

PROJECT-2. Подгрузка новых данных. Уточнение анализа.Отчёт по проекту.**Задание 2.1**

Рассчитайте максимальный возраст (max_age) кандидата в таблице.

Текст запроса:

```
SELECT
    MAX(c.age) max_age
FROM hh.candidate c
```

Результат:

max_age
100

Выводы:

Результат запроса говорит о том, что скорее всего, это ошибочные данные и их нужно будет учитывать/отфильтровывать при дальнейшей работе с этим набором данных.

Задание 2.2

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

Текст запроса:

```
SELECT
    MIN(c.age) min_age
FROM hh.candidate c
```

Результат:

min_age
14

Выводы:

Результат данного запроса говорит о том, что возможно это тоже ошибочные данные. И если посмотреть полностью данные кандидата, то он претендует на должность “Ведущий инженер-программист”, что в 14 лет явно не соответствует действительности. Даже в случае если кандидат является вундеркиндом, вряд ли он находится в активном поиске работы и тем более через интернет ресурсы. В данном случае 100-летний Frontend-разработчик выглядит более правдоподобно и ретро стиль в оформлении это иногда модно.

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
    c.age age,
    COUNT(age) cnt
FROM hh.candidate c
GROUP BY age
ORDER BY age DESC
```

Результат:

age	cnt
100	1
77	1
76	1
73	4
72	3
71	4
70	3

Выводы:

Наиболее активный возраст в поиске работы по нашим данным это с 24 до 33 лет, что вполне объяснимо и наиболее правдоподобно.

Кандидаты с возрастом до 19-20 лет это редкое исключение, так же как и после 60.

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:

```
SELECT
    COUNT(age) cnt
FROM hh.candidate c
WHERE c.age > 40 AND c.age != 100 -- отсекаем долгожителей
```

Результат:

cnt
6,263

Выводы:

Исходя из того, что в наших данных всего 44744 записи, то результат запроса слегка не совпадает с данными Росстата. 6263 это всего около 14 процентов и средний возраст кандидатов из нашей базы значительно ниже 40 лет. А точнее это 32.2 года. Это говорит возможно о том, что после 40 лет люди значительно реже меняют своё место работы или пользуются другими каналами поиска вакансий.

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
    ci.title city,
    COUNT(c.city_id) cnt
FROM hh.candidate c
    JOIN hh.city ci ON ci.id = c.city_id
GROUP BY city
ORDER BY cnt DESC
```

Результат:

city	cnt
Москва	16,622
Санкт-Петербург	4,937
Краснодар	1,066
Новосибирск	958
Казань	872
Екатеринбург	734
Самара	703
Ростов-на-Дону	607

Выводы:

Результаты данного запроса вполне объяснимы и должны совпадать с реальностью. С большим отрывом идут столицы, далее города миллионники и потом уже остальные. Интересно, что количество вакансий Санкт-Петербурга примерно пропорционально численности населения в 5 млн., а у Москвы значительно выше. Возможно это связано с большим количеством трудовых мигрантов в столице.

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
    c.gender gender,
    c.age age,
    c.desirable_occupation desirable_occupation,
    cy.title city,
    c.employment_type employment_type
FROM hh.candidate c
    JOIN hh.city cy ON cy.id = c.city_id
-- отбираем Москву и все варианты проектной работы
WHERE (cy.title LIKE 'Москва') AND (c.employment_type LIKE '%проектная
работа%')
ORDER BY c.id
```

Результат:

gender	age	desirable_occupation	city	employment_type
M	38	Веб-разработчик (HTML / CSS / JS / PHP / базы данных; фреймворки, дизайн, интерфейсы, CMS)	Москва	частичная занятость, проектная работа, полная занятость
M	31	Специалист	Москва	частичная занятость, проектная работа, полная занятость
F	42	pre-sale инженер, pre-sale менеджер	Москва	частичная занятость, проектная работа, полная занятость
M	49	Дежурный администратор	Москва	частичная занятость, проектная работа, полная занятость
M	29	Главный инженер проекта	Москва	частичная занятость, проектная работа, полная занятость
M	22	Программист C++	Москва	проектная работа, частичная занятость
F	29	Технический специалист	Москва	частичная занятость, проектная работа, полная занятость
M	32	IT Operations Coordinator	Москва	частичная занятость, проектная работа, полная занятость
M	23	Инженер-связист, системный администратор	Москва	частичная занятость, проектная работа, полная занятость

Выводы:

Всего кандидатов в запросе 2950, это почти 18 процентов от общего количества кандидатов г. Москвы, т.е. не так много людей претендуют на проектную работу, менее 1/5. Максимальное количество отобранных кандидатов претендуют на должность системного администратора - 145 человек.

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
    c.gender gender,
    c.age age,
    c.desirable_occupation desirable_occupation,
    cy.title city,
    c.employment_type employment_type
FROM hh.candidate c
JOIN hh.city cy ON c.city_id = cy.id
-- отбираем Москву и все варианты проектной работы
WHERE (cy.title = 'Москва')AND(c.employment_type LIKE '%проектная работа%')
-- также все варианты разработчиков, аналитиков и программистов
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:

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	28	Программист микроконтроллеров	Москва	стажировка, частичная занятость, проектная работа, полная занятость
М	36	Программист-разработчик	Москва	частичная занятость, проектная работа, полная занятость
М	25	Аналитик	Москва	проектная работа, стажировка, частичная занятость, полная занятость

Выводы:

Сделав эту выборку мы получили 778 кандидатов, это всё равно много для каких либо индивидуальных действий (собеседований/чтений резюме и т.п.)

Вообще то, по результатам запроса -

```

SELECT
    c.desirable_occupation desirable_occupation,
    count(c.id) cnt
FROM hh.candidate c
GROUP BY desirable_occupation
ORDER BY cnt DESC

```

мы получаем самую популярную на hh.ru IT-профессию и это системный администратор.

А если чуть изменить запрос -

```

SELECT
    count(c.id)
FROM hh.candidate c
-- отбираем все варианты системных администраторов
WHERE (lower(c.desirable_occupation) LIKE '%системный администратор%')

```

тогда цифры будут такими 5285 кандидатов, в то время как разработчиков 2366 и аналитиков 2206.

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.

Текст запроса:

```

SELECT
    c.id id,
    cy.title city
FROM hh.candidate c
    JOIN hh.city cy ON c.city_id = cy.id
WHERE c.current_occupation = c.desirable_occupation
ORDER BY city, id

```

Результат:

id	city
2,009	Абакан
10,340	Абакан
14,449	Абакан
20,261	Абакан
13,705	Агрыз
967	Адлер
4,276	Адлер
26,878	Адлер

Выводы:

Кандидатов у которых занимаемая должность совпадает с желаемой - 5104, это около 11,4% от общего количества. Это говорит о том, что большинство хотят при смене места работы перейти на другую должность. И после просмотра пару десятков записей в таблице hh.candidate видно, что большинство претендентов желает поступить на более высокую должность в сравнении с занимаемой на данный момент, что является нормальным для большинства людей.

Задание 3.5

Определите количество кандидатов пенсионного возраста.

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

Текст запроса:

```
SELECT
    COUNT(*)
FROM hh.candidate c
WHERE -- Пенсионный возраст мужчины 65 лет, женщины 60 лет
      ((c.gender = 'M' AND c.age > 64) OR (c.gender = 'F' AND c.age > 59))
      AND(c.age < 100) -- отсекаем долгожителей
```

Результат:

Количество
75

Выводы:

75 кандидатов пенсионного возраста это 0,16% от общего количества имеющихся в базе резюме. Это говорит о том, что люди пенсионного возраста очень редко пользуются услугами hh.ru, возможно они просто не работают или пользуются другими источниками поиска работы.

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
    c.gender gender,
    c.age age,
    c.desirable_occupation desirable_occupation,
    cy.title city,
    c.employment_type employment_type,
    tt.title timetable_type
FROM hh.candidate c
    JOIN hh.city cy ON c.city_id = cy.id
    JOIN hh.candidate_timetable_type ct ON c.id = ct.candidate_id
    JOIN hh.timetable_type tt ON ct.timetable_id = tt.id
WHERE -- отбираем нужные города
    cy.title in ('Новосибирск','Омск','Томск','Тюмень')
    AND tt.title = 'вахтовый метод' -- выделяем вахтовый метод
ORDER BY city, c.id
```

Результат:

gender	age	desirable_occupation	city	employment_type	timetable_type
М	29	ИТ Инженер	Новосибирск	полная занятость	вахтовый метод
М	25	Заместитель начальника лаборатории	Новосибирск	проектная работа, стажировка, частичная занятость, полная занятость	вахтовый метод
М	30	Ведущий инженер, Специалист по защите информации,	Новосибирск	частичная занятость, полная занятость	вахтовый метод
М	23	Программист	Новосибирск	полная занятость	вахтовый метод
М	35	Инженер АСУТП, инженер-электроник	Омск	полная занятость	вахтовый метод
М	25	Тестирующий 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

Выводы:

Получили всего 11 кандидатов. Даже изменив чуть запрос в части выбора графика работы на *AND lower(tt.title) LIKE '%вахтовый метод%'* получаем те же 11 человек, при том, что всего кандидатов из этих городов 1295.

Всего 0,85% претендентов готовы к работе вахтовым методом, понятно, что для этого нужна и привычка и определенный склад характера.

Исходя из результата, возможно у работодателей есть большой спрос на данную категорию специалистов, но окончательно это можно утверждать после изучения списка вакансий.

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

Текст запроса:

```
(SELECT -- основной запрос
    ca.desirable_occupation,
    ca.age
FROM hh.candidate ca
    JOIN hh.city ci ON ca.city_id = ci.id
WHERE
    ci.title = 'Санкт-Петербург'
    AND ca.age between 16 AND 21 -- возраст от 16 до 21 включительно
ORDER BY ca.age
LIMIT 10)
union all
SELECT -- выводим строку 'Total' с общим кол-вом по тем же параметрам
    'Total',
    COUNT(ca.id)
FROM
    hh.candidate ca
    JOIN hh.city ci ON ca.city_id = ci.id
WHERE
    ci.title = 'Санкт-Петербург'
    AND ca.age between 16 AND 21
```

Результат:

desirable_occupation ▾	age ▾
Системный администратор	16
Junior Разработчик C++/C#	18
Программист	18
Junior Data Scientist	18
Руководитель web-разработки	18
Специалист по IT	18
Unity3D developer Junior/middle	18
HTML-верстальщик	18
3D-дизайнер	18
Java-разработчик	18
Total	161

Выводы:

Всего в данной выборке 161 кандидат, это 3,26% всех кандидатов из Санкт-Петербурга. По данным результатам можно оценить на какие должности претендуют кандидаты скорее всего без высшего образования, сразу после школ, специальных курсов или студенты вузов.

Примечание к заданию 4.2 :

Возможно, задание лучше переписать, т.к. у многих пользователей есть вопросы к формулировке.

Как вариант:

Для заказчиков из Санкт-Петербурга нам необходимо собрать список 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. Вывести первые 10 строк, а также добавить строку Total с общим количеством таких кандидатов.

Общий вывод по проекту:

Из рассмотренного видно:

1. Данные требуют очистки, это видно на примере возраста претендентов, наверняка есть и другие некорректные данные.
2. Размер рынка труда напрямую зависит от численности населения в определённом городе и количества рабочих мест. Все эти три фактора связаны между собой. Исходя из этого столичные города с большим отрывом находятся на лидирующих позициях.
3. Можно также говорить о популярности той или иной специальности, в разрезе географического положения или возраста претендентов, если сделать дополнительные запросы.
4. Также мы увидели, что некоторые позиции (например: вахтовый метод) находятся в явном дефиците, что может повлиять на предложения оплаты труда и возможность старта карьеры для молодых специалистов.
5. Если добавить к нашим данным ещё и таблицы вакансий, то можно было бы увидеть, где и каких специалистов избыток, а где дефицит. На сколько отличаются требования по оплате труда от предложений работодателей в разрезе географии претендентов и вакансий, это могло бы повлиять на политику найма работодателей и на их поиск новых сотрудников.
6. Также можно добавить данные по учебным заведениям, чтобы прогнозировать недостаток или избыток необходимых специалистов в будущем.

В итоге - чем большим количеством качественных данных мы можем оперировать, тем больше полезных выводов из них мы можем сделать, прогнозов на будущие ситуации в данной области, определения путей развития и т.п.