

# Решение задачи регрессии с помощью методов машинного обучения

Голиков М. О.

«ОРЛОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
ИМЕНИ И. С. ТУРГЕНЕВА»

19 декабря 2020

# Содержание

1 Описание исходных данных

2 Анализ данных

# Описание исходных данных

В данной работе используются **данные** о страховых выплатах медицинскому персоналу.

1	age	sex	bmi	children	smoker	region	charges
2	19	female	27.9	0	yes	southwest	16884.924
3	18	male	33.77	1	no	southeast	1725.5523
4	28	male	33	3	no	southeast	4449.462
5	33	male	22.705	0	no	northwest	21984.47061
6	32	male	28.88	0	no	northwest	3866.8552
7	31	female	25.74	0	no	southeast	3756.6216
8	46	female	33.44	1	no	southeast	8240.5896
9	37	female	27.74	3	no	northwest	7281.5056

Рисунок: Пример данных

# Описание исходных данных

## Описание признаков:

- age - возраст бенефициара (лица, получающего страховые выплаты);
- sex - пол бенефициара;
- bmi - индексы массы тела;
- children - количество детей бенефициара, на которых распространяется страхование / количество иждивенцов бенефициара;
- smoker - является ли бенефициар курильщиком;
- region - жилой район бенефициара в США: северо-восток, юго-восток, юго-запад, северо-запад;
- charges - индивидуальные медицинские расходы, выставленные на счет медицинского страхования.

# Анализ данных

## Гистограммы признаков sex и region



Рисунок: Гистограмма признака sex

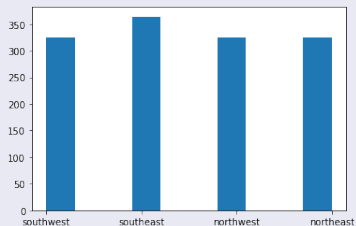


Рисунок: Гистограмма признака region

# Анализ данных

## Гистограммы признаков children и smoker

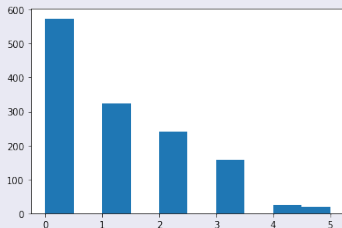


Рисунок: Гистограмма признака children

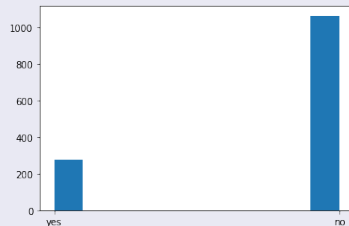


Рисунок: Гистограмма признака smoker

# Анализ данных

С помощью гистограмм удалось выяснить, что в категориальных признаках нет категорий, у которых была бы значительная количественная разница по отношению к другим значениям.

# Анализ данных

## Гистограммы признаков sex и region

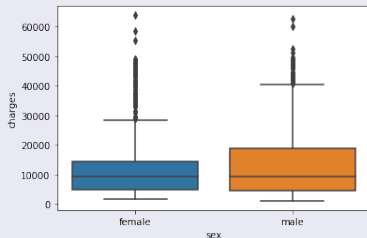


Рисунок: Гистограмма признака sex

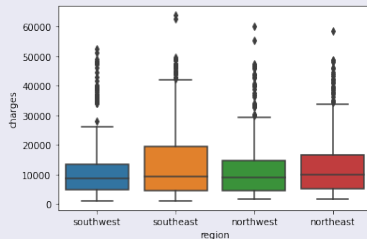


Рисунок: Гистограмма признака region



# Анализ данных

## Гистограммы признаков children и smoker

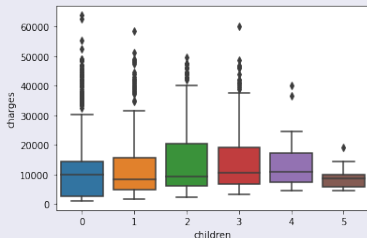


Рисунок: Гистограмма признака children

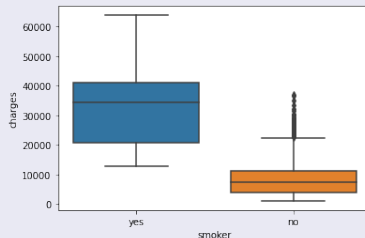
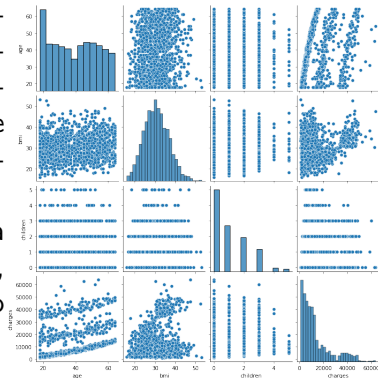


Рисунок: Гистограмма признака smoker

# Анализ данных

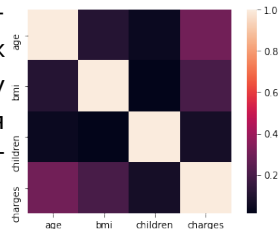
Справа можно увидеть диаграмму рассеяния и распределения признаков. Можно заметить, что между признаками не наблюдается линейной зависимости.

Стоит обратить внимание на распределение признака `charges`, которое сильно скошено в левую сторону.



# Анализ данных

На рисунке справа представлена визуализация матрицы корреляции. На основе этих данных можно сделать вывод, что между вещественными признаками не наблюдается линейной зависимости, а также отсутствует мультиколлинеарность.



Спасибо за внимание!