Можно ли знать себе цену?

Команда ANOD и hh.ru

Описание датасета

Взаимодействий	**500 000**
Резюме	**190 000**
Вакансий	**320 000**

- Датасет — набор резюме и вакансий, размещенных на hh.ru, а также их характеристики (дата создания, город, профессия, уровень зарплаты, тип занятости и т.д.)
- Временной срез 3 месяца (июнь, июль, август 2023 года)

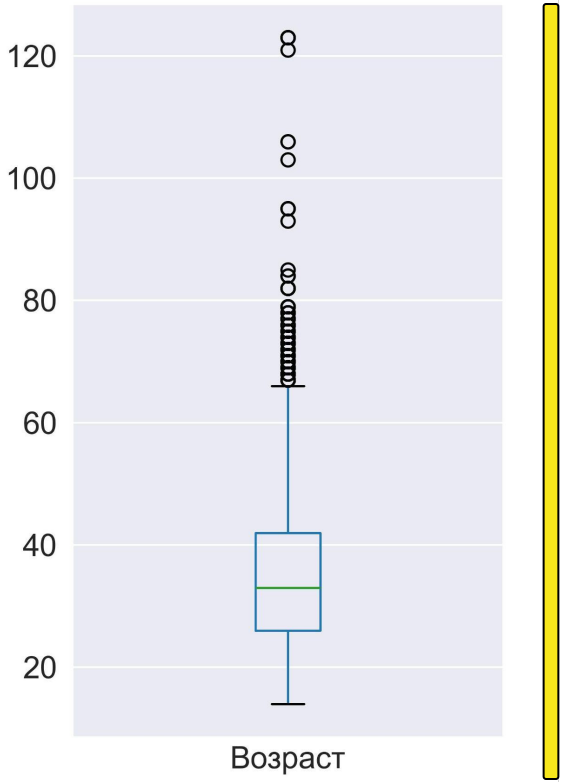
Наша мотивация

План ~~исследования~~ на жизнь:

1. Закончить школу
2. Выпустится из универа
3. ???
4. эммммм??
5. hh.ru
6. Любимая работа, где я получаю много деняк

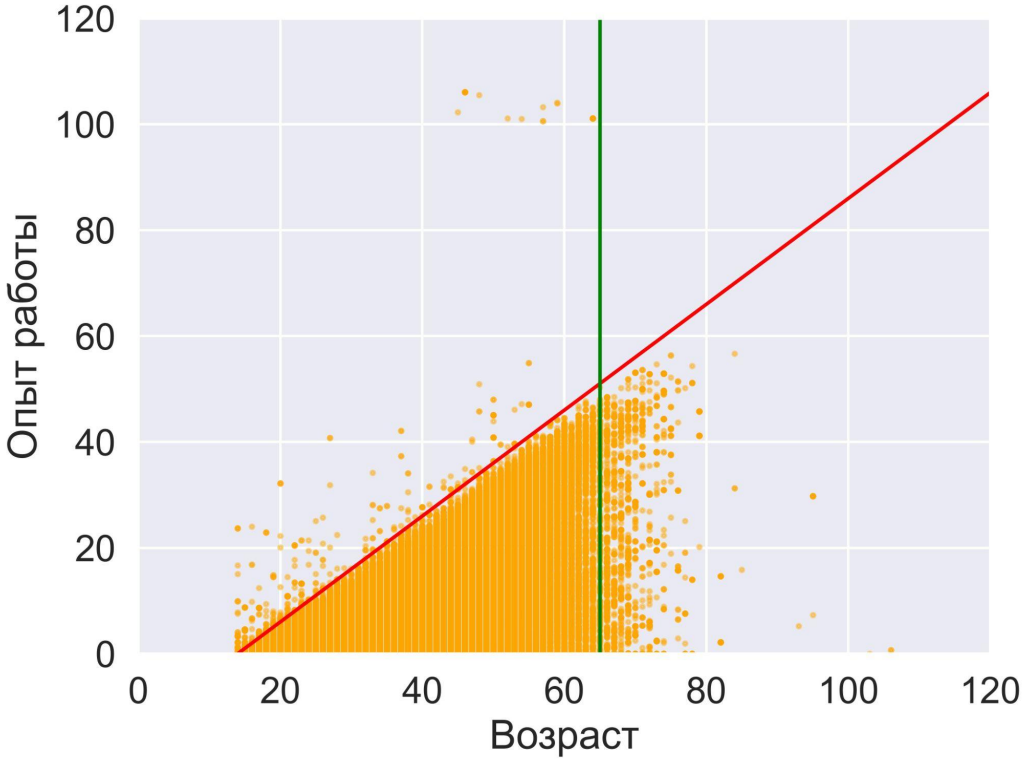
- Исследование будет полезно большому кругу лиц. В том числе, нам самим
- Платформа известна по всей России
- Данные помогают понять нынешнее положение рынка труда

Предварительный анализ → Распределение возрастов



Предварительный анализ

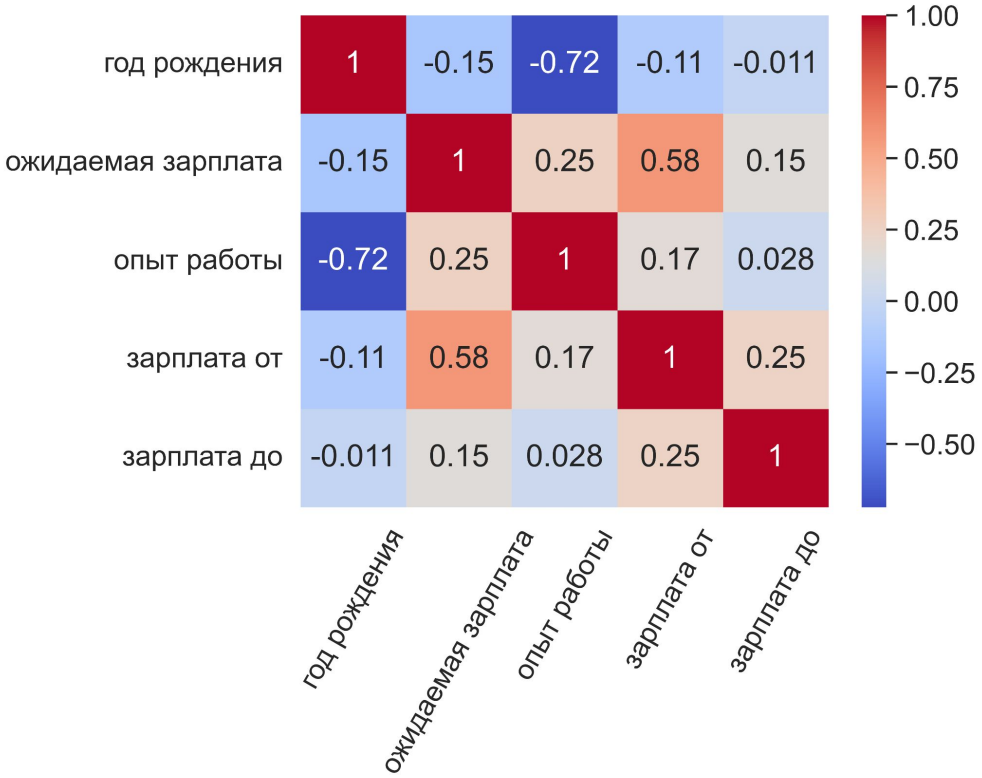
Возраст и опыт работы



Разница в возрасте работы и опыте 14 лет

Возраст в 65 лет

Предварительный анализ → Матрица корреляции



Предварительный анализ → **Очистка выбросов**

500 000 строк

200 000 строк

Были удалены взаимодействия:

- Минимальная зарплата ниже МРОТ (нулевая заменена на None)
- Соискатели старше 65 и младше 18 лет
- Возраст и количество опыта работы соискателя отличаются менее, чем на 14 лет
- Нулевые зарплатные ожидания кандидата
- Нет данных о верхней границе предлагаемой зарплаты
- Нет значения в статусах взаимодействий
- Соискатели, представляющие маленькие подвыборки (доктора наук, волонтёры, представители специфических регионов и т.д.)
- Очистка зарплатных ожиданий и вилки проводилась по boxplot отдельно для каждой профессии

Исследовательский вопрос

Какие характеристики кандидата с зарплатными ожиданиями выше вилки могут сподвигнуть рекрутера пригласить названного соискателя на собеседование?

Предварительный анализ → **Коэффициент Ивана Терентьева**

1. Разбиваем ген. совокупность по профессиям и характеристикам
2. Изучаем их влияние на приглашение кандидатов с зарплатными ожиданиями выше зарплатной вилки относительно приема при других зп. ожиданиях

На основе критерия считаем разницу между максимальным и минимальным значением по характеристике в различных профессиях, которая отображает значимость характеристики в этом случае

Предварительный анализ

Коэффициент Ивана Терентьева

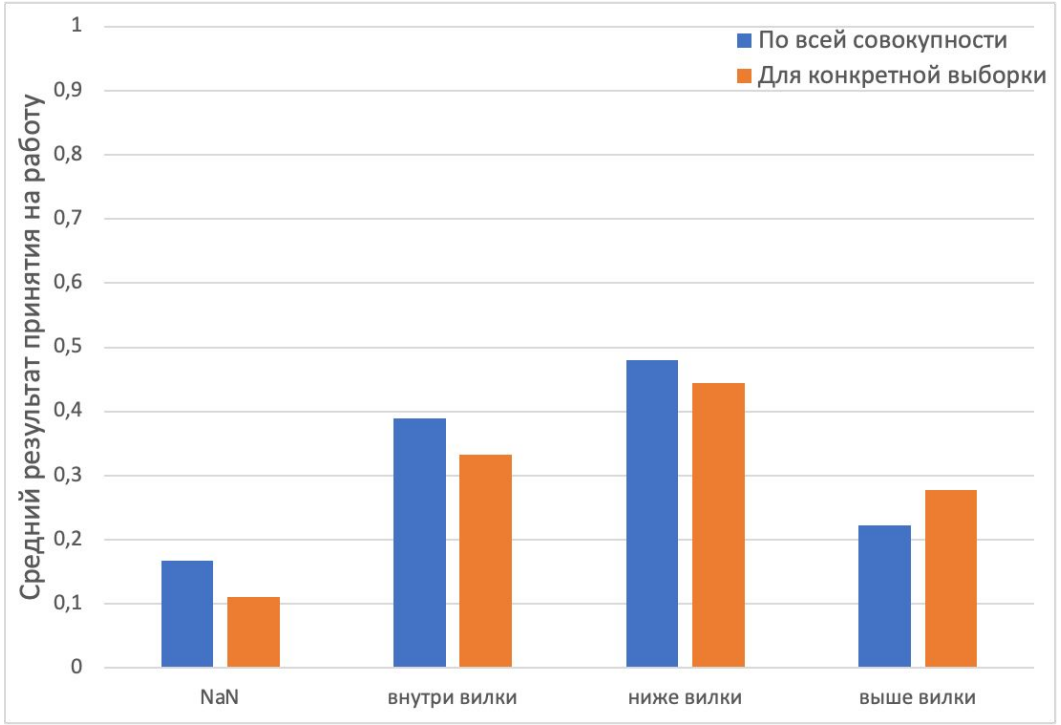
$$x_{high} = \frac{res\_high\_sample}{res\_high\_general}$$

$$x_{nan} = \frac{res\_nan\_sample}{res\_nan\_general}$$

$$x_{norm} = \frac{res\_norm\_sample}{res\_norm\_general}$$

$$x_{low} = \frac{res\_low\_sample}{res\_low\_general}$$

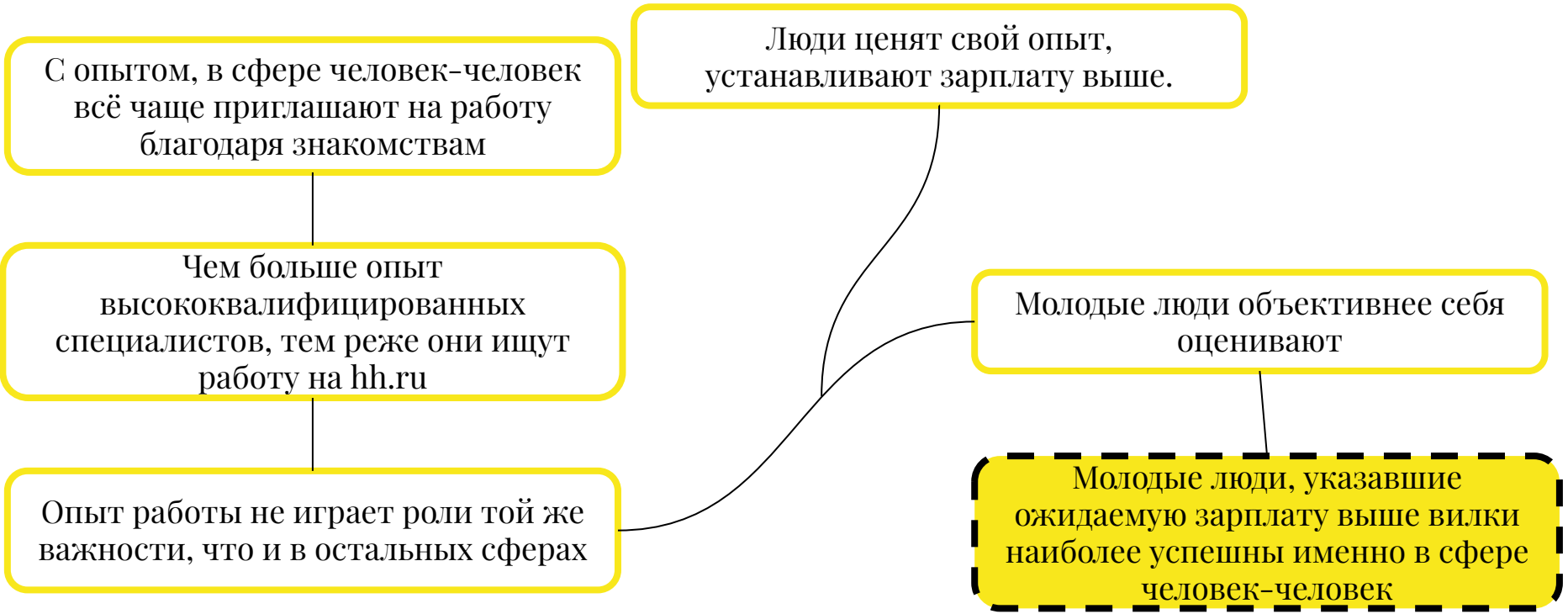
$$k = \frac{x_{high}}{x_{high} + x_{nan} + x_{norm} + x_{low}}$$



Гипотеза

Люди, успешно подавшие резюме на вакансии профиля человек-человек с зарплатными ожиданиями выше предлагаемых, моложе других людей, приглашённых на собеседование с тем же отношением зарплатных ожиданий к “вилке”

Механизм


С опытом, в сфере человек-человек всё чаще приглашают на работу благодаря знакомствам
Чем больше опыт высококвалифицированных специалистов, тем реже они ищут работу на hh.ru
Опыт работы не играет роли той же важности, что и в остальных сферах
Люди ценят свой опыт, устанавливают зарплату выше.
Молодые люди объективнее себя оценивают
Молодые люди, указавшие ожидаемую зарплату выше вилки наиболее успешны именно в сфере человек-человек

Исследуемые показатели

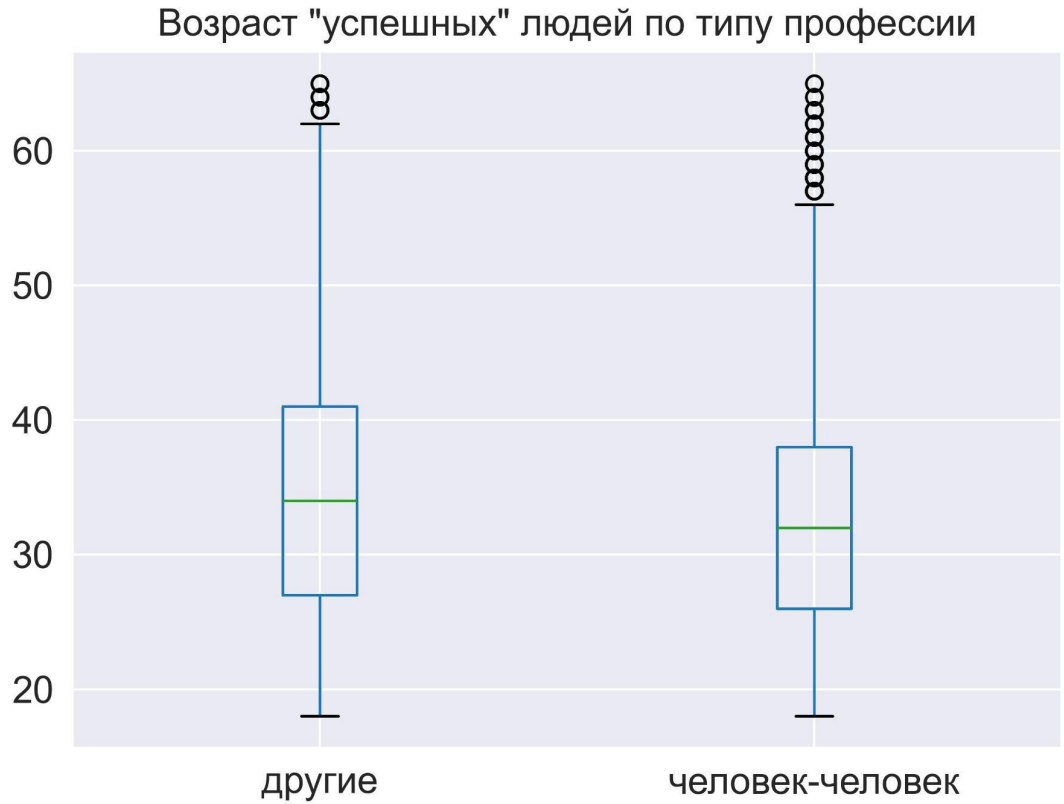
Объясняемая переменная	Единица измерения	Обозначение
Ожидаемая зарплата	₽	expected_salary
Верхняя граница	₽	compensation_to
Начальный статус взаимодействия (приглашение, отклик, отказ)	—	inititial_state
Финальный статус взаимодействия (приглашение, отклик, отказ)	—	final_state

Математическая модель

1. Соискателя пригласили на работу, хотя он указал зарплату выше зарплатной вилки
2. Две группы профессий: человек-человек; другие типы
3. С помощью U-критерия Манна-Уитни измеряем различие этих групп по **возрастам**
4. Убеждаемся в статистической значимости фактора ($p=0.023<0.050$)
5. Определяем у какой **группы профессий** возраст меньше

Математическая модель

Медиана
человек-человек — 32
другие — 34



Проверка устойчивости матмодели

- Пропуски в **верхней границе зарплатной вилки** заполняем с помощью модели машинного обучения CatBoost*
- Повторяем построение математической модели для новой выборки
- Проверяем, что гипотеза подтверждается для неё ($p=0.016<0.050$)
- Математическая модель устойчива!

**дополнительная информация об обучении модели в приложении*

Интерпретация математического подтверждения

Приглашённые на собеседование люди, указавшие ожидаемую заработную плату выше предлагаемой нанимателем, в среднем моложе в профессиях типа человек-человек

Перспективы

- Собрать данные о большем количестве профессий. В разных сферах важны разные факторы
- Собрать данные о компаниях различного размера. Это может повлиять на специфику отбора кандидатов
- Собрать данные о результатах собеседования. Взяли ли соискателя на работу
- Собрать данные об итоговой зарплате после собеседования

Ограничения

- Исследование нельзя обобщить для других методов рекрутинга
- Исследование нельзя обобщить для специфических представителей данной сферы деятельности
- Исследование нельзя обобщить для различных состояний экономики
- В выборке нет профессионалов “высшей категории”, исследование для них не применимо

Практическая польза

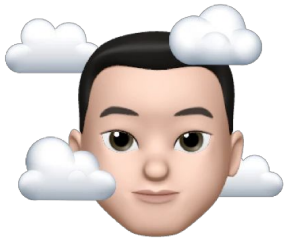
Анализ возможности указания ожидаемой заработной платы выше предлагаемой в сфере человек-человек

Policy implication

Молодые люди, могут смелее претендовать на заработную плату, выше “зарплатной вилки” на вакансии, изучаемого профиля

Возрастные соискатели с высокими зарплатными ожиданиями, ищущие работу в сфере человек-человек снизят ожидаемую зарплату (переквалифицируются) и будут успешнее устраиваться на работу

Участники команды ANOD



Кадырзанов
Борис

Головач
Владимир

Терентьев
Иван

Гасаненко
Арсений

Кудрявцева
Дарья

Any questions?

Приложение. Большие таблички

https://docs.google.com/spreadsheets/d/1cChlfnHYx10qiNxs8KExAVmif2yQSxa1frZaPmMZ_ww/edit?usp=sharing

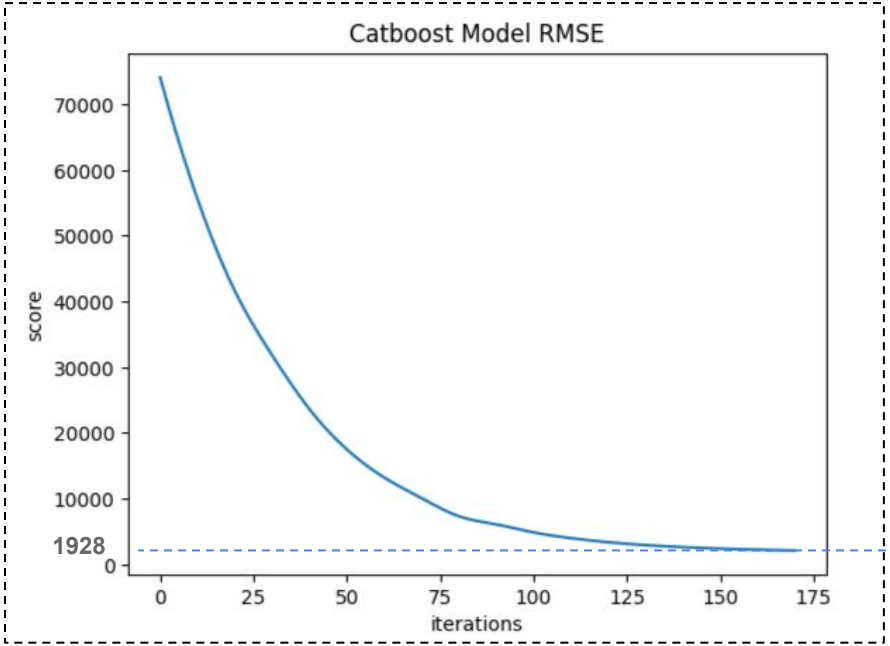
Приложение. О работе CatBoost

CatBoost— модель градиентного бустинга, обучаемого на:

- количественных признаках **вакансии** (нижняя граница зарплатной вилки)
- текстовых характеристиках **вакансии** (необходимые навыки)
- категориальных(регион, год создания, тип занятость и название профессии) признаках **вакансии**

Модель не обучается на данных из других строк, поэтому полученная благодаря ml выборка —новая

Полученный RMSE финальной регрессионной модели - 1928 рублей, гиперпараметры подбираются с помощью Grid Search



Приложение. Классификация Климова

По классификации, предложеннной Е.А. Климовым, выделяют 5 типов профессий:

1. Человек – живая природа. Работа с растительными и животными организмами, микроорганизмами и условиями их существования. (ветеринар, агрохимик)
2. Человек – техника (и неживая природа). Работа с неживыми, техническими объектами труда. (сварщик, водитель)
3. Человек – человек. Работа с социальными системами, сообществами, группами населения, людьми разного возраста. (няня, учитель)
4. Человек – знаковая система. Естественные и искусственными языками, условными знаками, символами, цифрами, формулами. (программист, экономист)
5. Человек – художественный образ. Работа с явлениями, фактами художественного отображения действительности. (художник, дизайнер)

Приложение. О малых подвыборках

Доктора. 60 штук. На собеседование 4. Возраст 57. Ожидаемая ЗП 150000.0

Волонтёры. Не подходят, мы изучаем зарплаты. Их всего 873 + 5

Ненецкий и Чукотские АО. Самая большая медианная ЗП, но всего 4 + 8 значений

Донецкая и Луганская области. 200 значений. По другому устроен рынок труда