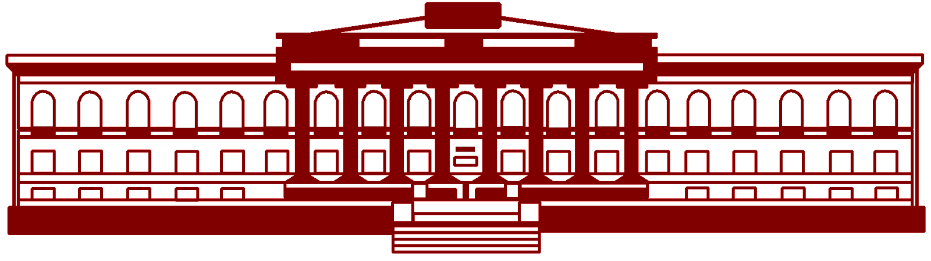
**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**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

Спеціальності 122 «Комп'ютерні науки»

ОП «Прикладне програмування»

Цимбала Дмитра

Викладач:

Білий Р.О.

**Київ – 2022**

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

**1. 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.

**2. 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.

### Джерела:

### Технічні характеристики імпортних автомобілів і вантажівок моделі 1985 року, автомобільний щорічник Уорда за 1985 рік.

### Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038

### Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037.

### Цей набір даних складається з трьох типів об’єктів: (а) специфікація автомобіля з точки зору різних характеристик, (б) присвоєний йому рейтинг страхового ризику, (в) його нормалізовані втрати під час використання порівняно з іншими автомобілями. Другий рейтинг відповідає ступеню ризику автомобіля, ніж вказує його ціна. Автомобілям спочатку присвоюється символ фактора ризику, пов'язаний з його ціною. Потім, якщо це більш ризиковано (або менше), цей символ коригується шляхом переміщення його вгору (або вниз) за шкалою. Актуарії називають цей процес «символізація». Значення +3 вказує на те, що авто є ризикованим, -3, що воно, ймовірно, досить безпечне.

### Третім фактором є відносна середня виплата збитку за рік страхування автомобіля. Це значення нормалізовано для всіх автомобілів певної класифікації розміру (дводверні маленькі, універсали, спортивні/спеціальні тощо) і являє собою середні втрати на автомобіль на рік.

### Примітка. Кілька атрибутів у базі даних можна використовувати як атрибут «класу».

### 3. Завдання до виконання

### 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:**

## Алгоритм відбору ознак

## Алгоритм ідентифікував наступні ключові ознаки: витрати палива на шосе, кількість кінських сил, хід двигуна та тип двигуна, впорядковані від найважливіших до менш важливих. Проведемо загальний аналіз цих ознак, розпочавши з числових показників.

## Середнє значення витрат палива на шосе становить приблизно 30, кількість кінських сил - близько 95, а хід двигуна - близько 3.

## 

## Медіана витрат палива на шосе дорівнює близько 30, кількості кінських сил - приблизно 85, а ходу двигуна - близько 3.

## 

## Максимальне значення витрат палива на шосе становить близько 50, кількість кінських сил - близько 200, а хід двигуна - близько 3.

## 

## Мінімальне значення витрат палива на шосе складає приблизно 18, кількості кінських сил - близько 48, а ходу двигуна - близько 3.

## 

## Тепер перейдемо до аналізу категорійних ознак. Ми можемо спостерігати, що найбільш популярним типом двигуна є "ohc", при цьому всі інші виявляють значний відступ.

## 

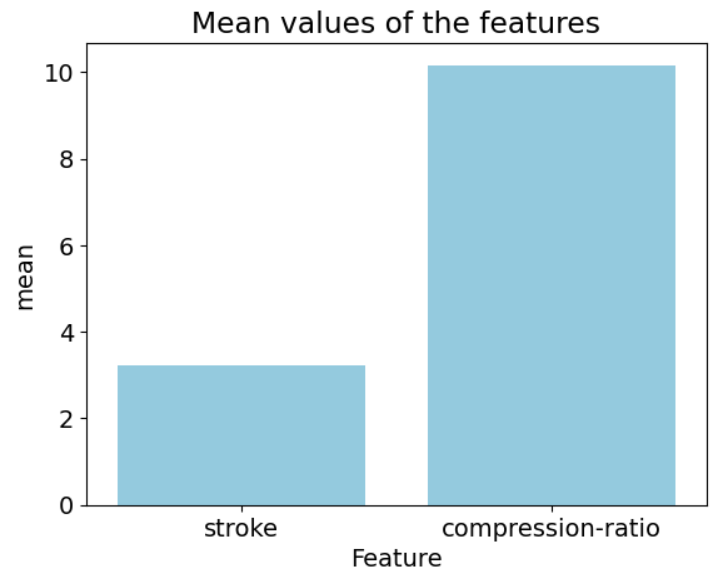
## Далі проаналізуємо графік кореляцій. Тут присутні як позитивні, так і негативні коефіцієнти, але більшість значень не перевищують ±0.2. Однак кореляція між кількістю кінських сил та витратами палива на шосе складає -0.83, що є надзвичайно великою величиною. Також важлива кореляція між типом двигуна та ходом двигуна, яка становить -0.26.

## 

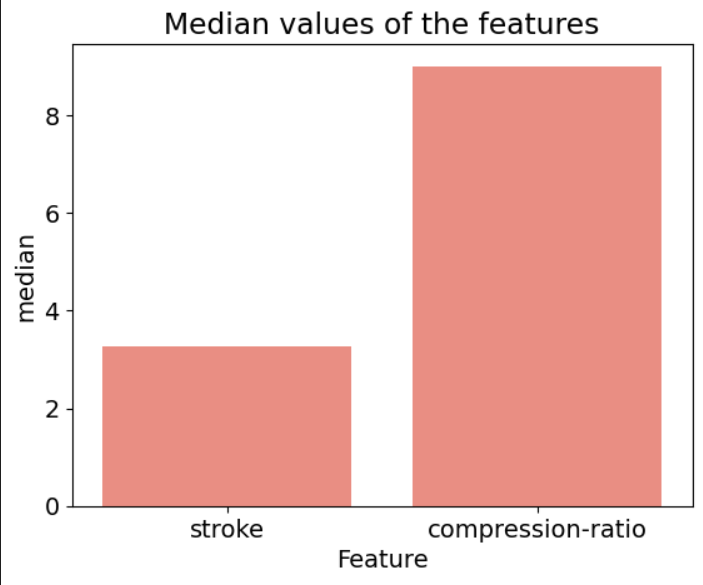
1. **Рекурсивний алгоритм**

Алгоритм виділив такі ключові ознаки як хід двигуна, ступінь стиснення, кількість циліндрів та тип подачі повітря у двигун, ранжуючи їх від найважливіших до менш важливих. Давайте проведемо загальний аналіз цих ознак, починаючи з числових.

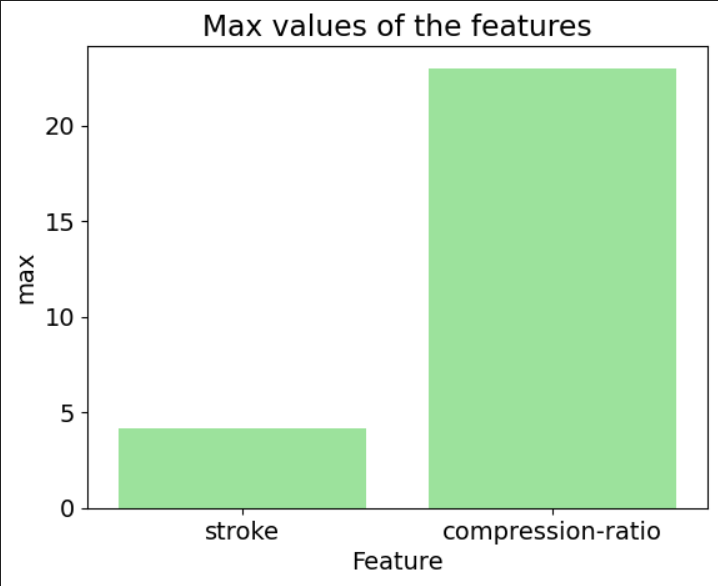
Середнє значення ступеню стистення дорівнює близько 10, а ходу двигуна близько 4.



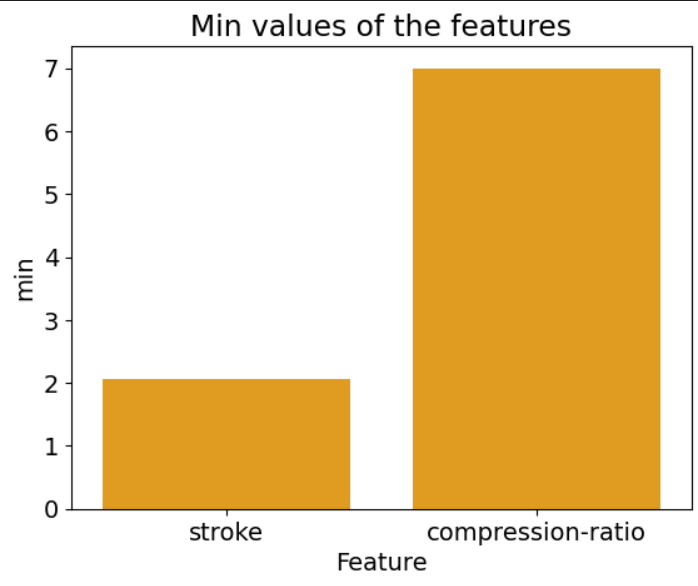
Середнє значення ступеня стиснення приблизно 10, а ходу двигуна - близько 4.



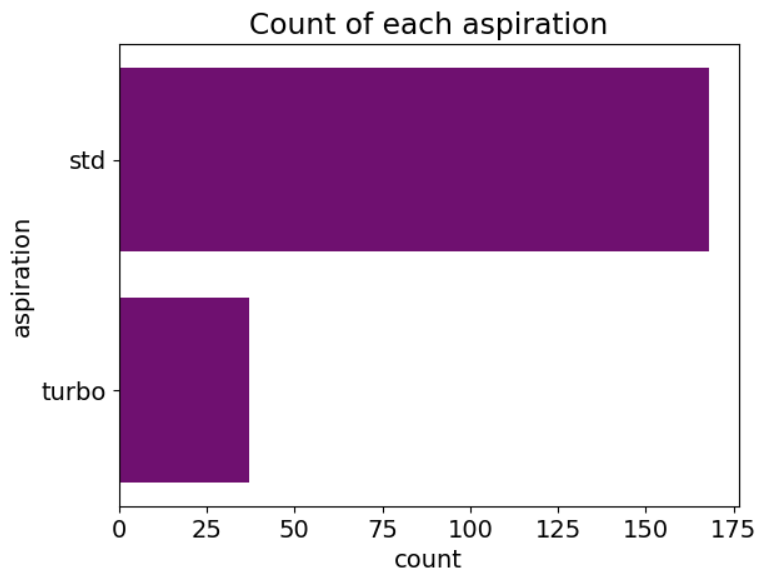
Максимальне значення ступеня стиснення досягає приблизно 23, тоді як хід двигуна становить близько 4.



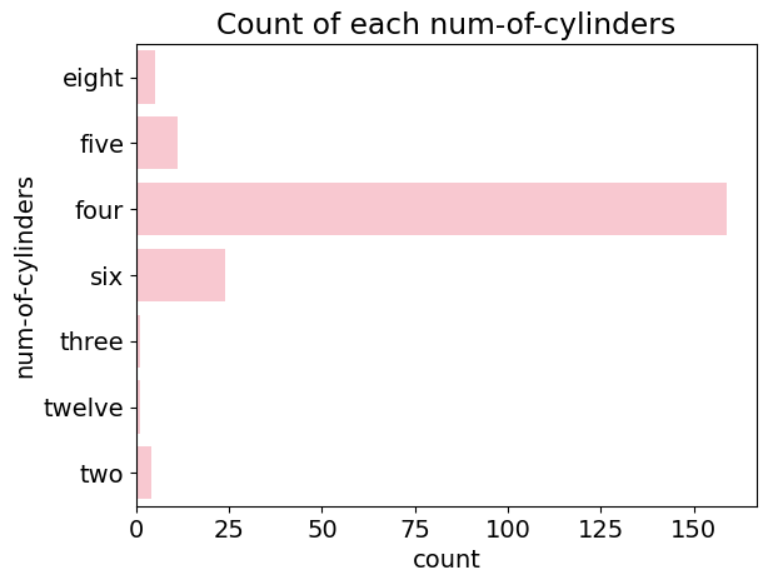
Мінімальне значення ступеня стиснення приблизно 7, а хід двигуна - близько 2.



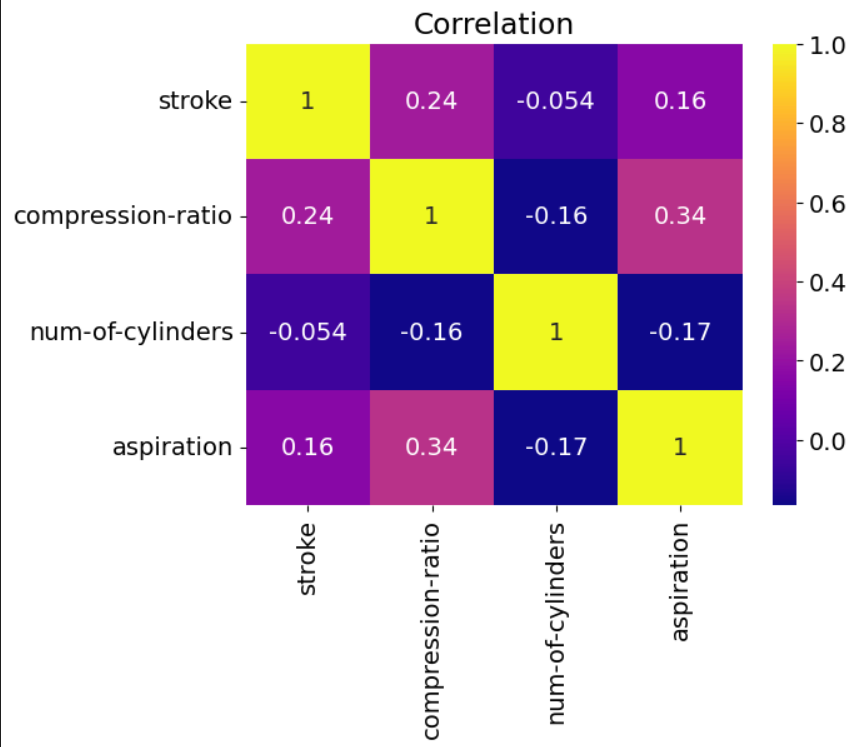
Перейдемо до аналізу категорійних ознак. З графіку видно, що найбільш популярною системою подачі повітря є "std", тоді як "турбо" суттєво поступається.



Щодо кількості циліндрів, 4 циліндри відзначаються значною перевагою, в той час як інші типи відчутно відстають і не користуються такою популярністю.



Далі розглянемо графік кореляцій. Виявлено як позитивні, так і негативні коефіцієнти, проте більшість значень не перевищують ±0.2. Зокрема, кореляція між системою подачі повітря та ступенем стиснення становить 0.34, а між ступенем стиснення та ходом двигуна - 0.24. Загалом, кореляція не має значної ваги.



## Однофакторний вибір ознак

## Алгоритм ідентифікував наступні ключові ознаки: витрати бензину на шосе, тип двигуна, пікові оберти та система заправки, відзначаючи їх важливість у порядку спадання. Давайте проведемо загальний аналіз цих ознак, розпочавши з числових значень.

## Середні оберти становлять приблизно 5000 обертів. В середньому витрати бензину складають 32 одиниці.

## 

## Середнє значення оборотів приблизно 5000 обертів за хвилину, а витрати бензину становлять 32 одиниці.

## 

## Максимальні оберти становлять близько 6300 обертів за хвилину, а витрати бензину досягають 54 одиниць.

## 

## Мінімальні оберти складають приблизно 4150 обертів за хвилину, а витрати бензину становлять 18 одиниць.

## 

## Оглянемо категорії строкових ознак. З графіка видно, що системи заправки "mpfi" та "2bbl" користуються найбільшою популярністю, у той час як інші мають невелику кількість.

## 

## На іншому графіку можна відзначити високу популярність типу двигуна "ohc", який перевищує 140 одиниць, в той час як інші типи не досягають навіть 20 одиниць.

## 

## Проведемо аналіз кореляції. Єдині показники з позитивною кореляцією — це системи заправки і пікові оберти, і ця кореляція становить 0.035. Усі інші мають негативні кореляції, хоча вони невеликі (більше -0.1 та менше 0), за винятком відносно великої кореляції між системою заправки та витратами палива на шосе, яка становить -0.63.

## 

**Висновок:** За результатами аналізу виявлено, що при визначенні того, які автомобілі продавати, важливо звертати увагу на ряд характеристик. У першому випадку, це витрати палива на шосе, кінські сили, хід двигуна та тип двигуна, зокрема ohc. Виділені значення показників становлять, відповідно, приблизно 30, 95, 3 та -0.83. У другому випадку, це кількість циліндрів, система постачання повітря, ступінь стиснення та хід двигуна. Відзначені середні значення для цих параметрів дорівнюють приблизно 10, 4, 4 та кореляція між системою постачання повітря та ступенем стиснення 0.34. У третьому випадку, важливо враховувати витрати бензину на шосе, тип двигуна, пікові обороти та систему заправки, зокрема 2bbl та mpfi. Відзначені середні значення становлять, відповідно, приблизно 32, 5000, а кореляція між системою заправки та витратами бензину -0.63.