

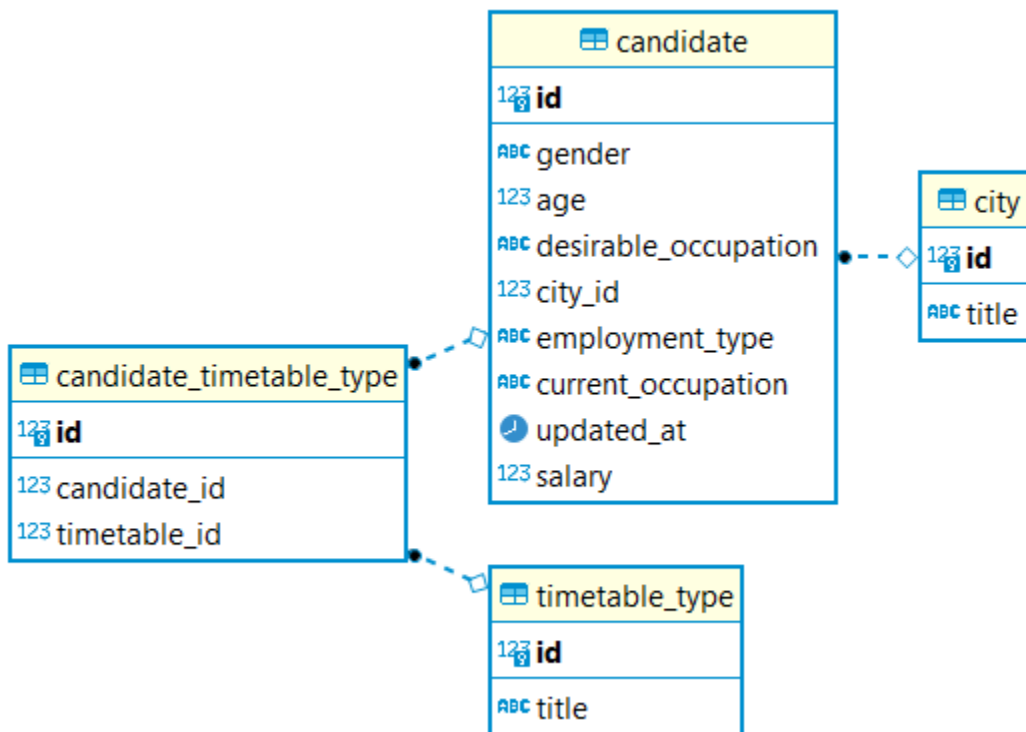
PROJECT-2. Подгрузка новых данных. Уточнение анализа

Этапы проекта:

1. знакомство с датасетом;
2. предварительный анализ данных;
3. анализ кандидатов;
4. глобальный анализ показателей.
5. Выводы










1. Знакомство с датасетом.

Структура данных:





CANDIDATE

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

Column Name	Data type	Not Null
 gender	varchar(1)	[v]
 age	int4	[v]
 desirable_occupation	varchar	[]
 city_id	int4	[v]
 employment_type	varchar	[v]
 current_occupation	varchar	[v]
 updated_at	date	[v]
 id	int4	[v]
 salary	numeric	[v]




CITY

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

Column Name	Data type	Not Null
 title	varchar	[v]
 id	int4	[v]


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.

Column Name	Data type	Not Null
 id	int4	[v]
 candidate_id	int4	[v]
 timetable_id	int4	[v]

TIMETABLE_TYPE

Это таблица-справочник вариантов рабочего графика, подходящего кандидату.

Column Name	Data type	Not Null
 id	int4	[v]
 title	varchar	[v]

2. Предварительный анализ данных.

Задание 2.1

Рассчитайте максимальный возраст (max_age) кандидата в таблице.

-- используем агрегирующую функцию для выборки максимального значения

SELECT

MAX(age)

FROM hh.candidate

+-----+

| MAX(age) |

+-----+

| 100 |

+-----+

Вывод 2.1.

Максимальный возраст кандидата составляет 100 лет, эти данные требуют дополнительной проверки.

Задание 2.2

Теперь давайте рассчитаем минимальный возраст (min_age) кандидата в таблице.

-- используем агрегирующую функцию для выборки минимального значения

SELECT

MIN(age)

FROM hh.candidate

+-----+

| MIN(age) |

+-----+

| 14 |

+-----+

Вывод 2.2.

Минимальный возраст кандидата составляет 14 лет, что в принципе противоречит трудовому законодательству РФ, минимальный возраст заключения трудового договора составляет 15 лет.

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.

-- группировка по возрасту и в разрезе возраста агрегация с подсчетом количества

SELECT

age,

COUNT(*) cnt

FROM hh.candidate

GROUP BY 1

ORDER BY 1 **DESC**

+-----+-----+

| age | cnt |

+-----+-----+

| 100 | 1 |

| 77 | 1 |

| 76 | 1 |

| 73 | 4 |

| 72 | 3 |

.....

| 21 | 754 |

| 20 | 405 |

| 19 | 173 |

| 18 | 61 |

| 17 | 14 |

| 16 | 4 |

| 15 | 2 |

| 14 | 1 |

+-----+-----+

Вывод 2.3.

Кандидат с возрастом 100 лет встречается только 1 раз, считаем это ошибочными данными и исключаем эту запись из дальнейшего анализа.

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.*

-- подсчет количества записей в выборке с условием

SELECT

 COUNT(*)

FROM hh.candidate

WHERE

 age > 40 AND age < 100

+-----+

| COUNT(*) |

+-----+

| 6263 |

+-----+

Вывод 2.4.

Найдено количество кандидатов в возрасте 40+ с исключением ложного кандидата в 100 лет. Чтобы сделать хоть какой то вывод, сделаем дополнительный запрос, чтобы посмотреть общее число кандидатов.

-- подсчет количества записей в выборке с условием

SELECT

 COUNT(*)

FROM hh.candidate

WHERE

 age < 100

+-----+

```
| COUNT(*) |  
+-----+  
| 44743 |  
+-----+
```

Теперь можно сказать, что кандидаты в возрасте 40+ составляют 14% от общего числа.

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.

```
-- группировка по городу и в разрезе города подсчет количества  
-- таблицы кандидатов и города связаны по ключу id города  
-- выборка без условия с последующей сортировкой
```

```
SELECT  
    city.title as city,  
    COUNT(candidate.id) as cnt  
FROM hh.city  
JOIN hh.candidate  
    ON candidate.city_id = city.id  
GROUP BY 1  
ORDER BY 2 DESC;
```

```
----  
+-----+-----+  
| city                                | cnt  |  
+-----+-----+  
| Москва                             | 16621 |  
| Санкт-Петербург                    | 4937  |  
| Краснодар                           | 1066  |
```

Новосибирск	958
Казань	872
...	
Ганцевичи	1
Ардатов	1
Урдома	1
Пыть-Ях	1
Адыгейск	1
Ессентукская	1
Пангоды	1
Хмельницкий	1

+-----+-----+

984 rows in set (0.671 sec)

Вывод 3.1:

В Москве значительно большее количество кандидатов.

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.

-- таблицы городов и кандидатов связаны по ключу `id` города

-- выборка с условием и последующей сортировкой

SELECT

`candidate.gender,`
 `candidate.age,`
 `candidate.desirable_occupation,`
 `city.title as city,`
 `candidate.employment_type`

FROM `hh.candidate`

JOIN `hh.city`

```

    ON candidate.city_id = city.id
WHERE city.title = 'Москва' AND LOWER(candidate.employment_type) LIKE '%проект%'
ORDER BY candidate.id;

```

```

----

| gender | age | desirable_occupation
| city   | employment_type
-----
| М      | 38 | Веб-разработчик (HTML / CSS / JS / PHP / базы данных; фреймворки,
дизайн, интерфейсы, CMS) | Москва | частичная занятость, проектная
работа, полная занятость |
| М      | 31 | Специалист
| Москва | частичная занятость, проектная работа, полная занятость
...
| М      | 26 | Начальник отдела
| Москва | стажировка, волонтерство, частичная занятость, проектная работа, полная
занятость |
| М      | 41 | Программист микроконтроллеров, Altium Designer
| Москва | проектная работа, частичная занятость, полная занятость
-----
2949 rows in set (0.044 sec)

```

Вывод 3.2

Взяв данные об общем числе кандидатов из Москвы мы можем сказать, что всего 17.7% кандидатов готовы к проектной 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.

-- таблицы городов и кандидатов связаны по ключу id города
-- выборка по трем условиям с последующей сортировкой

```
SELECT
    candidate.gender,
    candidate.age,
    candidate.desirable_occupation,
    city.title city,
    candidate.employment_type
FROM hh.candidate
JOIN hh.city
    ON candidate.city_id = city.id
WHERE
    city.title = 'Москва' AND
    LOWER(candidate.employment_type) LIKE '%проект%' AND
    (
        LOWER(candidate.desirable_occupation) LIKE '%разработчик%' OR
        LOWER(candidate.desirable_occupation) LIKE '%аналитик%' OR
        LOWER(candidate.desirable_occupation) LIKE '%программист%'
    )
ORDER BY candidate.id;
```

gender	age	desirable_occupation
city	employment_type	

М	38	Веб-разработчик (HTML / CSS / JS / PHP / базы данных; фреймворки, дизайн, интерфейсы, CMS)	Москва	частичная занятость, проектная работа, полная занятость
М	22	Программист C++	Москва	проектная работа, частичная занятость
М	25	Frontend-разработчик	Москва	стажировка, волонтерство, частичная занятость, проектная работа, полная занятость
М	30	Программист	Москва	частичная занятость, проектная работа

```

| М      | 35 | Ruby / Rails разработчик
| Москва | частичная занятость, проектная работа, полная занятость
|
.....
| М      | 30 | Инженер-программист
| Москва | стажировка, частичная занятость, проектная работа, полная занятость
|
| М      | 22 | Программист .NET junior
| Москва | частичная занятость, проектная работа, полная занятость
|
| М      | 41 | Программист микроконтроллеров, Altium Designer
| Москва | проектная работа, частичная занятость, полная занятость
|

```

777 rows in set (0.047 sec)

Вывод 3.3.

Сложно сделать какой то четкий вывод, глядя на простыню из 777 записей, но предварительно скажем так, что в самых популярных ит профессиях стремятся работать мужчины в возрасте 35-45 лет.

Задание 3.4

1 point possible (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.

-- таблицы городов и кандидатов связаны по ключу `id` города

-- выборка с условием с последующей сортировкой

SELECT

```

    candidate.id,
    city.title

```

```

FROM hh.candidate
JOIN hh.city
    ON candidate.city_id = city.id
WHERE
    candidate.current_occupation = candidate.desirable_occupation
ORDER BY 2, 1;

```

id	title
2009	Абакан
10340	Абакан
14449	Абакан
20261	Абакан
13705	Агрыз
967	Адлер
37880	Ярославль
37886	Ярославль
39508	Ярославль
40481	Ярославль
41970	Ярославль
1462	Ясногорск
9701	Яхрома
32063	Яшалта

5373 rows in set (0.223 sec)

Вывод 3.4

Как минимум 5373 кандидата не собираются менять область своей работы.

Задание 3.5

Определите количество кандидатов пенсионного возраста.

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

```
-- подсчет количества записей удовлетворяющих одному из условий
SELECT
    COUNT(candidate.id) as cnt
FROM hh.candidate
WHERE
    (candidate.gender = 'M' AND candidate.age >= 65 AND candidate.age < 100) OR
    (candidate.gender = 'F' AND candidate.age >= 60 AND candidate.age < 100);
```

```
----
+-----+
| cnt |
+-----+
| 75 |
+-----+
1 row in set (0.019 sec)
```

Вывод 3.5

В выборке присутствует 75 кандидатов пенсионного возраста, что составляет 0.2% от общего числа в выборке.

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.

```
-- таблицы городов и кандидатов связаны по ключу id города
-- для реализации связи многие ко многим варианты занятости
-- связаны с кандидатами через промежуточную таблицу
-- выборка с двумя условиями с последующей сортировкой
SELECT
```

```

        candidate.gender,
        candidate.age,
        candidate.desirable_occupation,
        city.title,
        candidate.employment_type,
        timetable_type.title timetable_type
FROM hh.candidate
JOIN hh.city
    ON candidate.city_id = city.id
JOIN hh.candidate_timetable_type
    ON candidate.id = candidate_timetable_type.candidate_id
JOIN hh.timetable_type
    ON timetable_type.id = candidate_timetable_type.timetable_id
WHERE
    city.title IN ('Новосибирск', 'Омск', 'Томск', 'Тюмень') AND
    timetable_type.title LIKE '%вахтовый метод%'
ORDER BY
    city.title, candidate.id;

```

gender	age	desirable_occupation	title
employment_type			
timetable_type			

M	29	ИТ Инженер	Новосибирск
полная занятость			вахтовый
метод			
M	25	Заместитель начальника лаборатории	Новосибирск
проектная работа, стажировка, частичная занятость, полная занятость			вахтовый
метод			
M	30	Ведущий инженер, Специалист по защите информации,	Новосибирск
частичная занятость, полная занятость			вахтовый
метод			
M	23	Программист	Новосибирск
полная занятость			вахтовый
метод			

М	35	Инженер АСУТП, инженер-электроник	Омск	
полная занятость			вахтовый	
метод				
М	25	Тестировщик ПО	Омск	
стажировка, полная занятость			вахтовый	
метод				
М	26	Специалист технической поддержки	Томск	
частичная занятость, полная занятость			вахтовый	
метод				
М	30	Менеджер проектов	Томск	
проектная работа, частичная занятость, полная занятость			вахтовый	
метод				
М	42	Инженер	Томск	
проектная работа, частичная занятость, полная занятость			вахтовый	
метод				
М	31	Инженер связи	Тюмень	
полная занятость			вахтовый	
метод				
М	31	Инженер АСУ ТП, АСУ, Мастер КИП, Программист АСУ	Тюмень	
полная занятость			вахтовый	
метод				

11 rows in set (1.258 sec)

Вывод 4.1

Число кандидатов из основных городов сибери, готовых работать вахтовым методом всего 11 человек, будущему работодателю не стоит рассматривать данный регион для поиска кандидатов на вахту.

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:

desirable_occupation	age
Системный администратор	16
Junior Разработчик C++/C#	18
3D-дизайнер	18
Unity3D developer Junior/middle	18
Специалист по IT	18
Java-разработчик	18
Программист	18
Руководитель web-разработки	18
HTML-верстальщик	18
Junior Data Scientist	18
Total	88

-- объединение двух запросов
-- первый запрос делает выборку с двумя условиями
-- сортирует результат и ограничивает вывод 10-ю записями
-- второй запрос и с теми же условиями делает подсчет
-- общего количества записей, удовлетворяющих условию
-- таблицы городов и кандидатов связаны по ключу id города

```
(
SELECT
    candidate.desirable_occupation,
    candidate.age
FROM hh.candidate
JOIN hh.city
    ON candidate.city_id = city.id
WHERE
    city.title = 'Санкт-Петербург' AND
    candidate.age BETWEEN 16 AND 21
ORDER BY 2
LIMIT 10
)
```

UNION ALL

```
(
SELECT
```

```

        'Total',
        COUNT(candidate.id)
FROM hh.candidate
JOIN hh.city
    ON candidate.city_id = city.id
WHERE
    city.title = 'Санкт-Петербург' AND
    candidate.age BETWEEN 16 AND 21
);

```

```

----
+-----+-----+
| desirable_occupation | age |
+-----+-----+
| Системный администратор | 16 |
| HTML-верстальщик | 18 |
| Junior Data Scientist | 18 |
| Специалист по IT | 18 |
| Модератор | 18 |
| Системный администратор | 18 |
| Unity3D developer Junior/middle | 18 |
| 3D-дизайнер | 18 |
| Java-разработчик | 18 |
| Программист | 18 |
| Total | 161 |
+-----+-----+
11 rows in set (0.031 sec)

```

Вывод 4.2

Можно сделать вывод, что молодые кандидаты ищут работу с начальных позиций.

5. Выводы

Был поверхностно проанализирован набор с данными 44 744 кандидатов сайта hh.ru.

Один кандидат был исключен из анализа на т.к. был указан не правдоподобный возраст. Еще один кандидат вызывал сомнения, но все же был оставлен.

Анализ возраста показал, что количество кандидатов в возрасте 40+ лет составляет всего 14% от общего числа соискателей, т.е. граждане России старше 40 лет стараются придерживаться того места работы, которое уже имеют.

Кандидаты пенсионного возраста составляют всего 0,2%. Люди пенсионного возраста наиболее лояльны к условиям работы и работодателю. Возможно стоит провести более детальное исследование среди специальностей, на которые идут пенсионеры.

Москва показала себя как город с максимальным числом соискателей, что в принципе соответствует демографическим показателям городов.

При анализе самых популярных ИТ-профессий был сделан вывод, что на эти места претендуют в основном мужчины в возрасте 35-45 лет.

При анализе специальностей молодых соискателей Санкт-Петербурга был сделан вывод, что они предпочитают начинать свою карьеру с начальных ИТ-специальностей.

При анализе кандидатов, готовых к работе вахтовым методом, проживающих в крупных городах Сибири, был сделан вывод, что они не сильно готовы к такому варианту работы. Т.е. работодателю не сильно стоит рассчитывать на кандидатов из этого региона.

P.s. в связи с не очень стабильной работой сервиса metabase (как то не сильно мне везло) анализ проводился на локальной копии БД.