Data Science: что такое анализ данных?

План курса

Введение в анализ данных

Поговорим про данные и как с ними работать

Python практика

Разберем самые популярные библиотеки

Python практика (2)

Продолжим практиковаться

Машинное обучение

Разберем простые алгоритмы и обучим свои модели

Практика

Применим все что узнали для работы с одним из датасетов

Данные – это что?

Это все что нас окружает!

Информация о покупках, фотографии товаров в маркетплейсах, ассортимент магазинов и вообще всё, что содержит в себе какую-то (не обязательно) полезную информацию

Так а почему все говорят про анализ данных?

Данных ОЧЕНЬ много!

Нам бы хотелось бы уметь вытаскивать то, что может нам полезно.

Задача 1 – определить, что нам действительно может быть полезно

Давайте представим, мы аналитики маркетплейса

Мы хотим начать оценить как хорошо будет продаваться какой-то товар

Какая информация может быть нам для этого полезна?

Давайте представим, мы аналитики маркетплейса

Мы хотим начать оценить как хорошо будет продаваться какой-то товар

Какая информация может быть нам для этого полезна?

1. Посмотреть схожие товары - какие люди покупают их - в целом полезно изучить аудиторию и конкурентов
2. Характеристики товара (может быть мы знаем какие-то предпочтения клиентов)
3. Сезонность

Основные этапы работы

1. Настроить хранение данных
    a. Если мы крупная компания - создаем огромные сервера с терабайтами данных
    b. Если это проект поменьше - собрать все данные воедино, настроить их удобное отображение - например, с помощью библиотеки Python ***pandas***

Основные этапы работы

1. Настроить хранение данных
   a. Если мы крупная компания - создаем огромные сервера с терабайтами данных
   b. Если это проект поменьше - собрать все данные воедино, настроить их удобное отображение - например, с помощью библиотеки Python ***pandas***
2. Обработка данных и их структурирование
   a. Убираем ненужные данные (можете ли вы привести пример?)
   b. Преобразуем данные в удобный для хранения формат
   c. Проверяем гипотезы

Основные этапы работы

1. Настроить хранение данных
   a. Если мы крупная компания - создаем огромные сервера с терабайтами данных
   b. Если это проект поменьше - собрать все данные воедино, настроить их удобное отображение - например, с помощью библиотеки Python ***pandas***
2. Обработка данных и их структурирование
   a. Убираем ненужные данные (можете ли вы привести пример?)
   b. Преобразуем данные в удобный для хранения формат
   c. Проверяем гипотезы
3. Построение моделей машинного обучения для решения нашей задачи
   a. Например, предсказание оценки студента по его успеваемости в прошлом

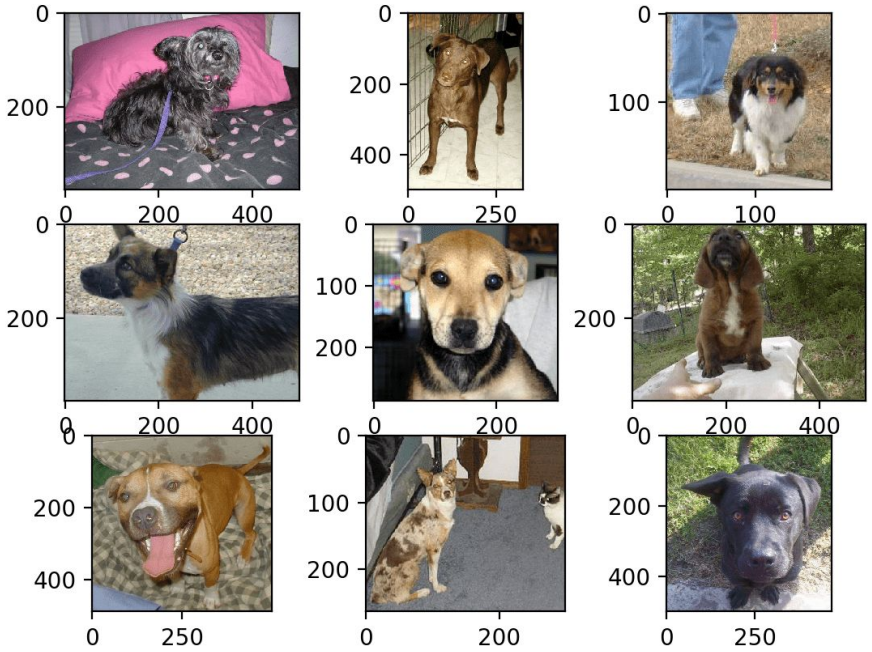
Типы данных

1. Табличные данные

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Типы данных

2. Изображения
3. Текст
4. Аудио



Компьютер лучше всего работает с числами!

Поэтому наша цель – преобразовывать все типы данных в числа

Рассмотрим пример!

[Таня пошла в магазин] – как это закодировать, используя только числа?

Идея

	Таня	Пошла	В	магазин
Таня	1	0	0	0
Ваня	0	0	0	0
идти	0	1	0	0
бежать	0	0	0	0
в	0	0	1	0
магазин	0	0	0	1
рынок	0	0	0	0

Идея

	Таня	Пошла	В	магазин
Таня	1	0	0	0
Ваня	0	0	0	0
идти	0	1	0	0
бежать	0	0	0	0
в	0	0	1	0
магазин	0	0	0	1
рынок	0	0	0	0

1
0
1
0
1
1
0

Идея

	Таня	Пошла	В	магазин
Таня	1	0	0	0
Ваня	0	0	0	0
идти	0	1	0	0
бежать	0	0	0	0
в	0	0	1	0
магазин	0	0	0	1
рынок	0	0	0	0

1
0
1
0
1
1
0

Получили вектор

[1, 5, 0, 45, 18, 64] - соответствует одному тексту / строке таблицы

Если мы возьмем много разных векторов и как бы заполним ими таблицу - получим матрицу

Вернемся к нашей табличке

Выжил пассажир или нет?

1. Идентификатор пассажира

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Выжил пассажир или нет?

1. Идентификатор пассажира
2. Класс обслуживания (эконом / бизнес / первый класс)

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Выжил пассажир или нет?

1. Идентификатор пассажира
2. Класс обслуживания (эконом / бизнес / первый класс)
3. Имя пассажира

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Выжил пассажир или нет?

1. Идентификатор пассажира
2. Класс обслуживания (эконом / бизнес / первый класс)
3. Имя пассажира
4. Пол пассажира

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Выжил пассажир или нет?

1. Идентификатор пассажира
2. Класс обслуживания (эконом / бизнес / первый класс)
3. Имя пассажира
4. Пол пассажира
5. Возраст пассажира

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Выжил пассажир или нет?

1. Идентификатор пассажира
2. Класс обслуживания (эконом / бизнес / первый класс)
3. Имя пассажира
4. Пол пассажира
5. Возраст пассажира
6. Количество братьев / сестер на борту + количество родителей / детей на борту

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Выжил пассажир или нет?

1. Идентификатор пассажира
2. Класс обслуживания (эконом / бизнес / первый класс)
3. Имя пассажира
4. Пол пассажира
5. Возраст пассажира
6. Количество братьев / сестер на борту + количество родителей / детей на борту
7. Номер билета

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Выжил пассажир или нет?

1. Идентификатор пассажира
2. Класс обслуживания (эконом / бизнес / первый класс)
3. Имя пассажира
4. Пол пассажира
5. Возраст пассажира
6. Количество братьев / сестер на борту + количество родителей / детей на борту
7. Номер билета
8. Стоимость билета

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Выжил пассажир или нет?

1. Идентификатор пассажира
2. Класс обслуживания (эконом / бизнес / первый класс)
3. Имя пассажира
4. Пол пассажира
5. Возраст пассажира
6. Количество братьев / сестер на борту + количество родителей / детей на борту
7. Номер билета
8. Стоимость билета
9. Номер каюты

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Выжил пассажир или нет?

1. Идентификатор пассажира
2. Класс обслуживания (эконом / бизнес / первый класс)
3. Имя пассажира
4. Пол пассажира
5. Возраст пассажира
6. Количество братьев / сестер на борту + количество родителей / детей на борту
7. Номер билета
8. Стоимость билета
9. Имя каюты
10. Порт, где пассажир взошел на борт

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

После преобразований – матрица

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	1	1	0	38.0	1	0	71.2833	0
1	1	1	0	35.0	1	0	53.1000	1
2	0	1	1	54.0	0	0	51.8625	1
3	1	3	0	4.0	1	1	16.7000	1
4	1	1	0	58.0	0	0	26.5500	1

Что нам было бы интересно посмотреть в данных?

Попробуем побыть аналитиками – предложите свои гипотезы, которые могли бы быть полезны для дальнейшей работы с данными.

Например,
Сколько в среднем выжило мужчин, а сколько женщин?

Ваши идеи?

О чем важно помнить?

1. Чтобы сделать данными идеальными нужно много работы, часто это настоящее искусство, требующее много знаний об области, из которой приходят данные
2. Прежде чем сделать выводы нужно убедиться, что мы видим не искаженную картину мира

На картинке показана доля выживших мужчин и доля выживших женщин - о чем нам говорят данные?

	Survived
Gender	
female	0.752896
male	0.205298

dtype: float64

На самом деле....

Процент выживаемости

Gender	
female	0.752896
male	0.205298

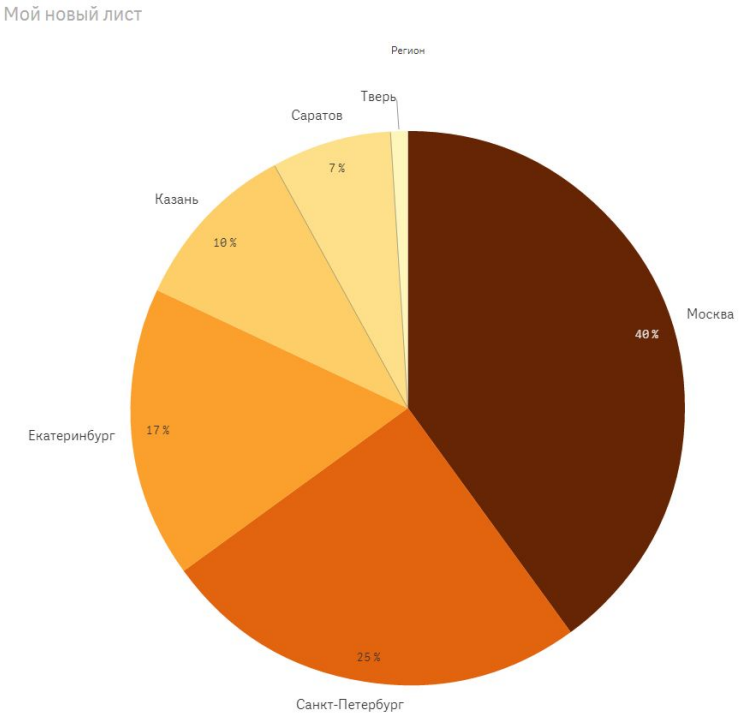
Количество человек

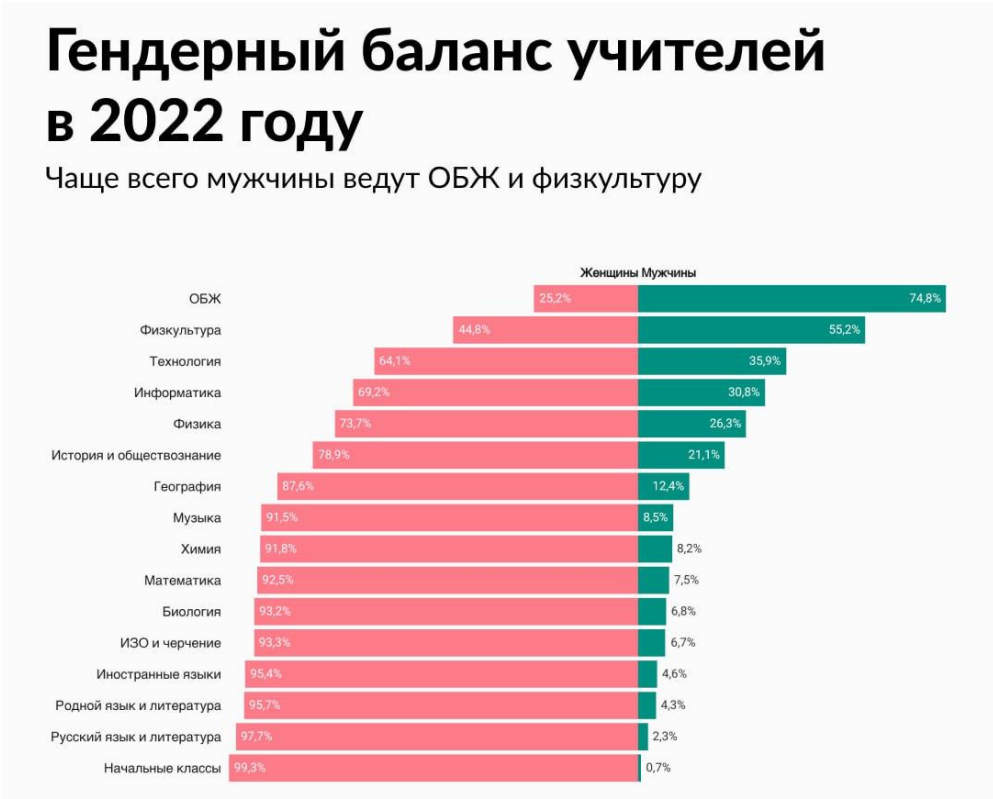
Gender	
female	259
male	453

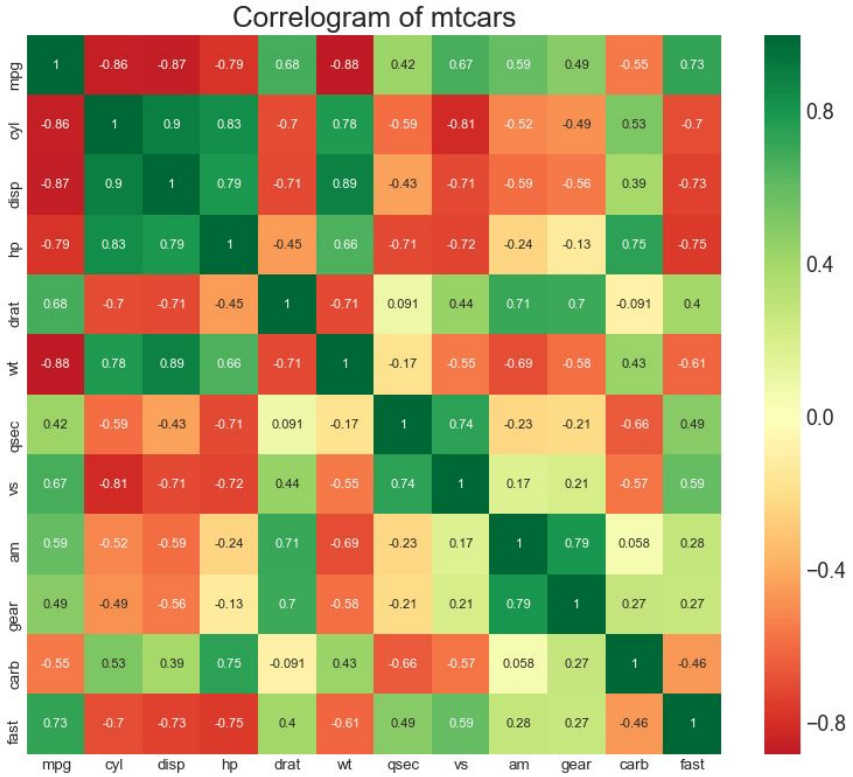
Мужчин было больше почти в два раза, поэтому понятно, что процент выживаемости у них будет меньше

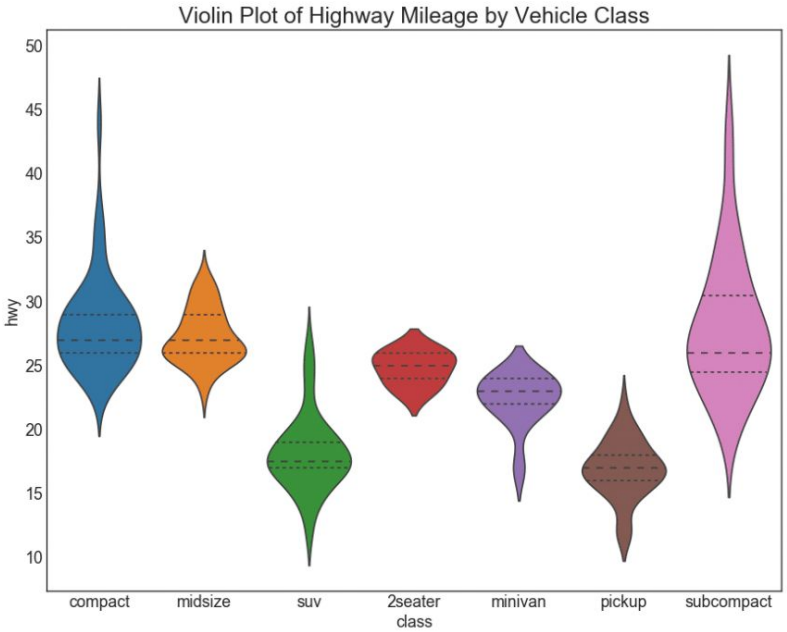
Примеры визуализаций

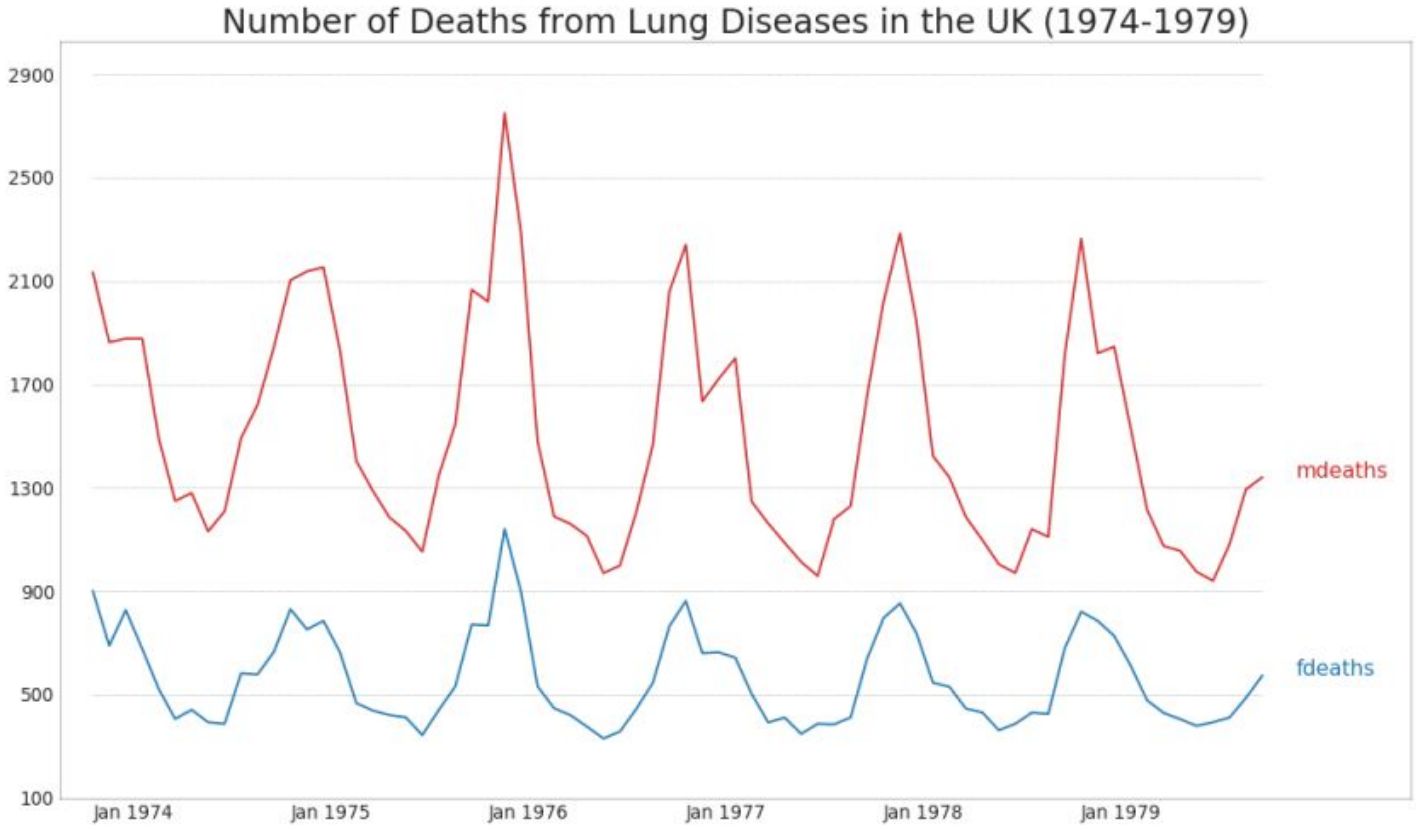
Для разных задач – разные типы визуализаций






Correlogram of mtcars
mpg 1 -0.86 -0.87 -0.79 0.68 -0.88 0.42 0.67 0.59 0.49 -0.55 0.73
cyl -0.86 1 0.9 0.83 -0.7 0.78 -0.59 -0.81 -0.52 -0.49 0.53 -0.7
disp -0.87 0.9 1 0.79 -0.71 0.89 -0.43 -0.71 -0.59 -0.56 0.39 -0.73
hp -0.79 0.83 0.79 1 -0.45 0.66 -0.71 -0.72 -0.24 -0.13 0.75 -0.75
drat 0.68 -0.7 -0.71 -0.45 1 -0.71 0.091 0.44 0.71 0.7 -0.091 0.4
wt -0.88 0.78 0.89 0.66 -0.71 1 -0.17 -0.55 -0.69 -0.58 0.43 -0.61
qsec 0.42 -0.59 -0.43 -0.71 0.091 -0.17 1 0.74 -0.23 -0.21 -0.66 0.49
vs 0.67 -0.81 -0.71 -0.72 0.44 -0.55 0.74 1 0.17 0.21 -0.57 0.59
am 0.59 -0.52 -0.59 -0.24 0.71 -0.69 -0.23 0.17 1 0.79 0.058 0.28
gear 0.49 -0.49 -0.56 -0.13 0.7 -0.58 -0.21 0.21 0.79 1 0.27 0.27
carb -0.55 0.53 0.39 0.75 -0.091 0.43 -0.66 -0.57 0.058 0.27 1 -0.46
fast 0.73 -0.7 -0.73 -0.75 0.4 -0.61 0.49 0.59 0.28 0.27 -0.46 1
mpg cyl disp hp drat wt qsec vs am gear carb fast
0.8
0.4
0.0
−0.4
−0.8


Violin Plot of Highway Mileage by Vehicle Class
hwy
50
45
40
35
30
25
20
15
10
compact
midsize
suv
2seater
minivan
pickup
subcompact
class


Number of Deaths from Lung Diseases in the UK (1974-1979)
2900
2500
2100
1700
1300
900
500
100
Jan 1974
Jan 1975
Jan 1976
Jan 1977
Jan 1978
Jan 1979
mdeaths
fdeaths

Вопросы?