**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**імені ТАРАСА ШЕВЧЕНКА**

****

**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра прикладних інформаційних систем**

**Звіт до лабораторної роботи №3**

**з курсу «Data Science»**

*студента 4 курсу*

*групи ПП-43*

*спеціальності 122 «Комп'ютерні науки»*

*ОП «Прикладне програмування»*

*Трофимчука Мирослава Ігоровича*

*Викладачі:*

*ас. проф. Білий Р. О.*

**Київ – 2023**

**Тема:** Методи аналізу та вибору значущих ознак.

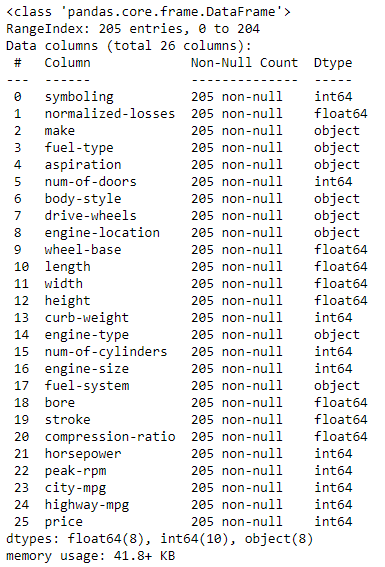
**Мета:** Метою лабораторної роботи є отримання практичних навичок аналізу та вибору значущих ознак для моделі за допомогою кореляційного аналізу, таблиць сопряжіння, аналізу багатомірні залежності та дихотомії, дисперсійного аналіз – ANOVA, критерій Хі-квадрат тощо.

***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.

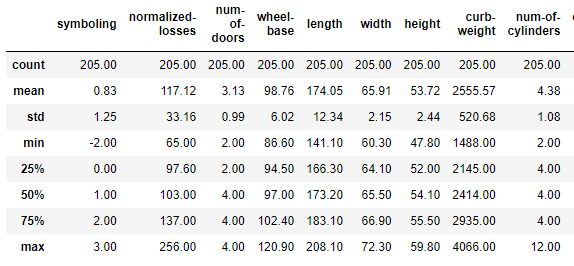
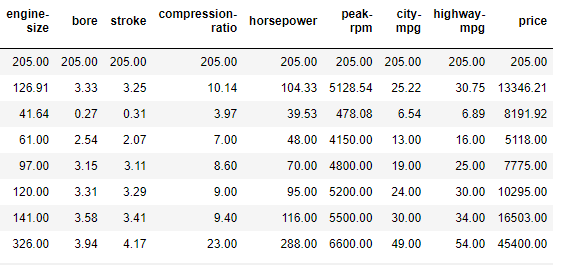
***Виконання***

1. Спочатку очистимо наші дані і проаналізуємо наш кінцевий датафрейм.



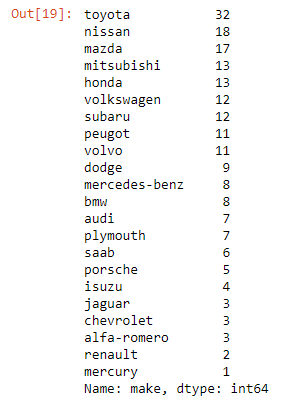
Можна побачити що всі наші дані розбиті на 26 стовпців і мають 205 рядків. Саме стільки було здійснено операцій продажу вживаних авто.

1. Проаналізуємо числові ознаки:



Можна побачити що мінімальна ціна вживаного авто 5118, а максимальна 45400. Середня 13350. Середня потужність двигуна 104 кінські сили. Прохідність миль/галон палива коливається від 13 до 49. Середнє значення 25.

1. Проведемо розподіл за брендами авто:



Можна побачити що найбільше продаються японські автомобілі:

Toyota, Nissan, Mazda, Mitshubishi, Honda, Subaru. Європейські автовиробники значно поступаються японським.

1. Проведемо розподіл за типом палива:

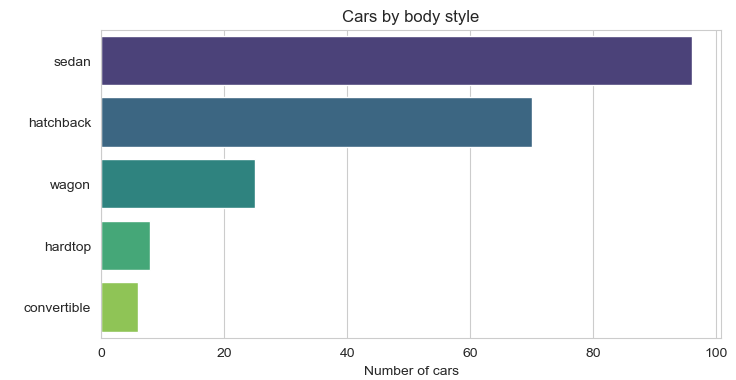


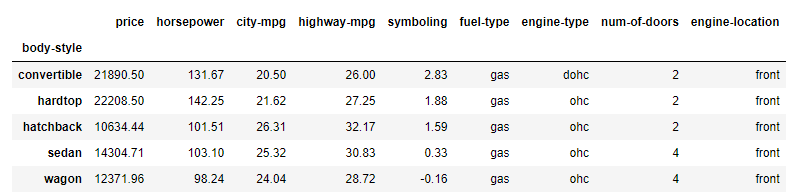
Можна побачиити що найпопулярнішим видом палива є бензин.

1. Проведемо розподіл за розташуванням двигуна:



Найпопулярнішим є переднє розташування двигуна.

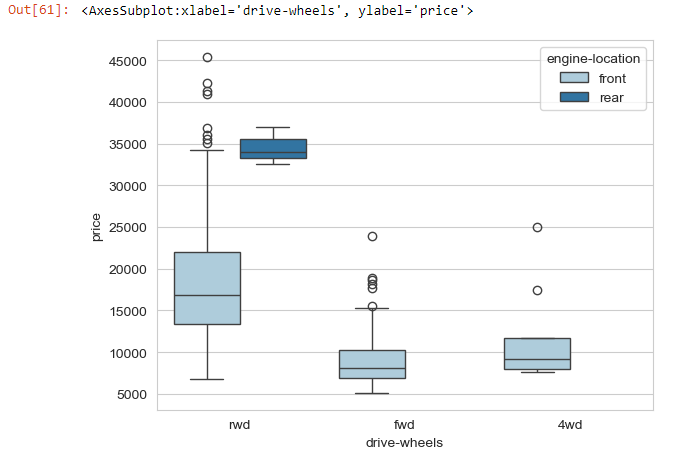
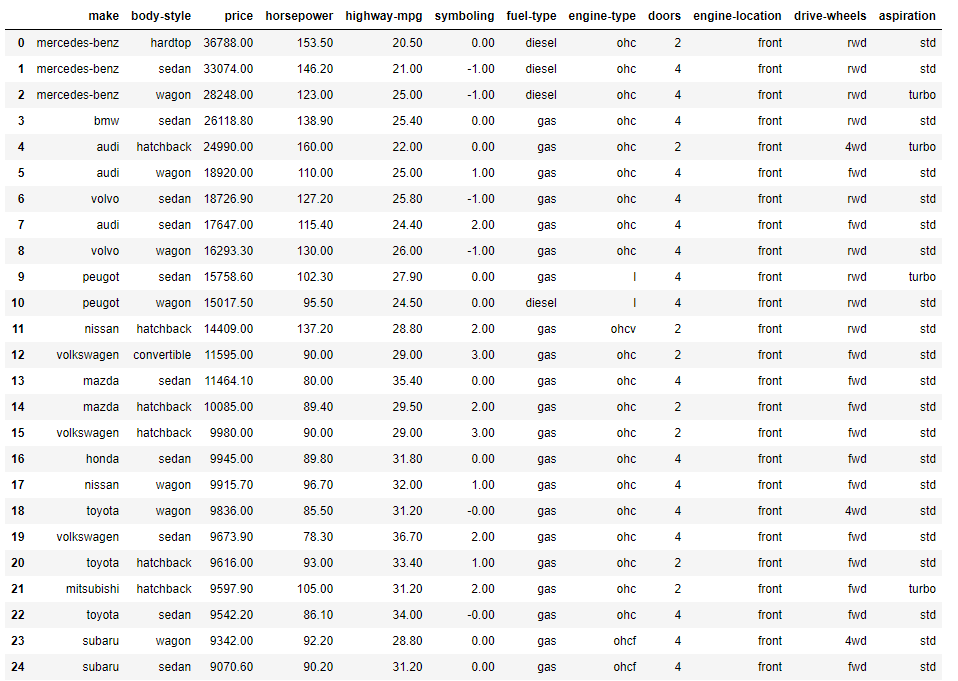
1. Тепер подивимося яку найпопулярнішу схему розміщення кузова мають вживані авто:

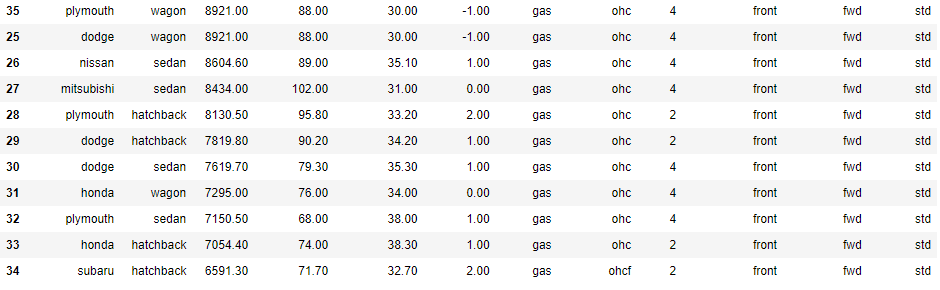
Найпопулярнішими є седани, їх близько 100. Тепер глянемо на середні показники для них:

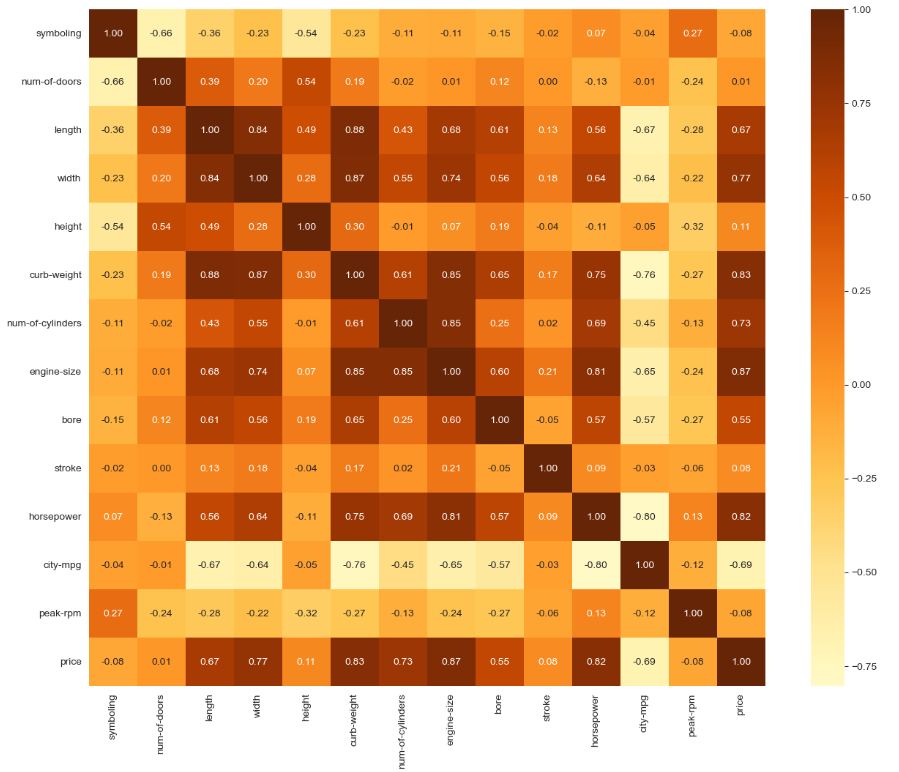
Найпотужніжими і найдорожчими авто є хардтопи. В них 2 дверей.

1. Глянемо на розподіл Price, Drive-Wheels, Engine-Location.

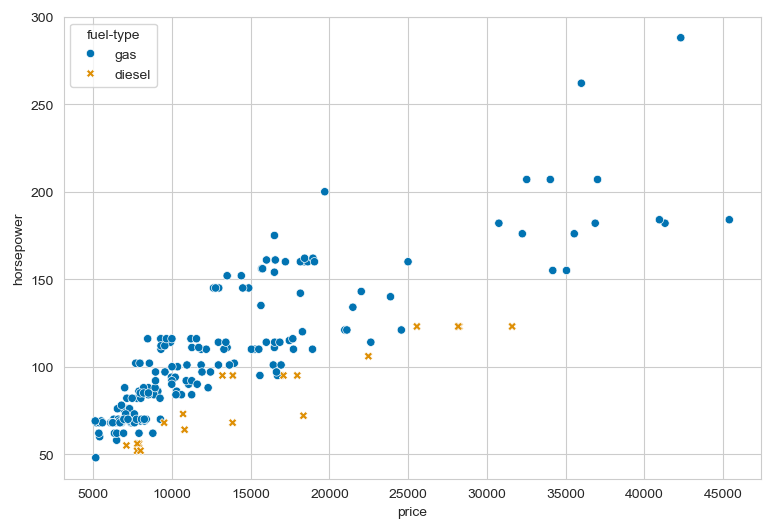
Можна побачити що авто з заднім розташуванням двигуна і заднім приводом є найдрожчими авто і їх найменше. Найдешевшими є авто з переднім розташуванням двигуна і привода.

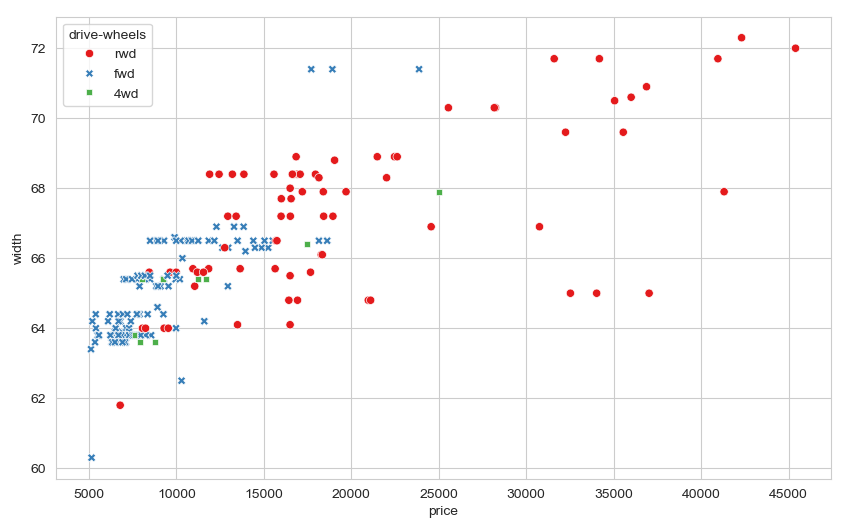
1. Проаналізуємо топ-14 брендів авто і топ-3 схем компонування:

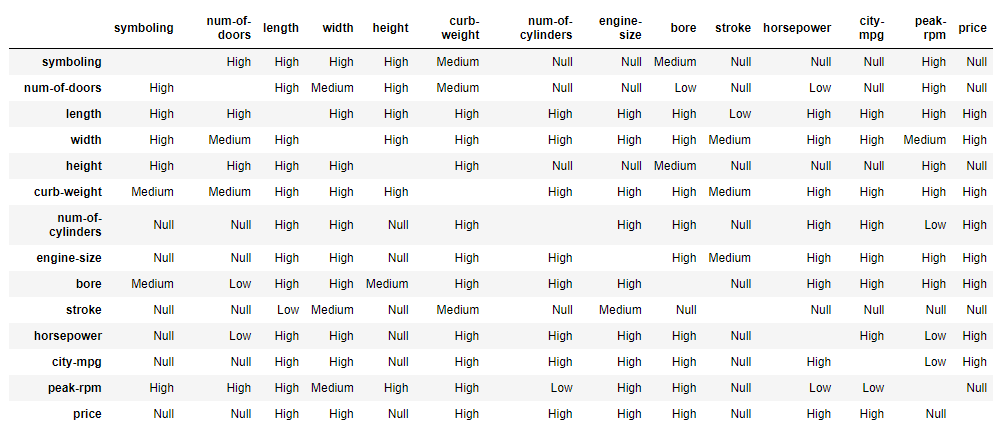
Можна побачити що найдорожчими авто є мерседеси всіх типів, вони мають низький ризик, найпопулярнішим паливом для них є дизель. Найдешевшими авто є японські honda, subaru та американські dodge.

1. Візуалізуємо зв’язки між ключевими ознаками:

Чим темніша область тим сильніший зв’язок.

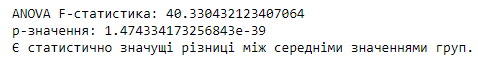
Можна побачити між ціною продажу і потужністю двигуна є сильний зв’язок. Діаграма розподілу це підтверджує.

Також можна побачити, що є зв’язок між розмірами авто та його вартістю.

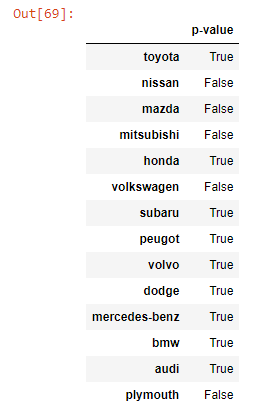
1. Проведемо кореляцію Пірсона для найважливіших числових ознак:

***High –*** означає значний зв’язок між ознаками, ***Medium –*** середній,  ***Low –*** низький, ***Null –*** зв’язок відсутній.

1. Проведемо ANOVA-test для груп авто та їх цін.



Для детальнішого ознайомлення можна провести Хі-квадрат тест:

p-value дорівнює True, якщо воно < 0.05

Якщо p-value менше визначеного рівня значущості (наприклад 0.05), то можна вважати, що існують статистично значущі різниці між середніми значеннями груп.