**# Аналитическая записка**

***Дата: 04.04.2024***

**## Введение**

 - Данная аналитическая записка была создана для анализа навыков, а также их кластеров(групп), на которые они распределяются. Цель, которая преследуется этой запиской - провести анализ востребованных навыков, самых оплачиваемых навыков, а также региональную специфику востребованных навыков. Для кандидатов в сфере BI-Аналитики я считаю, что навыки анализа необходимы в-первую очередь.

**## Методы и источники данных**

 - Выбранные методы анализа в работе:

  Анализ был проведен на основе построенного дашборда по данным, собранным по всей России, по компетенции - BI-Аналитик, также были рассмотрены методы программного анализа наиболее востребованных и наименее востребованных навыков, наиболее оплачиваемые навыки, региональная специфика востребованных навыков по регионам.

  Хочу выделить этапы работы, которые были пройдены перед проведением всех анализов:

  1. Для начала я выгрузил данные из БД

  2. После пред обработал их:

  \* Удалил ненужные знаки из описаний

  \* Лемматизировал слова - привел к инфинитиву

  3. На следующем этапе я уже начал реализовать алгоритмы Natural Language Processing(NLP):

  \* Сначала я обучил классификатор, используя в нем инструменты: модель типа BERT для векторизации слов, а также алгоритм K-Nearest Neighbors для поиска слов по двум классам: "Навык" и "Не Навык"

  \* Дальше я реализовал кластеризатор, используя для него тот же инструмент для векторизации слов - BERT модель, а для алгоритма кластеризации я выбрал - KMeans

  4.  После проведения всех манипуляций с данными, я получил слова из описаний, которые являются навыками и дальше провел с этими словами анализы:

  \* Определил наиболее востребованные и наименее востребованные навыки, суть анализа - я посчитал вхождение всех выявленных навыков из описаний и вывел 10 наиболее востребованных навыков, и 10 наименее востребованных навыков.

  \* Определил наиболее оплачиваемые навыки, суть анализа - я посчитал среднюю зарплату по вакансии, дальше сравнивал зарплаты каждой вакансии, если зарплата вакансии в два раза больше, чем средняя, то маркировал все определенные навыки из описания как - "Высокооплачиваемые навыки".

  \* Определил региональную специфику востребованных навыков - суть анализа, я посчитал встречаемость навыков, относительно каждого региона и вывел для каждого региона 5 самых востребованных навыков.

 - Выбранные источники данных:

  Всю необходимую информацию о представленных данных можно найти в Аналитической записки из предыдущего модуля Но хочется дополнить

**## Результаты проведенного анализа**

 - Результаты, проведенного анализа, которые бы хотелось отметить:

  Во-первых, давайте взглянем на анализ наиболее востребованных навыков и наименее востребованных навыков. Модель определила наиболее востребованные навыки в основном связанные с опытом или работой. Отсюда я бы хотел сказать, что компания не слишком сильно смотрят на технические навыки, думаю компания готова обучить своего сотрудника, если он готов работать, стараться и получать опыт.

  Во-вторых, давайте посмотрим на анализ наиболее оплачиваемых навыков, в которых очень часто встречаются мало знакомые слова, видимо это из-за того, что это более профессиональные инструменты, которыми не нужно пользоваться в повседневной жизни: betboom, senior(уровень работника относительно его опыта работы), debezium, 'iceberg. Данные инструменты я не встречал на своем пути, хочется повториться, что скорее всего они предусмотрены для более узко уровня работы и поэтому являются высокооплачиваемыми навыками.

  В-третьих, давайте разберем финальный анализ, который я вывел с помощью кода - определение наиболее востребованных навыков, относительно региона. В нем можно выделить, что относительно всех регионов требования различаются не критично, а также относятся к группе: "Если ты готов обучиться и будешь стараться, то мы с радостью тебя ждем".

  В-четвертых, давайте взглянем на дашборд, который был построен относительно кластеров и количества навыков в нем.

  \* Первый кластер содержит в себе 1063 навыка, которые я бы хотел отнести к категории: "Иноязычных навыков".

  \* Второй кластер содержит в себе 2103 навыка, которые я бы хотел отнести к категории: "Не технические навыки". Наверное, этот кластер я бы отнес к soft-skills.

  \* Третий кластер содержит в себе 2725 навыка, которые я бы хотел отнести к категории: "Технические навыки". В этом кластере преобладают технические навыки связанные с выполнением работы в техническом плане.

  В целом хочется отметить, что разбитые навыки очень схожи к друг другу. А также преобладание в количестве у технических навыков. Второй же по количество навыков идет кластер "Не технических навыков"/"Soft Skills".

**## Выводы и рекомендации**

 - Давайте подведем итоги проведенного анализа и его результатов. Хочется вернуться к кластерам, на которые были разделены полученные навыки из описаний. Давайте вспомним факт, что преобладают в количестве навыков кластеры: "Технические навыки"/"Hard Skills" и "Не технические навыки"/"Soft Skills".

 - Под конец хотелось бы вынести рекомендации для всех кандидатов на работу в этой сфере, а также для людей, которые уже связали свою жизнь с этой сферой. Относительно полученных выводов, хочется выделить умение работать с техникой, разными технологичными инструментами: ЯП(Языки программирования), языки запросов(SQL, DAX), визуальные инструменты представления данных(Дашборды, презентации). Но не стоит забывать про "Не технические навыки", которые включают в себя все навыки, чтобы работник качественно презентовал свою бизнес модель или убедил в внесение новых инструментов в бизнес.

 Также хочу выделить несколько самых необходимых навыков относительно проведенного анализа и моего мнения: знание SQL, знание специальных ЯП(Python, R), инструменты визуального представления данных(Power BI, Yandex DataLense, Power Point, Figma и др.)

**## Ограничение анализа**

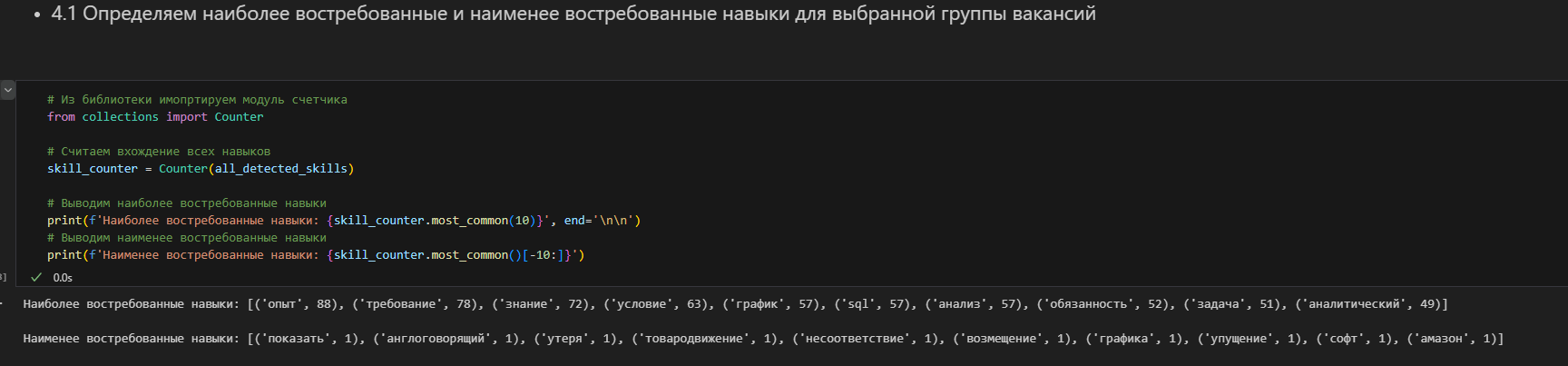
 - Хочу отметить ограничения, с которыми я столкнулся во время анализа: ограничение, с которыми я столкнулся до анализа навыков присутствуют в аналитической записке: . Но также хочется пополнить список:

 1. Слабое железо, которые не позволяет проводить качественный и корректный анализ, так как компьютер не выдерживает всех вычислений.

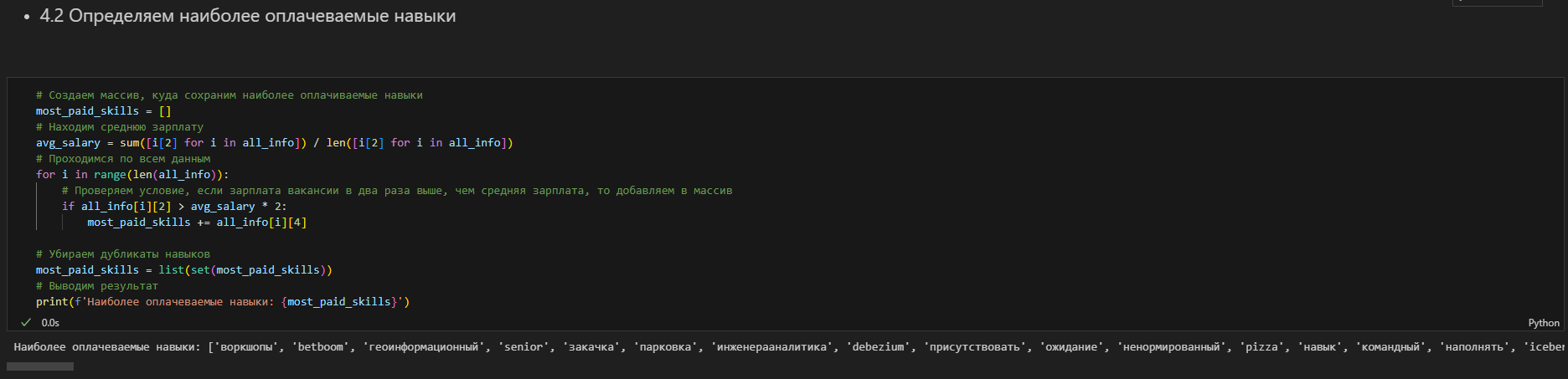
 2. Несовершенство полученных моделей - полученные модели не являются идеальными и в лучшем счете было реализовать не машинное обучение, а глубокое, которые позволяет улавливать более тонкие зависимости между данными, но из-за слабого железа и ограниченности по времени, это реализовать невозможно.

**## Данные и графики, использовавшиеся в анализе**

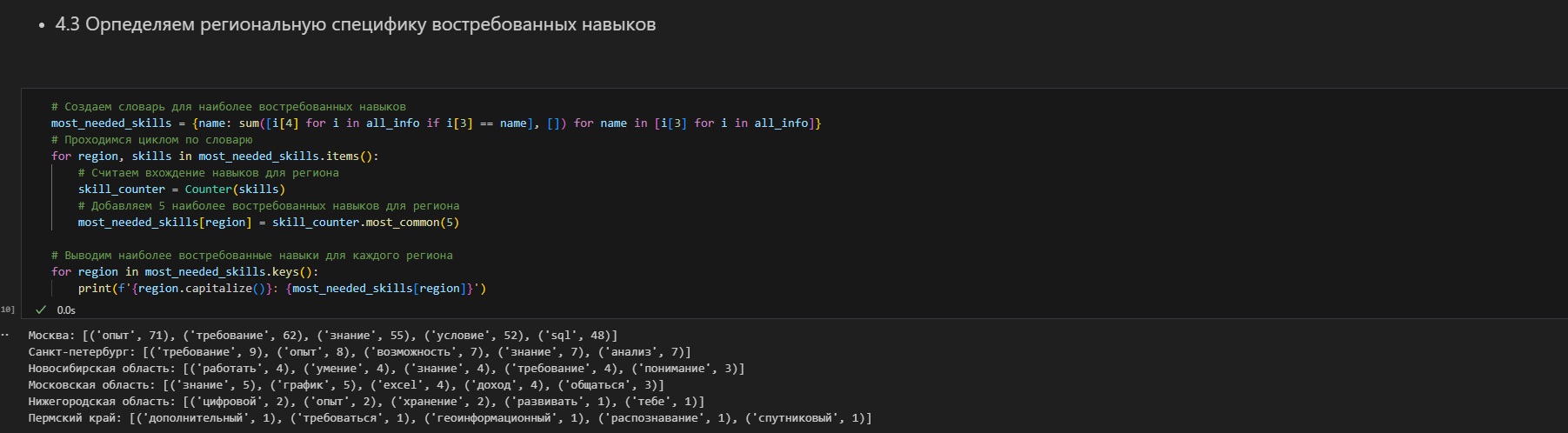
1. Данные для анализа наиболее востребованных и наименее востребованных навыков в пункте: "Результаты проведенного анализа" -> 1-й проведенный анализ.



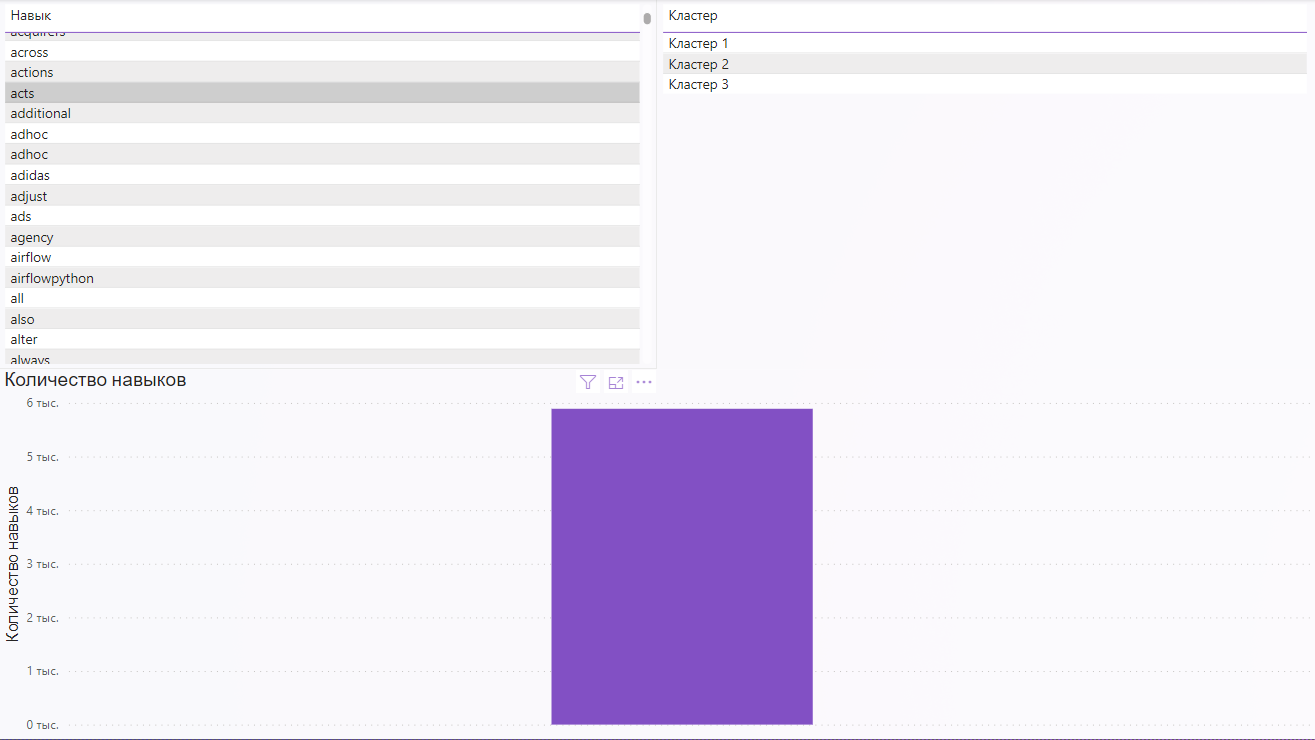
2. Данные для анализа наиболее высокооплачиваемых навыков в пункте: "Результаты проведенного анализа" -> 2-й проведенный анализ.

! 

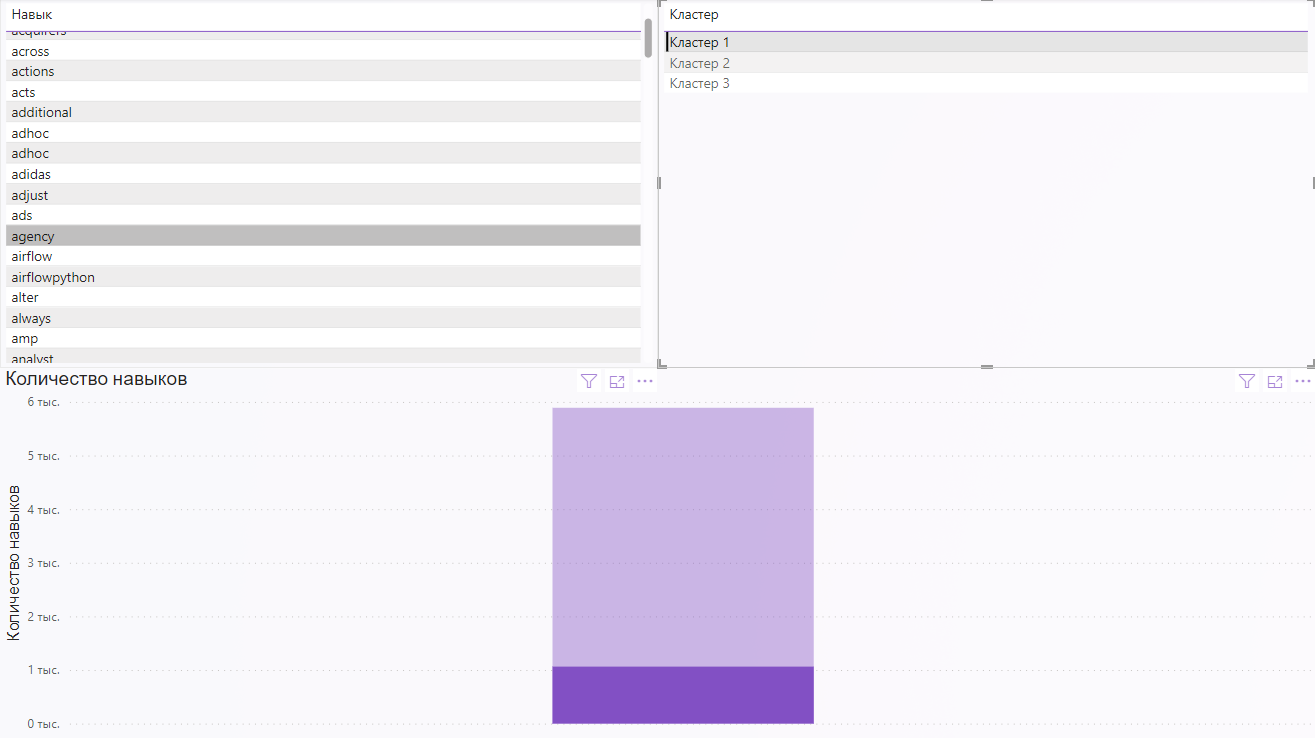
3. Данные для анализа наиболее востребованных навыков относительно регионов в пункте: "Результаты проведенного анализа" -> 3-й проведенный анализ.



4. График для анализа навыков по кластерам: "Результаты проведенного анализа" -> 4-й проведенный анализ.



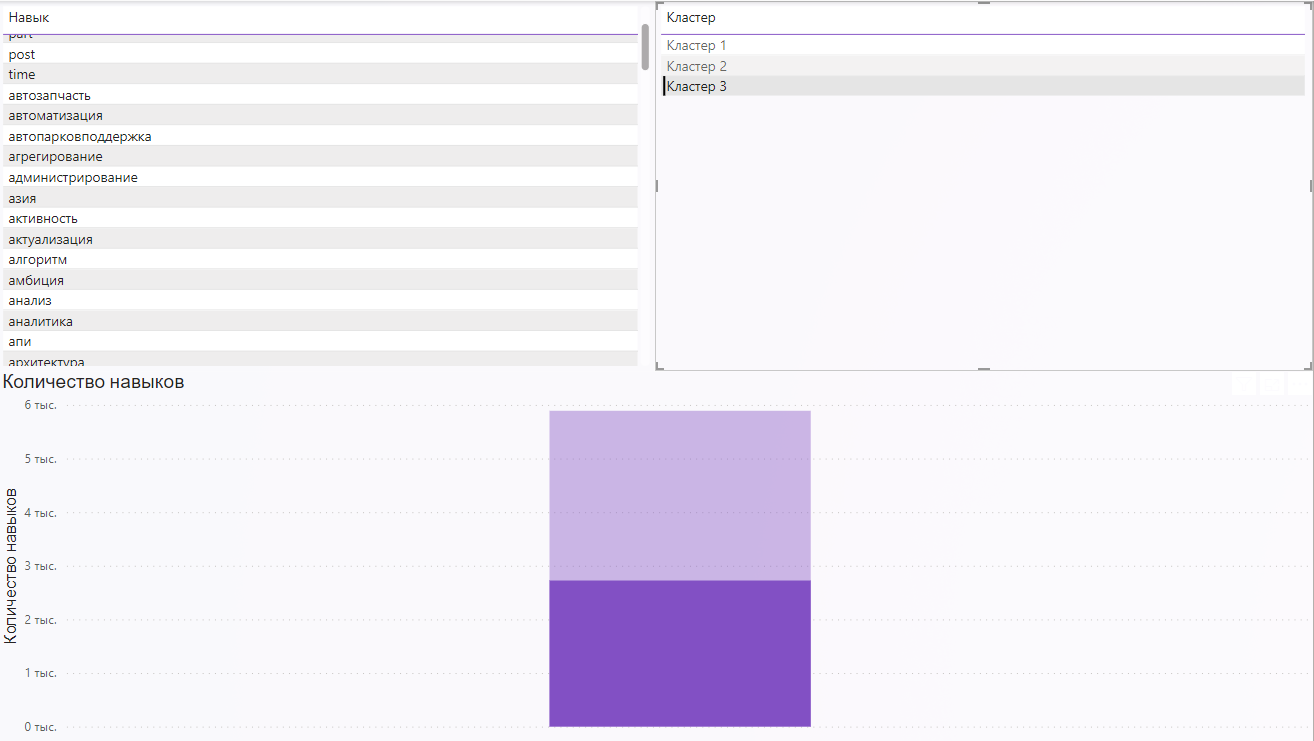
5. График для анализа навыков по первому кластеру: "Результаты проведенного анализа" -> 4-й проведенный анализ.



6. График для анализа навыков по второму кластеру: "Результаты проведенного анализа" -> 4-й проведенный анализ.



7. График для анализа навыков по третьему кластеру: "Результаты проведенного анализа" -> 4-й проведенный анализ.



***Дата и подпись:***

***04.04.2024 Rabov A.A***