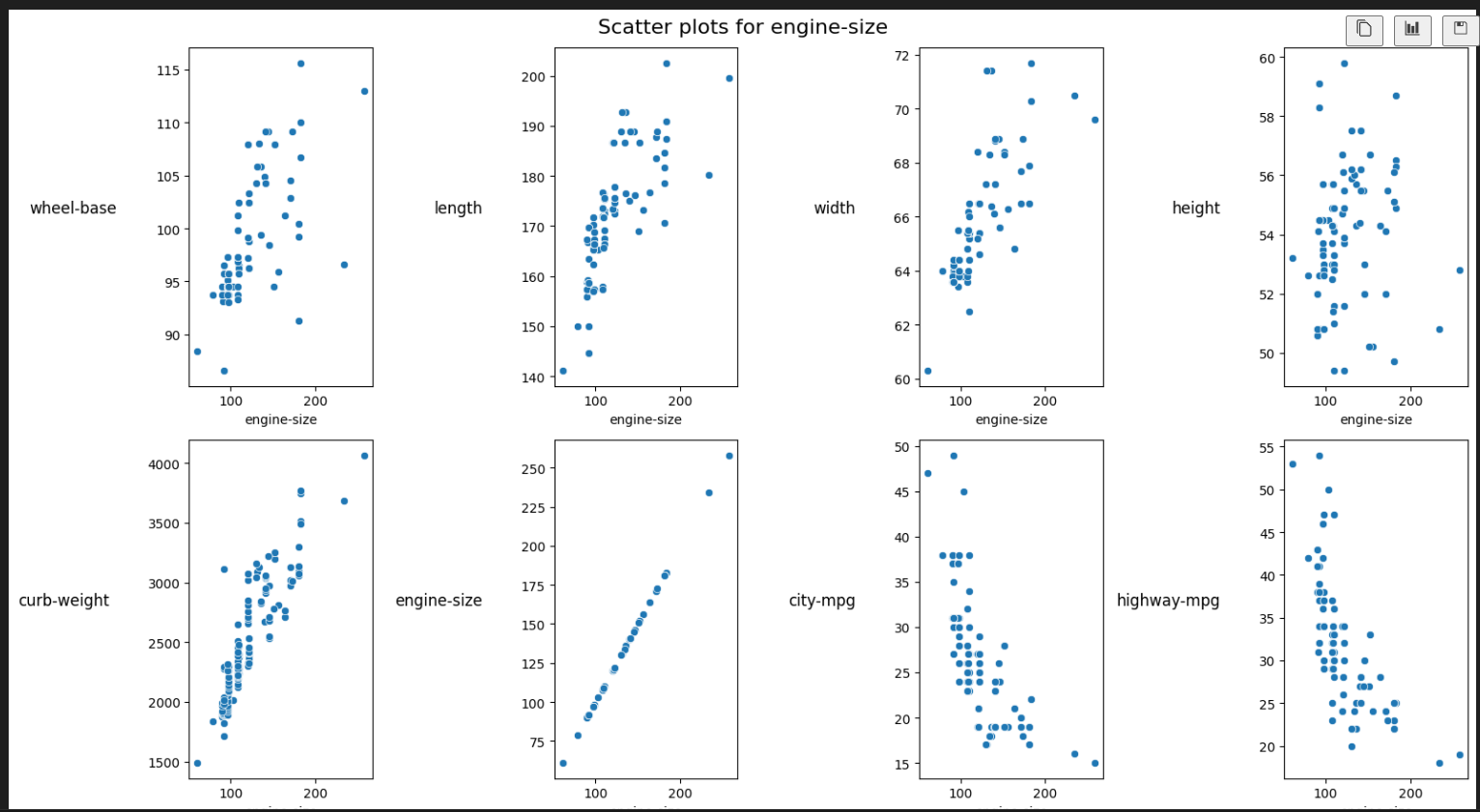


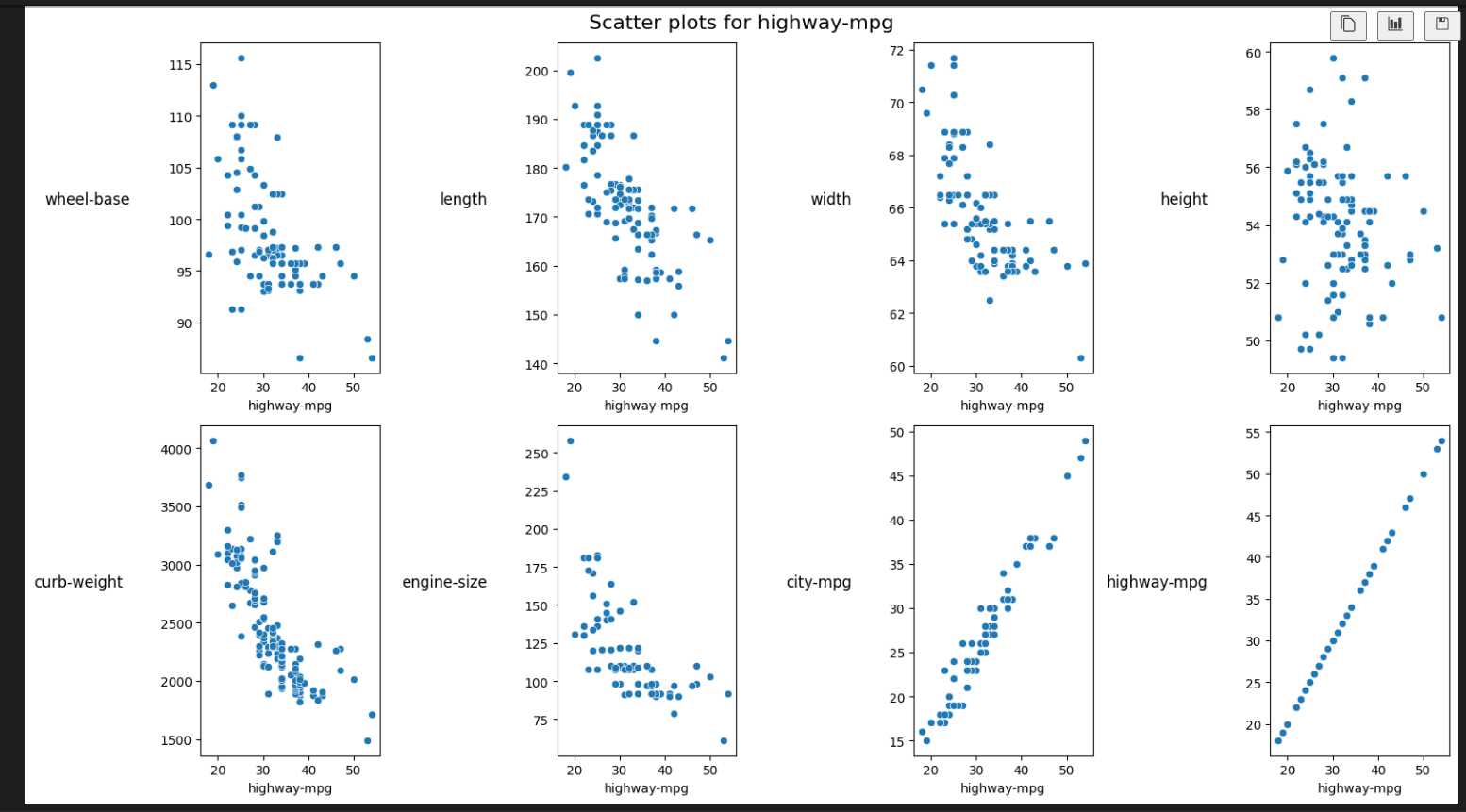


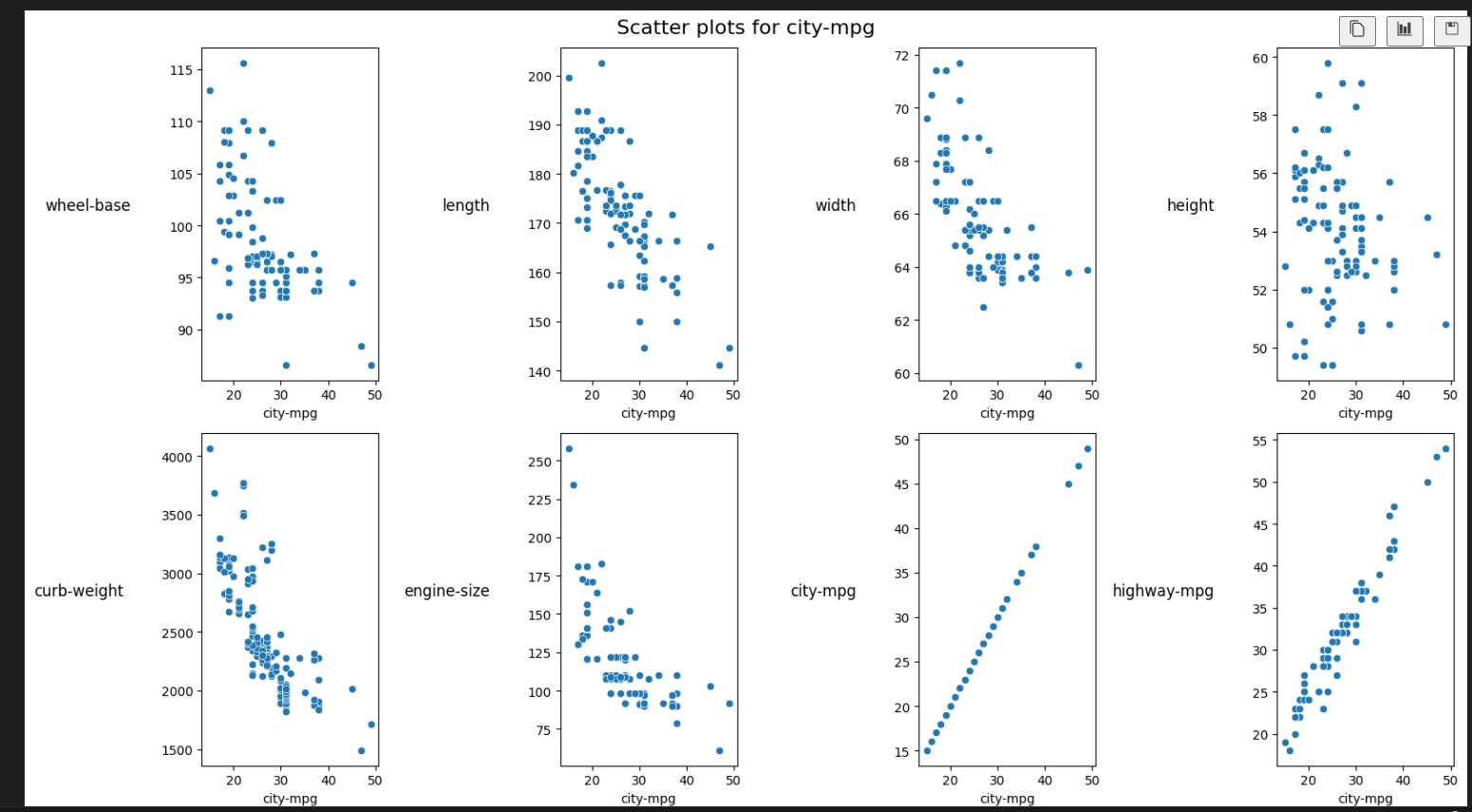






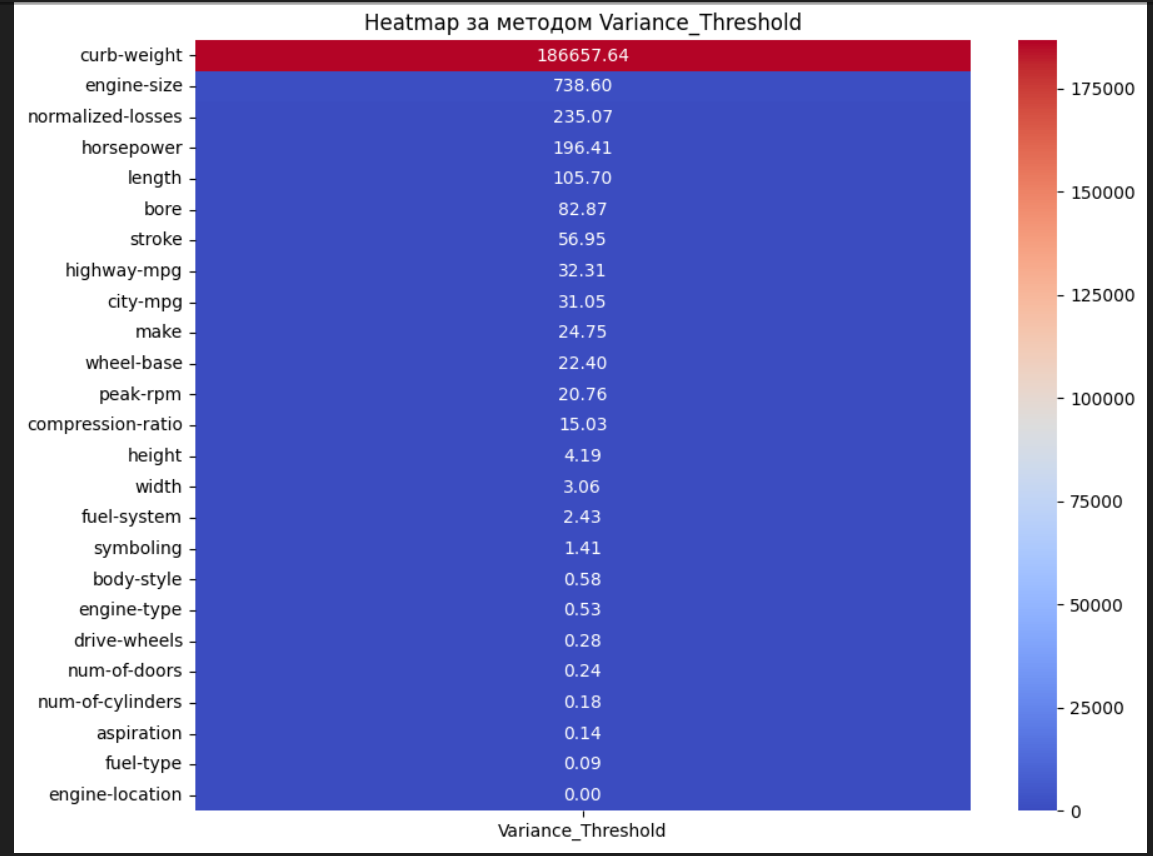




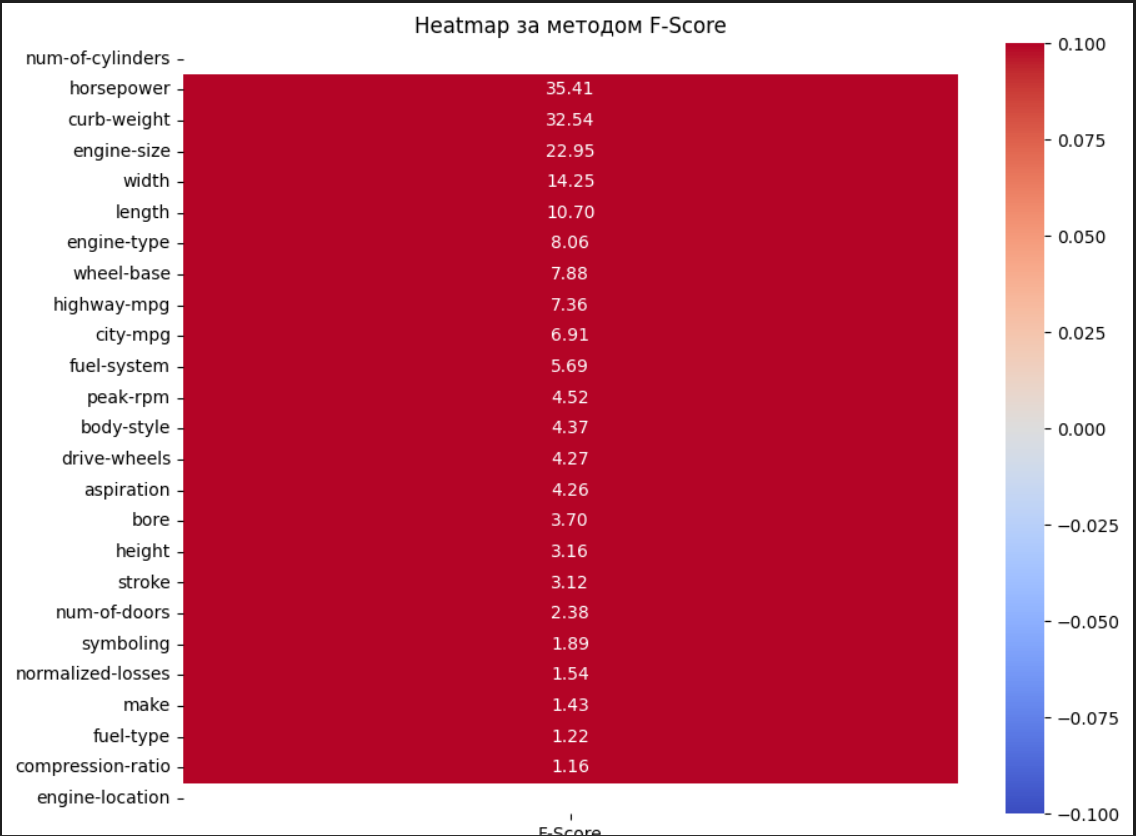


Проведено дослідження за методами відбору значущих ознак:

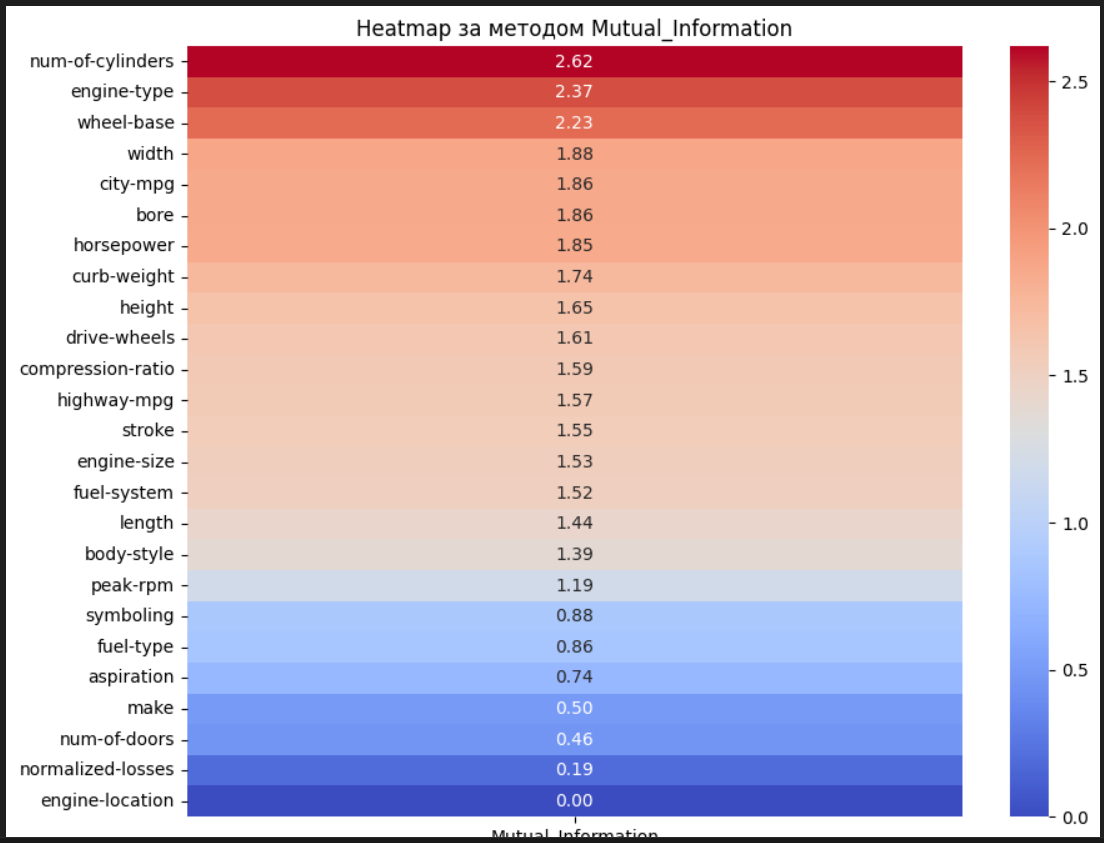
За допомогою Variance Threshold розглянемо дисперсію та варіативність для кожної ознаки.



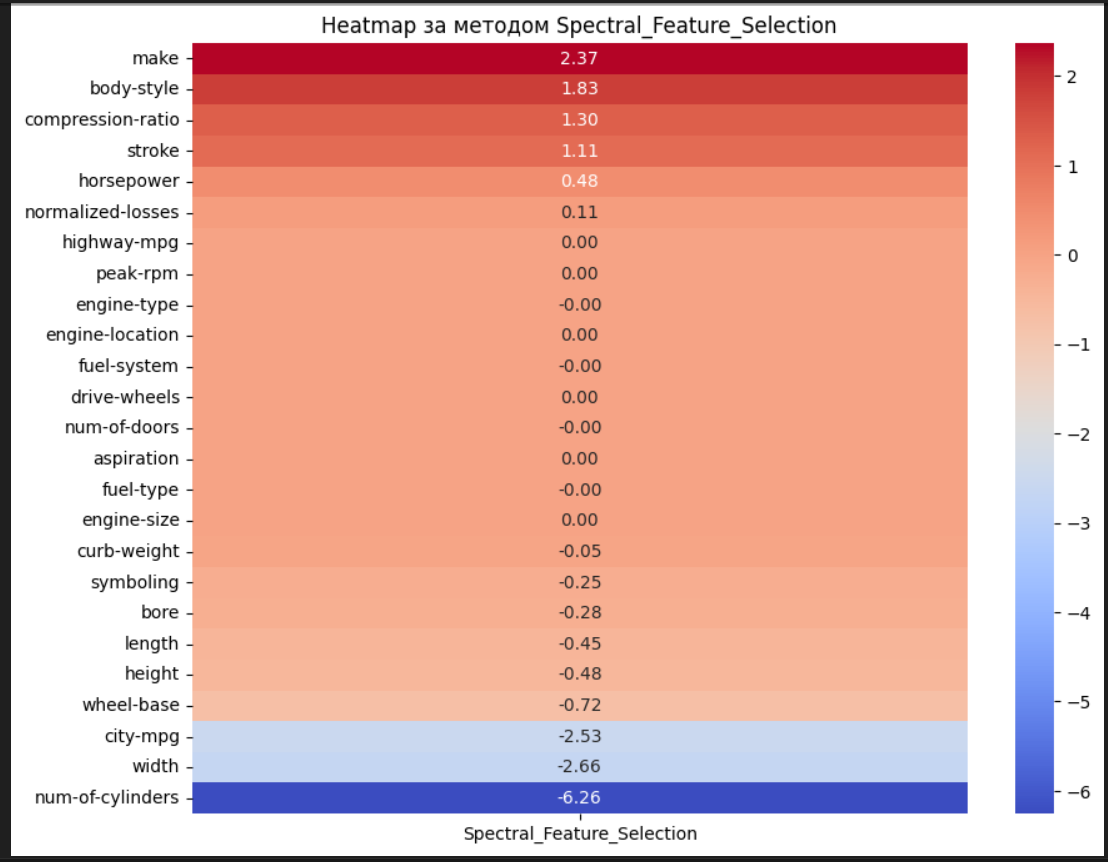
Далі за допомогою F-Score кожній ознаці надається бал, який базується на тому як добре вона розрізняє класи в залежних змінних



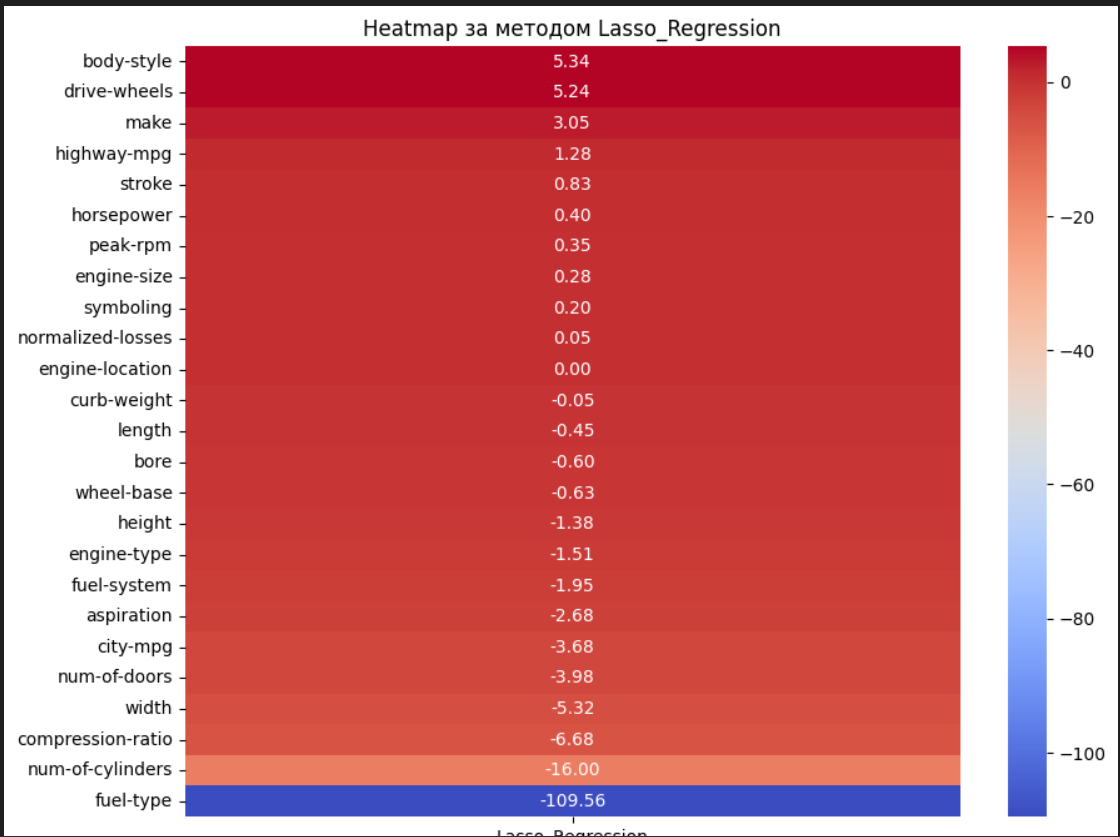
Також Mutual Information, який я використав для того щоб дізнатися інформативність та деяку унікальність кожної ознаки



Spectral feature selection був використаний також для додаткового визначення важливості ознак



І Lasso regression у якості ще одного додаткового параметра, який мінімізує суму квадратів помилок.



Висновки дослідження:

Дослідження дозволяє нам зробити кілька висновків щодо важливості представлених ознак у визначенні ціни автомобілів:

Make (Марка): Ця ознака виявилася дуже важливою. Ймовірно, вона має суттєвий вплив на модель, але важливість може залежати від конкретного способу кодування марок автомобілів.

Числові ознаки (wheel-base, length, width, curb-weight, engine-size, horsepower): Ці ознаки мають високу важливість, особливо engine-size і horsepower.

city-mpg, highway-mpg, fuel-type, aspiration, body-style, drive-wheels, engine-type, fuel-system: За результатами дослідження можна вважати, що ці числові ознаки також є важливими.

Числові ознаки з середньою важливістю (normalized-losses, bore, stroke, compression-ratio, peak-rpm): Важливість цих ознак є середньою, і можливо, варто детальніше розглянути інші аспекти.

Symboling: Ця ознака може вважатися менш важливою для моделі, оскільки важливість виявилася низькою.

engine-location: Має нульову важливість у двох методах. Це може свідчити про те, що ця ознака не має великого впливу на модель і ціноутворення.

Отже, на основі цих висновків ми можемо виділити змінні, які мають велике значення при визначенні цін на автомобілі під час їх покупки або продажу.

**Висновок:** Під час виконання цієї лабораторної роботи мною було отримано 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.