СБЕР УНИВЕРСИТЕТ

Школа Data scientist

Занятие 7

Анализ данных в Python

Тема 3

Disclaimer

Все формулировки далее нестрогие, за более строгими определениями обращайтесь к специализированной литературе

План занятия

- Первичный анализ и очистка данных
- Когортный анализ
- Корреляционный анализ
- Exploratory data analysis

Первичный анализ и очистка данных

Описательная статистика

Описательная статистика (descriptive statistics) занимается обработкой эмпирических данных, их систематизацией, наглядным представлением в форме графиков и таблиц, а также их количественным описанием посредством основных статистических показателей.

В отличии от статистического вывода , не делает выводов о генеральной совокупности на основании результатов исследования частных случаев.

Описательная статистика

```
data = pd.read_csv("iris.csv")
data["petal.width"].iloc[145:149] = None
data.tail(7)
```

Index	sepal.length	sepal.width	petal.length	petal.width	variety
143	6.8	3.2	5.9	2.3	Virginica
144	6.7	3.3	5.7	2.5	Virginica
145	6.7	3.0	5.2	NaN	Virginica
146	6.3	2.5	5.0	NaN	Virginica
147	6.5	3.0	5.2	NaN	Virginica
148	6.2	3.4	5.4	NaN	Virginica
149	5.9	3.0	5.1	1.8	Virginica

Описательная статистика

data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   sepal.length  150 non-null    float64
 1   sepal.width   150 non-null    float64
 2   petal.length  150 non-null    float64
 3   petal.width   146 non-null    float64
 4   variety       150 non-null    object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

data.size

750

data.count()

```
sepal.length    150
sepal.width     150
petal.length    150
petal.width     146
variety         150
dtype: int64
```

Описательная статистика

data.nunique()

```
sepal.length    35
sepal.width     23
petal.length    43
petal.width     22
variety          3
dtype: int64
```

data.sum()

```
sepal.length                                                    876.5
sepal.width                                                     458.6
petal.length                                                    563.7
petal.width                                                     171.4
variety         SetosaSetosaSetosaSetosaSetosaSetosaSetosaSeto...
dtype: object
```

data.isna().sum()
data.isnull().sum()

```
sepal.length    0
sepal.width     0
petal.length    0
petal.width     4
variety         0
dtype: int64
```

data.notna().sum()
data.notnull().sum()

```
sepal.length    150
sepal.width     150
petal.length    150
petal.width     146
variety         150
dtype: int64
```

Описательная статистика

data.mean()

```
sepal.length    5.843333
sepal.width     3.057333
petal.length    3.758000
petal.width     1.173973
dtype: float64
```

data.median()

```
sepal.length    5.80
sepal.width     3.00
petal.length    4.35
petal.width     1.30
dtype: float64
```

data.min()

```
sepal.length        4.3
sepal.width           2
petal.length          1
petal.width         0.1
variety          Setosa
dtype: object
```

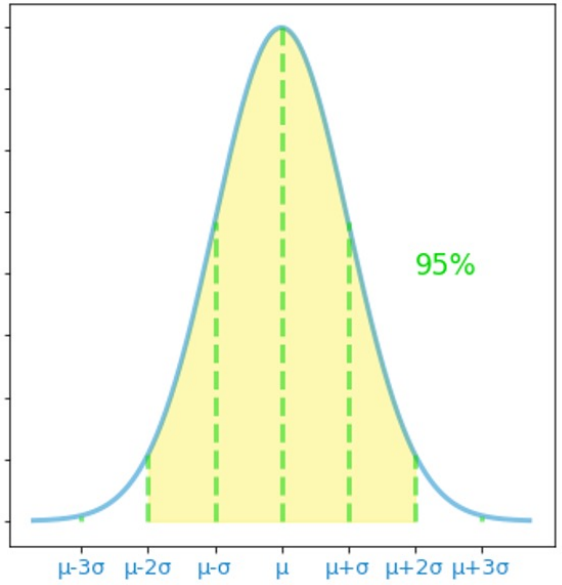
data.max()

```
sepal.length          7.9
sepal.width           4.4
petal.length          6.9
petal.width           2.5
variety         Virginica
dtype: object
```

Описательная статистика

Дисперсия σ^2
data.var()

```
sepal.length    0.685694
sepal.width     0.189979
petal.length    3.116278
petal.width     0.571870
dtype: float64
```



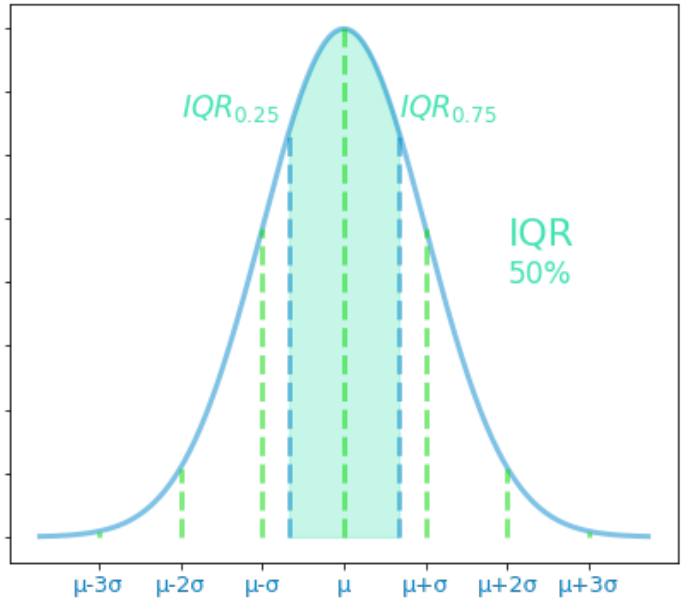
Стандартное отклонение σ
data.std()

```
sepal.length    0.828066
sepal.width     0.435866
petal.length    1.765298
petal.width     0.756221
dtype: float64
```

Описательная статистика

Квантили (квартили)
data.quantile()

```
sepal.length    5.80
sepal.width     3.00
petal.length    4.35
petal.width     1.30
Name: 0.5, dtype: float64
```



data.quantile(q=0.25)

```
sepal.length    5.1
sepal.width     2.8
petal.length    1.6
petal.width     0.3
Name: 0.25, dtype: float64
```

data.quantile(q=0.75)

```
sepal.length    6.4
sepal.width     3.3
petal.length    5.1
petal.width     1.8
Name: 0.75, dtype: float64
```

Описательная статистика

data.describe(percentiles=[.25, .5, .75,)

	sepal.length	sepal.width	petal.length	petal.width
count	150.000000	150.000000	150.000000	146.000000
mean	5.843333	3.057333	3.758000	1.173973
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

data.sem()

```
sepal.length    0.067611
sepal.width     0.035588
petal.length    0.144136
petal.width     0.062585
dtype: float64
```

Описательная статистика

data.describe(include='all')

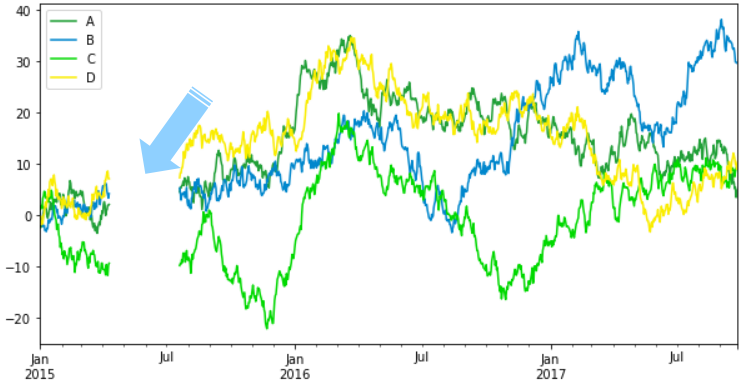
	sepal.length	sepal.width	petal.length	petal.width	variety
count	150.000000	150.000000	150.000000	146.000000	150
unique	NaN	NaN	NaN	NaN	3
top	NaN	NaN	NaN	NaN	Setosa
freq	NaN	NaN	NaN	NaN	50
mean	5.843333	3.057333	3.758000	1.173973	NaN
min	4.300000	2.000000	1.000000	0.100000	NaN
25%	5.100000	2.800000	1.600000	0.300000	NaN
50%	5.800000	3.000000	4.350000	1.300000	NaN
75%	6.400000	3.300000	5.100000	1.800000	NaN
max	7.900000	4.400000	6.900000	2.500000	NaN

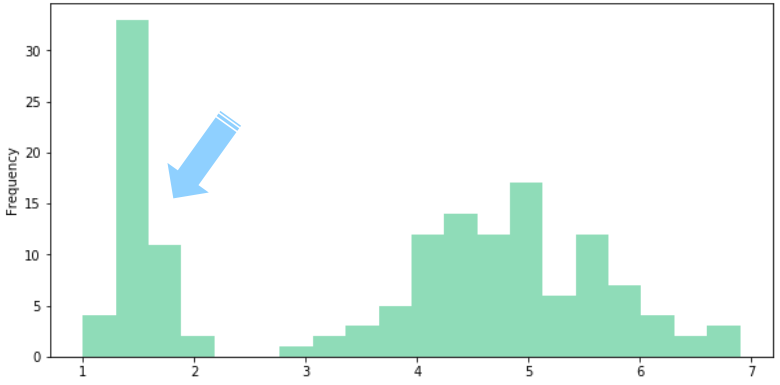
Описательная статистика

```
data. agg(
   {
      "sepal.length": ["min", "max", "median", "sem"],
      "petal.length": ["min", "max", "median", "mean", "sum", "sem"]
   })
```

	sepal.length	petal.length
max	7.900000	6.900000
mean	NaN	3.758000
median	5.800000	4.350000
min	4.300000	1.000000
sem	0.067611	0.144136
sum	NaN	563.700000

Обработка пропусков и выбросов в процессе первичного анализа статистики





- Пропуски: *NaN (null), inf, -inf, любая величина*
- Выбросы: *любая величина*

- Способы реагирования:
  - Разбиение на подгруппы: *.iloc[a:b]) и т.п.*
  - Заполнение: *.fillna(), .fillnull(), = и т.п.*
  - Удаление: *.dropna(), .dropnull, и т.п.*

СБЕР

Практика? Практика!

Когортный анализ
Корреляционный анализ
Exploratory data analysis

Когортный анализ

Когорта — это группа сущностей, имеющих общее свойство.

Например: группа людей, которая совершила нужное действие в определенный промежуток времени.

Когортный анализ — это наблюдение за когортами. Выбираем одну или несколько метрик, измеряем их и делаем выводы.

		sum	count
first_order	**order_date**		
2014-01-03	**2014-01-03**	16.448	1
	2014-11-12	153.112	1
2014-01-04	**2014-01-04**	288.060	1
2014-01-05	**2014-01-05**	19.536	1
2014-01-06	**2014-01-06**	4407.100	3

СБЕР

Практика? Практика!

Ковариация

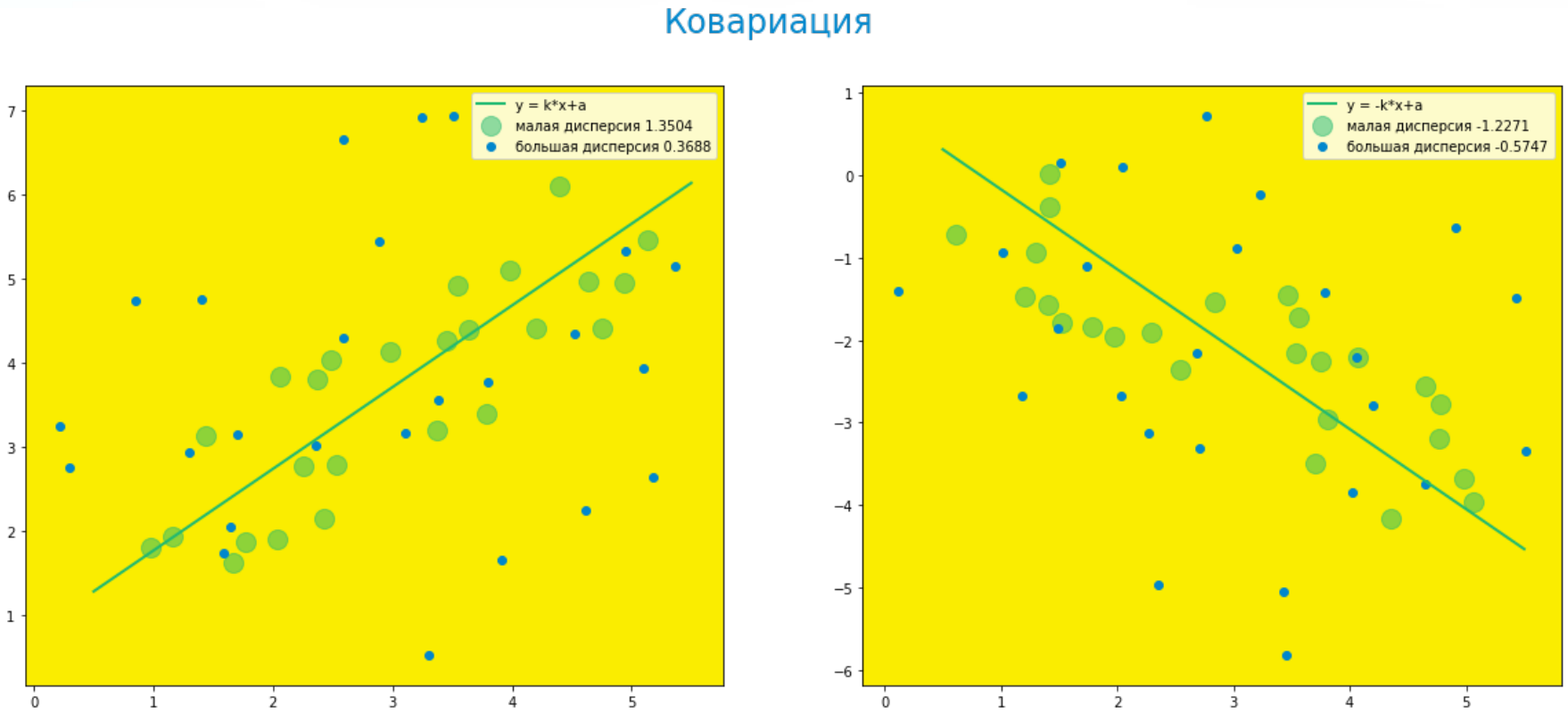
Ковариация – это мера того, как изменения одной переменной связаны с изменениями второй переменной. В частности, ковариация измеряет степень, в которой две переменные связаны линейно. Тем не менее, она также часто используется неформально как общая мера того, насколько монотонно связаны две переменные.

$$cov(X,Y) = \mathbb{E}\big((X - \mathbb{E}X)(Y - \mathbb{E}Y)\big)$$

{-1……..1}

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X - \bar{X})(Y - \bar{Y})$$

Ковариация



pd.....cov(...)

Корреляционный анализ

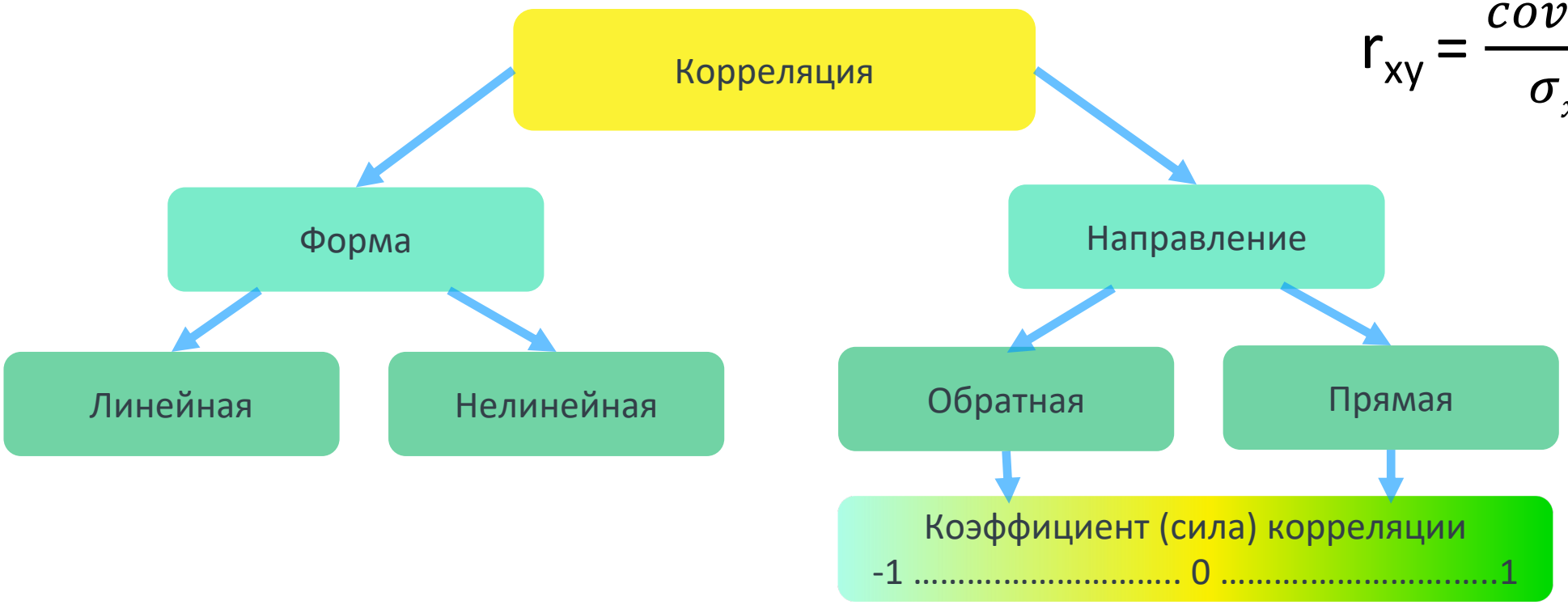
Изучение связей между переменными с точки зрения отражения соответствующих причинно-следственных отношений.

Корреляционная зависимость – это согласованные изменения двух (парная корреляционная связь) или большего количества признаков (множественная корреляционная связь). Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) другой(-их) переменной(-ых).

Корреляционный анализ – статистический метод, позволяет определить, существует ли зависимость между переменными и насколько она сильна.

Коэффициент корреляции – двумерная описательная статистика, количественная мера взаимосвязи (совместной изменчивости) двух переменных.

Корреляционный анализ

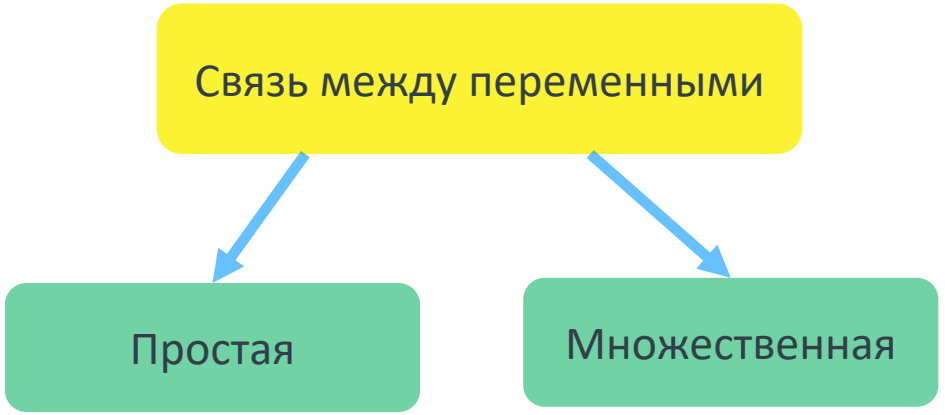


$$r_{xy} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

Корреляционный анализ

Прямая причинно-следственная связь - переменная X определяет значение переменной У.
Пример: Высота над поверхностью Земли прямо влияет на концентрацию воздуха.
Обратная причинно-следственная связь - переменная У определяет значение переменной X.
Пример: Чрезмерное потребление кофе вызывает нервозность. Или, может быть, кофе выпивается, чтобы успокоить свои нервы?
Связь, вызванная третьей (скрытой) переменной
Пример: имеется зависимость между числом утонувших людей и объёмом выпитых безалкогольных напитков летом. Однако, обе переменные связаны с жарой и потребностью людей во влаге?
Связь, вызванная несколькими скрытыми переменными
Пример: Наблюдается значимая связь между оценками студентов в университете и оценками в школе. Но влияют другие переменные: IQ, количество часов занятий, участие родителей, мотивация, квалификация преподавателей.
Связи нет, наблюдаемая зависимость случайна
Пример: Снижение количества пиратов ведет к росту средней температуры Земли .

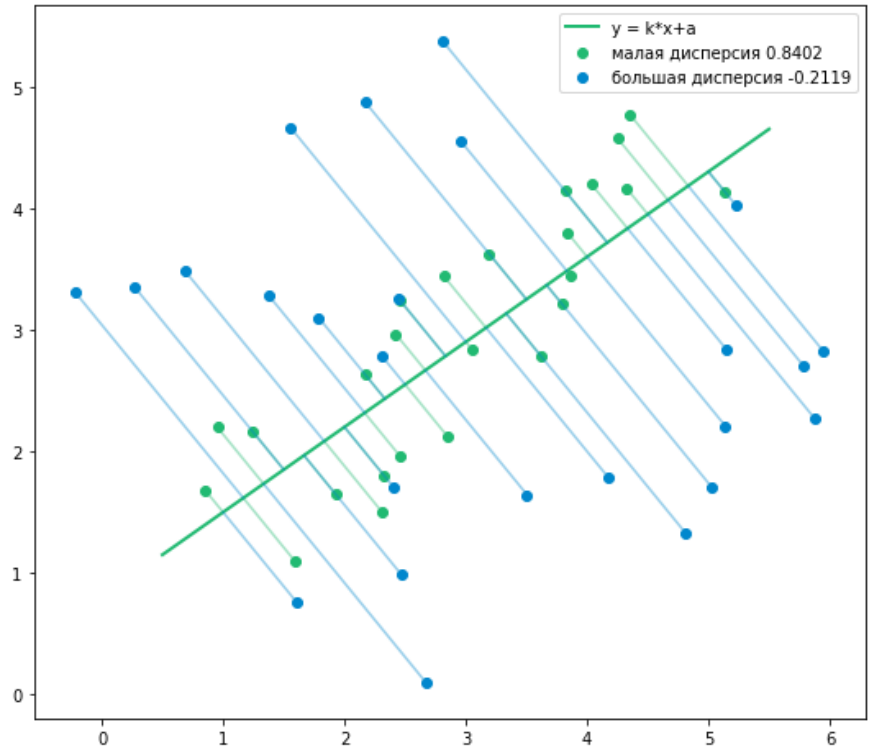
Корреляционный анализ

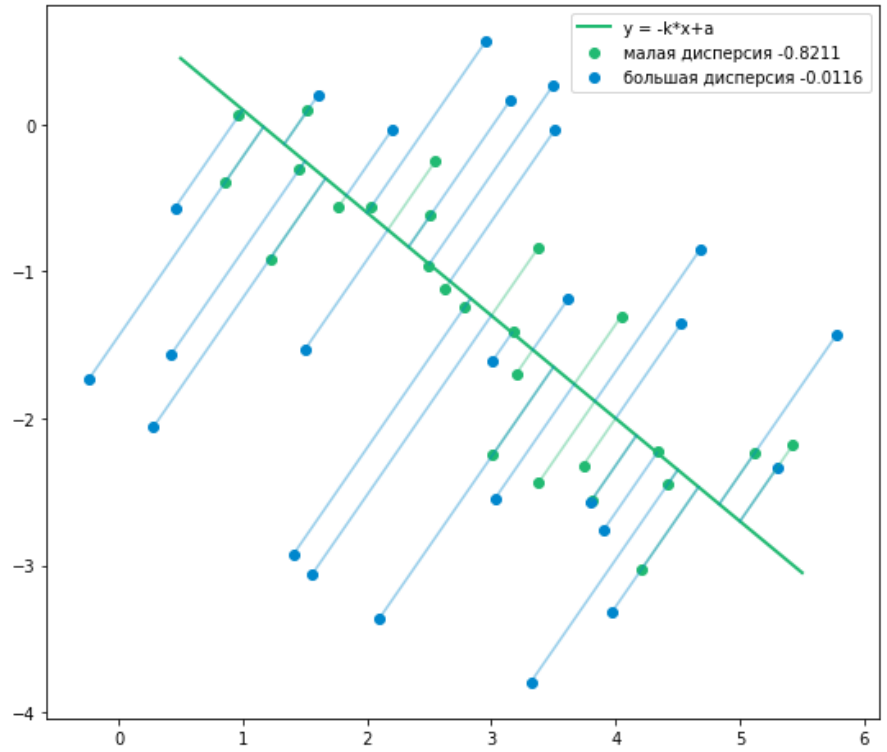


Коэфф. корреляции по модулю	Интерпретация
до 0.2	очень слабая корреляция
0.2…0.5	слабая корреляция
0.5…0.7	средняя корреляция
0.7…0.9	сильная корреляция
более 0.9	очень сильная корреляция

Корреляционный анализ

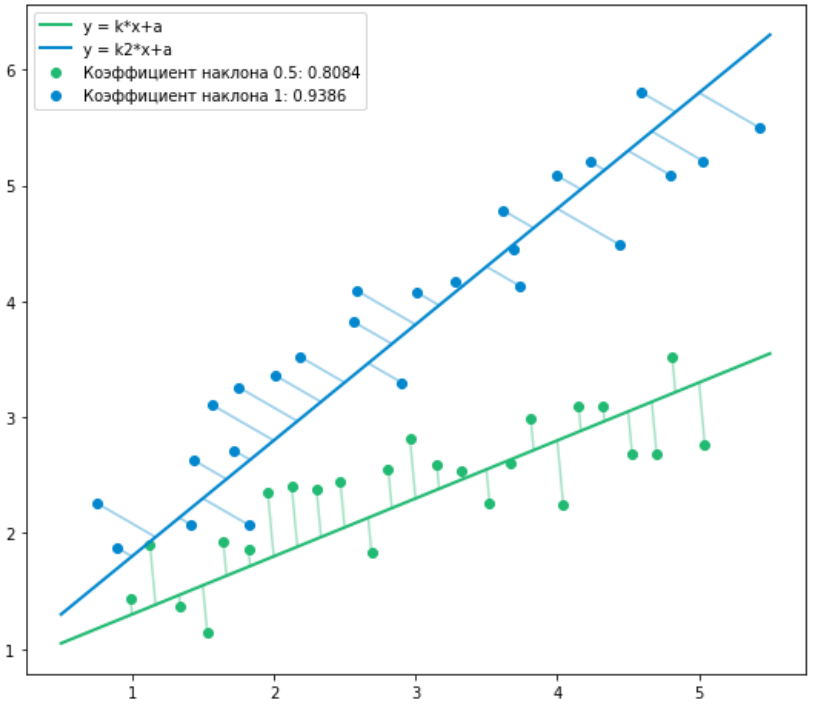
Корреляция Пирсона

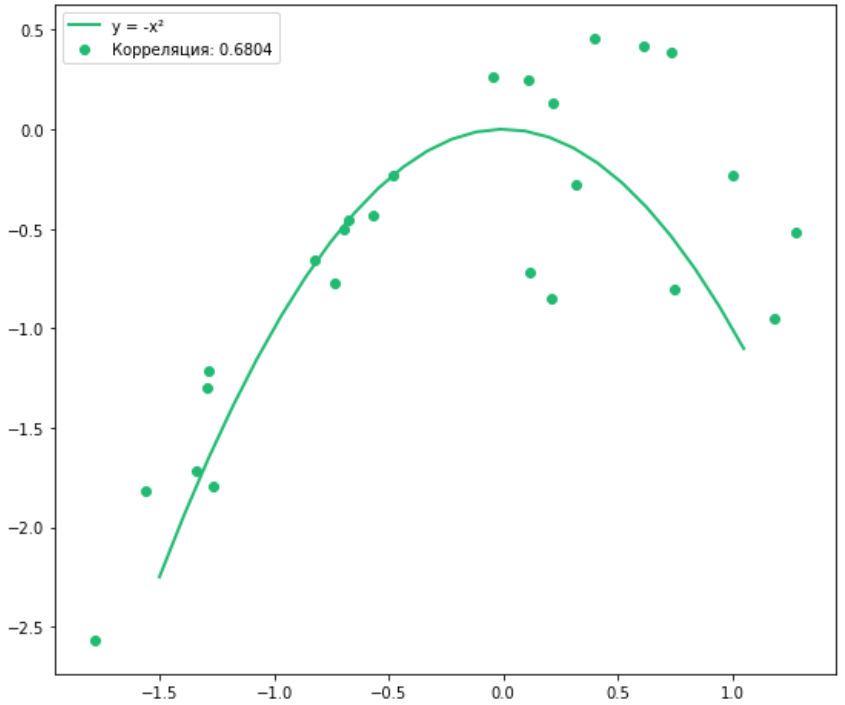




Корреляционный анализ

Корреляция Пирсона



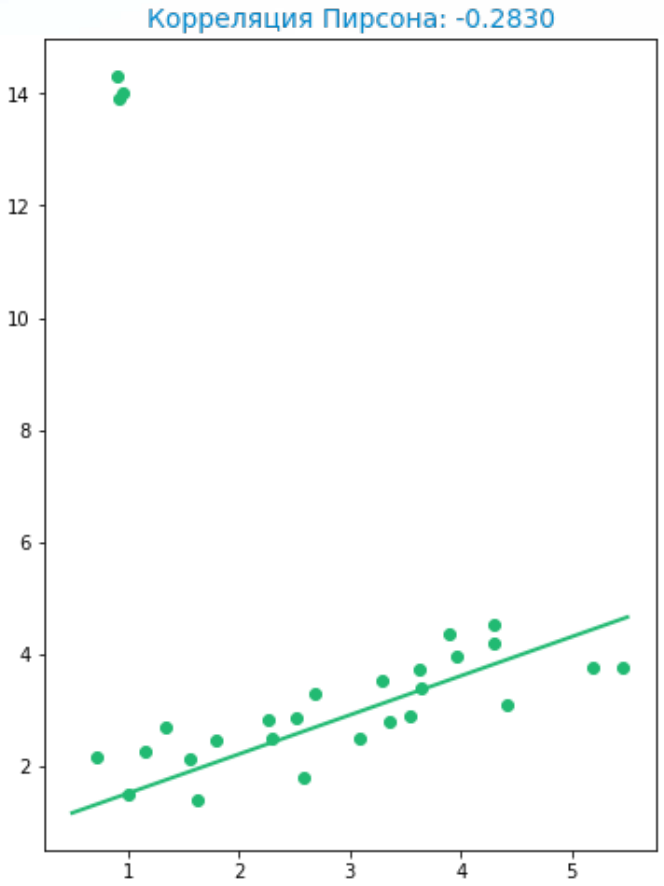


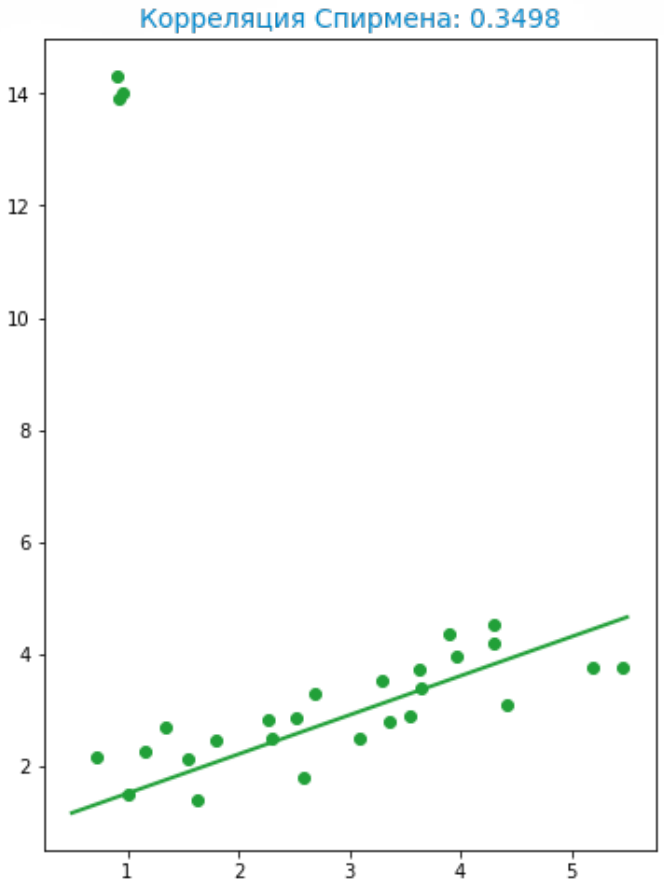
Корреляционный анализ

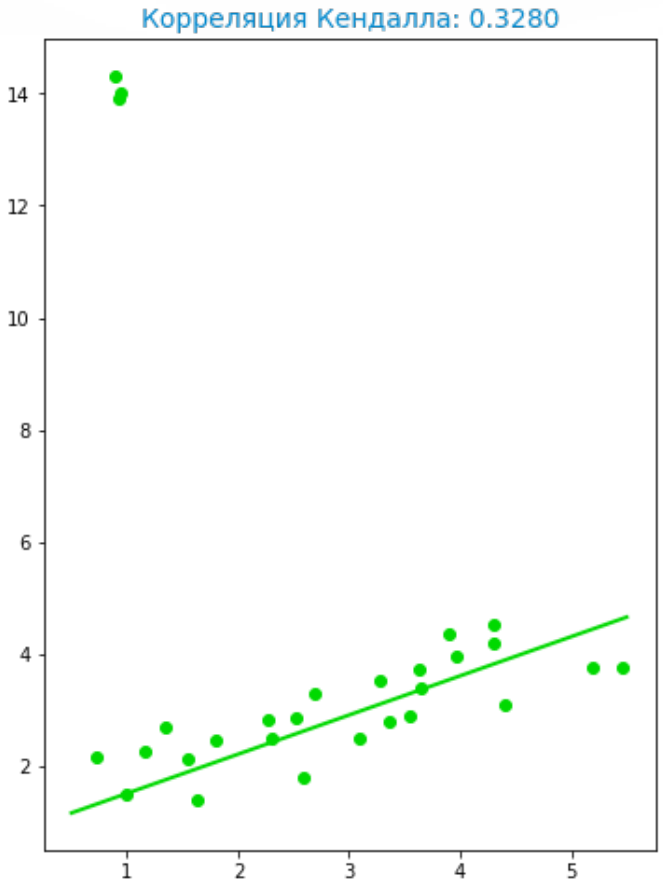
1. Для переменных с интервальной и номинальной шкалой используется коэффициент корреляции **Пирсона**.
2. Если, по меньшеё мере, одна из двух переменных имеет порядковую шкалу, либо не является нормально распределенной, используется ранговая корреляция **Спирмана** или **Кендалла**.

Тип шкал		**Мера связи**
Переменная 1	**Переменная 2**	
Интервальная Номинальная	Интервальная Номинальная	Коэффициент Пирсона
Ранговая Интервальная Номинальная	Ранговая Интервальная Номинальная	Коэффициент Спирмена
Ранговая	Ранговая	Коэффициент Кендалла

Корреляционный анализ







Корреляционный анализ

Коэффициент корреляции Пирсона оценивает только **линейную связь** переменных. Нелинейную связь данный коэффициент выявить не может.

Коэффициент корреляции Пирсона очень **чувствителен к выбросам** (outliers).

Корреляция **не подразумевает** наличия **причинно-следственной связи** между переменными.

Нельзя путать коэффициент корреляции Пирсона с критерием Пирсона ХИ-квадрат.

Корреляционный анализ

Коэффициент корреляции Пирсона оценивает только **линейную связь** переменных. Нелинейную связь данный коэффициент выявить не может.

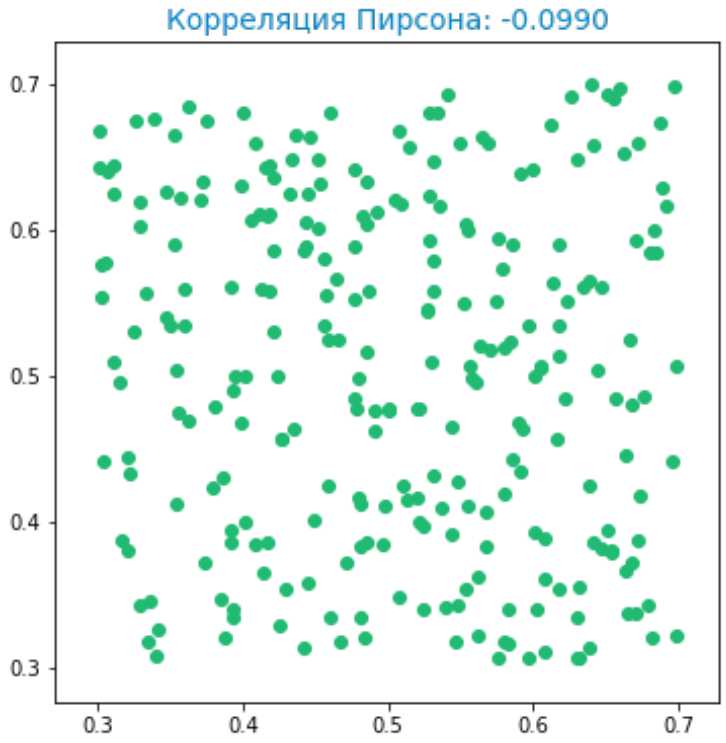
Коэффициент корреляции Пирсона очень **чувствителен к выбросам** (outliers).

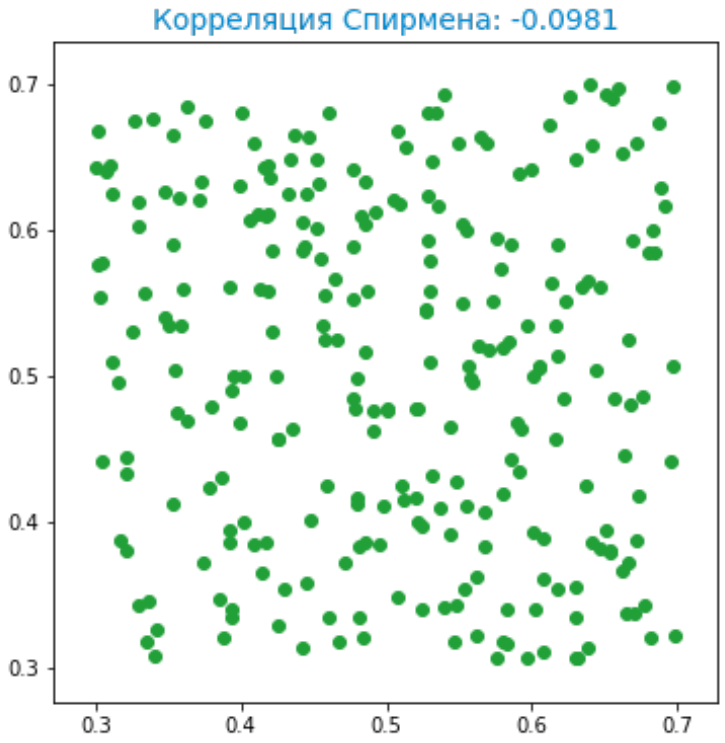
Корреляция **не подразумевает** наличия **причинно-следственной связи** между переменными.

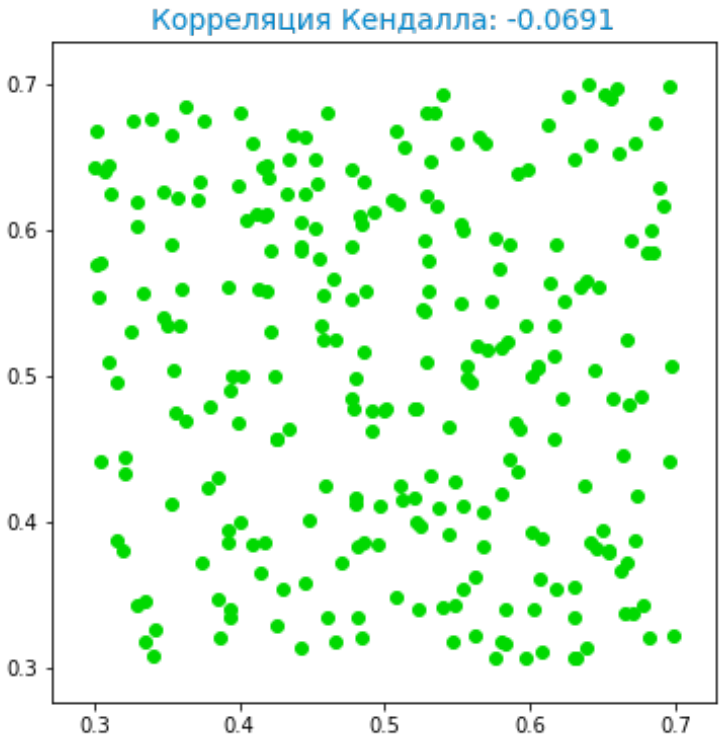
Коэффициенты корреляции Спирмена и Кендалла используются как меры взаимозависимости между **рядами рангов**, а не как меры связи между самими переменными.

Коэффициенты Спирмена и Кендалла обладают примерно одинаковыми свойствами, но коэффициент Кендалла в случае **многих рангов**, а также при введении **дополнительных объектов** в ходе исследования имеет определенные вычислительные преимущества.

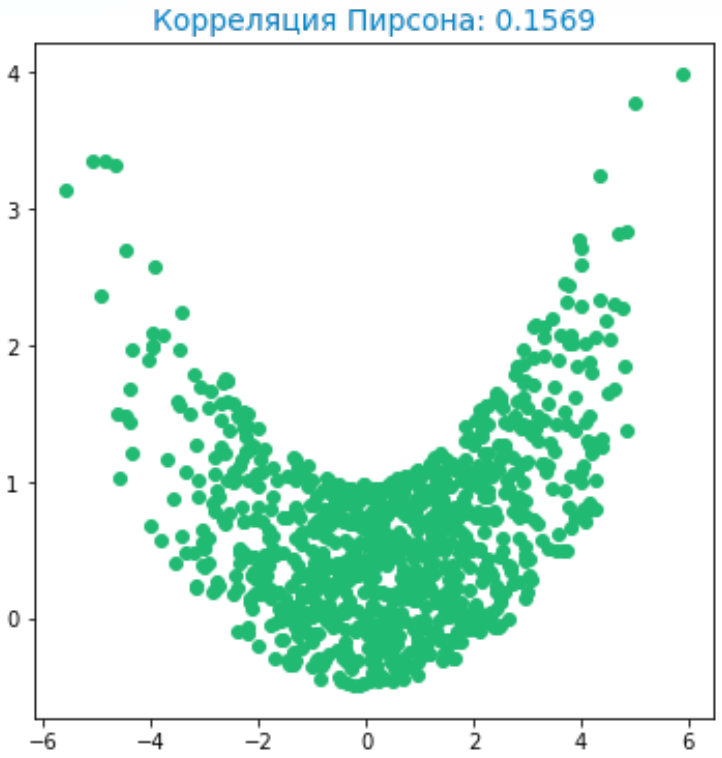
Корреляционный анализ

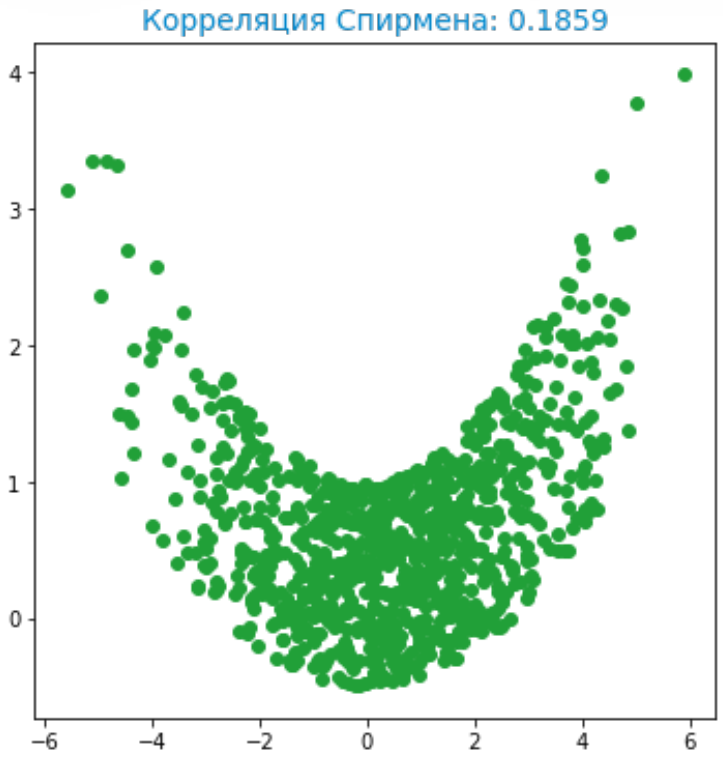


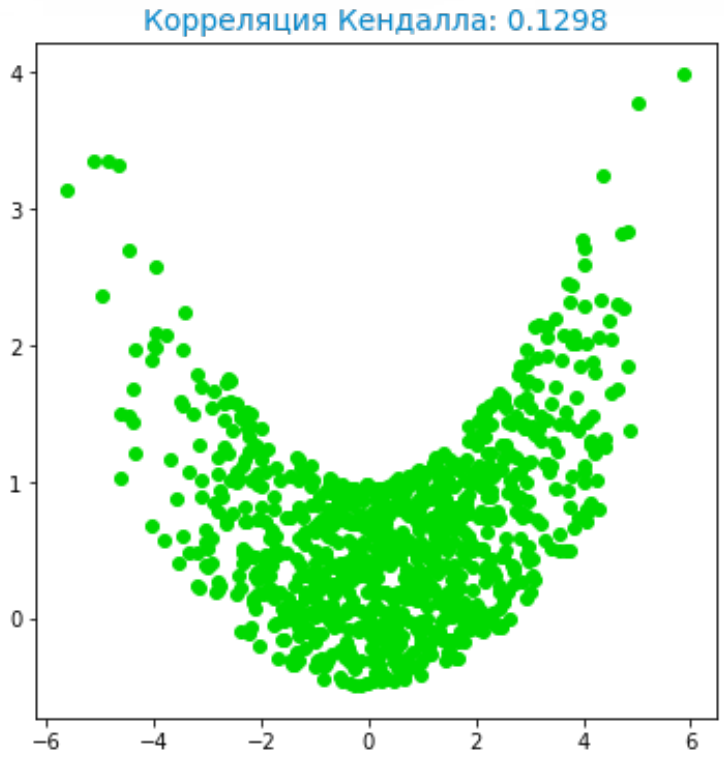




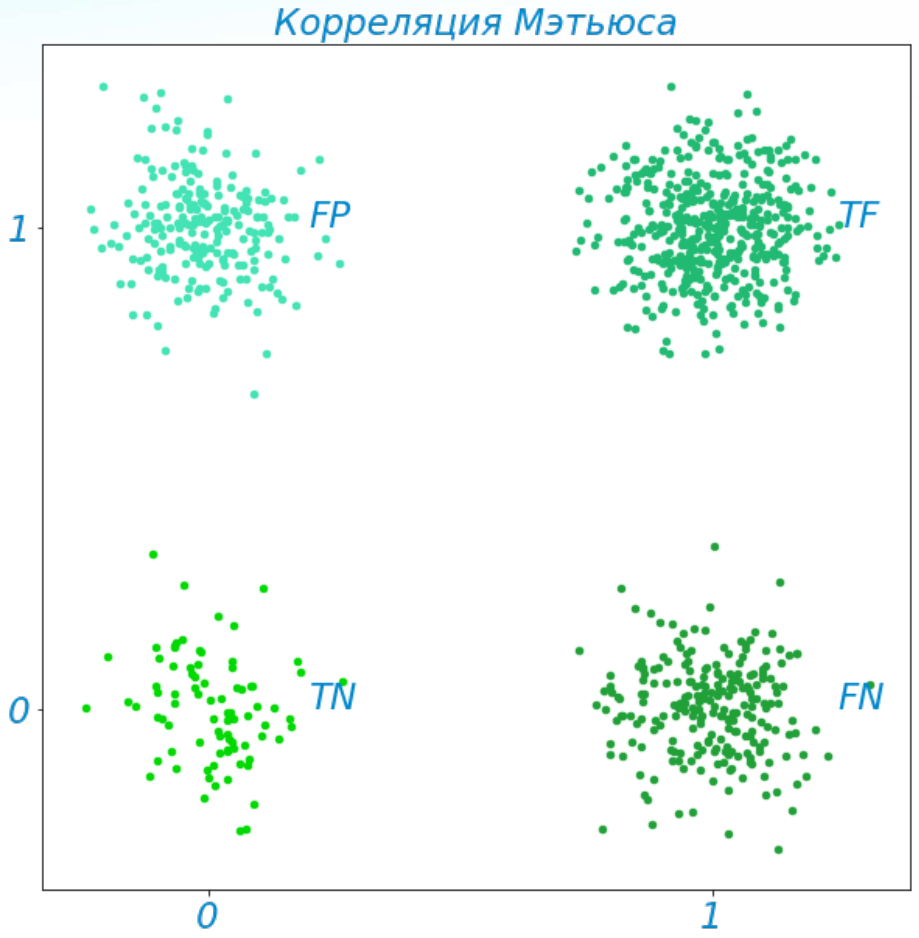
Корреляционный анализ







Корреляционный анализ

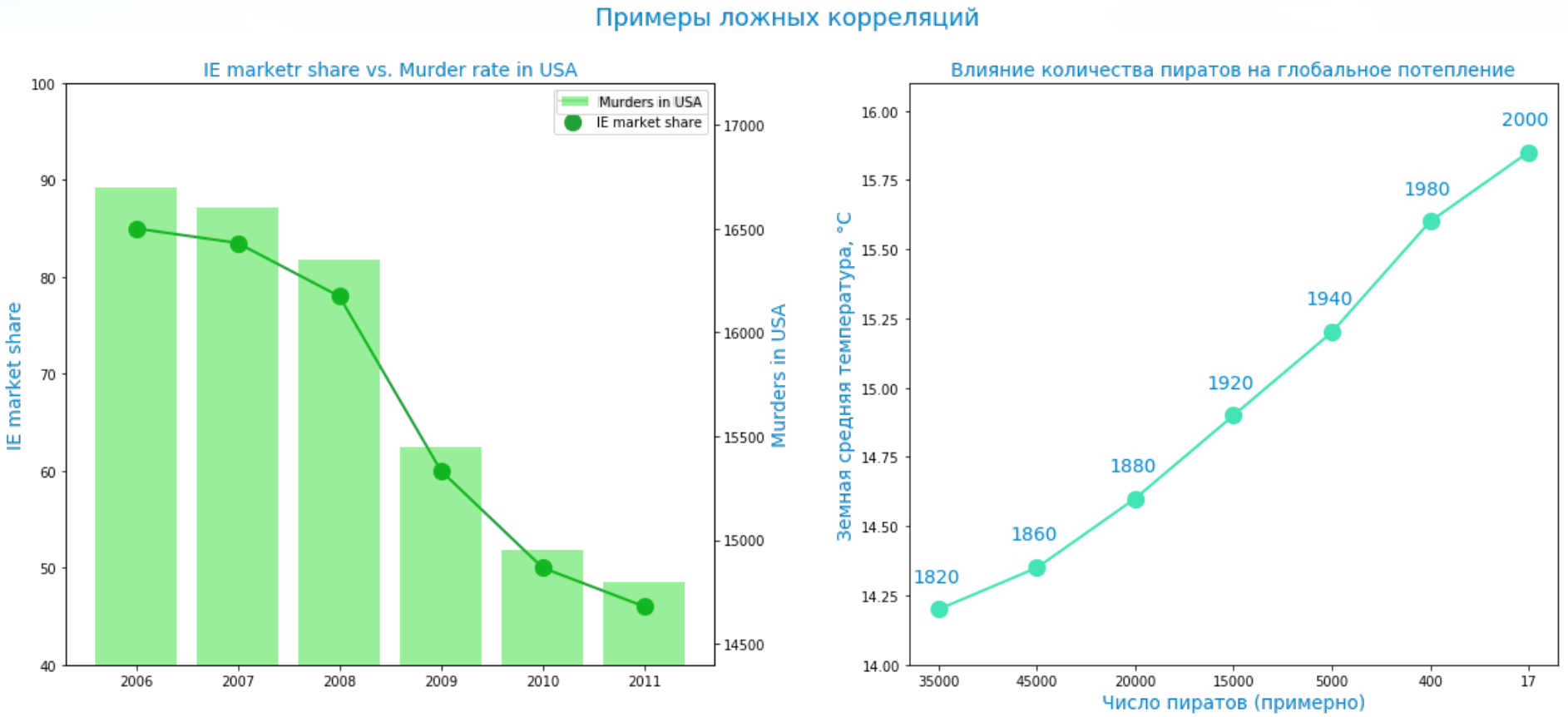


	X2=False	X2=True
X1=False	TN	FP
X1=True	FN	TP

$$MCC = \frac{(TN * TP) - (FN * FP)}{\sqrt{(TN + FT)(TN + FN)(TP + FT)(TP + FP)}}$$

pd……corr(…..)
sklearn.metrics.matthews_corrcoef(X1,X2)

Корреляционный анализ



СБЕР

Практика? Практика!

Exploratory data analysis

EDA - анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, в том числе с использованием инструментов визуализации

Основные цели EDA:

- максимальное «проникновение» в данные
- выявление основных структур
- выбор наиболее важных переменных
- обнаружение отклонений и аномалий
- проверка основных гипотез
- разработка начальных моделей

Exploratory data analysis

EDA основные шаги:

- Получение данных о переменных *(Understanding your variables)*
  - Описательная статистика
- Очистка данных *(Cleaning your dataset)*
  - Removing Redundant variables
  - Variable Selection
  - Removing Outliers
- Анализ взаимосвязей между переменными *(Analyzing relationships between variables)*
  - Correlation Matrix
  - Scatterplot
  - Histogram
  - Boxplot…

СБЕР

Практика? Практика!

Резюме

- Изучили первичный анализ и очистка данных
- Познакомились с когортным анализом
- Прошли корреляционный анализ
- Понимаем, что подразумевает Exploratory data analysis

Полезные ссылки

Примеры ложных корреляций
https://mi3ch.livejournal.com/2559227.html
Функции для работы с распределениями в Pandas:
https://pandas.pydata.org/docs/user_guide/basics.html

Обратная связь

?

СБЕР

Спасибо за внимание!