ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

ВЫСШАЯ ИНЖЕНЕРНАЯ ШКОЛА

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ПРОГНОЗИРОВАНИЕ ЗАРАБОТНОЙ ПЛАТЫ НА ОСНОВЕ ДАННЫХ О ВАКАНСИЯХ С САЙТА HH.RU И АНАЛИЗА ФАКТОРОВ, ВЛИЯЮЩИХ НА ЕЕ УРОВЕНЬ

по программе профессиональной переподготовки:
«Анализ данных на языке Python»

Выполнила: Смурыгина Марина Вячеславовна
Руководитель: к.т.н. Семендяев Родион Юрьевич

Санкт-Петербург, 2026

АКТУАЛЬНОСТЬ

Современный рынок труда быстро изменяется, поэтому требуется оперативный прогноз заработной платы. Данные онлайн-платформы hh.ru позволяют учитывать профессиональные, региональные и организационные характеристики вакансий. Дополнение этих данных макроэкономическими показателями - уровнем безработицы, численностью населения и особенностями федерального округа - повышает точность прогнозов. Применение методов машинного обучения позволяет выявлять устойчивые закономерности рынка труда. Поэтому прогнозирование заработной платы на основе объединённых микро- и макроданных является актуальной задачей для участников рынка труда.

ИСХОДНЫЕ ДАННЫЕ

Источник:

- вакансии с hh.ru
- официальная статистика с rosstat.gov.ru

Объём данных:

- Более 600 000 строк
- 25 признаков

Макроданные:

- уровень безработицы
- численность населения
- данные по Федеральным округам

Микроданные:

- зарплата
- профессиональная категория
- опыт работы
- график работы
- тип занятости
- регион работодателя

Дополнительно:

- расчёт целевой переменной *зарплата*
- группировка профессий
- количество вакансий у работодателя

ЦЕЛИ И ЗАДАЧИ

Объект исследования:

рынок труда России, представленный вакансиями с указанной заработной платой.

Предмет исследования:

факторы, определяющие уровень заработной платы: характеристики вакансии, требования к кандидатам, особенности работодателя, а также социально-экономические параметры региона.

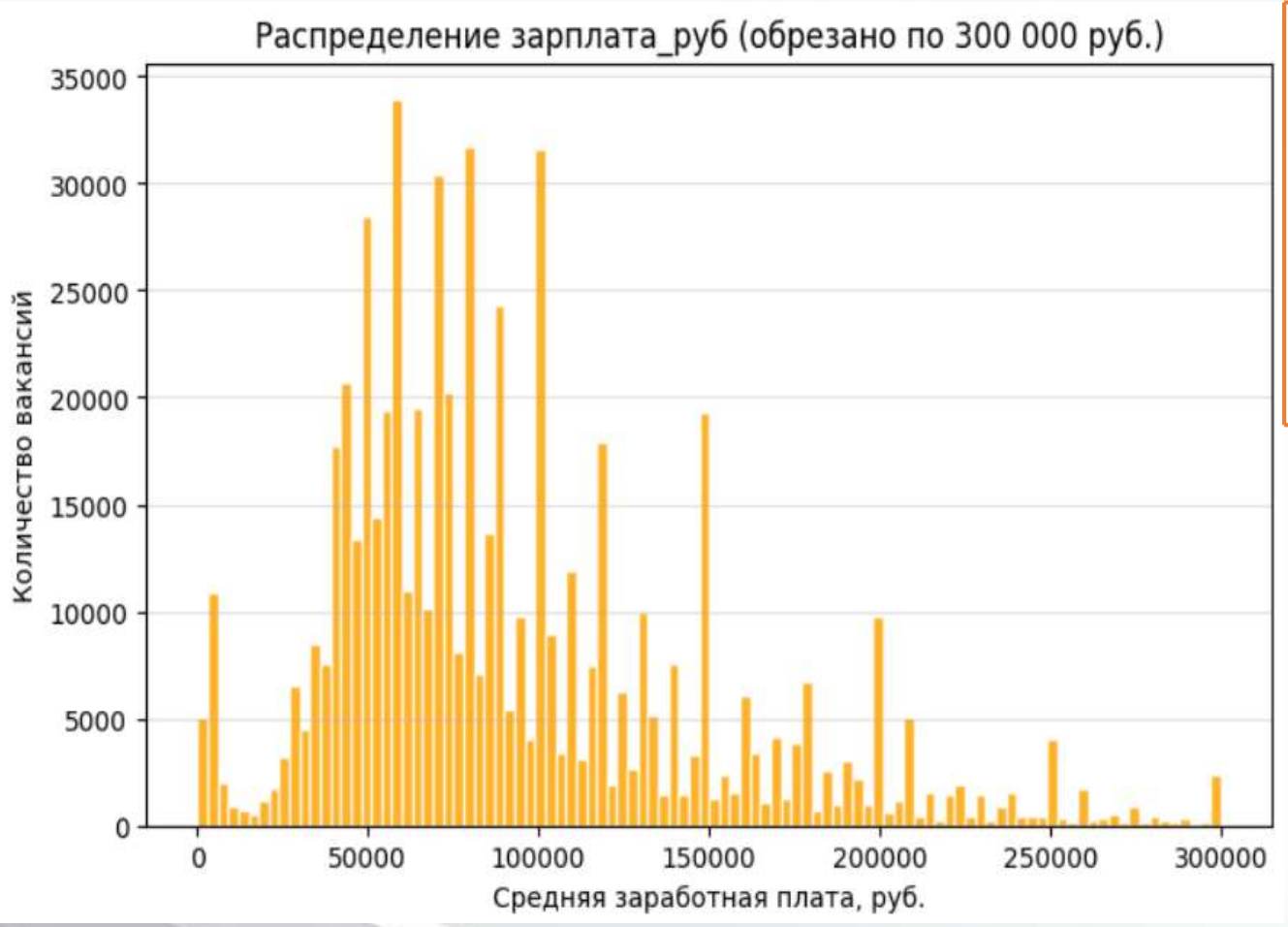
Цель исследования:

построение и сравнительный анализ моделей прогнозирования заработной платы на основе данных о вакансиях hh.ru с учётом влияющих на неё факторов.

Задачи:

1. Выполнить сбор данных с сайта hh.ru и rosstat.gov.ru ;
2. Провести очистку и подготовку данных;
3. Обогатить данные социально-экономическими характеристиками регионов;
4. Провести разведочный анализ данных;
5. Оценить значимость факторов и интерпретировать вклад различных характеристик вакансии и региона в формирование уровня заработной платы;
6. Построить и сравнить несколько моделей прогнозирования заработной платы, используя методы машинного обучения;
7. Сформулировать выводы и рекомендации на основе полученных результатов.

РАСПРЕДЕЛЕНИЕ ЗАРПЛАТЫ



ключевые показатели:

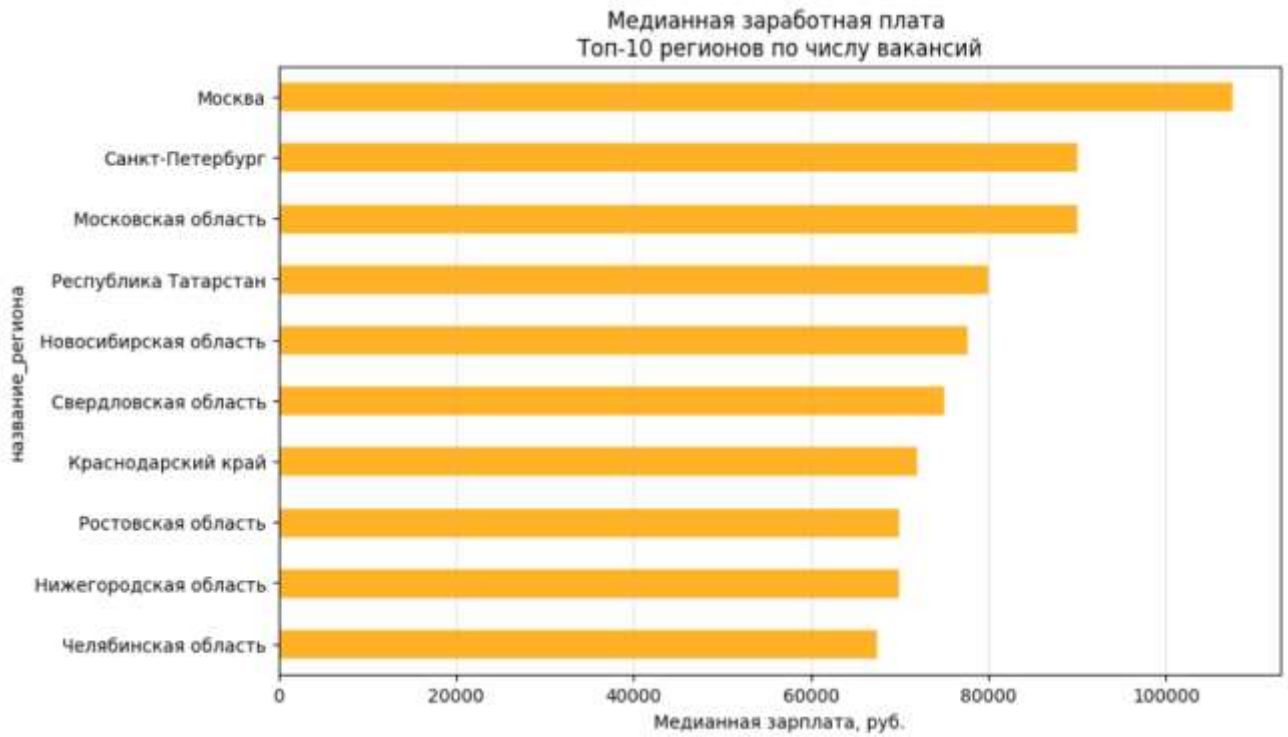
- минимальное значение: 1 руб.;
- медиана: 80 000 руб.;
- максимальное значение: 56 305 450 руб.;
- средняя зарплата: ~98 000 руб.

Основная масса наблюдений лежит в диапазоне от 30 тыс. до 150 тыс. руб., с несколькими локальными пиками, обусловленными типичными вилками зарплат на рынке труда. Значения выше 300 тыс. руб. встречаются реже (менее 5 % выборки) и образуют «длинный хвост», типичный для зарплатных данных.

РАСПРЕДЕЛЕНИЕ ЗАРПЛАТЫ ПО РЕГИОНАМ

ТОП-3 региона по медианной заработной плате:

- Чукотский АО - 145 000 рублей
- Магаданская область - 125 000 рублей
- Ямало-Ненецкий АО - 112 500 рублей

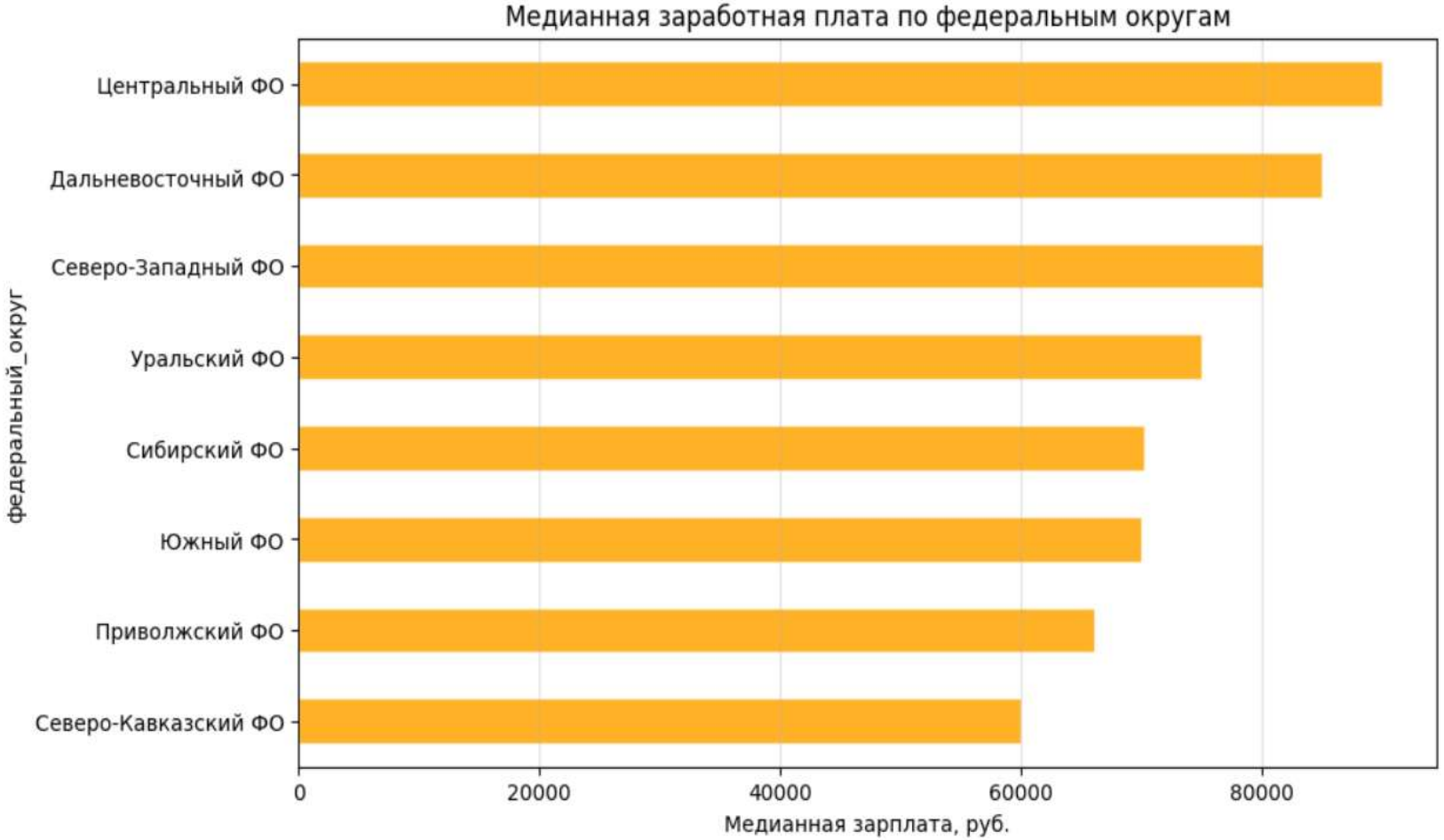


ТОП-10 регионов по количеству вакансий:

	Регион	Количество вакансий
1	Москва	101 861
2	Московская область	52 323
3	Санкт-Петербург	47 956
4	Краснодарский край	24 819
5	Свердловская область	24 512
6	Республика Татарстан	18 639
7	Новосибирская область	15 077
8	Челябинская область	14 334
9	Ростовская область	14 244
10	Нижегородская область	14 084

- Наибольшее количество вакансий сосредоточено в крупнейших экономических центрах России, из них Москва и Санкт-Петербург - это города с самой большой зарплатой.

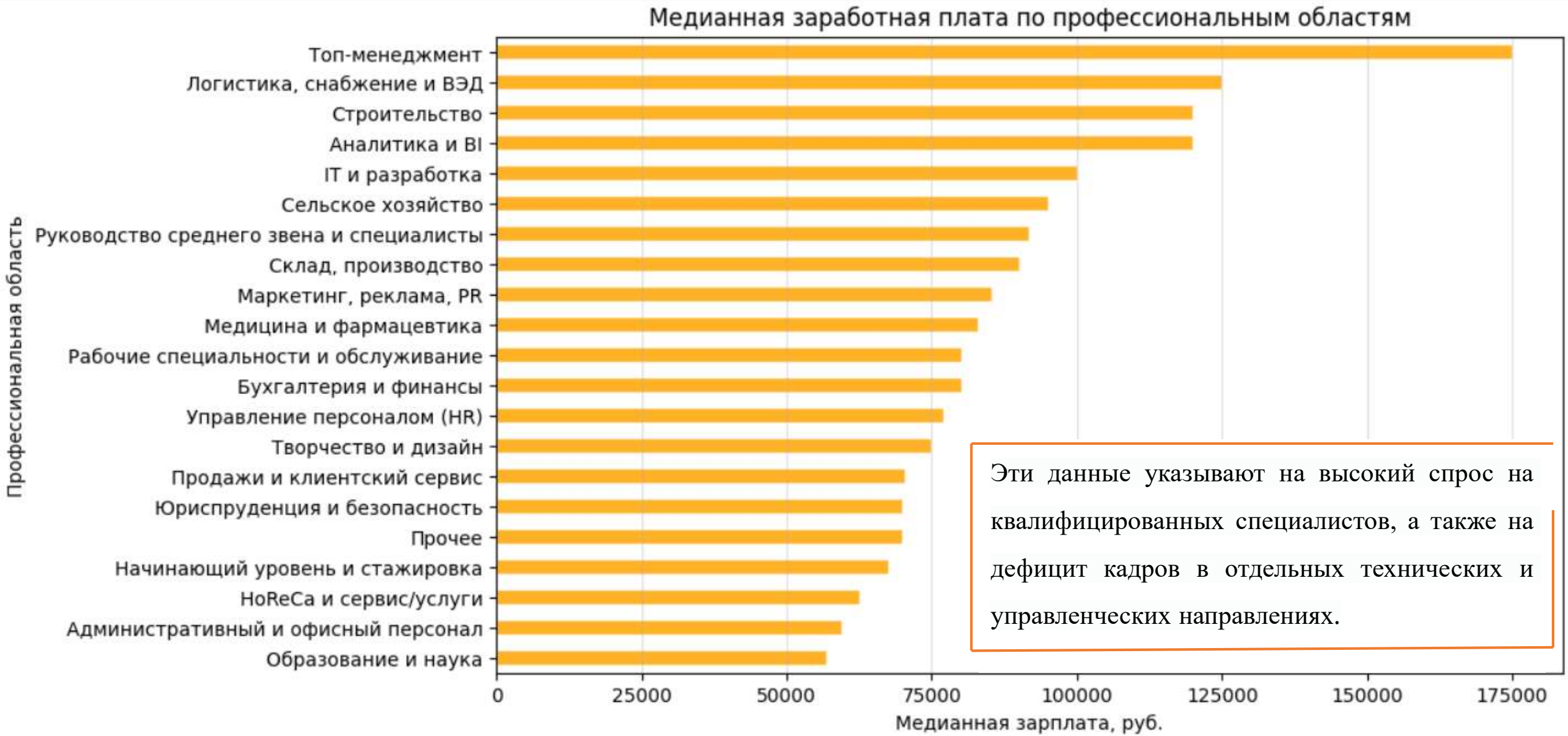
РАСПРЕДЕЛЕНИЕ ЗАРПЛАТЫ ПО ФЕДЕРАЛЬНЫМ ОКРУГАМ



Самыми высокими медианными зарплатами обладают регионы Центра и Дальнего Востока, что связано с высокой концентрацией высокооплачиваемых вакансий в Москве, а также с действием региональных коэффициентов и специфической отраслевой структурой Дальнего Востока. Наименьшие значения характерны для Приволжского и Северо-Кавказского округов.

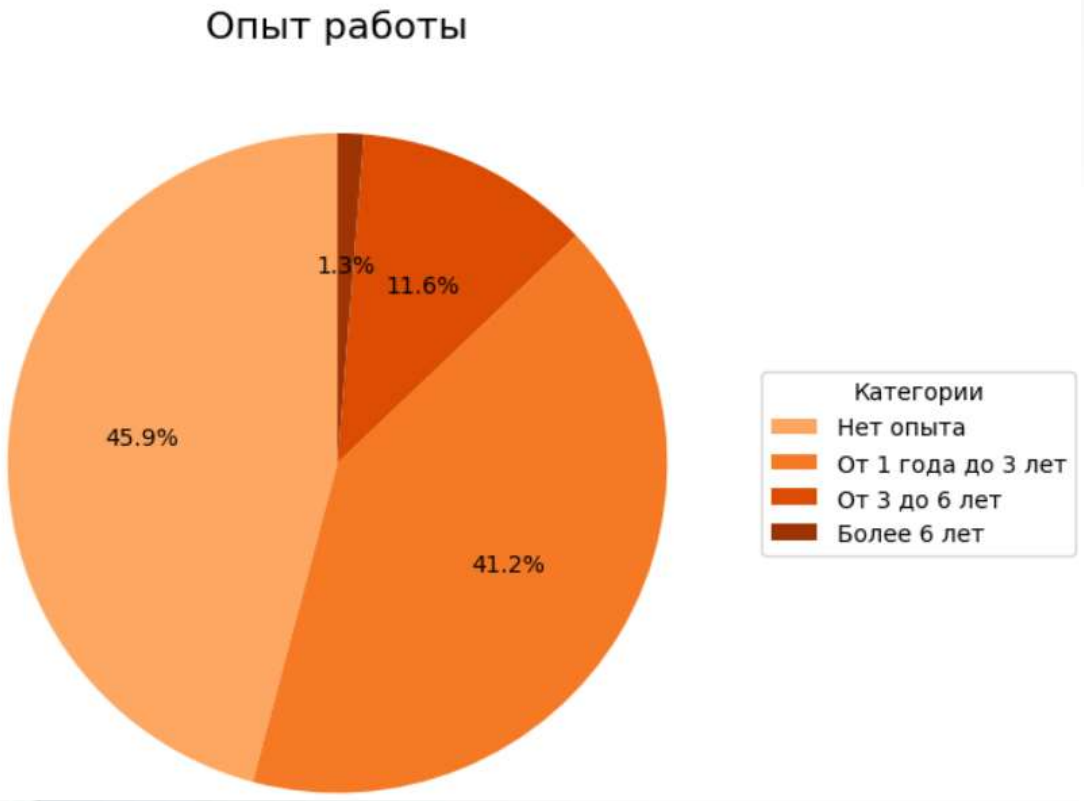
- Данное распределение соответствует особенностям социально-экономического развития округов.

РАСПРЕДЕЛЕНИЕ ЗАРПЛАТЫ ПО ПРОФЕССИОНАЛЬНЫМ ОБЛАСТЯМ

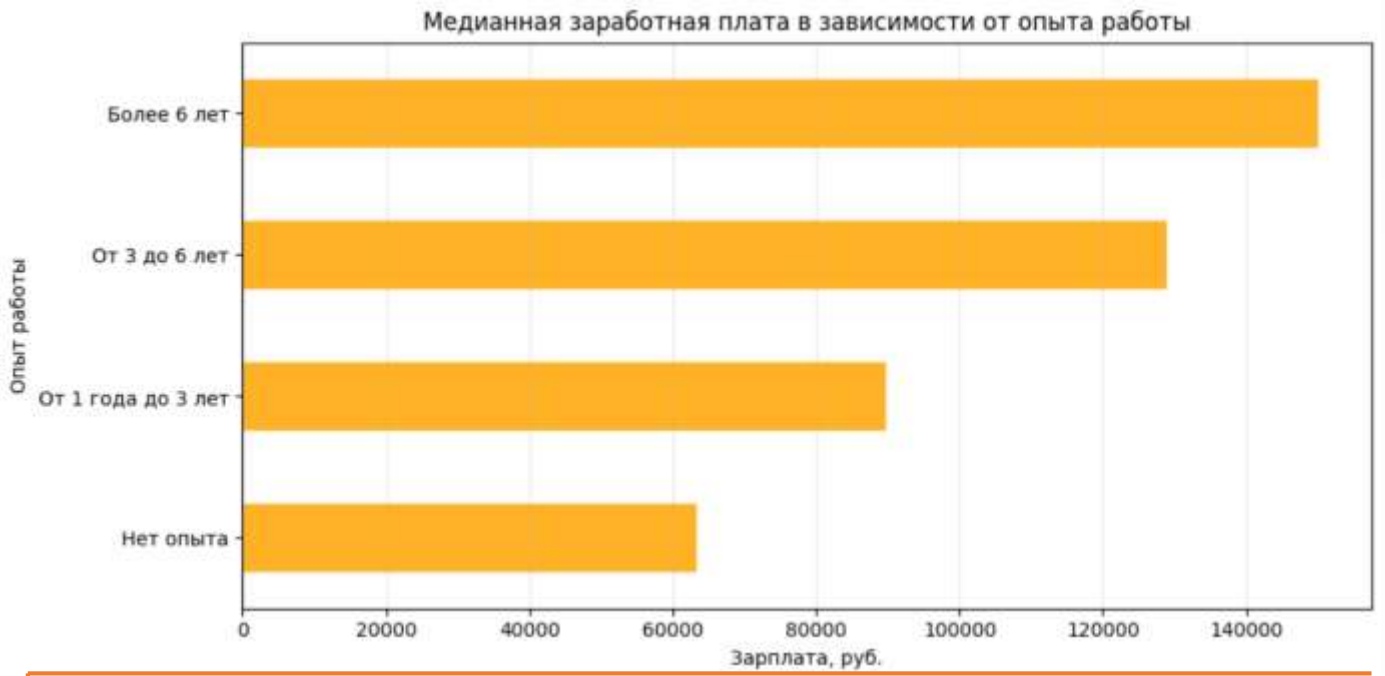


Эти данные указывают на высокий спрос на квалифицированных специалистов, а также на дефицит кадров в отдельных технических и управленческих направлениях.

РАСПРЕДЕЛЕНИЕ ВАКАНСИЙ ПО ОПЫТУ РАБОТЫ



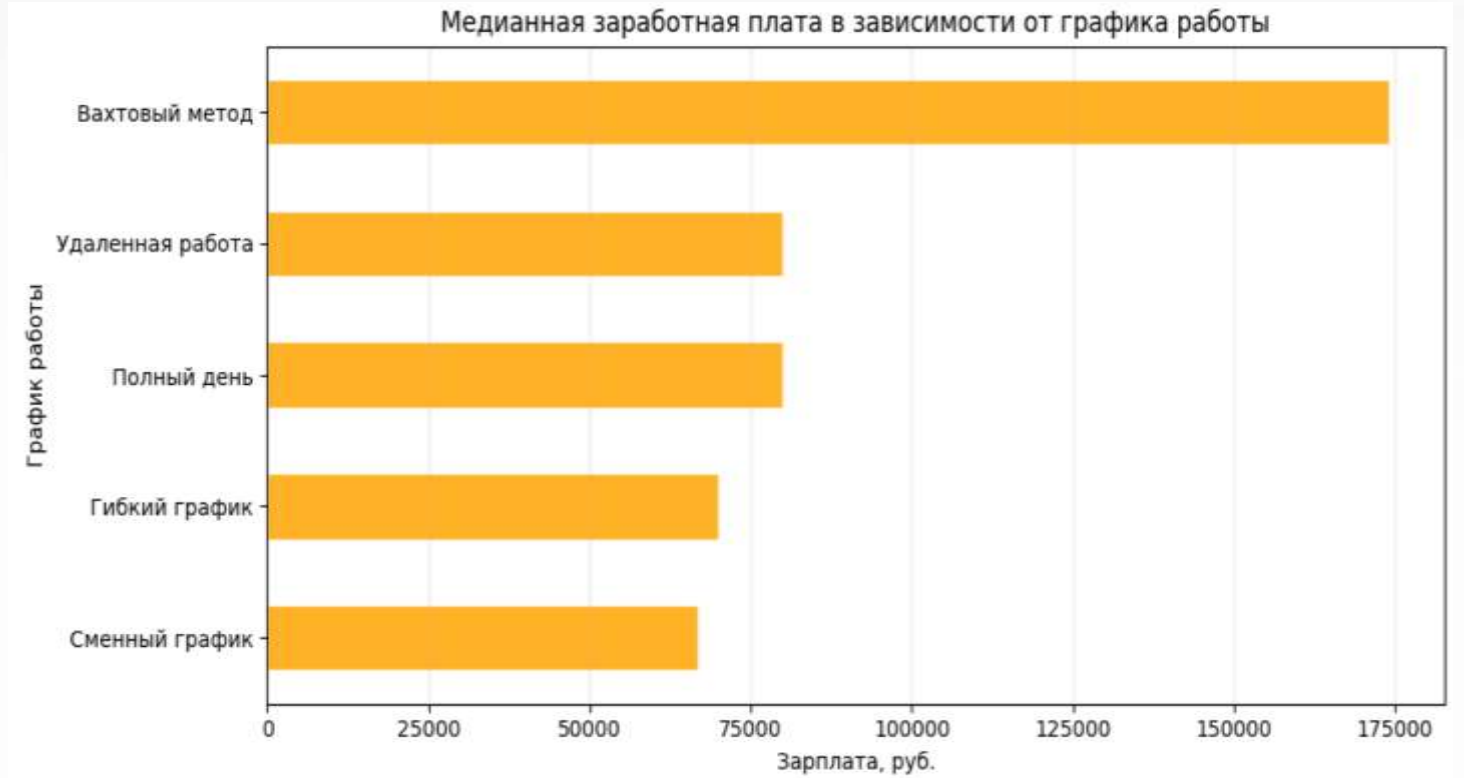
Опыт работы является одним из ключевых факторов, определяющих уровень оплаты труда.



Доля вакансий, требующих более высокой квалификации сравнительно невелика. Эти категории характерны для руководящих, финансовых, юридических, аналитических и технически сложных позиций, предложение которых ограничено.

Разрыв между медианой начальных позиций («нет опыта») и вакансиями для высококвалифицированных специалистов («более 6 лет») превышает в 2,3 раза, что подчёркивает ценность профессионального стажа для работодателей.

РАСПРЕДЕЛЕНИЕ ВАКАНСИЙ ПО ГРАФИКУ РАБОТЫ

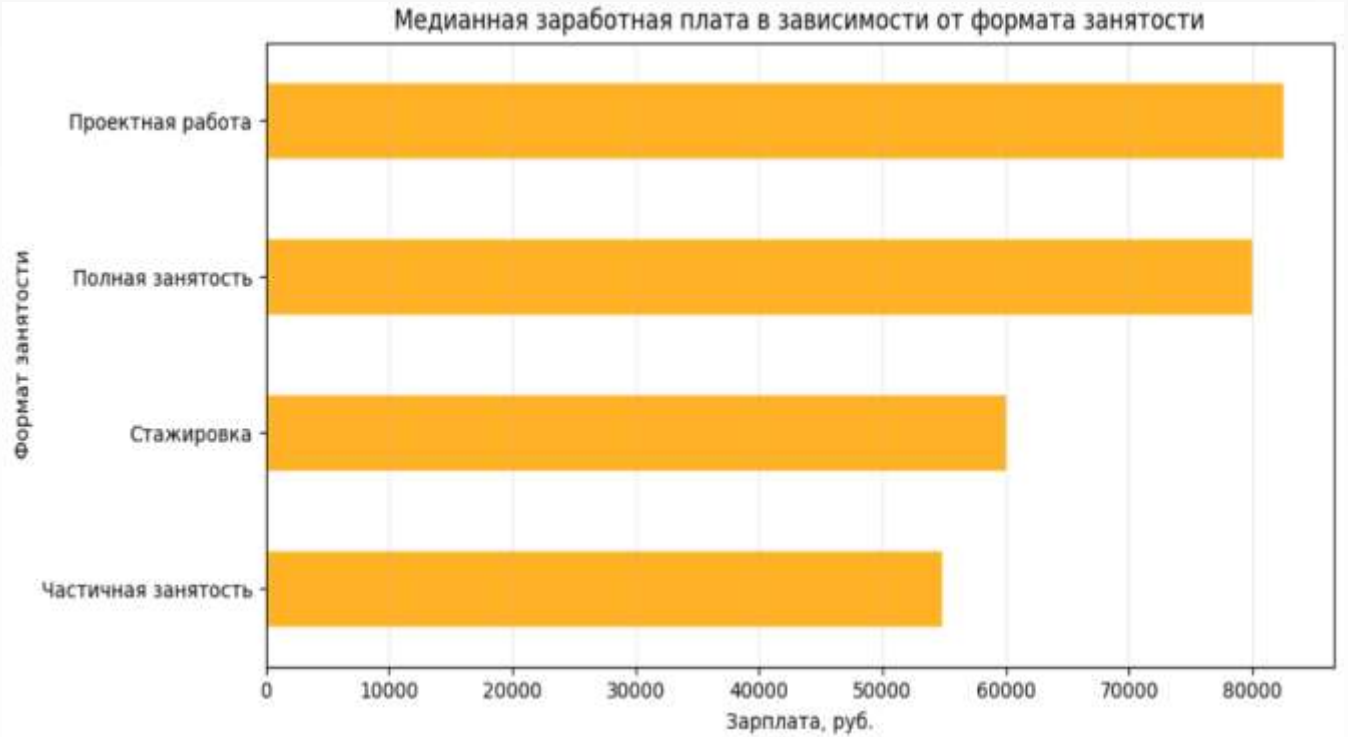


Удалённая работа занимает лишь 4,8%.

Больше всего удаленная работа преобладает в маркетинге, IT сфере и у административного персонала

Вакансии с вахтовым методом показали медианную заработную плату около 175 тыс. руб., что является максимальным значением среди всех видов графиков. Помимо того, что это связано с тяжелыми условиями труда вдали от дома зарплата может быть указана за период вахты, который может быть равен неделям, месяцам, кварталам и т.п.

РАСПРЕДЕЛЕНИЕ ВАКАНСИЙ ПО ФОРМЕ ЗАНЯТОСТИ

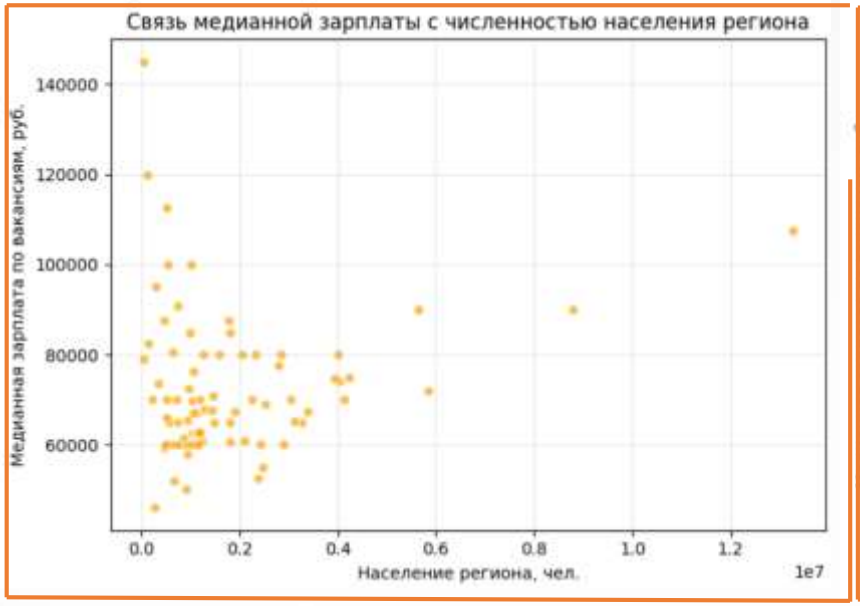


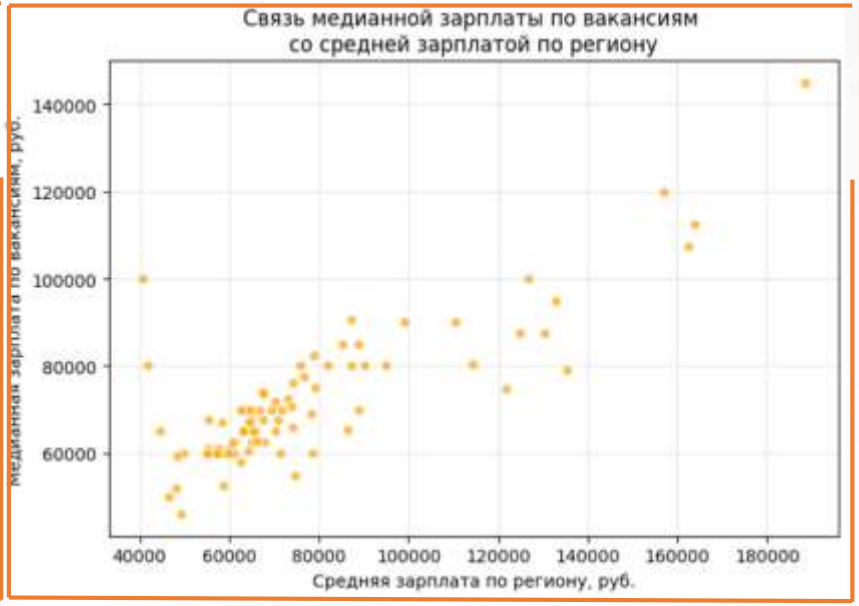
Проектная работа, временная занятость, стажировка -представлены долями менее 1%

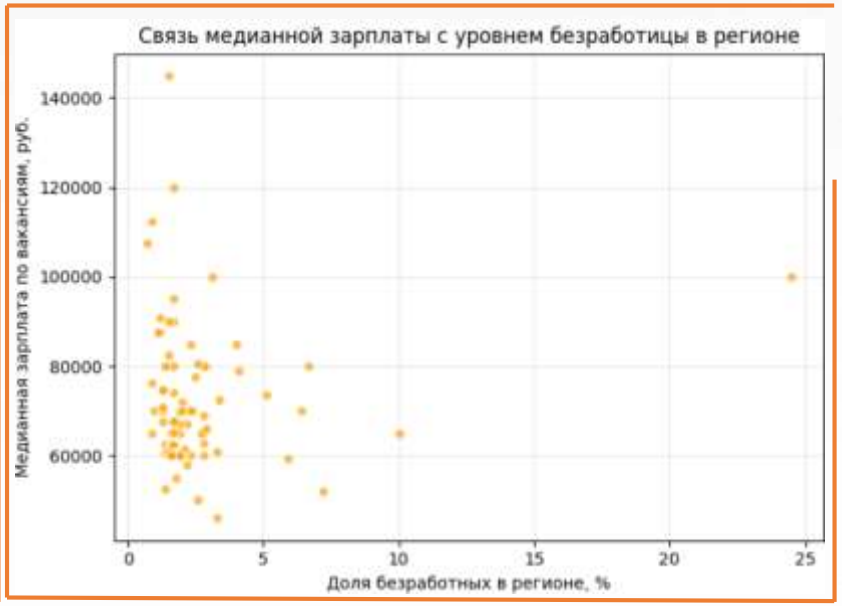
Проектная и полная занятость — наиболее высокооплачиваемые форматы, что объясняется требованиями к квалификации и вовлечённостью.

Частичная занятость и стажировки — существенно менее оплачиваемые, отражая ориентацию на массовые, вспомогательные или учебные форматы работы.

ЗАВИСИМОСТЬ ЗАРПЛАТЫ ОТ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРИЗНАКОВ





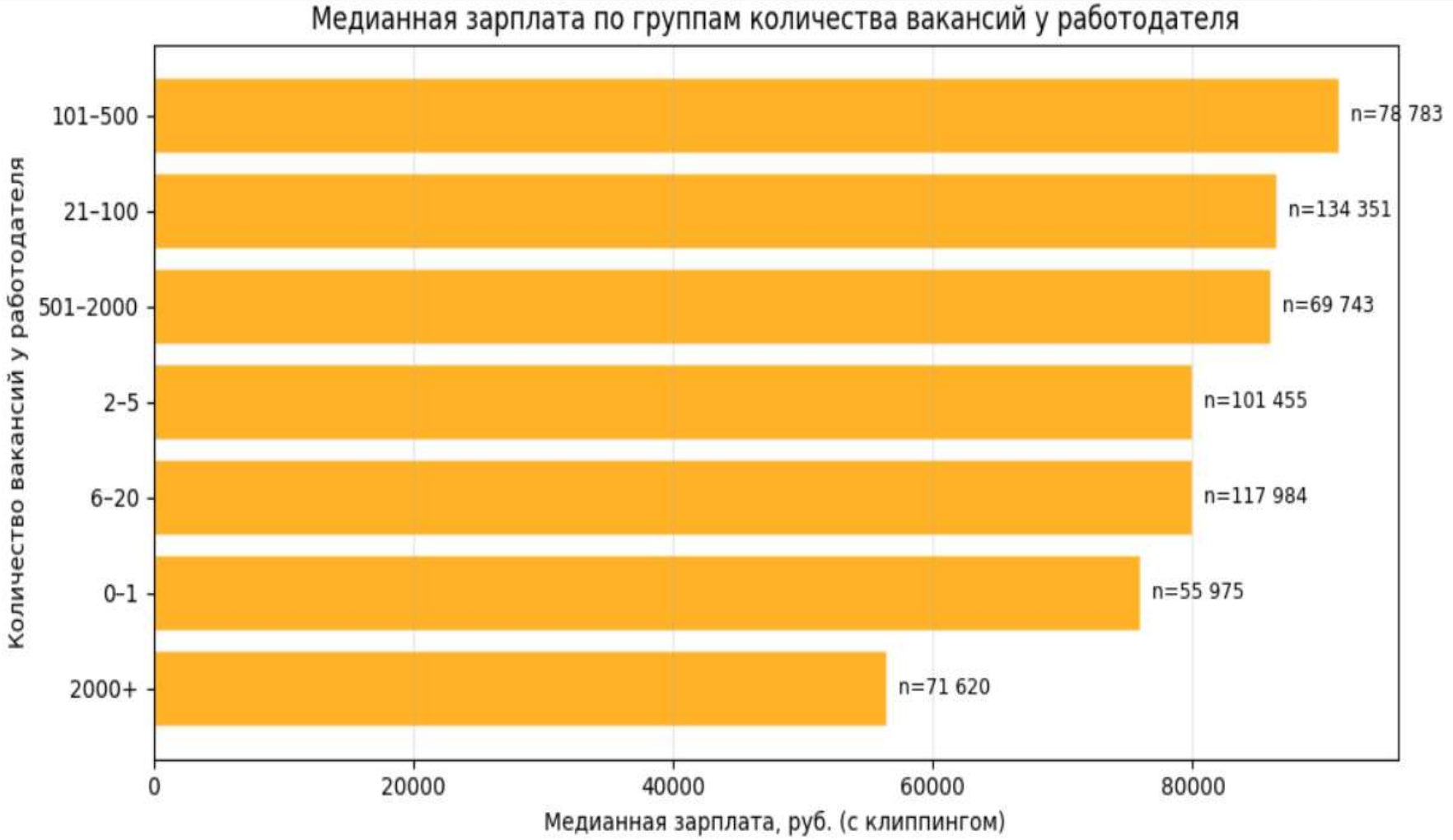


- Корреляция Пирсона = 0.181
- Слабая положительную связь между рассматриваемыми показателями
- Численность населения региона не является определяющим фактором уровня заработной платы.

- Корреляция Пирсона = 0.831
- Сильная положительная связь между показателями
- Средняя зарплата региона является важнейшим социально-экономическим фактором, определяющим предложение заработной платы в вакансиях.

- Корреляция Пирсона = 0.081
- Очень слабая положительную связь между рассматриваемыми показателями
- Уровень безработицы региона не является определяющим фактором уровня заработной платы.

РАСПРЕДЕЛЕНИЕ ЗАРПЛАТЫ ПО КОЛИЧЕСТВУ ВАКАНСИЙ У РАБОТОДАТЕЛЯ



- Корреляция Спирмена = –0,025
- Очень слабая отрицательная связь между рассматриваемыми показателями
- Количество вакансий не является прямым линейным фактором роста зарплаты.

КОРРЕЛЯЦИЯ ЧИСЛОВЫХ ПРИЗНАКОВ

Корреляция Пирсона

- Сильная мультиколлинеарность:

  Население региона и средняя зарплата = 0.86

  Доля безработных и средняя зарплата = -0.54

  Доля безработных и население региона = -0.6

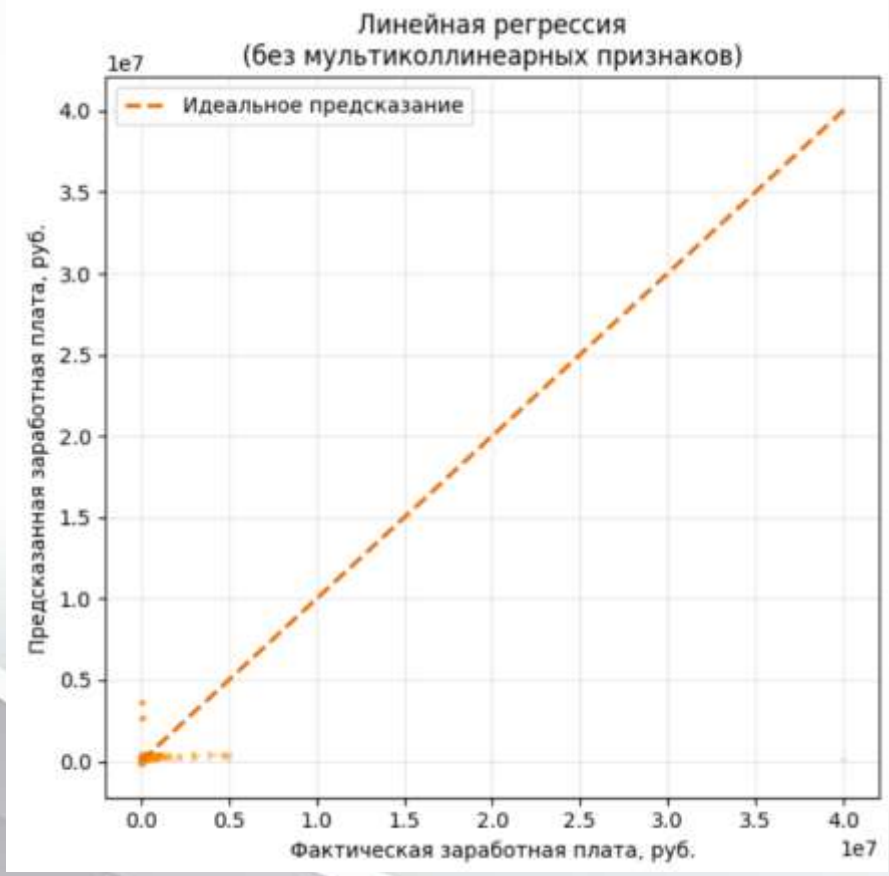
- Числовые признаки почти не объясняют зарплату напрямую.

- Удалим:

  население региона и долю безработных округа

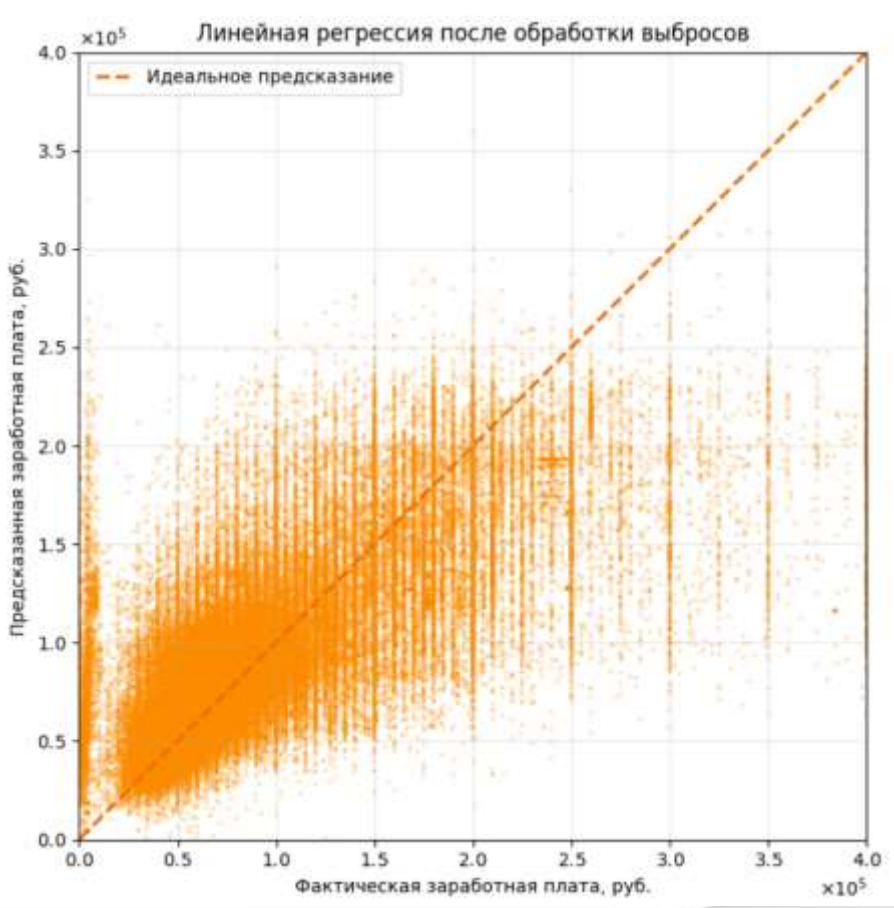
ЛИНЕЙНАЯ РЕГРЕССИЯ

Линейная регрессия до обработки выбросов



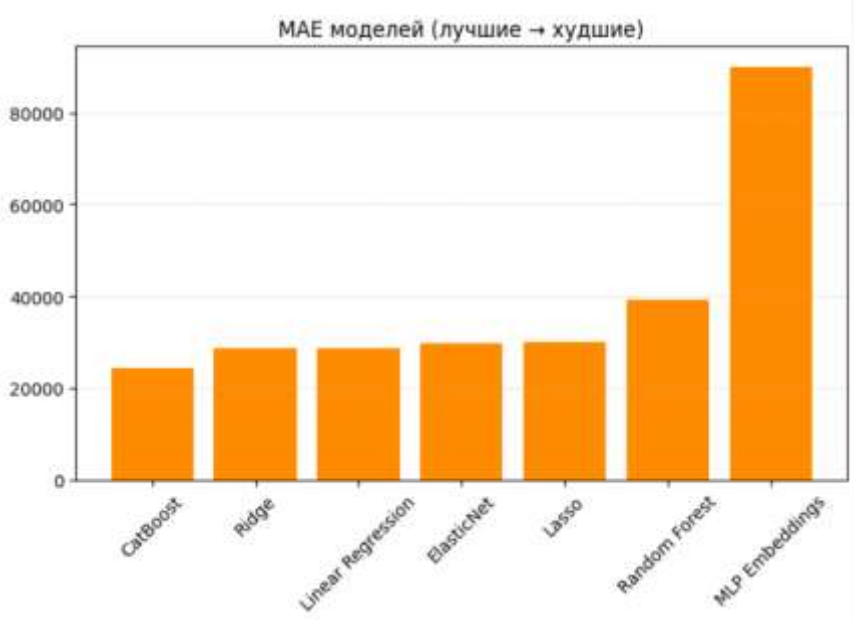
Целесообразно применение винсоризации таргета по 99.5-му процентилю для дальнейшего сравнения и обучения более сложных моделей машинного обучения

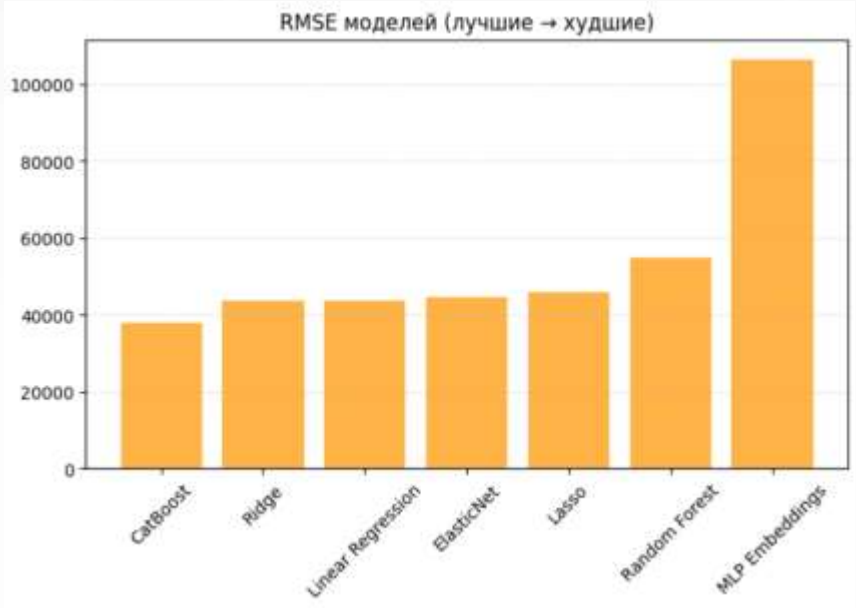
Линейная регрессия после обработки выбросов

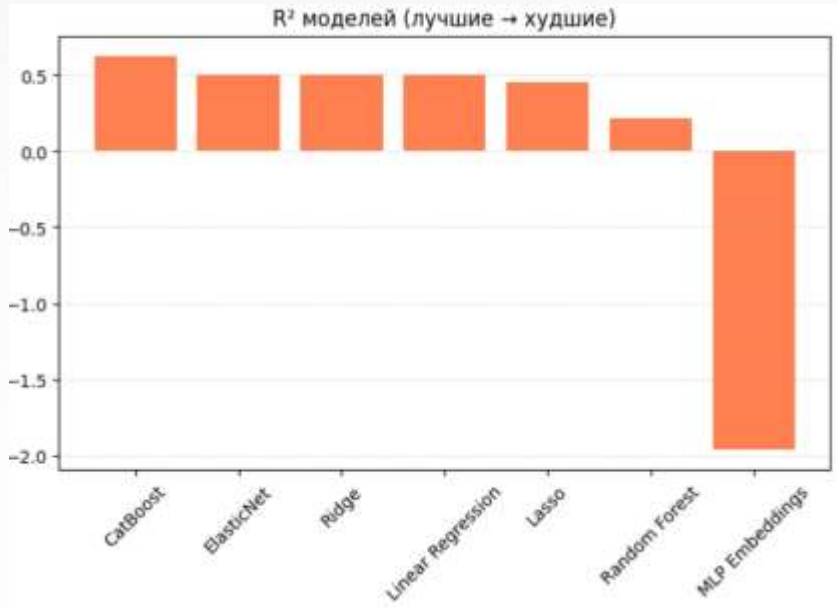


MAE: 31,657.80 руб.	**-10%**	MAE: 28,584.32 руб.
RMSE: 139,356.47 руб.	**-69%**	RMSE: 43,716.07 руб.
R^2 : 0.0450	**+1007%**	R^2: 0.4980

СРАВНЕНИЕ МОДЕЛЕЙ







CatBoost — лучшая модель

CatBoost обеспечил минимальные значения ошибок MAE и RMSE и максимальное значение коэффициента детерминации R^2:

MAE: 24 280 руб.

RMSE: 37 727 руб.

R^2: 0,626

Таким образом, данная модель оказалась наиболее точной и устойчивой среди всех рассмотренных.

ВЫВОДЫ

- CatBoost — лучшая модель
- Медианная зарплата = 80 000 рублей
- MAE = 24 280 рублей

30%

В эконометрике и статистике доходов отмечается, что заработная плата характеризуется высокой вариативностью и асимметричным распределением, что ограничивает точность прогнозных моделей. В связи с этим качество моделей рекомендуется оценивать также в относительных величинах, а не только в абсолютных значениях. Полученное значение MAE, составляющее 30% от медианной заработной платы, соответствует допустимому уровню ошибки для социально-экономических показателей.

Цель выпускной квалификационной работы была достигнута: выполнено построение и сравнительный анализ моделей прогнозирования заработной платы на основе данных о вакансиях, размещённых на платформе hh.ru, с учётом факторов, оказывающих влияние на уровень оплаты труда. В результате исследования был обоснован выбор наиболее эффективной модели машинного обучения для решения поставленной задачи. Полученные результаты могут быть использованы в аналитических и прикладных задачах мониторинга и оценки рынка труда, а также служить основой для дальнейших исследований в области прогнозирования социально-экономических показателей.

ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

ВЫСШАЯ ИНЖЕНЕРНАЯ ШКОЛА

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ПРОГНОЗИРОВАНИЕ ЗАРАБОТНОЙ ПЛАТЫ НА ОСНОВЕ ДАННЫХ О ВАКАНСИЯХ С САЙТА HH.RU И АНАЛИЗА ФАКТОРОВ, ВЛИЯЮЩИХ НА ЕЕ УРОВЕНЬ

по программе профессиональной переподготовки:
«Анализ данных на языке Python»

Выполнила: Смурыгина Марина Вячеславовна
Руководитель: к.т.н. Семендяев Родион Юрьевич

Санкт-Петербург, 2026