**PROJECT-2. Подгрузка новых данных. Уточнение анализа**

\*в исследовании в большей степени буду пользоваться термином “соискатели” против термина “кандидаты” (оставляя его в оригинальном тексте задач), примененного постановщиком задач, так как “соискатель” становится “кандидатом” лишь в момент поступления информации о нем заказчику как результат “первичной” выборки.

## **2. Предварительный анализ данных**

### Задание 2.1

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

**Текст запроса:**

SELECT /\*выбор столбцов\*/

'max\_age' Параметр, /\*создаем название параметра 'max\_age'; назначаем алиас "Параметр"\*/

MAX(cand.age) Значение\_параметра /\*ищем максимальное значение столбца age; назначаем алиас "Значение\_параметра"\*/

FROM hh.candidate cand /\*из таблицы hh.candidate, назначаем алиас "cand"\*/

### 

**Вывод:** значение возраста 100 лет вероятнее всего является ошибкой в данных, поэтому в дальнейшем анализе это значение следует исключить.

### Задание 2.2

Теперь давайте рассчитаем минимальный возраст (min\_age) кандидата в таблице.

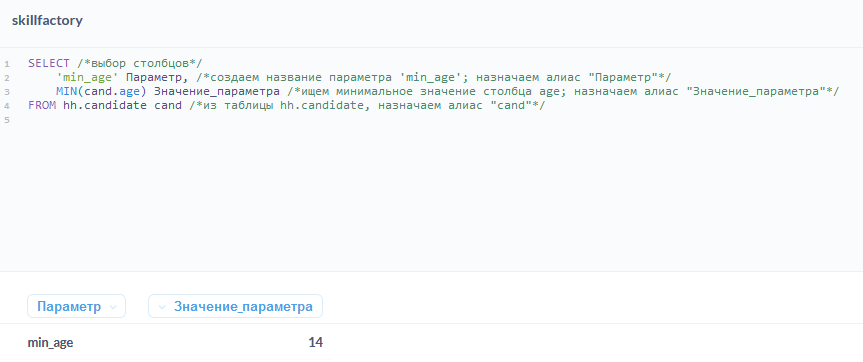
**Текст запроса:**

SELECT /\*выбор столбцов\*/

'min\_age' Параметр, /\*создаем название параметра 'min\_age'; назначаем алиас "Параметр"\*/

MIN(cand.age) Значение\_параметра /\*ищем минимальное значение столбца age; назначаем алиас "Значение\_параметра"\*/

FROM hh.candidate cand /\*из таблицы hh.candidate, назначаем алиас "cand"\*/



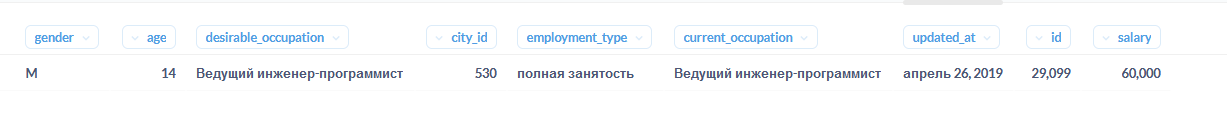
**Вывод:** значение возраста 14 лет в принципе может выступать в качестве минимального значения возраста, но, думаю, необходима дополнительная проверка этих данных следующим запросом.

**Текст запроса:**

SELECT \*

FROM hh.candidate cand /\*из таблицы hh.candidate, назначаем алиас "cand"\*/

WHERE cand.age = 14 /\*фильтруем по значению age, равному 14\*/



В полученном результате запроса мы видим, что соискатель с возрастом 14 лет в настоящее время занимает должность Ведущего инженера-программиста (желаемая должность такая же), что маловероятно для человека данного возраста. Другими словами, возраст в 14 лет весьма вероятно также является ошибкой в данных. Однако в последующем анализе исключать эти данные не будем, так как это не предусмотрено условиями задач.

### Задание 2.3

Попробуем «почистить» данные. Напишите запрос, который позволит посчитать для каждого возраста (age) сколько (cnt) человек этого возраста у нас есть.

Отсортируйте результат по возрасту в обратном порядке.

**Текст запроса:**

SELECT

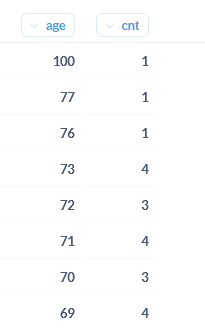
cand.age, /\*выбор столбца age из таблицы candidate с назначенным алиасом “cand”\*/

COUNT(cand.age) cnt /\*подсчёт непустых строк в столбце age; назначаем алиас "cnt"\*/

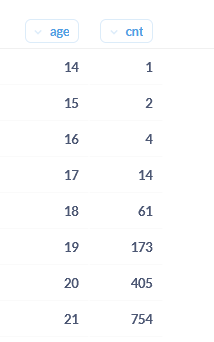
FROM hh.candidate cand /\*из таблицы hh.candidate, назначаем алиас "cand"\*/

GROUP BY 1 /\*группировка по столбцу age\*/

ORDER BY 1 DESC /\*сортировка по столбцу cnt в обратном порядке\*/



Выведем аналогичным образом количество человек для каждого возраста, но по возрастанию возраста:



**Вывод:** 100 среди значений возраста встречается только один раз, в то время как остальные данные похожи на правду. Значит, при необходимости анализировать возраст в дальнейших подсчётах лучше отбрасывать это число - соглашусь с комментарием, данным в данном учебном модуле.

Не менее “подозрительным” остается возраст 14 лет, встречающийся также только один раз. Но по причинам, указанным в задании 2.2, в последующем анализе исключать данные соискателя возрастом 14 лет мы не будем.

### 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.*

**Текст запроса (вариант 1):**

SELECT

'Число кандидатов, старше 40 лет' Параметр, /\*создаем название параметра 'Число кандидатов, старше 40 лет'; назначаем алиас "Параметр"\*/

COUNT(cand.id) Значение\_параметра /\*подсчёт непустых строк в столбце id\*; назначаем алиас "Значение\_параметра"\*/

FROM hh.candidate cand

WHERE cand.age > 40 AND cand.age < 100 /\*фильтруем по значениям age в диапазоне от 40 (не включая 40 в диапазон) до 100 (не включая 100 в диапазон) \*/

**Текст запроса (вариант 2):**

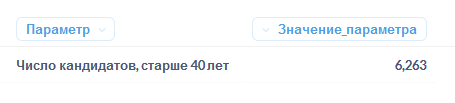
SELECT

'Число кандидатов, старше 40 лет' Параметр, /\*создаем название параметра 'Число кандидатов, старше 40 лет'; назначаем алиас "Параметр"\*/

COUNT(cand.id) Значение\_параметра /\*подсчёт непустых строк в столбце id\*; назначаем алиас "Значение\_параметра"\*/

FROM hh.candidate cand

WHERE cand.age BETWEEN 41 AND 99 /\*фильтруем по значениям age в диапазоне от 41 (включительно) до 99 (включительно)\*/



Найдем полное число кандидатов:

**Текст запроса:**

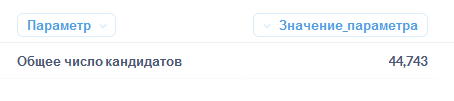
SELECT

'Общее число кандидатов' Параметр, /\*создаем название параметра 'Общее число кандидатов'; назначаем алиас "Параметр"\*/

COUNT(cand.id) Значение\_параметра /\*подсчёт непустых строк в столбце id\*; назначаем алиас "Значение\_параметра"\*/

FROM hh.candidate cand /\*из таблицы hh.candidate, назначаем алиас "cand"\*/

WHERE cand.age < 100 /\*фильтруем по значениям age до 100 (не включая 100 в диапазон) - исключаем значение 100 \*/

****

**Вывод:** количество соискателей старше среднего возраста занятых в экономике России (в расчетах 40 лет) составляет всего 14% от общего числа соискателей.

## **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.

Группировку таблицы необходимо провести по столбцу title, результат отсортируйте по количеству в обратном порядке.

**Текст запроса:**

SELECT

city.title city, /\*выбираем из таблицы city столбец title, назначаем алиас "city"\*/

COUNT(cand.id) cnt /\*подсчёт непустых строк в столбце id таблицы с алиасом “cand”, назначаем алиас "cnt"\*/

FROM hh.city /\*таблица с данными по городам\*/

JOIN hh.candidate cand ON cand.city\_id = city.id /\*оператор соединения таблиц: добавляем данные из таблицы candidate с алиасом “cand”; условие: city\_id таблицы candidate равен id таблицы city\*/

GROUP BY 1 /\*группировка по столбцу title\*/

ORDER BY 2 DESC /\*сортировка по столбцу cnt в обратном порядке\*/



**Вывод:** у Москвы значительный отрыв от остальных городов по числу соискателей, разместивших свои резюме.

### 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.

**Текст запроса:**

SELECT

cand.gender, /\*выбираем из таблицы с алиасом “cand” столбец gender\*/

cand.age age, /\*выбираем из таблицы с алиасом “cand” столбец age\*/

cand.desirable\_occupation, /\*выбираем из таблицы с алиасом “cand” столбец desirable\_occupation\*/

city.title city, /\*выбираем из таблицы city столбец title, назначаем алиас "city"\*/

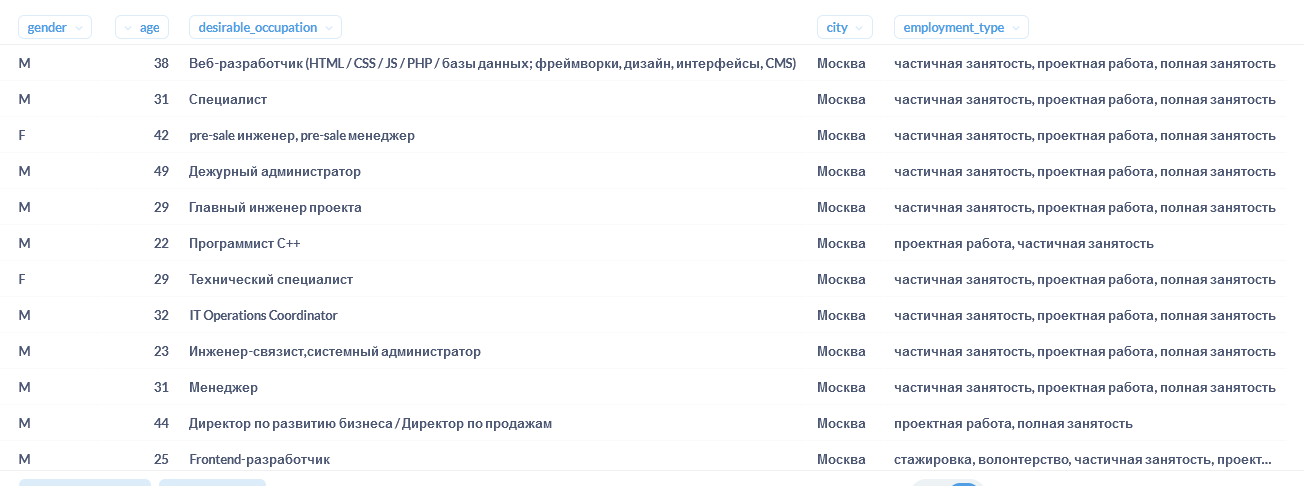
cand.employment\_type /\*выбираем из таблицы с алиасом “cand” столбец employment\_type\*/

FROM hh.candidate cand

JOIN hh.city ON cand.city\_id = city.id /\*оператор соединения таблиц: добавляем данные из таблицы city; условие: city\_id таблицы candidate равен id таблицы city\*/

WHERE city.title = 'Москва' AND LOWER(cand.employment\_type) LIKE '%проект%' /\*фильтруем по значениям столбца title таблицы city, равным 'Москва', и (логическое “И”) значениям столбца employment\_type таблицы candidate, содержащим в себе хотя бы часть слова “проект” независимо от регистра букв \*/

ORDER BY cand.id /\*сортировка по столбцу id таблицы candidate\*/



**Вывод:** большинство соискателей в г. Москве, готовые к “проектной работе”, желали бы продолжить или связать свою трудовую деятельность с IT-отраслью: программирование, аналитика, разработка, …

### 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.

**Текст запроса:**

SELECT

cand.gender gender, /\*выбираем из таблицы с алиасом “cand” столбец gender\*/

cand.age age, /\*выбираем из таблицы с алиасом “cand” столбец age\*/

cand.desirable\_occupation desirable\_occupation, /\*выбираем из таблицы с алиасом “cand” столбец desirable\_occupation\*/

city.title city, /\*выбираем из таблицы city столбец title, назначаем алиас "city"\*/

cand.employment\_type employment\_type /\*выбираем из таблицы city столбец title, назначаем алиас "city"\*/

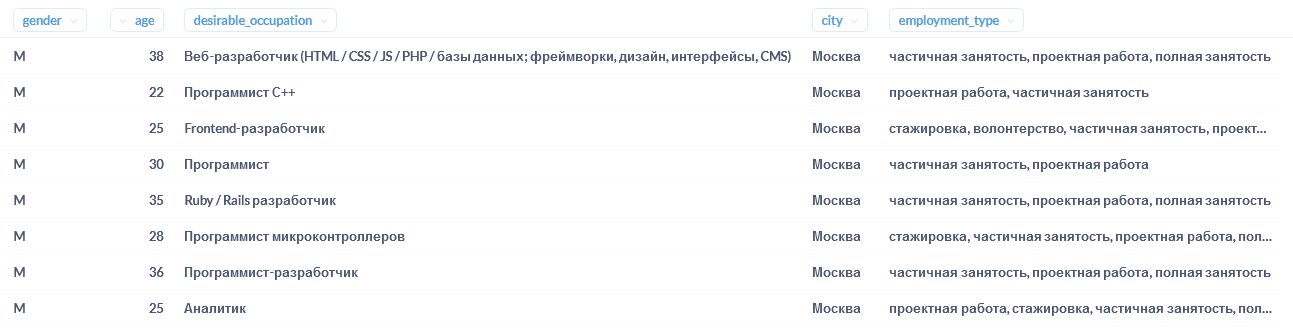
FROM hh.candidate cand

JOIN hh.city ON cand.city\_id = city.id /\*оператор соединения таблиц: добавляем данные из таблицы city; условие: city\_id таблицы candidate равен id таблицы city\*/

WHERE city.title = 'Москва' AND LOWER(cand.employment\_type) LIKE '%проект%' AND (LOWER(cand.desirable\_occupation) LIKE '%разработчик%' OR

LOWER(cand.desirable\_occupation) LIKE '%аналитик%' OR LOWER(cand.desirable\_occupation) LIKE '%программист%') /\*фильтруем по значениям столбца title таблицы city, равным 'Москва', и (логическое “И”) значениям столбца desirable\_occupation таблицы candidate, содержащим в себе слова “проект” или (логическое “ИЛИ”) “разработчик” или (логическое “ИЛИ”) “аналитик” или (логическое “ИЛИ”) “программист” независимо от регистра букв\*/

ORDER BY cand.id /\*сортировка по столбцу id таблицы candidate\*/



**Вывод:** согласно полученным данным основную массу соискателей на самые популярные *IT*-профессии (разработчик, аналитик, программист) составляют мужчины возрастом до 40 лет.

### Задание 3.4

1/1 point (graded)

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.

**Текст запроса:**

SELECT /\*выбираем нужные столбцы согласно формату выборки и назначаем соответствующие алиасы\*/

cand.id id,

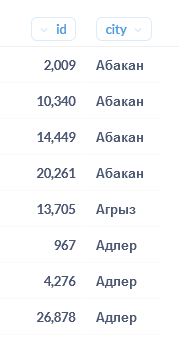
city.title city

FROM hh.candidate cand

JOIN hh.city ON cand.city\_id = city.id /\*оператор соединения таблиц: добавляем данные из таблицы city; условие: city\_id таблицы candidate равен id таблицы city\*/

WHERE cand.current\_occupation = cand.desirable\_occupation /\*фильтруем по условию равенства значений столбцов current\_occupation и (логическое “И”) desirable\_occupation таблицы candidate\*/

ORDER BY 2, 1



**Вывод:** полученная информация пригодна для дальнейших исследований лишь при наличии определенной практической потребности.

### Задание 3.5

Определите количество кандидатов пенсионного возраста.

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

**Текст запроса:**

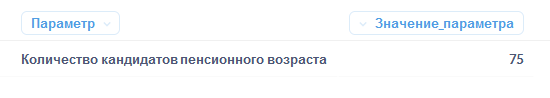
SELECT

'Количество кандидатов пенсионного возраста' Параметр, /\*создаем название параметра 'Количество кандидатов пенсионного возраста'; назначаем алиас "Параметр"\*/

COUNT(cand.id) Значение\_параметра /\*подсчёт непустых строк в столбце id\*; назначаем алиас "Значение\_параметра"\*/

FROM hh.candidate cand

WHERE (cand.gender = 'M' AND cand.age >= 65 AND cand.age < 100) OR (cand.gender = 'F' AND cand.age >= 60 AND cand.age < 100) /\*фильтруем по условиям значения столбца gender, равном “M”, и (логическое “И”) значениям столбца age, лежащих в диапазоне от 65 до 99 лет (включительно), или (логическое “ИЛИ”) по значениям столбца gender, равном “F” и (логическое “И”) значениям столбца age, лежащих в диапазоне от 60 до 99 лет (включительно), таблицы candidate\*/



**Вывод:** среди общего числа (44743) доля соискателей пенсионного возраста составляет 0,2%.

## **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.

**Текст запроса:**

SELECT /\*выбираем нужные столбцы согласно формату выборки и назначаем соответствующие алиасы\*/

cand.gender gender,

cand.age age,

cand.desirable\_occupation desirable\_occupation,

city.title city,

cand.employment\_type employment\_type,

t\_table.title timetable\_type

FROM hh.candidate cand

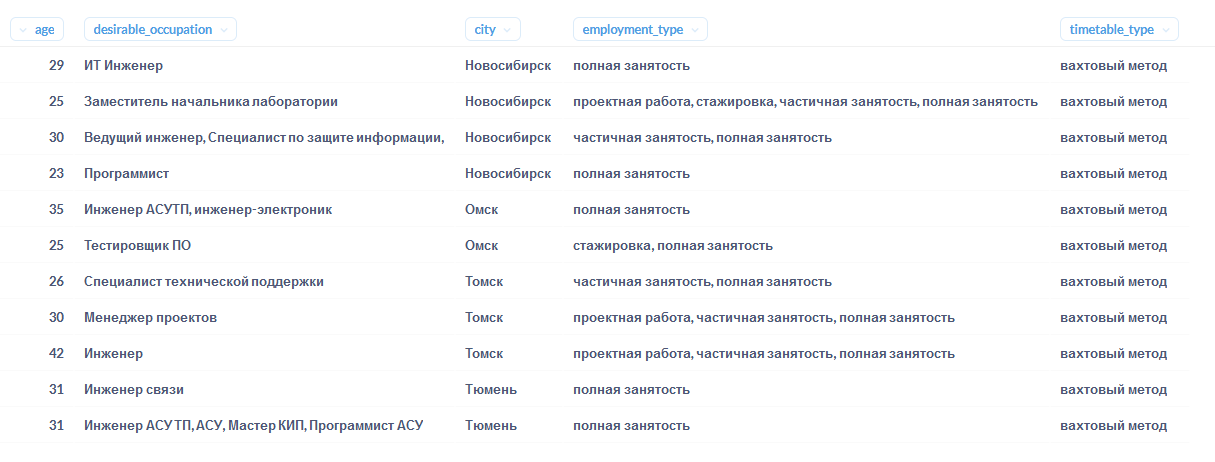
JOIN hh.city ON cand.city\_id = city.id /\*оператор соединения таблиц: добавляем данные из таблицы city; условие: city\_id таблицы candidate равен id таблицы city\*/

JOIN hh.candidate\_timetable\_type ctt ON cand.id = ctt.candidate\_id /\*оператор соединения таблиц: добавляем данные из таблицы candidate\_timetable\_type; условие: candidate\_id таблицы candidate\_timetable\_type равен id таблицы candidate\*/

JOIN hh.timetable\_type t\_table ON t\_table.id = ctt.timetable\_id /\*оператор соединения таблиц: добавляем данные из таблицы timetable\_type; условие: id таблицы timetable\_type равен timetable\_id таблицы candidate\_timetable\_type\*/

WHERE city.title IN ('Новосибирск', 'Омск', 'Томск', 'Тюмень') AND t\_table.title LIKE '%вахтовый метод%' /\*фильтруем по условиям значения столбца title таблицы city, равном “Новосибирск”, “Омск”, “Томск”, “Тюмень” и (логическое “И”) значениям столбца title таблицы timetable\_type, содержащем сочетание символов “вахтовый метод”\*/

ORDER BY city.title, cand.id /\*сортируем результат по городу и номеру кандидата\*/



Выведем количество кандидатов, соответствующих требованиям заказчиков.

**Текст запроса:**

SELECT

'Количество кандидатов, соответствующих требованиям заказчиков' Параметр, /\*создаем название параметра 'Количество кандидатов, соответствующих требованиям заказчиков'; назначаем алиас "Параметр"\*/

COUNT(cand.id) Значение\_параметра /\*подсчёт непустых строк в столбце id; назначаем алиас "Значение\_параметра"\*/

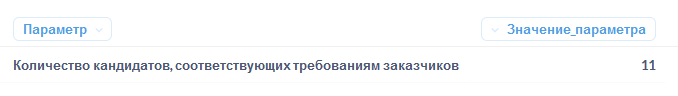
FROM hh.candidate cand

JOIN hh.city ON cand.city\_id = city.id

JOIN hh.candidate\_timetable\_type ctt ON cand.id = ctt.candidate\_id

JOIN hh.timetable\_type t\_table ON t\_table.id = ctt.timetable\_id

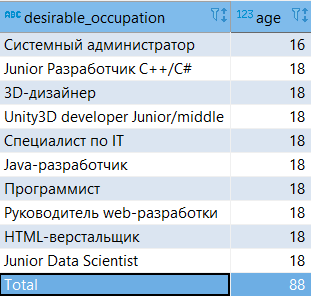
WHERE city.title IN ('Новосибирск', 'Омск', 'Томск', 'Тюмень') AND t\_table.title LIKE '%вахтовый метод%'

****

**Вывод:** количество соискателей из Новосибирска, Омска, Томска и Тюмени, готовых работать вахтовым методом небольшое количество (11). Вероятно, малое число кандидатов не позволит произвести выбор работника с нужной квалификацией и опытом, в связи с чем заказчику рекомендуется расширить географию поиска.

### 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:



**Текст запроса:**

(SELECT /\*выбираем нужные столбцы согласно формату выборки и назначаем соответствующие алиасы\*/

cand.desirable\_occupation desirable\_occupation,

cand.age age

FROM hh.candidate cand

JOIN hh.city ON cand.city\_id = city.id /\*оператор соединения таблиц: добавляем данные из таблицы city; условие: city\_id таблицы candidate равен id таблицы city\*/

WHERE city.title = 'Санкт-Петербург' AND cand.age BETWEEN 16 AND 21 /\*фильтруем по условиям значения столбца title таблицы city, равном “Санкт-Петербург” и (логическое “И”) значениям столбца age таблицы candidate, расположенных в интервале от 16 (включительно) до 21 (включительно)\*/

ORDER BY 2 /\*сортируем результат по возрасту кандидата\*/

LIMIT 10) /\*выводим список из 10 желаемых профессий кандидатов\*/

UNION ALL /\*оператор присоединения\*/

(SELECT

'Total', /\*создаем значение 'Total' в первом столбце присоединяемой таблицы, состоящей из одной строки\*/

COUNT(cand.id) /\*подсчёт непустых строк в столбце id - это значение во втором столбце присоединяемой таблицы\*/

FROM hh.candidate cand

JOIN hh.city ON cand.city\_id = city.id /\*оператор соединения таблиц: добавляем данные из таблицы city; условие: city\_id таблицы candidate равен id таблицы city\*/

WHERE city.title = 'Санкт-Петербург' AND cand.age BETWEEN 16 AND 21) /\*фильтруем по условиям значения столбца title таблицы city, равном “Санкт-Петербург” и (логическое “И”) значениям столбца age таблицы candidate, расположенных в интервале от 16 (включительно) до 21 (включительно) - для подсчета общего количества таких кандидатов\*/



Для дополнительного исследования полученной выборки с помощью дополнительного запроса сделаем вывод о количестве соискателей из г. Санкт-Петербург возрастом от 16 до 21 года на каждой из желаемых профессий (понимая, что часть информации будет упущена по причине неидентичности названия желаемых профессий или - в некоторых случаях - перечисление соискателями нескольких желаемых профессий (требуется дополнительное преобразование данных по признаку “желаемая профессия”).

**Текст запроса:**

(SELECT

cand.desirable\_occupation desirable\_occupation,

COUNT(cand.id) Количество\_соискателей /\*подсчёт непустых строк в столбце id; назначаем алиас "Количество\_соискателей"\*/

FROM hh.candidate cand

JOIN hh.city ON cand.city\_id = city.id

WHERE city.title = 'Санкт-Петербург' AND cand.age BETWEEN 16 AND 21

GROUP BY cand.desirable\_occupation

ORDER BY 2 DESC /\*сортируем результат по количеству кандидатов в обратном порядке\*/

LIMIT 10) /\*выводим список из 10 желаемых профессий с наибольшим числом кандидатов\*/

UNION ALL

(SELECT

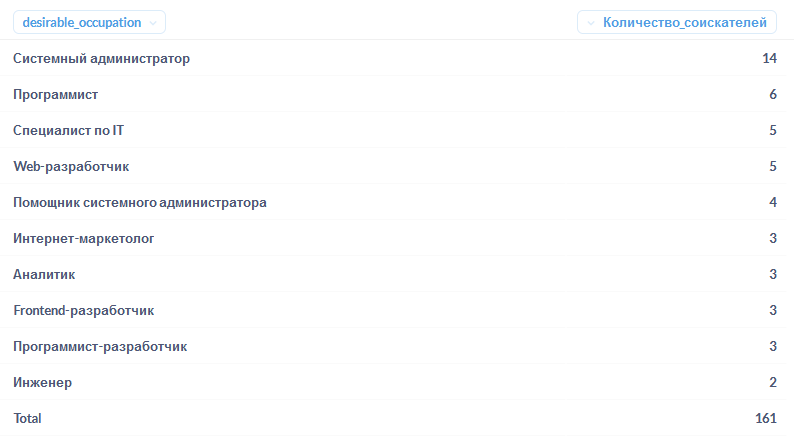
'Total',

COUNT(cand.id)

FROM hh.candidate cand

JOIN hh.city ON cand.city\_id = city.id

WHERE city.title = 'Санкт-Петербург' AND cand.age BETWEEN 16 AND 21)



**Вывод:** таким образом, в выборке согласно требованиям заказчиков из Санкт-Петербурга в качестве желаемой профессии наибольшее число из IT-отрасли, из них превалирует “Системный администратор”, занимая долю в 7% от общего числа кандидатов.

### Общий вывод по анализу данных

Для анализа представлены данные по 44744 соискателям. Из них два кандидата указали возраст, выходящий за возрастные рамки “среднестатистического” работника: 100 лет (принято решение исключить эти данные из дальнейшего анализа) и 14 лет.

При дополнительном анализе было выявлено, что вероятнее всего и указанный возраст 14 лет является ошибкой, однако данные по этому кандидату исключены не были для соответствия полученных ответов.

Дальнейший анализ соискателей в разрезе возраста показал, что количество соискателей старше среднего возраста занятых в экономике России (в расчетах 40 лет) составляет всего 14% от общего числа соискателей, т.е. граждане России старше 40 лет по тем или иным причинам имеют более консервативный взгляд на смену места работы (устраивает место работы, нет веры в возможность смены места работы, …).

Еще меньшую долю составляют соискателей пенсионного возраста (0,2%), что является закономерным. В связи с тем, что соискатели пенсионного возраста обычно более лояльны и к условиям работы, и к работодателю, следует провести дополнительные исследования по желаемым профессиям таких соискателей с целью предоставления полученной выборки соответствующим заказчикам.

При анализе данных в разрезе городов был сделан вывод, что у Москвы значительный отрыв от остальных городов по числу соискателей, разместивших свои резюме (второе место за Санкт-Петербургом). В целом полученный вывод соответствует как картине рынка труда по России, так и демографическим показателям городов (число жителей) Однако по представленным данным нет возможности дать точную оценку полученному выводу, так как нет ясности о каком именно городе представлены данные (о том, где проживает соискатель, или о том, в каком городе он ищет работу). Для этого необходимо уточнение данных о городах в таблице candidate, и, вероятнее всего, разделение данных на два признака: “город проживания”, “город, в котором ищет работу”.

Анализ данных по желаемым профессиям было выявлено, что большое число желаемых профессий составляют профессии из *IT*-отрасли. Проанализировав данные по самым популярным *IT*-профессиям (разработчик, аналитик, программист) мы получили вывод, что основную масса соискателей на эти профессии составляют мужчины возрастом до 40 лет.

Схожая картина и в Москве по соискателям, готовым к “проектной работе”: большинство желали бы продолжить или связать свою трудовую деятельность с IT-отраслью: программирование, аналитика, разработка, …