



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

ВИЗУАЛИЗАЦИЯ

Теванян Элен

19.04.2019

Москва 2019

The background of the slide features a large, abstract blue shape resembling a wave or a splash, set against a white background with smaller blue dots and splatters.

Визуализация: зачем?



ОКЛИ ТАК?

Расходы по проекту	6 270 044	6 456 628	8 665 664	6 789 156	332 527	-1 876 509	24 333 646	25 607 508	24 333 998	352	-1 273 510
Прямые расходы	6 270 044	6 456 628	8 665 664	6 789 156	332 527	-1 876 509	24 333 646	25 607 508	24 333 998	352	-1 273 510
00.01.00 Заработкая плата основного персонала	3 290 559	3 174 576	5 542 452	3 387 378	212 802	-2 155 074	12 497 403	16 316 907	12 864 520	367 117	-3 452 387
00.02.00 Заработкая плата персонала не участвующего в производстве продуктов реализации	626 031	660 000	635 887	454 070	-205 930	-181 816	2 202 449	1 936 628	1 925 396	-277 053	-11 231
00.11.00 Налоги на заработную плату	1 217 001	1 162 706	895 860	1 015 801	-146 905	119 941	4 637 623	2 646 690	4 600 562	-37 061	1 953 872
17.23.10 Бумажные канцелярские принадлежности	0	22 800	34 338	91 227	68 427	56 889	80 599	101 399	118 626	38 027	17 227
55.10.00 Услуги гостиниц и аналогичные услуги по предоставлению временного жилья			387	0	0	-387	0	387	0		
58.14.12 Печатный бизнес, проф.и академ.журналы и периодич.издания	0	7 500	14 488	51 132	43 632	36 644	24 489	43 993	58 121	33 632	14 128
61.20.11 Услуги мобильной телекоммуникационной связи – доступ и пользование	13	900	2 823	4	-896	-2 819	2 140	8 559	59	-2 082	-8 501
62.01.11 Услуги по проектированию и разработке информационных технологий для прикладных задач	472 134	587 705	664 343	454 236	-133 469	-210 107	2 174 849	1 951 038	1 899 857	-274 992	-51 181
64.19.30 Прочие услуги по посредничеству в денежно-кредитной сфере, не включенные в другие группировки	0	0	67	0	0	-67	0	67	0	0	-67
68.20.12 Услуги по сдаче в аренду (внаем) собственных или арендованных нежилых помещений	542 880	407 160	384 832	407 160	0	22 328	1 418 040	1 146 212	1 418 040	0	271 828
70.22.11 Консультативные услуги по вопросам стратегического управления	0	15 000	0	0	-15 000	0	54 000	0	19 000	-35 000	19 000
73.11.1 Услуги, предоставляемые рекламными агентствами	0	0	317	0	0	-317	0	317	0	0	-317
77.29.12 Услуги по аренде и лизингу мебели и прочих бытовых приборов	105 902	300 000	222 972	906 743	606 743	683 772	901 285	656 123	1 321 110	419 826	664 987
78.10.1 Услуги, предоставляемые агентствами по трудоустройству	0	35 593	52 552	0	-35 593	-52 552	83 050	154 747	0	-83 050	-154 747
82.19.13 Услуги по подготовке документов и прочие услуги по специализированному конторскому обслуживанию	17 522	18 338	27 817	17 604	-734	-10 213	66 268	79 065	58 606	-7 662	-20 459
84.13.18 Общие административные услуги, связанные с вопросами экономики, торговли и рабочей силы	-2 000	45 000	145 214	0	-45 000	-145 214	146 300	441 348	46 300	-100 000	-395 048
85.41.14 Прочие услуги в области посредничного невысшего тех.и проф.образования	0	11 850	28 196	3 800	-8 050	-24 396	27 650	86 474	3 800	-23 850	-82 674
94.11.10 Услуги, предоставляемые коммерческими и предпринимательскими организациями	0	7 500	13 120	0	-7 500	-13 120	17 500	37 553	0	-17 500	-37 553

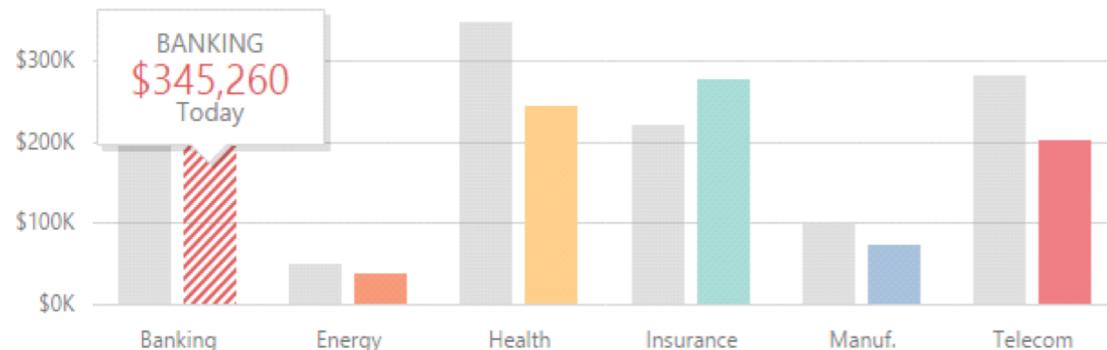


ОДНА КАРТИНКА ВМЕСТО 1000 СЛОВ

DAILY SALES PERFORMANCE

Today \$303K Yesterday \$1,279K Last Week \$7,896K

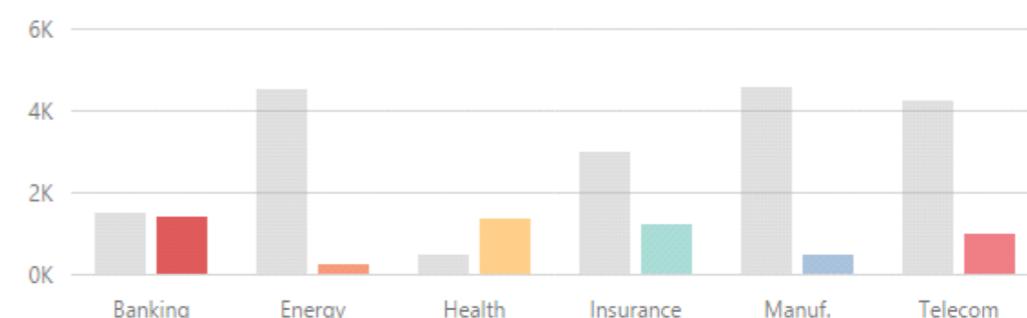
OCT 11, 2014



UNIT SALES BY SECTOR

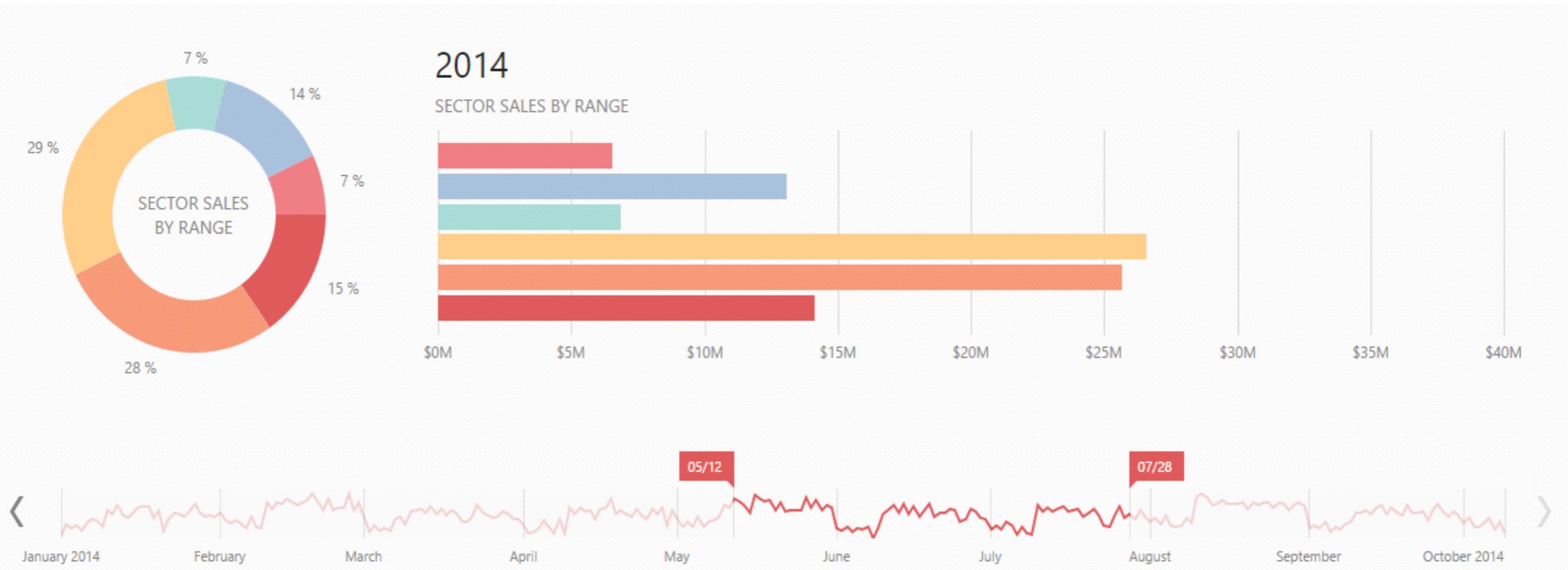
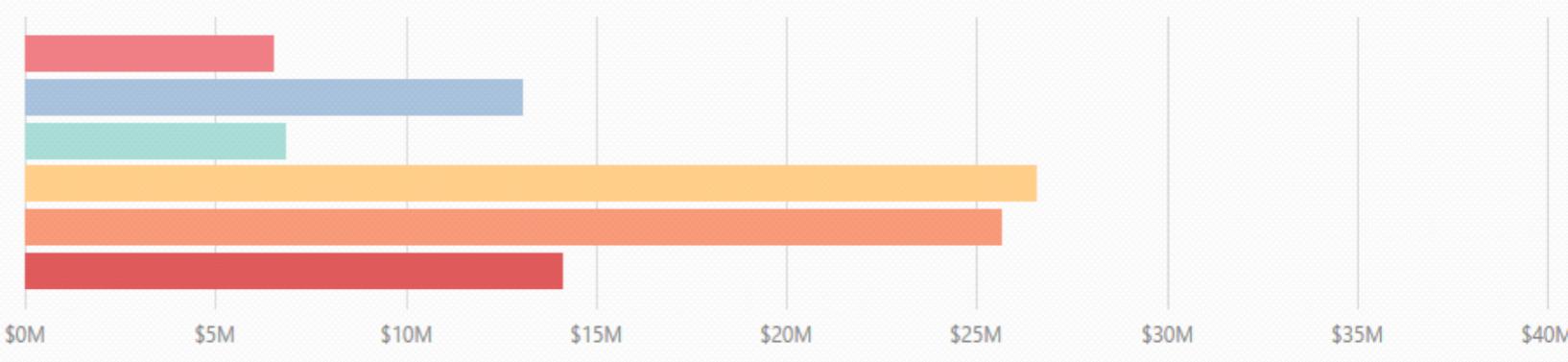
This Month 5,771 Last Month 18,224 YTD 171,063

OCTOBER



2014

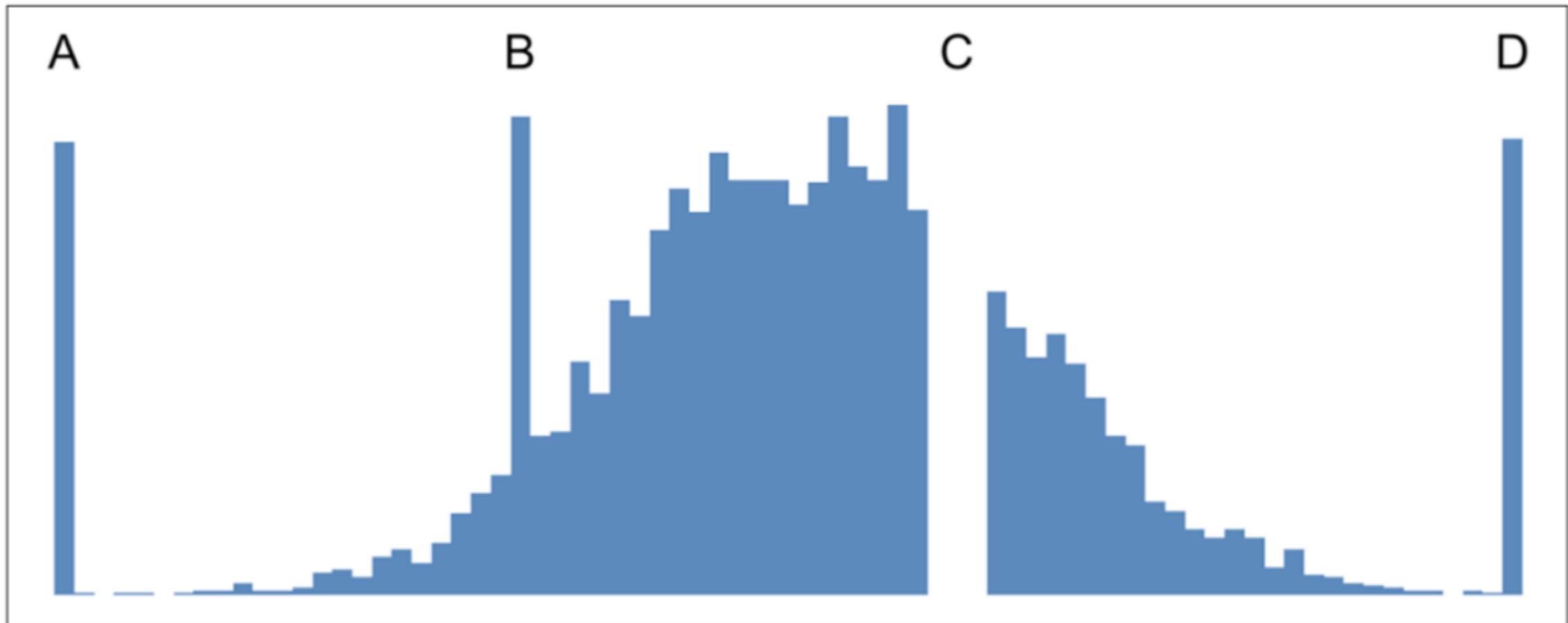
SECTOR SALES BY RANGE

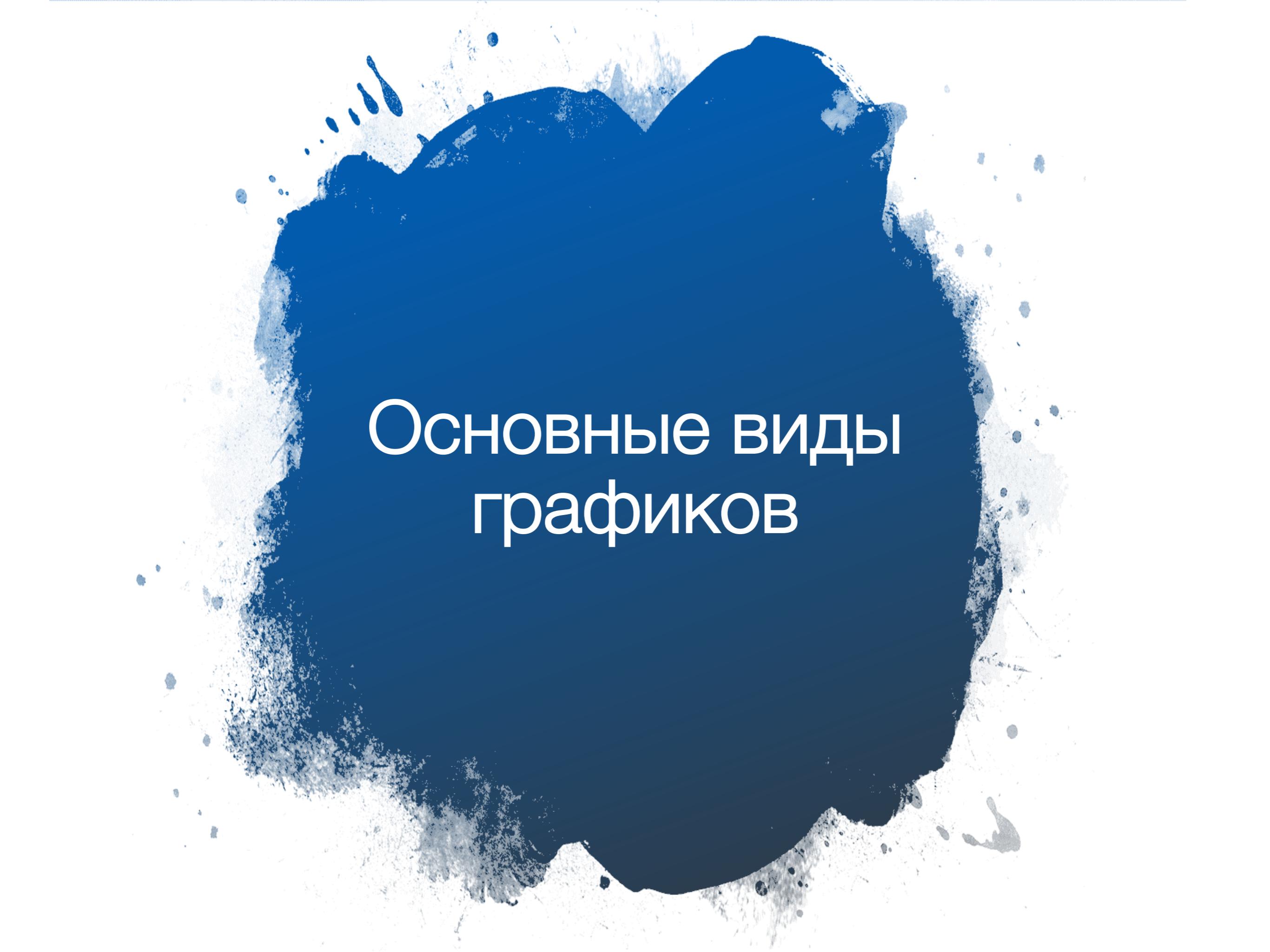




РАЗВЕДОЧНЫЙ АНАЛИЗ

- Найти паттерны в данных, сформулировать гипотезы о новых закономерностях



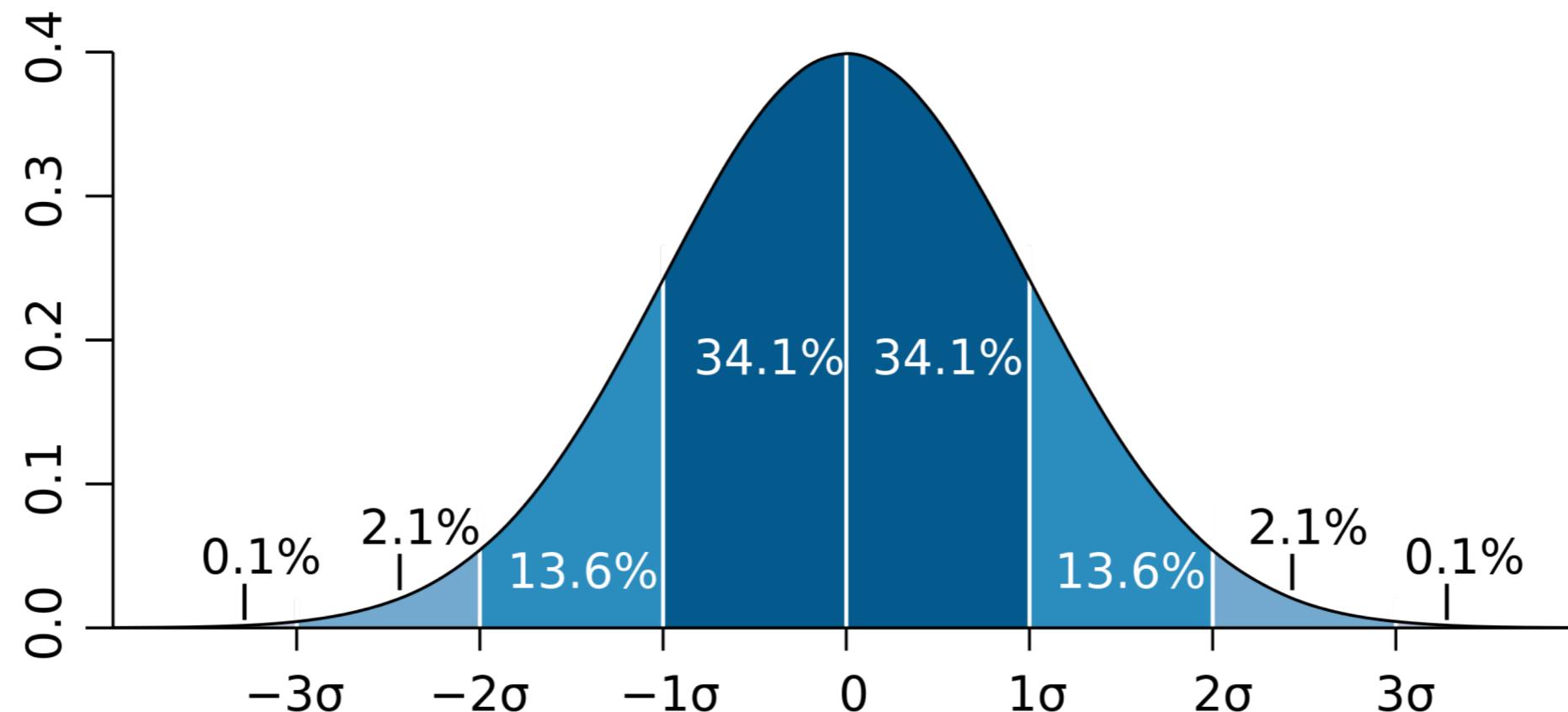


Основные виды графиков



А ЧТО ТАКОЕ РАСПРЕДЕЛЕНИЕ?

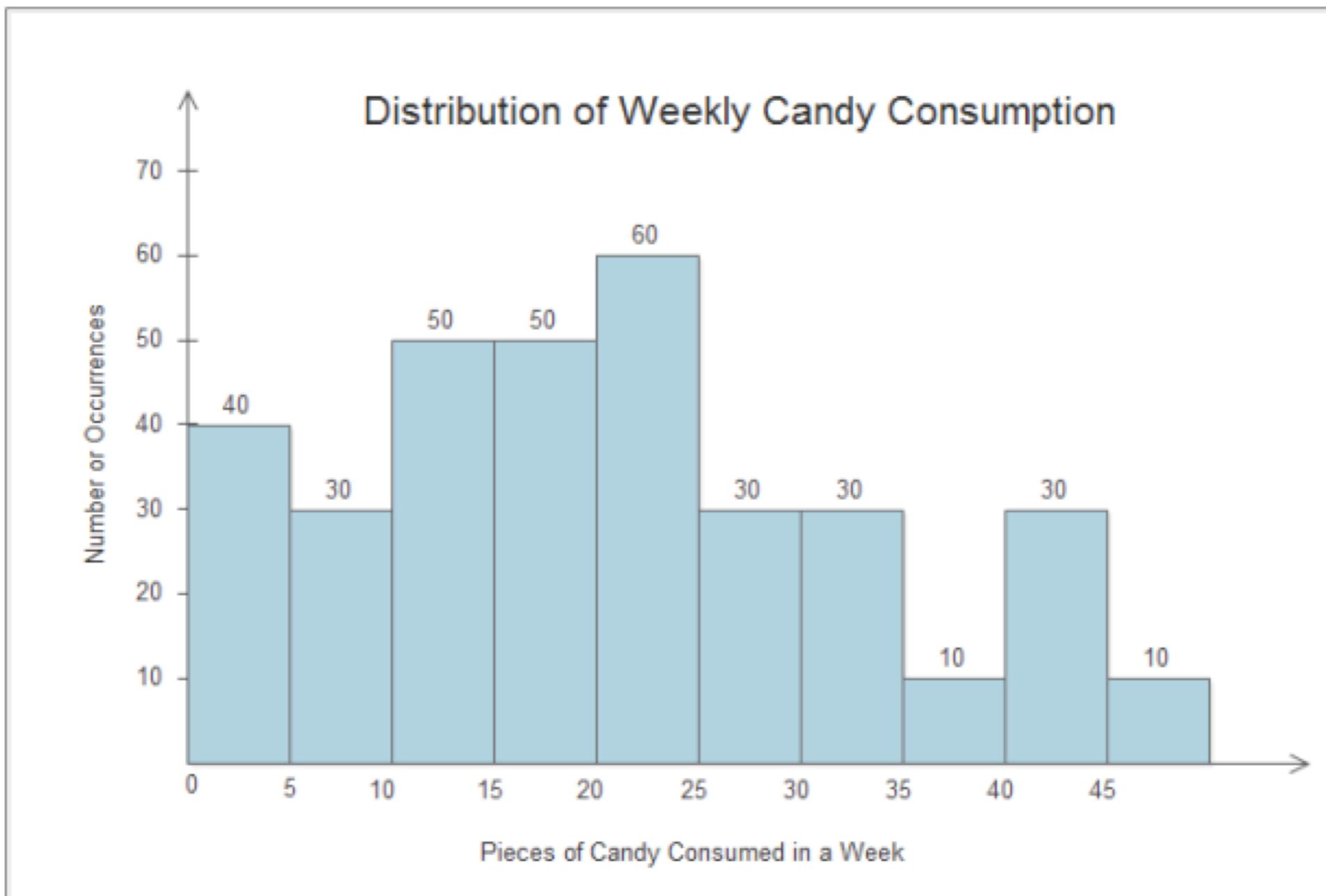
ряд значений, которые принимает изучаемая нами величина





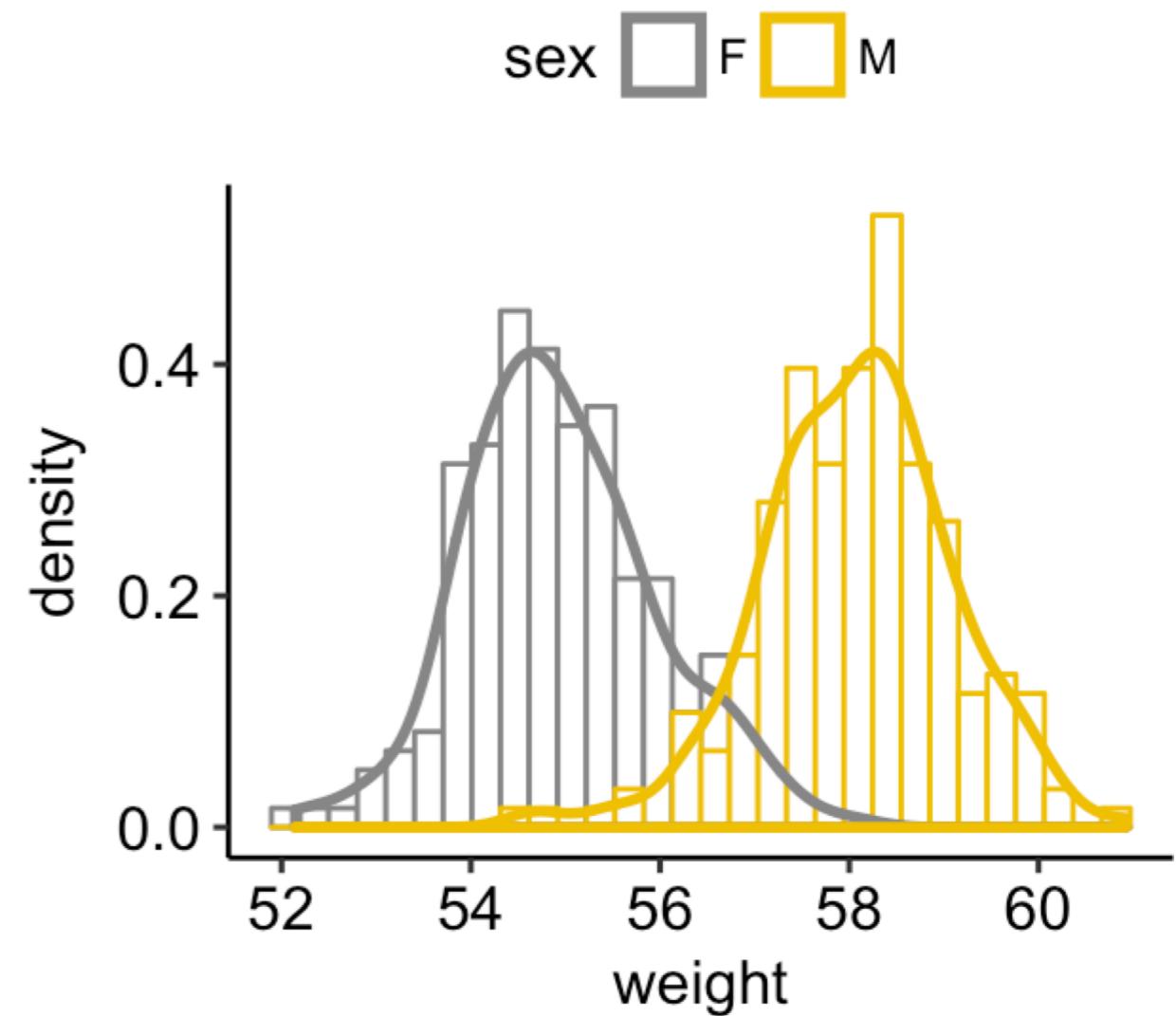
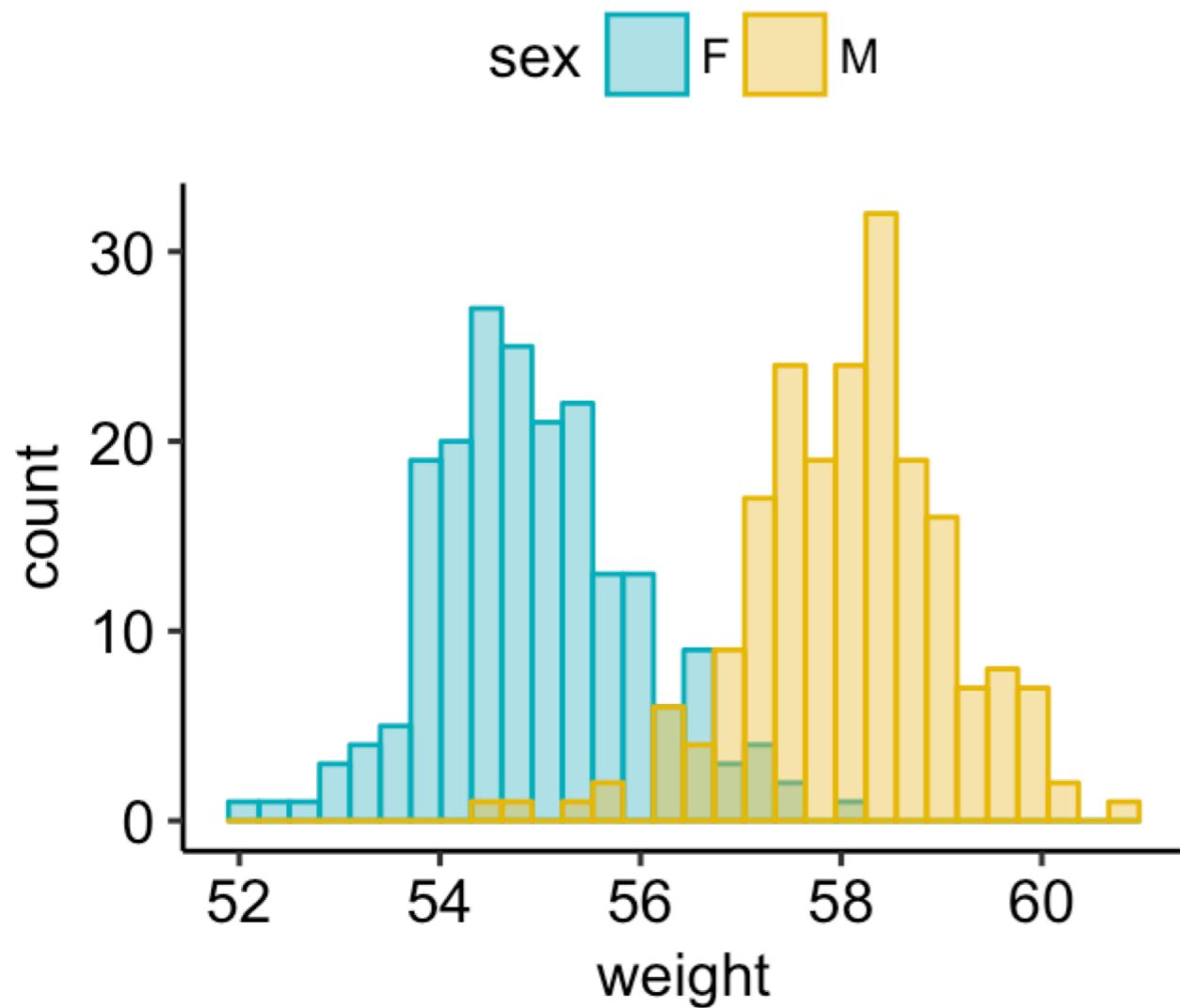
ГИСТОГРАММА

- Отлично подойдет для набора числовых данных
- Наиболее часто используется





ГИСТОГРАММА





ТИПЫ ГРАФИКОВ: ГИСТОГРАММА

- Отлично подходит для демонстрации распределения переменной
- Можно сравнить с нормальным распределением
- Ось ординат практически всегда используется как частота: легко интерпретировать
- Подходит для любого количества наблюдений



ГИСТОГРАММА

- Используется только для количественных признаков
- Часто показывает количество значений для интервалов, но не для конкретных значений

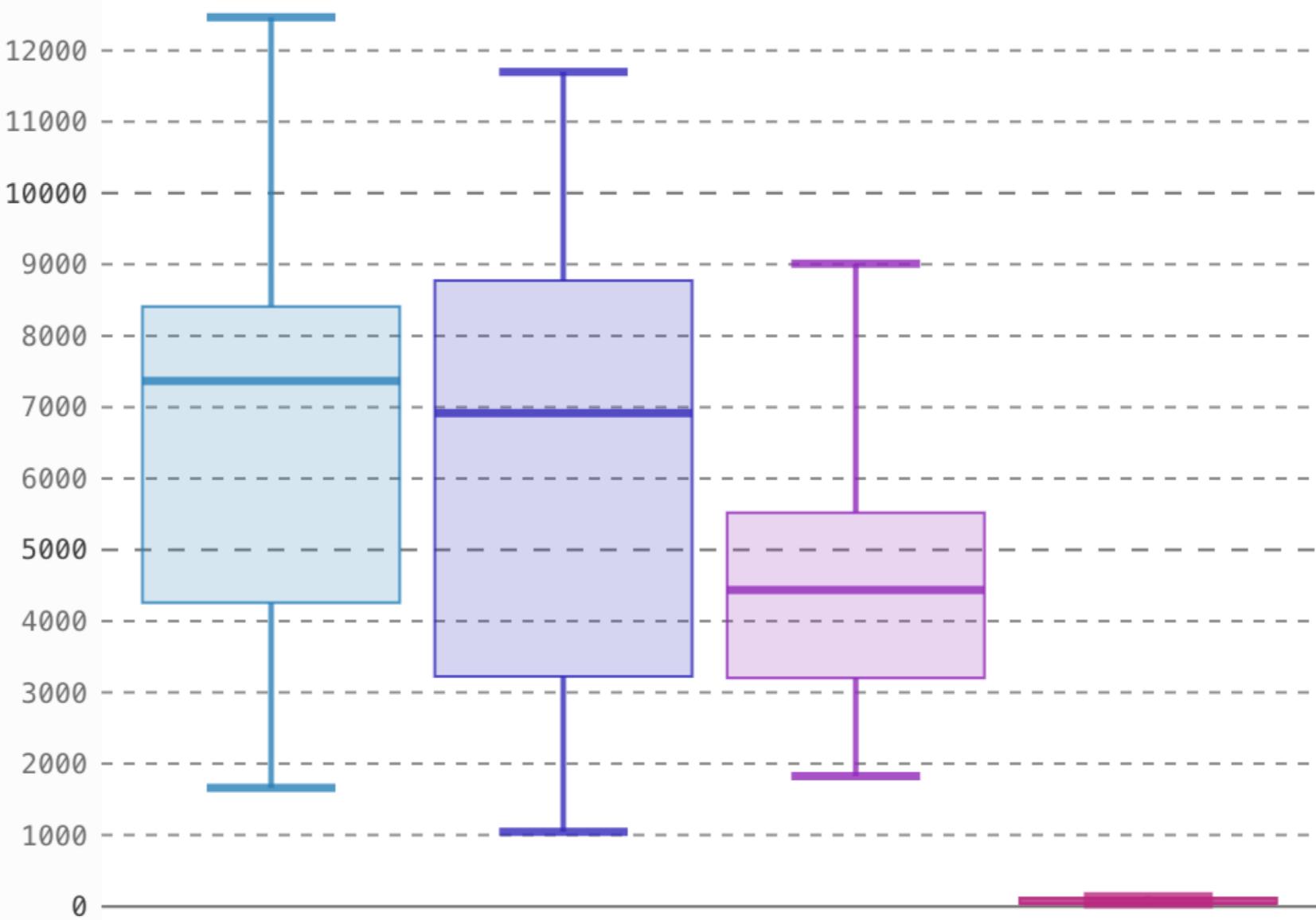


ЯЩИК С УСАМИ (BOXPLOT)



V8 benchmark results

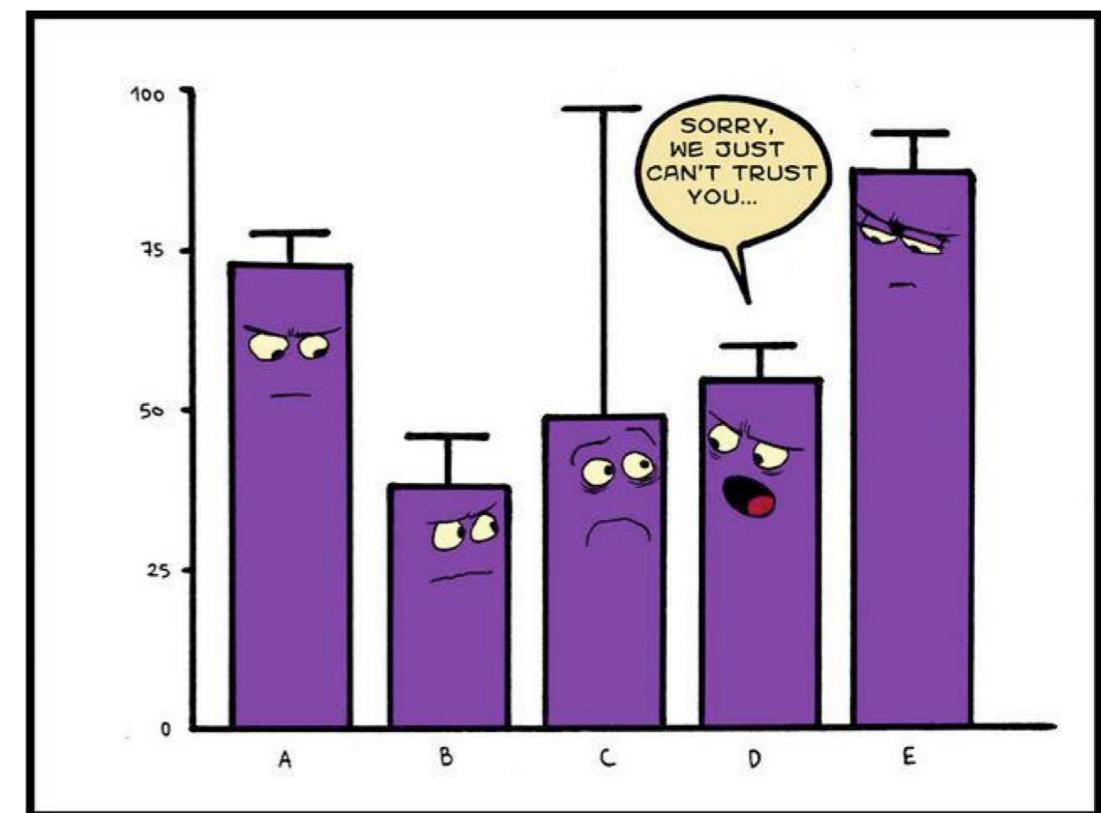
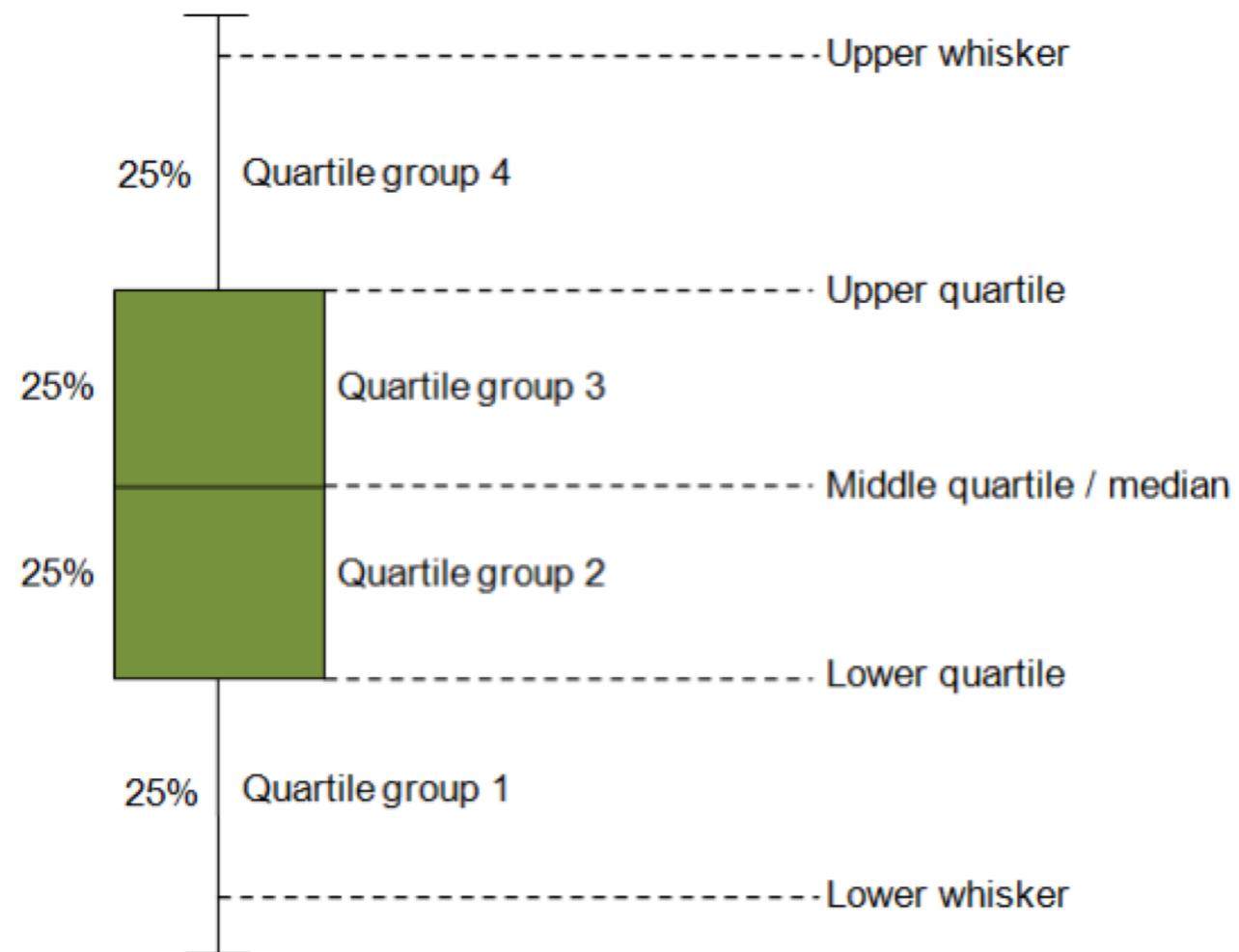
Chrome
Firefox
Opera
IE





ЯЩИК С УСАМИ

Data





ЯЩИК С УСАМИ

- Прекрасный способ обобщить большое количество данных
- Дает некоторое представление о симметрии данных, о разбросе
- Показывает наличие выбросов
- Подходит для сравнения нескольких переменных

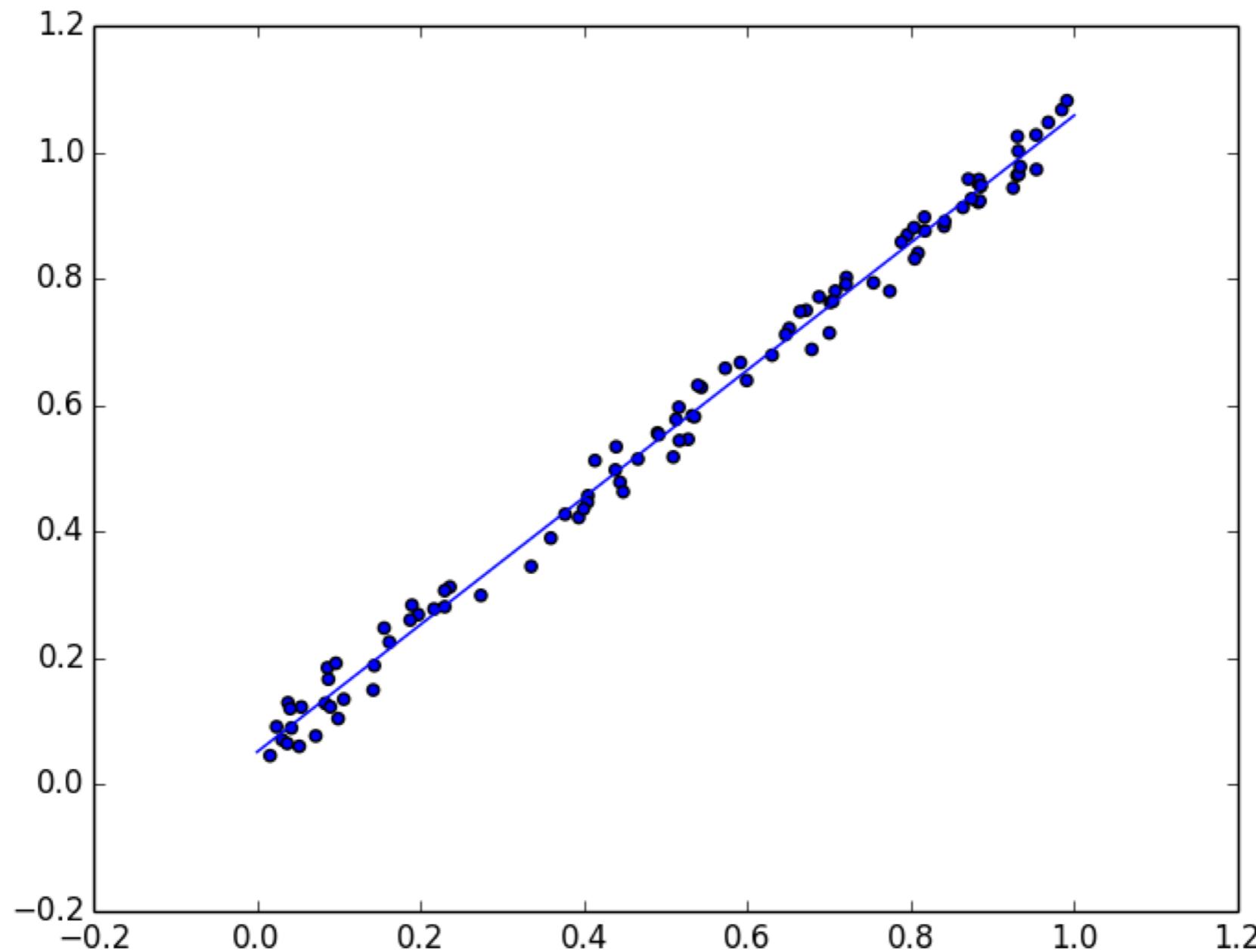


ЯЩИК С УСАМИ

- Нельзя оценить среднее и моду
- Может быть использован только для количественных признаков
- Требует дополнительных пояснений

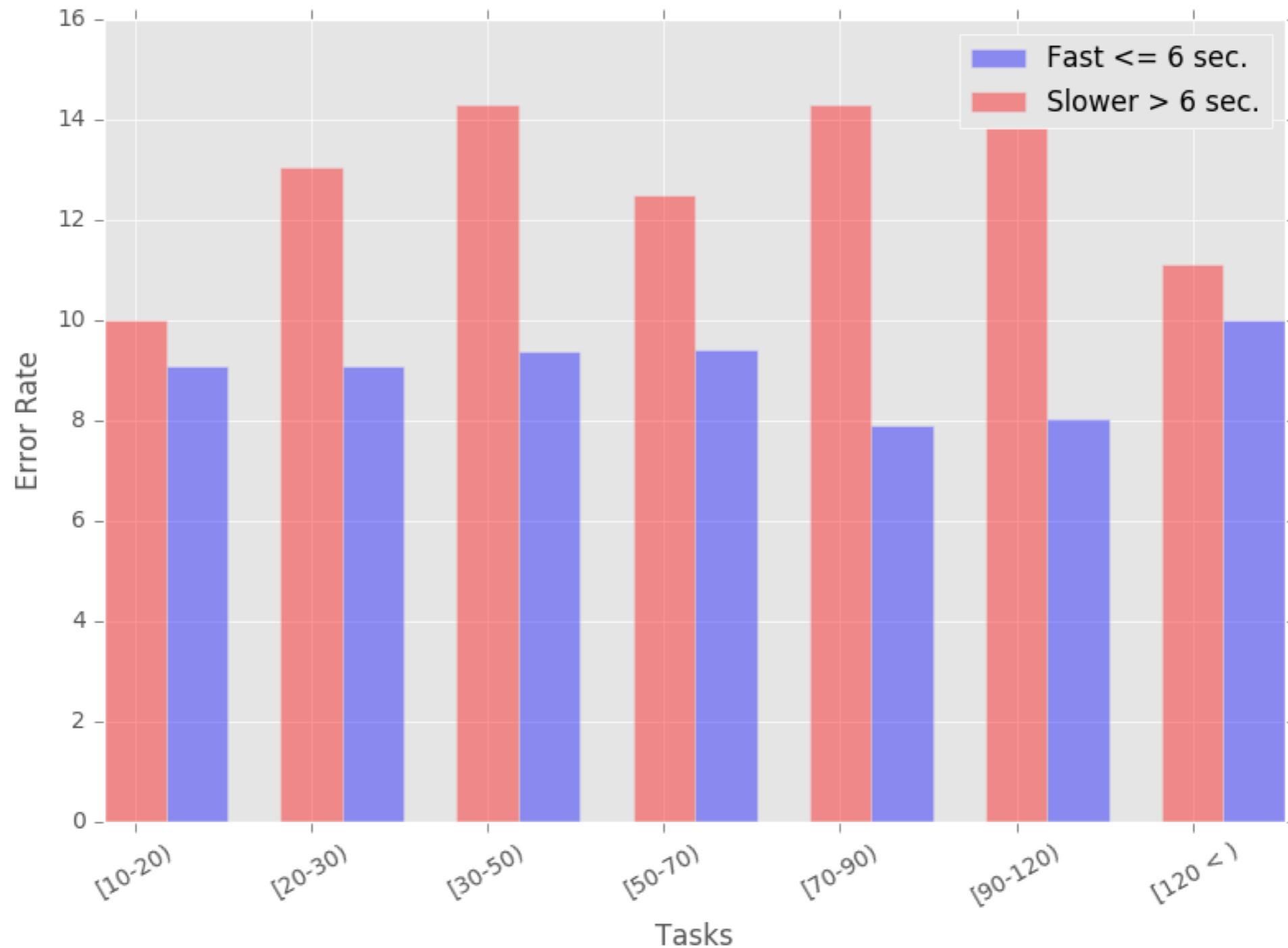


SCATTER PLOT



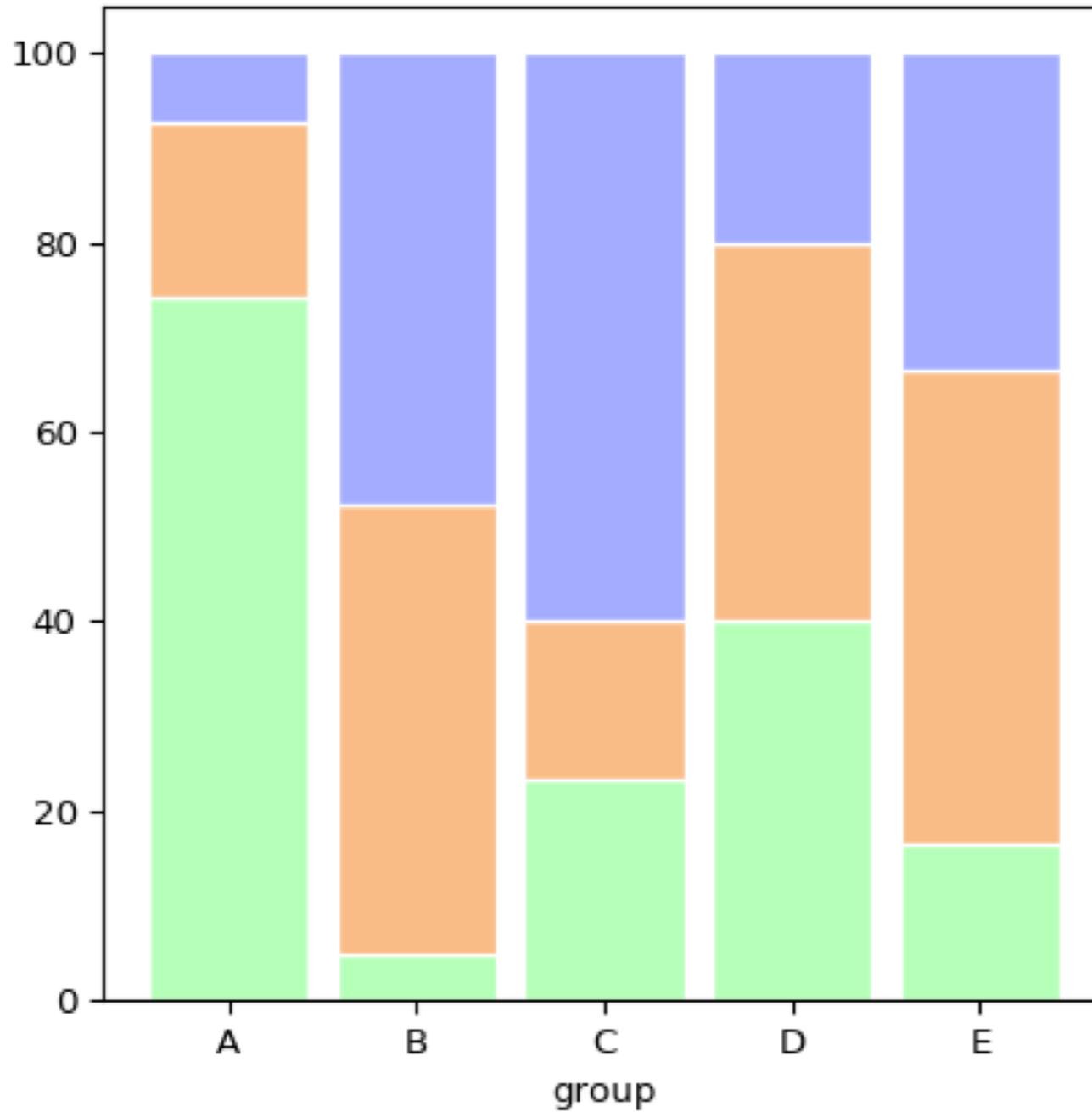


BAR PLOT



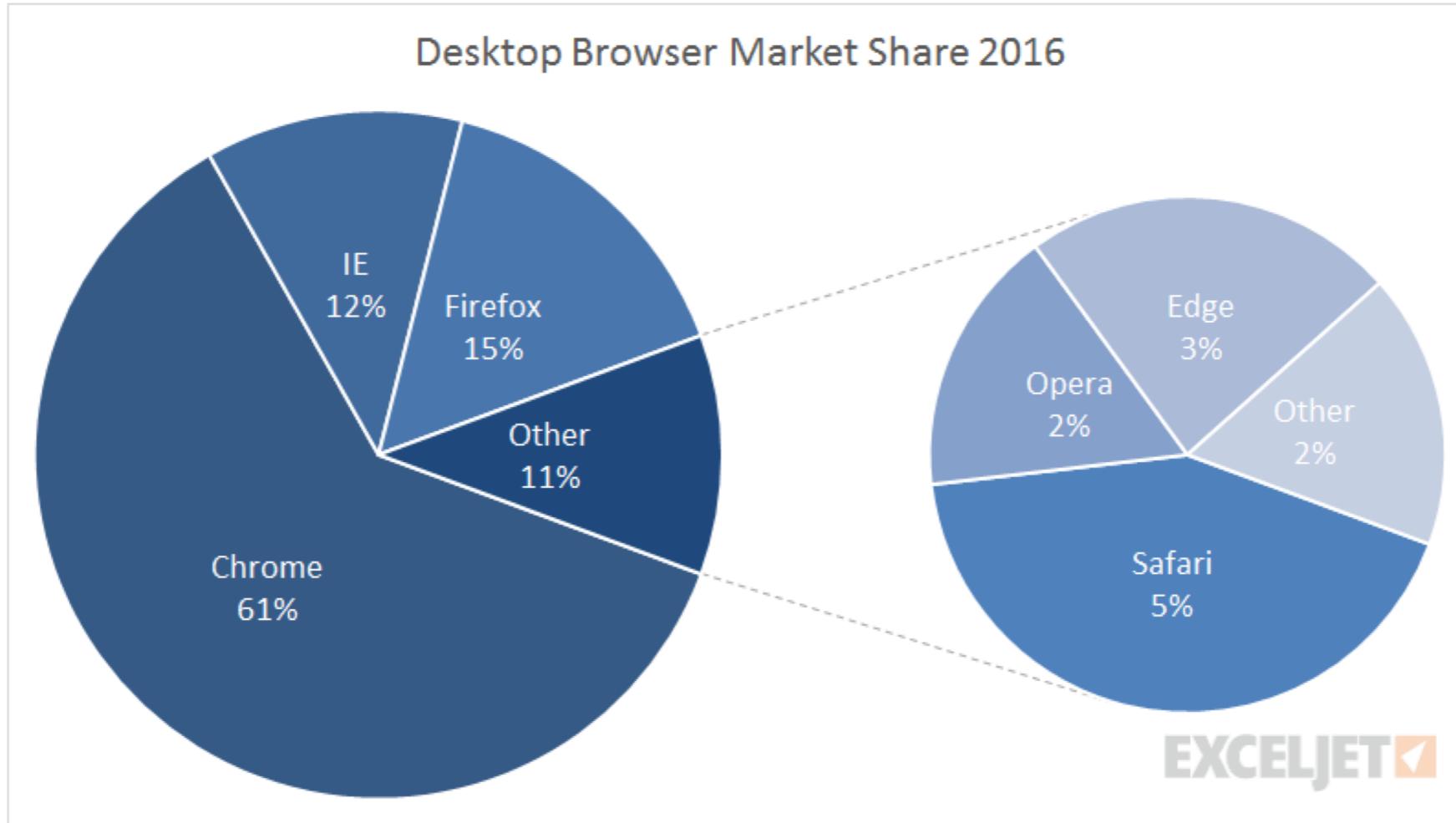


STACKED BAR PLOT



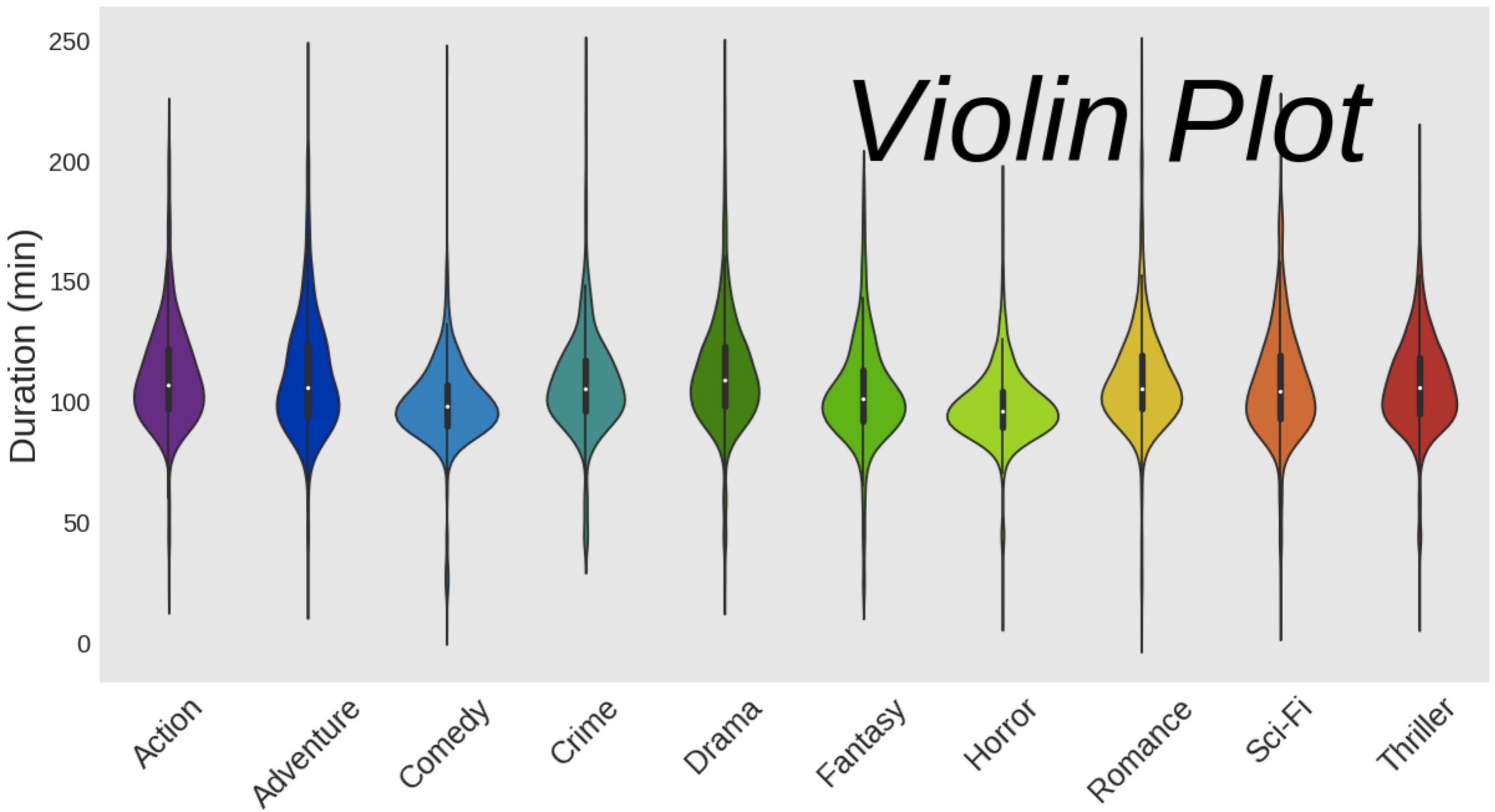


PIE PLOT





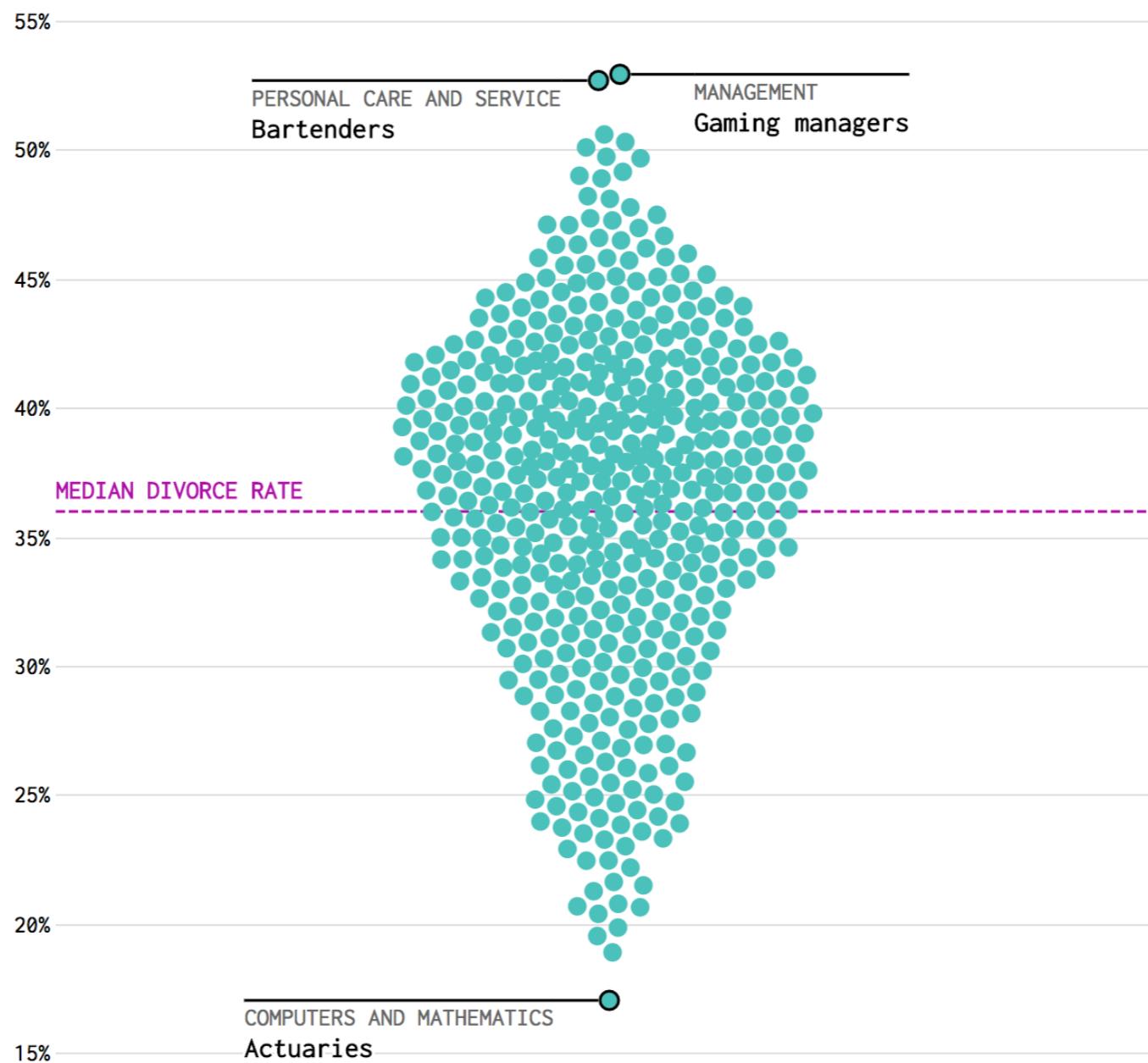
VIOLIN PLOT





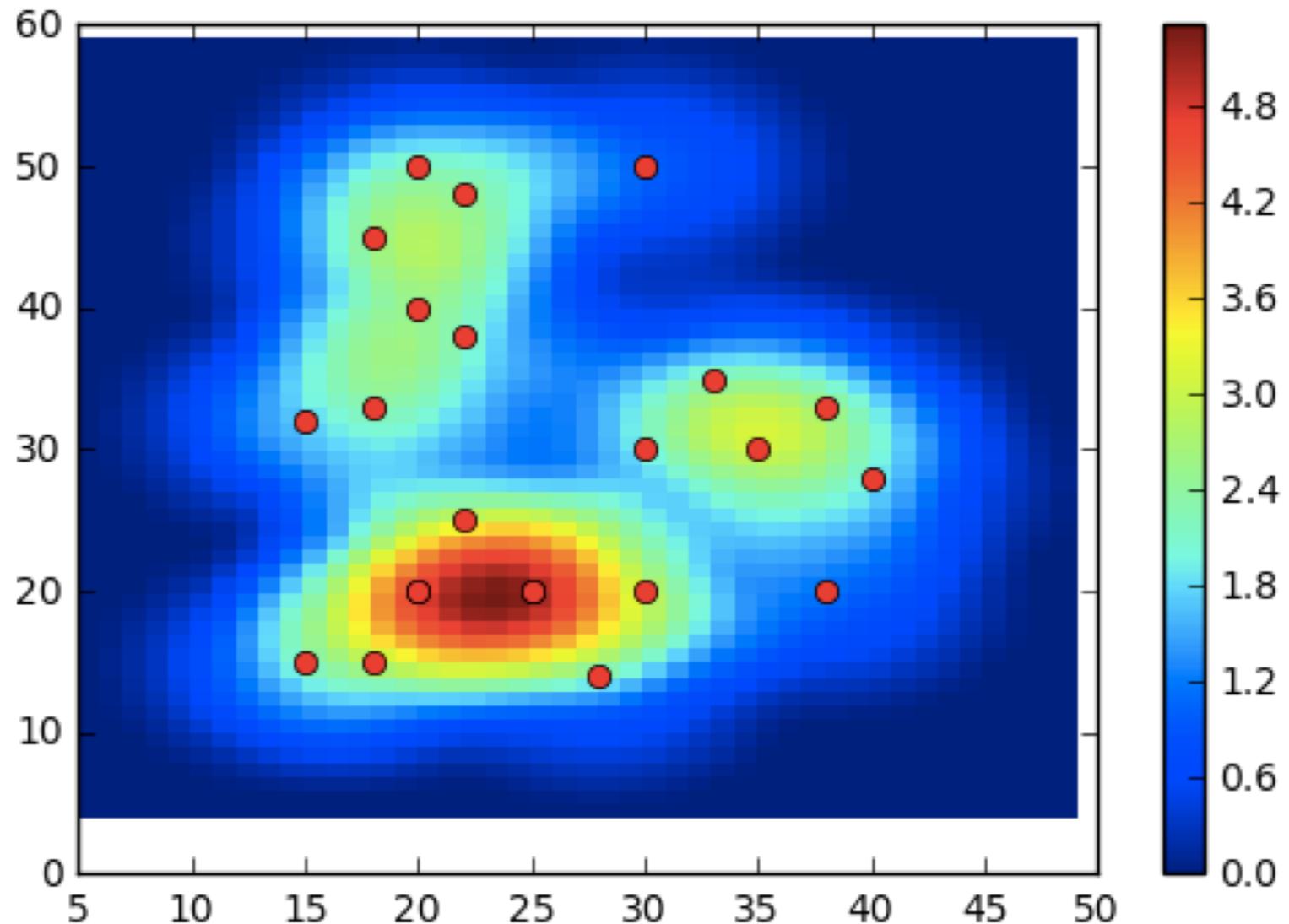
VIOLIN PLOT

DIVORCE RATE BY OCCUPATION





HEATMAP





КОРРЕЛЯЦИЯ

- Взаимосвязь двух случайных величин

$$r_{XY} = \frac{\text{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}.$$



КОРРЕЛЯЦИЯ

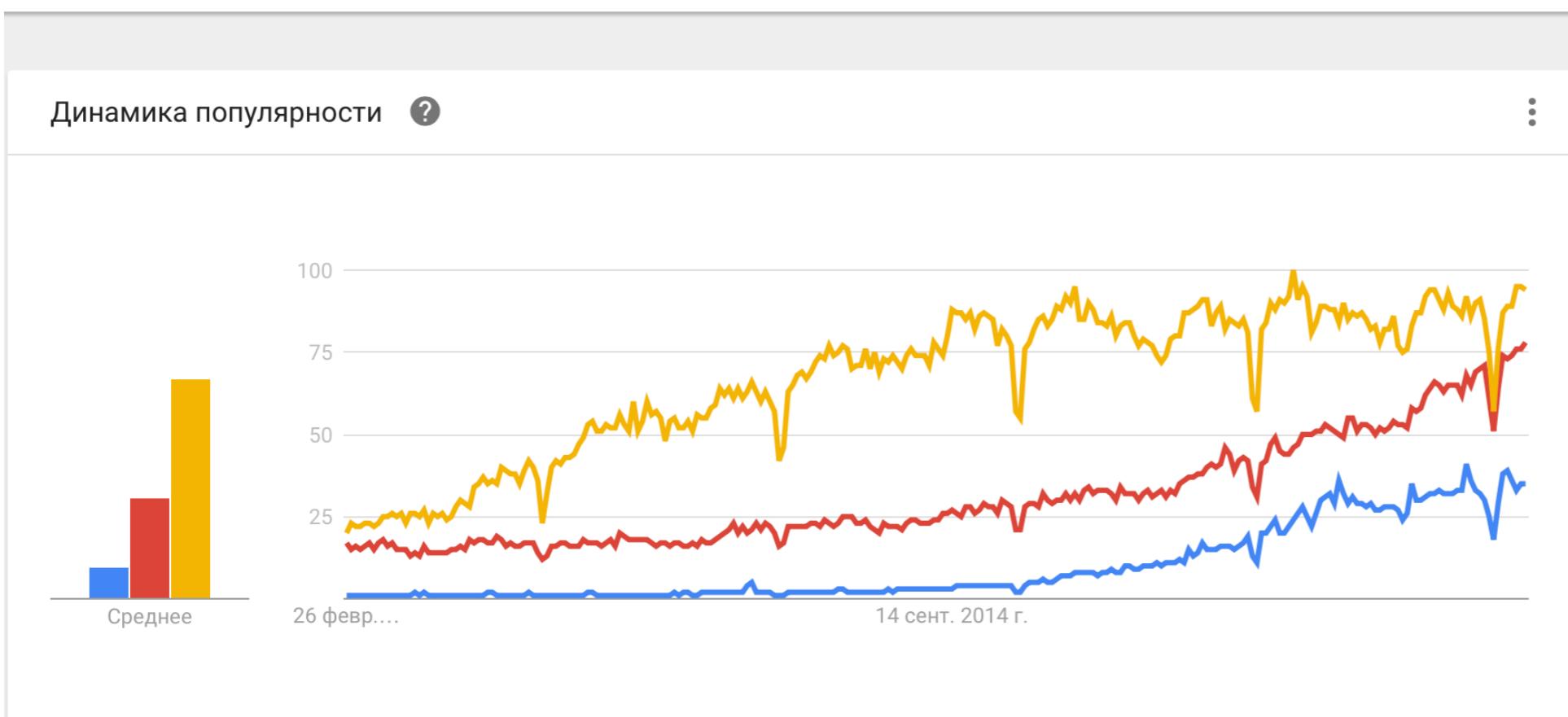
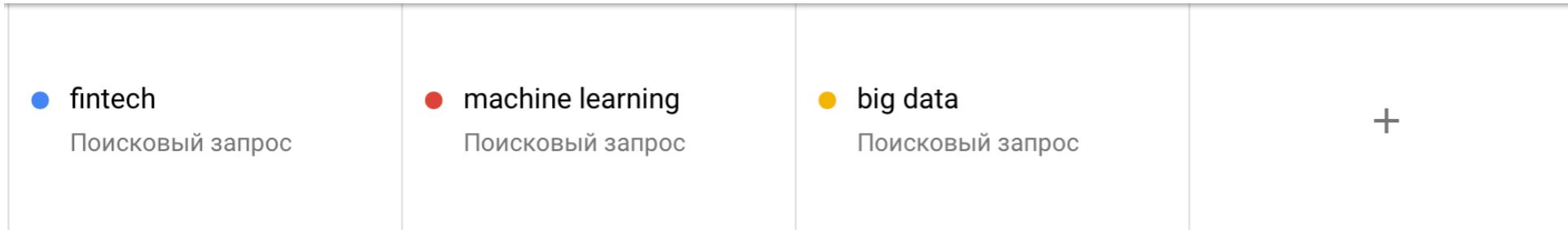
- Взаимосвязь двух случайных величин

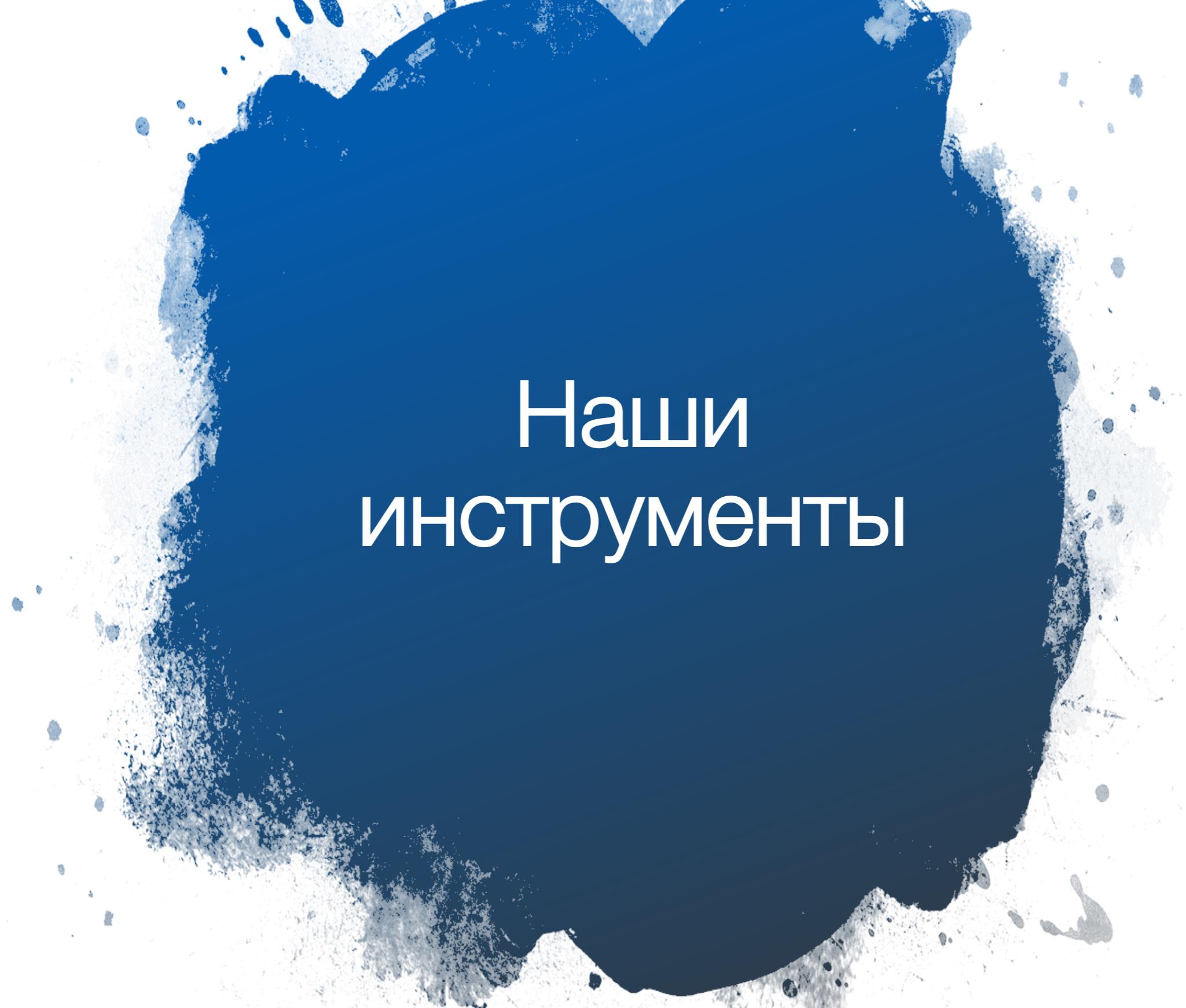
$$r_{XY} = \frac{\text{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}.$$

- Изменяется от -1 до 1
- Ассоциативная (!!!!) взаимосвязь
- Нулевая корреляция – отсутствие линейной взаимосвязи, но не отсутствие взаимосвязи как таковой!



КОРРЕЛЯЦИЯ





Наши инструменты

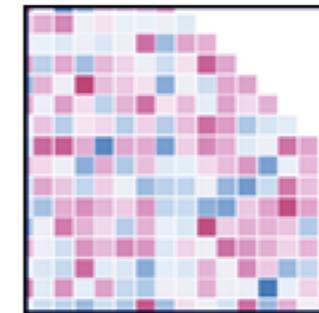


БИБЛИОТЕКИ

Pandas



matplotlib



Seaborn



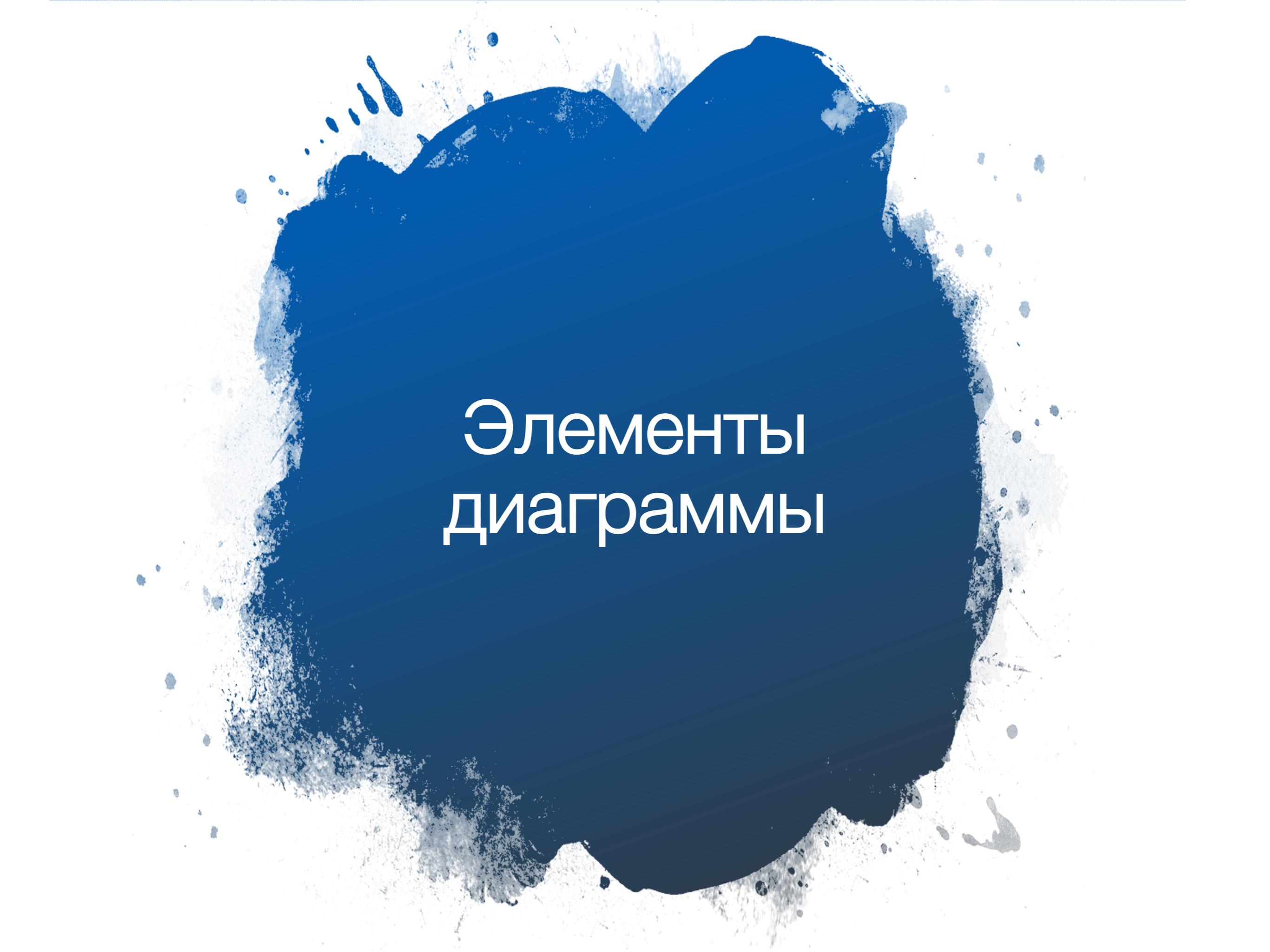
matplotlib.pyplot

- import matplotlib.pyplot as plt
- plt.plot([x], y)
- plt.bar()
- plt.hist(cumulative=False)
- plt.boxplot()
- plt.scatter()
- plt.violinplot()



seaborn

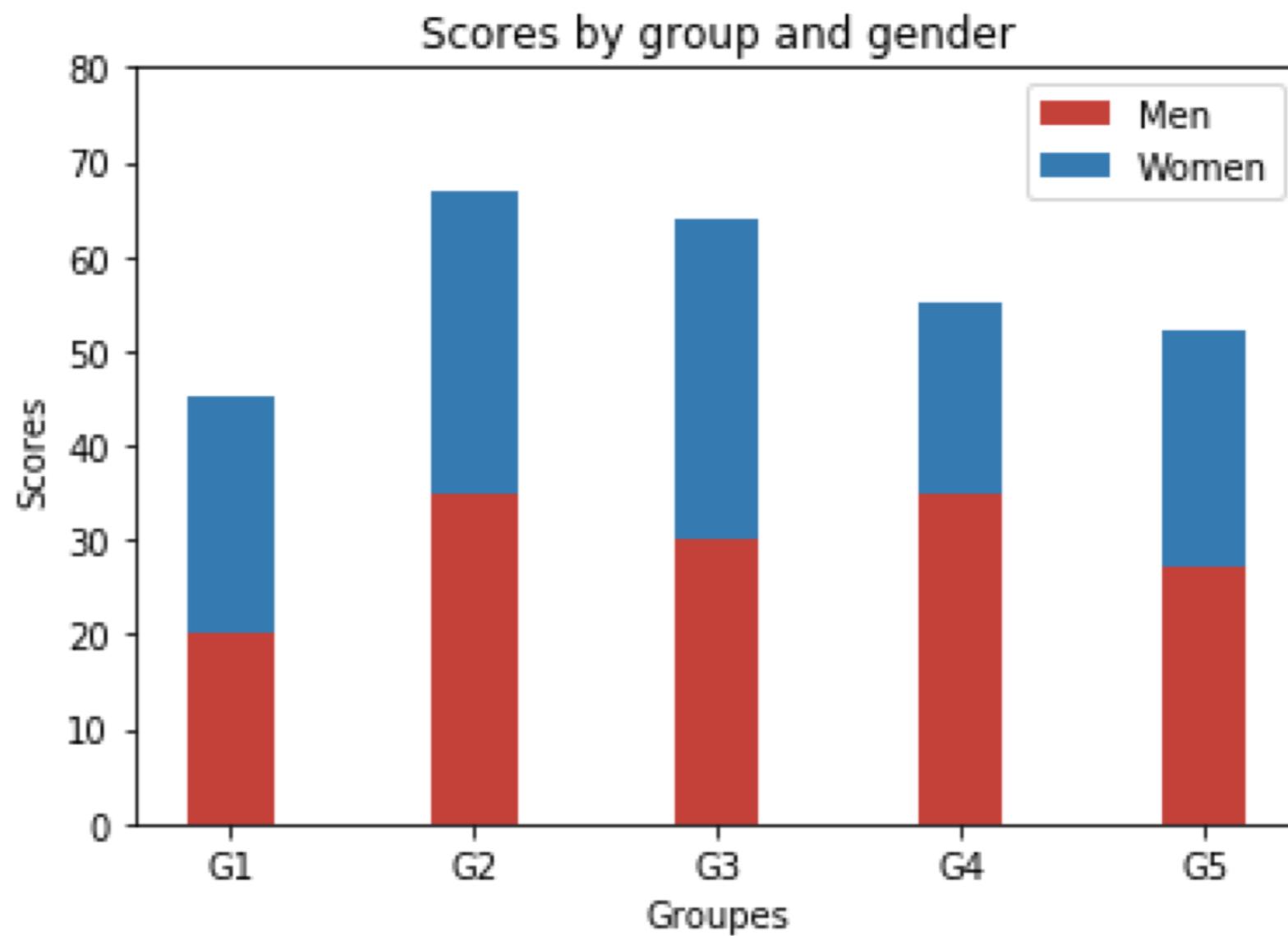
- import seaborn as sns
- sns.lineplot()
- sns.barplot()
- sns.distplot()
- sns.boxplot()
- sns.scatterplot()
- sns.violinplot()



Элементы диаграммы



ПРИМЕР





КОД ПРИМЕРА

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

N = 5
menMeans = (20, 35, 30, 35, 27)
womenMeans = (25, 32, 34, 20, 25)
ind = np.arange(N)    # the x locations for the groups
width = 0.35      # the width of the bars: can also be len(x) sequence

p1 = plt.bar(ind, menMeans, width, color='#d62728')
p2 = plt.bar(ind, womenMeans, width,
             bottom=menMeans)

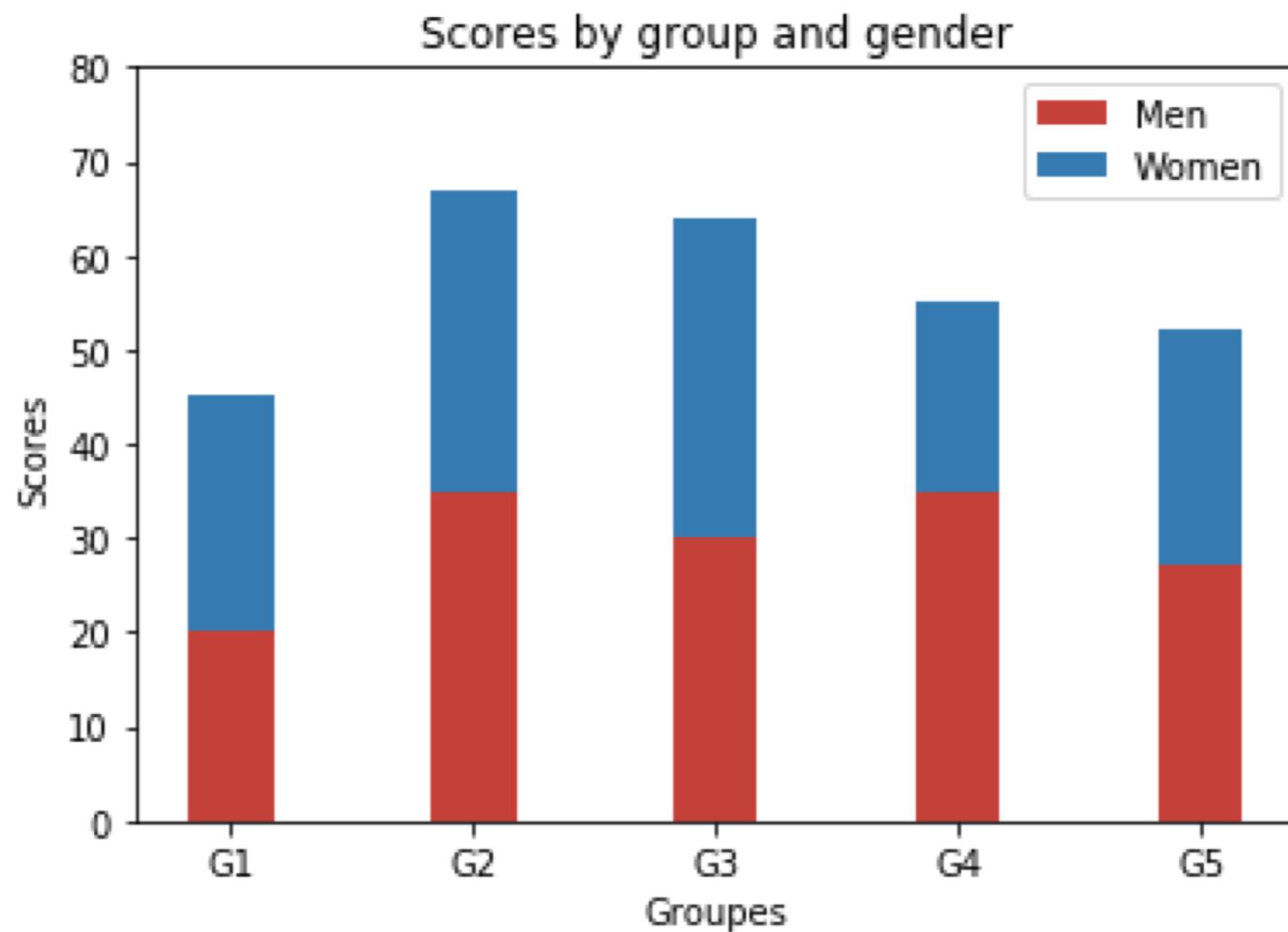
plt.xlabel('Groups')
plt.ylabel('Scores')
plt.title('Scores by group and gender')
plt.xticks(ind, ('G1', 'G2', 'G3', 'G4', 'G5'))
plt.yticks(np.arange(0, 81, 10))
plt.legend((p1[0], p2[0]), ('Men', 'Women'))

plt.show()
```



ЗАГОЛОВОК

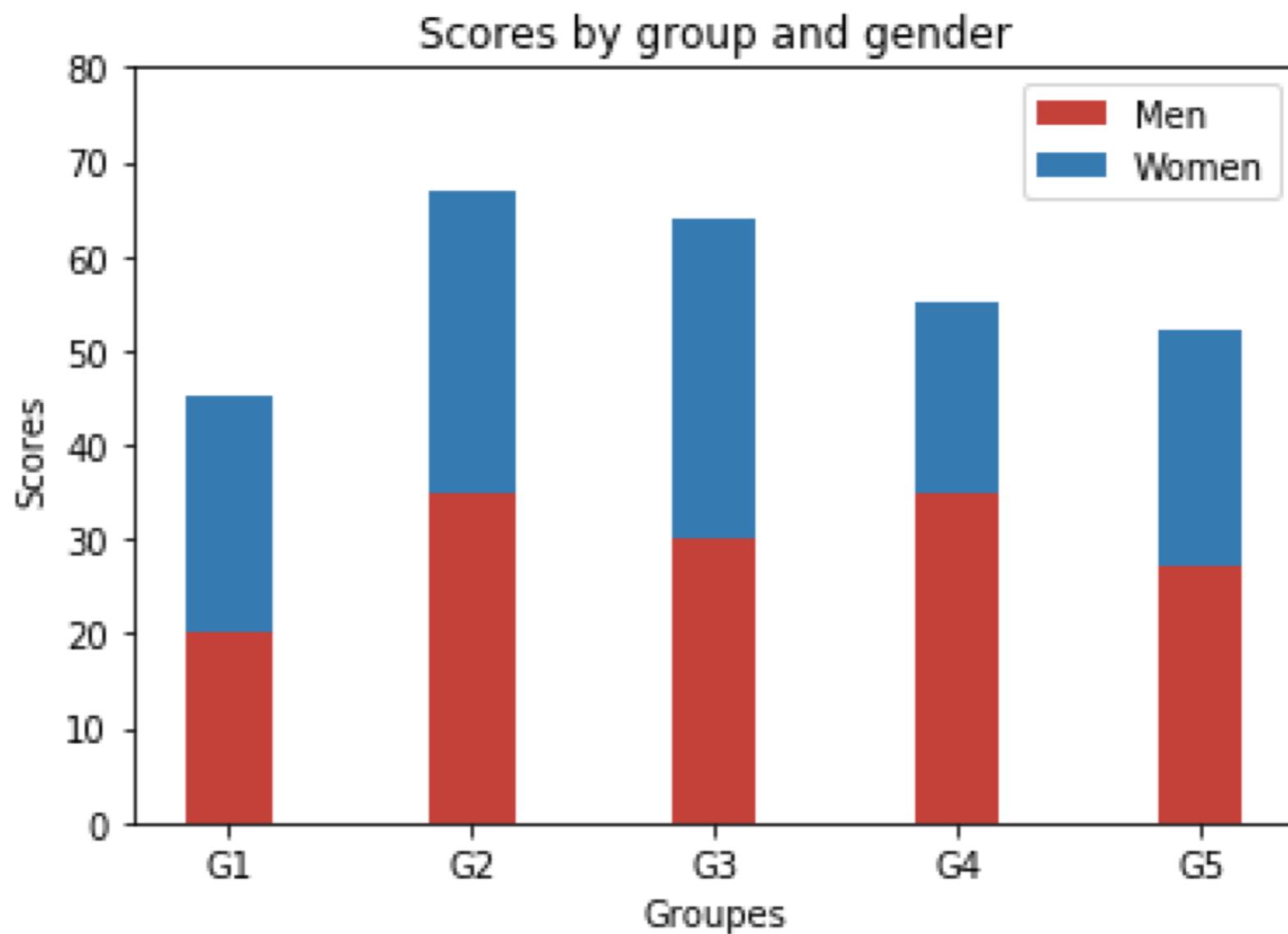
```
plt.title('Scores by group and gender')
```





НАЗВАНИЕ ОСИ АБСЦИСС

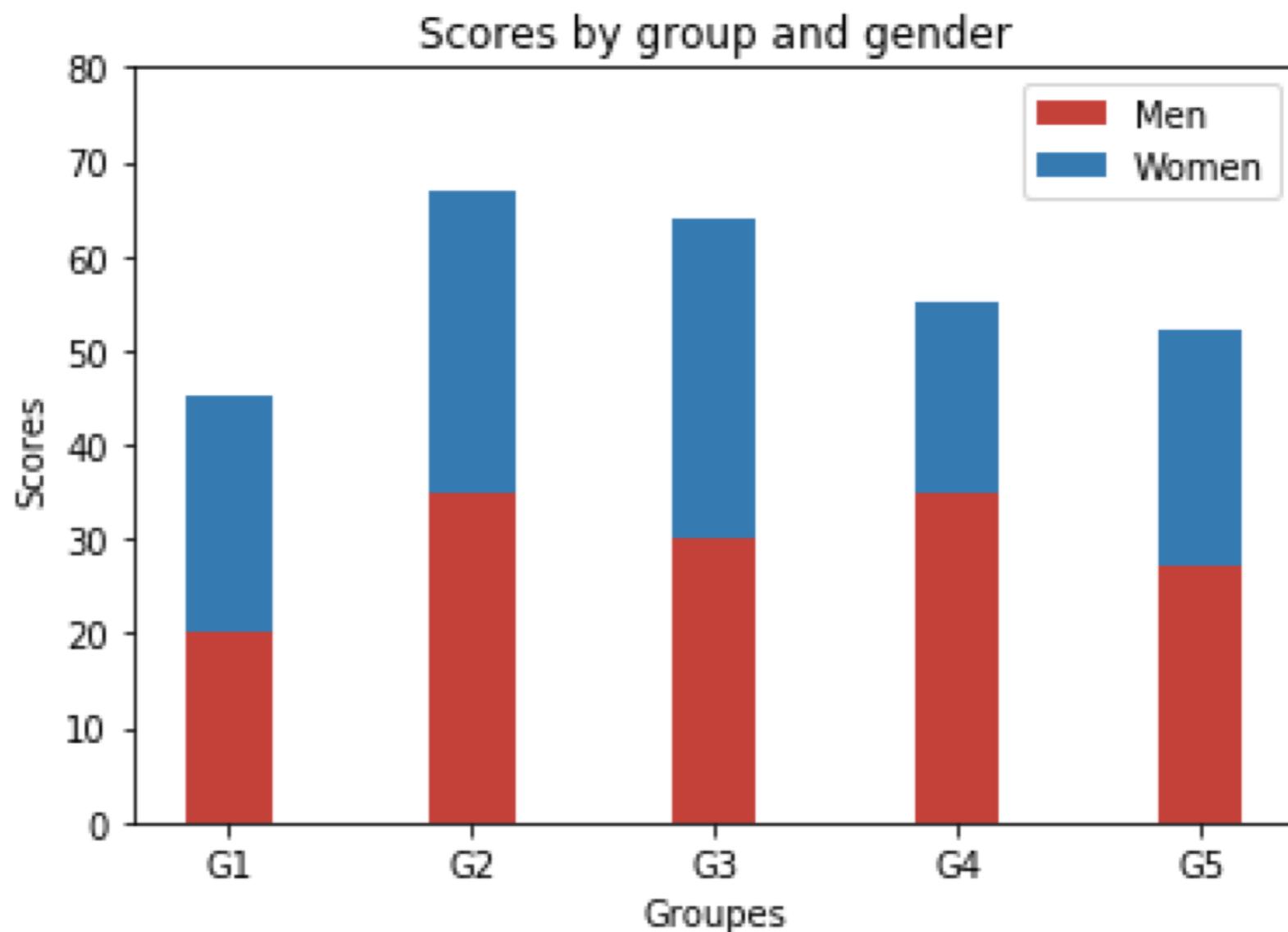
```
plt.xlabel('Groupes')
```





ПОДПИСЬ ОСИ АБСЦИСС

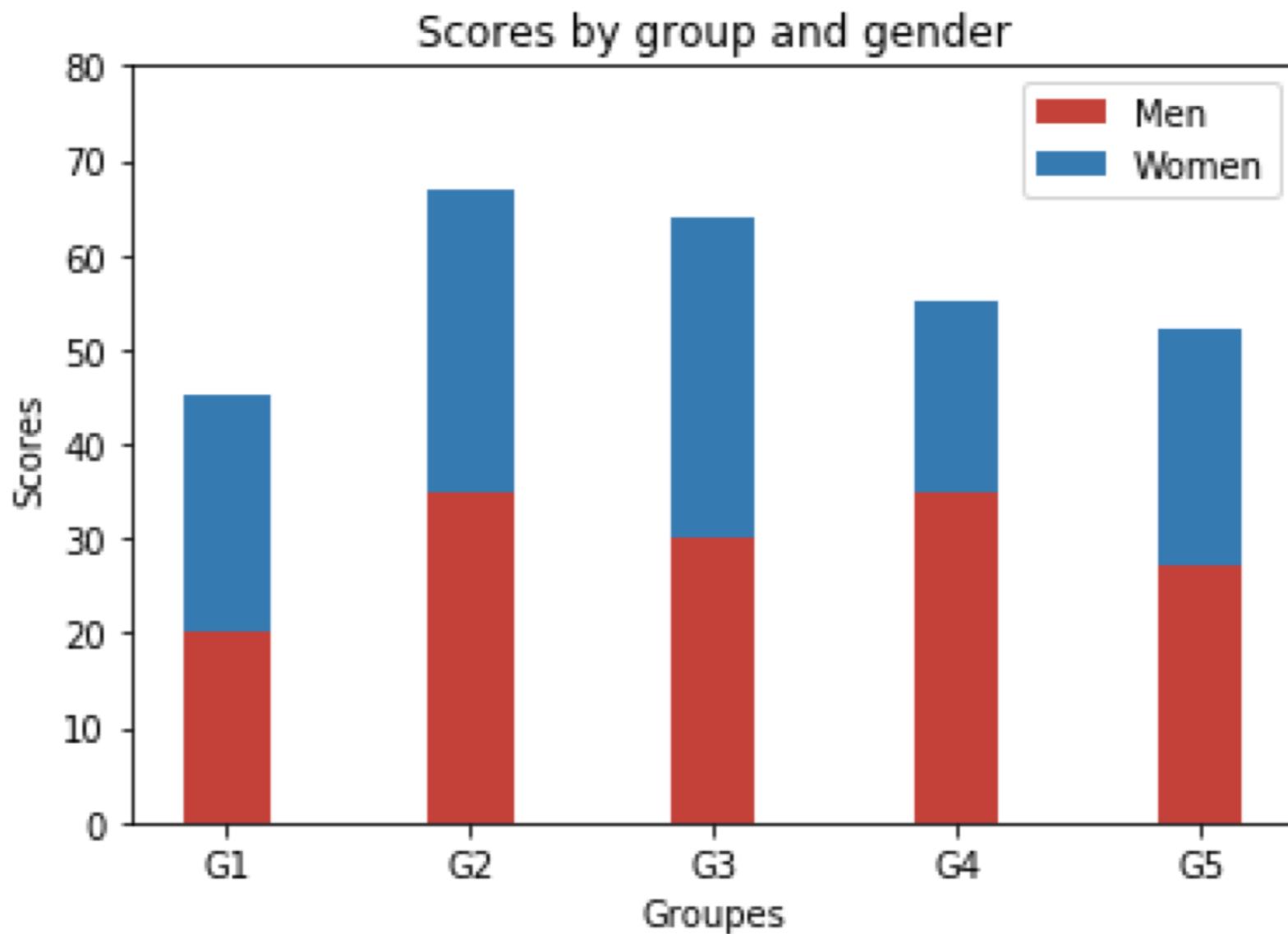
```
plt.xticks(ind, ('G1', 'G2', 'G3', 'G4', 'G5'))
```





НАЗВАНИЕ ОСИ ОРДИНАТ

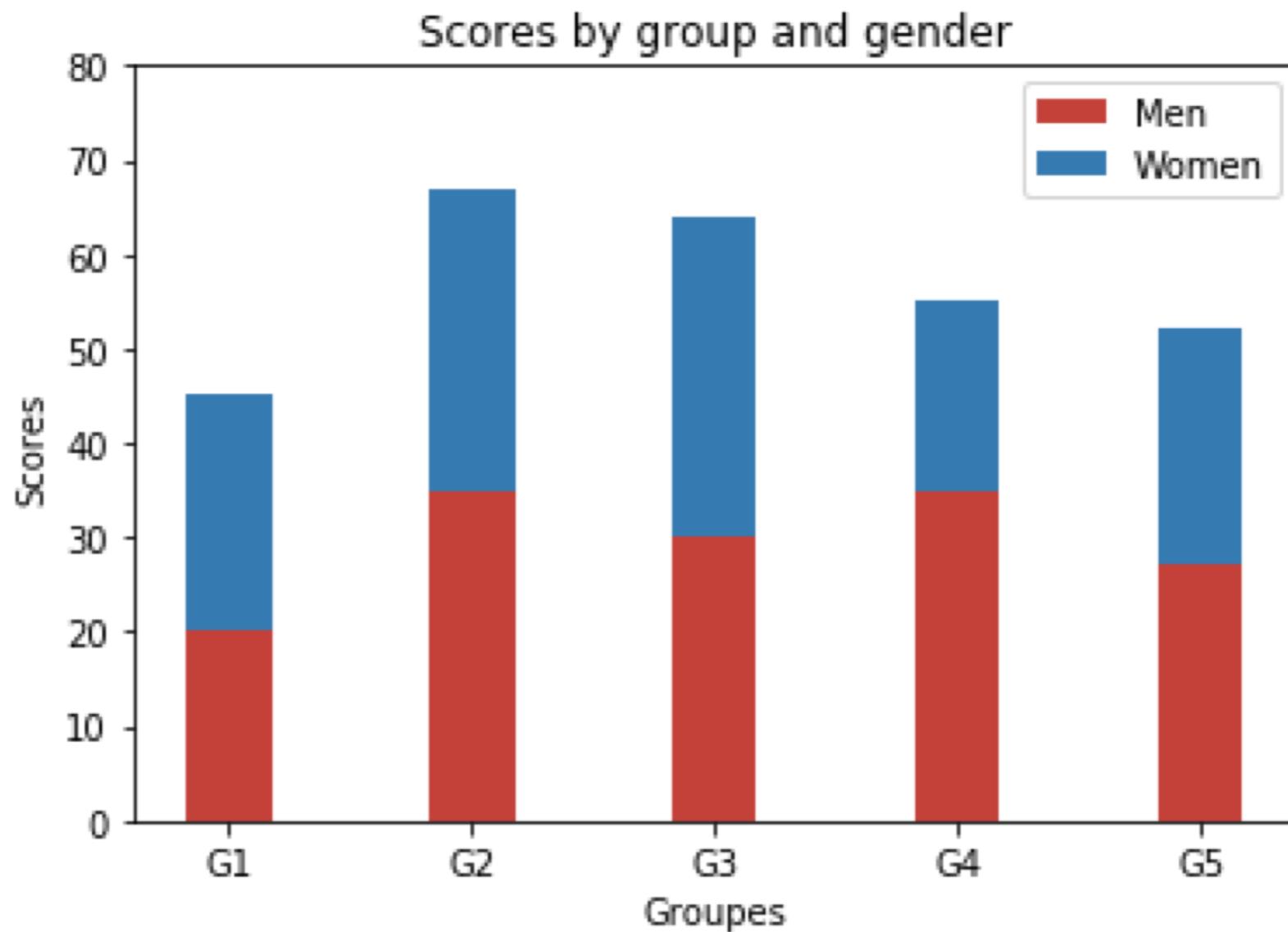
```
plt.ylabel('Scores')
```





ПОДПИСЬ ОСИ ОРДИНАТ

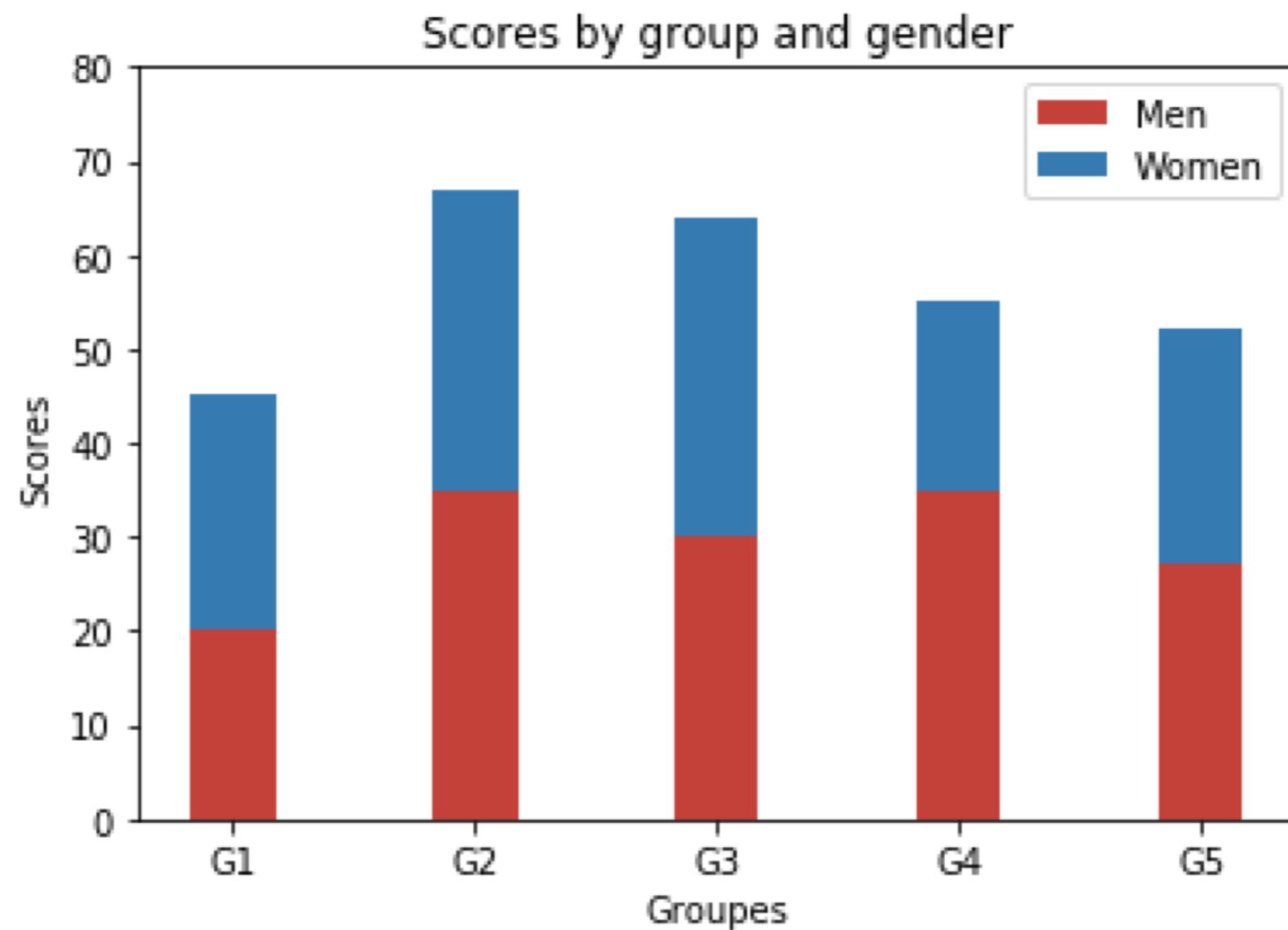
```
plt.yticks(np.arange(0, 81, 10))
```

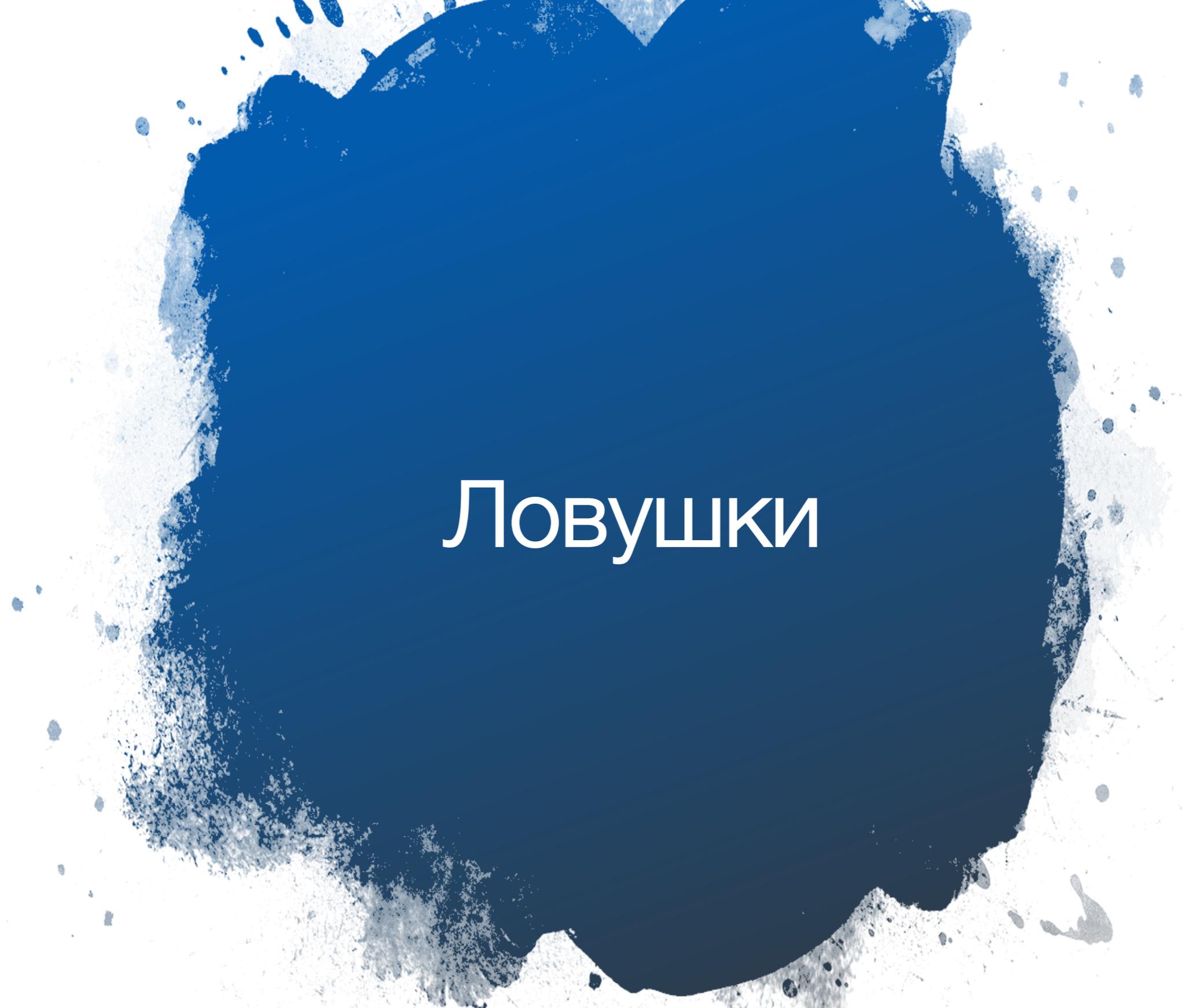




ЛЕГЕНДА

```
plt.legend((p1[0], p2[0]), ('Men', 'Women'))
```





Ловушки



ЛОВУШКИ ПРИ ПОСТРОЕНИИ ГРАФИКОВ

Важно не искажать интерпретацию данных при визуализации

Раньше было три вида лжи:
ложь, наглая ложь
и статистика.

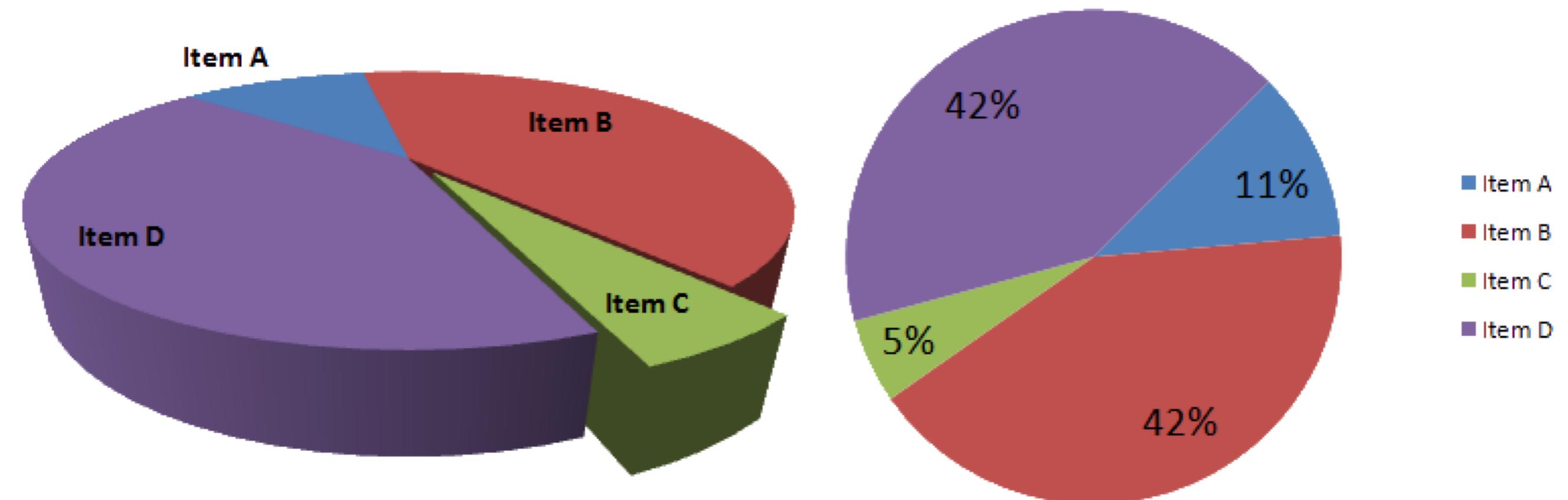
Теперь у нас есть
Big Data.

Atkritka.com



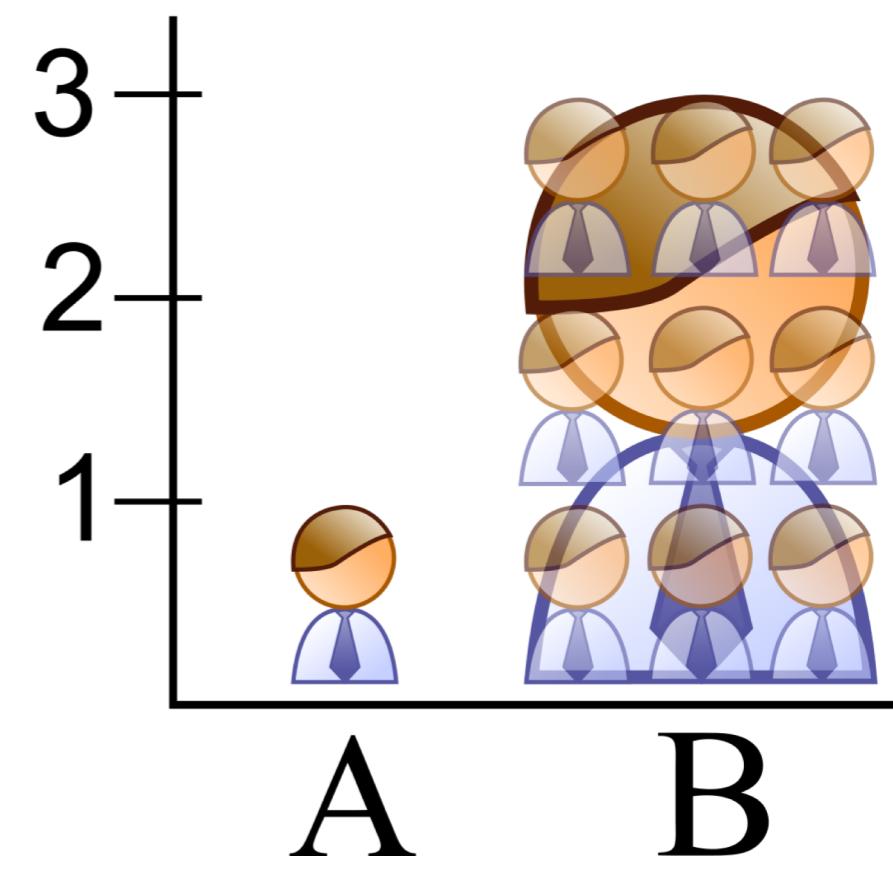
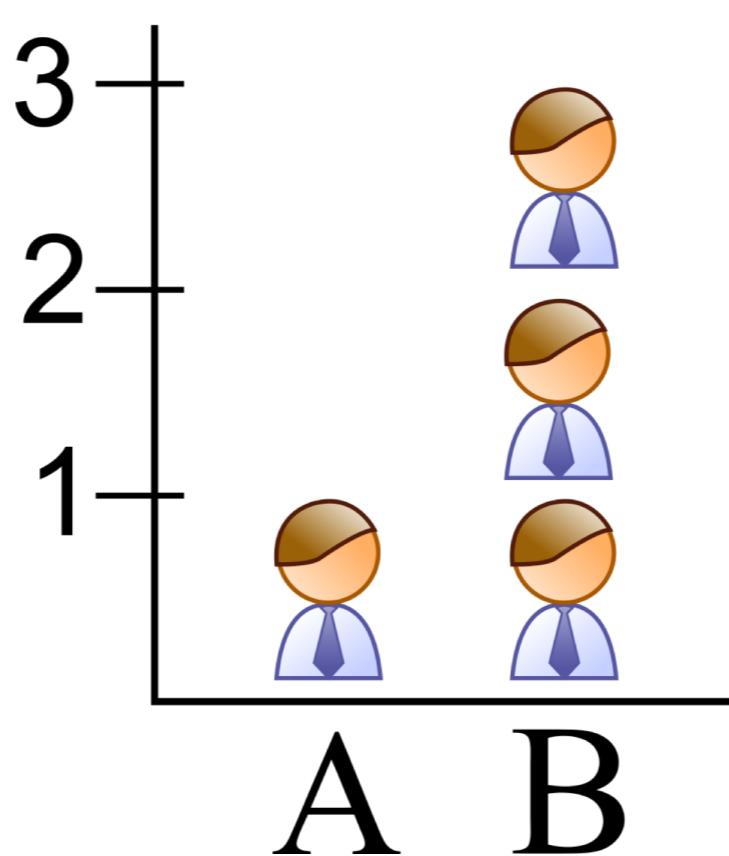
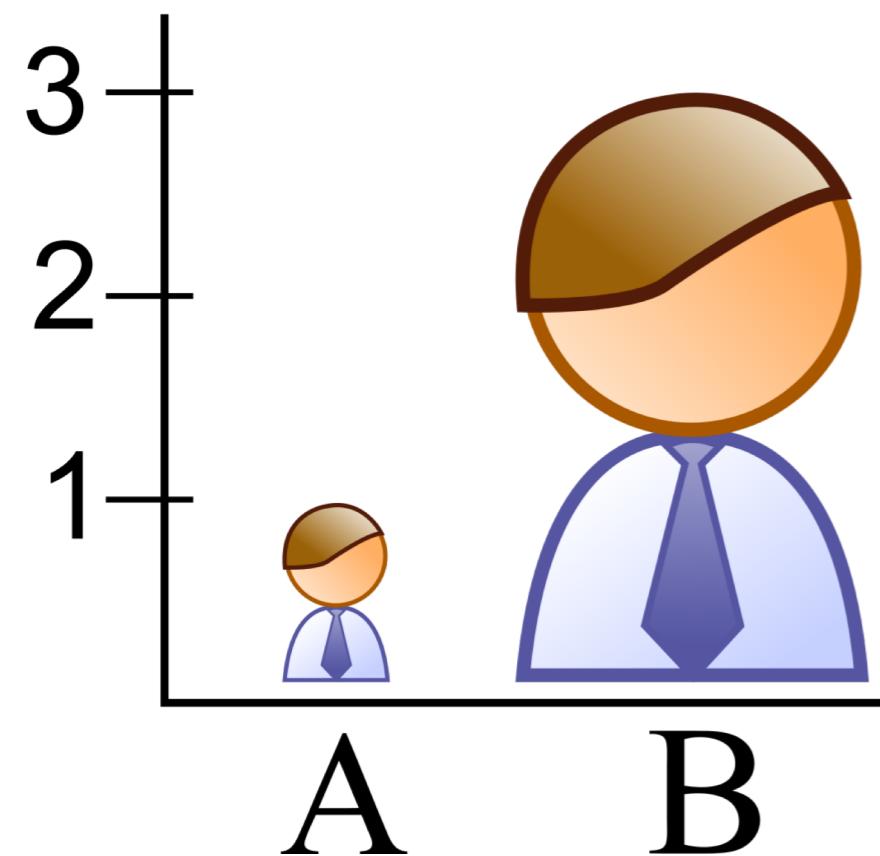


ЛОЖЬ ПРИ ПОМОЩИ ГРАФИКОВ



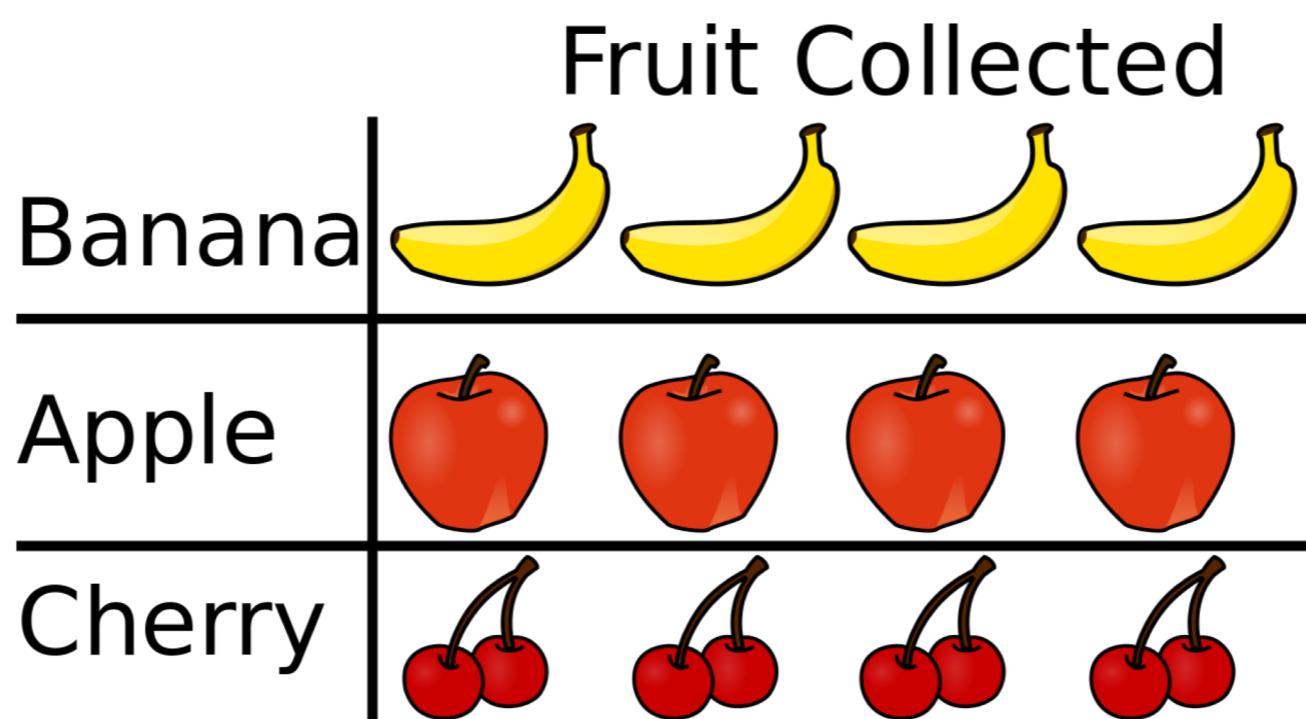
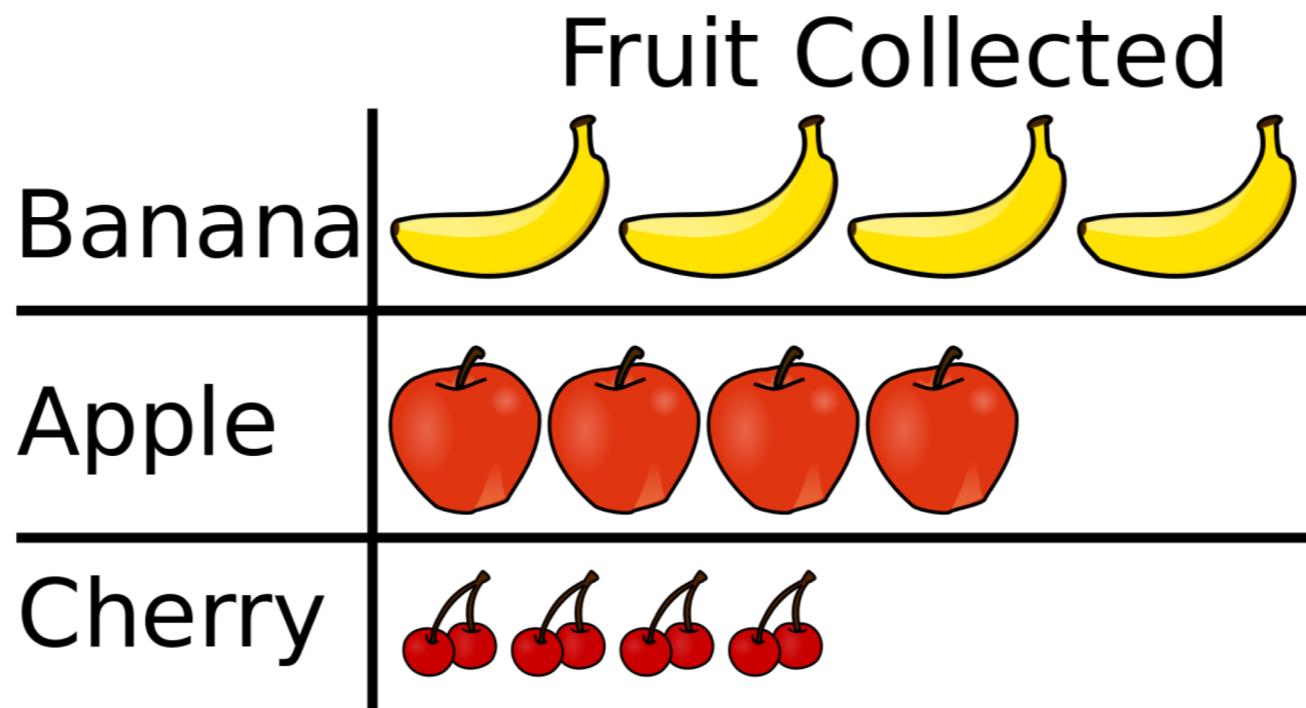


ЛОЖЬ ПРИ ПОМОЩИ ГРАФИКОВ



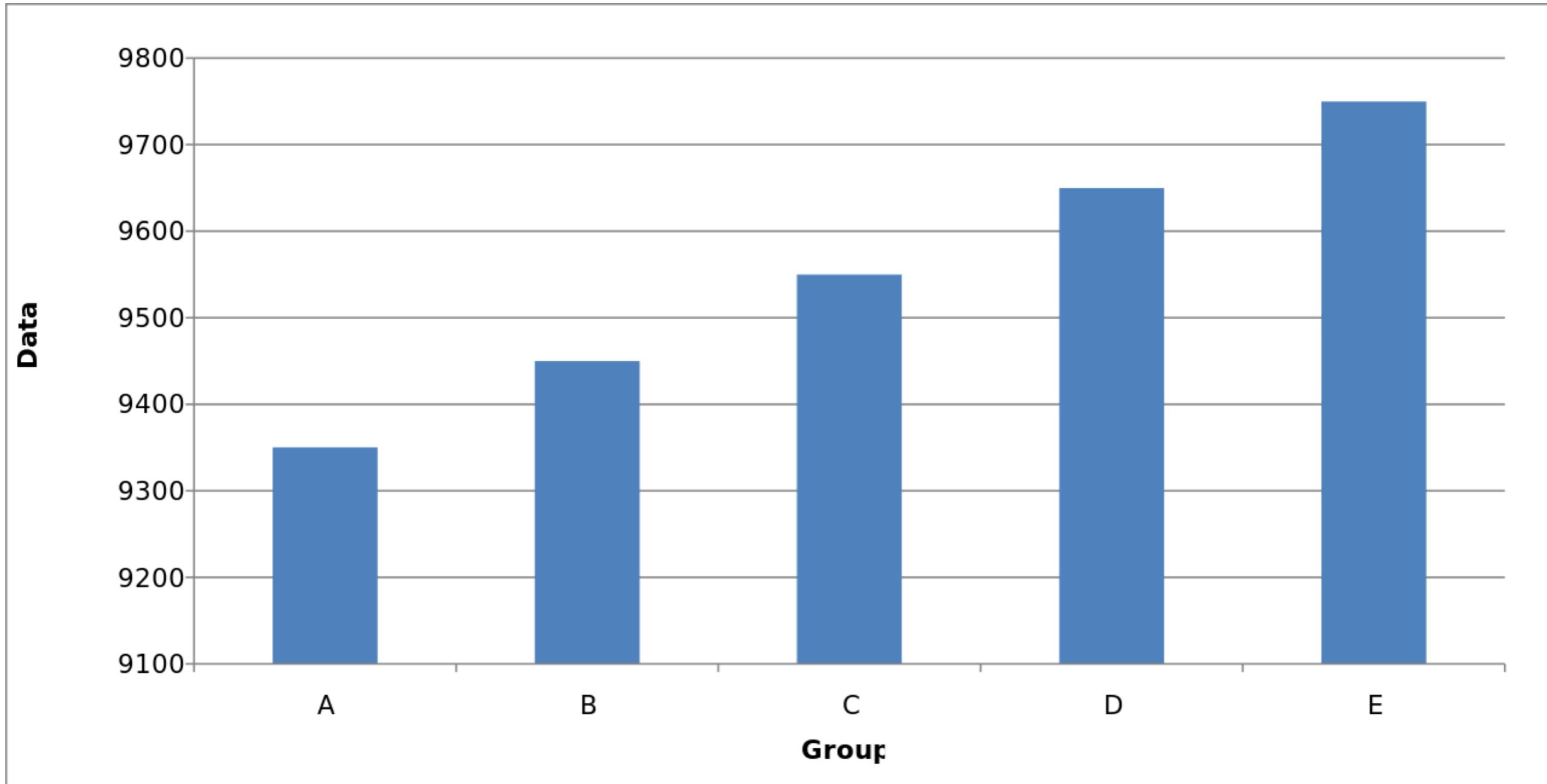


ЛОЖЬ ПРИ ПОМОЩИ ГРАФИКОВ



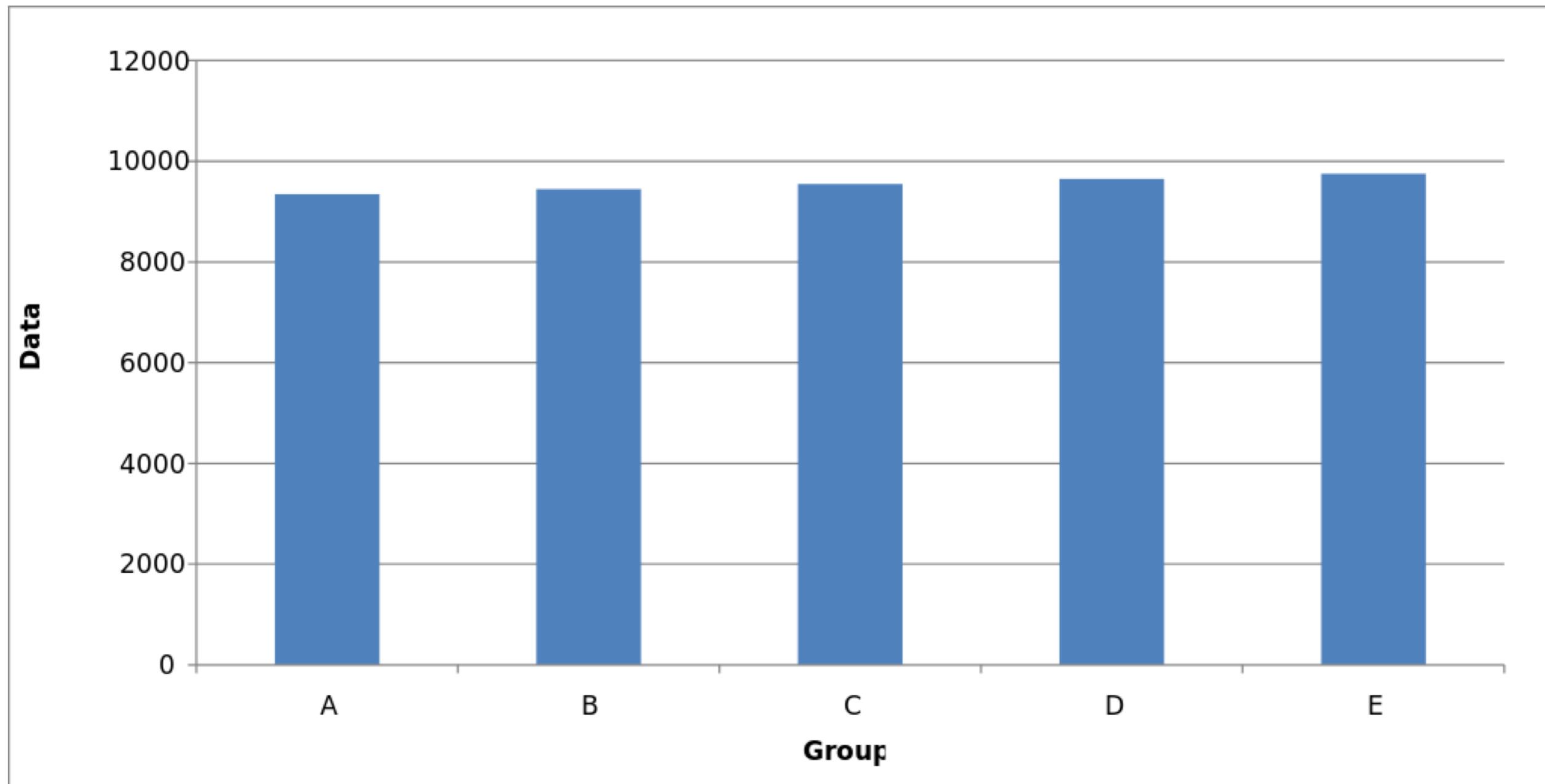


ЛОЖЬ ПРИ ПОМОЩИ ГРАФИКОВ





ЛОЖЬ ПРИ ПОМОЩИ ГРАФИКОВ

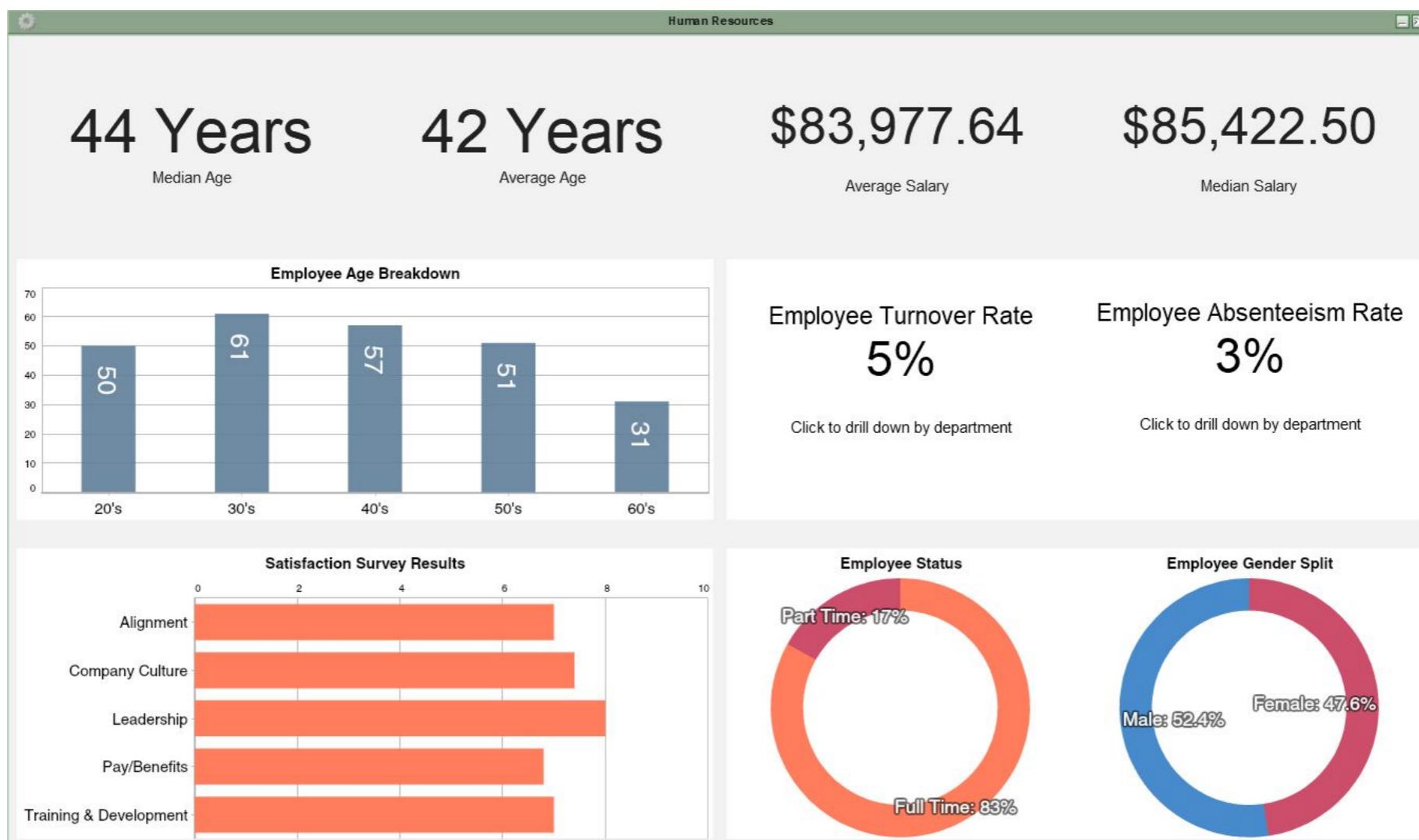


Дашборды



ДАШБОРД

- Интерактивная и динамическая визуализация с помощью виджетов





ЗАЧЕМ ДАШБОРДЫ

- Цель — лёгкий доступ для всех сотрудников к основным показателям
- Мотивация: видеть влияние всех внедрений и решений на KPI
- Мониторинг: быстро замечать сбои в работе сервиса
- Целеполагание: всегда видно, где есть недостатки, можно правильно концентрировать усилия



ДАШБОРДЫ СТРОЯТСЯ В СПЕЦСОФТАХ

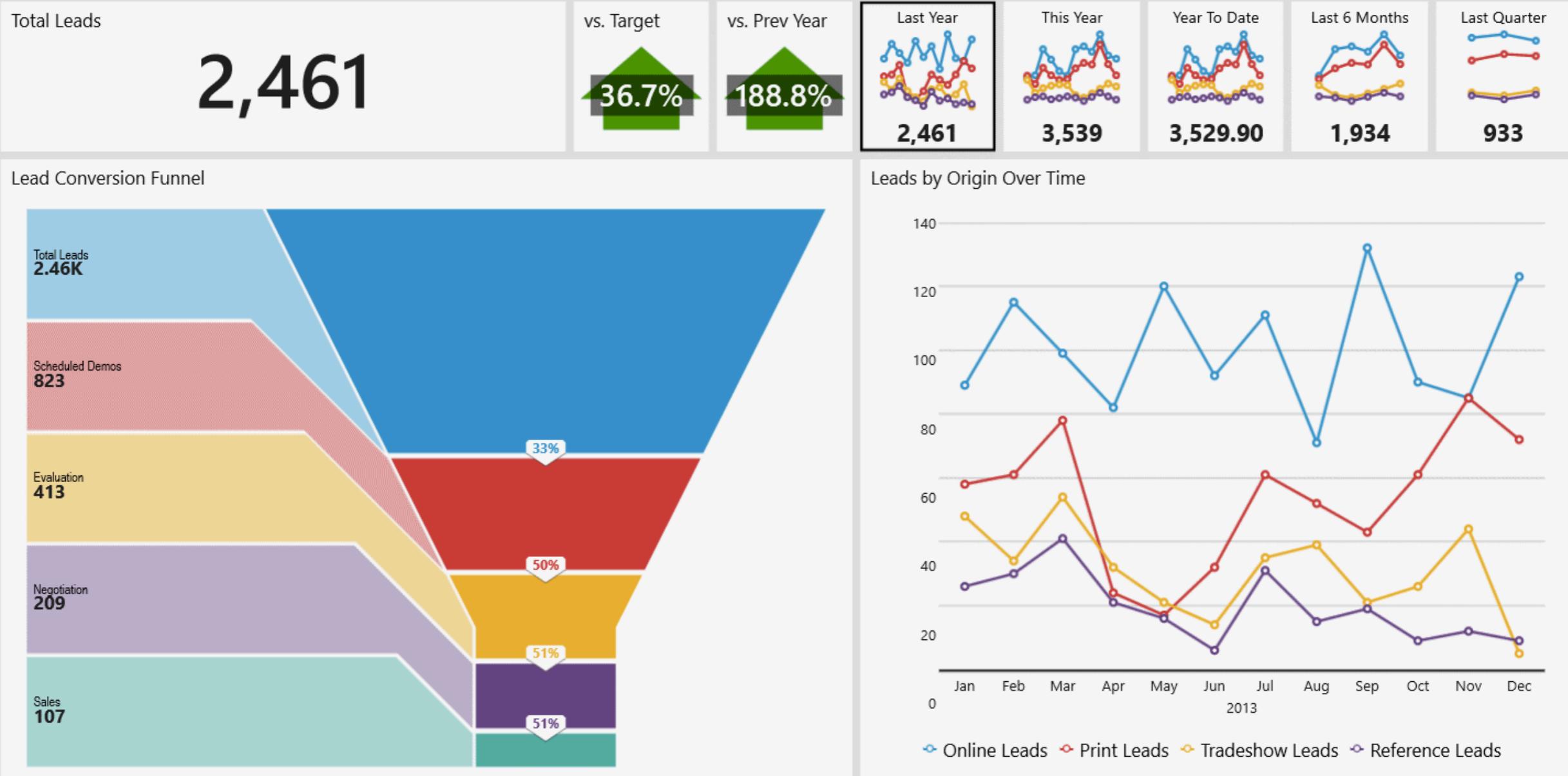




ЕЩЕ ПРИМЕР

Marketing Dashboard 2013

Activity





ОСНОВНЫЕ ПОНЯТИЯ

- группировка
- агрегация
- сортировка
- фильтрация
- вычисляемая колонка
- топовые (лучшие) значения
- виджеты



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ