

Машинное обучение

Лекция 2. Разведочный анализ данных

Автор: Рустам Азимов

Санкт-Петербург, 2023г.

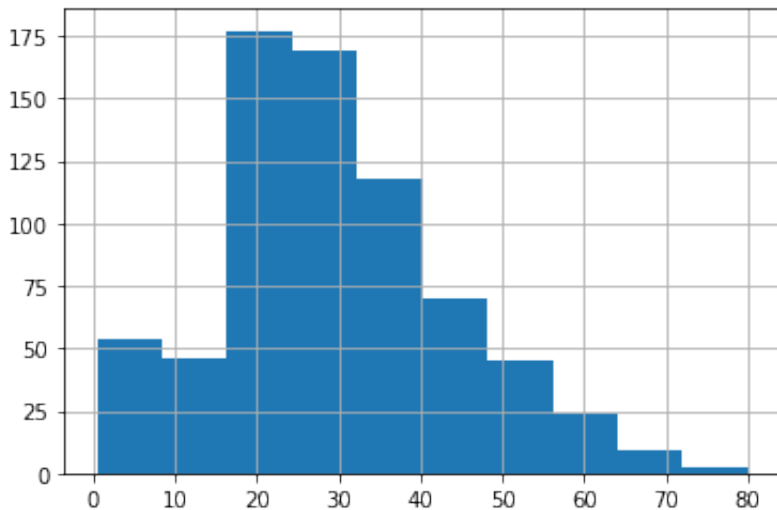
- ▶ **Разведочный анализ данных (Exploratory Data Analysis)** — предварительное исследование выборки с целью
 - ▶ определения его основных характеристик
 - ▶ выявления некоторых зависимостей в данных
 - ▶ определения важности признаков для поставленной задачи
 - ▶ создания новых признаков, преобразований данных
 - ▶ сужения набора семейств алгоритмов ML для рассмотрения

- ▶ Обработка пропущенных значений
- ▶ Откладывание избыточной информации в сторону
- ▶ Преобразования над значениями признака, приведение к нужному виду, нормализация
- ▶ Количественный признак → Категориальный признак

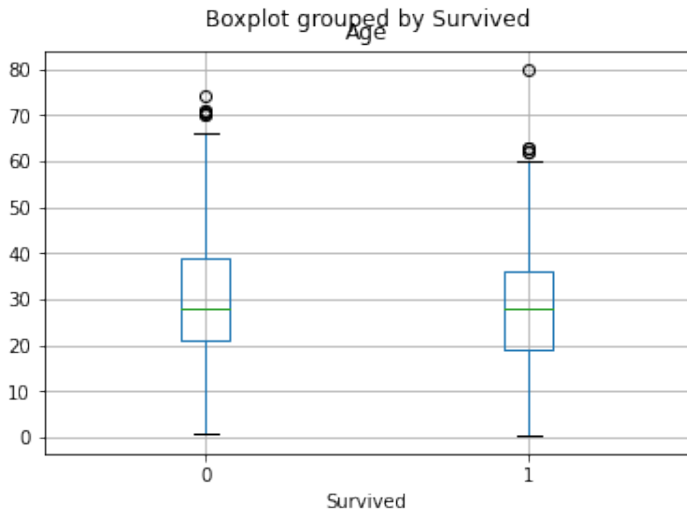
- ▶ Анализ признаков, их важности, зависимостей
 - ▶ Одномерный анализ
 - ▶ Двумерный анализ
- ▶ Создание новых признаков, их анализ

- ▶ Количественные признаки
 - ▶ Гистограммы (`hist()`)
 - ▶ Ящики с усами (`boxplot()`)
- ▶ Категориальные признаки
 - ▶ `countplot()`, `barplot()`

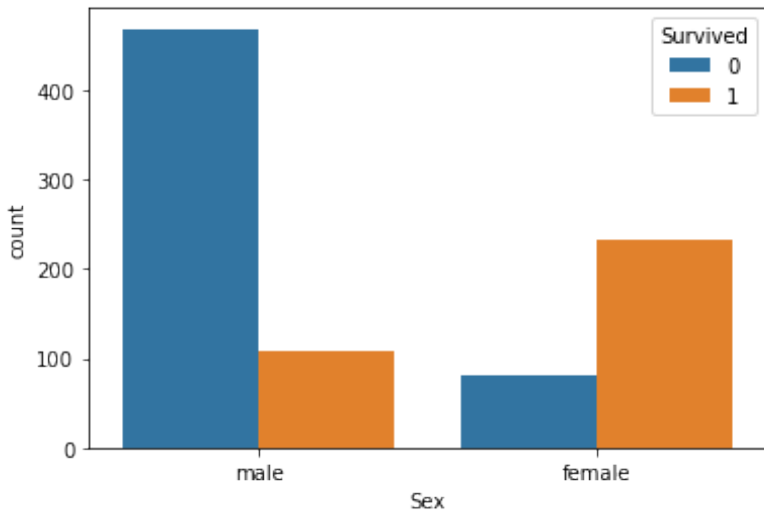
Гистограммы



Boxplot

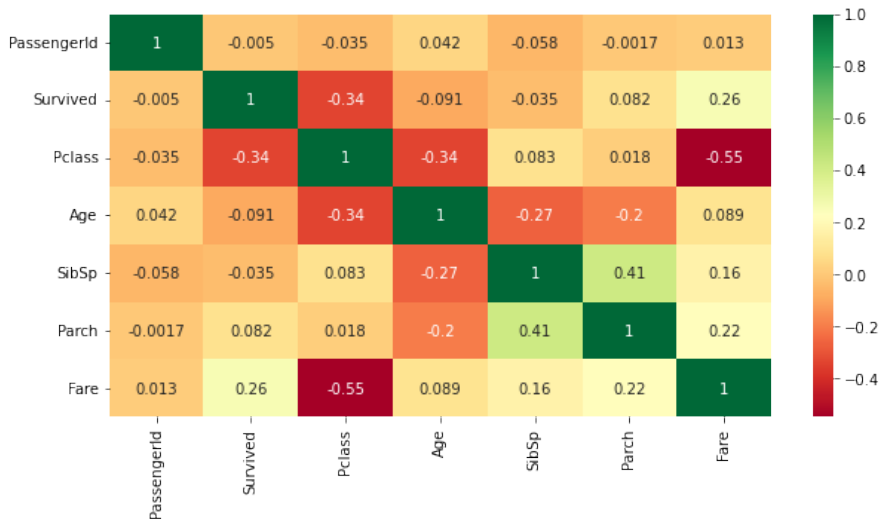


Countplot

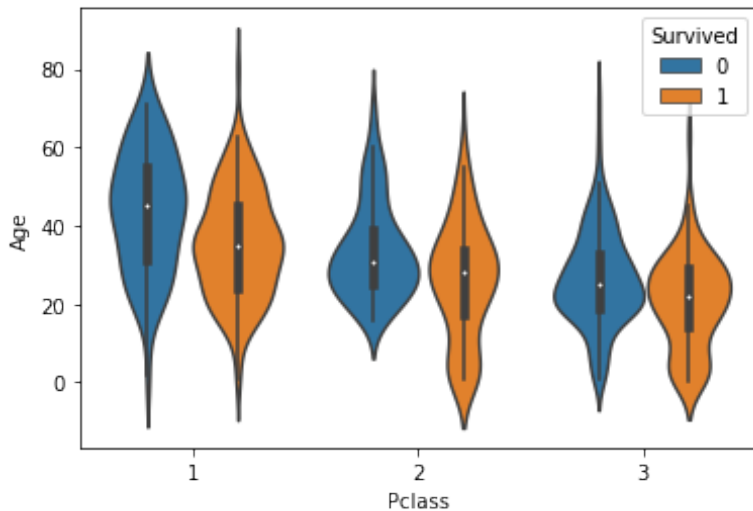


- ▶ Количественный vs Количественный
- ▶ Количественный vs Категориальный
- ▶ Категориальный vs Категориальный

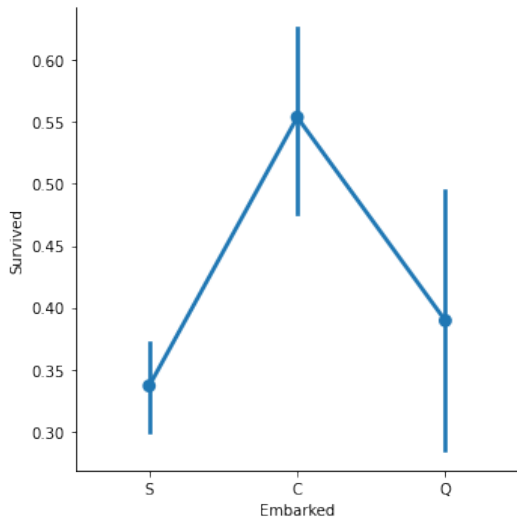
Количественный vs Количественный (corr)



Количественный vs Категориальный (violinplot)



Категориальный vs Категориальный (factorplot)



- ▶ machinelearning.ru
- ▶ mlcourse.ai, dlcourse.ai
- ▶ [kaggle](https://www.kaggle.com/)