**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-43*

*спеціальності 122 «Комп'ютерні науки»*

*ОП «Прикладне програмування»*

*Чернецький Р.В*

*Викладач:*

*Шолохов О.В*

*Білий Р.О*

**Київ – 2022**

**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.

**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.

**Хід роботи**

Провівши дослідження було виявлено найбільшу та найменшу кількість автомобілів по кожному з категоріальних критерії, приклад частини з них:

* Марка: найбільше авто з марки Toyota, найменше авто марки Porsche
* Тип палива: найбільше бензинових авто, найменше – дизельних
* По типу кузова маємо: найбільше седанів, найменшк – кабріолетів

A screenshot of a computer

Description automatically generatedМаємо наступні графіки:

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

З графіків розподілу можемо бачити що вхідні дані мають доволі нерівномірний розподіл значень. Розподіл дозволяє отримати уявлення про те, як різні значення розподілені у межах цієї ознаки. Цей аналіз є корисним для розуміння характеристик ознак та їхнього впливу на інші параметри в датасеті.

A graph with a line

Description automatically generatedA graph with lines and numbers

Description automatically generated

A graph of a graph

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated with medium confidence

Під час дослідження ми згенерували теплову карту, що візуалізує матрицю кореляції між числовими ознаками датасету. Кожен квадрат у тепловій карті представляє кореляцію між двома ознаками: чим світліший колір квадрата, тим сильніша позитивна кореляція між ознаками, тоді як темніший колір вказує на сильнішу негативну кореляцію.

A screenshot of a computer

Description automatically generated

За цією матрицею кореляції можна зробити наступні висновки:

* Ознаки length, width, curb-weight, engine-size та horsepower мають високу позитивну кореляцію з ознакою price. Це означає, що ці ознаки можуть впливати на цінову категорію автомобіля. Зазвичай автомобілі з більшими розмірами, потужними двигунами та вагою коштують дорожче.
* Ознака highway-mpg має високу негативну кореляцію з price. Це може вказувати на те, що автомобілі з більшою кількістю миль на галон на шосе коштують менше. Тобто, більш паливоефективні автомобілі можуть мати меншу ціну.

Було проаналізовано дані наступними методами: Variance Threshold, Pearson Correlation, F-Score, Mutual Information.

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

Виходячи з результатів:

* Важливості деяких ознак: Ознаки width, curb-weight, engine-size, та horsepower виявилися важливими за всіма методами відбору ознак. Це може свідчити про їх значущість у передбаченні ціни автомобіля. Ці ознаки, швидше за все, мають сильний вплив на кінцеву ціну.
* Різні методи, різні важливість ознак: Ознаки можуть мати різний ступінь важливості в залежності від методу відбору. Наприклад, wheel-base має високу важливість за допомогою Pearson Correlation, але нижчу залишаючи Variance Threshold. Це вказує на різницю в їх оцінці впливу на ціну авто в кожному з методів.
* Ознаки з низькою важливістю: Ознаки, які мають низькі показники в усіх методах відбору, можуть бути менш впливовими на ціну. Наприклад, symboling та peak-rpm не показали високої важливості у жодному з методів, що може означати менший вплив цих ознак на цінову категорію авто.

**Висновок**

Оцінюючи результати цієї роботи, можна виділити кілька ключових висновків:

1. Аналіз категоріальних даних:

* Марка авто: Найбільша кількість автомобілів у наборі даних відноситься до марки Toyota, а найменша - до Porsche.
* Тип палива: Більшість авто використовують бензин, тоді як дизельні автомобілі складають менший відсоток.
* Тип кузова: Седани є найпоширенішим типом кузова, у той час як кабріолети зустрічаються рідше.

1. Аналіз розподілу значень:

* Графіки розподілу показали нерівномірний характер даних. Цей аналіз корисний для розуміння того, як різні значення розподілені у межах кожної ознаки.

1. Матриця кореляції:

* Дослідження матриці кореляції показало, як числові ознаки взаємозв'язані. Ознаки, такі як length, width, curb-weight, engine-size, та horsepower, мають високу позитивну кореляцію з ціною автомобіля.
* Ознака highway-mpg показала високу негативну кореляцію з ціною, вказуючи, що автомобілі з більшою кількістю миль на галон на шосе можуть коштувати менше.

1. Аналіз методів відбору ознак:

* Variance Threshold: Важливість ознак може варіюватися в залежності від цього методу. Деякі ознаки, які виявилися важливими за одним методом, можуть бути менш важливими за іншим.
* Pearson Correlation, F-Score, Mutual Information: Кожен метод виявляє важливість різних ознак. Ознаки width, curb-weight, engine-size, та horsepower показалися важливими за всіма методами.

В цілому, аналіз різноманітних аспектів даних дозволяє зрозуміти, як певні ознаки впливають на ціну автомобіля. Оцінка важливості ознак через різні методи відбору дозволяє вибрати найбільш інформативні фактори для подальшого моделювання та передбачення.