Лабораторная работ 3.
Разведочный анализ данных.

Цель: познакомиться с методами разведочного анализа пред обработанных данных.
Задачи:

1. Загрузить пред обработанных данные из первой лабораторной;
2. Проанализировать распределение переменных;
3. Исследовать корреляцию между переменными;
4. Исследовать вбросы и аномалии;
5. Исследовать категориальные переменные;
6. Визуализировать результат в общий дешборд.

Пример выполнения лабораторной работы.

Вспомним какие данные находятся в нашем датасете. Импортируем необходимые библиотеки, загружаем данные и выводим DataFrame.

```
1. import pandas as pd
2. import numpy as np
3. import matplotlib.pyplot as plt
4. import seaborn as sns
5. df = pd.read_csv('lab2.csv',encoding='cp1251')
6. print(df.head(5))
```

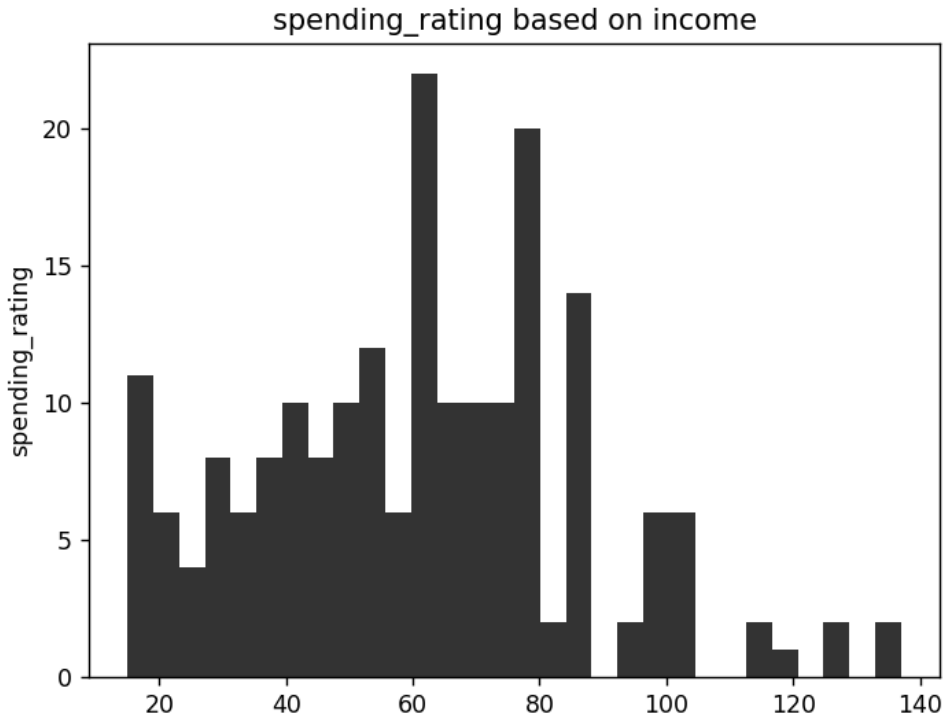
Результат:

	id	gender	age	income	spending_rating
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Наши данные уже предобработанны по этому приступим сразу к распределению числовых переменных. Построим гистограмму рейтинга трат в зависимости от дохода.

```
1.df = pd.read_csv('lab2.csv',encoding='cp1251')
2.plt.hist(df['income'], bins=30, color='black', alpha=0.8)
3.plt.xlabel('income')
4.plt.ylabel('spending_rating')
5.plt.title('spending_rating based on income')
6.plt.show()
```

Результат:



Наиболее высокий рейтинг трат наблюдается при доходе с показателем 60.

Далее рассмотрим корреляции между числовыми переменными. Используем коэффициент корреляции, чтобы определить, существует ли связь между возрастом и доходом покупателя.

```
1.df = pd.read_csv('lab2.csv',encoding='cp1251')
2.correlation = df['age'].corr(df['income'])
3.print("correlation between buyer age and income:", correlation)
```

Результат:

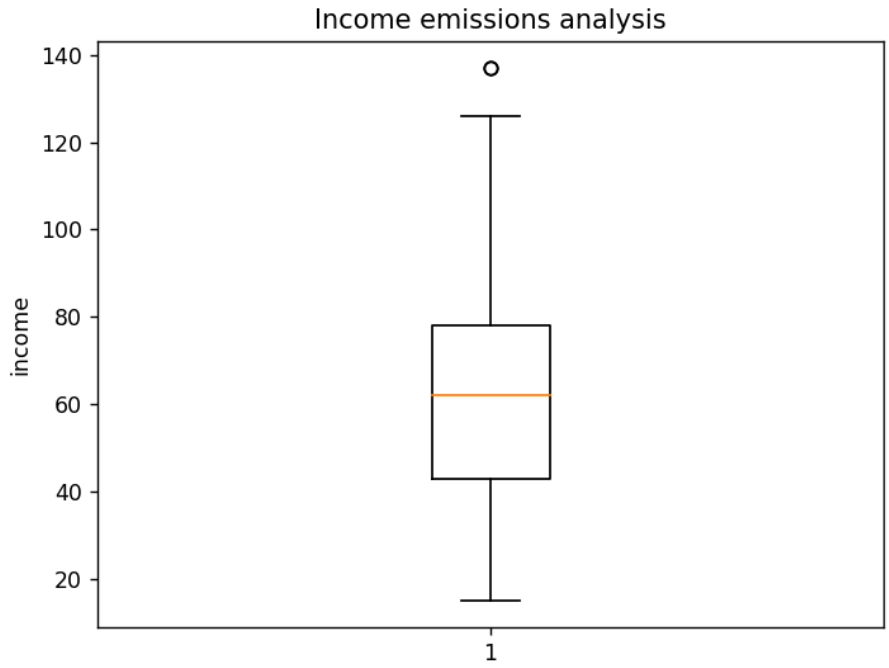
```
correlation between buyer age and income: -0.011487922703556497
```

Коэффициент корреляции отрицательный. Это значит, что при увеличении переменной возраста, переменная дохода уменьшается.

А теперь с помощью «Ящика с усами» найдем выбросы в данных в столбце дохода.

```
1.df = pd.read_csv('lab2.csv',encoding='cp1251')
2.plt.boxplot(df['income'])
3.plt.ylabel('income')
4.plt.title('Income emissions analysis')
5.plt.show()
```

Результат:

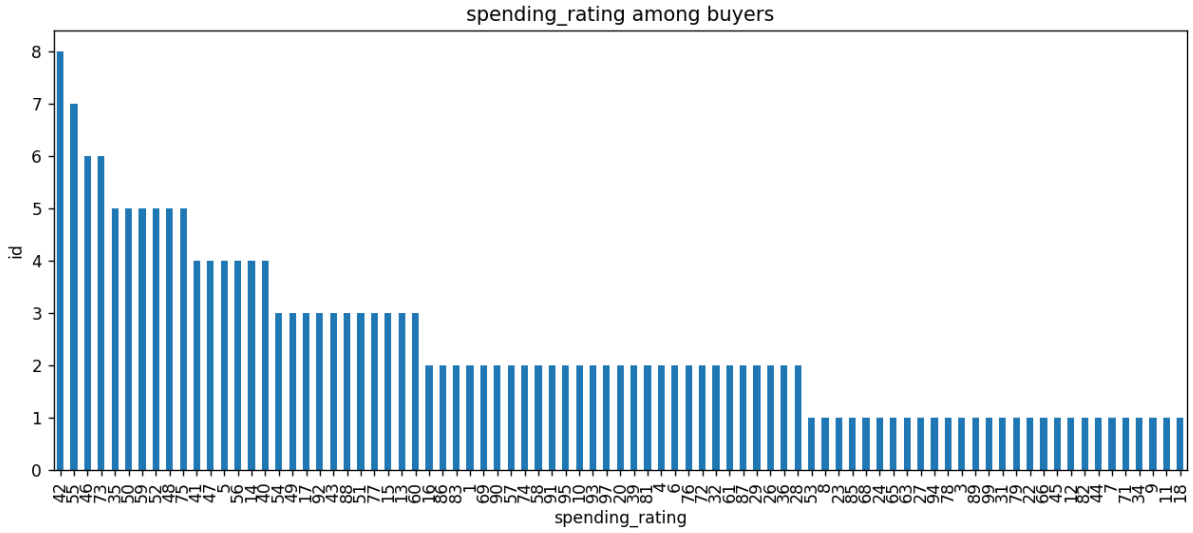


На диаграмме видно, что нижнее значение дохода - 20, медиана на уровне - 60, верхнее значение -130, и есть выброс в районе - 140.

Исследуем распределение частот категориальных переменных. Построим график рейтинга трат среди всех посетителей магазина.

```
1.df = pd.read_csv('lab2.csv',encoding='cp1251')
2.s_r = df['spending_rating'].value_counts()
3.s_r.plot(kind='bar')
4.plt.xlabel('spending_rating')
5.plt.ylabel('id')
6.plt.title('spending_rating among buyers')
7.plt.show()
```

Результат:

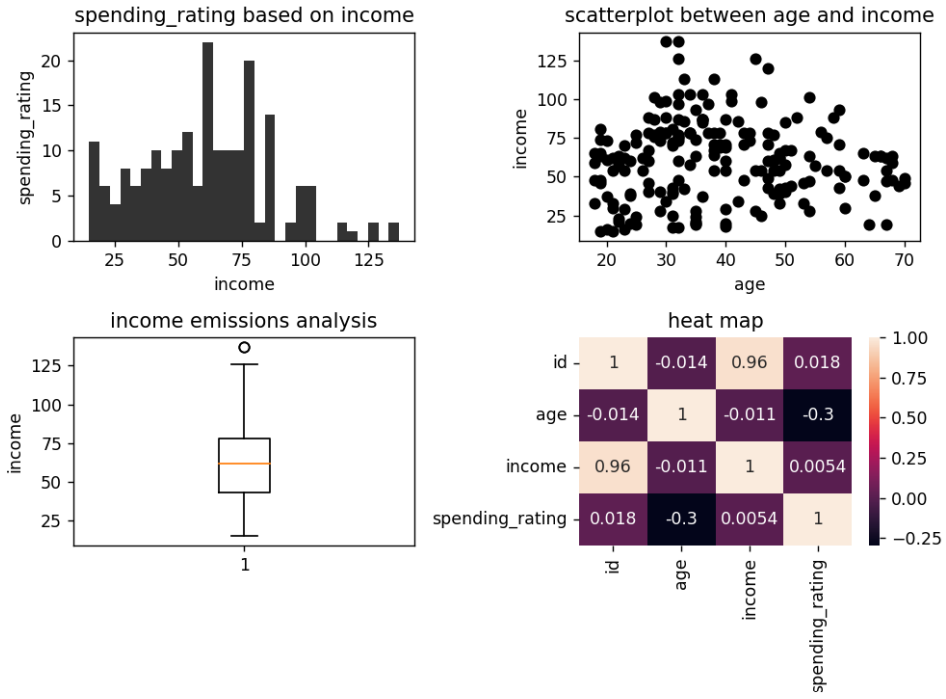


Чаще всего посетители магазина тратят сумму - 42.

Визуализируем результаты разведочного анализа в общем дешборде, который включает: гистограмму, диаграмму рассеяния, ящик с усами и тепловую карту.

```
df = pd.read_csv('lab2.csv',encoding='cp1251')
plt.figure(figsize=(8, 6))
plt.subplot(2, 2, 1)
plt.hist(df['income'], bins=30, color='black', alpha=0.8)
plt.xlabel('income')
plt.ylabel('spending_rating')
plt.title('spending_rating based on income')
plt.subplot(2, 2, 2)
plt.scatter(df['age'], df['income'], color='black')
plt.xlabel('age')
plt.ylabel('income')
plt.title('scatterplot between age and income')
plt.subplot(2, 2, 3)
plt.boxplot(df['income'])
plt.ylabel('income')
plt.title('income emissions analysis')
plt.subplot(2, 2, 4)
sns.heatmap(df.corr(), annot=True, fmt='.2g' )
plt.title('heat map')
plt.tight_layout()
plt.show()
```

Результат:



В ходе лабораторной работы мы провели разведочный анализ данных и определили, что:

1. Наиболее высокий рейтинг трат наблюдается при доходе с показателем 60;
2. Корреляция отрицательная, чем выше возраст, тем меньше доход;
3. «Диаграмма с усами» показала выброс в столбце дохода;
4. Наиболее частой суммой покупок является 42.

Задание.

1. Вывести на экран первые 5 строк вашего пред обработанного датасета;
2. Проанализировать распределение переменных;
3. Исследовать корреляцию между переменными;
4. Исследовать вбросы и аномалии;
5. Исследовать категориальные переменные;
6. Визуализировать результат в общий дешборд.

Формат отчета.

Протокол лабораторной работы в формате PDF, который должен содержать, поэтапное выполнение всех задач с текстовым описанием ваших действий и экранными формами, отображающими данные действия.