**[PROJECT-2. Подгрузка новых данных. Уточнение анализа](https://lms.skillfactory.ru/courses/course-v1:SkillFactory+DST-3.0+28FEB2021/course/" \l "block-v1:SkillFactory+DST-3.0+28FEB2021+type@sequential+block@71f6d23f9235497fa89b66e31d2f27f1)**

**1. Введение. Знакомство с датасетом**

 Добавить страницу в мои закладки

✍ Большинство крупных компаний хранят данные в реляционных базах, так как в таком виде они более структурированы и адаптированы для внешних систем, с помощью которых, как правило, данные в базу и заносятся. В системах с такими базами данных имеют значение именно отношения между сущностями. Если вам понадобилось, например, посчитать статистику по заказам, то здесь важны отношения: заказ, клиент, товар и т. д.

→ Мы переделали уже знакомый вам по предыдущему проекту датасет *hh.ru* в реляционный, чтобы вы могли попрактиковаться и в этой области.

Представьте, что вы работаете в некотором дата-центре по трудоустройству, куда стекаются данные с различных платформ для найма. Вам необходимо познакомиться с данными, понять, с какими резюме вы имеете дело, а дальше собрать статистику для различных кадровых агентств и статистических центров.

→ Вы получите данные по возрасту кандидатов, по городам с наиболее активным рынком труда, поймёте специфику найма (где какие вакансии более активны), а также узнаете желаемые позиции для тех или иных кандидатов.

**Проект включает в себя несколько этапов:**

* + знакомство с датасетом;
  + предварительный анализ данных;
  + анализ кандидатов;
  + глобальный анализ показателей.

⚡ В качестве итогового задания проекта вы подготовите **текстовый Google-документ**, в котором по каждой встречающейся вам в этом модуле задаче приведёте ваш ответ в формате:

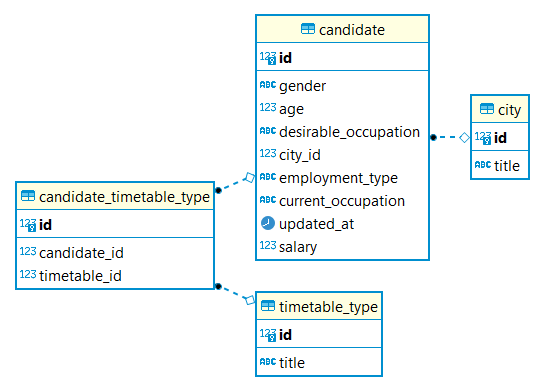
* + номер задания;
  + код с пояснениями;
  + результат запроса (в виде таблицы);
  + общий вывод в конце по результатам анализа рынка труда.

**ЗНАКОМСТВО С ДАТАСЕТОМ**

Напоминаем, что датасет, с которым мы будем работать, уже вам знаком. Мы взяли те же данные *HeadHunter*, только разбили их по таблицам и перенесли в [БД SQL](http://sql.skillfactory.ru:3000/browse/2/schema/hh). Такой процесс называется **нормализацией** (вы можете почитать про неё дополнительно, если вам понадобятся знания по проектированию баз данных).

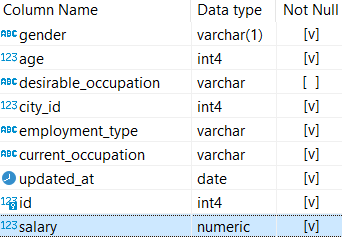
**ПОКАЗАТЬ УЧЁТНЫЕ ДАННЫЕ**

На диаграмме мы видим четыре таблицы:



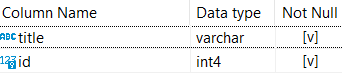
**CANDIDATE**

Таблица хранит в себе общие данные по кандидатам: *id*, пол, возраст, желаемая должность, город, вид занятости, текущая должность, дата обновления записи и зарплата.



**CITY**

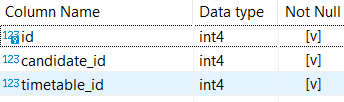
city — таблица-справочник для наших кандидатов — хранит код города и его название.



**CANDIDATE\_TIMETABLE\_TYPE**

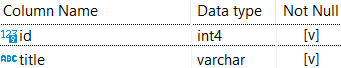
Это дополнительная таблица. Она существует для организации связи *многие-ко-многим*, так как у нас есть много кандидатов и у них может быть несколько подходящих типов рабочего графика.

**Примечание.** Нам стоило бы вынести в дополнительную таблицу и колонку employment\_type из таблицы candidate, так как вид связи здесь одинаковый, но мы решили оставить всё как есть, так как это даст нам дополнительные возможности для постановки интересных задач и проведения изысканий.



**TIMETABLE\_TYPE**

Это таблица-справочник вариантов рабочего графика, подходящего кандидату.



## 2. Предварительный анализ данных

 Добавить страницу в мои закладки

✍ Для начала давайте познакомимся с общей информацией по кандидатам и попытаемся оценить адекватность данных. Начнём с возраста.

### Задание 2.1

1/1 point (graded)

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

SELECT

MAX(cand.age) max\_age

FROM hh.candidate cand

  верно

Loading

Отправить

Correct (1/1 point)

Просмотреть

### Задание 2.2

1/1 point (graded)

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

SELECT

MIN(cand.age) min\_age

FROM hh.candidate cand

  верно

Loading

**Ответ**

Верно:

Какие выводы мы можем сделать? Если 14 лет в качестве минимального значения возраста хоть как-то может претендовать на адекватность, то 100 лет в качестве значения максимального возраста — это явно какая-то ошибка.

Correct (1/1 point)

Просмотреть

### Задание 2.3

1/1 point (graded)

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.

SELECT

cand.age age,

COUNT(cand.age)

FROM hh.candidate cand

GROUP BY 1

ORDER BY 1 DESC

Верно

Test results

**CORRECT**

Отправить

В некоторых задачах доступны следующие действия: сохранение, сброс, показ подсказки или ответа. Соответствующие кнопки расположены рядом с кнопкой «Отправить».

Correct (1/1 point)

Просмотреть

Что в итоге?

### Задание 2.4

1/1 point (graded)

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.*

SELECT

COUNT(cand.id)

FROM hh.candidate cand

WHERE cand.age > 40 AND cand.age < 100

ИЛИ

SELECT

COUNT(cand.id)

FROM hh.candidate cand

WHERE cand.age BETWEEN 41 AND 99

  верно

Loading

Отправить

В некоторых задачах доступны следующие действия: сохранение, сброс, показ подсказки или ответа. Соответствующие кнопки расположены рядом с кнопкой «Отправить».

Correct (1/1 point)

## 3. Глобальный анализ показателей

 Добавить страницу в мои закладки

✍ Теперь нам хотелось бы получить общие показатели по данным.

### Задание 3.1

1/1 point (graded)

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.

SELECT

city.title city,

COUNT(cand.id) cnt

FROM hh.city

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

GROUP BY 1

ORDER BY 2 DESC

Верно

Test results

**CORRECT**[**See full outputSee full output**](https://lms.skillfactory.ru/courses/course-v1:SkillFactory+DST-3.0+28FEB2021/courseware/cb850086b7984f298ab01a4fbe38e417/71f6d23f9235497fa89b66e31d2f27f1/?child=first)

Отправить

В некоторых задачах доступны следующие действия: сохранение, сброс, показ подсказки или ответа. Соответствующие кнопки расположены рядом с кнопкой «Отправить».

Correct (1/1 point)

Просмотреть

### Задание 3.2

1/1 point (graded)

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.

SELECT

cand.gender gender,

cand.age age,

cand.desirable\_occupation desirable\_occupation,

city.title city,

cand.employment\_type employment\_type

FROM hh.candidate cand

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

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

ORDER BY cand.id

Верно

Test results

**CORRECT**[**See full outputSee full output**](https://lms.skillfactory.ru/courses/course-v1:SkillFactory+DST-3.0+28FEB2021/courseware/cb850086b7984f298ab01a4fbe38e417/71f6d23f9235497fa89b66e31d2f27f1/?child=first)

Отправить

В некоторых задачах доступны следующие действия: сохранение, сброс, показ подсказки или ответа. Соответствующие кнопки расположены рядом с кнопкой «Отправить».

Correct (1/1 point)

Просмотреть

### Задание 3.3

1/1 point (graded)

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.

SELECT

cand.gender gender,

cand.age age,

cand.desirable\_occupation desirable\_occupation,

city.title city,

cand.employment\_type employment\_type

FROM hh.candidate cand

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

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%')

ORDER BY cand.id

Верно

Test results

**CORRECT**[**See full outputSee full output**](https://lms.skillfactory.ru/courses/course-v1:SkillFactory+DST-3.0+28FEB2021/courseware/cb850086b7984f298ab01a4fbe38e417/71f6d23f9235497fa89b66e31d2f27f1/?child=first)

Отправить

В некоторых задачах доступны следующие действия: сохранение, сброс, показ подсказки или ответа. Соответствующие кнопки расположены рядом с кнопкой «Отправить».

Correct (1/1 point)

Просмотреть

### 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.

SELECT

cand.id id,

city.title city

FROM hh.candidate cand

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

WHERE cand.current\_occupation = cand.desirable\_occupation

ORDER BY 2, 1

Верно

Test results

**CORRECT**[**See full outputSee full output**](https://lms.skillfactory.ru/courses/course-v1:SkillFactory+DST-3.0+28FEB2021/courseware/cb850086b7984f298ab01a4fbe38e417/71f6d23f9235497fa89b66e31d2f27f1/?child=first)

Отправить

В некоторых задачах доступны следующие действия: сохранение, сброс, показ подсказки или ответа. Соответствующие кнопки расположены рядом с кнопкой «Отправить».

Correct (1/1 point)

Просмотреть

### Задание 3.5

1/1 point (graded)

Определите количество кандидатов пенсионного возраста.  
*Пенсионный возраст для мужчин наступает в 65 лет, для женщин — в 60 лет.*

SELECT

COUNT(cand.id)

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)

  верно

Loading

Correct (1/1 point)

## 4. Анализ кандидатов для заказчиков

 Добавить страницу в мои закладки

### Задание 4.1

1/1 point (graded)

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.

SELECT

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

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%'

ORDER BY city.title, cand.id

Верно

Test results

**CORRECT**[**See full outputSee full output**](https://lms.skillfactory.ru/courses/course-v1:SkillFactory+DST-3.0+28FEB2021/courseware/cb850086b7984f298ab01a4fbe38e417/71f6d23f9235497fa89b66e31d2f27f1/?child=first)

Отправить

В некоторых задачах доступны следующие действия: сохранение, сброс, показ подсказки или ответа. Соответствующие кнопки расположены рядом с кнопкой «Отправить».

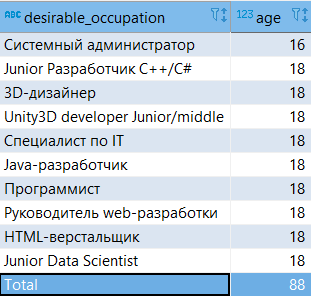
Correct (1/1 point)

Просмотреть

### Задание 4.2

1/1 point (graded)

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:



(SELECT

cand.desirable\_occupation desirable\_occupation,

cand.age age

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

ORDER BY 2

LIMIT 10)

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)

Верно

Test results

**CORRECT**[**See full outputSee full output**](https://lms.skillfactory.ru/courses/course-v1:SkillFactory+DST-3.0+28FEB2021/courseware/cb850086b7984f298ab01a4fbe38e417/71f6d23f9235497fa89b66e31d2f27f1/?child=first)

Отправить

В некоторых задачах доступны следующие действия: сохранение, сброс, показ подсказки или ответа. Соответствующие кнопки расположены рядом с кнопкой «Отправить».

Correct (1/1 point)

Просмотреть

✍ Мы с вами проделали огромную работу: проанализировали данные, выбрали кандидатов для заказчиков. Но расчёты, запросы или данные сами по себе, какими бы ценными они ни были, бесполезны без должных выводов. Для развёрнутого анализа кандидатов у нас уже есть всё необходимое:

* + возраст;
  + статистика по городам;
  + активные вакансии в городах;
  + желаемые должности.

Осталось лишь привести всё это в законченную, логичную форму.

Постарайтесь сделать вывод не только на основе данных — попробуйте также проанализировать, за счёт чего получились именно такие цифры, постарайтесь увидеть тенденции, сделать прогнозы.

Необходимо сделать как **отдельные выводы по результату каждого задания**, так и **один общий вывод по всем кандидатам**, имеющимся в базе данных:

* + Каких кандидатов много?
  + Что можно сказать о рынке труда в различных городах?
  + Какие ещё интересные исследования можно провести?

Вы можете добавить любые свои рассуждения и умозаключения, к которым пришли в процессе работы.

**Пример** того, какие выводы можно сделать, если наша задача — узнать, какой товар приносит максимум прибыли компании.

Запрос даёт нам ответ — максимум прибыли приносит принтер Canon.

**?**Какой вывод мы здесь можем сделать? Например, такой ↓

→ По результатам запроса мы поняли, что больше всего прибыли приносит принтер Canon. В 2012 году он принёс N у.е.

**?**Этого достаточно? Не совсем.

→ Стоит указать также, что данный принтер принёс так много денег не потому, что он очень популярен, а потому, что у него достаточно высокая цена.

→ Для совсем уж отличного вывода можно добавить, к примеру, что его продажи падают из месяца в месяц и, возможно, нам стоит задуматься о закупках более современной модели.

⚡ В качестве итогового задания вставьте в текстовое окно формы ниже **ссылку на текстовый Google-документ (не забудьте открыть Доступ по ссылке!)**, в котором на каждую задачу приведён ответ в формате:

* 1. Номер задания.
  2. Код с пояснениями.

Это важно, так как недостаточно просто высчитать результат — важно сделать оптимальный запрос (для этого мы делаем **код-ревью**), а также оформить его корректно, так, как оформляется документация и спецификации в реальных кейсах.

* 1. Результат запроса (в виде таблицы).
  2. Выводы.
  3. Общий вывод в конце по результатам анализа рынка труда.