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»

Студентки 4 курсу
групи ПП-41
спеціальності 122 «Комп'ютерні науки»
ОП «Прикладне програмування»
Штонь Софії Максимівни

Викладач:
Білий Р. О.

Київ – 2023

Тема роботи: Методи аналізу та вибору значущих ознак (Features' Selection Procedures).

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.

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».

Інформація про атрибути:

1. symboling: -3, -2, -1, 0, 1, 2, 3
2. normalized-losses: continuous from 65 to 256
3. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas
5. aspiration: std, turbo
6. num-of-doors: four, two
7. body-style: hardtop, wagon, sedan, hatchback, convertible
8. drive-wheels: 4wd, fwd, rwd
9. engine-location: front, rear
10. wheel-base: continuous from 86.6 120.9
11. length: continuous from 141.1 to 208.1
12. width: continuous from 60.3 to 72.3
13. height: continuous from 47.8 to 59.8
14. curb-weight: continuous from 1488 to 4066
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor
16. num-of-cylinders: eight, five, four, six, three, twelve, two
17. engine-size: continuous from 61 to 326
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi

19. bore: continuous from 2.54 to 3.94
20. stroke: continuous from 2.07 to 4.17
21. compression-ratio: continuous from 7 to 23
22. horsepower: continuous from 48 to 288
23. peak-rpm: continuous from 4150 to 6600
24. city-mpg: continuous from 13 to 49
25. highway-mpg: continuous from 16 to 54
26. price: continuous from 5118 to 45400.

Завдання для 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.

Хід роботи

Data cleaning:

Першим чином, замінимо всі знаки питання на NaN.

```
In [4]: df.replace("?", np.NaN, inplace=True) if "?" in df.values else None

        df
```

Out[4]:

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	whee ba
0	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88
1	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88
2	1	NaN	alfa-romero	gas	std	two	hatchback	rwd	front	94
3	2	164	audi	gas	std	four	sedan	fwd	front	99
4	2	164	audi	gas	std	four	sedan	4wd	front	99
...	...	...	...	...	...	...	...	...	...	
200	-1	95	volvo	gas	std	four	sedan	rwd	front	109
201	-1	95	volvo	gas	turbo	four	sedan	rwd	front	109

Тепер дізнаємось, скільки NaN значень в кожному стовпчику датасету.

```
normalized-losses    41
num-of-doors          2
bore                  4
stroke                4
horsepower            2
peak-rpm              2
price                 4
dtype: int64
```

Спочатку заповнимо всі стовпчики окрім normalized-losses, оскільки в них відносно небагато NaN значень. Оскільки ці стовпці репрезентують категорії, використовуємо моду (найпоширеніше значення). Для числових стовпців використаємо медіану.

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsep
0	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	
1	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	
2	1	NaN	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.4	10.0	
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.4	8.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
200	-1	95	volvo	gas	std	four	sedan	rwd	front	109.1	...	141	mpfi	3.78	3.15	9.5	
201	-1	95	volvo	gas	turbo	four	sedan	rwd	front	109.1	...	141	mpfi	3.78	3.15	8.7	
202	-1	95	volvo	gas	std	four	sedan	rwd	front	109.1	...	173	mpfi	3.58	2.87	8.8	
203	-1	95	volvo	diesel	turbo	four	sedan	rwd	front	109.1	...	145	idi	3.01	3.4	23.0	
204	-1	95	volvo	gas	turbo	four	sedan	rwd	front	109.1	...	141	mpfi	3.78	3.15	9.5	

205 rows × 26 columns

Перевіримо типи даних для кожного стовпчика:

```
symboling              int64
normalized-losses     object
make                  object
fuel-type             object
aspiration            object
num-of-doors          object
body-style            object
drive-wheels          object
engine-location       object
wheel-base           float64
length               float64
width                float64
height               float64
curb-weight            int64
engine-type           object
num-of-cylinders      object
engine-size            int64
fuel-system           object
bore                  object
stroke                object
compression-ratio    float64
horsepower            object
peak-rpm              object
city-mpg               int64
highway-mpg            int64
price                 object
dtype: object
```

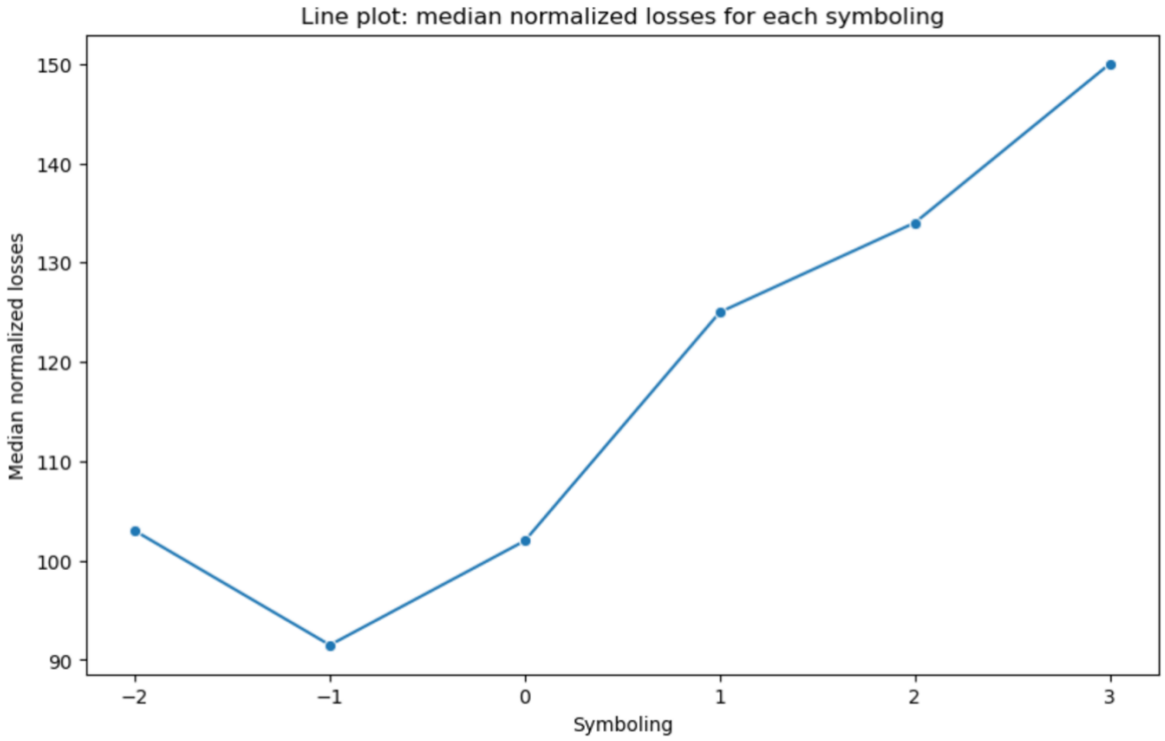
Колонки 'normalized-losses', 'bore', 'stroke', 'horsepower', 'peak-rpm', 'price' повинні бути числовими, тож конвертуємо їх:

```
symboling              int64
normalized-losses    float64
make                  object
fuel-type             object
aspiration            object
num-of-doors          object
body-style            object
drive-wheels          object
engine-location       object
wheel-base           float64
length               float64
width                float64
height               float64
curb-weight            int64
engine-type           object
num-of-cylinders      object
engine-size            int64
fuel-system           object
bore                 float64
stroke               float64
compression-ratio    float64
horsepower             int64
peak-rpm               int64
city-mpg               int64
highway-mpg            int64
price                float64
dtype: object
```

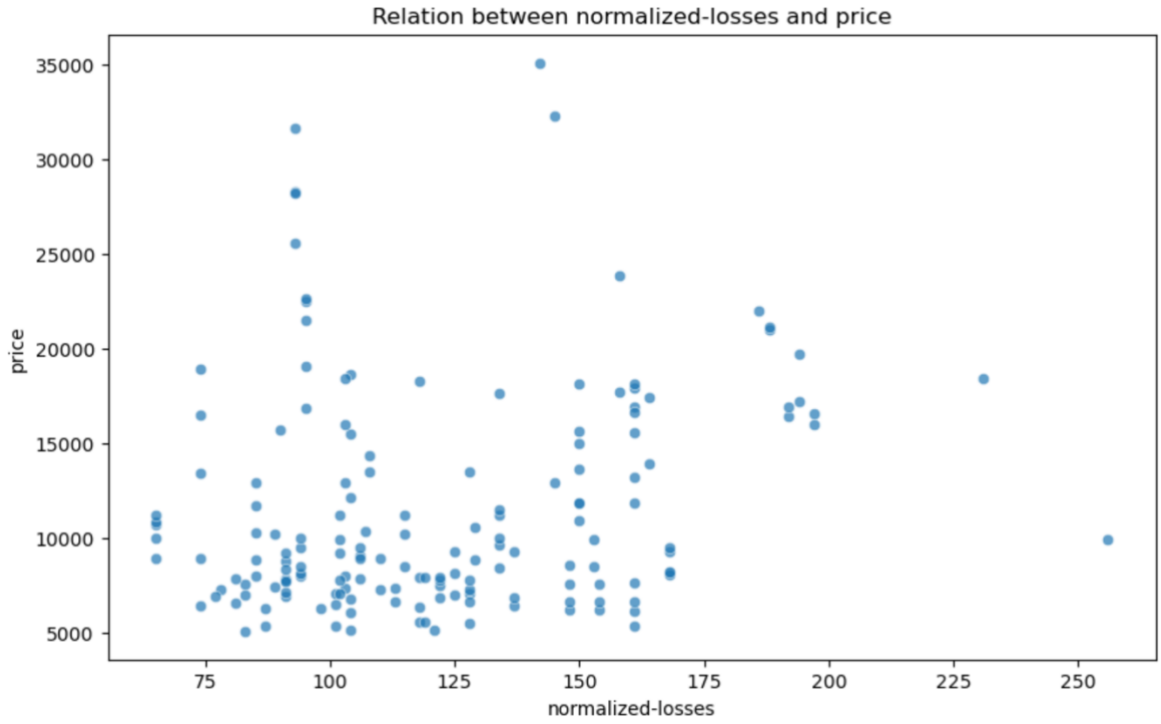
Тепер типи даних виглядають правильно. Перевіримо, які ще стовпчики мають NaN значення.

```
normalized-losses    41
dtype: int64
```

Бачимо, що залишилось заповнити стовпчик normalized-losses.



З цього графіку бачимо, що чим вищий "символ", тим вища відносна середня виплата збитку за рік страхування автомобіля. Проаналізуємо відношення normalized losses до інших ознак.



Можна сказати, що відношення між normalized-losses та ціною немає. Відношення з іншими стовпчиками також немає. Тому, заповнимо NaN значення normalized-losses медіаною для кожного значення 'symboling'.

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsep
0	3	150.0	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	
1	3	150.0	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	
2	1	125.0	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	
3	2	164.0	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0	
4	2	164.0	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
200	-1	95.0	volvo	gas	std	four	sedan	rwd	front	109.1	...	141	mpfi	3.78	3.15	9.5	
201	-1	95.0	volvo	gas	turbo	four	sedan	rwd	front	109.1	...	141	mpfi	3.78	3.15	8.7	
202	-1	95.0	volvo	gas	std	four	sedan	rwd	front	109.1	...	173	mpfi	3.58	2.87	8.8	
203	-1	95.0	volvo	diesel	turbo	four	sedan	rwd	front	109.1	...	145	idi	3.01	3.40	23.0	
204	-1	95.0	volvo	gas	turbo	four	sedan	rwd	front	109.1	...	141	mpfi	3.78	3.15	9.5	

205 rows × 26 columns

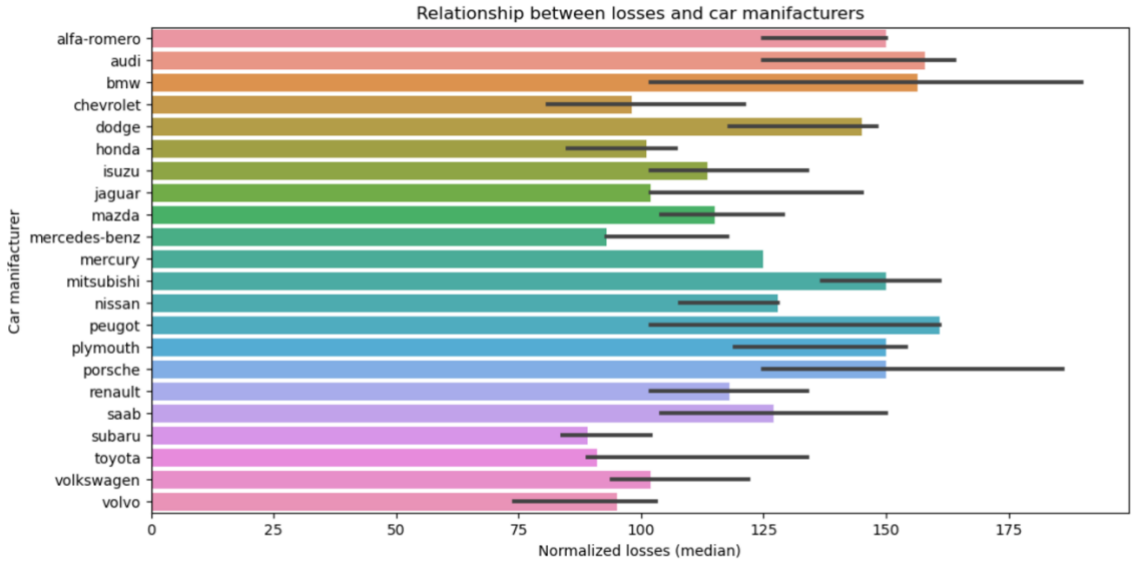
Отже, датасет був очищений.

EDA:

```
Numeric Statistics:
        symboling  normalized-losses  wheel-base      length       width  \
count  205.000000         205.000000  205.000000  205.000000  205.000000
mean     0.834146         121.492683   98.756585  174.049268   65.907805
std      1.245307          33.025149    6.021776   12.337289    2.145204
min     -2.000000          65.000000   86.600000  141.100000   60.300000
25%      0.000000          98.000000   94.500000  166.300000   64.100000
50%      1.000000         115.000000   97.000000  173.200000   65.500000
75%      2.000000         150.000000  102.400000  183.100000   66.900000
max      3.000000         256.000000  120.900000  208.100000   72.300000

           height  curb-weight  engine-size        bore      stroke  \
count  205.000000   205.000000   205.000000  205.000000  205.000000
mean    53.724878  2555.565854   126.907317    3.329366    3.256098
std      2.443522   520.680204    41.642693    0.270858    0.313634
min     47.800000  1488.000000    61.000000    2.540000    2.070000
25%     52.000000  2145.000000    97.000000    3.150000    3.110000
50%     54.100000  2414.000000   120.000000    3.310000    3.290000
75%     55.500000  2935.000000   141.000000    3.580000    3.410000
max     59.800000  4066.000000   326.000000    3.940000    4.170000

       compression-ratio  horsepower     peak-rpm     city-mpg  highway-mpg  \
count         205.000000  205.000000   205.00000  205.000000   205.000000
mean           10.142537  103.902439  5129.02439   25.219512    30.751220
std             3.972040   39.680343   478.40526    6.542142     6.886443
min             7.000000   48.000000  4150.00000   13.000000    16.000000
25%             8.600000   70.000000  4800.00000   19.000000    25.000000
50%             9.000000   95.000000  5200.00000   24.000000    30.000000
75%             9.400000  116.000000  5500.00000   30.000000    34.000000
max            23.000000  288.000000  6600.00000   49.000000    54.000000

              price
count    205.000000
mean   13150.307317
std     7879.121326
min     5118.000000
25%     7788.000000
50%    10295.000000
75%    16500.000000
max    45400.000000
```

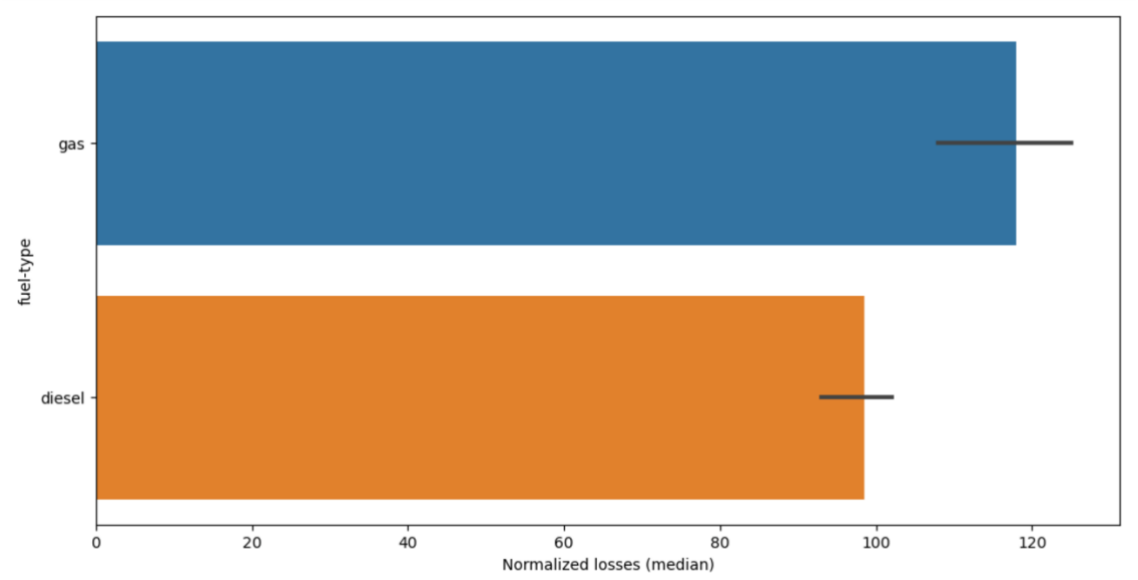
```
Categorical Statistics:
          make fuel-type aspiration num-of-doors body-style drive-wheels  \
count      205       205        205          205        205          205
unique      22         2          2            2          5            3
top     toyota       gas        std         four      sedan          fwd
freq        32       185        168          116         96          120

       engine-location engine-type num-of-cylinders fuel-system
count              205         205               205         205
unique               2           7                 7           8
top              front         ohc              four        mpfi
freq               202         148               159          94
```

Проаналізуємо відношення між збитками та марками автомобілів:



Бачимо, що найбільші втрати спостерігаються у марок Peugot, Audi, BMW.

Проаналізуємо збитки за типом пального:



Бачимо, що в середньому, автомобілі на дизелі мають менший показник збитку.

Features' Selection Procedures

Вибір значущих ознак описується у файлі DSBD_Звіт_Лаб3_ПП41_ШтоньС.

Висновок: Отже, у ході цієї лабораторної роботи було отримано навички аналізу та вибору значущих ознак для моделі.