Машинное обучение

Лекция 2. Разведочный анализ данных

Автор: Рустам Азимов

Санкт-Петербургский государственный университет

Санкт-Петербург

Exploratory Data Analysis

- **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

Data Cleaning

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

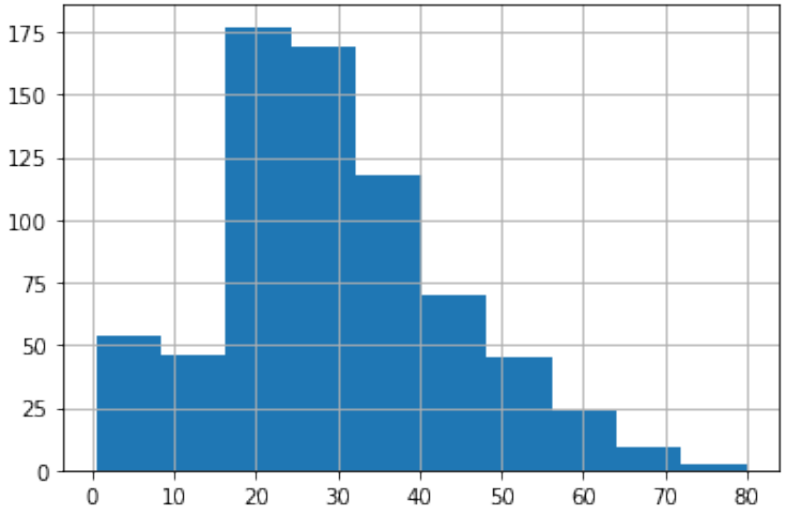
Feature Engineering

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

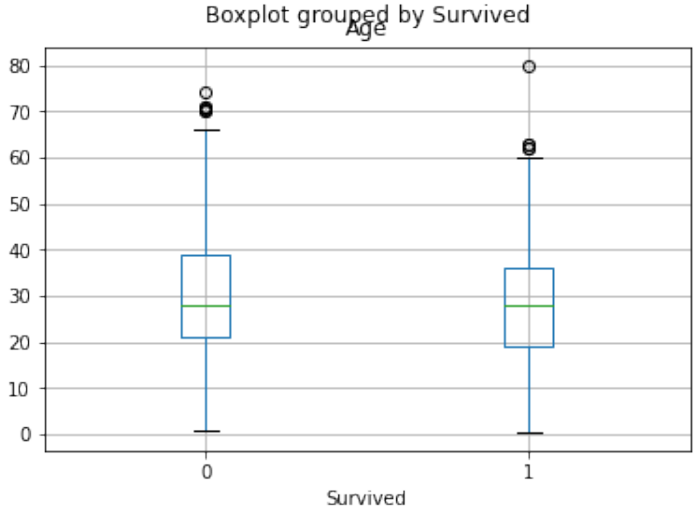
Одномерный анализ

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

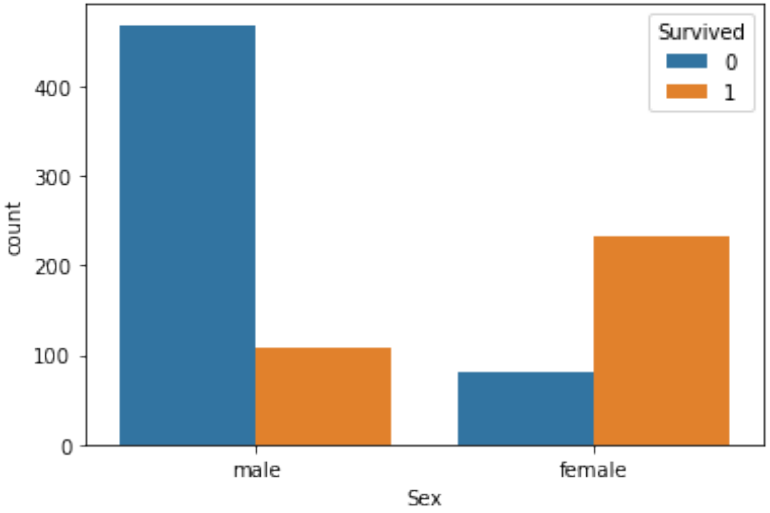
Гистограммы



Boxplot


Boxplot grouped by Survived
Age
80
70
60
50
40
30
20
10
0
0
1
Survived

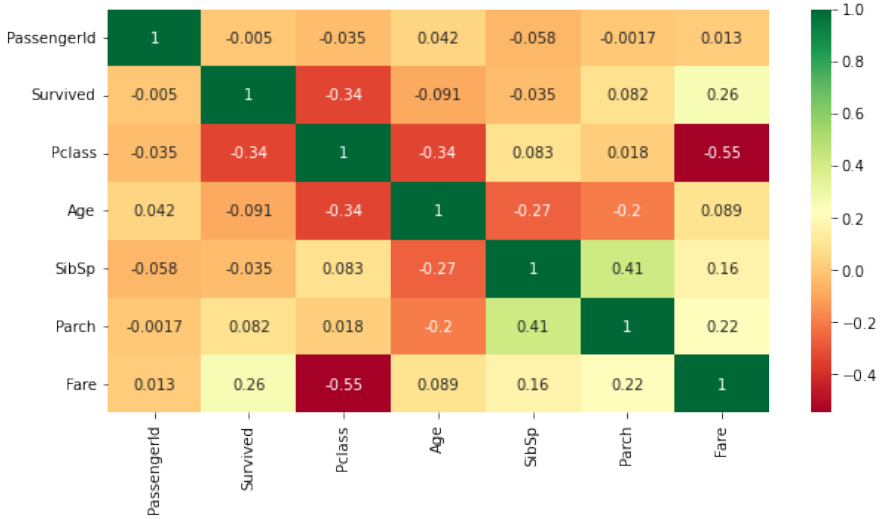
Countplot



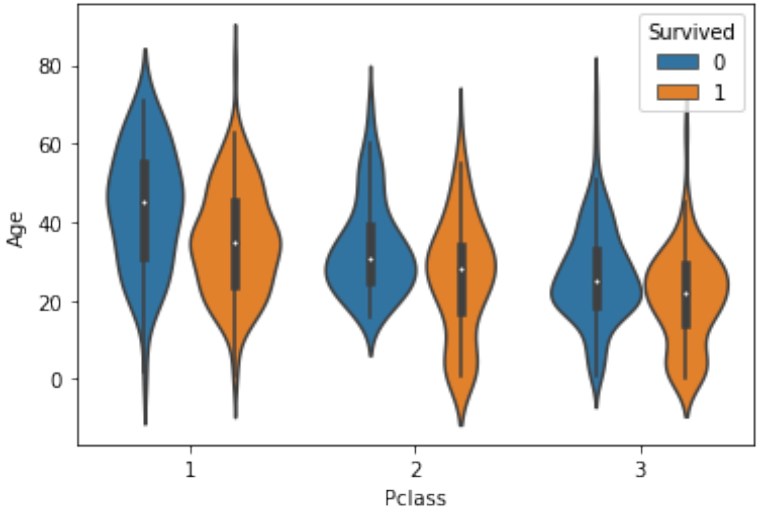
Двумерный анализ

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

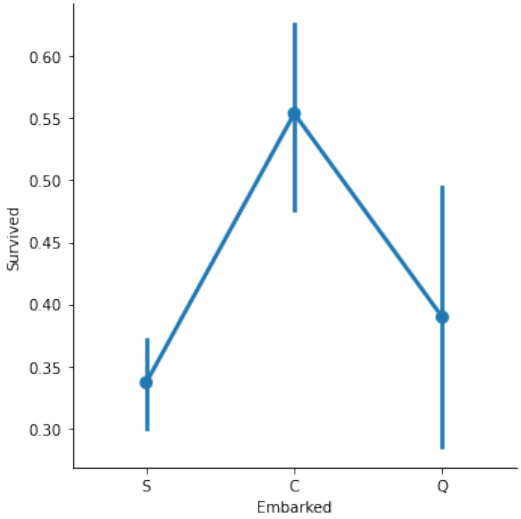
Количественный vs Количественный (corr)



Количественный vs Категориальный (violinplot)



Категориальный vs Категориальный (factorplot)



Дополнительные источники

- machinelearning.ru
- mlcourse.ai, dlcourse.ai
- kaggle