Отчет по анализу и моделированию страховых выплат

Данный отчет описывает процесс анализа страховых выплат и разработку модели для их прогнозирования.

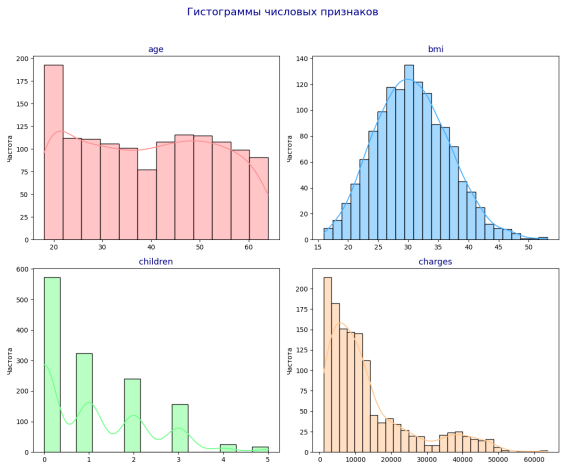
Модель основана на характеристиках клиентов: **возрасте, поле, индексе массы тела (BMI), количестве детей, статусе курения** и **регионе проживания.**

Отчет включает этапы:

- Подготовка
- Визуализация
- Предобработка
- Обучение моделей
- Интерпретация полученных результатов.

▼ Визуализация данных

Для анализа были построены гистограммы числовых признаков (возраст, BMI, количество детей, сумма выплат) и графики для категориальных признаков (пол, курение, регион).



- Возрастное распределение клиентов показало пик около 20 лет.
- BMI имел нормальное распределение с пиком около 30. Большинство клиентов бездетны или имеют одного ребенка.

▼ Анализ данных

В наборе данных выделены три категориальных признака (sex, smoker, region) и три числовых признака (age, bmi, children). Целевая переменная charges имеет тип данных float64.

При проверке данных был обнаружен и удален один дубликат для обеспечения чистоты выборки. Пропущенных значений в данных не обнаружено.

Дескриптивная статистика числовых признаков

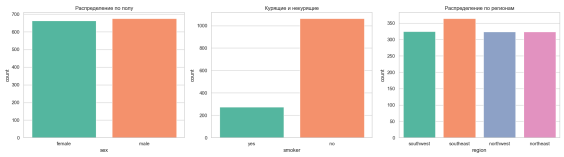
- В выборке 1337 человек.

Возраст (age):

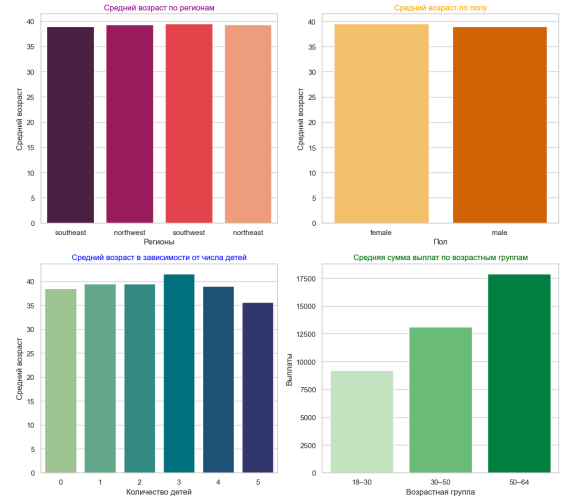
- **Средний возраст** — 39 лет, данные разбросаны ±14 лет от среднего.
- **Минимальный возраст** — 18 лет, **максимальный возраст** — 64 года.
- **Распределение симметричное** — равное количество молодых, зрелых и пожилых людей.

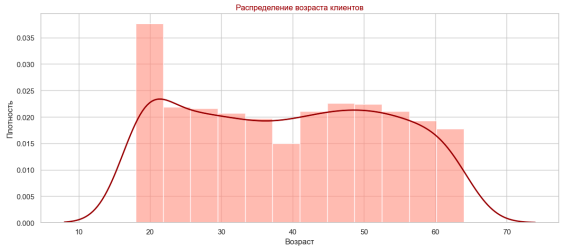
Индекс массы тела (BMI):

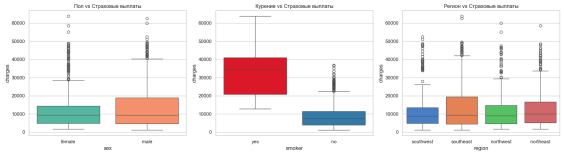
Распределение по полу сбалансировано, а некурящих клиентов значительно больше, чем курящих.



- Выплаты у курящих выше.
- С возрастом выплаты по страховке растут.







- **Средний BMI** близок к границе ожирения (≥30).
- Диапазон BMI от **сильного дефицита массы** до **тяжелого ожирения**: [15.96 - 53.13].
- **50% людей** имеют BMI в пределах **нормы/избыточного веса**.

Дети (children):

- **В среднем** — наличие **1 ребенка**.
- Есть клиенты без детей (0) и с максимальным количеством (5).
- Половина клиентов имеют не более **1 ребенка**, а **25% не имеют детей**.

Сумма страховых выплат (charges):

- **Средняя выплата** по страхованию — **13 тысяч**.
- **Большой разброс** данных.
- **Медиана** — 9386, и только **25% клиентов** платят больше **16657**.
- **Минимум** — 1121, **максимум** — 63770.

▼ Предобработка данных

Для подготовки данных к моделированию были выполнены следующие шаги:

1. Применено One-Hot Encoding для категориальных признаков

2. Проведено логарифмирование целевой переменной charges для уменьшения правостороннего смещения
3. Выполнено масштабирование числовых признаков с помощью StandardScaler

▼ Обучение моделей

Для прогнозирования страховых выплат были применены модели:

- **Линейная регрессия (Linear Regression)**
- **Градиентный бустинг (Gradient Boosting)**
- **Случайный лес (Random Forest)**
- **Полиномиальная регрессия (Polynomial Regression)**

Каждая модель обучалась на тренировочной выборке и оценивалась на тестовой с использованием метрик R2, MSE, RMSE и MAE.

▼ Результаты моделей

Модель	R2	MSE	RMSE	MAE
Linear Regression	0.8291	0.16	0.40	0.27
Gradient Boosting	0.8830	0.11	0.33	0.19
Random Forest	0.8566	0.13	0.36	0.19
Polynomial Regression	0.8796	0.11	0.33	0.20

Лучшие результаты по всем метрикам показали **Gradient Boosting** и **Polynomial Regression**. Обе модели демонстрируют близкие показатели, но **градиентный бустинг** немного превосходит по метрикам **R2** и **MAE**.

Лучшие модели:

- **Gradient Boosting** и **Polynomial Regression** продемонстрировали наилучшие результаты по всем метрикам.
- **Gradient Boosting** показал небольшое преимущество по метрикам **R2** и **MAE**, что делает его оптимальным выбором для данной задачи.