Лабораторна робота №2

Візуальний аналіз даних

Мета роботи: Набути практичних навичок обробки даних і аналізу отриманих результатів з метою прийняття коректних обґрунтованих рішень технічного, ділового, політичного, персонального або іншого характеру, використовував можливості Python.

Література

numpy.std: https://numpy.org/doc/stable/reference/generated/numpy.std.html
https://seaborn.pydata.org/generated/seaborn.kdeplot.html
https://seaborn.pydata.org/index.html

Зміст роботи

Завдання. Графічне представлення елементів описової статистики

– З бібліотек знадобляться matplotlib і seaborn^

```
import matplotlib.pyplot as plt

import seaborn as sns
```

– Побудуємо стовпчикову діаграму для змінної - кількість кімнат:

```
graphic = df['Кімнат'].value_counts().sort_index().plot(kind='bar
', figsize=(4,4), color='blue')

plt.show()
```

– Побудуємо гістограму для змінної загальна площа:

```
histogram1 = plt.subplot(121)

histogram1 = df['Загальна_площа'].sort_index().plot(kind='hist',
figsize=(20,10), color='blue')
```

– Побудуємо графік розсіювання для визначення залежності ціни від загальної площі :

```
fig, xy = plt.subplots(figsize=(10,10))
xy.scatter(x=df['Ціна'], y=df['Загальна_площа'])
plt.show()
```

– Порівняємо розподіл цін по містах (побудуємо коробчасту діаграму) та розмістити коробчасті діаграми горизонтально.

```
figure = plt.figure(1, figsize=(9, 6))
sns.boxplot(x='Ціна',y='Місто',data=df)
```

– Побудуємо кореляційну матрицю (heatmap). Матриця формується засобами Pandas, зі стандартним значенням параметрів.

```
sns.heatmap(df.corr(), cmap="crest")
```

– Визначте які дві ознаки найбільше корелюють?

– Побудуємо діаграму оцінки щільності. Діаграма оцінки щільності (KDE) — це метод візуалізації розподілу спостережень у наборі даних, аналогічний гістограмі.

```
sns.kdeplot(df[df['Кімнат'] == 1]['Ціна'])
plt.show()
```

– Побудуйте діаграму оцінки щільності для різних кімнат на одному графіку

– Побудуйте коробчасту діаграму для візуалізації розподілу цін в залежності від кількості кімнат.

– Побудуйте графік розсіювання, який відображатиме залежність загальної кількості кімнат від ціни.

– Побудуйте гістограму для оцінки розподілу ціни квартир

– Побудуйте гістограму (використовується для оцінки форми розподілу кількісної змінної) розподілу квартир, які продаються за загальною площею. Залежно від розміру інтервалу форма гістограми може змінюватися. Змініть інтервал з 25 м.кв. до 50.

Методичні рекомендації

Візуалізація даних - це наочне уявлення масивів різної інформації. Існує кілька типів візуалізації:

Стовпчикова діаграма *використовується для візуалізації категоріальних або кількісних дискретних даних*

Кругова діаграма *використовується для візуалізації категоріальних або кількісних дискретних даних з метою зрозуміти відношення складових до загального значення.*

Гістограма *використовується у статистиці для графічного представлення розподілу ймовірностей значень випадкової величини.*

На діаграмах розсіювання ряд точок, розміщених в декартовій системі координат, ***відображає значення за двома змінними****. Присвоївши кожної осі змінну, можна визначити, чи існують відносини або кореляція між цими двома змінними.*

У графічному вигляді міри центральної тенденції та міри розсіювання в одній або декількох групах зручно представляти за допомогою графіків «***скриня з вусами***» (рис.4).

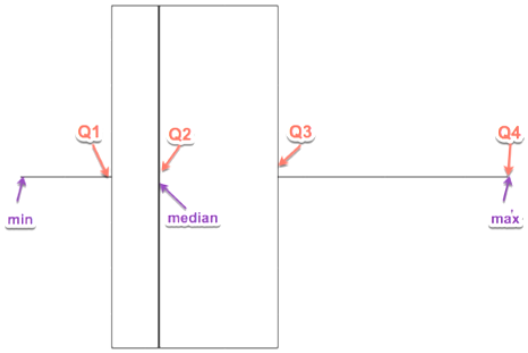


Рис.4. «Скриня з вусами»

Matplotlib - бібліотека Python для візуалізації даних двовимірної (2D) графіки, тривимірна (3D) графіка також підтримується.

При виборі типу графіка для візуалізації потрібно розуміти тип даних та що ви хочете зрозуміти.

Порівнювати значення: стовпчикова діаграма, лінійний графік або графік розсіювання.

Зрозуміти композицію(виділити складові): стовпчикова діаграма, кругова діаграма.

Оцінити розподіл даних: лінійний графік, стовпчикова діаграма, гістограма, графік розсіювання.

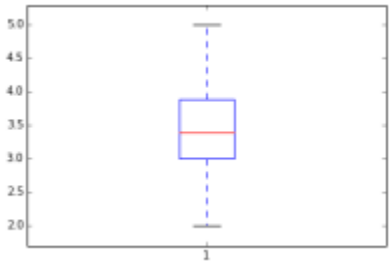
Зрозуміти тренд: лінійний графік, стовпчикова діаграма.

Зрозуміти відношення між даними: лінійний або графік розсіювання.

Контрольні запитання

1. За допомогою якого графіка можна дізнатися середнє значення (мат. очікування) і розкид значень (дисперсію) для різних категорій даних.

2. Опишіть основні елементи наступної діаграми і яким чином можна її отримати:



3. Якщо використовувати метод *DataFrame.plot()* з параметром kind = 'bar', який вид діаграми можна отримати?

4. Опишіть призначення графіка *HeatMap* (Теплова карта).

5. Для чого призначені функції *plt.show() і plt.draw()*?