**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-32*

*спеціальності 122 «Комп'ютерні науки»*

*ОП «Прикладне програмування»*

*Симоніка Д.О*

*Викладач:*

*Шолохов О.В*

*Білий Р.О*

**Київ – 2022**

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

**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.

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.

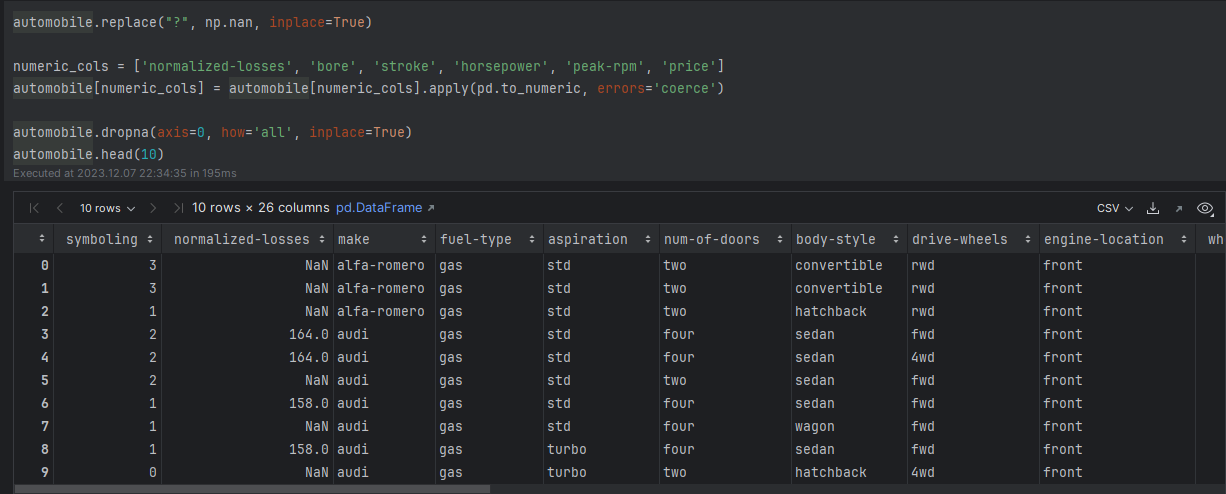
Третім фактором є відносна середня виплата збитку за рік страхування автомобіля. Це значення нормалізовано для всіх автомобілів певної класифікації розміру (дводверні маленькі, універсали, спортивні/спеціальні тощо) і являє собою середні втрати на автомобіль на рік.

**Завдання:**

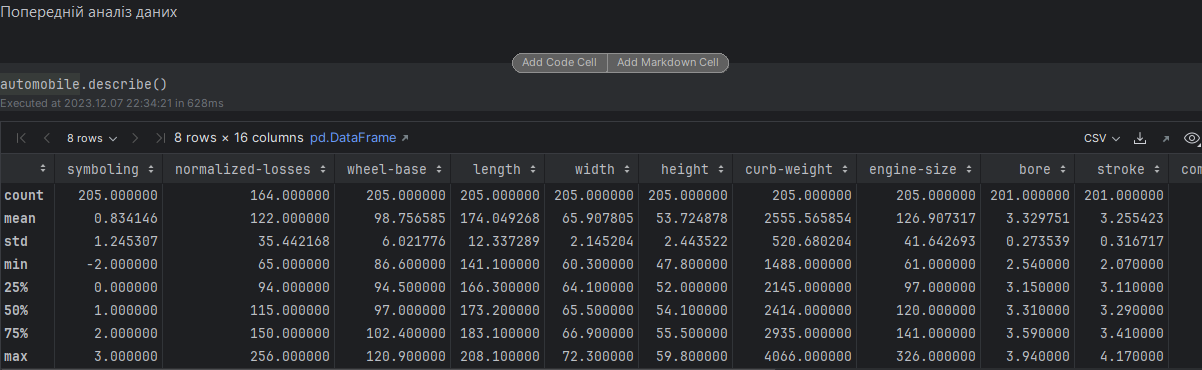
1. Завантажити файли з даними у папку проекту з посилання:
2. Очистити дані та обробити відсутні дані.
3. Зробити EDA по ознаках.
4. Проаналізуйте надані дані, використовуючи методи з прикладу та документації, та зберіть результати аналізу у результуючий ранжируваний датафрейм, в якому лівим індексом 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).
5. Проаналізуйте ознаки на взаємозалежність, та побудуйте відповідні heatmap засобами seaborn по кожному з використаних методів дослідження.
6. Зберіть висновки у звіт (графіки, висновки текстом у окремому файлі), який потребує належного оформлення, структури тощо.

**Хід роботи:**

Очистимо дані



Отримаємо попередній аналіз даних, а саме основні статистичні показники для кожного числового стовпця, такі як середнє значення (mean), стандартне відхилення (std), мінімальне і максимальне значення, а також квартилі (25%, 50%, 75%).



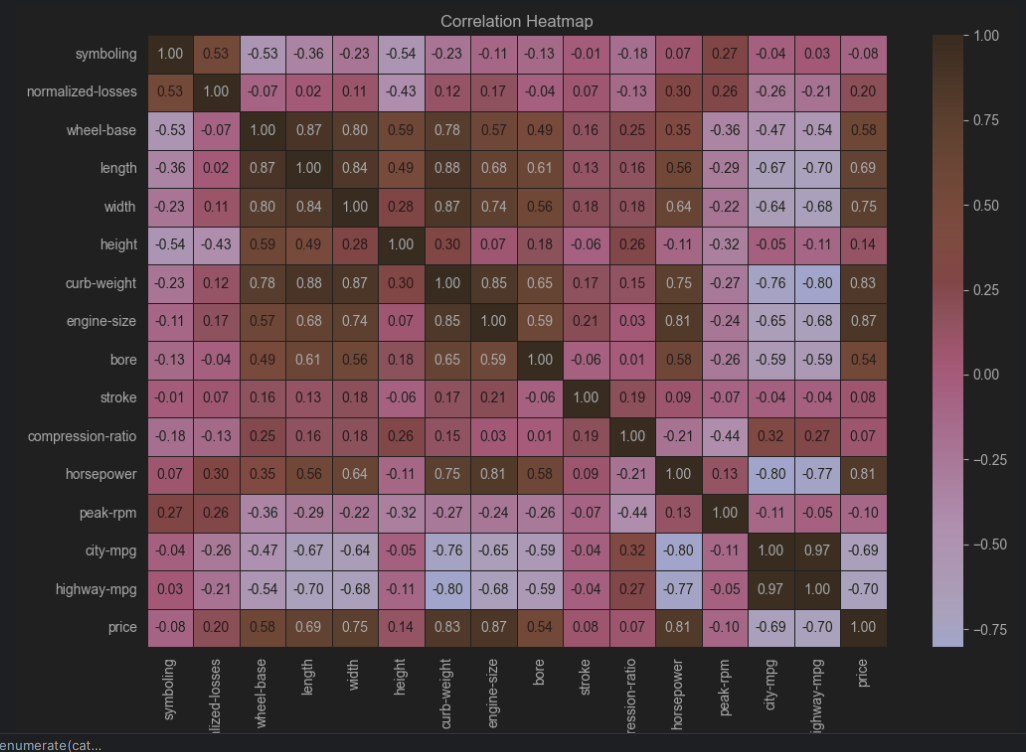
Далі проведемо аналіз розподілу по категоріям:



Бачимо, що на ринку переважають машини Toyota, бензинові машини, седани, більшість клієнтів обирають 4-дверні машини, машини з двигунами внутрішнього згорання, з чотирма циліндрами.

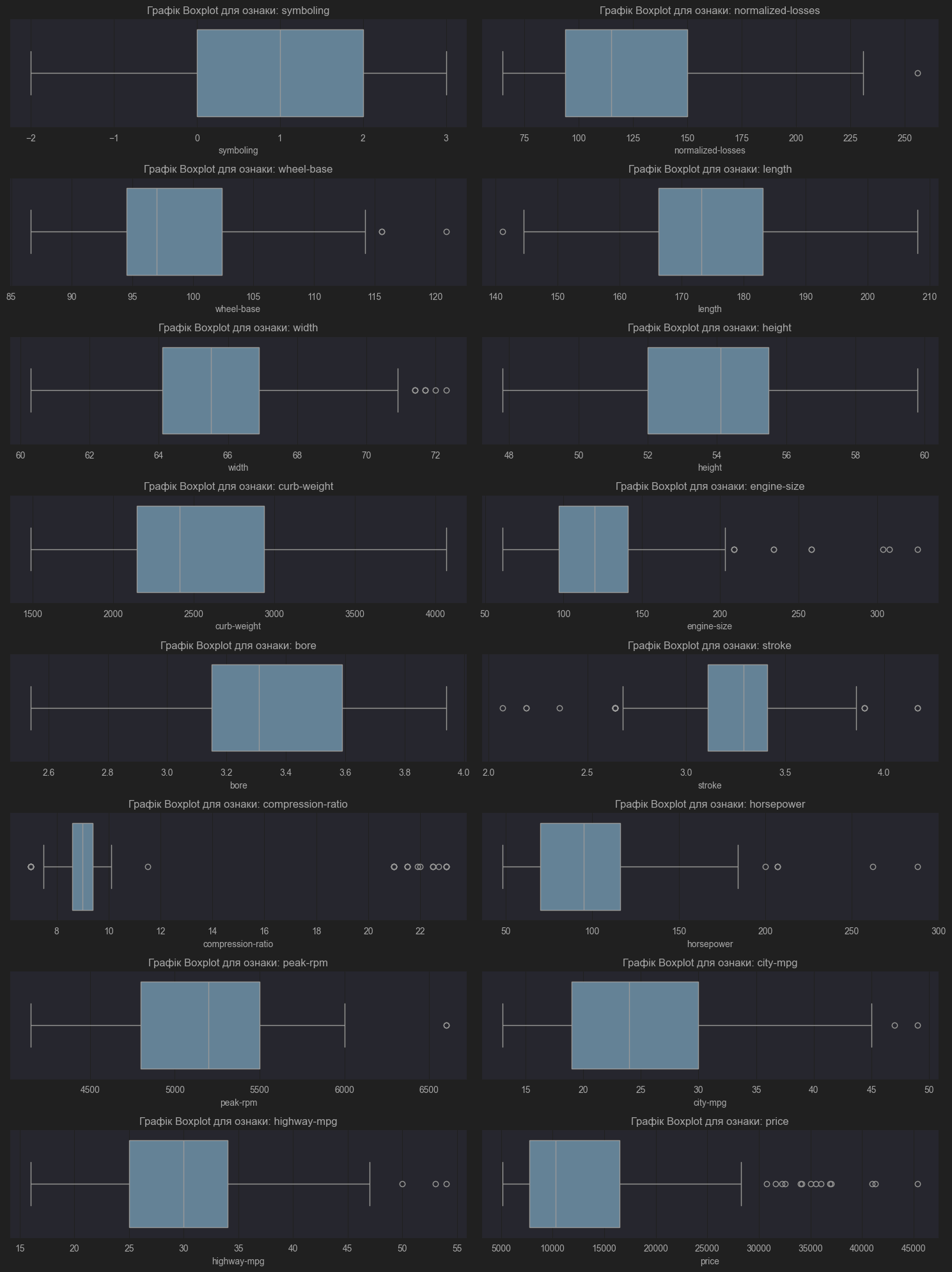
Найменшою популярність користуються машини марки Mercury, Renault, машини з 3 чи 12 циліндрами, машини, які працюють на дизельному паливі, та машини з повним приводом.

Далі проведемо кореляційний аналіз, який покаже наскільки значення однієї категорії пов’язані з іншою:



Бачимо звичайну пряму пропорційні довжини автівки до її висоти, кількість кінських сил до об’єму двигину, до ціни, яку клієнт віддасть за автівку. Також бачимо, пряму залежність розміру циліндра до об’єму двигуна. Можемо прослідкувати оберенену залежність: чим більша відстань між передньою та задньою осями автомобіля робить авто менш аварійним, таку ж ситуацію спостерігаємо при збільшенні висоти, ширини автівки, проте зі збільшення об’єму двигуна росте і аварійніст машини. Щодо ціни, то з її ростом збільшуються розміри автівки, об’єми двигуна, швидкість і тд.

Далі дослідимо викиди



Згідно даних графіків, ці викиди показують, що автомобілі у наборі даних мають широкий діапазон характеристик. Середні значення та квартили вказують на те, що автомобілі у наборі даних відносно однорідні, але є деякі екстремальні значення. Машини в наборі даних мають широкий діапазон розмірів.

* Довжина автомобіля коливається від 100 до 210 см, ширина - від 60 до 72 см, а висота - від 125 до 225 см.
* Машини в наборі даних мають широкий діапазон двигунів. Об'єм двигуна коливається від 1,2 до 3,5 л, а потужність - від 80 до 250 к.с.
* Машини в наборі даних мають широкий діапазон паливної ефективності. Витрата палива в місті коливається від 15 до 35 л/100 км, а витрата палива на шосе - від 30 до 45 л/100 км.

Проаналізуємо вплив ознак на ціну

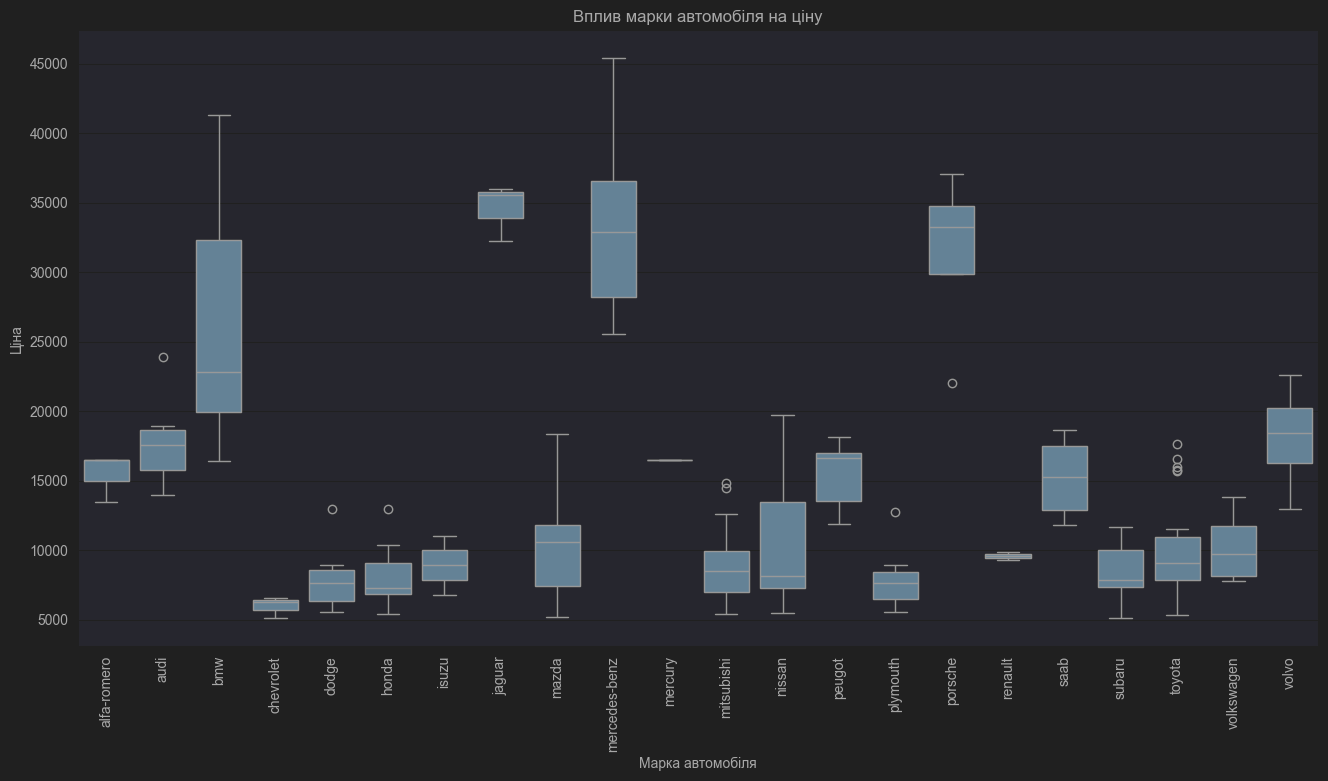


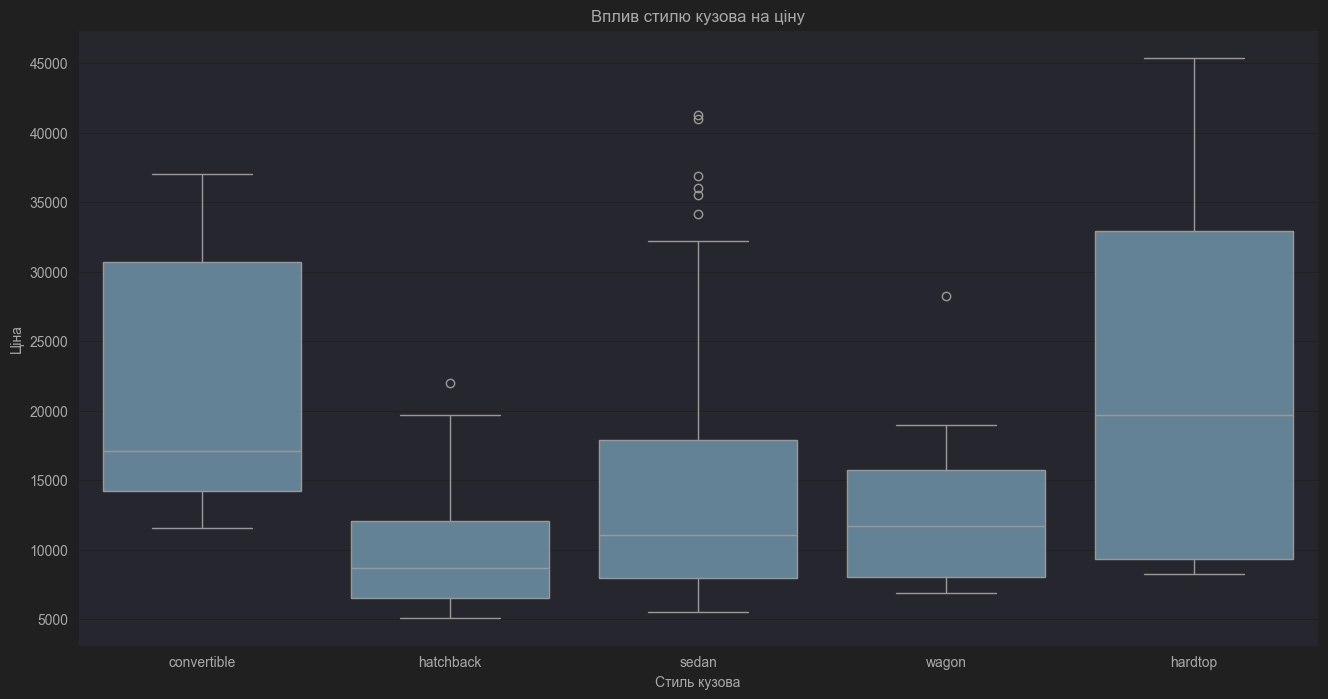
Можна зробити висновок, що

* Автомобілі з двигуном об'ємом 2,5 л або більше, як правило, коштують дорожче, ніж автомобілі з двигунами меншого розміру.
* Автомобілі з потужністю 150 к.с. або більше, як правило, коштують дорожче, ніж автомобілі з меншою потужністю.
* Автомобілі вагою 1500 кг або більше, як правило, коштують дешевше, ніж автомобілі з меншою вагою.

Вищенаведені дані можуть допомогти компанії знайти правильний підхід до широкої аудиторії покупців та для знаходження автомобілей, з якими потрібно провести мінімум роботи задля перепродажі. Наприклад, якщо автомобіль має значно більшу або меншу ціну, ніж можна очікувати на основі його характеристик, це може бути ознакою проблеми з двигуном або іншими компонентами автомобіля.

Зв’язок марки/ціни та стилю кузова/ціни



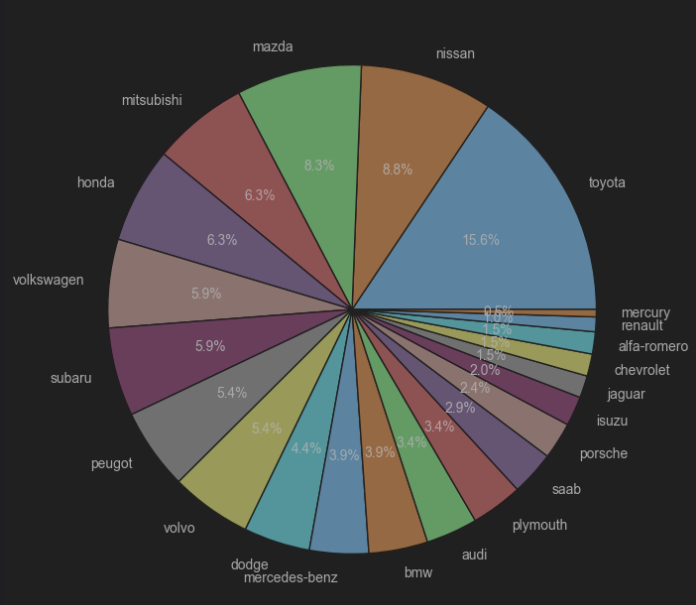
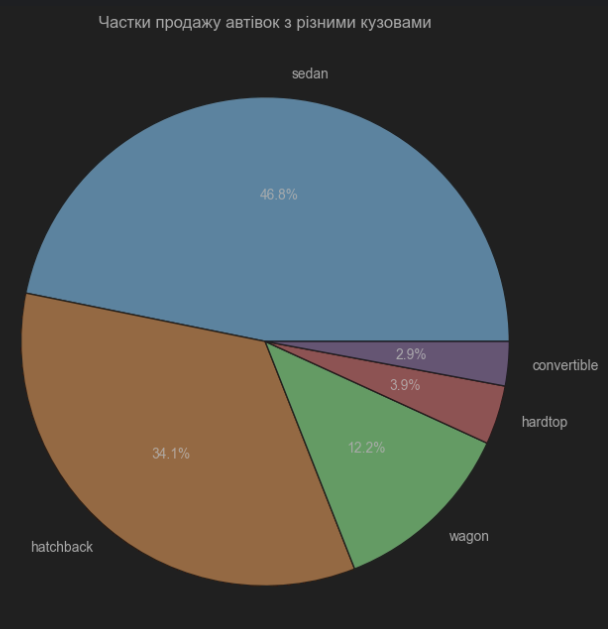


Звідси, бачимо, що марка автомобіля, стиль кузова та рік випуску є одними з найважливіших факторів, які впливають на ціну автомобіля.

Автомобілі від престижних марок, як правило, коштують дорожче, ніж автомобілі від менш престижних марок. Це пов'язано з тим, що престижні марки пропонують більш високу якість, надійність та імідж.

Автомобілі з більш престижними стилями кузова, як правило, коштують дорожче, ніж автомобілі з менш престижними стилями кузова. Це пов'язано з тим, що більш престижні стилі кузова вважаються більш привабливими та іміджево привабливим

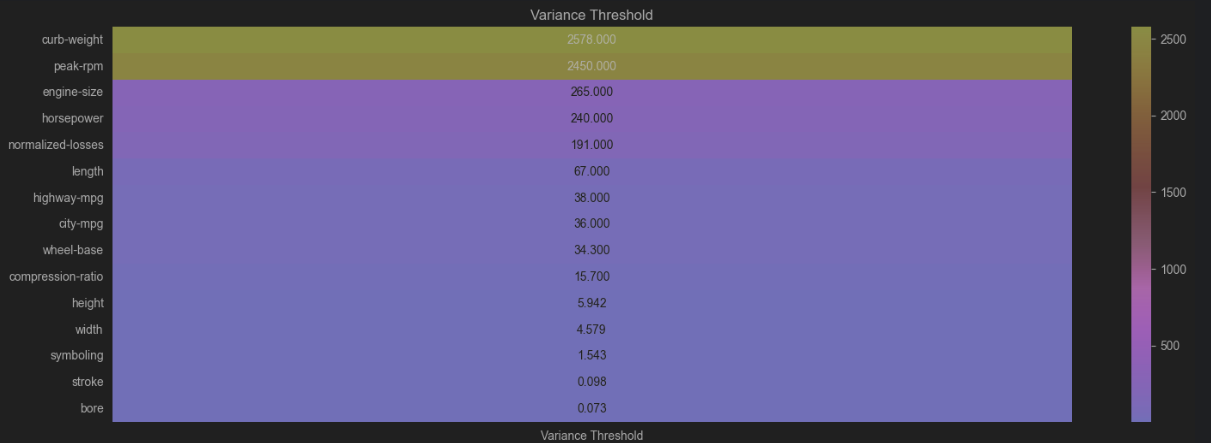
Далі проаналізуємо частки кузовів та брендів



Тобто, бачимо, що на продажах переважаю седани та марка Тойота, що може слугувати ключем до середньостатистичного клієнта, та відповіно росту прибутку.

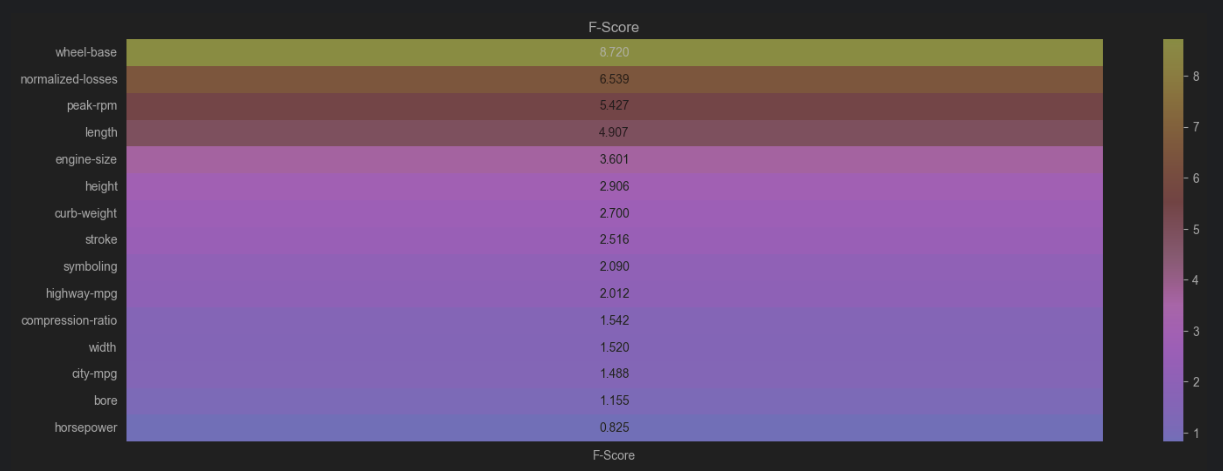
Далі проведемо дослідження за методами:

Variance Threshold - це простий метод вибору ознак, який видаляє ознаки з варіацією нижче певного порога. Варіація ознаки - це міра того, як поширені значення ознаки. Ознаки з низькою варіацією мають менший інформаційний виграш і тому менш корисні для прогнозування.



Хіп-мап порогового рівня варіації показує, що ознаки "wheel-base", "peek-rpm", "engine-size", "horsepower" мають найвищі оцінки порогового рівня варіації. Це означає, що ці ознаки мають найбільшу варіацію і тому є найбільш інформативними ознаками для прогнозування.

F-Score - це статистична міра, яка оцінює продуктивність класифікаційної моделі. Він обчислюється шляхом взяття гармонічного середнього значення точності та відкликання моделі. Ознаки з високим F-score є важливішими для прогнозування, оскільки вони є більш передбачуваними для змінної цілі.

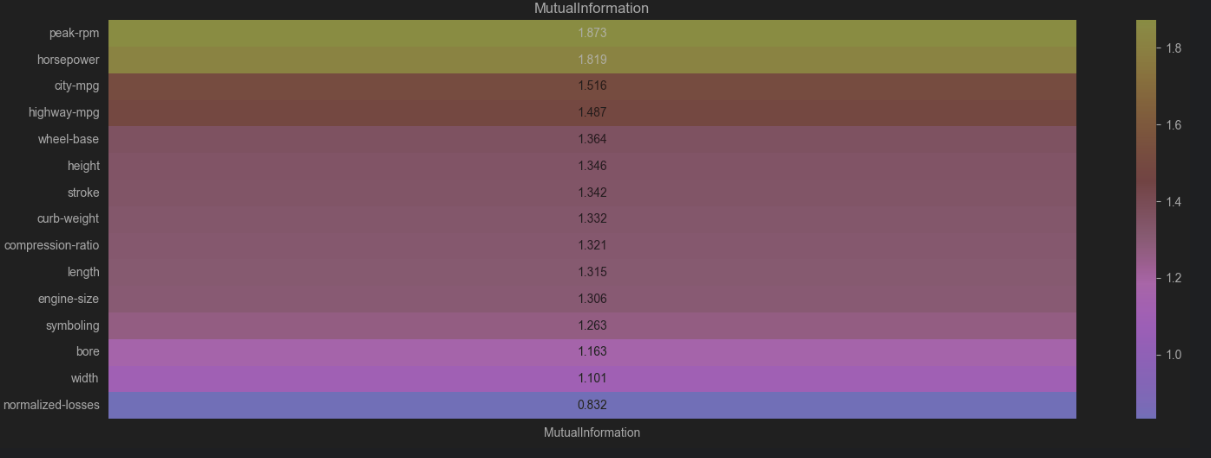


Бачимо, що ознаки "horsepower", "wheel-base", "length" та "width" мають найвищі оцінки F-score. Це означає, що ці ознаки є найбільш передбачуваними для ціни автомобіля.

Ознака "wheel-base" (потужність двигуна) має найвищу оцінку F-score, що означає, що вона є найбільш передбачуваною для ціни автомобіля. Це пов'язано з тим, що aвтомобілі з більшою базою коліс, як правило, більш просторі та комфортні, що може призвести до вищої ціни.

Загалом, хіп-мап показує, що ознаки "horsepower", "normalized-losses", "peak-rpm" та "length" є найбільш важливими для прогнозування ціни автомобіля.

MutualInformation - це міра взаємної залежності між двома змінними. Вона обчислюється шляхом вимірювання кількості інформації, яку одна змінна містить про іншу змінну. Ознаки з високою взаємною інформацією є важливішими для прогнозування, оскільки вони є більш інформативними про змінну цілі.



Звідси, бачимо, що ознаки "peak-rpm", "horsepower" та "city-mpg" є важливими для прогнозування ціни автомобіля .

Ознака "peak-rpm" (максимальні обороти в хвилину) - це міра того, наскільки швидко двигун може обертатися. Автомобілі з високими максимальними оборотами в хвилину, як правило, мають більш потужні двигуни, які коштують дорожче.

Ознака "horsepower" (потужність двигуна) - це міра того, скільки енергії може виробити двигун. Автомобілі з потужнішим двигуном, як правило, коштують дорожче, оскільки вони швидші та мають кращі характеристики.

Ознака "city-mpg" (міський витрата палива) - це міра того, скільки палива споживає автомобіль при їзді в міському циклі. Автомобілі з більш низьким витратою палива, як правило, коштують дорожче, оскільки вони вважаються більш економічними та екологічно чистими.

**Висновок**

У результаті аналізу ринку підтриманих автомобілів в Америці було визначено цінні висновки, які сприятимуть більш ефективному плануванню стратегій продажів та маркетингу.

По-перше, було визначено, що автомобілі марки Toyota,автомобілі з бензиновим двигуном та автомобілі з чотирма дверима, седани є найбільш популярними серед покупців. З іншого боку, марки Mercury та Renault, а також автомобілі на дизельному паливі чи з повним приводом, менше затребувані.

Важливими факторами, впливаючи на ціну автомобілів, є їхні розміри, об'єм двигуна та потужність. Великі автомобілі з потужними двигунами мають тенденцію коштувати дорожче. Крім того, престиж марки та стиль кузова впливають на цінову категорію: автомобілі відомих брендів та стильні кузови частіше коштують дорожче.

Аналіз взаємозв'язків та впливових факторів виявив ключові ознаки, такі як об'єм двигуна, потужність та розміри автомобіля, які слід враховувати при визначенні стратегій ціноутворення та маркетингу. Методи відбору ознак, такі як Variance Threshold, F-Score та Mutual Information, дозволяють визначити найбільш інформативні фактори для прогнозування ціни.

Отримані результати не лише допомагають у вирішенні конкретних завдань, але й служать основою для розробки стратегій залучення клієнтів, позиціонування продукту та вдосконалення асортименту. Зазначені висновки допоможуть компанії бути більш конкурентоспроможною на ринку підтриманих автомобілів та точно відповідати потребам різних сегментів клієнтів.