Нижегородский государственный университет им. Н.И. Лобачевского
Институт информационных технологий, математики и механики

Наглядный вероятностно-статистический анализ данных

Практическое задание 4

«Числовые характеристики многомерных случайных величин. Основы корреляционного анализа»

Пройдакова Екатерина Вадимовна,
доцент кафедры ТВиАД ИИТММ

Содержание

- Анализ многомерных данных: выделение зависимости
  - Пример: данные о результатах экзаменов
  - Числовые характеристики многомерных данных
  - Диаграммы рассеивания
  - Пример: данные о рейтингах покупателей ТЦ
- Практическое задание

1. АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ: ВЫЯВЛЕНИЕ ЗАВИСИМОСТИ

1. Анализ многомерных данных: выделение зависимости
1.1 Пример: данные о результатах экзаменов

Рассмотрим данные об итоговых экзаменах студентов США (Источник данных https://www.kaggle.com/spscientist/students-performance-in-exams) по столбцам:

A - пол (бинарные данные),

B - расовая принадлежность (одна из 5 этнических групп – номинальные данные),

C - уровень образования родителей (ординальные данные, можно упорядочить),

D - наличие льгот (бинарные данные),

E - прохождение подготовительного курса (бинарные данные),

F, G, H - оценки за экзамены по математике, чтению, письму (0-100, дискретные данные).

Цель:
определить степень зависимости между характеристиками

	A	B	C	D	E	F	G	H
1	gender	race/ ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
2	female	group B	bachelor's degree	standard	none	72	72	74
3	female	group C	some college	standard	completed	69	90	88
4	female	group B	master's degree	standard	none	90	95	93
5	male	group A	associate's degree	free/reduced	none	47	57	44
6	male	group C	some college	standard	none	76	78	75
7	female	group B	associate's degree	standard	none	71	83	78
8	female	group B	some college	standard	completed	88	95	92

1. Анализ многомерных данных: выделение зависимости
1.1 Пример: данные о результатах экзаменов

- Подготовим данные к анализу: импортируем для работы в Python

```python
import xlrd

def importData(bookName, sheetInd):
    res = []
    #указывается полный путь до файла
    book = xlrd.open_workbook(bookName)
    #извлекаем лист с данными по индексу
    sheet = book.sheet_by_index(sheetInd)
    # подгружаем данные построчно в список
    for i in range(1, sheet.nrows): # с единицы, так как не подгружаем заголовки
        res.append(sheet.row_values(i, 0, sheet.ncols))
    return res

# ===========  Данные о тестировании учеников =================

examsData = importData('04_Экзамены.xlsx', 0)
```

1. Анализ многомерных данных: выделение зависимости
1.1 Пример: данные о результатах экзаменов

- Преобразуем данные в более удобный для нас формат, введем порядок для ординальных данных

```
types = {0:"Пол", 1:"Расовая группа", 2:"Образование родителей",
         3:"Наличие льгот", 4:"Подготовительный курс",
         5:"Оценка по математике", 6:"Оценка по чтению", 7:"Оценка по письму"}

genderTypes = {"male": "М", "female": "Ж"}
raceTypes = {"group A": 0, "group B": 1, "group C": 2, "group D": 3, "group E": 4}
edTypes = {"some high school": 0, "high school": 1, "some college": 2,
           "associate's degree": 3, "bachelor's degree": 4, "master's degree": 5}
lunchTypes = {"standard": "нет", "free/reduced": "да"}
prepTypes = {"none": "нет", "completed": "да"}

for sample in examsData:
    sample[0] = genderTypes[sample[0]]
    sample[1] = raceTypes[sample[1]]
    sample[2] = edTypes[sample[2]]
    sample[3] = lunchTypes[sample[3]]
    sample[4] = prepTypes[sample[4]]

# функция, возвращающая наблюдения для конкретного показателя
def dataByType(data, type):
    res = []
    for rec in data:
        res.append(rec[type])
    return res

# Индексы числовых данных
numData = [5, 6, 7]
```

1. Анализ многомерных данных: выделение зависимости
1.2 Числовые характеристики многомерных данных

- Необходимые библиотеки: `numpy`, `matplotlib`
- Рассмотрим 3 случайные величины ξ_1, ξ_2, ξ_3 – оценки по математике, чтению, письму. Для них проводим разведывательный анализ.

```
import matplotlib.pyplot as plt
import numpy as np

# Вычисление числовых характеристик

scores = [] # массив выборочных значений для оценок
Mscores = [] # математическое ожидание оценок
for i in numData:
    scores.append(dataByType(examsData, i))
    Mscores.append(np.mean(dataByType(examsData, i)))

scoresCov = np.cov(scores) # ковариацинная матрица
scoresCorr = np.corrcoef(scores) # матрица коэффициентов корреляции
print("Математическое ожидание оценок по соответвующим предметам = " + str(Mscores))
print("Ковариационная матрица = \n" + str(scoresCov))
print("Матрица коэффициентов корреляции = \n" + str(scoresCorr))

print("Смещенная оценка дисперсии оценок по математике = " +
      str(np.var(dataByType(examsData, 5))))
print("Несмещенная оценка дисперсии оценок по математике = " +
      str(np.var(dataByType(examsData, 5), ddof = 1)))
```

1. Анализ многомерных данных: выделение зависимости
1.2 Числовые характеристики многомерных данных

```
Математическое ожидание оценок по соответвующим предметам = [66.089, 69.169, 68.054]
Ковариационная матрица =
[[229.918998   180.99895796 184.93913313]
 [180.99895796 213.1656046  211.78666066]
 [184.93913313 211.78666066 230.90799199]]
Матрица коэффициентов корреляции =
[[1.         0.81757966 0.80264205]
 [0.81757966 1.         0.95459808]
 [0.80264205 0.95459808 1.        ]]
Смещенная оценка дисперсии оценок по математике = 229.68907899999996
Несмещенная оценка дисперсии оценок по математике = 229.91899799799796
```

- Математическое ожидание многомерной случайной величины является вектором из математических ожиданий одномерных величин.
- На главной диагонали ковариационной матрицы располагаются дисперсии одномерных величин.

1. Анализ многомерных данных: выделение зависимости
1.2 Числовые характеристики многомерных данных

- По умолчанию функция `numpy.var()` считает смещенную оценку дисперсии. Использование параметра `ddof = 1` (delta degree of freedom) приводит к вычислению несмещенной оценки дисперсии.
- Стандартизированный вариант характеристики рассеивания – коэффициент корреляции. Набольшее значение коэффициента корреляции между ξ_2 **и** ξ_3 **– оценки по чтению и письму** (значительный вклад вносит линейная зависимость).
- Недиагональные элементы ковариационной матрицы и матрицы коэффициентов корреляции положительны, то есть **между оценками по любым двум предметам имеется положительная связь**: высокие значения оценки по одному предмету сопровождаются высокими значениями по другому предмету.

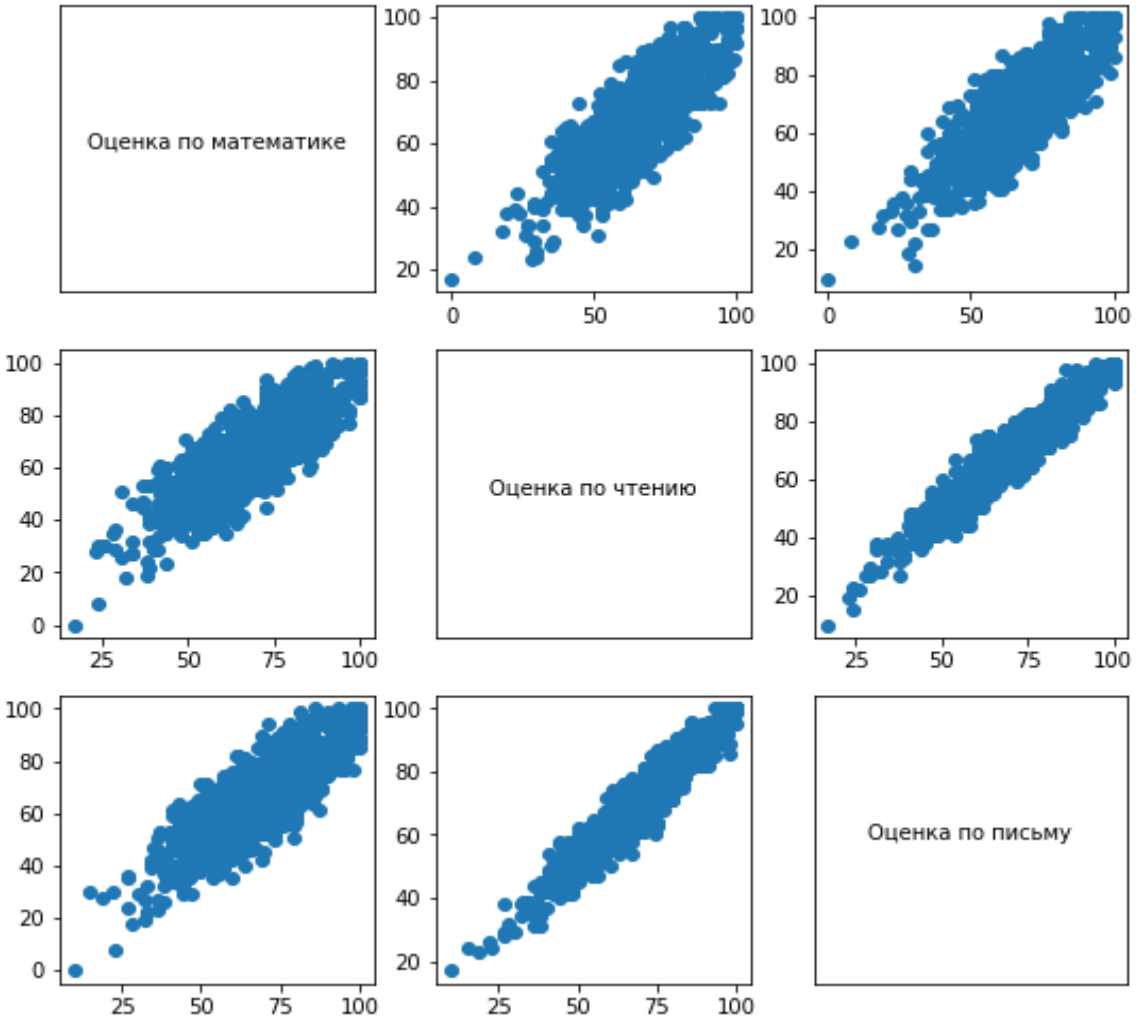
1. Анализ многомерных данных: выделение зависимости
1.3 Диаграммы рассеивания (scatterplot)

Визуализация многомерных данных – *диаграмма рассеивания* (**scatter, scatterplot**)

Каждая точка на диаграмме отвечает одному значению наблюдения за случайными векторами

```
68 # Диаграмма рассеивания для оценок по различным предметам
69 def scatterMatrix(data, ind, names):
70     fig, axes = plt.subplots(len(ind), len(ind), figsize = (3*len(ind), 3*len(ind)))
71     for i in range(len(ind)):
72         for j in range(len(ind)):
73             if (i == j):
74                 axes[i][j].annotate(types[ind[i]], (0.5, 0.5), ha = "center")
75                 axes[i][j].xaxis.set_visible(False)
76                 axes[i][j].yaxis.set_visible(False)
77             else:
78                 axes[i][j].scatter(dataByType(data, ind[i]), dataByType(data, ind[j]))
79     plt.show()
80
81 scatterMatrix(examsData, numData, types)
```

1. Анализ многомерных данных: выделение зависимости
1.3 Диаграммы рассеивания (scatterplot)



- Довольно сильная положительная связь между оценками по двум любым предметам
- Линейная зависимость между случайными величинами ξ_2 и ξ_3 (оценки по чтению и письму) наиболее явна

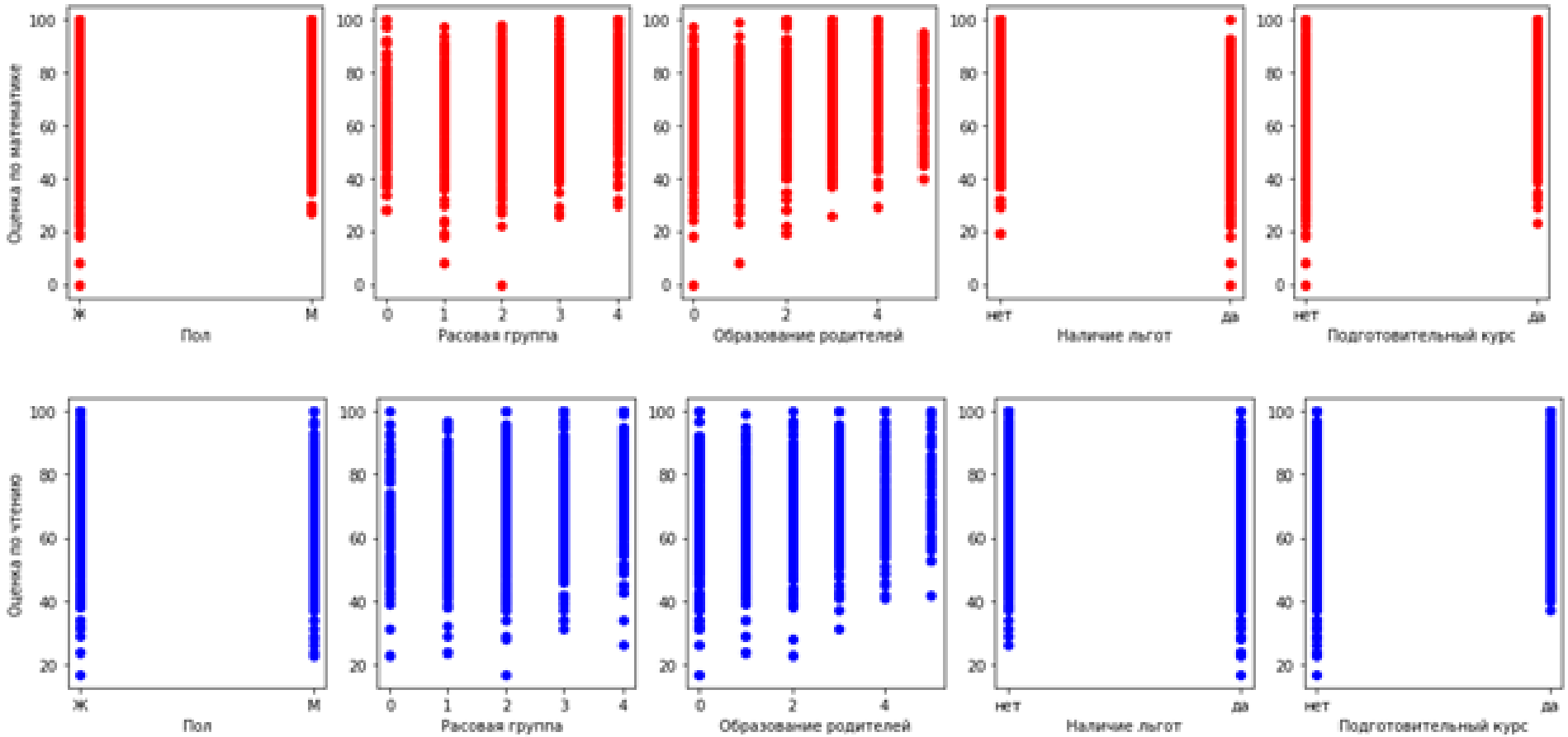
1. Анализ многомерных данных: выделение зависимости
1.3 Диаграммы рассеивания (scatterplot)

Диаграммы рассеивания можно строить не только для количественных, но и **для качественных данных**.

```
83 # Диаграммы рассеивания для качественных данных
84 colors = ['red', 'blue']
85 fig, axes = plt.subplots(2, 5, figsize = (18, 8))
86 for i in range(5):
87     for j in range(2):
88         axes[j][i].scatter(dataByType(examsData, i), dataByType(examsData, 5+j),
89             c = colors[j])
90         axes[j][i].set_xlabel(types[i])
91         if (i == 0):
92             axes[j][i].set_ylabel(types[5+j])
93 plt.show()
94
```

1. Анализ многомерных данных: выделение зависимости
1.3 Диаграммы рассеивания (scatterplot)

Результаты построения:



1. Анализ многомерных данных: выделение зависимости
1.4 Пример: данные о рейтингах покупателей ТЦ

- Имеются данные о клиентах некоторого торгового центра (ТЦ): пол, возраст, годовой доход.
- На основе информации, собранной через членские карточки клиентов, каждому клиенту выставляется рейтинговая оценка (spending score).
- Задача: формирование представлений о целевой аудитории ТЦ. Для этого необходимо первоначально провести простой корреляционный анализ для определения влияния трех наблюдаемых факторов на рейтинговую оценку покупателя.

	A	B	C	D
1	Gender	Age	Annual Income (k$)	Spending Score (1-100)
2	Male	19	15	39
3	Male	21	15	81
4	Female	20	16	6
5	Female	23	16	77
6	Female	31	17	40
7	Female	22	17	76
8	Female	35	18	6

- Источник данных: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

1. Анализ многомерных данных: выделение зависимости
1.4 Пример: данные о рейтингах покупателей ТЦ

Рассматриваем **три случайных величины:** η_1, η_2, η_3 – возраст, годовой доход и рейтинг покупателя. Найдем для них числовые характеристики

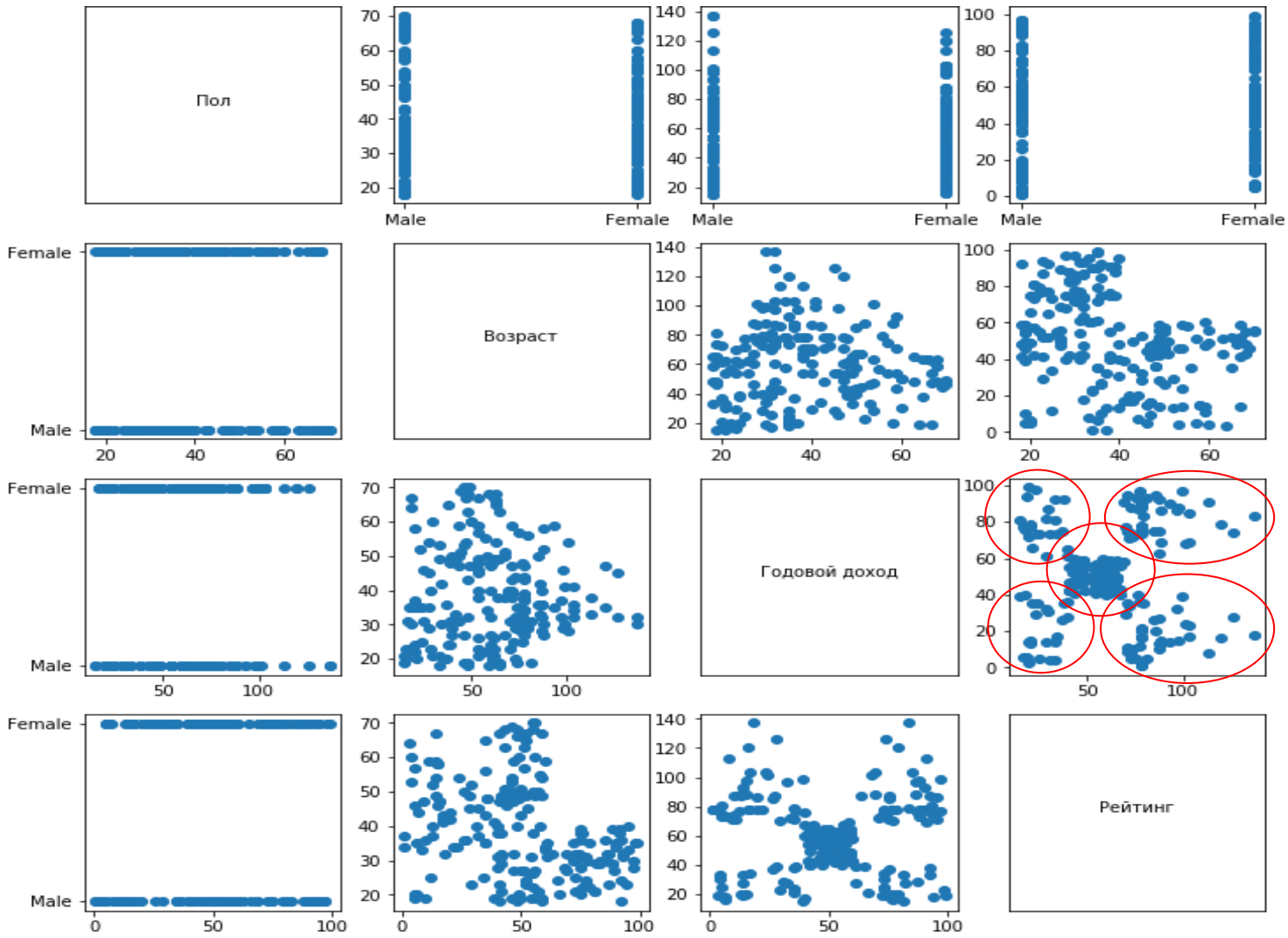
```
# ============  Данные о рейтинге покупателей =================

mallData = importData('04_КлиентыТЦ.xlsx', 0)

types = {0:"Пол", 1:"Возраст", 2:"Годовой доход", 3:"Рейтинг"}

numData = [1, 2, 3]

mallNum = []
Mmalls = []
for i in numData:
    mallNum.append(dataByType(mallData, i))
    Mmalls.append(np.mean(dataByType(mallData, i)))

mallCov = np.cov(mallNum) # ковариацинная матрица
mallCorr = np.corrcoef(mallNum) # матрица коэффициентов корреляции
print("Математическое ожидание = " + str(Mmalls))
print("Ковариационная матрица = \n" + str(mallCov))
print("Матрица коэффициентов корреляции = \n" + str(mallCorr))

scatterMatrix(mallData, [0, 1, 2, 3], types)

```

1. Анализ многомерных данных: выделение зависимости
1.4 Пример: данные о рейтингах покупателей ТЦ

```
Математическое ожидание = [38.85, 60.56, 50.2]
Ковариационная матрица =
[[ 195.13316583   -4.54874372 -118.04020101]
 [  -4.54874372  689.83557789    6.71658291]
 [-118.04020101    6.71658291  666.85427136]]
Матрица коэффициентов корреляции =
[[ 1.         -0.01239804 -0.32722685]
 [-0.01239804  1.          0.00990285]
 [-0.32722685  0.00990285  1.        ]]
```

- Возраст (η_1) имеет отрицательную умеренную корреляцию с рейтингом (η_3), а годовой доход (η_2) – положительную корреляцию с рейтингом (η_3), хотя и очень слабую
- Далее строим диаграммы рассеивания.

1. Анализ многомерных данных: выделение зависимости
1.4 Пример: данные о рейтингах покупателей ТЦ



1. Анализ многомерных данных: выделение зависимости
1.4 Пример: данные о рейтингах покупателей ТЦ

Первичные статистические выводы:

- Покупательский рейтинг несколько больше у женщин.
- С возрастом рейтинг покупателей падает.
- Торговый центр заинтересован, скорее, в покупателях, с большим годовым доходом.
- Вид диаграмм рассеивания рейтинга и годового дохода свидетельствует о существовании определенных целевых групп покупателей

2. ПРАКТИЧЕСКОЕ ЗАДАНИЕ

2. Практическое задание

Данные – из практического задания № 2: файл *02_Автоаварии.xls*

1. **Написать функции для подсчета** следующих выборочных числовых характеристик многомерных случайных величин: математическое ожидание, ковариация, ковариационная матрица, коэффициент корреляции, матрица коэффициентов корреляции.
Для подсчета каждой характеристики необходима отдельная функция. При написании **не использовать библиотечные функции** подсчета числовых характеристик.

2. **Проверить правильность работы** функций, сравнив их выходы с выходами функций библиотеки `numpy` на примере выборочных данных, полученных по результатам наблюдений за величинами χ – степень серьезности аварии, ξ – видимость дороги, ζ – скорость ветра, ς – влажность, γ – температура в момент совершения аварии.

2. Практическое задание

3. **Построить диаграммы рассеивания** для всевозможных пар случайных величин $\chi, \xi, \zeta, \varsigma, \gamma$.
Сделать выводы о влиянии величин друг на друга.

4. **Построить диаграммы рассеивания** для определения зависимости величины χ – **степень серьезности аварии от факторов, выраженных в качественных данных**: отметки о наличии вблизи места аварии лежачего полицейского, перекрестка, знака «Уступи дорогу», транспортной развязки, знака «нет выхода», железнодорожных путей, кругового движения, остановки общественного транспорта (автобусов, поездов и т.п.), знака «стоп», знаков или других мер успокоения движения, светофоров, поворотной петли.
Сделать выводы о влиянии указанных факторов на степень серьезности аварии