# КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

# імені ТАРАСА ШЕВЧЕНКА



## ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

**Кафедра прикладних інформаційних систем**

**Звіт до лабораторної роботи №3**

# з курсу

##### **«Data Science та Big Data»**

*Студента 4 курсу*

*групи ПП-41/1*

*спеціальності 122 «Комп'ютерні науки»*

*ОП «Прикладне програмування»*

Компанійця П. О.

*Викладач:*

Білий Р.О.

## Київ – 2023

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

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.

Зміст

[**1.** **Розробка процесів для запланованого функціоналу** 4](#_Toc152922189)

[**1.1** **Графіки розподілу для числових значень** 4](#_Toc152922190)

[**1.2** **Матриця кореляцій для числових значень** 5](#_Toc152922191)

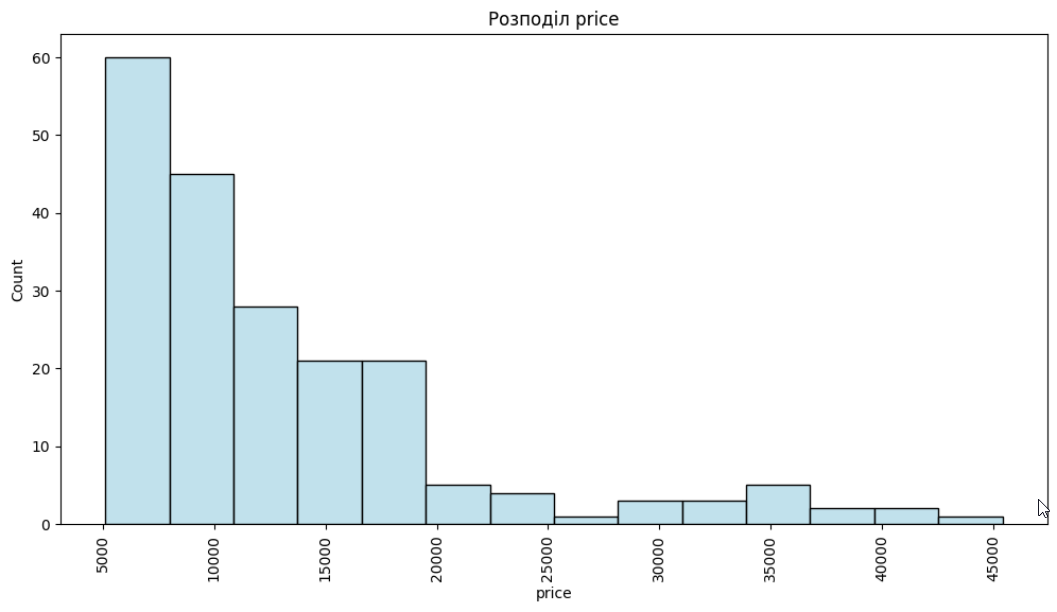
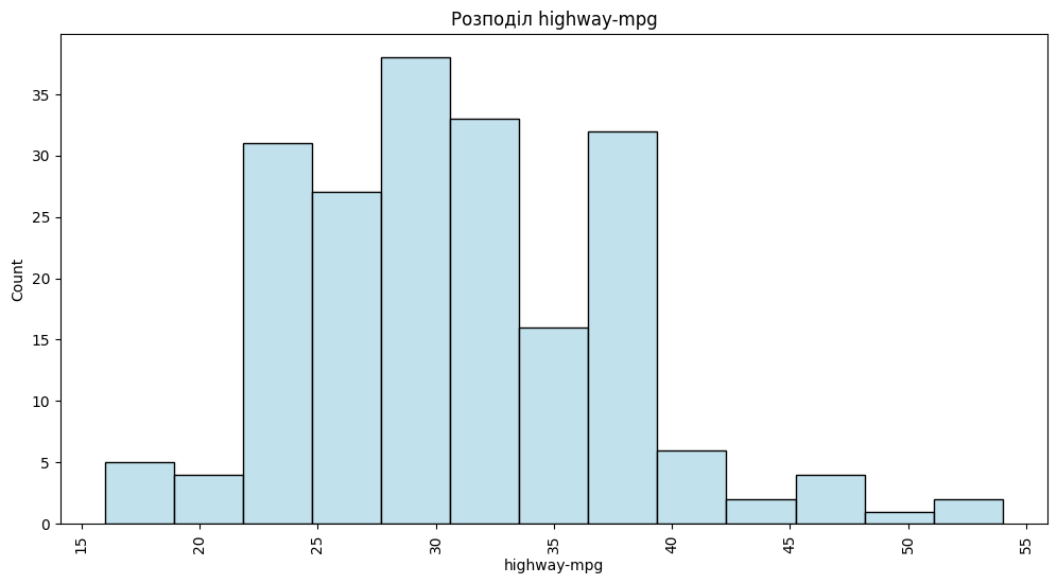
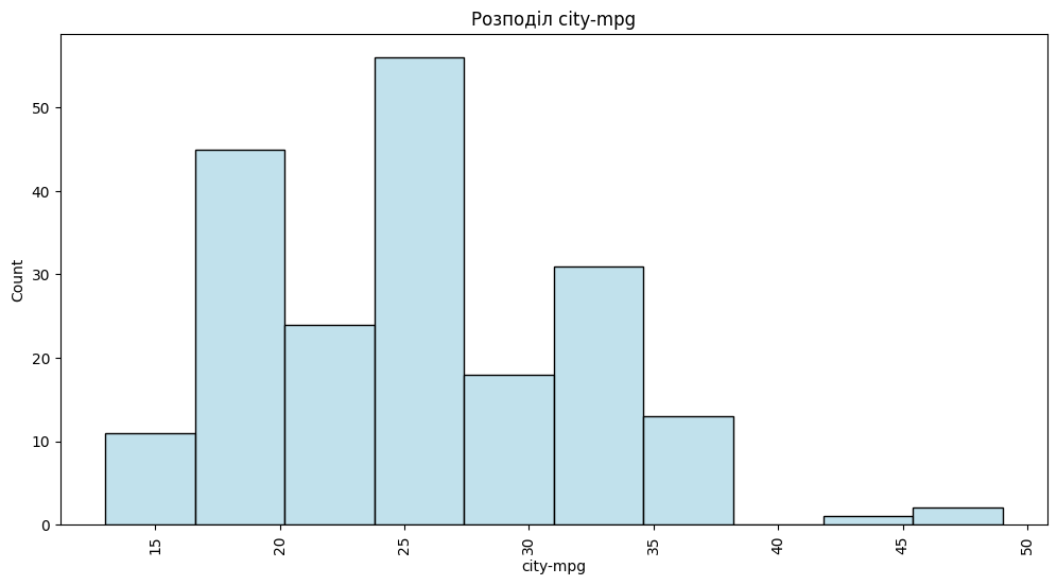
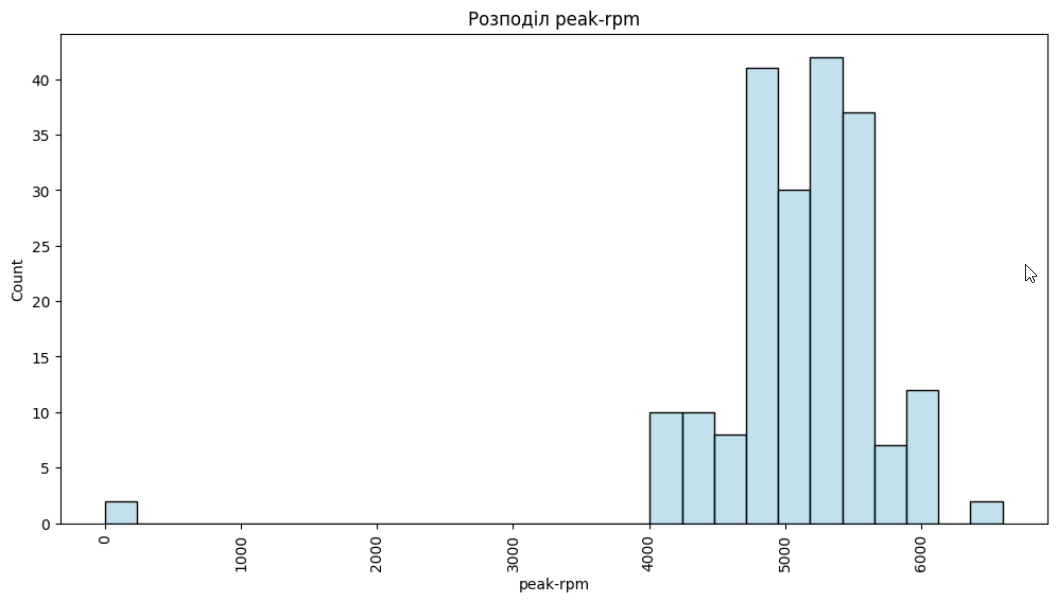
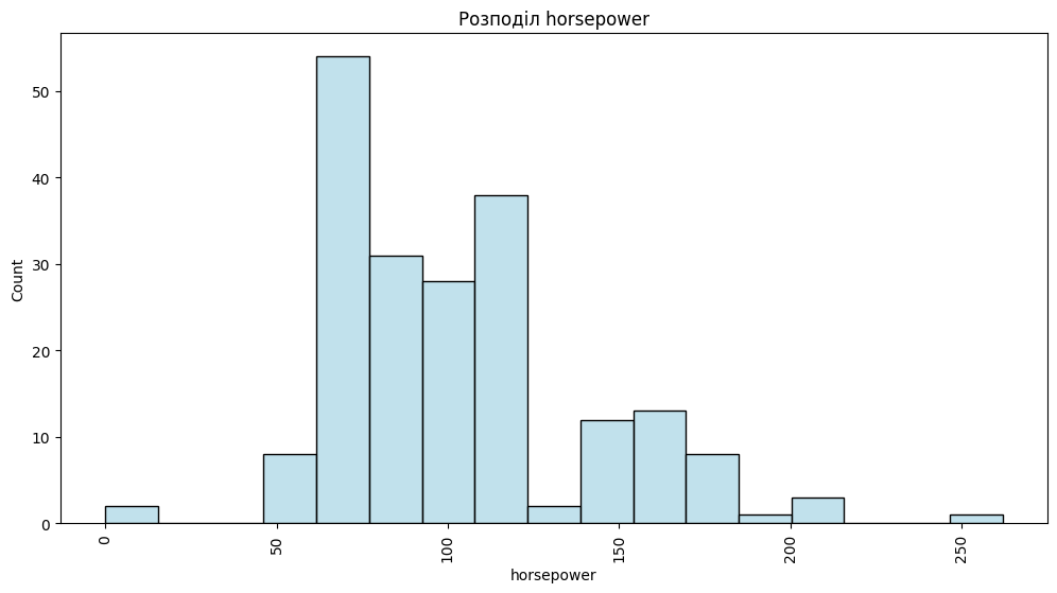
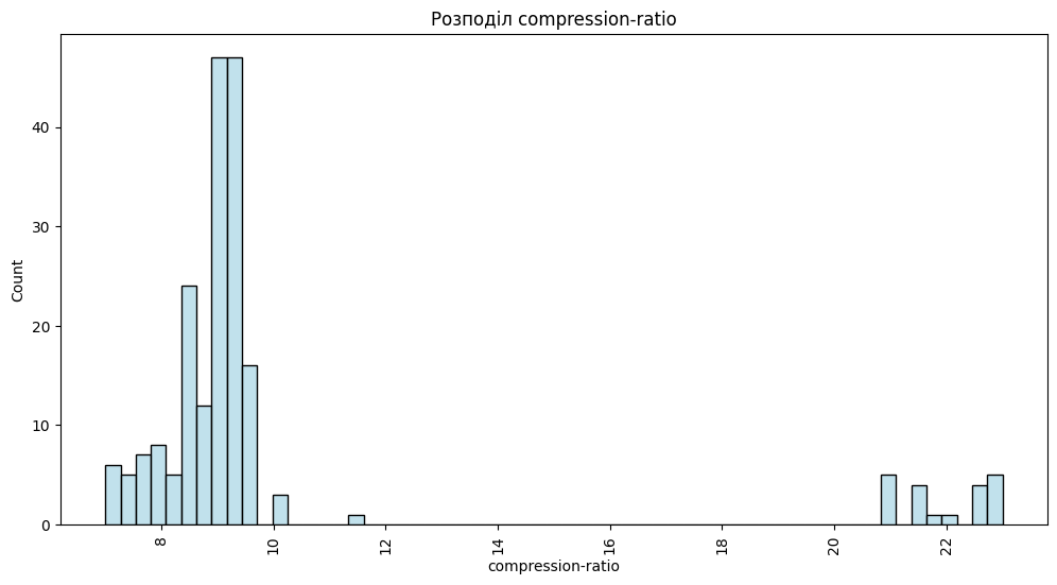
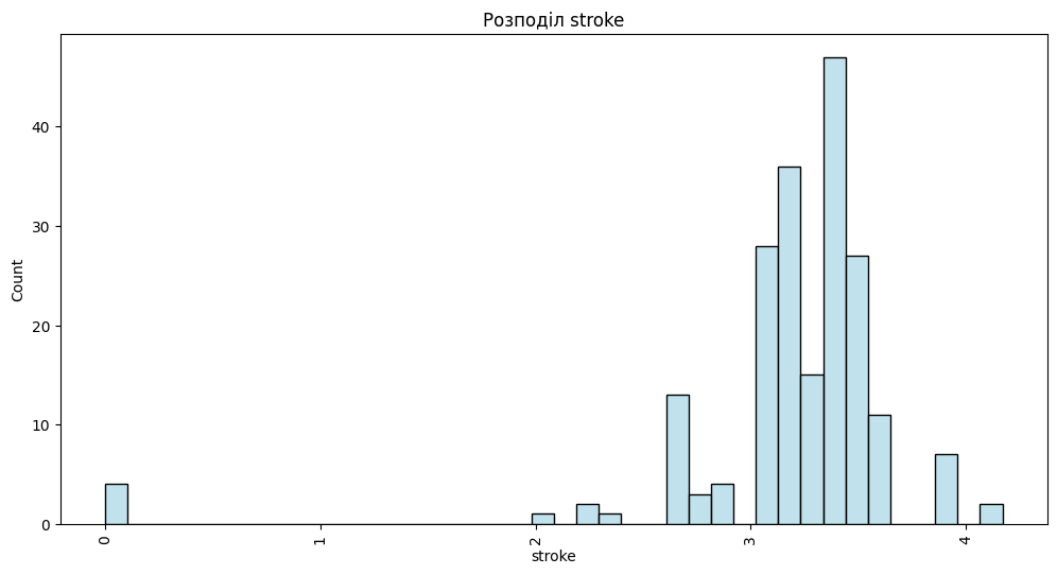
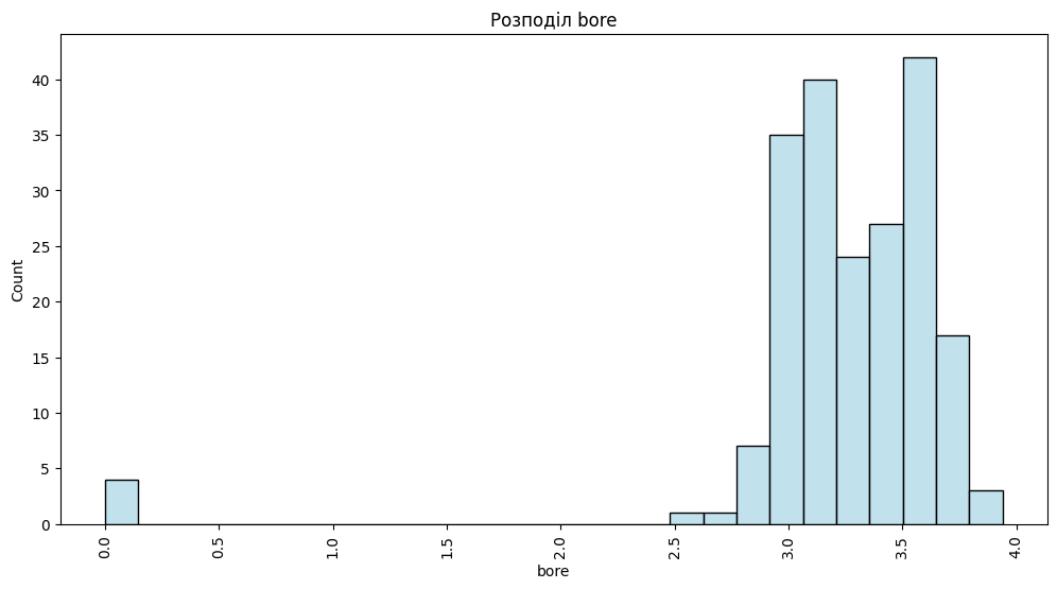
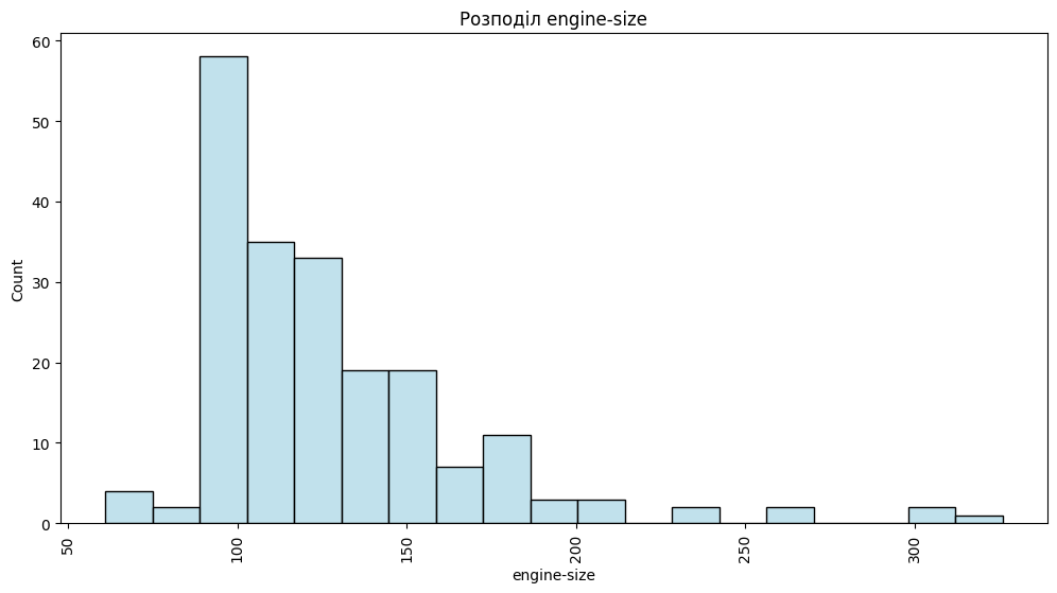
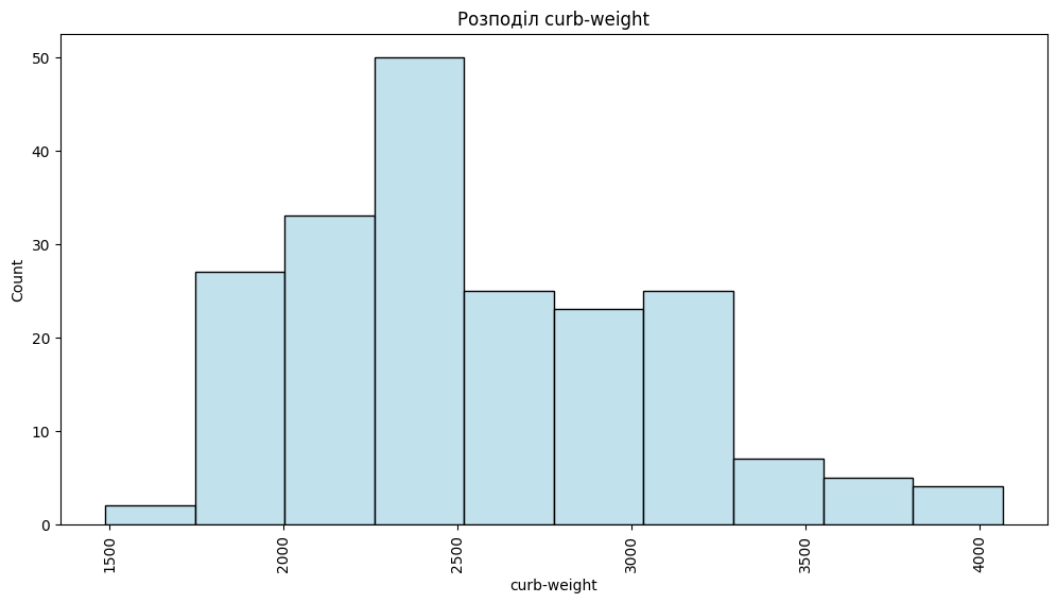
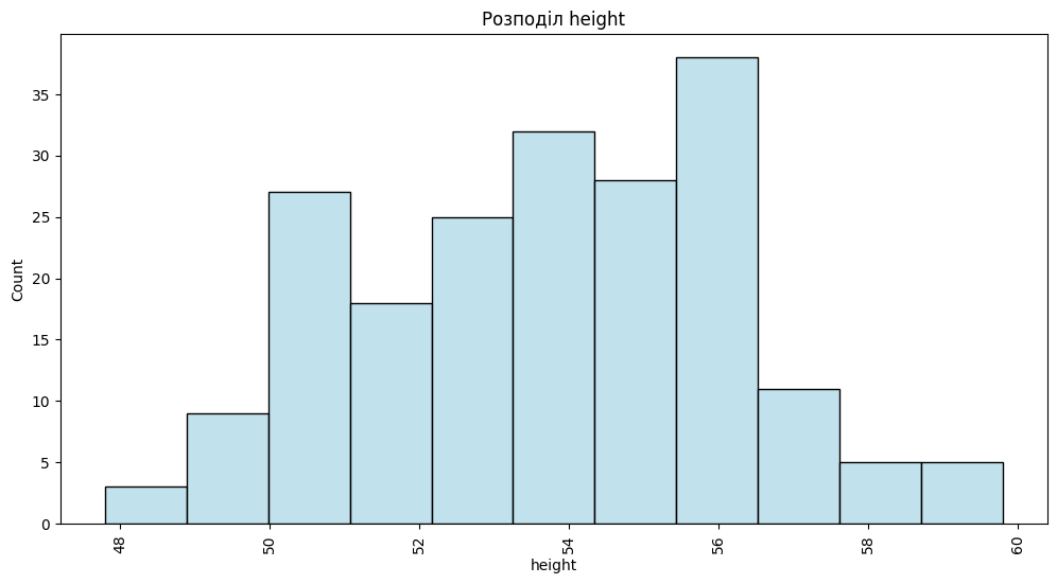
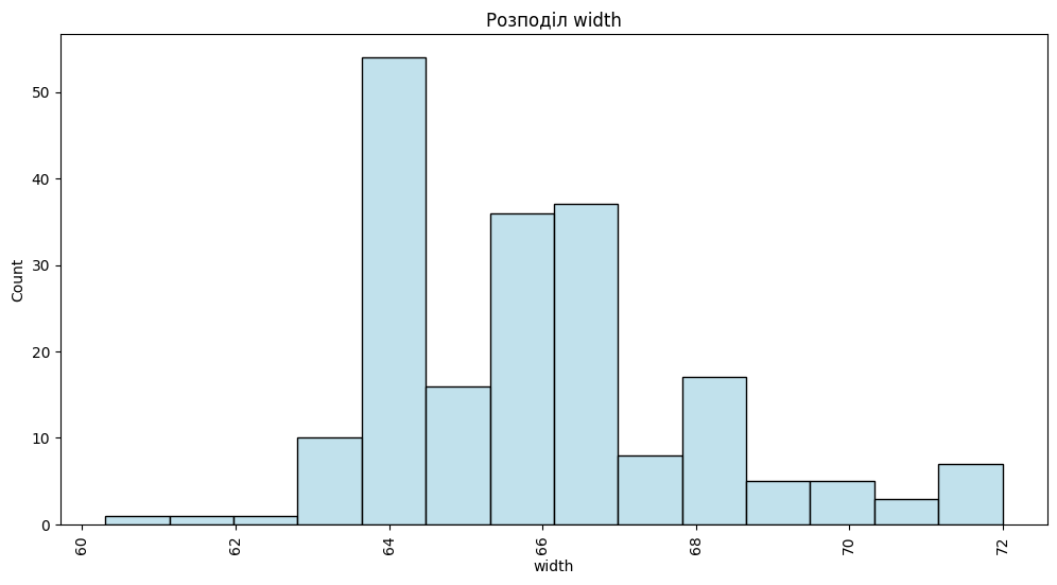
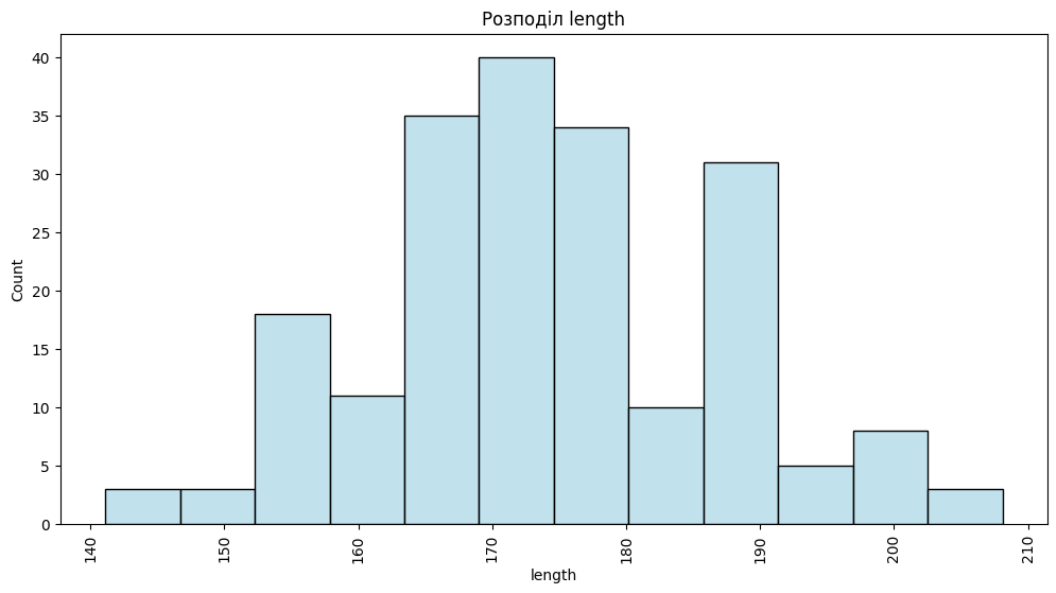
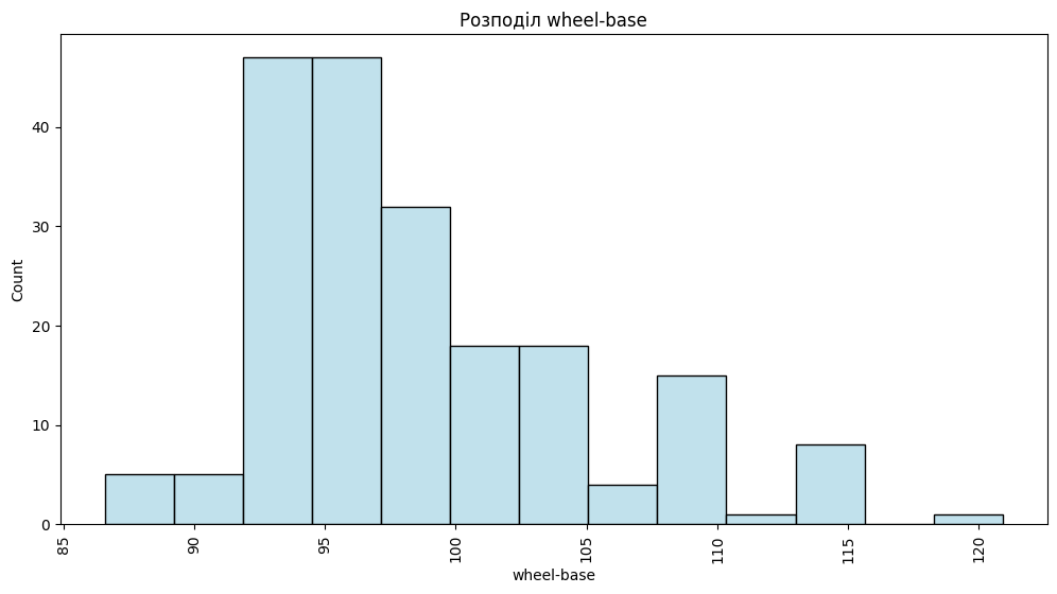
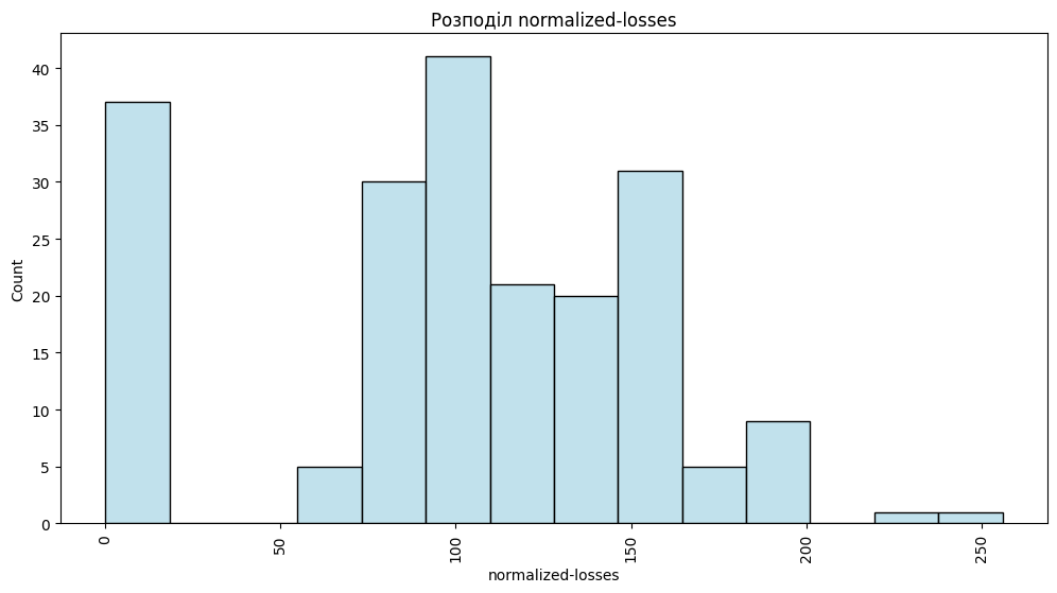
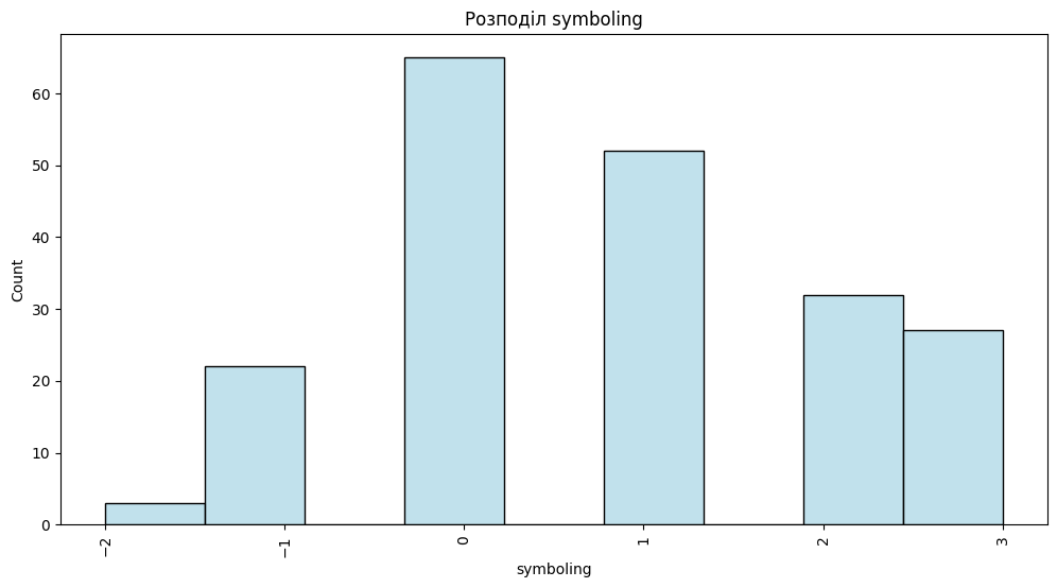
[**1.3** **Графіки для строкових значень** 6](#_Toc152922192)

[**1.4** **Матриця розсіювання для строкових значень** 7](#_Toc152922193)

[**1.5** **Аналіз даних методами** 8](#_Toc152922194)

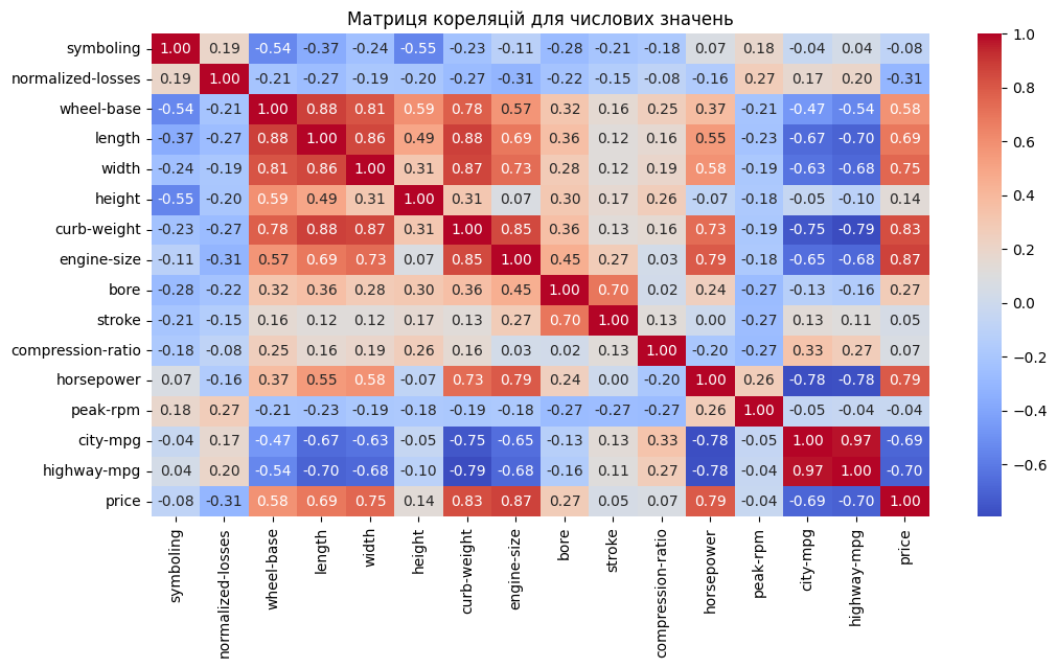
[**2.** **Висновок** 11](#_Toc152922195)

1. **Розробка процесів для запланованого функціоналу**
   1. **Графіки розподілу для числових значень**



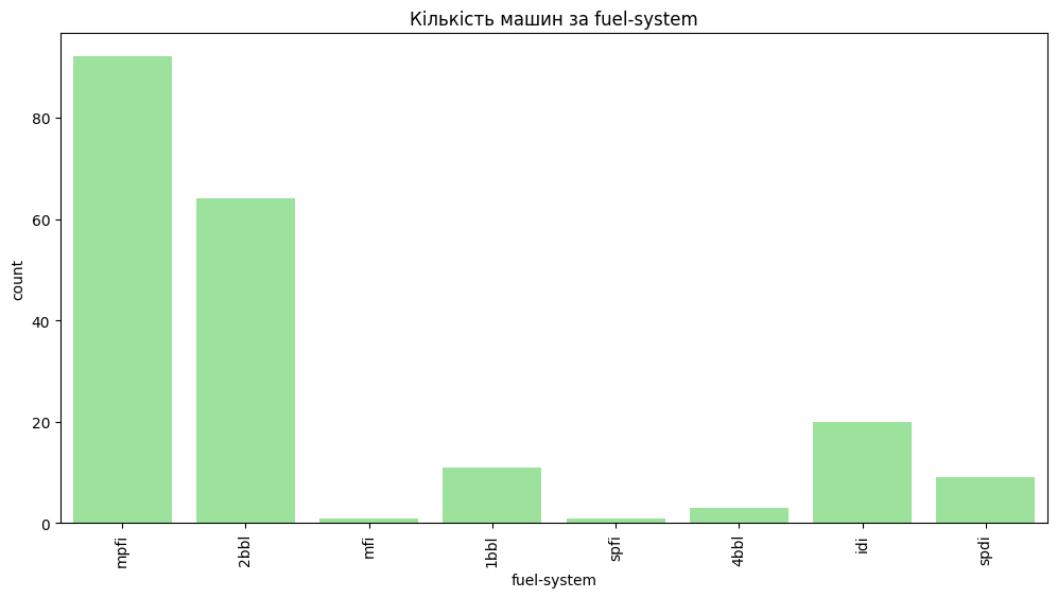
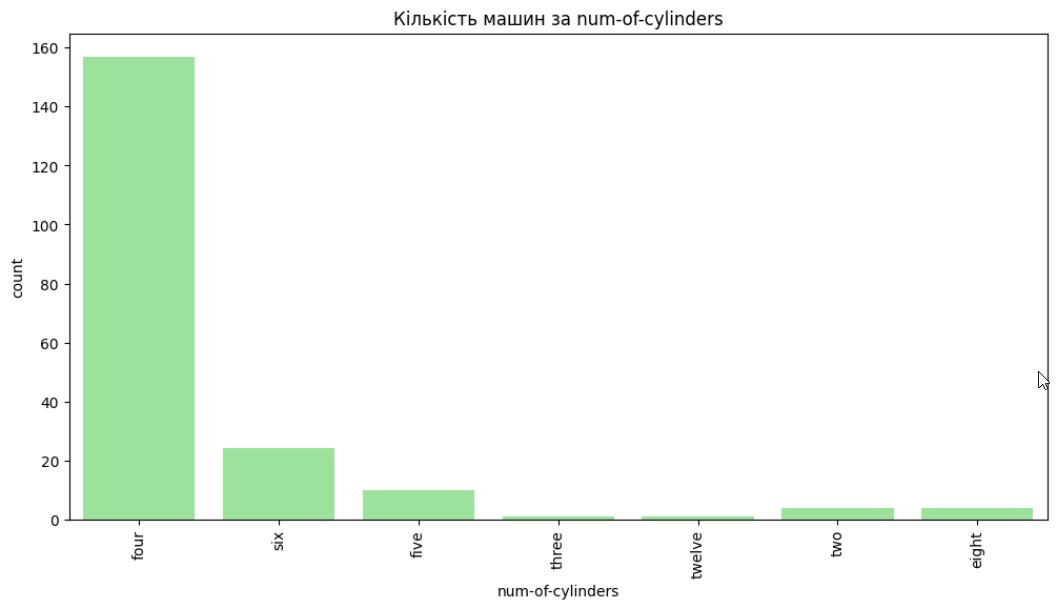
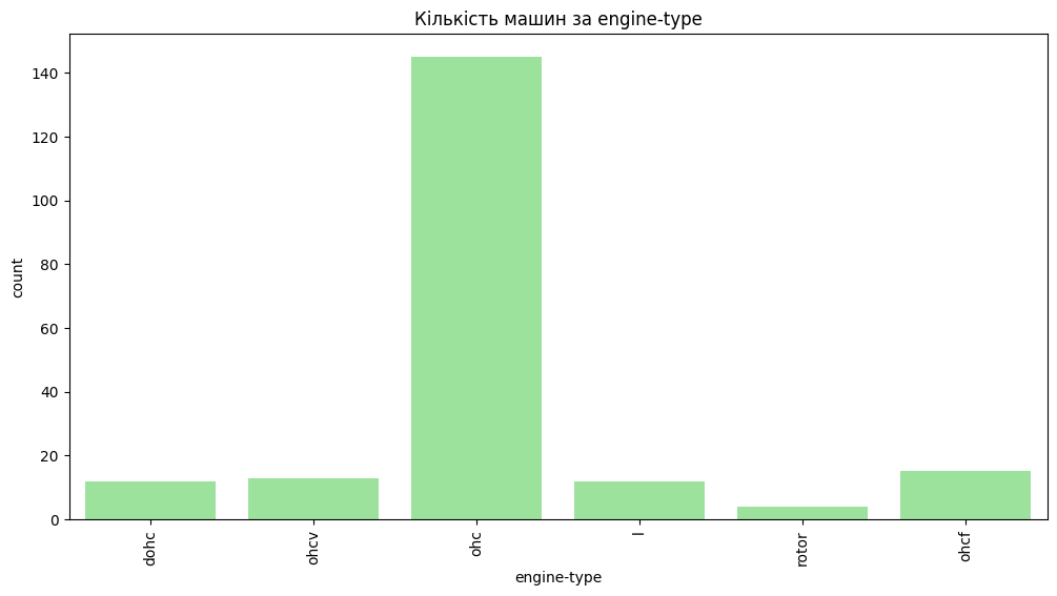
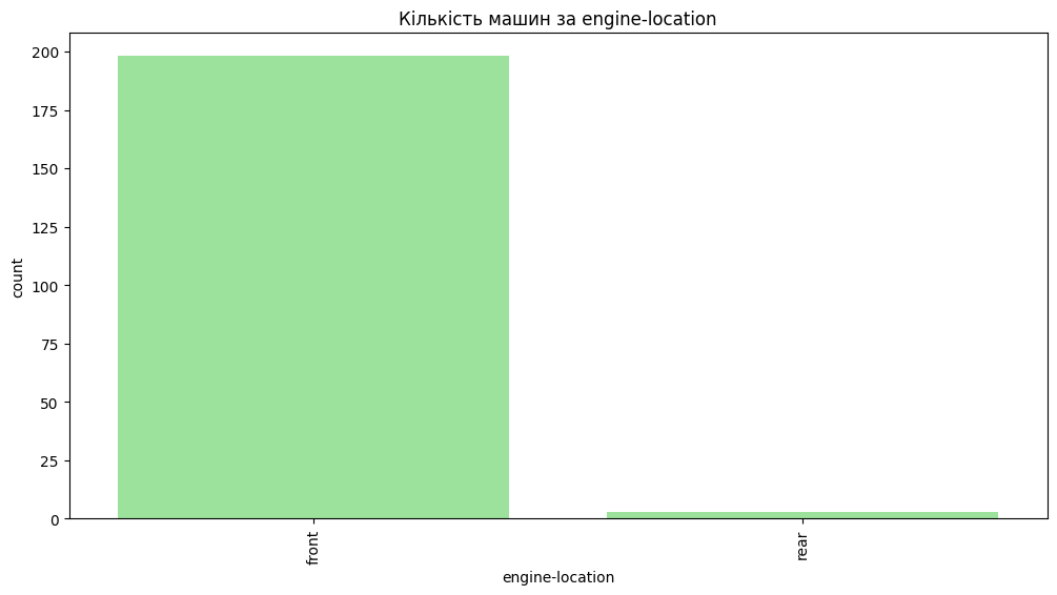
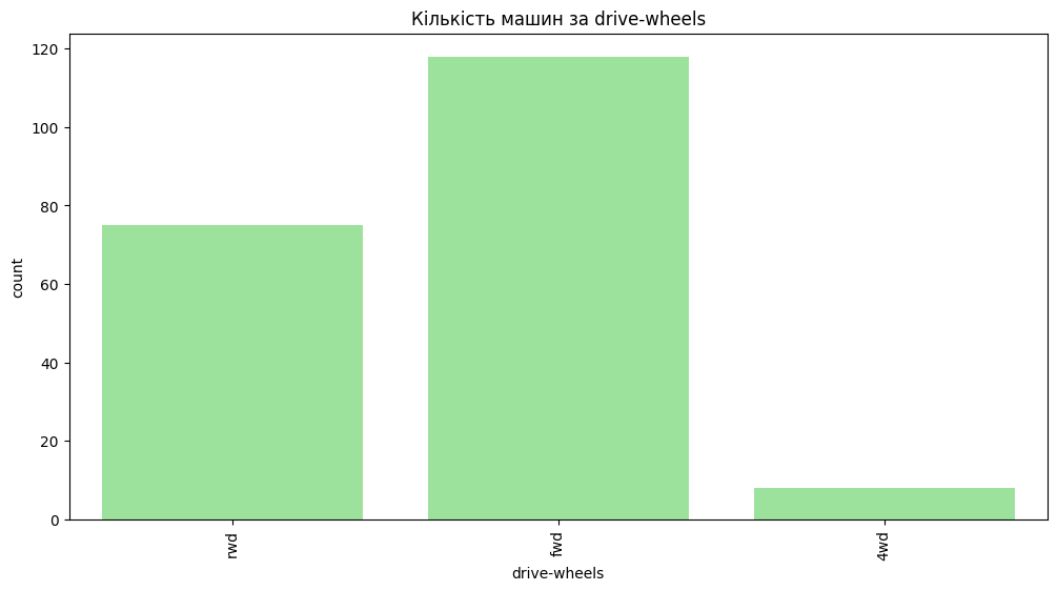
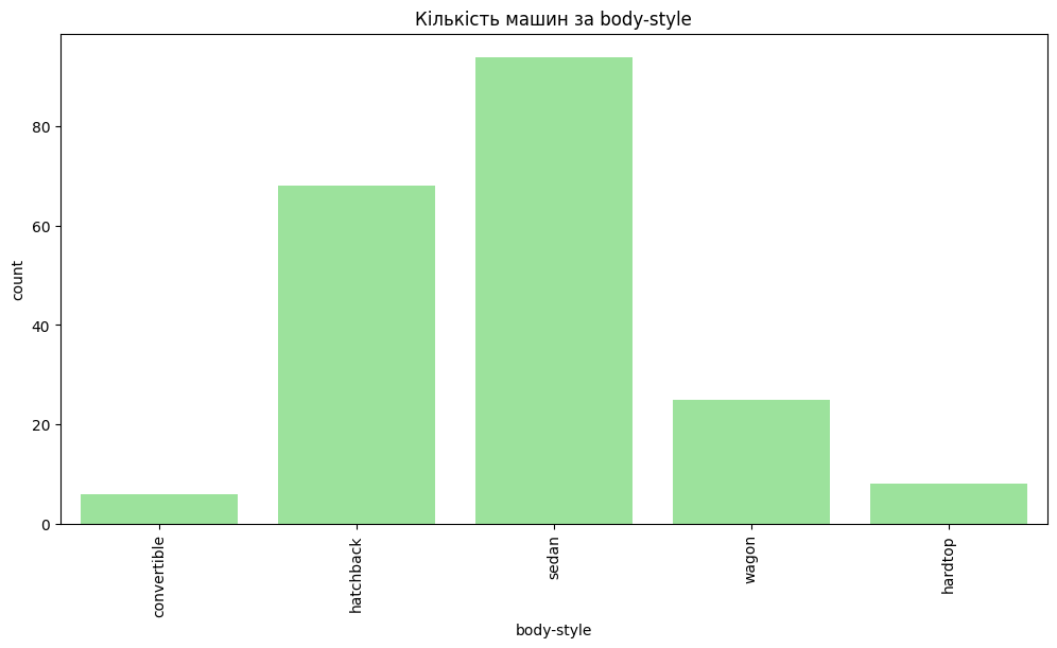
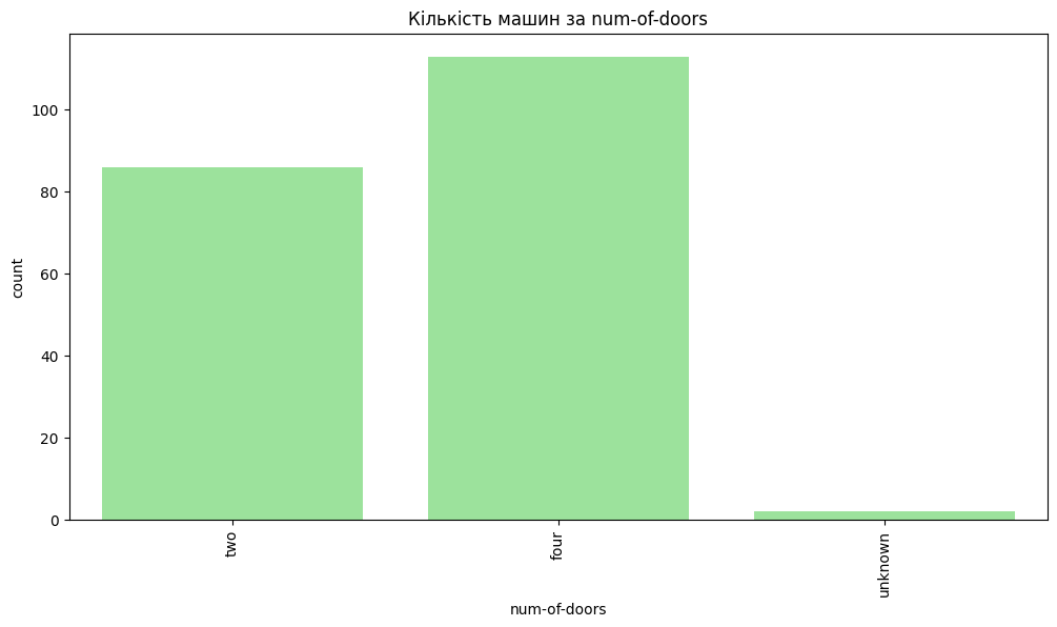
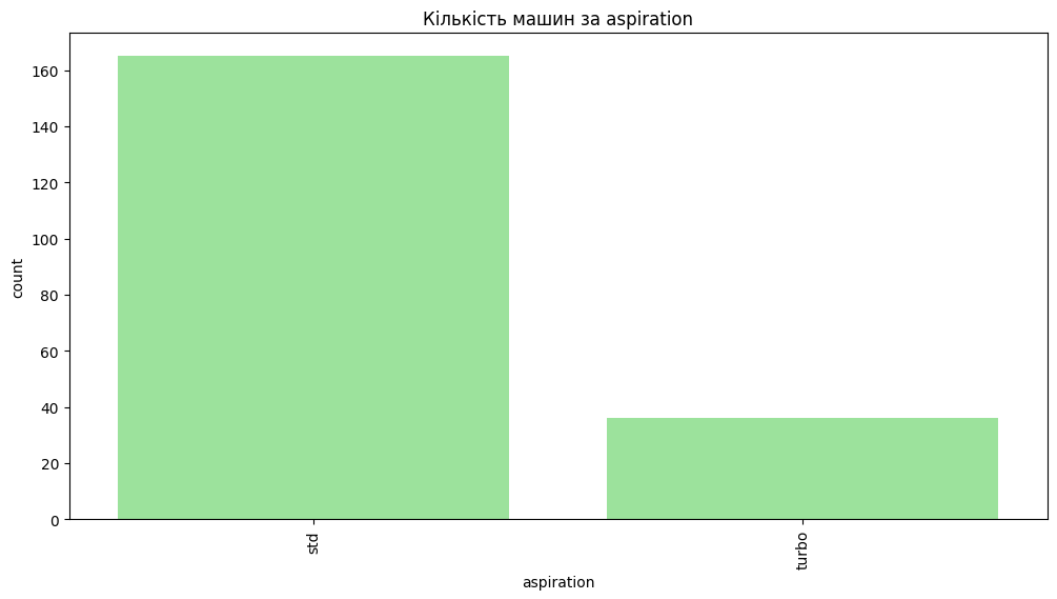
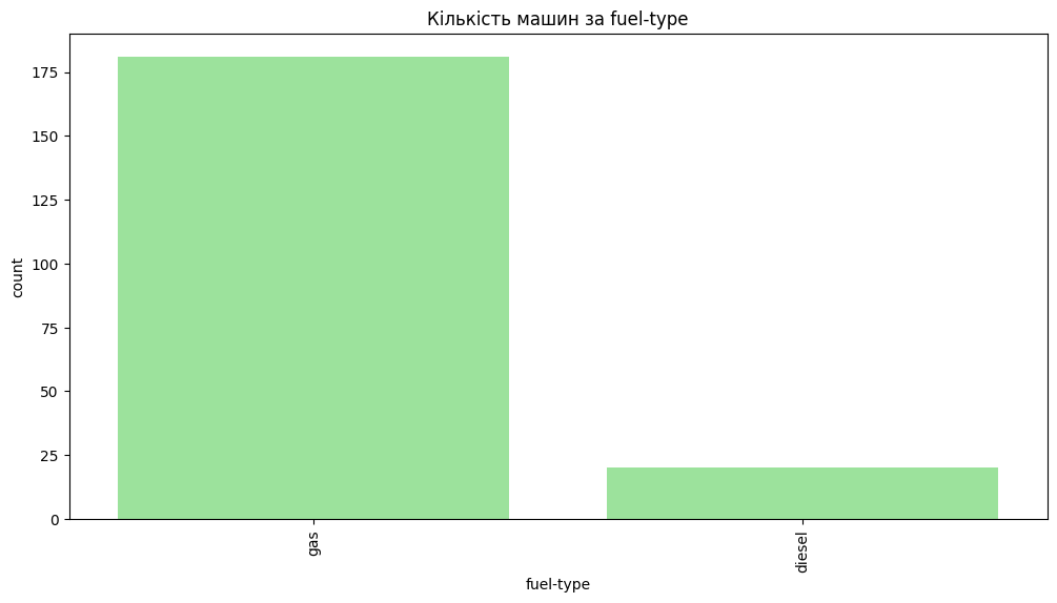
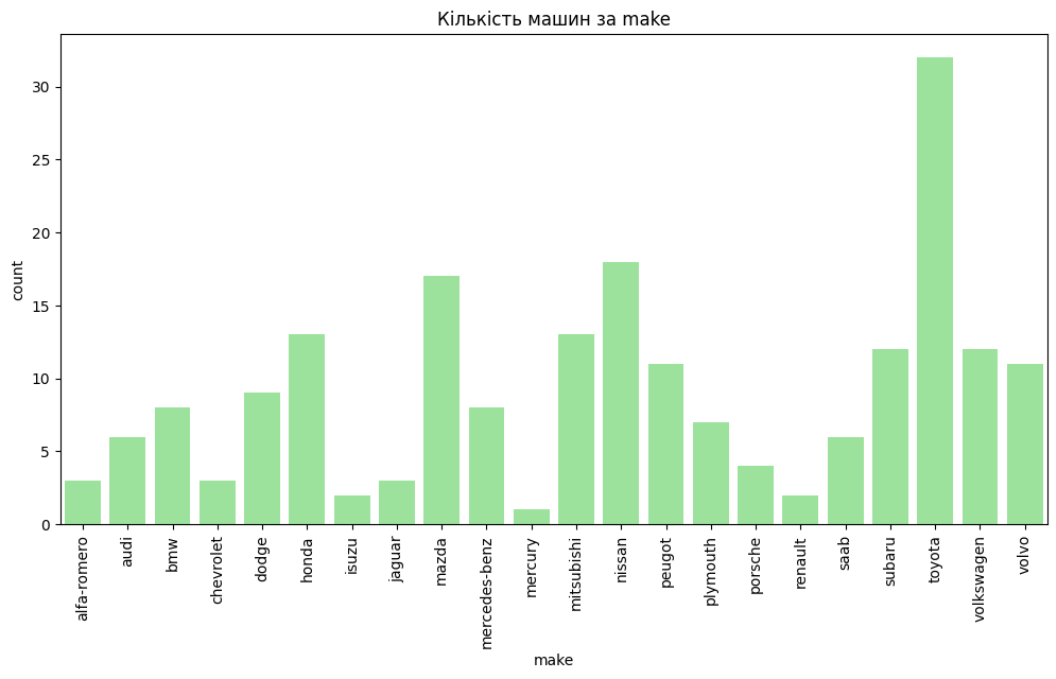
З наведених графіків вище можна побачити графіки розділу для числових колонок. Також можна побачити, вхідні дані мають нерівномірний розподіл значень. Розглянем, наприклад price: найбільша кількість цін – в межах від 5000-10000.

* 1. **Матриця кореляцій для числових значень**



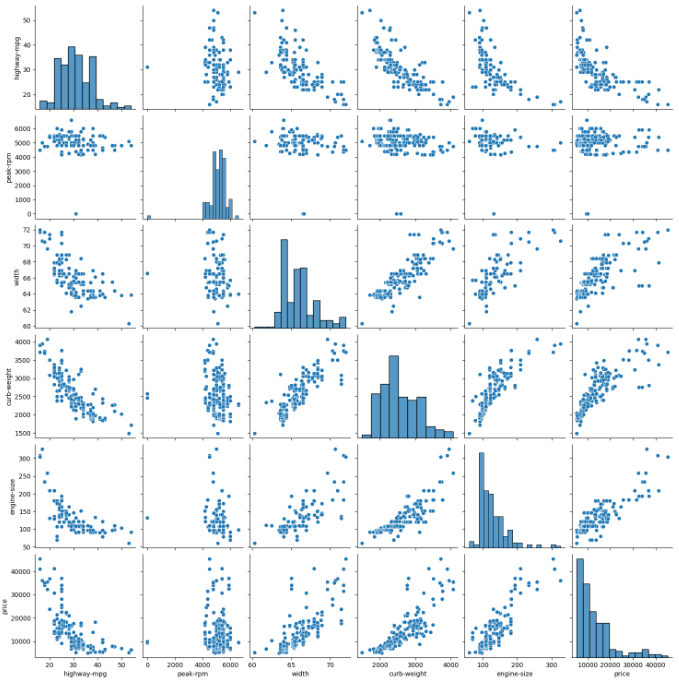
З наведених графіків вище можна побачити, як стовпці корелюють один з одним. Однак, опишемо лише кореляцію стовпця price з іншими. price дуже позитивно корелює з колонками wheel-base, length, width, curb-weight, engine-size, horsepower. Стовпець price дуже негативно корелює city-mpg i highway-mpg.

* 1. **Графіки для строкових значень**



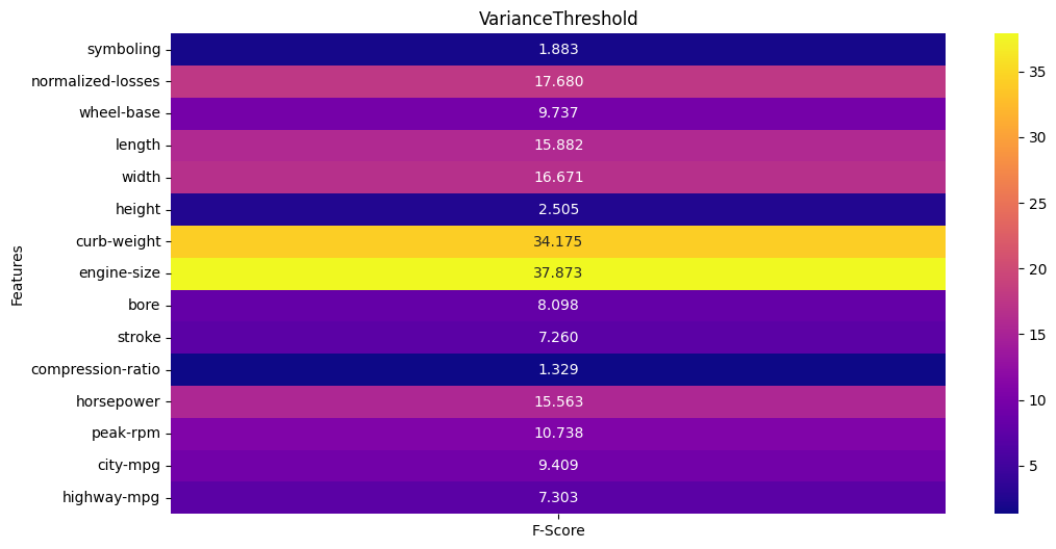
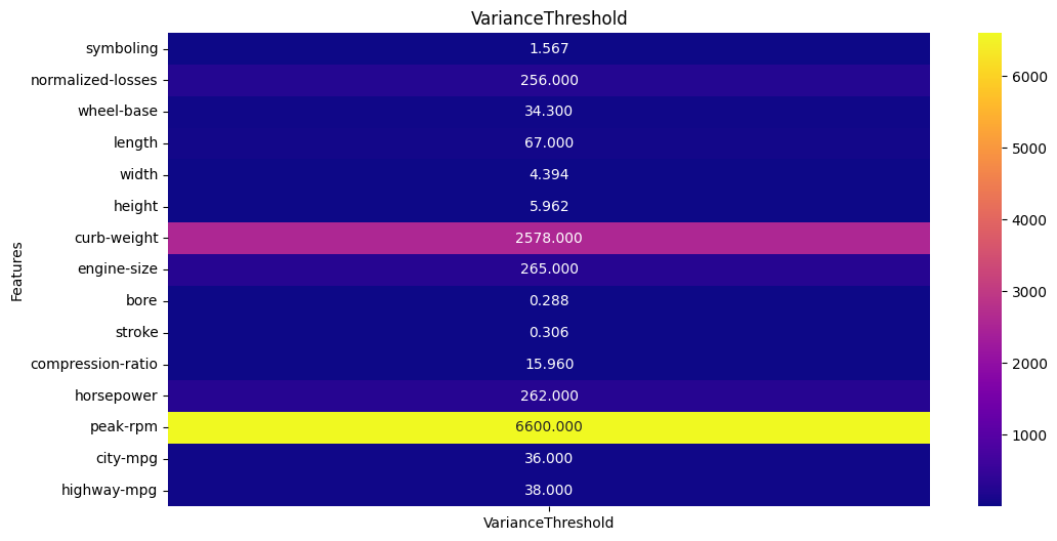
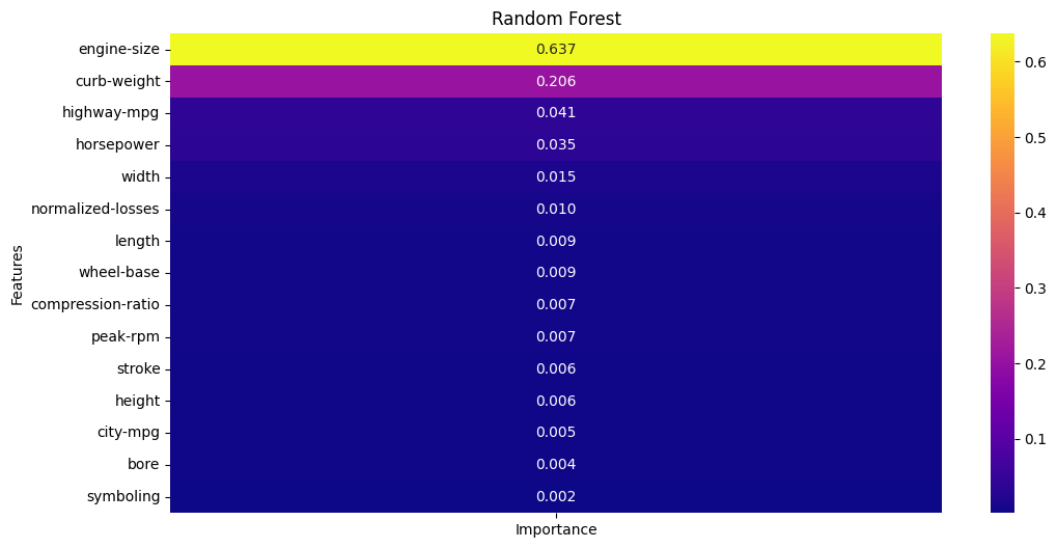
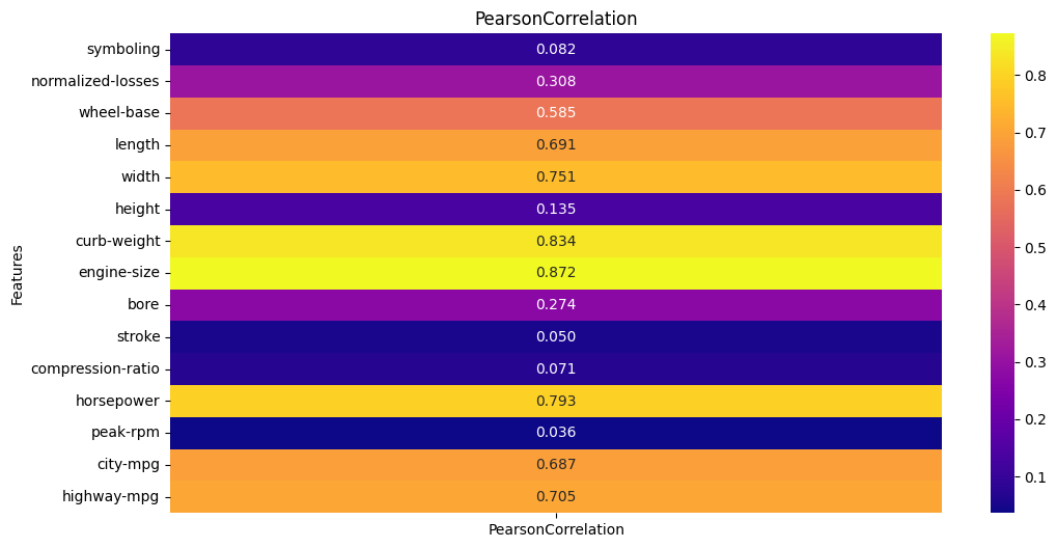
З наведених графіків вище можна побачити графіки, такі як в 1.1 підрозділі. Також можна побачити, вхідні дані мають нерівномірний розподіл значень.

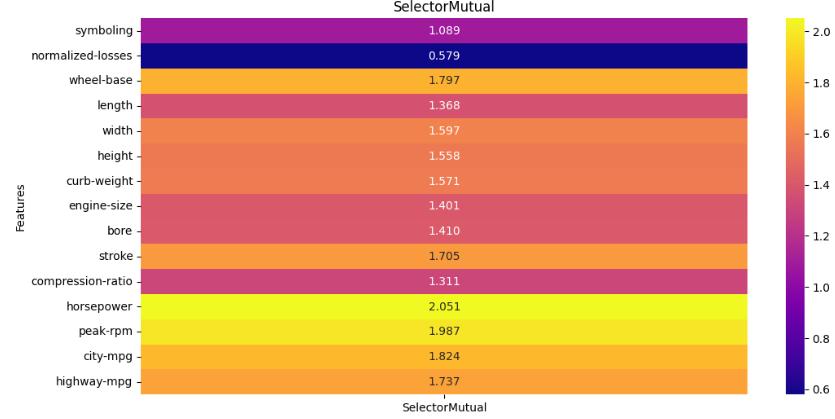
* 1. **Матриця розсіювання для строкових значень**



З наведених графіків вище можна побачити, як вигляд матриці кореляцій для деяких числових значень. Розглянемо, наприклад, price. Сам по собі price і normalized-losses має негативну кореляцію, тому на графіку можна побачити, ніби, спад. Для peak-rpm можна побачити, ніби, рівну пряму, оскільки вони мають дуже слабку низьку кореляцію. Для всіх інших елементів, кореляція висока, тому і маємо такі графіки.

* 1. **Аналіз даних методами**





З наведених графіків вище було проаналізовано дані наступними методами: Pearson Correlation, Random Forest, VarianceThreshold, F-Scope, Selector Mutual.

В наведених heatmap чим більший показник, тим важливіша ознака.

Наприклад, серед найважливіших ознаків є curb-weigh, який можна визначити за всіма методами. Далі по важливості йде engine-size, який можна визначити за 4/5 методами і так далі. Також можна побачити, що ознаки можуть мати різний ступінь важливості в залежності від методу відбору. Наприклад, peak-rpm має дуже високу важливість за методом Varience Threeshold, але у всіх інших він залишається з нижньою важливістю. Але є і такі ознаки, які при всіх методах залишалися з нижньою важливістю. Як приклад, symboling

1. **Висновок**

Під час виконання лабораторної роботи ознайомлено з методами аналізу даних. Для дослідження були взяті 5 методи, а саме Pearson Correlation, Random Forest, VarianceThreshold, F-Scope, Selector Mutual. Як результат, дослідження показало, що найбільш важливим показником виявився curb-weigh, який показав високі результати у всіх 5 методів. Однак, деякі дані, такі як, engine-size, width, length, horsepower змогли показати ефективність лише у двох методах із п’ти, хоча згідно з інформації з матриці кореляції стовпці мають високу кореляцію по відношенню до price. Причиною цього може бути те, що Random Forest - це широко використовуваний алгоритм машинного навчання, і неправильне налаштування, або ж через інші причини може призвести до непередбачуваних результатів або недооцінки його потенціалу. А щодо методу VarianceThreshold – то він передбачає, що висока дисперсія є показником корисності ознаки, що може не завжди відповідати реальним умовам даних.

Серед усіх використаних методів, на мою думку, найкращим виявився F-Scope.