МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение

высшего образования

«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра «Измерительно-вычислительные комплексы»

«Методы искусственного интеллекта»

Отчёт по лабораторной работе №3

Вариант - 9

Выполнил:

студент группы ИСТбд-42

Исхаков Тимур

Ульяновск

2022

1) Создать симулированный набор данных и записать его на диск в виде csv файла.

titles = ["Табельный номер", "Фамилия И.О.", "Пол", "Год рождения", "Год начала работы в компании", "Подразделение",

"Должность", "Оклад", "Кол-во выполненных проектов"]

male\_surnames = ["Некрасов", "Михеев", "Королев", "Белоусов", "Соколов", "Устинов", "Карпов", "Куликов", "Блинов",

"Петров"]

female\_surnames = ["Васильева", "Иванова", "Морозова", "Медведева", "Соловьёва", "Сидорова", "Михайлова", "Козловская",

"Воробьева", "Гурова"]

initials = ["А.", "Б.", "В.", "Г.", "Д.", "Е.", "Ж.", "З.", "И.", "К.", "Л.", "М.", "Н.", "О.", "П.", "Р.", "С.", "Т.",

"У.", "Ф.", "Э.",

"Ю.", "Я."]

genders = ["М", "Ж"]

company\_divisions = {

"маркетинг": ["Менеджер по маркетингу", "Интернет-маркетолог"],

"разработка": ["Разработчик ПО", "Тимлид", "Веб-дизайнер", "Мобильный разработчик", "Бизнес-аналитик", "DevOps"],

"тестирование": ["Тестировщик", "QA Manager"],

"HR": ["HR Аналитик", "HR Manager"]

}

data = []

count = random.randint(1000, 1100)

with open("data.csv", "w") as file:

writer = csv.writer(file, delimiter="|", lineterminator="\n")

writer.writerow(titles)

for i in range(1, count + 1):

gender = genders[random.randint(0, 1)]

if gender == "M":

name = random.choice(male\_surnames) + " " + random.choice(initials) + random.choice(initials)

else:

name = random.choice(female\_surnames) + " " + random.choice(initials) + random.choice(initials)

year\_of\_birth = random.randint(1975, 1995)

year\_of\_joining = random.randint(year\_of\_birth + random.randint(18, 26), 2022)

division = random.choice(list(company\_divisions.keys()))

job\_title = random.choice(company\_divisions[division])

if year\_of\_joining < 2015:

projects = random.randint(8, 12)

salary = random.randrange(60000, 80000, 5000)

elif year\_of\_joining > 2019:

projects = random.randint(1, 3)

salary = random.randrange(20000, 40000, 5000)

else:

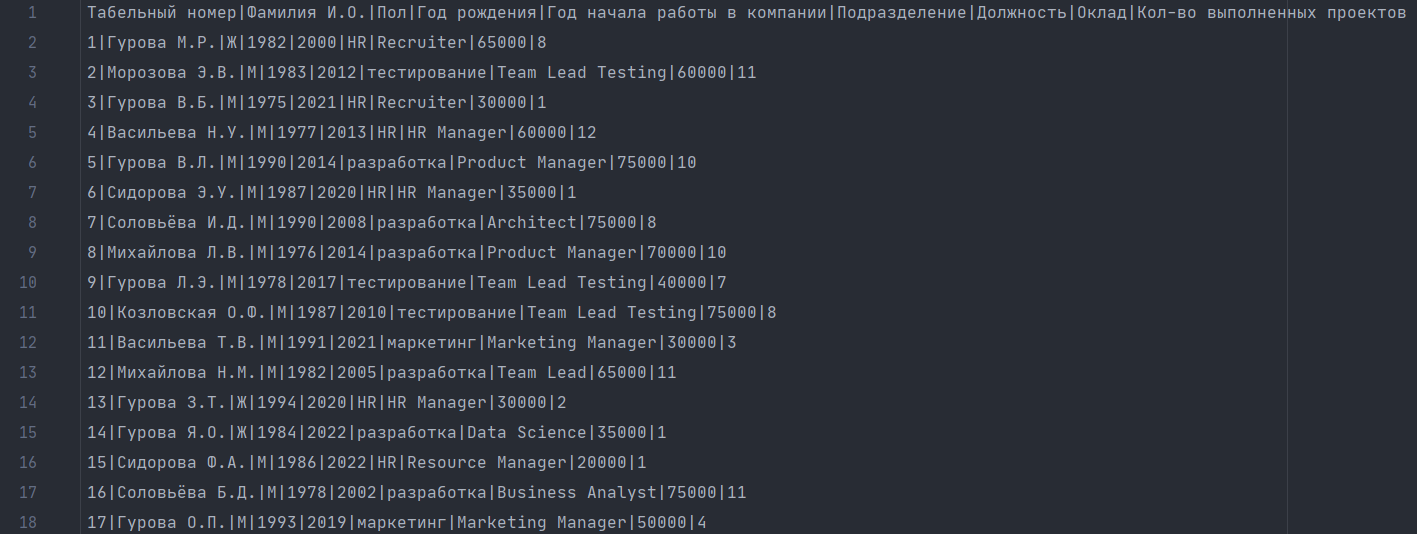
projects = random.randint(4, 7)

salary = random.randrange(40000, 60000, 5000)

employee = [i, name, gender, year\_of\_birth, year\_of\_joining, division, job\_title, salary, projects]

writer.writerow(employee)

Полученный csv-файл:



2) Прочитать сгенерированный набор данных в виде списков и получить с помощью программирования и методов библиотеки numpy для разных по типу признаков столбцов основные статистические характеристики.

with open("data.csv", "r") as file:

years\_of\_birth = []

salaries = []

projects = []

genders = []

reader = csv.reader(file, delimiter="|", lineterminator="\n")

headers = next(reader)

for employee in reader:

years\_of\_birth.append(int(employee[3]))

genders.append(employee[2])

salaries.append(int(employee[7]))

projects.append(int(employee[8]))

def process\_data\_numpy(data\_list, column):

print('-------------Использование numpy-------------')

print('Столбец: ', column)

print('Максимальное значение: ', numpy.max(data\_list))

print('Минимальное значение: ', numpy.min(data\_list))

print('Математическое ожидание: ', numpy.mean(data\_list))

print('Стандартное отклонение: ', numpy.std(data\_list))

print('Дисперсия: ', numpy.var(data\_list))

print('Медиана: ', numpy.median(data\_list))

print('Мода: ', statistics.mode(data\_list))

process\_data\_numpy(salaries, titles[7])

process\_data\_numpy(years\_of\_birth, titles[3])

process\_data\_numpy(projects, titles[8])

male\_employees = numpy.sum(genders.count('М'))

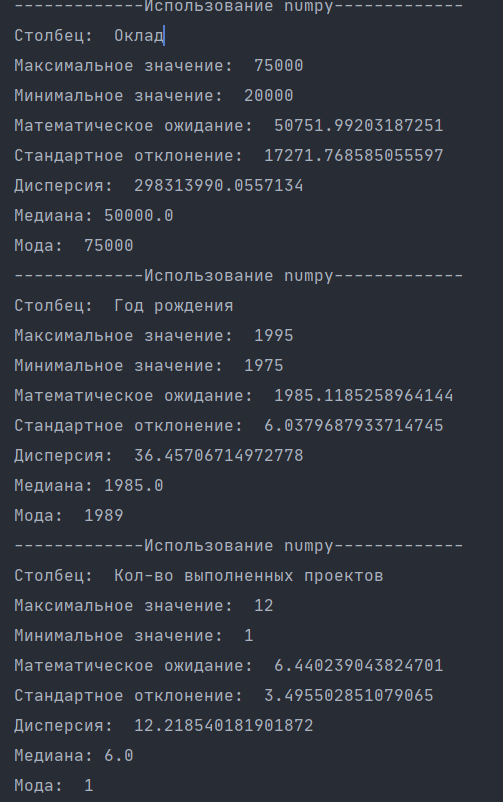
female\_employees = numpy.sum(genders.count('Ж'))

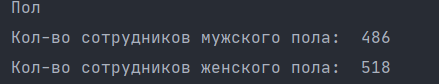
print('Пол')

print('Кол-во сотрудников мужского пола: ', male\_employees)

print('Кол-во сотрудников женского пола: ', female\_employees)

Результат работы:





3) Прочитать сгенерированный набор данных в виде датафрейма и получить с помощью методов библиотеки pandas для тех же столбцов те же статистические характеристики.

def process\_data\_pandas(df, column):

print('----------------Использование pandas--------------')

print('Столбец: ', column)

print('Максимальное значение: ', df[column].max())

print('Минимальное значение: ', df[column].min())

print('Математическое ожидание: ', df[column].mean())

print('Стандартное отклонение: ', df[column].std())

print('Дисперсия: ', df[column].var())

print('Медиана: ', df[column].median())

print('Мода: ', df[column].mode())

# работа с pandas

dataframe = pandas.read\_csv("data.csv", delimiter="|", lineterminator="\r", header=0, index\_col=0, encoding="cp1251")

process\_data\_pandas(dataframe, "Оклад")

process\_data\_pandas(dataframe, "Год рождения")

process\_data\_pandas(dataframe, titles[8])

gender = dataframe['Пол']

male\_employees\_pandas = gender.value\_counts()['М']

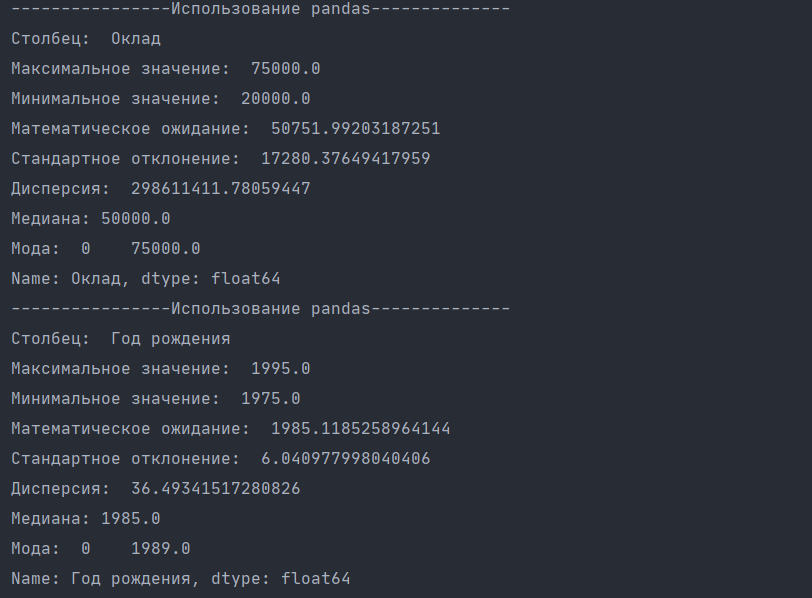
female\_employees\_pandas = gender.value\_counts()['Ж']

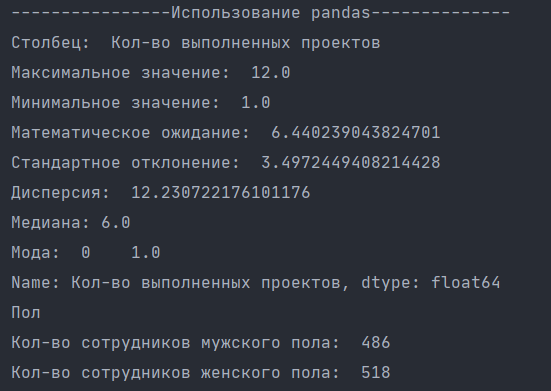
print('Пол')

print('Кол-во сотрудников мужского пола: ', male\_employees\_pandas)

print('Кол-во сотрудников женского пола: ', female\_employees\_pandas)

Результат работы:





4) Построить не менее 3 разнотипных графиков.

graf = dataframe['Подразделение'].hist()

plt.xlabel('Подразделение')

plt.ylabel('Кол-во сотрудников')

plt.xticks(rotation=90)

plt.title('Распределние работников по отделам')

plt.show()

# Вывод графика зависимости Оклада от года начала работы в компании

plt.subplot(2, 2, 2)

plt.xlabel("Количество выполненных проектов")

plt.ylabel("Оклад")

plt.plot(dataframe["Кол-во выполненных проектов"].sort\_values(),

dataframe["Оклад"].sort\_values())

plt.show()

# Вывод диаграммы полов в компании

data = [dataframe["Пол"].value\_counts()["М"],

dataframe["Пол"].value\_counts()["Ж"]]

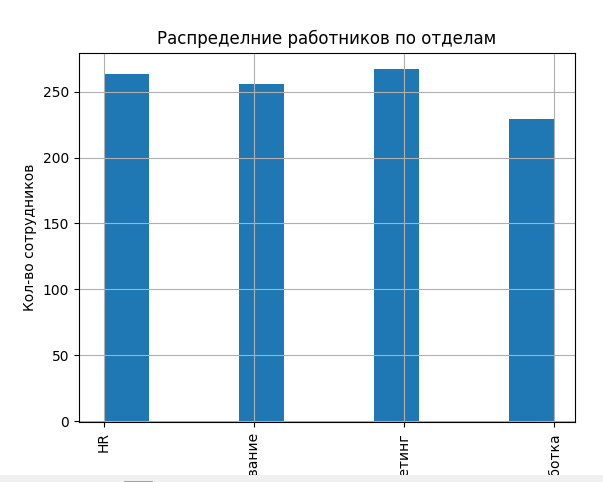
plt.pie(data, labels=["M", "Ж"])

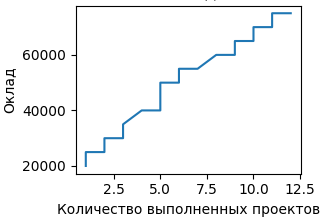
plt.title("Круговая диаграмма полов в компании")

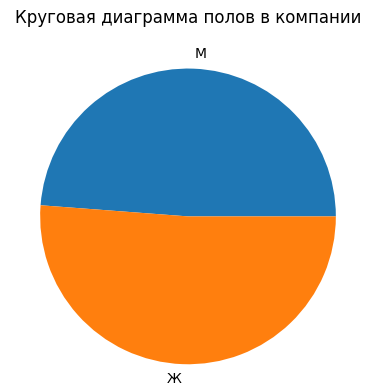
plt.ylabel("")

plt.show()

Результат работы:







5)

Библиотека CSV позволяет работать с csv файлами удобным образом с помощью различных методов, так в лабораторной работе были использованы методы writer и reader для записи и чтения набора данных из файла, был применен специальный разделитель. Также с помощью библиотеки можно построчно считывать и записывать данные, устанавливать нужную кодировку при записи/чтении. В целом библиотека достаточно проста и удобна в применении.

Модуль numpy удобен для работы при анализе данных, он достаточно быстрый и содержит множество методов для обработки больших массивов информации, имеются методы для расчета статистических характеристик (минимум, максимум, среднее, дисперсия, стандартное отклонение, медиана, мода). Центральный объект в NumPy является массив, поэтому взаимодействие со списками в Python происходит легко и понятно.

Модуль Pandas также был использован для анализа и обработки данных. Библиотека удобна при работе со статистическими данными, с помощью нее взаимодействие с ними можно производить как с таблицами в виде датафреймов, которые упрощают работу с полученной информацией. Использование датафреймов позволяет обращаться к отдельным столбцам или строкам таблицы.