Лабораторная работа №3

«Исследование библиотек CSV, pandas»

Полученный csv (отрывок):

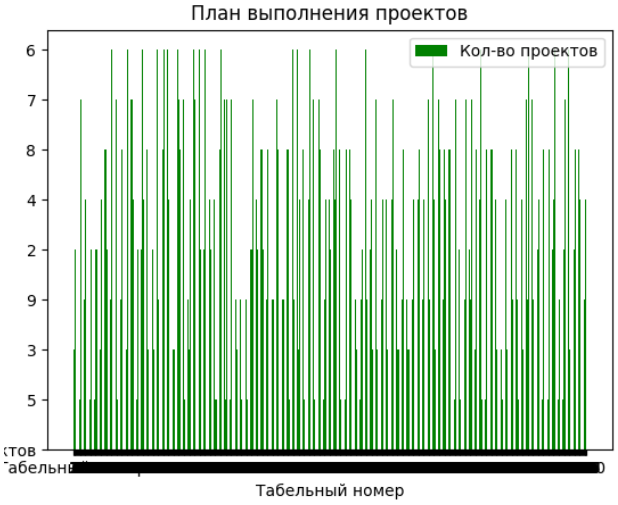
Console × data.csv × +

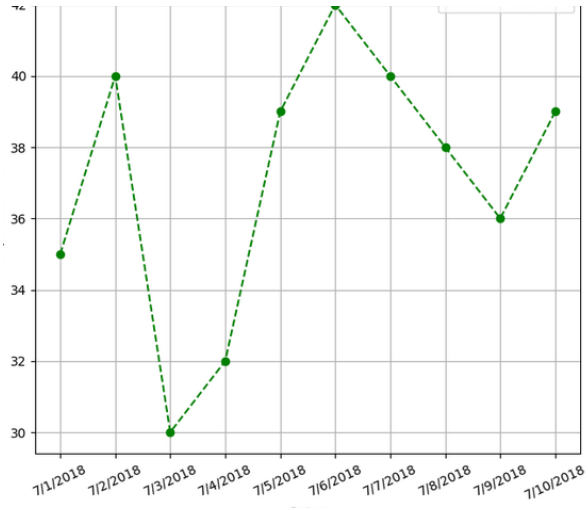
```
Табельный номер,Фамилия И. О.,Пол,Год рождения,Начало работы (год),Подразделение,Должность,Оклад,Кол-во проектов
1,Э. Г. Буровкин,М,1980,2012,Административный отдел,Деканат,45381,7
2,С. М. Сокортов,М,2001,2016,Научный отдел,Ректорат,45096,2
3,Г. Н. Спиченкова,Ж,1994,2009,Административный отдел,Ректорат,45565,7
4,В. П. Хамутский,М,1987,2019,Обслуживающий персонал,Бухгалтер,45116,6
5,Е. М. Буригина,Ж,1988,2020,Технический отдел,Бухгалтер,45865,3
6,С. М. Хажиева,Ж,1974,2006,Отдел кадров,Деканат,45647,9
7,В. М. Терлеева,Ж,1982,2014,Технический отдел,Юрист,45825,3
8,С. Р. Капашова,Ж,1986,2018,Отдел маркетинга,Преподаватель,45931,5
9,Г. А. Шинелев,М,2006,2021,Отдел кадров,Аспирант,45407,9
10,Д. Т. Пульсов,М,1994,2009,Административный отдел,Юрист,45293,7
11,А. П. Алючаева,Ж,2000,2015,Отдел кадров,Юрист,45503,9
12,К. Э. Завзятова,Ж,1986,2018,Административный отдел,Юрист,45965,7
13,В. Л. Корбина,Ж,1982,2014,Отдел кадров,Аспирант,45519,9
14,В. О. Рассветов,М,1978,2010,Отдел кадров,Преподаватель,45175,9
15,А. М. Мырылева,Ж,1985,2017,Научный отдел,Бухгалтер,45760,2
16,Н. В. Армеев,М,1983,2015,Обслуживающий персонал,Ректорат,45180,6
17,В. Я. Вильдяев,М,1990,2005,Административный отдел,Аспирант,45085,7
18,Е. Ф. Небошина,Ж,1992,2007,Отдел кадров,Аспирант,45631,9
19,А. Б. Нуждинова,Ж,1996,2011,Отдел маркетинга,Аспирант,45771,5
20,Д. Л. Коняев,М,1987,2019,Исследовательский отдел,Бухгалтер,45354,4
21,Д. А. Репневский,М,2003,2018,Исследовательский отдел,Деканат,45234,4
22,П. В. Алушпов,М,2001,2016,Бухгалтерский отдел,Юрист,45062,8
23,Ф. О. Хубларов,М,1977,2009,Исследовательский отдел,Бухгалтер,45242,4
24,В. Б. Кузоватов,М,1983,2015,Научный отдел,Деканат,45248,2
25,Д. А. Ашелькин,М,1975,2007,Бухгалтерский отдел,Ассистент,45478,8
26,Н. В. Тарабукин,М,1980,2012,Отдел кадров,Юрист,45279,9
27,С. Д. Шушковская,Ж,2001,2016,Научный отдел,Преподаватель,45504,2
28,Н. В. Подчапаева,Ж,2000,2015,Отдел кадров,Ректорат,45775,9
29,Д. Т. Винюков,М,1985,2017,Научный отдел,Ректорат,45488,2
30,А. Е. Грицков,М,1989,2021,Обслуживающий персонал,Ассистент,45492,6
31,К. А. Кружилина,Ж,1977,2009,Обслуживающий персонал,Бухгалтер,45956,6
32,С. Н. Латифова,Ж,1989,2021,Обслуживающий персонал,Аспирант,45764,6
33,Г. Э. Франтова,Ж,2006,2021,Отдел кадров,Преподаватель,45679,9
34,А. О. Алина,Ж,1977,2009,Исследовательский отдел,Юрист,45650,4
35,Л. В. Даньшова,Ж,1990,2005,Технический отдел,Аспирант,45561,3
36,З. Н. Форшенева,Ж,1988,2020,Технический отдел,Ректорат,45593,3
37,О. П. Скопцова,Ж,1994,2009,Административный отдел,Аспирант,45701,7
38,Ю. Е. Квицинская,Ж,1973,2005,Исследовательский отдел,Бухгалтер,45578,4
39,Т. В. Биккузина,Ж,2006,2021,Технический отдел,Юрист,45713,3
```

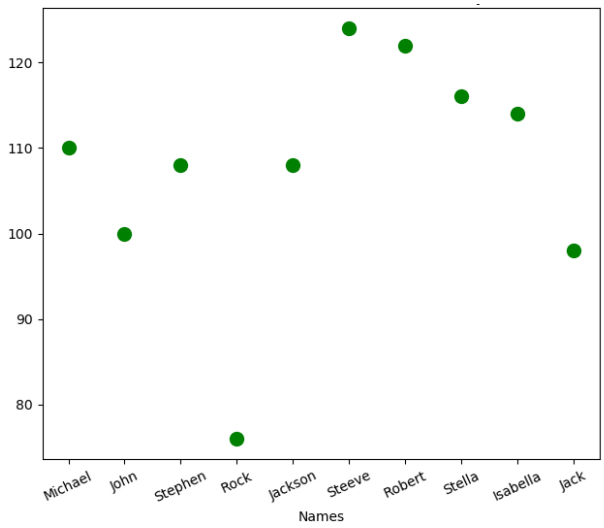
Вывод по 2 части:

```
---------------- Основная информация ----------------
Количество сотрудников: 1000 человек
Максимальный оклад: 45999 рублей
Минимальный оклад: 45000 рублей
Бюджет компании на ЗП: 45499500 рублей
Среднее кол-во проектов на 1го человека: 5.5 проекта
Медиана возраста: 1989.0 год

       Табельный номер  Год рождения         Оклад  Кол-во проектов
count      1000.000000   1000.000000   1000.000000      1000.000000
mean        500.500000   1989.360000  45499.500000         5.500000
std         288.819436      9.828886    288.819436         2.292434
min           1.000000   1973.000000  45000.000000         2.000000
25%         250.750000   1981.000000  45249.750000         3.750000
50%         500.500000   1989.000000  45499.500000         5.500000
75%         750.250000   1998.000000  45749.250000         7.250000
max        1000.000000   2006.000000  45999.000000         9.000000
```

Полученные графики:







Оценка:

NumPy

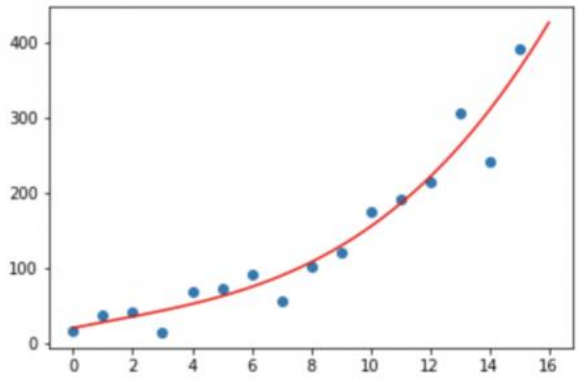
NumPy позволяет очень эффективно обрабатывать многомерные массивы. Многие другие библиотеки построены на NumPy, и без неё было бы невозможно использовать pandas, Matplotlib, SciPy или scikit-learn — именно поэтому она занимает первое место в списке.

```
In [1]: import numpy as np

In [2]: a = np.arange(12).reshape(2, 2, 3)

In [3]: a

Out[3]: array([[[ 0,  1,  2],
                [ 3,  4,  5]],

               [[ 6,  7,  8],
                [ 9, 10, 11]]])
```

Трёхмерный массив в NumPy. Источник: Data36

Также в ней есть несколько хорошо реализованных методов, например, функция random, которая гораздо качественнее модуля случайных чисел из стандартной библиотеки. Функция polyfit отлично подходит для простых задач по прогнозной аналитике, например, по линейной или полиномиальной регрессии.

```
In [31]: coefs = np.polyfit(x,y,1)
         predict = np.poly1d(coefs)

In [32]: x_test = np.linspace(0,16)
         y_pred = predict(x_test[:,None])
         plt.scatter(x,y)
         plt.plot(x_test,y_pred,c='r')
         plt.show()
```



Прогнозирование с использованием функции polyfit. Источник: Data36

pandas

Аналитики данных обычно используют плоские таблицы, такие, как в SQL и Excel. Изначально в Python такой возможности не было. Библиотека pandas позволяет работать с двухмерными таблицами на Python.

```
In [6]: super_tree.head()
```

Out[6]:

	day	my_date	user_id	event_type
0	day_1	2017-12-01	1000007	sent_a_super_tree
1	day_1	2017-12-01	1000010	sent_a_super_tree
2	day_1	2017-12-01	1000011	sent_a_super_tree
3	day_1	2017-12-01	1000019	sent_a_super_tree
4	day_1	2017-12-01	1000022	sent_a_super_tree

Таблица в pandas. Источник: Data36

Эта высокоуровневая библиотека позволяет строить сводные таблицы, выделять колонки, использовать фильтры по параметрам, выполнять группировку по параметрам, запускать функции (сложение, нахождение медианы, среднего, минимального, максимального значений), объединять таблицы и многое другое. В pandas можно создавать и многомерные таблицы.

Matplotlib

Визуализация данных позволяет представить их в наглядном виде, изучить более подробно, чем это можно сделать в обычном формате, и доступно изложить другим людям. Matplotlib — лучшая и самая популярная Python-библиотека для этой цели. Она не так проста в использовании, но с помощью 4-5 наиболее распространённых блоков кода для простых линейных диаграмм и точечных графиков можно научиться создавать их очень быстро.

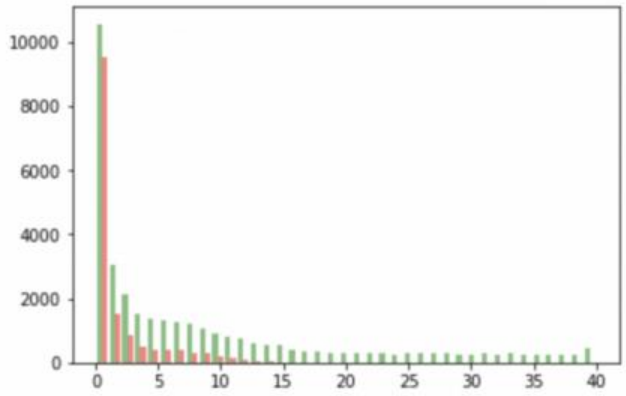
Файл **CSV** (значения, разделённые запятыми) является по сути текстовым файлам, данные в котором разделены с помощью специального разделителя. В качестве разделителя обычно используется запятая или символ «;», но могут использоваться и другие разделители. Каждая новая запись в файле CSV начинается с новой строки.

Формат файлов CSV можно легко экспортировать в электронные таблицы или базы данных.

VISUALIZATION

In [14]:
```
android = big_table[big_table.phone_type == 'android'].reset_index()
ios = big_table[big_table.phone_type == 'ios'].reset_index()
```

In [15]:
```
bins = np.linspace(0, 40, 40)
x = android['free']
y = ios['free']
data = [x,y]
plt.hist(data, bins, alpha = 0.5, color = ['g','r'])
plt.show()
```



Пример визуализации данных в Matplotlib. Источник: Data36