*Проект 2. SQL*

*Подгрузка новых данных. Уточнение анализа*

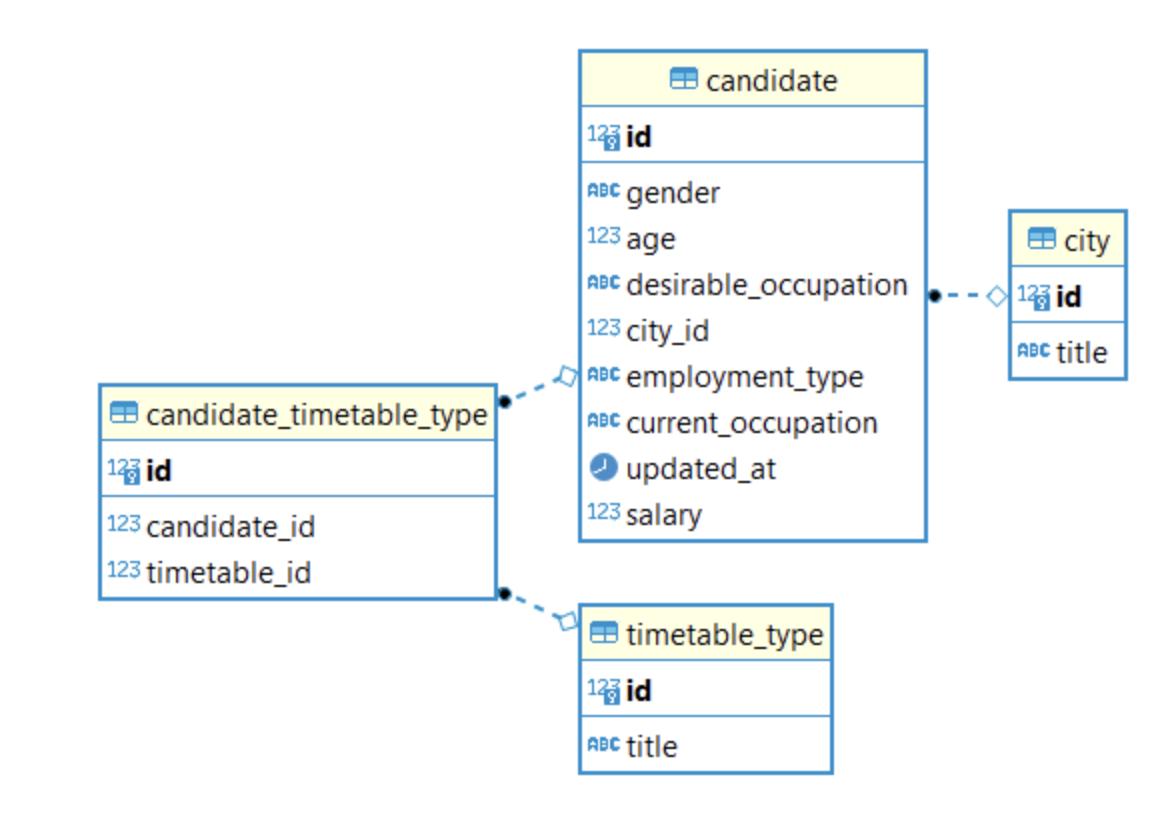
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Подготовлен: Кравченко Маргаритой Михайловной (группа DSPR-78)*

## **1. Введение. Знакомство с датасетом**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## 



## **2. Предварительный анализ данных** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## 

### Задание 2.1

*Рассчитайте максимальный возраст (max\_age) кандидата в таблице.*

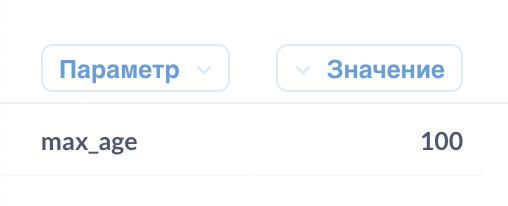
* **Код запроса[[1]](#footnote-1):**

SELECT 'max\_age' AS Параметр,

MAX(age) AS Значение

FROM hh.candidate

* **Результат:**



* **Вывод:**

Почти наверняка, данное значение является аномалией. Можно предположить, что кандидат заполняющий в резюме параметр возраст цифрой “100” - или не хочет его указывать (но не имеет возможности оставить поле пустым) или пытается привлечь внимание работодателя своей оригинальностью.

### Задание 2.2

### *Теперь давайте рассчитаем минимальный возраст (min\_ age) кандидата в таблице.*

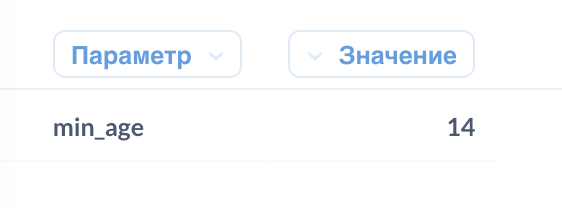
* **Код запроса:**

SELECT 'min\_age' AS Параметр,

MIN(age) AS Значение

FROM hh.candidate

* **Результат:**

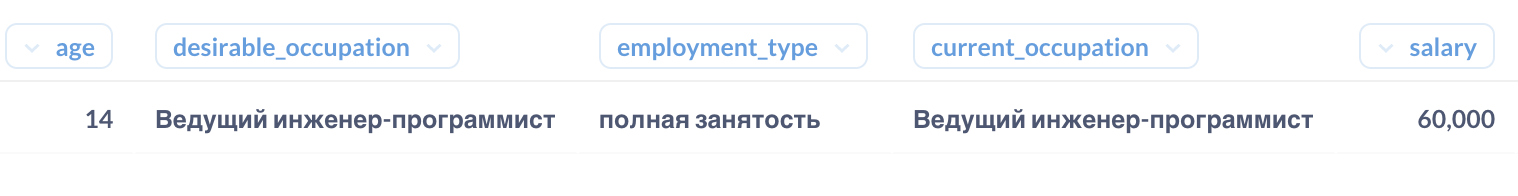


* **Вывод:**

Минимальное значение возраста - 14 вполне может быть как ошибкой ввода данных, так и вполне реальными данными. Реальность для сферы IT такова, что даже очень молодые люди могут чувствовать себя компетентными для некоторых вакансий.

* **Дополнительно:**

Интересно оценить остальные параметры кандидата с минимальным возрастом:



Из этих данных видно, что, скорее всего, цифра 14 - ошибка. Маловероятно, что 14 летний кандидат ищет работу и в данный момент работает на должности “Ведущий инженер-программист” на полный рабочий день.

### Задание 2.3

### *Попробуем «очистить» данные. Напишите запрос, который позволит рассчитать для каждого возраста (age) сколько (cnt) человек этого возраста у нас есть. Отсортируйте результат по возрасту в обратном порядке.*

* **Код запроса:**

SELECT age,

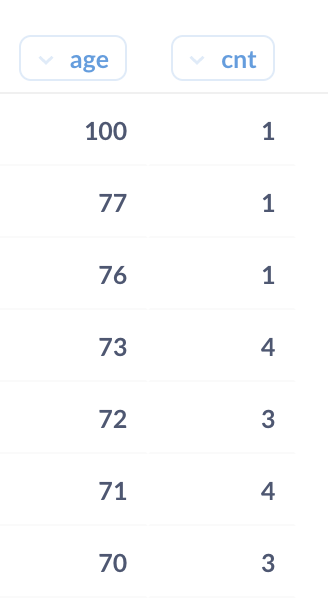
COUNT(id) AS cnt

FROM hh.candidate

GROUP BY age

ORDER BY age DESC

* **Результат:**



* **Вывод:**

Теперь очевидно, что значение возраста - 100, это аномалия. Оно встречается однажды и через внушительный интервал после других значений.

### Задание 2.4

### *По данным Росстата, средний возраст занятых в экономике России составляет 39.7 лет. Мы округлим это значение до 40. Найдите количество кандидатов, которые старше данного возраста. Не забудьте отфильтровать «ошибочный» возраст 100.*

* **Код запроса:**

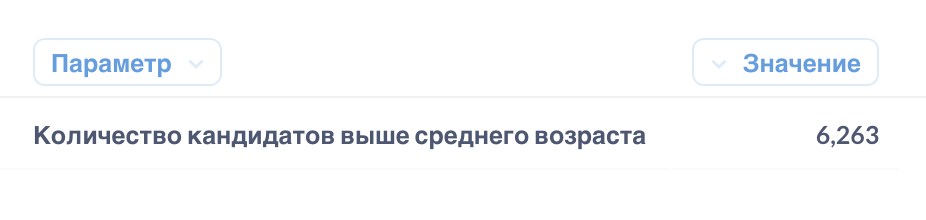
SELECT 'Количество кандидатов выше среднего возраста' AS Параметр,

COUNT(id) AS Значение

FROM hh.candidate

WHERE age > 40 AND age < 100

* **Результат:**



* **Вывод:**

Количество соискателей выше среднего возраста - совсем не велико, если обратиться к общему числу резюме в базе hh, то можно понять, что это менее 15% кандидатов. Эти данные вполне отвечают специфике рынка труда в сфере IT - молодая активно развивающаяся отрасль, большинство сотрудников которой - это молодые люди.

* **Дополнительно:**

Посмотрим количество кандидатов выше среднего возраста в соотношении с общим количеством кандидатов:

## 

## **3. Глобальный анализ показателей**

## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## 

### Задание 3.1

*Для начала напишите запрос, который позволит узнать, сколько (cnt) у нас кандидатов из каждого города (city). Формат выборки: city, cnt. Отсортируйте результат по количеству в обратном порядке.*

* **Код запроса:**

SELECT city.title AS city,

COUNT(c.id) AS cnt

FROM hh.candidate AS c

JOIN hh.city AS city

ON c.city\_id = city.id

GROUP BY city.title

ORDER BY cnt DESC

* **Результат:**



* **Вывод:**

Количество соискателей ожидаемо распределяется по городам примерно в соответствии с размером города и объемом предложений в сфере IT. Наиболее развит этот рынок в двух крупнейших городах России - Москва и Санкт-Петербург, далее с заметным отставанием следуют города миллионники и остальные.

### Задание 3.2

*Москва бросается в глаза как, пожалуй, самый активный рынок труда. Напишите запрос, который позволит понять, каких кандидатов из Москвы устроит «проектная работа».*

***Формат выборки:*** *gender, age, desirable\_occupation, city, employment\_type.*

*Отсортируйте результат по id кандидата.*

* **Код запроса:**

SELECT c.gender,

c.age,

c.desirable\_occupation,

city.title AS city,

c.employment\_type

FROM hh.candidate AS c

JOIN hh.city AS city

ON c.city\_id = city.id

WHERE city.title = 'Москва'

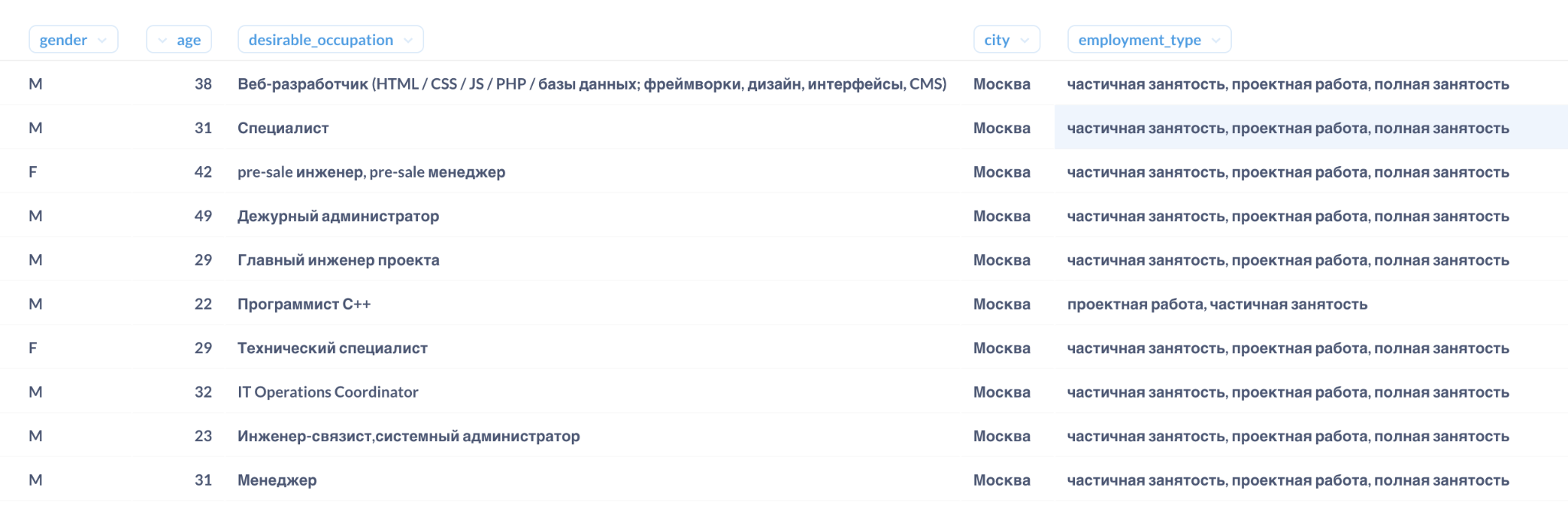
/\*тип занятости “проектная работа” может быть как единственным вариантом,

так и в числе других, поэтому используем LIKE \*/

AND c.employment\_type LIKE '%проектная работа%'

ORDER BY c.id

* **Результат:**



* **Вывод:**

Получена довольно объемная выборка, в которой можно встретить самые разнообразные должности от руководителей до операторов ПК. “Проектная работа” перечисляется соискателями в числе многих прочих типов занятости очевидно, чтобы охватить как можно больший объем предложений по работе.

### Задание 3.3

*Данных оказалось многовато. Отфильтруйте только самые популярные IT-профессии — разработчик, аналитик, программист.*

*Обратите внимание, что данные названия могут быть написаны как с большой, так и с маленькой буквы.*

*Отсортируйте результат по id кандидата.*

* **Код запроса:**

SELECT c.gender,

c.age,

c.desirable\_occupation,

city.title AS city,

c.employment\_type

FROM hh.candidate AS c

JOIN hh.city AS city

ON c.city\_id = city.id

WHERE city.title = 'Москва'

AND c.employment\_type LIKE '%проектная работа%'

/\*оставляем все варианты, содержащие искомые слова (игнорируя регистр)\*/

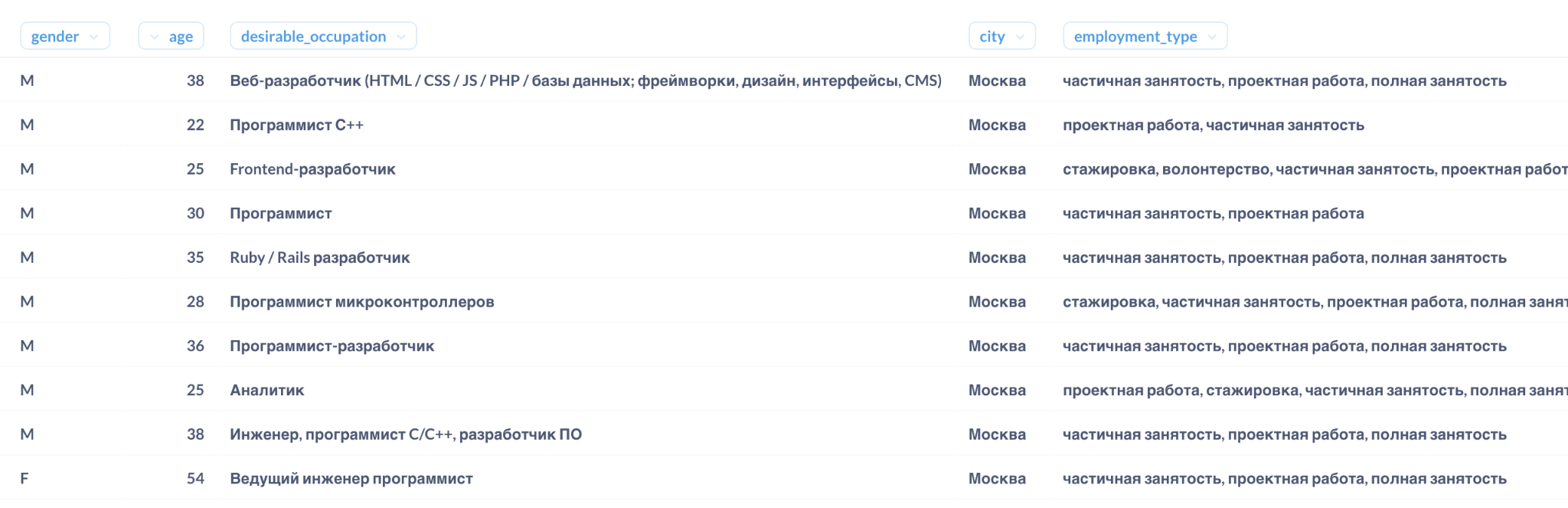
AND (LOWER(c.desirable\_occupation) LIKE '%разработчик%'

OR LOWER(c.desirable\_occupation) LIKE '%аналитик%'

OR LOWER(c.desirable\_occupation) LIKE '%программист%')

ORDER BY c.id

* **Результат:**

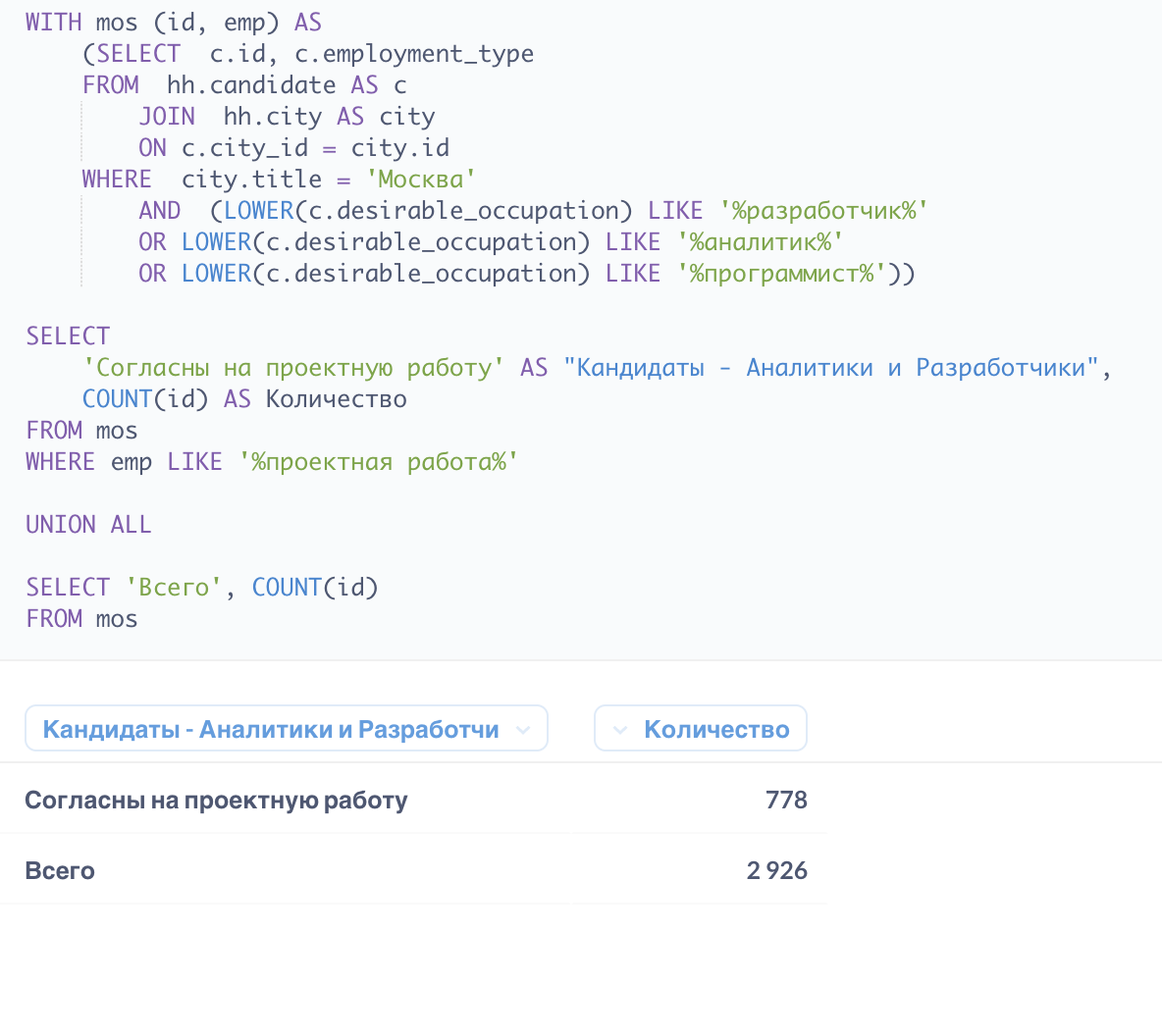
****

* **Вывод:**

Видно, что в этой выборке действуют те же закономерности, что и в предыдущей. Кандидаты не целенаправленно ищут проектную работу, а указывают ее как один из возможных вариантов трудоустройства.

Тут интересно посмотреть, какая часть от всех разработчиков и аналитиков рассматривает проектную работу в качестве возможной.

* **Дополнительно:**



То есть, каждый четвертый кандидат выбранных вакансий рассматривает проектную работу в качестве возможного для себя типа занятости.

### Задание 3.4

*Для общей информации попробуйте выбрать номера и города кандидатов, у которых занимаемая должность совпадает с желаемой.*

*Формат выборки: id, city.*

*Отсортируйте результат по городу и id кандидата.*

* **Код запроса:**

SELECT c.id,

city.title AS city

FROM hh.candidate AS c

JOIN hh.city AS city

ON c.city\_id = city.id

WHERE c.current\_occupation = c.desirable\_occupation

ORDER BY city, c.id

* **Результат:**



* **Вывод:**

Проверяем выборку, видно, что она соответствует и отсортирована согласно заявленным параметрам. Опираясь только на эти данные сложно сделать какой-то более или менее осмысленный вывод (разве что констатировать факт, что есть внушительное количество кандидатов желающих сменить работу, не рассматривая изменение занимаемой должности).

### Задание 3.5

*Определите количество кандидатов пенсионного возраста.*

*Пенсионный возраст для мужчин наступает в 65 лет, для женщин — в 60 лет.*

* **Код запроса:**

SELECT 'Количество кандидатов пенсионного возраста' AS Параметр,

COUNT(id) AS Значение

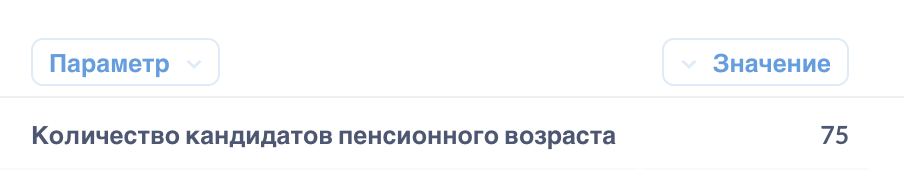
FROM hh.candidate

WHERE ((gender = 'F' AND age >= 60) /\*фильтр для женщин (пенсионный возраст с 60)\*/

OR (gender = 'M' AND age >= 65)) /\*для мужчин (пенсионный возраст с 65)\*/

AND age < 100 /\*удаляем аномалию - 100-летнего кандидата\*/

* **Результат:**



* **Вывод:**

Полученные данные подтверждают ранее сделанный вывод - что соискатели в сфере IT за явным преимуществом - это молодые люди. Кандидаты пенсионного возраста - заметная редкость. Такие резюме составляю долю - значительно меньше одного процента от общего числа.

## **4. Анализ кандидатов для заказчиков**

## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## 

### 

### Задание 4.1

*Для добывающей компании нам необходимо подобрать кандидатов из Новосибирска, Омска, Томска и Тюмени, которые готовы работать вахтовым методом.* ***Формат выборки:*** *gender, age, desirable\_occupation, city, employment\_type, timetable\_type. Отсортируйте результат по городу и номеру кандидата.*

* **Код запроса:**

SELECT c.gender,

c.age,

c.desirable\_occupation,

city.title AS city,

c.employment\_type,

t.title AS timetable\_type

FROM hh.candidate AS c

/\*через дополнительную таблицу устанавливаем связь с id графика работы\*/

JOIN hh.candidate\_timetable\_type AS ct

ON c.id = ct.candidate\_id

/\*присоединяем таблицу-справочник названий графиков работы\*/

JOIN hh.timetable\_type AS t

ON ct.timetable\_id = t.id

JOIN hh.city AS city

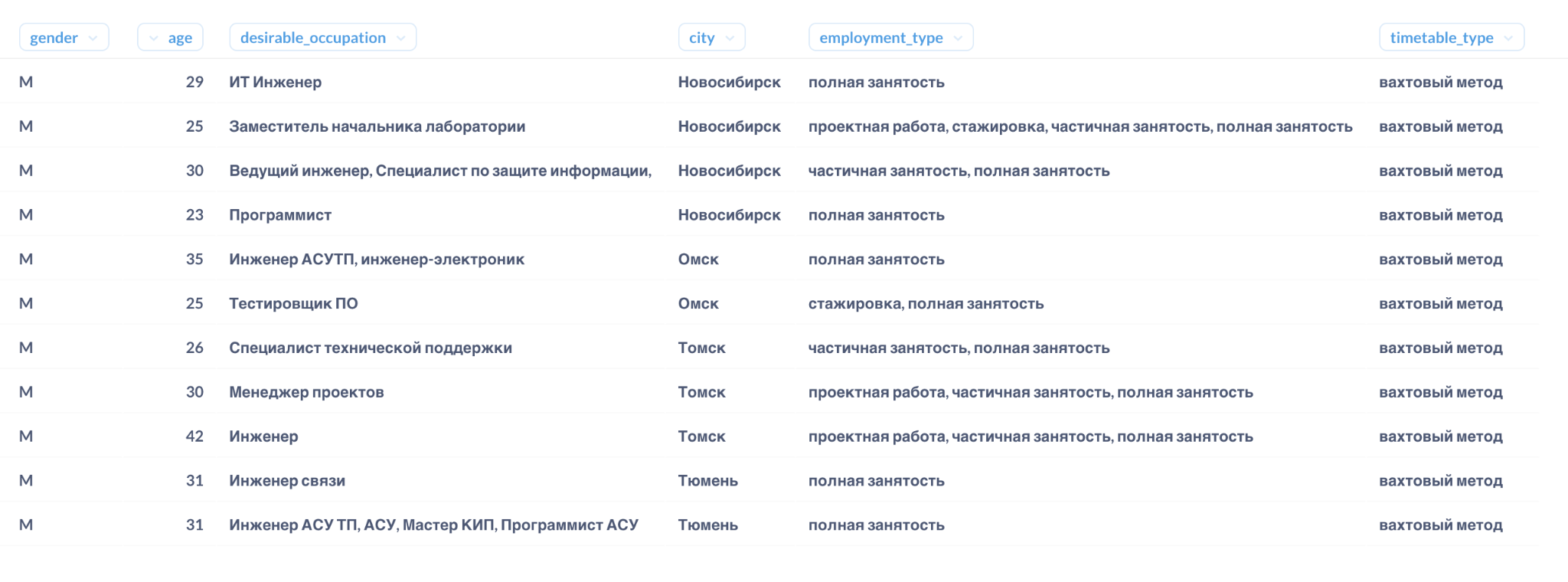
ON c.city\_id = city.id

WHERE city.title IN ('Новосибирск', 'Омск', 'Томск', 'Тюмень')

AND t.title = 'вахтовый метод'

ORDER BY city, c.id

* **Результат:**

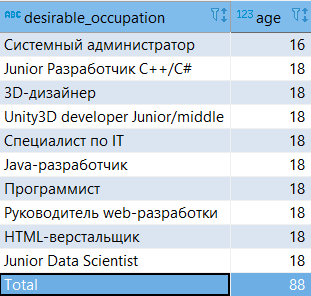


* **Вывод:**

Кандидатов, отвечающих заявленным требованиям совсем не много (все 11 резюме выборки уместились в скрин таблицы), и, для извлечения их данных из базы необходимо составить довольно детализированный запрос.

### Задание 4.2

*Для заказчиков из Санкт-Петербурга нам необходимо собрать список из 10 желаемых профессий кандидатов из того же города от 16 до 21 года (в выборку включается 16 и 21, сортировка производится по возрасту) с указанием их возраста, а также добавить строку Total с общим количеством таких кандидатов. Напишите запрос, который позволит получить выборку вида:*

**

* **Код запроса:**

/\*во избежание дублирования кода формируем дополнительную таблицу (CTE)\*/

WITH sp AS (SELECT c.id,

c.desirable\_occupation,

c.age

FROM hh.candidate AS c

JOIN hh.city AS city

ON c.city\_id = city.id

WHERE city.title = 'Санкт-Петербург'

AND age BETWEEN 16 AND 21)

(SELECT desirable\_occupation, age

FROM sp

ORDER BY age

LIMIT 10)

UNION ALL

(SELECT 'Total', COUNT(id)

FROM sp)

* **Результат:**



* **Вывод:**

Получен список, соответствующий всем заявленным параметрам, а также общее количество кандидатов в этом сегменте. При беглой оценке сферы интересов молодых специалистов, можно сказать, что преимущественно это предложения для начинающих (очевидно) и разнообразные вакансии из области разработки, в частности WEB. Произвольный список из 10 профессий в сделанной выборке хорошо иллюстрирует эти закономерности.

## **ИТОГИ**

## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## 

Опираясь на имеющийся анализ данных рынка труда в сфере IT, и дополняя его некоторыми аналитическими данными можно сделать следующие выводы:

1. Соискатели - это, преимущественно очень молодые люди. Средний возраст кандидата в базе - 32.2 года, в то время, как по данным Росстата, средний возраст занятых в экономике России составляет 39.7 лет.

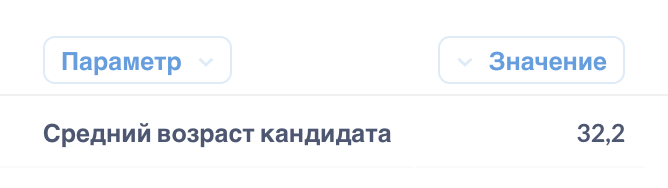
* **Код запроса:**

SELECT 'Средний возраст кандидата' AS Параметр,

AVG(age) AS Значение

FROM hh.candidate

* **Результат:**



Рынок IT в России сейчас развивается очень активно, новые компании сравнительно быстро растут, привлекая молодых сотрудников. В свою очередь, молодые люди, выбирая сферу профессиональной деятельности (часто впервые) все больше отдают предпочтение именно работе в информационных технологиях, как наиболее перспективной.

1. География проживания соискателей - невероятно широка. Всего представлено резюме кандидатов из почти тысячи населенных пунктов.

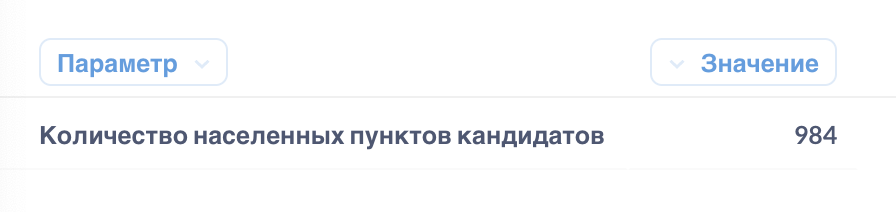
* **Код запроса:**

SELECT 'Количество населенных пунктов кандидатов' AS Параметр,

COUNT (DISTINCT city\_id) AS Значение

FROM hh.candidate

* **Результат:**



При этом распределение количества кандидатов по городам чрезвычайно неравномерно. Более половины(!) кандидатов сосредоточено всего в 3 из почти тысячи населенных пунктов.

* **Код запроса:**

WITH top3 ("ТОП-3 Городов", "Количество резюме", "Доля от общего")

AS (SELECT city.title,

COUNT (c.id) AS cnt,

COUNT (c.id) / (SELECT COUNT(id)

FROM hh.candidate)::numeric

FROM hh.candidate AS c

JOIN hh.city AS city

ON c.city\_id = city.id

GROUP BY city.title

ORDER BY cnt DESC

LIMIT 3)

SELECT \*

FROM top3

UNION ALL

SELECT 'ИТОГО ТОП-3 Городов',

SUM("Количество резюме"),

SUM("Доля от общего")

FROM top3

* **Результат:**



Наиболее перспективными как для соискателя, так и для работодателя сегодня являются крупнейшие города России, это с большим отрывом - две столицы, в меньшей степени - города-миллионники.

1. Используя представленные данные можно детализировать исследование в любом направлении, в зависимости от целей. Например посмотреть средний уровень желаемой заработной платы соискателей в сфере аналитики с разбивкой по городам.

Оставим только те, где таких кандидатов больше 10, чтобы сделать среднии показатель более репрезентативный и оценим наиболее перспективные регионы для соискателя (ТОП - 10 по уровню заработных плат городов).

* **Код запроса:**

SELECT city.title AS "Город" ,

COUNT(c.id) AS "Соискатели-аналитики" ,

AVG(c.salary) AS "Средняя желаемая ЗП"

FROM hh.candidate AS c

JOIN hh.city AS city

ON c.city\_id = city.id

WHERE LOWER(c.desirable\_occupation) LIKE '%data%'

OR LOWER(c.desirable\_occupation) LIKE '%анали%'

GROUP BY city.title

HAVING COUNT(c.id) >=10

ORDER BY AVG(c.salary) DESC

LIMIT 10

* **Результат:**



Получается, что города лидеры по объемам рынка труда в сфере IT (Москва, Санкт-Петербург, Краснодар) даже не в первой пятерке по уровню желаемых зарплат соискателей аналитиков. Скорее всего, высокий уровень конкуренции в главных городах отрасли сказывается на средних зарплатах специалистов, ищущих работу.

## *Приложение (код запросов без форматирования):*

### Задание 2.1

SELECT MAX(age)

FROM hh.candidate

### Задание 2.2

SELECT MIN(age)

FROM hh.candidate

### Задание 2.3

### SELECT age, COUNT(id) AS cnt

FROM hh.candidate

GROUP BY age

ORDER BY age DESC

### Задание 2.4

SELECT COUNT(id)

FROM hh.candidate

WHERE age > 40 AND age < 100

### Задание 3.1

SELECT city.title AS city, COUNT(c.id) AS cnt

FROM hh.candidate AS c

JOIN hh.city AS city ON c.city\_id = city.id

GROUP BY city.title

ORDER BY cnt DESC

### Задание 3.2

SELECT

c.gender,

c.age,

c.desirable\_occupation,

city.title AS city,

c.employment\_type

FROM hh.candidate AS c

JOIN hh.city AS city ON c.city\_id = city.id

WHERE city.title = 'Москва'

AND c.employment\_type LIKE '%проектная работа%'

ORDER BY c.id

### Задание 3.3

SELECT

c.gender,

c.age,

c.desirable\_occupation,

city.title AS city,

c.employment\_type

FROM hh.candidate AS c

JOIN hh.city AS city ON c.city\_id = city.id

WHERE city.title = 'Москва'

AND c.employment\_type LIKE '%проектная работа%'

AND (LOWER(c.desirable\_occupation) LIKE '%разработчик%'

OR LOWER(c.desirable\_occupation) LIKE '%аналитик%'

OR LOWER(c.desirable\_occupation) LIKE '%программист%')

ORDER BY c.id

### Задание 3.4

SELECT c.id, city.title AS city

FROM hh.candidate AS c

JOIN hh.city AS city ON c.city\_id = city.id

WHERE c.current\_occupation = c.desirable\_occupation

ORDER BY city, c.id

### Задание 3.5

SELECT COUNT(id)

FROM hh.candidate

WHERE ((gender = 'F' AND age >= 60)

OR (gender = 'M' AND age >= 65))

AND age < 100

### 

### Задание 4.1

SELECT

c.gender,

c.age,

c.desirable\_occupation,

city.title AS city,

c.employment\_type,

t.title AS timetable\_type

FROM hh.candidate AS c

JOIN hh.candidate\_timetable\_type AS ct ON c.id = ct.candidate\_id

JOIN hh.timetable\_type AS t ON ct.timetable\_id = t.id

JOIN hh.city AS city ON c.city\_id = city.id

WHERE city.title IN ('Новосибирск', 'Омск', 'Томск', 'Тюмень')

AND t.title = 'вахтовый метод'

ORDER BY city, c.id

### Задание 4.2

WITH sp AS

(SELECT c.id, c.desirable\_occupation, c.age

FROM hh.candidate AS c

JOIN hh.city AS city ON c.city\_id = city.id

WHERE city.title = 'Санкт-Петербург'

AND age BETWEEN 16 AND 21)

(SELECT desirable\_occupation, age

FROM sp

ORDER BY age

LIMIT 10)

UNION ALL

(SELECT 'Total', COUNT(id)

FROM sp)

1. Для улучшения читаемости текстового документа - код запросов отформатирован согласно “Руководству по стилю SQL” (<https://www.sqlstyle.guide/ru/>)

   При копировании отформатированного кода в Metabase смещение табуляции ухудшает читаемость кода!

   При необходимости, можно воспользоваться Приложением к проекту, где представлен код запросов SQL без форматирования для копирования в Metabase [↑](#footnote-ref-1)