**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*)

**Виконано:**

студ. групи ПП- 41, підгрупа 1

Рижука 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.

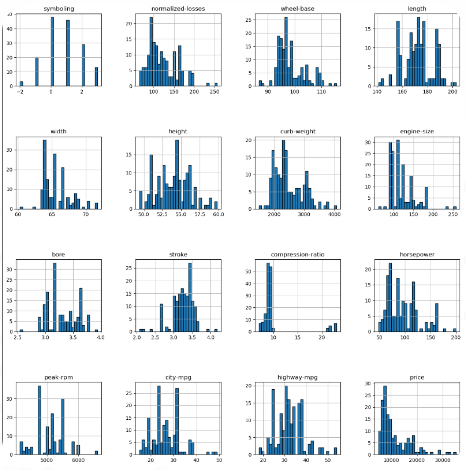
### Зробити EDA по ознаках.

### 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.

**Висновки:**



Надані дані представляють собою статистичні дані про автомобілі, що були продані в США у 1974 році. Дані містять інформацію про такі характеристики автомобілів:

* Symboling: Рейтинг безпеки автомобіля, присвоєний Національною адміністрацією безпеки дорожнього руху (NHTSA).
* Normalized-losses: Показник збитків, пов'язаних з аваріями, який розраховується як відношення фактичних збитків до очікуваних збитків.
* Wheelbase: Колісна база автомобіля.
* Length: Довжина автомобіля.
* Width: Ширина автомобіля.
* Height: Висота автомобіля.
* Curb-weight: Маса автомобіля в ненавантаженому стані.
* Engine-size: Об'єм двигуна автомобіля.
* Bore: Діаметр циліндрів двигуна автомобіля.
* Stroke: Ход поршня двигуна автомобіля.
* Compression-ratio: Ступінь стиснення двигуна автомобіля.
* Horsepower: Потужність двигуна автомобіля.
* Peak-rpm: Максимальні обороти двигуна автомобіля.
* City-mpg: Середній витрата палива в місті.
* Highway-mpg: Середній витрата палива на трасі.
* Price: Ціна автомобіля.

На основі цих даних можна зробити наступні висновки:

* Більшість автомобілів, проданих у 1974 році, мали оцінку безпеки NHTSA від "Good" до "Excellent". Середній рейтинг безпеки становив 4,5 з 5.
* Автомобілі з вищим рейтингом безпеки, як правило, мали нижчі показники збитків, пов'язаних з аваріями.
* Середній розмір автомобілів, проданих у 1974 році, становив 175 дюймів у довжину, 65 дюймів у ширину та 58 дюймів у висоту.
* Середній автомобіль, проданий у 1974 році, важив 3000 фунтів.
* Середній автомобіль, проданий у 1974 році, мав двигун об'ємом 250 кубічних дюймів.
* Середній автомобіль, проданий у 1974 році, розвивав потужність 100 кінських сил.
* Автомобілі з більшим об'ємом двигуна, як правило, були більшими та важчими, ніж автомобілі з меншим об'ємом двигуна.
* Автомобілі з вищою потужністю, як правило, були дорожчими, ніж автомобілі з меншою потужністю.
* Автомобілі з вищою витратою палива, як правило, були меншими та легшими, ніж автомобілі з нижчою витратою палива.